Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

Cab-Investment-firm-Data-Analysis-and-Insights

Igor Azevedo de Queiroz

06/08/2021

Business Problem/Understanding

- XYZ it is a private firm located is 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 object of this project is providing answer of the main questions made by the company’s CEO, which are:

- Which Cab companies XYZ firm should invest, in general?
- What is maximum amount of money made by each company?
- Which company provided more taxi travels?
- Which company has more customers?

- Also there are secondary questions that could help the CEO decision, which are:

- Which area of US the Cab companies provide higher incomes and taxi travels?
- Which season of US the Cab companies provide higher incomes and taxi travels?

Data Understanding

There are 6 different datasets providing a total of 20 attributes:

- Cab_Data.csv – this file includes details of transaction for 2 cab companies;
- Customer_ID.csv – this 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.
- States – this file contains list of US cities, regions and divisions. –
- US-federal-holidays-2011-2020 – this file contains list of US holidays from 2011 to 2020.

A merge method is used to merge all datasets in one big dataset.

Business Assumptions

Assumptions:

For further analysis, it is necessary to know the period of time of each season on US, those information will be included in a variable called Season. According to the website: https://www.timeanddate.com/calendar/aboutseasons.html, each season is specified on the followed days:

- spring runs from March 1 to May 31;
- summer runs from June 1 to August 31;
- fall (autumn) runs from September 1 to November 30; and
- winter runs from December 1 to February 28 (February 29 in a leap year).

Due the dataset City.csv does not inform the State of two cities: SILICON VALLEY and ORANGE COUNTY, it is necessary specify the States of them. According to the website: https://www.britannica.com/place/Silicon-Valley-region-California, The SILICON VALLEY State is California, CA. Also according to the website: https://en.wikipedia.org/wiki/Orange_County,_California, The ORANGE COUNTY State is California, CA.

Profit attribute is calculated based on the attributes Price_Charged and Cost_of_Trip.

Solution Strategy

The answer for which Cab company XYZ firm should invest can be answered as 2 different ways:

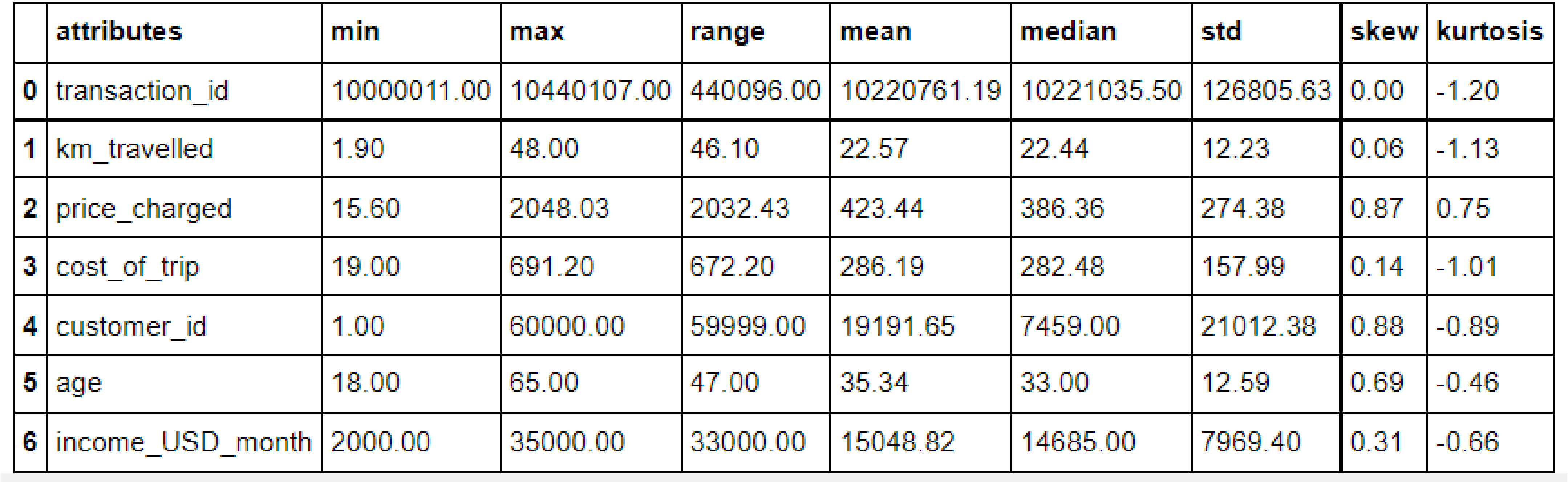
First part: How much money, taxi travels and customers each company made?
1º - Group the variable Company by the variable Profit, and sum the amount of price for each company. This method shows a general perspective about each company in all cities and years.
2º - For taxi travels, it is possible do the same method, however, it is necessary replace the variable Profit by the variable Transaction Id.
3º - For customers, it is possible do the same method, however, it is necessary eliminate the duplications and replace the variable Profit by the variable Customer Id.

Second part: Here it is made a verification of the amount of money and taxi travels by region and season.
1º - Group the variable Company by the variable Profit and Region, and sum the amount of price for each company. This method shows a specific perspective about each company in each region. For taxi travels, it is possible do the same method; however, it is necessary replace the variable Profit by the variable Transaction Id.
2º - Same method is used than first one, however replace the variable Region by Season. This method shows a specific perspective about each company in each season.

The general and specific overview about price and clients will help the decision about which company XYZ should invest.

Furthermore hypotheses will be created to influence that decision as well.

Data Numerical Descriptive Statistics

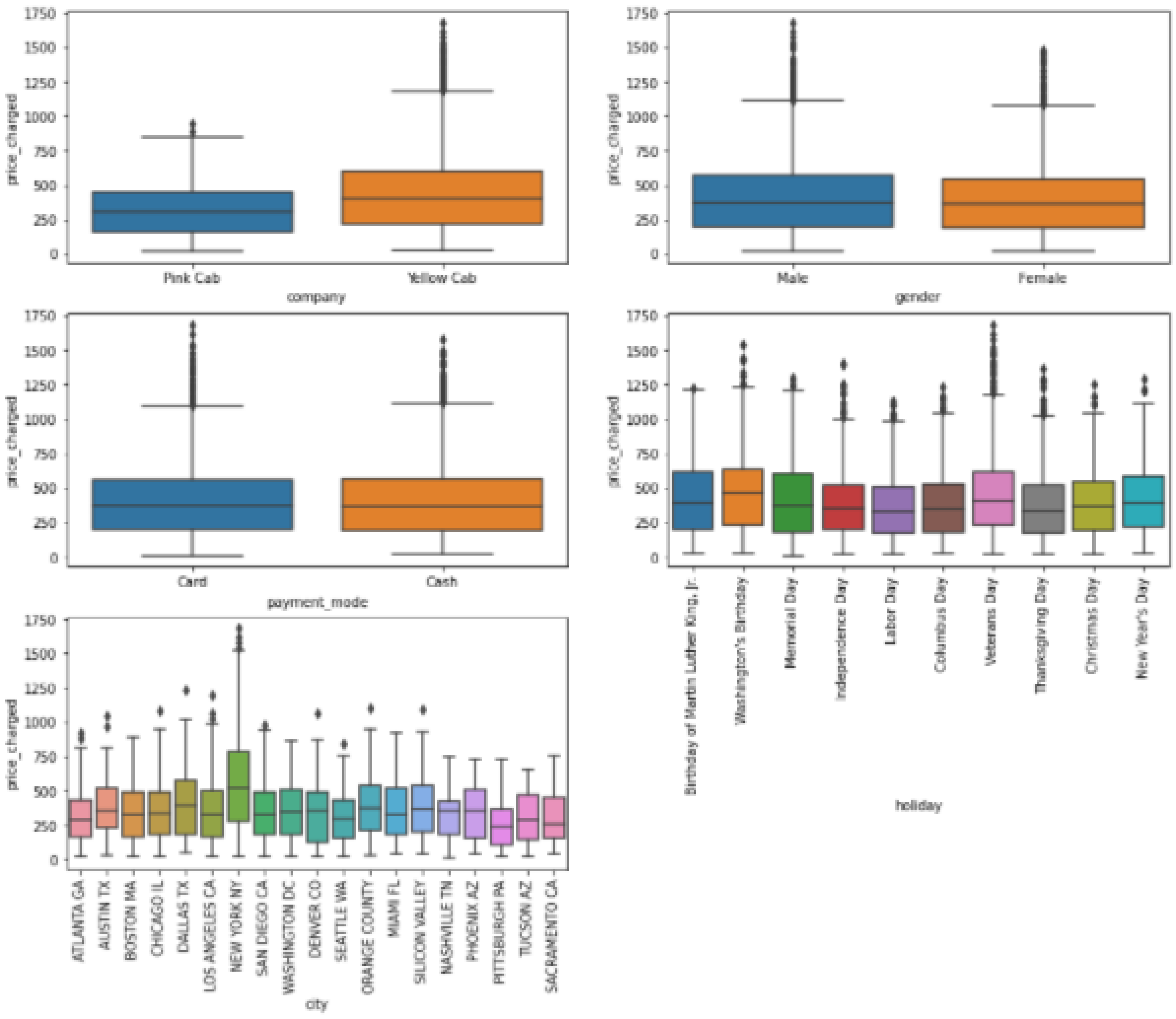


	attributes	min	max	range	mean	median	std	skew	kurtosis
0	transaction_id	10000011.00	10440107.00	440096.00	10220761.19	10221035.50	126805.63	0.00	-1.20
1	km_travelled	1.90	48.00	46.10	22.57	22.44	12.23	0.06	-1.13
2	price_charged	15.60	2048.03	2032.43	423.44	386.36	274.38	0.87	0.75
3	cost_of_trip	19.00	691.20	672.20	286.19	282.48	157.99	0.14	-1.01
4	customer_id	1.00	60000.00	59999.00	19191.65	7459.00	21012.38	0.88	-0.89
5	age	18.00	65.00	47.00	35.34	33.00	12.59	0.69	-0.46
6	income_USD_month	2000.00	35000.00	33000.00	15048.82	14685.00	7969.40	0.31	-0.66

Numerical descriptions of all integer and float attributes.

Data Categorical Descriptive Statistics

Categorical descriptions of all no integer and no float attributes.



Feature Engineering

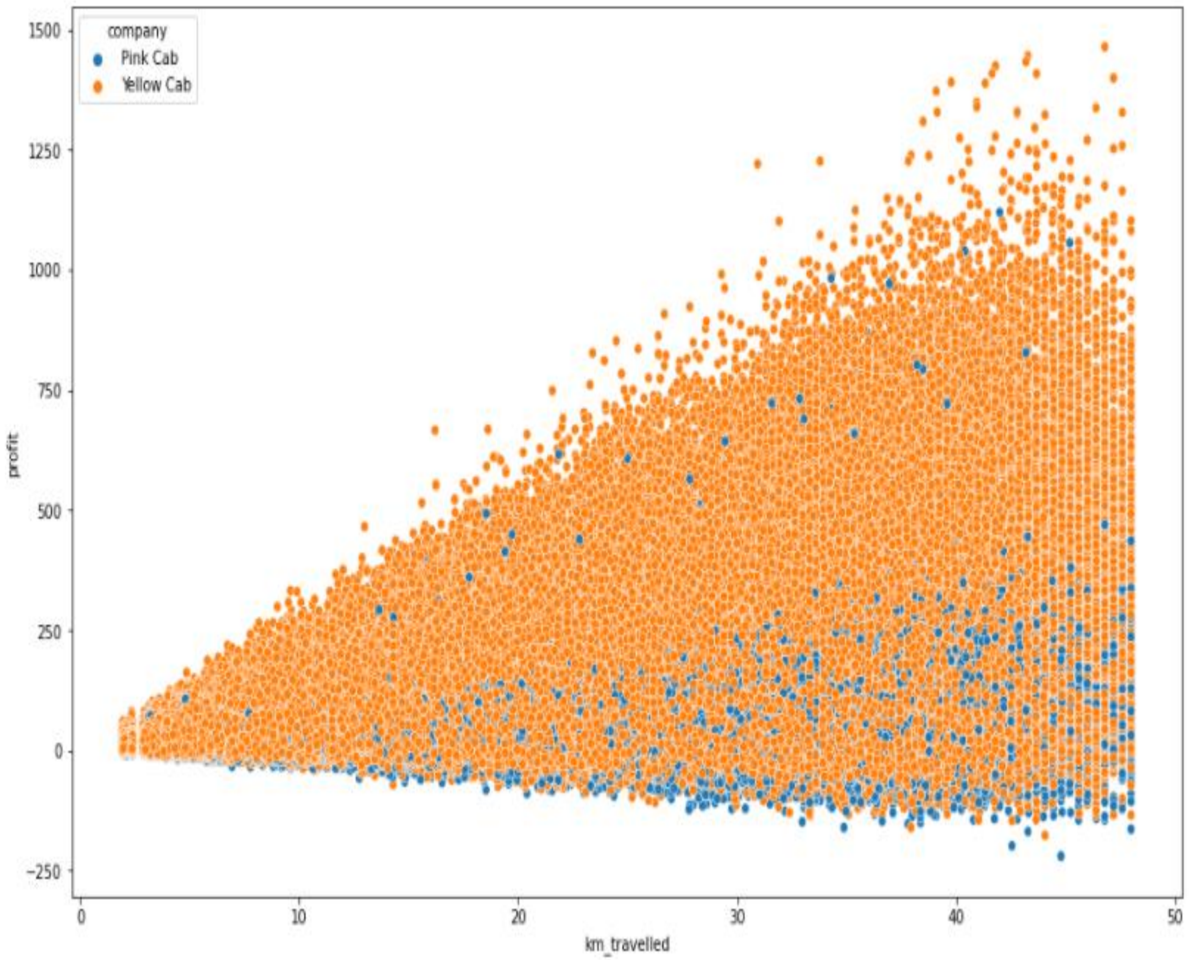
Here new attributes are created, such as:

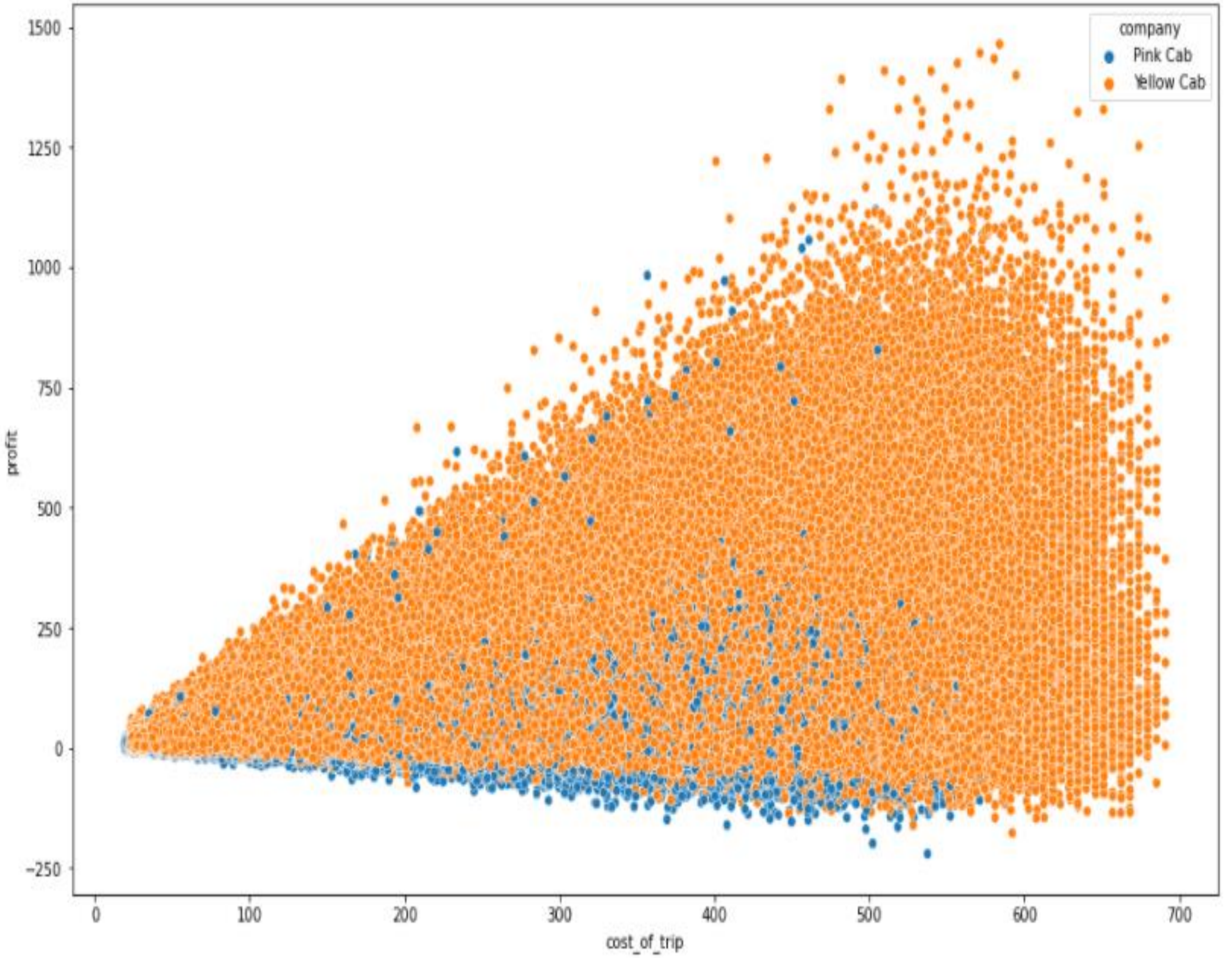
- Profit;
- Year;
- Month;
- Day;
- Year – Month;
- Year – Week;
- Month – Day;
- Season;
- State – Code

After the application of feature engineering on dataset, the total number of attributes that is useful is 24.

This attributes is responsible to answer the business question and for the creation of several hypotheses.

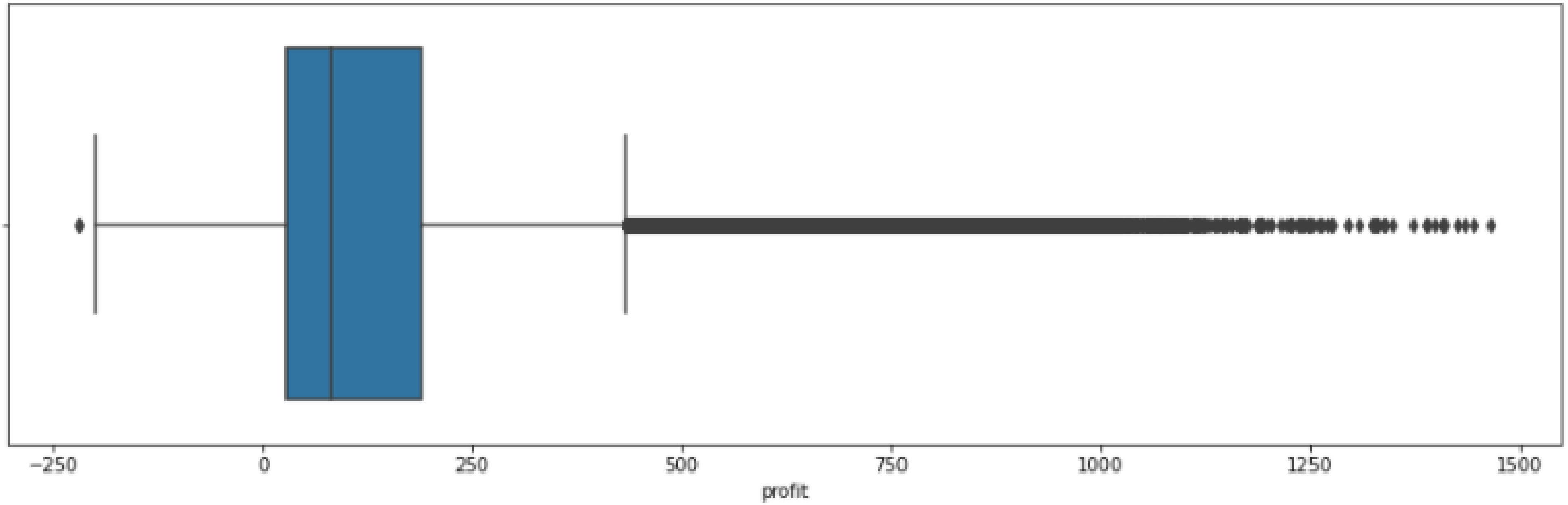
Outliers Detection - Scatterplot





Outliers Detection - Boxplot

The attributes that present outliers are Price_Charged and Profit. However, as Profit is a derivation of Price_Charged and Cost_of_Trip, the attribute Profit is the only one to be considered.



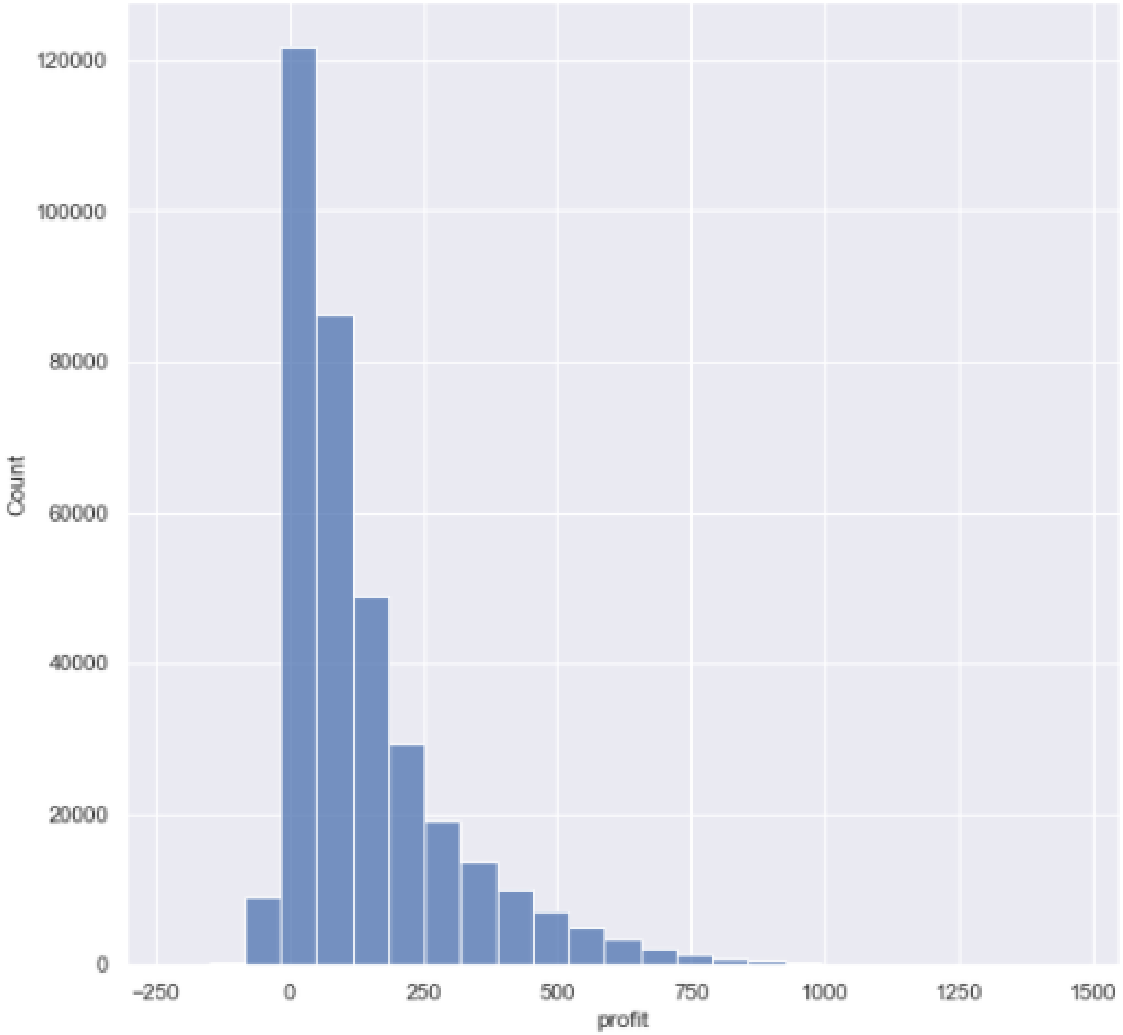
Despite the boxplot shows that there are a lots of outliers from a value of 433, the others scatterplots shows a pattern of distribution, with few outliers from value of around 1250 just for the Yellow Cab company.

Also the yellow company shows a number of travel so much higher than the Pink Cab.

For those reasons, it is not necessary exclude any outlier of this dataset.

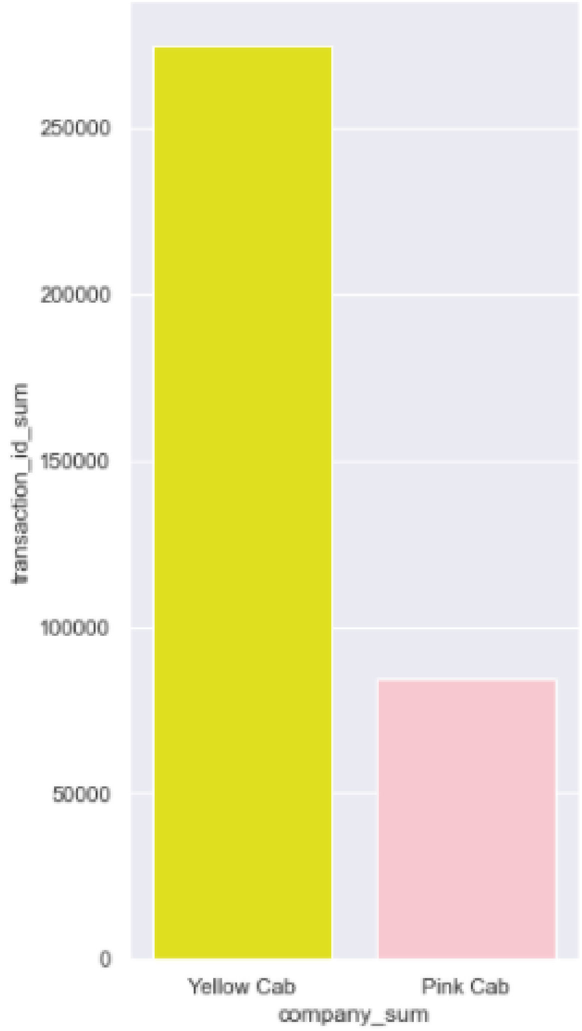
Exploratory Data Analysis (EDA)

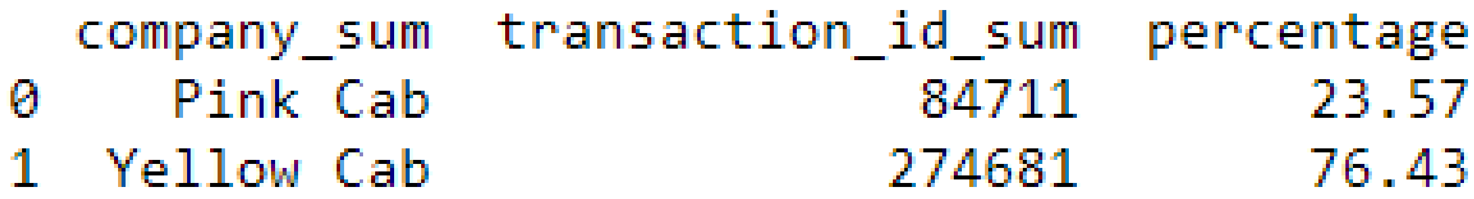
The attributes Profit has some kurtosis and presents a positive skew, which is to the left.



