Exploratory Data Analysis

G2M insight for Cab Investment firm

Gideon Osei Bonsu

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Executive Summary

- The two cab companies have a lot of potential for growth.
- The companies have a much better presence in Washington DC, San Francisco, and Boston, serving over 30% of the population in these cities.
- The Yellow Cab company has much better profits.
- The Pink Cab company has had a better average profit trend since 2018.
- The Yellow Cab company has better customer retention

Data Glacier
Your Deep Learning Partner

Problem Statement

- 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, it is planning for an investment in Cab industry and as per their Go-to-Market(G2M) strategy they want to understand the market before taking final decision.

- Having been provided with multiple datasets that contain information on 2 cab companies from 31/01/2016 to 31/12/2018, XYZ is interested in using actionable insights derived from my analysis to help them identify the right company to make their investment.

Approach Taken

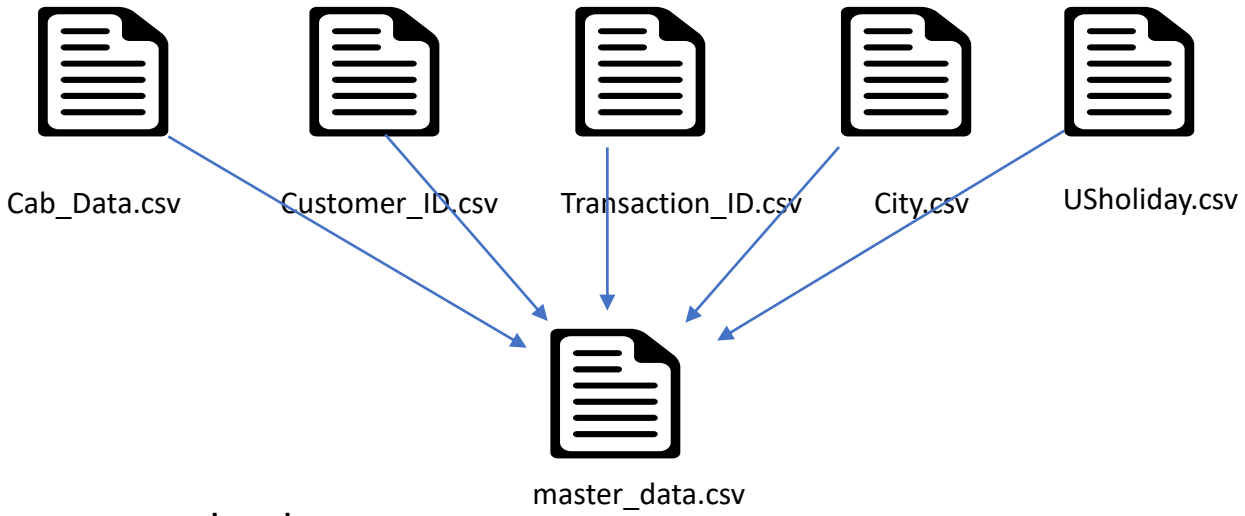
I followed the following steps to analyze the data:

- Ask: At this stage, I asked pressing questions regarding the purpose of the analysis and what the customer is hoping to get from the analysis.
- Prepare: this is where I imported the required libraries and read datasets.
- Process: This is where I cleaned the data, checked for null values, blank cells, cleaned column names, and grouped datasets to create a master dataset.
- Analyze: At this stage, I analyzed the data to draw insights from it.
- Share: This stage ran through “Analyze” stage as I shared visualizations of various sections of analysis of the data.
- Act: Here, I made recommendations.

Exploratory Data Analysis

Master Data contains:

- 18 Features (including 4 derived features)
- Timeframe of the data: 2016-01-31 to 01/29/2019
- Total data points: 359,392

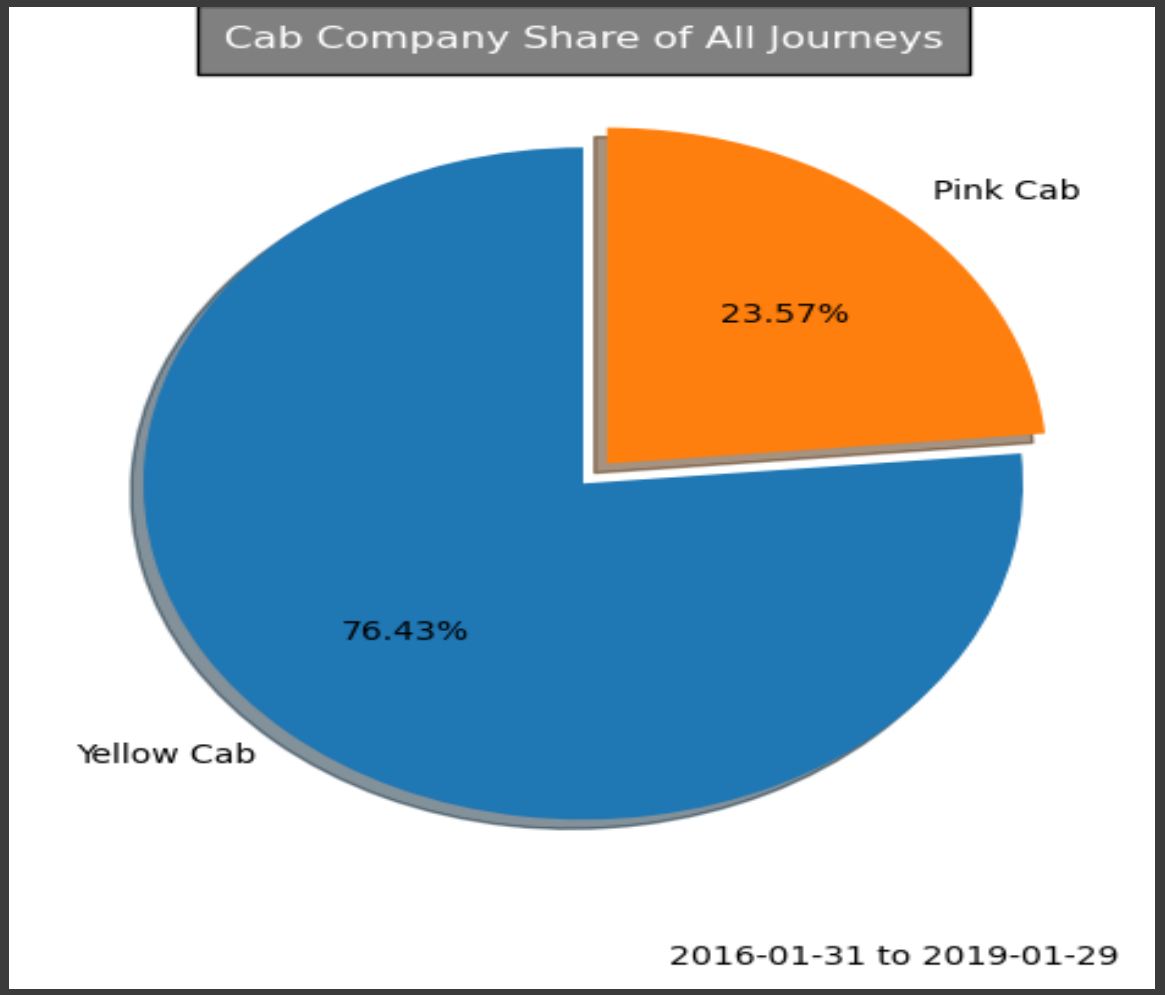


Assumptions:

- Profit of rides are calculated keeping other factors constant and only Price_Charged and Cost_of_Trip features used to calculate profit.

