Checkpoint 1 - Grupo 23

Análisis Exploratorio

El dataset original está conformado por 61913 registros de reservas y 31 columnas. Además, está estructurado por prácticamente una mitad de variables cuantitativas y otra mitad cualitativas.

Features Destacados

- Las reservas de PRT son mayoría y suelen ser más canceladas que las de otros países.
- Las reservas predominan en las semanas de verano en el hemisferio norte.
- El city hotel suele ser más cancelado que el resort hotel.
- Los agentes están fuertemente relacionados con el hotel y país de origen de reserva.
- El *lead_time* tiene cierta correlación con nuestra variable target. Este significa la cantidad de tiempo entre el día que se hizo la reserva y el arribo al hotel.

Supuestos

- Los hoteles son de Portugal ya que todas las instituciones del estudio eran de allí.
- Cuando el *distribution_channel* era “Direct”, es correcto que *agent_id* sea nulo. De caso contrario, si o si tiene que tener un *agent_id*.
- Undefined es solo un valor posible en *meal_type* tal como decía el paper.

Preprocesamiento de Datos

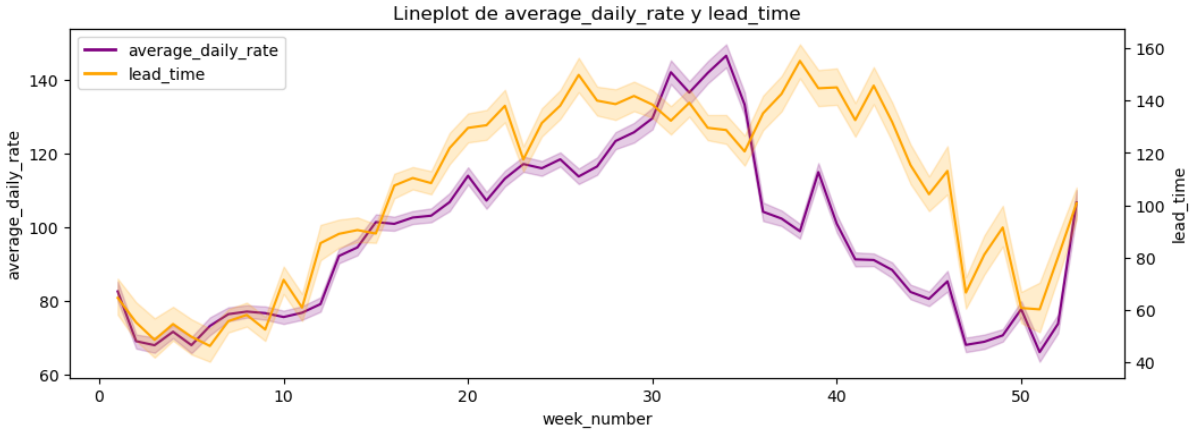
Detallar como mínimo los siguientes puntos:

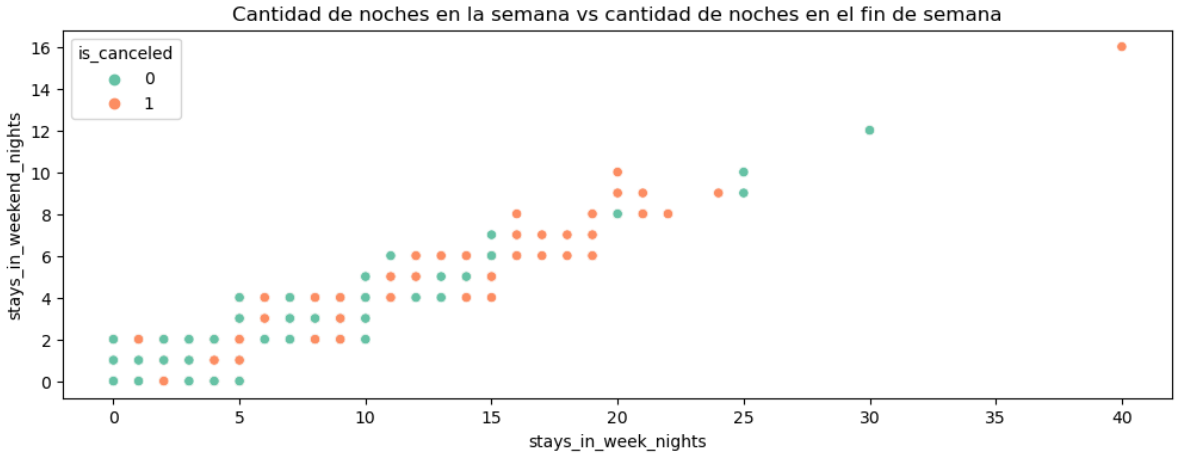
1. Columnas eliminadas:

Decidimos eliminar la columna *Company* ya que en su mayoría eran datos nulos y no creemos que vaya a aportar información. Analizamos otras variables irrelevantes (year, day_of_month, days_in_waiting_list) pero no las eliminamos ya que a priori no sabemos si realmente son irrelevantes o no.

2. Correlaciones detectadas

La primera correlación es entre el *lead_time* y el *average_daily_rate* a medida que pasan las semanas del año. Otra correlación es entre la cantidad de personas en la reserva y el *average_daily_rate*. Además, observamos ciertos patrones entre el *agent_id*, el *country* y el *hotel*. Otra correlación un tanto obvia es entre *stays_in_week_nights* y *stays_in_weekend_nights*. Por último, otra correlación entre tantas a nombrar es la de el *average_daily_rate* y el *assigned_room_type*.





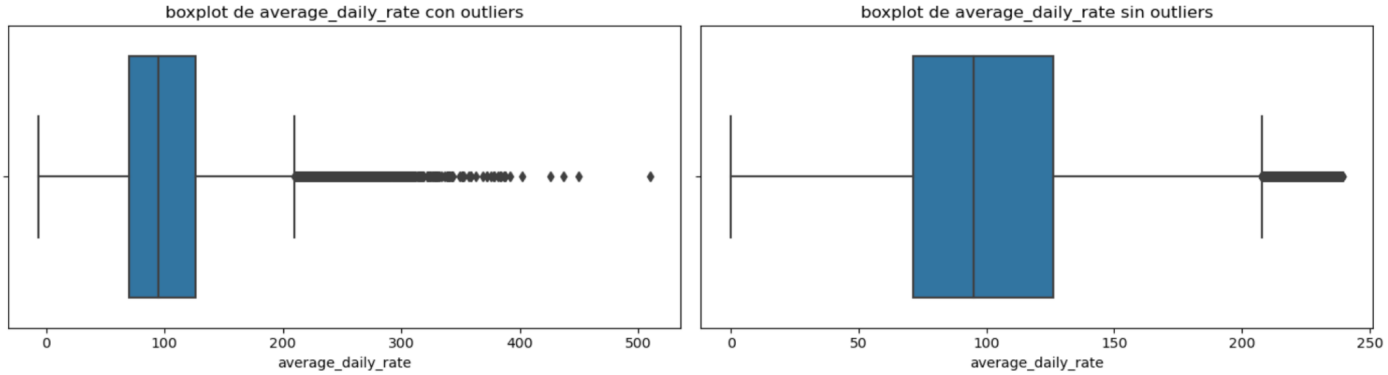
3. Columnas recodificadas:

Creamos la columna people ya que nos parecía útil tener la cantidad total de personas en la reserva. También creamos la columna *total_nights* por la misma razón, tiene sentido analizar la cantidad de días totales que se quedan los huéspedes. Más adelante veremos si estas adiciones son de utilidad o no. La última modificación que realizamos fue en agent_id, en los casos donde consideramos que tenía que ser un valor nulo, reemplazamos el NaN con '-1'.

