



Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

DAVID PAUL NALUMENYA

G2M insight for Cab Investment firm

25th/02/2024

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Executive summary

- XYZ, a private firm in the US, is considering an investment in the Cab industry, which has experienced remarkable growth in recent years with the presence of multiple key players in the market. In line with their Go-to-Market (G2M) strategy, XYZ aims to gain a comprehensive understanding of the market landscape before finalizing their investment decision.
- To achieve this, XYZ has been provided with multiple data sets containing information on two cab companies, each representing different aspects of the customer profile. The objective is to leverage actionable insights from the data to identify the right company for investment that aligns with their strategic goals.
- By analyzing the customer profile data sets and extracting valuable insights, XYZ can make an informed decision on which cab company presents the most promising investment opportunity. This executive summary highlights the importance of data-driven decision-making in selecting the right investment target within a competitive market environment.

Problem statement

XYZ, a private firm in the US, is seeking to make an investment in the Cab industry amidst significant growth and competition. They need to analyze the customer profile data sets of two cab companies to determine which company aligns best with their investment goals and market strategy. The challenge is to extract actionable insights from the data to guide their decision-making process and select the most promising investment opportunity.

Exploratory data analysis approach

For exploratory data analysis (EDA) on the provided data sets, the following approach will be considered:

1. Data Integration:

- Merge the Cab_Data.csv with the Transaction_ID.csv and Customer_ID.csv with the City.csv based on unique identifiers to create a comprehensive dataset that includes transaction details, customer demographics, and payment information.

2. Data Cleaning:

- Check for missing values, outliers, and inconsistencies in the merged dataset.
- Standardize data formats and address any data quality issues to ensure accuracy in the analysis.

3. Descriptive Analysis:

- Calculate summary statistics for key variables such as transaction amounts, customer demographics, and payment modes to understand the distribution and trends in the data.
- Explore the temporal trends in transaction volumes and customer behavior over the time period from 31/01/2016 to 31/12/2018.

4. Customer Profiling:

- Segment customers based on demographic characteristics such as age group, gender, and transaction history to identify patterns and preferences.
- Analyze customer behavior, frequency of transactions, and payment modes to gain insights into customer segments.

5. Market Analysis:

- Evaluate the performance of the two cab companies based on transaction data, market share, and customer retention rates.
- Use the City.csv data to analyze the distribution of cab users across different US cities and identify potential market opportunities.

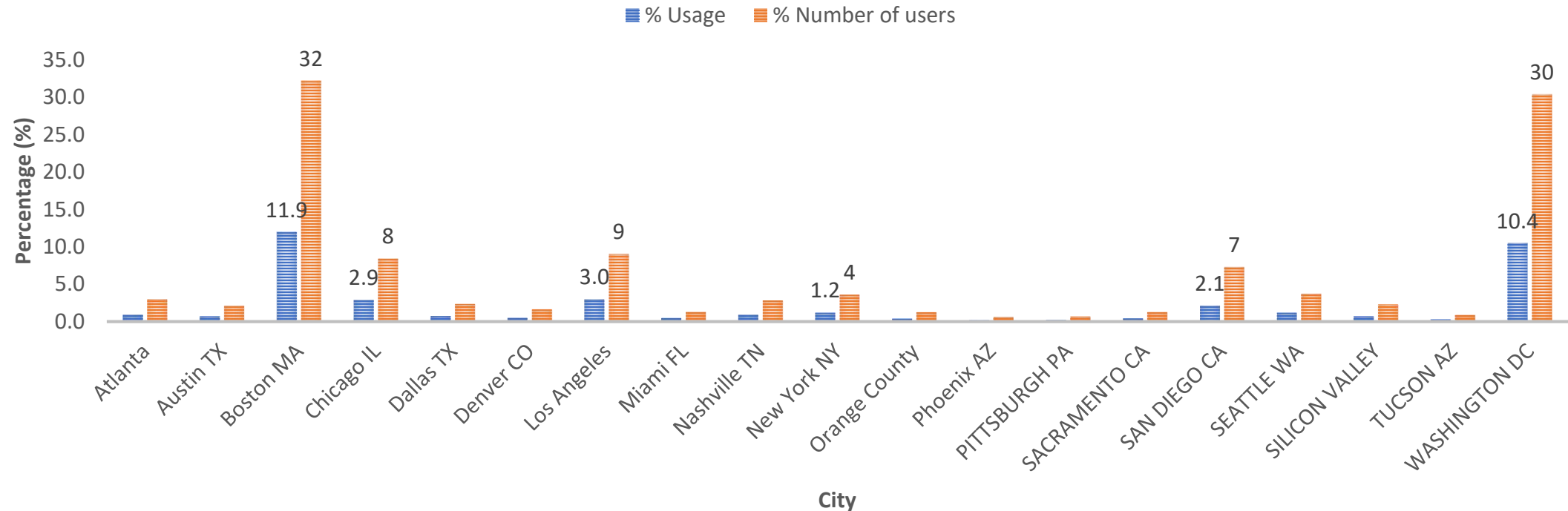
6. Data Visualization:

- Create visualizations such as bar charts, pie charts, and line graphs to represent key findings and trends in the data.
- Use geographical maps to visualize the concentration of cab users in different cities and assess market penetration.

