US Research University Prediction Model

Philip Gabriel Andrada

November 15, 2016

Preparation

```r
# loading necessary libraries
library(rpart)
library(randomForest)
```

```
## randomForest 4.6-12
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```r
library(tree)
library(party)
```

```
## Loading required package: grid
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

```
## Loading required package: strucchange
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```r
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:randomForest':
##
##     margin
```

```r
library(Boruta)
```

```
## Loading required package: ranger

##
## Attaching package: 'ranger'

## The following object is masked from 'package:randomForest':
##
##     importance
```

```r
library(e1071)
library(ROCR)
```

```
## Loading required package: gplots

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```r
library(corrplot)
library(ggplot2)
```

```r
#Reading Data Files
usuniv2010 <- read.csv("C:\\Users\\pandrada\\Desktop\\Capstone\\MERGED2010_11_PP.csv")
usuniv2011 <- read.csv("C:\\Users\\pandrada\\Desktop\\Capstone\\MERGED2011_12_PP.csv")
usuniv2012 <- read.csv("C:\\Users\\pandrada\\Desktop\\Capstone\\MERGED2012_13_PP.csv")
usuniv2013 <- read.csv("C:\\Users\\pandrada\\Desktop\\Capstone\\MERGED2013_14_PP.csv")
usuniv2014 <- read.csv("C:\\Users\\pandrada\\Desktop\\Capstone\\MERGED2014_15_PP.csv")

#Binding All Data Files into One Data Frame
usuniv <- rbind(usuniv2010,usuniv2011,usuniv2012,usuniv2013,usuniv2014)
```

```
## Warning in `[<-.factor`(`*tmp*`, ri, value = c(100200L, 105200L,
## 2503400L, : invalid factor level, NA generated

## Warning in `[<-.factor`(`*tmp*`, ri, value = c(100200L, 105200L,
## 2503400L, : invalid factor level, NA generated

## Warning in `[<-.factor`(`*tmp*`, ri, value = c(100200L, 105200L,
```

```
## 2503400L, : invalid factor level, NA generated

## Warning in `[<-.factor`(`*tmp*`, ri, value = c(100200L, 105200L,
## 2503400L, : invalid factor level, NA generated

## Warning in `[<-.factor`(`*tmp*`, ri, value = c(100200L, 105200L,
## 2503400L, : invalid factor level, NA generated

## Warning in `[<-.factor`(`*tmp*`, ri, value = c(100200L, 105200L,
## 2503400L, : invalid factor level, NA generated
```

```
#Since there are some incomplete Carnegie Classifications, we use usuniv2014 as basis for the classific
usuniv$CCBASIC2 <- usuniv2014$CCBASIC[match(usuniv$OPEID6,usuniv2014$OPEID6)]

#added the ACCEPTED column for those that are research universities (CCBASIC2 is equal to 15 or 16), as
usuniv$ACCEPTED <- ifelse(usuniv$CCBASIC2 %in% c(15,16), 1, 0)

#number of rows in the usuniv data frame
rows_usuniv <- nrow(usuniv)
rows_usuniv
```

```
## [1] 38389
```

```
#number of columns that are in the usuniv data frame
ncol(usuniv)
```

```
## [1] 1745
```

```
#number of rows that are research universities in the data frame before cleansing
rows_usunivaccepted <- nrow(usuniv[usuniv$ACCEPTED == 1,])
rows_usunivaccepted
```

```
## [1] 1154
```

```
#grab a head of research universities to see if we got the correct ones
head(usuniv[usuniv$ACCEPTED == 1,c(4,1744:1745)], 30)
```

```
##                                                   INSTNM CCBASIC2
## 2                     University of Alabama at Birmingham       15
## 4                     University of Alabama in Huntsville       16
## 6                               The University of Alabama       16
## 10                                      Auburn University       16
## 50                             University of South Alabama       16
## 61                          University of Alaska Fairbanks       16
## 82                          Arizona State University-Tempe       15
## 84                                   University of Arizona       15
## 113                            Northern Arizona University       16
## 144                                 University of Arkansas       15
## 237                    California Institute of Technology       15
## 254                    University of California-Berkeley       15
## 255                       University of California-Davis       15
```

```
## 256                    University of California-Irvine       15
## 257                University of California-Los Angeles       15
## 258                  University of California-Riverside       15
## 259                   University of California-San Diego       15
## 261               University of California-Santa Barbara       15
## 262                  University of California-Santa Cruz       15
## 294                       Claremont Graduate University       16
## 518                          San Diego State University       16
## 567                   University of Southern California       15
## 604 University of Colorado Denver/Anschutz Medical Campus       16
## 607                     University of Colorado Boulder       15
## 614                           Colorado School of Mines       16
## 616               Colorado State University-Fort Collins       15
## 627                               University of Denver       16
## 644                    University of Northern Colorado       16
## 675                          University of Connecticut       15
## 720                                    Yale University       15
##     ACCEPTED
## 2          1
## 4          1
## 6          1
## 10         1
## 50         1
## 61         1
## 82         1
## 84         1
## 113        1
## 144        1
## 237        1
## 254        1
## 255        1
## 256        1
## 257        1
## 258        1
## 259        1
## 261        1
## 262        1
## 294        1
## 518        1
## 567        1
## 604        1
## 607        1
## 614        1
## 616        1
## 627        1
## 644        1
## 675        1
## 720        1
```

```
#Create a vector with the columns that is needed from the study
# 19 - institution region (1-New England, 2-Mid East, 3-Great Lakes, 4-Plains, 5-Southeast, 6-Southwest
# 37-38 - admission rate
# 39-61 - SAT and ACT Scores
# 62-99 - percentage of degrees awarded for each field of study
```

```r
# 293-299 - total share of enrollment for different ethnicities
# 300 - total share of enrollment that are non-resident aliens (i.e. international students)
# 301 - total share of enrollment that have unknown race
# 314 - share of undergraduate, degree-/certificate-seeking students who are part-time
# 377 - average cost of attendance in an academic year institution
# 379 - in-state tuition and fees
# 380 - out-of-state tuition and fees
# 387 - completion rate of first-time, full-time students at four-year institutions with 150% of expect
# 397-403 - completion rate for first-time, full-time students for different ethnicities
# 404 - completion rate for first-time, full-time students for non-resident aliens
# 405 - completion rate for first-time, full-time students that have unknown race
# 429 - retention rate for first-time, full time students at four-year institutions
# 438 - percent of all federal undergraduate students receiving a federal student loan
# 1412 - percentage of first-generation students
# 1740-1741 - total share of enrollment per gender
# 1745 - acceptance flag
col_select <- c(19,37:38,61:99,293:301,314,377,379:380,387,397:405,429,438,1412,1740:1741, 1744, 1745)

# Create a new data frame with the columns that will be filtered out
usunivfilter <- usuniv[,col_select]

# Change the factor columns to numeric for faster processing
for (i in 1:ncol(usunivfilter)){
  usunivfilter[,i] <- as.numeric(as.character(usunivfilter[,i]))
}
```

```
## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion
```

```
## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion
```

```
## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion
```

```
## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion
```

```r
# Clean the results to have all complete
usunivfilter <- usunivfilter[!is.na(usunivfilter$C150_4),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$C150_4_ASIAN),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$C150_4_WHITE),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$C150_4_BLACK),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$C150_4_NRA),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$ADM_RATE_ALL),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$SAT_AVG_ALL),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$UGDS_ASIAN),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$UGDS_WHITE),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$UGDS_BLACK),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$UGDS_NRA),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$UGDS_WOMEN),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$UGDS_MEN),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$COSTT4_A),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP11),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP12),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP14),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP15),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP24),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP26),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP27),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP40),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP45),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP51),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCIP52),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PCTFLOAN),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PPTUG_EF),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$RET_FT4),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$PAR_ED_PCT_1STGEN),]

#We will create another data frame for the research universities only
usresearchuniv <- usunivfilter[usunivfilter$CCBASIC2 %in% c(15,16),]

#show number of rows in the filtered usuniv
rows_usunivfilter <- nrow(usunivfilter)
rows_usunivfilter
```

```
## [1] 4247
```

```r
#percentage of data from filtered to unfiltered
rows_usunivfilter / rows_usuniv
```

```
## [1] 0.1106306
```

```r
#show number of rows of filtered research universities
rows_usresearchuniv <- nrow(usresearchuniv)
rows_usresearchuniv
```

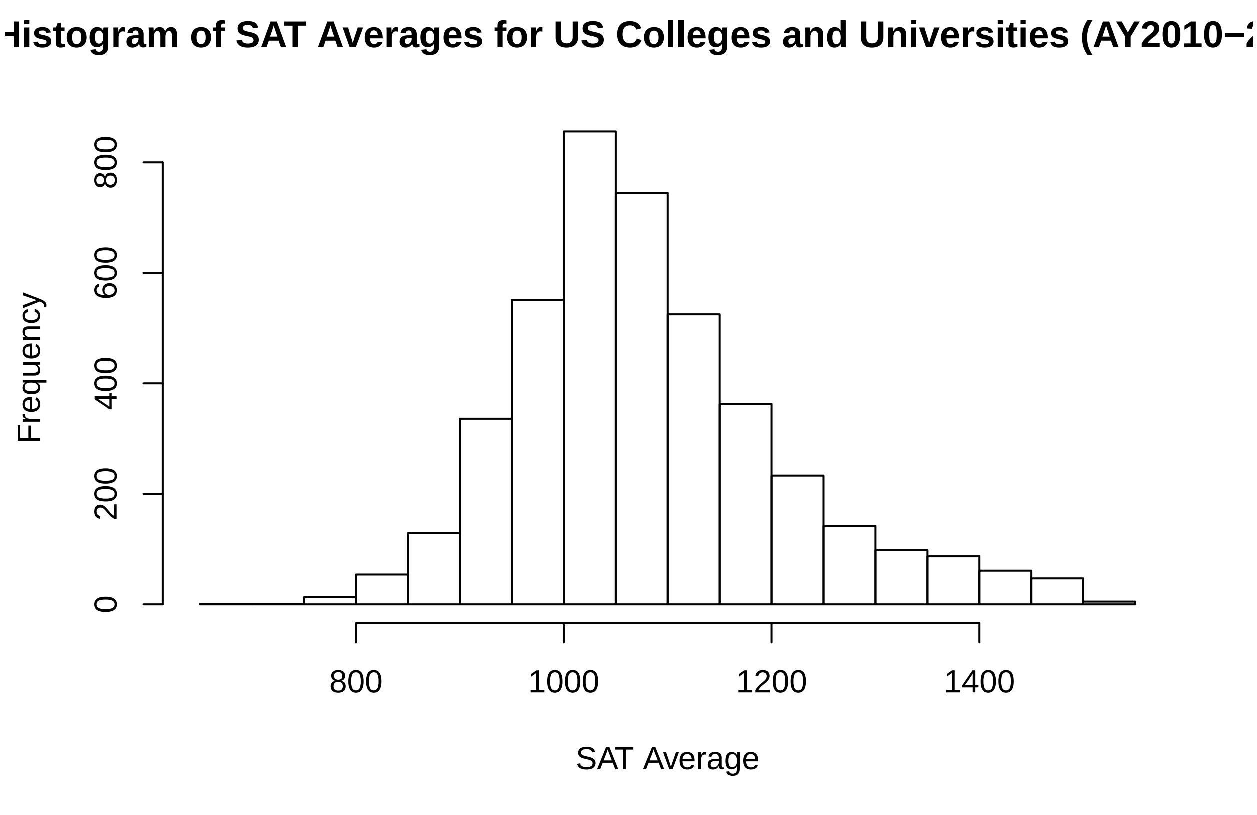
```
## [1] 815
```

