Título: Mercado de alquiler turístico no hostelero en Málaga. Análisis descriptivo, componentes explicativas del precio y modelo de comprobación del precio del anuncio sobre el precio de mercado. *Nota: El título es provisional pero de momento no se me ha ocurrido nada mejor.*

Título en inglés:\_\_\_\_\_

Nº de palabras\_\_\_\_

Curso académico: 2023/2024

Autor: Jose Luis Vicaria Cañaveras

Correo electrónico: [JoseLuis.Vicaria@alu.uclm.es](mailto:JoseLuis.Vicaria@alu.uclm.es)

Tutor: Miguel Ángel Tarancón Morán

Fecha\_\_\_\_

Firma tutor\_\_\_\_

“Como la hierba, la idea busca la luz,

ama las multitudes,

el mestizaje la enriquece,

crece más vigorosa cuando se la pisa.”

*Úrsula K. Le Guin*

Índice

[Marco y justificación de la relevancia del objetivo planteado 3](#_Toc177400747)

[Definición del objetivo planteado 4](#_Toc177400748)

[Metodología Empleada 5](#_Toc177400749)

[Desarrollo y resultados 6](#_Toc177400750)

[Análisis descriptivo 9](#_Toc177400751)

[Modelo de regresión 13](#_Toc177400752)

[Análisis Clúster 15](#_Toc177400753)

[Modelo logit 18](#_Toc177400754)

[Conclusiones 24](#_Toc177400755)

[Anexo I – Tratamiento de las variables 26](#_Toc177400756)

[Anexo II – Descriptivo de las variables 28](#_Toc177400757)

## Marco y justificación de la relevancia del objetivo planteado

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

Durante el siglo XX, especialmente en el primer tercio del siglo y posteriormente a partir de los años 60 con el aperturismo del régimen franquista, hubo un incentivo al crecimiento del turismo, viéndose la Costa del Sol fuertemente influenciada debido a sus características climáticas y sus playas.

Desde entonces, el impacto y el peso del turismo en Málaga no han dejado de crecer. 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. Es sobre este último porcentaje donde se centrará este análisis.

La inmensa mayoría de estos alojamientos turísticos, cuyo número en la ciudad de Málaga no para de crecer, se encuentran anunciados en la plataforma Airbnb. Estos se conocen como alojamientos turísticos no hoteleros.

Este crecimiento no está exento de controversia. El pasado mes de julio se llevaron a cabo grandes manifestaciones en contra del turismo masivo, culpando a Airbnb del aumento de los precios del alquiler.

No cabe duda de que los alojamientos turísticos no hoteleros son un tema de actualidad en Málaga.

<https://es.euronews.com/viajes/2024/07/07/malaga-para-vivir-no-para-sobrevivir-los-malaguenos-protestan-contra-el-turismo-en-medio-d>

<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.

## Definición del objetivo planteado

El objetivo de este trabajo es realizar un análisis del mercado de alquiler de alojamientos turísticos no hoteleros en la ciudad de Málaga, específicamente aquellos que se anuncian en la plataforma Airbnb. Este análisis busca identificar y comprender qué características presentes en los anuncios de la plataforma resultan más interesantes, importantes y significativas desde la perspectiva del posible cliente. Asimismo, se pretende desarrollar un modelo que permita responder a la pregunta: ¿Es el precio de un determinado anuncio adecuado o excesivo?

Para alcanzar estos objetivos, se llevará a cabo un estudio detallado que incluirá varias fases. En primer lugar, se procederá a la recopilación y limpieza de datos, extrayendo los mismos de la base de datos de AirBNB. Los datos recogidos incluirán variables como la ubicación, el precio por noche, la capacidad de alojamiento, las comodidades ofrecidas, tasas de respuestas a los anuncios y otros factores que puedan influir en la percepción del valor por parte de los clientes.

Una vez obtenidos los datos, se llevará a cabo un análisis descriptivo para identificar patrones y tendencias en el mercado de alquiler de alojamientos turísticos en Málaga. Se prestará especial atención a la relación entre el precio y las características de los alojamientos, buscando determinar cuáles de estas características tienen un mayor impacto en la fijación de precios.

Posteriormente, se desarrollará un modelo econométrico que permita evaluar la adecuación de los precios de los anuncios en función de las características identificadas. Este modelo tendrá como objetivo ofrecer una herramienta práctica para los potenciales clientes, ayudándoles a determinar si el precio de un anuncio es justo o si, por el contrario, es excesivo.

En resumen, este trabajo pretende proporcionar una visión integral del mercado de alquiler de alojamientos turísticos no hoteleros en Málaga, aportando feedbacks valiosos sobre las características que más valoran los clientes y ofreciendo herramientas para la evaluación justa de los precios en este mercado dinámico y en constante crecimiento.

## 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.

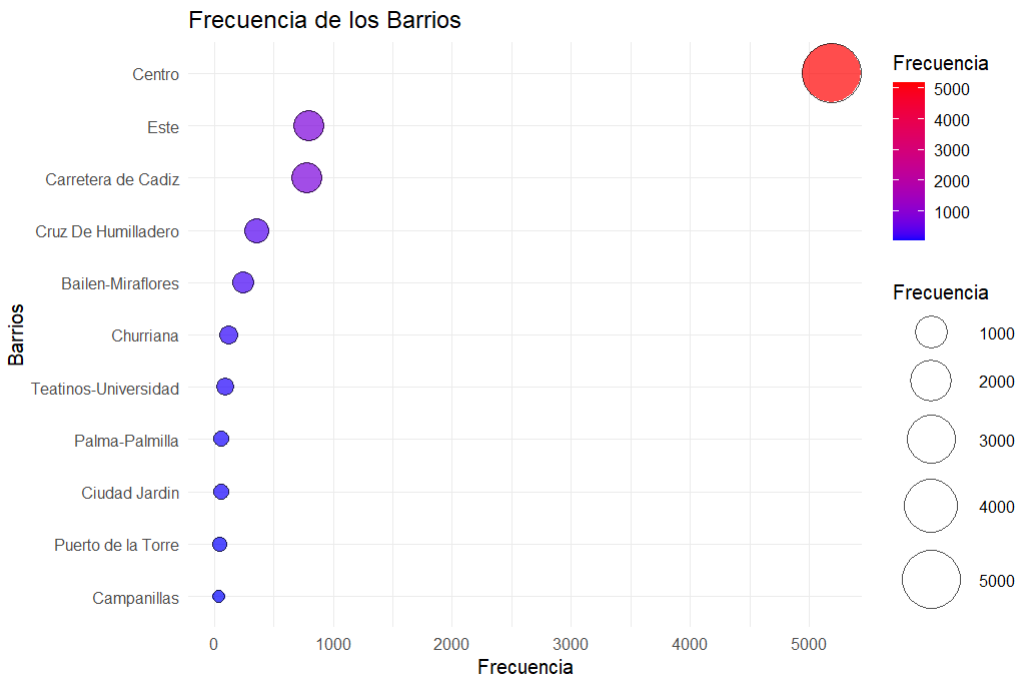
Las filas de los datos con estos missing values fueron eliminadas, con lo que la muestra de datos resultante ya era adecuada para trabajar con ella.

Análisis de la oferta

Para comenzar, se quiere responder a la pregunta *“¿Qué barrios son más atractivos turísticamente?*”.

Para ello, el procedimiento que se sigue en este trabajo será estudiar en que barrios hay más anuncios, ya que, es en aquellos barrios con más demanda de alojamientos donde habrá mayor oferta de los mismo.

A continuación, se muestra un gráfico de burbuja en el que se representa el número de anuncios en los barrios de Málaga de forma que a mayor número de anuncios, mayor tamaño y mayor tendencia al rojo tendrá la burbuja representada.



Representados todos los barrios, se precia una amplia diferencia del barrio **Centro** frente al resto.

Excepto **Centro**, todos los barrios tienen menos de 1000 anuncios, mientras que **Centro** supera los 5000. Esta diferencia ya de por si es significativa, pero se desarrollará más a fondo a continuación, una vez presentados 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 número de anuncios presente en **Centro** es de 5189, frente a 792 anuncios que hay presente en **Este**, siendo este el segundo barrio más poblado.

Visto desde otra perspectiva, de los 7783 anuncios de alojamientos turísticos no hosteleros presentes en la plataforma AirBNB, 5189 son solamente en el barrio **Centro**, por tanto, este barrio acapara el 66.67% de la oferta de anuncios presentes en toda la ciudad de Málaga.

