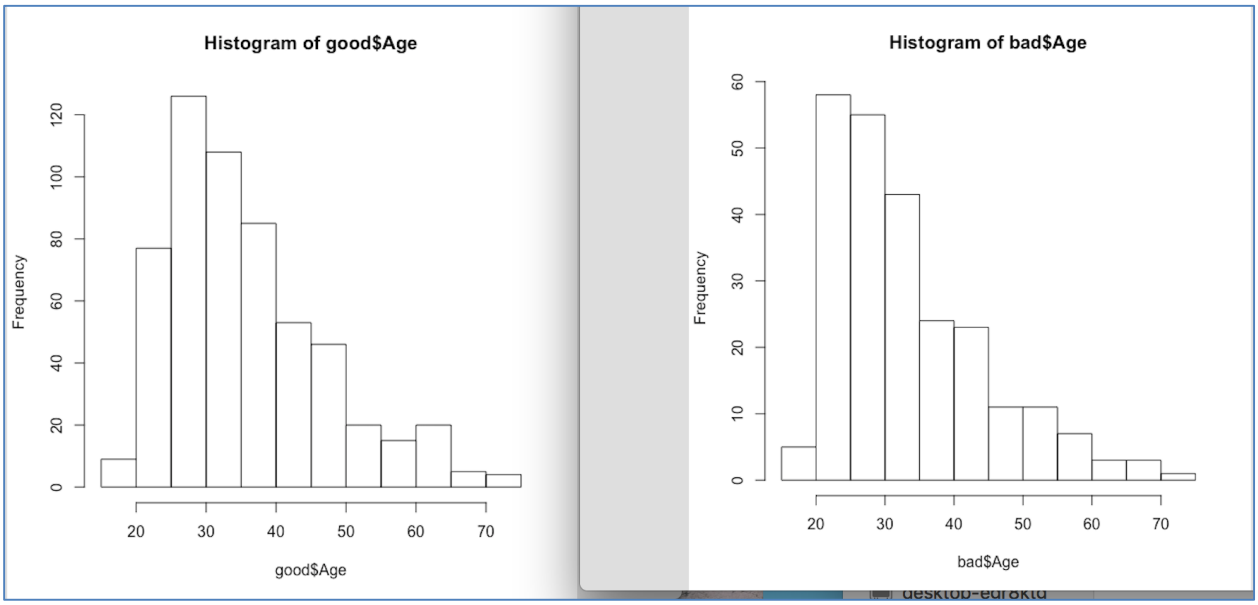
1. **What is the proportion of 'Good credit risk' to 'Bad credit risk' cases? Obtain descriptions of the predictor (independent) variables – mean, standard deviations, etc. for real-valued attributes. Is there anything noteworthy in the data?**

