Checkpoint 1 - Grupo 05

Análisis Exploratorio

Comenzamos analizando el contenido de nuestro dataset y encontramos que contiene inicialmente 61913 registros y 31 variables, siendo 11 de ellas cualitativas y 20 cuantitativas.
Analizamos cada variable individualmente, observando los valores que toma cada una. Para las cualitativas, calculamos la frecuencia con la que aparece cada uno de estos valores en total. En el caso de los gráficos de las variables cualitativas con una gran cantidad de registros, donde se imposibilitaba la visualización correcta del mismo, elegimos mostrar los primeros diez registros. Utilizamos principalmente gráficos de barras.
En el caso de las variables cuantitativas, calculamos la frecuencia con la que aparecía cada uno de los valores presentes en cada columna. Utilizamos principalmente histogramas.

Preprocesamiento de Datos

1. Columnas eliminadas:

Eliminamos las columnas 'arrival_date_year', 'arrival_date_month' y 'arrival_date_day_of_month' ya que decidimos juntarlas y generar una nueva columna 'arrival_date' que contenga toda esta información unificada.

2. Correlaciones detectadas:

A partir del mapa de calor podemos ver relación entre:
- stays_in_weekend_nights con stays_in_week_nights (0.49)
- children con adr (0.35)
- Is_repeated_guest y previous_bookings_not_cancelled (0.41)

La variable is canceled muestra mayor correlación (positiva o negativa) con:
- lead_time (0.29)
- required car parking spaces (-0,23)
- total of special requests (-0,24)

3. Columnas recodificadas:

La columna children era de tipo float, sin embargo sus datos representan enteros (int). Por eso la convertimos.

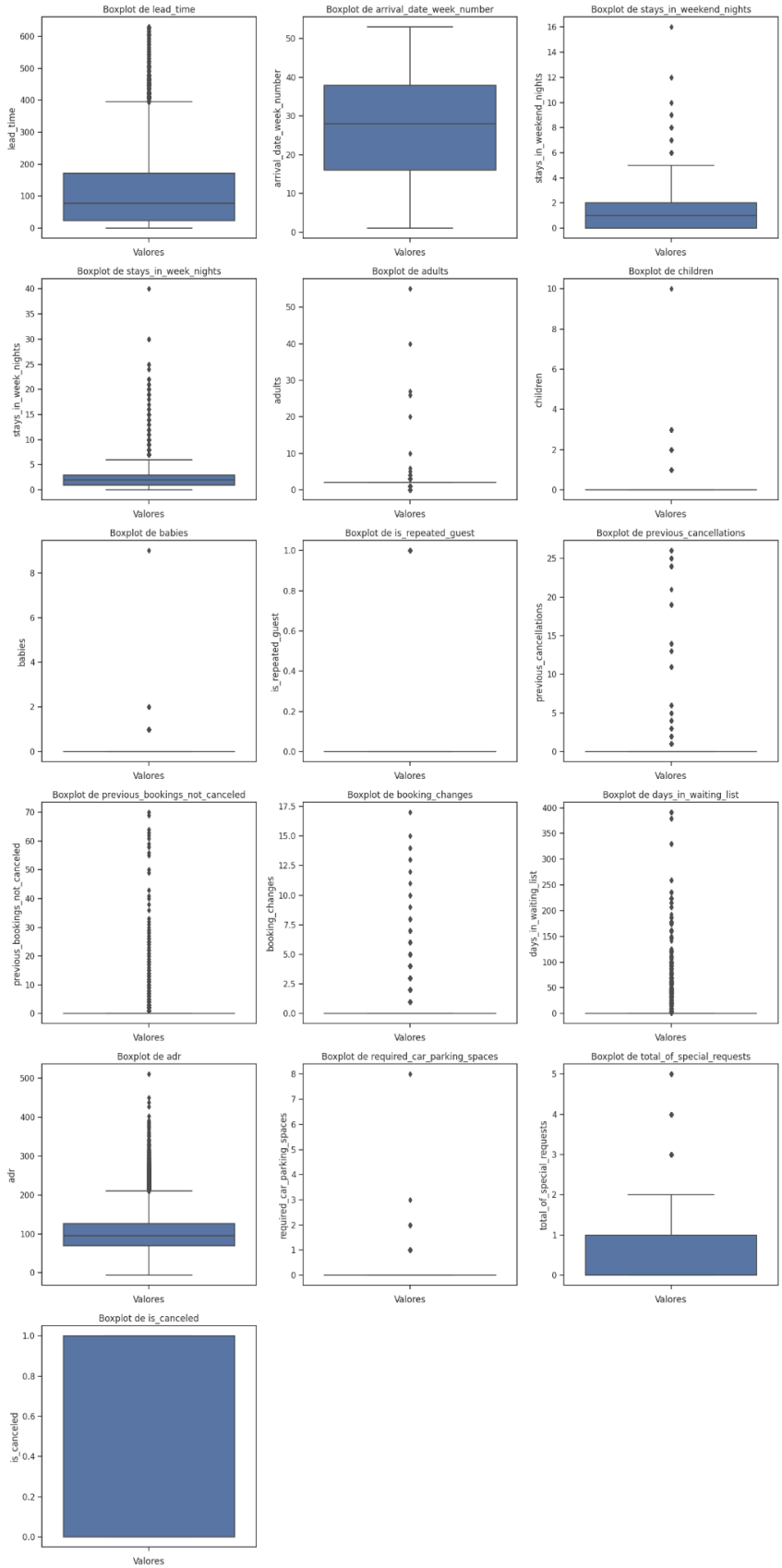
4. Valores atípicos:

Analizamos los valores atípicos con gráficos de box plot y violin plot. A partir de esto analizamos posibles outliers de stays in weekend nights, stays in week nights, total_of_special_requests, adults, children, babies, required_car_parking_spaces.
Por otro lado buscamos en adr si había valores negativos, los cuales no tendrían sentido, y eliminamos esos registros (outliers univariados).
Usamos z-score para total of special requests. Y eliminamos esos registros que superan el umbral (outliers univariados).
Intentamos hacer lo mismo para required car parking spaces pero el porcentaje de outliers era 5%, lo cual nos pareció bastante, por lo que decidimos analizar la variable junto con adults, para luego aplicar el z-score y encontrar un umbral adecuado para los outliers multivariados.
Para las variables stat in week nights y stay in weekend nights analizamos los registros que tenían en ambas variables valor 0. Luego, mostramos un scatter plot con los valores de stays in week nights y stays in weekend nights relacionados para poder detectar outliers utilizando la distancia mahalanobis, y así quitarlos del dataset (outliers multivariados).
Para adults y minors eliminamos aquellos registros que tienen 0 adultos y 0 o más menores (babies y children juntos). Luego analizamos los que tenían muchos adultos junto con la variable customer type para determinar si esos también eran outliers (outliers multivariados).

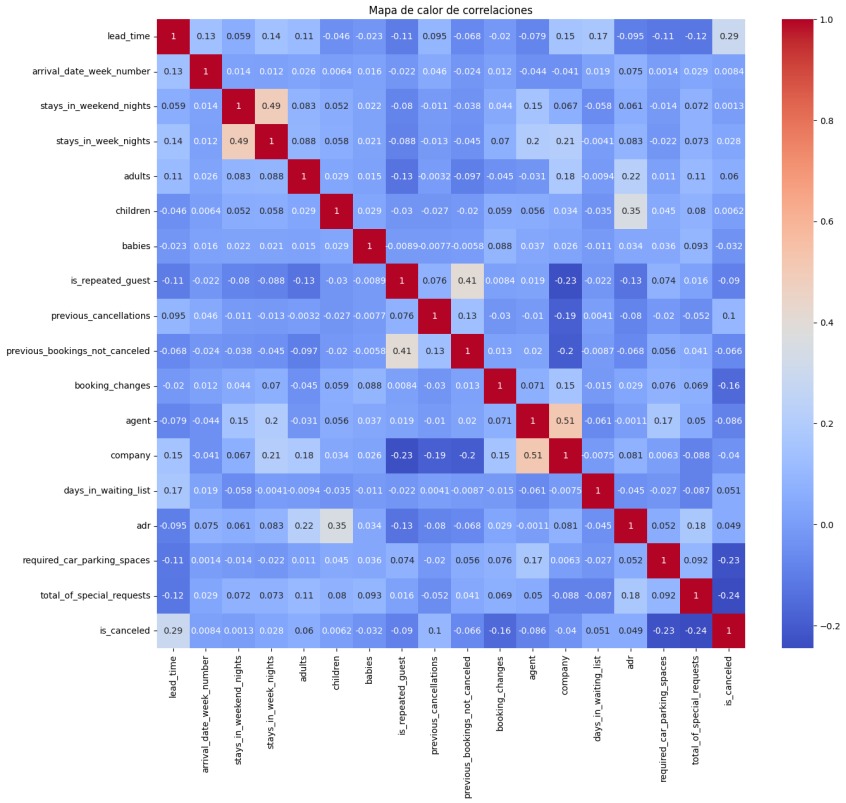
5. Valores faltantes:

Las columnas company y agent tenían altos porcentajes de valores faltantes, 94.91% y 12.74% respectivamente. Es por esto que tomamos la decisión de eliminar las columnas.
También las columnas children y country tenían datos faltantes. Country de un 0,35% y children un 0,0064%.
Para country decidimos utilizar MICE para hacer la imputación de estos datos faltantes.
Para children como solamente eran 4 registros sin datos decidimos eliminarlos.

Visualizaciones



Realizamos un boxplot para analizar posibles outliers



Elegimos un heatmap para visualizar la matriz de correlación entre las variables numéricas.

Tareas Realizadas

Integrante	Tarea
Martín Abramovich	Correlaciones detectadas Análisis de Valores Faltantes Imputación de Datos Armado de Reporte
Iara Jolodovsky	Correlaciones detectadas Análisis de Valores Faltantes Valores Atípicos Armado de Reporte
Tomás Vainstein Aranguren	Desarrollo de variables Correlaciones detectadas Valores Atípicos Armado de Reporte