Exploratory Data Analysis

G2M Insight Cab Investment Firm

21/03/24

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Executive Summary

The Client

XYZ is a private firm in US. Due to remarkable growth in the Cab Industry in last few years and multiple key players in the market, it is planning for an investment in Cab industry and as per their Go-to Market(G2M) strategy they want to understand the market before taking final decision.

The Dataset

Below are the list of datasets which are provided for the analysis:

1. **Cab_Data.csv :** this file includes details of transaction for 2 cab companies
2. **Customer_ID.csv :** this is a mapping table that contains a unique identifier which links the customer's demographic details
3. **Transaction_ID.csv :** this is a mapping table that contains transaction to customer mapping and payment mode
4. **City.csv :** this file contains list of US cities, their population and number of cab users

Problem Statement and Approach

Objective

Provide insights to help XYZ firm to choose the right company for making an investment.

Data Analysis

The analysis has been done in the following parts:

1. Data visualizations and summary
2. Data analysis on the basis of place, location, gender, age and profit.

Assumptions

It was assumed that Date of Travel data was consistently recorded for trend analysis, formats were standardized or could be standardized without information loss and Transaction ID and Customer ID were correctly assigned and unique.

It was also assumed that the transactional data (e.g., KM Travelled, Price Charged, Cost of Trip) were accurately recorded at the source.

Data Summary

- There are no duplicate values and no NA values in the dataset. All four datasets have been merged into one such that the resulting data frame will contain only the rows where the merge key exists in all data frames.
- When merging datasets, they were checked for inconsistencies that could introduce duplicates, such as differing formats or multiple entries for a single identifier across tables. For instance, a single Transaction ID should not be linked to multiple Customer IDs unless the business logic explicitly allows for this scenario.
- Next, it was ensured that supposed unique identifiers, such as Transaction ID and Customer ID, do not have duplicates within their respective datasets. This can be achieved using aggregation methods to count occurrences of each identifier and flagging any counts greater than one.
- Time period of dataset is from 2016 to 2018.

EDA: KM Travelled

Conclusion: As it can bee seen from Fig 1, most rides are from 2 to 48 Km.

