**MACHINE LEARNING EXTENDED ASSIGNMENT**

**BY**

**SUDEEP KUMAR DAS**

**PGPDSBA.O. SEP22.B**

**TABLE OF CONTENT**

**LIST OF FIGURES \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 04**

**LIST OF TABLES \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 05**

**Problem 1:**

* 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. \_\_\_\_\_\_\_\_\_\_ 06**
  2. **Split the data into train and test in the ratio 70:30. Is scaling necessary or not? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 18**
  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.: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_19**

**a. Logistic Regression Model \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_19**

**b. Linear Discriminant Analysis \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20**

**c. Decision Tree Classifier – CART model \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 21**

**d. Naïve Bayes Model \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 21**

**e. KNN Model \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 21**

**f. Random Forest Model \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_22**

**g. Boosting Classifier Model using Gradient boost. \_\_\_\_\_\_\_\_\_22**

* 1. **Which model performs the best? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_22**
  2. **What are your business insights? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_23**

**Problem 2: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_25**

**2.1) Pick out the Deal (Dependent Variable) and Description columns into a separate data frame. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_25**

**2.2) Create two corpora, one with those who secured a Deal, the other with those who did not secure a deal. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_26**

**2.3) The following exercise is to be done for both the corpora: \_\_\_27**

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

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

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

**d) Plot the Word Cloud for both the corpora.**

**2.4 Refer to both the word clouds. What do you infer? \_\_\_\_\_\_30**

**2.5) 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? \_30**

**LIST OF FIGURES**

**Figure 1: Distribution of Age \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_­­\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_08**

**Figure 2: Distribution of Gender \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_09**

**Figure 3: Distribution of Engineer \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_09**

**Figure 4: Distribution of MBA \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_10**

**Figure 5: Distribution of Work Experience \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_10**

**Figure 6: Distribution of Salary \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_11**

**Figure 7: Distribution of Distance \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_11**

**Figure 8: Distribution of License \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_12**

**Figure 9: Distribution of Transport \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_12**

**Figure 10: Age vs Transport \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_13**

**Figure 11: Gender vs Transport \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_14**

**Figure 12: Engineers vs Transport \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_14**

**Figure 13: MBA vs Transport \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_15**

**Figure 14: Work Experience vs Transport \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_15**

**Figure 15: Salary vs Transport \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_16**

**Figure 16: Distance vs Transport \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_16**

**Figure 17: License vs Transport \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_17**

**Figure 18: Heatmap of the Dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_17**

**Figure 19: ROC Curve of Logistic Regression Model \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_19**

**Figure 20: ROC Curve of Linear Discriminant Analysis \_\_\_\_\_\_\_\_\_\_\_\_\_20**

**Figure 21: ROC Curve of Decision Tree Classifier Model \_\_\_\_\_\_\_\_\_\_\_\_20**

**Figure 22: ROC Curve of KNN Model \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_21**

**Figure 23: ROC Curve of Random Forest Classifier Model \_\_\_\_\_\_\_\_\_\_21**

**Figure 24: ROC curve of Gradient Boosting Classifier \_\_\_\_\_\_\_\_\_\_\_\_\_\_22**

**Figure 25: Ranking of the models \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_23**

**Figure 26: Word Cloud True (Secured a Deal) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_28**

**Figure 27: Word Cloud False (Did not Secured a Deal) \_\_\_\_\_\_\_\_\_\_\_\_\_29**

**LIST OF TABLES**

**Table 1: First and last few rows of the dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_06**

**Table 2: Info of the dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 07**

**Table 3: Description of the dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_07**

**Table 4: Null values \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 08**

**Table 5: First and last few rows of train data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 18**

**Table 6: First and last few rows of the test data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_18**

**Table 7: First few rows of the dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_25**

**Table 8: Info of the dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_25**

**Table 9: Description of the Dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_25**

**Table 10: First Few rows, Info & Description of True \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 26**

**Table 11: First Few Rows, info & Description of False \_\_\_\_\_\_\_\_\_\_\_\_\_\_27**

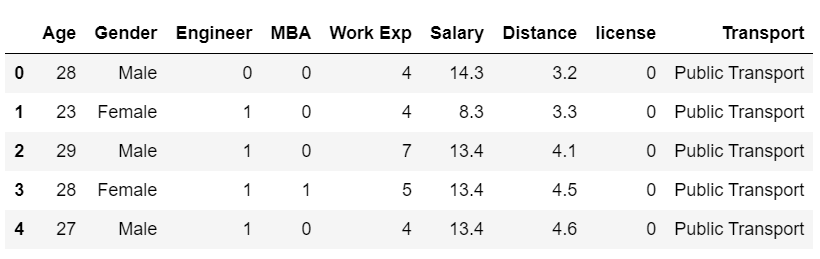
**Problem 1:**

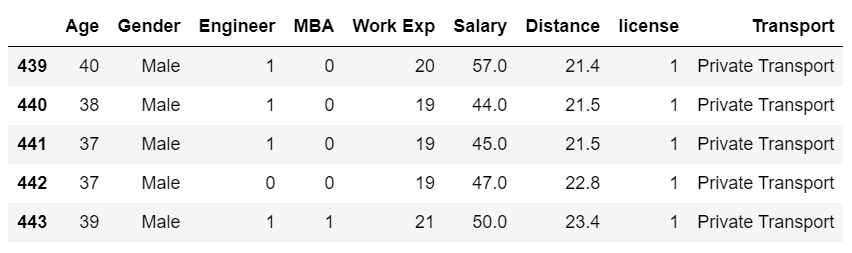
You work for an office transport company. You are in discussions with ABC Consulting company for providing transport for their employees. For this purpose, you are tasked with understanding how do the employees of ABC Consulting prefer to commute presently (between home and office). Based on the parameters 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.

**1.1) 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.**

Solution:

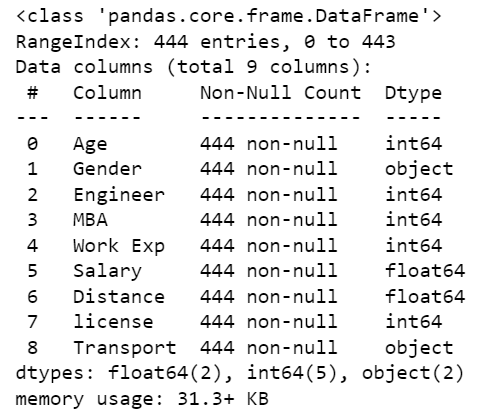
After importing the necessary libraries, we will read the dataset. The following are the first and last few rows of the dataset.





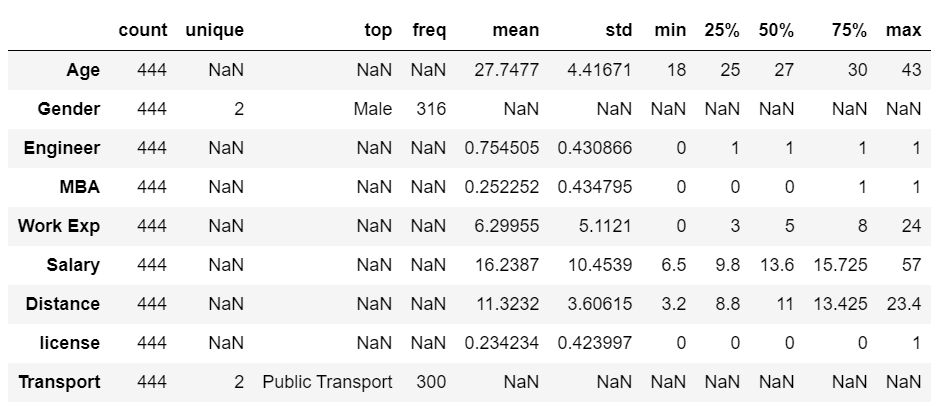
***Table 1: First and last few rows of the dataset***

The info of the dataset is as follows:



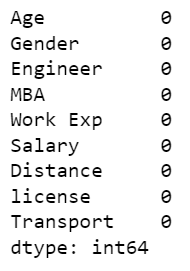
***Table 2: Info of the dataset***

The description of the dataset is as follows:



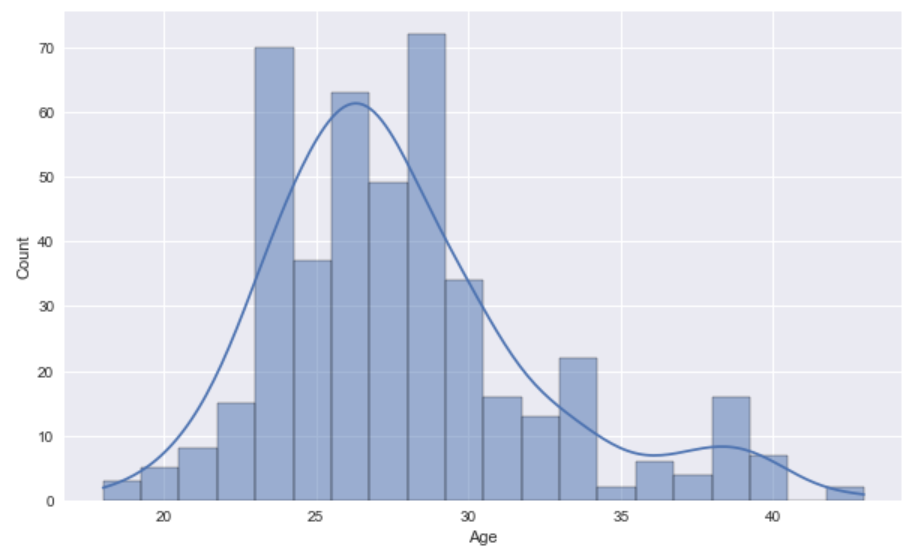
***Table 3: Description of the dataset***

Null value check of the dataset reveals that there are no null values in the dataset.



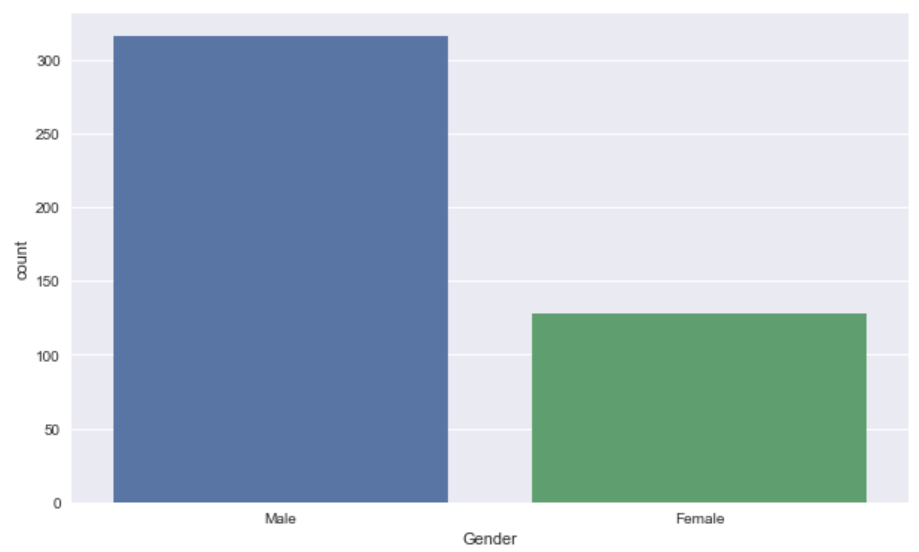
***Table 4: Null values***

Univariate Analysis:



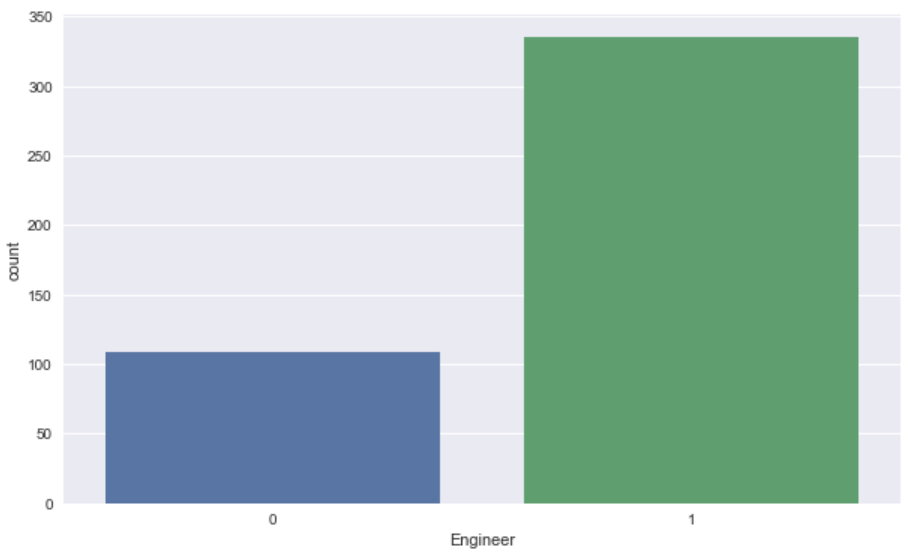
***Figure 1: Distribution of Age***

The age variable is approximately normally distributed. The mean age is 27.7 years with minimum age at 18 and maximum age at 43.



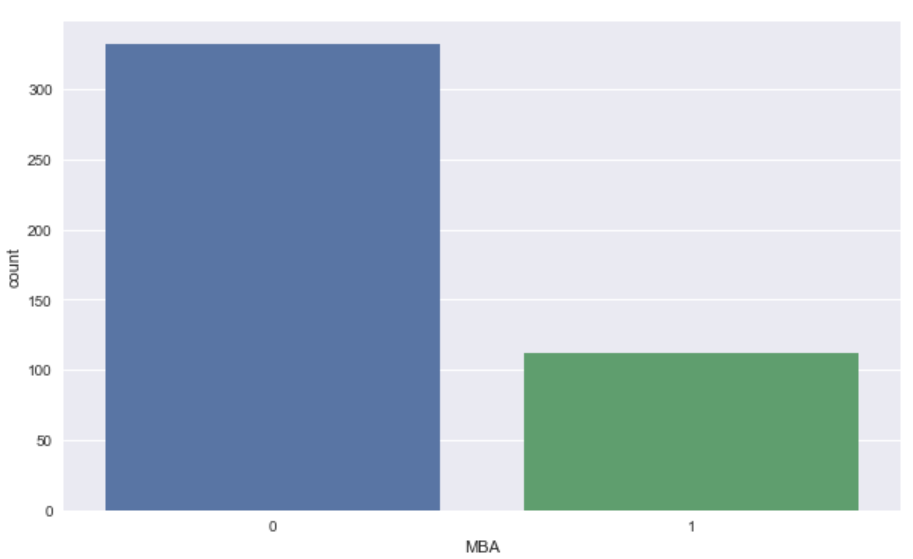
***Figure 2: Distribution of Gender***

71% of the dataset is male and only 29% are female.



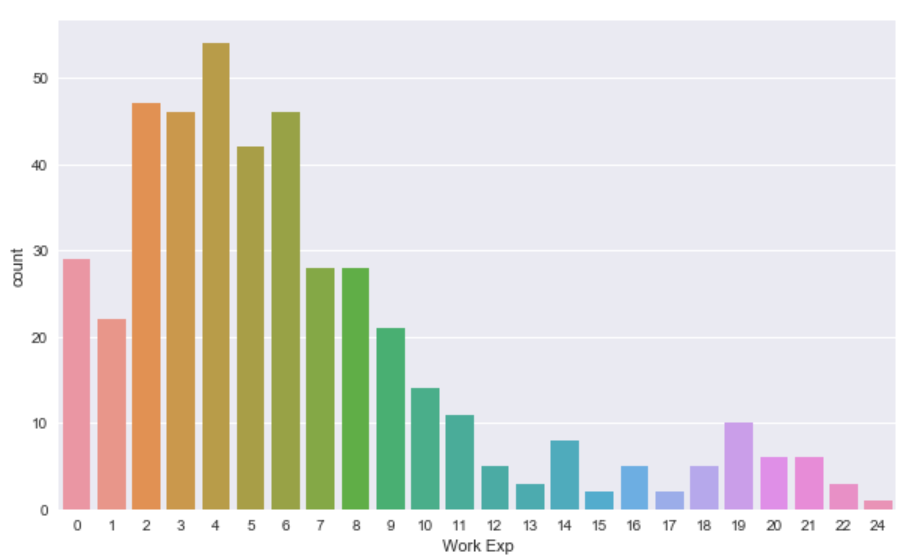
***Figure 3: Distribution of Engineer***

335 are engineer i.e. 75.4% are engineer and the rest are non engineer.