7. Correlation Analysis:

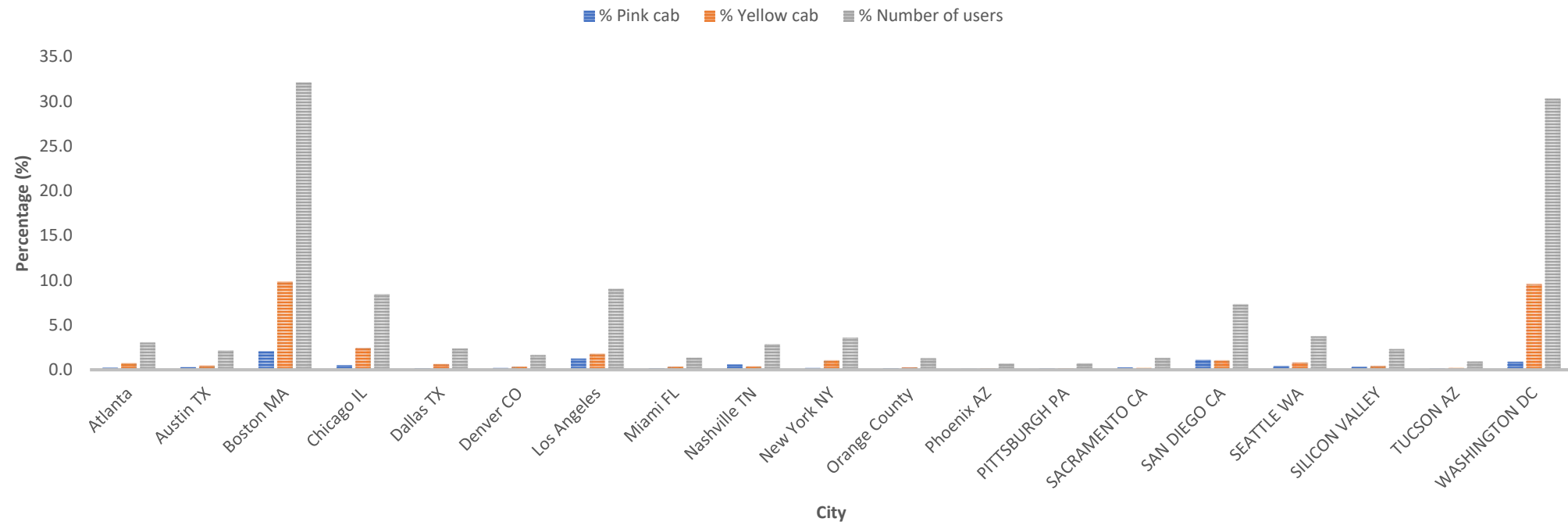
- Conduct correlation analysis to identify relationships between variables and determine factors that influence customer preferences, transaction behavior, and market dynamics.

Question 1: what is the percentage usage of cabs in the reported cities with respect to the percentage number of recorded users



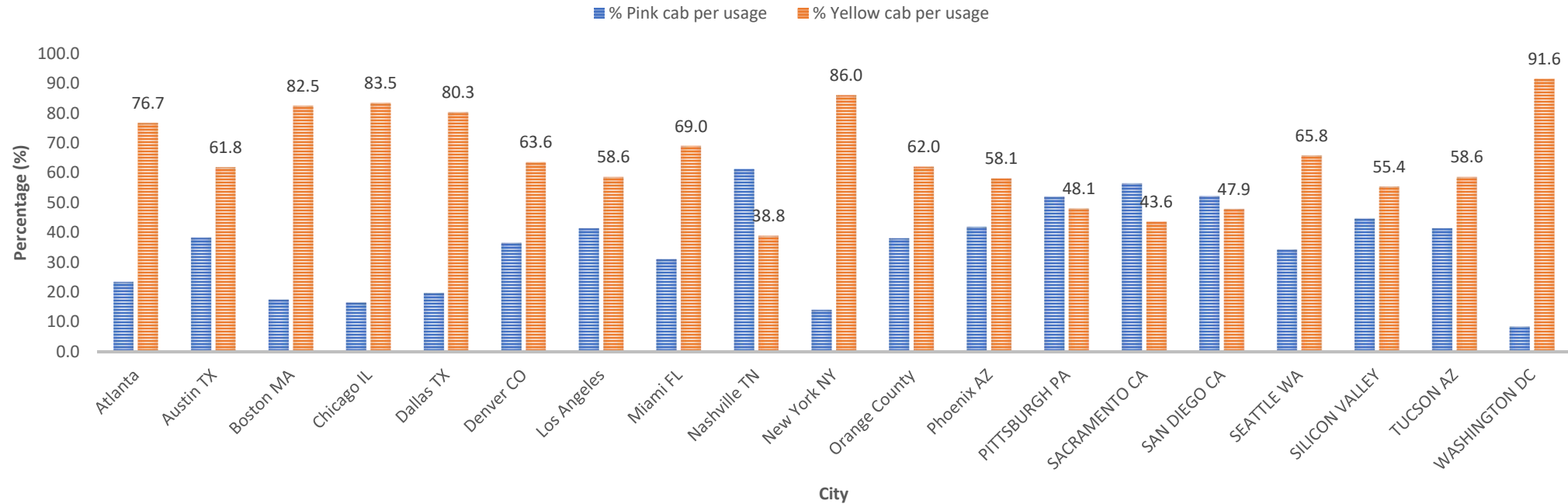
- Los Angeles, Chicago IL , Washington DC , Boston MA and San Diego CA have the highest percentage of users per the total population of each city, reported to use cabs.
- There is an approximately 30% usage of the cabs per the population in each of all the cities.

