**MARCO Y JUSTIFICACIÓN DE LA RELEVANCIA Y NOVEDAD DEL TRABAJO REALIZADO**

<https://www.malaga.es/noticias/com1_md-3/com1_ct-0/com1_fb-0/com1_cb-0/com1_md3_cd-50341/el-sector-turistico-de-la-costa-del-sol-evidencia-el-peso-del-turismo-sobre-otros-sectores-y-lanza-un-potente-mensaje-contra-la-turismofobia>

<https://cadenaser.com/andalucia/2024/05/08/el-empleo-del-sector-turistico-crece-casi-un-8-en-malaga-con-125000-ocupados-en-la-hosteleria-alojamientos-y-agencias-ser-malaga/>

https://www.malaga.es/noticias/com1\_prt-1/com1\_md3\_cd-50000/la-costa-del-sol-continua-despuntando-en-el-primer-trimestre-de-2024-con-cerca-de-millon-y-medio-de-viajeros#:~:text=As%C3%AD%2C%20de%20este%20%C3%BAltimo%20informe,en%20la%20cifra%20de%205.423.

El sector turístico es uno de los pilares más fuertes de la provincia de Málaga, y de su ciudad.

Durante el primer cuatrimestre de 2024, la costa del sol registró un incremento del 8,1% en la estancia de viajeros en establecimientos reglados, y un 13,5% en alojamientos turísticos durante el primer cuatrimestre de 2024. Es sobre ese 13,5% donde va a centrarse este análisis.

La inmensa mayoría de estos alojamientos turísticos, cuyo número en la ciudad de Málaga no para de ir en crecimiento, se encuentran anunciados bajo la plataforma AirBNB.

Este crecimiento no está exento de controversia. El pasado mes de Julio se dieron grandes manifestaciones en contra del turismo masivo, culpándose durante las mismas a AirBNB del aumento de los precios del alquiler. <https://es.euronews.com/viajes/2024/07/07/malaga-para-vivir-no-para-sobrevivir-los-malaguenos-protestan-contra-el-turismo-en-medio-d>

De lo que no cabe duda es que los alojamientos turísticos no hosteleros es un tema que está al orden del día en Málaga.

**Definición del objetivo planteado**

El objetivo de este trabajo es el análisis del mercado de alquiler de alojamientos turísticos no hosteleros de la ciudad de Málaga que se anuncian en la plataforma AirBNB, así como el hallazgo de que características de las presentes en los anuncios de la plataforma son más interesantes, importantes y significativas desde el punto de vista del posible cliente, y el desarrollo de un modelo según el cual, tomando en cuenta las características apreciables del inmueble expuestas en su anuncio, muestre si tiene un precio acorde al mercado o excesivo.

**Metodología Empleada**

Este trabajo consiste en un análisis de una base de datos de AirBNB para el cual se han empleado las siguientes técnicas:

Análisis descriptivo:

* Cálculo de la frecuencia de anuncios por barrio y presentación comparativa en forma de gráfico de burbuja.
* Histograma de precios por barrio
* Bloxplot de precios por barrio
* Media y varianza de las distintas variables por barrio
* Histograma de precios por barrio, con y sin limitaciones de propiedades a 200.

Modelo de regresión:

* Creación de un modelo de regresión con las variables que aparecen en los anuncios de AirBNB para clientes con un conjunto de entrenamiento (Se detallará más a fondo en el siguiente apartado)
* Comparación con los valores del conjunto de test.

Análisis Clúster

* Método del codo para determinación del nº óptimo de clusters.
* Análisis clúster del conjunto de datos.
* Obtención y representación de las componentes principales

Modelo Logit:

* Creación de un modelo logit con las variables que aparecen en los anuncios de AirBNB para clientes con un conjunto de entrenamiento (Se detallará más a fondo en el siguiente apartado)
* Comparación con los valores del conjunto de test.
* Caracterización de la curva ROC del modelo.

**Desarrollo y resultados**

Para este trabajo se ha comenzado extrayendo las variables más interesantes para estudiar de la base de datos, siendo estas:

* id: Identificador del alojamiento
* accommodates: Nº de plazas disponibles en el alojamiento
* antigüedad: Antigüedad del anuncio
* host\_response\_rate: Tasa de respuesta del anunciante
* host\_response\_time: Tiempo medio del anunciante en responder
* neighborhood\_overview\_flag: Descripción del vecindario
* bathrooms: Número de baños completos (Los aseos cuentan como 0,5)
* beds: Nº de camas disponibles en el alojamiento
* price: Precio por noche
* host\_is\_superhost: Muestra si el anunciante entra en la categoría de superhost (Anunciantes con más de 4,8 de valoración, host response rate superior al 90% y tasa de cancelación inferior al 1%).
* host\_listings\_count: Número de alojamientos anunciados en Airbnb del mismo propietario
* host\_has\_profile\_pic: Disponibilidad de foto en el perfil del anunciante
* host\_identity\_verified: Verificación, o no, de la identidad del propietario.
* neighbourhood\_cleansed: Barrio en que se enceuntra el alojamiento
* property\_type: Tipo de propiedad anunciada
* room\_type: Tipo de los cuartos de la propiedad
* minimum\_nights: Mínimo de noches de las que se puede disponer el alojamiento
* maximum\_nights: Máximo de noches de las que se puede disponer el alojamiento

A priori, no era posible trabajar con todas las variables tal y como eran presentadas en la base de datos. Para ello primero fue necesario realizar algunas transformaciones en las mismas que se pueden encontrar detalladas en el Anexo I.

