



**Data Glacier**

Your Deep Learning Partner

# Exploratory Data Analysis

G2M Case Study

Name: Freddy F. Tapia C. / XYZ

Location: Caracas, Venezuela

Team: Data and Analytics

Date: 14-March-2021

# Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

# Executive Summary

- 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.
- The main objective is to find the best company in terms of profitability in order to invest in it. To achieve this end we have information about each company, its costs, prices, profits, cities where it operates.
- To determine which company is more profitable, the EDA approach will be used, which allows us to test whether a hypothesis that we believe to be true is true or not, all this from the data.

# Problem Statement

XYZ is interested in using your actionable insights to help them identify the right company to make their investment. In order to do that we can begin with certain hypothesis which are important to achieve our goal,

- There is a quarter of the year in which the profits of a company have been higher than in any other quarter.
- In which city according to the type of company there are higher profits.
- What generation (age segmentation) and in which year, according to the type of company generates more profits.
- What social class (income segmentation) and in which year, according to the type of company generates more profits.
- According to the gender, the year and the type of company who generates more profits.

# Approach

In the elaboration of this analysis, several EDA techniques were used, which is defined as a process through which we want to find patterns, validate hypotheses, find insights, all from the data and its graphic representations. The steps that were carried out in this analysis are shown below,

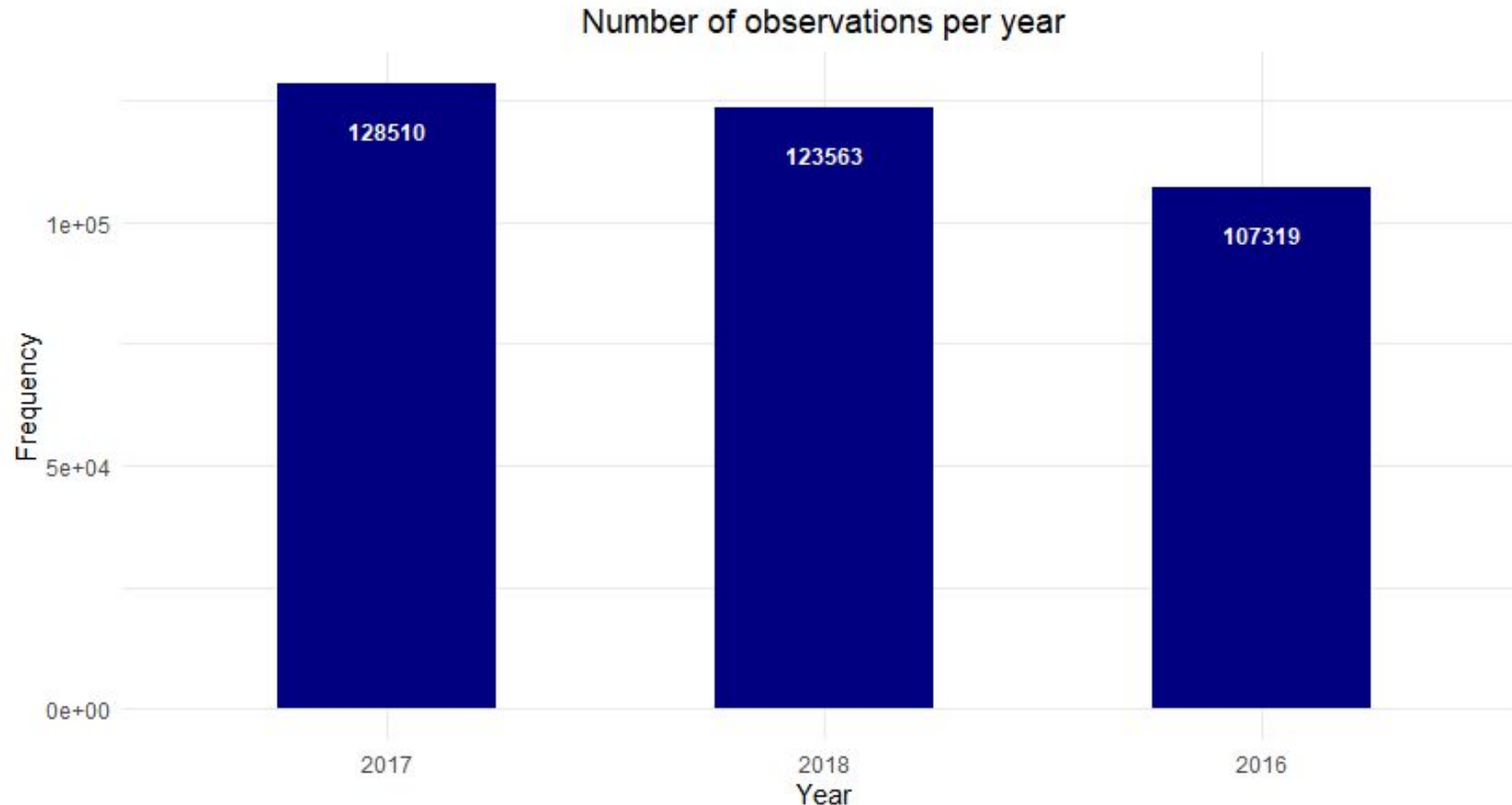
- Understand and validate each and every one of the data that was assigned to us.
- Perform an analysis and search for NA values as well as outliers.
- Create a master database in order to have all the information available in one place
- Create hypotheses that allow us to respond to the business problem and thus determine which company is the most profitable.
- Generate graphics in order to validate the hypothesis previously raised.
- Based on the results, generate final conclusions and recommendations.

