Universidad de
SanAndrés

Big Data

Profesora: Romero, María Noelia

Tutora: Oubiña, Victoria

Trabajo Práctico 2

Cucher Maximiliano, Saucedo Federico & Soares Gache Manuel

Fecha de entrega: 21/4/2024

Parte I: Limpieza de la base

1.b) Aquellas variables que decidimos eliminar para realizar nuestro análisis son: id, name, host id y host name. El motivo por el cual decidimos no tener en cuenta estas variables es porque al realizar un análisis con una gran cantidad de observaciones y características, la información que podemos obtener no resulta relevante a nivel granular tales como características del nombre del dueño y su id.

c) Al visualizar nuestra base de datos, observamos que había tres categorías con *missings values* (last_review, review_per_month y *price*). En el caso de las primeras dos variables, las variables que contaban con *missing values* eran aquellas que no tenían *reviews*, es decir, lo correcto para ese caso era que tomen el valor de 0 por lo cual realizamos eso.

En lo referente a la variable *price* al leer el *artículo "Missing Data imputation"* consideramos que la manera adecuada de resolver este problema de *missing data* es en lo que los autores denominan *Nonresponse weighting* (p.532). Es decir, construimos un modelo para predecir la categoría no respondida utilizando el resto de las variables. Dado que solo nos quedó una variable con *missing data* consideramos que esta herramienta podría solucionar nuestro potencial problema dado que si tuviéramos más de una variable con *missing data* esta metodología se complicaría. Además, creemos que las otras variables pueden ser de buena utilidad para predecir el precio, o por lo menos predecirlo mejor que poniendo el promedio, o otras alternativas similares.

d) Para determinar cómo manejar los valores atípicos, comenzamos realizando un diagrama de caja para todas las variables numéricas. Esto nos permitió identificar observaciones que podrían tomar valores inusuales en cada variable.

En primer lugar, eliminamos los valores negativos de la variable "availability_365", ya que carece de sentido que esta variable tenga valores negativos.

En segundo lugar, decidimos eliminar los valores atípicos de "reviews_per_month" que superaran los 30, ya que observamos una clara separación de estos valores con respecto al resto de la muestra. Esto sugiere que no resulta plausible que haya tantas revisiones por mes.

En tercer lugar, notamos que es poco común que los alojamientos de Airbnb tengan un mínimo de noches muy elevado, ya que la plataforma suele utilizarse principalmente para alquileres de corta duración. Observamos que la cantidad de alojamientos con un mínimo de noches muy elevado era baja. Por lo tanto, decidimos eliminar el 1% más alto de la muestra, quedándonos así con alojamientos de Airbnb con valores mínimos de 45 noches, lo cual parece más razonable.

En cuarto lugar, vamos a eliminar las 11 observaciones que marcan que el precio del AIRBNB es 0, ya que obviamente esto no tiene sentido.

Finalmente, decidimos no eliminar más valores atípicos, ya que en las demás variables no encontramos observaciones tan alejadas de la muestra en general. Además, los valores de las

variables no son muy elevados, por lo que no deberíamos enfrentar problemas significativos con la estadística descriptiva.

La variable “calculated_host_listings_count” presentaba observaciones que parecían ser valores atípicos, pero al analizar más detenidamente, encontramos que había muchas observaciones con valores muy elevados. Estos valores altos pueden corresponder a agencias que se encargan de gestionar propiedades en nombre de sus clientes. Por lo tanto, decidimos conservar estos valores, ya que representan un comportamiento válido dentro del contexto de la plataforma Airbnb.

De todos modos, a la hora de hacer estadística descriptiva, podemos hacer histogramas y en el caso que no sean muy informativos ya que hay observaciones que toman valores muy elevados, podemos tomar decisiones en ese momento.

Parte II: Gráficos y visualizaciones

1.

