
Universidad de
SanAndrés

# BIG DATA - TP2

Di Costanzo Pereira, Nina - Morozumi, Juan Ignacio - Profesores: Gibbons,
Universidad de San Andrés - Otoño 2024

**PARTE I:**

A) En este apartado, nos encargamos de eliminar los datos duplicados, lo que nos da es que 10 datos fueron eliminados ya que alguno de sus datos era iguales en todas sus variables.

B) En este caso eliminamos las siguientes columnas:
   - La variable “id”: al solo identificar el dato con un número no consideramos de relevancia la variable, ya que no aporta información, por ende, eliminamos;
   - La variable “name” ya que es el nombre o título del anuncio, pasa el mismo caso que con la variable anterior, por ende, la eliminamos.
   - La variable “host_id” pasa exactamente lo mismo que con la primera variable, por ende, la eliminamos.
   - La variable “Host_name” es lo mismo que la primera variable, por ende, la eliminamos.
   - La variable “neighbourhood” es lo mismo que la primera variable, por ende, la eliminamos.
   - Por último, last review la eliminamos ya que no es un dato de interés, tomar solo el valor de la última persona no aporta a nuestro estudio, ya que puede estar sesgado por la persona, si fuese un promedio de las últimas 100 reviews, el promedio total y el primer promedio, podríamos capaz observar para ver si los hospedajes mejoran, pero un dato no nos dice nada, por ende, eliminamos.

C) En el apartado de missing data, lo que optamos hacer fue en principio crear unas nuevas variables, en donde las mismas toman el valor promedio de los datos de la variable original y rellenan los espacios en blanco con esa variable. Después, cuando realicemos los estudios, los haremos en torno a la variable original y la variable con la missing data arreglada, así podemos eliminar problemas de sesgo que se puedan generar por tomar el promedio o la mediana. Creemos que es el mejor ya que solo eliminar datos nos cambiaría el resto de las variables, asi que no lo optamos por hacer. La eliminación de estilo Pairwise no sirve ya que no sabemos si el modelo es lineal y el resto de métodos tienen demasiada desventaja. Eso si, solo elegimos eliminar los datos donde tanto la variable “price” como la variable “revies_per_month” no hay datos ya sería modificar demasiado las variables. Como solo son 15 datos, no cambiaría mucho la estadística.

