2024

Trabajo Practico 2

PROGRAMACION

EUGENIA SAINI

INTRODUCCIÓN

Este trabajo busca interpretar datos proveniente de la empresa Airbnb en la ciudad de Nueva York. Para ello se utilizó el sistema Python, se importaron librerías y se desarrollaron códigos para implementar las tareas solicitadas.

PARTE I: LIMPIEZA DE LA BASE DE DATOS

EJERCICIO 1

En esta sección se realizó la limpieza de la base de datos que refirió a:

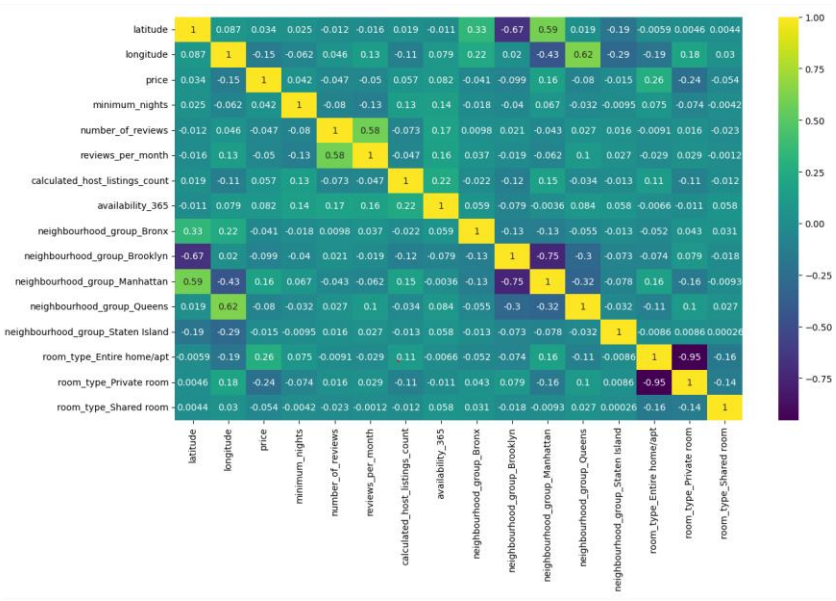
- **i)** **la identificación y eliminación de valores duplicados y de columnas sin información (Tabla 1 y Tabla 2),**
- **ii)** **la identificación y eliminación de columnas sin información (Tabla 3),**
- **iii)** **a la identificación de "missing data" y su tratamiento (Tabla 4, Tabla 5, Tabla 6 y Tabla 7).**
  Se reviso la bibliografía indicada y se analizó este caso en particular. Con base a la información de la Tabla 5, existe "missing data" (en total 10,064 filas) y en específico en las variables "reviews_per_month" (10,052) y en la variable "prices" (15). Para limpiar la base, se tomo como criterio eliminar las filas con missing data en "prices" por su baja significatividad sobre el total de los datos de la base (Tabla 6), y en el caso de "reviews_per_mont" se reemplazó el "missing data" (nan) con valores "0" dado que estaban vinculados a la variable "reviews" que, para esas filas, tambien eran "0", es decir no tenían reviews, por tanto, no tenías reviews_per_month). La nueva base de datos tiene ahora 48,880 filas (Tabla 7).
- **iv)** **identificación y eliminación de "outliers" (Tabla 8, Tabla 9 y Tabla 10)**
  Se verifico la presencia de outliers, considerándose estos como valores en las columnas que están a 3 desvíos de la media (de ambos lados), y se tomo como criterio eliminar las filas en donde los outliers estaban presentes en el 50% de las variables. Se creo un nuevo df "airbandb_3".
- **v)** **la transformación de variables (neighbourhood_group y room_type a variables numéricas) (Tabla 11)**
- **vi)** **la creación de una columna "Offer_group" que tenga la cantidad de oferentes por "Neighbourhood group" (Tabla 12)**

PARTE II. GRÁFICOS Y VISUALIZACIONES

EJERCICIO 2. MATRIZ DE CORRELACIÓN

La matriz de correlación (Tabla 13) identifica alta correlación positiva entre las variables "Latitude vs Manhattan", "longitud vs Queens", denotando la importancia de la localización y "reviews per month vs Number of reviews". Luego sigue "precio vs. Manhattan "y "precio vs room_type_entire home".