```r
#percentage of data from filtered research universities to unfiltered
rows_usresearchuniv / rows_usunivaccepted
```

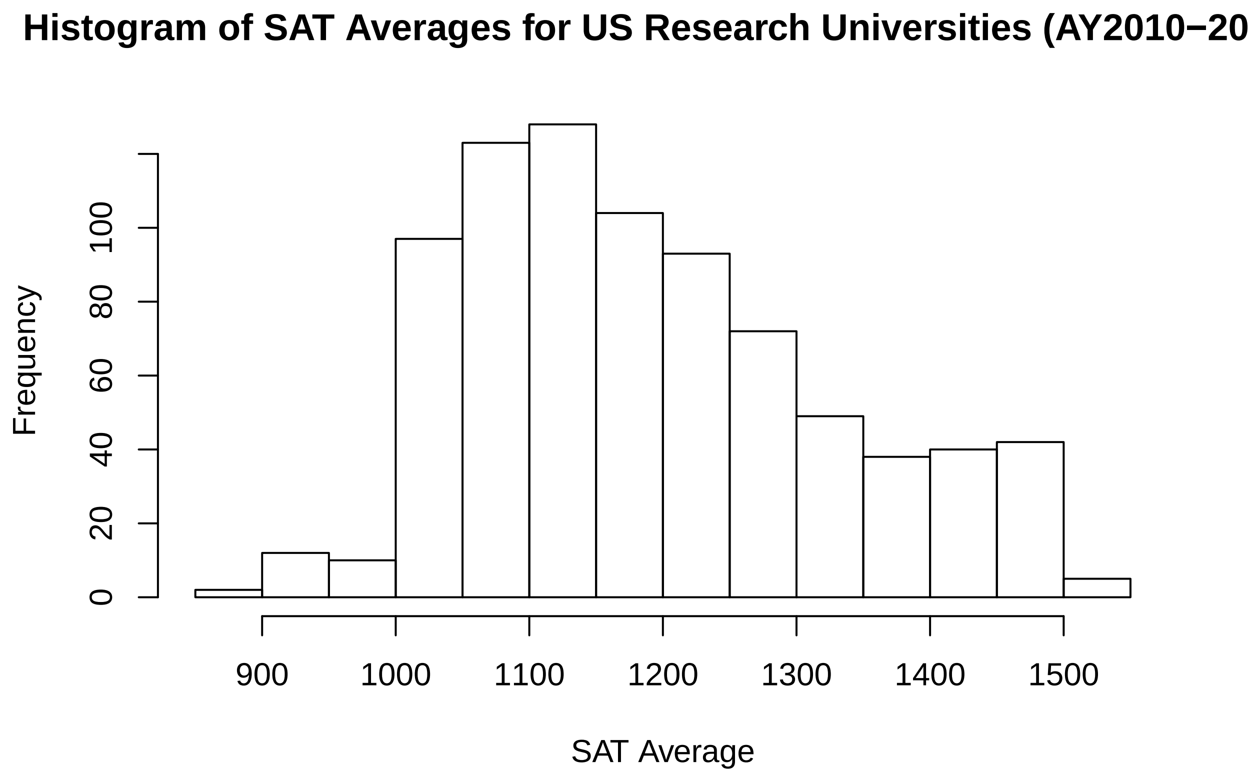
```
## [1] 0.7062392
```

Distributions and Box and Whisker Plots

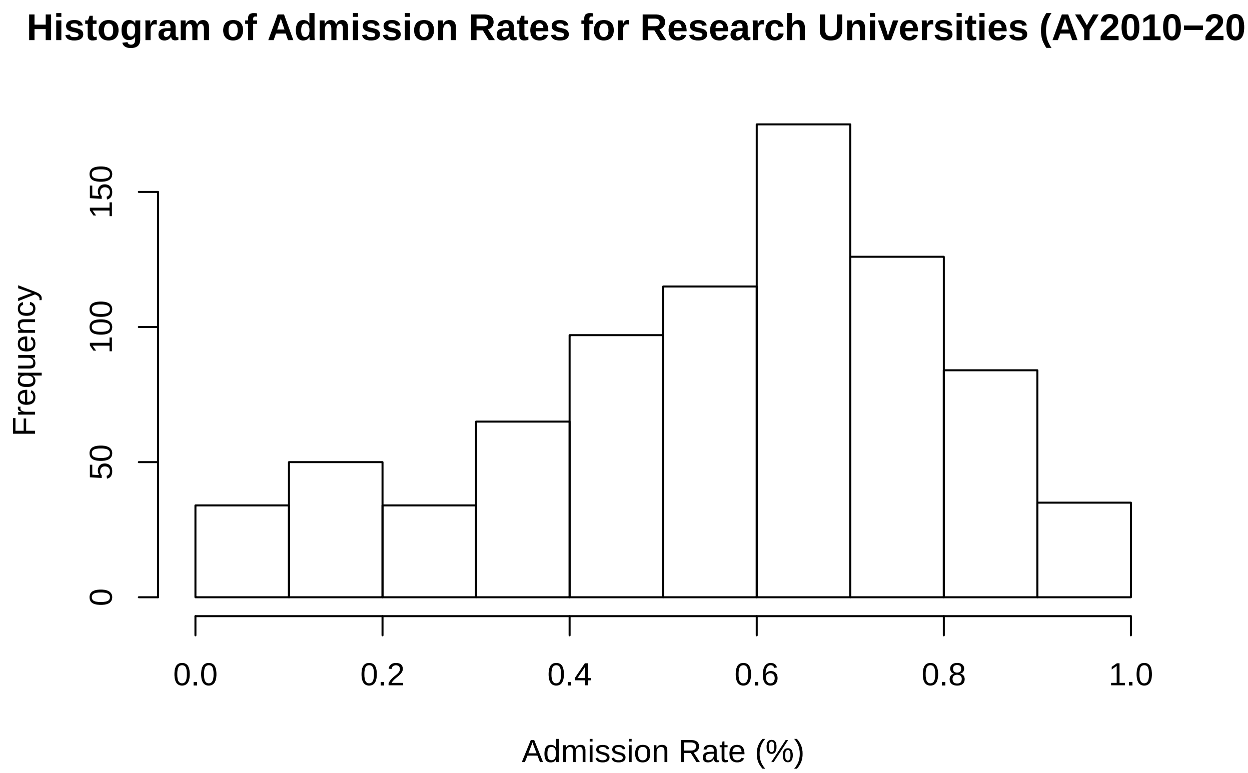
```r
# Histogram of SAT Averages for US Colleges and Universities
hist(usunivfilter$SAT_AVG_ALL, main = "Histogram of SAT Averages for US Colleges and Universities (AY20
```



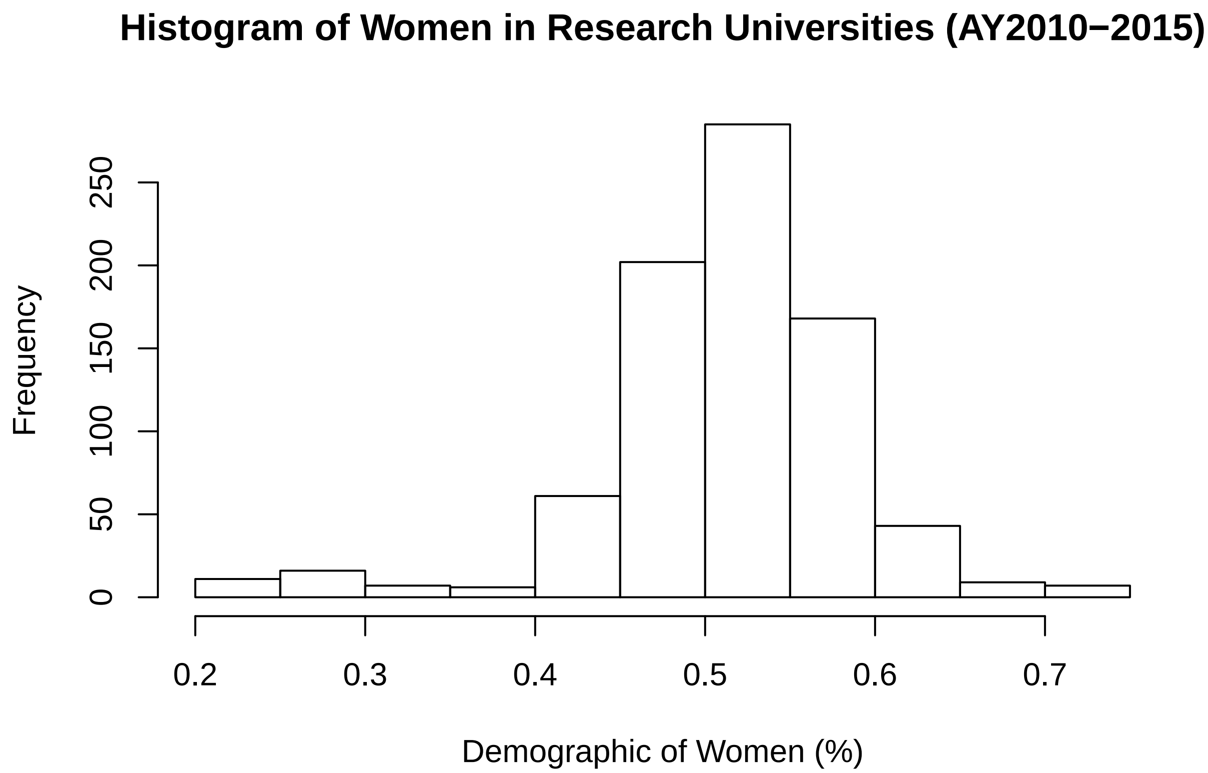

```r
# Histogram of SAT Averages for US Research Universities
hist(usresearchuniv$SAT_AVG_ALL, main = "Histogram of SAT Averages for US Research Universities (AY2010-
```



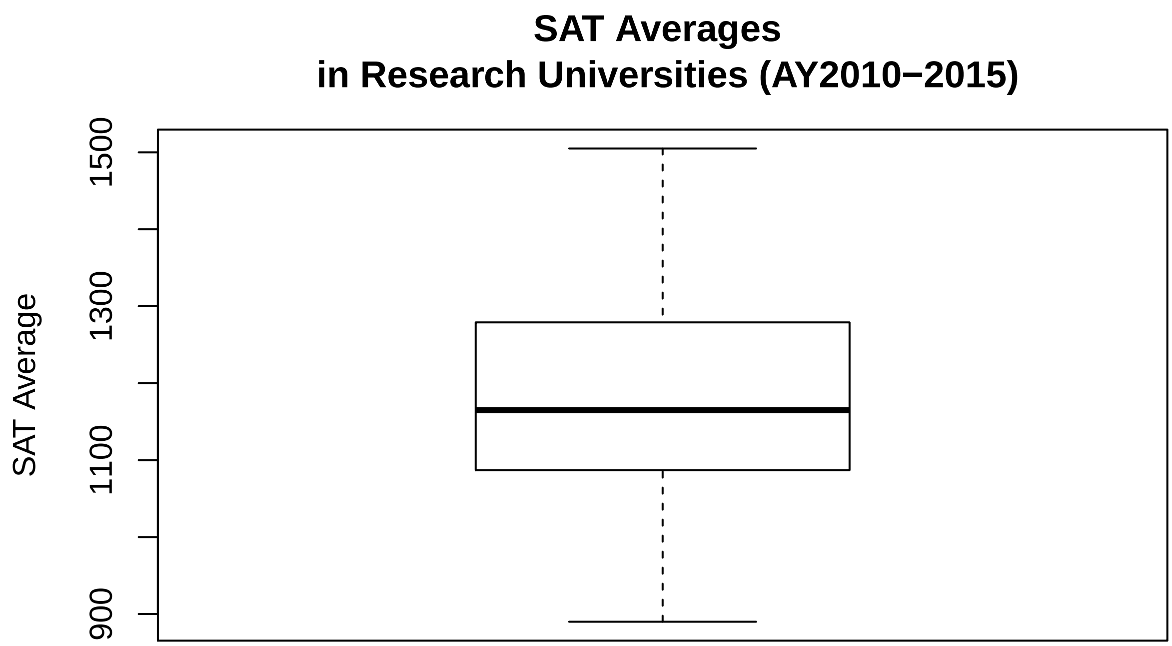

```
# Histogram of Admission Rates for US Research Universities
hist(usresearchuniv$ADM_RATE_ALL, main = "Histogram of Admission Rates for Research Universities (AY201
```



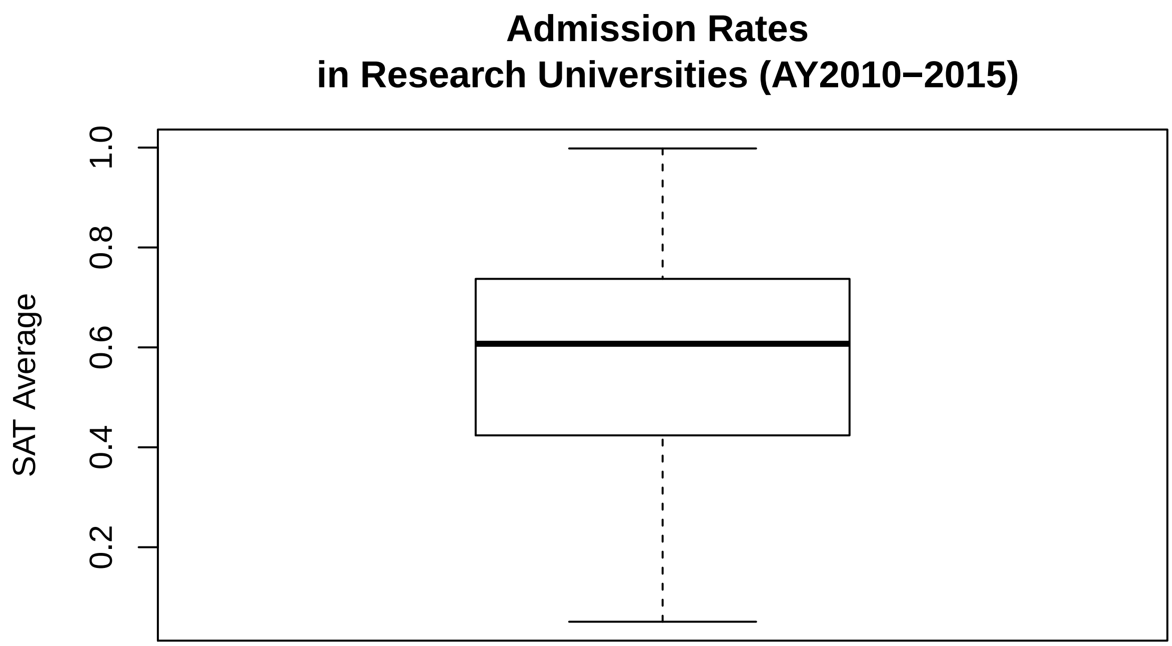

