

BUSINESS REPORT ON
MACHINE LEARNING
EXTENDED PROJECT

BY NANCY GUPTA

Part 1: Machine Learning Models You work for an office transport company

Dataset for part 1: [Transport.csv](#)

You are in discussions with ABC Consulting company for providing transport for their employees. For this purpose, you are tasked with understanding how do the employees of ABC Consulting prefer to commute presently (between home and office). Based on the parameters like age, salary, work experience etc. given in the data set 'Transport.csv', you are required to predict the preferred mode of transport. The project requires you to build several Machine Learning models and compare them so that the model can be finalised.

Data Dictionary :

Age: Age of the Employee in Years

Gender: Gender of the Employee

Engineer: For Engineer =1 , Non Engineer =0

MBA: For MBA =1 , Non-MBA =0

Work Exp: Experience in years

Salary: Salary in Lakhs per Annum

Distance: Distance in km from Home to Office

license: If Employee has Driving Licence -1, If not, then 0

Transport: Mode of Transport

The objective is to build various Machine Learning models on this data set and based on the accuracy metrics decide which model is to be finalised for finally predicting the mode of transport chosen by the employee.

Questions:

1. **Basic data summary, Univariate, Bivariate analysis, graphs, checking correlations, outliers and missing values treatment (if necessary) and check the basic descriptive statistics of the dataset.**

Ans. To begin with the analysis let us go through the dataset by summarizing it. After importing the necessary libraries and the dataset we checked the head or top 5 and last 5 rows of the dataset.

	Age	Gender	Engineer	MBA	Work Exp	Salary	Distance	license	Transport
0	28	Male	0	0	4	14.3	3.2	0	Public Transport
1	23	Female	1	0	4	8.3	3.3	0	Public Transport
2	29	Male	1	0	7	13.4	4.1	0	Public Transport
3	28	Female	1	1	5	13.4	4.5	0	Public Transport
4	27	Male	1	0	4	13.4	4.6	0	Public Transport

```
[ ] df.tail()
```

	Age	Gender	Engineer	MBA	Work Exp	Salary	Distance	license	Transport
439	40	Male	1	0	20	57.0	21.4	1	Private Transport
440	38	Male	1	0	19	44.0	21.5	1	Private Transport
441	37	Male	1	0	19	45.0	21.5	1	Private Transport
442	37	Male	0	0	19	47.0	22.8	1	Private Transport
443	39	Male	1	1	21	50.0	23.4	1	Private Transport

Now as we can check the number of rows are 444 and columns are 9 in the dataset.

```
no. of rows: 444
no. of columns: 9
```

On further analysis on the dataset, we see that the dataset different types of variables including float which is 2, integer which is 5 and 2 out of 9 variables are object type. We do not find any kind of null values in the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 444 entries, 0 to 443
Data columns (total 9 columns):
#   Column      Non-Null Count  Dtype
---  -
0   Age         444 non-null    int64
1   Gender      444 non-null    object
2   Engineer    444 non-null    int64
3   MBA         444 non-null    int64
4   Work Exp    444 non-null    int64
5   Salary      444 non-null    float64
6   Distance    444 non-null    float64
7   license     444 non-null    int64
8   Transport   444 non-null    object
dtypes: float64(2), int64(5), object(2)
memory usage: 31.3+ KB
```

There are no duplicate values in the dataset or no null values in the dataset either.


