Fernández Montes Mariel Quetzali / A01707797

Actividad 5 (Evaluación)

De los documentos csv “California” y “Girona\_Esp” y el xlsx “DF\_Mx” ubicados en la carpeta (<https://drive.google.com/drive/folders/18ZivSYijH6jiknL-V0sGtoCUtfVCPAeJ?usp=sharing>), se realizó una limpieza de datos, eliminando valores nulos y datos atípicos para posteriormente realizar filtros a los valores cuantitativos para obtener información específica del conjunto de datos y un análisis univariado de los valores cualitativos y poder obtener conclusiones. El análisis de las bases de datos se realizó por separado. A continuación, se presenta a modo de instrucciones los pasos que se llevaron a cabo para poder realizar el presente análisis de datos:

Etapa 1: Procesamiento de datos

1. Instalamos / importamos las siguientes librerías:

* Instalamos funpymodeling:
  + Obtenemos la librería “freq\_tbl”: Nos ayudará a graficar los valores univariados de las columnas.
* Importamos librería “pandas” y la nombramos “pd”: Ayuda a visualizar el análisis de las variables.
* Importamos librería “numpy” y la nombramos “pd”: Ayuda a realizar operaciones sencillas con las variables que requiramos.
* Importamos librería “matplotlib.pyplot” y la nombramos “plt”: Ayuda a graficar las variables que requiramos.

1. Importamos el archivo y lo convertimos en Data Frame
2. Con la función “info()” obtenemos la información del Data Frame
3. Realizamos el conteo de valores nulos por columna del Data Frame
4. Generamos una limpieza individual por documento y por columna para tener la información lo más específica posible

* Limpieza de archivo “California”
* De la línea 2 a la 5, de la 9 a la 11, la línea 14, la 24, 25 y 33 de código, sustituimos los valores nulos por el texto, ya que cuenta con variables cualitativas.
* De la línea 6 a 8, de la 17 a la 22 y 34 de código contienen variables cuantitativas cuyos valores nulos sustituimos por un 0, no se utilizaron diferentes métodos de sustitución de valores nulos pues no son datos que se puedan relacionar entre sí, por lo que asignar el promedio de la columna o los valores siguientes o anteriores al valor nulo, afectaría el análisis de los datos.
* Las líneas 12, 13 y 23 de código fueron eliminadas ya que no contaban con ningún tipo de información en su interior
* Para las líneas 15 y 16 y de la 26 a la 32 de código contienen variables cuantitativas cuyos valores nulos les asignamos el promedio de toda la columna ya que, en este caso, los valores no tienen gran diferencia entre sí y sí podrían servir como referencia para sustituir los valores nulos.
* Limpieza de archivo “Girona\_esp”
* De la línea 2 a la 7, de la 12 a la 15, de la 18 a la 20, la línea 23, 36, 41, 45, 46, 54 y 55 de código, sustituimos los valores nulos por el texto, ya que cuenta con variables cualitativas.
* De la línea 8 a la 11, la 16 y 17, de la 26 a la 34, de la 37 a la 40, de la 42 a la 44 y la 60 de código contienen variables cuantitativas cuyos valores nulos sustituimos por un 0, no se utilizaron diferentes métodos de sustitución de valores nulos pues no son datos que se puedan relacionar entre sí, por lo que asignar el promedio de la columna o los valores siguientes o anteriores al valor nulo, afectaría el análisis de los datos.
* Las líneas 21, 22 y 35 de código fueron eliminadas ya que no contaban con ningún tipo de información en su interior
* Para las líneas 24 y 25 y de la 47 a la 53 de código contienen variables cuantitativas cuyos valores nulos les asignamos el promedio de toda la columna ya que, en este caso, los valores no tienen gran diferencia entre sí y sí podrían servir como referencia para sustituir los valores nulos.
* Limpieza de archivo “DF\_Mx”
* De la línea 2 a la 6, de la 11 a la 13, la línea 16, la 26 y 27 de código, sustituimos los valores nulos por el texto, ya que cuenta con variables cualitativas.
* De la línea 7 a la 10, de la 19 a la 24 y 36 de código contienen variables cuantitativas cuyos valores nulos sustituimos por un 0, no se utilizaron diferentes métodos de sustitución de valores nulos pues no son datos que se puedan relacionar entre sí, por lo que asignar el promedio de la columna o los valores siguientes o anteriores al valor nulo, afectaría el análisis de los datos.
* Las líneas 14, 15, 25 y 35 de código fueron eliminadas ya que no contaban con ningún tipo de información en su interior
* Para las líneas 17 y 18 y de la 28 a la 34 de código contienen variables cuantitativas cuyos valores nulos les asignamos el promedio de toda la columna ya que, en este caso, los valores no tienen gran diferencia entre sí y sí podrían servir como referencia para sustituir los valores nulos.