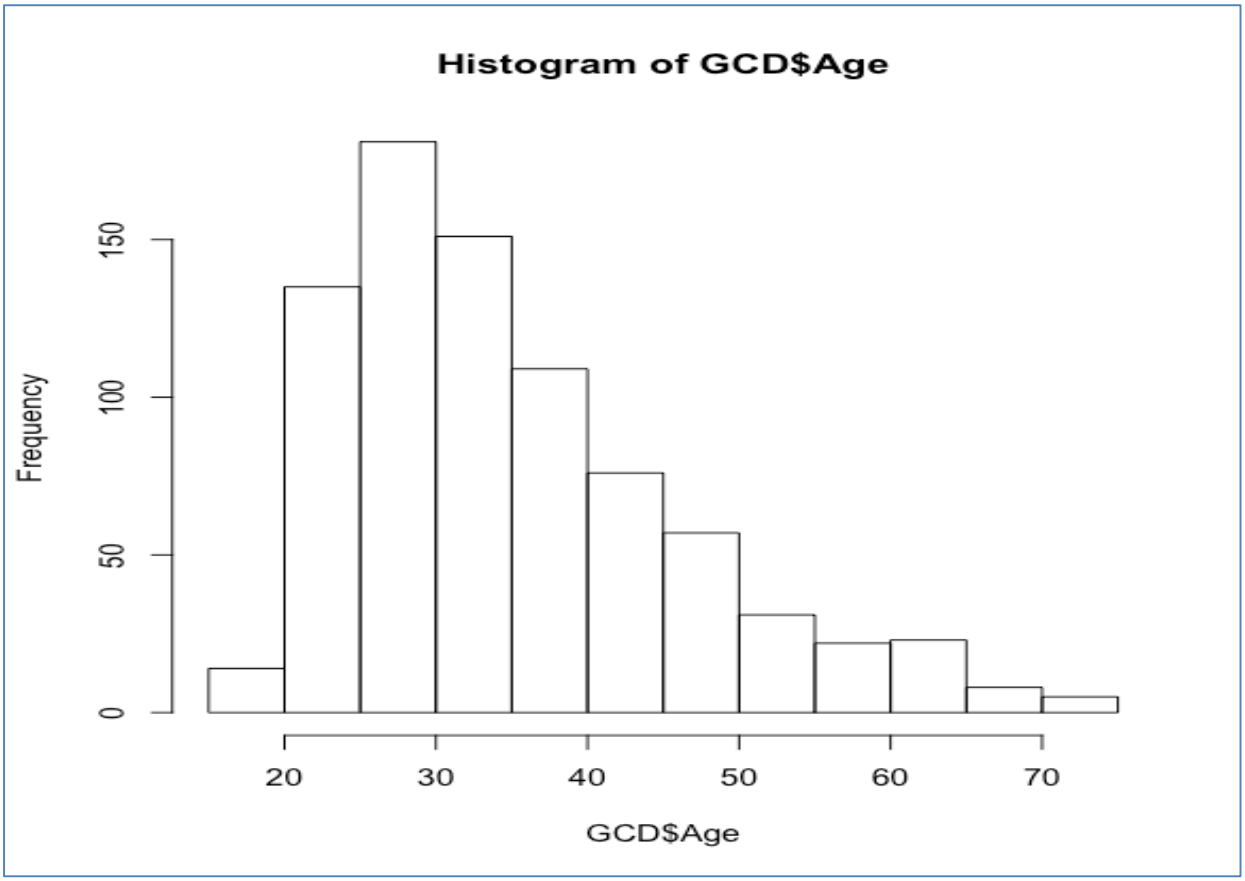
Good Credit Risk Count	568
Bad Credit Risk Count	266
Total Count	812
Good Credit Risk Proportion of whole	69.95074%
Bad Credit Risk Proportion of whole	30.04926%

Note-worth Insights gained from the data.

- I divided the data into good and bad credit risk. Then I plotted histograms for the age column for both good and bad credit risk. I found that people with good credit risk, age count was most for people having age in the range 27-35. For bad credit risk, age count was most for people having age in the range 20-26. It seems logical to think that people at a younger age will tend to have bad credit risk because they are not yet settled in life compared to people at an older age who will be settled and will have a better chance of having a good credit risk.



- In general, looking at the full dataset, it seems that 30 is the age where people tend to borrow the most.



- Also it is seen that people with bad credit risk (bad credit mean – 3865) tend to borrow in average more money compared to people with good credit risk (good credit mean – 3012)

2. **Using the test data, classify each of the test cases as 'Good credit risk' or 'Bad credit risk'. Create a confusion matrix and report the accuracy of each model.**

Confusion matrices

- Decision Tree

	Predicted Class	
Actual Class	1 (Good Credit Risk)	2 (Bad Credit Risk)
1 (Good Credit Risk)	138	28
2 (Bad Credit Risk)	39	39

