Exploratory Data Analysis

G2M insight for Cab investment firm

By Brennan Clinch
Data Science Intern at Data Glacier

8/21/22

Agenda

Understanding the Business Problem

Information on Datasets

EDA

Hypothesis Testing

Conclusions

Understanding the Business Problem

Understanding the Business Problem

XYZ is a private firm that is located in the United States. It has recently decided that it plans to invest in the Cab industry. The Cab companies that XYZ is making a decision on are Yellow Cab and Pink Cab. So we, as a part of XYZ's executive team, must help XYZ determine what company would be a better choice to invest in. We will determine the best Cab company through factors such as profit, popularity, user preferences, etc.

Information on Datasets

Information on Datasets

The Datasets we are provided:

- Cab_Data.csv : Contains information on the transactions for the 2 Cab companies. The transactions include the date of travel, city in the US, the price that was charged for a trip along with the true cost of the trip, and the trip mileage.
- City.csv : Contains information on the demographics for the cities selected in the US for both Cab companies.
- Customer_ID.csv : Contains information on the demographic details on customers. This includes details such as age, income, and gender.
- Transaction_ID.csv: Maps together the Customer ID and Transaction ID and contains information on the mode of payment made by customers.

- For this project, the 4 datasets were all joined together into one dataset (Cab.csv) for the EDA and hypothesis testing.

EDA

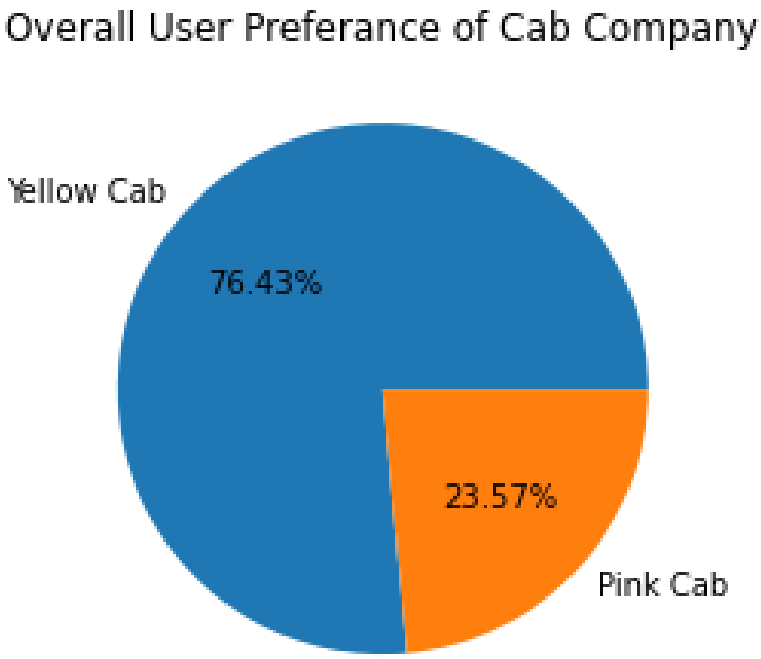
Correlation Between Variables

	Transaction ID	KM Travelled	Price Charged	Cost of Trip	Customer ID	Age	Income (USD/Month)	Population	Users
Transaction ID	1.000000	-0.001429	-0.052902	-0.003462	-0.016912	-0.001267	-0.001570	0.023868	0.013526
KM Travelled	-0.001429	1.000000	0.835753	0.981848	0.000389	-0.000369	-0.000544	-0.002311	-0.000428
Price Charged	-0.052902	0.835753	1.000000	0.859812	-0.177324	-0.003084	0.003228	0.326589	0.281061
Cost of Trip	-0.003462	0.981848	0.859812	1.000000	0.003077	-0.000189	-0.000633	0.015108	0.023628
Customer ID	-0.016912	0.000389	-0.177324	0.003077	1.000000	-0.004735	-0.013608	-0.647052	-0.610742
Age	-0.001267	-0.000369	-0.003084	-0.000189	-0.004735	1.000000	0.003907	-0.009002	-0.005906
Income (USD/Month)	-0.001570	-0.000544	0.003228	-0.000633	-0.013608	0.003907	1.000000	0.011868	0.010464
Population	0.023868	-0.002311	0.326589	0.015108	-0.647052	-0.009002	0.011868	1.000000	0.915490
Users	0.013526	-0.000428	0.281061	0.023628	-0.610742	-0.005906	0.010464	0.915490	1.000000

- From this table, we see that there is a strong correlation between:
  1. Price Charged and KM Travelled
  2. Population and Users
  3. KM Travelled and Cost of Trip