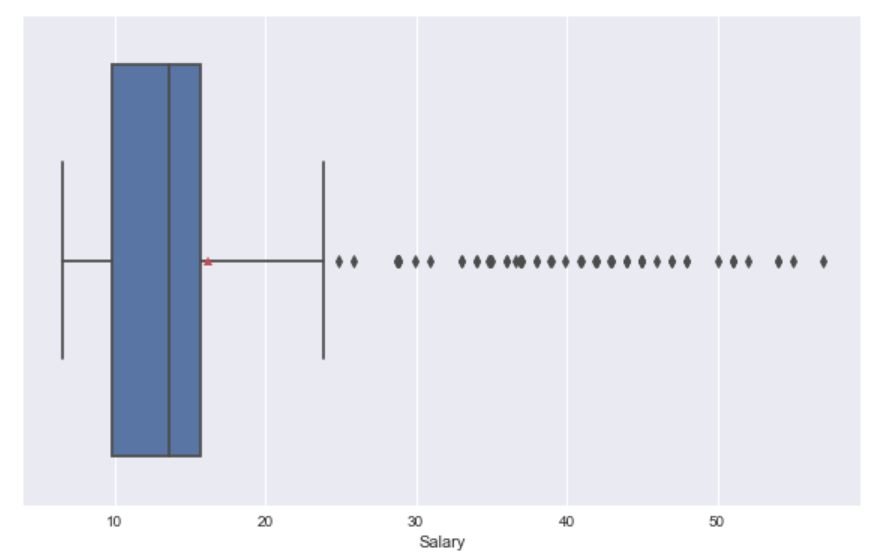
***Figure 4: Distribution of MBA***

Only 25% i.e. 112 people of the dataset are MBA qualified and the rest 75% are non MBA.



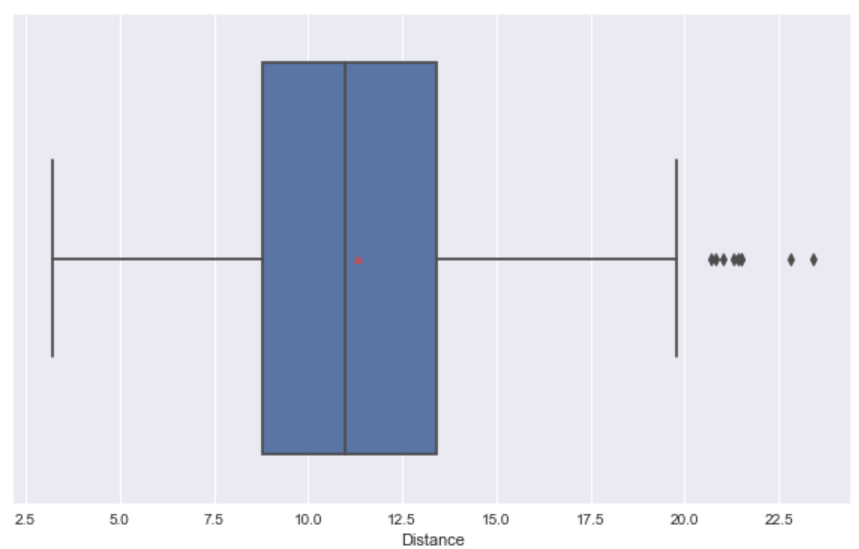
***Figure 5: Distribution of Work Experience***

The average work experience is approx. 6 years with minimum being 0 years and maximum work experience being 24 years. The data is left skewed.



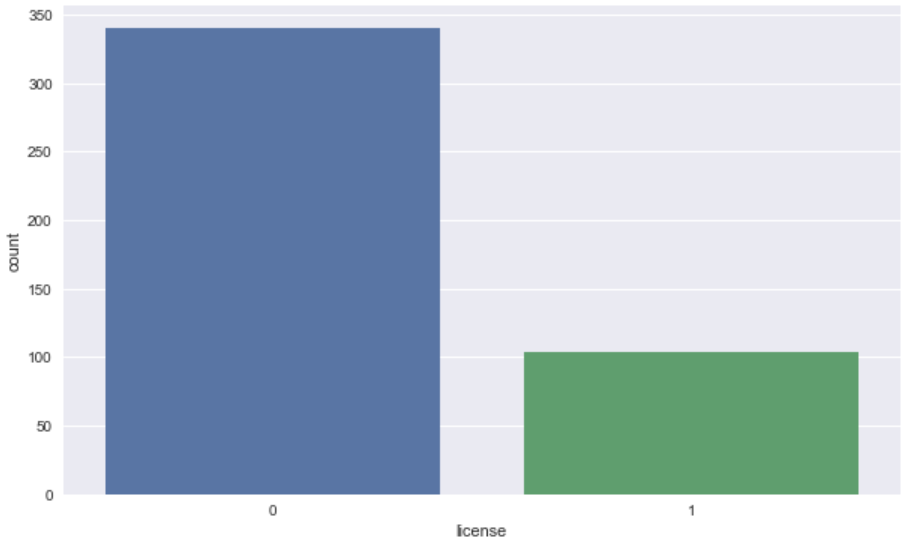
***Figure 6: Distribution of Salary***

The salary amount ranges from 10 to 57 with a mean salary amount of 16. There are outliers present in the dataset.



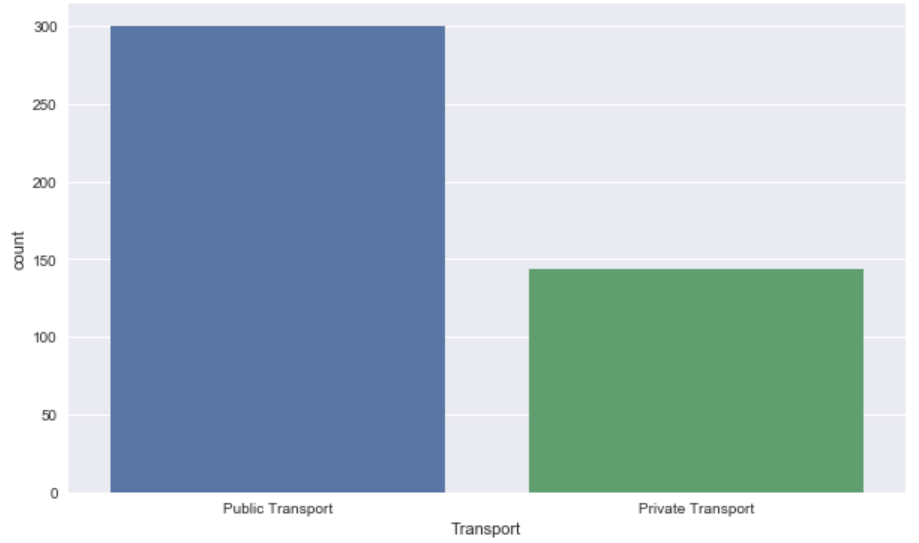
***Figure 7: Distribution of Distance***

The distance of office from residence of individual ranges from 3.2 to 23.4 with an average distance of 11. There are outliers present in the dataset.



***Figure 8: Distribution of License***

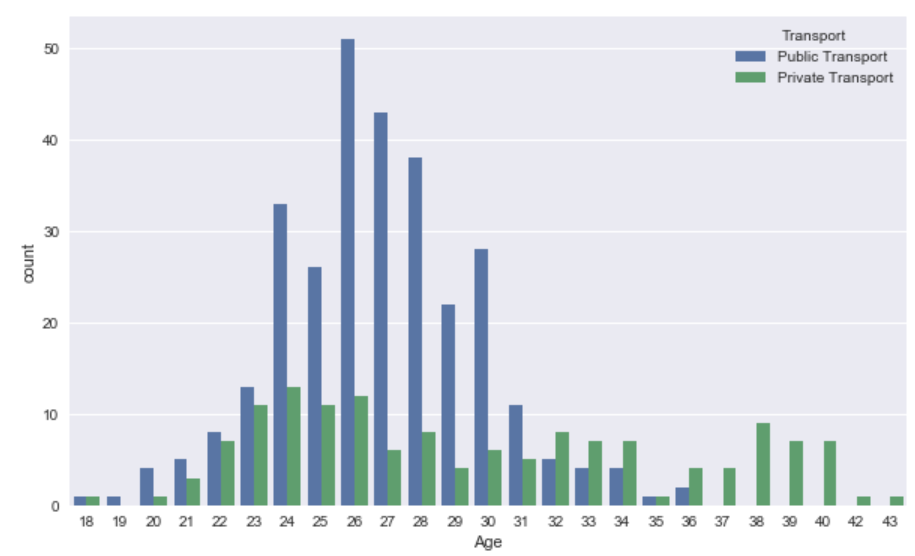
Only 23% of the dataset have License to drive a vehicle rest 77% do not have a license.



