Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

<G2M Case Study>

Name: Robin Masawi
Location: Harare, Zimbabwe
Team: Data Science
Date: 14-Mar-2021

Agenda

Background

Data Exploration and Approach

EDA and Summary

Hypothesis Testing

Recommendations

Background

- XYZ is a private firm in US and 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.

- Objective:

Summarize your analysis and recommendations and identify which company is performing better and is a better investment opportunity for XYZ.

- Data Available:

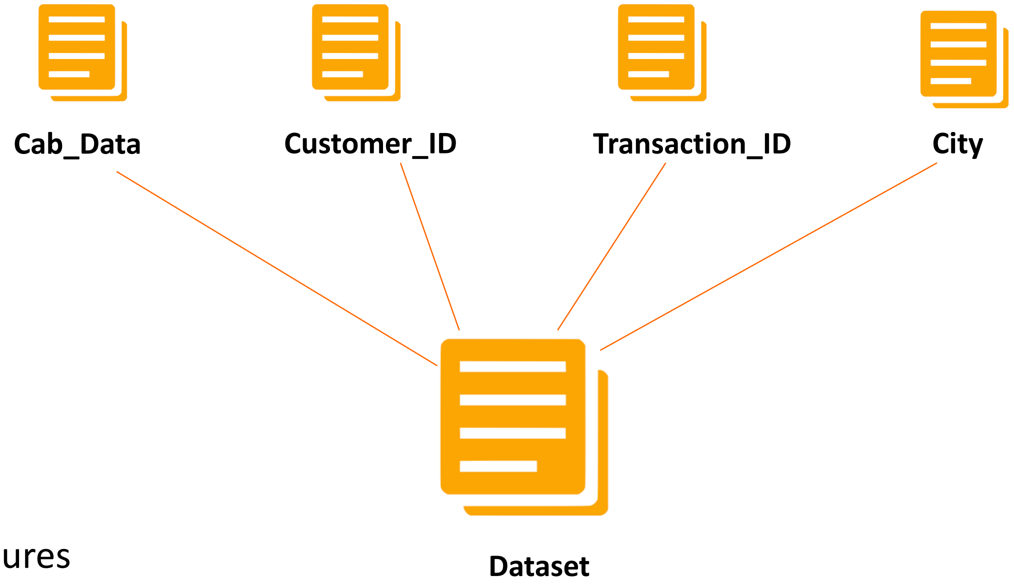
+ Multiple datasets for two companies have been provided.
+ Each data set provides different aspects of the customer's profile:

1. Cab Data: Includes details of transaction for the two cab companies.
2. Transaction ID: Mapping table that contains transaction to customer mapping and payment mode.
3. Customer ID: Mapping table that contains a unique identifier which links the customer's demographic details.
4. City: Contains list of US cities, their population and number of cab users.

Data Exploration and Approach

Dataset

- 4 datasets with 19 unique features (5 derived).
- Time period for data: 31/01/2016 to 31/12/2018.
- Total data points: 359,393.



Approach

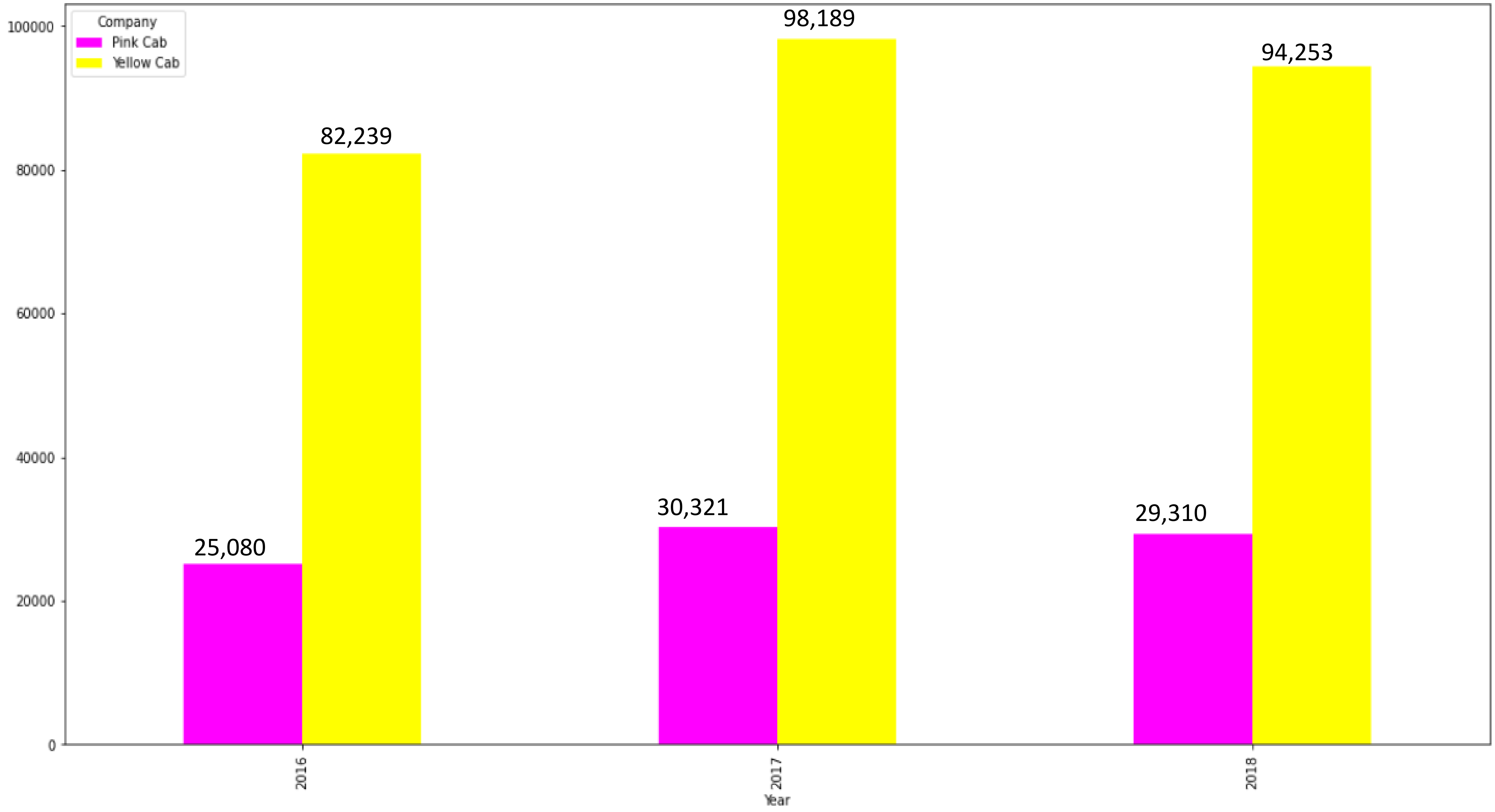
- The datasets were all combined to create one master dataset.
- 5 features were derived from the datasets available:

1. Month and Year: these were derived from the date_of_travel feature.
2. Profit: this is the difference between price charged and cost_of_trip features
3. Age_range: the ages of the customers were allocated to different bins.
4. Percentage_users: this is a ratio in percentile of the users in each city to the population of that city.

- Exploratory Data Analysis approach utilized to draw insights from the data.

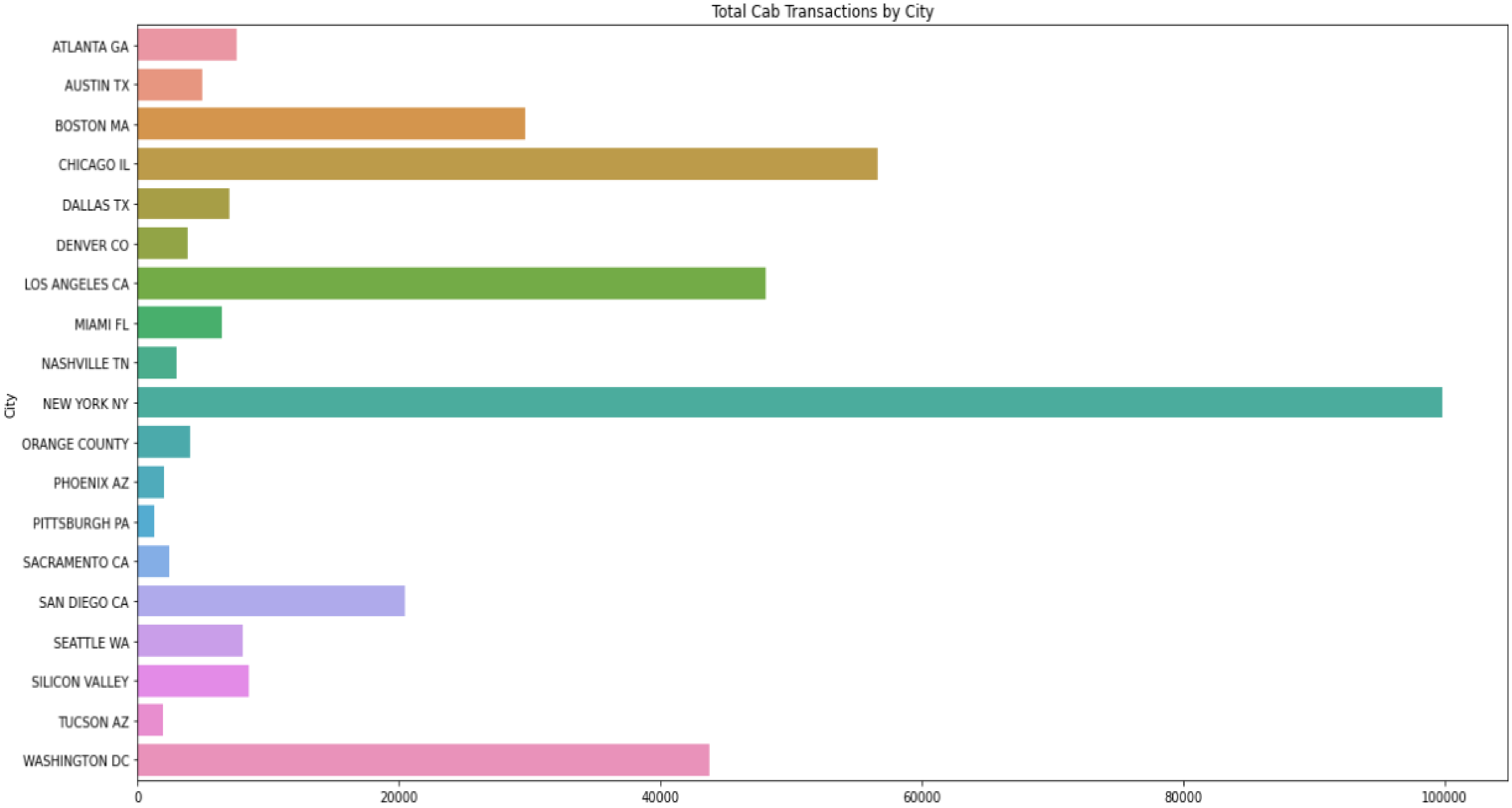
This refers to the critical process of performing initial investigations on data so as to discover patterns, spot anomalies, test hypothesis and check assumptions with the help of summary statistics and graphical representations.