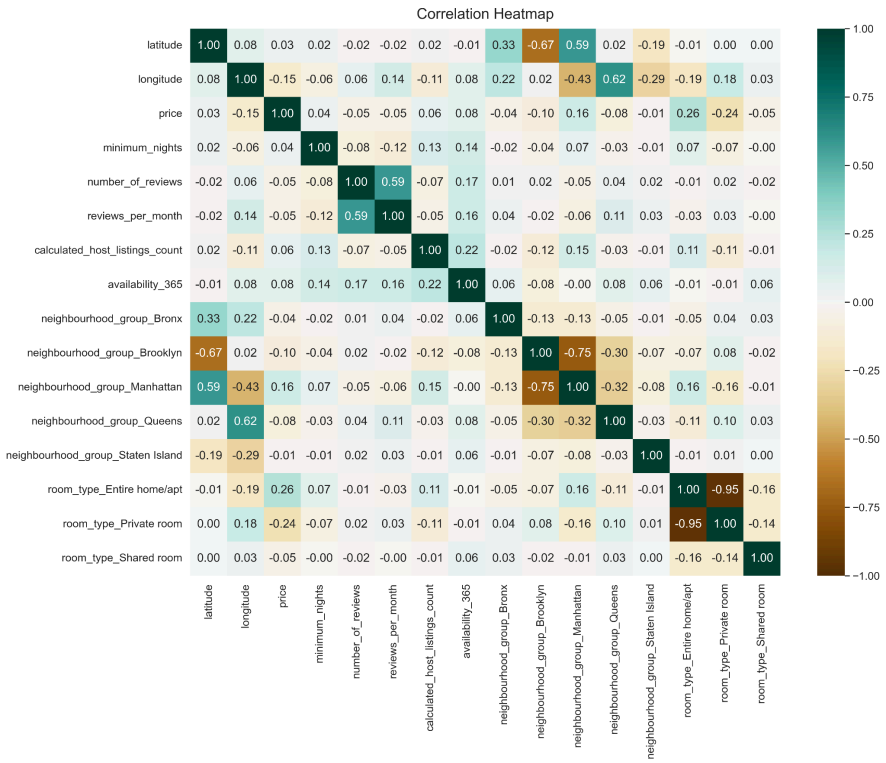

Gráfico 1: Correlation heatmap

En el correlation heatmap encontramos la correlación entre todas nuestras variables. Notamos que las correlaciones más altas son las siguientes. Obviamente va a haber correlación fuerte y negativa entre las dummies del mismo tipo de categoría ya que estas son excluyentes. Vemos que no hay correlación entre el barrio Staten Island y el tipo de cuarto compartido. Esto puede pasar porque la mitad de los cuartos de ese lugar son de este tipo y la otra mitad no. Por otro lado, es lógico ver correlaciones fuertes con latitud, longitud y barrios. Por ejemplo, Queens y latitud tienen una correlación fuerte y alta, porque Queens queda en una latitud alta

en el mapa. Vemos que la variable precio y Manhattan tienen una correlación bastante grande y positiva (0.16), lo que muestra que el precio en esa zona suele ser grande. Lo mismo pasa entre las variables precio y apartamento completo (correlación de 0.26), lo cual es lógico.

2.

