Ciencia de Datos

Trabajo Práctico № 2

Ignacio Spiousas

Noelia Romero

Ignacia Auguste

Florencia Reverendo

Trinidad Morán

Parte I: Limpieza de la base

En esta sección del trabajo realizamos la preparación de la base de datos para su posterior análisis. La base está compuesta por 48905 observaciones de publicaciones de la página de alojamientos *Airbnb* para la ciudad de Nueva York, y se registran las siguientes 15 variables de cada una: nombre; *id* del *host*, nombre del *host*; barrio al que pertenece; latitud, longitud; tipo de cuarto (compartido, individual, casa o departamento entero); precio; mínimo de noches; cantidad de reseñas; última reseña; reseñas por mes; cantidad de publicaciones del *host*; y días de disponibilidad al año.

Para la limpieza, eliminamos las columnas nombre, nombre del *host* y última reseña, ya que no eran relevantes. Utilizamos *drop.duplicates* de pandas para eliminar duplicados, conservando la primera fila. Identificamos y reemplazamos los NA en reseñas por mes con 0, ya que eran las publicaciones con 0 reseñas. También eliminamos 15 filas con valores faltantes en la columna precio. Basamos nuestra decisión en que la base de datos es numerosa, los datos parecen ser *missing completely at random* y eliminarlos no compromete una disminución significativa del volumen de la muestra (Patrician, 2002). Con el paquete *missingno* graficamos una matriz de valores faltantes antes y después del proceso para asegurarnos que no quedara alguno sin solucionar.

En vistas al control de outliers, filtramos las columnas de disponibilidad, precio, noches mínimas y cantidad de publicaciones del *host* por valores mayores a 0, ya que no tienen sentido valores iguales o menores a 0 en esas variables. Después, graficamos la distribución de cada una de las columnas a través de histogramas de frecuencia. Al hacerlo, vimos que las variables precio, noches mínimas, y cantidad de publicaciones del *host* tienen una forma logarítmica. Esto representa un problema a la hora de eliminar *outliers,* por lo que transformamos estas variables aplicando función logarítmica para acercarnos a distribuciones más simétricas y así facilitar la detección de los outliers sin estar sesgados hacia un lado. Las variables cantidad de reseñas y reseñas por mes también tienen forma logarítmica, pero

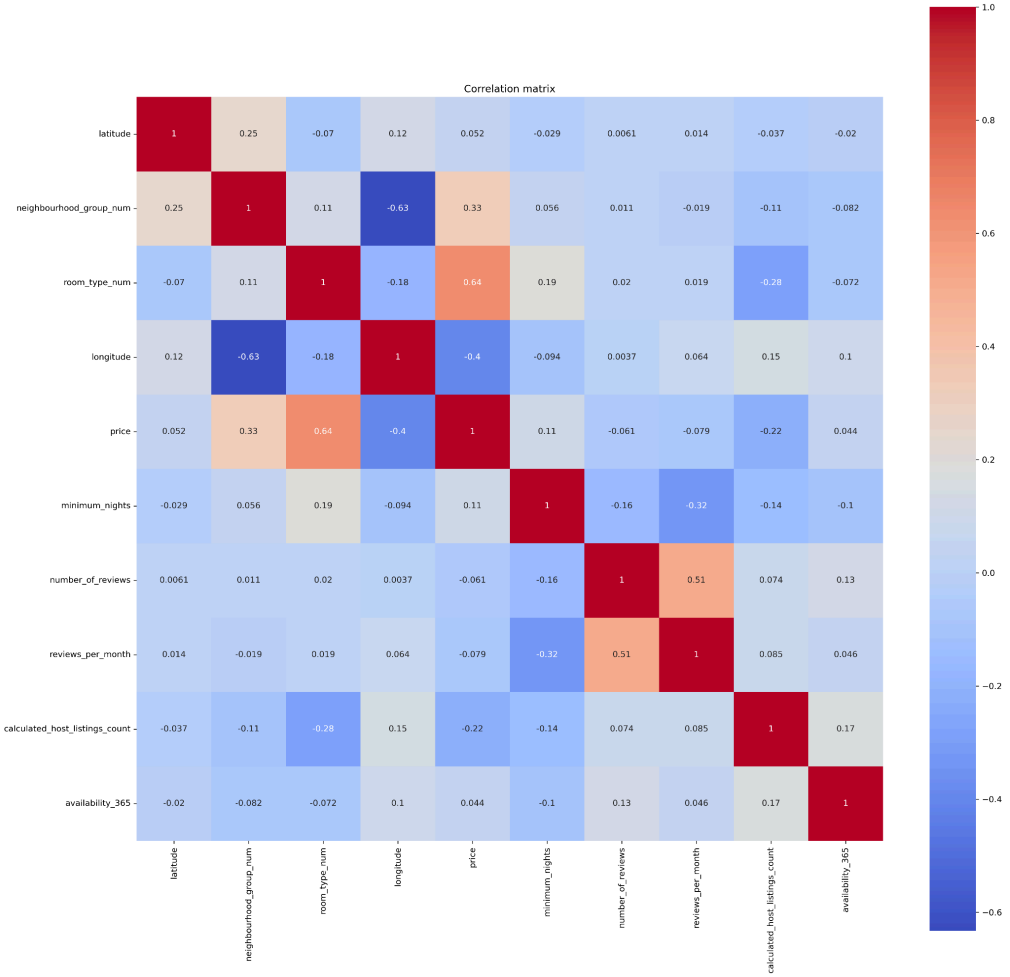
decidimos no transformarlas ya que tienen muchas observaciones que valen 0. De todos modos, no las incluimos como variables relevantes para eliminar valores atípicos. Una vez hecho esto, definimos e implementamos una función para eliminar *outliers* utilizando el criterio del rango intercuartílico, utilizando las columnas de precio, latitud, longitud, noches mínimas, cantidad de publicaciones del *host* y disponibilidad.