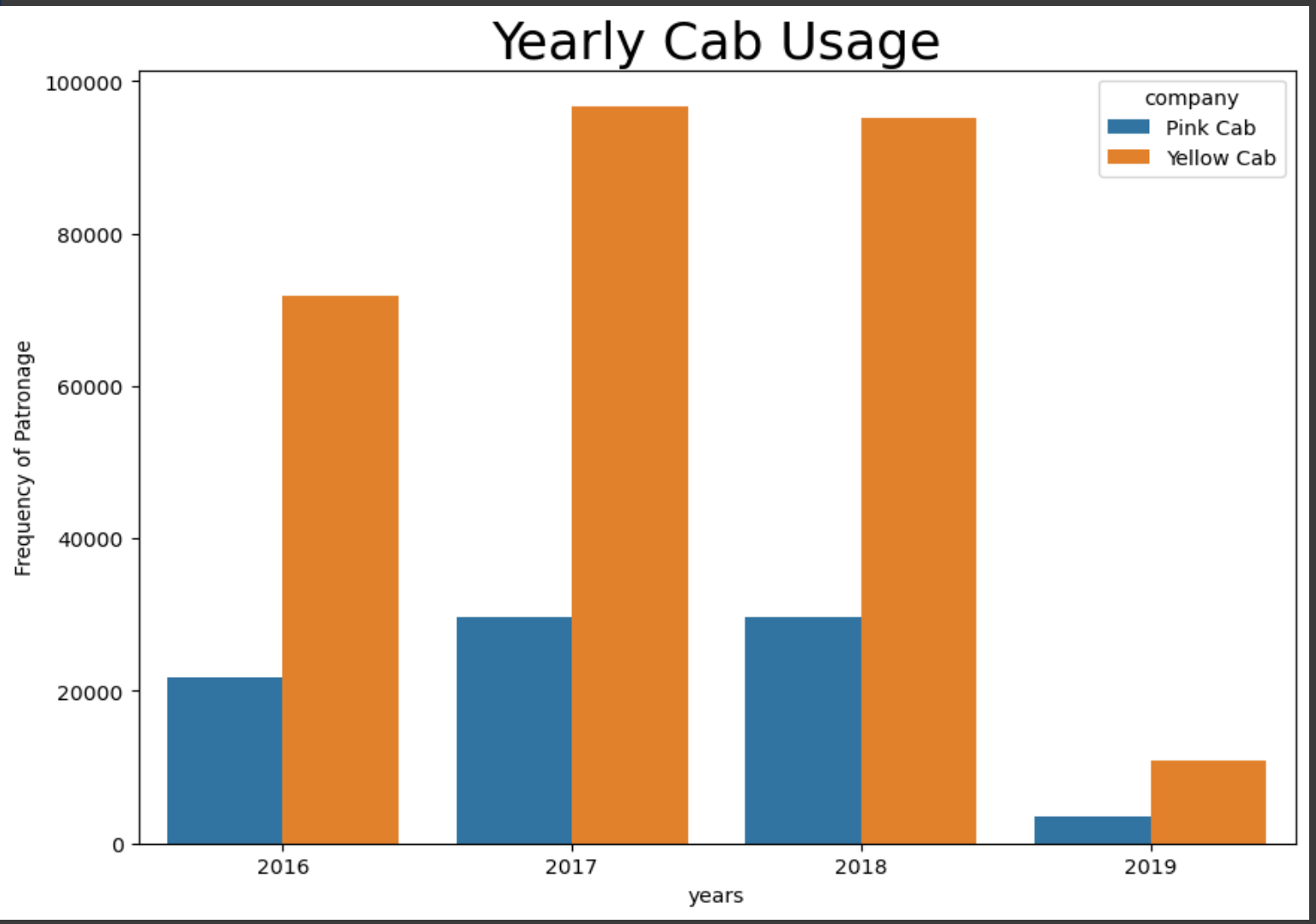
- The end date was adjusted to be 01/29/2019, instead of the stated 2018-12-31 because the numerical date values given in the dataset do not add up to 2018-12-31 Being the end date when converting to Date format (More explanation is given in Jupyter notebook)

