Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 2/8/2023
Internship Batch: LISUM24
Version: 1.0
Data intake by: Alexander Quesada Quesada
Data intake reviewer: NA
Data storage location:
https://github.com/papitaAlgodonCplusplus/LISUM24/tree/main/Week%202/datasets

Tabular data details:

File name	Cab_Data.csv
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

File name	City.csv
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

File name	Customer_ID.csv
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

File name	Transaction_ID.csv
Total number of observations	440008
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	8788 KB

Proposed Approach:

Using datasets:

1. Cab_Data.csv
2. Transaction_ID.csv
3. Customer_ID.csv

I discarded duplicated values by doing inner merge between them and creating a new dataset:

1. Merged_df

By doing inner join, only columns which index "Consumer ID" and "Transaction ID" being equal are taking into account for EDA, else, discarded.

Assumptions:

- The amount of rides provided for both companies are as in total amount, I'm assuming there was no bias towards pink cab company by not including % of rides given timeline.