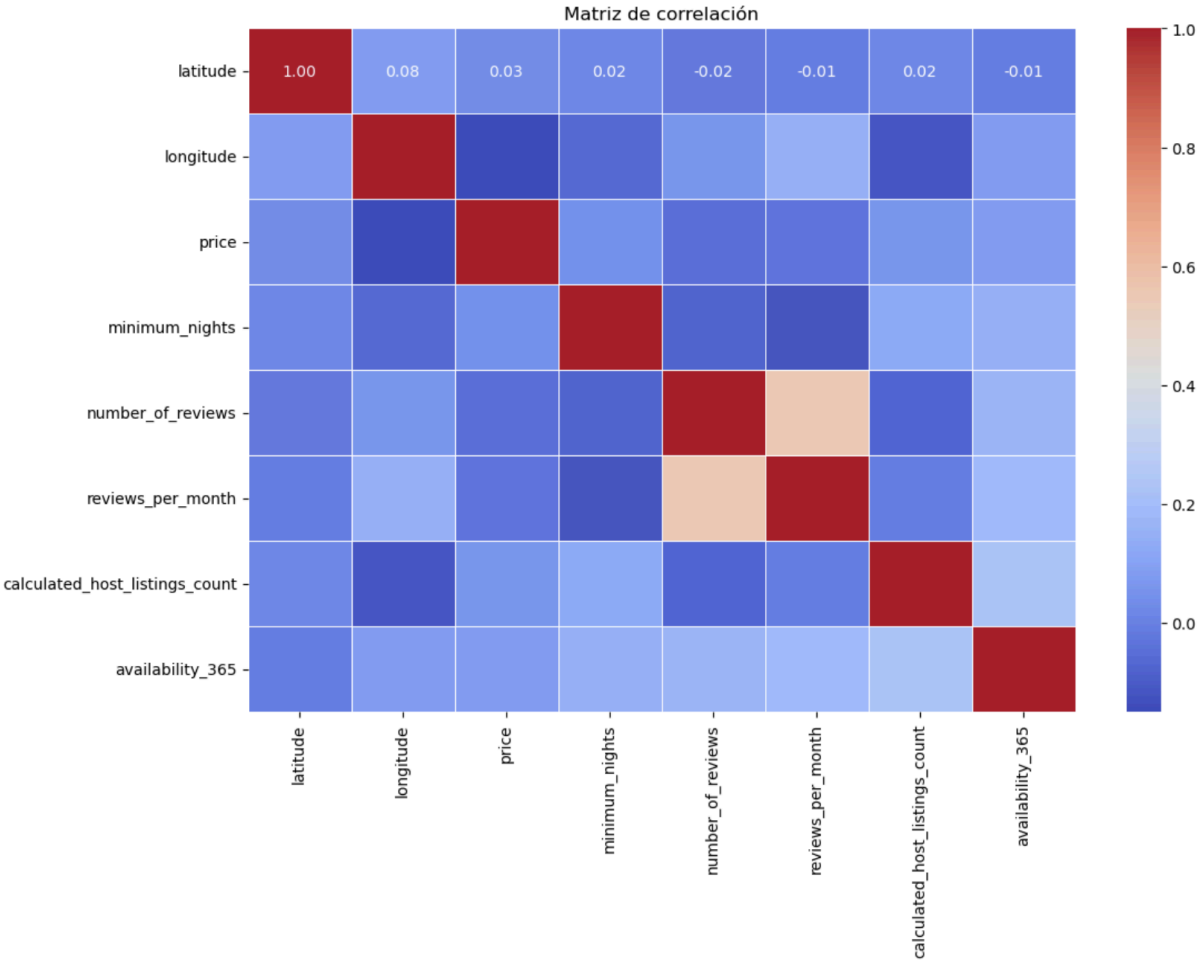
D) En este caso, con los outliers optamos por dejarlos como datos, ya que no consideramos eliminar datos que sean extremadamente grandes. Consideramos que son parte de la estadística, y que no sesgan el estudio sino que son parte de él. En cambio, si lidiamos con los valores absurdos, el único caso donde se dio fue la disponibilidad, que donde nos daban datos negativos decidimos eliminarlos ya que uno no puede tener disponibilidad negativa.

E) En este apartado transformamos los datos a valores numéricos, lo hicimos antes de lo que se debería así cuando trazamos los histogramas no generamos problemas.

F) En este caso, usamos los comandos “groupby” y “merge” para hacer lo que se nos pide en este inciso.

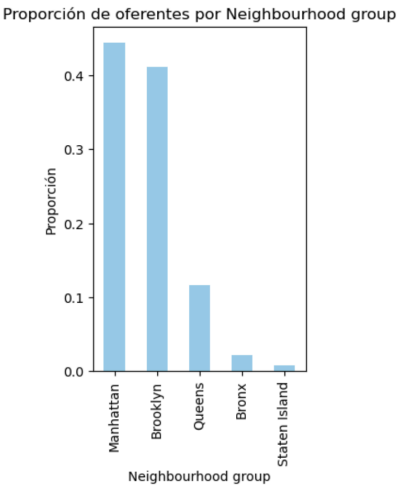
PARTE II:

1.



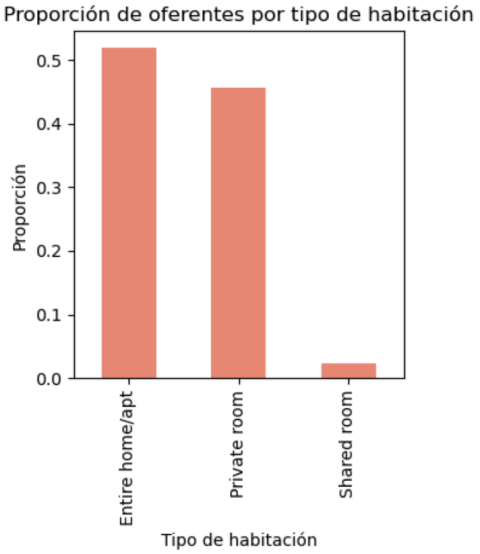

Como podemos observar en la matriz de correlacion propuesta, las unicas dos variables que presentan algun signo de una correlacion alta entre si, son las de “number_of_reviews” y la de “reviews_per_month”. Ya que, por el color de sus cuadros, podemos apreciar una correlación mayor al 0,5.

2.