Customer Preference Analysis



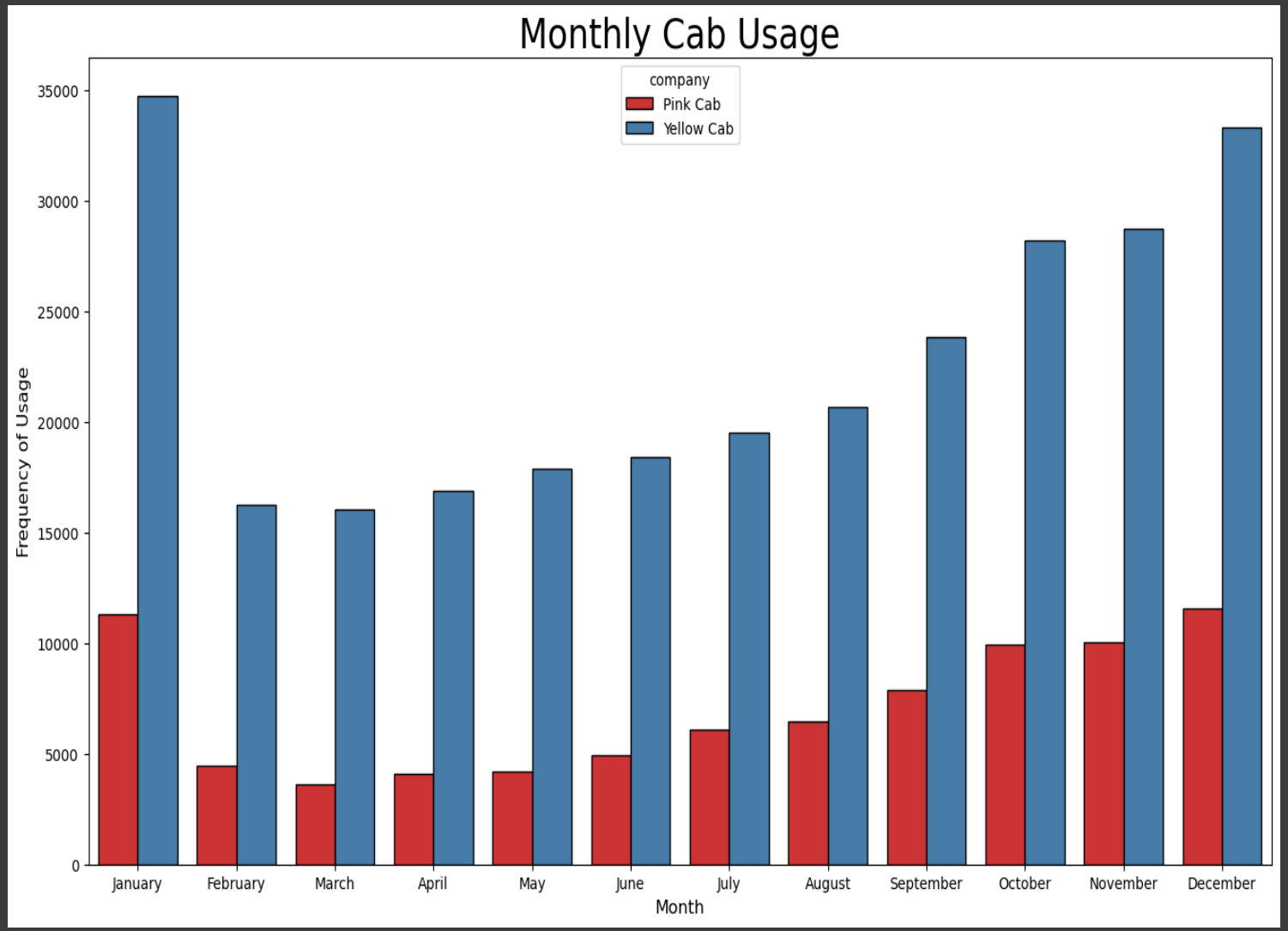
- The Yellow Cab is used a lot more by customers than the Pink Cab.
- The Yellow Cab had 274,681 out of the total 359,392 transactions in the dataset. This dominance translates to 76.43% of all cab journeys, outshining the Yellow Cab's 84,711 journeys, which make up a mere 23.57%.

Seasonality Analysis



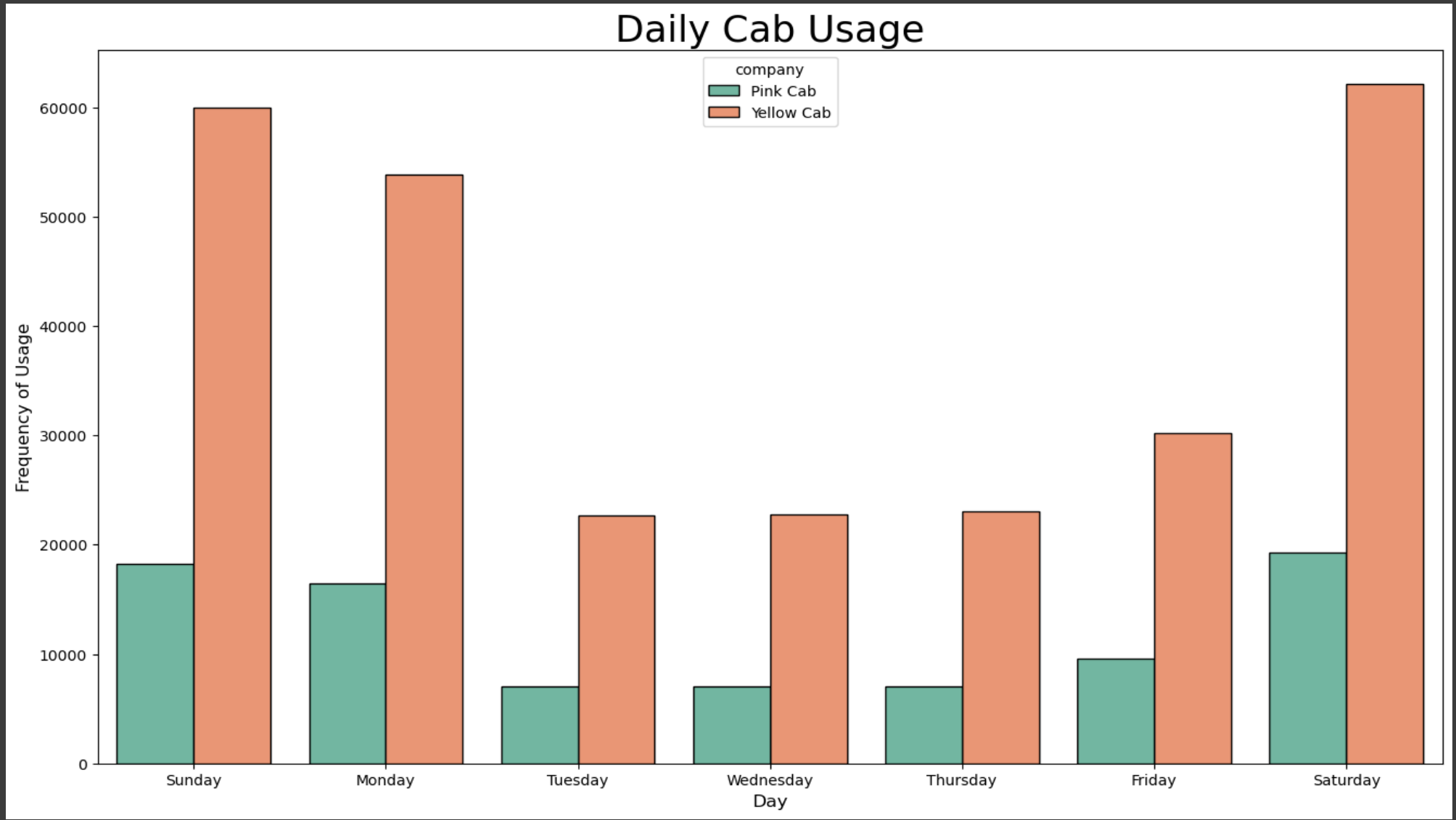
- Despite the Yellow Cab having a lot more transactions, its number of transactions saw a 1.49% decrease from 96,626 in 2017 to 95186 in 2018.
- The Pink Cab only had a 0.07% decrease from 29,750 in 2017 to 29,730 in 2018.

