Trabajo Práctico 2

Otoño 2024

Big Data

Integrantes: Martina Murga, Valentina Benitez, Benicio García Oliver

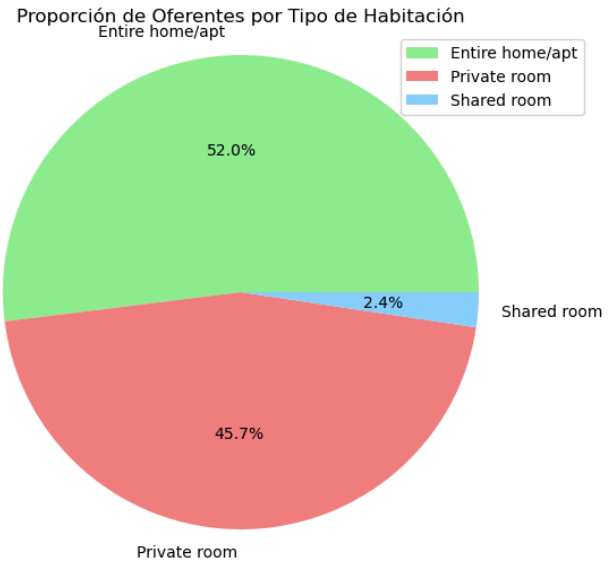
Parte I: Limpieza de la base

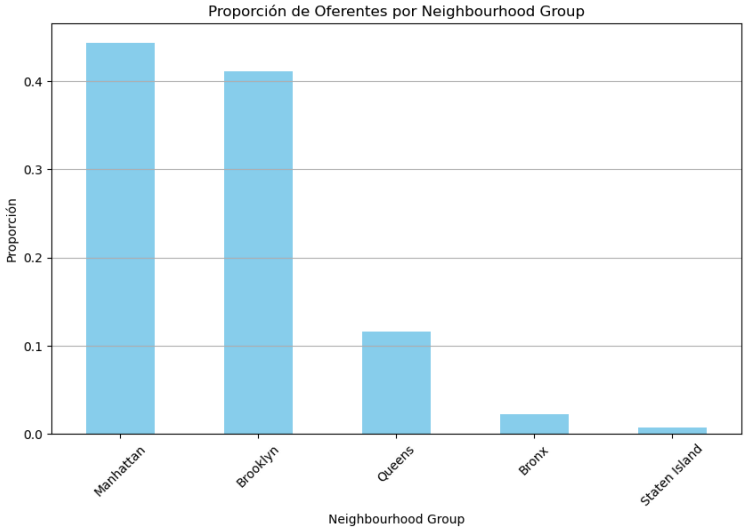
c) El objetivo del paper es limpiar los *datasets* para que puedan ser analizados como si no hubieran *missing val*ues. Si los *missing values* no son completamente aleatorios, es necesario modelarlos explícitamente o aceptar cierto sesgo en las inferencias. Desafortunadamente, no podemos estar seguros de si los datos realmente faltan de manera azarosa, o si el dato falta porque depende de predictores no observados o de los propios datos faltantes. La dificultad radica en que, por definición, estas posibles "variables ocultas" son inobservables, por lo que nunca podemos descartarlas. Dicho en otras palabras, puede conducir a estimaciones sesgadas y conducir a errores estándares mayores porque la muestra que excluye a los missing values puede no ser representativa de la muestra en su totalidad. En este trabajo optaremos por los métodos de imputación (llenar los valores faltantes) con el objetivo de conservar el tamaño de la muestra completo.

d) Si hay observaciones con *outliers* o valores que no tienen sentido, optamos por reemplazarlos por el promedio de las variables observadas. Tal como lo menciona el paper, esta estrategia puede distorsionar la distribución de esta variable, lo que conlleva complicaciones con medidas resumidas, incluidas subestimaciones de la desviación estándar.

Parte II: Gráficos y visualizaciones

2) Respondan las siguientes preguntas: ¿Cuál es la proporción de oferentes por “Neighbourhood group”? ¿Y por tipo de habitación? Además, realicen gráficos para mostrar estas composiciones y comenten los resultados.

