Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 14 December 2023
Internship Batch: LISUM28
Version:1.0
Data intake by: Trofymova Anastasiia
Data intake reviewer: Data Glacier
Data storage location:
https://github.com/Arsiry/data-glacier-internship/tree/main/Week2/DataSet

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: 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	759 B

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: 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

Proposed Approach:

- Table Data_Cab. Correcting the "Date of Travel" column data type.
- Table City. Convert columns “Population” and “Users” from string to numeric.
- Union All Data Table to one master DataFrame. Rename Columns.
- Identify and remove duplicates. Perform NA values and outtiers detection. We have data on 440,098 user transactions, but only 359,392 pertain to the Cab Industry during the period of interest. Therefore, our dataset contains missing values. For our research, we can delete rows with missing data. Also let's remove the data for the year 2019; there's very little of them and it might distort the overall picture.
- Let's make the data convenient for analysis. Add the month and year. Include the profit. Incorporate age and income level intervals of customers. Include intervals for city populations and the number of customers in cities.