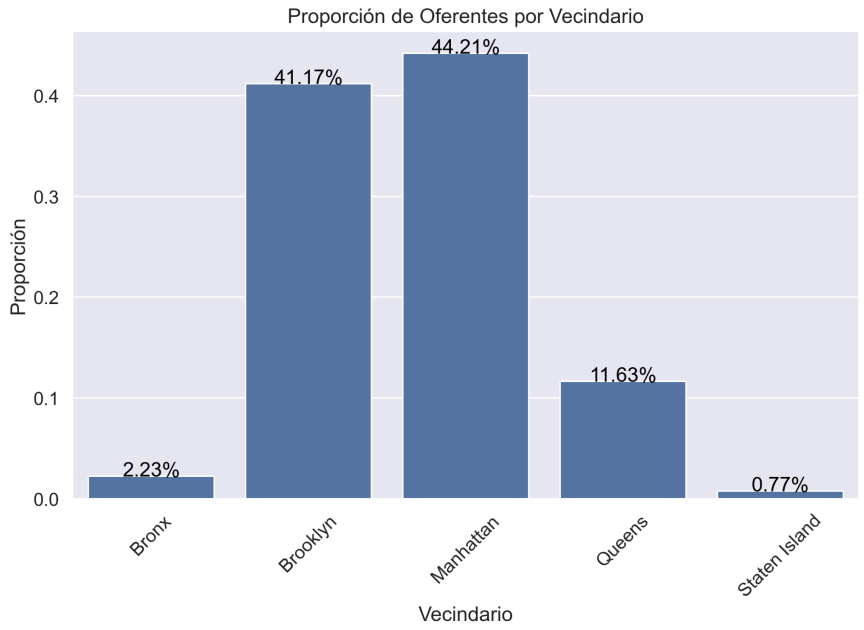

Gráfico 2: Proporción de Oferentes por vecindario

Podemos observar que aquellos lugares que en porcentaje equiparan el 85% de los oferentes son Manhattan y Brooklyn, los dos destinos que a priori parecieran más turísticos de Nueva York, por ello, el volumen relativo de ofertas de Airbnb resulta alto. A diferencia de estos dos lugares observamos porcentajes muy bajos de oferentes en Bronx (2,23%) y Staten Island (0,77%). Un posible motivo de porcentaje tan bajo de ofertas de Airbnb podría ser que no son destinos destacados por su turismo, en el caso del Bronx resulta un vecindario en el cual se cosenza un gran número de población de descendencia afroamericana y es caracterizado por no ser un lugar del todo válido para alquilar un Airbnb dado su nivel de violencia e inseguridad. Para el caso de Staten Island resulta un lugar residencial de Nueva York y no tanto un destino turístico a priori.

Estas conclusiones las realizan suponiendo que, principalmente el volumen de demanda de Airbnb son en parte de turistas y por lo tanto su oferta se ve condicionada a esta característica. Además consideramos relevante comentar que, para obtener conclusiones más informativas resulta útil controlar por el tamaño de población y de metros cuadrados.

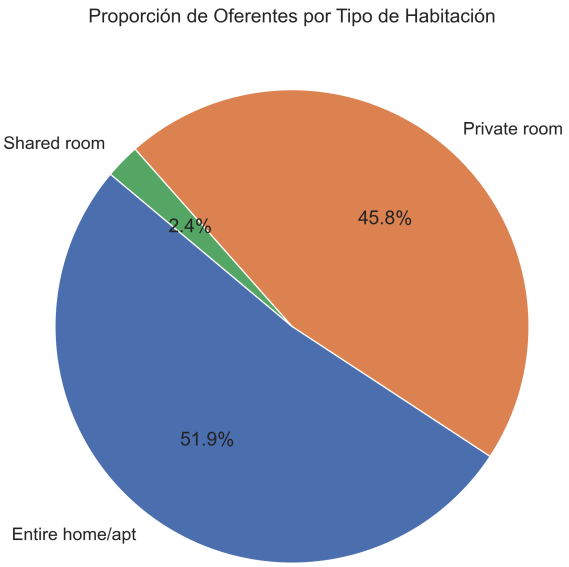

Gráfico 3: Proporción de oferentes por tipo de habitación.

En este gráfico el cual segrega las ofertas de Airbnb por tipo de habitación observamos que en su gran medida se componen de apartamentos enteros (51,9%) y de cuartos privados (45,8%). Tan solo el 2,4% de la muestra corresponde a cuartos compartidos. Esto podría ser ya que las habitaciones compartidas suelen corresponder a hostels que publican sus cuartos por otros medios, no por Airbnb.

3. Cuando hicimos el histograma nos dimos cuenta que si tomábamos todo el rango de valores de precio este no era muy informativo ya que las observaciones con precios altos imposibilitaba una visualización comprensible del histograma. Es por eso que decidimos quedarnos en este caso con las observaciones en donde el precio es menor a 2000.