```
Number of duplicate rows = 0
```



	0
Age	0
Gender	0
Engineer	0
MBA	0
Work Exp	0
Salary	0
Distance	0
license	0
Transport	0


dtype: int64

If we look at the description of the dataset below we get the mean value , standard deviation and distribution of the dataset in quarters along with the min and max values of all the 9 attributes.




	count	mean	std	min	25%	50%	75%	max
Age	444.0	27.747748	4.416710	18.0	25.0	27.0	30.000	43.0
Engineer	444.0	0.754505	0.430866	0.0	1.0	1.0	1.000	1.0
MBA	444.0	0.252252	0.434795	0.0	0.0	0.0	1.000	1.0
Work Exp	444.0	6.299550	5.112098	0.0	3.0	5.0	8.000	24.0
Salary	444.0	16.238739	10.453851	6.5	9.8	13.6	15.725	57.0
Distance	444.0	11.323198	3.606149	3.2	8.8	11.0	13.425	23.4
license	444.0	0.234234	0.423997	0.0	0.0	0.0	0.000	1.0

Now let us look at the value counts of Transport and Gender.



Transport	
Public Transport	300
Private Transport	144

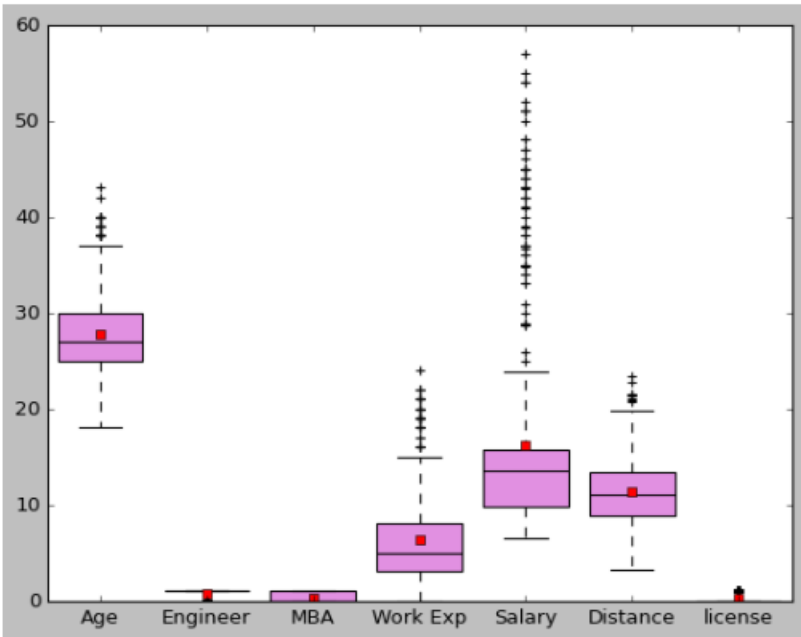
dtype: int64



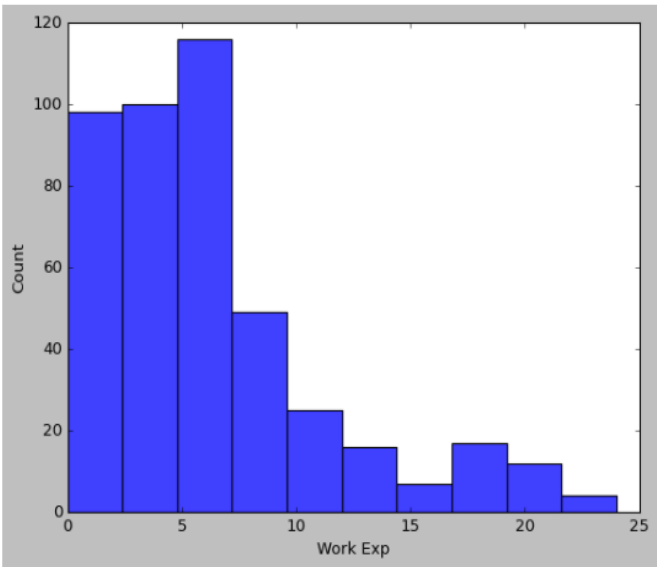
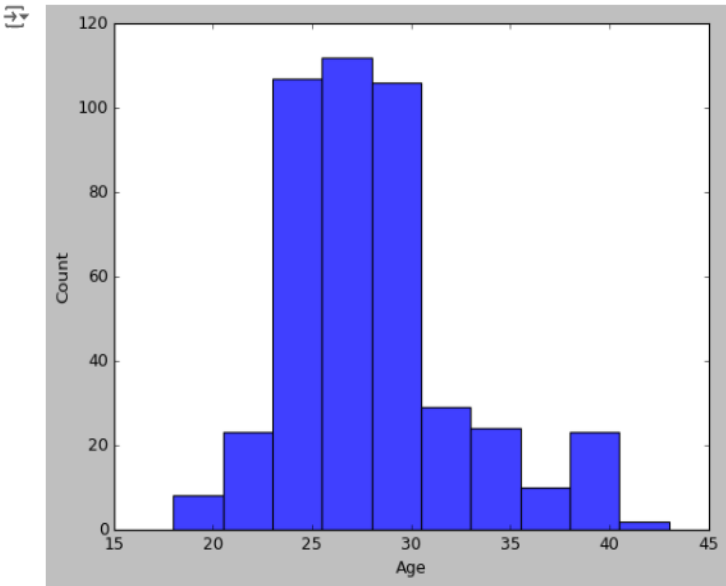
Gender	
Male	316
Female	128

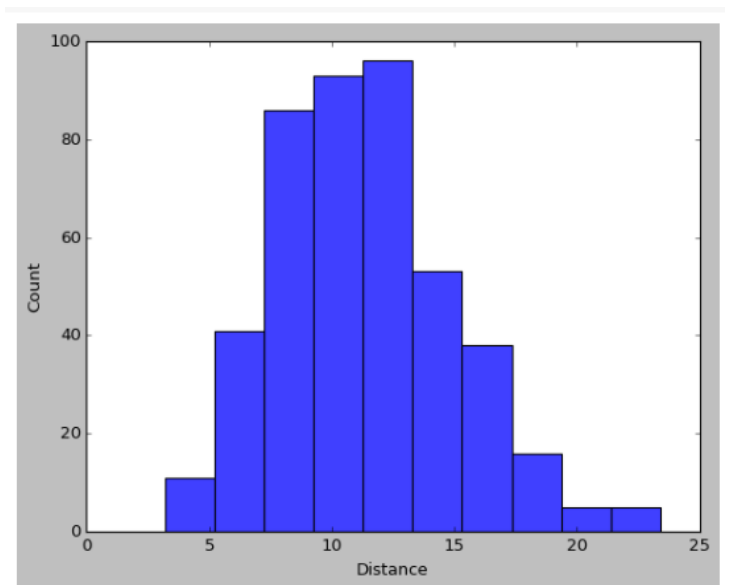
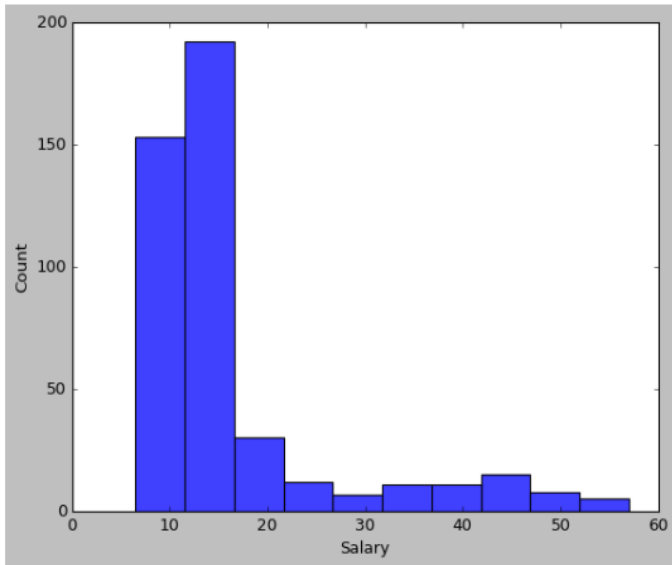
dtype: int64

If we look at the boxplot, we find that the dataset contains outliers which needs to be treated. The column Salary has been seen with huge outliers as compared to other columns. Work experience, Distance and Age also has outliers in them.

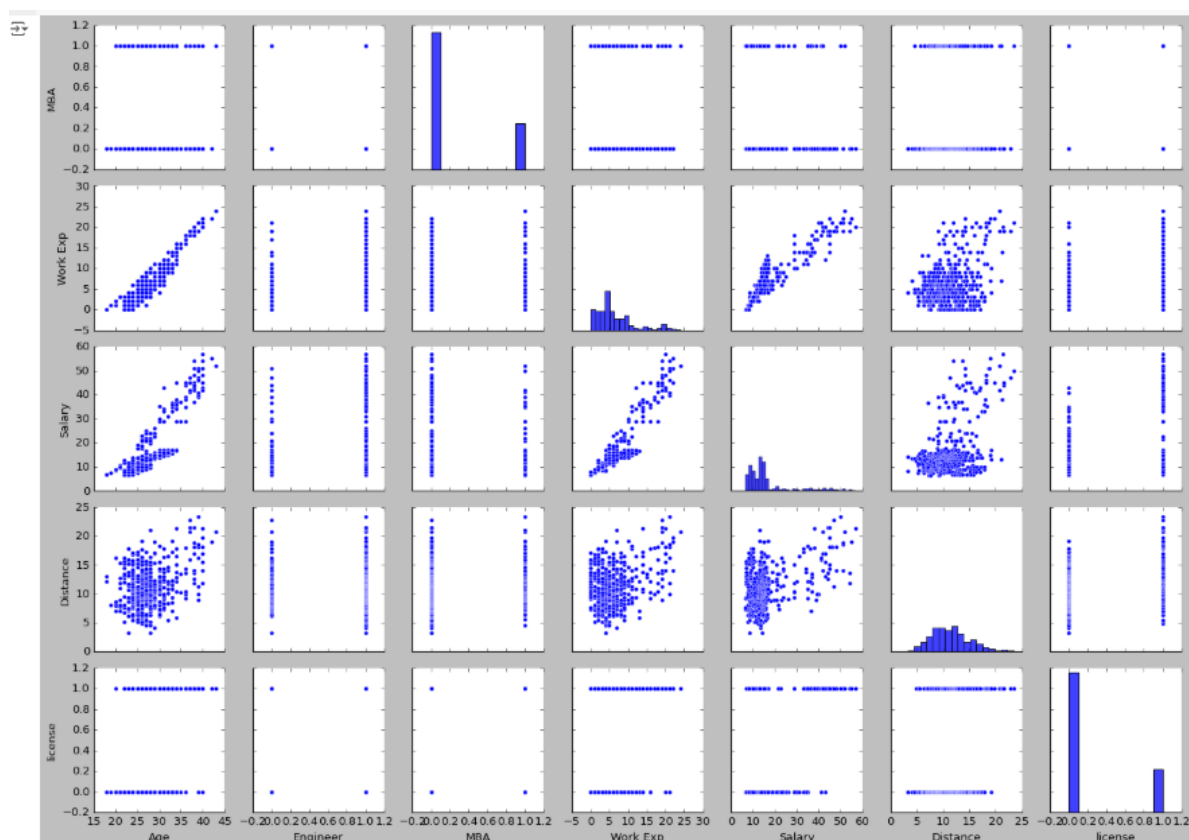


Following are the histograms of different variables:





Let us check the pair plot:



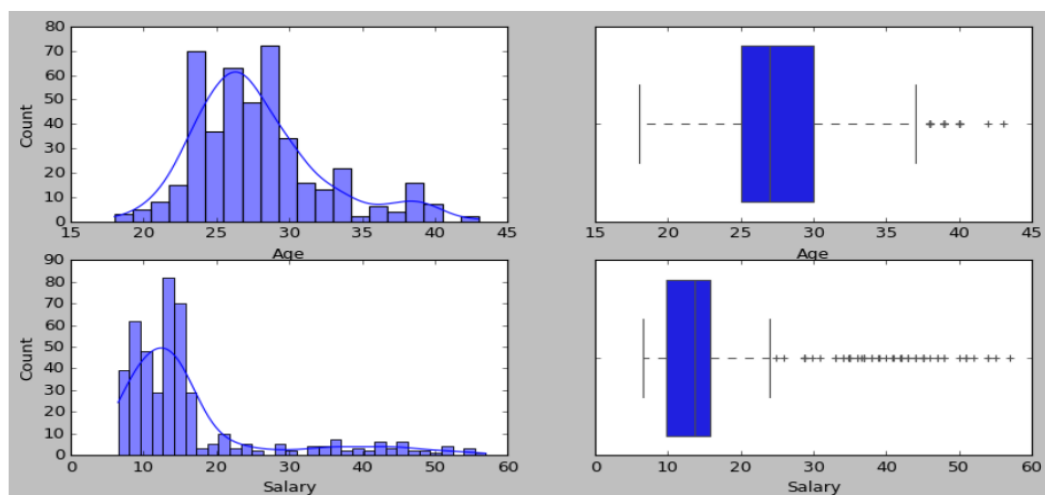
Let us check the Correlation Matrix: we see a strong relation between salary and work experience. We also see a strong relation between Age, Salary and Work Experience.

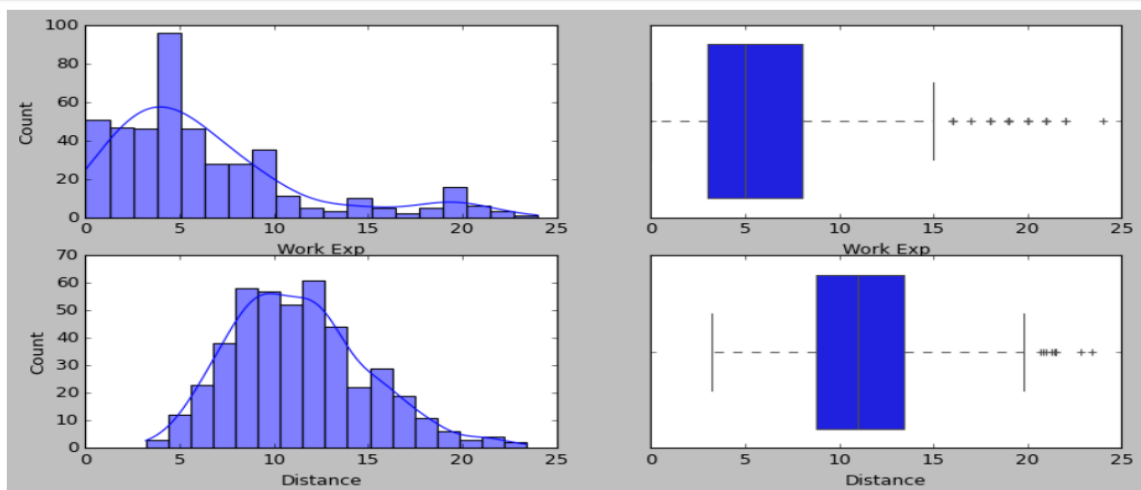


[↕]

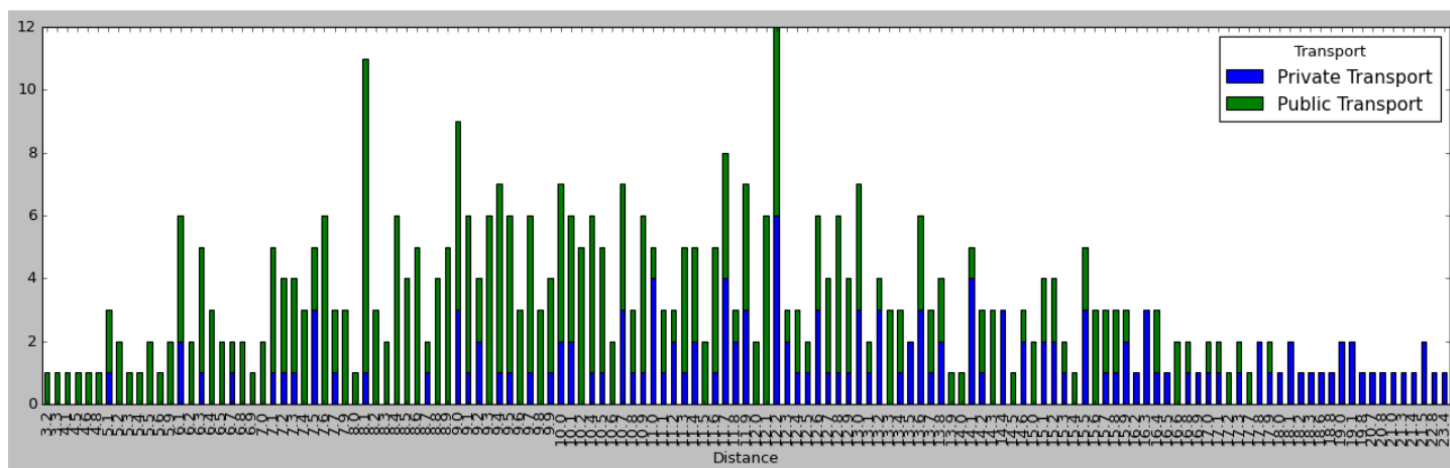
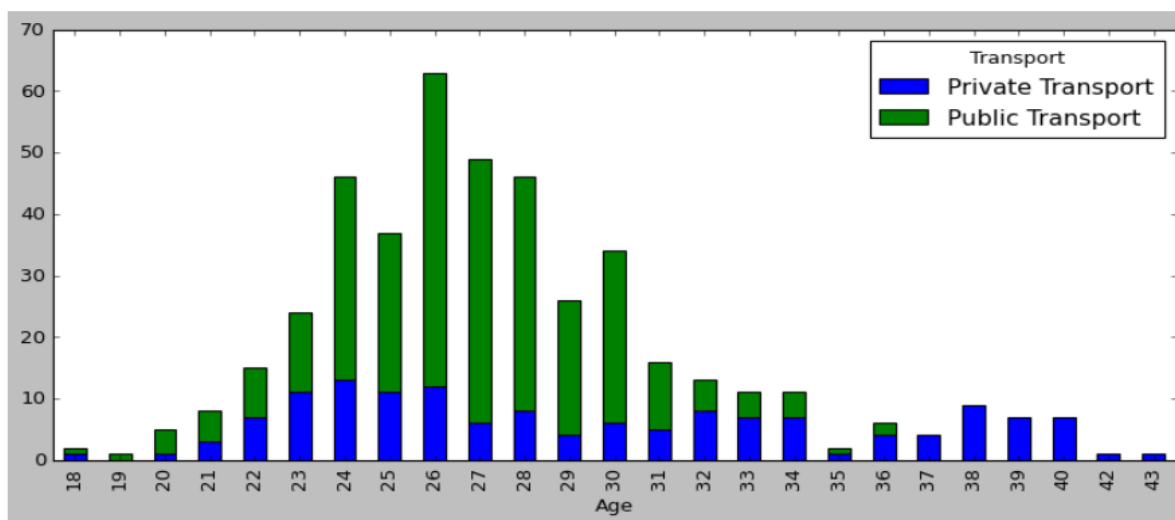
	Age	Engineer	MBA	Work Exp	Salary	Distance	license
Age	1.000000	0.091935	-0.029090	0.932236	0.860673	0.352872	0.452311
Engineer	0.091935	1.000000	0.066218	0.085729	0.086762	0.059316	0.018924
MBA	-0.029090	0.066218	1.000000	0.008582	-0.007270	0.036427	-0.027358
Work Exp	0.932236	0.085729	0.008582	1.000000	0.931974	0.372735	0.452867
Salary	0.860673	0.086762	-0.007270	0.931974	1.000000	0.442359	0.508095
Distance	0.352872	0.059316	0.036427	0.372735	0.442359	1.000000	0.290084
license	0.452311	0.018924	-0.027358	0.452867	0.508095	0.290084	1.000000

Let us go ahead with some more Univariate and Bivariate Analysis:

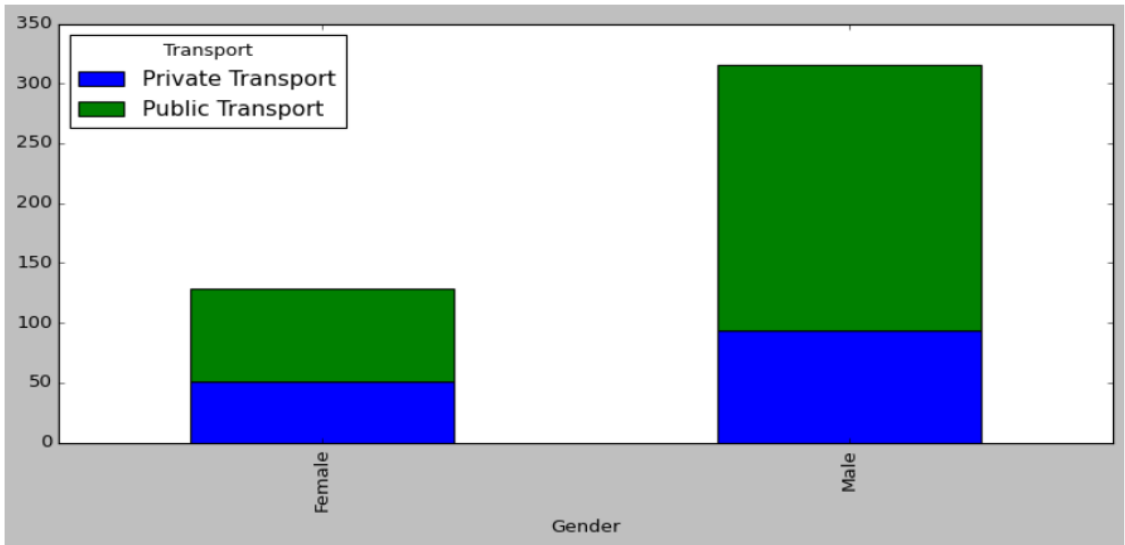
