Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 11/13/2024
Internship Batch: LISUM39
Version:<1.0>
Data intake by: Devin Chau
Data intake reviewer: Data Glacier
Data storage location:
https://github.com/mynameisdevinchau/Data-Glacier-Internship/tree/main/Week%202

Tabular data details: Cab_Data

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	20.2 MB

Tabular data details: City

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	759 Bytes

Tabular data details: Customer_ID

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	1.00 MB

Tabular data details: Transaction_ID

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	8.58 MB

Proposed Approach:

- Mention approach of dedup validation (identification)
  - For each dataset, I looked through their primary keys and combined them if their primary keys were contained.
    - Using Cab_Data, I merged the Transaction_ID table with its ‘Transaction ID’ and combined the two to create the new data frame ‘merged_df’.
    - Next, I merged the City table with the merged_df with ‘City’ to simplify the data frames.
    - Finally, after doing that, I merged merged_df with the Customer_ID table on ‘Customer ID’ to create a singular data frame for all values, avoiding anomalies.
- Mention your assumptions (if you assume any other thing for data quality analysis)
  - I assumed Transaction ID and Customer ID’s to be unique.
    - I assumed that Customer ID’s to be unique to all users but can reoccur
    - I assumed that Transaction ID’s to be unique and can not reoccur.

**Note: Convert this doc in pdf and provide the link of pdf file in your dashboard.
Please do not forget to remove this section while converting the file into pdf.**