7506R_TP1_GRUPO06_CHP1_REPORTE

Sebastian Makkos, Rodolfo Albornoz, Valeria Brzoza

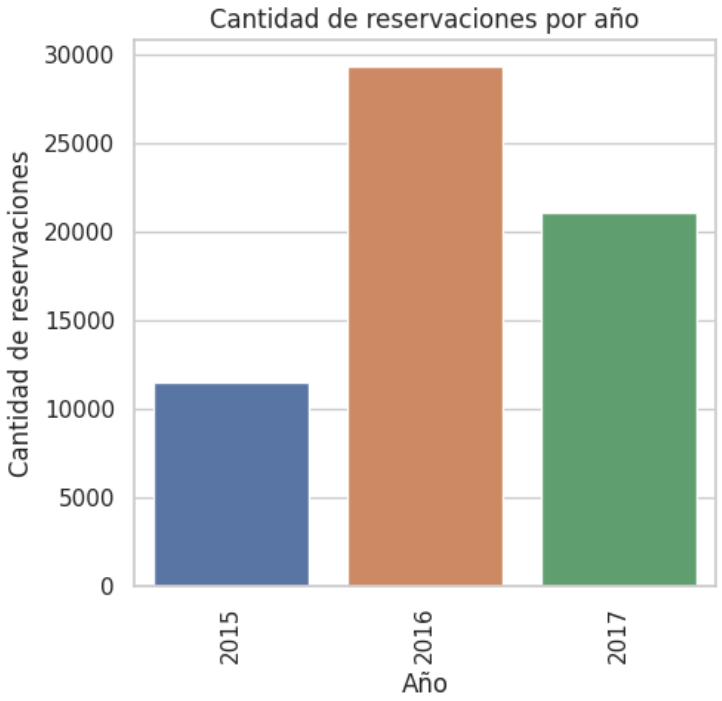
Organización de datos, 1C23

A partir del dataset hotels_train.csv, pudimos realizar un análisis exploratorio de las variables mismas y las relaciones que pudimos encontrar.

Pudimos diferenciar los tipos de variables, tanto cuantitativas como cualitativas y a partir de ellas, analizar cuán frecuente aparecían sus valores y sus medidas de resumen.

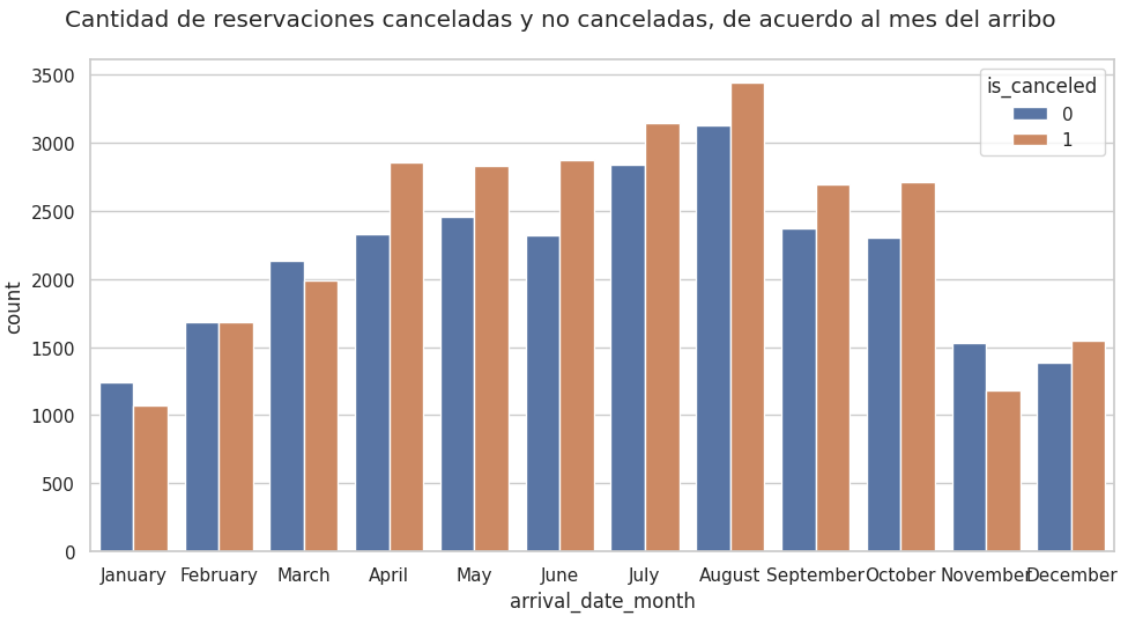
Al tener que realizar un análisis de las variables con relación al **TARGET**, que para nosotros era la variable **is_canceled**, consideramos ciertas variables irrelevantes para nuestro análisis, como market_segment y distribution_channel, ya que, para el análisis de una reserva cancelada, veíamos sin importancia analizar esos valores. Por lo tanto, eliminamos esas variables para poder seguir avanzando con el análisis exploratorio.

Pudimos realizar gráficos para analizar la distribución de las variables y poder visualizar la cantidad de valores diferentes que tomaban las mismas.



Y finalizando la exploración inicial, pudimos analizar la correlación entre las variables con gráficos y la relación de las mismas con el **TARGET**.

Realizamos algunos gráficos más, que nos servían para visualizar ciertos elementos para nuestro análisis exploratorio, como por ejemplo, la cantidad de reservaciones canceladas y no canceladas por mes



También analizamos los datos faltantes, es decir, dentro de las variables, analizar los valores nulos y realizar la imputación necesaria para cada variable, ya sea eliminar la misma o eliminar solamente sus valores, a partir de un porcentaje de nulos. Por ejemplo en la variable Company que tenía un porcentaje de nulos mayor del 94%, decidimos eliminar la variable y en otras, solamente eliminar ciertos valores nulos.

Y, por último, durante todo el análisis exploratorio, estuvimos analizando que había ciertos valores atípicos de acuerdo a ciertas variables. Por lo tanto, realizamos un análisis univariado, para determinar esos valores atípicos de acuerdo a ciertos criterios que determinamos y si se salían de esos criterios, eliminarlos.

En conclusión, podemos ver que entre abril y octubre, hubo muchas más reservaciones canceladas que comparado con el resto de los meses y que en 2017 fue el año con mayor cancelaciones comparado con los otros años. También notamos que company era una variable con muchos valores nulos, por ahí por falta de carga de datos, entonces decidimos eliminar esa variable, que no sumaba para nuestro análisis. Y por último, encontramos casos donde la relación entre adultos, children y babies, era desparejo, como fue el caso de encontrar una reservación con 5 children pero 0 adultos. En relación al target, podemos ver que casi todas las reservaciones con más de 500 días de anticipación fueron canceladas, así como todos los clientes que hicieron más de 11 reservas previas que cancelaron, van a seguir el patrón de cancelar. Por otro lado, si el cliente ya fue 20 veces al hotel, o pidió cambios en su reserva, probablemente no cancele la reservación.