



**Data Glacier**

Your Deep Learning Partner

# **G2M Insight For Cab Investment Firm Exploratory Data Analysis**

**LISUM02  
06<sup>th</sup> August 2021**

**By,  
Joseph Antony**

# Agenda

- Executive Summary**
- Problem Statement**
- Methodology**
- Tools Used**
- Data Preparation**
- EDA**
- Hypothesis Results**
- Forecast**
- Conclusion**

# Executive Summary

- This report was commissioned to summarize my findings to the client, **XYZ company**, for the best Cab company to invest in.
- This report focuses on two Cab companies - **Yellow Cab** and **Pink Cab**.
- An extensive Exploratory Data Analysis supported with Hypothesis Testing and Forecasting methodologies were utilized for this study.

## Recommendation:

According to the data and two-year forecasts, **Yellow Cab** has been identified to perform better in terms of:

- ✓  High daily number of trips on any given day
- ✓  Higher Market share on most Cities
- ✓  Better Customer Retention
- ✓  Higher Profitability

# Problem Statement

## The Client

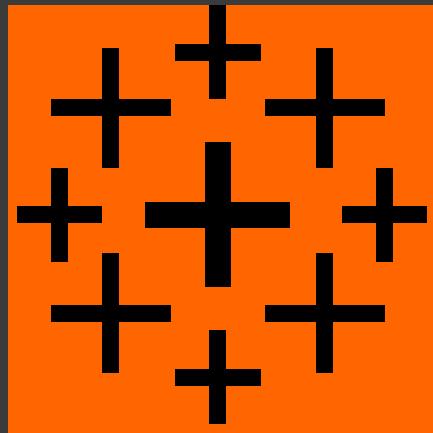
**XYZ** is a private firm in US. Due to remarkable growth in the Cab Industry in last few years and multiple key players in the market, **XYZ** is planning to invest in Cab industry and as per their Go-to-Market (G2M) strategy, they want to understand the market before taking final decision.

**XYZ** has provided multiple data sets that contains information on two cab companies. Each data set provided represents different aspects of the customer profile. **XYZ** is interested in using the actionable insights obtained to help them identify the right company to make their investment.

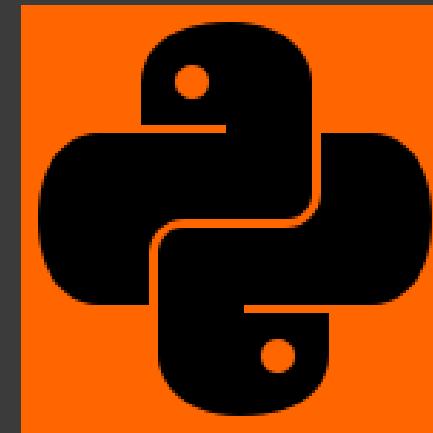
# Methodology

-  **Exploratory Data Analysis (EDA)**
-  **Hypothesis Testing**
-  **Forecasting**

# Tools Used



Tableau



Python

# Data Preparation

- Four data sets that contains information on Two cab companies had been provided.
- Each data set represents different aspects of the customer profile.
- Time period of data is from **02<sup>nd</sup> January 2016 until 31<sup>st</sup> December 2018**.
- Master Dataset was created by inner joining:
  - **Cab\_Data.csv**: Includes details of transaction for both cab companies.
  - **Customer\_ID.csv**: Mapping table that contains a unique identifier which links the customer's demographic details.
  - **Transaction\_ID.csv**: Mapping table that contains transaction to customer mapping and payment mode.

The fourth Dataset, **City.csv** has been left out as it does not provide information about daily trips for both Cab companies.

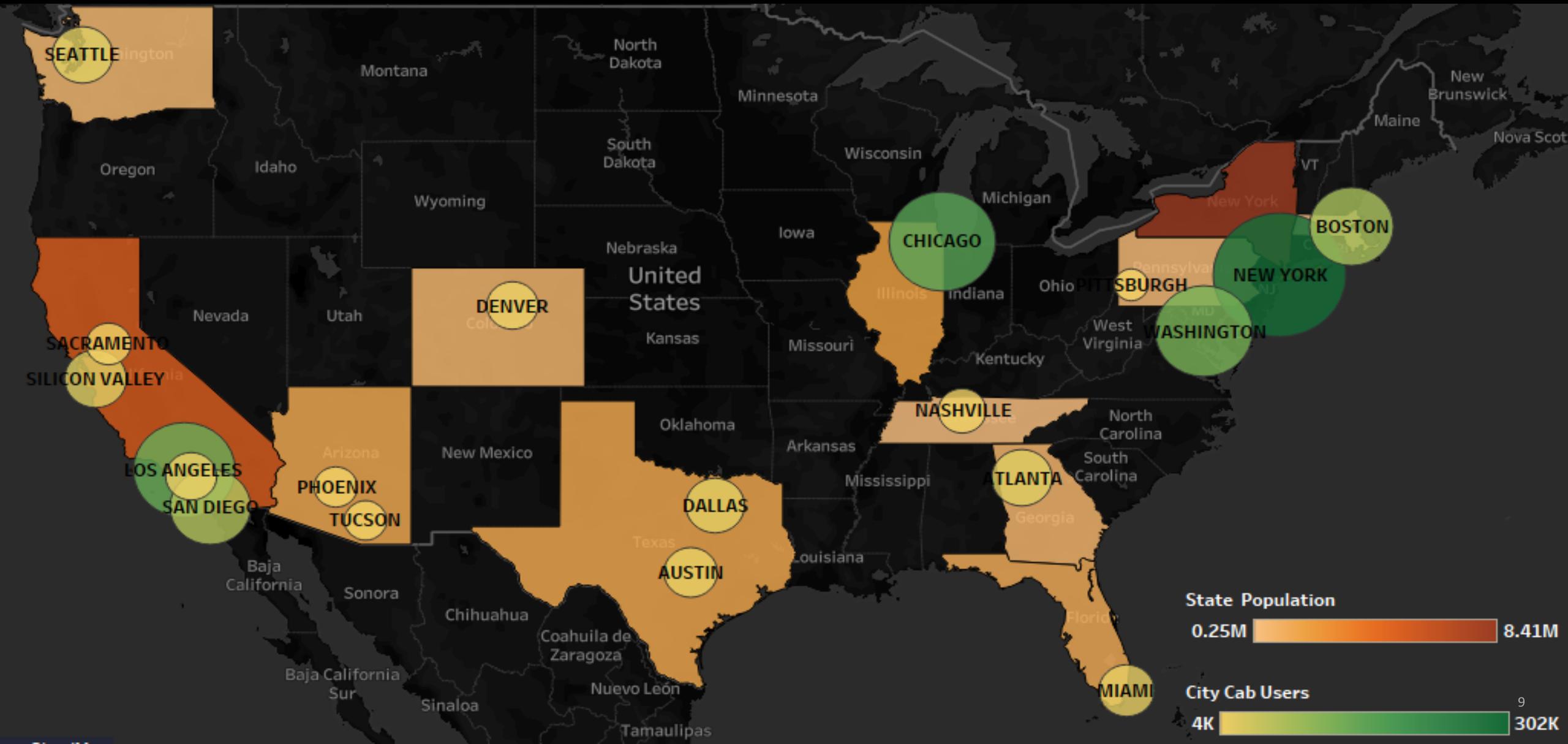


Schema of Master Dataset

# EDA Key Insights

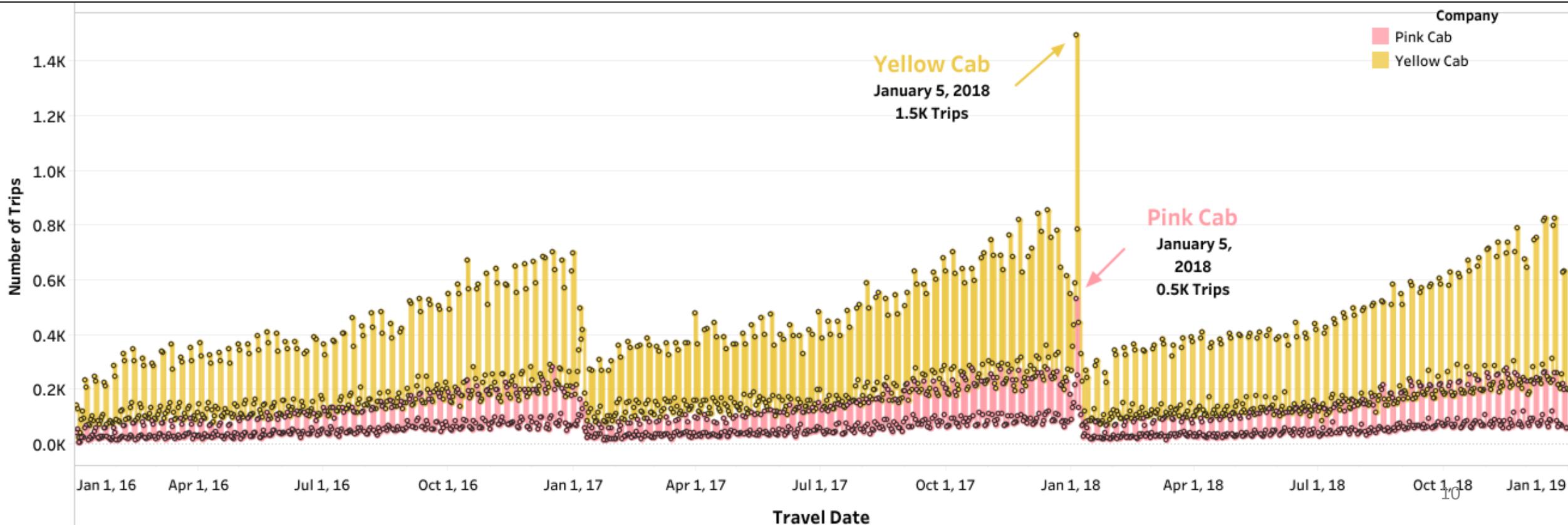


# Population and Cab Users



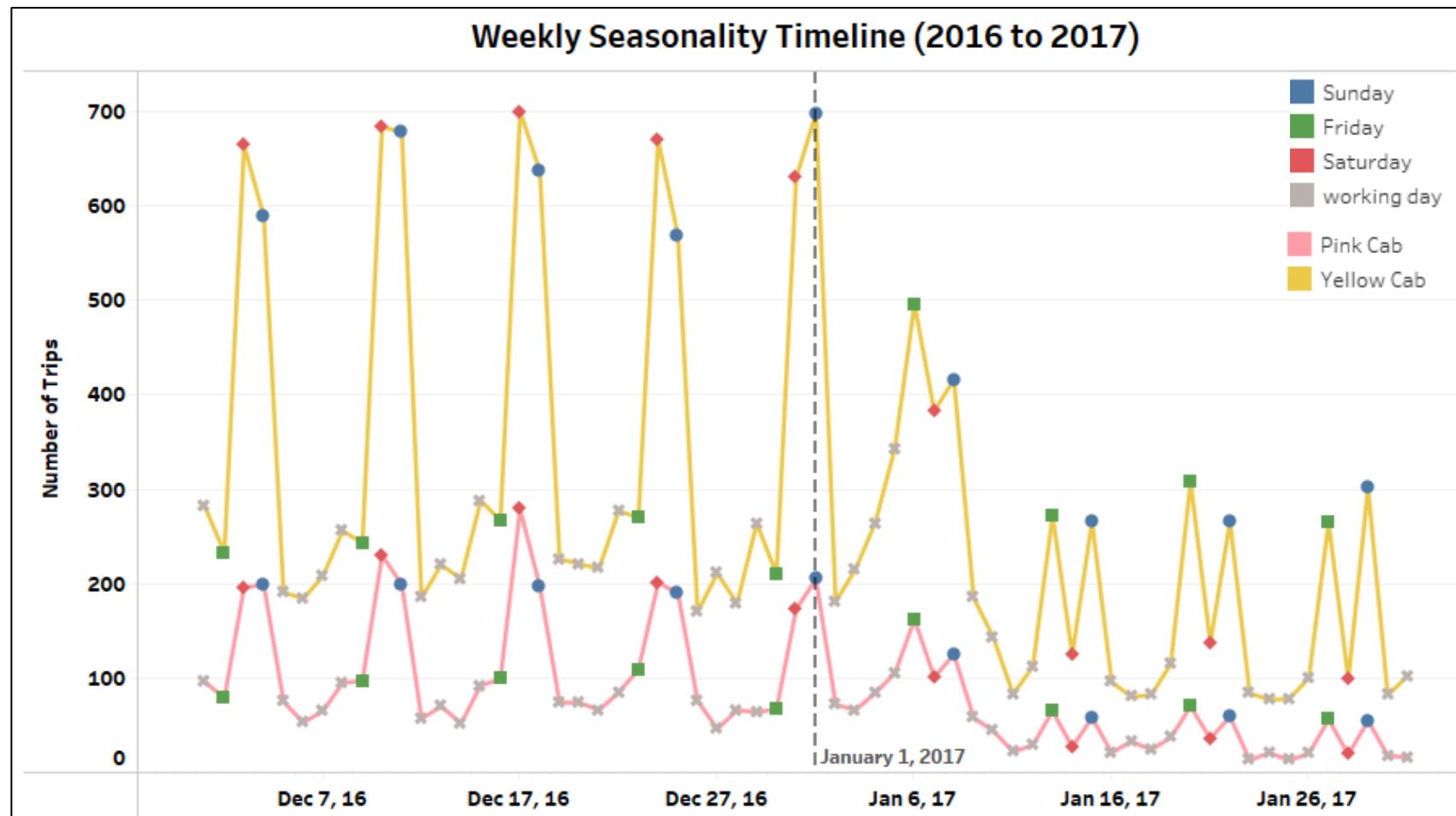
# Daily Trips (2016 - 2018)

- Same trip pattern for both companies.
- Yellow Cab has highest trips every single day.
- Jan 5<sup>th</sup>, 2018, saw the highest number of trips for both cab companies.
- Beginning of every year, the trips are very low.
- Conversely, at the end of the year, trips are very high.



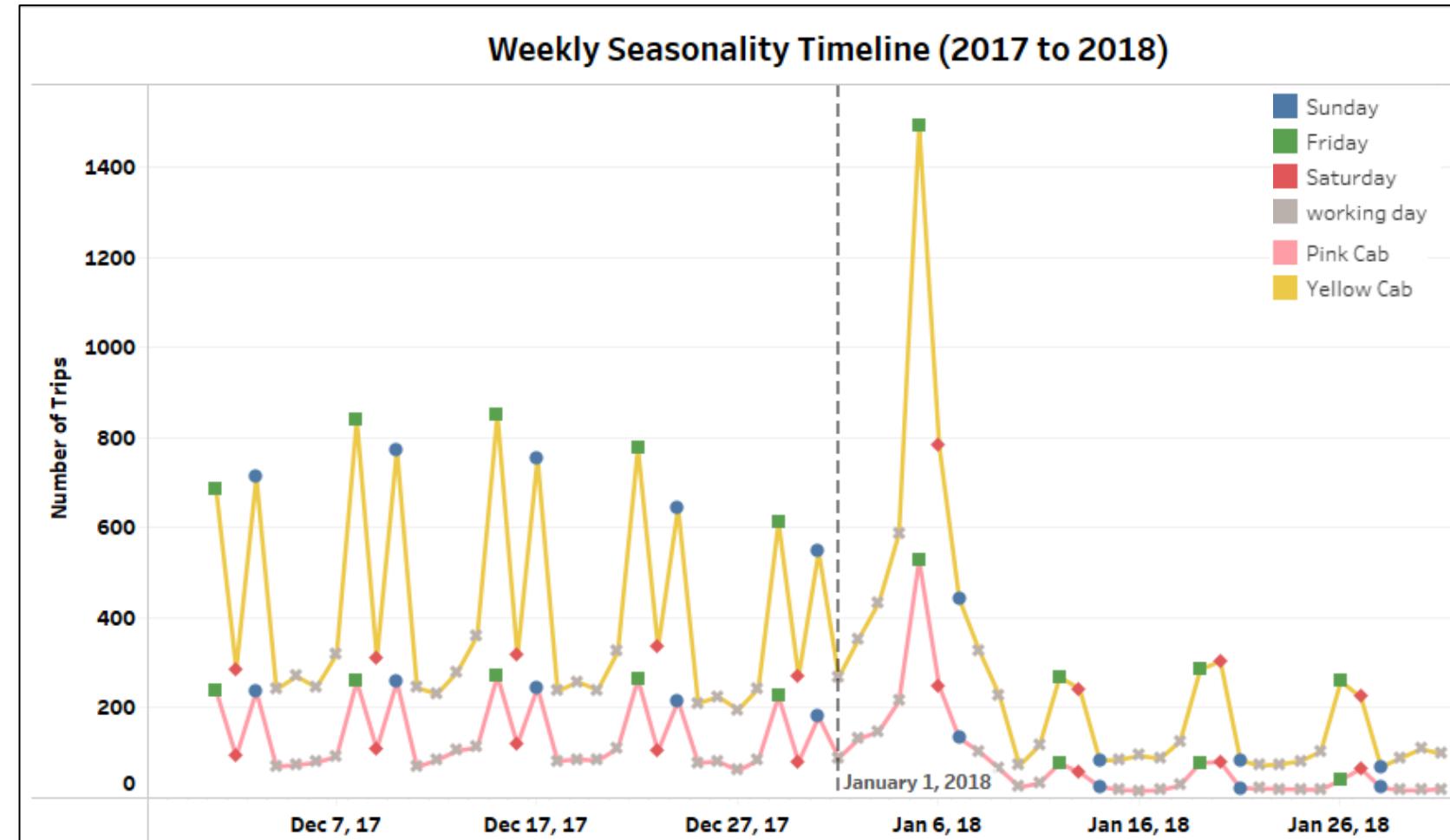
# Weekly Seasonality (2016 - 2017)

- Low trips on Friday, and higher trips on Saturday and Sunday in 2016.
- Pattern changes in 2017, where higher trips on Friday, then lowers on Saturday and high again on Sunday.



