Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 12 June 2024
Internship Batch: 9572138
Version: 1.0
Data intake by: Armel Moumbe
Data intake reviewer:
Data storage location:

Tabular data details: Cab_Data

Total number of observations	359,392
Total number of files	
Total number of features	7
Base format of the file	.csv
Size of the data	20.663 MB

The **Cab_Data.csv** file contains the following columns:
Transaction ID: Unique identifier for each transaction.
Date of Travel: Date when the travel occurred, represented as an Excel serial date number.
Company: Name of the cab company ("Pink Cab" or "Yellow Cab").
City: City where the travel occurred.
KM Travelled: Distance travelled in kilometers.
Price Charged: Amount charged to the customer.
Cost of Trip: Cost incurred for the trip.

Tabular data details: Customer_ID

Total number of observations	491,372
Total number of files	
Total number of features	4
Base format of the file	.csv
Size of the data	1.027 MB etc>

The **Customer_ID.csv** file contains the following columns:
Customer ID: Unique identifier for each customer.
Gender: Gender of the customer.
Age: Age of the customer.
Income (USD/Month): Monthly income of the customer in USD.

Tabular data details: Transaction_ID

Total number of observations	359,392
Total number of files	
Total number of features	3
Base format of the file	.csv
Size of the data	8.788 MB

The **Transaction_ID.csv** file contains the following columns:

Transaction ID: Unique identifier for each transaction.
Customer ID: Unique identifier for each customer.
Payment Mode: Payment method used for the transaction (Card or Cash).

Tabular data details: City

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

The **City.csv** file contains the following columns:
City: Name of the city.
Population: Population of the city.
Users: Number of cab users in the city.

Tabular data details: Master Data

Total number of observations	359,392
Total number of files	4
Total number of features	14
Base format of the file	.csv
Size of the data	41.942 MB

The **Master Data.csv** file, which is a merged file, contains the following columns:
Transaction ID: Unique identifier for each transaction.
Date of Travel: Date when the travel occurred (converted to a readable date format).
Company: Name of the cab company ("Pink Cab" or "Yellow Cab").
City: City where the travel occurred.
KM Travelled: Distance travelled in kilometers.
Price Charged: Amount charged to the customer.
Cost of Trip: Cost incurred for the trip.
Customer ID: Unique identifier for each customer.
Payment Mode: Payment method used for the transaction (Card or Cash).
Gender: Gender of the customer.
Age: Age of the customer.
Income (USD/Month): Monthly income of the customer in USD.
Population: Population of the city.
Users: Number of cab users in the city.

Proposed Approach:

The approach to validate and deduplicate the data was:
Step 1: Check for Duplicates
For each dataset, we use the duplicated() method to check for duplicate rows.
Step 2: Remove Duplicates
If any duplicates are found, we use the drop_duplicates() method to remove them.
There are no missing values or duplicate records in any of the datasets.

Assumptions

Identifiers:
Transaction ID in the Cab_Data.csv is unique and correctly represents a single transaction.
Customer ID in the Customer_ID.csv is unique and correctly represents a single customer.
Transaction ID in the Transaction_ID.csv is unique and correctly maps to a customer.
City entry in the City.csv represents a unique city.

Date Format: The Date of Travel in Cab_Data.csv was in the wrong format, which needs to be converted to a readable date format.

Data Completeness: All necessary fields are present in each dataset, and no crucial columns are missing. There are no missing values in columns critical for analysis, such as Transaction ID, Customer ID, Price Charged, Cost of Trip, Date of Travel, City, Gender, Age, and Income.

Consistency: The datasets are consistent in terms of the entities they represent. For example, the same Customer ID across different datasets refers to the same individual. The City names in Cab_Data.csv match those in City.csv.

Accuracy: The financial values such as Price Charged and Cost of Trip are accurate and recorded correctly. The demographic information (e.g., Age, Income) for customers is accurate and reasonable.

Data Validity: The data provided spans from 31/01/2016 to 31/12/2018, and it is assumed to be representative of the actual cab usage trends during this period.

Financial Data: Price Charged and Cost of Trip are assumed to be recorded in the same currency (USD).

Demographic Data: The Age and Income columns in Customer_ID.csv accurately reflect the customers' demographics.