Facultad de Ciencias Económicas y Empresariales.

Grado en Administración y Dirección de Empresas.

Trabajo Fin de Grado

Los datos y su análisis como aspecto fundamental para revalorizar las empresas familiares

Un estudio práctico sobre las oportunidades que brindan las nuevas herramientas de análisis de Big Data

Clave: **201504697**

MADRID | Junio, 2020

RESUMEN

Este trabajo de investigación tiene como principal objetivo demostrar la eficiencia y las utilidades que tienen el uso de innovadoras herramientas de análisis de Big Data, pretendiendo demostrar el gran valor que pueden aportar a múltiples departamentos dentro de las entidades correspondientes: desde el proceso productivo hasta la venta final del producto o servicio, pasando por todas las fases de la cadena de valor.

Como consecuencia de la situación económica de total incertidumbre que vivimos actualmente y que afecta con especial dureza a las pequeñas y medianas empresas, el foco del estudio se pone en esas empresas familiares que conforman el principal motor y pulmón de la economía española. Así, se pretende dar respuesta a esta pregunta: ¿Pueden estás técnicas implementarse con éxito en empresas que, a priori, tienen muy difícil competir contra las gigantes multinacionales?

Una vez analizadas las potenciales ventajas que pueden aportar el uso de estas herramientas, se examina si su implementación sería factible en las PYMES, teniendo muy en cuenta los posibles retos y limitaciones que presentan estas herramientas actualmente. Para ello, se realiza un análisis empírico de los datos cedidos por una empresa familiar de tamaño mediano para demostrar cómo este tipo de análisis puede generar un inmenso valor.

Palabras clave: *Big Data*, análisis de datos, empresa familiar, Rstudio, análisis exploratorio, análisis de conglomerados, modelos predictivos.

ABSTRACT

The main objective of this research work is to demonstrate the efficiency and benefits of the use of innovative Big Data analysis tools, trying to show the great value they can bring to multiple departments within the corresponding entities: from the production process to the final sale of the product or service, through all the phases of the value chain.

As a consequence of the current economic situation of total uncertainty that affects small and medium sized companies with particular severity, the focus of the study is placed on those family businesses that make up the main engine and lung of the Spanish economy. The aim is to answer this question: Can these techniques be successfully implemented in companies that, a priori, find it very difficult to compete against the multinational giants?

Once the potential advantages of using these tools have been analysed, it is examined whether their implementation would be feasible in SMEs, taking into account the possible challenges and limitations that these tools currently present. To this end, an empirical analysis of the data provided by a medium-sized family business is carried out to demonstrate how this type of analysis can generate immense value.

Keywords: Big Data, data analysis, family business, Rstudio, exploratory analysis, cluster analysis, predictive models.

“La adaptación parece ser, fundamentalmente, un proceso de relocalización de nuestra atención”

Daniel Kahneman

Premio Nobel Conmemorativo de Economía en el año 2002

ÍNDICE DE CONTENIDOS

ÍNDICE DE ACRÓNIMOS

API	Application Programming Onterface
CEO	Chief Executive Officer
EDA	Exploratoy Data Analysis
GFS	Google File System
ISO	International Organization for Standardization
NA	Not available (utilizado en Rstudio)
PYME	Pequeña Y Mediana Empresa
RAE	Real Academia Española
SAS	Statistical Analysis System
SQL	Structured Query Language
WCSS	Within Cluster Sum of Square

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

1. Introducción.

1.1 Objetivo.

Este trabajo de investigación tiene como principal objetivo demostrar la eficiencia y las utilidades que tienen el uso de innovadoras herramientas de análisis de Big Data, pretendiendo demostrar el gran valor que pueden aportar a múltiples departamentos dentro de las entidades correspondientes: desde el proceso productivo hasta la venta final del producto o servicio, pasando por todas las fases de la cadena de valor.

Y no sólo en las grandes empresas, sino de modo particular en las PYMES, que conforman más del 95% del total de empresas en España y son un importante motor de la economía española. Así, trataremos de dar respuesta a esta pregunta: ¿Pueden estás técnicas implementarse con éxito en empresas que, a priori, tienen muy difícil competir contra las gigantes multinacionales? Una vez analizadas las potenciales ventajas que pueden aportar el uso de estas herramientas, será el momento de examinar si su implementación sería factible en las PYMES. Para ello, realizaremos un análisis empírico de los datos cedidos por una empresa familiar de tamaño mediano para demostrar cómo este tipo de análisis puede generar un inmenso valor no solo en empresas de gran tamaño sino en cualquier empresa, en particular las empresas familiares de pequeño y mediano tamaño.

1.2 Justificación de tema objeto de investigación.

El nuevo paradigma tecnológico avanza imparable en todos los ámbitos profesionales y personales. Ante una realidad cada vez más compleja y donde la competencia es feroz, las empresas familiares resisten frente a las grandes multinacionales gracias a los esfuerzos de cada uno de los individuos que las conforman. Sin embargo, no se me ocurre mejor momento -dada la actual crisis que asola el panorama internacional- para demostrar que el uso de los avances tecnológicos puede suponer una ventaja competitiva no solo para las gigantes multinacionales, sino también para aquellas empresas familiares que necesitan reinventarse y optimizar procesos para sobrevivir. Esta será la única forma mediante la cual este tipo de negocios logrará consolidar su posición dentro del mercado nacional como el grueso fundamental del tejido empresarial español.

1.3 Metodología.

La metodología que se utilizará en la realización de este trabajo está basada en las técnicas de análisis de Big Data. En este sentido, se llevarán a cabo los análisis exploratorios, de conglomerados y predictivos. El análisis predictivo, probablemente el que genera mayor valor añadido, permite agrupar una variedad de técnicas estadísticas de modelización, aprendizaje automático y minería de datos que analiza los datos actuales e históricos reales para hacer predicciones acerca del futuro o acontecimientos no conocidos.

La utilización de herramientas de programación, principalmente Rstudio, servirá como base para el correcto tratamiento de los datos que se pretenden utilizar, de tal forma que podamos limpiarlos, corregirlos, explorarlos, modelizarlos y analizarlos. Asimismo, otras herramientas que se utilizarán para obtener soluciones a los problemas de negocio propuestos son Microsoft Azure, para entrenar los modelos, mediante complejos algoritmos, y conseguir mejorar su eficiencia y Orange, para realizar un análisis de conglomerados sobre el mapa físico de la Comunidad de Madrid.

1.4 Estructura.

La estructura de este trabajo se dividirá en tres partes. En primer lugar, trataremos los aspectos más teóricos e históricos del análisis de Big Data estableciendo el marco teórico del presente trabajo. En segundo lugar, realizaremos un análisis empírico de los datos cedidos por una empresa de tamaño mediano para demostrar cómo este tipo de análisis puede generar un inmenso valor no solo en empresas de gran tamaño sino en cualquier empresa. Para terminar, reuniremos las conclusiones más relevantes obtenidas a través del estudio y realizaremos la valoración práctica de este último.

2 Fundamentos teóricos.

2.1 Big Data

2.1.1 ¿Qué es?

La primera acepción que la RAE otorga a dato es "*la información sobre algo concreto que permite su conocimiento exacto o sirve para deducir las consecuencias derivadas de un hecho".* Para entender la génesis de este concepto hay, sin embargo, que remontarse a 2003. Google publicó un documento en el que explicaban el entonces revolucionario Sistema de Ficheros Distribuidos (GFS, por sus siglas en inglés) y, por primera vez, se mencionaba un término que años más tarde revolucionaría todos los sectores y generaría un nuevo paradigma: Big Data (Ghemawat, Gobioff, & Leung, 2003).

El origen de este término resulta revelador *per se* puesto que proviene de una compañía, no tratándose de un concepto académico o intelectual. La iniciativa privada en aras de generar la mayor rentabilidad posible, objetivo que siempre ha perseguido, se percata ya casi hace 20 años de que el almacenamiento y el análisis de un gran volumen de datos puede aportar numerosos beneficios a la actividad empresarial.

Para aportar rigor a esta investigación es preciso intentar definir un concepto tan complejo y diverso como es Big Data (Aguilar, 2013). Partiendo de la premisa de que no existe una definición que pudiera contentar a todos los expertos en la materia, desde un punto de vista técnico podría decirse que Big Data se refiere "*al conjunto de datos o combinaciones de conjuntos de datos cuyo tamaño (volumen), complejidad (variabilidad) y velocidad de crecimiento dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales"* (PowerData, 2017).

El universo de los datos es un ecosistema no solo complejo sino en constante cambio pero que también aporta un gran valor a las empresas u organizaciones que deciden utilizarlos. Precisamente, no son pocos los que defienden que Big Data es y se consolidará a lo largo del siglo como la *commodity* que marcará la diferencia y, en último térrmino, se convertirá en el nuevo petróleo, la mercancía más codiciada del plantea (Marcireau, 2019).

Finalmente, para comprender verdaderamente el concepto de Big Data es necesario entender las notas características más importantes de este. Para ello, hay quien describe este concepto en torno a las 3V's (Russom, 2011). Sin embargo, esas notas se quedan actualmente cortas para definir un fenómeno tan complejo, por ello utilizaremos las 5V's para explicar el concepto (BBVA, 2017):

- **Volumen:** La cantidad de datos que se generan año tras año aumenta de forma exponencial con el paso del tiempo. Esto provoca que sea extremadamente difícil la obtención, limpieza, estructuración de los datos. En muchas ocasiones, se necesita demasiado tiempo para transformar los tipos no estructurados (documentos, vídeos o audios) en tipos estructurados y procesar esos datos. Es en este punto donde las nuevas herramientas que van surgiendo consiguen aportar valor, reduciendo los tiempos de espera y generando ventajas competitivas.

- **Velocidad:** Como decíamos no solo se generan muchos datos, sino que además cada vez lo hacen más rápidamente. Únicamente las empresas que consigan adecuarse a esa rápidez de generación de grandes volúmenes de información podrán utilizarlos para obtener una factor diferenciador ya que, en no pocas situaciones, su validez también se acorta en el tiempo como consecuencia de una sociedad donde los cambios se producen a un ritmo vertiginoso. La única solución, ante este problema, parece ser disponer de un poder de procesamiento mucho más alto de lo habitual, lo que implica un coste superior a corto plazo pero que, sin duda, disminuirá con el paso del tiempo.

- **Variedad:** Los datos que se generan proceden de múltiples fuentes, tipos de datos y estructuras complejas por lo que es fundamental tener en cuenta la naturaleza de nuestro negocio para poder darle prioridad a un tipo de fuente o a otro. De esta forma, se podrían integrar aquellos que sean más importantes y reducir así la dificultad de este proceso.

- **Veracidad:** La mera tenencia de datos no asegura el éxito. En este apartado hay que resaltar la importancia de obtener datos de calidad que nos permitan extraer conclusiones que sean realmente importantes. Nótese que, en la actualidad, no existen estándares de calidad de datos unificados. Si bien en 1987 la Organización Internacional de Normalización (por sus siglas en inglés, ISO) publicó las normas ISO 9000 para garantizar la calidad de productos y servicios, el estudio de los estándares de calidad de los datos no comenzó hasta los años noventa y no fue hasta 2011 cuando ISO publicó las normas de calidad de datos ISO 8000. En definitiva, y sin ánimo de examinar la actual regulación que existe en ese sentido, las normas requieren de un mayor proceso de maduración así que como diversas reformas de calado para que se adapten a las necesidades cambiantes que surgen en este campo (Gomez Carretero & Piattini Velthuis, 2018).

- **Valor:** Íntimamente relacionado con la nota anterior, supone la capacidad de generar *insights* útiles para la empresa que valor a nuestro negocio. A este respecto, no cualquier dato es válido y no cualquier conclusión es útil.

Una vez alcanzado este punto y teniendo un conocimiento mucho más amplio de lo que se entiende por Big Data es fundamental entender el alcance del crecimiento de este en la última década por lo que lo trataremos a continuación.

2.1.2 La generación de datos: una máquina imparable.

En un entorno cada vez más digital, el ser humano se ha convertido en una máquina imparable de generar de datos, los datos se han convertido en el nuevo aire porque estamos rodeados de ellos. Se prevé que, en 2020, nuestra civilización habrá generado 40 zettabytes de información, lo que equivale a 57 veces el número de granos de arena presentes en todas las playas del mundo (International Campus, 2017). Estas cantidades son tan desorbitadas que provocan incertidumbre y desconocimiento. Todos creamos información, y lo hacemos constantemente, ya sea viendo un video en YouTube, subiendo una foto a Instagram o comprando un producto en Amazon. La cantidad de datos generada por minuto supera cualquier expectativa situándonos en alrededor de 4.000 millones de interacciones por minuto (Domo Consulting, 2017).

Más impactante aún, es la revelación que hace la International Data Corporation: *"[...] la cantidad de datos generados en los últimos ocho meses supera la cantidad de todos los datos creados desde la Revolución Industrial"* (Villamizar, 2019).

Todas estas afirmaciones podrían hacernos caer en un error típico, un error que *de facto* cometen numerosos CEOs alrededor de todo el mundo, pensar que solo las grandes compañías pueden obtener, estructurar, analizar y extraer conclusiones de los conjuntos de datos que se generan diariamente en todo el mundo. Mencionaba que era un error muy común porque entre el 60% y 73% de todos los datos que obtienen las compañías no se utilizan para obtener valor de ellos (Forrester, 2016). Los costes de esto son increíblemente altos, tanto desde el punto de vista económico como, también, desde el punto de vista de la oportunidad que están perdiendo de adecuarse a las nuevas tecnologías y consecuentemente, ser más competitivos. Por lo tanto, independientemente de si eres Facebook, Alibaba o una mera panadería, probablemente estés generando un volumen de datos lo suficientemente grande para que merezca la pena tratarlos a través de las nuevas herramientas de análisis. De hecho, la Library of Congress de Washington está considerada como la mayor biblioteca del mundo. Si se digitalizara todo su fondo bibliográfico, que supera los veinte millones de libros, toda la información cabría en apenas 235 terabytes. Una cifra no muy alta, teniendo en cuenta que una empresa de mil trabajadores genera anualmente el mismo volumen de datos (Brown, Chui, & Manyika, 2011).

Ante esto, es preciso recordar que los datos en sí mismos no valen nada, es su correcto tratamiento lo que permite traducirlos en información y más tarde en conocimientos que generen soluciones o *insights* para los problemas de negocio que puedan surgir. La mejora de la toma de decisiones será el factor distintivo para solucionar los problemas del presente y del futuro (Zamora, 2016).

En cualquier caso, Big Data tiene el potencial de transformar negocios e industrias y generar una alta cantidad de valor. De tal forma que transforme cómo las compañías se organizan, las tecnologías que usan o, que construya ecosistemas totalmente nuevos con clientes y proveedores. Todo ello tiene un claro impacto en las cuentas anuales de muchos negocios que ya han implementado este tipo de estrategias (BCG, 2020). La inversión, si bien en muchos casos no debe ser excesivamente alta, es totalmente rentable y, a día de hoy, es también necesaria.

2.2 Análisis de datos

2.2.1 ¿Cómo se puede hacer? ¿Qué tipos de análisis existen?

El análisis de datos, de forma muy simplificada, es la técnica que permite el correcto tratamiento de los datos en todas sus fases (limpieza, transformación, modelación y extracción de conclusiones) para poder descubrir importantes soluciones a problemas de negocio o *insights* que puedan generar ventajas competitivas. En definitiva, esta técnica extrae conocimientos útiles que permiten a los negocios y organizaciones tomar decisiones en función de estos.

En realidad, el análisis de datos se ha realizado desde hace siglos y es una de las grandes razones que nos han permitido avanzar como conjunto permitiéndonos aprender de nuestros errores y ayudándonos a mejorar la toma de decisiones. Sin embargo, ante la perspectiva de una generación tan abundante, rápida y heterogénea de datos, las herramientas tradicionales tienen los días contados debido a la aparición de nuevas técnicas de análisis que permiten el tratamiento de un mayor volumen de datos en un tiempo mucho menor.

En la actualidad existen múltiples tipos de herramientas de este tipo, algunas de pago, pero en su gran mayoría en formato *Open Source* (Código Abierto, en español), lo que supone que muchos usuarios introducen mejoras con cierta frecuenta ayudando a la plataforma a desarrollarse y ser más competitiva, siendo su uso 100% gratuito (Lerner & Tirole, 2002). Esto genera una relación simbiótica entre los usarios de estas herramientas dada la continua innovación, actualización y mejora que se produce en la plataforma.

Los dos lenguajes de programación más usados para la ciencia de datos son R y Python. Para el análisis y comprobación que realizaremos más adelante en este trabajo utilizaremos Rstudio como herramienta principal para el tratamiento de datos junto con la ayuda de otras herramientas que se aplicarán subsidiariamente. Por supuesto que existen multitud de otras herramientas como SQL, SAS o Java que tienen una utilidad similar, si bien su uso no está tan extendido entre los *data scientists*.

A continuación, se explican los tipos de análisis de datos que se pueden realizar con las herramientas disponibles y que se someterán a examen a lo largo del estudio empírico para probrar su eficacia:

- Análisis exploratorio: se trata de un concepto que engloba técnicas muy diversas y que permite conocer la naturaleza y estructura de una colección de datos en poco tiempo. A través de este tipo de análisis se obtiene, de forma sencilla, una amplia variedad de gráficos y estadísticos diferentes que han hecho posible la aparición de una nueva filosofía en los estudios estadísticos. Este nuevo concepto fue introducido por el estadístico norteamericano John Turkey en el siglo XX (Turkey, 1977).

- Análisis de conglomerados (o *cluster analisys*): es una técnica estadística multivariante que busca agrupar elementos (o variables) tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos (Lopez, 2009). Esta clasificación es un método muy conveniente para organizar grandes volúmenes de datos y, de esa forma, poder entender la información y extraer conclusiones de gran utilidad para el negocio (Everitt, Leese, Landau, & Stahl, 2011).

- Análisis predictivo: se trata de un análisis que se basa en la creación de modelos que son entrenados mediante algoritmos a través de un conjunto de datos que permite entrenar el modelo (*train*) y otro conjunto de datos que se utiliza para su validación (*test*). De esta forma se logra predecir los resultados futuros buscando maximizar o minimizar -según el caso- las métricas de error. (Mayor, 2015)

  Si bien la herramienta Rstudio permite realizar este análisis, en este trabajo se utilizará también la herramienta Microsoft Azure, que permite realizar este tipo de análisis con eficacia y rápidez. En este sentido, otra opción que tiene un gran potencial, en cuanto a la obtención de *insights* útiles se refiere, es la realización de un análisis de *cross selling*, que también será tenido en cuenta.

2.2.2 ¿Qué puede aportar el análisis de Big Data?

Llegados a este punto es momento de examinar hasta que punto estás técnicas de análisis que requieren de novedosas herramientas para el tratamiento de Big Data pueden aportar valor a las empresas actualmente. Ahora bien, el impacto es tan grande que tan solo el paso del tiempo podrá desvelarnos su verdadero potenical (Fosso Wamba, Akter, Edwards, Chopin, & Gnanzou, 2015).

Los beneficios que puede aportar esta actividad son innumerables por lo que vamos a centrarnos en varios aspectos fundamentales:

- *Corrección y optimización de procesos operativos.* La ola de avances tecnológicos ha dado lugar a la cuarta revolución industrial, también conocida como Industria 4.0 que ha provocado la interconexión de las máquinas entre sí gracias al conocido como "internet de las cosas" (IoT). Esto ha sido uno de los principales motivos por los que la generación de datos a nivel interno ha aumentado exponencialmente en las empresas. Gracias a un correcto tratamiento y uso de los datos, se puede mejorar en términos de eficiencia y productividad la cadena de valor de cualquier negocio. Hay multitud de ejemplos, pero algunas de las ventajas que pueden aportar a los procesos son: la reducción de tiempos (logística), la reducción de costes (como el de almacenamiento u oportunidad), la disminución de mermas y desperdicios, el aumento de la calidad final del producto, la flexibilidad en la producción, la eficiencia energética o el aumento de la seguridad personal (Infaimon, 2017).

- *Tácticas de venta más rápidas, personalizadas y efectivas*. Las técnicas de marketing son hoy en día más efectivas que nunca. Precisamente, el análisis en tiempo real de los datos nos permite conocer mejor que nunca al cliente y apostar por un mensaje personalizado de tal forma que sea consciente de sus necesidades. Este aspecto permite también desarrollar nuevos productos y/o servicios que se adecuen a los gustos que conocemos de clientes, o potenciales clientes, gracias al volumen de datos e información que tenemos de ellos. Todo esto, además, con una toma de decisiones notablemente mejorada, que permite anticiparnos a los cambios en el mercado o industria con rapidez y con un nivel de confianza mucho mayor.