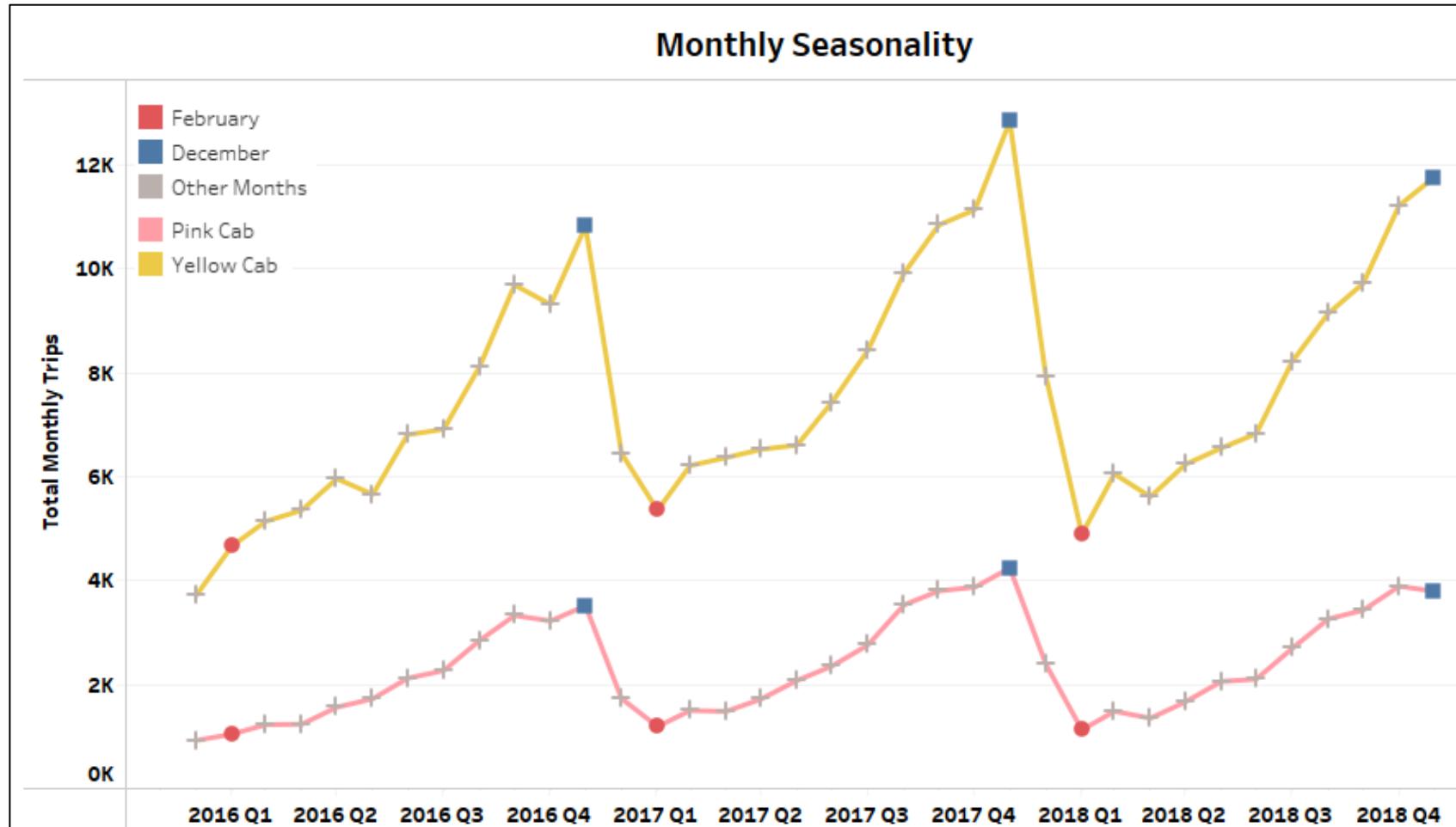
# Weekly Seasonality (2017 - 2018)

- In 2017, weekly seasonality changes again.
- High trips on Friday, then lower trips on Saturday, followed by even lower trips on Sunday



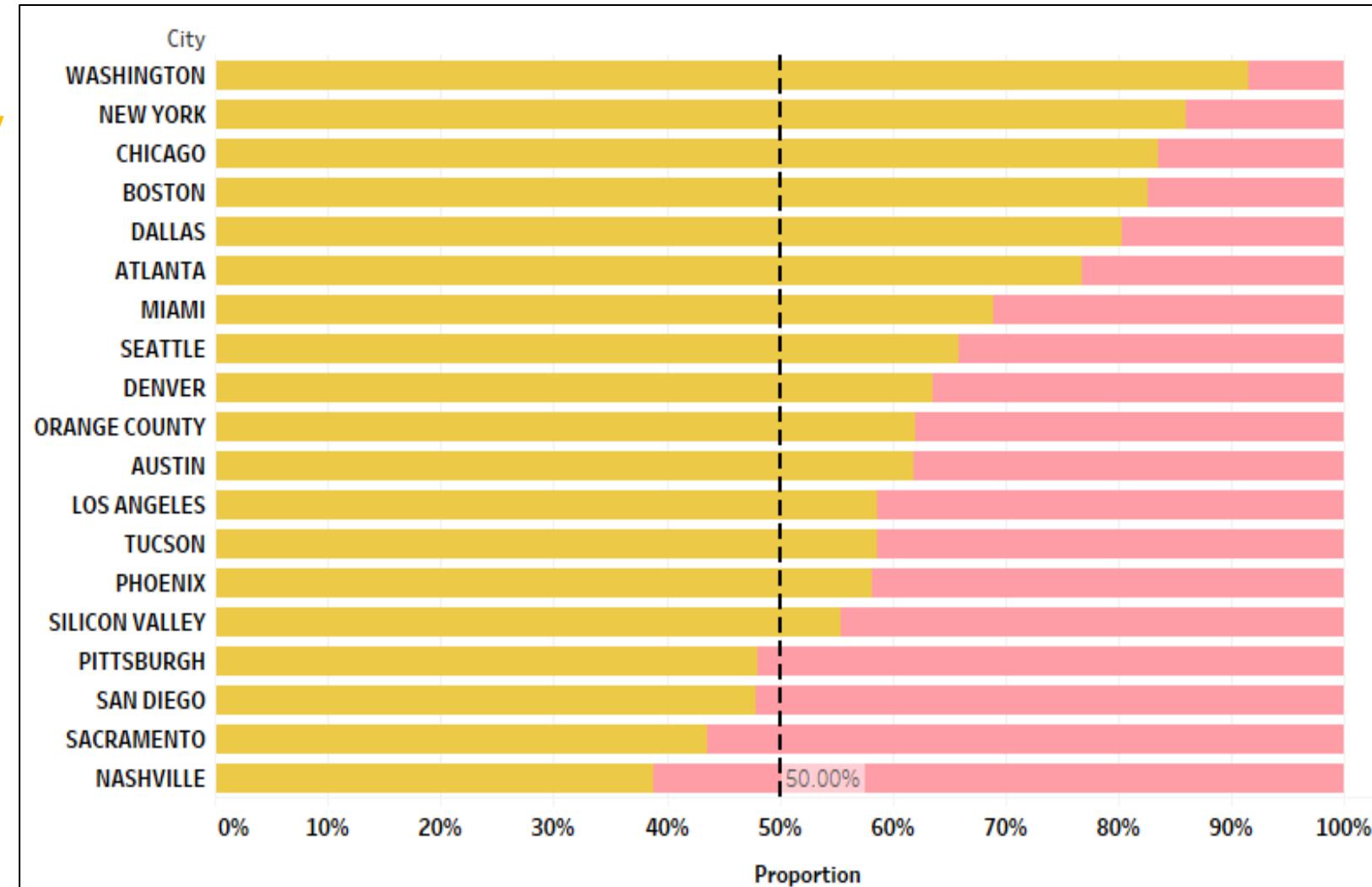
# Monthly Seasonality

- Highest trips in December
- Lowest Trips in February.



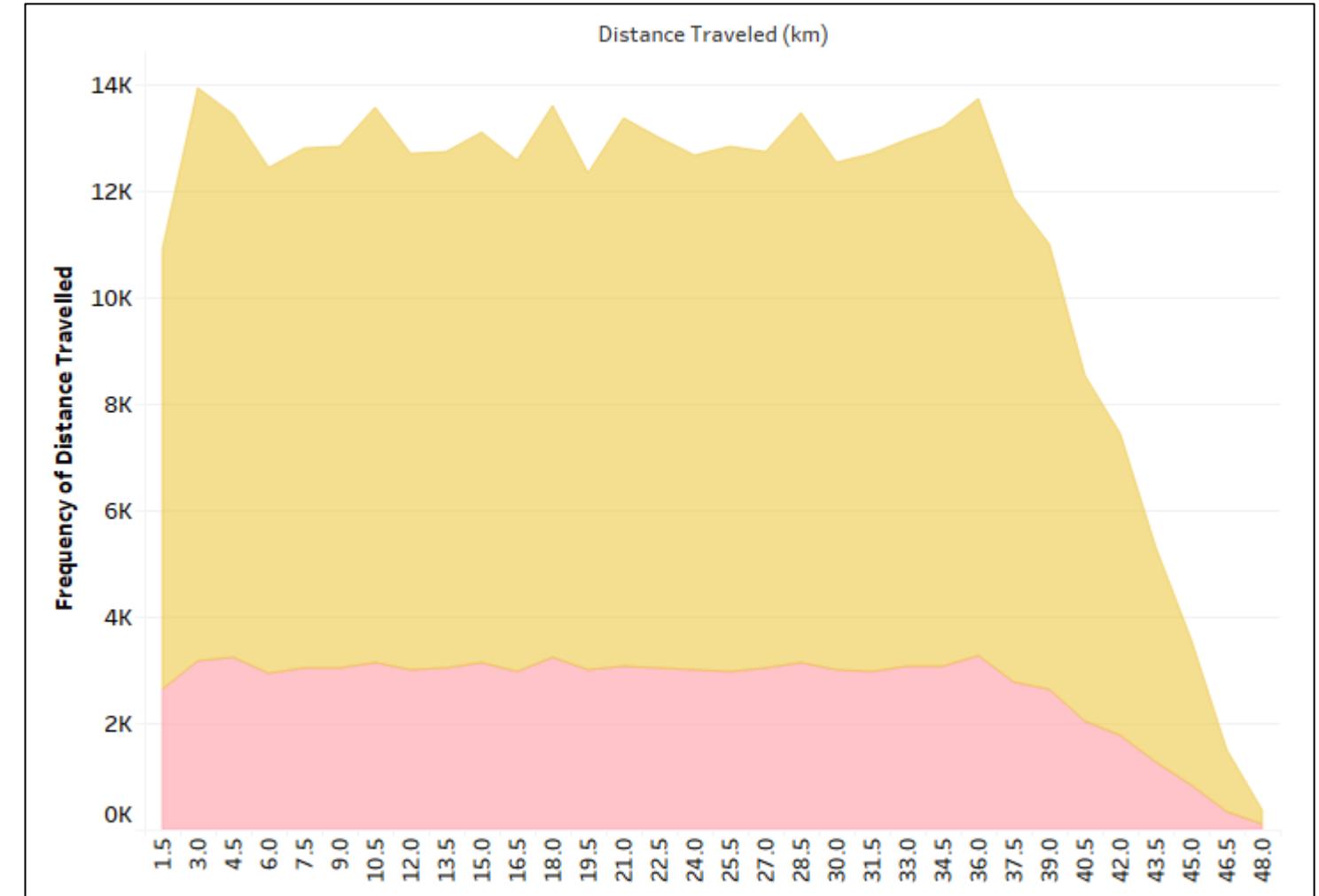
# Market Share by City

- Market share of Yellow Cab higher across majority of the Cities.
- Pink Cab has higher market share on few cities - Nashville, Sacramento, San Diego, Pittsburgh.



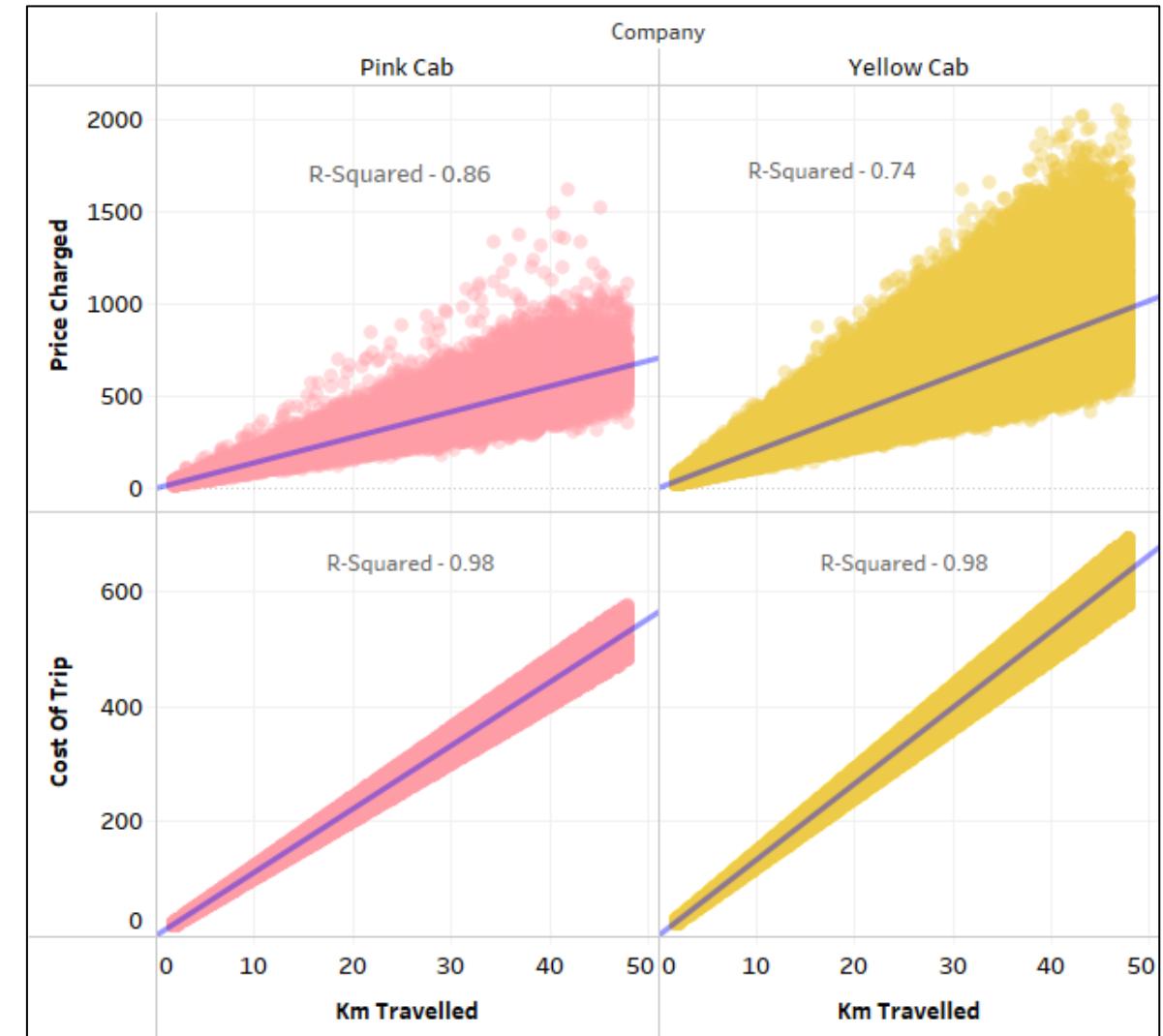
# Frequency of Distance Traveled

- **Distance travelled remain uniform for both Cab companies.**
- **Trips that involves travelling more than 36 km are rarer.**



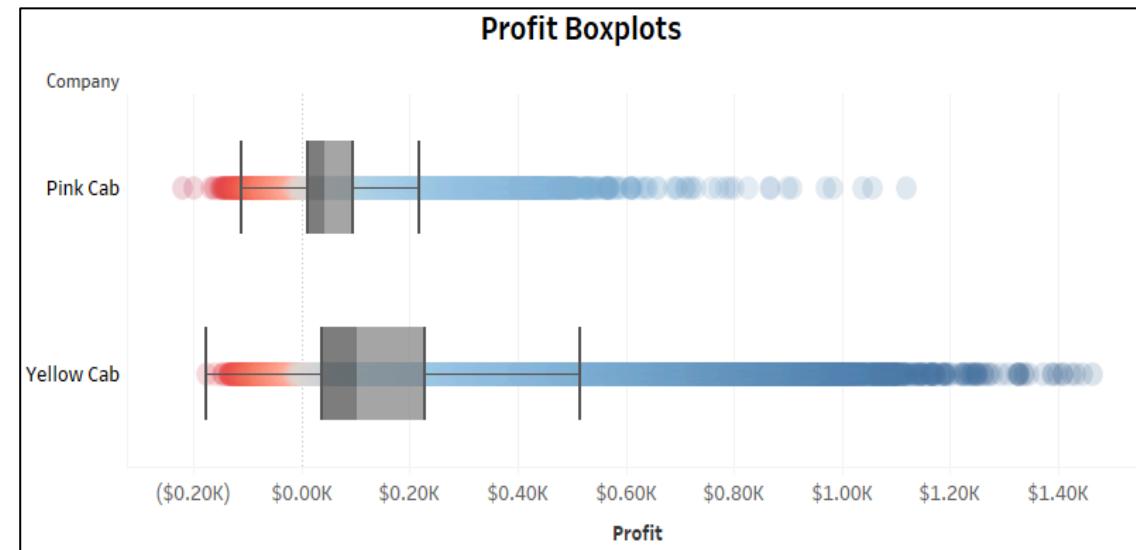
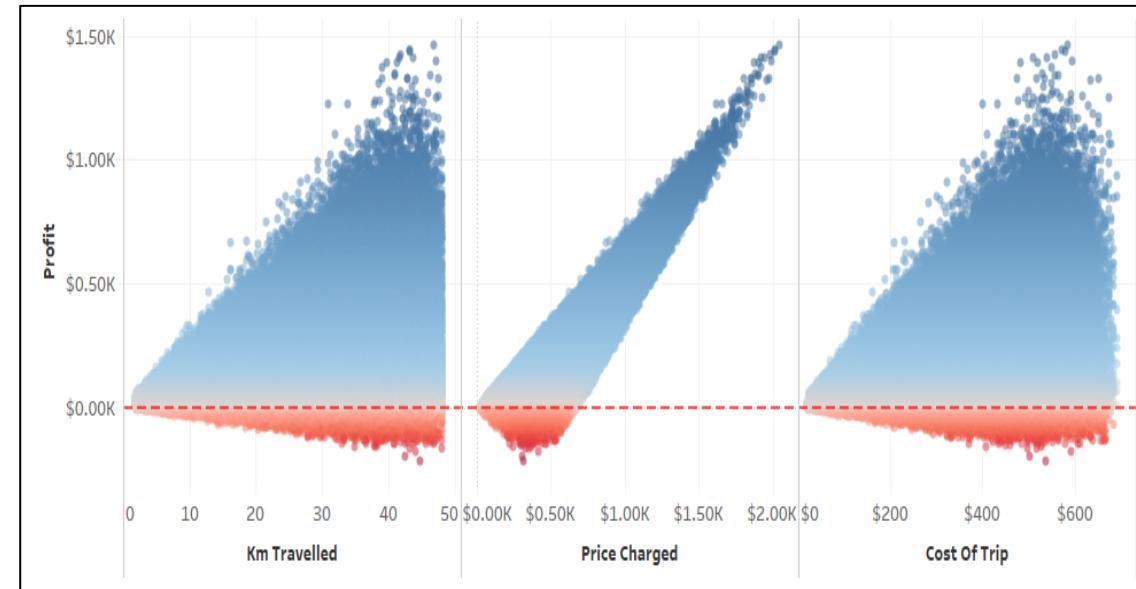
# Distance vs Revenue & Expense