# EDA

All the results were obtained from joined four files,

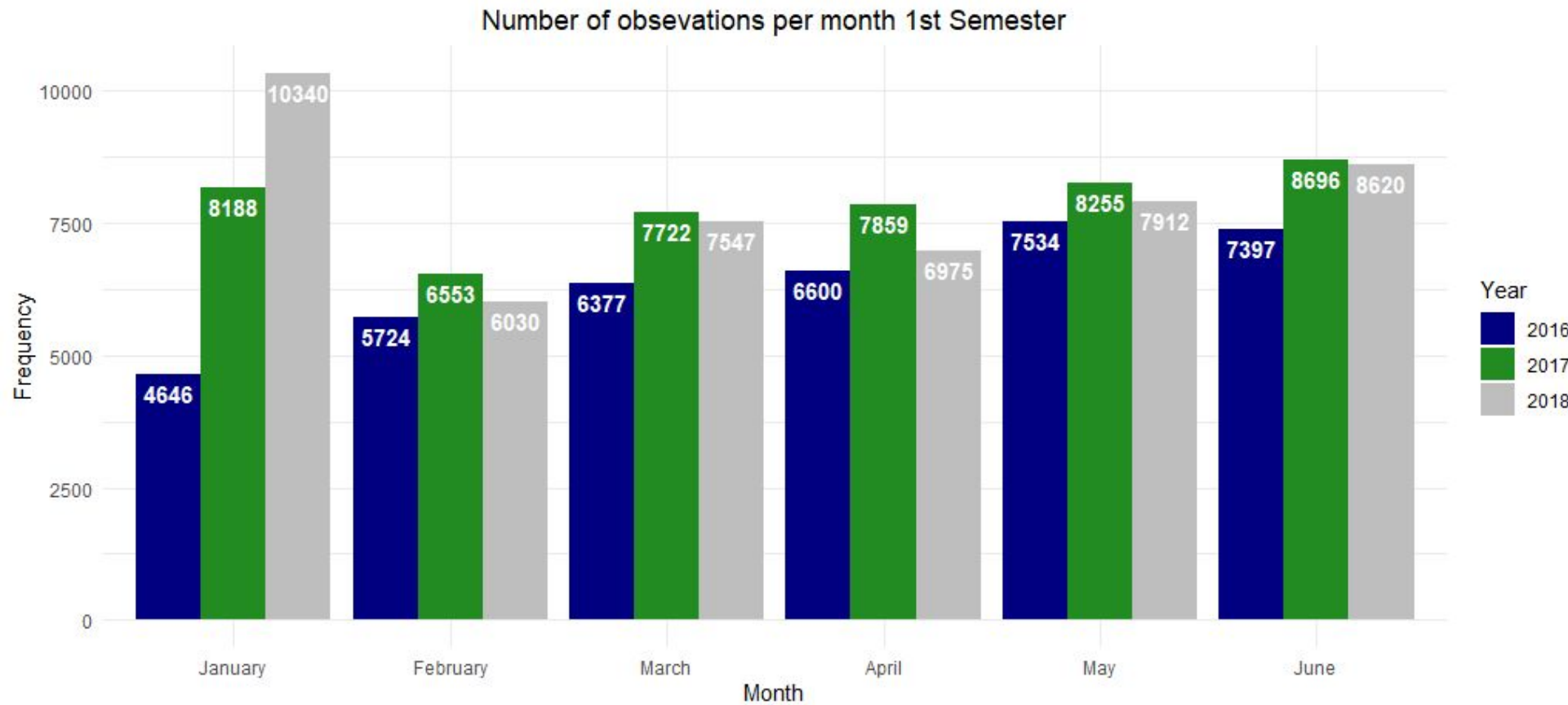
- City.csv (20 rows and 3 columns), Customer\_ID.csv (49,171 rows and 4 columns), Transaction\_ID.csv (440,098 rows and 3 columns), Cab\_Data.csv (359,392 rows and 7 columns).
- The final data has 359,392 rows and 15 columns. This data was obtained by merging the Customer\_ID.csv, Transaction\_ID.csv and Cab\_Data.csv files.
- Profit column was calculated as the difference between Price and Cost columns. For the finals results two important and interesting segmentations were made, the first was with generation and the second was with income salary, the sources of these segmentation are in the nexts slides.

# EDA - Number of observations



- The year 2017 has the mayor number of observations.
- The year 2016 has the minor number of observations.

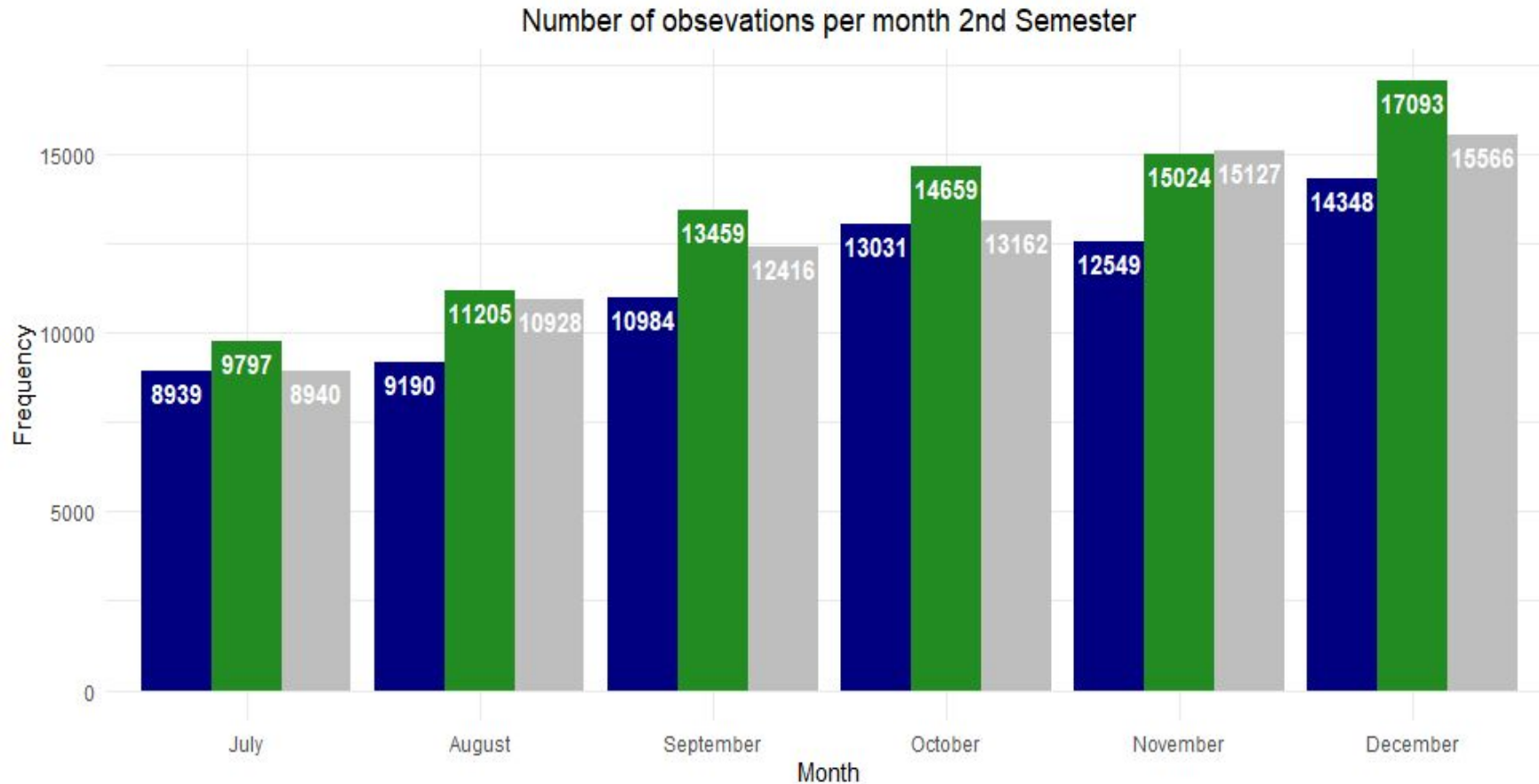
# EDA - Number of observations



- January 2018 has the mayor number of observations in the first semester.