Figura 1. Matriz de correlación



EJERCICIO 3.

La proporción de oferentes por “Neighbourhood group” (Tabla 14 y Tabla 15) muestra una mayor proporción de ofertas en Manhattan (44,3%) y en el Brooklyn (41,1%) respecto a Queens (11,5%), Bronx (2,3%) y Staten Island (0,765). Si se Analiza por tipo de habitación (Tabla 16 y Tabla 17) se destaca mayor preferencia por casas o departamentos completos (51,9%) y habitaciones de uso exclusivo (45,6%), versus habitaciones compartidas (2,3%).

Figura 2. Proporción de ofertas por zona

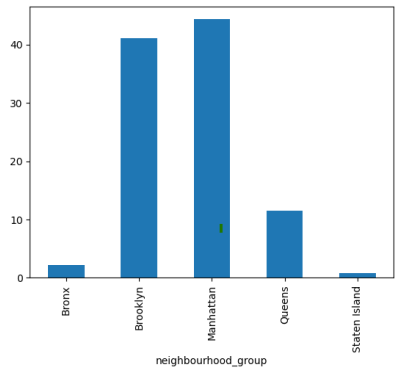
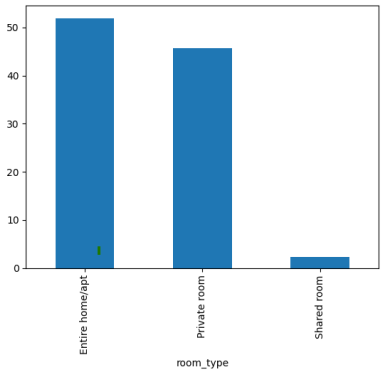


Figura 3. Proporción por tipo de habitación



EJERCICIO 4.

A continuación, se presentan tres figuras de histograma sobre la variable precios, en donde se observa distribución de “long tail” dado que existe un valor máximo de $10,000, mientras que la media es de $152,71 (Tabla 18). Se agruparon las medias de precio por zona (barrio) y por tipo de

habitación (Tabla 19 y Tabla 20) siendo Manhattan y Brooklyn las áreas de mayor precio ($196,8 y $124,4, respectivamente) y para casas o apartamentos completos ($211,7) versus habitaciones privadas ($89,8) y habitaciones compartidas ($70,12).

Figura 4. Histograma de Precios

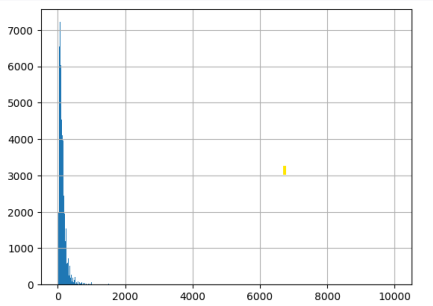
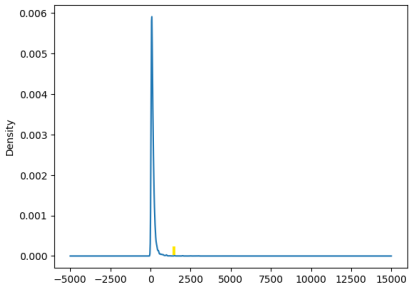
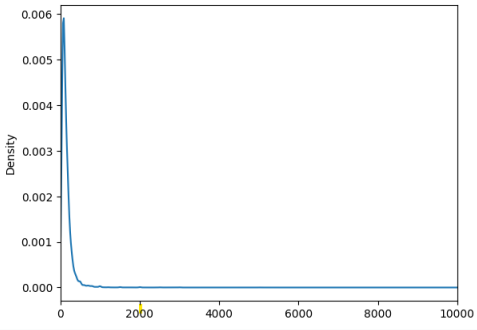


Figura 5. Histograma de Precios con función de Kernel



Se utilizo un bins = 500 y función de kernel “gaussiana

Figura 6. Histograma de Precios con función de Kernel



EJERCICIO 5.