Question 2: What is the percentage usage of the yellow and pink cab in the different cities with respect to the percentage number of users



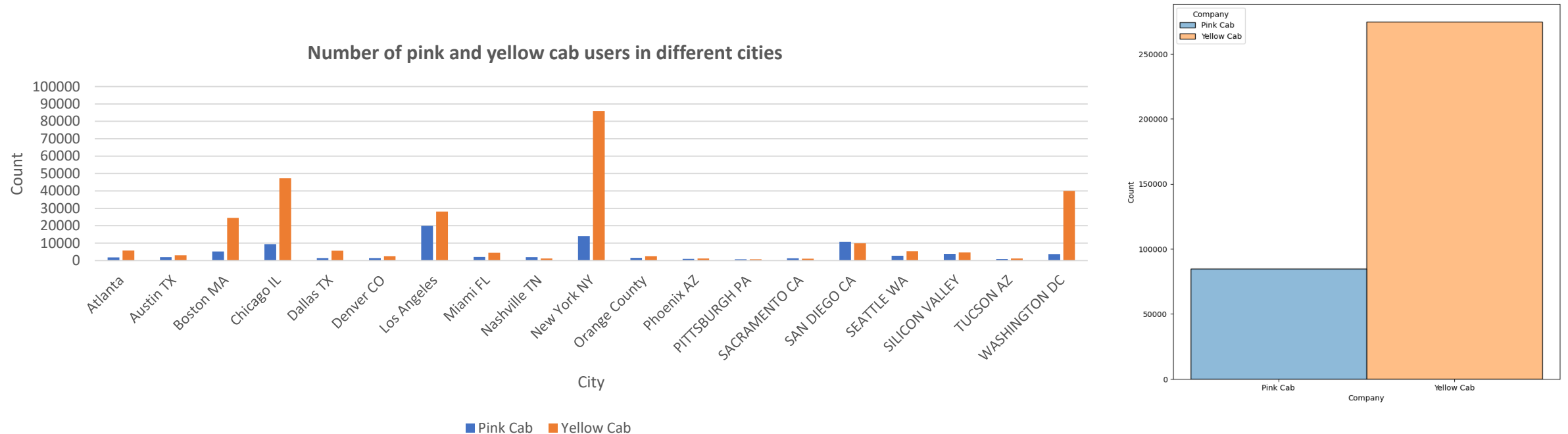
- The Yellow cab company predominated and had the highest percentage of users in all these cities except Nashville TN, Phoenix AZ, Pittsburgh PA, Sacramento CA, San Diego CA.
- San Diego CA has one of the highest number of users however has low percentage usage of the yellow cab.

Question 3: What is the percentage usage of the yellow and pink cab in the different cities with respect to the reported percentage usage from each different city



- On average both the pink and yellow cab companies are used concurrently in most cities in the USA, however the yellow cab company had a higher percentage usage in 79% of the cities above.
- The Red cab company predominated and had a lower percentage usage in all these cities except in 21% of all the cities, which are Nashville TN, Phoenix AZ, Pittsburgh PA, Sacramento CA and San Diego CA where it predominated over the yellow cab company.

Question 4: What is the actual number of the yellow and pink cab users in the different cities with respect to the reported usage from each different city

