Universidad de San Andrés

Escuela de Economía

BIG DATA

Docentes: María Noelia Romero y Victoria Oubina

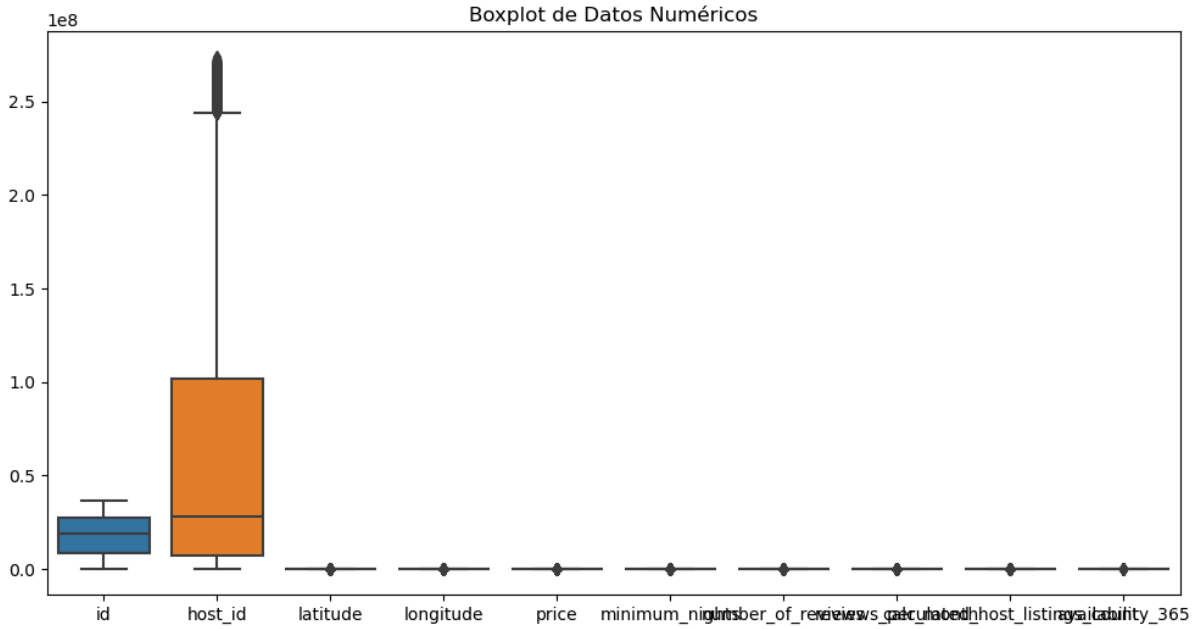
TP2

21/04/2024

Alumnos: Sofía Ellenberg, Sophie Schulzen y Vicente Zervino

Parte 1:

d)



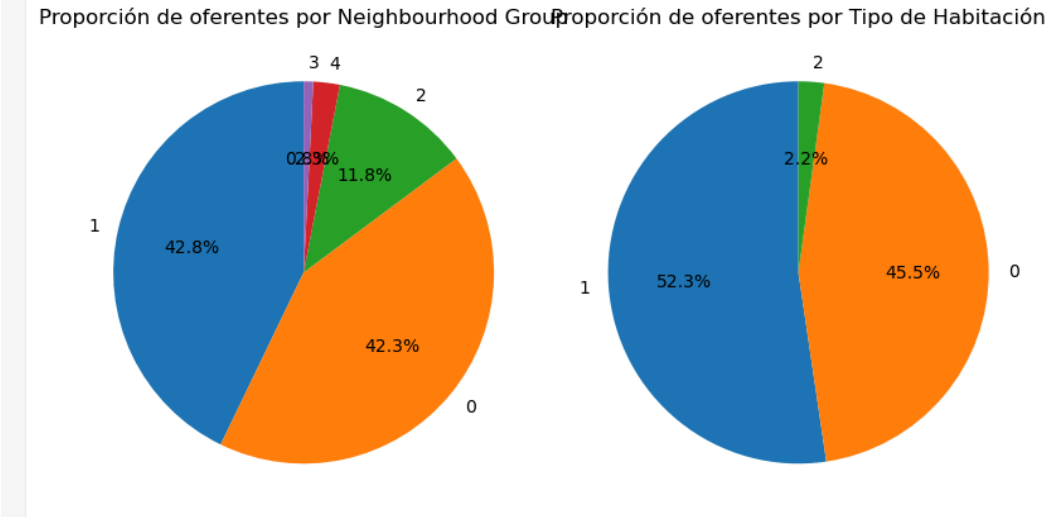
A partir del gráfico vemos que las variables id y host_id tienen una alta variabilidad en los datos. Sin embargo, por la naturaleza de las variables, estas no son útiles para el análisis estadístico, por ende no nos preocupan.

