Exploratory Data Analysis

G2M insight for Cab Investment firm

21/10/2022

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Executive Summary

- The investor : XYZ Company
- Investment category : Cab Investment
- The investment on : Pink Cab or Yellow Cab
- Strategy : G2M strategy
- Location: the US

The Problem

- Cab market is a fast growing market
- We want to understand the market before XYZ executives take final decision.

Approach – Data Understanding

A glimpse on the provided data

	City	Population	Users
0	NEW YORK NY	8,405,837	302,149
1	CHICAGO IL	1,955,130	164,468
2	LOS ANGELES CA	1,595,037	144,132
3	MIAMI FL	1,339,155	17,675
4	SILICON VALLEY	1,177,609	27,247
5	ORANGE COUNTY	1,030,185	12,994
6	SAN DIEGO CA	959,307	69,995
7	PHOENIX AZ	943,999	6,133
8	DALLAS TX	942,908	22,157

	Transaction ID	Customer ID	Payment_Mode
0	10000011	29290	Card
1	10000012	27703	Card
2	10000013	28712	Cash
3	10000014	28020	Cash
4	10000015	27182	Card
...	...	...	...
440093	10440104	53286	Cash
440094	10440105	52265	Cash
440095	10440106	52175	Card
440096	10440107	52917	Card
440097	10440108	51587	Card

	Customer ID	Gender	Age	Income (USD/Month)
0	29290	Male	28	10813
1	27703	Male	27	9237
2	28712	Male	53	11242
3	28020	Male	23	23327
4	27182	Male	33	8536
...	...	...	...	...
49166	12490	Male	33	18713
49167	14971	Male	30	15346
49168	41414	Male	38	3960
49169	41677	Male	23	19454
49170	39761	Female	32	10128

49171 rows × 4 columns

Approach – Data Understanding

A glimpse on the provided data

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip
0	10000011	2016-01-08	Pink Cab	ATLANTA GA	30.45	370.95	313.6350
1	10000012	2016-01-06	Pink Cab	ATLANTA GA	28.62	358.52	334.8540
2	10000013	2016-01-02	Pink Cab	ATLANTA GA	9.04	125.20	97.6320
3	10000014	2016-01-07	Pink Cab	ATLANTA GA	33.17	377.40	351.6020
4	10000015	2016-01-03	Pink Cab	ATLANTA GA	8.73	114.62	97.7760
...	...	...	...	...	...	...	...
359387	10440101	2018-01-08	Yellow Cab	WASHINGTON DC	4.80	69.24	63.3600
359388	10440104	2018-01-04	Yellow Cab	WASHINGTON DC	8.40	113.75	106.8480
359389	10440105	2018-01-05	Yellow Cab	WASHINGTON DC	27.75	437.07	349.6500
359390	10440106	2018-01-05	Yellow Cab	WASHINGTON DC	8.80	146.19	114.0480
359391	10440107	2018-01-02	Yellow Cab	WASHINGTON DC	12.76	191.58	177.6192

Approach – Data Understanding

A glimpse on the data which could be found also online

	Type	Date of Travel	City
0	Cold	2016-01-08	NEW YORK NY
1	Cold	2016-01-09	NEW YORK NY
2	Rain	2016-01-09	NEW YORK NY
3	Rain	2016-01-10	NEW YORK NY
4	Rain	2016-01-10	NEW YORK NY
...	...	...	...
52008	Snow	2018-12-24	EAST BOSTON MA
52009	Snow	2018-12-28	EAST BOSTON MA
52010	Rain	2018-12-28	EAST BOSTON MA
52011	Rain	2018-12-28	EAST BOSTON MA
52012	Rain	2018-12-31	EAST BOSTON MA

	Date of Travel	Holiday
0	2016-07-04	4th of July
1	2017-07-04	4th of July
2	2018-07-04	4th of July
3	2016-12-25	Christmas Day
4	2017-12-25	Christmas Day
5	2018-12-25	Christmas Day
6	2016-12-24	Christmas Eve
7	2017-12-24	Christmas Eve
8	2018-12-24	Christmas Eve
9	2018-10-08	Columbus Day

Approach- Data Understanding Summary

- We are Provided with four files containing data
- We will also leverage some of our analysis on two extra datasets available online(US weather data and US holidays data)
- Considering dataset combined of our four main datasets, we will have a dataset with 14 features 359392 rows
- Extra features can be deduced based on our main features later on upon need.
- The main dataframe will look as the next slide:

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip	Customer ID	Payment_Mode	Gender	Age	Income (USD/Month)	Population	Users
0	10000011	2016-01-08	Pink Cab	ATLANTA GA	30.45	370.95	313.6350	29290	Card	Male	28	10813	814,885	24,701
1	10351127	2018-07-21	Yellow Cab	ATLANTA GA	26.19	598.70	317.4228	29290	Cash	Male	28	10813	814,885	24,701
2	10412921	2018-11-23	Yellow Cab	ATLANTA GA	42.55	792.05	597.4020	29290	Card	Male	28	10813	814,885	24,701
3	10000012	2016-01-06	Pink Cab	ATLANTA GA	28.62	358.52	334.8540	27703	Card	Male	27	9237	814,885	24,701
4	10320494	2018-04-21	Yellow Cab	ATLANTA GA	36.38	721.10	467.1192	27703	Card	Male	27	9237	814,885	24,701
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
359387	10307228	2018-03-03	Yellow Cab	WASHINGTON DC	38.40	668.93	525.3120	51406	Cash	Female	29	6829	418,859	127,001
359388	10319775	2018-04-13	Yellow Cab	WASHINGTON DC	3.57	67.60	44.5536	51406	Cash	Female	29	6829	418,859	127,001
359389	10347676	2018-07-06	Yellow Cab	WASHINGTON DC	23.46	331.97	337.8240	51406	Card	Female	29	6829	418,859	127,001
359390	10358624	2018-08-02	Yellow Cab	WASHINGTON DC	27.60	358.23	364.3200	51406	Cash	Female	29	6829	418,859	127,001
359391	10370709	2018-08-30	Yellow Cab	WASHINGTON DC	34.24	453.11	427.3152	51406	Card	Female	29	6829	418,859	127,001

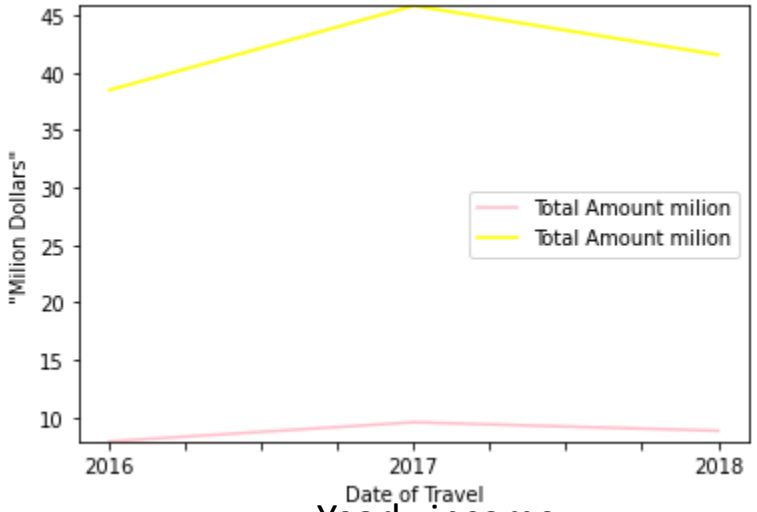
Approach

- Pink Cab and Yellow Cab comparison will leverage on these Data
- Some Hypothesis will be checked to evaluate profitability of either of companies based on EDA
- Some assumption will be made in each of evaluation

