MACHINE LEARNING

Project

Submitted By- Gunjar Fuley

Contact- 9938126651

Part 1: Machine Learning Models

You work for an office transport company. 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 Kms from Home to Office

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

Transport : Mode of Transport

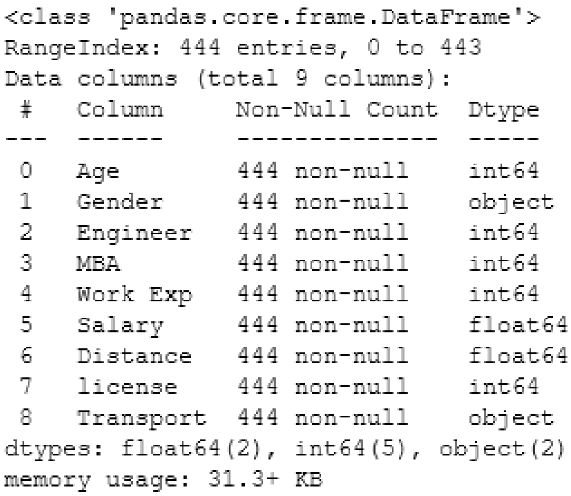
The objective is to build various Machine Learning models on this data set and based on the accuracy metrics decide which model is to be finalised for finally predicting the mode of transport chosen by the employee.

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.

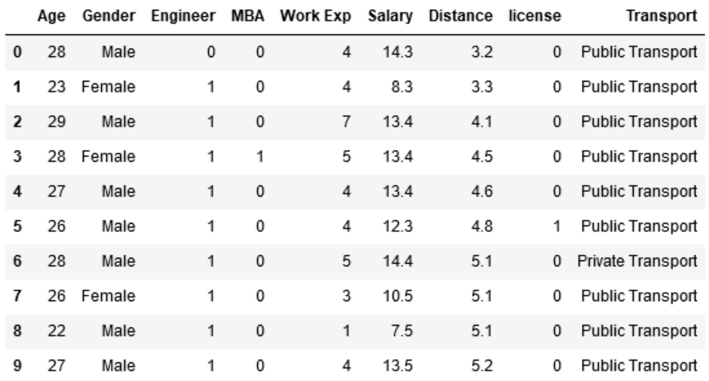
Basic Data Summary

- The data was uploaded in the Jupyter Notebook and all the necessary libraries were imported.
- There are 444 rows and 9 columns 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
```

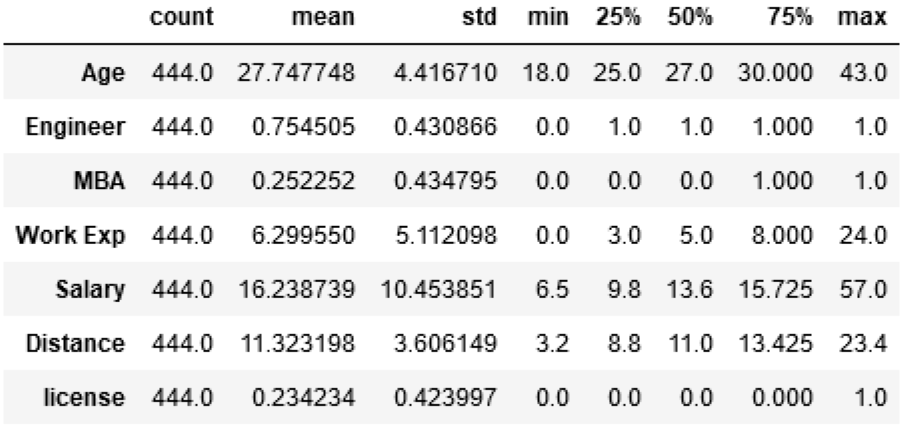
- There are 0 NULL values
- Datatype for all is fine except GENDER & TRANSPORT. It should be categorical.



	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
5	26	Male	1	0	4	12.3	4.8	1	Public Transport
6	28	Male	1	0	5	14.4	5.1	0	Private Transport
7	26	Female	1	0	3	10.5	5.1	0	Public Transport
8	22	Male	1	0	1	7.5	5.1	0	Public Transport
9	27	Male	1	0	4	13.5	5.2	0	Public Transport

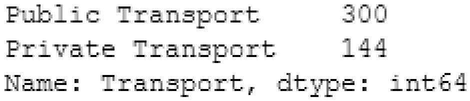
- We need to make private transport as 0 and public transport as 1 because public transport is favourable condition
- We need to treat column 'GENDER' because it is Male/ Female

Basic Descriptive Statistics



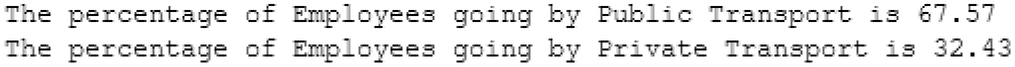
	count	mean	std	min	25%	50%	75%	max
Age	444.0	27.747748	4.416710	18.0	25.0	27.0	30.000	43.0
Engineer	444.0	0.754505	0.430866	0.0	1.0	1.0	1.000	1.0
MBA	444.0	0.252252	0.434795	0.0	0.0	0.0	1.000	1.0
Work Exp	444.0	6.299550	5.112098	0.0	3.0	5.0	8.000	24.0
Salary	444.0	16.238739	10.453851	6.5	9.8	13.6	15.725	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

- Average Age is 27 years, minimum age is 18 & maximum is 43, this means that population is young
- Average work experience is 6 years
- Average salary is 16 Lakhs per annum, min is 6.5 Lakh & max is 57 Lakh. This means that they have good spending capacity
- Average distance is 11 km, which is good for taking a cab to office



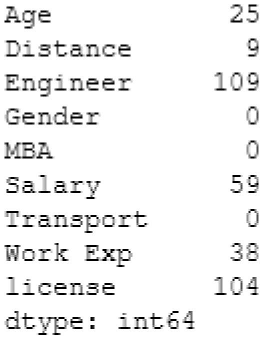
```
Public Transport     300
Private Transport    144
Name: Transport, dtype: int64
```

- Target variable is Transport