```r
# Histogram of Women in US Research Universities
hist(usresearchuniv$UGDS_WOMEN, main = "Histogram of Women in Research Universities (AY2010-2015)", xlab
```



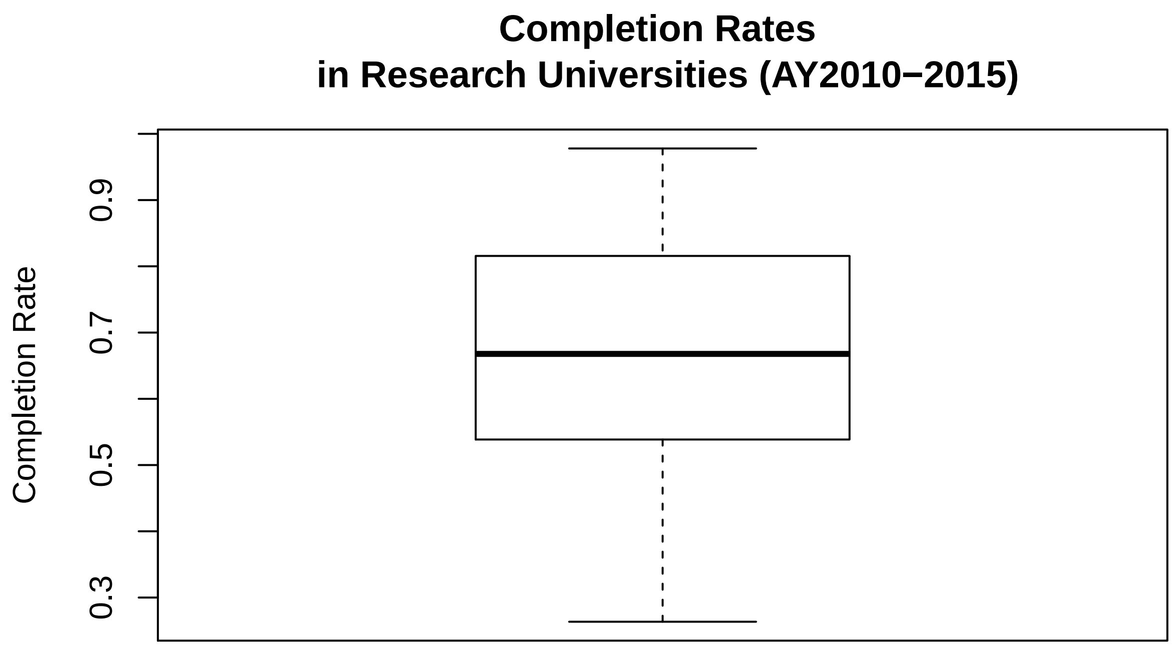

```r
#Boxplot of SAT Average in all US Research Universities
boxplot(usresearchuniv$SAT_AVG_ALL, main = "SAT Averages \n in Research Universities (AY2010-2015)", yla
```



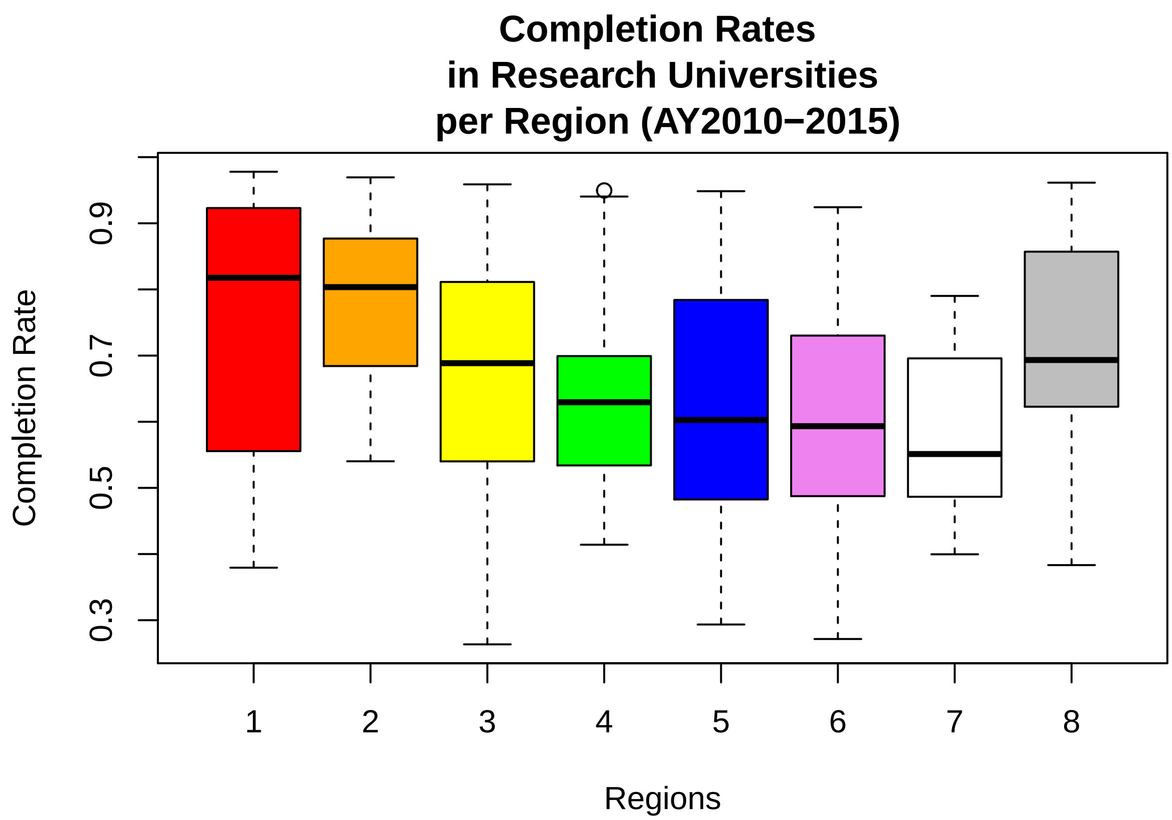

```
#Boxplot of admission rates in all US Research Universities
boxplot(usresearchuniv$ADM_RATE_ALL, main = "Admission Rates \n in Research Universities (AY2010-2015)"
```



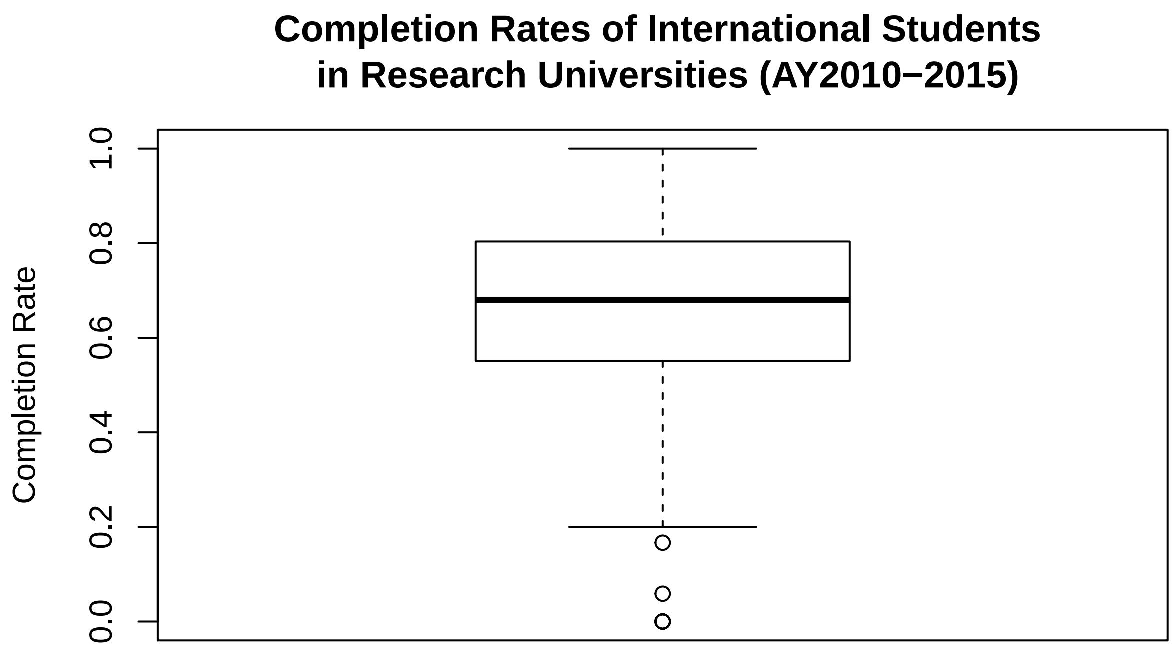

```r
#Boxplot of Completion Rates in all US Research Universities
boxplot(usresearchuniv$C150_4, main = "Completion Rates \n in Research Universities (AY2010-2015)", ylab
```



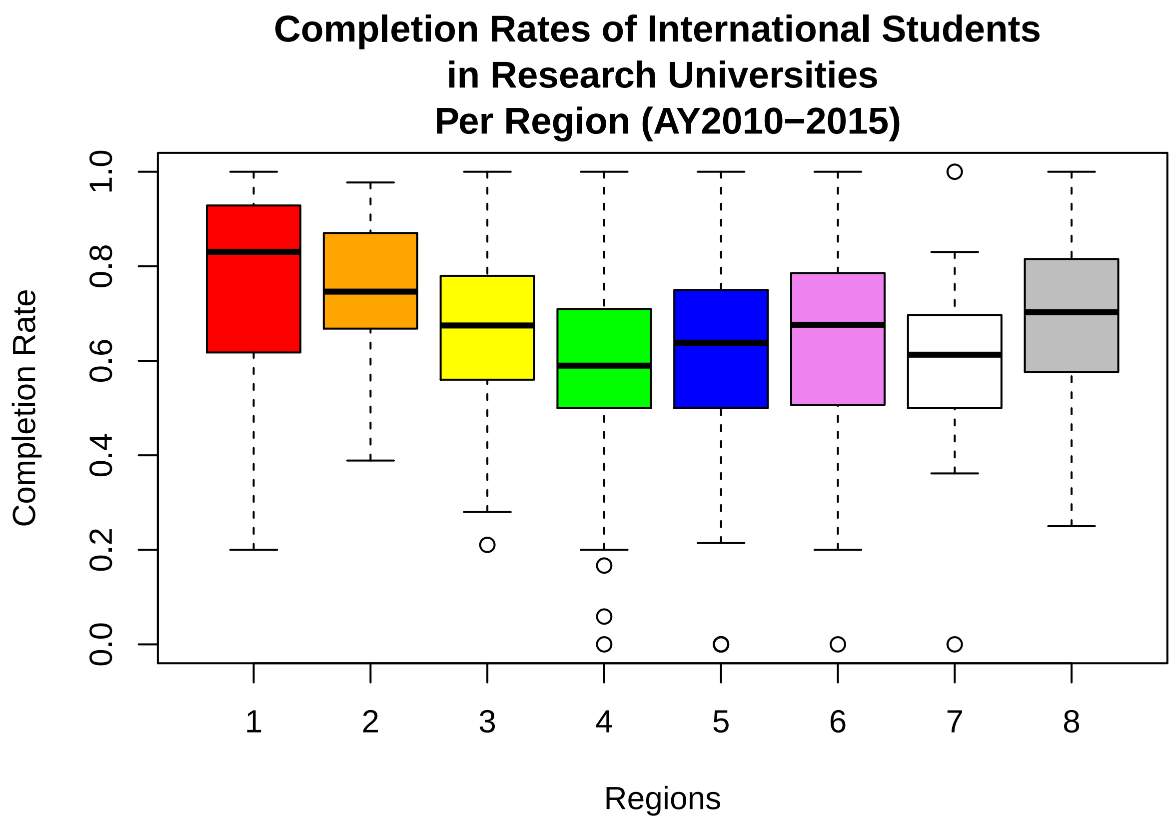

```
# Boxplot of Completion Rates per Region in US Research Universities
boxplot(C150_4 ~ REGION, usresearchuniv, main = "Completion Rates \n in Research Universities \n per Reg
```




```
#Boxplot of Completion Rates of International Students in all US Research Universities
boxplot(usresearchuniv$C150_4_NRA, main = "Completion Rates of International Students \n in Research Uni
```




