Nicolás Franco Celano Minig - 107666
Julián Agustín Rando - 108642
Alejo Gomez - 106421

Objetivos

El objetivo de esta primera entrega fue analizar cada variable y hallar relaciones entre ellas y con el *target 'is_canceled'*, además de limpiar variables irrelevantes para el análisis, visualizar los datos obtenidos y decidir qué hacer sobre los datos faltantes del *dataset*.

Desarrollo

Comenzamos duplicando el *dataframe* original para tener una copia de resguardo y se borró la variable *status_date*, ya que se nos indicó que es irrelevante para esta entrega. Además, para evitar inconsistencias se castea la columna *'children'* a *int*, ya que originalmente almacenaba *floats*.

A continuación, se identifican los tipos de variable y se realiza la siguiente manipulación de los datos:

- Se calculan las medidas de resumen de las variables cuantitativas
- Se describen los valores tomados por las variables cualitativas utilizando *barplots* de la biblioteca Matplot.
- Se grafican las distribuciones de las variables cuantitativas

Además, se cambiaron los valores de '*is_repeated_guest'* e '*is_canceled'* por *True* y *False* en vez de ceros y unos.

Finalizada esta visualización inicial se crean las matrices de correlación de las variables cuantitativas y se grafican las relaciones destacables.

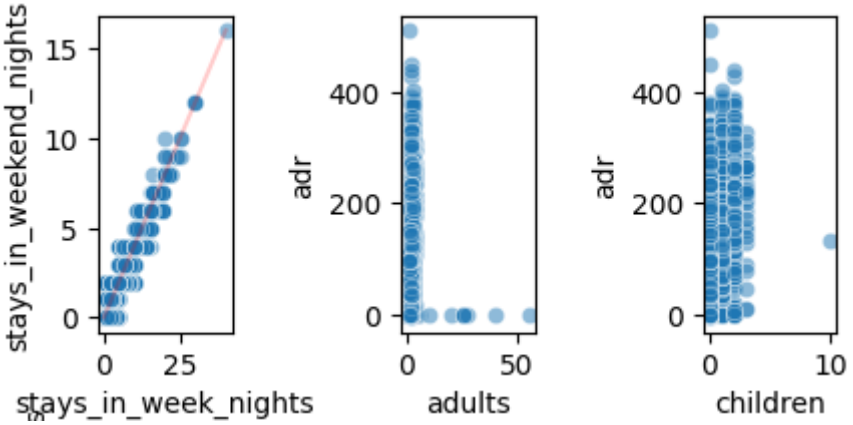

Fig 1: Algunas relaciones destacables entre variables cuantitativas

En cuanto al análisis de las relaciones de las variables con el *target*, creamos una serie de *barplots* y observamos ciertos patrones en los datos obtenidos, como que la gran mayoría de los que utilizaron un tipo de depósito sin reembolso cancelaron su reserva, o que más del doble de las reservas realizadas desde Portugal son canceladas.

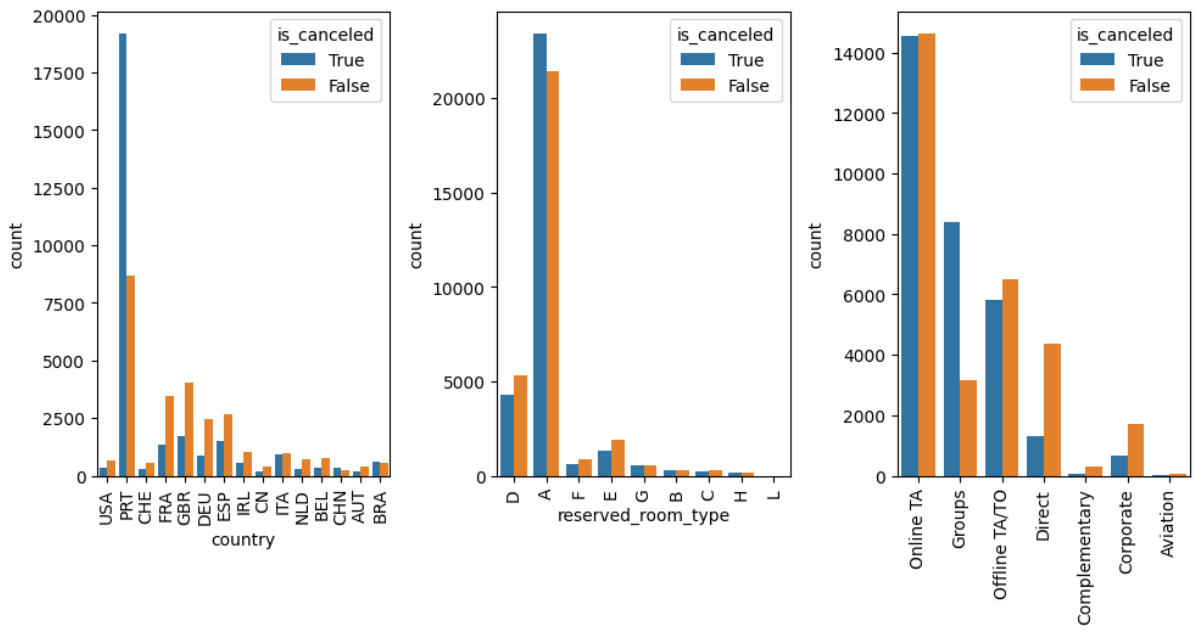

Fig 2: Algunas relaciones entre las variables y el target

También se hizo un análisis de los datos faltantes, donde se decidió que las entradas donde las columnas *country* y *children* son nulas sean eliminadas completamente, ya que conforman un porcentaje ínfimo de la cantidad de datos totales.

Además, los valores *NaN* de las columnas '*company'* y *'agent'* son reemplazados por *'not company'* y *'not agent'*, respetando lo dicho en el *paper*.

Para el análisis de valores atípicos se graficaron tanto las relaciones univariadas como multivariadas

Para las univariadas se hicieron gráficos de boxplot, separando los outliers en moderados (que superan el 1.5**IQR*) y los severos (que superan el 3**IQR*). También se buscan los outliers con un z-score mayor a 3 (o mayor a 5 en caso de que la cantidad de resultados sea demasiado grande para analizar). Además, evitamos eliminar los datos que tienen alguna relación con *is_canceled*, ya que podrían ser importantes para la predicción del target en la próxima etapa del trabajo. Eliminamos los outliers severos que se alejan mucho de los patrones observados, son datos insignificantes para el análisis y suponen una cantidad poco significativa comparada al total de datos.

Para las multivariadas usamos *boxplots* para graficar la distancia de Mahalanobis y posteriormente mostramos los valores atípicos que superan cierto umbral haciendo uso de gráficos de dispersión. Posteriormente se analizan algunas relaciones usando *Isolation Forests*. Para tratar con outliers utilizamos un criterio similar al punto anterior, y nos encargamos de eliminar entradas con datos sin sentido (como que la cantidad de adultos y niños sea cero o *adr* negativo).