Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

G2M insight for Cab investment firm

27 June 2021

Problem statement/Case Study

- **The Client XYZ is a private firm in 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.**

- **Objective:** Provide actionable insights to help XYZ firm in identifying the right company for making investment.

- **Cab Companies:**

- **Yellow Cab**
- **Pink Cab**

- The analysis is divided into the following sections:
  - Data understanding and data exploration
  - Exploratory data analysis
  - Hypothesis testing
  - Recommendations
  - Model building

Data Understanding and Data exploration

- There are 4 datasets :
- **Cab_Data.csv-**The dataset contains **359392 observations/rows and 7 fields/columns**. This dataset contains transaction details for each cab type.
- **Customer_ID.csv-**The dataset contains **49171 rows/observations and 4 fields/columns**.This dataset contains demographic details of each customer.The column **Customer ID** is the unique identifier or sometimes called Primary Key for this dataset.
- **Transaction_ID.csv-**The dataset contains **440098 rows/observations and 3 fields/columns**.This dataset maps with the **Customer_ID.csv** dataset on the **Customer ID** field/column. **Column ID** is a Foreign Key to the **Customer_ID.csv** dataset and the **Transaction ID** column is a Primary Key.
- **City.csv-**The dataset contains **20 rows/observations and 3 fields/columns**. Its contains a list of cities, the population of the cities and the number of cab users in U.S.

Joining datasets

- Merging is required to join datasets.
- First merge performed between the **Cab_Data.csv** and **Transaction_ID.csv** datasets.
- Merge on **Transaction ID** field/column is required.
- New dataset called **cab_and_transaction_merge**.

```
cab_and_transaction_merge
```

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip	Customer ID	Payment_Mode
0	10000011	2016-08-01	Pink Cab	ATLANTA GA	30.45	370.95	313.6350	29290	Card
1	10000012	2016-06-01	Pink Cab	ATLANTA GA	28.62	358.52	334.8540	27703	Card
2	10000013	2016-02-01	Pink Cab	ATLANTA GA	9.04	125.20	97.6320	28712	Cash
3	10000014	2016-07-01	Pink Cab	ATLANTA GA	33.17	377.40	351.6020	28020	Cash
4	10000015	2016-03-01	Pink Cab	ATLANTA GA	8.73	114.62	97.7760	27182	Card
...	...	...	...	...	...	...	...	...	...
359387	10440101	2018-08-01	Yellow Cab	WASHINGTON DC	4.80	69.24	63.3600	52392	Cash
359388	10440104	2018-04-01	Yellow Cab	WASHINGTON DC	8.40	113.75	106.8480	53286	Cash
359389	10440105	2018-05-01	Yellow Cab	WASHINGTON DC	27.75	437.07	349.6500	52265	Cash
359390	10440106	2018-05-01	Yellow Cab	WASHINGTON DC	8.80	146.19	114.0480	52175	Card
359391	10440107	2018-02-01	Yellow Cab	WASHINGTON DC	12.76	191.58	177.6192	52917	Card

359392 rows × 9 columns

- Next merge performed between the **cabtransaction_and_customer_merge** and **the Customer_ID.csv** datasets.
- Merge on **Customer ID** field/column is required.
- A new dataset called **cabtransaction_and_customer_merge**.

```
In [168]: cabtransaction_and_customer_merge
```

Out[168]:

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip	Customer ID	Payment_Mode	Gender	Age	Income (USD/Month)
0	10000011	2016-08-01	Pink Cab	ATLANTA GA	30.45	370.95	313.6350	29290	Card	Male	28	10813
1	10351127	2018-07-21	Yellow Cab	ATLANTA GA	26.19	598.70	317.4228	29290	Cash	Male	28	10813
2	10412921	2018-11-23	Yellow Cab	ATLANTA GA	42.55	792.05	597.4020	29290	Card	Male	28	10813
3	10000012	2016-06-01	Pink Cab	ATLANTA GA	28.62	358.52	334.8540	27703	Card	Male	27	9237
4	10320494	2018-04-21	Yellow Cab	ATLANTA GA	36.38	721.10	467.1192	27703	Card	Male	27	9237
...	...	...	...	...	...	...	...	...	...	...	...	...
359387	10439790	2018-07-01	Yellow Cab	SEATTLE WA	16.66	261.18	213.9144	38520	Card	Female	42	19417
359388	10439799	2018-03-01	Yellow Cab	SILICON VALLEY	13.72	277.97	172.8720	12490	Cash	Male	33	18713
359389	10439838	2018-04-01	Yellow Cab	TUCSON AZ	19.00	303.77	232.5600	41414	Card	Male	38	3960
359390	10439840	2018-06-01	Yellow Cab	TUCSON AZ	5.60	92.42	70.5600	41677	Cash	Male	23	19454
359391	10439846	2018-04-01	Yellow Cab	TUCSON AZ	13.30	244.65	180.3480	39761	Card	Female	32	10128

359392 rows × 12 columns

- Next merge between the **cabtransaction_and_customer_merge** and **the City.csv** dataset.
- Merge is performed on **City** field/column.
- The new final dataset called **master_data**.

```
master_data
```

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip	Customer ID	Payment_Mode	Gender	Age	Income (USD/Month)	Population	User
0	10000011	2016-08-01	Pink Cab	ATLANTA GA	30.45	370.95	313.6350	29290	Card	Male	28	10813	814,885	24,70
1	10351127	2018-07-21	Yellow Cab	ATLANTA GA	26.19	598.70	317.4228	29290	Cash	Male	28	10813	814,885	24,70
2	10412921	2018-11-23	Yellow Cab	ATLANTA GA	42.55	792.05	597.4020	29290	Card	Male	28	10813	814,885	24,70
3	10000012	2016-06-01	Pink Cab	ATLANTA GA	28.62	358.52	334.8540	27703	Card	Male	27	9237	814,885	24,70
4	10320494	2018-04-21	Yellow Cab	ATLANTA GA	36.38	721.10	467.1192	27703	Card	Male	27	9237	814,885	24,70
...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
359387	10307228	2018-03-03	Yellow Cab	WASHINGTON DC	38.40	668.93	525.3120	51406	Cash	Female	29	6829	418,859	127,00
359388	10319775	2018-04-13	Yellow Cab	WASHINGTON DC	3.57	67.60	44.5536	51406	Cash	Female	29	6829	418,859	127,00
359389	10347676	2018-06-07	Yellow Cab	WASHINGTON DC	23.46	331.97	337.8240	51406	Card	Female	29	6829	418,859	127,00
359390	10358624	2018-02-08	Yellow Cab	WASHINGTON DC	27.60	358.23	364.3200	51406	Cash	Female	29	6829	418,859	127,00
359391	10370709	2018-08-30	Yellow Cab	WASHINGTON DC	34.24	453.11	427.3152	51406	Card	Female	29	6829	418,859	127,00

359392 rows × 14 columns

- Insertion of new columns and renaming columns
- 3 new columns/fields inserted:
- **Month**-Month number of the year 1-12.
- **Year**-from 2016 to 2018.
- **Margin**-The profit made.To calculate Margin/profit=**Price Charged**-**Cost of Trip**.
- The new updated master_data now has 17 columns.

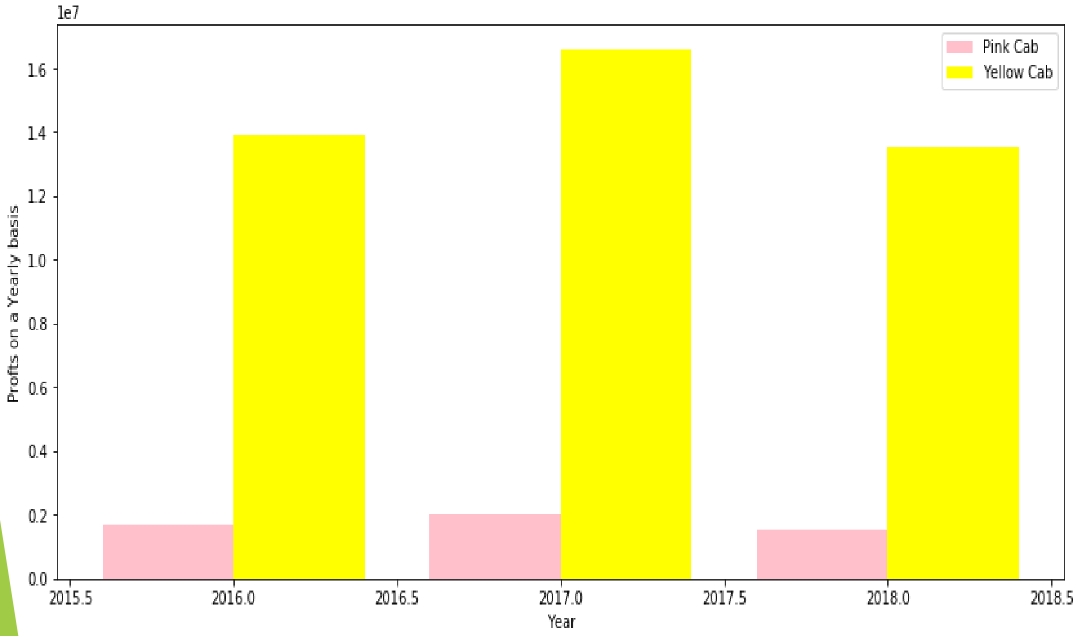
Exploratory Data Analysis(EDA)