```
# Boxplot of Completion Rates of International Students per Region in US Research Universities
boxplot(C150_4_NRA ~ REGION, usresearchuniv, main = "Completion Rates of International Students \n in Re
```

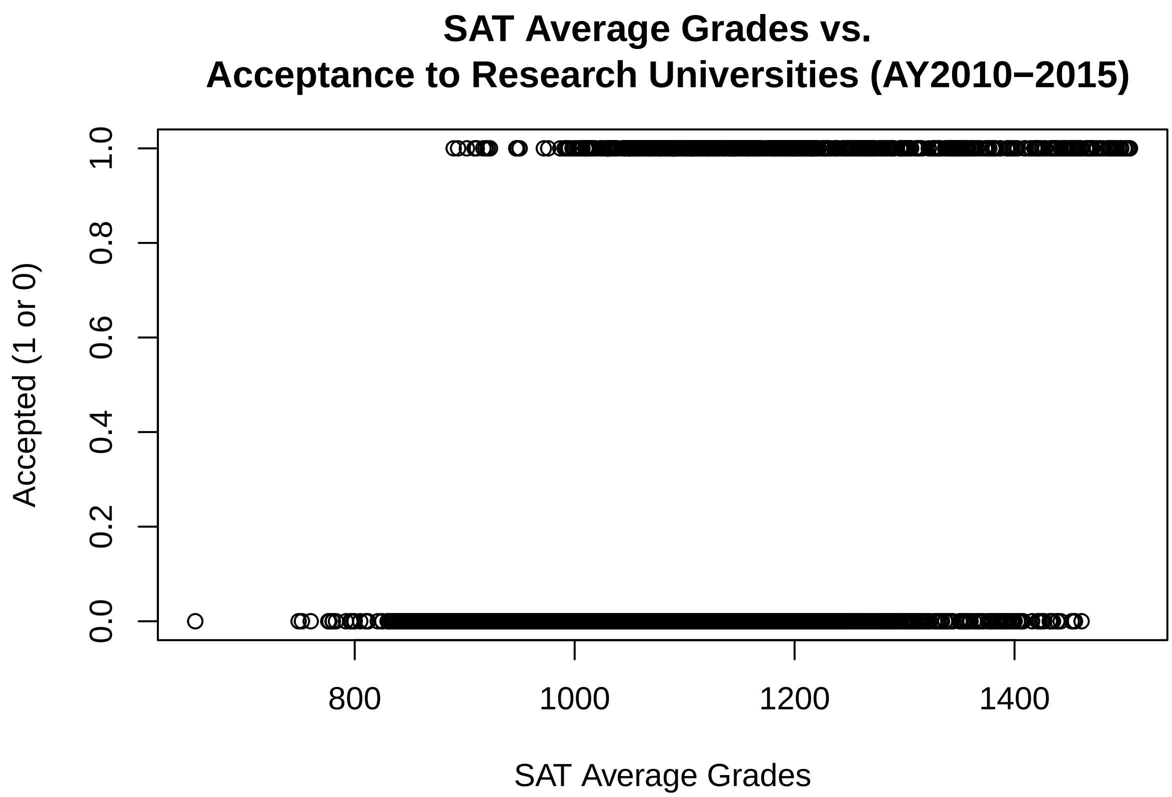



```
nrow(usresearchuniv[usresearchuniv$C150_4_NRA < 0.2,])
```

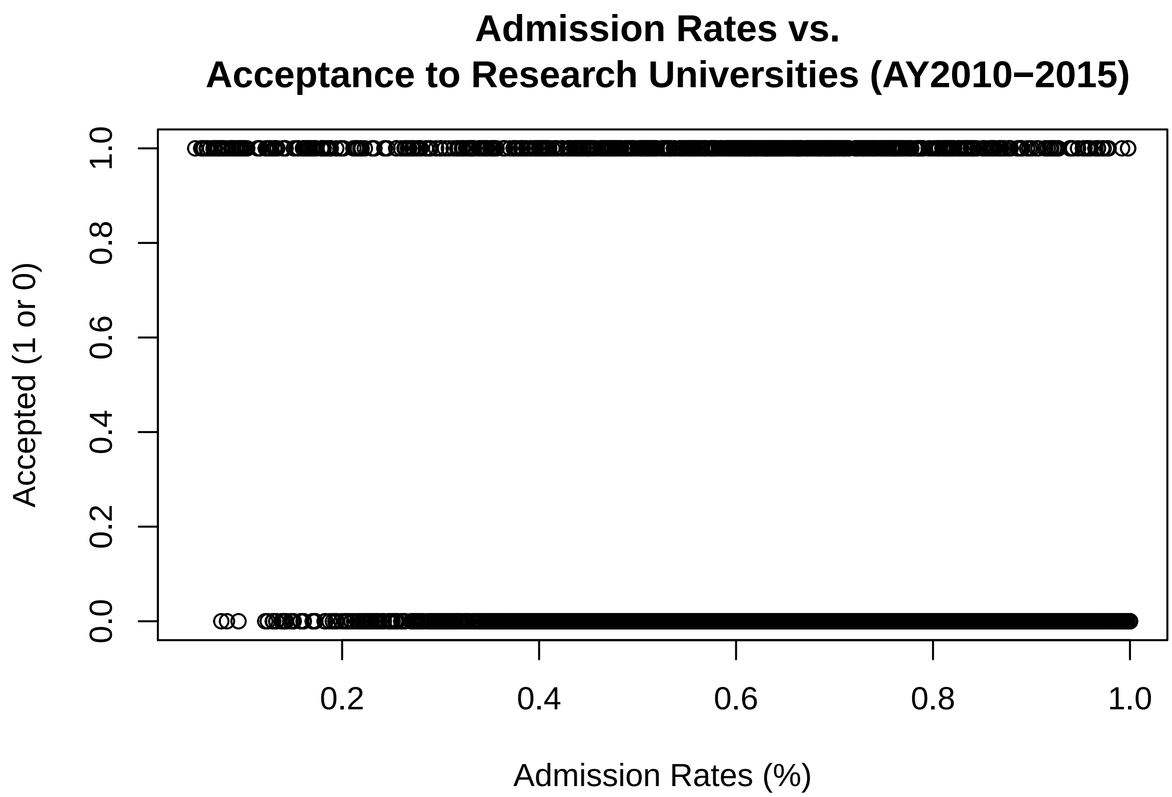
```
## [1] 9
```

Correlations

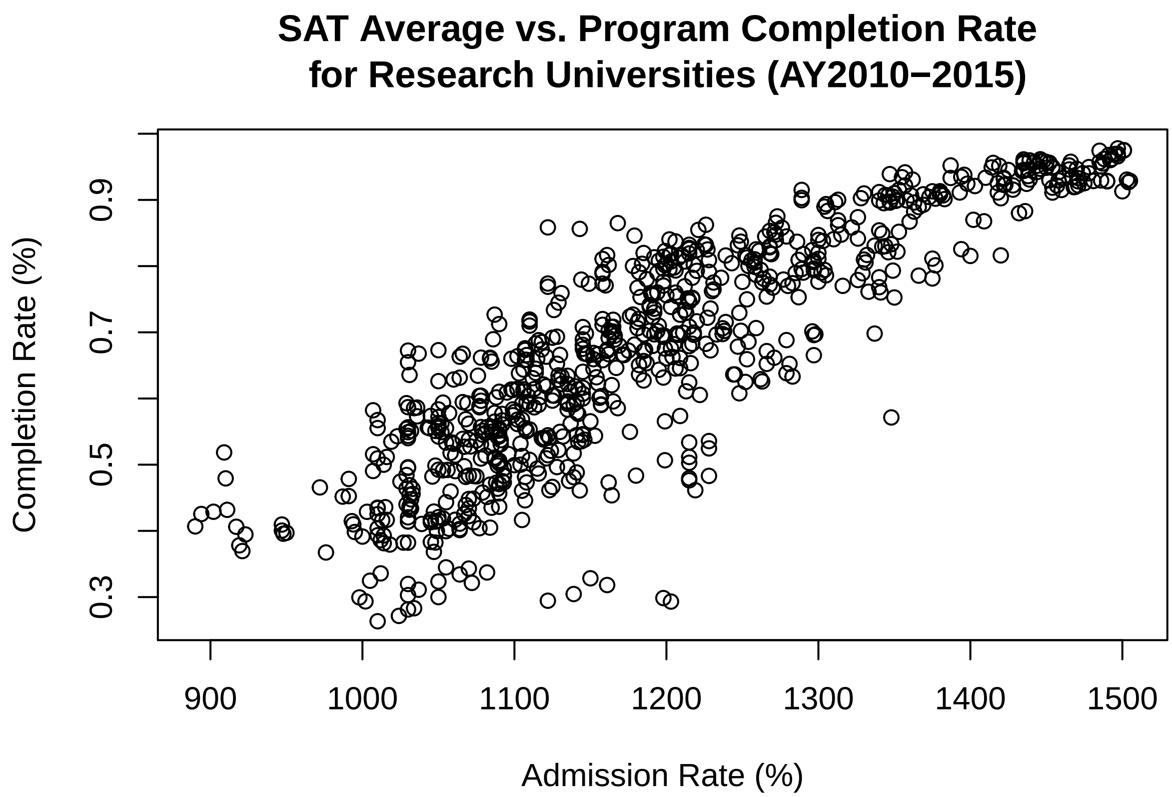
```
#Correlation between the SAT grades and the acceptance for the research universities
plot(usunivfilter$SAT_AVG_ALL, usunivfilter$ACCEPTED, main="SAT Average Grades vs. \n Acceptance to Rese
```




```
#Correlation between the admission rates and the acceptance for the research universities
plot(usunivfilter$ADM_RATE_ALL, usunivfilter$ACCEPTED, main="Admission Rates vs. \n Acceptance to Resea
```




```
#Correlation between admission rate for research universities and program completion rate
plot(usresearchuniv$SAT_AVG_ALL, usresearchuniv$C150_4, main="SAT Average vs. Program Completion Rate \n
```

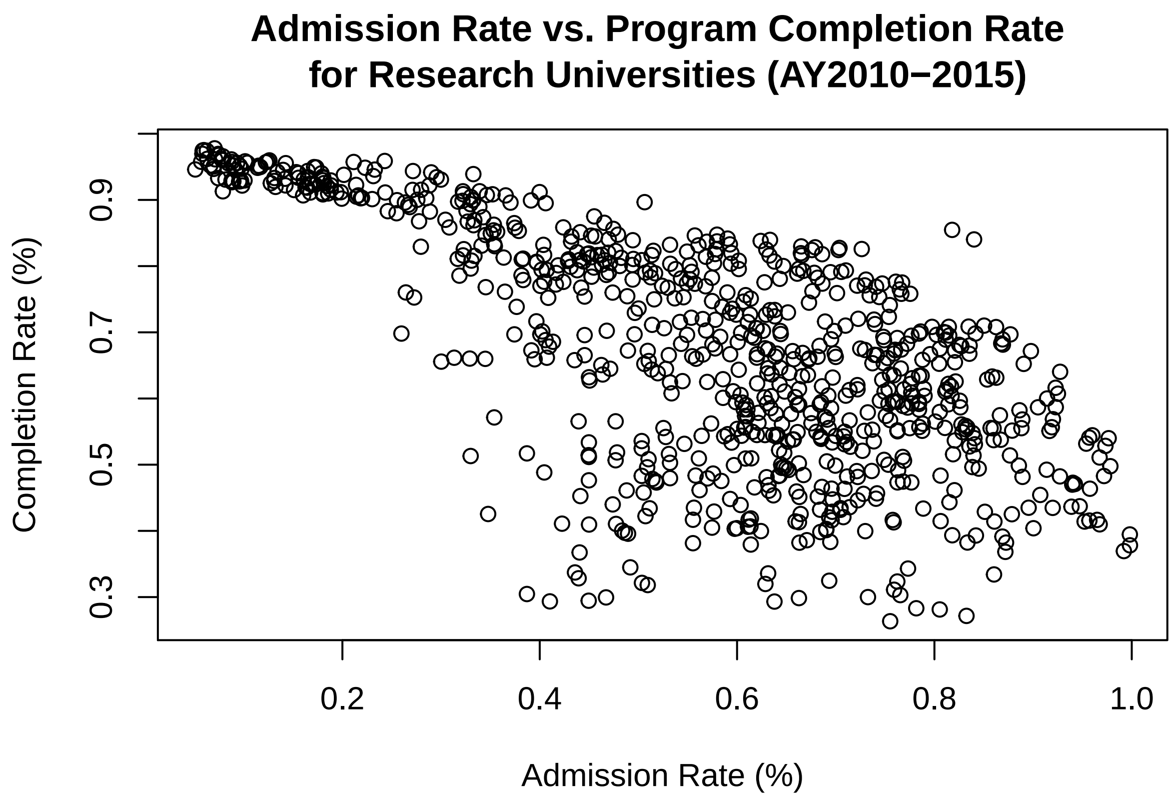



```r
#Correlation coefficient between admission rate and completion rate
cor(usresearchuniv$SAT_AVG_ALL, usresearchuniv$C150_4, method = "pearson")
```

```
## [1] 0.8702261
```

This means that there is a strong positive correlation between the SAT average scores and the completion rate for all students.

