Materia: Taller de Programación
Grupo 2: Jimena Teran, Juan Lynch, Renzo Falciglia
Fecha: 11/11/2024

Trabajo Práctico 2

Parte I: Limpieza de la base

En esta primera parte realizamos un análisis de la base de datos, de sus variables y características. Luego, realizamos una "limpieza" de los datos, de tal manera de prepararlos para poder realizar posteriormente las estimaciones necesarias. En cada bloque de código en el archivo .ipynb se explica el análisis o las transformaciones que le realizamos a los datos.

Luego del análisis descriptivo, se eliminan los valores duplicados y los valores nulos y NAs, para que no afecten el análisis. A su vez, como lo solicita la consigna, trabajamos con las variables Neighbourhood_group y con room_type. Estas dos variables son categóricas y describen el grupo en el que se encuentra el barrio y el "tipo" de habitación que tiene el alojamiento. A partir de las transformaciones realizadas, estas variables categóricas se codifican como numéricas.

En el caso de los NA'S de la variable reviews_per_month, tomamos la decisión de reemplazarlos por la mediana agrupando por neighbourhood_group_encoded. Suponemos que los alojamientos dentro de un mismo barrio tendrán precios similares entre sí.

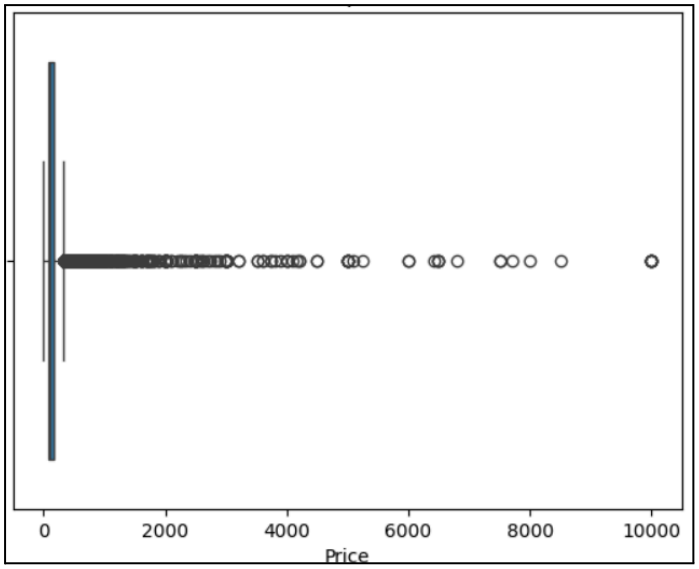
También eliminamos las columnas que no son de interés para el análisis que realizamos. Específicamente, se eliminas las columnas de 'id', 'name', 'neighbourhood_group','room_type', 'last_review', 'host_name' y 'availability_365'.

Por último, analizamos la distribución de nuestra variable de interés price, y detectamos que hay pocas observaciones con un precio mayor a USD 6.000, por lo que decidimos excluir estas observaciones del análisis. El data frame quedó con la siguiente forma:

Tabla 1 - Forma del data frame

	host_id	neighbourhood	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	neighbourhood_group_encoded	room_type_encoded
9151	20582832	Astoria	40.76810	-73.91651	10000.0	100	2	0.04	1	3	1
17692	5143901	Greenpoint	40.73260	-73.95739	10000.0	5	5	0.16	1	1	0
29238	72390391	Upper West Side	40.77213	-73.98665	10000.0	30	0	0.61	1	2	0
12342	3906464	Lower East Side	40.71355	-73.98507	9999.0	99	6	0.14	1	2	1
6530	1235070	East Harlem	40.79264	-73.93898	9999.0	5	1	0.02	1	2	0
40433	4382127	Lower East Side	40.71980	-73.98566	9999.0	30	0	0.61	1	2	0
30268	18128455	Tribeca	40.72197	-74.00633	8500.0	30	2	0.18	1	2	0
4377	1177497	Clinton Hill	40.69137	-73.96723	8000.0	1	1	0.03	11	1	0
29662	156158778	Upper East Side	40.76824	-73.95989	7703.0	1	0	0.61	12	2	0
45666	262534951	East Flatbush	40.65724	-73.92450	7500.0	1	8	6.15	2	1	1