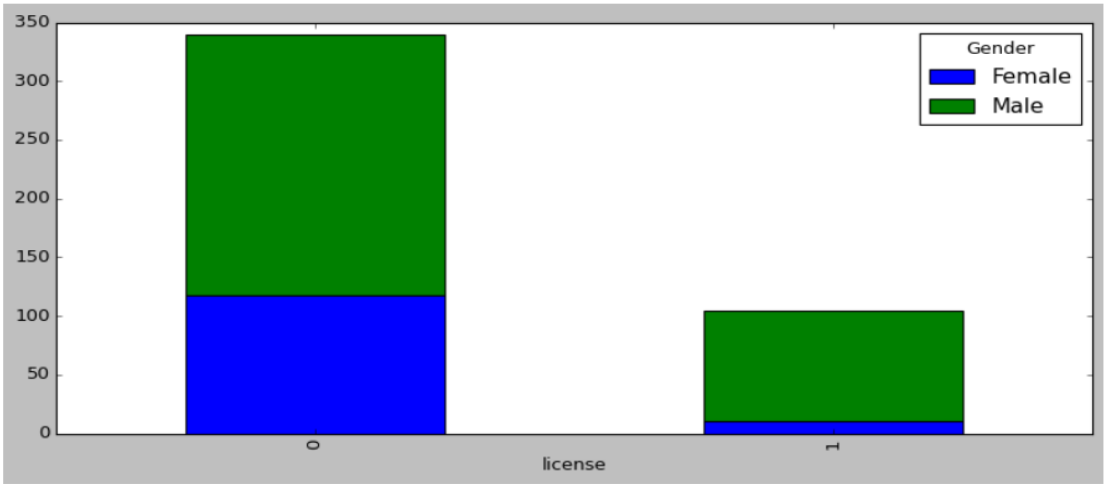
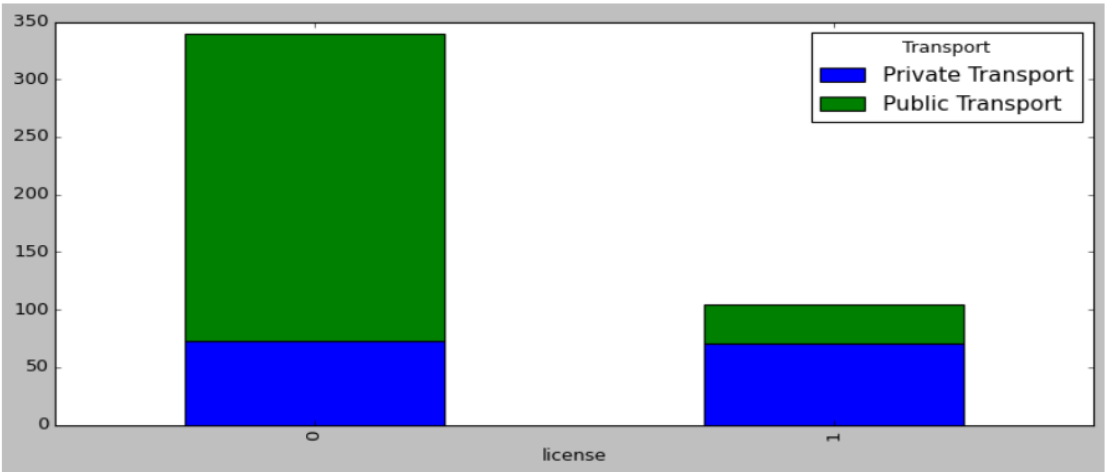




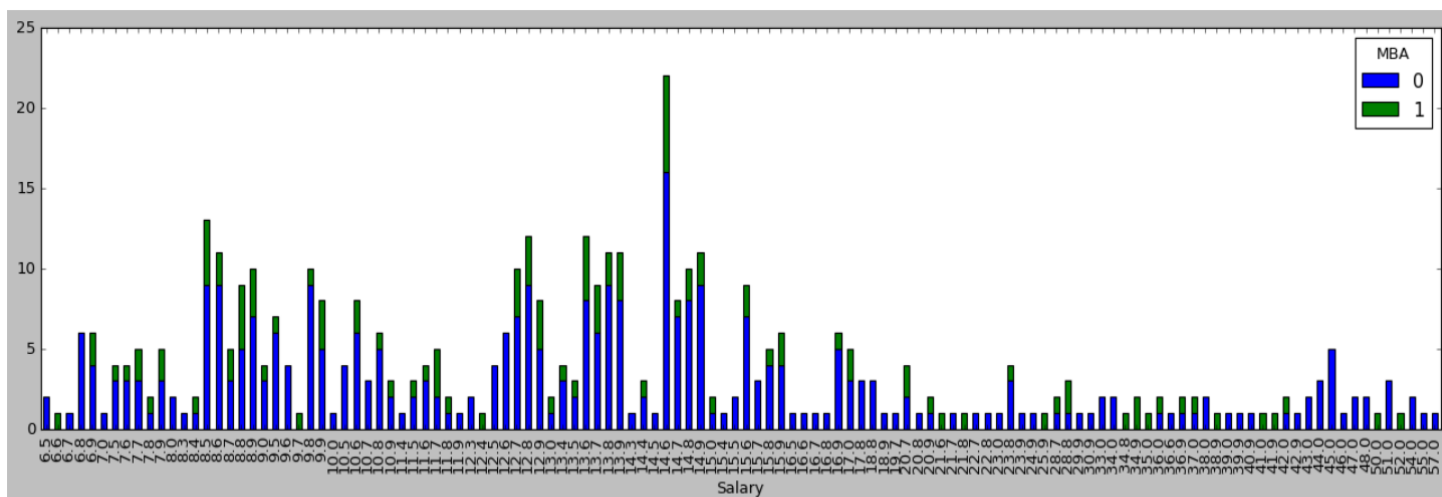
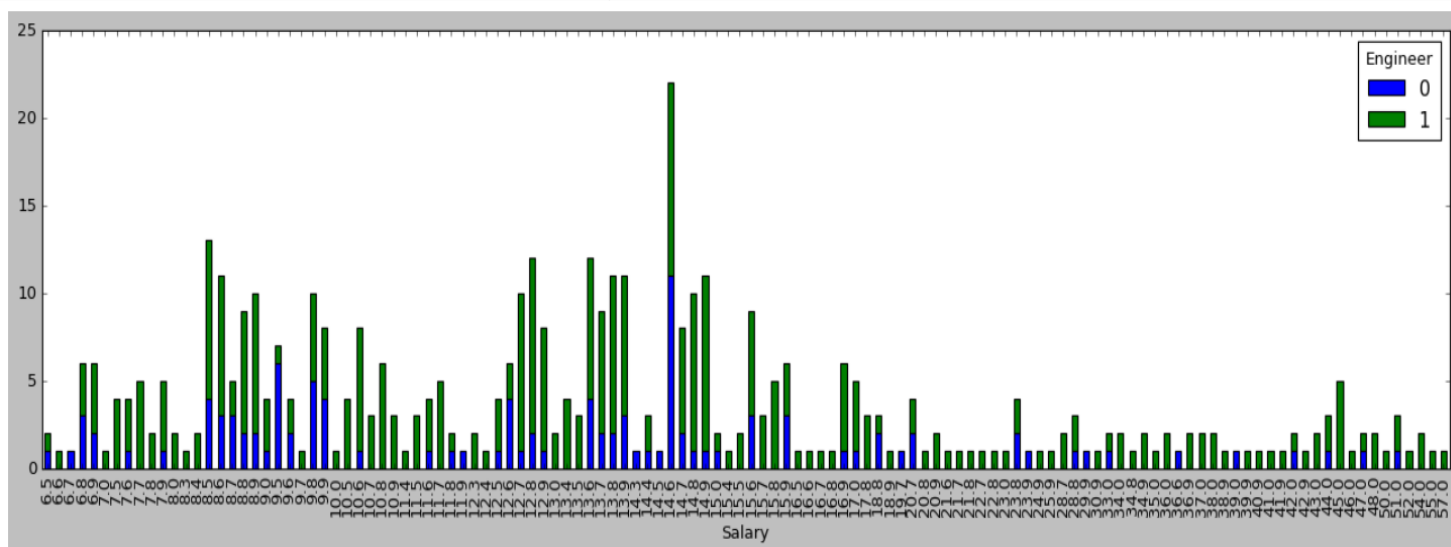
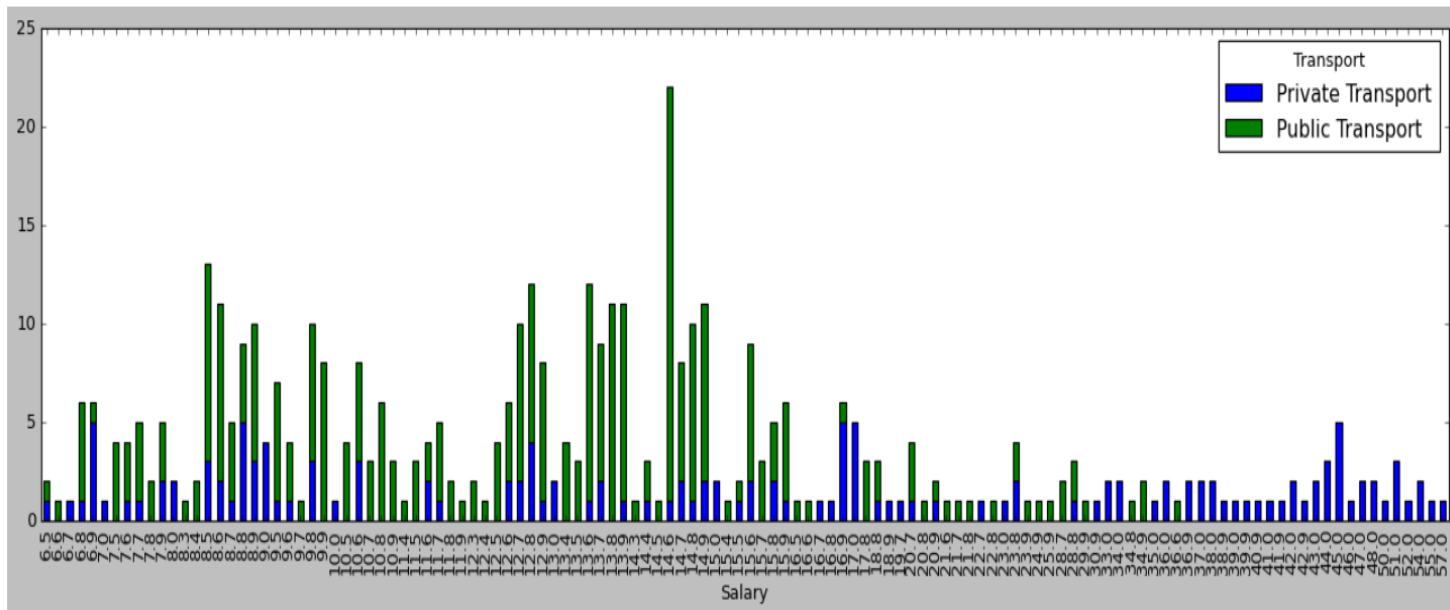
Here is the crosstab of Age against Transport and Distance against Transport:



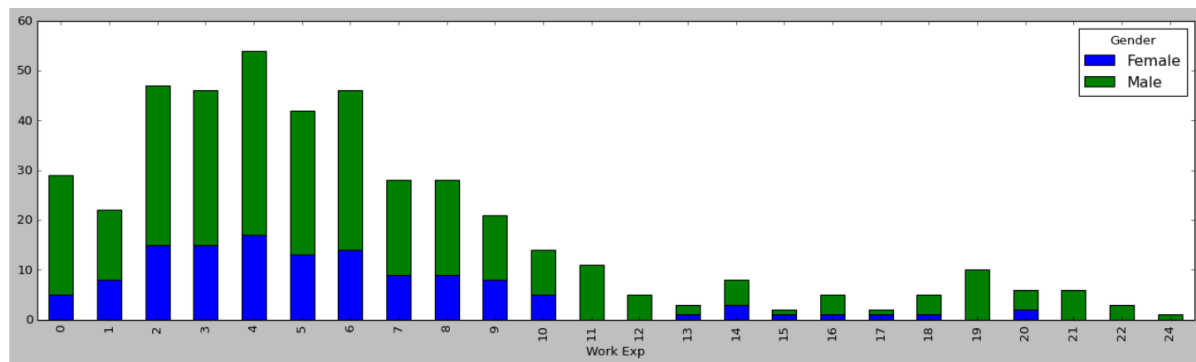
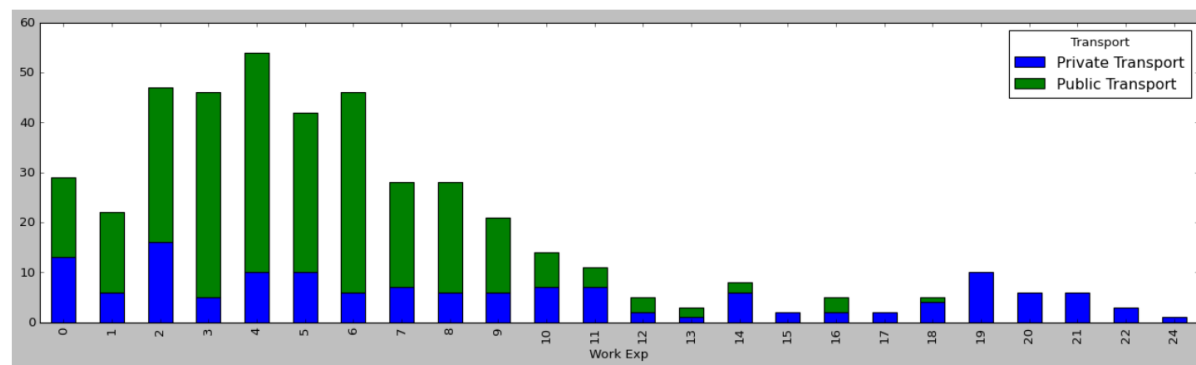
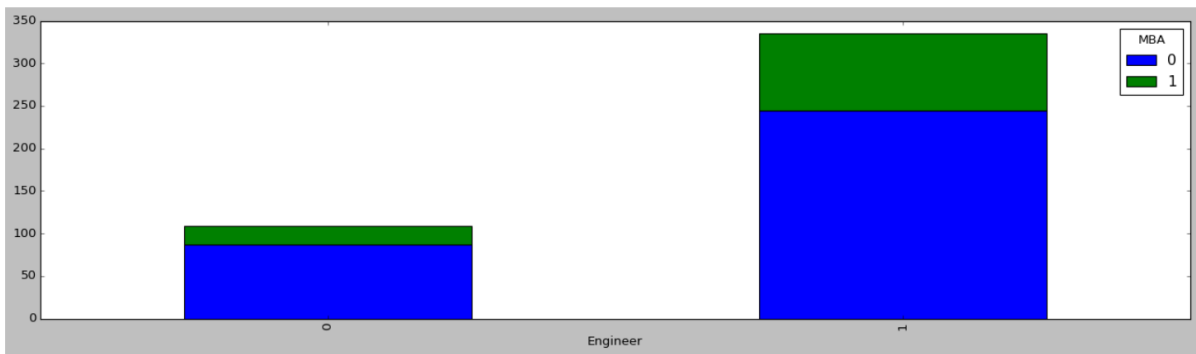
Here are the crosstabs of License against Transport, License against Gender and Gender against Transport.



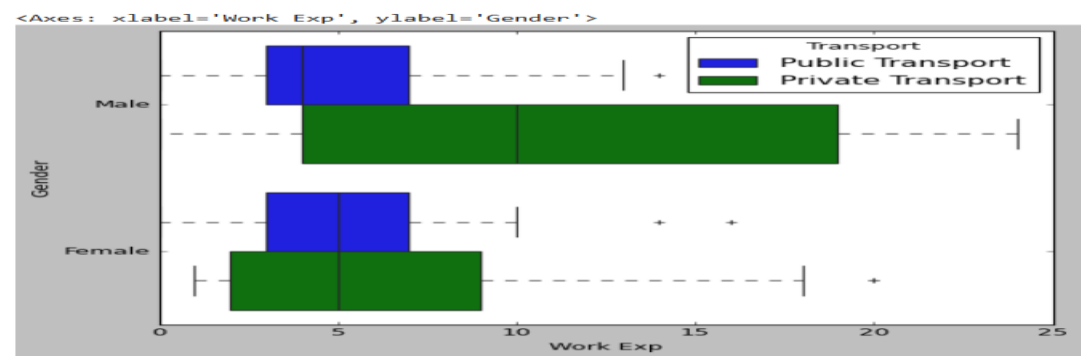
Moving on with further analysis we have crosstab of Salary against Transport, Salary against Engineer and Salary against MBA.

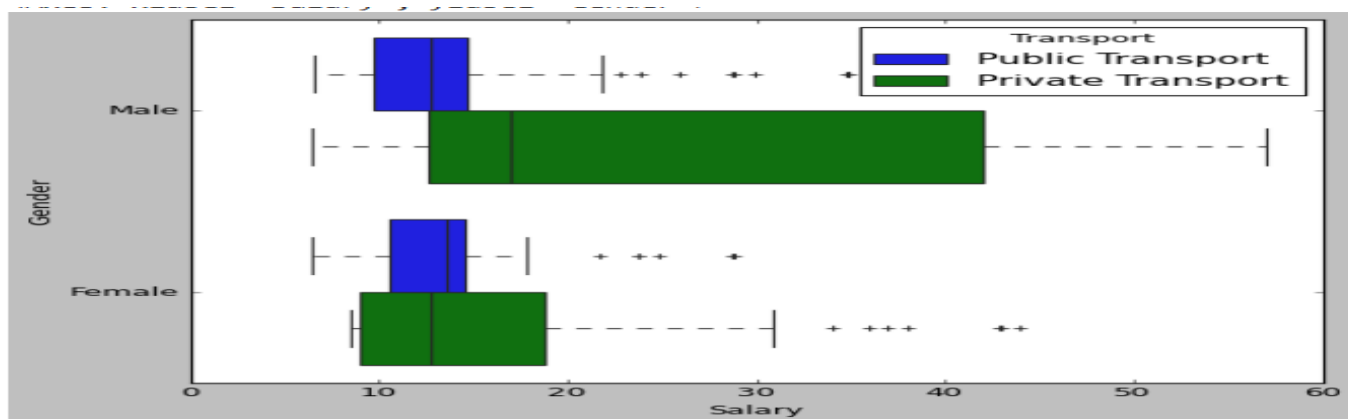


Below are the crosstabs of Engineer against MBA, Work Experience against Transport, Work Experience against Gender.

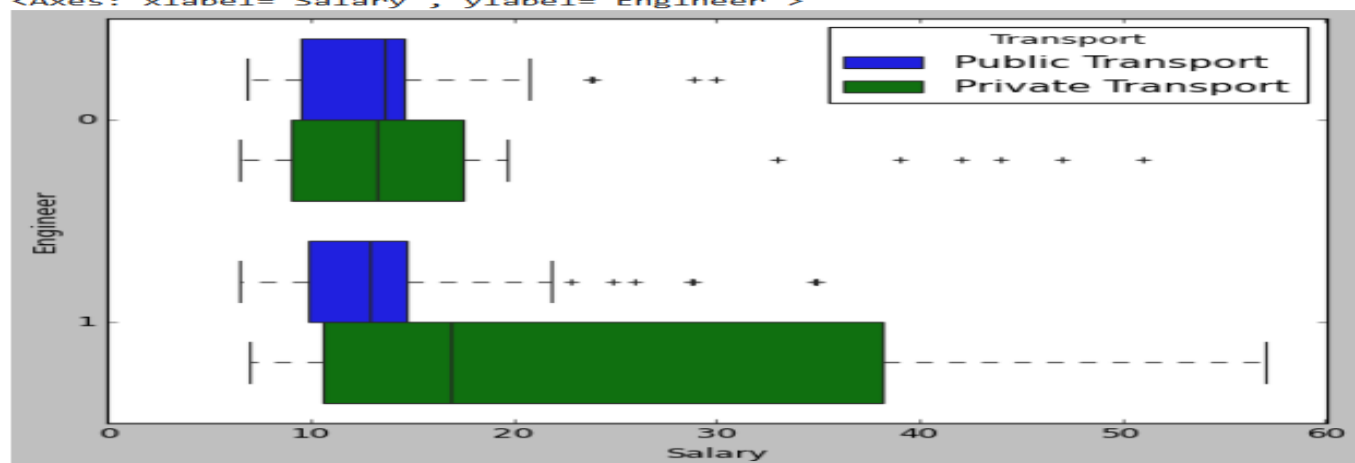


As we dive deep into the dataset, we create some boxplots for multivariate analysis such as:

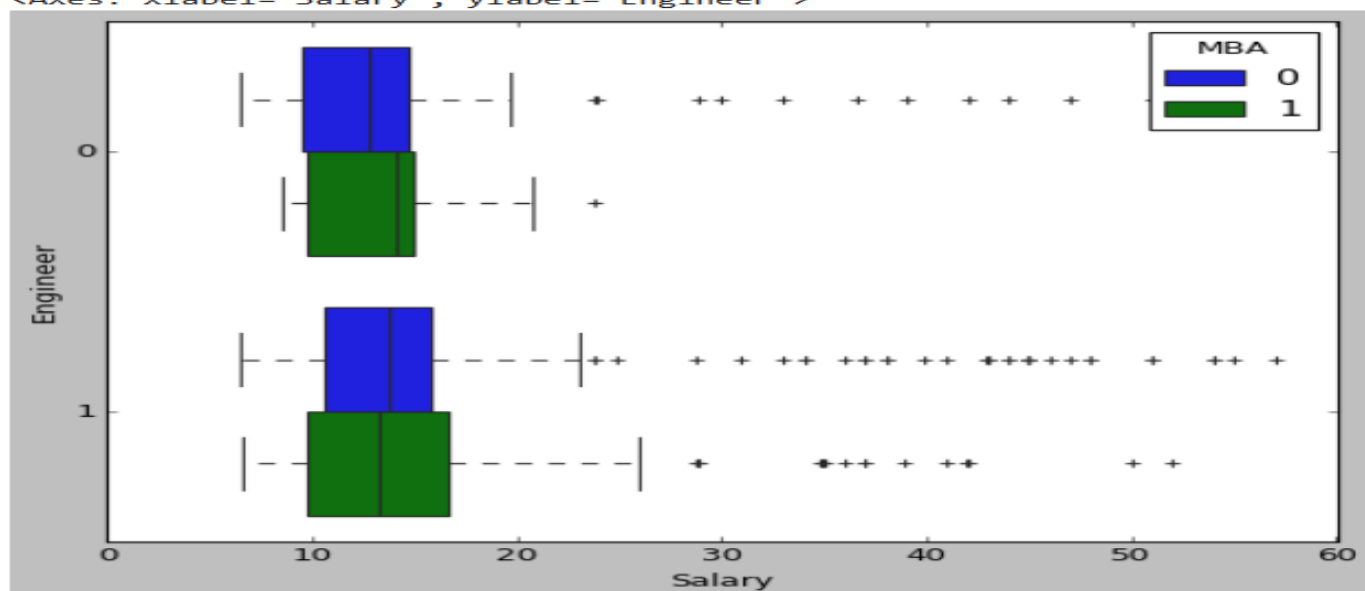




<Axes: xlabel='Salary', ylabel='Engineer'>



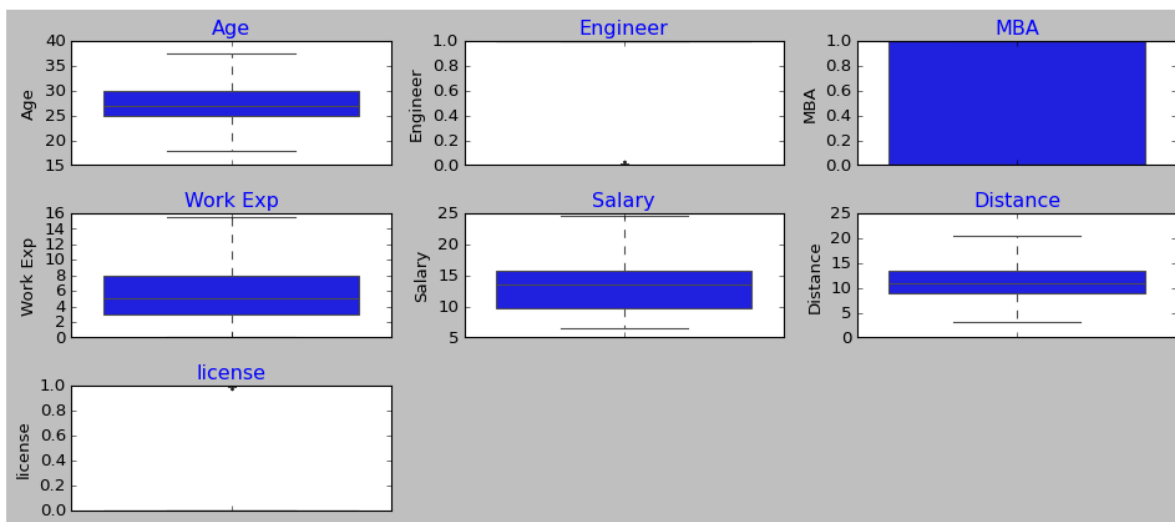
<Axes: xlabel='Salary', ylabel='Engineer'>



INSIGHTS ON EDA:

- There are outliers in work experience, salary, distance and Age which needs to be treated.
- Positive correlation between work experience, age and salary.
- Employees who do not have Driving License prefer Public Transport.
- Male Employees have more Driving Licenses as compared to Female Employees.
- Female Employees use Public Transport and Private Transport equally however male employees tend to use Public Transport more than Private Transport.
- Employees with higher Salary tend to use Private Transport over Public Transport.
- Employees are mostly engineers but not necessarily possess MBA Degree.

Next after treating the outliers by capping and flooring we see that the dataset is now clean and ready for further analysis.



2. Split the data into train and test in the ratio 70:30. Is scaling necessary or not?

After splitting the dataset into Train and Test set in 70:30 ratio.

	Age	Engineer	MBA	Work Exp	Salary	Distance	license	Gender_Male
0	28.0	0	0	4.0	14.3	3.2	0	True
1	23.0	1	0	4.0	8.3	3.3	0	False
2	29.0	1	0	7.0	13.4	4.1	0	True
3	28.0	1	1	5.0	13.4	4.5	0	False
4	27.0	1	0	4.0	13.4	4.6	0	True

	Transport_Public	Transport
0		True
1		True
2		True
3		True
4		True

3. Build the following models on the 70% training data and check the performance of these models on the Training as well as the 30% Test data using the various inferences from the Confusion Matrix and plotting a AUC-ROC curve along with the AUC values. Tune the models wherever required for optimum performance.:

a. Logistic Regression Model

Ans. After creating the logistic regression model have checked the performance matrix on train and test dataset.

0.8290322580645161

[[61 40]

[13 196]]

	precision	recall	f1-score	support
False	0.82	0.60	0.70	101
True	0.83	0.94	0.88	209
accuracy			0.83	310
macro avg	0.83	0.77	0.79	310
weighted avg	0.83	0.83	0.82	310

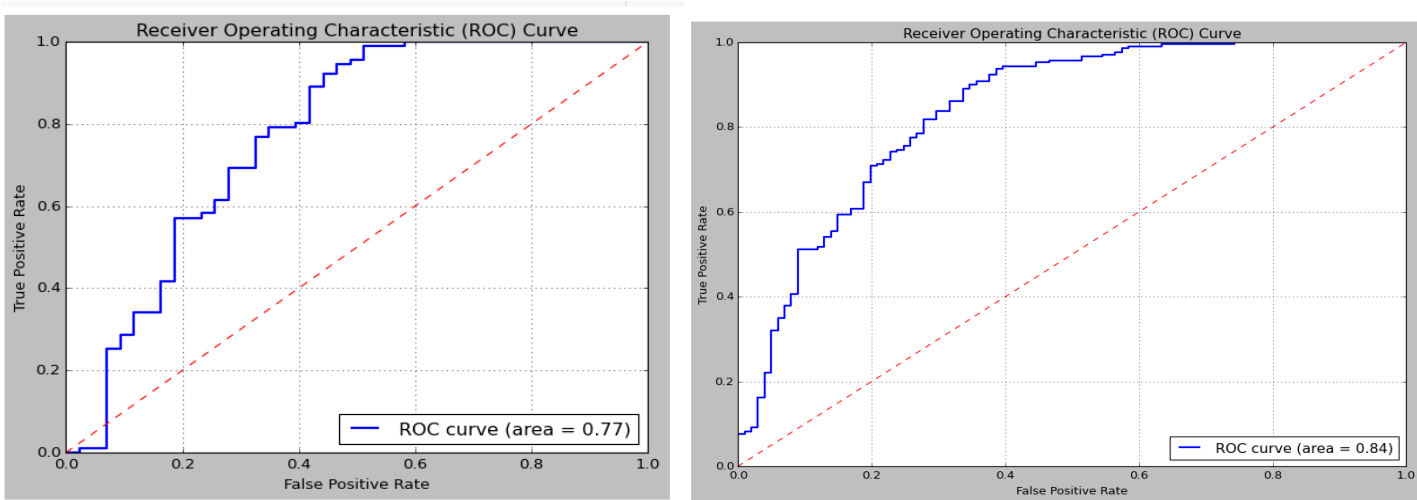
0.8059701492537313

[[23 20]

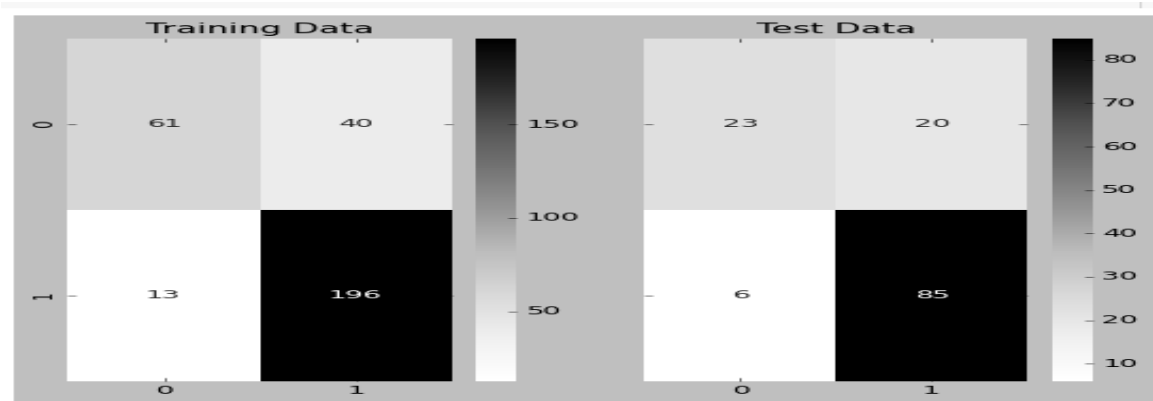
[6 85]]

	precision	recall	f1-score	support
False	0.79	0.53	0.64	43
True	0.81	0.93	0.87	91
accuracy			0.81	134
macro avg	0.80	0.73	0.75	134
weighted avg	0.80	0.81	0.79	134

ROC-AUC Curve for Train and Test Data



Confusion Matrix on Train and Test Data:



b. Linear Discriminant Analysis

Ans. Performance matrix on Linear Discriminant Analysis of Train and Test data.

0.8096774193548387

[[58 43]
[16 193]]

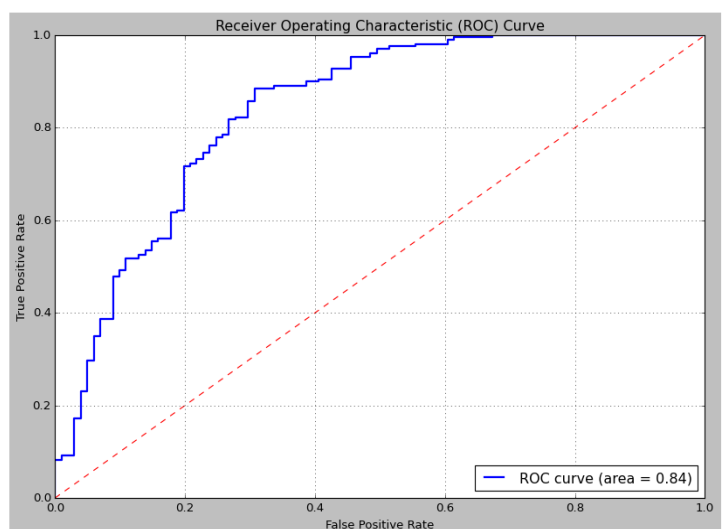
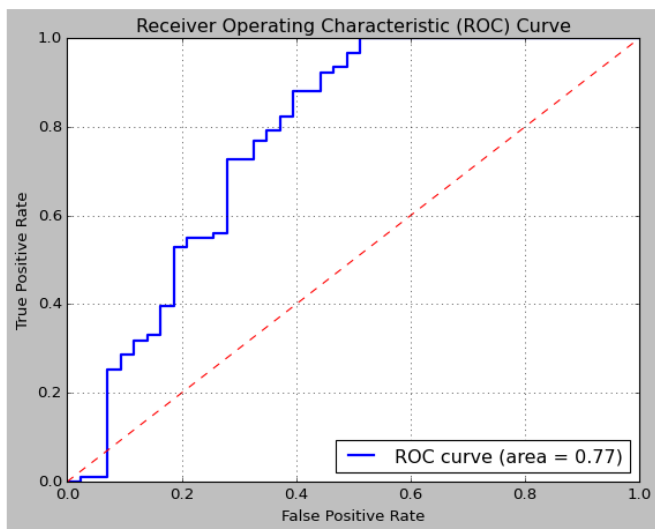
	precision	recall	f1-score	support
False	0.78	0.57	0.66	101
True	0.82	0.92	0.87	209
accuracy			0.81	310
macro avg	0.80	0.75	0.77	310
weighted avg	0.81	0.81	0.80	310

0.7985074626865671

[[22 21]
[6 85]]

	precision	recall	f1-score	support
False	0.79	0.51	0.62	43
True	0.80	0.93	0.86	91
accuracy			0.80	134
macro avg	0.79	0.72	0.74	134
weighted avg	0.80	0.80	0.78	134

ROC-AUC Curve on Train and Test Data



c. Decision Tree Classifier – CART model

Ans. Performance matrix on Decision Tree Classifier of Train and Test data

1.0

[[101 0]
[0 209]]

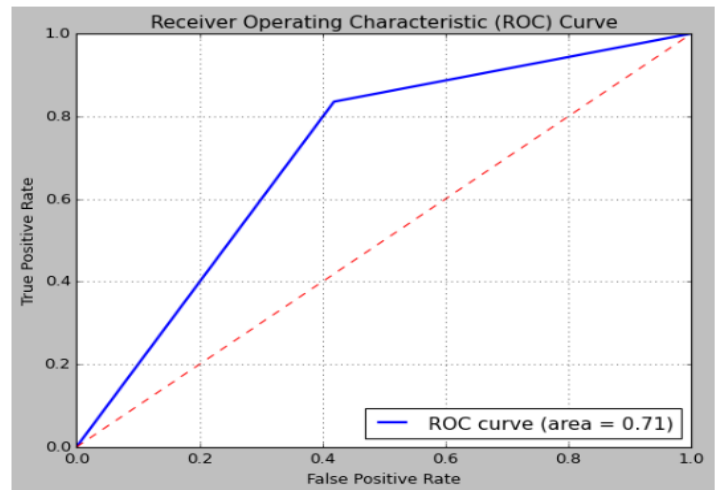
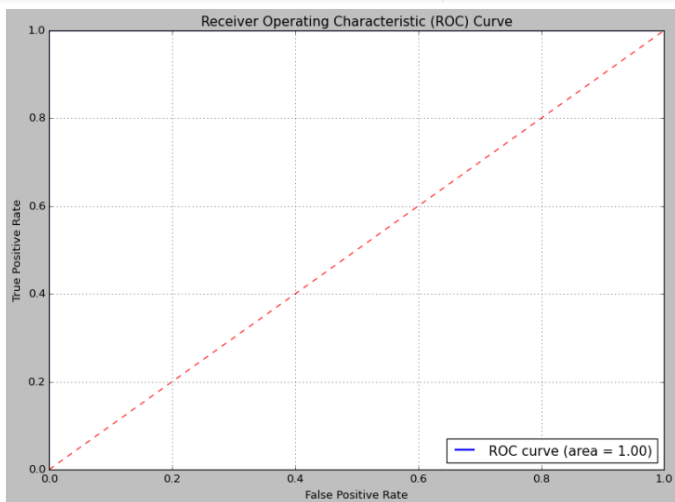
	precision	recall	f1-score	support
False	1.00	1.00	1.00	101
True	1.00	1.00	1.00	209
accuracy			1.00	310
macro avg	1.00	1.00	1.00	310
weighted avg	1.00	1.00	1.00	310

0.753731343283582

[[25 18]
[15 76]]

	precision	recall	f1-score	support
False	0.62	0.58	0.60	43
True	0.81	0.84	0.82	91
accuracy			0.75	134
macro avg	0.72	0.71	0.71	134
weighted avg	0.75	0.75	0.75	134

ROC-AUC Curve on Train and Test Data



d. Naïve Bayes Model

Ans. Performance matrix on Naïve Bayes Model of Train and Test data

0.7967741935483871

[[55 46]
[17 192]]

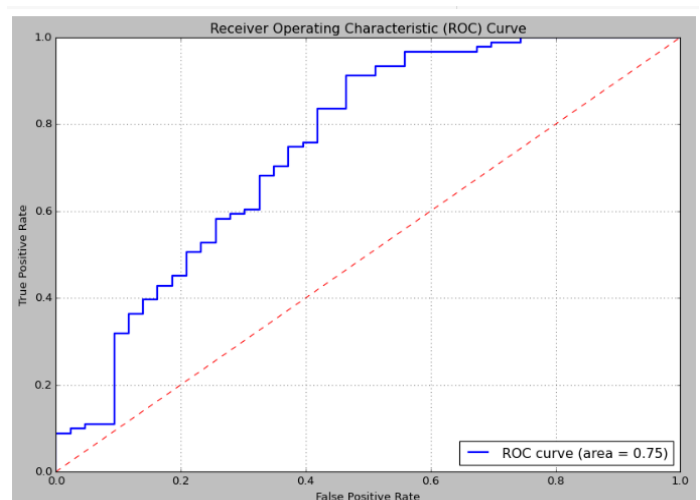
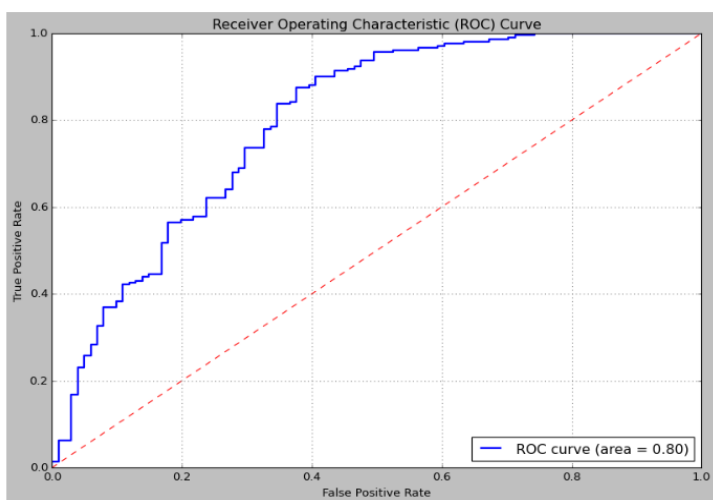
	precision	recall	f1-score	support
False	0.76	0.54	0.64	101
True	0.81	0.92	0.86	209
accuracy			0.80	310
macro avg	0.79	0.73	0.75	310
weighted avg	0.79	0.80	0.79	310

0.7761194029850746

[[21 22]
[8 83]]

