

Máster en Business Intelligence y Big Data
2019/2020

PROYECTO FIN DE MÁSTER

GeoPymes

RECOMENDADOR PARA APERTURA DE PYMES



AUTORES

*Nuria Turégano Schirmer
Rafael Acosta Sandoval
Elena Gutiérrez Alonso
Matilde Mondéjar Montagud*

TUTOR

Rafael Zambrano López

Índice

| | |
|---|-----------|
| 1. Introducción | 4 |
| 2. Planteamiento del problema | 5 |
| 2.1. La situación de la pyme en España | 5 |
| 2.2. Identificación de las causas | 8 |
| 2.3. Objetivo, oportunidad y motivación de la propuesta | 10 |
| 3. Validación de hipótesis | 11 |
| 3.1. Identificación de hipótesis a validar | 11 |
| 3.2. Proceso de validación | 12 |
| 3.3. Análisis y conclusiones derivadas de las encuestas y entrevistas | 13 |
| 3.3.1. Encuestas | 13 |
| 3.3.2. Entrevistas | 15 |
| 3.4. Análisis preliminar de datos | 16 |
| 3.5. Análisis del entorno | 19 |
| 3.5.1. Factores político-jurídicos | 19 |
| 3.5.2. Factores culturales | 20 |
| 3.5.3. Factores económicos | 20 |
| 3.5.4. Factores socio-demográficos | 24 |
| 3.5.5. Factores tecnológicos | 26 |
| 3.6. Análisis de los competidores | 27 |
| 4. Analisis y diagnostico / Plan estrategico-acción | 31 |
| 4.1. Analisis y diagnostico (DAFO) | 31 |
| 4.2. Definición del modelo de negocio | 34 |
| 4.2.1. Propuesta de valor | 35 |
| 4.2.2. Socios clave | 36 |
| 4.2.3. Recursos clave | 36 |
| 4.2.4. Segmentos de mercado | 37 |
| 4.2.5. Relaciones con clientes | 37 |
| 4.2.6 Canales de distribución | 38 |
| 4.2.7. Actividades Clave | 38 |
| 4.2.8. Estructura de costes | 39 |
| 4.2.9. Fuentes de ingresos y beneficios | 41 |
| 4.3. Plan de acción | 43 |
| 4.3.1 Definición del alcance del proyecto: objetivos y métricas | 43 |
| 4.3.2 Análisis de actividades: modelo lógico - arquitectura técnica | 45 |
| 4.3.2.1 Modelo lógico | 45 |
| 4.3.2.1.1. Identificación de los datos | 45 |
| 4.3.2.1.2. Extracción, transformación y validación de los datos (ETL) | 46 |
| 4.3.2.1.3. Almacenamiento de los datos | 49 |

| | |
|---|------------|
| 4.3.2.1.4. Análisis exploratorio y modelado | 49 |
| 4.3.2.1.5. Visualización | 57 |
| 4.3.2.2 Solución tecnológica: arquitectura técnica | 61 |
| 4.3.3 Análisis de recursos: talento humano y recursos físicos | 62 |
| 4.3.3.1 Talento humano | 62 |
| 4.3.3.2. Recursos físicos | 64 |
| 4.3.3.3. Recursos financieros | 65 |
| 4.3.4 Gestión del tiempo (cronograma) | 65 |
| 5. Rentabilidad | 65 |
| 5.1. Ingresos | 66 |
| 5.2. Gastos | 67 |
| 5.3. Resultados Financieros | 70 |
| Bibliografía | 71 |
| Anexo I: Encuestas para la validación de hipótesis | 75 |
| Emprendedores | 75 |
| Administraciones Públicas | 87 |
| Entidades bancarias | 95 |
| Anexo II: Código de la aplicación | 103 |

1. Introducción

Cuándo se va a crear una empresa, hay varias decisiones estratégicas que se han de tomar para asegurar la viabilidad del negocio. Tras tener claro **qué** se quiere hacer (idea de negocio) y **cómo** se va a hacer (plan de negocio), el tercer paso es decidir el **dónde**, encontrar el lugar adecuado para iniciar la actividad.

La localización de un negocio es un punto fundamental en referencia a la viabilidad económica de la empresa. Una buena ubicación estratégica puede llegar a ofrecer muchos beneficios a la empresa o, por el contrario, llevarla a fracasar.

A la hora de elegir un local en el que llevar a cabo la actividad empresarial, no se puede basar dicha elección en creencias o corazonadas que no estén fundadas en razonamientos lógicos y beneficiosos para la empresa. Por ello, hay que saber la importancia de la localización y tener en cuenta una serie de criterios o factores para decidir dónde hay que instalar la sede del negocio:

- A. **Factores de localización industrial:** en el ámbito industrial, sobre todo las fábricas y las empresas manufactureras, se basan sobre todo en cuestiones logísticas que les permitan realizar sus procesos productivos con normalidad.
 - 1. **Accesibilidad a materias primas:** estar cerca de las materias primas necesarias ayudará a tener un mayor control sobre ellas y reducir el coste de aprovisionamiento.
 - 2. **Mano de obra:** es necesario estar en un lugar donde se pueda encontrar mano de obra cualificada acorde al desarrollo de la actividad.
 - 3. **Red de comunicaciones y transportes:** se necesita tener unas buenas comunicaciones con las redes de carreteras y una buena accesibilidad para conseguir una logística y distribución óptimas.
 - 4. **Ayudas o beneficios fiscales:** si las corporaciones locales ofrecen ayudas fiscales a las empresas para que lleven los negocios a sus municipios, es probable que muchas compañías opten por establecerse en dichas localidades.
- B. **Factores de localización comercial:** para las empresas del sector servicios, las cuales dependen mayormente del lugar en el que se encuentren y la accesibilidad que tengan los establecimientos para sus clientes, tener una buena localización es vital.
 - 1. **Coste del establecimiento:** las mejores localizaciones son las más caras, sin incluir el coste de la posible reforma para adecuar el local a la actividad.
 - 2. **Proximidad a la demanda:** estar cerca de la demanda y del target de clientes es imprescindible para que elijan los productos y servicios ofrecidos.
 - 3. **Visibilidad del local:** un negocio difícil de encontrar o que no destaque no atraerá a la clientela, que optará por otras alternativas para satisfacer sus necesidades.
 - 4. **Comunicaciones y facilidad de acceso:** si es difícil acceder a un negocio o empresas, los clientes pueden optar por otras opciones.
 - 5. **Actividades complementarias:** si alrededor de la empresa existen otros tipos de actividades, puede ayudar a aumentar el número de clientes en nuestro negocio.

Por otro lado, hay que conocer al segmento de clientes al que se quiere captar: de nada sirve tener el local más céntrico de una capital, si nuestra clientela potencial no se mueve en las inmediaciones. Hay que conocer si el producto o servicio que se ofrece es el adecuado para el tipo de cliente que se encuentra en la zona donde se quiere ubicar el negocio.

Geopymes es una startup que surge de la necesidad de aquellos emprendedores de recibir información y asesoramiento a la hora de elegir el emplazamiento geográfico de su negocio.

Su **misión** es la de ayudar a la hora de tomar decisiones referentes a la mejor ubicación de una empresa de cara a que la inversión sea rentable y le asegure el éxito a largo plazo. Para ello, se sirve de las últimas aportaciones de la Inteligencia de Negocio y del Big Data para ofrecer a sus clientes una solución adecuada a sus requerimientos.

2. Planteamiento del problema

2.1. La situación de la pyme en España

Las pymes desempeñan un papel fundamental en el crecimiento económico de un país. Su relevancia implica que cualquier política orientada a la mejora del posicionamiento en el entorno económico global, tome en consideración de forma prioritaria a las pequeñas y medianas empresas. Por tanto, entender la evolución económica de España de los últimos años, es comprender los desafíos a los que se enfrenta la pyme.

La crisis económica y financiera que afectó a nuestro país en 2008, se reflejó en las pymes españolas, consecuencia de baja productividad, tamaño reducido, limitada capacidad exportadora, poca inversión en innovación, escasa digitalización, reducida capacidad para generar empleo estable... Los efectos de esta crisis hicieron más urgente reorientar el modelo de competitividad y crecimiento. En los últimos años, las Administraciones Públicas han llevado a cabo distintas actuaciones con el objetivo de impulsar la recuperación económica y la creación de empleo, como el cumplimiento de la Small Business Act (SBA), recomendación que reúne las líneas para potenciar la creación, el desarrollo y el crecimiento de la pyme en la Unión Europea.

La crisis desatada por el coronavirus marcará un antes y un después en muchos aspectos de nuestra sociedad. En el ámbito empresarial también lo hará, especialmente en la pequeña y mediana empresa, en el que se estima un cierre en torno al 20-30% del total actual.

Muchos expertos coinciden en que esta situación ha acelerado el proceso de digitalización de las pymes en España, y que la tecnología será el mejor aliado de este tipo de organizaciones para salir adelante ante esta crisis económica.

¿Qué criterios deben cumplirse necesariamente para ser considerado pyme?



En España, las pymes suponen el 99,8% de las empresas. Las pymes sin asalariados son mayoritarias y las microempresas se ponen a la cabeza de las pymes asalariadas.

| Empresas por tamaño | Número de empresas | Tasa de variación % | |
|---|--------------------|---------------------|--------------|
| | | mensual | anual |
| PYME (0-249 asalariados) | 2.867.732 | 0,49 | -0,09 |
| PYME sin asalariados (0 asalariados) ² | 1.539.708 | 0,00 | -0,68 |
| PYME con asalariados (1-249 asalariados) | 1.328.024 | 1,06 | 0,60 |
| Microempresas (1-9 asalariados) | 1.144.636 | 0,96 | 0,22 |
| Pequeñas (10-49 asalariados) | 157.765 | 1,90 | 2,90 |
| Medianas (50-249 asalariados) | 25.623 | 0,53 | 3,61 |
| Grandes (250 o más asalariados) | 4.959 | 2,14 | 4,93 |
| Total Empresas | 2.872.691 | 0,49 | -0,09 |

Gráfico 1.
Distribución de empresas por tamaño

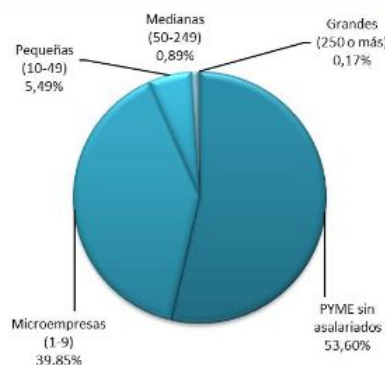


Gráfico 2.
Evolución de las empresas por tamaño

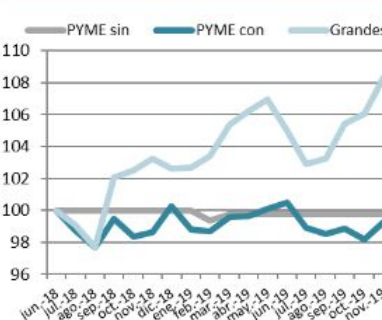


Figura 1: Distribución de número de empresas por tamaño. Fuente de datos: Informe 'Cifras pyme' (Ministerio de Industria, Comercio y Turismo)

Y generan el 66% del empleo empresarial total:

| Empresas por tamaño | Empleo | Tasa de variación % | |
|---|-------------------|---------------------|-------------|
| | | mensual | anual |
| PYME (0-249 asalariados) | 10.592.278 | 1,69 | 4,98 |
| PYME sin asalariados (0 asalariados) ³ | 1.539.708 | 0,00 | -0,68 |
| PYME con asalariados (1-249 asalariados) | 9.052.570 | 1,25 | 2,32 |
| Microempresas (1-9 asalariados) ⁴ | 3.466.704 | 1,39 | 0,80 |
| Pequeñas (10-49 asalariados) | 3.076.535 | 1,69 | 3,09 |
| Medianas (50-249 asalariados) | 2.509.331 | 0,53 | 3,52 |
| Grandes (250 o más asalariados) | 5.683.057 | 2,26 | 4,98 |
| Total Empleo | 16.275.335 | 1,48 | 2,94 |

Gráfico 3.
Distribución del empleo por tamaño.

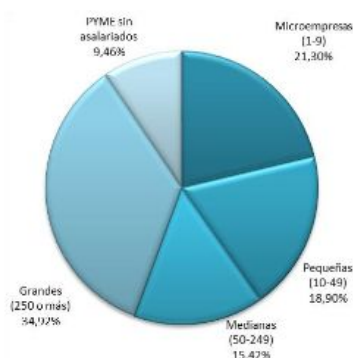


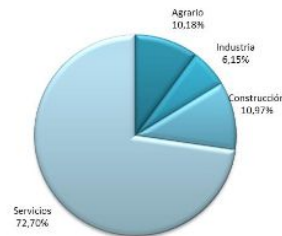
Tabla 3.
Empleo medio por tamaño de empresa

| Empresas por tamaño | Nov 2018 | Nov 2019 |
|----------------------|----------------|----------------|
| PYME | 3,6 | 3,7 |
| PYME sin asalariados | 1,0 | 1,0 |
| PYME con asalariados | 6,7 | 6,8 |
| Microempresas | 3,0 | 3,0 |
| Pequeñas | 19,5 | 19,5 |
| Medianas | 98,0 | 97,9 |
| Grandes | 1.145,5 | 1.146,0 |
| Empleo medio | 5,5 | 5,7 |

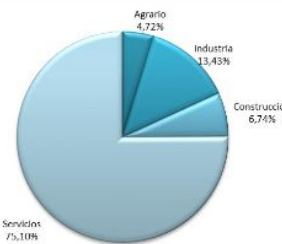
Figura 2: Número de empleos generados por tipo de empresas. Fuente de datos: Informe 'Cifras pyme' (Ministerio de Industria, Comercio y Turismo)

Tabla 4. Empresas por sectores

| Sectores | Empresas | Tasa de variación % | |
|-----------------------|------------------|---------------------|--------------|
| | | Mensual | Anual |
| Agrario | 292.422 | 3,06 | 2,02 |
| Industria | 176.711 | 0,23 | -1,20 |
| Construcción | 315.085 | 0,27 | 0,86 |
| Servicios | 2.088.473 | 0,20 | -0,42 |
| Total Empresas | 2.872.691 | 0,49 | -0,09 |



| Sectores | Empleo | Tasa de variación | |
|---------------------|-------------------|-------------------|-------------|
| | | Mensual | % Anual |
| Agrario | 768.811 | 8,85 | 1,17 |
| Industria | 2.186.045 | 0,88 | 2,18 |
| Construcción | 1.096.967 | 0,79 | 2,77 |
| Servicios | 12.223.512 | 1,22 | 3,20 |
| Total Empleo | 16.275.335 | 1,48 | 2,94 |



| Tabla 6. Tamaño medio empresarial por sectores | | | | | |
|--|---------|-----------|--------------|-----------|-------|
| | Agrario | Industria | Construcción | Servicios | Total |
| Noviembre 2019 | 2,6 | 12,4 | 3,5 | 5,9 | 5,7 |
| Noviembre 2018 | 2,7 | 12,0 | 3,4 | 5,6 | 5,5 |

Explotación Estadística del Directorio Central de Empresas, Provincias, Total, 2019

| Provincia | Número de Empresas |
|----------------|--------------------|
| 02 Albacete | 18,667 |
| 03 Almería | 144,455 |
| 04 Asturias | 18,667 |
| 05 Baleares | 18,667 |
| 06 Barcelona | 497,779 |
| 07 Bizkaia | 18,667 |
| 08 Burgos | 18,667 |
| 09 Cádiz | 18,667 |
| 10 Cantabria | 18,667 |
| 11 Castellón | 18,667 |
| 12 Ciudad Real | 18,667 |
| 13 Córdoba | 18,667 |
| 14 Cuenca | 18,667 |
| 15 Girona | 18,667 |
| 16 Granada | 18,667 |
| 17 Guipúzcoa | 18,667 |
| 18 Huelva | 18,667 |
| 19 Jaén | 18,667 |
| 20 León | 18,667 |
| 21 Lugo | 18,667 |
| 22 Madrid | 560,001 |
| 23 Málaga | 144,455 |
| 24 Murcia | 18,667 |
| 25 Navarra | 18,667 |
| 26 Ourense | 18,667 |
| 27 Palencia | 18,667 |
| 28 Salamanca | 18,667 |
| 29 Sevilla | 18,667 |
| 30 Sorja | 18,667 |
| 31 Tarragona | 18,667 |
| 32 Teruel | 18,667 |
| 33 Toledo | 18,667 |
| 34 Valencia | 186,667 |
| 35 Valladolid | 18,667 |
| 36 Zamora | 18,667 |
| 37 Ceuta | 18,667 |
| 38 Melilla | 18,667 |

Página 7

Es importante resaltar que la condición jurídica más mayoritaria a nivel nacional son las personas físicas seguida de Sociedades de responsabilidad.

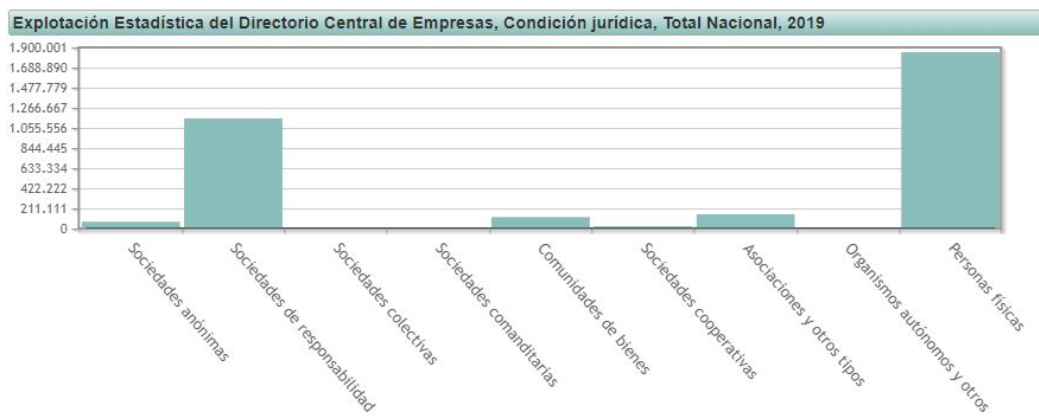


Figura 5: Empresas por tipo de condición jurídica. Fuente de datos: INE

El emprendimiento es uno de los sectores en auge en nuestro país y en todo el mundo. En la actualidad, nueve mujeres inician negocios por cada diez hombres españoles, una cifra que supera la media de Europa. La edad media del emprendedor en España se sitúa en torno a los 38,4 años. Algo que, en los próximos años, irá bajando por la entrada de más jóvenes emprendedores. Precisamente, el 35% de los menores de 30 años en la actualidad se verían emprendiendo un negocio por su cuenta. Casi el 50% de estos emprendedores en España tienen formación universitaria. A pesar de esto, esta cifra quizás comience a remitir, ya que hay muchos jóvenes que piensan que no es necesario disponer de una titulación de este tipo para iniciar un negocio. Un reciente estudio asegura que las ciudades españolas con mejores condiciones para el emprendimiento son Madrid, Barcelona y Valencia. La tasa de actividad emprendedora en España está situada en el 6,4%, aún por debajo del promedio europeo (8,4%). Por el contrario, la tasa de personas involucradas en abandonos empresariales se mantuvo más o menos estable en el último año, en torno al 1,6%.

Los principales obstáculos que hallan los emprendedores para poder iniciar sus proyectos son la financiación, políticas gubernamentales, cultura emprendedora y falta de información. Hace falta que los potenciales emprendedores se encuentren con un camino más llano y más oportunidades para desarrollar su negocio.

2.2. Identificación de las causas

A continuación analizaremos algunas de las causas que impiden a una pyme llegar a mantenerse en el mercado, sucumbiendo tarde o temprano ante sus competidores o ante su misma mala administración que la empuja a cerrar sus operaciones rápidamente.

Para el análisis se ha decidido utilizar el diagrama de Ishikawa (véase Figura 6), conocido también como Diagrama de Causa y Efecto, de manera que nos permita llegar a identificar el principal problema que tienen las pymes en sus primeros días de vida y que influye fuertemente en su posterior fracaso.

Se han identificado varias categorías de problemas, y en cada una de esas categorías se han enumerado las posibles causas que provocan estos problemas. Las categorías son las siguientes:

- Información
- Instituciones

- Asesoría
- Tecnología
- Emprendedores
- Mercado

Cada categoría tiene varias sub-causas que impactan directamente en esta y que al final todas influyen directa o indirectamente en el problema principal o efecto que es la *“alta incertidumbre y desconocimiento del mercado al momento de iniciar un negocio con un impacto fuerte en el fracaso del mismo”*.

Podemos ver como el poco o mal uso de la tecnología por una pyme influye de forma negativa en su falta de visibilidad de variables importantes en su entorno que le permita identificar potenciales problemas y ayude en la formulación de un plan de negocios más acorde a las condiciones actuales del mercado en el que está inmersa.

Por otro lado, la falta de planificación que conlleva a no formular planes de negocio realistas y consistentes hace extremadamente difícil la supervivencia de la pyme, ocasionando su cierre prematuro.

Si a lo anterior sumamos la falta de experiencia de muchos emprendedores en varias áreas claves de negocios como la planificación, gestión y manejo del recurso humano, es muy difícil asegurar la continuidad de la empresa más allá de los cinco años.

La asesoría tecnológica es bastante útil para las pymes pero se ve que realmente es un aspecto que debe mejorarse y de esta manera guiar a estas empresas a hacer un uso más eficiente de sus recursos financieros al momento de invertir en soluciones informáticas.

Para finalizar, aunque la mediana empresa en España solamente es del 0,6% del tejido empresarial total, la misma es muy importante para el país, ya que, más de un 13% de los asalariados trabajan para una de estas, sin embargo, tienen un problema serio al carecer de suficiente apoyo gubernamental para seguir creciendo y pasar a ser una empresa consolidada.

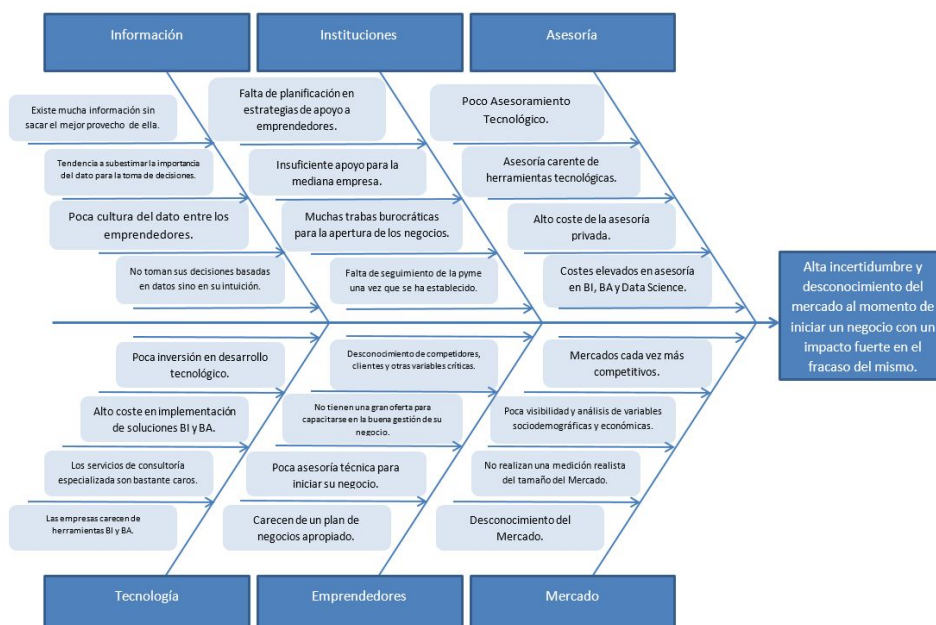


Figura 6: Diagrama de Causa-Efecto de Ishikawa. Fuente: Elaboración propia.

A lo largo del desarrollo de este proyecto se trabajará en eliminar o atenuar esas causas que provocan la alta incertidumbre que muchas veces paraliza a las pymes en sus inicios. Asimismo, se mejorará el análisis y visibilidad de las variables del entorno a través de herramientas de Inteligencia de Negocios y Big Data.

2.3. Objetivo, oportunidad y motivación de la propuesta

En el contexto actual de crisis económica desencadenada por la pandemia del coronavirus, donde las pymes, uno de los principales agentes económicos, será duramente golpeado, nos planteamos cómo podríamos contribuir a minimizar este impacto mediante el uso del “BI & Big Data”, de cara a contribuir en la mejora de la creación de empleo, el desarrollo económico y el bienestar social.

Hoy día existen multitud de fuentes de datos públicas de todo tipo (INE, Ayuntamientos, Cámaras de Comercio, estudios publicados en Internet, Google maps, ...) que ponen a disposición de quien lo necesite, información de diferente naturaleza (variables sociodemográficas, económicas, urbanísticas, de ubicación, movilidad, etc).

Analizando los factores que hacen que las empresas tengan éxito, se observa que la **ubicación** es uno de los más determinantes, ya que ésta determina en gran medida su clientela, que la fijará tanto las variables demográficas y económicas de su perímetro geográfico como variables de acceso y comunicaciones. Así mismo, la ubicación determinará la relación con competidores y proveedores.

Un estudio de la Comunidad de Madrid determina que un mal emplazamiento provoca el 70% del cierre de los comercios.

De ahí que sea clave, antes de iniciar la apertura de una actividad, realizar un estudio sobre cuál es la mejor ubicación posible para ese tipo de negocio, teniendo en cuenta nuestros potenciales clientes, accesos, competencia, etc.

Las grandes empresas saben que el factor ubicación es clave y por ello invierten en estudios de este tipo para decidir dónde abrir o dónde no. Con esta información de valor, lamentablemente, no suelen contar las pequeñas y medianas empresas, pues muchas de ellas no tienen ni los recursos ni el tiempo necesario para poder realizar estos estudios.

Por otro lado, consideramos que sería de gran valor poder hacer un estudio de estas características, no sólo para un cliente particular en un momento puntual de apertura o expansión de una actividad, sino en modo *macro*, de forma que se puedan tomar decisiones de apertura de actividades que satisfagan las necesidades tanto de emprendedores como de ciudadanos, beneficiando así el desarrollo económico a la vez que el bienestar social.

El **objetivo** de nuestra propuesta es cubrir esta falta de asesoramiento, poner al alcance de las pymes esta información clave para sus negocios.

Para ello, se desarrollará una herramienta online que recomiende qué tipo de comercio puede ser más apropiado y en qué ubicación puede tener una mayor probabilidad de éxito.

Esta recomendación se basará en:

- Potenciales clientes: basándose en el estudio de variables sociodemográficas, económicas, urbanísticas, accesos a redes de transporte y comunicaciones, etc.
- Competidores directos o indirectos existentes.

Los clientes potenciales identificados se dividen en tres líneas:

- Emprendedores: que puedan consultar a título personal en esta herramienta su propuesta de negocio y obtener la información relacionada. Su incertidumbre se verá reducida y aumentarán sus posibilidades de éxito.
- Administraciones Públicas: organismos de asesoramiento y apoyo al emprendedor, quienes dispondrán de esta herramienta para poder dar soporte de asesoramiento a los emprendedores que les soliciten información para iniciar una actividad. Tendrán una visión más amplia de las oportunidades de negocio de sus poblaciones.
- Empresas privadas: aquellas que vean de interés esta información de cara a mejorar su toma de decisiones. A modo de ejemplo: bancos (proceso de concesión de financiación), inmobiliarias (recomendar apertura de tipo de negocio en los locales en alquiler disponibles), empresas en modalidad de franquicia, etc.

Con esta propuesta pretendemos aportar un servicio de valor, partiendo de la información oculta en los datos publicados en abierto por las distintas organizaciones, que mejore el proceso de toma de decisiones de los diferentes agentes descritos, contribuyendo con ello a mejorar el éxito de las pymes y como consecuencia el desarrollo económico y social.

3. Validación de hipótesis

3.1. Identificación de hipótesis a validar

La identificación de las hipótesis representa un elemento fundamental una vez detectado el problema de nuestro proyecto.

Vamos a tratar de responder de forma amplia a las dudas que nos surgen acerca de la relación que existe entre las variables consideradas (sociodemográficas, económicas y urbanísticas). Para ello realizaremos ciertas suposiciones sobre los resultados de nuestra tesis que nos permitirán orientar el trabajo a una conclusión determinada.

En este proceso de investigación se propone comprobar mediante experimentación las siguientes afirmaciones:

1. El 80% del éxito de un negocio se basa en su ubicación (tipo de población, transporte, negocios cercanos, accesibilidad...)
2. Al menos 6 de cada 10 clientes desconocen la zona apropiada para abrir una PYME.
3. Al menos 6 de cada 10 clientes consideran valioso recibir asesoramiento sobre la “ubicación” en la que establecer su negocio basándose en variables sociodemográficas, económicas y urbanísticas de las zonas más propensas al éxito de un negocio.
4. Al menos 6 de cada 10 clientes valoran positivamente tener una herramienta que les pueda brindar una sugerencia sobre qué “tipos de negocio” son más apropiados para una ubicación específica.
5. Al menos 5 de cada 10 emprendedores pagarían por un servicio de asesoramiento sobre la ubicación de su negocio.

3.2. Proceso de validación

Para validar la primera hipótesis, nos hemos basado en estudios de mercado ya realizados donde se ha determinado que la ubicación es un factor clave al momento de asegurar el éxito de un negocio y que por lo menos ésta determina en un 80% las probabilidades de éxito que éste tiene.

En el caso de los sistemas de franquicias se ha comprobado mediante estudios con empresas especializadas en Location Analytics y Big Data como T4 Franquicias, que no hay un aspecto tan fundamental para que un negocio sea sostenible en el tiempo como una buena ubicación y no es exagerado afirmar que el 80% del éxito dependa de que se ubique en la zona adecuada, por lo que este tipo de negocios busca ofrecer a sus nuevos franquiciados información relevante sobre capacidad de gasto de la población cercana a la zona de apertura de cada establecimiento, así como su nivel de educación o la reputación online que tienen los locales que les rodean.

También, según el informe anual “El Futuro del Retail Intelligence 2019” elaborado por la empresa Location Intelligence Geoblink, el 85% de los retailers considera a la ubicación, un factor vital para el éxito de su negocio, después de encuestar alrededor de 600 profesionales en España, Reino Unido y Estados Unidos, donde dicho estudio constató que el 37% de los establecimientos en 2018 cerraron debido a una ubicación deficiente.

De cara a validar las siguientes hipótesis, se realizaron encuestas y entrevistas a los 3 segmentos de clientes identificados: emprendedores, Administraciones Públicas y entidades bancarias por ser la muestra de población más representativa.

Para determinar el número de encuestas a realizar para validar las hipótesis empleamos la herramienta SurveyMonkey.

Calculate your sample size

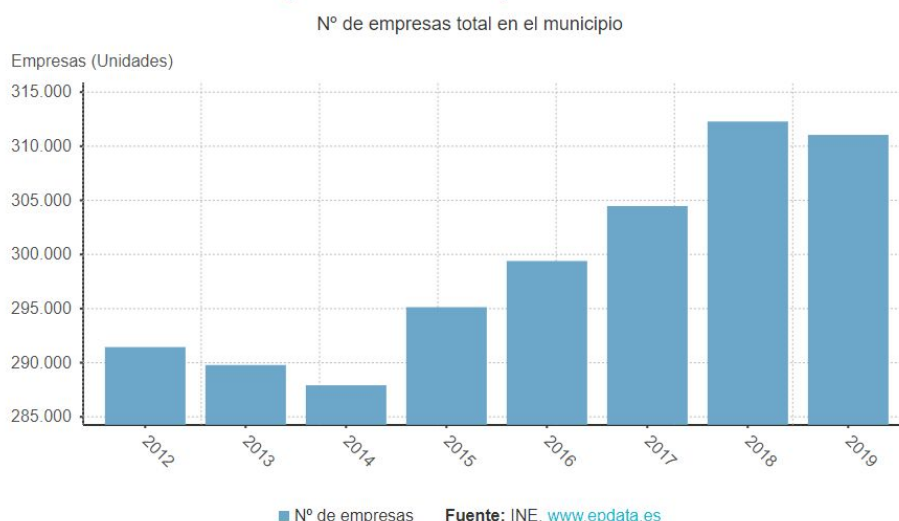
| Population Size ⓘ | Confidence Level (%) ⓘ | Margin of Error (%) ⓘ |
|-------------------|------------------------|-----------------------|
| 311000 | 90 | 10 |

Sample size

69

Se toma como tamaño de la muestra el número de empresas en el municipio de Madrid (311.000 tal y como puede verse en la imagen más abajo). Las Administraciones Públicas que prestan servicio de asesoramiento son pocas de modo que quedarán incluidas en este valor por no provocar una variación significativa del mismo.

¿Cuál es el número de empresas en Madrid?



En base a criterios de homogeneidad de los diferentes segmentos, decidimos que las 69 encuestas/entrevistas a realizar se distribuirán del siguiente modo:

- Emprendedores / Pymes: 55 (personas con idea de emprender o ya emprendedores)
- AAPP: 5 (Emprendedores Madrid, Juntas de Distrito, Instituto Mujer y Cámara Comercio)
- Entidades bancarias: 10 (BBVA, Banco Sabadell, Caixabank, Santander...)

Para la realización y medición de las diferentes encuestas se ha utilizado la aplicación Google Forms.

3.3. Análisis y conclusiones derivadas de las encuestas y entrevistas

El detalle de las preguntas de las encuestas realizadas y los resultados obtenidos por segmento puede consultarse en [Anexo I](#)

3.3.1. Encuestas

A modo de resumen se destacan las principales conclusiones en las diferentes tipologías:

Emprendedores

A pesar que el 64% de los encuestados tuvieron clara la ubicación de su negocio antes de iniciarlo, sólo el 36% de ellos se asesoró en la elección de esta. Dicha asesoría se tuvo en un 80% a través de empresas especializadas a través de un servicio presencial en un 55% y el resto a través de medios digitales.

Se les consultó si el asesoramiento recibido fue prestado por una entidad pública, el 76% indicó que no y solo un 24% indicó que sí.

Por ende, la **hipótesis 2** queda validada ya que los clientes carecen de un servicio de asesoramiento sobre la ubicación al momento de iniciar su negocio.

Por otra parte, cuando se les consultó cómo valoraría una herramienta online que les guiara sobre la elección de la ubicación de su negocio, el 91% de ellos la valoran positivamente. Es importante hacer notar que el 80% de ellos indicaron que no recibieron información sobre datos sociodemográficos, económicos ni urbanísticos de la zona elegida para su establecimiento

comercial. Asimismo, consideran que estos factores sociodemográficos, económicos y urbanísticos, son el segundo factor condicionante sobre la ubicación de un negocio, solo por debajo del coste del alquiler del local, tasas y licencias. De esta manera, la **hipótesis 3** queda validada.

Cuando se consultó entre los encuestados, sobre la valoración que le darían a una herramienta online que les ofreciera una recomendación según el tipo de industria más apropiada para invertir en una zona específica de la ciudad, el 89% de ellos le otorgaron una valoración positiva, validándose así la **hipótesis 4**.

Es importante mencionar que los emprendedores valoran de mejor manera una herramienta que les sugiera la mejor ubicación de un negocio, donde el 71% le otorgaron la máxima calificación para ello, a una herramienta que les sugiera el tipo de industria más apropiado para invertir en base a una ubicación, donde solo el 53% de ellos le otorgaron la máxima calificación.

Está claro que los emprendedores están ávidos de una herramienta con las características mencionadas anteriormente, ya que el 78% de ellos están dispuestos a pagar por una que le ofrezca la mejor ubicación de acuerdo al tipo de negocio que tienen en mente y un 62% de ellos están dispuestos a pagar por una herramienta que les sugiera el tipo de industria más acorde a una zona específica de la ciudad. La **hipótesis 5** es validada de esta forma.

Las formas de pago por orden de preferencia que mejor se ajustan a las necesidades de los clientes son: Servicio On Demand, Suscripción Mensual y Suscripción Anual.

Administraciones Públicas

La mala ubicación, alta incertidumbre del mercado, desconocimiento del mercado potencial y poca digitalización encabezan los factores que pueden afectar negativamente a un negocio con un 80%.

En relación a los factores que afectan negativamente al éxito de un negocio, destacar que el 100% de los encuestados consideran que los factores sociodemográficos están muy relacionados, en un 80% el coste de alquiler del local y la cercanía a negocios complementarios y un 60% le da importancia a la cercanía a calles principales de la ciudad y competidores.

No se ofrece ningún tipo de asesoramiento, orientación o recomendación por parte de los organismos públicos sobre cuestiones como la zona más apropiada en la que abrir un negocio. Validamos en este punto la **hipótesis 2**.

El 40% de las AAPP encuestadas creen que sería muy necesario por parte de quien emprende recibir algún tipo de asesoramiento por su parte sobre la ubicación del negocio, otro 20% poco necesario y el 40% se mantiene neutral.

A pesar que no todos los encuestados creen que debería hacerse desde las AAPP, el 100% de los encuestados valoran bien o muy bien que los emprendedores reciban asesoramiento sobre la “ubicación” en la que establecer su negocio basándose en variables sociodemográficas, económicas, urbanísticas de las zonas más propensas al éxito de un negocio. Y el 80% lo hace sobre el tipo de negocio quedándose el 20% neutral. Estos resultados validan las **hipótesis 3 y 4**.

El 80% de las Administraciones Públicas encuestadas pagarían por un servicio de asesoramiento sobre la ubicación de su negocio eligiendo forma de pago anual el 75%. Queda validada por tanto la **hipótesis 5**.

Entidades bancarias

La mala ubicación se pone a la cabeza de los factores que pueden afectar negativamente a un negocio (82%), seguido de la alta incertidumbre del mercado (73%) y del desconocimiento del mercado potencial (64%).

El 64% de los emprendedores que acuden a una entidad bancaria tiene clara la ubicación de su negocio desde el inicio pero no reciben ningún asesoramiento sobre la zona idónea para abrir su negocio. Validamos por tanto la **hipótesis 2**.

Los factores que condicionan la ubicación de un establecimiento son mayoritariamente que sea accesible por transporte privado con fácil estacionamiento o que esté cercano a una calle principal de la ciudad (ambas con el 64%)

El 91% de los encuestados creen necesario o muy necesario que los emprendedores reciban algún tipo de asesoramiento sobre la ubicación o el tipo de comercio más propicio para su negocio y si existiera una herramienta que asesorara en este sentido el 100% la ofrecería a sus clientes como parte de sus servicios. Así mismo la totalidad de los encuestados estaría dispuesto a pagar por esta herramienta en modalidad on demand (54%) o mensual (36%).

Por tanto quedan validadas las **hipótesis 3, 4 y 5**.

3.3.2. Entrevistas

Las entrevistas realizadas nos han podido aportar la visión global necesaria para nuestro proyecto y nos han permitido validar las hipótesis apoyando por tanto el resultado de las encuestas.

Emprendedor: Fundadora tienda alquiler y reparación de bicicletas en zona Retiro

Tras una larga conversación, en relación a la elección de la ubicación de su empresa, indica que la idea surgió a raíz de ver que en la zona donde vivía no existía ningún negocio que ofreciera este servicio y vió un potencial claro para esta actividad. En esa época, hace 10 años, no existía como ahora la filosofía de circular en bici, ni los puntos de alquiler de bicicletas públicos y privados que ahora son tan comunes. Estando cerca del parque del Retiro vio clara la oportunidad de éxito. Realizó un amplio estudio de mercado de la zona y al resultar positivo, se decidió a abrir.

No solicitó asesoramiento en ningún organismo público ni privado a la hora de iniciar la actividad pero indica que habría valorado muchísimo la existencia de una herramienta que le recomendara la mejor ubicación para su negocio. No duda en que hubiera pagado por una información tan valiosa; poder conocer de entre todas las zonas, la más apropiada para su idea de negocio, no sólo si la zona que se planteó podría ser una buena opción.

Entidad bancaria - Director sucursal Caixabank (oficina Madrid)

Durante la entrevista, explica que las variables “ubicación” y “tipo de negocio” actualmente ya son tenidas en cuenta a la hora de estimar el scoring de riesgo en la concesión de un crédito financiero vinculado a la creación de una empresa.

Sobre la ubicación que les informan los emprendedores en los estudios de mercado que presentan, ellos ya estiman la probabilidad de éxito o no. Considera que un asesoramiento más específico sobre ubicación, debería ser previo a ellos como Entidad bancaria, realizándose bien por parte de las AAPP o por parte de los propios emprendedores.

A pesar de indicar que a título personal le parecería interesante poder tener información más precisa sobre el éxito en la ubicación elegida para iniciar una actividad, cree que la Entidad no invertiría en estos momentos en dicho servicio.

Entidad bancaria - Apoderada comercial del Banco Sabadell (oficina Madrid)

Por su parte, esta entidad bancaria no ofrece servicio de asesoramiento sobre la localización de un negocio, es el cliente el que aporta dentro del plan de negocio su propio estudio de localización. Esta ubicación es tomada en cuenta en el rating interno de la entidad para el cálculo del riesgo de crédito.

Al tenerlo ya contemplado dentro del plan negocio a nivel de entidad bancaria no consideran por su parte pagar por este servicio.

Administración Pública - Área de Empleo Apoyo al emprendedor Ayuntamiento Alcobendas

A pesar de no haber podido realizar una entrevista a organismos del municipio de Madrid, hemos conseguido contactar con uno de los municipios de la Comunidad. El objetivo era conocer la visión desde un servicio público de asesoramiento a emprendedores.

Después de explicarnos los diferentes servicios que ofrecen a los futuros emprendedores, nos confirma que el asesoramiento sobre la ubicación la realizan utilizando hipótesis generales (por ejemplo, si quieren abrir una papelería, les sugieren que intenten abrir cerca de un centro educativo o de una biblioteca o que busquen una zona con población en edad escolar), pero que no disponen de ninguna herramienta que les proporcione esa recomendación concreta calculada mediante la relación de variables sociodemográficas, económicas y urbanísticas.

Manifiesta que sería muy interesante disponer de este tipo de información sobre la ubicación e indica que por su experiencia, le ve menos valor a la recomendación de tipo de negocio ya que el hecho de que un emprendedor cambie de idea no es algo común, van con esa decisión ya meditada y tomada.

En relación a si pagarían por un servicio de este tipo, lo desconoce. Tendría que iniciar los trámites correspondientes con otras áreas responsables para la gestión de este tipo de temas.

3.4. Análisis preliminar de datos

En los últimos 10 años se han dado grandes avances en la filosofía de Datos abiertos, donde cada vez se comparte más información para que esté disponible para todo el mundo de forma gratuita.

Pero... ¿Qué son los datos abiertos?

Datos Abiertos (*open data* en inglés) es una iniciativa global, ligada a las políticas de Gobierno Abierto, que persigue que los datos y la información, en especial las que poseen las Administraciones Públicas, se publiquen de forma abierta, regular y reutilizable para todo el mundo, sin restricciones de acceso, copyright, patentes u otros mecanismos de control.

La filosofía origen de estas iniciativas es fomentar la transparencia, la eficiencia, la participación ciudadana y el desarrollo económico.

Los Datos Abiertos fomentan dar servicio a personas individuales y a empresas, para que puedan utilizar la información pública con las siguientes finalidades:

- realizar simples consultas.

- enriquecer cierta información con nuevos datos (dar un valor añadido).
- generar aplicaciones y servicios.
- crear nuevos negocios.

Además permite dotar de mayor transparencia al conjunto de las Administraciones Públicas mediante:

- la publicación de los datos que poseen las Administraciones en formatos abiertos para generar confianza en las instituciones, porque expone el trabajo que se realiza en las distintas instituciones y muestra cómo se gestionan e invierten los recursos públicos.
- ayudas al desarrollo económico general, a la generación de nuevos sectores y de nuevos servicios para la ciudadanía.

En España en 2009, arrancó la iniciativa Aporta dentro de datos.gob.es, plataforma que supone un punto de encuentro entre las administraciones, las empresas y los ciudadanos que forman parte del ecosistema de los datos abiertos en España. Busca promocionar la apertura de la información pública y desarrollar servicios avanzados basados en datos. Está promovida por el Ministerio de Asuntos Económicos y Transformación Digital y la Entidad Pública Empresarial Red.es.



Figura 7: Mapa conceptual de los beneficios de los datos abiertos. Fuente de datos: datos.gob.es

También podemos obtener otros datos muy útiles como información sobre el catastro o aforo de peatones en <https://datos.madrid.es/> y datos estadísticos sobre las pymes en <https://www.ine.es/>. Podemos también encontrar datos sobre la movilidad de la ciudad en <http://datos-movilidad.crtm.es/>, fuente que utilizaremos para valorar la comunicación de la zona.

Utilizaremos estas fuentes de información para generar valor a las pymes desarrollando la herramienta de un recomendador de mejor ubicación y tipo de negocio.

