Checkpoint 1 - Grupo Alan Taylor

Análisis Exploratorio

Los datos que nosotros recopilamos del dataset son: 31 variables donde 19 son variables de tipo cuantitativas y las restantes de tipo cualitativas, 61.913 filas.
A primera vista separamos las variables que nos parecieron más relevantes y las menos relevantes y las graficamos con un pairplot para ver si estábamos encaminados.

Preprocesamiento de Datos

1. Columnas eliminadas:

Después de analizar cada variable eliminamos las menos relevantes como “agent”, ”arrival_date_year”, “arrival_date_week_number”, “arrival_date_month”,, “arrival_date_day_of_month” y “meal”.

2. Correlaciones detectadas:

Nombres de las columnas correlacionadas.
Gráfico de dispersión. Coeficiente de correlación de Pearson.
Notamos que había una correlación entre la cantidad de personas en la habitación (people) y su costo (adr) de 0.43, lo que nos pareció relevante.

3. Columnas recodificadas:

Company: La convertimos en una variable booleana que dice únicamente si la reservación tiene compañía o no, sin importar su id.
Nights: Es el total de noches reservadas, sin importar si es día de semana o fin de semana.
Change Room: Es una columna booleana que dice si la habitación reservada es la misma que la habitación asignada, si lo fue se le asigna 0, de lo contrario se le asigna 1.
Country: La subdividimos en dos variables booleanas que identifica si los países tienen alto porcentaje de cancelaciones (most canceled country) o bajo porcentaje de cancelaciones (less canceled country).

4. Valores atípicos:

Primero limpiamos el dataset de valores ilógicos o que se podían ver a simple vista que fueron mal cargados, como variables cuantitativas que deben ser estrictamente números positivos como adults, minors, adr, etc. tengan cargados números negativos o 0 en los casos que no podían ser 0.
Para analizar outliers univariados utilizamos gráficas de boxplot y z-score modificado en columnas como lead time, nights y days in waiting list. Para el caso

de days in waiting list no eliminamos los outliers, los truncamos al valor máximo de espera sin ser considerado outliers.
Para analizar outliers multivariados utilizamos la distancia de mahalanobis en las columnas minors y adults.

5. Valores faltantes:

Company: Convertimos los datos faltantes (94.5%) en que no tenían compañía en vez de tomarlos como mal ingresados ya que en el paper indicaba eso y nos parecía una variable relevante para el análisis.

Agent: La eliminamos por completo ya que no nos parecía relevante y tenía más del 10% faltante.

Minors: Eliminamos los registros con datos faltantes ya que nos parecía relevante la variable y el porcentaje de datos faltantes (0.006%) nos pareció que era lo suficientemente bajo para despreciar esos registros.

