**Madridrbnb - Tabla de contenidos**

[1. Descripción del dataset 2](#_Toc91513440)

[listings.csv 2](#_Toc91513441)

[reviews\_detailed.csv 3](#_Toc91513442)

[C5000121.xls 3](#_Toc91513443)

[Importancia y objetivos del análisis 4](#_Toc91513444)

[2. Integración y selección de los datos de interés a analizar. 5](#_Toc91513445)

[3. Limpieza de los datos 6](#_Toc91513446)

[Ceros y elementos vacíos 6](#_Toc91513447)

[listings.csv 6](#_Toc91513448)

[reviews\_detailed.csv 7](#_Toc91513449)

[Valores extremos 7](#_Toc91513450)

[listing.csv 7](#_Toc91513451)

[Otros tratamientos 9](#_Toc91513452)

[4. Análisis de los datos y representación de los resultados a partir de tablas y gráficas 10](#_Toc91513453)

[Selección de los grupos de datos a analizar 10](#_Toc91513454)

[Exportación de los datos preprocesados 11](#_Toc91513455)

[Comprobación de la normalidad y homogeneidad de la varianza 11](#_Toc91513456)

[Pruebas estadísticas 11](#_Toc91513457)

[¿Qué distritos tienen más alojamientos? 11](#_Toc91513458)

[¿Qué tipo de alojamiento es el más frecuente? 12](#_Toc91513459)

[¿Cuáles son las palabras más utilizadas en el título de los alojamientos? 12](#_Toc91513460)

[¿Qué tipo de alojamiento es más frecuente por distrito? 12](#_Toc91513461)

[¿Qué distrito es el más caro? ¿Cuál es el más barato? 13](#_Toc91513462)

[¿Cuál es el precio medio de cada tipo de alojamiento? 14](#_Toc91513463)

[¿Cuántos alojamientos por tipo de habitación y precio hay? 14](#_Toc91513464)

[¿Qué distritos tienen una mayor densidad de alojamientos por habitante? 15](#_Toc91513465)

[¿Existe una diferencia significativa entre los tipos de habitación por distrito? 15](#_Toc91513466)

[¿Cómo es la estacionalidad en el alquiler de alojamientos turísticos? 16](#_Toc91513467)

[¿Se podría construir un modelo de regresión para predecir el precio del alojamiento en función de otras variables? 17](#_Toc91513468)

[5. Conclusiones 19](#_Toc91513469)

[6. Código 20](#_Toc91513470)

[7. Contribución 20](#_Toc91513471)

# 1. Descripción del dataset

Los datos utilizados en esta práctica proceden de varios datasets:

|  |  |
| --- | --- |
| Dataset | Descripción |
| [listings.csv](https://www.kaggle.com/rusiano/madrid-airbnb-data?select=listings.csv) | Contiene información sobre los anuncios de alojamientos turísticos de Airbnb en Madrid |
| [reviews\_detailed.csv](https://www.kaggle.com/rusiano/madrid-airbnb-data?select=reviews_detailed.csv) | Contiene todas las reseñas realizadas para todos los alojamientos turísticos de Airbnb en Madrid. |
| [C5000121.xls](https://www.madrid.es/UnidadesDescentralizadas/UDCEstadistica/Nuevaweb/Demograf%C3%ADa%20y%20población/Indicadores%20Demográficos/C5000121.xls) | Contiene datos de población por distrito en la ciudad de Madrid. |

A continuación, se describe cada uno ellos.

## listings.csv

Este dataset contiene todos los datos críticos para esta práctica. Se ha obtenido de Kaggle y está formado por 19.618 registros con la información sobre los anuncios de alojamientos y 16 atributos. La descripción del dataset proporcionada por el dueño de los datos se puede consultar [aquí](https://docs.google.com/spreadsheets/d/1iWCNJcSutYqpULSQHlNyGInUvHg2BoUGoNRIGa6Szc4/edit#gid=982310896).

|  |  |  |
| --- | --- | --- |
| Nombre atributo | Tipo Dato | Descripción |
| id | Entero | Identificador único para el alojamiento de Airbnb |
| name | Cadena de caracteres | Título del alojamiento |
| host\_id | Entero | Identificador único del dueño del alojamiento |
| host\_name | Cadena de caracteres | Nombre del dueño del alojamiento |
| neighbourhood\_group | Cadena de caracteres | Distrito |
| neighbourhood | Cadena de caracteres | Barrio |
| latitude | Numérico | Latitud en WGS84 |
| longitude | Numérico | Longitud en WGS84 |
| room\_type | Cadena de caracteres | Tipo de alojamiento |
| price | Entero | Precio en euros |
| minimum\_nights | Entero | Número mínimo de noches en el alojamiento |
| number\_of\_reviews | Entero | Número de reseñas del alojamiento |
| last\_review | Fecha | Fecha de la última reseña |
| reviews\_per\_month | Numérico | Media de reseñas al mes |
| calculated\_host\_listings\_count | Entero | Número de alojamientos de que dispone el dueño |
| availability\_365 | Entero | Disponibilidad del alojamiento |

