Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

G2M insight for Cab Investment firm

9 August 2021

Agenda

Approach and Background

EDA & EDA Summary

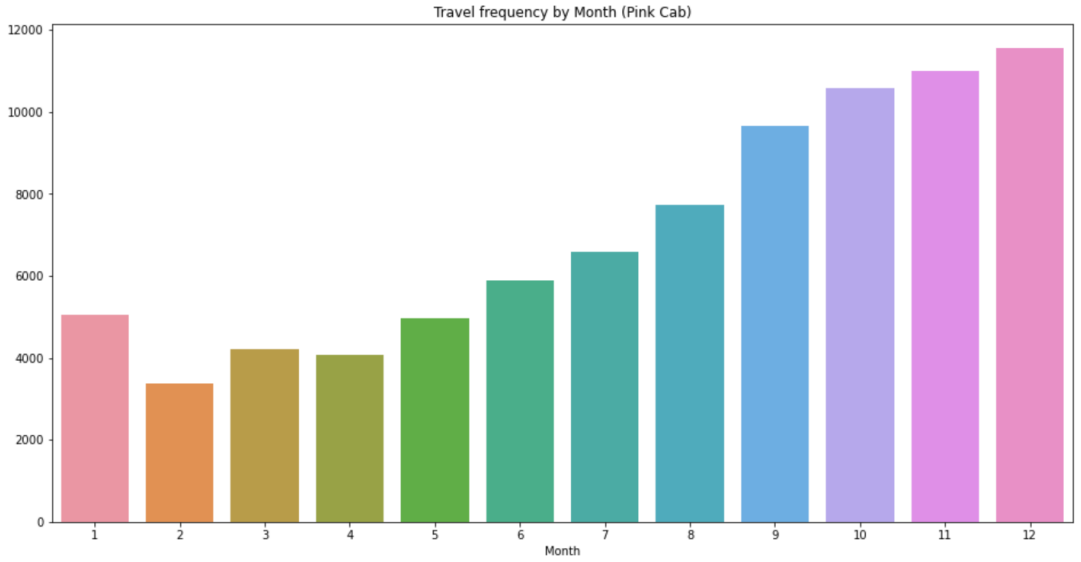
Hypothesis testing

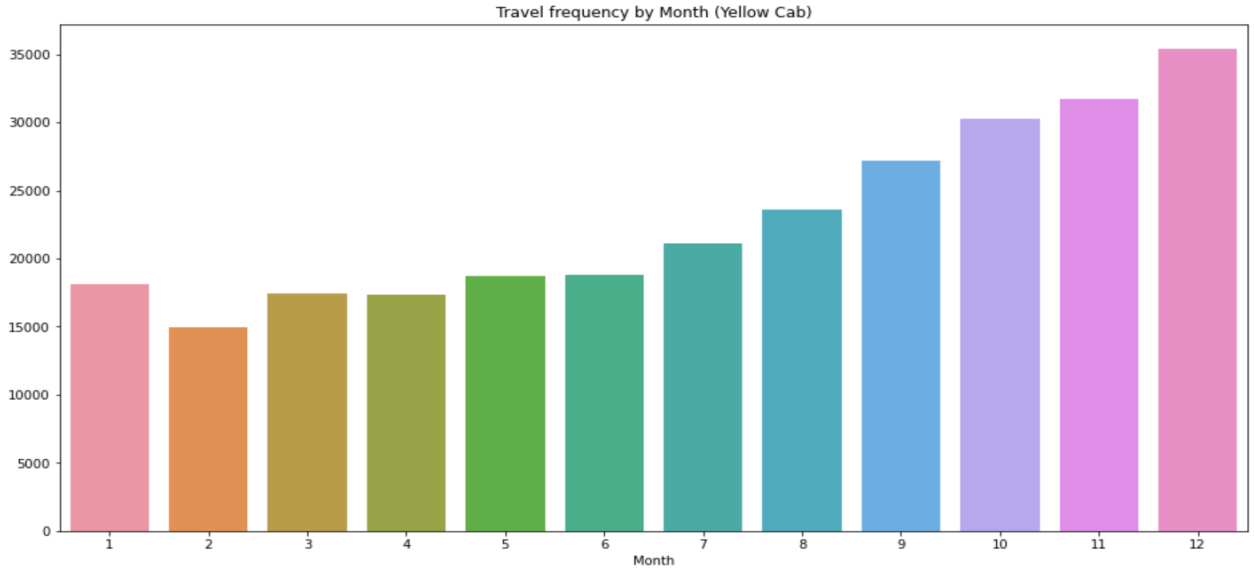
Recommendations

Approach and Background

- XYZ is a private firm in US and 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. We must summarize our analysis and recommendations and identify which company is performing better and is a better investment opportunity for XYZ.
- The datasets were combined when necessary to create datasets to do the analysis.

