Trabajo Practico N°2

Lucila Brisighelli y Maria Constanza Cerundolo

Fecha de entrega: 21 de abril de 2024

1 Limpieza de Datos

Tomando el articulo de 'Missing-data imputation" tratamos los valores faltantes con la *mean imputation*. En el articulo nos explican las implicancias de considerar esta estrategia. Por un lado, es la manera más fácil de tratar los valores faltantes, simplemente reemplazamos cada uno de los valores que faltan con la media de los valores observados. Sin embargo, esta estrategia puede ser un poco agresiva ya que ditorsiona la distribución de cada variable, particularmente realiza subestimaciones de la desviación estándar. Esto sucede especialmente si nuestros datos tienen muchos valores atípicos (outliers).

Para tratar los outliers utilizamos el rango intercuartílico (IQR) y luego los establecimos como NaN. Primero, calculamos los cuartiles Q1 (primer cuartil) y Q3 (tercer cuartil) para cada variable numérica en el conjunto de datos. Estos cuartiles se dividen los datos en cuatro partes. Para calcular el rango distinguimos el primer cuartil que es el valor por debajo del cual cae el 25% de los datos, y el tercer cuartil que es el valor por debajo del cual cae el 75% de los datos. Luego calculamos el rango intercuartílico (IQR), que es la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1). Este rango es una medida de dispersión robusta que se utiliza para identificar los valores que se desvían significativamente de la "tendencia" de los datos. Para ello, definimos el umbral para identificar outliers, siendo aquellos valores que están por debajo de Q1 - 1.5 * IQR o por encima de Q3 + 1.5 * IQR. Creamos *outlier_mask*, un booleano que identifica los outliers en el conjunto de datos, comparando cada valor numérico con los umbrales. Por ultimo, actualizamos *airbnb_data*, estableciendo como NaN aquellos valores que son identificados como outliers.

