Exploratory Data Analysis

G2M insight for Cab Investment firm

10/16/2022

Agenda

- Executive Summary
- Problem Statement
- Approach
- EDA
- EDA Summary
- Recommendations

Executive Summary

Background:

Due to remarkable growth in the Cab Industry in the last few years and multiple key players in the market, XYZ is planning an investment in the Cab industry and wants to understand the market.

Purpose:

Identify the right company to make more profits between two companies (Pink Cab and Yellow Cab)

Method:

EDA (explore distributions, comparisons, and time series analysis using stacked bar charts, line charts)
Forecast

Finding:

The **yellow cab** performs better than the pink one in many ways and should be chosen for Investment

Problem Statement

Original Data:

- Cab_Data.csv: this file includes details of transaction for 2 cab companies
- Customer_ID.csv: this is a mapping table that contains a unique identifier which links the customer's demographic details
- Transaction_ID.csv: this is a mapping table that contains transaction to customer mapping and payment mode
- City.csv: this file contains list of US cities, their population and number of cab users

- Time period: from 31/01/2016 to 31/12/2018

Post-processed Data:

- Features: 'Date of Travel', 'Company', 'City', 'Users', 'Payment_Mode', 'Gender', 'KM Travelled', 'Price Charged', 'Cost of Trip', 'Population', 'Customer ID', 'Age', 'Income (USD/Month)', 'Profit', 'Cost per KM', 'Price per KM', 'Profit per KM'
- Size: 359392

Approach

- Data clean for int type, date type, and string type
- Merge files together based on primary keys and foreign keys
- Data quality check, including missing one
- Extract more features by feature engineering
- Forecast based on time series

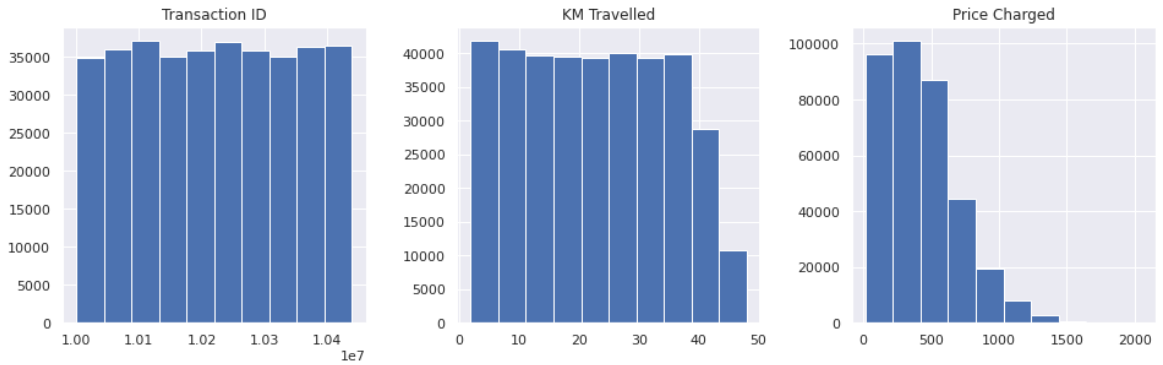
Analysis directions:

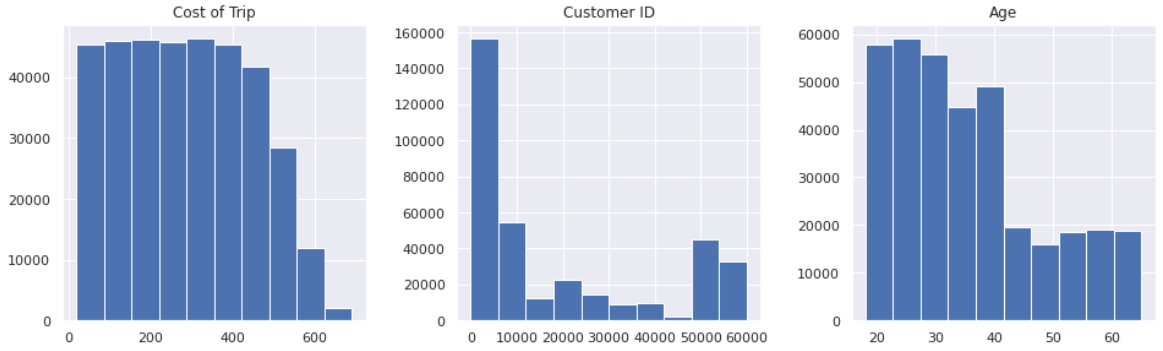
- Profit analysis
- City wise
- Gender wise
- Income wise
- Age wise
- Distance wise
- Time series analysis

EDA

- Profit analysis
- City wise
- Gender wise
- Income wise
- Age wise
- Distance wise
- Time series analysis

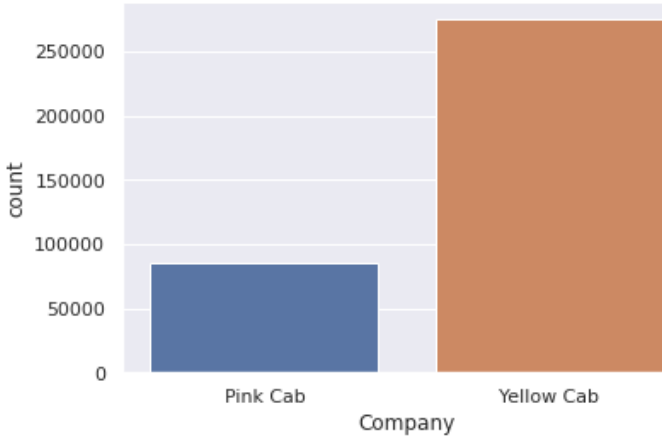
Summary



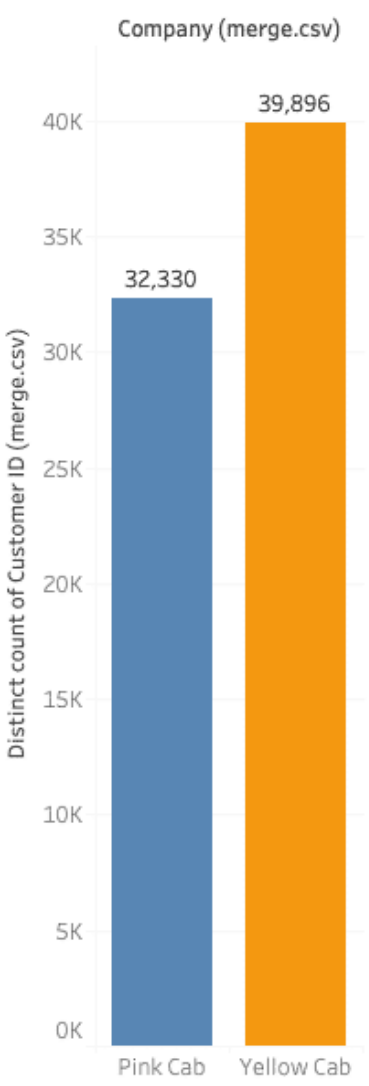


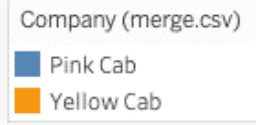
No missing values

Transaction numbers



User numbers

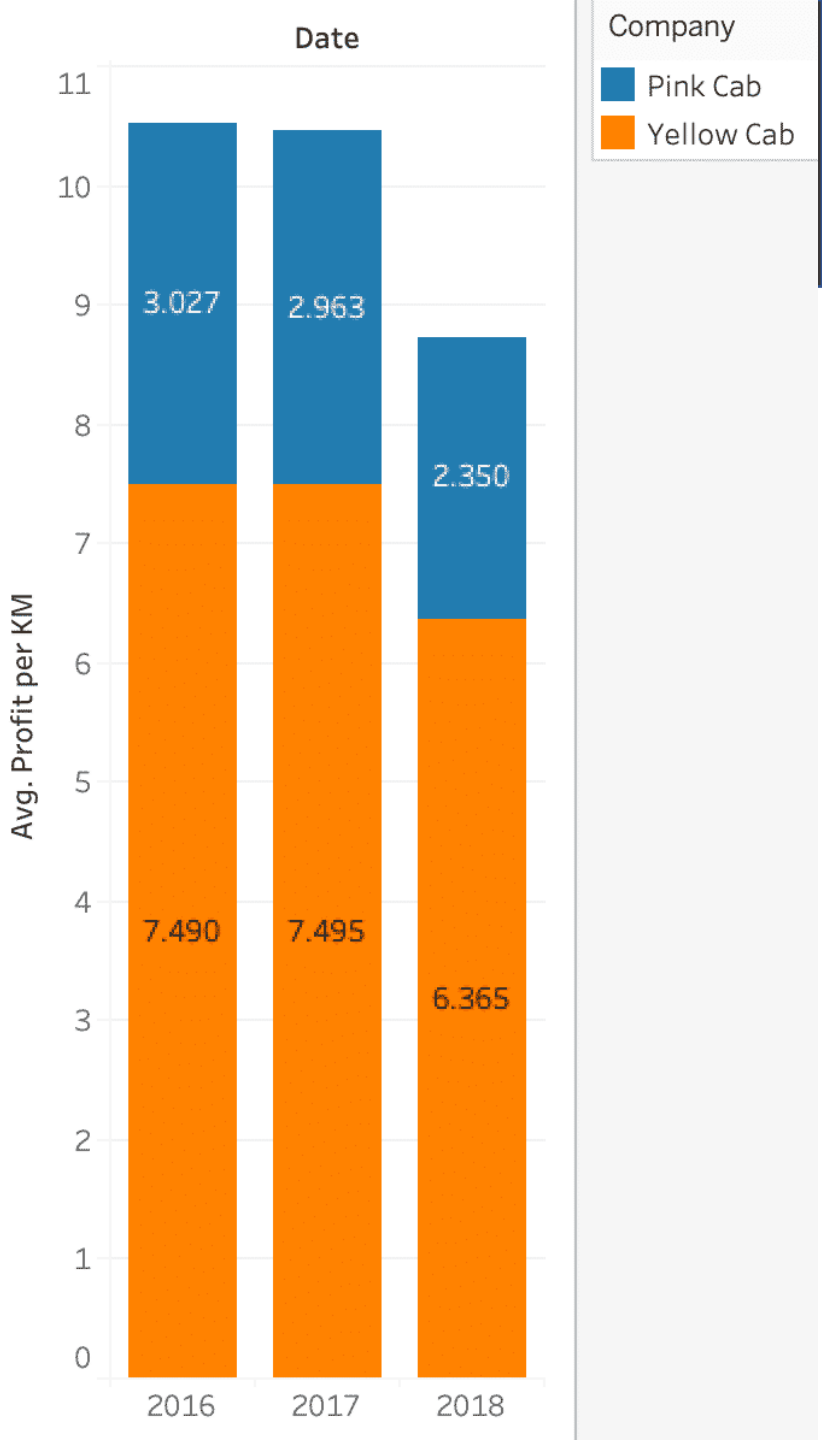




Profit Analysis

Hypothesis: How much will cabs earn per KM?

