Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

G2M insight for Cab Investment firm

16 May, 2022

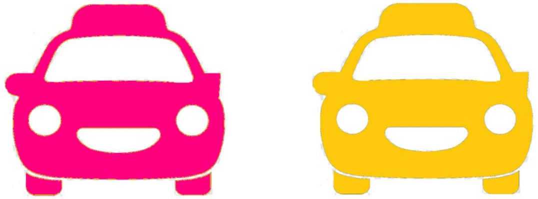
Agenda

- Problem statement
- Data exploration
- EDA
- EDA Summary and recommendations

Problem statement

Conduct analysis of Cab market data to help XYZ take a right investment decision.

Options for investment:
- Yellow Cab company
- Pink Cab company.



Important factors for companies in Cab industry:
maximize profit (especially profit per km because each additional km is connected with additiona
costs (fuel, amortization, etc.).
increase number of customers, particularly number of loyal customers
increase presence in all income, age and gender groups
expand geographically
maintain good results for long time.

To take the right decision, company performance analysis should be conducted. Further EDA chapters will contain analysis of these factors. However, first step of analysis is data exploration.

Data exploration

Data sets provided:

Cab_Data.csv	City.csv	Customer_ID.csv	Transactions.csv
Info about trip details	Info about cities	Info about customers’ feature	Info about payment details

The data sets are merged into one for further analysis.

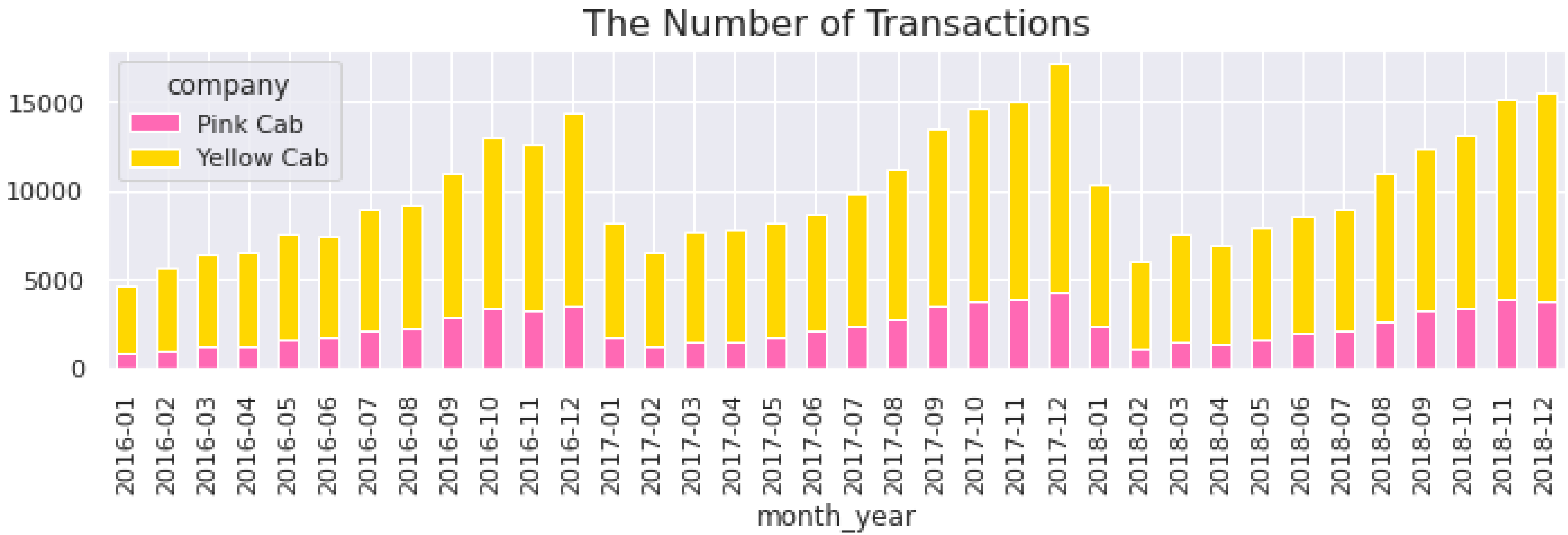
Data summary:

- no null values, no duplicates
- some columns need data type transformation
- outliers are present only in price_charged. This parameter depends on distance and time.

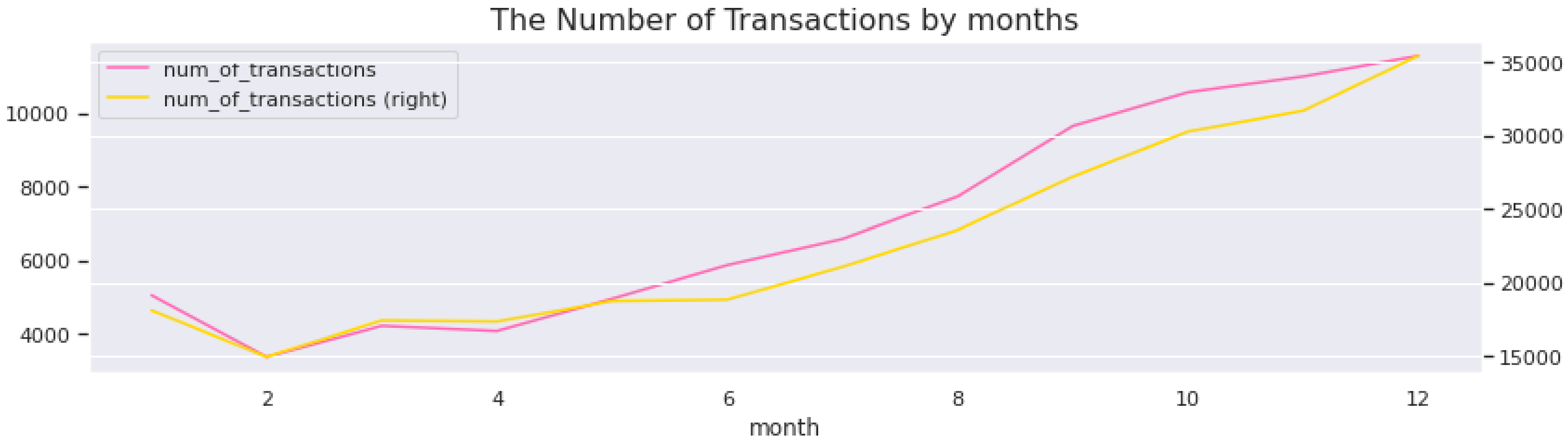
Since there’s no time information, we don’t remove outliers.

Additional assumptions:
we assume that users in city data set is total number of users, including Pink and Yellow Cab users

Transactions



- There is seasonality in number of transactions. Minimum is in February, maximum is in December.
- Maximum number was in December 2017, 2018 is comparable to 2016.