- **High Correlation between Distance traveled and Cab Expenses.**
- **Correlation between Distance and Cab Revenue less compared to Cost of Trip.**
- **Signifies other factors involved in determining prices charged to customers apart from distance traveled.**



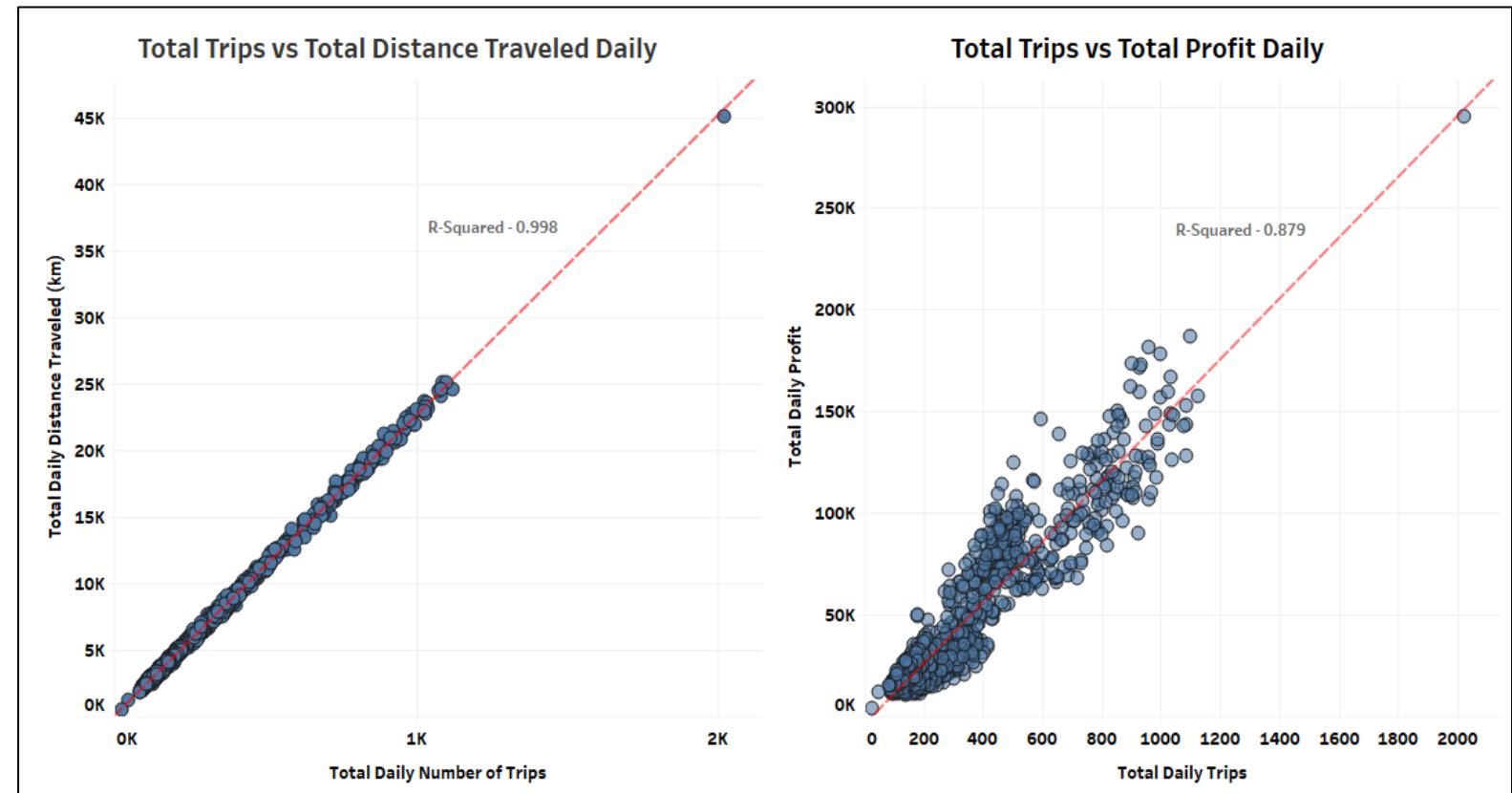
# Profit

- Profit = Price Charged – Cost of Trip.
- Profit more correlated with Price Charged.
- Profit varies at low Price charges but varies less as Price Charges increases.
- Heteroscedasticity increases with increase in distance traveled. Same pattern for Cost of Trip.
- Significant trips did not end in Profit.
- Median Profit for Yellow Cab higher than Pink Cab.

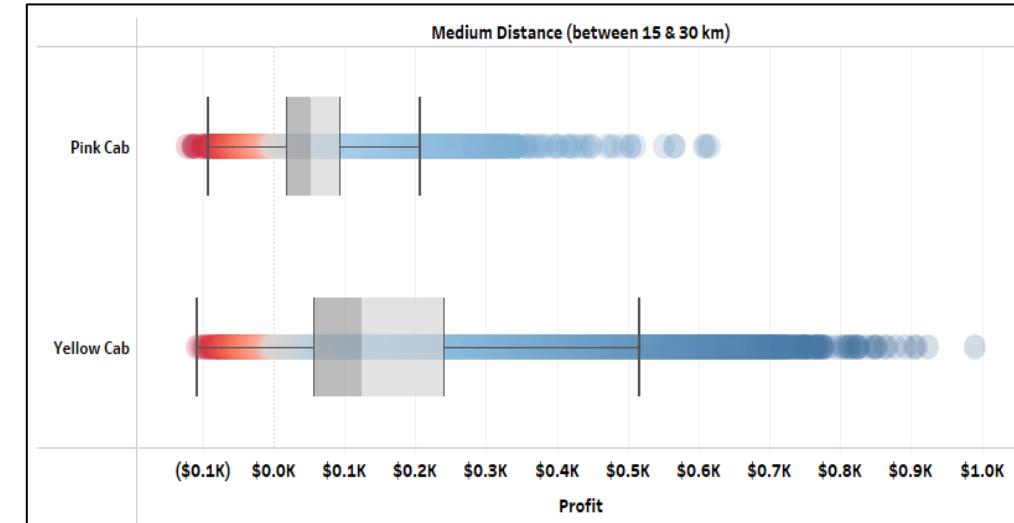
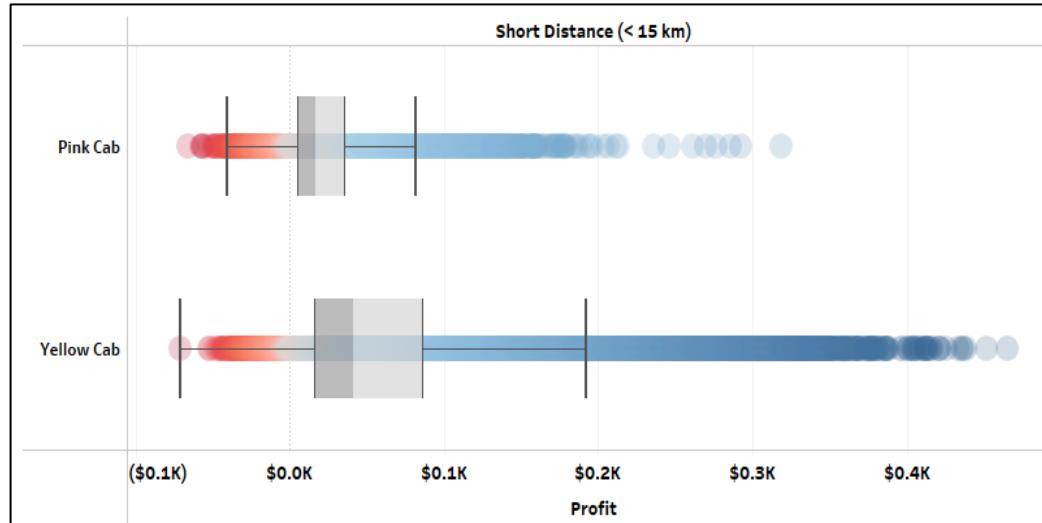


# Key Performance Indicator (KPI)

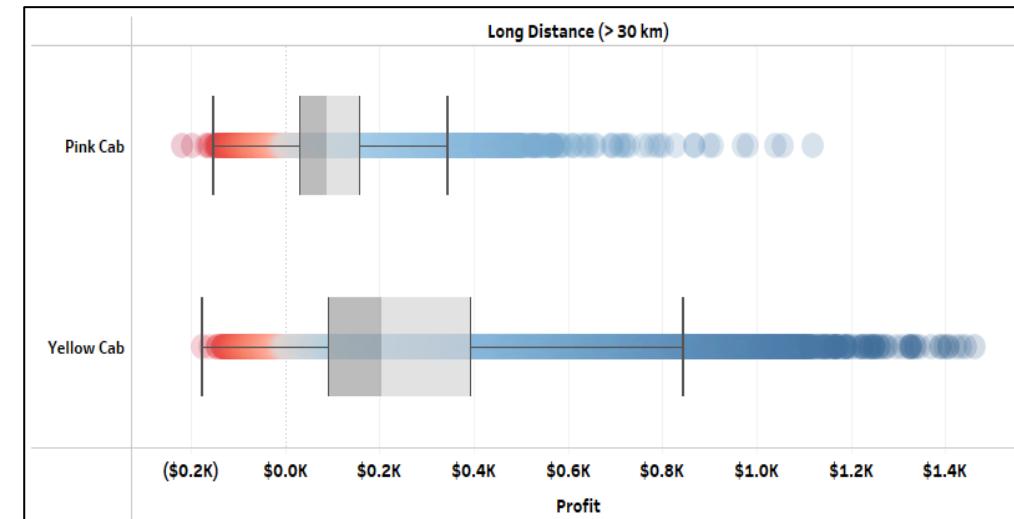
- Previous slides show the relations between Profit and Cab's Fares and Expenses.
- Both Cab fares and Expenses depend on distance traveled.
- There is a **perfect correlation between daily total trips and Daily total Distance traveled**.
- Moreover, **daily total trips also have a strong correlation with total daily profit**.
- Therefore, the **number of trips a cab company travels** in a day can determine the **company's financial performance**.



# Profit by Trip Type

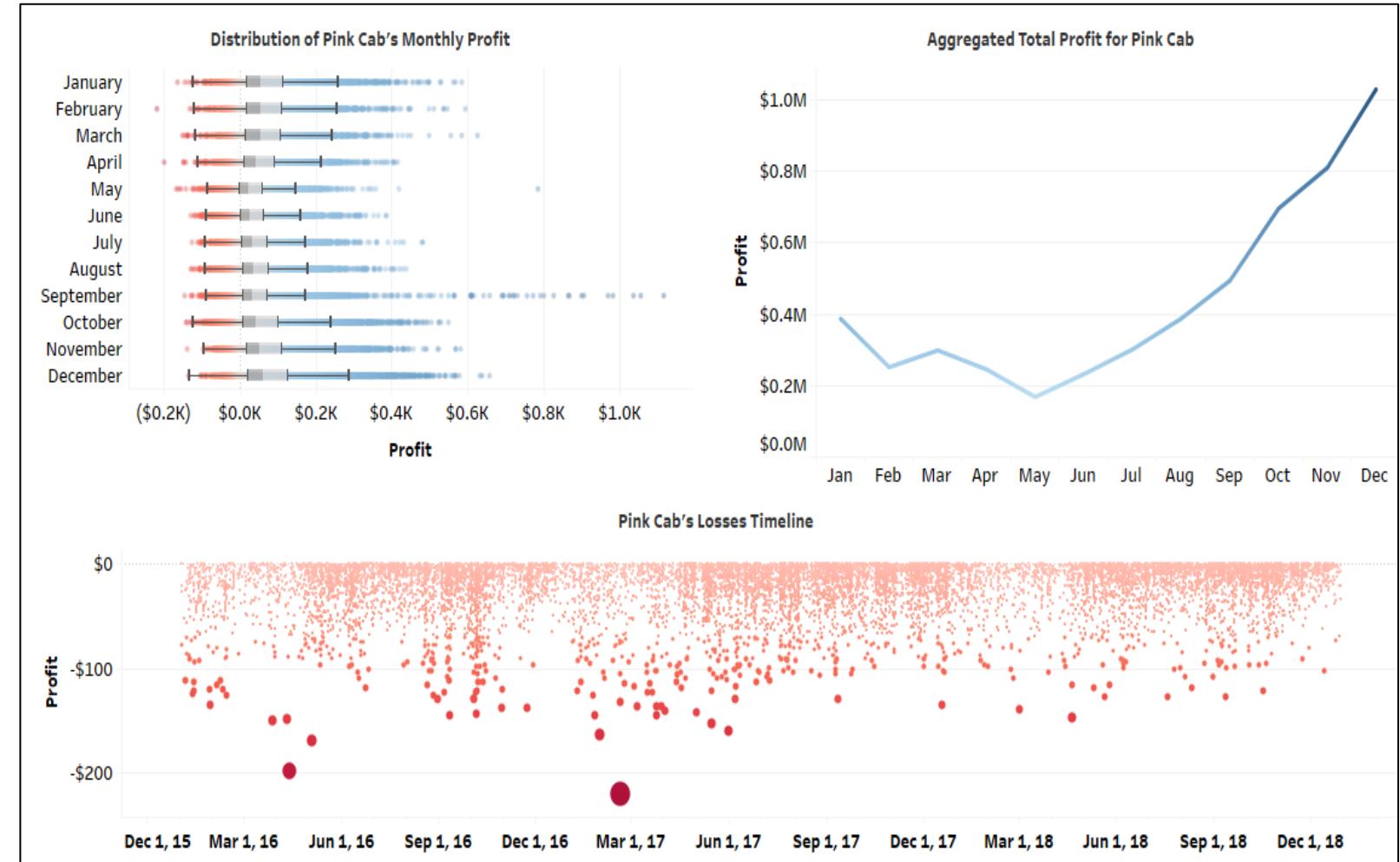


- Trips has been split into three categories based on distance.
- Across all kinds of trips, **Yellow Cab** makes higher median Profit compared to **Pink Cab**.



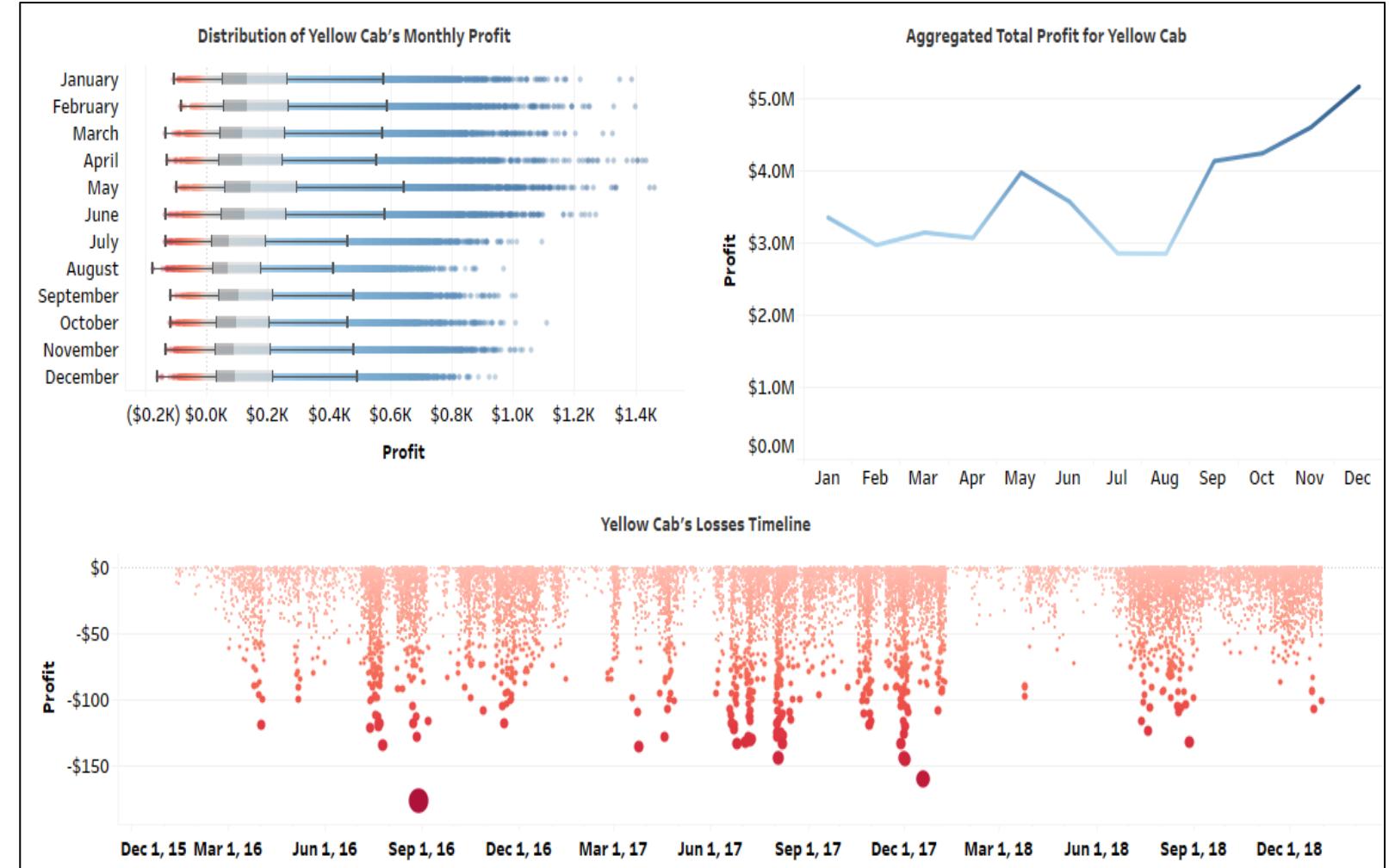
# Pink Cab – Profit Analysis

- Month likely associated with Profit.
- Median Profit lowest in May for **Pink Cab** and highest during December.
- Higher Profit outliers for month of September.
- Total Profit highest in December (2016 to 2018).
- Trip that ended in highest loss of \$220 made in February 2017, followed by \$199 in April 2016.