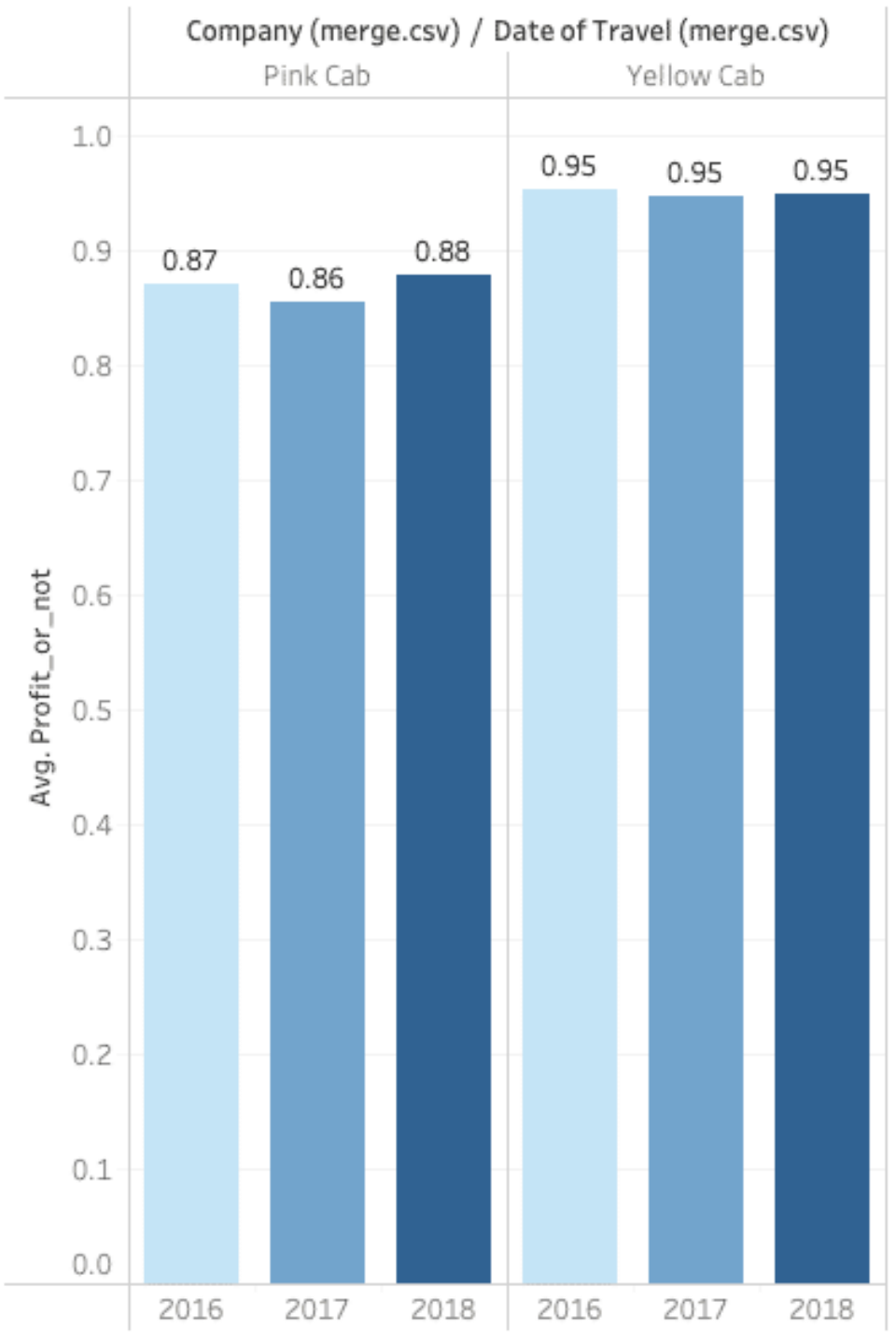
Conclusion: yellow cab earns at least two times profits than pink cab per KM.

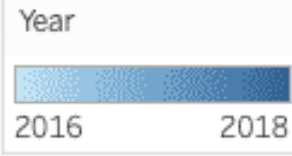


Profit Analysis

Hypothesis: What are the percentages that a ride can make a profit?

Conclusion: yellow cab has a higher percentage, which means it is more possible for the yellow cab to make a profit for every ride.

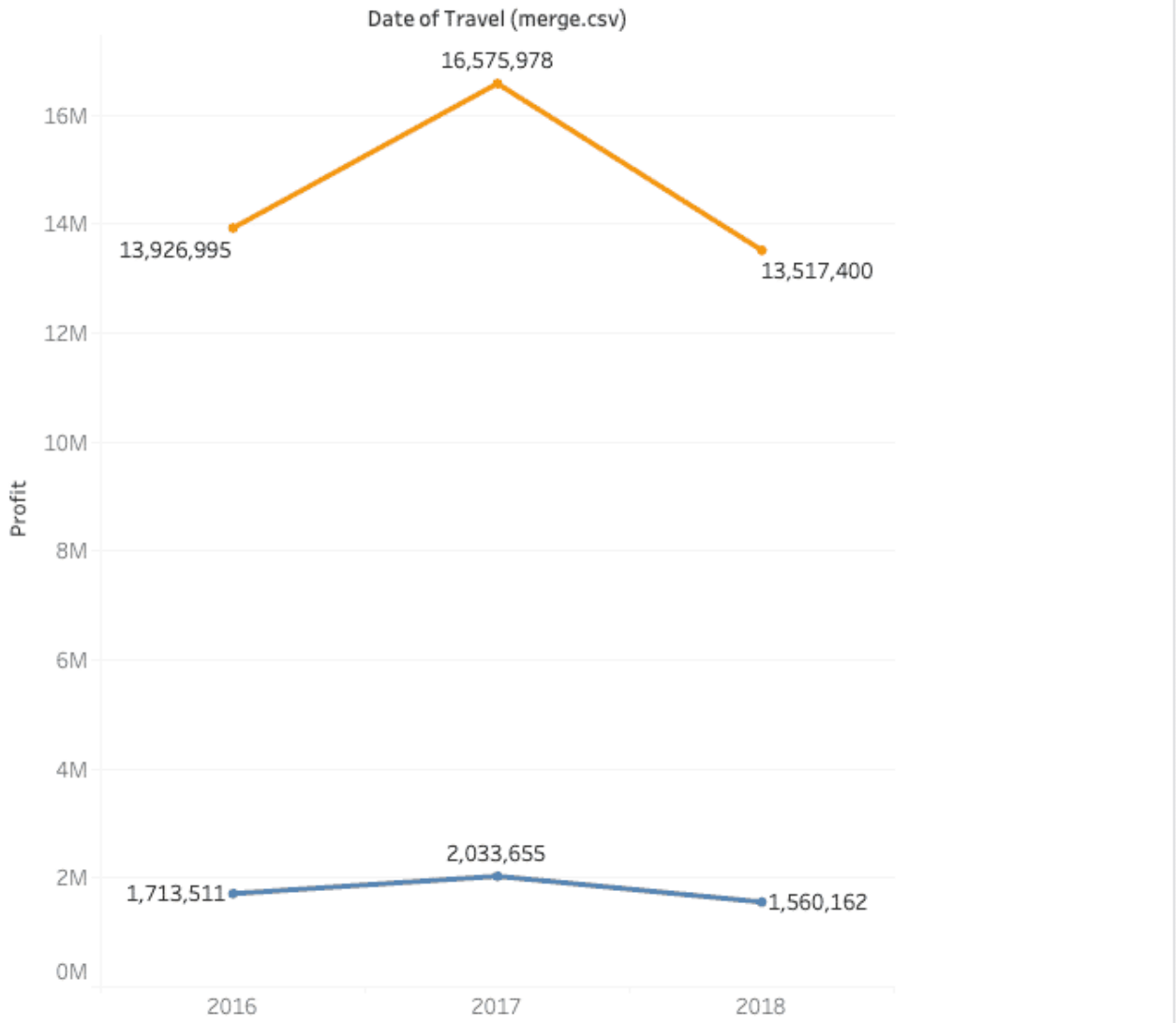




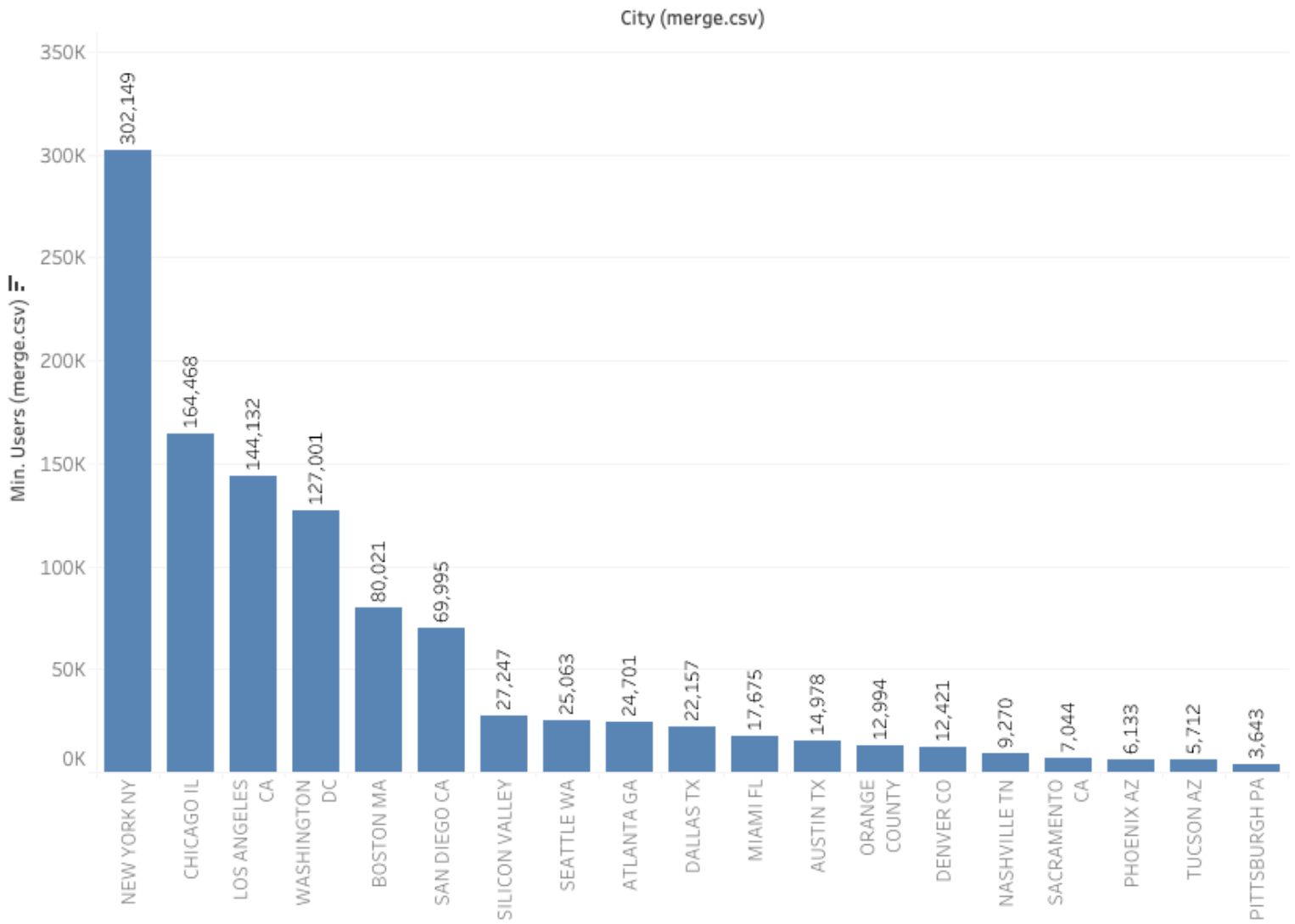
Profit Analysis

Profit comparison in total for each year (2016, 2017, 2018)

Conclusion: Yellow cab makes much higher profits in total than the pink cab.