Una vez adecuadas las variables para poder trabajar con ellas, se procedió a la limpieza de los datos.

Se comenzó con la identificación y eliminación de los **missing values.**

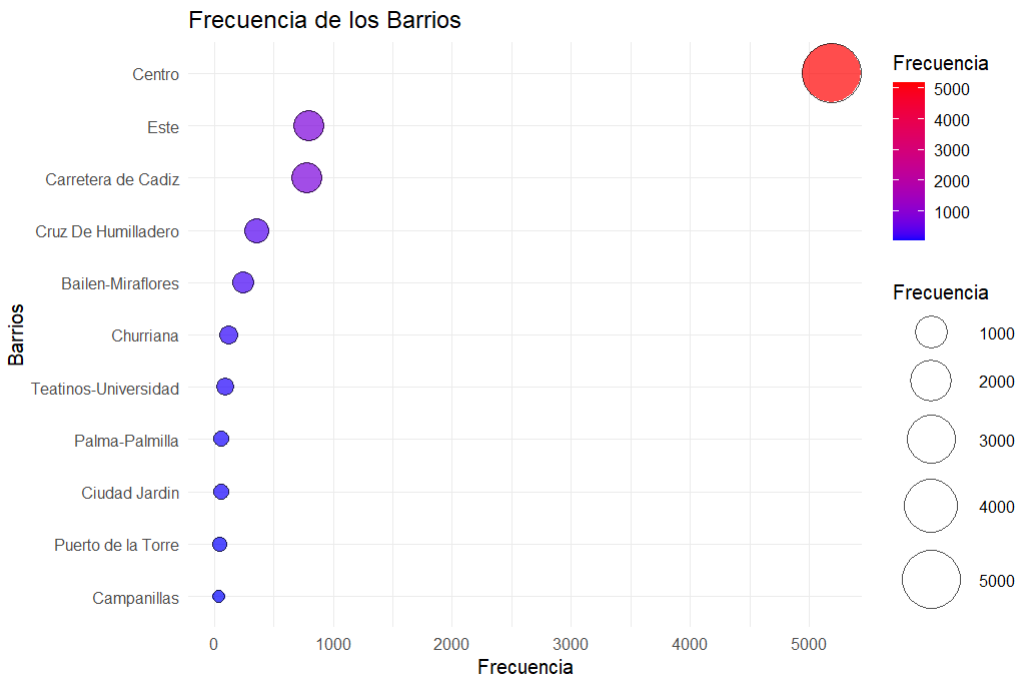


Es en las variables *host\_response\_rate* y *host\_response\_time* donde se dan la mayor parte de los missing values, siendo estos el 8% de los datos de cada una, y en las variables *beds* y *price* también se dan missing values, siendo estos un 1% de los datos de las mismas. El conjunto el total de missing values es un 1% de los datos.

Las filas de los datos con estos missing values fueron eliminadas, con lo que la muestra de datos resultante ya era adecuada para trabajar con ella.

Análisis de la oferta

Para comenzar, se quiere responder a la pregunta ¿Qué barrios son más atractivos turísticamente? Para ello, se estudiará en que barrios hay más anuncios, ya que, es en aquellos barrios con más demanda de alojamientos donde habrá mayor oferta de los mismo.



Es el barrio **Centro** de Málaga el que destaca por encima de los demás con una amplia diferencia, siendo seguido por los barrios **Este** y **Carretera de Cádiz**.

Mostrando los resultados numéricos:

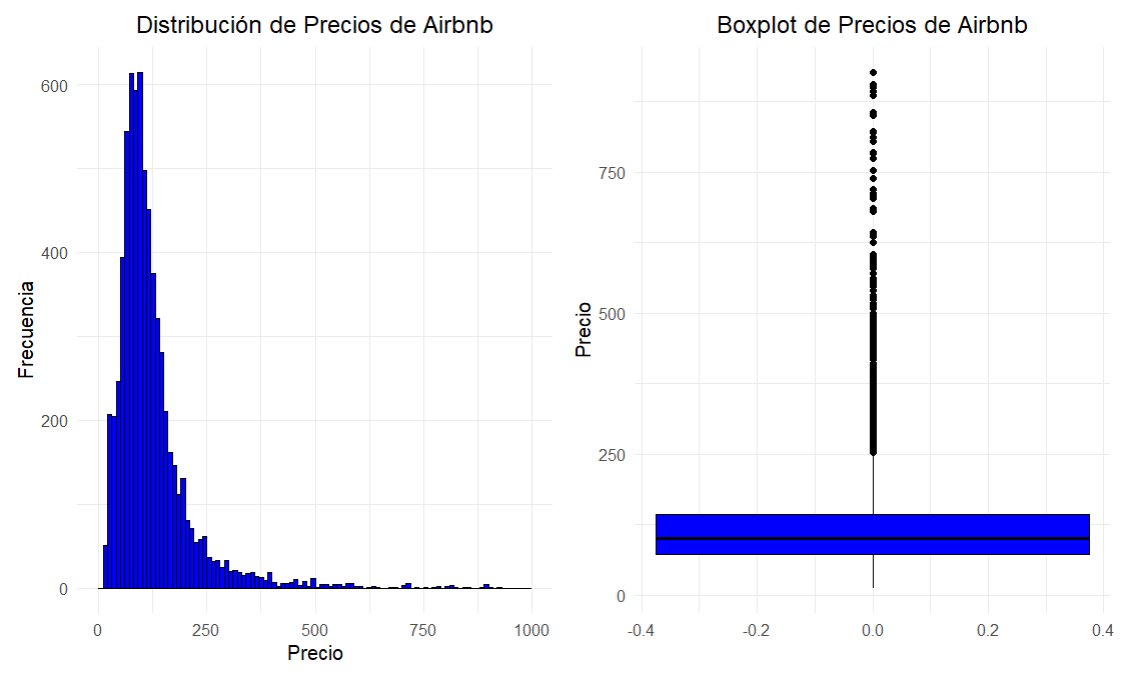
| **barrio** | **frecuencia** |
| --- | --- |
| Centro | 5,189 |
| Este | 792 |
| Carretera de Cadiz | 774 |
| Cruz De Humilladero | 353 |
| Bailen-Miraflores | 245 |
| Churriana | 125 |
| Teatinos-Universidad | 92 |
| Ciudad Jardin | 64 |
| Palma-Palmilla | 64 |
| Puerto de la Torre | 46 |
| Campanillas | 39 |

