



**Data Glacier**

Your Deep Learning Partner

# Exploratory Data Analysis

G2M Case Study

15/July/2022

# Background of Study

---



## Who are we studying?

XYZ is a private equity firm in the US.

The main point of this study is to create a feasible investment opportunity within the Cab Industry.

The two Cab Companies at focus are Yellow and Pink Cab co.



## Objective:

How can we create valuable insight for which Cab Company that XYZ should invest in?

# Agenda

---

- How will this presentation be organized?
  - Exploratory Data Analysis (Data Format/Manipulation)
  - Spatial Consumer Density
  - Consumer Characteristics
  - Profit Analysis/forecasting
  - Recommendations for investment opportunity

# Data Manipulation

- In total, 4 .CSV files were used.
  - Cab\_Data.csv
  - City.csv
  - Customer\_ID.csv
  - Transaction\_ID.csv
- All files were merged and joined to create a final DataSet

	Index	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip	Customer ID	Payment_Mode	Gender	Age	Income (USD/Month)
0	1	2018-07-21	Yellow Cab	ATLANTA GA	26.19	598.70	317.4228	29290	Cash	Male	28	10813
1	2	2018-11-23	Yellow Cab	ATLANTA GA	42.55	792.05	597.4020	29290	Card	Male	28	10813
2	4	2018-04-21	Yellow Cab	ATLANTA GA	36.38	721.10	467.1192	27703	Card	Male	27	9237
3	5	2018-05-04	Yellow Cab	ATLANTA GA	6.18	138.40	87.5088	27703	Cash	Male	27	9237
4	8	2016-09-21	Yellow Cab	ATLANTA GA	39.60	704.30	494.2080	28712	Card	Male	53	11242
...	...	...	...	...	...	...	...	...	...	...	...	...
274676	440092	2018-01-07	Yellow Cab	SEATTLE WA	16.66	261.18	213.9144	38520	Card	Female	42	19417
274677	440093	2018-01-03	Yellow Cab	SILICON VALLEY	13.72	277.97	172.8720	12490	Cash	Male	33	18713
274678	440095	2018-01-04	Yellow Cab	TUCSON AZ	19.00	303.77	232.5600	41414	Card	Male	38	3960
274679	440096	2018-01-06	Yellow Cab	TUCSON AZ	5.60	92.42	70.5600	41677	Cash	Male	23	19454
274680	440097	2018-01-04	Yellow Cab	TUCSON AZ	13.30	244.65	180.3480	39761	Card	Female	32	10128

## Data Format

# of Features: 11

# of Observations: 359,392

Time frame: 31/01/2016 – 31/12/2018

# Assumptions



## Dedupe Validation:

Every observation will be checked for duplication in an iterative manner. If any arise, they will be dropped.



Our data is not skewed. (No outliers)



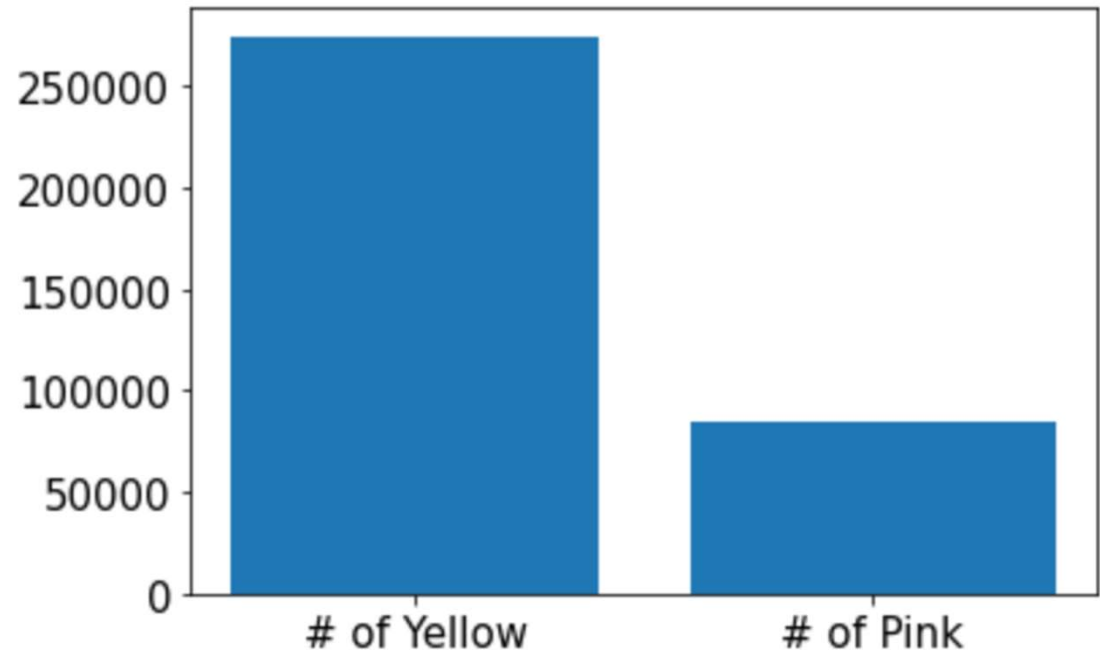
All missing data will be dropped from the dataset.