2 Gráficos y visualizaciones

1. La matriz de correlación impresa puede verse impresa en el jupyter notebook.

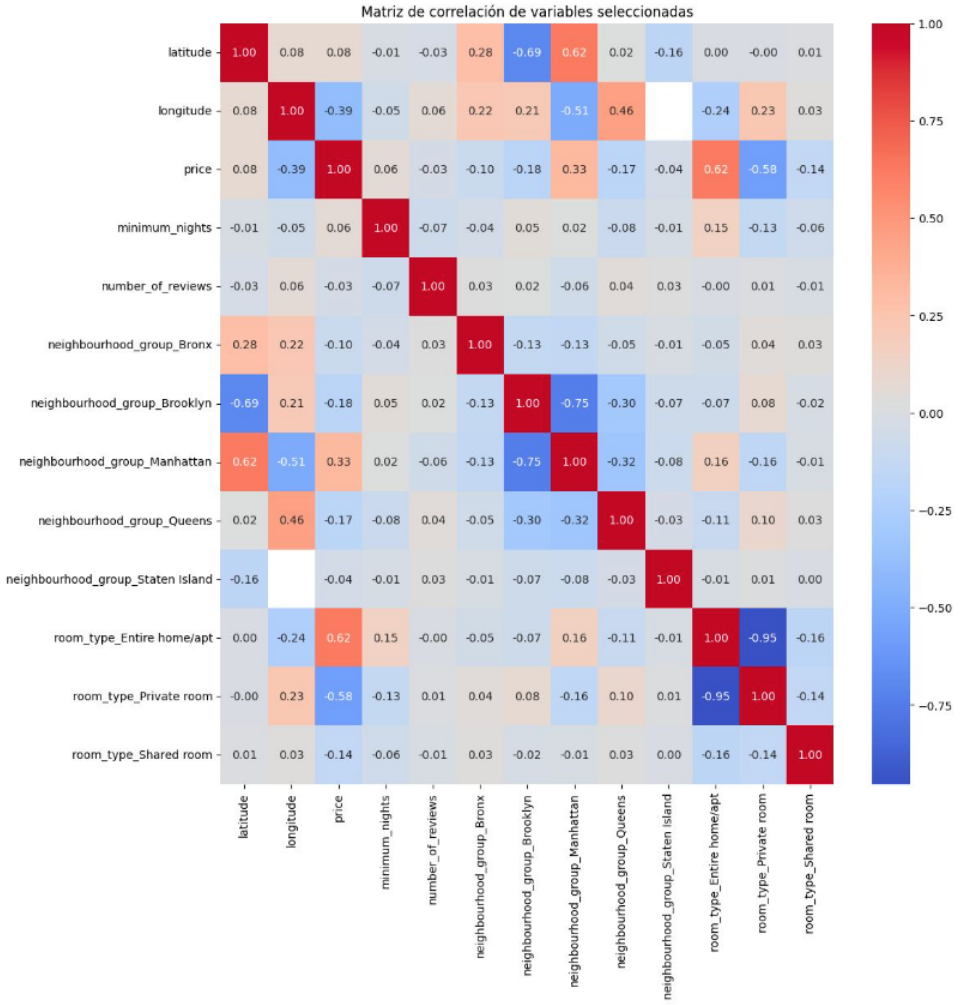

Figure 1: Mapa de color de la matriz de correlación.

2. La proporción de oferentes por neighbourhood group está dada de la siguiente manera: Un 44,3% corresponde a Manhattan mientras que el porcentaje restante se divide en un 41,1% en Brooklyn, 11,6% en Queens, 2,2% en Bronx y un 0,8% en Staten Island. Estos valores pueden verse en el siguiente gráfico de tortas también

Por su parte, la proporción de oferentes por tipo de habitación se compone de la siguiente manera. Un 52% corresponde a entire home/ apartment, un 45,7% a private room y el 2,4% restante a shared rooms. Esta distribución puede verse representada a partir del siguiente gráfico de tortas.

Obtenidos estos valores, podría pensarse que si cruzamos los datos la mayor oferta corresponde a departamentos en Manhattan. Esto también se ve en el mapa de calor de la matriz de correlación (0,12).

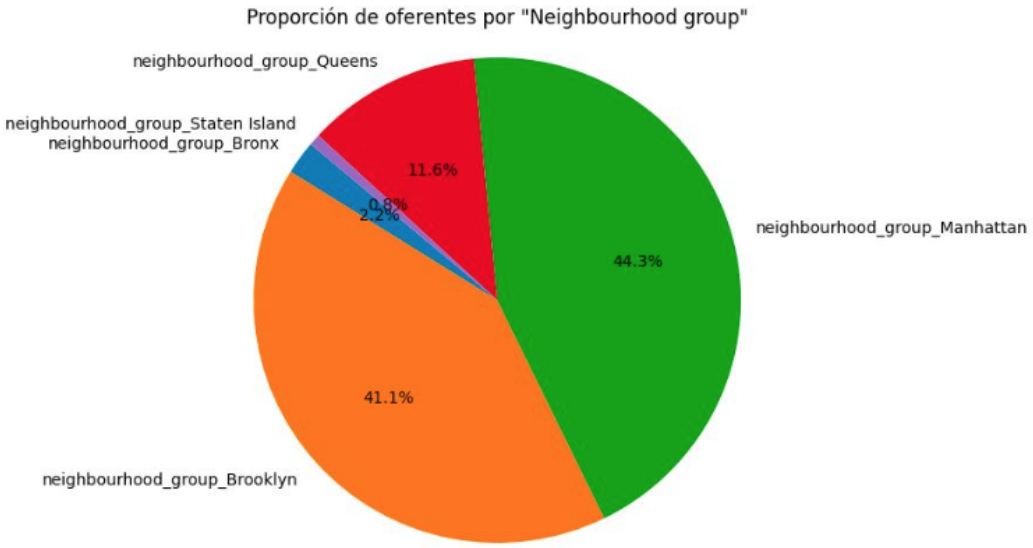

Figure 2: Proporción de oferentes por 'neighbourhood group'

