BASE DE DATOS SOBRE OFERENTES DE AIRBNB EN LA CIUDAD DE NUEVA YORK

PARTE I - LIMPIEZA DE LA BASE DE DATOS

El proceso de limpieza y transformación del dataset de Airbnb se llevó a cabo en varias etapas. En primer lugar, se verificó la existencia de filas duplicadas, encontrándose 10 de ellas, las cuales fueron eliminadas. A continuación, se eliminaron columnas no relevantes, como 'id', 'name', 'host_id', 'host_name' y 'last_review'.

En cuanto a la detección de valores atípicos (outliers), se analizó la distribución de densidad de las variables de interés. Se observó que las variables “number_of_reviews”, “reviews_per_month” y “availability_365” presentaban en su mayoría valores de 0, lo que concentraba la distribución en dichos valores (ver Figura 1).

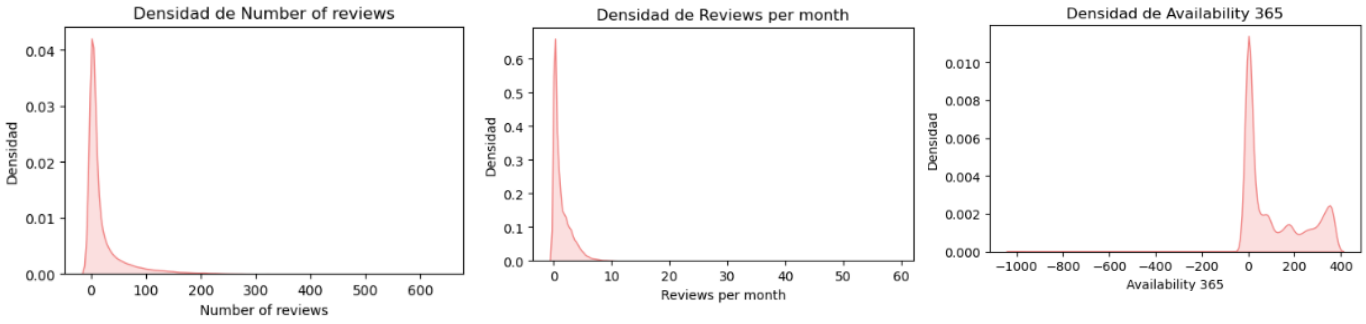

Figura 1. Distribución de densidad de Number of reviews, Reviews per moth y Availability 365

Mediante la aplicación de una transformación logarítmica en estas tres variables, se pudo mejorar considerablemente la distribución de los datos (ver Figura 2).

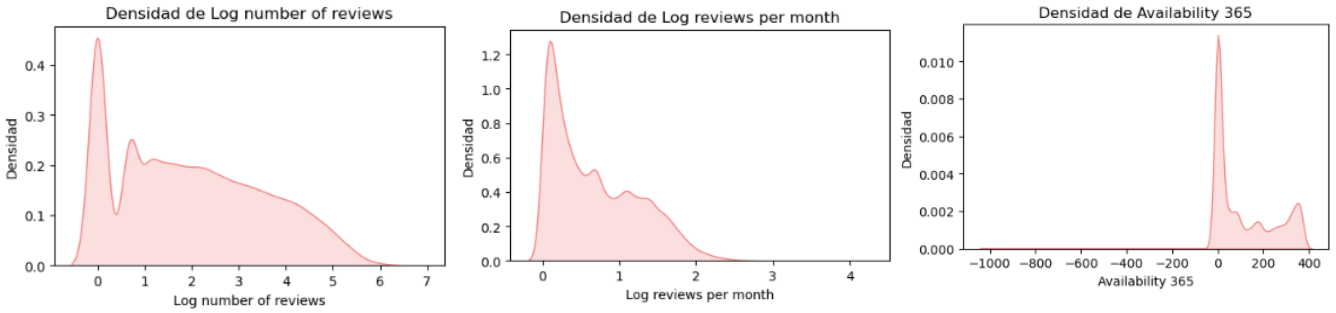

Figura 2. Distribución de densidad de Number of reviews, Reviews per month y Availability 365 luego de las transformaciones logarítmicas

Posteriormente, se identificaron y eliminaron los valores atípicos utilizando el método del rango intercuartílico (IQR) en las columnas relevantes: “log_number_of_reviews”, “log_reviews_per_month”, “latitude”, “longitude” y “price”. Además, se eliminaron los registros en los que el precio era igual a 0. Como resultado, se eliminaron un total de 15.548 datos, pasando de tener 48.895 a 33.347 datos.

