Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 09/02/2022
Internship Batch: LISUM06
Version: 1.0
Data intake by: Noé Gracida Hernández
Data intake reviewer:
Data storage location: https://github.com/DataGlacier/DataSets

Tabular data details: Cab_Data.csv

Total number of observations	359,392
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	759 Bytes

Tabular data details: Customer_ID.csv

Total number of observations	49,171
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1 MB

Tabular data details: Transaction_ID.csv

Total number of observations	440,098
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:

- The Cab_Data.csv dataset can be joined with the Transaction_ID.csv dataset, and then with the Customer_ID.csv dataset, both with full joins. Then, the resulting data frame can be joined to the City.csv dataset using a left join to exclude the data of the cities that have no observations.
- With the four datasets merged, it's proposed to drop the resulting observations with null values. They represent almost the 20% of the observations of the dataset, so imputing techniques are not recommended. Without these observations, the number of rows is still high (more than 350,000).
- The Master dataset did not show duplicates.
- The Master dataset presents outliers on the Price Charged variable. For some models, it would be a good idea to treat them.
- There is a big difference in the proportion of the categories of the Company variable. For some models, it would be a good idea to apply some balancing techniques.