Data Glacier

Your Deep Learning Partner

G2M Case Study: Exploratory Data Analysis

Virtual Internship

Company: G2M Insights for XYZ Cab Investment Firm

Author: Ammar Sidhu

Date: 10/21/2022

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Correlation Analysis

Recommendations

Background – G2M (Cab Industry) Case Study

- **Context:** XYZ is a private equity 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.

- **Objective:** Provide actionable insights to help XYZ firm in identifying the right company for making investment.

- **Components of Analysis:**
  - Data Exploration
  - Univariate & Bivariate Data Visualizations
  - Correlation Analysis
  - Hypothesis Testing
  - Recommendations

Data Exploration

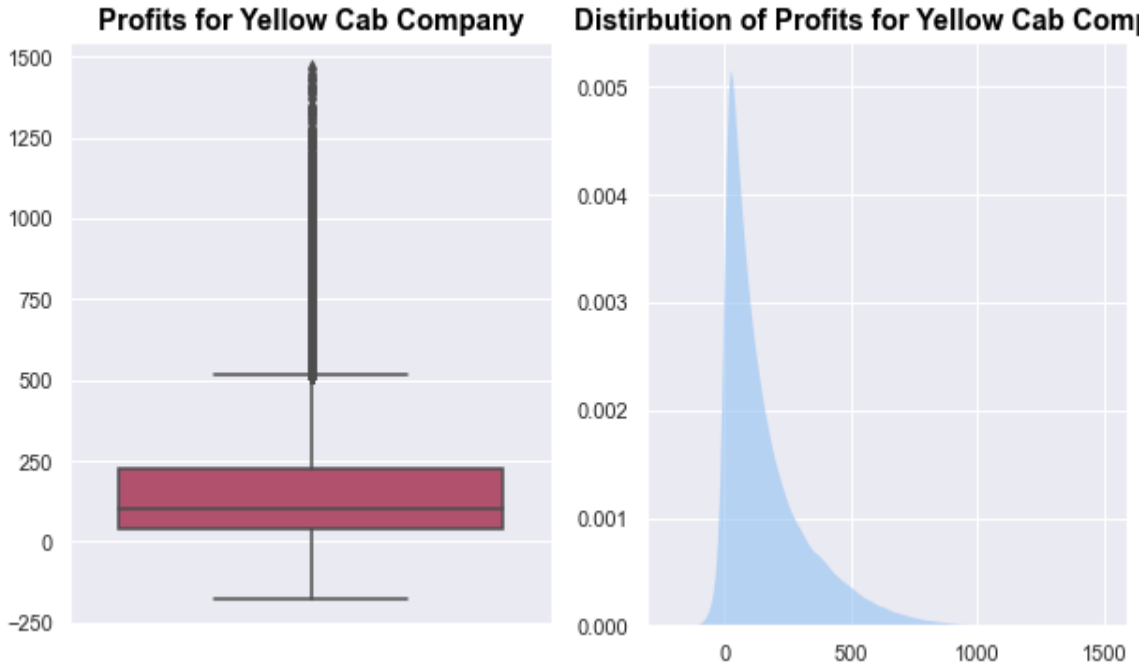
Data Description:

- **16 Features** (Cab Ride Info) + **1 Target** (Profits ($) per Ride)
- Timeframe of the data: **2016-01-31 to 2018-12-31**
- Number of Rows (Transactions): **359,392 entries**

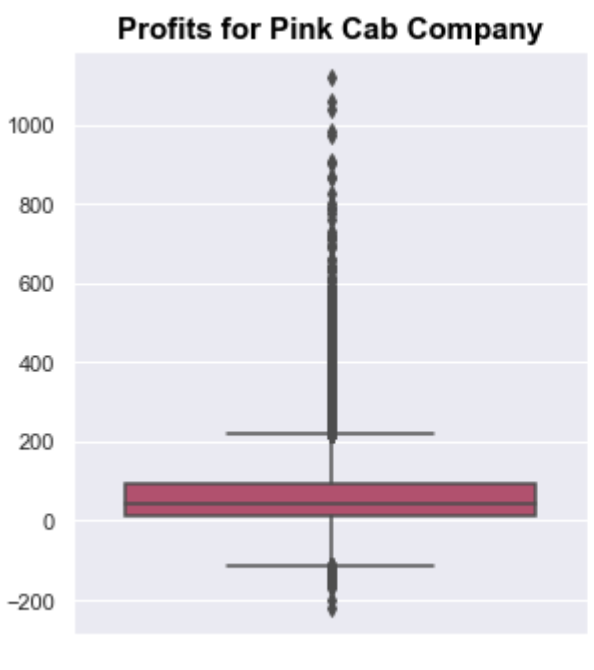
Assumptions & Facts:

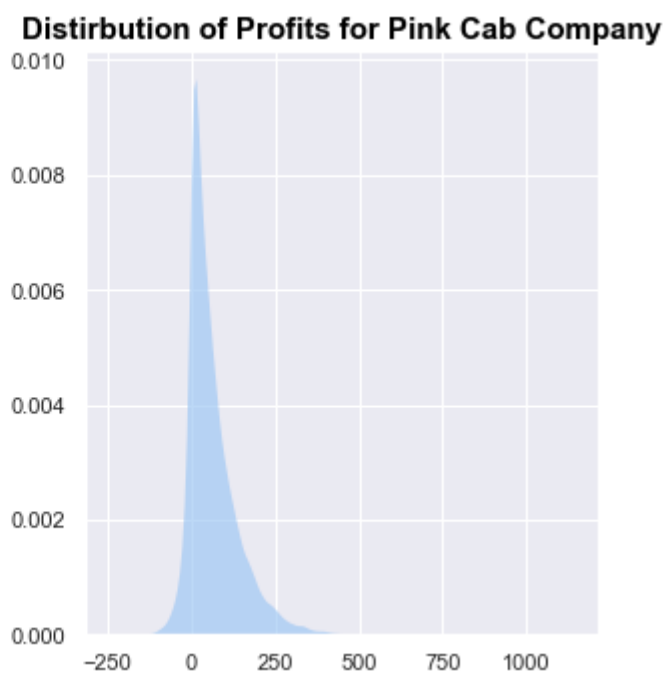
- Outliers are present in the Profits feature but are relevant to the analysis to explore where and why some rides are more profitable.
- None of the transactions are duplicates and are separate customer transactions with a given cab company.
- Cab rides that had Profits below $0 are a consequence of discounts or poor cab services and are not simply data entry issues.
- There are no missing values or information for a given transaction.

Profits (Margins) of Both Cab Companies



The upper and lower bounds for the target are: (515.906, -250.05559999999997)



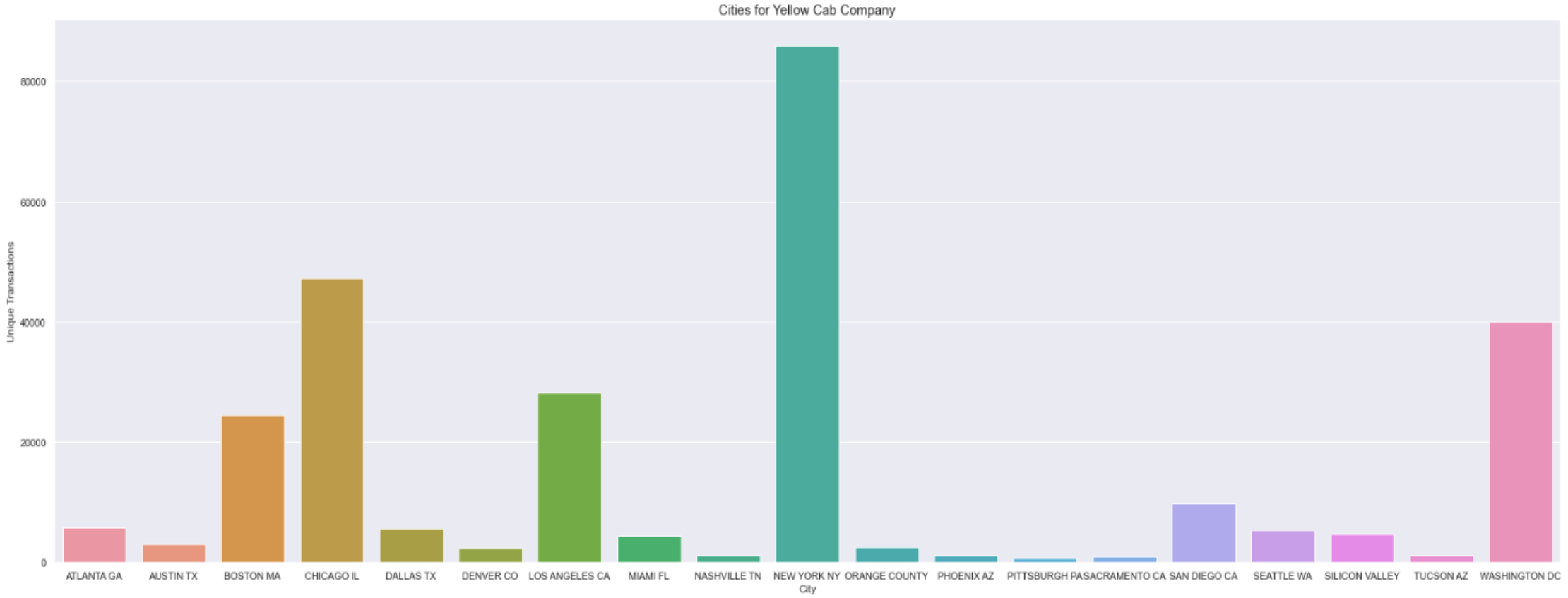


The upper and lower bounds for the target are: (218.1035, -112.93249999999999)

All outliers in both datasets that are above and below the upper/lower bounds of profits (margins) for both companies will be kept because they tell us 2 possible things about the cab ride:

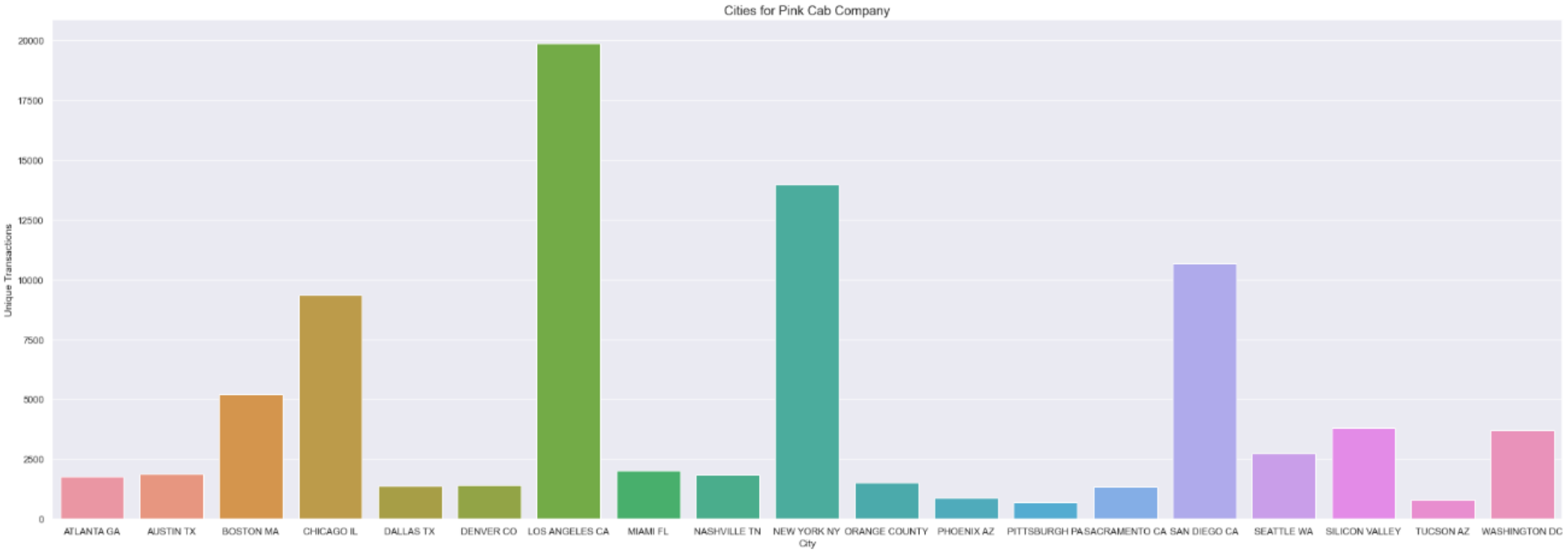
- If the profits are above the upper bound for a company, the cab is likely an expensive one due to the driver's vehicle, and the region they are traveling in has higher rates due to traffic and demand.
- If the profits are below the lower bound for a company, the passenger had some form of a discount token for the ride that reduced the amount they had to pay or they had a bad experience/issue that put the company at fault for the ride that resulted in less payment.

Yellow Cab Company Customers by City



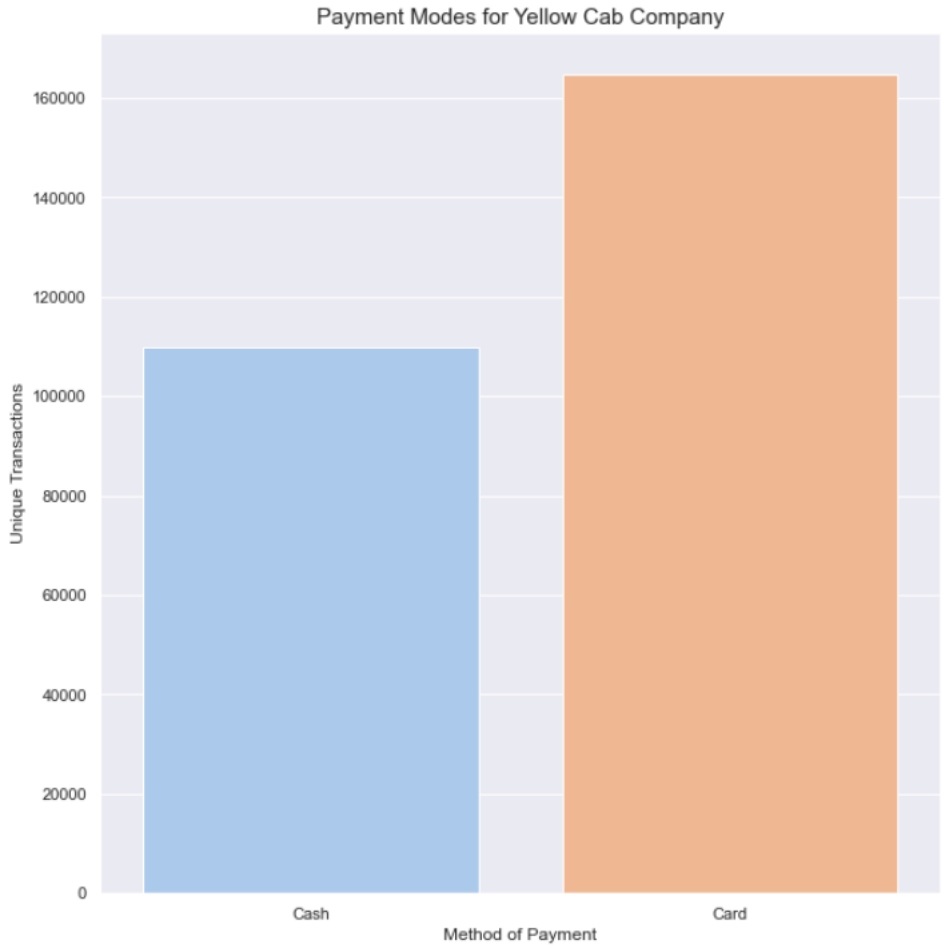
For the Yellow Cab Company, New York has far more customers than all other cities with a total of over 8000 customers during the time period of this data.

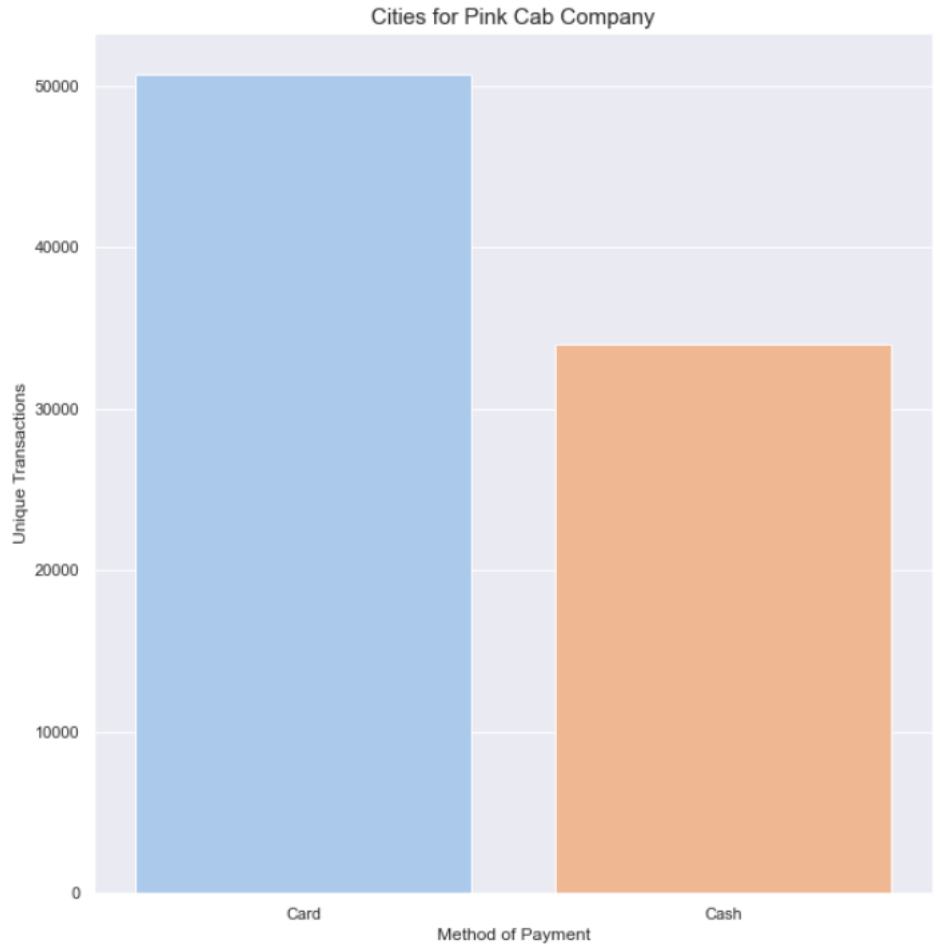
Pink Cab Company Customers by City



- For the Pink Cab Company, Los Angeles (Yellow Cab Company still outperforms Pink Cab Company here) is the most in-demand city followed by New York City.
- Cities like Phoenix, Pittsburgh, and Tucson are all not very much in-demand by both companies.
- Overall, Yellow Cab Company seems to attract more customers regardless of the city when compared to the Pink Cab Company because the transactions are just far greater.

