# Checkpoint 1 - Grupo 39

## **Análisis Exploratorio**

Realizar una breve descripción del dataset: cantidad de registros y columnas, etc.

Comentar los features más destacables: tipo de dato, qué representa y por qué se destacan. Listar hipótesis o supuestos que tomaron (por ejemplo si difieren de las definiciones que hay en el paper.

Informacion del dataset:

* Cantidad de registros: 61913
* Cantidad de columnas: 31
* Cantidad de variables numericas: 19
* Cantidad de variables Object (string): 11

Features destacadas:

* Adults: representa la cantidad de adultos de la reserva, es de tipo numerico y fué muy importante para varios de los analisis de valores atipicos, ya que se relaciona de una manera logica con las demas variables
* Babies y Children: representan los bebes y chicos de la reserva, son de tipo numerico y en ambos casos pasa algo parecido que con adults se relaciona de manera logica con las demas variables. Gracias a ellas pudimos deducir cosas sobre el conjunto de datos
* Lead\_time: representa el tiempo entre la reserva y el chek-in, es de tipo numerica y nos parece importante ya que el tiempo de espera de la llegada tiene mucho que ver con la cancelacion de la reserva, ya que en el medio pueden pasar varias cosas. Ademas la usamos para encontrar valores atipicos multivariados
* Is\_repeated\_guest: representa si la reserva es de un persona que ya ha hecho una reserva, es de tipo categorica y vimos que hay relacion entre esta y algunas de las variables (por ejemplo: previous\_booking\_not\_canceled)

Hipotesis tomadas:

* Los valores faltantes en la columna agent fueron tomados como reservas que no fueron hechas con un agente, sino por cuenta propia, a estos los imputamos asignandoles un valor de id -1 para poder usar los registros mas adelante

## **Preprocesamiento de Datos**

Detallar como mínimo los siguientes puntos:

## Columnas eliminadas:

Company: Se eliminó debido a la gran cantidad de valores faltantes y ademas proporcionaba un id el cual no era de utilidad para nuestro analisis

## Correlaciones detectadas:

Nombres de las columnas correlacionadas.

Gráfico de dispersión. Coeficiente de correlación de Pearson.

* Stay\_in\_weekend\_nigths(eje y) y stays\_in\_week\_nightsnigths(eje\_x):

Valor de correlacion de Pearson: 0.45

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

* previous\_booking\_not\_canceled (eje x) y is\_repeated\_guest(eje y):

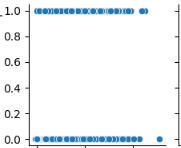
Valor de correlacion de Pearson: 0.59

Gráfico

Descripción generada automáticamente

* is\_canceled(eje y) y lead\_time(eje x):

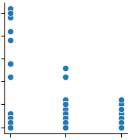
Valor de correlacion de Pearson: 0.28



Tienen una correlacion mediamente baja, pero nos interesa porque lead\_time es la variable que tiene mas relacion con is\_canceled, que es nuestra variable a predecir. Nos interesa estudiarla mas adelante

* previous\_cancellations(eje y) y arrival\_date\_year(eje x):

Valor de correlacion de Pearson: -0.32



## Columnas recodificadas:

Nombre de la columna

En qué consistió la recodificación y por qué se hizo.

## Valores atípicos:

Columnas encontradas.

Valor o rango de valores considerados atípicos

¿Son multivariados o univariados?

Diagramas de cajas.

Qué técnicas utilizaron y qué decisiones tomaron

## Valores faltantes:

¿Qué columnas tenían datos faltantes?

¿En qué proporción? ¿Qué se hizo con estos registros?

**Visualizaciones**

Se piden dos gráficos realizados: dispersión entre variables, histogramas, heatmaps, etc. que sean descriptivos del problema. Seleccionar aquellos que permitan entender cómo se distribuyen los datos, cómo se relacionan la variables etc.

Comentar brevemente qué se está visualizando en cada caso.

## **Tareas Realizadas**

Indicar brevemente en qué tarea trabajo cada integrante del equipo, si trabajaron en las mismas tareas lo detallan en cada caso (como en el ejemplo el armado de reporte).

|  |  |
| --- | --- |
| Integrante | Tarea |
| Azul Villanueva | Detección de Outliers Armado de Reporte |
| Mateo Suster | Análisis de Correlaciones Armado de Reporte |
| Ignacio Velasco | Análisis de Valores Faltantes Imputación de Datos |