- We need to first split a data between the **Yellow** and the **Pink cab company** before we perform our exploratory data analysis.
- Once that is done than it becomes easier to make some comparisons.
- Some aggregate functions can make comparisons much easier:
- **Sum**
- **Average**
- Total **Price Charged** for **Yellow cab and Pink cab:**
- **Yellow cab:125853887.18999998 $(U.S Dollars)**
- **Pink cab: 26328251.329999994 $(U.S Dollars)**

We find a higher Total **Price Charged** in a **Yellow cab** and the difference is 99525635.85999998 $(U.S Dollars).

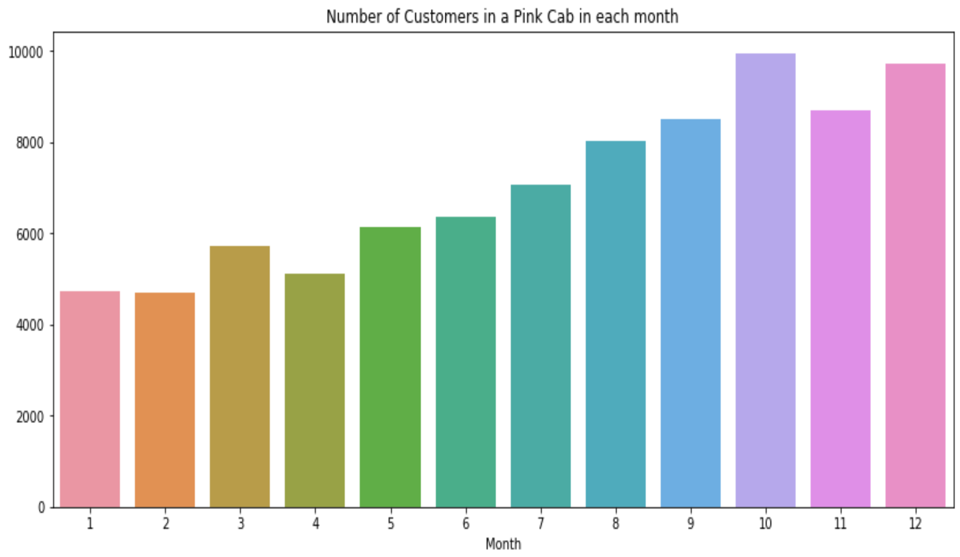
- Total **Margin** for **Yellow cab** and **Pink cab:**
- **Yellow cab: 44020373.17080002 $(U.S Dollars).**
- **Pink cab: 5307328.321 $(U.S Dollars).**
- We find a higher **Margin/Profit** in a **Yellow cab** and the difference is **38713044.84980002$(U.S Dollars)**

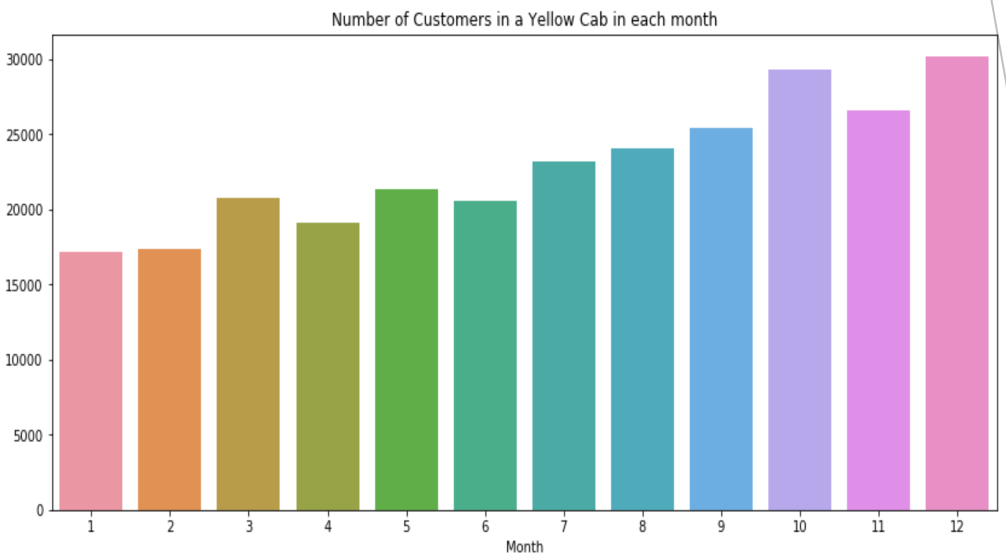
Profit analysis on a yearly basis



- We see a very high total profits generated each year for the **Yellow cab** compared with that of the **Pink cab**.
- Impacts a lot on the business performance.

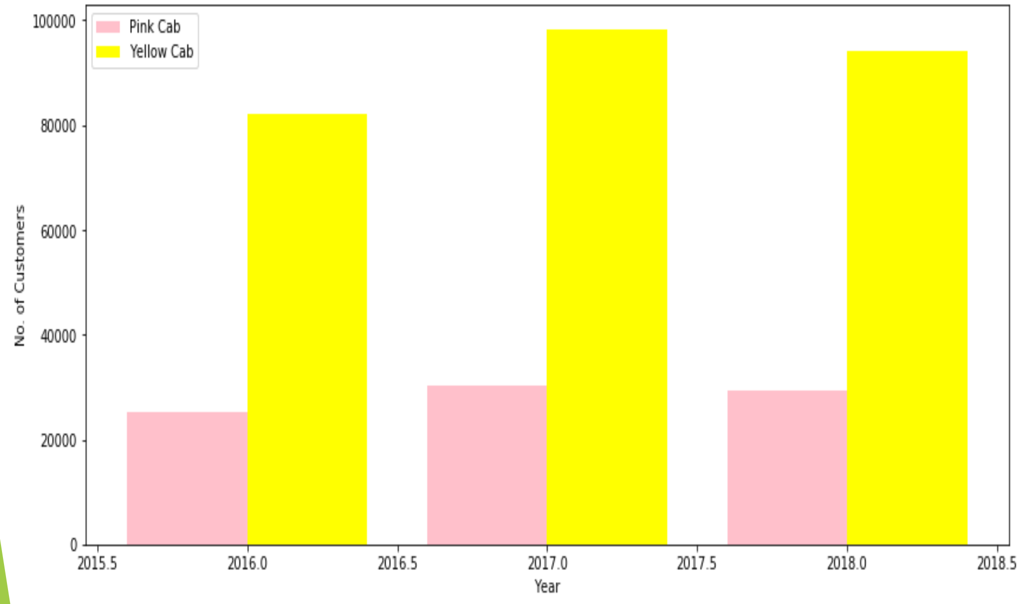
Analysis of a number of cab users in a monthly basis





- From the graph above we see the Yellow cab has a higher number of cab users
- The range in a number of cab users for the Yellow cab is higher than that of the Pink cab.
- Yellow cab ranges from 17108 to 30135 cab users.
- Pink cab ranges from 4734 to 9729 cab users.