TRAVEL FREQUENCY

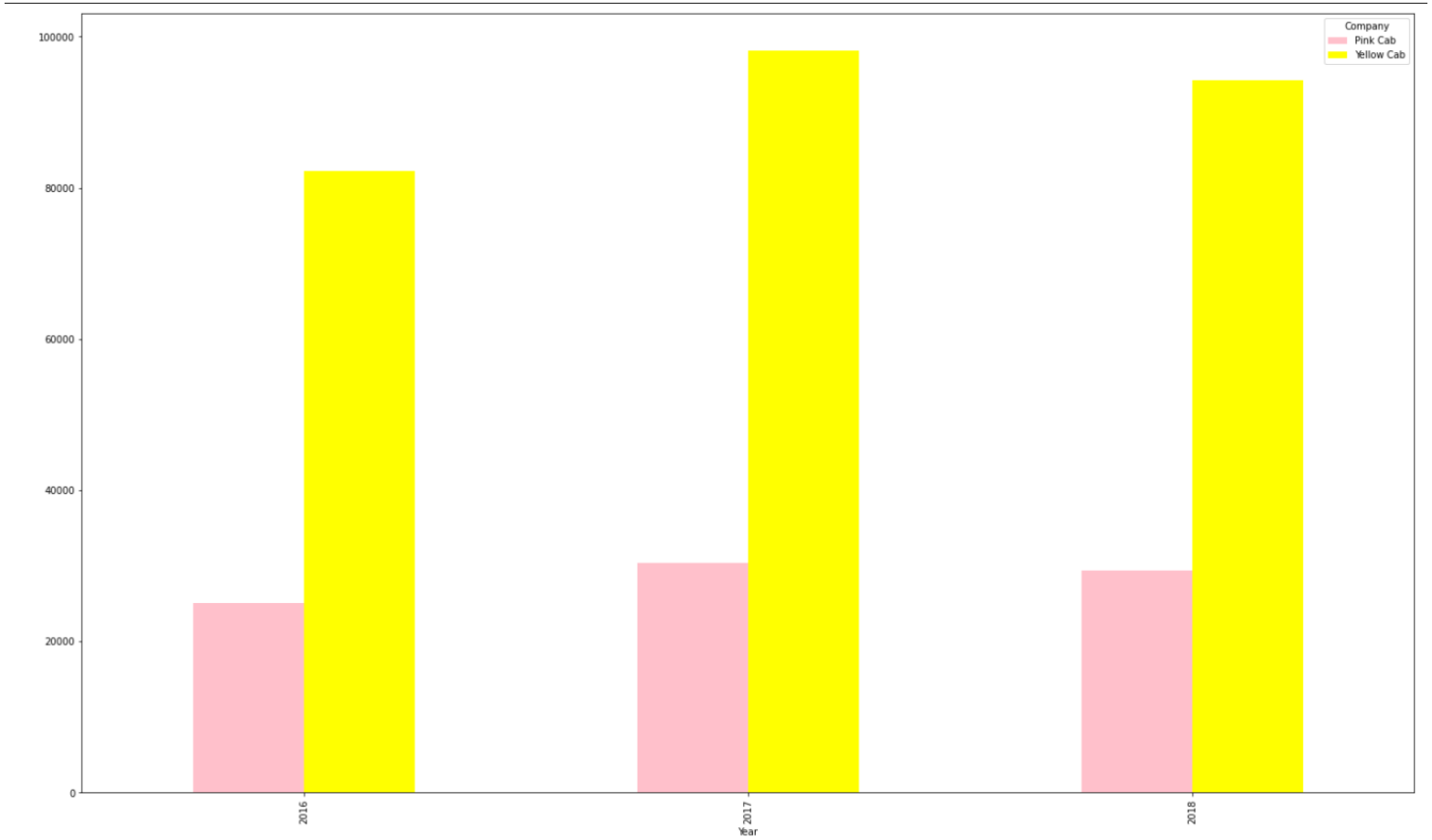




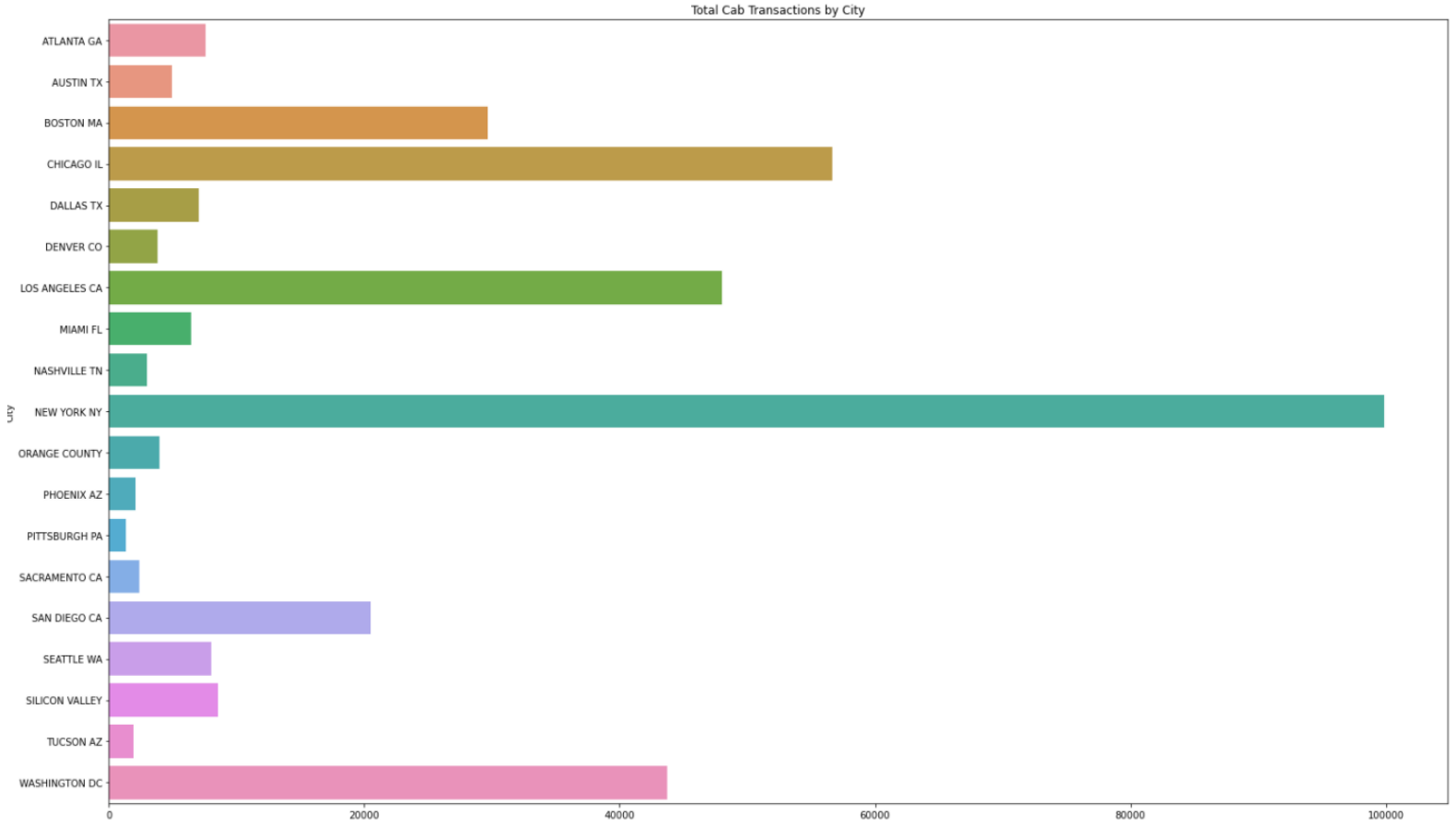
Here we can see that the travel frequency per month is more for Yellow when compared to Pink Cab. Specially towards the end of year in the months of November and December.

