Data Intake Report

Name: G2M Insights for Cab Investment Firm
Report date: 20/08/22
Internship Batch: LISUM12
Version: 1.0
Data intake by: Carl Somers
Data intake reviewer: Carl Somers
Data storage location: https://github.com/DataGlacier/DataSets

Tabular data details:

Total number of observations	359393
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	21.2MB

Total number of observations	49172
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1.1MB

Total number of observations	440099
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	9.0MB

Total number of observations	21
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	759KB

Proposed Approach:

- **Basic Analysis** of each dataset involving searching for null & duplicate values.
- **Data Transformation** to create a main dataset, this can be performed using the transaction IDs and customer IDs as well as the City location. Date of travel converted into more usable datetime.
- **Feature Extraction** from the dataset, providing more valuable insights – profit, cost per km, price per km.
- **Exploratory Analysis** to further understand the distribution functions and the correlations which exist in the dataset. Breaking down the topic into 3 distinct main categories – location, time, and customer.
- **Comparative Analysis** comparing yellow and pink taxi companies to gauge their individual performance and scale. Includes time series analysis.
- **Hypothesis Testing** on several key and interesting questions to prove their statistical significance.