G2M - Data Insight Report on Cab Industry Investment

Abstract

There are many cab industries across the world. This particular paper will focus on the US cab industry. A private equity firm named XYZ has an interest in cab industry and wants to invest in the market as it has seen a remarkable growth and wants some insight and recommendations for the purpose of decision making for investment. This paper focuses on the data provided, data analysis using python, and gives different insights which could be considered with high executives who will be making decisions.

1. Business Problem

A private United States based firm named XYZ, has an eye for the cab investment industry to make an investment due to understandable and detectable growth in the industry. Their business strategy is a G2M abbreviation for Go-to-market strategy which is basically a business plan to bring a new product or service into the industry. "For the sole reason of investment, they want an insight into the market so that they can make a final strategic decision."

2. Objective

"Providing actionable insights and recommendations using data analysis techniques and coming up with simple, effective and easily understandable visualizations which can help the stakeholders in decision making."

3. Properties of Data (Data Intake Report)

There are 2 cab companies known as yellow cab and pink cab where the stakeholders can decide which cab company, they can invest in. Below mentioned is a detailed dataset information for better understanding of the data which answers what data, how much data, and what type of data for all 4 datasets. In total there are 4 types of data which includes cab company dataset, Customer Dataset, Transaction dataset, and city dataset.

3.1 Cab Data CSV Details

Total Number of Observations	359393
Total Number of Files	4
Total number of Features	7 Columns
Base format of File	CSV
Size of Data	20.1 Megabytes

3.2 Customer Data CSV Details

Total Number of Observations	29290
Total Number of Files	4
Total number of Features	4 Columns
Base format of File	CSV
Size of Data	1 Megabytes

3.3 Transaction Data CSV Details

Total Number of Observations	440099
Total Number of Files	4
Total number of Features	3 Columns
Base format of File	CSV
Size of Data	8.58 Megabytes

3.4 City Data CSV Details

Total Number of Observations	21
Total Number of Files	4
Total number of Features	3 Columns
Base format of File	CSV
Size of Data	1 Kilobyte

4. Steps to create Applicable final dataset

As the final dataset has to have data points which can be moulded however you want but still give the needful result, there are some points I noted using pen and paper at first even before starting the analysis on the datasets.

As some papers showed before investing in any industry there are 5 factors that should be considered:

1. Growth of the industry and Value
2. Market Capitalization
3. Credit Rating
4. Stock Price Volatility
5. Performance over years

These points if seen from an analytics point of view, they have a lot of sub classes in them which can show great insights at great lengths. I derived the points, related to these as value, as probability of questions to answer to create a dataset with respect to Profit, Customers, Transactions, City, and Company. After good amount of time of permutation and combinations with the data, it was seen that value and growth could be portrayed using Profit which was divided into 5 different parts such as City wise profit, Yearly Profit, Gender wise profit, and Age group wise profit. With respect to the customers the Customer age group was one of the factors and gender of the customers. This could give the demand for the company as to who and what age group are most likely to pay or get a cab. In Transactions, insights were needed for Company wise, city wise and customers age group wise which could give the market capitalization and the performance over the years for the company. City was also one of the main factors as City wise yearly profit, top 5 cities who have most sales regarding cab, and most transactions could also give the market ratio of profit, growth and value.

Then I started using the python library known as Pandas for reading, and manipulation where for all the datasets there were some common fields which I checked such as null values, duplicate values, and unique values. After understanding each dataset, I started to analyse what fields will be needed to be added to give the answers to the above questions. Also, there were some columns which needed to be changed for example, the date of travel was dated from 2016 to 2018 but was in dtype[n64] format which had to be transformed into the datetime format. Then there were some columns which were needed but not in one dataset which I started to join from other datasets. After that was completed, I got a clear idea of which Columns or fields to merge of datasets to create the final dataset.

4.1 Final dataset details

Total Number of Observations	359392
Total Number of Files	4
Total number of Features	17 Columns
Base format of File	CSV
Size of Data	38.9 Megabytes

5. How analysis was prepared and performed

Analysis was prepared using 5 parts:

1. Creating rough hypothesis
    a. This phase involved reading papers, watching videos, understanding major factors before investing in any market. The questions which I came up with and the major factors are:
        i. Growth and Value
            1. What is the analysis after calculating the profit with respect to city, year, gender, and age group?
            2. What amount of company have most customers?
            3. Which city has most amount of cabs – yellow or pink? Which are the top 5 cities contributing the most?

ii. Market Capitalization
    1. What amount of people have taken Yellow Cab and what amount have used Pink Cab?
    2. Which city has the highest transaction rate?
iii. Credit rating
    1. What amount of cab's profit do customers contribute with respect to age group?
iv. Stock Price Volatility
v. Performance over the years
    1. Which company has good performance over the years?