Por último, transformamos las variables categóricas "barrio" y "tipo de habitación" a variables categóricas numéricas. Para la primera variable definimos cada barrio con un número del 0 al 4 ordenado en función de su atractivo turístico, de menor a mayor ('Staten Island': 0, 'Bronx': 1, 'Queens': 2, 'Brooklyn': 3, 'Manhattan': 4); para la segunda variable establecimos un número del 0 al 2 ordenado en función de la privacidad que ofrece la publicación, de menor a mayor ('Cuarto compartido': 0, 'Cuarto individual': 1, 'Casa/dpto completo': 2). Agregamos esta codificación como columnas nuevas para conservar también la información original. También creamos una nueva columna correspondiente a la oferta por grupo de barrio para cada barrio. Finalizado el proceso, nuestra base de datos cuenta con 21.680 filas de observaciones y 15 columnas de variables.

Parte II: Gráficos y representaciones

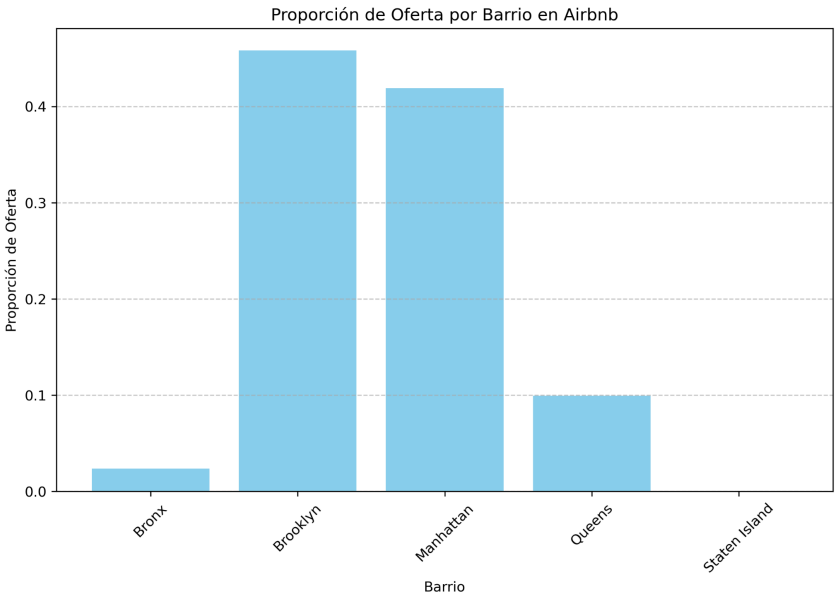
Ejercicio 2: Matriz de correlación

Realizamos una matriz de correlación (Figura 1) con variables de interés. En la figura podemos observar que las tonalidades cálidas indican un coeficiente de correlación positivo, mientras que las frías un coeficiente de correlación negativo. Es a partir de esta matriz que podemos resaltar la fuerte correlación positiva entre precio y tipo de habitación. Por la codificación de la variable, entendemos que privacidad/individualidad de los cuartos y precio aumentan o disminuyen en la misma dirección, con un coeficiente de correlación de 0,64. Otra correlación positiva similar pero menor es la de precio y grupo de barrio, la cual tiene un coeficiente de 0,33. Esta correlación implica que a medida que nos desplazamos de manera creciente a través de los valores de nuestra variable codificada en función del atractivo turístico, los precios aumentan. El