Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 8/14/2022
Internship Batch: 5068942
Version:<1.0>
Data intake by: Zhan Shi
Data intake reviewer: Zhan Shi
Data storage location: https://github.com/zhanshi1997/dgintern/tree/week2

Tabular data details:

Total number of observations	4312704
Total number of files	5
Total number of features	14
Base format of the file	.csv
Size of the data	39.5MB

Cab_Data.csv – this file includes details of transaction for 2 cab companies
Customer_ID.csv – this is a mapping table that contains a unique identifier which links the customer's demographic details
Transaction_ID.csv – this is a mapping table that contains transaction to customer mapping and payment mode
City.csv – this file contains list of US cities, their population and number of cab users
us-federal_holidays_2011_2020.csv – this file contains list of US holidays and their dates

Proposed Approach:

- I used the foreign keys (transaction ID and customer ID) in different tables as references to join interrelated data into a single table with every single row including all the attributes. Therefore, only one copy of each record is stored to improve the data quality.
- By searching in the table and doing matching, i also found the redundant data in the table.

Assumptions:

- Is there any seasonality in profits of these two cab services?
- What's the profit per ride of these two cab services?
- Is there any seasonality in number of customers using the cab service?
- Is there any customer preference on holiday?
- How is the distribution of customers based on income, age and city?
- Is there any city-wise difference in number of customers? How does it look like?
- Is there any difference of customer retention of these two cab service?
- What are the expectation of profits and customer population in the next year for two cab services?
- **... ...**