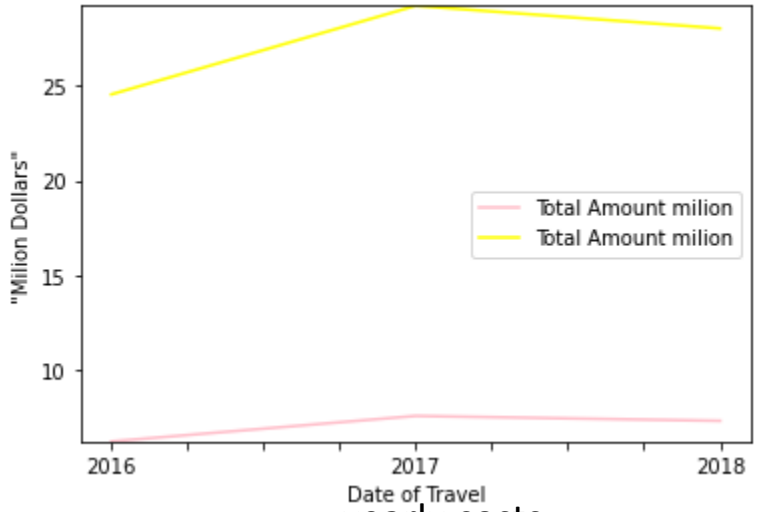
EDA end Hypothesis evaluations

1st Hypothesis : Does a company receive larger profit than the other?

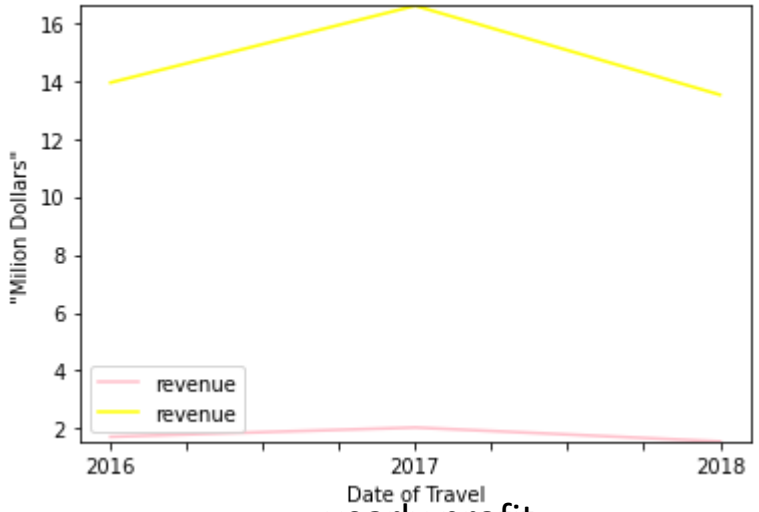
- The first point that may cross to everybody's mind is to compare the monthly and yearly costs , incomes and revenue(profit) of each individual company.



Yearly income



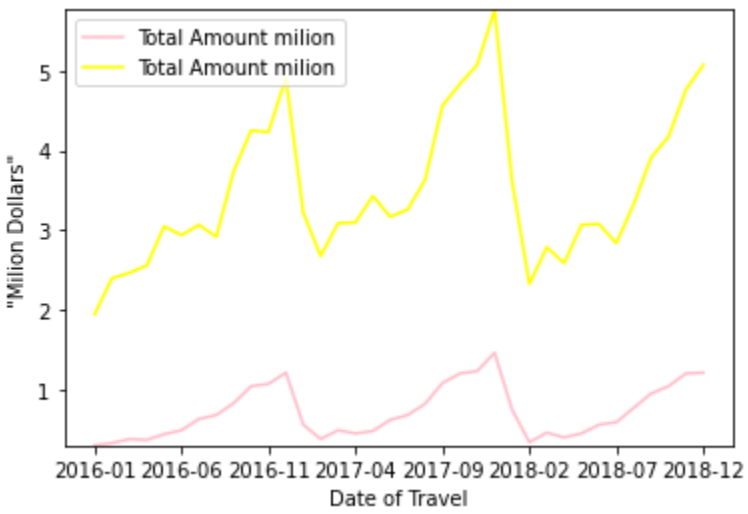
yearly costs



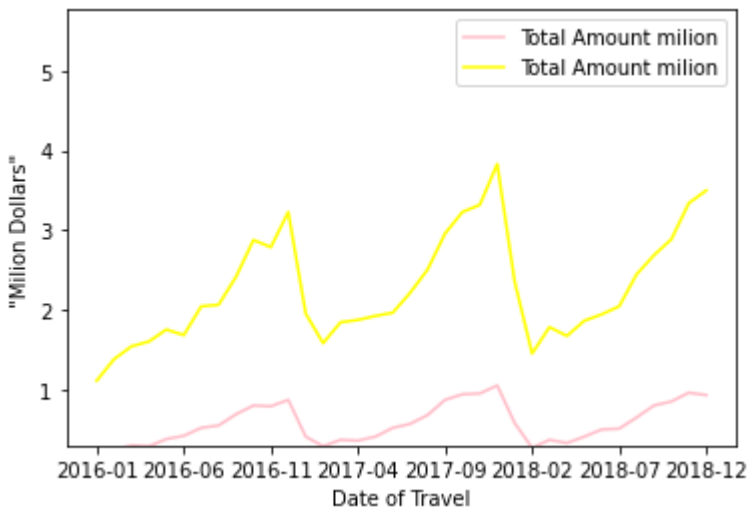
yearly profit

- This shows a meaningful difference in Pink and Yellow Cab incomes where both reach to a peak in 2017 and Yellow Cab is capturing 5 times more than the Pink Cab
- When it comes to yearly costs , Yellow cab has also around 5 times more expenses than Pink Cab
- what is clear is that the Yellow cab is almost 7 times larger in term of profit

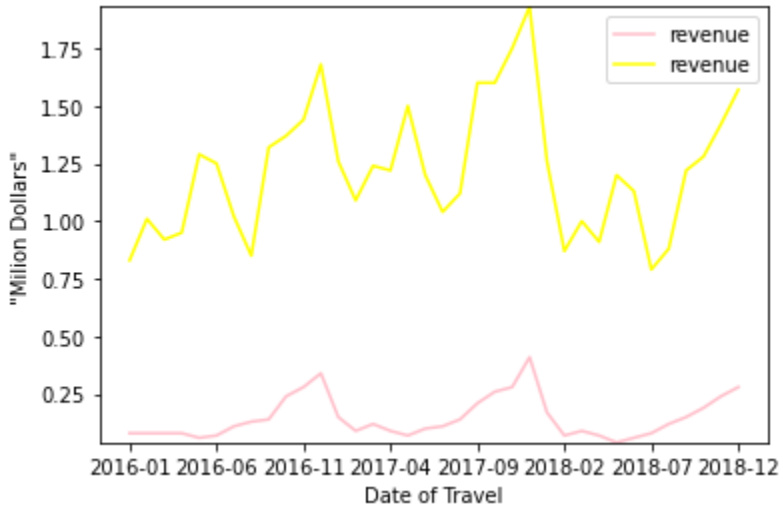
The evaluations can be done also on a monthly basis:



Monthly income



Monthly costs

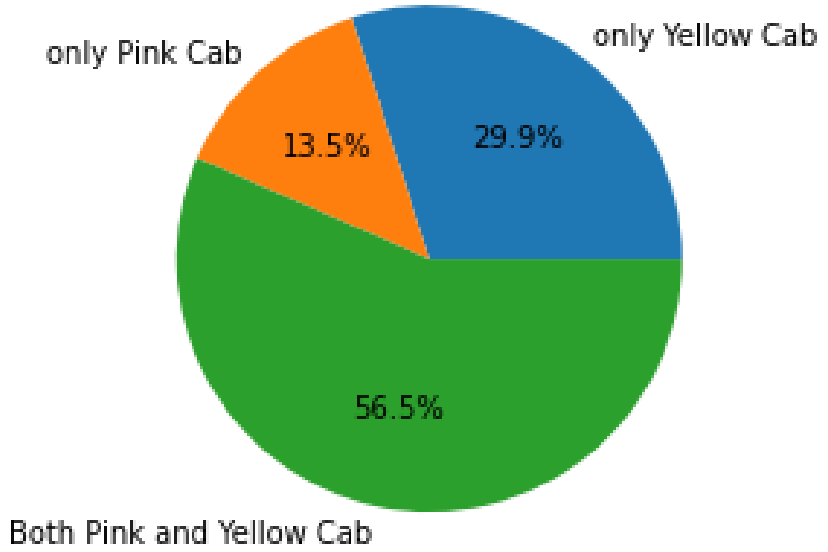


Monthly profit

The monthly evaluations shows that The yellow cab is still profiting well and ithe Pink Cab reaches to almost a very low revenue on some months while the Yellow cab was always better Profitting during these period of investigation

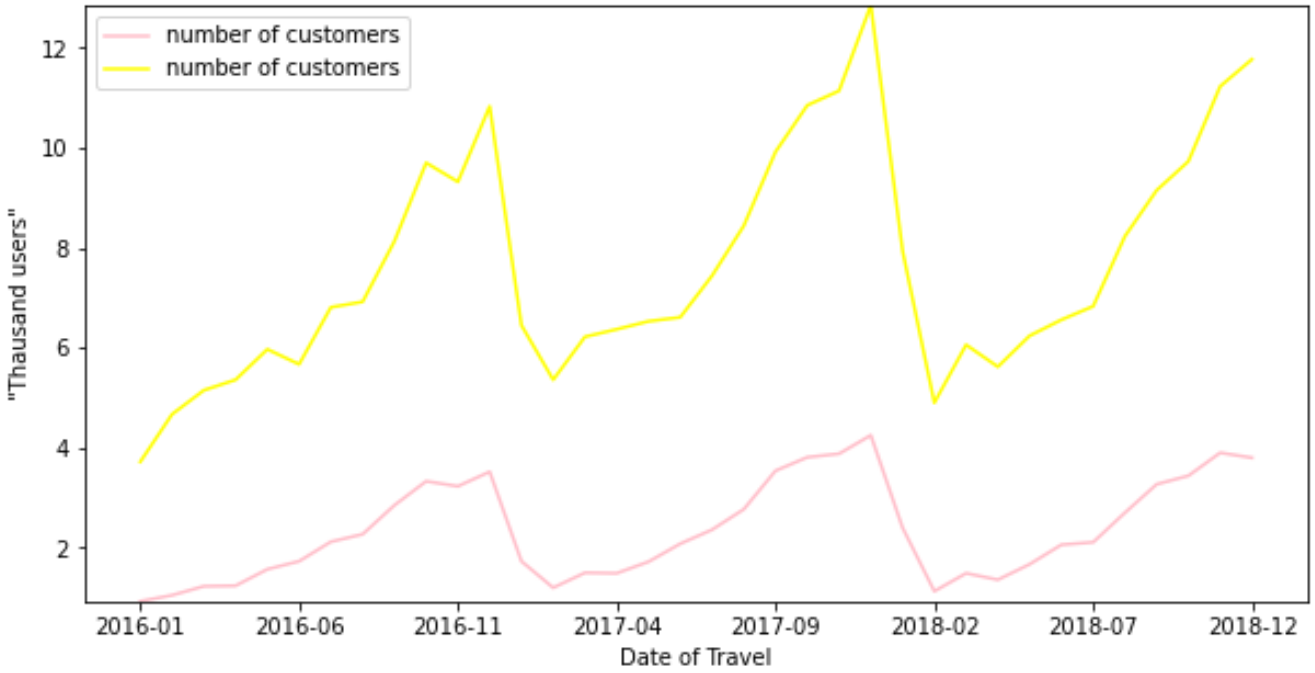
2nd Hypothesis : Are there Customers who Prefer one Company to another?

- knowing one company is more profiting than another can depend on many factor, but one important factor is how many users it has comparing to its rivals.Lets see which company has the maximum cab users. Number of unique users in entire Pink and Yellow Database are: 46148 and based on the following plot:



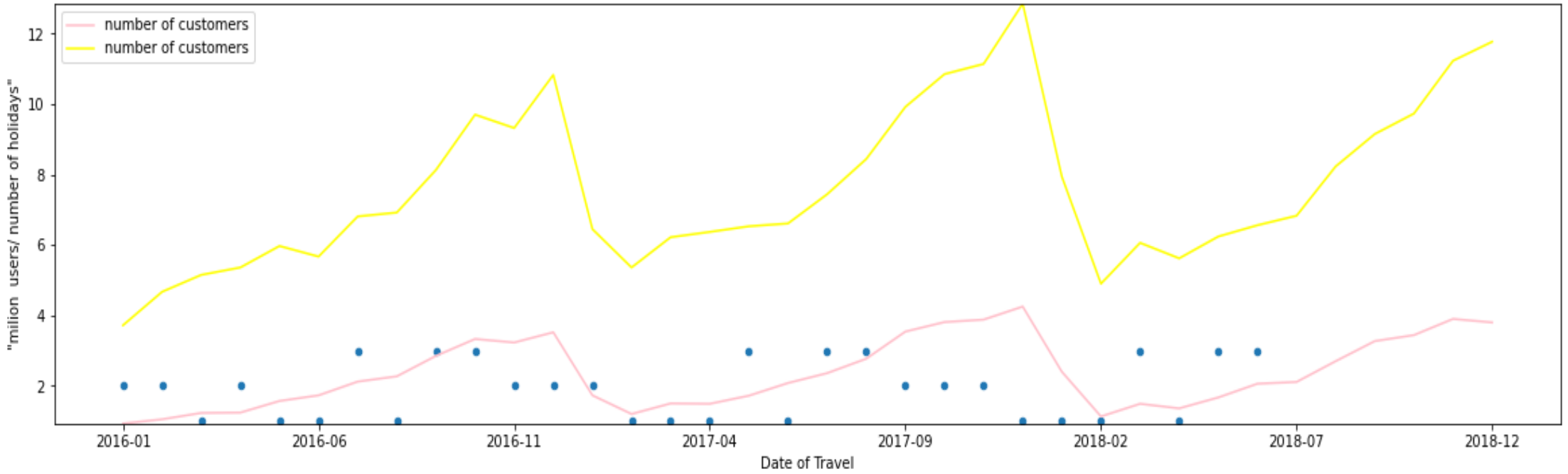
It turns out that Pink cab has around half of unique users of the yellow cab's unique users (users who do not use the other company) and obviou sly it can be deduced that the Yellow Cab has more users, but the difference in number of unique users is not too much to contribute in more profit of the Yellow cab. In other words, there are users of Pink Cab that are more willing to use the Yellow cab while they are also Pink Cab customers, but there must be something more that makes Yellow cab more profiting ,Lets investigate more