Seasonality Analysis

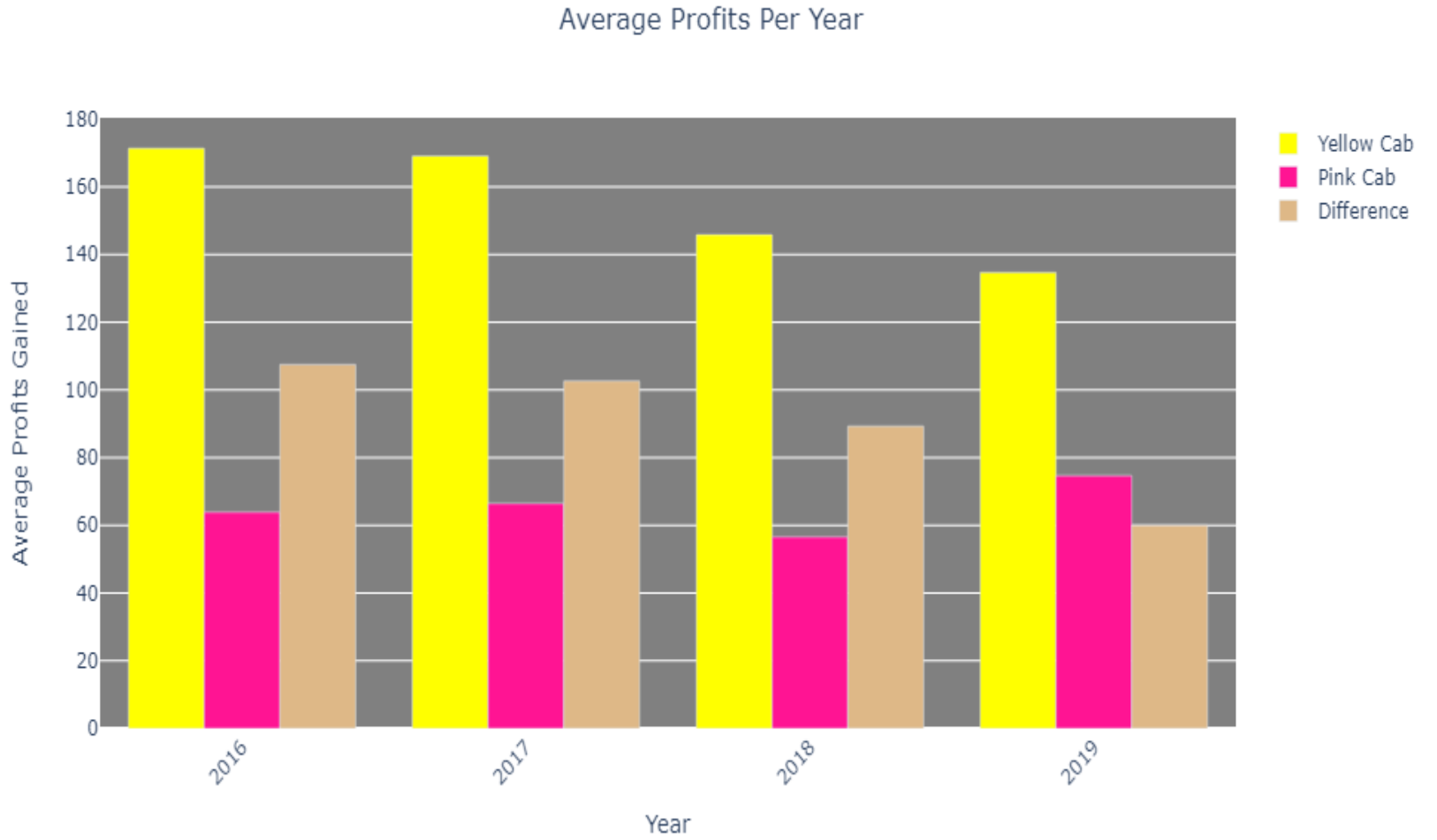


- There is considerable seasonality in Cab Usage.
- Cab usage rises almost steadily from February to January of the next year.
- Cabs are busiest in January and December, which are generally holiday months.