An ongoing calculation of profit will be the difference between cost of trip and price charged.

## Company EDA

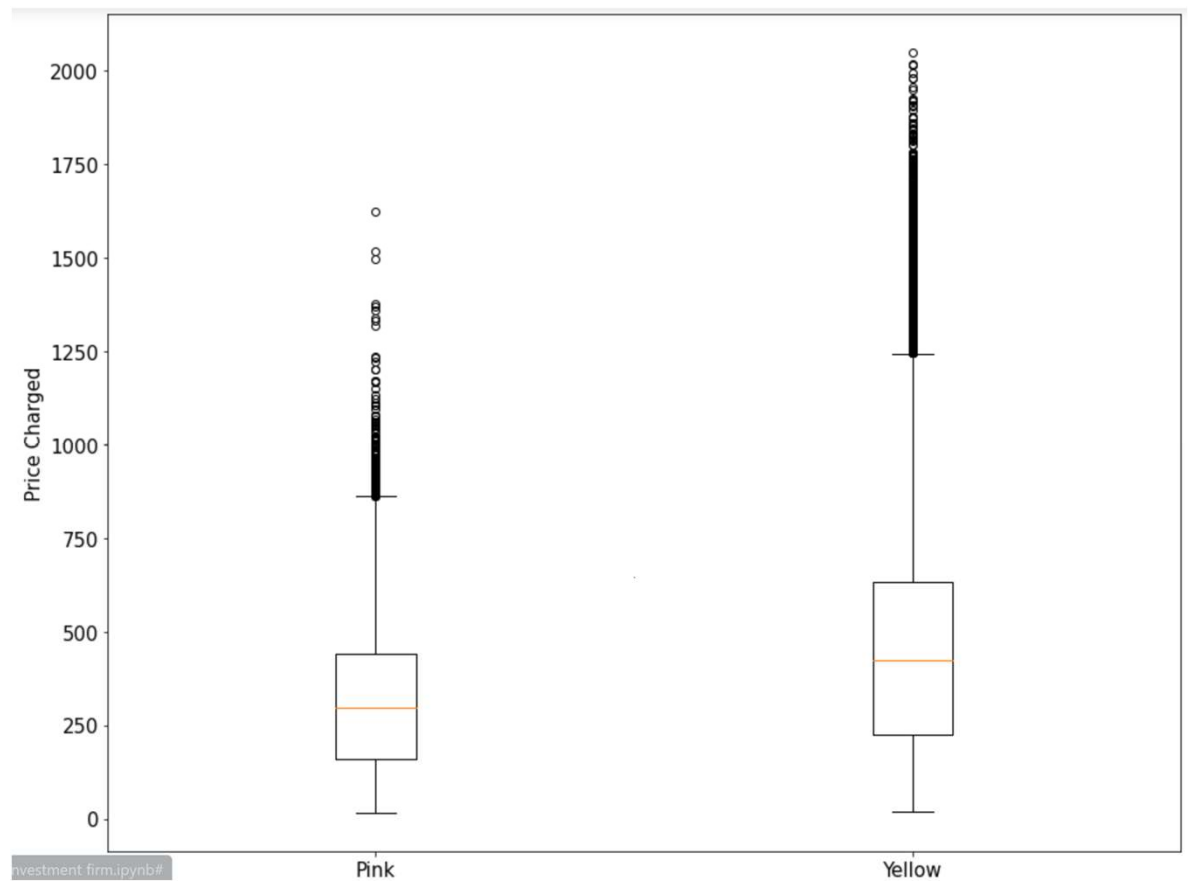
We can see that there is a larger number of Yellow cabs compared to Pink Cabs. This analysis may yield some interesting results in the customer frequency per state.



Yellow Cab	274681
Pink Cab	84711

## Company EDA 2

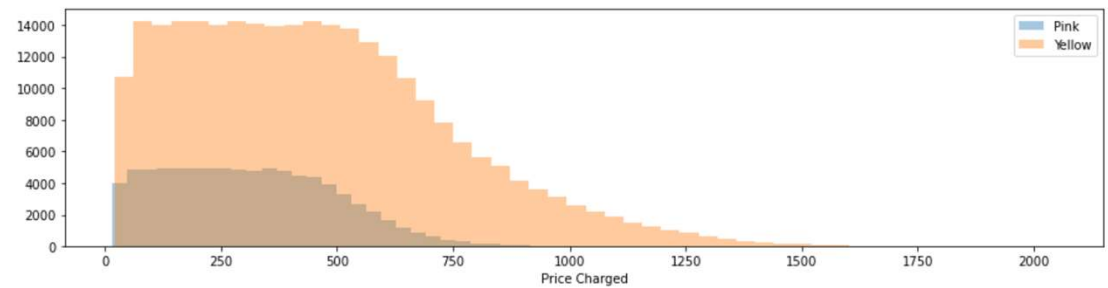
Looking more in-depth in the distribution of price charged, we can see that Yellow has a more skewed distribution than Pink. This means that overall, Yellow charges, on average, more than Pink.