Business Questions 1: How many travels each company made?

In total, The company Yellow Cab has 189970 (52.86%) more taxi travels than the company Pink Cab.

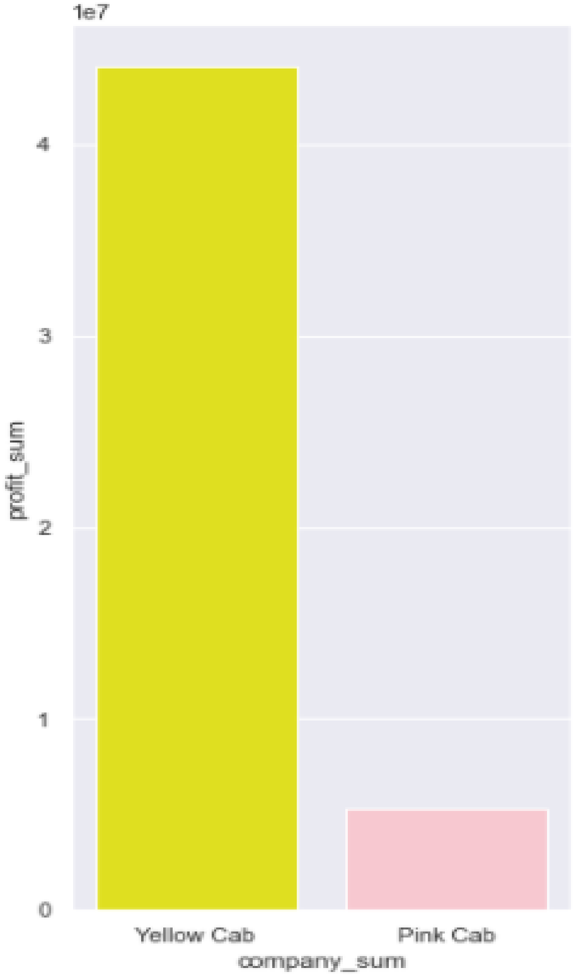


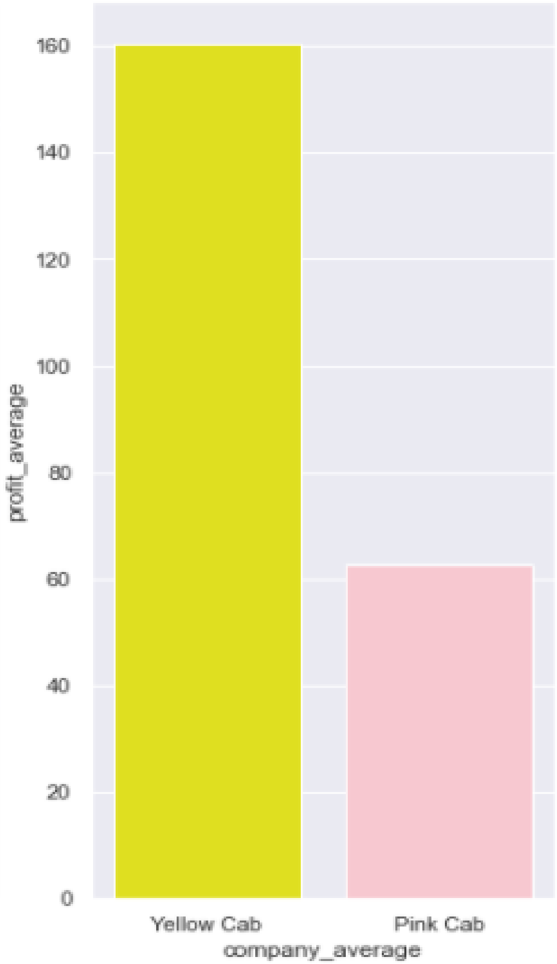


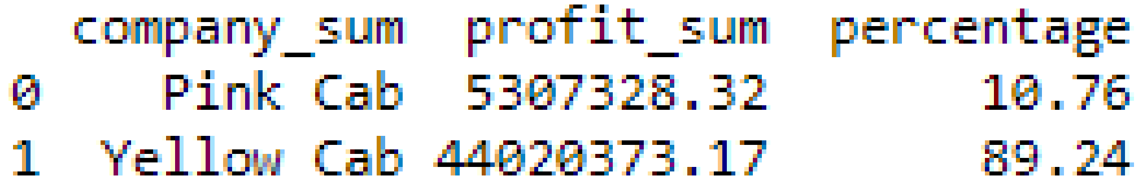
```
    company_sum  transaction_id_sum  percentage
0      Pink Cab               84711       23.57
1    Yellow Cab              274681       76.43
```

Business Questions 2: What is the total of money made by each company?

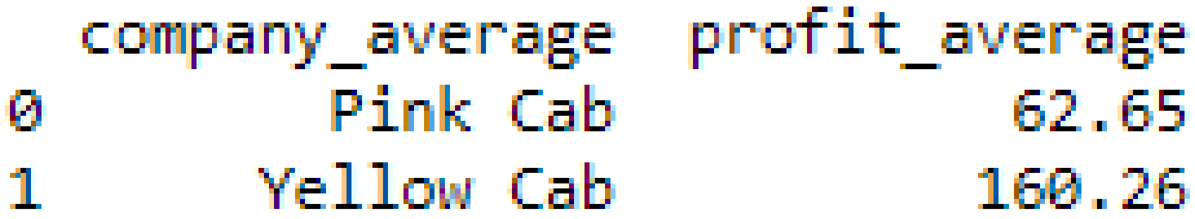
In total, the company Yellow Cab has $38713044.8498 (78.48%) more profit than the company Pink Cab.
On average, also, the company Yellow Cab has $97.61 (155.79%) more profit than the company Pink Cab.





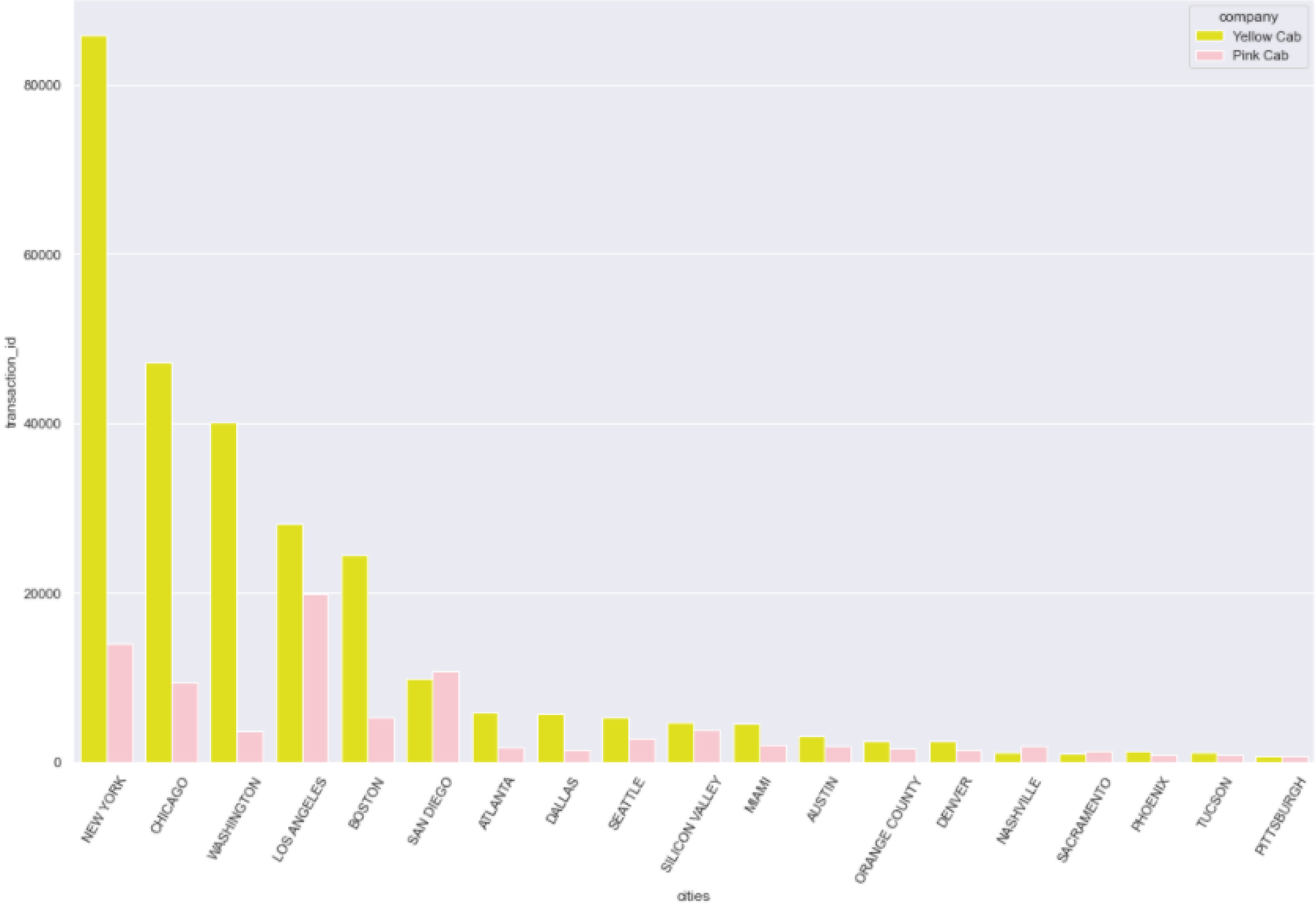


```
  company_sum   profit_sum  percentage
0    Pink Cab   5307328.32       10.76
1  Yellow Cab  44020373.17       89.24
```



```
  company_average  profit_average
0        Pink Cab           62.65
1      Yellow Cab          160.26
```

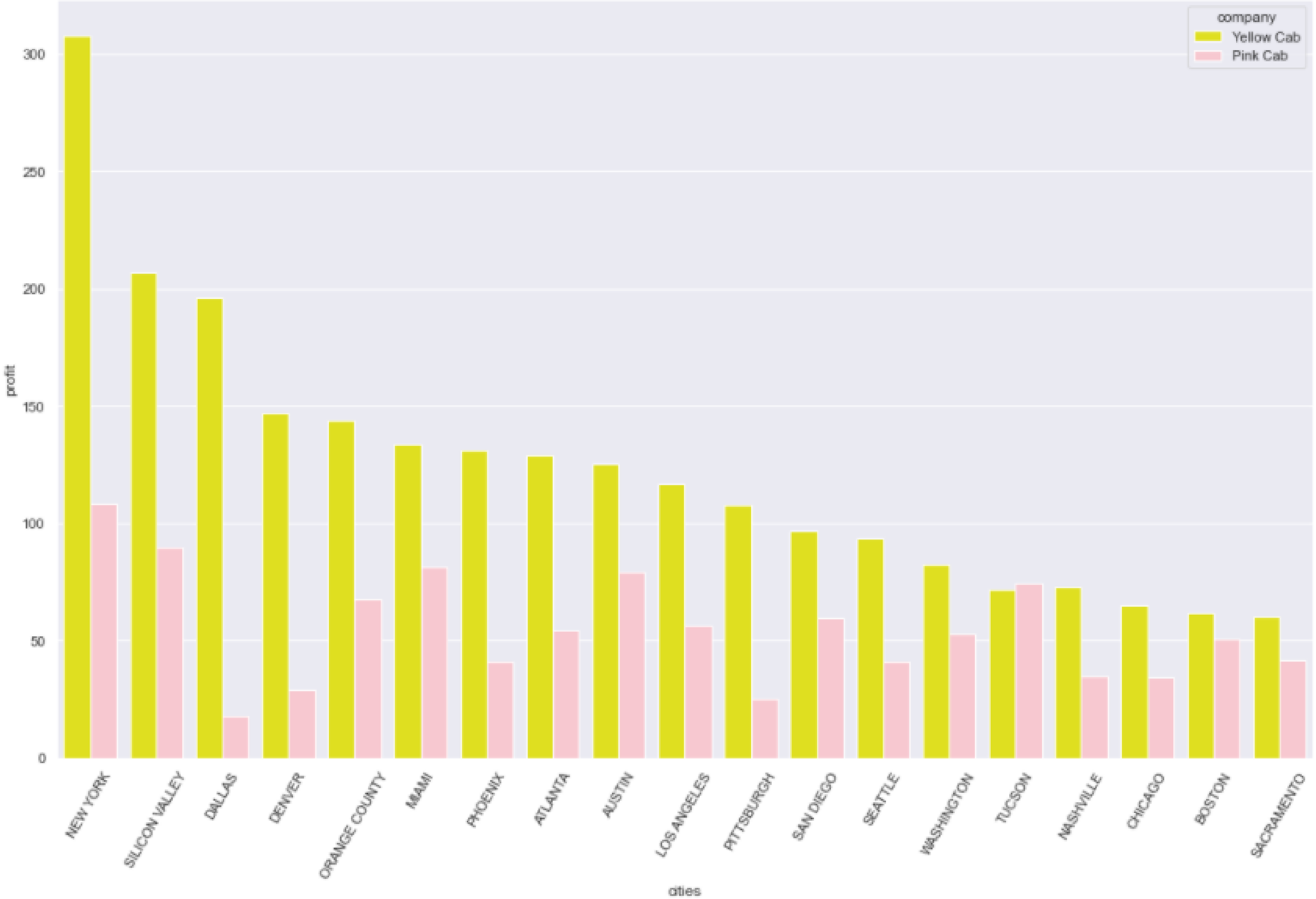
Business Questions 3: In total, how many travels each company made in each city?