City wise

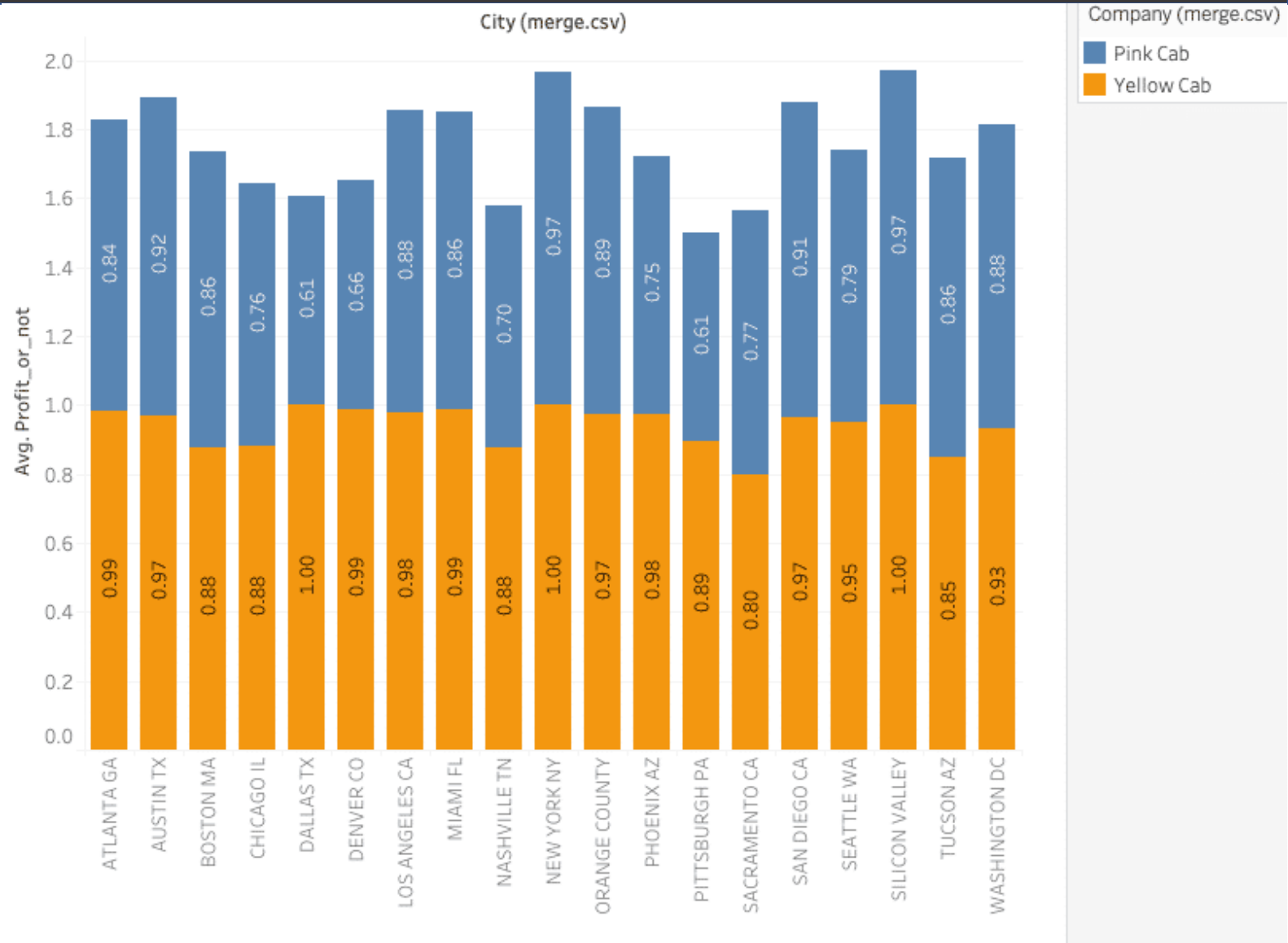


Number of cab users in each city, including yellow and pink cab

City wise

Hypothesis: What are the percentages of rides that can make profit in each city?

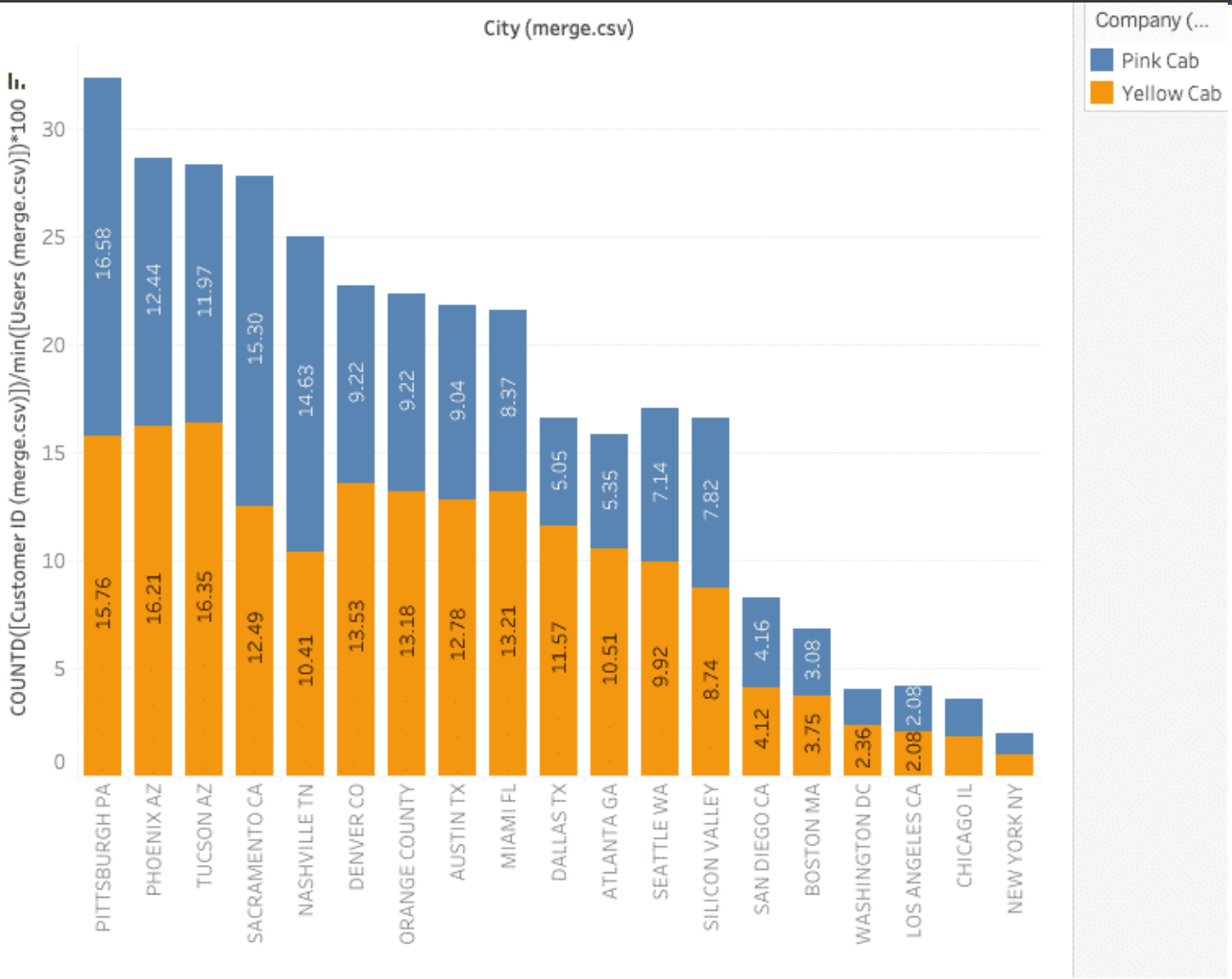
Conclusion: In comparison to the pink cab, yellow cab performs better with relatively higher percentages of profitable rides.



City wise

Hypothesis: Which cab has more users in each city?

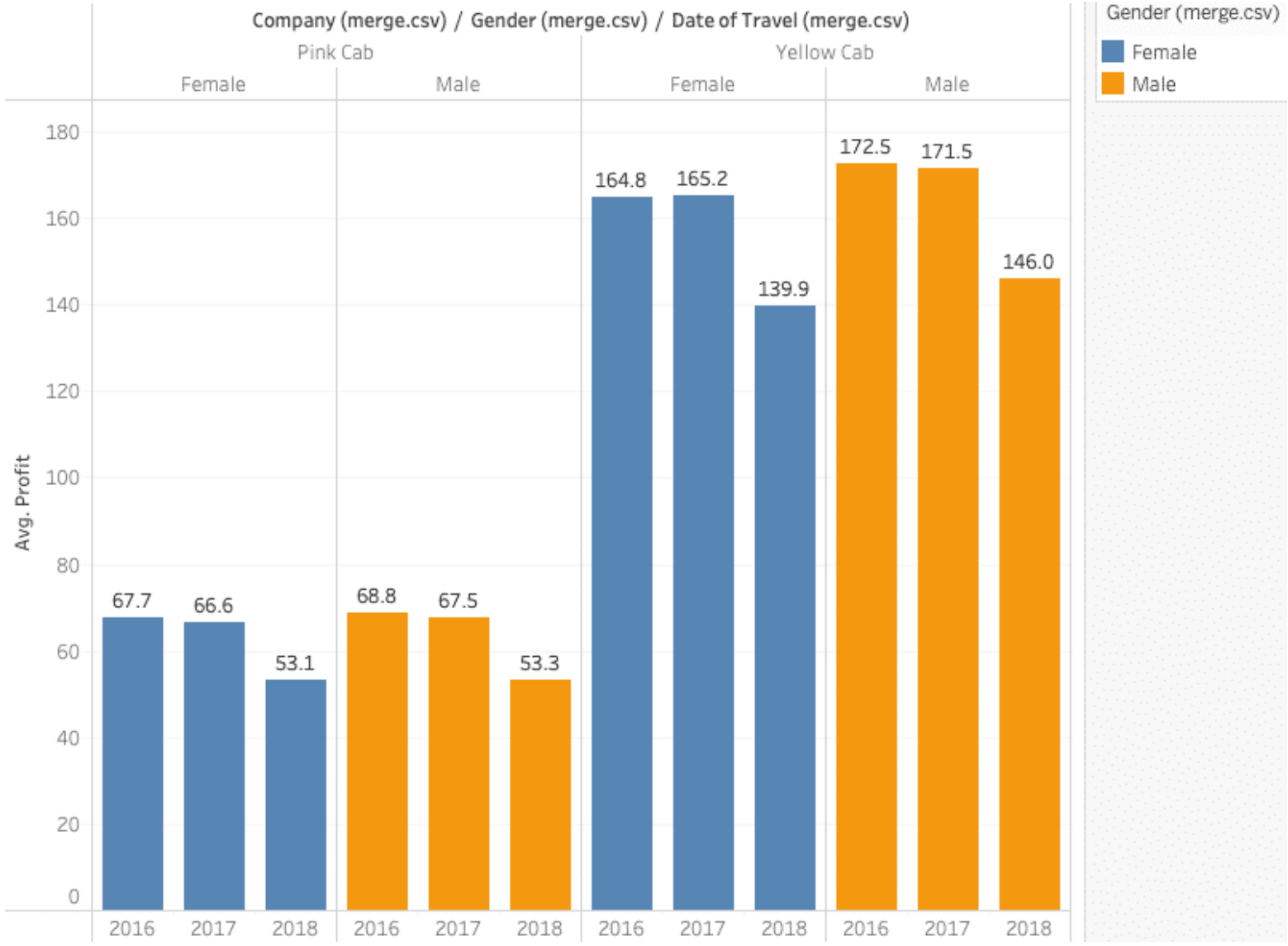
Conclusion: The percentage of yellow cab in each city is bigger than pink one in 15 cities out of 19.