- *Reducción de fraudes y riesgos*. No son pocas las voces autorizadas que relatan cómo el uso de estas nuevas técnicas pueden permitir una mejora en la protección de activos físicos, de propiedad industrial o financieros (Tabares & Hernandez, 2013). A través de un buen uso de las revolucionarias herramientas analíticas y los métodos estadísticos adecuados se pueden generar modelos para medir la propensión al fraude o incluso generar procedimientos automatizadas para alertar cuando un riesgo de cualquier tipo (desde crediticio a climático) sobrepase el umbral preestablecido.

- *Prevención de ciberataques o similares.* Las amenazas de ciberseguridad, que afectan a todo tipo de empresas, independientemente de su tamaño, no hacen más que crecer. De hecho, el mismo año de publicación del informe número de incidentes de este tipo ascendió un 38% respecto al anterior año (Price Waterhouse Coopers, 2015). Diversos expertos e investigadores en ciberseguridad, como Santiago Hernández Ramos, han expresado la utilidad que tiene el análisis de Big Data para la prevención de ataques de este tipo (Ramos, 2018).

- *Transformación de la lógica de negocio.* Las nuevas técnicas de análisis de datos permiten generar estrategias totalmente distintas y realizar giros de 360 grados en los modelos de negocio actuales. La industria, en general, se encuentra en un continuo estado de transición por culpa del nuevo paradigma de los datos. En este punto, destaco la oportunidad de crear nuevas lógicas de negocio. En este sentido, no se genera un nuevo modelo de negocio, sino que se desarrolla una nueva lógica de este, centrada precisamente en el uso de los datos para la generación de valor. Airbnb, Uber, Amazon o Netflix son ejemplos paradigmáticos de lo que se pretende ejemplificar en este punto. Modelos de negocio que ya existían, como el alojamiento, transporte, compraventa de productos o la visualización de contenido de entretenimiento que, gracias al uso de los datos en sus diferentes modalidades, generan una importante ventaja competitiva.

Por si estos aspectos fueran poco para justificar el uso de estas herramientas, todas las consultoras más importantes a nivel nacional e internacional se han posicionado a favor del uso y tratamiento de datos para generar valor en las compañías. La importancia de los datos disponibles es cada vez mayor para la creación de valor, esto se debe principalmente a que los

datos generan importantes *insights* a lo largo de la cadena de valor y estos, a su vez, se convierten en estrategias de precios, operaciones de optimización de la cadena de valor, prevención del fraude o creación de nuevos modelos de negocio más atractivos (Hurtgen & Mohr, 2018). Asimismo, todos los esfuerzos de la mayoría de los directivos de todas las empresas encuestadas centran sus esfuerzos en mejorar la captación, el análisis y el tratamiento de los datos, mostrándonos, una vez más, la importancia de ellos (Baltassis, Coulin, Gourevitch, Khendek, & Lucas, 2019). Un ejemplo del auge del uso de la Big Data para crear valor, se puede observar en sectores tan estratégicos como el financiero; se augura no tanto la posibilidad sino la necesidad de transformación del sector para poder procesar y almacenar información, resolviendo así los problemas a través del análisis de Big Data (Management Solutions, 2015).

Sin embargo, y como no podía ser de otra forma, hay voces críticas respecto de esta materia (Cumbley & Church, 2013). El precio de las herramientas, la posible brecha en la privacidad de los usuarios por el uso de la información que generan, el uso incorrecto de los datos obtenidos o la selección equivocada de las herramientas, en una industria todavía primeriza, son algunos de los argumentos que esgrimen las contrapartes. Otros problemas como la adaptación de negocios no digitalizados o el importante coste energético que tiene el procesamiento de los datos y su repercusión negativa en la emergencia climática también están sobre la mesa.

En este sentido, las razones más reiteradas que explican el por qué determinadas empresas todavía no han avanzado en lo relativo a iniciativas de análisis de datos son: que el equipo directivo no está comprometido con el uso de técnicas de *Data Analytics*, los equipos de *data science* y los de negocio no se comunican correctamente, la implementación de estas técnicas es difícil o los empleados de las empresas todavía no tienen la suficiente formación para ello (Brahm, Sherer, Fleming, & Bennett, 2017). Todo ello refuerza la idea de que existen capacidades tecnológicas suficientes pero, sin la voluntad necesaria, los cambios no se implementarán.

En definitiva, y con ánimo de concluir este punto, los beneficios del uso de estas técnicas parecen evidentes en su conjunto, si bien es cierto que existen determinadas complicaciones para llevarlo a cabo y poder generar beneficios a corto plazo. Es precisamente este motivo lo que despierta el interés del que suscribe estas palabras… ¿Pueden estás técnicas implementarse con éxito en empresas que, a priori, tienen muy difícil competir contra las gigantes multinacionales? Las PYMES, que conforman más del 95% del total de empresas en España, son un importante motor de la economía española y una vez vistas las potenciales ventajas que pueden aportar el uso de estas herramientas, es momento de examinar si su implementación sería factible (Ministerio de Industria, Comercio y Turismo, 2019).

2.3 Conexión entre análisis de datos y valoración de las compañías.

En este apartado, que pone punto final al marco teórico en el que se encuadra este trabajo, se pretende esclarecer hasta que punto el anteriormente nombrado *Data Analytics* supone no solo una ventaja competitiva sino además tiende a crear valor a quién lo utiliza.

Para ello, se propone examinar cómo se han comportado las empresas con mayor capitalización bursátil en el panorama internacional. La elección de este parámetro se debe a que la cotización de una compañía refleja un nutrido grupo de aspectos fundamentales como el valor de sus activos presentes, la generación de flujos de caja, los beneficios obtenidos o las expectativas de futuro. En otras palabras, y como ya defendía el padre del capitalismo moderno Adam Smith en el siglo XVIII, el mercado se ajusta automáticamente a través de la mano invisible, por lo tanto, ¿Qué mejor medidor que la cotización de las empresas en el mercado de valores?

La evolución de la cotización de las compañías se muestra a continuación:

Tabla 1: Evolución de la cotización de grandes compañías en el año 2000

Rank	Name	COTIZACIÓN ANUAL MEDIA 2000	INDUSTRIA	PAÍS
1	General Electric	$ 477.406,00	Conglomarate	EEUU
2	Cisco Systems	$ 304.699,00	Networking hardware	EEUU
3	Exxon Mobil	$ 286.367,00	Oil and gas	EEUU
4	Pfizer	$ 263.996,00	Health care	EEUU
5	Microsoft	$ 258.436,00	New Tech	EEUU

Fuente: Elaboración propia. Los datos obtenidos para la elaboración de esta tabla provienen de la lista publicada por el Financial Times el 31 de marzo del año 2000.

La Tabla1 muestra la cotización de las compañías más importantes en cuanto a cotización bursátil se refiere, en el año 2000. La cotización máxima asciende ese año a 477.406 millones de dólares para la empresa General Electric que en ese momento dedicaba su actividad económica a servicios muy diveros, pero innovadores para su tiempo. Por lo demás, cabe destacar que las industrias en las que operan las compañías de la Tabla 1son muy diversas, mientras que todas tienen sede en EE. UU., lo que nos permite apreciar su predominio a nivel internacional.

Tabla 2:Evolución de la cotización de grandes compañías en el año 2009

RANK	COMPANY	COTIZACIÓN ANUAL MEDIA 2009	INDUSTRIA	PAÍS
1	PetroChina	$ 333.021,38	Oil and gas	CHINA
2	Exxon Mobil	$ 332.777,35	Oil and gas	EEUU
3	ICBC	$ 237.949,28	Banking	CHINA
4	Microsoft	$ 218.783,08	New Tech	EEUU
5	Wal-Mart	$ 196.525,55	Retail	EEUU

Fuente: Elaboración propia. Los *datos obtenidos para la elaboración de esta tabla han sido obtenidos de la página web https://ycharts.com.*

En la Tabla 2, casi diez años después, en 2009, el panorama ha cambiado sustancialmente. En primer lugar, un claro y nuevo patrón se deduce de observar estos datos: la *commodity* más importante es el petróleo. De hecho, las dos empresas con mayor cotización bursátil están especializadas en este sector. Además, la cotización anual media es inferior a los datos del año 2000 como consecuencia de la irrupción de una fuerte crisis (Campello, Graham, & Harvey, 2010). Finalmente, China comienza a ganar influencia situando a dos compañías en el ranking de las empresas con mayor cotización a nivel mundial.

Tabla 3: Evolución de la cotización de grandes compañías en el año 2019

RANK	COMPANY	COTIZACIÓN ANUAL MEDIA 2019	INDUSTRIA	PAÍS
1	Micrososft	$ 1.049.465,00	New Tech	EEUU
2	Apple Inc	$ 1.030.977,50	New Tech	EEUU
3	Amazon.com	$ 894.520,00	New Tech	EEUU
4	Alphabet Inc	$ 832.370,00	Conglomerate	EEUU
5	Facebook, Inc.	$ 530.147,50	Social media	EEUU

Fuente: Elaboración propia. Los datos obtenidos para la elaboración de esta tabla han sido obtenidos de la página web https://ycharts.com.

En la Tabla 3, que data del año 2019, el panorama ha cambiado de forma todavía más drástica. La cotización media anual entre las cinco primeras compañías ha aumentado casi un 230% pero no solo eso, sino que además las industrias son mucho más homogéneas poniendo especial énfasis en las nuevas tecnologías. Las GAFAM (en referencia a Google, Apple, Facebook, Amazon y Microsoft) se localizan físicamente en Estados Unidos pero se encuentran globalizadas en cuanto a actividades se refiere y conforman un oligopolio mundial cuyo eje fundamental es el uso intensivo que le dan a los datos para crear valor y riqueza (Smryrnaios, 2016).

Respecto a las tres primeras, independientemente de que focalizan esfuerzos en distintos ámbitos, todas ofrecen servicios de *cloud computing, digital distribution* o *artificial intelligence.* Podemos afirmar entonces lo que ya mencionamos anteriormente en este mismo trabajo, los datos han sustituido al petróleo como la *commodity* más importante.

Todo esto, junto con lo anterior, no hace sino apoyar la tesis principal de este trabajo: desde un punto de vista teórico y desde un punto de vista práctico el análisis de Big Data no solo es una oportunidad a muy corto plazo, sino más bien una necesidad. Solo es cuestión de tiempo que deje de ser una ventaja competitiva para comenzar a ser un mínimo indispensable para competir con las demás compañías del sector.

3 Estudio empírico.

3.1 Base de datos.

3.1.1 Obtención de los datos a analizar y motivación del estudio.

A la hora de poder valorar hasta qué punto los beneficios que ofrecen la utilización de las disruptivas herramientas de análisis de datos en las medianas y pequeñas empresas son una realidad o, por otro lado, son una fantasía que solo está al alcance de los gigantes de la industria, se pretende implementar algunas de estas nuevas técnicas en un caso real.

Para la realización de este estudio ha sido imprescindible la participación de una de las empresas líderes en su sector a nivel nacional: Toldos la Estrella (con su marca comercial Solstore). Esta compañía de propiedad familiar nació como una empresa local situada en la provincia de Salamanca hace ya más de 30 años y, hoy en día, es un referente a nivel nacional. Desde el año 2016, la empresa ofrece sus servicios de instalación de toldos, pérgolas y cofres de gran calidad a todas las poblaciones de Madrid y alrededores. La idea de la colaboración surge ante la posibilidad de lograr una relación simbiótica entre empresa y alumno pudiendo obtener *insights* de gran valor para el negocio de Madrid y alrededores, al mismo tiempo que se elabora un estudio empírico desde un punto de vista estrictamente académico. Para ello, la empresa ha facilitado una base de datos de las ventas realizadas en la zona de Madrid y alrededores durante el año 2019 (en adelante "*Data Set*").

En palabras de su actual propietario, los toldos son "*un traje personalizado*" que nos ofrece protección solar ecológica, funciones de valor añadido, decoración y diseño, creación de nuevos espacios y confort. Es importante conocer el producto puesto que en torno a este girará el estudio. Este mimo y personalización del producto es lo que hace única la oferta de productos de las empresas familiares.

3.1.2 Descripción de la base de datos

El *Data Set* consta de once variables que es necesario explicar antes de comenzar el estudio:

Albarán: número del documento mercantil que acredita la entrega de un pedido.

Fecha: momento en el que se efectúa la compra.

N/FRA: el número de factura de la compra.

AG: número de factura (solo para efectos organizativos a nivel interno).

Cliente: Nombre del cliente que efectúa la compra.

Dirección: Lugar en el que se produce la instalación del producto adquirido.

Localidad: municipio donde se produce la instalación del producto.

Artículo: producto específico que ha sido adquirido.

Cantidad: número de productos de ese tipo que se han adquirido en ese pedido.

Precio: precio en euros (€).

Dto: el descuento que obtiene un determinado cliente en una determinada transacción.

El data set original constaba de 1.569 observaciones que se produjeron en la Comunidad de Madrid durante el año 2019.

3.1.3 Descripción del problema de negocio.

Existe una doble vertiente que motiva este estudio. Por un lado, la compañía Toldos la Estrella decidió, hace menos de cinco años, inaugurar su tienda en Madrid (concretamente, en San Sebastián de los Reyes) de cara a expandir su actividad por todo el territorio nacional comenzando por un lugar estratégico: Comunidad de Madrid y alrededores.

- Desde un punto de vista de localización, Madrid está a menos de 200 kilómetros de su sede, en Salamanca, y además tiene muy buenas conexiones con el resto de las ciudades importantes, lo que a largo plazo puede servir como bastión de almacenamiento y logística de productos.
- Desde un punto de vista de rentabilidad. El número de potenciales clientes en la Comunidad de Madrid es muy elevado debido al alto poder adquisitivo de la zona y al sinfín de nuevas edificaciones que se construyen con relativa frecuencia, debido al fenómeno de concentración en nuevas ciudades (Joint Research Centre, 2015).
- Desde un punto de vista demográfico. La densidad demográfica en la Comunidad de Madrid (alrededor de 6.600.000 de personas) es muy superior a la que existe en la provincia de Salamanca (alrededor de 330.000 personas).

Sin embargo, si bien la actividad ha crecido con fuerza desde los primeros años en Madrid y se ha consolidado el acierto que demostró ser la apertura de esta nueva tienda, todavía se trata de un negocio que tiene mucha capacidad de crecimiento como consecuencia de su corto periodo de vida. En este punto, surge la necesidad de acometer un importante análisis de cara a responder algunas preguntas que tienen una importancia elevada en el corto plazo. La solución de estas cuestiones supondría una mejora considerable en los procesos operativos, la reducción de costes y, en definitiva, el aumento de la rentabilidad y un mejor posicionamiento de la compañía. Estas son algunas de las cuestiones a las que se pretende dar repuesta a la finalización del estudio:

- ¿Cuál es el mejor momento del año para vender el producto?
- ¿Qué influye en los clientes a la hora de comprar uno o varios productos?
- ¿Se vende algún producto más que otro dependiendo de la zona?
- ¿Existe la posibilidad de realizar un análisis de *cross selling* para aumentar las ventas?
- ¿Cuáles son las variables que tienen mayor correlación?
- ¿Influyen los descuentos de forma trascendente a la hora de vender el producto?
- ¿Cuáles son las zonas que mayor nivel de ventas tienen?
- ¿Existen otras variables que, habiendo sido observadas, podrían haber mejorado las conclusiones de este análisis?

En este sentido, tras una productiva reunión con el equipo directivo de la empresa, se pueden esbozar tres grandes problemas de negocio.

En primer lugar, la empresa necesita ayuda para fijar cuáles son los territorios más interesantes para centrar los esfuerzos de promoción y marketing a corto plazo, dado el gran tamaño del mercado actual en Madrid que provoca un importante exceso de demanda. Por otro lado, la empresa también nos señala su principal problema dentro de la cadena de valor, que es lo que realmente ralentiza el número de ventas diarias: el montaje de toldos. Esto se debe no solo a la dificultad de realizar esta tarea, sino también a las complicaciones añadidas de la logística como consecuencia de una producción que, si bien esta inspirada en los gustos y modas de años anteriores, no puede adelantarse y predecir las modas de años futuros.

En última instancia, dadas las actuales condiciones del mercado y la buena situación financiera, en términos de solvencia, liquidez y rentabilidad, de la compañía, se pretende iniciar una estrategia de desarrollo de mercados basada en la ampliación geográfica en la península. Sin embargo, existe disparidad de opiniones sobre la nueva zona geográfica que se debería priorizar.

Para terminar, la realización de este estudio pretende no solo afrontar y solucionar los problemas de negocio comentados sino también demostrar la posibilidad real que tienen las PYMES de realizar un correcto tratamiento de sus datos que suponga un factor diferencial respecto a sus competidores. El estudio se realizará principalmente en Rstudio, pero también se utilizarán otras herramientas como Orange y Microsoft Azure de tal forma que se pueda observar el amplio elenco de herramientas de las que actualmente dispone cualquier entidad, así como su eficacia a la hora de abordar problemas reales.

3.2 Metodología.

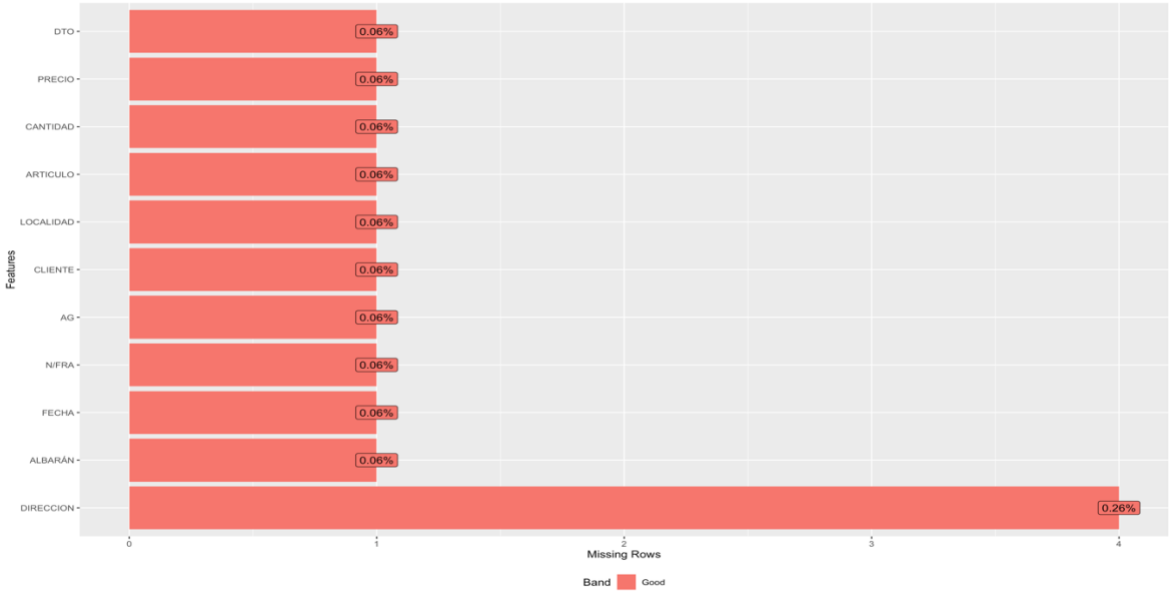
3.2.1 Análisis

3.2.1.1 Tratamiento y limpieza de datos.

Un óptimo análisis de datos requiere de un paso previo antes de comenzar con la exploración de los msimos. En efecto, se está haciendo referencia a la limpieza y tratamiento de datos que nos permite eliminar los datos que sean incompletos, inconsistentes o que contengan errores (Jugulum, 2016).

El primer paso a realizar es la búsqueda de valores ausentes (también conocidos como *NA's* o *missing values*). Para ello se ha creado la *Ilustración 1* :

Ilustración 1: Valores ausentes del dataset



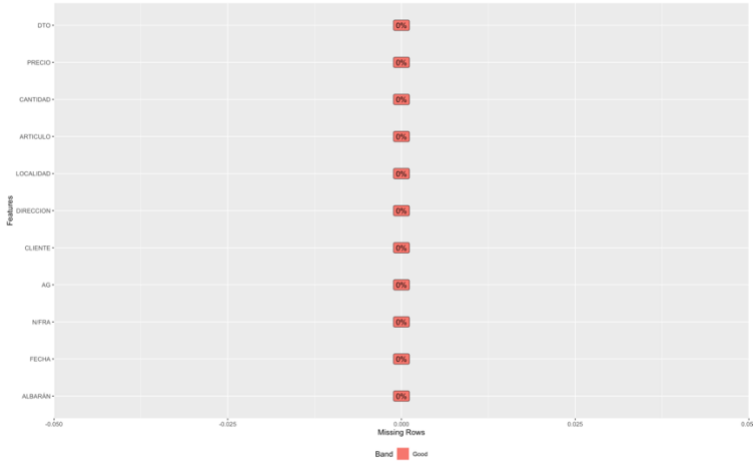
Fuente: Elaboración propia en Rstudio

Esta ilustración muestra que existen valores ausentes en nuestro conjunto de datos, pero son mínimos. Concretamente hay 1 una observación por variable que no tiene valores, a excepción de la variable Dirección donde hay hasta 4 observaciones que tienen valores ausentes.