```
The percentage of Employees going by Public Transport is 67.57
The percentage of Employees going by Private Transport is 32.43
```

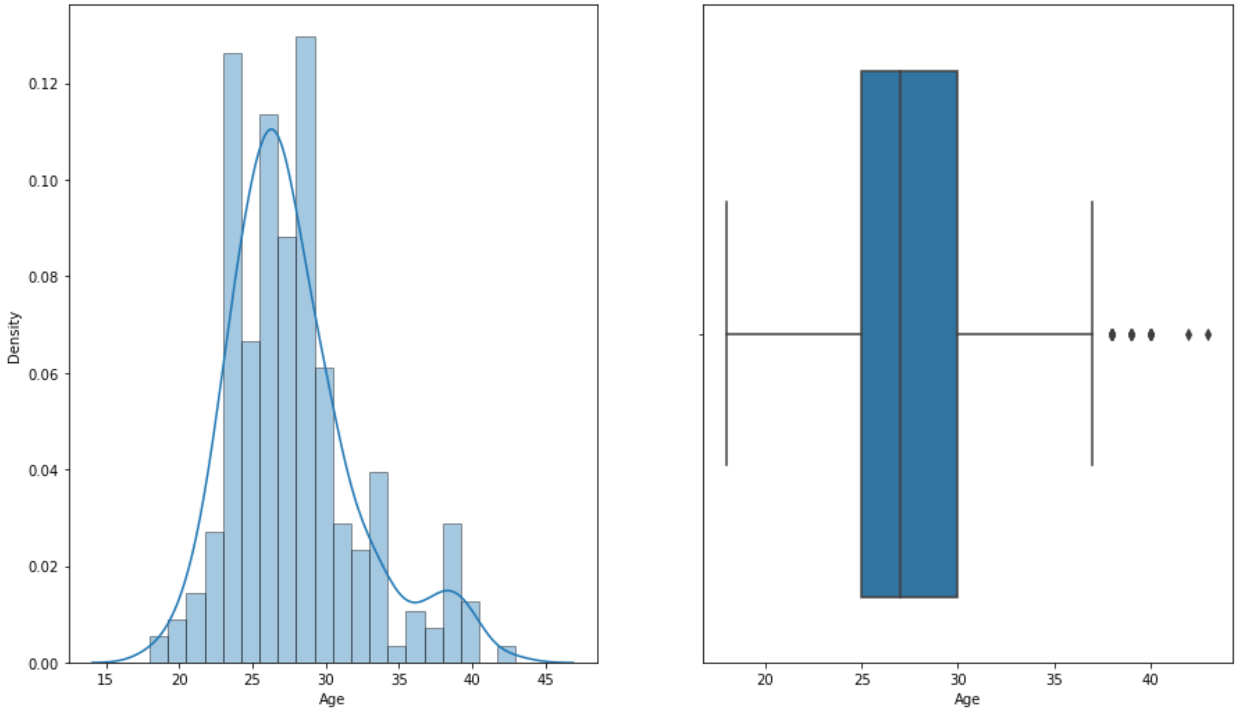
- The data does not seems to be unbalanced



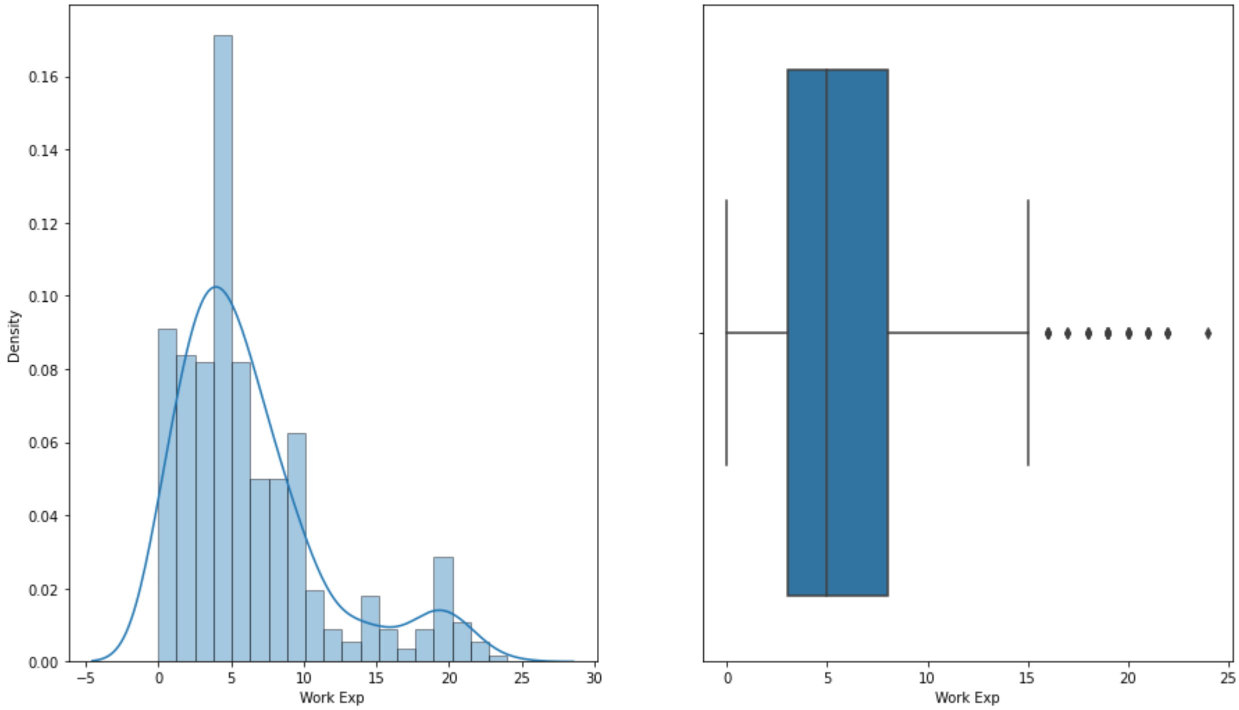
```
Age           25
Distance       9
Engineer     109
Gender         0
MBA            0
Salary        59
Transport      0
Work Exp      38
license      104
dtype: int64
```

- We can see presence of outliers in columns Age, Distance, Engineer, Salary, Work Exp, license (refer above figure)

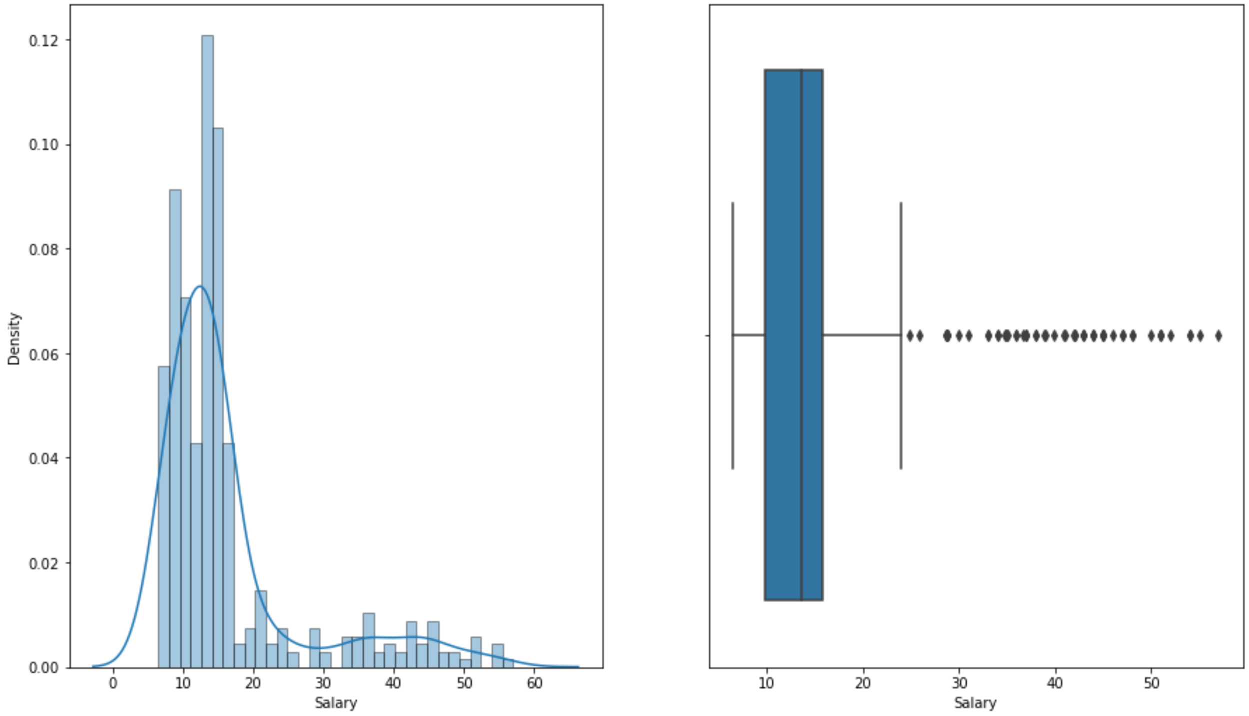
Univariate Analysis



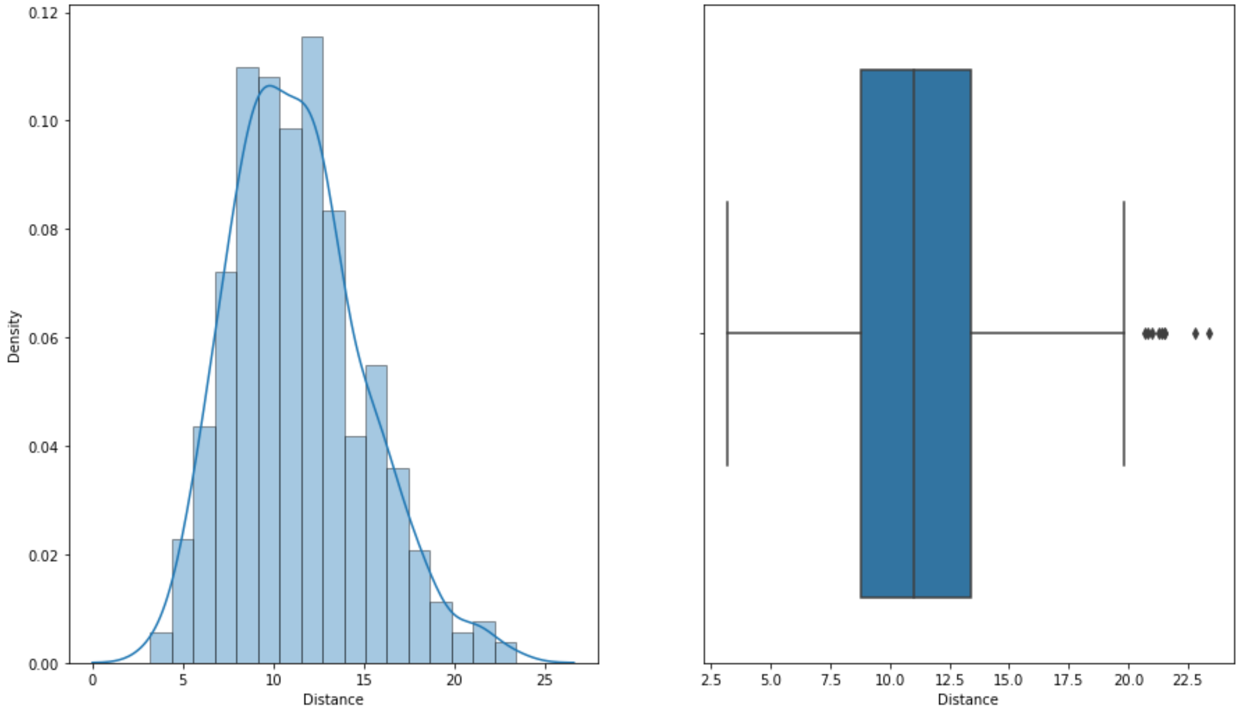
- We can see that variable 'Age' has outliers present in it. It is continuous variable. The data is right skewed.



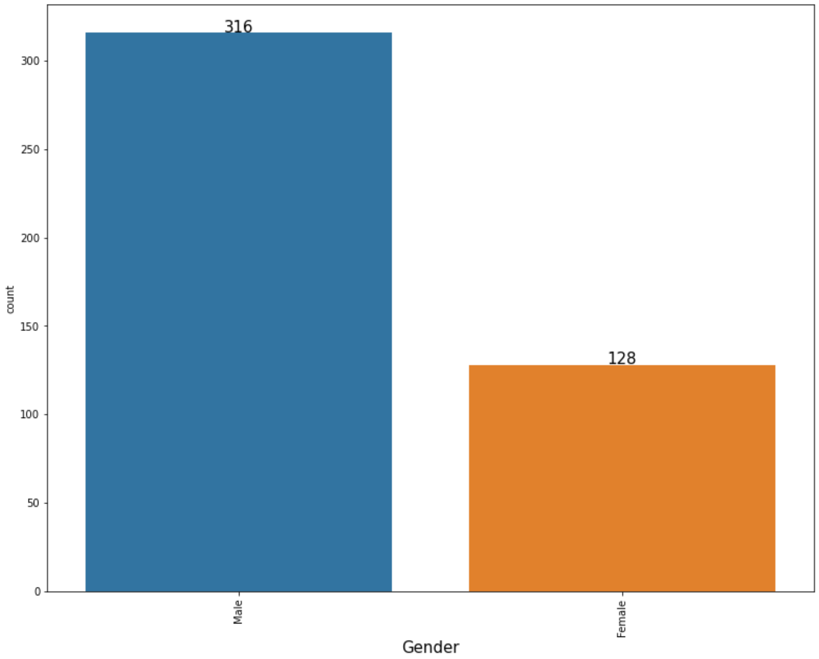
- The variable 'Work Exp' has also outliers present in it. The data is right skewed.



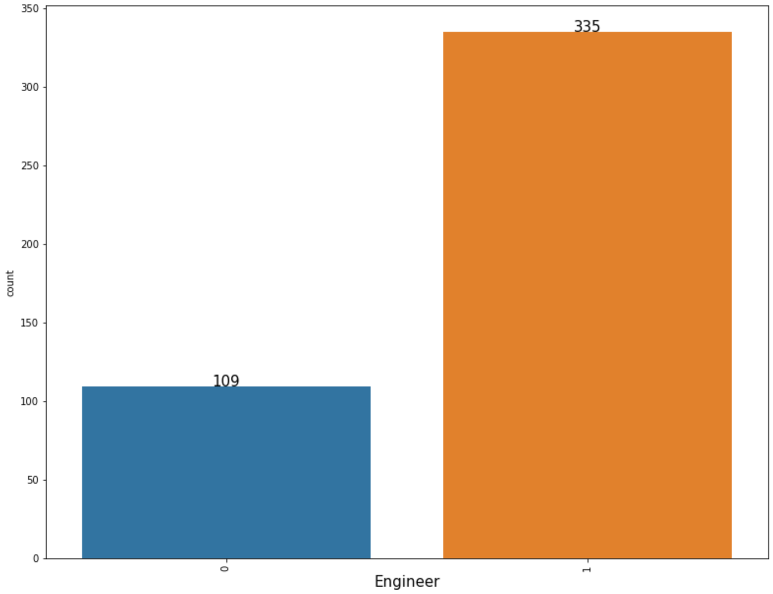
- The variable ‘Salary’ also has outliers present. The data is right skewed.



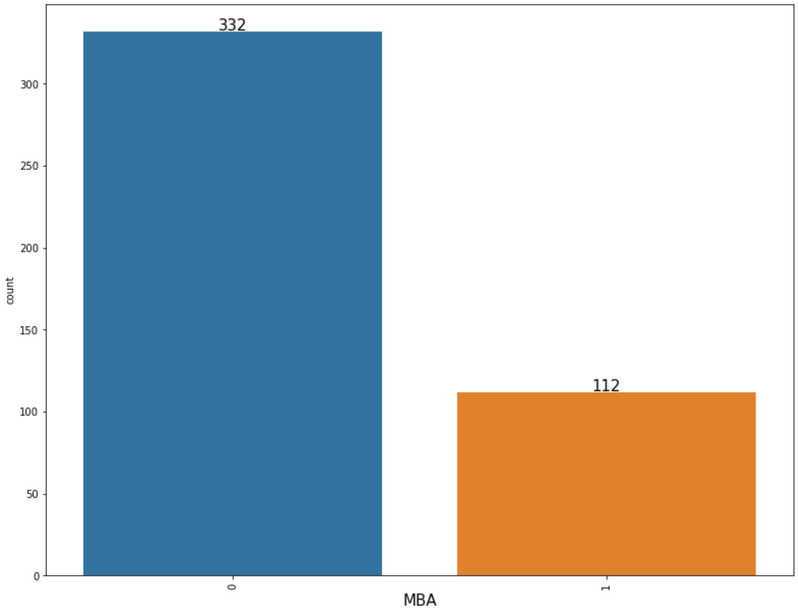
- The data is normally distributed for variable ‘Distance’. It also has outliers present in it.



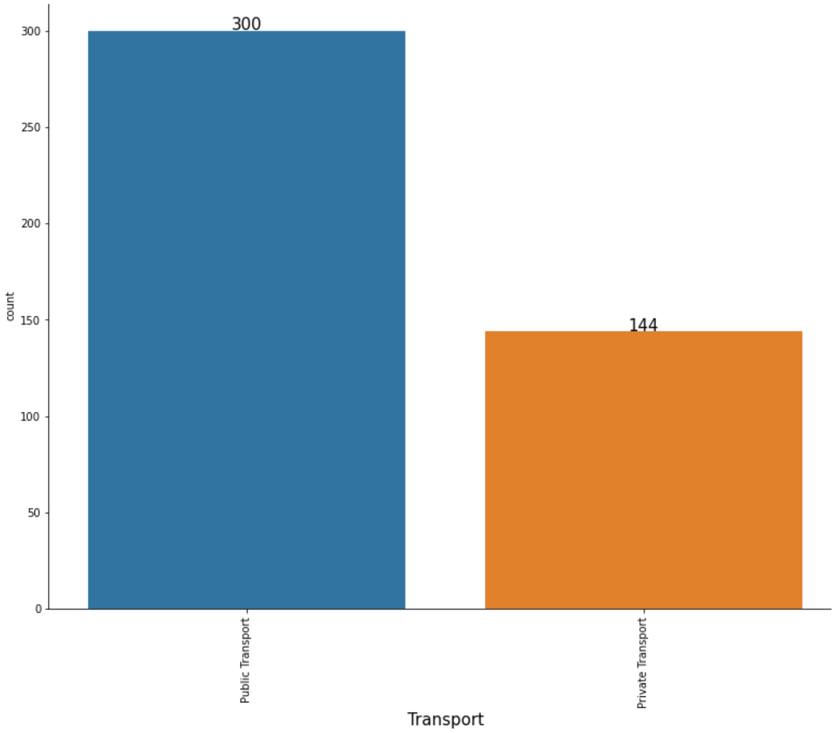
- The variable ‘Gender’ is categorical in nature. The number of Male is very high as compared to Female employees.



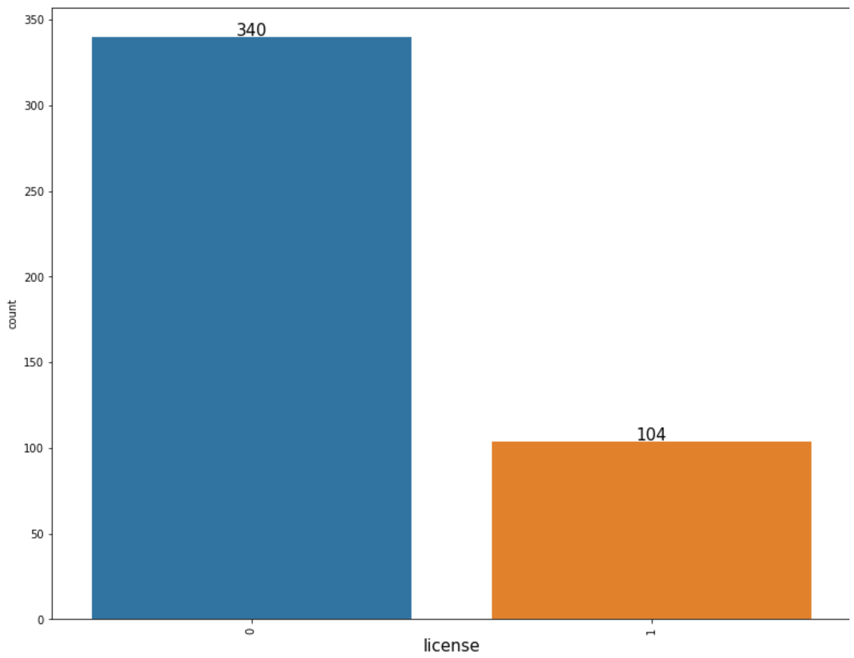
- The number of employees with Engineering degree is very high as compared to the non-engineers.



- The employees with MBA degree are less in number.

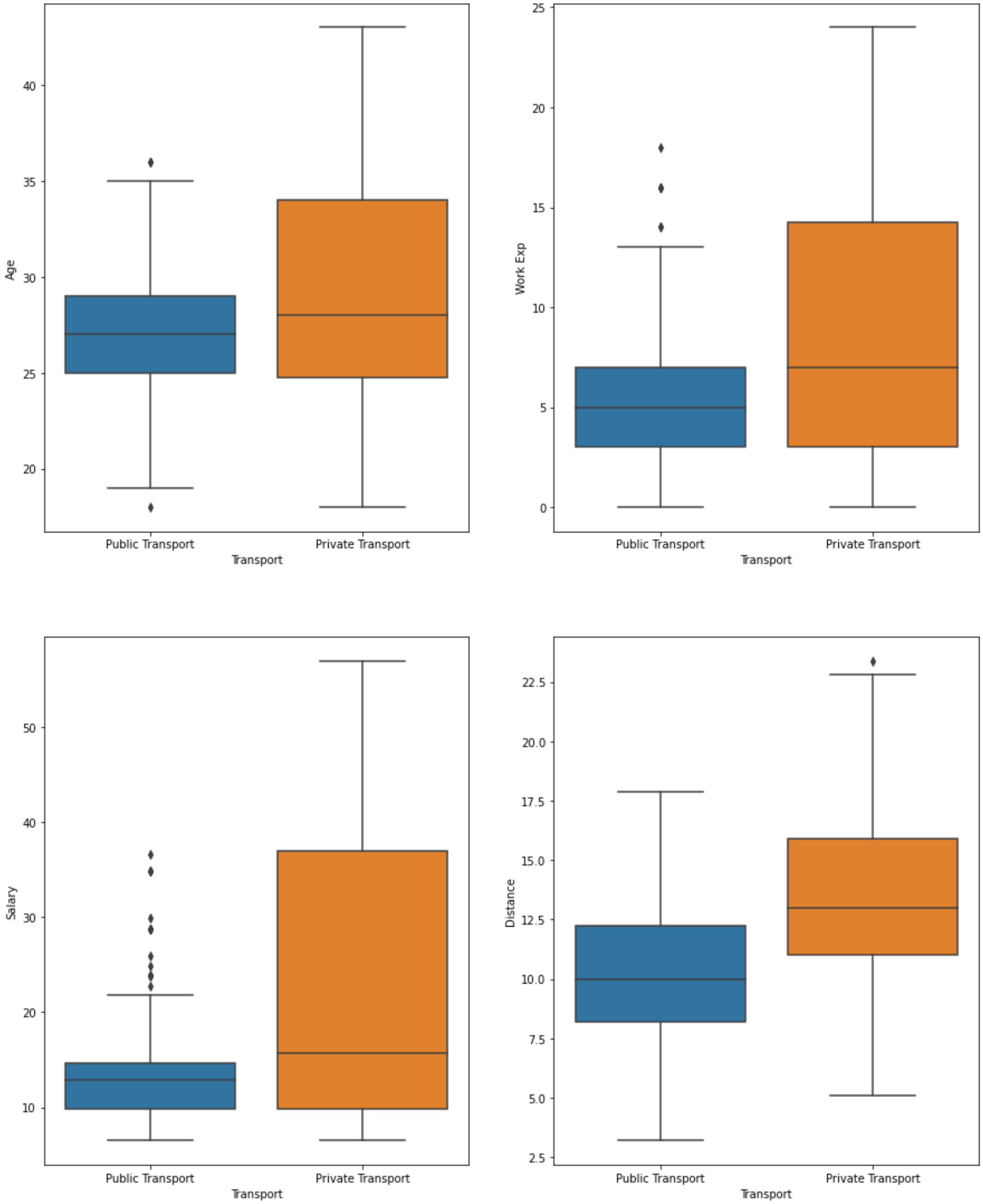


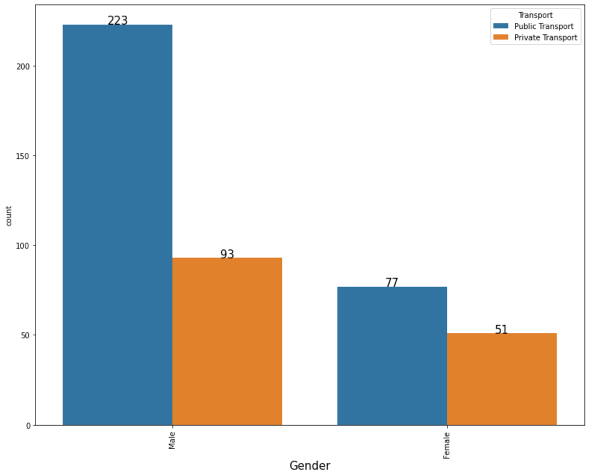
- Transport is the target variable. The number of employees using Public transport is 300 and employees using Private Transport is only 144.

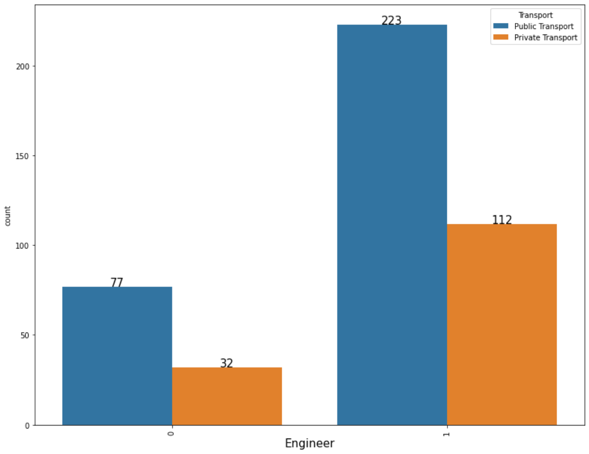


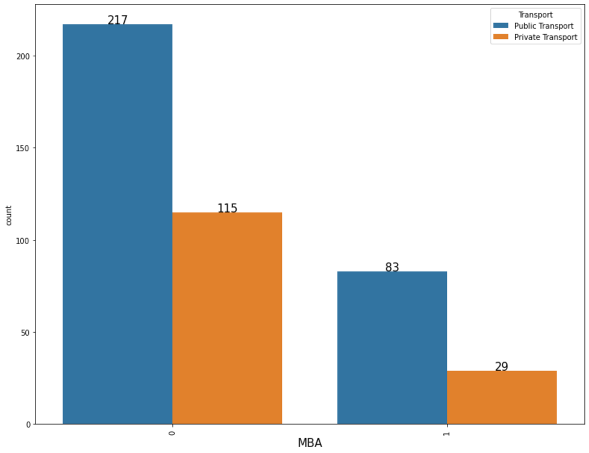
- Less number of employees in the company has license. The number of employees without license is 340.

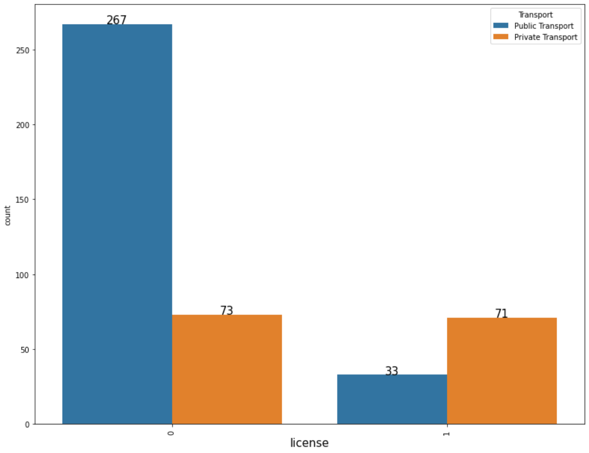
Bivariate Analysis

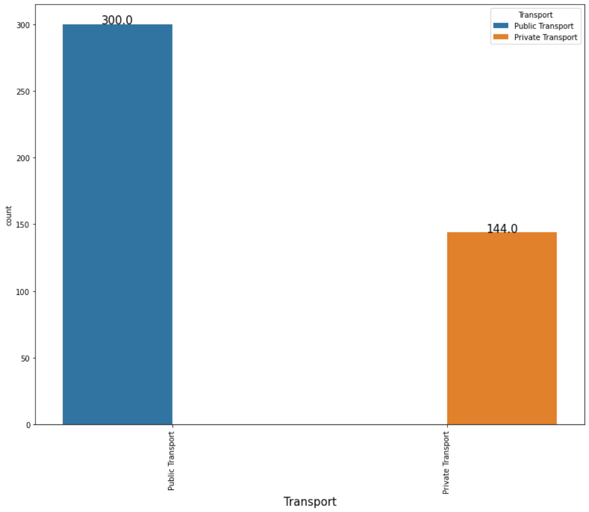








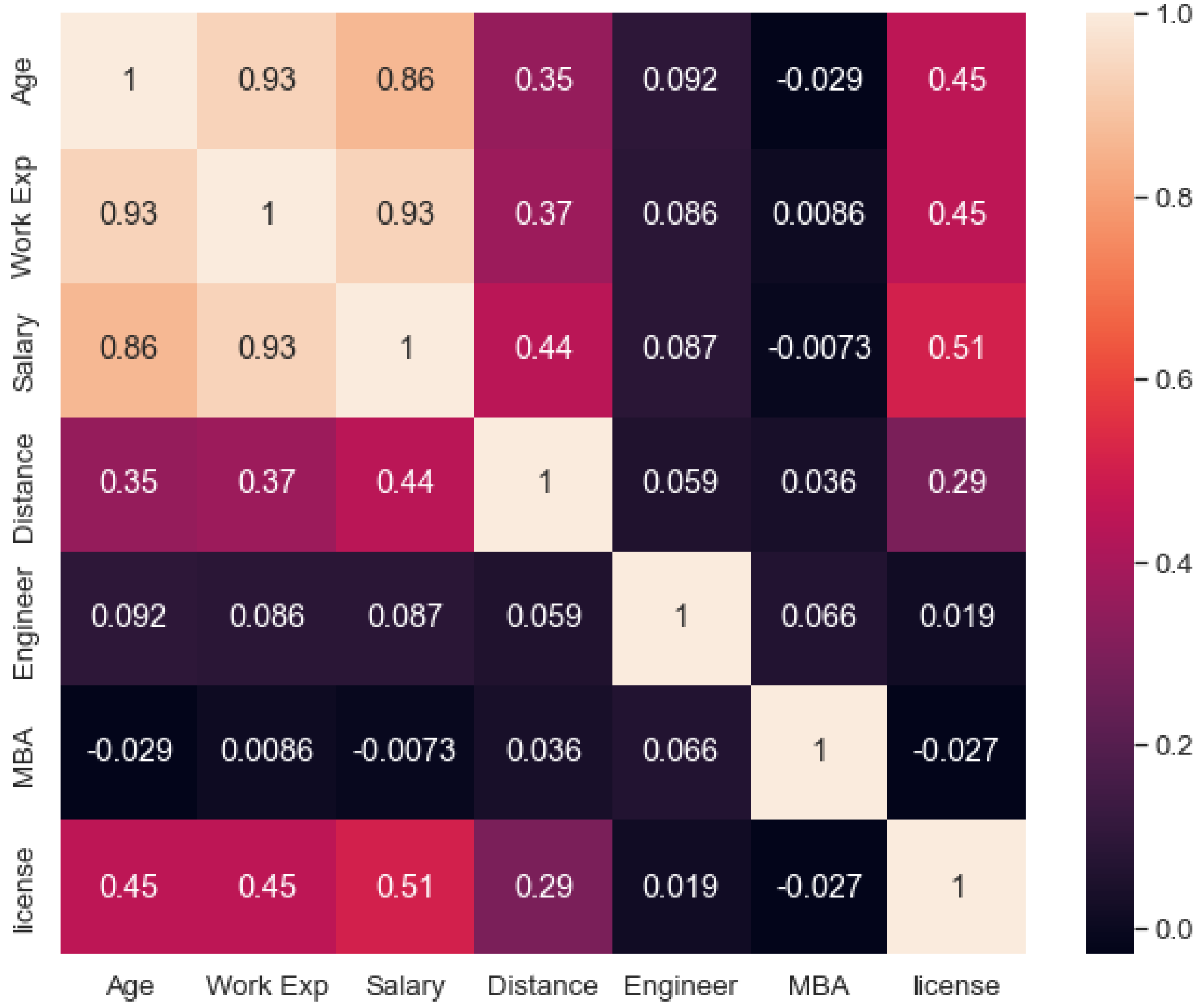




- We can understand employees having less ‘Age’ travel by public transport