Con el objetivo de acotar el amplio campo de aplicación que puede tener este servicio, hemos decidido limitar el estudio a la ciudad de Madrid, por ser una de las ciudades con mayor número de emprendedores por habitante y el mayor volumen de negocios/comercios del país.

En este contexto, además de las citadas fuentes, acudiremos a datos abiertos del Ayuntamiento de Madrid ([Portal de datos abiertos del Ayuntamiento de Madrid](#)) para obtener datos relevantes para nuestro estudio.

Este portal está dedicado a promover el acceso a los datos del gobierno municipal e impulsar el desarrollo de herramientas creativas para atraer y servir a la ciudadanía de Madrid.

A continuación listamos información que hemos identificado como necesaria para poder llevar a cabo esta iniciativa:

A. Variables Sociodemográficas

- Edad
- Género
- Densidad población
- Estructura del hogar
- Tipos de familia
- Población económicamente activa
- Nivel de estudios

B. Variables económicas

- Renta per cápita
- Gasto per cápita
- Endeudamiento

C. Variables urbanísticas

- Callejero Madrid
- Medios transporte públicos
- Aparcamientos privados (parking)
- Ubicación de CC. Comerciales
- Instituciones públicas (centros educativos, Ayuntamientos, teatros, bibliotecas...)
- Accesos: escaleras, rampas...
- Información de movilidad en la ciudad

D. Otras variables de interés:

- Número de competidores por zona
- Número de peatones por distrito
- Facturación de los comercios
- Nº de inmuebles en catastro y su uso por barrio

Este tipo de variables resultan interesantes a la hora de determinar la afluencia a un negocio ya que el perfil de personas que vive en una determinada zona influye en el gasto, en el tipo de negocio al que acuden y en definitiva en la manera en el que los ciudadanos transitan por la zona.

La edad y el tipo de estructura familiar serán variables claves, ya que una familia con hijos acude a establecimientos muy diferentes que personas mayores o adolescentes. Por ejemplo una tienda de ropa infantil a priori tendría más afluencia en una zona donde viven familias con hijos.

La manera en la que se llega a un comercio también puede darnos información valiosa, no tendrá la misma afluencia de personas un negocio en un polígono industrial con mala comunicación que una zona comercial peatonal en el centro de la ciudad.

Si bien actualmente los datos están disponibles en “datos abiertos” para la ciudadanía, se tratan de fuentes independientes y se hace necesario procesarla para poder obtener valor de ellos. Será imprescindible la construcción de un dataset propio con todas las variables citadas anteriormente.

Sin embargo, detectamos otro tipo de información que aun siendo tremendamente útil, no la tenemos a nuestro alcance.

Algunos datos que nos serían de utilidad de organismos de asesoramiento al emprendedor podrían ser histórico del número de consultas recibidas, tipo de comercio indicado y ubicación del mismo, perfil del emprendedor, tipo de consultas que realizan...

En cuanto a las Entidades financieras, tener visibilidad de ciertas interacciones con emprendedores como son histórico del número de peticiones de créditos para la apertura de negocios, perfil del emprendedor y tipo de negocio a iniciar, cantidad de la financiación solicitada, porcentaje de concesiones, motivos de rechazo... podrían aportar gran valor al estudio.

3.5. Análisis del entorno

3.5.1. Factores político-jurídicos

Las pymes en España, como en cualquier parte del mundo, están sujetas a una serie de factores políticos y legales que condicionan y limitan su acción en muchos aspectos, los cuales pueden afectar a la empresa en su concepción, operación y desaparición.

Existen normativas que aplican en la creación de un negocio, los impuestos que deben honrar de manera única o periódica, cómo deben de gestionarse los contratos laborales de los trabajadores, pagos de salarios y hasta los despidos.

Los factores políticos juegan un papel esencial en la vida de las pymes, ya que de no ser favorables pueden afectar a estas empresas al grado de propiciar su desaparecimiento forzado en casos extremos derivados de una inestabilidad política por parte de un gobierno.

La inestabilidad política de una región o país puede alejar la inversión extranjera en la zona, ya que al no tener garantías en la operación de la empresa para la cual apoyan, prefieren no asumir tales riesgos y buscan otros horizontes más seguros para rentabilizar su dinero.

Con relación a la normativa aplicable a las pymes, existen muchos reglamentos, leyes o decretos que regulan el funcionamiento de este tipo de negocios. Asimismo, existe normativa relacionada a la creación de empresas, creación de empresas por internet, cese de empresas por internet, emprendedores y PYME, Innovación, ejercicio a la actividad empresarial, normativa laboral, entre otras.

En cuanto al impacto en el desarrollo del proyecto, existen factores políticos y jurídicos que sin duda alguna inciden en muchas variables económicas, financieras o sociodemográficas, y que podría afectar la predicción del comportamiento de estas, afectando el desempeño y exactitud de las recomendaciones de los modelos estadístico-computacionales empleados. Muchos de esos factores fueron ampliados al inicio de este apartado, y en la medida incrementemente la

incertidumbre en el comportamiento de los mismos, se tendrían que hacer ciertos ajustes a los modelos para que estos sigan siendo útiles en el logro de su cometido.

3.5.2. Factores culturales

Existen un sinnúmero de factores culturales que inciden ya sea positiva o negativamente en el nacimiento y desempeño de las pymes, y a medida que potenciamos esos factores que inciden positivamente se puede mejorar la actitud emprendedora en una sociedad.

Comúnmente en el sector de las pymes las empresas familiares son las más predominantes, ya que cerca del 88% de las empresas en el país son de este tipo. Las empresas familiares imprimen al colaborador una identidad, valores y cultura empresarial bastante sólidos, que los convierte a estos en verdaderos embajadores de la marca, creyendo que el hecho de que una familia sea la propietaria de la empresa los diferencia positivamente de sus competidores dándoles al mismo tiempo una ventaja competitiva importante.

En cuanto a las características generales de las pymes en España se pueden mencionar las siguientes:

1. Las pymes están gestionadas por gerentes con una edad media de 52 años y 17 años de experiencia, siendo aproximadamente el 90% hombres. El 49% de ellos tienen estudios universitarios, donde existe una correlación positiva entre la cantidad de directivos con educación universitaria y el tamaño de la pyme.
2. En cuanto al personal que conforma la pyme, el 16% no tiene estudios, el 60% tiene estudios básicos, Bachillerato o Formación Profesional y el resto que es el 24% posee estudios Universitarios.
3. Las pymes tienen una orientación hacia la innovación en la organización y en la independencia o autonomía de sus colaboradores, más que una orientación basada en la toma de riesgos.

Analizando cómo estos factores pueden impactar el desarrollo de este proyecto, el hecho que exista un cierto grado de conservadurismo en la toma de decisiones, siendo un poco adversos a los riesgos, podría significar un impacto negativo en el uso o aceptación de la herramienta, desde el punto de vista financiero, aunque también se indica que la innovación juega un papel importante en la estrategia de este tipo de empresas, lo que podría favorecer al uso de la misma en la búsqueda de insights que promuevan este valor específicamente.

3.5.3. Factores económicos

En toda Europa, si existe un país donde su economía depende fuertemente de las pymes, ese país es España. Éstas aportan el 63% al Valor Añadido Bruto de la nación, por ende, la forma en que se desempeñan condiciona muchísimo la salud de la economía nacional.

Es indudable que la pandemia del Covid-19 va a devastar la economía mundial y la economía española será una de las más afectadas debido al impacto severo que está teniendo actualmente esta terrible enfermedad. Acorde a indicadores aproximados proporcionados por el Banco Bilbao Vizcaya Argentaria (BBVA) el Producto Interno Bruto (PIB) caerá un 8% en este año y el Índice de Precios de Consumo (IPC) caerá un 0.6%. Sin embargo, se prevé que, en 2021, exista una mejora considerable en estos dos indicadores macroeconómicos (véase Figura 8).

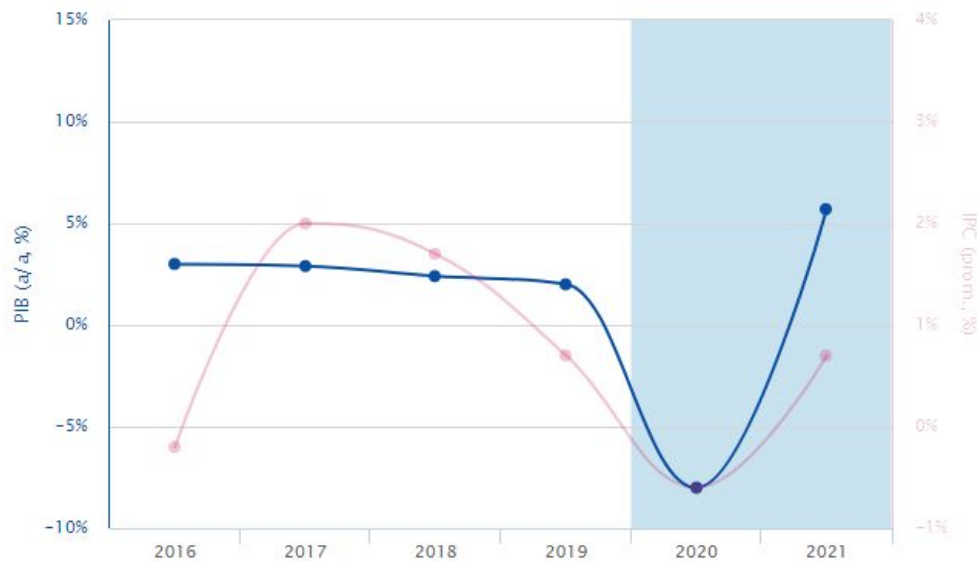


Figura 8: Crecimiento Porcentual Anual de Indicadores PIB y IPC. Fuente: (Indicadores | BBVA Research 2020).

En 2020, habrá un golpe muy fuerte al consumo privado el cual se espera que presente un decrecimiento de 8,7% vs un incremento de 2,6% del consumo público (véase Figura 9). Los clientes potenciales de la solución pueden ser tanto administraciones públicas, empresa privada o emprendedores, por lo que se estima un impacto en el cálculo de nuestros clientes potenciales inicialmente, aunque en 2021 se ve que el consumo privado experimentará un crecimiento de 3,4%, lo cual es un aliciente para captar un poco más de clientes del sector privado.

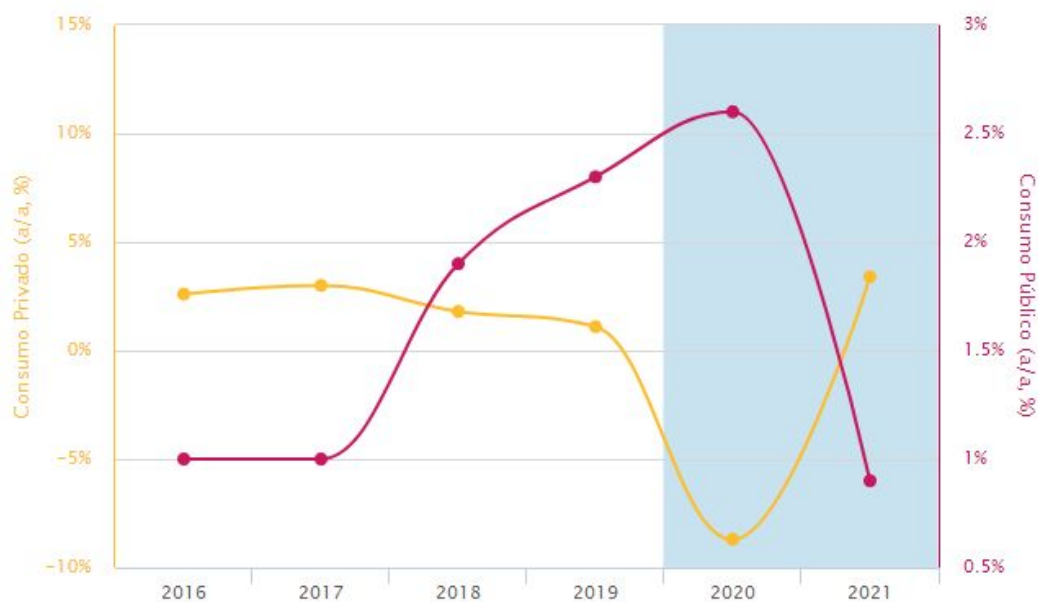


Figura 9: Evolutivo Anual de Indicadores Consumo Privado vs Consumo Público. Fuente: (Indicadores | BBVA Research 2020).

De todas las empresas españolas que exportan productos y servicios, el 72,21% son pymes, por lo que el indicador macroeconómico de las exportaciones es muy importante para asegurar la continuidad operativa de cada pyme y así mejorar el PIB de la nación. Al analizar la caída que

tendrán las exportaciones en 2020, esta será de 26,6%, ya en 2021 se espera un crecimiento de 9,3% (véase Figura 10).

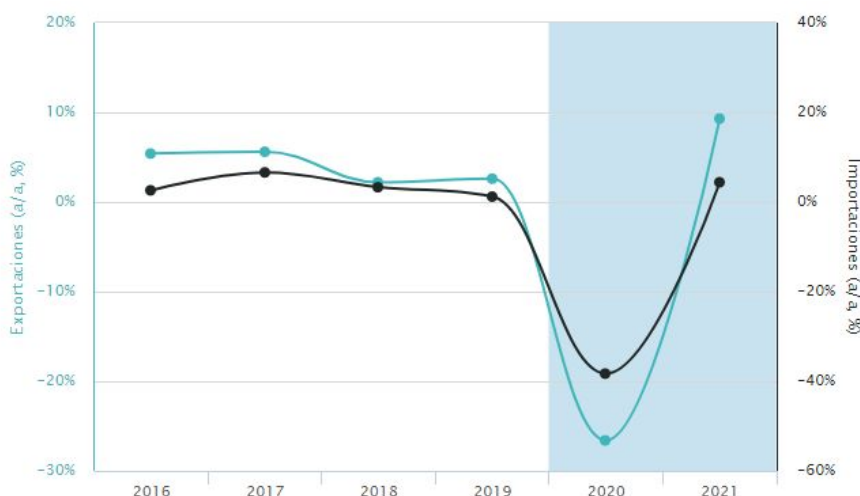


Figura 10: Evolutivo Anual de Importaciones y Exportaciones. Fuente: (Indicadores | BBVA Research 2020).

Uno de los indicadores críticos de toda economía es el empleo y este se vuelve todavía más crítico cuando las pymes generan el 65,9% de los empleos de toda España. Se prevé que el empleo caerá en 2020 alrededor 6,9 puntos porcentuales y la productividad aparente del sector trabajo un 1,2% (véase Figura 11). Por otra parte, la tasa de paro crecerá más de 20 puntos porcentuales durante este año, similar a la que se tenía hace cinco años atrás, aunque ese crecimiento no será tan alto en 2021 con un 17,3.

De aquí podemos inferir el gran impacto recibido en el sector de las pymes por la pandemia del Covid-19, lo que está provocando el cierre de muchas empresas y fuertes recortes de personal.



Figura 11: Evolutivo Anual de Empleo y Productividad aparente del factor trabajo. Fuente: (Indicadores | BBVA Research 2020).

La caída en la tasa de empleo de España en 2020, lógicamente incidirá en la renta disponible nominal de cada hogar y estimulará la tasa de ahorro debido a la incertidumbre generada por la situación actual (véase Figura 12).



Figura 12: Evolutivo Anual de Renta disponible nominal vs Tasa de ahorro. Fuente: (Indicadores | BBVA Research 2020).

En cuanto a la parte fiscal, el déficit público pasará de 2,6% a 10,8% en 2020 (véase Figura 13) con respecto al PIB, algo catastrófico para el gobierno. En 2021 se prevé una mejora, donde se estima será el 6,7% del PIB. Este indicador es bastante delicado porque reducirá seguramente el presupuesto que tiene el Gobierno para el apoyo a las pymes, ya que tendrá menos dinero disponible al estar pagando mayores intereses y capital por la contratación de mayor deuda externa para la financiación de ese déficit.



Figura 13: Evolutivo Anual del Saldo Fiscal vs Deuda, ambos como porcentaje del PIB. Fuente: (Indicadores | BBVA Research 2020).

Todos estos factores incidirán unos más que otros en el proyecto Geopymes, pero al fin y al cabo determinarán la estrategia en la selección de los clientes potenciales y en la determinación de la importancia de las variables del entorno a tomar en cuenta para optimizar las recomendaciones o valor agregado que recibirán los usuarios del aplicativo.

3.5.4. Factores socio-demográficos

Los factores socio-demográficos condicionan ya sea para bien o para mal el éxito de una pyme. No es nada nuevo para nadie que las ciudades o zonas del país donde sus habitantes cuentan con mayor ingreso per cápita potencia la creación y sostenimiento de las pymes.

No siempre las zonas o mercados condicionan por sí solos el éxito o fracaso de una pyme, sino que existe un factor determinado por las características sociodemográficas del empresario que al final de cuentas puede llevar al fracaso a su propia empresa.

Según un estudio que recibió el Premio Internacional Reina Maria Cristina, los autónomos que tienen empresas sin asalariados, los empresarios de la construcción frente a servicios, industria o agricultura, los emprendedores más jóvenes y menos experimentados, o los que carecen de formación versus los que tienen formación, tienen una mayor propensión al fracaso al frente de una pyme.

GeoPymes es un proyecto que será desarrollado para servir a las personas que desean abrir un negocio en la ciudad de Madrid cuya población es de aproximadamente 3,1 millones de habitantes y posee una superficie de 605,77 km². La ciudad de Madrid es el municipio más grande de la Comunidad Autónoma de Madrid tanto en población como en extensión territorial.

La Comunidad Autónoma de Madrid tiene aproximadamente 6,6 millones de habitantes, de los cuales la mayoría ronda entre los 40 a 45 años de edad (véase Figura 14). Se observa una tendencia creciente en el envejecimiento de la población y una tendencia decreciente en la presencia de población infantil entre 0-14 años (véase Figura 15). La población entre 15 - 64 años se mantiene relativamente estable.

La población femenina es de 52,06% y la población masculina es de 47,94%. Su tasa de natalidad es de 8,4%, la cual es alta comparada al resto del país y se encuentra entre las comunidades autónomas con menor tasa de natalidad. Su esperanza de vida es de 85 años, la cual es la más alta de toda España.

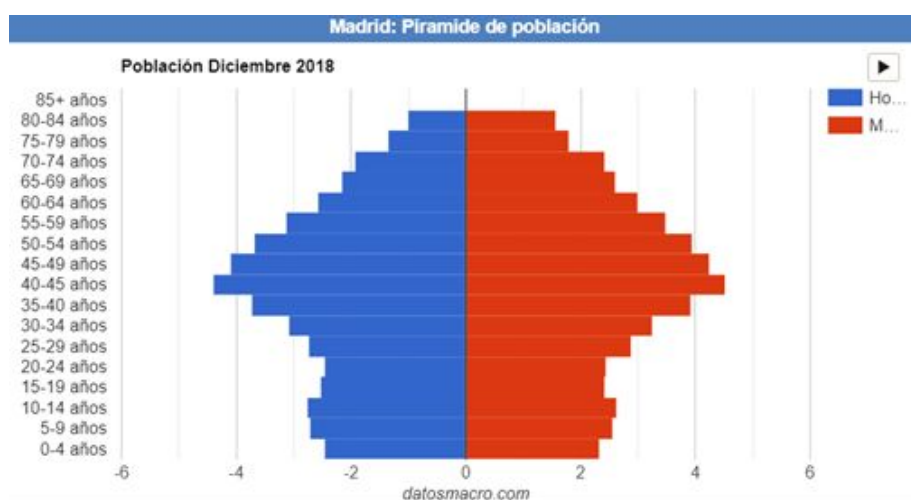


Figura 14: Distribución por rango de edades de la población de la Comunidad de Madrid. Fuente: (Pirámide de población de Madrid 2019 - datosmacro.com)

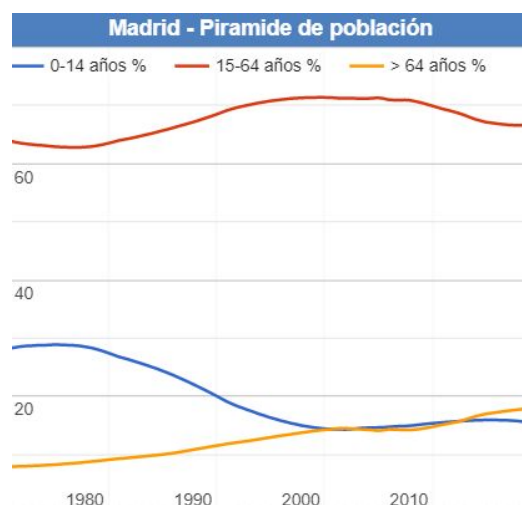


Figura 15: Evolutivo de los últimos 50 años de la población de la Comunidad de Madrid por rango de edades.

Analizando algunos indicadores importantes (véase Figura 16) relacionados a las pymes en la Comunidad de Madrid, se observa que su tasa de paro es de 14,6% menor a la media de España que es de 18,6%.

El sector productivo predominante es el sector de Servicios, donde 86,8% de las pymes pertenecen a este sector específico, el cual se encuentra por arriba de la media del país que es de 75%. Por otra parte, los sectores Agrícola e Industrial no son tan predominantes y están muy por debajo de la media española. También se observa un sector industrial que está casi al nivel de la media del país teniendo alrededor de un 5% de las pymes.

Los habitantes de la Comunidad de Madrid poseen un PIB per cápita de 136 (3 euros por encima de la media del país que es de 100 euros) y una renta disponible bruta por habitante de 125,9 euros.

| DEMOGRAFÍA | MADRID | ESPAÑA |
|--|-----------|------------|
| Superficie (Km2) | 8.028 | 505.938 |
| Población (1-1-17), Padrón municipal de habitantes | 6.507.184 | 46.572.132 |
| Densidad de población (Hab./ Km2), Padrón municipal (1-1-17) | 810,6 | 92,1 |
| MERCADO DE TRABAJO (EPA, IV trimestre 2016) | | |
| Tasa de actividad (%) | 63,2 | 59,0 |
| Tasa de paro (%) | 14,6 | 18,6 |
| Empleo en el sector agrícola (%). Año 2016 | 0,2 | 4,8 |
| Empleo en el sector industrial (%). Año 2016 | 8,0 | 14,9 |
| Empleo en el sector de la construcción (%). Año 2016 | 5,0 | 5,3 |
| Empleo en el sector de los servicios (%). Año 2016 | 86,8 | 75,0 |
| PRODUCCIÓN | | |
| PIBpm (millones de euros). Año 2016 | 211.528 | 1.118.522 |
| Variación en volumen del PIBpm 2016-2015 (%) | 3,9 | 3,8 |
| Participación del PIBpm regional en el total nacional (%). Año 2016 | 18,9 | 100 |
| PIBpm per cápita (Media de España=100), 2016 | 136,3 | 100 |
| Variación del PIBpm/Hab. 2016-2015 (%) | 3,2 | 3,5 |
| Renta disponible bruta/Hab. (Media de España=100). Año 2015 | 125,9 | 100 |
| PRODUCTIVIDAD DISTRIBUCIÓN VAB POR OCUPADO. Año 2016 | | |
| Total | 49,2 | 100 |
| Agricultura | 123,0 | 100 |
| Industria | 113,2 | 100 |
| Construcción | 123,0 | 100 |
| Servicios | 122,4 | 100 |
| TECNOLOGÍAS INFORMACIÓN Y COMUNICACIONES, EMPRESAS. Año 2016 | | |
| Empresas que utilizaron Medios Sociales para: Desarrollar la imagen de la empresa o productos de mercado (publicidad, lanzamiento de productos...) % | 53,9 | 49,6 |
| Empresas con conexión a Internet y sitio/página web (%) | 78,5 | 77,7 |
| Empresas que envían facturas electrónicas en un formato estándar adecuado para el procesamiento automático (p.e. EDI, UBL, XML, facturas...) (%) | 35,2 | 37,8 |
| Empresas que envían facturas electrónicas en un formato estándar adecuado para el procesamiento automático (p.e. EDI, UBL, XML, facturas...) (%) | 20,9 | 22,3 |
| INVESTIGACIÓN Y DESARROLLO. Año 2016 | | |
| Gastos internos en I+D (en % del PIB) | 1,7 | 1,2 |
| Personal en equivalencia a jornada completa en I+D (% por mil ocupados) | 17,0 | 11,1 |

Figura 16: Indicadores demográficos y productivos de las pymes en la Comunidad de Madrid. Fuente: (IPYME 2018).

La ciudad de Madrid es una de las zonas más apropiadas para el establecimiento de una pyme en toda España por su alta densidad poblacional y el alto poder adquisitivo de sus habitantes, sin embargo, presenta un alto desafío para estas, con relación a su éxito y sostenibilidad en el tiempo, debido a la alta cantidad de competidores que podría tener y los altos costes operativos en los que podrían incurrir, por ende, la necesidad de las personas que quieren emprender un negocio en cuanto a recibir una recomendación o asesoramiento para su inversión es bastante grande.

3.5.5. Factores tecnológicos

En esta era de la información que estamos viviendo en el siglo XXI se ha visto la aceleración exponencial en la generación de los datos y con ello la creación de una inmensa cantidad de herramientas para sacar el mejor provecho de ellos.

En el estudio **Top Technology Trends 2019** llevado a cabo por Gartner, el 69% de las pymes en España considera crucial para los negocios la tecnología de Seguridad de datos e información. Desde este punto de vista, se pueden tener algunos obstáculos en la recolección de algunas variables importantes debido a que las pymes son bastante conservadoras en la revelación de cierta información sensitiva que podrían ser de utilidad para el desarrollo, mantenimiento y escalabilidad del proyecto.

Por otra parte, más del 75% de las pymes considera a las tendencias tecnológicas como un factor importante (véase Figura 17), con esto, favoreciendo al uso de tecnologías Big Data en la toma de decisiones, las cuales se emplearán en la realización de este proyecto.

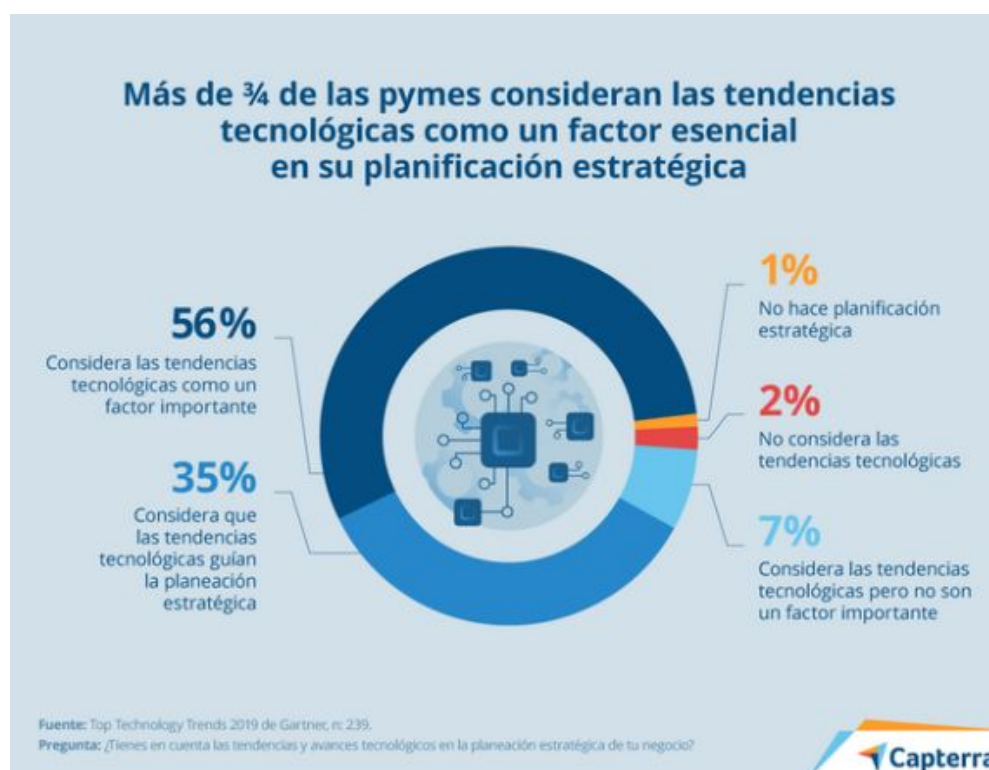


Figura 17: Tendencias tecnológicas como un factor esencial en la planificación estratégica de las pymes. Fuente de datos: Gartner.

Se conoce que existe mucha tecnología disponible, sin embargo, por la composición del nivel académico de los empleados de las pymes, supone un reto, la implantación y utilización de esta tecnología. Existe mucho apego a lo tradicional, resistencia al cambio, que puede reducir el uso

de la solución planteada en este documento, sin embargo, la misma se tratara de diseñar lo más amigable posible para el usuario final.

Otro reto importante es el bajo grado de digitalización promedio en las pymes (véase Figura 18). Este bajo grado de digitalización promedio viene determinado en gran parte por la baja formación en Tecnologías de Información y Comunicación (TIC) en el personal de las empresas, muy pocas poseen especialistas en TIC como parte de su grupo de colaboradores y el uso de tecnologías Big Data sólo está presente en el 10% de este tipo de empresas. Asimismo, se puede observar que menos de un 25% de ellas tienen servicios de Cloud Computing. Sin embargo, es importante que cerca del 90% de ellas posean un ordenador con acceso a internet de banda ancha fija o móvil, lo que posibilitará el acceso a la herramienta online propuesta.



Figura 18: Digitalización de la empresa española. (Smartpoint 2019).

3.6. Análisis de los competidores

Al estudiar el ámbito de actividad de Geopymes, hay que definir primero los conceptos de **Geomarketing** y de **Location Analytics**, para que así sea más fácil identificar a los competidores de la compañía.

El **geomarketing** se puede definir como el conjunto de técnicas que permiten analizar la realidad económico-social desde un punto de vista geográfico, a través de instrumentos cartográficos y herramientas de estadística espacial. Dichas técnicas permiten analizar las características de la oferta y la demanda de una actividad económica mediante la localización de clientes potenciales, competencia, puntos de venta y descubrir localizaciones de mayor potencialidad. Para ello, se toma como referencia los comportamientos, consumos y actitudes de personas que comparten un mismo entorno o entornos geográficos cercanos, aportando datos sobre el consumidor y mercado.

Por lo tanto, el **geomarketing** se puede entender desde dos puntos de vista:

1. **Utilizar elementos geográficos como variables de análisis** del mercado para ayudar en la toma de la decisión de localización de un negocio en una zona geográfica concreta. Está muy relacionado con los Sistemas de Información Geográfica (SIG), aplicados a una de las variables básicas del marketing: la distribución. En definitiva, da respuesta a la siguiente pregunta: ¿existe demanda en esta zona geográfica?
2. **Aplicar la geolocalización a las estrategias comerciales.** Se trata de conocer la ubicación geográfica exacta de un objeto o una persona para, en función de su situación, determinar la oferta comercial que sea más adecuada. Con esto se logra que un producto o servicio llegue al consumidor adecuado, en el momento idóneo.

Actualmente, el geomarketing se puede dividir en varios niveles: regional, nacional, local y micro local. A pesar de ello, cuando se alude a geomarketing, tradicionalmente se habla a nivel local.

En definitiva, la idea es cubrir una necesidad inmediata, basada en la proximidad geográfica, desarrollada gracias a la evolución tecnológica, a los sistemas de localización vía satélite y a la gestión de grandes bases de datos. Algunos autores lo resumen en las 4 C's del Geomarketing:

1. **Cientes.** Zona geográfica en la que se encuentra el público objetivo y los clientes potenciales.
2. **Competencia.** Conocer dónde se sitúa la competencia y su penetración en el mercado local.
3. **Concentración.** Información valiosa sobre las zonas donde exista mayor densidad de clientes.
4. **Cartografía.** El hecho de representar a los clientes y a la competencia en un mapa.

Tal es su evolución, que el geomarketing se está convirtiendo en una estrategia principal para la supervivencia de empresas, tanto pymes como multinacionales. El análisis geográfico va a evitar que se sitúen en mercados muy saturados o donde no exista demanda. El objetivo será cubrir una necesidad de una zona geográfica, de una forma óptima.

Las **principales ventajas** del geomarketing en las empresas son:

1. Optimizar la localización de un negocio. Situar el negocio donde exista mayor concentración de demanda y accesibilidad. Desde el punto de vista del aprovisionamiento también debe estudiarse el emplazamiento, para optimizar y facilitar el intercambio.
2. Aumentar la visibilidad del negocio. La segmentación de las campañas de marketing son más precisas, por lo se impacta sobre un público que presenta mayor interés. Además, el geomarketing influye en el posicionamiento orgánico de una web, llamado SEO Local.
3. Mejorar las estrategias de marketing y publicitarias en la web y redes sociales. Muy relacionado con el punto anterior, definir el público objetivo de una campaña web o en RRSS conlleva una mayor tasa de conversión y ventas. En el caso de las redes sociales, las estrategias de geomarketing repercuten positivamente en la fidelización de los clientes locales.
4. Análisis más profundo de la competencia directa. La mayoría de las herramientas de geomarketing ofrece en sus bases de datos información sobre la competencia directa.

Dicha información es muy valiosa para decidir el público objetivo y la zona geográfica de actuación.

5. Conocer las zonas geográficas a las que expandirse. El geomarketing aporta datos sobre aquellos nichos de mercado o región con potencial para expandir un negocio.

Entre los principales **beneficios del Geomarketing** se pueden destacar los siguientes:

1. Optimización de la inversión en acciones de marketing.
2. Un mayor conocimiento de mercados y la habilidad de focalizar esfuerzos en determinados segmentos del mercado.
3. Diseñar zonas de ventas, rutas de despacho, rutas de merchandising y rutas de cobranza.
4. Visualizar las ventas por cliente, zona de ventas, sectores, parroquias, municipios, poblaciones, estados, y nación, que sirve como insumo para los análisis de las ventas.
5. Identificar puntos de ventas, oficinas, sucursales, distribuidores, competencia, etc.
6. Localizar oficinas más cercanas, análisis de rutas óptimas y alternativas.
7. Determinar el área de influencia para precisar la población a la cual se está cubriendo.
8. Responder a preguntas como: ¿Es óptima la localización actual de mi negocio? ¿Dónde se podría ubicar una nueva sucursal? ¿Dónde dirigir una campaña publicitaria?
9. Análisis del potencial del mercado - domicilios por rango de ingresos.
10. Añadir valor en procesos de marketing directo o de atención al cliente.

Dentro de las múltiples ramas o áreas que conforman el Geomarketing, se puede definir **Location Analytics** como toda aquella tecnología que esté enfocada a aportar información de valor procedente de datos geolocalizados, añadiendo datos en un mapa y descubriendo tendencias y patrones que son invisibles con tablas, diagramas y gráficos. De esta forma se facilita la toma de decisiones, haciendo que estas sean más efectivas.

Una vez definidos estos dos conceptos, se pueden dividir los competidores de Geopymes en dos categorías: directos e indirectos.

A. Competidores directos: estarían integrados por todas aquellas empresas cuya actividad principal está relacionada o englobada dentro de la definición de Geomarketing y, en concreto, con la Location Analytics. Estas empresas utilizan tecnologías de Sistemas de Información Geográfica (GIS), y herramientas de Big Data para obtener y analizar información geográfica dentro de los negocios, con el objeto de extraer conclusiones y predicciones que sean útiles a la hora de tomar decisiones, definir estrategias, planificarlas, desarrollarlas y ejecutarlas. Para ello, desarrollan cuadros de mando, aplicaciones web o móviles a partir de softwares basados en GIS, como CARTO (startup española), ESRI, Google Maps, Mapbox, etc. De todas las empresas que se pueden incluir aquí, señalamos las más destacadas:

- **inAtlas:** [inAtlas | BIG DATA and Location Analytics](#)

Empresa española especializada en Data Analytics y Geomarketing, está asociada con la base de datos de Informa, empresa líder en el suministro de Información Comercial, Financiera, Sectorial y de Marketing de empresas y empresarios. Ofrece una solución web que ofrece información sobre reportes económico-financieros, comparativa de competidores, análisis sobre mejor localización, predicción de patrones, etc. Entre sus clientes hay Bancos y Sociedades de Seguros, Retail y Horeca, Energía y Telecomunicaciones, Bienes raíces, empresas del Sector público, ...

- **Geoblink:** [Geoblink: The Location Management Platform](#)

Startup española nacida en 2015 especializada en Location Intelligence. Ofrece sus servicios a empresas del sector Retail, Gran Consumo y Sector inmobiliario. Entre sus clientes hay empresas de Retail, Bancos, Supermercados y Franquicias, principalmente.

- **Geographica:** <https://geographica.com/>

Se definen como “entusiastas de los datos espaciales, el diseño y el software libre. Hacemos que los datos cuenten con soluciones intuitivas basadas en GIS, BI y Location Intelligence. Mostramos, mejoramos y analizamos tus datos”. Basan su aplicación en Carto. De entre sus clientes hay AAPP, empresas de Retail, Energéticas, Telecom, Seguros, ...

- **Geograma:** [Servicios de captura, tratamiento de geoinformación y desarrollo GIS](#)

Ofrecen servicios de captura, tratamiento y gestión de GeoInformación, para su desarrollo e integración con GIS: servicios de topografía e inventario, consultoría y asistencia técnica, geomarketing, ... Entre sus clientes destacan la DGT, Ferrovial, AAPP, consultoras de ingeniería, etc.

- **Moca:** [MOCA Platform: AI-powered marketing for deep personalization](#)

Empresa española especializada en marketing y en big data, ofrece sus servicios a empresas de retail, eventos y de turismo para favorecer la captación y fidelización basándose en el estudio georeferenciado tanto de los clientes como de los establecimientos que visitan.

- **Quantic Solutions:** [Quantic: HOME](#)

Empresa especializada en Big Data, Business Intelligence, AI y Geomarketing, ofrecen servicios de Análisis espacial, GIS aplicado a marketing y estudios de zonas óptimas. Las soluciones de geomarketing de Quantic integran el geoposicionamiento con el marketing y la inteligencia de datos para analizar todas las variables espaciales que tengan relación con la toma de decisiones, tanto en términos de marketing como de negocio.

- **Geomarketing Xpert:** [GeoMarketing Xpert, la mejor plataforma de Location Intelligence para un Análisis Experto](#)

Es la aplicación desarrollada por la agencia de Rating Axesor, consistente en una plataforma web de análisis de información, en la que se combina el acceso al repositorio de datos de mercado de Axesor, con la tecnología de GIS y el Business Intelligence. Con esta aplicación se pueden analizar puntos de venta actuales y nuevas ubicaciones, áreas comerciales, clientes y potenciales, y competencia.

- **Flame Analytics:** [Flame Analytics | Instore Analytics | Guest Wifi | Location Based Marketing](#)

Empresa española especializada en Geomarketing, que ofrece a sus clientes una solución de analítica para espacios físicos en base a los datos que estos generan. Es una solución modular, es decir, cada cliente se suscribe solo a aquello que le interesa, integrándose con la tecnología más apropiada para cada propósito. Ofrecen sus servicios a Centros Comerciales, cadenas de retail, restaurantes, hoteles y museos.

B. Competidores indirectos: su actividad principal no está relacionada con el análisis de datos georreferenciados ni con el Big Data, pero al conocer el mundo empresarial, o al tratar temas relacionados con el marketing, la arquitectura, o la ingeniería ambiental, expanden sus ámbito de actuación a la elección de mejor ubicación de un negocio. Entre estas empresas podemos citar:

1. **Asesorías de empresas:** muchos emprendedores y pequeñas empresas recurren a asesorías y gestorías para que administren su contabilidad, facturación, sus relaciones con las AAPP y con terceros, por lo que estas empresas tienen un conocimiento bastante profundo de la actividad de sus clientes y pueden orientarlos en la elección de la ubicación de sus negocios a fin de que esta sea lo más rentable posible.
2. **Consultoras y agencias de marketing:** aquellas empresas que estén especializadas en investigación de mercados y análisis de distribución comercial (Retail), pueden aprovechar sus conocimientos para realizar estudios de localización dentro de las áreas de influencia y evaluar cual es la mejor posición respecto a los posibles competidores.
3. **Estudios de arquitectura:** con la crisis financiera del año 2008 y el estallido de la burbuja inmobiliaria, muchos estudios se han reconvertido y, aparte de ofrecer servicios como el diseño y la construcción de edificios, también se dedican a actividades relacionadas con la apertura de nuevos negocios, como por ejemplo la tramitación de las licencias de obra y los permisos de reformas, estudios de mercado y de posicionamiento, así como elegir la ubicación más propicia para la empresa.

Si bien parece que al hacer este análisis la sensación que da es que el sector puede estar saturado y que no hay opción de introducirse, nos hemos dado cuenta de que muchas de estas empresas se centran en captar solamente a clientes grandes, obviando el potencial de los autónomos y de las pymes, lo cual representa un nicho de mercado el cual puede ser cubierto por Geopymes. Por otro lado, hemos visto que la colaboración con las AAPP es bastante anecdótica, por lo que también es una opción interesante a la hora de ofrecer nuestros servicios.

4. Analisis y diagnostico / Plan estrategico-acción

4.1. Analisis y diagnostico (DAFO)

En esta sección se realizará un análisis DAFO (Debilidades, Amenazas, Fortalezas y Oportunidades) (véase *Figura 19*) de los principales factores internos y externos que pueden afectar o beneficiar el desarrollo del proyecto Geopymes.

Respecto a los factores internos se analizarán las fortalezas y debilidades con las que cuenta el proyecto y en la parte externa se hará un análisis de las oportunidades y amenazas a las cuales se enfrentará para el desarrollo del mismo.

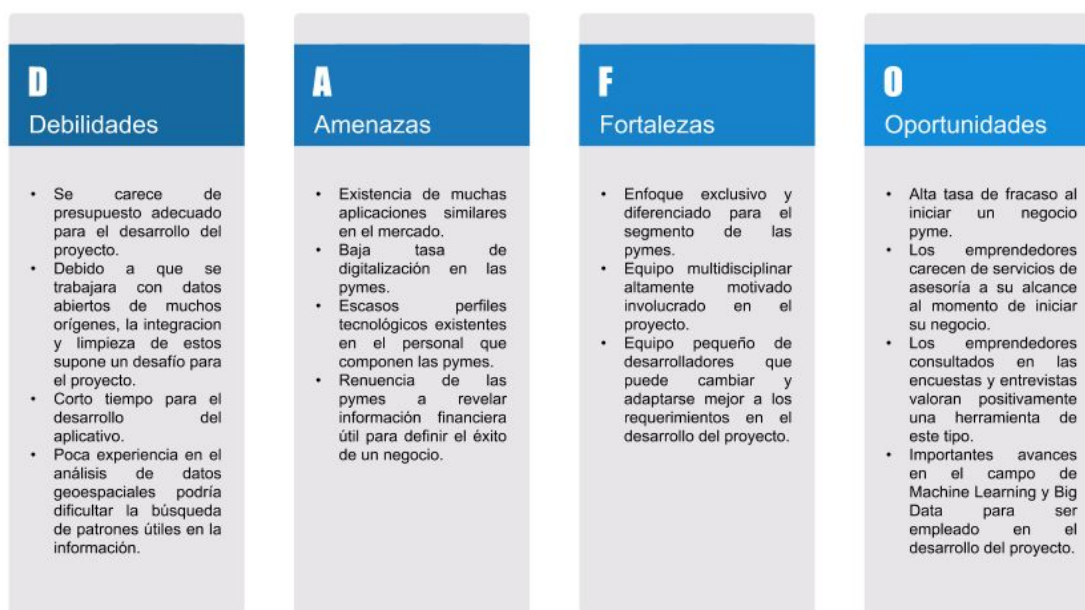


Figura 19: Análisis DAFO. Fuente: Elaboración propia.

Al analizar los factores internos se verán los recursos y habilidades con las cuales se cuenta para sacar adelante la idea y plasmarla en un proyecto empresarial, y en los factores externos se verán todos aquellos factores de los cuales no se tiene control y que pueden ayudar o perjudicar la puesta en marcha de la propuesta a desarrollar.

A partir de los resultados obtenidos en esta parte, se definirá la estrategia a seguir para el desarrollo del proyecto.

A continuación se hará una descripción breve de cada una de las partes que componen el análisis DAFO de Geopymes:

Fortalezas

1. **Enfoque exclusivo y diferenciado para el segmento de las pymes:** la solución tendrá un enfoque exclusivo para el segmento de las pymes, lo que permitirá la canalización de todos los esfuerzos en la búsqueda de esos componentes vitales en la recomendación de una ubicación o tipo de industria en la cual una pyme pueda tener mayor probabilidad de éxito.
2. **Equipo multidisciplinar altamente motivado involucrado en el proyecto:** el equipo está compuesto por personas cuyos perfiles comerciales y técnicos pueden facilitar la materialización de la idea empresarial.
3. **Equipo pequeño de desarrolladores que puede cambiar y adaptarse mejor a los requerimientos en el desarrollo del proyecto:** el hecho de que el grupo de desarrolladores está compuesto por solo cuatro personas, facilita la comunicación y la distribución de las tareas que componen el proyecto, así como la adaptación al cambio en caso que este sea requerido a medida que avanzan los tiempos en la construcción de la solución.