EDA and Summary – Yearly Transaction Analysis



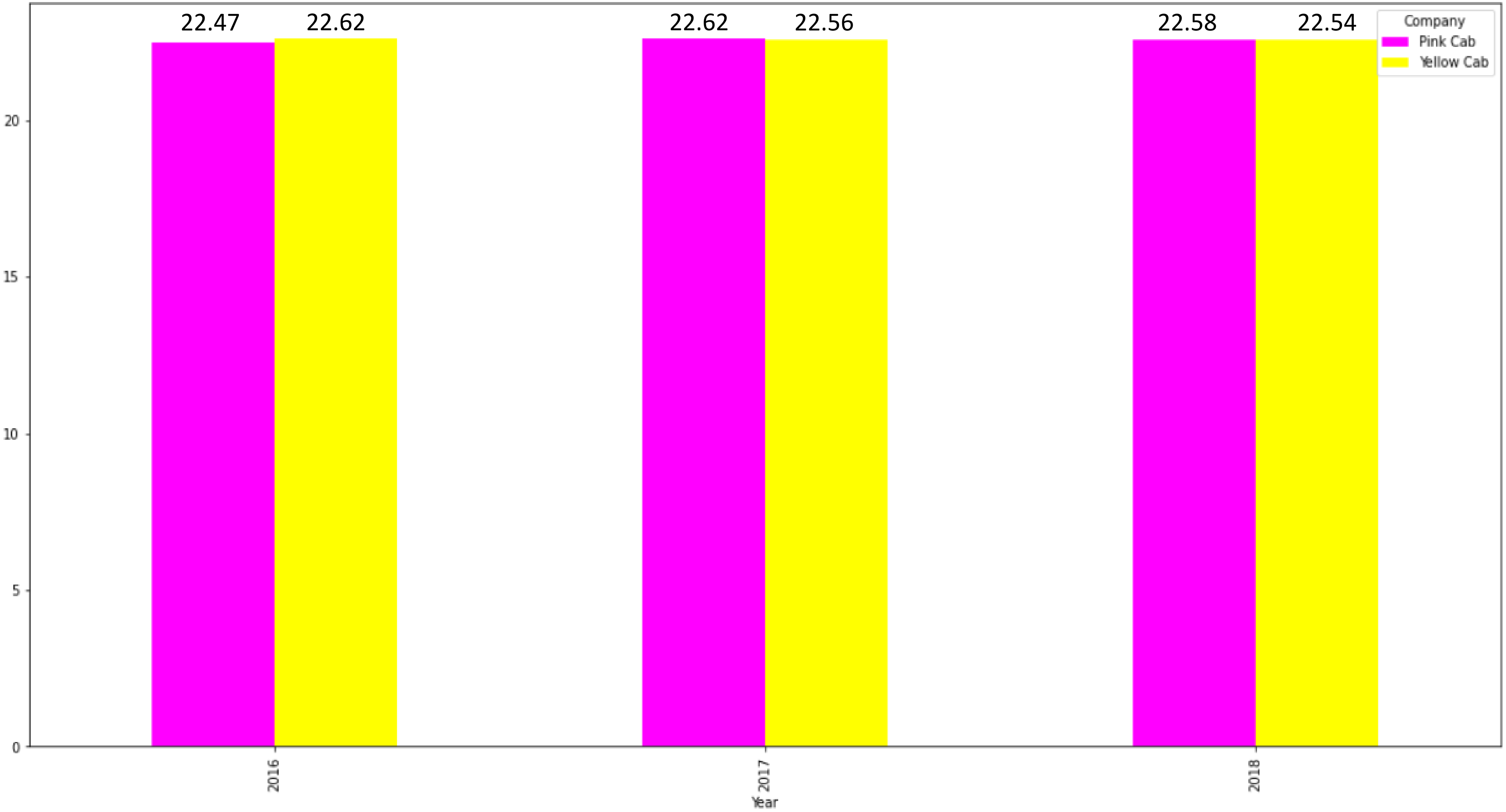
- Yellow Cab seems to dominate the market with the most transactions on a yearly basis compared to Pink Cab.

EDA and Summary - Transaction Analysis by City



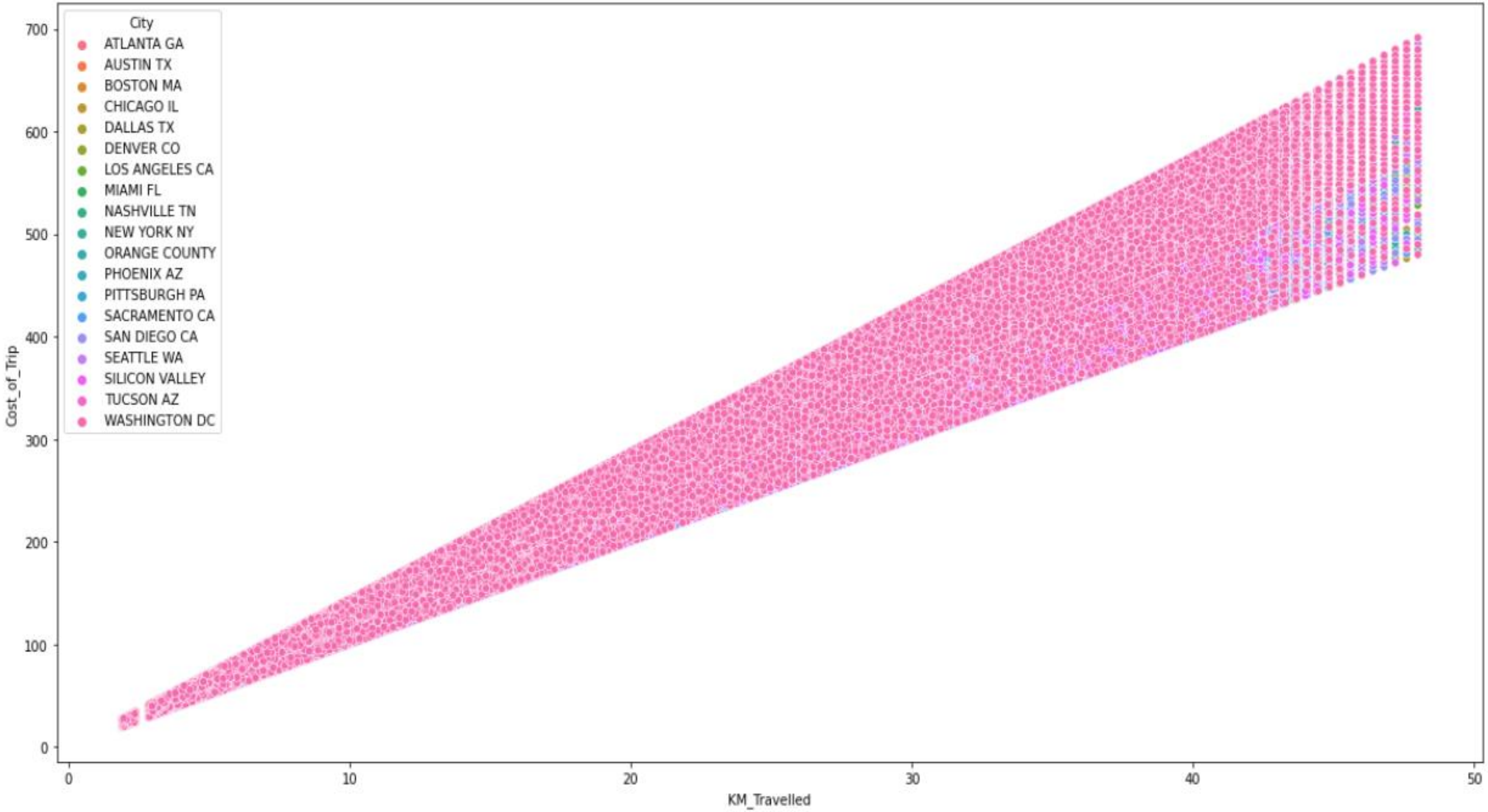
- New York has the highest number of Transactions followed by Chicago, Los Angeles and Washington DC.