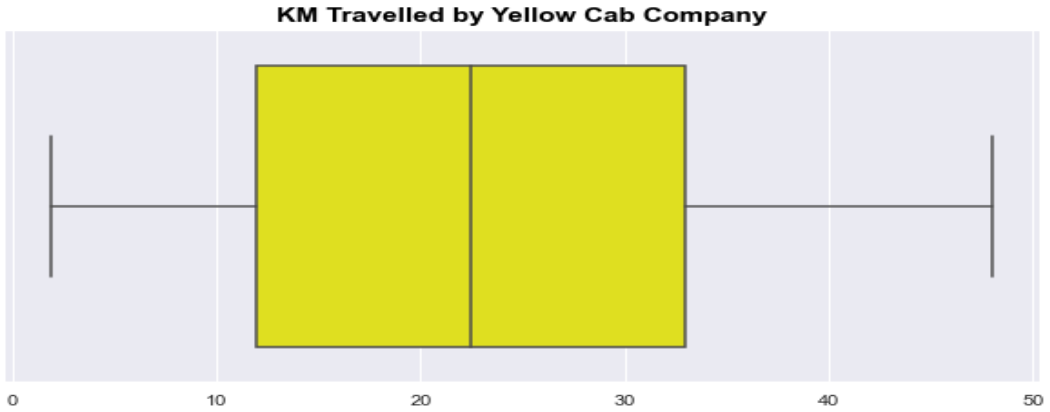
Payment Methods for Cab Companies

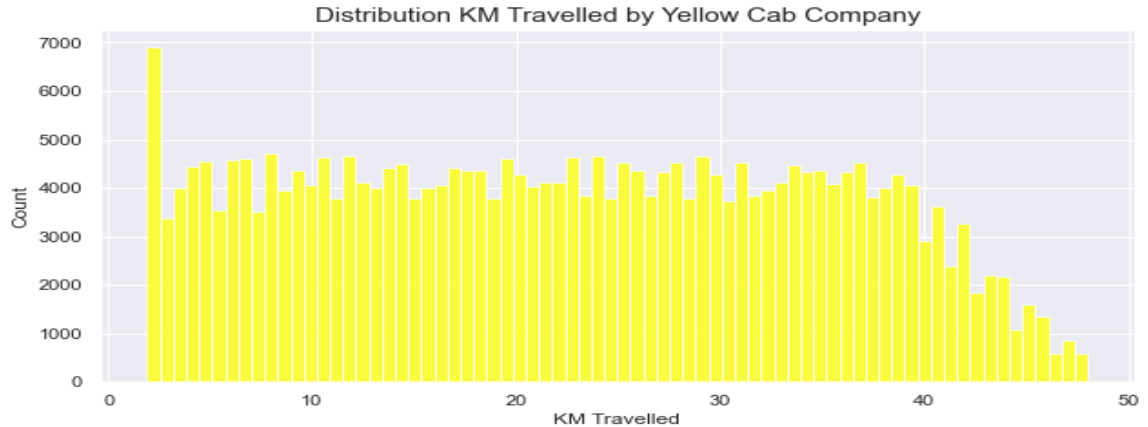


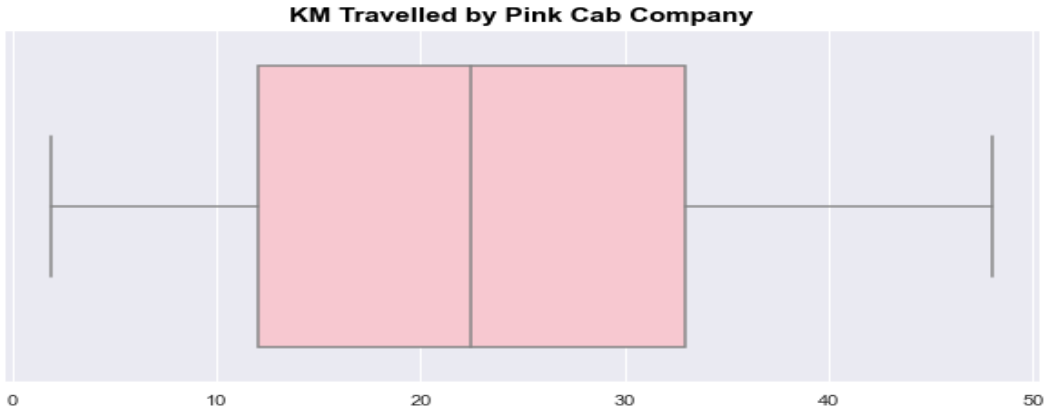


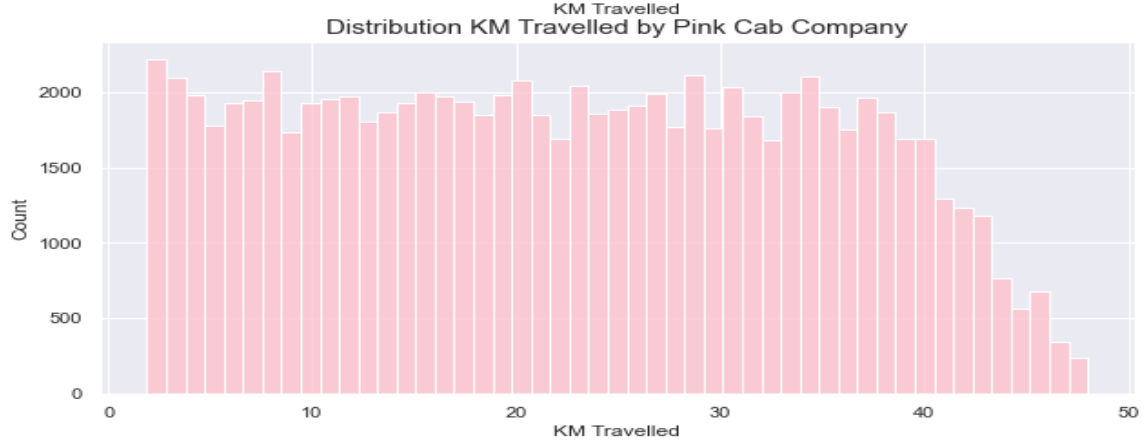
- For both companies, it is apparent that the most popular payment method is by card.
- This is likely due to more expensive fares requiring high amounts of cash that most customers do not feel comfortable carrying.

Distribution of Distance (KM) Travelled by Both Cab Companies



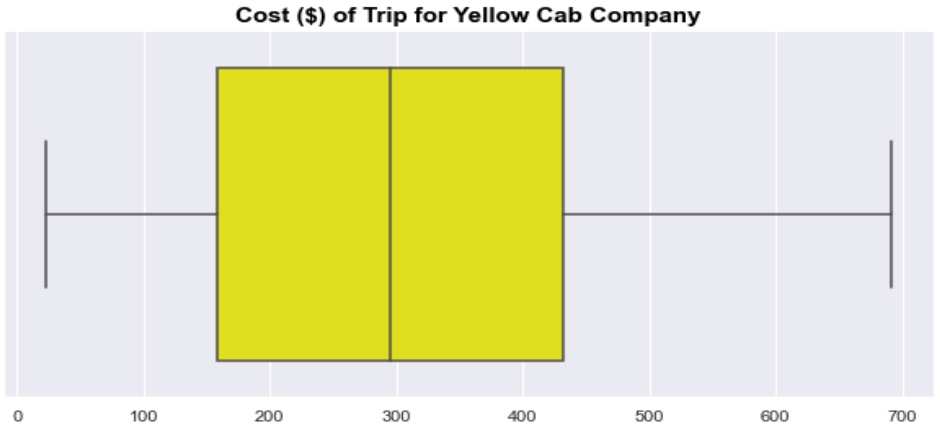


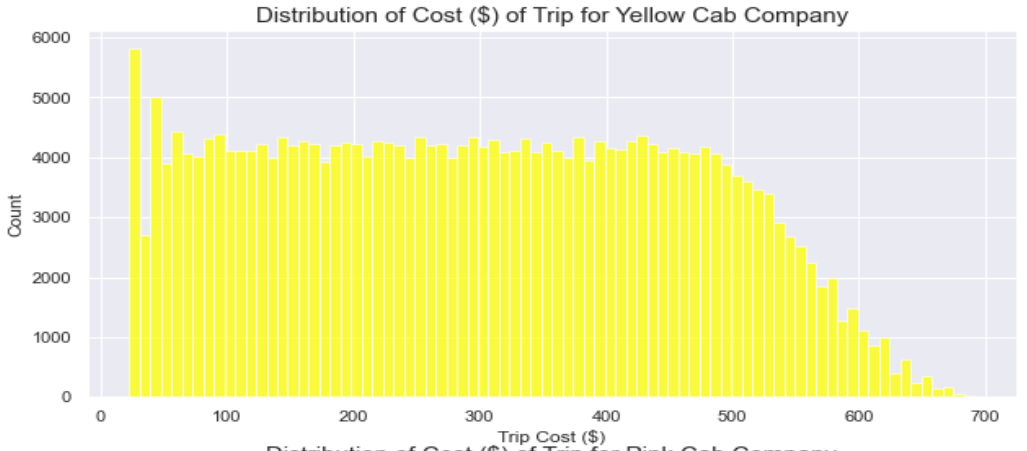


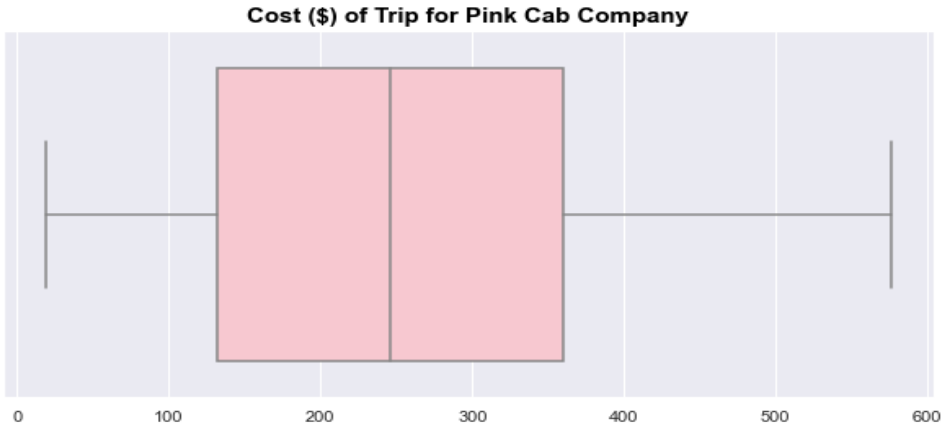


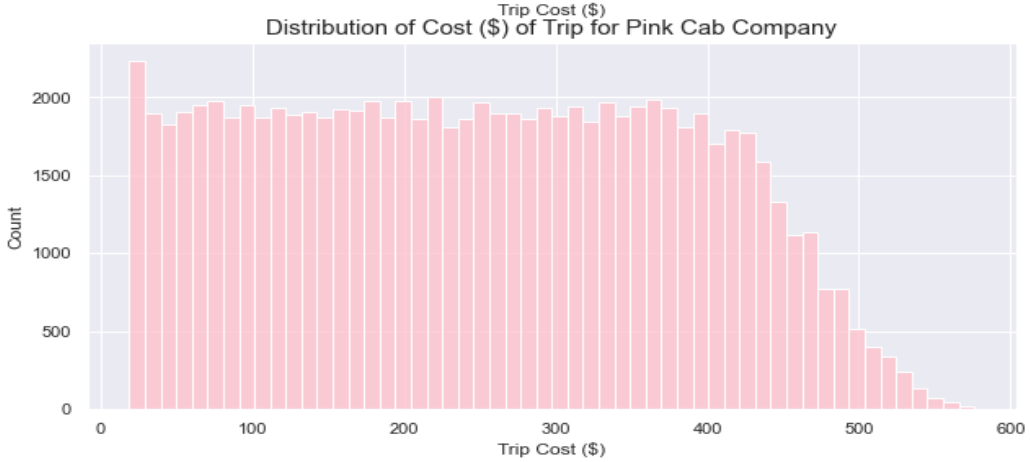
- For both cab companies, the range of traveling in KM per ride is approximately between ~2km and ~49km.
- The distributions appear to be very identical for the two companies with no outliers present in either company.
- However, the Yellow Cab Company has completed a lot more cab rides that are around or under ~4km when compared to other travel ranges for both companies.

