



Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

G2M Cab Investment Project

Team Muadh

Muadh Faizan

muadhfaizan@gmail.com

Data Science

20 July 2021

Agenda

Introduction

Project Structure

Assumptions

EDA





Recommendations

Further Insights

Introduction

An investment firm wanted to invest in the cab industry, which comprised of two taxi companies. They needed a study to be conducted, in order to analyse and find out which of the companies would be the more profitable investment.

4 datasets were provided:

-  Transaction data from 2016 to 2018 with details of trip fare, date of trip, city etc.
-  Customer ID and corresponding age, gender and income of customer.
-  Transaction ID with the corresponding customer ID for each transaction.
-  City data with some population data of each city.

Project Structure

The project was divided into 4 parts:

1. Understanding the structure of datasets.
2. Data processing and merging of datasets.
3. Exploratory Data Analysis
4. Findings and Recommendations

The datasets were merged into one master dataset, apart from the city dataset, as I did not use it in the analysis.

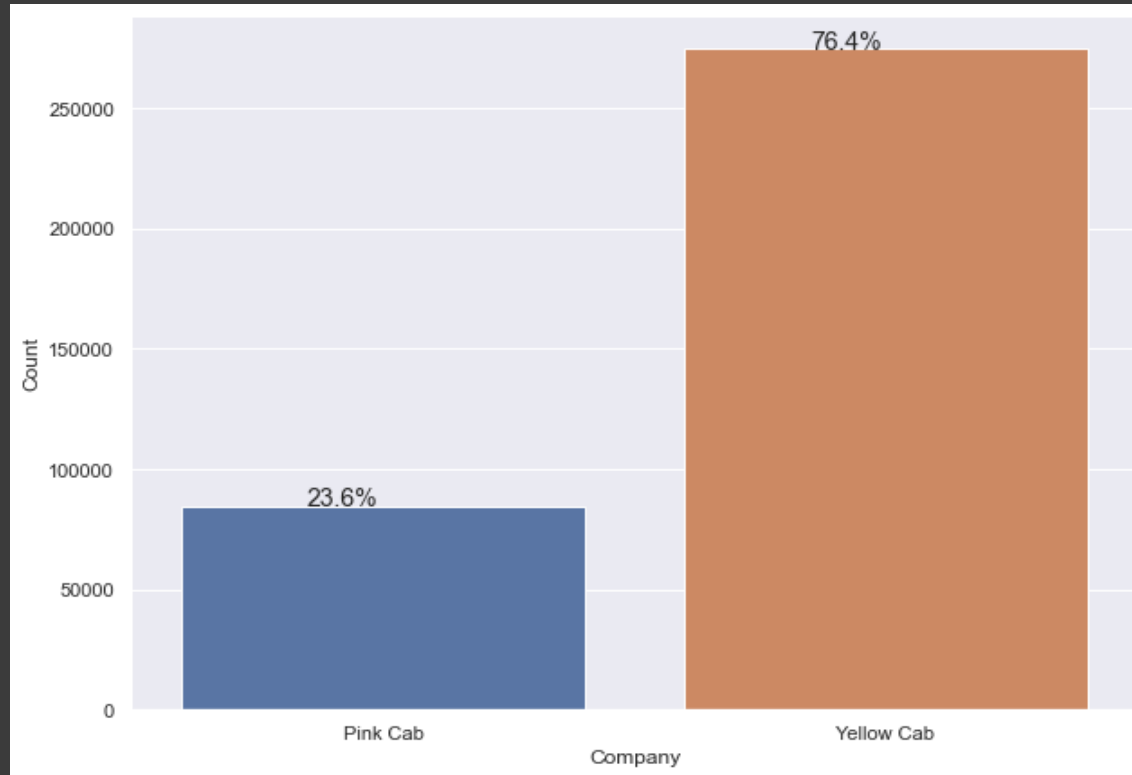
One additional dataset was imported from Kaggle, which was the US Weather Data which provided some information on the weather each day. This was also merged into the master dataset.