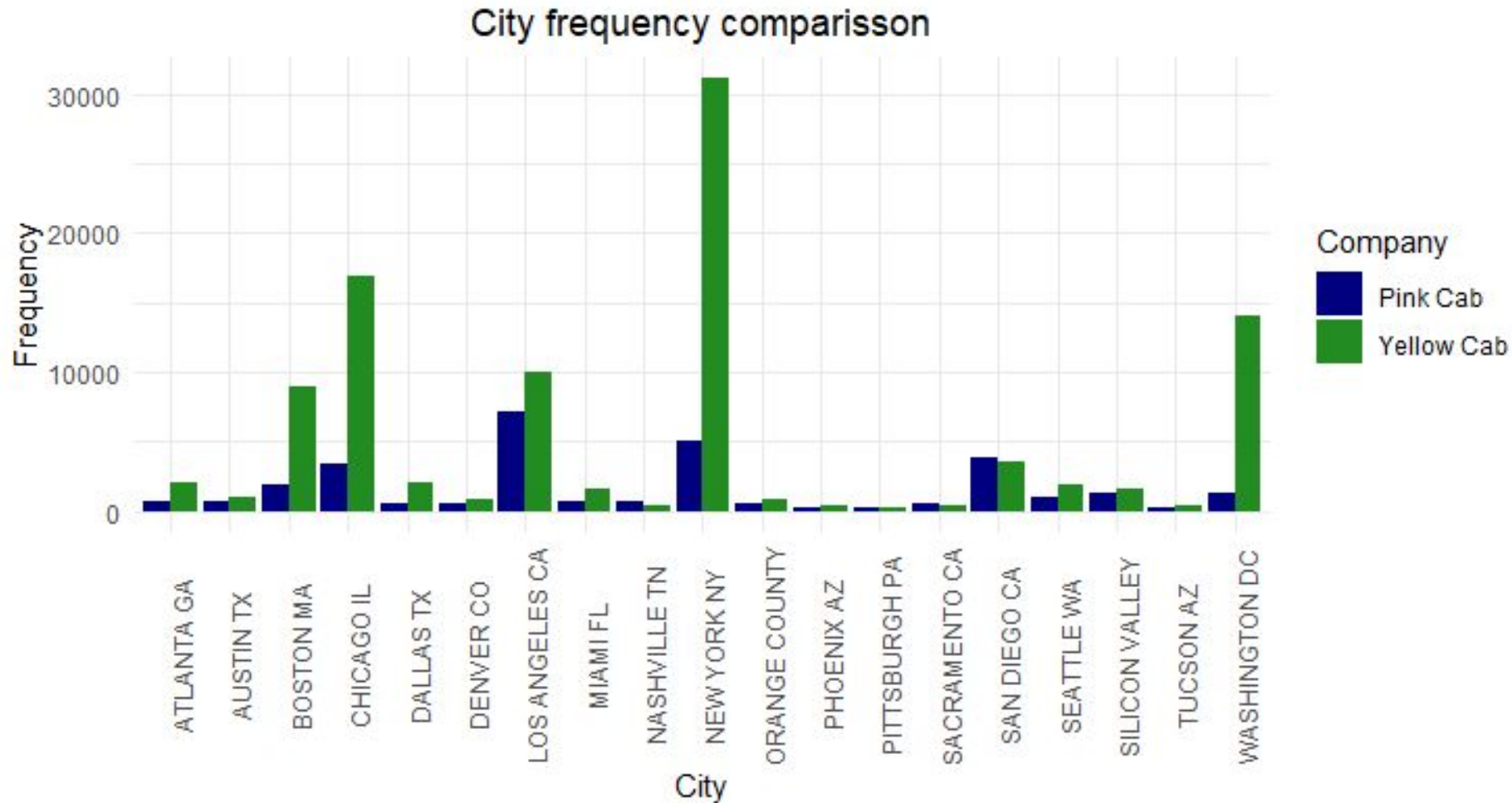


# EDA - Number of observations



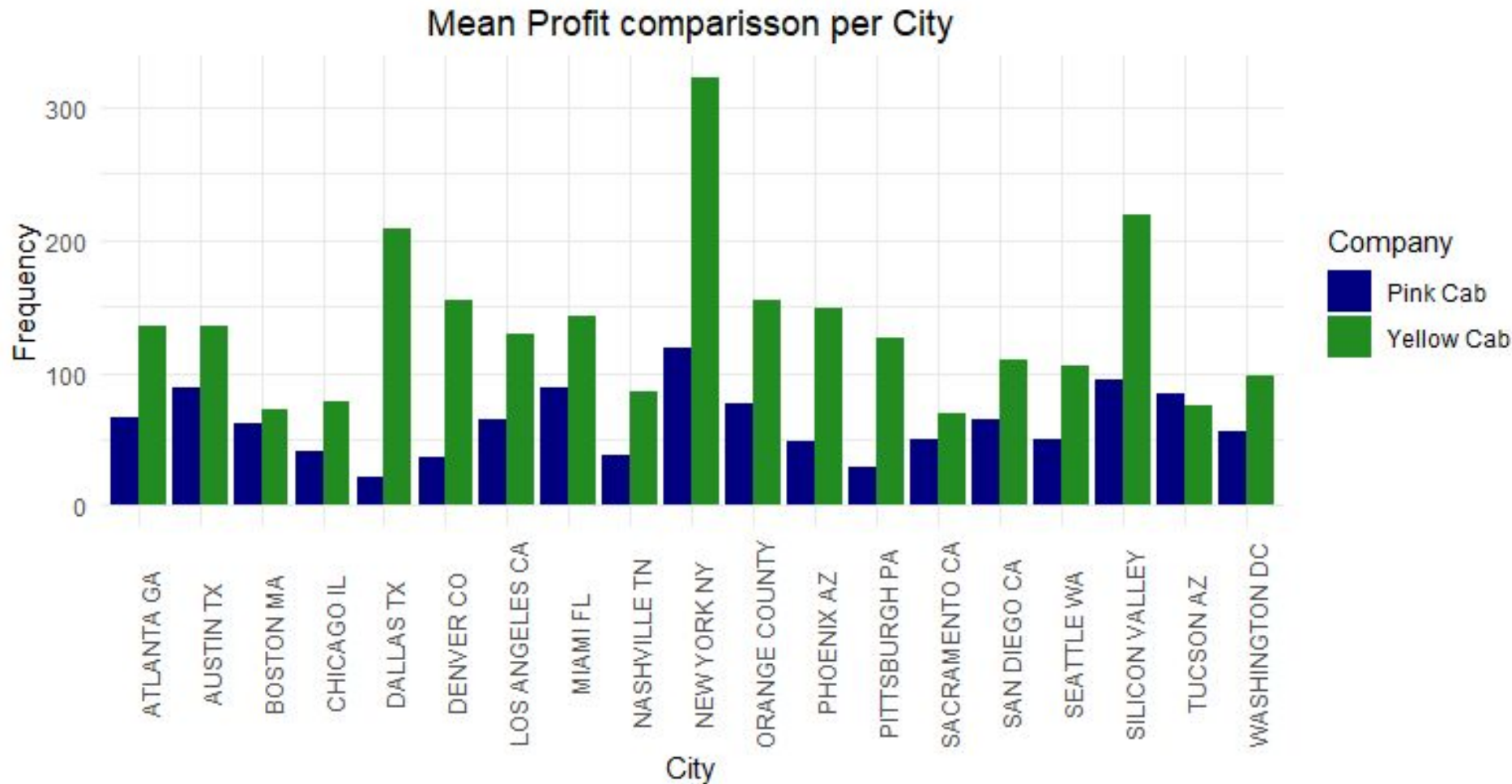
- December 2017 has the mayor number of observations in the second semester.

# EDA - City frequency



- New York, Washington DC. and Chicago are important cities for Yellow Cab company.
- Los Angeles, New York and San Diego are important cities for Pink Cab company.

# EDA - Mean Profit comparisson



- Nueva York, Dallas and Silicon Valley generates the highest mean profits for Yellow Cab company.
- Nueva York, Miami and Silicon Valley generates the highest mean profits for Pink Cab company.