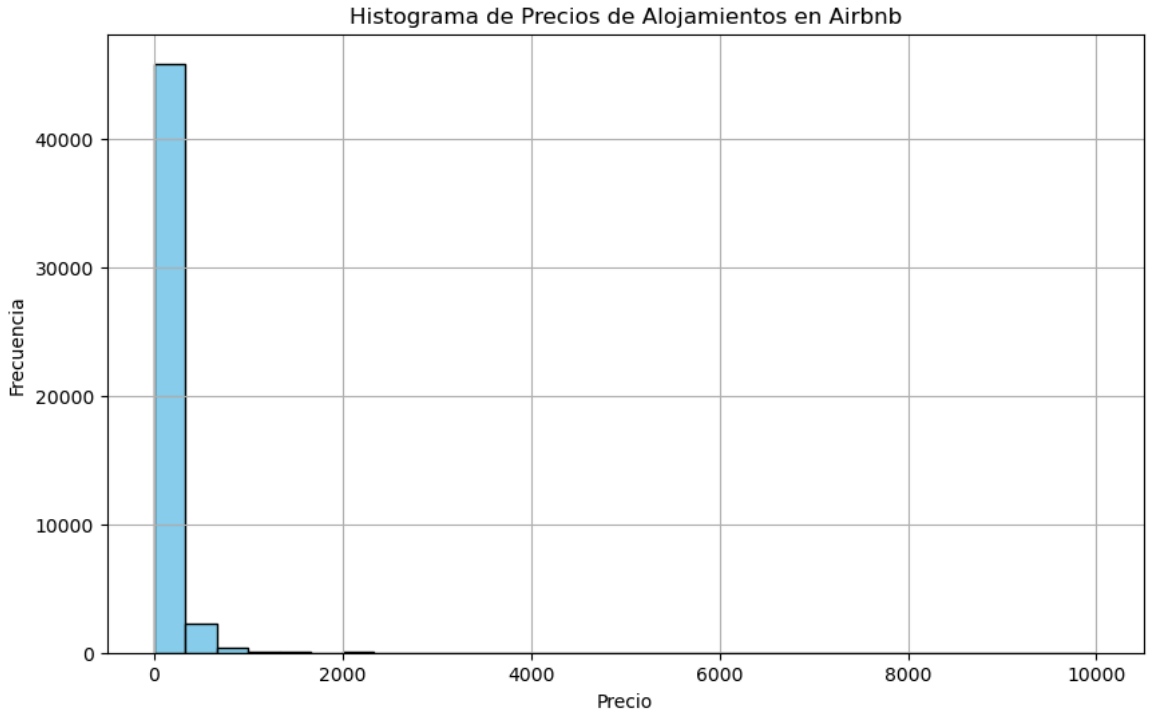




Proporción de Oferentes por Tipo de Habitación y por Vecindario

- Este gráfico nos muestra que la mayoría de las ofertas son para habitaciones privadas (65.7%), seguidas por casas o apartamentos enteros (32.0%) y una pequeña fracción para habitaciones compartidas (2.4%). Esto indica que los anfitriones prefieren ofrecer privacidad o espacios completos en lugar de ofrecer habitaciones compartidas. A su vez, el histograma indica que hay una mayor proporción de oferentes en Manhattan, Brooklyn y Queens. Mientras que en Bronx y Staten Island la proporción es considerablemente menor.

3) Realicen un histograma de los precios de los alojamientos. Comenten el gráfico obtenido. Además, respondan las siguientes preguntas: ¿cuál es el precio mínimo, máximo y promedio? ¿Cuál es la media de precio por “Neighbourhood group” y por tipo de habitación?



Precio Mínimo: $0.0

Precio Máximo: $10000.0

Precio Promedio: $152.73

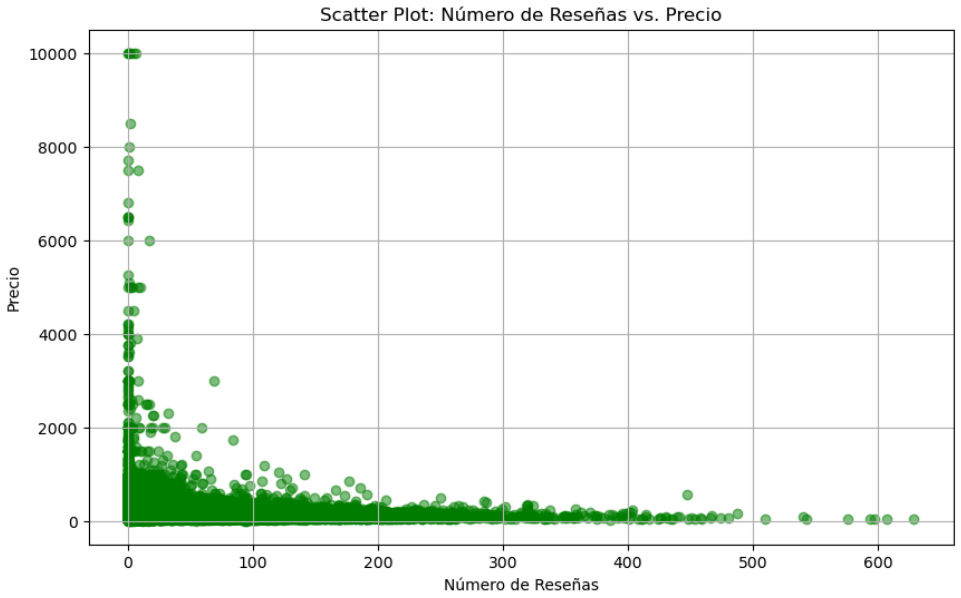
Histograma de Precios de Alojamientos en Airbnb

