Data Intake Report

1. Information

Name: ***G2M insight for Cab Investment firm***
Report date: 14th May 2022
Internship Batch: LISUM09
Version: 1.0
Data intake by: Huu Thien Nguyen
Data intake reviewer:
Data storage location:

2. Tabular data details:

2.1. Cab data

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	21,2MB

2.2. Transaction data

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	9MB

2.3. Customer data

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	1,1MB

2.4. City data

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	4KB

3. Proposed Approach

The initial steps of investigation of the dataset will be perform in week 2. It covers a general view of data instead of compare between the two company Yellow Cab and Pink Cab. The exploratory provided several critical insights for both cab business. Additionally, week 3 will provide full hypothesizes and recommendation for the board of the director to make a decision which company to invest.

3.1. Data preprocessing

Since the data after join will be difficult to feature engineer, therefore, feature engineering step will be performed before consolidating the data source.

Firstly, the Date of Travel in the Cab file was converted to “Date of Week” feature with date type instead of numeric in raw source. The “Profit” and “Profit/km” features were generated using (['Price Charged'] - ['Cost of Trip']) / [‘KM Travelled’]. It helped reducing the amount of column needed for the exploratory step. Secondly, “Age Bracket” was created from [‘Age’] of Customer data using binning technique. Similarly, “Income Range” was utilized from [‘Income’]. Thirdly, the [‘Population’] and [‘User’] was combined into “User/pop rate %” in the feature engineering step. Details of the implementation is fully explained in the coding notebook.

3.2. Data consolidating

After preprocessed the raw source, all 4 files are consolidated into a single dataset. Basically, the dataset was joined together using same common column. The “Transaction ID”, “Customer ID”, “City” columns are used as the joined variables.

3.3. Data exploratory

The exploration was divided into 2 main sections, univariate and bivariate analysis. For the former part, pandas profiling was utilized to automate the simple investigation. For the latter, python function was implemented to compare categorical vs categorical, and categorical vs numeric type. Several insights were found after the examination.

4. Insights/Use case

There are in total of 6 interesting insights was found. This is a general important insight after I made a simple throughout exploratory of the dataset. More hypothesis, forecast related to Profit and Date of Travel will be provided as well as comparison between 2 companies to help BOM decide which to invest in Week 3.

- Insight#1: Weekends (Fri, Sat, Sun) accounted for 64% of the transaction. Businesses should prepare more workforce and resources on the weekend occasion
- Insight#2: In a big city the user of cab companies is higher than the city which has low population
- Insight#3: The majority of the transaction is from Yellow Cab from 3 cities: New York (24%), Chicago (13%), Washington DC (11%)
- Insight#4: The majority of the transaction is at weekends (Fri, Sat, Sun) for both companies.
- Insight#5: The Yellow Cab made more Profit/km than the Pink Cab
- Insight#6: New York has the highest Profit/km out of any city (mean > 10$)