La mejor opción en este caso es eliminar dichas filas. De esta forma apenas perdemos información, no manipulamos la información del Data Set y, además, mejoramos la calidad de las observaciones. El resultado final solo supone la pérdida de 4 observaciones, pasando de las 1.567 iniciales a las 1.563 pero, a cambio, desaparecen los valores ausentes:

La *Ilustración 2* muestra cómo, tras la debida limpieza, no existe ningún valor ausente en el *Data Set*.

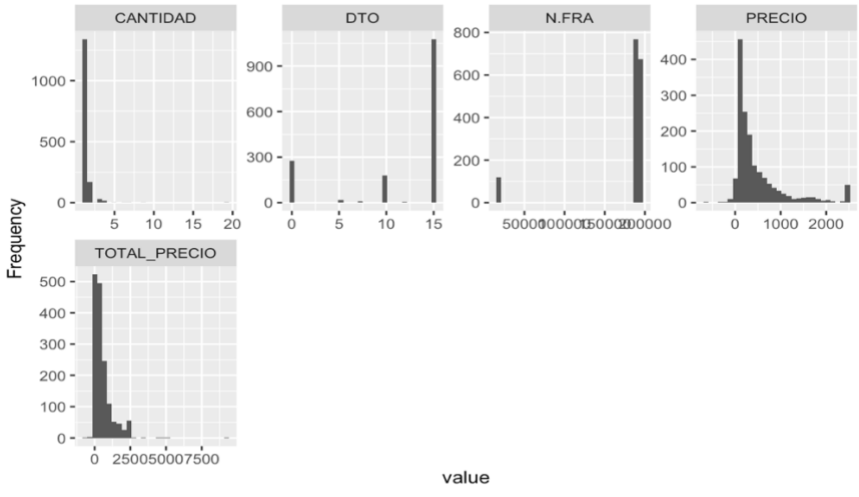
Ilustración 2: Limpieza de valores ausentes del dataset



Fuente: Elaboración propia en Rstudio

A continuación, es preciso analizar la posible existencia de valores atípicos (también conocidos como *outliers*). Su principal problema es que son elementos que pueden no ser representativos de la población pudiendo distorsionar seriamente el comportamiento de los contrastes estadísticos o bien, pueden ser indicativos de las características de un segmento válido de la población y, por consiguiente, una señal de la falta de representatividad de la muestra. En cualquier caso, son datos que merece la pena examinar para determinar si deben ser contemplados en el análisis o por el contrario, deben ser modificados o incluso eliminados *(Hawkins, 1980)*.

Ilustración 3: Conjunto de valores atípicos en el dataset



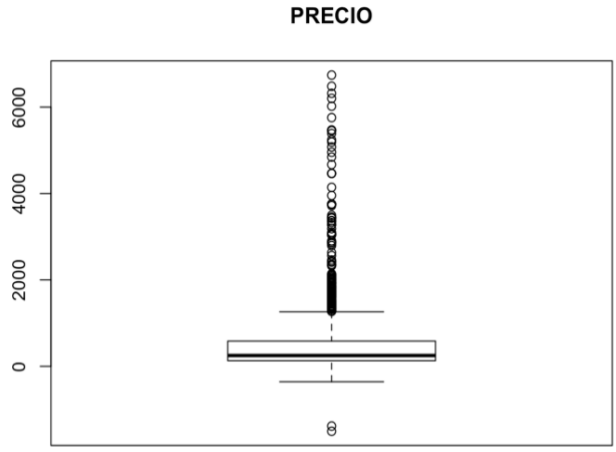
Fuente: Elaboración propia en Rstudio

La Ilustración 3 es un primer paso para determinar hasta qué punto pueden existir valores atípicos en nuestra base de datos. Si bien no parece que exista ningún outlier, es preciso un análisis más detallado en cada caso:

Respecto de la variable precio, podemos destacar dos aspectos fundamentalmente, después de examinar la información proveniente de la *Ilustración 4*:

- En primer lugar, la diferencia entre la media y el máximo es bastante alta. Esta dispersión podría indicar la presencia de algún outlier que examinaremos.
- Por otro lado, se aprecia que en la variable precio existen también observaciones de precio en negativo lo que implica algún tipo de descuento especial a mayores. Habrá que examinar hasta que punto nos resulta útil la creación de otra variable que recoja estos descuentos dejando el precio en positivo.

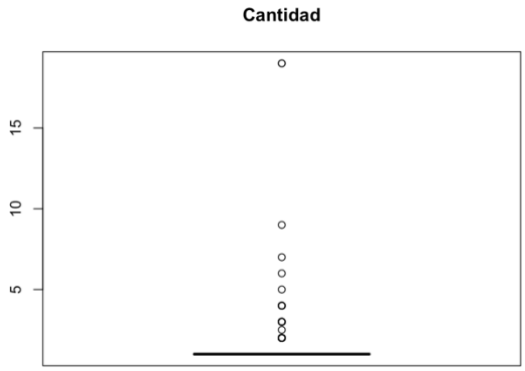
Ilustración 4:Boxplot de dispersión de la variable PRECIO del dataset



Fuente: Elaboración propia en Rstudio

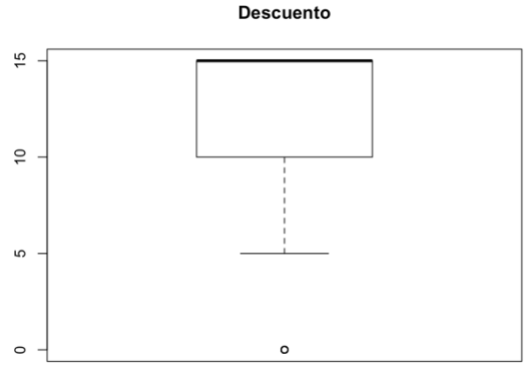
Respecto al resto de las variables, no hay ningún valor extraño que merezca la pena ser estudiado. Sin embargo, la variable N.FRA debería desaparecer puesto que no aporta información relevante. En este sentido, tras la observación de *las Ilustraciones 5 y 6*, podemos confirmar nuestra idea de que no existe outlier alguno en esas variables.

Ilustración 5: Boxplot de dispersión de la variable CANTIDAD del dataset



Fuente: Elaboración propia en Rstudio

Ilustración 6: Boxplot de dispersión de la variable DTO del dataset



Fuente: Elaboración propia en Rstudio

3.2.1.2 Análisis Exploratorio

Una vez detectados los valores ausentes y los valores atípicos que pueden afectar, y de hecho afectan, a la validez de las conclusiones que se pretende obtener, y antes de pasar a realizar el correspondiente EDA, se busca realizar alguna transformación para mejorar la calidad de las conclusiones que se extraigan del estudio.

En primer lugar, con el Data Set actual se muestra el precio de cada objeto y la cantidad de productos que se compra en cada transacción de forma separada. Para evitar perder información se crea una nueva variable con el nombre de TOTAL_PRECIO que básicamente es una operación que consiste en multiplicar la variable PRECIO y la variable CANTIDAD para

obtener el montante total. La *Ilustración 3* nos muestra que ha sido una operación exitosa de transformación.

Por otro lado, la variable FECHA, aunque aporta mucha información, de cara a que adquieran mayor valor las visualizaciones que se pretenden realizar en los siguientes apartados, se utiliza para crear nuevas variables. Con esta variable y a través de sencillas operaciones se crean la variable MES y la variable ESTACION_DEL_AÑO. Gracias a esto, se podrá explotar de forma más eficiente la potencia gráfica de la herramienta Rstudio. En relación con lo expuesto en estas líneas, se pretende simplificar el innumerable y muy heterogéneo elenco de productos con los que la empresa comercia y simplificarlo a cinco categorías (accionamiento, tejido, estructura, personal y otro) bajo la variable TIPO_DE_ARTICULO.

Después de esta breve transformación y creación de nuevas variables se procede a comenzar con el análisis exploratorio. Una de las grandes ventajas que tiene R es que te permite obtener información muy útil con fórmulas muy sencillas. Las funciones *summary* y *str* son un gran ejemplo de esto:

Ilustración 7: Extracto sobre el resumen del dataset

```
> summary(VENTAS_SOLSTORE_2019_sinNAs )
     FECHA                         MES             ESTACION_DEL_AÑO       N/FRA          CLIENTE
 Min.   :2019-01-09 00:00:00   Length:1563        Length:1563        Min.   : 19001   Length:1563
 1st Qu.:2019-03-25 00:00:00   Class :character   Class :character   1st Qu.:190561   Class :character
 Median :2019-05-14 00:00:00   Mode  :character   Mode  :character   Median :191224   Mode  :character
 Mean   :2019-05-16 11:30:03                                        Mean   :178506
 3rd Qu.:2019-06-21 00:00:00                                        3rd Qu.:192140
 Max.   :2019-11-08 00:00:00                                        Max.   :195373
  DIRECCION          LOCALIDAD           ARTICULO         TIPO_DE_ARTICULO      CANTIDAD          PRECIO
 Length:1563        Length:1563        Length:1563        Length:1563        Min.   : 1.000   Min.   :-628.7
 Class :character   Class :character   Class :character   Class :character   1st Qu.: 1.000   1st Qu.: 129.6
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median : 1.000   Median : 249.9
                                                                             Mean   : 1.214   Mean   : 470.7
                                                                             3rd Qu.: 1.000   3rd Qu.: 584.0
                                                                             Max.   :19.000   Max.   :2482.4

      DTO          TOTAL_PRECIO
 Min.   : 0.00   Min.   :-628.7
 1st Qu.:10.00   1st Qu.: 130.6
 Median :15.00   Median : 289.3
 Mean   :11.63   Mean   : 538.9
 3rd Qu.:15.00   3rd Qu.: 670.2
 Max.   :15.00   Max.   :9295.9
```

Fuente: Elaboración propia en Rstudio

Como se puede observar en la *Ilustración 7* la función *summary* no supone gran complejidad y te aporta mucha información para conocer en profundidad el conjunto de datos. En la primera línea, de color azul, aparece la fórmula que solo requiere de la palabra summary y el Data Set que se pretende analizar entre paréntesis. La información que se obtiene es toda la que aparece justo debajo.

Gracias a estos resultados salen a la luz algunos aspectos importantes:

- La variable FECHA explica que las transacciones se producen desde el 1 de enero de 2019 hasta el 8 de noviembre de 2019. Esto se debe a que los datos fueron cedidos por la empresa Toldos la Estrella en el mes de noviembre y todavía no estaban todos los datos del año 2019.
- El siguiente punto que merece la pena comentar llega en la variable CANTIDAD donde hay un mínimo de 1 producto por transacción y un máximo de 19 productos por transacción. Esto es interesante porque, si tenemos en cuenta que la media de productos obtenidos por clientes es 1,214 por pedido, eso significa que hay, como mínimo, un cliente que ha comprado mucho más que la media y sería útil intentar entender el porqué.
- Por último, hay un apunte que hacer respecto de las variables PRECIO y TOTAL_PRECIO. Si como es lógico, los cuartiles cambian como consecuencia de la aplicación de la operación matemática que antes se ha explicado, es útil conocer que existen "precios en negativo". Estos precios que parecen no tener sentido no son más que un descuento especial que se utiliza para obtener más ventas de determinados clientes. Al no ser descuentos programáticos, estos no se recogen en una variable distinta como podría ser la variable DTO, sino en negativo para reducirlas del precio total.

Ilustración 8: Extracto sobre las características de las variables del dataset

```
> str(VENTAS_SOLSTORE_2019_sinNAs )
'data.frame':   1563 obs. of  13 variables:
 $ FECHA           : POSIXct, format: "2019-04-10" "2019-04-24" "2019-04-30" "2019-05-06" ...
 $ MES             : chr  "ABRIL" "ABRIL" "ABRIL" "MAYO" ...
 $ ESTACION_DEL_AÑO: chr  "PRIMAVERA" "PRIMAVERA" "PRIMAVERA" "PRIMAVERA" ...
 $ N/FRA           : num  19001 19002 19003 19004 19005 ...
 $ CLIENTE         : chr  "Anónimo" "Anónimo" "Anónimo" "Anónimo" ...
 $ DIRECCION       : chr  "C/LUIS RODRIGUEZ ONTIVEROS, 114" "C/ SUEÑOS, 5 P.5 BJ A" "C/ALCAÑIZ, 9 - P.B - 4ºF" "CTRA. FUE
NCARRAL, 56 - OF.70" ...
 $ LOCALIDAD       : chr  "ALCOBENDAS" "ALCOBENDAS" "BARAJAS" "ALCOBENDAS" ...
 $ ARTICULO        : chr  "TEJADILLO" "TEJADILLO" "TEJADILLO" "TEJADILLO" ...
 $ TIPO_DE_ARTICULO: chr  "ESTRUCTURA" "ESTRUCTURA" "ESTRUCTURA" "ESTRUCTURA" ...
 $ CANTIDAD        : num  1 1 1 1 1 1 1 1 1 1 ...
 $ PRECIO          : num  114.1 325.6 80.3 70.4 227.3 ...
 $ DTO             : num  15 15 15 15 15 15 15 15 10 0 ...
 $ TOTAL_PRECIO    : num  114.1 325.6 80.3 70.4 227.3 ...
```

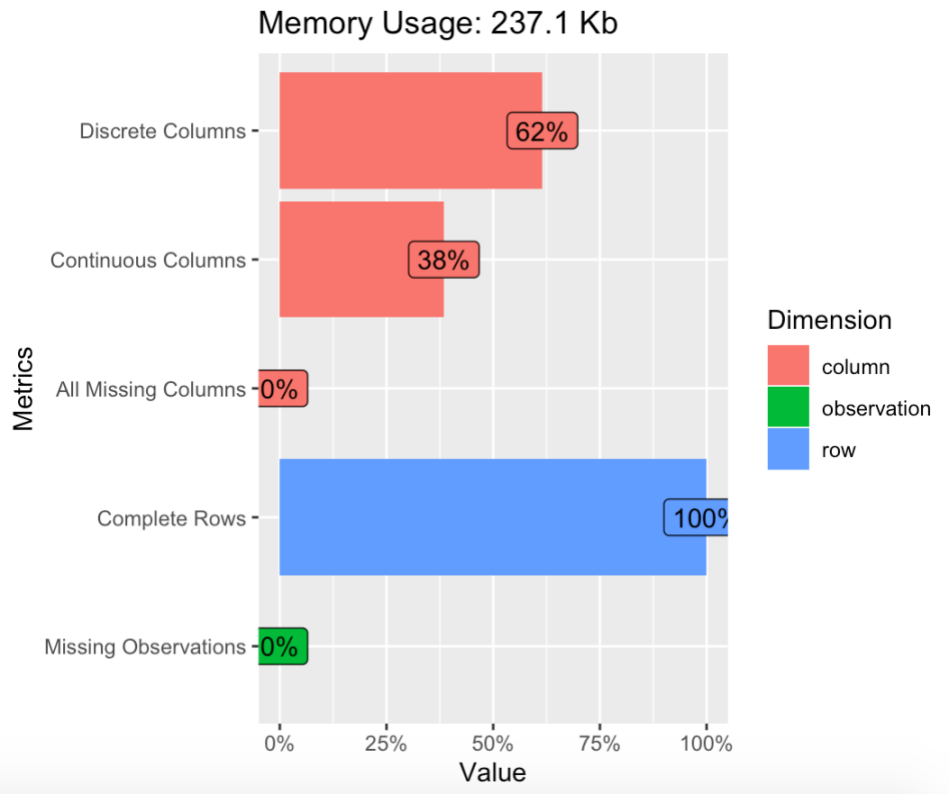
Fuente: Elaboración propia en Rstudio

La función *str (Ilustración 8)* es también de uso muy sencillo y en este caso muestra las características de cada variable distinguiendo entre variables categóricas y numéricas, así como

mostrándonos algún ejemplo de sus observaciones. Es de utilidad para entender el conjunto de observaciones que están recogidas en el Data Set y saber qué herramientas o técnicas pueden generar soluciones más eficazes.

Seguidamente, y mediante la función *plot_intro* observamos algunos de los puntos de interés que habíamos comentado en apartados anteriores representados gráficamente para un mejor entendimiento de estos:

Ilustración 9: Clasificación de variables del dataset



Fuente: Elaboración propia en Rstudio

De la *Ilustración 9* se puede deducir que hay un mayor número de variables de carácter discreto, esto es, variables que no aceptan cualquier valor sino solo aquellos que pertenecen al conjunto, representando el 62% de las variables. Por otro lado, las variables continuas, es decir, aquellas que toman valores a lo largo de un continuo dentro de un intervalo, únicamente representan el 38% de las variables.

Aquí también se evidencia que no hay ninguna columna o línea sin observación dentro del Data Set gracias al trabajo de limpieza que realizamos anteriormente.

Prosiguiendo con el análisis, se busca analizar la correlación ya que es un medidor importante en todo estudio empírico de estas características. Esto se debe a que a la hora de analizar datos siempre hay dos objetivos que sobresalen por encima del resto con carácter general: comparar grupos y estudiar las posibles relaciones que se producen entre variables. De este modo, la correlación nos indica la fuerza y la dirección de una relación lineal, así como la proporcionalidad entre dos variables estadísticas (Velickovic, 2015). Precisamente de aquí se deduce el alto valor informativo que tiene para la probabilidad y la estadística.

En la *Ilustración 10* se muestra la correlación que existe entre las diferentes variables numéricas que hay en este Data Set mediante la función *plot_correlation*:

Ilustración 10: Correlación entre las variables numéricas del dataset

Fuente: Elaboración propia en Rstudio

Como se puede observar en la *Ilustración 10*, siendo 1 el resultado que se produce cuando dos variables están directamente correlacionadas y, siendo -1 el resultado que se obtiene cuando dos variables están inversamente correlacionadas, no se puede destacar que exista ninguna correlación importante entre ninguna variable en este conjunto de datos. Más allá del hecho de que las variables TOTAL_PRECIO y PRECIO se expliquen la una a la otra como consecuencia de que una es resultado directa de la otra, debido a las operaciones matemáticas.

Pues bien, llegados a este punto, y antes de pasar a una de las partes centrales del análisis exploratorio como es la visualización de los datos mediante gráficas, puede ser de gran utilidad crear diversas tablas para poder ordenar los datos y tener un mejor conocimiento de los datos de que disponemos para después plasmarlos sobre gráficas.

En primer lugar, a través de operaciones, ahora sí con mayor dificultad a la hora de ejecutarlas en Rstudio, se agrupan las observaciones en función de la variable MES y se muestra mediante la nueva variable n, creada *ad hoc* para esta tabla, el número de pedidos por mes y mediante la variable TOTAL_PRECIO, el precio medio que se paga por esas operaciones cada mes.

En esta *Tabla 4* se ha ordenado de mayor precio a menor precio medio. De esta forma, la tabla ofrece una información única y no conocida hasta este momento.

Tabla 4: Agrupación de observaciones en función de la variable MES (orden según TOTAL_PRECIO)

	MES	n	TOTAL_PRECIO
	<chr>	<int>	<dbl>
1	SEPTIEMBRE	95	638.
2	MARZO	204	612.
3	OCTUBRE	80	575.
4	MAYO	375	557.
5	ABRIL	263	548.
6	JUNIO	190	514.
7	JULIO	97	503.
8	ENERO	103	445.
9	FEBRERO	109	442.
10	AGOSTO	32	354.

Fuente: Elaboración propia en Rstudio

El mes donde mayor precio tienen las transacciones es el mes de septiembre, mientras que, curiosamente, uno de los meses que menor precio medio tienen las transacciones que efectúa la empresa es justo agosto, época que pese al carácter estacional del negocio, no tiene demasiado peso.