	precision	recall	f1-score	support
False	0.72	0.49	0.58	43
True	0.79	0.91	0.85	91
accuracy			0.78	134
macro avg	0.76	0.70	0.72	134
weighted avg	0.77	0.78	0.76	134

ROC-AUC Curve on Train and Test Data



e. KNN Model

Ans. Performance matrix on KNN Model of Train and Test data

0.8451612903225807

[[67 34]
[14 195]]

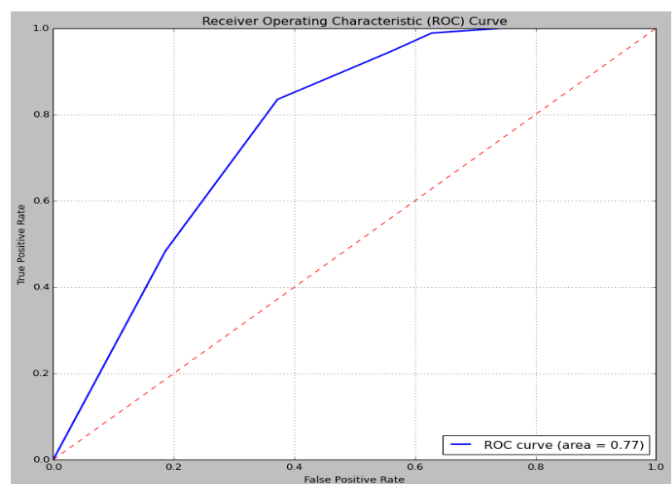
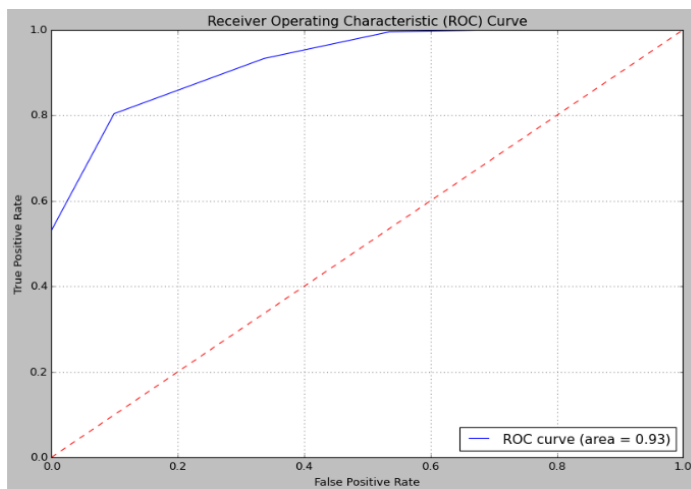
	precision	recall	f1-score	support
False	0.83	0.66	0.74	101
True	0.85	0.93	0.89	209
accuracy			0.85	310
macro avg	0.84	0.80	0.81	310
weighted avg	0.84	0.85	0.84	310

0.7835820895522388

[[19 24]
[5 86]]

	precision	recall	f1-score	support
False	0.79	0.44	0.57	43
True	0.78	0.95	0.86	91
accuracy			0.78	134
macro avg	0.79	0.69	0.71	134
weighted avg	0.78	0.78	0.76	134

ROC-AUC Curve on Train and Test Data



f. Random Forest Model

Ans. Performance matrix on Random Forest Model of Train and Test data

0.8208955223880597

[[24 19]
[5 86]]

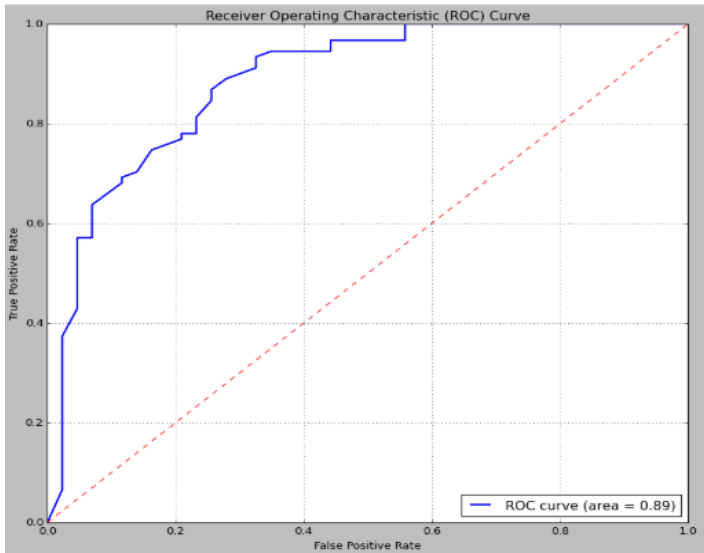
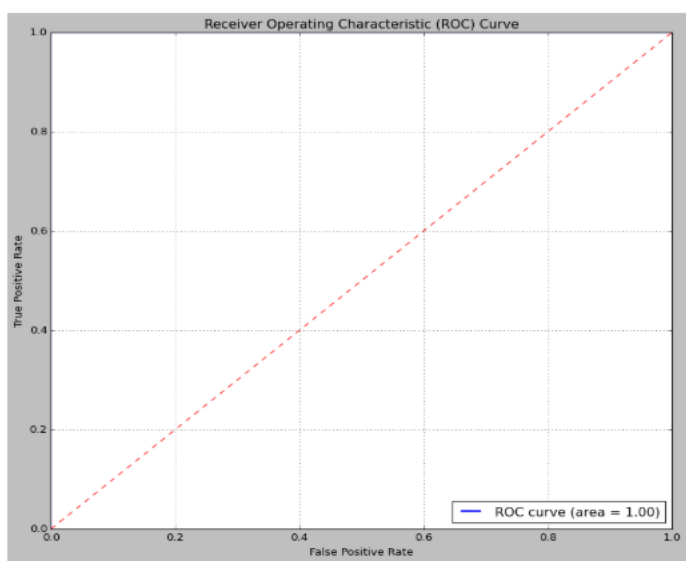
	precision	recall	f1-score	support
False	0.83	0.56	0.67	43
True	0.82	0.95	0.88	91
accuracy			0.82	134
macro avg	0.82	0.75	0.77	134
weighted avg	0.82	0.82	0.81	134

1.0

[[101 0]
[0 209]]

	precision	recall	f1-score	support
False	1.00	1.00	1.00	101
True	1.00	1.00	1.00	209
accuracy			1.00	310
macro avg	1.00	1.00	1.00	310
weighted avg	1.00	1.00	1.00	310

ROC-AUC Curve on Train and Test Data



g. Boosting Classifier Model using Gradient boost.

Ans. Performance matrix on Gradient Boost of Train and Test data

0.967741935483871

[[93 8]
[2 207]]

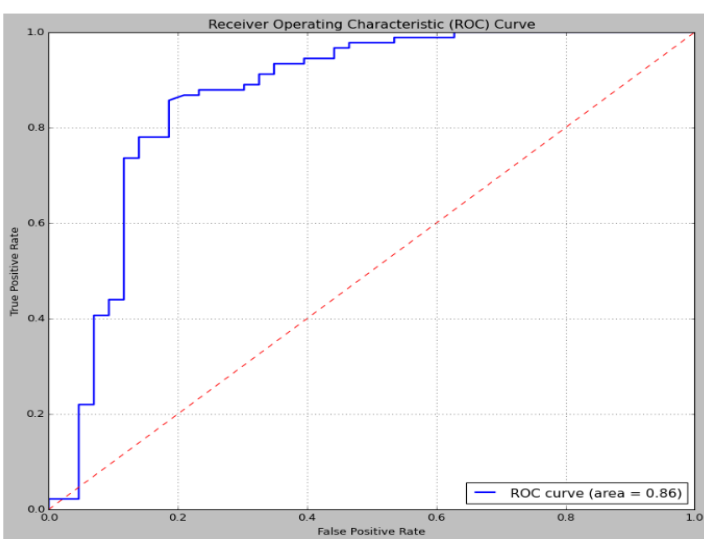
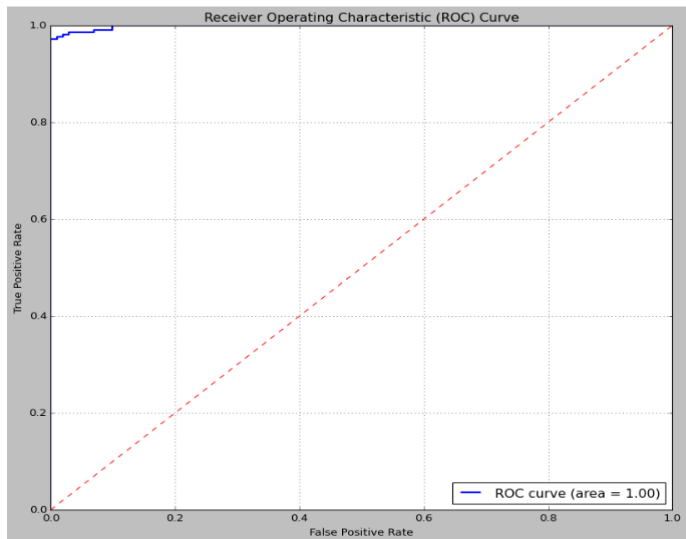
	precision	recall	f1-score	support
False	0.98	0.92	0.95	101
True	0.96	0.99	0.98	209
accuracy			0.97	310
macro avg	0.97	0.96	0.96	310
weighted avg	0.97	0.97	0.97	310

0.835820895522388

[[26 17]
[5 86]]

	precision	recall	f1-score	support
False	0.84	0.60	0.70	43
True	0.83	0.95	0.89	91
accuracy			0.84	134
macro avg	0.84	0.77	0.79	134
weighted avg	0.84	0.84	0.83	134

ROC-AUC Curve on Train and Test Data



INSIGHTS:

- The recall value for Logistic Regression model for Train and Test Data are 94% and 93% respectively which indicates that the model has been created really well and is performing decently on both train and test data giving an accuracy of 84% and 81% respectively.
- If we look at the precision and recall values for training and test data for LDA, they are 80% and 92% on average on both the sets which is also fair. Accuracy on both the sets on this model gives around 80%.
