Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 14th June 2023
Internship Batch: LISUM22
Version:
Data intake by: Sri Pallavi Chittimalla
Data intake reviewer:
Data storage location: https://github.com/DataGlacier/DataSets

Tabular data details: Cab_Data.csv

Total number of observations	359392
Total number of files	1
Total number of features	7 [Data types: float64(3), int64(1), object(3)]
Base format of the file	.csv
Size of the data	19.2 MB

Tabular data details: City.csv

Total number of observations	20
Total number of files	1
Total number of features	3 [Data types: object(3)]
Base format of the file	.csv
Size of the data	608 bytes

Tabular data details: Customer ID.csv

Total number of observations	49171
Total number of files	1
Total number of features	4 [Data types: int64(3), object(1)]
Base format of the file	.csv
Size of the data	105 MB

Tabular data details: Transaction ID.csv

Total number of observations	440098
Total number of files	1

Total number of features	3 [Data types: int64(2), object(1)]
Base format of the file	.csv
Size of the data	10.1 MB

Proposed Approach:

Data Cleaning and Preprocessing

- We have split the data into training, validation, and test set in a 40:40:20 ratio to identify the outliers, we had a boxplot for the columns KM Travelled, Price Charged, Cost of Trip, and Income.

Univariate analysis

- Outliers are present in the Price_Charged feature, but we are not treating this as an outlier due to the unavailability of trip duration details.

Key findings and insights

- Profit of rides is calculated keeping other factors constant and only Price Charged and Cost of Trip features are used to calculate profit