***Figure 9: Distribution of Transport***

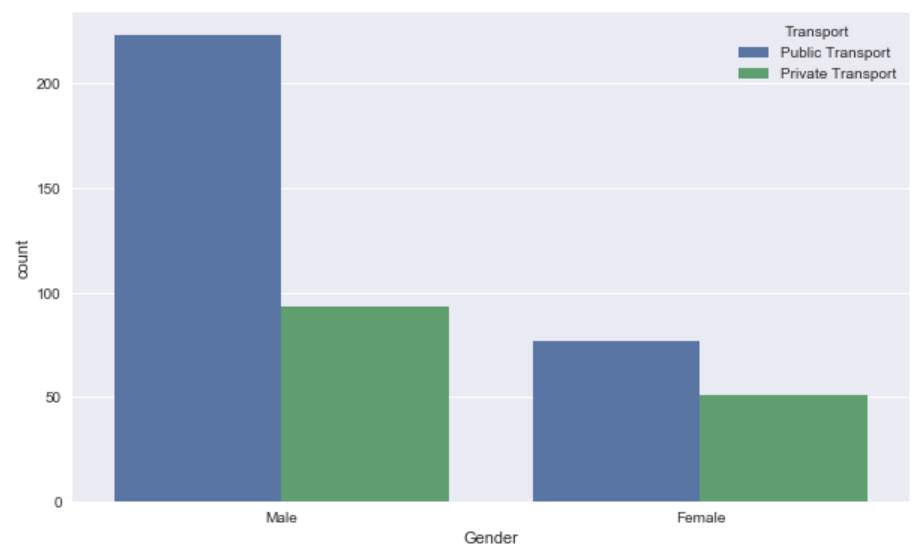
Majority of the individual prefer to use public transport. 67.5% of the people prefer to use public transport and the rest 32.5% use private transport.

Bivariate Analysis:



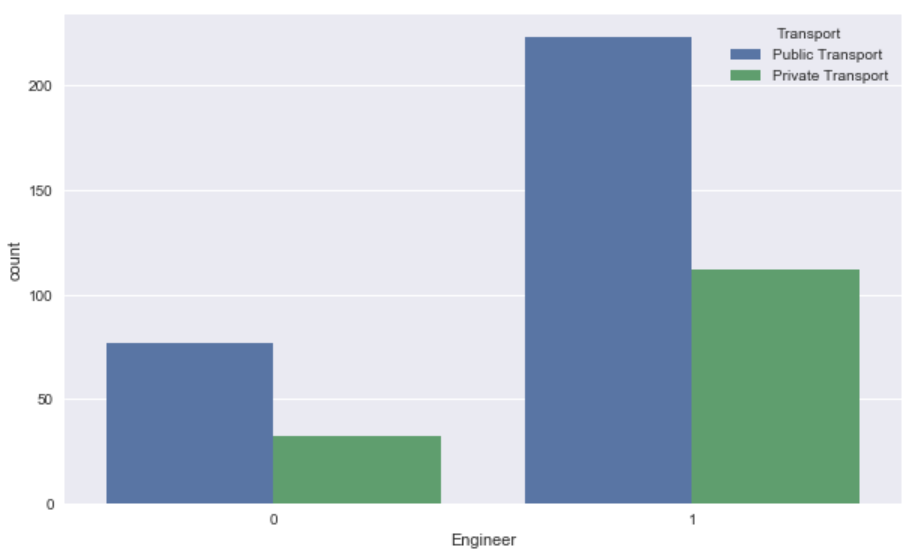
***Figure 10: Age vs Transport***

The above figure shows the relationship between the age of the individual and the preferred mode of transport. People between the age of 24 to 31 prefer to use public transportation more than the private transport but as the age increases people shift to private mode of transport. Individuals between the age of 36 to 43 prefer to use private mode of transport.



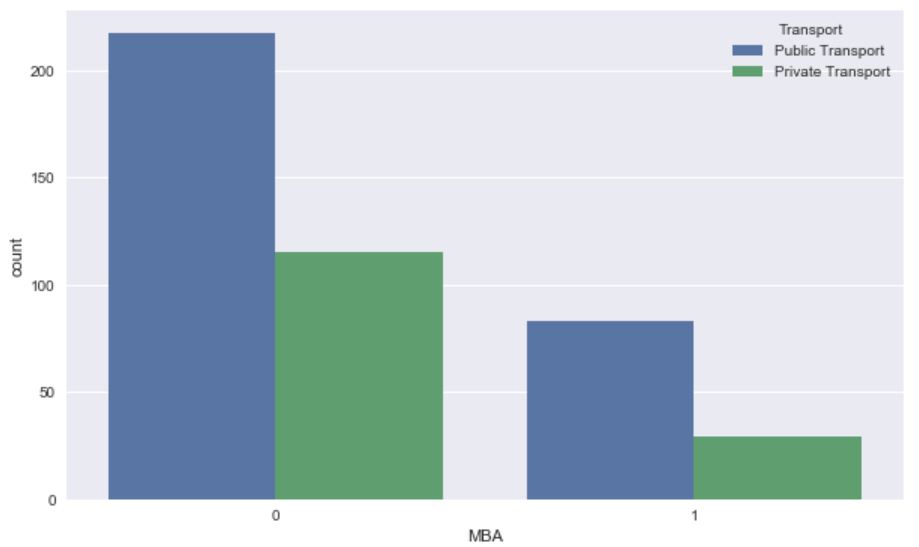
***Figure 11: Gender vs Transport***

The above figure shows the relation between the gender of an individual and the preferred mode of transport. The male population largely prefer public transport than private transport where as the female population only slightly favours public transport over private mode of transportation.



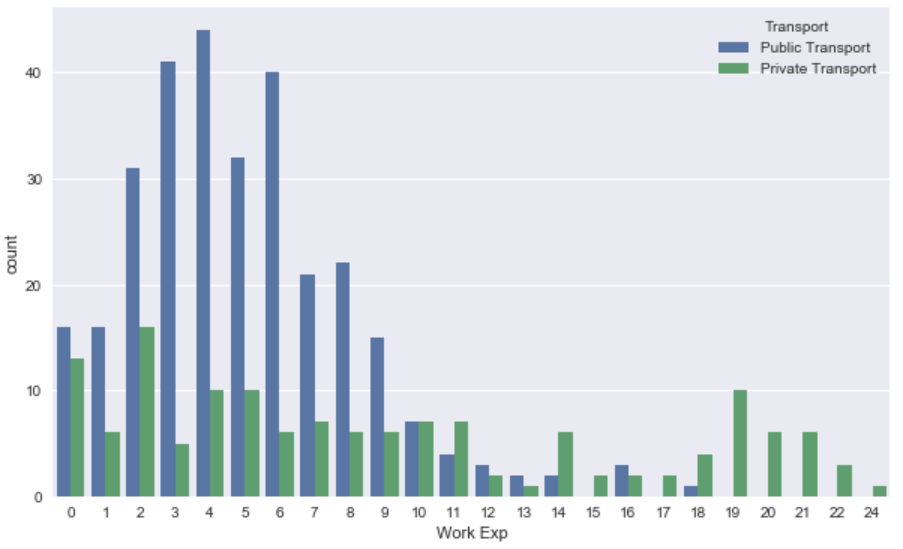
***Figure 12: Engineers vs Transport***

The engineers prefer to us the public mode of transportation more than the private mode of transportation.