```r
#Correlation between admission rate for research universities and program completion rate
plot(usresearchuniv$ADM_RATE_ALL, usresearchuniv$C150_4, main="Admission Rate vs. Program Completion Ra
```

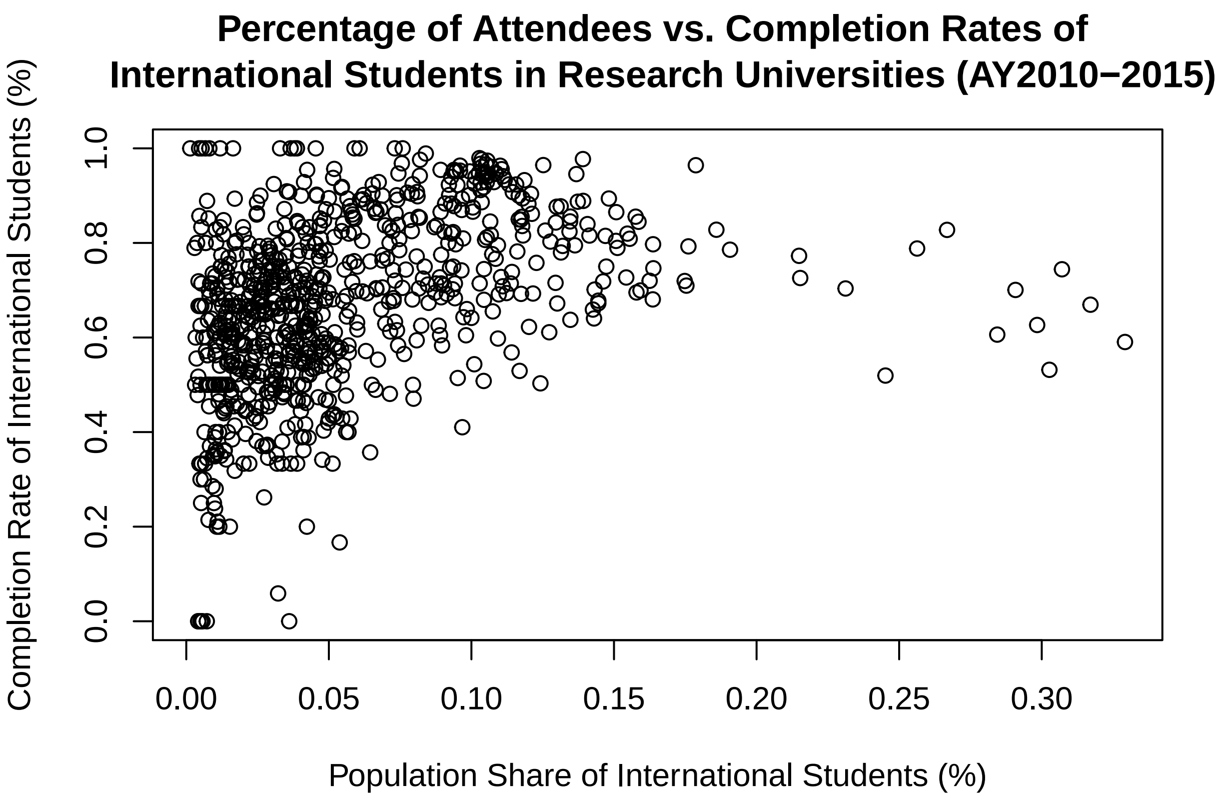



```r
#Correlation coefficient between admission rate and completion rate
cor(usresearchuniv$ADM_RATE_ALL, usresearchuniv$C150_4, method = "pearson")
```

```
## [1] -0.6825525
```

This means that there is a strong negative correlation between the admission rates and the completion rates for the research universities.

```r
#Correlation between attendees and completion rate of non-resident aliens (International Students)
plot(usresearchuniv$UGDS_NRA, usresearchuniv$C150_4_NRA, main="Percentage of Attendees vs. Completion Ra
```

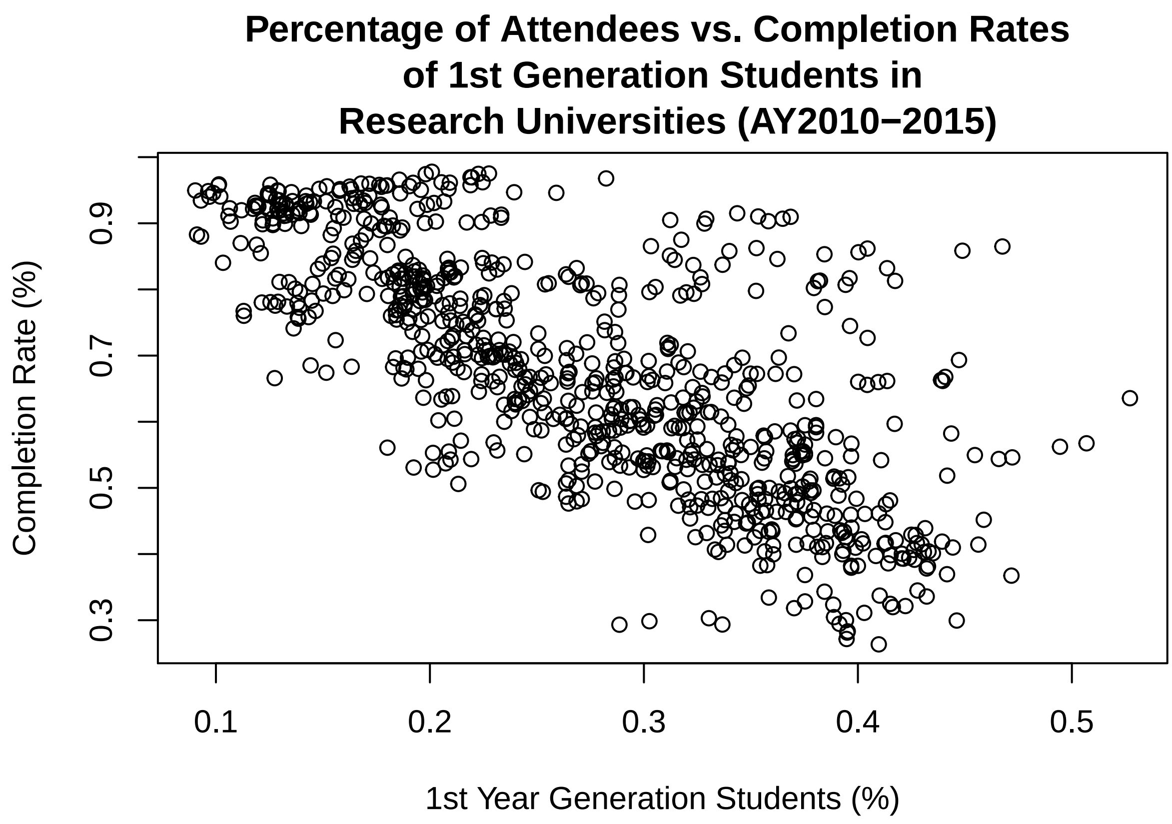



```r
#Correlation coefficient between admission rate and completion rate of international students
cor(usresearchuniv$UGDS_NRA, usresearchuniv$C150_4_NRA, method = "pearson")
```

```
## [1] 0.370641
```

This means that there is a weak positive correlation between international student population and their completion rate.

```r
#Correlation between attendees and completion rate of 1st Generation students in Research Universities
plot(usresearchuniv$PAR_ED_PCT_1STGEN, usresearchuniv$C150_4, main="Percentage of Attendees vs. Complet
```




```
#Correlation coefficient between admission rate and completion rate of 1st Generation students
cor(usresearchuniv$PAR_ED_PCT_1STGEN, usresearchuniv$C150_4, method = "pearson")
```

```
## [1] -0.7419477
```

This means that there is a strong negative correlation between 1st generation students and completion rates in research universities.

U.S. Research University Acceptance Model

In this report section, we are going to create a formula on getting an acceptance to a US Research University based on the College Scorecard statistics. We will try different methods of regression, and find the best regression technique from the following sources.

We will also consider another formula based on an international student taking up science degree/major.

```
# create a training and test model using a 75%/25% from the data set
rm_train <- sample(nrow(usunivfilter), floor(nrow(usunivfilter)*0.75))
univ_train <- usunivfilter[rm_train,]
univ_test <- usunivfilter[-rm_train,]

# create a generic formula for the US research university acceptance model for International Students b
formula_ISAcceptance <- formula(ACCEPTED ~ REGION + ADM_RATE_ALL + SAT_AVG_ALL + UGDS_NRA + COSTT4_A + I
```

We will do a generalized logistic regression formula.

```
# create a logistic regression
fit1 <- glm(formula_ISAcceptance, data = usunivfilter, family  = binomial())
summary(fit1)
```

```
##
## Call:
## glm(formula = formula_ISAcceptance, family = binomial(), data = usunivfilter)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2091  -0.5400  -0.2922  -0.1192   2.7993
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.478e+01  1.029e+00 -14.362  < 2e-16 ***
## REGION        1.246e-01  2.550e-02   4.886 1.03e-06 ***
## ADM_RATE_ALL  7.036e-01  3.297e-01   2.134   0.0328 *
## SAT_AVG_ALL   1.462e-02  7.312e-04  19.999  < 2e-16 ***
## UGDS_NRA      6.637e+00  1.147e+00   5.784 7.28e-09 ***
## COSTT4_A     -9.181e-05  5.441e-06 -16.872  < 2e-16 ***
## PCTFLOAN     -7.486e-01  4.247e-01  -1.763   0.0779 .
## UGDS_WOMEN   -1.995e+00  4.619e-01  -4.318 1.57e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4153.3  on 4246  degrees of freedom
## Residual deviance: 2838.4  on 4239  degrees of freedom
## AIC: 2854.4
##
## Number of Fisher Scoring iterations: 6
```

Based on the logistic regression, the formula will be

$$\frac{1}{1+e^{-x}}$$

where

$x = -14.8 + 0.125 REGION + 0.704 ADM\_RATE\_ALL + 0.0146 SAT\_AVG\_ALL + 6.64 UGDS\_NRA - 0.0000918 COSTT4$

.

We will test this regression with some data types.

```
# this will not accept the person because of the SAT average
df_accept <- data.frame(REGION = 5, SAT_AVG_ALL = 900, ADM_RATE_ALL = .55, UGDS_NRA=.010, COSTT4_A = 20
predict(fit1, type = "response", newdata = df_accept)
```

```
##          1
## 0.03356807
```

```r
# this will accept because of the SAT average and the cost
df_accept2 <- data.frame(REGION = 3, SAT_AVG_ALL = 1350, ADM_RATE_ALL = .35, UGDS_NRA=.25, COSTT4_A = 25
predict(fit1, type = "response", newdata = df_accept2)
```

```
##         1
## 0.9667774
```

Now, we will do some testing of performance with the logistic regression. Since we have split the dataset into training and testing set, we will see how the performance will be done.

```r
# do a logistic regression model based on this
glm_ISAcceptance <- glm(formula_ISAcceptance, data = univ_train, family = binomial())
summary(glm_ISAcceptance)
```

```
##
## Call:
## glm(formula = formula_ISAcceptance, family = binomial(), data = univ_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3403  -0.5322  -0.2863  -0.1158   2.8098
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.508e+01  1.187e+00 -12.710  < 2e-16 ***
## REGION        1.315e-01  2.926e-02   4.493 7.03e-06 ***
## ADM_RATE_ALL  6.881e-01  3.836e-01   1.794  0.07286 .
## SAT_AVG_ALL   1.469e-02  8.362e-04  17.573  < 2e-16 ***
## UGDS_NRA      7.948e+00  1.326e+00   5.993 2.06e-09 ***
## COSTT4_A     -9.166e-05  6.248e-06 -14.670  < 2e-16 ***
## PCTFLOAN     -7.710e-01  4.839e-01  -1.593  0.11109
## UGDS_WOMEN   -1.744e+00  5.353e-01  -3.258  0.00112 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3128.5  on 3184  degrees of freedom
## Residual deviance: 2108.8  on 3177  degrees of freedom
## AIC: 2124.8
##
## Number of Fisher Scoring iterations: 6
```

```r
# do the first testing with the prediction model
accepted_ind <- predict(glm_ISAcceptance, type="response", newdata = univ_test)
pred1 <- prediction(accepted_ind, univ_test$ACCEPTED)

# create the confusion matrix and accuracy for this prediction model
c1 <- confusionMatrix(as.integer(accepted_ind > 0.5), univ_test$ACCEPTED)
c1$table
```

```
##           Reference
```

```
## Prediction   0   1
##          0 830 121
##          1  33  78
```

```r
#Accuracy of the logistic regression model
c1$overall['Accuracy']
```

```
##  Accuracy
## 0.8549906
```

```r
#Precision of the logistic regression model
c1$byClass['Neg Pred Value']
```