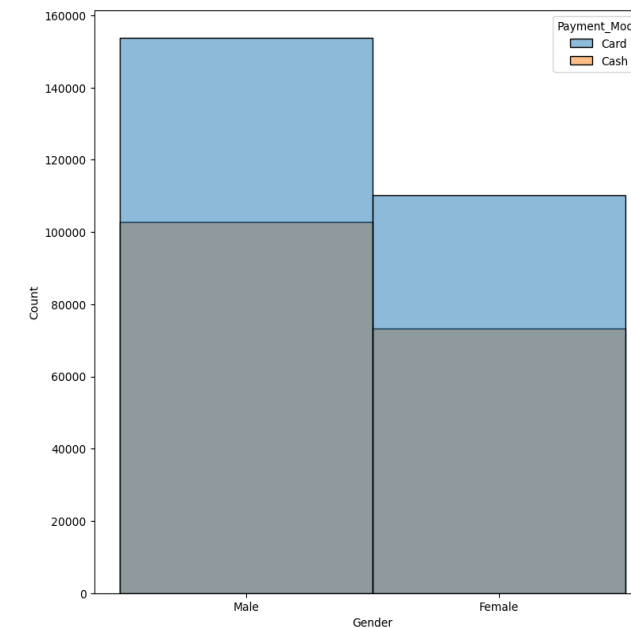
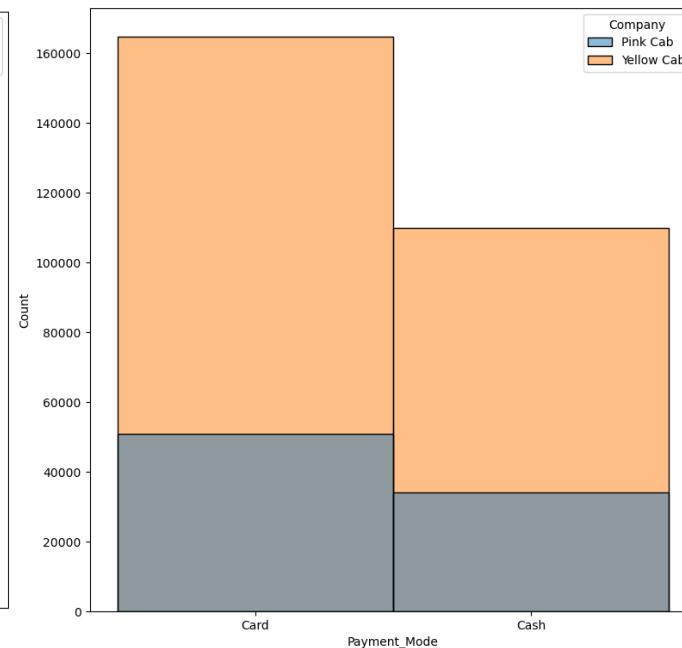
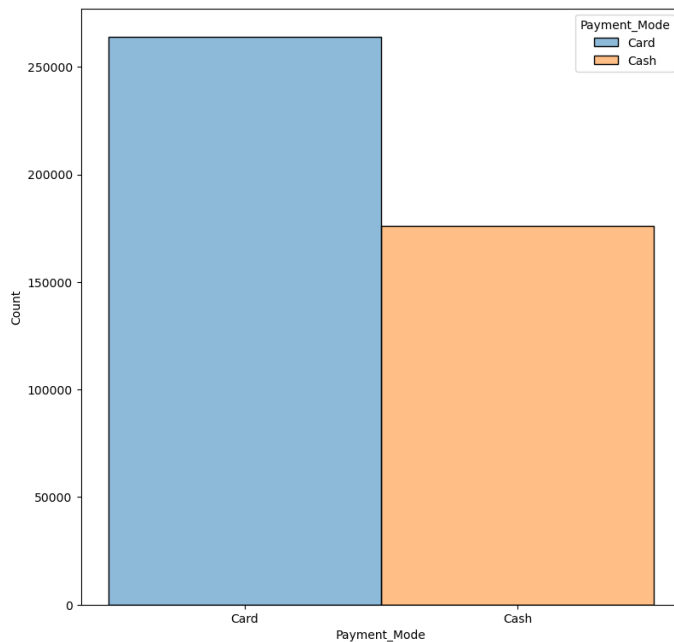


- The yellow cab company has the highest number of users compared to the pink cab company.
- Generally Los Angeles, Chicago IL , Washington DC , Boston MA and New York NY have the highest number of users reported to use the yellow cabs.

Question 5: What is the most preferred mode of payment between the male and female customers throughout all the cities?