# EDA - Total profit per year



- The best year for both Cab companies in terms of profit was 2017.

# EDA - Total rides per year



- 2017 was the year with more rides for both companies.

# EDA - Profit per ride



- For Yellow Cab company 2017 was the year with the best proportion ride / profit.
- For Pink Cab company 2016 was the year with the best proportion ride / profit.

# EDA - Profit per Generation

Profit sum per year and generation

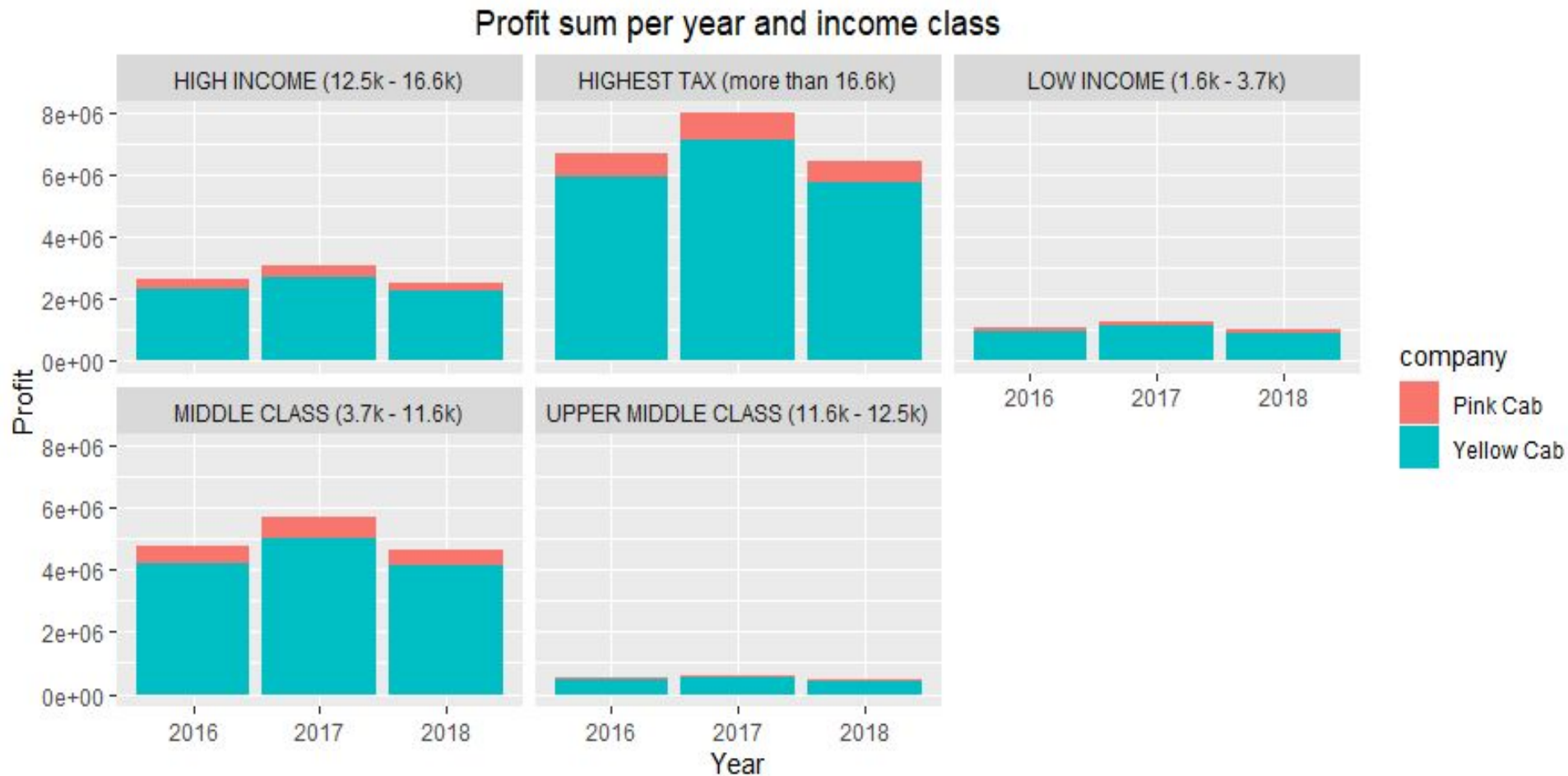


- For both companies the Millennials generation generates the highest profits.
- On the other hand, the Boomers II generation generates the lowest profits.