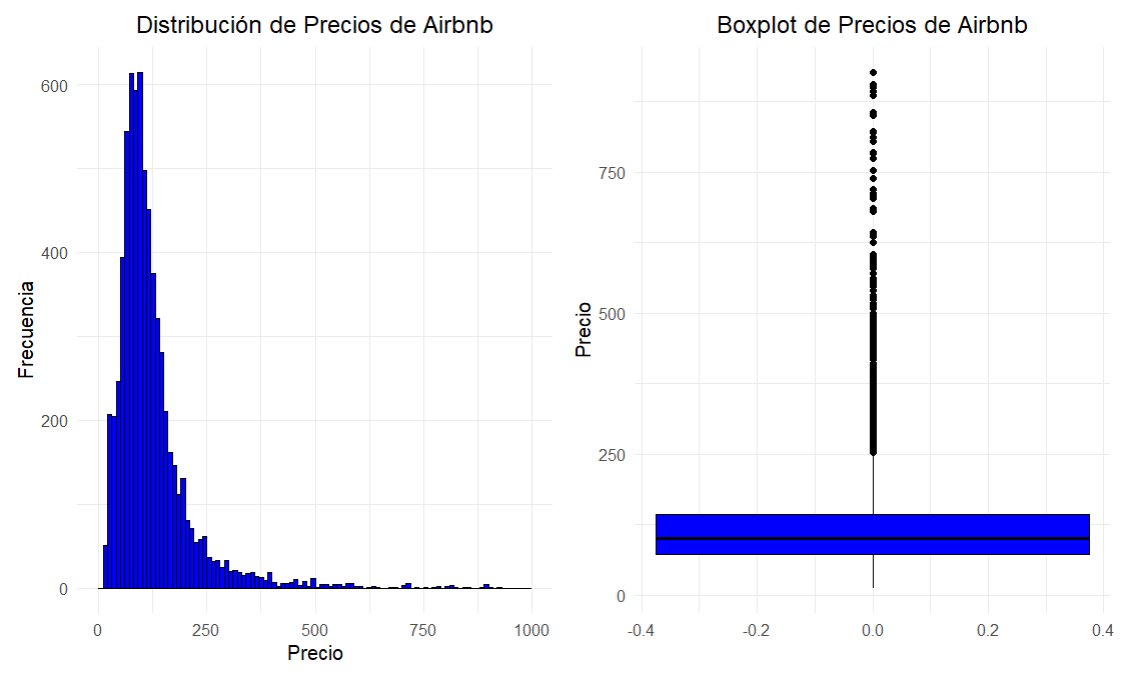
El que la mayoría de la oferta se acumule en este barrio ya muestra que es el barrio más demandado de toda Málaga para los turistas que busquen una estancia en alojamientos no hosteleros en la ciudad.

Además, debe tenerse en cuenta que, al ser el barrio más demandado por una amplia diferencia, los precios de los alojamientos en él tenderán a ser más altos que en el resto, por lo que parece que puede suponerse con estas pistas que el barrio en que se encuentre el alojamiento turístico no hostelero anunciado será uno de los factores importantes en la determinación del precio.

### 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:

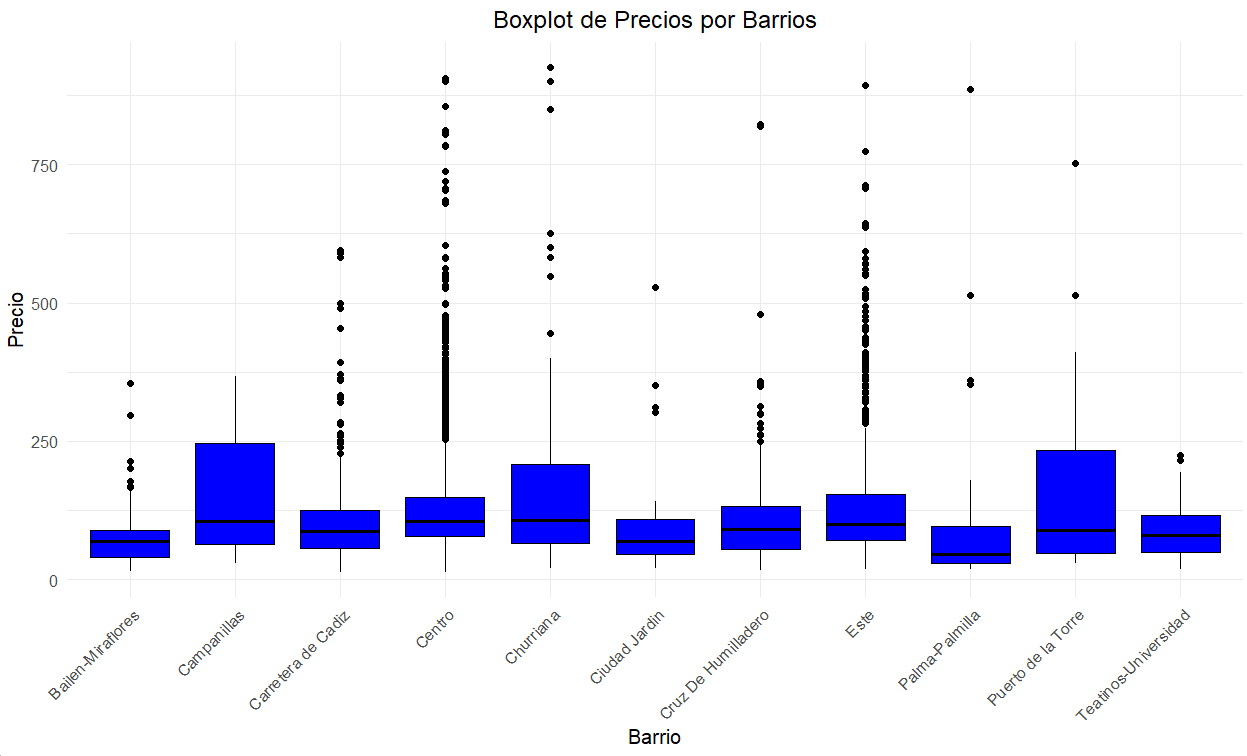


La distribución de precios permite apreciar que la mayor parte de los anuncios tienen precios inferiores a 250 euros por noche, siendo la concentración de los mismos significativamente menor a partir de los 250 euros por noche.

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

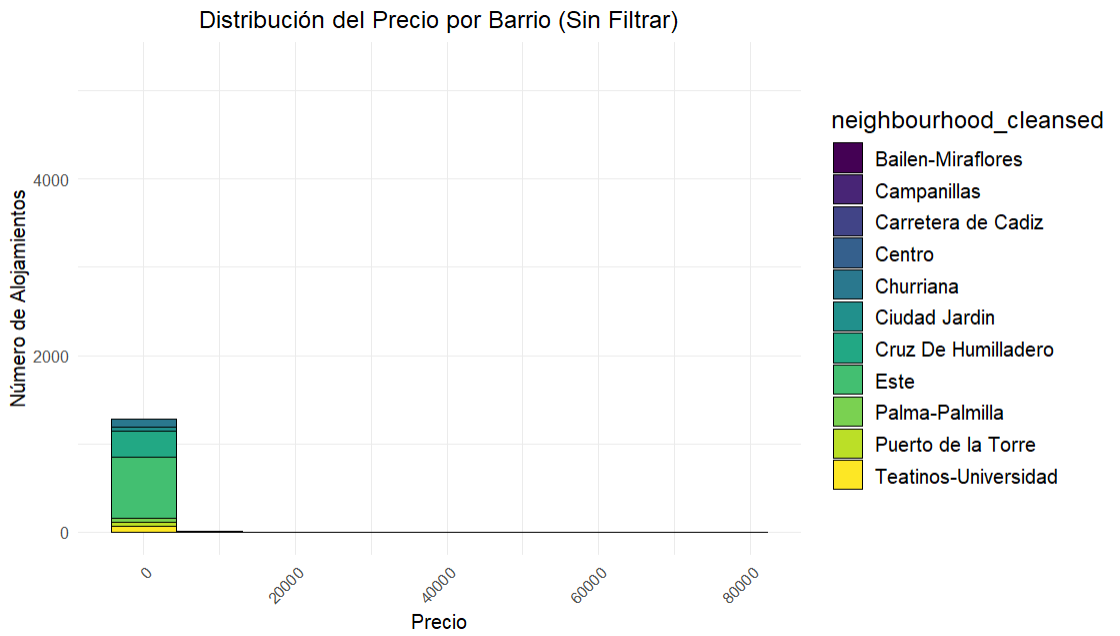
Cabe destacar, y se verá en el resto del análisis más claramente, que todos los outliers se dan por precio superior, es decir, en toda la muestra de datos, no existe ningún caso en el que un alojamiento se oferte por debajo del precio de mercado, sino que los casos que destacan es porque están por encima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

