Informe del Trabajo Práctico 2 - Big Data

Parte I:

a - Eliminamos los datos con id duplicado. Encontramos 20 datos repetidos que fueron extraídos del dataset con el que vamos a trabajar.

b - Las columnas name y host_name, que indican el nombre del alojamiento y dueño respectivamente, no agregan información relevante para el tipo de análisis que vamos a realizar más adelante.

c - Tabla 1: missing values por columna

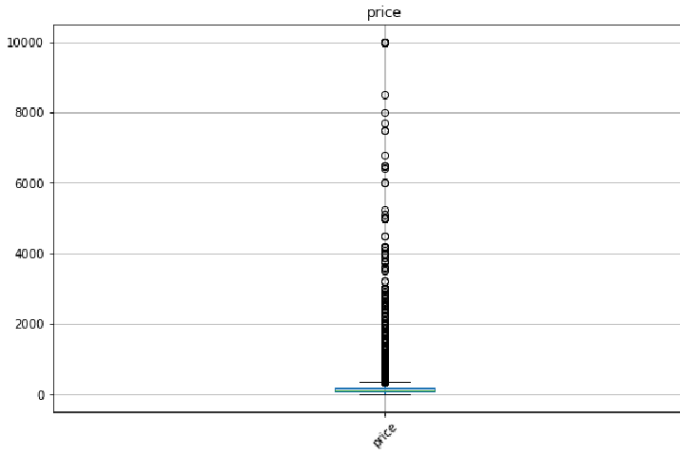
Variable	Missing Values
price	15
last_review	10052
reviews_per_month	10052

Como podemos ver en la Tabla 1 hay sólo tres variables que tienen missing values. Si bien *price* es la que menos missing values tiene de las tres, es la variable más importante. Al ser tan solo 15 datos en una muestra de más de cuarenta mil observaciones, decidimos dropear estas observaciones ya que no creemos que vaya a tener un efecto significativo sobre los resultados.

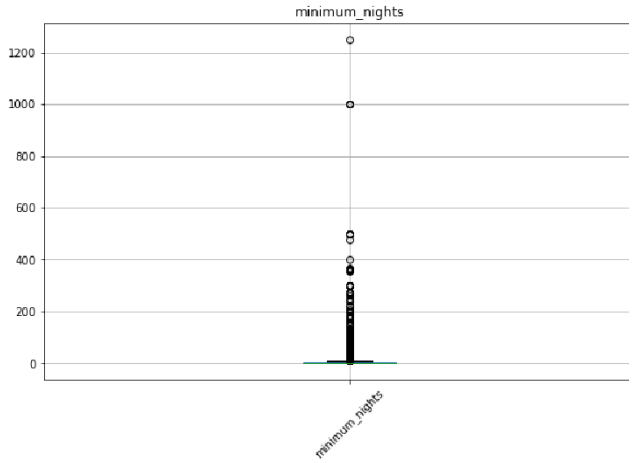
Para la variable *last_review* no tiene sentido realizar un imputation ya que los missing values vienen del hecho de que estas observaciones recibieron 0 reviews. Podemos ver que hay la misma cantidad de observaciones con 0 reviews que la cantidad de missing values para esta variable. Dicho esto, sería apropiado no tener en cuenta esta variable para realizar un análisis.

En cuanto a *reviews_per_month* podemos reemplazar a los missing value con un 0. Esto tiene sentido ya que son observaciones con cero reviews con lo cual sería lógico que los review por mes también sean cero.

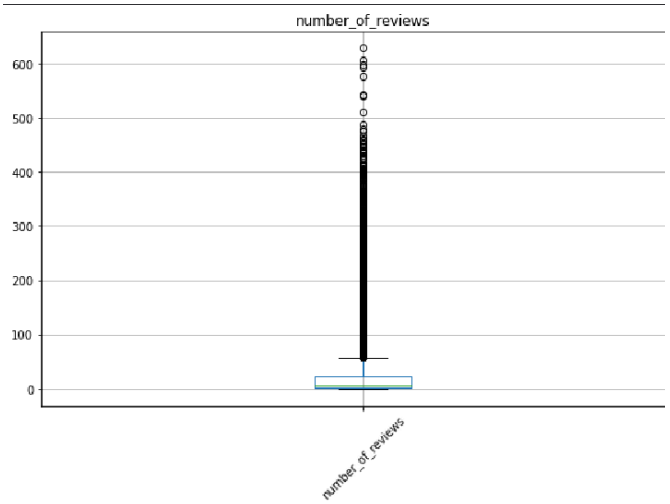
d - Un método muy útil para indagar en el dataset y averiguar si hay outliers o valores sin sentido es un boxplot. El boxplot permite realizar un análisis visual de ambas medidas. Las variables analizadas son *price, minimum_nights, number_of_reviews* y *availability_365*. Para estas variables graficamos su boxplot:



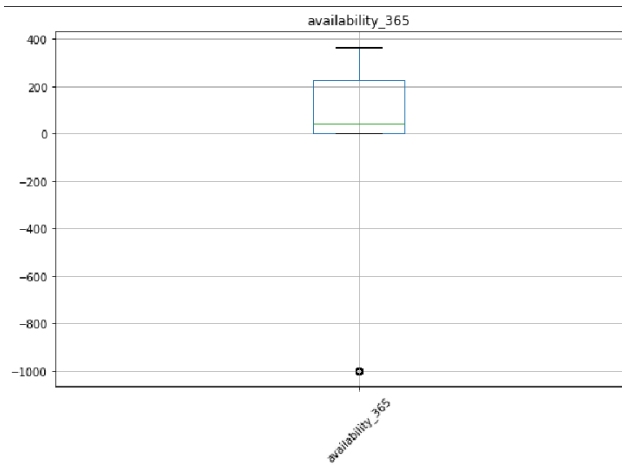
Para el precio podemos observar que hay una gran cantidad de valores que están por encima de 1,5 veces el rango intercuartil. Estos valores no presentan un problema para análisis. Por otro lado podemos observar varias observaciones con precio 0. Se puede argumentar que no tiene sentido que sean gratis entonces estas observaciones se eliminan.



Si bien no hay nada intrínsecamente incorrecto con que hayan valores de *minimum_nights* que sean igual o mayor a 1000 son datos muy alejados a la gran mayoría de los otros. Decidimos eliminar todos los datos que tiene minimum_nights >= 900.



No observamos nada incorrecto con la cantidad de reviews de los Airbnb.



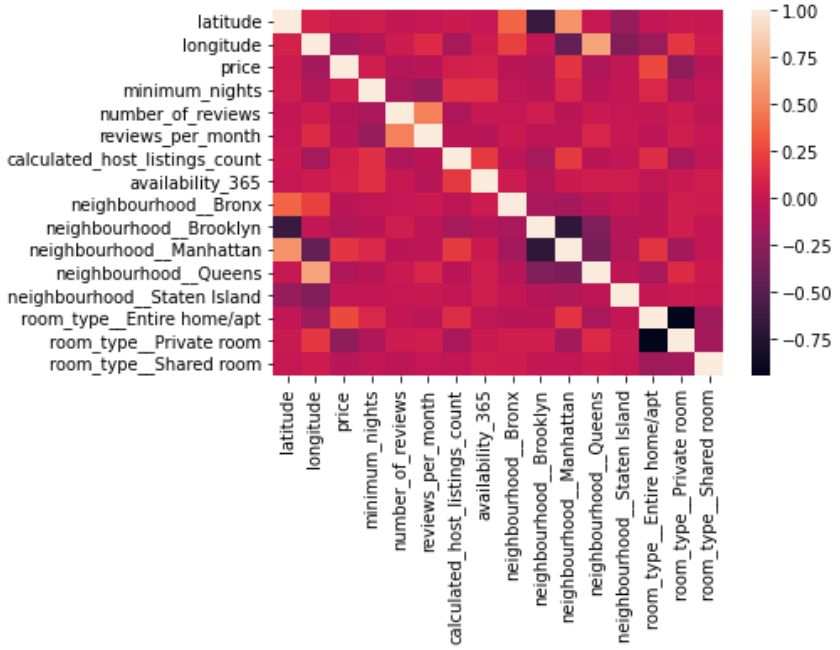
Finalmente, miramos la variable *availability_365* e inmediatamente resaltan las observaciones negativas. Además, nos encontramos con varias observaciones que su disponibilidad en días al año es igual a cero. Ponemos la condición de que la disponibilidad tiene que ser positiva.

e - Creamos dummies para cada tipo de neighbourhood y room.

f - Procedimiento en el código.