## reviews\_detailed.csv

Este dataset contiene todas las reseñas realizadas sobre los alojamientos de Airbnb. Tiene 625.006 registros y 6 atributos.

|  |  |  |
| --- | --- | --- |
| Nombre atributo | Tipo Dato | Descripción |
| listing\_id | Entero | Identificador único para el alojamiento de Airbnb |
| id | Entero | Identificador de la reseña |
| date | Fecha | Fecha de la reseña |
| reviewer\_name | Cadena de caracteres | Nombre de quién hizo la reseña |
| comments | Cadena de caracteres | Reseña |

## C5000121.xls

En la página web del Ayuntamiento de Madrid podemos descargarnos este fichero con indicadores demográficos por cada distrito. Este conjunto de datos contiene 21 distritos y 16 atributos:

|  |  |  |
| --- | --- | --- |
| Nombre atributo | Tipo Dato | Descripción |
| Distrito | Cadena de caracteres | Nombre del distrito |
| Población a 1/1/2021 | Entero | Población |
| 0 a 15 años | Numérico | % de población en esa franja de edad |
| 16 a 64 años | Numérico | % de población en esa franja de edad |
| 65 años y más | Numérico | % de población en esa franja de edad |
| 80 años y más | Numérico | % de población en esa franja de edad |
| Índice de juventud | Numérico | Población de 0 a 15 años / Población de 65 años y mas |
| Índice de reemplazo de la población activa | Numérico | Población de 16 a 19 años / Población de 60 a 64 años |
| Razón de progresividad | Numérico | Población de 0 a 4 años / Población de 5 a 9 años |
| Edad promedio | Numérico | Edad media |
| Natalidad | Numérico | Tasa de natalidad (por mil habitantes) |
| Mortalidad | Numérico | Tasa de mortalidad (por mil habitantes |
| Crecimiento vegetativo | Numérico | Natalidad – Mortalidad (por mil habitantes) |
| Inmigración | Numérico | Población inmigrante (por mil habitantes) |
| Emigración | Numérico | Población emigrante (por mil habitantes) |
| Migración neta | Numérico | Inmigración – Emigración (por mil habitantes) |

Para la realización de esta práctica no se requieren todos los datos que están presentes en estos datasets. Se explica con más detalle en el punto 2.

## Importancia y objetivos del análisis

Madrid se ha convertido en uno de los destinos predilectos de la inversión inmobiliaria mexicana. El idioma, los lazos culturales entre ambos países, la oferta gastronómica, así como la seguridad jurídica y ciudadana, entre otros atractivos, han provocado el interés del capital inversor mexicano en la capital de España.

Una conocida firma de inversión inmobiliaria mexicana ha pedido realizar un estudio sobre la situación del mercado inmobiliario en la ciudad de Madrid. El problema por resolver es sencillo: ¿Dónde invertir?

El inversor quiere respuesta a estas preguntas:

* ¿Qué tipo de vivienda comprar?
* ¿En qué zona de la capital?
* ¿Qué destino se le va a dar al inmueble? Entre las opciones de alquiler turístico y tradicional.

Es un hecho conocido que el mercado de alquiler se encuentra en franco retroceso debido al auge de los alquileres turísticos. Los propietarios han cambiado el alquiler tradicional por el turístico, espoleado por compañías de impacto mundial como Airbnb. El efecto sobre el alquiler tradicional ha sido nefasto. Por un lado, la oferta de alquileres se ha reducido, lo que ha tenido un gran impacto en los precios, y por otro lado ha generado un flujo de habitantes hacia las zonas periféricas de la capital. El centro de la ciudad es de los turistas.

Debido a esto, y para dar respuesta a nuestro inversor mexicano, se utilizan datos de Airbnb.

El mejor enfoque para poder proporcionar una respuesta adecuada a un problema complejo es dividir el problema en partes más pequeñas, más sencillas de responder, es por ello por lo que se definen una serie de preguntas más específicas. Estas preguntas se hacen desde dos ángulos distintos: por un lado, queremos analizar el aspecto geográfico, ¿dónde invertir?, y por otro lado lo concerniente al tipo de alojamiento a comprar ¿Qué comprar y para qué?

Un poco de geografía: La ciudad de Madrid se divide en 21 distritos. La presencia de la carretera de circunvalación M-30 actúa como barrera geográfica entre los distritos de la almendra central (interior) de los del extrarradio.

|  |  |  |
| --- | --- | --- |
| Distrito | Localización | Precio alquiler (eur/m2) (\*) |
| Centro | Interior | 17,5 |
| Tetuán | Interior | 15,0 |
| Chamartín | Interior | 15,4 |
| Chamberí | Interior | 17,1 |
| Salamanca | Interior | 17,5 |
| Retiro | Interior | 15,0 |
| Arganzuela | Interior | 14,6 |
| Barajas | Interior | 11,4 |
| Carabanchel | Interior | 11,7 |
| Ciudad Lineal | Interior | 12,8 |
| Fuencarral – El Pardo | Extrarradio | 15,4 |
| Hortaleza | Extrarradio | 12,5 |
| Latina | Extrarradio | 11,8 |
| Moncloa - Aravaca | Extrarradio | 14,3 |
| Moratalaz | Extrarradio | 10,9 |
| Puente de Vallecas | Extrarradio | 12,1 |
| San Blas | Extrarradio | 11,4 |
| Usera | Extrarradio | 11,5 |
| Vicálvaro | Extrarradio | 10,4 |
| Villa de Vallecas | Extrarradio | 11,2 |
| Villaverde | Extrarradio | 10,9 |