3rd Hypothesis : Is the use of the Cab companies consistent during different months of the year ?



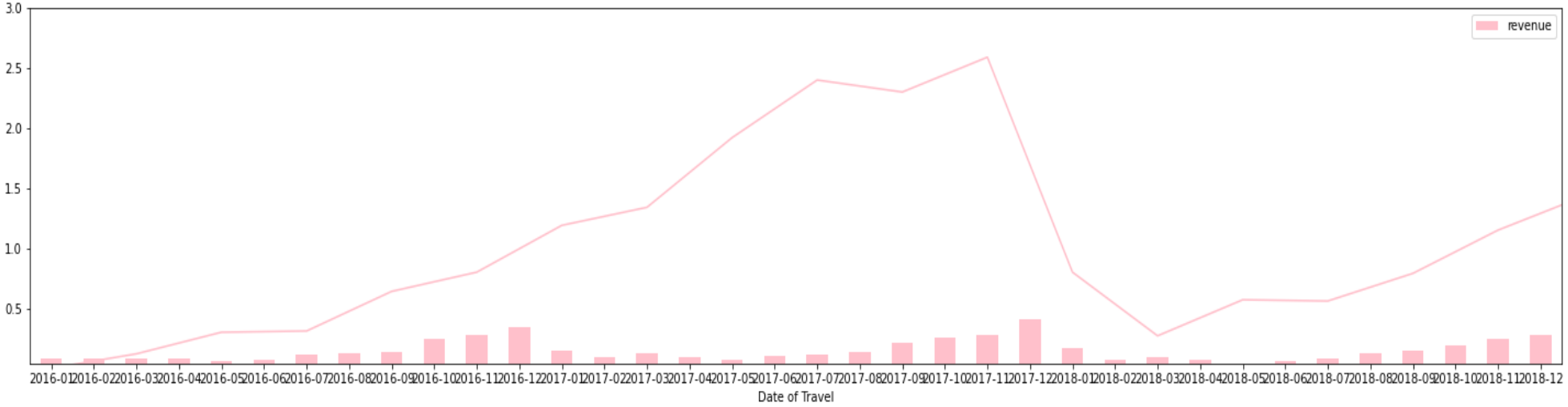
It is clear that the number of customers of each cab company fluctuates with almost the same pattern in different months of the years

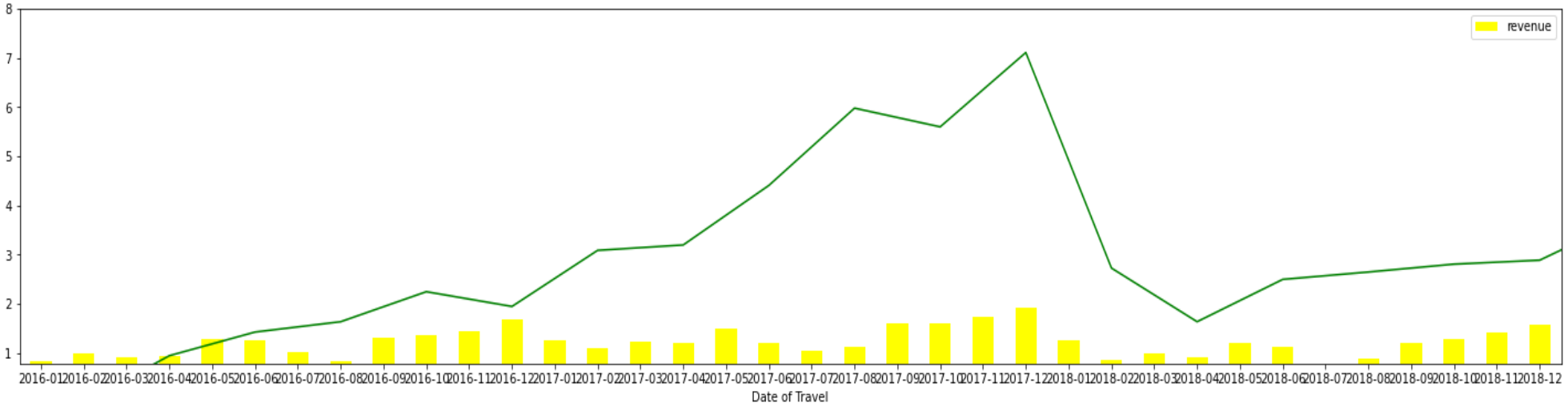
4th Hypothesis : Do Holidays have any effect in number of travels made by customers in the specifc time periods?



as shown the graph, having dots as number of holidays in that month it seems that there is a relation in the number of holidays in month and the increase in users of the both companies. the closer and more the holidays are, the number of users of the both companies tend to use them more and they are affected with almost the same pattern.

5th Hypothesis : Does margin proportionally increase with increase in number of customers?





- comparing each company's monthly income (bar) with number of their monthly users(line): as it is expected
- number of rides have relation with amount of revenue