- Employees with higher ‘Work Exp’ travel by private transport
- Employees with less salary travel by public transport and those with high salary travel by private transport

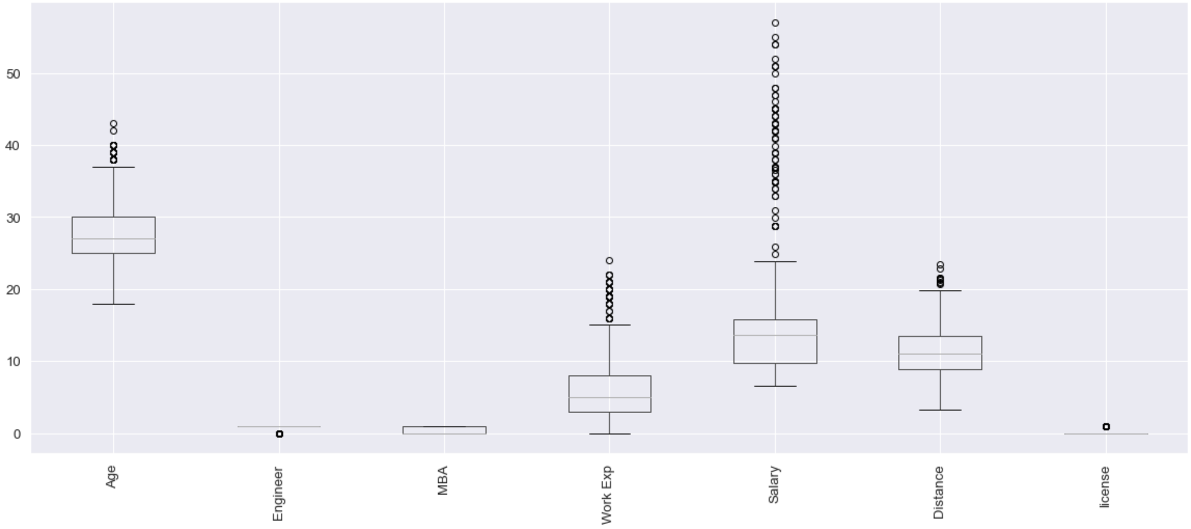
- Employees having less ‘Distance’ from office travel by public and employees with more ‘Distance’ travel by private transport
- Highest number of employees traveling by public transport are Males
- Employees having qualification as ‘Engineer’ travel mostly by public transport
- Employees having qualification as ‘MBA’ travel mostly by private transport
- Most of the people using public transport do not have license

Checking Correlations

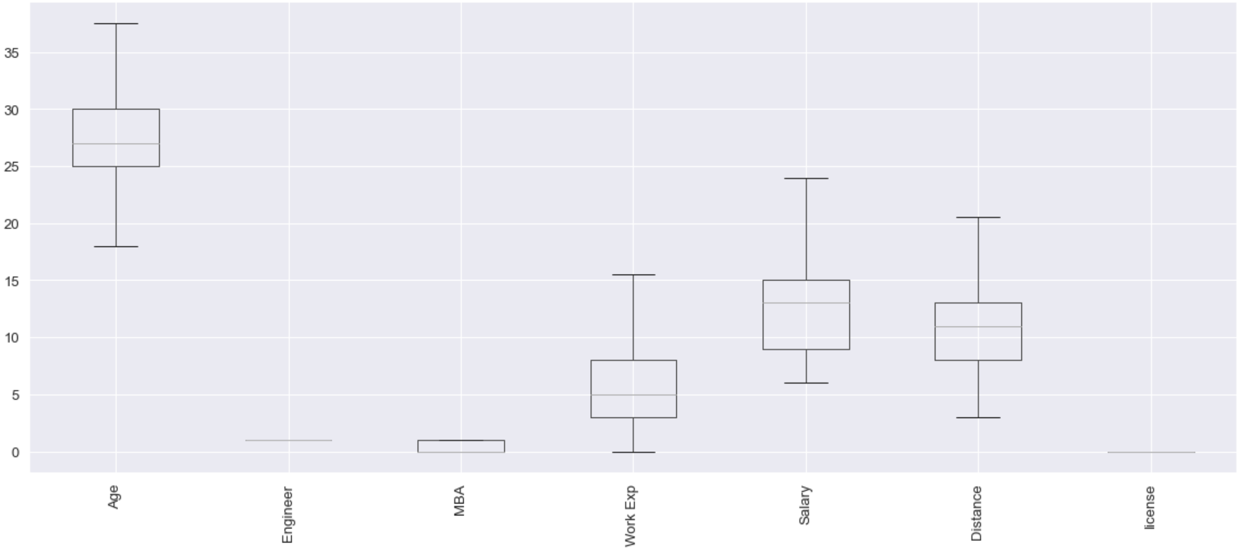


- High correlation of 0.93 between 'Work Exp' and 'Age'
- High correlation of 0.93 between 'Salary' and 'Work Exp'
- High correlation of 0.86 between 'Salary' and 'Age'

Outliers



- In the above graph we can see that all the variables have outliers.
- We shall be treating the outliers by imputing them with the standard technique of imputing with upper quantile and lower quantile limits.
- The upper value is calculated by Q3+(1.5 * IQR) & lower value is calculated by Q1-(1.5 * IQR).
- After imputation the data looks like the following image.



Missing Value Treatment

- There is no missing value in the data

Variable Treatment

- The target variable Transport had data type 'Object' which was converted into 'Float' to feed in the machine learning algorithms.
- One hot coding was done on feature 'Gender' which was labelled as Male/ Female.
- New columns were appended for the same. They are 'Gender_Female' and 'Gender_Male'.
- The datatype was made 'float' for new columns.

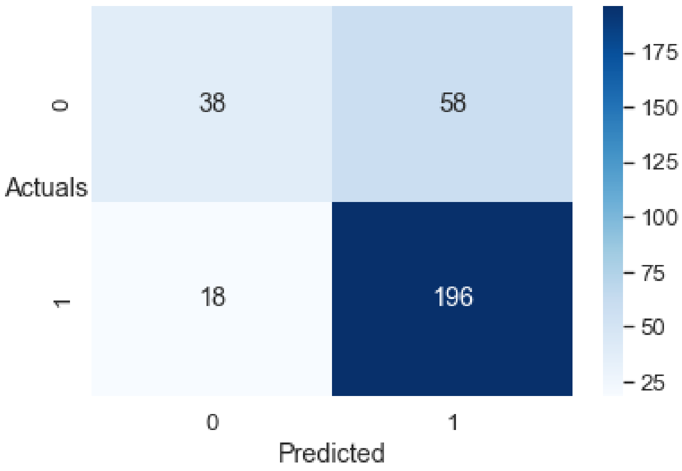
2. Split the data into train and test in the ratio 70:30. Is scaling necessary or not?
   The data was split into train and test in the ratio 70:30. Scaling is not necessary because features or variables like Age, Work Exp, Salary, and distance are in the range 0 to 100 & others variables are categorical in nature.

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.

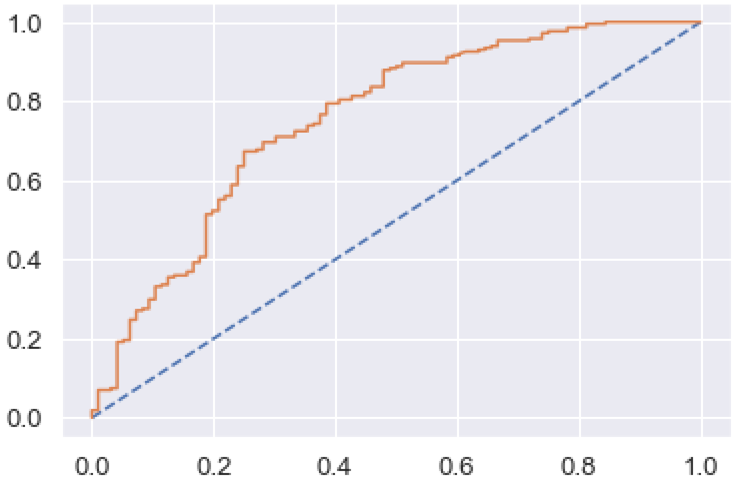
a. Logistic Regression Model