Analysis of a number of cab users on a yearly basis

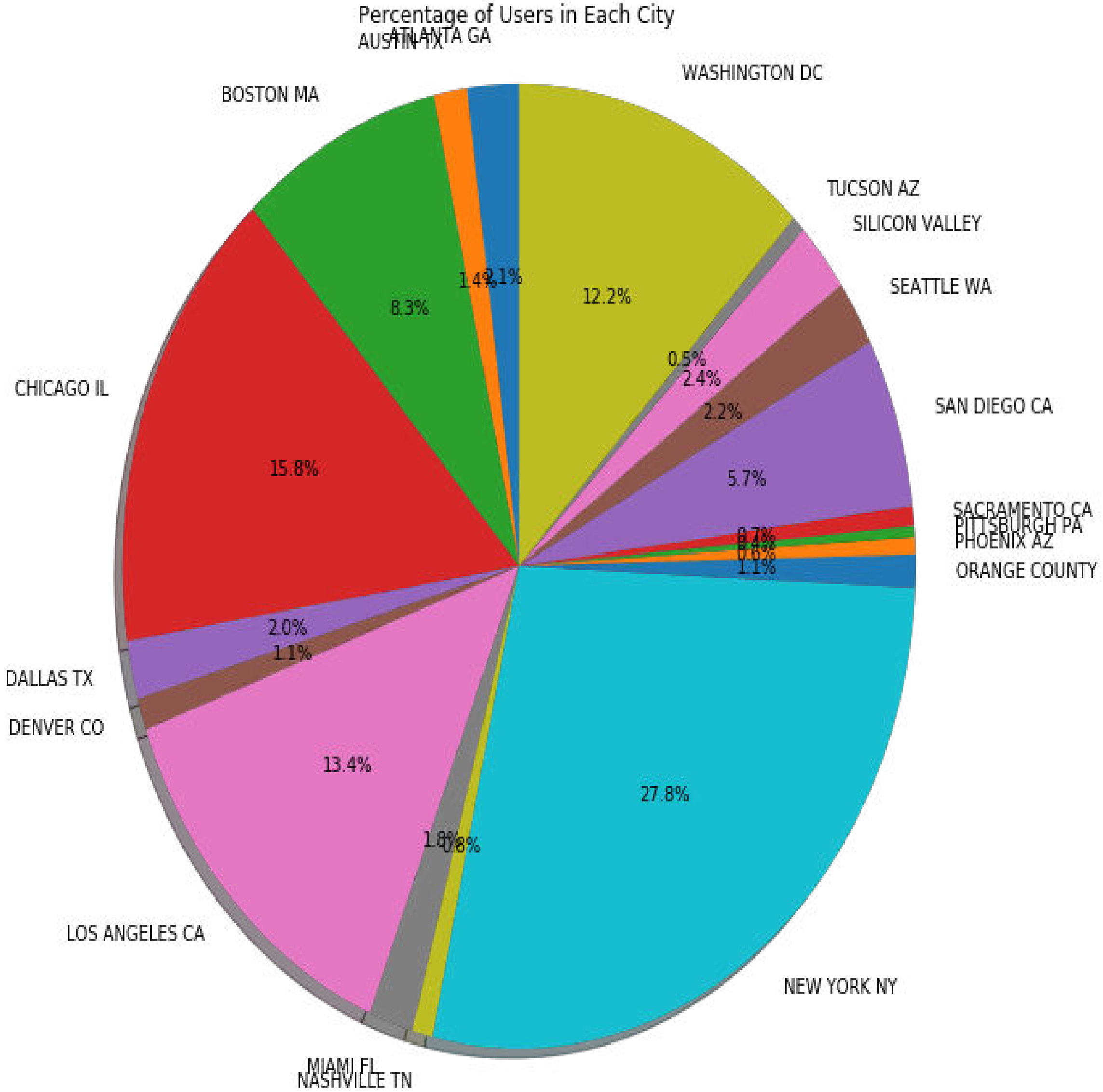


- Higher number of cab users travelling in a Yellow cab.
- There is a very high range in between the number of cab users in a Yellow cab and a Pink cab.
- Pink cab ranges from 25080 to 30321 cab users from the year 2016 to the year 2018.
- Yellow cab ranges from 82239 to 98189 cab users.

Limitations in analysis for a few cab users

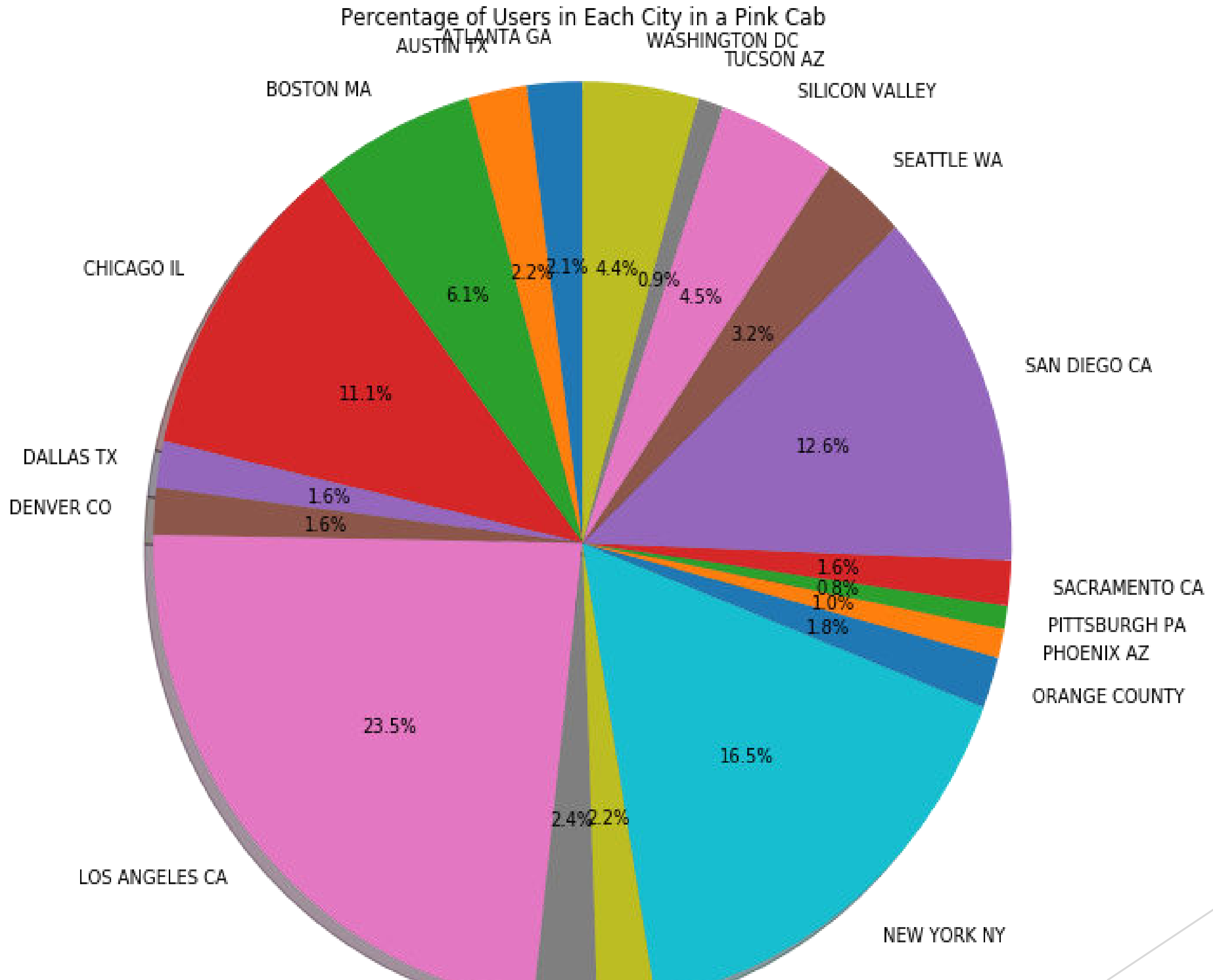
- Determining the number of cab users by identifying them by their unique Customer ID has some limitations.
- Does not provide accurate statistical analysis.
- Not all cab users are included, no representation of the whole population
- Need to determine number of cab users for the whole population in U.S.
- We can reveal the number of cab users by visualizing using a Pie chart showing the percentages of cab users in each city.

