Data Intake Report

Name: G2M Insight for Cab Investment Firm
Report date: 2025.02.12
Internship Batch: LISUM42
Version: 1.0
Data intake by: Abdukhakimov Asatilla
Data intake reviewer: Data Glacier Team
Data storage location:
https://github.com/asat94/Data-Glacier-Internship/tree/main/WEEK%2002

Tabular data details:

Cab_data dataset details:

Total number of observations	359392
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	19.2MB

City dataset details:

Total number of observations	20
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	0.612KB

Customer_ID dataset details:

Total number of observations	49171
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1.5MB

Transaction_ID dataset details:

Total number of observations	440098
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	10.1MB

Proposed Approach:

Deduplication Validation Approach

- The combined dataset was created by merging three key datasets: Cab, Customer, and Transaction data. The City data was also cleaned and used for additional analysis.
- Each row in the merged dataset was uniquely identified by Transaction ID and Customer ID
- The `drop_duplicates()` function was applied to the combined dataset to eliminate exact duplicate records.

Assumptions for Data Quality Analysis

- The dataset is time-bounded between 2016 and 2018, ensuring that all trends and projections are within this period.
- The dataset is assumed to represent the complete set of transactions for both companies during the given time period, providing a comprehensive view of their performance.
- External market factors (e.g., fuel prices, economic conditions) are not considered beyond the available data.
- Payment methods are limited to cash and card transactions, excluding other digital or alternative payment options such as cryptocurrency.
- Profit calculation strictly follows:

  $$\textit{Profit} = \textit{Price Charged} - \textit{Cost of Trip}$$

  without considering additional operational expenses or discounts.