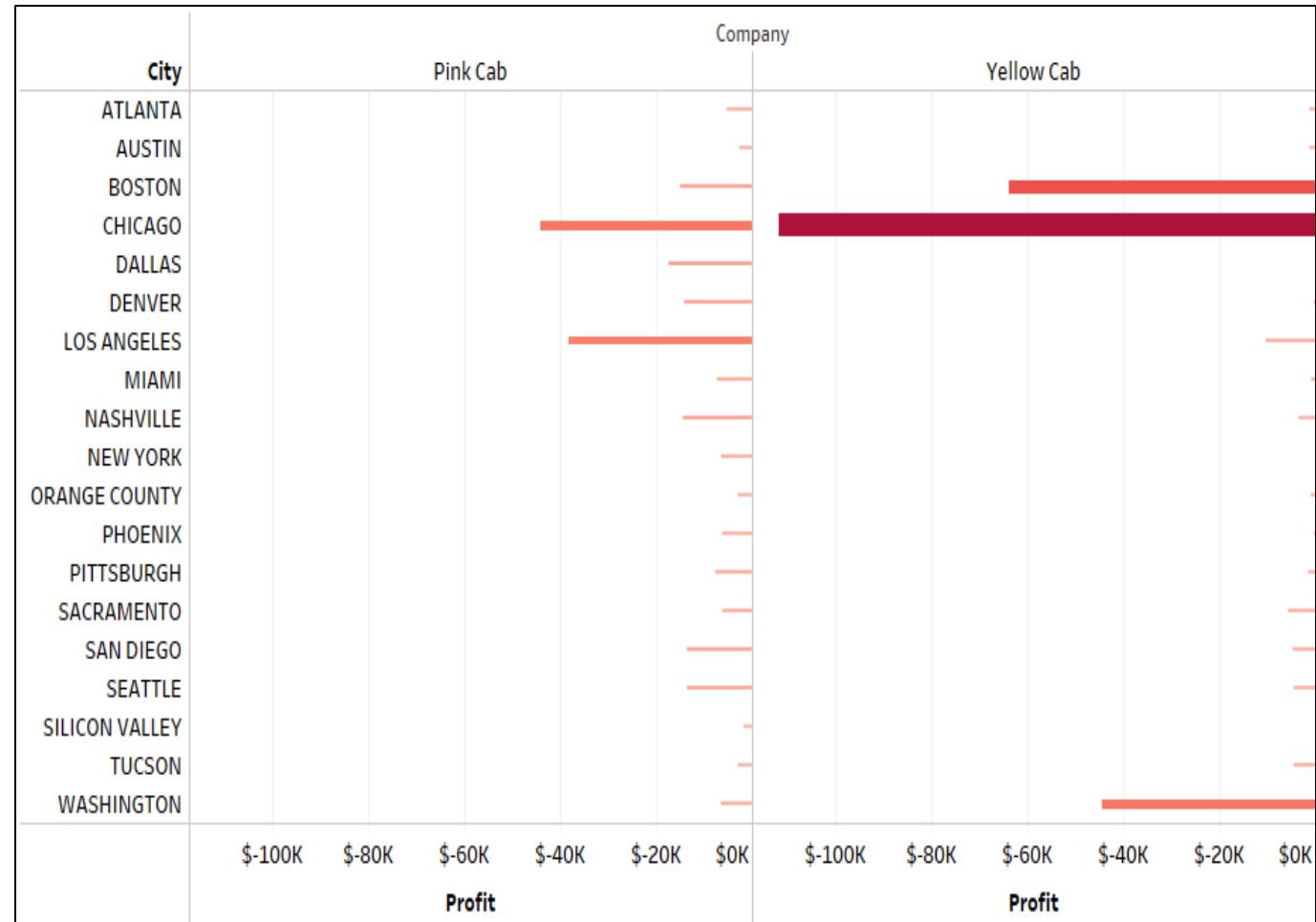
# Yellow Cab – Profit Analysis

- Yellow Cab profit significantly higher across all months compared to its rival.
- Median Profit lowest during July and August.
- Highest Total Profit in December (2016 to 2018),
- Trip that ended in high loss of \$177 made in August 2016, followed by \$166 on December 2017.



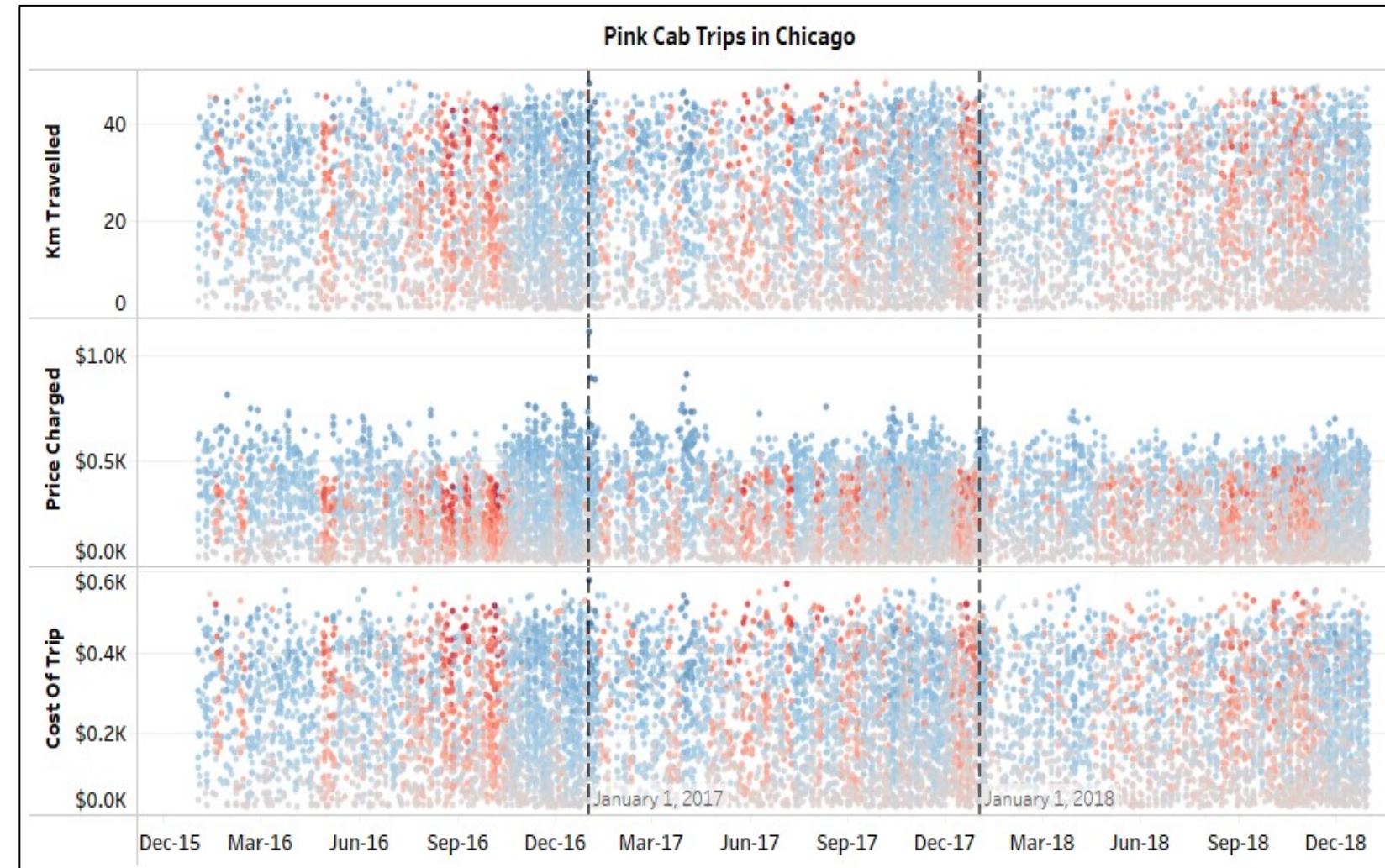
# Loses by City

- Yellow Cab makes highest total loses in Chicago, followed by Boston and Washington.
- Pink Cab makes highest total loses in Chicago and Los Angeles.



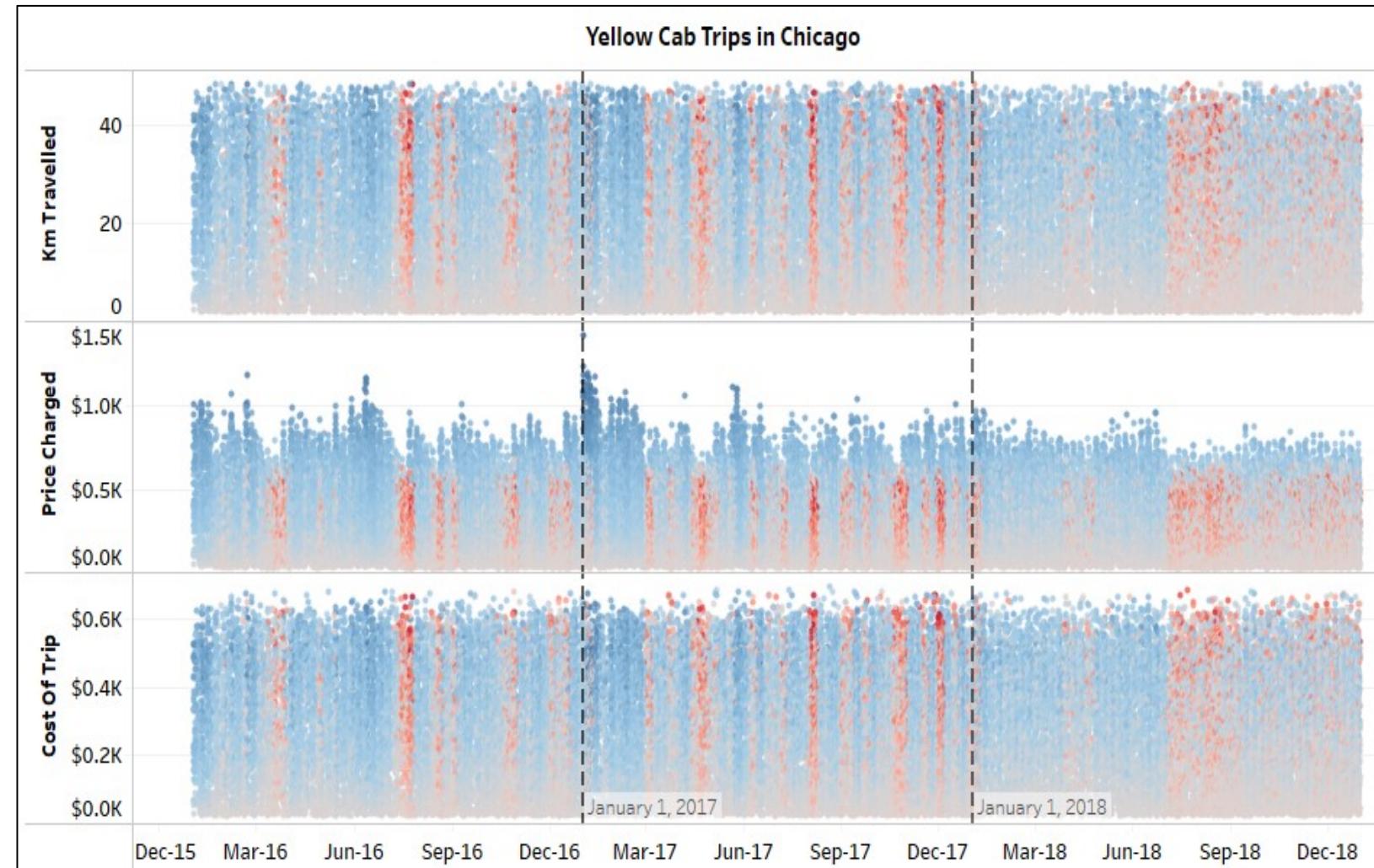
# Analyzing Chicago Non-Profitable Trips – Pink Cab

- Red bands represents trips that did not make profit (**Loss**).
- Cab expenses was higher than Cab revenue for **non-profitable** trips.
- In 2016, consistent **non-profitable** trips made during month of April, July, August, September, October.
- In 2017 and 2018, although several **non-profitable** trips were made across many months, they are less in comparison with the year 2016.



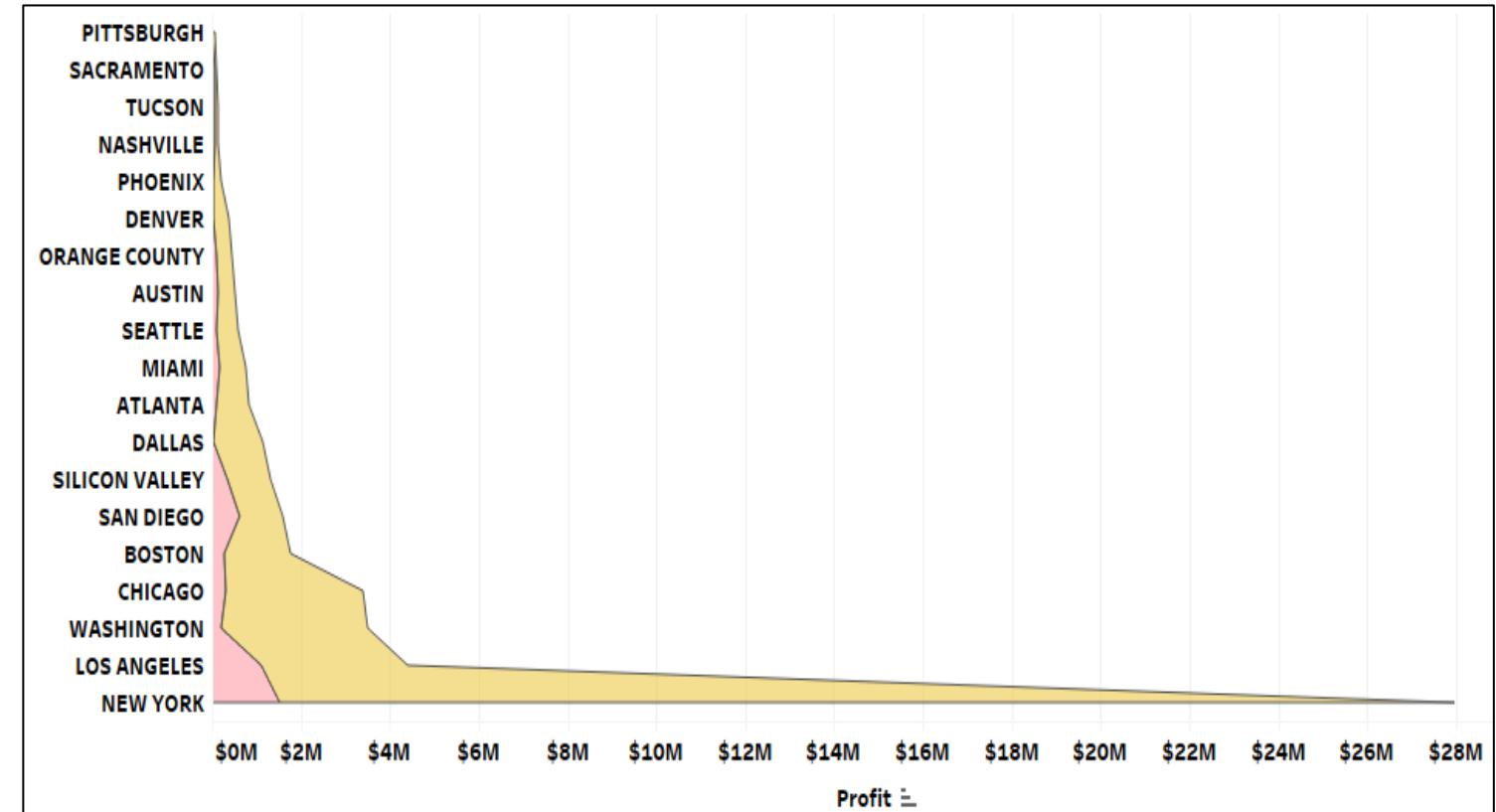
# Analyzing Chicago Non-Profitable Trips – Yellow Cab

- For **Yellow Cab**, July and August has highest number of **non-profitable** trips across all years.
- 2017 saw highest number of **non-profitable** trips across many months and 2018 saw the lowest number of **non-profitable** trips compared to other years.

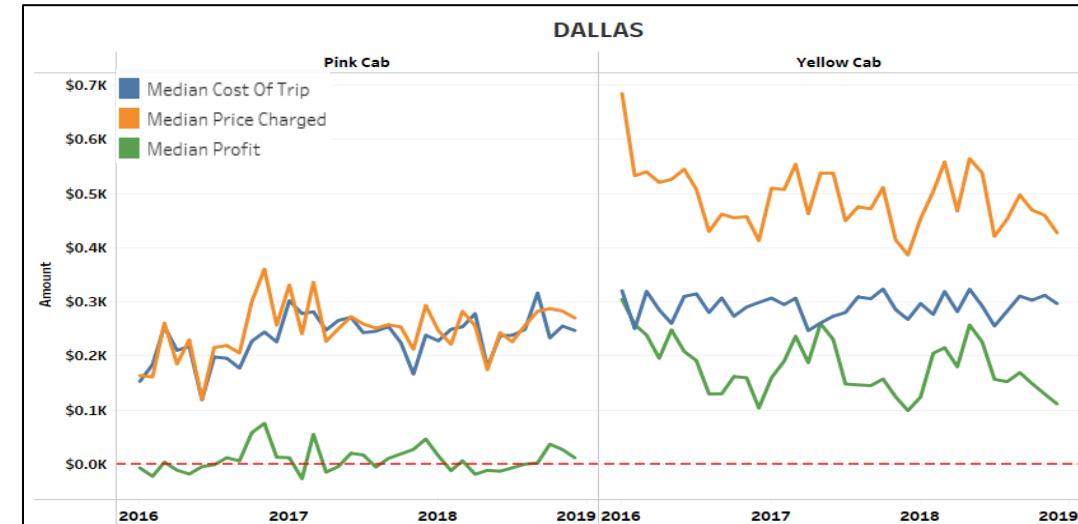
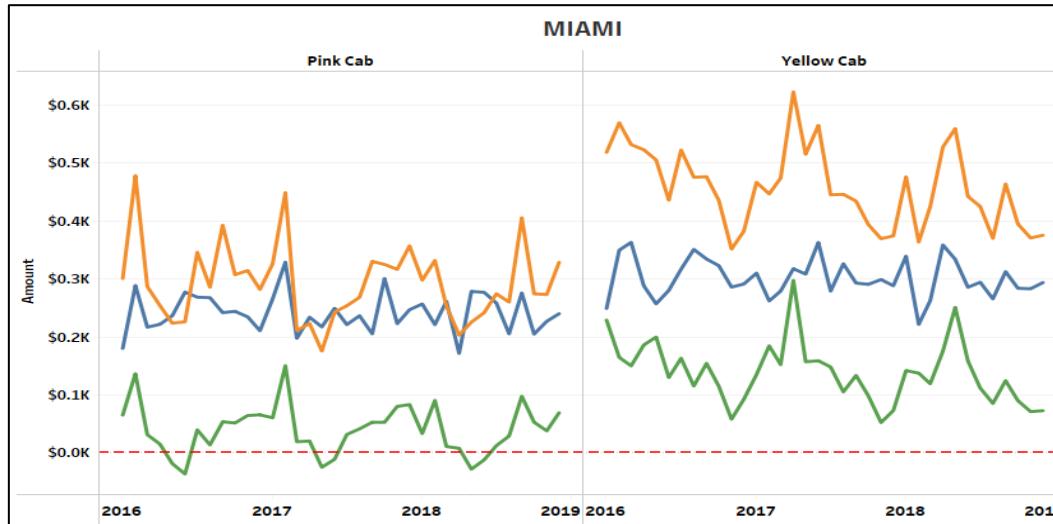


# Profit by City (2016 to 2018)

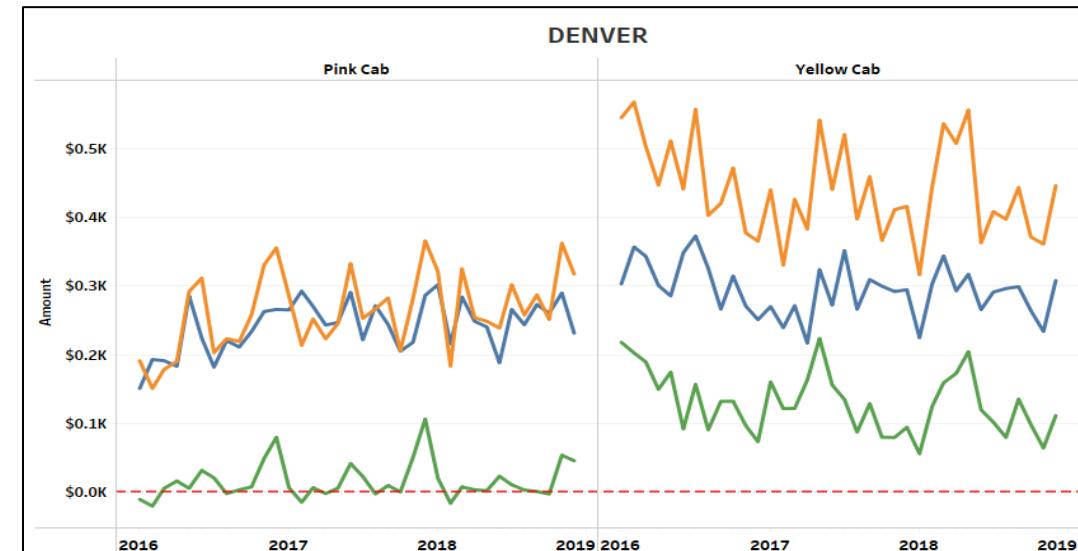
- Yellow Cab's total Profit outperforms Pink Cab across all cities.
- Highest total profit for both Cab companies is in New York.