Gráfico 1 - Boxplot variable Price

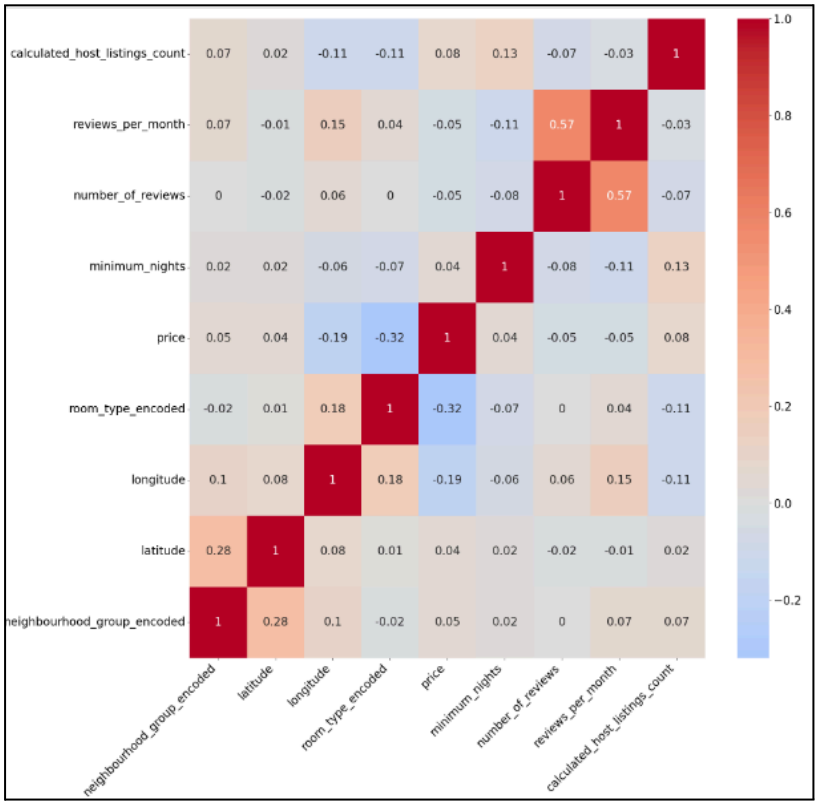


Parte II: Limpieza de la base

En esta parte realizamos un conjunto de gráficos y visualizaciones que nos ayudan a entender los datos y a prepararnos para la estimación posterior.

Como primer paso, realizamos una matriz de correlación con las siguientes variables ‘neighbourhood group’, ‘latitude’, ‘longitude’, ‘room type’,‘price’, ‘minimum nights’, ‘number of reviews’, ‘reviews per month’,‘calculated host listings count’.

