UNIVERSIDAD CEU SAN PABLO

FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

CARACTERÍSTICAS ÓPTIMAS QUE DEBE TENER UN APARTAMENTO DE ALQUILER A CORTO PLAZO EN MADRID PARA AUMENTAR SUS POSIBILIDADES DE ÉXITO

OPTIMAL CHARACTERISTICS THAT A SHORT-TERM RENTAL APARTMENT IN MADRID SHOULD HAVE TO INCREASE ITS CHANCES OF SUCCESS

TRABAJO DE FIN DE GRADO

GRADO EN ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS

AUTOR: ELENA SANTAMARÍA IZQUIERDO

DIRECTOR: PABLO ARÉS GASTESI

DEPARTAMENTO DE MATEMÁTICAS Y CIENCIAS DE DATOS

Madrid, 11 de mayo 2023

ÍNDICE

RESUMEN

En el presente trabajo se realiza un análisis estadístico, con un gran componente visual, de los apartamentos de alquiler de Airbnb en la ciudad de Madrid. El objetivo es proporcionar recomendaciones sobre las características óptimas que debe tener un apartamento de Madrid cuando se pone en alquiler (a corto plazo) para que tenga mayores posibilidades de éxito. Se ha hecho uso de una herramienta específica de visualización de datos e interactiva para analizar el contexto inicial y extraer conclusiones más generales, y posteriormente se ha continuado con un análisis más específico, usando algoritmos de aprendizaje supervisado (árboles de decisión, regresiones lineales simples y múltiples), y no supervisado (agrupaciones en clúster), mostrando así las relaciones entre varias variables, tanto cuantitativas como cualitativas.

PALABRAS CLAVE: Predicción, apartamento de alquiler, visualización de datos, árboles de decisión, clústeres.

ABSTRACT:

The present work consists of a statistical analysis with a large visual component of the Airbnb rental apartments in the city of Madrid. The objective of this work is to provide recommendations on the optimal characteristics that an apartment in Madrid should have, when it is put up for rent, so that it has a better chance of success. A specific and interactive data visualization tool has been used to analyse the initial context, and draw more general conclusions, and later a more specific analysis has been continued, using supervised learning algorithms (decision trees, simple and multiple linear regressions), as well as unsupervised learning techniques (clustering), thus showing the relationships between various variables, both quantitative and qualitative.

KEYWORDS: Prediction, rental apartment, data visualization, decision trees, clusters.

1. INTRODUCCIÓN

1.1. La economía colaborativa

La economía colaborativa se define como una forma de conectar a varios individuos a través de una plataforma virtual, creando, distribuyendo y consumiendo bienes y servicios sin el uso de intermediarios (Raffini, 2016). Esto se conoce como comunicación P2P (peer to peer). Como apunta, C.Y. Heo (2016) este tipo de comunicación ha tenido mucho éxito en diversos sectores tales como alojamiento, ocio, restauración y transportes. Ejemplos empresas conocidas en estos sectores son Airbnb, Cabify y BlaBlaCar

La economía colaborativa fue considerada por la revista Time en 2011 como una idea que revolucionaría el mundo del turismo, entendiéndose como una continuación de la revolución tecnológica. Más del 50% de los millenials[1] reconocen hacer uso de la economía colaborativa, siendo una alternativa a las formas más tradicionales, generando una competencia entre empresas y usuarios, que pueden ser vendedores y compradores al mismo tiempo. (Bardhi y Eckhardt, 2012).

Otro factor que ha impulsado la economía colaborativa en el turismo es el desarrollo tecnológico, considerando las plataformas que hay en internet como un sitio de confianza donde poder ver los comentarios y recomendaciones de otros usuarios (Buhalis y Law, 2008).

Dentro de la economía colaborativa este estudio se va a enfocar en las del turismo, concretamente en la plataforma de Airbnb.

AirBed & Breakfast, ahora conocida como Airbnb, se fundó en 2007 de la mano de Brian Chesky y Joe Gebbia, dos compañeros que compartían piso en San Francisco. Un año más tarde se les unió Nathan Blecharczyk y sacaron la página web, "*Airbedandbreakfast.com"*. Esta se presenta como una plataforma donde se puede alquilar alojamiento online, como anfitrión o huésped (Airbnb, 2023 a). La empresa se expandió internacionalmente en 2011, empezando en Alemania. En España se comenzó a usar el término economía colaborativa en 2012, y en una encuesta de 2020 Airbnb fue la segunda plataforma online más usada por los españoles para reservar hoteles o viviendas turísticas (Statista, 2023 a). Airbnb gana dinero de las tarifas que

[1] Nacidos entre 1981 y 1995

añade a los usuarios al alquilar el inmueble, actualmente tiene unos ingresos de 110.000 millones de dólares anuales, y más de 100.000 anuncios activos (Statista, 2021), alcanzando en 2021 su valor bruto más alto, con 46.880 millones de dólares. (Statista, 2022).

Teniendo en cuenta que muchos consumidores buscan abaratar costes en los viajes, compartir alojamiento es una muy buena opción, esta idea se ha ido expandiendo por todo el mundo, y a día de hoy, Airbnb se posiciona como líder en este mercado, disponiendo de ofertas de alojamientos en la mayoría de países.

1.2. Justificación del análisis del proyecto

La motivación para elegir el sector hotelero como tema para mi trabajo es por la importancia que tiene el turismo en España, que está directamente relacionado con la hostelería, ya que una gran parte de turistas son internacionales (71.659.281 turistas internacionales en 2022) y se alojan en hoteles, apartamentos o similares (Statista, 2023 b). Además, la pandemia y la compleja situación internacional son factores a tener en cuenta, porque han supuesto un cambio radical en las estadísticas del sector en los últimos años.

Por todo esto, me parece interesante realizar un análisis numérico y visual de la hostelería en España (concretamente de los apartamentos de Airbnb en la Comunidad de Madrid), que está en constante cambio, y ver cómo afecta a aquellas personas que tienen pensado alquilar su inmueble. Los datos los extraje de la página web "*Inside Airbnb*", que recopila información disponible en Airbnb sobre docenas de ciudades y países, tanto de Europa como de otros continentes, usando para ello tecnologías como Python o Bootstrap y Mapbox (para el diseño de los mapas).

2. METODOLOGÍA

2.1 Metodología CRISP-DM

Para el análisis realizado en el proyecto se ha utilizado la metodología CRISP-DM (Chapman, 1999), que está formado por cuatro niveles, organizados jerárquicamente desde el más general al más específico.

En el primer nivel se encuentra el número de fases del proceso, en el segundo nivel están las tareas más genéricas (limpieza de datos), en el tercero están las tareas más específicas, que explican cómo llevar a cabo las tareas genéricas en situaciones

concretas, y en el último la instancia de proceso, que registra los resultados de una situación en particular.

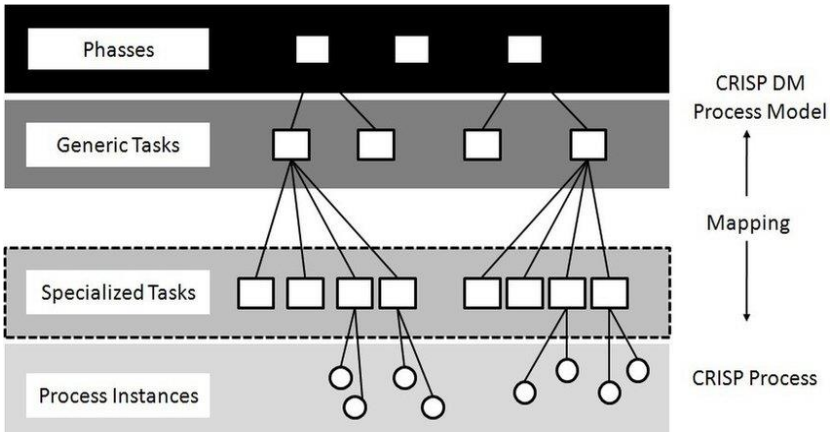


Ilustración 1: Esquema de los cuatro niveles de abstracción de la metodología CRISP-DM. Fuente: (Chapman, 1999)

A nivel más general, el proceso está organizado en seis fases, cada una estructurada en varias tareas secundarias. Esta metodología estructura el ciclo de vida de un proyecto de manera que las fases interactúan entre sí de forma iterativa.

Descripción de las fases:

1. Comprensión del negocio: entender los objetivos y requerimientos desde el punto de vista empresarial
2. Comprensión de los datos: recolección de los datos e identificación de los primeros problemas que se presentan, visualizando las relaciones más intuitivas entre variables
3. Preparación de los datos: incluye la depuración de las bases de datos, eliminando las variables que no se van a utilizar para el análisis, e implementando otros que pueden ser útiles. Esta fase está muy relacionada con la siguiente de modelado, por lo que están en constante interacción
4. Modelado: Se eligen las técnicas de aprendizaje automático (machine learning) más adecuadas para el trabajo, dependiendo del problema que se presenta, los datos disponibles, tiempo necesario para hacer el modelo y el grado de conocimiento de esa técnica.
5. Evaluación: Se evalúa el modelo en función de los criterios expuestos en la primera fase, de lo que se considera éxito en la resolución del problema. En

función de los resultados o caso de cometer algún error, se repetirán los pasos pertinentes y se volverá a evaluar el modelo

6. Despliegue: En caso de que el modelo sea válido, se realizará el despliegue, que trata de explicar y documentar los resultados obtenidos de forma clara, difundiendo los resultados

Dado que el objetivo de este trabajo es una exploración estadística de los datos junto con sus respectivas visualizaciones y no el desarrollo del modelo de negocio, el proyecto abarcará solamente las cuatro primeras fases, dejando la cinco y seis para otro trabajo si procede.

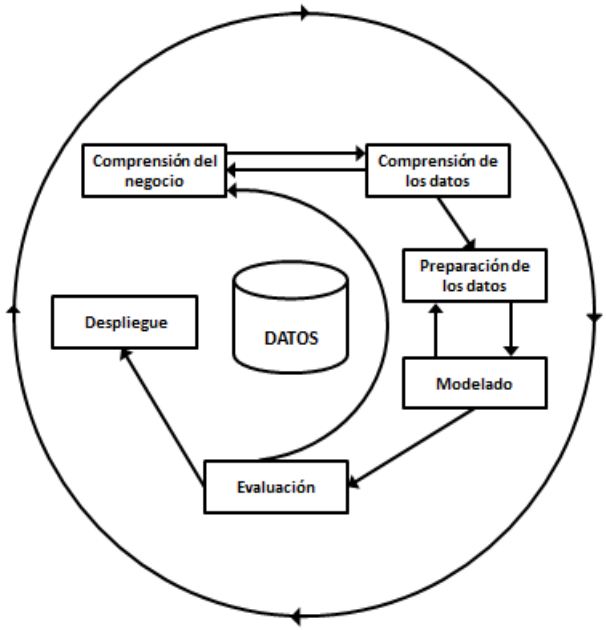


Ilustración 2: Seis fases de la metodología CRISP-DM. Fuente: (Chapman, 1999)

2.2 Herramientas y técnicas utilizadas

El análisis se ha sido realizado en lenguaje R, que permite crear gráficos avanzados y es muy adecuado para el análisis estadístico. Otra herramienta utilizada en este proyecto que se ha considerado apta para el tema elegido es Tableau, ya que es específica para el análisis visual, tanto de datos estructurados como no estructurados y siendo una interfaz interactiva que será utilizada para algunos gráficos.

La razón por la que se usan algoritmos de minería de datos es para identificar patrones en grandes bases de datos, estas técnicas están en constante evolución gracias a la inteligencia artificial, la estadística y la visualización de los datos entre otros.

Los algoritmos de minería de datos se clasifican en supervisados o predictivos y no supervisados. (M. Weiss, S., Indurkhya, N,1998)

"El aprendizaje supervisado se trata de, partiendo de unos datos con etiqueta conocida, se induce una relación entre dicha etiqueta y otra serie de atributos, para realizar predicciones en datos cuya etiqueta es desconocida. Este tipo de aprendizaje consta de dos fases: Entrenamiento (construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos). En caso de no ser posible realizar una predicción a partir de los datos originales, se recurre a los métodos no supervisados, que encuentran patrones y tendencias en los datos reales (no históricos)" (M. Weiss, S., Indurkhya, N,1998)

En la tabla que se muestra a continuación aparecen algunas de las técnicas de minería de ambos métodos, supervisado y no supervisado.

Supervisado	No supervisado
Árboles de decisión	Agrupamiento ("clustering")
Redes neuronales	Segmentación
Regresión	Minería de textos
Series temporales	Patrones secuenciales
	Visualización de datos

Tabla 1: Clasificación de algoritmos en Minería de datos. Fuente: (Weiss & Indurkhya, 1998)

En este estudio, realizado con la herramienta RStudio, utilizaremos las siguientes técnicas:

- Visualización de datos

Según Lev Manovich, (2010), la visualización es "*una transformación de datos cuantificados no visuales en una representación visual*", siendo una técnica para presentar la información obtenida de forma clara y concisa a los usuarios finales. Es una buena técnica para comenzar a analizar los datos, ya que se pueden encontrar patrones, tendencias y/o anomalías en un gráfico a primera vista (Minguillón, 2016) y se agrupa dentro del aprendizaje no supervisado.

- Regresiones lineales (simples y múltiples)

"La regresión es una técnica utilizada para inferir datos a partir de otros y hallar una respuesta de lo que puede suceder" (James, G., Witten, D., Hastie, T., Tibshirani, R, 2013).

Las regresiones se encuadran dentro del aprendizaje supervisado; en una regresión lineal simple se incluye una variable independiente $[x]$ que puede estar relacionada con la dependiente Y, y modificar su valor. Existen varios tipos de regresiones, y con un modelo adecuado se puede explicar y predecir el comportamiento de la variable dependiente.

Usando datos reales estos no siempre forman exactamente una recta, con lo cual hay que añadirle un error, quedando la fórmula de regresión lineal simple (fórmula de una recta) de la siguiente manera:

$$Y = \beta_o + \beta_1 x + \varepsilon \qquad (1)$$

En la ecuación (1), "Y *es la variable dependiente,* x *es la variable independiente,* β_o *es la ordenada en el origen y* β_1 *indica el cambio esperado de la respuesta* Y *por cambio unitario en* x *y* ε ***es el error"*** (Montgomery, 2006)

Cuando en el modelo de regresión interviene más de una variable independiente, se llama modelo de regresión lineal múltiple y una ecuación correspondiente con este modelo es:

$$Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \qquad (2)$$

En la función (2) β_2 mide el cambio de Y cada vez que cambia x_2 y $x_1$permanece constante.

El coeficiente de correlación lineal, o coeficiente de correlación de Pearson, indica el grado de relación lineal entre dos variables, está comprendido entre -1 y 1. Dicha puede ser directa (ambas variables aumentan o disminuyen en la misma dirección), inversa (una variable aumenta y la otra disminuye) o nula (no existe relación entre las variables o no existe correlación lineal), y el signo indica si la relación es positiva o negativa. Las fórmulas estadísticas se encuentran en el Apéndice A

❖ Árboles de decisión

El árbol de decisión es un método no paramétrico de aprendizaje supervisado que utiliza los principios de clasificación para predecir el resultado de una decisión. Cada árbol se compone de nodos, que son cada una de las opciones a tener en cuenta antes de tomar una decisión, las ramas (representadas por los segmentos que conectan los nodos), son la probabilidad de ocurrencia de la decisión elegida y los nodos finales el resultado de esa decisión. (Ali, Khan, Ahmad y Maqsood, 2012).

Con este algoritmo se consigue clasificar los datos en grupos y subgrupos, agrupando los datos más homogéneos y separando los más heterogéneos entre sí.

Se pueden distinguir dos tipos de métodos, en función del tipo de variable a discriminar:

- Árboles de clasificación: se usa para variables categóricas.
- Árboles de regresión: se usa para a variables continuas.

❖ Agrupamiento o clustering

El método de clustering forma parte del aprendizaje no supervisado y su objetivo es dividir los datos en grupos similares. Al utilizar este método existe el riesgo de perder información sobre los datos, pero se consigue simplificarlos, de forma que sean más sencillos de analizar. (Garre, M., Cuadrado, J. J., Sicilia, M. A., Rodríguez, D., & Rejas, R, 2007)

Para llevar a cabo este método se va a usar el algortimo k-means, que se trata de seleccionar k puntos aleatorios, calcular la distancia entre ellos para medir qué datos son más parecidos entre sí y el centroide[2] de cada grupo, asignando a cada punto su grupo más cercano. Este proceso se repite hasta llegar al número máximo de iteraciones (fijado por el usuario) o hasta que los clústeres no varíen. El mejor modelo encontrado será el que tenga menor suma de cuadrados total[3] (WCSS). Finalmente se representan en un gráfico, el cual cada grupo corresponde a un clúster. (Datacamp, 2023)

[2] Es el punto más representativo dentro de cada clúster. Por lo general, es la media de los valores de los puntos de cada grupo. (Microsoft, 2022)
[3] La suma de las distancias entre los puntos de datos y el centroide correspondiente para cada grupo (Datacamp, 2023)

3. COMPRENSIÓN DE LOS DATOS

3.1 Recolección de los datos

Los datos utilizados en este trabajo han sido extraídos de la web *Inside Airbnb*, una página online que hace webscrapping (extrae y almacena datos de páginas web para analizarlos o utilizarlos en otra parte) de la página de Airbnb; fue fundada por Murray Cox (Airbnb, 2023 b). Para el trabajo se han utilizado datos históricos recogidos entre enero de 2012 y septiembre de 2022 de la ciudad de Madrid.