Percentage of cab users in each city



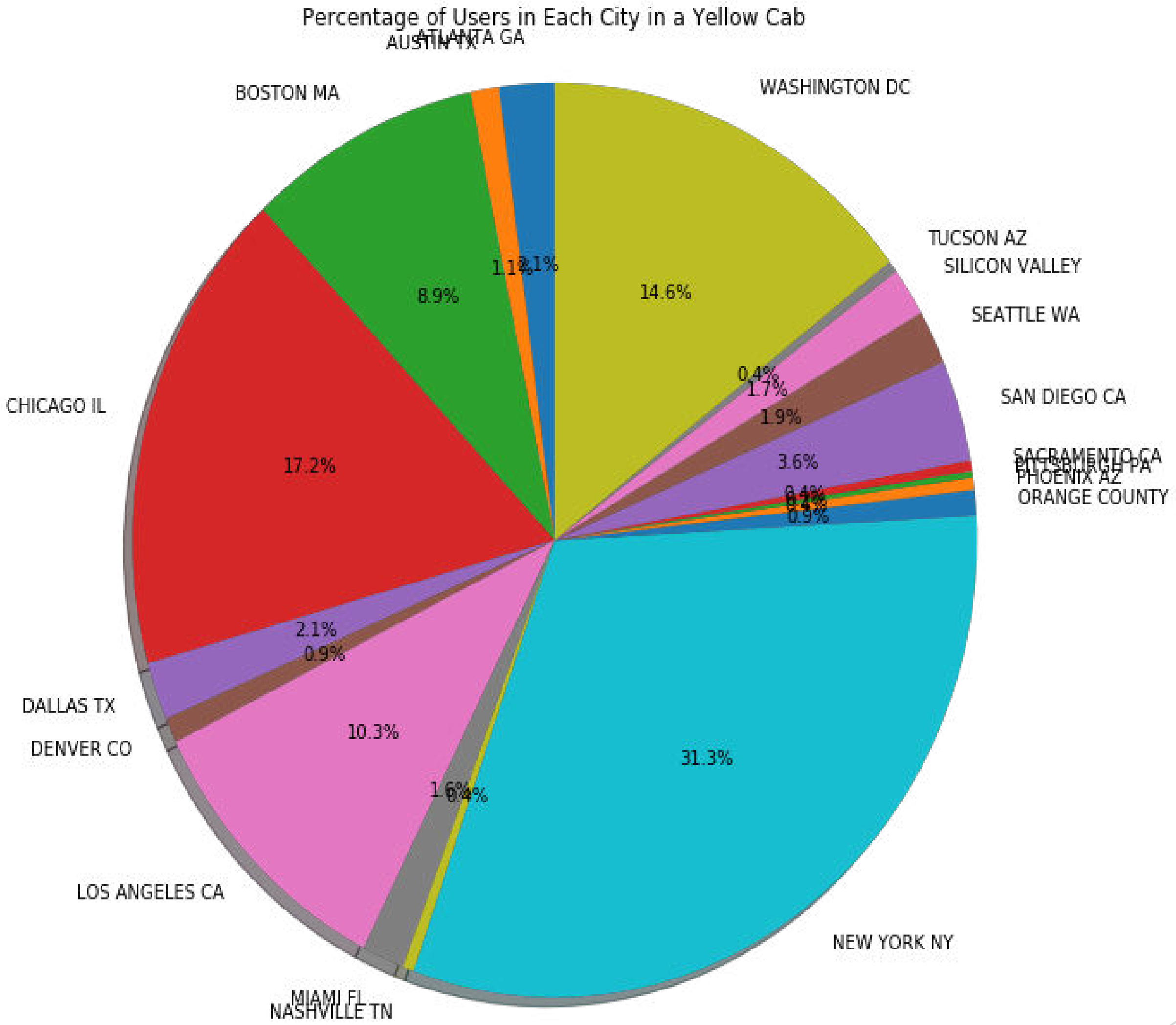
- The New York City has a high percentage of cab users with 27.8% followed by Chicago With 15.8% and Los Angeles With 13.4%

Percentage of cab users in each city travelling in Pink cab



- In the LOS ANGELES city the highest percentage of cab users are travelling in a Pink cab with 23.5%.

Percentage of cab users in each city travelling in a Yellow cab



- In the New York city the highest percentage of cab users are travelling In a Yellow cab with 31.3%.

Which cab has most cab users out of whole population in U.S?

- Visualizing using a pie chart revealing percentages does not provide answer for determining which cab has the majority of cab users.
- Only comparing between cities in U.S.
- Need to sum up all the number of cab users in each city for both cabs.

Table representation for a number of cab users travelling in Pink cab in each city

```
users_percity_in_pinkcab
```

```
City
ATLANTA GA          1762
AUSTIN TX           1868
BOSTON MA           5186
CHICAGO IL          9361
DALLAS TX           1380
DENVER CO           1394
LOS ANGELES CA     19865
MIAMI FL            2002
NASHVILLE TN        1841
NEW YORK NY        13967
ORANGE COUNTY       1513
PHOENIX AZ           864
PITTSBURGH PA        682
SACRAMENTO CA       1334
SAN DIEGO CA       10672
SEATTLE WA          2732
SILICON VALLEY      3797
TUCSON AZ            799
WASHINGTON DC       3692
```

Table representation for a number of cab users travelling in Yellow cab in each city

```
users_percity_in_yellowcab
```

City	
ATLANTA GA	5795
AUSTIN TX	3028
BOSTON MA	24506
CHICAGO IL	47264
DALLAS TX	5637
DENVER CO	2431
LOS ANGELES CA	28168
MIAMI FL	4452
NASHVILLE TN	1169
NEW YORK NY	85918
ORANGE COUNTY	2469
PHOENIX AZ	1200
PITTSBURGH PA	631
SACRAMENTO CA	1033
SAN DIEGO CA	9816
SEATTLE WA	5265
SILICON VALLEY	4722
TUCSON AZ	1132
WASHINGTON DC	40045

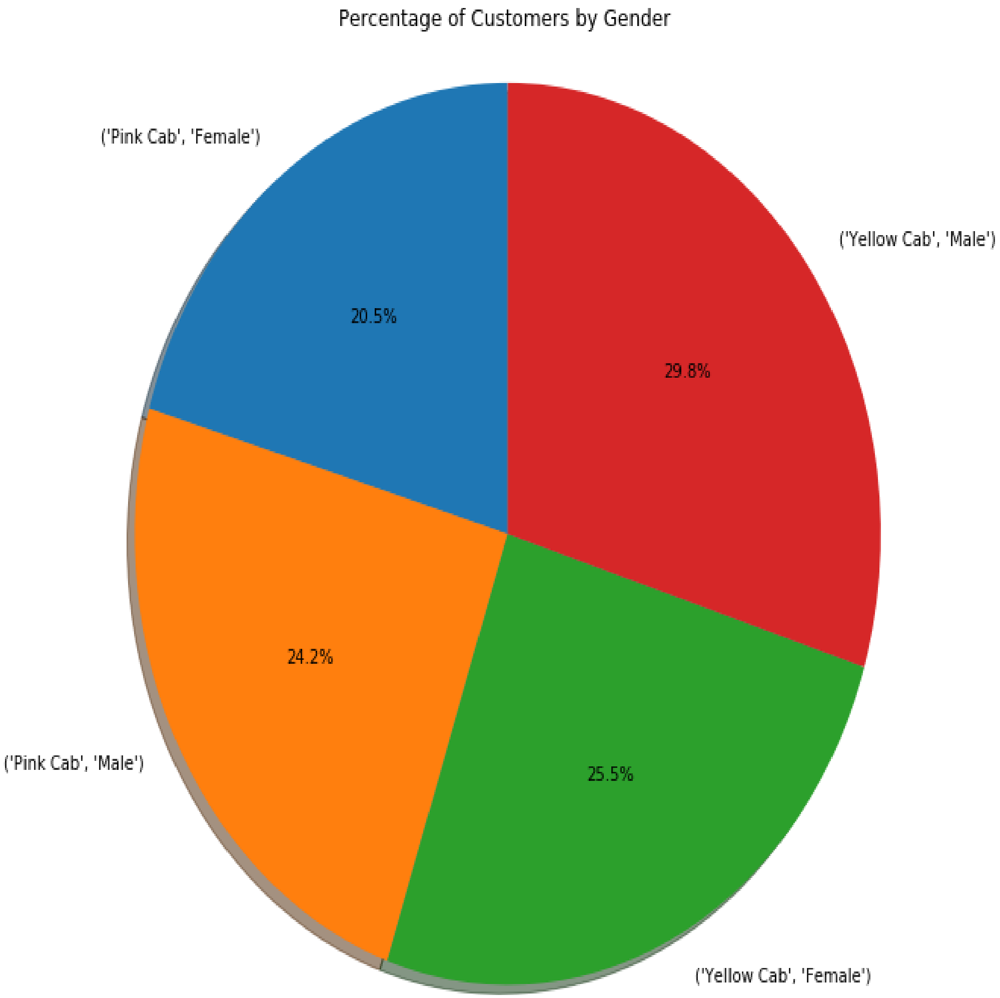
- With the table representation illustrated above, it’s now easier to calculate the total number of cab users representing the whole population.
- We sum up all the number of cab users in all cities altogether .

- Total number of   cab users travelling in Pink cab:84711.

Total number of cab users travelling in Yellow cab:274681.