(\*) Según el portal [idealista](https://www.idealista.com/sala-de-prensa/informes-precio-vivienda/alquiler/madrid-comunidad/madrid-provincia/madrid/)

Buscamos respuesta a estas preguntas relacionadas con la distribución geográfica de los alojamientos en Madrid:

* ¿Qué distritos tienen más alojamientos?
* ¿Qué tipo de alojamiento es el más frecuente por distrito?
* ¿Existen diferencias significativas de precio para los diferentes distritos?
* ¿Cuál es la densidad de alojamientos por distrito? ¿Qué distritos tienen una mayor densidad de alojamientos por habitante?
* ¿Existe una diferencia significativa entre los tipos de alojamiento por distrito?

Airbnb oferta 4 tipos de alojamientos: alojamiento completo, habitación privada, habitación compartida y habitación de hotel. Las preguntas por responder en este grupo son:

* ¿Qué tipo de alojamiento es el más frecuente?
* ¿Cuál es el precio medio de cada tipo de alojamiento?
* ¿Cuáles son las palabras más utilizadas en los títulos de los anuncios de alojamientos?
* ¿Cuantos alojamientos por tipo de habitación y precio hay? ¿Existen diferencias significativas de precio para los diferentes tipos de habitación?
* ¿Existe una diferencia significativa entre el promedio de reseñas por mes para los alojamientos de tipo habitación privada y apartamento?
* ¿Cómo es la estacionalidad en el alquiler de alojamientos turísticos?
* ¿Se podría construir un modelo para predecir el precio del alojamiento en función de otras variables?

Todas estas preguntas se responden utilizando técnicas de ciencia de datos.

# 2. Integración y selección de los datos de interés a analizar.

Del dataset listings nos quedamos con los siguientes datos, los eliminados no son relevantes para resolver nuestro problema.

|  |  |  |
| --- | --- | --- |
| Nombre atributo | Tipo Dato | Descripción |
| id | Entero | Identificador único para el alojamiento de Airbnb |
| name | Cadena de caracteres | Título del alojamiento |
| neighbourhood\_group | Cadena de caracteres | Distrito |
| latitude | Numérico | Latitud en WGS84 |
| longitude | Numérico | Longitud en WGS84 |
| room\_type | Cadena de caracteres | Tipo de alojamiento |
| price | Entero | Precio en euros |
| minimum\_nights | Entero | Número mínimo de noches en el alojamiento |
| reviews\_per\_month | Numérico | Media de reseñas al mes |
| availability\_365 | Entero | Disponibilidad del alojamiento |

Del dataset reviews\_detailed sólo necesitamos el id y la fecha en la que se han hecho las reseñas para poder analizar cuándo se hacen reseñas.

|  |  |  |
| --- | --- | --- |
| Nombre atributo | Tipo Dato | Descripción |
| listing\_id | Entero | Identificador único para el alojamiento de Airbnb |
| date | Fecha | Fecha de la reseña |

Por último, del dataset de población, sólo nos quedaremos con el distrito y la población, el resto de los atributos no son relevantes

|  |  |  |
| --- | --- | --- |
| Nombre atributo | Tipo Dato | Descripción |
| Distrito | Cadena de caracteres | Nombre del distrito |
| Población a 1/1/2021 | Entero | Población |

Los datasets no se integran en único conjunto de datos, ya que responden por separado a las distintas preguntas y su integración supondría la repetición de datos (por ejemplo, en el caso de integrar los datos de población con los de alojamientos, estaríamos repitiendo los datos de población en cada registro, por lo que mantenemos el modelo relacional).

# 3. Limpieza de los datos

En esta fase se limpian los datasets de valores nulos, ceros, y valores extremos.

## Ceros y elementos vacíos

### listings.csv

**reviews\_per\_month** tiene valores nulos. Se acepta que haya alojamientos en los que los turistas no hayan hecha reseñas y se sustituyen los valores nulos por ceros.

En los campos **name**, **host\_name** y **last\_review** también hay valores nulos. El campo **name** sólo tiene 3 registros con valores nulos, pero este hecho no afecta al estudio. Los campos **host\_name** y **last\_review** no se integran en el dataset final por lo que no los tratamos.

Se detectan valores cero en los campos **number\_of\_reviews**, **reviews\_per\_month**, **availability\_365** y **price**. Se consideran valores válidos para los campos **number\_of\_reviews**, **reviews\_per\_month** y **availability\_365**, ya que puede haber alojamientos sin reseñas, y también alojamientos listados en Airbnb pero que están bloqueados por el anfitrión. Sin embargo, **price** no puede tener valor cero, por lo que se eliminan los registros que cumplen esta condición.

