Data Intake Report

Name: G2M insight for cab Investment firm
Report date: 14/05/2022
Internship Batch: LISUM09
Version: 0.1
Data intake by: Shwetha Basavanagowda
Data intake reviewer: NA
Data storage location: NA

Tabular data details:

Cab_Data.csv

Total number of observations	359392
Total number of files	<Number of files received>
Total number of features	7
Base format of the file	.csv
Size of the data	19.2MB

Customer_ID.csv

Total number of observations	49171
Total number of files	<Number of files received>
Total number of features	4
Base format of the file	.csv
Size of the data	1.5MB

Transaction_ID.csv

Total number of observations	440098
Total number of files	<Number of files received>
Total number of features	3
Base format of the file	.csv
Size of the data	10.1MB

City.csv

Total number of observations	20
Total number of files	<Number of files received>
Total number of features	3
Base format of the file	.csv
Size of the data	608 bytes

Proposed Approach:

- First Check the number of rows and columns and its data types and size of all the dataset files provided.
- Convert the data-types as required and rename the coloumn names as required
- Check for any null values and if any remove those records or replace with suitable values as per the scenario.
- Check for any duplicate records and if any remove the duplicate records
- Then for analysis merge or append the files as required to create master data and create new features/columns which required for analysis.
- Convert some numeric data to categorical values for analysis.
- Then check for outliers in numeric attributes by plotting box plots. If you find any outliers, treat those values by removing those records or replacing it with suitable values as per the scenario. In some cases leave the outliers as it is if you have a strong reason/assumptions that support.
- Check the Counts of all the categorical data and plot them if you required.
- Then depending on the plots and hypothesis get your insights.

Assumptions:

In our Analysis of the given datasets I am assuming

- "Cost of Trip" in Cab_Data.csv as expenses incurred by the company for that particular ride.
- "Price Charged"in Cab_Data.csv as Price which the customer/user will be paying to the company.
- "Users" in City.csv as number of cab users in that city.
- Outliers in "Profit" and "Price per KM" columns in master dataset are due to availing of premium services and use of luxury cars by the customers.