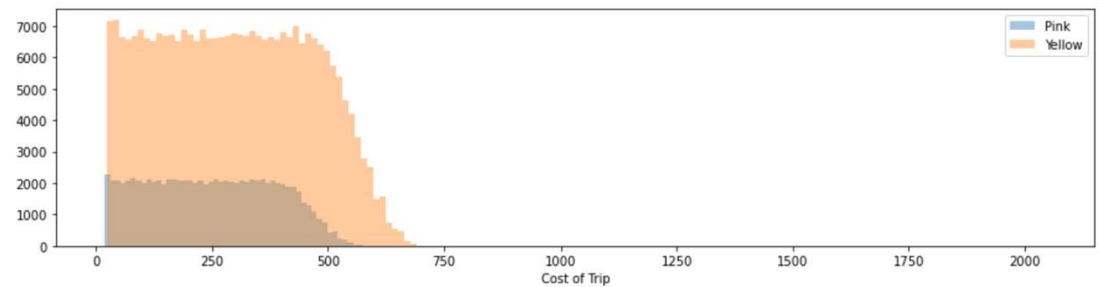


## Company EDA 3

Since profit will be an essential factor towards investment feasibility, looking at the distributions of cost of trip and price charged will be essential.



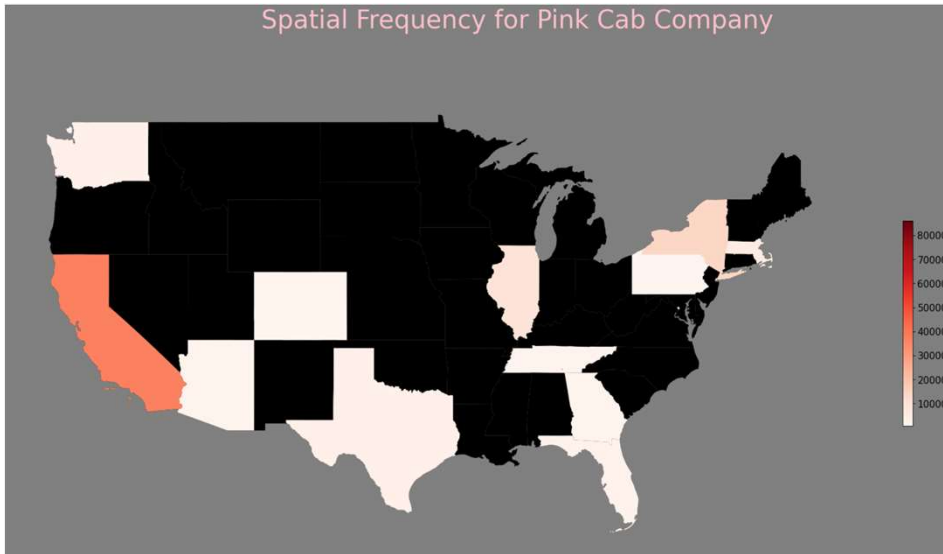
Yellow has a larger range than pink, meaning yellow has a more variable price charged.



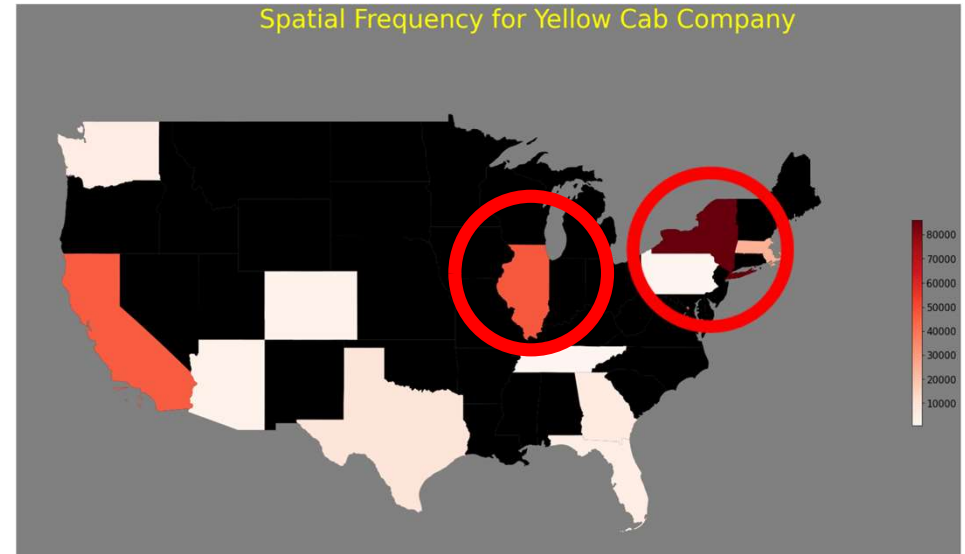
Cost of trip seems to be higher, but with less range. When noticing the difference between both, we can see that yellow will end up having more revenue within this study.