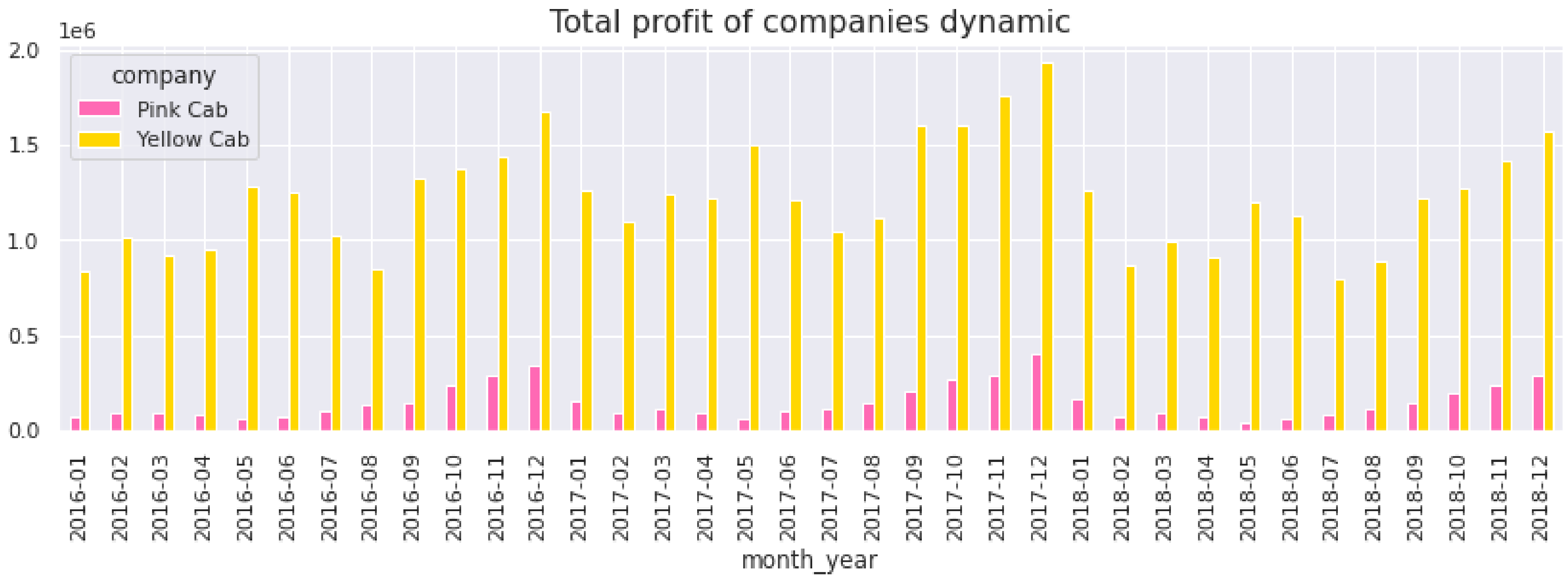


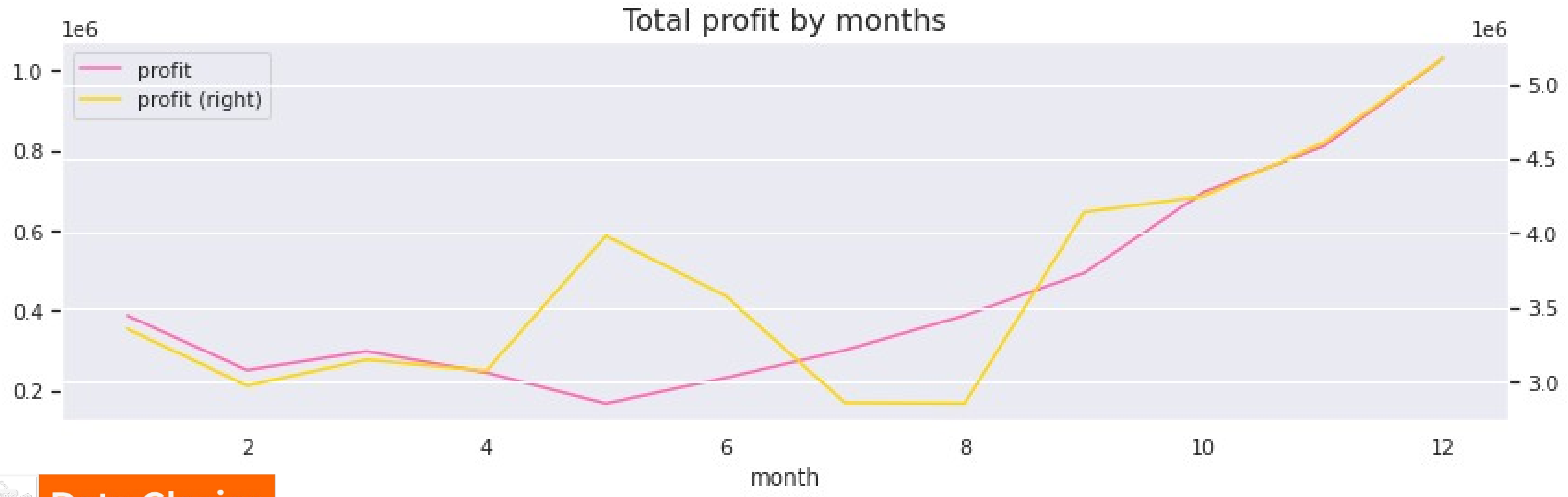
- Number of transactions grows smoothly by the end of a year.
- Number of transactions in Yellow Cab significantly exceeds number in Pink Cab but trend is the same.

Data Glacier

Your Deep Learning Partner

Profit analysis


Total profit of companies dynamic


Total profit by months

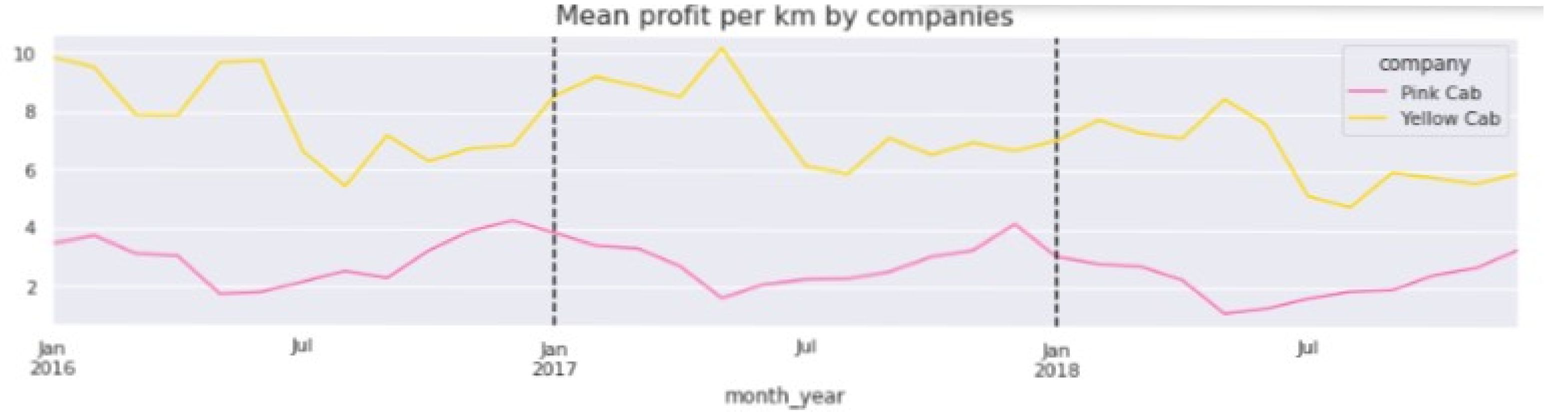
- Profit is calculated as price minus cost of the trip.
- There is also fluctuations in total profit but pattern is different from number of transactions and customers dynamic, especially in Yellow Cab.
- Total profit of Yellow Cab by months significantly exceeds total profit of Pink Cab.
- There's a light decreasing trend: in December 2018 total profit is 1,18 times less compared to December 2017.

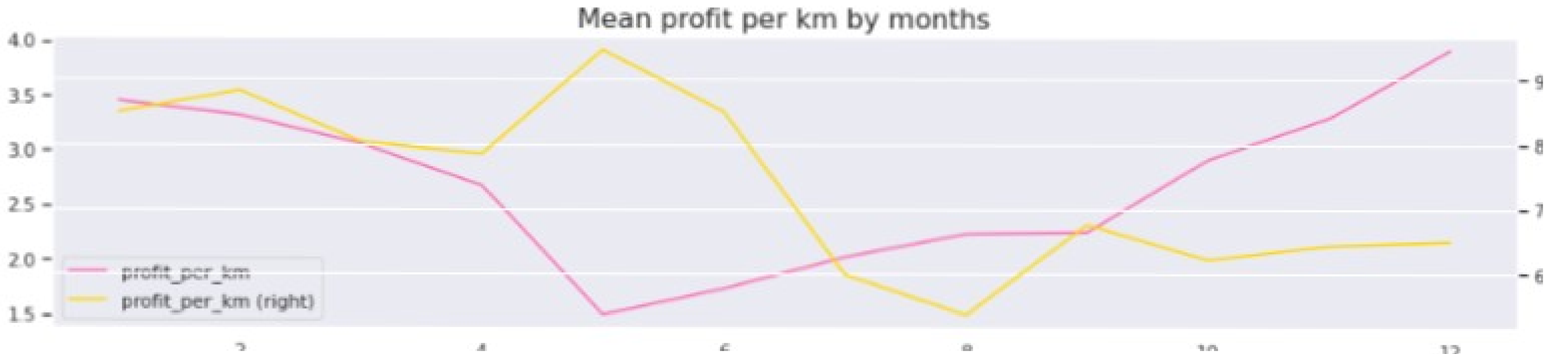
- Sum of total profit of Yellow Cab for analyzed period – more than 8 times compared to Pink Cab, mean total profit – more than 2 times.
- Dynamic of total profit by months are different for 2 companies, especially in 2 and 3 quarter. Profit of Yellow Cab fluctuates more.