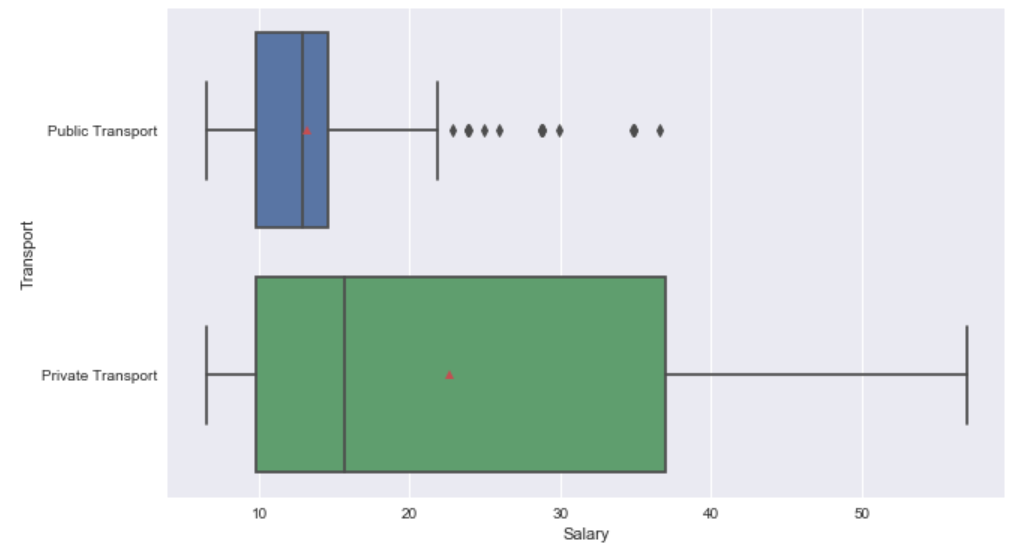
***Figure 13: MBA vs Transport***

The MBAs prefer to use the public mode of transportation more than the private transportation.



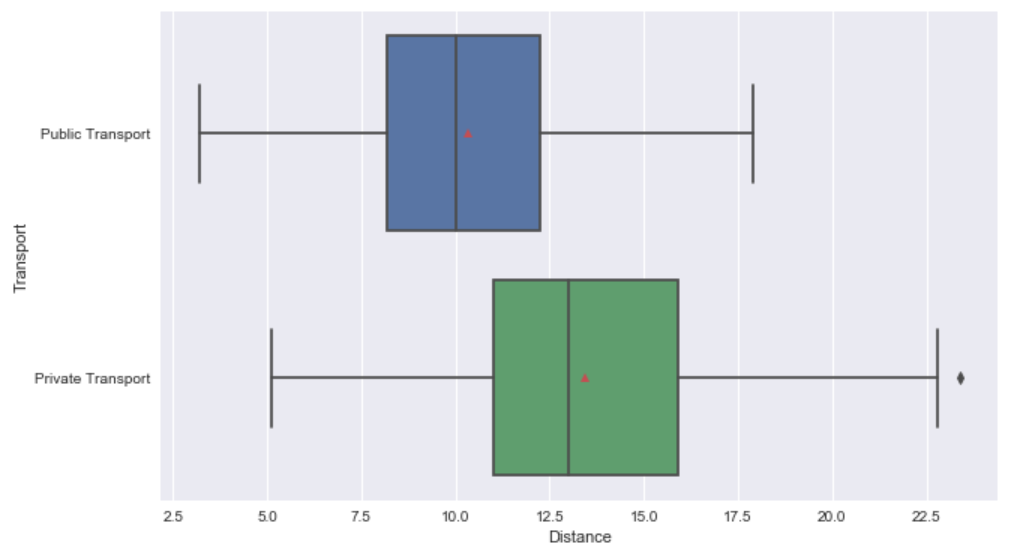
***Figure 14: Work Experience vs Transport***

From the above figure we can say that individuals with less work experience prefers the public transportation over private transport but as the work experience increases, they tend to shift to the private mode of transportation.



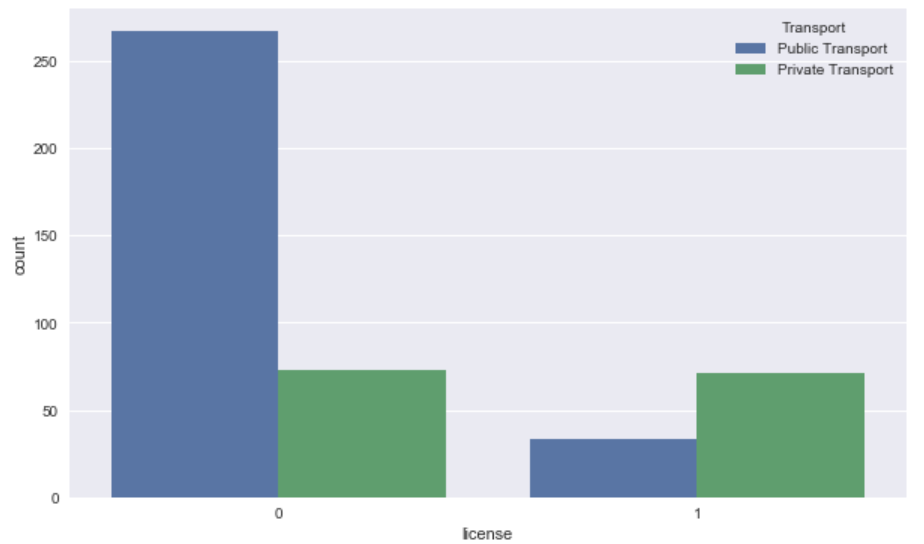
***Figure 15: Salary vs Transport***

Individuals with less salary use the public transport more over private transport but as the salary increases, the individuals uses the private transport more over public transport.



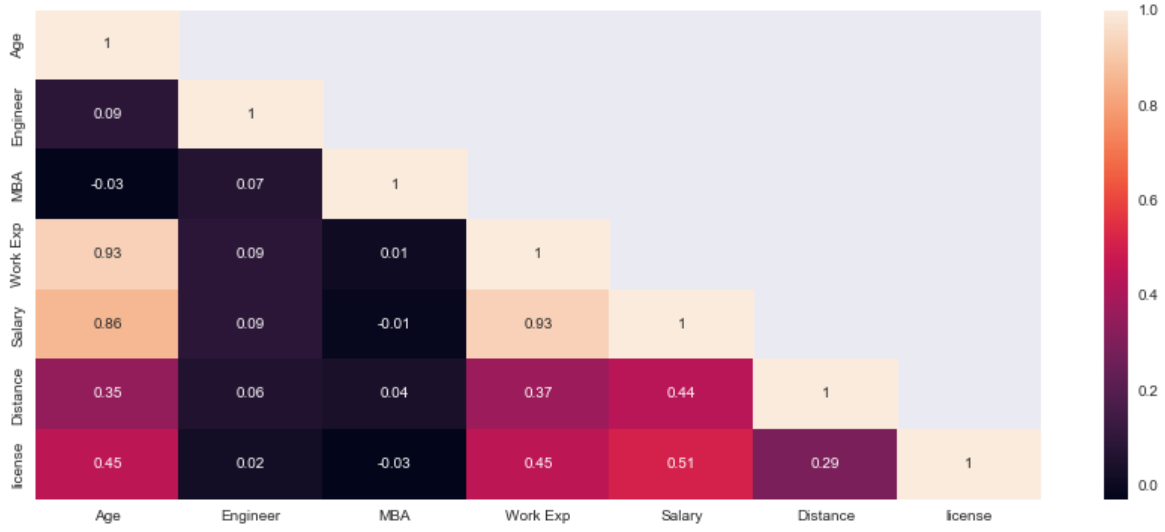
***Figure 16: Distance vs Transport***

The individuals’ choice of transport depends upon the distance to be travelled. Public transport is used when the distance to be travelled is less but when the distance to be travelled increases, the individuals shifts to the private mode of transport.



***Figure 17: License vs Transport***

The choice of the mode of transportation largely depends upon license. Individuals who do not have license choses to travel by public transport where as those individuals who have a license prefer to travel by private transport.