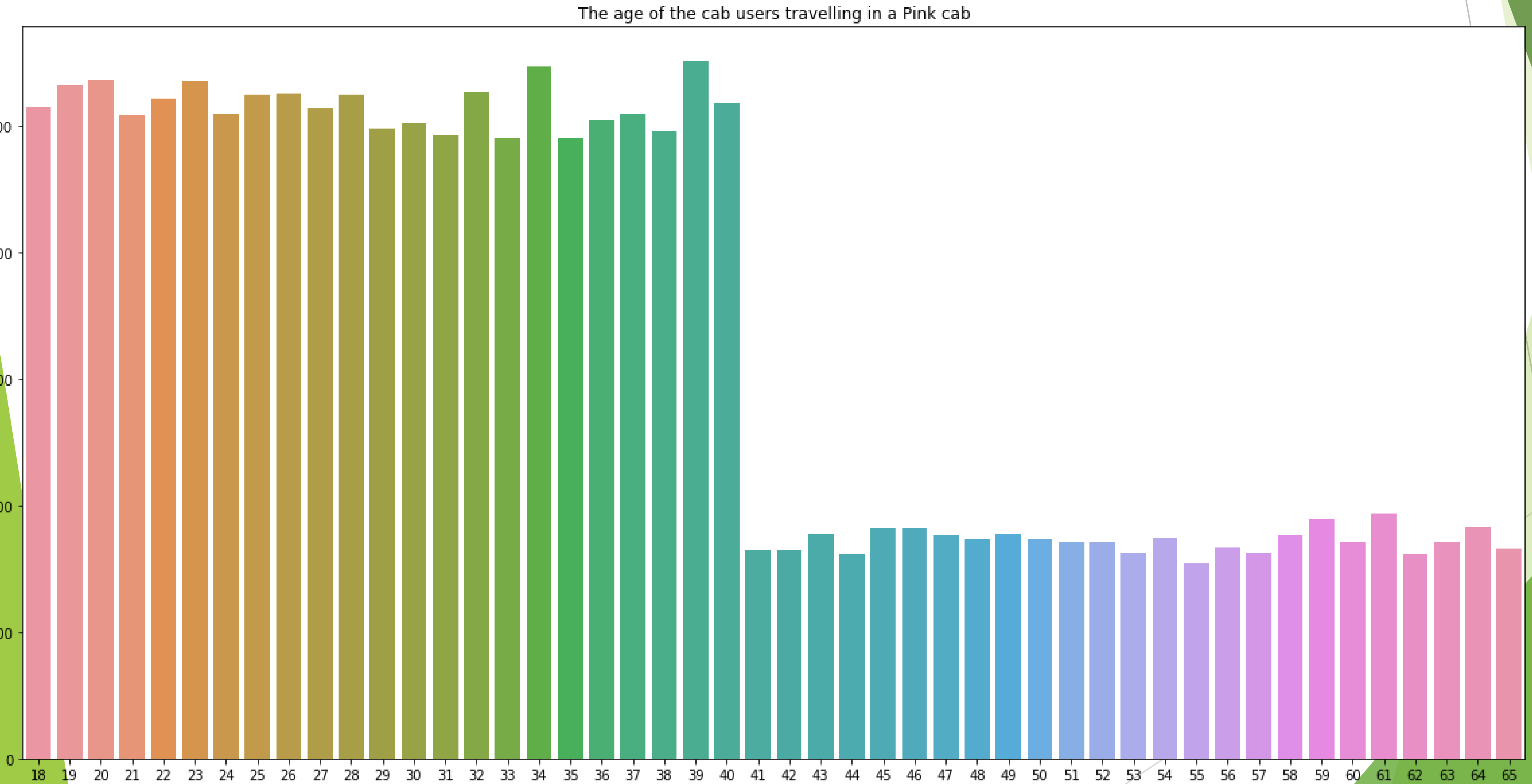
As we can see, the total number of cab users travelling in a Yellow cab are higher than the ones travelling in a Pink cab.

Gender Analysis

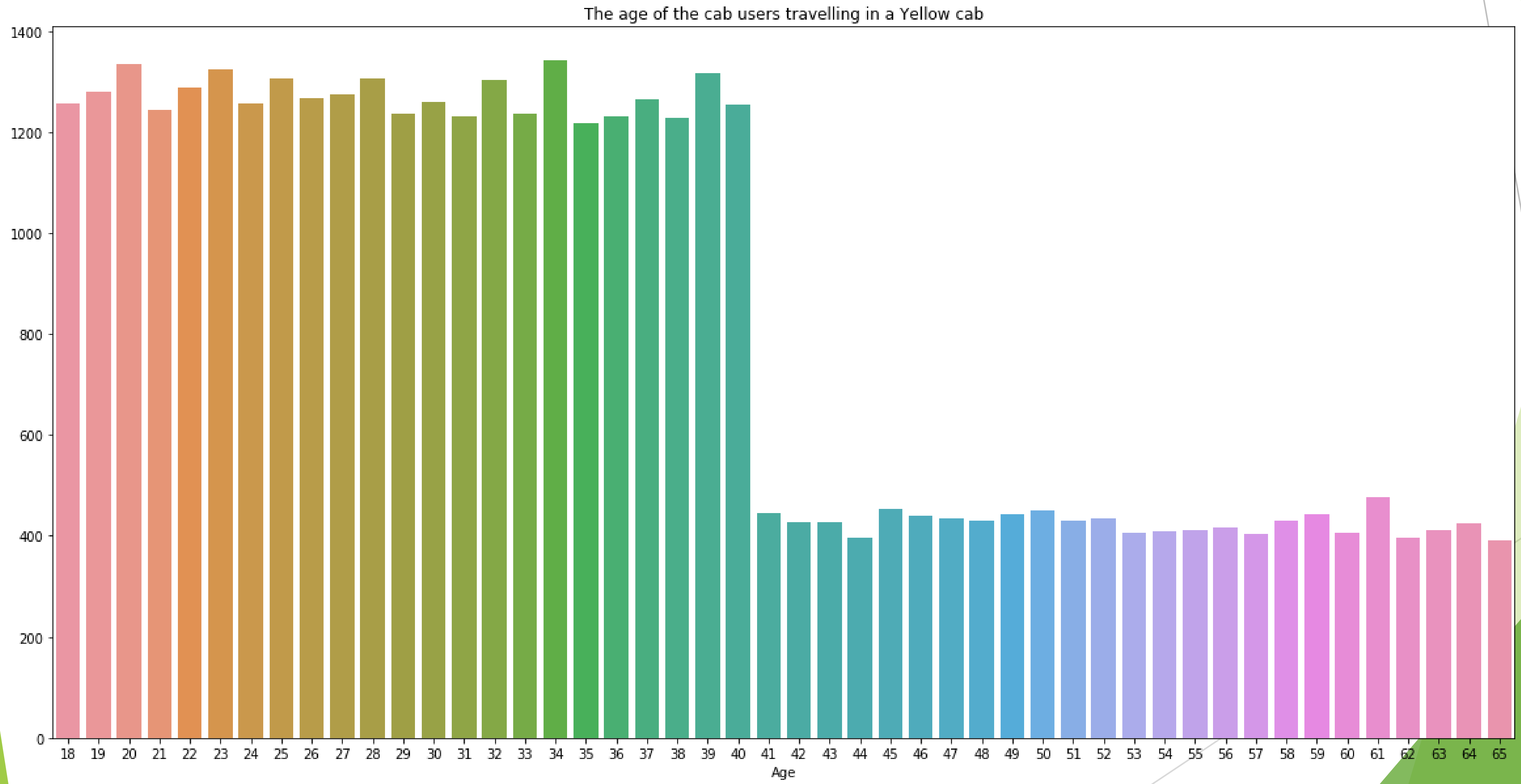


- Gender is one of the characteristics in determining customer behavior.
- Decision needs to be taken on what gender groups preference.
- Therefore, insights are required to reveal this.
- As shown on a pie chart illustrated the following is revealed:
  - In the Pink cab there are 20.5% females and 24.2% males.
  - In the Yellow cab there are 29.8% males
  - and 25.5% females.
- There is a higher percentage of males travelling in both Yellow and Pink cab.
- This means the male gender group is dominant

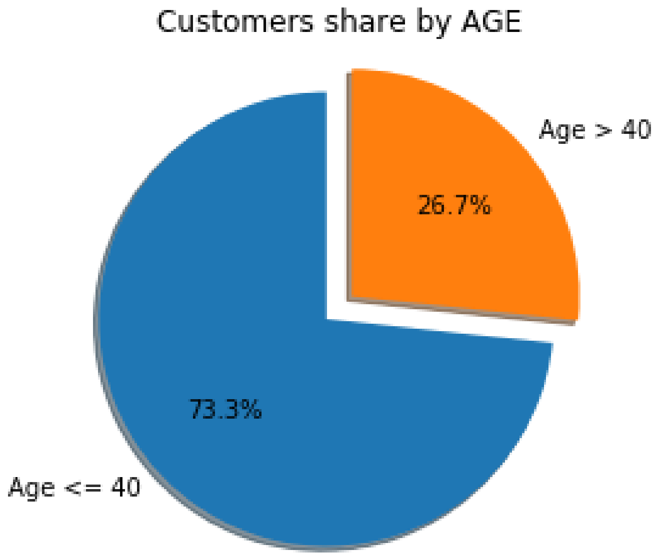
Age analysis for cab users travelling in a Pink cab