Se realizaron gráficos scatters con las variables precios, barrio, cantidad mínima de noches, y localización (longitud y latitud) (Figuras 7, 8 y 9). Los precios mas altos se dan en Manhattan y Brooklyn, con un outlier en Queens. La mayor proporcion de noches se presenta en cantidades menores a 200 noches al año, y si se compara precios según latitud y longitud, se observa una mayor prporcion en la region de Manhattan y Brooklyn.

Figura 7. Precios por barrio

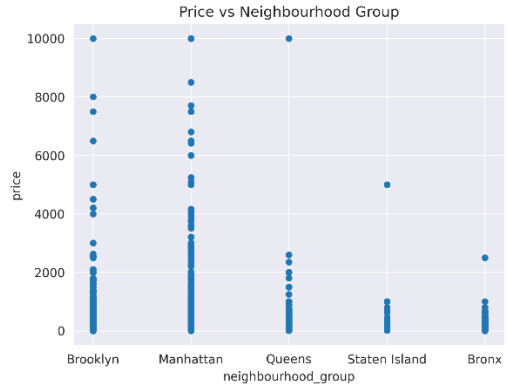
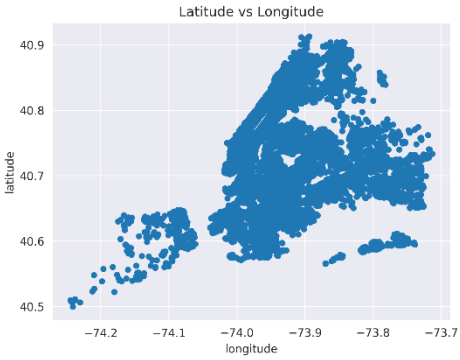


Figura 8. Precios según cantidad de noches

Figura 9. Precios según latitud y longitud



EJERCICIO 6.

Para trabajar en Componentes Principales, se importó la librería PCA, y resulta que la variación de los datos puede explicarse a través de dos CP (Figura 10, Figura 11 y Figura 12).

En el caso de la Figura 11, un gráfico “biplot de ACP” (Análisis de Componentes Principales), la longitud y orientación de las flechas indican qué características contribuyen más a la variabilidad en los datos. La longitud de la flecha (vector) para cada característica indica cuánto de la variabilidad explica esa característica en la dirección de los componentes principales. En este caso, si analizamos la longitud de la llecha y su contribución, se observa que una flecha más larga significa que la característica contribuye más a la variabilidad a lo largo de ese componente principal, como en el caso de las variables “Neighbourhood Group Manhattan”, “Neighbourhood Group Brooklyn” “latitude”, y “room type entire home/apt”. La flecha es más larga en la dirección del primer componente principal (CP1), esa característica explica más varianza a lo largo de ese componente y son las variables “Neighbourhood Group Brooklyn” y “room type entire home/apt”.

Respecto a la dirección de las flechas, las flechas que apuntan en una dirección similar indican que las características están positivamente correlacionadas, y las flechas que apuntan en direcciones opuestas, sugieren correlaciones negativas entre esas características. En la Figura 11 pueden observarse 4 cuadrantes que describen la relación entre las variables y su importancia en cuanto a explicar la variabilidad.

Por otro lado, el ángulo entre dos flechas proporciona información sobre la correlación entre las características. Un ángulo pequeño (o flechas que apuntan en la misma dirección) sugiere una correlación positiva, mientras que ángulos mayores (cerca de 180°) sugieren una correlación negativa. En la Figura 11 puede entonces agruparse por grupo de variables que poseen correlación positiva como por ejemplo las variables del cuadrante superior izquierdo, comparado con las variables que se presentan en los otros cuadrantes.