Gender wise

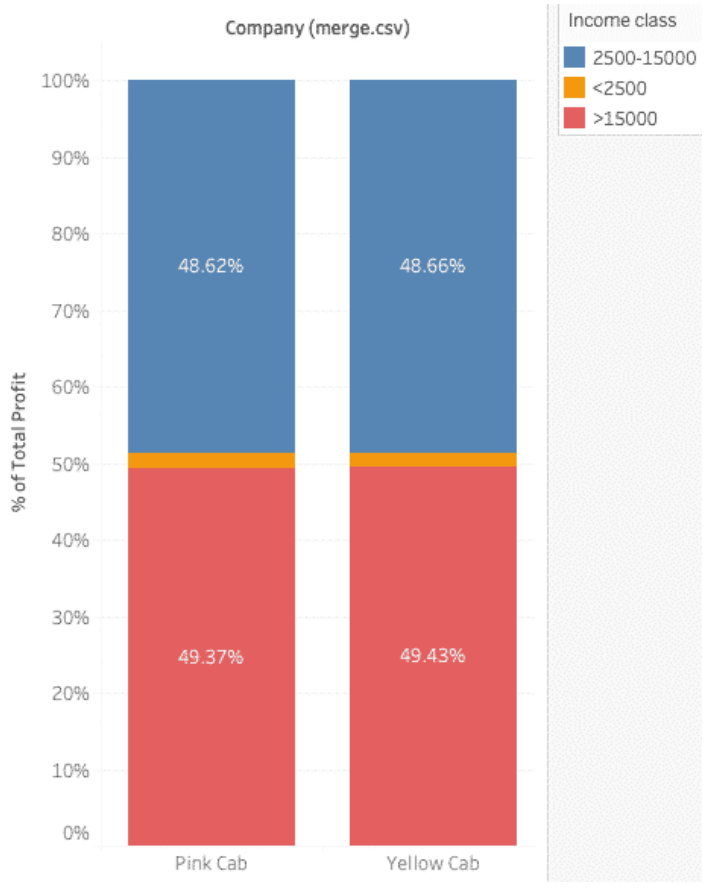
Hypothesis: Does gender effects the profits?

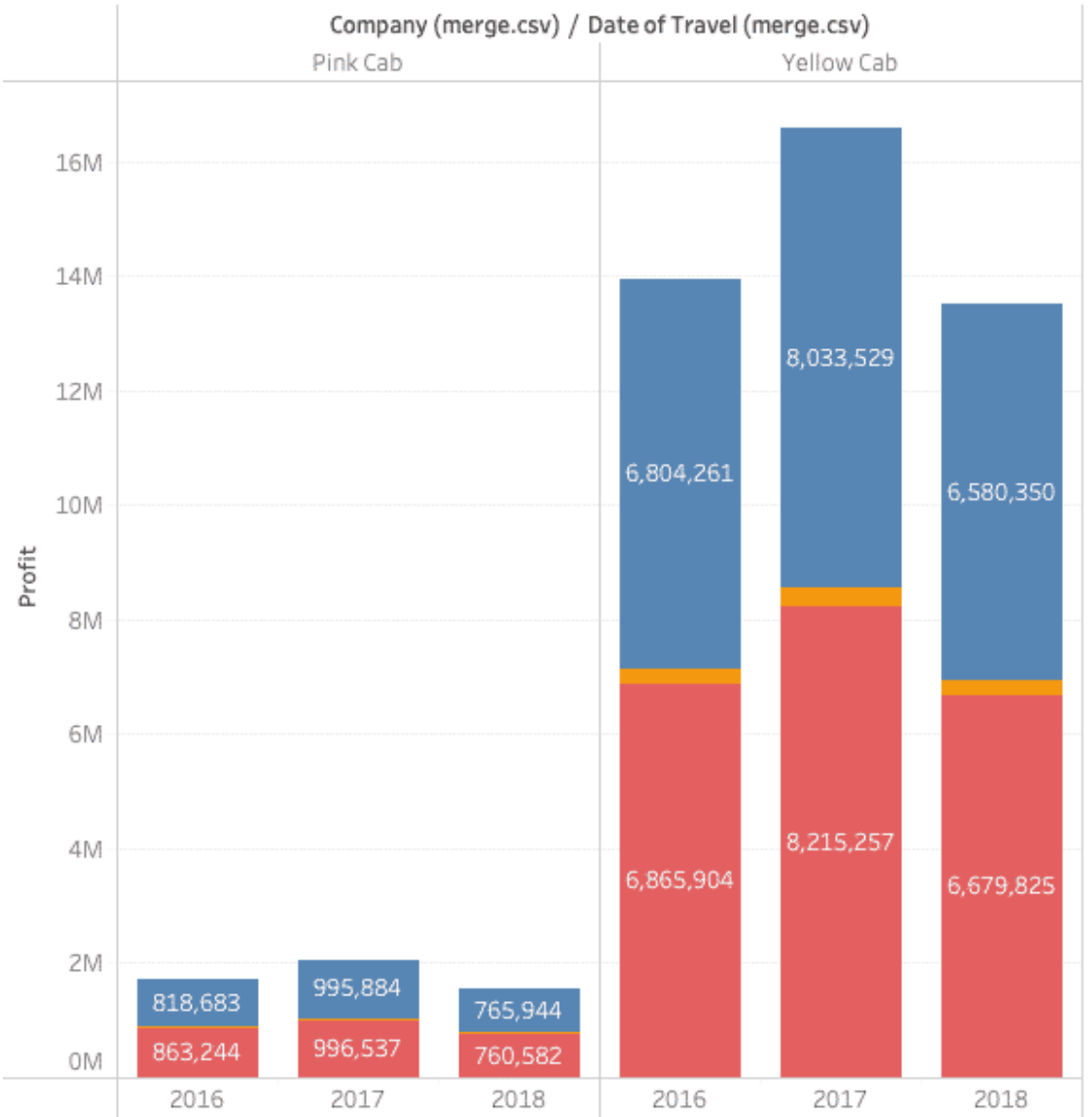
Conclusion: Gender does not make difference and there is not a distinct difference between two cabs.

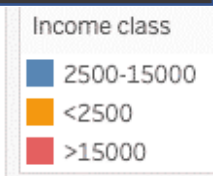


Income wise

Hypothesis: Is there any difference of distributions of customer income groups between two cabs?

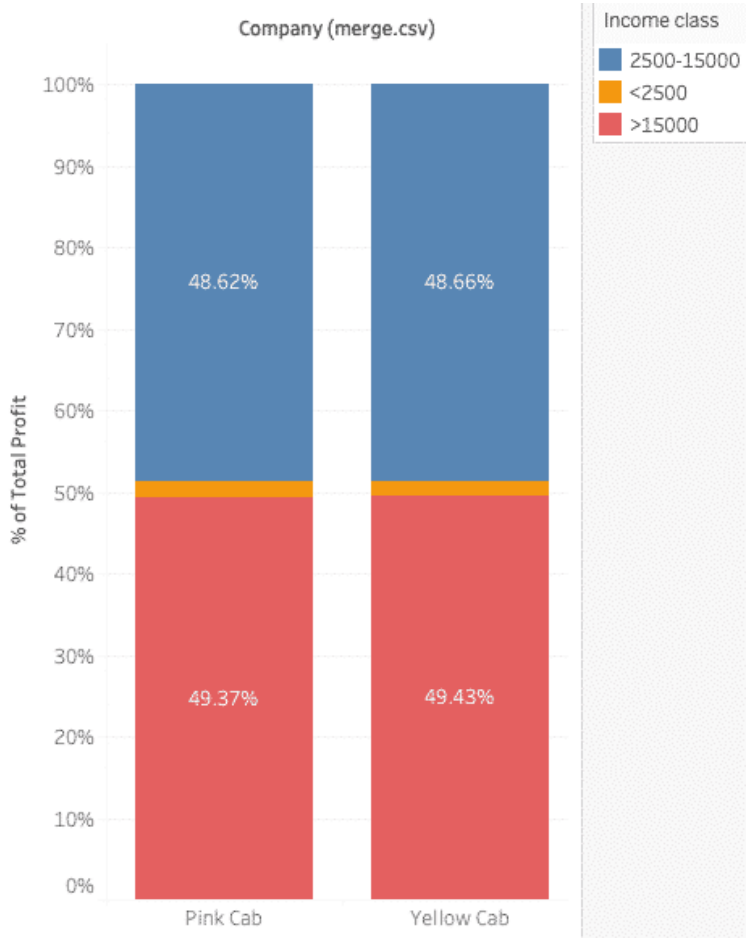


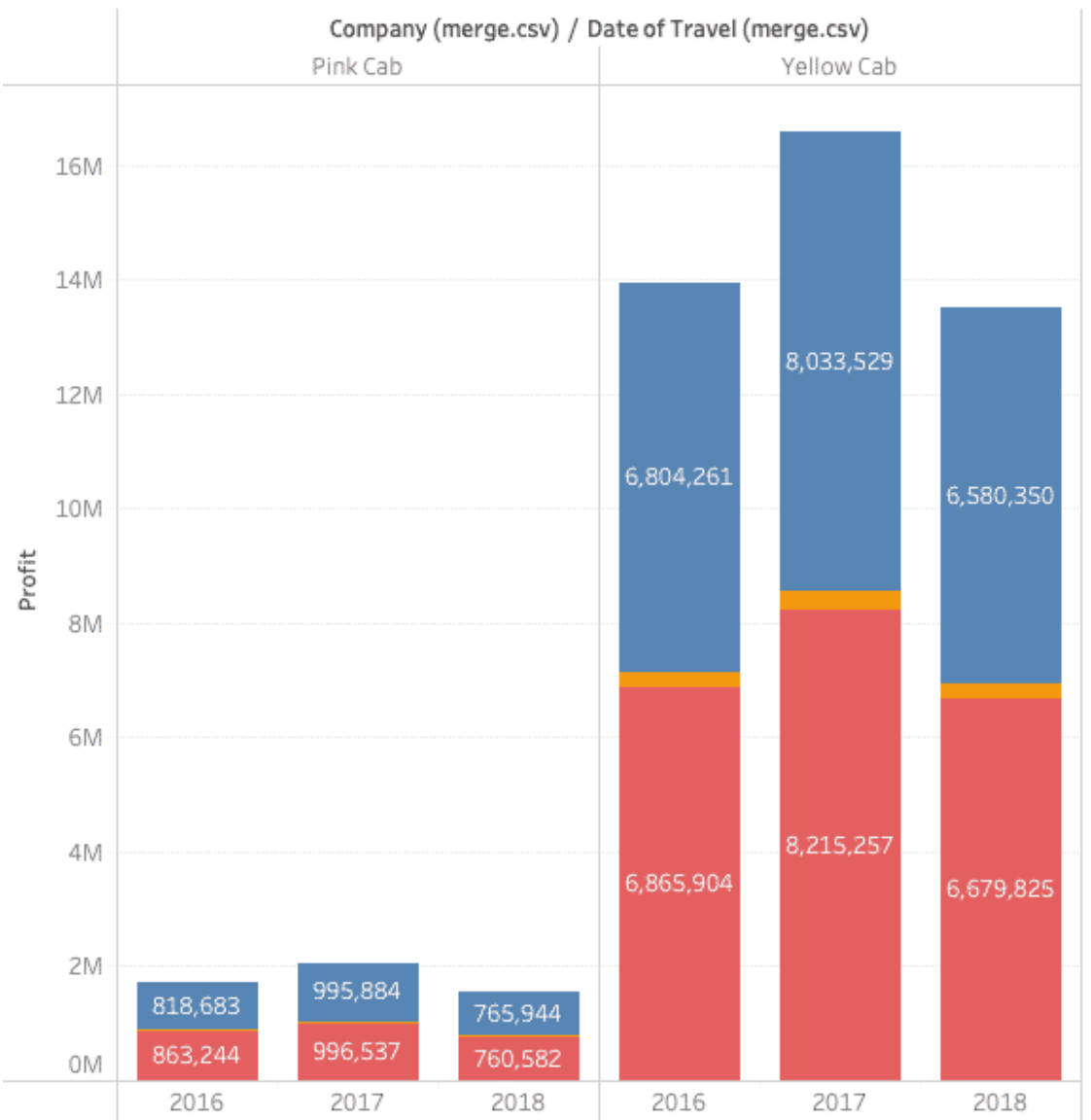


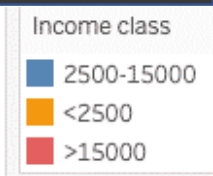


Income wise

- **Conclusion: As shown, they are similar. These three classes provide the same contribution to profits.**

