Lic. en Ciencias del Comportamiento

Ciencia de Datos

Trabajo Práctico 2

Comisión

19:00 Martes

Juan Costa

DNI: 44.640.563

Legajo: 33569

Sol Ruiz Coines

DNI: 44553147

Legajo: 33513

Facundo Kanonicz

DNI: 43724736

Legajo: 33574

A partir de la base de datos sobre oferentes de Airbnb en la ciudad de Nueva York, realizaremos un análisis con el objetivo de construir un modelo de predicción de precios. Se realizará una limpieza de la base de datos, diferentes gráficos que permitan visualizar la correlación entre las distintas variables, y por último, construiremos un modelo de machine learning de predicción de precios para poder entender qué factores influyen en la renta en Airbnb.

Parte I: Limpieza de la base

En primer lugar, llevamos a cabo la limpieza de la base de datos. La base contaba con 48905 datos, e incluía 16 variables relacionadas al alquiler de propiedades: *'id', 'name', 'host_id', 'host_name', 'neighbourhood_group', 'neighborhood', 'latitude', 'longitude', 'room_type', 'price', 'minimum_nights', 'number_of_reviews', 'last_review', 'reviews_per_month', 'calculated_host_listings_count' y 'availability_365'*.

Para comenzar, eliminamos todos los valores duplicados con el comando *.drop_duplicates()*, que remueve del DataFrame todos los valores repetidos. Posteriormente, descartamos las columnas que no tenían información de interés para el modelo de predicción de precios, utilizando el comando *.drop(columns = [...]). Determinamos que las columnas irrelevantes eran 'host_id', 'host_name', 'last_review' y 'reviews_per_month'*, ya que no proporcionan valor para los indicadores clave del análisis. Luego de eliminar las columnas de poco interés, evaluamos el problema de los missing values, y tomamos la decisión de realizar una limpieza de listwise deletion, a partir de lo establecido en el libro "A Guide on Data Analysis" (Nguyen, 2020). Consideramos que esta técnica era la más adecuada debido a su fácil implementación, y contemplamos que los missing values eran Faltantes Completamente al Azar (MCAR). Esto significa que la falta de datos es completamente aleatoria y no está relacionada con los valores observados ni no observados. A partir de esto, utilizamos listwise deletion para que no se produzcan parámetros y errores estándar sesgados. Con el comando *.dropna()*, eliminamos todos los datos que contengan un valor vacío (NaN) en alguna de las columnas.

Respecto a los outliers o valores que no tienen sentido, primero graficamos en boxplots para cada columna numérica, para identificar visualmente la densidad de datos outliers. A partir de los gráficos, tomamos la decisión de transformar las variables numéricas en logaritmo natural, para aproximar la muestra a una distribución normal, y reducir la heterocedasticidad, así disminuyendo su varianza. Luego de transformar las variables, utilizamos el método de IQR para la eliminación de outliers. Recurrimos a una multiplicación de 2.5 por el IQR para realizar un análisis menos estricto, a fin de no eliminar una gran cantidad de datos.

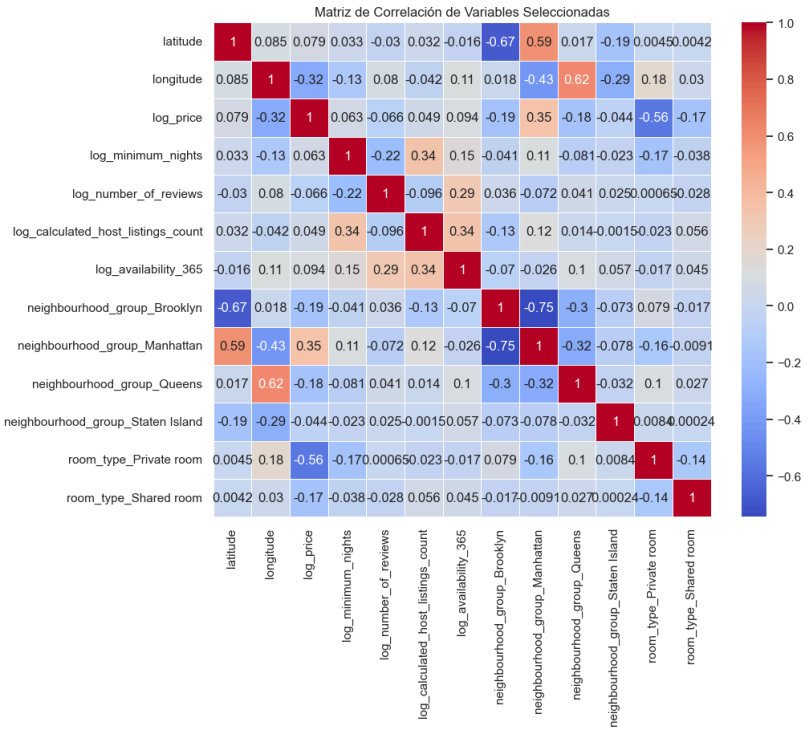
Por último, transformamos las variables *'neighbourhood_group'* y *'room_type'* a variables numéricas con el comando *LabelEncoder(),* y le asignamos un valor a cada uno de los datos de cada columna. Finalmente, con la ayuda de los comandos *.groupby()* y *.merge()*, creamos una nueva columna con la cantidad de oferentes por grupo, y la llamamos *oferentes_por_grupo*.

