Documentación ETL

Introducción

En este proyecto el proceso de Extracción, Transformación y Carga se llevó a cabo para la gestión y análisis de los datos ya que nos permitió recopilar datos de diversas fuentes y cargarlos en GCP.

Objetivo

Dar tratamiento a los datasets obtenidos desde la web y de esta forma evitar outliers o datos erróneos que pudieran afectar nuestro modelo de Machine Learning y el análisis de estos datos.

Metodología y Tecnologías usadas

Se extrajo la data por medio de un webscrapping y se almacenó en un bucket en gcp el cual toma los datasets en crudo, posterior a esto se llevó a cabo la transformación de estos datos el cual se llevó a cabo en cloud functions pasándole un script el cual pedía que tomara los datos del bucket con los datos en crudo, antes de comenzar las transformaciones el script verifica si ese datasets ya fue procesado antes, una vez que los tomaba agregaba una columna para la identificación del color, luego verificaba nulos y duplicados, los últimos se eliminaban y los nulos se filtraban segun el caso, para la columna de congestion_surcharge se tomaban los valores de la mediana para reemplazar los nulos y en otros casos se tomaban como 0.

También se hizo una función para verificar las fechas y que éstas tuvieran un formato según el año de los datos proporcionados; aquí se extrajo la columna de las fechas y se obtuvieron los años, meses y días, y la hora de inicio y fin de los viajes, en columnas distintivas.

Lo siguiente que procede a hacer es realizar un id de los viajes para poder tener una Primary key al momento de consumir los datos en PBI.

Se eliminan las columnas innecesarias y se modifican los nombres de las columnas de forma que sean entendibles.