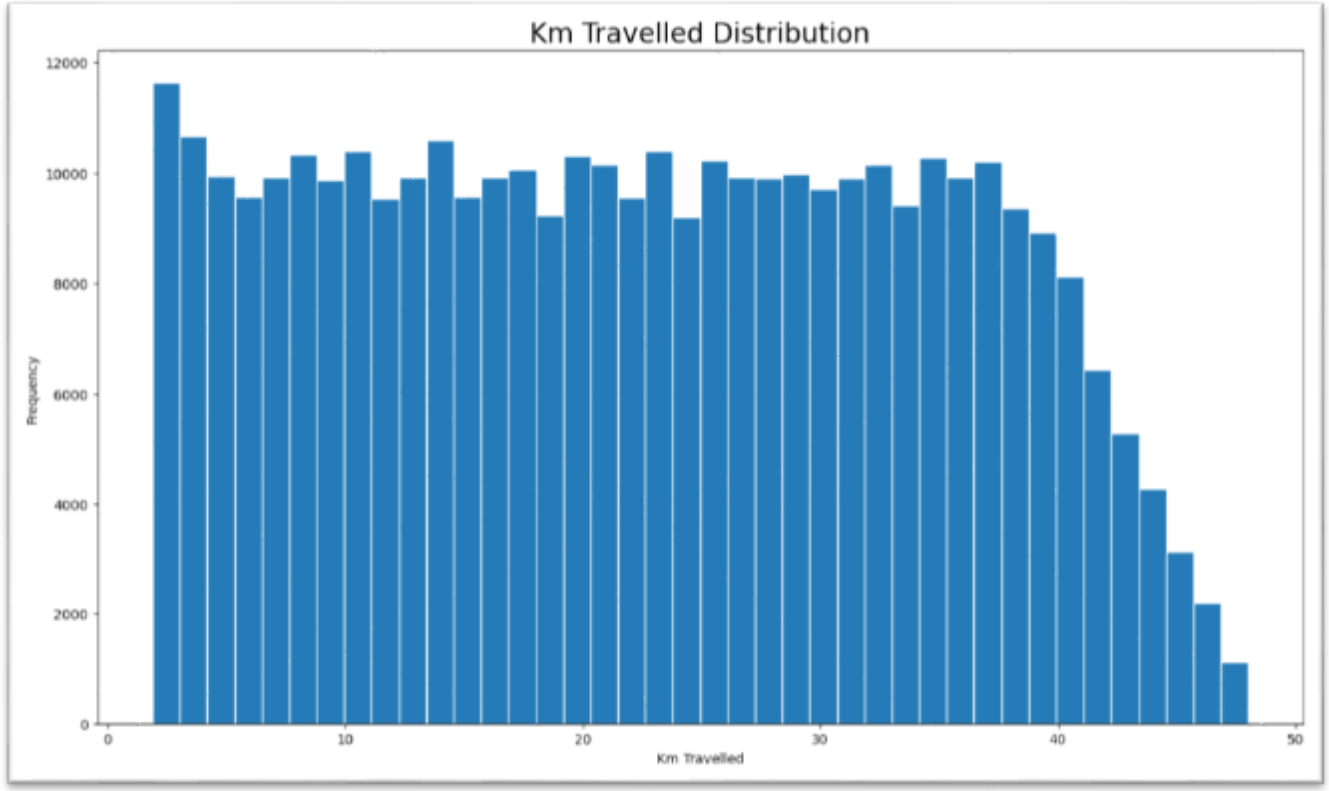

Fig 1: Km travelled by customers of both the companies.

EDA: Payment Mode

Conclusion: From Fig 2, we can see that, users prefer card as mode of payment rather than cash. In both companies, card was preferred more as compared to cash.