**Segmentation source:** <https://www.beresfordresearch.com/age-range-by-generation/>



# EDA - Profit per Income class



- For both companies the highest Tax class generates the highest profits.
- On the other hand the Upper middle class generates the lowest profits.

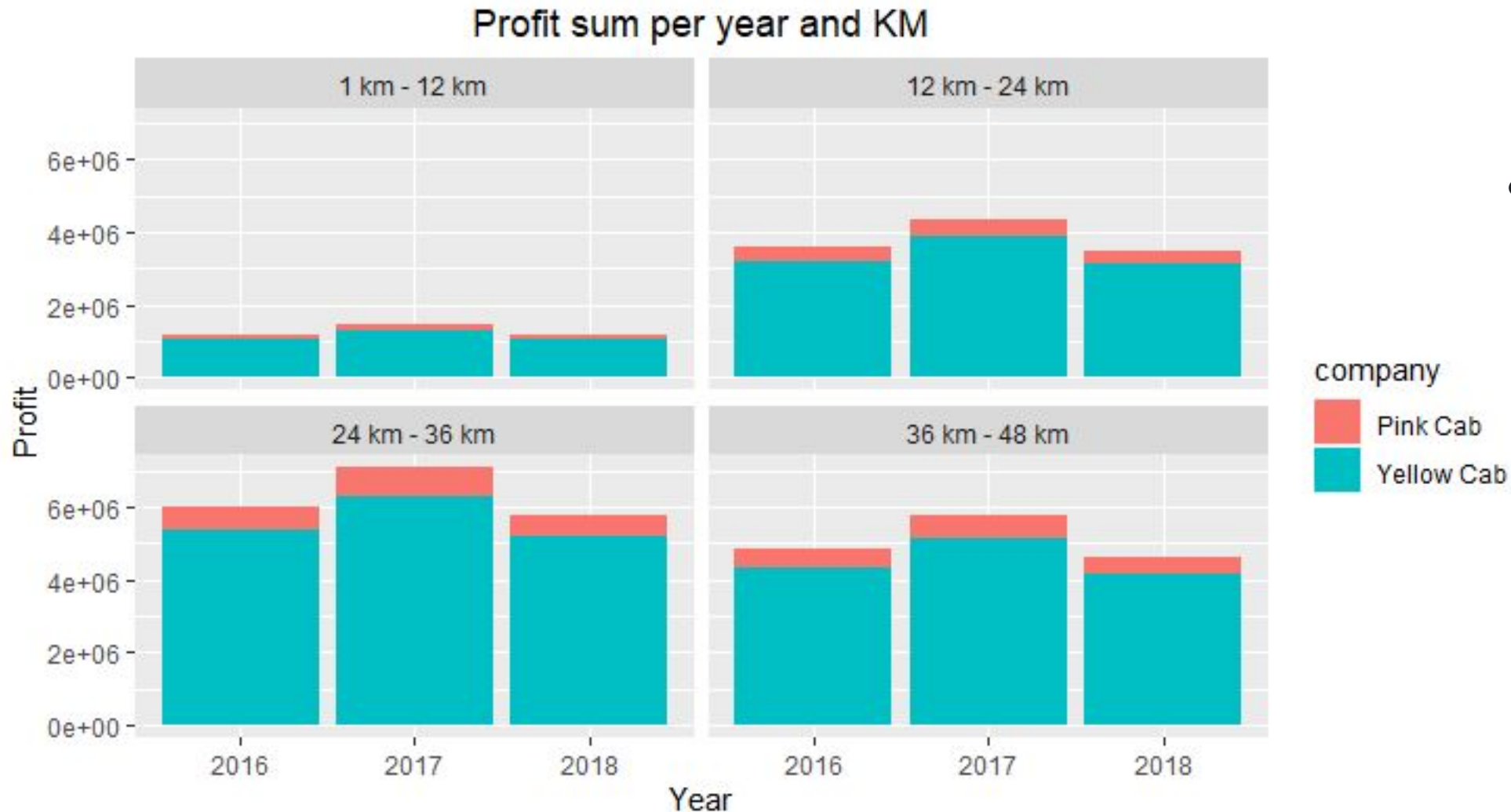


# EDA - Profit per Gender



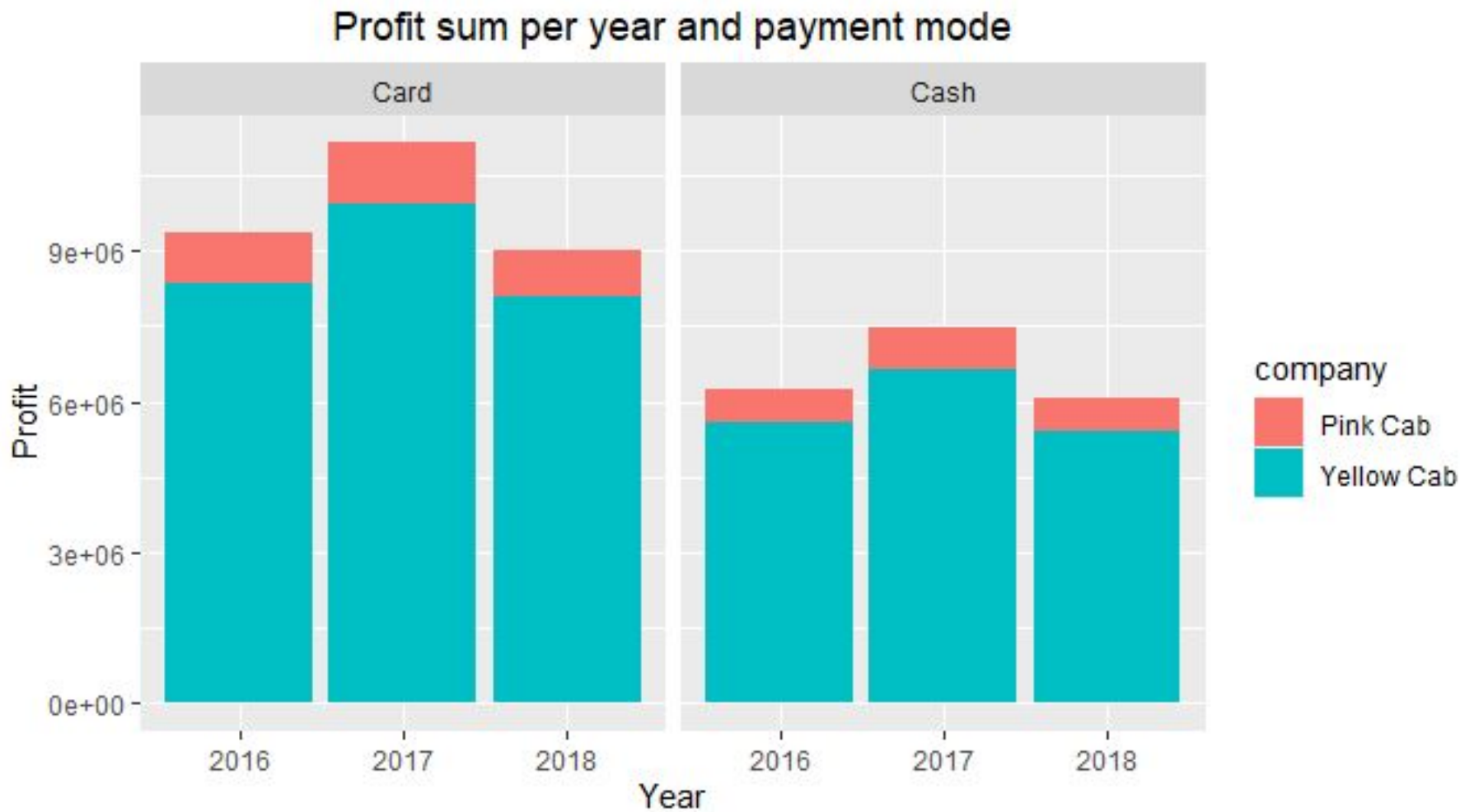