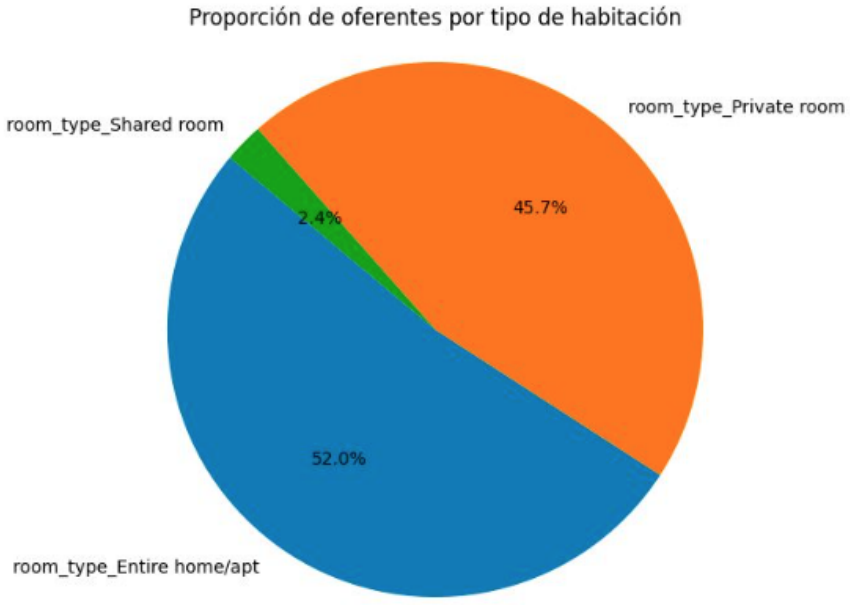

Figure 3: Proporción de oferentes por tipo de habitación

3.Histogramas de precios de alojamiento A simple vista con este gráfico podemos ver que a un precio de entre 90 y 100, aproximadamente, es en donde más demanda habría,

ya que la frecuencia es poco más de 500. La mayo frecuencia está entre los precios 50 y 100. A partir de 100 puede verse una tendencia bajista ya que la frecuencia disminuye en la medida que los precios aumentan, exceptuando ciertos picos puntuales en los precios 150, 200, 250 y 300 aproximadamente. El precio mínimo es 0, el precio máximo es 334 y el precio promedio es de 119.98251381876223. La media de precio por neighbourhood group es, para cada grupo:

- *neighbourhood_group_Bronx*: 75.963762
- *neighbourhood_group_Brooklyn*: 102.095211
- *neighbourhood_group_Manhattan*: 131.427031
- *neighbourhood_group_Queens*: 87.389069

Por su parte, la media de precios por tipo de habitación, para cada habitación, es:

- *room_type_Entirehome/apt*: 145.762236
- *room_type_Privateroom*: 77.882662
- *room_type_Sharedroom*: 58.168966