Cost ($) of Trip for Both Cab Companies





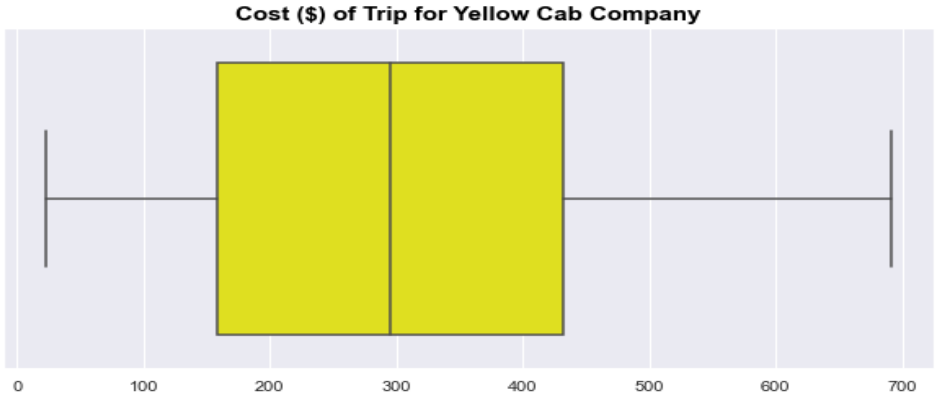


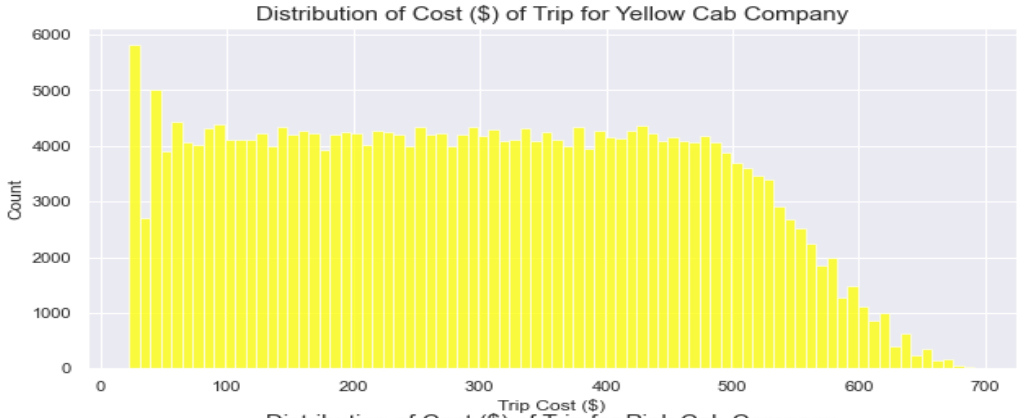


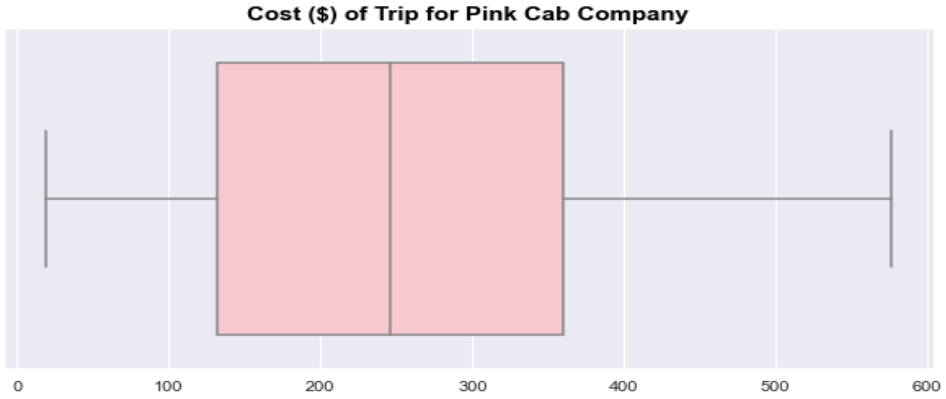
•The distribution of the Cost of trips are relatively identical for both companies.

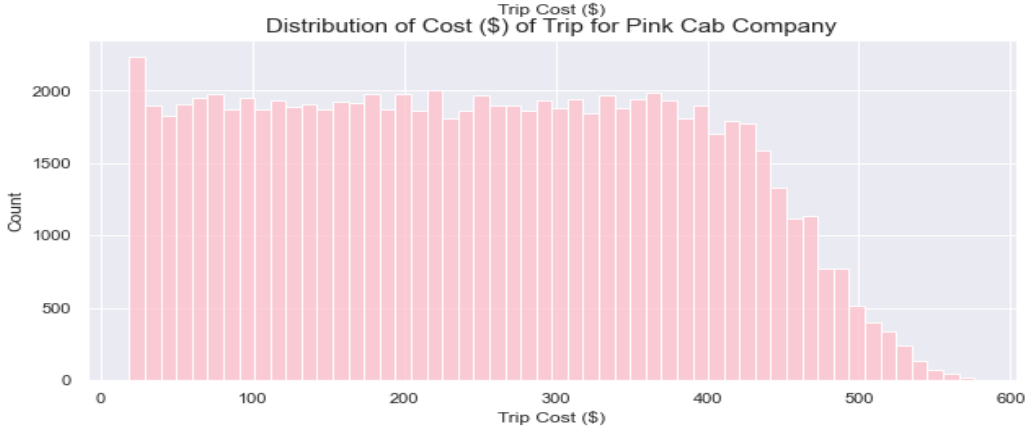
•The Yellow Cab Company has completed some trips that cost more than the Pink Cab Company. It is possible that these are longer trips or the Yellow Cab Company has more expensive and larger vehicles for transportation. We will explore this later and in more depth.