Yearly Usage

We can see that over the years, Yellow taxi has dominated in terms of usage but in 2018 dropped when compared to 2017.

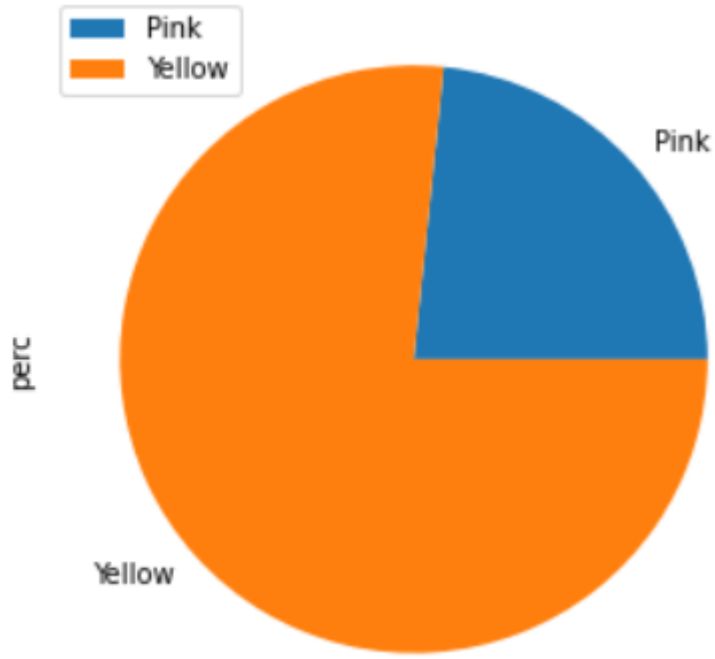


Total transactions divided by City

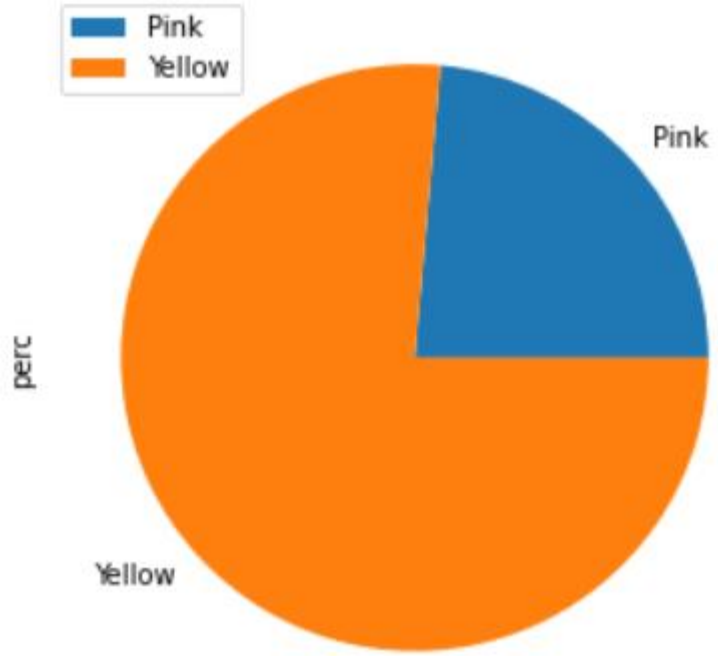


So, we can see that New York city was the city with most transaction(rides) comparatively.