Assumptions

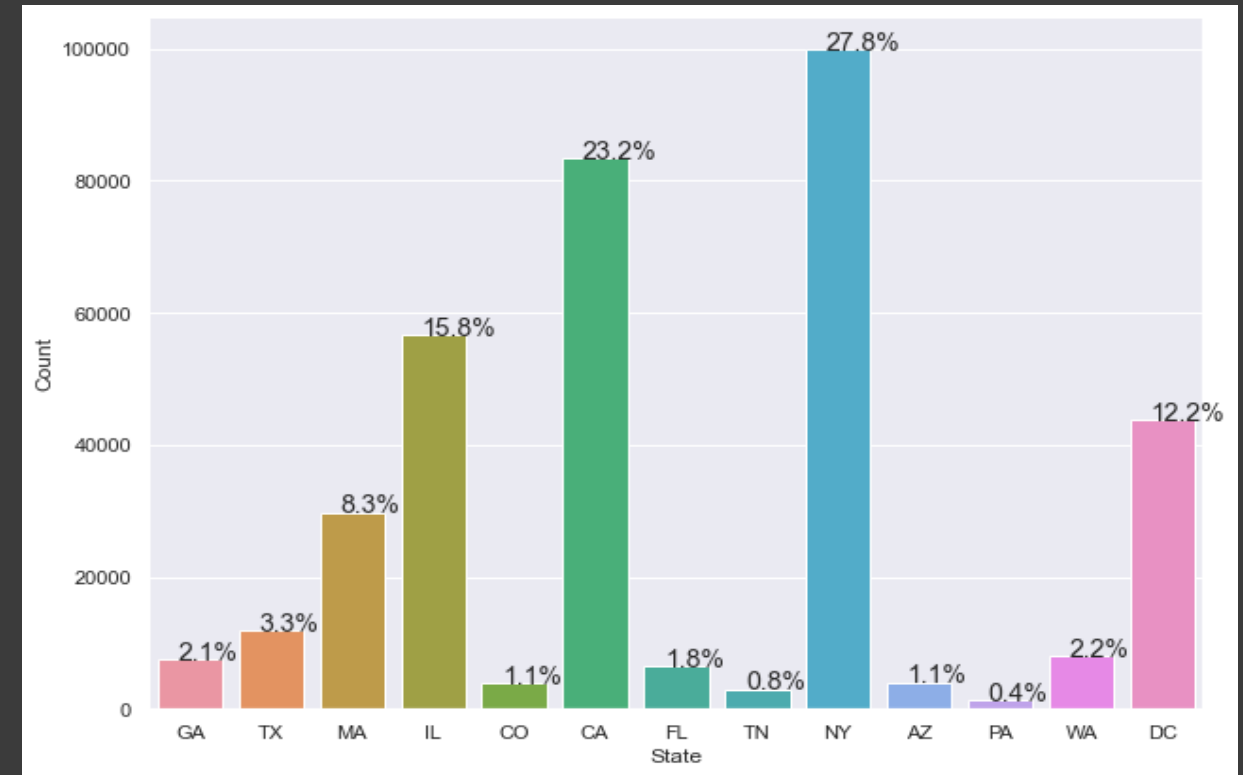
Some assumptions were made during the project:

- The difference between columns 'cost' and 'charged' is the profit per transaction.
- Some transactions had negative profit, so I assumed that the customer would have availed some sort of discount or coupon, which reduced the charged amount.
- The weather data did not have records for certain days, I assumed that the weather was normal on these days, as there was no 'normal' value in the weather column.
- On days where multiple types of weather was present, I only took the first value that occurred. I could not perform better filtration processes due to time constraints.