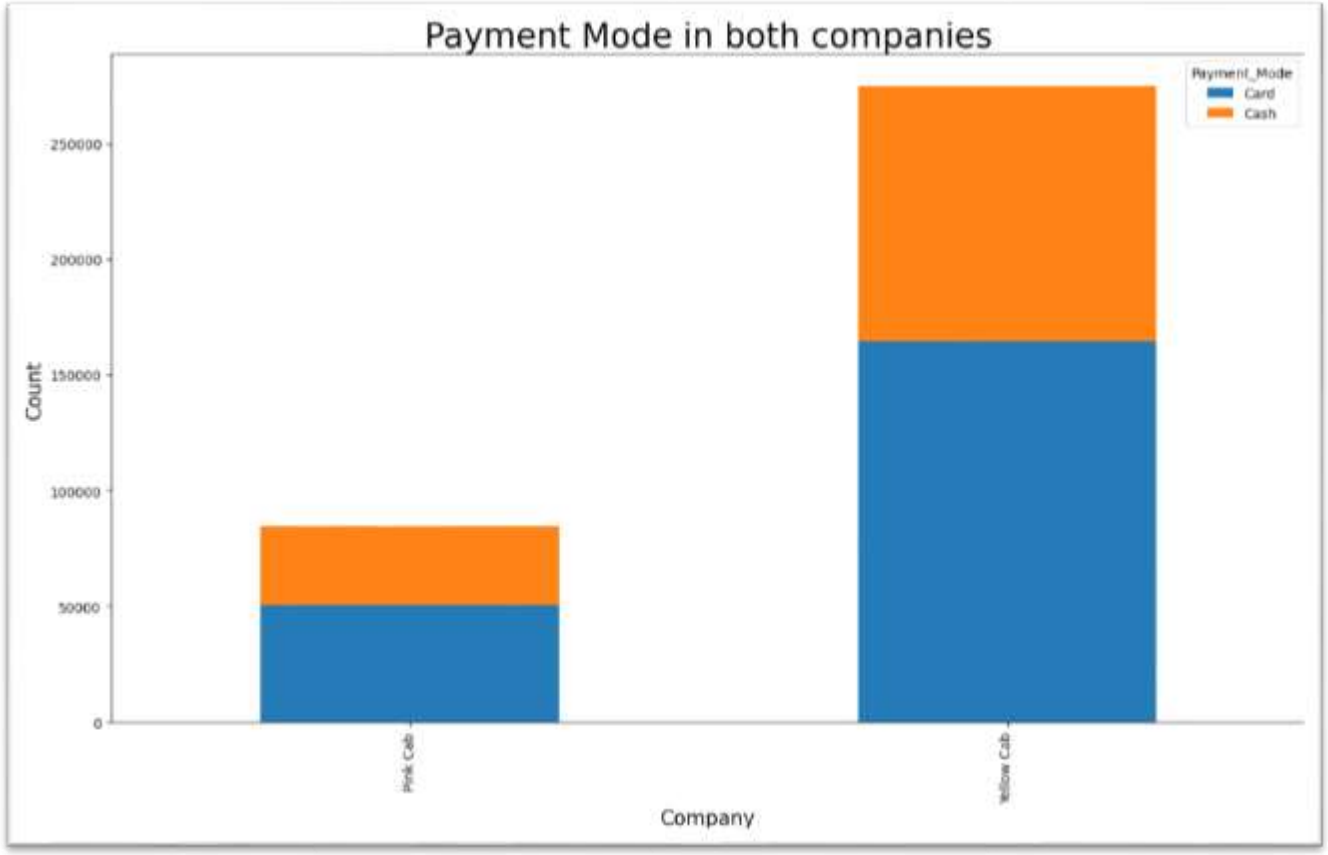

Fig 2: Preferred payment mode in both companies

EDA: Yearly Profit

Conclusion: From Fig 3, it can be seen the yearly profit by both companies over the years. Pink Cab had consistently low profit as compared to yellow cab, over the years.