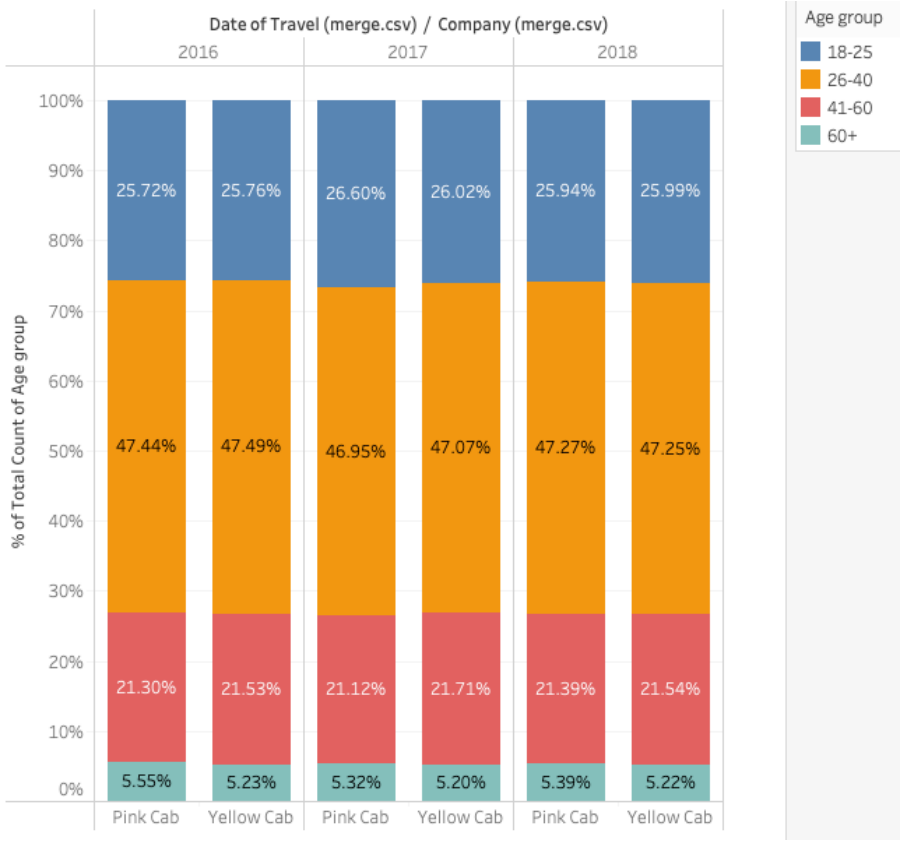


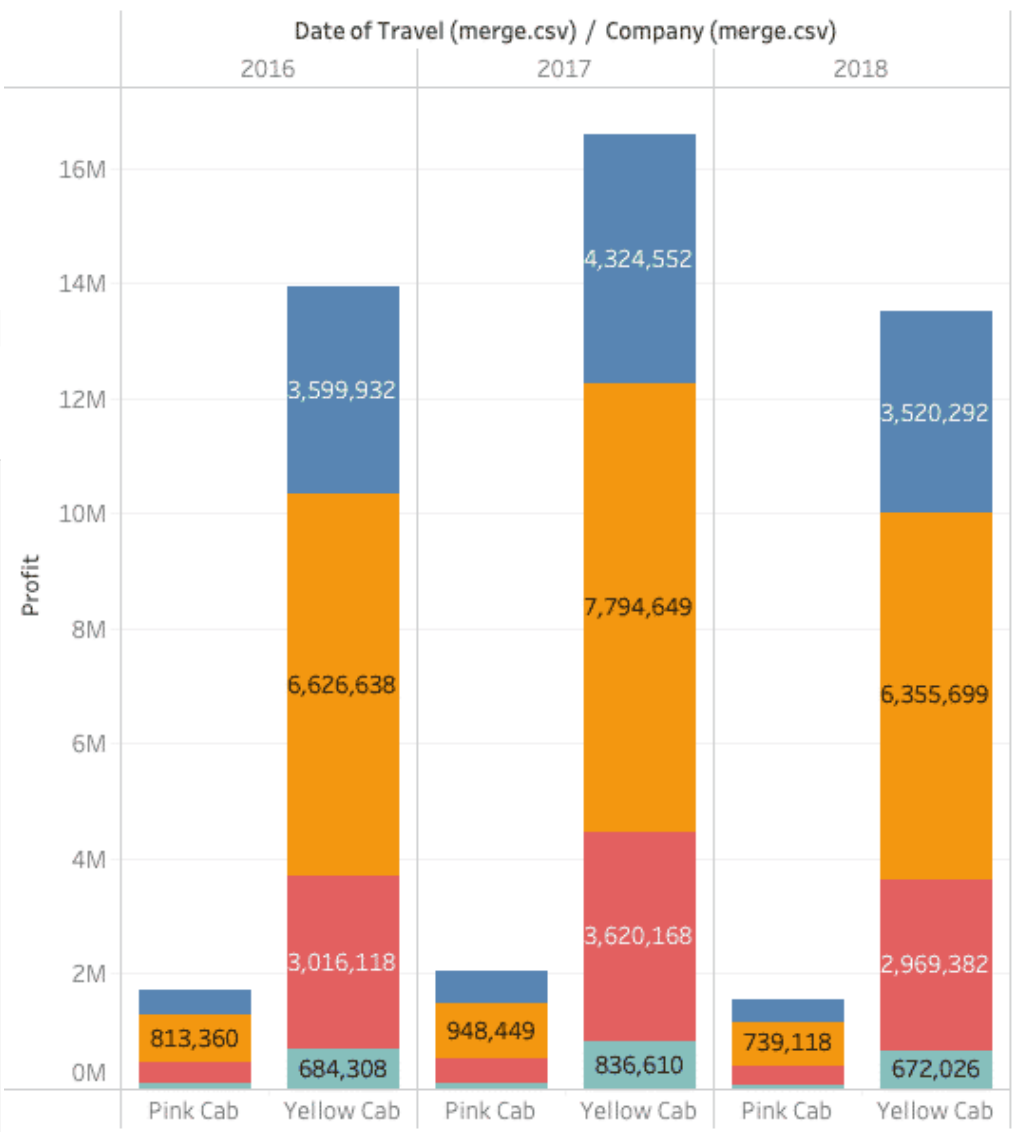




Age wise

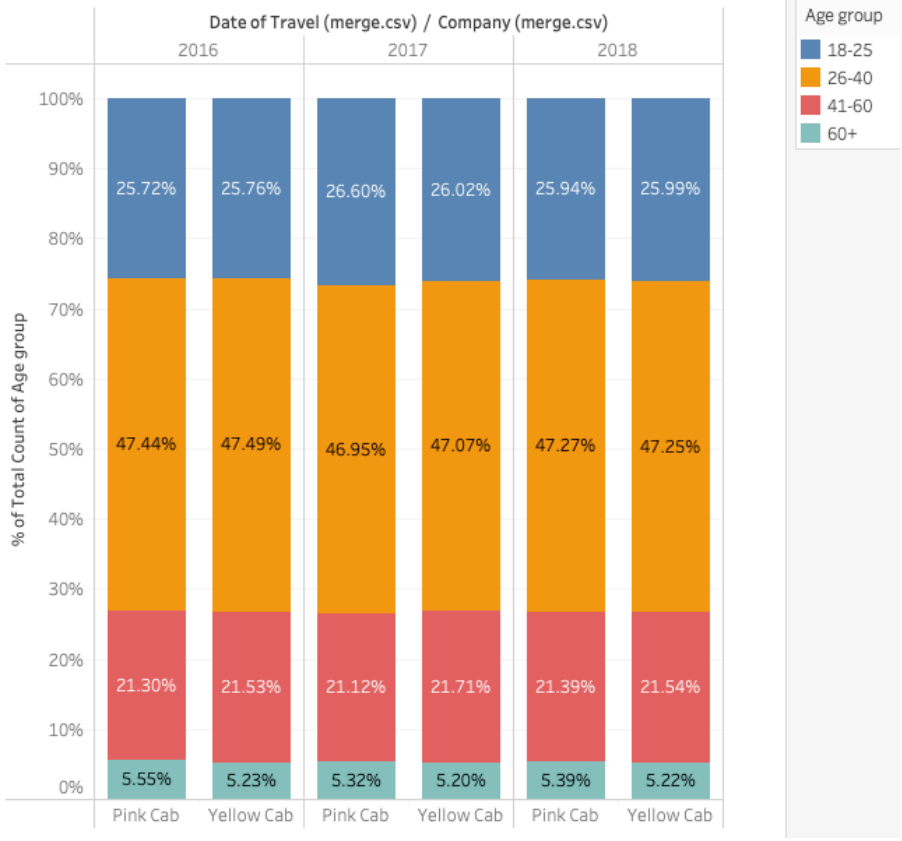
Hypothesis: Is there any difference of distributions of customer age groups between two cabs?