EDA and Summary – KM Travelled



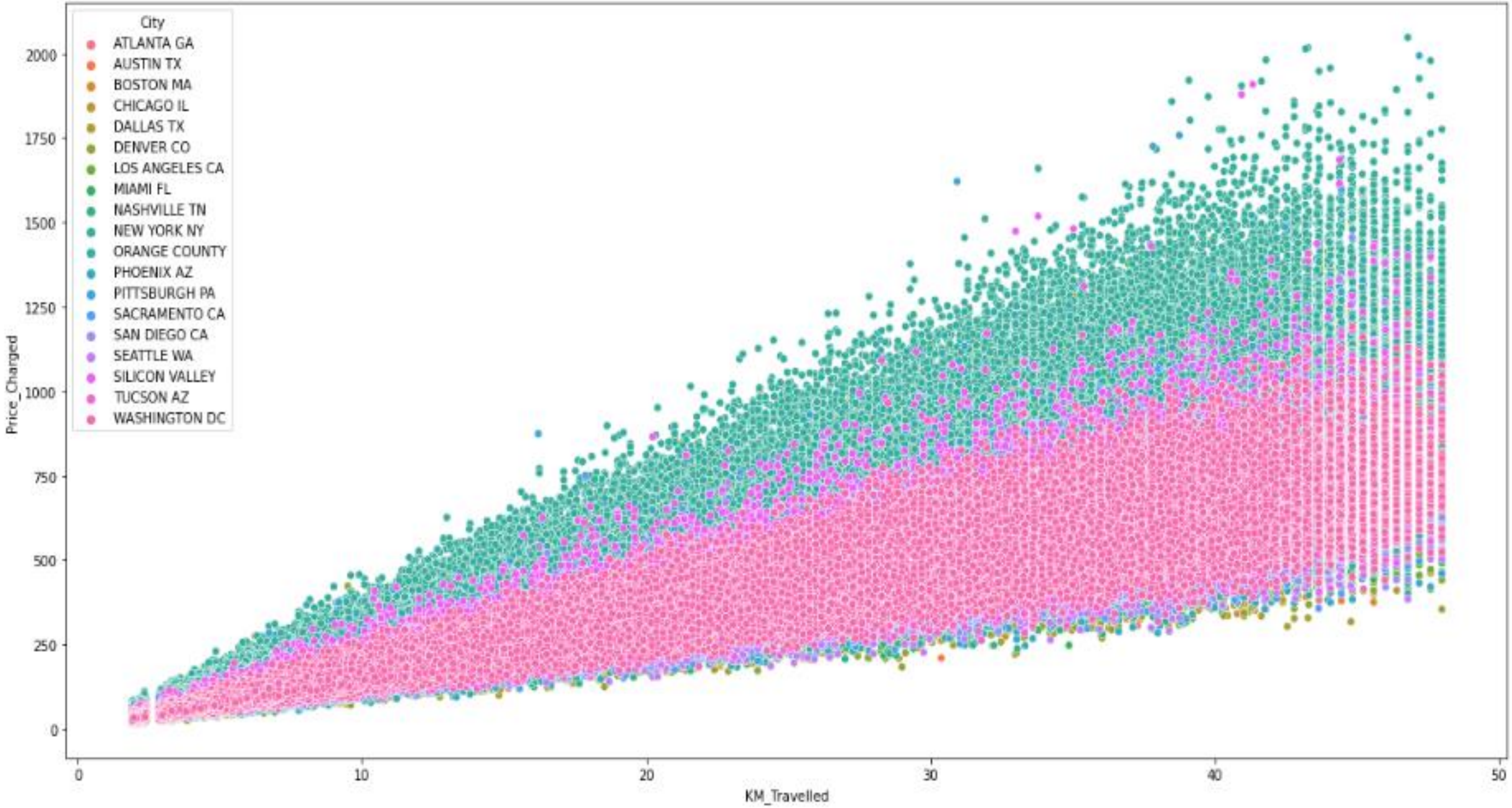
- There is an equal distribution of mean KM Travelled for both cab companies.

EDA and Summary – Cost of Trip and KM Travelled



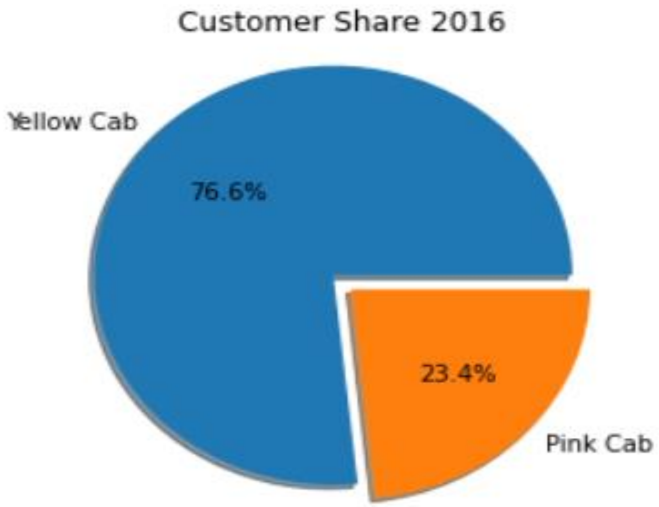
- Cost of Trip and KM Travelled are directly proportional but not depending according to city.

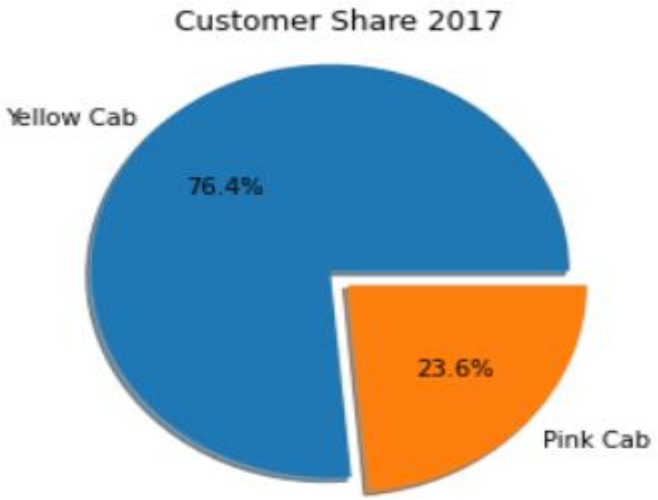
EDA and Summary – Price Charged and KM Travelled

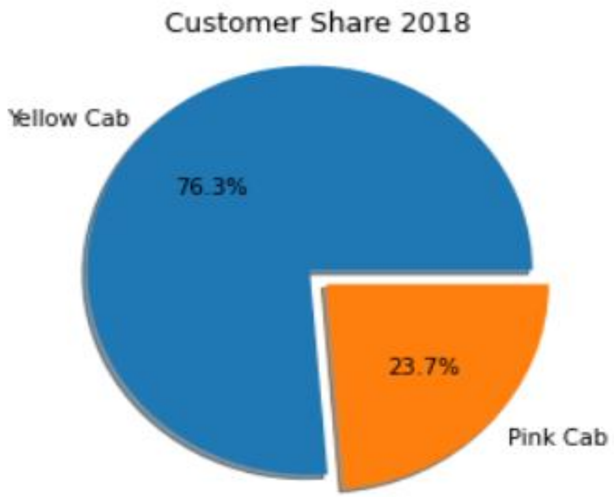


- New York and Silicon Valley cost more in cab fare than other cities.