Accuracy = (138 + 39) / (138 + 28 + 39 + 39) = 0.7254 = in percentage = 72.54%

- Naïve Bayes

	Predicted Class	
Actual Class	1 (Good Credit Risk)	2 (Bad Credit Risk)
1 (Good Credit Risk)	146	20
2 (Bad Credit Risk)	39	39

Accuracy = (146 + 39) / (146 + 20 + 39 + 39) = 0.7582 = in percentage = 75.82%

- Bagging

	Predicted Class	
Actual Class	1 (Good Credit Risk)	2 (Bad Credit Risk)
1 (Good Credit Risk)	158	8
2 (Bad Credit Risk)	46	32

Accuracy = (158 + 32) / (158 + 8 + 46 + 32) = 0.7787 = in percentage = 77.87%

- Boosting

	Predicted Class	
Actual Class	1 (Good Credit Risk)	2 (Bad Credit Risk)
1 (Good Credit Risk)	140	26
2 (Bad Credit Risk)	30	48

Accuracy = (140 + 48) / (140 + 26 + 30 + 48) = 0.7705 = in percentage = 77.05%

- Random Forest

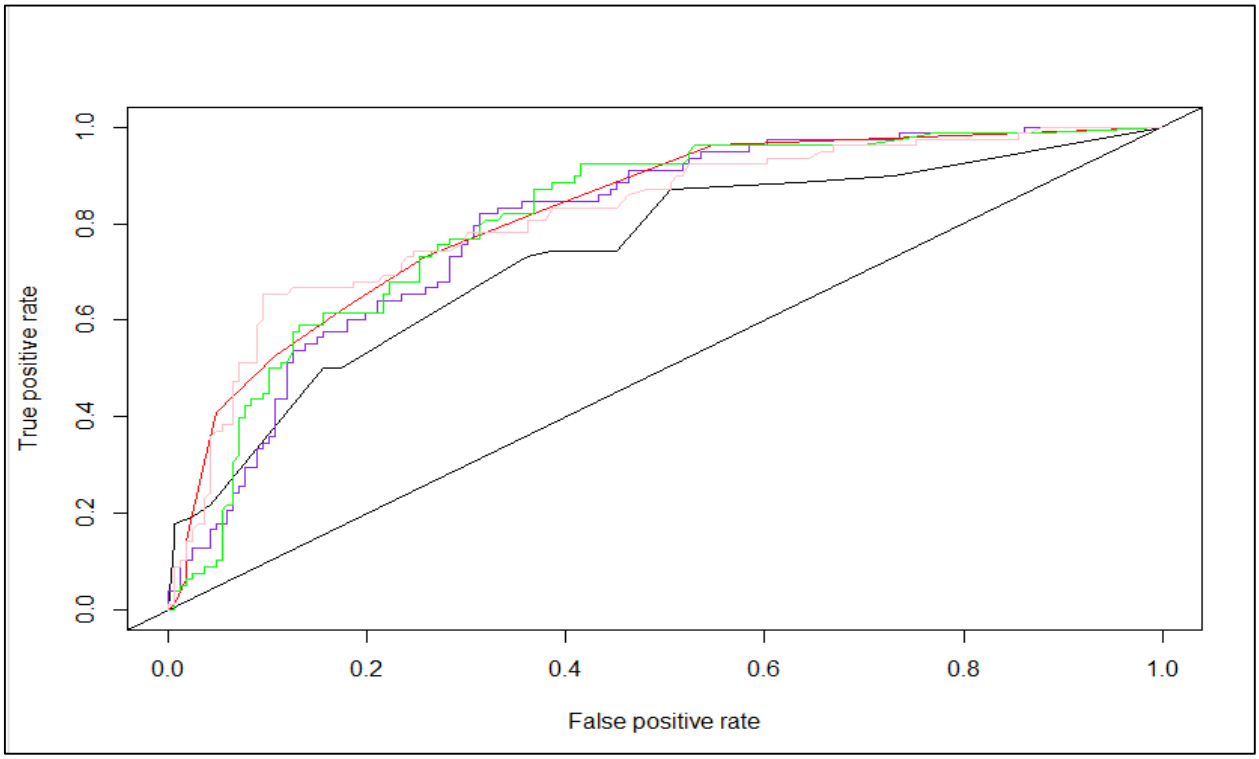
	Predicted Class	
Actual Class	1 (Good Credit Risk)	2 (Bad Credit Risk)
1 (Good Credit Risk)	155	11
2 (Bad Credit Risk)	44	34