La razón que puede explicar este fenómeno es que la instalación de toldos se produce normalmente fuera de temporada alta de ocupación de casas o apartamentos de verano. Para justificar esta afirmación ordenamos *Tabla 4* en función del número de transacciones que se producen cada mes:

Tabla 5: Agrupación de observaciones en función de la variable MES (orden según n)

	MES	n	TOTAL_PRECIO
	<chr>	<int>	<dbl>
1	MAYO	375	557.
2	ABRIL	263	548.
3	MARZO	204	612.
4	JUNIO	190	514.
5	FEBRERO	109	442.
6	ENERO	103	445.
7	JULIO	97	503.
8	SEPTIEMBRE	95	638.
9	OCTUBRE	80	575.
10	AGOSTO	32	354.

Fuente: Elaboración propia en Rstudio

En la *Tabla 5* se puede deducir algunas de las posibles explicaciones que se esbozaban con anterioridad. En efecto, el número de compraventas de toldos (englobando todo tipo de estructuras complejas) y material de recambio para estos, crece significativamente los meses antes de la temporada alta de verano. De esta forma los meses de mayo, abril y marzo es donde se producen con mayor frecuencia estas compras. Esto se puede deber a que los clientes pretenden tener la casa o apartamento de verano listo para la temporada de ocupación y, de esta forma, se realizan las obras con anterioridad a su llegada. De esta misma forma, es en septiembre cuando se realizan los mayores desembolsos como consecuencia de la mayor utilización que se produce en toldos en las fechas anteriores.

Continuando con este análisis que pretende organizar los datos a través de tablas creadas en R para entender todavía mejor este conjunto de datos, se pretenden agrupar las tablas en función de las localidades y el precio medio de cada una. A continuación de este texto, se muestran dichos extractos, recogidos ambos en la *Tabla 6*, ordenados en función del precio medio (izquierda) o en función del número de transacciones en cada localidad (derecha).

Tabla 6: Agrupación de datos en función de la LOCALIDAD (orden según TOTAL_PRECIO y según n, respectivamente)

	LOCALIDAD	n	TOTAL_PRECIO
	<chr>	<int>	<dbl>
1	TRESCANTOS	58	720.
2	POZUELO DE ALARCON	27	646.
3	COLMENAR VIEJO	32	620.
4	PARACUELLOS DEL JARAMA	47	589.
5	MADRID	655	550.
6	SAN SEBASTIAN DE LOS REYES	243	549.
7	SAN AGUSTIN DE GUADALIX	42	521.
8	LAS ROZAS	33	498.
9	ALCOBENDAS	132	488.
10	TORRELODONES	22	417.
11	FUENLABRADA	26	375.
12	VILLANUEVA DE LA CAÑADA	28	308.

	LOCALIDAD	n	TOTAL_PRECIO
	<chr>	<int>	<dbl>
1	MADRID	655	550.
2	SAN SEBASTIAN DE LOS REYES	243	549.
3	ALCOBENDAS	132	488.
4	TRESCANTOS	58	720.
5	PARACUELLOS DEL JARAMA	47	589.
6	SAN AGUSTIN DE GUADALIX	42	521.
7	LAS ROZAS	33	498.
8	COLMENAR VIEJO	32	620.
9	VILLANUEVA DE LA CAÑADA	28	308.
10	POZUELO DE ALARCON	27	646.
# … with 30 more rows			

Fuente: Elaboración propia en Rstudio

Pues bien, de la tabla de la izquierda se extraen datos que son de vital importancia a la hora de extraer conclusiones útiles en los próximos apartados. En primer lugar, Trescantos (45.000 habitantes), Pozuelo de Alarcón (85.000 habitantes) y Colmenar Viejo (50.000 habitantes) son las localidades, dentro del grupo de aquellas que cuentan con al menos 20 transacciones por localidad (las que no llegaban a estos estándares no las hemos seleccionado por entender que podían ser compraventas aisladas de particulares que no representaban una realidad), que tienen un mayor precio medio superando todas ellas los 600 euros por envío. Esto revela que son lugares muy atractivos para invertir, no solo ya porque son localidades con población alta, casas preparadas para pasar veranos (amplias, con piscinas y jardines) o simplemente porque tienen unas rentas per cápita bastante altas (como por ejemplo Pozuelo de Alarcón que es la primera en el ranking nacional), sino porque además solo con observar la tabla se puede deducir que suelen ser obras bastantes importantes y consecuentemente la rentabilidad también es mayor en estas zonas.

Por otro lado, en la tabla de la derecha se muestran las dos variables anteriormente comentadas, pero se ordena en función del número de transacciones por localidad (n). En este caso, las localidades que más productos de este tipo demandan son Madrid (6.5 millones de habitantes), San Sebastián de los Reyes (85.000 habitantes) y Alcobendas (120.000 habitantes). A este respecto, parecía evidente que Madrid, por el número tan elevado de población, iba a tener un

gran número de ventas. Por otro lado, tanto San Sebastián de los Reyes como Alcobendas se sitúan en los siguientes lugares por volumen de transacciones. Aquí es reseñable ilustrar cómo San Sebastián de los Reyes, con una población no demasiado alta, tiene un inmenso número de compraventas. La explicación de este hecho no puede ser otra que la existencia de una tienda física en esta localidad que hace que el número de ventas aumente significativamente. Por lo tanto, en futuros análisis habrá que valorar la importancia de abrir tiendas físicas para incrementar las ventas en aquellos lugares donde se entienda más efectivo por el número de potenciales clientes.

Por último, y antes de continuar con la visualización de algunos de estos datos, elaboramos la *Tabla 7* en la que incluimos las variables TIPO_DE_ARTICULO y TOTAL_PRECIO.

Tabla 7: Agrupación de datos en función del TIPO DE ARTÍCULO (orden según n)

	`TIPO DE ARTÍCULO`	n	TOTAL_PRECIO
	<chr>	<int>	<dbl>
1	TEJIDO	686	901.
2	ACCIONAMIENTO	591	250.
3	ESTRUCTURA	194	305.
4	OTRO	42	363.
5	PERSONAL	20	236.

Fuente: Elaboración propia en Rstudio

Claramente de aquí se puede destacar que los artículos que se engloban dentro de la categoría TEJIDO son los más numerosos en cuanto a ventas, pero también los que más ingresos generan ya que su precio medio es mucho mayor al resto (900 euros frente a los demás que ninguno supera los 400 euros).

Exploraremos más adelante el porqué de esto, pero a priori, la razón lógica que puede explicar esto es el hecho de que, aunque normalmente el grueso de ventas se dé por la instalación de nuevos toldos, hay una cantidad de transacciones bastante significativas que se originan como consecuencia del deterioro de los toldos e infraestructuras que necesitan de recambios en forma de tejido, mayoritariamente.

Respecto a los artículos más comprados, también se ha elaborado la *Tabla 8* para identificarlos:

Tabla 8: ARTICULO más demandado de 2019

	ARTICULO	n	TOTAL_PRECIO
	<chr>	<int>	<dbl>
1	OPERADOR SUNEA 15/17 IO	77	465.
2	INCREMENTO LONA SOLTIS	76	274.
3	MANDO SITUO 5 IO	74	95.1
4	MANDO SITUO 1 IO	64	66.6
5	OPERADOR SUNEA 35/17 IO	54	405.
6	AUTOMATISMO VIENTO EOLIS VELETA IO	51	155.
7	TEJADILLO	46	250.
8	AUTOMATISMO EOLIS 3D IO NEGRO	35	187.
9	OPERADOR SUNEA 55/17 IO	34	368.
10	TOLDO MOD. VERANDA	29	2645.

Fuente: Elaboración propia en Rstudio

Las tres ventas más habituales que se producen en la Comunidad de Madrid superando los 70 pedidos son: el Operador Sune 15/17 IO (pertenece a la familia que ha sido denominado ACCIONAMIENTO), un incremento de Lona Soltis (que pertenece a la familia TEJIDO) y en último lugar del ranking se sitúa el Mando Situo 1 IO (que también pertenece a la familia de ACCIONAMIENTO).

3.2.1.3 Visualización

Dentro de los posibles usos que se le pueden dar a los datos, la visualización ha sido y es una herramienta muy potente que permite extraer conclusiones de información compleja para mejorar la toma de decisiones.

Para la realización de este apartado, además de proseguir con el análisis a través de Rstudio, que tiene funciones muy efectivas para la representación de datos, se han utilizado otras herramientas de visualización más sencillas de utilizar para los usuarios sin experiencia, pero igualmente útiles a la hora de mejorar la toma de decisiones. Concretamente, la aplicación RAW graphs, que se puede encontrar en internet y es libre de pago, permite a sus usuarios

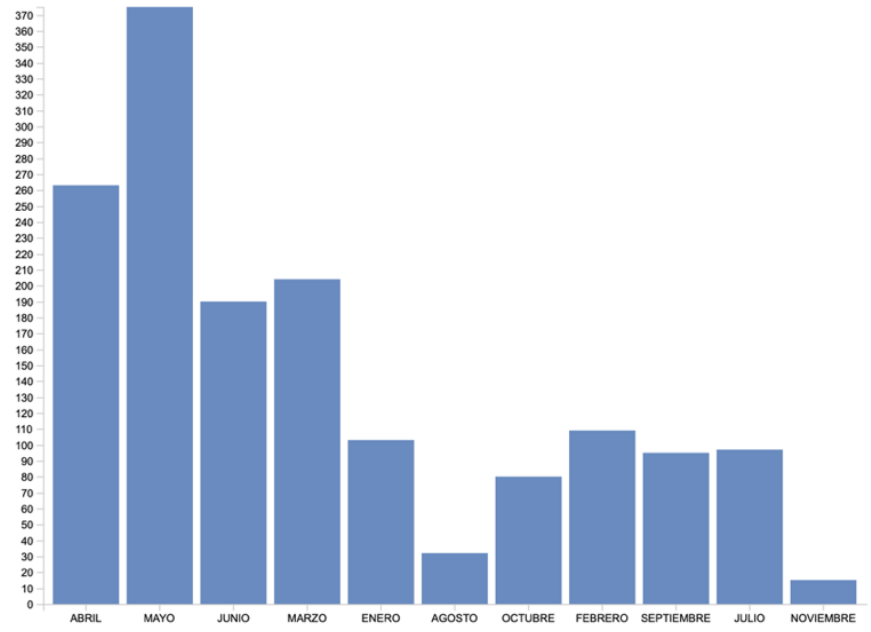
utilizar un buen número de gráficos para la representación de datos sin necesidad de tener conocimientos en lenguajes de programación[1].

Se procede a comenzar esta visualización con la app que se mencionaba en las líneas anteriores.

Los *diagramas de barras* y los *diagramas de círculos* (también conocidos como *circle parkings)* son dos de las formas de representación más útiles para averiguar la trascendencia de cada variable dentro de un conjunto de datos. Por eso utilizaremos estas para representar las variables que a estas alturas del estudio conocemos de mayor importancia: MES, TIPO DE ARTICULO, LOCALIDAD, Y TOTAL_PRECIO.

El diagrama de barras de la *Ilustración 11* muestra cómo se reparten las transacciones a lo largo de los meses estudiados.

Ilustración 11: Diagrama de barras sobre las transacciones en función de la variable MES



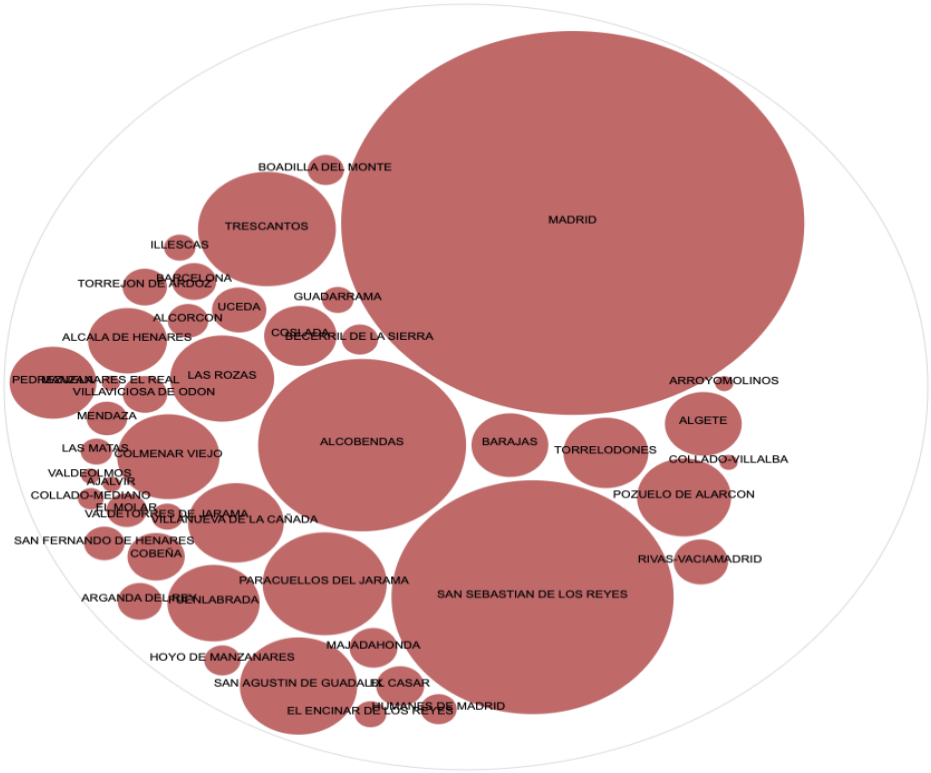
Fuente: Elaboración propia en RAW graphs

[1] Link a la página web https://app.rawgraphs.io

Todo esto con la especial particularidad que se ha comentado en el apartado anterior, noviembre se encuentra incompleto y de diciembre no se han obtenido datos. Por lo demás, mayo destaca por ser el mes con mayor número de transacciones con una diferencia sustancial sobre el resto.

En el diagrama de círculos (*Ilustración 12*) se muestran los pesos relativos de cada LOCALIDAD en función del número de veces que aparecen en las transacciones.

Ilustración 12: Diagrama de círculos sobre las LOCALIDADES en función de las transacciones



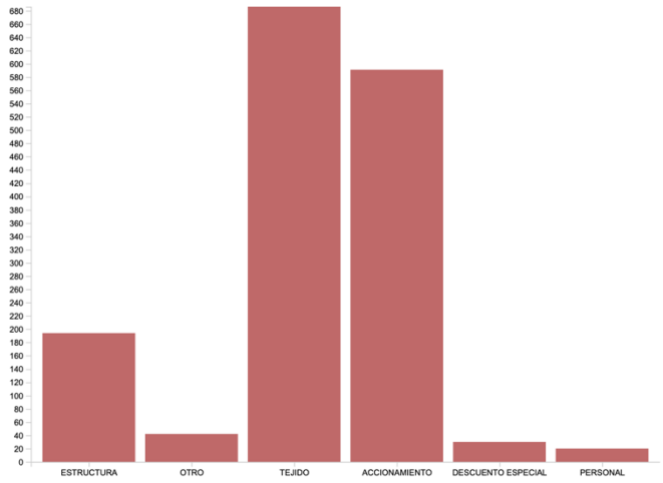
Fuente: Elaboración propia en RAW graphs

Este diagrama permite reconocer, de forma rápida y muy visual, cuáles son los lugares donde se producen el mayor número de ventas. Destacan la ciudad de Madrid, San Sebastián de los Reyes y Alcobendas. Anteriormente se había mostrado esta clasificación mediante una tabla que permitía ordenar y agrupar por los medidores que se preferían, pero con estas fáciles

representaciones se obtiene una información mucho más visual, haciendo más sencilla la tarea de comparar las localizaciones más sugerentes.

Continuando con este anális, en la *Ilustración 13* se muestran las transacciones totales divididas en función del TIPO DE ARTICULO que se compra.

Ilustración 13: Diagrama de barras sobre las transacciones totales en función del TIPO DE ARTÍCULO

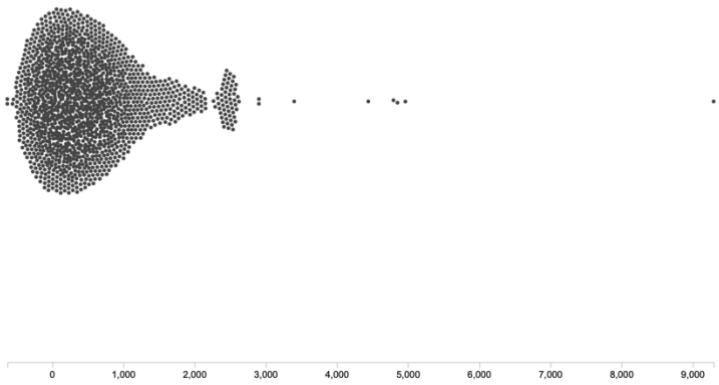


Fuente: Elaboración propia en RAW graphs

De aquí se extrae como conclusión que el mayor número de ventas provienen de las familias del TEJIDO y el ACCIONAMIENTO. Además de esto, también nos encontramos con otras categorías como PERSONAL Y DESCUENTO ESPECIAL. Respecto a la primera, esta solo se carga al cliente cuando es un gasto extraordinario por motivo de agilizar la instalación o bien porque la instalación es compleja y requiere más mano de obra de lo habitual, ya que de otro modo, este gasto de instalación suele ir incorporado al precio de compra. Por otro lado, también hay una categoría especial que, como ya se ha comentado en páginas anteriores, no constituye un ingreso como tal, sino más bien lo contrario, es una rebaja de carácter extraordinario que se realiza solo a determinados clientes en función del cumplimiento de una serie de requisitos. Independientemente de esto, son categorías que no representan una cuota significativa en el apartado de Ingresos Netos a la hora de analizar la cuenta de resultados de la compañía, tal y como también muestra la *Ilustración 13*.

Otra de las muchas posibilidades que nos ofrece esta aplicación gratuita es la de representar la dispersión de los precios en una gráfica que en inglés se denomina *beeswarm plot*, mediante la representación de la variable TOTAL_PRECIO. Gracias a esta representación se puede observar la dispersión de los precios de forma muy visual en la *Ilustración 14*. El grueso de precios se sitúa entre los 0 euros y los 2.000 euros, superando o reduciendo este intervalo en muy pocas ocasiones.

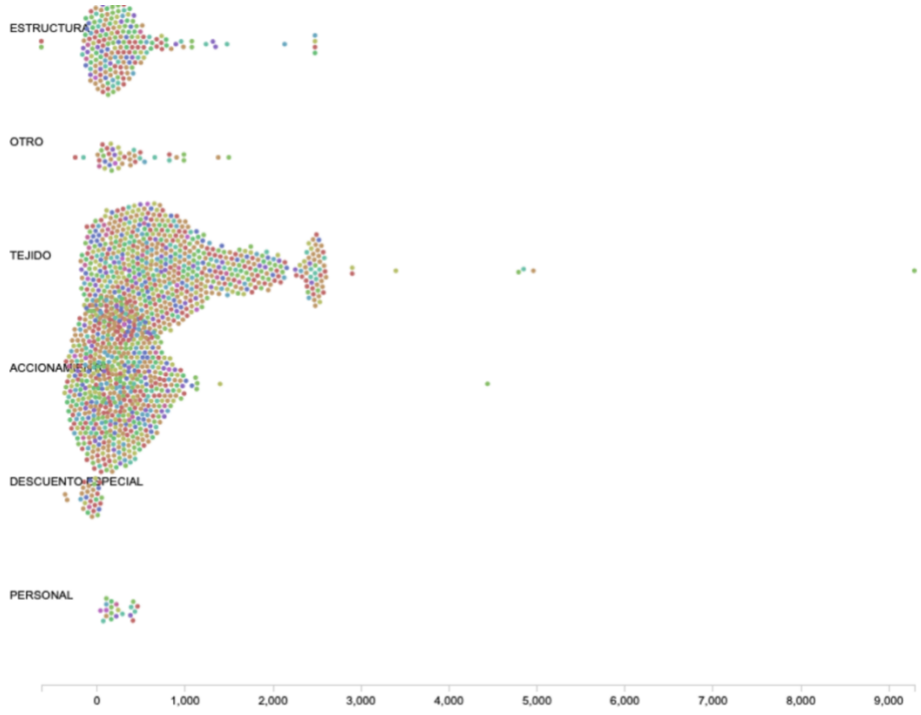
Ilustración 14: Diagrama de dispersión de la variable TOTAL_PRECIO



Fuente: Elaboración propia en RAW graphs

Sin embargo, las posibilidades de este tipo de representación son infinitas por lo que a modo ilustrativo y también para conocer mejor la distribución de precios, se incluyen nuevas reglas a la hora de representar esta.

