Checkpoint 1 - Grupo 26

Análisis Exploratorio

En el dataset encontramos 61913 registros y 31 columnas. Las columnas que destacamos fueron adults (int), adr (int), lead_time(int) e is_canceled (int). Se destacan porque son las variables con las que más trabajamos del dataset y que presentan un mayor grado de correlación con una mayor cantidad de columnas.

Entre las hipótesis o supuestos que tomamos, podemos destacar la de adultos, es inusual que en una reserva haya 0 adultos. Lo mismo para variables que nos indican fechas, si el año de la reserva era mayor al del año actual, o si el día de llegada se excede de los 31 posibles.

Preprocesamiento de Datos

1. Columnas eliminadas:

arrival_date_week_number: dato que no aporta información relevante, puesto que ya contamos con el día, mes y año de la reserva.

stays_in_week_nights y *stays_in_weekend_nights*: decidimos fusionar estas variables para crear otra nueva llamada stays_in_nights, que nos dice la cantidad de noches de la reserva. Una vez creada esta variable, decidimos eliminar las ya mencionadas.

previous_cancellations y *previous_bookings_not_canceled*: misma idea, fusionamos estas variables para crear otra llamada reservations, donde nos refleja la cantidad de reservas que habían hecho antes de la actual. Luego las eliminamos.

2. Correlaciones detectadas:

Las columnas con un mayor coeficiente de correlación de Pearson, y por ende, mayor correlación entre sí son: children y adr (0.36); is_repeated_guest y reservations (0.36); agent y company (0.51); is_canceled y lead_time (0.29).

3. Columnas recodificadas:

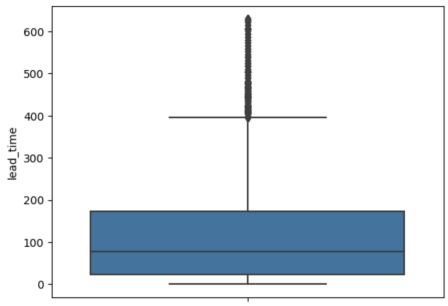
Children: Para la columna “children” reemplazamos los registros que tenían valor NaN por su mediana. Optamos por esta modificación ya que nos parecía pertinente para el análisis que todos los registros tengan un valor en esta columna. El hecho de contar con niños en la reserva es determinante para relacionar los datos.

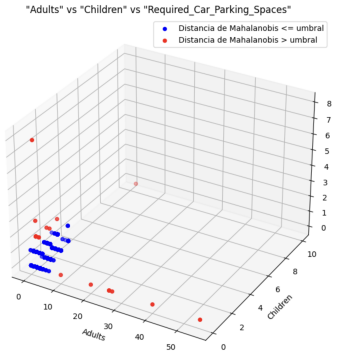
Adults: En esta columna, nos inclinamos por eliminar los registros que contuvieran valor 0, ya que no tenía sentido analizar una reserva que no contiene adultos. Los niños no podrían realizar una reserva por sus propios medios.

Country: En la columna “country” eliminamos las filas que tuvieran valor “null”.

4. Valores atípicos:

Realizamos un análisis de los valores atípicos univariados sobre algunas columnas con valores cuantitativos. Estás serían: ['lead_time', 'arrival_date_day_of_month', 'nights_of_stay', 'adr']. Para el análisis de valores atípicos univariados utilizamos el método Z-Score Modificado. Para el de valores multivariados utilizamos la Distancia de Mahalanobis de a tres variables. Para “lead_time” encontramos atípico el valor 600. Para “arrival_date_day_of_month” no encontramos ninguno, en “nights_of_stay” 56 y por último para “adr” destacamos $510. En el análisis multivariado pusimos el foco en el grupo formado entre [“adults”, “children” y “required_car_parking_spaces”], encontrando como valor atípico por ejemplo una reserva de 2 adultos, 0 niños y la solicitud de 8 espacios de estacionamiento.





5. Valores faltantes:

Las columnas que tenían datos faltantes eran company (3.06%), agent (0.41%), country (0.01%) y children (0.00%). Estos porcentajes son con respecto al total de datos del dataset. En company y agent se decidió imputarlos con otro valor, ya que sus NaNs eran posibles valores, a parte de que nos lo decía él papers. En country decidimos simplemente eliminarlos, ya que eran registros pocos confiables. Finalmente en children se decidió reemplazarlos por su mediana, ya que no podíamos deducir que valor podría tomar.

Visualizaciones

Gráfico de relación entre is_canceled y adults. En este heatmap apreciamos que a menor cantidad de adultos, existe un mayor número de reservaciones tanto canceladas como no canceladas, con predominancia de las no canceladas cuando se trata de un solo adulto.

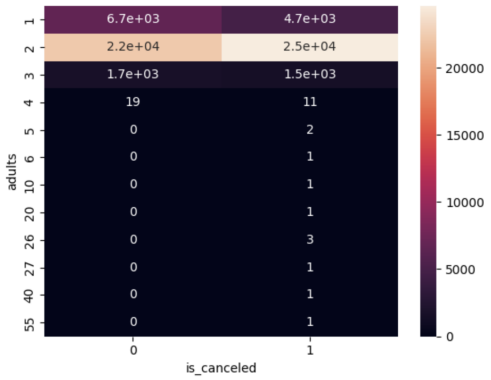
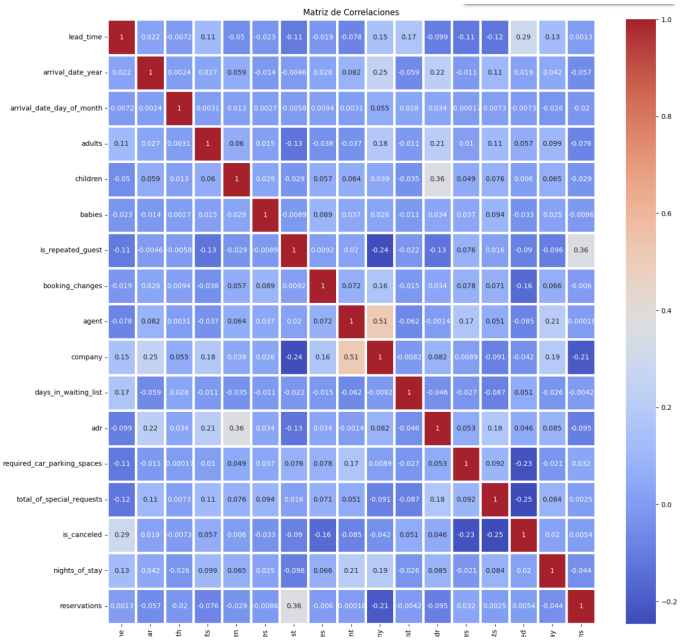


Gráfico de matriz de correlaciones entre todas las variables. Los valores más cercanos a uno y de color más cercano a rojo son los que presentan un mayor grado de correlación entre sí.



Tareas Realizadas

Integrante	Tarea
Agustín Vallcorba	Análisis de calidad de datos Transformación de datos Relación entre las variables Análisis de variables cualitativas y cuantitativas Armado de Reporte
Kevin Carbajal	Medidas de resumen (var cuantitativas) Valores atípicos Transformación de datos Relación entre las variables Análisis de variables cualitativas y cuantitativas Armado de Reporte
Nicolás García	Valores atípicos Transformación de datos Relación entre las variables Análisis de variables cualitativas y cuantitativas Armado de Reporte