Seasonality Analysis

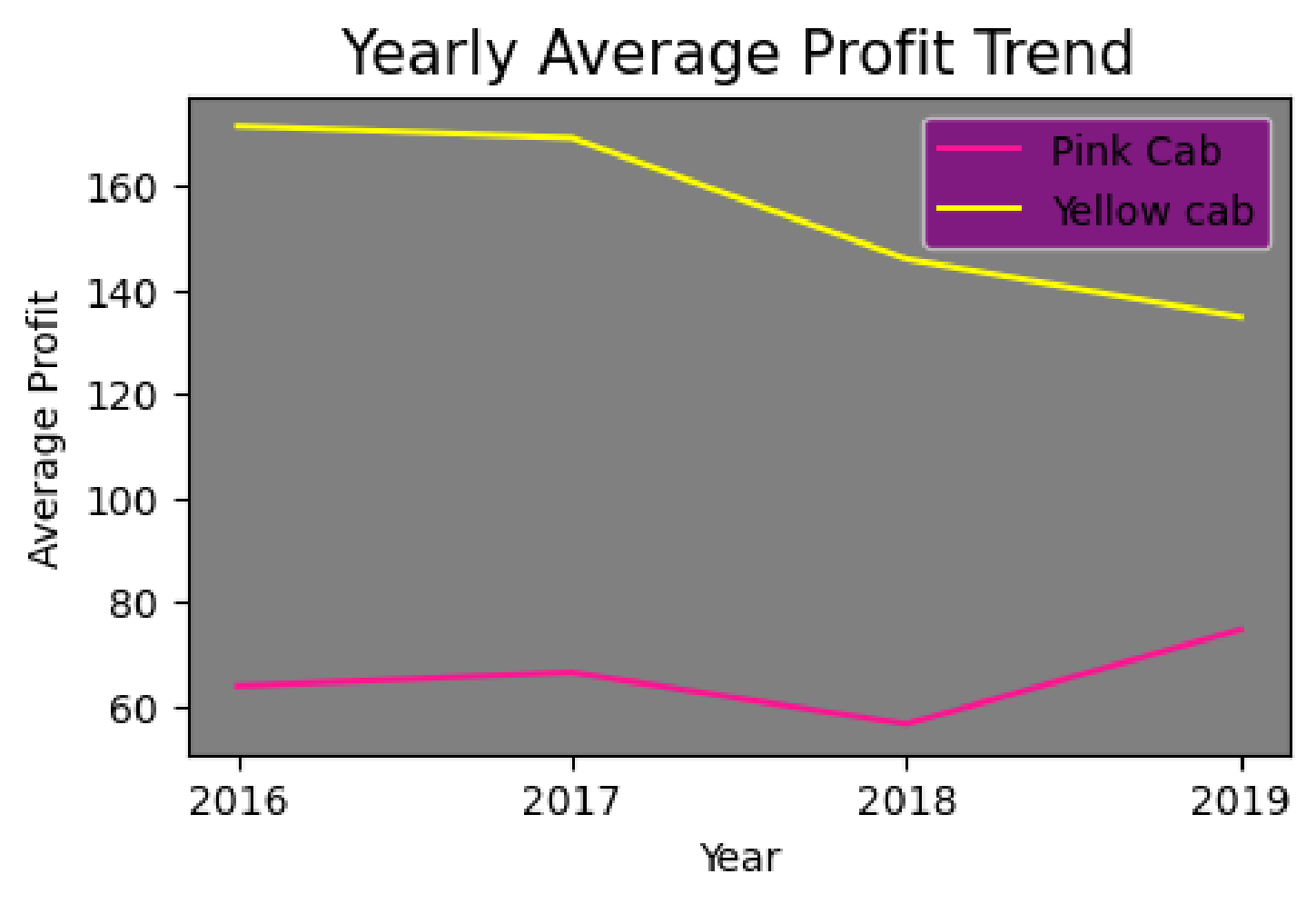


Profit Analysis



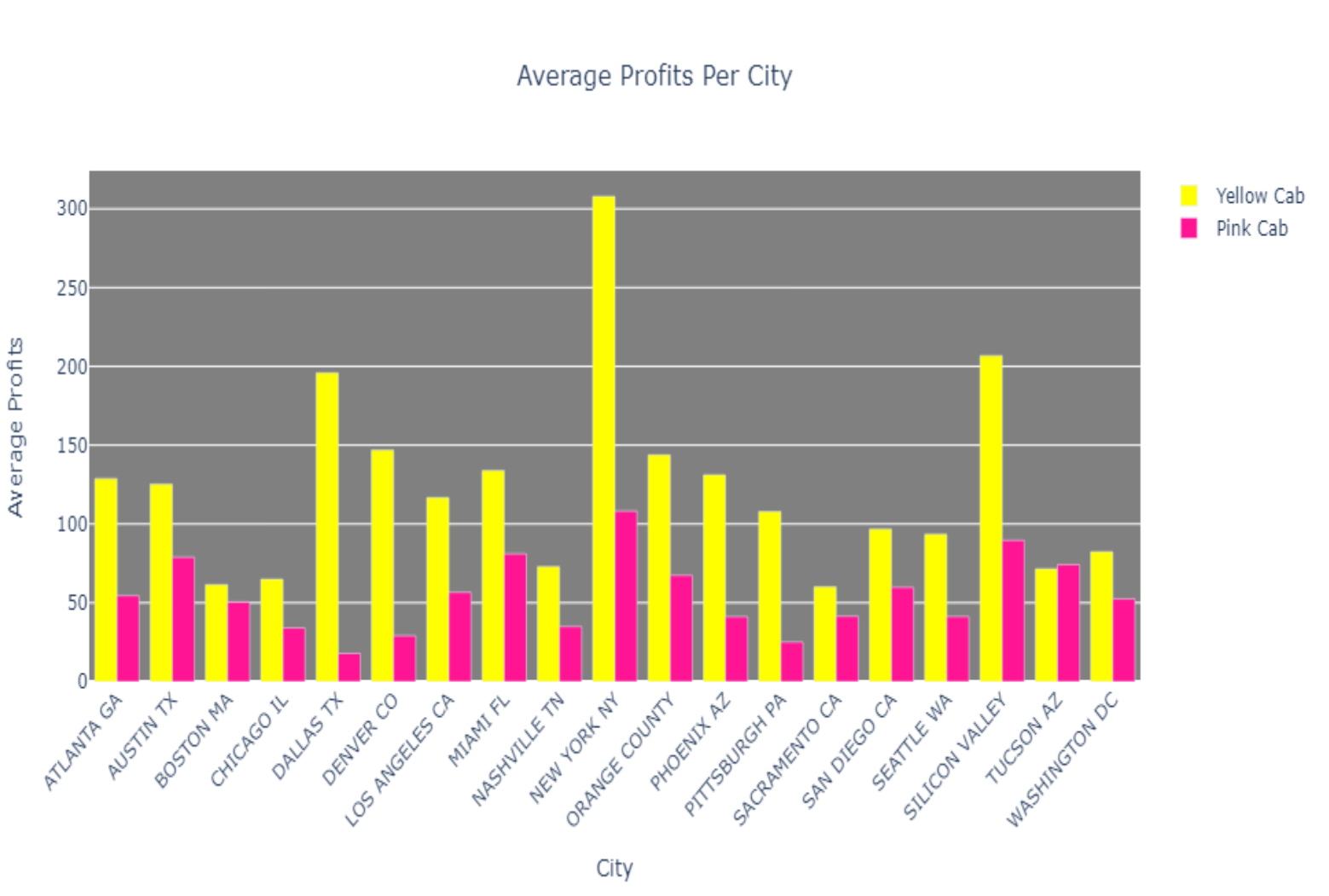
- The Yellow Cab has its average profits being significantly higher than the Pink Cab's.
- From 2016 to 2019, there exists an average profit gap of $89.86 in favor of the Yellow Cab.

Profit Analysis



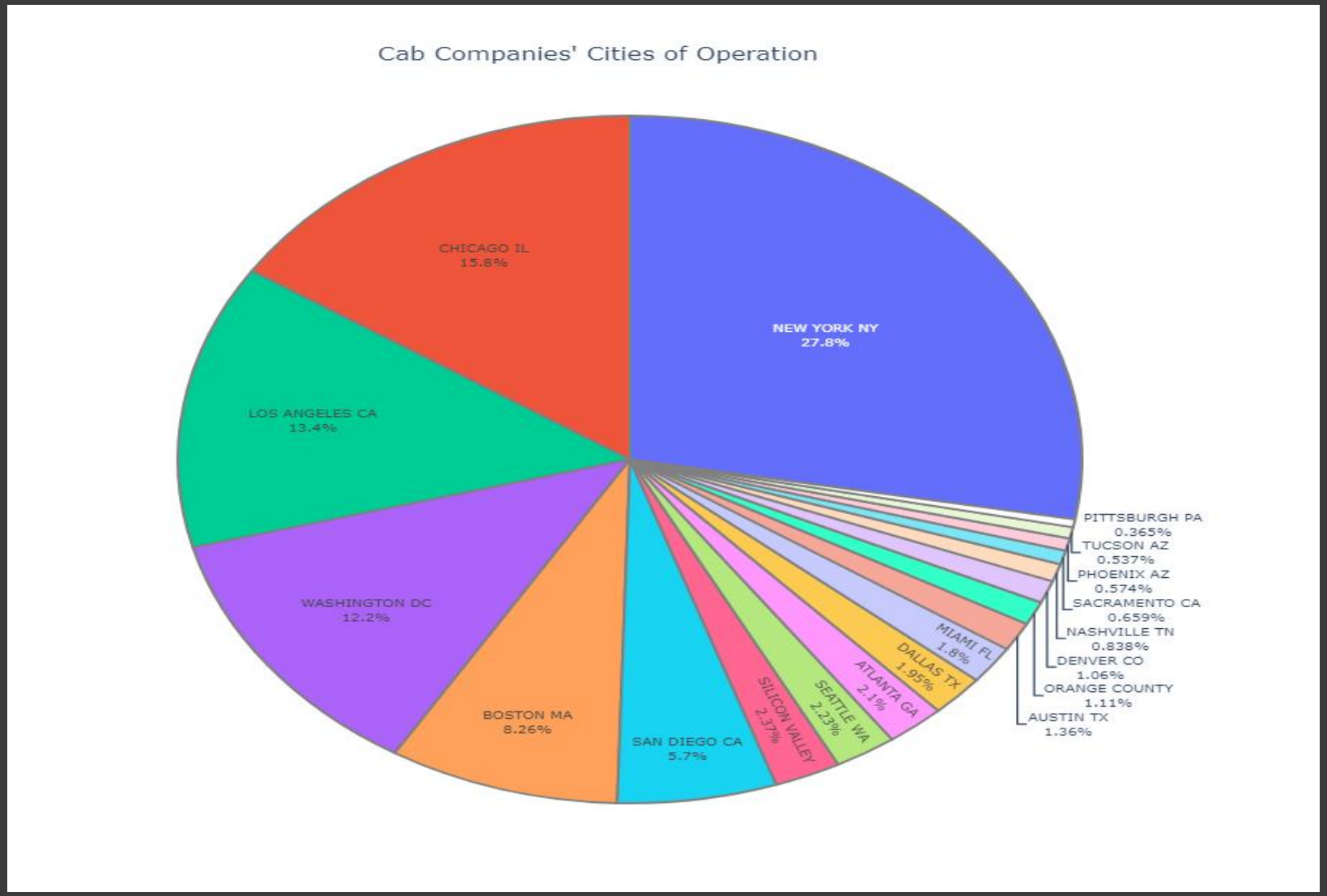
- Despite the Yellow Cab's notably higher annual profits, its earnings have shown a consistent decline from 2017 to 2019. Conversely, the Pink Cab experienced a surge in profits between 2018 and 2019.