Asimismo, se transformaron las columnas 'neighbourhood_group' y 'room_type' en variables numéricas mediante *Label Encoding* y se añadió una nueva columna que cuenta la cantidad de oferentes por 'neighbourhood_group'. Por último, los valores faltantes en las columnas 'reviews_per_month' y 'price' fueron imputados utilizando la mediana.

PARTE II - GRÁFICOS Y VISUALIZACIONES

EJERCICIO 2 - MATRIZ DE CORRELACIONES

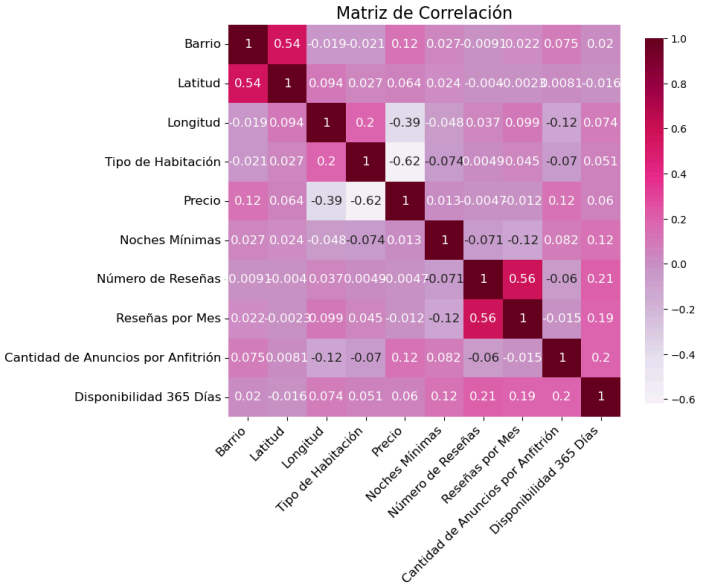

Figura 3. Matriz de correlación del dataset

En la matriz de correlación (ver Figura 3), por ejemplo observamos una correlación moderada entre el número de reseñas y las reseñas por mes (corr = 0.56). Asimismo, existe una correlación moderada entre el barrio o la ubicación del alojamiento y la latitud (corr = 0.54). Por otro lado, encontramos una correlación muy baja entre el precio y las noches mínimas (corr = 0.013), así como entre el precio y la disponibilidad del anuncio (corr = 0.06).

EJERCICIO 3 - PROPORCIÓN DE OFERENTES

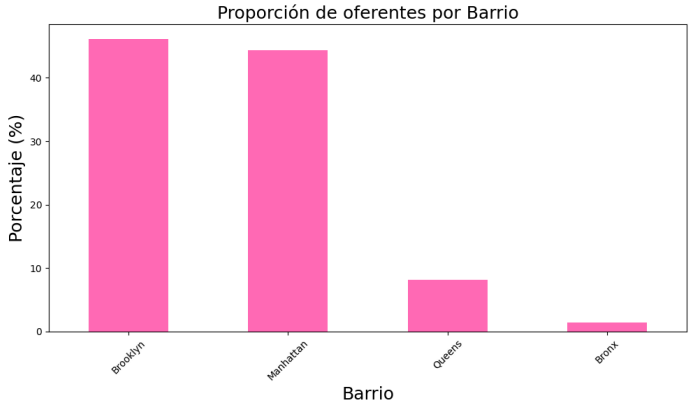

Figura 4. Proporción de oferentes por barrio

En cuanto a la proporción de oferentes por barrio, la **Figura 4** muestra como Brooklyn y Manhattan representan el 46.16% y el 44.31% de las ofertas, respectivamente, sumando un total del 90.5%, lo que indica su alta popularidad entre los anfitriones y la intensa competencia en estas zonas. Por otro lado, Queens y el Bronx tienen una representación mucho menor, con un 8.14% y un 1.40%.

