Checkpoint 1 - Análisis Exploratorio y Preprocesamiento de Datos

Exploración inicial

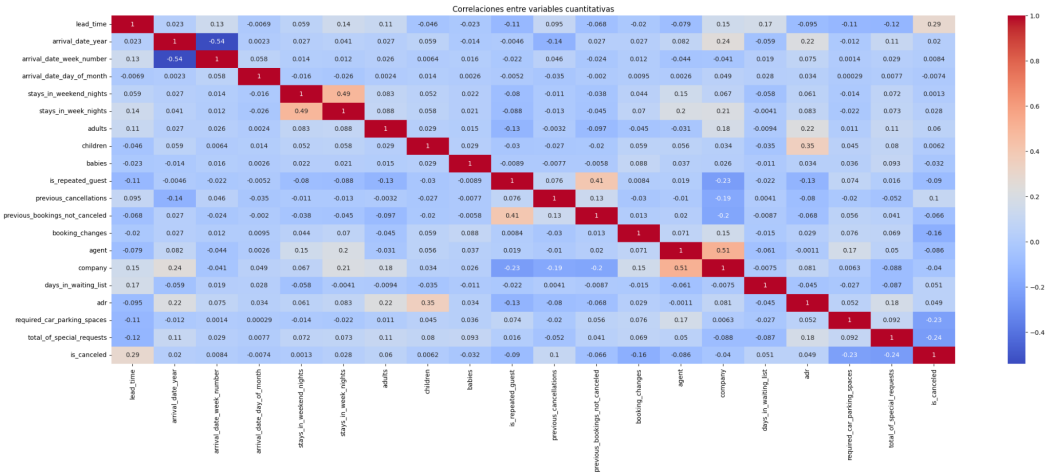
Empezamos analizando las variables del dataset *df_hotels_train*, determinado el tipo de variable, y llegando a la siguiente conclusión:

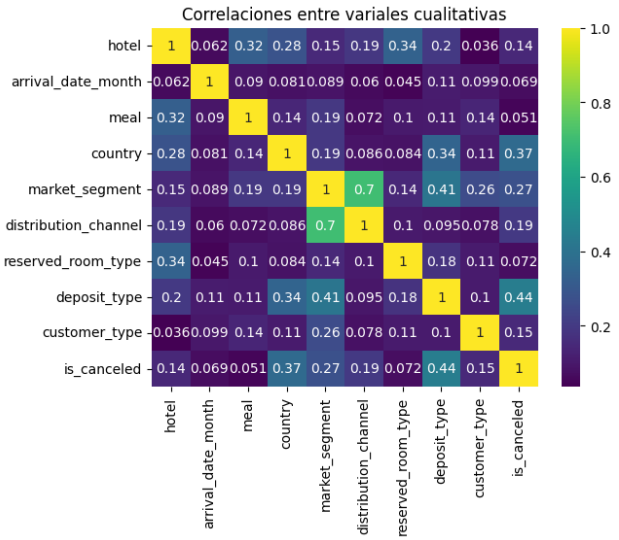
- Variables cuantitativas: *lead_time, arrival_date_year, arrival_date_week_number, arrival_date_day_of_month, stays_in_weekend_nights, stays_in_weeks_nights, adults, children, babies, previous_cancellations, previous_bookings_not_canceled, booking_changes, days_in_waiting_list, adr, required_car_parking_spaces, total_of_special_requests, reservation_status_date, id*
- Variables cualitativas: *hotel, arrival_date_month, meal, country, market_segment, distribution_channel, is_repeated_guest, reserved_room_type, assigned_room_type, deposit_type, agent, company, customer_type, is_canceled*

Luego analizamos los valores posibles y la frecuencia de las variables cualitativas y las medidas de resumen y la distribución de las cuantitativas. También encontramos las siguientes variables que consideramos irrelevantes para el análisis, y por lo tanto las eliminamos del dataset de training y test: *agent, company, reservation_status_date*.

Correlaciones

Luego de haber realizado la exploración inicial graficamos la correlación entre las variables, para ver cuáles están más relacionadas a la variable a predecir *is_canceled.* A continuación se pueden ver los resultados.





En base a estos gráficos se decidieron las variables a analizar más profundamente y se realizaron visualizaciones adicionales sobre los datos.

Datos faltantes

Para determinar qué hacer con los datos faltantes a nivel de columna, empezamos graficando para cada variable el porcentaje de datos faltantes con respecto al total del dataset. Haciendo esto pudimos observar que los datos *children, country, agent* y *company* tienen valores nulos. En el caso de *children* decidimos reemplazar los valores nulos con ceros, ya que este tipo de dato afectará los futuros análisis y solo cuenta con cuatro casos. Con respecto a *country, agent,* y *company*, hay 216 reservas que no tienen un país registrado, y en base a las correlaciones y nuestro criterio decidimos borrar la columna, ya que consideramos que no aporta información sobre la cancelación de la reserva.

Por otro lado, encontramos algunas variables con valores absurdos. El primero de estos es *meals*, que en ciertas filas está registrado como *Undefined*. Según el paper, *SC* y *Undefined* representan lo mismo, por lo tanto decidimos reemplazar los *Undefined* por *SC*. Esto puede ser beneficioso para el encoding a la hora de entrenar los modelos ya que sería una variable menos.

La otra variable que se modificó fue *adr*, ya que también tenía valores absurdos. Hay una sola reserva que tiene un *adr* negativo, por lo tanto es considerado como un error de registro, por lo tanto se procedió a borrarlo. Los otros 873 casos eran valores nulos, en comparación a la cantidad de reservas es una proporción muy chica, pero puede ser un factor que afecte a la hora de la predicción del modelo. Por lo tanto se decidió imputar los datos mediante una interpolación del *adr* basándose en el hotel y la semana de la reserva.

Valores atípicos

Por último, se analizaron los valores atípicos, graficandolos para poder visualizarlos y se decidió como tratarlos. En cuanto al análisis univariado, se utilizaron boxplots para identificar los valores atípicos en cada variable por separado y el z-score para analizarlos. En el caso de *lead_time* y *previous_cancellations* se decidió eliminar los valores atípicos ya que determinamos que son situaciones anómalas, mientras que las otras variables analizadas, *adr, children, babies, stays_in_weekend_nights, stays_in_week_nights, market_segment* y *distribution_channel*, se dejaron ya que consideramos que los valores atípicos aportan información relevante. En el caso del análisis multivariado, se hizo uso del algoritmo de Mahalanobis, observamos las distancias asignadas a cada entrada y en base a eso seteamos un umbral para considerarlos como outliers, evaluamos los casos de: *lead_time* con *days_in_waiting_list* y *adr* con *children*.