EDA and Summary – Customer Share



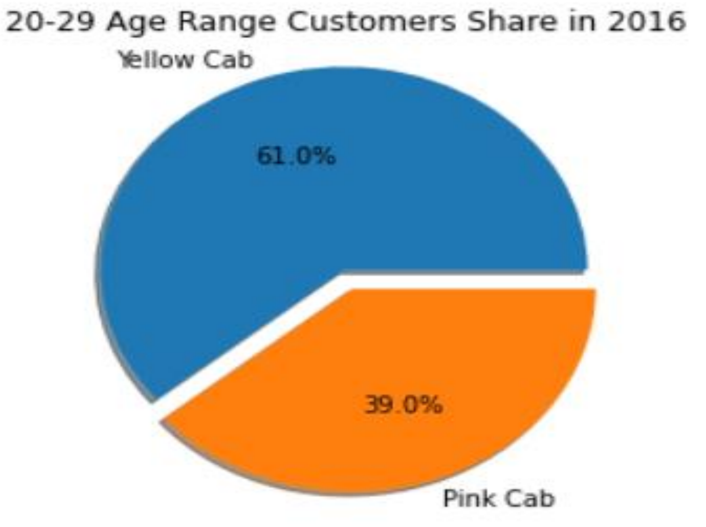


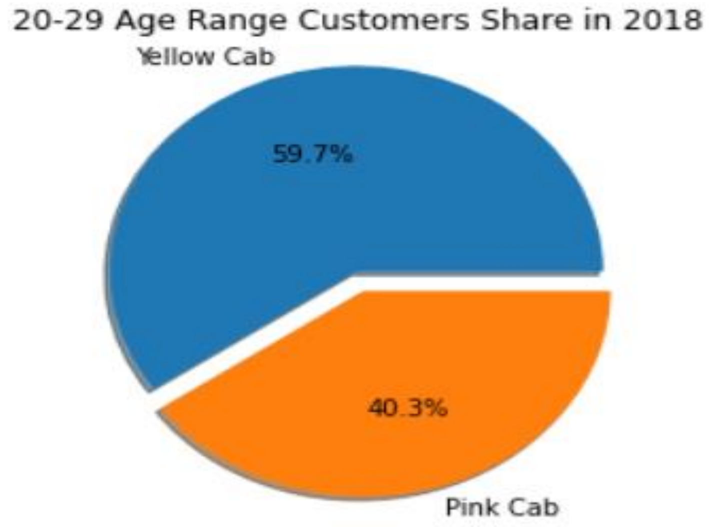


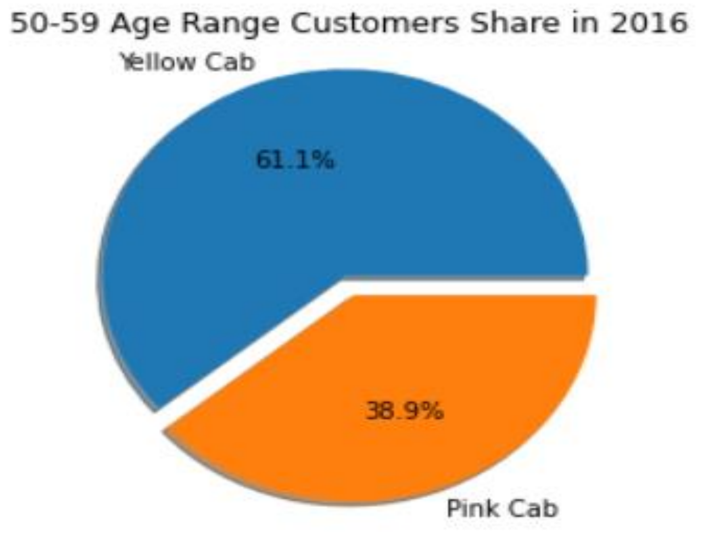
Year	Company	Total Customers
2016	Pink Cab	16,661
	Yellow Cab	25,937
2017	Pink Cab	18,643
	Yellow Cab	27,789
2018	Pink Cab	18,400
	Yellow Cab	27,470

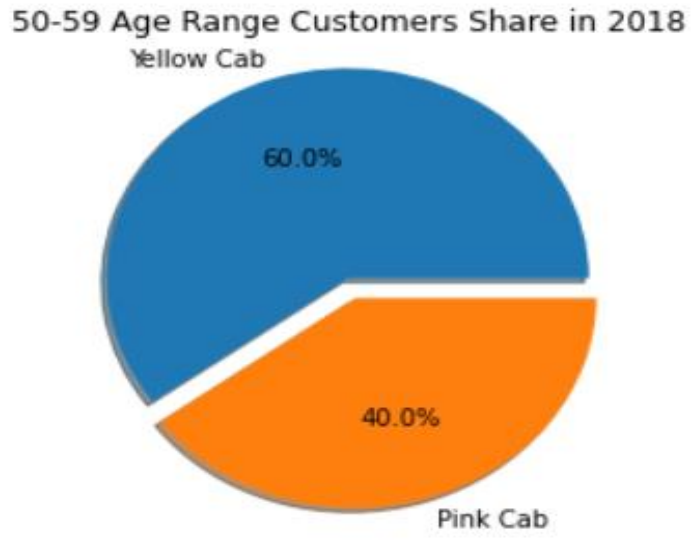
- Yellow Cab dominates more than half of the customer base for the three years.

EDA and Summary – Age and Customer Share Analysis


20-29 Age Range Customers Share in 2016

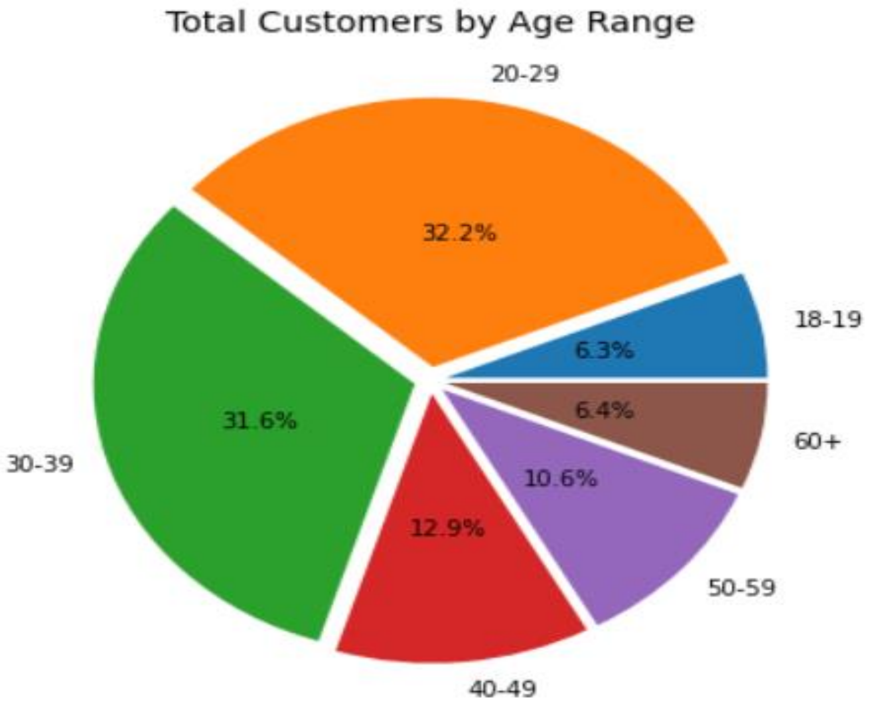

20-29 Age Range Customers Share in 2018


50-59 Age Range Customers Share in 2016

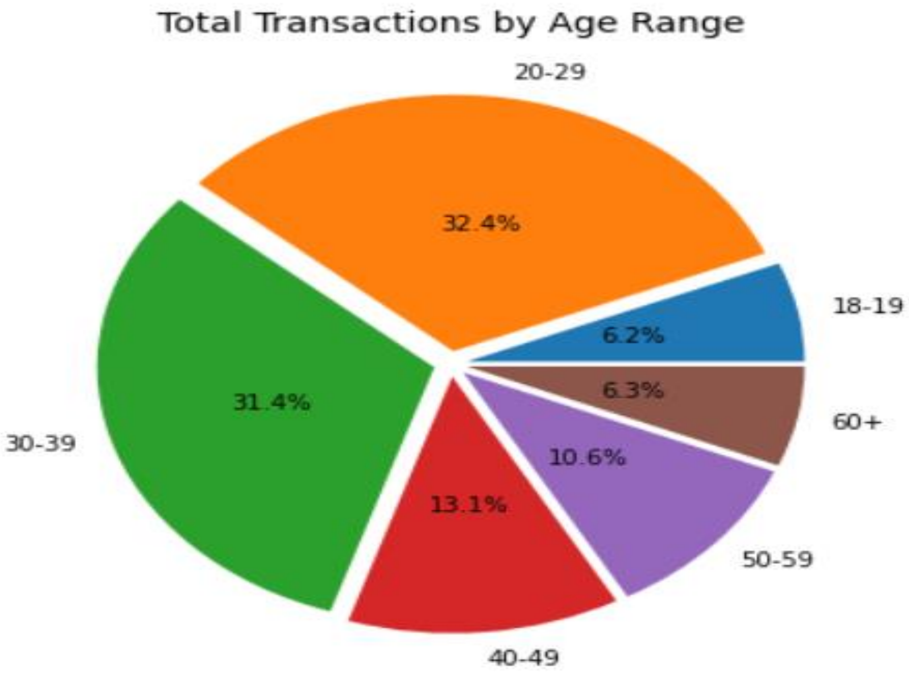

50-59 Age Range Customers Share in 2018

- As highlighted for these age groups Yellow Cab dropped in customers share for 2018 compared to 2016.

EDA and Summary – Age Analysis



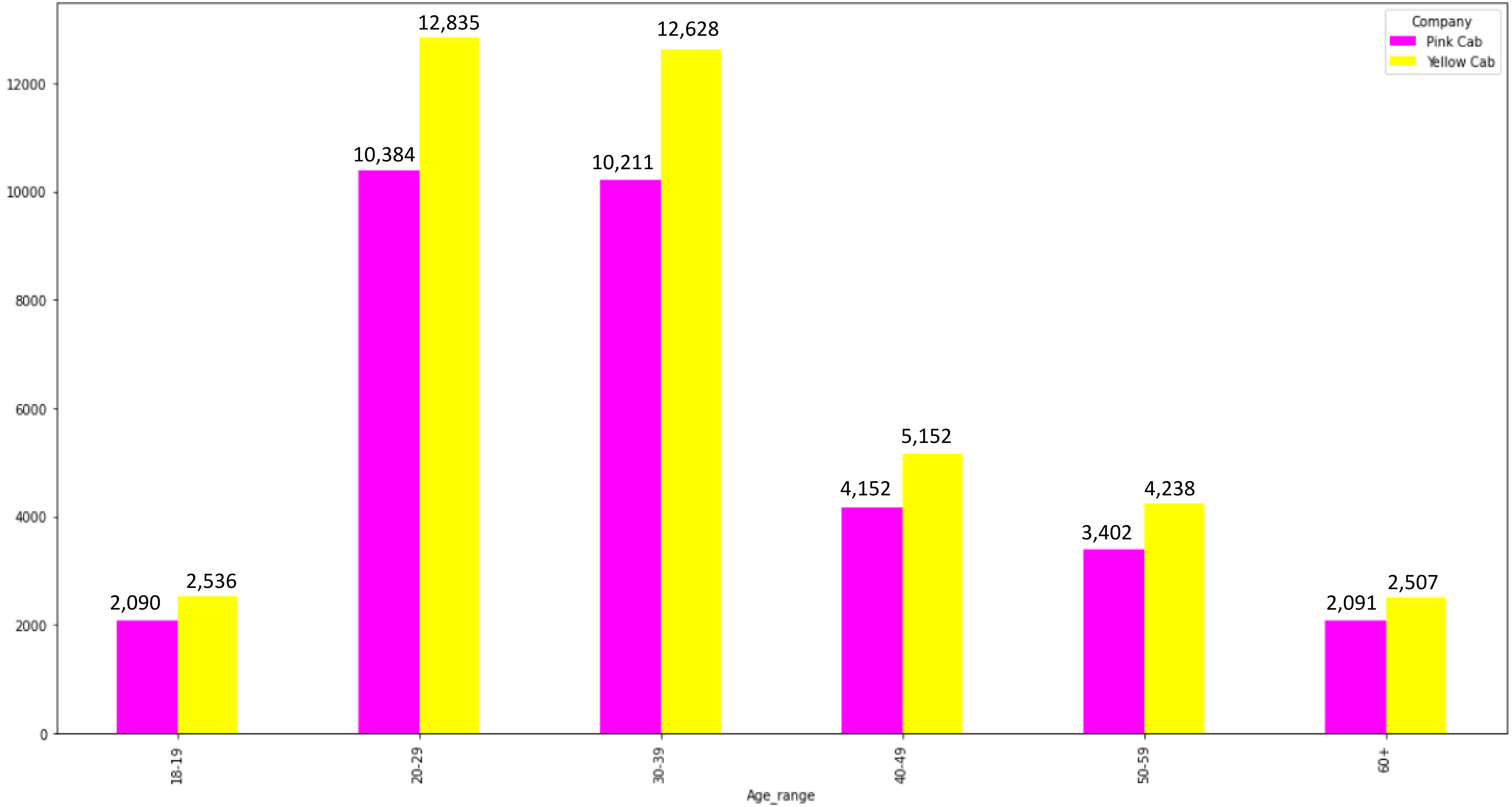
Age_Range	Total Customers
18-19	2,925
20-29	14,853
30-39	14,598
40-49	5,935
50-59	4,899
60+	2,938