Oportunidades

1. **Alta tasa de fracaso al iniciar un negocio pyme:** la mayoría de las pymes fracasan mucho antes de lo esperado. Por ejemplo, más del 62% de los autónomos fracasan antes de que su negocio cumpla los cinco años de vida. Este hecho es indudablemente una oportunidad para Geopymes, debido a que este segmento de empresarios requieren de algo más para asegurar la continuidad de su negocio y la clave podría ser una recomendación de una buena ubicación para el funcionamiento de la pyme, dado que la ubicación es uno de los principales factores por los que una pyme fracasa.
2. **Los emprendedores carecen de servicios de asesoría a su alcance al momento de iniciar su negocio:** de acuerdo a lo que se observó en las encuestas realizadas para la validación de las hipótesis del proyecto, los emprendedores de la ciudad de Madrid adolecen del servicio de asesoría al momento de abrir su negocio, con lo que Geopymes podría convertirse en la asesoría idónea para muchos de ellos en cuanto a la ubicación y tipo de industria apropiada.
3. **Los emprendedores consultados en las encuestas y entrevistas valoran positivamente una herramienta de este tipo:** la buena aceptación de la idea de Geopymes hace pensar que es una oportunidad valiosa para que el proyecto tenga éxito una vez que este sea comercializado.
4. **Importantes avances en el campo de Machine Learning y Big Data para ser empleado en el desarrollo del proyecto:** el auge que tienen en la actualidad la aplicación de modelos de Machine Learning en la resolución de problemas complejos, así como el uso de herramientas de Big Data en el análisis de datos tipo geoespaciales, hace posible un proyecto como este, donde a partir de información geolocalizada se buscarán insights para dar una recomendación de valor a un emprendedor antes de iniciar su empresa.

Debilidades

1. **Se carece de presupuesto adecuado para el desarrollo del proyecto:** por ser una Start Up donde los fondos provienen de los fundadores, familiares o amigos, se maneja un presupuesto muy reducido que limita muchos aspectos básicos para el desarrollo del proyecto como la adquisición de software para desarrollo, logística para estudio del mercado, contratación de recursos humanos, entre otros, que coloca en desventaja a Geopymes frente a competidores con más recursos.
2. **Debido a que se trabajará con datos abiertos de muchos orígenes, la integración y limpieza de estos supone un desafío para el proyecto:** la totalidad de los datos que se utilizarán para el entrenamiento de los modelos computacionales que darán la recomendación a los clientes serán Datos Abiertos. Este tipo de datos por naturaleza están dispersos en muchos sitios web, careciendo de una adecuada documentación y por tanto haciendo difícil su consolidación en un solo dataset que sea funcional para alimentar un modelo de Machine Learning.
3. **Corto tiempo para el desarrollo del aplicativo:** muchos proyectos de Big Data requieren de muchos meses de desarrollo para encontrar los patrones idóneos en la data y poder resolver de una forma óptima un problema determinado, por lo tanto, este es sin duda un reto que podría convertirse en una debilidad para la elaboración de un producto mínimo viable.

4. **Poca experiencia en el análisis de datos geoespaciales podría dificultar la búsqueda de patrones útiles en la información:** el equipo de desarrolladores no posee mucha experiencia con el tratamiento y análisis de datos geoespaciales, significando un reto muy interesante de cara al proyecto.

Amenazas

1. **Existencia de muchas aplicaciones similares en el mercado:** en el mercado se pueden encontrar muchas aplicaciones similares que utilizan las ventajas ofrecidas por el location analytics y el Big Data para la realización de predicciones de valor a sus clientes, significando competidores muy serios para Geopymes, que le podrían impedir posicionarse adecuadamente en un mercado sumamente competitivo.
2. **Baja tasa de digitalización en las pymes:** la baja tasa de digitalización promedio en la pyme española (véase Figura 18) supone un obstáculo importante al tratar de posicionar una solución como Geopyme, ya que el uso de Big Data es sumamente bajo. En este mismo sentido, un indicador que no es muy halagador, es que estas empresas poseen un presupuesto bastante reducido en la parte tecnológica.
3. **Escasos perfiles tecnológicos existentes en el personal que componen las pymes:** dado que las personas que componen las pymes cuentan con pocos especialistas en TIC (véase Figura 18) y la mayoría de sus empleados tienen poca formación en las mismas, en caso de no cambiar ese comportamiento en los próximos años, se podrían tener inconvenientes en la aceptación de la tecnología empleada en Geopymes, debido al desconocimiento de la misma.
4. **Resistencia de las pymes a revelar información financiera útil para definir el éxito de un negocio:** de acuerdo al estudio Top Technology Trends 2019 llevado a cabo por Gartner, el 69% de las pymes considera importante la protección de sus datos, con lo que se tendría un obstáculo importante en la optimización y mejora de los resultados que brinda el aplicativo.

4.2. Definición del modelo de negocio

A continuación se detallan los elementos que componen nuestro modelo de negocio. Para ello, vamos a utilizar la metodología del Business Model Canvas para ir desarrollando los distintos componentes e integrantes de nuestra propuesta:

| SOCIOS CLAVE | RECURSOS CLAVE | PROPUESTA DE VALOR | RELACIÓN CON CLIENTES | SEGMENTOS DE MERCADO |
|---|--|---|---|---|
| <ul style="list-style-type: none">● Repositorio de fuentes de datos abiertos● Proveedores de software y hardware● Profesionales expertos en BI y Big Data | <ul style="list-style-type: none">● Fuentes de datos● Hardware y software● Recursos humanos | Desarrollo de una herramienta online de recomendación de negocios y ubicación para pymes. | <ul style="list-style-type: none">● Reuniones periódicas con AAPP y otras entidades● Asesoría y soporte | <ul style="list-style-type: none">● Clientes particulares● AAPP● Otras entidades: bancos, inmobiliarias, aseguradoras |
| | ACTIVIDADES CLAVE | | CANALES DE DISTRIBUCIÓN | |
| | <ul style="list-style-type: none">● Identificar los datos a utilizar● Recogida e integración de los datos● Procesamiento de los datos● Construcción del dataset● Modelado y clusters de recomendación● Dashboard o cuadro de mando● Otras: Captación de grandes clientes, soporte y atención al cliente, I+D | | <ul style="list-style-type: none">● Visitas con los clientes potenciales● Redes sociales● Contacto via email● Eventos especializados, ferias profesionales, ... | |
| ESTRUCTURA DE COSTES | | | FUENTES DE INGRESOS / BENEFICIOS | |
| <ul style="list-style-type: none">● Horas de trabajo● Infraestructura IT y coste licencias● Administrativos y Marketing / Comercial● Coste periodico plataforma Cloud / DWH● Soporte áreas implicadas | | | <ul style="list-style-type: none">● Contrato de servicio por consulta (clientes particulares) o por suscripción mensual o anual (Grandes cuentas)● Opción Premium de informes con más información● Ampliación de las variables analizadas, servicios adicionales. | |

Figura 20: Model Canvas. Fuente: Elaboración propia.

4.2.1. Propuesta de valor

La propuesta de valor de Geopymes es la de desarrollar una herramienta online para las personas que quieran emprender un negocio que les recomiende qué tipo de comercio puede ser más apropiado y en qué ubicación puede tener una mayor probabilidad de éxito, a fin de optimizar el proceso de decisión de la localización de su empresa.

Dada la complejidad del proyecto, se hace necesario dividir el desarrollo de estas dos líneas de recomendación: mejor ubicación y tipo de negocio. Atendiendo al resultado de las encuestas y entrevistas realizadas en la fase de validación de hipótesis, se decide comenzar por la mejor ubicación, puesto que fue la que mostró más interés por parte de los encuestados, dejando así para una etapa posterior, la recomendación del mejor tipo de negocio en una zona geográfica elegida.

Para hacer dicha recomendación, la herramienta tomará en consideración las siguientes variables:

- Cuáles son las características sociodemográficas, económicas y urbanísticas de la zona.
- Quienes serán sus potenciales competidores directos o indirectos.
- Qué clientes formarán parte de su mercado objetivo.

Basándose en las tecnologías de geolocalización, a través del análisis geográfico y de la representación de datos sobre mapas, y de la utilización de modelos descriptivos y predictivos que estudian e integran las diferentes fuentes de datos con las que se realizan las comparaciones, se desarrolla una aplicación cuyo principal objetivo es el de la optimización del lugar idóneo para iniciar un negocio.

La herramienta se presentará como una plataforma de reporting, con mapas y cuadros de mando, adaptada a las necesidades y requerimientos de cada cliente, que mostrará más o menos información en función del servicio al que se haya suscrito.

4.2.2. Socios clave

Los principales socios clave de la propuesta de Geopymes son aquellos organismos e instituciones de los que a través de sus páginas web se extrae la información con las que se configura nuestro repositorio de datos:

- Ministerio de Asuntos Económicos y Transformación Digital y la Entidad Pública Empresarial Red.es, impulsando la web datos.gob.es
- Ayuntamiento de Madrid, a través de su portal de datos abiertos
- El Instituto Nacional de Estadística (INE)

Otros proveedores clave serían aquellas empresas que proveen las licencias del software de BI, el almacenamiento en la nube, servidores en línea, etc con los que se desarrollará la herramienta.

Por otra parte, Geopymes cuenta con una red de contactos profesionales con amplios conocimientos en Business Intelligence y Big Data, los cuales forman parte de una red de apoyo y asesoramiento a la que acudir a la hora de resolver dudas y obtener información para avanzar en el progreso del proyecto. De entre ellos, cabe destacar la colaboración de la Escuela de Organización Industrial, y todo su personal docente.

4.2.3. Recursos clave

El principal recurso clave para la propuesta de valor de Geopymes son los **datos**, ya que gracias al tratamiento, interpretación y análisis de los mismos, nos permitirá establecer patrones y pronósticos que ayuden a los clientes en su proceso de decisión sobre la mejor ubicación de su negocio.

Otro recurso clave para el desarrollo del proyecto es el **equipo humano**, el cual está integrado por los miembros del grupo. Entre los integrantes, contamos con analistas y programadores con amplios conocimientos en BI y Big Data, que además de conocer cómo interpretar los datos y desarrollar el modelo (know how), elaboran cuadros de mando y visualizaciones fáciles de usar e interpretar que permiten al cliente obtener la información que necesita de forma clara y concisa.

Además, para desarrollar el proyecto, es preciso contar con **equipos e infraestructura tecnológica**, que permita a los integrantes trabajar de forma remota y colaborativa. Entre el equipamiento necesario, podemos citar los siguientes recursos:

- Hardware: ordenadores y servidor con conexión a Internet
- Servicio de almacenamiento de los datos en la nube: Google Cloud Platform
- Software para la gestión y tratamiento de bases de datos: MySQL, HeidiSQL
- Lenguaje de programación para el análisis del dataset: R
- Software de geolocalización y análisis de datos geográficos: QGIS

- Software de BI, para la visualización de los datos y su presentación en un cuadro de mando: Tableau

| FUENTES DE DATOS | BBDD: INTEGRACIÓN Y TRANSFORMACIÓN | ANÁLISIS DEL DATASET: CLUSTERIZACIÓN | GEOLOCALIZACIÓN Y VISUALIZACIÓN |
|---|--|--|--|
|    |  Google Cloud Platform   |  |   |

Figura 21: Recursos utilizados en el proyecto. Fuente: Elaboración propia.

4.2.4. Segmentos de mercado

El segmento de mercado al que se dirige Geopymes está integrado por dos grandes grupos:

1. **Cientes particulares** que desean iniciar un negocio y necesitan ayuda para elegir la ubicación del mismo
2. **Grandes cuentas** (como Administraciones Públicas, bancos, inmobiliarias, aseguradoras, ...), que pueden integrar la herramienta desarrollada dentro de su catálogo de servicios con el objetivo de ofrecer apoyo y asesoramiento a los futuros emprendedores o mejorar la toma de decisiones de sus propias actividades.

4.2.5. Relaciones con clientes

Dependiendo del tipo de cliente al que nos dirijamos, la comunicación con ellos podrá ser más o menos directa, en función del objetivo que se quiera conseguir.

Por ejemplo, con Grandes Cuentas es prioritario crear con ellos una relación de colaboración que se prolongue en el largo plazo. Para conseguirlo, será necesario crear una primera toma de contacto en una reunión en la que se les presente nuestra herramienta y se les explique los fundamentos de la propuesta de valor, resaltando cómo ésta puede adaptarse a sus necesidades específicas de información y cómo puede integrarse dentro de su catálogo de servicios como un producto que aporte valor añadido a sus clientes.

Una vez establecida la relación, será necesario mantener un contacto regular y cercano con los clientes, ofreciendo mantenimiento y soporte técnico (para resolver las posibles incidencias que puedan surgir con el manejo de la herramienta) así como para asesorar e informar de las actualizaciones y novedades que potencien el desarrollo de la herramienta.

El servicio de soporte podrá realizarse vía email, RRSS o por teléfono, mientras que para informar de las nuevas funcionalidades y de las mejoras de la tecnología sería recomendable organizar reuniones periódicas, que además nos sirvan para obtener opiniones, reacciones, sugerencias, ... de cara a adaptar nuestro servicio a las exigencias de cada cliente.

Con los clientes particulares, al ser una relación más limitada tanto en contenido como en el tiempo, ya que muchos solo realizarán una o dos consultas relativas a su actividad, el contacto se realizará sobre todo de forma indirecta. Los clientes potenciales podrán obtener información a través de la página web corporativa o de las cuentas en RRSS, sobre las funcionalidades de nuestra herramienta. Asimismo, pueden suscribirse a la newsletter periódica a las actualizaciones del blog de la compañía para mantenerse informados sobre las novedades en el sector de las nuevas tecnologías y a su aplicación en el mundo del emprendimiento. Y, en el caso de que necesiten ayuda sobre cuestiones técnicas o por incidencias por el uso, podrán contactar y recibir soporte utilizando la dirección de correo electrónico o el número de teléfono de atención al cliente.

4.2.6 Canales de distribución

Al tener dos segmentos de mercado diferenciados, hay que diversificar los canales a través de los cuales haremos llegar nuestra propuesta a los potenciales clientes:

- **Grandes Cuentas:** la relación consultiva a través de visitas concertadas será la más apropiada para este tipo de clientes. Para ello, se puede contactar vía email a los responsables de las áreas de interés identificadas para convenir una cita presencial o telemática.
- **Emprendedores particulares:** para establecer contacto con este grupo, se puede promocionar nuestra tecnología a través de acciones comerciales dirigidas (como eventos profesionales, ferias del sector, workshops,...) así como en RRSS profesionales (Linkedin, Viadeo, Xing, ...) y en blogs o páginas webs especializadas.

4.2.7. Actividades Clave

El proyecto de Geopymes está intrínsecamente relacionado con la explotación, interpretación y visualización de datos. De todo este proceso, se pueden distinguir las siguientes actividades clave:

1. **Identificación de las fuentes de datos a utilizar:** se trata de determinar los diversos tipos de datos con los que vamos a trabajar. La información relativa a las variables sociodemográficas, económicas, actividades empresariales y demás se guardan en formato .csv o .xlsx, mientras que todos aquellos datos relacionados con coordenadas geográficas (distritos, barrios, códigos postales, secciones censales, ...) se hallan en formato shp.
2. **Recogida e integración de los datos:** una vez identificadas y descargadas las fuentes de datos necesarias, se cargan en la plataforma de Google Cloud, dónde se almacenará la información que vayamos a usar. Para configurar, mantener y administrar las bases de datos relacionales en la plataforma, se usará Cloud SQL (a través de MySQL).
3. **Procesamiento de los datos:** con todos los datos subidos en la nube, se procederá a analizarlos para transformarlos en datos óptimos con los que construyamos un dataset final con el que se realizará el análisis de clusters y se elaborarán los gráficos que integrarán el cuadro de mando de la herramienta.

4. **Clusterización de las actividades:** basándonos en la utilización de varios paquetes y algoritmos de R como el análisis de la matriz de correlaciones de las variables del modelo, “Random Forest”, “H2O”, “Boruta”, ... así como en la creación de fórmulas propias, se ha realizado el estudio y análisis de aquellas zonas geográficas del municipio de Madrid en las que, tras elegir una actividad económica, se muestran varios clusters o segmentos que agrupan las empresas y negocios en función de que obtengan valores similares de las variables que se han incluido para entrenar el modelo de recomendación. Una vez obtenidos los resultados de la clusterización, se perfilan los clusters con el objetivo de buscar diferencias entre las variables y poder clasificarlos en una escala de recomendación. Se integran de nuevo en el dataset final para su posterior visualización a través de mapas y puntos geográficos.
5. **Construcción del dashboard o cuadro de mando:** con los datos resultantes tras el proceso de ETL y del análisis con R, así como los análisis geoespaciales en QGIS sobre toda la información geoespacial proveniente de archivos shp, se procede a elaborar un cuadro de mando con Tableau, en el que se visualizarán en un mapa del área urbana de Madrid los clusters agrupados por secciones censales, así como diversos gráficos con las medidas e índices de las variables utilizadas en el proceso de segmentación.

Otras actividades que serán necesarias para la puesta a punto del proyecto, tienen que ver con la captación de clientes:

- **Captación de grandes clientes.** Atraer a organismos públicos o grandes entidades para que integren nuestra herramienta de visualización dentro de sus sistemas o de su catálogo de servicios proporcionará a nuestro proyecto una base sólida que garantice un aval de cara a ampliar nuestra cartera de clientes, obtener financiación y garantizar la rentabilidad de la empresa en el medio y largo plazo.
- **Servicios postventa de soporte, mantenimiento y atención al cliente.** El contacto continuo con los clientes nos permitirá recabar información relevante sobre su opinión para adaptar nuestra solución.
- **Investigación y Desarrollo.** Para poder ofrecer un servicio competitivo que retenga a nuestros clientes será imprescindible estar al tanto de las novedades en el campo del Big Data y BI, ya sea por medio de cursos de formación, suscripción a blogs y webs especializadas en la materia, como por asistencia a eventos profesionales del sector de las nuevas tecnologías.

4.2.8. Estructura de costes

La puesta en marcha de la propuesta de Geopymes requiere de una serie de gastos iniciales, cuyos conceptos principales se detallan a continuación:

| Descripción | Cantidad | Comentar |
|--|-----------------------------------|--|
| Gastos de establecimiento | 3.500 € | Gastos de gestoría y capital social |
| Gastos de personal | 80.000€/año | Sueldos y salarios de los miembros del equipo |
| Gestoría (contable, fiscal, laboral) | 1.800 € / anuales | Cuota de 150€/mes |
| Costes accs. comerciales / marketing / web y RRSS | 4.500 € / anuales | 350 € mensuales + fee inicial |
| Coste almacenamiento Google Cloud Platform - MySQL v.5.7 | 887,64 \$ / anuales ≈ 750 € / año | 73,97 \$ al mes (2vCPUs, Memoria 7.5 GB, Almacenamiento 31GB) |
| Licencias (Tableau) | 15.240 €/anuales | <ul style="list-style-type: none"> ❑ 1 licencia Tableau Creator, que incluye Tableau Desktop, Tableau Prep Builder y Tableau Server ❑ 100 Licencias Viewer |
| Hardware | 4.515 € | 4 portátiles de trabajo + servidor |
| Total costes - Inversión inicial | 110.305 € | |

Figura 22: Costes del proyecto. Fuente: Elaboración propia.

El principal coste sería el tiempo medible en horas que cada implicado en el proyecto destinaría para la ejecución del mismo, que consistiría tanto en el tratamiento, análisis e interpretación de los datos y su utilización para la generación de informes que aporten valor, como en otras gestiones de organización para que la empresa pueda consolidarse.

Debido a la situación extraordinaria ocasionada por la pandemia del Covid-19, el trabajo se desarrollará en remoto, por lo que no será necesario utilizar ningún espacio de trabajo o coworking, ya que cada miembro del equipo desarrollará sus funciones desde su propio hogar. No se imputa como coste ya que lo asumirá cada uno de los integrantes del grupo.

El desarrollo de acciones comerciales y de campañas de marketing online a través de RRSS y páginas web supondrá de media unos 350€ mensuales, que supondrán principalmente la creación y mantenimiento de una página web corporativa, así como una estrategia de posicionamiento SEO a través de Google Adwords.

En cuanto a los costes de la tecnología usada, se componen básicamente del coste de las licencias del servicio de almacenamiento en la nube y del gestor de BBDD y de las licencias del programa de visualización Tableau, y un ordenador portátil para cada miembro del equipo adecuado para el desempeño de sus funciones.

Los costes del primer año de actividad serán más elevados, debido a los costes de constitución de la sociedad y a la adquisición de aquellos bienes y servicios necesarios para arrancar la empresa. Se estima que en los años siguientes los costes se reduzcan, al haber amortizado parte de los gastos de la inversión inicial.

4.2.9. Fuentes de ingresos y beneficios

Los ingresos de Geopymes procederán de los dos tipos de contratos que se establezcan con los clientes:

- Contrato de **servicio por consulta** (para los clientes particulares)
- Contrato por **suscripción mensual o anual** (para Grandes Cuentas)

Adicionalmente, para los servicios de Grandes Cuentas, se puede incluir una opción **Premium** en la que, además de las visualizaciones incluidas en las otras opciones, se puedan incluir informes más detallados, ampliando el número de variables analizadas como precio medio alquiler de la zona, e incluyendo servicios adicionales como búsqueda de locales en alquiler en la zona recomendada, elaboración del presupuesto de puesta en marcha del futuro negocio, listar empresas de proveedores y suministros, inclusión de variables financieras sobre empresas etc.



| Especificaciones de la herramienta Geopymes | Plan Individual | Plan GGCC | Premium |
|--|-----------------|-----------|---------|
| Acceso online a la plataforma | ✓ | ✓ | ✓ |
| Perfiles demográficos y socioeconómicos por secciones censales | ✓ | ✓ | ✓ |
| Segmentación de actividades por clusters de recomendación | ✓ | ✓ | ✓ |
| Evaluación comparativa de ubicaciones | ✓ | ✓ | ✓ |
| Puntuación de las localizaciones en función de las variables seleccionadas | ✓ | ✓ | ✓ |
| Visualizaciones interactivas de datos: mapas, cuadros, gráficos | ✓ | ✓ | ✓ |
| Indicadores clave de rendimiento (KPIs) | ✓ | ✓ | ✓ |
| Descarga de informes en PDF | ✓ | ✓ | ✓ |
| Soporte básico por teléfono y correo electrónico | ✓ | ✓ | ✓ |
| Notificaciones con las actualizaciones del mercado (precios alquiler, altas y bajas negocios...) | | ✓ | ✓ |
| Elaboración de nuevos perfiles de clientes | | ✓ | ✓ |

| Especificaciones de la herramienta Geopymes | Plan Individual | Plan GGCC | Premium |
|---|-----------------|-----------|---------|
| Análisis de penetración de mercado | | | ✓ |
| Análisis de la competencia | | | ✓ |
| Evaluación del potencial y pronóstico de ventas de la ubicación elegida | | | ✓ |
| Inclusión de datos económicos de empresas | | | ✓ |
| Información sobre alquiler de locales | | | ✓ |

Figura 23: Especificaciones de la herramienta Geopymes. Fuente: Elaboración propia.

La estimación sobre la captación anual de cada tipo de cliente para un escenario de 5 años sería la siguiente:

| Clientes | Año 1 | Año 2 | Año 3 | Año 4 | Año 5 |
|----------------------|-------|-------|-------|-------|-------|
| Particulares | 200 | 300 | 400 | 600 | 900 |
| GGCC | 2 | 6 | 9 | 12 | 16 |
| Suscriptores Premium | 0 | 0 | 0 | 4 | 7 |

En función de los costes e ingresos anuales anteriormente descritos, se muestra a continuación una previsión del margen para los próximos cinco años:

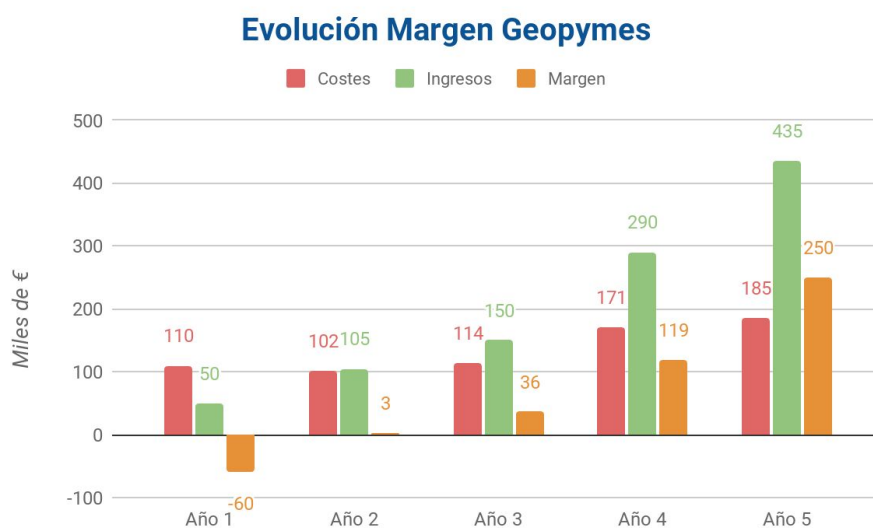


Figura 24: Evolutivo del margen esperado. Fuente: Elaboración propia.

4.3. Plan de acción

4.3.1 Definición del alcance del proyecto: objetivos y métricas

Misión, visión y objetivos

Misión

Nuestra misión consiste en asesorar y guiar a las personas que se aventuran en abrir un negocio en una de las cuestiones clave en las que basa su decisión de emprender: dónde situarlo para que tenga mayor probabilidad de éxito.

Visión

Geopymes pretende ser un referente de ayuda y apoyo para las pequeñas y medianas empresas en su proceso de inicio y consolidación de sus negocios.

Objetivo General

Presentar una herramienta de visualización geográfica basada en metodologías de Big Data y Business Intelligence que, mediante el análisis de datos sociodemográficos, económicos y urbanísticos de las zonas consideradas, presente a los emprendedores información relevante sobre sus potenciales competidores y su mercado objetivo.

Objetivos y estrategias específicos:

Objetivo 1: Conseguir una cartera sólida de clientes con suscripción mensual / anual o Premium, que nos aporte el 70% de nuestros ingresos anuales y aporte solidez financiera al proyecto.

- Realizar un plan de marketing intenso en el primer año, para captar más clientes, de ambos segmentos.
- Fidelizar y retener a los clientes con servicios de valor añadido adicionales.

Objetivo 2: Aumentar el margen neto en torno al 20% en los 5 primeros años.

- Aumentar el volumen de ingresos a través de las ventas, priorizando las suscripciones anuales frente a las mensuales o a las consultas individuales.
- Reducir los gastos corrientes, para mejorar el margen de explotación.

Objetivo 3: Posicionarnos como una empresa clave del sector en el corto o medio plazo.

- Ofrecer el mejor producto posible, actualizando nuestra herramienta con las últimas novedades tecnológicas e incorporando nuevas funcionalidades.
- Proporcionar en todo momento un servicio de atención al cliente de calidad, tanto en el proceso de captación como en el soporte postventa y la resolución de incidencias.
- Adquirir nuevas competencias y habilidades que permitan abarcar otras áreas de negocio, mediante la formación continua de nuestro equipo.

Métricas

| Objetivos | Estrategias | Indicador clave (KPIs) | Evaluación |
|---|---|---|------------|
| Consolidación de la cartera de clientes | Realizar un plan de marketing intensivo en el primer año para captar clientes de ambos segmentos. | <ul style="list-style-type: none"> Análisis del marketing vía email: nº de personas que abren el correo o que hacen click en dicho correo, bajas en el servicio, correos en spam, suscripciones a la newsletter, ... Análisis del impacto de nuestra actividad en RRSS: nº likes, tweets, respuestas, hashtags, suscriptores, ... Menciones y/o apariciones en medios de comunicación (físicos o digitales): blogs, webs especializadas, prensa, RRSS,... ROI campañas marketing: (Beneficio - Inversión)/Inversión | Trimestral |
| | Fidelizar y retener a los clientes | <ul style="list-style-type: none"> % Retención clientes: Clientes renovados/Clientes totales. % Churn o tasa de rotación de clientes: Clientes perdidos / Clientes totales Análisis del uso de la herramienta por los clientes: Nº Clientes que cambian su tipo de suscripción por una superior, frecuencia de uso de la aplicación, que funcionalidades o características les resultan más útiles, ... | Mensual |
| Aumentar el margen neto en torno al 20%/anual en los 5 primeros años | Aumentar el volumen de ingresos a través de las ventas, priorizando las suscripciones anuales frente a las mensuales o a las consultas individuales. | <ul style="list-style-type: none"> Customer Lifetime Value (LTV): indica el margen bruto que sacamos a cada cliente a lo largo del tiempo que está con nosotros = (Ingresos medios/mes x margen bruto medio) x ciclo vida en meses Average Revenue Per User (ARPU) = Total ingresos obtenidos / Total de usuarios activos | Mensual |
| | Mejora la composición de la estructura de costes para mejorar el margen de explotación. | <ul style="list-style-type: none"> Coste de adquisición de nuevos Clientes (CAC) = Costes comerciales / Clientes nuevos Gastos de explotación / Total de gastos Costes fijos / Costes variables | Mensual |
| Posicionarnos como una empresa clave del sector en el corto o medio plazo | Ofrecer el mejor producto posible, actualizando nuestra herramienta con las últimas novedades tecnológicas e incorporando nuevas funcionalidades. | <ul style="list-style-type: none"> Reuniones con nuestra cartera de clientes a largo plazo para informales de las nuevas funcionalidades de la herramienta | Trimestral |
| | Proporcionar en todo momento un servicio de atención al cliente de calidad, tanto en el proceso de captación como en el soporte postventa y la resolución de incidencias. | <p>Encuestas de satisfacción al cliente, que evalúen si la atención recibida ha sido deficiente, correcta o sugiera algún cambio.</p> <p>Resolución de incidencias:</p> <ul style="list-style-type: none"> Backlog de incidencias y categorización Nº incidencias resueltas / Total incidencias Tiempo medio resolución incidencia | Mensual |
| | Adquirir nuevas competencias y habilidades que permitan abarcar otras áreas de negocio, mediante la formación continua de nuestro equipo. | <p>Realización de cursos:</p> <ul style="list-style-type: none"> (Cursos completados / Total cursos) por empleado Número horas formación/personal total <p>Obtención de certificaciones:</p> <ul style="list-style-type: none"> Certificaciones conseguidas por empleado | Anual |

4.3.2 Análisis de actividades: modelo lógico - arquitectura técnica

4.3.2.1 Modelo lógico

A continuación se detallan las diferentes fases que se abarcan dentro del modelo lógico de nuestro proyecto:



Figura 25: Modelo lógico. Fuente: Elaboración propia.

Se definen cinco procesos clave por los que iremos avanzando hasta llegar a la solución, desde la identificación de los datos hasta el proceso de visualización final:



Figura 26: Modelo procesos. Fuente: Elaboración propia.

4.3.2.1.1. Identificación de los datos

Los datos utilizados para el desarrollo de la herramienta provienen en su totalidad de fuentes Opensource, es decir, son de dominio público, pudiendo ser utilizados sin tener que pagar por ellos. A continuación, se detallan las fuentes de las que hemos obtenido la información:

- ❑ Datos sociodemográficos: recopilados del INE y del Portal de datos abiertos del Ayuntamiento de Madrid.

- Información geoespacial de puntos de interés (POI): se han recogido puntos de educación, ocio, turismo, seguridad y transporte entre otros, de madrid.org/nomecalles.

Los datos obtenidos pueden descargarse en ficheros con formato .csv o .xls directamente desde la web, mediante técnicas de scraping o utilizando APIs para la consulta directa de la información, que pueden programarse para aquellos datos cuya actualización sea periódica en el corto y medio plazo. Para aquellos datos cuya actualización no sea tan frecuente, bastará con tener en cuenta en qué fecha se publican los nuevos datos para programar su integración en la base de datos.

La propuesta de Geopymes se desarrollará como un servicio en la nube, para que así pueda estar siempre disponible, actualizada y accesible a través de un navegador de internet, permitiendo el acceso a las distintas funcionalidades del sistema en función del perfil de usuario que acceda.

4.3.2.1.2. Extracción, transformación y validación de los datos (ETL)

El proceso de Extracción, Transformación y Carga no fue fácil debido a que la información no estaba en el formato adecuado y el encoding correcto, por lo que en la mayoría de los casos se tuvo que colocar la información en el Encoding UTF-8 para que la misma fuera leída apropiadamente y subida al repositorio.

Para subir la información se utilizó el comando LOAD DATA (véase Figura 27), leyendo los archivos .CSV de una carpeta local de la computadora.

```
LOAD DATA LOCAL INFILE 'C:/...../datos_censales_pob_activa_parados.csv'
INTO TABLE datos_censales_pob_activa_parados
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 ROWS;
```

Figura 27: Ejemplo de script para carga de datos al repositorio MySQL. Fuente: Elaboración propia.

Otra de las fuentes de datos importantes son los archivos espaciales o Shapefiles cuya carga al repositorio MySQL se hizo con la herramienta ogr2ogr (véase Figura 28), la cual viene incluida al instalar QGIS o se puede hacer su instalación por separado.

```
C:\Program Files\QGIS 3.14\bin>ogr2ogr -f "MySQL"
MYSQL:"geopymes,host=34.72.143.222,user=geopymes_workspace,password=geoeoi07$,port=3306" -
nln "mercadona" -a_srs "EPSG:3857" "C:\Users\Rafa\Documents\GIS Projects\Proyecto Fin
Master\Comercio_Mercadillos - MySQL Load Test\mercadillos - MySQL Load Test.shp"
```

Figura 28: Ejemplo de script para carga de datos de archivos .SHP al repositorio MySQL. Fuente: Elaboración propia.

Muchos de los problemas a los cuales nos enfrentamos fue que la información venía con caracteres no permitidos en la base de datos, números con comas, valores mal formateados en los campos, y distintos niveles de agregación (barrio, distrito, coordenadas local, sección censal, actividad...)

Una de las ventajas de almacenar información espacial directamente en la base de datos es que facilita su visualización en software de tratamiento de datos geoespacial como QGIS y su posterior lectura en software de análisis de datos como R.

Entre la información de la cual se hizo un proceso ETL fue:

1. Información de locales del Ayuntamiento de Madrid

El Ayuntamiento de Madrid pone a disposición tres ficheros con información sobre locales y actividades. Tras realizar un análisis en detalle de la información contenida en éstos, se procedió a elegir el más adecuado para nuestro objetivo, en función de las variables que contenían y la consistencia de las mismas.

Se cargó el archivo .CSV donde se almacenaba información relevante para cada local como son sus coordenadas geográficas, si el local está abierto o cerrado, el tipo de acceso, información de epígrafe, entre otras.

Este archivo contiene información de todos los locales y actividades. Dado que nuestro proyecto se centra en comercios, tuvimos que hacer una clasificación en función del epígrafe (clasificación cuya base parte del código CNAE) de modo que quedara identificado qué era un comercio (papelería, ferretería, bar-restaurante, etc) y qué era, lo que hemos denominado, punto de interés (escuelas, transportes, residencias, ...). De este modo pudimos aislar los comercios, nuestros datos de partida para el proyecto.

2. Información de Secciones Censales

Para poder integrar la información de las secciones censales al dataset de locales del municipio de Madrid, se presentó un problema, dado que la información de los locales no tenía esta agrupación. Para solucionar dicho problema se procedió a determinar el polígono de la sección censal que contenía cada local, auxiliándose para esto por medio de análisis geoespaciales realizados en QGIS (véase Figura 29) y en R, donde finalmente se hizo la integración a partir de las coordenadas de los locales.

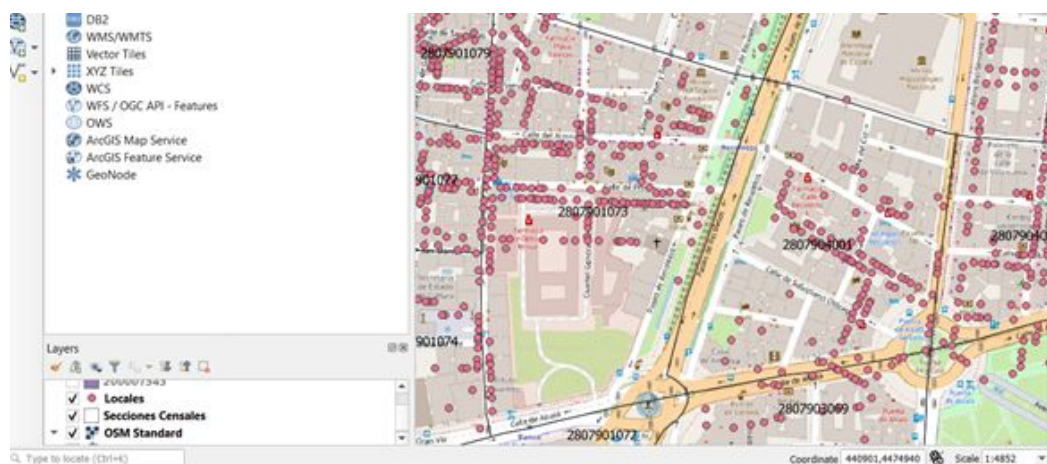


Figura 29: Visualización de locales sobre capa de secciones censales del municipio de Madrid. Fuente: Elaboración propia.

El procedimiento que se realizó en R consistió en colocar en un mismo CRS las capas de Secciones Censales y Locales, y posteriormente aplicar el método Over de la librería SP, de esta manera obteniendo el valor de la Sección Censal del polígono donde se encontraba contenida la coordenada UTM de cada local.

3. Puntos de Interés

Se cargaron las capas (archivos .SHP) de puntos de interés del portal madrid.org en la base de datos y posteriormente se hizo la consolidación de los mismos en una sola tabla. Se utilizó la herramienta R para analizar la cantidad de puntos de interés en un radio de 250 metros cercano a cada local (ver Anexo II).

Posteriormente estos puntos de interés fueron agrupados en categorías macros con mira a introducirlas al modelo y adicionalmente se generaron variables categóricas con el nivel de tránsito de acuerdo al tipo de punto de interés, por ejemplo la variable tránsito alto está construida con puntos de ocio (cines, teatros...), puntos comerciales (centros comerciales, mercados...), puntos de turismo (museos, hoteles...) y puntos de transporte (metros, parking...), consideramos tránsito medio los puntos de educación (escuelas infantiles, colegios...) y salud (ambulatorios...) y tránsito bajo los puntos de seguridad (policía, bomberos...) y organismos públicos (ver Anexo II).

4. Información Sociodemográfica

Construimos un dataset con variables sociodemográficas muy amplio a nivel barrio y distrito del que pudimos estudiar variables tales como estructura de la población y de los hogares, indicadores económicos, de desempleo, educación, salud, calidad de vida... La alta correlación entre las variables, la distribución de los datos y el nivel de agregación, que nos pareció muy amplio en términos geoespaciales para hacer una recomendación de ubicación, nos hizo descartar este dataset y proceder a la carga de información sociodemográfica procedente del INE a nivel de sección censal. Incluimos variables como superficie, población, renta media, rango de edades de los habitantes,... entre otras variables importantes ya identificadas como tales en el estudio anterior.

A partir de algunas de estas variables hicimos transformaciones para obtener otras nuevas como son la tasa de paro o la densidad de población.

5. Información de Competidores

Para obtener la información de competidores (ver Anexo II), se hizo un análisis geoespacial de los locales por tipo de actividad que estaban en un radio de 250 metros a cada local, utilizando como información base la fuente de locales explicada anteriormente. El proceso de transformación consistió en relacionar las distintas tablas que se cargaron al repositorio y la agregación de las métricas más relevantes a nivel de Sección Censal, consiguiendo con ello al final un conjunto de datos consolidado en una tabla, lista para ser leída en la herramienta R.

Debido a los requerimientos actuales y limitaciones de tiempo, para llevar a cabo el cálculo de los clusters, la carga de los ficheros de texto y archivos espaciales se hizo de forma manual mediante lenguaje SQL, sin embargo, dichos scripts están listos para ser codificados en procedimientos almacenados de MySQL y calendarizados con la herramienta Cloud Scheduler de Google Cloud si así se requiere.

Otra alternativa podría ser utilizar el método csv del objeto read desde R (véase Figura 30), el cual toma un fichero directamente desde una URL y lo carga en un dataframe, posteriormente se escribe la información en el repositorio MySQL para su almacenamiento.

```
read.csv(file="https://datos.madrid.es/egob/catalogo/200085-6-censo-locales.csv",encoding="utf8")
```

Figura 30: Ejemplo de código en R para lectura de datos abiertos desde un sitio web. Fuente: Elaboración propia.

4.3.2.1.3. Almacenamiento de los datos

El almacenamiento de los datos se hizo en una instancia MySQL versión 5.7 alojada en la nube, utilizando para ello la plataforma Google Cloud.

Esto facilitó el acceso a la base de datos desde cualquier parte del mundo, teniendo en cuenta que MySQL tiene muchas facilidades para que cualquier herramienta de análisis o visualización de datos lea de ella de forma amigable.

Para el almacenamiento de la información se genera un modelo relacional con tablas de hechos y dimensiones.

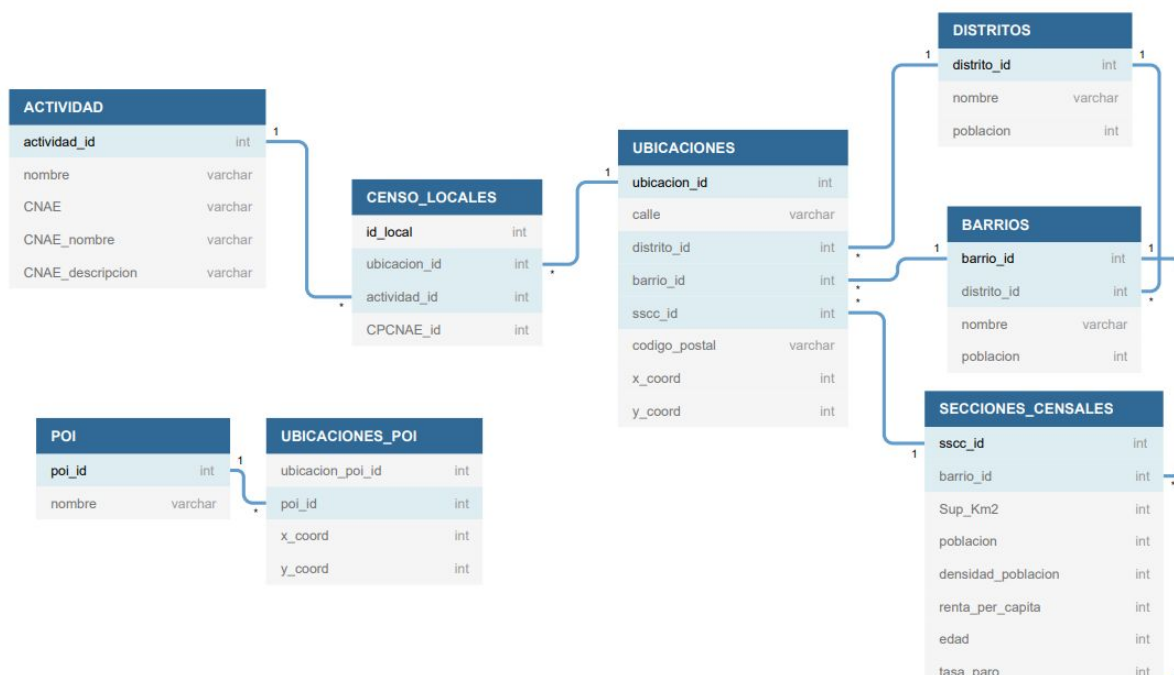


Figura 31: Modelo relacional usado en el proyecto. Fuente: Elaboración propia.

4.3.2.1.4. Análisis exploratorio y modelado

Ya con las distintas tablas del modelo relacional almacenadas en la base de datos MySQL, se procedió con la carga del dataset final hacia Rstudio, el cual será la herramienta de modelado que se utilizará.

Se hizo un análisis exploratorio de las variables de la siguiente manera:

1. Eliminación de valores atípicos (Outliers)

Se define como valor atípico leve aquel que dista 1,5 veces el rango intercuartílico por debajo de Q1 o por encima de Q3.