Parte II: Gráficos y visualizaciones

Para la segunda parte del trabajo, visualizamos las variables principales y la relación entre sí de las mismas. Fue un análisis exploratorio ad hoc.

En primer lugar, se realizó la matriz de correlación para comprender los coeficientes y la relación a simple vista de la fuerza de relación entre variables. En una primera observación, se entiende que las siguientes variables están correlacionadas de manera negativa en más de un r=0.5.

- (“*private_rooms*” y “*log_price*”)
- (“*Brooklyn* “ y “*latitude*”)
- (“*availability_365*” y “*latitude*”)
- (“*Brooklin*” y “*Manhattan*”)



Por un lado, podemos preguntarnos si hay una relación inversamente proporcional entre *private rooms* y *pricing*. Por otro lado, podemos observar comportamientos significativos con respecto a la ubicación de las viviendas. Luego, evaluaremos la proporción de oferentes por barrio y por habitación encontrando los siguientes resultados.

Manhattan	0.443026
Brooklyn	0.411162
Queens	0.115914
Bronx	0.022266
Staten Island	0.007633

Figura 1

En la figura 1.1. Se observa la proporción de oferentes según el barrio; Manhattan y Brooklyn teniendo más del 85% total de oferentes dentro de la plataforma. Luego se observa que está correlacionado con los precios más altos, haciendo alusión a que sí son los lugares más turísticos. En la figura 1.2. se observa cómo no existe mucha oferta de cuartos compartidos. *¿Será porque el target que apunta NY como turistas no es el tipo de usuario que use shared rooms?*

Con respecto a estos números, lo que podemos concluir es que en Manhattan y Brooklyn se oferta más, por lo que asumimos que se demanda más también. *¿Estará correlacionado con los precios?*

Entire home/apt	0.519749
Private room	0.456532
Shared room	0.023719

Figura 2

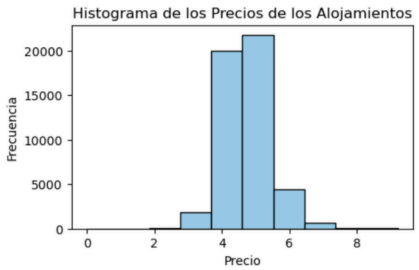


Figura 3

Los precios estandarizados son relativamente normales, y esto implica que “el mercado” de precios funciona. Si bien existen datos fuera de lo normal para ambos lados, ya que existen tanto departamentos sumamente lujosos como otros con varias carencias, la mayoría de las propiedades se sitúan en un rango de precios predecible y coherente con las características del inmueble y la ubicación. (Figura 1.3.)

El precio promedio por vivienda es de aproximadamente 153 USD por noche. No obstante, al analizar más detenidamente los precios, se observa que Manhattan y Brooklyn son las áreas que más influyen en el incremento del promedio general en Nueva York. Por su parte, en Staten Island, dado que es una isla, las propiedades resultan ser aún más costosas que en el Bronx y Queens, a pesar de la mayor oferta en estos últimos. Staten Island, por lo tanto, parece mantener un carácter más exclusivo.

