Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 04/14/2024
Internship Batch: LISUM32
Version: 1.0
Data intake by: Apurwa Bhausaheb Sontakke
Data intake reviewer:
Data storage location: https://github.com/DataGlacier/DataSets

Tabular data details:
Table Name : Cab Data

Total number of observations	359,392
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	20.1 MB

Table Name : City.csv

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 bytes

Table Name : 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	1.00 MB

Table Name : 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	8.58 MB

Proposed Approach:

The approach took involves a comprehensive examination of cab usage data to extract insights and confirm underlying patterns that could inform business decisions. The methodology proceeds as follows:

Data Wrangling:

- Cleaning and formatting the dataset, especially ensuring date fields are correctly converted from Excel serial date format to datetime objects in Python for accurate time series analysis.
- Calculating additional fields such as 'Weekday' from the 'Date of Travel' to facilitate granular analysis by day of the week.

Seasonality Analysis:

- Grouping the data by various time frames including month and weekday to analyze patterns.
- Employing visualizations to illustrate trends, such as the number of rides by month for different years and cab usage by day of the week, which can highlight seasonal behavior and weekly preferences.

Statistical Testing:

- Utilizing one-way ANOVA tests to determine if there are statistically significant differences between the means of different groups (such as months or weekdays).
- These tests help to validate whether observed patterns are due to random variation or are indicative of underlying trends.

Visual Exploratory Analysis:

- Creating line plots to observe trends over time, such as total rides and average profit by month across different years, to identify both growth trends and seasonal fluctuations.
- Bar plots and scatter plots are used to evaluate the revenue by company against city user density and to investigate the relationship between user density and revenue.

Hypothesis Formulation:

- Each visualization and statistical test is aligned with a specific hypothesis, such as the impact of user density on company performance, or the effect of seasonality on cab usage and revenue.
- The hypothesis-driven approach allows for a focused analysis where each graph or test is a deliberate step towards confirming or refuting the proposed hypotheses.

Revenue and Profit Analysis:

- Aggregating total revenue by city and by month to assess company performance in different markets.
- Calculating average profit to understand the profitability trends, which may differ from revenue trends.

This structured approach, combining rigorous data preparation, visual exploration, and statistical testing, facilitates a detailed understanding of the dataset. It enables the identification of key factors that influence cab usage and company performance, providing a data-driven basis for strategic business decisions.