Organización de datos - TP 1

Grupo 10 - **Checkpoint 1:**

En cuanto a las variables cualitativas, se analizó la cantidad de apariciones de cada categoría de cada variable y se utilizaron gráficos para visualizarlas. También se analizó la relación entre las variables cualitativas y la variable target*: is_canceled*. Se obtuvieron algunos hallazgos interesantes, como el hecho de que Portugal es el país que predomina en las reservas. Por lo tanto, se generó una nueva variable booleana llamada *is_portuguese*. Se encontró que las reservas provenientes de Portugal tienden a cancelar mucho más que las del resto de países.

Además, se creó una variable llamada *deposit_boolean* que define un 1 si se hizo un depósito (total o parcial) y un 0 si no. Se observó que va en contra de lo que se creería mediante la lógica, ya que se podría pensar que, al depositar, el cliente sería menos propenso a cancelar. Lo encontrado es que, dentro de las reservas con algún tipo de depósito, el 99% canceló. Son más de 10000 casos así que es muy relevante

También se creó una variable nueva llamada *company_boolean* y se concluyó que no hay casi diferencia entre las empresas con company en cuanto a si cancelan o no. Por último, se creó la variable *had_room_type_change* que define si hubo un cambio de habitación o no. Encontramos que más del 90% de las reservas donde hubo cambios de habitación no cancelaron.

Analizamos los valores nulos. En la variable *country* se decidió eliminar esas filas porque eran muy pocas y no se consideró relevante. Quizás más adelante cambiemos esto, pero queremos probar. En la variable *agent* se agregó una categoría nueva: "Sin agente", representada con el id 0, ya que representaba el 13% aproximadamente y nos pareció importante. Encontramos que aproximadamente el 65% de las reservas que no tienen asignado agent no fueron canceladas. En la variable *company* había muchos datos nulos (cerca de un 94%) y se decidió crear una nueva categoría “Sin Company” con id=0. Luego analizamos los valores “Undefined”. En la variable *meal* se reemplazaron los valores de SC por Undefined, ya que representan lo mismo (ningún plan de comidas asignado).

En cuanto a las variables cuantitativas, se analizó la cantidad de apariciones de cada variable y sus frecuencias, y se utilizaron gráficos para visualizarlas. Hicimos un análisis de la correlación entre la variable *is_canceled* y el resto.

Se encontró un valor negativo en la variable *adr* y se cambió por la media ya que no tenía sentido. Al analizar *lead_time* se pudo observar cómo, a medida que va creciendo el lead time, va creciendo el porcentaje de reservas canceladas.

Se creó una variable nueva llamada *booking_changes_boolean,* donde se pudo ver que la gran mayoría de los clientes que hicieron por lo menos un cambio en la reserva no cancelaron. Específicamente, solo el 24% canceló. Se creó otra variable nueva *special_request_boolean* donde se vio que el 68% de los clientes que hicieron al menos una special request no cancelaron. Por último, se creó otra variable llamada

special_request_and_changes_made que junta las dos anteriores, y se concluyó que, si el cliente hizo algún cambio en la reserva y además hizo una special request, tiene una probabilidad de 77% de no cancelar. Dos variables que nos parecieron interesante son stays_in_weekend_nights y stays_in_week_nights, estas variables presentan una relación bastante lineal por lo cual, en caso de querer disminuir el tamaño del dataset, puede servir. La variable *previous_cancelation* muestra que en los casos donde sí hubo, la probabilidad de cancelación es muy grande.

En cuanto a los outliers, al graficar outliers univariados (especialmente con boxplots) se encontró un par, pero no se decidió eliminarlos/modificarlos ya que tienen sentido y por ahora no molestan. Al analizar outliers multivariados ocurrió algo similar. Encontramos valores que se destacan por distintas razones, pero decidimos dejarlos por ahora. Lo interesante es que quedan marcados para luego en caso de ser necesaria una transformación ya tenerlos analizados.

En conclusión, tras analizar las variables, creemos que las que tienen mayor impacto en las cancelaciones son: is_portuguese, *deposit_boolean, had_room_type_change, agent, lead_time* y *special_request_and_changes_made.*