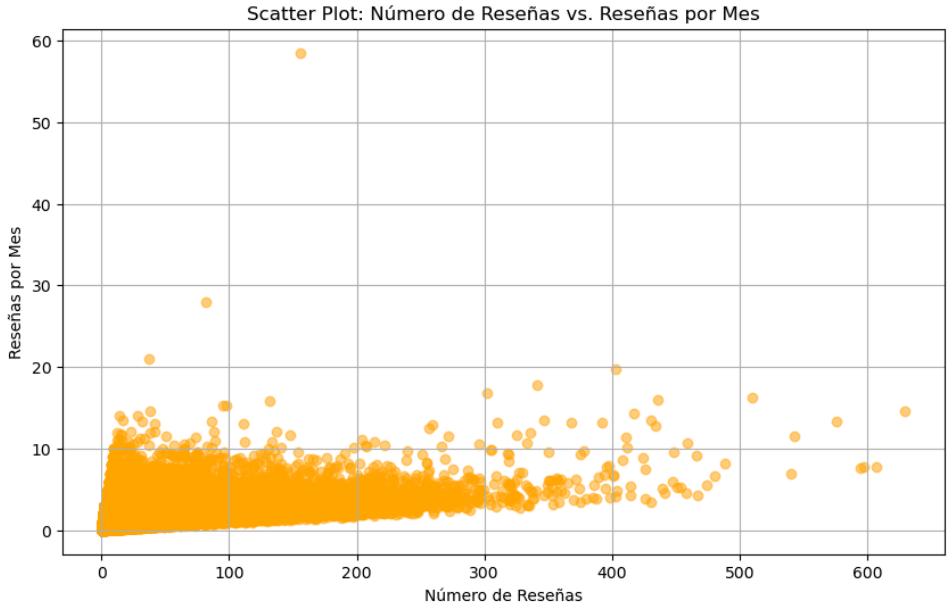
- El histograma nos indica que la mayoría de los alojamientos tienen precios bajos, con una cantidad significativa de alojamientos cerca del precio mínimo. Hay una disminución muy rápida en la cantidad de alojamientos a medida que aumenta el precio, lo que sugiere que los alojamientos caros son mucho menos comunes, esto tal vez indica también una menor demanda por parte de los clientes hacia alojamientos costosos.
- No obstante hay que tener en cuenta que los precios mínimos y máximos son extremos y podrían incluir valores atípicos que distorsionan el precio promedio.

Media de Precio por Neighbourhood Group:	
Bronx	$87.464646
Brooklyn	$124.380597
Manhattan	$196.862352
Queens	$99.536900
Staten Island	$114.812332

Media de Precio por Tipo de Habitación:	
Entire home/apt	$211.788107
Private room	$89.783388
Shared room	$70.127586

4) Realicen dos scatter plots con dos variables de interés en cada uno.