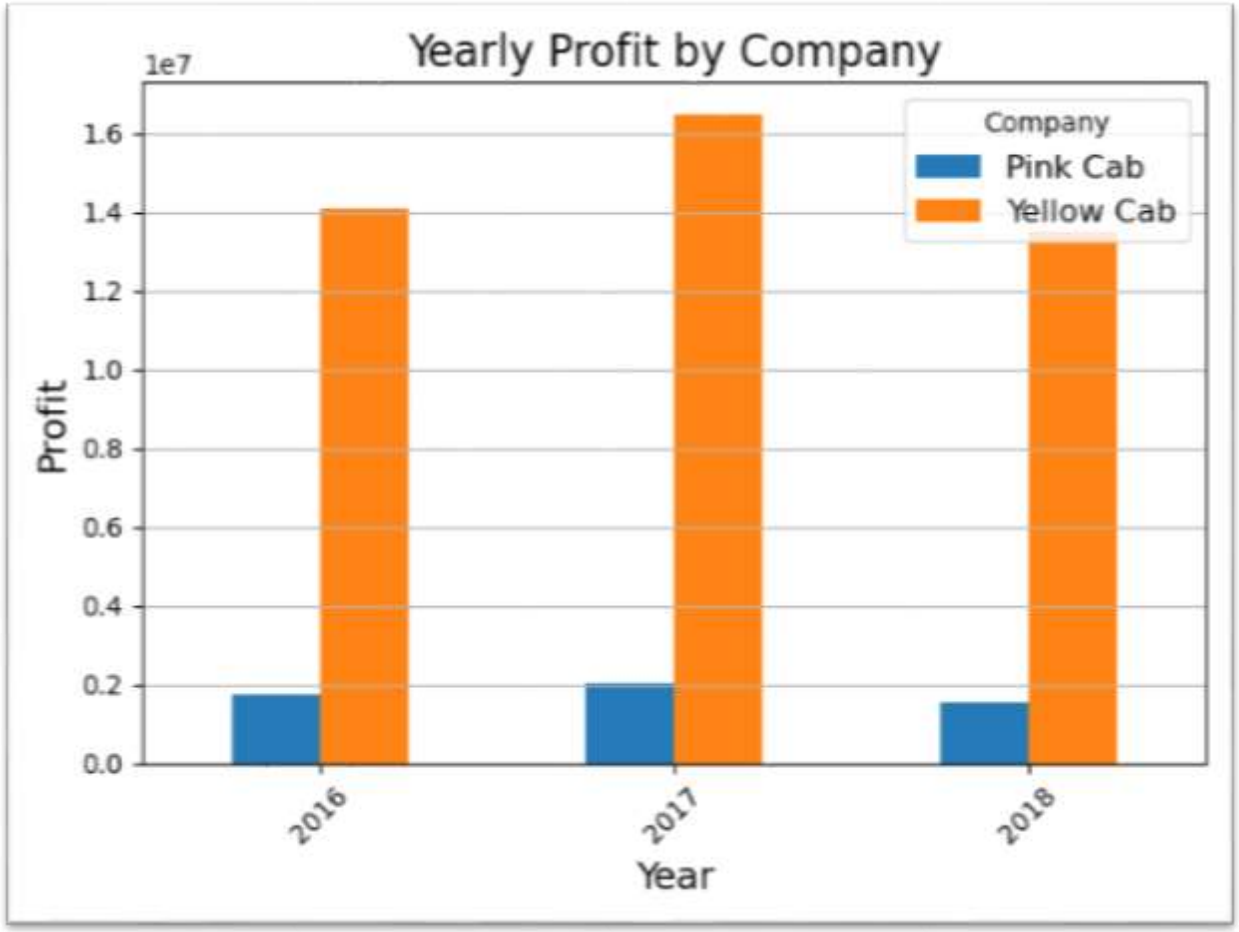

Fig 3: Yearly Profit by both company

EDA: Gender

Conclusion: From Fig 4, Both male and female prefer Yellow cab more by approximately 5%