Profit Analysis



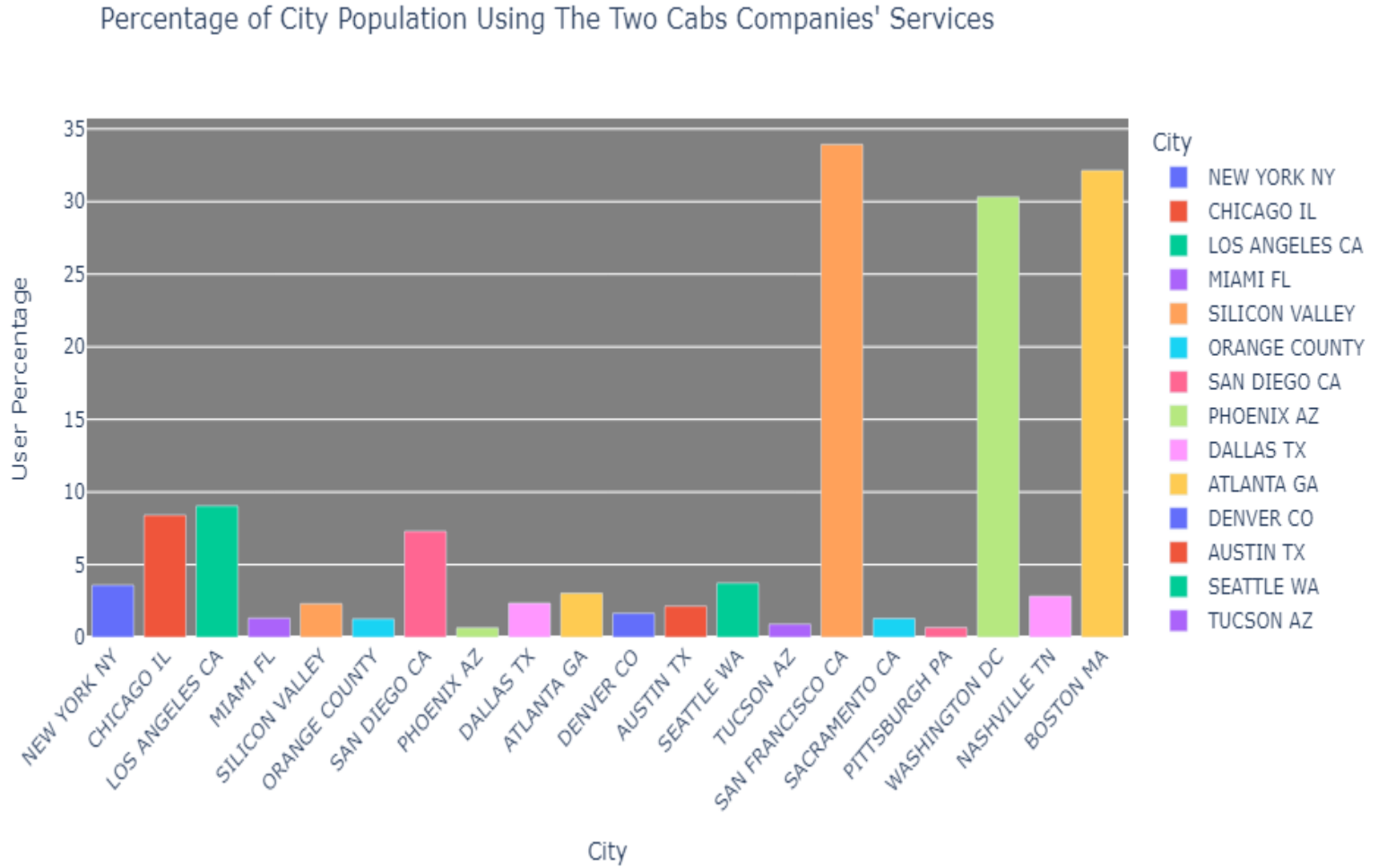
- New York, Silicon Valley and Dallas yield the highest average profits for the Yellow Cab company and Dallas is replaced by Miami for the Pink Cab company.
- We will see in the next slide that cities with most users are not necessarily the cities where the companies make the most average profit.