EDA - Histogram

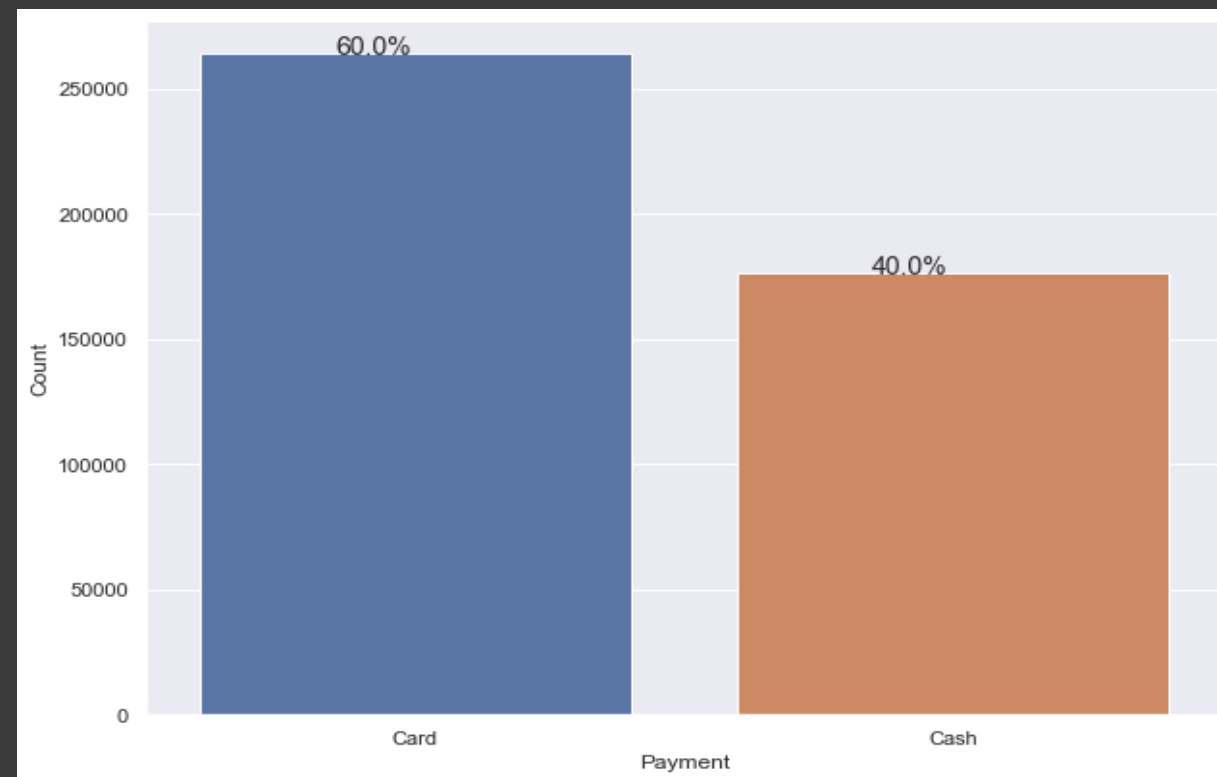


- Yellow cab seems to have much larger share of the taxi market.
- One explanation could be that pink cabs are more expensive and are tailored for higher income audience. We will have to check this later.



- New York followed by California have the highest number of taxi trips.
- Illinois and Washington DC are third and fourth respectively with 15.8% and 12.2% of the market share.

EDA