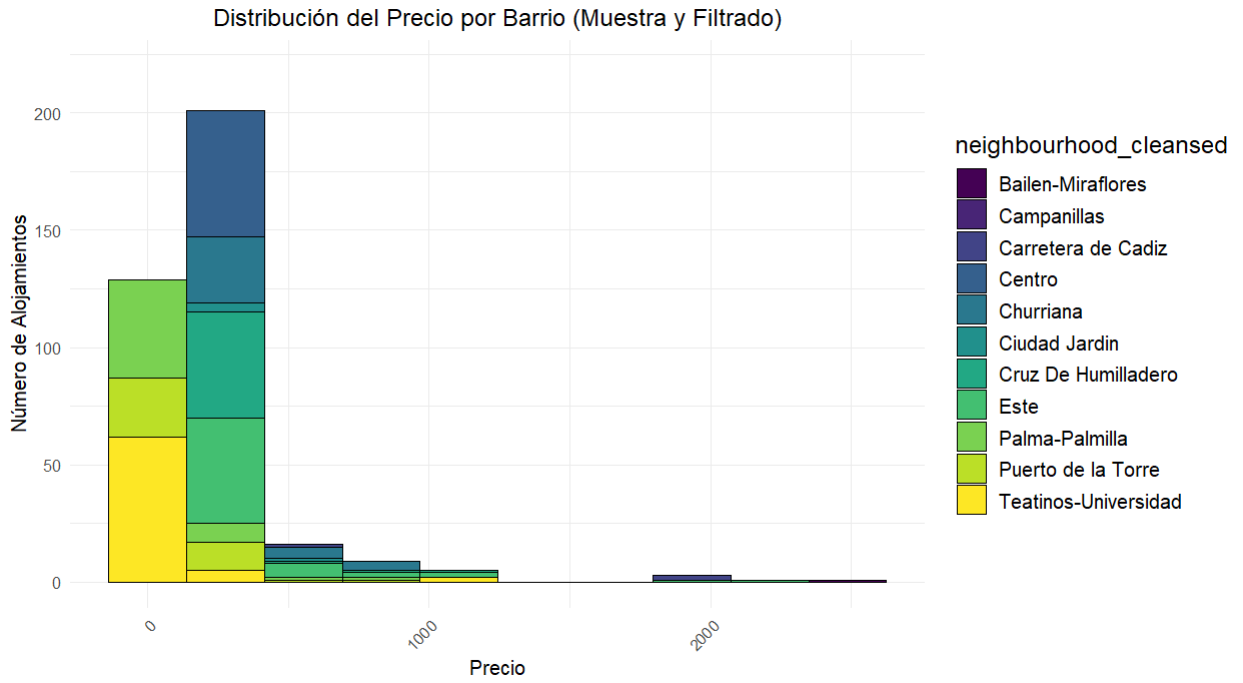


Visto así, no es posible extraer demasiada información. Para ello, en visos de una mejor apreciación, se ha extraído una muestra aleatoria de 200 alojamientos de aquellos barrios que superaban dicho número para apreciar mejor la distribución de los precios.

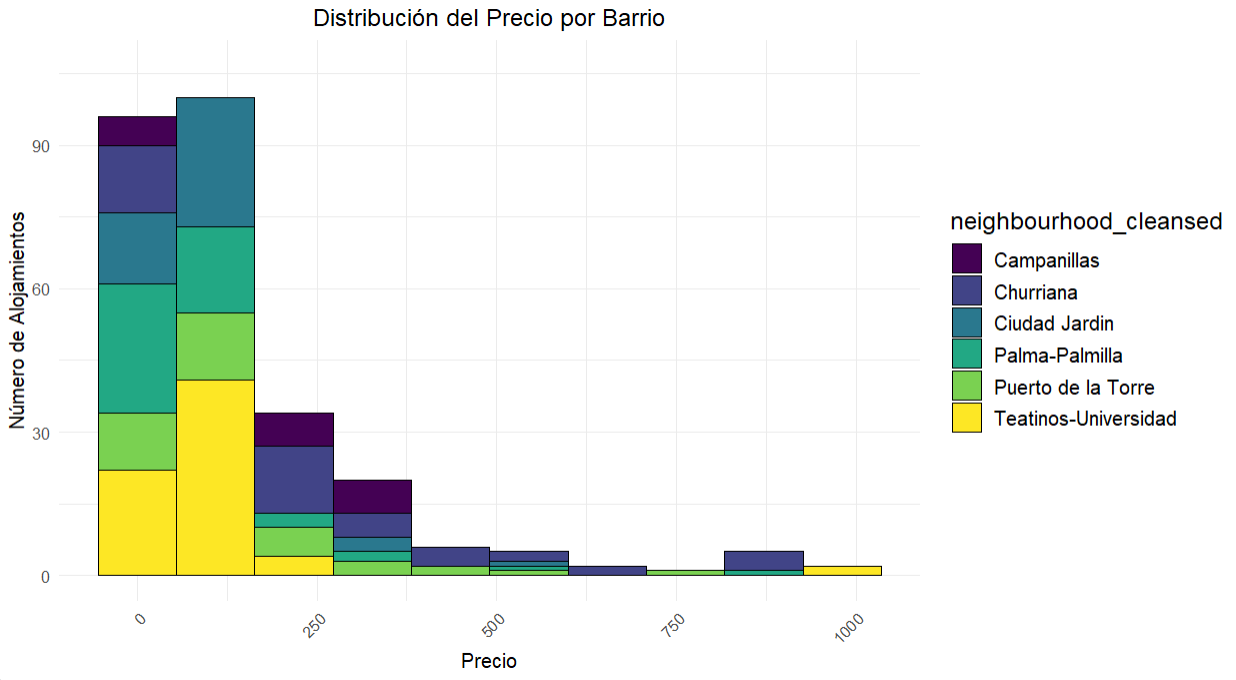
Dichos barrios son:

|  |  |  |
| --- | --- | --- |
| **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, por lo 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 **6984** filas a **6073** filas (Un 13.04% de los datos).

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:

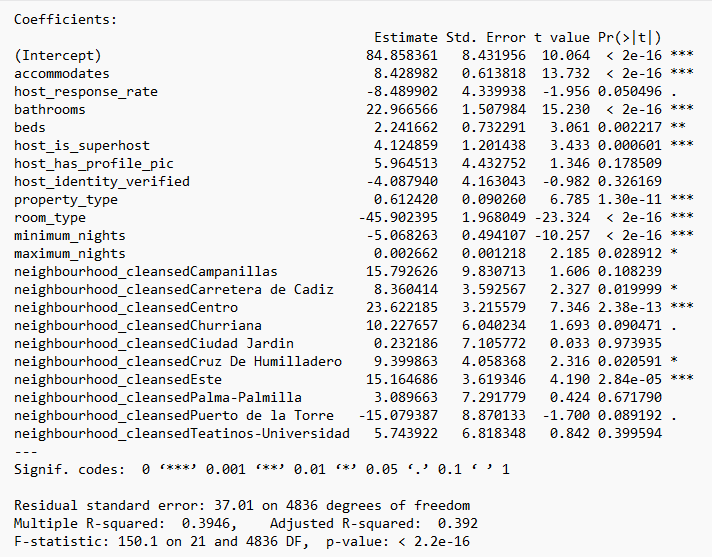
* Price

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\_performance.txt**[[1]](#footnote-1)



Resumen del modelo:

* Residual standard error: La desviación estándar de los residuos de este modelo es de 37.01 en promedio.
* Multiple : 0.3946. La variabilidad en el precio explicada por las variables de este modelo es del 39.46%.
* Adjusted : 0.392. LA variabilidad en el precio explicada por el número de variables de este modelo es del 39.2%.
* F-statistic: 150.1 y p-valor asociado . Muestran que el modelo es significativo en su conjunto.

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\_cleansedCarretera de Cádiz
* neighbourhood\_cleansedCentro
* neighbourhood\_cleansedCruz de Humilladero
* neighbourhood\_cleansedEste

Mirando más en detalle las variables significativas y prestando atención a sus coeficientes:

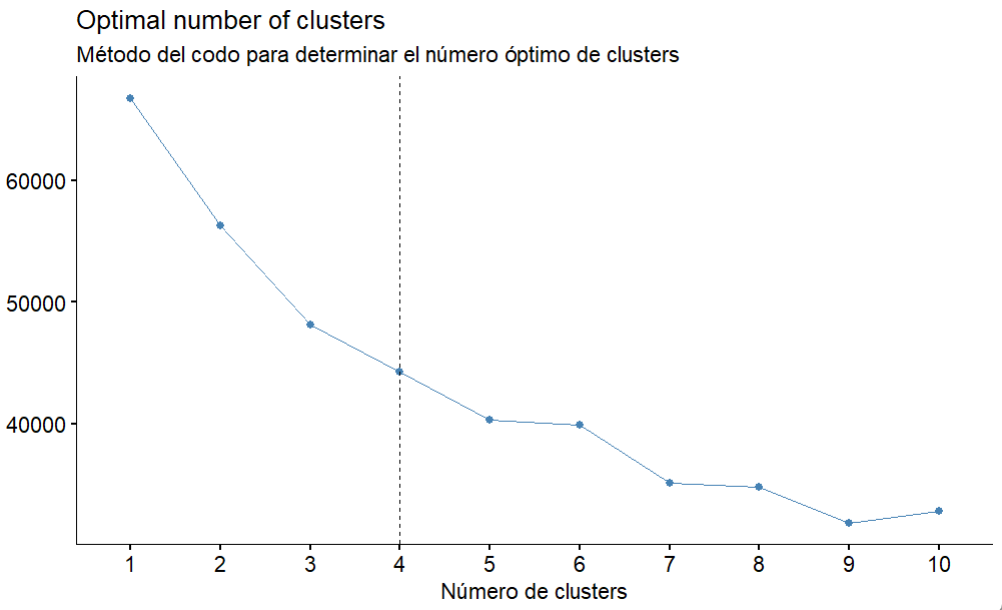
* *accommodates, bathrooms, beds, host\_is\_superhost, property\_type, maximum\_nights, neighbourhood\_cleansedCarretera de Cádiz, neighbourhood\_cleansedCentro, neighbourhood\_cleansedCruz de Humilladero, neighbourhood\_cleansedEste* tienen coeficientes positivos, por lo que tienen un efecto creciente en el precio por noche, sobre todo *bathrooms*, que es la que tiene un mayor coeficiente, siendo este 22.9666.
* *host\_response\_rate, room\_type, minimum\_nights* tienen coeficientes negativos, por lo que tienen un efecto decreciente en el precio por noche, sobre todo *room\_type*, con un coeficiente de -45.9204, que es la que tiene mayor coeficiente. Destacar que no es solo la variable con mayor coeficiente negativo, sino que en valor absoluto es la variable con el mayor coeficiente, por lo que es la variable con más peso de todas.