Al igual que la mediana, los cuartiles dejan resultados a un lado y al otro pero dividiendo la muestra en cuatro partes. El 25% de los resultados son menores que el primer cuartil (Q1), el 50% son menores que Q2 (es la mediana) y el 75% menores que el tercer cuartil (Q3).

Una forma fácil de detectar outliers es con un diagrama de caja y bigotes. En el diagrama vemos la caja, cuyo borde superior es Q3 y el inferior es Q1. Entre medias están el 50% de las ocurrencias. La altura de la caja es el rango intercuartílico. Y el bigote, la línea gruesa, es la mediana. Por encima y debajo se ven dos límites, que son los umbrales para los valores atípicos.

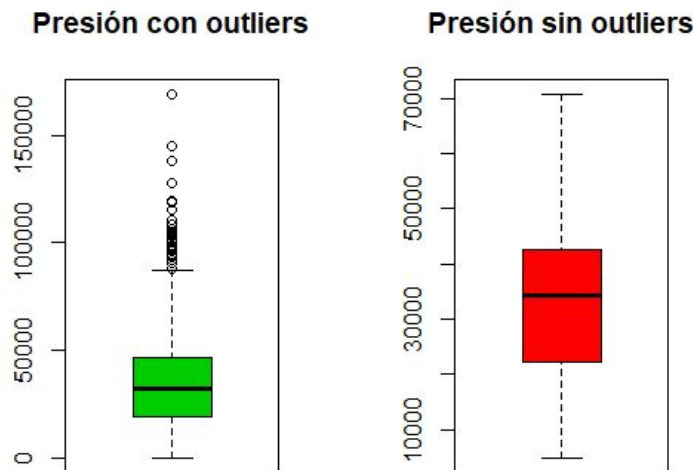


Figura 32: Diagrama de cajas y bigotes. Fuente: Elaboración propia.

Una vez detectados los outliers de nuestro dataset definimos nuestra propia función que recibe como parámetros el conjunto de datos, el umbral inferior y el umbral superior. Y en la función decidimos reemplazar por la media aquellos valores atípicos que estén por debajo del umbral inferior, y por la mediana aquellos que estén por encima del umbral superior.

Realizamos este proceso con todas las variables sociodemográficas donde detectamos outliers.

2. Análisis de correlación de variables

Nuestro dataset parte de más de 135 variables del que sería muy complicado analizar y perfilar nuestros clusters en la parte de modelado y además habría variables que nos estarían aportando la misma información. Nuestro objetivo en este apartado es eliminar las variables con una alta correlación.

La correlación es la forma numérica en la que la estadística ha podido evaluar la relación de dos o más variables, es decir, mide la dependencia de una variable con respecto de otra variable independiente. Para ello usamos el índice de correlación que puede variar entre -1 y +1, ambos extremos indicando correlaciones perfectas, negativa y positiva respectivamente. Un valor de $r = 0$ indica que no existe relación lineal entre las dos variables. Una correlación positiva indica que ambas variables varían en el mismo sentido y negativa en sentido contrario. Si hiciéramos este ejercicio con las 135 variables no podríamos sacar unas conclusiones claras. Para ello, calcularemos la matriz de correlación e iremos analizando gráficamente por los diferentes grupos de variables cuál es su correlación.

Para las variables sociodemográficas observamos que sí existen correlaciones entre las variables de sexo-edad-nacionalidad y con estudios-edad de 45 a 64 años.

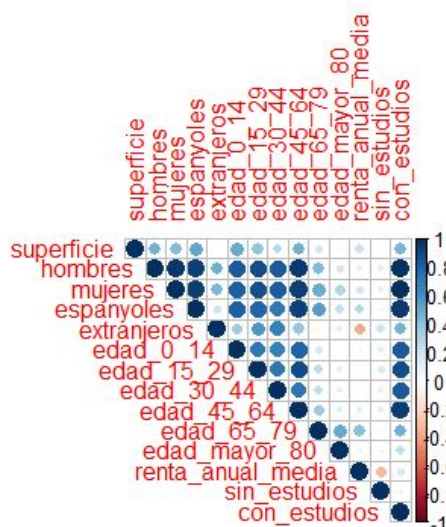


Figura 33: Salida de la matriz de correlación. Fuente: Elaboración propia.

Para las variables de todos los comercios a nivel SSCC, vemos correlaciones entre alimentación, bazar, bar-restaurant y peluquería seguramente asociado a la densidad de población de la zona donde es habitual que concurran este tipo de comercios. Por lo que esperamos a la importancia de las variables para tomar una decisión.

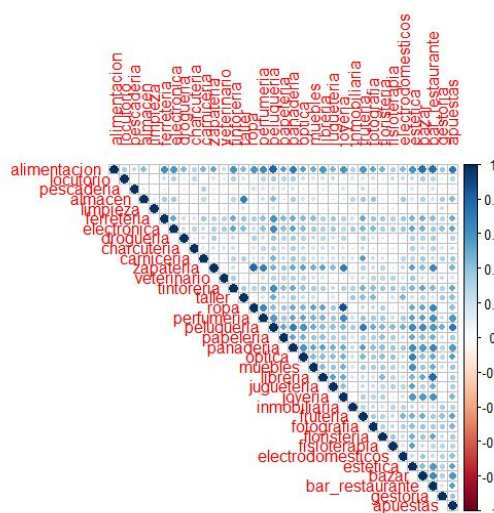


Figura 34: Salida de la matriz de correlación. Fuente: Elaboración propia.

Finalmente se procede a generar una matriz de correlación de actividades agrupadas a nivel código postal buscando más representatividad de comercios geográficamente, de modo que nos facilite la elección de variables de importancia que explicaremos en el siguiente apartado.

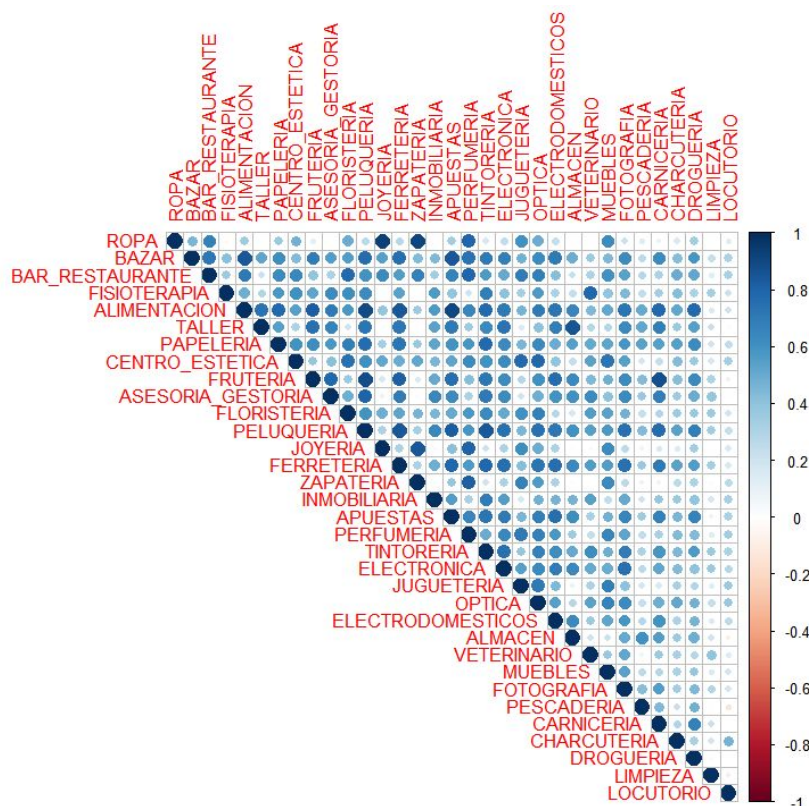


Figura 35: Salida de la matriz de correlación. Fuente: Elaboración propia.

3. Selección de variables según su importancia

Una vez finalizados los análisis de correlación, se procede a realizar análisis de importancia de variables de cara a obtener las variables más significativas para cada actividad. Estas serán las que incluiremos en el proceso posterior de clusterización.

El objetivo de realizar esta selección es facilitar la agrupación y perfilado del clustering, permitiendo así discriminar de una manera clara entre los diferentes cluster y por tanto facilitando una explicación clara y entendible a nuestro cliente sobre los criterios en los que se fundamenta la recomendación de ubicación.

Con la combinación del resultado obtenido de los análisis de importancia de variables (para los que se han utilizado diferentes métodos), correlaciones y nuestro conocimiento sobre la ubicación de los comercios en el municipio de Madrid, se procede a seleccionar las variables finales a incluir en el proceso de clusterización.

Para poder ejecutar los análisis de importancia, se genera previamente un dataset por cada actividad, agrupando a nivel código postal y actividad las variables. A pesar de que la unidad de detalle del proyecto es a nivel sección censal, se hace necesario para este

ejercicio generar la vista código postal para poder obtener mayor representatividad de las actividades en una zona geográfica.

R nos ofrece un amplio abanico de algoritmos para ayudarnos en este sentido, algunos de ellos son Boruta, Random Forest, H2O...

A continuación se muestra, a modo de ejemplo, el resultado de ejecución de los diferentes métodos y elección final de variables para la actividad “joyería”.

H2O

Gracias a su forma de comprimir y almacenar los datos, H2O es capaz de trabajar con millones de registros en un único ordenador. Aunque la principal ventaja de H2O frente a otras herramientas es su escalabilidad, sus algoritmos son igualmente útiles cuando se trabaja con un volumen de datos reducido.

El bosque aleatorio distribuido (DRF) es una potente herramienta de clasificación y regresión. DRF genera un bosque de árboles de clasificación o regresión, en lugar de un solo árbol. Cada uno de estos árboles es un aprendiz débil construido sobre un subconjunto de filas y columnas. Más árboles reducirán la variación (véase figura 36 y 37 variando el número de ntrees). Tanto la clasificación como la regresión toman la predicción promedio de todos sus árboles para hacer una predicción final, ya sea para una clase o para un valor numérico.

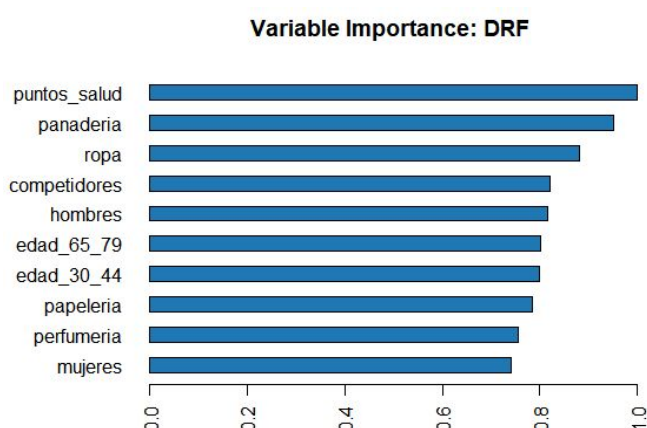


Figura 36: Gráfico de variables de importancia, actividad joyería (ntrees=1.000). Fuente: Elaboración propia.

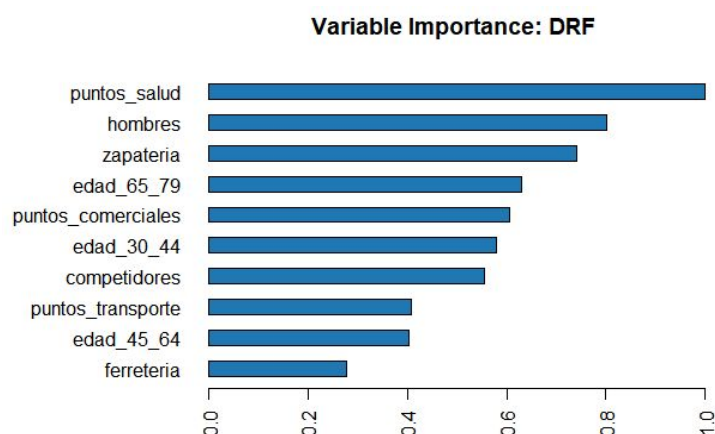


Figura 37: Gráfico de variables de importancia, actividad joyería (ntrees=50). Fuente: Elaboración propia.

Random Forest

Otro método que hemos utilizado es Random Forest con R para medir la importancia de las variables cuantitativas (véase figura 37). Es un método de clasificación basado en la realización de múltiples árboles de decisión sobre nuestro conjunto de datos. La idea es realizar varias clasificaciones con menos variables y menos observaciones y al final quedarse con un promedio de estas clasificaciones. La característica que hace de este método muy interesante es la posibilidad de incluir un gran número de variables input en nuestro modelo.

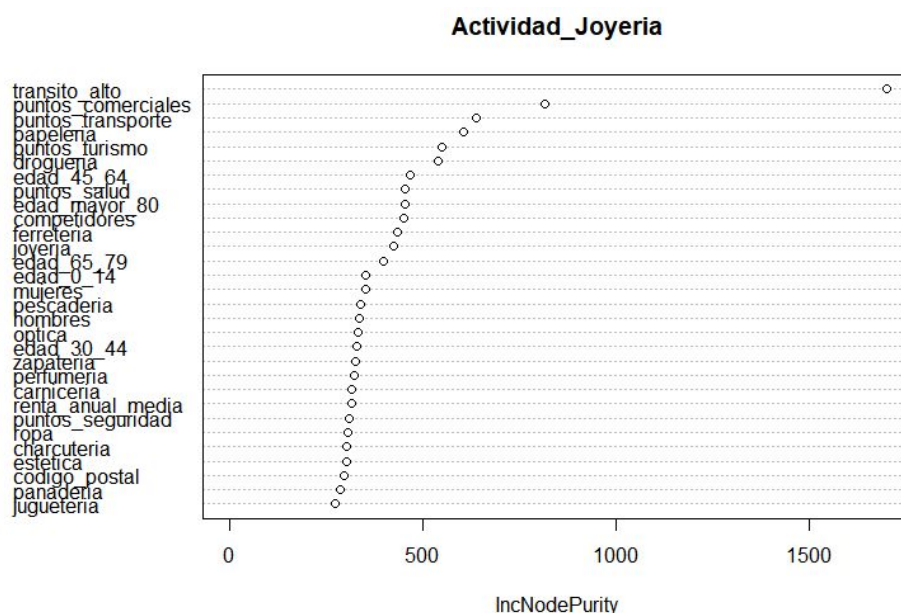


Figura 38: Importancia de variables Random Forest, actividad joyeria (ntrees=1.000). Fuente: Elaboración propia.

Boruta

Boruta es un método que utiliza random forest como algoritmo subyacente. La idea es generar en cada iteración una serie de variables sombra a partir de los predictores, copiando cada uno de ellos y permutando entre sí los elementos de cada nueva columna. Las columnas en verde son las que valoraremos para tomar como variables más importantes.

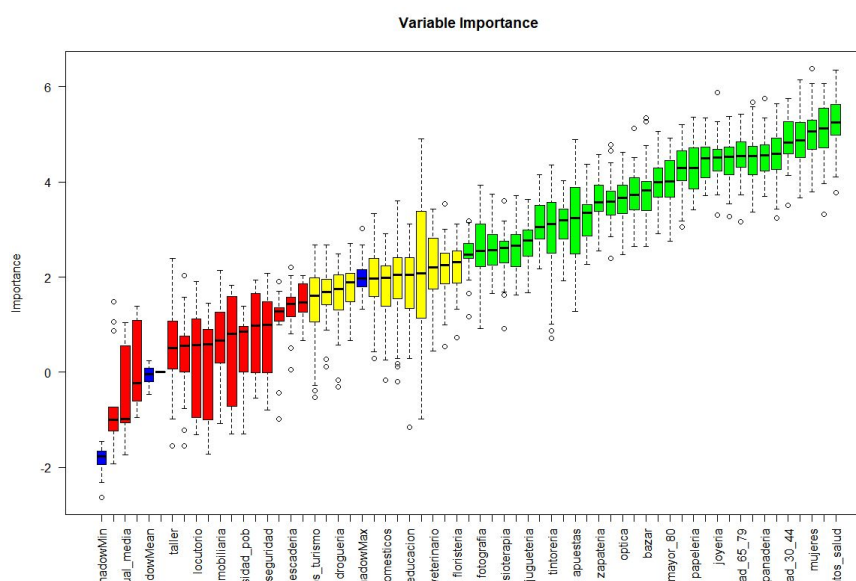


Figura 39: Algoritmo Boruta. Fuente: Elaboración propia.

En función de lo anterior, hemos seleccionado las variables que entrarán en la parte de modelado, donde se entrenará el clustering tomando estas variables.

- Actividad “Joyerías” seleccionamos las siguientes: puntos salud, panadería, ropa, competidores, perfumería, zapatería, puntos comerciales

Del mismo modo, quedarían así las variables para los otros comercios:

- Actividad “Óptica”: *puntos salud, muebles, papelería, edad_65_79, ferretería, puntos educacion, competidores*
- Actividad “Papelería”: *edad_45_64, edad_30_44, puntos educacion, peluquería, frutería, alimentacion, competidores*

Para la prueba de concepto se han seleccionado estas tres actividades, pero en principio no habría ninguna limitación para realizar el ejercicio con el resto de actividades.

Modelado

El proceso de modelado del proyecto Geopymes está basado en técnicas de Aprendizaje Automático No Supervisado, específicamente utilizando un algoritmo de Clustering denominado K-Means. El algoritmo K-Means agrupa las observaciones en K clusters distintos, donde K es determinado por el analista antes de ejecutar el algoritmo. K-Means encuentra los mejores clusters, entendiéndose como mejor aquellos cuya varianza interna es la menor posible. (véase Figura 40)

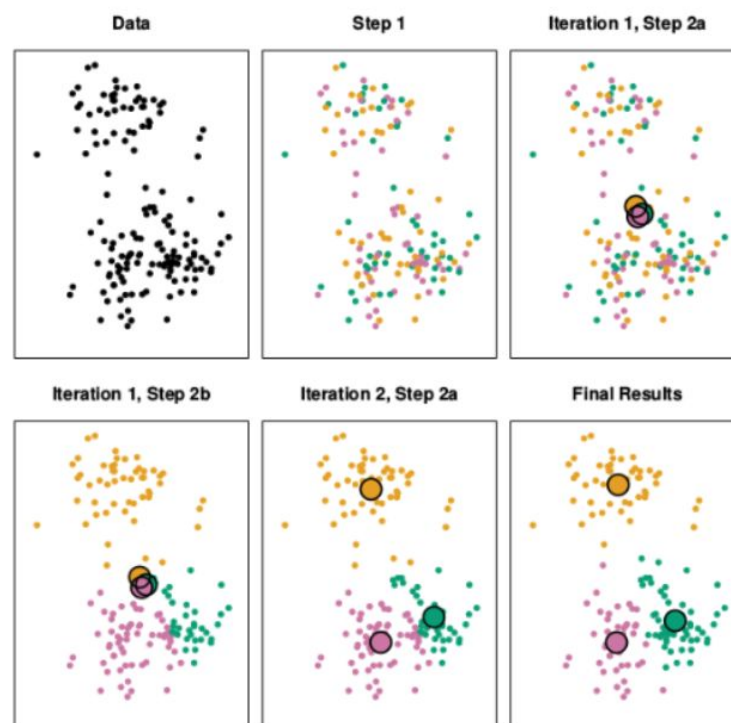


Figura 40: Algoritmo de clustering basado en K-Means. Fuente: (Rodrigo, 2017)

El proceso de Clustering se hará en función de cada actividad como ya vimos en el apartado anterior, teniendo en cuenta que las features varían según la actividad a la que pertenece un local.

El primer paso del proceso de modelado que se hizo fue la estandarización de variables, de manera de colocar todas las variables a una misma escala. El método básicamente obtiene la media y desviación estándar de todo el conjunto de datos y luego normaliza cada elemento restando de la media y dividiéndolo por la desviación estándar.

El segundo paso es la identificación del número idóneo de clusters de acuerdo a la actividad a clusterizar. Este método consiste en ejecutar el algoritmo de K-Means k veces donde k es el número de clusters y analizar la suma del cuadrado de los errores en el algoritmo de clustering con cada número de clusters. (véase Figura 41). Al graficar los resultados del algoritmo obtendremos una gráfica en forma de un brazo, y justo en el “codo” del brazo se encuentra localizado el número óptimo de clusters donde se minimiza el error de los puntos de cada cluster respecto al centroide de este.

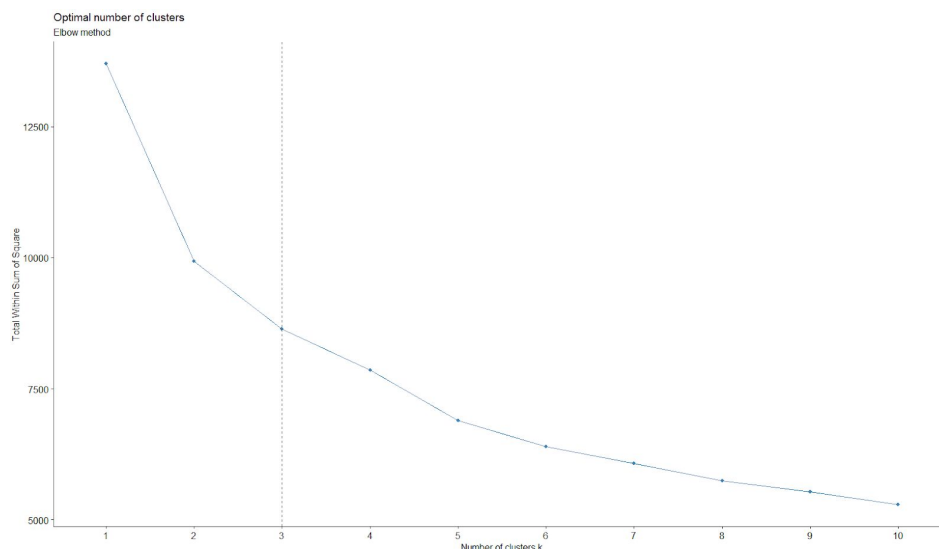


Figura 41: Representación gráfica del resultado de la ejecución del método Elbow con un conjunto de datos de prueba. Fuente: Elaboración propia.

Como se mencionó anteriormente, el número óptimo de cluster podría variar según cada actividad empresarial, debido a que las variables con las que se entrena el modelo de K-Means puede variar en función de la actividad.

Como tercer paso, con el número de clúster elegido se procederá a entrenar el algoritmo de K-Means con las variables seleccionadas anteriormente debidamente normalizadas.

El modelo devuelve una lista con los clusters y los centroides de estos, así como otras métricas afines al modelo.

A continuación se muestran los centros de las variables para cada cluster según la actividad:

Joyerías

| | cluster | cant_sccc | puntos_salud | panaderia | ropa | competidores | perfumeria | zapateria | puntos_comerciales |
|---|---------|-----------|--------------|-----------|------|--------------|------------|-----------|--------------------|
| ▶ | 1 | 76 | 6 | 4 | 72 | 78 | 5 | 15 | 1 |
| | 2 | 1588 | 2 | 1 | 2 | 8 | 0 | 0 | 0 |
| | 3 | 620 | 5 | 4 | 15 | 27 | 2 | 3 | 0 |

Ópticas

| | cluster | cant_sccc | puntos_salud | papeleria | edad_65_79 | ferreteria | puntos_educacion | muebles | competidores |
|---|---------|-----------|--------------|-----------|------------|------------|------------------|---------|--------------|
| ▶ | 1 | 553 | 3 | 1 | 267 | 1 | 3 | 1 | 9 |
| | 2 | 994 | 3 | 1 | 132 | 1 | 2 | 1 | 8 |
| | 3 | 737 | 5 | 2 | 163 | 2 | 3 | 4 | 32 |

Papelerías

| | cluster | cant_sccc | edad_45_64 | edad_30_44 | puntos_educacion | competidores | peluqueria | fruteria | alimentacion |
|---|---------|-----------|------------|------------|------------------|--------------|------------|----------|--------------|
| ▶ | 1 | 695 | 348 | 324 | 3 | 32 | 17 | 4 | 22 |
| | 2 | 534 | 567 | 396 | 2 | 8 | 4 | 1 | 7 |
| | 3 | 1055 | 294 | 220 | 2 | 10 | 6 | 2 | 8 |

Al asociar cada clúster a cada Sección Censal del municipio de Madrid y representar los datos a nivel geoespacial en un mapa de secciones censales (ver Figura 42) se observa claramente cada agrupación de secciones censales de acuerdo al clúster al que pertenecen. En este caso específico, se ejecutó el algoritmo con tres clusters, los cuales vemos representados por los colores rosa, naranja y azul.

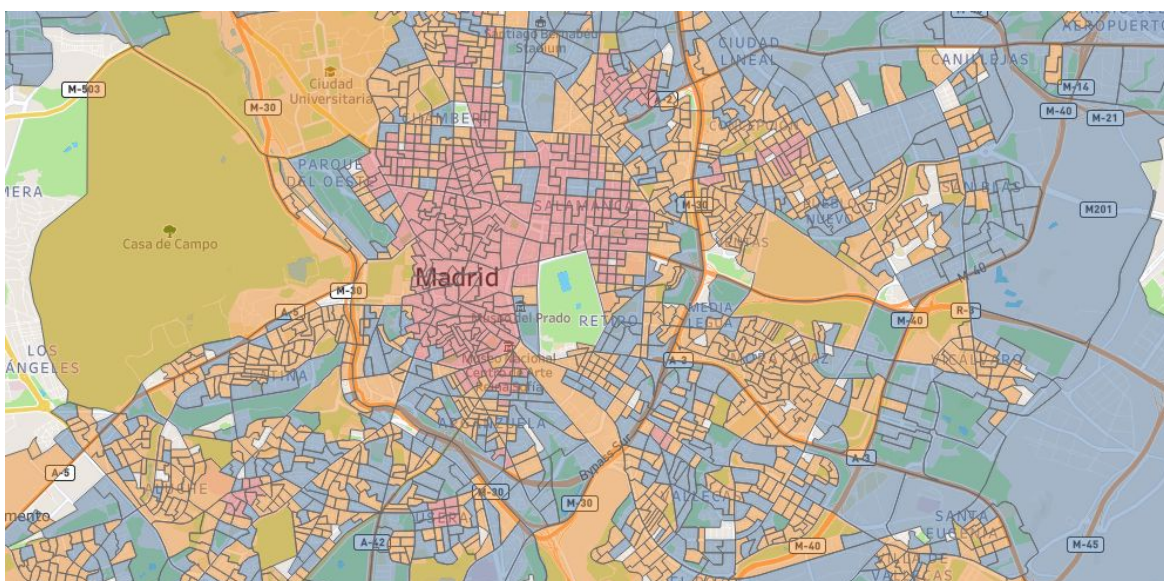


Figura 42: Representación gráfica de los clusters resultantes al ejecutar el modelo K-Means sobre el conjunto de datos de secciones censales del Municipio de Madrid. Fuente: Elaboración propia.

4.3.2.1.5. Visualización

Para mostrar los resultados de los análisis de segmentación realizados y las recomendaciones de zonas que el modelo de datos establece, se ha optado por construir un cuadro de mando en Tableau, el cual estará publicado en el servidor de Tableau de Geopymes, desde donde se gestionarán los accesos de los usuarios, personalizados en función del tipo de servicio o suscripción que hayan contratado.

La herramienta permitirá al usuario, de forma fácil e intuitiva, mostrar la información solicitada de acuerdo a sus preferencias, desplegada tanto de forma georreferenciada en mapas como en tablas, índices y gráficos descriptivos, pudiendo descargar toda esta información en un dossier en pdf para una lectura más detallada.

Una vez el usuario accede a la herramienta, se encuentra con una primera pantalla en la que se muestran en un mapa los clústeres recomendados a nivel de sección censal y, debajo, una descripción de los 3 clusters, clasificados en función del nivel de recomendación sobre qué zona es la más idónea para su negocio.

| Recomendación | Nº SSCC | Puntos Salud | Panaderías | Tiendas Ropa | Competidores | Perfumerías | Zapaterías | Puntos Comerciale |
|---------------|---------|--------------|------------|--------------|--------------|-------------|------------|-------------------|
| Alta | 76 | 5.5 | 4.3 | 71.9 | 78.0 | 5.1 | 14.5 | 0.84 |
| Media | 620 | 5.0 | 3.7 | 14.9 | 27.3 | 2.1 | 3.3 | 0.25 |
| Baja | 1588 | 2.5 | 0.8 | 2.2 | 8.0 | 0.3 | 0.3 | 0.05 |

| Recomendacion | Alta | Media | Baja |
|---------------|---|--|--|
| Cluster | 1 | 3 | 2 |
| Renta Media | 19.878 € | 16.173 € | 14.463 € |
| Competidores | 78 | 27 | 8 |
| POI | Predominan los puntos de Turismo, Ocio, Comercio y Transporte | Segundo en puntos de Turismo, Ocio, Comercios y Transporte | Tiene mas puntos de Salud y Educacion |
| Edad | El que más tiene de 30-64. El que menos de mayores de 65 | Segundo en rango de edad 30-64. | El que más tiene de 0-29 y más mayores pr encima de 65 |
| Transito | Alto | Bajo | Medio |
| Tasa de Paro | 13.95% | 15.07% | 15.21% |
| Genero | Igual proporción de hombres y mujeres | Es el primero que tiene más mujeres que hombres | Segundo en rango de mas mujeres que hombres |
| Nacionalidad | Tiene más extranjeros | Segundo en rango con mas población nacional | Es el que tiene más población nacional |

En concreto, si se selecciona la actividad de “Joyería”, aparecen ordenados los clústeres por el nivel de criterios que han condicionado que el clúster 1 sea el elegido, que son, entre otros, el nivel de Renta Media, la tasa de paro, el número de competidores, el número de Puntos de Interés (POIs), el tránsito por dicha zona,...

Esta zona tiene un gran número de establecimientos comerciales (tiendas de ropa, perfumerías, zapaterías,...) así como un alto número de puntos de interés relacionados con el ocio, turismo y transporte, es decir, es una zona muy bien comunicada y con gran afluencia de gente, al estar en pleno centro y por ser un gran foco comercial, por lo que de cara a recomendarla sería la más ventajosa, lo que la hace interesante de cara a la venta al público.

El rango de edad de los clientes potenciales está entre los 30 y 64 años, con una renta media superior al de las otras opciones, y una tasa de desempleo menor que las otras alternativas, lo que nos puede indicar que los clientes de esta zona son personas con ingresos estables y con una situación familiar consolidada, que pueden permitirse el comprar un producto de más valor añadido.

En cuanto al género, todas las zonas tienen más cantidad de mujeres que de hombres, aunque en la zona de alta recomendación la proporción entre unos y otros es prácticamente la misma, por lo que no se considera que en esta zona sea una variable que influya a la hora de decidirse por este área. Si el cliente desea enfocarse a un público más femenino, entendiendo que son las mujeres las que llevan más joyas, puede elegir ir a la segunda opción, donde la proporción de mujeres es superior al resto de áreas. Pero si opta por la zona de alta recomendación, entiende

que tanto hombres como mujeres adquieren artículos de joyería, ya que tienen un poder adquisitivo alto.

Respecto a la nacionalidad, la zona recomendada es la que tiene mayor proporción de extranjeros residentes que el resto, siendo las otras opciones las que tienen mayor proporción de población nacional residente. Teniendo en cuenta estos factores, nos pueden indicar la elección de la zona de recomendación alta por la mayor presencia de extranjeros residentes en el centro, los cuales tienen que tener una renta media superior a los de otras zonas de residencia para poder vivir allí, y por tanto, pueden gastar más de sus ingresos en productos de lujo.

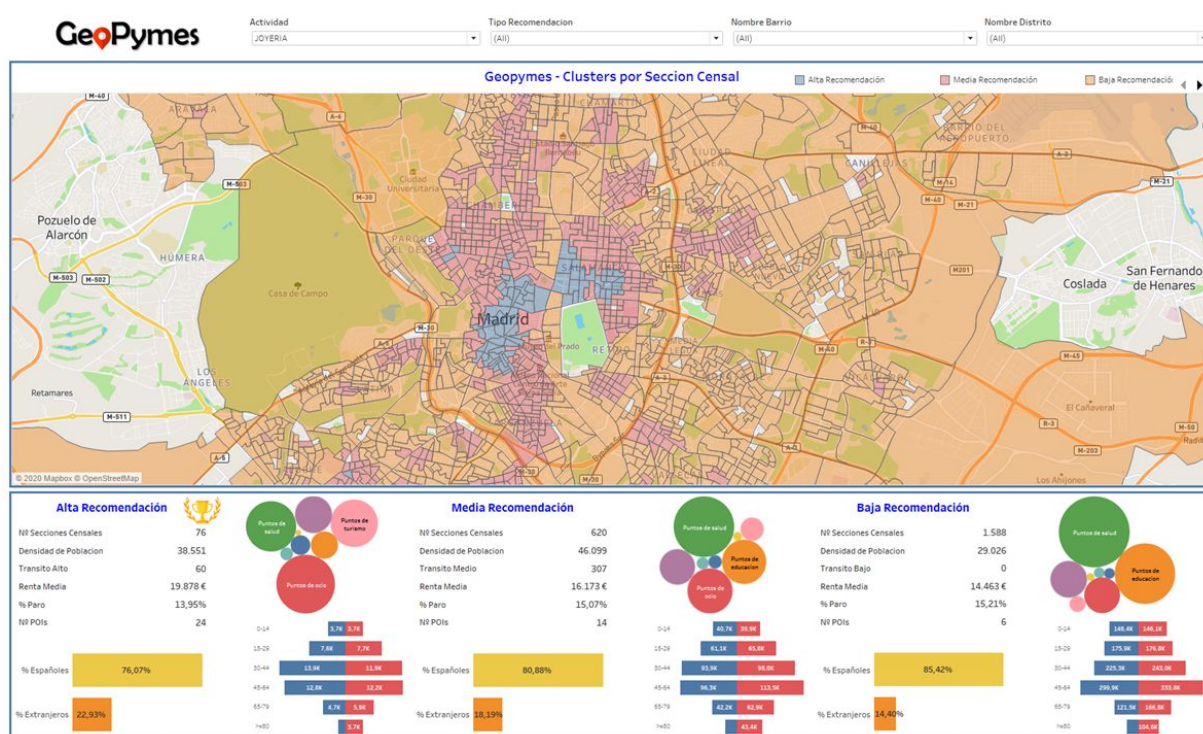


Figura 43: Ejemplo de visualización de la herramienta. Fuente: Elaboración propia.

Si vamos a la siguiente pantalla, se muestra el detalle de la zona deseada, pudiendo seleccionar en el filtro “Tipo Recomendación” de la parte superior cualquiera de las tres alternativas. Siguiendo con el ejemplo anterior, si seleccionamos para la actividad de joyería la zona de “Alta Recomendación”, nos muestra en el mapa el detalle de las secciones censales que conforman dicha zona, así como las características sociodemográficas y del entorno del área mostrada: la población nacional y extranjera, la proporción de hombres y mujeres, el rango de edades que hay en dicha zona, el tipo y la cantidad de comercios, así como de puntos de interés.

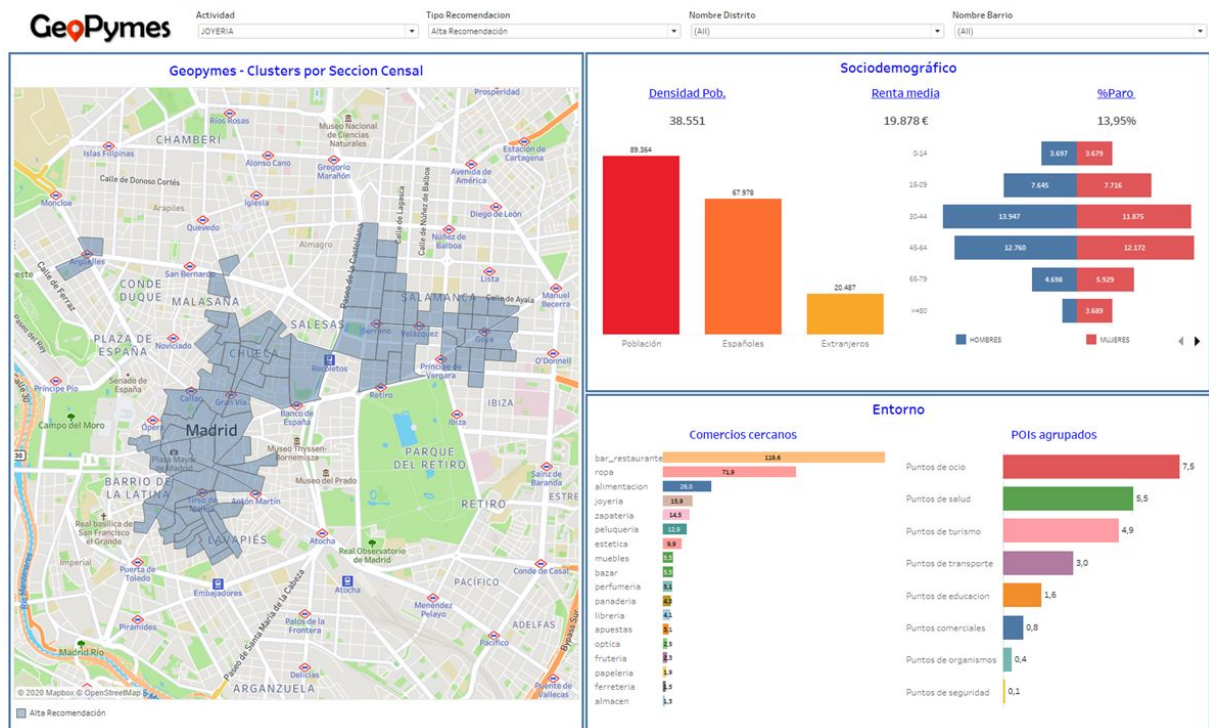


Figura 44: Ejemplo de visualización de la herramienta. Fuente: Elaboración propia.

También es posible seleccionar un distrito o un barrio seleccionandolos en los filtros de arriba, para ver en qué áreas de recomendación están divididos:

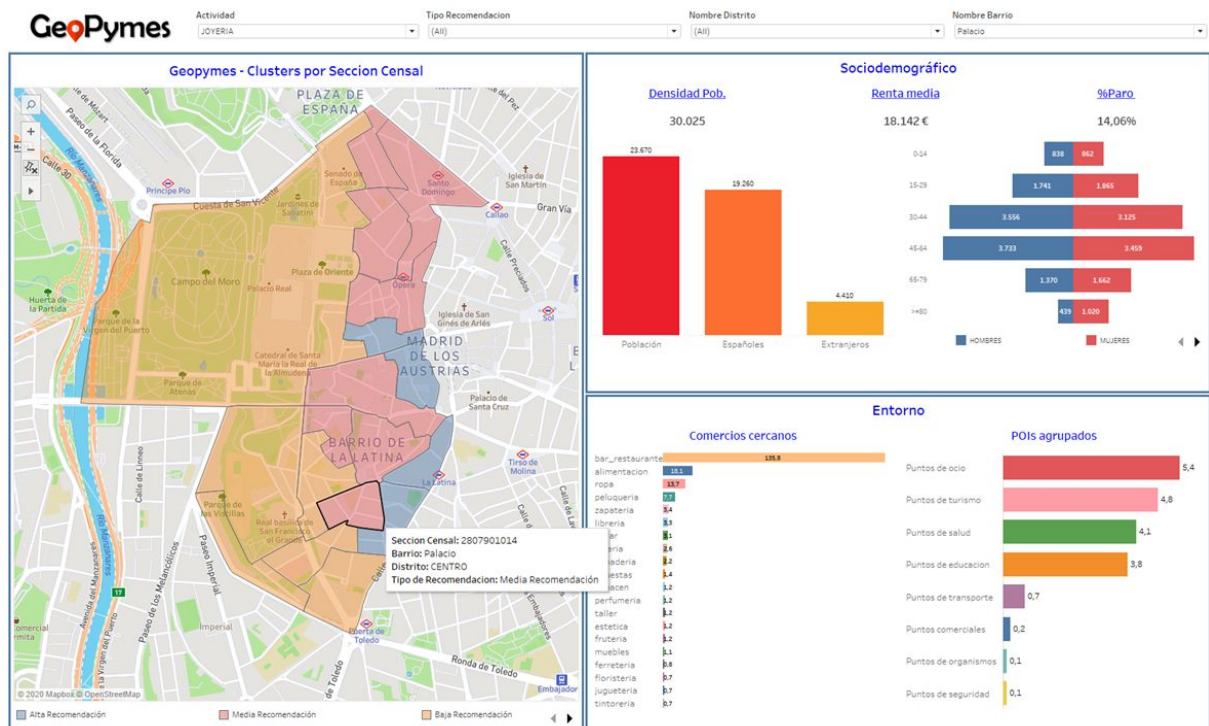


Figura 45: Ejemplo de visualización de la herramienta. Fuente: Elaboración propia.

Por otro lado, la herramienta permite navegar al detalle de la sección censal, pinchando sobre el mapa para seleccionarla y ver cuales son sus características demográficas y de su entorno más cercano :

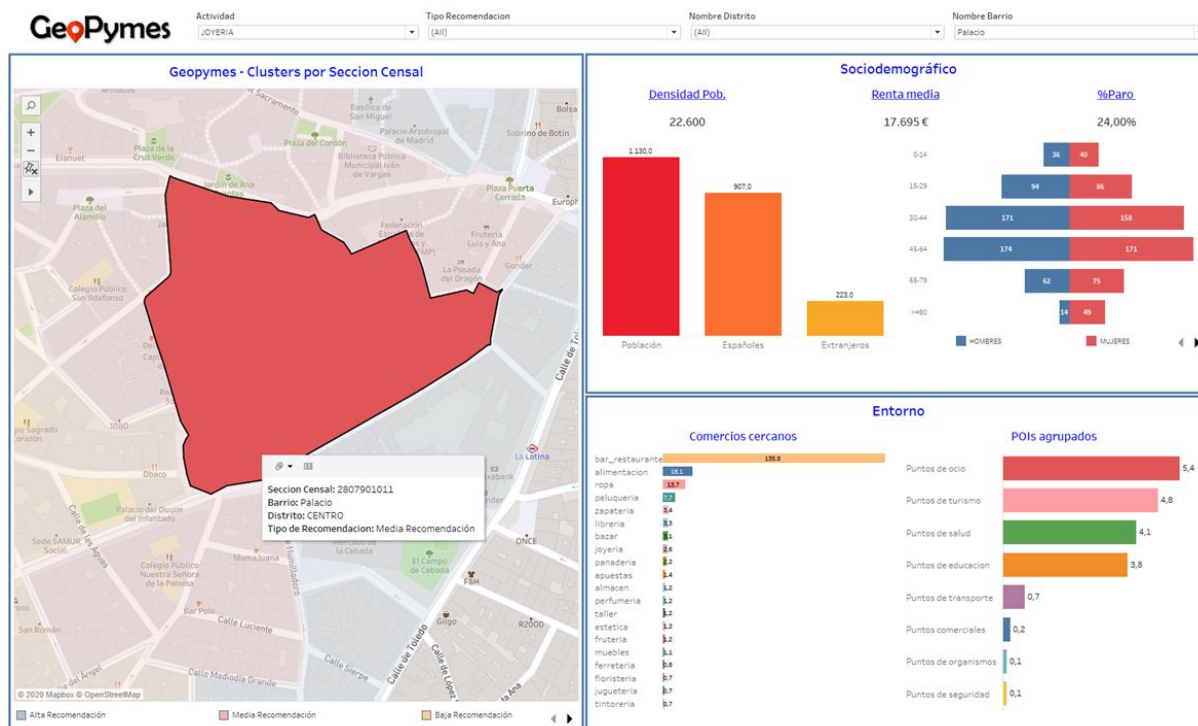


Figura 46: Ejemplo de visualización de la herramienta. Fuente: Elaboración propia.

Tanto los gráficos como los mapas son interactivos, sirviendo de filtros que muestran u ocultan la información mostrada al usuario, quien puede navegar por la herramienta de forma intuitiva no siendo necesario tener que recibir una formación detallada o específica para poder usarla.

4.3.2.2 Solución tecnológica: arquitectura técnica

Definición de la solución

Geopymes estará alojada en Google Cloud y estará disponible al público inicialmente a través de un servidor de Tableau. Antes de llegar la información en clusters a Tableau, se pasará por la arquitectura de Google Cloud, donde se tiene una instancia MySQL version 5.7 (última versión disponible en Google Cloud), y en cuya instancia se hará todo el proceso de ETL de las distintas fuentes de Open Data que alimentarán la solución de Geopymes (ver Figura 45).

Toda la parte de ciencia de datos se hará con la herramienta R, donde se cuenta con la suite de Rstudio para todo lo concerniente a la parte de desarrollo. Entre R y MySQL se estará transfiriendo la información pre procesada, de manera de hacerla persistente en el repositorio de datos y que esta se encuentre lista para su posterior visualización en Tableau.