Age analysis for cab users travelling in a Yellow cab



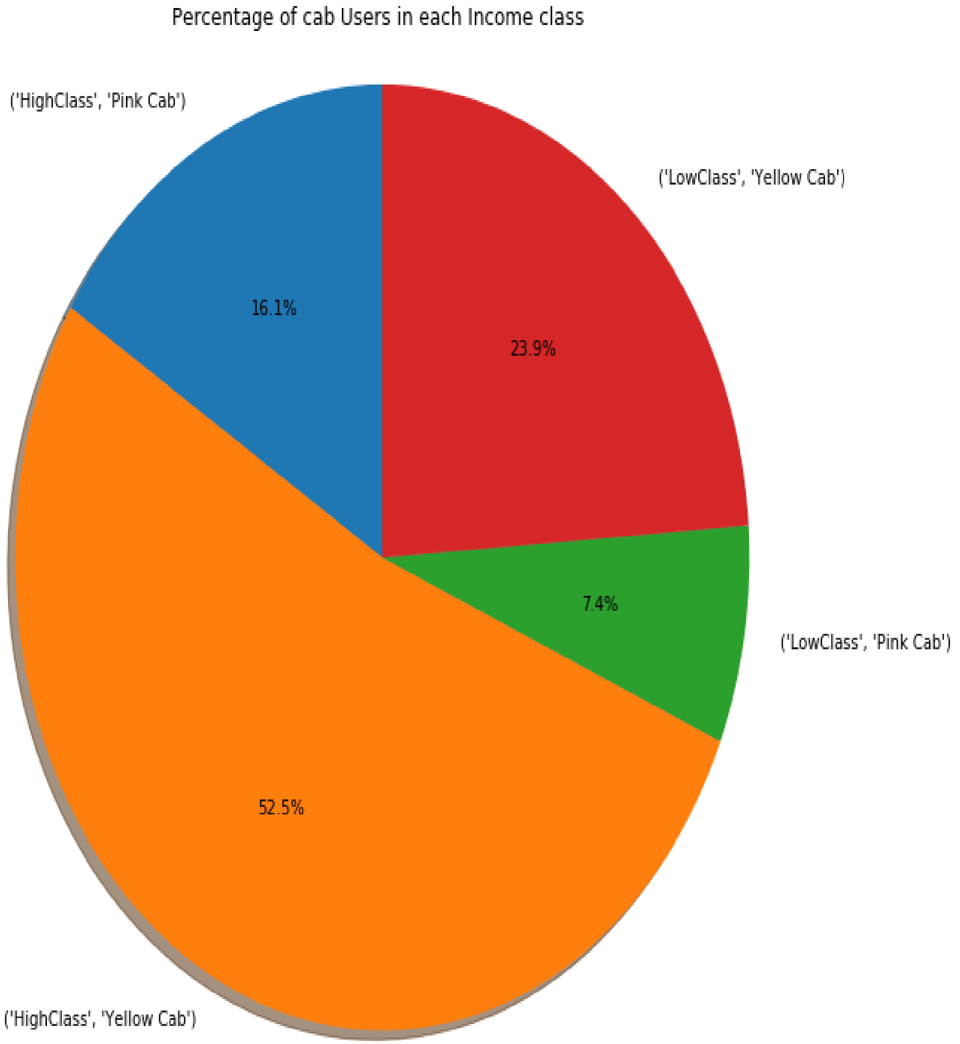
- Looking at the graphs above as illustrated we can see as we reach customers who are more than 40 years of age we see a decrease in a number of cab users
- The alternative is by visualizing using a pie chart .We can set the age limit by visualizing as: Age<=40 years or Age>40 years
- The Customers share by Age is visualized using a pie chart as illustrated below
- We can see that there is 73.3% customers of age less than 40 years and 26.7% customers of age more than 40 years



Income Analysis

- Income analysis is one of the characteristics to determine customer behavior.
- We need to group the Income Class of cab users into High class and Low class
- High class are classified as high-income earners and Low class as low-income earners
- Cab users earning less than 10 000 $(U.S Dollars) per month would be classified as Low class and earning above 10 000$(U.S Dollars) per month would be classified as High class

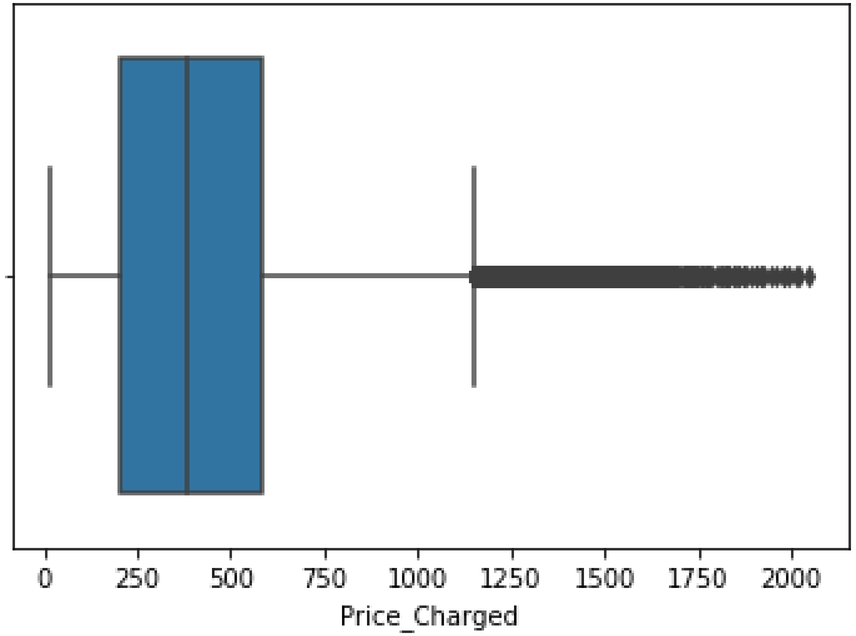
Percentage of cab users in each Income group

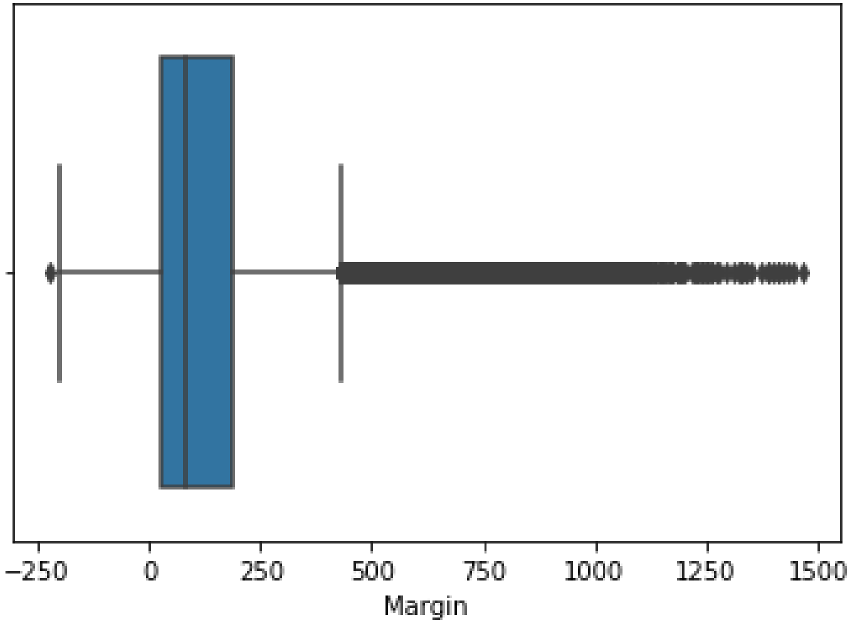


- Looking at the pie chart as illustrated, we can see that there Is a majority of Income group of cab users belonging to an Income group of a High class travelling in Yellow cab

