Data Intake Report

Name: <G2M Insight for Cab Investment Firm Project>
Report date: <14/03/2023>
Internship Batch:<LISUM19 course>
Version:<1.0>
Data intake by:<Harold Wilson>
Data intake reviewer:<intern who reviewed the report>
Data storage location: <https://github.com/DataGlacier/DataSets>

Tabular data details: Cab data

Total number of observations	359392
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	20.1MB

Tabular data details: City data

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	<4.0KB>

Tabular data details: Customer data

Total number of observations	49171
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	<1.0 MB>

Tabular data details: Transaction data

Total number of observations	440098
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	8.58MB

Proposed Approach:

Tabular data details: Cab data

- There are 7 features in this data set which has two data types as objects and five data types as numeric.
- Columns Date of Travel was renamed to Date to help in analysis.
- The Date is in excel serial format was converted to datetime in python.

Tabular data details: City data

- There are 3 features in this data set which all have data types as objects
- The data types of column population and users was converted to numeric to help in analysis.
- Also commas and spaces were removed from columns population and users for easy data transformation and creation of feature engineering.

Tabular data details: Customer data

- Columns Income (USD/Months) was renamed to Income to help in analysis.
- There are 4 features in this data set which has one data types as objects and three data types as numeric.

Tabular data details: Transaction data

- There are 3 features in this data set which has two data types as numeric and one data types as object.
- There are no missing values in this data set
- There are no duplicates.

In general:

- There are no missing values in this data set
- There are no duplicates.