Data Intake Report

Name: G2M Case Study
Report date: 09 Dec 2022
Internship Batch: LISUM16
Version: 1.0
Data intake by: Adrian Baysa
Data intake reviewer: N/A
Data storage location: https://github.com/agbaysa/dataglacier_week2/blob/main/df_120922.7z

Tabular data details:

Total number of observations	359,392
Total number of files	6 csv files consolidated into 1 main dataframe (df_120922.csv)
Total number of features	31 columns (with 17 addl features)
Base format of the file	csv
Size of the data	87.9MB (unzipped)

Proposed Approach:

- Import using pandas
- Data Cleanup
    - Convert integers to date
    - Join dataframes including data for US holidays and US weather
    - Drop and rename columns
    - Get Month and Weekday columns from Date of Travel
    - Include Additional Features
        - Income: Price Charged – Cost of Trip
        - Price Per KM: Price Charged / KM Travelled
        - Cost per KM: Cost of Trip / KM Travelled
        - Income per KM: Income / KM Travelled
        - Profitable: 1 if profitable
        - Profit_perc: Income / Price Charged
        - Day of Weel
        - Month
        - Year
        - Age Categories
        - Income_salary Categories
        - Market Penetration
        - City Size Categories
    - Convert Population and User columns to numeric
    - Remove duplicate rows using drop_duplicates()
    - Validate data to ensure rows match the number of rows for Cab_Data.csv
    - Use pandas_profiling to profile the data
- Analyze
- Make Recommendations