In total, the Yellow Cab Company has higher number of travel in almost all cities, whit exception on the cities: San Diego, Nashville and Sacramento.

And New York is the city with higher number of travels, with more than 80000 travels for the Yellow Cab Company.

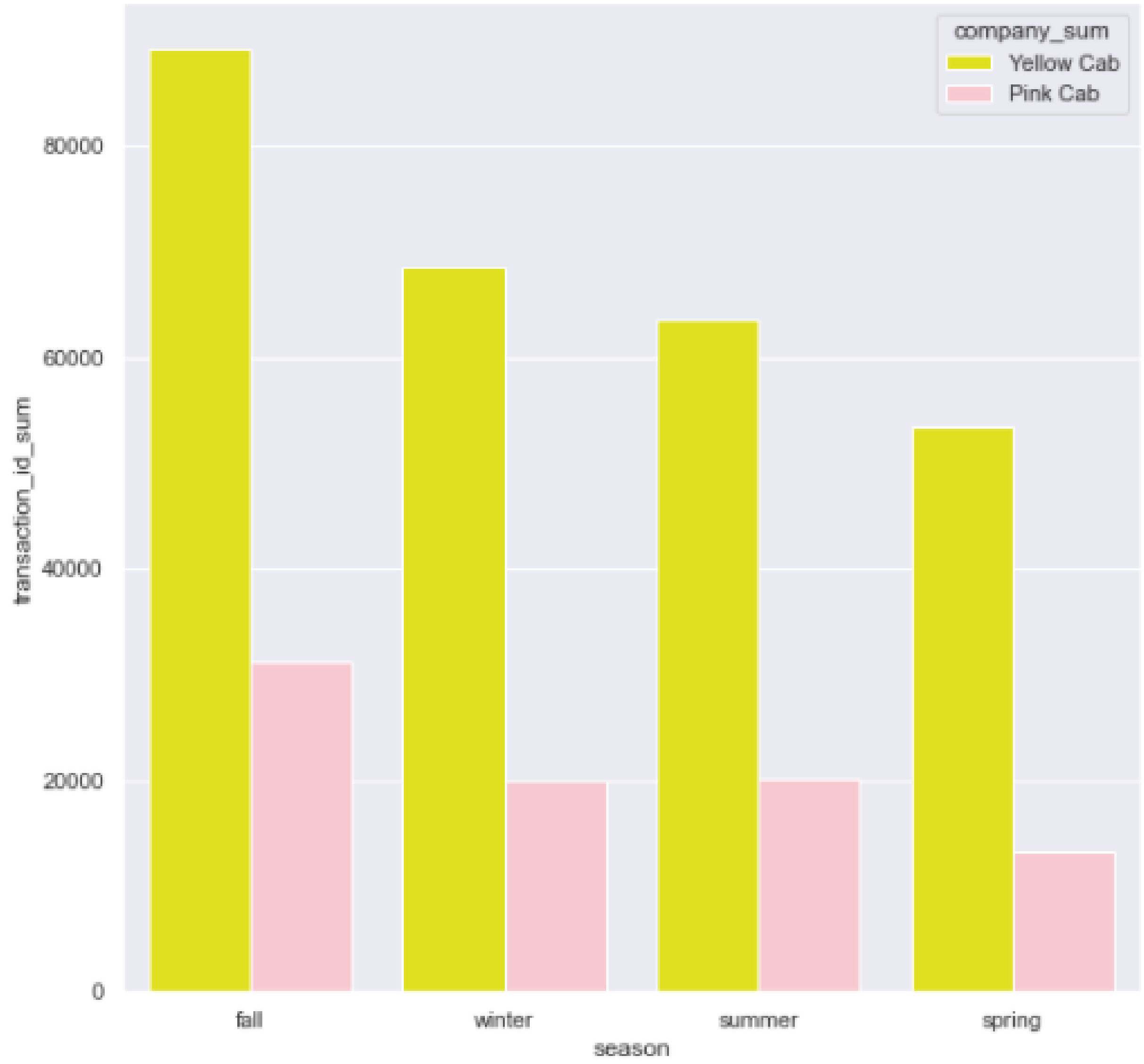
Business Questions 4: In average, which city the cab companies provide higher income?



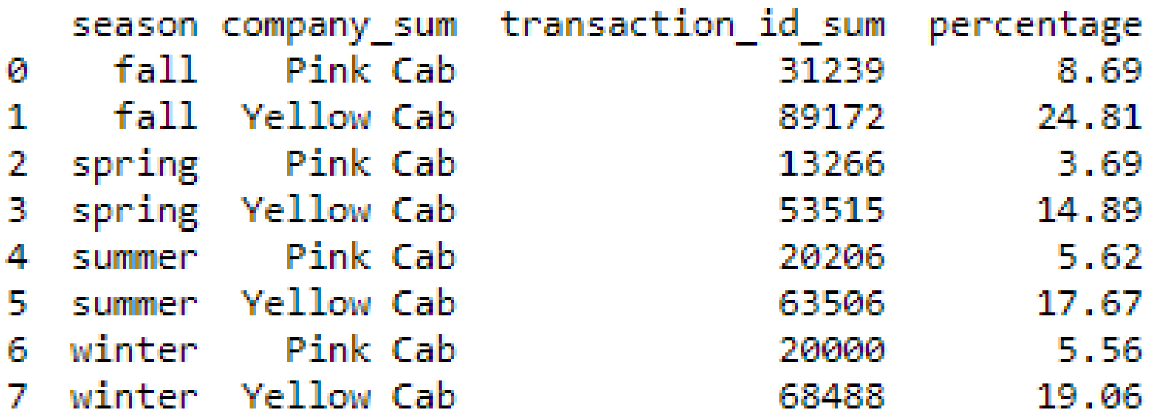
In average the Yellow Cab Company has the higher profit in almost all cities, whit exception on the city: Tucson.

And New York is the city with higher profit, with more than $300 in average for the Yellow Cab Company.

Business Questions 5: How many travels each company has on each season, in total?

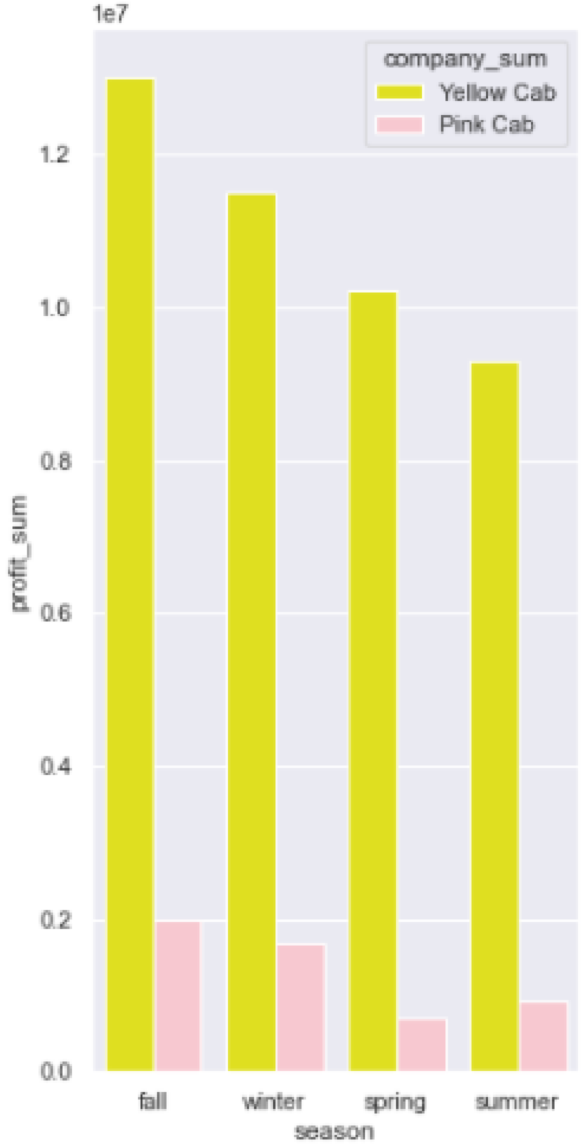


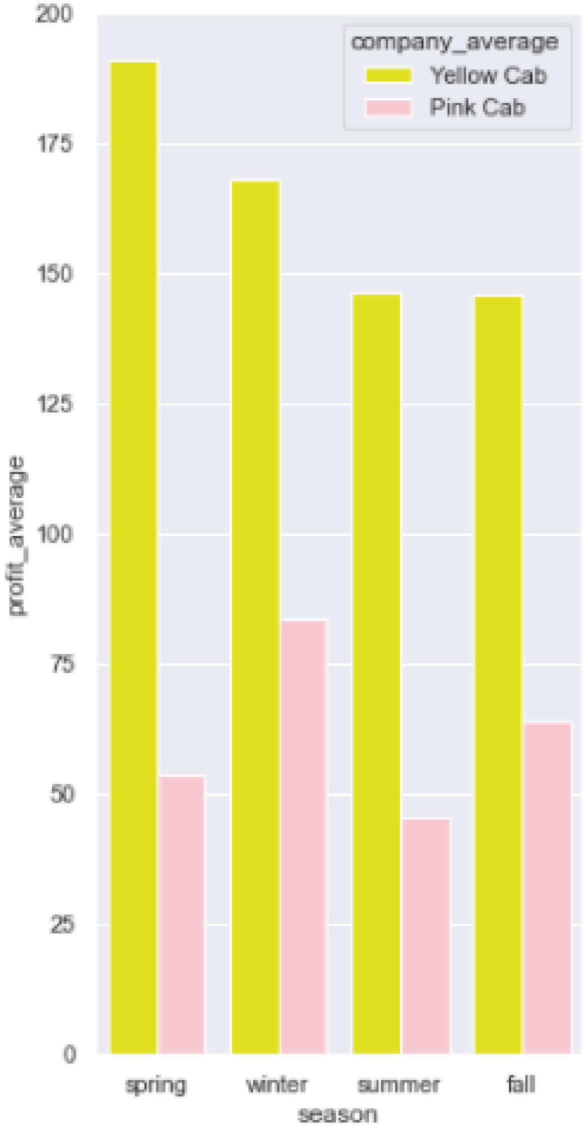
On total, the season fall for the company Yellow Cab has the highest percentage of taxi travels among all seasons, with 24.81%.



```
   season company_sum  transaction_id_sum  percentage
0    fall    Pink Cab               31239        8.69
1    fall  Yellow Cab               89172       24.81
2  spring    Pink Cab               13266        3.69
3  spring  Yellow Cab               53515       14.89
4  summer    Pink Cab               20206        5.62
5  summer  Yellow Cab               63506       17.67
6  winter    Pink Cab               20000        5.56
7  winter  Yellow Cab               68488       19.06
```

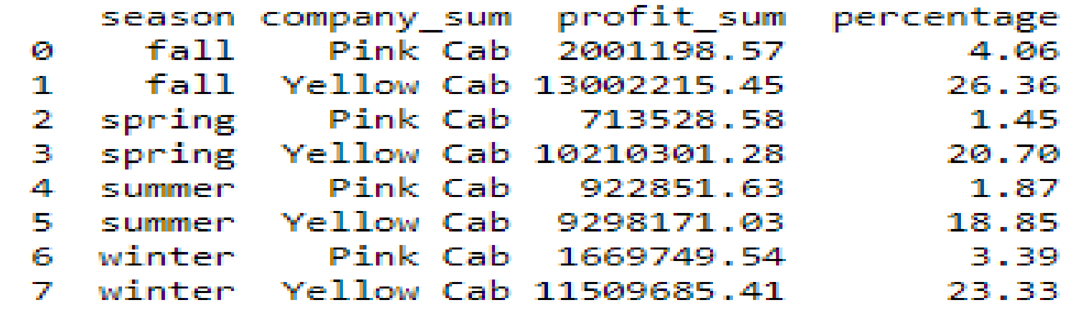
Business Questions 6: How many travels each company has on each season, in total?



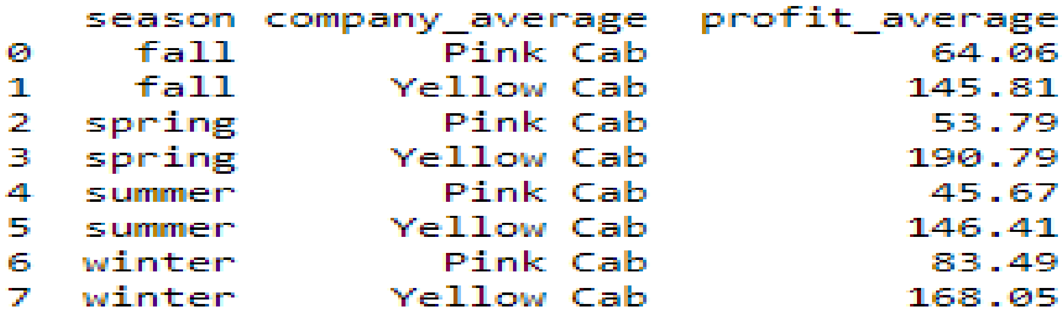


On total, the season fall for the company Yellow Cab has the highest amount of profit among all seasons, with $13002215.45 (26.36%).

However, on average, the season spring for the company Yellow Cab has the highest amount of profit among all seasons, with $190.79.

