Data Intake Report

Name: G2M Cab Investment Project
Report date: January 19th, 2023
Internship Batch: LISUM17
Version: 1.0
Data intake by: Amogh Vig
Data intake reviewer: N/A
Data storage location: https://github.com/avig00/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	21.2 MB

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	759 bytes

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	1.1 MB

File name: Transaction_ID.csv

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	9 MB

Proposed Approach:

- Duplicate rows were removed for each raw data file using the .drop_duplicates() function of Pandas.
- To remove NA values from the data, using the .drop_na() function of Pandas was used.
- A key assumption made for the analysis is that profits can be calculated using only the cost of the trip and the price charged for the trip, no other factors are accounted for when calculating profit.
- The “Price Charged” feature of the raw cab data (Cab_Data.csv) did appear to have some outliers. However, due to having insufficient evidence to prove that these outliers were mistakes, they were included in the analysis so that the provided data set is fully represented.