En la página web están disponibles dos bases de datos: listings.csv.gz, que incluye una versión completa y otra resumida con las variables más importantes que van a analizarse por separado, y reviews.csv, base de datos con los comentarios de los huéspedes.

3.2 Selección de datos y descripción

Las bases de datos limpias quedan de la siguiente forma:

listings_detalle	9.196 observaciones	19 variables
listings_madrid	15.842 observaciones	13 variables
nube_amenities	877 observaciones	2 variables
reviews_madrid	24.668 observaciones	2 variables

Tabla 2: Cantidad de variables utilizadas en el estudio. Fuente: Elaboración propia

Para facilitar el análisis se han omitido todas las variables que no aportaban información al estudio o que no se han considerado relevantes, quedándonos con las siguientes variables a analizar:

Diccionario de la tabla de listings_detalle:

Campo	**Tipo de variable**	**Descripción**
Id	integer	Identificador único para las observaciones del Airbnb
room_type	text	Tipo de habitación: Apartamento completo:

		Se dispone de todo el espacio del apartamento para el huésped que lo alquila; esto suele incluir: cocina, baño, habitación y entrada por separado. El anfitrión debe notificar en la web en caso de ocupar el apartamento también. Habitación privada: Se dispone de un dormitorio entero para el huésped y puede que comparta otros espacios del apartamento con más gente. Habitación compartida: Se dispone de un dormitorio compartido entre varias personas, al igual que el resto de espacios del apartamento. Habitación de hotel: Alojamiento más elegante con servicios excelentes.
bedrooms	integer	Número de habitaciones del apartamento
beds	integer	Número de camas de apartamento
amenities	text	Servicios disponibles en el apartamento
price	currency	Precio por día en moneda local
last_review	date	Fecha (dd/mm/aa) del último comentario que recibió el apartamento
review_scores_rating	integer	Valoración de la experiencia total

review_scores_accuracy	integer	Valoración de la veracidad del apartamento con lo que se oferta en la web
review_scores_cleanliness	integer	Valoración de la limpieza del apartamento
review_scores_checkin	integer	Valoración de la llegada al apartamento
review_scores_communication	integer	Valoración de la comunicación entre el huésped y el anfitrión
review_scores_location	integer	Valoración del conocimiento del huésped sobre la información más importante acerca de: seguridad, transporte y otras causas de interés que afecten a su estancia
review_scores_value	integer	Valoración de la calidad/precio del apartamento

Tabla 3: Variables utilizadas en la base de listings_detalle. Fuente: Elaboración propia

Diccionario de la tabla de listings_madrid:

Campo	**Tipo de variable**	**Descripción**
id	integer	Identificador único para las observaciones del Airbnb
neighbourhood_group	text	Distrito de la Comunidad de Madrid
neighbourhood	text	Barrio de la Comunidad de Madrid
latitude	numeric	Latitud usando el Sistema geolocalización (WGS84) Uses the World Geodetic System (WGS84) projection for latitude and longitude.
longitude	numeric	Longitud usando el Sistema geolocalización (WGS84)

		Uses the World Geodetic System (WGS84) projection for latitude and longitude.
room_type	string	Tipo de habitación
price	currency	Precio por día en moneda local
minimum_nights	integer	Número mínimo de noches para alojarse en el apartamento
number_of_reviews	integer	Número de comentarios que ha recibido el apartamento. Se ha considerado el número de comentarios como una aproximación a la demanda
last_review	date	Fecha (dd/mm/aa) del último comentario que recibió el apartamento

Tabla 4: variables utilizadas en la base de listings_madrid. Fuente: Elaboración propia

Diccionario de la tabla de reviews_madrid:

Campo	**Tipo de variable**	**Descripción**
Id	integer	Identificador único para los comentarios del Airbnb
Reviews	text	Comentarios de los huéspedes

Tabla 5: variables utilizadas en la base de reviews_madrid. Fuente: Elaboración propia

3.3 Construcción, integración y formateo de datos

En base de datos de listings_madrid se ha añadido el campo código postal a partir de la página de *www.codigo-postal.info* que incluye los códigos postales de los diferentes barrios de España. Esto se ha utilizado para la visualización en Tableau de los distritos de Madrid en los mapas de precios y demanda.

No se han creado datos derivados de otros ya existentes ya que no se ha considerado necesario, analizando las variables cualitativas en su formato original.

No se realiza la integración de las tres bases de datos, si no que se van a analizar las bases de datos por separado, ya que eso facilita el análisis.

Tampoco ha sido necesario cambiar el orden de los campos ni de los registros. Así como tampoco se ha modificado el formato de los campos, ya que su original es el más apto y es compatible con el análisis que se quiere realizar.

4. PREPARACIÓN DE LOS DATOS

4.1 Exploración de los datos

Tras haber obtenido los datos se procede a la exploración de estos, esta técnica se usa entre otras cosas para encontrar valores atípicos o valores extremos, comprobar suposiciones y hacer agrupaciones. (López, P., Santín, C., González, D. (2007)

Para ello, lo primero que se hacen son los diagramas de cajas, ya que así se pueden identificar fácilmente los valores atípicos.

De la base de datos de listings_madrid se representan las siguientes variables cuantitativas:

El precio de los apartamentos (Apéndice B) que varía desde 1 hasta 10.000 euros; el número mínimo de noches que el anfitrión solicita para poder alquilar el apartamento (Apéndice B), que varía desde 1 hasta 1.000 aproximadamente, y el número de comentarios que han escrito los huéspedes sobre el apartamento (Gráfico 1), con valores desde 1 hasta 800 aproximadamente.

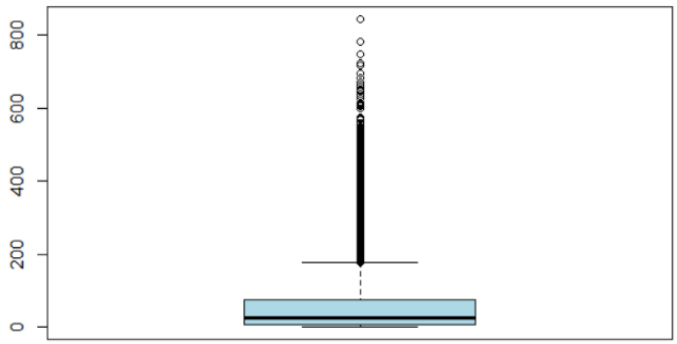


Gráfico 1: Diagrama de cajas del número de comentarios. Fuente: Elaboración propia

En los tres diagramas de cajas se aprecia la presencia de valores atípicos en la parte superior, con lo cual son distribuciones asimétricas positivas, con los valores del precio y el número mínimo de noches más dispersos que el número de comentarios.

De la base de datos de listings_detalle se presentan las siguientes variables cuantitativas:

En los siguientes diagramas de cajas (Gráficos 2, 3, 4, 5, 6, 7 y 8) se muestran las calificaciones que han dado los huéspedes al apartamento según: general, precisión, checkin, limpieza, comunicación, localización y valor del alojamiento. Estas calificaciones varían desde 1 hasta 5 (desde pésimo hasta excelente) y, al contrario que con el resto de las variables, se aprecia la presencia de valores atípicos en la parte inferior en todos los gráficos, con lo cual indica que son distribuciones asimétricas negativas.

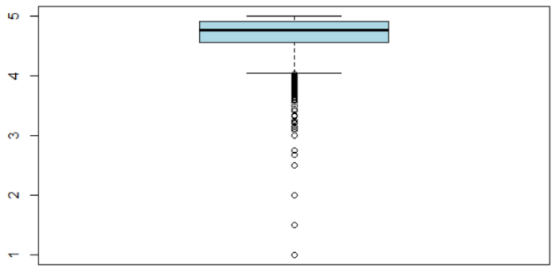


Gráfico 2: Diagrama de cajas de la calificación general. Fuente: Elaboración propia

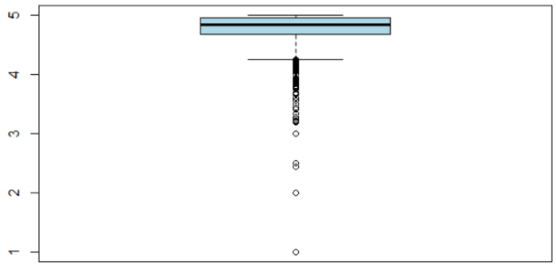


Gráfico 3: Diagrama de cajas de calificación de la precisión. Fuente: Elaboración propia

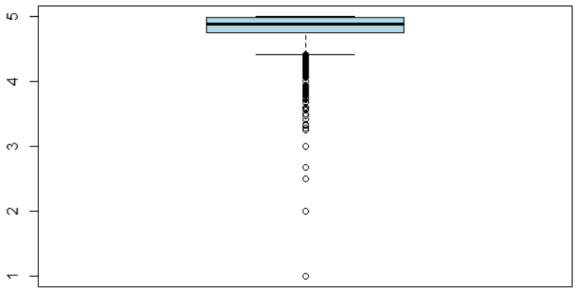


Gráfico 4: Diagrama de cajas de la calificación del checkin. Fuente: Elaboración propia

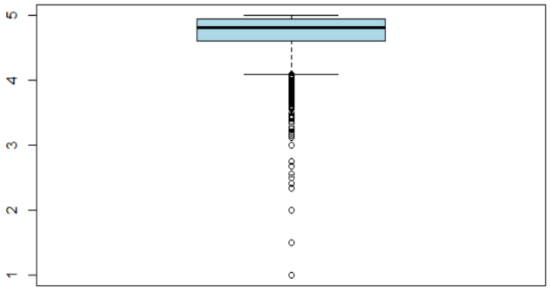


Gráfico 5: Diagrama de cajas de calificación de la limpieza. Fuente: Elaboración propia

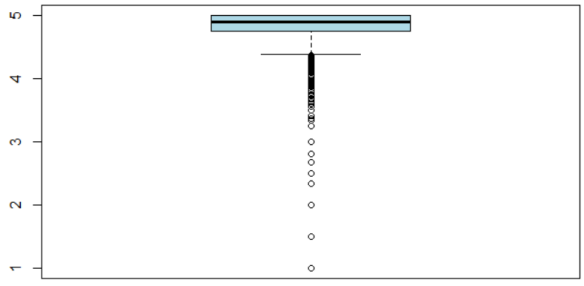


Gráfico 6: Diagrama de cajas de la calificación de la comunicación. Fuente: Elaboración propia

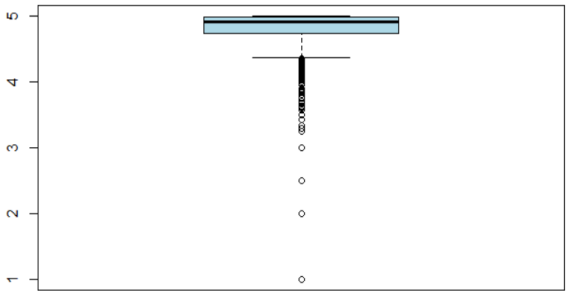


Gráfico 7: Diagrama de cajas de la calificación de la localización. Fuente: Elaboración propia

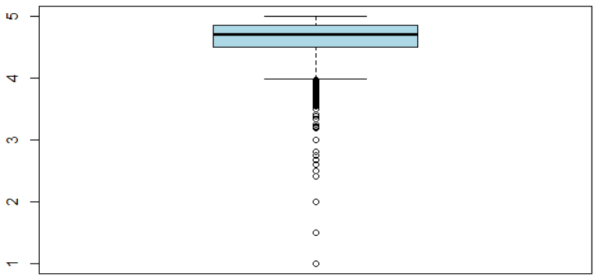


Gráfico 8: Diagrama de cajas de la calificación del valor. Fuente: Elaboración propia

También se presentan los gráficos del número de camas y habitaciones que hay en los apartamentos, que varía entre 1 y 17, con algunos valores atípicos en la parte superior, siendo distribuciones asimétricas positivas (Gráficos 9 y 10) y el precio de los apartamentos, que varía entre 1 y 997, pudiéndose apreciar gran cantidad de puntos atípicos en la parte superior, siendo una distribución asimétrica positiva (Gráfico 11)

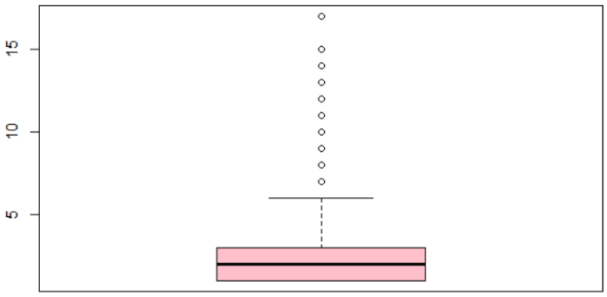


Gráfico 9: Diagrama de cajas del número de camas. Fuente: Elaboración propia

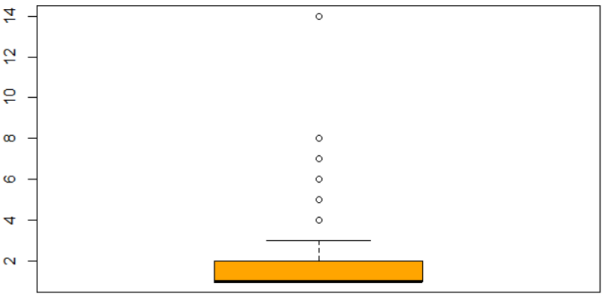


Gráfico 10: Diagrama de cajas del número de habitaciones. Fuente: Elaboración propia

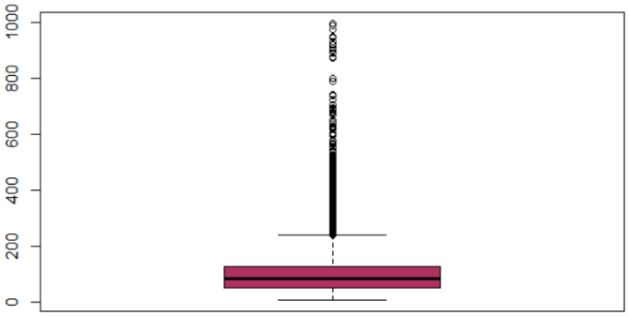


Gráfico 11: Histograma del número del precio. Fuente: Elaboración propia

4.2 Verificación

Tras realizar la exploración de los datos, se han encentrado valores atípicos en todos los diagramas de cajas, que al ser bases de datos abundantes es comprensibles que los haya. Se ha decidido centrarse en el estudio del comportamiento de los alquileres más comunes para un turista (para precios, número mínimo de noches, comentarios, habitaciones y camas son valores bajos y para puntuaciones son altos) para poder realizar un estudio más concreto y fiable. Al tener gran cantidad de observaciones la eliminación de estos no supone más de un 10% del número de datos de cada variable. En el caso de la variable precio se ve una clara distorsión de los datos cuando se compara el gráfico que contiene valores atípicos del que está acotado a valores entre 0 y 200 (Gráficos 12 y 13)

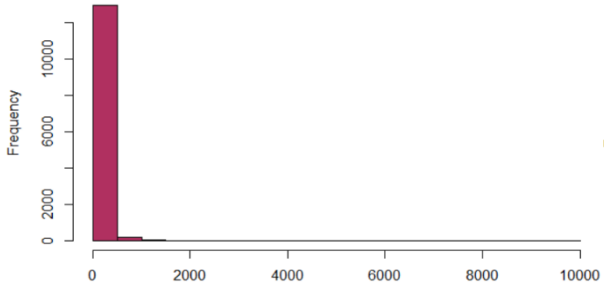


Gráfico 12: Histograma del precio de los apartamentos con valores atípicos. Fuente: Elaboración propia

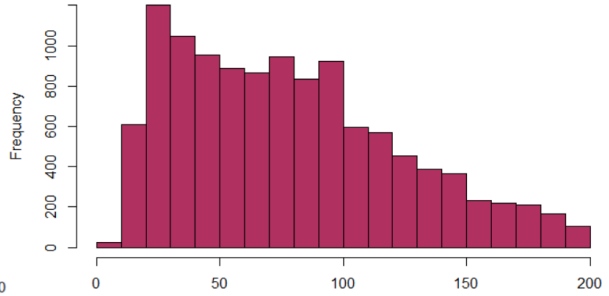


Gráfico 13: Histograma del precio de los apartamentos sin valores atípicos. Fuente: Elaboración propia

4.3 Limpieza de los datos

Las bases de datos con las que se va a trabajar contienen datos suficientes para aplicar diferentes técnicas de aprendizaje automático, pero es necesario hacer una limpieza de los datos, porque como se ha mostrado, se encuentran valores extremos en todas las variables numéricas.