```
## Neg Pred Value
##      0.7027027
```

```r
#Recall of the logistic regression model
c1$byClass['Specificity']
```

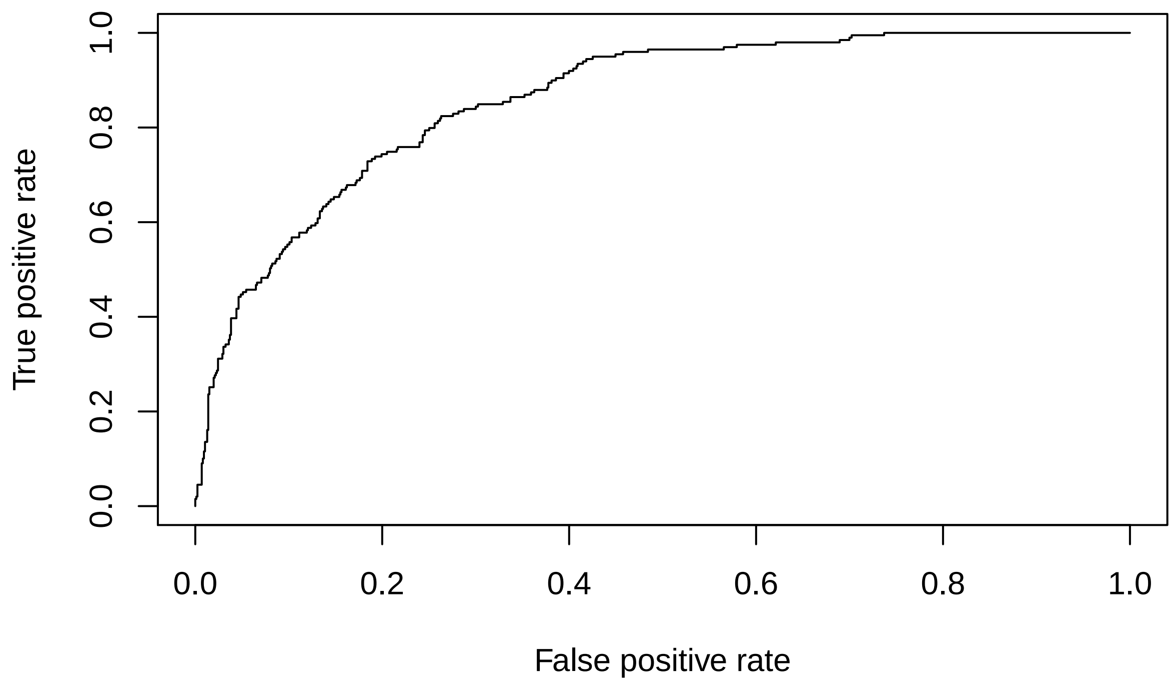
```
## Specificity
##   0.3919598
```

Accuracy shows the correct value. But in precision and recall, it is using "Neg Pred Value" and "Specificity" respectively. It should have been "Pos Pred Value" and "Sensitivity", as defined before. However, I manually calculated for the precision and recall for these values, and they are displayed correctly as it should be.

Precision: TP / (FP + TP) Recall: TP / (FN + TP)

As I show the precision and recall, it would be done the same thing, and verified manually that these are the correct percentages.

```r
# show the curve on the performance
perf1 <- performance(pred1, "tpr", "fpr")
plot(perf1, lty = 1)
```




```
# Now we check on what acceptable ways we could do for regression
# doing single decision tree
model_dtree1 <- rpart(formula_ISAcceptance, method="anova",data = univ_train)
pred_dtree1 <- predict(model_dtree1, newdata = univ_test)
accu1 = abs(pred_dtree1 - univ_test$ACCEPTED) < 0.5
frac1 = sum(accu1)/length(accu1)
print(frac1)
```

```
## [1] 0.8700565
```

```
# doing random forest
model_forest1 <- randomForest(formula_ISAcceptance, data = univ_train)
```

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values. Are you sure you want to do regression?
```

```
pred_forest1 <- predict(model_forest1, newdata = univ_test)
accu2 <- abs(pred_forest1 - univ_test$ACCEPTED) < 0.5
frac2 <- sum(accu2)/length(accu2)
print(frac2)
```

```
## [1] 0.9190207
```

```r
# doing support vector machine
model_svm1 <- svm(formula_ISAcceptance, data = univ_train)
pred_svm1 <- predict(model_svm1, newdata = univ_test)
accu3 <- abs(pred_svm1 - univ_test$ACCEPTED) < 0.5
frac3 <- sum(accu3)/length(accu3)
print(frac3)
```

```
## [1] 0.8691149
```

```r
# doing simple tree
model_tree1 <- tree(formula_ISAcceptance, data = univ_train)
pred_tree1 <- predict(model_tree1, newdata = univ_test)
accu4 <- abs(pred_tree1 - univ_test$ACCEPTED) < 0.5
frac4 <- sum(accu4)/length(accu4)
print(frac4)
```

```
## [1] 0.8700565
```

```r
# doing conditional inference tree
model_party1 <- ctree(formula_ISAcceptance, data = univ_train)
pred_party1 <- predict(model_party1, newdata = univ_test)
accu5 <- abs(pred_party1 - univ_test$ACCEPTED) < 0.5
frac5 <- sum(accu5)/length(accu5)
print(frac5)
```

```
## [1] 0.86629
```

Based on the run, random forest is the best regression method to use in this model.

Next, another formula is created. This is an acceptance model for an international student that wants to take up Science degree/major

```r
# create a formula for the US research university acceptance model for International Students taking up
formula_ISSciAcceptance <- formula(ACCEPTED ~ REGION + ADM_RATE_ALL + SAT_AVG_ALL + PCIP11 + PCIP12 + P

# do a logistic regression model based on the formula created
glm_ISSciAcceptance <- glm(formula_ISSciAcceptance, data=univ_train,family=binomial())
summary(glm_ISSciAcceptance)
```

```
##
## Call:
## glm(formula = formula_ISSciAcceptance, family = binomial(), data = univ_train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.52753  -0.47116  -0.23872  -0.07852   3.07079
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.744e+01  1.420e+00 -12.286  < 2e-16 ***
## REGION        1.393e-01  3.170e-02   4.395 1.11e-05 ***
```

```
## ADM_RATE_ALL  1.065e+00  4.263e-01   2.497 0.012524 *
## SAT_AVG_ALL   1.521e-02  9.981e-04  15.244  < 2e-16 ***
## PCIP11        5.866e-01  1.970e+00   0.298 0.765854
## PCIP12       -5.158e+00  2.025e+01  -0.255 0.798975
## PCIP14        5.644e+00  7.778e-01   7.256 3.97e-13 ***
## PCIP15        3.281e-01  2.224e+00   0.148 0.882736
## PCIP24       -5.223e+00  1.204e+00  -4.339 1.43e-05 ***
## PCIP26        8.390e+00  1.790e+00   4.688 2.76e-06 ***
## PCIP27       -2.289e+01  6.896e+00  -3.320 0.000900 ***
## PCIP40       -3.562e+01  4.793e+00  -7.431 1.07e-13 ***
## PCIP45        7.950e+00  1.212e+00   6.561 5.33e-11 ***
## PCIP51        2.113e+00  6.072e-01   3.479 0.000502 ***
## PCIP52        8.241e-01  6.648e-01   1.240 0.215119
## UGDS_NRA      1.024e+01  1.467e+00   6.981 2.94e-12 ***
## UGDS_UNKN    -2.408e+00  1.598e+00  -1.507 0.131760
## COSTT4_A     -1.037e-04  7.176e-06 -14.456  < 2e-16 ***
## PCTFLOAN     -7.394e-01  5.598e-01  -1.321 0.186586
## UGDS_WOMEN    2.602e-01  7.851e-01   0.331 0.740328
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3128.5  on 3184  degrees of freedom
## Residual deviance: 1873.0  on 3165  degrees of freedom
## AIC: 1913
##
## Number of Fisher Scoring iterations: 6
```

```r
# do the testing with the prediction model
accepted_ind2 <- predict(glm_ISSciAcceptance, type="response", newdata = univ_test)
pred2 <- prediction(accepted_ind2, univ_test$ACCEPTED)

# prepare confusion matrix and accuracy to see the scores
c2 <- confusionMatrix(as.integer(accepted_ind2 > 0.5), univ_test$ACCEPTED)
c2$table
```

```
##           Reference
## Prediction   0   1
##          0 831 100
##          1  32  99
```

```r
c2$overall['Accuracy']
```

```
##  Accuracy
## 0.8757062
```

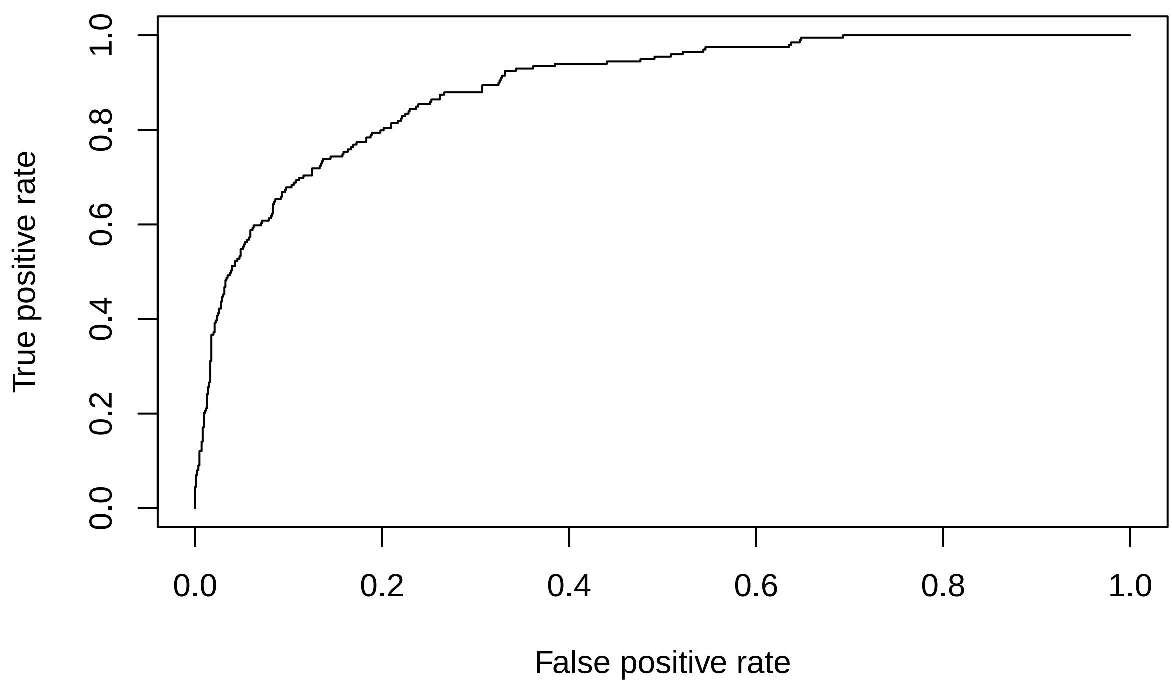
```r
#Precision of the logistic regression model
c2$byClass['Neg Pred Value']
```

```
## Neg Pred Value
##      0.7557252
```

```r
#Recall of the logistic regression model
c2$byClass['Specificity']
```

```
## Specificity
##   0.4974874
```

```r
# show the curve on the performance
perf2 <- performance(pred2,"tpr","fpr")
plot(perf2, lty = 1)
```




```r
# Now we check on what acceptable ways we could do for regression
# doing single decision tree
model_dtree2 <- rpart(formula_ISSciAcceptance, method="anova",data = univ_train)
pred_dtree2 <- predict(model_dtree2, newdata = univ_test)
accu6 <- abs(pred_dtree2 - univ_test$ACCEPTED) < 0.5
frac6 <- sum(accu6)/length(accu6)
print(frac6)
```

```
## [1] 0.9001883
```

```r
# doing random forest
model_forest2 <- randomForest(formula_ISSciAcceptance, data = univ_train)
```

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values. Are you sure you want to do regression?
```

```r
pred_forest2 <- predict(model_forest2, newdata = univ_test)
accu7 <- abs(pred_forest2 - univ_test$ACCEPTED) < 0.5
frac7 <- sum(accu7)/length(accu7)
print(frac7)
```

```
## [1] 0.952919
```

```r
# doing support vector machine
model_svm2 <- svm(formula_ISSciAcceptance, data = univ_train)
pred_svm2 <- predict(model_svm2, newdata = univ_test)
accu8 <- abs(pred_svm2 - univ_test$ACCEPTED) < 0.5
frac8 <- sum(accu8)/length(accu8)
print(frac8)
```

```
## [1] 0.9114878
```

```r
# doing simple tree
model_tree2 <- tree(formula_ISSciAcceptance, data = univ_train)
pred_tree2 <- predict(model_tree2, newdata = univ_test)
accu9 <- abs(pred_tree2 - univ_test$ACCEPTED) < 0.5
frac9 <- sum(accu9)/length(accu9)
print(frac9)
```

```
## [1] 0.8992467
```

```r
# doing conditional inference tree
model_party2 <- ctree(formula_ISSciAcceptance, data = univ_train)
pred_party2 <- predict(model_party2, newdata = univ_test)
accu10 <- abs(pred_party2 - univ_test$ACCEPTED) < 0.5
frac10 <- sum(accu10)/length(accu10)
print(frac10)
```

```
## [1] 0.8983051
```

Based on this, random forest is the best regression method to use.

In this project, I have selected a couple of variables that we could use in this model. However, we could use more than a few variables to get the optimal result.

With this in mind, feature selection is very essential, especially with datasets that have many variables for model selection. Although in this report, we have 1745 variables, and deduced it to 72 variables, we have to check which variables will be very useful in doing our research model.

In this portion, we will consider all variables, and use Boruta and RFE to use what variables we could use for doing a better outcome of the model.

Boruta is a package created was written by Miron B. Kursa and Witold R. Rudnicki to use an all relevant feature selection wrapper algorithm. According to their description, it "finds relevant features by comparing original attributes' importance with importance achievable at random, estimated using their permuted copies". (Source: https://cran.r-project.org/web/packages/Boruta/Boruta.pdf)

The Recursive Feature Elimination, or RFE, is a function in R's Caret package that uses the random forest algorithm to evaluate the attributes needed to be able to get an optimal result in the data that we have. (Source: http://machinelearningmastery.com/feature-selection-with-the-caret-r-package/)

Now, we will be doing some feature eliminations using Boruta and RFE.

