Data Intake Report

Name: G2M Insight for Cab Investment Firm
Report date: 07/09/22
Internship Batch: LISUM11:30
Version:<1.0>
Data intake by: Alex Lindberg
Data intake reviewer:
Data storage location:

Tabular data details:

Total number of observations	359392
Total number of files	
Total number of features	19 (Transaction_ID: int64, Company: object, Date: int64, Customer_ID: int64, Gender: int64, Age: int64, Distance(KM): float64, Profit: float64, Revenue: float64, Cost: float64, Price_per_KM: float64, Payment_Mode: object, Monthly_Income: int64, Income_Percentage: float64, City: object, City_Percentage: float64, City_Users: float64, City_Population: float64)
Base format of the file	.csv
Size of the data	52.1 MB

Note: Replicate same table with file name if you have more than one file.

Proposed Approach:

- In order to identify possible duplicates in the data, I searched for rows that had the same Customer_ID, Date, Distance, Cost, and Revenue. It is very unlikely that a customer took two trips in the same day that had all of these factors the same.
- I assumed that the actual time period for the data was 01/01/2016 – 12/31/2018 instead of 01/31/2016 – 12/31/2018 as listed in the prompt. I did this because Date_Of_Travel had 1095 unique dates and there are only 1066 calendar days in the listed time period.
- I assumed the the 'Users' feature referred to the number of unique users combined between the two companies in any given city.