Checkpoint 1 - Grupo 32

Análisis Exploratorio

Durante el análisis exploratorio encontramos que en un principio el dataset cuenta con casi 62 mil registros y 31 columnas. La mayoría de datos son numéricos (ya sea solo 0's y 1's ó ID's únicos para identificar usuarios/empresas, por ejemplo).

Preprocesamiento de Datos

1. Columnas eliminadas

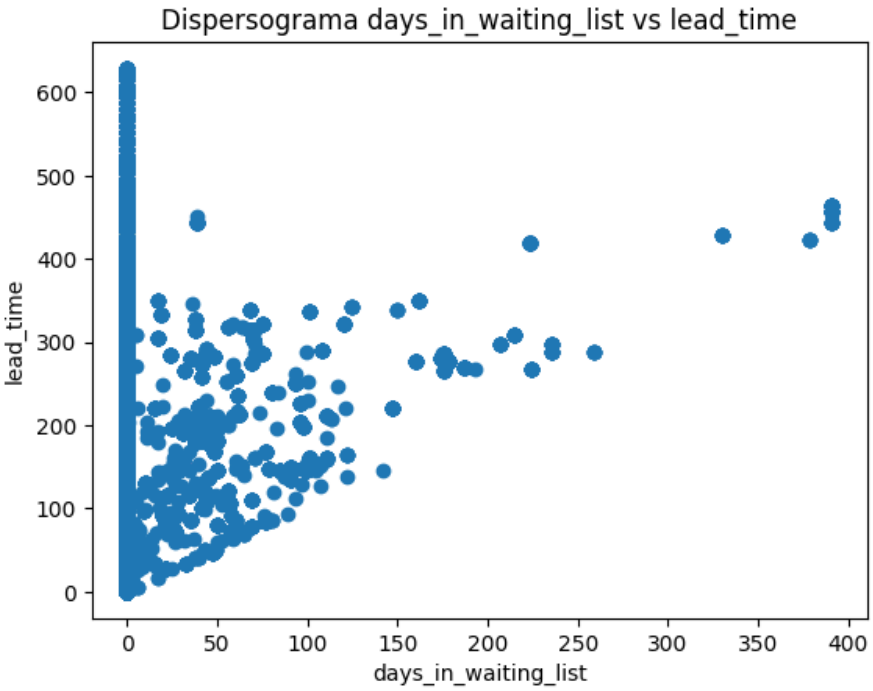
Si bien encontramos ciertas columnas cómo por ejemplo "meal", "car_parking_spaces" o "customer_type" que parecen no afectar directamente al target, decidimos no eliminarlas por el momento, sin antes haber probado el modelo teniendo en cuenta estas variables.

2. Columnas agregadas

Agregamos una nueva columna "room_changed", la cual es un booleano que nos dice si efectivamente hubo un cambio de habitación, basándonos en la comparación entre la habitación reservada, y la habitación entregada al cliente.

3. Correlaciones detectadas

Analizando el Heat-map encontramos que *is_repetead_guest* y *previous_booking_not_canceled* tienen un valor de 0.41 el cual nos dice que existe una cierta relación positiva.
También realizamos un primer análisis a través de un dispersograma entre las variables days_in_waiting_list y lead_time para ver la existencia de una posible correlación



Encontramos que las variables parecieran tener una correlación lineal y calculando la correlación de pearson, filtrando por los días en los que se espero un día o más para no afectar el resultado, nos da un valor de 0.584 el cual nos confirma que existe una relación.

4. Columnas recodificadas

La columna “company” fue recodificada, asignando un booleano según si tiene o no una company. La columna “arrival_date_month” pasó de ser un object a un número que representa dicho mes. La columna “is_repeated_guest” le cambiamos el valor 0 o 1 por un booleano True o False.

5. Valores atípicos

Encontramos que las columnas *stays_in_weekend_nights, stays_in_week_nights, adults, children, babies, previous_cancellations, days_in_waiting_list y requird_car_parking_spaces* contenian valores atípicos.

Por ejemplo, en el análisis de la columna previous_cancellations, se encontraron 125 casos con valores mayores a 10. De estos, 116 casos volvieron a cancelar, lo que sugiere

una posible relación y por eso no se eliminaron. Además, se identificaron 192 casos en los que no había ningún adulto registrado, y estos se eliminaron ya que se consideró que una reserva debe tener al menos un adulto

6. Valores faltantes

Encontramos que la columna “company” poseía ≈ 95% de valores faltantes. La columna “agent” poseía ≈ 13% de valores faltantes. Luego en medida < 1% las columnas “children” y “country” poseían valores faltantes. Otras columnas cómo “meal”, “market_segment” y “distribution_channel” poseian < 1% de valores inválidos. A la columna “company” se le asignó un valor booleano (0 o 1) el cual nos dice si tiene o no una company asignada. A la columna “agent” se le asignó el id de la fila a dichos valores nulos. Todos los valores nulos o inválidos con porcentajes pequeños fueron eliminados.