Visualizaciones

Gráfico de distribución de las variables más relevantes

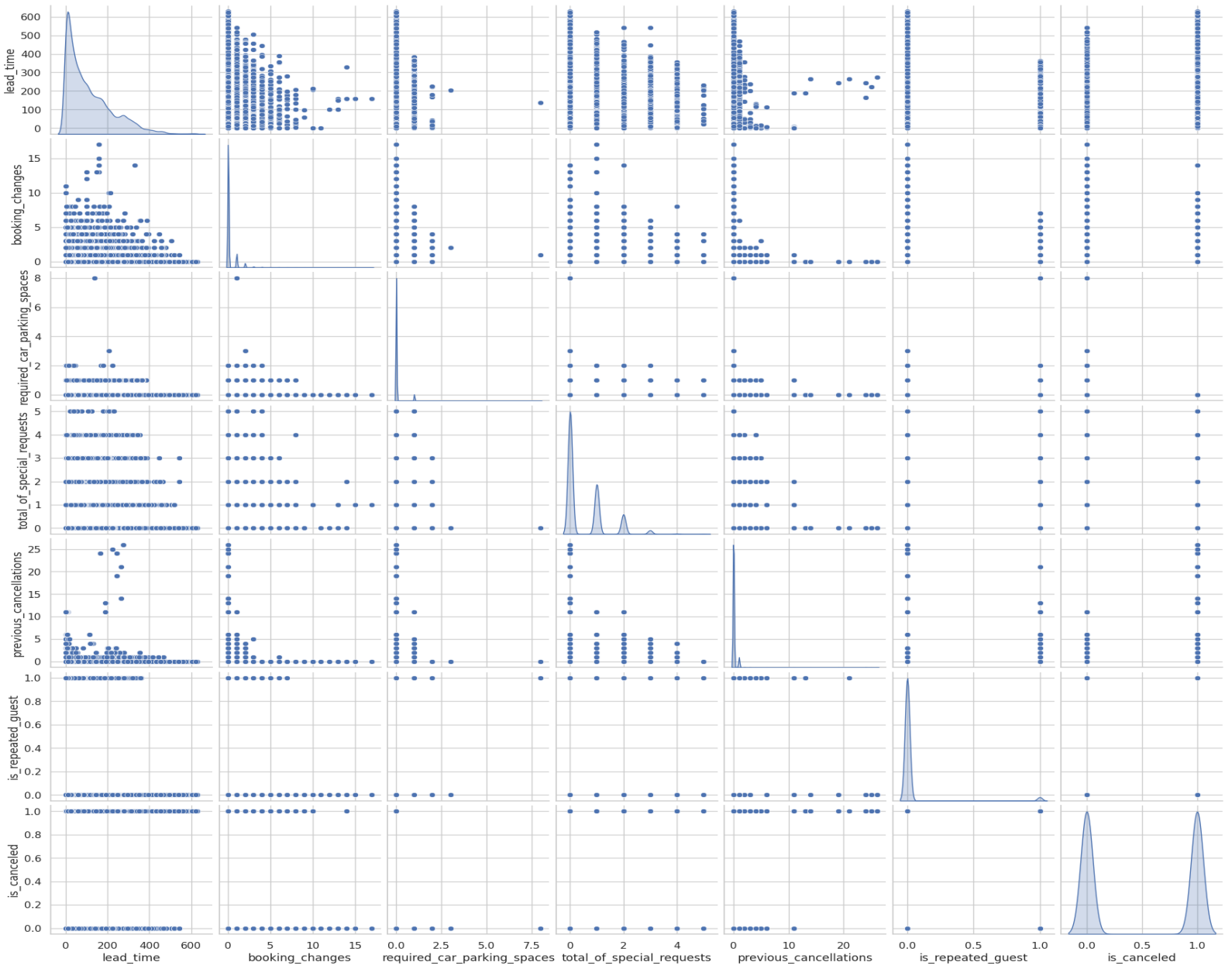
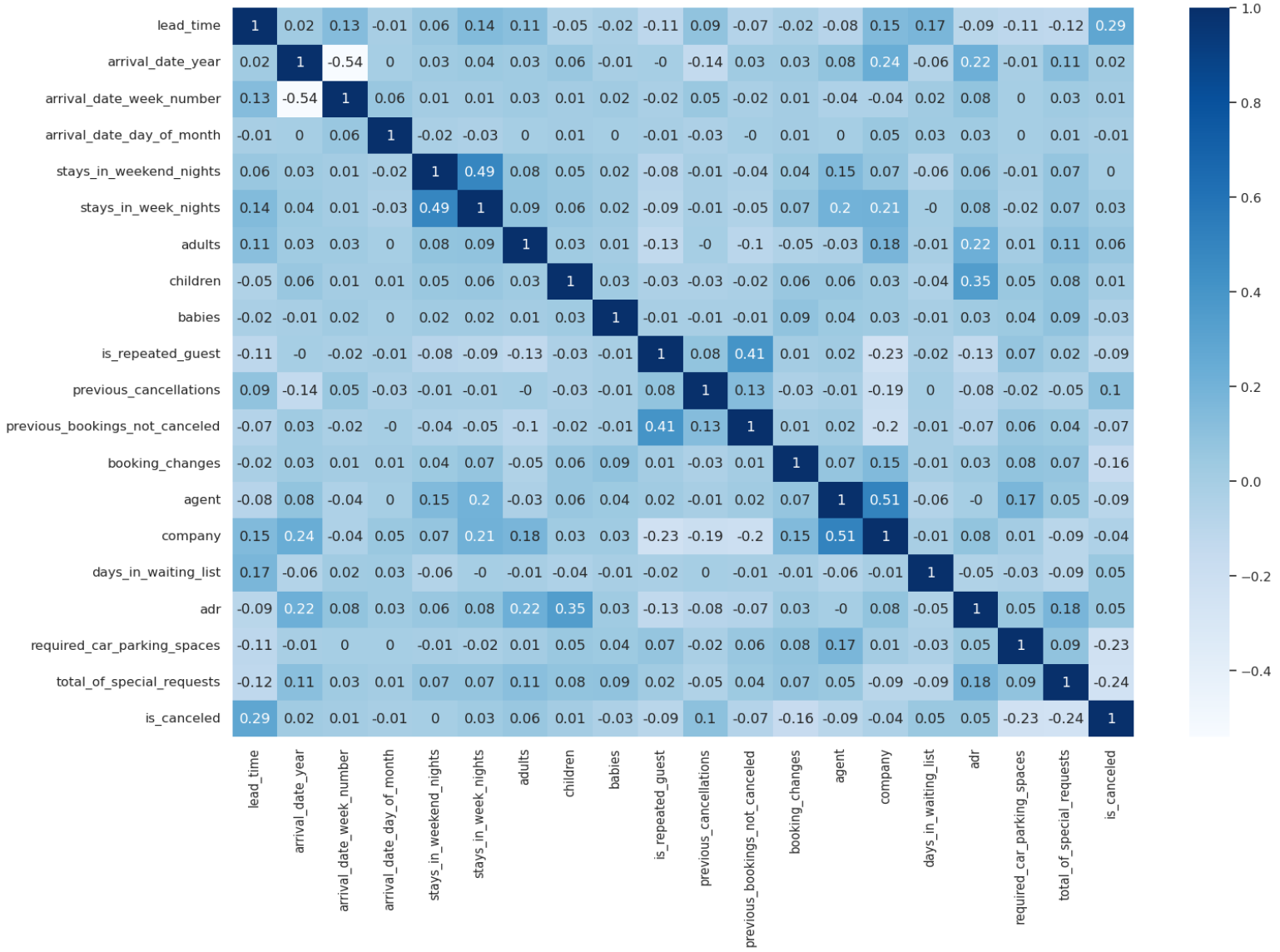


Gráfico de correlaciones



Tareas Realizadas

De manera general, en el trabajo todos los integrantes del grupo formamos parte del desarrollo del análisis de los datos, debido a que en conjunto fuimos trabajando en simultáneo. Por lo tanto todos trabajamos en todas las áreas, un poco menos en algunos aspectos, y más en otros, en resumen se específica más en el siguiente marco.

Integrante	Tarea
Ignacio Sebastian Oviedo	Análisis de Correlaciones Análisis de Variables Análisis de Valores Faltantes Imputación de Datos Armado de Reporte
Estefano Polizzi	Análisis de Correlaciones Análisis de Variables Visualización de Datos Imputación de Datos Detección de Outliers
Santiago Bautista Trezeguet	Análisis de Correlaciones Análisis de Valores Faltantes Visualización de Datos Imputación de Datos Armado de Reporte