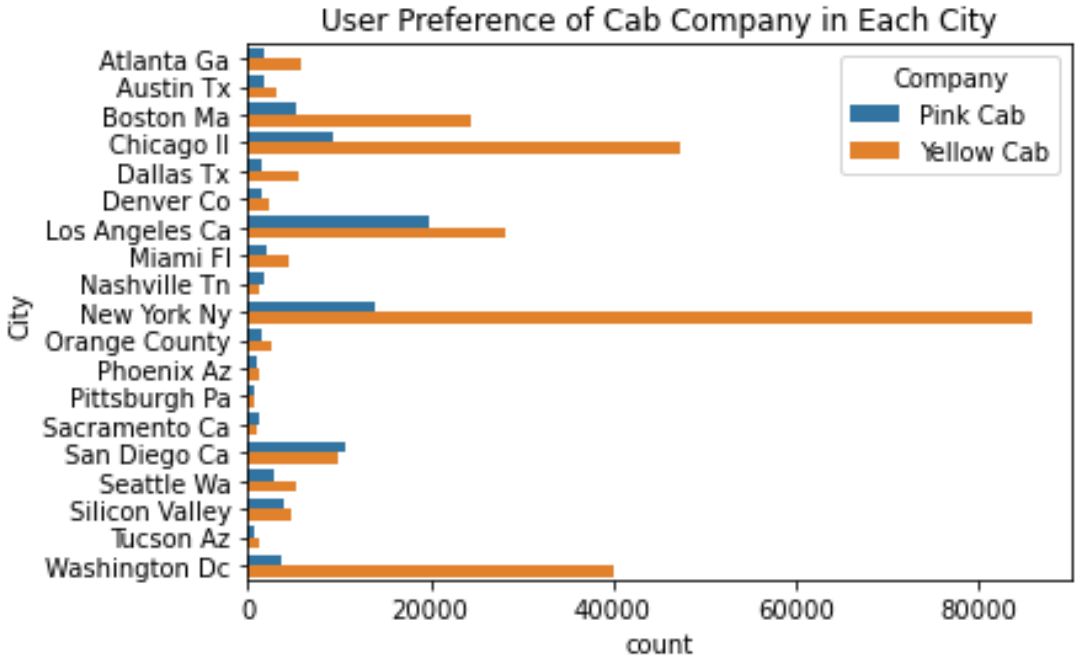
User Preference of Cab Company

- Overall, more than ¾ of all users prefer Yellow Cab.



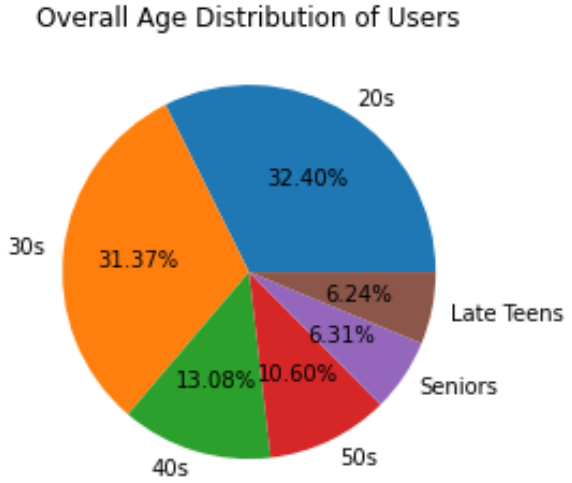
User Preference of Cab Company By City

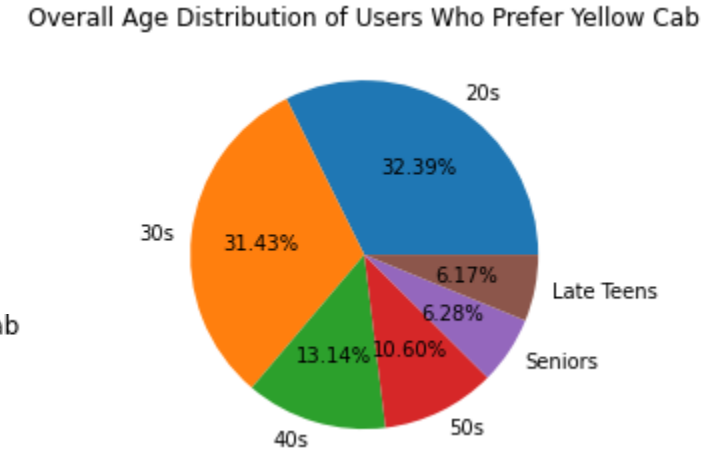
- New York receives the most users out of all the cities. With Chicago and Washington DC being the next top cities.
- A few cities have a preference for Pink Cab. These cities include San Diego, Sacramento, and Nashville.

