TRABAJO PRÁCTICO N°2

1. LIMPIEZA DE BASE DE DATOS

Para el análisis de oferentes de Airbnb de la ciudad de Nueva York se importó la base de datos "Base Airbnb NY.csv". Esta contiene 16 variables para 48.905 registros de lugares para hospedarse. Para poder hacer un correcto análisis primero se realizó una limpieza de la base. Para esto, primero se eliminaron los valores duplicados (10 filas fueron eliminadas); segundo se sacaron las columnas que no tienen información de interes (*'id', 'name', 'host_id', 'host_name', 'neighbourhood', 'last_review*') quedando así 10 columnas (variables).

Para los *missing values* decidimos realizar una imputación, que es un procedimiento en el cual se reemplazan los datos faltantes con algunos valores razonables (Nguyen, 2020). En nuestro caso, reemplazamos los valores NaN por la mediana, ya que es mucho más resistente a outliers, a diferencia de la media que puede ser llevada a esos datos extremos que luego en el análisis de datos generen predicciones menos precisas. Además, como no conocemos la distribución de nuestros datos, usar la media en lugar de la mediana podría no ser representativa de la tendencia central (Nguyen, 2020).

En cuanto a los *outliers,* se obtuvieron a partir del uso del Rango Intercuartil (IQR). Decidimos usar este método que es más robusto frente a distribuciones sesgadas ya que no conocemos la distribución de los datos, entonces no podemos asumir normalidad. Luego de analizar los datos que eran outliers se decidió eliminar los de la columna "*availability_365*" ya que son pocos datos (15) a los que no se les podía atribuir ninguna interpretación (-999), por lo tanto, no nos aportan información relevante para nuestro análisis.

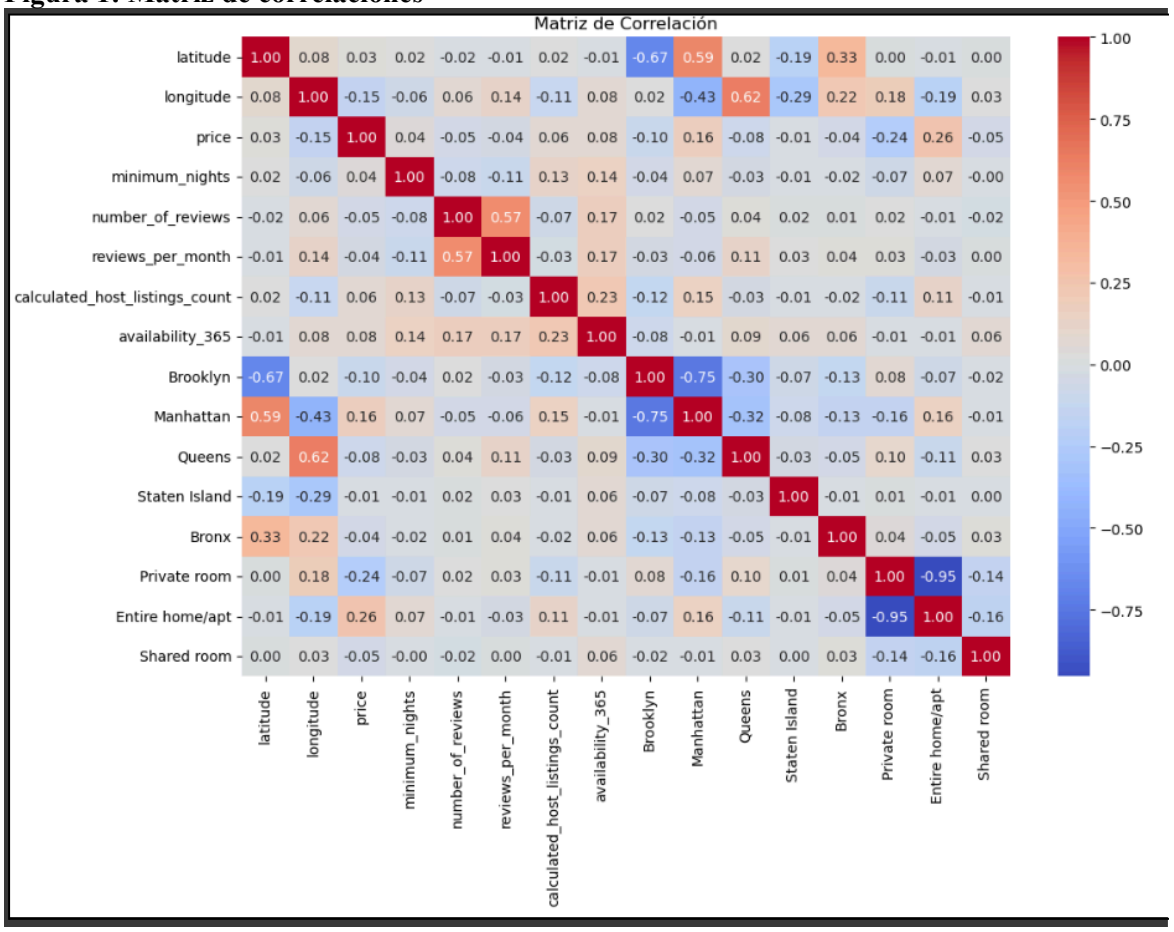
Luego, para terminar la limpieza de datos, transformamos las variables "*neighbourhood_group*" y "*room_type*" a variables numéricas y utilizamos la función *merge*, para crear una columna ("*offer_group*") que contiene la cantidad de oferentes por "*neighbourhood group*".

