Data Intake Report

Name: Go to Market Insight for Company XYZ
Report date: 7/7/2022
Internship Batch: LISUM11
Version: 1.0
Data intake by: Noah Igram
Data intake reviewer:<intern who reviewed the report>
Data storage location: https://github.com/DataGlacier/DataSets.git

Tabular data details:

- 4 .csv files

Cab_Data.csv

Cab_Data.csv contains transaction for 2 cab companies, Yellow and Pink.

Total number of observations	359,392
Total number of features	7
Base format of the file	.csv
Size of the data	21.2 MB

City.csv

Customer_ID.csv is a mapping table that maps a customer's demographic details to a unique ID.

Total number of observations	20
Total number of features	3
Base format of the file	.csv
Size of the data	759 bytes

Customer_ID.csv

Customer_ID.csv is a mapping table that maps a customer's demographic details to a unique ID.

Total number of observations	49,171
Total number of features	4
Base format of the file	.csv
Size of the data	1.1 MB

Transaction_ID.csv

Transaction_ID.csv is a mapping table containing transaction ID details, customer ID details and payment mode details.

Total number of observations	440,098
Total number of features	3
Base format of the file	.csv
Size of the data	9 MB

Features of Data

Features of Cab_Data.csv:

- Transaction ID
- Date of Travel
- Company
- City
- KM travelled
- Price Charged
- Cost of Trip

Features of City.csv:

- City
- Population
- Cab Users

Features of Customer_ID.csv:

- Customer ID
- Gender
- Age
- Income (USD/month)

Features of Transaction_ID.csv:

- Transaction ID
- Customer ID
- Payment Mode

Dedupe validation

While none of the individual datasets themselves contain any missing values, there are some overlapping entries between the 4 datasets. A master dataset was created to navigate past this issue and the details are shown in the python notebook submitted alongside this document.

The Customer ID dataset contains information on customer gender, age, and income, which are mapped to a corresponding customer ID number. Therefore when creating a master dataset on transaction information we can include gender, age, and income information for some of the rows.

Similarly, the transaction ID dataset contains information of cab users' mode of payment, which is mapped to a corresponding transaction ID number as well as a customer ID number. This allows us to add payment mode data into our master dataset on overall transactions.