- Similar to age, the income distribution is also uniform and divided into two parts, with the change in frequency occurring at \$25,000.
- All incomes before \$25,000 have a frequency of around 1,400 customers, while an average of 300 to 400 customers have incomes more than \$25,000.

More people tend to pay with card than cash

Hypothesis 1

H0: mean trips of each quarter is the same.

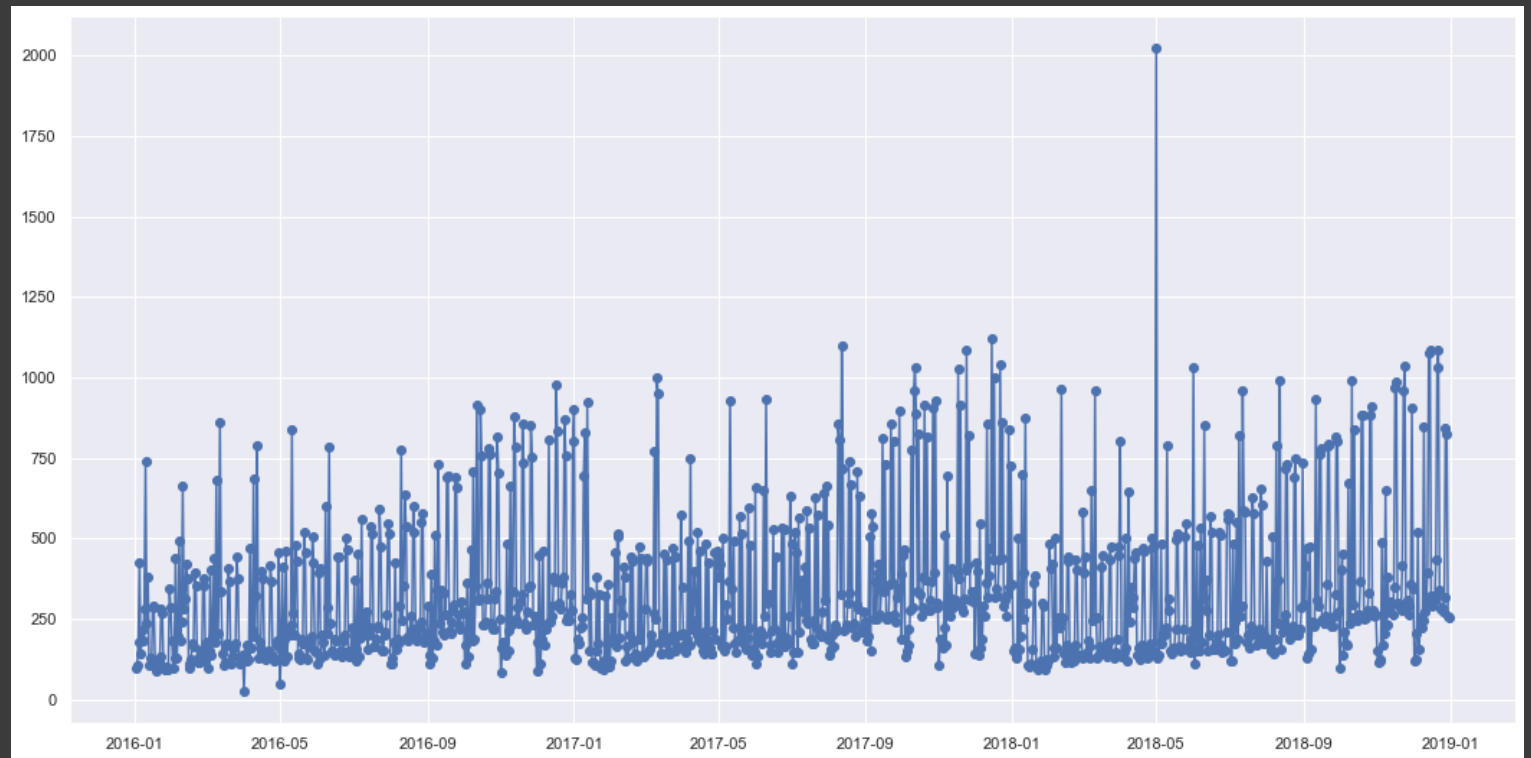
H1: at least one pair of quarters have different mean trips.