### reviews\_detailed.csv

**comments**: hay 4 reseñas con valores nulos y 329 con valores en blanco. Se eliminan al no aportar valor.

**reviewer\_name**: hay 1 registro con un valor en blanco, pero no se trata al no utilizar este campo en el análisis.

## Valores extremos

### listing.csv

En el dataset listing se aprecian valores extremos en los siguientes campos:

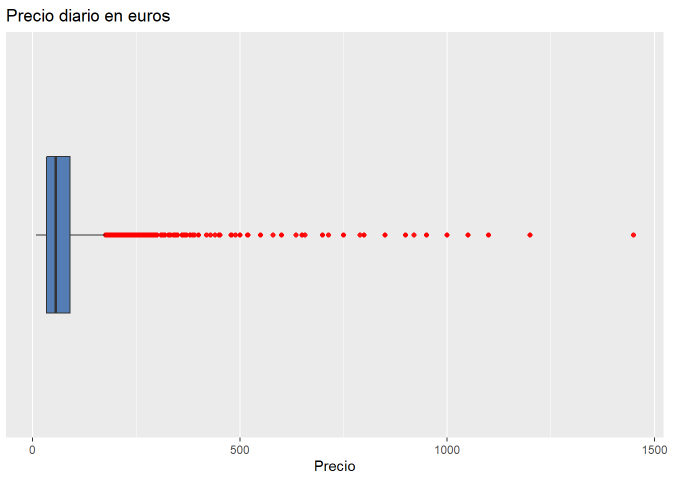
**price**: el precio máximo es 9999, lo que podría indicar que la ausencia de valores se ha codificado con este número, o un error en la captura de los datos. No resulta obvio detectar valores extremos en el precio de los alojamientos, por dos motivos:

* La localización va a limitar el valor máximo del precio del alojamiento.
* Airbnb oferta inmuebles de lujo. Se trata de alojamientos exclusivos. De ahí que exista una gran variabilidad en los precios.

Chart

Description automatically generated

Para tratar los valores extremos se ha determinado un límite de precio máximo en el percentil 95 de los precios de cada distrito. Los registros por encima de ese valor se eliminan. De esta forma se pretende resolver los dos problemas que plantea el precio: la componente geográfica y el lujo.



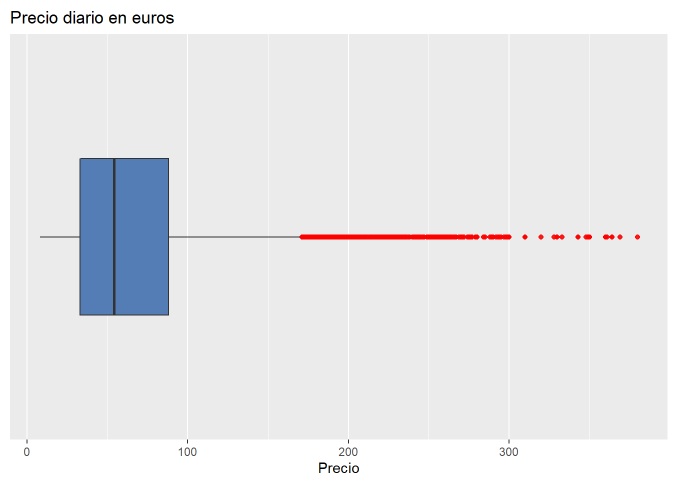
Sin embargo, al representar los diagramas de cajas de precio por distrito se observan valores extremos en distritos que generalmente tienen un precio medio de alquiler más barato: San Blas-Canillejas, Vicálvaro, Moratalaz, Fuencarral-El Pardo, Ciudad Lineal y Barajas.

A picture containing graphical user interface

Description automatically generated

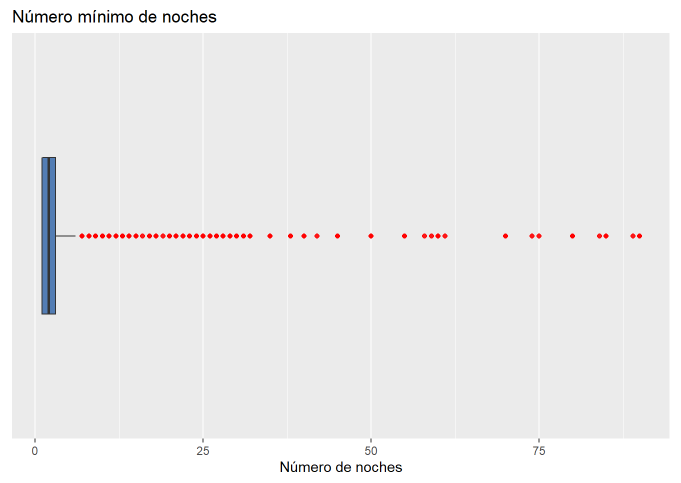
El motivo de esta variabilidad se identifica en la final de la Champions League que se jugó en 2019 en el estadio Wanda-Metropolitano que provocó un incremento exponencial de los precios de los alojamientos en los alrededores. La variable **price** almacena el mayor precio al que se alquila el alojamiento, y de ahí la necesidad de volver a corregir estos valores extremos.