Age_Range	Total Transactions
18-19	22,437
20-29	116,430
30-39	112,735
40-49	47,017
50-59	38,087
60+	22,686

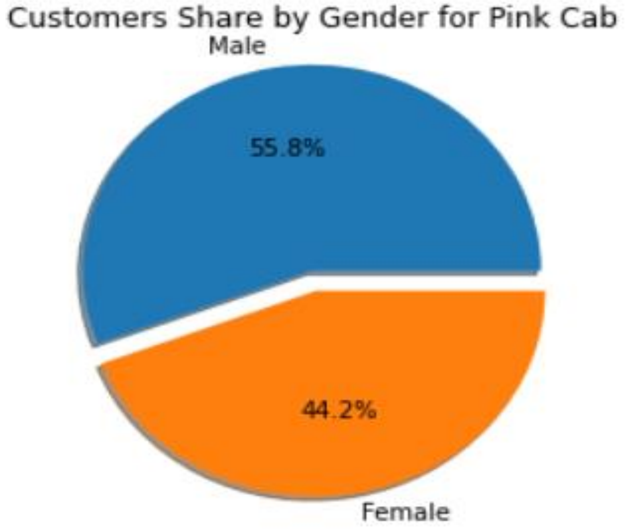
- The 20-29 and 30-39 age groups dominate in terms of total customers and transactions for both cab companies.

EDA and Summary – Age Analysis by Company



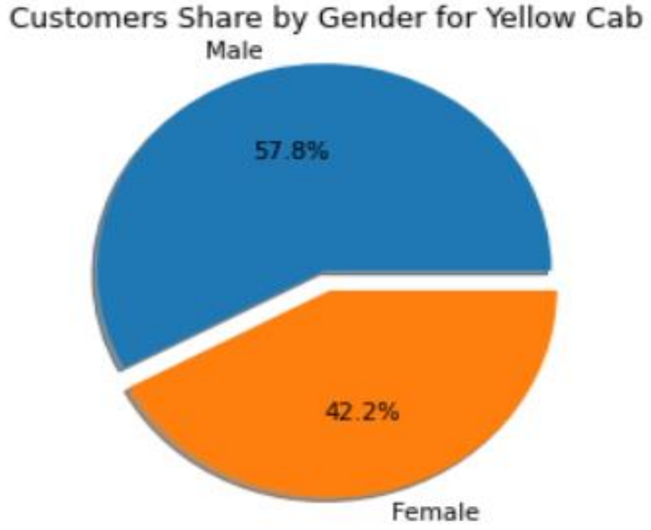
- Yellow Cab has a larger customer base for each age group compared to Pink Cab.

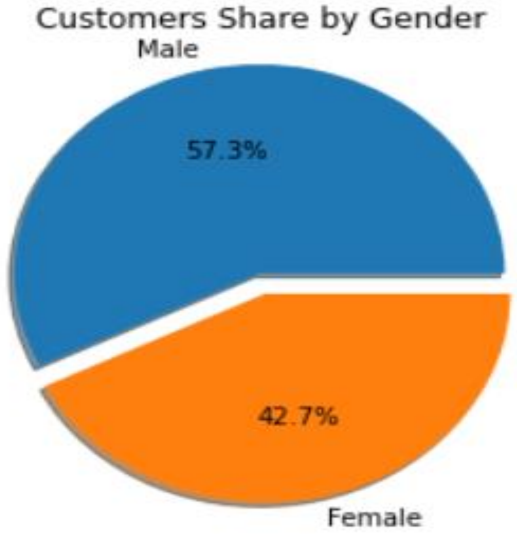
EDA and Summary – Gender Analysis



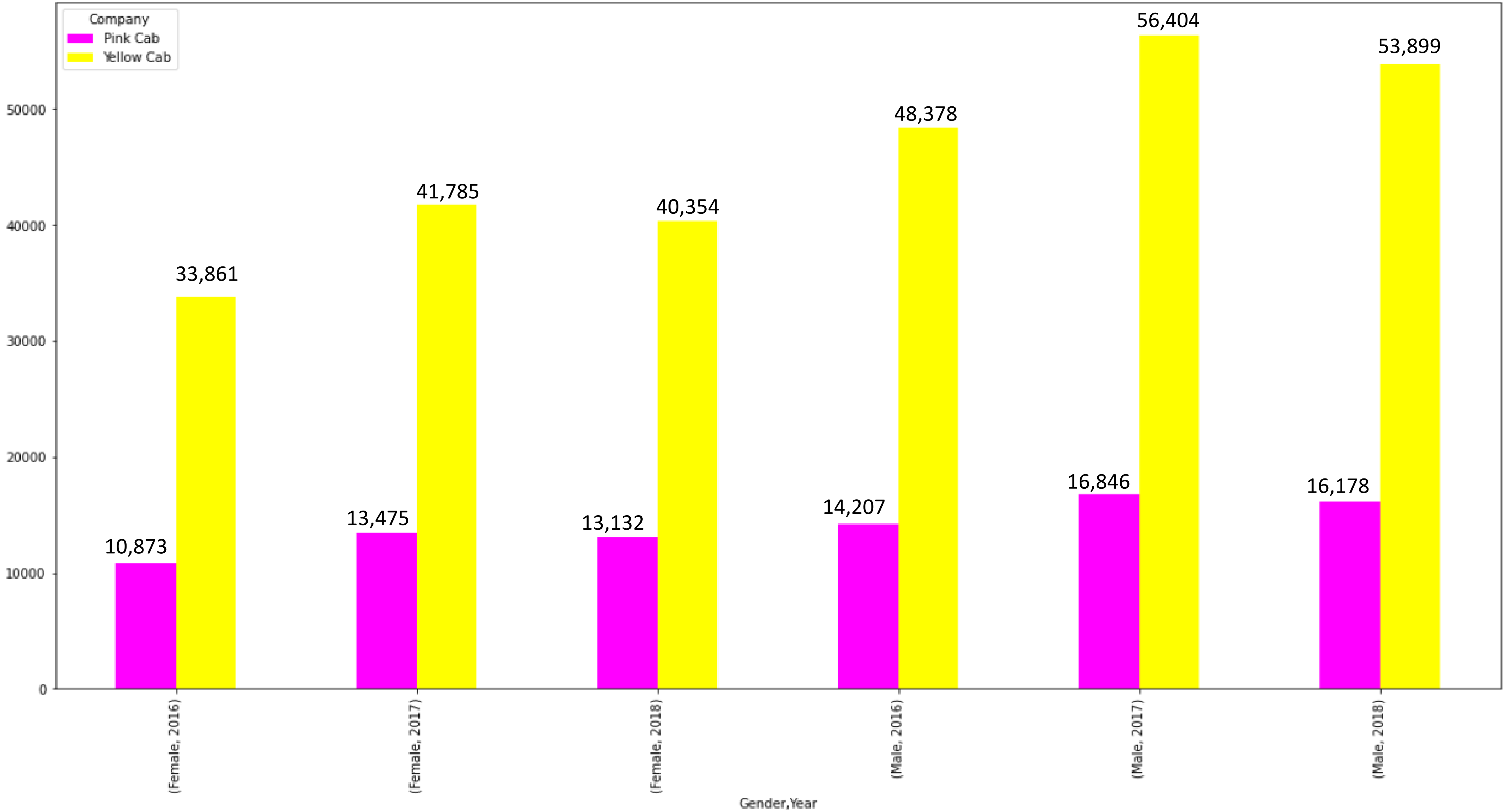
Cab Company	Male	Female
Yellow Cab	21,502	18,394
Pink Cab	17,511	14,819

- Overall the male customers use the cab frequently with Yellow Cab having the most males and females.