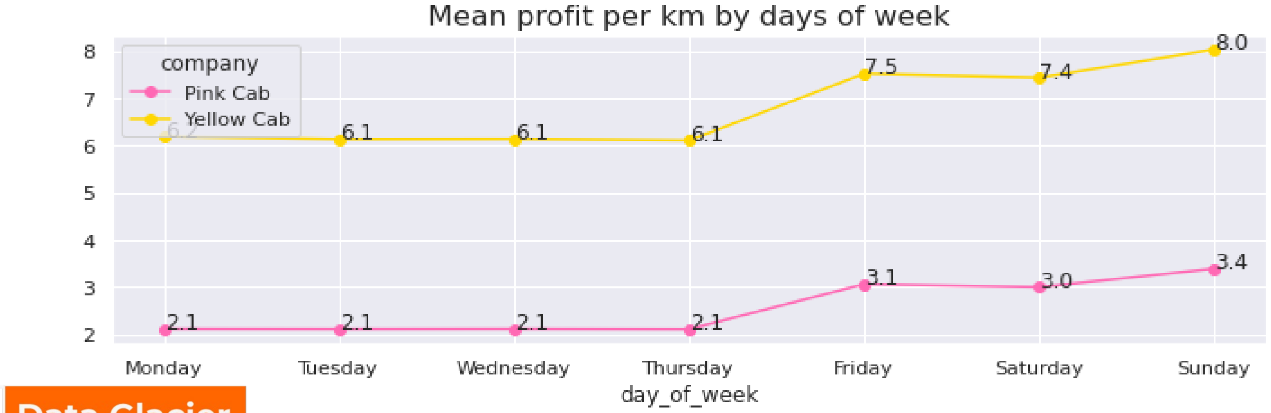
Data Glacier

Your Deep Learning Partner

Profit per km analysis


Mean profit per km by companies


Mean profit per km by months


Mean profit per km by days of week

- Since profit and distance of a trip may vary, profit per km is more representative metric.
- Fluctuations of profit per km during analyzed period are more often and different between companies. E.g. in May Yellow Cab has maximum profit per km while Pink Cab has minimal value in this month.
- During a week value smoothly grows starting from Thursday and reaches maximum on Sundays.

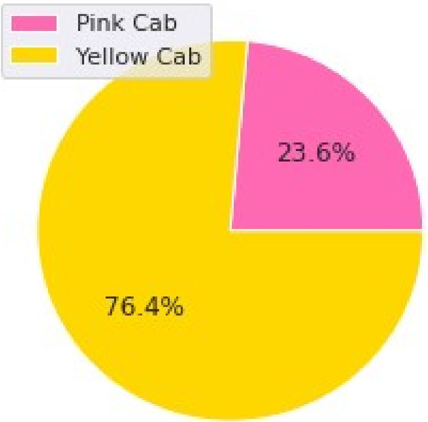
company	Yellow Cab	Pink Cab
profit_per_km	7.11	2.77

Mean profit per km of Yellow Cab is almost 3 times higher than mean value of Pink Cab.

Data Glacier

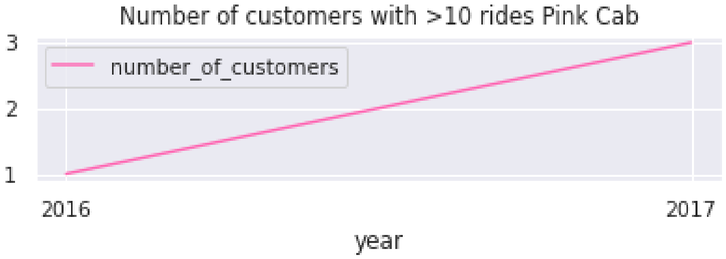
Your Deep Learning Partner

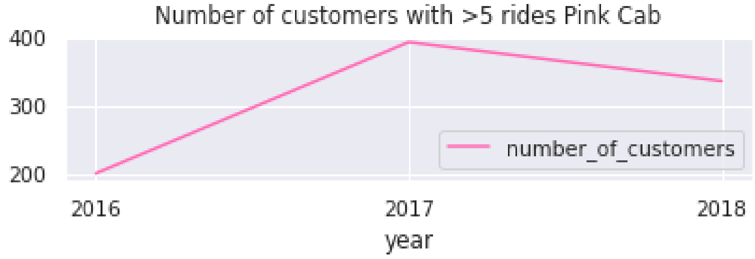
Customer analysis

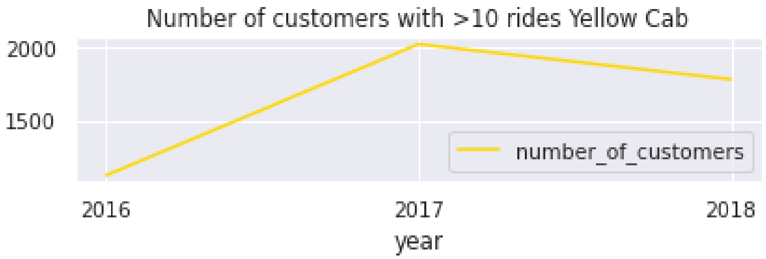


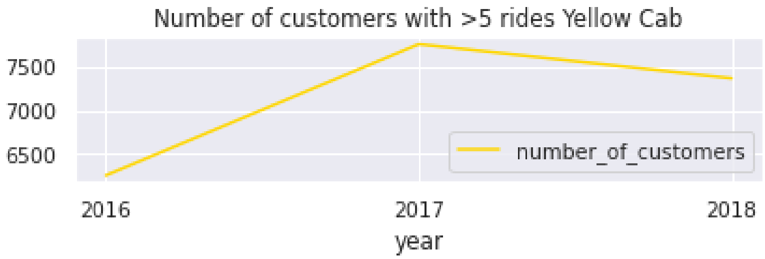
Ratio between number of customers of 2 companies: Yellow Cab has 76,4 percent, Pink Cab – 23,6 percent. Thus, if we consider market as only 2 players, Yellow Cab has ¾ of the customers.

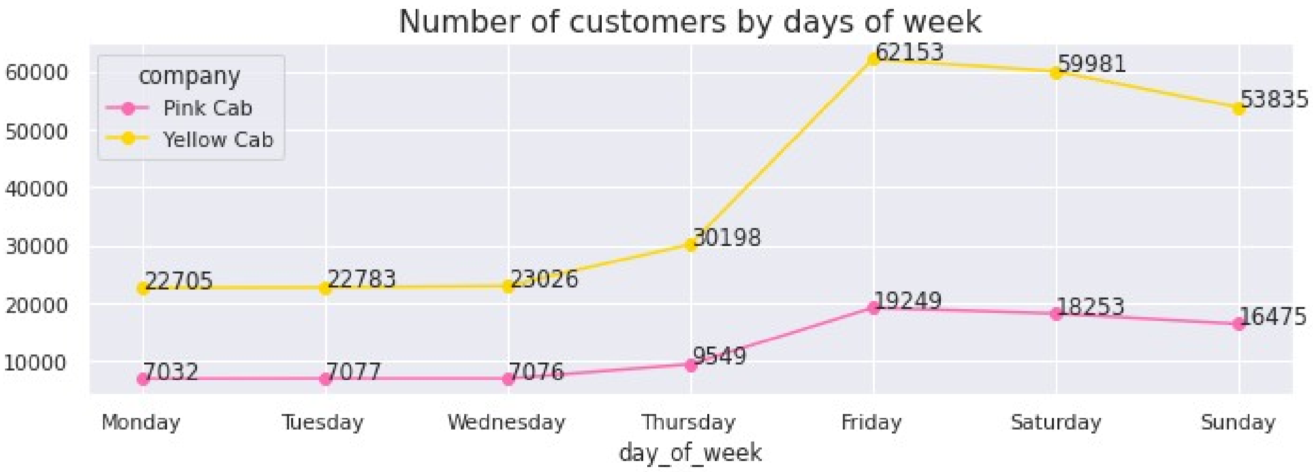
Yellow Cab has higher customer retention. It has many times higher number of customers which had more than 5 and more than 10 trips using one company.