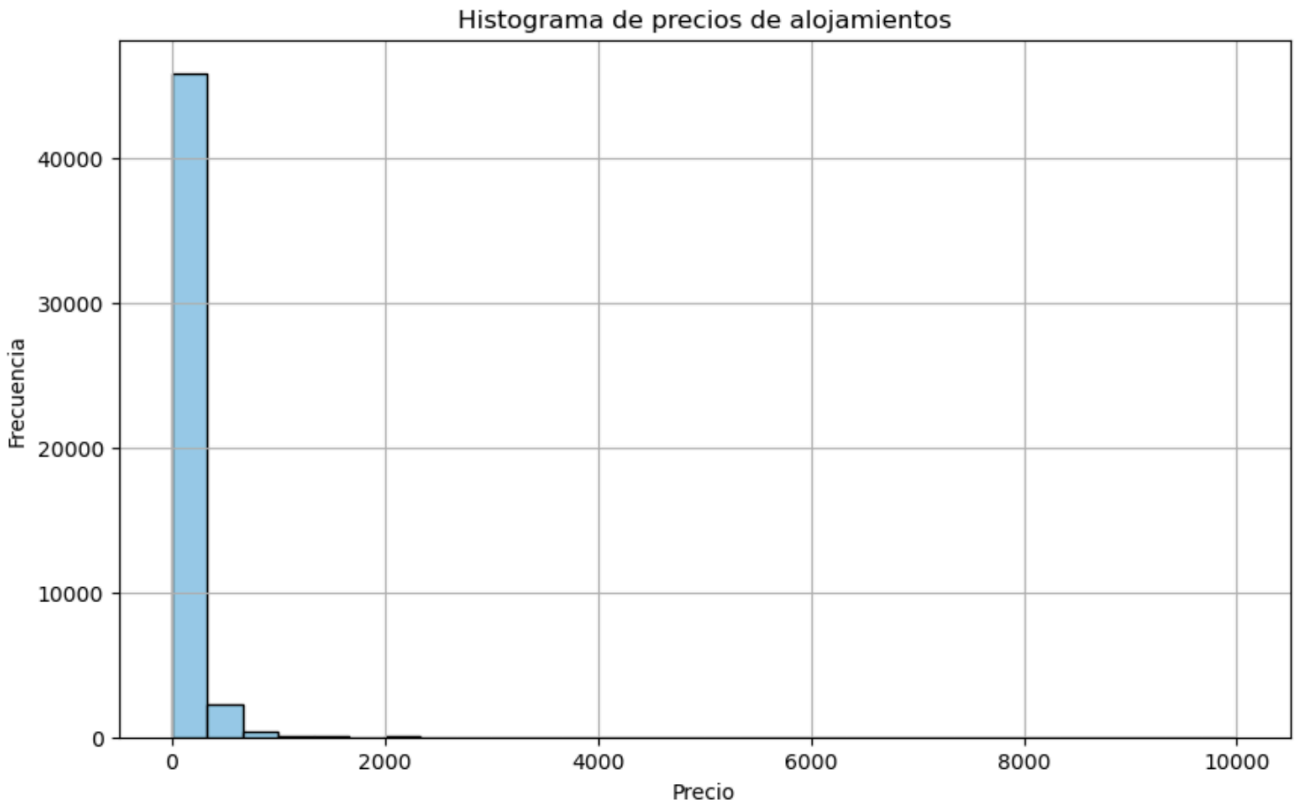

La proporción de los oferentes de “ Neighbourhood group”, son principalmente de las zonas de Manhattan y de Brooklyn, que se encuentran por encima del 40%. Luego,

un poco por encima del 10% nos encontramos con el barrio de Queens. Y por último, completando los barrios del estado de Nueva York, nos encontramos con el barrio de Bronx y de Staten Island con un porcentaje menor al 5%.



