Data Intake Report

Name: G2M insight for Cab investment
Report date: 08/01/2022
Internship Batch: LISUM05
Version:<1.0>
Data intake by: Arnold Amusengeri
Data intake reviewer: N/A
Data storage location: https://github.com/ahnoamu/DataSets

Tabular data details:

Total number of observations	Number of rows including header Cab_data.csv - 359393 rows City.csv - 21 rows Customer_ID – 49172 rows Transaction_ID – 440099
Total number of files	4 files
Total number of features	Number of columns Cab_data.csv - 7 cols City.csv - 3 cols Customer_ID – 4 cols Transaction_ID – 3 cols
Base format of the file	.csv
Size of the data	Cab_data.csv - 21Mb City.csv - 759 bytes Customer_ID – 1.1 Mb Transaction_ID – 8.6 Mb

Note: Replicate same table with file name if you have more than one file.

Proposed Approach:

- Mention approach of dedup validation (identification)

Data exploration
-Total features - 17
-Time frame - 2016-2018
-Total number of observations - 848681

- Mention your assumptions (if you assume any other thing for data quality analysis)

1. Transaction IDs are unique, Customer IDs can be repetitive
2. The price charged data is normally distributed. i.e Outliers are present in Price_Charged feature, but due to unavailability of trip duration details, we are not treating any value as outlier.

3. Keeping other factors constant, only the Price_Charged and Cost_of_Trip features are used to calculate the profit of rides.
4. Users feature of city dataset is treated to be the total number of cab users in the city, this is assumed to be all other cabs in addition to Yellow and Pink cab users.

Hypotheses

- Pink cab generates more profit in general and per ride
- Pink cab contributes larger values therefore do better business.
- More males than females use Pink Cabs as their mode of transport
- Females largely contribute to profits in both companies
- Medium class of income earners use pink cab more than low-income earners
- Middle income class (25000-35000 usd/month) contribute larger values to profit
- Age 15-25 are frequent customers for pink cab
- Age 35-45 contribute a larger amount to the profit
- Pink Cab has a higher customer retention
- Pink cab covers a larger number of cab users in Los Angeles
- Pink cabs contribute the largest percentage of total population of cab users in Washington DC
- Pink cab has a wider reach to customers compared to yellow cab
- Demand is high on 3rd day of month
- Demand decreases exponentially towards the end of the year
- Profit is high on 3rd quarter of the year
- Pink cab records higher number of customers on Friday, Saturday and Sunday
- Pink cab records larger number of transactions on Friday, Saturday and Sunday
- Pink customers ride for short distances. Therefore, the company is offering a better plan for short trips
- Yellow Cab attracts more customers on holidays the Pink Cab
- Pink Cab profit is likely to decrease in the year 2019
- Pink Cab rides are likely to decrease in the year 2019

Analysis

Profit Analysis

- Average profit per KM (for each cab in each year)
- Percentage profit per annum (for each cab in each year)
- Profitable rides per city (percentage profit for each cab in each city all years)

Profit trend for each cab (from one year to the next)

Customer base gender-wise (for each cab in each year)

Profit contribution per gender of customers (for each cab in each year)

Customer base class-wise (for each cab in each year)

Profit contribution class-wise (considering customer income)

Customer base per age group

Profit contribution age-group-wise

Customer retention

- Customers who have taken at least 5 rides of the same cab
- Customers who have taken at least 10 rides of the same cab

Comparison of cab preference by users city-wise throughout the period

City-wise cab users covered by each company as a percentage of the total population of cab users.

Cab user numbers covered by each company as a percentage considering their sum across all 19 Cities (National coverage) per year.

Seasonality in demand:

- Daily
- Monthly

Seasonality in profit (quarter cycle)

Day-wise customer (and ride) preference for each cab

- Number of customers
- Number of rides

Customer analysis based on ride distance (Distance (Km travelled) VS number of customers)

Customer preference on holiday

Forecasting for 2019

- Profit
- Number of rides

Recommendations