4. Valores atípicos:

Encontramos muchas columnas con valores atípicos, las siguientes son: lead time, stays_in_week_nights, stays_in_weekend_nights, adults, children, babies, previous cancellation, previous bookings not canceled, booking changes, days in waiting list, average daily rate, total special requests. En su gran mayoría utilizamos z-score modificado o su hermano, z-score, no nombramos variable por variable ya que son una gran cantidad. Para todas las variables hicimos análisis univariados salvo para *total_of_special_requests y stays_in_week_nights*, que utilizamos la distancia de

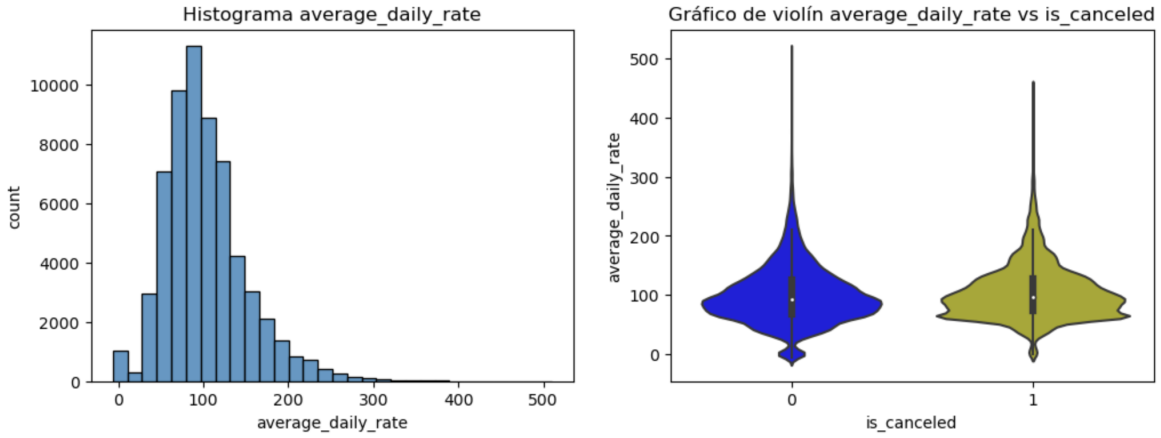
Mahalanobis, comparándolos con *is_canceled* y *stays_in_weekend_nights* respectivamente. Otro caso para destacar es el del *average_daily_rate,* donde aplicamos el z-score modificado, y luego observamos que habían valores menores a 0, lo cual no es posible así que también eliminamos esos valores. Finalmente en *distribution_channel* y *market_segment* contenían valores 'Undefined' lo cual el paper no nombra, entonces los eliminamos.