### Análisis Clúster

Una vez calculado el modelo y determinadas las variables significativas, se utiliza esta información para realizar un análisis de clúster en el conjunto de datos. El objetivo de este análisis es investigar si el barrio donde se encuentra un anuncio afecta de manera significativa la distribución de los datos según su ubicación.

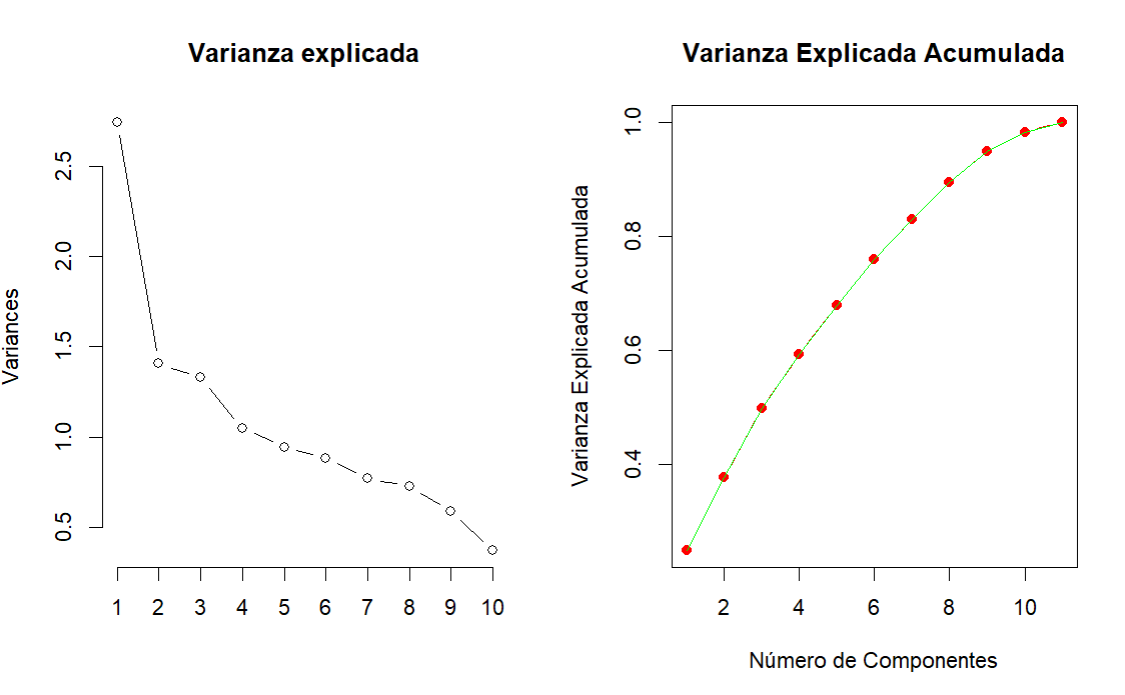
Se añade la columna de *price* al conjunto tomado de variables significativas para tenerlo en cuenta para el análisis clúster, ya que es una variable de mucha importancia en el análisis.

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

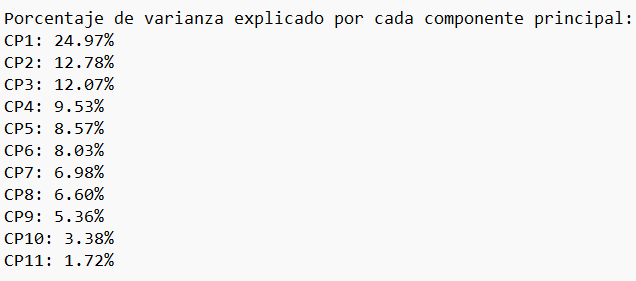
Para determinar el número óptimo de clúster en que dividir el conjunto, se realiza el **Método del codo**. 

El método del Codo indica que la distribución óptima en clusters es de 1 clúster (Al tener el barrio **Centro** la mayoría de los anuncios es probable que el peso de estos provoque que el número óptimo sea 1). Sin embargo, a efectos de estudiar como reaccionan los datos ante divisiones, y si la misma puede proveer alguna idea interesante al estudio, se van a realizar 4 clusters.

Una vez realizado, se introducen los resultados de los clusters a los datos y se procede a calcular las componentes principales, obteniéndose los siguientes resultados del análisis:



Siendo los valores numéricos:



De estos resultados se extraen las siguientes conclusiones:

* La primera componente principal, **CP1**, captura el 24.97% de la varianza total de los datos, lo que implica que casi una cuarta parte de la información de los datos originales se encuentra representada por ella.
* La segunda componente principal, **CP2**, captura el 12.78% de la varianza. Sumando al acumulado (**CP1** solamente en este punto**)**, capturan el 37.75% de la varianza total. Estas dos componentes ya explican por sí solas una parte significativa de la variabilidad.
* La tercera componente, **CP3**, representa el 12.07%, y la cuarta componente, **CP4**, el 9.53%. Cada componente sucesiva va capturando cada vez menos varianza. En este caso, las cuatro primeras componentes principales explican el 59.35% de la variabilidad total de los datos, lo que implica que estos ya se pueden representar razonablemente bien con ellas.
* El resto de componentes explican fracciones progresivamente más pequeñas y, por ende, menos importantes. Para **CP10** y **CP11**, la varianza explicada es muy pequeña (3.38% y 1.72% respectivamente), lo que indica que apenas aportan información adicional sobre la variabilidad de los datos.

Ya que se ha concluido que las dos primeras componentes principales explican por si solas una parte significativa de la variabilidad total de los datos, se procederá a continuar el análisis con ellas:

[[2]](#footnote-2)

La explicación de los resultados de las cargas de la componente principal 1 (**CP1**) es:

