Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 14/03/2024
Internship Batch: LISUM31
Version: 1.0
Data intake by: Lakshita Bansal
Data intake reviewer: Data Glacier
Data storage location: <location URL eg: github, cloud>

Tabular data details:

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	20663 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

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

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

Proposed Approach:

There are no duplicate values and no NA values in the dataset. All four datasets have been merged into one such that the resulting data frame will contain only the rows where the merge key exists in all data frames.

When merging datasets, I checked for inconsistencies that could introduce duplicates, such as differing formats or multiple entries for a single identifier across tables. For instance, a single Transaction ID should not be linked to multiple Customer IDs unless the business logic explicitly allows for this scenario.

Next, I ensured that supposed unique identifiers, such as Transaction ID and Customer ID, do not have duplicates within their respective datasets. This can be achieved using aggregation methods to count occurrences of each identifier and flagging any counts greater than one.

<u>Assumptions:</u>

It was assumed that Date of Travel data was consistently recorded for trend analysis, formats were standardized or could be standardized without information loss and Transaction ID and Customer ID were correctly assigned and unique.

It was also assumed that the transactional data (e.g., KM Travelled, Price Charged, Cost of Trip) were accurately recorded at the source.

<u>Hypotheses:</u>

1. Gender plays a role in profit for yellow cab

2. Age plays a role in profit for pink cab

3. Payment Mode Preference

4. Preference for Cab Company Based on Trip Cost

5. Number of rides is correlated with Busy Cities

6. Customer with high income are more loyal