```
The model score for Logistic Regression Training set is 0.7548387096774194


The model score for Logistic Regression Testing set is 0.753731343283582
```

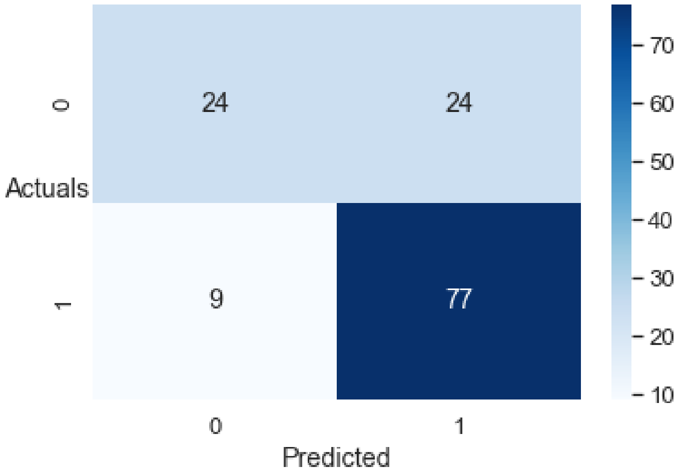
```
The classification report & Confution matrix for Logistic Regression training set is
              precision    recall  f1-score   support

         0.0       0.68      0.40      0.50        96
         1.0       0.77      0.92      0.84       214

    accuracy                           0.75       310
   macro avg       0.73      0.66      0.67       310
weighted avg       0.74      0.75      0.73       310
```



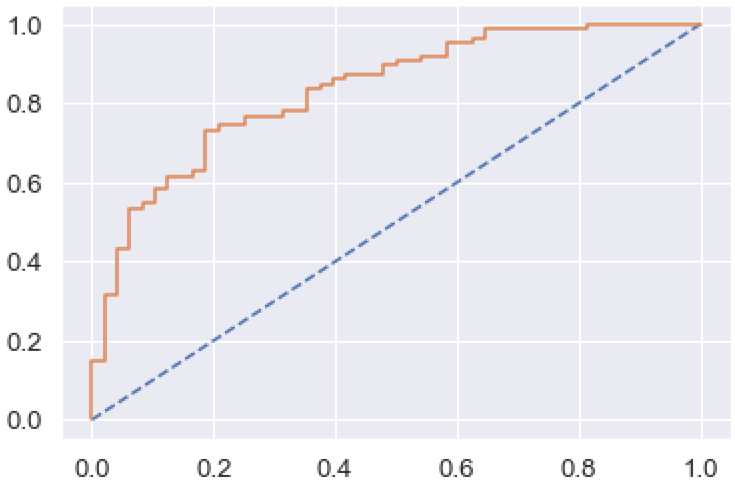
```
The AUC score for Logistic Regression Training dataset is: 0.7544
```



```
The Classification Report & Confusion Matrix for Logistic Regression testing set is
              precision    recall  f1-score   support

         0.0       0.73      0.50      0.59        48
         1.0       0.76      0.90      0.82        86

    accuracy                           0.75       134
   macro avg       0.74      0.70      0.71       134
weighted avg       0.75      0.75      0.74       134
```



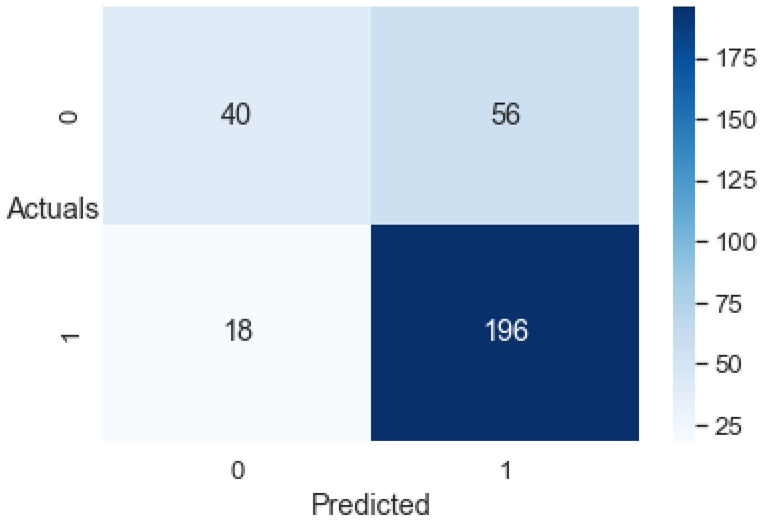
```
The AUC score for Logistic Regression testing set is: 0.8331
```



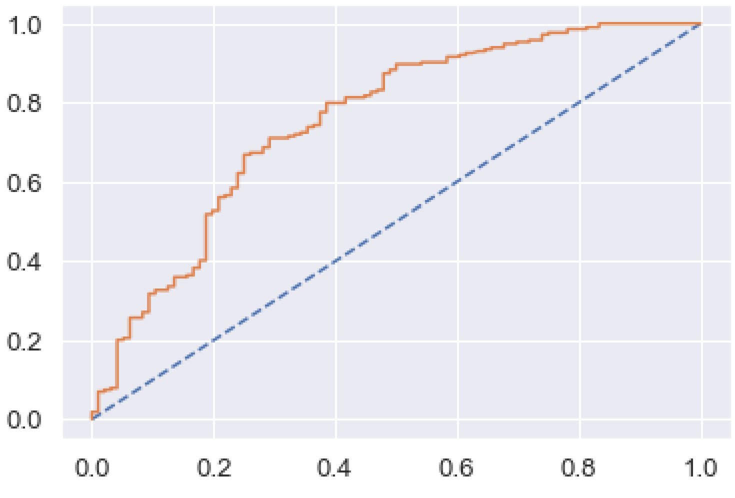
b. Linear Discriminant Analysis

```
The model score for Linear Discriminant Analysis training set is 0.7612903225806451
```

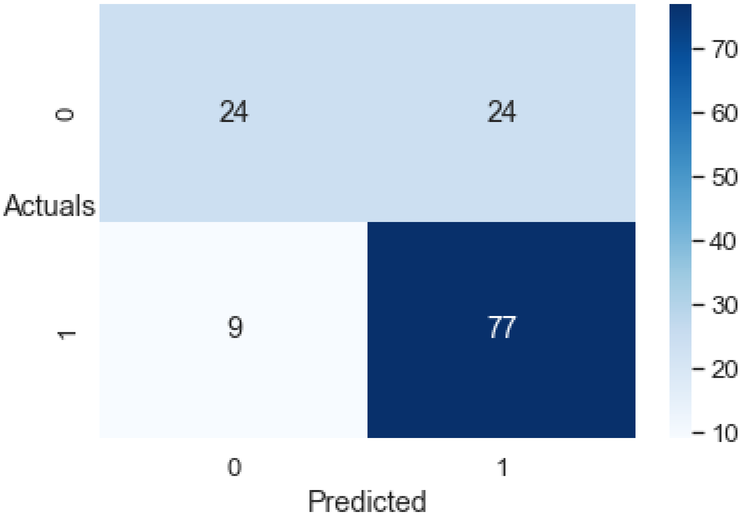
```
The model score for Linear Discriminant Analysis testing set is 0.753731343283582
The classification report for Linear Discriminant Analysis training set is
              precision    recall  f1-score   support

         0.0       0.69      0.42      0.52        96
         1.0       0.78      0.92      0.84       214

    accuracy                           0.76       310
   macro avg       0.73      0.67      0.68       310
weighted avg       0.75      0.76      0.74       310
```



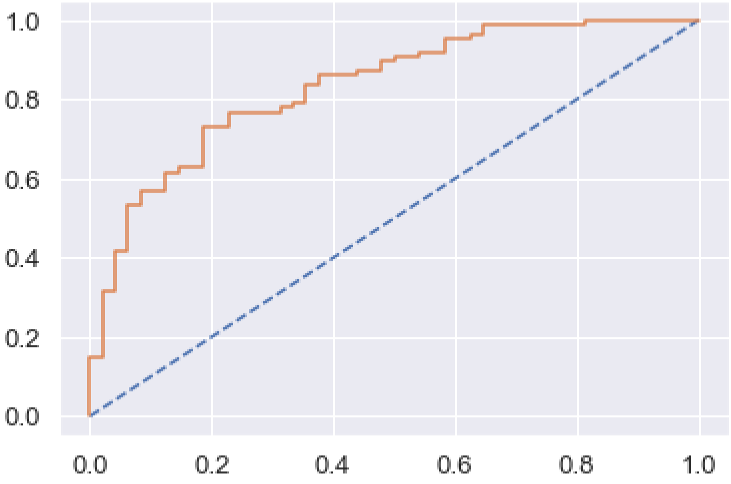
```
The AUC score for Linear Discriminant Analysis training set is: 0.754
```



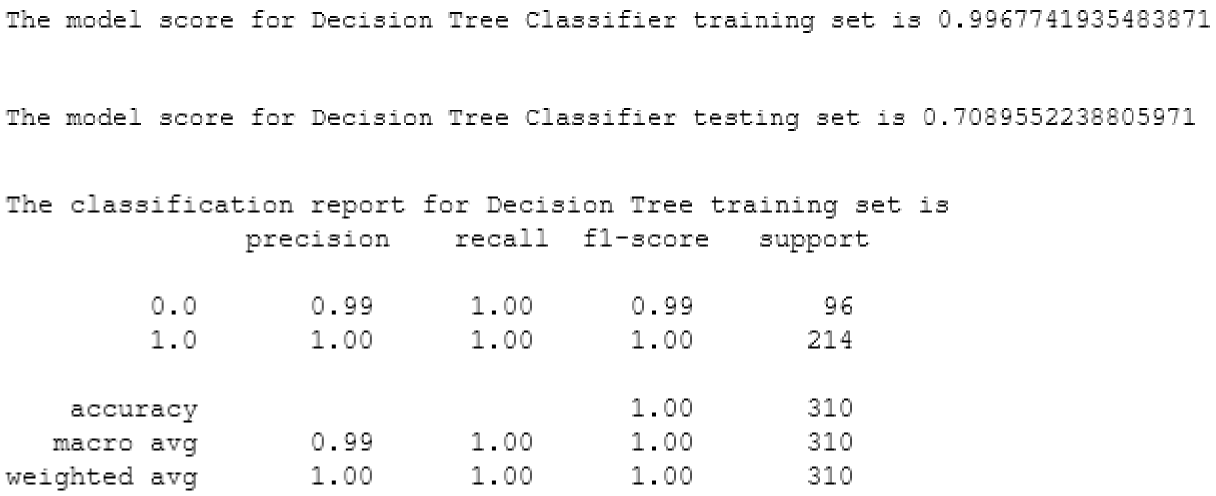
```
The classification report for Linear Discriminant Analysis testing set is
               precision    recall  f1-score   support