Age Distribution for Both Cab Companies



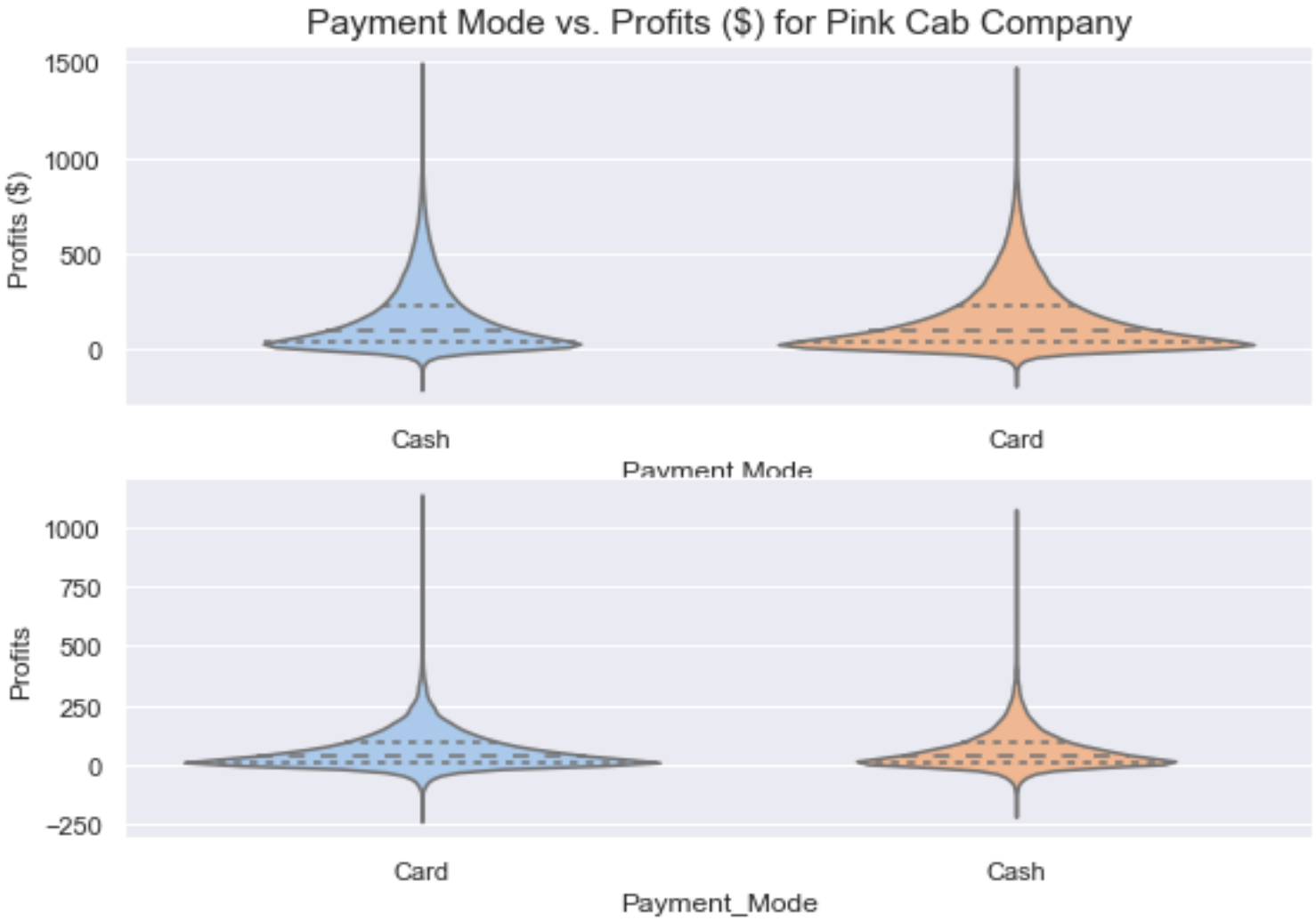






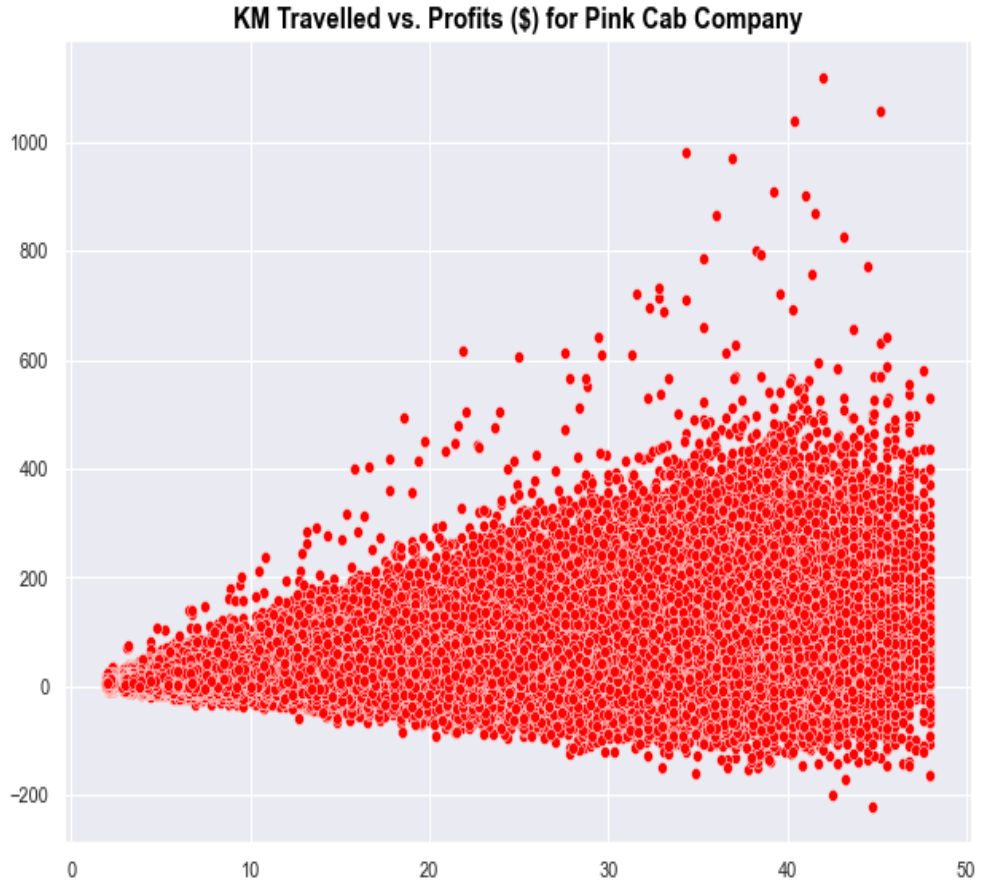
- For both companies, most customers are in their late-20s to early-40s in terms of age.
- Some cab riders are over 50+ but hardly any are below 20.
- These numbers make sense because seniors and young adults are unlikely to travel in cabs for safety reasons.
- Additionally, adults in their 20s and 40s are likely working and have more control/freedom with money.

Profits by Payment Method for Both Companies



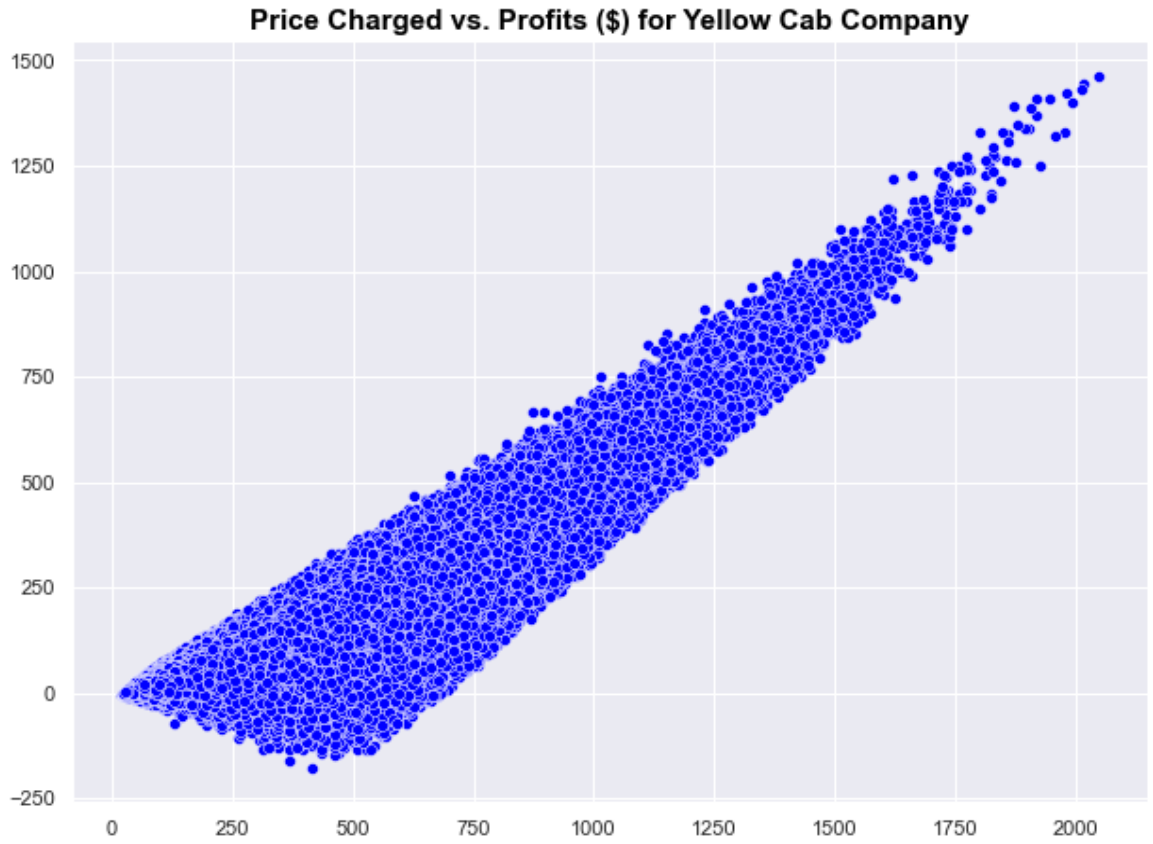
Both payment methods are used to pay for low and high profit cab rides equally. Using cards to pay for cab rides is a more popular choice especially when the fares are less than $100 for both companies.

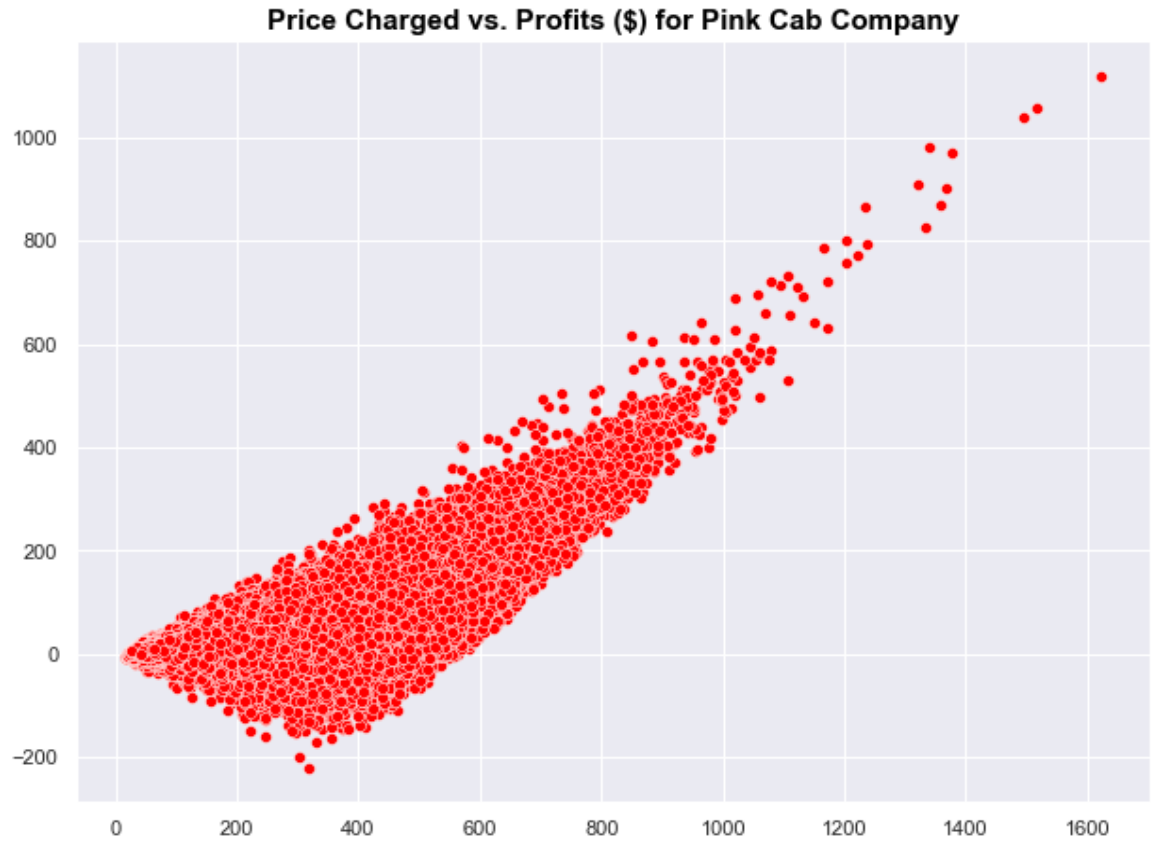
Distance (KM) Travelled vs. Profits ($)



Generally, as the distance travelled by a cab company for either of the cab companies, the profits of the transaction increases. There appears to be somewhat of a positive linear relationship between the distance traveled and profits earned.

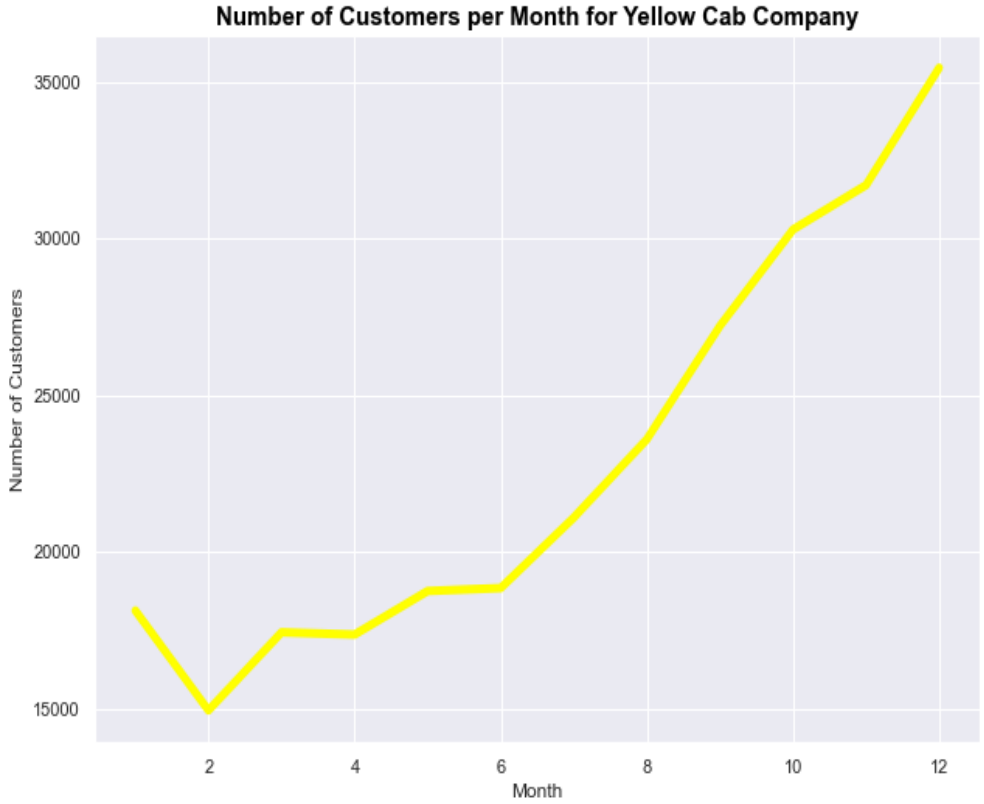
Price ($) Charged vs. Profits ($)