La corrección se hace limitando el precio máximo a 150 euros la noche en estos distritos. La nueva distribución de precios tras la corrección es la siguiente:



**minimum\_nights**: esta variable toma un valor máximo excesivamente alto (1125 días), lo que podría estar indicando una captura errónea del dato. El Plan Especial de Hospedaje (PEH), que entró en vigor tras su aprobación definitiva en el Pleno municipal del 27 de marzo de 2019, limita a 90 días la posibilidad de alquilar una vivienda con fines turísticos sin permiso y a partir de ese plazo obliga a obtener una licencia de uso terciario de hospedaje. Esta medida afecta al 95% de los pisos de uso turístico.

Se decide no eliminar registros, en este caso de valores superiores a 90, sino limitar estos registros a un máximo de 90 noches al año.



## Otros tratamientos

En el dataframe **df\_population,** donde se cargan los datos de población por distrito, hay que normalizar los nombres de los distritos para poder unir esos datos a los distritos del dataframe **df\_listings**.

# 4. Análisis de los datos y representación de los resultados a partir de tablas y gráficas

## Selección de los grupos de datos a analizar

En esta fase se seleccionan los diferentes análisis a realizar, así como los datos que van a participar en dicho análisis.

|  |  |  |  |
| --- | --- | --- | --- |
| Pregunta | Tipo de Análisis | Variables utilizadas | Link |
| ¿Qué distritos tienen más alojamientos? | Univariante | listing.csv: neighbourhood\_group | [1](#_¿Qué_distritos_tienen) |
| ¿Qué tipo de alojamiento es el más frecuente? | Univariante | listing.csv: room\_type | [2](#_¿Qué_tipo_de) |
| ¿Cuáles son las palabras más utilizadas en el título de los anuncios de alojamientos? | Univariante | listing.csv: name | [3](#_¿Cuáles_son_las) |
| ¿Qué tipo de alojamiento es el más frecuente por distrito? | Bivariante | listing.csv: neighbourhood\_group, room\_type | [4](#_¿Qué_tipo_de_1) |
| ¿Existen diferencias significativas de precio en los distintos distritos? | Bivariante: Test de normalidad. Test de homocedasticidad. Test de hipótesis | listing.csv: id, neighbourhood\_group, price, price\_group (categorización de price) | [5](#_¿Qué_distrito_es) |
| ¿Cuál es el precio medio para cada tipo de alojamiento? | Bivariante | listing.csv: price, room\_type | [6](#_¿Cuál_es_el) |
| ¿Cuántos alojamientos por tipo de habitación y precio hay? ¿Existen diferencias significativas de precio para los distintos tipos de alojamientos? | Bivariante: Test de normalidad. Test de homocedasticidad. Test de hipótesis. | listing.csv: room\_type, price, price\_group (categorización de price) | [7](#_¿Cuántos_alojamientos_por) |
| ¿Qué distritos tienen una mayor densidad de alojamientos por habitante? | Multivariante | C5000121.xls:neighbourhood\_group, population  listing.csv: neighbourhood\_group, longitude,latitude | [9](#_¿Qué_distritos_tienen_1) |
| ¿Existe una diferencia significativa entre los tipos de alojamiento por distrito? | Bivariante: Test chi cuadrado. | listing.csv: neighbourhood\_group, room\_type | [10](#_¿Existe_una_diferencia) |
| ¿Cómo es la estacionalidad en el alquiler de alojamientos turísticos? | Univariante | reviews\_detailed.csv: date | [11](#_¿Cómo_es_la) |
| ¿Se podría construir un modelo de regresión para predecir el precio del alojamiento en función de otras variables? | Multivariante | listing.csv: price, minimum\_nights, reviews\_per\_month, availability\_365, latitude, longitude, room\_type, neighbourhood\_group | [12](#_¿Se_podría_construir) |

## Exportación de los datos preprocesados

Una vez realizada la limpieza y determinados los datos que se necesitan para el análisis, se guardan.



## Comprobación de la normalidad y homogeneidad de la varianza

Comprobamos la normalidad y homocedasticidad de las variables necesariaspara contestar a las siguientes preguntas:

* ¿Existen diferencias significativas de precio para los diferentes distritos?
* ¿Existen diferencias significativas de precio para los diferentes tipos de habitación?

Se comprueba la normalidad de la variable **price** con el test de **Kolmogorov-Smirnov**. Como la distribución no es normal, se comprueba la homocedasticidad de la variable con el test de **Fligner-Killeen**. La prueba indica que no hay homocedasticidad, por lo que usamos el test de **Kruskal-Wallis** para determinar si hay diferencias significativas de precio para los diferentes distritos y para los diferentes tipos de habitación. Este test indica que hay diferencias significativas en ambos casos.

* ¿Existen diferencias significativas entre el promedio de reseñas por mes por tipo de alojamiento?