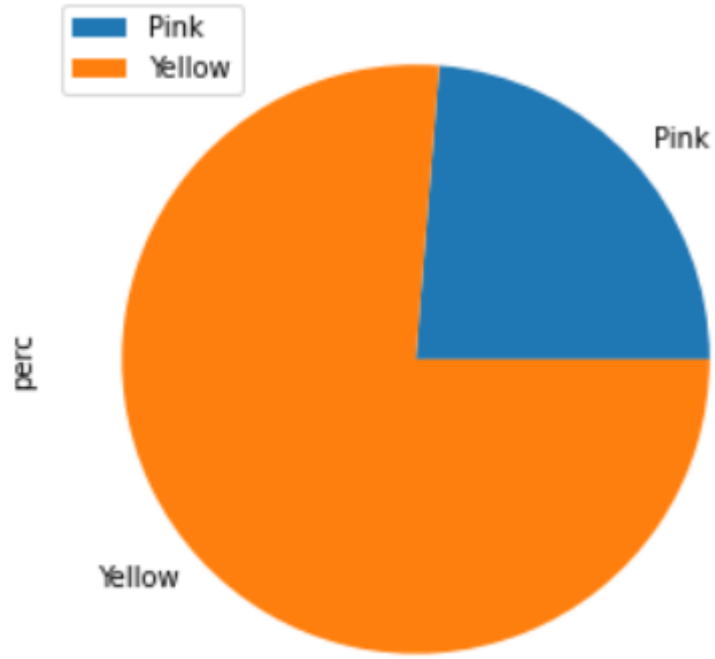
Percentage usage over the years



2016



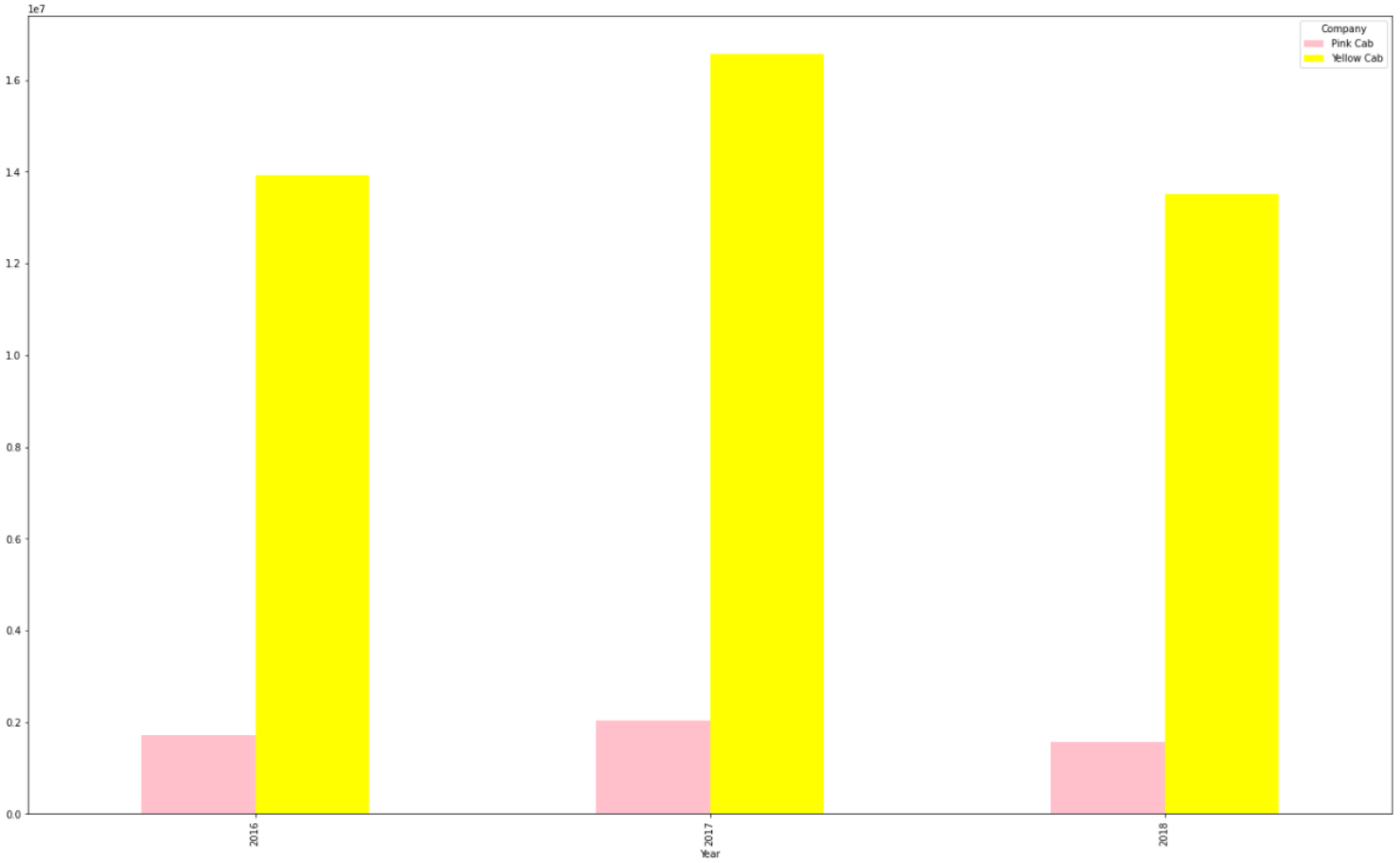
2017



2018

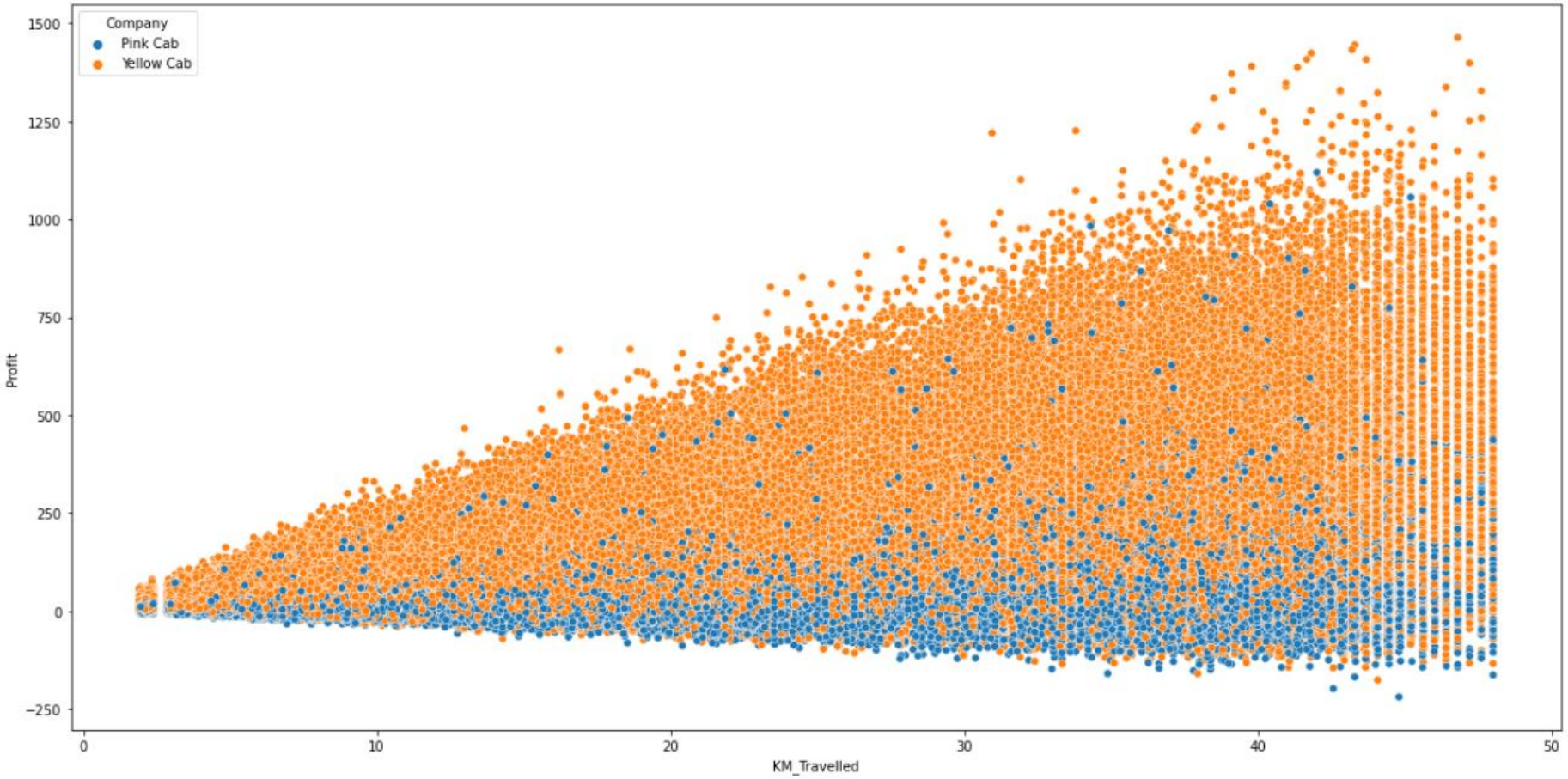
