Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 9/13/2024
Internship Batch:
Version: 1.0
Data intake by: David Kalen
Data intake reviewer:
Data storage location: https://github.com/D-Kalen/Week-2

Tabular data details:

Total number of files: 4

Cab_Data.csv

Total number of observations	359392
Total number of features	7
Base format of the file	.csv
Size of the data	21.2 MB

City.csv

Total number of observations	20
Total number of features	3
Base format of the file	.csv
Size of the data	759 bytes

Customer_ID.csv

Total number of observations	49,171
Total number of features	4
Base format of the file	.csv
Size of the data	1.1 MB

Transaction_ID.csv

Total number of observations	440,098
Total number of features	3
Base format of the file	.csv
Size of the data	9 MB

Data Understanding/Summary

This data consists of multiple datasets related to cab services provided by Pink Cab and Yellow Cab. The Cab Data contains information on each trip, such as transaction details, distance traveled, cost, and profit. This dataset is important for analyzing both companies' financial performance and operational metrics.

The Transaction Data includes payment modes and customer-related details, linked via Transaction ID. Additionally, the Customer_ID.csv dataset holds customer information, allowing for analysis of customer behavior, loyalty, and ride frequency. The City.csv dataset provides city-specific data, helping to assess geographical performance and identify which cities generate more revenue or have higher ride volumes. Together, these datasets enable a comprehensive analysis of profitability, customer trends, and regional performance for both companies.

I verified that there were no duplicate rows by first checking each dataset as a data frame in Python pandas. I printed the ".duplicated().sum()" of each data frame to verify that no rows were an exact copy of each other. I also checked the duplicates of values in columns with the same process for different columns that shouldn't have duplicate values, like the Transaction ID in Cab Data.