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)

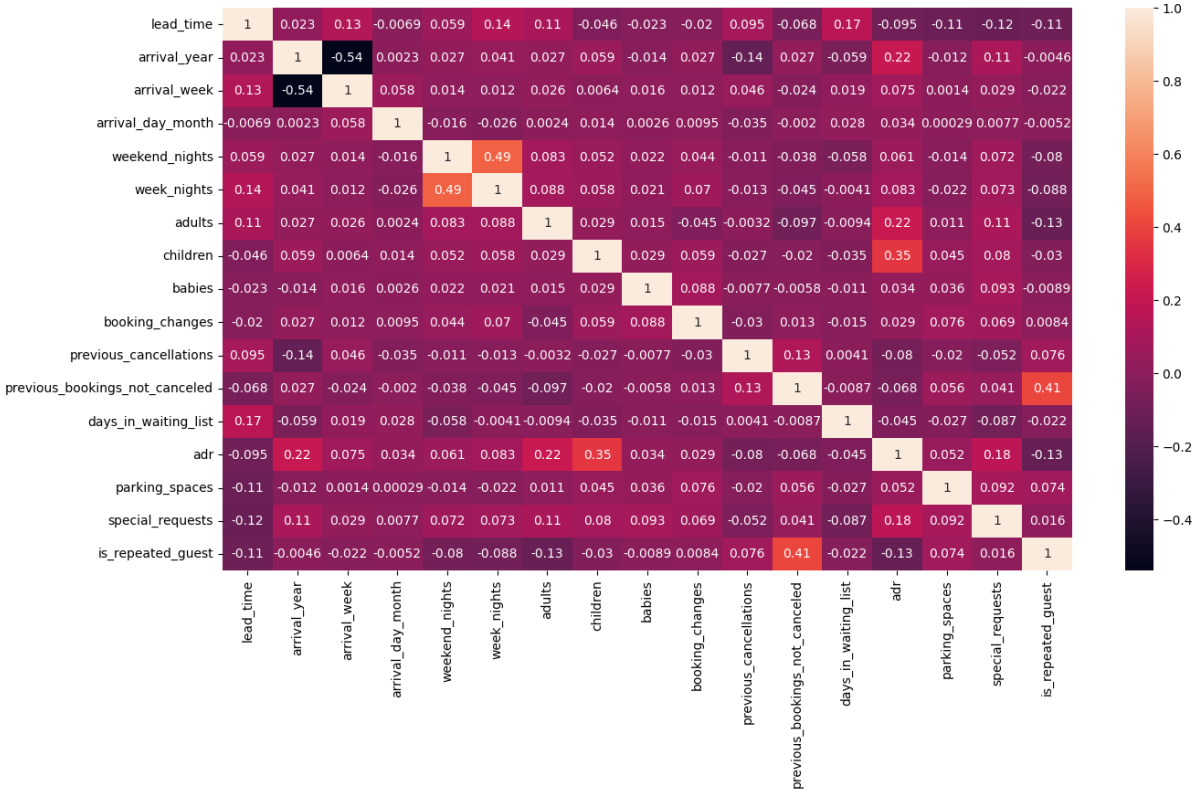

Figura 1

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.

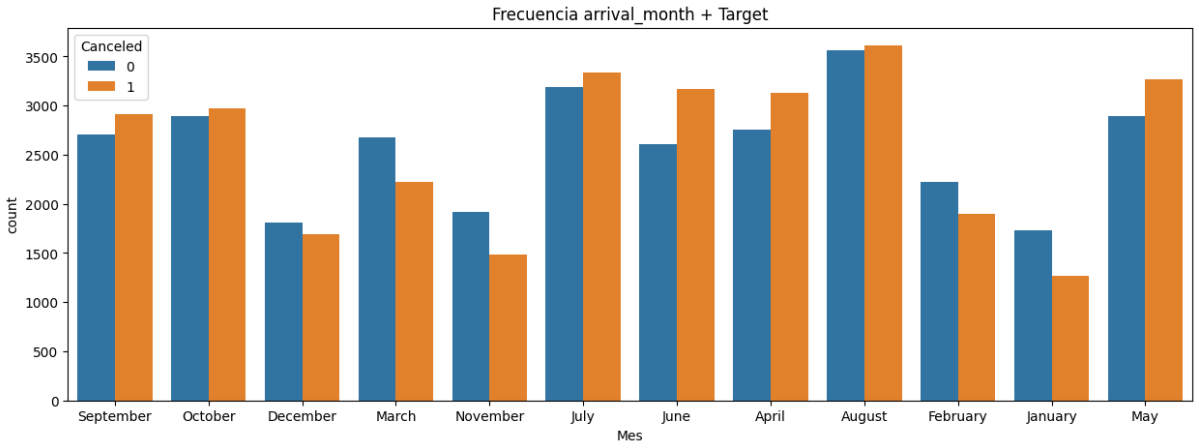

Figura 2

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, de manera que utilizando distintos métodos determinamos cuáles variables presentaban valores atípicos.

Información relevante

Para el manejo de valores atípicos 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: La elección de un modelo de regresión logística multivariada para imputar los datos faltantes, se debe a que esta técnica es adecuada para manejar datos categóricos y binarios. Como la variable presentaba un 12,74% de datos faltantes, la utilización de un modelo de éste tipo podría mejorar significativamente la calidad de los datos.
- company: Ante la gran cantidad de datos faltantes, aquí se decidió directamente eliminar la variable pues no presentaba información de utilidad

Durante el análisis exploratorio, se encontró que el ADR es la única variable numérica que presenta valores negativos. Esto no representa un problema ya que al ser la tarifa media por día, los valores menores a 0 nos indican que hay pérdida de ganancias. Por lo tanto, en éste caso, esos valores tienen sentido.