El nº de anuncios en el barrio **Centro** es ampliamente superior al de anuncios en el resto de barrios, por lo que a priori se puede intuir que este será el barrio más interesante para aquellas personas que busquen un alojamiento no hostelero en la plataforma AirBNB.

Análisis descriptivo

En este apartado, se va a mostrar el análisis realizado al conjunto de datos. La variable sobre la que se va a seguir el estudio en todo momento es el precio.

Para comenzar, se estudia la distribución del precio según la frecuencia de anuncios, así como el boxplot del precio considerando todos los casos:

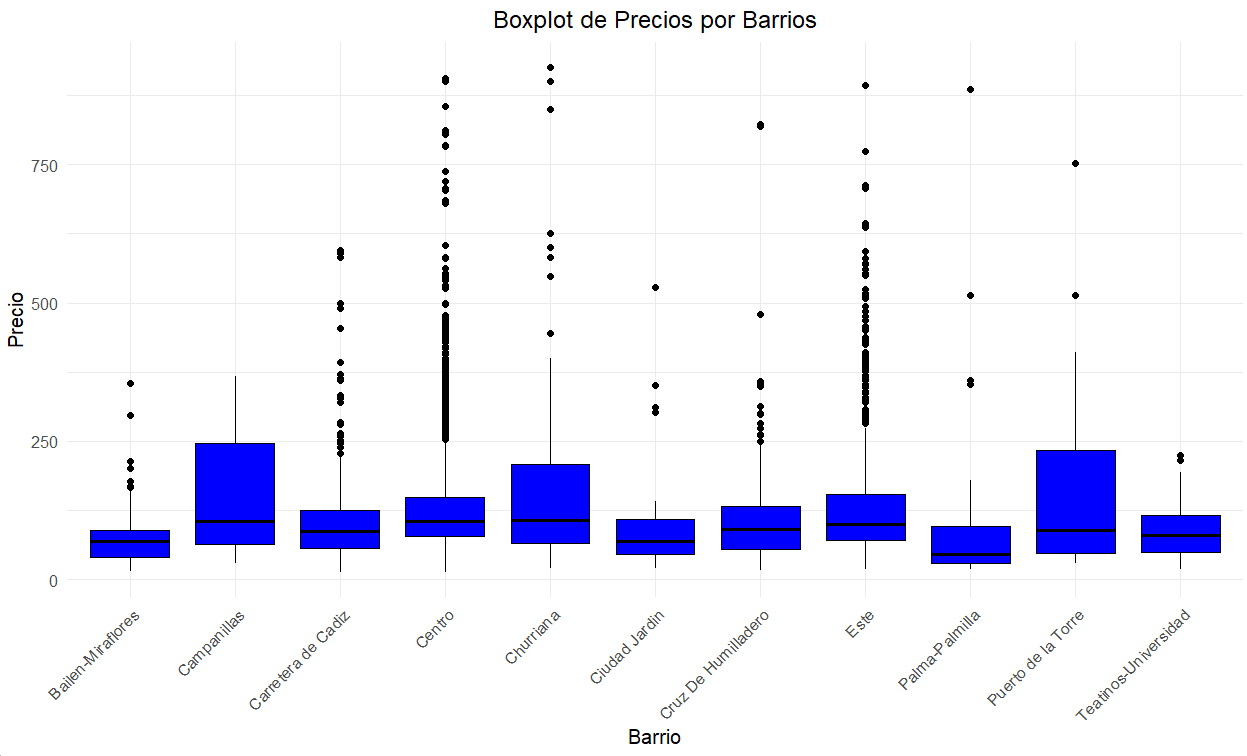


La distribución de precios permite apreciar que la mayor parte de los anuncios tienen precios inferiores a 250 euros por noche, siendo la concentración de los mismos significativamente menor a partir de los 250 euros por noche.

El gráfico boxplot permite precisar aún más, y ver que los outliers comienzan a partir de los 250 euros por noche.

