Organización de datos, Curso Rodriguez
1er Cuatrimestre 2023

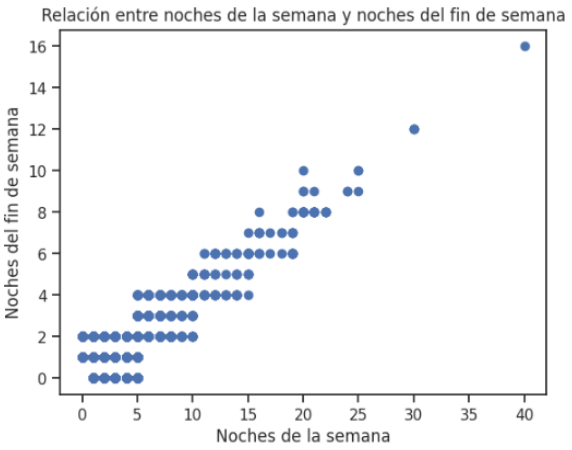
Docente: Rodriguez, Juan Manuel.
Estudiantes:

- Mendoza Hernandez, Sabrina Scarlet (108524)
- Valeriani, Matias Gabriel (108570)
- Jang, Lucas (109151)

Checkpoint 1: Análisis exploratorio e Ingeniería de features

En esta primera entrega, se realizaron tareas como: cargar datos del data sets en un dataframe llamado “hotels” y analizar cada una de las variables que forman parte del mismo. Se identificaron los tipos de las variables, se calcularon la cantidad de valores únicos por variable cualitativa y para las variables de tipo cuantitativas se calculó la mediana, moda, máximo, mínimo, etc. Luego de identificar las variables seguimos con la correlación existente que podría tener cada una, para ello nos resultó conveniente usar un heatmap que nos ayudará visualmente a identificar las variables con correlaciones positivas y negativas. Para la sección, variables irrelevantes, las analizamos por medio de visualizaciones relacionadas con el target, sacamos las variables con comportamientos como “babies” que tenía un alto porcentaje de 0’s y no parecía útil para el análisis, así mismo con las otras columnas borradas. Para la relación de variables vs target, discutimos cuáles son las principales columnas con mayor relación y las graficamos. Para visualización realizamos un pairplot que nos muestra todas las variables y obtener así la primera aproximación del trabajo. Datos faltantes fue identificar los datos nulos por variables y en base a el porcentaje discutimos el método de imputación. Por último, Valores atípicos, Analizamos cada boxplot de las variables cuantitativas para observar los posibles outliers y combinamos las variables para observar mejor la distribución de las mismas

1er Hallazgo:
Como las variables stays_in_week_night y stays_in_weeked_night tienen una alta correlación y representan las noches hospedadas, decidimos sumar ambas variables para tener una representación de todas las noches en total.



2do Hallazgo:
Realizamos un heatmap entre arrival_date_week_number y arrival_date_year, observamos la falta de datos hasta a mediados del 2015 y la misma situación para finales del 2017, suponemos, nos ayudará en futuras aproximaciones