# Monthly Timeline of Trips (1)

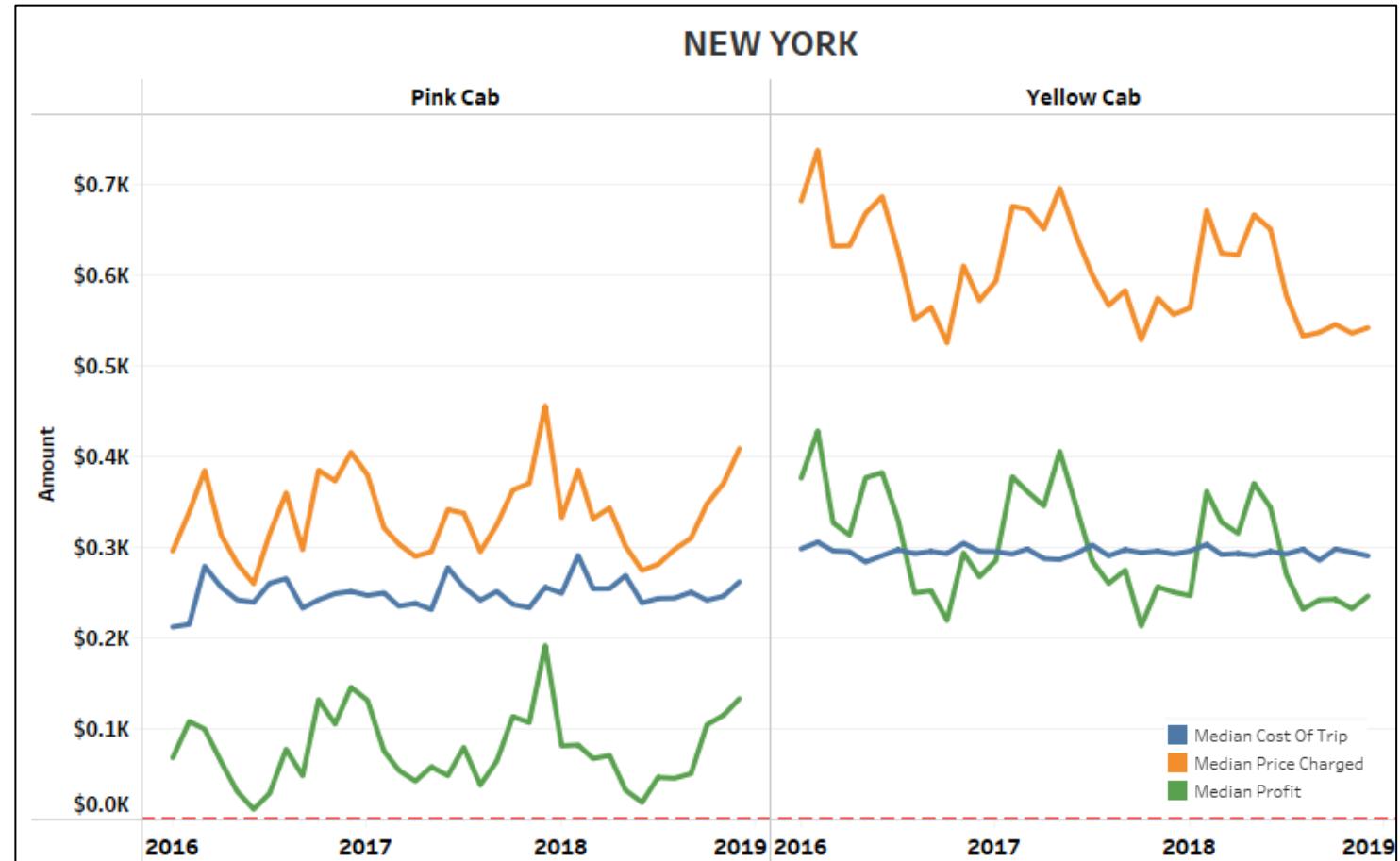


- The plots illustrates median monthly prices from 2016 to 2018 across selected cities.
- In general, **Yellow Cab's** expenses is relatively higher than **Pink Cab's**.
- In the above cities, **Pink Cab's** median profits for some months has been consistently in **loss**.
- Yellow Cab's** Profit is comparatively much higher than **Pink Cab**.



# Monthly Timeline of Trips (2)

- In New York City, although the Cab expense is similar in other cities, **Yellow Cab** charges significantly higher Cab Fares.
- Hence, **Yellow Cab** makes the highest profit in New York compared to any other city.
- Comparatively, **Pink Cab** charges very low Cab fares to its customers.



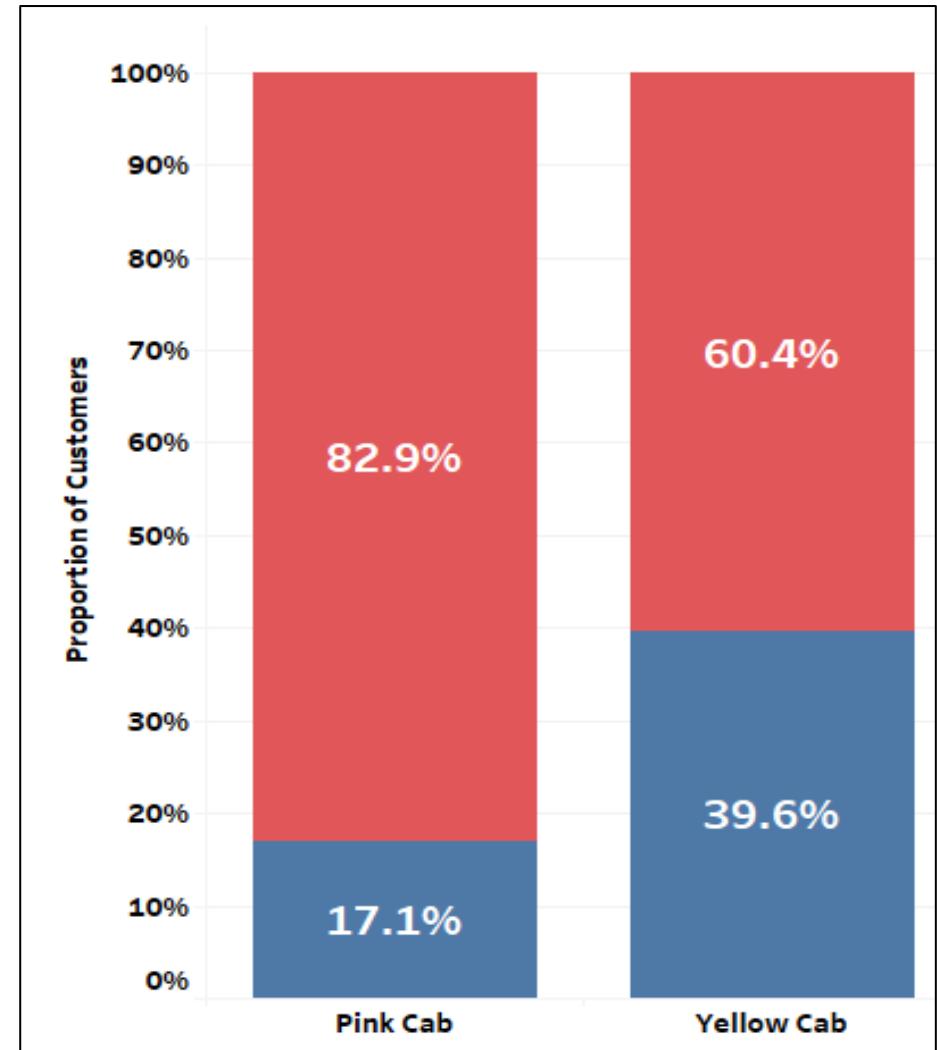
# Monthly Timeline of Trips (3)

In some cities, both companies seems to have non-seasonal fluctuating cab fares and expenses.



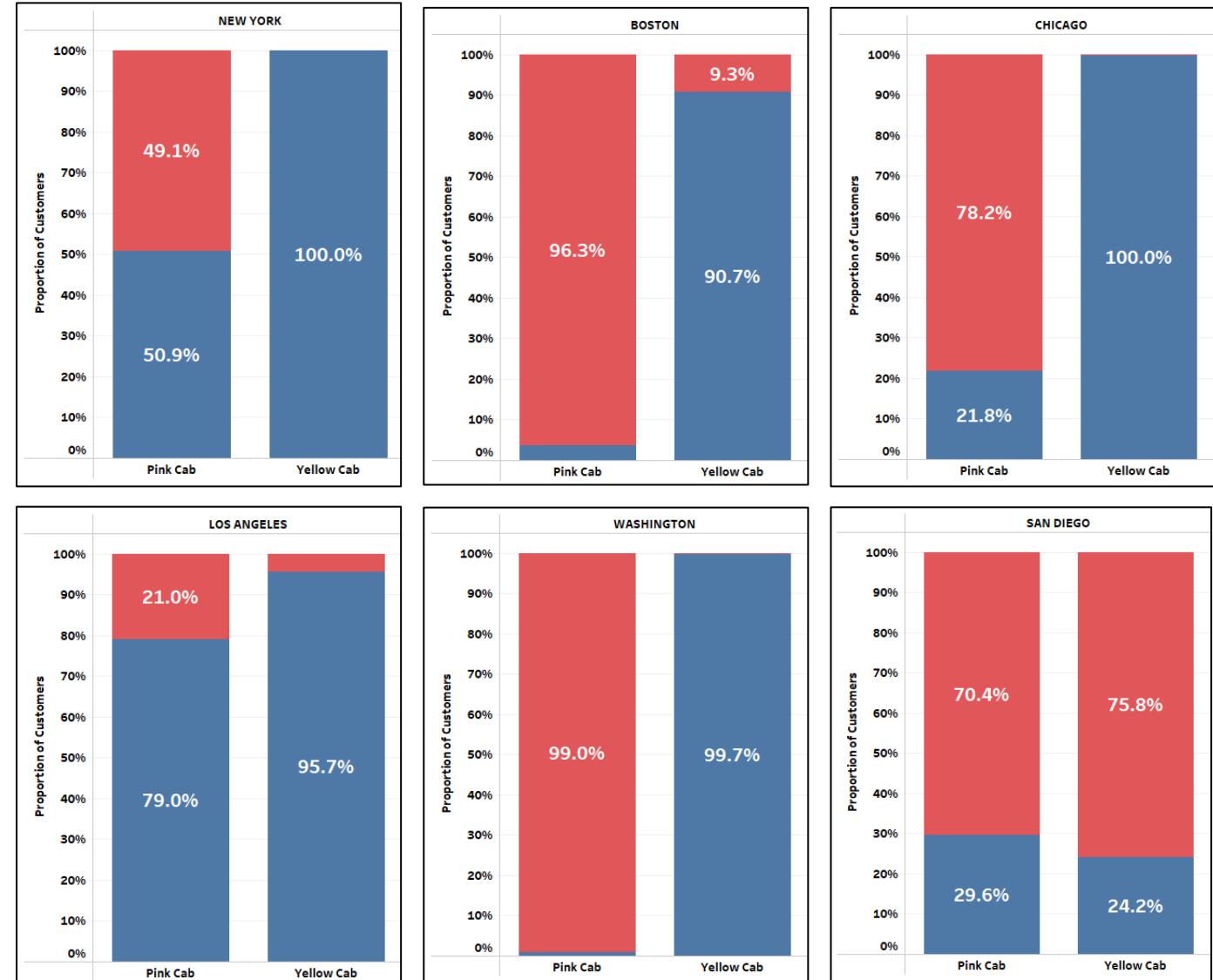
# Loyal Customers

- The plots illustrates the proportion of most loyal customer's who has used a Cab service **at least five times**.
- Overall, **Yellow Cab** has the highest proportion of loyal customer's by about 40%.



# Loyal Customers by City

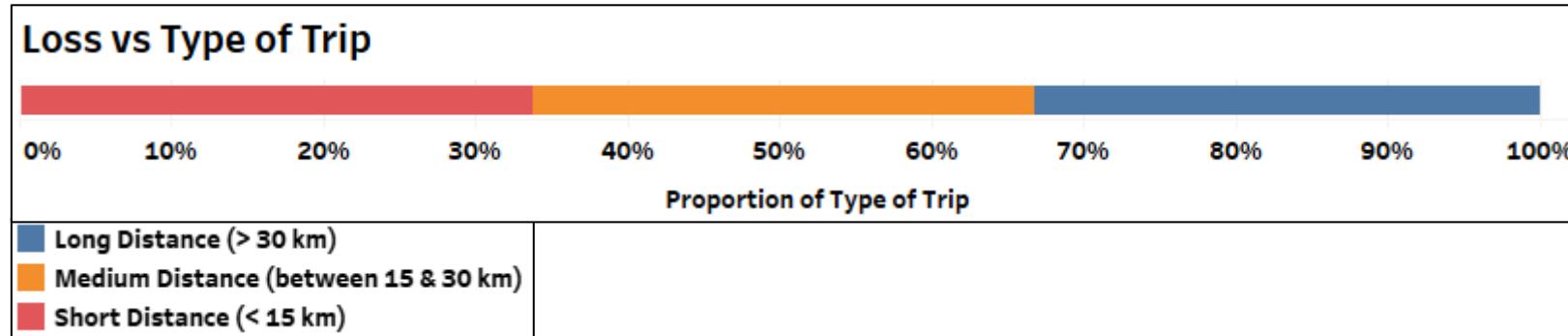
- Yellow Cab has the highest loyal customers in New York, Boston, Chicago, Los Angeles and Washington.
- Although lower than its rival, Pink Cab has significant loyal customers in Los Angeles, followed by New York.
- Only in San Diego, Pink Cab has comparatively higher loyal customers compared to its rival, at about 30%.



# Hypothesis 1:

Is there any association between Profit being at a loss and the type of trip?





- **H0: There is no association between Loss and type of trip.**
- **H1: There is an association between loss and type of trip.**

### Test Utilized : Chi-squared test

Alpha: **0.05**

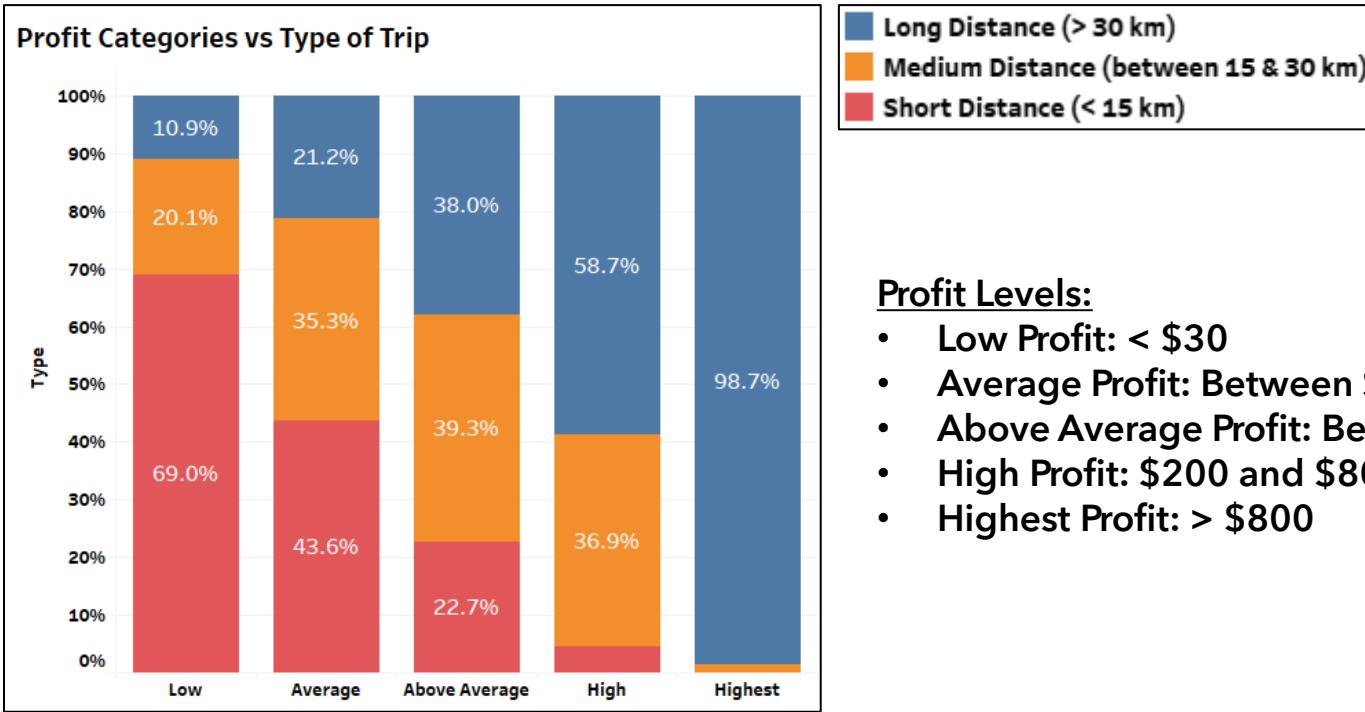
P-Value: **1.0**

**Verdict:** If the Null Hypothesis is true, **the probability that the observed number of trips is the same as the expected number of trips from a randomly sampled data is 100%**. Therefore, as the P-value is higher than the alpha value, **we fail to reject the null hypothesis**. There is **no association between Loss and type of trip**.