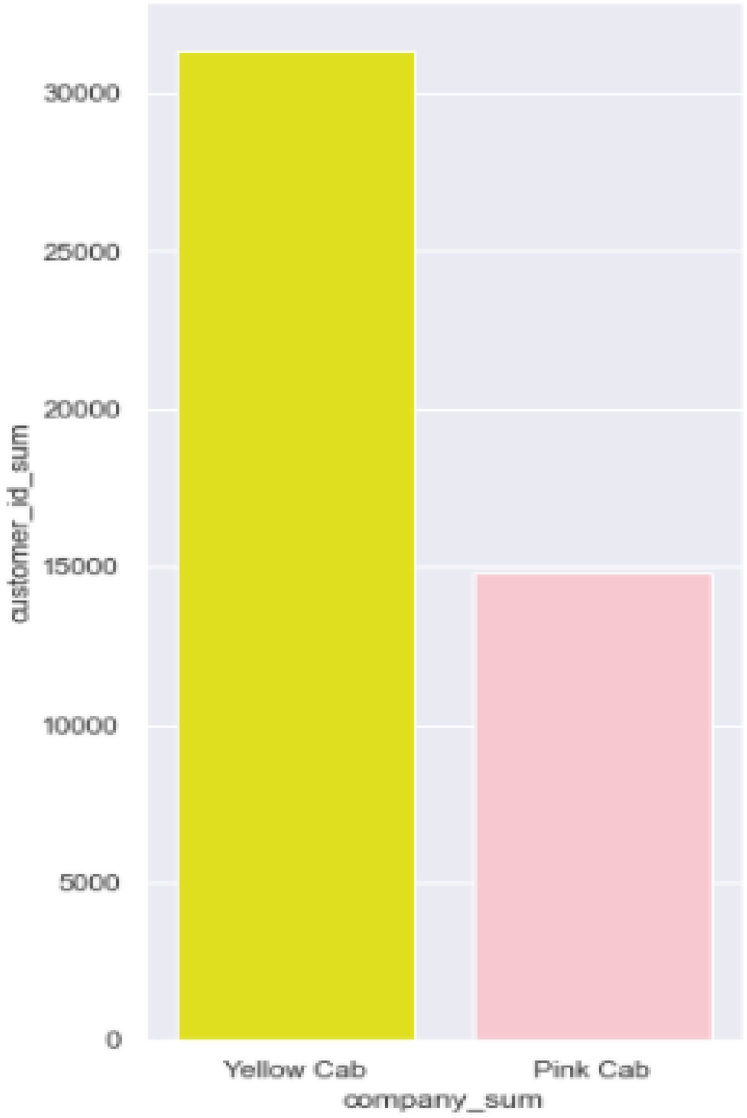


```
   season company_sum   profit_sum  percentage
0    fall    Pink Cab   2001198.57        4.06
1    fall  Yellow Cab  13002215.45       26.36
2  spring    Pink Cab    713528.58        1.45
3  spring  Yellow Cab  10210301.28       20.70
4  summer    Pink Cab    922851.63        1.87
5  summer  Yellow Cab   9298171.03       18.85
6  winter    Pink Cab   1669749.54        3.39
7  winter  Yellow Cab  11509685.41       23.33
```

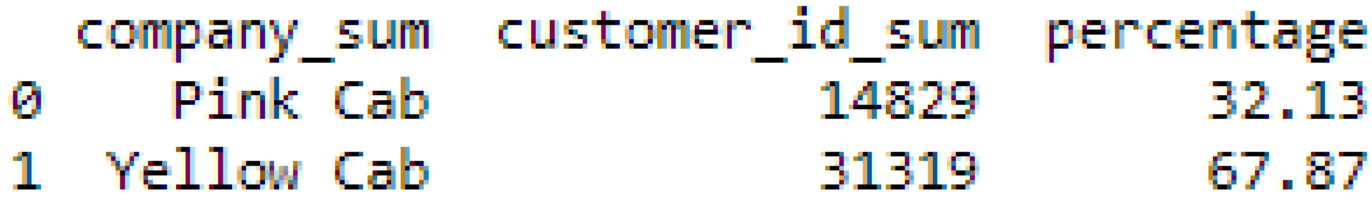


```
   season company_average  profit_average
0    fall        Pink Cab           64.06
1    fall      Yellow Cab          145.81
2  spring        Pink Cab           53.79
3  spring      Yellow Cab          190.79
4  summer        Pink Cab           45.67
5  summer      Yellow Cab          146.41
6  winter        Pink Cab           83.49
7  winter      Yellow Cab          168.05
```

Business Questions 7: How many customers each company has?



In total, the company Yellow Cab has 16490 (35.73%) more customers than the company Pink Cab.



```
  company_sum  customer_id_sum  percentage
0    Pink Cab            14829       32.13
1  Yellow Cab            31319       67.87
```

Answering the Business Questions

Yellow Cab company has a higher number of taxi trips than Pink Cab company, as well as a higher profit and higher number of customers.

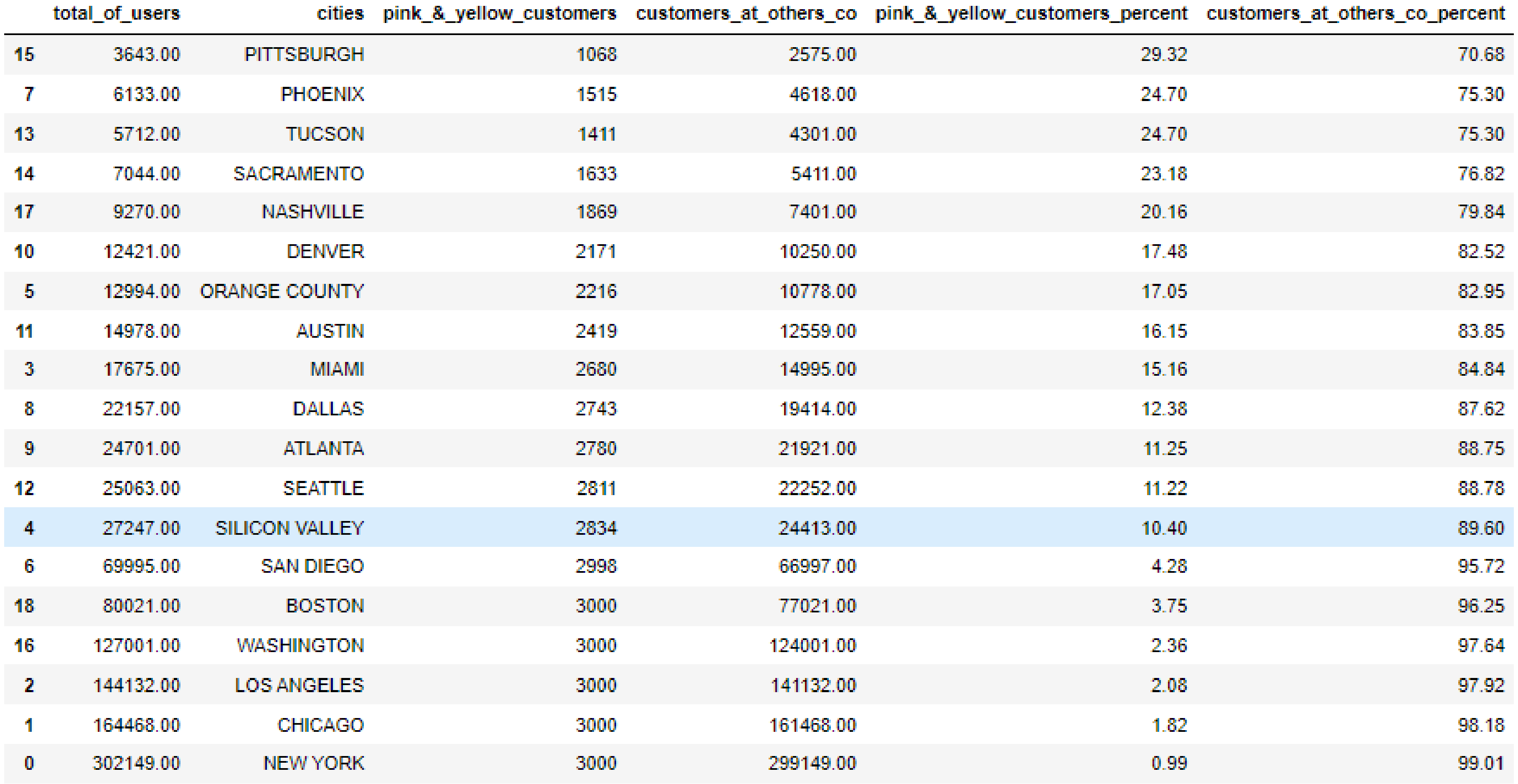
The data also show that Yellow Cab has superior results in almost all cities, especially in New York, which is the city with the highest profit and higher number of taxi trips.

The same company also stands out from the other company in relation to the season, where the fall or spring (depending which measure is more important) season shows the highest profit and number of taxi trips as well.

Due those all information, it is recommended to invest in Yellow Cab company.

Hypothesis 1: The number of users of the Yellow Cab plus Pink Cab is the same number of the Total Cab users at every city.

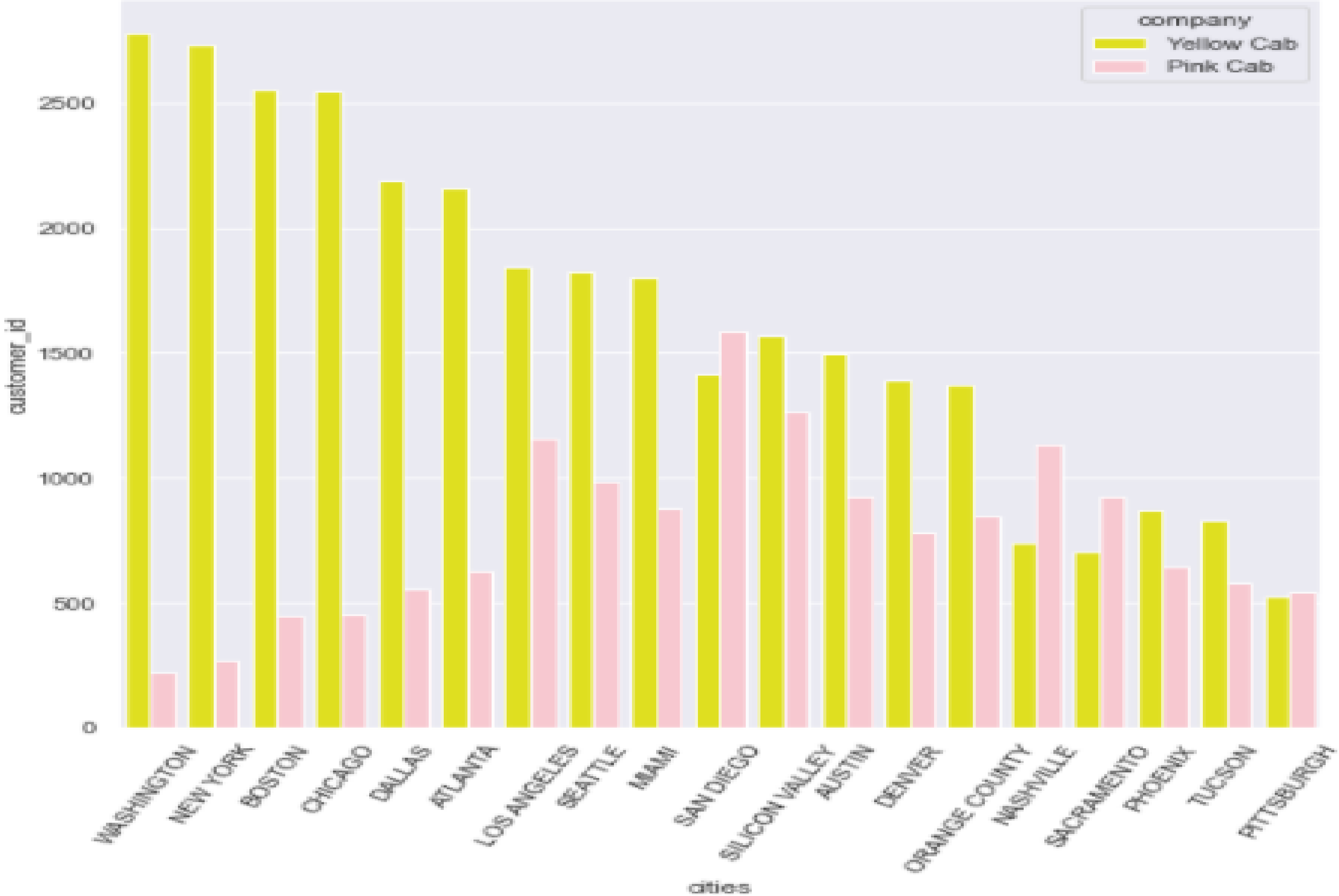
FALSE There are others Cab companies at each state, which the range of customers in others companies floats between 70% and 99%, depending of the state.



	total_of_users	cities	pink_&_yellow_customers	customers_at_others_co	pink_&_yellow_customers_percent	customers_at_others_co_percent
15	3643.00	PITTSBURGH	1068	2575.00	29.32	70.68
7	6133.00	PHOENIX	1515	4618.00	24.70	75.30
13	5712.00	TUCSON	1411	4301.00	24.70	75.30
14	7044.00	SACRAMENTO	1633	5411.00	23.18	76.82
17	9270.00	NASHVILLE	1869	7401.00	20.16	79.84
10	12421.00	DENVER	2171	10250.00	17.48	82.52
5	12994.00	ORANGE COUNTY	2216	10778.00	17.05	82.95
11	14978.00	AUSTIN	2419	12559.00	16.15	83.85
3	17675.00	MIAMI	2680	14995.00	15.16	84.84
8	22157.00	DALLAS	2743	19414.00	12.38	87.62
9	24701.00	ATLANTA	2780	21921.00	11.25	88.75
12	25063.00	SEATTLE	2811	22252.00	11.22	88.78
4	27247.00	SILICON VALLEY	2834	24413.00	10.40	89.60
6	69995.00	SAN DIEGO	2998	66997.00	4.28	95.72
18	80021.00	BOSTON	3000	77021.00	3.75	96.25
16	127001.00	WASHINGTON	3000	124001.00	2.36	97.64
2	144132.00	LOS ANGELES	3000	141132.00	2.08	97.92
1	164468.00	CHICAGO	3000	161468.00	1.82	98.18
0	302149.00	NEW YORK	3000	299149.00	0.99	99.01