This means that the High Class Contributes to high investment

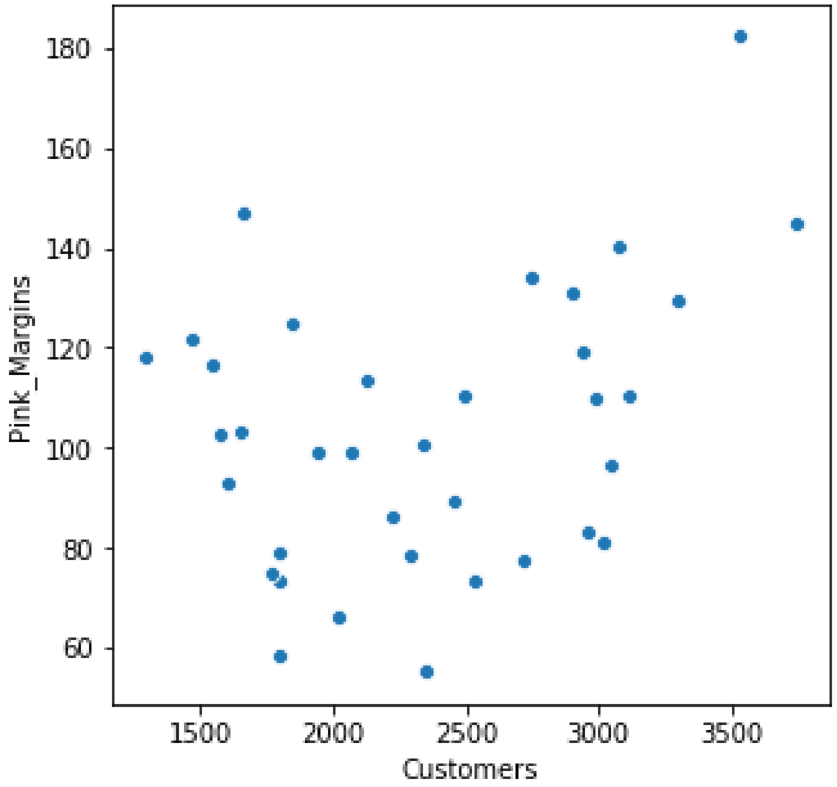
Detection of outliers



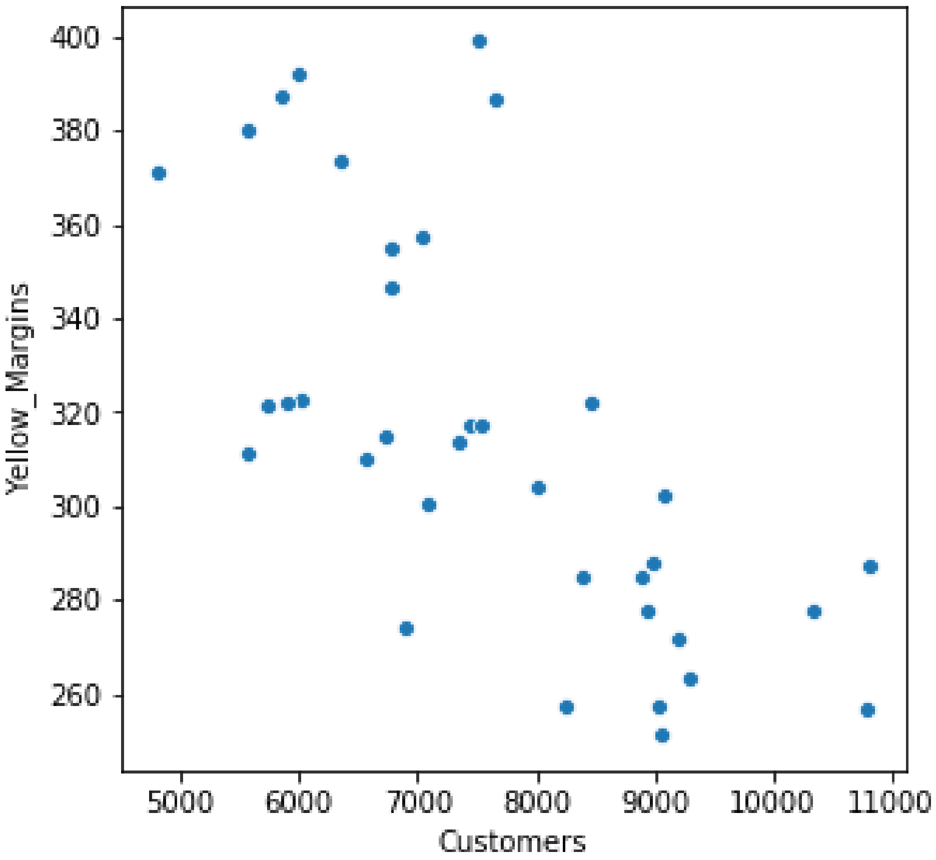


- We need to check for outliers in numerical values
- There are no outliers found in the columns/fields with numerical values except for the columns Price Charged and Margin
- The presence of outliers is due to high –end cars

Does margin proportionally increase with increase in number of customers?



- As illustrated on the diagram, we see that the Pink cabs increase their margins with an increase in number of customers



- Here as we see, the Yellow cabs decrease their margins with an increase in number of customers

Hypothesis Testing

Is there a difference in margin/profit between male and female customers for Yellow cabs?

```
print('P value is ', p_value)
```

```
18394 21502
We accept alternate hypothesis that there is a statistical difference
P value is  0.005345131177690902
```

- There is a difference in margin/profit for the Yellow cab between male and female customers,and therefore we accept the alternate hypothesis.

Is there a difference between gender and KM Travelled for Pink cabs?

Is there a difference in margin/profit between male and female customers for Pink cabs?

```
print('P value is ', p_value)
```

```
14819 17511
We accept null hypothesis that there is no statistical difference
P value is  0.5889424154257326
```

- We accept the null hypothesis,and therefore there is no difference in margin/profit for Pink cabs between male and female customers.

Is there a difference in margins/profit due to age of customers for Pink cab?

```
print('P value is ', p_value)
```

```
23721 8609
We accept null hypothesis that theres no difference
P value is  0.7840727156471817
```

- There is no difference is no difference in margin/profit for Pink cab due to the age of customers,and therefore we accept the null
- It doesn't make any difference whether the customer is less than equal to 40 or greater than 40.

Is there a difference in profit/margin due to the age of customers for Yellow cabs?

```
print('P value is ', p_value)
```

```
29254 6295
We accept null hypothesis that theres no difference
P value is  0.5813353417655228
```

- There is no difference in profit/margin for Yellow cabs due to the age of customers,so we accept the null hypothesis.
- It aslo doesn't make any difference whether the customer is less than equal to 40 or greater than 40.

Is there a difference between gender and KM Travelled for Yellow cabs?

```
print('P value is ', p_value)
```

```
18394 21502
We accept null hypothesis that theres no difference
P value is  0.20078753614825767
```

- There is no difference between gender and KM Travelled for yellow cab, so we accept the null hypothesis.

Is there a difference between gender and KM Travelled for Pink cab?

```
print('P value is ', p_value)
```

```
14819 17511
We accept alternate hypothesis that theres a difference
P value is  0.045565869972749515
```

- There is a difference between gender and KM Travelled for Pink cabs.

Recommendations

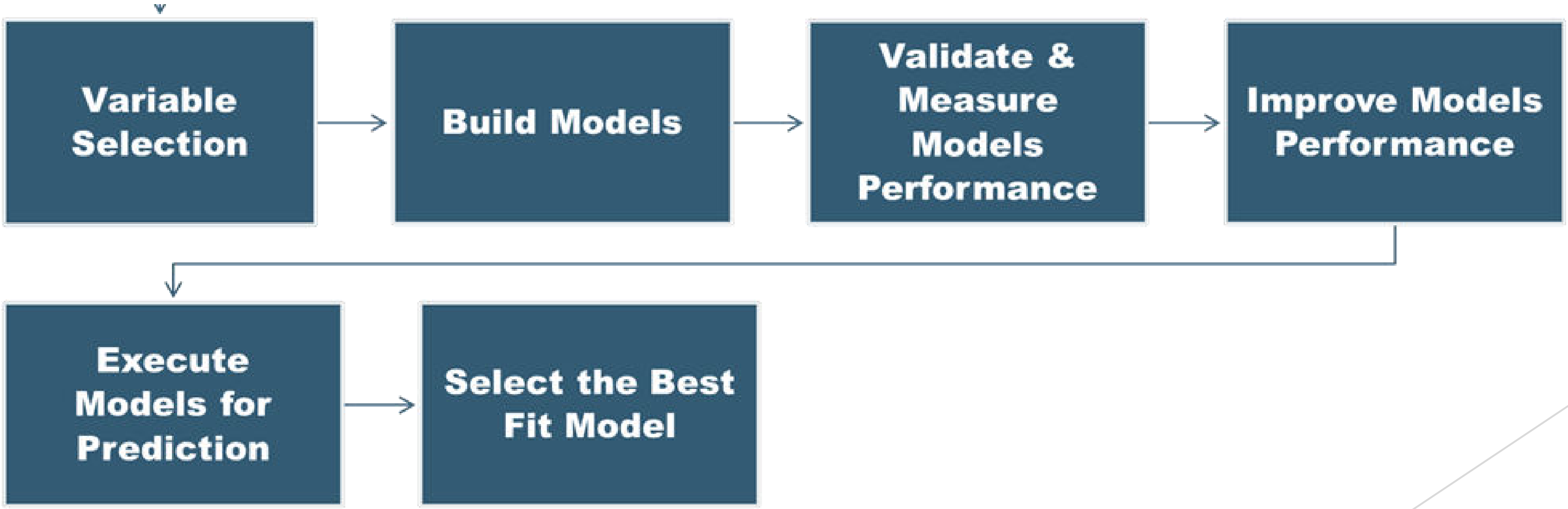
- To make a precise decision in which company would be a better investment opportunity,we need to clearly review our figures revealed from the exploratory data analysis
- The first exploration was determining the number cab users periodically(monthly and yearly) and our insights revealed as follows:
- The Yellow cab resulted in a higher cab users than that of the Pink cab travelling on a monthly basis. The Yellow cab revealed a higher range of cab users than that of the Pink cab
- The Yellow cab also resulted in a higher cab users than that of the Pink cab travelling on a yearly basis and revealed a higher range of cab users
- On top of that we found higher cab users travelling on a Yellow cab in representation of the whole country in U.S.
- Yellow cab had 274681 cab users and the Pink cab had 84711