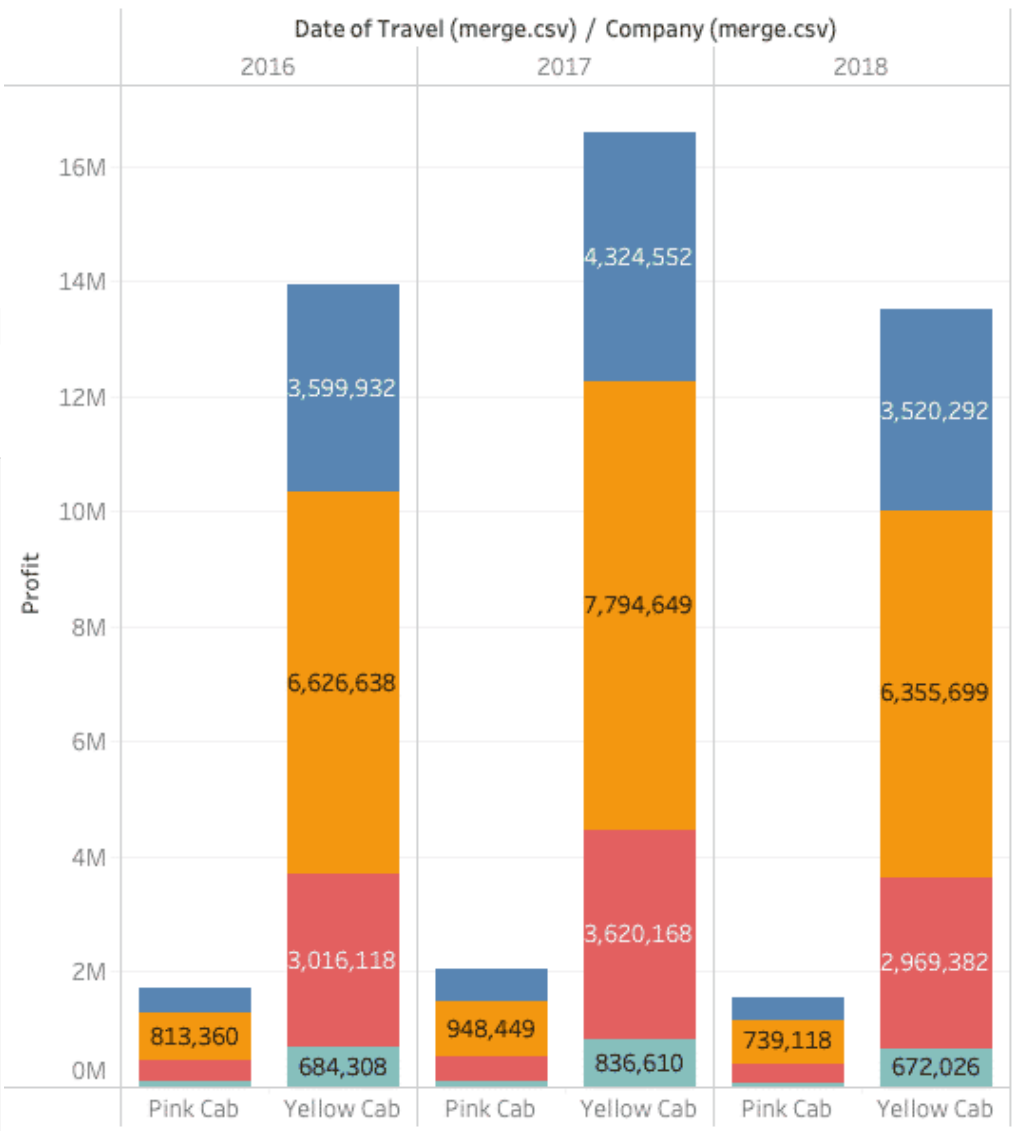




Age wise

Conclusion: There is no big difference. Both of two companies cover all age groups with similar distributions.

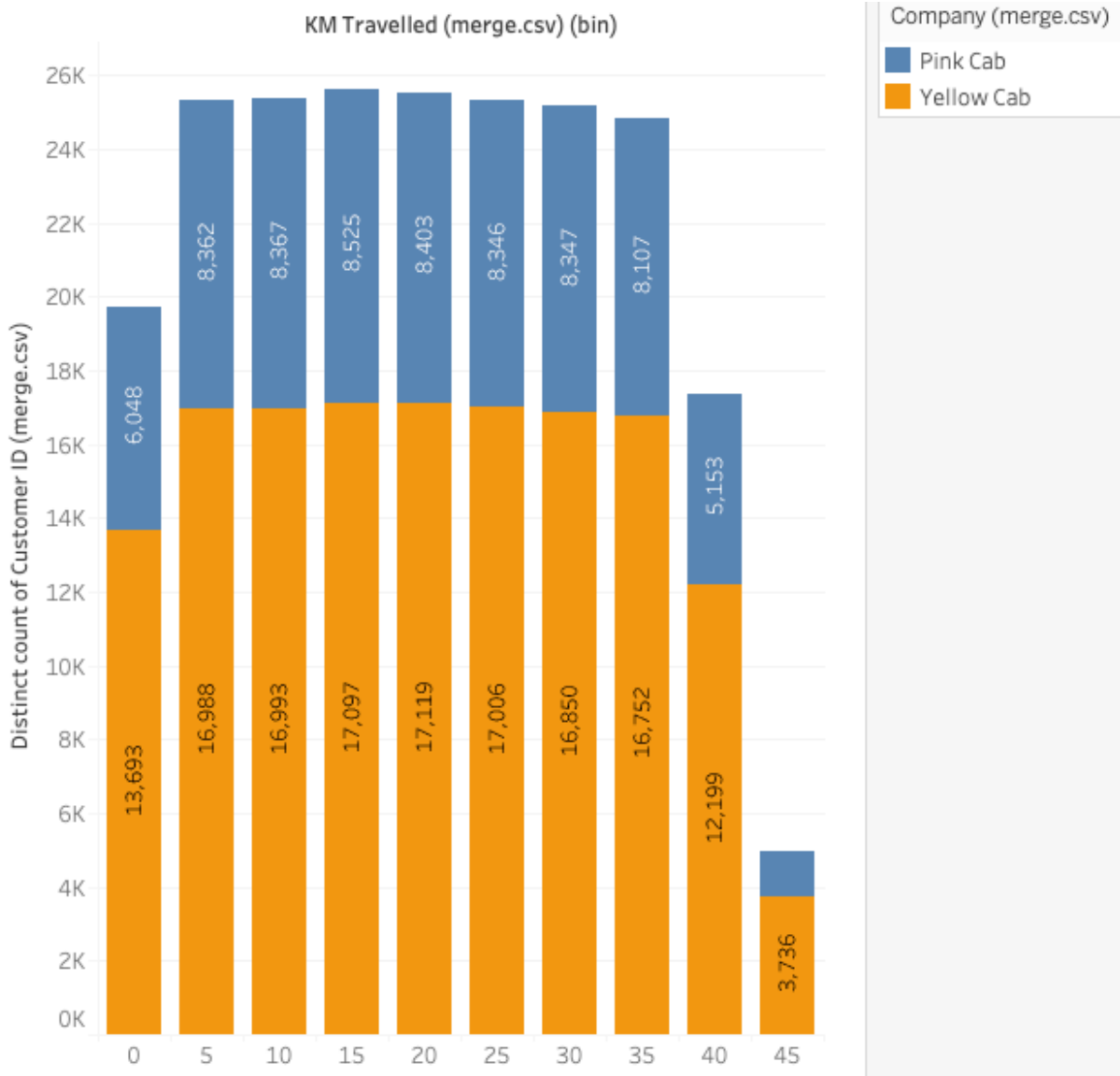




Distance Analysis

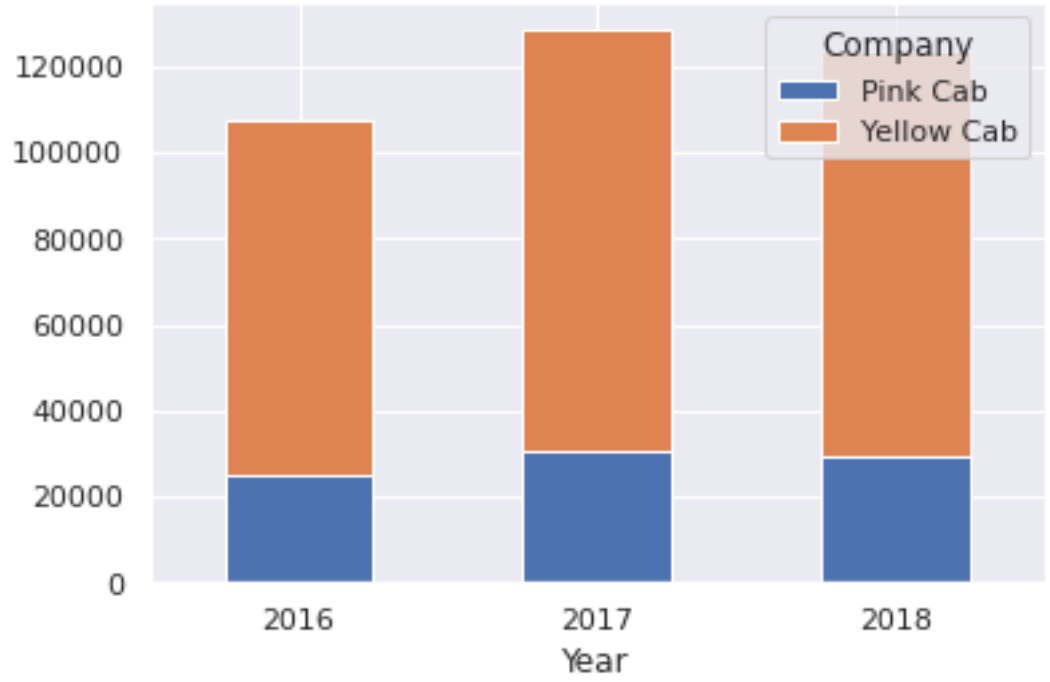
Hypothesis: Is there any preference of cab services among various distances?

Conclusion: Yellow cab has customers almost uniform for all kinds of Trip, while Pink cab has relatively long trip customers, which means that its service to long trips is not so attractive.