# Which company has more spatial control in the United States?

Spatial Frequency for Pink Cab Company



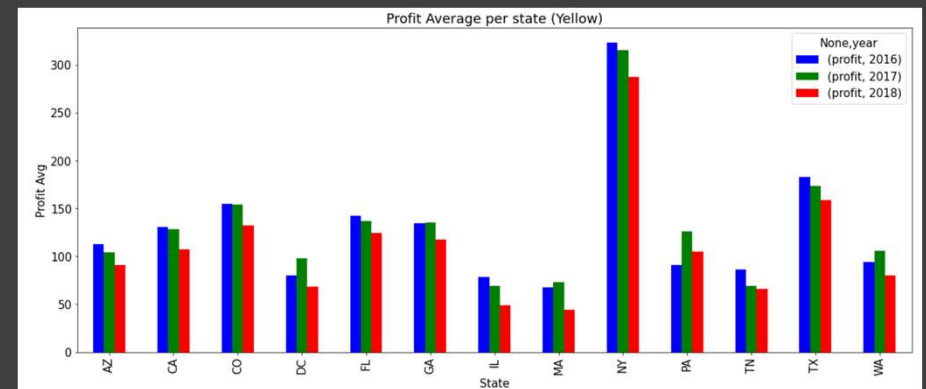
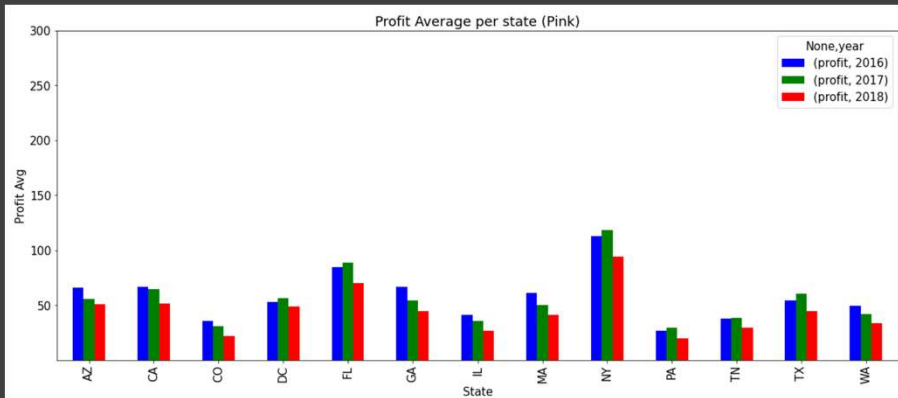
Spatial Frequency for Yellow Cab Company



Overall, Yellow is seen to have more customer control throughout all regions of the United States. The more prominent differences are seen in states such as California, Illinois and New York.

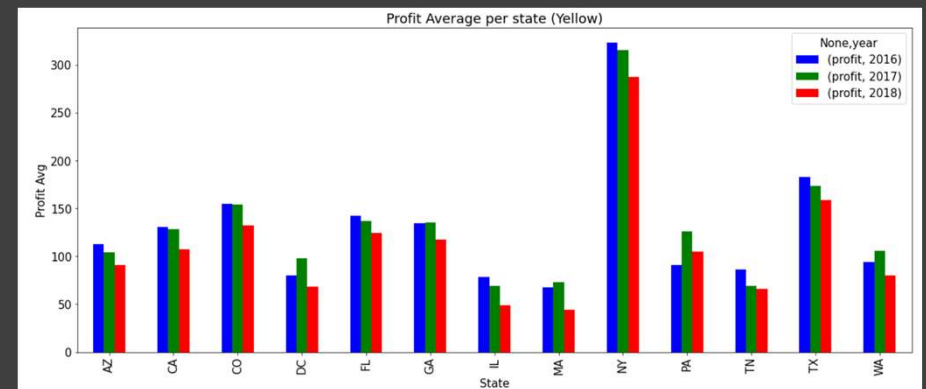
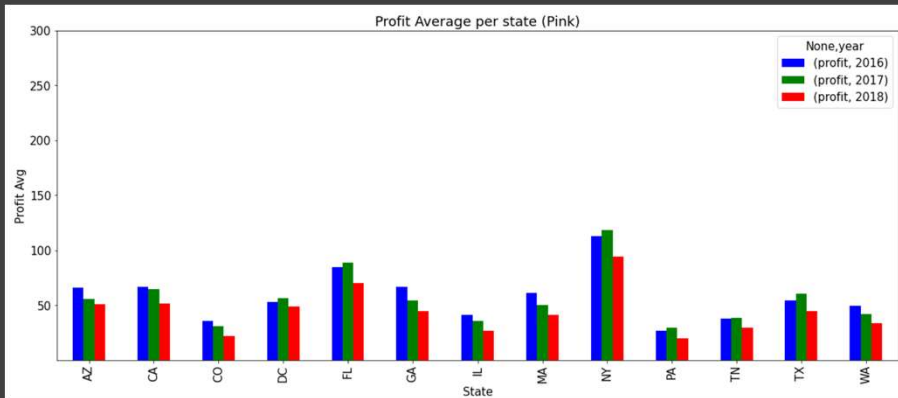


# Profit Based on State




Profit Average  
Based on State,  
Per Year

- Between both companies, there is a definite difference between the amount of average profit. New York is seen to be the most profitable, however, considering the holistic view of all states, Yellow has gained a more noticeable profit margin.