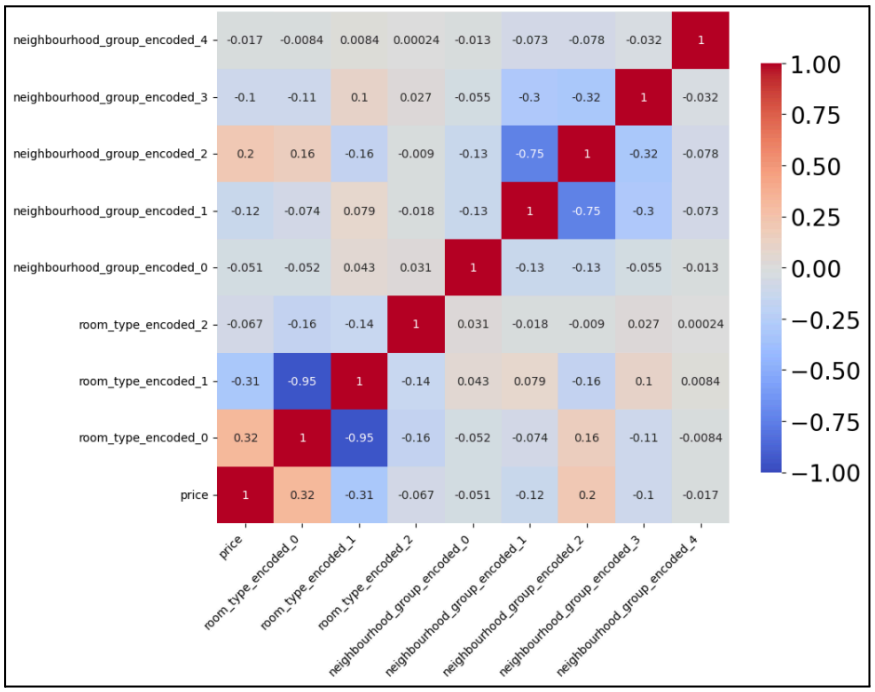
Gráfico 2 - Matriz de correlación



Los resultados de esta matriz no sugieren correlaciones muy fuertes entre la variable price y el resto de las variables. Sin embargo, las variables 'neighbourhood group', 'room type' fueron incluídas de manera ordinal, de acuerdo a los números que se le asignó a cada tipo de barrio y a cada tipo de alojamiento. Esto implica que no se pueda interpretar correctamente la correlación con la variable price.

Por lo tanto, para solucionar este problema y observar cómo se correlaciona la variable de tipo de barrio y de tipo de alojamiento, se utiliza la técnica de *one hot encoding*, que asigna variables binarias a cada una de las categorías de ambas variables. Luego, se realiza una matríz de correlación de la variable price junto a neighbourhood group y a room type.

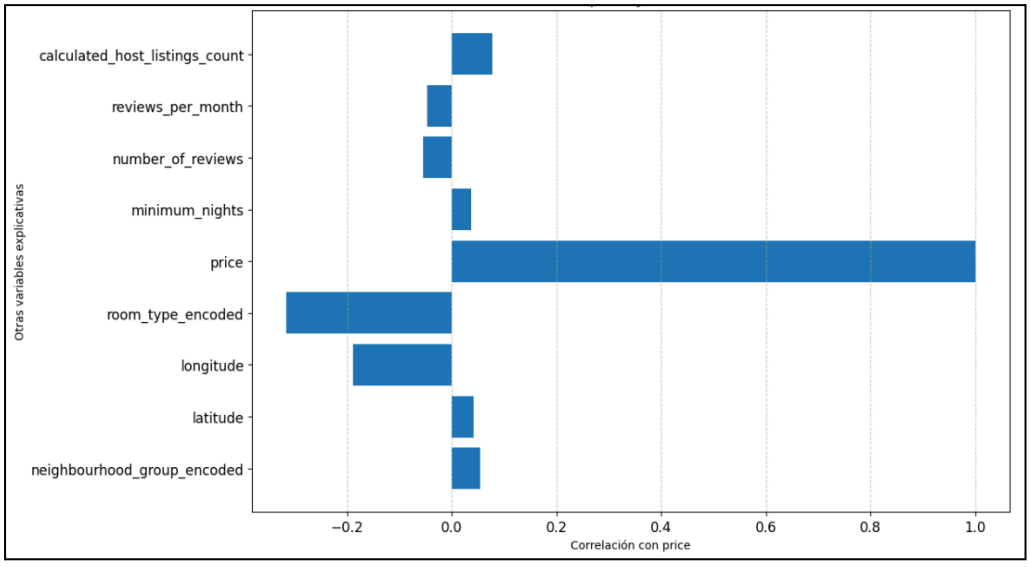
Gráfico 3 - Matriz de correlación con One Hot Encoding



Con esta segunda matriz puede interpretarse mejor la correlación con el barrio y el tipo de habitación. Por ejemplo, se observa que tienen una correlación fuerte y positiva con el precio cuando el alojamiento es un departamento entero y cuando está ubicado en Manhattan, lo que resulta intuitivo ya que es uno de los barrios más buscados por los turistas. A su vez, se observa una correlación negativa entre el precio cuando el alojamiento se trata de una habitación dentro de una casa o departamento y cuando está ubicado en Brooklyn.

