Data Intake Report

Name: G2M Insight for cab Investment firm
Report date: 11/12/2024
Internship Batch: LISUM39
Version: 1.0
Data intake by: Patrick Otieno
Data intake reviewer:
Data storage location:

Tabular data details:

Cab Data Dataset:

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	20663 KB

City Dataset:

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	1 KB

Customer ID Dataset:

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	1027 KB

Transaction ID Dataset:

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	8788 KB

Proposed Approach:

Unique Row Identification: To enhance the analytical process, all four datasets were consolidated, ensuring each row was distinctly identified by its Transaction ID. This method facilitated a comprehensive examination of individual transactions, enabling a granular analysis of customer behaviors across the dataset.

Handling Duplicate Rows: To maintain data integrity, the code dataset.drop_duplicates() was utilized across the combined dataset to remove redundant entries. Additionally, dataset.dropna() was applied to cleanse the dataset by discarding rows with missing values.

Dataset Overview: The merged dataset contained critical columns, including Transaction ID, Customer ID, Payment Mode, Date of Travel, Company, City, KM Travelled, Price Charged, Cost of Trip, Gender, Age, Income (USD per Month), Population, Users, and Year. For deeper insights, additional calculated fields such as Net profit, Price per KM, Profit per KM, and Number of Rides were created. This comprehensive structure allowed for focused analysis on metrics like profitability, geographic distribution, demographic segmentation, payment preferences, and customer retention, guiding investment strategy.

Assumptions:

- The analysis accounted for potential noise outside the provided data.
- The timeframe for analysis was limited to 2016–2018.
- Random sampling was presumed for data collection.
- Only cash and card payments were considered in this analysis.
- The profit calculation relied on the formula: Price Charged - Cost of Trip.