2. MATRIZ DE CORRELACIÓN

Una vez realizada la limpieza de la base de datos generamos una matriz de correlación (fig.1) entre las variables seleccionadas previamente, omitiendo "*offer_group*". Para esto la variable *neighbourhood_group* que previamente había sido transformada a variable numérica, se

transformó en una variable binaria (Dummy) para poder facilitar la interpretación de la matriz.

Figura 1: Matriz de correlaciones



En la matriz se encuentran correlaciones significativas entre la latitud y Brooklyn (-0.67). Esto pasa porque si los barrios se concentran en una región geográfica específica, debería resultar en una correlación significativa entre las coordenadas geográficas (latitud o longitud) y las variables dummy que representan los barrios, como Manhattan. Sin embargo, si las correlaciones no son muy significativas, pueden existir varias razones que lo expliquen, por ejemplo que haya una distribución geográfica diversa dentro del barrio (mismo análisis para latitud y longitud con distintos barrios).

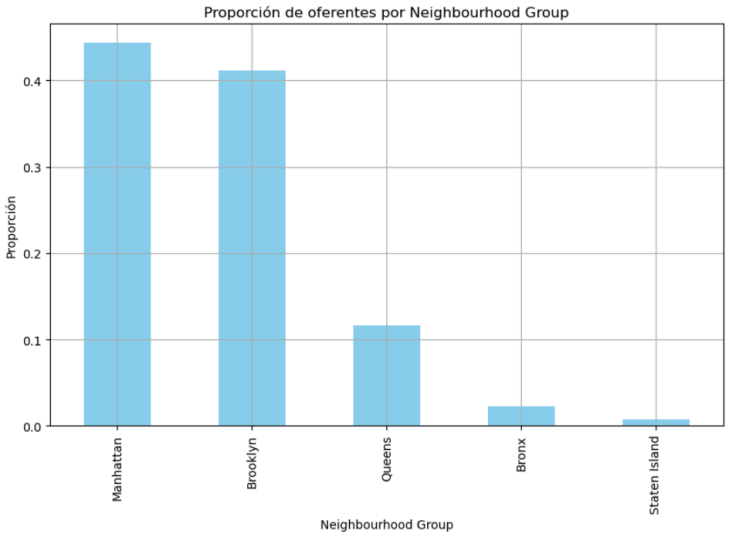
Por otra parte hay una correlación débil entre entire room y price (0,26), lo cual puede implicar que hay una leve tendencia a que este tipo de ofertas tengan un precio más elevado y que , otros factores (como la ubicación, los servicios ofrecidos, o la demanda en ciertos barrios) son importantes en la determinación del precio.

