Reservas de Hotel

Introducción:

El presente trabajo práctico consiste en analizar un dataset de reservas de hoteles con el fin de entrenar modelos predictivos y ensambles de modelos (en futuros checkpoints) el cual nos permita predecir si una reserva nueva será cancelada o no. Para ello en este primer checkpoint nos concentramos en realizar el análisis exploratorio del dataset, donde trabajaremos los siguientes puntos: Duplicados, Distribución de las variables, Correlación de variables, Datos faltantes, Imputación de datos y Outliers

Checkpoint 1: Análisis Exploratorio y Preprocesamiento de Datos:

Comenzamos realizando un **análisis de duplicados**, para esto usamos el método *duplicated()* de pandas, si lo haciamos sobre todas las variables nos daba cero, pero si le sacamos el id (porque al fin y al cabo no nos interesa a nivel análisis por sus caracteristica) nos daban aprox 17000 filas, los cual nos pareció una cifra muy alta por lo que procedimos a estudiarlas un poco manual y observamos que habían algunas que no eran repetidas y otras que sí pero de a muy pocos bloques, por lo que podrían ser casualidades, entonces al final decidimos no eliminar ninguna.

Luego seguimos con el **análisis de las variables** que tenía el dataset, de las cuales se observaron que tenemos 31 distintas y que podían ser de 3 tipos, object (string), float64 e int64. Estas fueron divididas en variables cuantitativas y cualitativas para su mejor y correspondiente tratamiento, para las primeras se observó su media, moda, mediana, mínimo y máximo. Y para las siguientes, se observó cuántos valores posibles podía tomar y cuales eran los más comunes.

Con respecto a la **distribución de variables** realizamos barplots las que tenían poca variación de valores e histogramas para las demás excepto id, country, debido a que tenían muchos valores distintos y queba poco legible el gráfico.

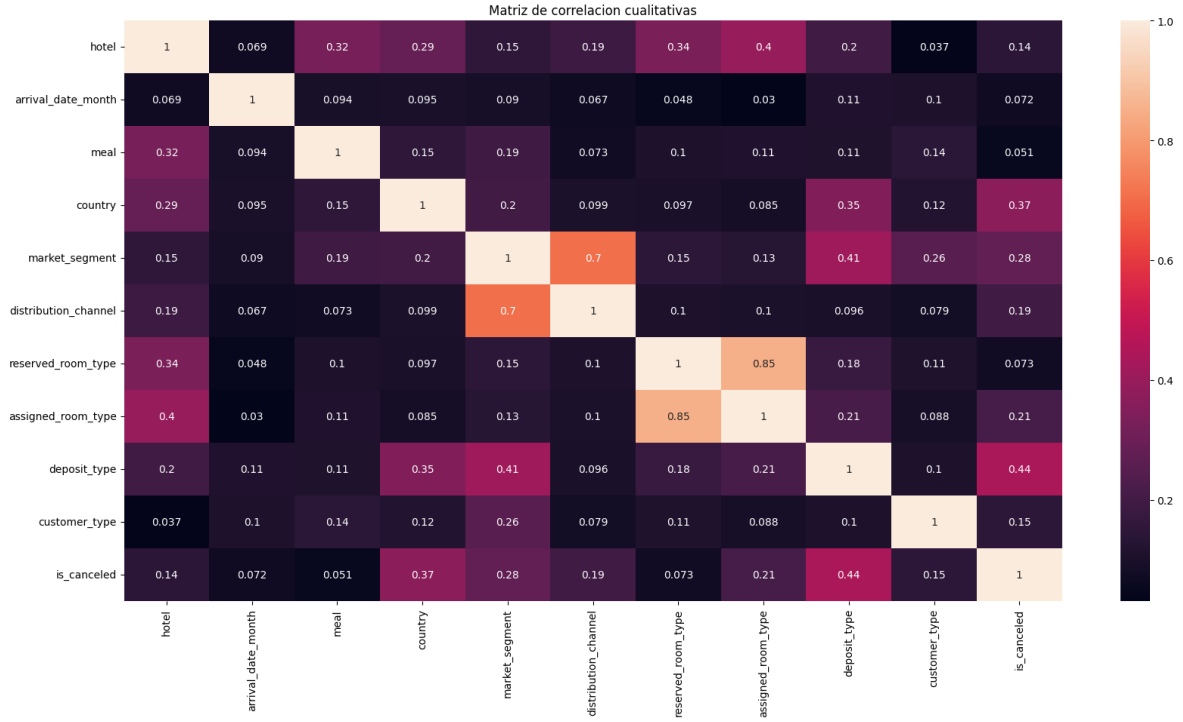
La **correlación de variables,** para las variables cuantitativas utilizamos el coeficiente de correlación de Pearson con el método *.corr()* de pandas y para las cualitativas implementamos la V de cramer, luego ambas matrices de correlación la graficamos en un heatmap con *sns*. A su vez las variables con mayor correlación fueron *lead_time* (Días entre la realización de la reserva y el día de llegada) con correlación positiva por el lado de las cuantitativas, luego *deposit_type* y *country* por el lado de las cualitativas