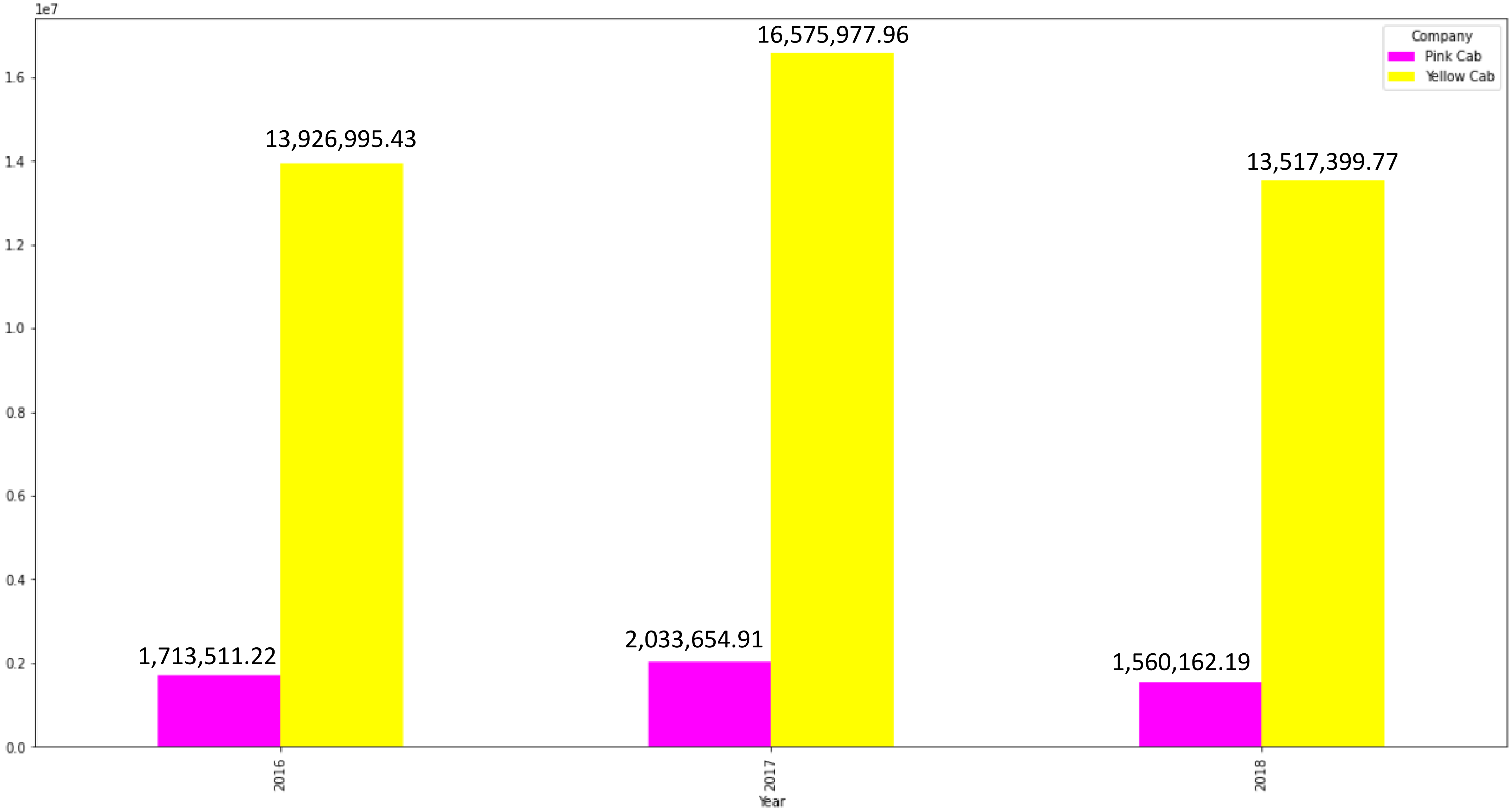
EDA and Summary - Gender and Transaction Analysis



- Yellow Cab has the higher number of transactions for both genders in the consecutive years.

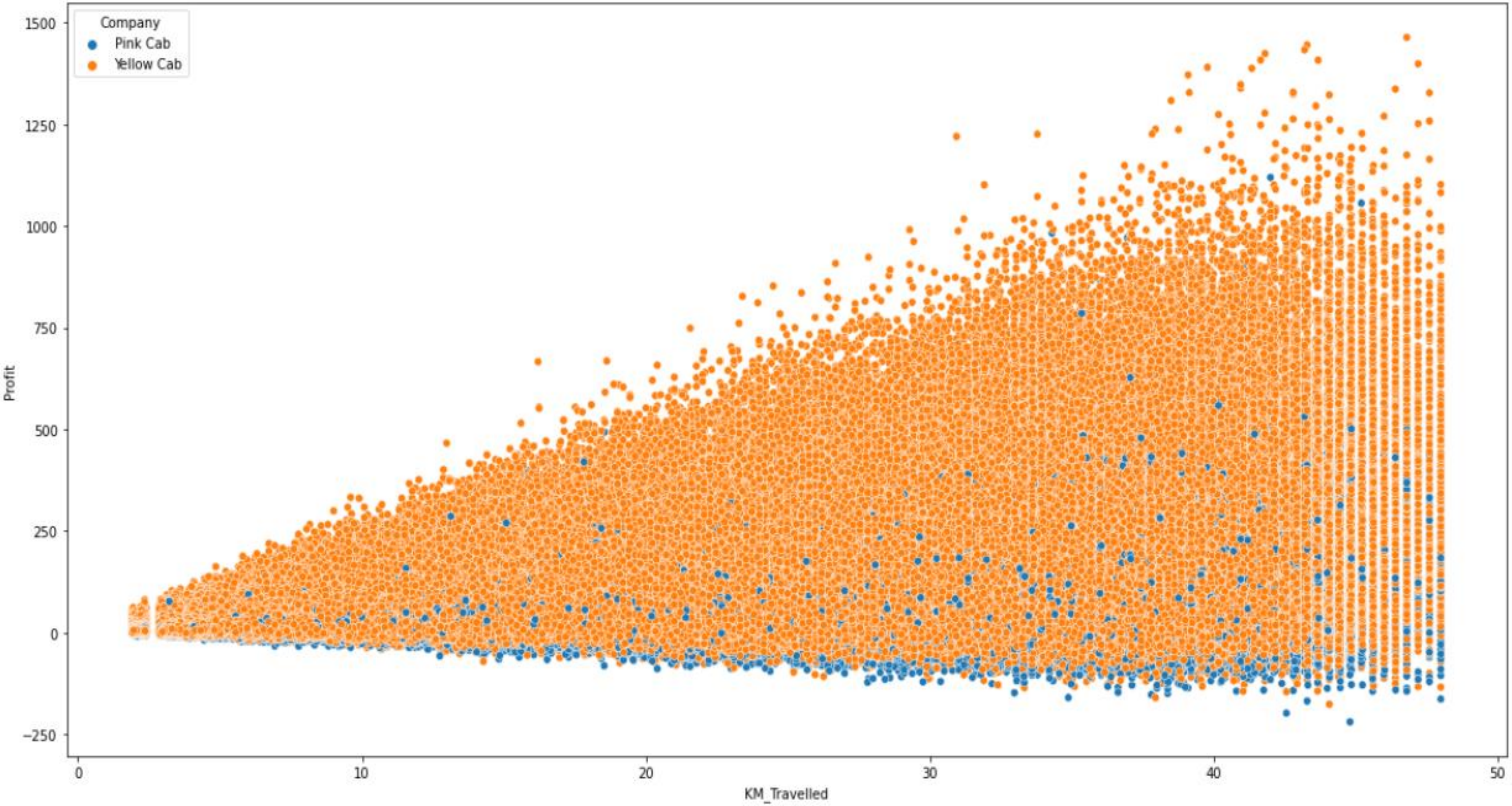
- Both companies experienced a drop in transaction numbers for both genders in 2018.

EDA and Summary – Profit Analysis



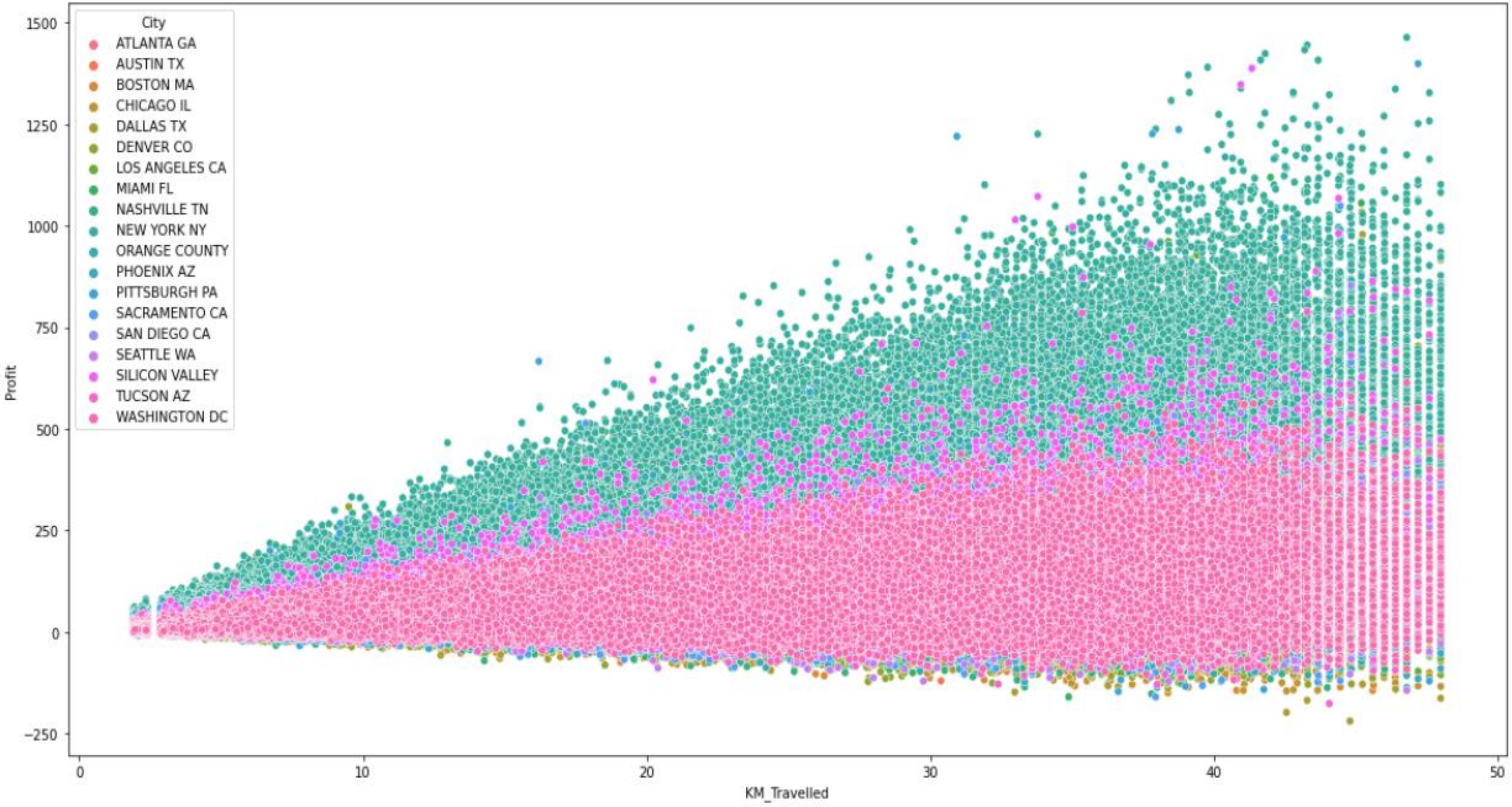
- On a year on year basis Yellow Cab exhibits the highest profits.