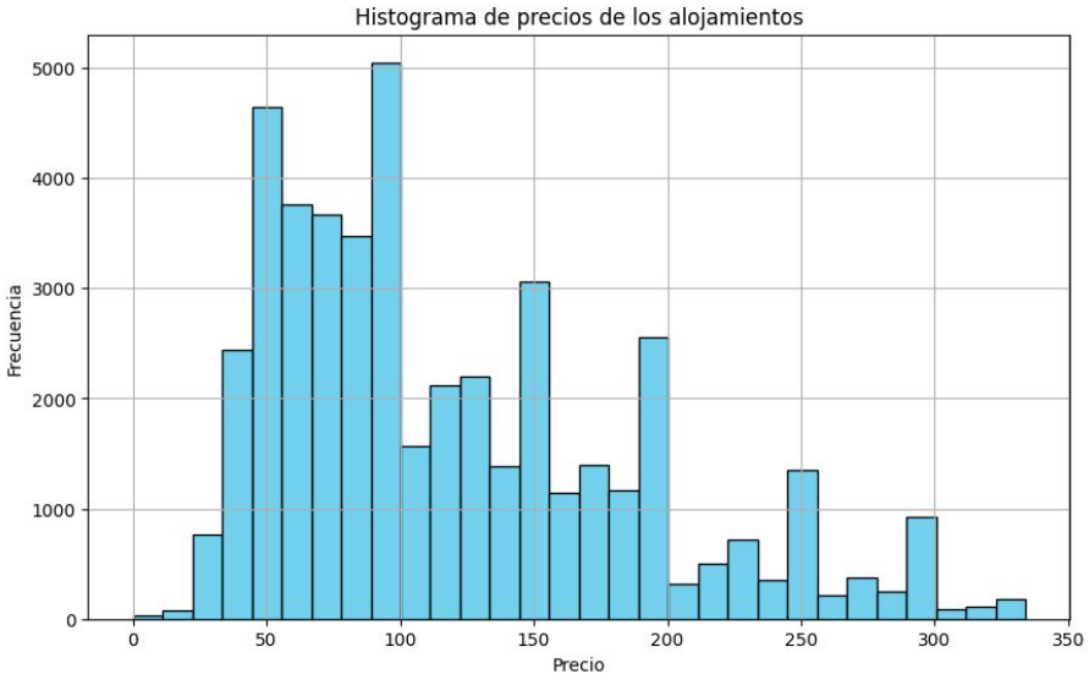

Figure 4: Histograma de precios de los alojamientos

4.Scatter Plot En primer lugar realizamos un scatter plot en donde las variables de interés son los precios y el número de reseñas. Precisamente, este gráfico de dispersión muestra la relación entre el precio y el número de reservas de alojamientos. A simple vista, se pueden observar algunas tendencias interesantes. Por un lado, la mayor concentración de puntos se encuentra en el rango de precios más bajos, lo que indica que la mayoría de

los alojamientos/ofertas a precios más económicos y reciben un alto número de reservas. Así, a medida que el precio aumenta, el número de reservas tiende a disminuir, lo que sugiere una relación inversamente proporcional entre ambas variables. Sin embargo, aún se pueden observar algunos puntos dispersos con precios más altos y un número moderado de reservas. Por otro lado, hay una cantidad considerable de puntos con precios muy bajos (cercanos a cero) y un número relativamente bajo de reservas, lo que podría indicar ofertas nuevas, con poca popularidad o con alguna característica particular que aún no genera tracción en el mercado. No se aprecian agrupaciones o patrones definidos más allá de la tendencia general mencionada, lo que sugiere una amplia variabilidad en los datos. El gráfico de dispersión no muestra una correlación perfecta, sino más bien una tendencia general, lo que implica que existen otros factores además del precio que influyen en el número de reservas, como por ejemplo la zona, el tipo de habitación, entre otros ya mencionados en puntos anteriores. En resumen, este scatter plot revela una relación inversamente proporcional general entre el precio y el número de reservas, con una mayor concentración de datos en el rango de precios más bajos y una dispersión más amplia a medida que el precio aumenta.