over the years, the ratio remained almost same for the companies usage.

Profit share per year



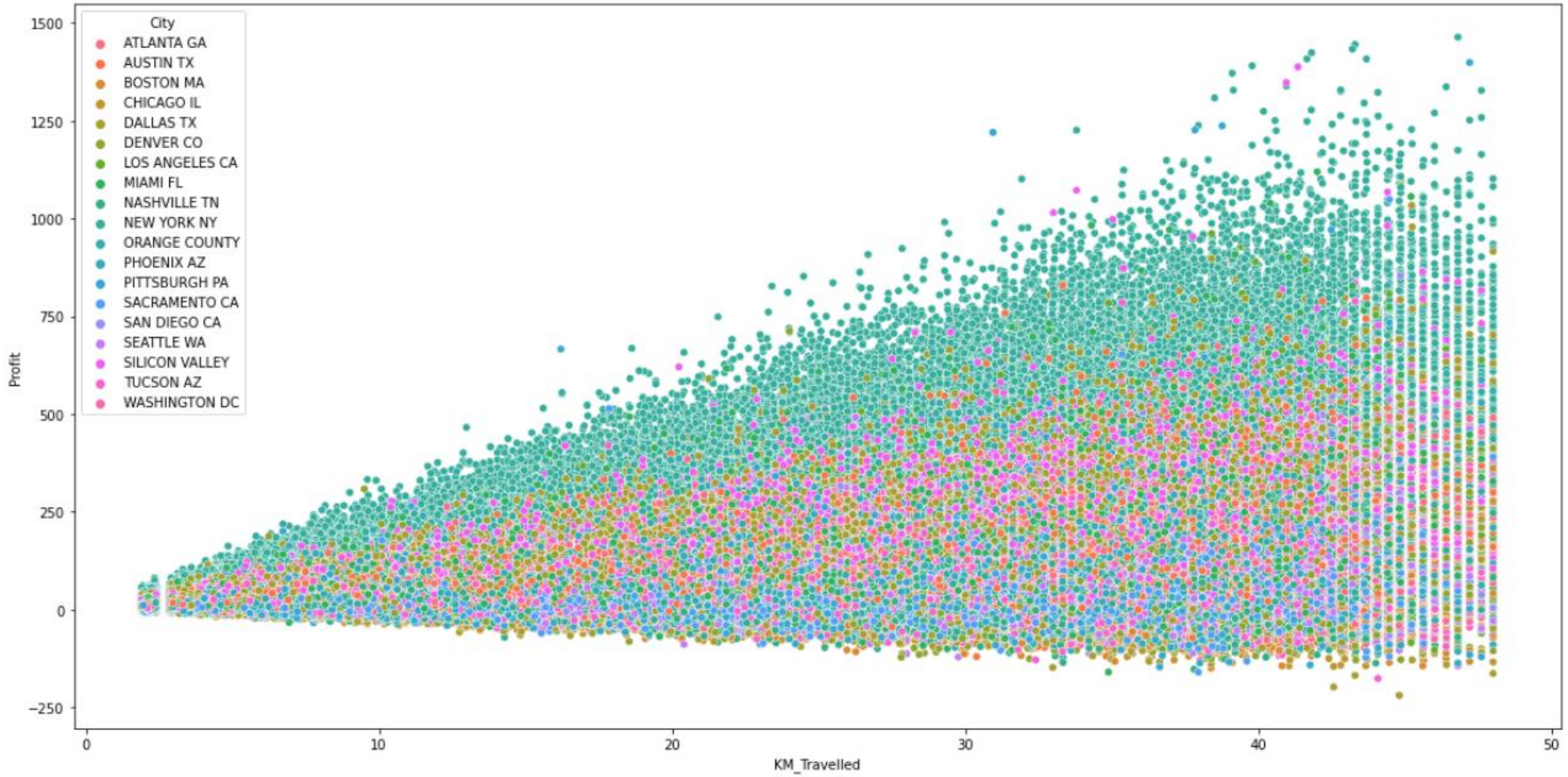
Profit share has been dominated over the years by yellow cab

