Data Intake Report

Name: G2M insight for Cab Investment firm (Must for all Specialization)
Report date: 7 August to 14 August 2023
Internship Batch: LISUM24 30 July - 30 Oct 2023
Version: 1.0
Data intake by: Abdullah Fatih Höcü
Data intake reviewer:<intern who reviewed the report>
Data storage location: https://github.com/hocuf/G2M-insight-for-Cab-Investment-firm--Data-Glacier.git

Tabular data details:

Total number of observations	19
Total number of files	11
Total number of features	7
Base format of the file	csv,.docx, jpynb
Size of the data	1.14 GB

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

Proposed Approach:

1. **Data Cleaning and Preparation:**
    - **Duplicate Handling:** The data will be sorted based on transaction ID, or any unique identifier present in the dataset. Duplicates will be identified based on these unique identifiers and removed.
    - **Handling Missing Data:** Rows with missing data will be flagged and analyzed. Depending on the nature and quantity of the missing data, imputation methods or removal might be applied.
    - **Data Transformation:** Convert categorical variables into a format suitable for analysis, possibly using encoding techniques. Normalize numerical data if required.
2. **Exploratory Data Analysis (EDA):**
    - **Demographic Analysis:** Identify which demographic group utilizes taxis the most. This will involve grouping the data by demographic and then calculating the sum or count of taxi uses.
    - **Company Popularity:** Assess which taxi company is more popular in each city. This will involve grouping by city and company and then comparing the count of uses.
3. **Assumptions:**
    - **Missing Data as Invalid:** Any missing data is considered as an invalid entry unless proven otherwise.
    - **Duplicate Transactions:** Any transaction ID appearing more than once is considered a duplicate.

- **Data Format:** Non-numeric values in numeric fields are considered errors and will be treated or removed.