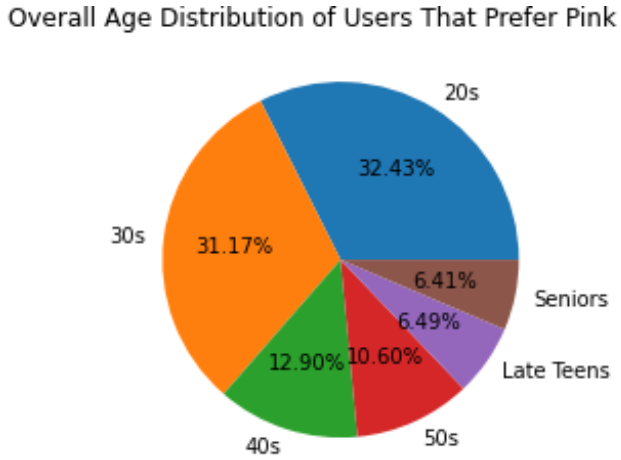


Age Distributions of Users

- Based on the information, we see that the age group that got the most users overall was those in their 20s.
- Pink Cab had the most users that were Seniors (60+).
- Overall, young adults are the target user population for both companies.

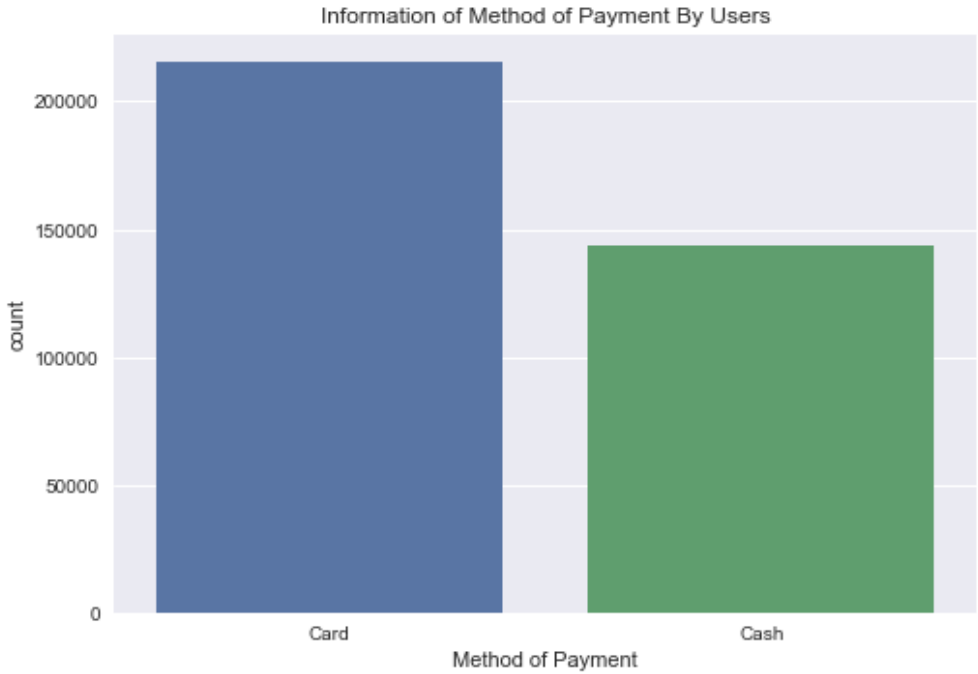


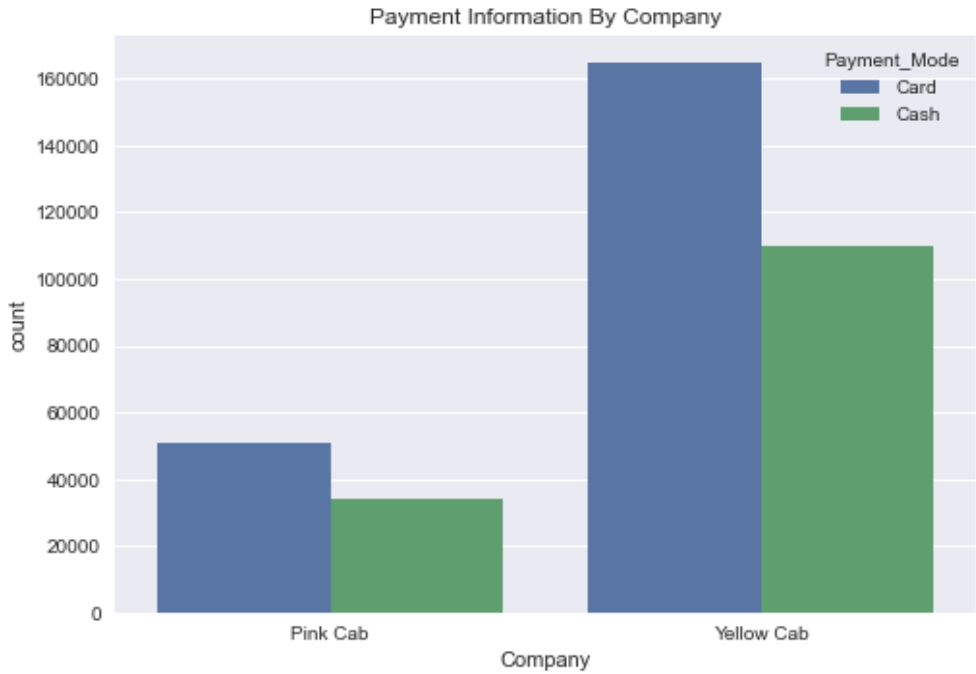




User Payment Preferences

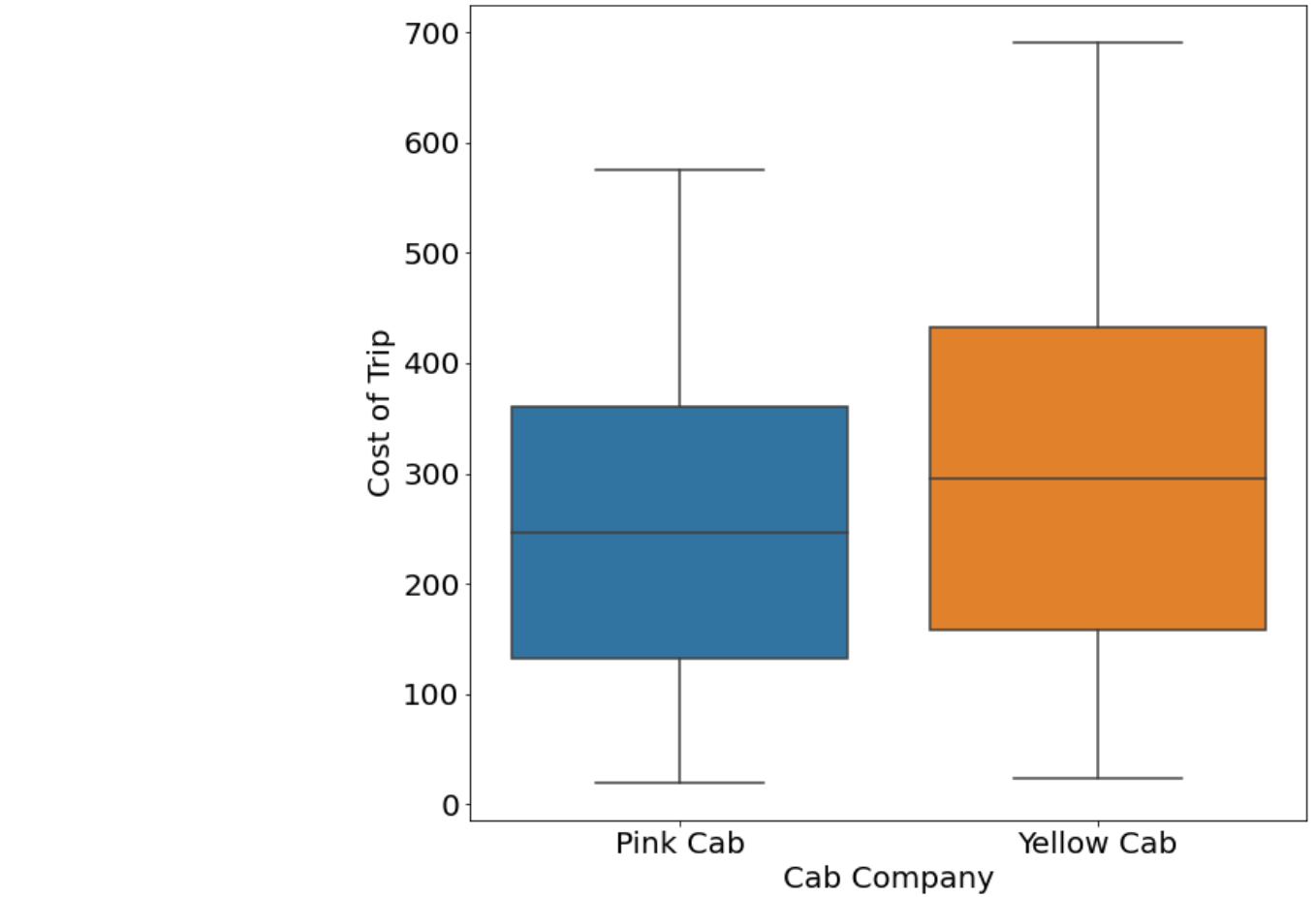
- Overall, by Cab company and as a whole, most users prefer to pay by credit card.