Payment mode	Female	Male	Income (USD/Month)
Card	25.03352	34.95108	15085.4242
Cash	16.65879	23.35662	15102.3121

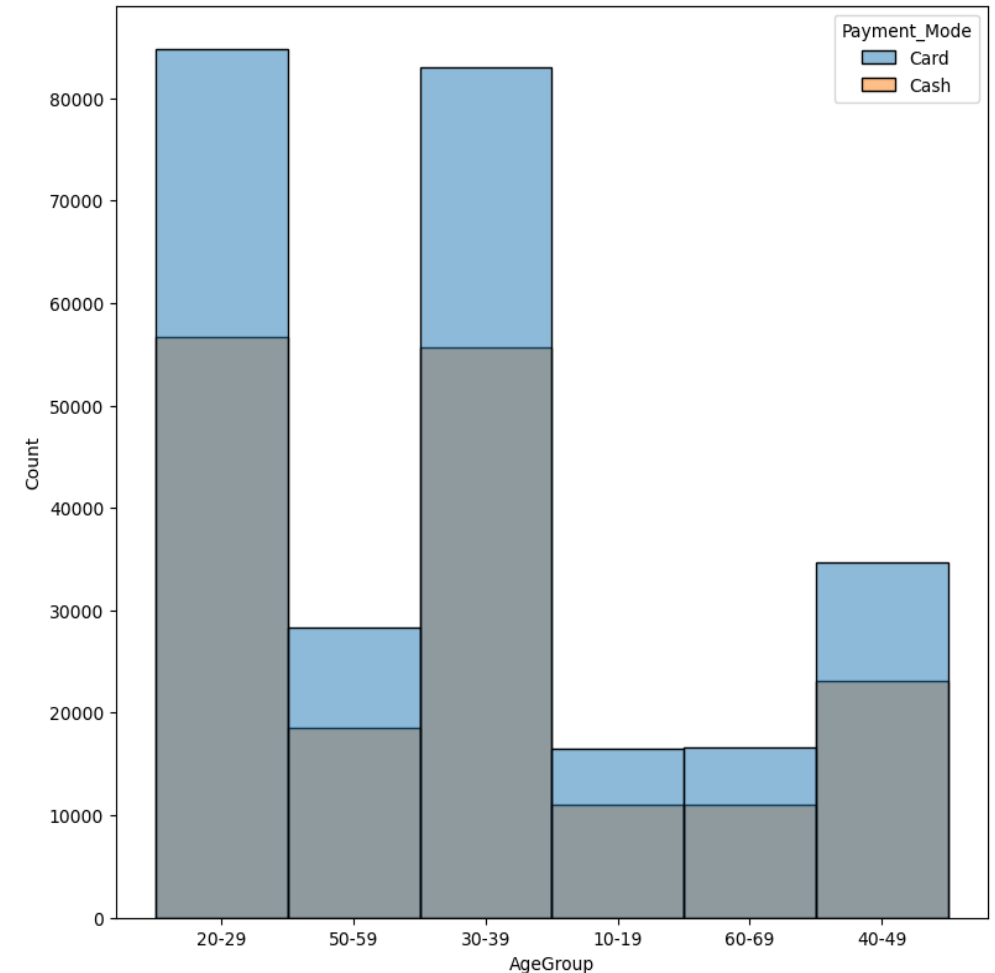
- 59.98% of the customers paid using card compared to the 40% who used cash, 34.95% of the males and 25.03% of the females paid using card.
- On average more users preferred using card than cash when paying for the yellow cab.