Tableau es la herramienta de visualización seleccionada, debido a su versatilidad en la toma de datos desde distintas fuentes geoespaciales y estructuradas. Como se mencionó al inicio de este apartado, esta herramienta estará expuesta a través de internet para ser accedida por cualquier usuario autorizado.

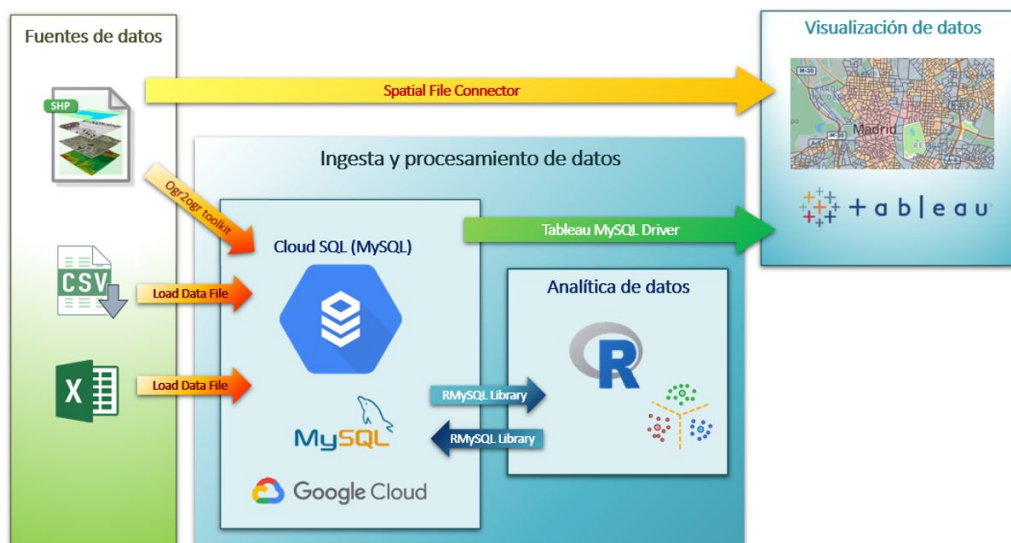


Figura 47: Representación gráfica de la arquitectura del proyecto. Fuente: Elaboración propia.

4.3.3 Análisis de recursos: talento humano y recursos físicos

4.3.3.1 Talento humano

El equipo fundador será el encargado inicialmente tanto de las tareas de desarrollo y soporte de la herramienta, como de aquellas funciones que no estén directamente relacionadas con todo el proceso tecnológico, como la Dirección Estratégica y la Gestión Comercial.

En caso de que por exigencias de los proyectos fuera necesario ampliar la plantilla, se considerará la contratación de personal capacitado para cubrir aquellos puestos que vayan surgiendo. Para ello, se optará por hacer uso de las Bolsas de Empleo de varias instituciones académicas, como la de la EOI, para buscar aquellos perfiles que mejor se adapten a los requerimientos de las vacantes. La estructura organizativa será horizontal y colaborativa, sin que haya ningún área que esté por encima del resto:

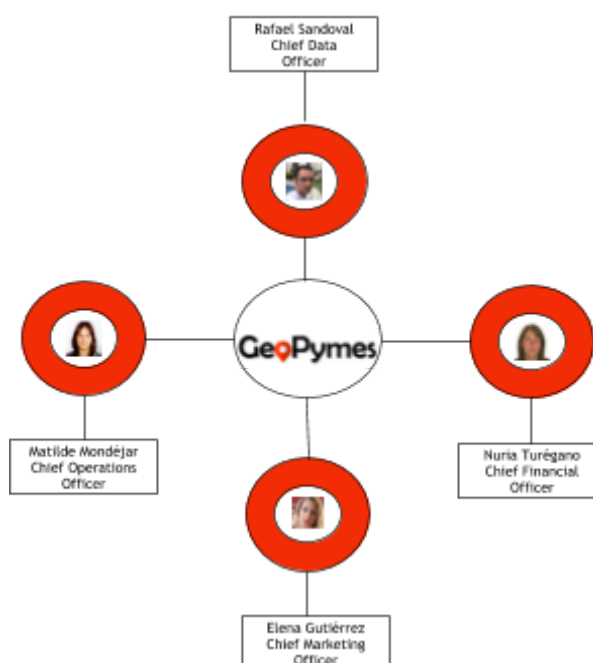


Figura 48: Estructura organizativa. Fuente: Elaboración propia.

En cuanto a las actividades clave dentro del proceso, todos los miembros realizarán las tareas relacionadas con el **tratamiento, análisis y visualización de los datos**, si bien es verdad que cada uno adicionalmente ocupará un cargo en el que desarrollará labores que contribuyan a la buena marcha del proyecto estableciéndose áreas de responsabilidad. Las funciones de cada miembro son:

- ❖ **Chief Data Officer (CDO):** Rafael Alejandro Acosta Sandoval será el responsable de los datos de la empresa al más alto nivel, tanto desde un punto de vista tecnológico como de negocio, incluyendo seguridad. Será quien ayude a gestionar el dato como activo corporativo. Entre sus funciones se incluyen la estrategia para la explotación del dato y gobierno del dato; es decir, definir políticas de seguridad en la gestión y almacenamiento de los datos, políticas de privacidad, así como mantenerse al día con las novedades en temas de regulación que marca el país o, en este caso, la Unión Europea.

Nuestra acción estratégica busca la especialización en el análisis geoespacial de los datos, ofreciendo una solución basada en el BI/BD, que ayude a los futuros emprendedores de las pymes a determinar aquellas áreas o zonas en las que se negocio puede tener una mayor probabilidad de éxito en el largo plazo.

Al tratarse de una solución tecnológica, es fundamental mantenerse al tanto de todas aquellas innovaciones y mejoras que nos permitan actualizar nuestra herramienta. De esta forma, al incorporar nuevas funcionalidades, podremos adaptarnos a las exigencias del sector y satisfacer las necesidades de nuestros clientes.

- ❖ **Chief Financial Officer (CFO):** Nuria Turégano Schirmer será responsable de la planificación económica de la empresa, así como la financiera. Llevará la proyección, el orden y el control de las inversiones, la financiación y los riesgos con el propósito de maximizar las ganancias de la empresa, evitando así las pérdidas. Creará indicadores para monitorizar y evaluar el rendimiento financiero y establecerá estrategias económicas que mejoren el aprovechamiento de los recursos.

Nuestra acción dentro del marco financiero tendrá como objetivo aumentar el margen neto en torno al 20% en los 5 primeros años, para ello se generará mensualmente el informe P&L (Profit and Loss) sobre el que se tomarán decisiones ejecutivas para no desviarnos de objetivo.

- ❖ **Chief Operations Officer (COO):** Matilde Mondéjar Montagud por su capacidad de análisis global se encargará de dar apoyo en labores estratégicas. Así mismo, será la responsable de que se ejecuten los procesos de forma óptima. Para ello, deberá aprovechar los recursos económicos, de instalaciones y operaciones con el fin de obtener un mayor rendimiento.
- ❖ **Chief Marketing Officer (CMO):** Elena Gutiérrez Alonso es la persona que se encargará de desarrollar y supervisar las acciones de marketing y publicidad de la empresa. Su misión principal es facilitar el crecimiento e incrementar las ventas a través del desarrollo de un plan de marketing integral que promueva el reconocimiento de marca y ayude a la empresa a obtener una ventaja competitiva.

Uno de los nichos de mercado clave para Geopymes es el grupo formado por las instituciones y organismos de las AAPP, así como por otras grandes entidades y corporaciones, ya que nos proporcionará alrededor del 80% de los ingresos estimados. Por ello, es vital diseñar una estrategia comercial eficiente y enfocada a este grupo de

clientes, consistente en contactar de forma directa y personalizada con cada uno de ellos, para explicarles las funcionalidades y beneficios de nuestra herramienta.

Para poder contactarlos y concertar con ellos una reunión, se creará una base de datos de clientes potenciales, con todos aquellos datos que podamos recabar, para conocer de antemano cuales son sus necesidades de información y cómo puede nuestra herramienta ayudar a solventarlas.

Por otro lado, para hacer llegar nuestra propuesta al segmento de clientes individuales, es necesario darnos a conocer a través de todos aquellos medios en los que podamos potenciar nuestra imagen de marca, ya sea a través de una página web corporativa (en la que se detallarán las características de nuestro servicio), como a través de publicaciones en RRSS, blogs especializados, newsletters, o participando en eventos profesionales, ferias y presentaciones del sector.

En relación a las cargas de trabajo y costes de participación, se fijan tareas por áreas acordes a las responsabilidades asignadas y a los objetivos fijados descritos en los puntos anteriores de este documento. Por ejemplo, el objetivo de “consolidación de la cartera de clientes (captación y fidelización)” lo asumirá el responsable de Marketing y velará porque sus acciones sirvan para cumplir los objetivos de porcentajes de clicks en web, likes en RRSS, ROI, porcentaje de retención, churn, etc.

Dado que desconocemos el volumen de negocio que vamos a tener, es muy difícil estimar las horas que requerirá cada tarea.

Se acuerda establecer un registro de horas, justificando las tareas desarrolladas, de modo que puedan remunerarse al precio/hora fijado previamente por los integrantes del equipo.

Al ser un equipo multidisciplinar, se balancearán los recursos según las necesidades que vayan surgiendo, recibiendo el responsable en cuestión el apoyo y soporte necesario del resto de miembros del equipo de modo que se garantice el cumplimiento en los plazos requeridos.

4.3.3.2. Recursos físicos

1. Infraestructura física y elementos tecnológicos

Para el desarrollo de los proyectos de los clientes, necesitamos establecer relaciones con proveedores de diferentes servicios:

- Servicios de gestión de BBDD y almacenamiento en la nube
- Empresas que aporten las licencias de software necesarias (Tableau, Windows,...)
- Suministros y servicios externos (electricidad, telefonía, Internet, etc.)

Según se vaya avanzando con el proyecto, se intentará renegociar los contratos con los proveedores, para obtener algún descuento o mejora, tanto por volumen de cliente como por fidelización, con el fin de mejorar el margen neto de beneficio.

2. Suministros y servicios externos

Todas aquellas cuestiones administrativas relacionadas con la contabilidad, fiscalidad, facturación, gestiones laborales, ... serán asumidas por una asesoría, quien se encargará de mantenernos al corriente periódicamente de la gestión del negocio.

4.3.3.3. Recursos financieros

Para poder poner en marcha nuestra idea de negocio, necesitamos contar con una serie de recursos financieros que cubran los costes iniciales y que permitan la solvencia de la empresa hasta el momento en que contemos con una cartera sólida de clientes que nos asegure un flujo constante de ingresos.

Nuestra estructura financiera estará compuesta tanto por fondos propios como ajenos. En la fase inicial del proyecto, cada miembro del equipo fundador de Geopymes aportará la cantidad de 20.000 € por persona, para formar el capital social de la nueva compañía, que tendrá un total de 80.000 €.

Los costes se han detallado en el anterior epígrafe y comprenden gastos de personal y gastos de servicios externos (acciones comerciales, licencias, hardware, ...)

4.3.4 Gestión del tiempo (cronograma)

A continuación se muestra el diagrama con el tiempo de dedicación a las diferentes tareas que componen los procesos del proyecto:

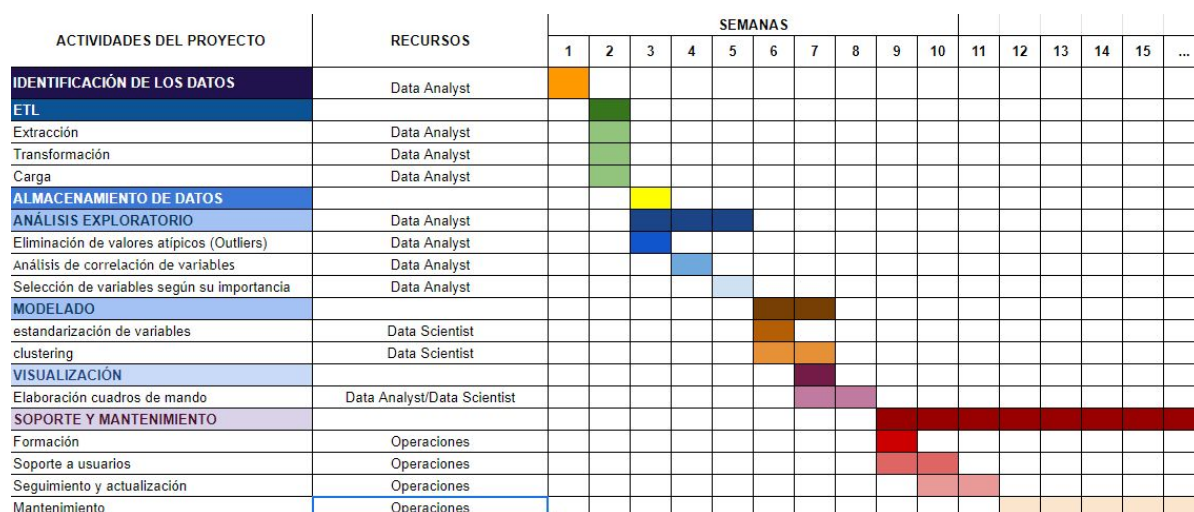


Figura 49: Diagrama de Gantt Geopymes. Fuente: Elaboración propia.

5. Rentabilidad

El beneficio que se espera obtener con el desarrollo de Geopymes tiene doble naturaleza:

- Tangible: generación de ingresos que se reflejen en beneficios para la startup.
- Intangible: poner a disposición de las pymes el valor de la información de modo que puedan tomar mejores decisiones respecto a sus negocios y aumenten las probabilidades de éxito. Esto derivará en un mayor desarrollo económico y bienestar social.

Con el objetivo de valorar la rentabilidad del proyecto analizaremos tres aspectos: la rentabilidad, la liquidez y el riesgo.

La rentabilidad se evaluará a través de los siguientes parámetros: VAN, TIR y TIRM.

Se tratan de dos fórmulas financieras empleadas habitualmente para analizar la viabilidad de un proyecto.

El valor actual neto (**VAN**) consiste en actualizar los pagos y cobros de una inversión o proyecto con el fin de conocer cuánto se va ganar o perder con dicha inversión. Si tras medir los flujos de los futuros ingresos y gastos descontados, es decir, calculando cuánto vale hoy el dinero que recibiremos en una fecha futura, da un valor positivo indicará que el proyecto generará beneficios y por tanto será rentable.

La tasa interna de retorno (**TIR**) es la tasa de interés o rentabilidad que ofrece un proyecto. Suele definirse también como el valor de la tasa de descuento que hace que el VAN sea igual a cero.

La tasa interna de retorno modificada (**TIRM**) indica la rentabilidad de un proyecto teniendo en cuenta que los flujos que genera dicho proyecto rinden o cuestan una tasa distinta de la interna del propio proyecto. La TIRM corrige el principal inconveniente de inconsistencia de la TIR, ya que considera que los flujos liberados por el proyecto se reinvierten a una tasa equivalente a la tasa de actualización impuesta (coste de oportunidad).

Cuando analizamos la liquidez de un proyecto hablamos del período de recuperación de la inversión o **payback**, que es la cantidad de tiempo que tarda una empresa en recuperar la inversión inicial. Este es un aspecto clave para decidir si embarcarse o no en un proyecto.

Para la valoración del riesgo, hemos considerado aumentar la tasa de descuentos al 20%, respecto al 15% habitual para proyectos de innovación, como consecuencia de la incertidumbre e impacto de la crisis derivada de la actual pandemia.

INVERSIÓN

Aportación de socios

Cada uno de los 4 socios realizará una contribución económica inicial igualitaria y que será en un principio la única inversión con la que contará la startup. Esta cantidad se ha calculado estimando que será suficiente para sufragar los costes iniciales necesarios para la estabilización de la empresa.

5.1. Ingresos

La tasa anual de emprendedores en España es del 6,4%, estaríamos hablando de unos 3 millones de emprendedores que si hacemos un reparto equitativo por población, a nivel de Madrid serían 422.400. Como ya vimos en las entrevistas realizadas, el 36% se informó sobre una posible ubicación de su negocio, por lo que tendríamos un público objetivo de unos 150.000 clientes particulares al año. Estimamos una penetración del negocio en el segmento de particulares del 1%, pero debido a la actual crisis hemos considerado bajar esta cifra a 0,14% siendo incremental cada año de actividad.

Para el segmento de Grandes Cuentas, contemplamos como público objetivo al Ayuntamiento de Madrid, Cámara de comercio, Instituto de la mujer, entidades bancarias, inmobiliarias...

Para la suscripción premium no podremos empezar a dar servicio hasta el 4º cuarto año de actividad, ya que hasta que la empresa no obtenga beneficios no podremos asumir el coste de invertir en bases de datos de terceros e implementar nuevas funcionalidades.

5.2. Gastos

Gastos de establecimiento

Para fundar nuestra startup deberemos hacer frente a una serie de gastos iniciales. Dado que se constituirá como Sociedad Limitada, el capital social mínimo exigido son 3.000 euros, a lo que habría que sumar gastos de notaría y de registro de la actividad.

Gastos de personal

Se acuerda asignar a cada miembro del equipo un salario bruto anual (incluye costes de Seguridad Social) revisable en función de la evolución de la actividad, barajando también la posibilidad de repartir dividendos con el objetivo de no incrementar la masa salarial. No se estima por el momento la contratación de personal adicional.

Gastos de Gestoría

La empresa contratará un servicio externo de gestoría que llevará a cabo todas las funciones contables, fiscales y laborales necesarias para desarrollar la actividad de la empresa.

Costes acciones comerciales y publicidad

Optaremos por contratar un servicio de Marketing Digital diseñado para empresas como el que ofrece por ejemplo la empresa Serseo para negocios con expansión nacional.



Recomendada para empresas de ámbito nacional, interesadas en generar más oportunidades de venta, incrementar su visibilidad general y los resultados de su negocio crecerán de forma controlada.

CLIENTES NACIONAL

Desde 325€ /mes

Figura 50: Ejemplo presupuesto marketing digital. Fuente: Serseo.

Coste de almacenamiento en la nube

Necesitaremos un servicio Google Cloud SQL que se encargará de configurar los servidores, aplicar parches y actualizaciones, configurar replicaciones, gestión de backups, etc. permitiendo que nuestra base de datos sea segura, de alta disponibilidad y flexible. Todo ello nos hará ser más productivos despreocupándonos de todas estas funciones y dedicando más tiempo al desarrollo de nuestro producto.

Licencias

Necesitaremos 1 licencia de Tableau Creator (incluye Desktop) y 100 licencias Tableau Viewer que podremos contratar en la página web de Tableau con el siguiente coste mensual:

Tableau Creator

Cada implementación requiere al menos una licencia de Creator

USD 70

por usuario por mes | facturado anualmente

Descubra información con un eficaz conjunto de productos que brinda soporte a todo el flujo de análisis.

Incluye: Tableau Desktop, Tableau Prep Builder y una licencia Creator para Tableau Server.

Tableau Explorer

USD 35

por usuario por mes | facturado anualmente
5 usuarios Explorer como mínimo

Explore datos confiables y responda sus propias preguntas más rápido con un análisis de autoservicio completo.

Incluye: una licencia Explorer para Tableau Server.

Tableau Viewer

USD 12

por usuario por mes | facturado anualmente
100 usuarios Viewer como mínimo

Vea dashboards y visualizaciones e interactúe con ellos a través de una plataforma segura y fácil de usar.

Incluye: una licencia Viewer para Tableau Server.

Figura 51: Tipos de licencias. Fuente: Tableau.

Hardware

Se adquirirá un servidor de configuración básica, con un procesador de cuatro núcleos 4GB de memoria RAM, y dos discos duros de 500GB así como 4 portátiles para desempeñar las funciones laborales requeridas.

Todos los equipos entrarán en un plan de amortización a 4 años.

BBDD elnforma

Para poder prestar las características del servicio Premium es necesario contratar una base de datos de empresas, en la que se detallen, además de los datos de contacto, los datos económicos y financieros y la clasificación de la actividad en el CNAE.



Acceda a su cuenta

¿Alguna duda? **900 10 30 20**

Contacte gratis con nosotros de Lunes a Viernes de 8.30h a 18.30h

Configure su listado de Empresas

Datos del listado
Filtrado de empresas

¿Qué datos de las empresas necesita?

| CONTACTO | Obligatorio ayuda | Coste |
|--|-------------------------------------|--------|
| <input checked="" type="checkbox"/> CIF | <input checked="" type="checkbox"/> | 0,05 € |
| <input checked="" type="checkbox"/> Nombre de la Empresa | <input checked="" type="checkbox"/> | 0,04 € |
| <input checked="" type="checkbox"/> Dirección completa | <input checked="" type="checkbox"/> | 0,04 € |
| <input checked="" type="checkbox"/> Teléfono | <input checked="" type="checkbox"/> | 0,04 € |
| <input checked="" type="checkbox"/> Página Web | <input checked="" type="checkbox"/> | 0,04 € |
| <input checked="" type="checkbox"/> Persona de contacto | <input checked="" type="checkbox"/> | |

Mi listado

Registros totales 192.430

Coste total 20.298,59 €

Datos del listado

Paquete Personalizado

- ☒ Actividad CNAE 192.386 x
- ☒ Dirección Completa 192.430 x
- ☒ Ventas 88.461 x
- ☒ Resultado del Ejercicio 100.572 x

Empresas

☒ MADRID x

Figura 52: Ejemplo base de datos de terceros. Fuente: elnforma

Gastos de conexión y comunicaciones

Dado que no consideramos tener una ubicación física del negocio y que cada socio dispone de conexión e internet en su domicilio, correrá a cargo de cada persona los gastos asociados a tal servicio.

Impuestos

El tipo impositivo aplicable a una Sociedad Limitada es del 25% sobre el beneficio neto.

5.3. Resultados Financieros

A continuación se muestra el detalle de cada partida con los resultados financieros:

| ANÁLISIS RENTABILIDAD DEL PROYECTO | | | | | |
|------------------------------------|----------------------|----------------------|----------------------|---------------------|---------------------|
| Aportación de los socios | 80.000 € | | | | |
| | AÑO 1 | AÑO 2 | AÑO 3 | AÑO 4 | AÑO 5 |
| INVERSIÓN | | | | | |
| Hardware | 4.515 € | 0 € | 0 € | 0 € | 0 € |
| BBDD Empresas | 0 € | 0 € | 0 € | 30.000 € | 0 € |
| Software - licencias | 15.240 € | 15.240 € | 15.240 € | 15.240 € | 15.240 € |
| Software - almacenamiento | 750 € | 750 € | 750 € | 750 € | 750 € |
| TOTAL INVERSIONES | 20.505,00 € | 15.990,00 € | 15.990,00 € | 45.990,00 € | 15.990,00 € |
| INGRESOS | | | | | |
| Precios por tipo cliente | | | | | |
| Particulares | 150 € | 150 € | 150 € | 150 € | 150 € |
| Grandes cuentas | 10.000 € | 10.000 € | 10.000 € | 10.000 € | 10.000 € |
| Suscriptores Premium | 0 € | 0 € | 0 € | 20.000 € | 20.000 € |
| Número de clientes | | | | | |
| Particulares | 200 | 300 | 400 | 600 | 900 |
| Grandes cuentas | 2 | 6 | 9 | 12 | 16 |
| Suscriptores Premium | 0 | 0 | 0 | 4 | 7 |
| Ingresos | | | | | |
| Particulares | 30.000 € | 45.000 € | 60.000 € | 90.000 € | 135.000 € |
| Grandes cuentas | 20.000 € | 60.000 € | 90.000 € | 120.000 € | 160.000 € |
| Suscriptores Premium | 0 € | 0 € | 0 € | 80.000 € | 140.000 € |
| TOTAL INGRESOS | 50.000,00 € | 105.000,00 € | 150.000,00 € | 290.000,00 € | 435.000,00 € |
| GASTOS | | | | | |
| Gastos de establecimiento | 3.500 € | 0 € | 0 € | 0 € | 0 € |
| Gastos de personal | 80.000 € | 80.000 € | 80.000 € | 80.000 € | 80.000 € |
| Gestoría | 1.800 € | 1.800 € | 1.800 € | 1.800 € | 1.800 € |
| Costes comerciales y publicidad | 4.500 € | 3.600 € | 3.600 € | 3.600 € | 3.600 € |
| Impuestos | 0,00 € | 902,50 € | 12.152,50 € | 39.652,50 € | 83.402,50 € |
| TOTAL GASTOS | 89.800,00 € | 86.302,50 € | 97.552,50 € | 125.052,50 € | 168.802,50 € |
| FLUJO DE CAJA OPERATIVO | - 60.305,00 € | 2.707,50 € | 36.457,50 € | 118.957,50 € | 250.207,50 € |
| VALOR ACTUAL | - 60.305,00 € | 2.461,36 € | 30.130,17 € | 89.374,53 € | 170.895,09 € |
| ACUMULADO | - 60.305,00 € | - 57.843,64 € | - 27.713,47 € | 61.661,06 € | 232.556,15 € |
| VAN: | 152.556,15 € | | | | |
| TIR: | 36,53% | | | | |
| TIRM: | 33,83% | | | | |

Figura 53: Cuadro Rentabilidad. Fuente: Elaboración propia.

Una vez analizados estos indicadores financieros podemos confirmar la viabilidad financiera de nuestro proyecto con un VAN positivo:

| | | | |
|---------------------------|----------------------|-----------------------|--------------------------------|
| VAN 152.556,15€ | TIR 36,53% | TIRM 33,83% | PAY-BACK 4º y 5º año |
|---------------------------|----------------------|-----------------------|--------------------------------|

Bibliografía

Publicaciones y estudios sobre la pyme en la España

<http://www.ipyme.org/>

Estadísticas relativas a las pymes en España actualizado a Mayo 2020

<https://industria.gob.es/>

Condiciones debe cumplir una empresa para ser una pyme

<http://www.creatuempresa.org/>

Distribución de empresas nacionales por provincia y condición jurídica

<https://www.ine.es/>

Informe GEm España 19/20

https://www.gem-spain.com/wp-content/uploads/2020/06/Informe-GEM-Espa%C3%B1a-2019_20.pdf

Datos de emprendedores en España, características y perfiles

<https://www.emprendedores.es/>

Principales problemas tecnológicos en las pymes españolas

https://cincodias.elpais.com/cincodias/2016/04/13/pyme/1460561503_411481.html

¿Por qué les cuesta tanto crecer a las Pymes?

<https://ideasparatuempresa.vodafone.es/por-que-les-cuesta-tanto-crecer-a-las-pymes/>

Los errores más comunes en las pymes

<https://revistapymes.es/los-errores-mas-comunes-en-las-pymes/>

La localización del negocio

<https://www.emprendepyme.net/crear-empresa/la-localizacion-del-negocio>

Factores de localización de una empresa

<https://www.emprendepyme.net/factores-de-localizacion-de-una-empresa.html>

Factores de localización de una empresa

<http://www.camarafp.org/portal/index.php/estudiantes/ubicacion-de-la-empresa.html>

Normativa relacionada con la PYME

<http://www.ipyme.org/es-ES/Normativa/Paginas/Normativa.aspx>

Factores determinantes en la creación de una empresa: valores culturales, redes sociales y ayudas públicas

https://www.researchgate.net/publication/28143602_Factores_determinantes_en_la_creacion_de_una_empresa_valores_culturales_redes_sociales_y_ayudas_publicas

Las empresas familiares, motor de la economía española

https://www.elconfidencial.com/empresas/2018-10-04/empresas-familiares-economia-espanola-bra_1624960/

Análisis estratégico para el desarrollo de la PYME en España: Internacionalización y orientación emprendedora

<https://www.cepyme.es/wp-content/uploads/2017/04/Informe-Pyme-Espa%C3%B1a-2016-Faedpyme-1.pdf>

Tendencias tecnológicas en las pymes para 2020-2021

<https://www.capterra.es/blog/1268/tendencias-tecnologicas-en-las-pymes-espanolas-2020-seguridad-de-datos-es-el-principal-foco-de-interes>

Transformación digital en las pymes y los autónomos españoles

<https://www.ithotelero.com/wp-content/uploads/2019/12/Informe-eAPyme-2019.pdf>

La economía española es una de las más dependientes de las pymes

<https://cepymenews.es/la-economia-espanola-una-las-mas-dependientes-las-pymes-toda-europa/>

Las pymes crean el 66% del empleo en España

<https://www.eleconomista.es/gestion-empresarial/noticias/9646458/01/19/Las-Pymes-espanolas-crean-el-66-del-empleo.html>

Indicadores | BBVA Research

<https://www.bbvaresearch.com/indicadores/>

La mitad de los emprendedores autónomos fracasa.

<http://theconversation.com/la-mitad-de-los-emprendedores-autonomos-fracasa-119332>

Emprendedores: una correcta ubicación es sinónimo de éxito

<https://www.eleconomista.es/gestion-franquicias/noticias/6470785/02/15/Emprendedores-una-correcta-ubicacion-es-sinonimo-de-exito.html>

La ubicación del negocio, factor clave de éxito

<https://www.juguetesb2b.com/tendencias/20190405/ubicacion-negocio-factor-clave-exito.aspx>

¿Quieres triunfar? Busca una buena localización para tu negocio

<https://www.pymesyautonomos.com/marketing-y-comercial/quieres-triunfar-busca-una-buena-localizacion-para-tu-negocio>

Estadísticas PYME - Evolución e Indicadores

<http://www.ipyme.org/Publicaciones/Estadisticas-PYME-2017.pdf>

Correlación entre variables

http://viref.udea.edu.co/contenido/menu_alterno/apuntes/ac36-correlacion-variables.pdf

¿Qué es el clustering en R?

<https://jarroba.com/que-es-el-clustering/#:~:text=El%20Clustering%20es%20una%20tarea,los%20objetos%20de%20otros%20Clusters.>

Descripción del paquete MICE

<https://datascienceplus.com/imputing-missing-data-with-r-mice-package/>

¿Qué es Boruta?

<http://www.revistaindice.com/numero68/p22.pdf>

Método Random Forest, características

<https://analisisydecision.es/medir-la-importancia-de-las-variables-con-random-forest/>

Árboles de decisión con H2O

<https://translate.google.com/translate?hl=es&sl=en&u=https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drf.html&prev=search&pto=aue>

Clustering y heatmaps: aprendizaje no supervisado

https://www.cienciadedatos.net/documentos/37_clustering_y_heatmaps

Rentabilidad del proyecto: VAN y TIR

<https://obsbusiness.school/es/blog-investigacion/finanzas/van-y-tir-dos-herramientas-para-la-viabilidad-de-una-inversion#:~:text=El%20VAN%20y%20el%20TIR%20son%20dos%20f%C3%B3rmulas%20financieras%20empleadas,la%20Tasa%20Interna%20de%20Retorno.>

Definición de Valor actual neto

<https://economipedia.com/definiciones/valor-actual-neto.html>

Gastos de establecimiento

<https://www.europapress.es/economia/noticia-cuanto-te-cuesta-fundar-startup-20150616135430.html>

Marketing online

https://serseo.es/tabla-tarifas-servicios-inbound-marketing-pymes/?utm_campaign=&utm_term=&utm_source=adwords&utm_medium=ppc&ctf_src=g&ctf_net=adwords&ctf_mt=b&ctf_grp=110554496558&ctf_ver=1&ctf_cam=10738841730&ctf_kw=&ctf_acc=384-790-3674&ctf_ad=466796025722&ctf_tgt=dsa-957778666282&gclid=CjwKCAjw8MD7BRAReIwAGZsrBUZFTnOY-engaqqc3PefD5i818lOsRVbn047l5cdWuR5s5sHQMNiPBoCKloQAvD_BwE

Qué es Google Cloud

<https://www.beservices.es/beneficios-administrar-mysql-postgresql-sql-en-cloudsql-n-5433-es>

Tableau

<https://www.tableau.com/es-es/about/press-releases/2017/tableau-introduces-new-subscription-pricing>

¿Cuánto cuesta un servidor?

<https://www.pymesyaautonomos.com/tecnologia/el-coste-del-servidor-en-la-pyme-el-hardware>

Anexo I: Encuestas para la validación de hipótesis

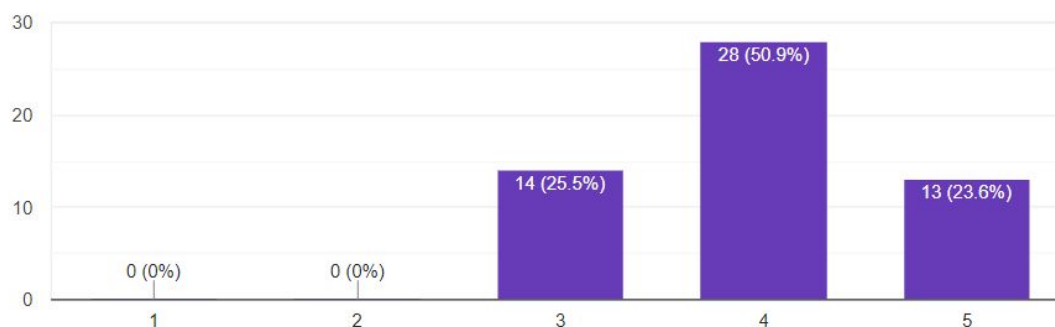
Emprendedores

De las 55 personas encuestadas hemos obtenido el 100% de las respuestas

1. ¿Qué tipo de factores piensa que afectan negativamente en el éxito de un negocio?

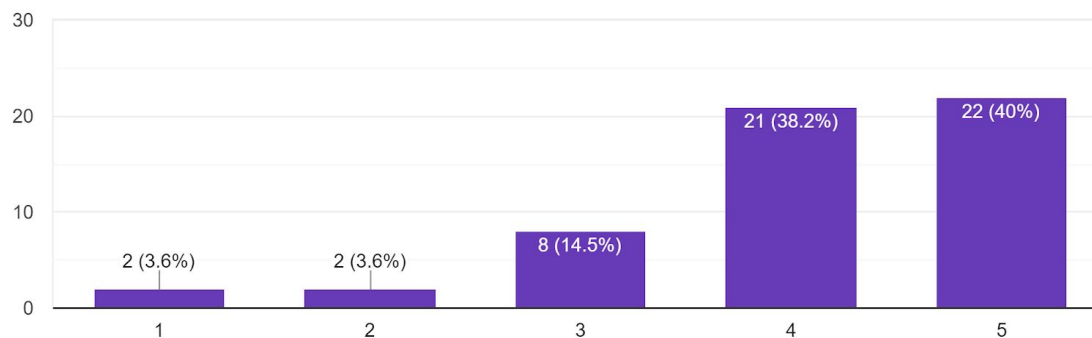
Alta incertidumbre en el mercado

55 respuestas



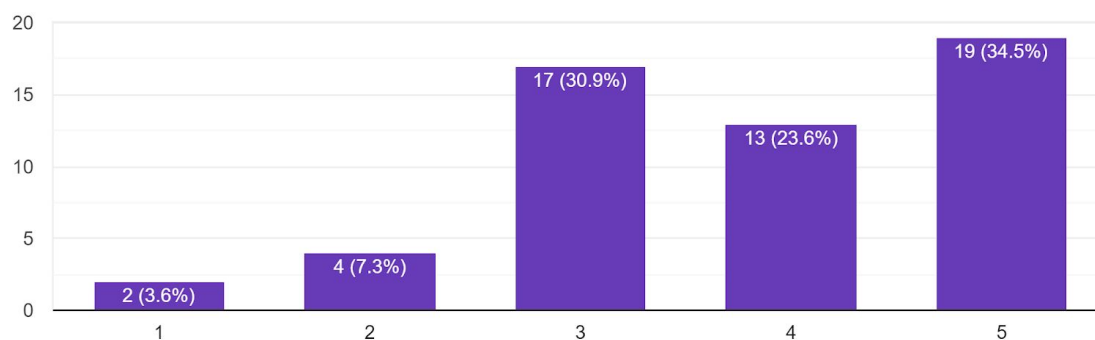
Desconocimiento del mercado potencial de clientes

55 respuestas



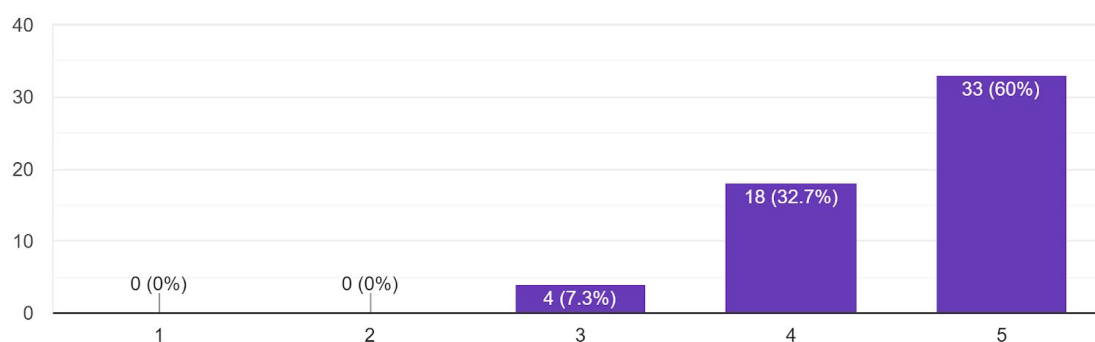
Mala ubicación

55 respuestas



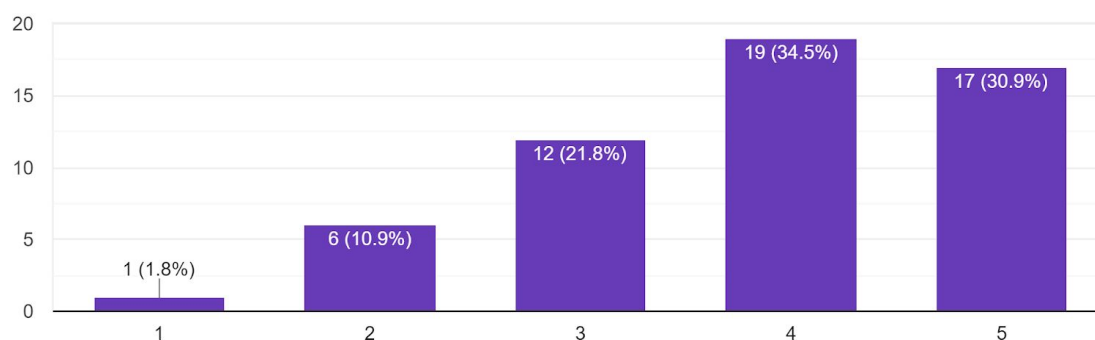
Mala administración/gestión

55 respuestas



El no uso de nuevas tecnologías o poca digitalización

55 respuestas



Otros (Especifique)

11 responses

Falta de profesionalidad

Personal desmotivado

Falta de imaginación. Falta de entrega.

Competición alta

No diversificarse

Fuentes de financiación

La capacidad logística y transporte de materias primas para la producción del producto que se quiera vender.