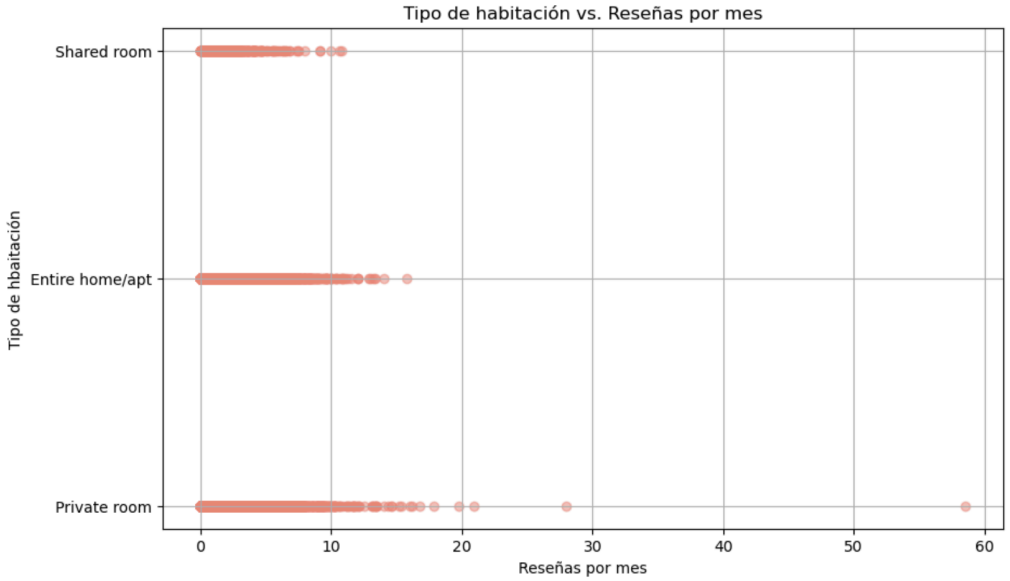

Las proporciones de los oferentes dadas las características de los inmuebles ofrecidos las podemos observar en el cuadro superior, donde las publicaciones que ofrecen departamentos o casas enteras se llevan un porcentaje un poco mayor al 50% de la totalidad. Luego, le siguen las ofertas por habitaciones privadas que se acerca al 45% de la totalidad ofrecida. Mientras que por último, encontramos las habitaciones compartidas, las cuales no superan el 5% del mercado.

3.

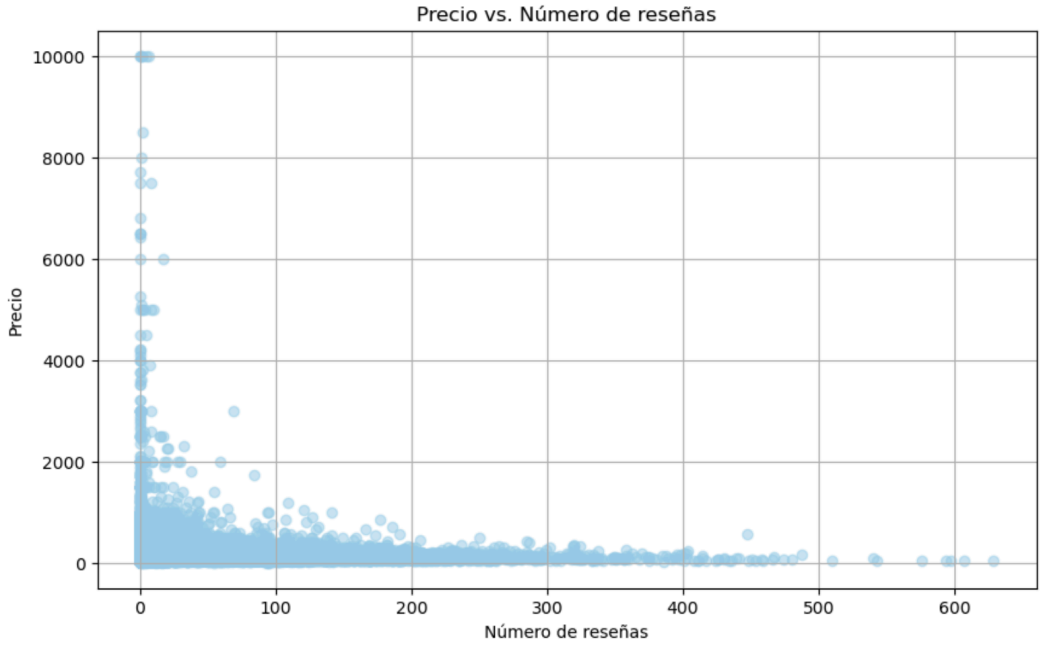


Podemos observar que los precios que se manejan en su mayoría, se encuentran por debajo de los 1000 dólares.
El precio mínimo es de 0, el precio máximo es de 10000. Mientras que el precio mínimo es de 152 dólares aproximadamente.
El promedio de precios en el barrio de Bronx es de 87 dólares, el del barrio de Brooklyn es de 114, el de Manhattan es de 196, el de Queens es de 99, y por último, nos encontramos con el barrio de Staten Island donde el promedio es de 114.
Ahora el precio promedio por una casa/departamento entero es de aproximadamente de 211 dólares, mientras que el de una habitación privada es de 89, y el de una habitación compartida ronda los 70 dólares.

