Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 06/14/2024
Internship Batch: LISUM34
Version:1.0
Data intake by: Aryaman Srivastava
Data intake reviewer:
Data storage location: https://github.com/AryaSriva/DataSets

Tabular data details:

File Name	finalCabData.csv
Total number of observations	353,434
Total number of files	1
Total number of features	17
Base format of the file	.csv
Size of the data	32,900 KB

File Name	City.csv
Total number of observations	20
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	1 KB

Proposed Approach:

- Remove duplicate customers(based on Customer IDs) and rows with missing data from the customer data before merging the customer data with transaction data by Customer ID.
- Next, remove duplicate transactions(based on Transaction IDs) and rows with missing rows from the transaction and cab data before merging the transaction data with cab data by Transaction ID
- In the new data set, dubbed finalCabData, remove outliers based on the Cost of Trip, Price Charged, and Number of Kilometers Travelled using the IQR rule(values greater than 1.5*IQR + 3rd quartile or less than 1st quartile – 1.5*IQR, where IQR is the 3rd quartile – first quartile, are treated as outliers)
- Add 3 new columns to the finalCabData dataset, Profit, Profit Margin, and Income Category.
- Set the values for Profit by subtracting the Cost of Trip column from the Price Charged column.
- Set the values for Profit Margin by dividing the Profit column by the Price Charged column and multiplying the quotient by 100.

- Set the values for the Income category column by each row into 3 categories based on the value in the Income(USD/Month) column(>10,000/month, 5,000 to 10,000/month, <5,000/month)
- For the city data, remove duplicate cities(based on city name) and rows with missing column values.
- Assumptions
    - **The Price_charged feature is the total revenue made on the transaction (including any tips)**
    - **The Price_charged feature is only affected by the Cost of the Trip feature and the Kilometers Travelled feature**
    - **The Cost of the Trip feature considers all costs associated with the trip**
    - **The date_of_travel feature uses mm/dd/yyyy format**