Data Intake Report

Name: G2M Insights for Cab Investment Firm
Report date: <11th February 2023>
Internship Batch: LISUM18
Version:<1.0>
Data intake by: Tanmay Umesh Potbhare
Data intake reviewer:<intern who reviewed the report>
Data storage location: <https://github.com/TanmayPotbhare/G2M-Insight-for-Cab-Industry-Investment >

Cab_Data.CSV details:

Total number of observations	359393 (3Lakh 59Thousand 93) Rows
Total number of files	4
Total number of features	7 Columns
Base format of the file	CSV
Size of the data	20.1 Megabytes (20.1 MB)
Out of 7 columns, 5 columns are Numerical and 2 are string or text or alphabetical.	

Customer_ID.CSV details:

Total number of observations	29290 (29Thousand 290) Rows
Total number of files	4
Total number of features	4 Columns
Base format of the file	CSV
Size of the data	1 Megabyte (1 MB)
Out of 4 columns, 2 are numerical values and 2 are alphabetical.	

Transaction_ID.CSV details:

Total number of observations	440099 (4Lakh 44Thousand 99) Rows
Total number of files	4
Total number of features	3 Columns
Base format of the file	CSV
Size of the data	8.58 Megabytes (8.58 MB)

City.CSV details:

Total number of observations	21 Rows
Total number of files	4
Total number of features	3 Columns
Base format of the file	CSV
Size of the data	1 Kilobytes (1 KB)

Proposed Approach:

- Mention the approach of dedupe validation (identification)
    - Before starting for Data analysis and exploring, my first action was to create a hypothesis no matter how rogue it may sound and see what all possibility of questions is this hypothesis answering to.
    - After the Selecting very few hypothesis which can tell many linked answer about the data, I started to analyses and explore the datasets.
    - Analysis was done in a way where I checked for null values, outliers, or creating new columns and joining them, etc.
    - After completing the dataset named as <final dataset> I merged all the datasets and then created one last column as Profit which is subtraction operation of two columns.
    - I divided visualization on some criteria:
        - Profit
        - Customers
        - City
        - Transaction

      On the basis of analysis on these four I was able to answer various important answers in regard to the Yellow and Pink Cab Company.

- Mention your assumptions (if you assume any other thing for data quality analysis)
    - I did not assume anything I tried to solve the answers to my hypothesis on the given data.
    - But, only the Price Charged field had outliers which was seen using boxplot but I did not do anything about it as there was not enough data.