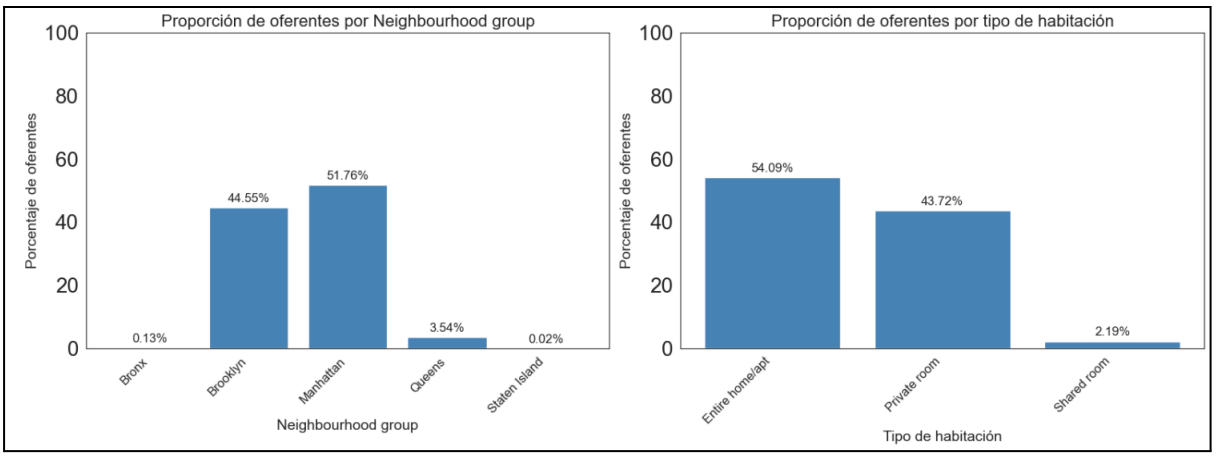
A continuación, realizamos otros gráficos de correlación para evaluar a la variable price.

Gráfico 4 - Correlación variable precio



Ahora realizamos un análisis para responder a las siguientes preguntas: ¿Cuál es la proporción de oferentes por “Neighbourhood group”? ¿Y por tipo de habitación?

Gráfico 5 - Proporciones



Cómo se desprende de los gráficos, casi el 52% de los oferentes están ubicados en Manhattan, mientras que el 44% está ubicado en Brooklyn. Por otro lado, el 54% de los alojamientos son del tipo departamento entero, mientras que casi el 44% es una habitación privada dentro de un departamento más grande.

A continuación se realiza un histograma y una función de kernel sobre los precios de los alojamientos. En cuanto a la configuración del histograma, elegimos filtrar valoren mayores a 2000 porque creemos que ya se entiende la tendencia en la frecuencia para valores altos. En cuanto a los bins, elegimos 100 porque de esta manera se muestra un gráfico claro.

En cuanto a la función de Kernel, usamos la misma regla de precios menores a 2000. Usamos la función de kernel Gaussiano que viene por default, ya que probamos con un epanechnikov pero no notamos gran diferencia. A su vez, utilizamos el ancho de banda que viene en Seaborn de forma predeterminada (calculado mediante una regla que minimiza el Error Cuadrático Medio):