Para obtener el precio máximo, mínimo, promedio y esos mismos valores corrimos el histograma sin borrar las observaciones más altas. El precio mínimo es de 10 dólares, el precio máximo es de 10.000 dólares, mientras que el precio promedio es de 151,64 dólares aproximadamente.

En cuanto a la media de precio por barrio, tenemos que los alquileres en Bronx tienen un precio promedio de 87,78 dólares, en Brooklyn de 123,74, en Manhattan 195,82,en Queens 97,17, y por último en Staten Island 114,85. Este promedio difiere bastante lo que implica que la variable del barrio es muy relevante a la hora de predecir el precio. En cuanto a la media de precio por tipo de habitación, también notamos ciertas discrepancias. En lo que respecta a la media de una casa/departamento entero es de 211,34. Luego, siguiendo con la media de un cuarto privado es de 88,21 y finalmente, la media de un cuarto compartido es de 70,33. Podemos ver que la mayor diferencia se encuentra entre alquilar una casa/departamento entero en comparación con las otras dos posibilidades.

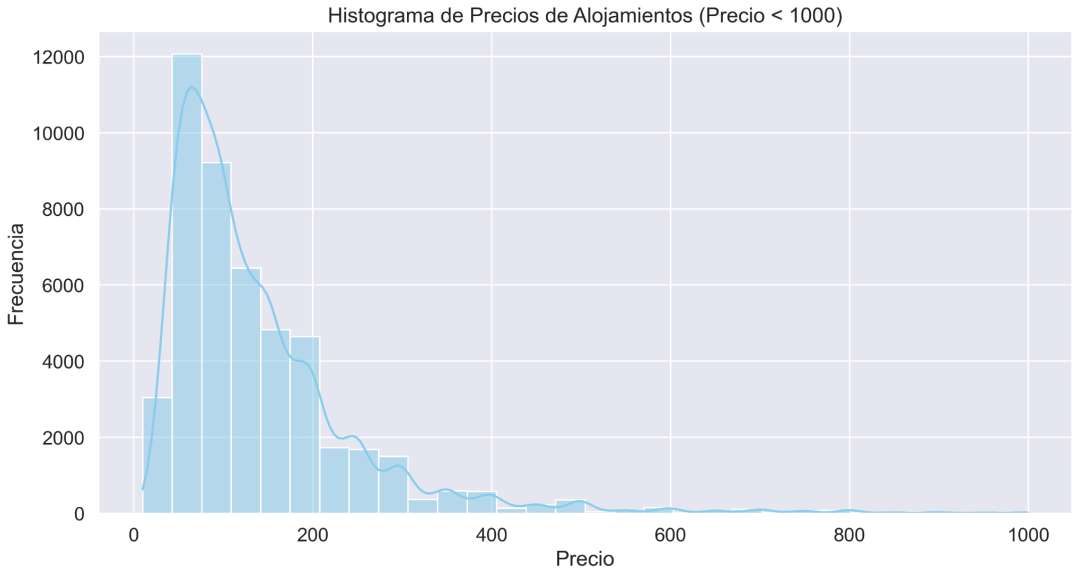

Gráfico 4: Histograma de precios de alojamientos (precio<1000)

Observando el histograma podemos concluir que la gran mayoría de las variables se encuentran entre 10 y 300 dólares por noche. Posterior a ese valor tenemos un porcentaje muy bajo de la muestra. Esto tiene sentido ya que estos serían los Airbnbs de lujo que obviamente son más escasos que los standard.

4.

Gráfico 5: Scatter plot precio y grupo vecindario.

Viendo el Gráfico 5, podríamos considerar la posibilidad de agrupar los vecindarios según sus precios en dos grupos. El primer grupo estaría formado por Manhattan y Brooklyn, mientras que el segundo incluiría a Queens, Staten Island y Bronx.

El primer grupo muestra una concentración de observaciones en precios entre 0 y 4000 dólares en el caso de Manhattan, y entre 0 y 2500 dólares en el caso de Brooklyn. Posteriormente, tienen un número considerable de observaciones dispersas hasta alcanzar un precio de 10000 dólares.