Accuracy = (155 + 34) / (155 + 11 + 44 + 34) = 0.7746 = in percentage = 77.46%

3. **Using the test data, calculate the confidence of predicting a 'Good credit risk' for each case and construct an ROC curve for each classifier. You should be able to plot all the curves on the same axis. Use a different colour for each classifier. Calculate the AUC for each classifier.**

Classification Model	AUC
Decision Trees	**0.7377 (BLACK)**
Naïve Bayes	**0.7978 (BLUE)**
Bagging	**0.8202 (RED)**
Boosting	**0.8099 (GREEN)**
Random Forest	**0.8174 (PINK)**

ROC CURVE



LEGEND

Decision Trees	
Naïve Bayes	
Bagging	
Boosting	
Random Forest	

4. Create a table comparing the results in parts 4 and 5 for all classifiers. Is there a single "best" classifier?

All Accuracies

Classification Model	Accuracy
Decision Trees	**72.54%**
Naïve Bayes	**75.82%**
Bagging	**77.87%**
Boosting	**77.05%**
Random Forest	**77.46%**

All AUC'S

Classification Model	AUC
Decision Trees	**0.7377 (BLACK)**
Naïve Bayes	**0.7978 (BLUE)**
Bagging	**0.8202 (RED)**
Boosting	**0.8099 (GREEN)**
Random Forest	**0.8174 (PINK)**

According to both the above tables, Bagging has the highest accuracy and the highest AUC. So the best performance is obtained using bagging.

5. **Examining each of the models, determine the most important variables in predicting whether or not an applicant is a good or bad credit risk. Which variables could be omitted from the data with very little effect on performance? Give reasons.**

```
Classification tree:
tree(formula = Class ~ ., data = GCD.train)
Variables actually used in tree construction:
 [1] "Status"   "Purpose"  "Age"      "Savings"  "Other"    "History"  "Duration"
 [8] "Debtors"  "Existing" "Credit"
Number of terminal nodes:  18
Residual mean deviance:  0.774 = 426 / 550
Misclassification error rate: 0.201 = 114 / 568
> cat("\n#Baging Attribute Importance\n")

#Baging Attribute Importance
> print(GCD.bag$importance)
       Age     Credit    Debtors   Duration Employment   Existing    Foreign
    4.3573     9.0009     1.6333    15.5741     5.9369     0.3293     0.0000
   History    Housing        Job     Liable      Other   Personal   Property
   10.8041     0.8529     0.6604     0.2834     3.6681     2.4784     1.0086
   Purpose       Rate  Residence    Savings     Status  Telephone
   12.9478     1.9023     0.3133     8.0283    19.7151     0.5053
> cat("\n#Boosting Attribute Importance\n")

#Boosting Attribute Importance
> print(GCD.Boost$importance)
       Age     Credit    Debtors   Duration Employment   Existing    Foreign
   10.0986    10.2591     1.6626    12.0028     3.6364     0.8240     0.4066
   History    Housing        Job     Liable      Other   Personal   Property
    9.1069     0.8924     2.3220     1.2547     4.2324     2.9007     7.4632
   Purpose       Rate  Residence    Savings     Status  Telephone
   13.0443     1.9938     2.1470     4.0019    11.1082     0.6425


#Random Forest Attribute Importance
> print(GCD.rf$importance)
           MeanDecreaseGini
Status              24.1176
Duration            22.0492
History             18.6433
Purpose             22.1139
Credit              26.3848
Savings             12.9255
Employment          12.8316
Rate                 8.8759
Personal             7.9314
Debtors              3.3641
Residence            7.9490
Property            10.6320
Age                 21.6762
Other                7.2872
Housing              6.0339
Existing             4.8665
Job                  6.9430
Liable               3.2522
Telephone            4.2338
Foreign              0.8765
>
```

