Checkpoint 1 - Grupo 16 GPWin

Análisis Exploratorio

El dataset a analizar posee 31 columnas y 61913 observaciones. Existen solo 4 columnas con valores nulos. De los features, consideramos más relevantes en principio, “deposite_type”, “country” y “lead_time” dado que son las que tienen una correlación más fuerte con el target.

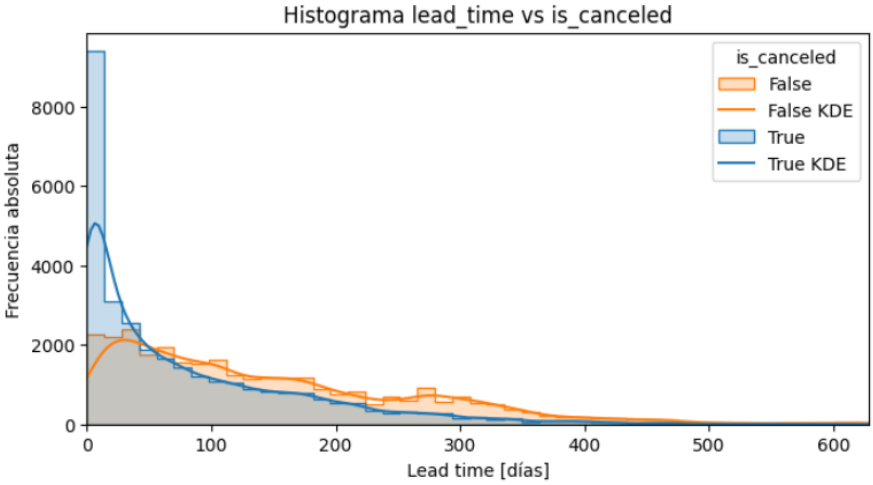
Preprocesamiento de Datos

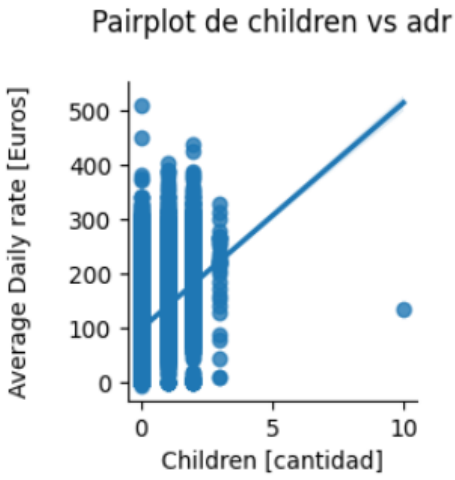
1. Columnas eliminadas:

   *La única variable que decidimos eliminar completamente fue “company” dado que al tener un 95% de nulos, no contiene información útil.
   *La columna Id si bien no se ha eliminado, no será tenida en cuenta dado que al ser una clave de identificación para cada observación, no aporta información.

2. Correlaciones detectadas:

   Como se explicó líneas arriba, las correlaciones más fuertes encontradas con el target son:
   “lead_time vs target” / Coef de Pearson: 0,29
   “deposite_type vs target” /Coef de Pearson: 0,43
   “Country vs target”/ Coef de Pearson: 0,28
   A su vez también detectamos variables con fuerte correlación entre sí, las cuales o bien podrían representar información redundante o bien ser un aporte significativo de información para la predicción del target, por ejemplo *“children vs adr”*.



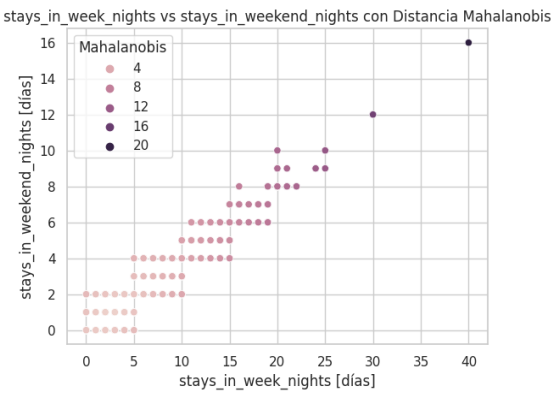


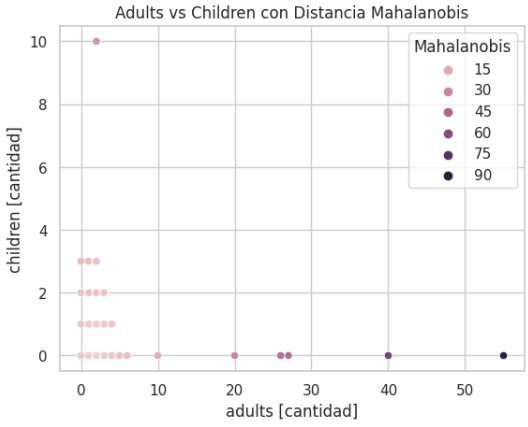
3. Columnas recodificadas:

Se realizó un encodeo de las columnas categóricas y se llevó a cabo utilizando la librería de sklearn instanciando el elemento LabelEncoder.

4. Valores atípicos:

Para el análisis de los valores atípicos de forma univariada, utilizamos los gráficos “boxplot” y el conocimiento del dominio del problema, mientras que para el análisis multivariado lo llevamos adelante utilizando la distancia de Mahalanobis
Las variables que presentan de forma univariada valores atípicos, bajo el criterio de quedar fuera de los “bigotes” del gráfico son: lead_time / stays_in_weekend_nights /stays_in_week_nights /adult/children/babies/previous_cancellations/previous_bookings_not_canceled/booking_changes/days_in_waiting_list/adr/total_of_special_requests.
Calculamos la distancia de Mahalanobis para las variables stays_in_week_nights y stays_in_weekend_nights encontrando visualmente una distancia óptima de 12, mientras que dicha distancia de corte para el análisis de adults y children no da como resultado un valor de 20.
Para el resto de las variables se consideró como criterio de eliminación de outliers, todos aquellos valores que se encuentran fuera de los “bigotes” del boxplot.