Por otro lado, podemos observar una diferencia significativa entre el alquiler de un departamento completo y una habitación compartida (Figura 1.4.). Es importante considerar que los departamentos completos están destinados, presumiblemente, a alojar a más personas que las habitaciones privadas, lo que influye en la variación de los precios.

```
Media de precio por Neighbourhood group:
neighbourhood_group
Bronx            87.425551
Brooklyn        124.403962
Manhattan       196.880266
Queens           99.536900
Staten Island   114.812332
Name: price, dtype: float64

Media de precio por Tipo de Habitación:
room_type
Entire home/apt    211.793480
Private room        89.790568
Shared room         70.075928
```

Figura 4

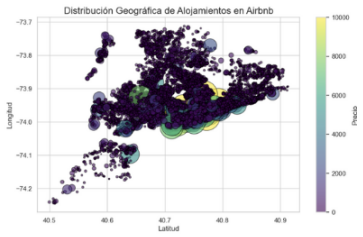


Figura 5

Luego, profundizamos en la relación entre precio y latitudes en la Figura 1.5.. Cada observación muestra una propiedad distinta, lo que permite ver un patrón que se asemeja al "mapa de Nueva York": [Mapa de NY]

La visualización del mapa pone en evidencia la gran cantidad de oferentes en las distintas zonas, lo cual coincide con lo presentado en tablas anteriores. Consideramos relevante observar la variación de los precios según el barrio. Manhattan y Brooklyn destacan como los barrios con las publicaciones más elevadas. Un aspecto llamativo es la presencia de un gran círculo en la parte inferior, en Staten Island. Aunque inicialmente podría interpretarse como un outlier, en realidad refleja la zona de mayor costo dentro de la isla.

A su vez, buscamos visualizar la relación entre el precio por noche y la cantidad de *reviews* por mes que tenía cada publicación. Esto lo buscamos con el fin de entender la efectividad de las reviews. La hipótesis que planteamos es que un departamento con buenas reviews generará mayor confianza entre los potenciales inquilinos, lo que permitiría justificar un aumento en el precio.

Figura 6

En el gráfico (Figura 1.6.) podemos observar que la mayoría de los departamentos ofrecidos tienen reviews. Claramente, hay casos donde los precios son muy altos o muy bajos en comparación con la media, y estos suelen tener 0 reviews, lo que los convierte en casos aislados o atípicos.

Esta información es valiosa para la aplicación Airbnb ya que le permite entender la importancia de las reviews para los usuarios. La mayor demanda se concentra en propiedades con precios dentro de la media y que cuentan con reviews, lo que sugiere que las evaluaciones juegan un papel clave en las decisiones de los usuarios. Consideramos que Airbnb debería incentivar a sus anfitriones a ofrecer un excelente servicio con el fin de obtener buenas reviews. Si lo logran, es más probable que los clientes los elijan nuevamente, incrementando su éxito en la plataforma.

Luego de estas visualizaciones, y de comprender mejor patrones de conducta de los oferentes, se realizó un análisis de componentes principales (PCA), con el fin de entender cuánto explican los

dos primeros componentes, como estan armados los *loadings* y qué variables son las que más pesan en cada varianza explicada.

PC1 está fuertemente asociada a factores relacionados con la cantidad de oferentes y la duración mínima de estadía, mientras que PC2 está más relacionada con la disponibilidad, las reseñas y la ubicación geográfica (longitud). Esto se logra observar por los *loadings*. (Figura 1.7.)

```
Varianza explicada por PC1: 23.27%
Varianza explicada por PC2: 20.29%
Loadings (correlaciones con las componentes principales):
                                      PC1       PC2
latitude                         0.063351  0.005786
longitude                       -0.272806 -0.471983
log_minimum_nights               0.560822  0.052699
log_price                        0.324772  0.294583
log_number_of_reviews           -0.189877 -0.524274
log_calculated_host_listings_count  0.579042 -0.239693
log_availability_365             0.360864 -0.596093
```

Figura 7

Algunas variables tienen altas cargas en ambos componentes, esto indica que esas variables originales están contribuyendo a ambos. Por ejemplo, log price tiene una contribución notable en ambas (PC1: 0.325, PC2: 0.295), lo que podría sugerir que esta variable tiene una influencia en ambos componentes, aunque los componentes en sí sean no correlacionados.

El siguiente gráfico (Figura 1.8.) muestra cómo las observaciones en los datos se proyectan en un espacio reducido de dos dimensiones, representado por las componentes principales PC1 y PC2. Estos dos componentes explican un 43.56% de la varianza total (PC1 explica el 23.27% y PC2 el 20.29%).