1. Realizamos nuevamente el conteo de valores nulos para asegurar que ya fueron sustituidos en su totalidad
2. Agrupamos los datos cualitativos en una variable que llamamos “cuali” y los datos cuantitativos a otra variable llamada “cuanti” para poder manipularlos con mayor facilidad por separado.
3. Para realizar el análisis de los valores atípicos, se utilizó la variable que solo contenía datos cuantitativos y se utilizó el método de desviación estándar que ayuda a limitar las 3 posiciones a la derecha y a la izquierda para identificar los valores que se encuentran dentro de este rango (inliers) y aquellas que no se encuentren dentro de este rango serán considerados como valores atípicos.
4. Imprimimos como NA en un nuevo Data Frame, los valores que son outliers para identificarlos rápidamente.
5. En otro Data Frame imprimimos los valores que no son outliers para identificarlos rápidamente.
6. Sustituimos los outliers por el promedio de la columna en el que se encuentra cada uno de los outliers.

Insigths:

* México fue la zona geográfica de la que se obtuvieron mayor cantidad de instancias para poder ser analizadas, con un total de 24,224 instancias, le sigue Girona con 19,655 instancias y por último California con 6,936 instancias.
* Las 3 bases de datos evaluaban las mismas características (columnas) teniendo un total de 75 columnas por cada base de datos, sin embargo, las bases de datos de California y Girona contaban con 3 columnas que no contenían ningún tipo de información (neighbourhood\_group\_cleansed, bathrooms, calendar\_updated), por lo que fueron eliminadas, utilizando únicamente 72 columnas para su estudio, mientras que CDMX contaba con una columna adicional además de las 3 anteriores que no contaba con información (license), por lo que el análisis de esta base de datos se realizó con 71 columnas.
* A pesar de que California es la base de datos con menor instancias, es la que cuenta con menor cantidad de valores nulos, lo que ayuda a tener conclusiones más certeras al momento de realizar el análisis, le sigue México y por último Girona, el cual duplica la cantidad de valores nulos que las otras 2 bases de datos.

Etapa 2: Extracción de datos

Una vez que se realizó la limpieza de los datos y ya contamos con una variable que guarda los datos cualitativos y otra que guarda los datos cuantitativos, ejecutamos filtros para obtener información relevante del Data Frame. Los filtros ejecutados fueron los siguientes:

* Para la columna “host\_acceptance\_rate” obtuvimos los valores de aceptación con un puntaje mayor al 50%.
* De la columna “host\_is\_superhost” se obtuvieron todos los usuarios considerados “superhost”.
* De la columna “host\_identify\_verified” filtramos los hosts que no han realizado el proceso de verificar su identidad dentro de la aplicación.
* Posteriormente se filtraron los tipos de propiedad que sean habitaciones de hotel y habitaciones individuales, esta información se obtuvo de la columna “property\_type”.
* Como tercer filtro, se filtraron las residencias que contaran con más de un baño, utilizando la columna “bathrooms\_text”.
* Se realizó un filtro en conjunto utilizando las columnas “Price” y “room\_type” que arrojó un Data Frame con las residencias que costaban más de $10,000 y fueran de tipo casa o apartamento.
* Se filtraron las residencias que tuvieran un puntaje mayor a 4.5 estrellas en la categoría de limpieza, utilizando la columna “review\_scores\_cleanliness”
* Se filtraron las residencias que tuvieran un puntaje mayor a 4.9 estrellas respecto al servicio en general, utilizando la columna “review\_scores\_value”
* Utilizando la columna “availability\_365”, se filtraron los valores con puntuación menor a 100 días de disponibilidad.
* Para la columna “host\_response\_time” se filtraron los usurarios que brindan una respuesta dentro de la primera hora en la que el huésped solicita información respecto a la residencia.