- The male gender generates the highest profits for both companies.

# EDA - Profit per KM



- The rides with a distance between 24 km and 36 km generates the highest profits for both companies.

# EDA - Profit per payment mode



- The card payment mode generates the highest profits for both companies.

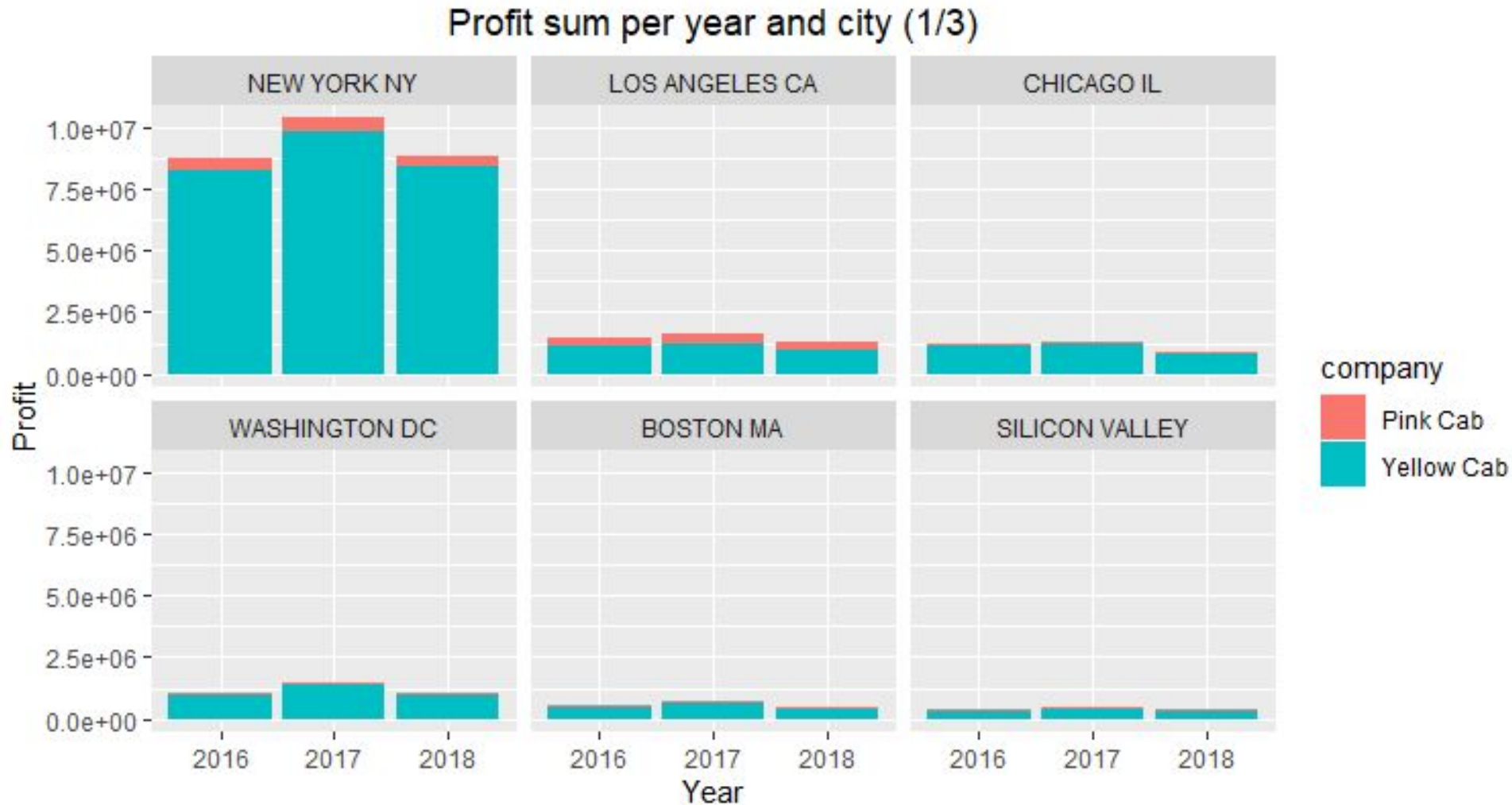
# EDA - Profit average cities

Profit sum per year and city (2/3)