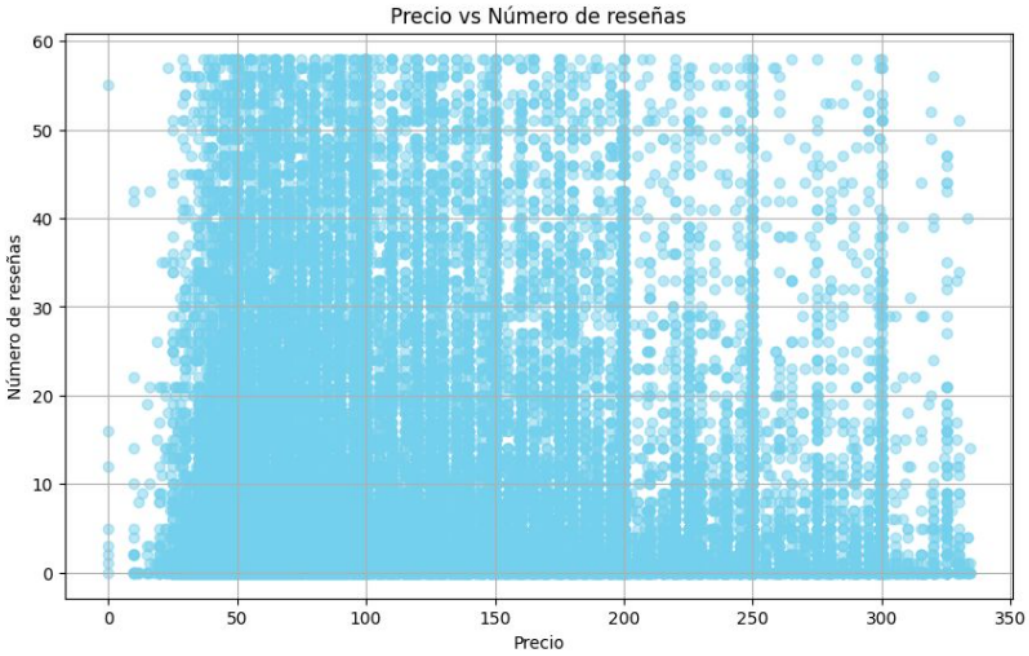

Figure 5: Scatter Plot 1

En segundo lugar realizamos un scatter plot en donde las variables de interés son los precios y el mínimo de noche por reseña. En este caso se observa la relación entre el precio y el mínimo de noches requeridas para reservas de alojamientos. Se puede ver que hay una alta concentración de puntos en el rango de precios más bajos (alrededor de 0 a 100), lo que sugiere que la mayoría de las ofertas económicas tienen un requisito de una noche de estadía mínima. A medida que el precio aumenta, se observa una mayor dispersión en el mínimo de noches requeridas. Hay algunas ofertas con precios más altos que exigen sólo una noche, mientras que otras solicitan estancias más prolongadas. Además, se distinguen

varias líneas horizontales o agrupaciones de puntos, lo que indica que existen requisitos comunes de mínimo de noches para ciertas ofertas o rangos de precios. Por ejemplo, se aprecia una línea horizontal prominente en 7 noches mínimas para un rango de precios. Aunque la tendencia general muestra una mayor dispersión a precios más altos, no hay una correlación perfecta entre el precio y el mínimo de noches. Existen ofertas de precios elevados que aceptan estancias cortas y viceversa. La mayoría de las ofertas parecen requerir un mínimo de noches entre 1 y 10, con algunas excepciones más allá de ese rango.