Por otro lado, el segundo grupo también presenta una concentración de observaciones, pero en este caso en un rango de precios más bajo, entre 0 y 1000 dólares. Luego, Queens muestra una dispersión mayor de observaciones hasta los 2500 dólares, mientras que Staten Island y Bronx tienen algunos valores atípicos en sus precios.

En líneas generales, observamos que mientras más "turístico", en promedio el precio de los Airbnb de los diferentes vecindarios son más caros

Gráfico 6: Scatter plot precio y tipo de cuarto.

Se observa una notable diferencia entre las tres categorías en términos de nivel de precios y su dispersión. Los cuartos compartidos presentan los precios más bajos y menor variabilidad. Seguidamente, los cuartos privados muestran un nivel de precios y dispersión intermedios. Finalmente, los apartamentos y casas completas tienen los precios más elevados y la mayor dispersión. Esta tendencia es lógica, considerando que puede existir una mayor diversidad de calidad en los cuartos privados en comparación con los compartidos, y esta variabilidad se acentúa aún más en los apartamentos y casas completos.

5. Lo primero que hicimos fue graficar un *PCA biplot,* es decir graficamos en un mismo gráfico los *loadings* (muestra que tan fuerte, cada variable influencia cada componente) y *PCA plot*. Destacar que aquí decidimos eliminar las variables: "neighbourhood_group", "neighbourhood" y "last_review" ya que las primeras dos eran dummies y si las graficábamos, íbamos a tener tantas flechas que el análisis del mismo no iba a ser fructífero. En cuanto a la variable last_review, al no poder ser transformada en una dummy, no podía ser graficada en el biplot. Una vez aclarado esto, procederemos a analizar el gráfico.

En primer lugar analizaremos los loadings de cada variable, podemos observar que “minimun_nights”,“calculated_host_listing_count” y “price” son tres variables que están ampliamente correlacionadas, esto se debe a que el ángulo entre las tres flechas es pequeño. Siguiendo esta línea, podemos ver que “availability_365” está correlacionado con estas variables también, pero destacar que en menor medida, ya que el ángulo es un poco más grande. Asimismo, podemos observar que “numebr_of_reviews” y “reviews_per_month” están altamente correlacionadas lo cual es muy intuitivo. Ya que a mayor “reviews_per_month”, mayor será el “numebr_of_reviews”.

Por otro lado, podemos observar que el primer grupo de variables tiene una correlación muy baja con la variable “longitude”, ya que el ángulo entre estas y “longitude” parecería ser de 90 grados. De forma similar, podemos observa que el primer grupo de variables parecen tener una correlación baja con “numebr_of_reviews” y “reviews_per_month”.

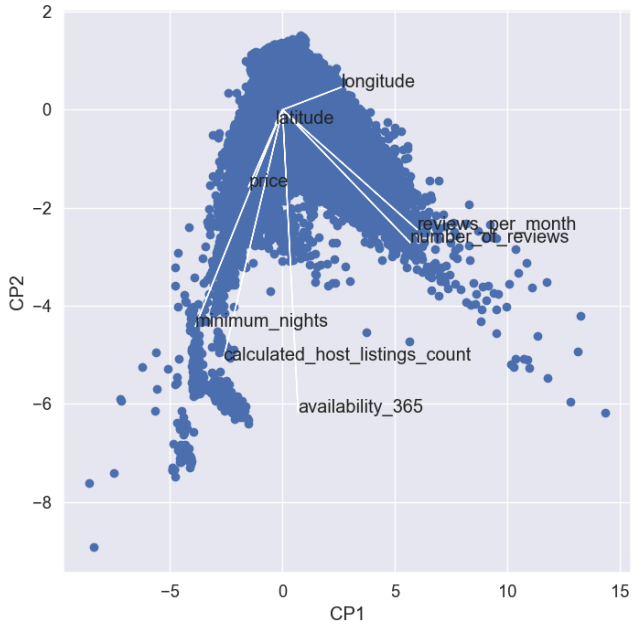

Gráfico 7: Visualización en dos dimensiones: Componente Principal 1 y 2.