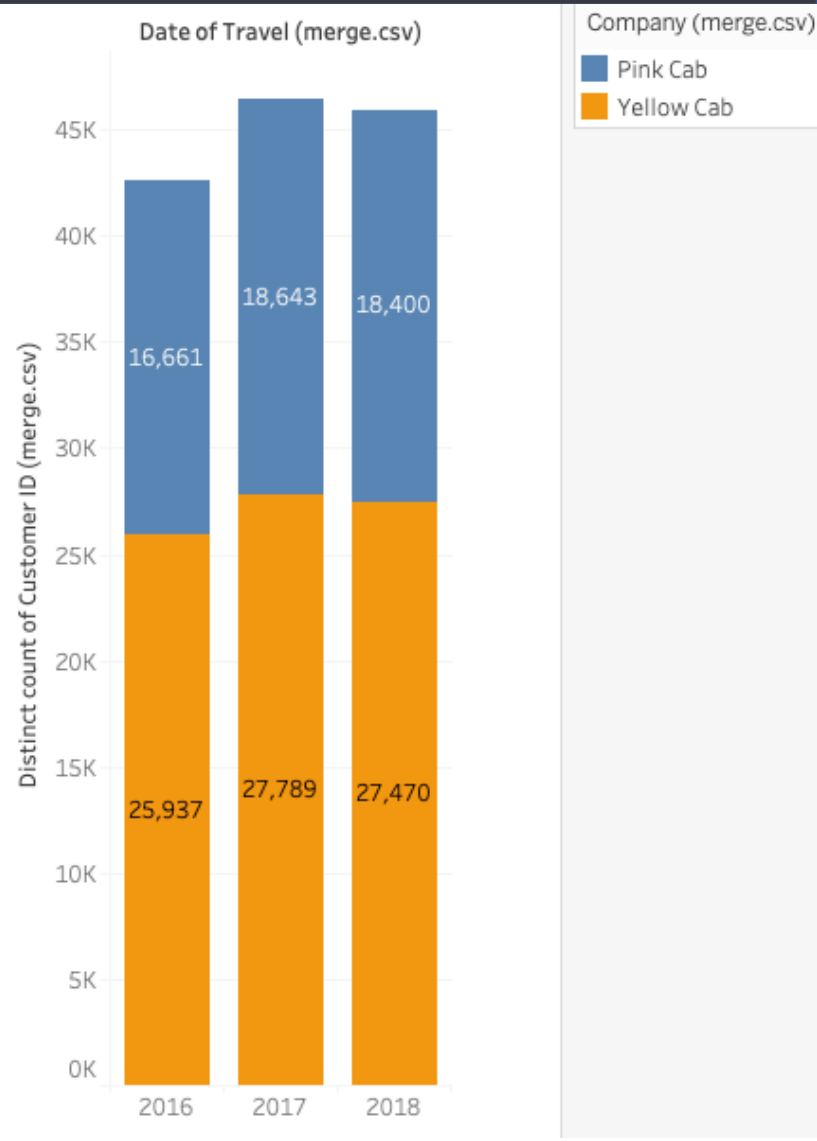


Time series Analysis

Number of transactions based on years



Number of customers based on years



Time series Analysis

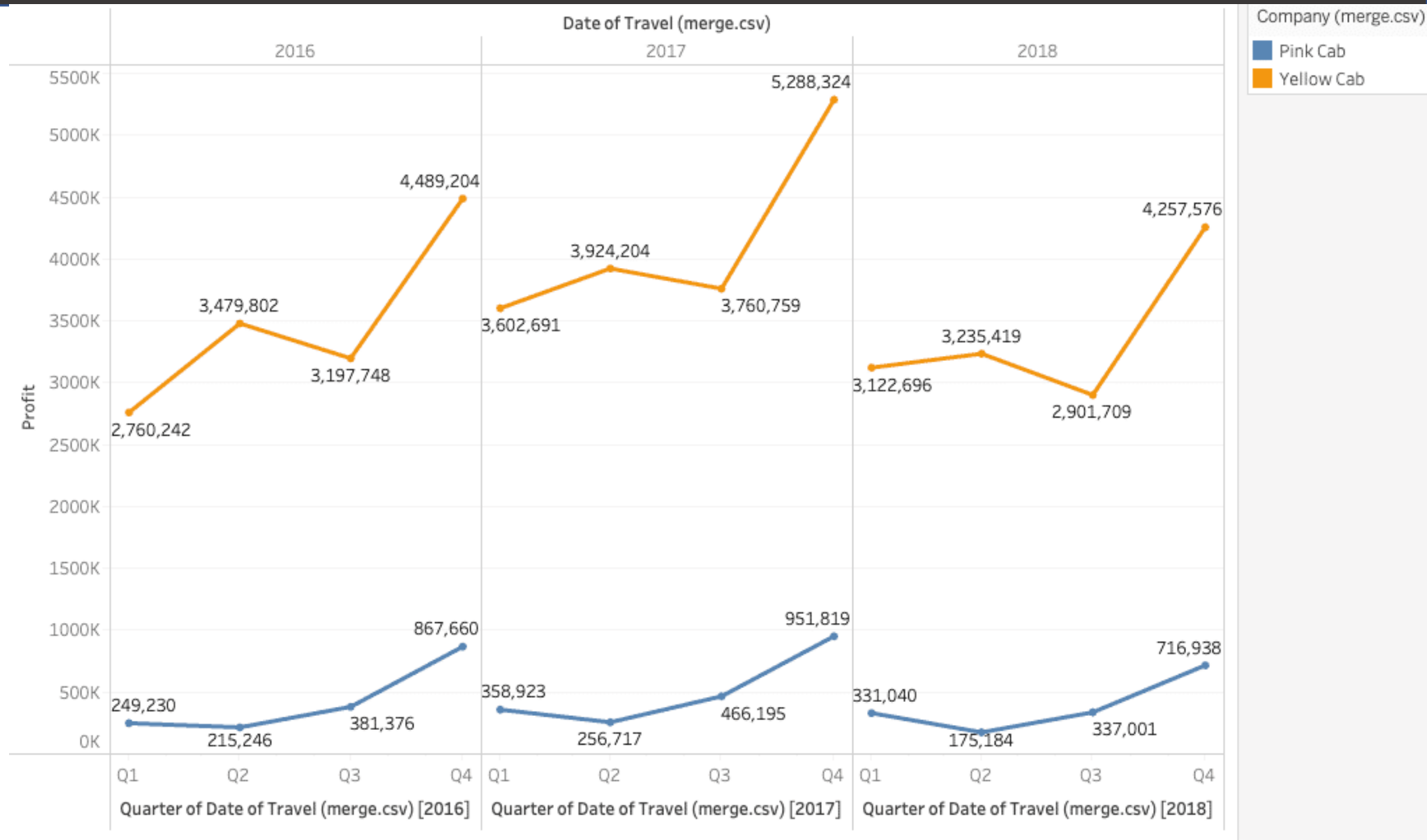
Hypothesis: Is there any seasonality in number of cab service and profits in Quarter scale?

Conclusion: Seasonality is shown in both cabs. Number of rides rises from Q1 to Q4 and decrease as Q1 comes, which means 4 quarters a cycle.

Time series Analysis

Hypothesis: Is there any seasonality in number of cab service and profits in Quarter scale?

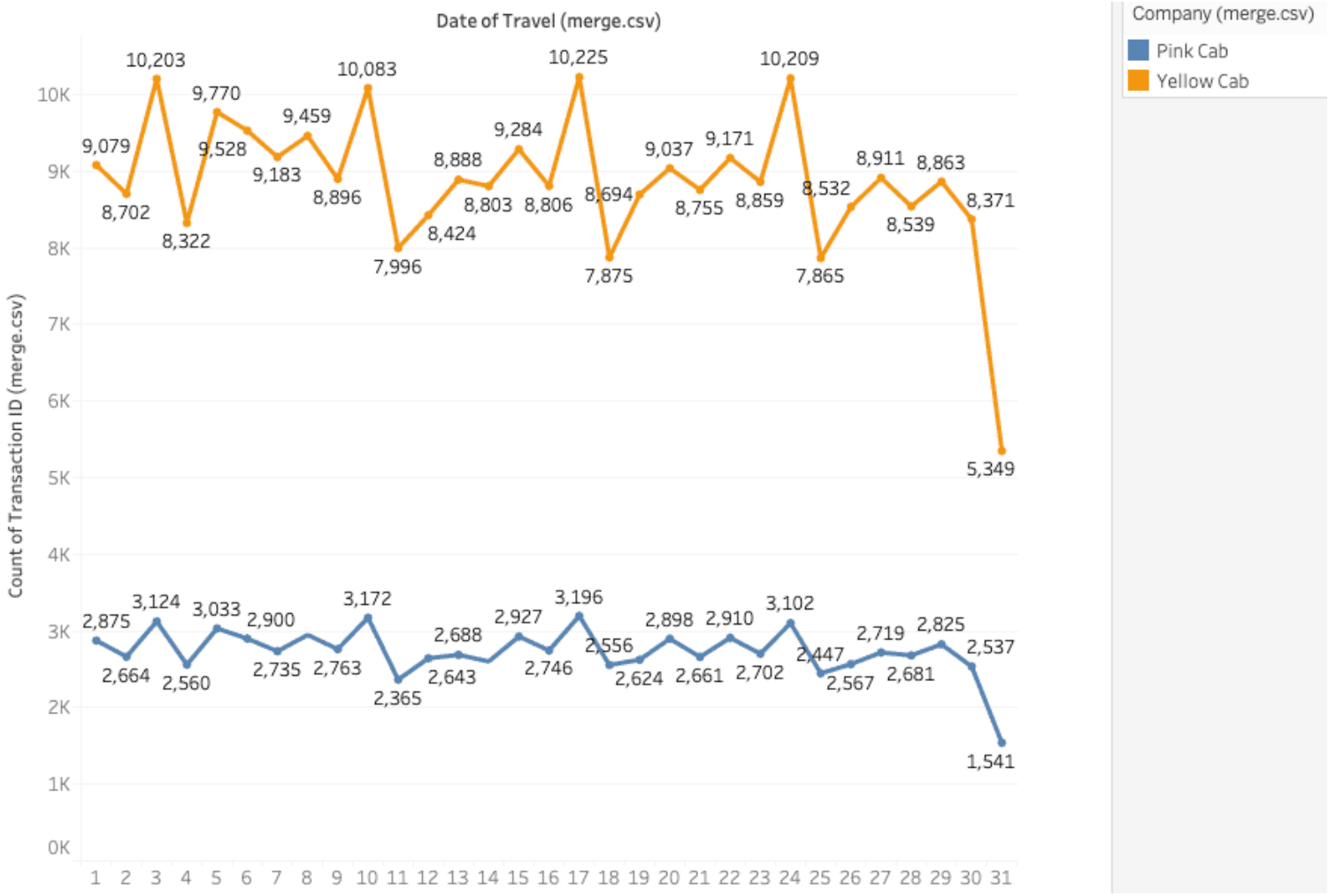
Conclusion: Similar quarter patterns can be found in terms of Profit.



Time series Analysis

Hypothesis: Is there any seasonality in number of cab service in month scale?

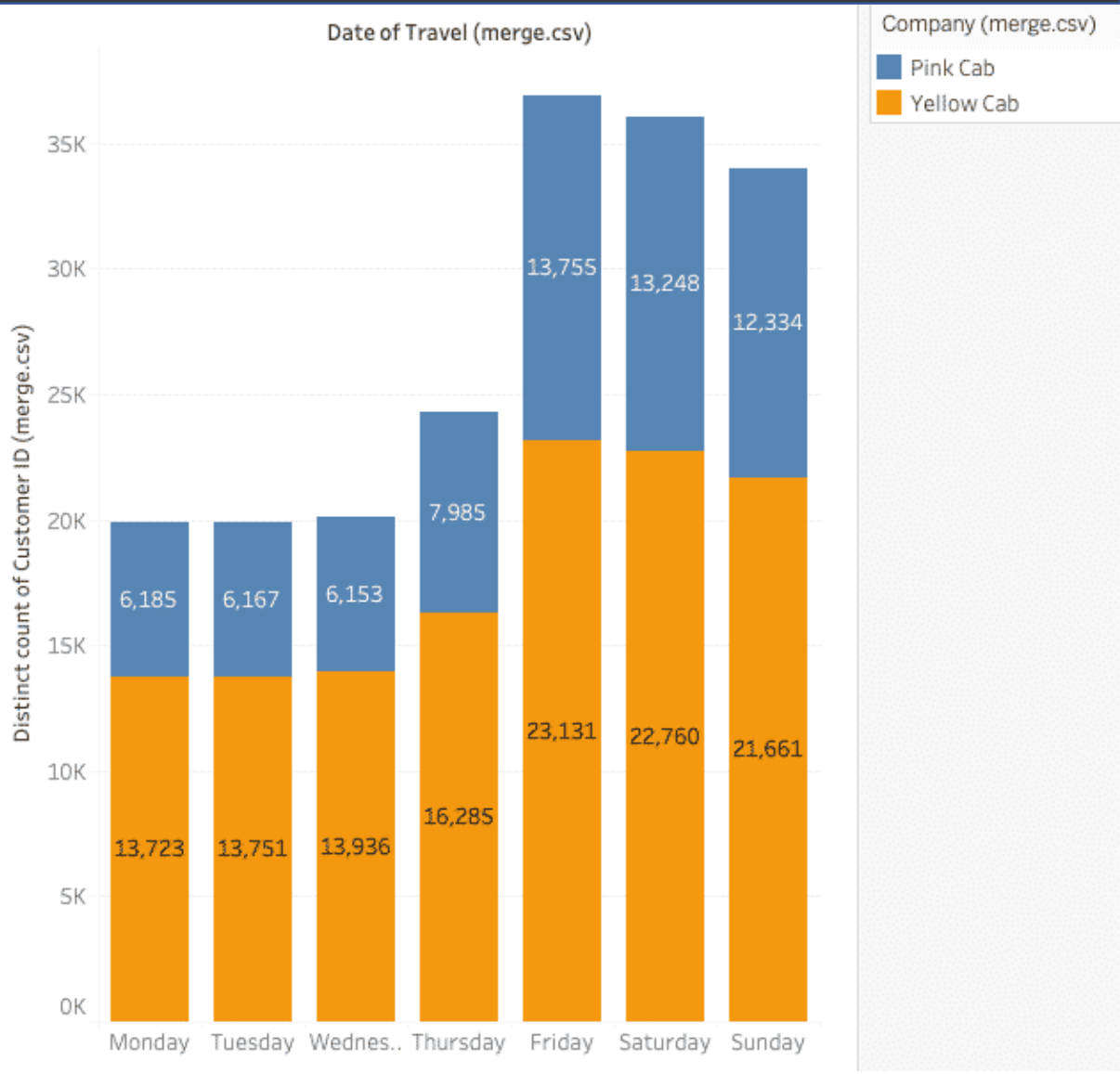
Conclusion: An interesting founding: Number of rides is big on 3rd day of the month and it repeats similar pattern every 7 days.