Question 6: Which age group predominates overall across all the cities and what is the most preferred mode of payment amongst all the different age groups?

Age Group	Age	Female	Male	Card	Cash	Income (USD/Month)
	count					mean
10-19	3112	11093	16420	16544	10969	14849.35701
20-29	15797	59913	81657	84823	56747	15068.13205
30-39	15579	57348	81440	83087	55701	14984.56274
40-49	6336	24141	33640	34624	23157	14992.38889
50-59	5229	19186	27641	28263	18564	15100.49761
60-69	3118	11806	15813	16650	10969	14975.74471

Many customers for the different cab companies are reported to be between 20 years to 39 years of age, mostly used card as the preferred method of payment for the cab.

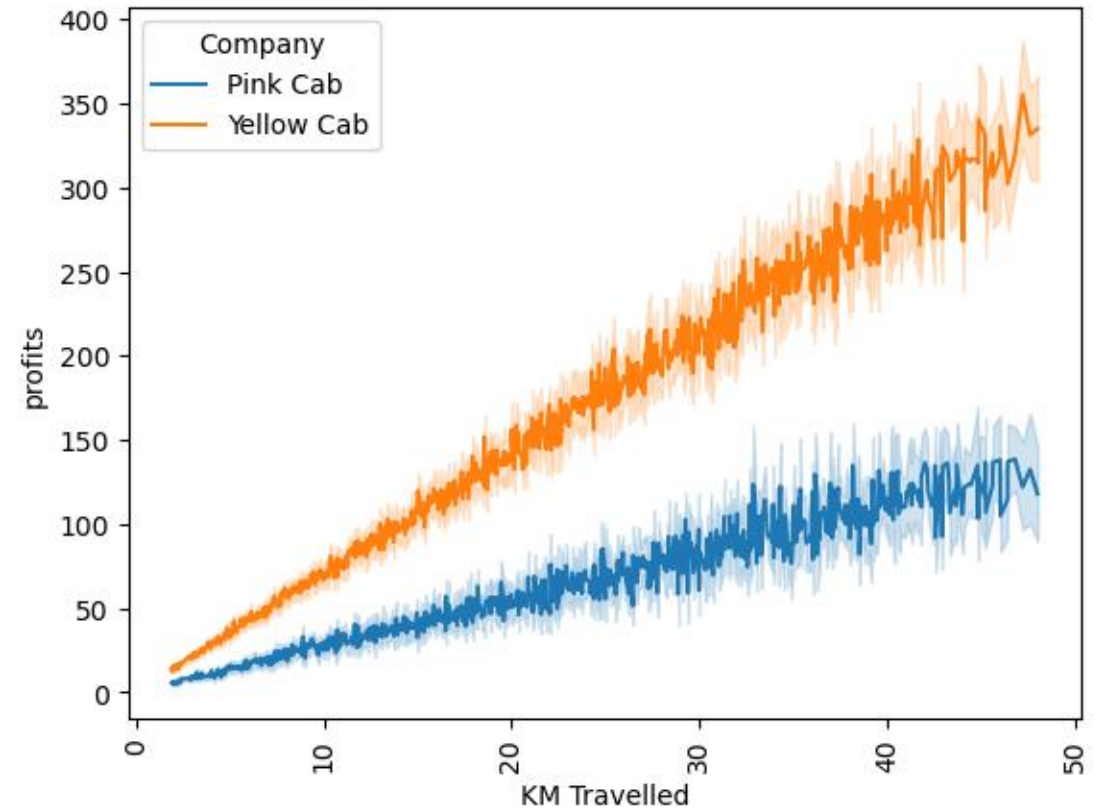
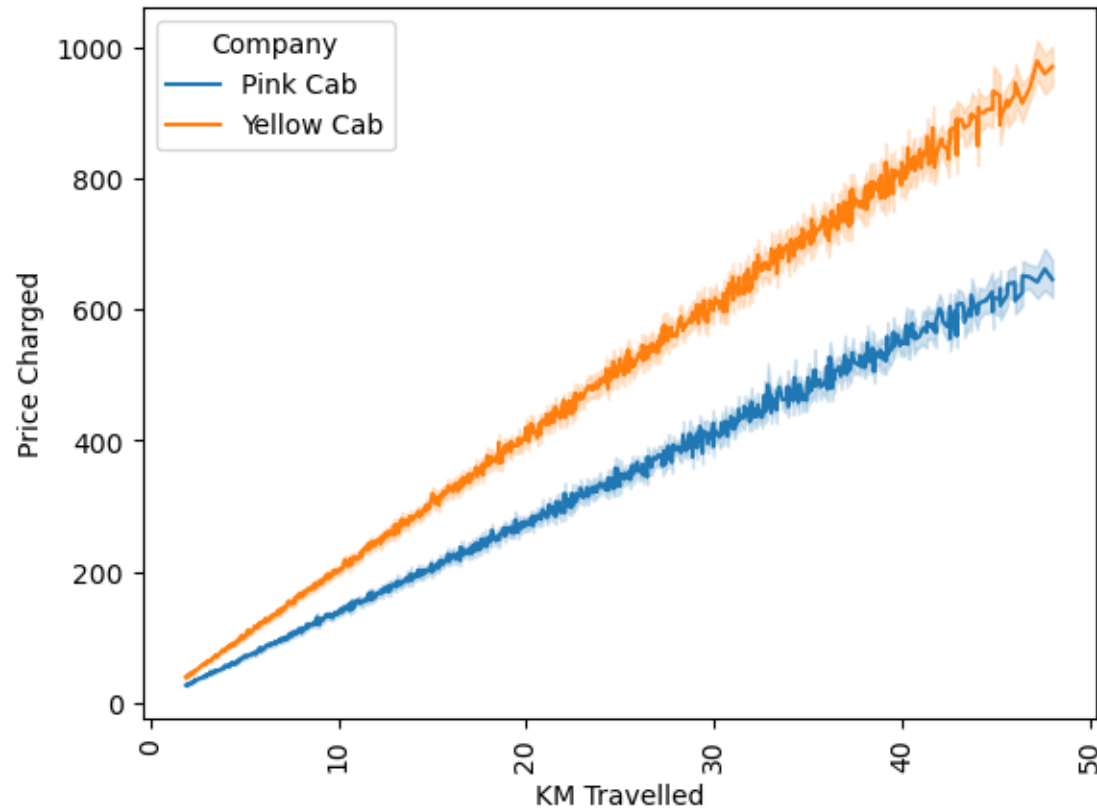