Insigths:

* La zona geográfica con mayor aceptación de parte de los huéspedes hacia los hosts es la de Girona, teniendo una aceptación del 87.94% (CDMX 86.03% y California 81.88%)
* Los “superhost” son host que han permanecido por mayor tiempo en la plataforma y han tenido calificaciones sobresalientes respecto a su servicio, tomando en cuenta lo anterior, California es la zona que cuenta con mayor cantidad de “Superhost” en la plataforma alcanzando un total de 41.36% de usuarios considerados como “superhost” (CDMX 36.30% y Girona 15.55%)
* CDMX cuenta con la menor cantidad de host que no han verificado su identidad, alcanzando un 7.94% (Girona 11.50% y California 11.97%)
* La zona que cuenta con más propiedades en renta como habitaciones de hotel o habitaciones individuales es CDMX con un .18% del total de propiedades, (California .06% y Girona .04%), sin embargo, es un dato que se considera poco relevante, ya que la representatividad de este tipo de propiedades vs el total es mínima.
* En las 3 zonas geográficas, más del 50% de las residencias cuentan únicamente con un baño, Girona es la zona que cuenta con más residencias que tienen más de un baño, llegando al 49.25% (California 48.66% y CDMX 37.95%)
* De la cantidad de residencias que son de tipo casa o apartamento y su costo es mayor a $10,000 son muy pocas, la zona que cuenta con mayor cantidad de residencias que cumplen con estas características es CDMX con 0.71% (Girona 0.05% y California 0.03%), lo interesante de esta información sería identificar la frecuencia con la que se rentan estas residencias para validar si tienen alto tráfico o simplemente están siendo ignoradas.
* CDMX es la zona que cuenta con mejor calificación superior a 4.5 estrellas en el aspecto de limpieza, alcanzando un 88.53% (California 88.34% y Girona 47.86%)
* A pesar de que el porcentaje es menor a la tercera parte del total de residencias, CDMX cuanta con la mayor cantidad de residencias con calificación casi perfecta (4.9 estrellas) alcanzo un total de 26.77% (California 19.84% y Girona 13.21%)
* Entra las 3 zonas, CDMX es la que cuenta con mayor demanda en sus residencias ya que solo el 21.18% se encuentra disponible los 365 días del año (Girona 25.34% y California 38.64%)
* CDMX cuenta con el índice de respuesta más rápida durante la primera hora que el huésped solicitó información, teniendo un 70.17% de respuestas durante la primera hora (Girona 62.38% y California 60.86%)

Etapa 3: Extracción de características y visualización de datos

1. De las variables “cuali” y “cuanti” en las que anteriormente habíamos separado los datos cualitativos de los cuantitativos, utilizamos la variable “cuali” para poder realizar un análisis univariado de las siguientes columnas y poder graficar sus resultados:

* De la columna “host\_response\_time” se utilizó una gráfica de barras para visualizar el tiempo de respuesta, el análisis se realizó principalmente de la frecuencia (instancias) de cada opción de tiempo de respuesta.
* Para la columna “host\_acceptence\_rate” se utilizó una gráfica de dispersión para identificar rápidamente en qué zona se aglomeran los puntajes de aceptación para los hosts tomando en cuenta la zona geográfica que estamos analizando, ya sea California - Estados Unidos, CDMX - México o Girona – España.
* Para la columna “host\_identify\_verified” se utilizó una gráfica de pie para tener una representación clara de los hosts que han verificado su identidad vs los que no.
* Para la columna “property\_type” se utilizó una gráfica de área para identificar las instancias que se encuentran en las diferentes categorías de “tipos de hogar” y visualizar la más ofertada en el mercado.
* Para la columna “number\_of\_revies” su utilizó nuevamente una gráfica de dispersión para visualizar rápidamente la aglomeración de los datos y ver la cantidad de reviews que existen en la zona.

Insights:

* CDMX cuenta con el porcentaje más alto de respuestas dentro de la primera hora en la que el huésped solicita información, sin embargo, su segundo indicador más alto es “0” haciendo referencia a que no se cuenta con información sobre el tiempo de respuesta, mientras que Girona y California son más estables en este sentido, ya que cuentan con mayor cantidad de registros sin valores nulos que ayudan a identificar el tiempo de respuesta.
* En las 3 zonas, el nivel de aceptación hacia los hosts se incrementa entre el 70% y el 100% de aceptación, lo que ayuda a identificar que el servicio en general es bueno. Tomar en cuenta que Girona tiene mejor puntuación de los 3, le sigue CDMX y por último California
* Girona cuenta con espacios sin información para poder identificar si todos sus huéspedes tienen identidad verificada o no, mientras que CDMX se pone a la cabeza con la mayor cantidad de host con identidad verificada y le sigue California.
* Girona es la zona que cuenta con más tipos de propiedad para ofrecer al mercado, llegando a un total de 8 tipos, sin embargo, para las 3 zonas geográficas el tipo de residencia más demandado es la renta de una residencia completa sobresaliendo de forma exponencial sobre el resto de las residencias.
* Girona de igual forma, cuenta con el mayor % de visitas dentro de la aplicación sobre las otras 2 zonas.