Primero se han eliminado los valores nulos y después se ha continuado con la eliminación de los valores atípicos, ambos se han hecho desde Rstudio, usando el rango intercuartílico[4] para cada diagrama de cajas y se ha decidido poner un límite superior (Gráficos: 14, 15, 16) e inferior (Gráficos: 18, 19, 20, 21, 22, 23 y 24) para

[4] El rango intercuartílico (IQR) es la diferencia entre el penúltimo y el primer cuartil de una distribución (Q3-Q1). El primer y tercer cuartil representan respectivamente el 25% y 75% del conjunto de datos. (Rodó, 2020)

evitar la aparición de estos valores en el estudio, ya que no corresponden con el comportamiento de los clientes más comunes.

De la base de datos de listings_madrid se representan los siguientes histogramas de frecuencias:

Se visualiza el histograma de la variable precio (Gráfico 13), habiendo establecido el límite en 200 euros, y se aprecia una distribución ligeramente asimétrica positiva, la mayor parte del precio se encuentra a la izquierda de la media, lo que indica que la mayoría de los apartamentos alquilados son de precios bajos. En el histograma de la demanda se ha puesto el límite en 100, y se aprecia una distribución claramente asimétrica positiva, (Gráfico 14) donde la mayoría de los apartamentos tienen entre 0 y 20 comentarios.

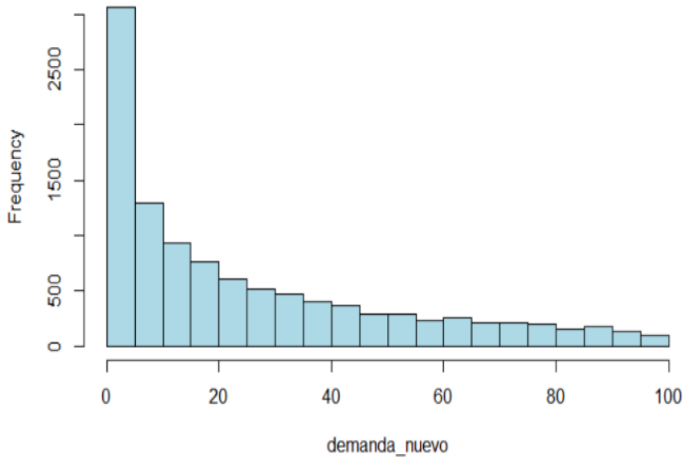


Gráfico 14: Histograma de la demanda sin valores atípicos. Fuente: Elaboración propia

En el diagrama de barras de la variable número mínimo de noches (Gráfico 15), con límite en 4 noches, se observa que el número mínimo de noches más común para alojarse en los apartamentos es de 1 o 2 noches, por lo que se entiende que gran cantidad de los clientes son turistas y no planean quedarse mucho tiempo para visitar la ciudad. Además, en el Gráfico 16, se observa que la mayor parte de los apartamentos disponibles se alquilan completos, o por habitaciones completas, siendo las habitaciones en los hoteles o compartidas una minoría en la base de datos. El distrito con más oferta es el distrito Centro (Gráfico 17).

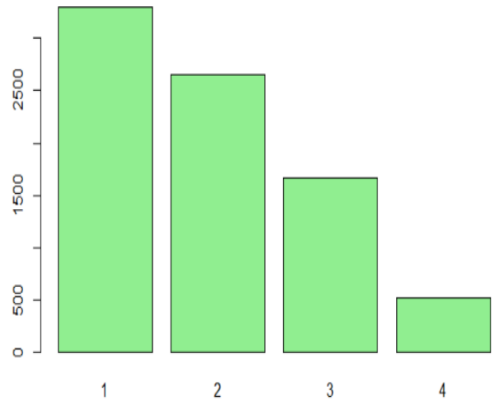


Gráfico 15: Diagrama de barras del número mínimo de noches sin valores atípicos. Fuente: Elaboración propia

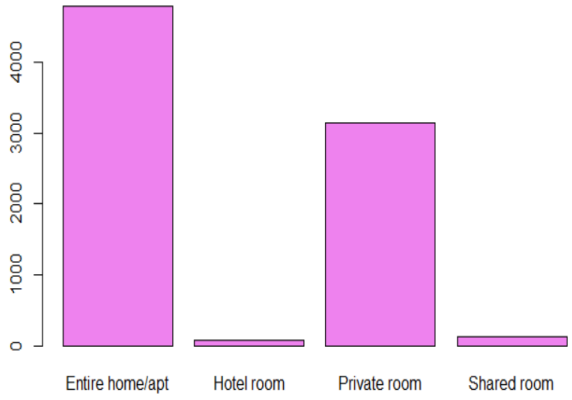


Gráfico 16: Diagrama de barras del tipo de apartamento sin valores atípicos. Fuente: Elaboración propia

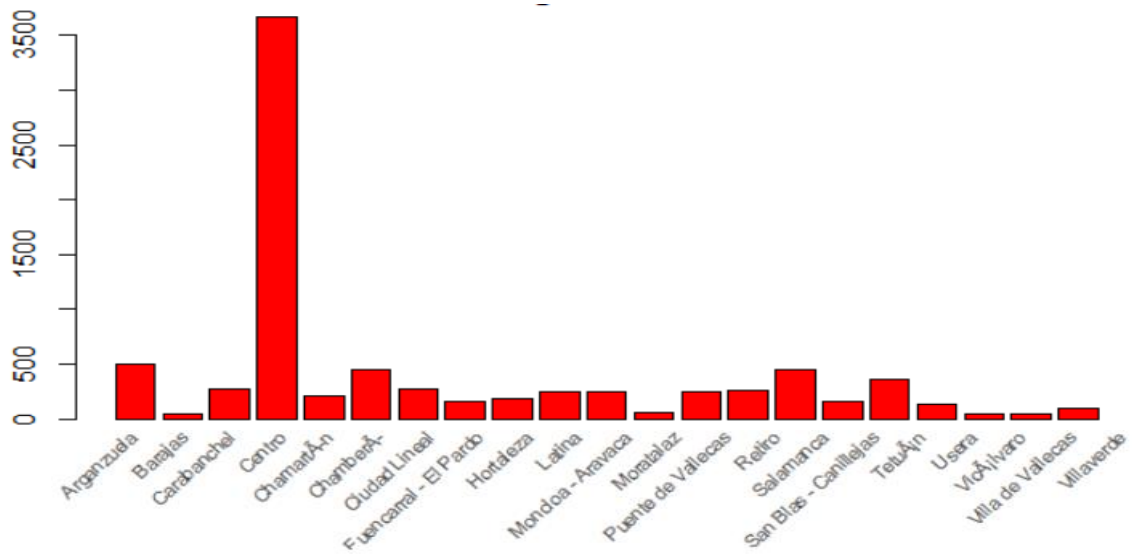


Gráfico 17: Diagrama de barras del distrito sin valores atípicos. Fuente: Elaboración propia

De la base de datos de listings_detalle se representan los siguientes gráficos:

En los siguientes histogramas (Gráficos 18, 19, 20, 21, 22, 23, y 24), se visualizan las reseñas que han dado los huéspedes al apartamento (intervalo entre 1 y 5), según: general, precisión, limpieza, checkin, comunicación, localización, valor y alojamiento. Todos los gráficos son asimétricos negativos y predominan las calificaciones altas, alrededor del 5, siendo esta puntuación la más repetida en comunicación, localización y valor. Así que, según los huéspedes, Madrid es una ciudad óptima para alquilar un apartamento en caso de se busque calidad/precio, servicios adecuados que debe

tener una ciudad y un propietario que esté pendiente de las necesidades del cliente y mantenga una comunicación frecuente.

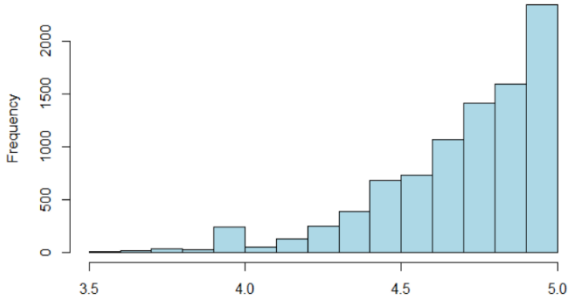


Gráfico 18: Histograma de calificación general precisión sin valores atípicos. Fuente: Elaboración propia

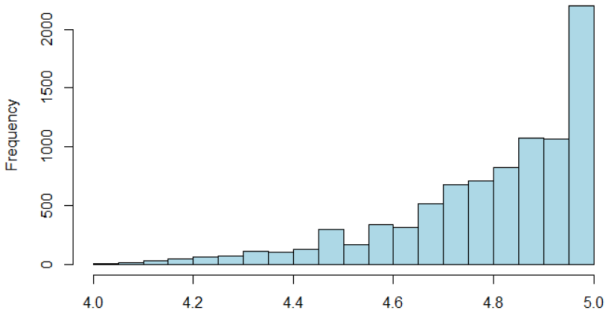


Gráfico 19: Histograma de calificación de la precisión sin valores atípicos. Fuente: Elaboración propia

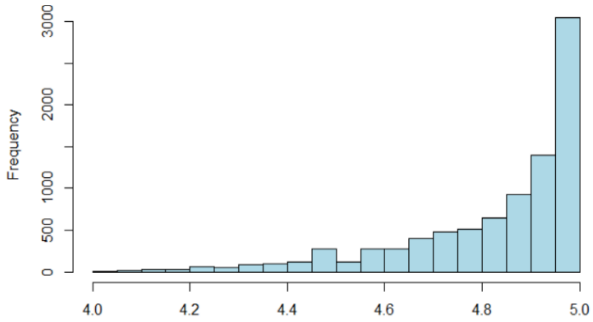


Gráfico 20: Histograma de calificación de la localización sin valores atípicos. Fuente: Elaboración propia

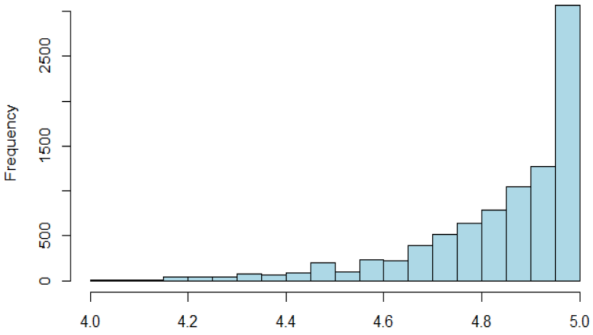


Gráfico 21: Histograma de calificación de la comunicación sin valores atípicos. Fuente: Elaboración propia

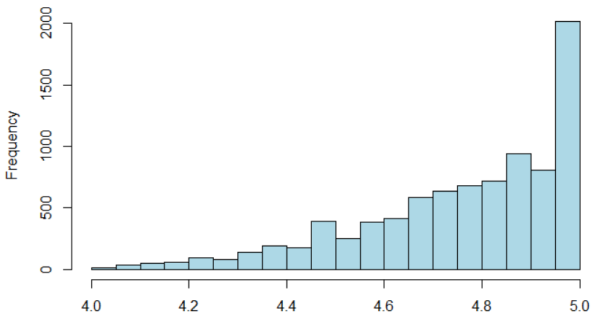


Gráfico 22: Histograma de calificación de la limpieza sin valores atípicos. Fuente: Elaboración propia

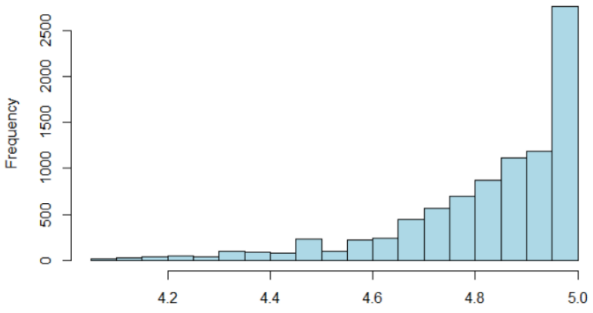


Gráfico 23: Histograma de calificación del checkin sin valores atípicos. Fuente: Elaboración propia

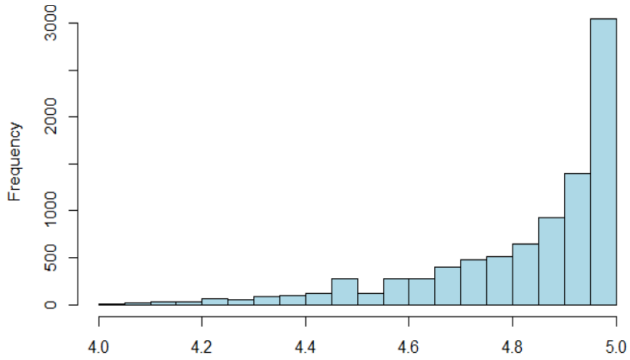


Gráfico 24: Histograma de calificación del valor sin valores atípicos. Fuente: Elaboración propia

También se visualizan el precio en el Gráfico 25 (vuelve a representarse porque no tiene el mismo número de observaciones que el precio de listings_madrid), con el límite en 200 y una distribución ligeramente asimétrica positiva, similar al precio de la base de datos de listings_madrid.

Y se aproxima que más de la mitad de los apartamentos tienen una única habitación para dormir (Gráfico 26) y suele haber 1 o 2 camas por apartamento (Grafico 27), con lo que se puede intuir que los alojamientos de Madrid no son muy grandes, no siendo aptos para familias numerosas.

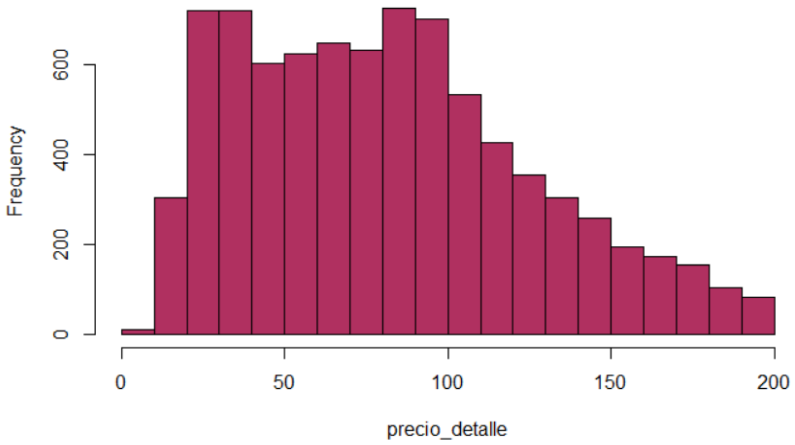


Gráfico 25: Histograma del precio de los apartamentos sin valores atípicos. Fuente: Elaboración propia

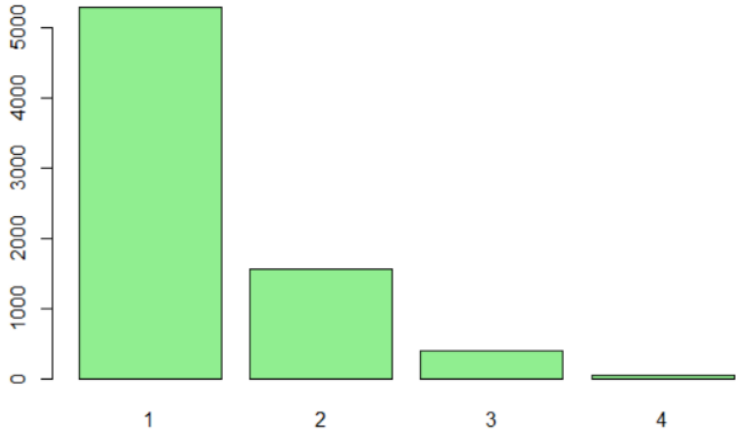


Gráfico 26: Diagrama de barras del número de habitaciones sin valores atípicos. Fuente: Elaboración propia

