Data Intake Report

Project Name: G2M Insight for Cab Investment Firm

Report Date: 05/12/2024

Internship Batch: LISUM33

Version: 1.0

Data Intake by: Marina Tsvetkova

Data Intake Reviewer: Data Glacier

Data Storage Location:
https://github.com/Marinatsv07/Data_Glacier_Internship/tree/main/Week_2

https://github.com/DataGlacier/DataSets

Tabular Data Details

Cab_Data Dataset:

Metric	**Details**
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	**20.2 MB**

City Dataset:

Metric	Details
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**

Customer ID Dataset:

Metric	Details
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 MB**

Transaction ID Dataset:

Metric	Details
Total number of observations	**44098**
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

Unique Row Identification:

- All datasets were merged using Transaction ID to perform comprehensive analysis.
- Each row in the combined dataset was uniquely identified by Transaction ID to analyze individual transactions, facilitating analysis of multiple transactions by the same customer.

Duplicate Rows:

- Utilized dataset.drop_duplicates() to eliminate any duplicate rows across all datasets.
- Applied dataset.dropna() to remove rows with missing (N/A) values.

Dataset Understanding:

- The merged dataset includes columns such as Transaction ID, Customer ID, Payment Mode, Date of Travel, Company, City, KM Travelled, Price Charged, Cost of Trip, Gender, Age, Income (USD/Month), Population, Users, Year, Month, Day, Avg_Profit_Per_KM, Profit_Percentage, Quarter.
- Additional columns created include Profit (Price Charged - Cost of Trip), Margin (Profit / Cost of Trip), and Number of Rides.

Assumptions:

- Analyses assumed external noise beyond the provided data.
- Data timeframe is constrained between 2016 and 2018.
- Datasets are assumed to be randomly selected.
- Only cash and card payment methods are considered.
- Profit calculation is based solely on Price Charged minus Cost of Trip.

Analysis Summary

1. Seasonality in Cab Usage:
    - Finding: Significant seasonal trends were observed, indicating variability in customer numbers across different months.
2. Company Preference by Time Period:
    - Finding: Significant differences in daily user counts between Yellow Cab and Pink Cab.
3. Relationship Between Margin and Number of Customers:
    - Finding: A modest positive correlation was found, suggesting that margins slightly increase with a higher number of customers.
4. Customer Attributes (Age):
    - Finding: No significant difference in age distribution between customers of Yellow Cab and Pink Cab.
5. Customer Attributes (Income):
    - Finding: No significant difference in income levels between customers of Yellow Cab and Pink Cab.
6. Customer Attributes (Gender):
    - Finding: A significant difference in gender distribution, indicating distinct gender preferences for each company.

Recommendations

- Seasonality Strategy: Adjust marketing campaigns and resource allocation to align with seasonal demand patterns.
- Peak Time Promotions: Develop promotions tailored to peak usage times specific to each company.
- Gender-Specific Marketing: Implement targeted marketing strategies to appeal to the dominant gender segment for each cab company.
- Service Customization: Customize services to better meet the needs and preferences of identified customer segments.