Ilustración 15: Diagramas de dispersión sobre la variable TOTAL_PRECIO en función de la variable TIPO DE ARTÍCULO

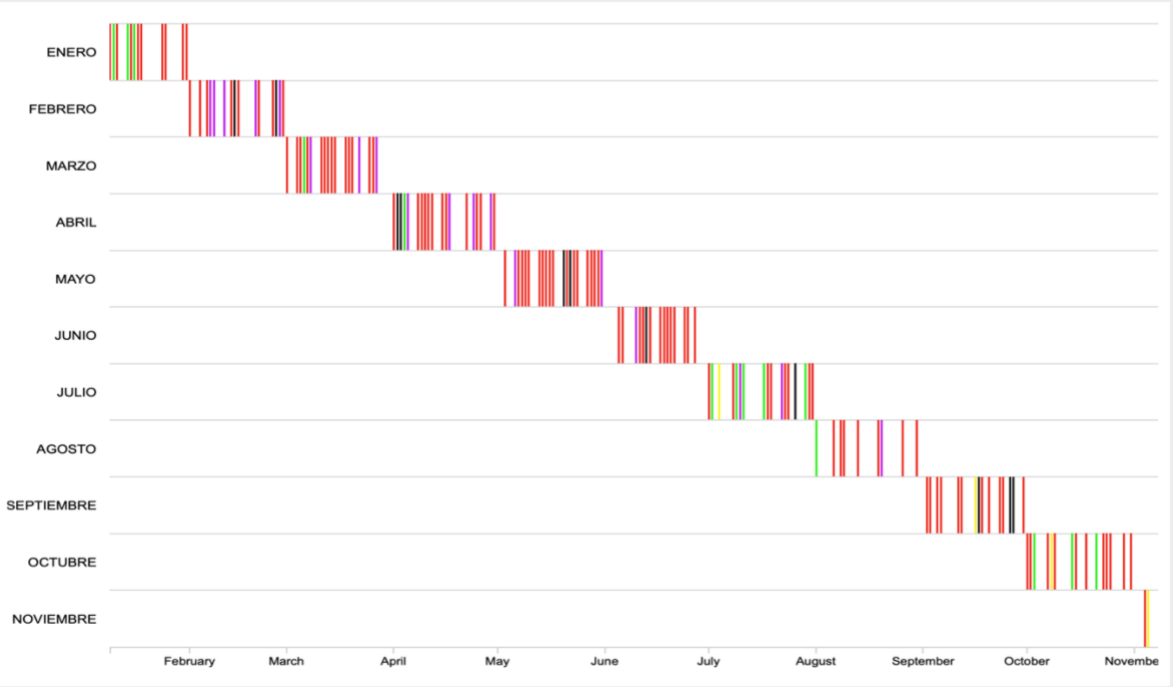


Fuente: Elaboración propia en RAW graphs

La *Ilustración 15* ya no solo explica la distribución de los precios recogidos en el Data Set sino que va más allá agrupando la dispersión de estos en función de la variable TIPO_DE_ARTICULO y además se añade la opción de colorear cada transacción en función de la variable MES. La estructura que aparece es similar a la de la *Ilustración 14* pero aportándonos una información mucho mayor gracias a un par de *clicks* más.

Otro ejemplo de ilustración muy interesante es la posibilidad de crear una serie temporal mediante una *Gantt chart* que nos permite ordenar las ventas según la variable MES y el TIPO_DE_ARTICULO (este último representado en color, siendo accionamiento en morado, y tejido en rojo, los más repetidos).

Ilustración 16: Serie temporal de las ventas en función de la variable MES y el TIPO DE ARTÍCULO

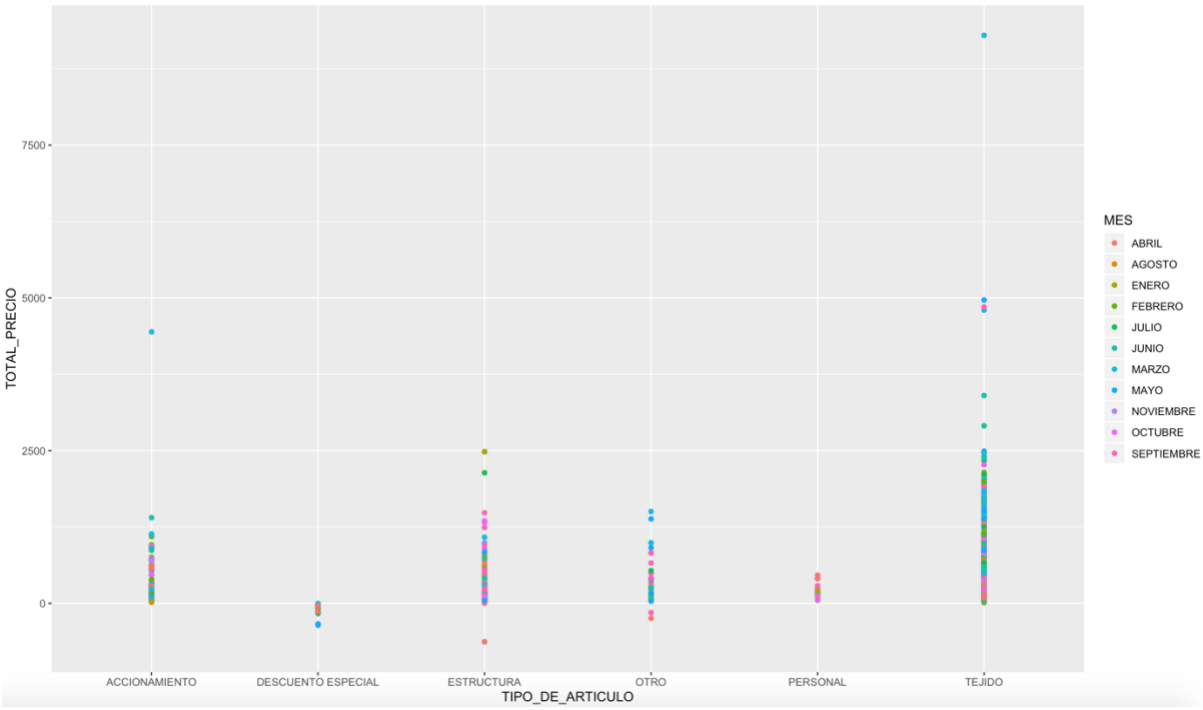


Fuente: Elaboración propia en RAW graphs

Después de haber observado todas las opciones que nos proporciona esta herramienta, es momento de continuar con el análisis en R para ver gráficas más sofisticadas y útiles, pero al mismo tiempo, de mayor complejidad técnica.

La primera gráfica que se genera a través de R es una gráfica de puntos que representa la variable TOTAL_PRECIO en función de la variable MES y la variable TIPO DE ARTICULO. Sin embargo, este tipo de análisis, no siempre es exitoso puesto que para encontrar qué tipo de gráfica se adecúa a las necesidades de representación de nuestros datos, se requiere tiempo y conocimientos. En la *Ilustración 17* se observa la gráfica de puntos obtenida mediante el paquete *ggplot* y más concretamente con la función *geom_point.* De ésta claramente no se obtiene demasiada información y tampoco es útil para usarla como referencia en futuras reuniones con proveedores, inversores o clientes en aras de buscar ejemplificar ningún hecho.

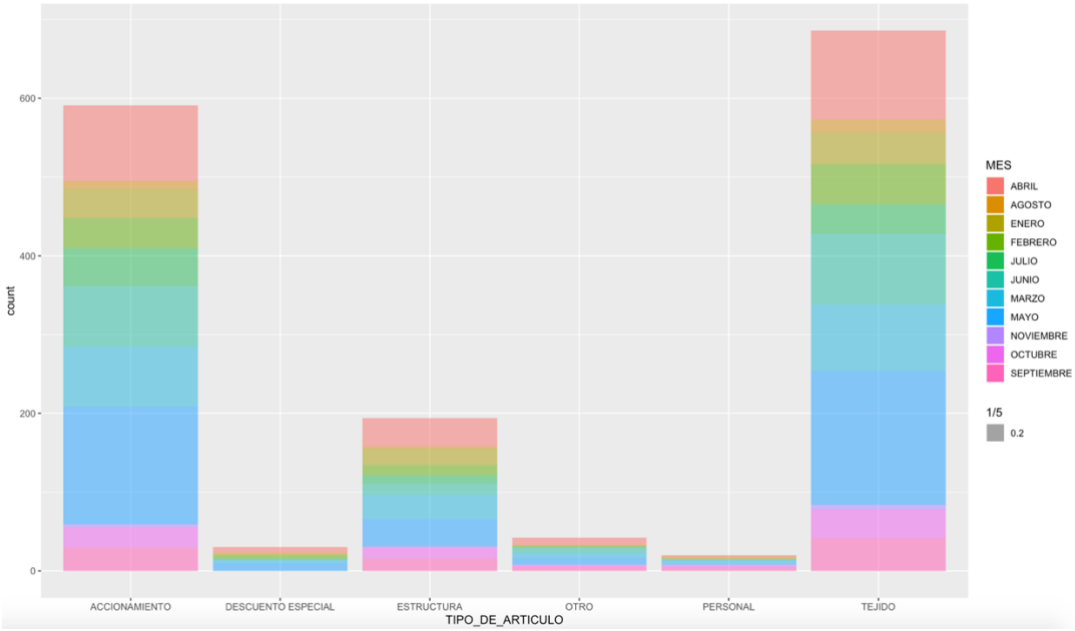
Ilustración 17: Gráfica de puntos que representa la variable TOTAL_PRECIO en función del MES y el TIPO DE ARTÍCULO



Fuente: Elaboración propia en Rstudio

Es en este momento cuando un mayor conocimiento de la herramienta permite obtener resultados mucho más interesantes. En este caso se utiliza también el paquete *ggplot* pero, a diferencia de la anterior gráfica, se utiliza la función *geom_bar* que en este caso sí es una potente herramienta de visualización de estos datos. El diagrama de barras (*Ilustración 18*), se centra en primera instancia en agrupar las observaciones en función de la variable TIPO_DE_ARTICULO. Si bien esto ya se había hecho con anterioridad, este nuevo diagrama de barras le da un nuevo enfoque utilizando los colores de las barras para representar de forma conjunta la variable MES. De esta forma también queda recogido cuándo se producen esas transacciones. En definitiva, nos permite mostrar más información y además es mucho más visual que el anterior diagrama de puntos.

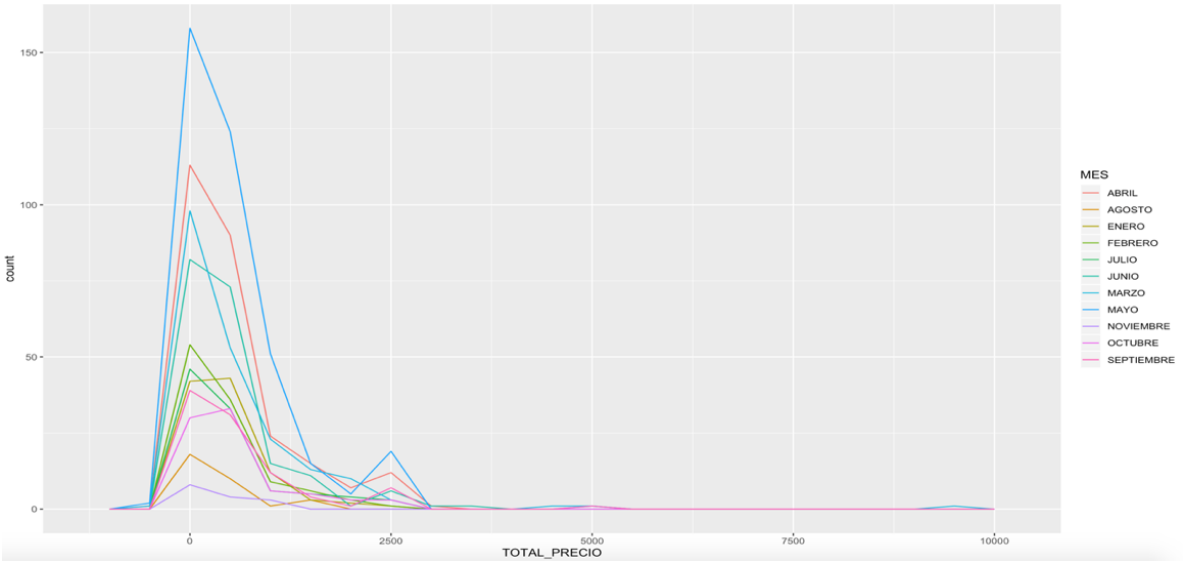
Ilustración 18: Diagrama de barras en función del tipo de articulo (meses en color)



Fuente: Elaboración propia en Rstudio

Otra posibilidad que nos ofrece el paquete *ggplot*, que como ya se ha podido comprobar en las demostraciones anteriores es una gran herramienta para la visualización de los datos, es la de crear un histograma que nos permite visualizar la variable TOTAL_PRECIO y la variable MES de forma conjunta obteniendo resultados con un potencial importante.

Ilustración 19: Gráfico lineal sobre la variable TOTAL_PRECIO y la variable MES



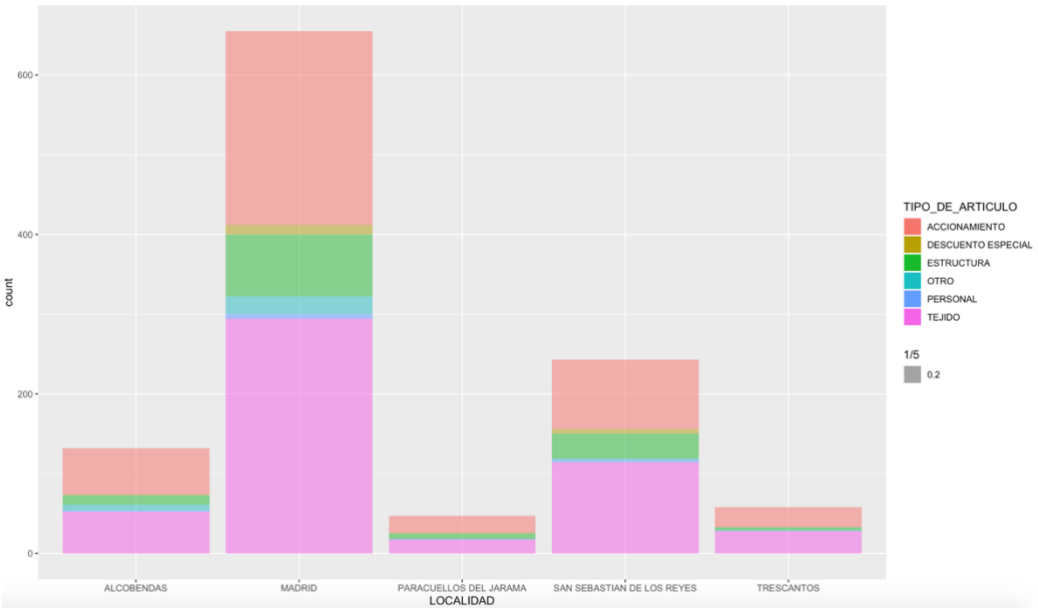
Fuente: Elaboración propia en Rstudio

Llegados a este punto surge un problema importante a la hora de representar datos en gráficas de las distintas localidades: hay demasiadas localidades en las que se han vendido productos dentro de la Comunidad de Madrid y el gráfico queda distorsionado. De esta forma, se ha creído oportuno generar una nueva variable únicamente con el propósito de poder obtener una representación gráfica de las localidades que mayor representatividad tienen en este Data Set.

Para ello se han seleccionado las cinco localidades que por volumen de transacciones eran las más importantes y, a continuación, se ha procedido a representarlas en diagramas de barras en función, primero, de la variable TIPO DE ARTICULO y, después, según el MES, representado en colores.

La *Ilustración 20* muestra cómo el tejido y los accionamientos son los artículos más comprados en todas las localidades importantes.

Ilustración 20: Gráfica de barras sobre localidades más importantes y TIPO DE ARTICULO

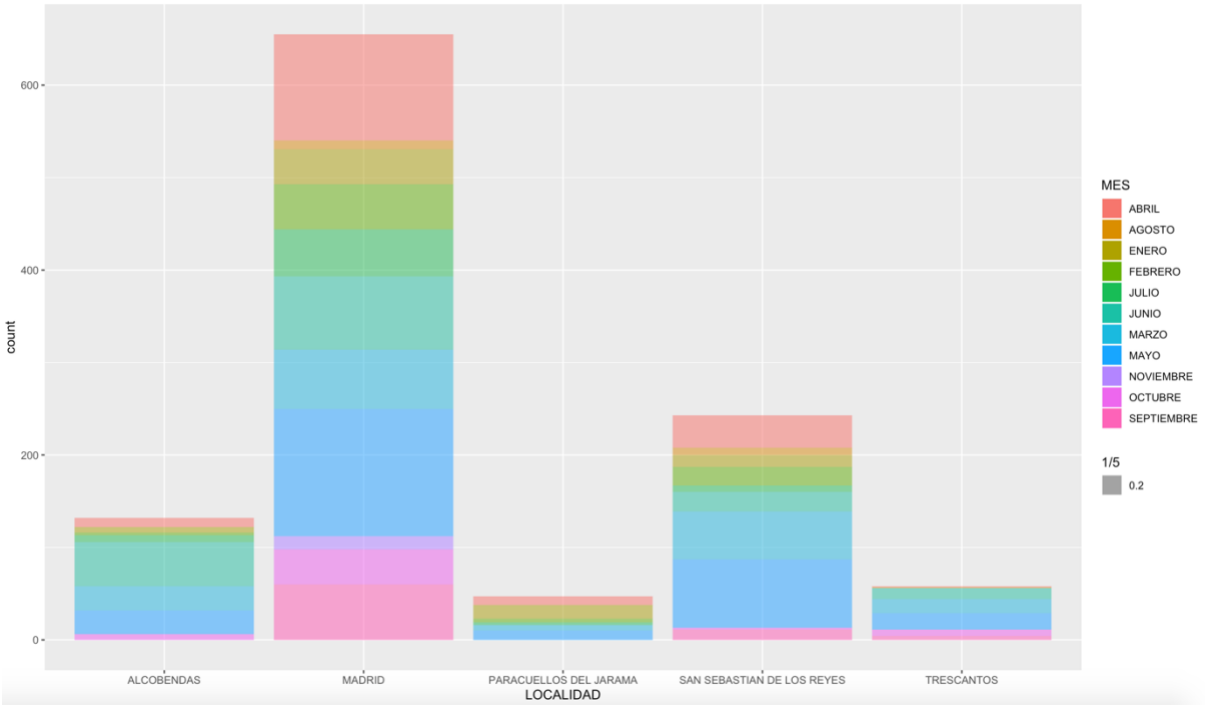


Fuente: Elaboración propia en Rstudio

A continuación, en la *Ilustración 21*, se utiliza un gráfico de barras muy similar para explorar si existe algún mes especialmente beneficioso, en cuanto a ventas se refiere, en las distintas

localidades con más compraventas. Más allá de confirmar la tendencia que demostraba que los meses de mayo y marzo erán los más importantes, no se extraen más conclusiones válidas.

Ilustración 21: Gráfico de barras relacionando las localidades más importantes y la variable MES



Fuente: Elaboración propia en Rstudio

3.3 Solución técnica

Tras todas las transformaciones, análisis y visualizaciones, ya tenemos un amplio conocimiento de los datos que permitirá obtener soluciones y recomendaciones no solo eficientes sino también con un fuerte respaldo en evidencias numéricas y visuales. Sin embargo, esto solo es el principio puesto que la parte fundamental de este estudio se situaba en dar respuesta a los diferentes problemas de negocio que ya mencionamos en el apartado correspondiente.

Para ello es necesario soluciones técnicas de alta calidad y precisión. A continuación, se pretende aplicar estas técnicas a través de las herramientas expuestas.

3.3.1 Análisis de *clustering*: alternativas factibles

El análisis de conglomerados (clusters) es una técnica estadística multivariante que busca agrupar elementos (o variables) tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos. Ahora bien, la realización de un análisis de este tipo no tiene una forma determinada, sino que existen múltiples soluciones para alcanzar un óptimo resultado.

Por esta razón, y como no se debería perder de vista el objetivo final de este trabajo que viene a ser demostrar la utilidad que tienen múltiples herramientas disruptivas en la generación de valor dentro de las PYMES, se van a explicar algunas de las posibilidades que existen para después centrarse en la solución que se ha considerado óptima. De esta forma, no solo mostraremos una solución para el caso *ad hoc* sino para muchas otras situaciones.

3.3.1.1 *Orange* como opción

En primer lugar, lo primero que hay que detectar es cómo está estructurado nuestro conjunto de datos porque en función de ello la solución será mas sencilla o, por lo contrario, más compleja. En este caso nos encontramos ante un conjunto de datos que nos aporta la dirección completa del comprador, su localidad y el precio total de la compra que realiza cada uno.