City Data Analysis



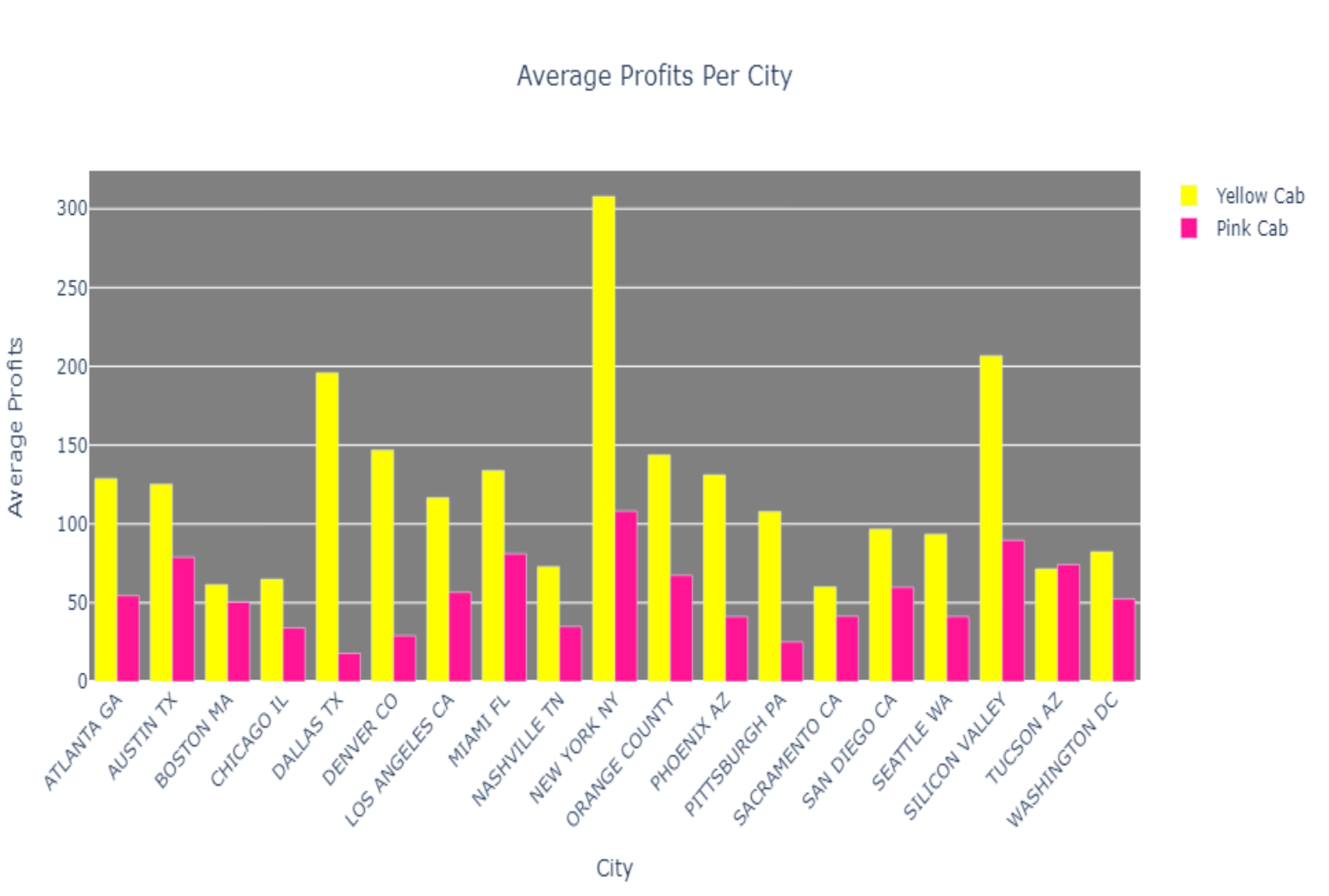
- New York, Chicago, Los Angeles, Washington, DC and Boston are the top 5 cities of operation for these two cab companies, with respect to number of users.

City Data Analysis



- The two Cab companies have over 30% of the entire population of Washington DC, San Francisco and Boston using their services.
- There is a lot of room for growth, especially considering that they have less than 10% of the people in each of the other cities using their services.

Profit Analysis



- New York, Silicon Valley and Dallas yield the highest average profits for the Yellow Cab company and Dallas is replaced by Miami for the Pink Cab company.
- We will see in the next slide that cities with most users are not necessarily the cities where the companies make the most average profit.

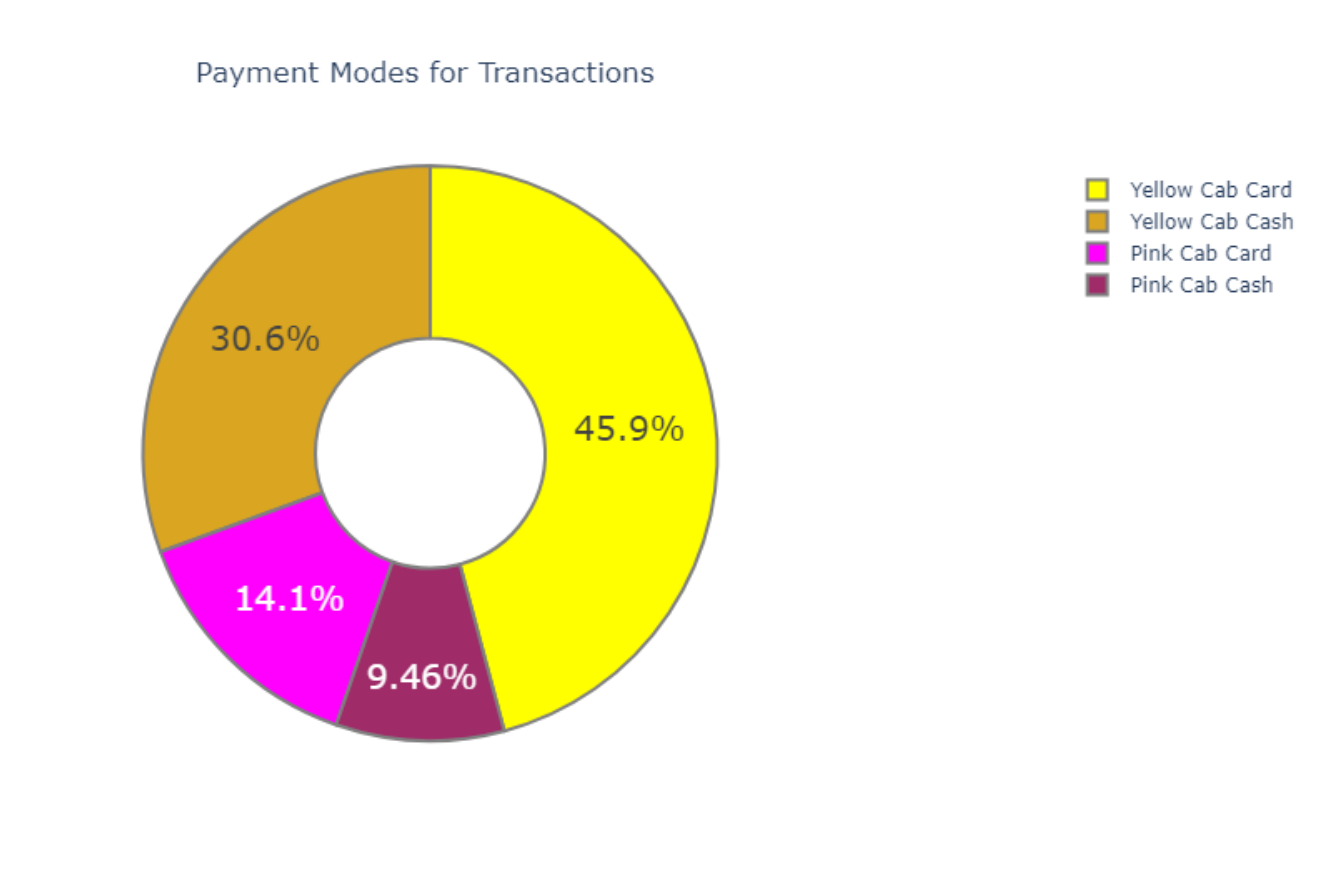
Customer Retention Analysis

Cab Company	5+ Transaction Customers	5+ Retention Rate	10+ Transaction Customers	10+ Retention Rate	15+ Transaction Customers	15+ Retention Rate
Pink	3690	4.36	240	0.28	7	0.01
Yellow	14631	5.33	9709	3.53	5474	1.99