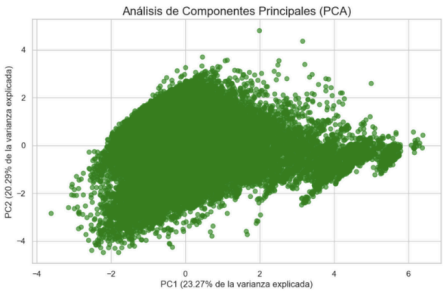


Figura 8

La mayor densidad de puntos se concentra cerca del origen(0,0), lo que indica que la mayoría de las observaciones están agrupadas en torno a valores promedio. Sin embargo, también se observa una dispersión de algunos datos hacia los extremos, lo que podría sugerir la presencia de outliers o casos particulares que se desvían del comportamiento general. Además, la forma ligeramente ovalada que toman los puntos podría indicar una correlación no lineal entre las componentes, a pesar de que estas son ortogonales por definición.

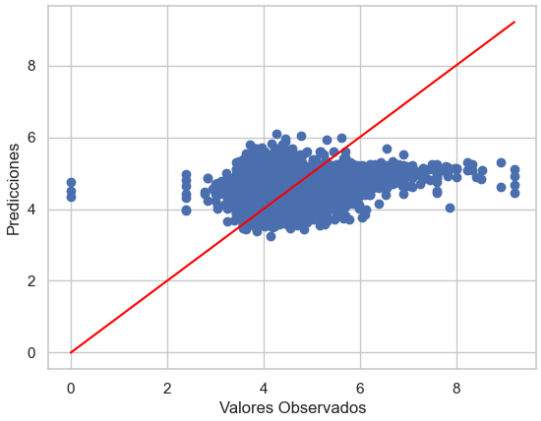
Parte III: Predicción

En esta parte del trabajo buscamos predecir los precios de los alojamientos en base a todas las otras variables numéricas. Con este objetivo, dividimos la base de datos en una base de entrenamiento, que contiene el 70% de los datos, y una base de prueba, con el 30% restante del total. La variable dependiente a predecir fue el logaritmo del precio por noche, mientras que las independientes fueron la versión logarítmica de: la latitud, la longitud, el mínimo de noches requerido para reservar, el de anuncios/alojamientos que un determinado anfitrión tiene en la plataforma Airbnb y el número de días al año que el anuncio está disponible para reservar.

En este sentido, se creó un modelo de regresión lineal que fue entrenado con la base de datos para el entrenamiento y se generaron las predicciones a partir de las variables dependientes de la base de prueba. Luego, calculamos algunos resultados:

- El valor de R^2 es de 0.138
- El intercepto del modelo es igual a -433.8
- Los coeficientes del modelo tienen los siguientes valores: [0.0, 1.366, -5.177, -0.017, -0.042, -0.022, 0.049]

- El Error cuadrático medio es igual a 0.41
- La raíz del error cuadrático medio es igual a 0.64
- El error absoluto medio es igual a 0.5



También realizamos una validación cruzada con 5 pliegues para seguir entrenando y evaluando el modelo, a partir de la cual calculamos el MSE para cada pliegue (0.41, 0.42, 0.41, 0.42, 0.42) y el MSE promedio (0.415). A su vez, calculamos el puntaje del modelo utilizando el comando *test_score*, y obtuvimos un resultado de 0.15.

El gráfico representa una comparación entre los valores reales u observados y las predicciones del modelo, con una línea de referencia de la regresión lineal.

A partir del análisis de estos resultados, concluimos que este modelo lineal tiene un bajo nivel de acierto en la predicción del precio según las variables seleccionadas. Esto se refleja tanto en los valores bajos del R^2 y el puntaje del modelo, como en los coeficientes cercanos a 0. A su vez, el gráfico muestra una gran dispersión de los datos, lo que también indica que el modelo no está haciendo predicciones precisas. Si bien en los valores observados entre 4 y 6 parecen poder ser bien predichos, los demás valores se encuentran alejados de la línea de referencia, por lo que presentan errores de predicción.

Una posible solución sería identificar un mejor ajuste del modelo considerando otras variables dependientes.

---

1. APA (7th edition):

Nguyen, M. (2020). *A Guide on Data Analysis*. Bookdown.

https://bookdown.org/mike/data_analysis/