Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 07/12/24/
Internship Batch: LISUM35
Version: 1.0
Data intake by: Griffin Palfrey
Data storage location: https://github.com/griffinpalfrey/DataSets

Tabular data details:

File name	Cab_data
Total number of observations	359,392
Total number of features	7 ‘Transaction ID’ and ‘Date of Travel’ features are integers; ’City’ & ‘Company’ features are strings; ’KM Travelled’, ‘Priced Charged’, ‘Cost of Trip’ features are type float.
Base format of the file	.csv
Size of the data	62.831935 MB

File name	City
Total number of observations	20
Total number of features	3 ‘City’ feature is a string; ‘Population’ & ‘Users’ features are integers
Base format of the file	.csv
Size of the data	0.004128 MB

File name	Customer_ID
Total number of observations	49,171

Total number of features	4 ‘Customer ID’ , ‘Income (USD/Month) and ‘Age’ features are integers’; ‘Gender’ feature is a string
Base format of the file	.csv
Size of the data	4.224885 MB

File name	Transaction_ID
Total number of observations	440,098
Total number of features	3 ‘Transaction ID’ & ‘Customer ID’ features are integers; ‘Payment Method’ feature is a string.
Base format of the file	.csv
Size of the data	33.887678 MB

Validation:
Deduplication is an important step in insuring the accuracy, integrity, and quality of our data sets. The idea is to check for redundancy, removing anything that may skew our data in unrealistic ways. The deduplication validation was performed using Python’s Pandas library, where the following steps were taken:

1. Exact Duplicate Identification: I identified records that were identical across all fields
2. Partial Duplication Identification: I identified records that had matching values in key fields but differed in other fields. This was completed by checking to see if the value for ‘Transaction ID’ was ever repeated. This would indicate that the same transaction had been recorded twice in the data set, which is not plausible.

Assumptions:
While the dataset is not current, we assume that the data is relevant to make decisions about future investment, and thus, is accurate and attuned to the modern day.