Data Intake Report

Name: G2M Case Study
Report date: 05 April 2023
Internship Batch: LISUM16
Version: 2.0
Data intake by: Adrian Baysa
Data intake reviewer: N/A
Data storage location: https://github.com/agbaysa/dataglacier_week2

Tabular data details (for each dataset):

Cab

Total number of observations	359,392
Total number of files	1 file
Total number of features	8 columns
Base format of the file	csv
Size of the data	`21219244 bytes`

City

Total number of observations	20
Total number of files	1
Total number of features	4 columns
Base format of the file	csv
Size of the data	`759 bytes`

Customer ID

Total number of observations	49171
Total number of files	1
Total number of features	5 columns
Base format of the file	csv
Size of the data	`1051215 bytes`

Transaction ID

Total number of observations	440098
Total number of files	1
Total number of features	4
Base format of the file	csv
Size of the data	`8998194 bytes`

Holidays (helper/additional csv)

Total number of observations	57
Total number of files	1
Total number of features	2 columns
Base format of the file	csv
Size of the data	1164 bytes

Weather (helper/additional csv)

Total number of observations	864
Total number of files	1
Total number of features	5 columns
Base format of the file	csv
Size of the data	28223 bytes

Combined File

Total number of observations	359,392
Total number of files	1
Total number of features	32 columns
Base format of the file	csv
Size of the data	21219244 bytes

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 Week
        - 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