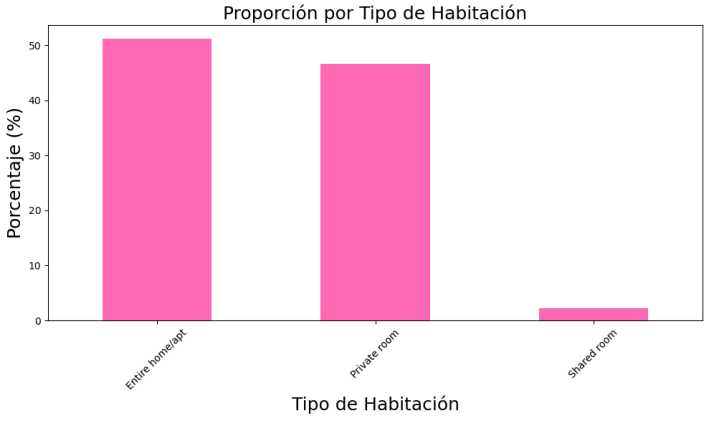

Figura 5. Proporción de oferentes por Tipo de Habitación

En cuanto a la proporción por tipo de habitación, la **Figura 5.** muestra cómo más del 51% de las ofertas son apartamentos completos, lo que indica una clara preferencia por la privacidad y el espacio. Las "Private rooms" representan el 46.65% de la oferta, reflejando un equilibrio entre costo y privacidad. Sin embargo, las "Shared rooms" son significativamente menos populares, ocupando solo el 2.20%, lo que sugiere una mayor preocupación por la privacidad en la actualidad.

EJERCICIO 4 - HISTOGRAMA DE LOS PRECIOS DE LOS ALOJAMIENTOS

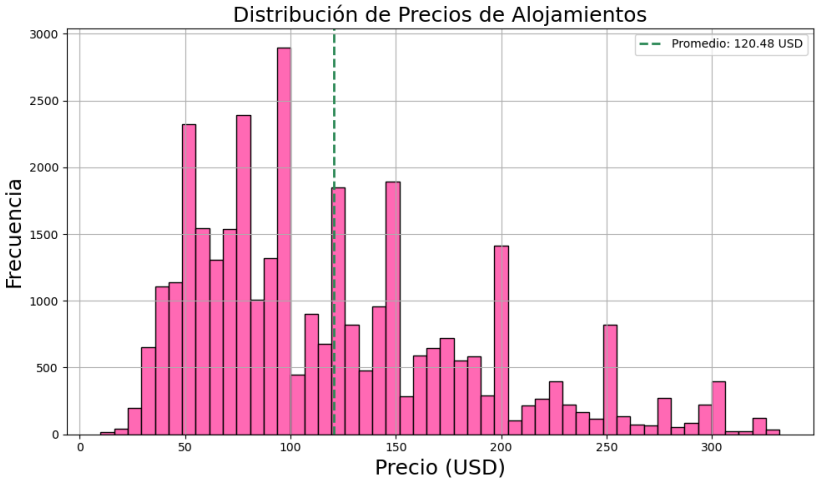

Figura 6. Histograma de la distribución de los precios de los alojamientos

La **Figura 6** muestra la distribución de los precios de los alojamientos en la plataforma Airbnb y muestra que la mayoría de los alojamientos se concentran en precios entre 50 y 150 USD, con un notable pico alrededor de los 100 USD. Además, se puede ver que hay una presencia significativa de precios más altos, con menos alojamientos en los rangos de 200 a 300 USD, lo que indica que aunque hay opciones premium o de lujo, estas son considerablemente menos comunes.

El precio mínimo de los alojamientos en AirBnB es de 10.0 USD, el máximo es de 332.0 USD y el promedio es de 120.48 USD. En cuanto a la media de precio por barrio, el Bronx tiene una media de 69.53 USD, Brooklyn de 106.46 USD, Manhattan de 142.34 USD y Queens de 89.58 USD.

A partir de estos datos podemos interpretar que Manhattan tiene el precio promedio más alto debido a su alta demanda turística y su prestigio como centro financiero y cultural; Brooklyn, al ser una opción popular para viajeros que buscan cercanía a Manhattan sin los altos costos, tiene un precio moderado; Queens y el Bronx tienen precios más bajos, lo que refleja el perfil residencial y una menor presión turística que tienen en comparación con Manhattan y Brooklyn.

En relación al precio según el tipo de habitación que ofrecen los alojamientos cuando se trata de una casa o departamento completo la media es de 161.16 USD, si es una habitación privada la media es de 78.86 USD y si es una habitación compartida la media es de 56.77 USD