         0.0       0.73      0.50      0.59        48
         1.0       0.76      0.90      0.82        86

    accuracy                           0.75       134
   macro avg       0.74      0.70      0.71       134
weighted avg       0.75      0.75      0.74       134
```



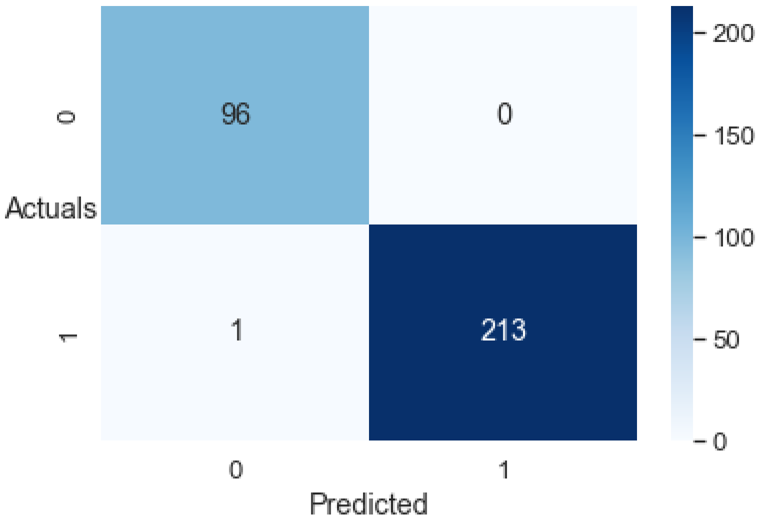
```
The AUC score for Linear Discriminant Analysis testing set is: 0.834
```



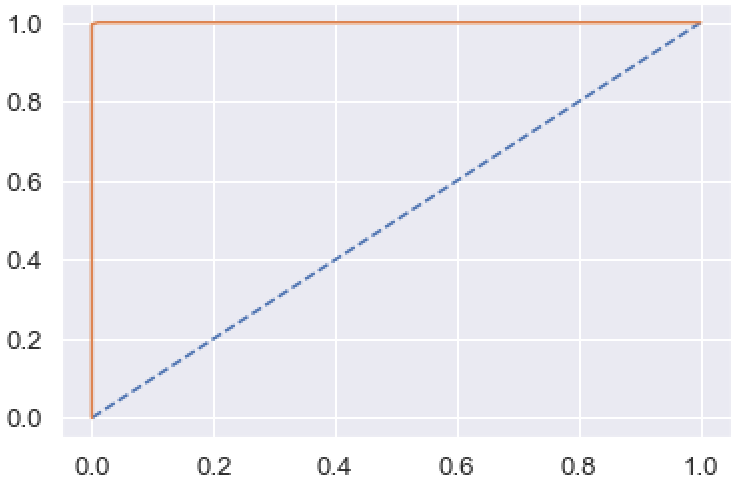
c. Decision Tree Classifier – CART model



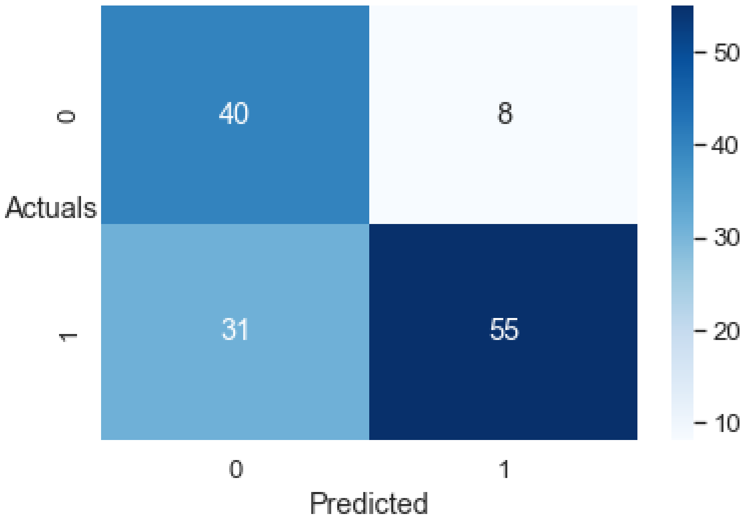
```
The model score for Decision Tree Classifier training set is 0.9967741935483871


The model score for Decision Tree Classifier testing set is 0.7089552238805971


The classification report for Decision Tree training set is
              precision    recall  f1-score   support

         0.0       0.99      1.00      0.99        96
         1.0       1.00      1.00      1.00       214

    accuracy                           1.00       310
   macro avg       0.99      1.00      1.00       310
weighted avg       1.00      1.00      1.00       310
```



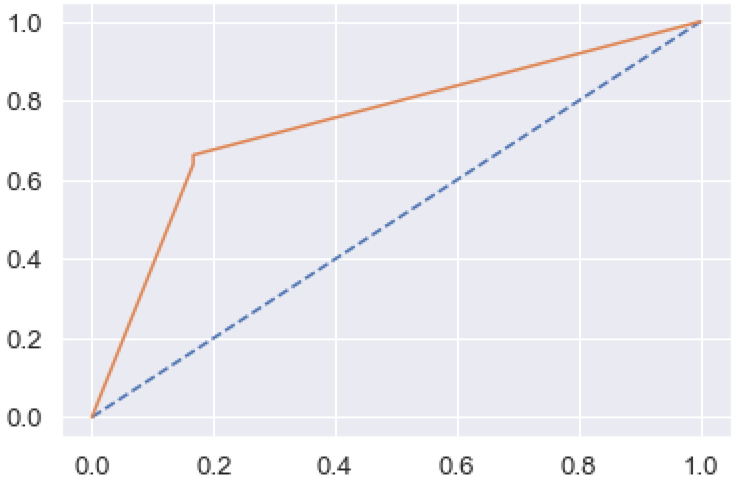
```
The AUC score for Decision Tree training set is: 1.000
```



```
The classification report for Decision Tree testing set is
              precision    recall  f1-score   support

         0.0       0.56      0.83      0.67        48
         1.0       0.87      0.64      0.74        86

    accuracy                           0.71       134
   macro avg       0.72      0.74      0.71       134
weighted avg       0.76      0.71      0.71       134
```

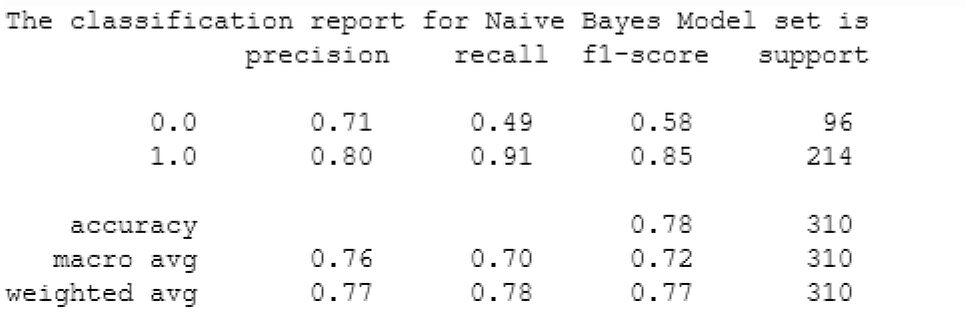


```
The AUC score for Decision Tree testing set is: 0.746
```

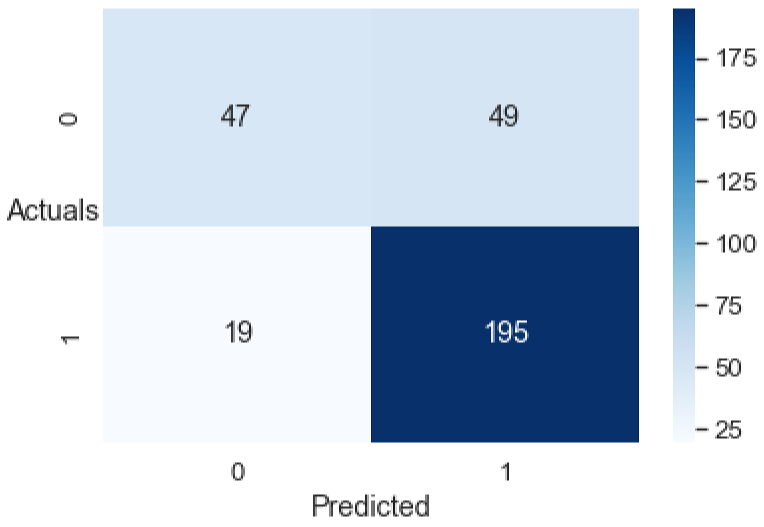


d. Naïve Bayes Model

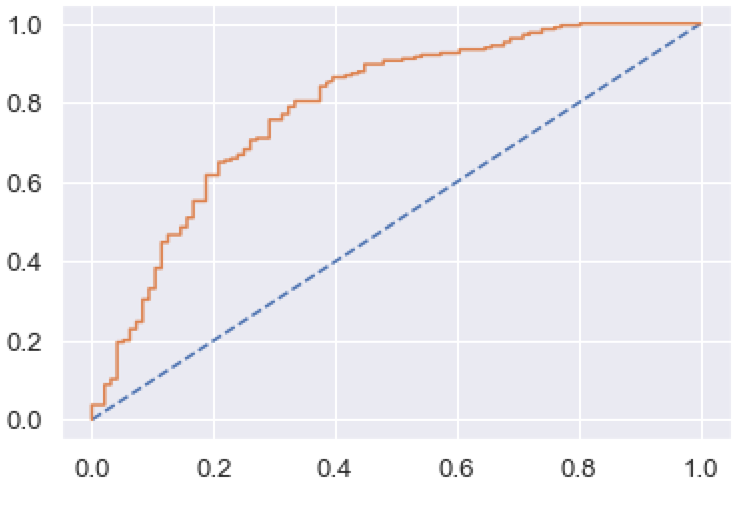
```
The model score for Naive Bayes Model training set is 0.7806451612903226


The model score for Naive Bayes Model testing set is 0.7835820895522388
```



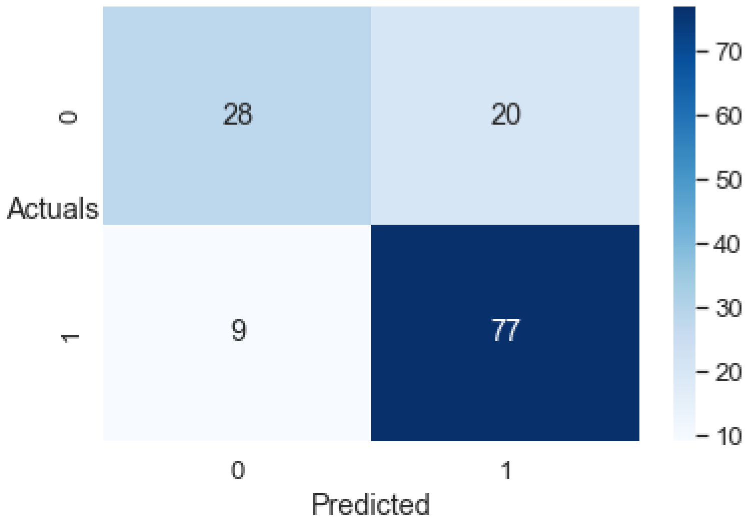
```
The classification report for Naive Bayes Model set is
              precision    recall  f1-score   support

         0.0       0.71      0.49      0.58        96
         1.0       0.80      0.91      0.85       214

    accuracy                           0.78       310
   macro avg       0.76      0.70      0.72       310
weighted avg       0.77      0.78      0.77       310
```



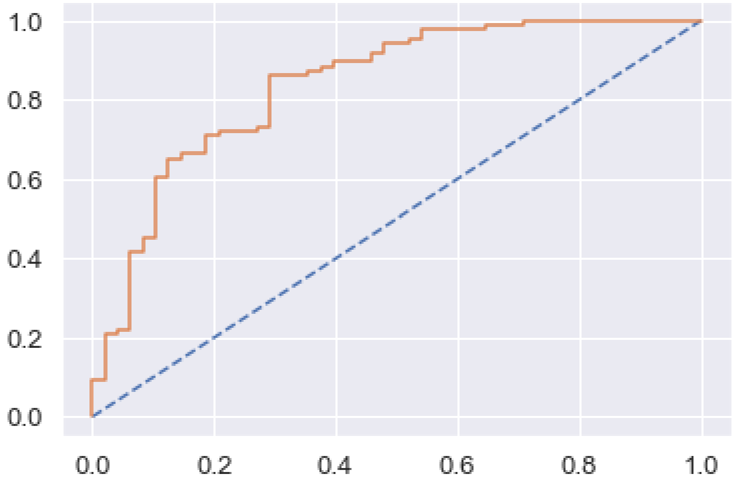
```
The AUC score for Naive Bayes training set is: 0.788
```



```
The classification report for Naive bayes Model testing set is
              precision    recall  f1-score   support