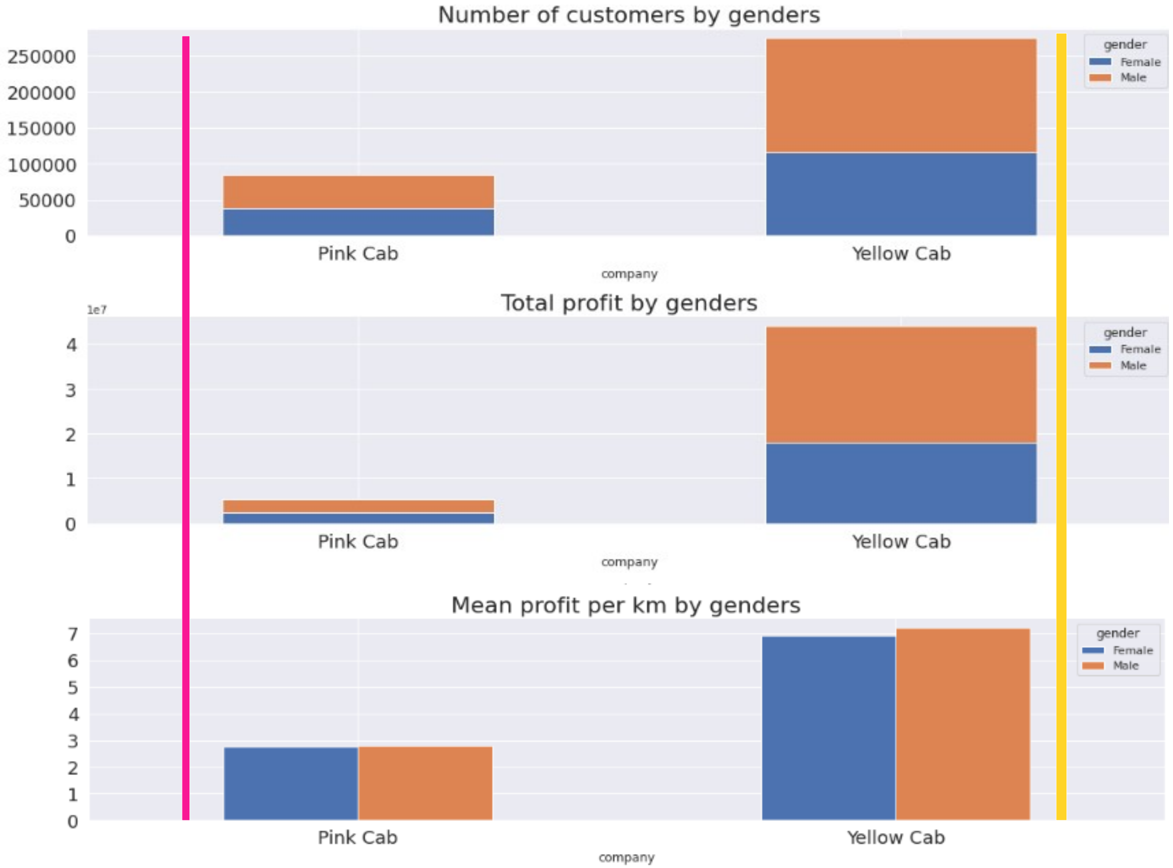
Dynamic of number of customers by months is pretty similar to number of transactions dynamic.

Weekly dynamic differs from dynamic of profit per km because maximum number of customers is on Fridays.

Data Glacier

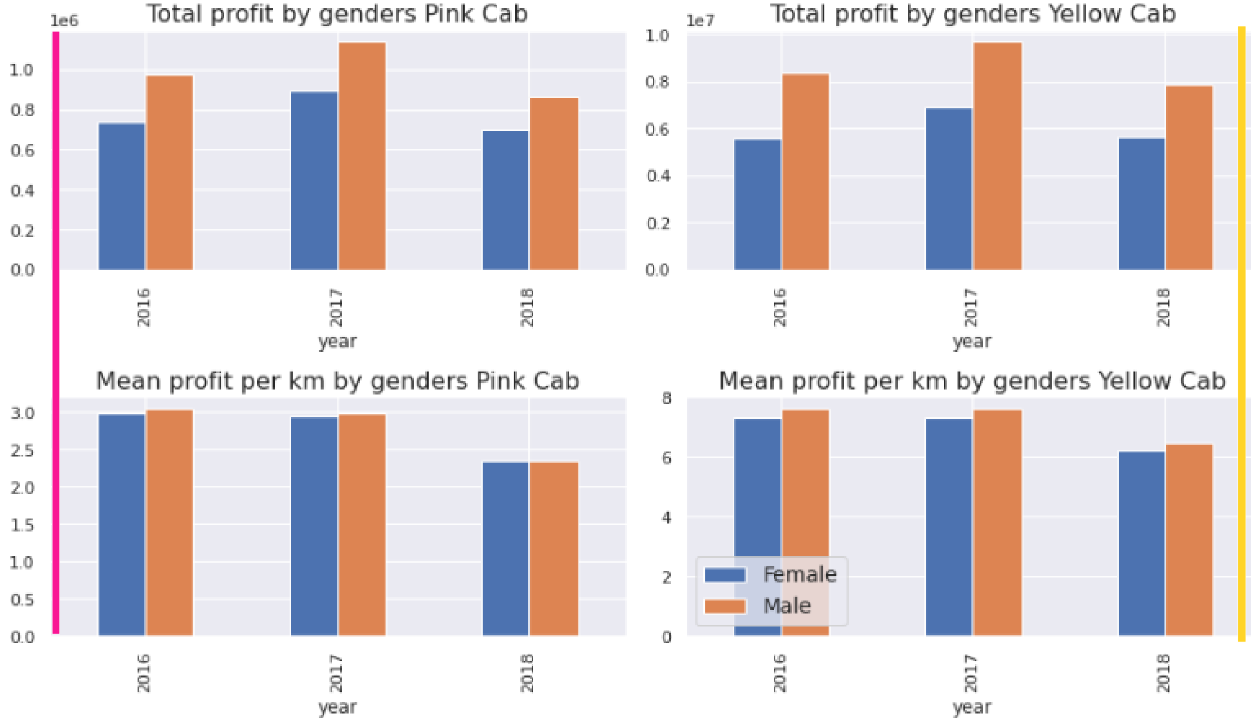
Your Deep Learning Partner

Customer analysis - gender



	customers number		profit		profit_per_km	
	Female	Male	Female	Male	Female	Male
Pink Cab	37480	47231	2,33M	2,98M	2.75	2.79
Yellow Cab	116000	158681	18,13M	25,89M	6.94	7.23
Comparison	Male 1.34 > Female		Male 1.41 > Female		Male 1.03 > Female	

The tendency doesn't change during the analyzed period.

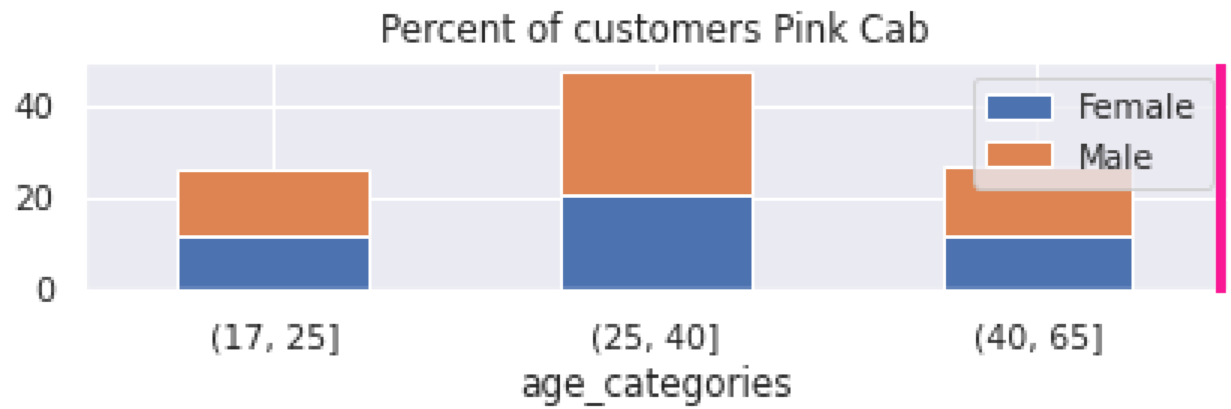


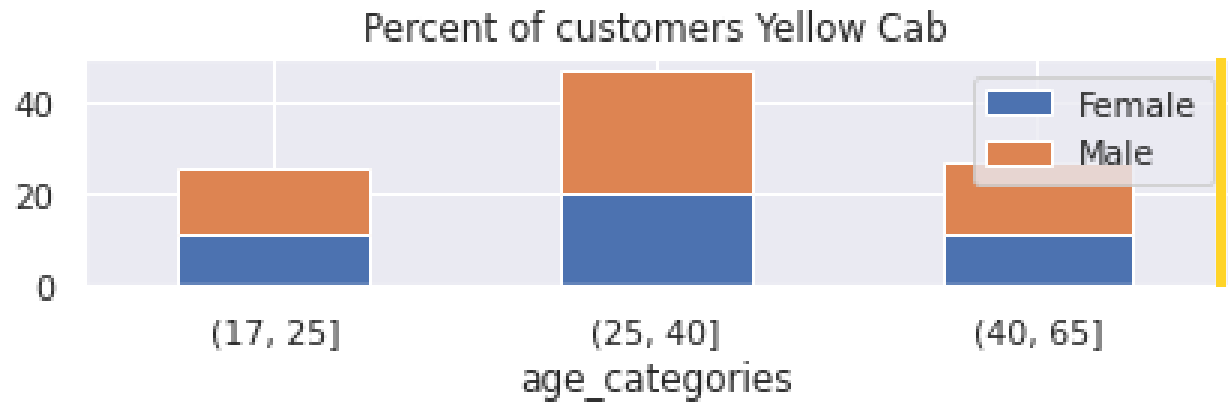
Distribution of customers and total profit between 2 genders is unequal – share of women-customers is less as well as their contribution to total profit. Visually, mean profit per km is a little bit less for women in Yellow Cab, although it also less in Pink Cab.

