DATA INTAKE REPORT

Name: G2M Insight for Cab Investment firm

Report Date:16th June 2021

Internship Batch: LISUM01

Version: 1.0

Data Intake by: Blaise Papa

Data Storage Location: https://github.com/DataGlacier/DataSets.git

Cab_Data.csv

Total number of observations	359,392
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	20.66 MB

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	1KB

Transaction_ID.csv

Total number of observations	440,098
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	8.788MB

Customer_ID.csv

Total number of observations	49,171
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1.027MB

Proposed Approach:

- **Load all the datasets and review them, identify relationships and merge them into one dataset**
- **Check for duplicate fields, null rows, and outliers. If duplicates are found drop them, if null rows are found impute them.**
  Since there were none of the above mentioned this approach was not used
- **Assuming the start date of collecting the data is 1-1-2016 and the end was 30-12-2018.**
- **The outliers are to kept as they help understand cab trips/transactions**
- **There are no other cab companies in the mentioned states**
- **The difference between the cost of the trip and the price charged is the profit; a negative profit is considered a loss**
- **In the day_week 0 is Monday through to 6 which is Sunday**
- **Each user uses only a single account (User ID)**