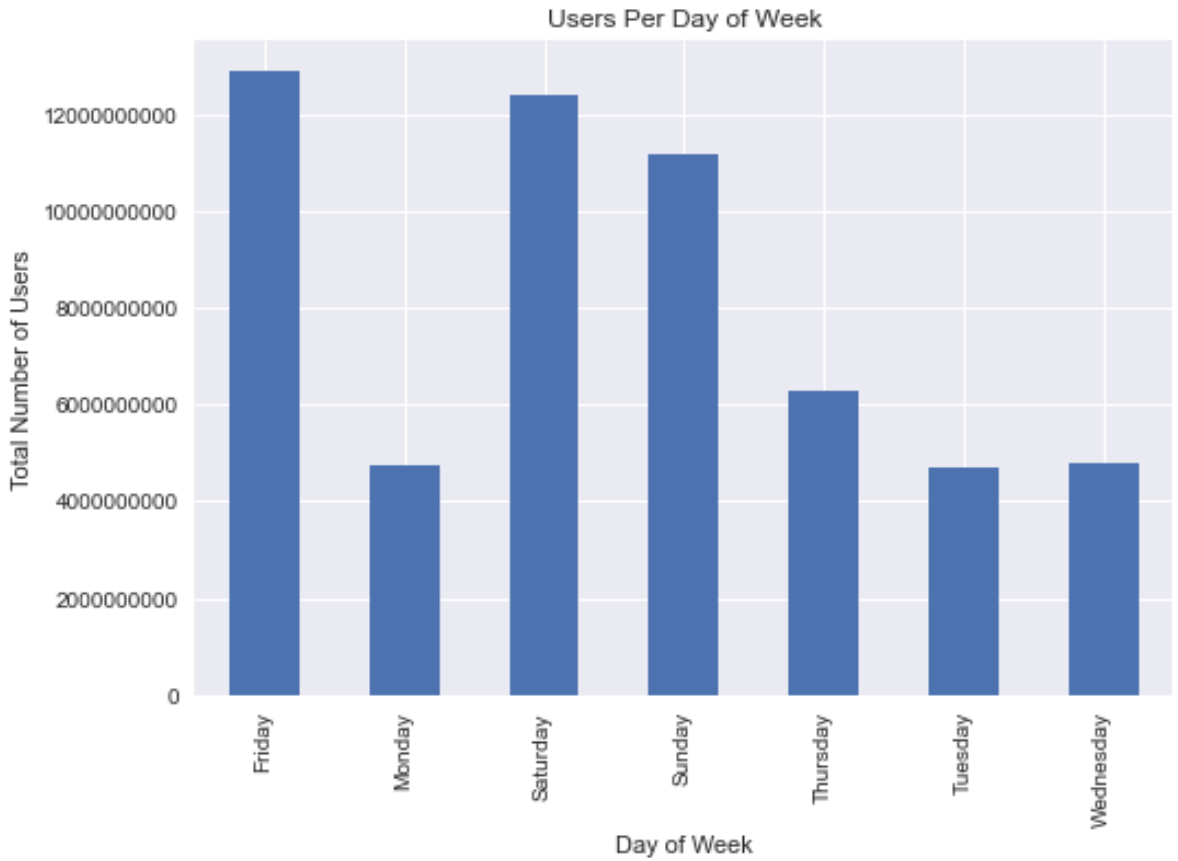
Cost of Trip

- For users who prefer Yellow Cab over Pink Cab, their cost of the trip is higher.
- This means that while Yellow Cab is the more popular option, Pink Cab might offer cheaper prices.



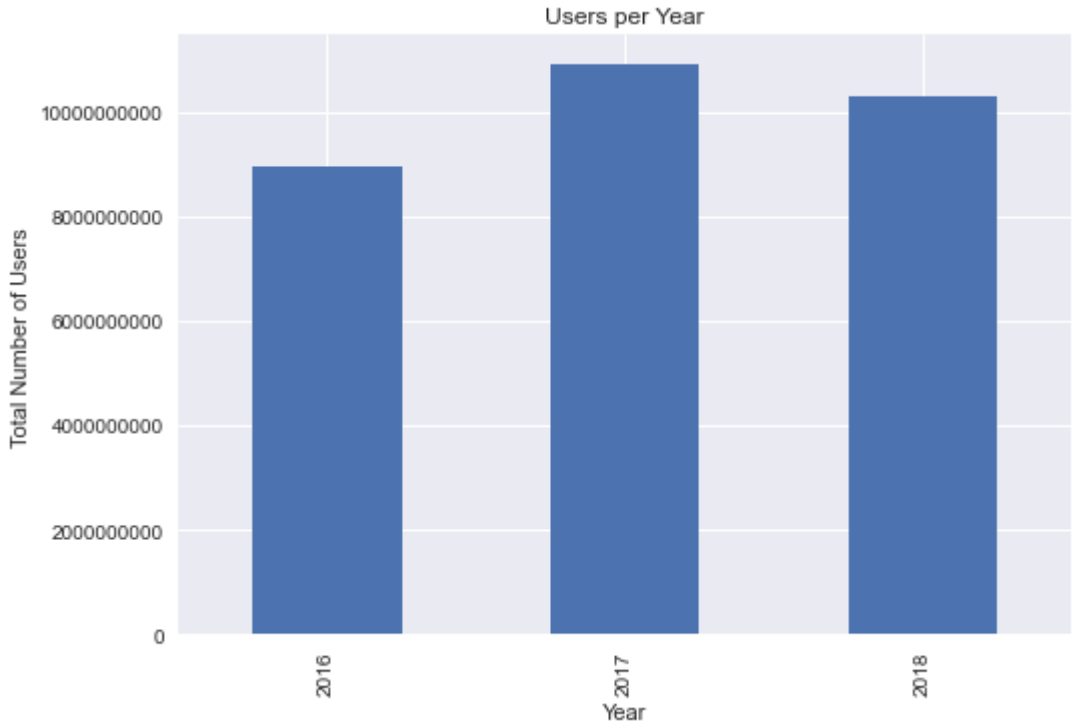
Amount of Users By Day of Week

- Overall, weekend days got the most users, with Fridays being the most popular.