         0.0       0.76      0.58      0.66        48
         1.0       0.79      0.90      0.84        86

    accuracy                           0.78       134
   macro avg       0.78      0.74      0.75       134
weighted avg       0.78      0.78      0.78       134
```

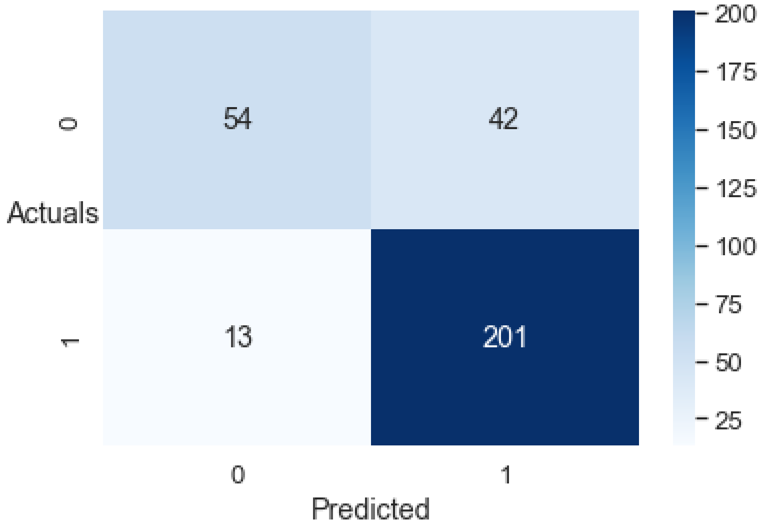


```
The AUC score for Naive Bayes testing set is: 0.838
```

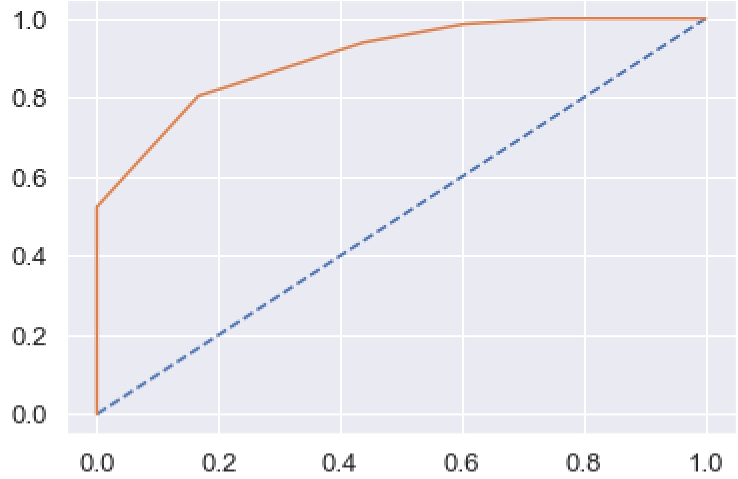


e. KNN Model

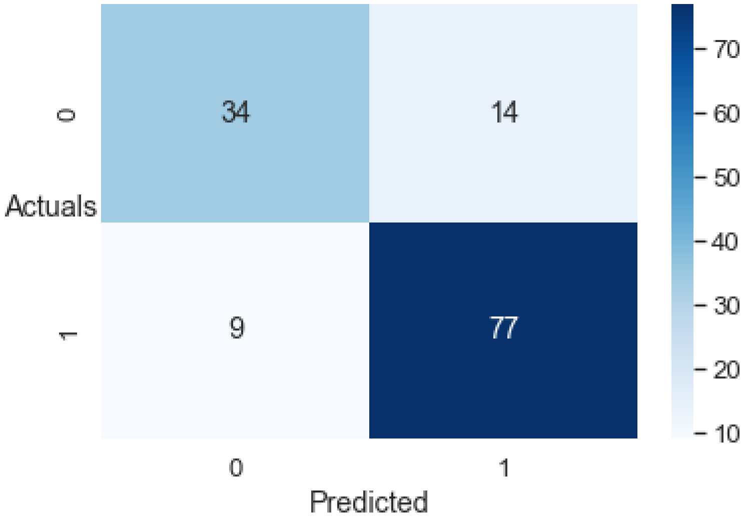
```
The model score for KNN training set is 0.8225806451612904


The model score for KNN testing set is 0.8283582089552238

The classification report for KNN set is
               precision    recall  f1-score   support

         0.0       0.81      0.56      0.66        96
         1.0       0.83      0.94      0.88       214

    accuracy                           0.82       310
   macro avg       0.82      0.75      0.77       310
weighted avg       0.82      0.82      0.81       310
```



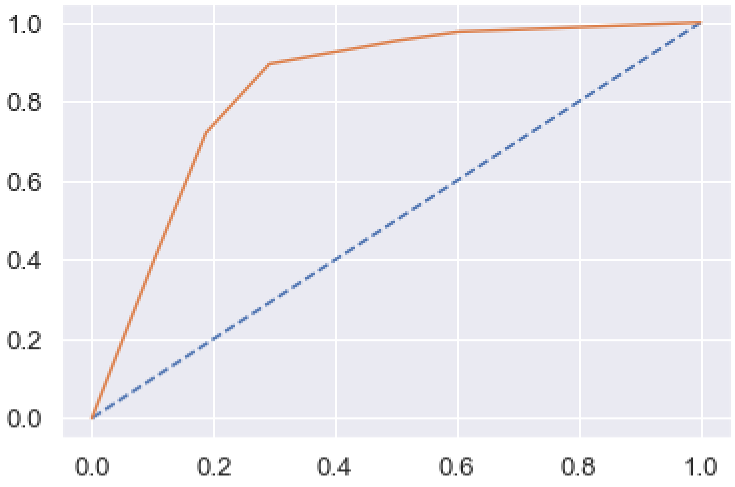
```
The AUC score for KNN training set is: 0.902
```



```
The classification report for KNN testing set is
              precision    recall  f1-score   support

         0.0       0.79      0.71      0.75        48
         1.0       0.85      0.90      0.87        86

    accuracy                           0.83       134
   macro avg       0.82      0.80      0.81       134
weighted avg       0.83      0.83      0.83       134
```



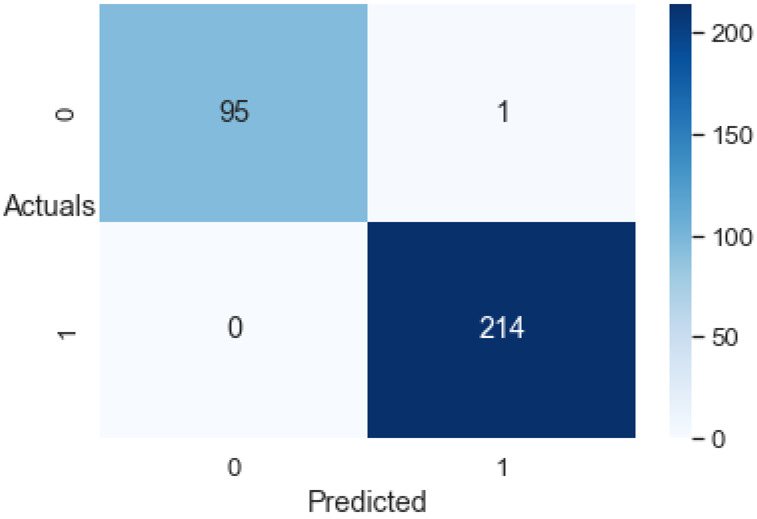
```
The AUC score for KNN testing set is: 0.837
```



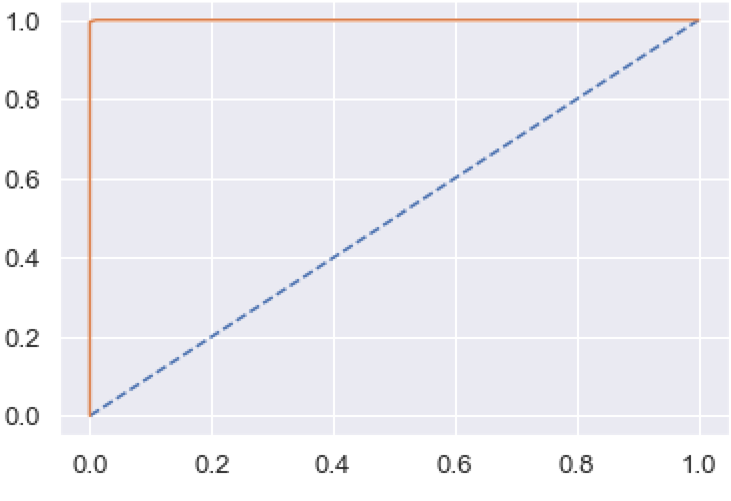
f. Random Forest Model

```
The model score for Random Forest Classifier training set is 0.9967741935483871
```

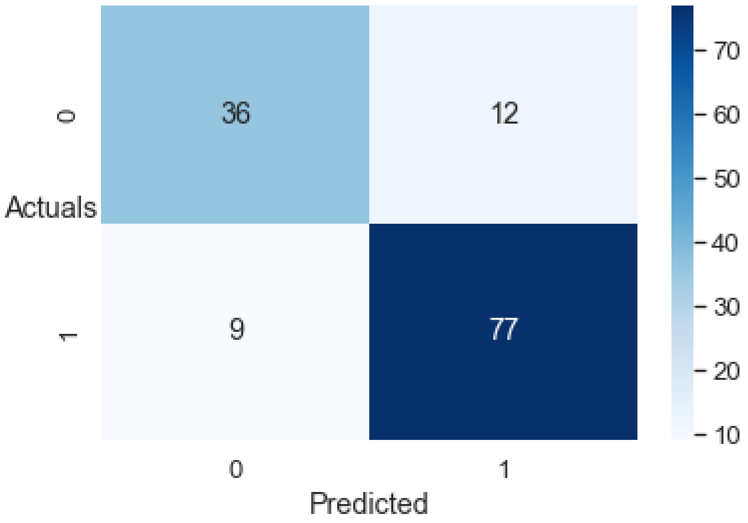
```
The model score for Random Forest Classifier testing set is 0.8432835820895522
The classification report for RFC training set is
              precision    recall  f1-score   support

         0.0       1.00      0.99      0.99        96
         1.0       1.00      1.00      1.00       214

    accuracy                           1.00       310
   macro avg       1.00      0.99      1.00       310
weighted avg       1.00      1.00      1.00       310
```



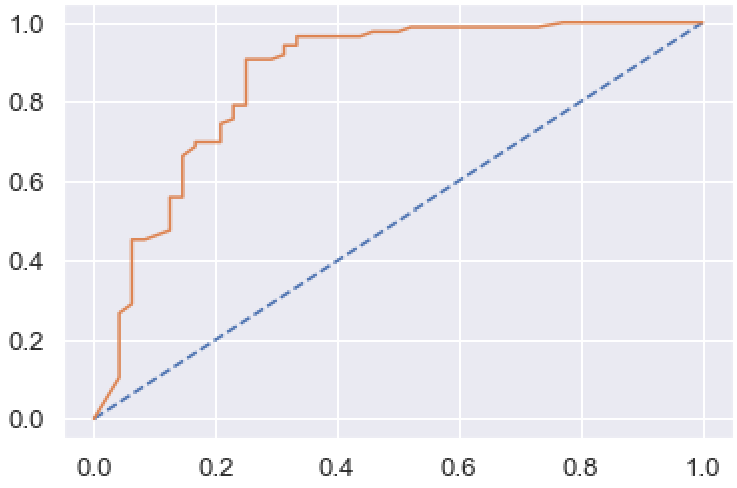
```
The AUC score for RFC training set is: 1.000
```



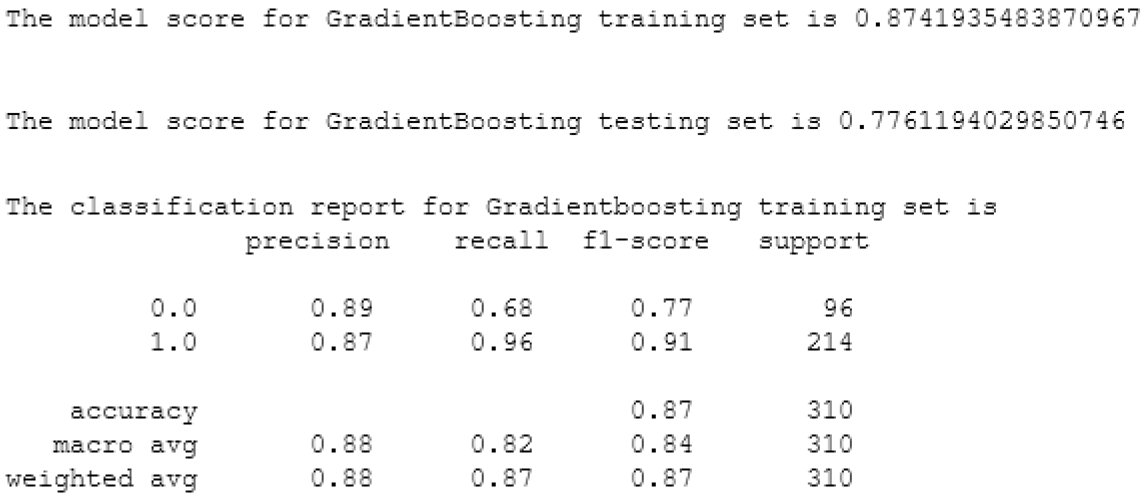
```
The classification report for RFC testing set is
              precision    recall  f1-score   support