This hypothesis will check to see if there is any seasonality or trend in the taxi trips per year - if the trips in each quarter are generally the same or different.

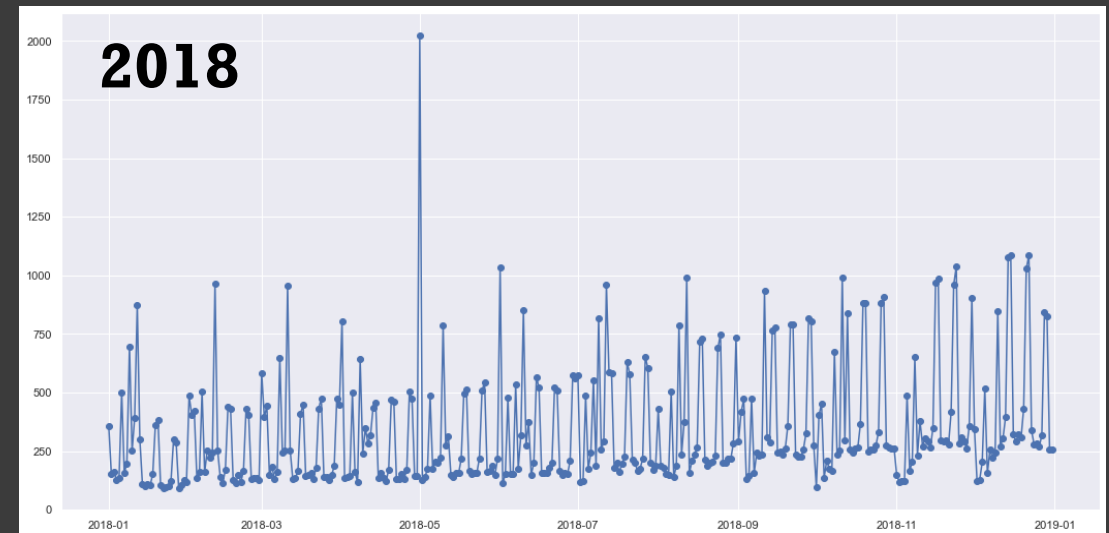
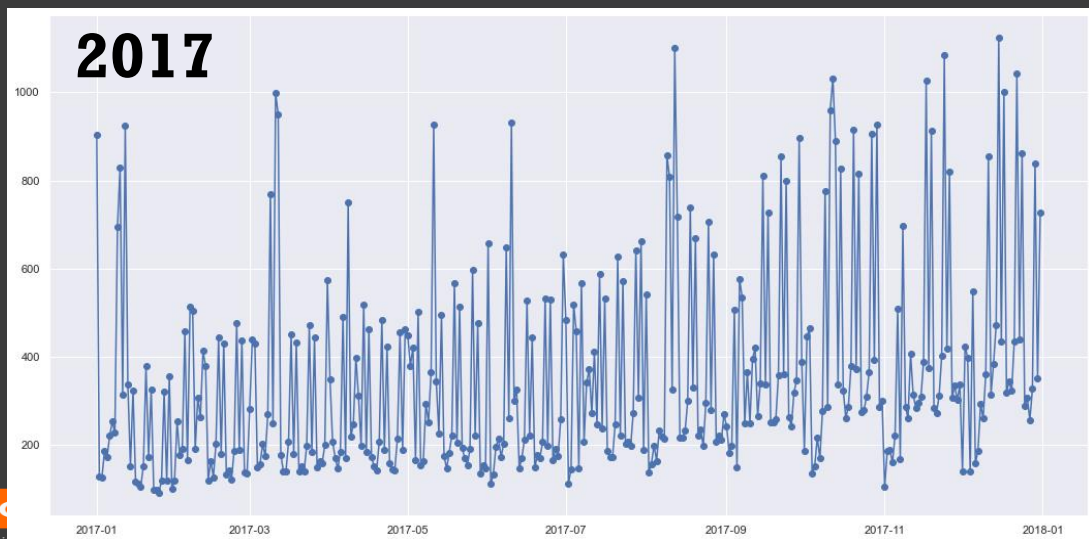
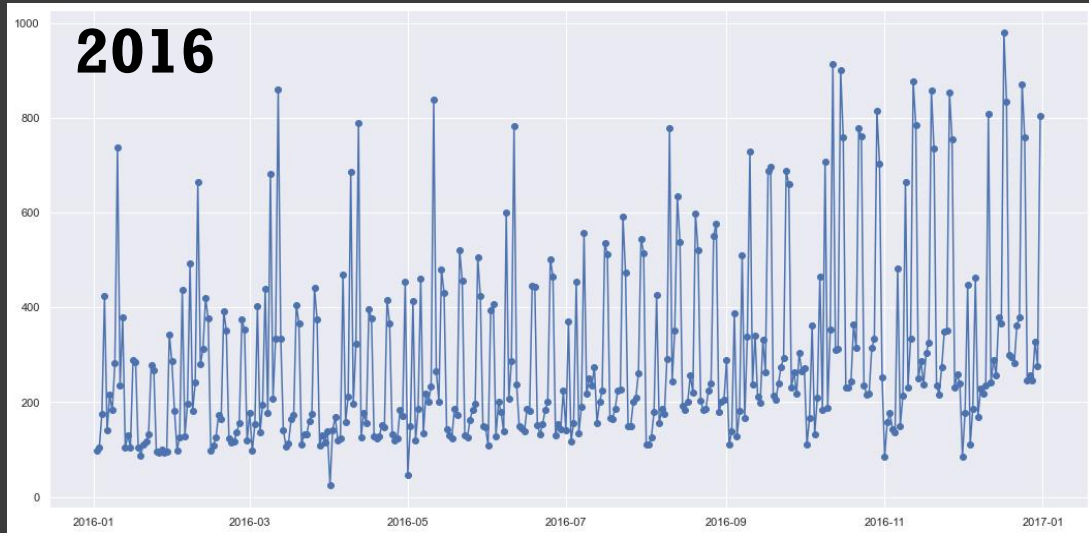
Observations

On a holistic picture, it does seem that there is a general trend over the 3 years. There seems to be a massive spike in trips somewhere in the middle of 2018.

Overall, there is no doubt that there is a seasonality trend.



Hypothesis 1



Hypothesis 1 - Results

Observations

- The 3 years are very similar as noted previously.
- Across the span of each year, the trips very gradually increases.
- In each month, there also seems to be a trend.

Results

- The graphs clearly indicate that there is seasonality in the taxi rides, and that the number increases gradually throughout the year, and then drops at the start of the year (except for New Year)
- Thus, H_0 is rejected.

Hypothesis 2

H0: average profit margin of each city is the same.

H1: average profit margins are different for at least one pair of cities.