Se comprueba la normalidad de la variable **reviews\_per\_month** con el test de **Kolmogorov-Smirnov**. Como la distribución no es normal, se comprueba la homocedasticidad de la variable con el test de **Fligner-Killeen**. El test indica que no hay homocedasticidad, por lo que usamos el test de **Kruskal-Wallis** para determinar si hay diferencias significativas entre el promedio de reseñas por mes por tipo de alojamiento. El test indica que hay diferencias significativas.

## Pruebas estadísticas

### ¿Qué distritos tienen más alojamientos?

El distrito Centro es, con mucha diferencia, el que más alojamientos turísticos tiene anunciados en Airbnb. Los siguientes distritos son limítrofes al centro.

A picture containing table

Description automatically generated

### ¿Qué tipo de alojamiento es el más frecuente?

El tipo de alojamiento más anunciado es “Entire home/Apt” seguido de “Private room”. Los usuarios de Airbnb dan mucha importancia a su privacidad, de ahí los bajos valores de las habitaciones compartidas. El resultado de las habitaciones de hotel es testimonial, ya que este tipo de alojamiento se ofertan en otras plataformas.

Chart, scatter chart

Description automatically generated

### ¿Cuáles son las palabras más utilizadas en el título de los alojamientos?

Las palabras más utilizadas en los títulos de los anuncios de alojamientos de Airbnb en Madrid son habitación, apartamento y centro, con sus variantes en inglés.



### ¿Qué tipo de alojamiento es más frecuente por distrito?

Con la excepción de Villaverde, Villa de Vallecas, Vicálvaro, San Blas-Canillejas, Moratalaz, Barajas y Ciudad Lineal, en el resto de Madrid predomina el alquiler del apartamento completo en vez de habitaciones privadas o compartidas.

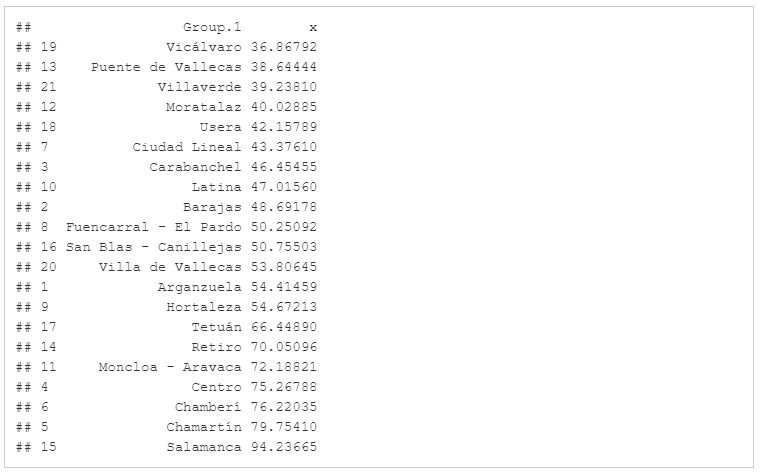
Chart, bar chart

Description automatically generated

### ¿Qué distrito es el más caro? ¿Cuál es el más barato?

Los distritos más caros son los situados dentro de la almendra central, con la excepción de Moncloa-Aravaca que es una de las zonas de lujo de la capital.

Los distritos más baratos son los situados en la parte exterior de la M-30.



Tras discretizar el precio en 5 categorías: Muy Alto, Alto, Moderado, Bajo, Muy Bajo, la distribución de alojamientos por distrito y categoría de precio se representa en el siguiente gráfico:

Chart, bar chart

Description automatically generated

En los distritos del extrarradio predominan los precios bajos-muy bajos, mientras que en la almendra central predominan los precios altos-muy altos. En cualquier caso, se encuentran alojamientos de todos los rangos de precios en cualquiera de los distritos.

### ¿Cuál es el precio medio de cada tipo de alojamiento?

Un alquiler de un apartamento completo tiene un precio similar al de una habitación de hotel, y es un 216% más caro que una habitación privada y un 228% más caro que una habitación compartida.

Chart, line chart, histogram

Description automatically generated

### ¿Cuántos alojamientos por tipo de habitación y precio hay?

Se aprecia que los apartamentos completos tienen un rango de precios entre alto y muy alto. Las habitaciones privadas suelen tener un precio bajo – muy bajo, siendo las habitaciones compartidas de precio muy bajo.

Chart, bar chart

Description automatically generated

### ¿Qué distritos tienen una mayor densidad de alojamientos por habitante?

Se crea una nueva variable que almacena el número de habitantes por alojamiento turístico para cada distrito. En el distrito Centro hay un alojamiento cada 17 habitantes. Los 4 siguientes distritos con mayor número de habitantes por alojamiento están en la almendra central. En Vicálvaro hay un alojamiento por cada 1424 habitantes. Todos los distritos con mayor número de habitantes por alojamiento turístico están en el extrarradio.

Chart, scatter chart

Description automatically generated

### ¿Existe una diferencia significativa entre los tipos de habitación por distrito?

Para comparar si existen diferencias significativas en una variable categórica entre los grupos definidos por otra categoría se aplica un test **chi-cuadrado** sobre la tabla de contingencia de ambas variables.