También hay una correlación alta entre dos barrios (Brooklyn y Manhattan) y negativa, lo cual indica que hay una relación inversa entre las características asociadas a esos barrios, quizás el precio, o tipo de alojamiento.

También existe una correlación muy alta negativa (-0.95) entre *private_room* y Entire room ya que son variables dummies y son mutuamente excluyentes (tienen una relación muy directa). Esto no pasa con shared room ya que "*private_room*" y "*Shared room*" no están tan estrechamente vinculadas como "*private_room*" y "*Entire room*", porque las dos primeras categorías no son igualmente comunes, y esto disminuye la correlación negativa.

3. GRÁFICOS (ejercicios 3, 4, 5)

Luego de la matriz de correlación se analizó la proporción de oferentes por "Neighbourhood group" y por tipo de habitación, dando como resultado los gráficos de la figura 2 y figura 3 respectivamente:



La figura 2 muestra la distribución de los oferentes en los distintos grupos de vecindarios de la ciudad. Los resultados indican que Manhattan concentra la mayor parte de los alojamientos ofertados, seguido por Brooklyn. Por su parte, Queens, Bronx y Staten Island tienen una representación significativamente menor.
Este patrón sugiere que la mayor parte de los alojamientos en Airbnb se concentran en las áreas más turísticas y céntricas de la ciudad, especialmente en Manhattan y Brooklyn.

Figura 2: Proporción de oferentes por barrio

La figura 3 muestra la distribución de los tipos de habitación que se ofrecen en la plataforma. Los resultados muestran que "Entire home/apt" es el tipo de alojamiento más común, y representa más de la mitad de las ofertas, seguido por "*private_room*" y "*shared rooms*" son menos comunes.
El dominio de los alojamientos completos puede estar relacionado con la preferencias de los usuarios de Airbnb, que buscan una mayor privacidad. Pero el hecho de que las habitaciones privadas tengan presencia significativa sugiere que hay una demanda por opciones más económicas.