La seguridad de los sistemas informáticos y de los datos que contienen en caso de los negocios digitales debe estar bien pensada y acorde a la legislación si la gente no ve seguro conectarse no comprar el producto aunque sea bueno

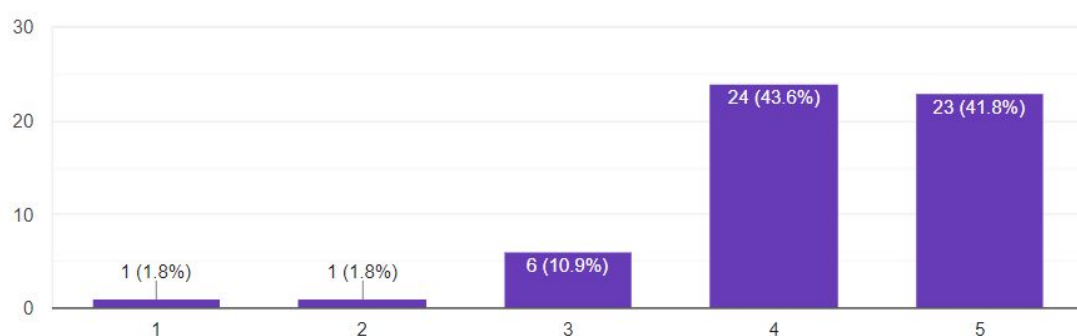
Relación de precio, cantidad y calidad

falta de innovacion o de una idea nueva, desconexion con la moda/tendencias de consumo

4. Valore aquellos factores que pueden condicionar la decisión sobre la ubicación de un negocio:

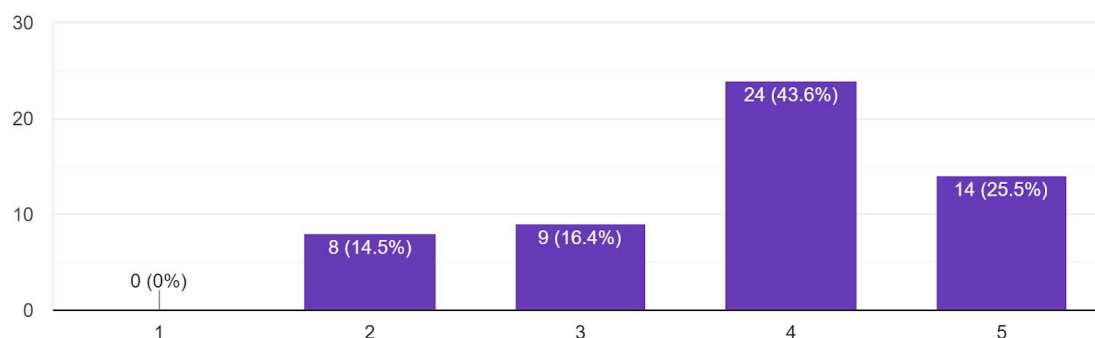
El coste del alquiler del local, así como de las tasas y licencias

55 responses



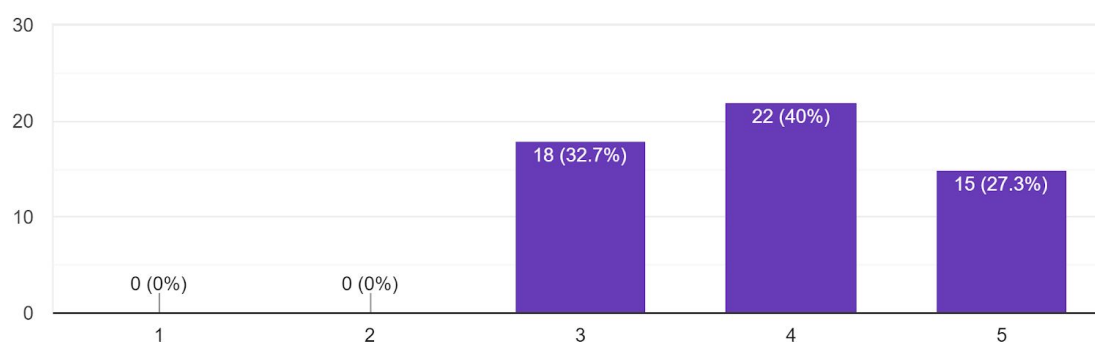
Que esté bien comunicado a través de la red de transporte público

55 respuestas



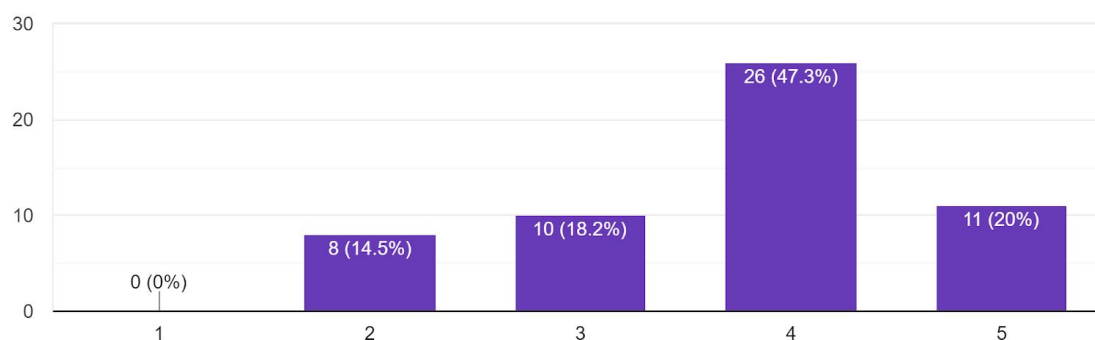
Que sea accesible por transporte privado y haya un buen servicio de estacionamiento

55 respuestas



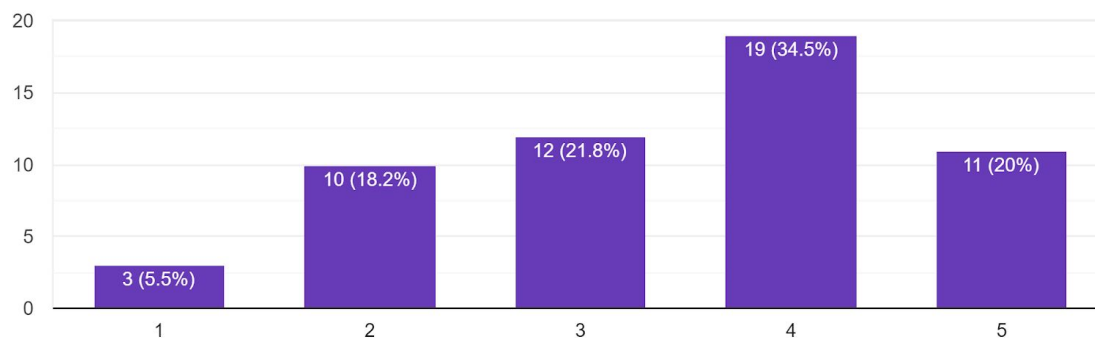
Que esté cercano a las calles o vías principales de la localidad

55 respuestas



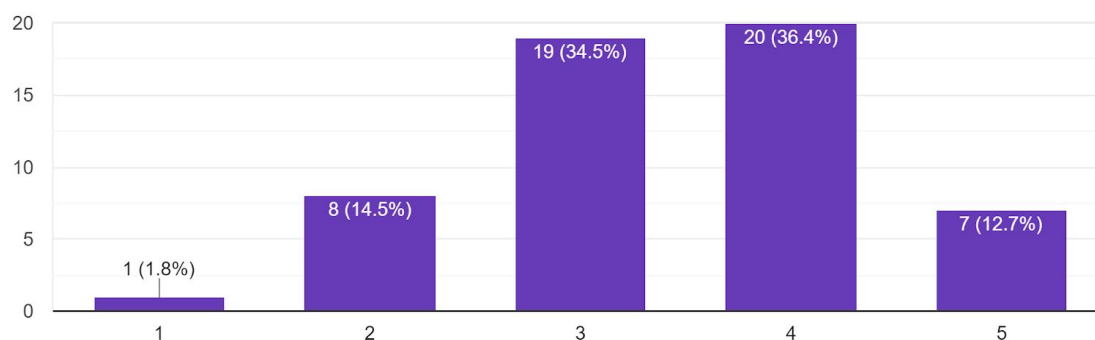
Que esté situado en una vía peatonal o en la que haya un importante número de transeúntes

55 respuestas



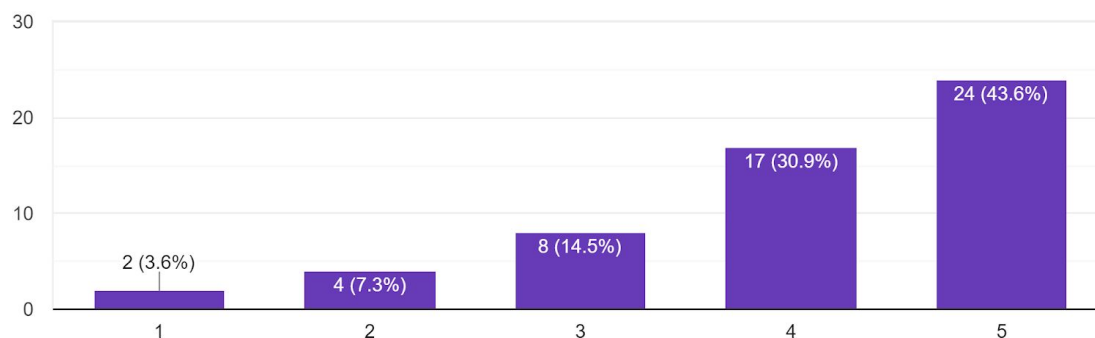
Que sea fácilmente accesible a pie y/o para aquellas personas con movilidad reducida

55 respuestas



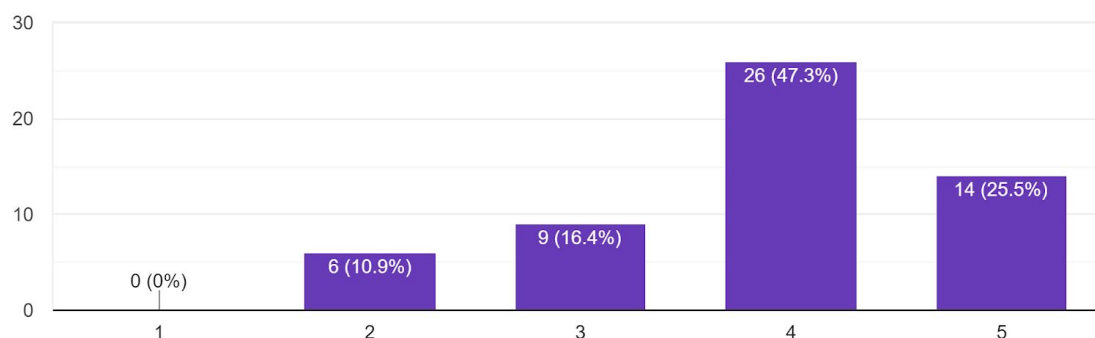
La existencia de competidores directos

55 respuestas



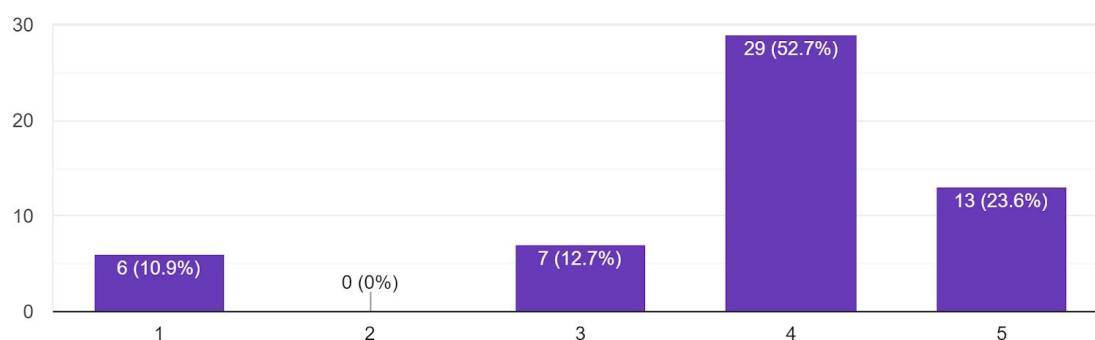
La cercanía de negocios o actividades complementarias

55 responses



Factores socio demográficos y económicos de la zona (edad de la población, nº miembros por familia, renta per cápita...)

55 responses



Otros (Especifique)

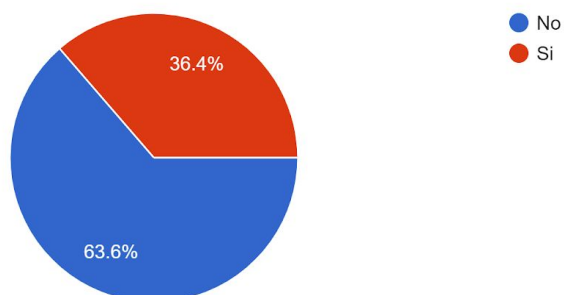
2 responses

Presencia digital/Creacion de actividades

En negocios totalmente digitales la ubicación es importante para temas fiscales y de legislación a cumplir

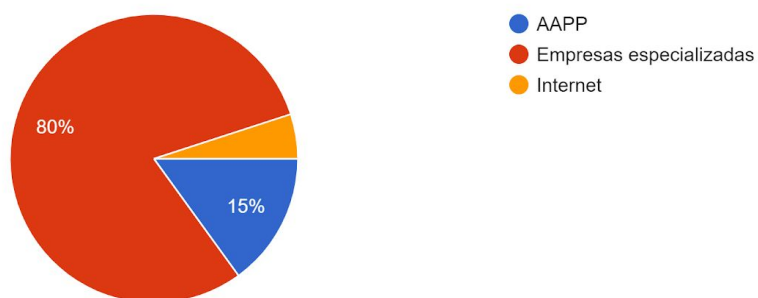
5. Antes de iniciar su negocio, ¿consultó o se asesoró para elegir la ubicación de su empresa?

55 responses



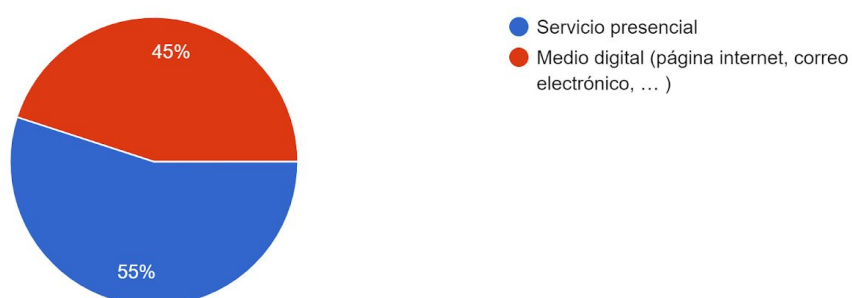
¿Quién le prestó el asesoramiento?

20 responses



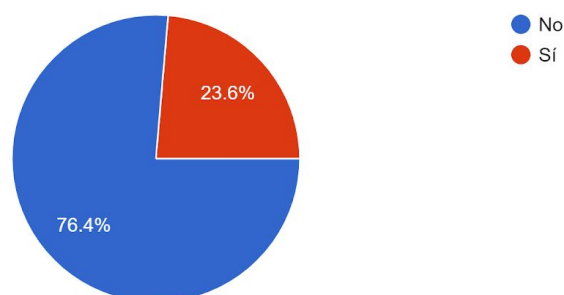
¿Cómo le prestó el asesoramiento?

20 responses



6. A la hora de solicitar la licencia de apertura o de obras por parte del ayuntamiento, ¿recibió algún tipo de asesoramiento, orientación o recomen...úblicos sobre la zona en la que abrir su negocio?

55 responses



Indique qué tipo de información le suministraron:

13 responses

Negocios de la Zona

Información integral sobre todos los aspectos relevantes para el correcto funcionamiento

Información sobre impuestos

Como registrar empresa, pagar IVA, declarar impuestos

Forma y plazo de presentación

Información catastral

toda la necesaria para iniciar la actividad del negocio

Fiscal

Tasas

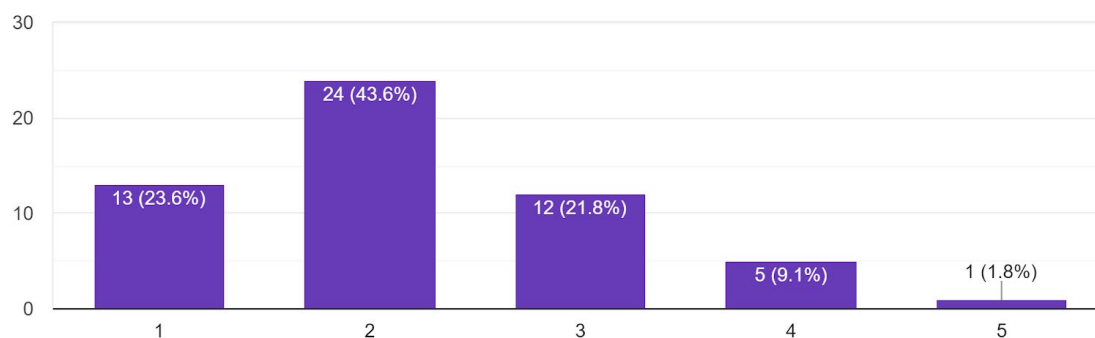
Asesoramiento legal

Jurídica

Información tasas, riesgos, etc

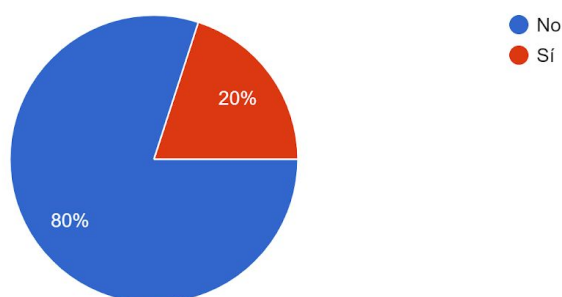
7. ¿Cree que a la hora de emprender o de iniciar un negocio los ciudadanos reciben el apoyo necesario por parte de las AAPP en cuanto a inform...ramiento sobre la ubicación y el tipo de negocio?

55 responses



8. Cuando planificó la apertura de su negocio, ¿usó algún tipo de herramienta que le proporcionará información con datos sociodemográficos, económi...icos de la zona elegida para su establecimiento?

55 responses



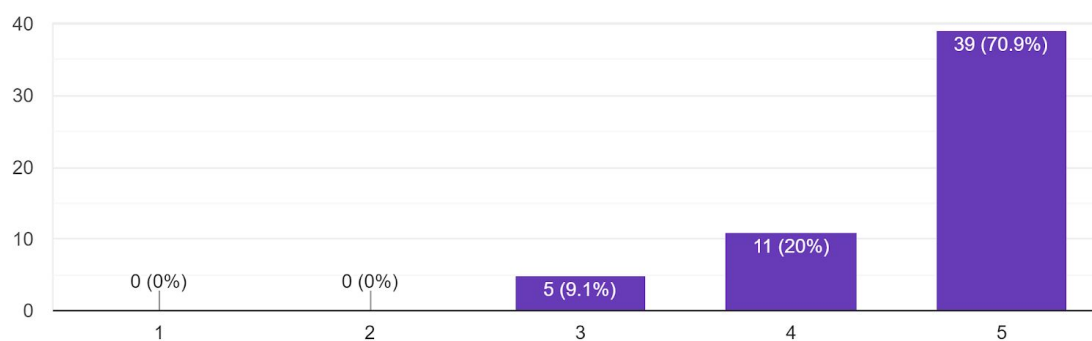
Especifique la herramienta o fuente de información:

11 responses

| |
|----------------------|
| Master Tool |
| Información privada |
| Organismos oficiales |
| google |
| Encuestas de mercado |
| Geomarketing |
| Viveros de empresas |
| ine |
| APP |
| Ayuntamiento, google |
| Google |

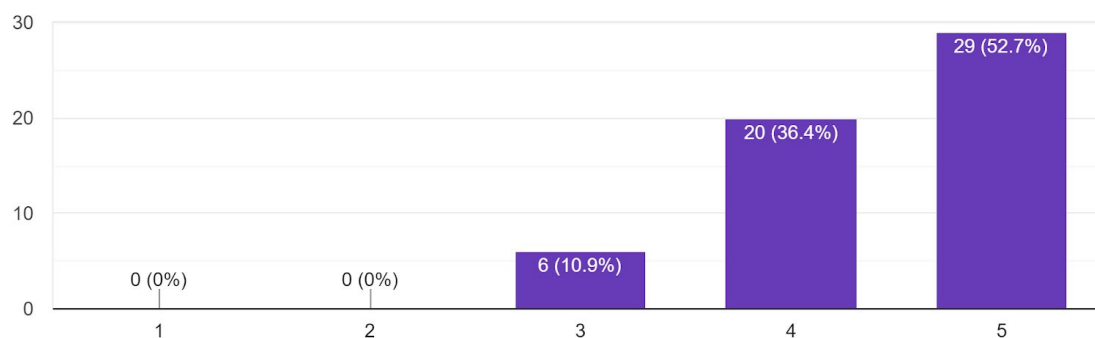
9. Si supiera de la existencia de una herramienta online que le sugiriera la mejor ubicación para abrir un negocio, ¿qué puntuación le daría?

55 responses



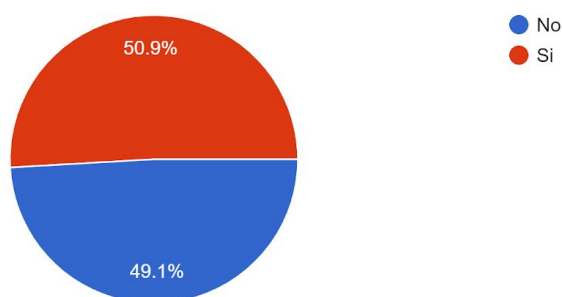
10. Asimismo, si dicha herramienta online le sugiriera el(los) tipo(s) de industria más acorde para abrir un negocio en una zona específica de la ciudad, ¿qué puntuación le daría?

55 responses



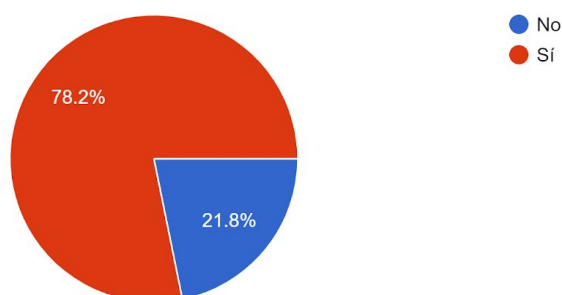
11. ¿Estaría dispuesto a cambiar el tipo de negocio que quiere emprender para asegurar un mayor éxito en una zona determinada?

55 responses



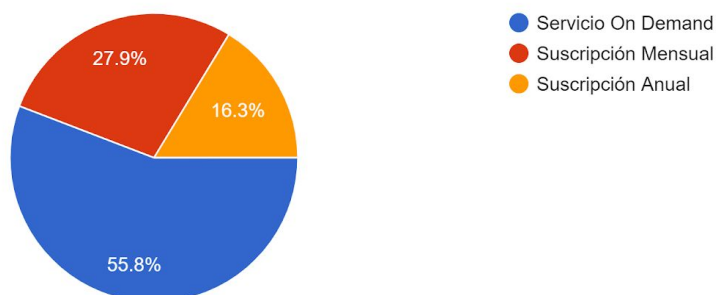
12. ¿Estaría dispuesto a pagar por un servicio que le sugiera la mejor ubicación para establecer un negocio de una industria específica? En caso que r...ad de pago que más se ajusta a su caso particular.

55 responses



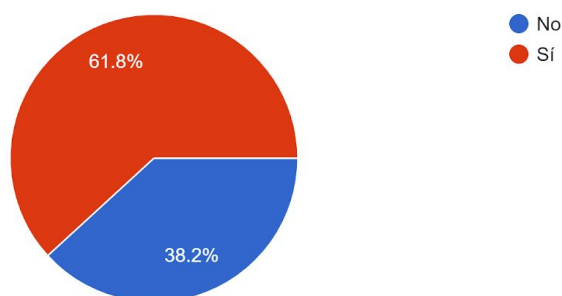
Especifique por favor la modalidad de pago que más se ajusta a su caso particular:

43 responses



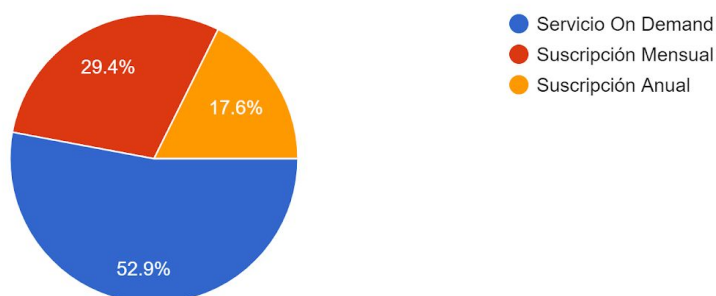
13. ¿Estaría dispuesto a pagar por un servicio que le sugiera los tipos de industria apropiados para establecer un negocio en una zona específica de la...d de pago que más se ajusta a su caso particular.

55 responses



Especifique por favor la modalidad de pago que más se ajusta a su caso particular:

34 responses



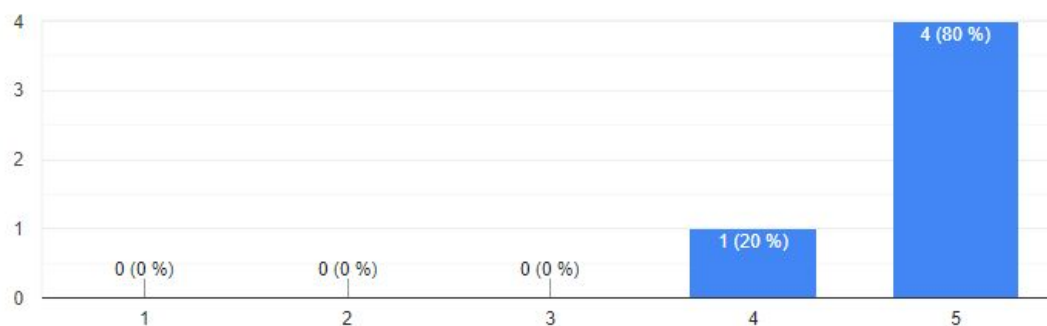
Administraciones Públicas

De las 5 administraciones públicas encuestadas hemos obtenido el 100% de las respuestas

¿Qué tipo de factores piensa que afectan negativamente en el éxito de un negocio?

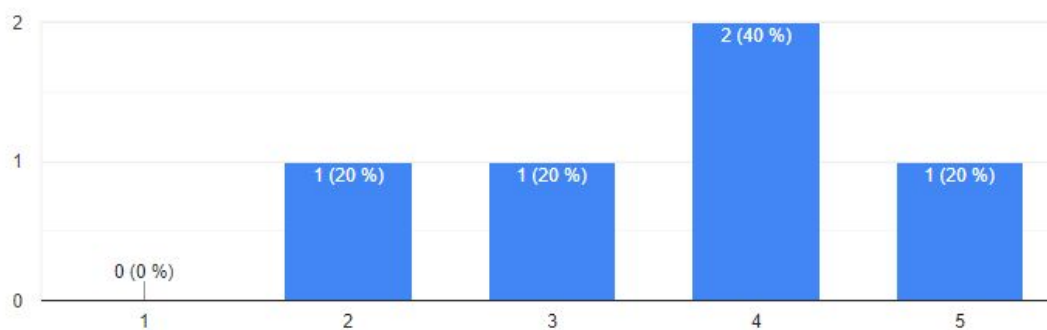
Desconocimiento del mercado potencial de clientes

5 respuestas



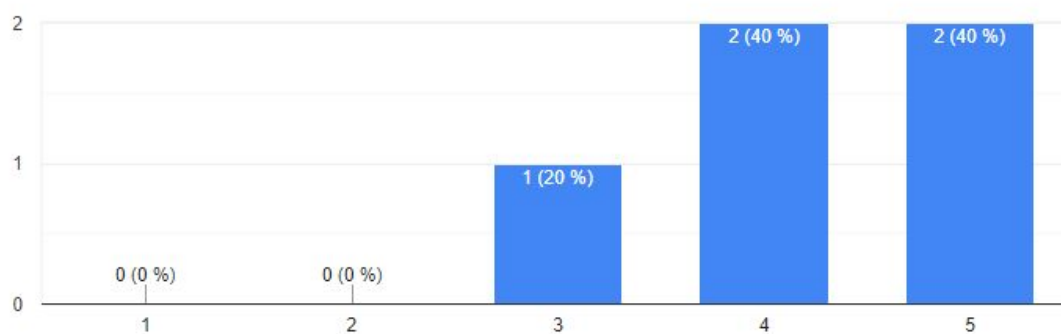
La alta incertidumbre en el mercado

5 respuestas



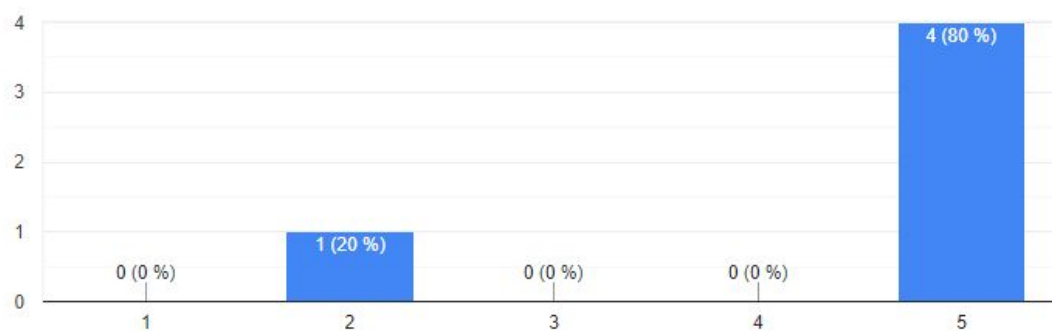
Mala ubicación

5 respuestas



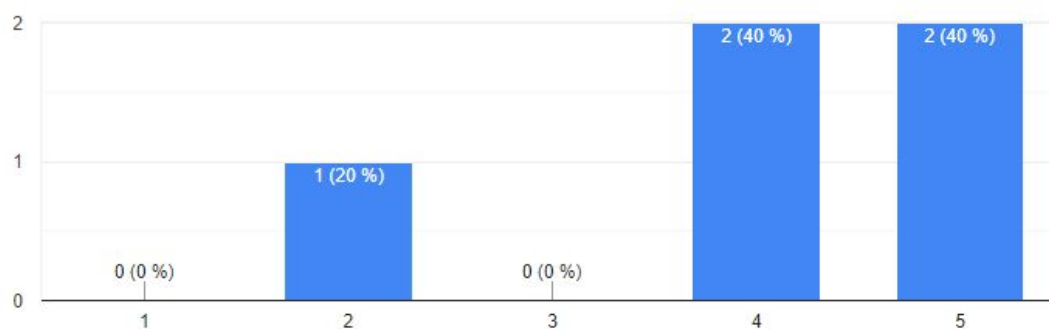
Mala administración

5 respuestas



El no uso de nuevas tecnologías y poca digitalización

5 respuestas



Otros: (Especifique)

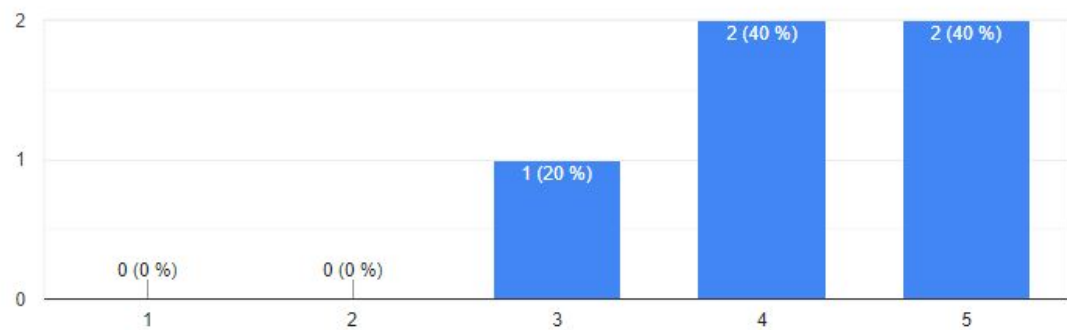
1 respuesta

Falta e formación específica o escasa experiencia en el sector

De los siguientes factores, ¿cuáles cree que pueden condicionar la decisión sobre la ubicación de un negocio?

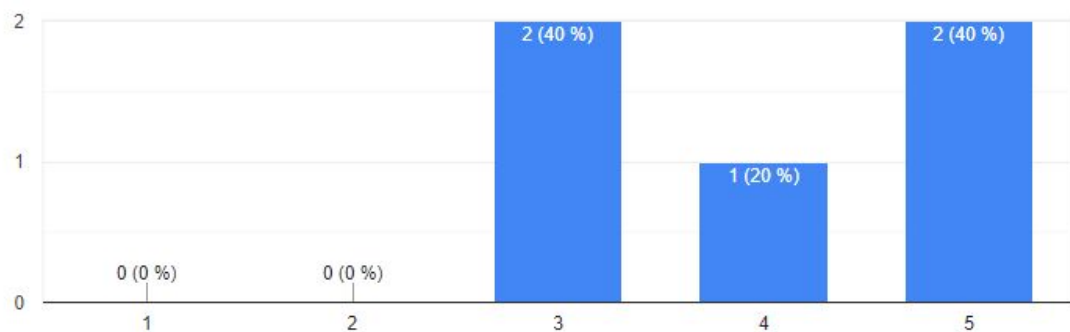
El coste del alquiler del local, así como de las tasas y licencias

5 respuestas



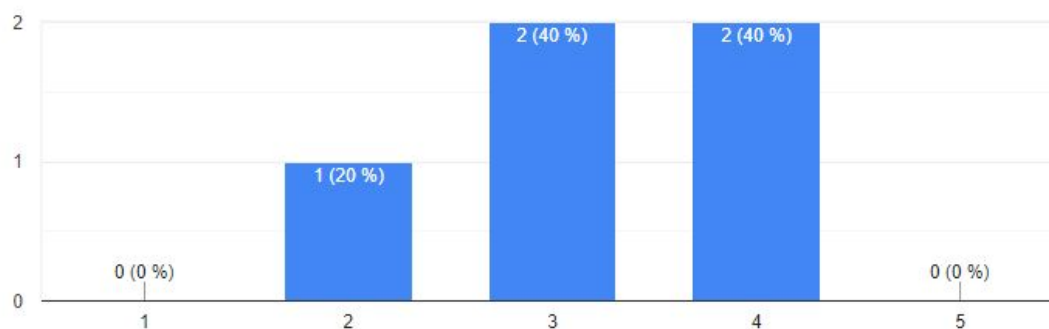
Que esté bien comunicado a través de la red de transporte público

5 respuestas



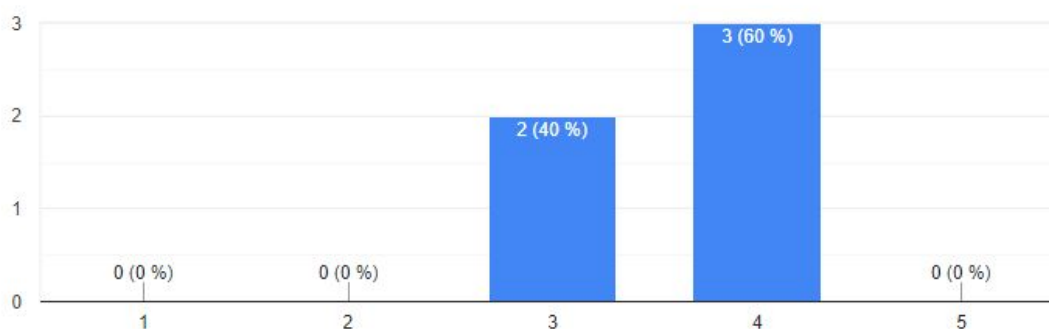
Que sea accesible por transporte privado y haya un buen servicio de estacionamiento

5 respuestas



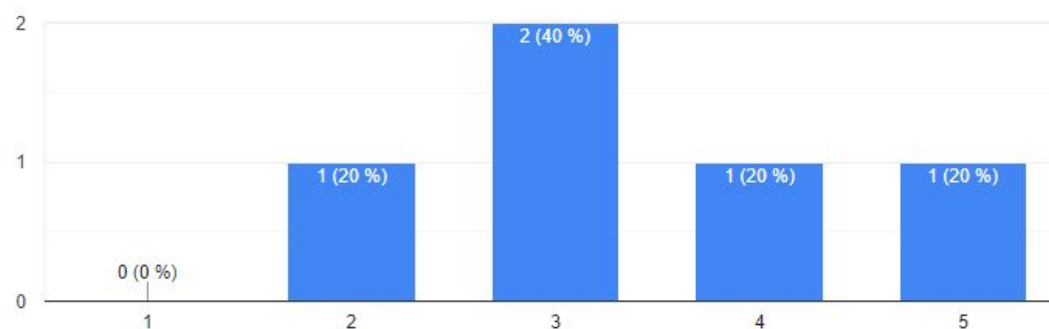
Que esté cercano a las calles o vías principales de la localidad

5 respuestas



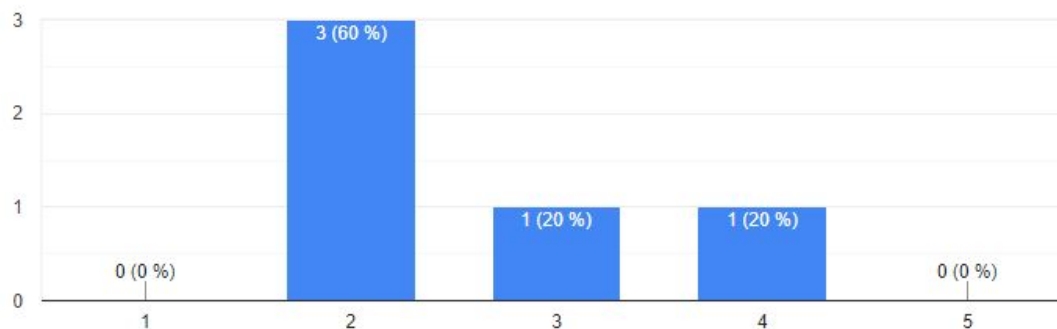
Que esté situado en una vía peatonal o en la que haya un importante número de transeúntes

5 respuestas



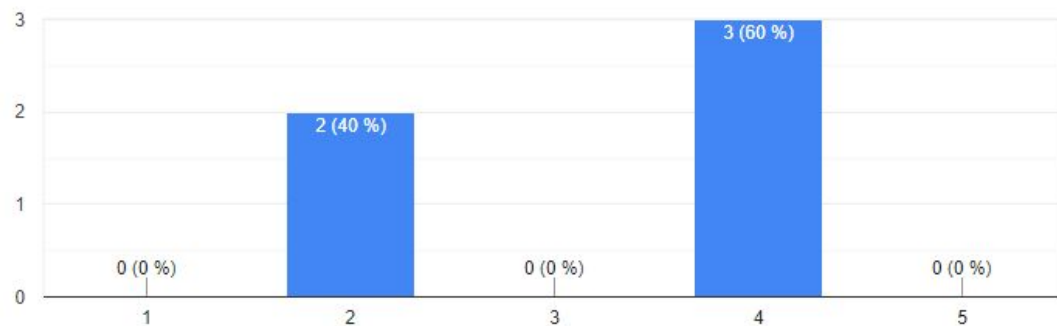
Que sea fácilmente accesible a pie y/o para aquellas personas con movilidad reducida

5 respuestas



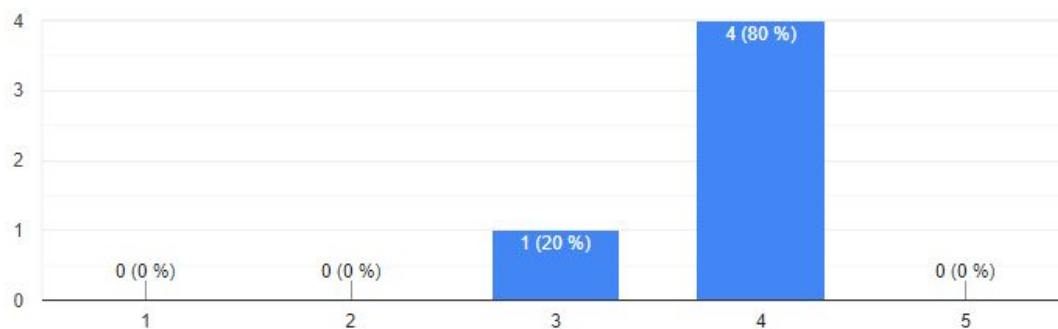
La existencia de competidores directos en la zona

5 respuestas



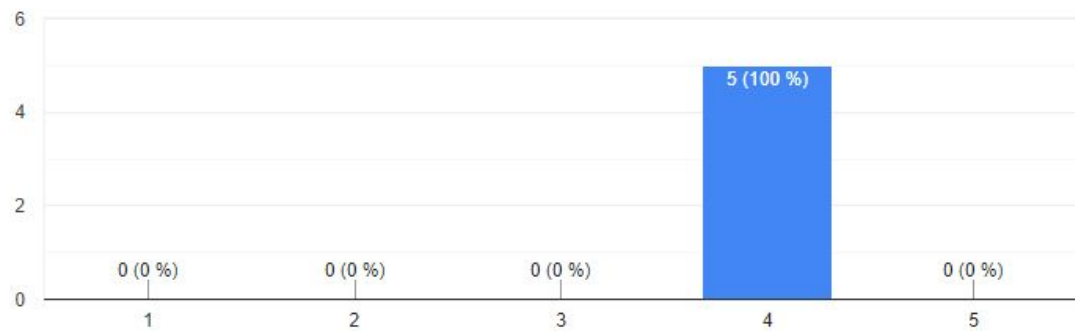
La cercanía de negocios o actividades complementarias

5 respuestas



Factores sociodemográficos y económicos de la zona (edad población, nº miembros por familia, renta per cápita...)

5 respuestas



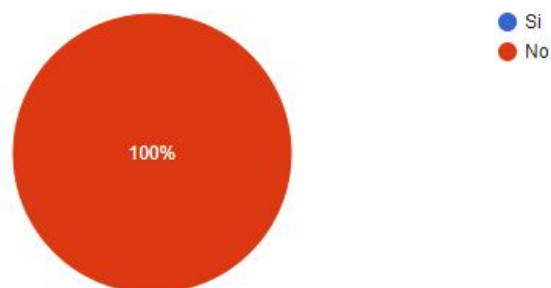
Otros: (Especifique)

0 respuestas

Aún no hay respuestas para esta pregunta.

En el momento de solicitar información sobre la licencia de apertura o de obras por parte del usuario, ¿se le ofrece algún tipo de asesoramiento, orientación o recomendación por parte de los organismos públicos sobre cuestiones como la zona en la que abrir su negocio?

5 respuestas



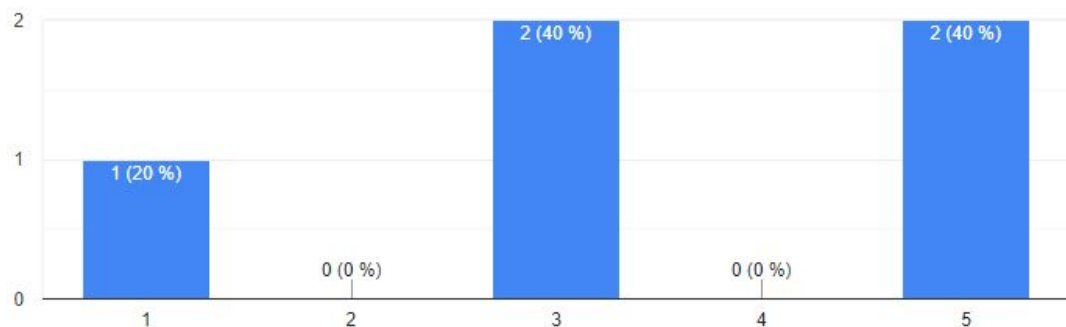
Indique qué tipo de información se proporciona sobre la zona en la que abrir su negocio:

0 respuestas

Aún no hay respuestas para esta pregunta.

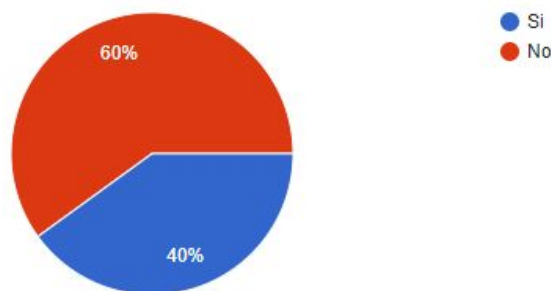
¿Cree que sería necesario o recomendable por parte de quien emprende recibir algún tipo de asesoramiento de las Administraciones Públicas sobre la ubicación más propicia para su negocio?

5 respuestas



Dentro de las herramientas que ponen a disposición para asesorar a los futuros emprendedores, ¿utilizan algún tipo de aplicación que proporcione información con datos sociodemográficos, económicos y urbanísticos de la zona elegida para su establecimiento?

5 respuestas



Especifique el tipo de herramienta o fuente de información que se utiliza

2 respuestas

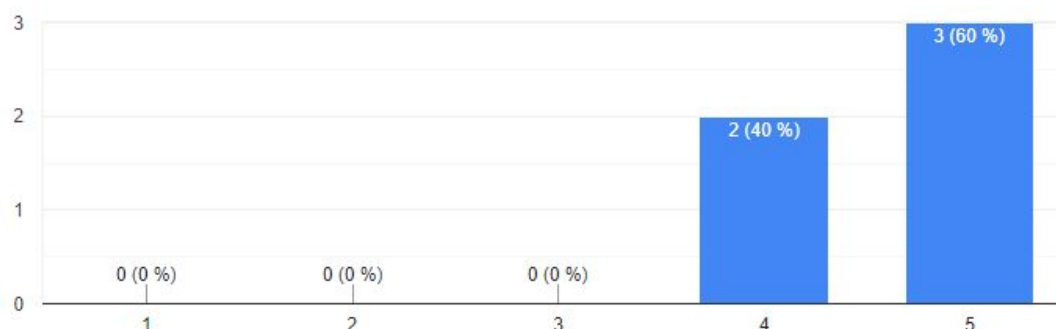
Datos Abiertos, en la página web del Ayuntamiento

Ediciones del Departamento de Estadística y páginas webs

¿Cree que disponer de esta información serviría para mejorar la gestión del tejido empresarial de su Comunidad/Municipio y satisfacer las necesidades de sus ciudadanos en este sentido?

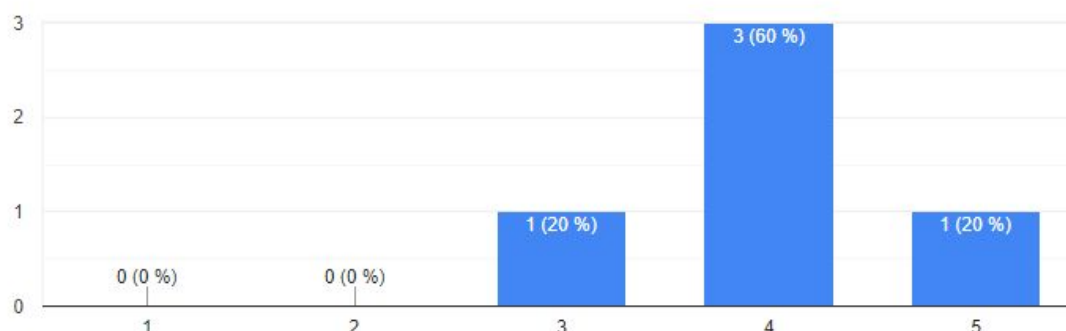
Valore la prestación de "mejor ubicación" en función del tipo de negocio

5 respuestas



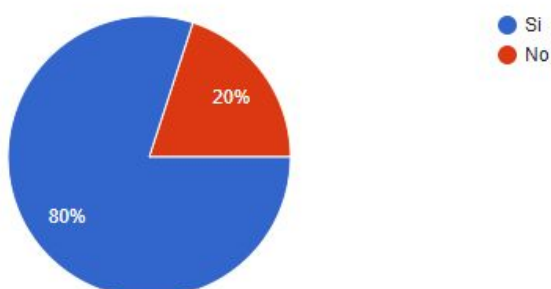
Valore la prestación de "mejor tipo de negocio en función de una ubicación dada"

5 respuestas



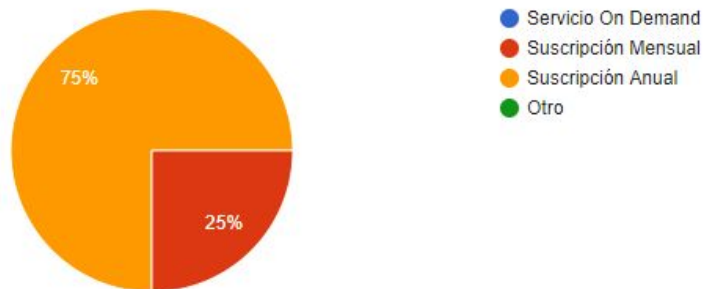
¿Consideraría la opción de pagar por un servicio o herramienta que se integrase dentro de sus sistemas de información y que prestase los servicios descritos en la pregunta anterior?

5 respuestas



Especifique por favor la modalidad de pago que más se ajusta a su caso particular.

4 respuestas



Especifique por favor la modalidad de pago que más se ajusta a su caso particular.

0 respuestas

Aún no hay respuestas para esta pregunta.

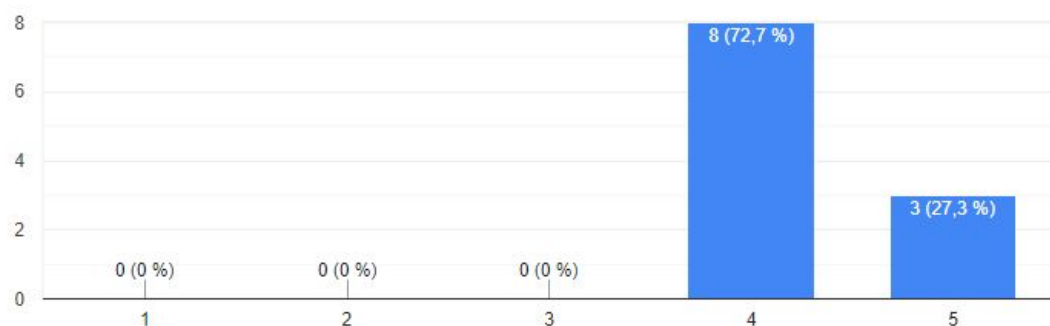
Entidades bancarias

De las 11 personas encuestadas hemos obtenido el 100% de las respuestas

¿Qué tipo de factores piensa que afectan negativamente en el éxito de un negocio?