Table

Description automatically generated

El resultado del test es que hay diferencias significativas entre el tipo de habitación y el distrito en el que se encuentra el alojamiento (p-value < 2.2e-16).

### ¿Cómo es la estacionalidad en el alquiler de alojamientos turísticos?

Se aprecia como el negocio de los alojamientos turísticos crecía de forma exponencial desde el año 2015. La brusca caída se debe a los efectos del COVID-19 en el año 2020, situación que continua en 2021. Los últimos datos del dataset son de Abril de 2021, por lo que no se puede apreciar en el gráfico si hay recuperación con respecto a 2020 o no.

Chart, histogram

Description automatically generated

Las reseñas se mantienen constantes a lo largo del año. Sorprende que en los meses de verano sea cuando menos reseñas se escriban.

Chart, bar chart

Description automatically generated

Las reseñas permanecen constantes a lo largo de la semana, y se ve un fuerte incremento en los domingos y lunes, lo cual evidencia que es en los fines de semana cuándo más se reservan estos alojamientos.

Chart, histogram

Description automatically generated

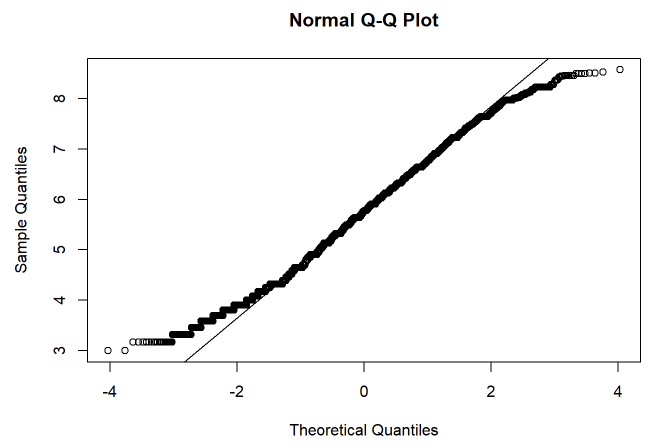
### ¿Se podría construir un modelo de regresión para predecir el precio del alojamiento en función de otras variables?

Se prueba con dos modelos: un modelo de regresión lineal múltiple y un modelo de regresión logística.

#### Modelo de regresión lineal múltiple

Probamos un modelo que nos permita predecir el precio de alquiler (variable dependiente cuantitativa) en función de las siguientes variables independientes: número de noches mínimas de alquiler, número de reseñas por mes, disponibilidad, longitud, latitud, distrito y tipo de habitación.

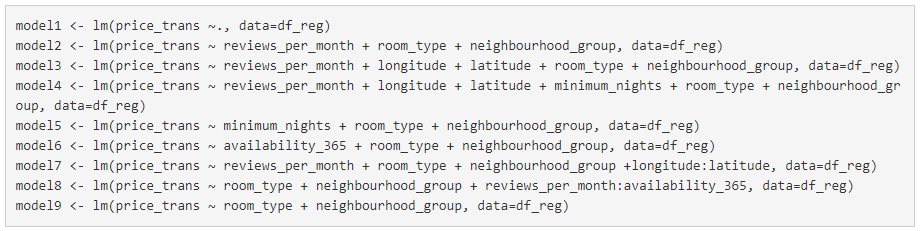
Anteriormente se comprobó que la variable **price** no seguía una distribución normal, afectada sobre todo, por valores extremos debido al lujo, por lo que en este modelo se utiliza una transformación logarítmica para acercar más su distribución a la normalidad.



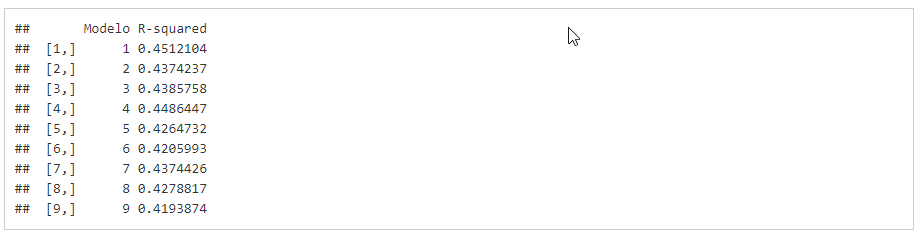
Posteriormente se hace un estudio de correlación para ver si existe correlación entre la variable price transformada (**price\_trans**) y el resto de variables obteniéndose una correlación muy baja como se muestra a continuación.

|  |  |
| --- | --- |
|  |  |

Se crean distintos modelos usando diferentes combinaciones de las variables cualitativas y cuantitativas.



Se comprueba el ajuste con la función **R-Squared**, se determina que el mejor modelo es el **modelo 1**, que **solo es capaz de explicar el 45% de la variabilidad observada en los precios**.