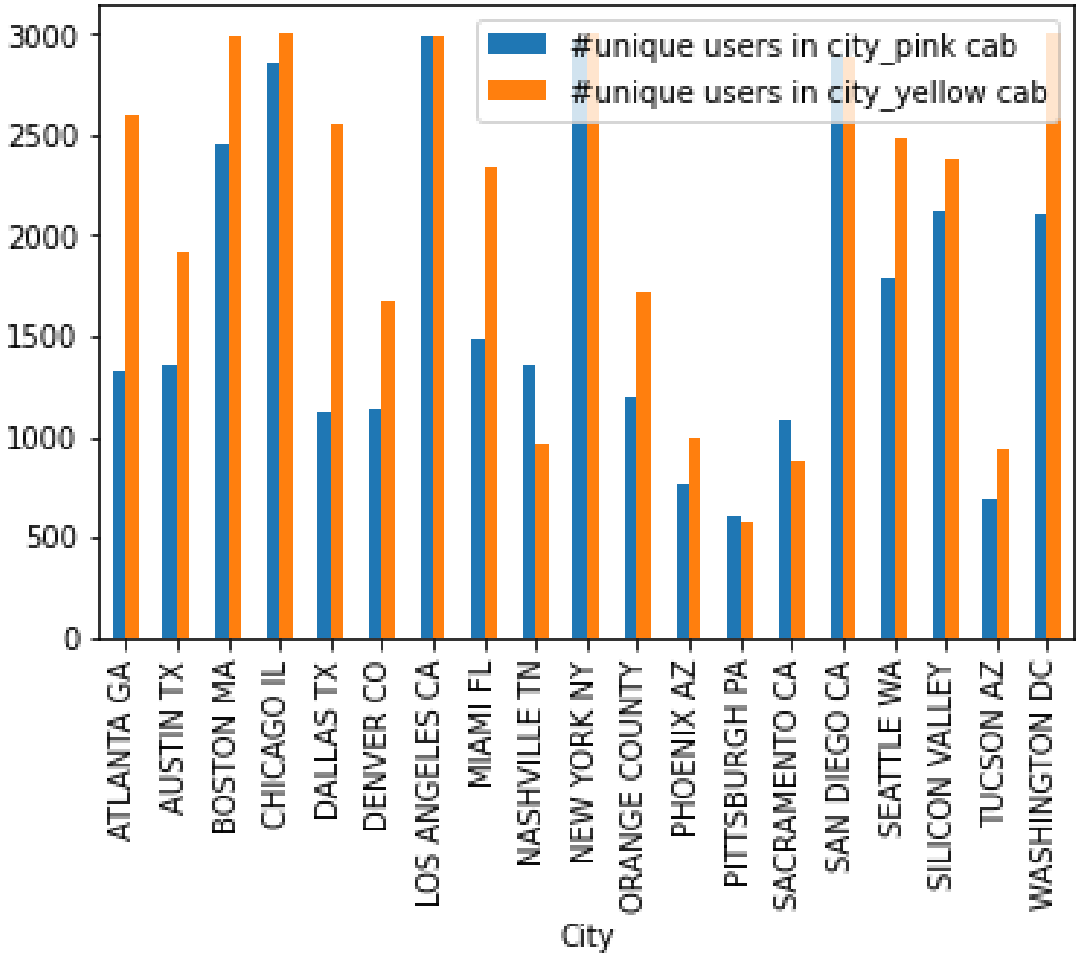
6th Hypothesis: how different attributes of the customer segments affect the business?

- different examples such as is Pink Cab invested enough in other cities ? for example , is all Pink Cab's revenue from one city and not invested enough in other cities ?

- or is there any gender/Age group that is not willing to use the the company ? (are genders equally divided for each company ? (evaluation based on unique users not number of rides)

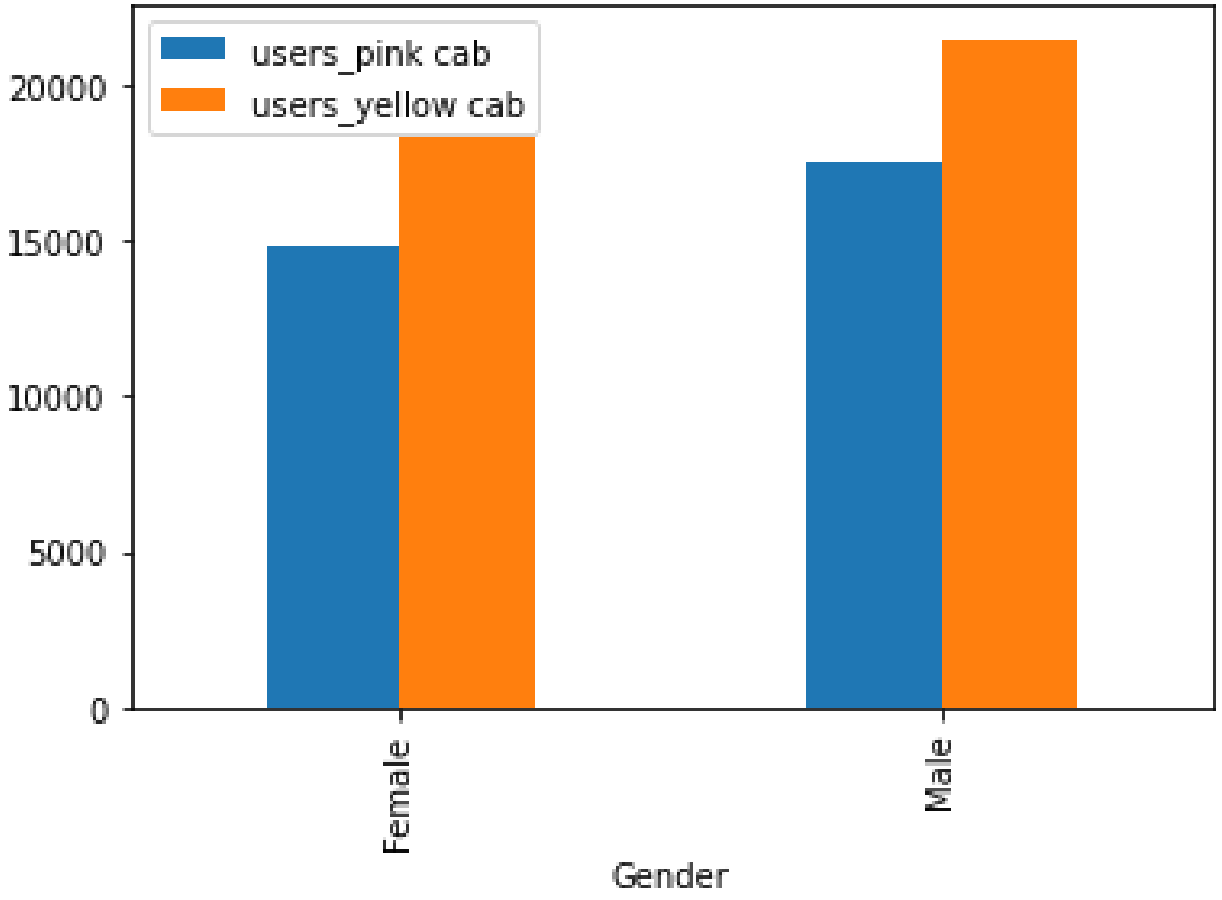
- or is there any problem related to Payment Mode ? do People use the same proportion of payment Mode for both companies ?

which company is leading in different cities?