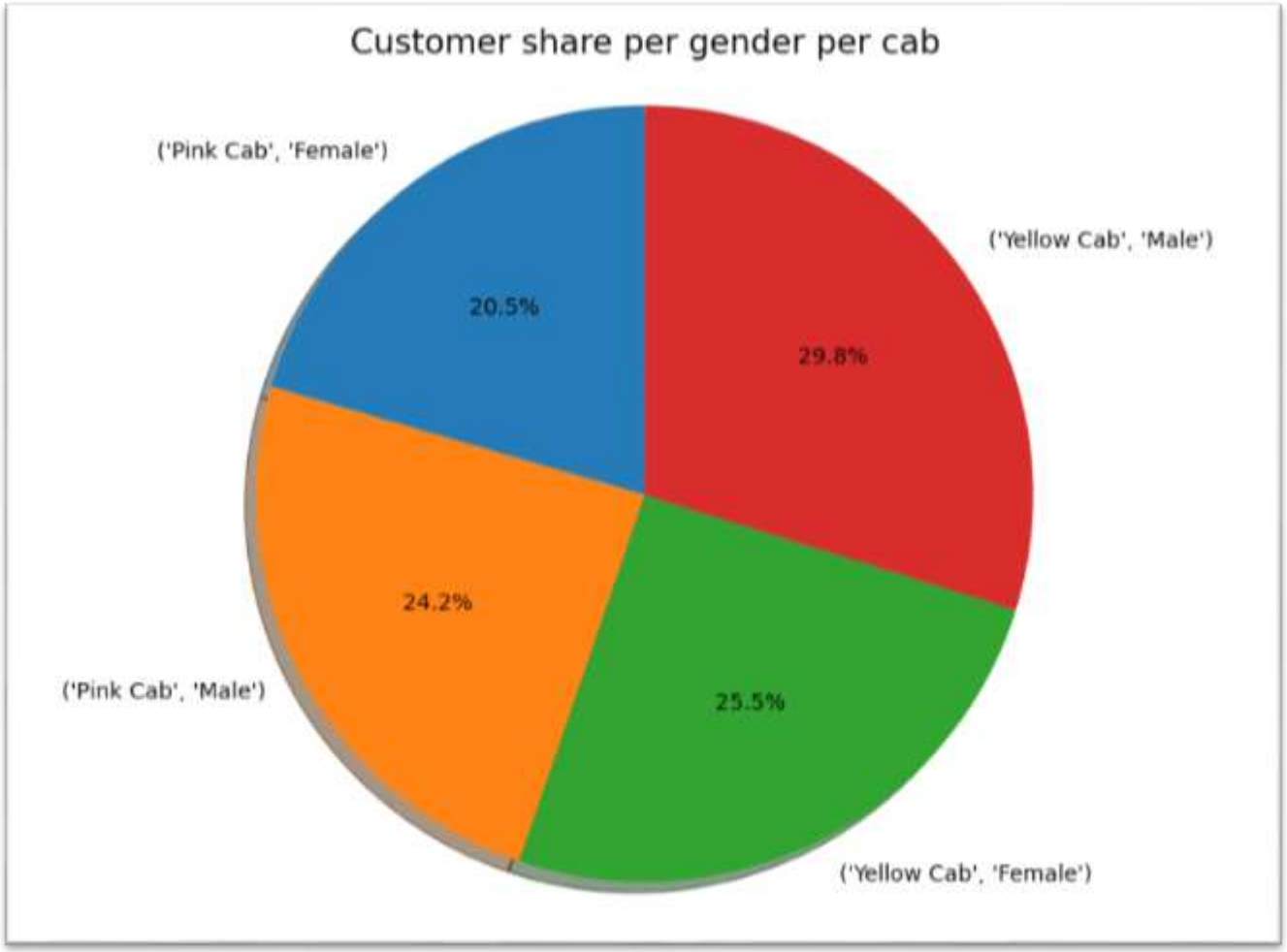

Fig 4: Customer share per gender per cab

EDA: Profit % per month

Conclusion: From Fig 5, yellow cab had significantly more profit percentage all years.