# Hypothesis 2:

**Is there any association between Profit and the type of trip?**





### Profit Levels:

- Low Profit: < \$30
- Average Profit: Between \$30 and \$85
- Above Average Profit: Between \$85 and \$200
- High Profit: \$200 and \$800
- Highest Profit: > \$800

- **H0: There is no association between Profit and type of trip.**  
 ➤ **H1: There is an association between Profit and type of trip.**

**Test Utilized : Chi-squared test**

Alpha: **0.05**

P-Value: **0.0**

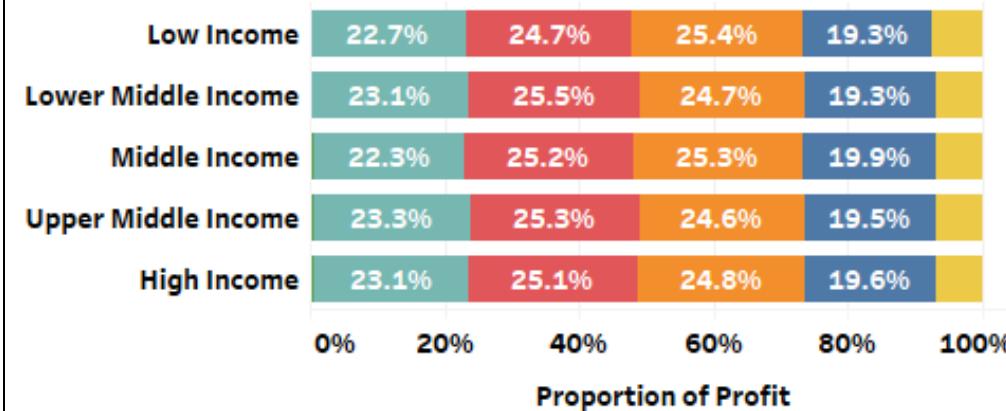
Verdict: If the Null Hypothesis is true, **the probability that the observed number of type of trips is the same as the expected number of type of trips for each profit category from a randomly sampled data is 0%.** Therefore, as the P-value is lower than the alpha value, **we reject the null hypothesis.** There **is** association between Profit and type of trip.

# Hypothesis 3:

**Is there any association between Profit  
and Customer income level?**



## Customer Income Vs Profit



Profit Level
Loss
Low
Average
Above Average
High
Highest

### Monthly Income Levels:

- Low Income: < \$2670
- Lower Middle Income: Between \$2670 and \$4451
- Middle Income: Between \$ 4451 and \$8903
- Upper Middle Income : \$ 8903 and \$20,030
- High Income : > \$ 20,030

- **H0: There is no association between Customer Income and Profit.**
- **H1: There is an association between Customer Income and Profit.**

### Test Utilized : Chi-squared test

Alpha: **0.05**

P-Value: **0.0**

**Verdict:** P-value is **statistically significant** since **P-value is lower than the alpha**. Therefore, we can **reject the null hypothesis**. There is an **association between Profit and Customer income level**.

# Hypothesis 4:

Is there an association between City and Profit?



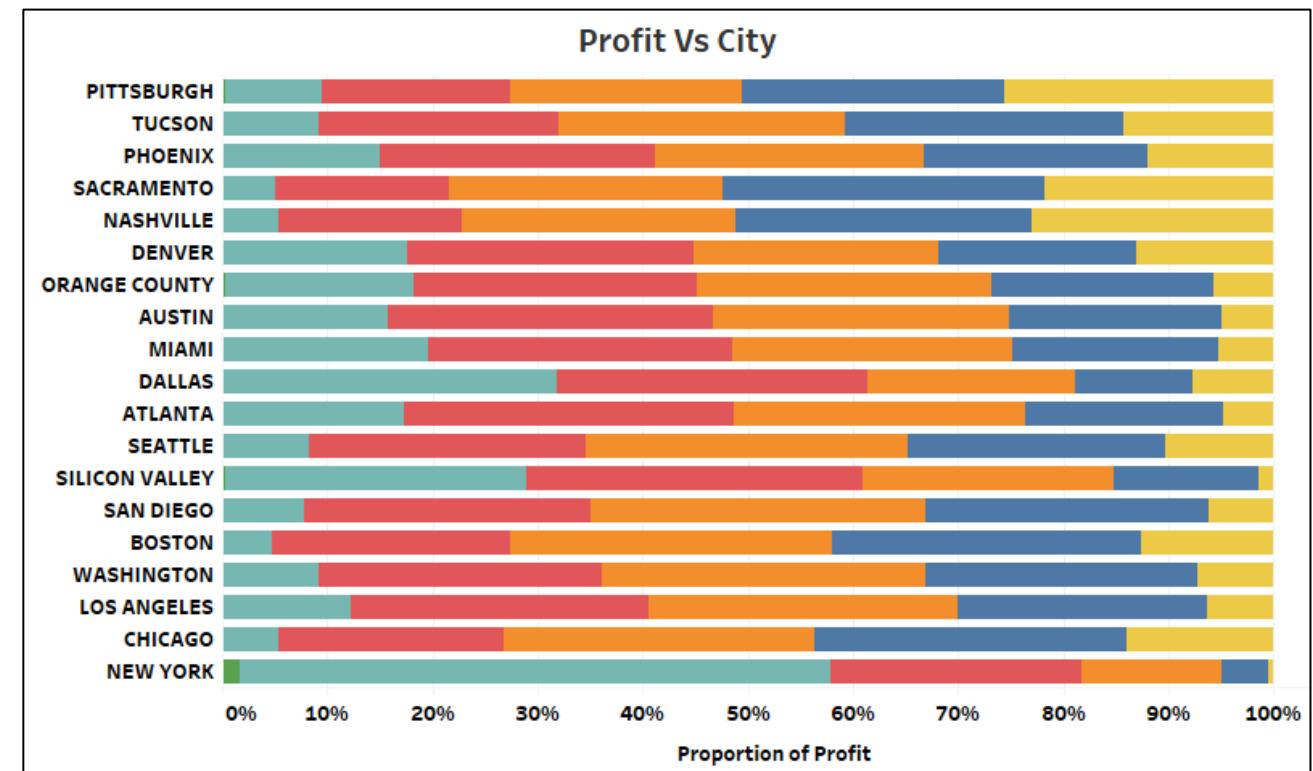
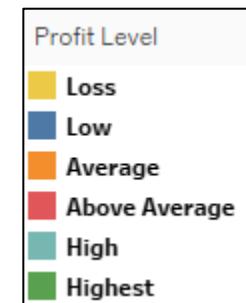
- **H0: There is no association between City and Profit.**
- **H1: There is an association between City and Profit.**

**Test Utilized : Chi-squared test**

Alpha: **0.05**

P-Value: **0.0**

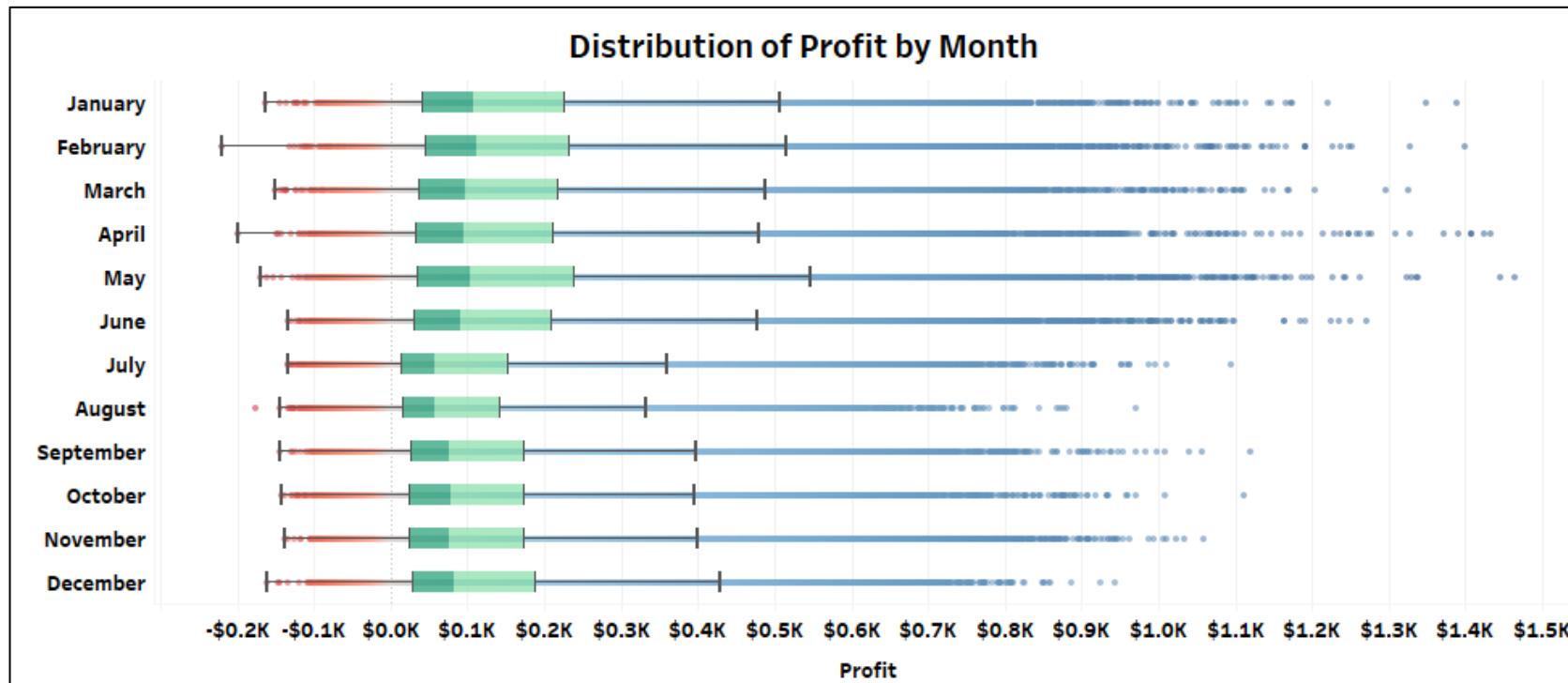
**Verdict:** P-value is **statistically significant** since **P-value is lower than the alpha.** Therefore, we can **reject the null hypothesis.** There is an **association between Profit and City.**



# Hypothesis 5:

**Does Profit vary by Month?**





- **H0: No difference in median profit across months**
- **H1: At least one month has different median profit from other months.**

### Test Utilized : Kruskal-Wallis H-test

Alpha: **0.05**

P-Value: **0.0**

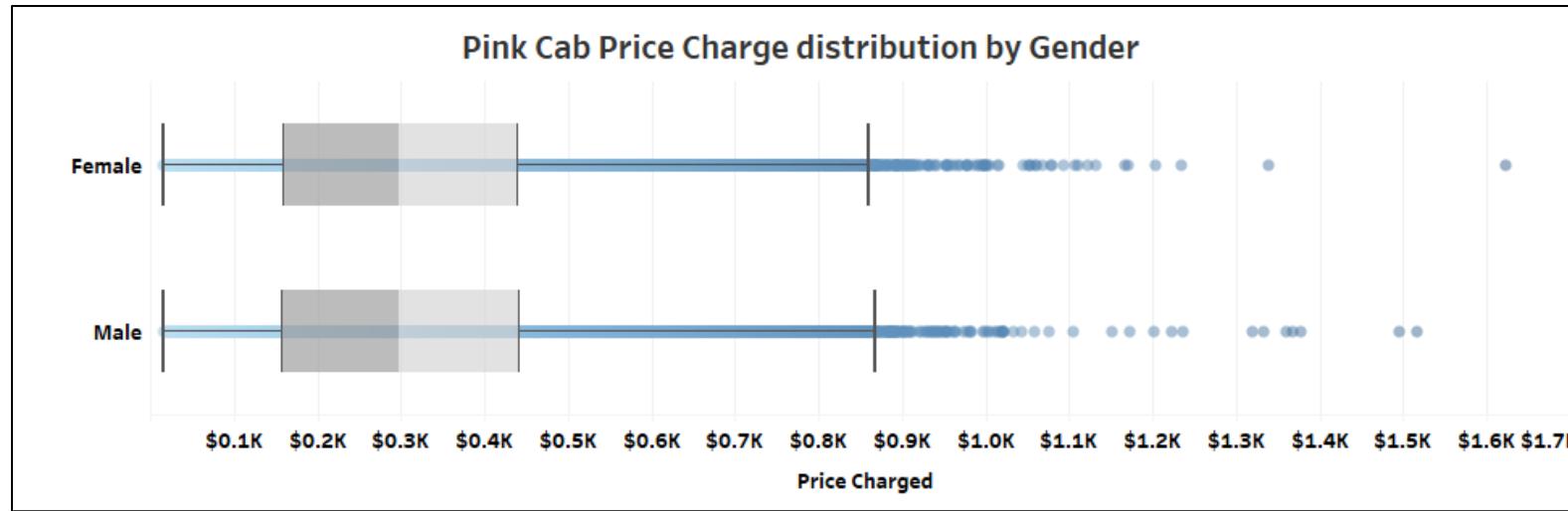
**Verdict:** If the null hypothesis is true, the probability of obtaining median profit such as shown here **will be 0%**. Since the P-value is lower than the alpha, **we reject the null hypothesis** in favor of the alternative hypothesis. **Median profit do change at least for one month.**

# Hypothesis

## 6.1:

Does Pink Cab charge its customer  
differently based on gender?





- **H0: No difference in median price charged among gender**
- **H1: There is a difference in median price charged among gender.**

### Test Utilized : Mood's median test

Alpha: **0.05**

P-Value: **0.278**

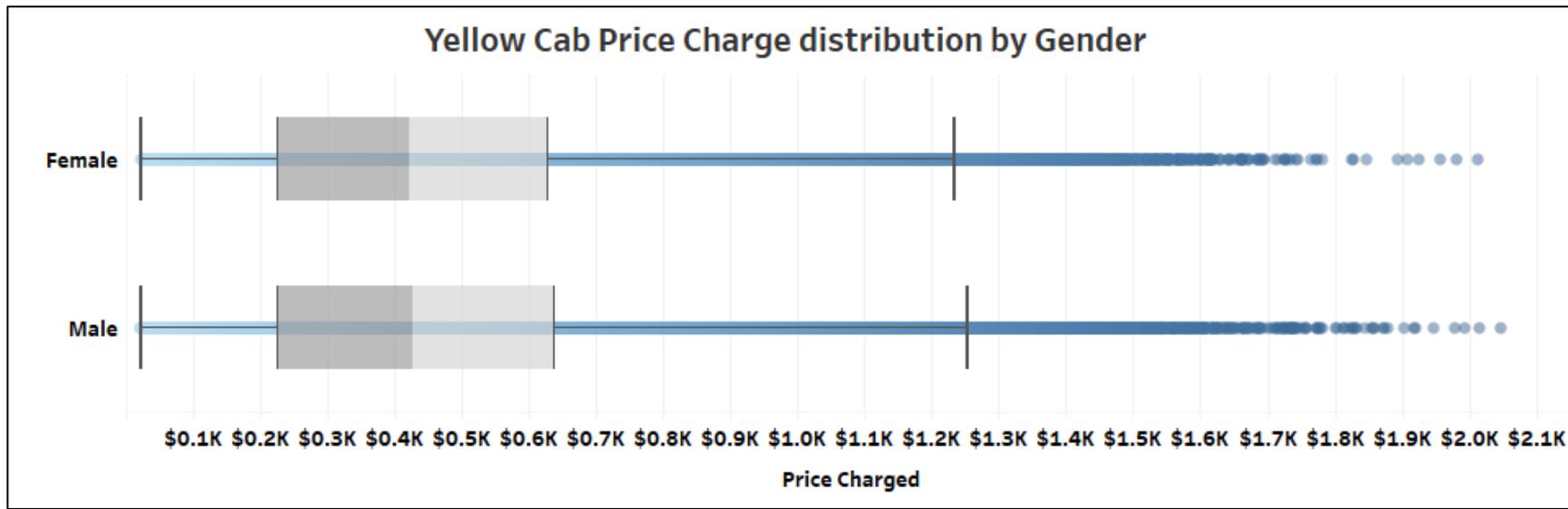
**Verdict:** If the null hypothesis is true, the probability of obtaining median price charged such as shown here **will be 27.8%**. Since the P-value is higher than the alpha, **we fail to reject the null hypothesis** in favor of the alternative hypothesis. **Median price charged remains the same for Pink Cab's customers regardless of gender.**

# Hypothesis

## 6.2:

Does Yellow Cab charge its customer  
differently based on gender?





