Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: June 11, 2021
Internship Batch: LISUM01
Version:1.0
Data intake by: Nicolette Peterkin
Data intake reviewer: Nicolette Peterkin
Data storage location: https://github.com/DataGlacier/DataSets

Tabular data details:

Name of file: Transaction_ID.csv

Total number of observations	440099
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	8.58 MB

Name of file: Customer_ID.csv

Total number of observations	49172
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1 MB

Name of file: City.csv

Total number of observations	21
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	759 Bytes

Name of file: Cab_Data.csv

Total number of observations	359393
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	20.2 MB

Proposed Approach:

- A structured approach of deduplicate validation will be used for this assignment. Where I will go through each .csv file to see the similarities if they can be joined; and on what levels they can be joined. The assumption made at the moment is that there are duplicates on a tabular level that can be sorted through using python and or EDA Notebook.