Data Intake Report

Name: G2M Cab Investment Report
Report date: 22nd June, 2021 (22/06/2021)
Internship Batch: LISUM01
Version: 1.0
Data intake by: Udbhav Balaji
Data intake reviewer: Udbhav Balaji
Data storage location: GitHub URL: https://github.com/udbhavbalaji/DataSets

Tabular data details:

File Name	Cab_Data.csv
Total number of observations	359392 rows
Total number of files	1
Total number of features	7 Columns
Base format of the file	Comma Separated Values (.csv)
Size of the data	21.2 MB

File Name	City.csv
Total number of observations	20 rows
Total number of files	1
Total number of features	3 Columns
Base format of the file	Comma Separated Values (.csv)
Size of the data	4 KB

File Name	Customer_ID.csv
Total number of observations	49171 rows
Total number of files	1
Total number of features	4 Columns
Base format of the file	Comma Separated Values (.csv)
Size of the data	1.1 MB

File Name	Transaction_ID.csv
Total number of observations	440098 rows
Total number of files	1
Total number of features	3 Columns
Base format of the file	Comma Separated Values (.csv)
Size of the data	9 MB

File Name	USHolidays.csv
Total number of observations	11 rows
Total number of files	1
Total number of features	4 Columns
Base format of the file	Comma Separated Values (.csv)
Size of the data	4 KB

Proposed Approach:

- **Assumption:** Profit is calculated for each ride individually.
- **Approach for Dedup Validation**: While creating the master dataset, we ensured no duplicates by performing a left merge on the records given in Cab_Data.csv and Transaction_ID.csv. This helped make sure that only the transactions relevant to Yellow and Pink companies were being used for the evaluation.
- We go through each of the datasets and according to its shape, we add it to a master dataset. For Example, we map a particular city's population and users to each record in the master data where the ride took place in that particular city. We do a similar process for cab users and customer details.
- Once the master dataset is prepared, Exploratory Data Analysis (EDA) is done to further explore the data in order to hopefully see some patterns that can help us make relevant and insightful decisions, that can be backed up by the data. Relationships were explored and plotted to increase understanding for the required stake-holders of the data and decision-makers.
- Some probable hypotheses were reported, and statistical hypothesis testing was performed to validate these hypotheses.