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:

- Used the foreign keys (transaction ID and customer ID) as references to join interrelated data into a single table with every single row including all the attributes.
- Used linear regression to roughly predict the profit in 2019
- Used Polynomial regression to roughly predict the customer in 2019

Assumptions:

- Profit of each ride is calculated by subtracting Price Changed with Cost of Trip.
- Total number of customers of both cab service is larger than Customer_ID.csv and we also found redundant fields in the Customer_ID dataset. We assumed that customers can be other cab customers as well (including Yellow and Pink cab) while some customers didn't use any cab service during the study time.
- Users feature of City.csv is treated as number of cab users in the city. We have assumed that this can be other cab users as well (including Yellow and Pink cab).