Por otro lado, cuanto más lejos estén los vectores del origen de un componente principal, más influencia tendrán en ese CP. Por lo tanto podemos ver que las variables más influyentes para el Componente Principal 1, son “numebr_of_reviews” y “reviews_per_month”. Asimismo, las variables más influyentes para el CP2 son todas exceptuando “latitude” y “longitude”, esto ya que estas últimas dos son las más cercanas al origen. Por último destacar que “latitude” no influye en ninguna de las dos variables ya que está posicionada en el (0,0).

Finalmente, como podemos observar en el Gráfico 8, vemos que el 41,6% de la varianza se logra explicar con dos componentes. Asimismo, podemos ver que si nuestro objetivo es que

nuestros componentes principales expliquen al menos el 80% de la varianza[1], esto se cumple recién con 5 componentes principales. Y si nuestro objetivo es explicar el 100% de la varianza, se cumple con 8 componentes principales. Por último, vemos que cada CP logra explicar los siguientes valores de la varianza: [0.22793762; 0.18996756; 0.1392926; 0.1272027; 0.10764468; 0.08667301; 0.07260065; 0.04868118].

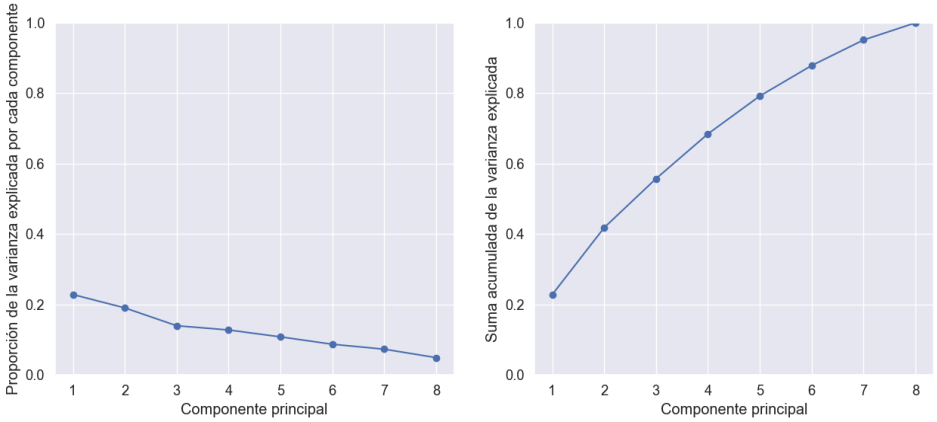

Gráfico 8: Proporción de la varianza y la suma acumulada de la misma explicada por cada CP.

Parte III: Predicción

3.Training MSE: 49125,76. Test MSE: 33932,02.

Creemos que lo más importante de interpretar son los *mean square error*. Los coeficientes no son muy informativos en modelos de predicción.

El error de entrenamiento es mayor al error de testeo. Esto es bastante inusual ya que no tiene mucho sentido que el error en la base de datos que usamos para entrenar el modelo sea mayor que para la parte de la base de datos que usamos para testearlo. Investigando un poco, nos dimos cuenta que esto podía pasar por distintas razones:

a) Se suele recomendar que el manejo de la data se haga después de la división en muestras de entrenamiento y testeo. Es decir, deberíamos haber creado las dummies después de haber separado la muestra.

b) Se recomienda hacer cross validation para que los resultados sean más robustos. Lo que podría estar sucediendo en nuestro caso es que al separar la muestra una sola vez, justo de la casualidad de que el modelo se ajusta mejor a la muestra de testeo.

Intentamos hacer un código para arreglar este problema y obtuvimos los siguientes errores cuadráticos medio. En este caso si obtuvimos los resultados que esperaríamos, pero los coeficientes nos dieron bastante parecido lo cual es raro.

MSE promedio de entrenamiento en validación cruzada: 44436,88. MSE promedio de prueba en validación cruzada: 44941,19.

[1] https://bioturing.medium.com/how-to-read-pca-biplots-and-scree-plots-186246aae063