- **H0: No difference in median price charged among gender**
- **H1: There is a difference in median price charged among gender.**

### **Test Utilized : Mood's median test**

Alpha: **0.05**

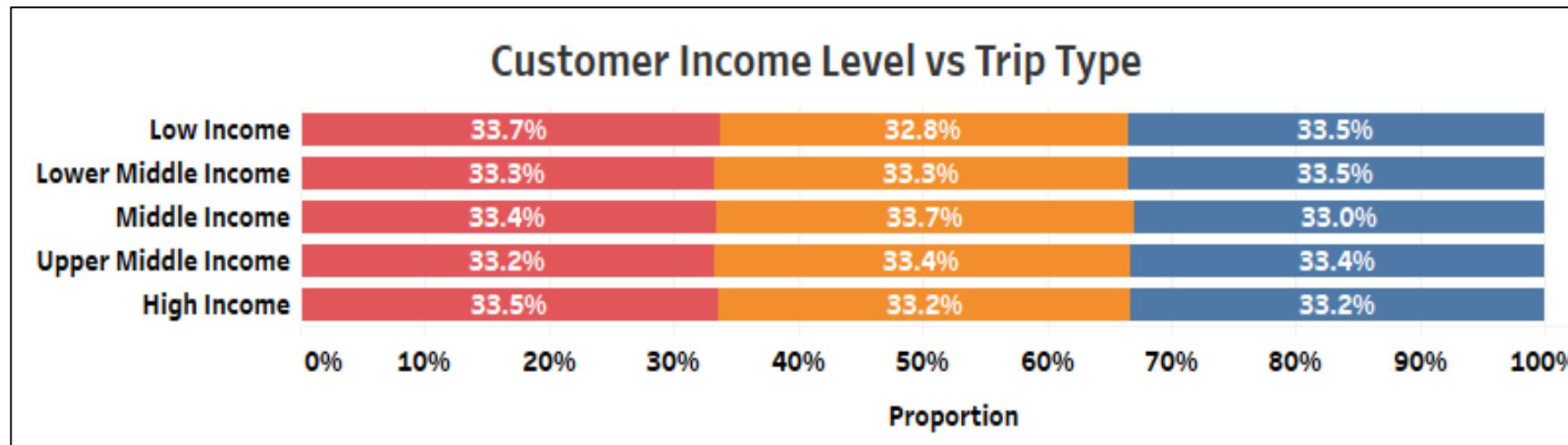
P-Value: **0.0**

**Verdict:** If the null hypothesis is true, the probability of obtaining median price charged such as shown here **will be 0%**. Since the P-value is lower than the alpha, **we reject the null hypothesis** in favor of the alternative hypothesis. **Median price charged is different for Yellow Cab's customers based on gender.**

# Hypothesis 7:

Is there an association between Customer Income Level and Type of Distance traveled?





- **H0: There is no association between Customer Income and Distance Traveled.**
- **H1: There is an association between Customer Income and Distance Traveled.**

**Test Utilized : Chi-squared test**

Alpha: **0.05**

P-Value: **0.427**

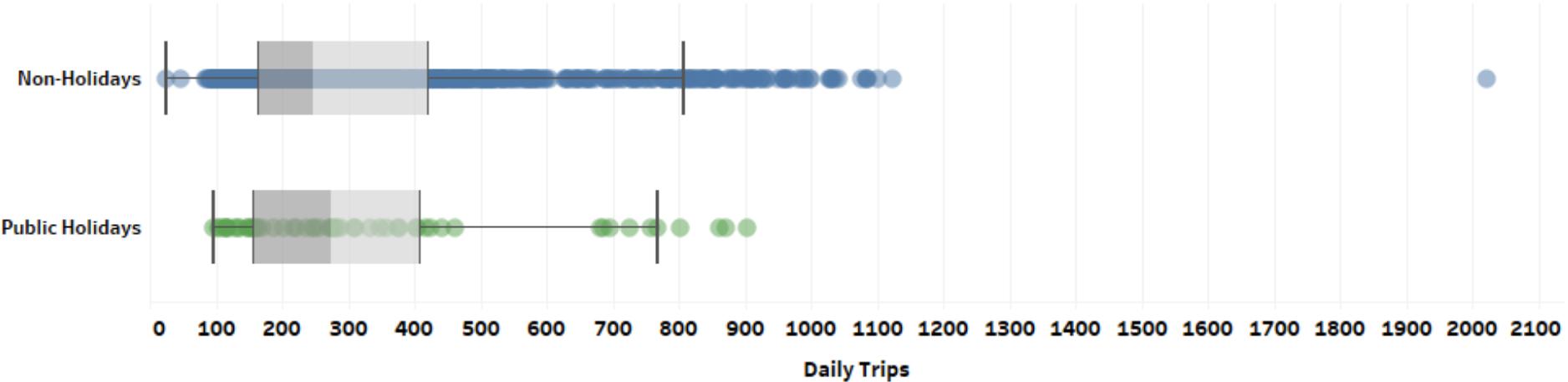
**Verdict:** P-value is **statistically insignificant** since **P-value is higher than the alpha**. Therefore, we **cannot reject the null hypothesis**. There is no **association between Customer Income and Distance Traveled**.

# Hypothesis 8:

Does Public Holidays affect daily number of trips?



## Daily trips during both Public and Non-Holidays



- H0: No difference in median daily trips during both types of days.
- H1: There is a difference in median daily trips during both types of days.

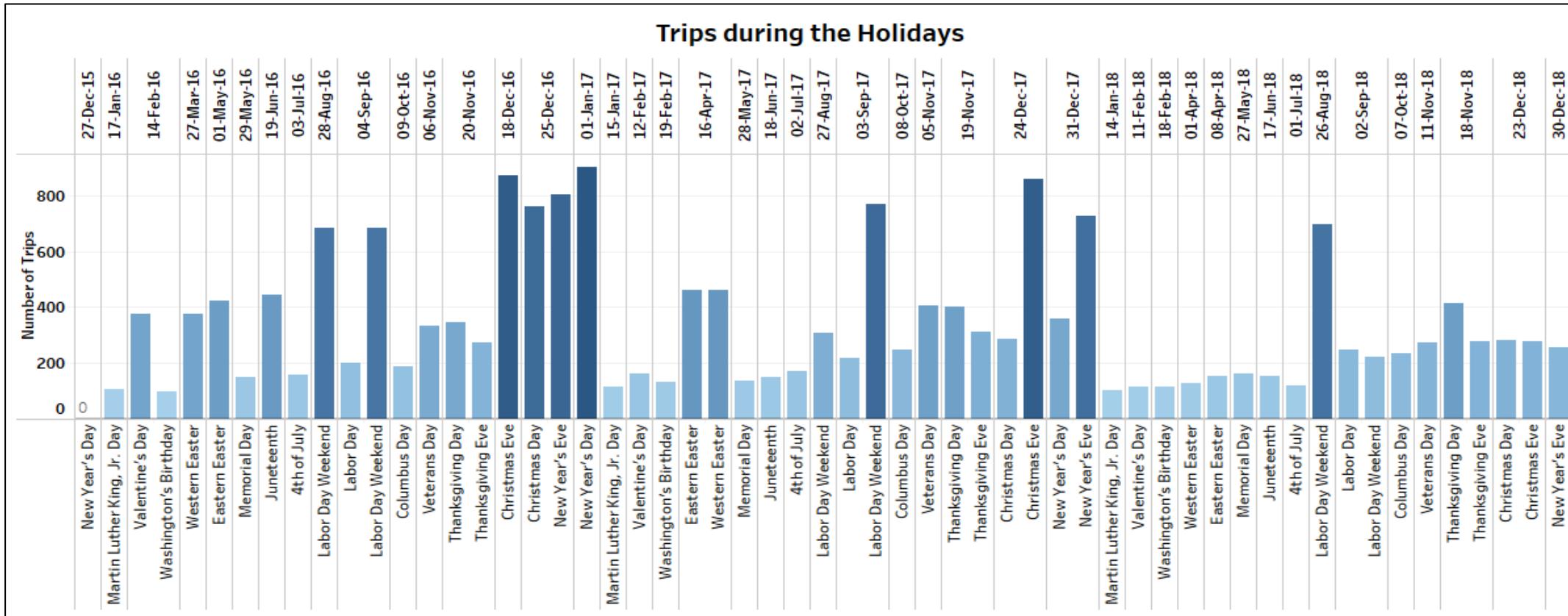
### Test Utilized : Mood's median test

Alpha: **0.05**

P-Value: **0.403**

**Verdict:** If the null hypothesis is true, the probability of obtaining the results of median daily trips such as shown here **is 40.3%**. Since the P-value is higher than the alpha, **we fail to reject the null hypothesis** in favor of the alternative hypothesis. **There is no significant difference in median trips on both types of days.**

# Trips during Public Holidays



- The **lowest trips** during public holidays for all three years were on **Martin Luther King Jr. Day** and **Washington's Birthday**.
- During Public Holidays, customers **travel the most** during **Labor Day Weekend**, followed by **Christmas Eve** and **New Year's Eve**.
- In 2016, there were almost no trips on New Year's Day according to the data. But New Year's Day in 2017 is observed to have a high spike in trips. In 2018, trips on this day has reduced significantly.

# Forecasting

Forecasting Daily Number of Trips for the  
Next Two Years



# Forecasting Strategy

**Goal:** Forecast Individual Company's Daily Number of Trips

**Model Used:** Facebook Prophet 

**Evaluation Metric:** Root Mean Squared Error (RMSE)

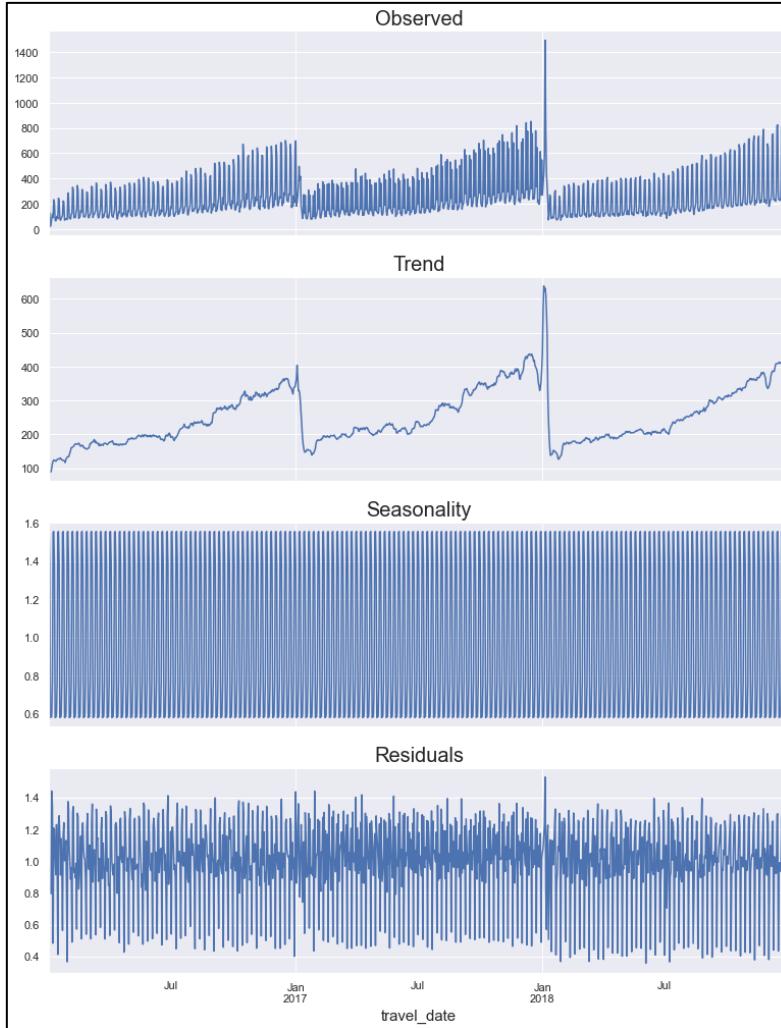
**Problem Type:** Supervised Learning

## Steps Involved:

1. Initially **decomposing** Daily Trip Time Series.
2. Analyze **Auto Correlation and Partial-Auto Correlation Plots**.
3. Create 'n' **lags as independent variables**.
4. **Split** the time series into **Train and Test sets**. Test set will be for a period of **last 90 days**.
5. Train Prophet Model **using only Train set**.
6. **Evaluate model performance** by comparing predictions with Test set.
7. Train on the whole dataset and **forecast daily trips for next two years**.
8. Compare forecasts of both Cab companies.

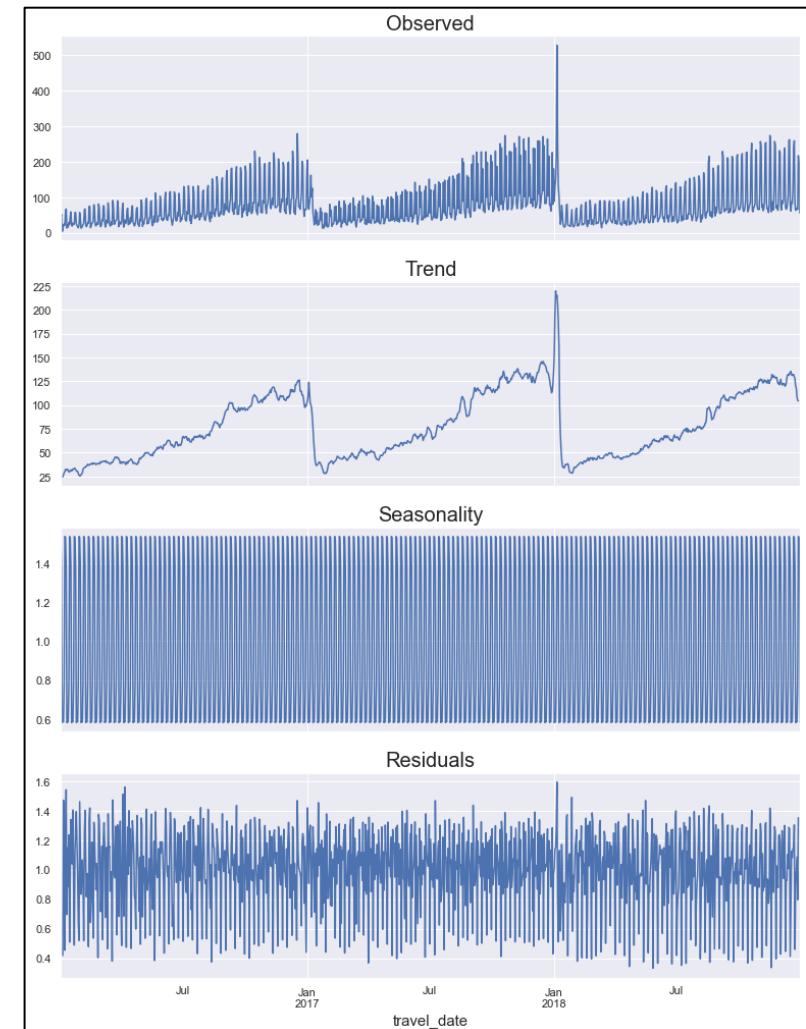
# Daily Trips Time Series Decomposition

Yellow Cab

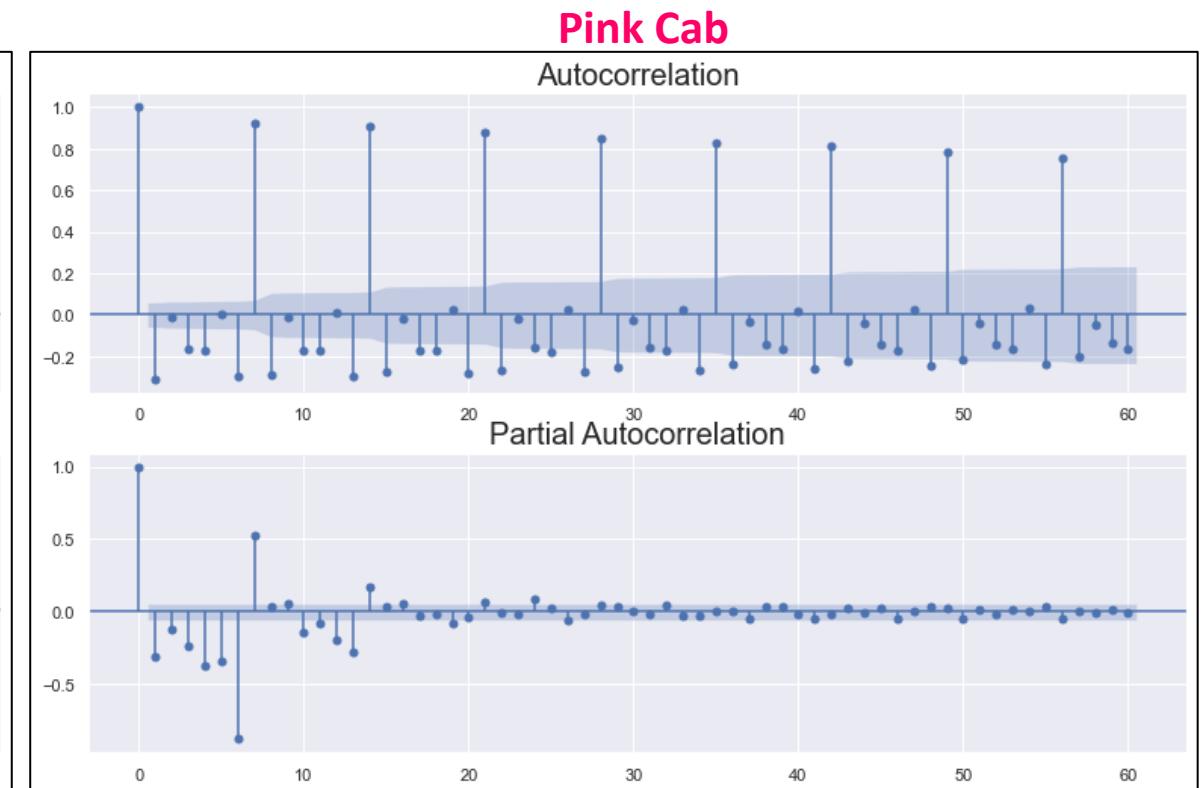
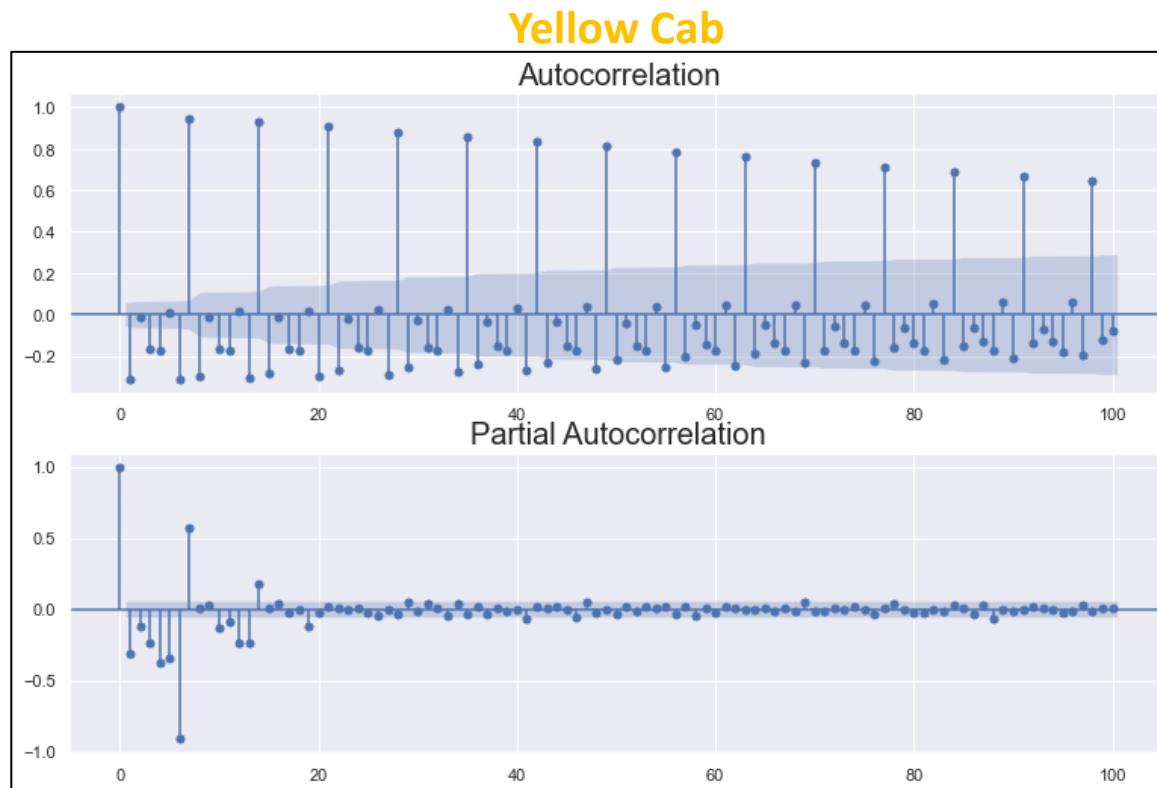


- The plots illustrates a breakdown of Daly trip time series into **trend**, **seasonal** and **residuals (error)** components.
- **Multiplicative model** used due non-constant monthly and yearly seasonality.
- **Trend:** Noticeable upward month-wise trend but the annual trend seems constant for both companies.
- **Seasonality:** Constant seasonality across all weeks.
- **Residuals:** No noticeable pattern.