Gráfico 6 - Histograma

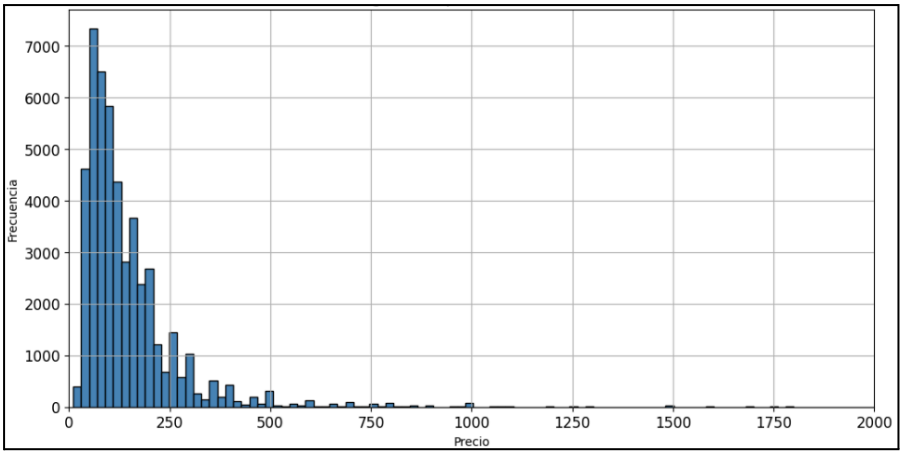
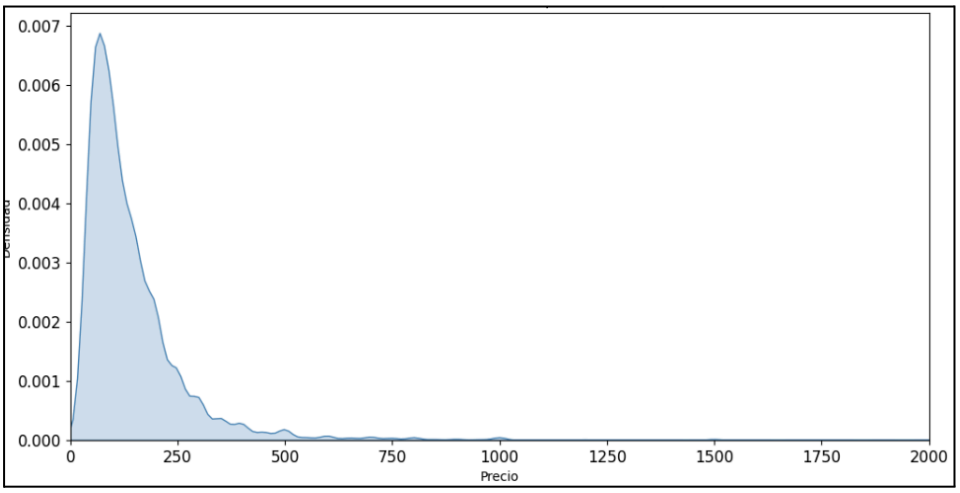


Gráfico 6 - Kernel



A continuación, realizamos los cálculos requeridos para responder las siguientes preguntas: ¿cuál es el precio mínimo, máximo y promedio? ¿Cuál es la media de precio por "Neighbourhood group" y por tipo de habitación? y mostramos las respuestas:

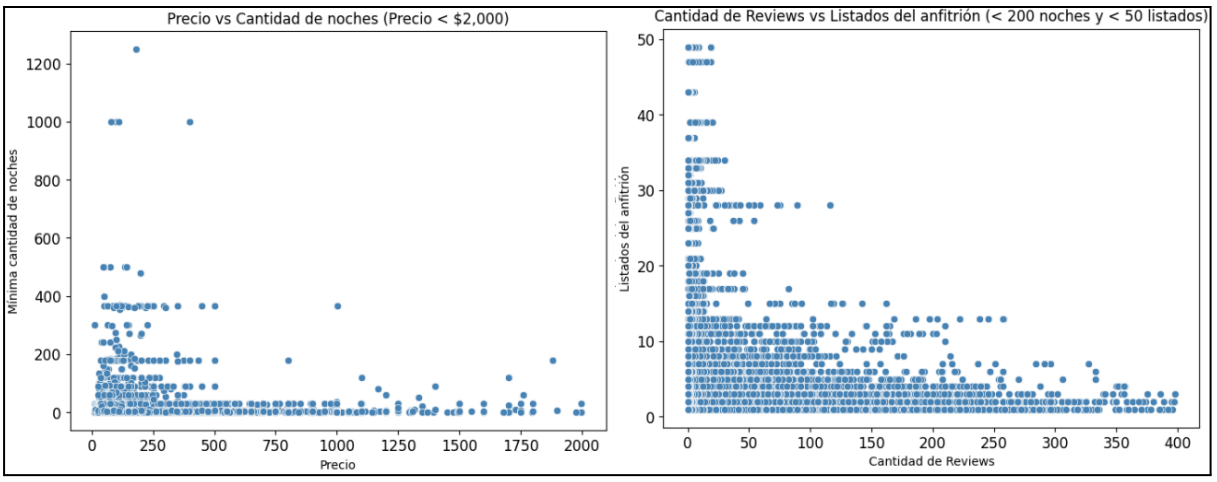
Tabla 2 - Precios por tipo de barrio y por tipo de alojamiento

```
Precio mínimo: 0.00
Precio máximo: 10000.00
Precio promedio: 152.73

Precio medio por 'Neighbourhood group':
neighbourhood_group_encoded
0     87.464646
1    124.380597
2    196.862352
3     99.536900
4    114.812332
Name: price, dtype: float64

Precio medio por tipo de habitación:
room_type_encoded
0    211.788107
1     89.783388
2     70.127586
Name: price, dtype: float64
```