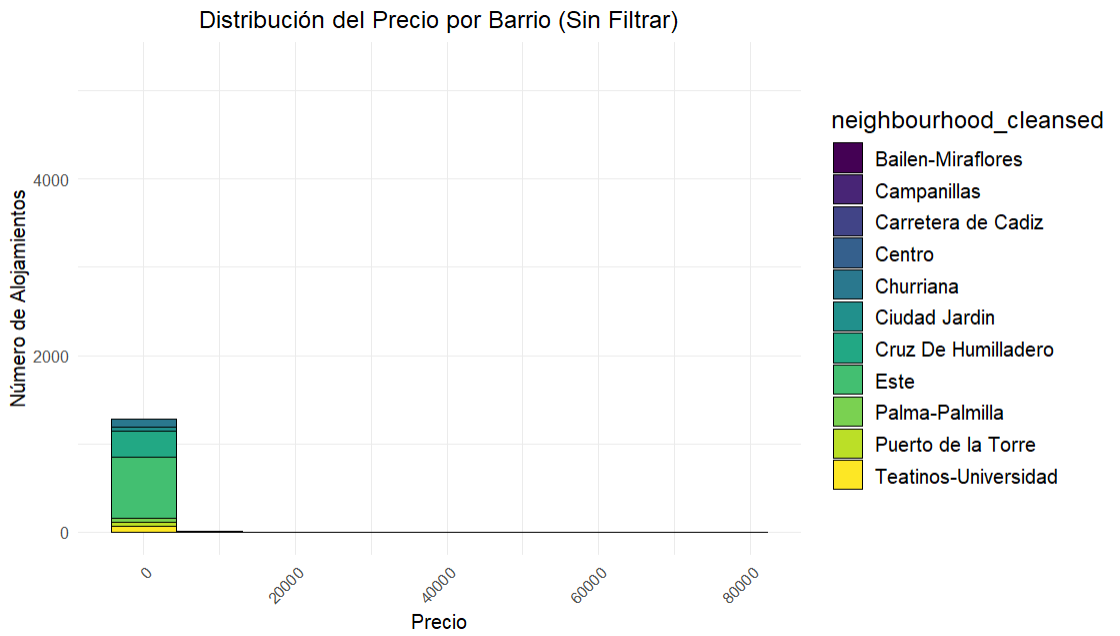
Cabe destacar, y se verá en el resto del análisis más claramente, que todos los outliers se dan por precio superior, es decir, en toda la muestra de datos, no existe ningún caso en el que un alojamiento se oferte por debajo del precio de mercado, sino que los casos que destacan es porque están por encima del mismo.

Caracterizando barrio a barrio, se obtiene:



* Bailen-Miraflores: Tiene un rango de precios bajo, con algunos outliers destacables.
* Campanillas: Es el barrio con más variabilidad de precios dentro del rango intercuartílico recogido por el boxplot. Destaca por ser el único barrio sin outliers.
* Carretera de Cádiz: Rango de precios bajos, presencia de outliers hasta cerca de los 625 euros por noche.
* Centro: El rango intercuartílico no es de los más elevados, sin embargo, destaca por ser el que tiene una mayor presencia de outliers. Debe recordarse que este es el barrio con el mayor nº de anuncios con una amplia diferencia.
* Churriana: Rango intercuartílico amplio y pocos outliers en él.
* Ciudad Jardín: Su rango intercuartílico no es muy anplio, pero sus outliers, pese a ser pocos, tienen una amplia diferencia en el orecio con respecto a sus precios normales.
* Cruz de Humilladero: Análogo a Ciudad Jardín
* Este: Rango intercuartílico destacable. Gran presencia de outliers que llegan a alxanzar el rango de precios más elevado.
* Palma-Palmilla: Rango de precios bajo. Sin embargo, destacan sus outliers, los cuales guardan una gran diferencia en el precios, llegando uno de ellos al rango más alto.
* Puerto de la Torre: Rango intercuartílico bastante amplio y con presencia de 2 outlier destacables, cuyos precios son altos.
* Teatinos-Universidad: Rango de precios no medio. Tiene 2 outliers, pero a diferencia del resto de barrios, estos no se alejan demasiado de su rango de precios.

Estos resultados, expuestos en forma de distribución, y calculando 10 intervalos entre el precio mínimo y el máximo dan:

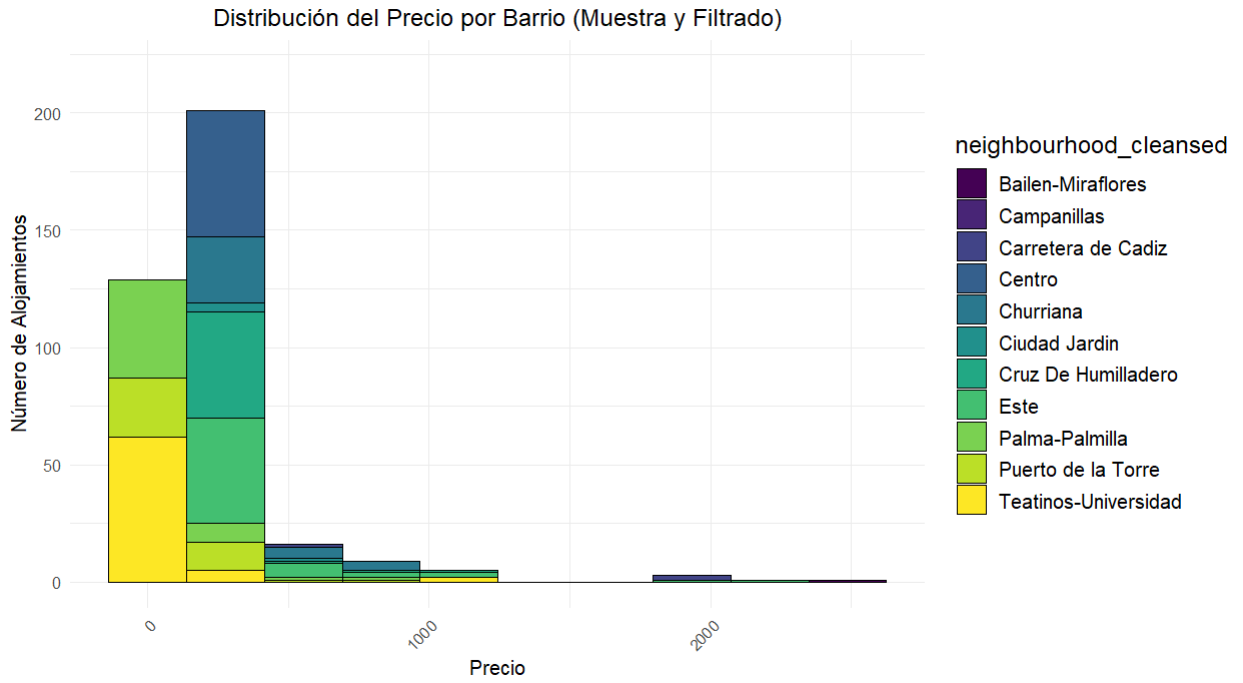


Visto así, no es posible extraer demasiada información. Para ello, en visos de una mejor apreciación, se ha extraído una muestra aleatoria de 200 alojamientos de aquellos barrios que superaban dicho número para apreciar mejor la distribución de los precios.

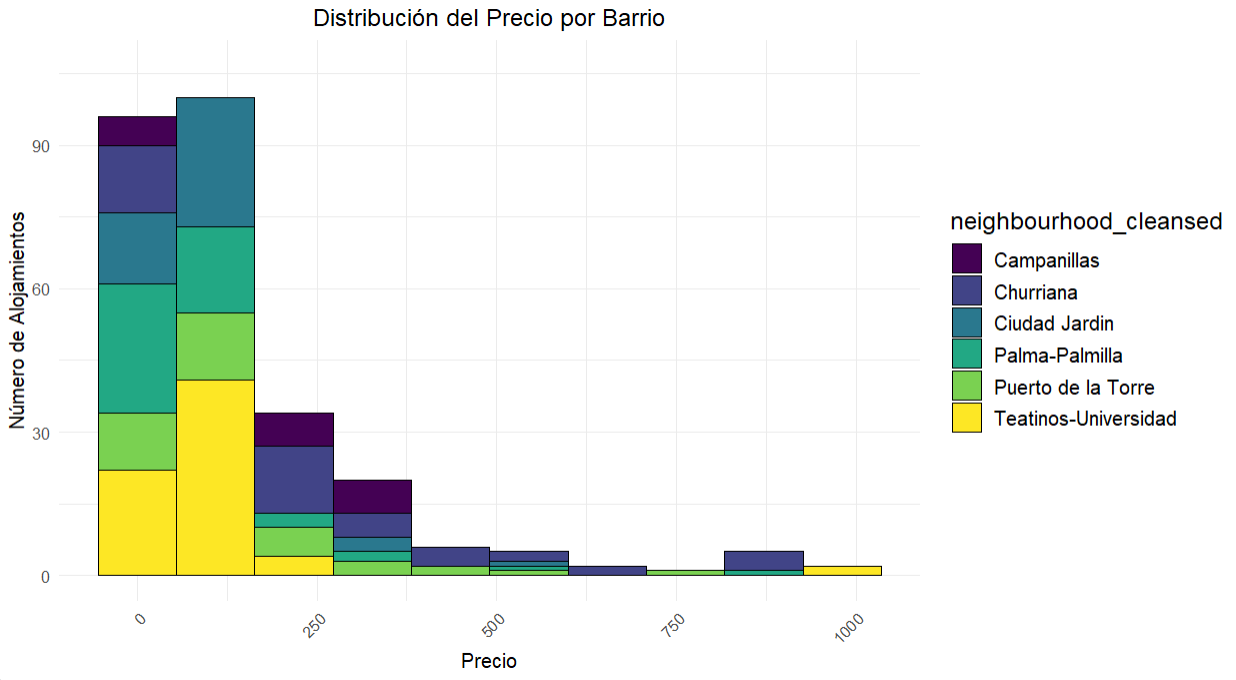
Dichos barrios son:

|  |  |  |
| --- | --- | --- |
| **Barrio** | **Nº de anuncios** | **Exceso sobre 200** |
| Bailen-Miraflores | 206 | 6 |
| Carretera de Cádiz | 652 | 452 |
| Centro | 4806 | 4606 |
| Cruz de Humilladero | 293 | 93 |
| Este | 690 | 490 |

La distribución ahora queda:



Aun así, sigue siendo difícil sacar algunas conclusiones, así que, con el objetivo de poder comparar, se ha realizado también la distribución sin contar con los barrios de más de 200 propiedades, resultando:



Modelo de regresión

Este trabajo está enfocado desde el punto de vista de un posible cliente, por lo que, las variables que se consideran para la construcción del modelo son aquellas que se pueden ver desde los anuncios de Airbnb. El objetivo no es tanto hallar un modelo de regresión que explique el precio de los anuncios de los alojamientos turísticos no hosteleros (Aunque esto se hará), sino encontrar cuales de estas variables son significativas para explicar los precios de los anuncios.