Pues bien, si quisiéramos encontrar una solución rápida y sencilla, pero a la vez representativa y potente, la primera recomendación sería utilizar el programa informático de minería de datos llamado Orange[2] una vez hubiéramos convertido las direcciones a coordenadas. Esta aplicación esta programada en Phyton, Cython, C y C++, pero tiene la gran ventaja de que tiene una interfaz muy simple y no es necesario tener conocimientos de programación para utilizarla. A modo meramente ejemplificativo, adjunto al trabajo una captura de su interfaz, así como el resultado que generó en un proyecto que pretendía realizar un análisis de clustering de los precios de los Airbnb en la ciudad de Nueva York del año 2019 (Castro & Castaño, 2019).

[2] Link al enlace de descarga https://orange.biolab.si.

En el Anexo 1 se encuentra la captura de la interfaz justo antes de ejecutar las órdenes para la creación del mapa con los puntos de cada vivienda de Airbnb, generando a posteriori diferentes áreas de influencia (o clústers) que son las que se pintan en el propio resultado. Y, en el Anexo 2, se puede observar el resultado que, como bien se aprecia, tiene un gran poder de visualización, mostrándonos los diferentes conglomerados que se formaban en la ciudad de Nueva York dependiendo del precio.

Si bien, este análisis hubiera sido sencillo, eficaz y muy rápido de realizar para nuestro estudio, existe un problema fundamental que se repite con asiduidad en el panorama nacional: apenas hay mapas físicos de España que puedan obtenerse de forma gratuita en Internet y, específicamente, en el caso de esta aplicación, no hay otra forma de utilizar esta herramienta por lo que no se puede utilizar para realizar el análisis. Sin embargo, si la empresa que se pretende analizar realiza pedidos a diversas partes del mundo como, por ejemplo, a Estados Unidos (país que cuenta con un sinfín de mapas gratuitos que se pueden obtener en Internet), esta herramienta sería muy útil y no requiere de una gran formación en aspectos de programación.

3.3.1.2 Georreferenciación mediante la API key de Google

El desarrollo de la georreferenciación de direcciones a través de las librerías *ggmaps* y *tidyverse* de Rstudio, volviendo a la herramienta principal del estudio, puede ser una solución muy eficaz para resolver los problemas de negocio. Ambas librerías se pueden descargar de forma gratuita, pero, para poder georreferenciar las direcciones necesitamos hacerlo a través de la API de google que es una herramienta de pago y requiere suscripción para conseguir la KEY[3], que se necesita para lograr el correcto funcionamiento de la función *mutate_geocade.* Esta opción es muy interesante para aquellas empresas que pretendan invertir una cantidad importante de dinero para obtener una mejor visualización de sus ventas y poder así extraer *insights* que

[3] Es un código alfanumérico que actúa como contraseña para identificarnos en los proyectos que deseamos realizar.

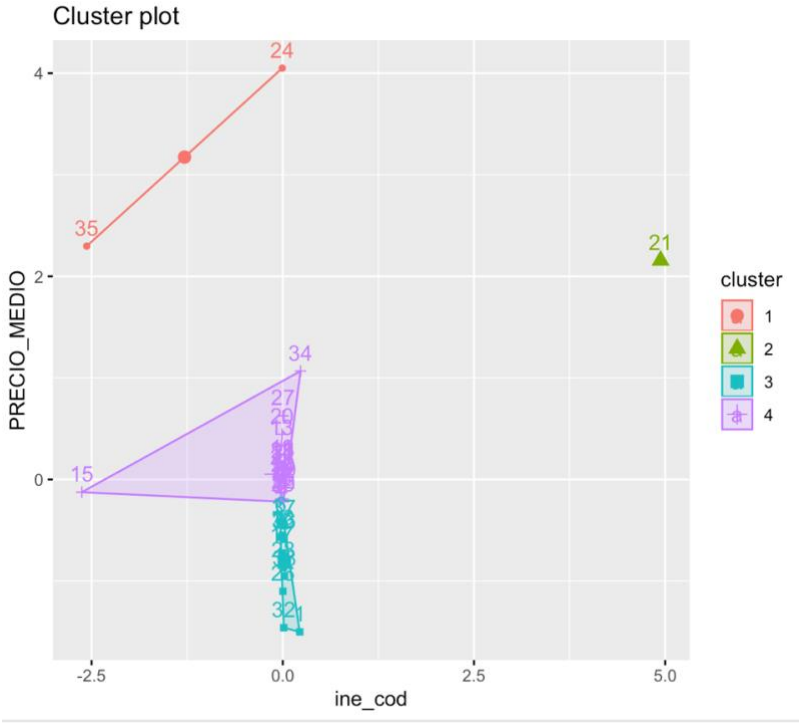
generen valor. Además, activar la suscripción en la mencionada plataforma de google supone ampliar el elenco de recursos para analizar grandes cantidades de datos. Sin embargo, dado que en este caso buscábamos una solución que acarrease el menor coste posible a las empresas y teniendo en cuenta que podemos tener el mismo problema que para el anterior caso si queremos representar el mapa de España con los diferentes *clusters*, no es la solución que hemos implementado para este caso, pero sí merece ser comentada dado su inmenso potencial.

3.3.1.3 Elbow criterion y Silhouette criterion

Llegamos ya a la tercera opción que proponemos para realizar el análisis de conglomerados a través de Rstudio. Este método se caracteriza porque nos desvela bastante información, pero no tiene tanto potencial visual como el último que mostraremos. Para realizar este análisis con el lenguaje de programción de R, se necesita descargar las librerías *cluster.datasets* y *factoextra.* Además, se necesita una estructura de datos específica para poder obtener resultados que resulten de utilidad. El formato de archivo elegido para realizar esta maniobra ha sido el de agrupar todas las observaciones por código postal y encontrar el precio de compra medio en cada una de esas localidades de tal forma que simplifiquemos mucho más el conjunto de datos que teníamos hasta este momento. La nueva composición del DataSet es de 39 observaciones dado que se han podido agrupar todas las localidades que aparecían en el conjunto de datos original en solo 39 códigos postales.

A partir de aquí comenzamos el análisis mediante las librerías mencionadas anteriormente. En primer lugar, escalamos las variables (para que sus valores sean comparables) y aplicamos la función *k-means* para poder observar el precio promedio de las operaciones de cada clúster. La representación gráfica de los clústeres formados se puede obtener de forma sencilla (véase la *Ilustración 22).* Aquí podemos apreciar la formación de cuatro clústeres que contienen 1, 2, 12 y 24 elementos respectivamente. Se han generado 4 clústeres porque en la propia función de *k-means* hemos introducido este parámetro (4) para hacer una primera observación. Después de comprobar la cantidad óptima de clústeres mediante el uso de dos criterios (el *Elbow criterion* y el *Silhouette criterion),* obtendremos el nuevo valor a introducir en la función.

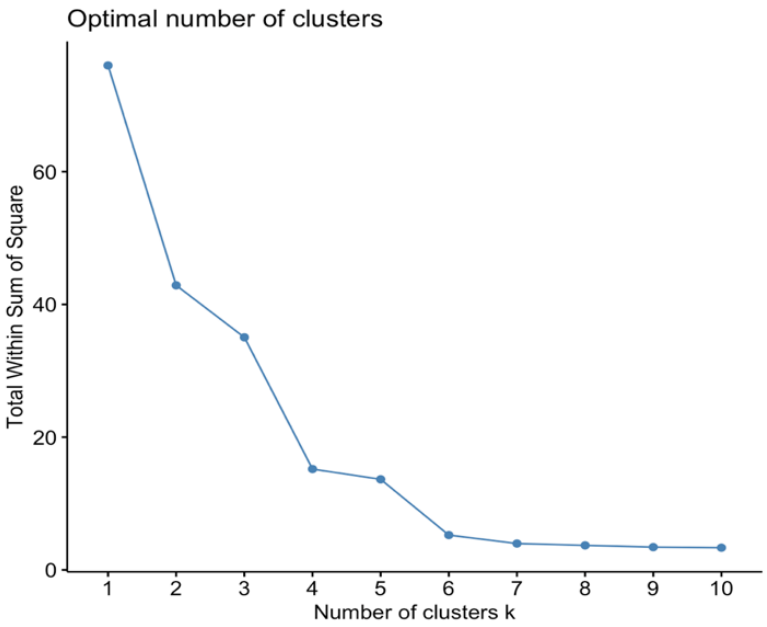
Ilustración 22: Clusterización a través de la función k-means



Fuente: Elaboración propia en Rstudio

Respecto del primero de estos criterios, el método Elbow, busca el número ideal de clúster a través de la optimización de la WCSS ("*within clusters sum of square*)". Se puede observar en la *Ilustración 23* cómo a medida que se aumenta el número de clúster disminuye el valor de WCSS. Se escoge el punto en que se dejan de producir grandes variaciones del valor de WCSS al aumentar K. En este caso, las opciones serían 2 ó 4 clústeres, que es donde la gráfica lineal coge forma de "codo". De aquí el nombre del método.

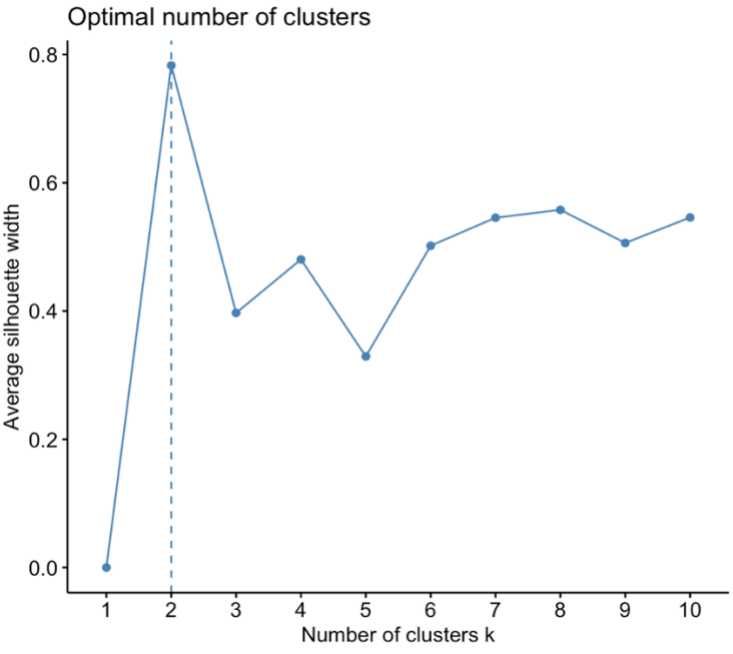
Ilustración 23: Cantidad óptima de clusters a través del método Elbow



Fuente: Elaboración propia en Rstudio

Por otro lado, el segundo de estos criterios para decidir el número óptimo de clústeres de este conjunto de datos, conocido como *Silhouette criterion*, arroja los resultados que se pueden apreciar en la *Ilustración 24*. El método de la silueta nos indica que el número óptimo de clústeres es 2 (aunque tambien se podría tener en consideración dibujar 4 clústeres por la forma de la gráfica).

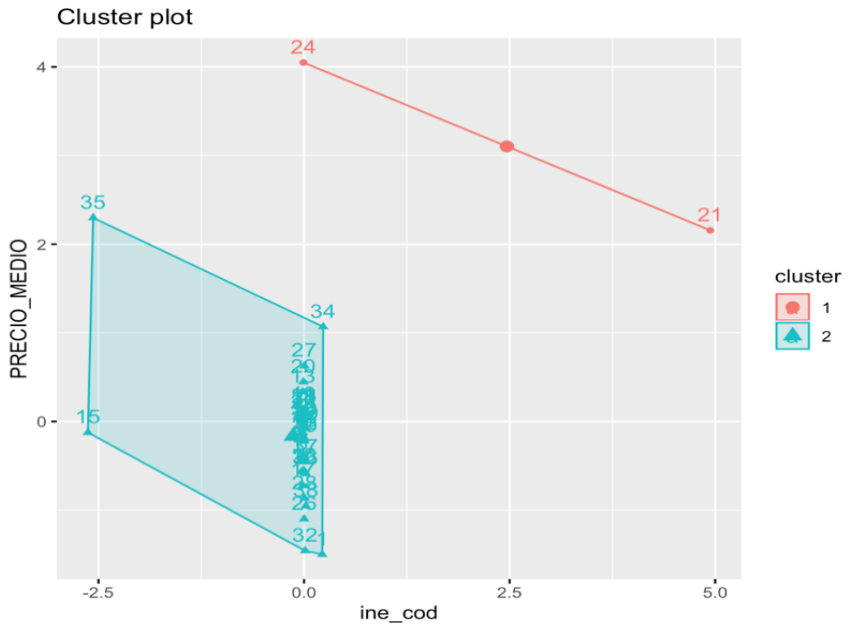
Ilustración 24: Cantidad óptima de clusters a través del método Silhouette



Fuente: Elaboración propia en Rstudio

En conclusión, introduciríamos en la función *kmeans* únicamente 2 clústeres para observar cómo sería la nueva clasificación de los precios medios por código postal.

Ilustración 25: Clusterización óptima a través de la función k-means



Fuente: Elaboración propia en Rstudio

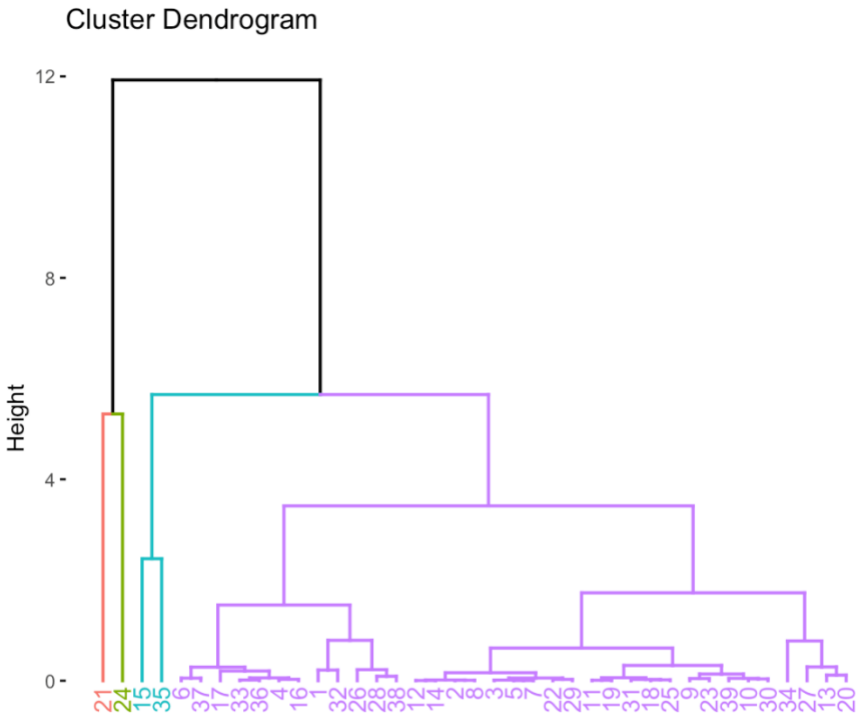
Así quedaría la nueva representación (*Ilustración 25*) con el cambio del parámetro "número de clústeres". Tal y como se puede observar, es una gráfica que, si bien sí nos aporta información acerca de los diferentes grupos que se pueden formar, entendemos que, de cara a organizar esfuerzos para la compañía, aporta menos valor que agruparlo en torno a 4 clústeres como se había hecho en un primer momento. Este es un ejemplo de cómo no siempre aportan valor los procesos de comprobación y análisis que se realizan.

3.3.1.4 Hierarchical clustering

Otro método que suele ser bastante útil para realizar este tipo de análisis es el de *hierarchical clustering*. Este método de clusterización trata de encontrar jerarquías en el *Data Sett* principal, generando grupos basados en la semejanza de esos datos (Johnson, 1967). En este caso, las

agrupaciones se generarían en torno a los precios medios de las transacciones que se realizan en función del código postal por el que se están examinando. Aquí también es necesario usar la función *scale* para escalar las variables y hacerlas comparables antes de acometer el análisis. Para observarlo de manera visual se utiliza el *dendrograma* que ayuda a decidir el número de grupos que podrían representar mejor los datos teniendo en cuenta cómo se van anidando los clústeres.

Ilustración 26: Dendrograma de clusterización por jerarquías



Fuente: Elaboración propia en Rstudio

El resultado sería el diagrama que aparece en la *Ilustración 26*. De nuevo puede aportar información útil pero no es el pretendido para este sector y caso en concreto. Nos confirma, eso sí, las sospechas de que la agrupación en torno a 4 clúster es francamente más acertada puesto que nos permite dividir y homogeneizar, en mayor medida, los grupos analizados.

3.3.2 Análisis de *clustering:* elección *ad hoc*

Llegados a este punto, y tras múltiples intentos que, o bien, no se han podido realizar por problemas técnicos (como la escasez de mapas dibujados por coordenadas), o bien, se han desechado por considerar que no aportaban el valor que se requería, es momento de comenzar con el análisis clustering más eficiente para este complejo caso.

El objetivo de este análisis pretende ser el de clasificar los municipios -gracias a la agrupación por código postal y precio medio de las transacciones que se producen en esas localidades- según categorías de interés para la empresa de tal manera que podamos aconsejar a la empresa la inversión o no en determinados municipios estratégicos. En realidad, se persigue realizar la misma función que un sistema de información geográfica[4] logrando un resultado mucho más explícito que con las famosas nubes de puntos.

Las librerías que se necesitan cargar en R para la creación del mapa interactivo que se pretende alcanzar son: *tidyverse, rgal, broom, wesanderson* y *leaflet*. Hay que recordar el problema al que nos habíamos enfrentado en ocasiones anteriores: apenas hay mapas físicos de España y, menos aún, de los municipios de la Comunidad de Madrid. La solución a este problema es que dibujemos los propios mapas nosotros mismos. Para ello hay que obtener los datos espaciales que nos permitan obtener las coordenadas de las fronteras de los municipios de nuestro mapa. Los conocidos como *shapefiles*[5] se pueden encontrar en internet y son bastante precisos y eficaces a la hora de dibujar mapas. Una vez hemos obtenido estos datos[6] podemos proseguir

[4] Conjunto de herramientas que integra y relaciona diversos componentes que permiten la organización, manipulación, análisis y modelación de cualquier tipo de información geográfica referenciada y asociada a un territorio.

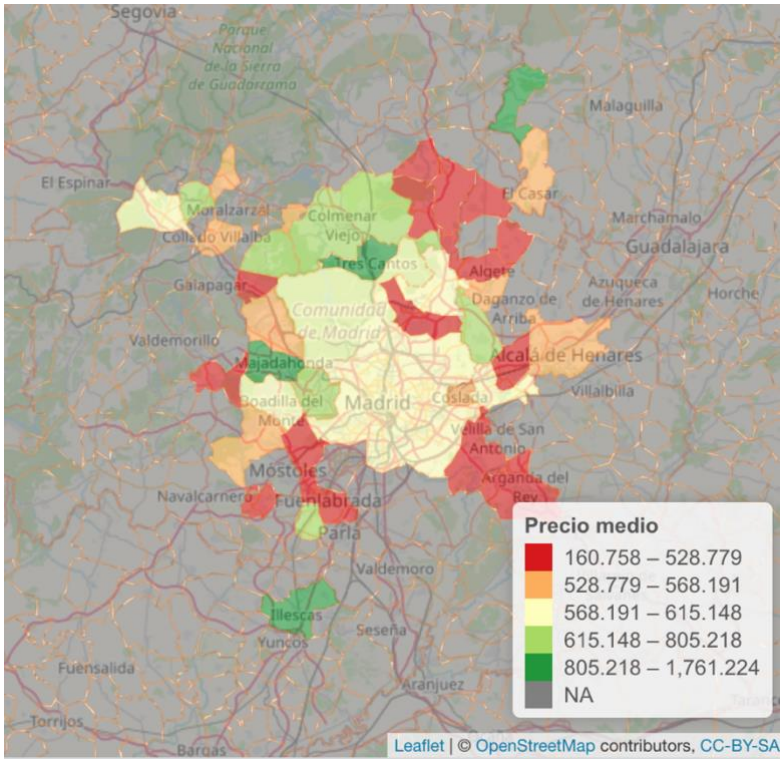
[5] Archivos que contienen información de datos espaciales

[6] Datos obtenidos en https://opendata.esri.es/datasets/53229f5912e04f1ba6dddb70a5abeb72_0, página totalmente gratuita.

y dibujar los mapas mediante las funciones *ggplot* y *geom_path*. Tras ello, unimos ambos Data Sets. Por un lado, el que contiene las coordenadas de cada municipio junto con el código postal correspondiente y, por otro lado, el que contiene la agrupación de las transacciones por código postal junto con su precio medio. Mientras tanto, creamos la paleta visual de colores y el formato de la etiqueta correspondiente.

