Data Intake Report

Name: G2M Insight for Cab Investment
Internship Batch: LISUM10: 30
Version: 1.0
Data intake by:
Fatemeh Bagheri
Data intake reviewer:
Data storage location: https://github.com/DataGlacier/DataSets

Tabular data details:

Cab_Data.csv

Total number of observations	359392
Total number of files	
Total number of features	7
Base format of the file	.csv
Size of the data	20.2 MB

City.csv

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

Customer_ID.csv

Total number of observations	49171
Total number of files	
Total number of features	4
Base format of the file	.csv
Size of the data	1 MB

Transaction_ID.csv

Total number of observations	440098
Total number of files	
Total number of features	3
Base format of the file	.csv
Size of the data	20.2 MB

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

Proposed Approach:

- Load all data as Pandas dataframe (python or Jupiter)
- This data assumes there are no duplicates in Transaction ID (Cab_Data and Transaction_ID), and in Customer ID (Customer_Data), City (City), so index these columns when loading each dataframe
- Cut the same rows and columns (include all the necessary information)
- Cab_ID and Transaction_ID merged into one Pandas dataframe using the join function
- Added additional columns to Cab_Data: Profit Generated, Trips, Yellow Cab Trips, Pink Cab Trips, Share of Yellow Cab Trips, Share of Pink Cab Trips – to make summaries easier to access
- Added additional columns to Customer_ID: Age Group, Income Bracket – for making a breakdown of users by demographic characteristic
- Correct some rows with nan or unreadable text
- Used the groupby function to show summaries of data, sometimes merging with other frames, such as breakdown of users by city, and the City dataframe
- Used Pandas Libraries to make several plots including, time series amd k means