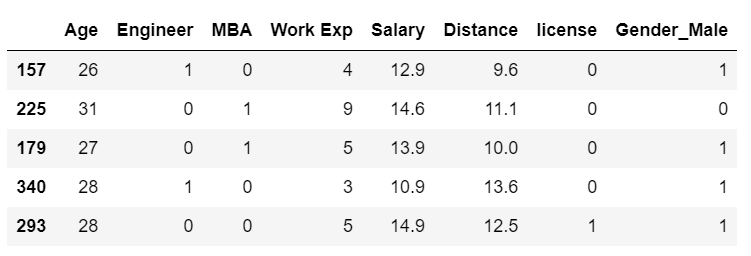


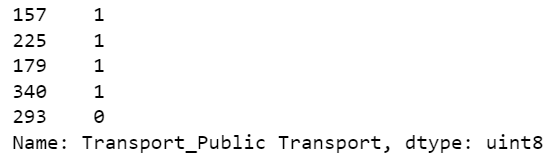
***Figure 18: Heatmap of the Dataset***

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

Solution:

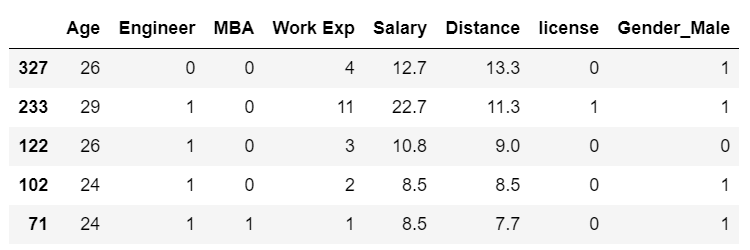
The train and test split of the data is as follows:

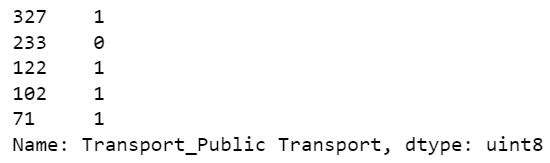




***Table 5: First and last few rows of train data***

The first and the last few rows of the test data is as follows:





***Table 6: First and last few rows of the test data***

**1.3)** **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**

**b. Linear Discriminant Analysis**

**c. Decision Tree Classifier – CART model**

**d. Naïve Bayes Model**

**e. KNN Model**

**f. Random Forest Model**

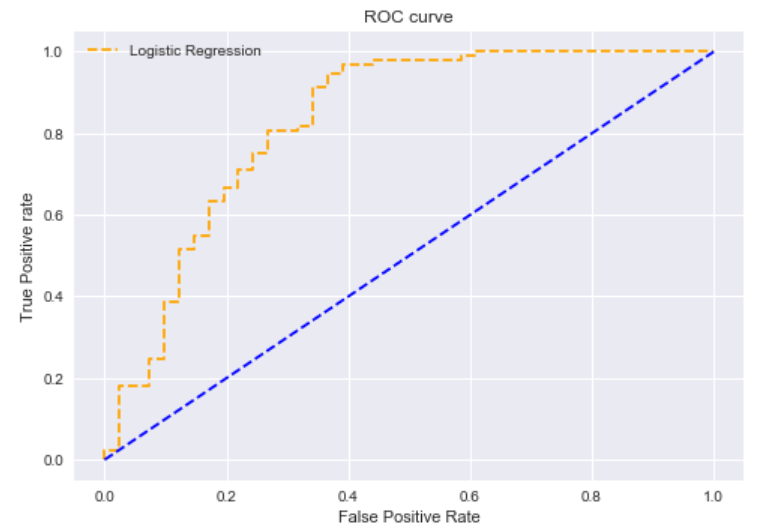
**g. Boosting Classifier Model using Gradient boost.**

Solution:

The dataset was split in 70:30 ratio as train and test dataset and then Models were applied to get the accuracy score. The following are the findings:

1. Logistic Regression Model:

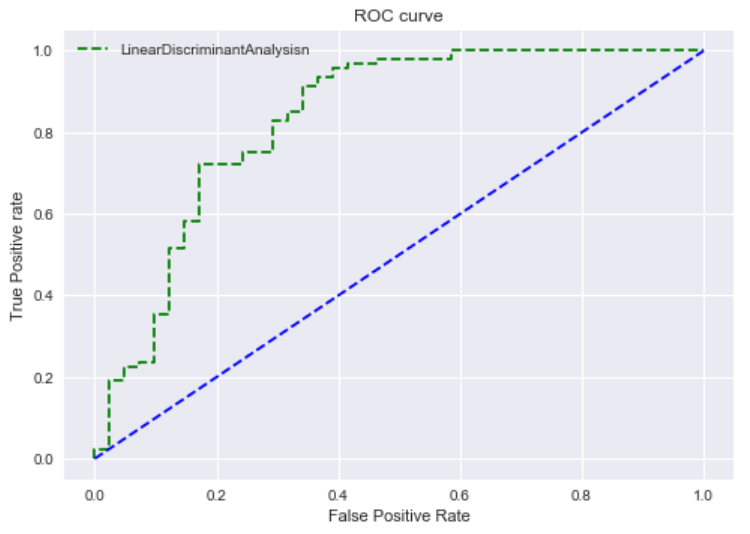
The accuracy score of the Logistic Regression Model is 82.09%. The AUC score of the Model is 0.83 and the ROC curve of the model is as follows:



***Figure 19: ROC Curve of Logistic Regression Model***

1. Linear Discriminant Analysis:

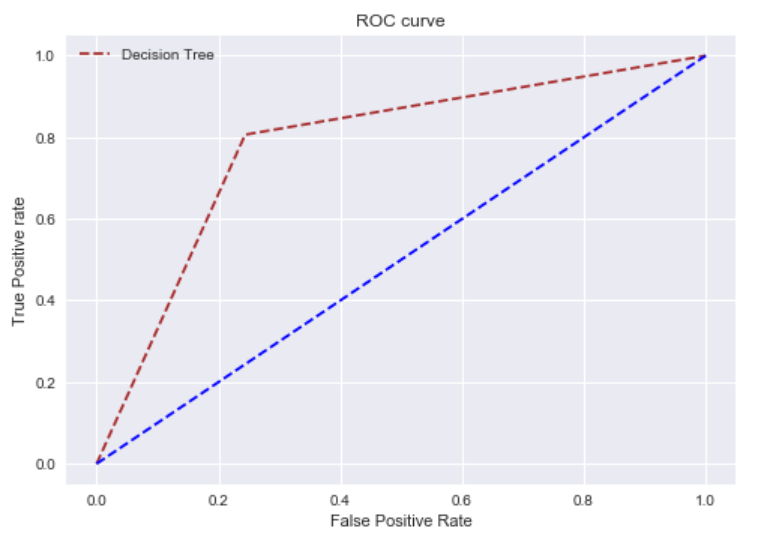
The accuracy score of the LDA model is 82.089%, the AUC score is 0.831 and the ROC curve of the model is as follows:



***Figure 20: ROC Curve of Linear Discriminant Analysis***

1. Decision Tree Classifier:

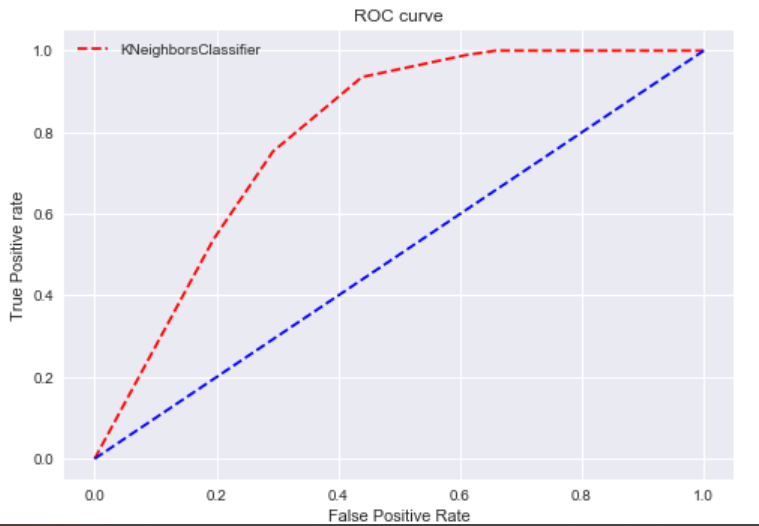
The accuracy score of Decision tree classifier model is 79.10% and the AUC score is 0.78. The ROC curve of the model is as follows.