EJERCICIO 5 - SCATTER PLOTS

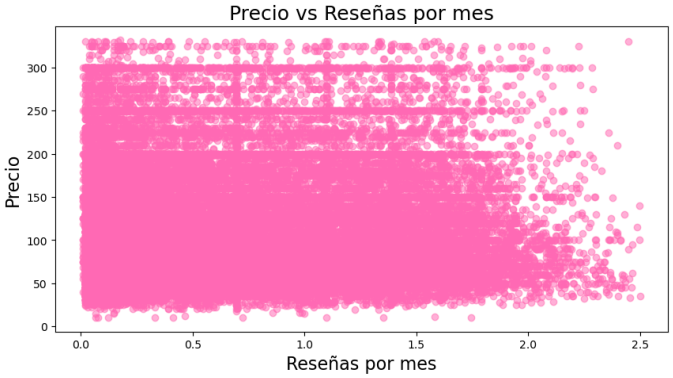

Figura 7. Scatter plot que compara reseñas por mes y precio

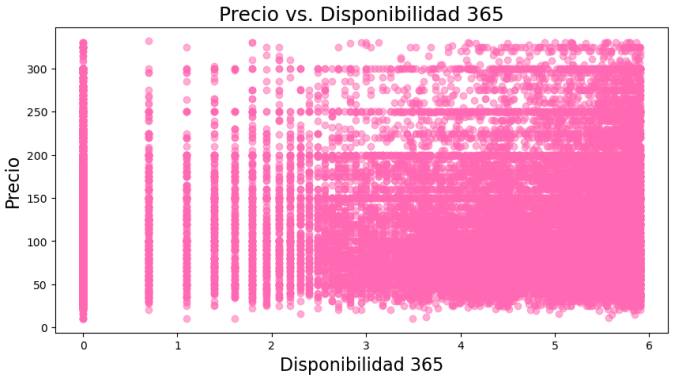

Figura 8. Scatter plot que compara número de disponibilidad de días al año del anuncio y precio

En la **Figura 7**, se observa que el número de reseñas es relativamente similar en los diferentes niveles de precios. Sin embargo, se puede identificar una tendencia en la que los alojamientos con precios más bajos tienden a tener un mayor número de reseñas debido a la mayor concentración de datos se encuentra por debajo precio = 200, mientras que aquellos con precios más altos suelen recibir menos reseñas.

En la **Figura 8,** podemos observar una gran concentración de puntos en la sección de 2 a 6 días de disponibilidad del anuncio. Esto indica que muchos alojamientos tienen una disponibilidad del anuncio por un corto período del año. Sin embargo, los precios en esta área varían notablemente, lo que sugiere que la disponibilidad del anuncio no varía en relación al precio.

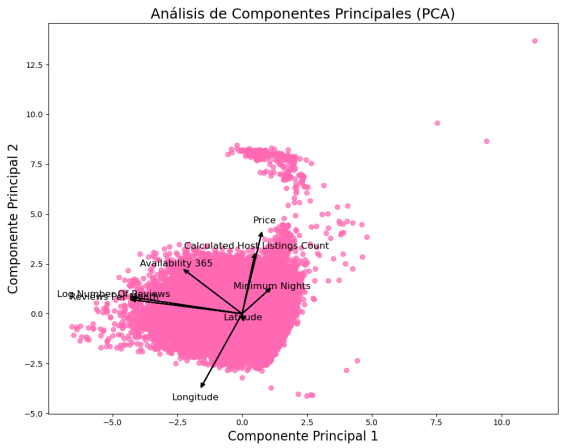

Figura 9. Análisis de Componentes Principales en dos dimensiones

EJERCICIO 6 - ANÁLISIS DE COMPONENTES PRINCIPALES

Los dos primeros componentes principales (CP1 y CP2) explican el **40.75%** de la varianza total en los datos. Variables como *Reseñas por mes* y *Número de reseñas* podrían estar correlacionadas positivamente si sus flechas apuntan en la misma dirección. Además, en el componente principal 1 number of reviews tiene un valor negativo (-0.6369) y *Precio* tiene un valor positivo (0.1090). Esto sugiere que en el primer componente podrían estar inversamente correlacionadas.