Data Glacier

Your Deep Learning Partner

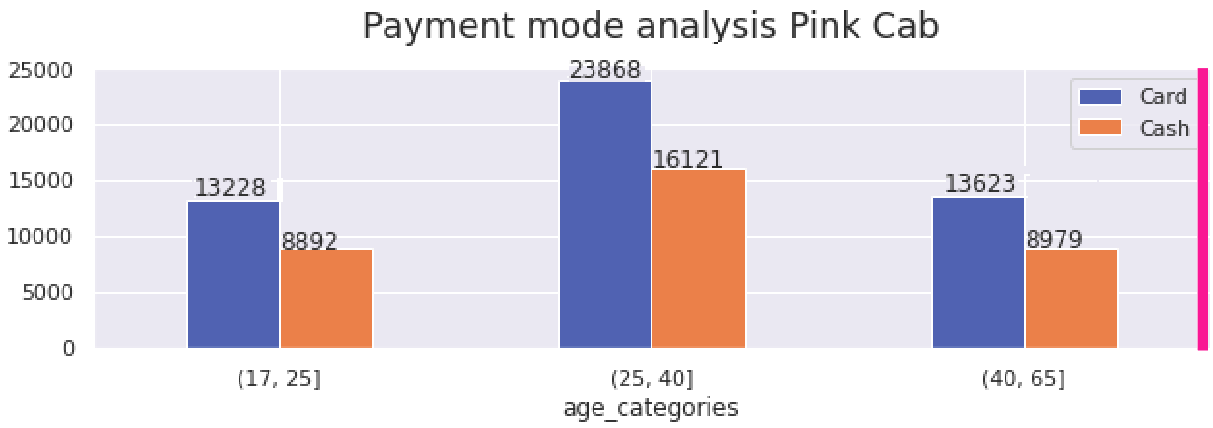
Customer analysis - age

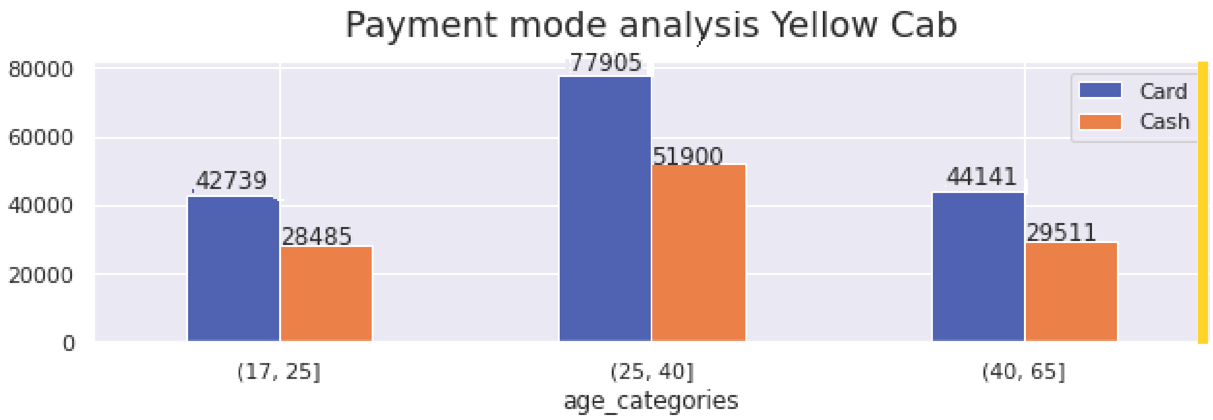

Percent of customers Pink Cab


Percent of customers Yellow Cab

In all age groups payment by card is more popular, in average in 1,5 times.

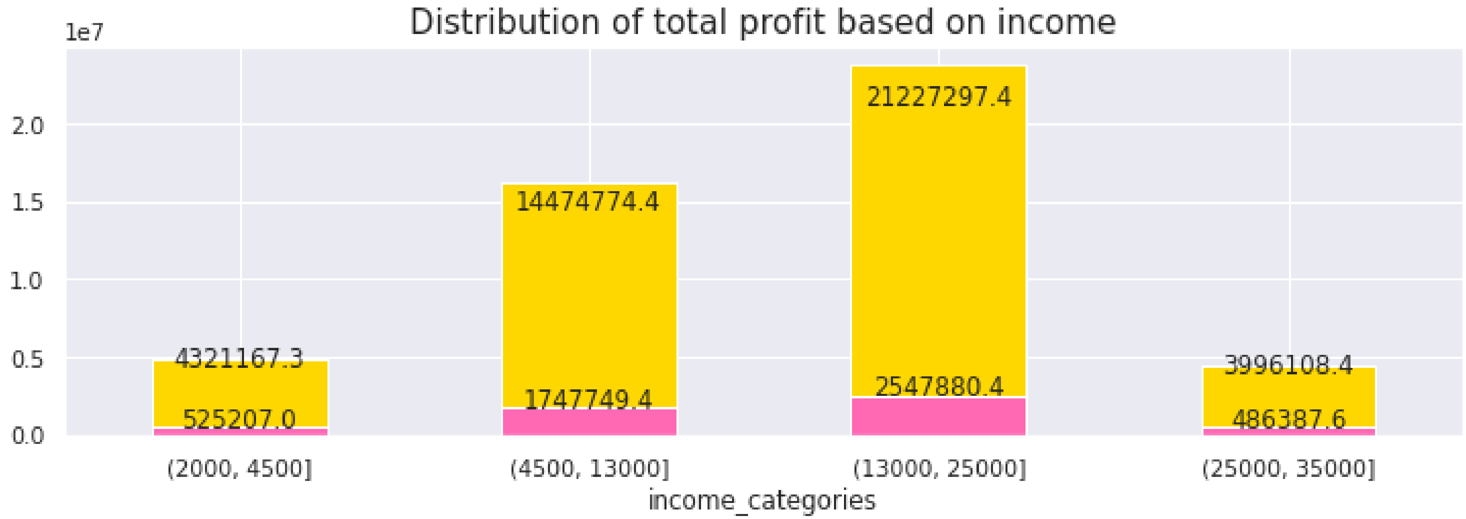
Across 3 age categories, almost a half of customers belongs to middle-aged customers. Shares of customers in 18-25 and 41-65 age groups are equal. We can see that in all age groups women's share is less.