Relationship between Profit and Distance travelled



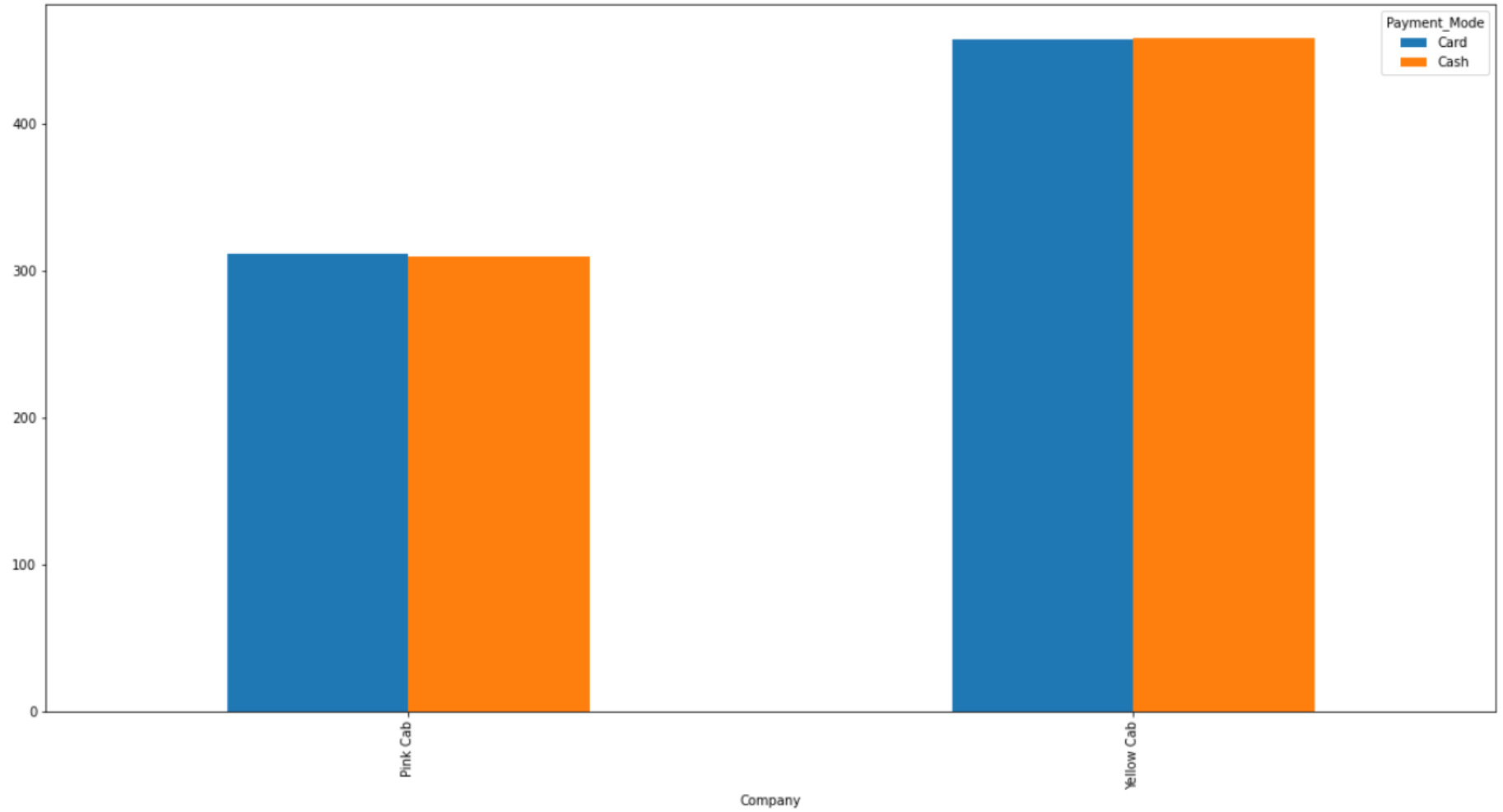
From the graph we can see that there is linear relationship between profit and distance for Yellow cab.