Pink Cab



# ACF & PACF



- Both time series were differenced one time to make the series stationary.
- High auto correlation with the 7th lag, illustrating that the number of trips at present is correlated with the number of trips within the past week (7 days).

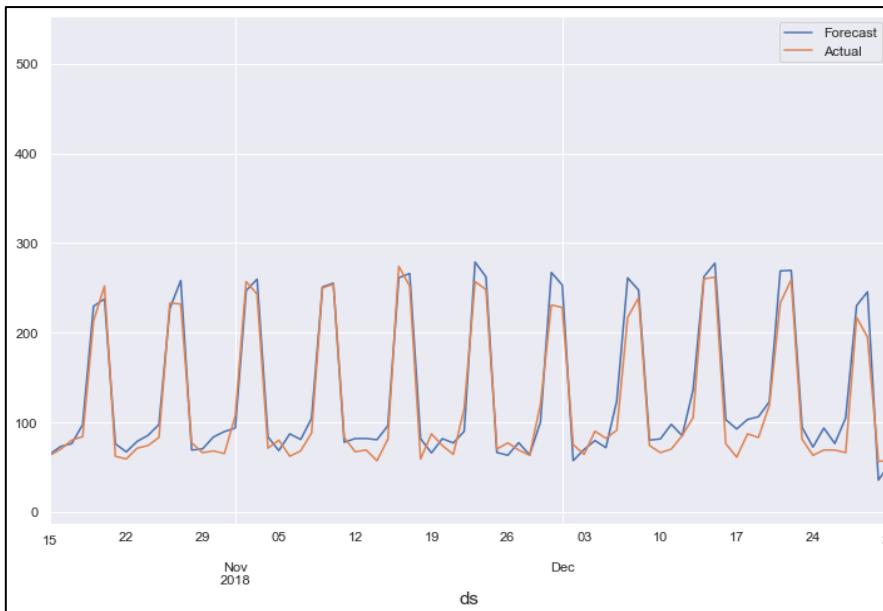
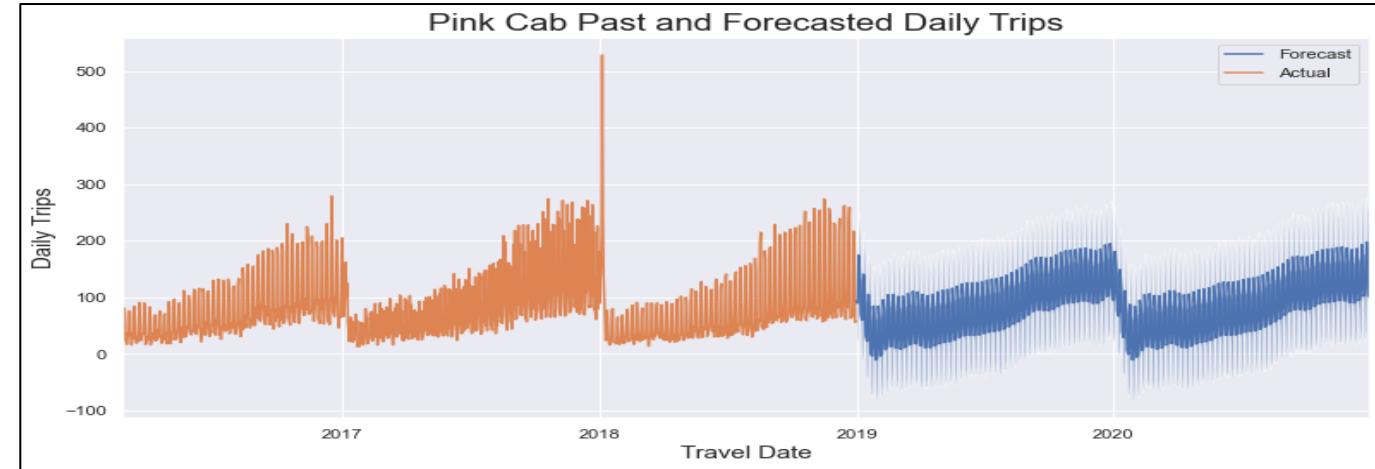
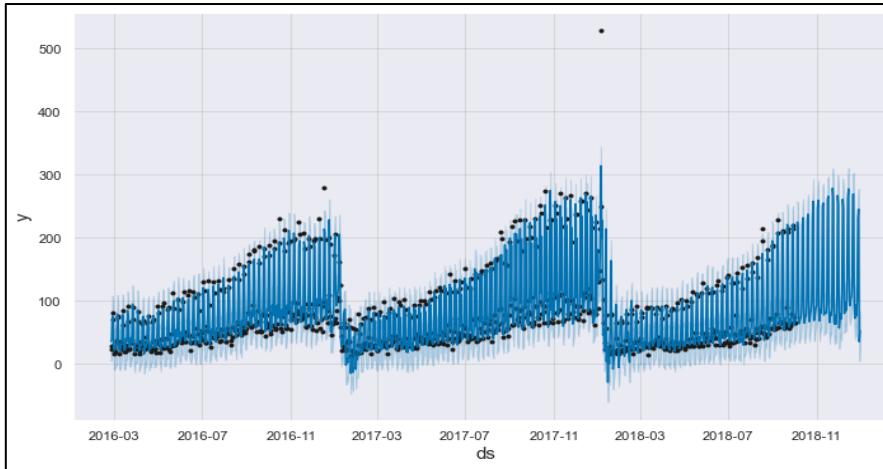
# Model Training

- Created **lags** as independent variables by differencing 54 days.
- Additionally, extracted **day, month, week, etc.** as additional independent variables.
- Data split into train (**01/01/2016 to 14/10/2018**) and test (**15/10/2018 to 31/12/2018**) sets.

	ds	y	year	month	week	date	day_of_week	t-54	t-53	t-52	...	t-10	t-9	t-8	t-7	t-6	t-5	t-4	t-3	t-2	t-1
0	2016-02-25	29	2016	2	8	25	3	41.0	52.0	4.0	...	14.0	18.0	27.0	28.0	32.0	79.0	65.0	23.0	19.0	23.0
1	2016-02-26	24	2016	2	8	26	4	52.0	4.0	6.0	...	18.0	27.0	28.0	32.0	79.0	65.0	23.0	19.0	23.0	29.0
2	2016-02-27	81	2016	2	8	27	5	4.0	6.0	23.0	...	27.0	28.0	32.0	79.0	65.0	23.0	19.0	23.0	29.0	24.0
3	2016-02-28	70	2016	2	8	28	6	6.0	23.0	24.0	...	28.0	32.0	79.0	65.0	23.0	19.0	23.0	29.0	24.0	81.0
4	2016-02-29	25	2016	2	9	29	0	23.0	24.0	21.0	...	32.0	79.0	65.0	23.0	19.0	23.0	29.0	24.0	81.0	70.0

Training Dataset

# Model Evaluation – Pink Cab

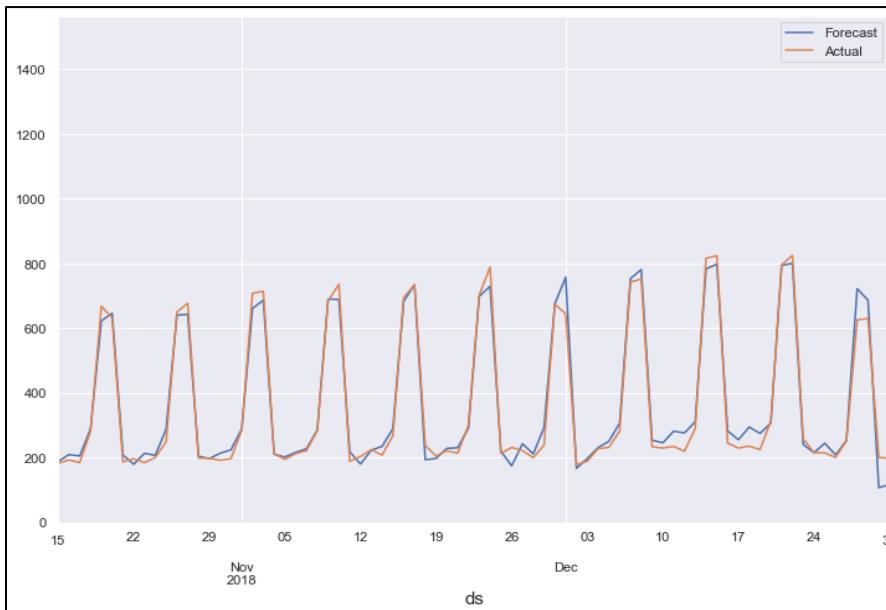
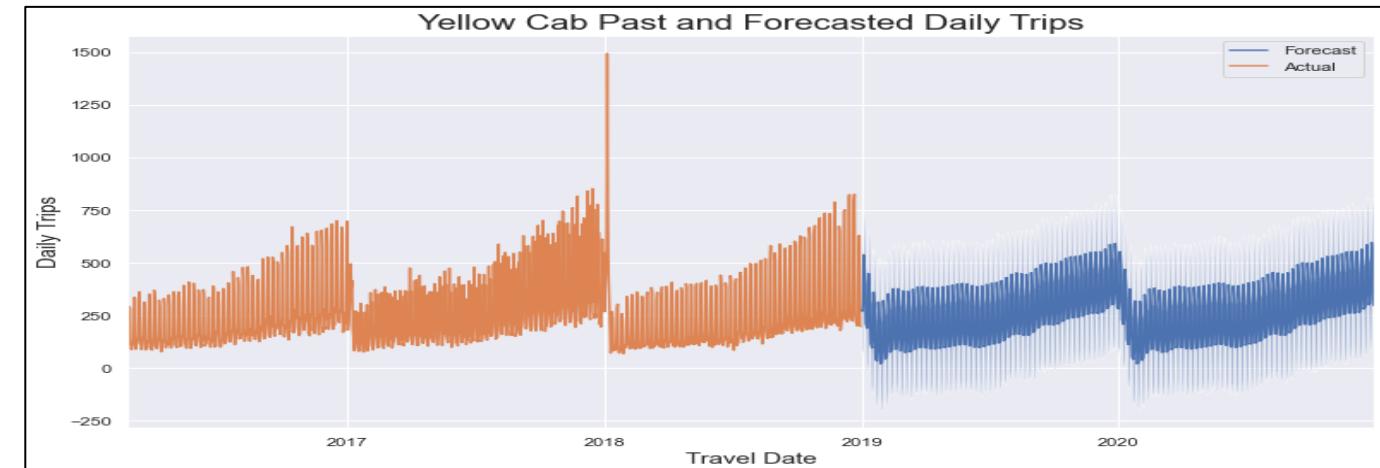
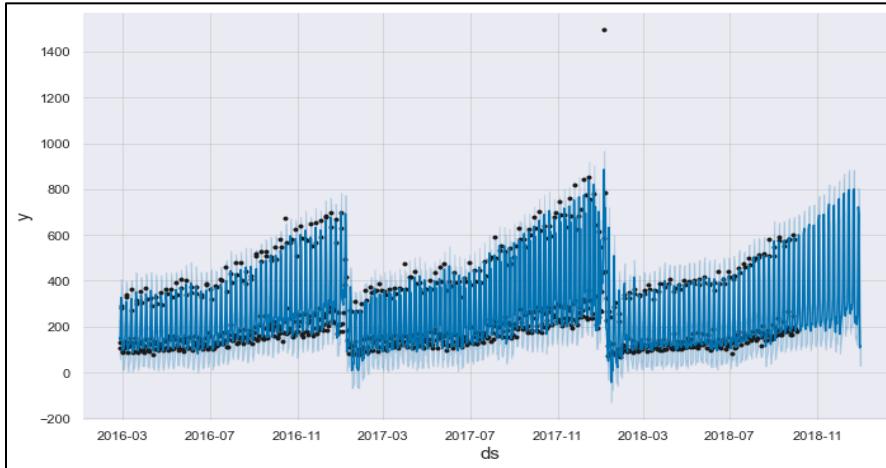


- **RMSE = 17.62 days**
- **Mean Test trips = 79.69 trips**
- **Model Accuracy = 88.36%**

The predicted line approximately follows the same pattern as the test set. The model seems to **over predict** trips on all days.

Forecasting for the next two years seems to illustrate same trend as before.

# Model Evaluation – Yellow Cab



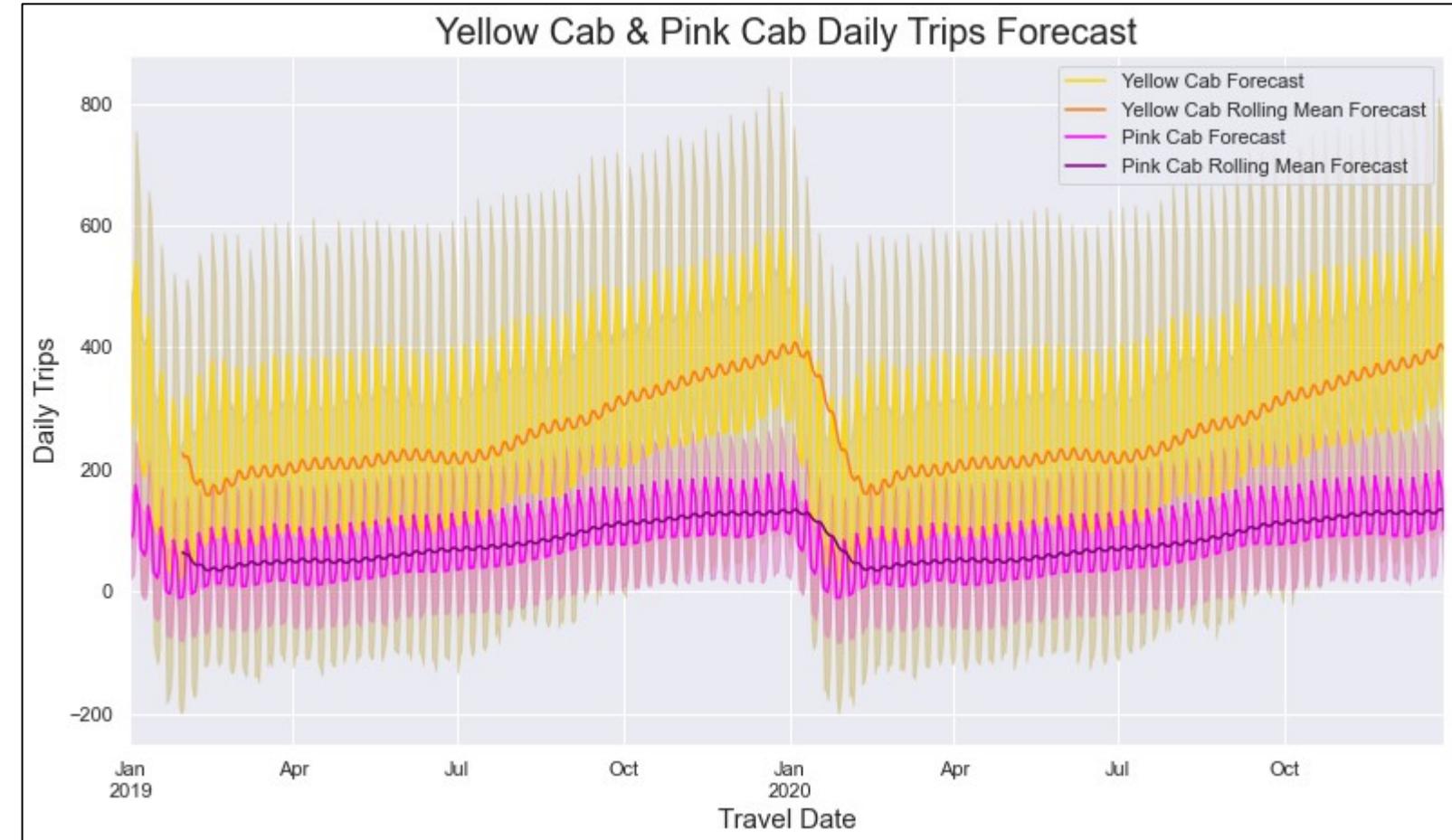
- **RMSE** = 32.5 trips
- **Mean Test trips** = 256.67 trips
- **Model Accuracy** = 93.42%

The predicted line nearly follows the same pattern as the test set, much better than **Pink Cab's** forecasts. The model also seems to have a better accuracy.

Forecasting for the next two years seems to illustrate same trend as before.

# Forecast Comparison

- The plots depicts the forecasts made by the model for the next two years (2019 to 2020).
- The confidence intervals of **Yellow Cabs** are wider compared to **Pink Cabs**.
- When comparing both forecasts side by side, we can see **Yellow Cab** company is projected to still lead in daily trips compared to its rival.



# Conclusion

To conclude, based on the extensive EDA, hypothesis testing, and forecasting done on the datasets provided, I highly recommend **XYZ** to invest in **Yellow Cab** for the following reasons:

- ✓ **Higher market share across all cities, especially in New York.**
- ✓ **Higher number of trips every single day.**
- ✓ Despite losses during specific months, **Yellow Cab** still **outperforms Profit wise.**
- ✓ **Higher customer loyalty.**
- ✓ **Forecasts for next two years shows Yellow Cab will still outperform its rival every single day.**

# Thank You!