Parte II:

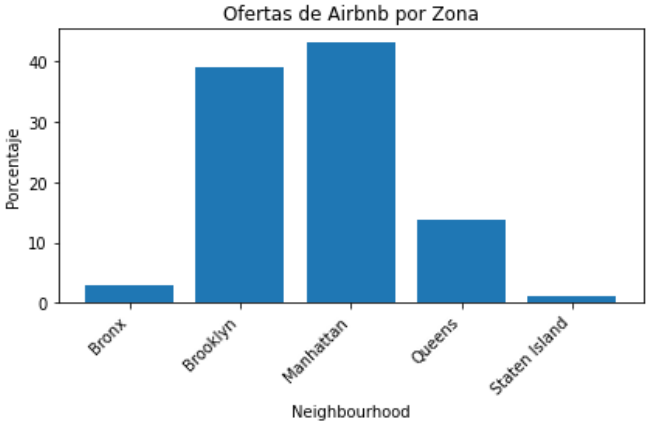
1 - Computamos las correlaciones de las variables de interés. Tanto *neighbourhood_group* como *room_type* eran variables categóricas, por lo que fueron eliminadas del correlograma y reemplazadas con dummies representando a cada tipo.



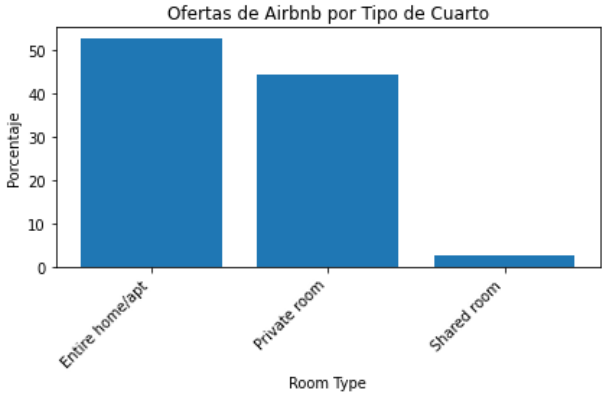
Encontramos que la mayoría de las correlaciones de las variables se encuentran entre 0.5 y -0.5 indicando que no hay mucha correlación entre las variables. Después encontramos que existe gran correlación negativa entre las variables que eran categóricas, algo que tiene

sentido ya que eran categorías excluyentes. Luego es interesante notar que la latitud y longitud se correlaciona con los barrios, algo que era de esperar.

2 - Primero analizamos el porcentaje de ofertas por zona de Nueva York:

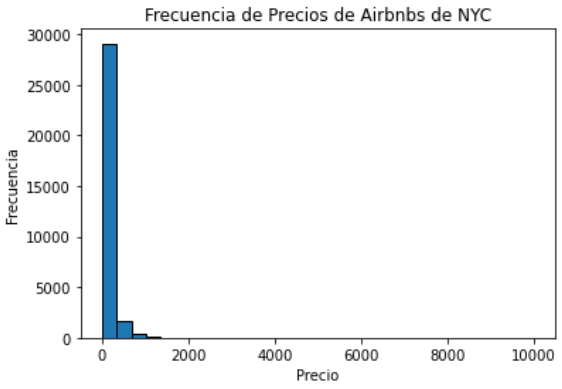


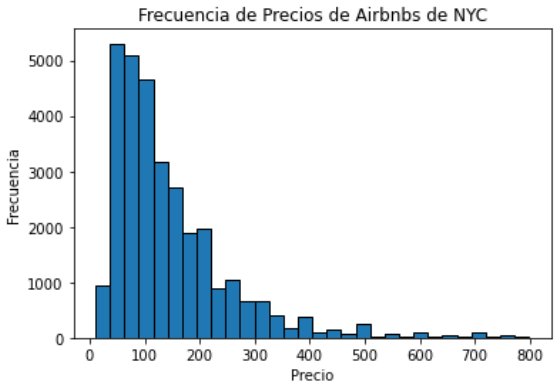
En el gráfico podemos observar como hay dos barrios de Nueva York que destacan en cuanto a cantidad de ofertas de Airbnb. Entre Brooklyn y Manhattan se encuentran más del 80% de las ofertas.



La composición de ofertas por tipo de cuarto está dominada por el departamento o casa entero. Esta categoría cuenta con más de la mitad del mercado mientras que la oferta de un cuarto privado la sigue muy cercanamente con un 44% de las ofertas. Por último, se ofrece un cuarto compartido solo el 2.7% de las veces.

3 - Graficamos el histograma de los precios. Mostramos dos histogramas con distintos límites de precios. El primero tiene en cuenta todos los datos mientras que el segundo hace “zoom” a los datos con precio menor a 800.





Podemos observar como los precios se concentran en el rango 50 a 200 aproximadamente. Tal como lo analizamos en la Parte I del informe, los precios cuentan con múltiples outliers que llegan incluso hasta $10.000. Por esta razón se distorsiona el histograma 1. El precio mínimo es $10, el máximo es $10.000 y el precio promedio es de $162.