Visualizaciones

Heat-map de porcentajes de cancelación, según el día del año.

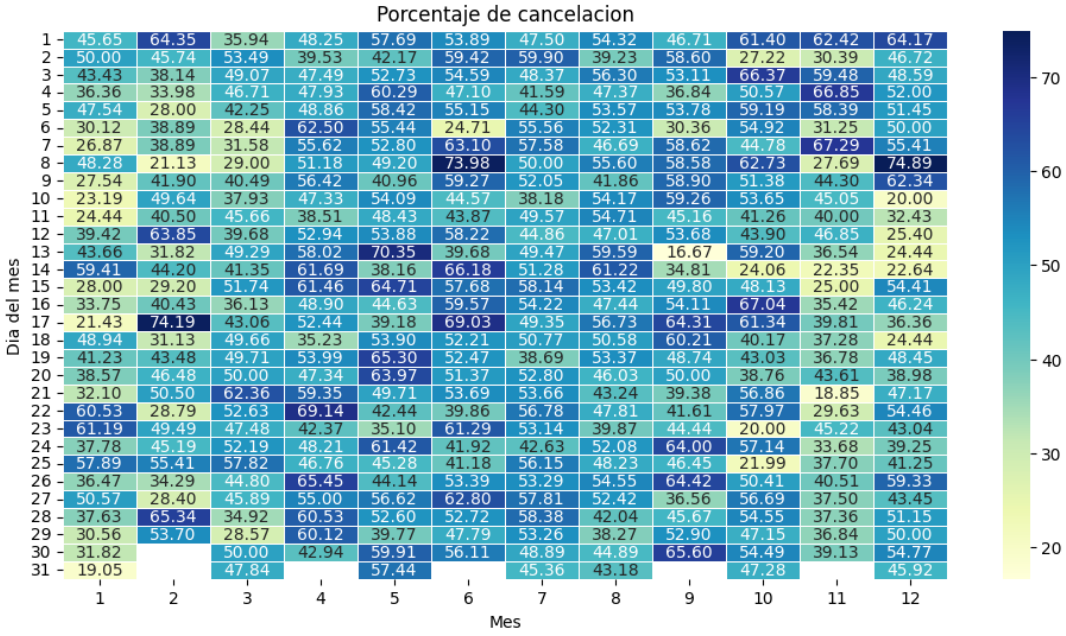
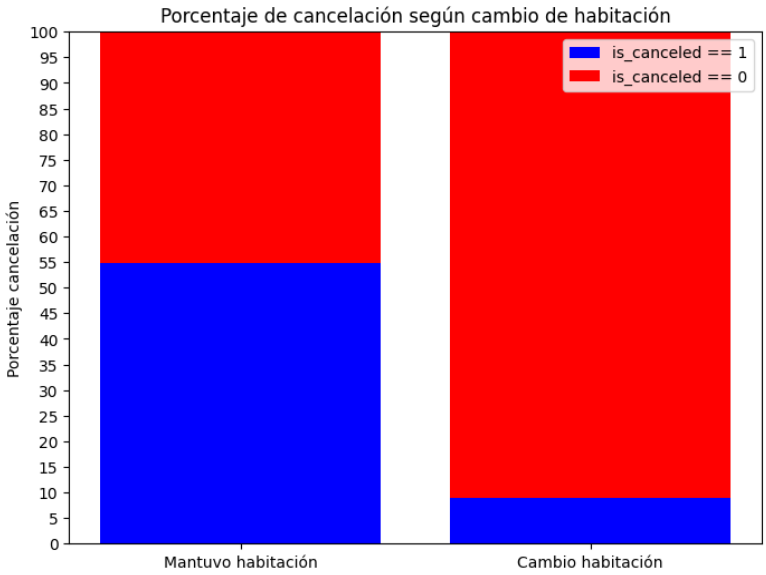


Gráfico de barras, utilizando nuestra nueva columna de datos, “room_changed”, la cual nos pareció interesante, debido a la diferencia que se aprecia si el hotel permite un cambio de habitación.



Tareas Realizadas

Integrante	Tarea
Daniel Agustin Marianetti	Imputación de Datos Análisis de Datos Armado de Reporte
Franco Ezequiel Rodriguez	Exploracion Inicial Visualización de datos Armado de reporte
Ezequiel Lazarte	Visualización de datos Detección de Outliers Armado de reporte