***Figure 21: ROC Curve of Decision Tree Classifier Model***

e. KNN Model

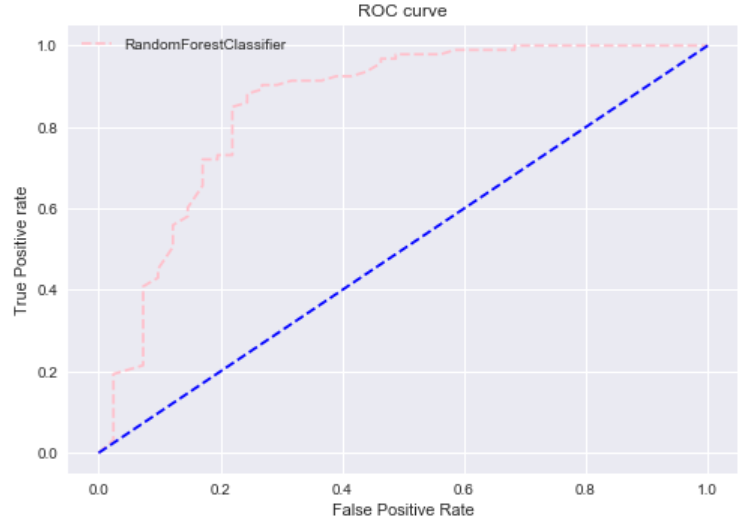
The accuracy score of the Naïve Bayse Model is 82.089 and the AUC score of the model is 0.793. The ROC curve of the model is as follows:



***Figure 22: ROC Curve of KNN Model***

f. Random Forest Model:

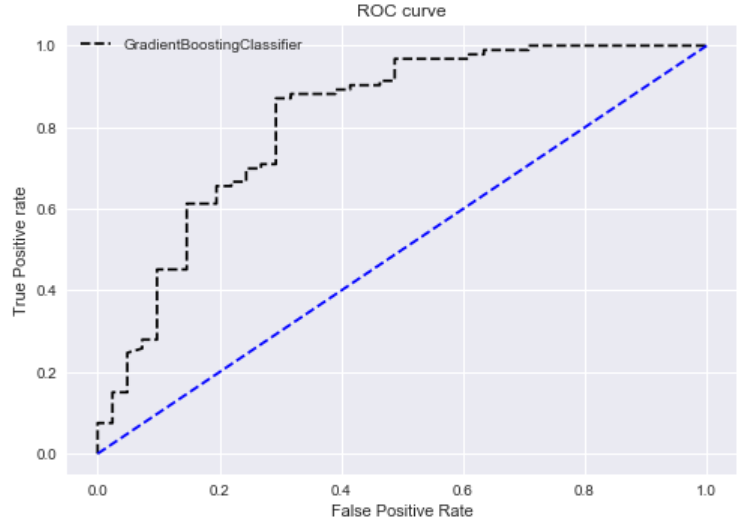
The accuracy score of the Random Forest Model is 84.33% and the AUC score of the model is 0.851. The ROC curve of the model is as follows:



***Figure 23: ROC Curve of Random Forest Classifier Model***

g. Boosting Classifier Model using Gradient boost.

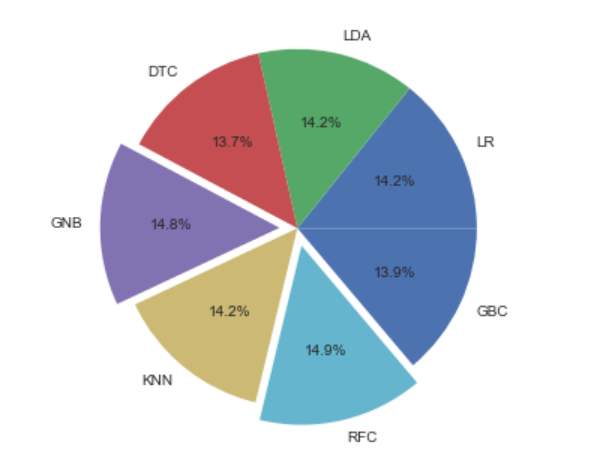
The Accuracy score of the Gradient Boosting model is 79.85% and the AUC score of the model is 0.818. The ROC curve of the model is as follows.

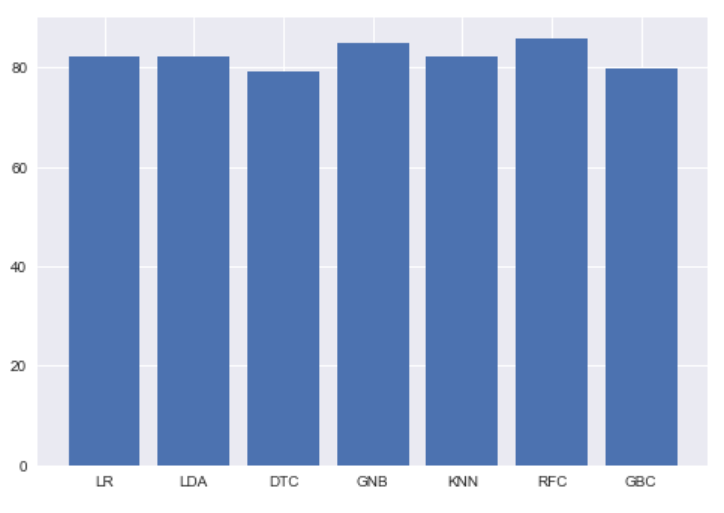


***Figure 24: ROC curve of Gradient Boosting Classifier***

* 1. **Which model performs the best?**

Solution:





***Figure 25: Ranking of the models***

The above figure shows the ranking of all the models that we have used. We can see that the Random Forest Classifier is the model followed by the Gradient Boosting Model.

* 1. **What are your business insights?**

Solution:

Business Insights:

* The dataset Consists of 444 rows and 9 columns and there are no missing values.
* The mean age is 27.7 years with minimum age at 18 and maximum age at 43.
* 71% of the dataset is male and only 29% are female.
* 335 are engineer i.e., 75.4% are engineer and the rest are non-engineer.
* Only 25% i.e., 112 people of the dataset are MBA qualified and the rest 75% are non-MBA.
* The average work experience is approx. 6 years with minimum being 0 years and maximum work experience being 24 years.
* The salary amount ranges from 10 to 57 with a mean salary amount of 16.
* The distance of office from residence of individual ranges from 3.2 to 23.4 with an average distance of 11.
* Only 23% of the dataset have License to drive a vehicle rest 77% do not have a license.
* Majority of the individual prefer to use public transport. 67.5% of the people prefer to use public transport and the rest 32.5% use private transport.
* People between the age of 24 to 31 prefer to use public transportation more than the private transport but as the age increases people shift to private mode of transport. Individuals between the age of 36 to 43 prefer to use private mode of transport.
* The male population largely prefer public transport than private transport whereas the female population only slightly favours public transport over private mode of transportation.
* The engineers prefer to us the public mode of transportation more than the private mode of transportation.
* The MBAs prefer to use the public mode of transportation more than the private transportation.
* that individuals with less work experience prefers the public transportation over private transport but as the work experience increases, they tend to shift to the private mode of transportation.
* Individuals with less salary use the public transport more over private transport but as the salary increases, the individuals uses the private transport more over public transport.
* The individuals’ choice of transport depends upon the distance to be travelled. Public transport is used when the distance to be travelled is less but when the distance to be travelled increases, the individuals shift to the private mode of transport.
* The choice of the mode of transportation largely depends upon license. Individuals who do not have license choses to travel by public transport where as those individuals who have a license prefer to travel by private transport.

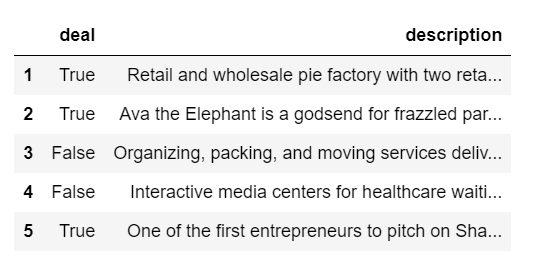
**Problem 2:**

A dataset of Shark Tank episodes is made available. It contains 495 entrepreneurs making their pitch to the VC sharks.

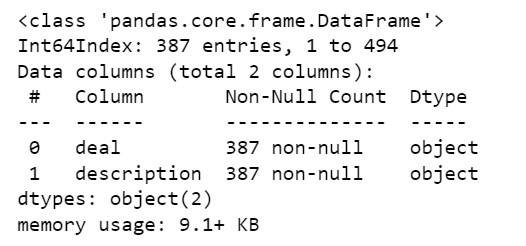
**2.1) Pick out the Deal (Dependent Variable) and Description columns into a separate data frame.**

Solution:

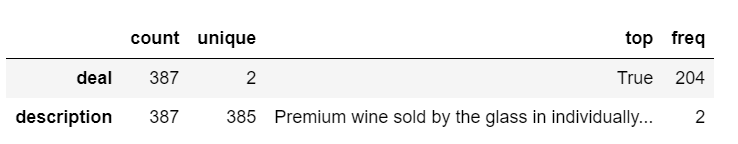
The variables Deal and Description are picked out into a separate dataframe. Following figures displays the output.