Figura 10. Numero de Componentes Principales

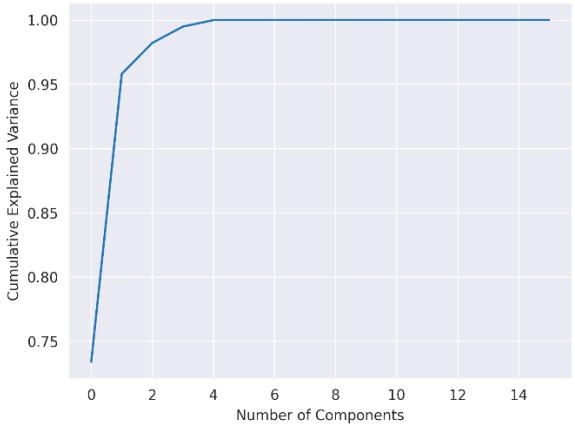


Figura 11. Numero de Componentes Principales

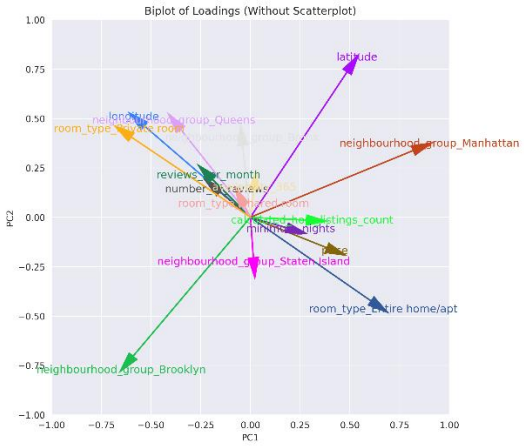
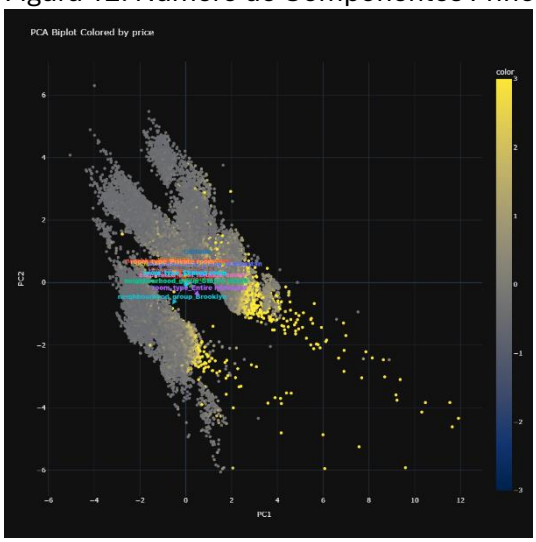


Figura 12. Numero de Componentes Principales



PARTE III. PREDICCIÓN Y VALIDACIÓN

EJERCICIO 7.

Se eliminaron de la base de datos las variables relacionadas al precio, “neighbourhood group Manhattan” y “room type Entire home/apt”.

EJERCICIO 8.

Se creo una base de prueba (test) y una de entrenamiento (train con 70% de los datos) y se tomó a la variable “precio” como variable dependiente en la base de entregamiento (vector Y), manteniendo el resto de las variables como independientes (matriz X). Se obtuvo un modelo de predicción.

Los resultados obtenidos al ejecutar X_train.shape, X_test.shape, Y_train.shape y Y_test.shape nos proporcionan información crucial sobre la estructura de nuestros datos divididos para entrenamiento y prueba de un modelo de machine learning. Estos números nos indican que el conjunto de entrenamiento (X_train) contiene 34,195 observaciones, cada una descrita por 13 características, mientras que el conjunto de prueba (X_test) tiene 14,656 observaciones con las mismas 13 características. Por su parte, los conjuntos de etiquetas (Y_train e Y_test) contienen los valores correspondientes a la variable objetivo que queremos predecir para cada observación en los conjuntos de entrenamiento y prueba, respectivamente. Esta división de los datos es fundamental en el proceso de machine learning, ya que permite entrenar el modelo en un conjunto de datos y luego evaluarlo en un conjunto independiente para obtener una estimación más precisa de su rendimiento en nuevos datos.

EJERCICIO 9.

