Data Intake Report

Name: Cab Investment Firm
Report date: 14/10/2022
Internship Batch: LISUM14
Version:<1.0>
Data intake by: Panala.Priyadharshini
Data intake reviewer:<intern who reviewed the report>
Data storage location: git hub

Tabular data details:

Total number of observations	355032
Total number of files	4
Total number of features	14
Base format of the file	.csv
Size of the data	32MB

Note: Replicate same table with file name if you have more than one file.

Proposed Approach:

- **Mention approach of dedup validation (identification)**

For this assignment we have four datasets. They are
Cab dataset: This data set having information about cab companies. We have to companies' data yellow cab and pink cab. And we have how many KM they travelled and how much they charged and also, we have date of travel information
City dataset: In this dataset we have cities names in the US and population. And also, we have information about how many cab users are there.
Customer dataset: In this data set we have information about customer Id, their gender and age and also monthly income of the users.
Transaction Dataset: In this data set we have transaction ID, customer Id and payment mode.

I merged all four datasets to make final dataset.

Validation:

- We don't have any duplicates in this dataset
- We don't have null values.
- We don't have outliers except price charged. but we don't have trip duration details. That's why I ignored these outliers.

- Mention your assumptions (if you assume any other thing for data quality analysis)

Assumptions:

- I have checked for the data distribution and correlation.