Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 11/11/2024
Internship Batch: LISUM39
Version:1.0
Data intake by: Vinicius Brun
Data intake reviewer:
Data storage location: https://github.com/vinbrun/data-glacier-virtual-internship/tree/main/week-2/DataSets

Tabular data details: Cab_Data.csv

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	20.1 MB

Tabular data details: City.csv

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	0.8 KB

Tabular data details: Customer_ID.csv

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	1.0 MB

Tabular data details: Transaction_ID.csv

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	8.6 MB

Proposed Approach:

- There appears to be no missing values in the existing columns.
- Hypothesis: investigate why there are more rows in Transaction IDs than in Cab_Data.
- Deduplication approach:
  - Identify exact duplicates based on the column Transaction_ID in the file Transaction_ID.csv (unique identifier).
  - Identify exact duplicates based on the column Customer_ID in the file Customer_ID.csv (unique identifier).
  - Identify duplicates usin Date_of_Travel, Company, City, KM Travelled, Price Charged, and Cost of Trip together in the file Cab_Data.csv.
- Data validation approach:
  - Join Cab_Data.csv with Transaction_ID.csv on Transaction_ID to ensure that each trip in Cab_Data.csv has a corresponding transaction in Transactoin_ID.csv.
  - Join Transaction_ID.csv with Customer_ID.csv on Customer_ID to ensure that transaction in Transactoin_ID.csv has a corresponding customer in Customer_ID.csv