         0.0       0.80      0.75      0.77        48
         1.0       0.87      0.90      0.88        86

    accuracy                           0.84       134
   macro avg       0.83      0.82      0.83       134
weighted avg       0.84      0.84      0.84       134
```



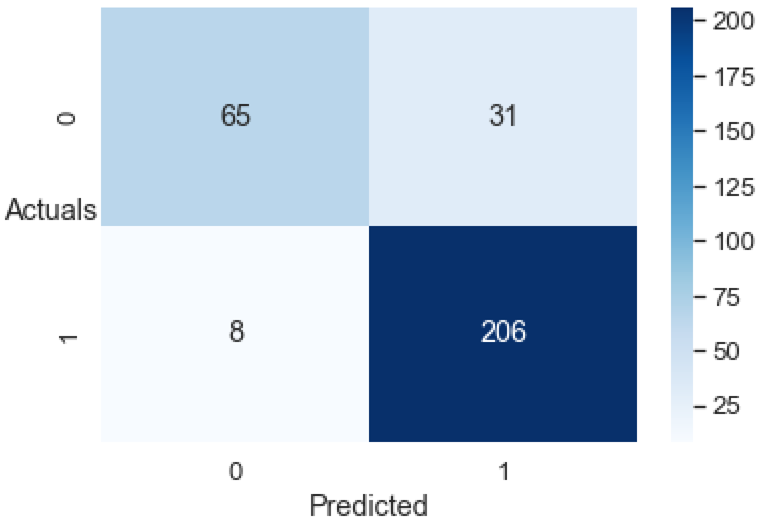
```
The AUC score for RFC testing set is: 0.859
```



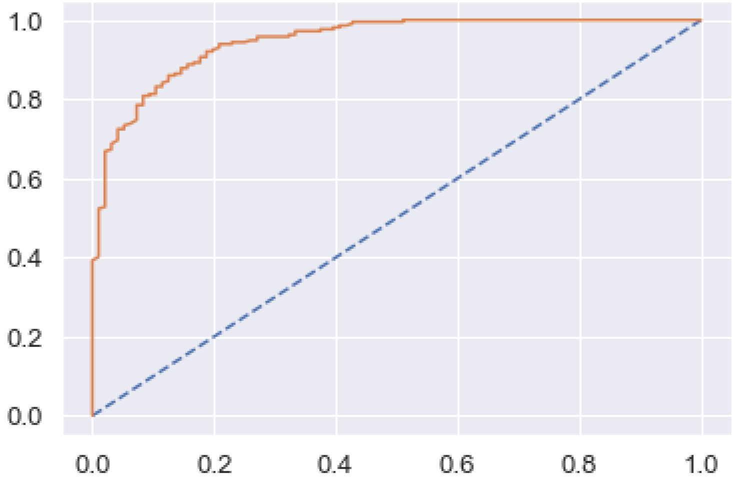
g. Boosting Classifier Model using Gradient boost.



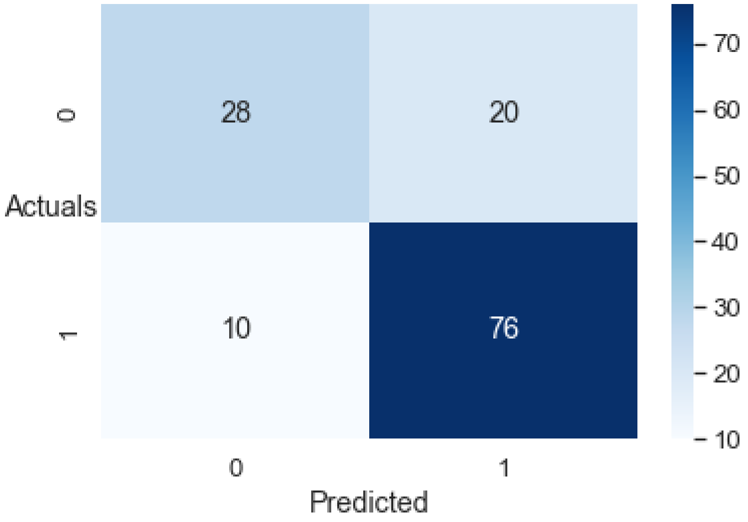
```
The model score for GradientBoosting training set is 0.8741935483870967


The model score for GradientBoosting testing set is 0.7761194029850746


The classification report for Gradientboosting training set is
               precision    recall  f1-score   support

         0.0       0.89      0.68      0.77        96
         1.0       0.87      0.96      0.91       214

    accuracy                           0.87       310
   macro avg       0.88      0.82      0.84       310
weighted avg       0.88      0.87      0.87       310
```



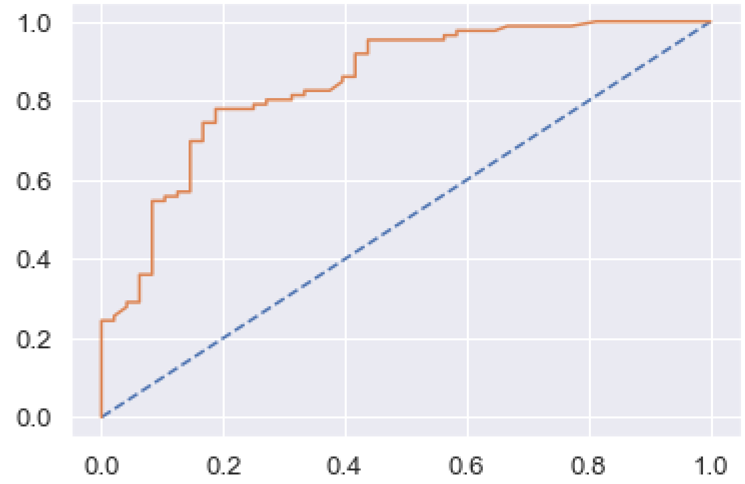
```
The AUC score for GradientBoosting training set is: 0.947
```



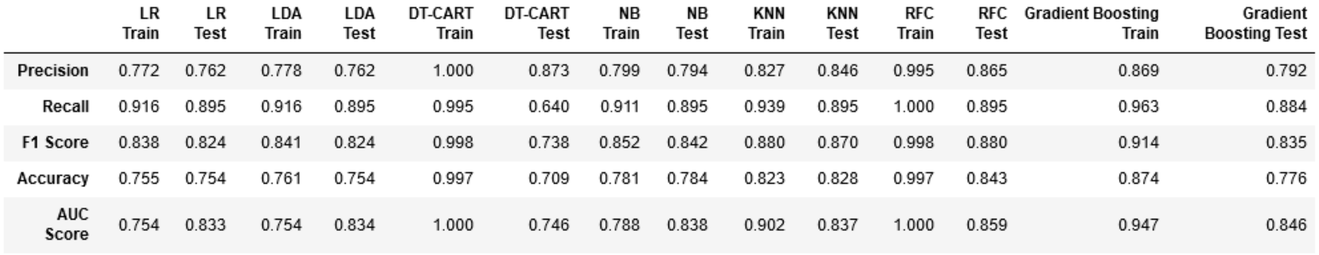
```
The classification report for Gradientboosting testing set is
               precision    recall  f1-score   support

         0.0       0.74      0.58      0.65        48
         1.0       0.79      0.88      0.84        86

    accuracy                           0.78       134
   macro avg       0.76      0.73      0.74       134
weighted avg       0.77      0.78      0.77       134
```



```
The AUC score for GradientBoosting testing set is: 0.846
```



4. Which model performs the best?



	LR Train	LR Test	LDA Train	LDA Test	DT-CART Train	DT-CART Test	NB Train	NB Test	KNN Train	KNN Test	RFC Train	RFC Test	Gradient Boosting Train	Gradient Boosting Test
Precision	0.772	0.762	0.778	0.762	1.000	0.873	0.799	0.794	0.827	0.846	0.995	0.865	0.869	0.792
Recall	0.916	0.895	0.916	0.895	0.995	0.640	0.911	0.895	0.939	0.895	1.000	0.895	0.963	0.884
F1 Score	0.838	0.824	0.841	0.824	0.998	0.738	0.852	0.842	0.880	0.870	0.998	0.880	0.914	0.835
Accuracy	0.755	0.754	0.761	0.754	0.997	0.709	0.781	0.784	0.823	0.828	0.997	0.843	0.874	0.776
AUC Score	0.754	0.833	0.754	0.834	1.000	0.746	0.788	0.838	0.902	0.837	1.000	0.859	0.947	0.846

- Logistic Regression model has performed poorly with training accuracy 75.5% and test accuracy of 75.4%.
- Linear Discriminant Analysis model has also performed poor with accuracy of 76.1% and 75.4% on train and test dataset respectively
- Decision Tree- CART has good accuracy score of 99.7% on train. But the accuracy on test dataset is only 70.9%. The number of False positives are 8 and False Negatives is 31.
- Naïve Bayes Model has performed poor with accuracy of 78.1% and 78.4% on train and test dataset respectively
- KNN Model has performed poor with accuracy of 82.3% and 82.8% on train and test dataset respectively
- **Random Forest Classifier has performed the best among all the models with accuracy on training set of 99.7% and on test set it is 84.3%. The number of False positives are 12 and False Negatives is 9.**
- After applying the model tuning technique Gradient boosting to the model the performance received was good. Gradient boosting gave accuracy of 94.7% on train and 84.6% on the test datasets.
- Random Forest Classifier has performed the best. Decision Tree- CART has performed fairly.

5. What are your business insights?

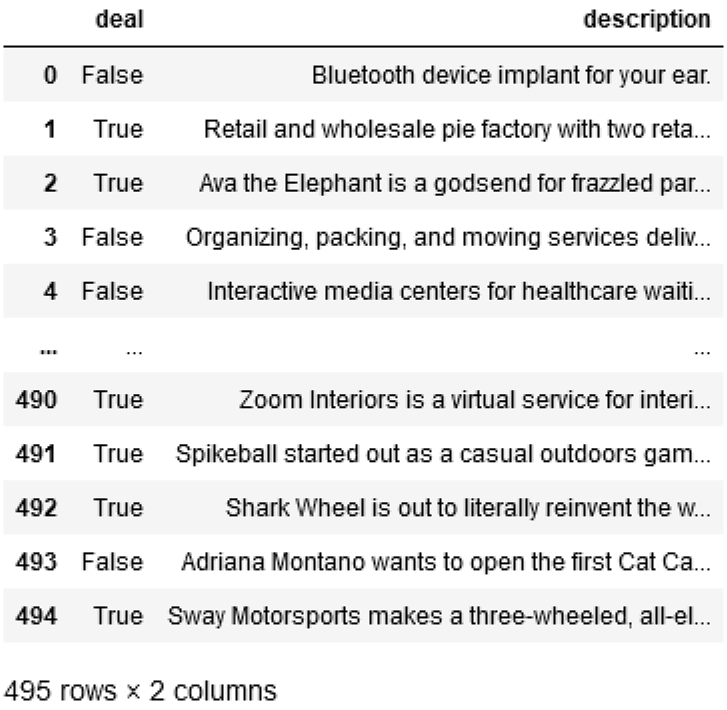
- We shall select and recommend Random Forest Classifier because it has best accuracy and also number of False Positives and False Negatives are much less compared to Decision Tree- CART.
- Presently most number of employees travels by public transport only. Therefore, if we provide better discounts and benefits then they shall definitely opt for our service.
- Marketing campaign should be for male Engineers as they travel mostly through public transport.
- Employees with high salary and more work experience should not be target customers as they prefer travelling by private transport.
- Most of the employees using public transport does not have license. Therefore, if marketing campaigns tells benefits of NOT driving a personal vehicle. Then, they might connect to such campaigns easily.

Part 2: Text Mining

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.

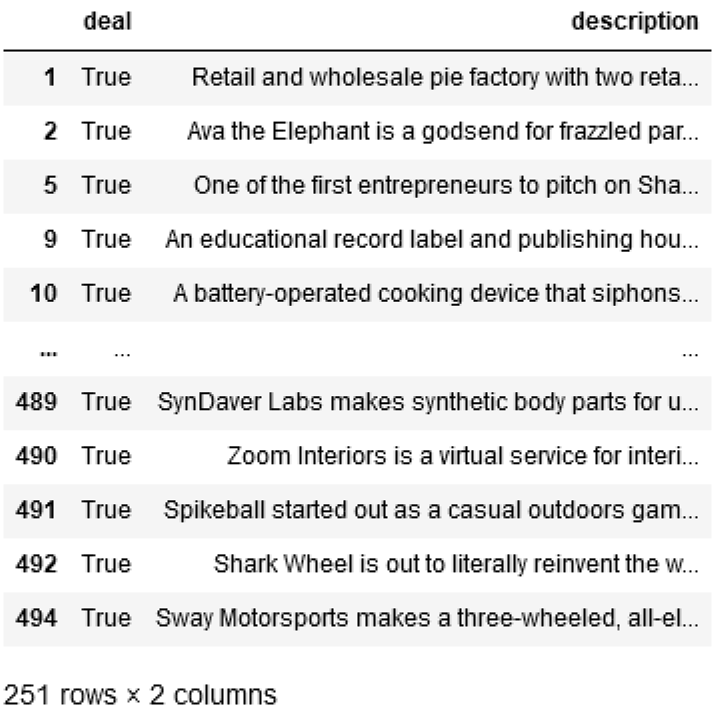
1. Pick out the Deal (Dependent Variable) and Description columns into a separate data frame.



	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 × 2 columns

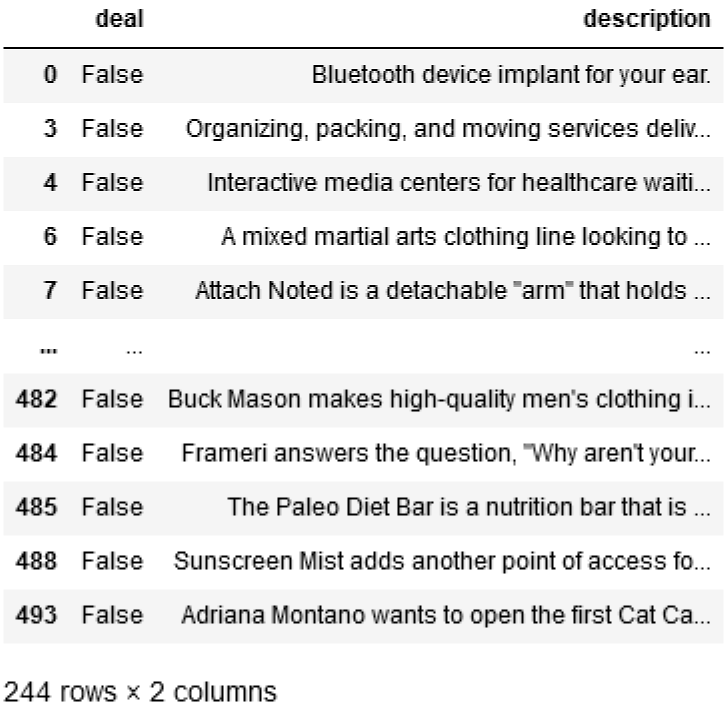
2. Create two corpora, one with those who secured a Deal, the other with those who did not secure a deal.



	deal	description
1	True	Retail and wholesale pie factory with two reta...
2	True	Ava the Elephant is a godsend for frazzled par...
5	True	One of the first entrepreneurs to pitch on Sha...
9	True	An educational record label and publishing hou...
10	True	A battery-operated cooking device that siphons...
...	...	...
489	True	SynDaver Labs makes synthetic body parts for u...
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...
494	True	Sway Motorsports makes a three-wheeled, all-el...

251 rows × 2 columns

The above entrepreneurs WON the deal.



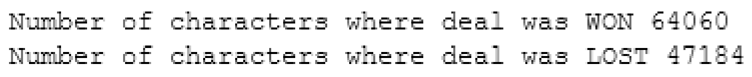
	deal	description
0	False	Bluetooth device implant for your ear.
3	False	Organizing, packing, and moving services deliv...
4	False	Interactive media centers for healthcare waiti...
6	False	A mixed martial arts clothing line looking to ...
7	False	Attach Noted is a detachable "arm" that holds ...
...	...	...
482	False	Buck Mason makes high-quality men's clothing i...
484	False	Frameri answers the question, "Why aren't your...
485	False	The Paleo Diet Bar is a nutrition bar that is ...
488	False	Sunscreen Mist adds another point of access fo...
493	False	Adriana Montano wants to open the first Cat Ca...

244 rows × 2 columns

The above entrepreneurs LOST the deal.

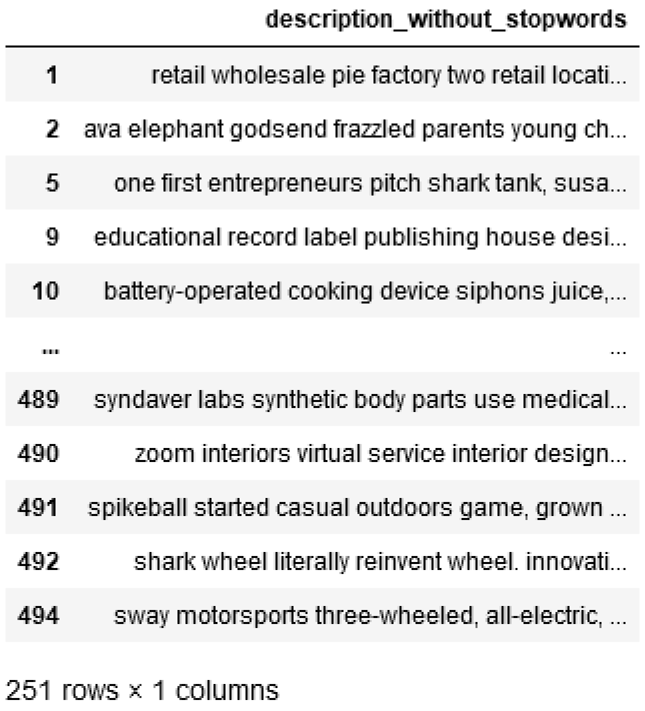
3. The following exercise is to be done for both the corpora:

a) Find the number of characters for both the corpuses.