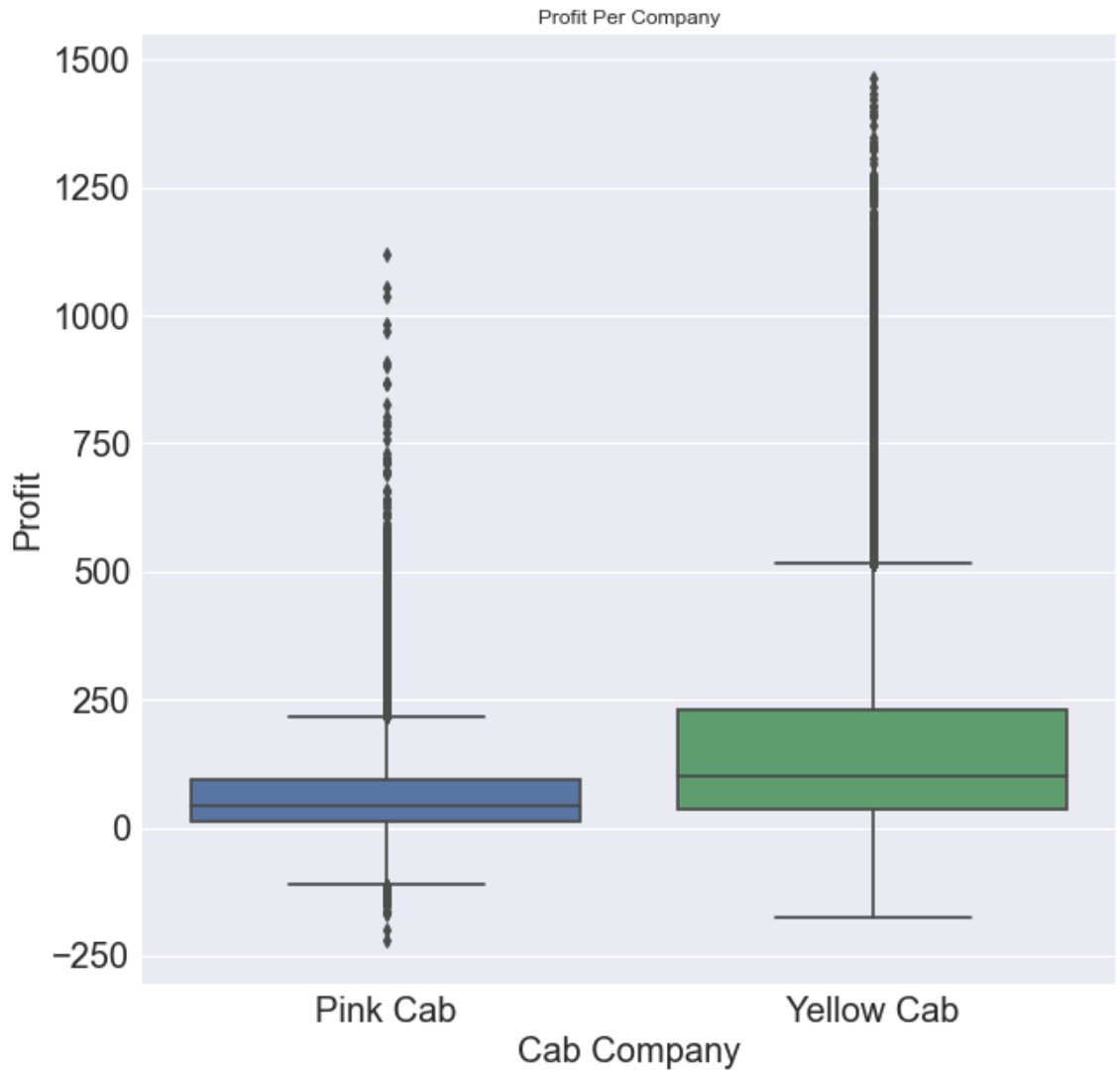
Amount of Users Per Year

- Overall, from the diagram, 2017 was the year that got the most users over both Cab companies.