Payment mode analysis Pink Cab

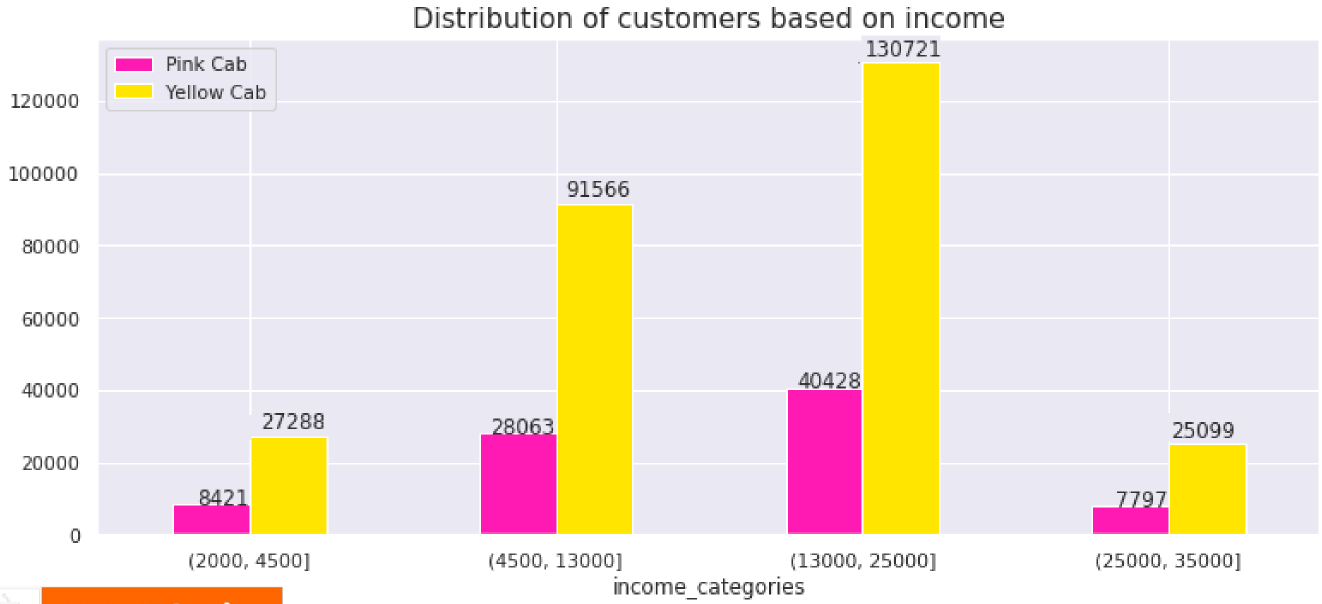

Payment mode analysis Yellow Cab

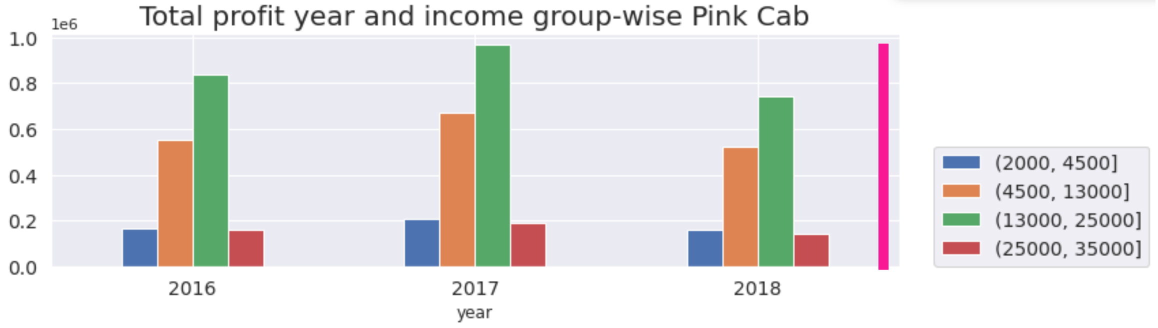
Data Glacier
Your Deep Learning Partner

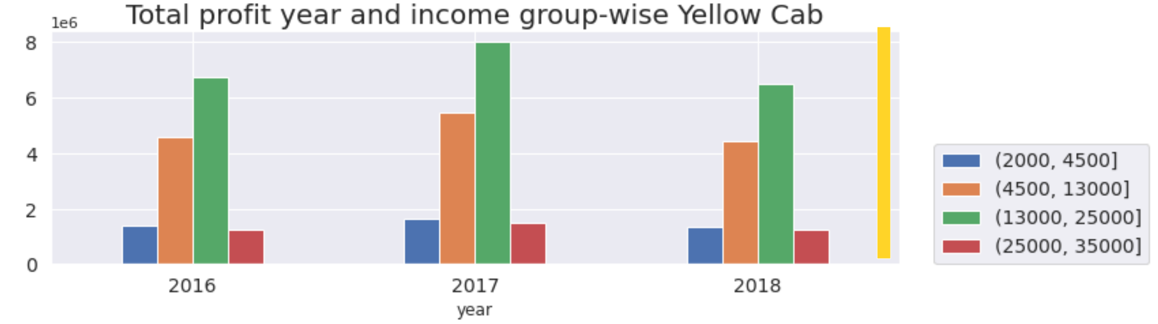
Customer analysis - income


Distribution of total profit based on income

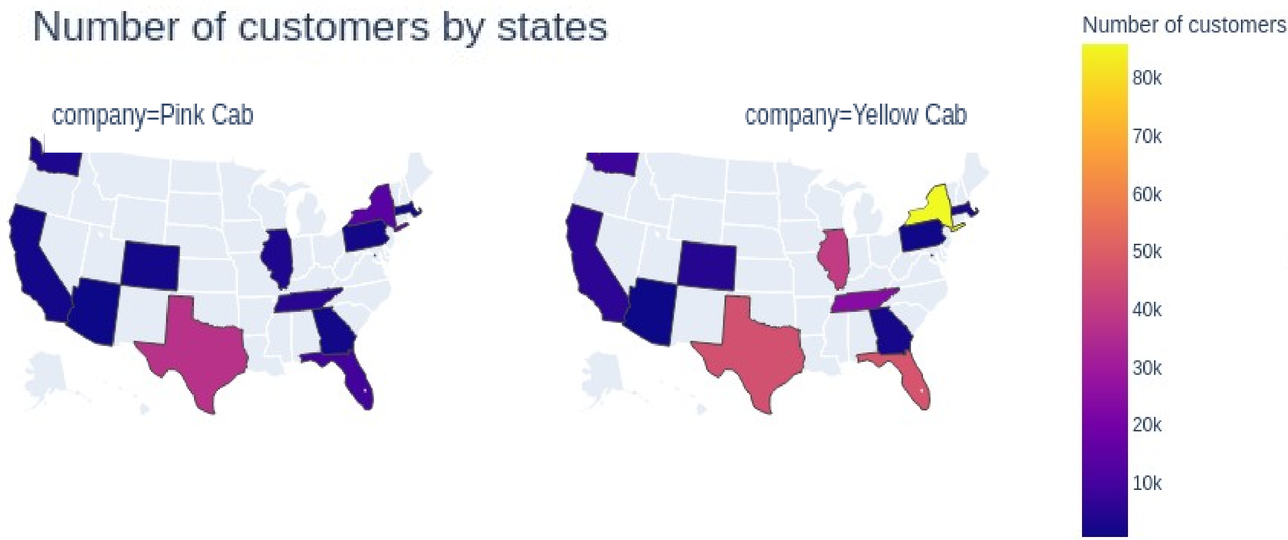
- The biggest share of total profit as well as number of customers belongs to higher-medium income group (13000,25000). The second place belongs to lower-medium group (4500,13000). Lower and higher-income groups have equal shares.
- Mean profit per km is almost equal for all the groups.
- The ratio between income groups is stable during the analyzed period

