Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 14/03/2023
Internship Batch: LISUM19
Version:<1.0>
Data intake by: Harshith Sakala Santhosh
Data intake reviewer:<intern who reviewed the report>
Data storage location: https://github.com/sakalaharshith/G2M-insight-for-Cab-Investment-firm.git

Cab_Data.csv:

Total number of observations	359393
Total number of files	4
Total number of features	7
Base format of the file	.csv
Size of the data	24,524 KB

City.csv:

Total number of observations	21
Total number of files	4
Total number of features	3
Base format of the file	.csv
Size of the data	1 KB

Customer_ID.csv:

Total number of observations	49172
Total number of files	4
Total number of features	4
Base format of the file	.csv
Size of the data	1027 KB

Transaction_ID.csv:

Total number of observations	440099
Total number of files	4
Total number of features	3
Base format of the file	.csv
Size of the data	8788 KB

Proposed Approach:

- Duplicate values are present in multiple features like Company, Gender, Age, Income, City, KM Travelled, Price Charged, Cost of Trip etc... The duplicate values were not removed as it makes sense that the categories for some of these features are 2 and less and if we take a feature like ‘City’ there will be only limited number of unique values in this case 20 unique cities and repetition of these city names is expected and same goes for all features.
- When it comes to outliers, outliers were present in price_charged features but without a proper context information about trip and under assumption that reason for outliers could be either bad weather or usage of high-end vehicles etc.… was considered and also there are some features generated from existing data also had outliers in them for example profit_percentage_per_ride etc... as these were again extracted from cost_of_trip and price_charged so these outlier values are considered valid.
- As mentioned before new features were generated from existing ones and they are profit_percentage_per_ride, price_charged_per_km, price_charged_per_km from cost_of_trip and price_charged to better understand the performance and importance of each cab company.
- Monthly usage of cabs for both companies was extracted to understand how these companies work in holiday months.
- Gender preferability of cab company was also calculated to better understand the performance of each company.