This hypothesis will check if the average profit margins are the same from city to city.

city	
New York	279.947491
Dallas	160.856957
Silicon Valley	154.561013
Miami	117.493220
Orange County	114.766920
Atlanta	111.477158
Austin	107.577824
Denver	103.943793
Phoenix	93.479109
Los Angeles	91.847452
Washington DC	79.860762
San Diego	77.467955
Seattle	75.613962
Tucson	72.636300
Pittsburgh	64.863638
Chicago	59.820104
Boston	59.568883
Nashville	49.678478
Sacramento	49.567466

city	
New York	12.437591
Dallas	7.119055
Silicon Valley	6.752115
Miami	5.170861
Orange County	5.135902
Atlanta	5.006862
Austin	4.821304
Denver	4.620508
Phoenix	4.179977
Los Angeles	4.050952
Washington DC	3.521579
San Diego	3.437747
Seattle	3.345344
Tucson	3.202417
Pittsburgh	2.769356
Boston	2.638993
Chicago	2.638358
Nashville	2.173062
Sacramento	2.161169

Results

- Straight away we see that the mean profit for each city is quite different. In New York it is extremely high. Other prominent cities such as Silicon Valley, Orange County and Miami also have profit margins on the higher side.
- The range of means is also quite large, even without considering New York.
- Based on what we have seen, we can safely say that the profit margins are not the same for each city.
- The null hypothesis is rejected

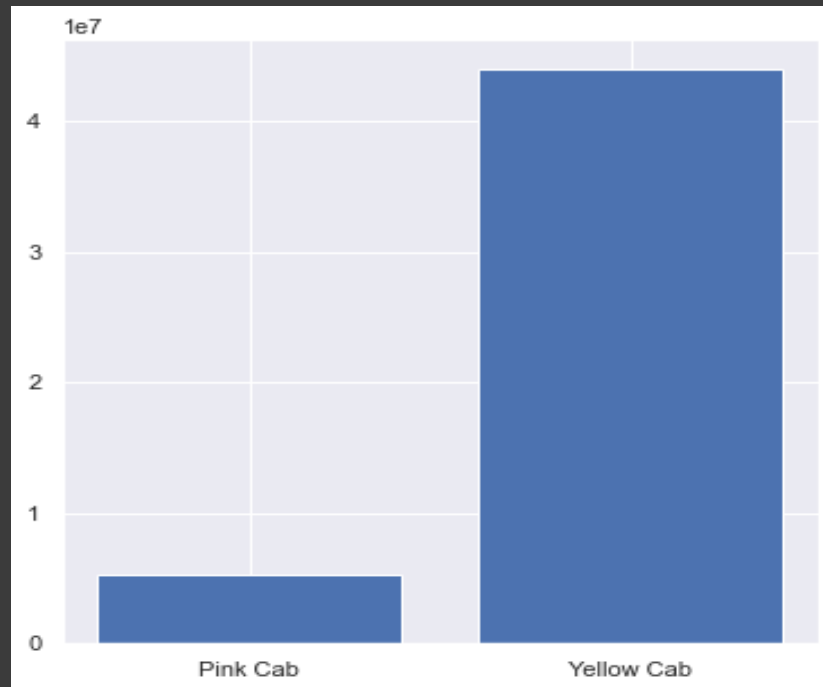
Hypothesis 3

H0: average profits for yellow cab is different to that of pink cab.

H1: average profits for both companies are different.

This hypothesis will check if the profit margins are different for each company.

Total Profit per company



Average Profit per company per year

year	company	
2016	Pink Cab	68.321819
	Yellow Cab	169.347821
2017	Pink Cab	67.070839
	Yellow Cab	168.817057
2018	Pink Cab	53.229689
	Yellow Cab	143.416122

Average Profit per company

company	
Pink Cab	62.652174
Yellow Cab	160.259986

Results

- This shows that yellow cab has made much more profit over the 3 year period than pink cab, almost \$38m more.
- The average profit is also higher, so yellow cab seems to charge more than pink cab.
- But in 2019, both companies' profits seem to slightly lower than the previous 2 years, as shown in the result annual profit chart.
- Therefore, we can say that the profits of both companies are not the same. So this hypothesis is also rejected.

Hypothesis 4

H0: mean income of customers is different for each company.

H1: mean income is different for each company.

This hypothesis will check if the taxi companies cater to different income groups of customers or not.

Results

- The first step was to check if customers use both companies or not. If they use both, then it would not help in this analysis since there will be a lot of overlapping.
- However, it was found that more than 25000 customers used both Pink and Yellow taxi services, therefore this hypothesis was not continued.