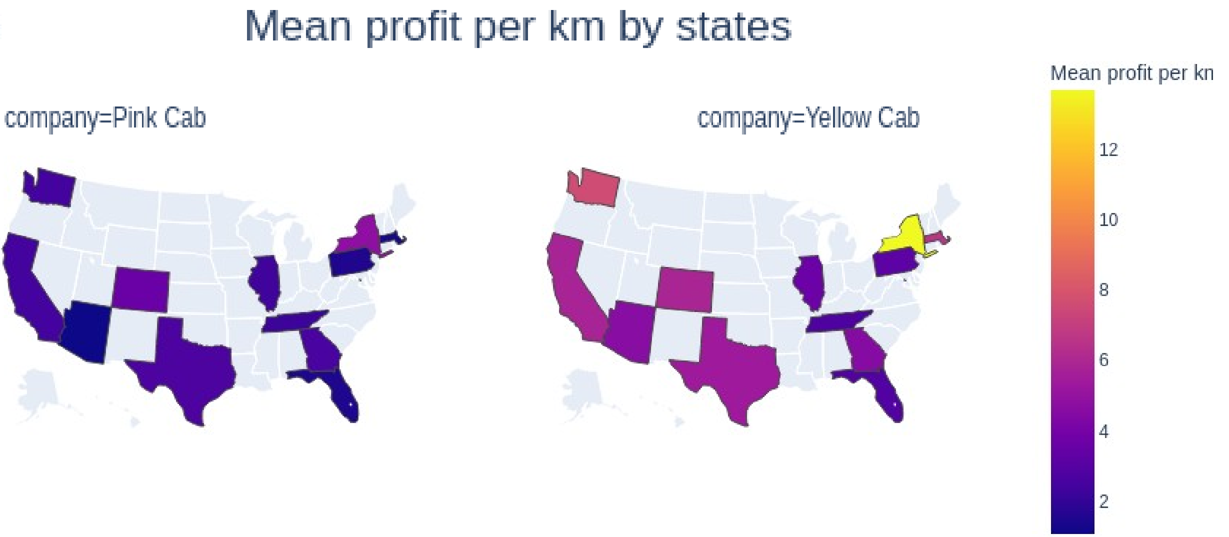

Distribution of customers based on income

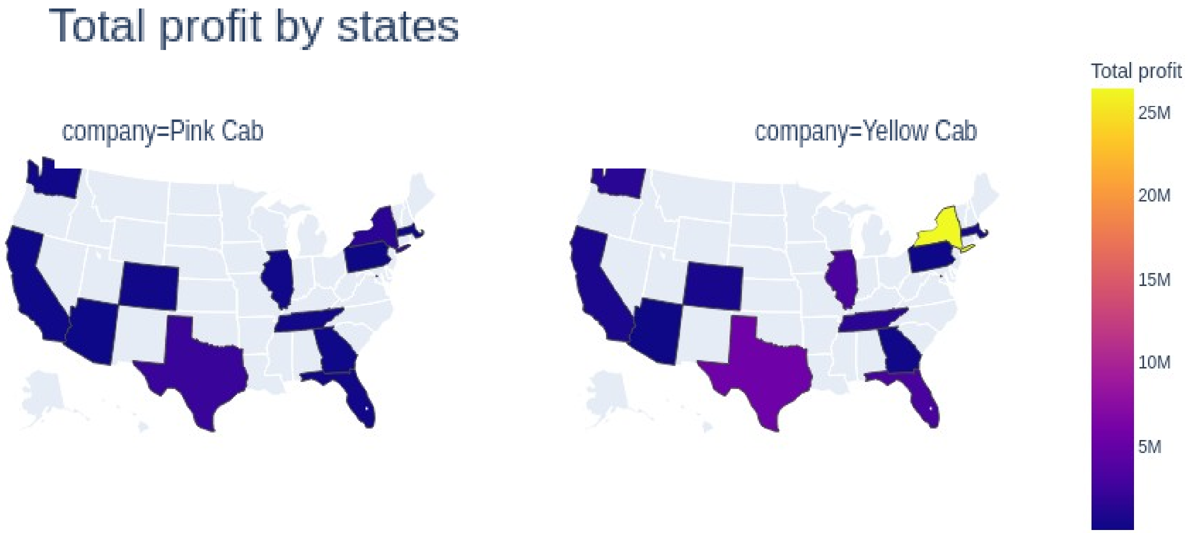

Total profit year and income group-wise Pink Cab


Total profit year and income group-wise Yellow Cab

Region analysis


Number of customers by states


Mean profit per km by states


Total profit by states

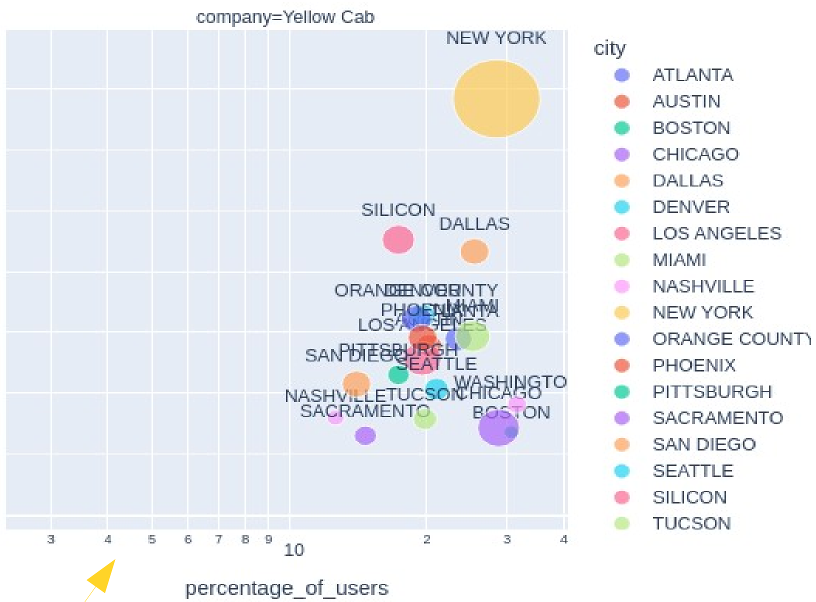
- NY state is the most profitable for Yellow Cab company. Both companies have the highest profit per km in this state. For Pink Cab, maximum number of customers is in Texas state. However, profits distribution is quite equal across the states for Pink Cab.

- It’s predictable that in average Yellow Cab has higher profit per km across states than Pink Cab.

.

Region analysis

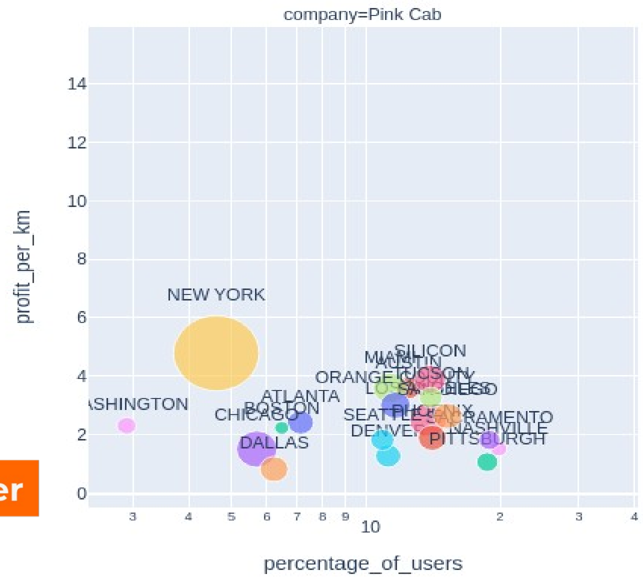
Potential of developing cab service in different cities