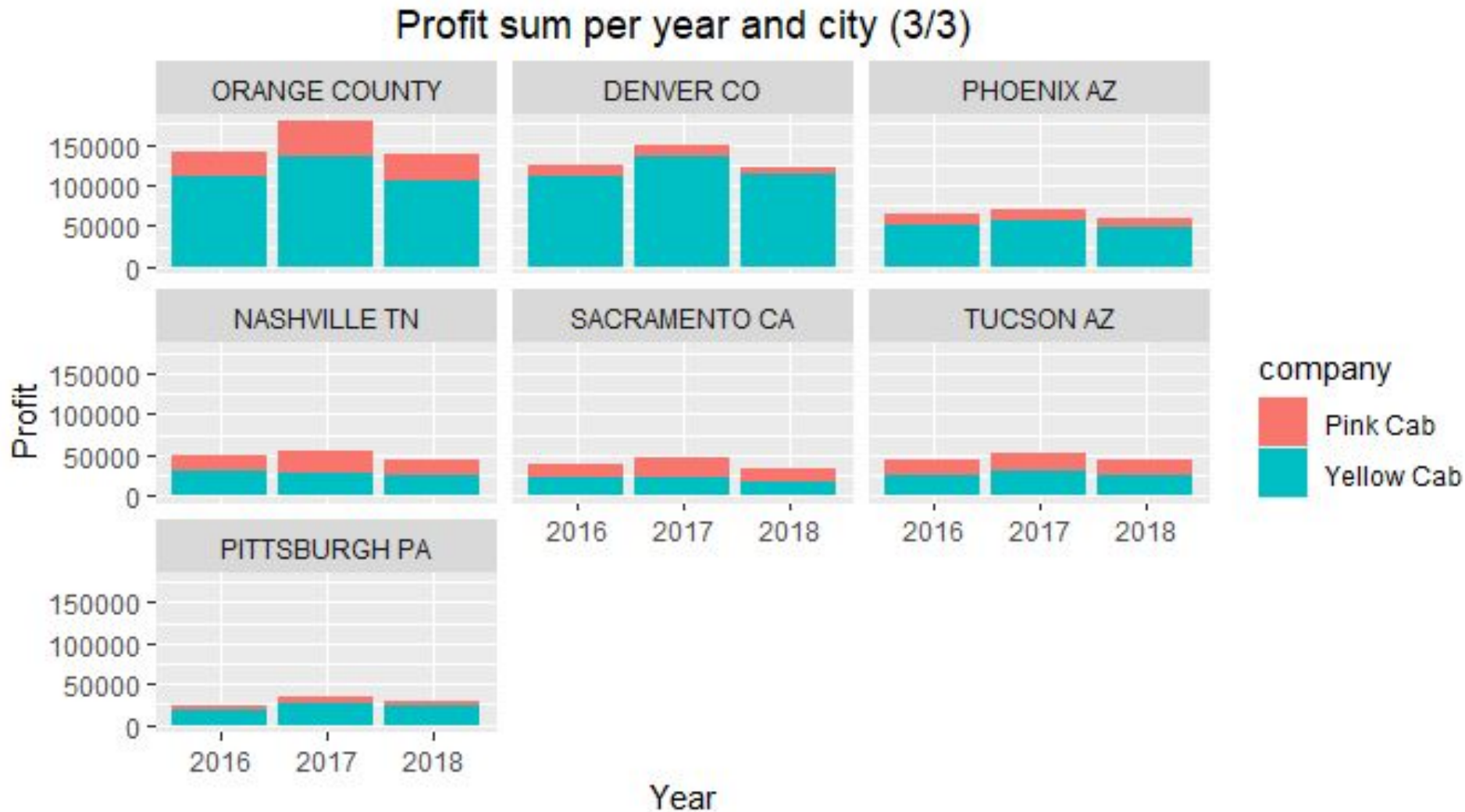
- San Diego is a important city to Pink Cab company, their profits are almost the same of Yellow Cab company.

# EDA - Profit top six cities



- The top six cities in terms of profits are in the graph.

# EDA - Seven low profit cities



- For both companies Pittsburgh represent the lowest profit.