Por último, ya solo queda configurar la imagen del mapa y dejar a la función *leaflet* hacer el resto. El resultado es el mapa que aparece en la *Ilustración 27*.

Ilustración 27: Análisis de conglomerados de los municipios de la Comunidad de Madrid



Fuente: Elaboración propia en Rstudio

3.3.2.1 Resultados y primeras conclusiones del análisis

Este mapa interactivo otorga una potente visualización de sus ventas a la empresa, por lo que se puede concluir que ha sido exitoso en múltiples aspectos:

- La geolocalización física de las ventas se ha generado con suficiencia. En este sentido, no importaba tanto cada venta física individualizada que se producía en cada municipio, ya que esto hubiera sido representado por una nube de puntos, sino lo importante aquí

era obtener una estimación del precio medio que cada municipio está dispuesto a pagar por venta, algo que se ha obtenido gracias a la representación conjunta de las transacciones mediante áreas de precios que coinciden con los municipios representados.

- La creación de hasta 5 conglomerados según el rango de precio medio de las transacciones que se producen en ese municipio. Este era el principal objetivo del *análisis clustering* en un primer momento, poder determinar qué zonas son más atractivas y, por lo tanto, dónde merece la pena centrar los esfuerzos de marketing o ventas respecto de las que, por el bajo precio medio de sus transacciones, no son tan recomendables.

  A este respecto, la zona este de la Comunidad de Madrid se podría configurar como el Clúster 1. Este conglomerado sería el menos atractivo para la venta de toldos.

  La zona suroeste de la Comunidad de Madrid recibe el nombre de Clúster 2. Se trata de una zona más atractiva que el Clúster 1 con algún municipio que registra precios medios bastante altos. Sin embargo, esa tendencia se entrelaza con municipios con un rango de precios poco atractivo. Es un área que podría tener cierto margen de mejora en un futuro próximo.

  Por último, la zona noroeste recibiría el nombre de Clúster 3. Se trata de la zona donde las transacciones realizadas por la empresa que estamos analizando tienen un rango de precios más alto. Es el área que más interesa a la empresa en términos de rentabilidad por transacción. Consecuentemente, sería la primera área donde la empresa debería centrar sus esfuerzos de marketing. Aumentar las ventas en esa área y consolidar la posición en esa zona sería un movimiento estratégico que aportaría múltiples beneficios a la compañía, tal y como muestra el *análisis clustering* desarrollado.

- El examen de correlación entre más cercanía a la tienda física y operaciones más rentables, desde el punto de vista de precios más altos, se ha probado inexistente. Si bien se había demostrado cómo la cercanía a la tienda física suponía un incremento considerable en el número de ventas, como era el caso de San Sebastian de los Reyes, no se ha podido probar que haya ninguna relación entre cercanía y mayor gasto por cliente.

- Asímismo, descartamos otras posibles ideas de análsis como examinar las mismas variables por meses para ver si existía diferencia temporal. Tras realizar el oportuno análisis no se ha descubierto ninguna correlación distinta a la ya obtenida en el párrafo anterior.

- De cara a concretar nuevas líneas de investigación para el futuro, se abre la posibilidad de aumentar el mapa físico a nivel nacional (por municipio o por CCAA) para tener una visualización de todas las ventas en el territorio nacional y determinar qué zonas son más atractivas para abrir próximas tiendas físicas o centrar esfuerzos del equipo comercial y de marketing. Actualmente no es posible dado que los datos se circunscriben a la Comunidad de Madrid y alrededores.

- Además, se plantean nuevas líneas de análisis que se podrían llevar a la práctica a raíz de este: Entremezclar un mapa que examine la renta per cápita del cada municipio junto con el precio medio pagado por cada transacción del producto analizado. Todo ello para observar, todavía con mayor detalle y certeza, que áreas están todavía por explotar y cuáles se están explotando correctamente.

- Finalmente, otra idea que coge mucho peso es la posibilidad de utilizar este análisis de forma anual para comparar los datos con los resultados que se obtengan tras el paso del tiempo. De esta forma, se podrán inferir tendencias hasta ahora desconocidas que los números no desvelan con la misma facilidad que una potente visualización.

3.3.3 Creación de modelos predictivos

Todo lo que se ha demostrado hasta ahora es ya una evidencia irrefutable de cómo los negocios más tradicionales, y también los más innovadores, pueden generar valor a través del uso de las nuevas herramientas de análisis de datos independientemente de su tamaño. No obstante, dado que este trabajo pretende mostrar las grandes capacidades y utilidades de las nuevas herramientas de análisis de datos, se decidió proseguir con el proyecto. En este caso, pretendimos examinar la posibilidad de acometer la creación de un modelo predictivo a través de la plataforma Microsoft Azure[7]. Si bien es cierto que esta plataforma tiene, como norma general, un coste relativamente bajo, los estudiantes de la Universidad Pontificia Comillas cuentan con la posibilidad de utilizar esta plataforma de forma gratuita por lo que entendimos que era una oportunidad magnífica.

Antes de entrar en los aspectos más prácticos de este análisis, es fundamental establecer un marco teórico para entender las posibilidades de esta herramienta. El análisis predictivo es un área de la minería de datos que consiste en la extracción de información existente en los datos y su utilización para predecir tendencias y patrones de comportamiento (Nyce, 2007). Para establecer y explotar las posibles relaciones y adelantarse a lo que pueda suceder en el futuro es indispensable la creación de un modelo predictivo.

Un modelo predictivo tiene por objetivo evaluar la probabilidad de que un determinado suceso se pueda repetir y a raíz de ello, poder adelantarse en la toma de decisiones. El proceso para crear un modelo predictivo, gracias a los avances en *machine learning* y *deep learning*, no es excesivamente costoso, ni tampoco complejo debido a aplicaciones muy intuitivas como Microsoft Azure. Los pasos consisten en la definición del objetivo, la preparación de los datos, la evaluación de algoritmos, el entrenamiento del modelo y la medición del modelo.

[7] Link a la plataforma de servicios de Microsoft Azure: https://studio.azureml.net/Home/ViewWorkspaceCached/a4de3fe547a84ed3b8fc99fda9930195?#Workspace/Experiments/ListExperiments

3.3.3.1 Aspectos previos de gran relevancia.

Para el caso que nos ocupa, hay una dicotomía evidente respecto de los objetivos a alcanzar. Por un lado, podríamos fijar como objetivo la averiguación del precio objetivo para cada producto. La pregunta a la que intentaría dar solución el modelo sería algo así como: ¿Cuánto estarían dispuestos a pagar como máximo mis clientes por cada uno de los productos que ofrezco dependiendo de variables como localización, época del año o cantidad de producto comprado? En otras palabras, intentar optimizar el precio de venta de tal forma que se maximicen los beneficios.

Por otro lado, otra opción a considerar como objetivo a fijar para el estudio -totalmente independiente de la anterior- podría ser la predicción del tipo de producto que los clientes estarían interesados en comprar. A tal efecto, el objetivo sería responder a esta pregunta: ¿Qué tipo de producto va a estar interesado en comprar un cliente dependiendo de la zona donde viva, el precio o la época del año? Esto sería de gran utilidad porque en un negocio como el del montaje de toldos, la logística y los centros de distribución son muy importantes debido a la complejidad para realizar cada montaje, como si de una obra arquitectónica se tratara en múltiples ocasiones. La capacidad de predecir los pedidos de los clientes sería una ventaja competitiva que abarataría costes y, en definitiva, generaría un alto valor añadido al negocio.

Sin embargo, y pese a que los estudios anteriores eran bastante prometedores, después de hablar con la dirección de la empresa e intentar crear los modelos correspondientes para obtener los resultados esperados mediante la plataforma de Microsoft Azure, el desenlace del análisis no fue el esperado y los resultados obtenidos se clasificarían como no válidos. Las razones detrás de semejante revés fueron variadas.

En primer lugar, las técnicas de modelización no tienen en cuenta la existencia de otras variables que difícilmente se pueden medir en un negocio eminentemente familiar y de trato muy cercano. La amistad con el comprador, el compromiso como consecuencia de pedidos anteriores, la publicidad boca a boca tan importante en este negocio que ya deja establecido el precio o el trato desigual dependiendo del poder adquisitivo que se le atribuye a un particular

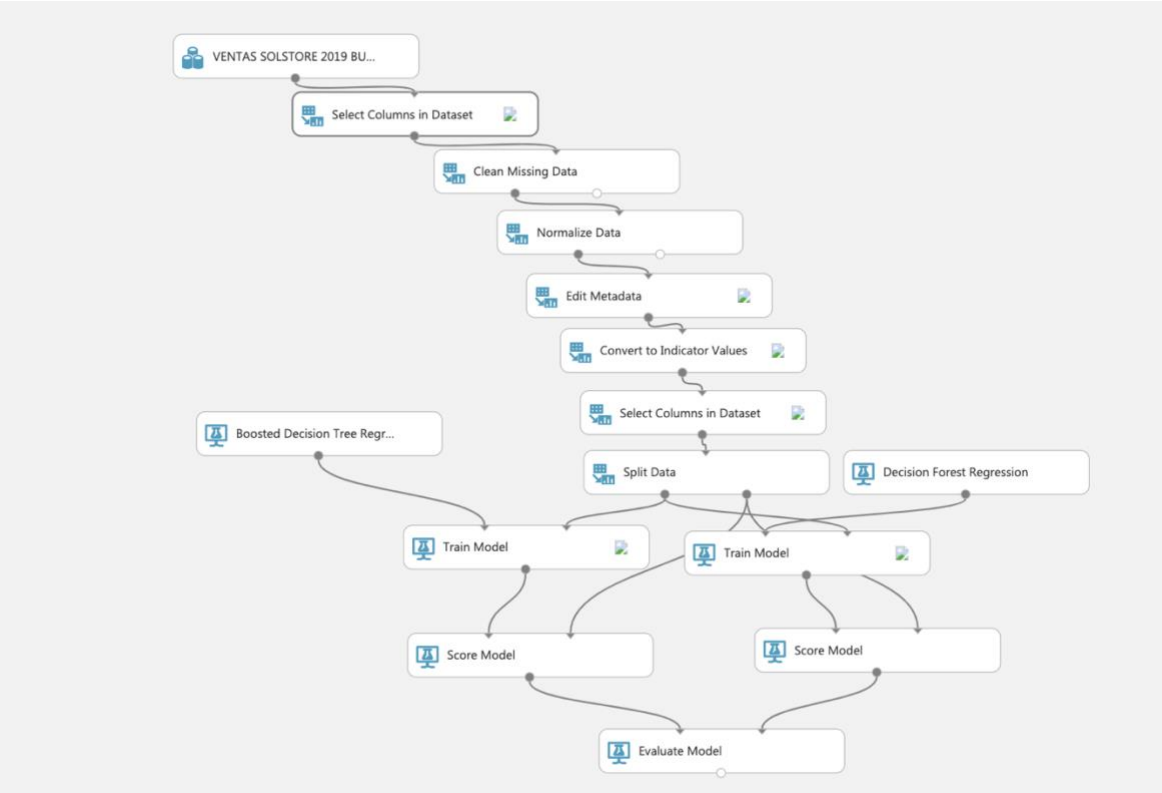
o a una u otra empresa, dependiendo de factores difícilmente cuantificables, hacen excesivamente laborioso poder establecer un precio de venta óptimo a través de un modelo que carece del número suficiente de variables posibles para obtenerlo. Así pues, los criterios que se manejan en muchas ocasiones son de carácter subjetivo lo que en último término hace inviable la creación de un modelo predictivo.

En segundo lugar, se había planteado la posibilidad de predecir el tipo de productos que nuestros potenciales clientes iban a estar interesados en adquirir. Sin embargo, después de conversaciones con el jefe de ventas, entendimos que esta idea, si bien es perfectamente aplicable a otro tipo de PYMES, no era factible en el marco de ventas para una empresa de toldos. Esto se debía a que si bien podíamos categorizar los diferentes productos que la empresa vende en varios apartados, como de hecho se hizo para obtener mejores visualizaciones, la realidad es que el único producto que se vende aquí son toldos, si bien, el catálogo de productos es excesivamente amplio. En este sentido, la compañía tiene casi 600 variaciones de toldos y todos ellos dependen de forma singular del tipo de vivienda, las dimensiones de la vivienda o los gustos del consumidor. En otras palabras, no es posible adelantarse al producto en el que estará interesado el cliente porque ese producto se ajusta tanto al consumidor que inevitablemente no se puede predecir con las variables que contiene el DataSet.

En tercer y último lugar, la realidad es que este tipo de análisis no siempre son exitosos, sino que depende del tipo de negocio y sus características, así como de la información de la que se dispone.

A pesar de que la creación del modelo predictivo no fue exitosa, a continuación se pretende, a meros efectos académicos, explicar muy brevemente cómo se llevo a cabo este análisis. En la Ilustración 28 aparece representado el esquema requerido para generar el modelo predictivo a través de la plataforma de Microsot Azure. En este supuesto, tenemos como objetivo el primero de los que habíamos mencionado: predecir el precio de venta óptimo dependiendo de las variables explicativas.

Ilustración 28: Extracto de ejemplo del modelo predictive con Microsoft Azure



Fuente: Elaboración propia en Microsoft Azure

3.3.3.2 Preparación y definición de las características

Para comenzar, incorporamos el conjunto de datos con el que queremos trabajar. Para ello podríamos utilizar el Data Set original pero dado que ya hemos limpiado y corregido ese Data Set, añadiendo otras nuevas variables gracias a las operaciones que hicimos con R, utilizamos directamente el Data Set modificado que obtenemos de R.

El siguiente paso sería seleccionar las columnas del conjunto de datos que queremos incluir para la creación del modelo. En este sentido, cuanto mayor correlación exista entre la variable *precio*, que pretendemos predecir, y el resto de variables explicativas, mejor. En este caso, seleccionamos las variables LOCALIDAD, TIPO_DE_ARTICULO, DTO, MES y ARTICULO.

Como parece lógico, a la hora de realizar el modelo predictivo es imprescindible disponer de un conjunto de datos sin valores extraños, si bien ya habíamos tratado los datos con Rstudio para dejarlos listos para su manipulación, siempre es importante ser precavido y por eso

utilizamos la opción de eliminar todos los valores ausentes para asegurarnos de que el conjunto de datos estaba bien. El siguiente paso es normalizar los datos que tenemos de tal forma que rescatamos los datos numéricos para restringir los valores de los conjuntos de datos a un rango estándar. De este modo, al usar una escala común, no distorsionamos la información del Data Set.

Llegados a este punto, incorporamos dos funcionalidades que, si bien no es necesario utilizar en este experimento, sí que pueden tener gran utilidad en situaciones similares. Hablamos de *Edit Metadata* y *Convert to Indicator Values*. La primera de ellas permite transformar valores que podrían ser interpretados como números en categorías[8], esto es necesario para que el modelo pueda funcionar. Por otro lado, la segunda de las funcionalidades permite convertir la columna que contenía una variable categórica en diversas columnas que contienen valores booleanos (0 y 1).

3.3.3.3 Construcción del modelo

La construcción del modelo continúa con la que probablemente es la decisión más complicada e importante hasta este punto, escoger qué algoritmos vamos a utilizar para entrenar el modelo. Es una decisión determinante porque dependiendo de cuán óptimo sea el algoritmo elegido, mejor serán los resultados que arrojará el experimento. A este respecto, los algoritmos de regresión fueron una elección clara desde el primer momento dado que son muy útiles para predecir números en concreto. Ahora bien, la decisión de qué tipo de algoritmo de regresión elegir no es sencilla. *Decision Forest Regression* y *Boosted Decision Tree Regression* fueron las opciones por las que nos decantamos en última instancia. Sin embargo, para otros casos pueden no ser los óptimos y lo ideal sería entrenar al modelo con todos los algoritmos de regresión posibles para encontrar la opción óptima. Para entrenar los modelos, probarlos y compararlos, hemos dividido el dataset en dos conjuntos de datos diferentes, uno de entrenamiento y otro de prueba con la realidad a través de la sección *Split Data*. En este módulo hemos establecido un 80% de los datos para entrenar el modelo y el 20% restante para probarlo.

[8] p. ej. Una variable que tiene 0, 1 y 2 pero que en realidad significa “Venta en día soleado”, “Venta en día nublado” y “Venta en día lluvioso”, respectivamente.

Siguiendo con el experimento, hemos entrenado los modelos usando esta división de datos con el módulo de *Train model*, en el cual es imprescindible establecer qué variable queremos predecir, en nuestro caso el precio, y el módulo de *Score model*.

En última instancia, hemos realizado un test de calidad de los resultados obtenidos a través de la sección *Evaluate Model*. Esto nos permitía comparar la aplicación de ambos logaritmos en el modelo.

Como avanzamos al inicio de este apartado, la evaluación de resultados no son determinantes debido a las razones que ya expusimos. Si bien, la comparación de los resultados se hubiera realizado a través del error[9] y el coeficiente de determinación[10].

3.3.4 Otra opción: Análisis de *cross selling*

El análisis de crosseling es una técnica de marketing utilizada hace ya varias décadas y consiste en "*[…] ofrecer y sugerir productos complementarios a los que el cliente consume o pretende consumir*" (Mondeja Jimenez, 2014). Este análisis depende de tres aspectos relacionados entre sí que serán observados al detalle: *"vender en el campo de una necesidad desvinculada, vender hacia arriba o hacia abajo y vender una solución integrada"* (Harding, 2004).

No obstante, a pesar de que lleva muchos años en funcionamiento, las técnicas de *cross selling* son mucho más eficientes y potentes en la actualidad. Gracias a las nuevas herramientas que venimos alabando durante este trabajo, podemos determinar qué combinación ganadora de productos genera más ventas en su tienda física o en el catálogo de su tienda online. Esto repercute directamente en la distribución del catálogo, ofertas y descuentos o, incluso, en las campañas de marketing.

Para la empresa analizada, esta técnica no genera valor añadido por el simple hecho de que solo comercializan un producto (los toldos), aunque vendan los elementos que conforman al

[9] Representa la diferencia entre el valor real y el valor predicho.

[10] Medida de como se adapta el modelo a los datos. Cuanto más cercano a 1 sea la proporción, mejor predicción será la del modelo.

producto (estructura, tejido o accionamiento) por separado. Además, el producto en sí mismo tiene muchas variaciones. Esto implica que no puedes ofrecer productos complementarios como tal, sino que la persona comprará el toldo que le guste y los accesorios serán de carácter necesario para su funcionamiento.

Sin perjuicio de lo explicado, esta técnica debe ser tenida en cuenta por la compañía para futuros análisis porque en el caso de que se planteen ampliar su gama de productos, vendiendo, por ejemplo, piscinas portátiles o muebles para la decoración de esas zonas, esta herramienta podrá ser de mucha utilidad.

La realización del análisis se podría hacer sin problemas mediante R, utilizando las librerías de *skimr, arules, arulesViz, ggplot2, tibble, plyr, dplyr, rstudioapi, igraph* y *lubridate.*

3.4 Conclusiones del estudio

Del exhaustivo análisis de este conjunto de datos, que representaba cerca de 1.600 ventas de componentes de toldos realizados por la empresa Toldos la Estrella en la Comunidad de Madrid y alrededores, podemos extraer las siguientes conclusiones:

- Apenas reducimos el número de observaciones tras la correcta limpieza de datos por lo que la calidad de extracción de datos es muy alta. Este hecho es encomiable y muy positivo para poder confirmar los resultados del estudio.

- No existe correlación positiva fuerte entre las variables examinadas.

- Tras la realización del análisis exploratorio, los grandes descubrimientos fueron los siguientes: el mes con más número de ventas fue septiembre, el mes con el precio medio más alto por pedido fue mayo, Tres Cantos es el municipio con el precio medio más alto por pedido y el operador suena 15/17 IO se consolidó como el producto estrella al ser el más vendido. Si bien es cierto que las tablas son de gran utilidad para extraer conclusiones, el poder de las nuevas herramientas de visualización hace que sea todavía más fácil y sencilla esa tarea.
- El análisis de *clustering* realizado nos facilitó la creación de hasta 5 conglomerados según los rangos de precio. A este respecto, la zona este de la Comunidad de Madrid se

podría configurar como el Clúster 1. Este conglomerado sería el menos atractivo para la venta de toldos.