- The precision, recall and accuracy score on the training data gives 100% for Decision tree classifier however the test data gives a recall value of 84% for this model with an accuracy of 75% only.
- The recall value of Naïve Bayes Model is also decent and similar for both training and test dataset with the 92% score. The Accuracy for this model is 80%.
- The KNN Model gives a recall value of 93% on the Training data and 95% on the test data.
- Random Forest Model gives a recall value of 100% on the training data and 95% on the test data with 82% accuracy.
- Gradient Boost gives a recall value of 99% and 95% on the training and test data respectively.

4. Which model performs the best?

Ans. After having compared all the models mentioned above, we have found that all the models are performing decently well with an average score of around 85 but we choose Random Forest Model as both precision & recall is higher for both classes in Training & test Set. All the models are having an Accuracy score where the Training set is higher as compared to Testing Set but within Industrial standards (within 10%).

5. What are your business insights?

Ans.

	Feature	Importance
4	Salary	0.229326
5	Distance	0.223205
0	Age	0.188184
3	Work Exp	0.170251
6	license	0.077625
7	Gender_Male	0.056103
1	Engineer	0.031495
2	MBA	0.023811

These are the important features sequentially that plays the most important role in forming the solution. Out of which the most important features are Salary of the employee then comes the Distance between office and home of an employee, then comes the age followed by work experience.

As we have seen from the data that a lot of respondents tend to choose public transport over Private Transport.

Part 2: Text Mining

Dataset for text mining: [Shart Tank Companies.csv](#)

A dataset of Shark Tank episodes is made available. It contains 495 entrepreneurs making their pitch to the VC sharks. You will ONLY use “Description” column for the initial text mining exercise.

1. Pick out the Deal (Dependent Variable) and Description columns into a separate data frame.

Ans. After importing the necessary libraries we check the dataset

	deal	description	episode	category	entrepreneurs	location	website	askedFor	exchangeForStake	valuation	season	shark1	shark2	shark3	shark4	shark5	title	episode-season	Multiple Entrepreneurs	
0	False	Bluetooth device implant for your ear.	1	Novelties	Darrin Johnson	St. Paul, MN	NaN	1000000		15	6666667	1	Barbara Corcoran	Robert Herjavec	Kevin O'Leary	Daymond John	Kevin Harrington	Ionic Ear	1-1	False