De los resultados se desprende que el valor medio más alto se da en Manhattan, con USD 197 por alojamiento. A su vez, el precio medio más alto es en aquellos alojamientos que se ofrecen enteros, con un promedio de USD 212.

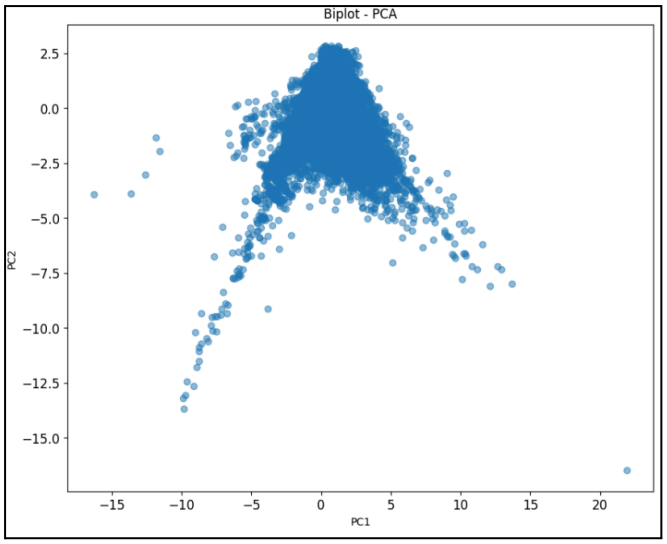
A continuación, se realizan dos *scatterplots* con variables de interés. En el primero, se muestra el precio con la cantidad de noche mínima que requieren los oferentes. En el segundo con la cantidad de anuncios publicados por el oferente.

Gráfico 7 - Scatterplots



Cómo último paso de esta sección 2, se realiza un análisis de componentes principales para graficar las ponderaciones (o loadings) de las variables en dos dimensiones (biplot con flechas).

Gráfico 8 - Biplot PCA



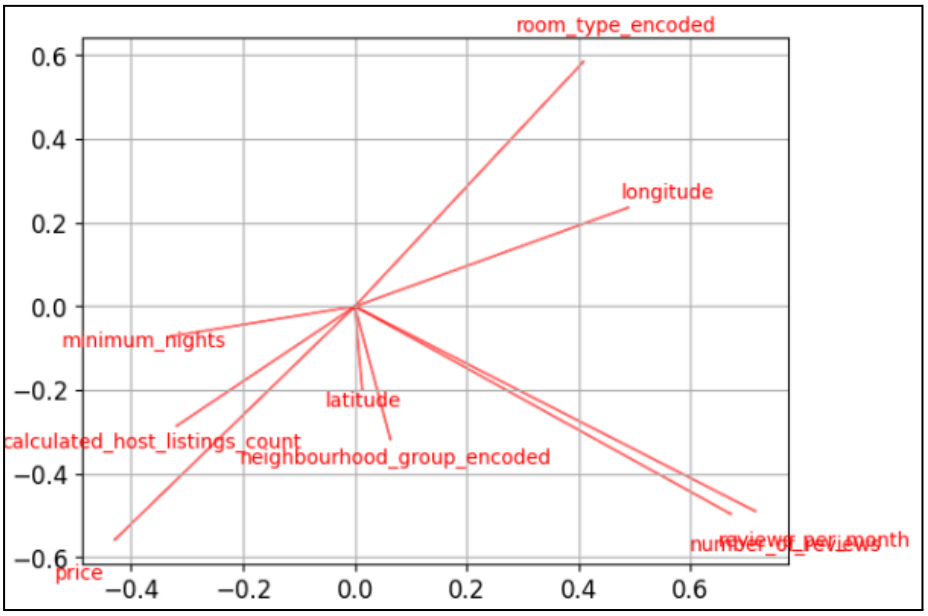
Cómo se desprende del gráfico de componentes principales, la mayoría de los puntos se concentran en el centro del gráfico, formando un patrón triangular o de abanico. El componente principal 1 (PC1) tiene un rango aproximadamente entre -15 y +20, mientras que PC2 varía de -15 a +2.5. Esto sugiere que PC1 capta una mayor parte de la variabilidad que PC2. El hecho de que PC1 tenga un rango mayor sugiere que este componente captura más información de la estructura de los datos originales.