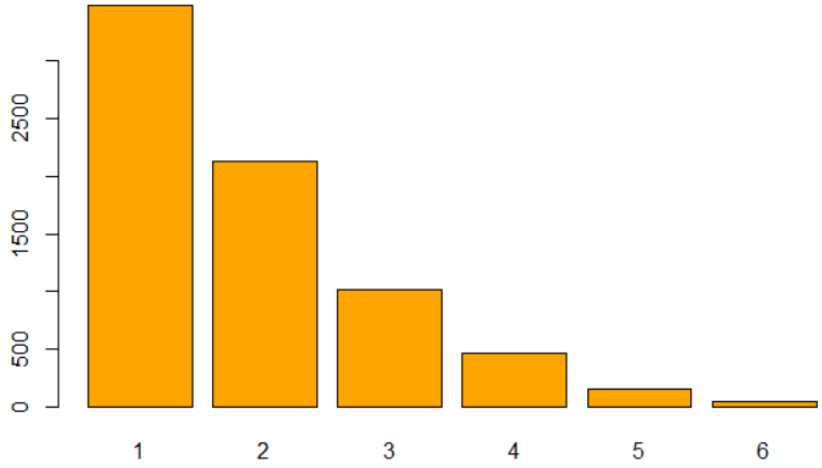


Gráfico 27: Diagrama de barras del número de camas sin valores atípicos. Fuente: Elaboración propia

5. MODELADO

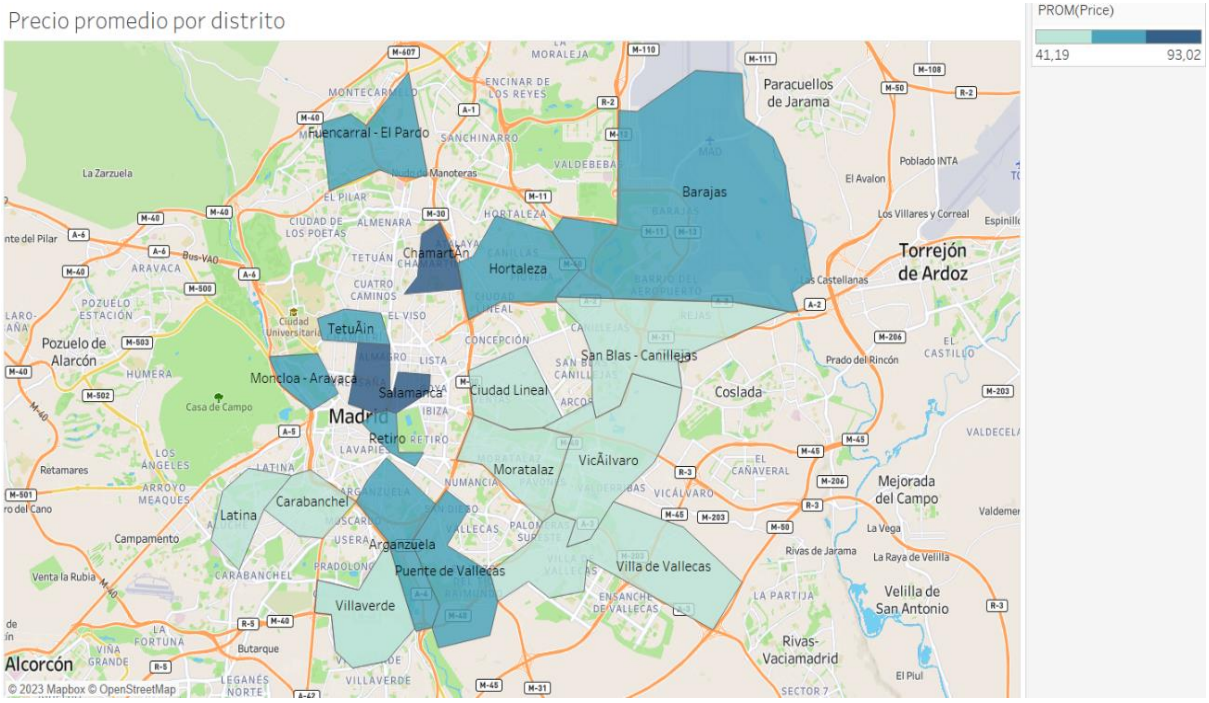
En este apartado se van a aplicar las técnicas explicadas anteriormente para hacer los modelos de aprendizaje automático.

5.1 Visualización de datos en Tableau

Para la realización de algunos gráficos (mapas de calles y densidad y diagramas de barras con variables cualitativas y cuantitativas) se ha decidido usar la herramienta de Tableau.

Una variable fundamental para este análisis es el precio[5], por lo que, para tener una visión general de dicha variable, se realiza un mapa de Madrid con los precios medios de cada distrito desde 2012 hasta 2022, asignándoles un color (más claro si el precio es menor y más oscuro si el precio es mayor), en una escala compuesta por tres tonalidades.

Según el banco de datos de la página *Madrid.es*, los distritos de Centro, Salamanca, Chamberí, Chamartín y Arganzuela tienen los precios promedio de alquiler de vivienda habitual más elevados de la ciudad de Madrid en 2022, en torno a $18€/m^2$, mientras que otros distritos como Villaverde, Vicálcaro o Villa de Vallecas tienen los precios promedio más bajos, alrededor de $11€/m^2$ (Véase Apéndice B). Esto mismo se ve reflejado en el Mapa 1, siendo los barrios más céntricos los de precio promedio más elevado. Esta información es de mucha utilidad para los huéspedes ya que pueden hacerse una idea de cuáles son las zonas más caras y cuáles las más asequibles.

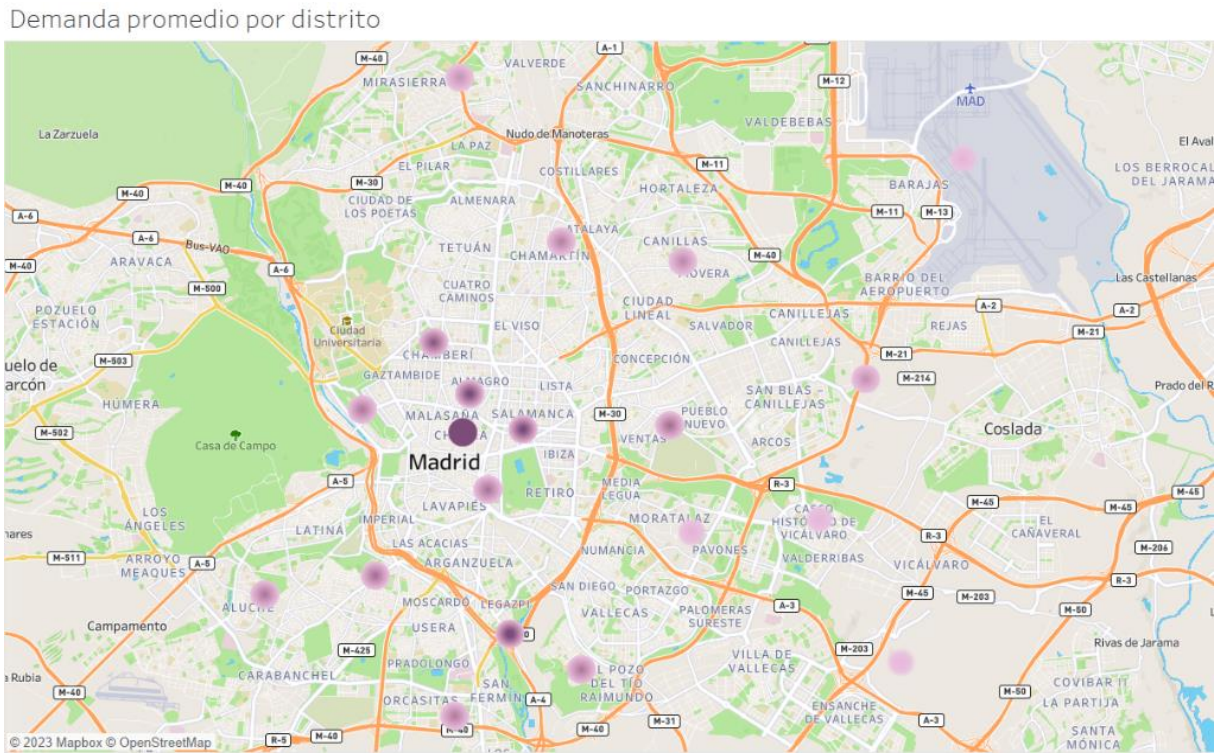


Mapa 1: Mapa de calles del precio promedio de cada distrito de Madrid.
Fuente: Elaboración propia en Tableau

[5] Las variaciones anuales de las tasas de inflación del IPC no son muy grandes de 2012 a 2022 (inflation.eu, 2023)

Otra variable que es considerada relevante cuando se estudian los alojamientos es la demanda. En este caso se ha decidido visualizar un mapa de densidad del promedio por distrito de la demanda de Madrid desde 2012 hasta 2022, y que indica la intensidad de la demanda por distrito (color más claro si la demanda es menor y más oscuro si la demanda es mayor).

Gracias a este mapa se puede intuir a simple vista cuáles son las mejores zonas donde alquilar un apartamento, ya que son las más elegidas por los huéspedes. En este caso, la zona más céntrica es la más concurrida, incluyendo distritos como Centro, Chamberí y Salamanca. Sin embargo, el distrito de Arganzuela que está más alejado de la zona centro, también es de los más elegidos por los huéspedes. Tomando como referencia los índices de alquiler de vivienda de 2018 (Ministerio de transportes movilidad y agenda urbana, 2023), los distritos más caros son Chamberí, Salamanca, Moncloa, Retiro y Chamartín, con un precio medio que ronda los 1.000€/mes, que justamente coinciden con los distritos que tienen mayor demanda.



Mapa 2: Mapa de densidad de la demanda promedio de cada distrito de Madrid. Fuente: Elaboración propia en Tableau

Comparando ambos mapas se podría decir que el precio promedio tiene cierta relación con la demanda promedio, ya que en los distritos donde el precio medio es mayor también lo es la demanda y viceversa. Esto se explica por la cantidad de ofertas de alojamiento en el distrito Centro, según *Inside Airbnb* hay más de 21.000 ofertas en la ciudad de Madrid, de las cuales el 45% están localizadas en este distrito.

En relación con la demanda, también puede ser interesante tener información sobre la ocupación mensual en la ciudad de Madrid desde 2012 a 2022 para tener una idea sobre cuándo sería el mejor momento para poner un alojamiento en alquiler. Según el INE, en la encuesta de ocupación de apartamentos turísticos, los meses de mayor afluencia de turistas corresponden con el segundo semestre del año, más concretamente en los meses de septiembre y octubre (Gráfico 28). La tendencia general de la demanda es creciente a pesar de las secuelas que dejó el Covid (año 2020 e inicios de 2021) y la guerra entre Rusia y Ucrania (2022). Por lo que este es un buen momento para cualquier individuo tenga pensado alquilar su propiedad por un periodo corto de tiempo.

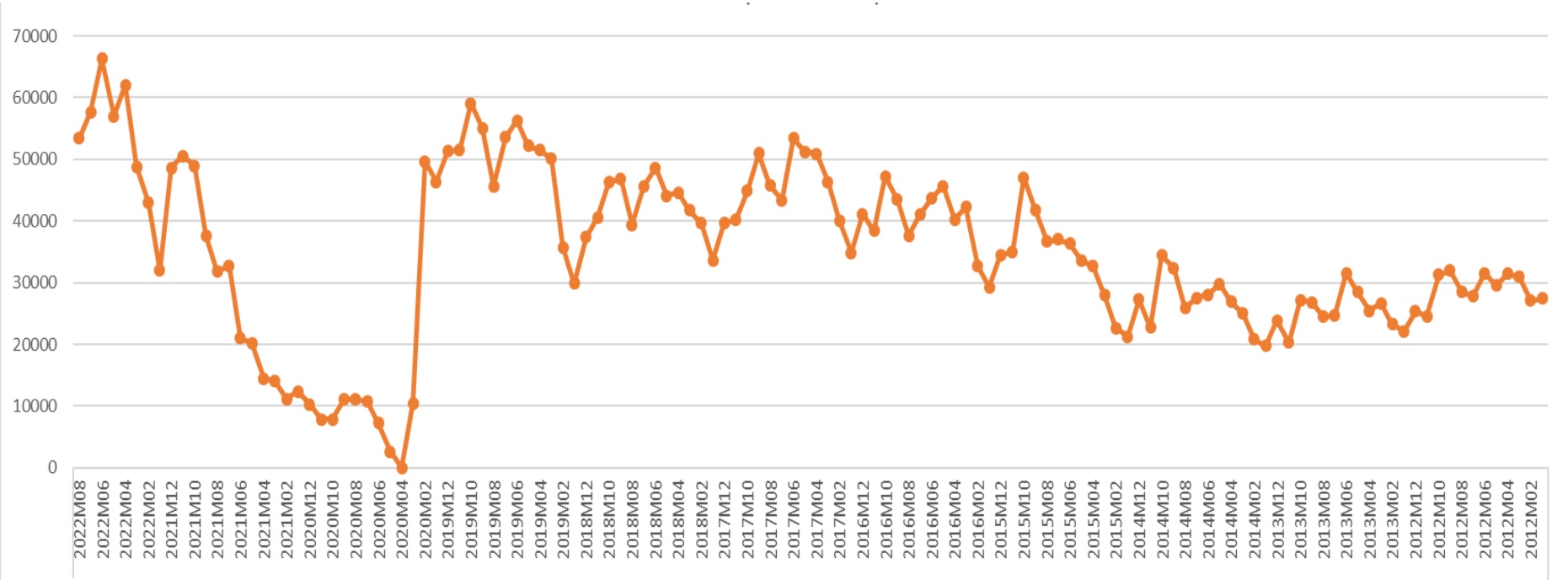


Gráfico 28: Encuesta de ocupación mensual de apartamentos turísticos en Madrid de 2012 a 2022. Fuente: Instituto Nacional de Estadística

También ha parecido atractivo visualizar en un mismo gráfico el tipo de habitación, separado por tres rangos de precio (bajo, medio y alto): menor a 50€/noche, entre 50 y 100€/noche y mayor a 100€/noche.

De esta forma se sabe que mayor parte de los apartamentos son completos, siendo las habitaciones de hotel y las compartidas las menos demandadas. Enfocándose en los apartamentos completos, la mitad tienen un precio alto (2.612 apartamentos) y el 40% un precio medio (2.226 apartamentos). Al contrario que en las habitaciones privadas, más del 60% tienen un precio bajo (2.648 apartamentos) y tan solo un 12% se consideran de precio elevado (4.524 apartamentos). Lo cual quiere decir que las habitaciones privadas de precio bajo y los apartamentos enteros de precio alto son los más demandados, estando bastante polarizado la cantidad de dinero que los clientes están dispuestos a destinar en su alojamiento.

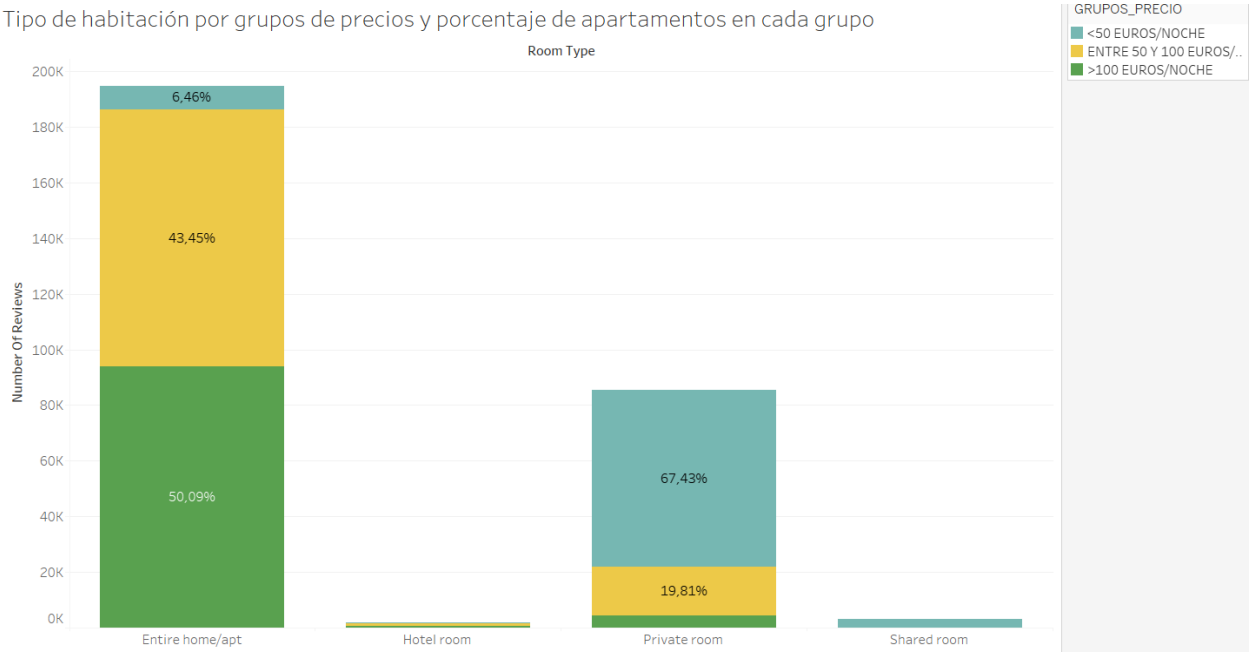


Gráfico 29: Diagrama de barras del tipo de habitación por grupos de precios y porcentaje de apartamentos en cada grupo. Fuente: Elaboración propia en Tableau

Para completar el anterior gráfico se han añadido los distritos pertenecientes a la ciudad de Madrid, separados por los rangos de precios y cuánto representan en porcentaje (Véase Apéndice B).