resto de los valores que más resaltan son esperables, como una correlación positiva entre barrio y latitud (0,25); o negativa entre barrio y longitud (-0,63) - por ende también entre precio y longitud, (-0,4)-; también una relación negativa entre cantidad de reseñas por mes y mínimo de noches (-0,32); y otra correlación negativa entre la cantidad de publicaciones del *host* y el tipo de la habitación (-0,28, a menor privacidad puede hacer más publicaciones de la misma propiedad). A su vez, hay una débil correlación negativa entre cantidad de publicaciones del *host* y precio (-0,22), que no esperábamos.



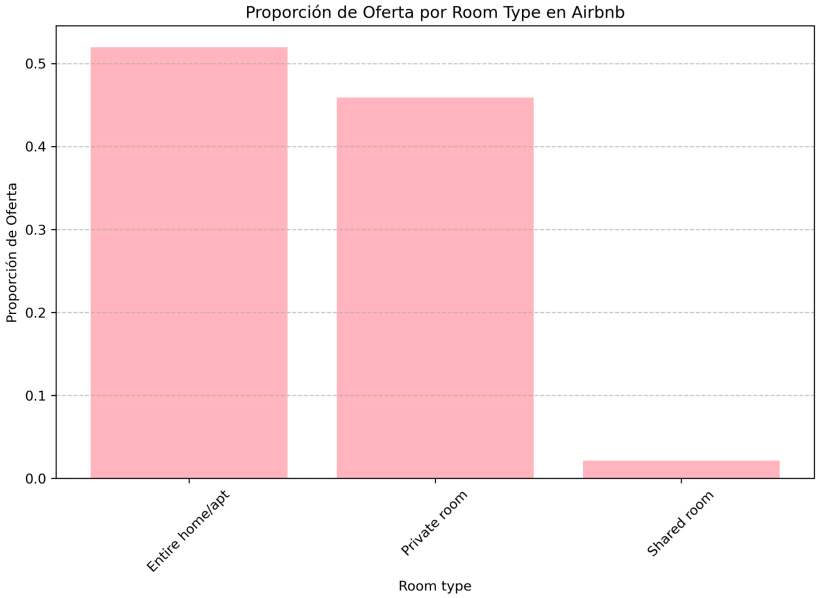
Ejercicio 3: Proporciones

Anteriormente habíamos calculado la cantidad de oferentes en cada tipo de barrio que se encontraba en la base. Para poder calcular la proporción de oferentes por tipo de barrio, dividimos la cantidad de oferentes en cada barrio en la cantidad total de oferentes y así obtuvimos el siguiente histograma con la distribución de la oferta por barrio.



Podemos observar que el barrio con mayor cantidad de oferentes es Brooklyn que tiene un 46% del total de alojamientos ofrecidos en Airbnb, seguido de Manhattan con el 42%. Queens tiene el 9% de los alojamientos y Bronx representa el 2.3%. Finalmente encontramos a Staten Island que contiene menos del 1% de la oferta disponible.

Realizamos el mismo procedimiento para cada tipo de habitación y obtuvimos la siguiente información:



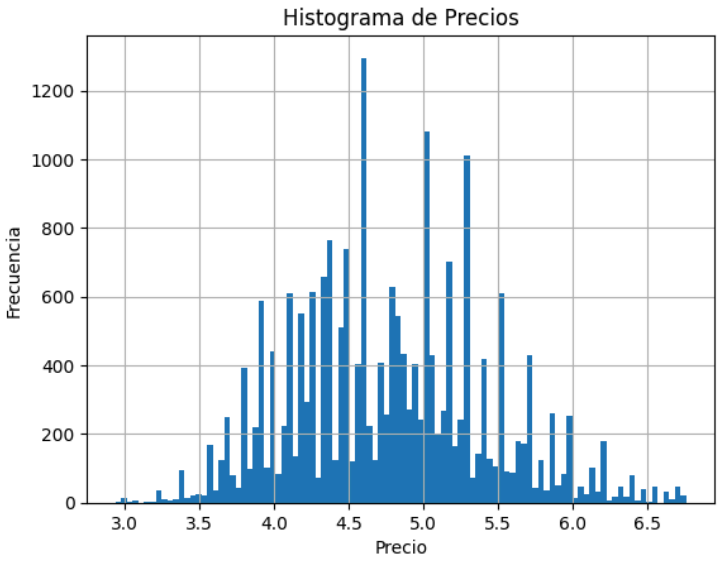
Podemos ver que más de la mitad de los oferentes disponen de un departamento completo (52%). Otro tipo de alojamiento altamente ofrecido son las habitaciones privadas que representan un 46% de las opciones. La proporción de habitaciones compartidas que se ofrecen es muy inferior en comparación al resto, representando solo el 2%.