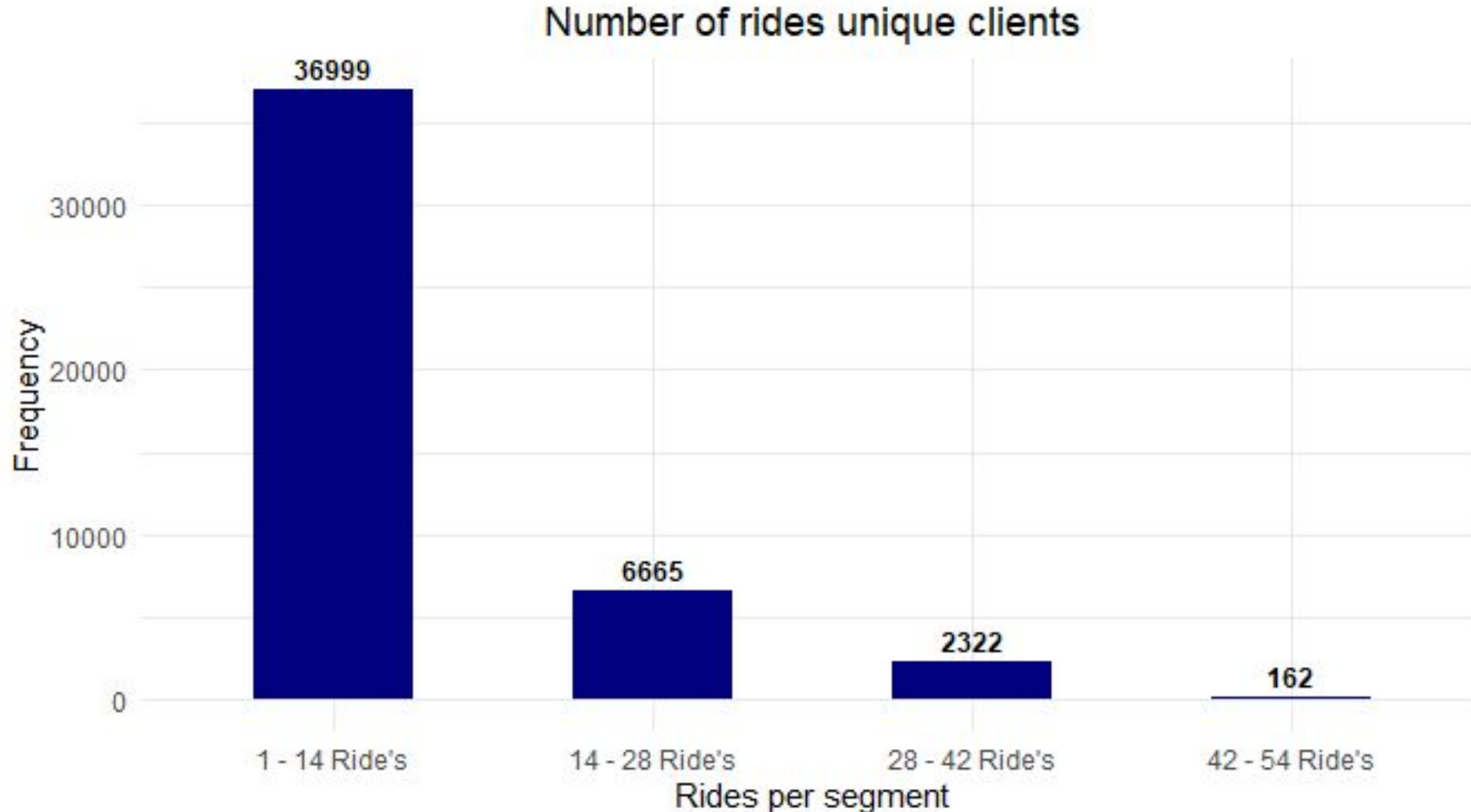


# EDA - Profit per quarter



- The four quarter of the year is the period of time where the profit are higher.

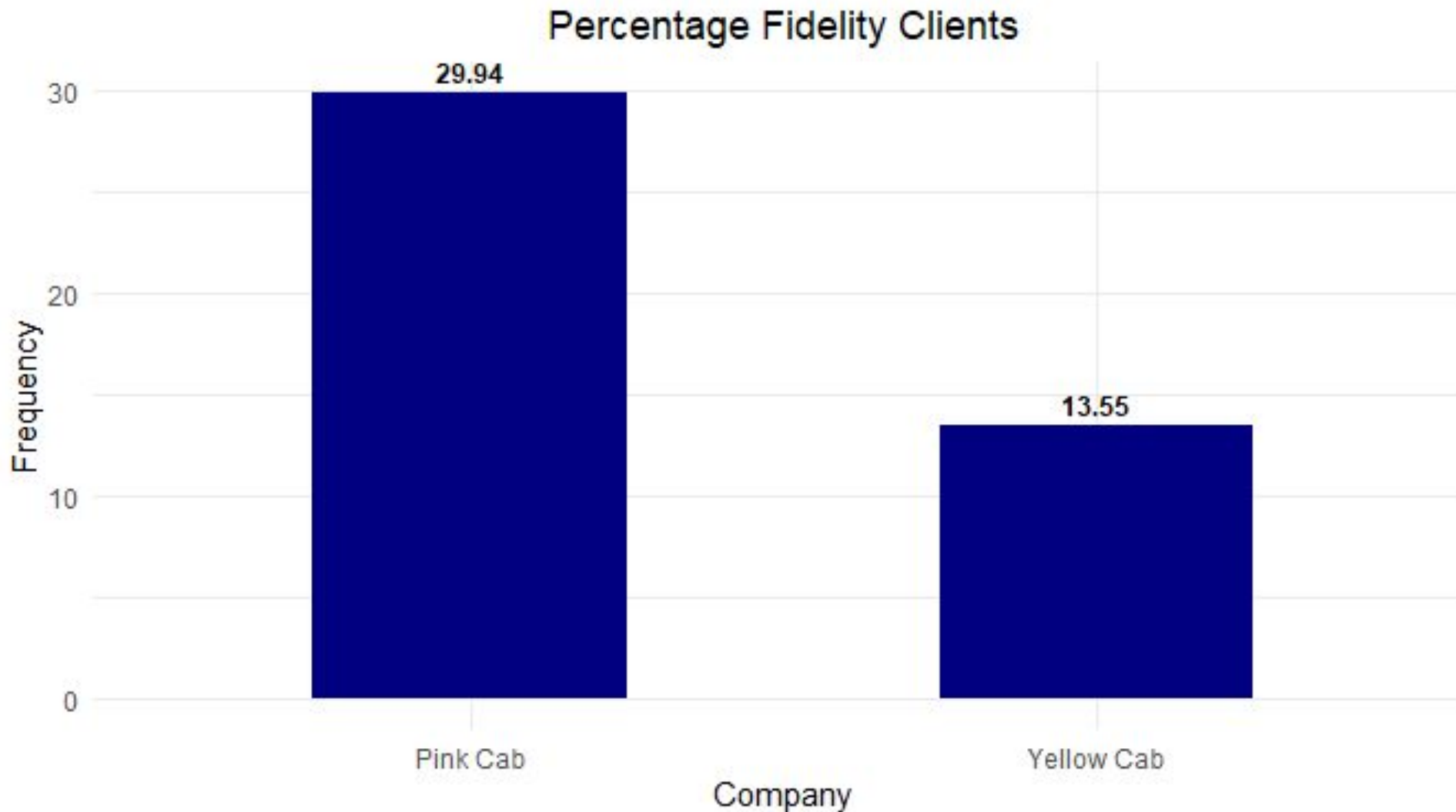
# EDA - Number of rides segmentation



- The majority of the customers have a number of rides between 1 and 14. Only 162 customers have more than 42 rides.



# EDA - Customers Fidelity



- The customers of the Pink Cab company are more “loyal” than the customers of Yellow Cab company.
- A 29.94% of the whole customers have been using this company.

# EDA Summary

The most important results that we obtain with this analysis are,

- The vast majority of observations belongs to Yellow Cab company and are in the year 2017 in the last Quarter.
- In general New York, Washington DC, Chicago and Los Angeles are the cities which have the majority of rides for both companies. New York in terms of profits is the best city for both companies.
- The best year for both Cab companies in terms of profit was 2017.
- For both companies the Millennials generation generates the highest profits.
- For both companies the highest Tax class generates the highest profits.
- Almost all customer have between 1 and 14 rides in the whole data.
- 29.94% of the whole customers have been using Pink Cab company, this show the loyalty of these customers to that company.

# Recommendations

According with the EDA and the results, the final recommendation is that the XYZ firm must to invert in the Yellow Cab company, following the main reasons,

- In general Yellow Cab company has better performance in terms of profits.
- The Yellow Cab company has the **majority of rides** in several cities this help to this company to obtain more profits.
- The **generation segment** that XYZ firm must to attack or offer this services is the **Millennials**, because this generates the higher profits.
- The **income class** that XYZ firm must to attack is the **highest Tax class**, because this generates the higher profits.
- It is recommend that XYZ firm consider the males customers, payment in card and the distance of a rides between 24 km and 36 km as important segment to consider to generates more profits.

# Thank You