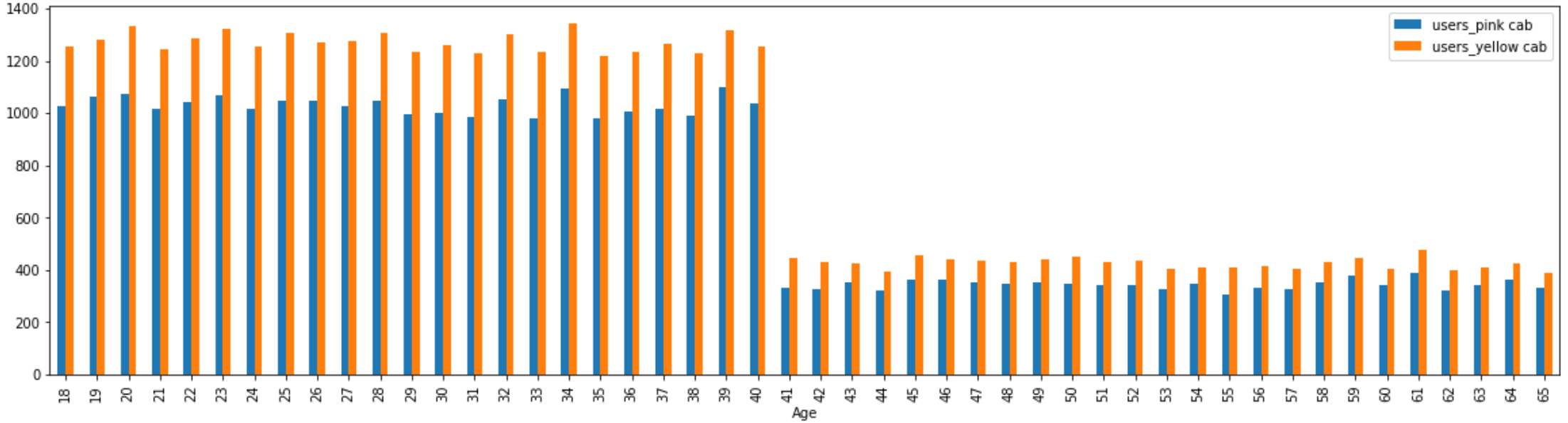
It turns out that Pink Cab is having almost as unique users as Yellow cab in cities like Chicago ,LA,NY, and San Diego and even slightly more users in Yellow cab in some cities Sacramento and Nashvile , but less unique users in other cities.
From this, one can deduce that people are less willing to comeback to Pink cab despite they know this company

what about different genders?



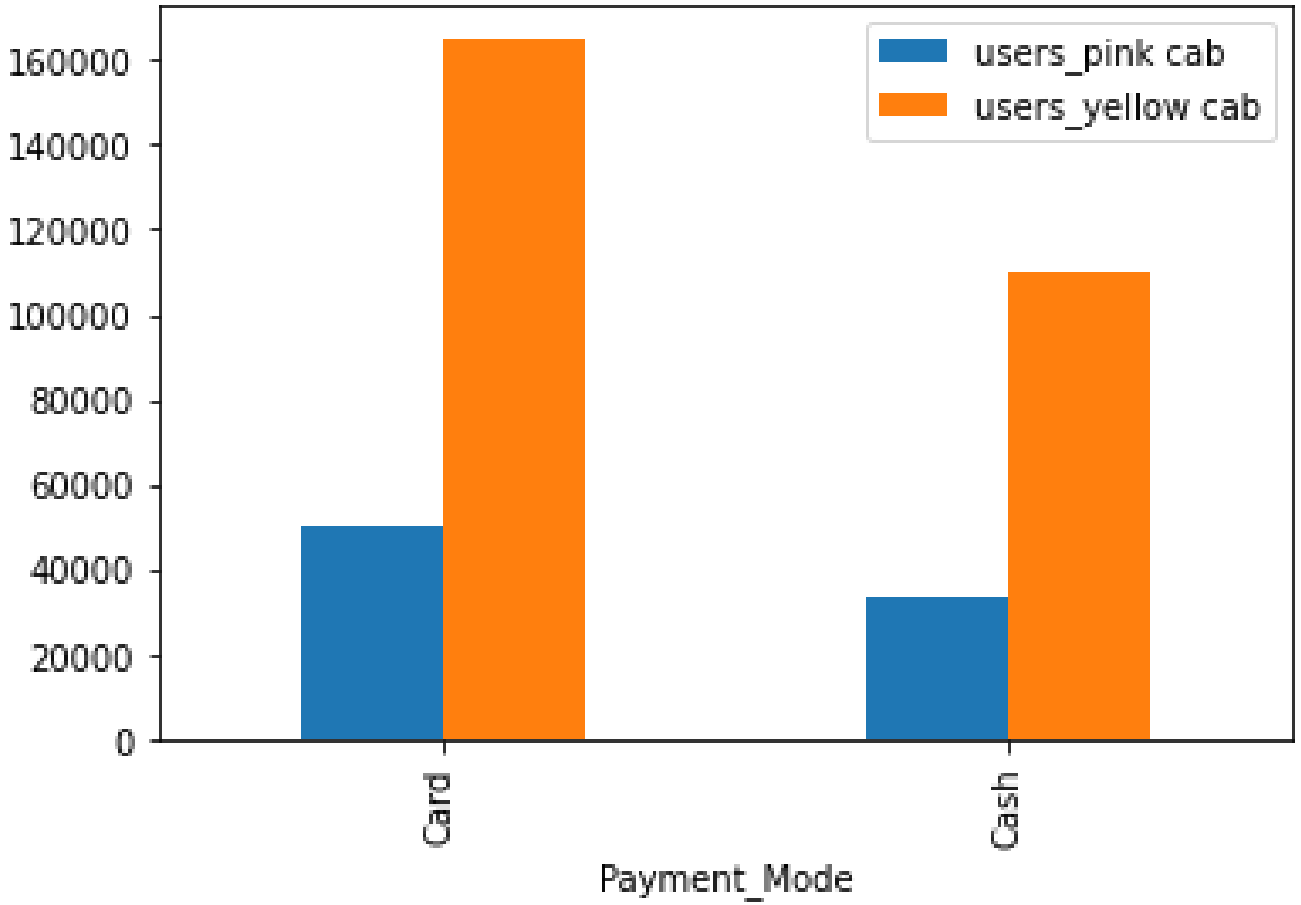
gender gap is not a big deal for both company as the plot shows that both male and female users show the same propotional interest to both companies. in other words for each company, there is no gap in number of users (though male users in both companies are slightly more)

Different ages



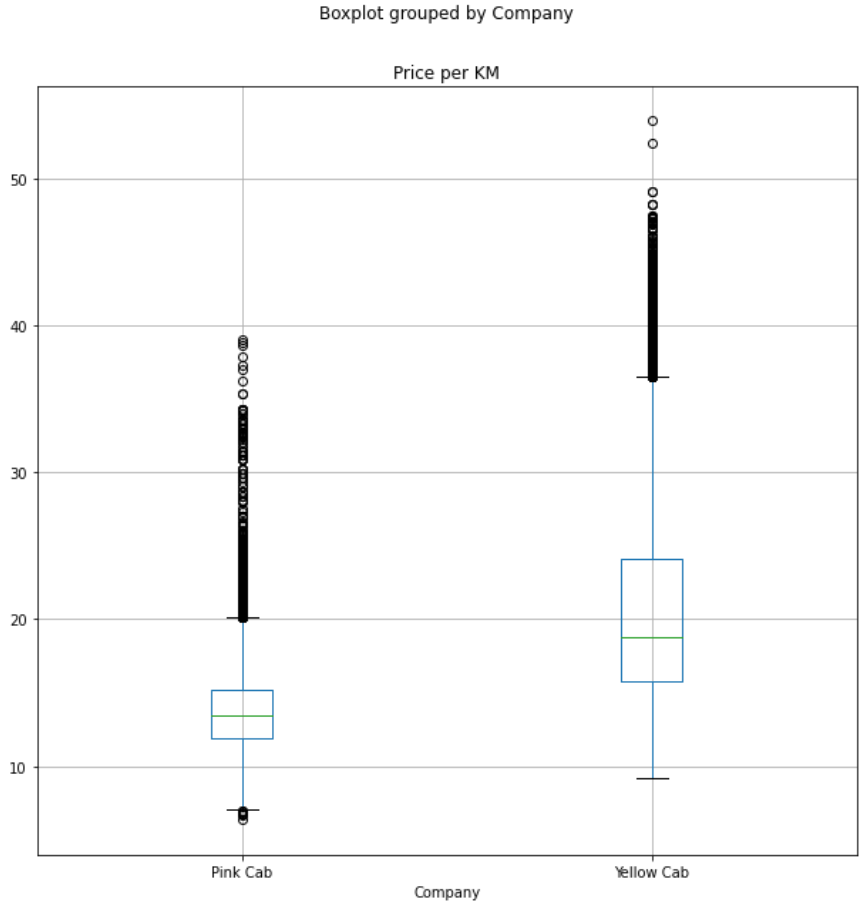
also different age groups are treating both companies almost the same way and people younger than 40 years old are leading in both companies