e)

Observamos que se transformaron las dos columnas de variables exigidas a variables numericas: 'neighbourhood_group' y 'room_type'.

Parte 2:

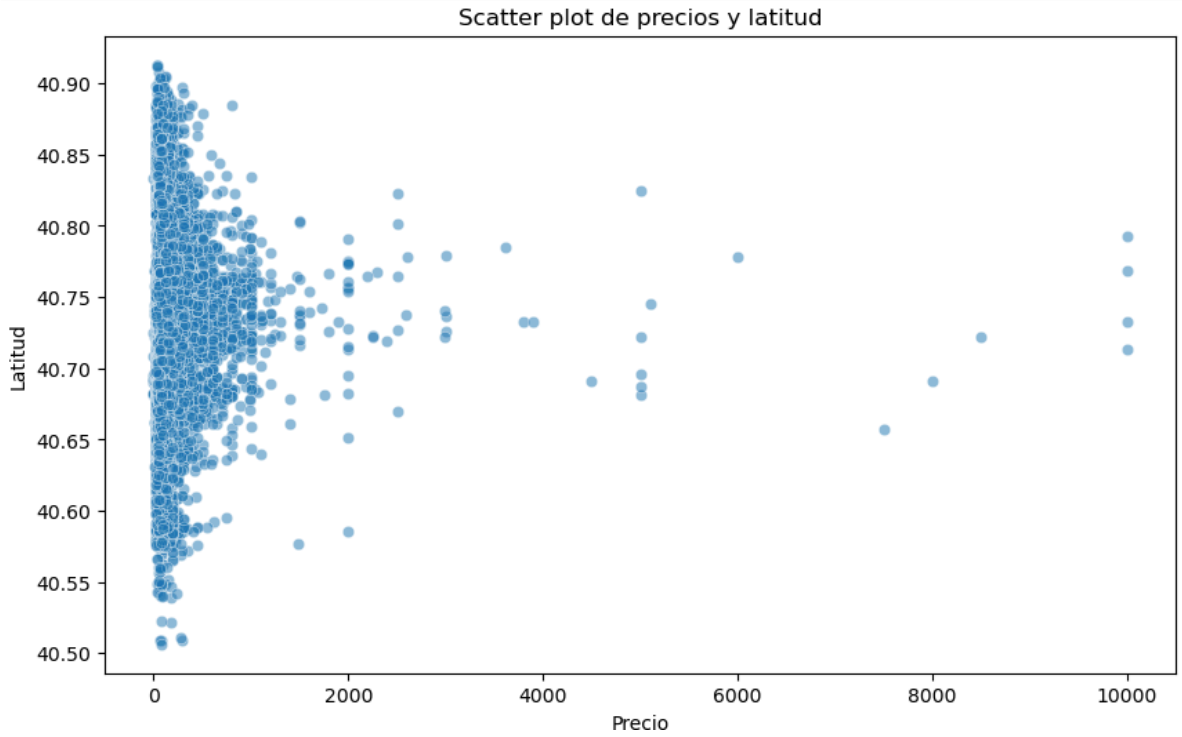
2)