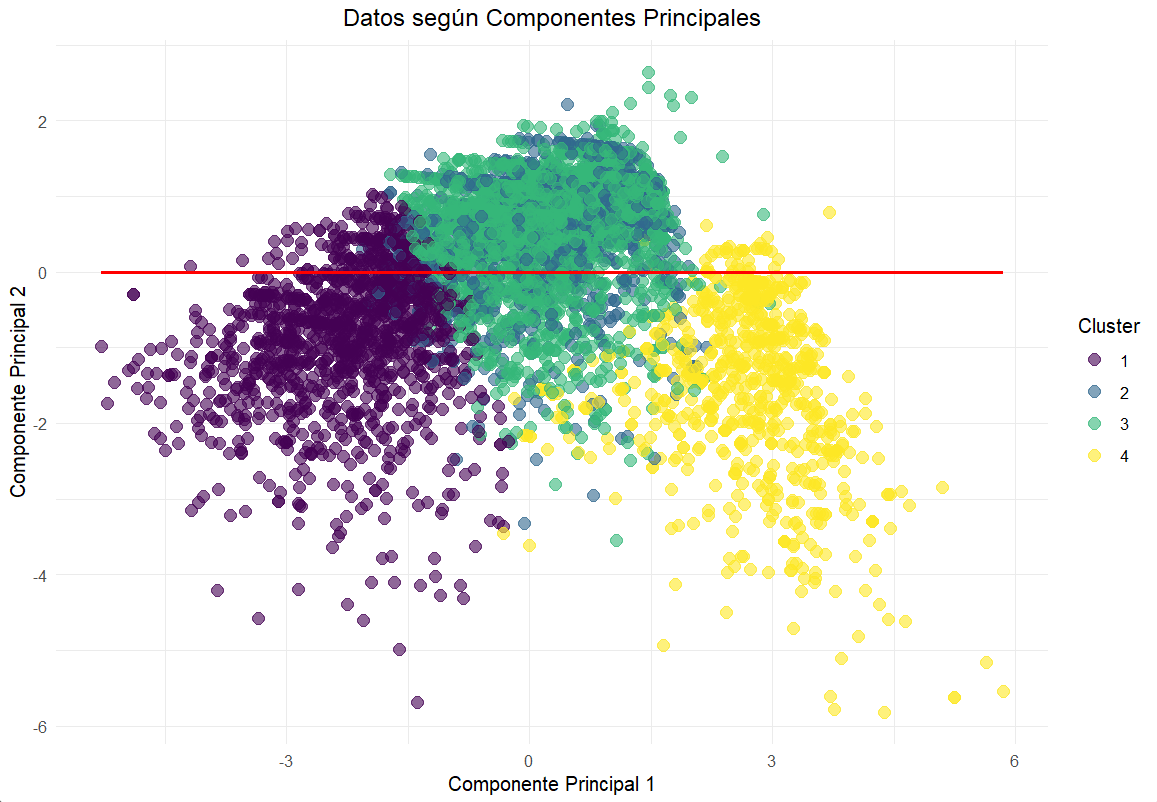
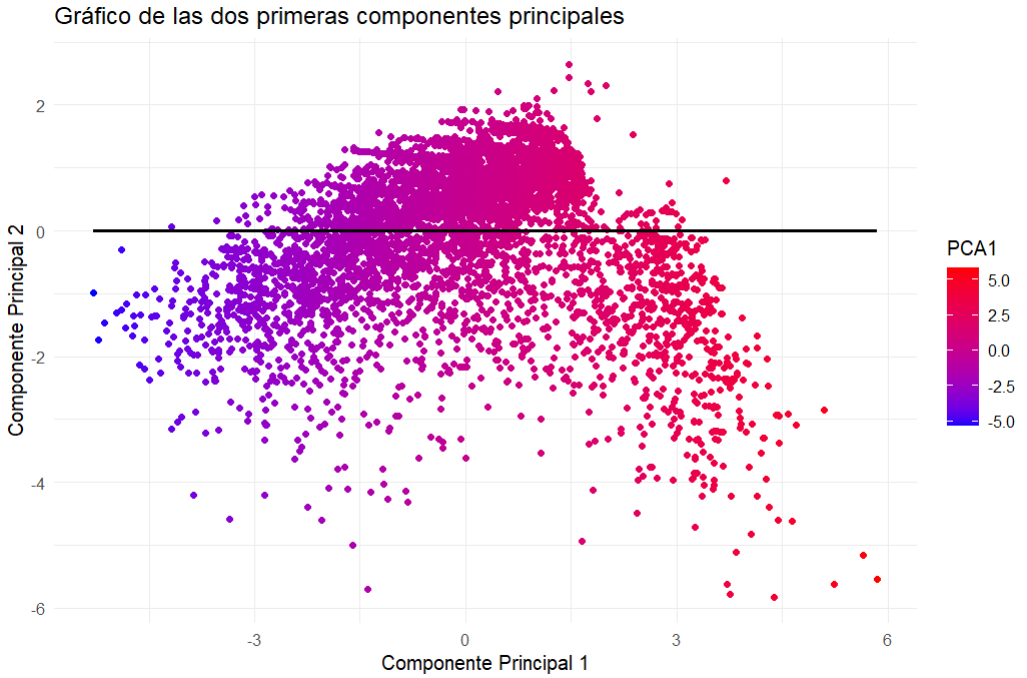
* **accommodates (-0.5257):** Esta variable tiene una carga negativa fuerte. Significa que a medida que la capacidad de alojamiento aumenta, el valor de la CP1 disminuye significativamente.
* **host\_response\_time (0.0385):** Esta variable tiene una carga positiva muy pequeña, lo que indica que tiene un impacto mínimo en la CP1.
* **bathrooms (-0.2816):** La cantidad de baños tiene una carga negativa moderada, contribuyendo a una disminución de la CP1.
* **beds (-0.4784):** Similar a la variable de alojamiento, un mayor número de camas reduce considerablemente el valor de la CP1.
* **host\_is\_superhost (0.0194):** Ser un superanfitrión tiene una carga positiva muy baja, indicando una contribución casi insignificante a la CP1.
* **property\_type (0.2177):** El tipo de propiedad tiene una carga positiva moderada, lo que sugiere que diferentes tipos de propiedad aumentan el valor de la CP1.
* **room\_type (0.3779):** El tipo de habitación tiene una carga positiva, indicando que esta variable aumenta el valor de la CP1.
* **maximum\_nights (-0.1229):** El número máximo de noches tiene una carga negativa pequeña, sugiriendo una leve disminución en la CP1 con un mayor número de noches.
* **minimum\_nights (0.0114):** El número mínimo de noches tiene una carga positiva muy baja, indicando un impacto casi nulo en la CP1.
* **neighbourhood\_cleansed (0.1075):** La variable que representa el vecindario tiene una carga positiva pequeña, sugiriendo una ligera influencia en el aumento de la CP1.
* **price (-0.4435):** El precio tiene una carga negativa considerable, indicando que precios más altos tienden a reducir el valor de la CP1.

Y de la componente principal 2 (**CP2**):

* **accommodates (-0.1830):** Esta variable tiene una carga negativa pequeña, lo que sugiere que a medida que la capacidad de alojamiento aumenta, el valor de la CP2 disminuye ligeramente.
* **host\_response\_time (-0.3081):** El tiempo de respuesta del anfitrión tiene una carga negativa moderada, indicando que tiempos de respuesta más largos reducen el valor de la CP2.
* **bathrooms (-0.4282):** El número de baños tiene una carga negativa significativa, lo que indica que más baños están asociados con una disminución considerable en la CP2.
* **beds (-0.2920):** El número de camas tiene una carga negativa moderada, sugiriendo que más camas reducen el valor de la CP2.
* **host\_is\_superhost (0.0186):** Ser un superanfitrión tiene una carga positiva muy baja, indicando una contribución insignificante a la CP2.
* **property\_type (-0.3874):** El tipo de propiedad tiene una carga negativa considerable, indicando que ciertos tipos de propiedad reducen el valor de la CP2.
* **room\_type (-0.4315):** El tipo de habitación tiene una carga negativa significativa, lo que sugiere que ciertos tipos de habitaciones disminuyen el valor de la CP2.
* **maximum\_nights (0.2197):** El número máximo de noches tiene una carga positiva pequeña, indicando que un mayor número de noches aumenta ligeramente el valor de la CP2.
* **minimum\_nights (-0.0727):** El número mínimo de noches tiene una carga negativa muy pequeña, sugiriendo un impacto leve en la disminución de la CP2.
* **neighbourhood\_cleansed (-0.4590):** El vecindario tiene una carga negativa significativa, indicando que ciertos vecindarios reducen considerablemente el valor de la CP2.
* **price (0.0459):** El precio tiene una carga positiva muy baja, indicando una contribución mínima a la CP2.

En resumen, las variables que más influyen en la primera componente principal son "*accommodates*" y "*beds*" (negativamente), así como "*price*" (también negativamente). Esto sugiere que la **CP1** puede estar representando un factor que combina elementos de capacidad y costo, donde una mayor capacidad de alojamiento y precio más alto se asocian con valores más bajos de **CP1**. En cambio, variables como el tipo de propiedad y el tipo de habitación tienen una influencia positiva moderada, sugiriendo que diferentes tipos de propiedades y habitaciones están asociados con valores más altos de **CP1**, mientras que las variables que más influyen en la segunda componente principal (CP2) son "*bathrooms*", "*room\_type*", y "*neighbourhood\_cleansed*" (todas negativamente). Esto sugiere que la **CP2** puede estar representando un factor relacionado con la calidad o el tipo de alojamiento y la ubicación, donde un mayor número de baños, ciertos tipos de habitaciones y ciertos vecindarios están asociados con valores más bajos de **CP2**. Por otro lado, variables como el número máximo de noches y el precio tienen una influencia positiva muy pequeña, sugiriendo un impacto mínimo en el aumento de la **CP2**.

Si se representan estas 2 primeras componentes principales, representando los clusters se tiene:

**

La representación de las dos componentes principales diferenciando por cluster aparece de esta forma. Se aprecian con cierta facilidad dos clusters, uno en la parte derecha de los gráficos y otro en la izquierda, habiendo 2 clusters en la parte central superior de los mismos superpuestos difíciles de percibir de forma visual.

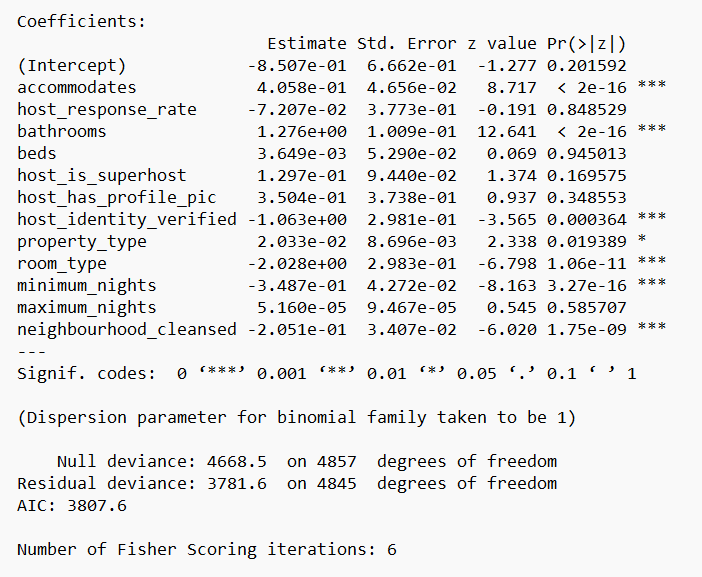
Simplemente a modo de comprobación extra de que el proceso de creación de las componentes principales se ha realizado correctamente, se les ha añadido la línea de tendencia, cuya horizontalidad indica que no existe relación lineal significativa entre la componente 1 y la componente 2, por lo que no son dependientes la una de la otra, lo cual encaja con lo buscado al diseñarse las componentes principales para ser linealmente independientes entre sí.