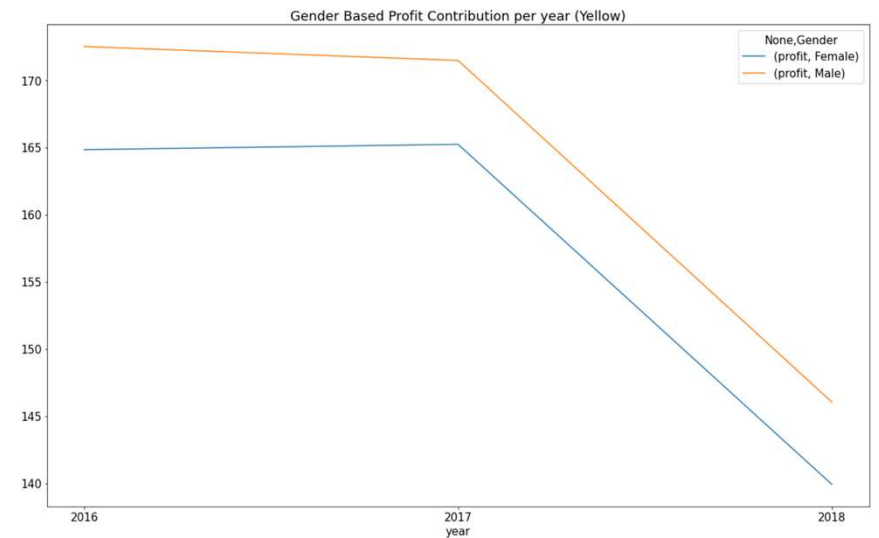
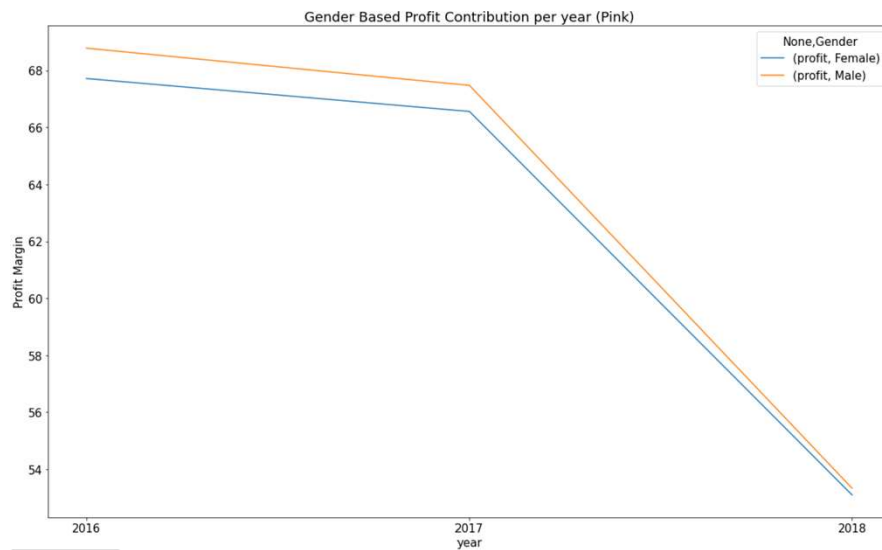


Profit Average  
based on State,  
per Year 2

- However, there is a noticeable downward trend for 2018's average profit compared to all of the other years. This may give a thorough prediction for future forecasting.



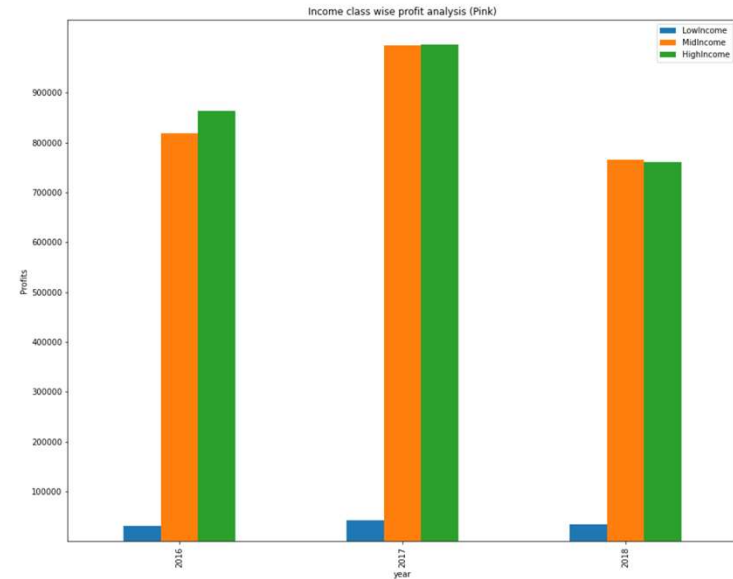
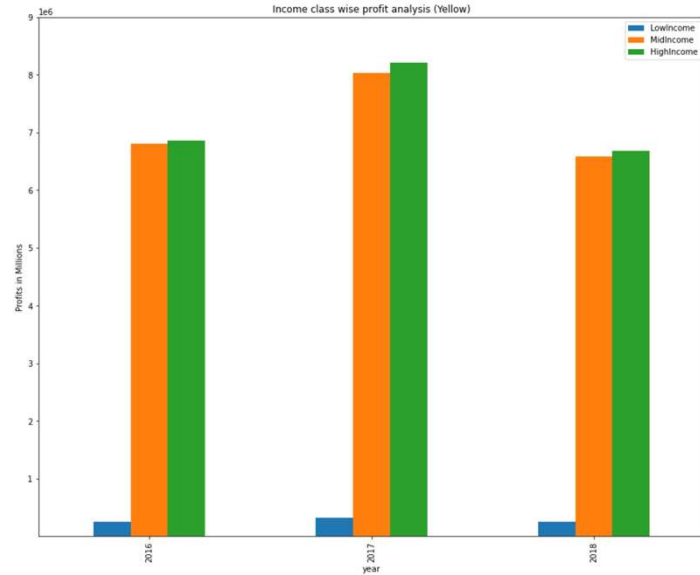
# Customer Characteristics Based on Profit



## Gender Based Profit Contribution Per Year

- Pink has a much closer distribution of genders, giving a similar profit margin per year compared to Yellow. However, both still have a higher male contribution, which may contribute to a higher population of males using these services.