4.

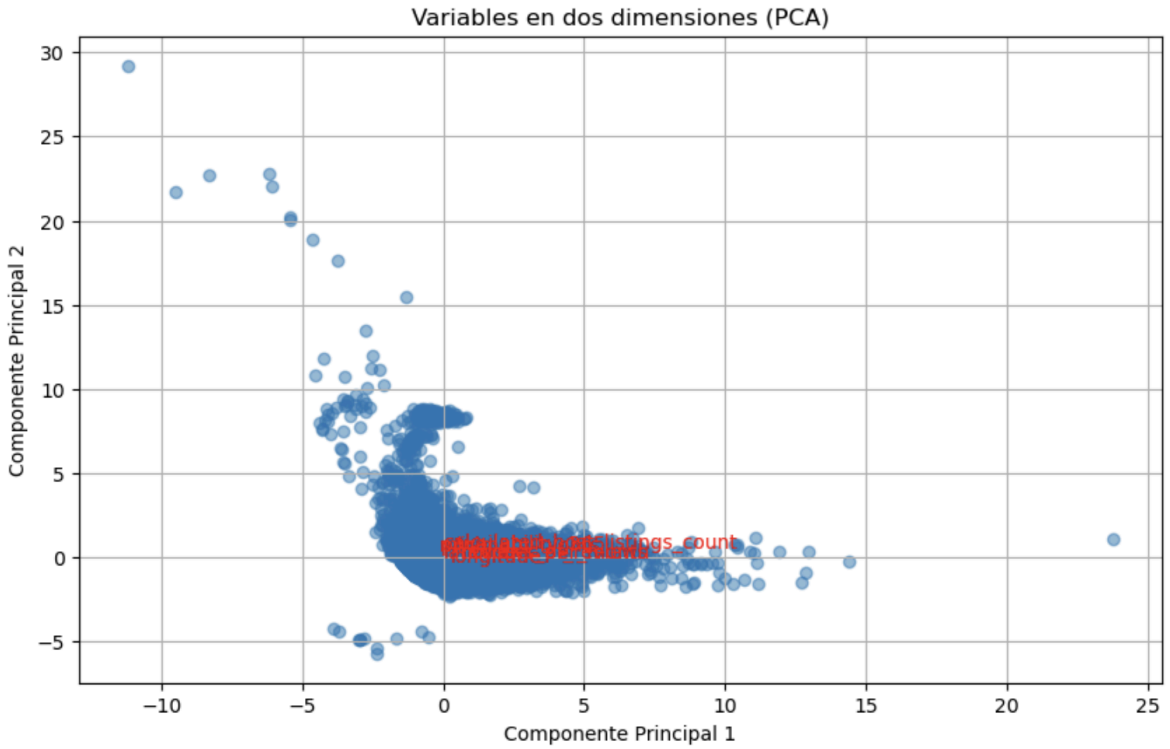



El primer scatter plot es de la variables de Reseñas por mes y el de tipo de habitación. En él, podemos observar como el tipo de habitación que se ofrece en la página que recibe más reseñas es el de las habitaciones individuales, le sigue el de la casa/departamento completo, y por último, el de la habitación compartida.



En este scatter plot, podemos observar como las variables que se comparan son los precios con el número de reseñas que tiene cada inmueble. En él, podemos observar que los que aquellas que reciban más reseñas son las que tienen los precios mas bajos del mercado, mientras que las que tienen un valor más elevado son las que tienen un menor número de reseñas.

5.




El porcentaje de la varianza explicada por los dos componentes es de un 37.69. Es decir, que los dos componentes sólo son capaces de analizar un 37,69% de la variabilidad de los datos originales de la base de datos.

**PARTE III:**

Primero, hemos eliminado del conjunto de datos todas las variables relacionadas con el precio. Para ello, hemos cargado el conjunto de datos desde el archivo CSV ubicado en el repositorio de GitHub y eliminamos las filas con valores faltantes.

