Data Intake Report

Name: G2M insight for Cab Investment firm (Must for all Specialization)
Report date: 28/04/2024
Internship Batch: LISUM32: 01 Apr 24 - 01 July 24
Version: 2.0
Data intake by: Yiu Hung Leong
Data intake reviewer:
Data storage location: https://github.com/ErnestCodeHub/Data-Glacier-Internship/tree/main/Week%202

Tabular data details:

Cab Data Dataset:

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	20663 KB

City Dataset:

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	1 KB

Customer ID Dataset:

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	1027 KB

Cab Data Dataset:

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	8788 KB

Proposed Approach:

Data Cleaning & Manipulation:

1. Duplication and NA value analysis have been performed on all four dataset using .dropduplicate() and .dropn()

2. Date of Travel column from Cab data dataset has been converted from Excel serial dates to pandas datetime objects in order to identify the actual date of travel.

3. All four dataset have been merged into one master data using inner join

4. 4 new columns were created in the master dataset, namely, Users, Percentage(%) of taxi driver in City, Profit, Profit per KM have been created for better analysis purpose

Assumptions:

1. There are only 2 cab companies in the market as illustrated.
2. There is no other extra costing factor for both companies.
3. The sample datasets provided are a fair representation of the population data.
4. Price charged is the sole source of profit for both companies.