Data Intake Report

Name: G2M insight for Cab Investment firm

Report date: 15/01/2023
Internship Batch: LISUM17
Version:<1.0>
Data intake by: Uday Singh
Data intake reviewer:
Data storage location: https://github.com/gitkym/dg_week_2.git

Tabular data details:

1. Cab_Data.csv

Total number of observations	359392
Total number of files	<Number of files received>
Total number of features	7
Base format of the file	.csv
Size of the data	21.2 MB

2. Customer_ID.csv

Total number of observations	49171
Total number of files	<Number of files received>
Total number of features	4
Base format of the file	.csv
Size of the data	1.1 MB

3. **Transaction_ID.csv**

Total number of observations	440098
Total number of files	<Number of files received>
Total number of features	3
Base format of the file	.csv
Size of the data	9 MB

4. **City.csv**

Total number of observations	20
Total number of files	<Number of files received>
Total number of features	3
Base format of the file	<.csv,.txt etc>
Size of the data	759 bytes

Proposed Approach:

- Store csv files in pandas database and use pd.Duplicated to check for duplicates
- Check for NA values
- Join datasets using the keys customer id, transaction id and city
- Convert to usable format (dates)
- Add useful columns (such as profit)
- Look at daily/monthly/weekly data to gain understanding
- Plot data to spot trends/seasonality
- Test for seasonality
- Compare the 2 companies