Reporte entrega 1

Tareas realizadas

Exploración de los datos

Se comenzó realizando un análisis exploratorio de los datos, donde investigamos por separado las variables cualitativas y cuantitativas del dataset. Analizamos valores únicos que podían tomar las variables, su distribución, correlación y covarianza, y relaciones con la variable target.
Utilizamos un heatmap para poder tener un pantallazo general de la correlación que hay entre las variables numéricas del dataset (ver Figura 1)

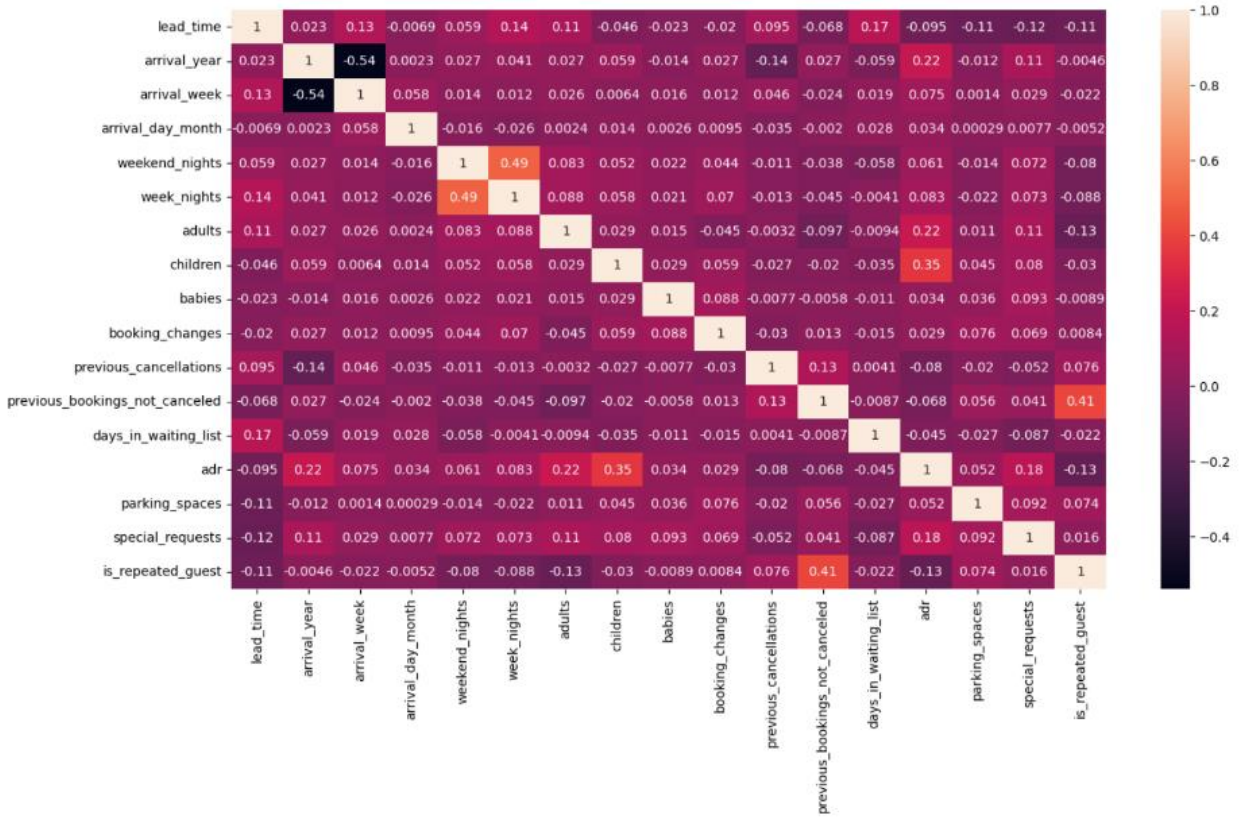

Figura 1

Para la relación de las variables con el target, no notamos que haya algún mes en el que haya más o menos cancelaciones pues, como se puede ver en la figura 2, en todos los meses la distribución del target es bastante equilibrada.

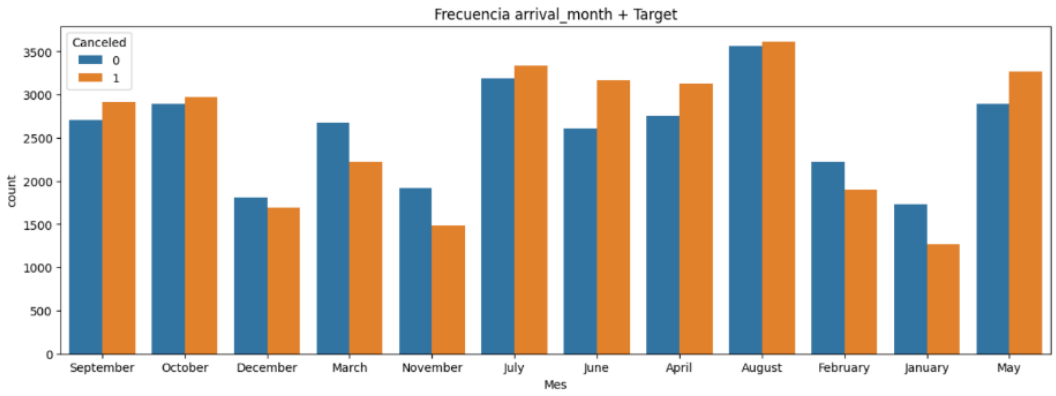

Figura 2

Valores faltantes

Se analizó la cantidad de valores faltantes e inválidos para cada variable y se tomaron distintas decisiones para manejarlos.

Valores atípicos

Para tratar con los valores atípicos en el dataset, se realizaron varias tareas. Primero fueron localizados en la etapa de análisis exploratorio. Luego se realizó un análisis univariado entre las variables que presentaban outliers y, en general, se decidio eliminar estos valores atípicos. Finalmente se realizó un análisis multivariado entre algunas variables que consideramos interesantes.

Información relevante

Para el manejo de valores faltantes y nulos en las variables *children, country, agent* y *company* se tomaron las siguientes decisiones:

- children: La cantidad de datos faltantes representaban menos del 0,01% de los registros así que decidimos eliminar las 4 filas con campo *children* nulo
- country: Dado que el 0,36% de sus datos eran nulos, en éste caso borrar los registros podría presentar un problema ya que no sería insignificante la pérdida de información. Por lo tanto, se optó por completar los datos con el valor "`sin_pais`"
- agent: Debido a que sus datos no son faltantes sino nulos, se optó por designarles una categoría especial a esos datos.
- company: Ante la gran cantidad de datos nulos, aquí se decidió directamente eliminar la variable pues no presentaba información de utilidad.

Se eliminaron la mayoría de valores atípicos encontrados para facilitar análisis posteriores.