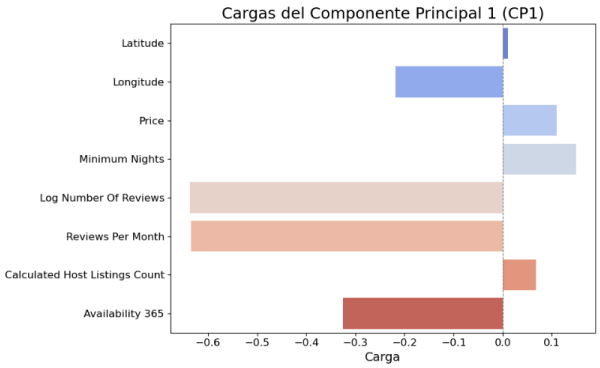

Figura 10. Loadings componente 1

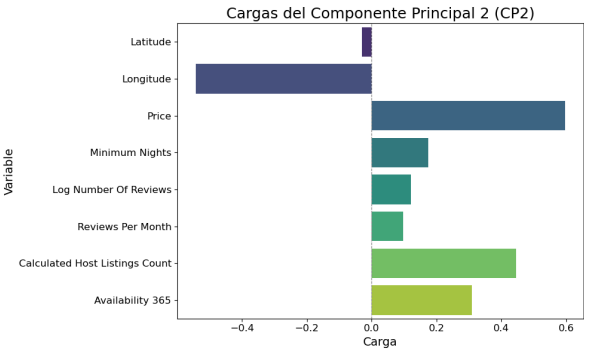

Figura 11. Loadings componente 2

Las cargas de los componentes muestran que number of reviews y reviews per month tienen influencias negativas significativas en el CP1, lo que indica que los listados con menos reseñas son prominentes. En el CP2, price tiene una carga positiva alta, sugiriendo que los listados más caros están bien representados en este componente.

PARTE III - PREDICCIÓN

EJERCICIO 9 - REGRESIÓN LINEAL

	Variable	Coeficiente
0	Latitud	95.38
1	Longitud	-595.5
2	Noches Mínimas	-0.23
3	Número de reseñas	-0.02
4	Reseñas por Mes	-3.22
5	Conteo de Anfitriones	0.06
6	Disponibilidad durante el año	0.02
7	(Log) Número de Reseñas	-2.08
8	(Log) Número de Reseñas por Mes	12.89
9	(Log) Disponibilidad 365	2.65
10	Barrio	8.25
11	Tipo de Habitación Codificado	-70.05
12	Cantidad de Oferentes	0

Tabla 1. Coeficientes de las variables a partir del modelo de regresión lineal

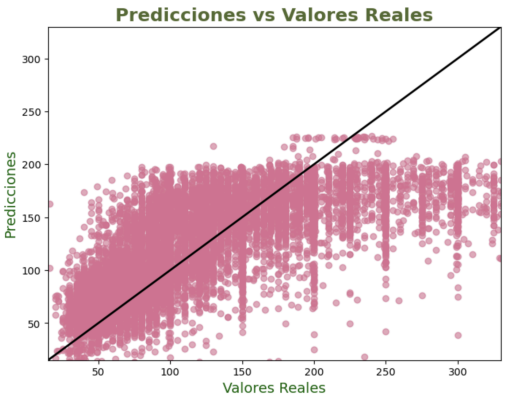

Figura 12. Valores predichos vs. Valores reales del modelo de regresión lineal

Comentarios

Los coeficientes en la **Tabla 1** muestran el impacto de cada una de las variables independientes (X) sobre el precio (Y), manteniendo constante el resto de las variables. Las variables que impactan más sobre el precio son aquellas que cuentan con un coeficiente alto, en este caso la longitud (-595.49) tiene un gran impacto negativo sobre el precio, esto se debe a que los departamentos en las zonas con mayor latitud (oeste) tienen precios más

bajos (disminuyen). El tipo de habitación (-70.04) también es muy relevante, ya que el precio varía significativamente de un tipo de habitación a otra. La latitud (95.37) también presenta impactos sobre el precio, aunque menos fuertes que la longitud, a medida que la latitud aumenta, los precios también.

El MSE del modelo indica las diferencias entre los valores reales y las predicciones del modelo, por lo que lo ideal sería obtener un MSE bajo, así las predicciones se acercan lo más posible a lo que ocurre en la realidad. En este caso, como se puede observar en la **Tabla 2**, el MSE es de 2290.64, que se podría considerar alto, por lo que la predicción del modelo no es muy adecuada.

	Métrica	Valor
0	MSE	2290.64
1	R^2 Train	0.49
2	R^2 Test	0.49

Tabla 2. Métricas del modelo de regresión lineal

Los R^2 del conjunto de entrenamiento y del conjunto de prueba son 0.49 y 0.49 respectivamente, e indican que el modelo explica aproximadamente el 49% de la variabilidad entre los precios del conjunto de entrenamiento y el 49% de la variabilidad de precios del conjunto de prueba. Esto quiere decir que su capacidad de explicar la variabilidad es moderada, y se podría mejorar. Para mejorar este modelo se podría aplicar un modelo no lineal o transformar las variables de otras formas.

En la **Figura 12** podemos interpretar que al principio las predicciones se parecen a los valores reales, pero a medida que los valores crecen, la predicción se vuelve cada vez menos precisa. Esto refleja una falta de precisión creciente en la capacidad del modelo en capturar la relación cuando los valores son mayores. La distribución de puntos tiene un patrón parecido al de una función logarítmica, por lo que se puede notar que un modelo lineal no sería el adecuado para hacer predicciones sobre este conjunto de datos.