El porcentaje de subtotales por rango de precios es bastante similar en todos los grupos, entorno al 33%. Pero al analizarlo por distritos, los apartamentos de Centro y Salamanca predominan en el rango de precios alto (mayor de 100€/noche),

Arganzuela, Tetuán, Chamartín y Chamberí tienen mayor demanda en el rango entre 50 y 100€/noche y Moratalaz, Villaverde, Usera y puente de Vallecas son más demandados en el rango de precios más bajo (menor de 50€/noche).

Lo último que se va a exponer de este apartado es una nube de palabras de los servicios que ofrecen los apartamentos, y diagramas de barras con las palabras más y menos recurrentes, tanto de los comentarios que dejan los huéspedes como de los servicios disponibles. Se debe aclarar que esto no es lo mismo que un análisis de sentimiento, ya que para ello habría que hacer un estudio más elaborado, simplemente se ha querido puntualizar los adjetivos (para la variable comentarios) y sustantivos y verbos (para la variable servicios) más repetidos, por si estos pueden aportar algo de información al estudio.

De la base de listings_detalle se ha usado la variable servicios y se ha hecho un conteo de las veces que aparece cada palabra, posteriormente se ha realizado un diagrama de barras de las diez palabras más repetidas (Gráfico 30) y otro para las diez menos repetidas (Gráfico 31). Las palabras más repetidas (alrededor de 10.000 veces) son “agua”, “wifi” y “secador” y algunas de las menos frecuentes (alrededor de 1.000 veces) “vistas a la ciudad”, “bañera” y “fumar” y “mascotas”. Esto da una idea de cuáles son los servicios más básicos y que tienen la mayoría de los apartamentos y cuáles son los más exclusivos. Por lo que, si se quiere que un apartamento tenga más posibilidades de ser alquilado, es necesario que disponga de los servicios más demandados. Sin embargo, los servicios menos demandados son un añadido al valor del apartamento, pero se puede prescindir de ellos.

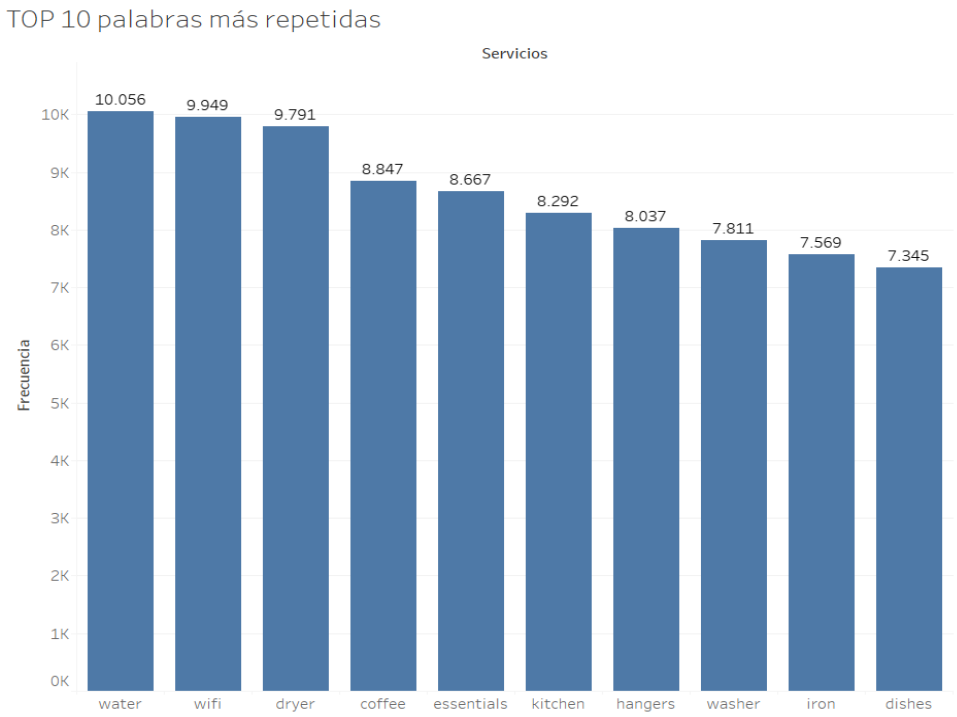


Gráfico 30: Diagrama de barras con los diez servicios más repetidos en los apartamentos alquilados. Fuente: Elaboración propia en Tableau

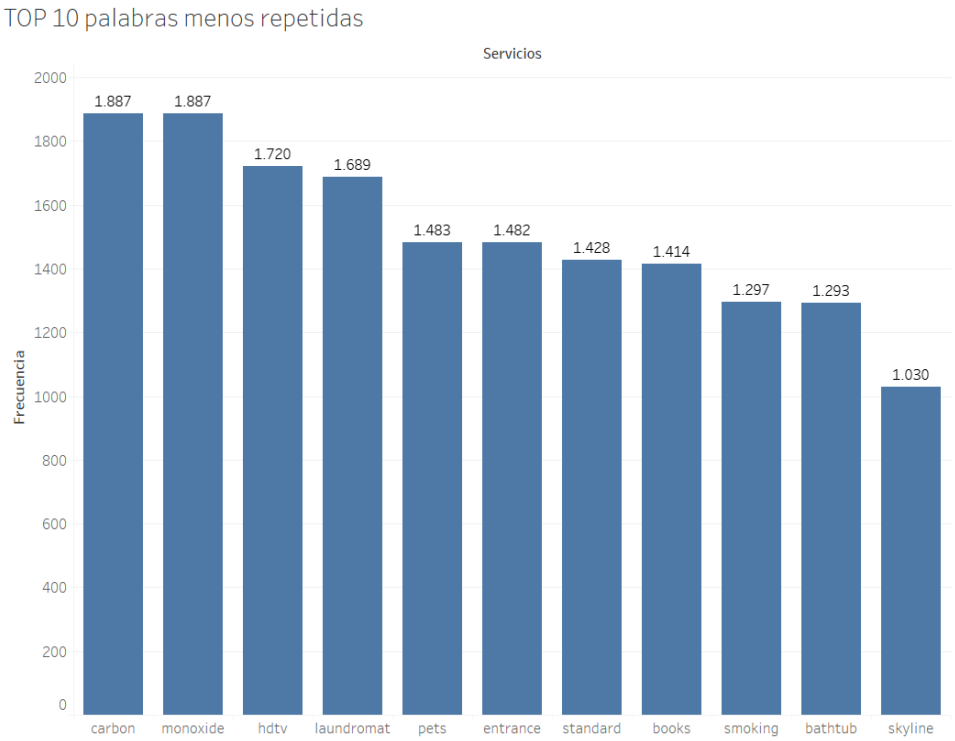


Gráfico 31: Diagrama de barras con los diez servicios menos repetidos en los apartamentos alquilados. Fuente: Elaboración propia en Tableau

A continuación, se visualiza la nube de las treinta palabras más repetidas dibujado con una frecuencia de saturación de color (a más claro sea el color, menos veces se repite y viceversa)

Gráfico 32: Nube de palabras con los treinta servicios más repetidos en los apartamentos alquilados Fuente: Elaboración propia en Tableau

De la base de reviews_madrid se ha usado la variable comentarios, se ha hecho un conteo de las veces que aparece cada palabra y se han visualizado los siete adjetivos más repetidos por los huéspedes (Gráfico 33). De estos todos son positivos, siendo "genial" el más usado por los clientes (más de 3.000 veces).

Al ser todos los adjetivos positivos, se puede pensar que los clientes están muy satisfechos con los apartamentos alquilados en Madrid, pero no tiene por qué ser así, ya que varía mucho en función de las palabras anteriores y posteriores de estos adjetivos, que pueden darle un significado u otro (Ejemplo: la experiencia en el apartamento ha sido muy buena o la experiencia en el apartamento no ha sido nada buena). Para poder analizar los cometarios correctamente haría falta hacer un análisis más exhaustivo, utilizando técnicas de procesamiento del lenguaje natural (NLP)[6], que puede valorarse para próximos trabajos.

[6] "*Rama de la inteligencia artificial que ayuda a las computadoras a entender, interpretar y manipular el lenguaje humano*" (SAS, 2023)

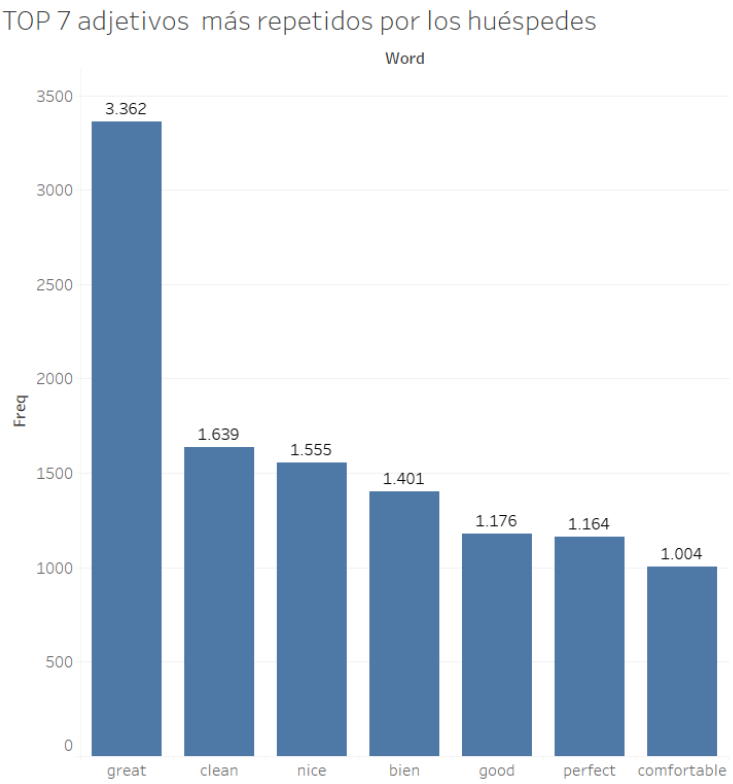


Gráfico 33: Diagrama de barras con los siete adjetivos más repetidos por los huéspedes en los comentarios. Fuente: Elaboración propia

5.1 Modelos de regresión lineales: simples y múltiples

Como se ha mencionado en el apartado 5.1, el precio y la demanda suelen estar relacionados. Además, el fuerte desarrollo turístico y el aumento del número de viajeros en Madrid (Gráfico 28), ha situado al alquiler turístico como un negocio muy rentable, haciendo las viviendas de alquiler vacacionales más atractivas que las habituales aumentando el precio de estas últimas. (Martínez del Olmo, 2018)

Para contrastar lo explicado se realiza una comprobación numérica del precio y demanda de la base de datos de listings_madrid, utilizando el coeficiente de determinación el resultado arroja un valor positivo pero muy cercano a cero.

Para cerciorarse de que la evidencia numérica es correcta, se realiza una comprobación gráfica, para ver si puede existir una correlación no lineal (Véase Apéndice B). En este caso, no se puede apreciar un patrón entre las dos variables, con lo que no podemos relacionarlos con ningún tipo de regresión.

Fijándose en la base de datos de listings_detalle, se encuentran más variables numéricas por lo que es mucho más útil usar los modelos de regresión lineales simples y múltiples.

Primero se estudia si existe alguna correlación lineal entre pares de variables de las puntuaciones y tras confirmar que existe dicha correlación, se visualiza una matriz de correlaciones en pares de todas las puntuaciones entre sí, excluyendo la puntuación total.

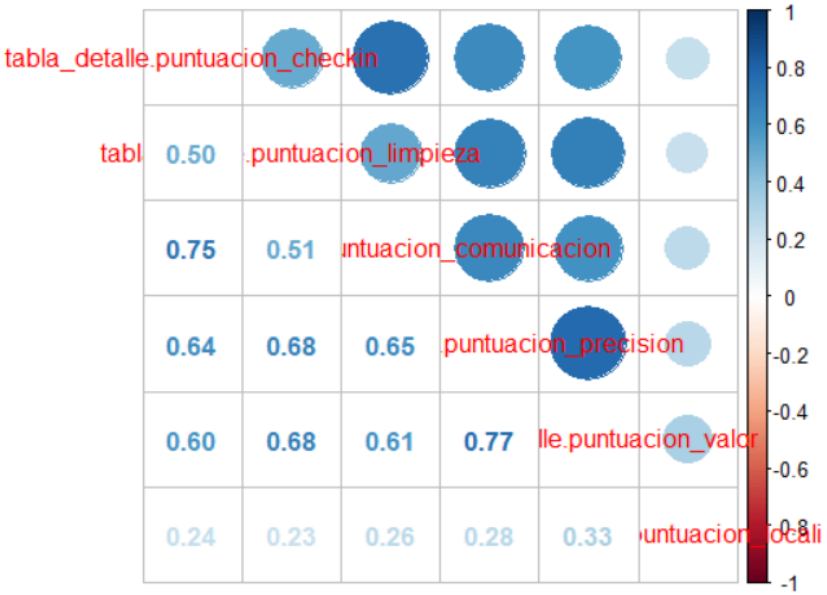


Gráfico 34: Matriz de correlaciones entre las puntuaciones. Fuente: elaboración propia

Del gráfico 34 se puede concluir que existe cierta dependencia entre las puntuaciones excepto para la puntuación de la localización, que no tiene una relación alta con ninguna otra puntuación. Con lo cual, cuando un potencial propietario esté alquilando su apartamento en Airbnb, no debería de fijarse donde esté localizado para determinar si su apartamento va a tener éxito en el mercado, ya que es indiferente.

Se ha obtenido que las puntuaciones con mayor correlación entre sí son:

- Puntuación de la comunicación con la puntuación del checkin del apartamento con un coeficiente de determinación (R^2) del 0.75, esto quiere decir que la variable dependiente (puntuación de comunicación) explica en un 75% la variable independiente (puntuación del valor)

- La puntuación de limpieza con la puntuación de precisión y de limpieza con valor, ambas con un coeficiente de determinación (R^2) del 0.68.
- Puntuación de la precisión con la puntuación del valor, con el coeficiente de determinación más alto, del 0.77.
- Número de habitaciones con el número de camas con un R^2 de 0.71.

Como el porcentaje de correlación es mayor o muy cercano al 70% en todos los casos se considera que hay una cierta correlación lineal entre las variables.

A continuación, se comprueba gráficamente si las variables continuas siguen una relación lineal y positiva.