1	True	Retail and wholesale pie factory with two reta...	1	Specialty Food	Tod Wilson	Somerset, NJ	http://whybake.com/	460000		10	4600000	1	Barbara Corcoran	Robert Herjavec	Kevin O'Leary	Daymond John	Kevin Harrington	Mr. Tod's Pie Factory	1-1	False
2	True	Ava the Elephant is a godsend for frazzled par...	1	Baby and Child Care	Tiffany Krumins	Atlanta, GA	http://www.avatheelephant.com/	50000		15	333333	1	Barbara Corcoran	Robert Herjavec	Kevin O'Leary	Daymond John	Kevin Harrington	Ava the Elephant	1-1	False
3	False	Organizing, packing, and moving services deliv...	1	Consumer Services	Nick Friedman, Omar Soliman	Tampa, FL	http://collegehunkshaulingjunk.com/	250000		25	1000000	1	Barbara Corcoran	Robert Herjavec	Kevin O'Leary	Daymond John	Kevin Harrington	College Foxes Packing Boxes	1-1	False
4	False	Interactive media centers for healthcare waiti...	1	Consumer Services	Kevin Flannery	Cary, NC	http://www.wispots.com/	1200000		10	12000000	1	Barbara Corcoran	Robert Herjavec	Kevin O'Leary	Daymond John	Kevin Harrington	Wispots	1-1	False

Now let us pick out the DEAL(Dependent Variable) and Description columns but before doing so let us check some basic summary of the dataset.

(495, 19)

If we check the shape of the dataset, there are 495 rows and 19 columns.

	0
deal	0
description	0
episode	0
category	0
entrepreneurs	72
location	0
website	38
askedFor	0
exchangeForStake	0
valuation	0
season	0
shark1	0
shark2	0
shark3	0
shark4	0
shark5	0
title	0
episode-season	0
Multiple Entrepreneurs	0

dtype: int64

And there are no such null values that will impact the problem solution. We are ready to pick out a separate data frame.

	deal	description
0	False	Bluetooth device implant for your ear.
1	True	Retail and wholesale pie factory with two reta...
2	True	Ava the Elephant is a godsend for frazzled par...
3	False	Organizing, packing, and moving services deliv...
4	False	Interactive media centers for healthcare waiti...
...
490	True	Zoom Interiors is a virtual service for interi...
491	True	Spikeball started out as a casual outdoors gam...
492	True	Shark Wheel is out to literally reinvent the w...
493	False	Adriana Montano wants to open the first Cat Ca...
494	True	Sway Motorsports makes a three-wheeled, all-el...

495 rows x 2 columns

This is how the data looks.

2. **Create two corpora, one with those who secured a Deal, the other with those who did not secure a deal.**
 Ans. After the separating the data into two corpus the distribution goes as 251 rows for deal_corpus and 244 for no_deal_corpus.