#### Modelo de regresión logística

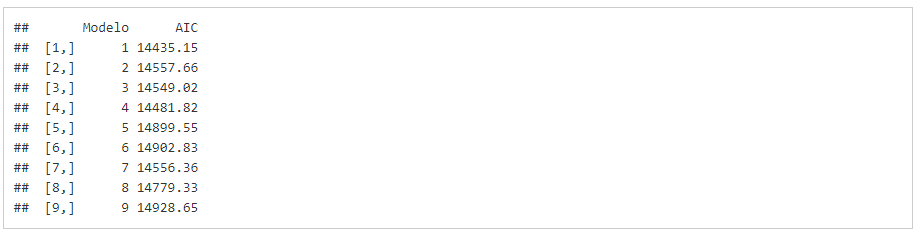
Probamos con un modelo de regresión logística que nos permita obtener la probabilidad de que el alquiler sea mayor de 100 euros la noche. Para ello discretizamos la variable precio en dos categorías.



Se usan las mismas variables independientes del modelo anterior y se crean los diferentes modelos.



Se mide la bondad del ajuste de los modelos con el **Criterio de Información de** **Akaike (AIC)**. El menor de los valores nos indicará cual es el mejor modelo. En este caso el **modelo 1**.



Se analiza la precisión del modelo, comparando la predicción del modelo contra un conjunto de prueba (train-test). La matriz de confusión arroja los siguientes resultados:

* **Exactitud** (Accuracy): 82.2%. Es la habilidad del modelo para detectar correctamente los alquileres por encima y debajo de 100 euros.
* **Sensibilidad** (Sensitivity): 82.5%. Es la habilidad de detectar verdaderos positivos en el total de positivos que hay en la realidad. Consideramos como positivo el que el alquiler sea mayor de 100 euros la noche.
* **Especificidad** (Specificity): 58.5%. Es la habilidad del modelo de detectar verdaderos negativos en el total de negativos reales. El modelo es solo capaz de predecir el 58.5% de los casos reales donde el precio es mayor de 100 euros.

# 5. Conclusiones

Atendiendo a las respuestas obtenidas en todas y cada una de las preguntas que se han realizado se llegan a las siguientes conclusiones:

**¿Qué tipo de vivienda comprar?**

Viviendas de segunda mano, de tamaño mediano-pequeño, a reformar.

El hecho de comprar viviendas de segundo mano se explica por la ausencia de promociones de obra nueva en las zonas de interés. Se explica en el siguiente punto cuáles son estas zonas.

Viviendas de tamaño mediano-pequeño, ya que la vivienda preferida por los usuarios de Airbnb son los apartamentos enteros. El precio se incrementa hasta un 250% con respecto a alquilar una habitación compartida. Para compensar esto, se requerirían varias habitaciones privadas en el mismo inmueble, pero implicaría inmuebles más caros de comprar y de reformar.

En el análisis de palabras más utilizadas en los anuncios se observan muchos términos relacionados con la localización y tipo de inmueble, pero también hay adjetivos como: stylish, acogedor, cozy, luxury, lovely, spacious, que indican una necesidad de transformar los inmuebles de acuerdo a las tendencias de decoración minimalista actuales.

**¿En qué zona de la capital?**

Viviendas en la zona de la almendra central pero fuera del Centro debido a la saturación de esta zona (17 habitantes por alojamiento turístico).

Se recomiendan oportunidades en barrios más caros como Chamberí, Salamanca, y Retiro, y centrar los esfuerzos en adquirir inmuebles en zonas de la almendra central con menor coste: Arganzuela, Tetuán, Chamartín.

Otra opción muy interesante por precio y su proximidad al aeropuerto sería Barajas.

Las oportunidades serían ventas de inmuebles ya destinados al alquiler turístico en el que la inversión por reforma no sería necesaria. Sería posible encontrar estas oportunidades debido a los efectos de la pandemia.

**¿Qué destino se le va a dar al inmueble?**

La inversión debe destinarse al alquiler turístico. Es un mercado que hasta la pandemia de COVID-19 crecía de forma exponencial y al que se ven signos de recuperación.

Según las [estadísticas del Ayuntamiento de Madrid](http://www-2.munimadrid.es/CSE6/control/seleccionDatos?numSerie=05040201030) la superficie media de la vivienda en Madrid es de 82 m2. Según el portal [idealista](https://www.idealista.com/sala-de-prensa/informes-precio-vivienda/alquiler/madrid-comunidad/madrid-provincia/madrid/) el precio medio del alquiler en Madrid en Noviembre de 2021 es de 14,6 euros/metro cuadrado. Por lo tanto, el precio medio del alquiler tradicional es de 1197 euros al mes.

En este ejercicio se ha comprobado que el precio medio del alquiler turístico es de 88 euros/noche, por lo que llegaría a los 2640 euros al mes (considerando 30 días). Un 220% más.

El análisis de estacionalidad del alquiler turístico también nos muestra homogeneidad a lo largo de los meses y una alta ocupación durante todos los días de la semana.

# 6. Código

El código se ha generado en R. Está documentado en este repositorio [github.](https://github.com/gromerof1974/Madridrbnb)

# 7. Contribución

|  |  |
| --- | --- |
| Contribuciones | Firma |
| Investigación previa | BLB, GRF |
| Redacción de las respuestas | BLB, GRF |
| Desarrollo del código | BLB, GRF |