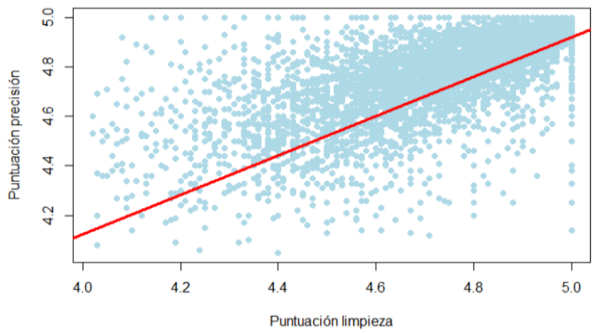


Gráfico 35: Gráfico de puntos entre las puntuaciones de limpieza y precisión. Fuente: Elaboración propia

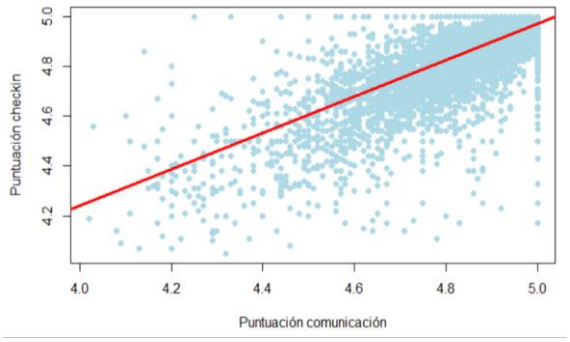


Gráfico 36: Gráfico de puntos entre las puntuaciones de comunicación y el checkin. Fuente: Elaboración propia

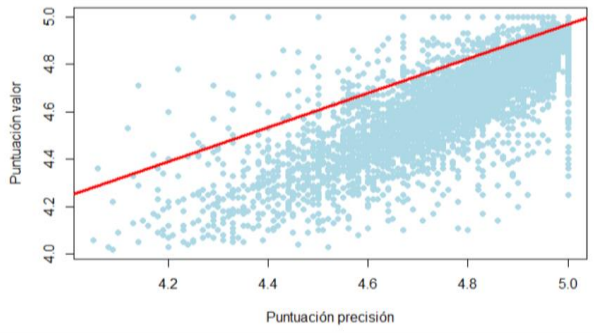


Gráfico 37: Gráfico de puntos entre las puntuaciones de precisión y valor. Fuente: Elaboración propia

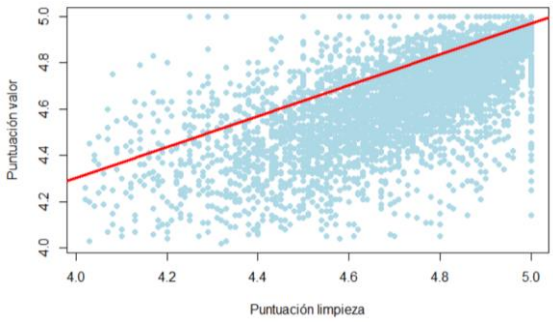


Gráfico 38: Gráfico de puntos entre las puntuaciones de limpieza y valor. Fuente: Elaboración propia

Como se esperaba por los resultados de la matriz de correlaciones, se puede ver claramente una correlación lineal y directa entre las variables de puntuación, variando los valores en un intervalo entre 4 y 5, acumulándose más en la puntuación máxima,

con lo que los huéspedes están muy satisfechos con la limpieza, la precisión, el valor y la comunicación de los apartamentos alquilados.

A partir de las variables aplicadas en los modelos de regresión simple se han realizado los siguientes modelos de regresión múltiple con sus correspondientes contrastes de hipótesis que pueden verse en el Apéndice B.

Fijándose en el p-valor de los estimadores y de los coeficientes de correlación, al tener valores menores a 0,05 en todos los modelos analizados, existe relación entre la variable dependiente con las variables independientes.

La diferencia entre los modelos de puntuación (Modelo B.1, Modelo B.2 y Modelo B.4), con el modelo que usa el precio como variable dependiente el precio (Modelo B.3), es que este último tiene una desviación estándar de los errores mayor que el resto y un coeficiente de determinación del 0,27, con lo cual, aunque el p-valor indique que existe relación entre las variables, no es un buen modelo y habría que descartarlo. Mientras que en los demás modelos se obtiene un coeficiente de determinación entorno el 0,6 y una desviación estándar de los residuos muy baja, con lo que son potenciales buenos modelos.

Aun teniendo una bondad del ajuste cercano al 70% y un p-valor bajo, no se puede asegurar que los modelos sean aceptables. Para ello, hay que tener también en cuenta la distribución de los residuos, esto puede hacerse de muchos métodos diferentes, en este trabajo se va a utilizar la visualización para ver su distribución.

En toda regresión lineal clásica los residuos deben de asemejarse a la distribución de una normal (como se explica en el Apéndice A, los residuos son la diferencia entre los valores reales y los valores predichos). Cuando se observa en los gráficos 39 y 41 no se ajustan a la forma de la curva normal, por lo que no se pueden calificar estos modelos como válidos. Sin embargo, los gráficos 40 y 42 sí tienen una distribución que se ajusta bastante bien a la curva, asique sus modelos correspondientes B.2 y B.4 pueden considerarse aceptables.

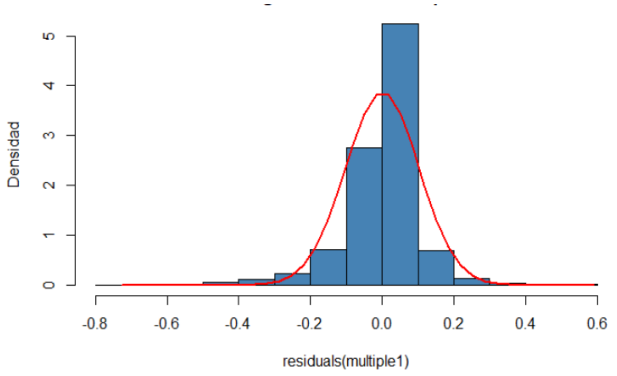


Gráfico 39: Histograma de los residuos de la regresión lineal múltiple entre las puntuaciones de checkin con comunicación y precisión

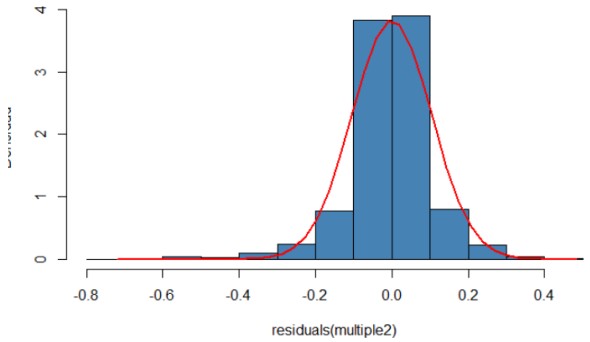


Gráfico 40: Histograma de los residuos de la regresión lineal múltiple entre las puntuaciones de precisión con valor y comunicación

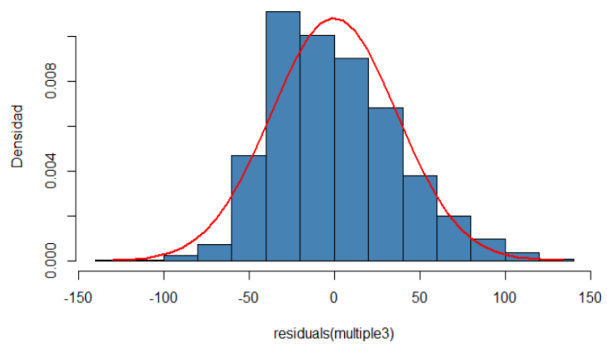


Gráfico 41: Histograma de los residuos de la regresión lineal múltiple entre las puntuaciones de precio con número de camas y de habitaciones

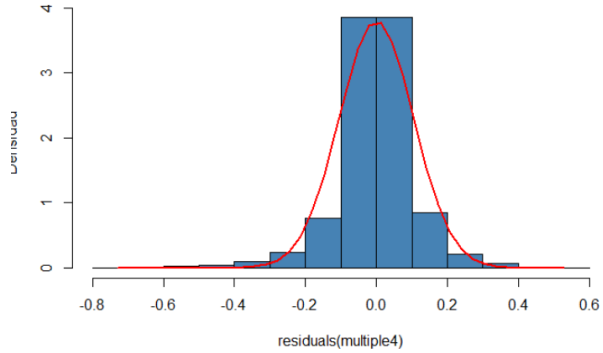


Gráfico 42: Histograma de los residuos de la regresión lineal múltiple entre las puntuaciones de precisión con limpieza y valor

Por consiguiente, se puede decir que cuanto mejor es la comunicación que tienen anfitrión y huésped y mejor es la calidad/precio del apartamento, más se asemejará el apartamento real al que aparece en la web (es decir, mejora la puntuación de la precisión). Por lo que es recomendable que, para ser lo más transparente posible con los clientes y evitar fomentar la publicidad engañosa, el propietario mantenga una comunicación constante con ellos y se asegure que las reseñas de anteriores huéspedes son positivas, y así dar a entender que calidad/precio de su apartamento es alta. Además, mantener el apartamento limpio también es de gran importancia tanto para mejorar la calidad/precio como para la precisión.

5.2 Árboles de decisión y clústeres

Una técnica muy utilizada para cuando se tienen variables cualitativas y cuantitativas son los clústeres o agrupaciones, aunque la principal diferencia con los árboles de decisión y cómo se ha mencionado en el apartado 2.2, es que se usa para datos no etiquetados y forman parte de los métodos no supervisados.

Se va a realizar un único estudio de clúster de la base de listings_madrid con las variables precio y demanda, ya que, como se ha explicado en la sección 5.2, suelen estar relacionadas. Al no visualizar una correlación fuerte en los modelos de regresión lineal, se ha probado a analizarlo desde una perspectiva más visual.

Antes de hacer los grupos es necesario mirar el número de medias (kmeans) óptimos para modelo. Utilizando el método del [7]codo (WSS), el número óptimo es tres, ya que a partir de ese número el total de la suma de los cuadrados comienza a estabilizarse (no hay apenas diferencia entre un modelo y el siguiente). Después se dibuja el gráfico escalando los ejes [8]para que las variables sean comparables y se asigna un color a cada clúster (Gráfico 43).

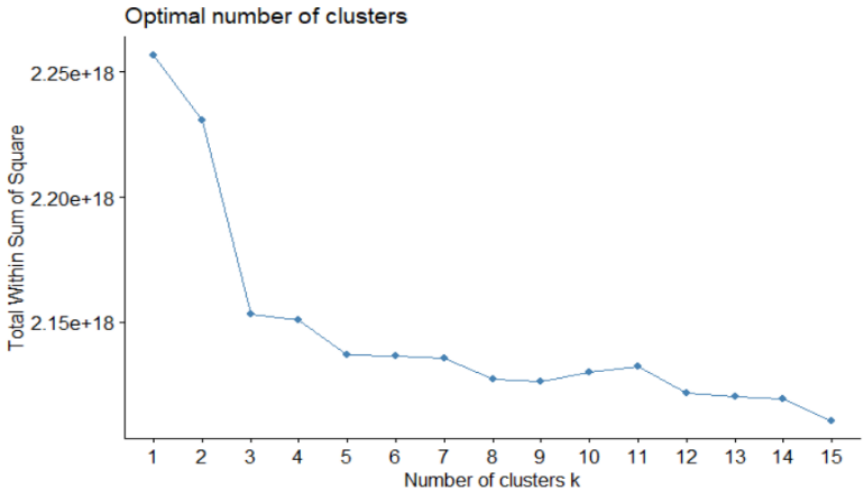


Gráfico 43: Gráfico de línea del número óptimo de clústeres entre las variables precio y demanda. Fuente: Elaboración propia

[7] "Método para encontrar el número óptimo de grupos para un conjunto de datos. Se seleccionan un rango de valores candidatos de k, luego se aplica el agrupamiento de K-medias usando cada uno de los valores de k" (Sphinx-Gallery, 2007). Se calcula la distancia promedio de cada punto en un grupo a su centroide y se representa en una gráfica, eligiendo el valor de k, donde la distancia promedio disminuye bruscamente

[8] "Las variables deben transformarse para que estén en una escala similar y puedan compararse correctamente. Existen diferentes métodos para abordar este problema. La más conocida y utilizada es la estandarización, que consiste en restar el valor promedio del valor de la característica y dividirlo por su desviación estándar" (Datacamp, 2023).

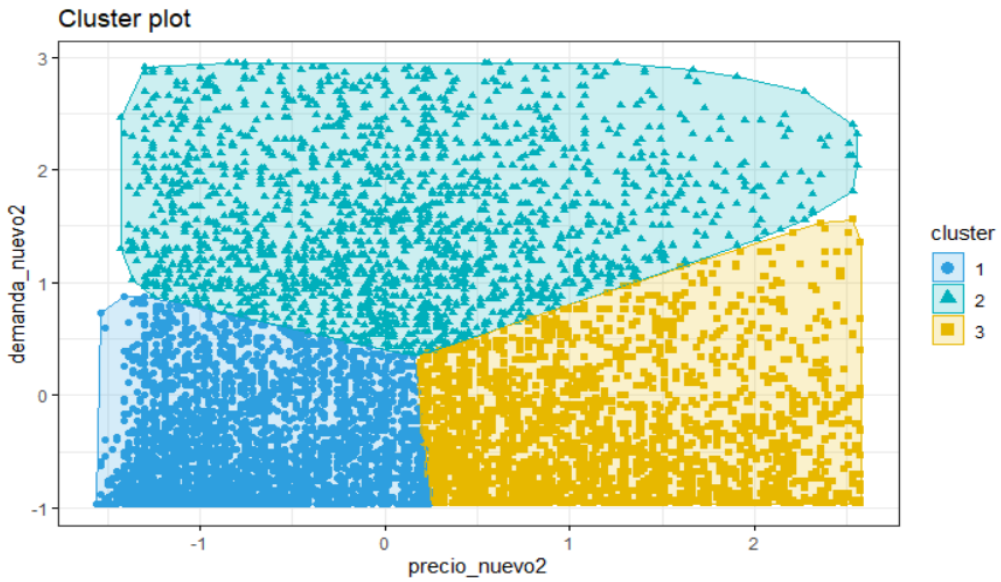


Gráfico 44: Gráfico de clústeres entre las variables precio y demanda. Fuente: Elaboración propia

En el primer grupo se encuentran valores más bajos tanto en precio como en demanda, en el segundo los valores más altos en demanda y con un precio variado y en el tercer grupo los datos con precio más alto y demanda baja.

Para entender el Gráfico 44 numéricamente, se calcula la media y la cantidad de variables que hay en cada grupo.

- Primer clúster: se obtiene una media de 80€/noche, con 1.503 apartamentos enteros, 556 habitaciones privadas, 6 habitaciones compartidas y ninguna habitación de hotel y el distrito más repetido es Centro con 1.007 apartamentos.
- Segundo clúster: se obtiene una media de 48€/noche, con 1.320 apartamentos enteros, 2.387 habitaciones privadas, 51 habitaciones de hotel y 113 habitaciones compartidas y el distrito más común es Centro con 1.335 apartamentos.
- Tercer clúster: se obtiene una media de 134€/noche, con 1.975 apartamentos enteros, 195 habitaciones privadas, 51 habitaciones de hotel y 1 habitación compartida y el distrito más repetido es el Centro con 1.328 apartamentos.

El distrito más repetido es el Centro, sin embargo, en el segundo grupo predominan las habitaciones privadas mientras que en los otros predominan los apartamentos enteros. Esto indica que un tercio de los inquilinos prefieren alquilar habitaciones individuales en el Centro y compartiendo apartamento con otras personas, dando prioridad a estar cerca de los lugares turísticos antes que a la privacidad.

Sabiendo que en el segundo clúster el Centro es el distrito donde se alquilan más apartamentos, puede ser interesante analizar la correlación entre el precio y la demanda para intentar encontrar una dependencia más clara entre las variables de esta forma. Por el contrario, se continúa sin encontrar una relación fuerte entre el precio y la demanda, siendo 0,14 el coeficiente de variación, y al visualizar la gráfica tampoco parece que los datos sigan un patrón claro (Gráfico 45).

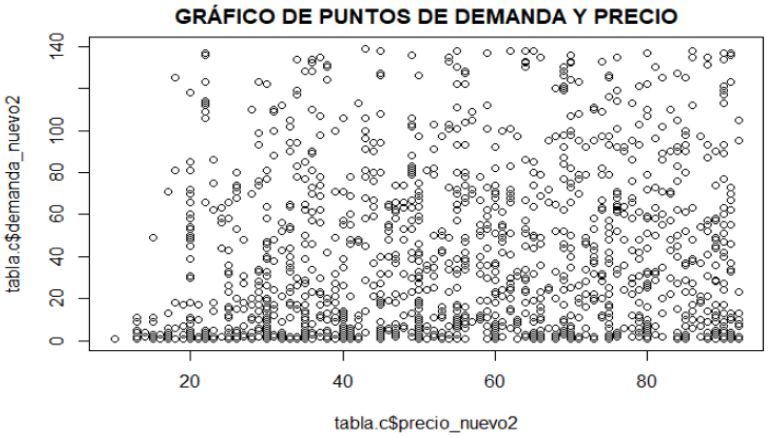


Gráfico 45: Gráfico de puntos de las variables precio y demanda discriminado por distrito Centro y clúster 2. Fuente: Elaboración propia

A pesar de ello, comparándolo con la correlación entre el precio y la demanda total (sin clústeres) del distrito Centro, hay aumento del R^2 entre las variables de más del 400%, con lo que el uso de agrupación por clústeres es de utilidad para mejorar la dependencia entre variables que están relacionadas entre sí. Se ha realizado el mismo estudio con los distritos más solicitados de Madrid (Arganzuela y Salamanca) del segundo clúster para ver cómo varía la bondad del ajuste de estas variables (precio y demanda) utilizando el agrupamiento y sin él. Para el distrito de Salamanca no se aprecia a penas diferencia entre los coeficientes de correlación variando de 12% (sin clústeres) a 14% (con clústeres), pero para el distrito de Arganzuela este coeficiente aumenta a más del doble (del 11% al 28%) cuando se analiza dentro del clúster. Aun así, cuando se realizan los gráficos de puntos se sigue sin encontrar un patrón definido (Mirar en Apéndice B).