```r
# First, we will create another copy of the dataset
usunivnoccbasic <- usunivfilter

# Next, we will change those that have "NA" to 0, since there is no data in it
usunivnoccbasic[usunivnoccbasic == "NA"] <- 0

# Next, we will choose rows that have complete cases
usunivnoccbasic <- usunivnoccbasic[complete.cases(usunivnoccbasic),]

# Now that we have the cleansed dataset, we will implement Boruta
boruta.train <- Boruta(ACCEPTED ~ .-CCBASIC2, data=usunivnoccbasic)
print(boruta.train)
```

```
## Boruta performed 99 iterations in 26.25063 secs.
##  60 attributes confirmed important: ADM_RATE, ADM_RATE_ALL,
## C150_4, C150_4_AIAN, C150_4_ASIAN and 55 more.
##  7 attributes confirmed unimportant: C150_4_NHPI, PCIP12, PCIP25,
## PCIP29, PCIP46 and 2 more.
##  3 tentative attributes left: C150_4_2MOR, PCIP10, PCIP22.
```

```r
getSelectedAttributes(boruta.train)
```

```
##  [1] "REGION"            "ADM_RATE"          "ADM_RATE_ALL"
##  [4] "SAT_AVG_ALL"       "PCIP01"            "PCIP03"
##  [7] "PCIP04"            "PCIP05"            "PCIP09"
## [10] "PCIP11"            "PCIP13"            "PCIP14"
## [13] "PCIP15"            "PCIP16"            "PCIP19"
## [16] "PCIP23"            "PCIP24"            "PCIP26"
## [19] "PCIP27"            "PCIP30"            "PCIP31"
## [22] "PCIP38"            "PCIP39"            "PCIP40"
## [25] "PCIP41"            "PCIP42"            "PCIP43"
## [28] "PCIP44"            "PCIP45"            "PCIP49"
## [31] "PCIP50"            "PCIP51"            "PCIP52"
## [34] "PCIP54"            "UGDS_WHITE"        "UGDS_BLACK"
## [37] "UGDS_HISP"         "UGDS_ASIAN"        "UGDS_AIAN"
## [40] "UGDS_NHPI"         "UGDS_2MOR"         "UGDS_NRA"
## [43] "UGDS_UNKN"         "PPTUG_EF"          "COSTT4_A"
## [46] "TUITIONFEE_IN"     "TUITIONFEE_OUT"    "C150_4"
## [49] "C150_4_WHITE"      "C150_4_BLACK"      "C150_4_HISP"
## [52] "C150_4_ASIAN"      "C150_4_AIAN"       "C150_4_NRA"
## [55] "C150_4_UNKN"       "RET_FT4"           "PCTFLOAN"
## [58] "PAR_ED_PCT_1STGEN" "UGDS_MEN"          "UGDS_WOMEN"
```

```r
# We will print the stats of the variables that would be accepted or not
stats <- attStats(boruta.train)
print(stats)
```

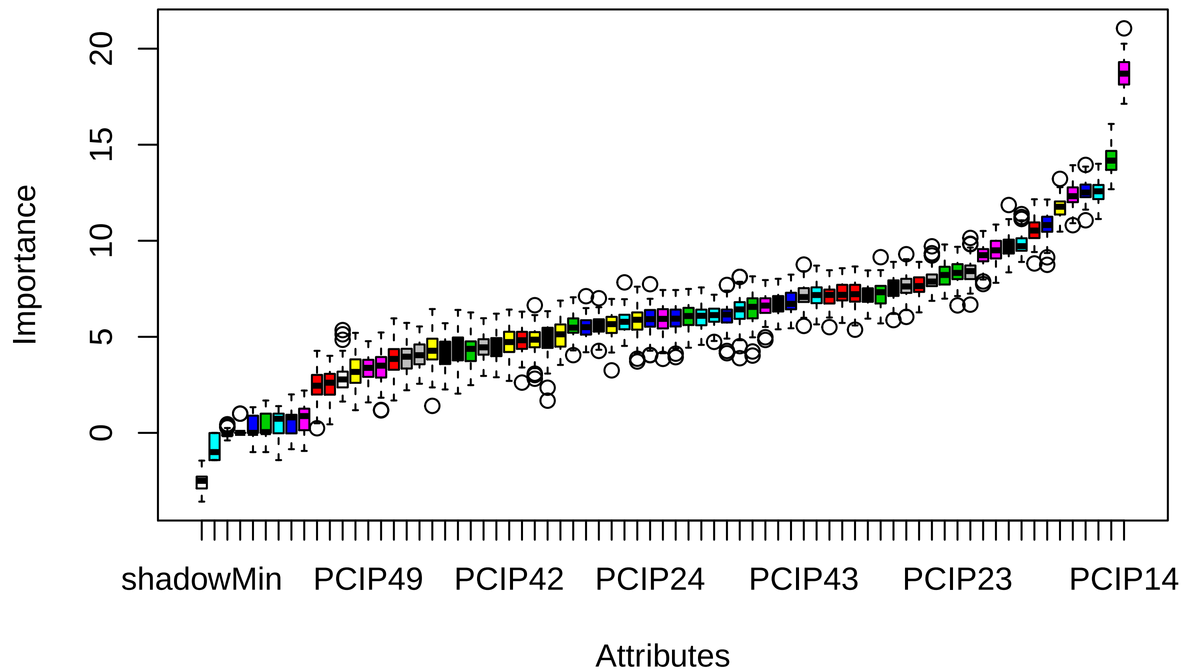
```
##                       meanImp   medianImp     minImp    maxImp   normHits
## REGION              5.5531479  5.65541262  4.2668951  7.007936 1.00000000
## ADM_RATE            7.2055049  7.18535776  5.7220003  8.570573 1.00000000
## ADM_RATE_ALL        7.2316599  7.32050357  5.6975400  9.145504 1.00000000
## SAT_AVG_ALL        12.5821102 12.51006784 11.0660764 13.948851 1.00000000
```

```
## PCIP01             6.1170645  6.13090594  4.7368300  7.193177 0.98989899
## PCIP03             6.6094378  6.62548190  4.8373601  7.954159 1.00000000
## PCIP04            11.7240810 11.77150784 10.4747367 13.221292 1.00000000
## PCIP05             8.3971911  8.42122854  6.6761523 10.145963 1.00000000
## PCIP09             4.8732652  4.90130338  1.6771974  6.334498 0.96969697
## PCIP10             2.5267989  2.61283219  0.4444349  4.012305 0.41414141
## PCIP11             6.4771426  6.55730806  4.0227266  8.149189 1.00000000
## PCIP12             0.5734679  0.80423019 -0.8494920  2.001767 0.01010101
## PCIP13             6.0278530  6.09595794  4.5798719  7.571188 1.00000000
## PCIP14            18.7074161 18.69853351 17.1310010 21.055979 1.00000000
## PCIP15             4.8460772  4.85235599  2.8178853  6.645656 0.95959596
## PCIP16             7.6534548  7.62147254  6.0375186  9.298945 1.00000000
## PCIP19             7.5229293  7.58319677  5.8706100  8.901044 1.00000000
## PCIP22             2.4394106  2.45620610  0.2399080  4.271884 0.38383838
## PCIP23             8.3437666  8.32606596  6.6350217  9.685576 1.00000000
## PCIP24             5.8762895  5.92226996  4.0415971  7.739864 1.00000000
## PCIP25            -0.8297707 -1.00100150 -1.4167771  0.000000 0.00000000
## PCIP26             5.8957724  5.93467473  3.8516397  7.437130 1.00000000
## PCIP27             5.0783790  5.12927297  3.5359865  6.887261 0.95959596
## PCIP29             0.1540002  0.00000000  0.0000000  1.001002 0.00000000
## PCIP30             4.1708356  4.27991026  2.2596071  5.707220 0.90909091
## PCIP31             4.8605202  4.81960321  2.6174955  6.312517 0.93939394
## PCIP38             4.2936303  4.37221601  2.4850088  6.268402 0.90909091
## PCIP39             5.4621037  5.50107030  4.1964203  7.115478 0.98989899
## PCIP40             5.8138261  5.78148398  4.5275635  7.833606 1.00000000
## PCIP41             3.3901914  3.49868622  1.1699086  5.225245 0.70707071
## PCIP42             4.6926004  4.72699041  2.7050698  6.418733 0.94949495
## PCIP43             7.1783737  7.07078366  5.5679934  8.761482 1.00000000
## PCIP44             4.4849767  4.48462918  2.8938664  6.207151 0.94949495
## PCIP45             7.6592305  7.64184414  6.2718828  8.900147 1.00000000
## PCIP46             0.3838771  0.04066599 -1.0010015  1.684197 0.00000000
## PCIP47             0.2279959  0.00000000 -1.0010015  1.336102 0.00000000
## PCIP48             0.3618668  0.73603447 -1.4170446  1.393093 0.00000000
## PCIP49             3.3384159  3.38569049  1.5864445  4.773530 0.69696970
## PCIP50             5.8159304  5.88996495  3.7140161  7.610976 0.98989899
## PCIP51             4.0552888  4.04338586  2.5566253  5.538055 0.88888889
## PCIP52             9.6883139  9.60777682  8.3555041 11.859644 1.00000000
## PCIP54             3.8286889  3.84958631  1.6869158  5.960554 0.85858586
## UGDS_WHITE         8.2243664  8.21962912  6.9876629  9.807068 1.00000000
## UGDS_BLACK        10.8090976 10.81822647  8.7465932 12.149841 1.00000000
## UGDS_HISP          6.3777429  6.42714810  3.8868634  8.119792 1.00000000
## UGDS_ASIAN         9.2552287  9.25077575  7.7457772 10.510764 1.00000000
## UGDS_AIAN          4.3302542  4.27557035  1.4061608  6.445591 0.90909091
## UGDS_NHPI          3.9004889  3.96546887  2.2170966  5.724992 0.85858586
## UGDS_2MOR          4.3270918  4.30971776  2.0429535  6.399436 0.90909091
## UGDS_NRA           7.2082551  7.21803280  5.3696143  8.661530 1.00000000
## UGDS_UNKN          6.0486315  6.08313690  4.4338040  7.491227 1.00000000
## PPTUG_EF           6.8141684  6.71340427  5.4444710  8.238734 1.00000000
## COSTT4_A           9.8213054  9.71768997  8.8985397 11.387079 1.00000000
## TUITIONFEE_IN      9.5157518  9.50141124  7.8107556 10.844340 1.00000000
## TUITIONFEE_OUT     5.6296827  5.66015201  3.2538503  6.971818 0.98989899
## C150_4             7.9237279  7.87766700  6.8721949  9.710931 1.00000000
## C150_4_WHITE       6.7106461  6.68934206  5.3903378  8.021079 1.00000000
## C150_4_BLACK       7.0774799  7.18173356  5.5040478  8.464933 1.00000000
```

```
## C150_4_HISP        5.5415321  5.47682242  4.0467530  7.057760 0.98989899
## C150_4_ASIAN       6.0905100  6.16211112  4.1461749  7.699960 0.98989899
## C150_4_AIAN        7.1692567  7.17478073  5.6466760  8.707349 1.00000000
## C150_4_NHPI        0.7936726  0.87556953 -0.9411222  2.204637 0.00000000
## C150_4_2MOR        3.2197431  3.17401567  1.1812294  5.205159 0.62626263
## C150_4_NRA         4.4865578  4.45174107  2.9641970  5.968019 0.93939394
## C150_4_UNKN        7.1544577  7.23198783  5.9472652  8.456507 1.00000000
## RET_FT4           10.5378981 10.53211635  8.8178658 12.160164 1.00000000
## PCTFLOAN          14.1974875 14.17234365 12.6826450 16.081587 1.00000000
## PAR_ED_PCT_1STGEN  5.9243467  5.94548362  3.9407762  7.435837 1.00000000
## UGDS_MEN          12.5477284 12.57507989 11.1325292 14.006876 1.00000000
## UGDS_WOMEN        12.3760193 12.32031704 10.8102542 13.932341 1.00000000
##                    decision
## REGION            Confirmed
## ADM_RATE          Confirmed
## ADM_RATE_ALL      Confirmed
## SAT_AVG_ALL       Confirmed
## PCIP01            Confirmed
## PCIP03            Confirmed
## PCIP04            Confirmed
## PCIP05            Confirmed
## PCIP09            Confirmed
## PCIP10            Tentative
## PCIP11            Confirmed
## PCIP12             Rejected
## PCIP13            Confirmed
## PCIP14            Confirmed
## PCIP15            Confirmed
## PCIP16            Confirmed
## PCIP19            Confirmed
## PCIP22            Tentative
## PCIP23            Confirmed
## PCIP24            Confirmed
## PCIP25             Rejected
## PCIP26            Confirmed
## PCIP27            Confirmed
## PCIP29             Rejected
## PCIP30            Confirmed
## PCIP31            Confirmed
## PCIP38            Confirmed
## PCIP39            Confirmed
## PCIP40            Confirmed
## PCIP41            Confirmed
## PCIP42            Confirmed
## PCIP43            Confirmed
## PCIP44            Confirmed
## PCIP45            Confirmed
## PCIP46             Rejected
## PCIP47             Rejected
## PCIP48             Rejected
## PCIP49            Confirmed
## PCIP50            Confirmed
## PCIP51            Confirmed
## PCIP52            Confirmed
```