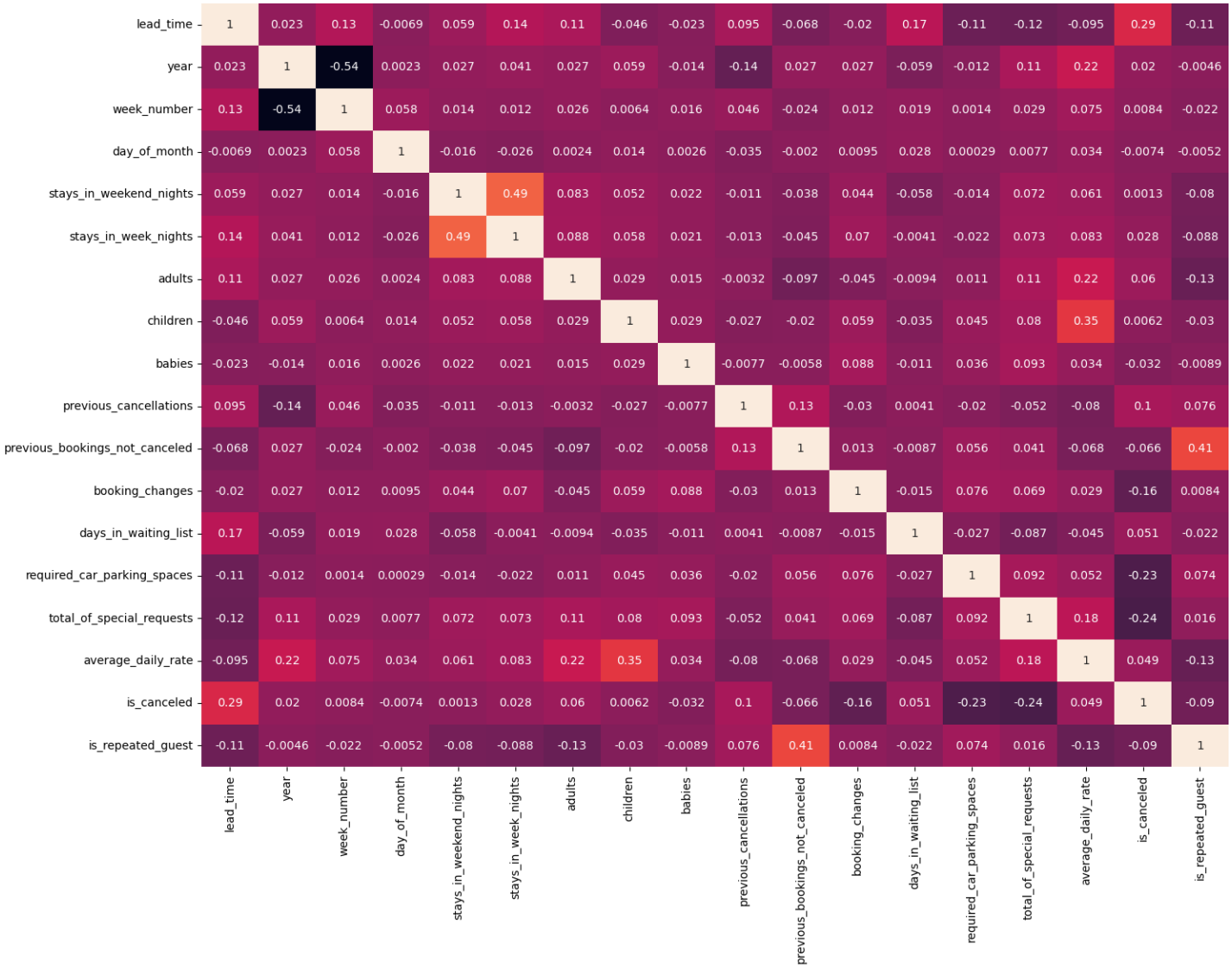
5. Valores faltantes:

Los primeros datos faltantes son en agent_id, en los casos donde consideramos que tenia que ser un valor no nulo, realizamos una especie de árbol de decisión manual, donde vimos que tenía una correlación con los hoteles y con el país, por lo tanto tomamos esos dos parámetros para inferir el agent_id más probable de cada reserva. El 12% originalmente eran datos faltantes. Algo similar realizamos con country, con correlaciones con el deposit_type y market_segment. La otra variable con datos faltantes fue la nombrada anteriormente, *Company*, con 94,9% de datos faltantes, la columna fue eliminada. En *children* había un 0,01% de datos faltantes, los cuales fueron interpretados como '0'.

Visualizaciones



El primer gráfico muestra cómo se distribuye nuestra variable en el dataset. Por otro lado, el segundo gráfico demuestra la relación entre la variable *average_daily_rate* y la *variable target*.



Este heatmap muestra el coeficiente de correlación de Pearson entre todas las variables numéricas. Cabe aclarar que esto es previo a la imputación de datos y eliminación de outliers.

Tareas Realizadas

Dado que somos solo dos integrantes, decidimos ir tarea a tarea, tomando decisiones en conjunto y dividiéndonos la carga de cada parte a la mitad.

Integrante	Tarea
Mateo Vroonland	Todo
Juan Pablo Carosi Warburg	Todo