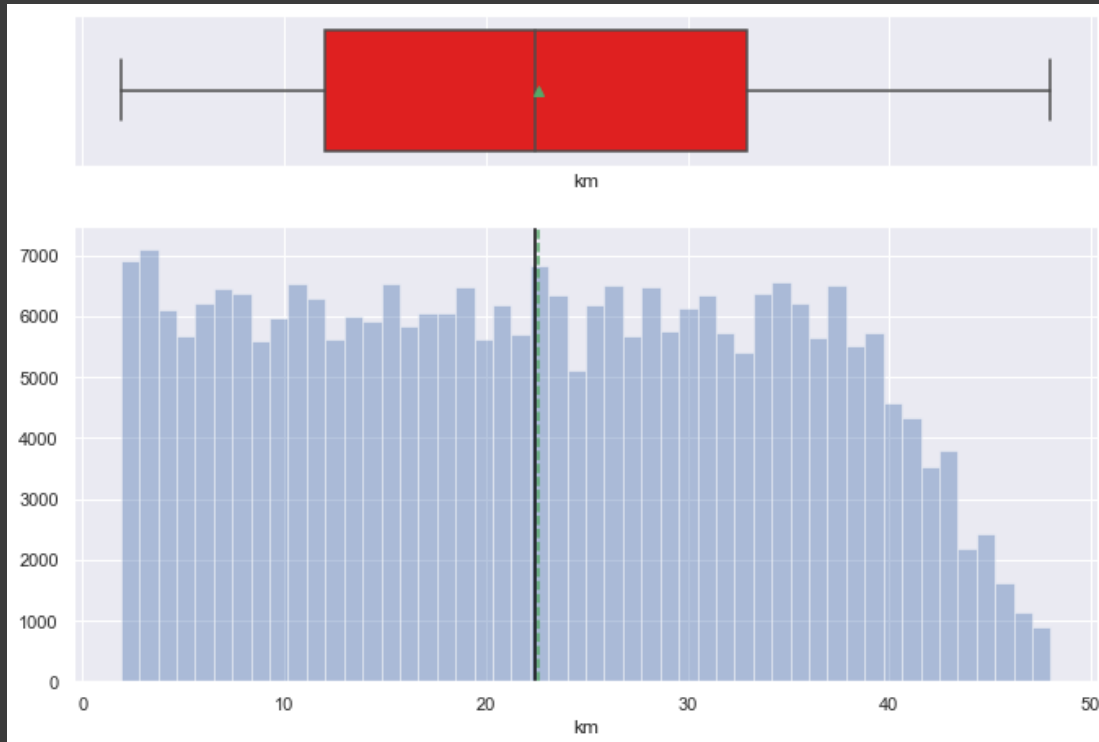
Hypothesis 5

H0: mean km of both companies are same.

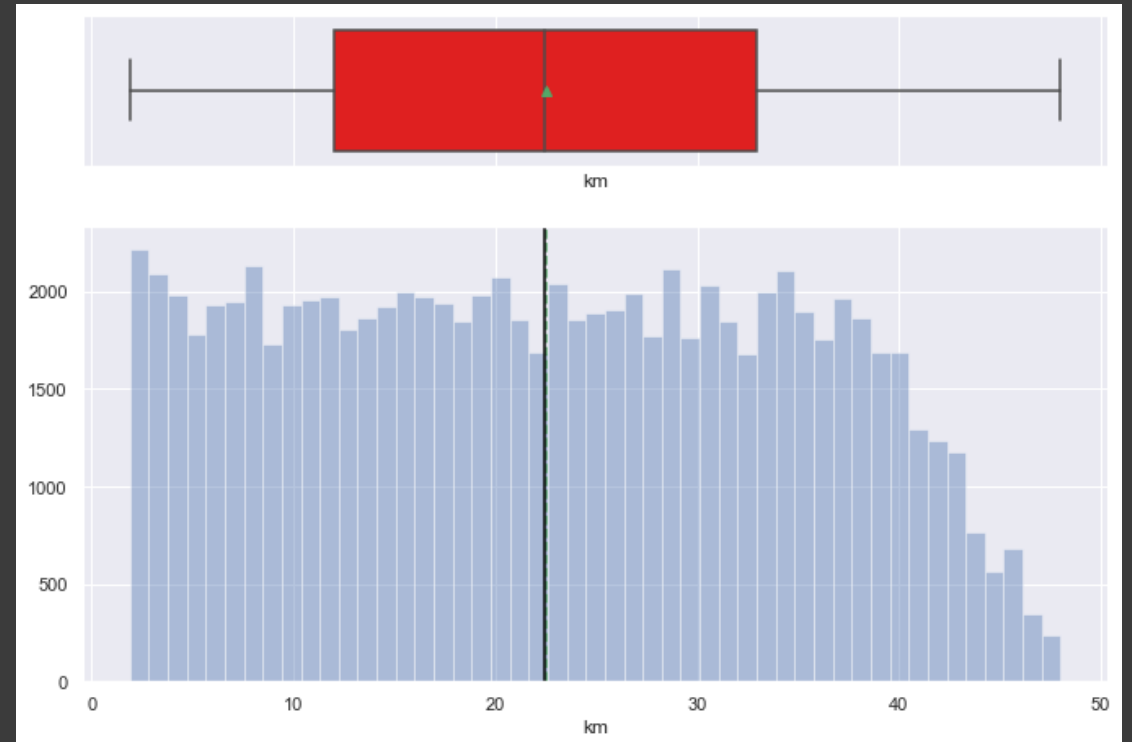
H1: mean kms of both companies are different.