Hemos dividido el conjunto de datos en una base de entrenamiento (70%) y una base de prueba (30%) utilizando el comando train_test_split. Establecimos la semilla (random_state) en 201 y la variable dependiente en la base de entrenamiento fue el precio (price), mientras que el resto de las variables fueron consideradas como variables independientes.

Hemos implementado un modelo de regresión lineal utilizando la biblioteca statsmodels.api. A continuación, se presenta el resumen del modelo obtenido:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.107
Model:                            OLS   Adj. R-squared:                  0.106
Method:                 Least Squares   F-statistic:                     515.0
Date:                Sun, 21 Apr 2024   Prob (F-statistic):               0.00
Time:                        21:04:04   Log-Likelihood:            -2.5806e+05
No. Observations:               38831   AIC:                         5.161e+05
Df Residuals:                   38821   BIC:                         5.162e+05
Df Model:                           9
Covariance Type:            nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
neighbourhood_group               8.3393      1.339      6.226      0.000       5.714      10.964
latitude                        152.4437     17.976      8.480      0.000     117.210     187.678
longitude                      -533.1654     21.097    -25.272      0.000    -574.515    -491.815
room_type                       -96.5124      1.779    -54.263      0.000     -99.999     -93.026
minimum_nights                   -0.1813      0.055     -3.270      0.001      -0.290      -0.073
number_of_reviews                -0.2045      0.024     -8.607      0.000      -0.251      -0.158
reviews_per_month                 0.2135      0.688      0.310      0.756      -1.135       1.562
calculated_host_listings_count   -0.0102      0.037     -0.273      0.785      -0.083       0.063
availability_365                  0.1677      0.008     21.920      0.000       0.153       0.183
reviews_per_month_dummy       -4.547e+04   1766.384    -25.743      0.000   -4.89e+04    -4.2e+04
==============================================================================
Omnibus:                    97250.398   Durbin-Watson:                   1.928
Prob(Omnibus):                  0.000   Jarque-Bera (JB):       2271223845.119
Skew:                          27.410   Prob(JB):                         0.00
Kurtosis:                    1186.535   Cond. No.                     3.48e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.48e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

R-squared: El coeficiente de determinación del modelo es 0.107, lo que indica que aproximadamente el 10.7% de la variabilidad en el precio de alquiler puede explicarse por las variables incluidas en el modelo.

**Interpretación de los Coeficientes**

- ➢ neighbourhood_group: Manteniendo todas las demás variables constantes, un cambio de una unidad en el grupo de vecindario se asocia con un aumento de $8.34 en el precio del alquiler.
- ➢ latitude: Un aumento de una unidad en la latitud se asocia con un aumento de $152.44 en el precio del alquiler, manteniendo todas las demás variables constantes.
- ➢ longitude: Un aumento de una unidad en la longitud se asocia con una disminución de $533.17 en el precio del alquiler, manteniendo todas las demás variables constantes.
- ➢ room_type: El alquiler de habitaciones privadas (room_type) tiene un efecto negativo en el precio del alquiler. Manteniendo todas las demás variables constantes, el precio del alquiler disminuye en $96.51 cuando la propiedad es una habitación privada en comparación con una vivienda entera.
- ➢ minimum_nights: Manteniendo todas las demás variables constantes, un aumento de una noche mínima de estancia está asociado con una disminución de $0.18 en el precio del alquiler.
- ➢ number_of_reviews: Manteniendo todas las demás variables constantes, un aumento de una revisión está asociado con una disminución de $0.20 en el precio del alquiler.
- ➢ availability_365: Manteniendo todas las demás variables constantes, un aumento de una unidad en la disponibilidad durante 365 días se asocia con un aumento de $0.17 en el precio del alquiler.

El error cuadrático medio (MSE, por sus siglas en inglés) se utiliza para evaluar la de un modelo de regresión. Cuanto menor sea el valor del MSE, mejor será la capacidad predictiva del modelo. Los MSE del modelo son:

- Error cuadrático medio en el conjunto de entrenamiento es 37232.13935725737
- Error cuadrático medio en el conjunto de prueba es 28674.679830454133

Dado que es menor  en el conjunto de prueba que en el conjunto de entrenamiento, indica que el modelo generaliza bien a nuevos datos y no está sobreajustado al conjunto de entrenamiento. Esto sugiere que el modelo tiene una buena capacidad predictiva en datos no vistos.