```
Number of characters where deal was WON 64060
Number of characters where deal was LOST 47184
```

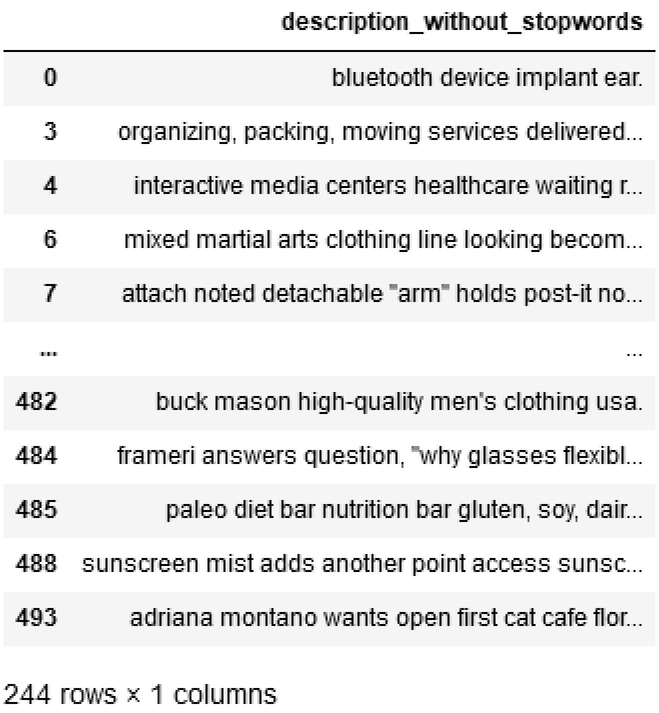
b) Remove Stop Words from the corpora. (Words like ‘also’, ‘made’, ‘makes’, ‘like’, ‘this’, ‘even’ and ‘company’ are to be removed)



	description_without_stopwords
1	retail wholesale pie factory two retail locati...
2	ava elephant godsend frazzled parents young ch...
5	one first entrepreneurs pitch shark tank, susa...
9	educational record label publishing house desi...
10	battery-operated cooking device siphons juice,...
...	...
489	syndaver labs synthetic body parts use medical...
490	zoom interiors virtual service interior design...
491	spikeball started casual outdoors game, grown ...
492	shark wheel literally reinvent wheel. innovati...
494	sway motorsports three-wheeled, all-electric, ...

251 rows × 1 columns

The above entrepreneurs WON the deal. The stop words were removed from the description.

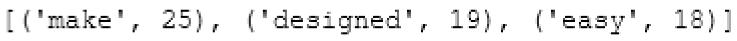


	description_without_stopwords
0	bluetooth device implant ear.
3	organizing, packing, moving services delivered...
4	interactive media centers healthcare waiting r...
6	mixed martial arts clothing line looking becom...
7	attach noted detachable "arm" holds post-it no...
...	...
482	buck mason high-quality men's clothing usa.
484	frameri answers question, "why glasses flexibl...
485	paleo diet bar nutrition bar gluten, soy, dair...
488	sunscreen mist adds another point access sunsc...
493	adriana montano wants open first cat cafe flor...

244 rows × 1 columns

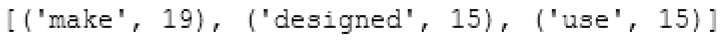
The above entrepreneurs LOST the deal. The stop words were removed from the description.

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



```
[('make', 25), ('designed', 19), ('easy', 18)]
```

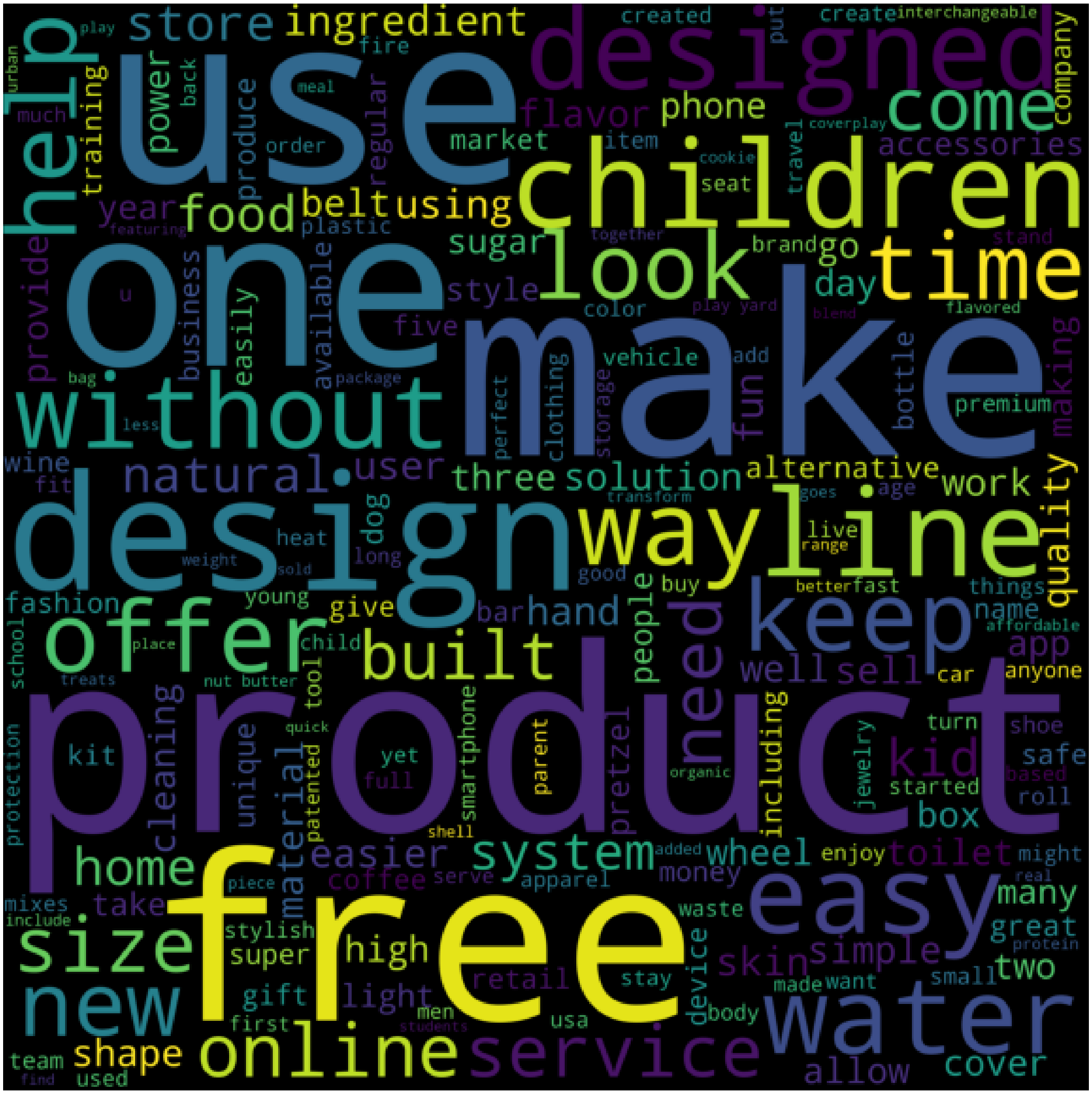
After removing the stop words, above are the 3 most frequently used words where entrepreneurs WON the deal.



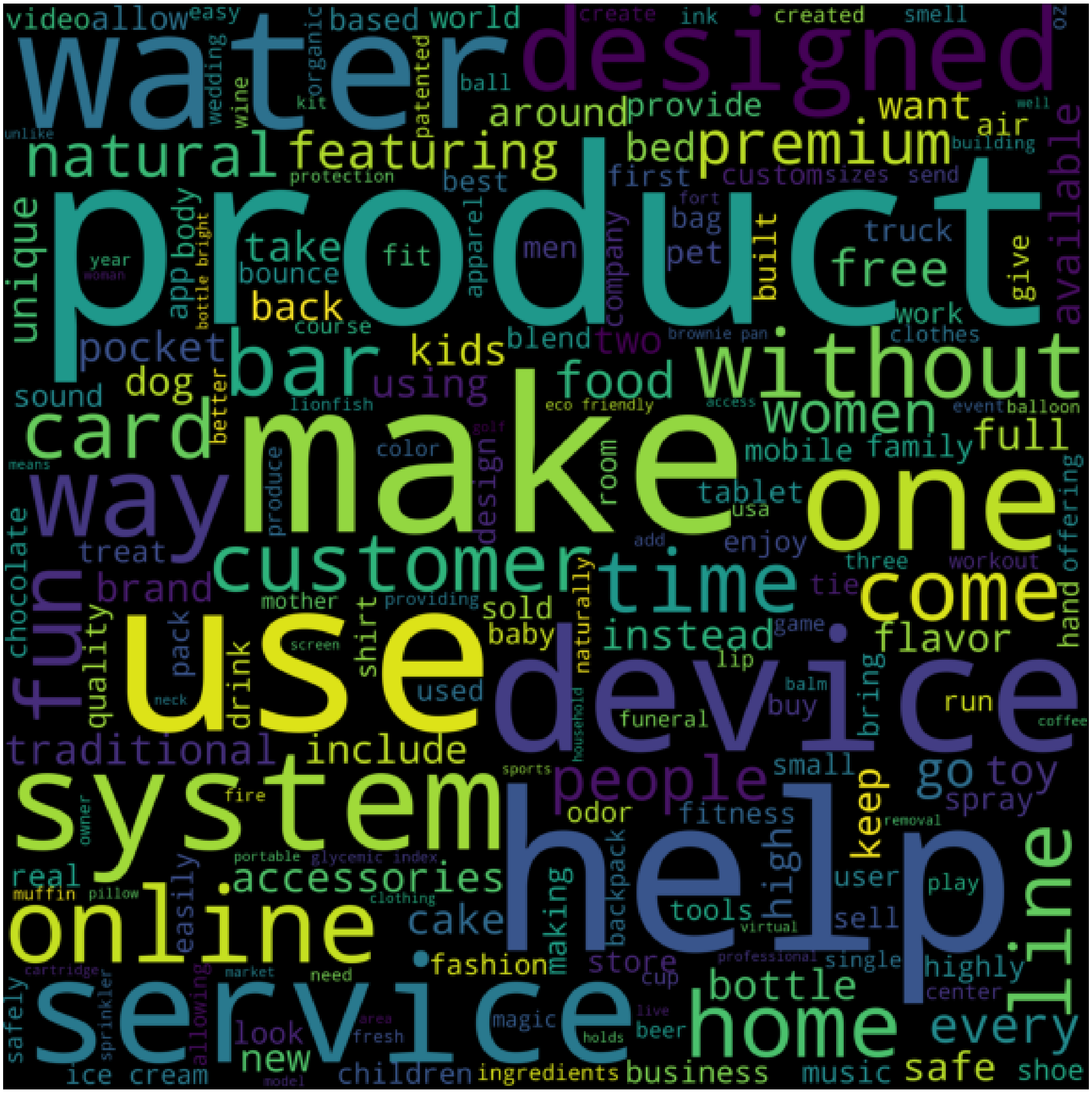
```
[('make', 19), ('designed', 15), ('use', 15)]
```

After removing the stop words, above are the 3 most frequently used words where entrepreneurs LOST the deal.

d) Plot the Word Cloud for both the corpora.



Above is the word cloud for the WON deals.



Above is the word cloud for the LOST deals.

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

Inference from Word Cloud of WON Deals

There are some prominent words visible from the word cloud of WON deals. These words are product, free, make, use, one, design, way, line, children, look, time, easy, help, water, online, size, new, offer etc. From above words, we can understand that most of the entrepreneurs who WON the deal have come with some product. This has more inclination towards solving the problem of society easily. This includes day to day problem faced by masses. They have designed to help various age groups like children. They have online strategy and have new offers.

Inference from Word Cloud of LOST Deals

The reason for failure of these entrepreneurs might be that most of them have tried resolve problems related to water. That might have sounded common and cliché to the sharks. Failed contestants seem to have more inclination towards the service industry which has high competition. There are some prominent words like device, use, system which indicates that they might have come up with ideas which are not suitable for this evolving tech savvy world.

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?

Yes, the word device is prominent in the world cloud of the failed contestants. However, the prominent words for the WON deals like design, free, offer etc. are missing in the LOST deals. From here we can easily understand that the products or services which has new designs, better offers and with freebies for customers are liked more by the shark investors.