Since customers tend to use both companies, let's check if the distance travelled has an effect on choice of cab.

Total kms - Yellow Cab



Total kms – Pink Cab



Hypothesis 5

company	year	
Pink Cab	2016	22.468488
	2017	22.618763
	2018	22.577275
Yellow Cab	2016	22.616742
	2017	22.557303
	2018	22.541036

Results

- The histograms are also very similar for both companies, just that yellow cabs have much more trips than pink cab.
- The average kms travelled is almost the same for each company, and also same per year.
- The means are obviously almost the same, so we fail to reject the null hypothesis.

Hypothesis 6

H0: trips in each weather type are same for both companies compared to normal conditions

H1: trips are affected by weather for one company differently than for the other company.

We will check if customers prefer a certain type of cab based on the weather.

weather	company	Count
Heavy Rain	Pink Cab	61
	Yellow Cab	240
Heavy Snow	Pink Cab	39
	Yellow Cab	200
Light Rain	Pink Cab	17574
	Yellow Cab	60194
Light Snow	Pink Cab	1733
	Yellow Cab	7658
Moderate Fog	Pink Cab	11355
	Yellow Cab	20549
Moderate Rain	Pink Cab	718
	Yellow Cab	2258
Moderate Snow	Pink Cab	149
	Yellow Cab	609
Normal	Pink Cab	44762
	Yellow Cab	154521
Other Hail	Pink Cab	72
	Yellow Cab	252
Severe Cold	Pink Cab	1868
	Yellow Cab	3970
Severe Fog	Pink Cab	3477
	Yellow Cab	6780
Severe Storm	Pink Cab	87
	Yellow Cab	285
UNK Precipitation	Pink Cab	2816
	Yellow Cab	17165

Ratio of pink to yellow cabs

Heavy Rain	0.25416666666666665
Heavy Snow	0.195
Light Rain	0.291956008904542
Light Snow	0.22629929485505354
Moderate Fog	0.5525816341427807
Moderate Rain	0.3179805137289637
Moderate Snow	0.24466338259441708
Normal	0.28968230855352994
Other Hail	0.2857142857142857
Severe Cold	0.47052896725440807
Severe Fog	0.5128318584070797
Severe Storm	0.30526315789473685
UNK Precipitation	0.1640547625983105

Results

- So the ratio of pink to yellow cabs used during normal weather is around 0.29. The ratios during other weather conditions also are around the same range, between 0.25 to 0.31, except for a few conditions.
- Among the exceptions are moderate fog, severe fog and severe cold, where the ratio is almost 0.5 for both. This almost feels like due to the conditions, the customers just wanted a cab and were willing to take whatever company's taxi was available.
- This theory is not evident in the case of severe storm or heavy rain, however.
- Due to the few weather conditions that showed a large increase in use of pink taxis, which increased the ratio in question, we can say that the null hypothesis is rejected, and that some weather conditions do affect the customers' choice of cab.

Recommendations

The recommendation after all the analyses and hypotheses is that XYZ firm should invest in the Yellow Cab company.

The reasons are as follows:

- The market share is dominated by Yellow cab, with around 76%.
- It has much higher profits than pink cab over the 3 years of data, regardless of the km travelled, time of year, customers etc.
- It has customers of all backgrounds and income groups. Regardless of their higher prices, customers still seem to flock to Yellow cab's services rather than pink cab.
- From the analyses conducted here, there has been no clear indicator as to why the high prices are not deterring the customers away. The only explanation would be that the quality of service is far better from Yellow Cab, that customers do not mind paying a bit more.

Further Insights

- Majority of customers are below 40 years and have income of 25000 and below.
- The gender of customers is fairly equally distributed.
- New York has the majority of taxi trips, unsurprisingly. It also has the highest profit margins.
- The taxi trips tend to increase gradually over the year, and then fall back slightly as the new year begins.
- New Year's day generally has a spike in taxi trip.

Thank You