Millions



## Income Class wise Profit Analysis

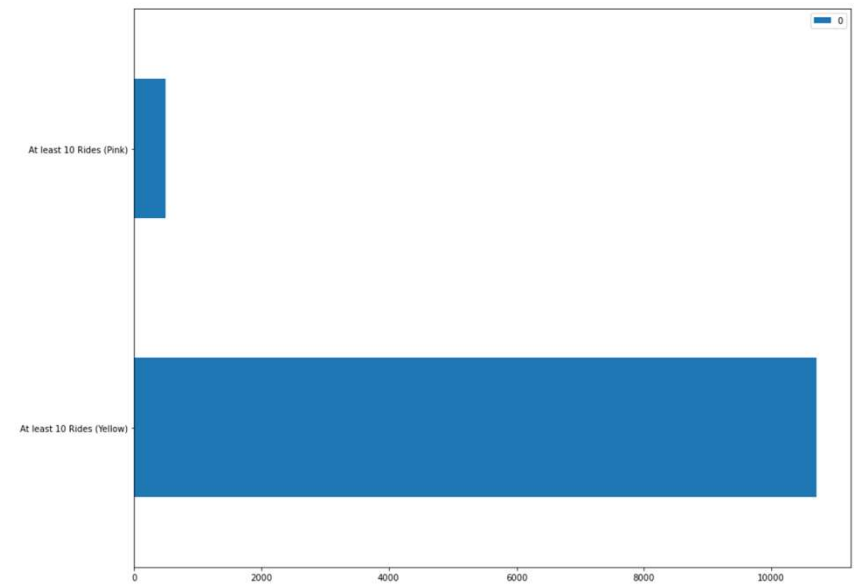
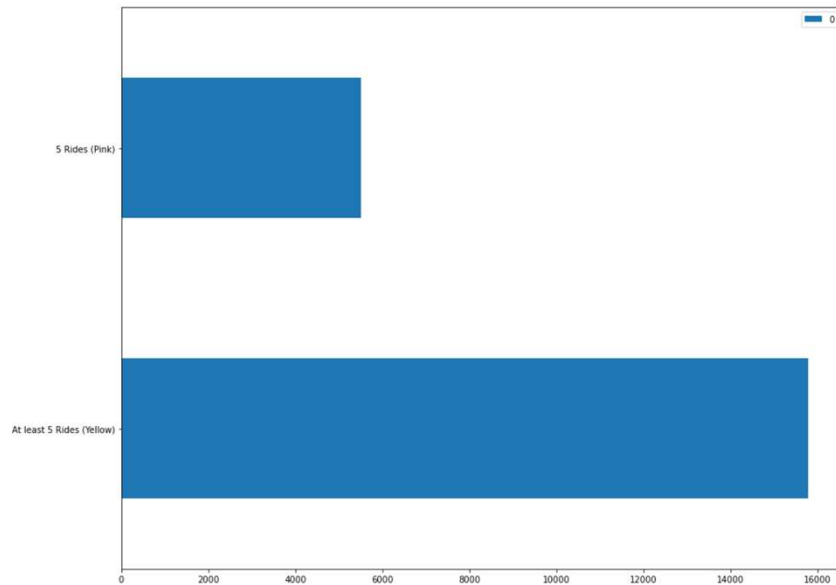
- Most of the middle and high-income classes generate the most revenue for both companies.
- The distribution per year is even across all income classes, however, Yellow has had the most profit overall.



