Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: September 5th 2022
Internship Batch: LISUM13
Version:1.0
Data intake by: Runtao Zhou
Data intake reviewer: N/A
Data storage location: https://github.com/DataGlacier/DataSets

Tabular data details:

Total number of observations	359392
Total number of files	5
Total number of features	14
Base format of the file	.csv
Size of the data	31 MB

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

Proposed Approach:

- Mention approach of dedup validation (identification)
  1. Checked each column to make sure they have the correct data type
  2. Checked for structure of the data to make sure each data point is consistent with other data point
  3. Checked for missing values or outliers
- Mention your assumptions (if you assume any other thing for data quality analysis)
  1. Customer's age distribution might be different for pink cab company and yellow cab company
  2. Male and female customer distribution might be different for pink cab company and yellow cab company.
  3. Customer payment method distribution might be different between pink cab company and yellow cab company
  4. The average price per KM might be different for pink cab company and yellow cab company
  5. Different cities have different profit margins for different cab companies
  6. Cash payment have bigger profit margin
  7. Different cab companies can have different customer base in different cities.