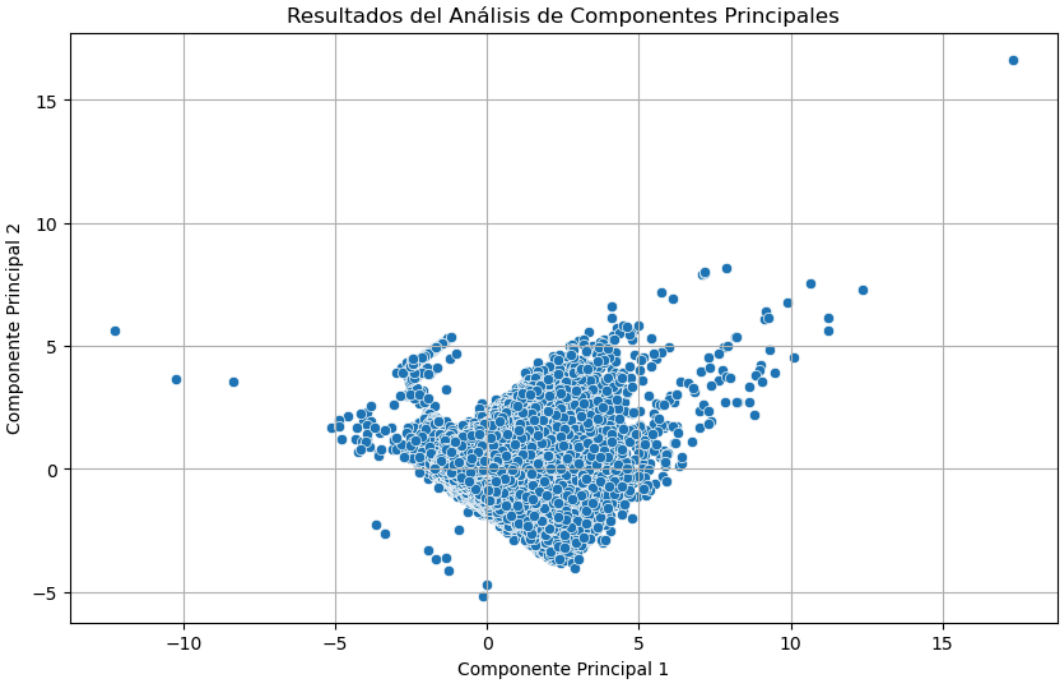
Número de Reseñas vs. Precio

En el primer scatter plot, se observa una concentración de alojamientos con un número menor de reseñas y precios bajos. No parece haber una correlación clara entre el precio y el número de reseñas, ya que alojamientos con precios altos y bajos tienen tanto pocas como muchas reseñas.

Número de Reseñas vs. Reseñas por Mes

El segundo scatter plot nos muestra que los alojamientos con muchas reseñas no necesariamente tienen un alto número de reseñas por mes, lo que podría indicar que son alojamientos que han estado disponibles durante más tiempo en lugar de ser necesariamente los más populares o frecuentemente reservados recientemente.

5) Utilicen el análisis de componentes principales para graficar las variables en dos dimensiones. Comenten los resultados obtenidos (qué porcentaje de la varianza se logra explicar con dos componentes, como son los loadings, si ven algún patrón en el gráfico).



Varianza explicada por el Componente Principal 1: 0.21

Varianza explicada por el Componente Principal 2: 0.16

Resultados del Análisis de Componentes Principales:

- El gráfico del PCA muestra una distribución de los alojamientos en dos dimensiones principales. Con el primer componente explicando el 21% y el segundo el 16% de la varianza. Juntos no capturan la mayoría de la varianza en

los datos, lo que sugiere que las características de los alojamientos son multidimensionales y no pueden simplificarse fácilmente a solo dos factores.

- Por otra parte, no hay un patrón claro y definido en el gráfico, lo que podría significar que no hay grupos distintos o que las variables incluidas en el PCA tienen una relación compleja que no se presta a una separación clara en estas dos dimensiones.
-

Parte III: Predicción

3) Implementen una regresión lineal y comenten los resultados obtenidos.

A continuación comentaremos sobre los valores que arrojó las regresión para cada valor de la estadística descriptiva

- Error Cuadrático Medio (MSE) es 35546.46. Su utilidad depende del contexto de los precios de los alojamientos; sin embargo, al ser un número bastante grande, puede indicar que hay una variabilidad significativa en los precios que el modelo no está capturando. Es importante comparar este valor con la varianza de los precios y ver qué tan grande es en proporción.

Coeficientes de la Regresión Lineal: El análisis será hecho asumiendo las demás variables constantes, es decir el análisis marginal es ceteris paribus y en promedio.

- id y host_id: Los coeficientes son extremadamente pequeños, lo que sugiere que la identificación única de un listado o un anfitrión tiene muy poco o ningún efecto sobre el precio.
- latitude: Tiene un coeficiente positivo, lo que indica que cuanto más al norte está el alojamiento (mayor latitud), mayor es el precio.
- longitude: Tiene un coeficiente negativo significativo, lo que sugiere que cuanto más al oeste está el alojamiento (mayor longitud), menor es el precio.

Esto podría reflejar diferencias en el precio promedio entre distintos vecindarios.

- minimum_nights, number_of_reviews, reviews_per_month, y calculated_host_listings_count: Todos tienen coeficientes negativos, lo que significa que un aumento en estos valores está asociado con una disminución en el precio. Esto podría ser contra intuitivo, especialmente para number_of_reviews y reviews_per_month, ya que uno podría esperar que una mayor popularidad (más reseñas) resulte en precios más altos. Sin embargo, puede haber factores no capturados que influyan en esta relación.
- availability_365: Con un coeficiente positivo, sugiere que los alojamientos con mayor disponibilidad a lo largo del año tienden a tener precios más altos, lo que podría ser interpretado como un signo de mayor demanda.
- neighbourhood_group_numeric y room_type_numeric: Ambos con coeficientes negativos bastante grandes, indican que hay grupos de vecindarios y tipos de habitación que tienden a tener precios más bajos.
- offer_group: Tiene un impacto negativo muy pequeño en el precio, casi insignificante.
- Intercepto: el intercepto es -29116.63, lo que es un poco inusual ya que un intercepto negativo en el contexto de precios sugiere que, en la ausencia de todas las otras variables, el precio sería negativo, lo cual no tiene sentido práctico. Esto puede sugerir problemas con los datos, como valores atípicos o la necesidad de transformación de variables para que el modelo sea más representativo.