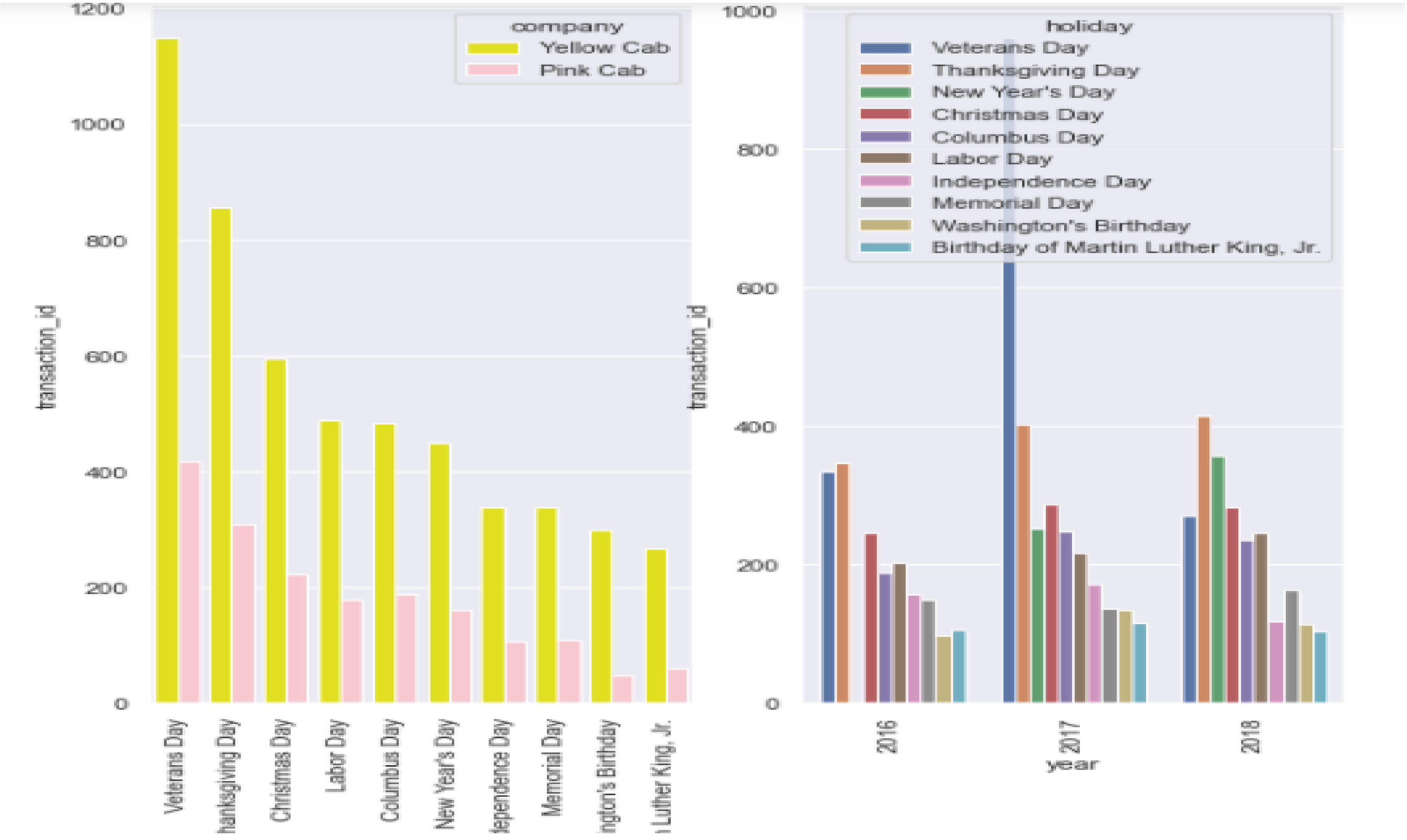
Hypothesis 2: The number of users of the Yellow Cab is higher than the company Pink Cab in each city.

FALSE There are cities where the Pink cab company has more customers than the Yellow cab company, cities such: San Diego, Nashville, Sacramento and Pitsburgh.



Hypothesis 3: People uses more taxi during New Year’s Day than other holiday in all 3 years.

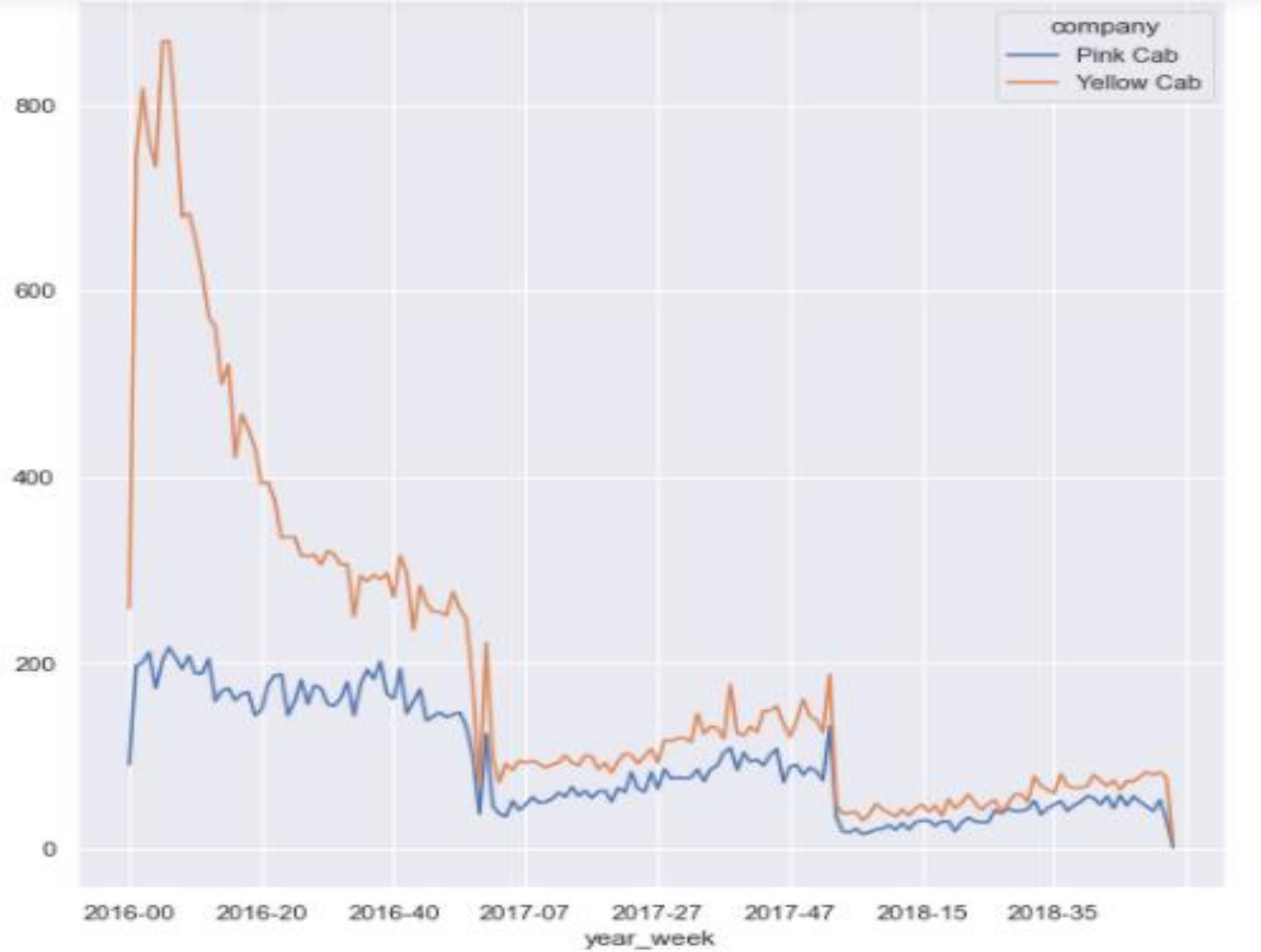
FALSE In total, the holiday which people uses more taxis is Veterans Day, However, splitting this in years, 2016 and 2018, Thanksgiving Day is the holiday which people uses more taxis, and in 2017 the holiday is Veterans Day.



Hypothesis 4: The number of users for each company increases WoW (Week over Week).

FALSE The number of users for the company Yellow Cab company WoW (Week over Week) suffered a decrease of 2.26 percent.

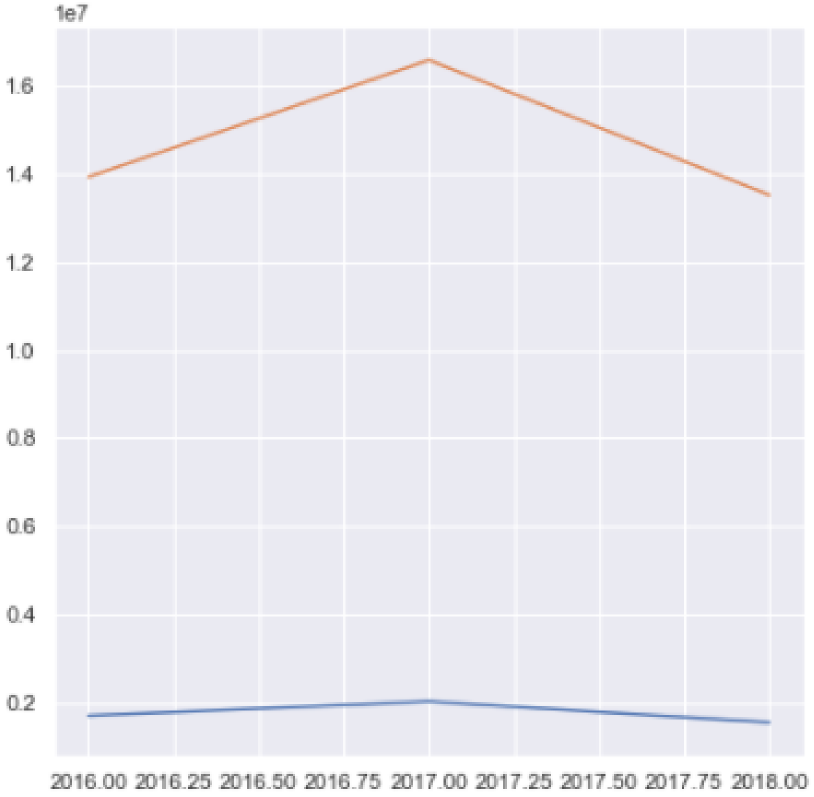
The number of users for the company Pink Cab company WoW (Week over Week) suffered a decrease of 2.39 percent.

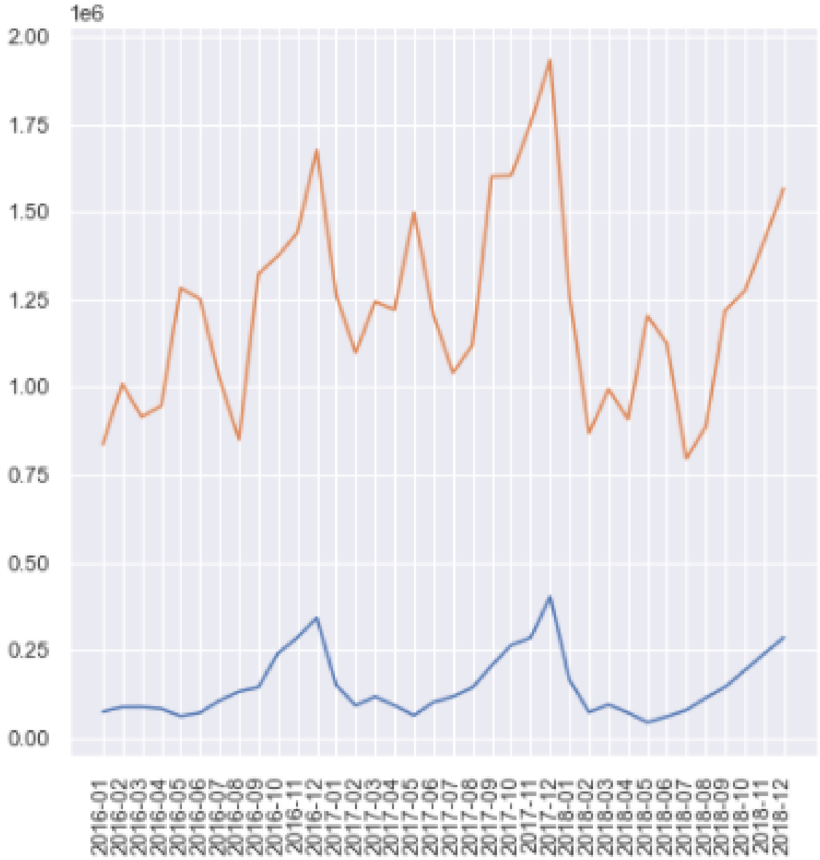


Hypothesis 5: The amount of money made by the cab companies increase YoY (Year over Year) and MoM (Month over Month).

