Data Intake Report

Name: EDA_G2M Insight for Cab Investment Firm
Report date: 10/14/2024
Internship Batch:
Version: 1.0
Data intake by: Ky Dang
Data intake reviewer:
Data storage location:

Tabular data details:

1. **Cab_Data.csv**

Total number of observations	359392
Total number of files	
Total number of features	7
Base format of the file	.csv
Size of the data	20663KB

2. **City.csv**

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

3. **Customer_ID.csv**

Total number of observations	49171
Total number of files	
Total number of features	4
Base format of the file	.csv
Size of the data	1027KB

4. **Transaction_ID.csv**

Total number of observations	440098
Total number of files	
Total number of features	3
Base format of the file	.csv
Size of the data	8788KB

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

Proposed Approach:

- Mention approach of dedup validation (identification)

  First Step: Dataset Assessment

  Before proceeding, the datasets must be validated against key data quality standards, including **Data Reliability**, **Data Integrity**, and **Data Timeliness**…

  However, these standards may require adaptation due to the specific nature of this project. Here are the key considerations:

  - **Data Reliability**: The datasets are ambiguous regarding their sources, which raises concerns about their reliability.
  - **Dummy Factors**: Upon thorough observation, dummy or placeholder factors are present in the dataset.
  - **Data Timeliness**: The datasets do not meet the timeliness standard, as they are outdated (the most recent data is from 2016).

  Second Step: Data Inspection

  After uploading the tables into Python and conducting initial checks using Jupyter Notebook, the data appears to be relatively clean and organized. The following steps were taken:

  - **Data Shape Verification**: I checked the shape of the tables to ensure all records were loaded correctly.
  - **Duplicate Detection**: Deduplication techniques were applied to identify duplicates and missing values. However, no significant issues were found. The data seems clean and possibly pre-cleaned beforehand.