- The Yellow cab generated a higher margin/profit than that of the Pink cab on a yearly basis
- From 2016 to 2018 has been generating a very higher profit than that of the Pink cab

So far, the Yellow cab company has been excelling and therefore is a better investment opportunity for XYZ.

Building Predictive Models using Linear Regression, Decision Tree and Random Forest.

Model Building steps



Model1: Linear Regression

- Linear Regression is a method for predicting target value and attempts to model the linear relationship between target and one or more predictors.
- In our dataset, Price Charge is the target value and all the other variables are predictors.
- **Splitting the data into a training set (75%), and test set (25%).**

- **Yellow Cab**

```
X_train.info()

<class 'pandas.core.frame.DataFrame'>
MultiIndex: 192276 entries, (10033146, 559, 'NEW YORK NY') to (10194745, 58301, 'BOSTON MA')
Data columns (total 6 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   KM_Travelled        192276 non-null  float64
 1   Cost_of_Trip        192276 non-null  float64
 2   Month               192276 non-null  int64
 3   Year                192276 non-null  int64
 4   Age                 192276 non-null  int64
 5   Income_(USD/Month)  192276 non-null  int64
dtypes: float64(2), int64(4)
memory usage: 24.8+ MB
```

```
X_test.info()

<class 'pandas.core.frame.DataFrame'>
MultiIndex: 82405 entries, (10263934, 7987, 'LOS ANGELES CA') to (10226996, 3195, 'CHICAGO IL')
Data columns (total 6 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   KM_Travelled        82405 non-null  float64
 1   Cost_of_Trip        82405 non-null  float64
 2   Month               82405 non-null  int64
 3   Year                82405 non-null  int64
 4   Age                 82405 non-null  int64
 5   Income_(USD/Month)  82405 non-null  int64
dtypes: float64(2), int64(4)
```

- **Pink Cab**

```
X_train.info()

<class 'pandas.core.frame.DataFrame'>
MultiIndex: 59297 entries, (10400049, 1517, 'NEW YORK NY') to (10085754, 13379, 'SILICON VALLEY')
Data columns (total 6 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   KM_Travelled        59297 non-null  float64
 1   Cost_of_Trip        59297 non-null  float64
 2   Month               59297 non-null  int64
 3   Year                59297 non-null  int64
 4   Age                 59297 non-null  int64
 5   Income_(USD/Month)  59297 non-null  int64
dtypes: float64(2), int64(4)
memory usage: 17.6+ MB
```

```
X_test.info()

<class 'pandas.core.frame.DataFrame'>
MultiIndex: 25414 entries, (10184224, 46628, 'SACRAMENTO CA') to (10158114, 8037, 'LOS ANGELES CA')
Data columns (total 6 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   KM_Travelled        25414 non-null  float64
 1   Cost_of_Trip        25414 non-null  float64
 2   Month               25414 non-null  int64
 3   Year                25414 non-null  int64
 4   Age                 25414 non-null  int64
 5   Income_(USD/Month)  25414 non-null  int64
```

Model2: Decision Tree

- **Decision tree** builds regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.
- The final result is a tree with decision nodes and leaf nodes.
- The topmost decision node in a tree which corresponds to the best predictor for the target value (Price Charged).

Model3: Random Forest

- A **Random Forest** operates by constructing several **Decision trees.**
- A prediction from the **Random Forest** is an average of the predictions produced by the **Decision trees** in the forest.

Base Model:

Yellow Cab

Dep. Variable:	Price_Charged	R-squared:	0.745
Model:	OLS	Adj. R-squared:	0.745
Method:	Least Squares	F-statistic:	1.336e+05
Date:	Sun, 14 Mar 2021	Prob (F-statistic):	0.00
Time:	09:57:32	Log-Likelihood:	-1.7581e+06
No. Observations:	274681	AIC:	3.516e+06
Df Residuals:	274674	BIC:	3.516e+06
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.774e+04	700.786	39.591	0.000	2.64e+04	2.91e+04
KM_Travelled	20.3282	0.198	102.704	0.000	19.940	20.716
Cost_of_Trip	-0.0052	0.015	-0.346	0.729	-0.034	0.024
Month	-5.5016	0.080	-68.649	0.000	-5.659	-5.345
Year	-13.7343	0.347	-39.532	0.000	-14.415	-13.053
Age	-0.0835	0.022	-3.780	0.000	-0.127	-0.040
Income_(USD/Month)	0.0002	3.49e-05	4.746	0.000	9.73e-05	0.000

Omnibus:	51903.377	Durbin-Watson:	0.652

Pink Cab

Dep. Variable:	Price_Charged	R-squared:	0.863
Model:	OLS	Adj. R-squared:	0.863
Method:	Least Squares	F-statistic:	8.871e+04
Date:	Sun, 14 Mar 2021	Prob (F-statistic):	0.00
Time:	09:57:34	Log-Likelihood:	-4.7693e+05
No. Observations:	84711	AIC:	9.539e+05
Df Residuals:	84704	BIC:	9.539e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.515e+04	584.885	25.903	0.000	1.4e+04	1.63e+04
KM_Travelled	13.4824	0.165	81.834	0.000	13.160	13.805
Cost_of_Trip	0.0295	0.015	1.985	0.047	0.000	0.059
Month	1.5216	0.069	21.950	0.000	1.386	1.657
Year	-7.5169	0.290	-25.924	0.000	-8.085	-6.949
Age	-0.0400	0.018	-2.185	0.029	-0.076	-0.004
Income_(USD/Month)	3.423e-05	2.9e-05	1.181	0.238	-2.26e-05	9.11e-05

Omnibus:	28936.298	Durbin-Watson:	0.887

- As shown above, As per Base Model:
- Cost of Trip, Month, Year, Age, Income are significant variable for **Yellow Cab** which are the best predictors for Price Charged.
- Cost_of_Trip, Year, Age, Income are significant variable for **Pink Cab** which are the best predictors for Price Charged. Month is not considered significant.

Best Fit Model: RMSE Value & Accuracy

- **RMSE or root mean square error** measures the error which is Prediction values – Actual values.
- **Lower the RMSE value the better is the Model.**

- RMSE values & Accuracy for Yellow Cab

	Train	Test
Linear Regression	145.4599	146.1994
Decision Tree	107.3967	109.4580
Random Forest	77.2731	78.4734

	Accuracy
Linear Regression	74.43906127028283%
Decision Tree	86.11582117196697%
Random Forest	92.85776861169764%

- RMSE values & Accuracy for Pink Cab

	Train	Test
Linear Regression	67.2351	67.9136
Decision Tree	80.7492	84.4882
Random Forest	57.4761	59.7556

	Accuracy
Linear Regression	86.06270464033021%
Decision Tree	79.66683587364297%
Random Forest	89.78196675241622%

- **As per the above RMSE data and Accuracy, Random Forest Model is the best fit model for further deployment.**

- **Interpreting Random Forest Model: Cost of Trip, Month, Year, Age, Income are the best predictors for Price Charged.**

Thank You

Data Glacier
Your Deep Learning Partner