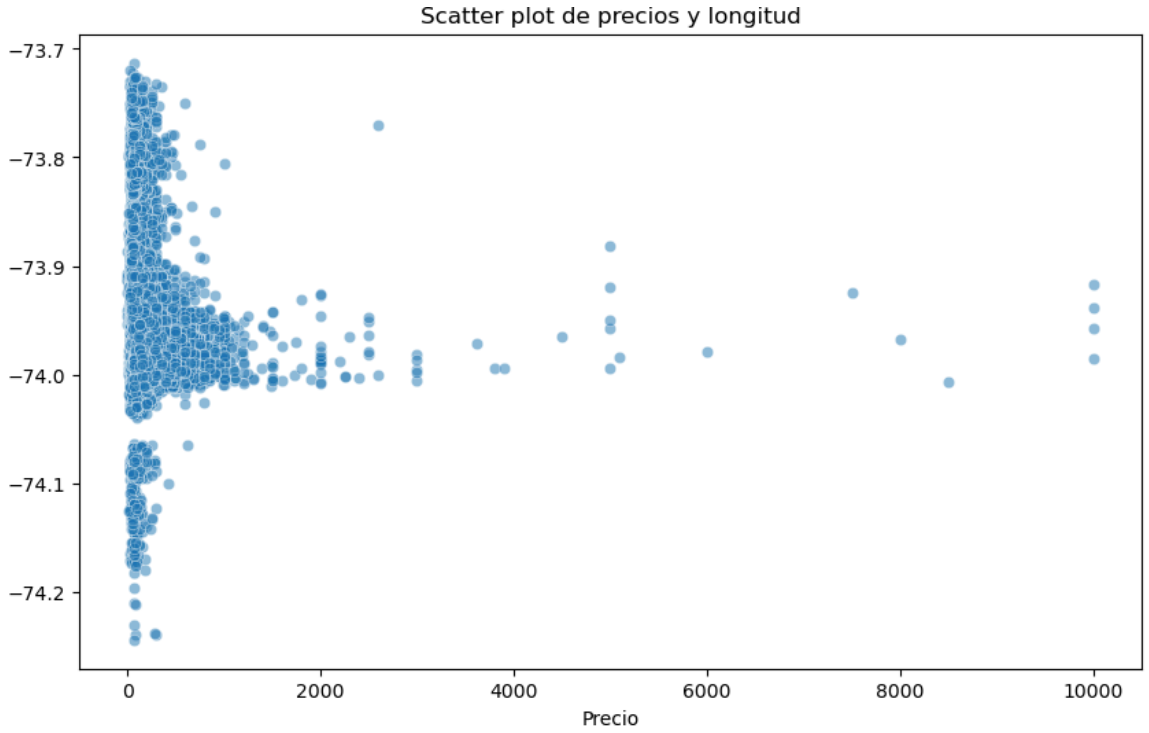


3)