El modelo no será solo creado, sino que por supuesto también será testeado. Para ello, primeramente, se limpiarán los outliers, pasando el conjunto de datos de \_\_\_\_ filas a \_\_\_filas.

A continuación, se dividirá el conjunto de datos entre **datos de entrenamiento** (80% de los datos) y **datos de prueba** (20% de los datos restantes).

Las variables a considerar para la creación del modelo de regresión son:

Variable explicada:

* Price

Variables explicativas:

* accommodates
* host\_response\_rate
* bathrooms
* beds
* minimum\_nights
* host\_has\_profile\_pic
* host\_identity\_verified
* property\_type
* room\_type
* maximum\_nights
* minimum\_nights
* neighbourhood\_cleansed

Estas son las variables que se pueden extraer de un alojamiento a partir de un anuncio de un alojamiento turístico no hostelero en AirBNB

El modelo resultante, así como sus métricas de evaluación se encuentran en el archivo Regresión\_model\_summary.txt[[1]](#footnote-1)

(*INSERTAR AQUÍ RESULTADOS; VARIABLES SIGNIFICATIVAS Y MÉTRICAS DE ERROR*)

Tomando un 95% de confianza, las variables significativas devueltas por el modelo son:

* accommodates
* host\_response\_time
* bathrooms
* beds
* host\_is\_superhost
* property\_type
* room\_type
* maximum\_nights
* minimum\_nights
* neighbourhood\_cleansed

Análisis Clúster

Una vez se tiene calculado el modelo y sabiendo cuales son las variables significativas, se parte de ellas para realizar un análisis clúster del conjunto de datos.

Se añade la columna de *price* al conjunto tomado de variables significatiavs para tenerlo en cuenta para el análisis clúster, ya que es una variable de mucha importancia en el análisis.

También, antes de la realización de la división por clústers, se procede a normalizar als variables para poder trabajar mejor con ellas.

Para determinar el número óptimo de clúster en que dividir el conjunto, se realiza el **Método del codo**.

(*INSERTAR RESULTADOS DEL MODELO DEL CODO Y GRÁFICA*)

Una vez realizado, se introducen los resultados de los clusters a los datos

(*INSERTAR CABECERA DE DATOS CLUSTERS(¿)*)

Con esto ya realizado, se procede calcular las componentes principales, seleccionando las dos primeras componentes principales, obteniéndose:

*(INSERTAR GRÁFICA COMPONENTES PRINCIPALES Y EXPLICARLA)*

Pero ¿Explican realmente bien estas componentes principales el comportamiento de los datos?

Para ver si esto es así, se puede ver la varianza explicada acumulada de las componentes principales:

*(INSERTAR GRÁFICA DE LA VARIANZA ACUMULADA EXPLICADA Y LA EXPLICACIÓN)*

Modelo logit

Teniendo en cuenta que el modelo logit es un modelo de elección binaria, primero hay que definir qué elección se quiere estudiar. Para el caso de este estudio, se va a crear una variable, que se llamará *precio\_alto*, la cual se define en este estudio como 1.5 veces la mediana de *price*.

La idea tras este modelo es, en base a las características que se pueden encontrar de cierto alojamiento turístico no hostelero en el anuncio de AirBNB, responder a la pregunta “¿Es el precio por noche de este anuncio excesivo?”[[2]](#footnote-2).

Al igual que con el modelo de regresión calculado anteriormente, se divide el conjunto de datos en **datos de entrenamiento datos de entrenamiento** (80%) y **datos de prueba** (20%), realizándose el modelo con los datos de entrenamiento y el testeo del mismo con los datos de prueba.[[3]](#footnote-3)

(*INSERTAR AQUÍ RESULTADOS; VARIABLES SIGNIFICATIVAS Y MÉTRICAS DE ERROR*)

Pero, ¿Es este modelo correcto para realizar identificaciones de los casos en qué el precio es alto? Para comprobar esto, se calcula la curva de ROC del modelo y su área entre la curva y la línea de 45º.

(*INSERTAR AQUÍ GRÁFICO DE LA CURVA DE ROC Y LA EXPLICACIÓN*)