Ejercicio 4: Histograma de precios

El histograma nos permite ver la dispersión de precios de los alojamientos. Podemos observar que los precios varían entre 2.9 y 6.5 USD. Por la dispersión del gráfico pareciera que la mayoría de los alojamientos cuestan entre 4.3 y 5.0 USD. Además, la frecuencia más alta nos indica que un poco más de 800 alojamientos tienen un valor de 5.0 USD, siendo éste el valor más común de encontrar. El precio que se repite con mayor frecuencia poco difiere del precio promedio de los alojamientos que es de 4.7 USD.

Si analizamos por cada barrio podemos encontrar que el precio promedio de un alojamiento en Staten Island es de 6.6 USD siendo el barrio con los precios más elevados. Manhattan tiene un costo promedio de 5.0 USD en cada alojamiento. Brooklyn tiene un valor promedio de alojamiento de 4.6 USD, seguido de Queen con 4.4 USD. El alojamiento que en promedio es más económico es en el barrio de Bronx donde el precio promedio es de 4.2 USD.

Por habitación, como era de esperar, podemos ver que el alojamiento más caro es en un departamento completo, siendo el valor promedio de 5.1 USD. Si se busca una habitación privada el precio promedio es de 4.3 USD y para una opción más económica se puede encontrar una habitación compartida por 4.0 USD en promedio.

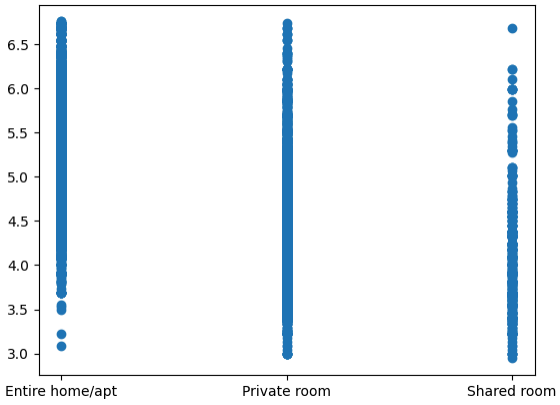


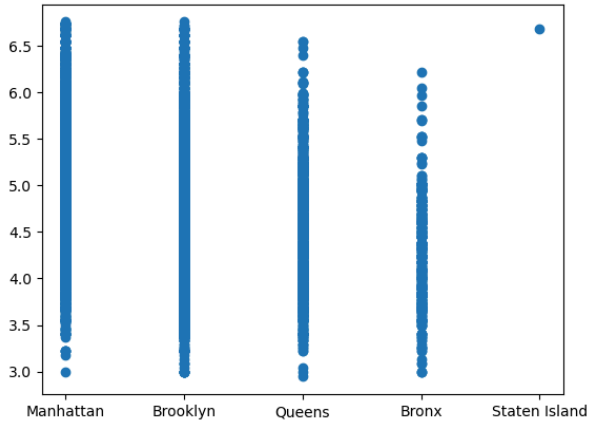
Ejercicio 5: Scatter plots

Elegimos las variables “Precio” y “Tipo de habitación" para realizar el primer scatter plot porque son las variables que nos dieron con valor de correlación más alto de todos y nos pareció interesante ver la dispersión del precio por cada tipo de habitación. Al analizar el gráfico confirmamos nuestras sospechas sobre la distribución del precio en cada tipo de habitación. Vemos que en el departamento completo la mayoría de los datos se encuentran en