- 5.33% of the Yellow Cab company's customers returned at least 5 times, compared to 4.36% for the Pink Cab.
- 3.53% of the Yellow Cab's customers appeared at least 10 times, only 0.28% for the Pink Cab company.
- At 15+ retention also, the Yellow Cab leads the way again, with 1.99% retention, compared to only 0.01% for the Pink Cab.

❖ We can confidently conclude that the Yellow Cab has better customer retention.

Mode of Payment Analysis

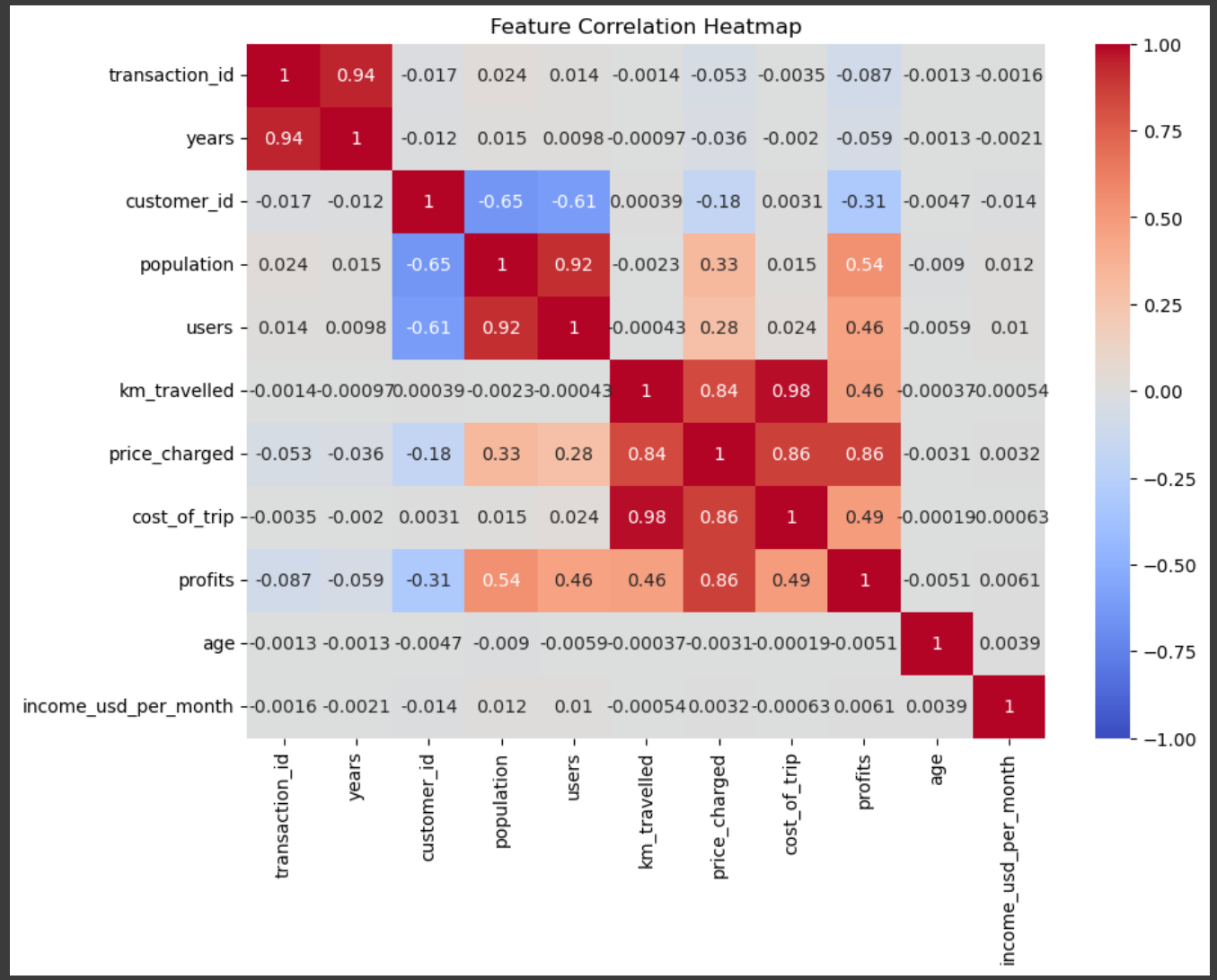


- Cards are the preferred payment mode for customers.
- Cards are used for payment 60% of the time, cash 40% - this is true for both Cab Companies.
- I suggest maintaining both card and cash payment alternatives since a considerable portion of customers utilizes each of these methods.

Price and Distance Correlation

- As expected, there is a positive correlation between Price Charged and the Distance travelled.

Feature Correlation Heatmap



- As expected, there is a strong correlation between Population and Users, as well as between Distance Travelled, Cost of Trip and Price Charged.

- Again, there is a strong correlation between Transaction ID and the Years of the transactions.
  - - A transaction at a later date would comprise of a greater number for its Transaction ID.

Recommendations

- The Cab companies collectively serve more than 30% of the total population in Washington DC, San Francisco, and Boston. Nevertheless, there exists significant potential for expansion, particularly given that their service utilization remains below 10% of the population in each of the other cities.
- I suggest that XYZ considers investing in the Yellow Cab company due to the following factors:
  - ❑ The Yellow Cab company exhibits a significantly superior customer retention rate compared to the Pink Cab. This implies that even during challenging periods, the Yellow Cab has a higher likelihood of enduring due to the strong customer loyalty it enjoys.
  - ❑ The Yellow Cab is the preferred choice for most customers, with 76.43% of all cab journeys, it trumps the Yellow Cab's 23.57%.
  - ❑ The Yellow Cab boasts notably higher average profits. Throughout the span of four years, there exists an average profit gap of $89.86 in favor of the Yellow Cab.
- Notable Drawback:
  - ❑ In spite of the Yellow Cab's notably higher annual profits, its earnings have shown a consistent decline from 2017 to 2019. Conversely, the Pink Cab experienced a surge in profits between 2018 and 2019.

Thank You