### 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?”[[3]](#footnote-3).

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.[[4]](#footnote-4)



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

* accommodates
* bathrooms
* property\_type
* host\_identity\_verified
* room\_type
* minimum\_nights
* neighbourhood\_cleansed

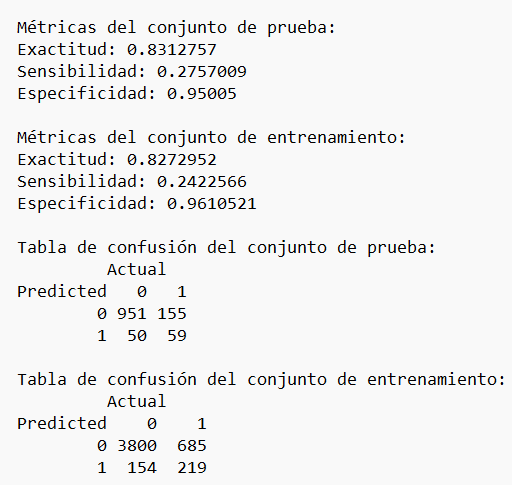
Mirando más en detalle las variables significativas y prestando atención a sus coeficientes:

* *accommodates, bathrooms, property\_type* tienen coeficientes positivos, por lo que, a mayor número de plazas disponibles, baños o el tipo de vivienda, mayor es el impacto positivo en la probabilidad de éxito. De estos, *bathrooms* es la variable con el mayor peso, con mucha diferencia con respecto a las demás.
* *host\_identity\_verified, room\_type, minimum\_nights* tienen coeficientes negativos, por lo que, en caso de que el anunciante tenga su identidad verificada, el tipo de habitación y a mayor número de noches mínimas, menor será la probabilidad de éxito. De estas, *room\_type* es la variable con mayor peso. Al igual que con el modelo de regresión calculado anteriormente, *room\_type* no es solo la variable con mayor coeficiente negativo, sino que es la variable que, en valor absoluto, tiene el mayor valor de coeficiente, indicando que es la variable con mayor peso de todas.

Además

* Deviance Residual: 3781.6. < Null Deviance: 4668.5. El que sea menor la residual que la nula implica que el modelo con los predictores proporciona una mejor representación de los datos que un modelo sin predictores, y además indica una mejora en el ajuste del modelo por la inclusión de las variables independientes.
* AIC:3807.6

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



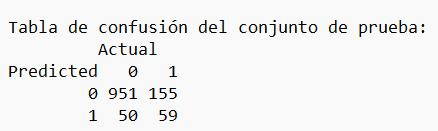
Métricas del conjunto de prueba:

* **Exactitud (0.8312757)**: La exactitud mide el porcentaje de predicciones correctas sobre el total de predicciones realizadas. En este caso, el modelo tuvo una exactitud del 83.13%, lo que significa que, de todas las predicciones en el conjunto de prueba, el 83.13% fueron correctas.
* **Sensibilidad (0.2757009)**: La sensibilidad mide la capacidad del modelo para identificar correctamente los casos positivos (clase 1). Aquí, el modelo identificó correctamente el 27.57% de los casos positivos reales. Esto sugiere que el modelo tiene dificultades para detectar correctamente las observaciones de la clase 1 (positivas).
* **Especificidad (0.95005)**: La especificidad mide la capacidad del modelo para identificar correctamente los casos negativos (clase 0). Una especificidad del 95% indica que el modelo es muy bueno para identificar los casos negativos (es decir, clase 0). Predice correctamente la mayoría de los casos negativos.

Métricas del conjunto de entrenamiento:

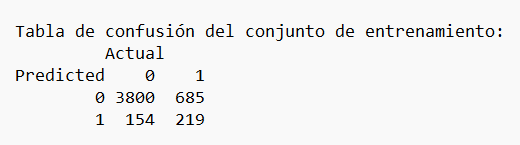
* **Exactitud (0.8272952)**: La exactitud del conjunto de entrenamiento es similar a la del conjunto de prueba, con un 82.73%. Esto indica que el modelo generaliza de manera razonablemente consistente entre el entrenamiento y la prueba, aunque hay una ligera variación.
* **Sensibilidad (0.2422566)**: La sensibilidad es aún más baja en el conjunto de entrenamiento, con solo un 24.22% de los positivos identificados correctamente. Al igual que en el conjunto de prueba, el modelo lucha para detectar los positivos.
* **Especificidad (0.9610521)**: La especificidad es ligeramente superior en el conjunto de entrenamiento (96.11%), lo que nuevamente indica que el modelo clasifica muy bien los casos negativos en ambas fases (entrenamiento y prueba).

Tabla de confusión del conjunto de prueba:



* **951 verdaderos negativos** (TN): El modelo predijo 0 (negativo) y la clase real era 0.
* **59 verdaderos positivos** (TP): El modelo predijo 1 (positivo) y la clase real era 1.
* **50 falsos positivos** (FP): El modelo predijo 1 (positivo) pero la clase real era 0.
* **155 falsos negativos** (FN): El modelo predijo 0 (negativo) pero la clase real era 1.

Tabla de confusión del conjunto de entrenamiento:



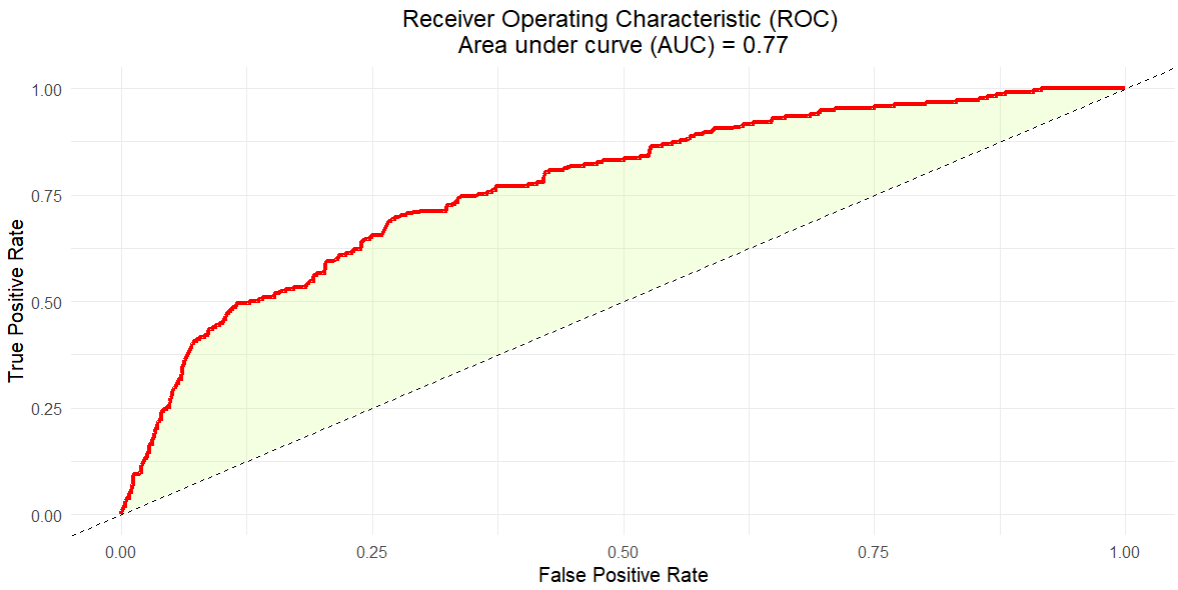
* **3800 verdaderos negativos** (TN): El modelo predijo 0 y la clase real era 0.
* **219 verdaderos positivos** (TP): El modelo predijo 1 y la clase real era 1.
* **154 falsos positivos** (FP): El modelo predijo 1 pero la clase real era 0.
* **685 falsos negativos** (FN): El modelo predijo 0 pero la clase real era 1.

Análisis general de los resultados

* **Alto desequilibrio entre sensibilidad y especificidad**: Aunque la especificidad es muy alta en ambos conjuntos (entrenamiento y prueba), lo que significa que el modelo es muy bueno en predecir los casos negativos (clase 0), la sensibilidad es bastante baja. Esto indica que el modelo no detecta bien los positivos (clase 1), por lo que pueden darse casos de que efectivamente el precio del alojamiento turístico no hostelero estudiado si sea de “Precio alto” y el modelo no lo recoja adecuadamente.
* **Posible desbalance de clases**: La baja sensibilidad sugiere que podría haber un **desequilibrio en las clases** (más ejemplos de la clase 0 que de la clase 1), lo que podría llevar al modelo a priorizar la predicción de la clase mayoritaria (clase 0) y subestimar la importancia de la clase minoritaria (clase 1), dando más peso por tanto a alojamientos que no son de “Precio alto”.
* **Modelo conservador en predicciones positivas**: El modelo parece ser conservador en cuanto a predecir la clase 1, pues tiene más falsos negativos que verdaderos positivos en ambos conjuntos.

A pesar de esto, la capacidad de distinción del modelo entre los casos de “Precio alto” y no es razonablemente buena.

Ello lo muestra el cálculo de su curva ROC, y del área bajo la misma:



En este caso se da que el área bajo la curva ROC es de 0.77, encontrándose en el rango que se considera entre Rendimiento aceptable a Bueno.

## Conclusiones

En cuanto al mercado del alquiler de alojamientos turísticos no hosteleros en la ciudad de Málaga, la mayor oferta se da en el barrio **Centro**, el cual acumula la mayor parte de los anuncios.

El rango más común de precios de los anuncios de alojamientos turísticos no hosteleros en Málaga se da en el intervalo ]0, 250] euros por noche.