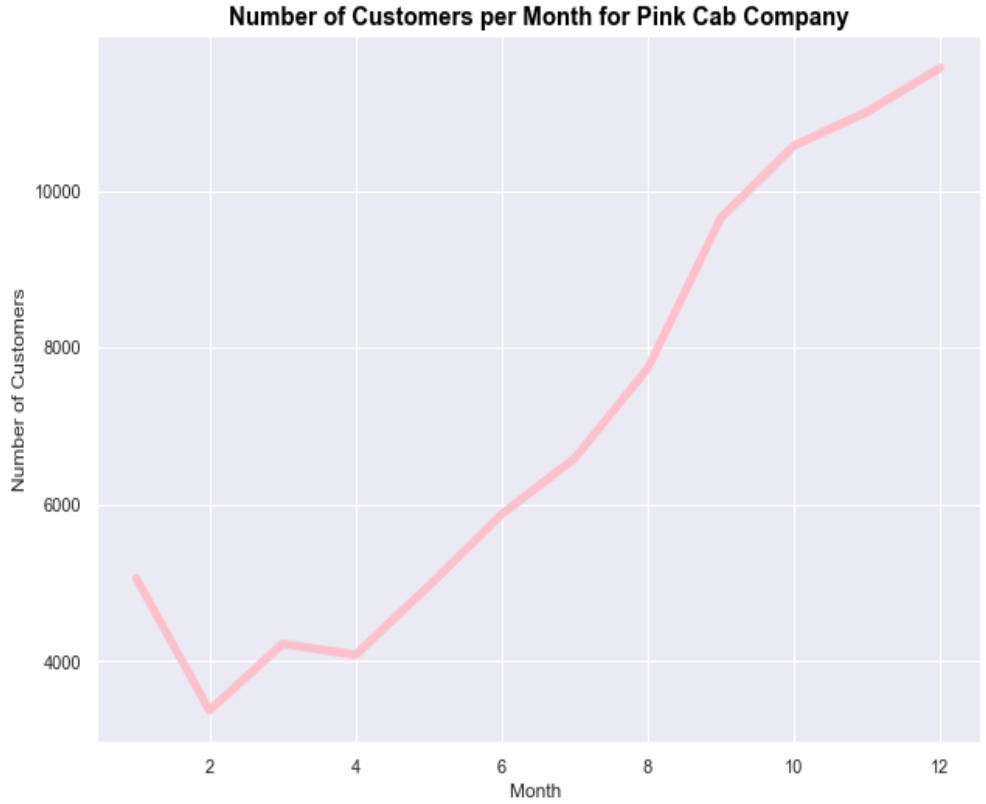




A **strong and positive linear relationship** exists between the price charged and profits for both cab companies. The **Yellow Cab Company** appears to have a **stronger linear relationship** than the Pink Cab Company.

Is there any seasonality in the number of customers using the cab service?



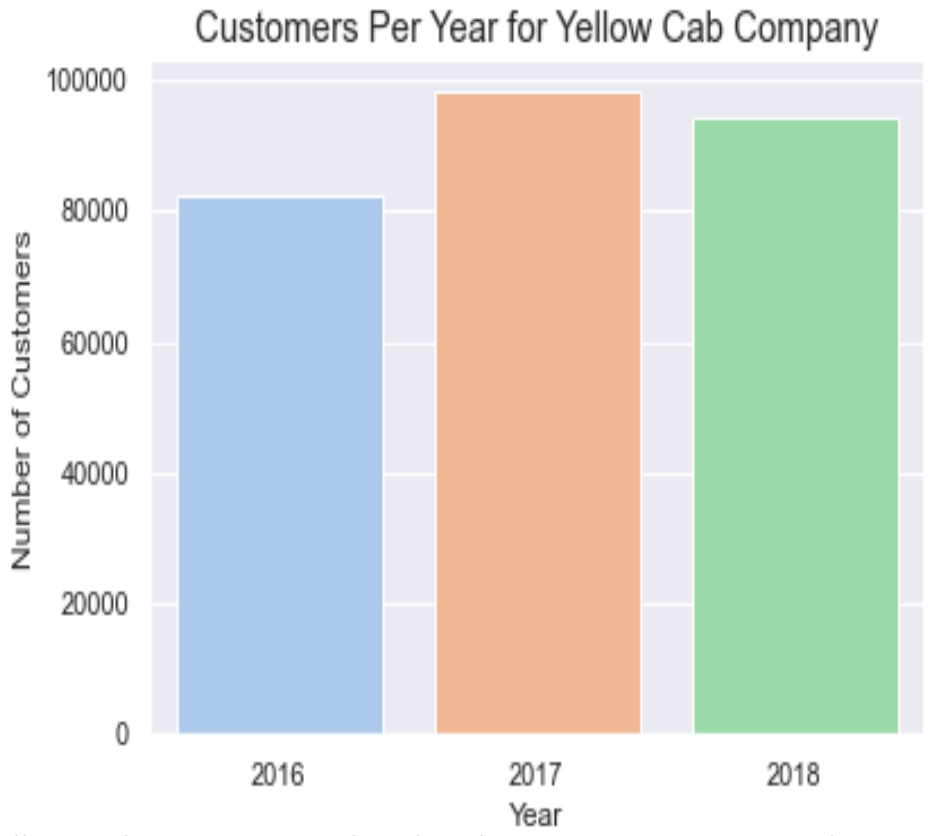


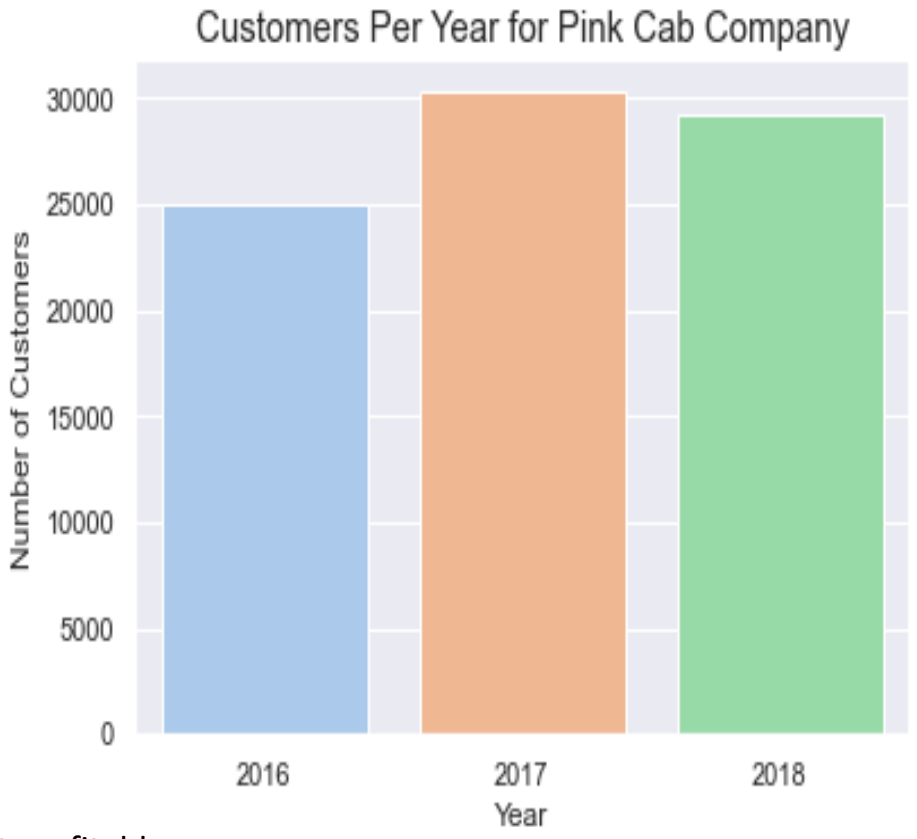
- Both companies perform well in the fall/winter seasons.
- This is because of the cooler weather during these months, which discourages people from walking, taking public transit or biking to destinations.
- Additionally, although January and February are cold months in the U.S.A., it seems the cab companies do not receive much business during these times of the year likely due to a lack of traveling throughout the peak winter seasons (in climate weather).
- The **Yellow Cab Company is drastically outperforming the Pink Cab Company** regardless of the month/season during the year.