EDA and Summary – Company Profit and KM Travelled Analysis



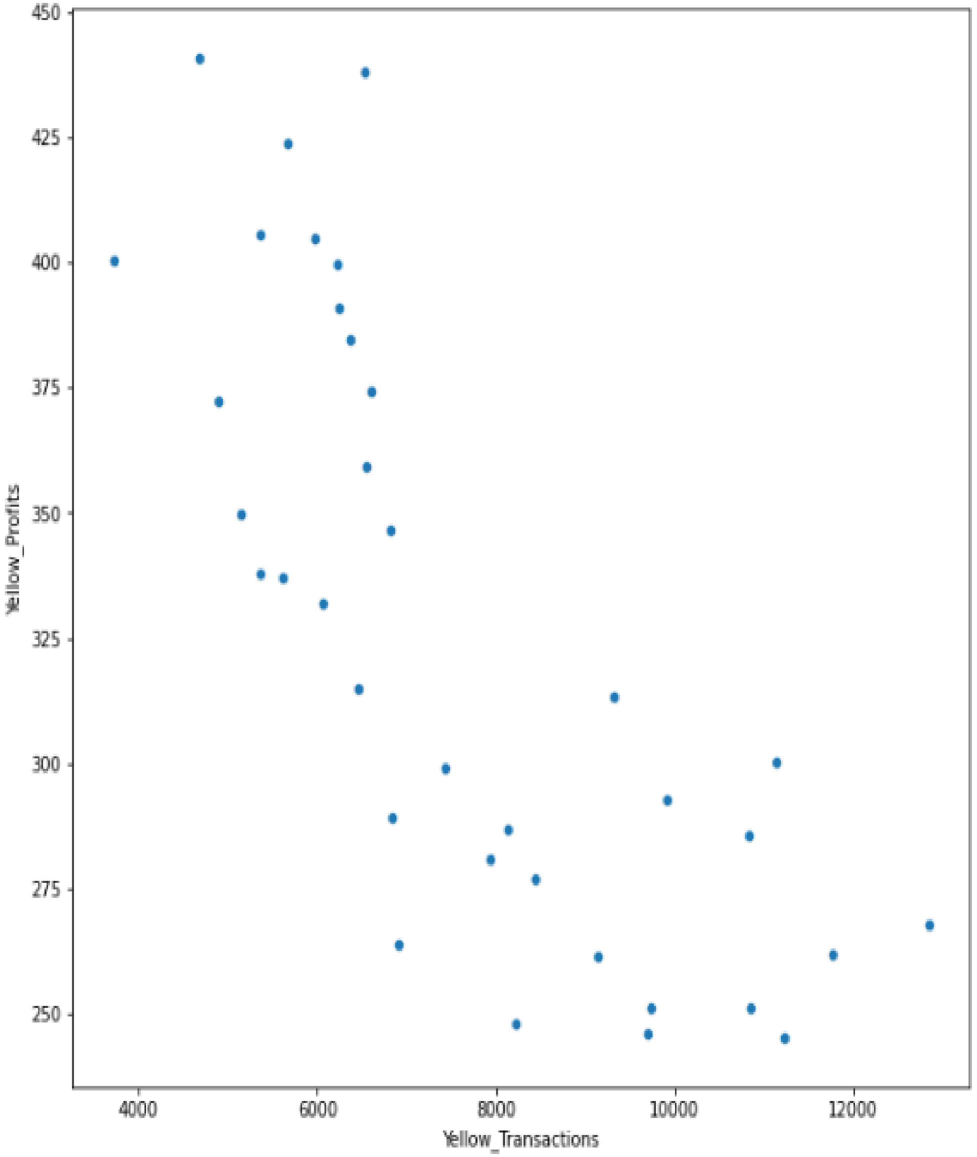
- The more the KM Travelled the higher the profits get.
- Yellow Cab shows the most profits, with a few outlying Pink Cab profits emerging.

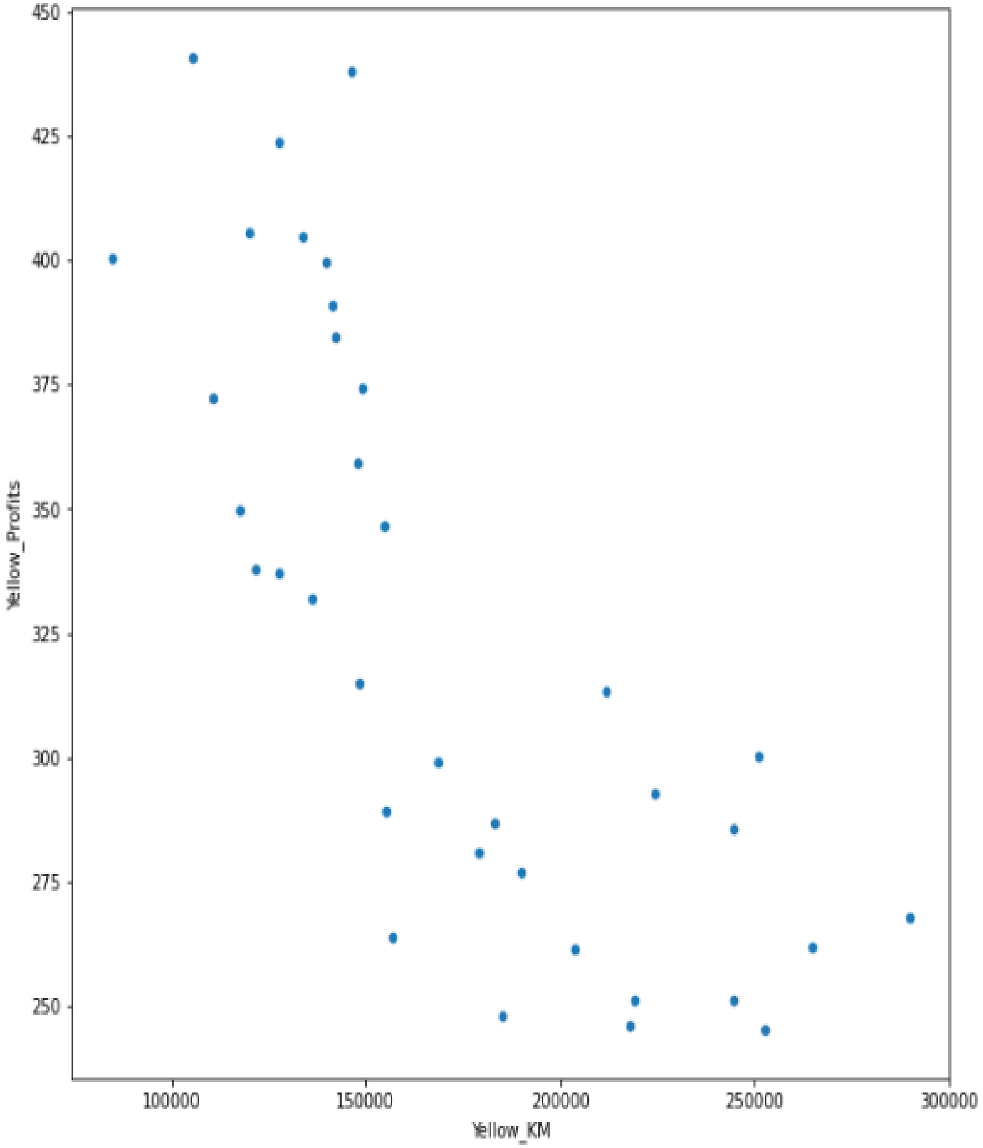
EDA and Summary – City Profit and KM Travelled Analysis



- Highest profits are generated in New York City and Silicon Valley.