Otra técnica muy utilizada para el análisis cuando hay datos numéricos y categóricos son los árboles de decisión. En este estudio se han aplicado únicamente en la base de listings_madrid, ya que esta contiene varias variables que podrían estar

relacionadas entre sí, y no se han podido analizar con modelos de regresión. En este caso se consideran árboles de clasificación ya que se usan variables cualitativas.

En todos los árboles se ha iniciado el primer nodo con la variable tipo de habitación, hotel_room y shared_room no aparecen en el resultado porque hay muy poca cantidad de datos con estas características. También se han añadido a los modelos las variables demanda, número mínimo de noches y distrito.

Los tipos de habitación de los apartamentos están formados por un 59% de apartamentos completos, 1% de habitaciones de hotel, 39% de habitaciones privadas y 1% de habitaciones compartidas.

El primer nodo del Gráfico 46 se divide en 2 subgrupos en función del número mínimo de noches que pasen los huéspedes en el alojamiento, hay un 59% de probabilidad de que sean igual o más de 2, en cuyo caso es un 69% de probable que sea un apartamento entero. Por el contrario, hay un 41% de probabilidad de que pasen menos de 2 noches, en tal caso es un 45% de probable que escojan una habitación privada. Este último subgrupo se divide en otros 2 subgrupos en función de la demanda, si la demanda del apartamento es menor que 7 hay más probabilidad de que el huésped escoja alojarse en una habitación privada (62%), en caso contrario hay más probabilidad de que escoja un apartamento entero (51%).

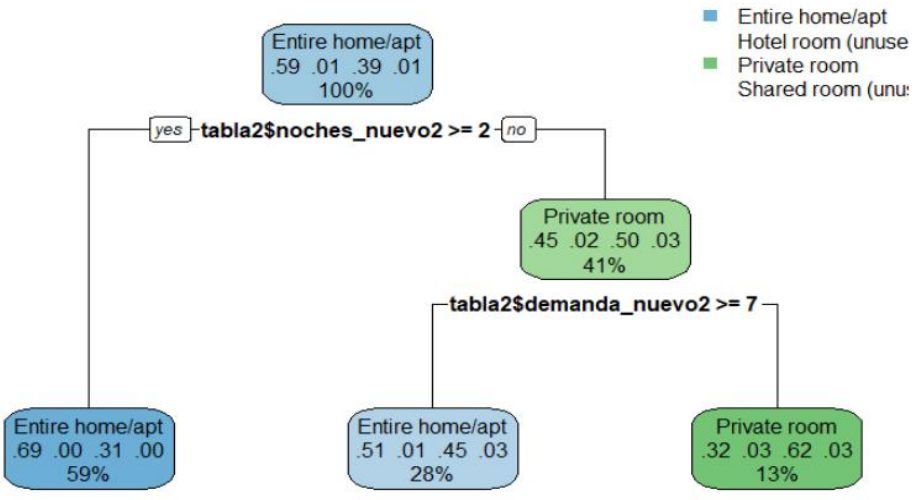


Gráfico 46: Diagrama de árbol con las variables tipo de apartamento y demanda. Fuente: Elaboración propia

El número mínimo de noches que los huéspedes pasan en un apartamento completo suele ser mayor que para las habitaciones privadas. Y si un apartamento está más

solicitado (mayor demanda) los huéspedes preferirán uno entero, en caso contrario optarán por una habitación privada.

En el Gráfico 47 el primer nodo se divide en 2 subgrupos en función del distrito en el que se encuentre el alojamiento hay un 73% de probabilidad de que esté situado en Arganzuela, Centro, Chamartín, Chamberí, Salamanca, Retiro o Tetuán, en cuyo caso el tipo de habitación es un 66% de probable que sea un apartamento entero. Por el contrario, hay un 21% de probabilidades de que se ubique en un distrito diferente a los anteriores, en tal caso hay un 59% de probabilidad de que escojan una habitación privada. Este último subgrupo se divide en otros 2 subgrupos en función del número mínimo de noches que decide quedarse el cliente, si es menor que 2 hay una gran probabilidad de que el huésped escoja alojarse en una habitación privada (74%), en caso contrario hay más probabilidad de que escoja un apartamento entero (51%). Si está dentro de este último subgrupo, se procede a mirar otra vez el número mínimo de noches, si fue es menor que 3, es más probable que haya alquilado una habitación privada (54%), y en caso de ser mayor o igual a 3 hay un 60% de probabilidad de que haya alquilado un apartamento entero.

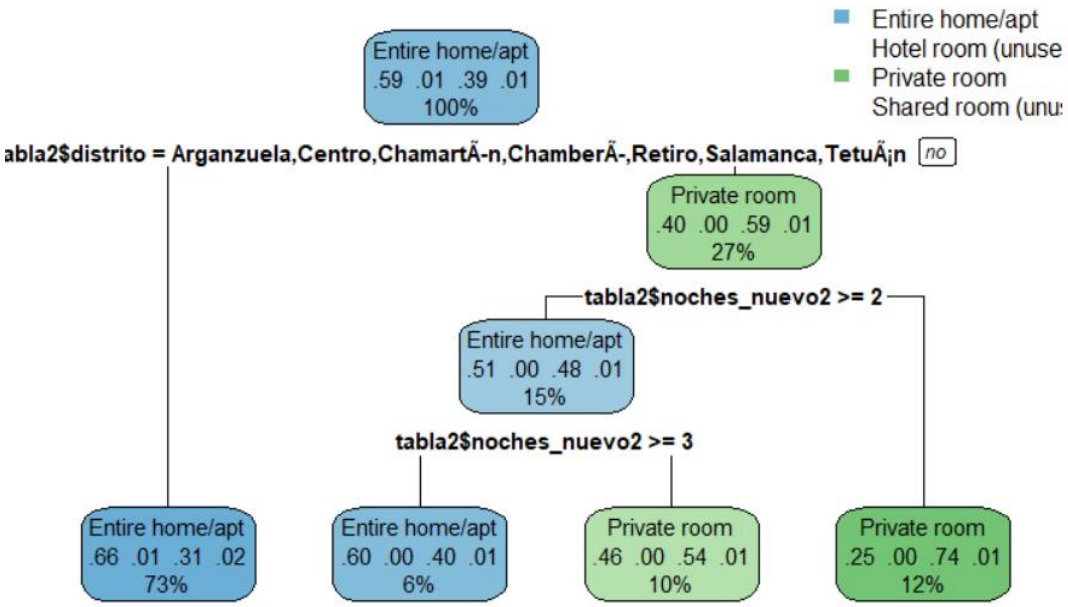


Gráfico 47: Diagrama de árbol con las variables distrito y número mínimo de noches. Fuente: Elaboración propia

Lo que se entiende de este diagrama es que la mayoría de los apartamentos se localizan en los distritos más céntricos de Madrid y suelen ser apartamentos enteros, mientras que, si el apartamento está fuera de estos distritos, pueden encontrarse en un porcentaje similar habitaciones privadas y apartamentos completos. Como aparece en el Gráfico 46 el número de noches en apartamento completo suele ser mayor que para las habitaciones privadas.

6. CONCLUSIONES

El sector hotelero español está en auge a pesar de las condiciones económicas en las que se encuentra el país, concretamente el alquiler de los apartamentos turísticos en la ciudad de Madrid. Esto indica que es buen momento para poner en alquiler su propiedad por un periodo corto de tiempo, pero fijarse en la demanda no es suficiente para lograr alquilar su propiedad.

Un factor fundamental y que suele estar relacionado con la demanda es el precio que decida poner de alquiler. A pesar de que los datos de este estudio no muestren una relación lineal significativa entre ellos, sí que se observa un aumento del coeficiente de determinación de estos tras aplicar los clústeres.

Para asesorar a cualquier individuo que quiera poner su inmueble en alquiler es necesario saber qué tipo de alojamiento dispone, ya que esto va a condicionar la decisión final (no se han tenido en cuenta las habitaciones compartidas y las habitaciones de hotel porque la demanda es muy baja), por ello se va a partir de 2 hipótesis:

- Primera hipótesis: El propietario desea poner en alquiler una habitación privada

Las habitaciones privadas son las más demandadas, pero el precio de estas suele ser bajo (menor a 50 €/noche), y los inquilinos no suelen pasar más de dos noches en ellas. Y, aunque en el distrito Centro se alquilen la mayor parte de los apartamentos, para una habitación privada es aconsejable que esté situada fuera del centro de Madrid, ya que es donde hay una mayor demanda de estas.

- Segunda hipótesis: El propietario desea poner en alquiler un apartamento completo

Los apartamentos completos son los más demandados en el centro, que es además el distrito más demandado en la ciudad de Madrid. El precio suele ser alto (mayor a 100€/noche) y los huéspedes las alquilan por más de dos noches, por lo que es más rentable que una habitación privada (fijándose únicamente en los ingresos). En este caso habrá que tener en cuenta el número de habitaciones que tiene el alojamiento, no siendo aconsejable poner en alquiler una propiedad con más de dos habitaciones por apartamento porque es menos probable que los clientes se fijen en esta.

En ambos casos es recomendable que se alquile en el segundo semestre del año, preferiblemente en septiembre y octubre, ya que son los meses con mayor número de apartamentos alquilados entre 2012 y 2022.

Otro factor relevante son las puntuaciones que dejan los huéspedes una vez han terminado su estancia, que son valoraciones (en estrellas) del 1 al 5 de todas las características importantes que debe tener un buen apartamento (siendo 1 pésimo y 5 excelente). Esto le sirve a usted (anfitrión) de referencia para mejorar en aquellas características donde la puntuación es más baja, y para los potenciales clientes, siendo un testimonio real que les ayuda a decidir si alquilar o no ese alojamiento. Tras el análisis de las puntuaciones, se concluye que a mejor comunicación entre anfitrión y huésped y mejor calidad/precio tenga el alojamiento, más alta será la puntuación de precisión. Por lo que es recomendable que sea lo más transparente posible con los clientes y evite fomentar publicidad engañosa para su propio beneficio, ya que esto le va a repercutir negativamente en las valoraciones. La comunicación con los clientes debe ser constante y debe asegurarse que las reseñas de anteriores huéspedes son positivas (en caso contrario mejorar sus puntos débiles), y así dar a entender que calidad/precio de su apartamento es alta. Además, mantener el apartamento limpio también influye tanto en la comunicación, como en el valor y la precisión.

Para cualquier tipo de alojamiento hay que tener en cuenta los servicios mínimos que debe de ofrecer el anfitrión, intentando en la medida de lo posible disponer de los más repetidos en los apartamentos que han sido alquilados (agua, wifi, perchas, lavadora, plancha…) y evitar centrarse en adquirir los menos recurrentes (mascotas, bañera, fumar, vistas a la ciudad…). Los comentarios sobre su apartamento tienen la misma

importancia que las valoraciones. Con la nube de palabras se puede intuir (teniendo en cuenta solo los adjetivos) que los comentarios más comunes son positivos, pero habría que estudiarlos en detalle a través de un análisis de sentimiento, para conocer el tipo de reacción general de los usuarios (positiva o negativa) y por qué motivo.

7. APÉNDICES

7.1 Apéndice A

En este apéndice se muestran las fórmulas estadísticas usadas en el trabajo.

- Coeficiente de correlación lineal o R^2

  Mide la varianza de la variable dependiente Y explicada por la variable independiente X

$$R^2 = \frac{TSS - RSS}{TSS} \qquad (A.1)$$

donde el numerador corresponde con la varianza explicada, formada por: $TSS = \sum(y_i - \bar{y})^2$ que mide la varianza total y $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ que mide la varianza de los residuos, todo ello está dividido entre la varianza total. (James, 2021)

Cuando R^2 está cerca de 1 indica que no hay varianza de los residuos, por lo tanto, la varianza total coincide con la varianza explicada, con lo cual la variable X predice con un alto porcentaje a la variable Y. Por el contrario, cuando R^2 está muy cercano al 0, puede significar que no existe relación entre las variables, o que si existe relación no es lineal.

- P-valor

El p-valor sirve para aceptar o rechazar una hipótesis, y sirve para cerciorarse de que el resultado de R^2 es estadísticamente significativo. Siendo el punto de corte crítico para rechazar o aceptar la hipótesis 0,05, cuando el p-valor es menor a esta cantidad se rechaza la hipótesis nula de $H_0: \beta_1 = 0$ y se acepta la hipótesis alternativa de $H_0: \beta_1 \neq 0$ asumiendo que hay relación entre las variables. En caso de que el p-valor sea mayor a 0,05 no se rechaza la hipótesis nula con lo cual hay una gran probabilidad de que las variables no estén relacionadas o que no exista una relación lineal entre ellas. (James, 2021)

- Desviación estándar de los residuos o RMSE

El RMSE es una estimación de la desviación estándar del error de los valores observados respecto a los valores predichos por el modelo, es decir, la cantidad promedio que la respuesta se desviará de la línea de regresión.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (A.2)$$

donde y_i son los valores observados y $\hat{y}_i$ son los valores predichos del modelo.

7.2 Apéndice B

En este apéndice se muestran las tablas y gráficos que complementan la información el estudio realizado.

GRÁFICOS

Gráfico B.1

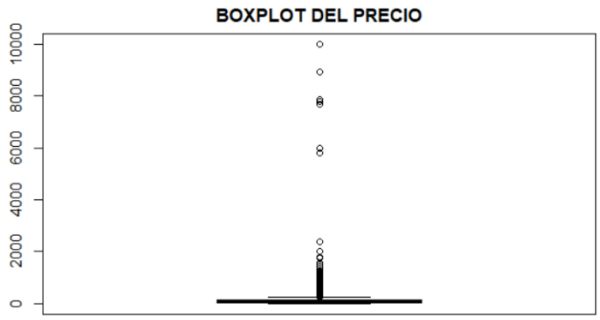


Diagrama de cajas del precio de la base de listings_madrid. Fuente: Elaboración propia

Gráfico B.2

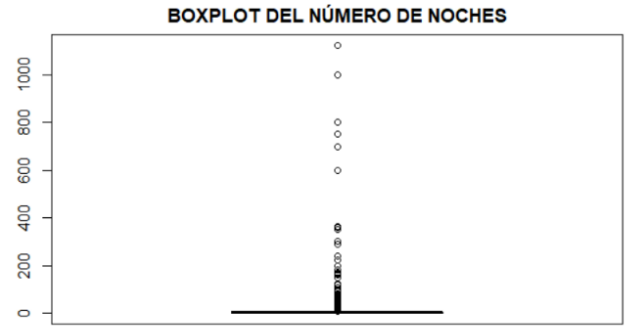


Diagrama de cajas del número mínimo de noches de la base de listings_madrid. Fuente: Elaboración propia

Gráfico B.3

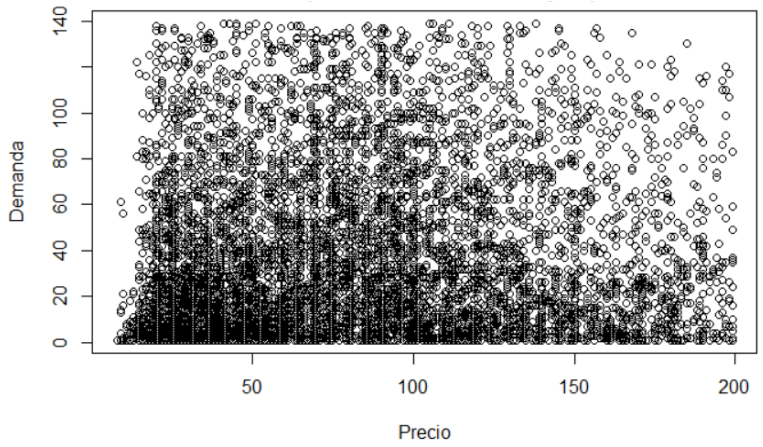


Gráfico de puntos entre el precio y la demanda. Fuente: Elaboración propia

Gráfico B.4

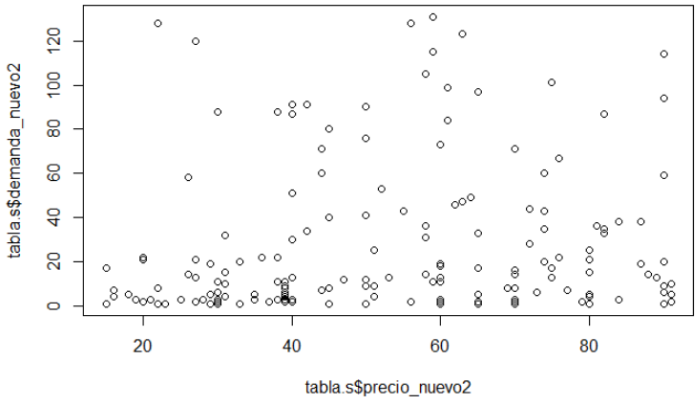


Gráfico de puntos de las variables precio y demanda discriminado por distrito Salamanca y clúster 2. Fuente: Elaboración propia