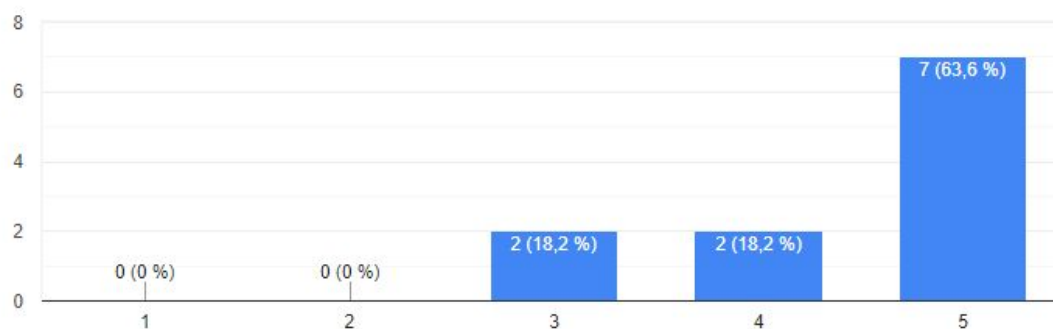
La alta incertidumbre en el mercado

11 respuestas



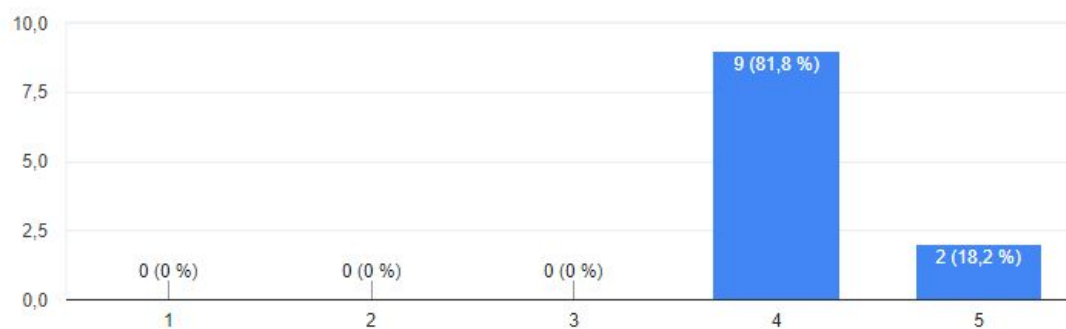
Desconocimiento del mercado potencial de clientes

11 respuestas



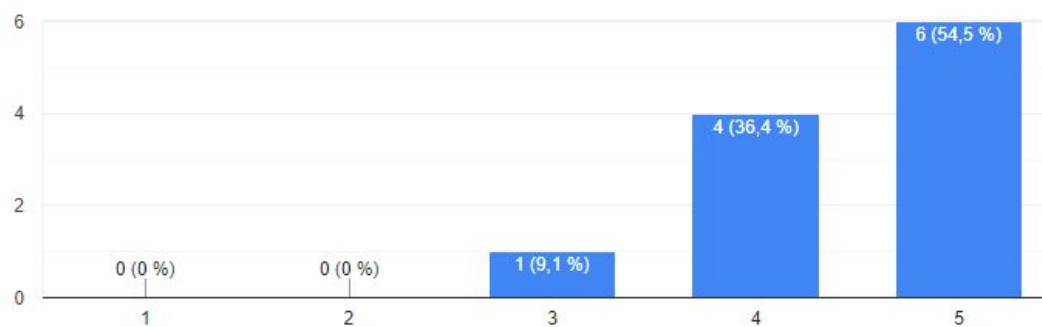
Mala ubicación

11 respuestas



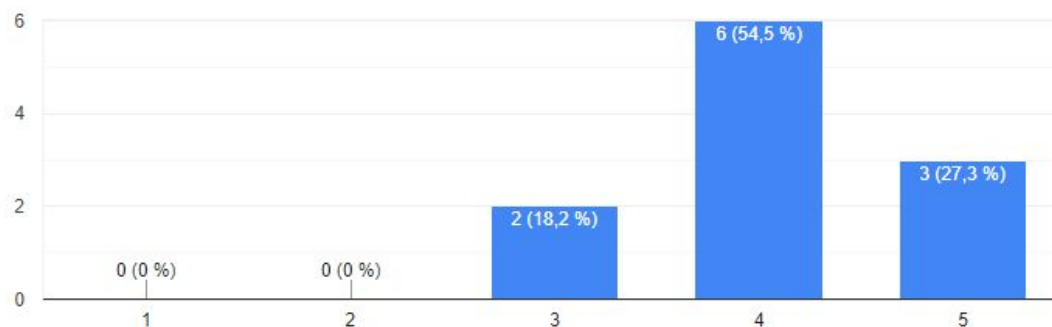
Mala administración

11 respuestas



El no uso de nuevas tecnologías y poca digitalización

11 respuestas



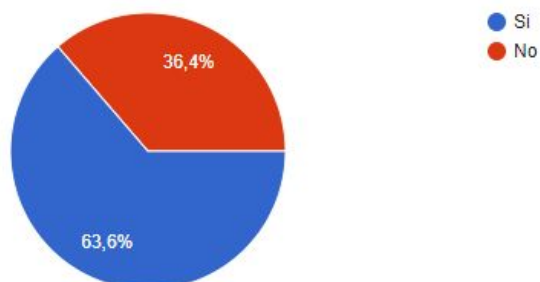
Otros (especifique)

0 respuestas

Aún no hay respuestas para esta pregunta.

Cuando un emprendedor acude a vosotros ¿tiene clara la ubicación de su negocio desde el inicio?

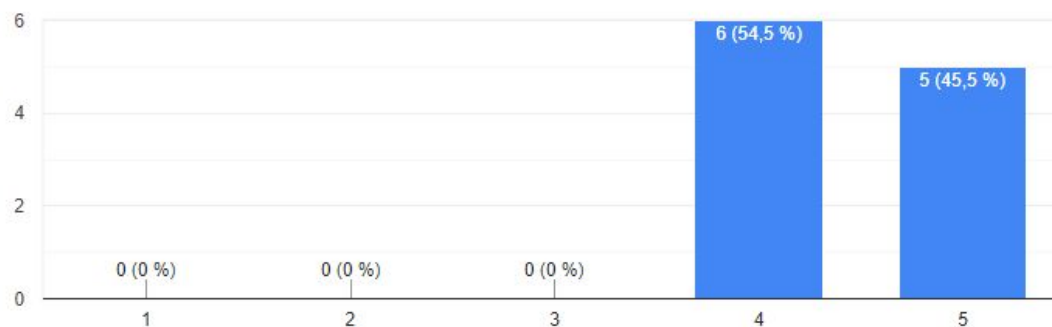
11 respuestas



Valore aquellos factores que pueden condicionar la decisión sobre la ubicación de un negocio

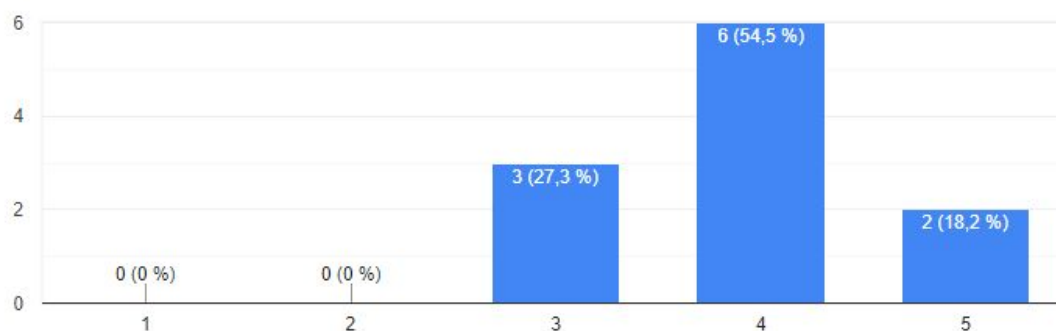
El coste del alquiler del local, así como de las tasas y licencias

11 respuestas



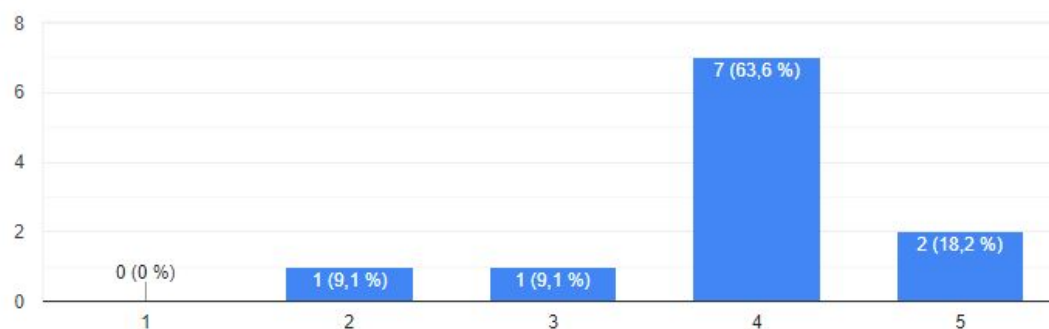
Que esté bien comunicado a través de la red de transporte público

11 respuestas



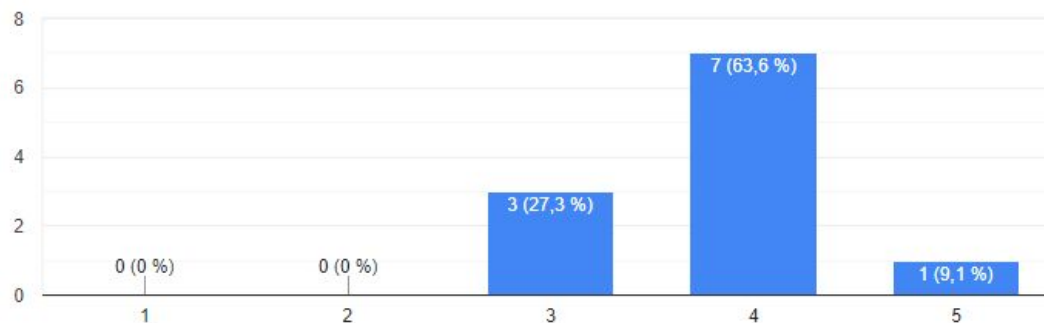
Que sea accesible por transporte privado y haya un buen servicio de estacionamiento

11 respuestas



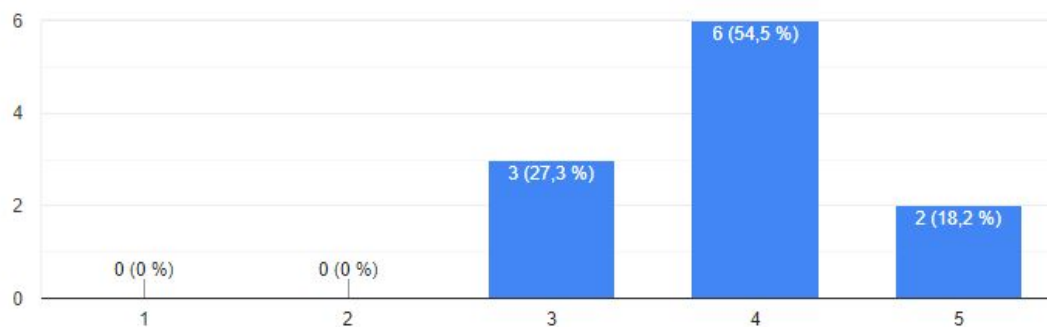
Que esté cercano a las calles o vías principales de la localidad

11 respuestas



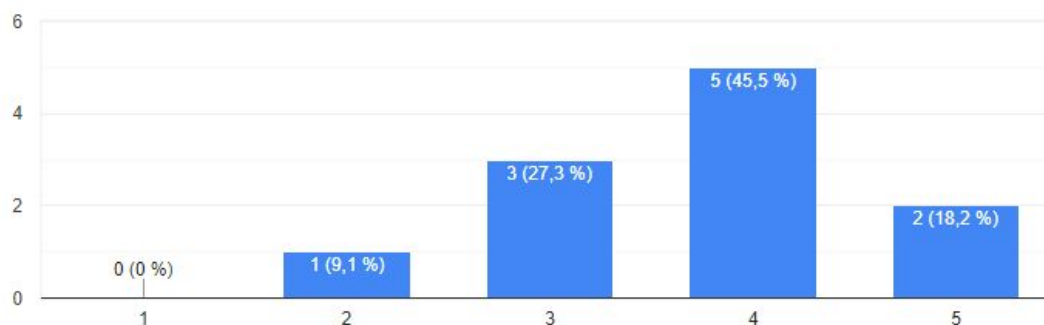
Que esté situado en una vía peatonal o en la que haya un importante número de transeúntes

11 respuestas



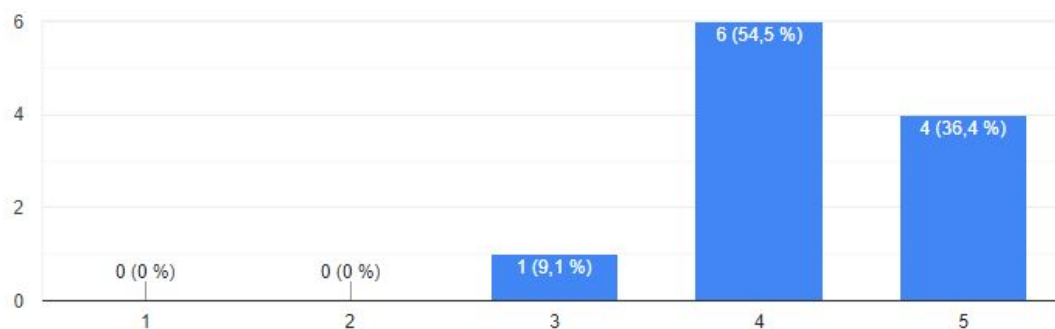
Que sea fácilmente accesible a pie y/o para aquellas personas con movilidad reducida

11 respuestas



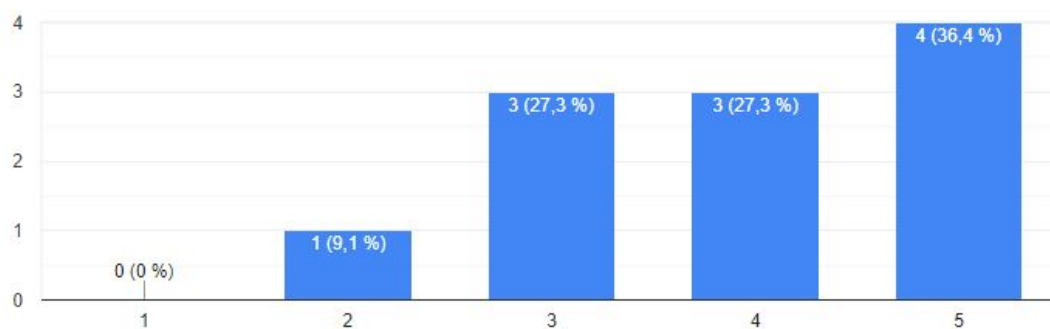
La existencia de competidores directos

11 respuestas



La cercanía de negocios o actividades complementarias

11 respuestas



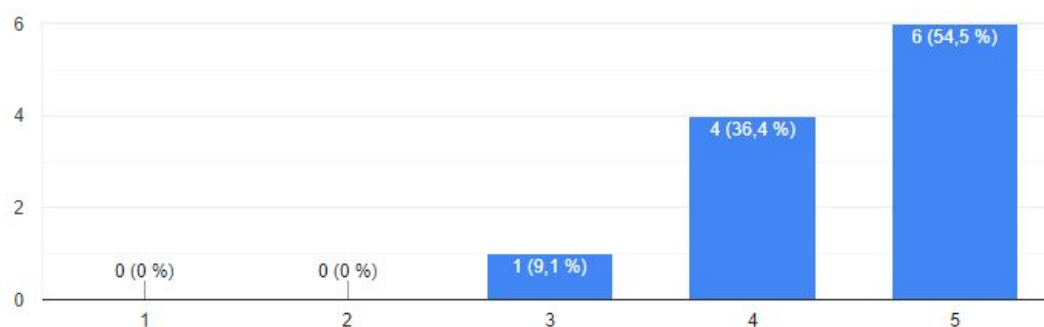
Otros: (Especifique)

0 respuestas

Aún no hay respuestas para esta pregunta.

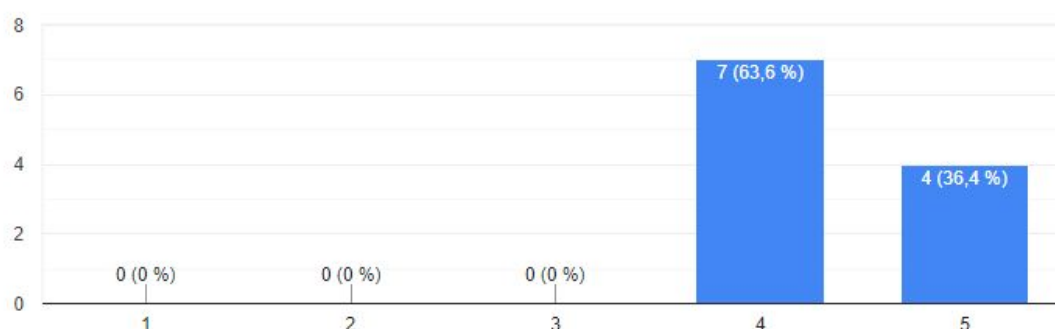
¿Cree que sería necesario o recomendable por parte de quien emprende recibir algún tipo de asesoramiento sobre la ubicación más propicia para su negocio?

11 respuestas



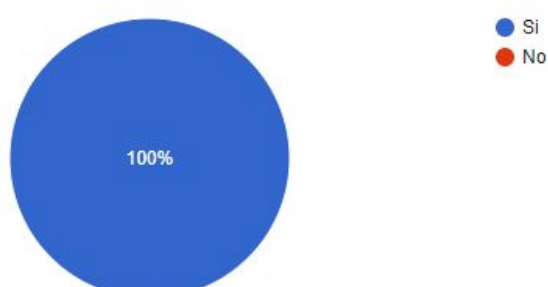
Si supiera de la existencia de una herramienta online que le sugiriera al emprendedor la mejor ubicación para abrir su negocio, o el tipo de industria más acorde para abrir un negocio en una zona específica de la ciudad ¿la utilizaría para su negocio? ¿La ofrecería a sus clientes como parte de los servicios que ofrece?

11 respuestas



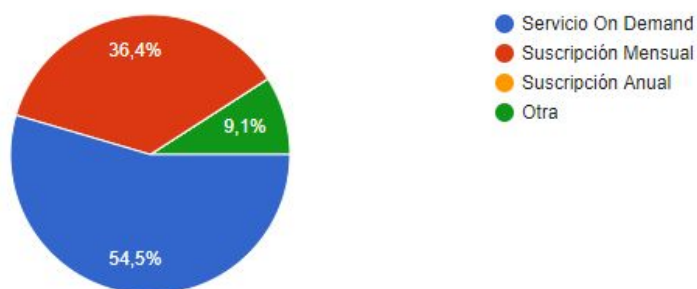
¿Consideraría la opción de pagar por un servicio o herramienta que se integrase dentro de sus sistemas de información y que prestase los servicios descritos en la pregunta anterior?

11 respuestas



Especifique la modalidad de pago que más se ajusta a su caso particular

11 respuestas



Especifique la modalidad de pago que más se ajusta a su caso particular

1 respuesta

Otra

Anexo II: Código de la aplicación

1. Funciones usadas en los distintos códigos en R

#Llamada a las librerías necesarias

```
if(!require('RMySQL')) install.packages('RMySQL')
library(RMySQL)
if(!require('VIM')) install.packages('VIM')
library(VIM)
if(!require('ggmap')) install.packages('ggmap')
library(ggmap)
if(!require('rgdal')) install.packages('rgdal')
library(rgdal)
if(!require('rgeos')) install.packages('rgeos')
library(rgeos)
if(!require('purrr')) install.packages('purrr')
library(purrr)
library(RMySQL)
if(!require('osrm')) install.packages('osrm')
library(osrm)
if(!require('geosphere')) install.packages('geosphere')
library(geosphere)
library(dplyr)
library(magrittr)
library(tidyr)
```

función que limpia el dataset base con los criterios definidos: locales abiertos, puerta calle, con coordenadas, actividades principales y sin NAs

```
f_get_cleaned_censo_locales <- function(x) {
```

```
  f_mydb = dbConnect(MySQL(), user='geopymes_workspace', password='geoeoi07$',
    dbname='geopymes', host='34.72.143.222')
```

```
  f_rs_censos_locales_final=dbSendQuery(f_mydb,"select * from ra_censo_locales_final where
    desc_tipo_acceso_local='Puerta Calle'")
  x=fetch(f_rs_censos_locales_final,n=-1)
```

```
  x<-x[x$coordenada_x_local!='0',]
```

```
  x<-x[x$desc_situacion_local=="Abierto",]
```

```
  x$cant_poi_totales<-x$Cantidad_centros_salud+      x$Cantidad_cines+
x$Cantidad_educapr+  x$Cantidad_educapu+  x$Cantidad_Farmacias+ x$Cantidad_Gasolineras+
    x$Cantidad_Hoteles+  x$Cantidad_librerias+  x$Cantidad_metro_bocas+
x$Cantidad_museos+  x$Cantidad_parroquias+x$Cantidad_juzgados+  x$Cantidad_parking+
x$Cantidad_teatros+  x$Cantidad_univers+  x$Cantidad_albergues+ x$Cantidad_libreriaao+
```

```
x$Cantidad_oficinas_turismo+ x$Cantidad_centros_comerciales+
x$Cantidad_comerespecial+ x$Cantidad_correos+ x$Cantidad_estcerca+
x$Cantidad_guardia_civil+ x$Cantidad_policia_local+ x$Cantidad_policia_nacional+
x$Cantidad_mercado_abastos+ x$Cantidad_saludotros+ x$Cantidad_bomberos+
x$Cantidad_hospitales+ x$Cantidad_mercadillos+ x$Cantidad_puentes+
x$Cantidad_educase+ x$Cantidad_helisuper+ x$Cantidad_hipermercados+
x$Cantidad_parques_tematicos+ x$Cantidad_itv+ x$Cantidad_cementerios+
x$Cantidad_cruz_roja
```

```
x<-x[!is.na(x$Empresas_alta_liquidez),]
```

#seleccionamos las 35 actividades más importantes (volumen significativo)

```
x<- x %>% filter(actividad %in% c('ALIMENTACION', 'ALMACEN', 'APUESTAS',
'ASESORIA_GESTORIA', 'BAR_RESTAURANTE', 'BAZAR', 'CARNICERIA', 'CENTRO_ESTETICA',
'CHARCUTERIA', 'DROGUERIA', 'ELECTRODOMESTICOS', 'ELECTRONICA', 'FERRETERIA',
'FISIOTERAPIA', 'FLORISTERIA', 'FOTOGRAFIA', 'FRUTERIA', 'INMOBILIARIA', 'JOYERIA',
'JUGUETERIA', 'LIBRERÍA', 'LIMPIEZA', 'LOCUTORIO', 'MUEBLES', 'OPTICA',
'PANADERIA', 'PAPELERIA', 'PELUQUERIA', 'PERFUMERIA', 'PESCADERIA', 'ROPA', 'TALLER',
'TINTORERIA', 'VETERINARIO', 'ZAPATERIA'))
```

```
rm(f_mydb)
rm(f_rs_censos_locales_final)
```

```
return (x)
```

```
}
```

#función necesaria para eliminar conexiones hacia la base de datos MySQL

```
f_killDbConnections <- function () {

  all_cons <- dbListConnections(MySQL())

  print(all_cons)

  for(con in all_cons)
    + dbDisconnect(con)

  print(paste(length(all_cons), " connections killed."))

}
```

2. Creación y limpieza del dataset final

```
# Crea dataset final
setwd("C:/Users/mgalonsg/documents/geopymes")
install.packages("RMySQL")
library(RMySQL)
mydb=dbConnect(MySQL(), user='geopymes_workspace', password='geoeoi07$',
```

```
dbname='geopymes', host='34.72.143.222')
dbListTables(mydb)
```

Aplica función para obtener datasetlimpio, aplicando criterios fijados.

```
source("/Dev Env - R Scripts/functions.R")
```

```
install.packages("data.table")
```

```
library(data.table)
```

```
dataset_limpio<-f_get_cleaned_censo_locales()
```

```
str(dataset_limpio)
```

```
dim(dataset_limpio)
```

```
#46729x170
```

Crea variables agrupadas de los Puntos de interés

```
dataset_limpio <-mutate(dataset_limpio,
```

```
  puntos_ocio =
```

```
Cantidad_cines+Cantidad_teatros+Cantidad_libreriaao+Cantidad_librerias+Cantidad_parroquias+
```

```
Cantidad_cementerios+Cantidad_parques_tematicos
```

```
)
```

```
dataset_limpio<-mutate(dataset_limpio,
```

```
  puntos_comerciales
```

```
=Cantidad_hipermercados+Cantidad_helisuper+Cantidad_mercadillos+Cantidad_mercado_abastos
```

```
+Cantidad_comerespecial+Cantidad_centros_comerciales
```

```
)
```

```
dataset_limpio<-mutate(dataset_limpio,
```

```
  puntos_salud =
```

```
Cantidad_centros_salud+Cantidad_Farmacias+Cantidad_cruz_roja+Cantidad_hospitales+Cantidad
```

```
_saludotros
```

```
)
```

```
dataset_limpio<-mutate(dataset_limpio,
```

```
  puntos_transporte =
```

```
Cantidad_parking+Cantidad_metro_bocas+Cantidad_Gasolineras+Cantidad_itv+Cantidad_estcerca
```

```
+Cantidad_puentes
```

```
)
```

```
dataset_limpio<-mutate(dataset_limpio,
```

```
  puntos_turismo =
```

```
Cantidad_museos+Cantidad_Hoteles+Cantidad_albergues+Cantidad_oficinas_turismo
```

```
)
```

```
dataset_limpio<-mutate(dataset_limpio,
```

```
  puntos_educacion =
```

```
Cantidad_educapr+Cantidad_univers+Cantidad_educapu+Cantidad_educase
```

```
)
```

```
dataset_limpio<-mutate(dataset_limpio,
```

```
  puntos_organismos = Cantidad_juzgados+Cantidad_correos
```

```
)
```

```
dataset_limpio<-mutate(dataset_limpio,
```

```
  puntos_seguridad =
```

```
Cantidad_bomberos+Cantidad_policia_nacional+Cantidad_policia_local+Cantidad_guardia_civil
```

```
)
```

```

dim(dataset_limpio)

# Incluye variable "transito" categorizada en función de los POI agrupados

dataset_limpio <- mutate(dataset_limpio,
                          num_transito_alto =
puntos_ocio+puntos_comerciales+puntos_turismo+puntos_transporte,
                          num_transito_medio = puntos_educacion+puntos_salud,
                          num_transito_bajo = puntos_seguridad+puntos_organismos
                          )

dataset_limpio <- mutate(dataset_limpio,
                          transito_alto =
case_when(num_transito_alto>=num_transito_medio&num_transito_alto>=num_transito_bajo ~
1,
TRUE ~
0),

                          transito_medio =
case_when(num_transito_medio>num_transito_alto&num_transito_medio>num_transito_bajo ~
1,
TRUE ~
0),

                          transito_bajo =
case_when(num_transito_bajo>num_transito_alto&num_transito_bajo>num_transito_medio ~ 1,
TRUE ~
0)
                          )

# Unimos al dataframe anterior (locales + POI + Transito + Sociodemograficas)
#nos traemos la tabla de SQL de sociodemograficas

f_mydb = dbConnect(MySQL(), user='geopymes_workspace', password='geoeoi07$',
dbname='geopymes', host='34.72.143.222')

eg3=dbSendQuery(f_mydb,"select * from mm_datos_sociodemograficos_censales_final")
data_socio=fetch(eg3,n=-1)

#quitamos outliers de la tabla sociodemograficos

#vble sup_km2

boxplot(data_socio$SUP_KM2,
        main = "sup",
        boxwex = 0.5,col="blue")

impute_outliers <- function(x, removeNA = TRUE){
  quantiles <- quantile(x, c(0.05, 0.95), na.rm = removeNA)

```

```

x[x<quantiles[1]] <- mean(x, na.rm = removeNA)
x[x>quantiles[2]] <- median(x, na.rm = removeNA)
x
}
imputed_data <- impute_outliers(data_socio$SUP_KM2)

par(mfrow = c(1,2))

boxplot(data_socio$SUP_KM2, main = "Presión con outliers",
        col = 3)
boxplot(imputed_data, main = "Presión sin outliers",col=2)

replace_outliers <- function(x, removeNA = TRUE){
  qrts <- quantile(x, probs = c(0.25, 0.75), na.rm = removeNA)
  caps <- quantile(x, probs = c(.05, .95), na.rm = removeNA)
  iqr <- qrts[2]-qrts[1]
  h <- 1.5 * iqr
  x[x<qrts[1]-h] <- caps[1]
  x[x>qrts[2]+h] <- caps[2]
  x
}

data_socio$SUP_KM2 <- replace_outliers(data_socio$SUP_KM2)

data_socio

#vble Poblacion

boxplot(data_socio$POBLACION,
        main = "pob",
        boxwex = 0.5,col="blue")

impute_outliers <- function(x, removeNA = TRUE){
  quantiles <- quantile(x, c(0.05, 0.95), na.rm = removeNA)
  x[x<quantiles[1]] <- mean(x, na.rm = removeNA)
  x[x>quantiles[2]] <- median(x, na.rm = removeNA)
  x
}
imputed_data <- impute_outliers(data_socio$POBLACION)

par(mfrow = c(1,2))

boxplot(data_socio$POBLACION, main = "Presión con outliers",
        col = 3)
boxplot(imputed_data, main = "Presión sin outliers",col=2)

replace_outliers <- function(x, removeNA = TRUE){
  qrts <- quantile(x, probs = c(0.25, 0.75), na.rm = removeNA)
  caps <- quantile(x, probs = c(.05, .95), na.rm = removeNA)
  iqr <- qrts[2]-qrts[1]
  h <- 1.5 * iqr
  x[x<qrts[1]-h] <- caps[1]

```



```

x[x>qrts[2]+h] <- caps[2]
x
}

data_socio$POBLACION <- replace_outliers(data_socio$POBLACION)

data_socio

#vble hombres

boxplot(data_socio$HOMBRES,
        main = "hom",
        boxwex = 0.5,col="blue")

impute_outliers <- function(x, removeNA = TRUE){
  quantiles <- quantile(x, c(0.05, 0.95), na.rm = removeNA)
  x[x<quantiles[1]] <- mean(x, na.rm = removeNA)
  x[x>quantiles[2]] <- median(x, na.rm = removeNA)
  x
}
imputed_data <- impute_outliers(data_socio$HOMBRES)

par(mfrow = c(1,2))

boxplot(data_socio$HOMBRES, main = "Presión con outliers",
        col = 3)
boxplot(imputed_data, main = "Presión sin outliers",col=2)

replace_outliers <- function(x, removeNA = TRUE){
  qrts <- quantile(x, probs = c(0.25, 0.75), na.rm = removeNA)
  caps <- quantile(x, probs = c(.05, .95), na.rm = removeNA)
  iqr <- qrts[2]-qrts[1]
  h <- 1.5 * iqr
  x[x<qrts[1]-h] <- caps[1]
  x[x>qrts[2]+h] <- caps[2]
  x
}

data_socio$HOMBRES <- replace_outliers(data_socio$HOMBRES)

#vble mujeres

boxplot(data_socio$MUJERES,
        main = "muj",
        boxwex = 0.5,col="blue")

impute_outliers <- function(x, removeNA = TRUE){
  quantiles <- quantile(x, c(0.05, 0.95), na.rm = removeNA)
  x[x<quantiles[1]] <- mean(x, na.rm = removeNA)
  x[x>quantiles[2]] <- median(x, na.rm = removeNA)
  x
}

```

```

imputed_data <- impute_outliers(data_socio$MUJERES)

par(mfrow = c(1,2))

boxplot(data_socio$MUJERES, main = "Presión con outliers",
        col = 3)
boxplot(imputed_data, main = "Presión sin outliers",col=2)

replace_outliers <- function(x, removeNA = TRUE){
  qrts <- quantile(x, probs = c(0.25, 0.75), na.rm = removeNA)
  caps <- quantile(x, probs = c(.05, .95), na.rm = removeNA)
  iqr <- qrts[2]-qrts[1]
  h <- 1.5 * iqr
  x[x<qrts[1]-h] <- caps[1]
  x[x>qrts[2]+h] <- caps[2]
  x
}

data_socio$MUJERES <- replace_outliers(data_socio$MUJERES)

data_socio

#vble ESPANYOLES

boxplot(data_socio$ESPANYOLES,
        main = "esp",
        boxwex = 0.5,col="blue")

impute_outliers <- function(x, removeNA = TRUE){
  quantiles <- quantile(x, c(0.05, 0.95), na.rm = removeNA)
  x[x<quantiles[1]] <- mean(x, na.rm = removeNA)
  x[x>quantiles[2]] <- median(x, na.rm = removeNA)
  x
}
imputed_data <- impute_outliers(data_socio$ESPANYOLES)

par(mfrow = c(1,2))

boxplot(data_socio$ESPANYOLES, main = "Presión con outliers",
        col = 3)
boxplot(imputed_data, main = "Presión sin outliers",col=2)

replace_outliers <- function(x, removeNA = TRUE){
  qrts <- quantile(x, probs = c(0.25, 0.75), na.rm = removeNA)
  caps <- quantile(x, probs = c(.05, .95), na.rm = removeNA)
  iqr <- qrts[2]-qrts[1]
  h <- 1.5 * iqr
  x[x<qrts[1]-h] <- caps[1]
  x[x>qrts[2]+h] <- caps[2]
  x
}

```

```
data_socio$ESPANYOLES <- replace_outliers(data_socio$ESPANYOLES)
```

```
data_socio
```

```
#vble EXTRANJEROS
```

```
boxplot(data_socio$EXTRANJEROS ,  
        main = "ext",  
        boxwex = 0.5,col="blue")
```

```
impute_outliers <- function(x, removeNA = TRUE){  
  quantiles <- quantile(x, c(0.05, 0.95), na.rm = removeNA)  
  x[x<quantiles[1]] <- mean(x, na.rm = removeNA)  
  x[x>quantiles[2]] <- median(x, na.rm = removeNA)  
  x  
}  
imputed_data <- impute_outliers(data_socio$EXTRANJEROS)
```

```
par(mfrow = c(1,2))
```

```
boxplot(data_socio$EXTRANJEROS, main = "Presión con outliers",  
        col = 3)  
boxplot(imputed_data, main = "Presión sin outliers",col=2)
```

```
replace_outliers <- function(x, removeNA = TRUE){  
  qrts <- quantile(x, probs = c(0.25, 0.75), na.rm = removeNA)  
  caps <- quantile(x, probs = c(.05, .95), na.rm = removeNA)  
  iqr <- qrts[2]-qrts[1]  
  h <- 1.5 * iqr  
  x[x<qrts[1]-h] <- caps[1]  
  x[x>qrts[2]+h] <- caps[2]  
  x  
}
```

```
data_socio$EXTRANJEROS<- replace_outliers(data_socio$EXTRANJEROS)
```

```
data_socio
```

```
#vble POB_0_14
```

```
boxplot(data_socio$POB_0_14 ,  
        main = "14",  
        boxwex = 0.5,col="blue")
```

```
impute_outliers <- function(x, removeNA = TRUE){  
  quantiles <- quantile(x, c(0.05, 0.95), na.rm = removeNA)  
  x[x<quantiles[1]] <- mean(x, na.rm = removeNA)  
  x[x>quantiles[2]] <- median(x, na.rm = removeNA)  
  x  
}  
imputed_data <- impute_outliers(data_socio$POB_0_14)
```

```

par(mfrow = c(1,2))

boxplot(data_socio$POB_0_14, main = "Presión con outliers",
        col = 3)
boxplot(imputed_data, main = "Presión sin outliers",col=2)

replace_outliers <- function(x, removeNA = TRUE){
  qrts <- quantile(x, probs = c(0.25, 0.75), na.rm = removeNA)
  caps <- quantile(x, probs = c(.05, .95), na.rm = removeNA)
  iqr <- qrts[2]-qrts[1]
  h <- 1.5 * iqr
  x[x<qrts[1]-h] <- caps[1]
  x[x>qrts[2]+h] <- caps[2]
  x
}

data_socio$POB_0_14<- replace_outliers(data_socio$POB_0_14)

data_socio

#vble POB_15_29

boxplot(data_socio$POB_15_29 ,
        main = "29",
        boxwex = 0.5,col="blue")

impute_outliers <- function(x, removeNA = TRUE){
  quantiles <- quantile(x, c(0.05, 0.95), na.rm = removeNA)
  x[x<quantiles[1]] <- mean(x, na.rm = removeNA)
  x[x>quantiles[2]] <- median(x, na.rm = removeNA)
  x
}
imputed_data <- impute_outliers(data_socio$POB_15_29)

par(mfrow = c(1,2))

boxplot(data_socio$POB_15_29, main = "Presión con outliers",
        col = 3)
boxplot(imputed_data, main = "Presión sin outliers",col=2)

replace_outliers <- function(x, removeNA = TRUE){
  qrts <- quantile(x, probs = c(0.25, 0.75), na.rm = removeNA)
  caps <- quantile(x, probs = c(.05, .95), na.rm = removeNA)
  iqr <- qrts[2]-qrts[1]
  h <- 1.5 * iqr
  x[x<qrts[1]-h] <- caps[1]
  x[x>qrts[2]+h] <- caps[2]
  x
}

data_socio$POB_15_29<- replace_outliers(data_socio$POB_15_29)

```

```
data_socio

#vble POB_30_44

boxplot(data_socio$POB_30_44 ,
        main = "44",
        boxwex = 0.5,col="blue")

impute_outliers <- function(x, removeNA = TRUE){
  quantiles <- quantile(x, c(0.05, 0.95), na.rm = removeNA)
  x[x<quantiles[1]] <- mean(x, na.rm = removeNA)
  x[x>quantiles[2]] <- median(x, na.rm = removeNA)
  x
}
imputed_data <- impute_outliers(data_socio$POB_30_44)

par(mfrow = c(1,2))

boxplot(data_socio$POB_30_44, main = "Presión con outliers",
        col = 3)
boxplot(imputed_data, main = "Presión sin outliers",col=2)

replace_outliers <- function(x, removeNA = TRUE){
  qrts <- quantile(x, probs = c(0.25, 0.75), na.rm = removeNA)
  caps <- quantile(x, probs = c(.05, .95), na.rm = removeNA)
  iqr <- qrts[2]-qrts[1]
  h <- 1.5 * iqr
  x[x<qrts[1]-h] <- caps[1]
  x[x>qrts[2]+h] <- caps[2]
  x
}

data_socio$POB_30_44<- replace_outliers(data_socio$POB_30_44)

data_socio

#vble POB_45_64

boxplot(data_socio$POB_45_64 ,
        main = "64",
        boxwex = 0.5,col="blue")

impute_outliers <- function(x, removeNA = TRUE){
  quantiles <- quantile(x, c(0.05, 0.95), na.rm = removeNA)
  x[x<quantiles[1]] <- mean(x, na.rm = removeNA)
  x[x>quantiles[2]] <- median(x, na.rm = removeNA)
  x
}
imputed_data <- impute_outliers(data_socio$POB_45_64)

par(mfrow = c(1,2))
```

```

boxplot(data_socio$POB_45_64, main = "Presión con outliers",
        col = 3)
boxplot(imputed_data, main = "Presión sin outliers",col=2)

replace_outliers <- function(x, removeNA = TRUE){
  qrts <- quantile(x, probs = c(0.25, 0.75), na.rm = removeNA)
  caps <- quantile(x, probs = c(.05, .95), na.rm = removeNA)
  iqr <- qrts[2]-qrts[1]
  h <- 1.5 * iqr
  x[x<qrts[1]-h] <- caps[1]
  x[x>qrts[2]+h] <- caps[2]
  x
}

data_socio$POB_45_64<- replace_outliers(data_socio$POB_45_64)

data_socio

#vble POB_65_79

boxplot(data_socio$POB_65_79 ,
        main = "79",
        boxwex = 0.5,col="blue")

impute_outliers <- function(x, removeNA = TRUE){
  quantiles <- quantile(x, c(0.05, 0.95), na.rm = removeNA)
  x[x<quantiles[1]] <- mean(x, na.rm = removeNA)
  x[x>quantiles[2]] <- median(x, na.rm = removeNA)
  x
}
imputed_data <- impute_outliers(data_socio$POB_65_79)

par(mfrow = c(1,2))

boxplot(data_socio$POB_65_79, main = "Presión con outliers",
        col = 3)
boxplot(imputed_data, main = "Presión sin outliers",col=2)

replace_outliers <- function(x, removeNA = TRUE){
  qrts <- quantile(x, probs = c(0.25, 0.75), na.rm = removeNA)
  caps <- quantile(x, probs = c(.05, .95), na.rm = removeNA)
  iqr <- qrts[2]-qrts[1]
  h <- 1.5 * iqr
  x[x<qrts[1]-h] <- caps[1]
  x[x>qrts[2]+h] <- caps[2]
  x
}

data_socio$POB_65_79<- replace_outliers(data_socio$POB_65_79)

data_socio

```

```
#vble POB_MAS_80

boxplot(data_socio$POB_MAS_80 ,
        main = "80",
        boxwex = 0.5,col="blue")

impute_outliers <- function(x, removeNA = TRUE){
  quantiles <- quantile(x, c(0.05, 0.95), na.rm = removeNA)
  x[x<quantiles[1]] <- mean(x, na.rm = removeNA)
  x[x>quantiles[2]] <- median(x, na.rm = removeNA)
  x
}
imputed_data <- impute_outliers(data_socio$POB_MAS_80)

par(mfrow = c(1,2))

boxplot(data_socio$POB_MAS_80, main = "Presión con outliers",
        col = 3)
boxplot(imputed_data, main = "Presión sin outliers",col=2)

replace_outliers <- function(x, removeNA = TRUE){
  qrts <- quantile(x, probs = c(0.25, 0.75), na.rm = removeNA)
  caps <- quantile(x, probs = c(.05, .95), na.rm = removeNA)
  iqr <- qrts[2]-qrts[1]
  h <- 1.5 * iqr
  x[x<qrts[1]-h] <- caps[1]
  x[x>qrts[2]+h] <- caps[2]
  x
}

data_socio$POB_MAS_80<- replace_outliers(data_socio$POB_MAS_80)

data_socio

#vble RENTA_ANUAL_MEDIA

boxplot(data_socio$RENTA_ANUAL_MEDIA ,
        main = "renta",
        boxwex = 0.5,col="blue")

impute_outliers <- function(x, removeNA = TRUE){
  quantiles <- quantile(x, c(0.05, 0.95), na.rm = removeNA)
  x[x<quantiles[1]] <- mean(x, na.rm = removeNA)
  x[x>quantiles[2]] <- median(x, na.rm = removeNA)
  x
}
imputed_data <- impute_outliers(data_socio$RENTA_ANUAL_MEDIA)

par(mfrow = c(1,2))

boxplot(data_socio$RENTA_ANUAL_MEDIA, main = "Presión con outliers",
```



```

col = 3)
boxplot(imputed_data, main = "Presión sin outliers",col=2)

replace_outliers <- function(x, removeNA = TRUE){
  qrts <- quantile(x, probs = c(0.25, 0.75), na.rm = removeNA)
  caps <- quantile(x, probs = c(.05, .95), na.rm = removeNA)
  iqr <- qrts[2]-qrts[1]
  h <- 1.5 * iqr
  x[x<qrts[1]-h] <- caps[1]
  x[x>qrts[2]+h] <- caps[2]
  x
}

data_socio$RENTA_ANUAL_MEDIA<- replace_outliers(data_socio$RENTA_ANUAL_MEDIA)

data_socio

#vble POB_SIN_ESTUDIOS

boxplot(data_socio$POB_SIN_ESTUDIOS ,
        main = "sin",
        boxwex = 0.5,col="blue")

impute_outliers <- function(x, removeNA = TRUE){
  quantiles <- quantile(x, c(0.05, 0.95), na.rm = removeNA)
  x[x<quantiles[1]] <- mean(x, na.rm = removeNA)
  x[x>quantiles[2]] <- median(x, na.rm = removeNA)
  x
}
imputed_data <- impute_outliers(data_socio$POB_SIN_ESTUDIOS)

par(mfrow = c(1,2))

boxplot(data_socio$POB_SIN_ESTUDIOS, main = "Presión con outliers",
        col = 3)
boxplot(imputed_data, main = "Presión sin outliers",col=2)

replace_outliers <- function(x, removeNA = TRUE){
  qrts <- quantile(x, probs = c(0.25, 0.75), na.rm = removeNA)
  caps <- quantile(x, probs = c(.05, .95), na.rm = removeNA)
  iqr <- qrts[2]-qrts[1]
  h <- 1.5 * iqr
  x[x<qrts[1]-h] <- caps[1]
  x[x>qrts[2]+h] <- caps[2]
  x
}

data_socio$POB_SIN_ESTUDIOS<- replace_outliers(data_socio$POB_SIN_ESTUDIOS)

data_socio

#vble POB_CON_ESTUDIOS