Se implementó una regresión lineal y se estimaron los coeficientes, siendo:

- Variable Dependiente (Y): Precio
- Variables Independientes (X):

X_variable	coefficients
latitude	-187.4211172
longitude	-506.8589027
minimum_nights	-0.025598448
number_of_reviews	-0.244548879
reviews_per_month	-3.956957333
calculated_host_listings_count	-0.175184527
availability_365	0.195067199
neighbourhood_group_Bronx	-29.49508139
neighbourhood_group_Brooklyn	-62.15810332
neighbourhood_group_Queens	-31.61104916
neighbourhood_group_Staten Island	-194.4983776
room_type_Private room	-106.3809441
room_type_Shared room	-145.4829308

Los resultados de la regresión nos ofrecen una visión interesante sobre los factores que influyen en el precio de los alojamientos en Airbnb. Por un lado, observamos que la ubicación geográfica juega un papel fundamental. Los alojamientos situados en zonas con mayores valores de latitud y longitud tienden a ser más económicos. Asimismo, los vecindarios más populares, como Manhattan, suelen tener precios más elevados en comparación con otros. Por otro lado, características como el tipo de habitación también impactan significativamente en el precio. Los alojamientos privados y compartidos son generalmente más baratos que los apartamentos completos. Es interesante notar

que a medida que aumenta el número de reseñas positivas y la frecuencia con la que se reciben, el precio tiende a disminuir, lo que sugiere que los huéspedes valoran la reputación de los anfitriones. Además, una mayor disponibilidad a lo largo del año se asocia con precios más altos, posiblemente debido a una mayor demanda en esos casos.

En resumen, este modelo de regresión nos permite identificar una serie de factores clave que influyen en la determinación del precio de los alojamientos en Airbnb. La ubicación geográfica, el tipo de habitación, el número de reseñas y la disponibilidad son solo algunos de los elementos que parecen jugar un papel importante. Sin embargo, es fundamental recordar que estos resultados se basan en el conjunto de datos específico utilizado en el análisis y pueden variar al considerar otros factores o diferentes mercados.

EJERCICIO 10.

En la siguiente table se presentan el Error Cuadrático Medio (MSE), la Raíz del Error Cuadrático Medio (RMSE), y el Error Absoluto Medio (MAE).

Métrica	Valor
MSE	49418.417038
RMSE	222.302535
MAE	73.441332

Cuando calculamos estas medidas dentro de la muestra (in-sample), estamos evaluando el desempeño del modelo en los datos que se usaron para entrenarlo. En este caso, es común que los valores de MSE, RMSE y MAE sean bajos, ya que el modelo está ajustado específicamente a estos datos y, en consecuencia, puede captar las relaciones presentes en ellos con precisión. Sin embargo, un buen ajuste dentro de la muestra no siempre indica que el modelo generalizará bien en datos nuevos. Un modelo que tiene un ajuste perfecto dentro de la muestra podría estar sobreajustado (overfitting), capturando ruido específico de los datos de entrenamiento en lugar de patrones generales.

Por otro lado, cuando calculamos estas medidas fuera de la muestra (out-of-sample) – es decir, en datos que el modelo no ha visto antes – evaluamos la capacidad del modelo para generalizar a datos nuevos. En esta situación, los valores de MSE, RMSE y MAE tienden a ser más altos que los obtenidos dentro de la muestra, ya que el modelo no está "entrenado" específicamente en estos datos y se enfrentará a un escenario donde debe aplicar los patrones aprendidos de manera general. Si los valores de error fuera de la muestra son significativamente más altos que los errores dentro de la muestra, podría indicar que el modelo no generaliza bien y está sobreajustado.

En conclusión, comparar las medidas de bondad de ajuste dentro y fuera de la muestra es fundamental para evaluar el rendimiento real del modelo. Un modelo que mantiene errores bajos en ambas muestras sugiere una buena capacidad de generalización, mientras que un modelo que presenta grandes diferencias podría estar sobreajustado y requerir ajustes para mejorar su capacidad predictiva en datos no vistos.

Los resultados obtenidos para las métricas de evaluación del modelo sugieren que, si bien el modelo logra realizar predicciones, estas presentan un nivel de error considerable. Un MSE de 49418.417038 indica una diferencia cuadrática media significativa entre los valores predichos por el modelo y los valores reales. El RMSE, que representa la raíz cuadrada del MSE, nos ofrece una medida del error en la misma escala que nuestra variable objetivo, y en este caso, un valor de 222.302535 sugiere un error promedio de esta magnitud. Por último, el MAE, que calcula el error absoluto medio, nos indica que, en promedio, las predicciones del modelo se desvían en 73.441332 unidades de los valores reales.