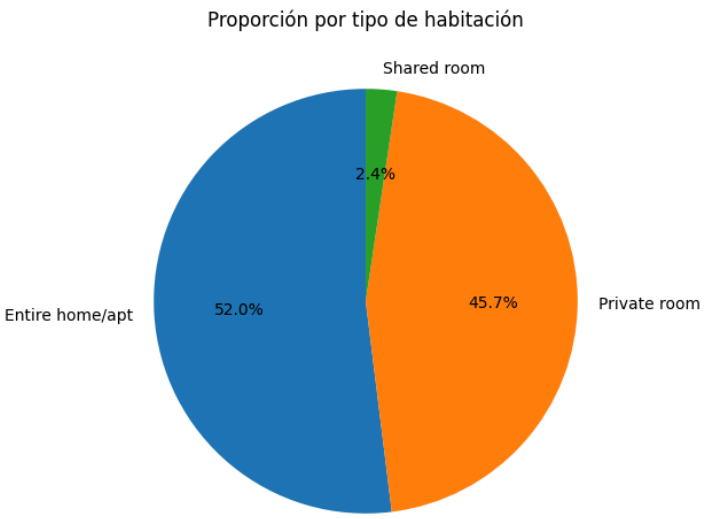


Figura 3: Distribución de tipos de habitación

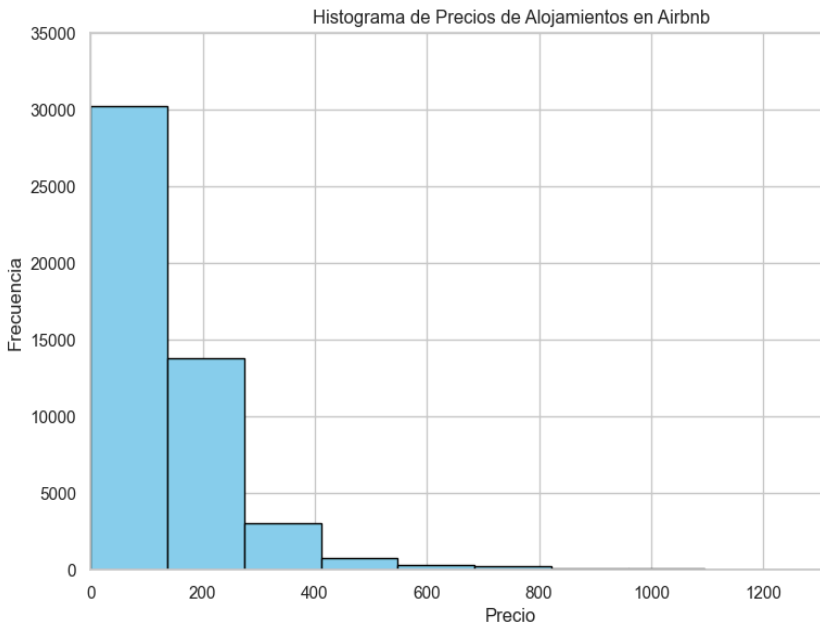


Figura 4: Precios de alojamientos

Por otro lado, el histograma de la figura 4 muestra que una gran cantidad de alojamientos tienen precios dentro de un rango que va desde desde un poco más que cero hasta aproximadamente 200 USD. Mientras mayor es el precio, se puede observar que disminuye la cantidad de publicaciones de hospedaje de una forma brusca, lo que indicaría que los precios elevados son menos comunes. La distribución es asimétrica y está sesgada hacia la derecha. Esto nos indica que son pocas las propiedades que tienen precio muy alto y por lo tanto elevan el promedio hacia arriba.

En la elaboración de este histograma (Fig.4) se consideró la Regla de Rice para estimar el número indicado de bins (intervalos) adecuado, considerando el tamaño de la muestra y el IQR. De este modo, se permite que los datos se puedan visualizar claramente al proporcionar un equilibrio entre la subrepresentación de datos y varios detalles visuales.

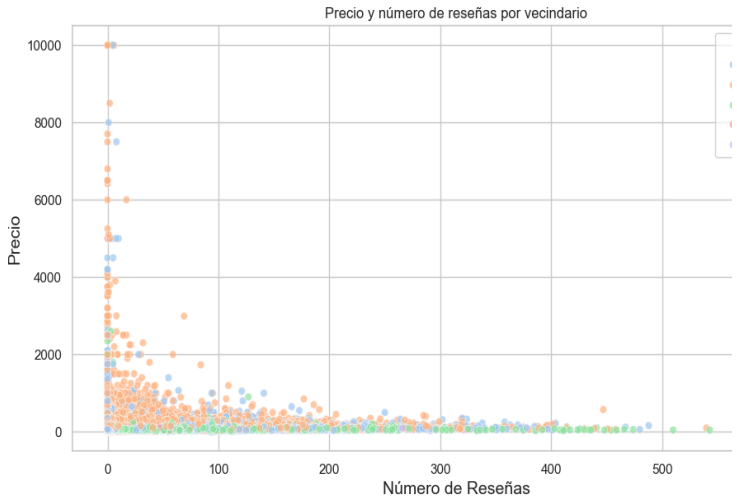


Figura 5: Precio y número de reseñas por vecindario

La figura 5 muestra un gráfico de dispersión cuya variable Y es el precio y la variable X el número de reseñas. Además, cada color representa algún distrito de New York. Sabemos que una tendencia ascendente indicaría que los alojamientos con más reseñas tienden a tener un precio más alto. Sin embargo, de esta figura podemos deducir que los lugares con menor número de reseñas suelen ser los más caros. Esto podría deberse a que quizás están destinados a un público con más poder adquisitivo, que representa una menor proporción de la población.

En el scatter plot de la figura 6 se examina si hay alguna relación entre el precio y cuántos listados tiene un anfitrión. Teniendo en cuenta la dispersión de estos datos, se puede interpretar claramente que cuanto menos propiedades tengan los anfitriones hay una mayor variabilidad en los precios, encontrando que los más altos están con aquellos que solo poseen una propiedad. Esto podría explicarse porque o bien los que manejan propiedades de lujo (más caras) tienden a tener menos listados o porque los que gestionan más propiedades operan de manera más eficiente, lo que les permite ofrecer precios más competitivos. Es decir, al tener más propiedades bajo su control, pueden reducir costos y bajar precios para atraer a más huéspedes.