Question 7: What is the correlation between kilometers travelled, price charged , cost of trip and profits from each trip?

	KM Travelled	Price Charged	Cost of Trip	profits
KM Travelled	1	0.835753	0.981848	0.462768
Price Charged	0.835753	1	0.859812	0.864154
Cost of Trip	0.981848	0.859812	1	0.486056
profits	0.462768	0.864154	0.486056	1

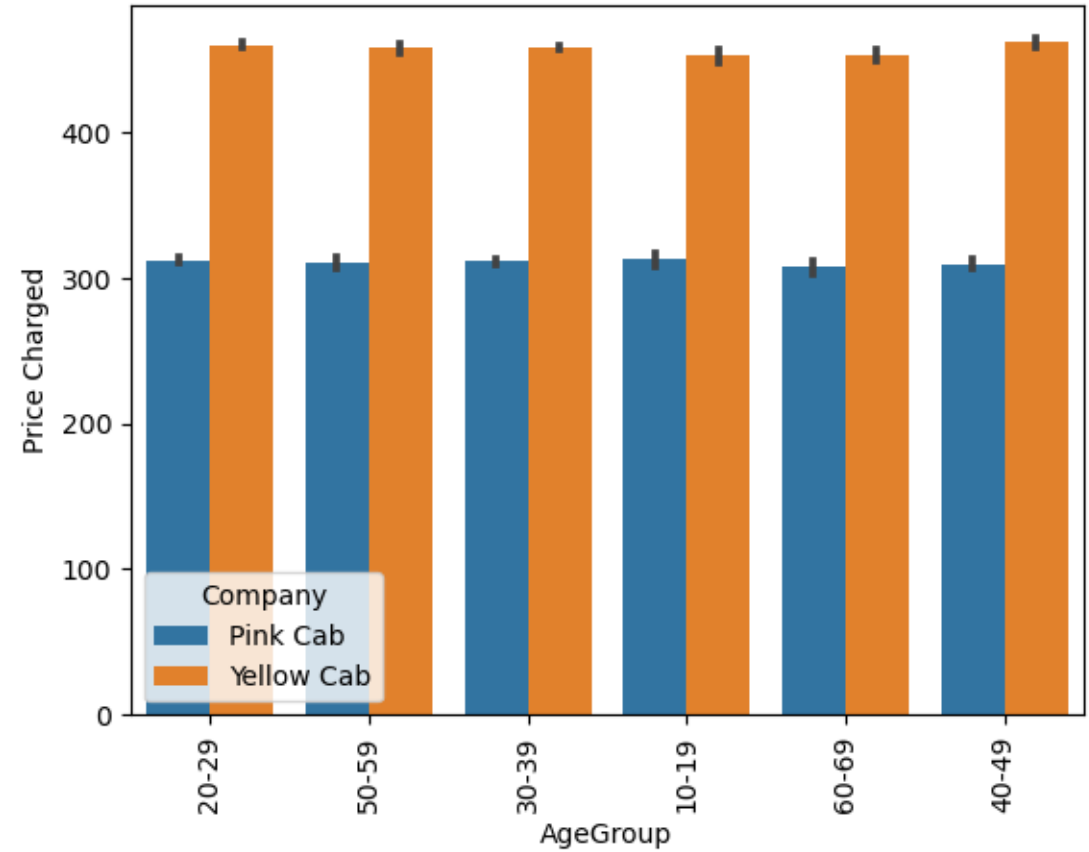
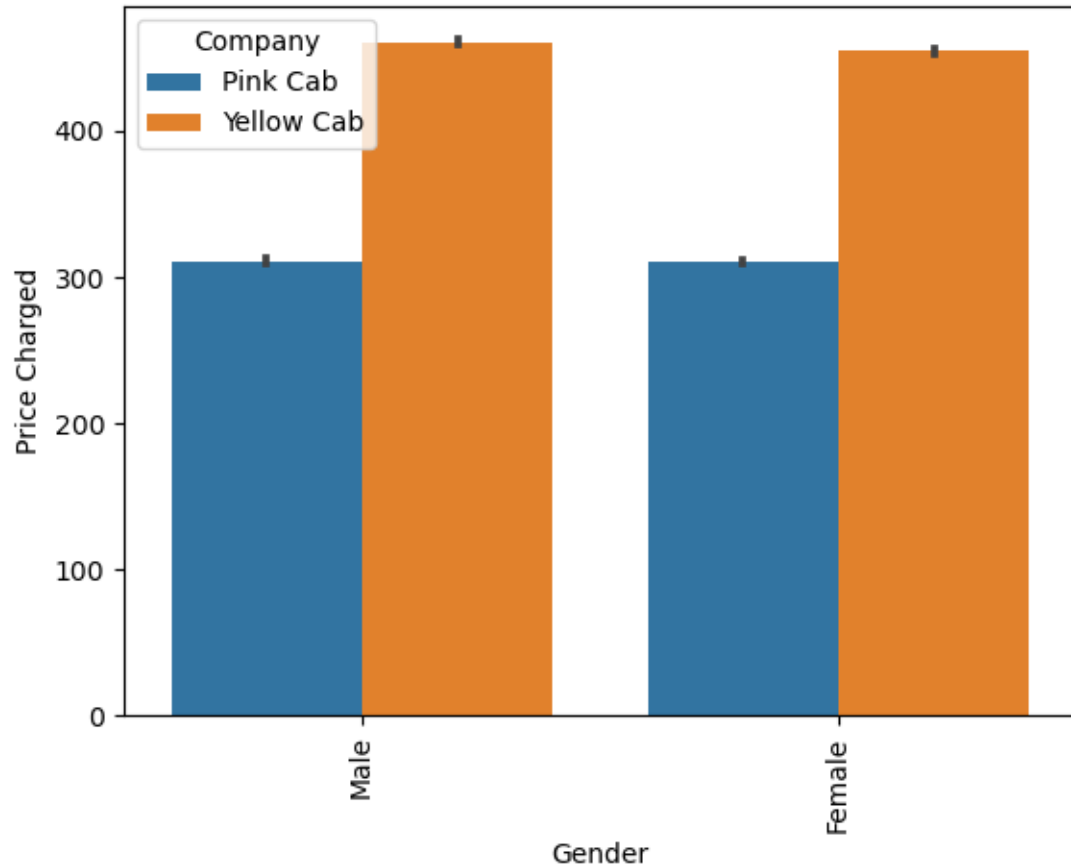
- 46.27% of the reported profits are accounted for by the change in kilometers travelled
- 48.61% of the reported profit (extra income due to over charging) is attributed to the change in cost of trip
- An increase in kilometers travelled affected the cost of trip and price to be charged for that trip