***Table 7: First few rows of the dataset***



***Table 8: Info of the dataset***

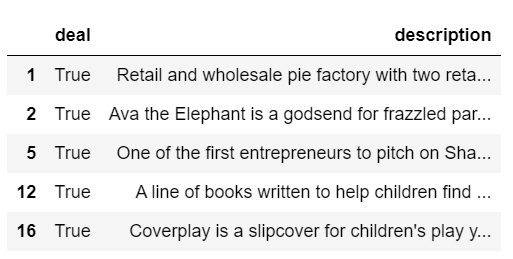


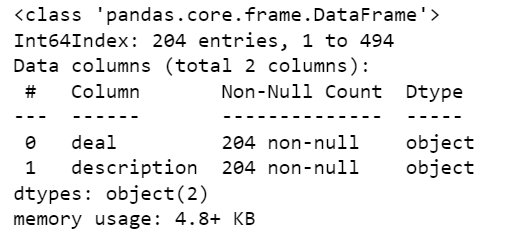
***Table 9: Description of the Dataset***

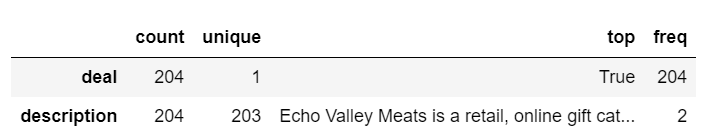
**2.2) Create two corpora, one with those who secured a Deal, the other with those who did not secure a deal.**

Solution:

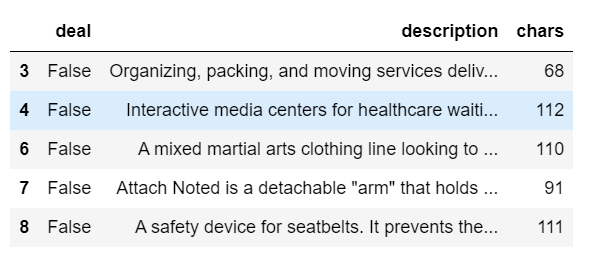
Two corpora were created one with those who secured a deal(true) and one with those who did not secure a deal(false).

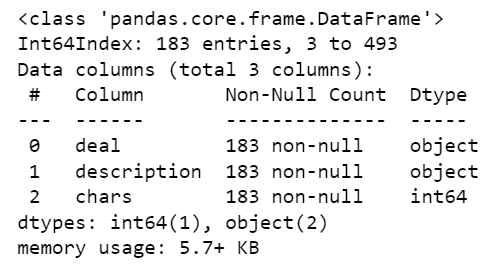


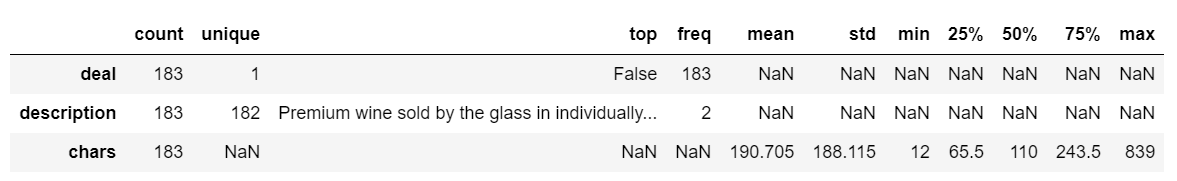




***Table 10: First Few rows, Info & Description of True***







***Table 11: First Few Rows, info & Description of False***

**2.3) The following exercise is to be done for both the corpora:**

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

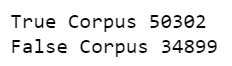
**b) Remove Stop Words from the corpora. (Words like ‘also’, ‘made’, ‘makes’, ‘like’, ‘this’, ‘even’ and ‘company’ are to be removed)**

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

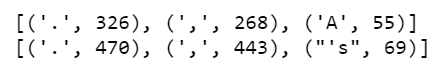
**d) Plot the Word Cloud for both the corpora.**

Solution:

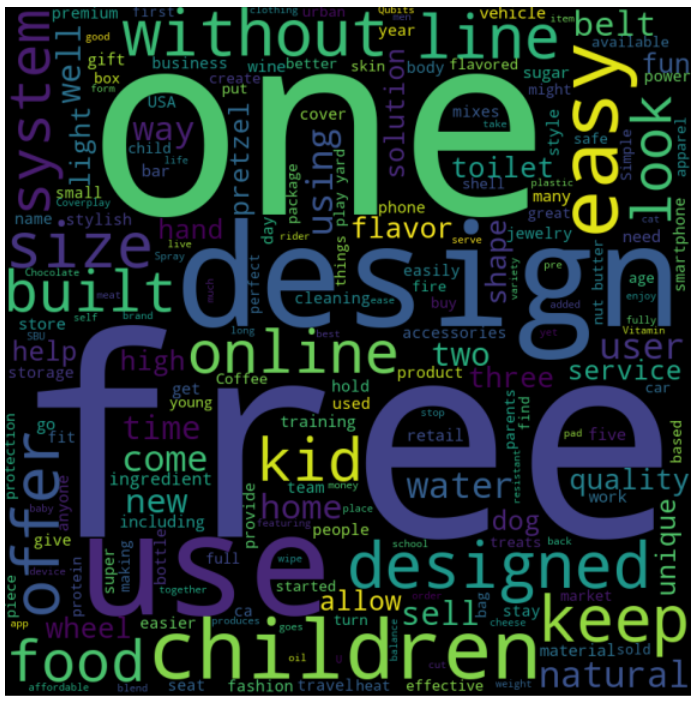
The number of characters for both the corpuses are:



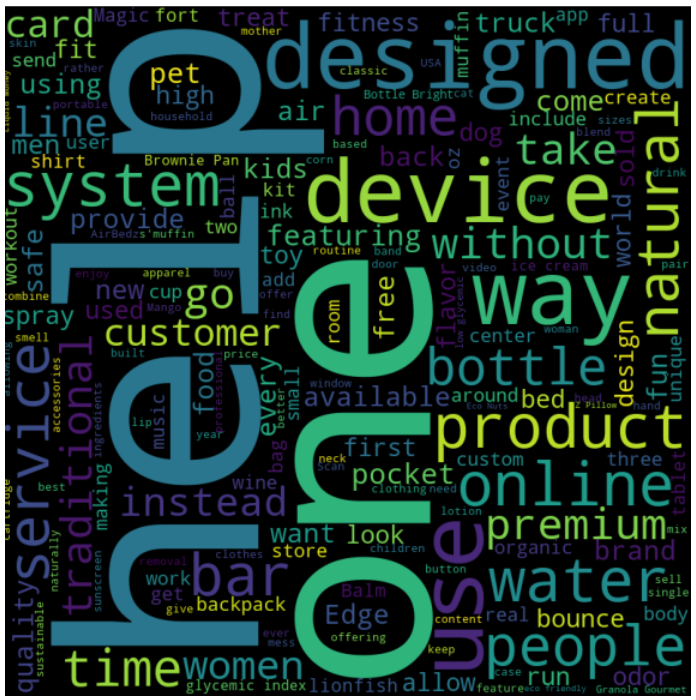
The stop words in both the corpuses are removed. The following are the top 3 frequently used stop words in both the corpuses.



The word cloud for both the corpuses are as follows:



***Figure 26: Word Cloud True (Secured a Deal)***



***Figure 27: Word Cloud False (Did not Secured a Deal)***

**2.4) Refer to both the word clouds. What do you infer?**

Solution:

The 'secured a deal' word cloud contains words such as '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 are more likely to secure a deal.

The 'Did not secure a deal' word cloud contains words such as 'one', 'designed' , 'help' ,'device' ,'bottle', 'premium' ,'use' .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.

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.

**2.5) 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?**

Solution:

The word 'device' is not easily found in the 'secured a deal' word cloud while it is easily spotted in the 'not secured a deal' 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.