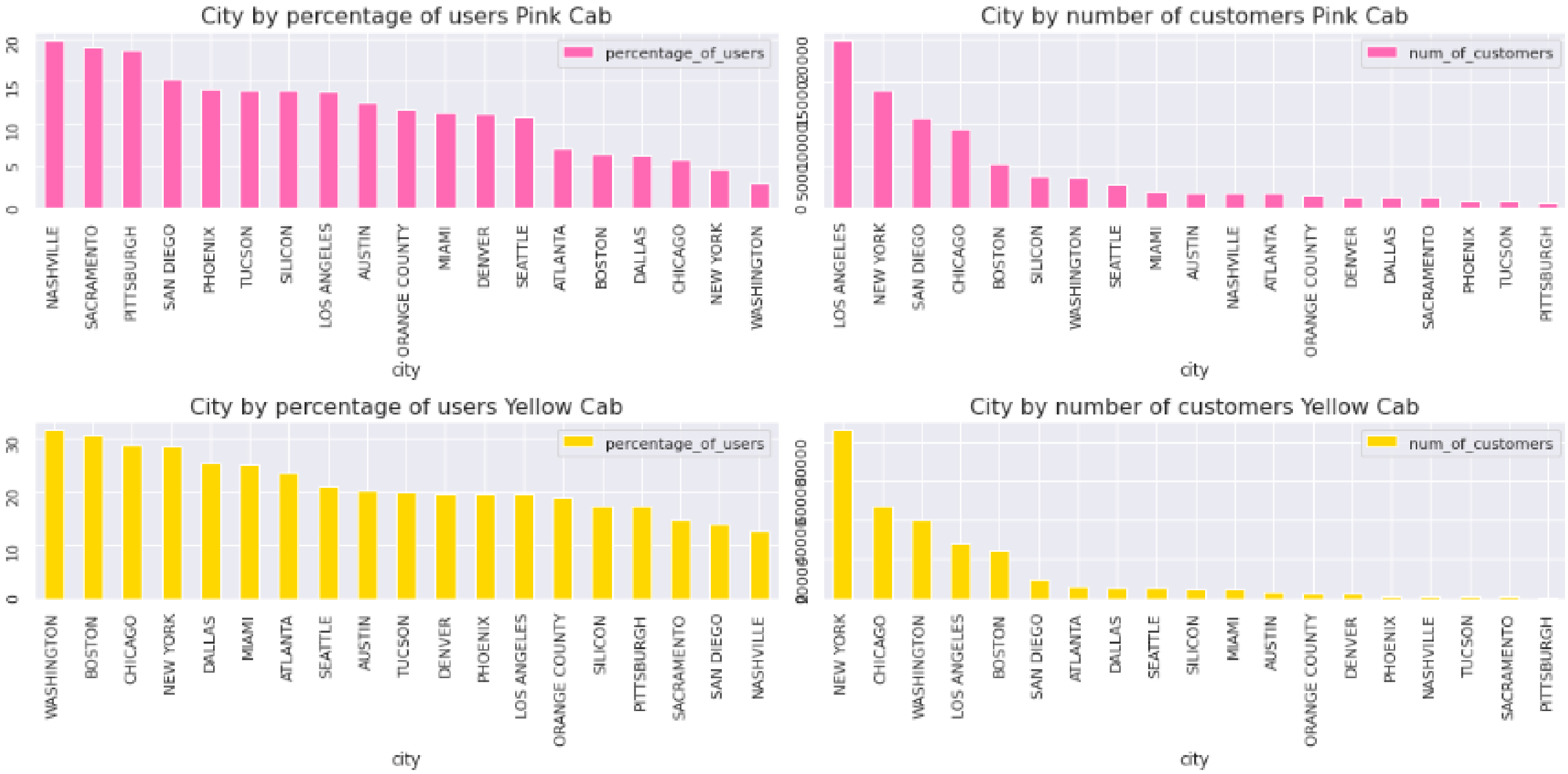


Yellow Cab

Pink Cab

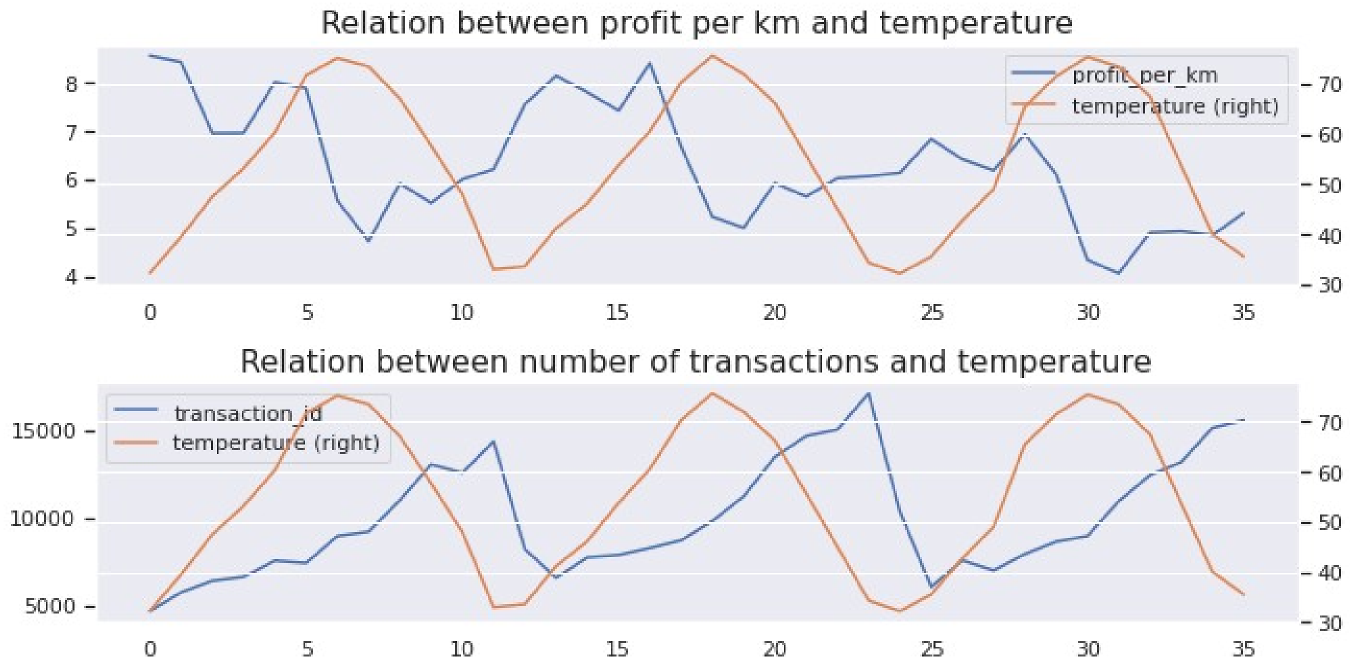
Potential of developing cab service in different cities





- We can see that New York significantly differs from the other cities due to its population and relatively high profit per km. More attractive cities are cities that are situated closer to upper-right corner and have a bigger 'bubble' size.
- For Yellow Cab it’s New York, Dallas, Silicon Valley, Denver, Miami, etc. For Pink Cab it’s Silicon Valley, Miami, Tucson, New York, etc.
- More detailed cities range by percentage of cab users and by number of customers for both companies is presented above.

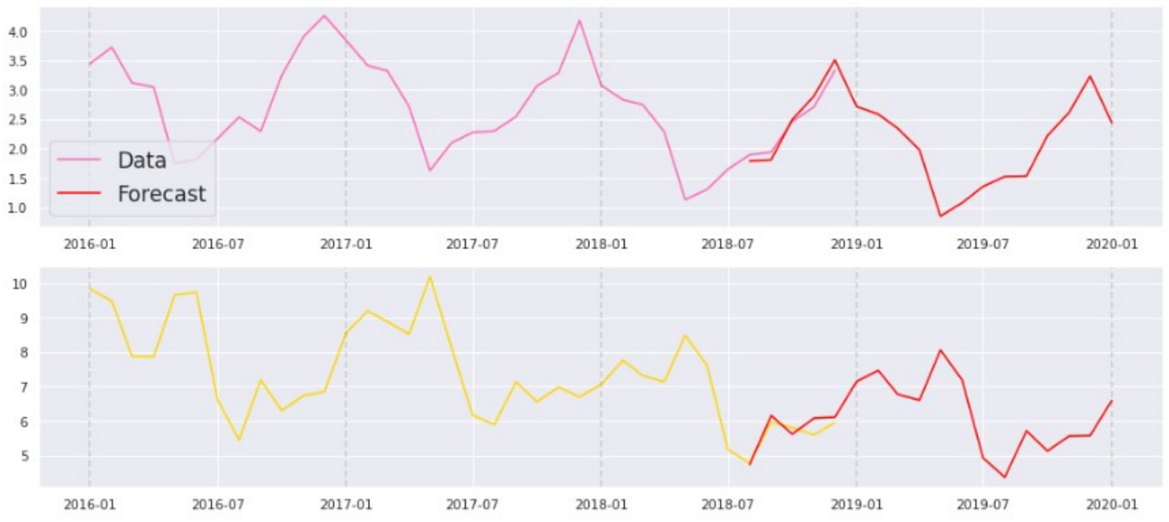
Climate factor analysis and forecast



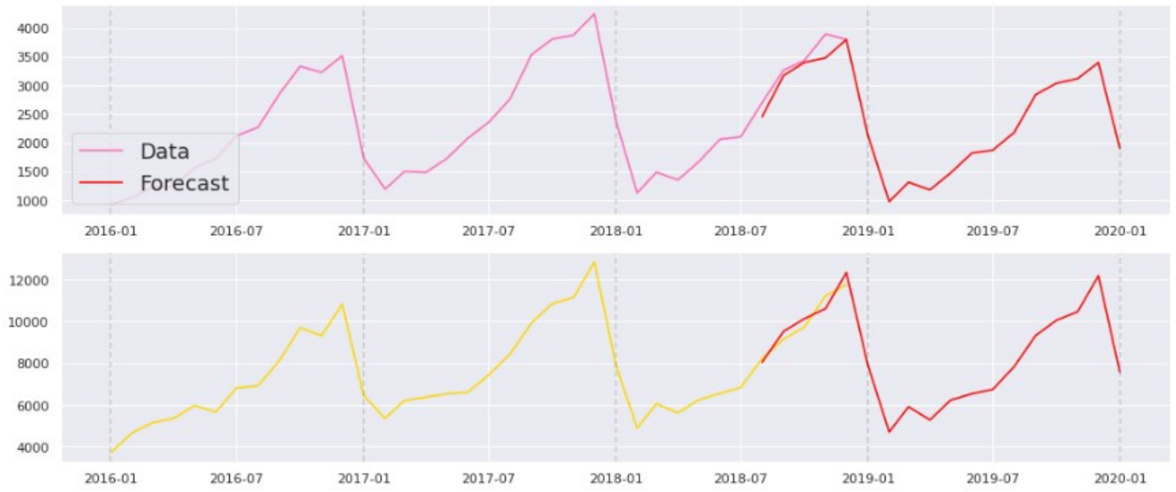
Relation between profit per km and temperature

Relation between number of transactions and temperature

Mean profit per km forecast



Total number of customers forecast



There is medium negative correlation between temperature and profit per km. This means that in colder days profit per km is higher. There's no connection between temperature and number of transactions. Probably, it's connected with a fact that we take average US temperature and it should be considered region-wisely.

EDA summary and recommendations

- There is seasonality in number of transactions and number of customers, as well as in profit and profit per km metrics in both companies. There is negative trend in mean profit per km, so, in near future companies may face decreasing profits.
- There is significant difference between sum of total profit and number of customers between 2 genders. Moreover, there is a difference in mean profit per km between genders.
- There is a strong negative correlation between age and number of customers and age and total profit. Elder people use cab service less often and have a smaller contribution to total profit.
- There's no strict connection between income and profit. However, the biggest share of customers and profit belongs to high-medium segment (13000-25000 USD/month).
- In all age groups payment by card is more popular, in average in 1,5 times.
- Number of customers, profit per km and total profit are connected with number of users and population of the city. More populated cities have higher potential for Cab service development.
- There's no strong connection between the companies metrics and temperature. However, there is medium negative correlation between temperature and profit per km.

Yellow Cab metrics compared to Pink Cab

- **Number of transactions:** 3.24 times higher;
- **Customers share**: 76.43%;
- **Total profit**: 8.29 times higher;
- **Profit per km**: 2.57 times higher;
- **Customer retention >5** – 23 times higher; **>10** – 822.5 times higher.

Yellow Cab has higher performance in many aspects.

- It has higher customer and transaction number, higher customer retention, higher mean profit per km.
- It's true for all age, gender and income groups, as well as for all months of analyzed period.
- Moreover, it has wider geographic coverage.
- That is why, **Yellow Cab is more recommended for investment.**

G2M insight for Cab Investment firm

Thank You