A su vez, la forma triangular o de abanico de la distribución puede indicar que los datos están estructurados de una manera específica y que las variables originales están correlacionadas de forma compleja. Puede ser que haya subgrupos o clústeres en los datos, con una gran densidad en ciertas áreas y una menor densidad en los extremos. Estos

puntos más extremos pueden representar valores atípicos o grupos de observaciones con características distintas.

A continuación se calculan los loadings o ponderaciones y se grafica.

Gráfico 9 - Ponderaciones PCA



Como se desprende de este último gráfico, PC1 parece estar influenciado por variables como price y number_of_reviews, ya que estas variables tienen componentes significativos en la dirección de PC1 (horizontal).

PC2 está influenciado por variables como room_type_encoded, que tiene una gran componente en la dirección vertical.

Parte III: Predicción y Validación

En esta sección se realizan transformaciones necesarias a la base para realizar la predicción del precio de los alojamientos. Para ello, se divide la base en una base de prueba (test) y una base de entrenamiento (train), de la siguiente manera:

- Dimensiones de X_train: (34195, 13)
- Dimensiones de X_test: (14656, 13)
- Dimensiones de y_train: (34195,)
- Dimensiones de y_test: (14656,)

Con estas transformaciones realizadas se implementa una regresión final con los siguientes resultados:

1. Variables Principales:

- `latitude` (-6.32): indica que las propiedades ubicadas más al sur tienen precios ligeramente más bajos.

- `longitude` (-1.19): Las propiedades más al oeste tienden a tener precios más bajos, aunque el impacto es moderado.
- `minimum_nights` (-2.07): aumentar el número mínimo de noches tiene un efecto negativo en el precio, aunque el impacto es pequeño.
- `last_review` (por ejemplo, `last_review_2019-07-06` con -2.099 y `last_review_2019-07-08` con -4.857): algunas fechas de la última reseña tienen un efecto leve en el precio.

El modelo muestra que la ubicación (latitud y longitud) y algunas características relacionadas con las reseñas influyen en el precio de las propiedades. No obstante, muchas variables tienen un impacto limitado, lo que sugiere que simplificar el modelo eliminando algunas variables podría mejorar su interpretabilidad.

A continuación se calcula el Error Cuadrático Medio (MSE), la Raíz del Error Cuadrático Medio (RMSE), y el Error Absoluto Medio (MAE) en la base de entrenamiento y testeo (usando coeficientes estimados con la base de entrenamiento). Los resultados se muestran en la siguiente tabla:

Tabla 3 - MSE, RMSE y MAE

Métrica	**Entrenamiento**	**Prueba**
MSE	2.081071e-08	101647.513752
RMSE	1.442592e-04	318.822072
MAE	1.011099e-04	88.250354

Las métricas de error muestran una gran diferencia entre el ajuste del modelo en los datos de entrenamiento y prueba. **MSE, RMSE** y **MAE** son extremadamente bajos en el conjunto de entrenamiento, indicando un ajuste casi perfecto dentro de la muestra. Sin embargo, en el conjunto de prueba, estos valores son mucho más altos, lo cual sugiere que el modelo no generaliza bien a datos nuevos.

La discrepancia entre las métricas de entrenamiento y prueba indica que el modelo podría estar sobreajustado (overfitting). Aunque el modelo funciona bien dentro de la muestra, su rendimiento fuera de la muestra es deficiente, por lo que sería recomendable simplificar el modelo o aplicar regularización para mejorar su capacidad de generalización.