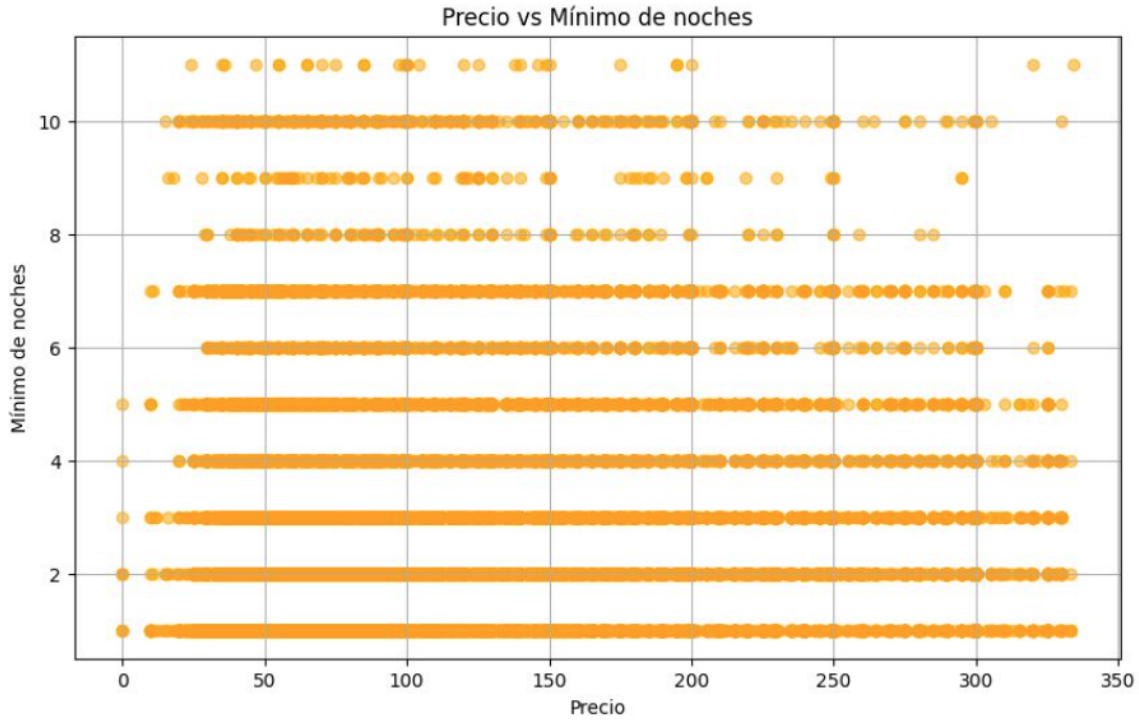

Figure 6: Scatter Plot 2

5.Análisis de Componentes Principales

El PCA es una técnica utilizada para reducir la dimensionalidad de los datos, conservando la mayor cantidad posible de la varianza presente. Observando el gráfico, pudimos hacer los siguientes comentarios: Obtuvimos que el porcentaje de varianza explicada por los dos primeros componentes principales es de [0.36671598 ; 0.32306442]. Logran explicar una porción significativa de la varianza total en los datos, lo que se deduce de la alta concentración de puntos en la región central del gráfico. Por su parte, los loadings de los componentes principales fueron [0.51907583 ; 0.63817926 -0.56858378] [0.7838123 ; -0.09010895 ; 0.61442547]. Las variables originales se combinan de manera diferente para formar las dos componentes principales visualizadas. Se observa un patrón distintivo en forma de "caparazón" o "herradura" en la distribución de los puntos. Este tipo de patrón suele ser indicativo de una estructura de agrupamiento o clustering subyacente en los datos, donde los puntos cercanos entre sí podrían representar observaciones similares o pertenecientes a un mismo grupo. Aunque la mayoría de los puntos se concentran en la región central, hay algunos puntos dispersos que podrían considerarse valores atípicos o outliers. Estos puntos podrían requerir un análisis adicional para comprender su naturaleza y determinar si deben ser tratados de manera especial.