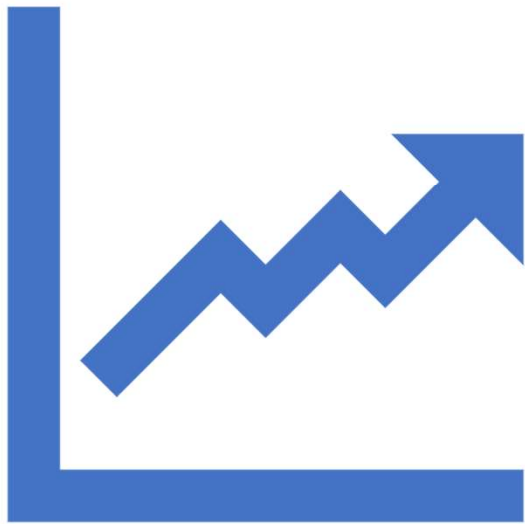


Customer Retention

## Customer Retention



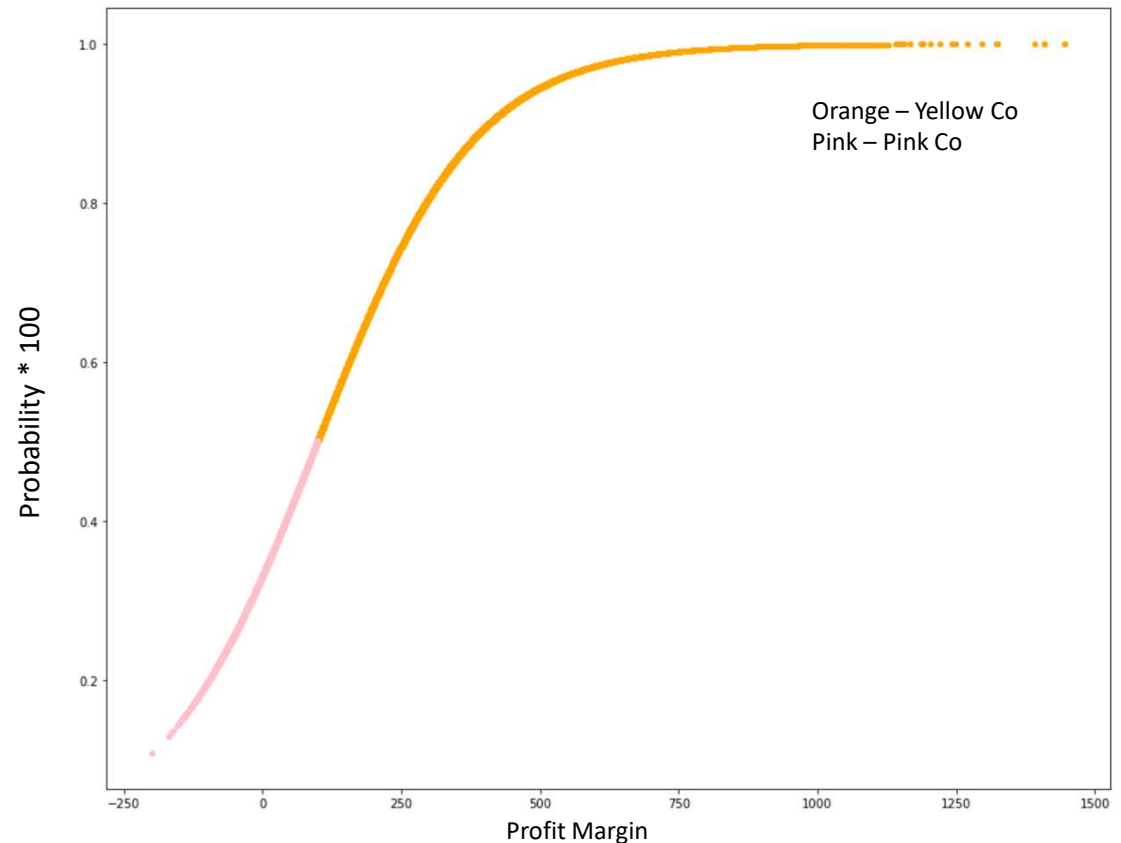
- Though the scaling is different, we are looking at the difference between people who have had at least 5 rides versus 10 rides.
- Customer loyalty has had more of an impact on Yellow's side compared to Pink.



Profit Analysis For  
Pink and Yellow

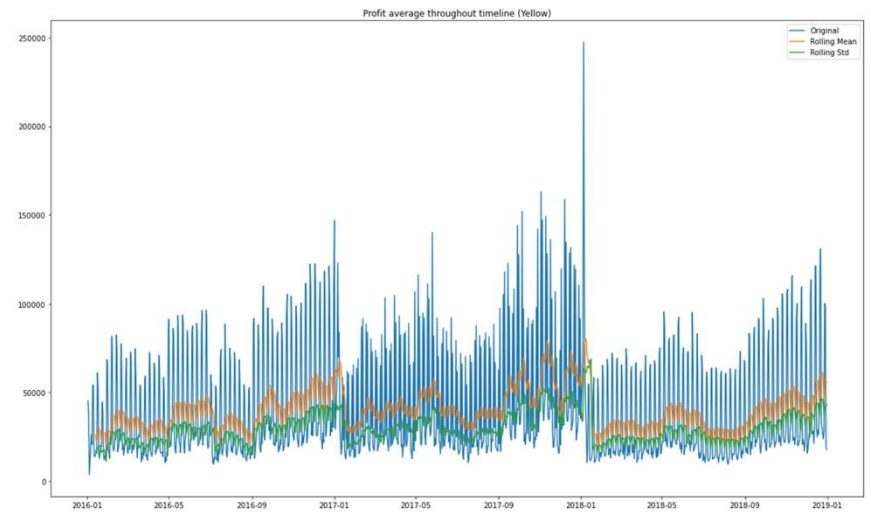
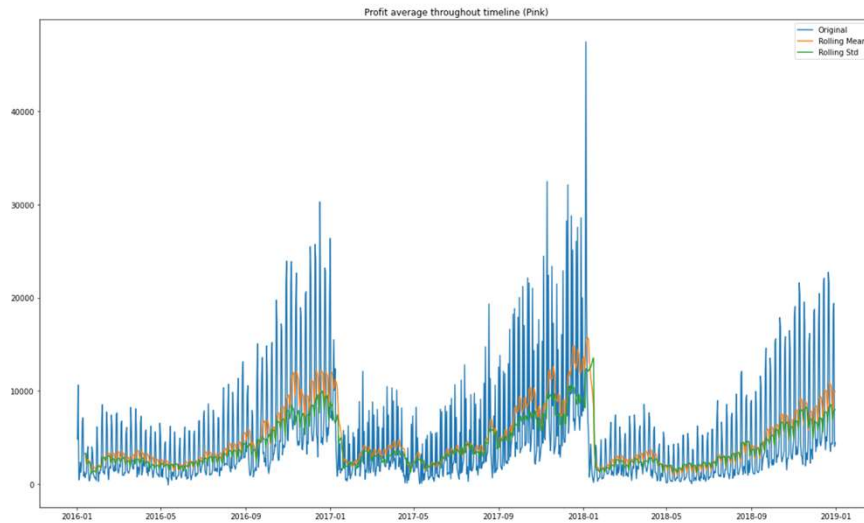
Probability  
Regression (Which  
company has the  
highest probability  
to gain the most  
profit)