Cuando analizamos el precio promedio por barrio nos encontramos con la siguiente tabla:

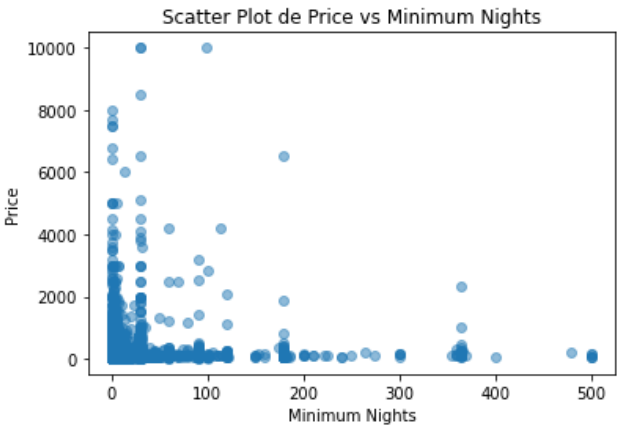
Barrio	**Precio Promedio**
Manhattan	$214
Brooklyn	$133
Staten Island	$114
Queens	$100
Bronx	$89

Precio promedio por tipo de cuarto:

Tipo de Oferta	**Precio Promedio**
Casa/Dpto Entero	$224
Cuarto Privado	$94
Cuarto Compartido	$66

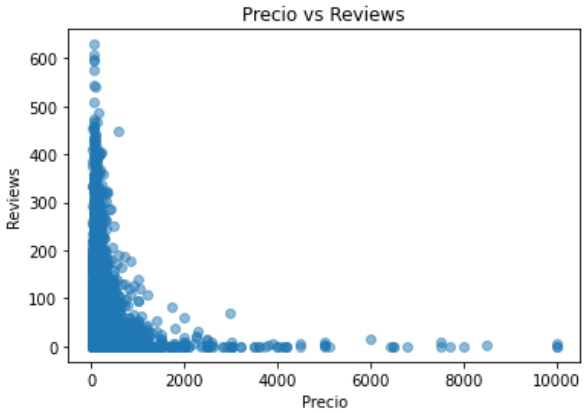
4 - Scatterplots

Precio vs Noches Mínimas:



Como el precio es por noche, tiene sentido que haya una relación negativa entre estas variables ya que se espera que haya un descuento por unidad consumida.

Precio vs Reviews:



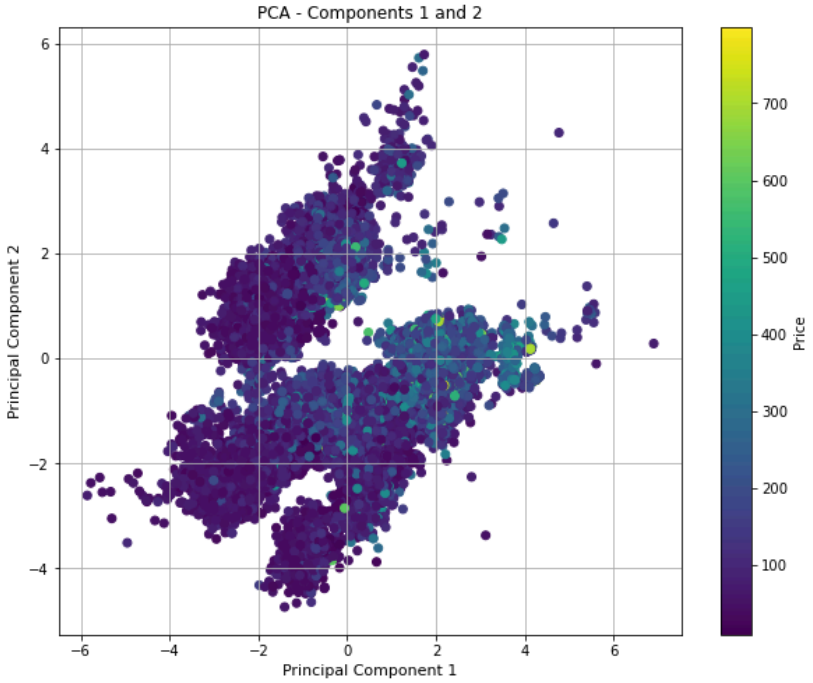
Encontramos una relación negativa entre el precio y la cantidad de reviews. Como economistas nos interesa ver cómo se cumple la dinámica precios y demanda, donde a mayor precio encontramos menor demanda.