```

```

boxplot(data_socio$POB_CON_ESTUDIOS ,
        main = "con",
        boxwex = 0.5,col="blue")

impute_outliers <- function(x, removeNA = TRUE){
  quantiles <- quantile(x, c(0.05, 0.95), na.rm = removeNA)
  x[x<quantiles[1]] <- mean(x, na.rm = removeNA)
  x[x>quantiles[2]] <- median(x, na.rm = removeNA)
  x
}
imputed_data <- impute_outliers(data_socio$POB_CON_ESTUDIOS)

par(mfrow = c(1,2))

boxplot(data_socio$POB_CON_ESTUDIOS, main = "Presión con outliers",
        col = 3)
boxplot(imputed_data, main = "Presión sin outliers",col=2)

replace_outliers <- function(x, removeNA = TRUE){
  qrts <- quantile(x, probs = c(0.25, 0.75), na.rm = removeNA)
  caps <- quantile(x, probs = c(.05, .95), na.rm = removeNA)
  iqr <- qrts[2]-qrts[1]
  h <- 1.5 * iqr
  x[x<qrts[1]-h] <- caps[1]
  x[x>qrts[2]+h] <- caps[2]
  x
}

data_socio$POB_CON_ESTUDIOS<- replace_outliers(data_socio$POB_CON_ESTUDIOS)

data_socio

#vble POB_PARADOS

boxplot(data_socio$POB_PARADOS ,
        main = "parados",
        boxwex = 0.5,col="blue")

impute_outliers <- function(x, removeNA = TRUE){
  quantiles <- quantile(x, c(0.05, 0.95), na.rm = removeNA)
  x[x<quantiles[1]] <- mean(x, na.rm = removeNA)
  x[x>quantiles[2]] <- median(x, na.rm = removeNA)
  x
}
imputed_data <- impute_outliers(data_socio$POB_PARADOS)

par(mfrow = c(1,2))

boxplot(data_socio$POB_PARADOS, main = "Presión con outliers",
        col = 3)
boxplot(imputed_data, main = "Presión sin outliers",col=2)

```

```

replace_outliers <- function(x, removeNA = TRUE){
  qrts <- quantile(x, probs = c(0.25, 0.75), na.rm = removeNA)
  caps <- quantile(x, probs = c(.05, .95), na.rm = removeNA)
  iqr <- qrts[2]-qrts[1]
  h <- 1.5 * iqr
  x[x<qrts[1]-h] <- caps[1]
  x[x>qrts[2]+h] <- caps[2]
  x
}

data_socio$POB_PARADOS<- replace_outliers(data_socio$POB_PARADOS)

data_socio

#vble DENSIDAD_POB

boxplot(data_socio$DENSIDAD_POB ,
        main = "den",
        boxwex = 0.5,col="blue")

impute_outliers <- function(x, removeNA = TRUE){
  quantiles <- quantile(x, c(0.05, 0.95), na.rm = removeNA)
  x[x<quantiles[1]] <- mean(x, na.rm = removeNA)
  x[x>quantiles[2]] <- median(x, na.rm = removeNA)
  x
}
imputed_data <- impute_outliers(data_socio$DENSIDAD_POB)

par(mfrow = c(1,2))

boxplot(data_socio$DENSIDAD_POB, main = "Presión con outliers",
        col = 3)
boxplot(imputed_data, main = "Presión sin outliers",col=2)

replace_outliers <- function(x, removeNA = TRUE){
  qrts <- quantile(x, probs = c(0.25, 0.75), na.rm = removeNA)
  caps <- quantile(x, probs = c(.05, .95), na.rm = removeNA)
  iqr <- qrts[2]-qrts[1]
  h <- 1.5 * iqr
  x[x<qrts[1]-h] <- caps[1]
  x[x>qrts[2]+h] <- caps[2]
  x
}

data_socio$DENSIDAD_POB<- replace_outliers(data_socio$DENSIDAD_POB)

data_socio

#añadimos vble seccion censal a local_poi_informa3

mydb = dbConnect(MySQL(), user='geopymes_workspace', password='geoeoi07$',

```

```
dbname='geopymes', host='34.72.143.222')
```

```
eg4=dbSendQuery(mydb,"select id_local,CUSEC from ra_censo_locales_x_sccc")
sscc=fetch(eg4,n=-1)
```

```
#unimos info sociodemo al dataset
```

```
dataset_sccc= merge (dataset_limpio,sscc,by="id_local", all.x=TRUE)
dataset_proyecto= merge (dataset_sccc,data_socio,by="CUSEC", all.x=TRUE)
```

```
#unimos info competidores directos al dataset
```

```
tmydb = dbConnect(MySQL(), user='geopymes_workspace', password='geoeoi07$',
dbname='geopymes', host='34.72.143.222')
```

```
eg6=dbSendQuery(tmydb,"select id_local,num_competidores from nt_columna_competidores")
competidores=fetch(eg6,n=-1)
dataset_proyecto_final= merge (dataset_proyecto,competidores,by="id_local", all.x=TRUE)
```

```
#añadir info de competidores indirectos
```

```
mydb = dbConnect(MySQL(), user='geopymes_workspace', password='geoeoi07$',
dbname='geopymes', host='34.72.143.222')
```

```
eg7=dbSendQuery(mydb,"select * from ra_censo_locales_poi_2")
data_competidores=fetch(eg7,n=-1)
```

```
#pegarle actividad
```

```
dataset_proyecto_final_2= merge
(dataset_proyecto_final,data_competidores,by.x="id_local",by.y="IdLocal",all.x=TRUE)
```

```
#añadir tasa de paro
```

```
mydb = dbConnect(MySQL(), user='geopymes_workspace', password='geoeoi07$',
dbname='geopymes', host='34.72.143.222')
```

```
eg10=dbSendQuery(f_mydb,"select * from ra_datos_censales_pob_activa_parados_2")
parados=fetch(eg10,n=-1)
```

```
parados<-mutate(parados,
tasa_paro=POB_PARADOS/POB_ACTIVAS)
```

```
dataset_proyecto_final_2= merge(dataset_proyecto_final,parados,by="CUSEC",all.x=TRUE)
```

```
# Select FINAL
```

```
dataset_final <-dataset_proyecto_final_2 %>%
filter(actividad %in% c('ALIMENTACION', 'ALMACEN', 'APUESTAS',
'ASESORIA_GESTORIA', 'BAR_RESTAURANTE', 'BAZAR','CARNICERIA', 'CENTRO_ESTETICA',
'CHARCUTERIA', 'DROGUERIA', 'ELECTRODOMESTICOS', 'ELECTRONICA',
```

```
'FERRETERIA', 'FISIOTERAPIA', 'FLORISTERIA', 'FOTOGRAFIA',
      'FRUTERIA', 'INMOBILIARIA', 'JOYERIA', 'JUGUETERIA', 'LIBRERÍA',
'LIMPIEZA', 'LOCUTORIO', 'MUEBLES', 'OPTICA', 'PANADERIA',
      'PAPELERIA', 'PELUQUERIA', 'PERFUMERIA', 'PESCADERIA', 'ROPA', 'TALLER',
      'TINTORERIA', 'VETERINARIO', 'ZAPATERIA')
```

```
)%>%
```

```
select(id_local,codigo_postal,id_distrito_local,desc_distrito_local,id_barrio_local,
      desc_barrio_local,desc_seccion_censal_local,CUSEC,
      coordenada_x_local,coordenada_y_local,id_epigrafe,CNAE_2009,actividad,puntos_ocio,
```

```
puntos_comerciales,puntos_seguridad,puntos_organismos,puntos_educacion,puntos_turismo,
      puntos_transporte,puntos_salud,transito_alto,transito_medio,transito_bajo,
      SUP_KM2,POBLACION.y,
      HOMBRES.y,MUJERES.y,
      ESPANYOLES,EXTRANJEROS,
      POB_0_14.y,POB_15_29.y,POB_30_44.y,POB_45_64.y,POB_65_79.y,POB_MAS_80,
      RENTA_ANUAL_MEDIA,
      POB_SIN_ESTUDIOS,POB_CON_ESTUDIOS,POB_PARADOS.x,
      Cantidad.ALIMENTACION,Cantidad.APUESTAS,
```

```
Cantidad.ASESORIA_GESTORIA,Cantidad.BAR_RESTAURANTE,Cantidad.BAZAR,Cantidad.CENTRO_E
STETICA,
```

```
Cantidad.ELECTRODOMESTICOS,Cantidad.FISIOTERAPIA,Cantidad.FLORISTERIA,Cantidad.FOTOG
RAFIA,
```

```
Cantidad.FRUTERIA,Cantidad.INMOBILIARIA,Cantidad.JOYERIA,Cantidad.JUGUETERIA,Cantidad.LI
BRERÍA,
```

```
Cantidad.MUEBLES,Cantidad.OPTICA,Cantidad.PANADERIA,Cantidad.PAPELERIA,Cantidad.PELUQU
ERIA,
```

```
Cantidad.PERFUMERIA,Cantidad.ROPA,Cantidad.TALLER,Cantidad.TINTORERIA,Cantidad.VETERIN
ARIO,
```

```
Cantidad.ZAPATERIA,Cantidad.CARNICERIA,Cantidad.CHARCUTERIA,Cantidad.DROGUERIA,Cantida
d.ELECTRONICA,
```

```
Cantidad.FERRETERIA,Cantidad.LIMPIEZA,Cantidad.ALMACEN,Cantidad.PESCADERIA,Cantidad.LOC
UTORIO,
```

```
      num_competidores,
```

```
prt_edad_0_14,prt_edad_15_29,prt_edad_30_44,prt_edad_45_64,prt_edad_65_79,prt_edad_may
or_80,prt_espanyoles,
```

```
      prt_extranjeros,prt_hombres,prt_mujeres,DENSIDAD_POB,tasa_paro)
```

```
names(dataset_final) <- c("id_local","codigo_postal","id_distrito","desc_distrito","id_barrio",
```

```
      "desc_barrio","desc_sccc","id_sccc",
```

```
      "coordenada_x","coordenada_y","id_epigrafe","cnae","actividad","puntos_ocio",
```

```
      "puntos_comerciales","puntos_seguridad","puntos_organismos",
```

```
      "puntos_educacion","puntos_turismo",
```

```
      "puntos_transporte","puntos_salud","transito_alto","transito_medio",
```

```
      "transito_bajo",
```

```
      "superficie",
```

```
      "poblacion",
```

```
      "hombres",
```

```
      "mujeres",
```

```
      "espanyoles",
```

```

"extranjeros",
"edad_0_14",
"edad_15_29",
"edad_30_44",
"edad_45_64",
"edad_65_79",
"edad_mayor_80",
"renta_anual_media",
"sin_estudios",
"con_estudios",
"parados",
"alimentacion",
"apuestas","gestoria","bar_restaurante","bazar","estetica",
"electrodomesticos","fisioterapia","floristeria","fotografia",
"fruteria","inmobiliaria","joyeria","jugueteria","libreria",
"muebles","optica","panaderia","papeleria","peluqueria",
"perfumeria","ropa","taller","tintoreria","veterinario",
"zapateria","carniceria","charcuteria","drogueria","electronica",
"ferreteria","limpieza","almacen","pescaderia","locutorio","competidores",
"prt_edad_0_14",
"prt_edad_15_29",
"prt_edad_30_44",
"prt_edad_45_64",
"prt_edad_65_79",
"prt_edad_mayor_80",
"prt_espanyoles",
"prt_extranjeros",
"prt_hombres",
"prt_mujeres",
"densidad_pob","tasa_paro")

```

#exportamos tabla a MySQL pero antes en excel reemplazamos la coma por el punto en todas las columnas

```
write.table(dataset_final, file = "dataset_final_locales.csv", sep=';', dec = '.', row.names = FALSE,
col.names = TRUE)
```

#Agrupamos por código postal cada actividad seleccionada para el estudio de importancia de variables. Serán distintas según las reglas de negocio de cada tipo de negocio.

```

dataset_joyerias <-dataset_final %>%
  filter(actividad == "JOYERIA") %>%
  group_by(codigo_postal) %>%
  summarize(num_joyerias=n(),
            puntos_ocio=mean(puntos_ocio),
            puntos_comerciales=mean(puntos_comerciales),
            puntos_seguridad=mean(puntos_seguridad),
            puntos_organismos=mean(puntos_organismos),
            puntos_educacion=mean(puntos_educacion),
            puntos_turismo=mean(puntos_turismo),
            puntos_transporte=mean(puntos_transporte),
            puntos_salud=mean(puntos_salud),

```

```
transito_alto=mean(transito_alto),
transito_medio=mean(transito_medio),
transito_bajo=mean(transito_bajo),
prt_hombres=mean(prt_hombres),
prt_mujeres=mean(prt_mujeres),
prt_edad_0_14=mean(prt_edad_0_14),
prt_edad_15_29=mean(prt_edad_15_29),
prt_edad_30_44=mean(prt_edad_30_44),
prt_edad_45_64=mean(prt_edad_45_64),
prt_edad_65_79=mean(prt_edad_65_79),
prt_edad_mayor_80=mean(prt_edad_mayor_80),
renta_anual_media=mean(renta_anual_media),
alimentacion=mean(alimentacion),
apuestas=mean(apuestas),
gestoria=mean(gestoria),
bar_restaurante=mean(bar_restaurante),
bazar=mean(bazar),
estetica=mean(estetica),
electrodomesticos=mean(electrodomesticos),
fisioterapia=mean(fisioterapia),
floristeria=mean(floristeria),
fotografia=mean(fotografia),
fruteria=mean(fruteria),
inmobiliaria=mean(inmobiliaria),
joyeria=mean(joyeria),
jugueteria=mean(jugueteria),
libreria=mean(libreria),
muebles=mean(muebles),
optica=mean(optica),
panaderia=mean(panaderia),
papeleria=mean(papeleria),
peluqueria=mean(peluqueria),
perfumeria=mean(perfumeria),
ropa=mean(ropa),
taller=mean(taller),
tintoreria=mean(tintoreria),
veterinario=mean(veterinario),
zapateria=mean(zapateria),
carniceria=mean(carniceria),
charcuteria=mean(charcuteria),
drogueria=mean(drogueria),
electronica=mean(electronica),
ferreteria=mean(ferreteria),
limpieza=mean(limpieza),
almacen=mean(almacen),
pescaderia=mean(pescaderia),
locutorio=mean(locutorio),
competidores=mean(competidores),
densidad_pob=mean(densidad_pob),
tasa_paro=mean(tasa_paro))
```

```
write.table(dataset_joyerias, file = "dataset_joyerias.csv", sep=';', dec = ',', row.names = FALSE,
```



```
col.names = TRUE)
```

```
dataset_papeleria <- dataset_final %>%  
  filter(actividad == "PAPELERIA") %>%  
  group_by(codigo_postal) %>%  
  summarize(num_papeleria=n(),  
    puntos_ocio=sum(puntos_ocio),  
    puntos_comerciales=sum(puntos_comerciales),  
    puntos_seguridad=sum(puntos_seguridad),  
    puntos_organismos=sum(puntos_organismos),  
    puntos_educacion=sum(puntos_educacion),  
    puntos_turismo=sum(puntos_turismo),  
    puntos_transporte=sum(puntos_transporte),  
    puntos_salud=sum(puntos_salud),  
    transito_alto=sum(transito_alto),  
    transito_medio=sum(transito_medio),  
    transito_bajo=sum(transito_bajo),  
    hombres=sum(hombres),  
    mujeres=sum(mujeres),  
    edad_0_14=sum(edad_0_14),  
    edad_15_29=sum(edad_15_29),  
    edad_30_44=sum(edad_30_44),  
    edad_45_64=sum(edad_45_64),  
    edad_65_79=sum(edad_65_79),  
    edad_mayor_80=sum(edad_mayor_80),  
    renta_anual_media=mean(renta_anual_media),  
    alimentacion=sum(alimentacion),  
    apuestas=sum(apuestas),  
    gestoria=sum(gestoria),  
    bar_restaurante=sum(bar_restaurante),  
    bazar=sum(bazar),  
    estetica=sum(estetica),  
    electrodomesticos=sum(electrodomesticos),  
    fisioterapia=sum(fisioterapia),  
    floristeria=sum(floristeria),  
    fotografia=sum(fotografia),  
    fruteria=sum(fruteria),  
    inmobiliaria=sum(inmobiliaria),  
    joyeria=sum(joyeria),  
    jugueteria=sum(jugueteria),  
    libreria=sum(libreria),  
    muebles=sum(muebles),  
    optica=sum(optica),  
    panaderia=sum(panaderia),  
    papeleria=sum(papeleria),  
    peluqueria=sum(peluqueria),  
    perfumeria=sum(perfumeria),  
    ropa=sum(ropa),  
    taller=sum(taller),  
    tintoreria=sum(tintoreria),  
    veterinario=sum(veterinario),
```

```
zapateria=sum(zapateria),
carniceria=sum(carniceria),
charcuteria=sum(charcuteria),
drogueria=sum(drogueria),
electronica=sum(electronica),
ferreteria=sum(ferreteria),
limpieza=sum(limpieza),
almacen=sum(almacen),
pescaderia=sum(pescaderia),
locutorio=sum(locutorio),
competidores=sum(competidores),
densidad_pob=mean(densidad_pob),
tasa_paro=mean(tasa_paro))
```

```
write.table(dataset_papeleria, file = "dataset_papeleria.csv", sep=';', dec = ',', row.names =
FALSE, col.names = TRUE)
```

```
dataset_optica <-dataset_final %>%
  filter(actividad == "OPTICA") %>%
  group_by(codigo_postal) %>%
  summarize(num_opticas=n(),
    puntos_ocio=sum(puntos_ocio),
    puntos_comerciales=sum(puntos_comerciales),
    puntos_seguridad=sum(puntos_seguridad),
    puntos_organismos=sum(puntos_organismos),
    puntos_educacion=sum(puntos_educacion),
    puntos_turismo=sum(puntos_turismo),
    puntos_transporte=sum(puntos_transporte),
    puntos_salud=sum(puntos_salud),
    transito_alto=sum(transito_alto),
    transito_medio=sum(transito_medio),
    transito_bajo=sum(transito_bajo),
    hombres=sum(hombres),
    mujeres=sum(mujeres),
    edad_0_14=sum(edad_0_14),
    edad_15_29=sum(edad_15_29),
    edad_30_44=sum(edad_30_44),
    edad_45_64=sum(edad_45_64),
    edad_65_79=sum(edad_65_79),
    edad_mayor_80=sum(edad_mayor_80),
    renta_anual_media=mean(renta_anual_media),
    alimentacion=sum(alimentacion),
    apuestas=sum(apuestas),
    gestoria=sum(gestoria),
    bar_restaurante=sum(bar_restaurante),
    bazar=sum(bazar),
    estetica=sum(estetica),
    electrodomesticos=sum(electrodomesticos),
    fisioterapia=sum(fisioterapia),
    floristeria=sum(floristeria),
    fotografia=sum(fotografia),
    fruteria=sum(fruteria),
```

```

inmobiliaria=sum(inmobiliaria),
joyeria=sum(joyeria),
jugueteria=sum(jugueteria),
libreria=sum(libreria),
muebles=sum(muebles),
optica=sum(optica),
panaderia=sum(panaderia),
papeleria=sum(papeleria),
peluqueria=sum(peluqueria),
perfumeria=sum(perfumeria),
ropa=sum(ropa),
taller=sum(taller),
tintoreria=sum(tintoreria),
veterinario=sum(veterinario),
zapateria=sum(zapateria),
carniceria=sum(carniceria),
charcuteria=sum(charcuteria),
drogueria=sum(drogueria),
electronica=sum(electronica),
ferreteria=sum(ferreteria),
limpieza=sum(limpieza),
almacen=sum(almacen),
pescaderia=sum(pescaderia),
locutorio=sum(locutorio),
competidores=sum(competidores),
densidad_pob=mean(densidad_pob),
tasa_paro=mean(tasa_paro))

```

```

write.table(dataset_optica, file = "dataset_optica.csv", sep=';', dec = ',', row.names = FALSE,
col.names = TRUE)

```

```

dataset_carniceria <-dataset_final %>%
  filter(actividad == "CARNICERIA") %>%
  group_by(codigo_postal) %>%
  summarize(num_carniceria=n(),
    puntos_ocio=sum(puntos_ocio),
    puntos_comerciales=sum(puntos_comerciales),
    puntos_seguridad=sum(puntos_seguridad),
    puntos_organismos=sum(puntos_organismos),
    puntos_educacion=sum(puntos_educacion),
    puntos_turismo=sum(puntos_turismo),
    puntos_transporte=sum(puntos_transporte),
    puntos_salud=sum(puntos_salud),
    transito_alto=sum(transito_alto),
    transito_medio=sum(transito_medio),
    transito_bajo=sum(transito_bajo),
    hombres=sum(hombres),
    mujeres=sum(mujeres),
    edad_0_14=sum(edad_0_14),
    edad_15_29=sum(edad_15_29),
    edad_30_44=sum(edad_30_44),
    edad_45_64=sum(edad_45_64),

```

```

edad_65_79=sum(edad_65_79),
edad_mayor_80=sum(edad_mayor_80),
renta_anual_media=mean(renta_anual_media),
alimentacion=sum(alimentacion),
apuestas=sum(apuestas),
gestoria=sum(gestoria),
bar_restaurante=sum(bar_restaurante),
bazar=sum(bazar),
estetica=sum(estetica),
electrodomesticos=sum(electrodomesticos),
fisioterapia=sum(fisioterapia),
floristeria=sum(floristeria),
fotografia=sum(fotografia),
fruteria=sum(fruteria),
inmobiliaria=sum(inmobiliaria),
joyeria=sum(joyeria),
jugueteria=sum(jugueteria),
libreria=sum(libreria),
muebles=sum(muebles),
optica=sum(optica),
panaderia=sum(panaderia),
papeleria=sum(papeleria),
peluqueria=sum(peluqueria),
perfumeria=sum(perfumeria),
ropa=sum(ropa),
taller=sum(taller),
tintoreria=sum(tintoreria),
veterinario=sum(veterinario),
zapateria=sum(zapateria),
carniceria=sum(carniceria),
charcuteria=sum(charcuteria),
drogueria=sum(drogueria),
electronica=sum(electronica),
ferreteria=sum(ferreteria),
limpieza=sum(limpieza),
almacen=sum(almacen),
pescaderia=sum(pescaderia),
locutorio=sum(locutorio),
competidores=sum(competidores),
densidad_pob=mean(densidad_pob),
tasa_paro=mean(tasa_paro))

```

```

write.table(dataset_carniceria, file = "dataset_carniceria.csv", sep=';', dec = ',', row.names =
FALSE, col.names = TRUE)

```

#Agrupamos nuestro dataset por sección censal para lanzar el clustering

```

dataset_sccc <-dataset_final %>%
  group_by(id_sccc) %>%
  summarize(puntos_ocio= mean(puntos_ocio),
            puntos_comerciales= mean(puntos_comerciales),

```

```
puntos_seguridad= mean(puntos_seguridad),
puntos_organismos= mean(puntos_organismos),
puntos_educacion= mean(puntos_educacion),
puntos_turismo= mean(puntos_turismo),
puntos_transporte= mean(puntos_transporte),
puntos_salud= mean(puntos_salud),
transito_alto= mean(transito_alto),
transito_medio= mean(transito_medio),
transito_bajo= mean(transito_bajo),
superficie= mean(superficie),
poblacion= mean(poblacion),
hombres= mean(hombres),
mujeres= mean(mujeres),
espanyoles= mean(espanyoles),
extranjeros= mean(extranjeros),
edad_0_14= mean(edad_0_14),
edad_15_29= mean(edad_15_29),
edad_30_44= mean(edad_30_44),
edad_45_64= mean(edad_45_64),
edad_65_79= mean(edad_65_79),
edad_mayor_80= mean(edad_mayor_80),
renta_anual_media= mean(renta_anual_media),
sin_estudios= mean(sin_estudios),
con_estudios= mean(con_estudios),
parados= mean(alimentacion),
alimentacion= mean(alimentacion),
apuestas= mean(apuestas),
gestoria= mean(gestoria),
bar_restaurante= mean(bar_restaurante),
bazar= mean(bazar),
estetica= mean(estetica),
electrodomesticos= mean(electrodomesticos),
fisioterapia= mean(fisioterapia),
floristeria= mean(floristeria),
fotografia= mean(fotografia),
fruteria= mean(fruteria),
inmobiliaria= mean(inmobiliaria),
joyeria= mean(joyeria),
jugueteria= mean(jugueteria),
libreria= mean(libreria),
muebles= mean(muebles),
optica= mean(optica),
panaderia= mean(panaderia),
papeleria= mean(papeleria),
peluqueria= mean(peluqueria),
perfumeria= mean(perfumeria),
ropa= mean(ropa),
taller= mean(taller),
tintoreria= mean(tintoreria),
veterinario= mean(veterinario),
zapateria= mean(zapateria),
carniceria= mean(carniceria),
```

```
charcuteria= mean(charcuteria),
drogueria= mean(drogueria),
electronica= mean(electronica),
ferreteria= mean(ferreteria),
limpieza= mean(limpieza),
almacen= mean(almacen),
pescaderia= mean(pescaderia),
locutorio= mean(locutorio),
competidores= mean(competidores),
prt_edad_0_14= round(mean(prt_edad_0_14),2),
prt_edad_15_29= round(mean(prt_edad_15_29),2),
prt_edad_30_44= round(mean(prt_edad_30_44),2),
prt_edad_45_64= round(mean(prt_edad_45_64),2),
prt_edad_65_79= round(mean(prt_edad_65_79),2),
prt_edad_mayor_80= round(mean(prt_edad_mayor_80),2),
prt_espanyoles= round(mean(prt_espanyoles),2),
prt_extranjeros= round(mean(prt_extranjeros),2),
prt_hombress= round(mean(prt_hombress),2),
prt_mujeres= round(mean(prt_mujeres),2),
densidad_poblacion= mean(densidad_pob),
tasa_paro= mean(tasa_paro))
```

#le agrego barrio y distrito con esta tabla ra_sccc_x_barrio_x_distrito

```
eg14=dbSendQuery(tmydb,"select * from ra_sccc_x_barrio_x_distrito_eg")
barrio_distrito=fetch(eg14,n=-1)
```

```
dataset_sccc_2<-dataset_sccc
```

```
dataset_sccc_2<-mutate(dataset_sccc_2,
  Distrito=substr(id_sccc,start=6,stop=7))
```

```
dataset_sccc_2<-mutate(dataset_sccc_2,
  Seccion = substr(id_sccc,start=8,stop=10))
```

```
dataset_sccc_3= merge (dataset_sccc_2,barrio_distrito,by=c("Distrito","Seccion"),all.x=TRUE)
```

#exportamos tabla a MySQL pero antes en excel reemplazamos la coma por el punto en todas las columnas

```
write.table(dataset_sccc_3, file = "dataset_sccc_distrito_ok.csv", sep=';', dec = '.', row.names =
FALSE, col.names = TRUE)
```

3. Importancia de variables y correlaciones

Con el objetivo de seleccionar las variables finales a incluir para lanzar los clusters, se realizan análisis de importancia de variables utilizando el paquete H2O (basado en Random Forest) así como Boruta y matriz correlación entre las diferentes actividades.

Seleccionaremos variables lo más representativas para cada actividad atendiendo al resultado de este análisis de importancia de variables, análisis de correlación y nuestro conocimiento sobre cómo se relacionan los comercios en el Municipio de Madrid.

El método utilizado a continuación se realizará para las diferentes actividades a analizar.

Se instala paquete H2O

```
install.packages("h2o")
library("h2o")
h2o.init() # Arranca la máquina virtual
```

Carga de los datos

```
data <- h2o.importFile("C:/Users/nturega/Desktop/dataset_joyerias.csv")

h2o.describe(data)
```

Sobre csv joyerias agrupadas por código postal

Carga de los datos

```
data <- h2o.importFile("C:/Users/nturega/Desktop/dataset_joyerias.csv")

h2o.describe(data)
```

Se pasan a numéricas las variables que ha cargado como string

```
data$renta_anual_media <- as.numeric(data$renta_anual_media)
data$parados <- as.numeric(data$parados)
data$densidad_pob <- as.numeric(data$densidad_pob)
data$new <- as.numeric(data$new)
```

Se comprueba que los cambios se han realizado correctamente

```
h2o.describe(data)
```

Se asigna predictores y variable respuesta (y).

```
y <- "num_joyerias"
x <- setdiff(names(data), c(y, "joyeria")) # omitir misma actividad, info ya en vble.
competidores.
print(x)
```

```
rf_fit <- h2o.randomForest(x = x,
                           y = y,
                           training_frame = data,
                           ntrees = 1000,
                           seed = 1234)
rf_fit@model$summary # vemos un resumen de rf_fit
```



```
# representar el top de variables según importancia

h2o.varimp_plot(rf_fit)

# Importancia de variables

library(readr)
install.packages("randomForest")
library(randomForest)
install.packages("Boruta")
library(Boruta)

##### JOYERIA #####

joyeria <- read_delim("C:/Users/nturega/Desktop/dataset_joyerias_cp_boruta.csv",
                    ";", escape_double = FALSE, trim_ws = TRUE)

# Utilizando random forest

model <- randomForest(num_joyerias ~ ., data=joyeria, ntree=1000)

importance(model)
varImpPlot(model)

# Utilizando el metodo Boruta

# Creamos una columna con datos aleatorios para comprobar que algoritmo detecte que no es
importante
joyeria$new <- runif(nrow(joyeria),0,1)

boruta_output <- Boruta(num_joyerias ~ ., data=joyeria, doTrace=2, maxRuns = 50) # Metodo
Boruta
boruta_signif<- names(boruta_output$finalDecision[boruta_output$finalDecision %in%
c("Confirmed", "Tentative")])

boruta_output$finalDecision

print(boruta_signif) # significant variables

plot(boruta_output, las=2, xlab="", main="Variable Importance") # Plot de la importancia de
variables

# Se importa la matriz_actividades

data <- read_delim("C:/Users/nturega/Desktop/matriz_actividades.csv",
                  ";", escape_double = FALSE, trim_ws = TRUE)

# Valido que la carga de los paquetes ha sido correcta

library(moments)
library(dplyr)
library(ggplot2)
```

```
library(corrplot)

# Muestro las primeras filas del fichero

data %>% head()

# Compruebo los tipos de las variables

str(data)

# Obtengo los estadísticos más importantes

summary(data)

# Matriz de correlaciones

cor_matrix <- cor(data)
round(cor_matrix, digits = 2)

# La dibujamos:

corrplot(cor_matrix, type="upper")
```

4. Código de ejecución del modelo Clustering

```
library(RMySQL)
library(VIM)
library(ggmap)
library(rgdal)
library(rgeos)
library(purrr)
library(factoextra)
library(NbClust)
library(proj4)

source("Prd Env - R Scripts/functions.R")

mydb = dbConnect(MySQL(), user='geopymes_workspace', password='geoeoi07$',
dbname='geopymes', host='34.72.143.222')
rs_censos_locales_final=dbSendQuery(mydb,"select * from eg_dataset_proyecto_final_sccc_3")
data_censos_locales_final=dbFetch(rs_censos_locales_final,n=-1)

new_df<-data.frame()

columns_to_use<-c("puntos_salud","panaderia","ropa","competidores","perfumeria","zapateria","puntos_comerciales")

new_df<-data.frame('JOYERIA',columns_to_use)
colnames(new_df)<-c('actividad','most_important_features')
```

```
columns_to_use<-c("puntos_salud","papeleria","edad_65_79","ferreteria","puntos_educacion","muebles")
```

```
new_df1<-data.frame('OPTICA',columns_to_use)
colnames(new_df1)<-c('actividad','most_important_features')
new_df<-rbind(new_df,new_df1)
```

```
columns_to_use<-c("edad_65_79","mujeres","puntos_educacion","fruteria","peluqueria","competidores")
```

```
new_df1<-data.frame('CARNICERIA',columns_to_use)
colnames(new_df1)<-c('actividad','most_important_features')
new_df<-rbind(new_df,new_df1)
```

```
columns_to_use<-c("edad_45_64","edad_30_44","puntos_educacion","peluqueria","fruteria","alimentacion")
```

```
new_df1<-data.frame('PAPELERIA',columns_to_use)
colnames(new_df1)<-c('actividad','most_important_features')
new_df<-rbind(new_df,new_df1)
```

```
data_censos_locales_final_for_clustering<-data_censos_locales_final
columns_to_use<-new_df[new_df$actividad=='OPTICA',c('most_important_features')]
```

```
data_censos_locales_final_for_clustering_scaled<-scale(data_censos_locales_final_for_clustering[,columns_to_use])
data_censos_locales_final_for_clustering_scaled<-as.data.frame(data_censos_locales_final_for_clustering_scaled)
```

```
#determine the optimal cluster's number
#-----#
```

```
# Elbow method
fviz_nbclust(data_censos_locales_final_for_clustering_scaled, kmeans, method = "wss") +
  geom_vline(xintercept = 5, linetype = 2)+
  labs(subtitle = "Elbow method")
```

```
# Silhouette method
fviz_nbclust(data_censos_locales_final_for_clustering_scaled, kmeans, method = "silhouette")+
  labs(subtitle = "Silhouette method")
```

```
MyKmeans <- function(x,k) list(cluster=kmeans(x, k, iter.max=50))
```

```
set.seed(123)
fviz_nbclust(data_censos_locales_final_for_clustering_scaled, kmeans, nstart = 25, method = "gap_stat", nboot = 50)+
  labs(subtitle = "Gap statistic method")
```

```
fviz_nbclust(data_censos_locales_final_for_clustering_scaled, MyKmeans, method = "gap_stat", nboot = 50)+
  labs(subtitle = "Gap statistic method")
```

```
#-----#
```

```
clusters <- kmeans(data_censos_locales_final_for_clustering_scaled,3)

data_censos_locales_final_for_clustering_scaled$cluster <- as.factor(clusters$cluster)
data_censos_locales_final_for_clustering$cluster <- as.factor(clusters$cluster)
data_censos_locales_final_for_clustering$actividad<-'JOYERIA'

dbWriteTable(mydb, value = data_censos_locales_final_for_clustering, name =
'ra_sscs_final_cleaned_clusters', append = TRUE )
```

5. Función para identificación de Puntos de Interés en un radio de 250 metros de cada local

```
f_set_poi_censo_locales <- function(query1, query2, table_name, x) {

  f_mydb = dbConnect(MySQL(), user='geopymes_workspace', password='geoeoi07$',
dbname='geopymes', host='34.72.143.222')

  #f_rs=dbSendQuery(f_mydb,"select * from eg_poi_censo_locales WHERE
convert(substr(cmun,1),SIGNED)=79 OR convert(substr(cmun,1),SIGNED)=796 limit 10000")
  f_rs=dbSendQuery(f_mydb,query1)
  f_data_poi=fetch(f_rs,n=-1)

  EPSG <- make_EPSG()
  EPSG[grepl("Spain$", EPSG$note), ]

  coordinates(f_data_poi)<--utm_x+utm_y

  proj4string(f_data_poi) <- CRS("+init=epsg:23030")

  f_data_poi <- spTransform(f_data_poi,proj4string(f_data_poi))

  #Se filtra el tipo de acceso a Puerta Calle debido a que solo esos locales tienen las coordenadas
correctas
  f_rs_censos_locales=dbSendQuery(f_mydb,query2)
  x=fetch(f_rs_censos_locales,n=-1)

  data_censos_locales_new=x

  data_censos_locales_new1<-x[x$coordenada_x_local!='0',]

  new_df3<-data.frame(IdLocal=character(0),Tipo_Negocio=character(0),Cantidad=integer(0))

  EPSG <- make_EPSG()
  EPSG[grepl("Spain$", EPSG$note), ]

  for (i in 1:nrow(data_censos_locales_new1)) {
```

```

new_row<-data_censos_locales_new1[i,]

coordinates(new_row)<--coordenada_x_local+coordenada_y_local

proj4string(new_row) <- CRS("+init=epsg:23030")

i_new_row <- spTransform(new_row,proj4string(new_row))

i_inf_new_row <- gBuffer(spgeom = i_new_row, width = 250)

# new_df1<-as.data.frame(table(f_data_poi[i_inf_new_row,]@data$tipo_negocio))

new_df1<-as.data.frame(table(f_data_poi[f_data_poi$id_local!=new_row$id_local,][i_inf_new_r
ow,]@data$tipo_negocio))

if (nrow(new_df1)>0) {

  new_df1$idlocal<-new_row$id_local
  colnames(new_df1)<-c('Tipo_Negocio','Cantidad','IdLocal')
  new_df2<-new_df1[,c('IdLocal','Tipo_Negocio','Cantidad')]
  new_df3 <- rbind(new_df3, new_df2)
  if (mod(i,1000)==0) {

    print(i)
  }
}

new_df4<-new_df3
new_df5<-reshape(new_df4, direction = "wide", idvar = "IdLocal", timevar = "Tipo_Negocio")

#colocar en cero todos los valores NAs
new_df5[is.na(new_df5)] = 0

mydb1 = dbConnect(MySQL(), user='geopymes_workspace', password='geoeoi07$',
dbname='geopymes', host='34.72.143.222')

#escribir resultado en tabla de MySQL
dbWriteTable(mydb1, value = new_df5, name = table_name, append = TRUE )

x<-new_df5

dbDisconnect(f_mydb)
dbDisconnect(mydb1)

rm(f_mydb)
rm(f_rs)
rm(f_data_poi)
rm(f_rs_censos_locales)
rm(data_censos_locales_new)
rm(data_censos_locales_new1)

```

```
rm(new_df3)
rm(new_row)
rm(i_new_row)
rm(i_inf_new_row)
rm(new_df1)
rm(new_df2)
rm(new_df4)
rm(new_df5)
rm(mydb1)

return(x)

}
```

6. Código para identificar la Sección Censal donde se encuentra ubicado un Negocio o Local

```
library(RMySQL)
if(!require('rgdal')) install.packages('rgdal')
library(rgdal)
if(!require('rgeos')) install.packages('rgeos')
library(rgeos)
if(!require('osrm')) install.packages('osrm')
library(osrm)
if(!require('geosphere')) install.packages('geosphere')
library(geosphere)
library(dplyr)
library(magrittr)
library(tidyr)

i_data_sccc <- readOGR(dsn = "C:\\Users\\Rafa\\Documents\\Master en Business Intelligence y Big
Data\\Fase Proyecto\\Shape Files\\cartografia_censo2011_nacional", layer =
"SECC_CPV_E_20111101_01_R_INE")

mydb = dbConnect(MySQL(), user='geopymes_workspace', password='geoeoi07$',
dbname='geopymes', host='34.72.143.222')

rs_censos_locales=dbSendQuery(mydb,"select * from ra_censo_locales_new where
desc_tipo_acceso_local='Puerta Calle'")
data_censos_locales=fetch(rs_censos_locales,n=-1)

data_censos_locales_new=data_censos_locales

EPSG <- make_EPSG()
EPSG[grepl("Spain$", EPSG$note), ]

data_censos_locales_new<-data_censos_locales_new[data_censos_locales_new$coordenada_x_lo
cal!='0',]

coordinates(data_censos_locales_new)<--coordenada_x_local+coordenada_y_local
```

```
proj4string(data_censos_locales_new) <- CRS("+init=epsg:23030")

i_data_censos_locales_new <- spTransform(data_censos_locales_new,proj4string(i_data_sccc))

a.data<-over(i_data_censos_locales_new,i_data_sccc[, "CUSEC"])

data_censos_locales_new$CUSEC<-a.data$CUSEC

censos_locales_con_SCCC<-data_censos_locales_new@data

dbWriteTable(mydb, value = censos_locales_con_SCCC, name = "ra_censo_locales_x_sccc",
append = TRUE )
```

7. Construcción de tabla consolidada al cargar archivos geoespaciales (.shp) de los puntos de interés desde el portal Madrid.org

```
CREATE TABLE ra_poi AS
select 'Gasolineras' tipo_negocio, cmun,nombre,utm_x,utm_y from ra_gasolineras union all
SELECT 'Farmacias',suca_cmun3,busca,utm_x,utm_y from ra_farmacias union ALL
SELECT 'Hoteles',cmun,etiqueta,utm_x,utm_y from ra_hoteles1 union all
select 'Hoteles',cmun,etiqueta,utm_x,utm_y from ra_hoteles2 union all
select 'Hoteles',cmun,etiqueta,utm_x,utm_y from ra_hoteles3 union all
select 'Hoteles',cmun,etiqueta,utm_x,utm_y from ra_hoteles4 union all
select 'Hoteles',cmun,etiqueta,utm_x,utm_y from ra_hoteles5 union all
select 'libreriaao', cmun,literal,utm_x,utm_y from ra_libreriaao union all
select 'librerias', cmun,nombre,utm_x,utm_y from ra_librerias union all
select 'teatros', codmun,nombre,utm_x,utm_y from ra_teatros union all
select 'puentes', cmun,etiqueta,utm_x,utm_y from ra_puentes union ALL
select 'parroquias', cmu,nombre,utm_x,utm_y from ra_parroquias union all
select 'parques_tematicos', cmun,etiqueta,utm_x,utm_y from ra_parques_tematicos union all
select 'museos', cmun,nombre,utm_x,utm_y from ra_museos union all
select 'cines', cmun,busca,utm_x,utm_y from ra_cines union all
select 'cementerios', cmun,nombre,utm_x,utm_y from ra_cementerios union all
select 'comerespecial', cmun,etiqueta,utm_x,utm_y from ra_comerespecial union ALL
select 'mercado_abastos', cmun,denominaci,utm_x,utm_y from ra_mercado_abastos union all
select 'mercadillos', cmun,nombre,utm_x,utm_y from ra_mercadillos union all
select 'hipermercados', cmun,eti,utm_x,utm_y from ra_hipermercados union ALL
select 'centros_comerciales', cmun,etiqueta,utm_x,utm_y from ra_centros_comerciales union all
select 'juzgados', cmun,nombre,utm_x,utm_y from ra_juzgados union all
select 'correos', cmun,busca,utm_x,utm_y from ra_correos union ALL
select 'estcerca', cmun,busca,utm_x,utm_y from ra_estcerca union ALL
select 'helisuper', cmun,etiqueta,utm_x,utm_y from ra_helisuper union all
select 'itv', cmun,etiqueta,utm_x,utm_y from ra_itv union all
select 'metro_bocas', cod_muni,salida,utm_x,utm_y from ra_metro_bocas union ALL
select 'parking', cmun,busca,utm_x,utm_y from ra_parking union all
select 'oficinas_turismo', cmun,nombre,utm_x,utm_y from ra_oficinas_turismo union all
select 'bomberos', cmun,etiqueta,utm_x,utm_y from ra_bomberos union all
select 'policia_nacional', cmun,nombre,utm_x,utm_y from ra_policia_nacional union all
select 'guardia_civil', cmun,etiqueta,utm_x,utm_y from ra_guardia_civil union all
select 'policia_local', cmun,etiqueta,utm_x,utm_y from ra_policia_local union all
```



```
select 'cruz_roja', cmun,etiqueta,utm_x,utm_y from ra_cruz_roja union all  
select 'centros_salud', codmun,nombre,utm_x,utm_y from ra_centros_salud union all  
select 'hospitales', cmun,centro_nom,utm_x,utm_y from ra_hospitales union all  
select 'educapu', cmuni,etiqueta,utm_x,utm_y from ra_educapu union all  
select 'educapr', cmun4,ds_largo,utm_x,utm_y from ra_educapr union all  
select 'educase', cmuni,busca,utm_x,utm_y from ra_educase union all  
select 'univers', cmun4,etiqueta,utm_x,utm_y from ra_univers union all  
select 'saludotros', codmun,nombre,utm_x,utm_y from ra_saludotros union all  
select 'albergues', cmun,nombre,utm_x,utm_y from ra_albergues;
```