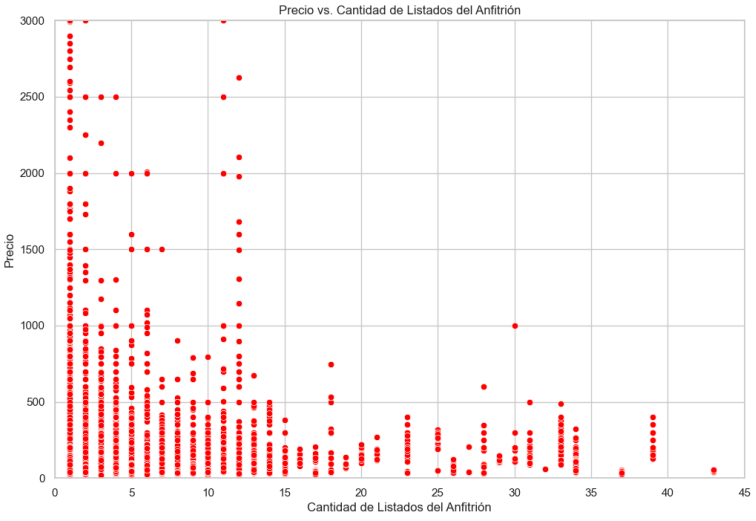
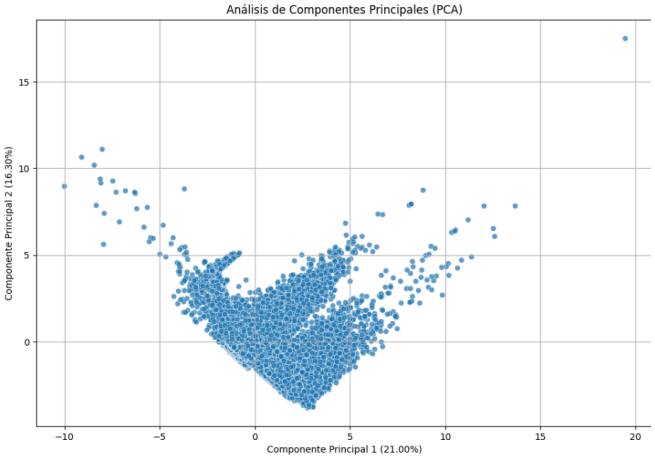


Figura 6: Precio y cantidad de listados del anfitrión

4. ANÁLISIS DE COMPONENTES PRINCIPALES



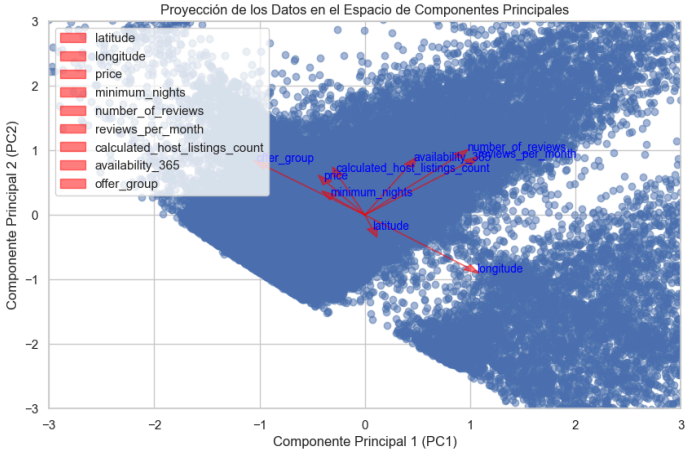


Figura 7.1 y 7.2: Análisis de componentes principales

En las figuras 7.1 y 7.2 se puede observar el resultado del Análisis de Componentes Principales. Cada punto representa observaciones del conjunto original proyectadas en los dos componentes principales (PC1 y PC2), mientras que los vectores o loadings indican, por una parte la contribución de cada variable a los componentes (largo del vector) y por otro lado la relación de dicha variable con los CP (dirección). Se puede observar que algunas variables, como por ejemplo *number_of_reviews* y *reviews_per_month* contribuyen a los mismos aspectos del conjunto de datos ya que tienen una similar longitud y dirección de vector.