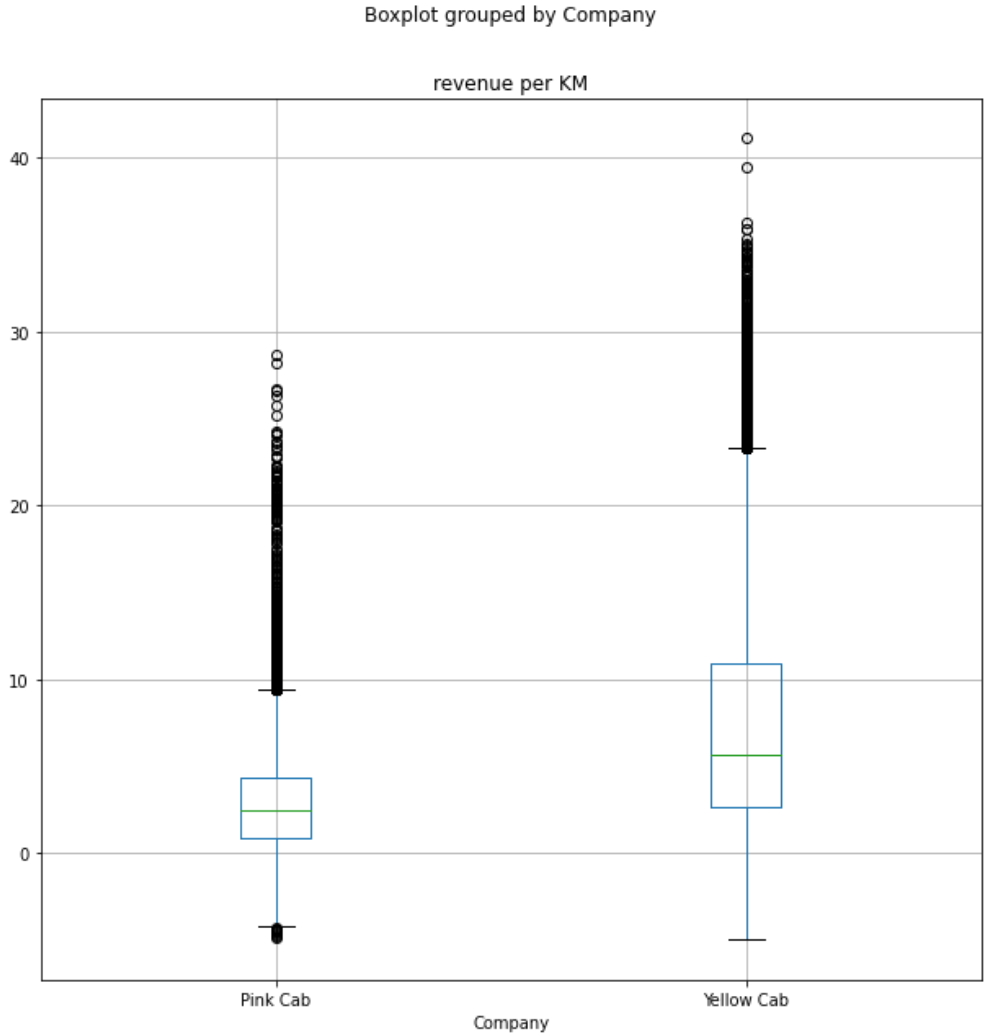
Payment methods



- and finally as the following plot shows , the payments made with the credit cards are preferred in both companies , and as it is expected pink cab transactions made with bode card and cash are less than Yellow cabs

7th Hypothesis : does Yellow have lower/higher price per KM which results in more profit due to its more customers ?

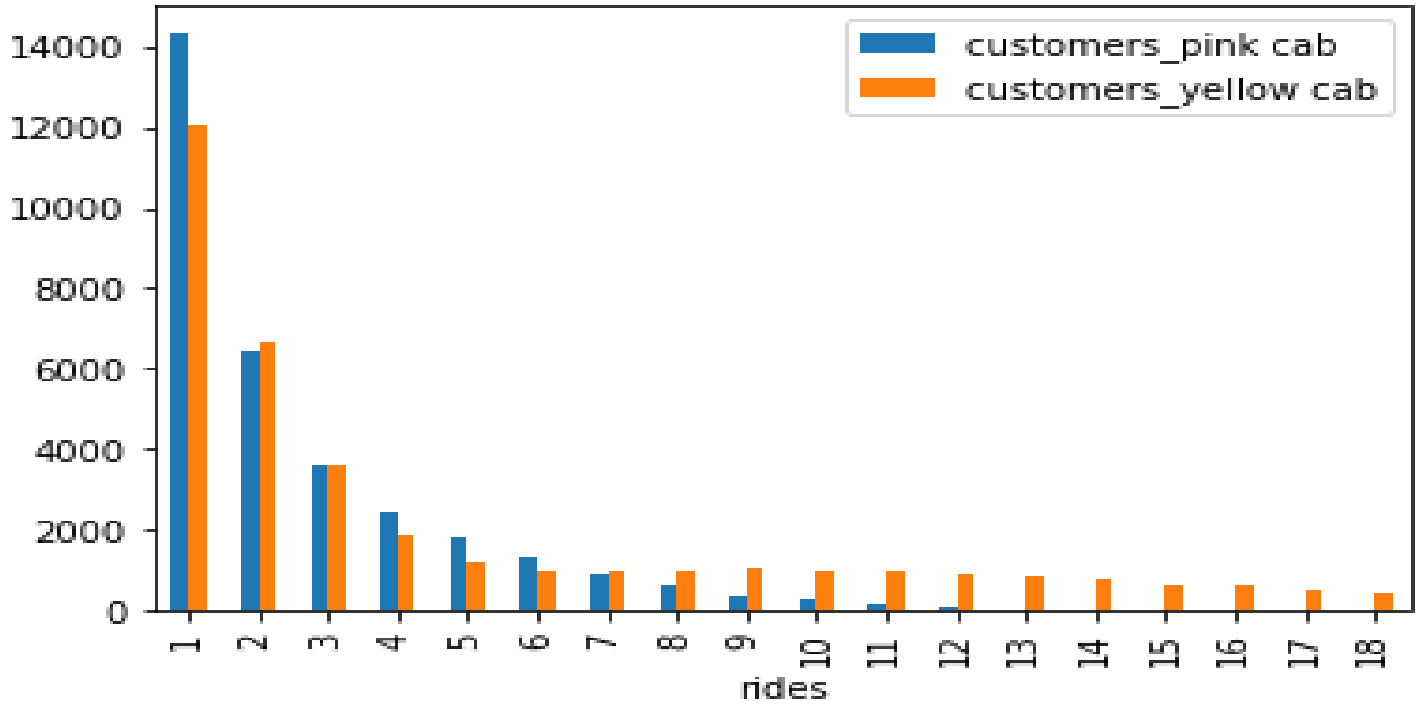




7th Hypothesis : does Yellow have lower/higher price per KM which results in more profit due to its more customers ?