5 - Realizamos un Análisis de Componente Principal (PCA) donde están predeterminados el uso de dos componentes. El componente 1 explica el 16% de la varianza mientras que el componente 2 explica el 13%.

Loading Vectors:

Variable	**Loading Vector CP 1**	**Loading Vector CP 2**
host_id	-0.01897331	-0.164
latitude	0.27211133	-0.499
longitude	-0.35083576	-0.308
minimum_nights	0.18226979	0.039
number_of_reviews	-0.1229867	0.015

reviews_per_month	-0.16210734	-0.064
calculated_host_listings_count	0.23353512	-0.01
availability_365	0.04111532	-0.068
neighbourhood__Bronx	-0.02803478	-0.256
neighbourhood__Brooklyn	-0.31392594	0.495
neighbourhood__Manhattan	0.49746544	-0.222
neighbourhood__Queens	-0.26007837	-0.301
neighbourhood__Staten	0.01534427	0.15
room_type__Entire home/apt	0.36018326	0.274
room_type__Private room	-0.35633208	-0.261



Tras graficar los dos componentes principales nos encontramos con una dispersión de datos de forma tal que podrían ser clustereados en dos grupos.

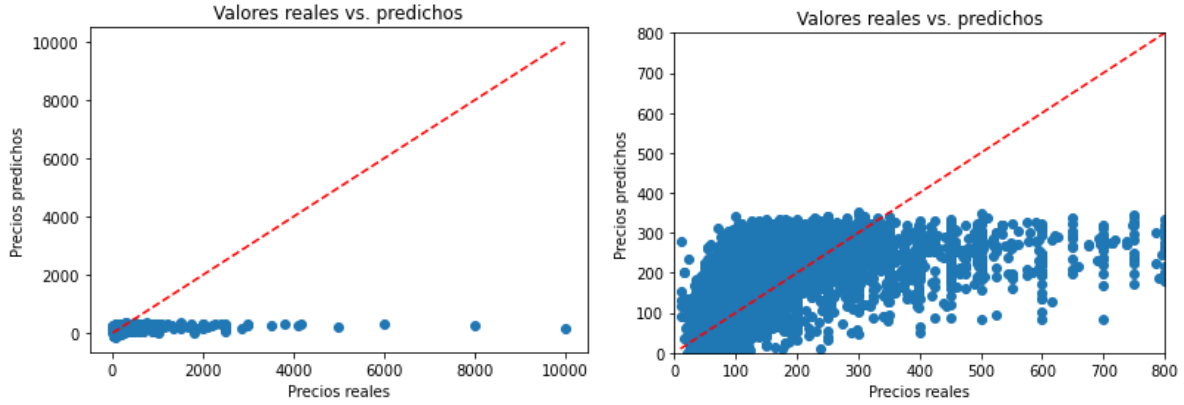
Parte III:

1 - Separamos la base de datos en features (X's) y la variable objetivo/dependiente (y).

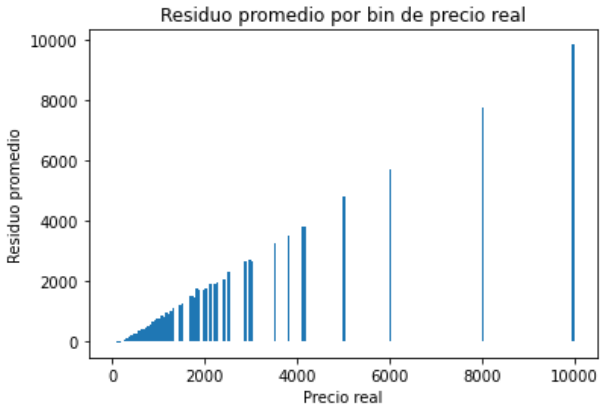
2 - Procedimiento en el código.

3 - Tras correr la regresión lineal encontramos un error promedio de $223 en cuanto al valor predecido del precio del Airbnb por noche en comparación al valor real.

Graficamos los valores predichos y los valores reales para comprar. El gráfico 1 no tiene límite de precio y el gráfico 2 se enfoca en los valores menores a 800:



Estos gráficos nos sugieren que el error de predicción es creciente en el precio real. Para verificar esto separamos los datos en bins con un ancho de $50. Luego computamos el error promedio de cada bin y graficamos:



Tal como sugiere el gráfico de valores reales vs predicho, el error promedio del modelo es creciente en el precio.