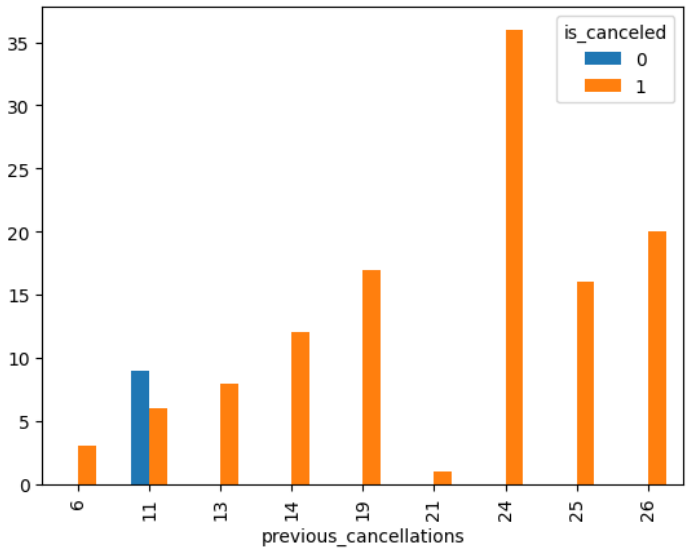
Luego, respecto a los **datos faltantes e imputación de datos** observamos faltantes sobre muy pocas variables (Children(4), country(221), agent(7890) y company(58761)) ver decisiones y justificación en el archivo (diff Reentrega). Luego también se hicieron algunas imputaciones sobre variables que tenían un string “Undefined”

Para el análisis de **outliers** se hicieron de forma univariada y multivariada, para la **univariada** realizamos Boxplots para tener una idea visual y a su vez implementamos una función que nos calculaba los límites que definían si un outlier es severo o no (Usamos el

rango intercuartílico 3) y las variables que tenían menos de 200 outliers severos le aplicamos una eliminación automática, luego para los demás analizamos manual, algunos los eliminamos y otros no porque en realidad eran valores super comunes y válidos. Luego en el análisis **multivariado** hicimos scatterplot entre variables que había mayor correlación y les calculamos la distancia de mahalanobis, acá nos vimos que algunos valores outliers ya habían sido eliminados en la forma univariada y también valores dentro de lo normal

Una vez estos procesos terminados, volvimos a graficar histogramas para verificar que no se hayan modificado casi nada las distribuciones. Cabe aclarar que durante todo el proceso se utilizaron 2 df sobre las cuantitativas y otros 2 sobre las cualitativas, en los filtrados trabajabamos y los final tenían los ids para luego hacer el merge.
Gráficos interesantes a mostrar:





En este último gráfico se observa que cuando tenían cancelaciones previas era muy probable que volvieran a cancelar 👀🤨