Data Intake Report

Name: G2M insight for Cab Investment Firm
Report date: July 14, 2022
Internship Batch: LISUM11: 30
Version:1.0
Data intake by: Baoze Lin
Data intake reviewer: _____
Data storage location: https://github.com/BaoGeist/CabDatasets

Tabular data details:

Cab Data

Total number of observations	359392
Total number of files	1
Total number of features	6
Base format of the file	.csv
Size of the data	20.663 MB

City

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

City Size (

Total number of observations	20
Total number of files	1
Total number of features	2
Base format of the file	.csv
Size of the data	1 KB

Transaction ID

Total number of observations	440098
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	8.788 MB

Customer ID

Total number of observations	49171
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	1.027 MB

Proposed Approach:

- First read in datasets, look for similar columns and join them together with outer joins to see how they connect, and what data is missing
    - Data that can be derived from our current columns are Profits
- Then utilize the different methods of displaying data to show findings
- An assumption is that in our master table, the entries with NaN for company and other related features are a result of it being from a different cab company
- There is no cab data for San Francisco. Moving forward, there is an assumption that although city information is provided for
- There is an assumption that the Cab Uses feature in Transaction_ID.csv include Yellow Cab, Pink Cab and other Cab data. It is assumed that entries that do not match up with the Pink Cab and Yellow Cab data are only the other Cab data, and not missing Pink Cab and Yellow Cab data. Furthermore, it is assumed that none of the other Cab companies mentioned hold a majority in the Cab uses, and thus cannot affect the investment decision.