Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 04/19/2024.
Internship Batch: LISUM32
Version: 1.0
Data intake by: Vedant Wagh
Data intake reviewer:
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	21.2MB

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	759 bytes

Customer ID

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	1.00 MB

Transaction ID

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	9 MB

Proposed Approach:

Data Preprocessing:

Date conversion:
The 'Date of Travel' in the cab data is converted from an integer format to a datetime object

Merging: Multiple datasets (transaction, customer, cab, city) are merged to form a comprehensive dataset that captures all aspects of cab rides, customer demographics, and city details.

Cleaning: The combined dataset is cleaned by checking for missing values and duplicates. Missing values are imputed with the mean, while duplicates are searched for but not yet removed.

Data Validation (De-duplication Identification):

De-duplication is identified as an important step, but the current code only includes a check for duplicates without handling them. The approach should involve removing duplicates to ensure data quality.

Handling Missing Values and Outliers:

Missing values are filled with mean values across the dataset. This assumes that the mean is an appropriate estimate for the missing data, which may not always hold true for all features.

Outlier detection is attempted by calculating the Interquartile Range (IQR), but the code snippet is incomplete and needs correction.

Exploratory Data Analysis (EDA):

Statistical analysis: Descriptive statistics are provided for the cleaned dataset to understand the central tendency, dispersion, and shape of the data's distribution.

Visual insights: A series of plots are created to visualize the distribution of cab trips across cities, revenue comparison between two companies, the distribution of customer age, income by company, and the relationship between price charged and cost of trip.

Hypothesis Testing:

Seasonality in cab usage: An analysis is conducted to identify monthly patterns in cab usage.

Customer age impact on cab usage: A T-test is performed to compare the frequency of cab usage between younger and older customers.

Company dominance by time period: The number of transactions for each company is compared monthly.

Margins and customer count: A scatterplot is created to examine the relationship between customer count and profit margins for each company.

Customer segment attributes: Customer usage is analyzed by age groups and compared between companies.

Assumptions:

The conversion of dates assumes the Excel format which counts days from January 1, 1900.
Filling missing values with the mean assumes that data is missing at random and that the mean is a representative value.

Outlier detection via IQR assumes a normal distribution and may not be suitable for all variables.

The merged dataset represents an accurate joining of the different datasets based on the common keys without loss of information.

Data Quality Analysis Recommendations:

De-duplication should be carried out to remove any duplicate records identified.
The outlier detection method should be validated to ensure it is appropriate for the data distribution of each variable.

Additional checks should be implemented to handle anomalies, such as extremely high values in the Price Charged or Cost of Trip, which could be errors or true outliers.
A more nuanced approach to missing data could include using median values or predictive imputation techniques, especially for non-normal distributions.