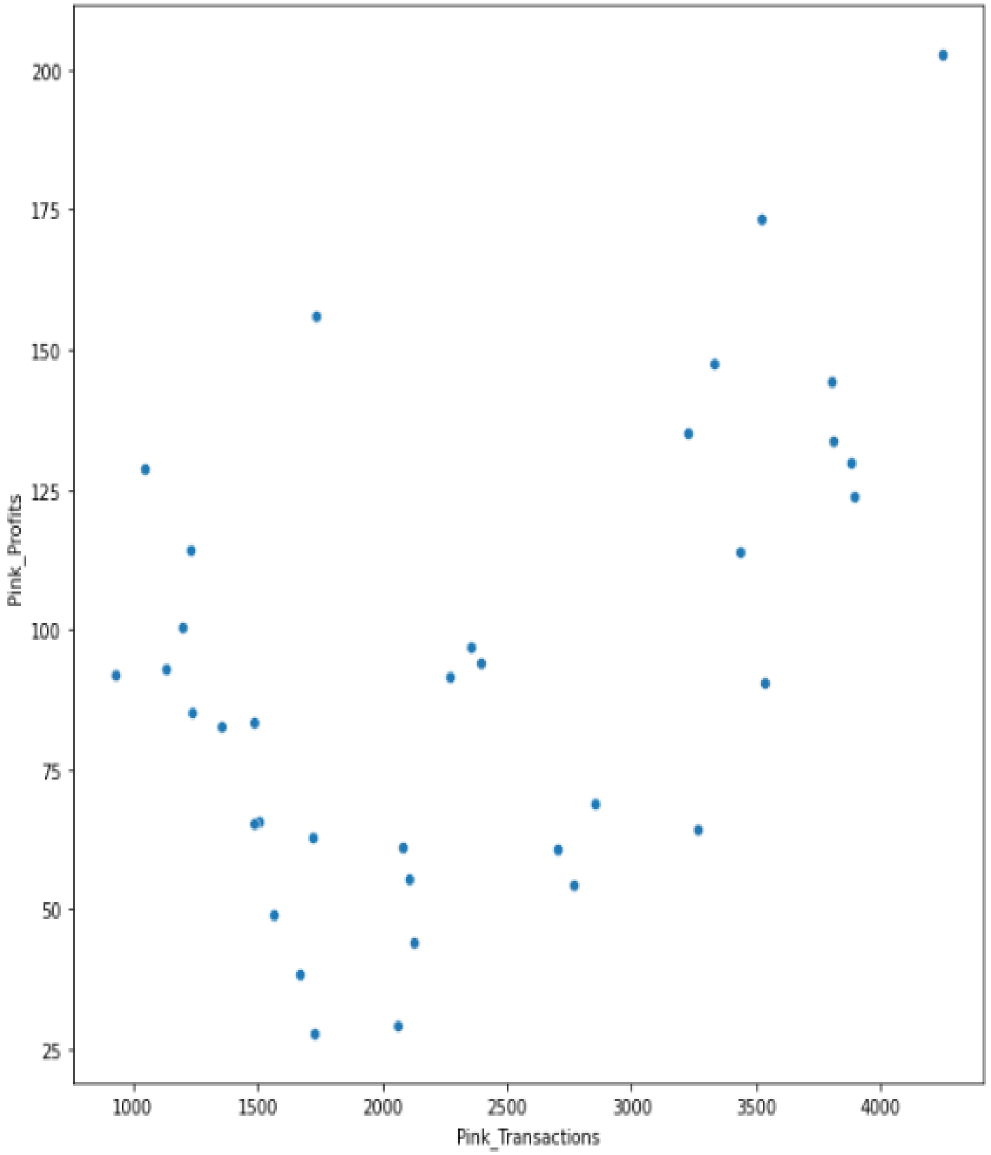
EDA and Summary – Yellow Cab Profit Analysis

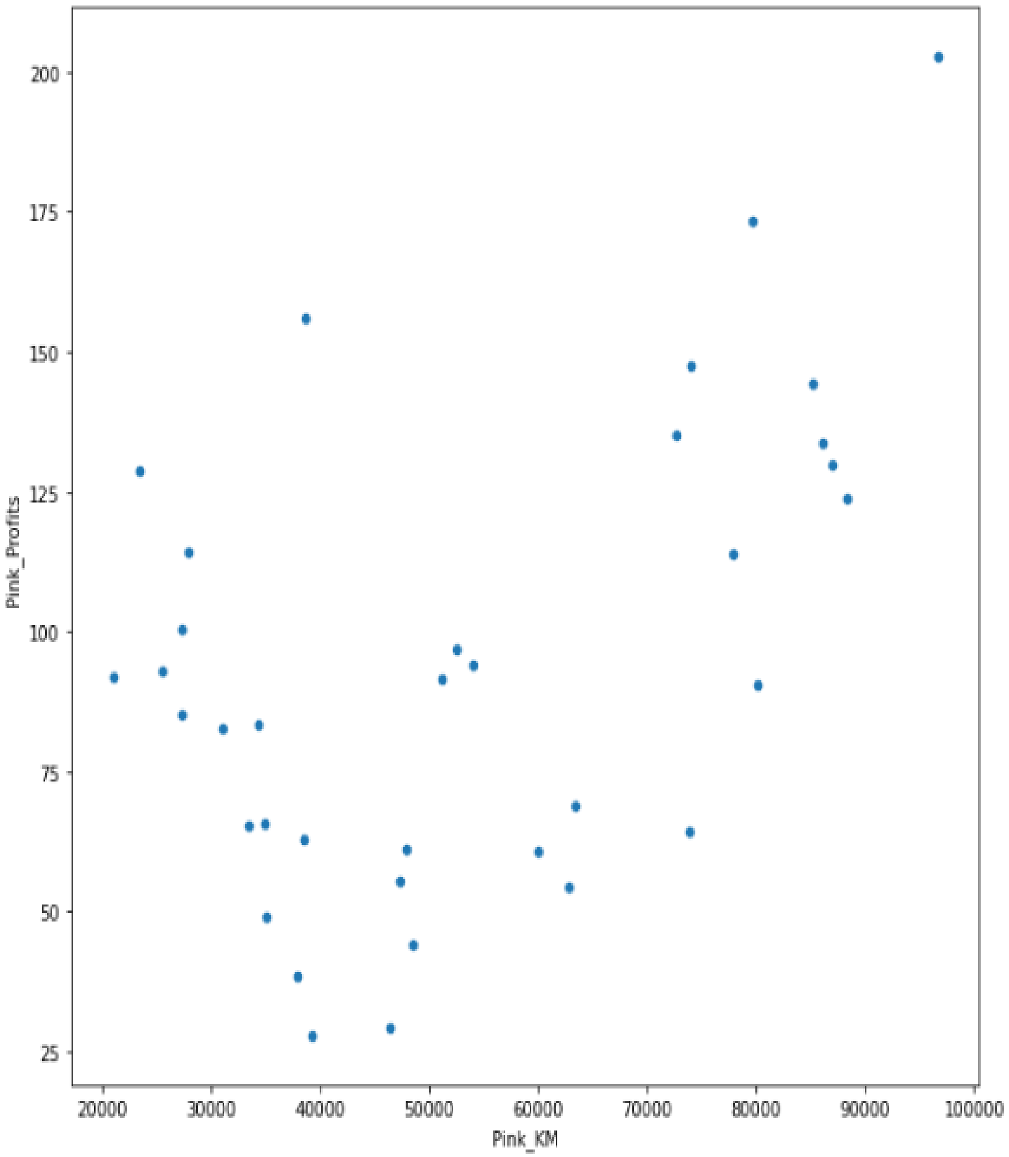




- Sum of profits, sum of transactions and sum of km travelled have a negative correlation.

EDA and Summary - Pink Cab Profit Analysis





- Sum of profits, sum of transactions and sum of km travelled have a positive correlation.

Hypothesis Testing

- **One**

  H0: KMs Travelled and Profit gained are not related. (p = 0)

  H1: KMs Travelled and Profit gained are related. (p != 0)

Conclusion: KMs Travelled and Profit gained are related.

- **Two**

  H0: There is no difference in KM Travelled by Females compared to Males for Yellow Cab.

  H1: There is a difference in KM Travelled by Females compared to Males for Yellow Cab.

Conclusion: There is no difference in KM Travelled by Females compared to Males for Yellow Cab.

- **Three**

  H0: Females bring in less profits than Males for Yellow Cab.

  H1: Females bring in more profits than Males for Yellow Cab.

Conclusion: Females bring in more profits than Males for Yellow Cab.

Hypothesis Testing

- **Four**

  H0: Females bring in less profits than Males for Pink Cab.

  H1: Females bring in more profits than Males for Pink Cab.

Conclusion: Females bring in less profits than Males for Pink Cab.

- **Five**

  H0: The mean Profit for the different Age groups for Yellow Cab are equal.

  H1: One or more of the mean Profits for the different Age groups for Yellow Cab are unequal.

Conclusion: There is a difference in Profit due to Age.

- **Six**

  H0: The mean Profit for the different Age groups for Pink Cab are equal.

  H1: One or more of the mean Profits for the different Age groups for Pink Cab are unequal.

Conclusion: There is no difference in Profit due to Age.

Recommendations

After an evaluation of both companies Yellow Cab was found to be better than Pink Cab based on the following points:

- **Transaction Analysis**: Yellow Cab has domination over the market with total transactions processed three times those for Pink Cab over the 3 years.

- **Customer Share**: From 2016 - 2018 Yellow Cab had a customer reach of more than 76% though there was a drop of 0.3% which can be considered insignificant.

- **Age Wise Reach**: For each age group Yellow Cab has larger numbers of customers with both companies having the most customers in the 20 - 29 and 30 - 39 age groups.

- **Gender Aspect**: Yellow Cab has the highest number of transactions for both male and female customers from 2016 – 2018.

- **Profit Wise**: Yellow Cab exhibits profits twelve times those for Pink Cab year on year. Also the more KM are travelled the higher the profits get.

For Yellow Cab there is a difference in profit per age group, which allows for versatility in campaigns that they may launch to increase profits.

We recommend Yellow Cab for investment.

Thank You

Data Glacier

Your Deep Learning Partner