Yearly Customer Retention

Year	Transaction_ID
2016	82239
2017	98189
2018	94253

Year	Profits
2016	169.347821
2017	168.817057
2018	143.416122





Year	Transaction_ID
2016	25080
2017	30321
2018	29310

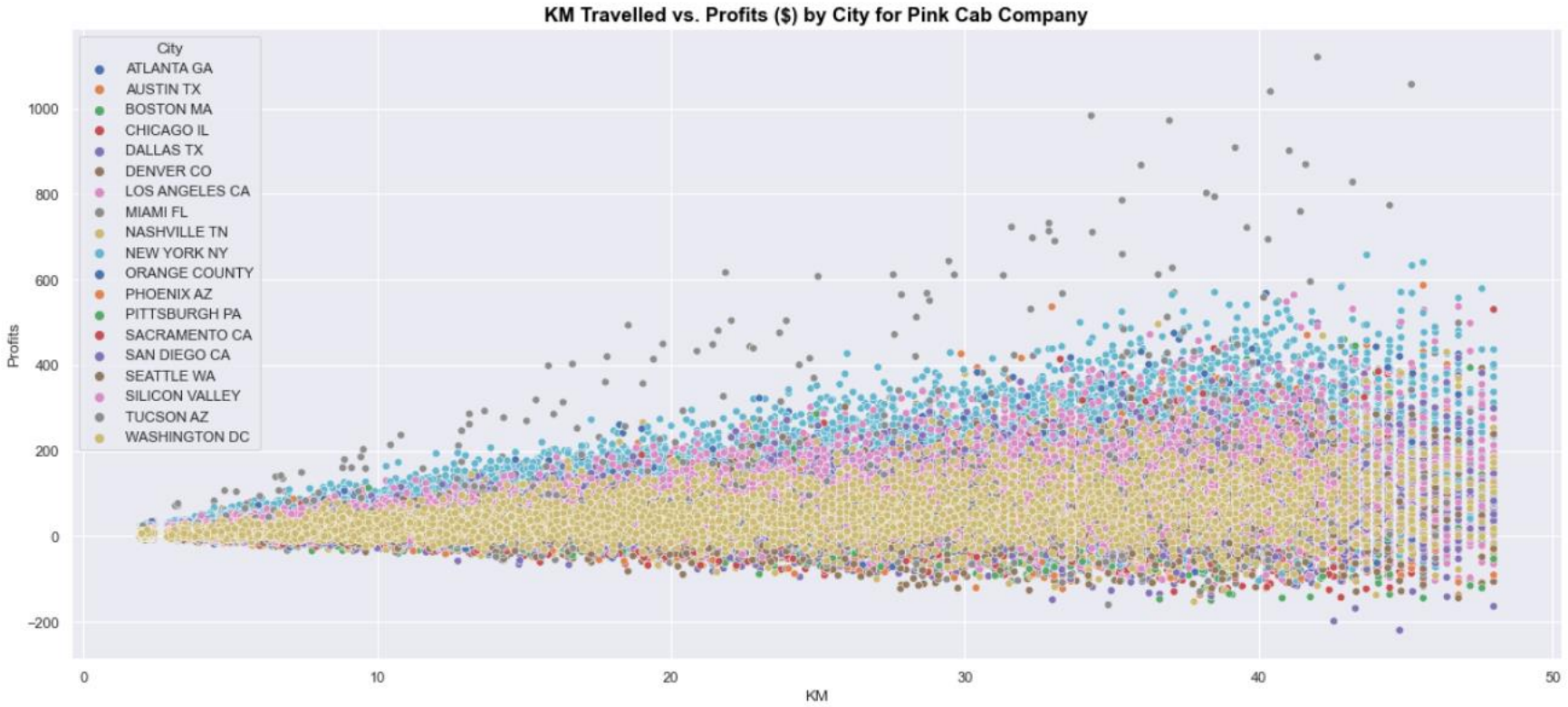
Year	Profits
2016	68.321819
2017	67.070839
2018	53.229689

- For both, the Yellow Cab Company and Pink Cab Company, 2016 was the most profitable year.
- Despite 2017 yielding the most customers for both companies, it was as profitable as the previous year. The cost of the trip is more than the company is charging.
- In 2018, both companies seem to be losing out on their average yearly profits because despite having more customers in 2018 than in 2016, the profits are noticeably lower.
- Both cab companies seem to be on a **decline** when it comes retaining profits.

Yellow Cab Company – What is the relationship between distance traveled in kilometers and profits earned for each city? Does profit increase for longer cab rides?

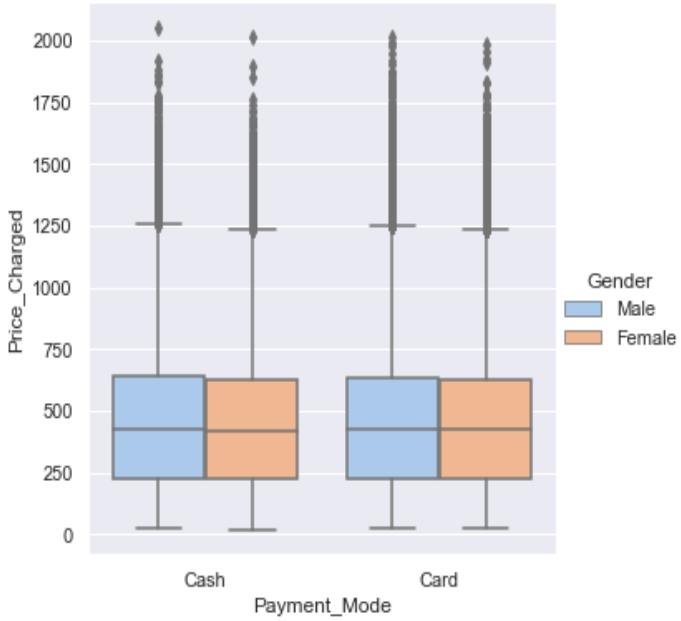
For the Yellow Cab Company, longer fairs with greater distance coverage yield larger profits for New York, Boston, and Orange County. Generally, the Yellow Cab Company profits well as the distance traveled increases.

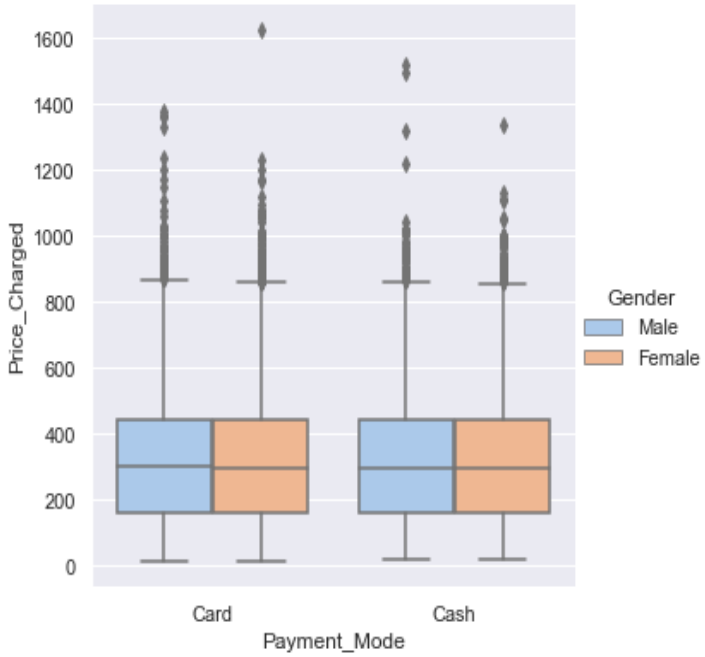
Yellow Cab Company – What is the relationship between distance traveled in kilometers and profits earned for each city? Does profit increase for longer cab rides?

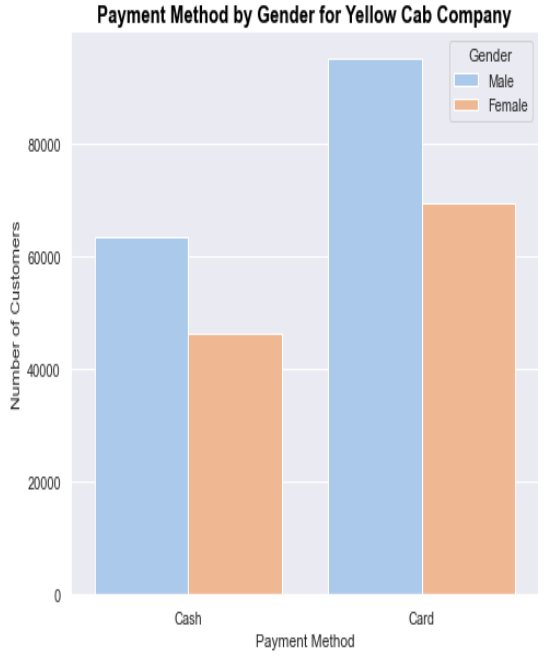


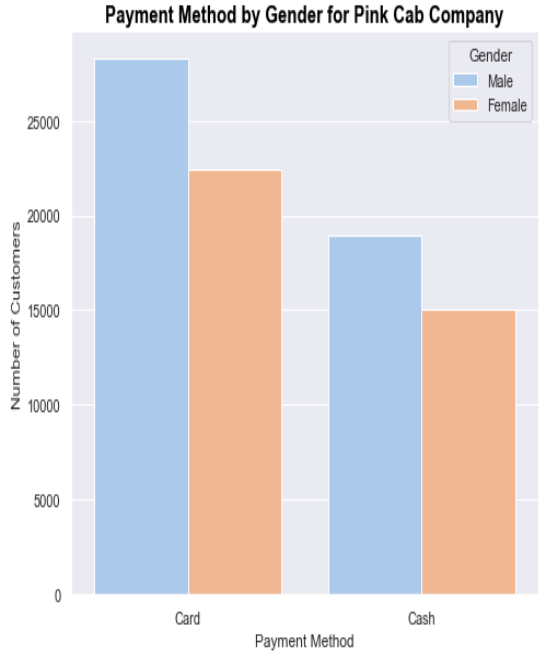
- For the Pink Cab Company, longer fairs with greater distance coverage yield substantially higher profits for Miami, FL. The distance traveled and profits for this city are linear with a steady increase in profit as distance of cab ride increases.
- For both cities, more distance traveled by a cab does not yield a larger profit for Sacramento, Washington, and Tucson. The profits seem to remain relatively stable with some decrease in profits as distance increases.

What's the most popular payment method by gender? Are card payments more expensive than cash payments?









- The most popular payment method regardless of gender is by card although it is not by much surprisingly. Usually, it is preferred to pay higher amounts of fees by card as opposed to cash because of the convenience of not carrying a lot of cash. This explains why card payments are more popular regardless of gender.
- For the Pink Cab Company, card payments are on average 1 dollar more, which is insignificant, but the there are over 17,000 more transactions paid with cards vs. cash.
- For the Yellow Cab Company, card payments are on average 1 dollar less, which is also insignificant, but there are over 54,000 more transactions paid with cards vs. cash.
- In general, both cab companies should be prepared to handle any payment method because the method a customer chooses to pay by is ambiguous.
- **Note:** Average price charged by the Yellow Cab Company is almost on average **$145 more** than the Pink Cab Company regardless of payment method.

