Organización de datos - TP 1

Grupo 10

Comenzamos con la exploración inicial del dataset, donde investigamos acerca de las variables cuantitativas por un lado y de las cualitativas por el otro.

Para las variables cualitativas, realizamos bar plots, para investigar su frecuencia. Al querer analizar la relación entre estas variables cualitativas y el target is_canceled, realizamos un nuevo gráfico bar plot para estas variables pero con los sub grupos de si fueron reservas canceladas o no.

Para las variables cuantitativas, realizamos histogramas para conocer las distribuciones de este tipo de variables, y para analizar su relación con is_canceled realizamos gráficos de dispersión con la diferenciación de color en los puntos según si habían cancelado o no.

Además en las variables cuantitativas, realizamos un pairplot para análisis en 2 dimensiones entre las variables de este tipo, sumado al dato de si canceló o no.
Y un gráfico de la correlación entre estas, donde no encontramos ninguna correlación lineal fuerte entre ninguna variable.

En cuanto a los valores nulos, encontramos que las siguientes variables cualitativas que presentaban datos faltantes.

- Country
- Agent
- Company
- Distribution Channel
- Meal
- Market segment

Y decidimos para los casos donde el dato null, representaba un “No aplica” dejar estos datos, ya que en realidad no era un valor faltante, si no una categoría más. Para Country y agent se utilizó el ID 0, y en Meal se unificaron las variables “Undefined y SC” ya que ambas representaban “Sin paquete de comida”

Para las demás variables cualitativas al ser pocas filas en relación al total del dataset, tomamos la decisión de eliminarlas

Las variables cuantitativas que presentaban valores nulos

- Children

Y procedimos a eliminar las filas debido a que tan solo eran 2 filas

En la búsqueda de outliers notamos ciertos valores atípicos, pero ninguno como valor atípico que no podría ser un dato real en el contexto dado del data set.
Por lo que decidimos no transformar estos valores ni eliminarlos, y simplemente tenerlos en consideración para el momento que tengamos que utilizar algún modelo, imputación o algoritmo donde estos puedan impactar de mala manera.

También notamos que en el adr había un valor negativo lo cual no tenía sentido en el contexto de lo que significa la variable, y decidimos modificarlo por su media.
En este caso no se detectó como valor atípico analizando el boxplot, ya que había muchos valores 0 y el número negativo era cercano a 0. A pesar de no estar muy alejado de los demás valores, si era un valor sin sentido en el contexto, como fue mencionado.