```
(251, 244)
```

3. **The following exercise is to be done for both the corpora:**

- a) **Find the number of characters for both the corporuses.**

Ans. The number of characters in deal_corpus is 64060 and for no_deal_corpus is 47184.

```
deal_characters, no_deal_characters
```

(64060, 47184)

- b) **Remove Stop Words from the corpora. (Words like ‘also’, ‘made’, ‘makes’, ‘like’, ‘this’, ‘even’ and ‘company’ are to be removed)**

Ans.

deal_cleaned_words	no_deal_cleaned_words
['retail', 'wholesale', 'pie', 'factory', 'two', 'retail', 'locations', 'new', 'jersey', 'ava', 'elephant', 'godsend', 'frazzled', 'parents', 'young', 'children',	['bluetooth', 'device', 'implant', 'ear', 'organizing', 'packing', 'moving', 'services', 'delivered', 'college', 'women', 'interactive', 'media', 'centers', 'healthcare', 'waiting', 'rooms', 'offering', 'patients',

Here are some of the words after cleaning both the corpus.

c) What were the top 3 most frequently occurring words in both corpuses (after removing stop words)?

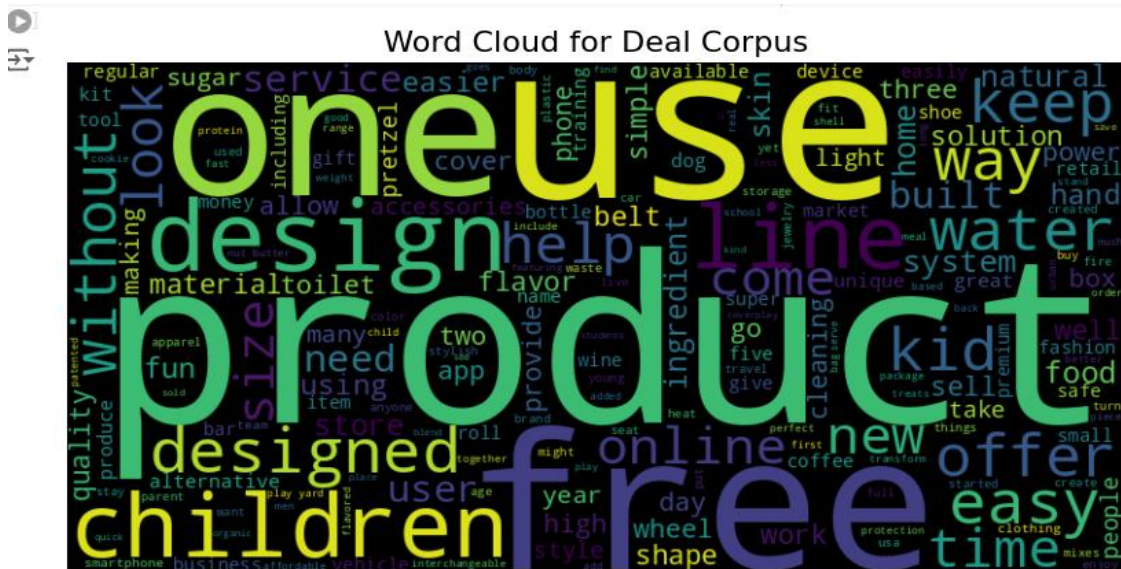
Ans. After cleaning the text on both the corpora we found that in deal corpus the top 3 most frequently used words are “free” which occurred 23 times, “children” that occurred 21 times and “designed” that occurred 21 times. In no deal corpus we see “designed” has occurred 19 times, “use” has occurred 17 times and “water” has occurred 17 times.

deal_top_words, no_deal_top_words

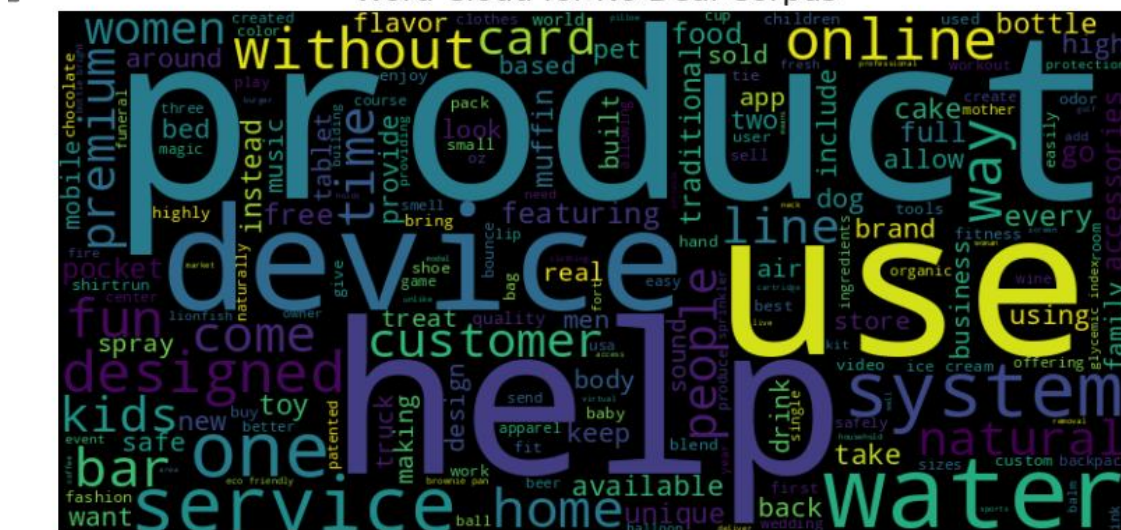
```
⇒ ([('free', 23), ('children', 21), ('designed', 21)],  
    [('designed', 19), ('use', 17), ('water', 17)])
```

d) Plot the Word Cloud for both the corpora

Ans.



Word Cloud for No Deal Corpus



4. Refer to both the word clouds. What do you infer?

Ans. When we see the word cloud of Deal corpus we see words like 'one', 'design', 'free', 'children', 'offer', 'easy', 'online', 'use'. These indicate that Deals aimed towards catering to the children, which provided offers or a free sample/product, was easy to use, had a good design and was unique in its creativity is more likely to secure a deal. We could see that deals which are focused on children's use seem to secure a deal in most of the cases.

The no Deal Corpus word cloud contains words such as 'one', 'designed', 'help', 'device', 'bottle', 'premium', 'use', 'service'. These indicate that Deals with a mediocre design, less suited to solve/help a problem, products involving water bottles, having a higher and premium price tag and less usability are less likely to secure a deal. We could see that the service needs to be better in of the products which made the users not to secure a deal.

It is also observed that words such as 'one', 'designed', 'system' and 'use' have a higher weight in both these word clouds. This indicates that either these were not the defining factors to whether a deal is made or not or might have been used in a different context in the description in each scenario.

5. Looking at the word clouds, is it true that the entrepreneurs who introduced devices are less likely to secure a deal based on your analysis?

Ans. The word 'device' is not easily found in the 'Deal corpus' word cloud while it is easily spotted in the 'No Deal Corpus' word cloud.

This indicates that the word 'device' occurred frequently when a deal was rejected hence implying the statement given in the question is true, which in turn means that the entrepreneurs who introduced devices are less likely to secure a deal due to various factors.

THE END