En nuestro análisis se obtuvo que, al reducir el conjunto de datos a solo dos componentes principales, se pudo captar el 37,3% de la variabilidad. Los loadings obtenidos se encuentran en la Tabla 1, pero a grandes rasgos podemos concluir que las variables *longitude*, *reviews_per_month* y *number_of_reviews* son quienes tienen un peso más importante en el PC1, mientras que *offer_group* también pero con una relación inversa. Para el PC2, las variables con más peso también son *reviews_per_month* y *number_of_reviews*, sin embargo el peso es moderado.

Loadings:

	PC1	PC2
latitude	0.047341	-0.123875
longitude	0.666863	-0.490461
price	-0.244506	0.299837
minimum_nights	-0.206265	0.164297
number_of_reviews	0.604410	0.550306
reviews_per_month	0.670381	0.490888
calculated_host_listings_count	-0.171767	0.365737
availability_365	0.287561	0.460552
offer_group	-0.643167	0.452716

Tabla 1

5. PREDICCIÓN

Para poder hacer la predicción del precio primero eliminamos esa columna, y como no había ninguna altamente relacionada no se eliminó ninguna otra. Se tomó la decisión de reemplazar los valores de los outliers por las medianas (ya se había probado con eliminarlos

de la base pero arrojó mejores resultados este método) ya que si se mantenían el R^2 era de 0.1.

Para entrenar el modelo se utilizó el 70% (34216) de los datos, y la semilla 201. Se estableció a price como la variable dependiente en la base de entrenamiento (vector y). El resto de las variables fueron las variables independientes (matriz X). Se agregó la columna "intercepto" para permitir que el modelo ajuste a esta, el valor de la variable dependiente cuando todas las variables independientes son cero.

Para todos estos análisis se utilizaron las librerías: sklearn, pandas, numpy, matplot, seaborn. Se obtuvo un $R^2 = 0.42$, por lo tanto el modelo entrenado no es bueno para predecir.

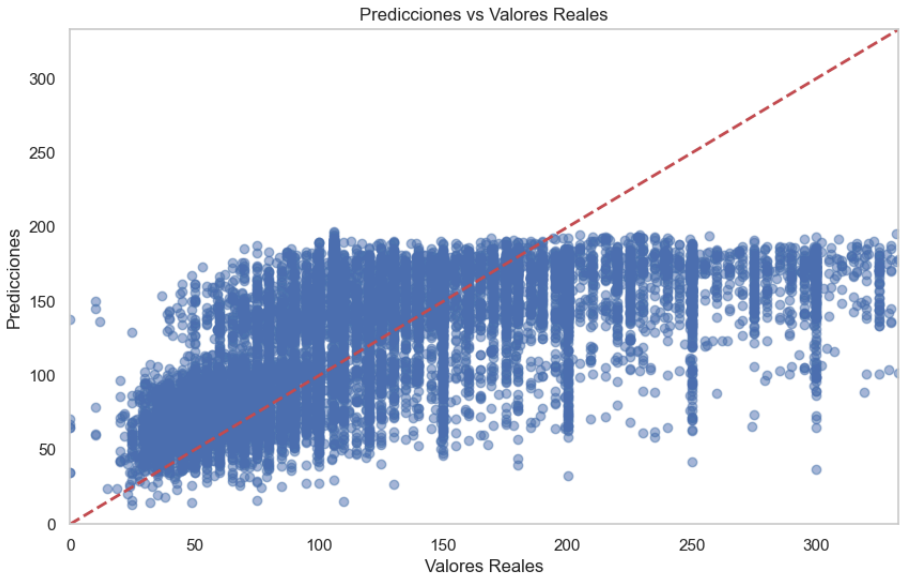


Figura 8: regresión lineal

BIBLIOGRAFÍA:

Nguyen, M. (2020). *A Guide on Data Analysis*. Bookdown. Disponible en: https://bookdown.org/mike/data_analysis/