El precio de los anuncios viene determinado por las siguientes variables explicativas al 95% de confianza:

* accommodates
* host\_response\_time
* bathrooms
* beds
* host\_is\_superhost
* property\_type
* room\_type
* maximum\_nights
* minimum\_nights
* neighbourhood\_cleansedCarretera de Cádiz
* neighbourhood\_cleansedCentro
* neighbourhood\_cleansedCruz de Humilladero
* neighbourhood\_cleansedEste

De estas variables, *bathroom* y *room\_type* son las dos variables con los coeficientes más grandes (En valor absoluto), implicando que son las dos variables con más peso a la hora de la determinación del precio de un anuncio.

Esto lo que quiere decir es que los dos principales factores que determinan el precio por noche de un alojamiento en la ciudad de Málaga sin tener en cuenta su barrio de localización son: El tipo de habitación que se oferta y el número de baños de los que va a tenerse disponibilidad en el alojamiento.

Dentro de los barrios, el peso del coeficiente asociado al barrio **Centro** es grande, y positivo, tanto por si mismo como por comparación con el resto de barrios, demostrando la importancia de este a la hora de determinar el precio de un alojamiento, indicando que el precio por noche de un alojamiento con unas ciertas características será significativamente mayor si se encuentra en el barrio **Centro** que el precio por noche de un alojamiento similar en otro barrio.

En relación al análisis clúster, pese a que es posible la división en tantos clusters como se desee, el nº óptimo de clusters en los que dividir estos datos es 1, ya que el gran nº de anuncios en **Centro** en comparación con el resto de barrios eclipsa los efectos del resto.

De las componentes principales calculadas después de este análisis clúster se tiene que:

* **Primera Componente Principal (CP1):** Las variables que más influyen son "*accommodates*" y "*beds*" (ambas negativamente) y "*price*" (también negativamente). Esto indica que la **CP1** refleja un factor combinado de capacidad y costo, donde una mayor capacidad de alojamiento y un precio más alto se asocian con valores más bajos de **CP1**. Además, el tipo de propiedad y el tipo de habitación tienen una influencia positiva moderada, lo que sugiere que diferentes tipos de propiedades y habitaciones están asociados con valores más altos de **CP1**.
* **Segunda Componente Principal (CP2):** Las variables más influyentes son "*bathrooms*", "*room\_type"* y "*neighbourhood\_cleansed*" (todas negativamente). Esto sugiere que la **CP2** representa un factor relacionado con la calidad del alojamiento y la ubicación, donde un mayor número de baños, ciertos tipos de habitaciones y ciertos vecindarios están asociados con valores más bajos de **CP2**. El número máximo de noches y el precio tienen una influencia positiva muy pequeña, indicando un impacto mínimo en el aumento de la **CP2**.

En resumen, la **CP1** está relacionada con la capacidad y el costo del alojamiento, mientras que la **CP2** está más asociada con la calidad del alojamiento y su ubicación.

De esto se pueden sacar varias conclusiones:

* Capacidad y Costo como Factores Clave **(CP1):**
  + Las variables "*accommodates*" y "*beds*" tienen una influencia negativa significativa en la **CP1**, al igual que el "*price*". Esto sugiere que los alojamientos con mayor capacidad (más personas y camas) y precios más altos tienden a estar asociados con valores más bajos de la primera componente principal.
  + Este patrón podría indicar que hay un segmento de mercado donde los alojamientos más grandes y costosos están menos preferidos o se comportan de manera diferente en comparación con otros tipos de alojamientos.
* Influencia del Tipo de Propiedad y Habitación **(CP1):**
  + Las variables "*property\_type*" y "*room\_type*" tienen una influencia positiva moderada en la ***CP1***, lo que sugiere que diferentes tipos de propiedades y habitaciones están asociados con valores más altos de esta componente.
  + Esto podría indicar que la variedad en el tipo de alojamiento juega un papel importante en la preferencia de los usuarios o en la manera en que se distribuyen estos datos.
* Calidad del Alojamiento y Ubicación **(CP2):**
* Las variables "*bathrooms*", "*room\_type*", y "*neighbourhood\_cleansed*" tienen una influencia negativa significativa en la ***CP2***. Esto sugiere que los alojamientos con más baños, ciertos tipos de habitaciones y ubicados en ciertos vecindarios tienden a estar asociados con valores más bajos de esta componente.
* Esto podría reflejar un factor de calidad del alojamiento, donde ciertas características de los alojamientos y su ubicación influyen en cómo se agrupan los datos, posiblemente indicando preferencias de los usuarios o características de mercado específicas.
* Impacto Mínimo de Otras Variables **(CP2):**
  + Las variables como el "*maximum\_nights*" y "*price*" tienen una influencia positiva muy pequeña en la **CP2**, sugiriendo que estos factores tienen un impacto mínimo en esta componente.
  + Esto podría indicar que el número máximo de noches permitidas y el precio no son tan determinantes en la segunda componente principal, quizás porque la calidad del alojamiento y la ubicación tienen un peso mayor en esta dimensión.

Como conclusión, el análisis sugiere que la primera componente principal está dominada por factores de capacidad y costo del alojamiento, mientras que la segunda componente principal está más influenciada por la calidad del alojamiento y la ubicación. Estos resultados pueden ser útiles para segmentar el mercado de alojamientos turísticos no hosteleros y entender mejor las preferencias de los usuarios. Por ejemplo, se podría enfocar en ofrecer diferentes estrategias de marketing o ajuste de precios según las características dominantes de cada segmento.

Finalmente, en relación al modelo Logit calculado también se pueden extraer diversas conclusiones y posibles utilidades del mismo.

El modelo logit desarrollado para identificar alojamientos con precios altos muestra una capacidad razonable de discriminación, con un área bajo la curva ROC (AUC) de 0.77. Este valor indica que el modelo tiene una habilidad moderada para distinguir entre alojamientos con precios altos y bajos. La exactitud general del modelo es buena, con valores similares en el conjunto de prueba (83.13%) y el conjunto de entrenamiento (82.73%), lo que sugiere que el modelo generaliza bien entre los diferentes conjuntos de datos.

El análisis de las variables significativas revela que ciertas características tienen un impacto positivo en la probabilidad de que un alojamiento tenga un precio alto. Las variables como el número de plazas disponibles ("*accommodates*"), el número de baños ("*bathrooms*") y el tipo de propiedad ("*property\_type*") están asociadas con un aumento en la probabilidad de precios altos. De estas, el número de baños es la variable más influyente, destacando su importancia en la evaluación del precio de los alojamientos, es decir, que el número de baños de un alojamiento es el factor más importante en el mercado de alquiler de alojamientos turísticos no hosteleros de Málaga de la plataforma Airbnb.

Por otro lado, variables como la verificación de identidad del anfitrión ("*host\_identity\_verified*"), el tipo de habitación ("*room\_type*") y el número mínimo de noches ("*minimum\_nights*") tienen coeficientes negativos. Esto sugiere que, aunque estas características pueden ser importantes, tienden a estar asociadas con una menor probabilidad de que el precio del alojamiento sea alto. En particular, el tipo de habitación ("*room\_type*") tiene el coeficiente negativo más significativo, lo que indica que su impacto es considerable en la determinación de si el precio por noche de un alojamiento es excesivo o no, por tanto, el tipo de habitación anunciada es determinante también para determinar si el precio es excesivo, o no.

El modelo no está exento de limitaciones. A pesar del buen desempeño general, el modelo presenta desafíos en términos de sensibilidad, que es relativamente baja (27.57% en el conjunto de prueba y 24.22% en el conjunto de entrenamiento). Esto significa que el modelo tiene dificultades para identificar correctamente todos los casos positivos (alojamientos con precios altos), lo que puede ser problemático en situaciones donde es crucial no pasar por alto estos casos. La alta especificidad (95% en el conjunto de prueba y 96.11% en el conjunto de entrenamiento) sugiere que el modelo es muy bueno para identificar los casos negativos, pero su capacidad para detectar casos positivos necesita mejora.

Para acabar, se van a presentar algunos posibles ejemplos de utilidad del modelo Logit calculado, para mostrar su utilidad a pesar de las limitaciones previamente expuestas:

* **Optimización de Estrategias de Precios:** Las plataformas de alquiler de alojamientos pueden utilizar este modelo para identificar propiedades con potencial para precios más altos. Esto puede ayudar a los anfitriones a ajustar sus precios de manera más competitiva y maximizar sus ingresos, basándose en las características significativas identificadas por el modelo, y por el contrario, desde el punto de vista del cliente, puede utilizar este modelo para comprobar si el precio por noche de un alojamiento de su interés es excesivo o no, ayudando en la toma de decisión de alquilar el alojamiento.
* **Segmentación de Mercado:** Las agencias de marketing y las plataformas de alquiler pueden segmentar mejor sus mercados utilizando las predicciones del modelo. Por ejemplo, pueden dirigir campañas de marketing específicas a propiedades con características que se correlacionan positivamente con precios altos, como un mayor número de baños y tipos específicos de propiedades.
* **Evaluación de Nuevas Propiedades:** Los nuevos anfitriones pueden usar el modelo para evaluar el potencial de sus propiedades en términos de precio. Al entender qué características aumentan la probabilidad de precios altos, los anfitriones pueden invertir en mejoras específicas, como aumentar el número de baños o cambiar el tipo de propiedad, para aumentar el valor percibido de su alojamiento.
* **Mejoras en la Plataforma:** Las plataformas pueden utilizar los resultados del modelo para mejorar sus algoritmos de recomendación y búsqueda, destacando propiedades que tienen una alta probabilidad de ser listadas a precios más altos. Esto puede mejorar la experiencia del usuario al proporcionar recomendaciones más precisas y relevantes.

