Data Intake Report

Name: G2M Insight for Cab Investment Firm
Report Date: 03/14/2023
Internship Batch: LISUM19
Version: 1.0
Data Intake By: Vincent Weng
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 Identication:

I merged all four datasets together in order to better perform certain anaylsis operations. I uniquely identified each row of the combined dataset by the key of (Transaction ID), because I wanted to analyze each row as a singular transaction. This way, we are able to analyze the different transactions that came from the same customer.

Duplicate Rows:

The following code snippet "dataset.drop_duplicates()" was performed on each dataset to eliminate any rows that were duplicates. Furthermore, the following code snippet "dataset.dropna()" was performed on each dataset to eliminate rows that had N/A values.

Dataset Understanding:

The merged datset has the following columns: 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. I've added in the following columns: Profit, Price per KM, Profit per KM, and Number of Rides. After looking at the data provided, the five main areas that I wanted to focus my analysis on, which were profits, locations, age & income, payment method, and customer retention. I believe these factors will best analyze which company is performing better and give a sense of how incentivizing investing in them will be.

Assumptions:

• Data analysis were performed for both companies under the assumption of external noise beyond the data that was provided.

• Data analysis were performed for both companies under the assumption that the dataset were time constrained between 2016 and 2018.

• The datasets were drawn from random selection.

• Payment by cash or card are the only payment methods considered.

• Price Charged - Cost of Trip were the only factors when calculating profit.