Time series Analysis

Hypothesis: Is there any seasonality in number of customers and rides in day-of-week scale?

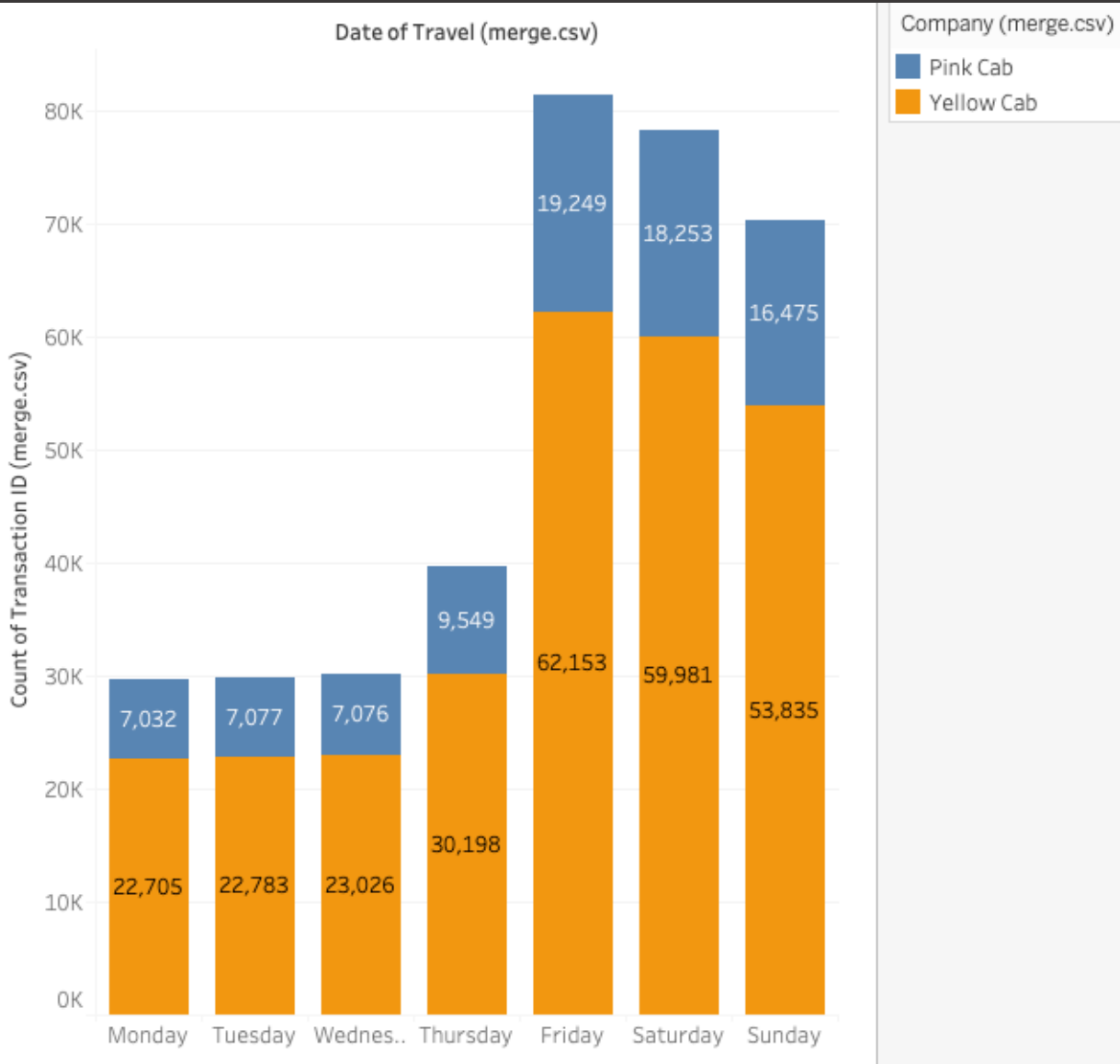
Conclusion: on Friday, Saturday, and Sunday, demands and profits are usually much higher than other days of week.



Time series Analysis

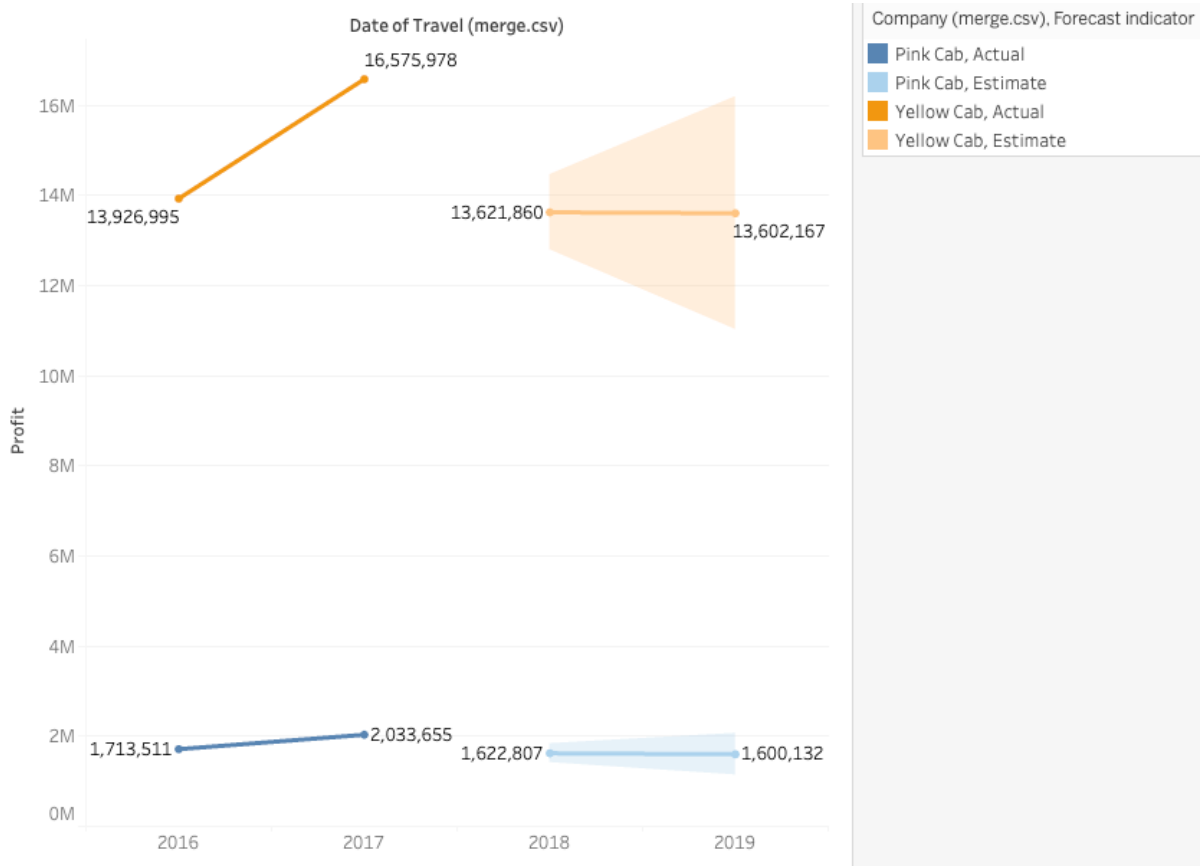
Hypothesis: Is there any seasonality in number of customers and rides in day-of-week scale?

Conclusion: on Friday, Saturday, and Sunday, demands and profits are usually much higher than other days of week.

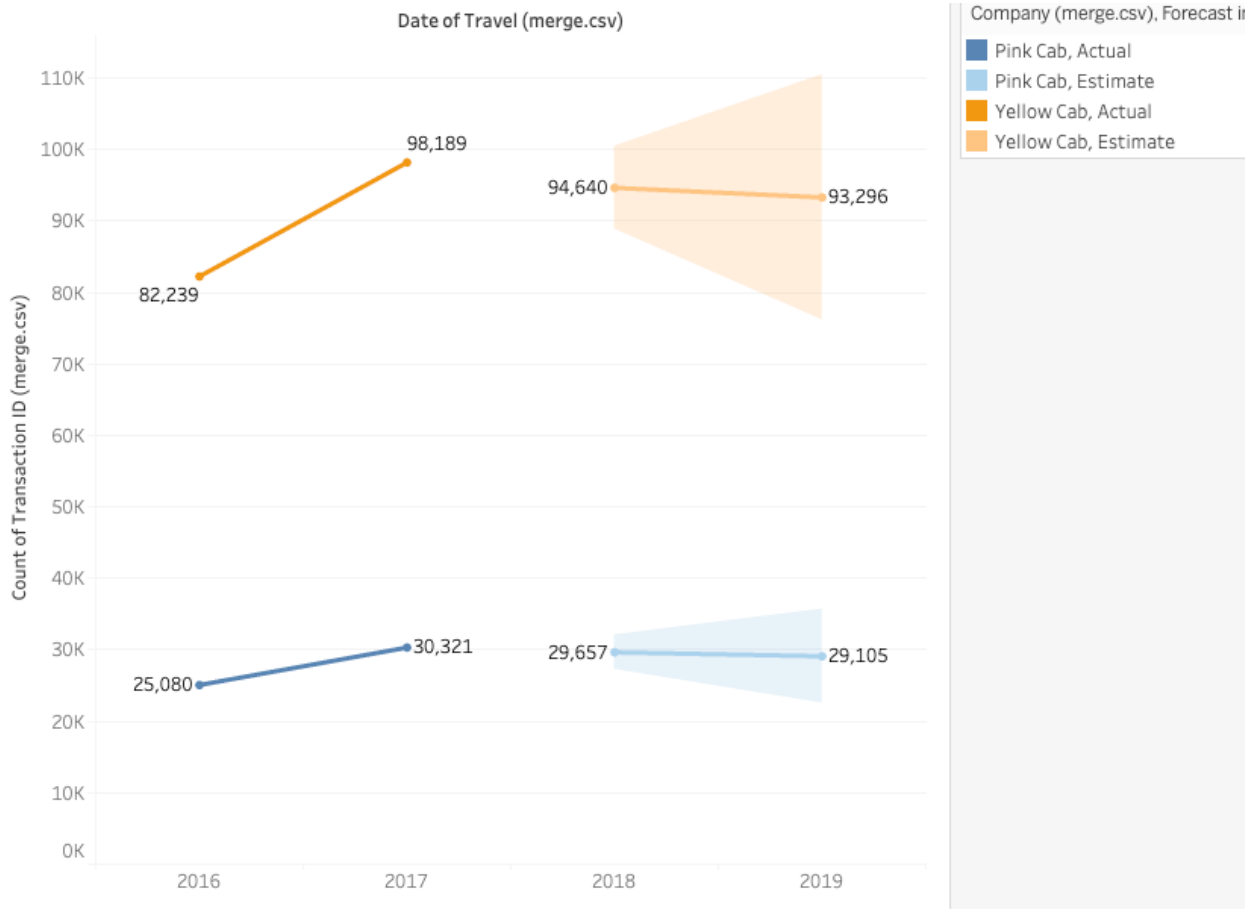


Forecast

Profit forecast



Number of Ride forecast



EDA Summary

1. Yellow cab has a much more competitive ability to make a profit.

2. Yellow cab covers more users in each city, which means higher potential.

3. Yellow cab has good services and covers all kinds of genders, incomes, and age groups.

4. There are obvious time patterns that we can dive deep to improve its profits.

5. Compared with pink cabs, yellow provides better services on long trips and thus more chosen by customers.

Recommendations

As a result, **yellow cab** performs better than pink one in many ways and should be chosen for investment.

Thank You