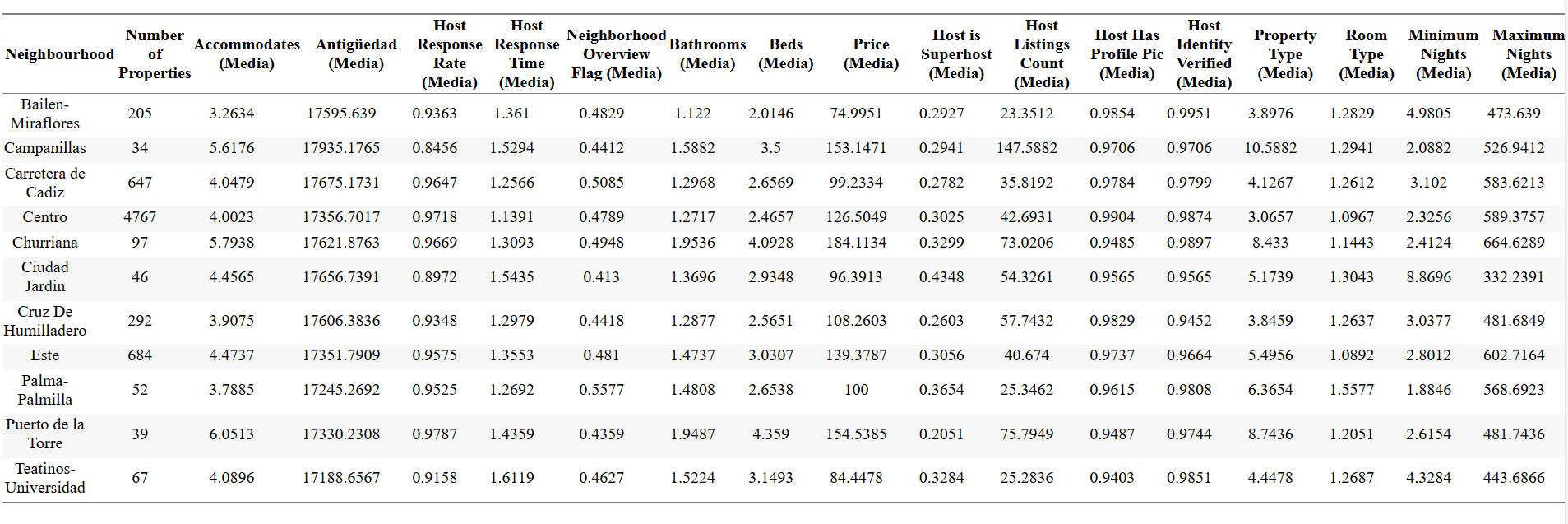
**Conclusiones**

**Anexo I – Tratamiento de las variables**

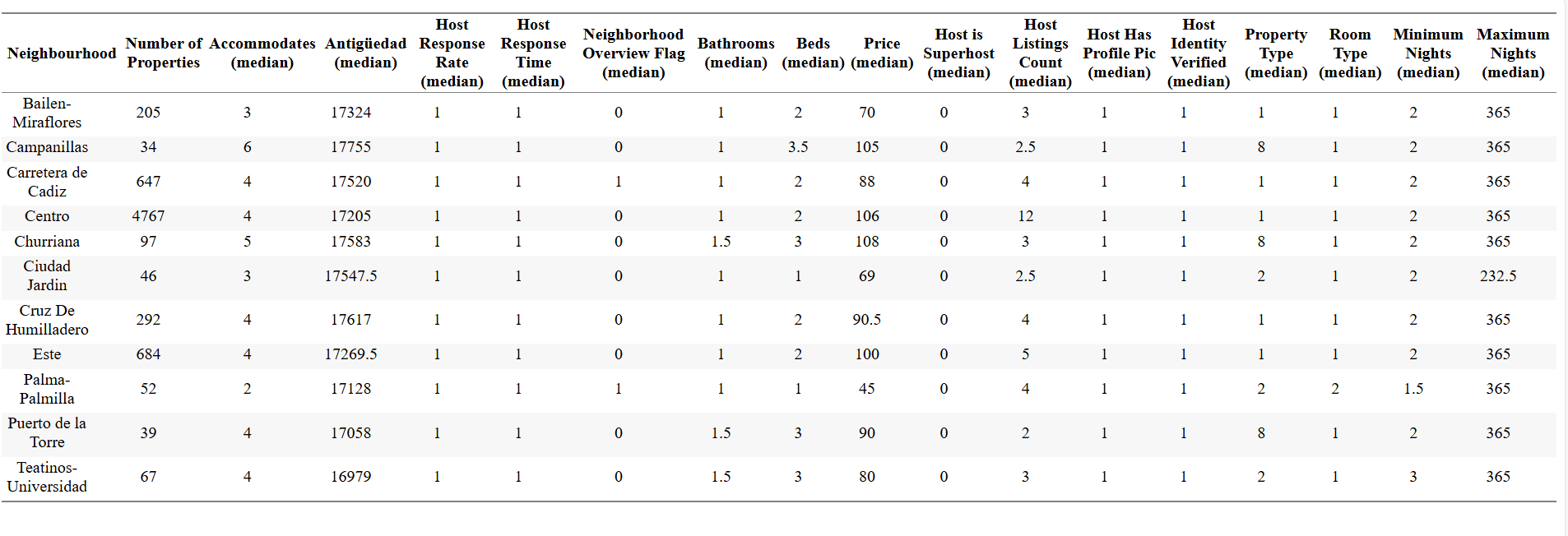
* La variable *host\_response\_time* se dividía en 4 categorías, por lo que a cada una de ellas se le asignó un valor numérico:
* Within an hour: 1
* Within a few hours: 2
* Within a day: 3
* A few days or more: 4
* Las variables *host\_is\_superhost*, *host\_has\_profile\_pic* y *host\_identify\_verified* son de naturaleza binaria, por lo que a las respuestas que ofrecían (**False** y **True**) se les adjudicaron los valores 0 y 1 respectivamente.
* La variable *room\_type* se dividía en 3 categorías, por lo que a cada una de ellas se le asignó un valor numérico:
* Entire home/apt: 1
* Private room: 2
* Shared room: 3
* La variable *property\_type* se divide en 46 categorías, por lo que a cada una de ellas se le asignó un valor numérico:
* Entire rental unit : 1
* Private room in rental unit : 2
* Shared room in rental unit : 3
* Private room in townhouse : 4
* Private room in home : 5
* Entire loft : 6
* Entire condo : 7
* Entire home : 8
* Entire serviced apartment : 9
* Entire townhouse : 10
* Entire cottage : 11
* Entire guest suite : 12
* Entire chalet : 13
* Private room in condo : 14
* Room in boutique hotel : 15
* Entire villa : 16
* Private room in guesthouse : 17
* Room in serviced apartment : 18
* Entire guesthouse : 19
* Floor : 20
* Camper/RV : 21
* Tiny home : 22
* Entire vacation home : 23
* Private room in guest suite : 24
* Room in aparthotel : 25
* Private room : 26
* Shared room in earthen home : 27
* Private room in loft : 28
* Private room in chalet : 29
* Private room in hostel : 30
* Private room in casa particular : 31
* Entire cabin : 32
* Private room in serviced apartment : 33
* Entire place : 34
* Shared room in chalet : 35
* Casa particular : 36
* Private room in bed and breakfast : 37
* Dome : 38
* Private room in villa : 39
* Room in hotel : 40
* Private room in vacation home : 41
* Private room in farm stay : 42
* Private room in yurt : 43
* Shared room in casa particular : 44
* Cave : 45
* Shared room in hotel : 46
* La variable *neighbourhood\_cleansed* se divide en 11 categorías, por lo que a cada una de ellas se le asignó un valor numérico:
* Este : 1
* Centro : 2
* Churriana : 3
* Carretera de Cadiz : 4
* Bailen-Miraflores : 5
* Cruz De Humilladero : 6
* Teatinos-Universidad : 7
* Puerto de la Torre : 8
* Ciudad Jardin : 9
* Campanillas : 10
* Palma-Palmilla : 11

