Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 9/5/2024
Internship Batch: LISUM33
Version: 1.0
Data intake by: Lewis Kiptoo
Data intake reviewer:
Data storage location: Github

Tabular data details:

Transaction ID datasets

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	8788KB

City datasets

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	1K

Customer ID datasets

Total number of observations	49171
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	1027KB

Cab 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	20663KB

Proposed Approach:

Unique Row Identification:

I combined the four datasets to enhance the performance of specific analysis tasks. Because I intended to examine each row as a separate transaction, I used the key of (Transaction ID) to uniquely identify each row in the aggregated dataset. In this manner, we are able to examine the several transactions that originated with the same client.

Duplicate Rows:

For every dataset, the code snippet "dataset.drop duplicates()" was run in order to remove any duplicate entries. Additionally, each dataset was subjected to the following code snippet, "dataset.dropna()," in order to remove any rows with N/A values.

Dataset Understanding:

Transaction ID, Customer ID, Payment Mode, Date of Travel, Company, City, KM Traveled, Price Charged, Cost of Trip, Gender, Age, Income (USD per Month), Population, Users, and Year are the columns that make up the combined dataset. The following columns have been added: Number of Rides, Profit, Price per KM, and Profit per KM. I decided to concentrate my analysis on five key areas after reviewing the available data: profits, locations, age groupings, payment methods, and client retention. These, in my opinion, are the best indicators of which firm is doing better and how profitable it will be to invest in them.

Assumptions:

1. Both companies' data analyses were conducted with the assumption that there was external noise in addition to the supplied data.
2. Data analysis was done for both businesses with the understanding that the datasets were
3. limited time frame from 2021 to 2023.
4. Fifth, the datasets were chosen at random.
5. The only payment options that are taken into consideration are cash and cards.
6. Price Charged: The only variables used to determine profit were the trip costs.