2. Data Understanding & Exploring Data
    a. The understanding of data was divided into four parts again:
        i. Cab dataset – Data was explored using the count for every column, changing the date from int64 to datetime format in D/M/Y format, looking for unique cities present, checking for the null values where no null value was present, adding columns such as day, minute, and year from date of travel column, after which the dataset was clear to be merged with others.
        ii. Customer Dataset – For this dataset I checked for null values, info about the datatypes and count of each column.
        iii. Transaction Dataset – Checking for how many payment modes are present using unique() method, then count, and checking null values.
        iv. City Dataset – Checking unique cities and also looking if the city dataset and cab dataset have equal count of unique cities.

3. Preparing Hypothesis
    a. Using the rough hypothesis questions at the starting and exploring the data more after exploring, all the questions in the rough hypothesis were used which could give accurate and effective recommendations and insights.

4. Answering hypothesis
    a. All the questions were written in markdown cells and after the final dataset was created the questions were answered mostly using seaborn library of python using different charts.

5. Recommendations
    a. Based on the analysis the recommendation was the yellow cab is good to invest in as all the five main points checked for yellow cab compared to pink cab.

6. Type of Analysis Performed

There are lot of analysis methods which can be used with data, in this process, the insight rely upon the sentiment analysis and regression analysis.

Mostly regression analysis is used as the data insights used, are analysed using relationships with one dependent and one independent data. There are a total of 7 different charts used such as, bar plot, dist plot, scatter plot, pie plot, joint plot, Heatmap and box plot. These charts have specific information to

show and these charts are used with specific type of datatypes or in simple terms when there is distribution, or Categories.

1. Bar plot – This chart is used to show categorical information.
    a. KM Travelled
    b. Price Charged
    c. Payments Count plot
    d. Gender Count
    e. City Wise Profit
    f. Gender wise profit
    g. Age group wise profit
    h. Users with respect to years and company
    i. Price Charged with respect to company
    j. Price Charged with respect to city
    k. Price Charged with respect to Age

2. Scatter plot – This chart is used to determine the relationship or correlation.
    a. Transactions with respect to Customer Age group
    b. KM travelled with US Holidays

3. Joint plot – This graph is used when you want to show univariate or bivariate graphs which is actually an interface for Joint Grid class.
    a. Company and Profit Plot
    b. KM travelled with US Holidays

4. Dist plot – Also known as Distribution plot is used to represent the histogram with univariate set of values.
    a. Profit with and without the KDE

5. Pie Plot – This is a categorical chart used to show different categories with their percent value.
    a. Customer with respect to company

6. Box plot – This plot shows the distribution where the values are numerical, and outliers can be detected using this plot.
    a. Year Wise profit

7. Heatmap – This is a map of all the fields which is used to show the correlation between them.
    a. Dataset Correlation

7. Results

The results seen showed that in most of the factors where the analysis was carried out yellow cab was the better one compared to pink cab.

1. Profit
    a. City wise profit – In 20 cities yellow cab has more profit compared to pink cab.
    b. Year wise profit – In 2016, 2017, 2018 yellow cab has managed more profit compared to pink cab.
    c. Gender wise profit – In both male and females it is seen that yellow cab is more preferred by both the genders.

   d. Age wise profit – Here as well profit margin of yellow cab with respect the age is way higher than pink cab.

As per the analysis done above it can be safely said that yellow cab has more preference and profit with respect to above 4 factors.

2. Users or Customers – It was seen that out of all the customers 18 percent of the users preferred pink cab and 81.3% of users preferred yellow cab. Also, if it is seen for all the 3 years from 2016 to 2018 yellow cab has a higher margin than pink cab with respect to users.

3. KM travelled on Holidays – It is seen that there are yellow cabs taken more on holidays as well compared to pink cabs.

4. Transactions

   a. Price charged with respect to company over the years – It is seen for all three years that the transactions happened with respect to cabs, yellow cab has more transactions.
   b. Price charged city wise – The transactions seen in all 20 cities have yellow cab on the higher ratio of transactions compared to the pink cab.
   c. Price charged age group wise – Age groups also prefer the yellow cab as it has more transactions.

On the above characteristics these are the recommendation based on the:

1. Profit
2. Transactions
3. Customers

The yellow cab is recommended for investment as it has growth and value, good market capitalization, good credit rating, and good performance over 3 years.
Stock volatility cannot be told unless there was data on any cab industry considered.

8. **Failed Experiments**

1. Holiday Dataset – The dataset could not be properly merged with the final dataset due some issue but if it could have been merged there could be more definitive insights for sure.

9. Future Enhancements

The stock data of any company can be assumed and on that stock data more insights can be derived such as risk to reward ratio, Price to sales ratio, price to earnings ratio and so on. This can definitely be done and can be effective as well.

Links

GitHub Link - https://github.com/TanmayPotbhare/G2M-Insight-for-Cab-Industry-Investment

(This is where the dataset, intake report and the ipynb file for this insight is)