Question 8: How is the price charged per trip differ for the two cab companies and is it affected by number of kilometers travelled?



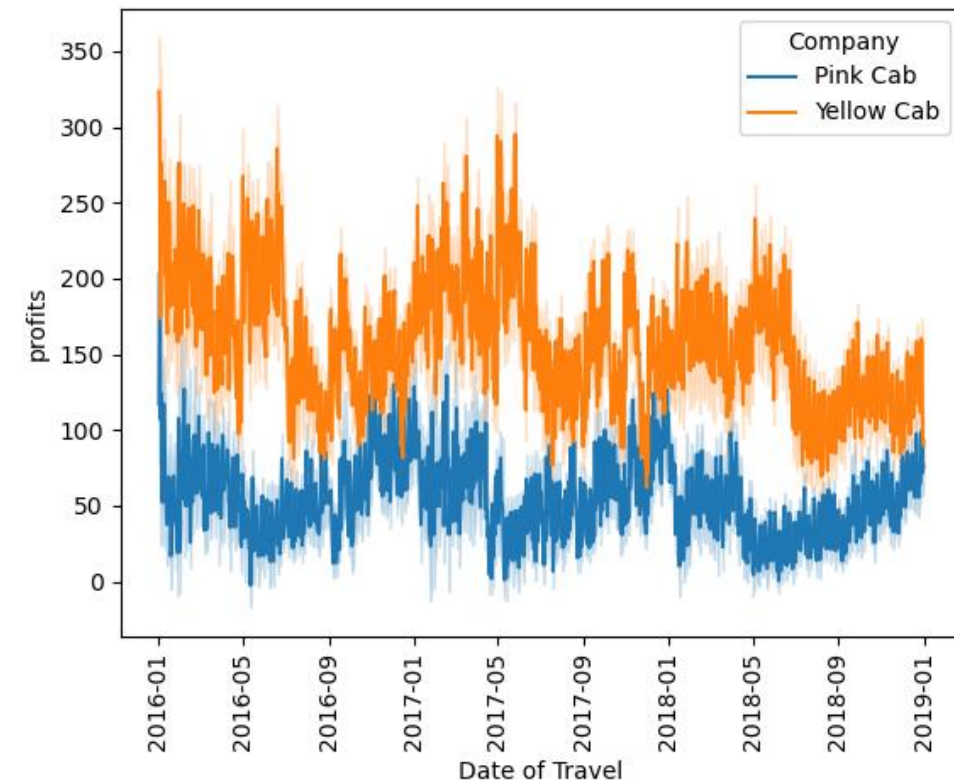
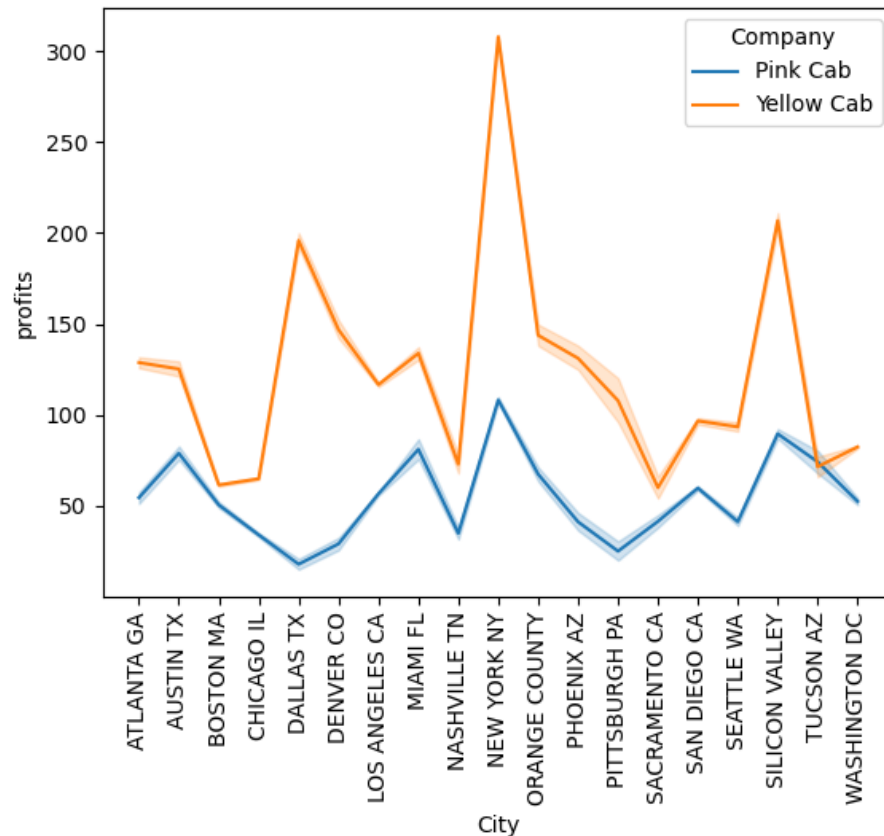
- The yellow cab company charged a bit more for the same number of kilometers than the pink cab
- Both cab companies have reported equal number of KM travelled, however in comparison to the amount of extra money earned per kilometers travelled, the yellow cab predominated over the pink cab
- However, with increase in number of KM travelled, the higher the amount of extra cash earned for the yellow cab than the pink cab

Question 9: Is the price charged for any cab company affected by gender and age group of the customer?



- Both male and female customers paid extra when using the yellow cab as for the pink cab
- All age groups were charged the same and all were charged the same for the yellow cab as it is for the pink cab
- All customers regardless of gender and age were willing to pay price charged by the cab companies

Question 9: Which cab company makes more extra income per trip over the period of three years reported?



- The yellow cab company predominates in the market share and consequently attains more profits over the pink cab
- The yellow cab charged a little more than the pink cab therefore it earns extra in all cities reported for a period of three years.
- In some major cities such as New York NY, Silicon valley, Dallas TX, the yellow cab earns extra money than the pink cab.

EDA summary

1. Descriptive analysis to calculate summary statistics for transaction amounts, customer demographics, and payment modes, as well as exploring temporal trends in transaction volumes and customer behavior from 31/01/2016 to 31/12/2018.
2. Customer profiling will involve segmenting customers based on demographic characteristics and analyzing behavior and payment modes to gain insights into customer segments.
3. Market analysis will evaluate the performance of two cab companies based on transaction data, market share, and customer retention rates, and analyze the distribution of cab users across US cities using City.csv data to identify market opportunities.
4. Correlation analysis will be conducted to identify relationships between variables influencing customer preferences, transaction behavior, and market dynamics.

Key takeaways and recommendations

- The yellow cab company has a higher market share in terms of number of reported users in most major cities compared to the pink cab, therefore consequently makes more money
- The yellow cab company is most preferred than the pink cab and this can be explained by the percentage usage in every city
- Mostly customers between the ages of 20-39 years use cabs, this population accounts more on the number of reported users.
- With the most preferred payment mode being card, this ensures more secure payment, more users of the yellow cab opted for the card payment.

Key takeaways and recommendations

- The yellow cab charged extra for the same number of kilometers travelled than the pink cab, consequently the yellow cab makes more extra cash than the pink cab.
- There has been consistent reported usage of the yellow cab for a period of three years, it still predominated over the pink cab in terms of market share and extra cash.
- All customers regardless of gender and age were willing to pay price charged by the yellow cab company.
- Generally Los Angeles, Chicago IL , Washington DC , Boston MA and New York NY have the highest number of users reported to use the yellow cabs and these offer great business opportunities and hotspots for customers

Recommendations

- The yellow cab has more market share, higher percentage usage and higher number of users in most of the major and crowded cities in the USA.
- It's use is consistent throughout the year and over years with more reported users and extra cash than the pink cab.
- Therefore, it's a better investment option over the pink cab company.

The end

Thank You