According to the above images which shows importance of each variable for each classifier model, I am measuring the importance of each variable with looking at their numeric count. The higher the numeric count, the higher their importance. So all the variables which are having very low numeric value can be omitted from the data.

Some of the variables that can be omitted are:

- Foreign
- Housing
- Telephone
- Liable

I choose these variables to be omitted because they have very low numeric values for each of the classifier model. Also for decision trees, these variables are not included in building the tree. These are only some of the variables that can be omitted. I have not included all of them.

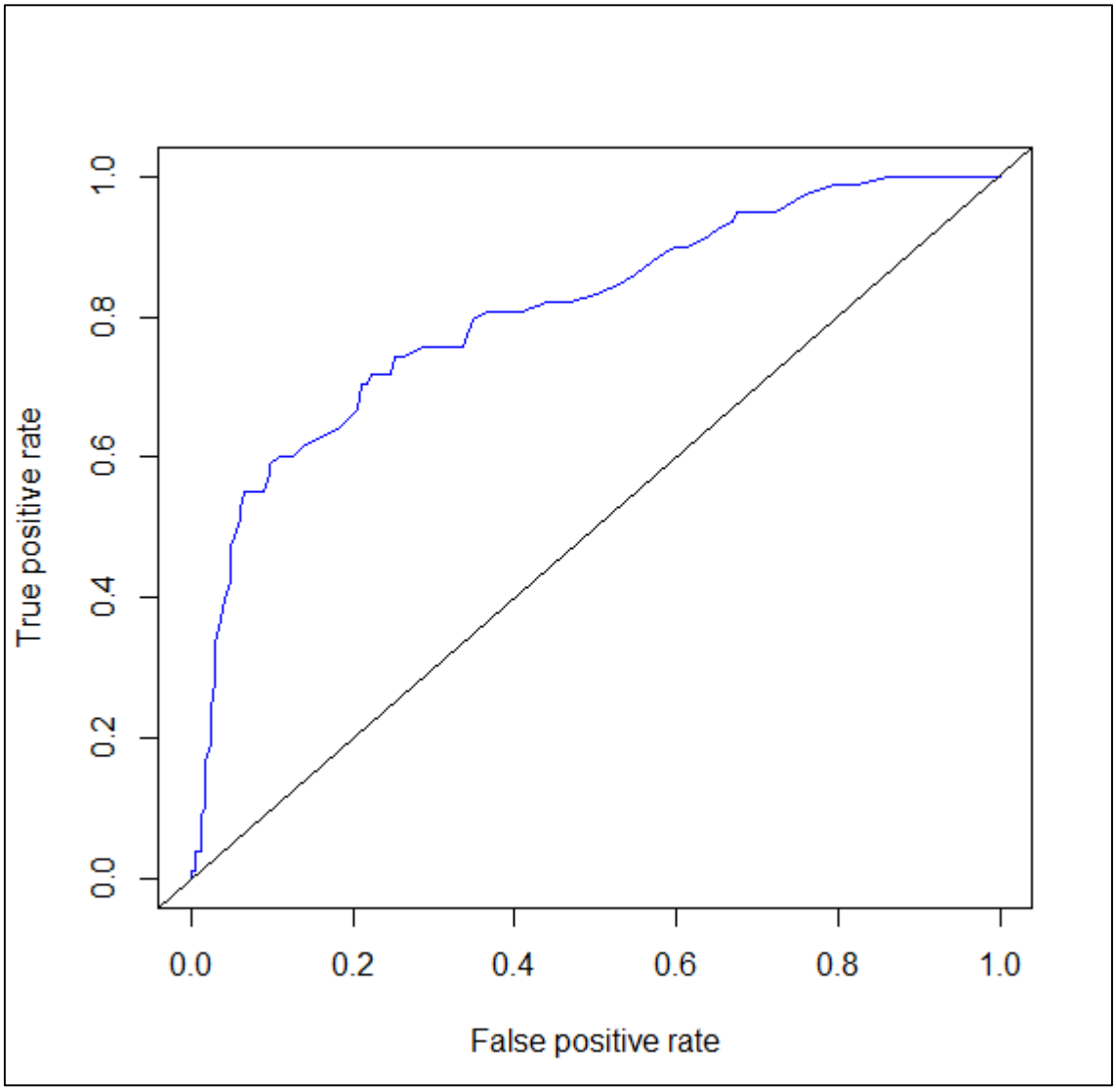
Some of the variables that are very important are:

- Status
- Duration
- Purpose
- Credit

I choose these variables to be important because they have very high numeric values for each of the classifier model. Also for decision trees, these variables are included in building the tree. These are only some of the variables that are important. I have not included all of them.

6. **By experimenting with parameter settings for at least one of the classifiers, create the best classifier you can – that is, one with an accuracy greater than the models you originally created in Part A. Demonstrate this improved accuracy using ROC, AUC, or other accuracy measures. Report the parameter settings and assumptions made in designing this classifier.**

I experimented with using random forest. I used three random forests rather than one and then combined all of them together to produce final model. For each random forest, I used 35 trees.



Confusion Matrix

	Predicted Class	
Actual Class	1 (Good Credit Risk)	2 (Bad Credit Risk)
1 (Good Credit Risk)	156	10
2 (Bad Credit Risk)	38	40

Accuracy = (156 + 40) / (156 + 10 + 38 + 40) = 0.8033 = in percentage = 80.33%

All Accuracies

Classification Model	Accuracy
Decision Trees	**72.54%**
Naïve Bayes	**75.82%**
Bagging	**77.87%**
Boosting	**77.05%**
Random Forest (OLD)	**77.46%**
Random Forest (NEW)	**80.33%**

All AUC'S

Classification Model	AUC
Decision Trees	**0.7377**
Naïve Bayes	**0.7978**
Bagging	**0.8202**
Boosting	**0.8099**
Random Forest (OLD)	**0.8174**
Random Forest (NEW)	**0.8398**

The new random forest has the highest accuracy and highest AUC compared to all other classifiers.

7. **Using the insights from your analysis so far, implement an artificial neural network classifier and report its performance. Comment on attributes used and your data pre-processing required. How does this classifier compare with the others? Can you give any reasons?**

Attributes used:

I used duration, age, credit, status, purpose and history. I used these because they have very high importance score which I found out from my previous analysis.

Pre-processing:

- I converted the Class attribute to numeric.

- I normalized the age. Credit and duration columns. I did this because the impact on response variables by the feature having greater numeric range could be more than the one having less numeric range, and this could, in turn, **impact prediction accuracy**. **The objective is to improve predictive accuracy** and not allow a particular feature impact the prediction due to large numeric value range.
- I augmented the categorical attributes status, purpose and history by making indicator columns.

Confusion matrix

	Predicted Class	
Actual Class	1 (Good Credit Risk)	2 (Bad Credit Risk)
1 (Good Credit Risk)	158	25
2 (Bad Credit Risk)	35	26

Accuracy = (158 + 26) / (158 + 25 + 35 + 26) = 0.7541 = in percentage = 75.41%

Compared to other classifier models especially random forest, ANN has got a lower accuracy.

Reasons for this

This could be because of too little training data. Hence the network has not been able to learn general features from the training data. So it was not able to predict the unseen test data too well.