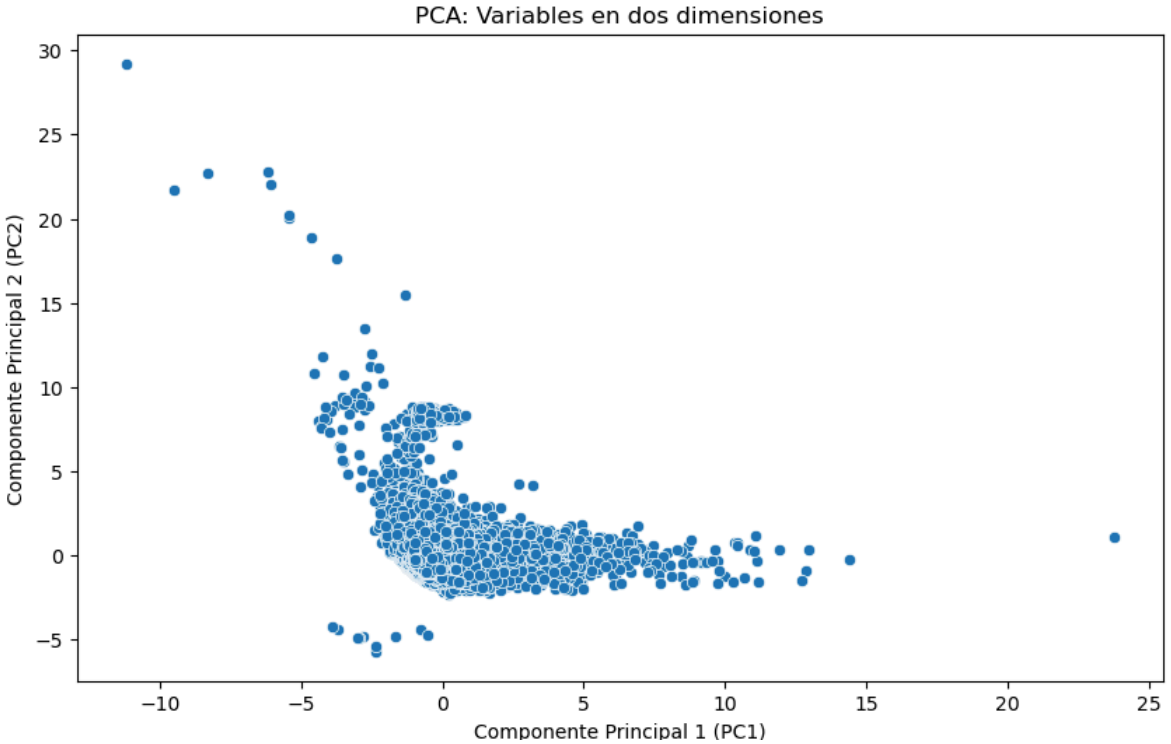
El histograma muestra una distribución de precios en los que la amplia mayoría se concentran en el intervalo (0,333), pero que a la vez llega a tener precios de hasta $10.000. El precio promedio es de $142,3 aproximadamente, los precios mínimos y máximos son $0 y $10.000 respectivamente. Se reportan también las medias filtradas por tipo de cuarto y vecindario.

4)





Tanto en dirección norte-sur como en dirección este-oeste parece que los precios mayores están en el centro y no en los bordes. Si bien hay precios "bajos" en todas las ubicaciones, solo hay precios "altos" en el centro de la ciudad de Nueva York, territorialmente hablando.
5)



#La varianza explicada por los primeros 2 componentes es del 37,7%. Los loadings muestran que el componente PC1 explica variabilidad relacionada con el número de reviews y el número por mes, junto con la disponibilidad 365 días. Esto tiene sentido porque las primeras 2 están lógicamente muy relacionadas entre si y el promedio por mes tiene relación obviamente con si la disponibilidad es todos los días del año. Variables como la latitud, longitud, precio, cantidad mínima de noches y conteo de hosts no están tan relacionadas con el componente 1. El componente 2 también agarra variabilidad relacionada con la disponibilidad 365 días pero que se relaciona aparte con el precio, la cantidad mínima de noches y la cantidad de hosts. El método PCA con 2 variables deja prácticamente sin explicar la variabilidad de la latitud, ya que ambos coeficientes son bajos en valor absoluto. El gráfico no muestra una relación clara entre los dos coeficientes principales, lo que tiene sentido porque los componentes son ortogonales.

Parte 3:

3)

En primer lugar observamos que el id y el host_id no tienen impacto en el precio. Es decir, no lo incluyen. A su vez, el número de anuncios del anfitrión, el minimo de noches y la disponibilidad parecerian influir de manera positiva pero no muestran un efecto tan grande. Luego, vemos que la latitud y el número de revisiones influyen de manera positiva teniendo la latitud un mayor impacto.. A diferencia, la longitud, tienen un impacto negativo en el precio lo que pareceria un resultado medio incoherente si lo comparamos con latitud que probablemente se deba a problemas de estimacion. Consideramos que el precio podria ser estimado de manera mas precisa si se incluyeran otras variables en el modelo.