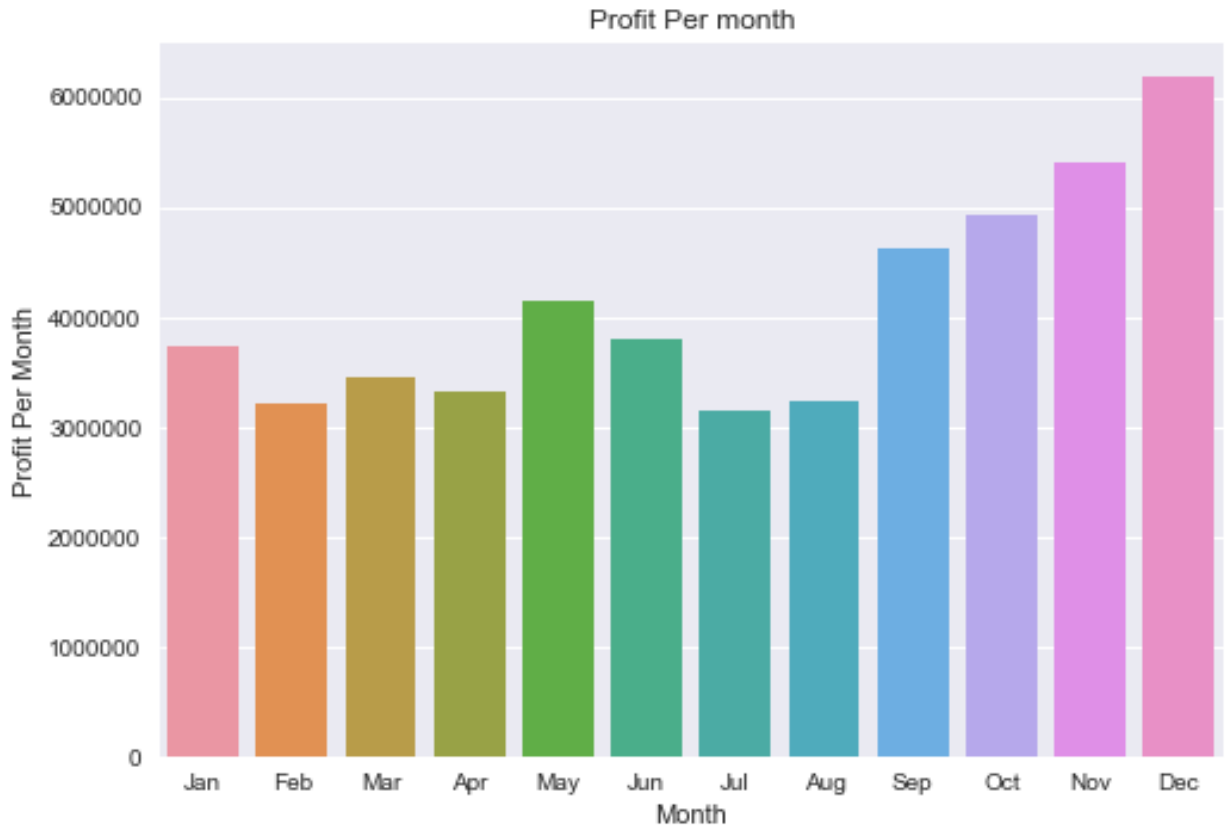
Profit Per Company

- We created a new variable which is Profit. It is equivalent to the difference between the Price Charged and the Cost of Trip.
- On average, Yellow Cab made more Profit per year than Pink Cab.
- From the boxplots, there appear to be many outliers with Profit being more than $ 500, which makes the median a much better estimate than the mean.



