Trabajo Práctico 1 : Reservas de Hotel

Grupo 11: Datatack

Exploración inicial:

Se trabajó realizando un análisis exploratorio del archivo **hotels_train.csv** reservas de un hotel, se clasificó las variables de la columna según sean cualitativas o cuantitativas usando los conceptos vistos en clase, los siguientes pasos fueron hacer un pequeña limpieza de datos; se vio que la variable "company" tiene 58761 de valores "NaN" versus el total de datos de 61913, se eliminó esta columna del dataset. Como siguiente paso se realizó gráficos tipo barra de la frecuencia de las variables cualitativas, en caso de la variable "country", no se pudo observar un gráfico entendible por lo tanto se pasó a un gráfico de tabla. En los siguientes pasos de cálculo la "moda", "mediana" y la “media” de las variables cuantitativas, falto calcular la correlación de variables y analizar la relación entre la variables con el target.

Gráficos

Se buscó analizar la información dada, construyendo gráficos que ayudarán con la comprensión visual.

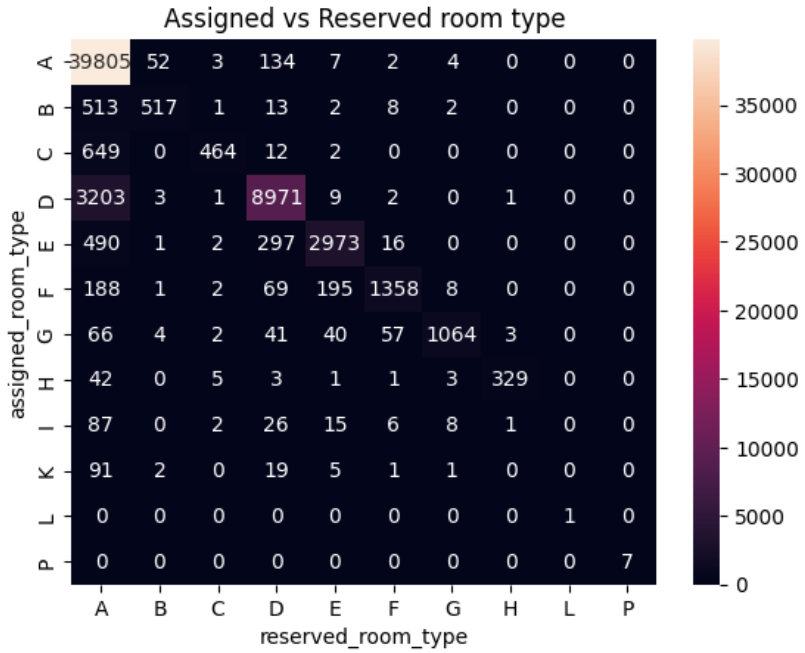
Al principio se jugó con la idea de realizar gráficos para todo y ver cómo se ajustaba la información a estos y elegir cual utilizar a partir de allí. Pero esto se descartó rápidamente debido a la cantidad que se podía llegar a armar. Consecuentemente primero se trabajó con datos más similares entre ellos y luego se expandió el alcance de los gráficos.

Por su parte, previo a realizar los gráficos se cambiaron los labels de dos de las columnas del set dado ya que tanto para cancelación como para sí había concurrido previamente al hotel se creyó que si o no eran más explicativo (por lo menos en gráficos) que 1 o 0.

Además también es importante notar, que muchos de los tipos de dato del dataset eran "objetos" y por lo tanto el módulo seaborn trabaja medio raro con ellos, entonces donde se pudo se utilizó seaborn.objects para evitar este problema.

Con respecto a los gráficos en sí, el primero realizado fue el de número de huéspedes por hotel y número de cancelaciones, ya que se cree que para cualquier hotel, lo más importante es tener clientes para poder tener ganancias. Luego se siguió con este hilo y se desarrollaron un número de gráficos que relacionaba el precio por noche con distintos datos presentes. A partir de allí, se analizó cómo afectaba distintos puntos de publicidad a la venta, agencias y compañías de turismo, entre otros.

Los gráficos que consideramos los más importantes son:



Donde el primero habla de la cantidad de clientes que tiene cada hotel y el segundo de su organización y tipo de clientes, ya que con un poco más de información sobre las habitaciones estas dicen mucho sobre quien se hospeda allí.

Análisis de datos faltantes:
Se buscó en el dataset los valores nulos o indefinidos.
Había valores nulos en las categorías "children", "country", "agent" y "company".

- Children: se decidió borrar los registros por ser una cantidad muy reducida.
- Country: se decidió asignarles el valor ‘NOC’ para representar el país “No informado”
- Company: Dado que falta el 95% de los datos, se decide eliminar la columna.
- Agent: Se le asignó el valor numérico 0 a los datos no informados.

Datos indefinidos
Se encontraron en las categorías "meal", "market_segment" y "distribution_channel".

- Meal: Se halló que tiene valores 'SC' u 'Undefined' que representan lo mismo (una reserva sin paquete de comida). Se decidió unificarlas en ‘SC’
- Distribution_channel: se decidió borrar los registros por ser una cantidad muy reducida.
- Market_segment: Los valores indefinidos de esta variable estaban incluidos en los de 'distribution_channel'. El análisis es análogo y se decide que permanezcan fuera del dataset.

Análisis de variables numéricas:
Se analizaron los valores numéricos en el dataset, buscando valores negativos.
Se halló uno en ‘adr’ y se lo imputó utilizando la media de valores similares.

Análisis de outliers

Se realizó un análisis univariado y multivariado, donde se hallaron outliers en las siguientes categorías:

- Children
- ADR
- Babies:
- Required_car_parking_spaces
- Lead time
- Total_of_special_requests

En general, la estrategia para resolver los outliers fue calcular la media de utilizando registros similares.

En el caso de children, también se descubrió un valor con 10 niños, que al comparar con los valores similares se concluyó que se trata de un error de tipeo y es 1 en realidad.