Martin Laurito, Lucas Franciulli, Dante Alejandro Finci

Informe Checkpoint 1

En el primer checkpoint del trabajo práctico realizamos las siguientes tareas de análisis exploratorio y preprocesamiento de datos:

- ## Exploración Inicial

  Detallallamos el tipo de las variables del Dataset y para las cuantitativas calculamos sus medidas de resumen mientras que para las cualitativas mostramos posibles valores que toman y cuán frecuentemente lo hacen.

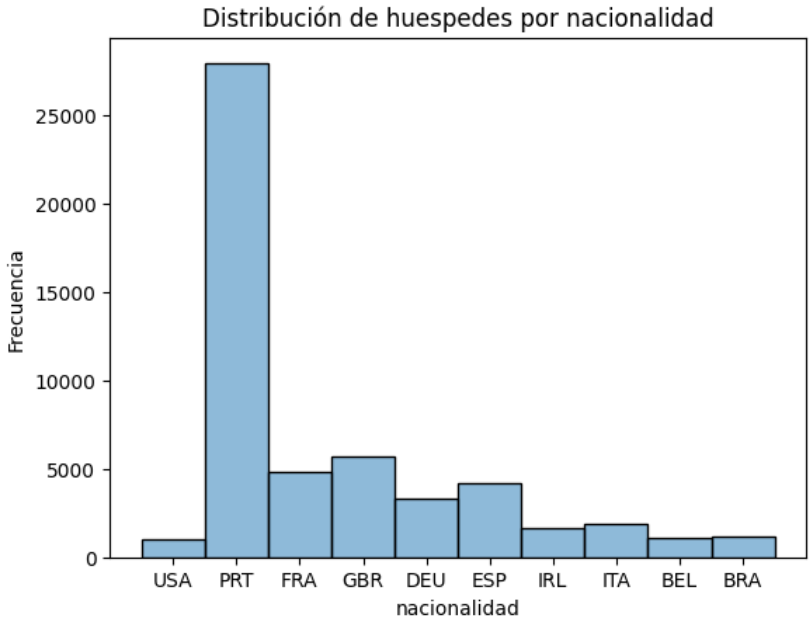
  Luego, Observamos cada variable y determinamos que las variables **meal** e **id** son irrelevantes para el análisis. Realizamos gráficos de las diversas distribuciones de las variables. Analizamos las correlaciones existentes entre las variables y la relación de las variables con el target: en el caso de las variables cuantitativas mediante la correlación de Pearson y en el caso de las cualitativas mediante gráficos.
  Algunas de las características del Dataset que encontramos en esta parte del análisis son:

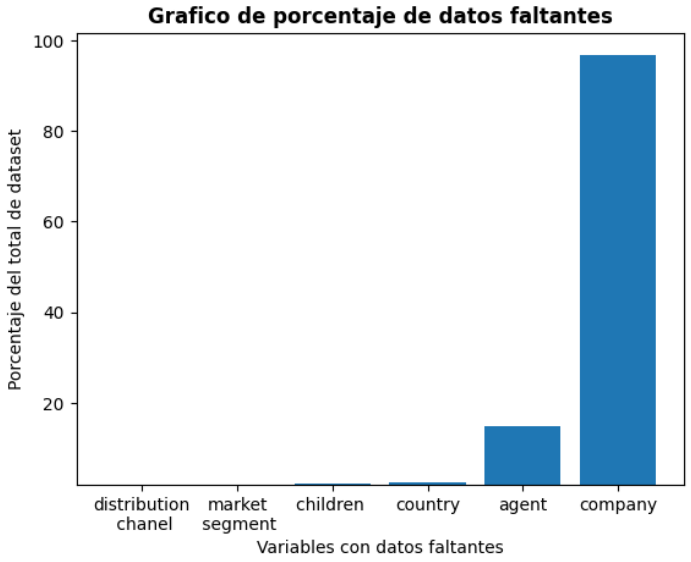
1. Para la variable distribution_channel cuando toma el valor **TA/TO** (que recordemos que representa al *83.8%* de los distribution_channel) cancela las reservas un *54.21%* de las veces.
2. Si la variable *required_car_parking_spaces* toma un valor diferente a 0 este NO cancela su reserva, en cambio en el caso contrario tiene un 52.58% de chances de hacerlo.
3. Que *lead_time* y *previous_cancellations* tienen cierta correlación positiva con el target y que *total_of_special_requests* y *total_of_special_requests* tienen cierta correlación negativa
4. Los huéspedes proveniente de PRT (representa un 45% de los valores que toma la variable country) tienen más del doble de chances de cancelar que de no hacerlo.
5. Para la variable deposit_type: No Deposit (representa un 83% de los valores) muestra considerablemente mayor chance de no cancelar que de hacerlo y Non Refund es solo para cancelaciones.

- ## Visualización de los datos

  En el apartado de visualización de datos decidimos graficar las variables que vimos más importantes para que sean vistas. Utilizamos varios tipos de gráficos, entre ellos: Bar Charts, HeatMaps, Box Plots y Scatter Plots.

  A continuación vamos a ver un ejemplo de alguno uno de estos:





A la izquierda podemos observar un Bar Chart mostrando las 10 más comunes nacionalidades de los huéspedes y a la derecha un Bar Chart mostrando las variables con datos faltantes y cuanto porcentaje de datos faltantes tienen.

- ## Datos Faltantes

Los datos faltantes del dataset estaban en variables children, country, agent y company que eran los que tenían Nans.Y también, había otro tipo de dato 'Undefined' que estaba en las variables market_segment y distribution_chanel por lo tanto se los podría definir como dato faltante ya que ese tipo de dato no está definida en categorías. Nos fijamos también que no haya filas duplicadas, pero no había.

Analizamos los datos con Nans, vimos el porcentaje de la cantidad de datos Nans de cada variable con respecto al total de datos. Cómo country contenía un 95% de datos Nans decidimos que lo mejor era eliminar esa variable y trabajar sin ella para la predicción. Para el resto de los casos, analizamos el porcentaje de la apariciones de cada dato en cada variable y, para cada una, imputamos el dato de mayor porcentaje.

- ## Valores atípicos

El análisis de valores atípicos no se puede aplicar a variables cualitativas por lo que estas se excluyen del análisis.
Calculando el z-score de cada variable cuantitativa separamos las que presentan valores atípicos y luego mediante boxplots analizamos los valores atípicos de cada una.
Luego comparamos varias variables juntas con un análisis multivariado para poder encontrar outliers de esta forma. Utilizamos el método de Isolation Trees para lograr esto.