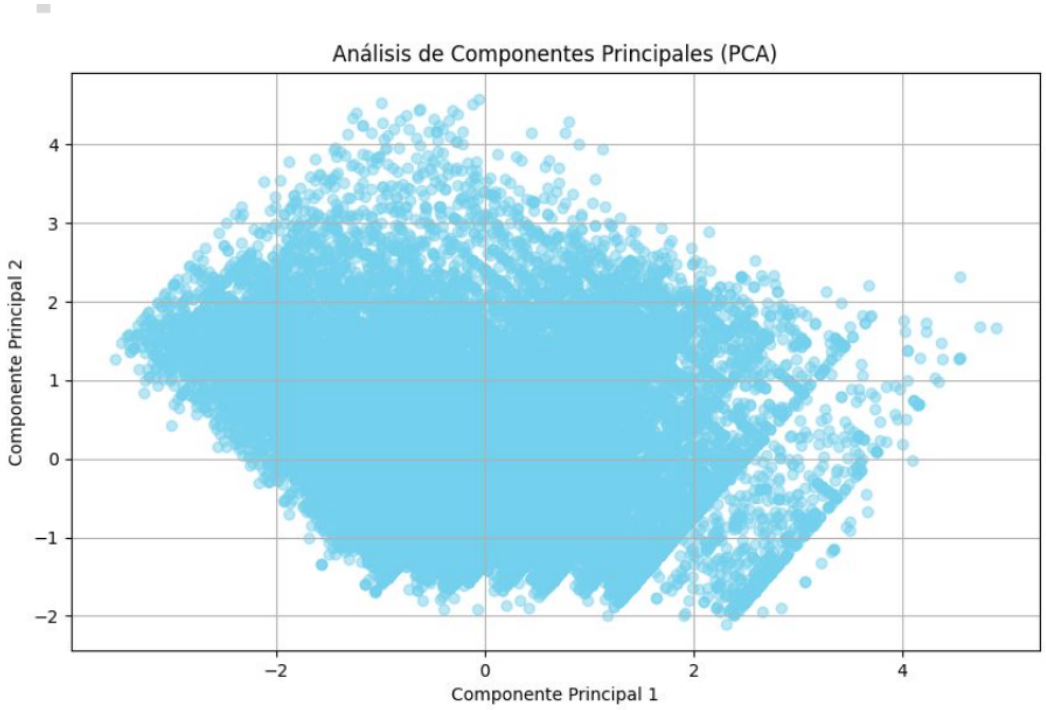

Figure 7: Análisis de Componentes Principales

3 Predicción

El objetivo de esta parte es desarrollar un modelo de regresión linea que pueda proporcionar estimaciones precisas de los precios de los alojamientos. Primeramente eliminamos todas las variables relacionadas con el precio de la base de datos.Luego, dividimoss el conjunto de datos en una base de entrenamiento y en una de prueba utilizando el método *train_test_split*. La base de entrenamiento comprendía el 70% de los datos, la semilla aleatoria (random state instance) fue de 201 para garantizar la reproducibilidad de los resultados. La variable dependiente en la base de entrenamiento es el precio (price), mientras que el resto de las variables se consideraron como variables independientes.

En una segunda parte, implementamos un modelo de regresión lineal utilizando la biblioteca $scikit-learn$. Se agregó una columna de unos (1) a las características para modelar el intercepto en la regresión. Luego, se ajustó el modelo a los datos de entrenamiento y se evaluaron los resultados obtenidos. El modelo de regresión lineal proporcionó los siguientes coeficientes:

El intercepto del modelo fue de -68848.503, lo cual no tiene una interpretación directa en este contexto (inclusión de la columna de unos en las características).

Para el coeficiente con respecto a la latitud se vio que por cada unidad de aumento en la latitud, el precio del alojamiento aumenta en 203.928. Para la longitud observamos que por cada unidad de aumento en la longitud, el precio del alojamiento disminuye en 820.599. Luego para "minimum_nights", por cada unidad de aumento en el número mínimo de noches, el precio del alojamiento aumenta en 0.14. Para "number_of_reviews", por cada unidad de aumento en el número de revisiones, el precio del alojamiento disminuye en 0.399.Para "reviews_per_month", por cada unidad de aumento en el número de revisiones por mes, el precio del alojamiento aumenta en 4.027. Luego, para "calculated_host_listings_count", por cada unidad de aumento en el recuento de listados calculados por el host, el precio del alojamiento aumenta en 0.079. Por último, para "availability_365", por cada unidad de aumento en la disponibilidad durante 365 días, el precio del alojamiento aumenta en 0.177.