- it turns out that despite having higher price per KM and consequently higher revenue per KM for Yellow cab, people still prefer to use the Yellow cab which is interesting. so the less prices of Pink cab has not attracted its customers enough to use Pink Cab more. .
- Note: however this should be considered into account that Pink cab might be less used due to unavailability or lack of taxis comparing to its competing company, but this issue cannot be investigated based on the information we have.

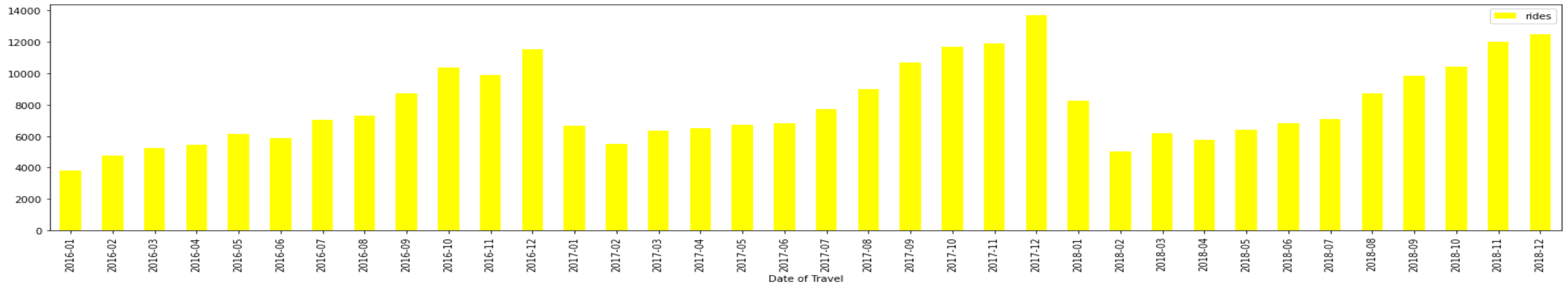
Hypothesis 8 : one company has more loyal customers than the other?

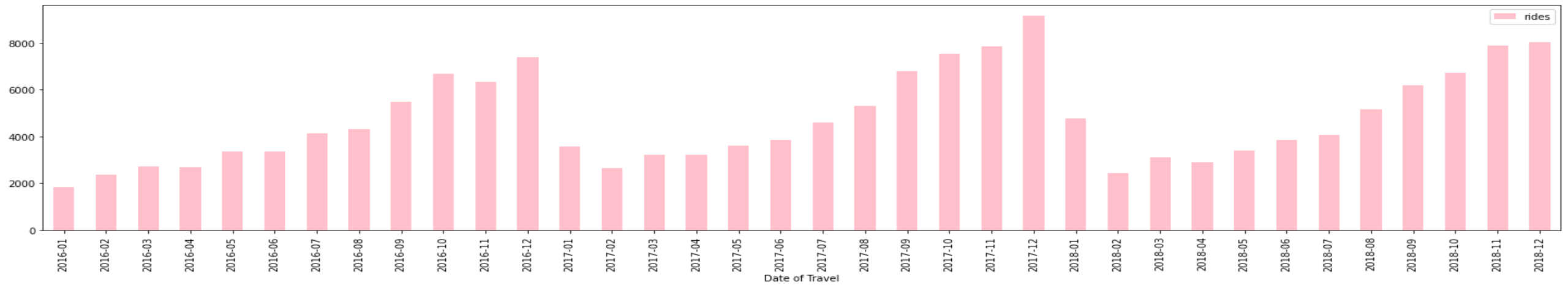


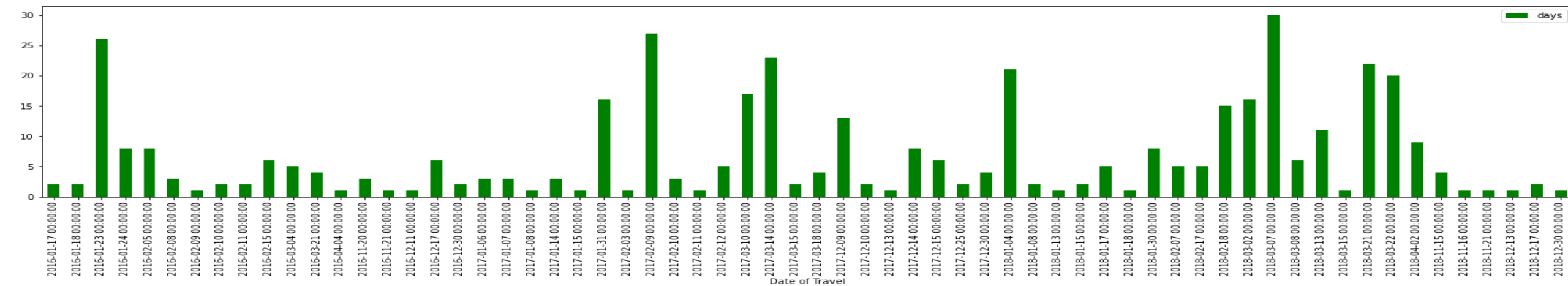
- it is also interesting to check the customers loyalty considering how many customers return to the cab companies for their rides. it is interesting that as the below plot illustrates , despite both companies serve major of their customers with only single ride, and even pink cab excels yellow cab in single rides, but yellow cab is more fortunate to have their customers back for their futures rides. to be more specific , customers are tend to use Yellow cab for their 12th ride and more, while Pink cab is not fortunate to serve customers for more rides.

9th hypothesis: Are the rides seasonal in cities depending on the weather condition ?

- lets evaluate if seasonality in a city such as NY where both companies perform the same in term of number of unique users have any impact. from the plots its clear that rides for both companies in new York are seasonal(fluctuates) but depending on the weather condition , it seems people use cabs when the weather is snowy.


rides
Date of Travel


rides
Date of Travel


days
Date of Travel

EDA summary and recomendation

- Finally based on the above finding It is recommended to the XYZ company to invest on Yellow cab, as :
- it has more loyal customers ,
- is more profiting ,
- already more popular in most of the major cities
- having more customers despite its higher prices per KM

Thank You