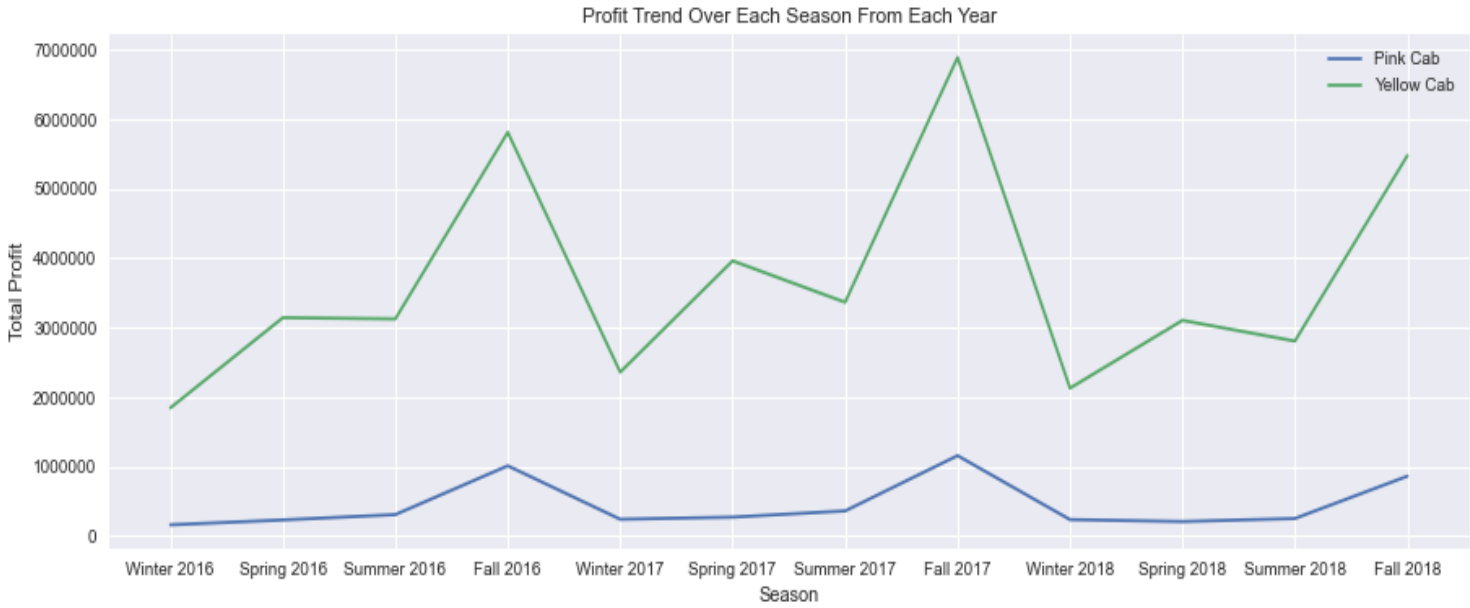
Profit Per Month

- Both Cab companies had the most amount of users during December through all the years.



Profit Trend Per Season

- For Pink Cab and Yellow Cab, Profit was the highest during Fall 2017, and lowest during Winter 2016.
- Profit had a very steady increase from the winter to the summer during all 3 years for Pink Cab.
- For Yellow Cab, Total Profit barely declined and was close to constant from Spring 2016 to Summer 2016.



Hypothesis Testing

Hypothesis #1: Is there a significant difference between the Cost of Trip from the Pink Cab company given the payment mode?

-Null Hypothesis H0: There is no difference between the Cost of Trip from the Pink Cab company given the payment mode

-Alternative Hypothesis HA: There is a difference between the Cost of Trip from the Pink Cab company given the payment mode

-μ1 = Mean value of Cost of Trip for Pink Cab users who pay with cash

-μ2 = Mean value of Cost of Trip for Pink Cab users who pay with card

```
P-Value: 0.37432686
We fail to reject the null hypothesis and conclude that there is no significant difference between the cost of the trip based on payment method for Pi
nk Cab company.
```

We can conclude that there is no difference between the cost of trip for Pink Cab users regarding the method of payment.

Hypothesis #2: Is there a significant difference between the amount of users in 2017 vs 2018?

-Null Hypothesis H0: There is no difference between the amount of users in 2017 and 2018

-Alternative Hypothesis HA: There is no difference between the amount of users in 2017 and 2018

-μ1 = Mean value of amount of users in 2017 (Yellow, Pink, or overall)

-μ2 = Mean value of amount of users in 2018 (Yellow, Pink, or overall)

Yellow Cab

```
P-Value: 0.01861653
We can reject the null hypothesis and conclude that there is a significant difference between the amount of users in 2017 and 2018 for Yellow Cab.
```

Pink Cab

```
P-Value: 0.38192002
We fail to reject the null hypothesis and conclude that there is no significant difference between the amount of users in 2017 and 2018 for Pink Cab.
```

We can conclude that Yellow Cab had a more significant difference between the total number of users in 2017 vs 2018 than Pink Cab.

Hypothesis #3: Is there a significant difference in the Profit Ratio (Profit/KM Travelled) during 2016 and 2017?

-Null Hypothesis H0: There is no difference between the Profit Ratio during 2016 and 2017

-Alternative Hypothesis HA: There is a difference between the Profit Ratio during 2016 and 2017

-μ1 = Mean value of Profit Ratio in 2016 (Yellow, Pink, or overall)

-μ2 = Mean value of Profit Ratio in 2017 (Yellow, Pink, or overall)

Pink Cab

```
P-Value: 0.00964014
We can reject the null hypothesis and conclude that there is a significant difference between the Profit Ratio's in 2016 and 2017 for Pink Cab.
```

Yellow Cab

```
P-Value: 0.86666235
We fail to reject the null hypothesis and conclude that there is no significant difference between the Profit Ratios in 2016 and 2017 for Yellow Cab.
```

We can conclude that Pink Cab had a significantly higher difference in their Profit to KM ratio from 2016 to 2017 vs Yellow Cab.

Conclusions

- Based on our analysis in the EDA, we can conclude that Yellow Cab is the best Cab company for XYZ to invest in due to it’s superior popularity from users to it’s higher amounts of Profit being made on average.
- Some other interesting insights that were gained:
  - Pink Cab is a good Cab company for users looking for cheaper prices charged, since average Cost of Trip was lower based on the EDA.
  - California could have more Pink Cab services than other cities in the US since 2 cities prefer Pink Cab over Yellow Cab. Whereas big cities such as New York and Chicago have an enormous preference for Yellow Cab.
  - Weekends are the most popular time for Cab usage.
  - Fall is the most popular season with Winter being the least popular for usage.
  - Young Adults are the most popular age group for both companies.
  - Credit Card is the most popular payment method.

Thank You