Which company has more negative profits (margins)? What might be causing lower profits (margins)?

```
# Exploring negative profits for the Yellow Cab Company
negative_profits_yellow_cab = yellow_cab_df[yellow_cab_df['Profits'] < 0]
negative_profits_yellow_cab = negative_profits_yellow_cab[['Transaction_ID', 'Profits', 'Year']]
print(negative_profits_yellow_cab.head(5))
print('The shape of the negative profits yellow cab company dataframe frame is: ' + str(negative_profits_yellow_cab.shape))
print('The mean of the negative profits yellow cab company dataframe is: ' + str(negative_profits_yellow_cab.Profits.mean()))
print('The propostion of negative profits for the yellow cab company is: ' + str(negative_profits_yellow_cab.shape[0] / yellow_cab_df.shape[0]))
```

```
     Transaction_ID  Profits  Year
69         10281043 -10.5800  2017
87         10209556  -0.1788  2017
292        10400687  -0.4204  2018
547        10089005 -38.6820  2016
567        10216015 -16.0080  2017
The shape of the negative profits yellow cab company dataframe frame is: (13690, 3)
The mean of the negative profits yellow cab company dataframe is: -18.926115325054724
The propostion of negative profits for the yellow cab company is: 0.04983963215511812
```

Year	Profits
2016	-20.081730
2017	-20.839714
2018	-15.902013

```
# Exploring negative profits for Pink Cab Company
negative_profits_pink_cab = pink_cab_df[pink_cab_df['Profits'] < 0]
negative_profits_pink_cab = negative_profits_pink_cab[['Transaction_ID', 'Profits', 'Year']]
print(negative_profits_pink_cab.head(5))
print('The shape of the negative profits pink cab company dataframe frame is: ' + str(negative_profits_pink_cab.shape))
print('The mean of the negative profits pink cab company dataframe is: ' + str(negative_profits_pink_cab.Profits.mean()))
print('The propostion of negative profits for the pink cab company is: ' + str(negative_profits_pink_cab.shape[0] / pink_cab_df.shape[0]))
```

```
    Transaction_ID  Profits  Year
39        10031719  -10.870  2016
42        10347704   -0.186  2018
45        10266097  -10.220  2017
47        10319969  -12.550  2018
59        10136342  -14.770  2017
The shape of the negative profits pink cab company dataframe frame is: (11129, 3)
The mean of the negative profits pink cab company dataframe is: -20.367706802048666
The propostion of negative profits for the pink cab company is: 0.13137609047231175
```

Year	Profits
2016	-21.968708
2017	-21.823338
2018	-17.095356

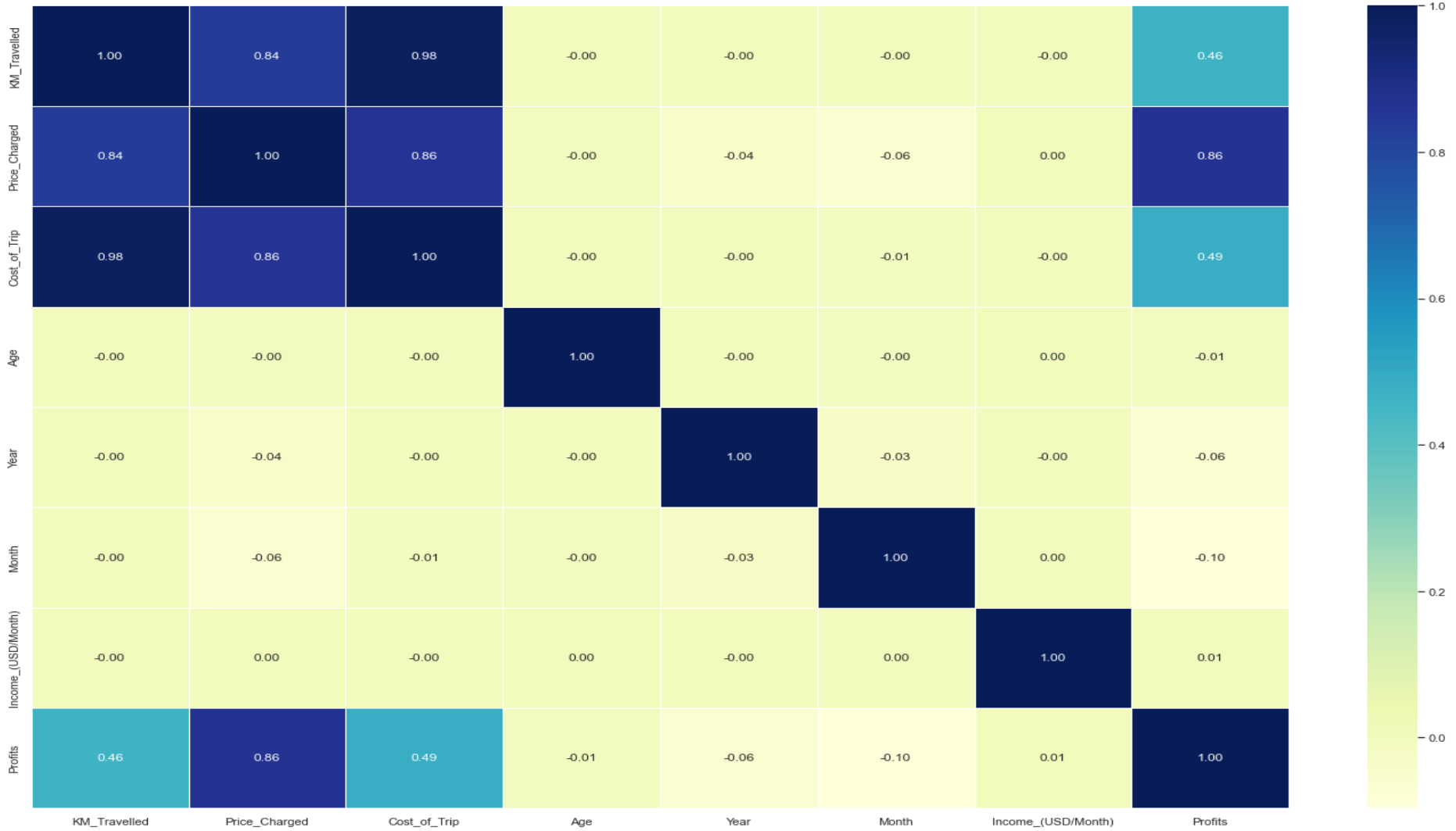
- The negative margins are caused due to some type of discount/token for the cab company or possibly poor service by the cab driver.

- The Pink Cab Company is suffering with lower average margins (-$20) than the Yellow Cab Company.

- The Yellow Cab Company has more customers who provided negative margins between 2016 to 2018, however, the proportion of negative profits to overall profits is much lower than that of the Pink Cab Company. Only about ~5% of The Yellow Cab Company's profits are negative whereas the 13% of the Pink Cab Company's profits are negative despite having less customers.

- The negative profits are on the decline from 2016 to 2018 for both companies, yet the Yellow Cab Company is still averaging less yearly negative profits than the Pink Cab Company.

Correlation Analysis



Inference from Correlation Analysis:

- There is a strong correlation between the Profits and Price Charged as expected.
- There is a moderate correlation between Profits and Kilometers Travelled.
- There is also a moderate correlation between Profits and Cost of Trip.
- There are strong correlations between Kilometers Travelled, Price Charged and Cost of Trip.

Hypothesis Testing: Do Linear Relationships Exists with Cab Company Margins (Profits ($)) and Cab Ride Features?

```
# Pearson's Correlation Test for Profits and Kilometers Traveled
stat, p = pearsonr(df['Profits'], df['KM_Travelled'])
print('stat=%.10f, p=%.10f' % (stat, p))

if p > 0.05:
    print('Retain the null hypothesis because there may not exist a linear relationship between the two variables (independence).')
else:
    print('Reject the null hypothesis because there likely exists a linear relationship between the two variables (dependence).')
```

```
stat=0.4627681979, p=0.0000000000
Reject the null hypothesis because there likely exists a linear relationship between the two variables (dependence).
```

```
# Pearson's Correlation Test for Profits and Price Charged
stat, p = pearsonr(df['Profits'], df['Price_Charged'])
print('stat=%.10f, p=%.10f' % (stat, p))

if p > 0.05:
    print('Retain the null hypothesis because there may not exist a linear relationship between the two variables (independence).')
else:
    print('Reject the null hypothesis because there likely exists a linear relationship between the two variables (dependence).')
```

```
stat=0.8641539468, p=0.0000000000
Reject the null hypothesis because there likely exists a linear relationship between the two variables (dependence).
```

```
# Pearson's Correlation Test for Profits and Cost of Trip
stat, p = pearsonr(df['Profits'], df['Cost_of_Trip'])
print('stat=%.10f, p=%.10f' % (stat, p))

if p > 0.05:
    print('Retain the null hypothesis because there may not exist a linear relationship between the two variables (independence).')
else:
    print('Reject the null hypothesis because there likely exists a linear relationship between the two variables (dependence).')
```

```
stat=0.4860560801, p=0.0000000000
Reject the null hypothesis because there likely exists a linear relationship between the two variables (dependence).
```

The Hypothesis Test:

- H0: The Target and Feature are independent (p = 0).

  vs.

- H1: The Target and Feature are dependent (p != 0).

- **There exists a linear relationship between all the moderate to strongly correlated features and target variable**.
- This further verifies that the Yellow Cab Company is the better company to invest in because distance travelled, cost of trip and price charged all have a linear relationship with the profits/margins per trip.
- Since the Yellow Cab Company takes more customers on longer cab rides with higher prices charged, the Yellow Cab Company is also earning more profits than the Pink Cab Company.
- Due to multicollinearity, we will have to remove the features that correlate strongly with each other to improve modeling.

Recommendations

We have evaluated both the cab companies on following points and found Yellow cab better than Pink cab:

- **Popularity and Cities:** The Yellow Cab Company is more popular regardless of income level and has more transactions between 2016 to 2018. The Yellow Cab Company is more popular regardless of age group because the distribution of age groups is identical between both companies, yet the Yellow Cab Company has over 3x as many customers. This is the same case for income groups too. The Yellow Cab Company is performing better in every single city, including the popular cities (e.g. New York), both companies are present in with exception to Miami, Florida, which is the only city the Pink Cab Company outperforms the Yellow Cab Company.

- **Overall Profits:** The Yellow Cab Company is earning more profits (margins) on average between 2016 to 2018 with significantly higher margins for outlier transactions in the Yellow Cab Company vs. Pink Cab Company.

- **Yearly Customer Retention:** The Yellow Cab Company has seen a decline in average profits per year from 2016 to 2018 but it is overall still more and less detrimental than the loss that the Pink Cab Company is suffering from.

- **Forecasts of Profits:** The Yellow Cab Company is suffering less from negative profits when compared to the Yellow Cab Company when comparing their average negative profits and proportion of negative profits.

- **Profits for Long Distance Rides:** The Yellow Cab Company is earning more profits as the distance travelled by the cab increases (positive linear relationship) regardless of the city.

- **Losses (Negative Margins):** The Yellow Cab Company is suffering less from negative profits when compared to the Yellow Cab Company when comparing their average negative profits and proportion of negative profits.

Therefore, the XYZ Investment Firm should strongly consider investing in the **Yellow Cab Company** if they had to chose one company over the other. However, it is important to note that both companies have been declining in margins from 2016 to 2018. It might be wise to hold off investments and wait for more recent data to better understand the trends and trajectories for the future of the company.

Thank You