**Anexo II – Descriptivo de las variables**

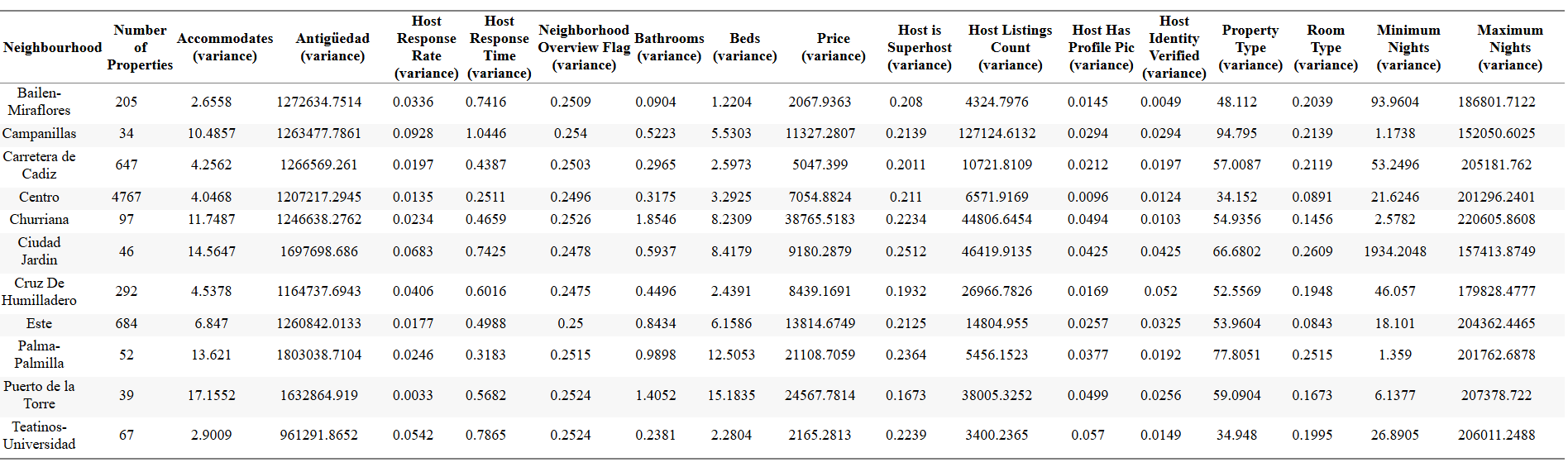
Medias de las variables por barrio



Medianas de las variables por barrio



Varianzas de las variables por barrio



1. El archivo de este modelo desglosado junto a las métricas de evaluación se encuentra en el siguiente enlace: https://github.com/Joseluisvic/Trabajo-Final-De-M-ster [↑](#footnote-ref-1)
2. Se considera solo el caso de que haya exceso porque se ha apreciado en el primer apartado que todos los outliers son por exceso, y ello parece indicar que la tendencia del mercado es al exceso y no al defecto con relación a los precios por noche de los alojamientos. [↑](#footnote-ref-2)
3. El archivo de este modelo desglosado junto a las métricas de evaluación se encuentra en el siguiente enlace: https://github.com/Joseluisvic/Trabajo-Final-De-M-ster [↑](#footnote-ref-3)