Renata Emilia Chávez Martínez
A01351716

Reporte de Evaluación

Durante todo el documento se van explicando cada una de las partes del código y se van especificando el funcionamiento del código así como comparaciones entre las bases de datos.

Etapa 1

La primera etapa cargue todas las bases de datos en las cuales primeramente visualice los datos nulos que se tienen en las bases de datos, y hice una limpieza global de los datos tanto numéricos como cualitativos para después visualizar las columnas y posteriormente ir columna por columna para arreglar la base de datos. El último recurso que se utilizó fue el dropna para poder tener bases de datos 100% limpias.

La eliminación de outliers usamos el método de “rango intercuartílico” considerando que este es un modelo mucho más robusto que el de desviación considerando que este puede manejar mejor esos valores atípicos que se tienen dado que tomamos los valores que están por debajo del primer cuartil menos 1.5 veces el rango intercuartílico o por encima del tercer cuartil más 1.5 veces el rango intercuartílico se consideran outliers. Para este proceso separamos las bases de datos entre datos numéricos y strings para solo hacer el análisis en las variables numéricas y no afectar todo el modelo matemático.

Finalmente creamos nuevos dataframes con esas bases de datos ya limpias para poder utilizarla en el análisis de la etapa 2 y 3.

Etapa 2

En esta etapa realizamos filtros para tener bases de datos específicas considerando parámetros específicos como:

a) host_acceptance_rate> 50%
b) Sólo los registros con categoría “superhost”
c) Los registros que no hayan verificado identidad “ not identity_verified”
d) Los registro cuyo property_type = “Private room” y “Hotel room”
e) Los registros que cuenten con bathroom > 1
f) Los registros cuyo precio sea mayor de $10,000 y que sean de tipo “Entire home”
g) Los registros cuyo review_scores_cleanliness > 4.5
h) Los registros cuyo review_scores_value > 4.9
i) Los registros cuya availability_365< 100
j) Los registros cuya host_response_time sea “within an hour”

Es importante considerar que para este tipo de registros fue importante asegurarse que las columnas tengan el tipo de variables adecuadas para que funcione.

Etapa 3

En esta etapa tomamos 3 variables las cuales usamos con las tablas de frecuencia y las otras usamos las funciones de las librerías de python para variables numéricas.

En la primera parte usando la variables de tiempo de respuesta por el host podemos ver que las respuestas el mayor porcentaje de tiempo en que la respuesta es dentro de una hora es en el DF. Esto nos indica que probablemente tenemos mejores servicios al cliente que en Girona y en California, esto considerando que le el valor que más importa tenemos mejor calificación, mientras que en él los demás podemos decir que el lugar con peor respuesta a los clientes de Airbnb es en Girona, España.

Usando los rangos de calificación del servicio, dado que no podemos usar las tablas de frecuencia, con ese tipo de comandos considerando que las variables dentro de la columna son de tipo numérico, mientras que lo que se busca es que sean de tipo alfabético o string. Por eso utilizamos comandos para bases de datos numéricas con gráficos. Podemos ver que el que tiene mejor calificación es México, después tenemos EUA y finalmente esta España con las mejores calificaciones. En España tenemos mucha más dispersión de las calificaciones en los rangos pero realmente las que se tienen mejores calificaciones esta en la ciudad de México.

En cuestión con la verificación del host vemos que en las dos gráficas podemos ver porcentaje así como de número en el DF tenemos mayor número de hosts verificados, posteriormente tenemos a Girona Espana y despues California, por más que hice las dos gráficas para ver la comparación con los números de la muestra de cada una de las bases de datos aún así en México se tiene mayor números de verificaciones así como de porcentajes.

Podemos ver que en el tipo de propiedad de DF y de Girona tenemos alrededor de 8000 o más del mismo y más repetido tipo de propiedad (este es renta de toda la propiedad). En México vemos muchos más tipos de rentas aún y cuando solamente elegí los tipos de propiedad con más de 35 selecciones en las tres bases de datos. En segundo lugar tenemos todo el condominio en California y DF mientras que en España es toda la casa. Vemos una tendencia en todas las bases de datos que en la plataforma de Airbnb de manera global las rentas más repetidas son las de propiedades enteras o cuartos privados y posteriormente otro tipos de propiedades.

Renata Emilia Chávez Martínez
A01351716

Finalmente en cuestiones de ubicación podemos ver que en Mexico tenemos mejor desempeño en cuestión de las ubicaciones así como California. Es importante recalcar que en cuestión de la calificación tenemos que tener en cuenta la muestra que se tiene de cada una de la base de datos.