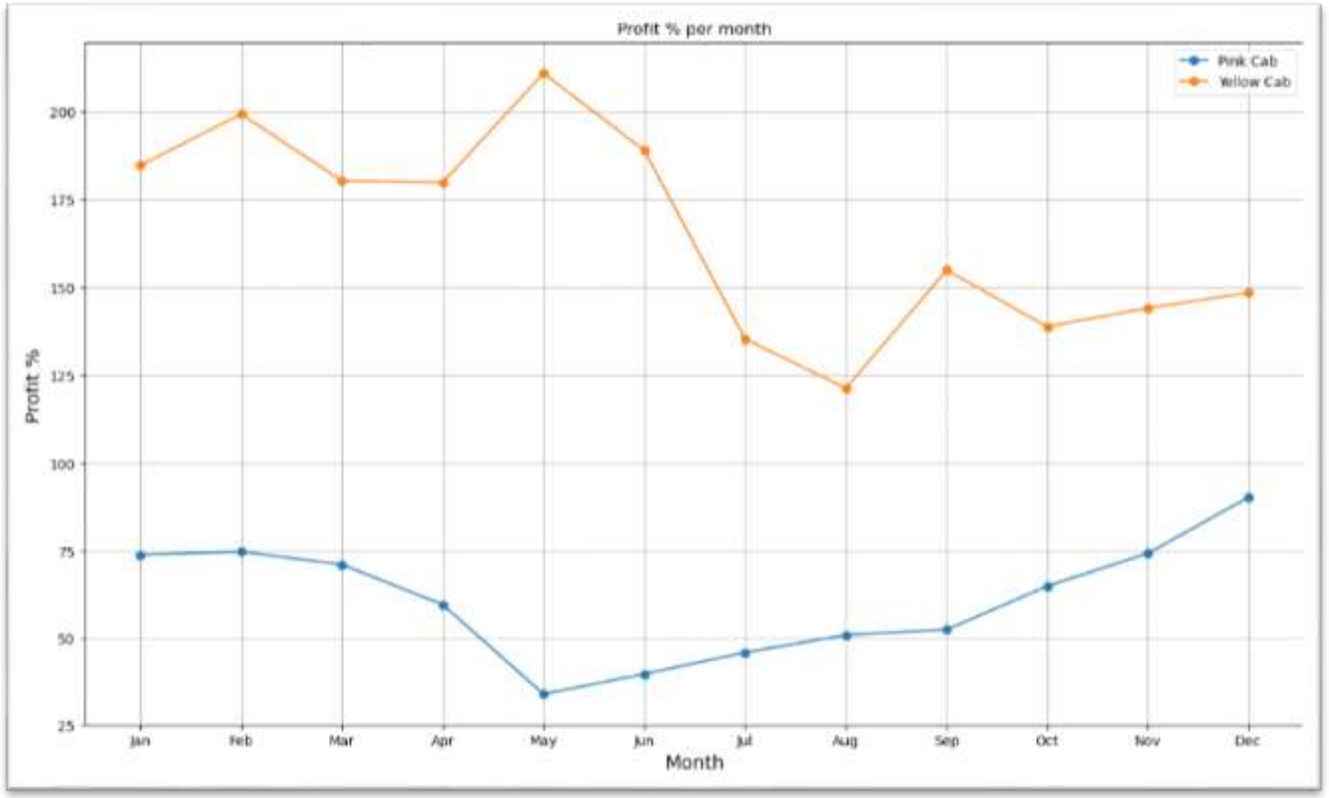

Fig 5: Profit % per month

Hypothesis

1.1 Gender plays a role in profit for pink cab

Ho : There is no difference in profit in terms of gender for pink cab

H1 : There is difference in profit in terms of gender for pink cab

Results: Pink Cab: p-value: 0.11515305900425798

We accept null hypothesis (H0) that there is no difference in terms of gender for pink cab

1.2 Gender plays a role in profit for yellow cab

Ho : There is no difference in profit in terms of gender for yellow cab

H1 : There is difference in profit in terms of gender for yellow cab

Results: Yellow Cab: p-value: 6.060473042494056e-25

We accept alternate hypothesis (H1) that there is a difference in terms of gender for yellow cab

2. The number of cab rides is positively correlated with the population of the city.

Results: Correlation between City Population and Number of Rides: 0.9039264586442209

3.1 Age plays a role in profit for pink cab

Ho : There is no difference in profit in terms of age for pink cab

H1 : There is difference in profit in terms of age for pink cab

Results: p-value for Pink Cab: 0.8618554671919694

We accept null hypothesis (H0) that there is no difference in terms of age for pink cab.

3.2 Age plays a role in profit for pink cab

Ho : There is no difference in profit in terms of age for pink cab

H1 : There is difference in profit in terms of age for pink cab

Results: p-value for Yellow Cab: 1.4177354970217492e-95

We accept alternate hypothesis (H1) that there is a difference in terms of gender for pink cab

4. There is a significant difference in the cost of trips between customers who choose Pink Cab and those who choose Yellow Cab.

Results: Average Cost of Trip: Yellow Cab: 297.92 , Average Cost of Trip: Pink Cab: 248.15

Recommendations

From above analysis and hypothesis testing, it is recommended that XYZ firm should invest in Yellow Cab.

Thank You