FALSE The amount of money made by the company Yellow Cab company YoY (Year over Year) suffered a decrease of 1.48 percent. And the amount of money made by the company Pink Cab company YoY (Year over Year) suffered a decrease of 4.58 percent.

However, the amount of money made by the company Yellow Cab company MoM (Month over Month) suffered an increase of 1.81 percent. And the amount of money made by the company Pink Cab company MoM (Month over Month) suffered an increase of 3.94 percent.

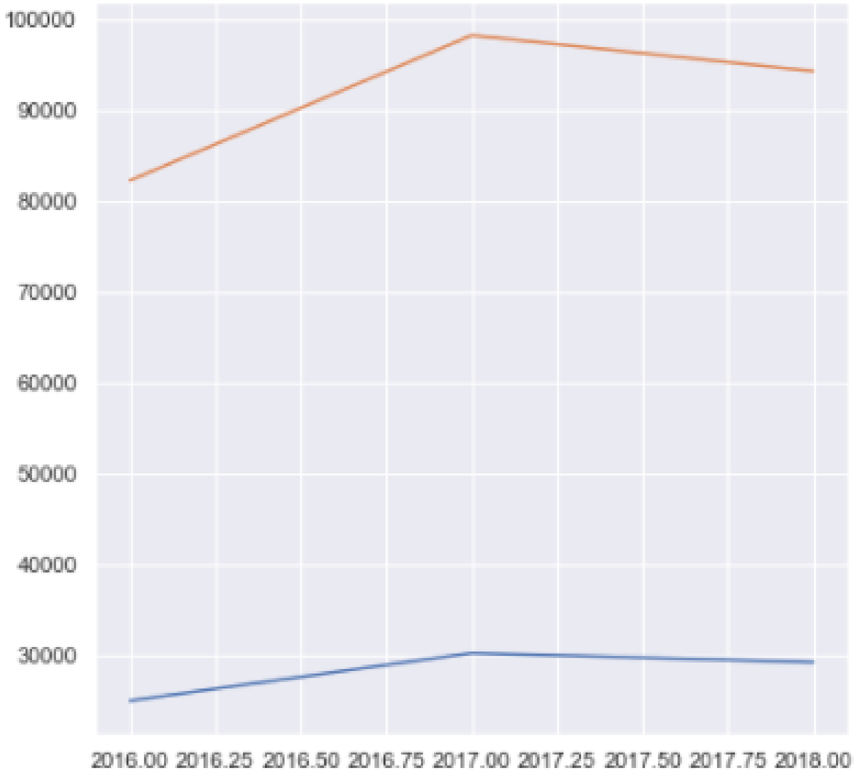


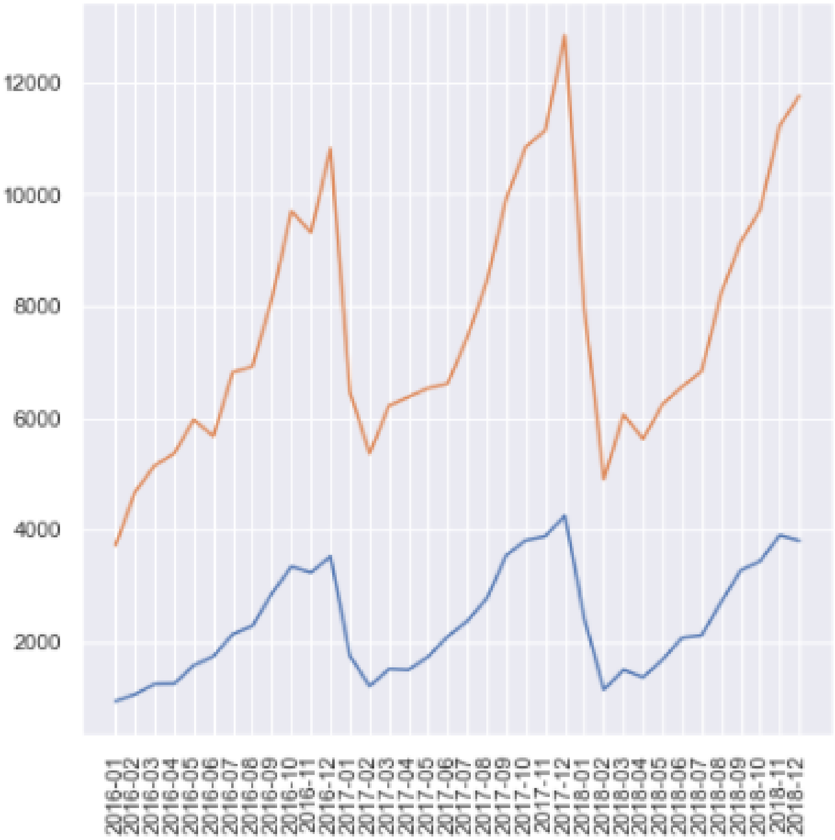


Hypothesis 6: The number of travels made by the cab companies increase YoY (Year over Year) and MoM (Month over Month).

TRUE The number of travels made by the company Yellow Cab company YoY (Year over Year) suffered a decrease of 7.06 percent. And the number of travels made by the company Pink Cab company YoY (Year over Year) suffered a decrease of 8.01 percent.

Also, the number of travels made by the company Yellow Cab company MoM (Month over Month) suffered an increase of 3.35 percent. And the number of travels of money made by the company Pink Cab company MoM (Month over Month) suffered an increase of 4.11 percent.

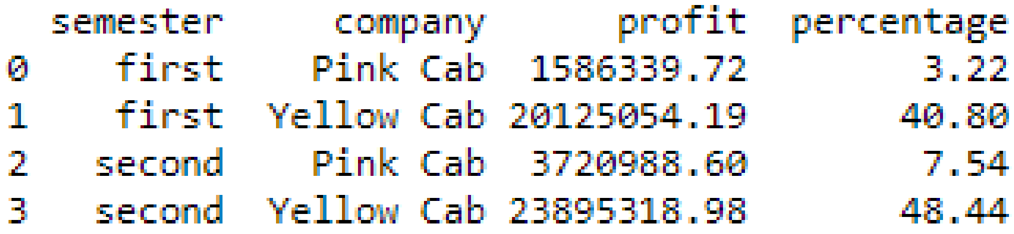




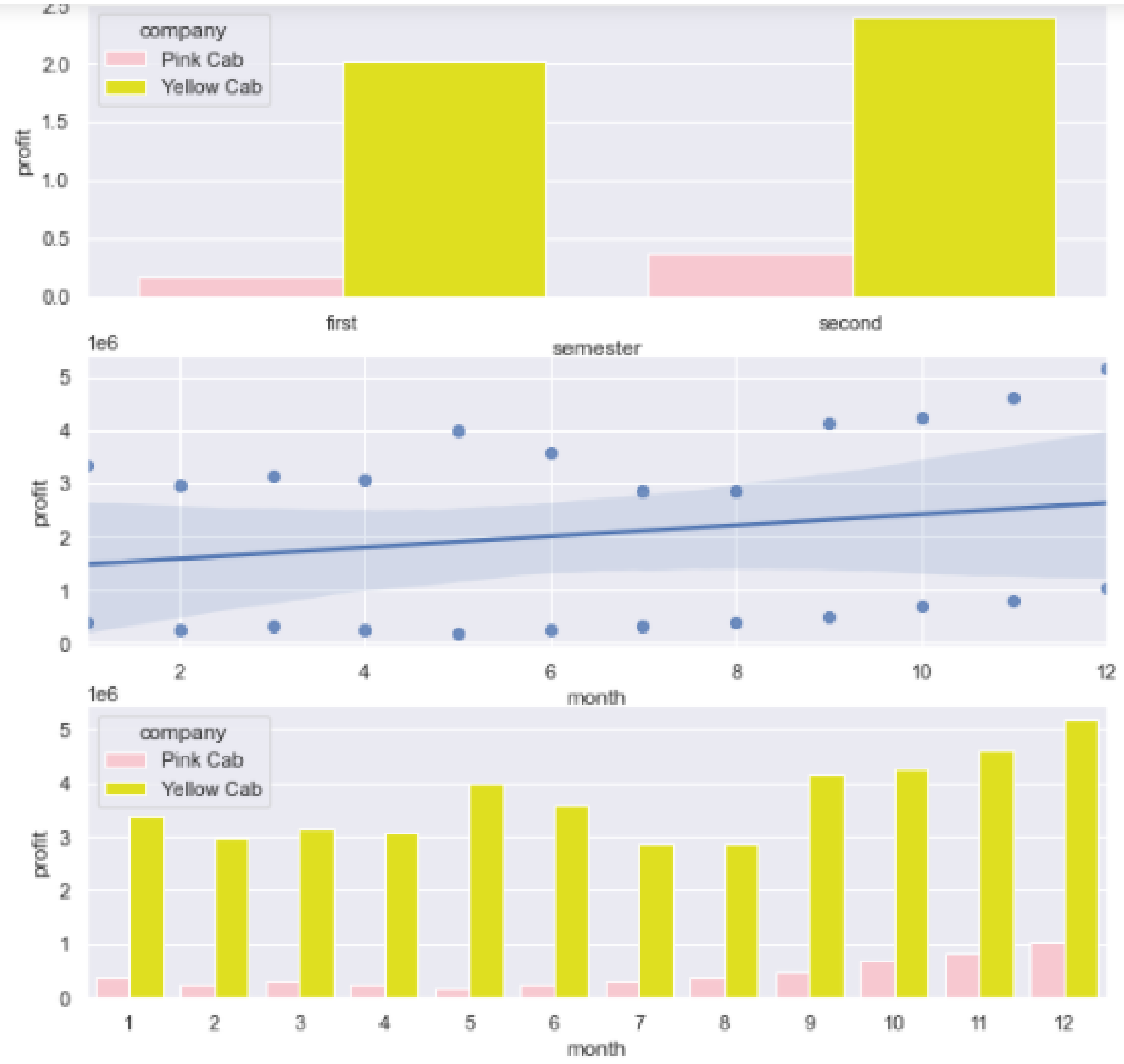
Hypothesis 7: Companies make more profit during the seconde semester of the year.

TRUE In total and in the second semester, the company Yellow Cab make $3770264.7852 (7.64%) more profit than the first semester.

And in the second semester, the company Pink Cab make $2134648.873 (4.33%) more profit than the first semester.

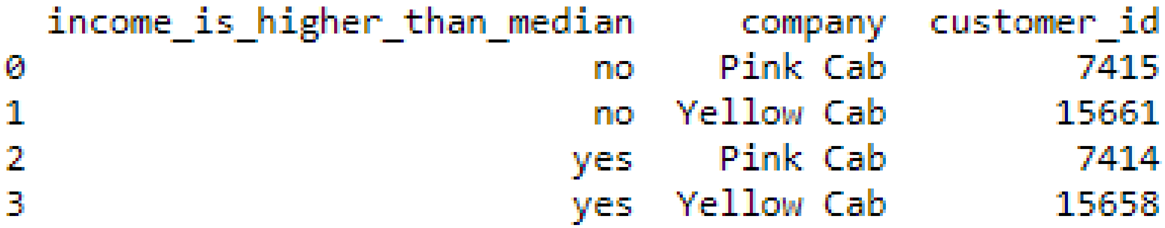


```
  semester     company       profit  percentage
0    first    Pink Cab   1586339.72        3.22
1    first  Yellow Cab  20125054.19       40.80
2   second    Pink Cab   3720988.60        7.54
3   second  Yellow Cab  23895318.98       48.44
```

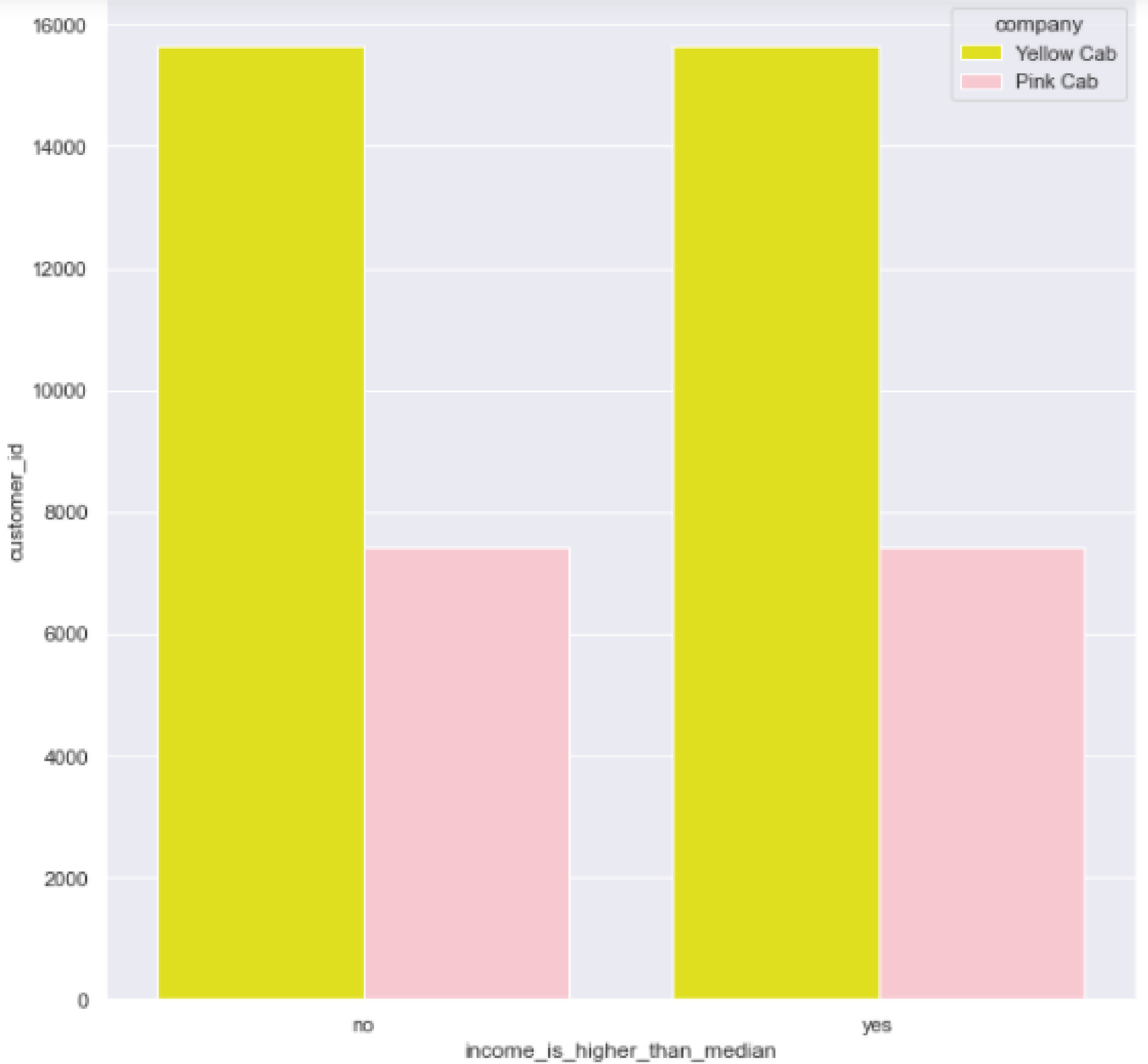


Hypothesis 8: People with the amount of income higher than the median uses more taxi.