Relationship between Profit and Distance travelled per city



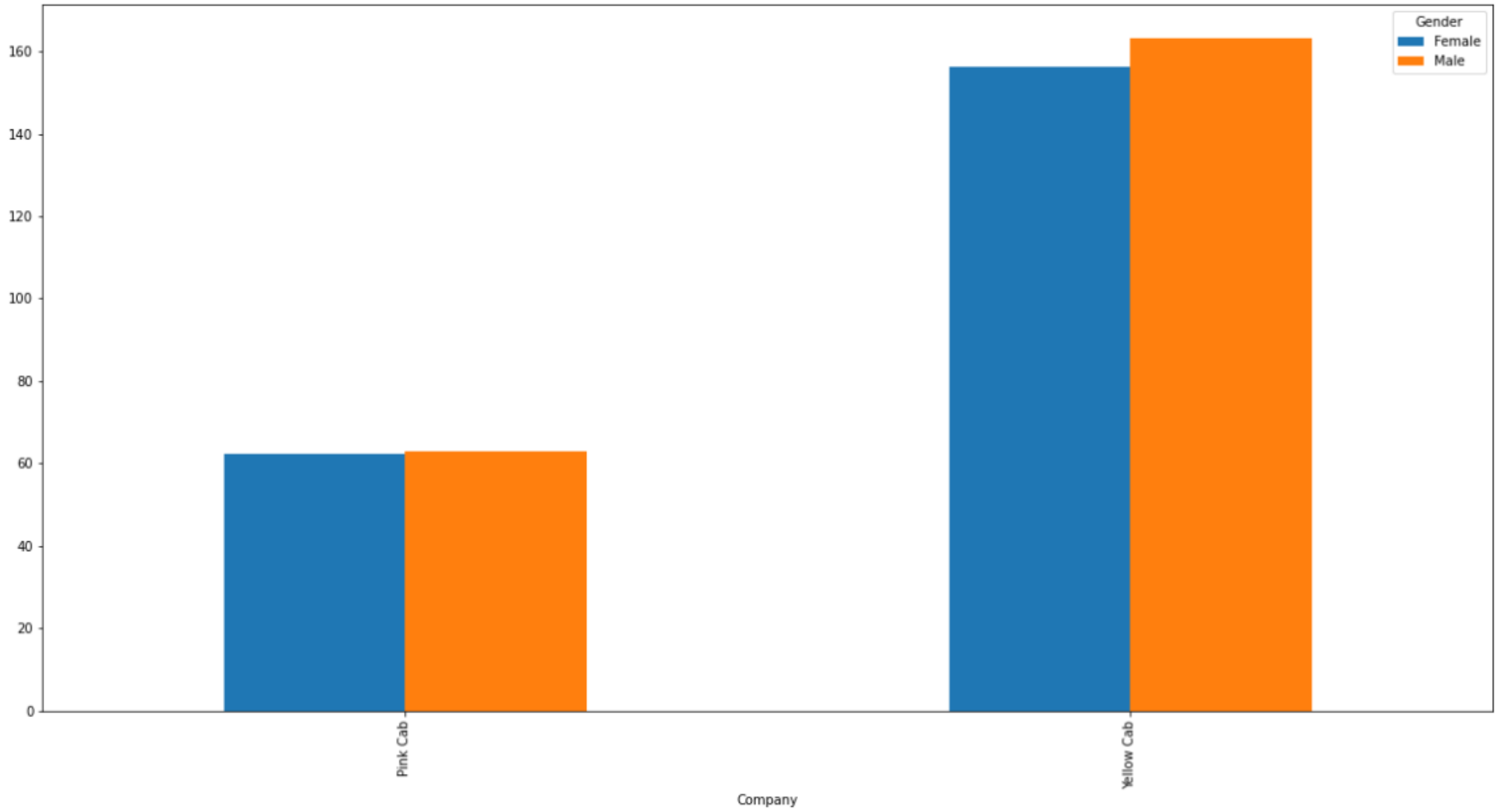
New York city has the highest profit as the KMs increased.

