Data Intake Report

Name: G2M Insights for Cab Investment Firm
Report date: 22-10-2023
Internship Batch: LISUM26
Version:<1.0>
Data intake by: Mohammad Shehzar Khan
Data intake reviewer:<intern who reviewed the report>
Data storage location: <location URL eg: github, cloud>

Tabular data details:

Name of the file	Cab_data.csv
Total number of observations	359,392
Total number of features	7
Base format of the file	.csv
Size of the data	19.2 MB

Name of the file	City.csv
Total number of observations	20
Total number of features	3
Base format of the file	.csv
Size of the data	608 Bytes

Name of the file	Customer_ID.csv
Total number of observations	49,171
Total number of features	4
Base format of the file	.csv
Size of the data	1.5 MB

Name of the file	Transaction_ID.csv
Total number of observations	440,097
Total number of features	3
Base format of the file	.csv
Size of the data	10.1 MB

Proposed Approach:

- Mention approach of dedup validation (identification)
- Mention your assumptions (if you assume any other thing for data quality analysis)

The approach has been done in the following manner:

- Data Exploration & Understanding
- Feature Engineering
- Data Merging
- Data Visualization
- Data Analysis
- Recommendations

Assumptions:

- Outliers are present in Price_Charged feature but due to unavailability of trip duration details ,we are not treating this as outlier.
- Profit of rides are calculated keeping other factors constant and only Price_Charged and Cost_of_Trip features used to calculate profit.
- Users feature of city dataset is treated as number of cab users in the city. We have assumed that this can be other cab users as well(including Yellow and Pink cab).
- When we looked at the Cost of the Trip for each ride, there were some rides that were unusually costly as compared to all other rides. Possible reasons for such expensive rides could be either Premium rides, or Overnight rides, or Long distance rides. We have called these people as premium customers.
- We have assumed that a trip upto 8km is short distance trip, a trip from 8km to 30km is Medium distance trip, and a trip above 30km is a long distance trip.
- We have assumed that any customer who uses the same cab service atleast every 4 months on average is a loyal customer to that cab service. Here, we have data for 36 months approx., so any customer who has 9 rides with a particular company is a loyal customer to that company.