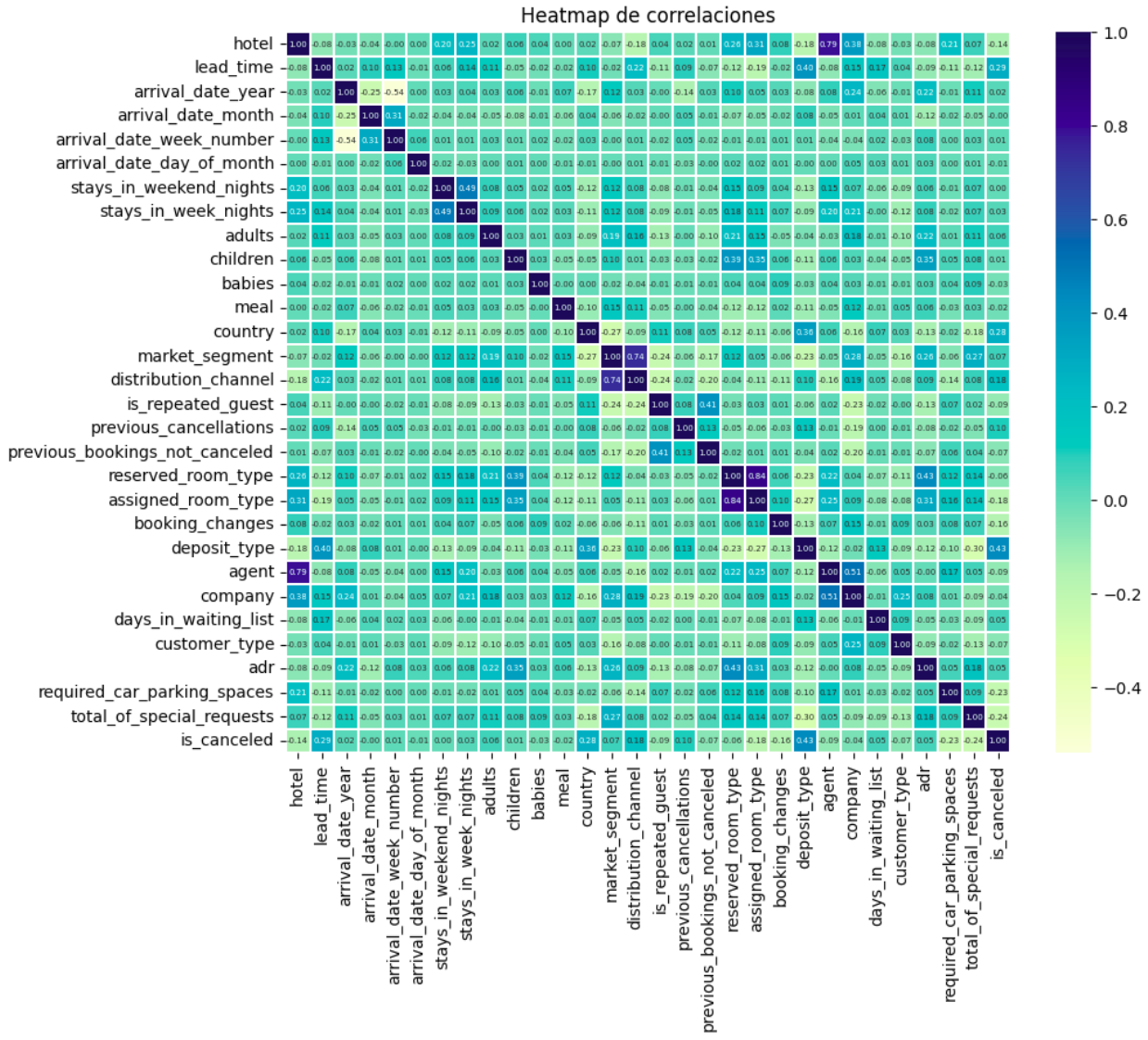
5. Valores faltantes:

*Solo las columnas “company”, “country” y “agent” tenían registros nulos. (95%,0,3%,13%).
Se decide eliminar la variable company debida a la escasa información que aporte y realizar una asignación utilizando MICE, para las variables “agent” y “country”

*Las variables “distribution_channel” y “market_segment” poseen valores “undefined”, sólo 6 observaciones entre ambas (cantidad no significativa) por lo que se decide eliminar dichos registros.

Visualizaciones

El primer gráfico y uno de los más representativos, consiste en un heatmap generado a partir de la matriz que posee los coeficientes de Pearson de las correlaciones entre todas las variables.
De esta forma, podemos obtener rápidamente y de forma cuantitativa la relación entre las variables unas con otras y fundamentalmente entre cada una y el target.



Tareas Realizadas

Indicar brevemente en qué tarea trabajo cada integrante del equipo, si trabajaron en las mismas tareas lo detallan en cada caso (como en el ejemplo el armado de reporte).

Integrante	Tarea
DIEM, Walter Gabriel	Detección de Outliers Calculo de Mahalanobis Análisis de Correlaciones Análisis de distribuciones
MAIOLO, Alejandro	Análisis de Correlaciones Armado de Reporte Encodeado de Datos Análisis de distribuciones
RUIZ, Karen Belén	Análisis de Correlaciones Análisis de Valores Faltantes Imputación de Datos Análisis de distribuciones