```
## PCIP54              Confirmed
## UGDS_WHITE          Confirmed
## UGDS_BLACK          Confirmed
## UGDS_HISP           Confirmed
## UGDS_ASIAN          Confirmed
## UGDS_AIAN           Confirmed
## UGDS_NHPI           Confirmed
## UGDS_2MOR           Confirmed
## UGDS_NRA            Confirmed
## UGDS_UNKN           Confirmed
## PPTUG_EF            Confirmed
## COSTT4_A            Confirmed
## TUITIONFEE_IN       Confirmed
## TUITIONFEE_OUT      Confirmed
## C150_4              Confirmed
## C150_4_WHITE        Confirmed
## C150_4_BLACK        Confirmed
## C150_4_HISP         Confirmed
## C150_4_ASIAN        Confirmed
## C150_4_AIAN         Confirmed
## C150_4_NHPI          Rejected
## C150_4_2MOR         Tentative
## C150_4_NRA          Confirmed
## C150_4_UNKN         Confirmed
## RET_FT4             Confirmed
## PCTFLOAN            Confirmed
## PAR_ED_PCT_1STGEN Confirmed
## UGDS_MEN            Confirmed
## UGDS_WOMEN          Confirmed
```

```
# We will plot on the number of variables and its importance for Boruta
plot(boruta.train, type = c("g","o"), cex = 1.0, col = 1:70)
```




```
#Now, let us try RFE
rfe_control <- rfeControl(functions=rfFuncs, method="cv", number = 10)
rfe.train <- rfe(usunivnoccbasic[,1:70], usunivnoccbasic[,72], sizes = 1:70, rfeControl = rfe_control)
```

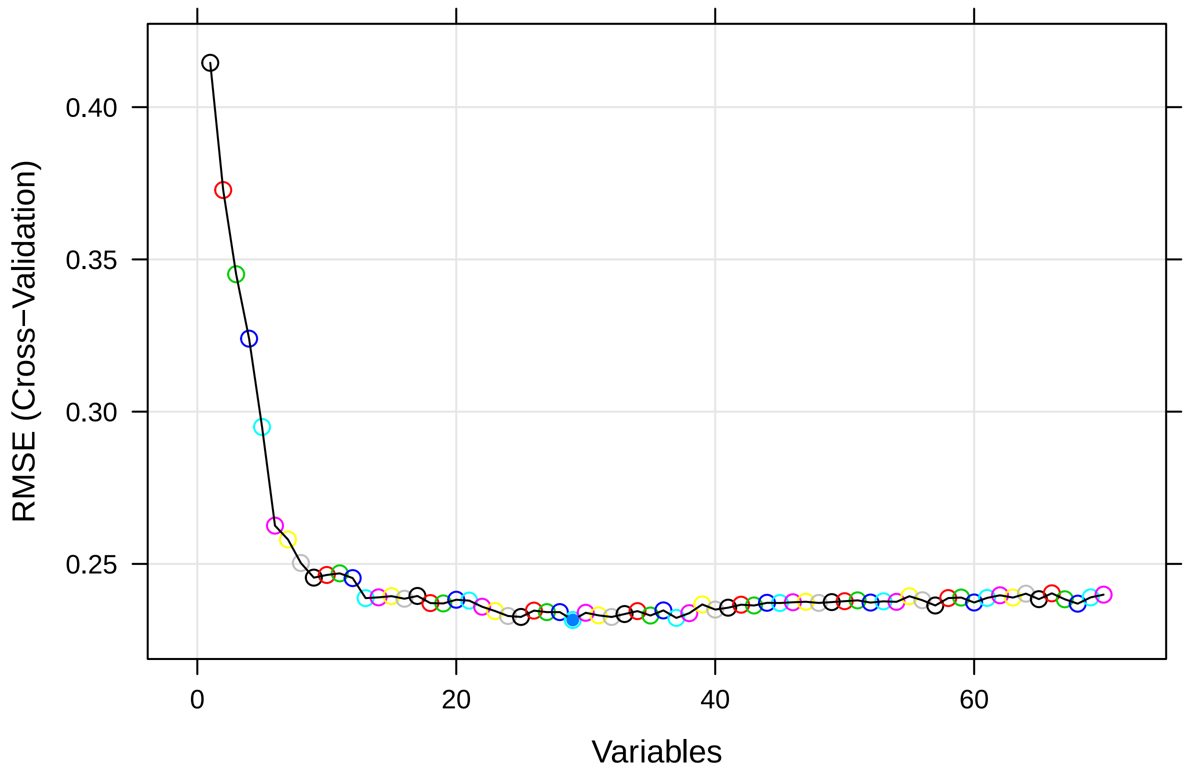
```
##
## Attaching package: 'plyr'

## The following object is masked from 'package:modeltools':
##
##     empty
```

```
predictors(rfe.train)
```

```
##  [1] "PCIP14"        "PCTFLOAN"      "PCIP04"        "PCIP52"
##  [5] "SAT_AVG_ALL"   "UGDS_BLACK"    "UGDS_MEN"      "UGDS_WOMEN"
##  [9] "PCIP45"        "PCIP43"        "COSTT4_A"      "PCIP23"
## [13] "TUITIONFEE_IN" "UGDS_HISP"     "RET_FT4"       "C150_4_AIAN"
## [17] "UGDS_ASIAN"    "PCIP39"        "PCIP16"        "UGDS_NRA"
## [21] "UGDS_WHITE"    "PCIP19"        "C150_4"        "PCIP03"
## [25] "PCIP05"        "PCIP26"        "PCIP24"        "PPTUG_EF"
## [29] "PCIP50"
```

```r
# We will plot on the number of variables and its importance for RFE
plot(rfe.train, type = c("g","o"), cex = 1.0, col = 1:70)
```



Based on these runs, RFE determines fewer variables needed for the prediction model than Boruta. There would be some cases that the Boruta package could be used, depending on the number of variables.

US Research University Completion Rate Prediction Model

```r
rm_train2 <- sample(nrow(usresearchuniv), floor(nrow(usresearchuniv)*0.75))
univ_train2 <- usresearchuniv[rm_train2,]
univ_test2 <- usresearchuniv[-rm_train2,]

formula_completionrate <- formula(C150_4_NRA ~ REGION + ADM_RATE_ALL + UGDS_NRA + PPTUG_EF + COSTT4_A +
```

We will do a generalized multivariate linear regression formula.

```r
# create a logistic regression
fit2 <- lm(formula_completionrate, data = usresearchuniv)
summary(fit2)
```

```
##
## Call:
```

```
## lm(formula = formula_completionrate, data = usresearchuniv)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62640 -0.05949  0.00907  0.07396  0.51024
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.323e-01  3.881e-02  24.021  < 2e-16 ***
## REGION            -2.791e-03  2.847e-03  -0.980  0.32728
## ADM_RATE_ALL      -1.472e-01  3.336e-02  -4.412 1.16e-05 ***
## UGDS_NRA           2.210e-01  1.274e-01   1.735  0.08314 .
## PPTUG_EF          -3.508e-01  7.451e-02  -4.708 2.94e-06 ***
## COSTT4_A           1.588e-06  5.358e-07   2.965  0.00312 **
## PCTFLOAN          -3.614e-01  5.114e-02  -7.068 3.41e-12 ***
## PAR_ED_PCT_1STGEN -9.581e-02  8.656e-02  -1.107  0.26865
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1408 on 807 degrees of freedom
## Multiple R-squared:  0.4242, Adjusted R-squared:  0.4192
## F-statistic: 84.94 on 7 and 807 DF,  p-value: < 2.2e-16
```

Based on the regression, the formula will be

$$C150\_4\_NRA = 0.932 - 0.00279REGION - 0.147ADM\_RATE\_ALL + 0.021UGDS\_NRA - 0.351PPTUG\_EF + 0.000001$$

.

We will test this regression with some data types.

```
# for Ivy League schools with high admission rates for all and international students
df_accept3 <- data.frame(REGION = 1, ADM_RATE_ALL = .55, UGDS_NRA=.25, PPTUG_EF = 0.07, COSTT4_A = 5000
predict(fit2, newdata = df_accept3)
```

```
##         1
## 0.7757938
```

```
# for Ivy League schools with less admission rates, but have high shares of students doing part-time
df_accept4 <- data.frame(REGION = 1, ADM_RATE_ALL = .05, UGDS_NRA=.05, PPTUG_EF = 0.46, COSTT4_A = 5000
predict(fit2, newdata = df_accept4)
```

```
##        1
## 0.612912
```

Now, we will do some testing of performance with the logistic regression. Since we have split the dataset into training and testing set, we will see how the performance will be done.

```
# using multivariate linear regression to calculate the completion rate for international students
lm_NRAcompletion <- lm(formula_completionrate, data = univ_train2)
summary(lm_NRAcompletion)
```

```
##
## Call:
## lm(formula = formula_completionrate, data = univ_train2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60966 -0.06432  0.01155  0.07153  0.50310
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.281e-01  4.134e-02  22.448  < 2e-16 ***
## REGION            -1.517e-03  3.087e-03  -0.491  0.62328
## ADM_RATE_ALL      -1.396e-01  3.703e-02  -3.769  0.00018 ***
## UGDS_NRA           1.166e-01  1.385e-01   0.842  0.40021
## PPTUG_EF          -3.588e-01  8.177e-02  -4.388 1.35e-05 ***
## COSTT4_A           1.571e-06  5.751e-07   2.732  0.00648 **
## PCTFLOAN          -3.624e-01  5.675e-02  -6.386 3.41e-10 ***
## PAR_ED_PCT_1STGEN -8.148e-02  9.693e-02  -0.841  0.40088
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.134 on 603 degrees of freedom
## Multiple R-squared:  0.4345, Adjusted R-squared:  0.4279
## F-statistic: 66.18 on 7 and 603 DF,  p-value: < 2.2e-16
```

```r
# do the testing with the prediction model
accepted_ind3 <- predict(lm_NRAcompletion, interval="prediction", newdata = univ_test2)

# Checking on PRED(25)
errors <- accepted_ind3[,"fit"] - univ_test2$C150_4_NRA
rel_change <- abs(errors) / univ_test2$C150_4_NRA
table(rel_change<0.25)["TRUE"] / nrow(univ_test2)
```

```
##      TRUE
## 0.7990196
```

```r
# Now we check on what acceptable ways we could do for regression
# Doing single decision tree
model_dtree3 <- rpart(formula_completionrate, method="anova",data = univ_train2)
pred_dtree3 <- predict(model_dtree3, newdata = univ_test2)
accu11 <- abs(pred_dtree3 - univ_test2$C150_4_NRA) < 0.25
frac11 <- sum(accu11)/length(accu11)
print(frac11)
```

```
## [1] 0.8872549
```

```r
# Doing random forest
model_forest3 <- randomForest(formula_completionrate, data = univ_train2)
pred_forest3 <- predict(model_forest3, newdata = univ_test2)
accu12 <- abs(pred_forest3 - univ_test2$C150_4_NRA) < 0.25
frac12 <- sum(accu12)/length(accu12)
print(frac12)
```

```
## [1] 0.9019608
```

```r
# Doing support vector machine
model_svm3 <- svm(formula_completionrate, data = univ_train2)
pred_svm3 <- predict(model_svm3, newdata = univ_test2)
accu13 <- abs(pred_svm3 - univ_test2$C150_4_NRA) < 0.25
frac13 <- sum(accu13)/length(accu13)
print(frac13)
```

```
## [1] 0.8970588
```

```r
# doing simple tree
model_tree3 <- tree(formula_completionrate, data = univ_train2)
pred_tree3 <- predict(model_tree3, newdata = univ_test2)
accu14 <- abs(pred_tree3 - univ_test2$C150_4_NRA) < 0.25
frac14 <- sum(accu14)/length(accu14)
print(frac14)
```

```
## [1] 0.8921569
```

```r
# doing conditional inference tree
model_party3 <- ctree(formula_completionrate, data = univ_train2)
pred_party3 <- predict(model_party3, newdata = univ_test2)
accu15 <- abs(pred_party3 - univ_test2$C150_4_NRA) < 0.25
frac15 <- sum(accu15)/length(accu15)
print(frac15)
```

```
## [1] 0.8823529
```

From the regressions that we have run, the random forest is the best regression model to use for determining completion rates for international students.