Data Intake Report

Name: G2M Insight for Cab Investment Firm
Report date: 13/02/2024
Internship Batch: LISUM31
Version: 1.0
Data intake by: Sarah Quayyum
Data intake reviewer: Data Glacier
Data storage location:

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	21.2 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	9 MB

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	1.1 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	4 KB

Proposed Approach:

- **Merge files**
    - **Merging all four datasets into one dataset allows us to analyze each row as a singular transaction. This makes it easy to analyze different transactions that may come from the same customer.**
- **Duplicate and drop rows**
    - **Rows that are duplicates and/or contain "NA" values should be dropped from the dataset. This can be done using the *dataset.drop_duplicates()* and *dataset.dropna()* functions in pandas.**
- **Investigating dataset**
    - **In order to inform our decision on which company is performing better and would be the most profitable to invest in, we need to look at the following key items: profit, price per KM, profit per KM, customer retention.**