Using a logistic regression, we can see that the Yellow company has a higher probability of obtaining a higher profit margin compared to the Pink Company.



Disclaimer: Negative numbers are included; however, this may be due to the loss of profit that Pink Company may have endured.

## Time Series Trend



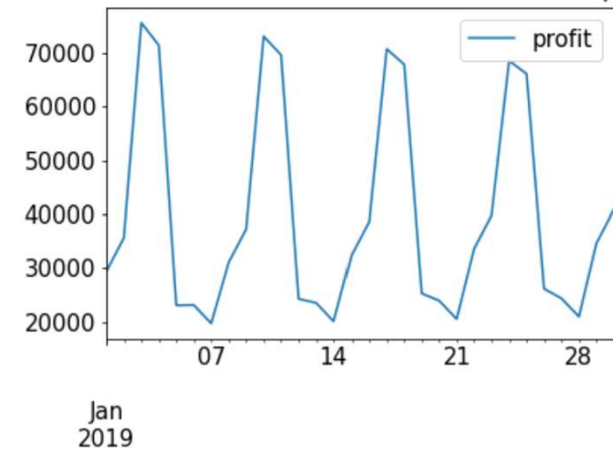
- Both taxi companies have seasons of increase towards the beginning of the year, however, we can see that Yellow has had more variations in the amount of profit that they obtained.



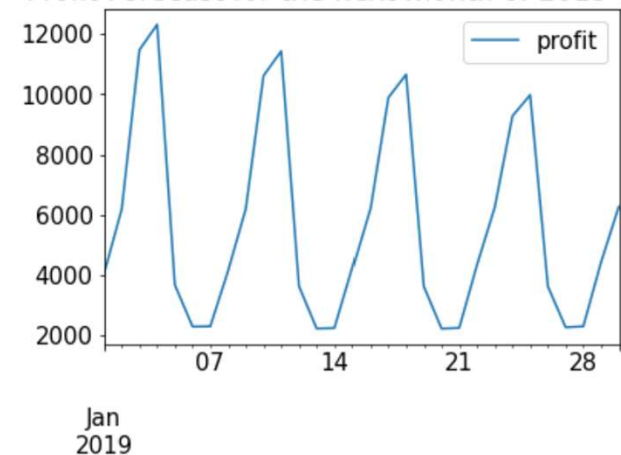
# Time Series Forecasting for the Next Month

- Using an autoregressive model, we can see that both companies seem to fluctuate in profits.
- This predicted profit is diminishing over time.

Profit Forecast for the next month of 2019 (Yellow)



Profit Forecast for the next month of 2019 (Pink)



# Spatial Factor Recommendations

---

## Spatial Control

- Having a holistic focus throughout the United States can be crucial for revenue.
- It is safe to say that the Yellow company has more of an outreach compared to Pink, by a large margin.

## Yearly State Profit Control

- Per year, Yellow makes a higher profit margin for each state compared to Pink.
  - Though there is a large cluster in New York, all the other states exhibit much larger proportions than Pink.
-

# Customer Characteristic Recommendations

---

## Gender Based on Profit

- There seems to be a distinctive gap between genders for Yellow compared to Pink.
- This may be a cause for concern, as Pink has had all genders contribute the same amount per year compared to Yellow.
- When it comes to scaling, Yellow still manages to bring more profit from both genders.

## Income Based on Profit

- Both companies have a majority of the Middle and High classes.
  - The distribution is almost indistinguishable, however, Yellow still manages to create an impressive profit from all social classes.
-



# Customer Retention Recommendation

---

- Yellow has had a larger frequency of more loyal customers compared to Pink, which would mean that the majority of customers will most likely stick with Yellow compared to Pink.

# Profit Recommendations

---

- Forecasting
  - Based on the given graphs, both companies seem to diminish over time.
  - This could be regarding other competitors coming into the market. (Uber)
  - However, since Yellow has larger scaling than Pink, they are still projected to lose, but not as much as Pink.

---

# Conclusion

---

- Based on the compelling insights, our conclusion for which company XYZ should invest in would be **Yellow**.

# Thank You