Type of payment for the different companies

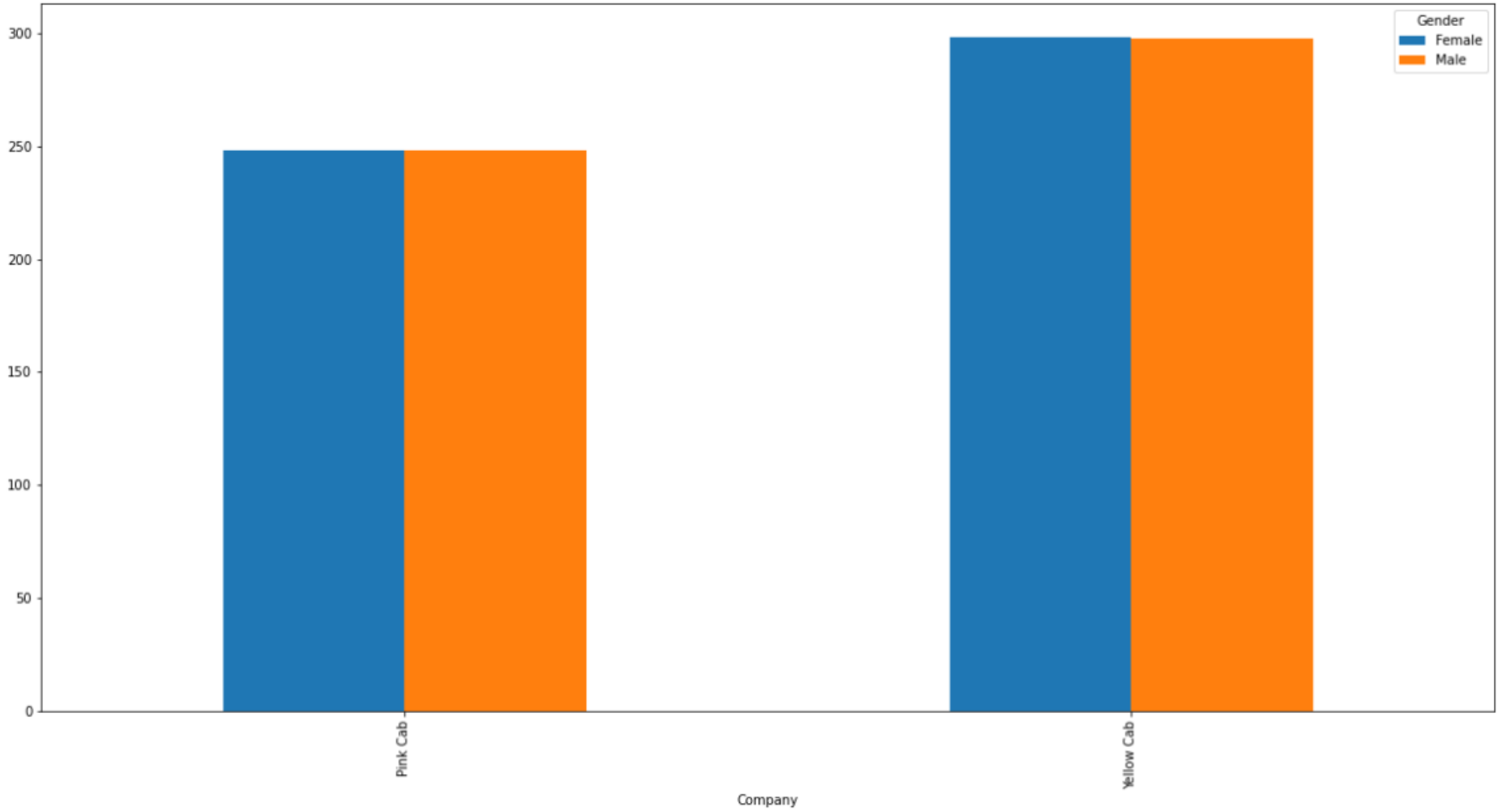


Both the companies recieved payments equally in card and cash.

Profit generation based on gender across the two companies



Cost of trip for the different genders across the two companies



The cost of trip is same for different genders in pink cab and almost same in yellow cab.

Hypothesis testing

Hypothesis 1:

Null: There is difference in number of people who pay by Card and Cash for yellow cab. Alternate: There is no difference in number of people who pay by Card and Cash for yellow cab.

```
m = merged2[['Company','Price_Charged','Payment_Mode']]
c = m[m['Company'] == 'Pink Cab']
cash_pink = c[c['Payment_Mode'] == 'Cash']
card_pink = c[c['Payment_Mode'] == 'Card']
```

```
stats.ttest_ind(cash_pink.Price_Charged,card_pink.Price_Charged)
```

35]: Ttest_indResult(statistic=-0.7744750456153074, pvalue=0.4386520812775526)

We accept the null hypothesis as p value is > 0.05.

Hypothesis testing

Hypothesis 2:

Null: There is difference in number of people who pay by Card and Cash for pink cab. Alternate: There is no difference in number of people who pay by Card and Cash for pink cab.

```
d = m[m['Company'] == 'Yellow Cab']
```

```
cash_yellow = d[d['Payment_Mode'] == 'Cash']
card_yellow = d[d['Payment_Mode'] == 'Card']
```

```
stats.ttest_ind(cash_yellow.Price_Charged,card_yellow.Price_Charged)
```

```
[88]: Ttest_indResult(statistic=0.5719506609006757, pvalue=0.5673558691551936)
```

We accept the null hypothesis as p value is > 0.05.

Hypothesis testing

Hypothesis 3:

Null: males bring in less profits than females for Yellow Cab. Alternate: males bring in more profits than females for Yellow Cab.

```
female_profit_data = merged2[(merged2.Gender=='Female')&(merged2.Company=='Yellow Cab')].groupby('Transaction_ID').Profit.mea
male_profit_data = merged2[(merged2.Gender=='Male')&(merged2.Company=='Yellow Cab')].groupby('Transaction_ID').Profit.mean()

stats.ttest_ind(female_profit_data,male_profit_data)
```

```
01]: Ttest_indResult(statistic=-10.31549420719532, pvalue=6.060473042494144e-25)
```

We accept the alternate hypothesis as p value is < 0.05.

Hypothesis testing

Hypothesis 4:

Null: Females bring in different profits as Males for pink Cab. Alternate: Females bring in same profits than Males for pink Cab.

```
female_profit_data1 = merged2[(merged2.Gender=='Female')&(merged2.Company=='Pink Cab')].groupby('Transaction_
male_profit_data1 = merged2[(merged2.Gender=='Male')&(merged2.Company=='Pink Cab')].groupby('Transaction_ID')

stats.ttest_ind(female_profit_data1,male_profit_data1)
```

```
00]: Ttest_indResult(statistic=-1.5754642478511207, pvalue=0.11515305900425798)
```

We accept the null hypothesis as p value is > 0.05.

Hypothesis 5:

Hypothesis testing

Hypothesis 5:

Null: Cost of trip for male and female is same for pink cab Alternate: Cost of trip for male and female is different for pink cab

```
female_cost_data_pink = merged2[(merged2.Gender=='Female')&(merged2.Company=='Pink Cab')].groupby('Transaction_ID').Cost_of_T
male_cost_data_pink = merged2[(merged2.Gender=='Male')&(merged2.Company=='Pink Cab')].groupby('Transaction_ID').Cost_of_Trip.

stats.ttest_ind(female_cost_data_pink,male_cost_data_pink)
```

```
95]: Ttest_indResult(statistic=0.5657091706365529, pvalue=0.5715929984922214)
```

We accept the null hypothesis as p value is > 0.05.

Hypothesis testing

Hypothesis 6:

Null: Cost of trip for male and female is different for Yellow cab Alternate: Cost of trip for male and female is same for Yellow cab

```
female_cost_data_yellow = merged2[(merged2.Gender=='Female')&(merged2.Company=='Yellow Cab')].groupby('Transac
male_cost_data_yellow = merged2[(merged2.Gender=='Male')&(merged2.Company=='Yellow Cab')].groupby('Transactior

stats.ttest_ind(female_cost_data_yellow,male_cost_data_yellow)
```

```
)6]: Ttest_indResult(statistic=0.948897235613972, pvalue=0.3426737155073213)
```

We accept the null hypothesis as p value is > 0.05.

Recommendations

- Through the research and analysis done, I would recommend to invest in Yellow cab company as it has huge potential and has also dominated over the years in terms of usage and profit generation too.

Thank You