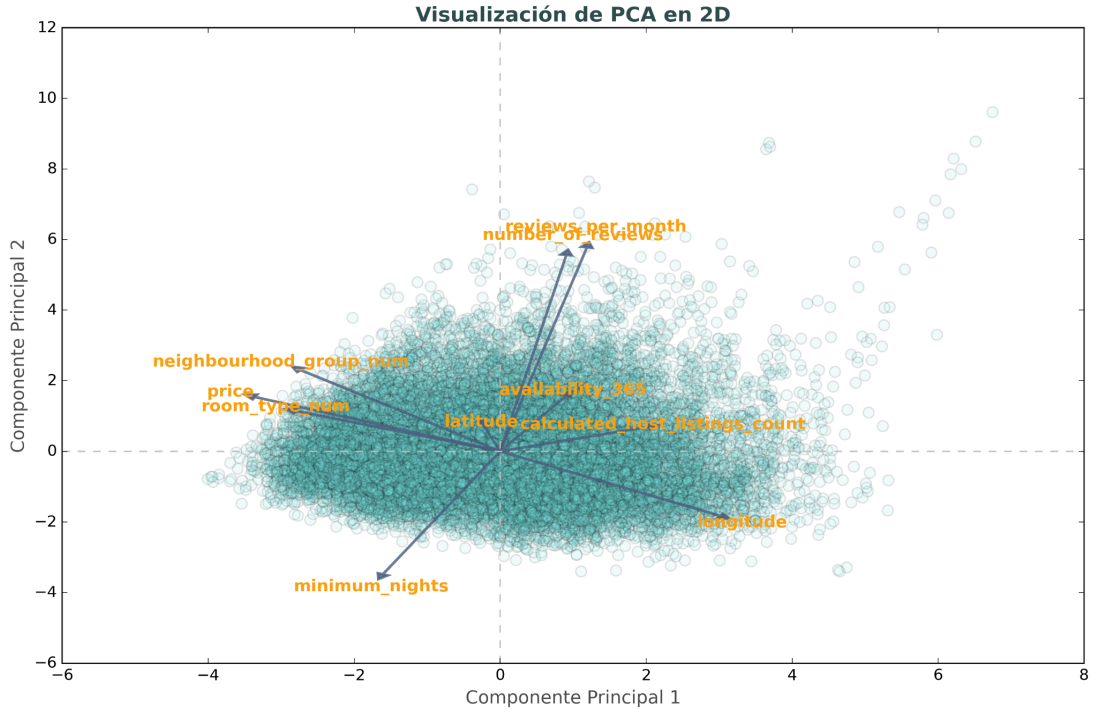
la parte superior del eje y ya que el precio de este tipo de alojamiento es el más elevado. La habitación privada también tiene valores altos al evaluarla respecto al precio pero se nota que son datos más dispersos, la mayor cantidad de datos pareciera concentrarse en los valores medios de precio. La habitación compartida tiene muy pocos datos ubicados en la parte superior, los datos se concentran en los valores de precios más bajos. Podemos comprobar lo que muestra la correlación positiva y es que a medida que suben los precios encontramos alojamientos que tienen una mayor privacidad. Si la privacidad del alojamiento disminuye también disminuye el valor del alojamiento como es esperable que suceda.

Quisimos hacer un análisis parecido con los distintos barrios que ofrecen alojamiento para analizar la varianza del precio y encontramos que los datos también condicen con lo esperado por los precios promedios de los alojamientos. Vemos que Manhattan tiene los precios más altos, con datos concentrados en la parte superior del eje Y, indicando una oferta de alojamientos más costosa. Brooklyn sigue una tendencia similar, pero con una mayor dispersión de precios en rangos más bajos. Queens muestra precios algo más bajos que Manhattan y Brooklyn, con menos dispersión en los valores altos. Finalmente, Bronx tiene los precios más bajos de todos los barrios, con pocos datos en la parte superior del gráfico.





Ejercicio 6: Gráfico de variables usando PCA.



La siguiente figura expone de manera gráfica nuestros datos en función de los dos componentes principales encontrados aplicando Análisis de Componentes Principales. El primer componente principal explica el 23.9% de la varianza, mientras que el segundo explica 16.7%, en conjunto, explican alrededor del 40% de la varianza total. Los siguientes componentes explican un 12.6%, 10.7%, 10.3%, 8.3%, 7.4%, 4.4%, 3.3%, hasta 2.4%. Recién en el 6to componente tenemos un 80% de la varianza explicada, por lo que es una estructura de datos compleja. A través del gráfico, podemos observar que el grupo de variables barrio, precio y tipo de habitación se proyectan en la misma dirección que el componente principal 1, con loadings negativos; mientras que la cantidad de publicaciones del *host* y longitud (como habíamos visto en la primer matriz de correlaciones) se proyectan manera inversa con loadings positivos. A su vez, el grupo de las variables que contemplan reseñas (por mes y en total) y -en menor medida- disponibilidad, se proyectan en la dirección del componente principal 2, con loadings positivos; mientras que en la misma dirección pero hacia el lado contrario, con loadings negativos, se dirige la variable de mínimo de noches.