Gráfico B.5

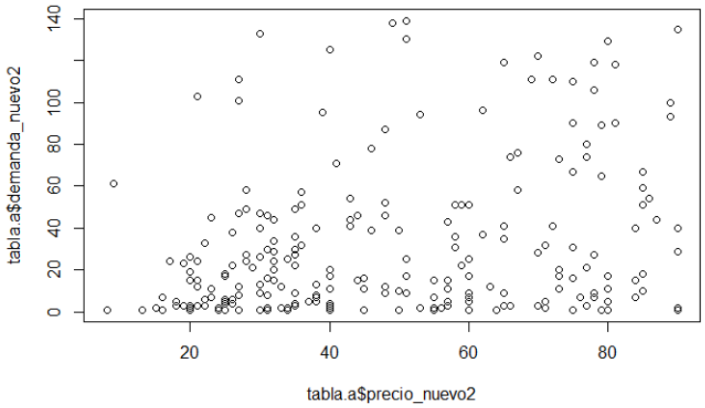


Gráfico de puntos de las variables precio y demanda discriminado por distrito Arganzuela y clúster 2. Fuente: Elaboración propia

MODELOS DE REGRESIÓN LINEAL MULTIPLE

Modelo B.1

```
Call:
lm(formula = tabla_detalle$puntuacion_checkin ~ tabla_detalle$puntuacion_comunicacion +
    tabla_detalle$puntuacion_precision)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72569 -0.03368  0.02188  0.04422  0.59054

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       0.786136   0.038921   20.20   <2e-16 ***
tabla_detalle$puntuacion_comunicacion 0.589090   0.009972   59.07   <2e-16 ***
tabla_detalle$puntuacion_precision    0.248820   0.009003   27.64   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1031 on 7288 degrees of freedom
Multiple R-squared:  0.5993,    Adjusted R-squared:  0.5992
F-statistic:  5450 on 2 and 7288 DF,  p-value: < 2.2e-16
```

Primer modelo múltiple: Puntuación del checkin respecto a las puntuaciones de comunicación y precisión. Fuente: Elaboración propia

Modelo B.2

```
Call:
lm(formula = tabla_detalle$puntuacion_precision ~ tabla_detalle$puntuacion_valor +
    tabla_detalle$puntuacion_comunicacion, data = tabla_detalle)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72060 -0.03569 -0.00020  0.04742  0.48393

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      0.958306   0.037764   25.38   <2e-16 ***
tabla_detalle$puntuacion_valor       0.502209   0.007238   69.38   <2e-16 ***
tabla_detalle$puntuacion_comunicacion 0.307379   0.009676   31.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.104 on 7288 degrees of freedom
Multiple R-squared:  0.6486,    Adjusted R-squared:  0.6485
F-statistic:  6725 on 2 and 7288 DF,  p-value: < 2.2e-16
```

Segundo modelo múltiple: Puntuación de la precisión respecto a las puntuaciones de valor y comunicación. Fuente: Elaboración propia

Modelo B.3

```
Call:
lm(formula = tabla_detalle$precio_detalle2 ~ tabla_detalle$camas2 +
    tabla_detalle$habitaciones2)

Residuals:
     Min       1Q   Median       3Q      Max
-129.872  -28.901   -4.317   23.683  134.099

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 35.2440     1.0495   33.58   <2e-16 ***
tabla_detalle$camas2        14.4158     0.5712   25.24   <2e-16 ***
tabla_detalle$habitaciones2 15.2412     0.9967   15.29   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.82 on 7288 degrees of freedom
Multiple R-squared:  0.2787,    Adjusted R-squared:  0.2785
F-statistic:  1408 on 2 and 7288 DF,  p-value: < 2.2e-16
```

Tercer modelo múltiple: Precio respecto a número de camas y número de habitaciones. Fuente: Elaboración propia

Modelo B.4

```
Call:
lm(formula = tabla_detalle$puntuacion_precision ~ tabla_detalle$puntuacion_limpieza +
    tabla_detalle$puntuacion_valor, data = tabla_detalle)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73028 -0.03598 -0.00025  0.04850  0.52866

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   1.412984   0.030464   46.38   <2e-16 ***
tabla_detalle$puntuacion_limpieza 0.234199   0.008193   28.58   <2e-16 ***
tabla_detalle$puntuacion_valor    0.485413   0.007987   60.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1053 on 7288 degrees of freedom
Multiple R-squared:  0.6402,    Adjusted R-squared:  0.6401
F-statistic:  6485 on 2 and 7288 DF,  p-value: < 2.2e-16
```

Cuarto modelo múltiple: Puntuación de la precisión respecto a las puntuaciones de limpieza y valor. Fuente: Elaboración propia

TABLAS

Tabla B.1

	2023		2022											
Distrito	**Febrero**	**Enero**	**Diciembre**	**Noviembre**	**Octubre**	**Septiembre**	**Agosto**	**Julio**	**Junio**	**Mayo**	**Abril**	**Marzo**	**Febrero**	**Enero**
Ciudad de Madrid	**16,6**	**16,5**	**16,2**	**16,2**	**16,3**	**16,3**	**16,1**	**15,7**	**15,7**	**15,4**	**15,2**	**15,0**	**15,0**	**14,7**
01. Centro	21,4	21,1	20,3	20,0	19,4	19,2	18,9	18,5	18,5	18,5	18,4	18,5	18,3	17,9
02. Arganzuela	16,3	16,3	16,1	16,1	15,8	15,7	15,7	15,3	15,3	15,2	15,3	15,1	15,0	14,7
03. Retiro	17,5	17,5	17,2	17,3	16,9	16,4	16,3	16,0	16,0	16,0	15,9	15,8	15,3	15,2
04. Salamanca	21,0	20,4	20,0	19,8	20,0	20,0	19,4	18,9	18,9	18,5	18,5	18,3	18,2	17,6
05. Chamartín	17,6	17,2	16,8	16,7	16,7	16,7	16,7	16,2	16,2	16,2	16,1	16,0	15,6	15,3
06. Tetuán	17,0	16,9	16,6	16,5	16,3	16,3	16,0	15,7	15,7	15,5	15,5	15,4	15,4	15,3
07. Chamberí	20,0	20,0	19,7	19,0	18,7	18,3	18,0	18,0	18,0	17,8	17,6	17,3	17,1	17,2
08. Fuencarral-El Pardo	13,5	13,2	13,1	13,1	13,0	12,9	12,8	12,7	12,7	12,6	12,6	12,6	12,4	12,4
09. Moncloa-Aravaca	15,4	15,3	15,3	15,6	15,8	15,8	15,6	15,2	15,2	15,1	14,7	14,7	14,3	14,2
10. Latina	13,0	13,1	13,0	13,1	12,9	12,9	12,7	12,5	12,5	12,3	12,0	12,0	12,0	12,0
11. Carabanchel	12,6	12,6	12,7	12,7	12,5	12,5	12,3	12,1	12,1	12,0	11,9	11,8	11,7	11,7
12. Usera	13,6	13,3	12,9	12,8	12,9	12,9	13,1	12,8	12,8	12,5	11,8	11,4	11,3	11,3
13. Puente de Vallecas	13,2	13,1	12,9	12,7	12,8	12,7	12,5	12,4	12,4	12,5	12,5	12,0	12,0	12,0
14. Moratalaz	12,1	12,1	12,0	11,9	11,9	12,0	11,7	11,3	11,3	11,4	11,2	11,2	11,2	11,1
15. Ciudad Lineal	14,2	14,3	14,3	14,1	13,8	13,8	13,7	13,6	13,6	13,4	13,3	13,1	13,0	13,0
16. Hortaleza	13,7	13,5	13,3	13,3	13,2	13,2	13,0	12,8	12,8	12,8	12,8	12,8	12,7	12,6
17. Villaverde	11,9	11,9	12,0	11,9	11,8	11,7	11,4	11,2	11,2	11,1	11,0	10,8	10,8	10,9
18. Villa de Vallecas	11,8	12	11,9	11,7	11,6	11,7	11,7	11,8	11,8	11,6	11,3	11,1	11,2	11,3
19. Vicálvaro	11,4	11,3	11,2	11,3	11,3	11,2	11,1	11,0	11,0	10,8	10,6	10,5	10,5	10,5
20. San Blas-Canillejas	12,7	12,4	12,5	12,5	12,4	12,3	12,2	12,1	12,1	12,0	12,0	11,8	11,6	11,5
21. Barajas	12,6	12,6	12,5	12,4	12,2	12,1	12,2	12,2	12,2	12,4	12,4	12,2	11,7	11,6

Tabla de la evolución del precio de alquiler de la vivienda habitual ($€/m^2$) según el mes por distrito. Fuente: Banco de datos del ayuntamiento de Madrid

Tabla B.2

Precios de cada apartamento separados por grupos de precio y porcetaje de apartamentos por cada grupo

GRUPOS..	Neighbourhood Gro..	% de total Recuento de Id..	Recuento definido de Id
<50 EUROS/ NOCHE	Centro	21,24%	868
	Arganzuela	35,92%	208
	Puente de Vallecas	61,64%	180
	Ciudad Lineal	57,32%	188
	Carabanchel	54,60%	190
	Retiro	39,54%	121
	TetuÃin	33,72%	145
	Latina	57,42%	178
	San Blas - Canillejas	50,26%	98
	Moncloa - Aravaca	43,36%	124
	Hortaleza	37,32%	78
	ChamberÃ	29,22%	154
	Usera	62,09%	95
	Salamanca	23,14%	118
	Fuencarral - El Pardo	44,79%	86
	Villaverde	63,16%	84
	ChamartÃn	27,56%	70
	Barajas	54,55%	30
	Moratalaz	64,79%	46
	VicÃilvaro	58,33%	35
	Villa de Vallecas	60,00%	33
	Total	33,36%	3.129
ENTRE 50 Y 100 EUROS/ NOCHE	Centro	34,77%	1.421
	Arganzuela	37,65%	218
	Puente de Vallecas	21,23%	62
	Ciudad Lineal	25,30%	83
	Carabanchel	26,15%	91
	Retiro	28,76%	88
	TetuÃin	39,30%	169
	Latina	28,71%	89
	San Blas - Canillejas	32,31%	63
	Moncloa - Aravaca	29,72%	85
	Hortaleza	34,93%	73
	ChamberÃ	35,48%	187
	Usera	24,18%	37
	Salamanca	33,92%	173
	Fuencarral - El Pardo	32,29%	62
	Villaverde	29,32%	39
	ChamartÃn	36,22%	92
	Barajas	23,64%	13
	Moratalaz	18,31%	13
	VicÃilvaro	18,33%	11
	Villa de Vallecas	18,18%	10
	Total	32,83%	3.079
>100 EUROS/ NOCHE	Centro	43,99%	1.798
	Arganzuela	26,42%	153
	Puente de Vallecas	17,12%	50
	Ciudad Lineal	17,38%	57
	Carabanchel	19,25%	67
	Retiro	31,70%	97
	TetuÃin	26,98%	116
	Latina	13,87%	43
	San Blas - Canillejas	17,44%	34
	Moncloa - Aravaca	26,92%	77
	Hortaleza	27,75%	58
	ChamberÃ	35,29%	186
	Usera	13,73%	21
	Salamanca	42,94%	219
	Fuencarral - El Pardo	22,92%	44
	Villaverde	7,52%	10
	ChamartÃn	36,22%	92
	Barajas	21,82%	12
	Moratalaz	16,90%	12
	VicÃilvaro	23,33%	14
	Villa de Vallecas	21,82%	12
	Total	33,82%	3.172

GRUPOS_PRECIO
- <50 EUROS/NOCHE
- ENTRE 50 Y 100 EUROS/..
- >100 EUROS/NOCHE

Tabla de precios de cada apartamento separados por grupos de precio y porcentaje de apartamentos por cada grupo. Fuente: Elaboración propia

8. BIBLIOGRAFÍA

Airbnb. (2023). Sobre nosotros. Obtenido de https://news.airbnb.com/es/about-us/ (Fecha de consulta: 04/02/2023)

---------. (2023). Explore the Data. Obtenido de http://insideairbnb.com/explore/ (Fecha de consulta: 20/02/2023)

Ali, J., Khan, R., Ahmad, N. Maqsood, I. (2012). Random Forests and Decision Trees. *International Journal of Computer Science Issues.*

Anello, E. (2023). K-Means Clustering in R Tutorial. Obtenido de https://www.datacamp.com/tutorial/k-means-clustering-r (Fecha de consulta: 20/02/2023)

Celdrán-Bernabeu, M.A.; Mazón, J.-N. y Giner Sánchez, D. (2018). Open Data y turismo. Implicaciones para la gestión turística en ciudades y destinos turísticos inteligentes. *Investigaciones Turísticas* (15), pp. 49-78

Cerdá, E., García, B., Such, M.J., (2020). "Análisis de la economía colaborativa en el turismo urbano. estudio de la implantación de Airbnb en Madrid y Barcelona". Universidad de Murcia

Cerdá Mansilla, E., García Henche, B., & Such Devesa, M. J. (2021). análisis de la economía colaborativa en el turismo urbano. estudio de la implantación de airbnb en madrid y barcelona. *Cuadernos de Turismo*, (47), 383–412.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T.P., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide* (pp. 9-10). U.S.A. SPSS.

Codigo-postal.info. (2023). Códigos postales de Madrid en Comunidad de Madrid. Obtenido de https://www.codigo-postal.info/madrid (Fecha de consulta: 10/12/2022)

Garre, M., Cuadrado, J. J., Sicilia, M. A., Rodríguez, D., & Rejas, R. (2007). Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software. REICIS. *Revista Española de Innovación, Calidad e Ingeniería del Software*, 3(1), 6-22.

González, E. I. (2006). *Introducción al Analisis de Regresón Lineal Tercera Edicion Montgomery Peck Vining.* Cecsa. México.

Heo, C. Y. (2016). Sharing economy and prospects in tourism research. *Annals of Tourism Research*, 58, 166-170

Inflation.eu. (2023). Inflacion de España. Obtenido de https://www.inflation.eu/es/tasas-de-inflacion/espana/inflacion-historica/ipc-inflacion-espana-2023.aspx (Fecha de consulta: 15/01/2023)

Instituto Nacional de Estadística (2023). *Encuesta de ocupación en apartamentos turísticos*. Obtenido de https://ine.es/jaxiT3/Datos.htm?t=3152#!tabs-tabla (Fecha de consulta: 28/04/2023)

Izquierdo, M., Ramón-Rodríguez, L., Devesa, J. (2016). *Economía del Turismo, Recursos Naturales y Nuevas Tecnologías (INNATUR)* 150: 107-119

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* Springer. USA

López, J. F. (2020). *Rango intercuartílico*. Obtenido de https://economipedia.com/definiciones/rango-intercuartilico.html (Fecha de consulta: 28/04/2023)

López, P.,Santín, C., González, D. (2007). *Minería de datos. Técnicas y herramientas: técnicas y herramientas*. Thompson. Madrid.

Manovich, L. (2010). What is Visualization? "*Creative Commons Attribution-NonCommercial-ShareAlikelicense (CC BY-NC-SA)*"

Martínez del Olmo, A. (2018). La explosión del alquiler y las desigualdades residenciales en Madrid. *ANDULI. Revista Andaluza De Ciencias Sociales*, (17), 109–132.

Microsoft. (2022). Componente: Agrupación en clústeres K-means. Obtenido de https://learn.microsoft.com/es-es/azure/machine-learning/component-reference/k-means-clustering (Fecha de consulta: 20/02/2023)

Minguillón Alfonso, Julià. (2016). "Introducción a la visualización de datos". Universitat Oberta de Catalunya

M. Weiss, S., Indurkhya, N. (1998). *Predictive Data Mining: A Practical Guide.* Morgan Kaufmann Publishers. San Francisco, CA.

Raffini, L. (2016). L'economia della condivisione tra retoriche, ambiguità e lati oscuri. Riflessioni a partire dal caso Airbnb. *La Rivista delle Politiche Sociali / Italian Journal of Social Policy*, 1/2016, 129-149

Statista (2023). Airbnb - estadísticas y hechos. Obtenido de https://www.statista.com/topics/2273/airbnb/#topicOverview (Fecha de consulta: 01/02/2023)

---------- (2023). El sector hotelero en España - Datos estadísticos. Obtenido de https://es.statista.com/temas/3875/sector-hotelero-en-espana/#topicOverview (Fecha de consulta: 01/02/2023).

---------- (2021). Principales datos de Airbnb a nivel mundial en 2020. Obtenido de https://es.statista.com/estadisticas/1218479/principales-indicadores-de-actividad-de-airbnb-en-el-mundo/ (Fecha de consulta: 01/02/2023)