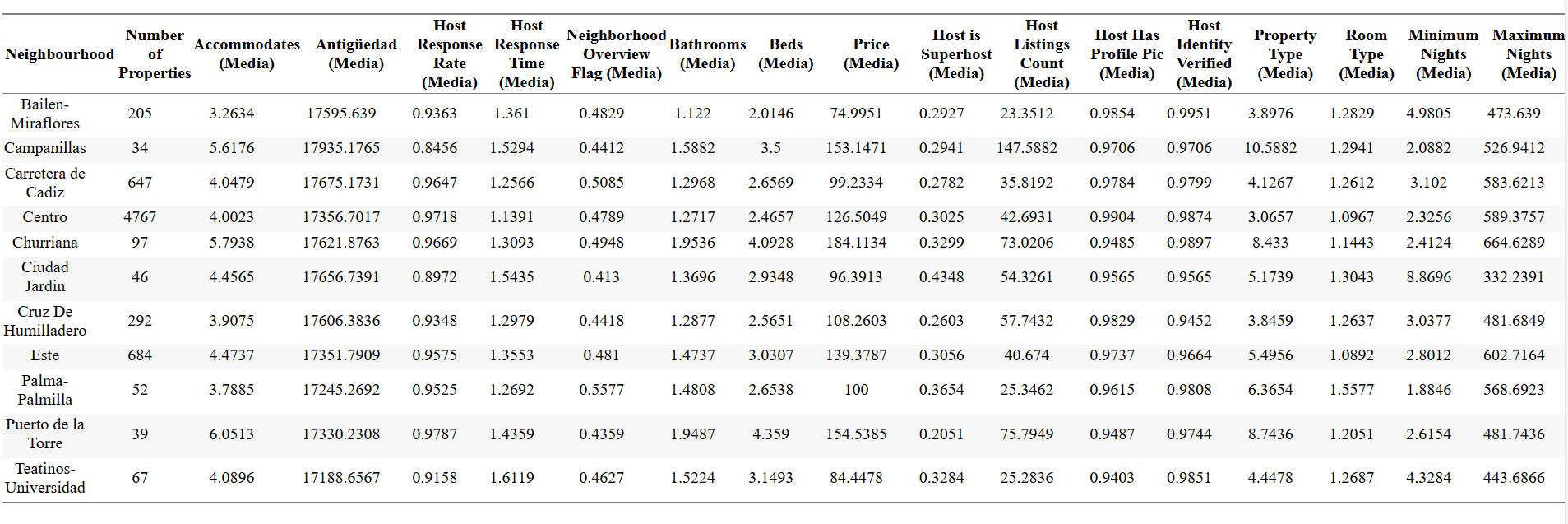
En resumen, aunque el modelo logit presenta ciertas limitaciones en su capacidad para detectar todos los casos de precios altos, su buen desempeño general y su capacidad para identificar características significativas lo hacen una herramienta valiosa para la optimización de precios, segmentación de mercado y evaluación de nuevas propiedades en el sector de alquiler de alojamientos.

## Anexo I – Tratamiento de las variables

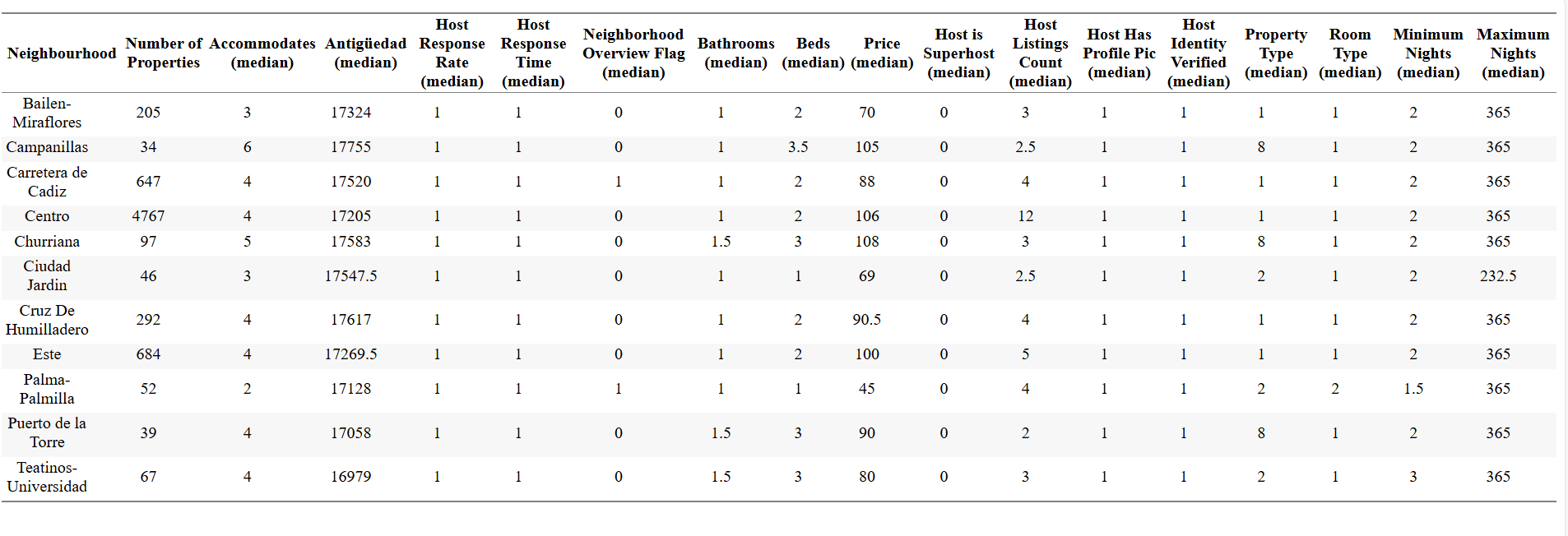
* La variable *host\_response\_time* se dividía en 4 categorías, por lo que a cada una de ellas se le asignó un valor numérico:
* Within an hour: 1
* Within a few hours: 2
* Within a day: 3
* A few days or more: 4
* Las variables *host\_is\_superhost*, *host\_has\_profile\_pic* y *host\_identify\_verified* 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

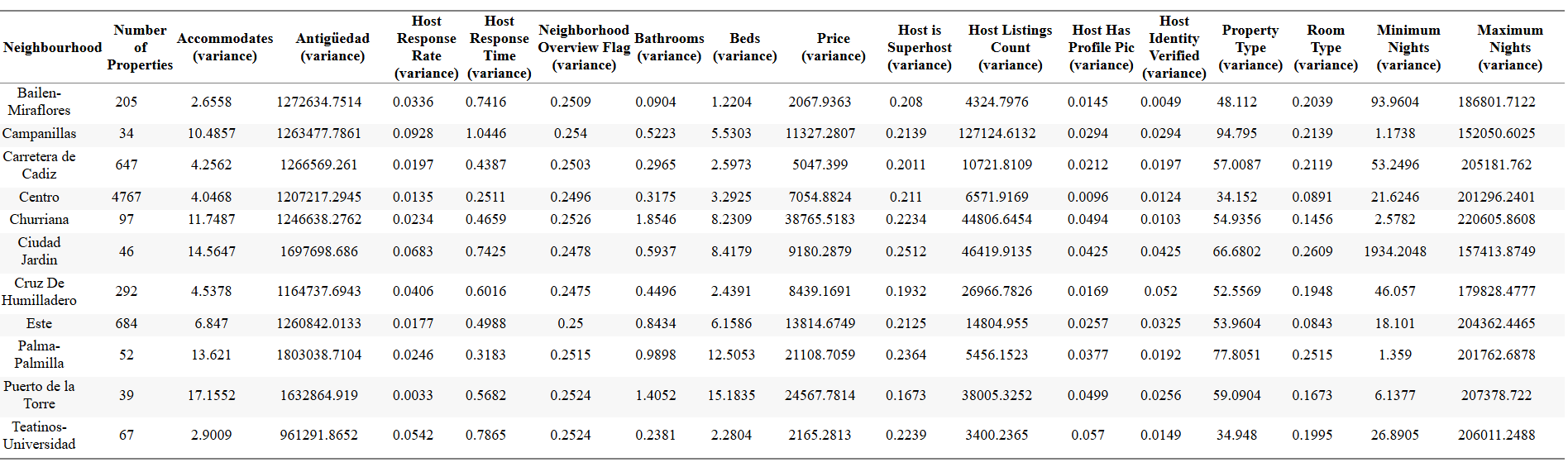
Medias de las variables por barrio



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. Las cargas para el resto de componentes pueden encontrarse en el documentos *resultados\_pca.txt* disponible en https://github.com/Joseluisvic/Trabajo-Final-De-M-ster [↑](#footnote-ref-2)
3. 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-3)
4. El archivo de este modelo desglosado junto a las métricas de evaluación se encuentran con los nombres **logit\_model\_evaluation.txt** y **logit\_model\_summary.txt** en el siguiente enlace <https://github.com/Joseluisvic/Trabajo-Final-De-M-ster> [↑](#footnote-ref-4)