Ejercicio 8: Test & Train

Dividimos la base en dos, creando una base de entrenamiento que posee el 70% de los datos y otra base de prueba que contiene el 30% restante. Asignamos la variable “Price” como variable dependiente en la base de entrenamiento y el resto de las variables fueron tomadas

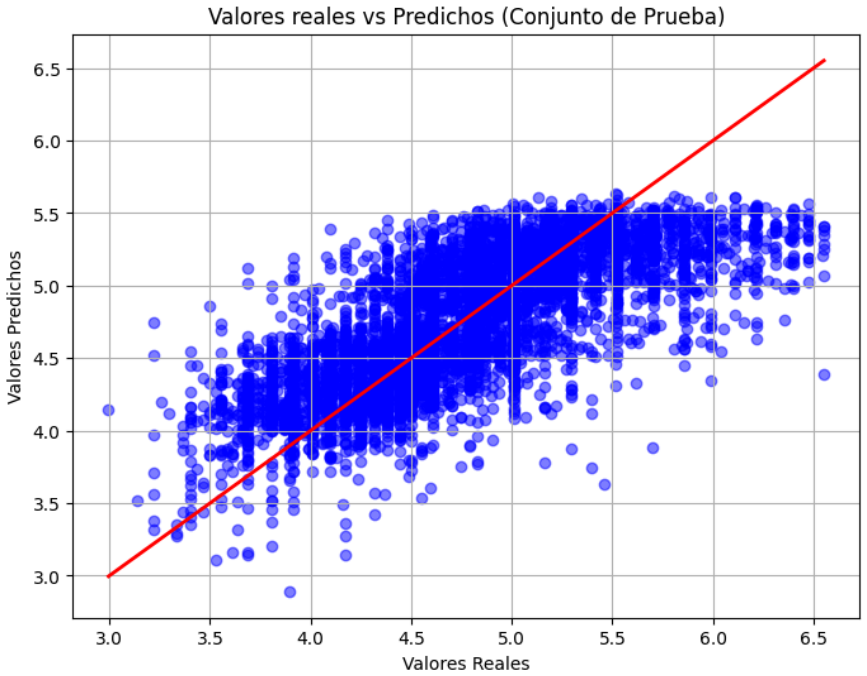
como variables independientes. El conjunto de prueba quedó comprendido por 15176 datos y el conjunto de entrenamiento por 6504 datos.

Ejercicio 9: Análisis de la regresión lineal

Al realizar la regresión lineal podemos afirmar que este modelo explica más de la mitad de la variabilidad de los datos ya que obtuvimos un R^2 de 0.54 en la evaluación del conjunto de entrenamiento. Esto indica que el modelo tiene una capacidad predictiva aceptable, podría ser mejor.

El MSE es bajo, lo que sugiere que la diferencia entre los valores predichos y los reales es pequeña. En el caso del conjunto de prueba el MSE es de 0.195, mientras que en el conjunto de entrenamiento es de 0.194. Dado que ambos valores son casi idénticos podemos decir que el modelo no solo ajusta bien los datos que conoce, sino que también generaliza adecuadamente a nuevos datos. Esto refuerza la idea de que el modelo es efectivo para esta tarea predictiva.

Los coeficientes nos indican cómo varía el precio en función de cada una de las variables independientes. La magnitud y signo de los coeficientes permiten interpretar qué variables tienen un mayor peso en la predicción del precio y en qué dirección influencian al modelo. Por ejemplo, la variable "room type num" tiene un coeficiente positivo (0.731), lo que significa que, manteniendo las demás variables constantes, un aumento en la el tipo de habitación (entendiendo aumento en el tipo de habitación como aumento de la privacidad) incrementa el precio. Por el contrario, la "minimum nights" tiene un coeficiente negativo (-0.064), lo que sugiere que un aumento en la el mínimo de noches disminuye el precio, es decir que esta variable afecta negativamente a la variable precio.



Estos coeficientes ayudan a entender qué factores son más relevantes para predecir los precios y cómo interactúan entre sí en el contexto del modelo.