Data Intake Report

Name: G2M Insight for Cab Investment Firm Project

Report date: 20/06/2021
Internship Batch: LISUM01
Version: 1.0
Data intake by: William Ogweli Okomba
Data intake reviewer: William Ogweli Okomba
Data storage location: github link

1. **Tabular data details: Cab_Data.csv**

Total number of observations	359392
Total number of files	5
Total number of features	7
Base format of the file	.csv
Size of the data	20.1MB

2. **Tabular data details: City.csv**

Total number of observations	20
Total number of files	5
Total number of features	3
Base format of the file	.csv
Size of the data	759 BYTES

3. **Tabular data details:** Customer_ID.csv

Total number of observations	47171
Total number of files	5
Total number of features	4
Base format of the file	.csv
Size of the data	1.00MB

4. **Tabular data details:** Transaction_ID.csv

Total number of observations	440098
Total number of files	5
Total number of features	3
Base format of the file	.csv
Size of the data	8.58 MB

5. **Tabular data details:** us_holidays.csv

Total number of observations	3288
Total number of files	5
Total number of features	10
Base format of the file	.csv
Size of the data	186 KB

Proposed Approach:

- The cab_data dataset dates start from 01/01/2016 and not 31/01/2016 as stated.
- The outlier was detected on price_charged variable and removed.
- The outlier was detected on holiday and long holiday variables in us holiday dataset, these looks genuine that I did not remove them.
- Date of travel variable in cab dataset was changed to datetime and split into other variables such as date, month, and day (I dropped date of travel variable after extracting other variables from it).
- Profit margin variable was realized by subtracting the cost of trip variable from price charged variable.
- All the datasets have no missing values as well as duplicates.
- I added Us holiday dataset to enhance the insights, I changed its date variable to datetime and split it into other variables for so that it can marge with other datasets.
- Some of the dataset’s columns/variable names were modified like changing cases and removing white space so that they merge correctly.
- I assumed the datasets provided were genuine and up to date.
- Categorical variables were more compared to numerical variables in all dataset.
- All the five datasets were merged after cleaning.
- I carried out univariate, bivariate and multivariate data analysis.