La zona suroeste de la Comunidad de Madrid recibe el nombre de Clúster 2. Se trata de una zona más atractiva que el Clúster 1 con algún municipio que registra precios medios bastante altos. Sin embargo, esa tendencia se entrelaza con municipios con un rango de precios poco atractivo. Es un área que podría tener cierto margen de mejora en un futuro próximo.

Por último, la zona noroeste recibiría el nombre de Clúster 3. Se trata de la zona donde las transacciones realizadas por la empresa que estamos analizando tienen un rango de precios más alto. Es el área que más interesa a la empresa en términos de rentabilidad por transacción. Consecuentemente, sería la primera área donde la empresa debería centrar sus esfuerzos de marketing. Aumentar las ventas en esa área y consolidar la posición en esa zona es clave en el mercado de toldos, tal y como muestra el análisis *clustering* desarrollado.

Esta herramienta es de gran utilidad tanto para las estrategias de precio a corto plazo, como también para desarrollar una estrategia de crecimiento a largo plazo. Por lo tanto, la visualización, en el mapa físico del territorio español, de las diferentes zonas según el rango de precios pagados por los clientes, nos podrán dar ideas de gran calado para determinar los precios óptimos de cada zona que permitan maximizar beneficios y para tomar decisiones acerca del crecimiento a nuevos mercados geográficos por parte de la empresa.

En este mismo sentido, un análisis de este tipo también se puede utilizar para examinar la evolución de las ventas a lo largo del año en las diversas zonas y así obtener patrones de consumo que generen *insights* útiles para la organización y estrategia en los niveles más altos de dirección.

- Si bien el análisis de *cross selling* no es un alternativa factible para este caso, la construcción de un modelo predictivo sí es una posibilidad a tener en cuenta en el largo plazo porque según se vaya aumentando el tamaño de la empresa y su área de influencia, las ventas se expandirán por toda la península y será de gran utilidad tener un modelo que puede predecir el precio de venta óptimo o, incluso, adelantarse a las necesidades de los potenciales clientes de modo que la logística sea más eficiente reduciendo el tiempo de espera para los clientes y abaratando costes de almacenamiento -entre otros- para la empresa. Sin embargo, para que funcione correctamente el modelo,

es necesario incluir nuevas variables que puedan explicar de forma más fehaciente las variables PRECIO y ARTICULO. A este respecto, algunas variables que debieron ser observadas son:

- Satisfacción del cliente (en una escala del 1 al 10).
- Medio de promoción que atrajo al cliente.
- Clima del municipio.
- Población del municipio.
- Decisión impulsada por la emergencia climática (el uso de toldos reduce significativamente el consumo de aire acondicionado).
- Dimensiones de la terraza o lugar donde se instaló el toldo (en m^2).
- Número de habitantes de la vivienda.
- Precio del m^2 de ese municipio o barrio.
- El color de los toldos vendidos.

Sin perjuicio de lo ya comentado, existen otras variables difícilmente cuantificables como la valoración personal que da el cliente a su hogar y por consiguiente el dinero que quiere invertir en él u otras desconocidas como la posibilidad de que se origine una pandemia mundial. Esta realidad hace que sea inviable crear un modelo predictivo o extraer unas conclusiones que no sean otra cosa distinta que una simplificación de la realidad con mayor o menor precisión.

En definitiva, podemos concluir que ha sido un estudio exitoso donde hemos obtenido información que, sin duda, generará valor a la compañía tanto en el corto plazo como también en proyectos o estrategias a largo plazo. Pese a ello, la obtención de más información por parte de la empresa vendedora sería favorable para una mayor eficiencia en el análisis.

3.5 Solución al problema de negocio.

3.5.1 Acciones.

Las acciones que recomendamos a la empresa, a raíz de los resultados arrojados por el estudio, para resolver sus problemas de negocio son:

Un buen punto de partida sería continuar con la recomendación en lo que a la recolecta de nueva información se refiere. La medición de otras variables explicativas que podrían correlacionarse de forma más severa con las variables examinadas ayudaría a obtener resultados más precisos. Esta nueva información podrá ser adquirida por diferentes canales. En primer lugar, la extracción de datos mediante técnicas de "*web scrapping*" (que podría servir para la obtención del precio del m^2 en las diferentes viviendas) es hoy una de las técnicas más útiles para obtener información regularmente. Otra opción es acudir a los medios tradicionales de investigación como pudiera ser los planes urbanísticos y los catastros que existan en los diversos municipios (para extraer los m^2 de las terrazas donde se hagan instalaciones). Por último, otro canal efectivo que permite obtener información es el cuestionario a clientes tras la finalización de la compra (para medir la satisfacción con la compra).

Todas estas técnicas tendrían como finalidad primordial la creación de una base de datos interna (cuanto más grande mejor) que permita ampliar las conclusiones alcanzadas, así como la fiabilidad de estas. Para ello, sería necesario la contratación de un pequeño equipo que se encargue de estas labores diariamente de tal forma que el *core business* pueda seguir realizando su actividad principal con la eficiencia necesaria.

A raíz de lo anterior, es necesario una acción anual de actualización de datos que permita observar en tiempo real los gustos y preferencias de los clientes, así como, el nivel de precios aceptable por estos. El carácter anual de la actualización se debe a que es un mercado que cambia de forma regular cada 2- 3 años como consecuencia del surgimiento de nuevas modas y tendencias. Por este hecho es recomendable que para futuros estudios se realicen los análisis con los datos históricos de, al menos, los últimos dos años.

Por último, el análisis de *clustering* realizado a través del mapa físico del territorio español nos muestra cuáles deben ser los municipios donde volquemos más esfuerzos promocionales según el rango de precios donde sea más interesante operar para la empresa. En términos de estrategias a largo plazo, nuestra recomendación es utilizar el mapa físico por rango de precios a nivel nacional para determinar qué zonas pueden ser más atractivas en función de variables como la renta per cápita dentro del municipio o el clima de la zona. La expansión del área de influencia de las empresas de este tamaño siempre es un reto y entraña dificultades, sin embargo, el análisis de estos datos incrementa considerablemente las posibilidades de éxito.

Del mismo modo, la obtención de las variables que determinamos anteriormente podría ser un paso fundamental para conseguir que los modelos predictivos funcionen y tengan un nivel de confianza suficientemente alto, así como un error pequeño para considerarse válidos.

3.5.2 Impacto.

Las recomendaciones realizadas requieren una inversión importante de tiempo y dinero para llevarlas a cabo. La construcción de una base de datos interna, junto con los procesos para reclutar a los profesionales que se encarguen de la recolección y análisis de datos diarios suponen un coste no solo en términos económicos sino también desde el punto de vista temporal. Es evidente que, si el mercado se encuentra en un momento favorable para el crecimiento y la expansión geográfica por toda la península, las propuestas de acción pueden ser un inconveniente ya que su implementación se podría demorar y, claro está, el mercado se autoregula solo y no espera a nadie. A su vez, habría que examinar las cuentas de la empresa para determinar cómo de saneadas están de cara a poder acometer inversiones de este tipo.

No obstante, el impacto de nuestras recomendaciones se estima muy positivo dado que la toma de decisiones respaldada por evidencias extraídas de los datos siempre será más precisa y adecuada a la situación. La objetivación de toma de decisiones, que por supuesto siempre tendrán un componente subjetivo, es fundamental para el buen funcionamiento de toda actividad empresarial. Por ello, el análisis de *clustering* nos ha servido para desarrollar las nuevas técnicas de marketing por localizaciones, así como, será la herramienta fundamental

para otorgar los *insights* necesarios de cara a una futura expansión geográfica que actualmente se ha visto paralizada como consecuencia directa de la crisis generada por el Covid-19.

Así pues, en un negocio de carácter tan estacional como el de venta de toldos, que ya sabemos cómo sus ventas se duplican o triplican en verano con respecto al resto del año, la posibilidad crear un modelo predictivo que ayude a predecir no solo el precio, por un lado, sino también el tipo de artículo en el que va a estar interesado el cliente generaría un impacto ciertamente positivo en el negocio.

4 Conclusiones.

La utilidad de estas nuevas técnicas y herramientas de análisis de grandes cantidades de datos ya no es puesta en duda por nadie. La optimización de algunas de las métricas más importantes gracias a las conclusiones de estos análisis es una realidad. Desde la mejor retención de clientes, pasando por el desarrollo de productos más ajustados a las necesidades de la demanda o la obtención de ventajas competitivas hasta llegar a una mayor rapidez de reacción a los cambios de mercado.

Además, el análisis coste/beneficio[11], una ratio fundamental en el devenir de toda actividad económica es positivo. Por un lado, las herramientas utilizadas para acometer este tipo de estudios son, mayoritariamente, de naturaleza Open Source y, consecuentemente, son totalmente gratuitas y se actualizan regularmente por los propios usuarios. La otra cara de la moneda son los beneficios aportados por estas. Podemos asegurar tras el análisis realizado en esta empresa familiar de tamaño mediano que los beneficios son importantes tanto desde el punto de vista económico como desde el punto de vista estratégico. El uso generalizado de este tipo de técnicas dentro de aquellas empresas con mayor capitalización bursátil no podía ser casualidad.

La importancia de los datos es cada día mayor en el nuevo paradigma tecnológico. La facilidad con la que se pueden extraer conclusiones de gran valor a través de ellos es reseñable. A lo largo de este trabajo se ha probado que no hace falta ser una empresa internacional o ni siquiera tener un tamaño excesivamente grande para generar grandes volúmenes de información. Cualquier tipo de empresa, desde una multinacional del sector automovilístico hasta un pequeño negocio dedicado a la carnicería, puede recopilar muchos datos acerca de actividad económica. En este caso, la empresa que nos atañe disponía de apenas 1.600 ventas en el último año dentro de la Comunidad de Madrid y alrededores. Sin embargo, esos datos han sido más que suficientes para ejemplificar la viabilidad de implantación de esta tecnología.

[11] Análisis que mide la relación entre el coste por unidad producida y el beneficio obtenido por su venta.

Así pues, las aplicaciones de estas herramientas no se podrá dar de forma uniforme en cada empresa, sino que dependerán del modelo de negocio, de sus características, de los datos recopilados, de la industria o, incluso, de la situación macroeconómica del país donde opera. Salvando estas diferencias, el impacto sería transversal en todas las áreas de negocio: dirección, producción, marketing o el departamento de ventas verían afectado su funcionamiento.

La época actual, asolada por la mayor pandemia mundial en los últimos 100 años, es un momento transcendental en la vida de muchas empresas de pequeño y mediano tamaño. Es un momento de incertidumbre sin precedentes donde los detalles marcarán la diferencia a la hora de subsistir en un mercado turbulento y muy castigado por los dos meses de parón económico. Sin embargo, en estos momentos de dificultad, la modernización y optimización de procesos a través de las nuevas herramientas mostradas, no solo no debe quedar en un segundo plano, sino que debe situarse como la opción principal para generar valor, tal y como demuestra este trabajo.

Pese a ello, todo buen estudio tiene limitaciones y complicaciones que necesariamente deben ser mencionados en este apartado de conclusiones. Como toda herramienta que pretende simplificar la realidad a través de modelos, no siempre es posible lograr explicar la variable objetivo como ha sido el caso en este estudio debido a la falta de variables que realmente se correlacionasen y explicasen esta.

En este sentido, durante el estudio también podrán surgir complicaciones a la hora de realizar el análisis como consecuencia de la falta de personal cualificado para su realización, resultados erróneos por la aplicación incorrecta de determinadas herramientas o problemas técnicos causados por la reticencia por parte de empresas tradicionales a afrontar cambios novedosos y disruptivos. Todos estos aspectos deben ser tenidos en cuenta para futuros análisis de este tipo, dado que estos han aparecido en menor escala en este ensayo.

En cualquier caso, este análisis ha demostrado ser de gran utilidad. La innovación que proporcionan las herramientas de análisis de datos será fundamental para poder competir en el actual mercado. Permitiéndome la libertad de modificar una de las grandes citas de Charles

Darwin, *"No es la más fuerte de las [empresas] la que sobrevive, tampoco la más inteligente. Es aquella que mejor se adapta al cambio".* El cambio de paradigma ya es una realidad, es momento de adaptarse y que las pequeñas y medianas empresas familiares encabecen esta transformación. Poner el foco de nuestra atención en las herramientas de análisis de datos debe ser el objetivo primordial.

5 Bibliografía

Aguilar, L. J. (2013). Big Data, Análisis de grandes volúmenes de datos en organizaciones. Alfaomega. México D.F.

Baltassis, E., Coulin, A.-D., Gourevitch, A., Khendek, Y., & Lucas, Q. (2019). A rough road to data maturity. Informe BCG.

BBVA. (2017). Las cinco uves del big data. Obtenido de BBVA - BIG DATA: https://www.bbva.com/es/las-cinco-uves-del-big-data/

BCG. (2020). Real-World Impact of Big Data and Advanced Analytics. Obtenido el 3 de abril de 2020 en https://www.bcg.com/capabilities/big-data-advanced-analytics/impact.aspx

Brahm, C., Sherer, L., Fleming, R., & Bennett, B. (2017). With Advanced Analytics, It's People (Not Data) That Stand in the Way of Change. Brief Insights, Bain & Company.

Brown, B., Chui, M., & Manyika, J. (2011). "Are you ready for the era of 'big data'?" McKinsey Quarterly, Octubre. Obtenido el 15 defebrero de 2020 en https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/are-you-ready-for-the-era-of-big-data#

Campello, M., Graham, J., & Harvey, C. (2010). "The real effects of financial constraints: Evidence from a financial crisis". Journal of Financial Economics, Volume 97, Issue 3, pp. 470-487.

Castro, B., & Castaño, P. (2019). Modelos Predictivos: Características y posible precio de alquiler de un Airbnb en Nueva York. Trabajo Fin de Diploma en Fundamentos en Business Analytics. Universidad Pontificia Comillas.

Comisión Europea. (2015). Evaluación del grado de urbanización a nivel global. Joint Research Centre.Cumbley, R., & Church, P. (2013). "Is "Big Data" creepy?" Computer Law & Security Review Volume 29, Issue 5, pp. 601-609.

Domo Consulting. (2018). Data never sleeps 6.0. Informe.

Everitt, B. S., , Landau, S., Leese, M., & Stahl, D. (2011). An introduction to classification and clustering. En Cluster Analysis, 5th ed. Chapter 1,pp. 15-16.Wiley.

Gualtieri, M., &Yuhanna, N. (2016). Hadoop Is Data's Darling For A Reason. Forrester report.

Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study". International Journal of Production Economics, Volume 165, pp.234-246.

Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003). "The Google File System". ACM SIGOPS Operating Systems Review Volume 37, Issue 5, pp. 29-43.

Gomez Carretero, A. I., & Piattini Velthuis, M. (2018). Importancia de la calidad de los datos en la transformación digital. RUIDERAe: Revista de Unidades de Información, número 13, pp. 8-12.

Harding, F. (2007). Venta cruzada: manual de cross-selling. Gestión 2000.Barcelona.

Hawkins, D. M. (1980). Multiple outliers. En D. M. Hawkins, Identification of Outliers, Chapter 5,pp. 51-73. Springer, Dordrecht.

Hürtgen, H., & Mohr, N. (2018). Achieving business impact with data. A comprehensive perspective on the insights value chain. McKinsey & Company report.

Infaimon. (2017). ¿Qué es la visión artificial y cómo puede mejorar tu aplicación? Obtenido el 11 de enero de 2020 en https://blog.infaimon.com/sistemas-de-vision-artificial-tipos-aplicaciones/

International Campus. (2017). El origen del Big Data. Obtenido de Big Data International Campus: https://www.campusbigdata.com/big-data-blog/item/106-origen-big-data

Johnson, S. (1967). Hierarchical clustering schemes. Psychometrika, Volume 32, Number 3, pp. 241-254.

Jugulum, R. (2016). Importance of Data Quality for Analytics. En Sampaio, P. & Saraiva, P. (Eds.), Quality in the 21st Century,pp. 23-31.Springer

Lerner, J., & Tirole, J. (2002). "Some Simple Economics of Open Source". The Journal of Industrial Economics, Volume 50, Issue 2, pp. 197-234.

Fuente, S. de la (20011). Análisis de Conglomerados (Cluster Analysis).Materiales del curso. Universidad Autónoma de Madrid.

Management Solutions. (2015). Data Science y la transformación del sector financiero.

Marcireau, J.-A. (2019). 2,5 millones de bytes al día: por qué los datos son el nuevo petróleo. El Confidencial. Obtenido el 29 de marzo de 2020 en El Confidencial: https://blogs.elconfidencial.com/economia/tribuna/2019-01-23/big-data-oportunidades-inversiones-fondo-msci-bra_1758462/

Mayor, E. (2015). Learning Predictive Analytics with R. Get to grips with key data visualization and predictive analytic skills using R. Packt Publishing.

Ministerio de Industria, Comercio y Turismo. (2019). Cifras PyME. Gobierno de España.

Mondéjar, J. A. (2014). Cross-selling y up-selling. En: La actividad turística española en 2013, pp. 407-410 Madrid. AECIT Editores.

Nyce, C. (2007). Predictive Analytics White Paper. American Institue for CPCU/Insurance Institute of America.

PowerData. (2017). Big Data: ¿En qué consiste? Su importancia, desafíos y gobernabilidad. Obtenido de PowerData: https://www.powerdata.es/big-data

Price Waterhouse Coopers. (2015). Temas candentes de la Ciberseguridad Un nuevo espacio lleno de incógnitas.

Ramos, S. H. (2018). La importancia del Big Data en la Ciberseguridad. Obtenido el 25 de febrero de 2020 en Telefónica: https://empresas.blogthinkbig.com/ciberseguridad-bigdata/

Russom, P. (2011). Big data analytics. TDWI Best Practices Report. Obtenido el 17 de abril de 2020 en https://tdwi.org/research/2011/09/best-practices-report-q4-big-data-analytics.aspx?tc=page0&tc=assetpg&m=1

Smryrnaios, N. (2016). "The GAFAM effect: Strategies and logics of the internet oligopoly". Communication & Langages, Volume 188, Issue 2, pp. 61-83.

Tabares, L., & Hernandez, J. (2014). Big Data Analytics: Oportunidades, Retos y Tendencias. Documento de Trabajo, Universidad de San Buenaventura. Colombia.

Turkey, J. (1977). Exploratory Data Analysis. Addisson Wesley.

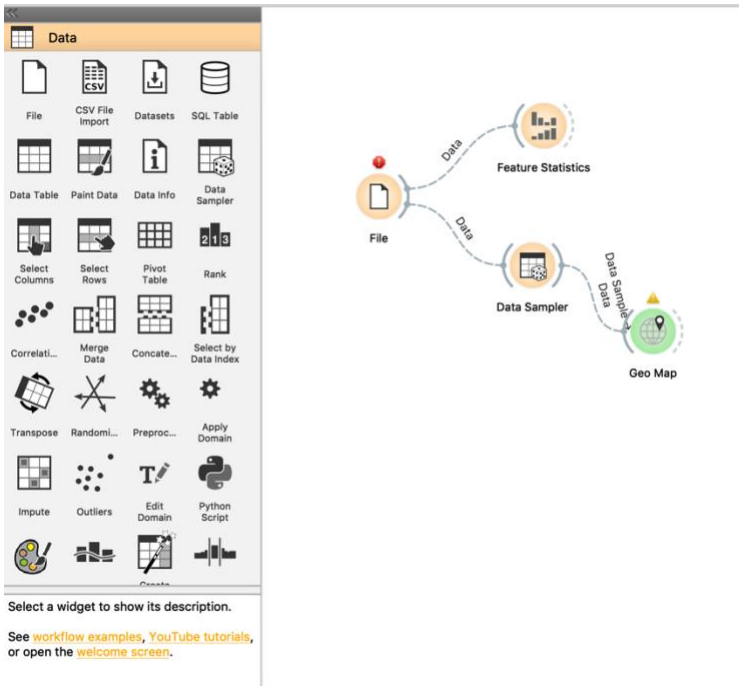
Velickovic, V. (2015). "What Everyone Should Know about Statistical Correlation". American Scientist, Volume 103, Issue 1, pp. 26-29.

Villamizar, R. (2019). 'Big data', el nuevo petróleo. Obtenido el 1 de marzo de 2020 en Portafolio: https://www.portafolio.co/opinion/otros-columnistas-1/big-data-el-nuevo-petroleo-529392

Zamora, J. (2016). Tomar mejores decisiones con el Big Data. *Harvard Deusto Business Review*.Número 256, pp. 6-14

6 Anexos.

Anexo 1: Captura de Orange para la creación de mapas



Anexo 2: Captura de Orange del análisis *clustering*