FALSE People with the amount of income higher than the median uses less taxi than people which has income inferior than median, however this difference is minimal, which is 1 for Pink Cab company and 3 for Yellow Cab company.

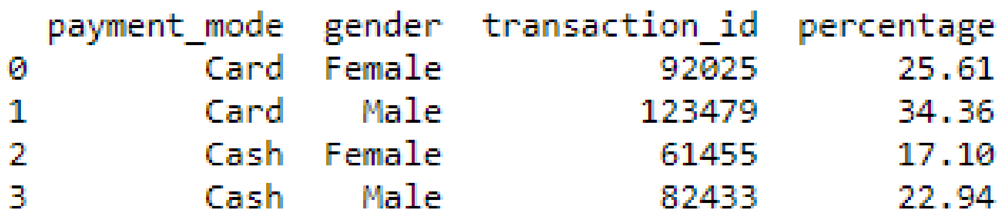


```
  income_is_higher_than_median     company  customer_id
0                           no    Pink Cab         7415
1                           no  Yellow Cab        15661
2                          yes    Pink Cab         7414
3                          yes  Yellow Cab        15658
```

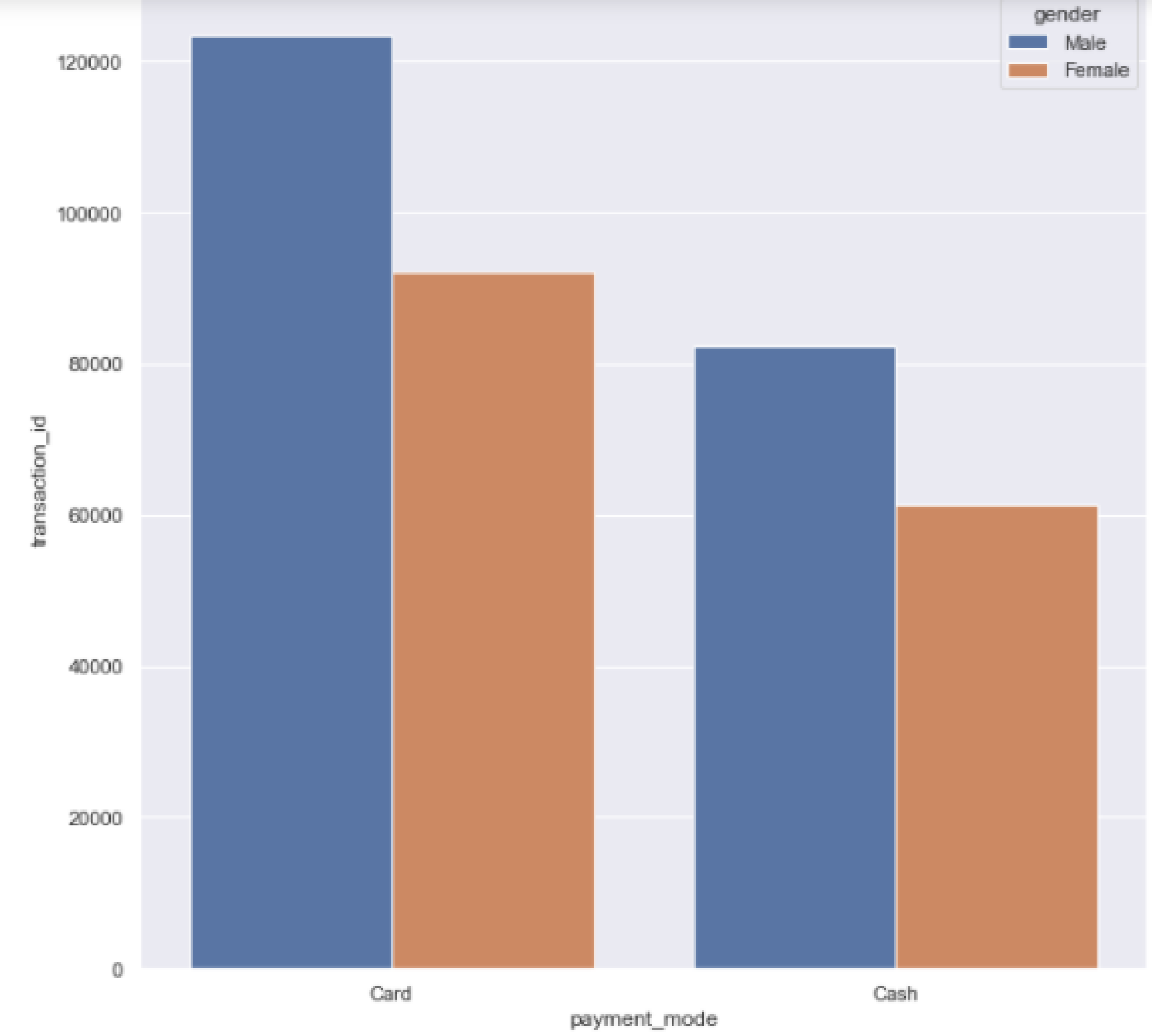


Hypothesis 9: Men use more cash than women.

TRUE In total, the number of men using cash is 20978 (5.84%) more than women.

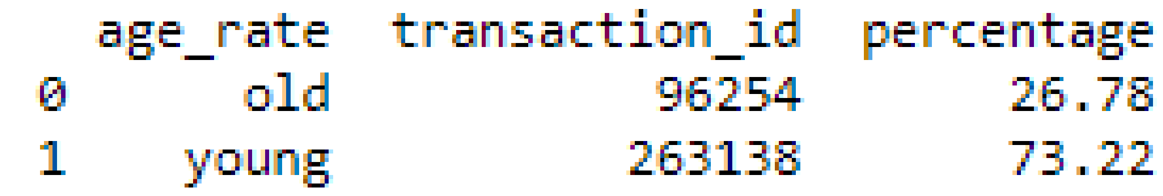


```
  payment_mode  gender  transaction_id  percentage
0         Card  Female           92025       25.61
1         Card    Male          123479       34.36
2         Cash  Female           61455       17.10
3         Cash    Male           82433       22.94
```

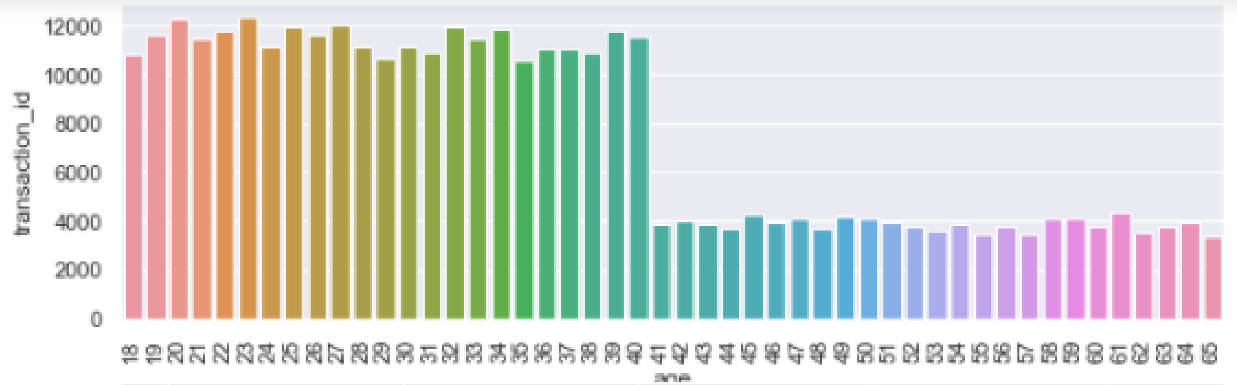


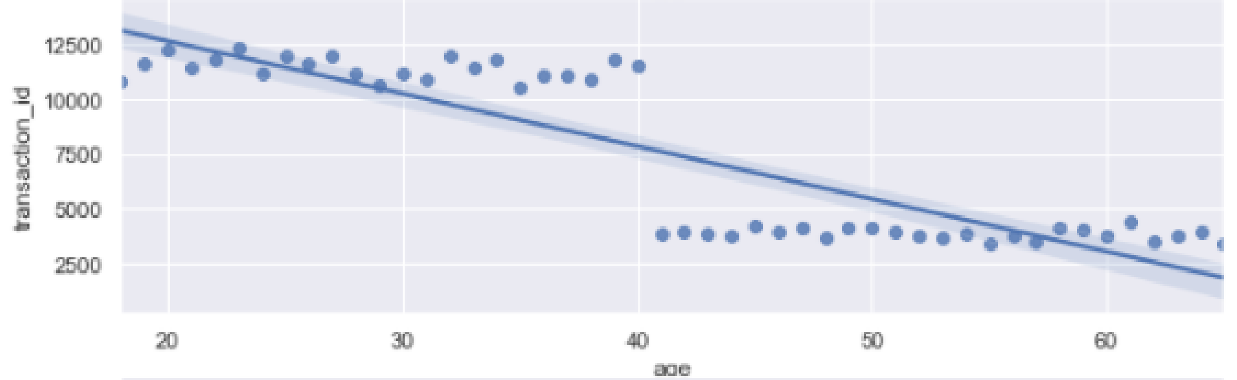
Hypothesis 10: Young people use more taxi than old people.

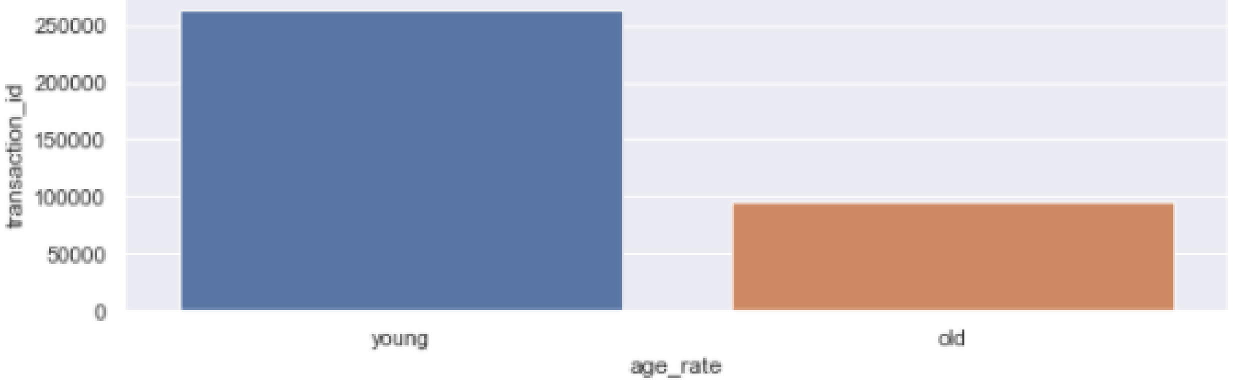
TRUE In total, there are 166884 (46.44%) young people more using taxi than old people.



```
   age_rate  transaction_id  percentage
0       old           96254       26.78
1     young          263138       73.22
```







Final Results

Yellow Cab Company has that percentage higher than Pink Cab Company:

- Number of taxi trips (52.86%);
- Profit (78.48%);
- Number of customers (35.73%).

Conclusion

In conclusion, Yellow Cab Company has a higher number of taxi trips than Pink Cab Company, as well as a higher profit and higher number of customers.

The data also show that Yellow Cab has superior results in almost all cities, especially in New York, which is the city with the highest profit and higher number of taxi trips.

The same company also stands out from the other company in relation to the season, where the fall or spring (depending which measure is more important) season shows the highest profit and number of taxi trips as well.

Also other insights provided by the hypotheses, inform that Yellow Cab Company present better results than Pink Cab Company, however some of them shows that the number of taxi trips, the profit and the number of customers, for both Companies, are decreasing over the year, giving space for other Cab companies.

Due those all information, it is recommended to invest in Yellow Cab Company.

Next Steps

Other project that can be made with this dataset is the exploration data analyses, which identify the best's attributes in order to apply machine learning algorithms, with the objective to predict the number of taxi trips, the profit or the number of customers for each company.

Thank You

Data Glacier

Your Deep Learning Partner