Data Intake Report

Name: G2M Insight for Cab Investment Firm
Report date: June 13, 2022
Internship Batch: LISUM 10
Version:1.0
Data intake by: Jeeyeon Shon
Data intake reviewer: Jeeyeon Shon
Data storage location:

1) https://github.com/shonjeeyeon/DataSets (for Cab_data, City, Transaction ID, and Customer ID)
2) https://www.kaggle.com/donnetew/us-holiday-dates-2004-2021 (for US Holiday Dates)

Tabular data details:

Cab_Data.csv

Total number of observations	359,392
Total number of files	1
Total number of features	8
Base format of the file	.csv
Size of the data	20.2 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	759 Bytes

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	1.0 MB

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	8.58 MB

US Holiday Dates (2004-2021).csv

Total number of observations	342
Total number of files	1
Total number of features	6
Base format of the file	.csv
Size of the data	15.3 KB

Proposed Approach:

- Customer ID can be used to remove duplicate data of a customer.
- Transaction ID is unique per trip and therefore can be used to deduplicate rides.
- Price and cost can be used to calculate profit.
- Because the datasets do not include types of cars used or customers' wait time spent in rides, I will assume the cars and wait times are similar in each trip and over companies.
- Customers' income information does not include their residing states or size of families, so I will assume the circumstances related to deciding income level is same over customers.