US Research University Prediction Model

Philip Gabriel Andrada

October 31, 2016

Preparation

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

```
## Warning: package 'randomForest' was built under R version 3.2.5

## randomForest 4.6-12

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

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.2.5

## Loading required package: lattice

## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'

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

```
library(Boruta)
```

```
## Warning: package 'Boruta' was built under R version 3.2.5

## Loading required package: ranger

## Warning: package 'ranger' was built under R version 3.2.5

##
## Attaching package: 'ranger'

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

```r
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.2.5
```

```r
library(ROCR)
```

```
## Warning: package 'ROCR' was built under R version 3.2.5
```

```
## Loading required package: gplots
```

```
## Warning: package 'gplots' was built under R version 3.2.5
```

```
## 
## Attaching package: 'gplots'
```

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

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.2.5
```

```r
library(ggplot2)

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

```r
#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)

#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
# 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 studnets 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]))
}
```

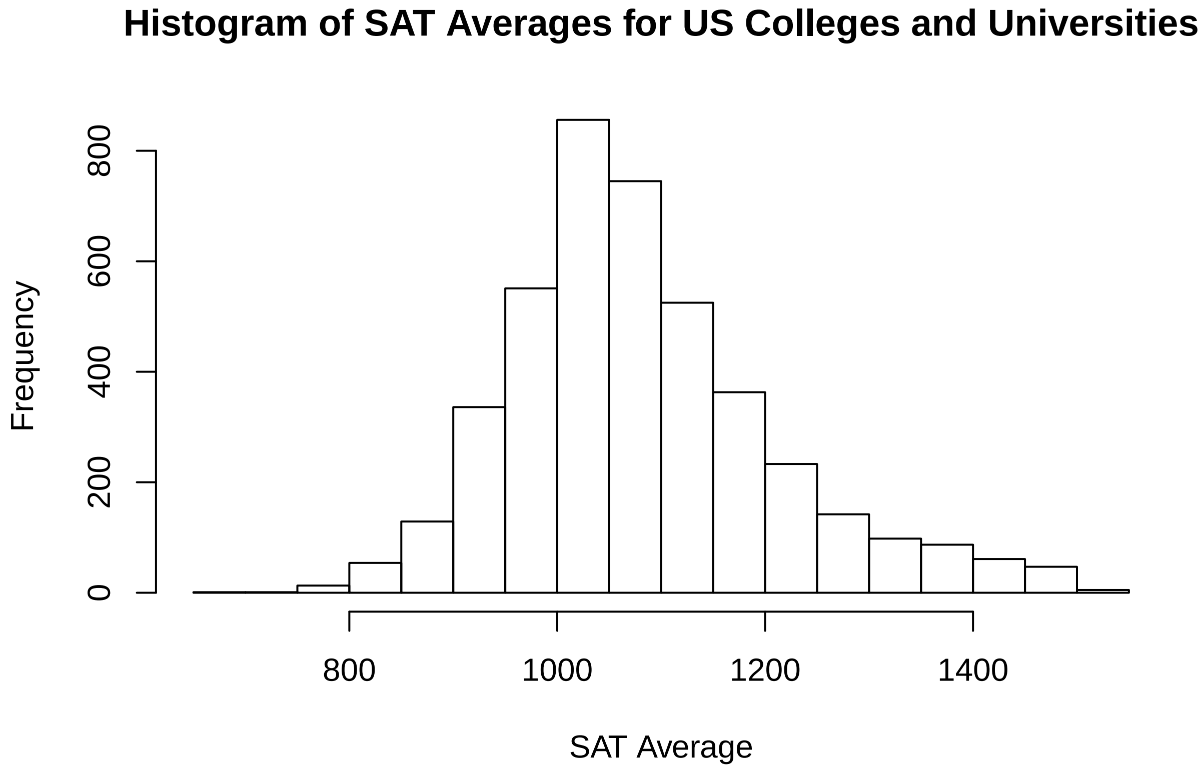
```
## Warning: NAs introduced by coercion
```

```
## 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
```

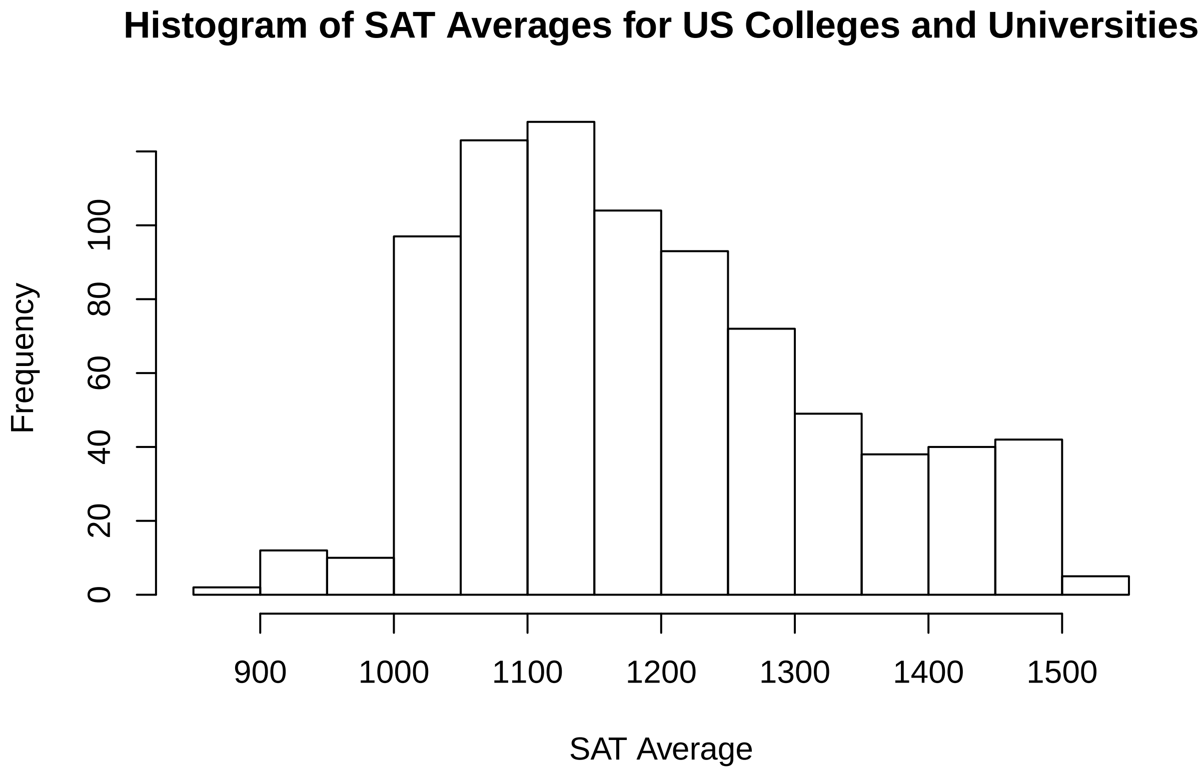
```r
# Clean the results to have all complete
usunivfilter <- usunivfilter[!is.na(usunivfilter$C150_4),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$C150_4_ASIAN),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$C150_4_WHITE),]
usunivfilter <- usunivfilter[!is.na(usunivfilter$C150_4_BLACK),]
usunivfilter <- 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),]
```

Distributions and Box and Whisker Plots

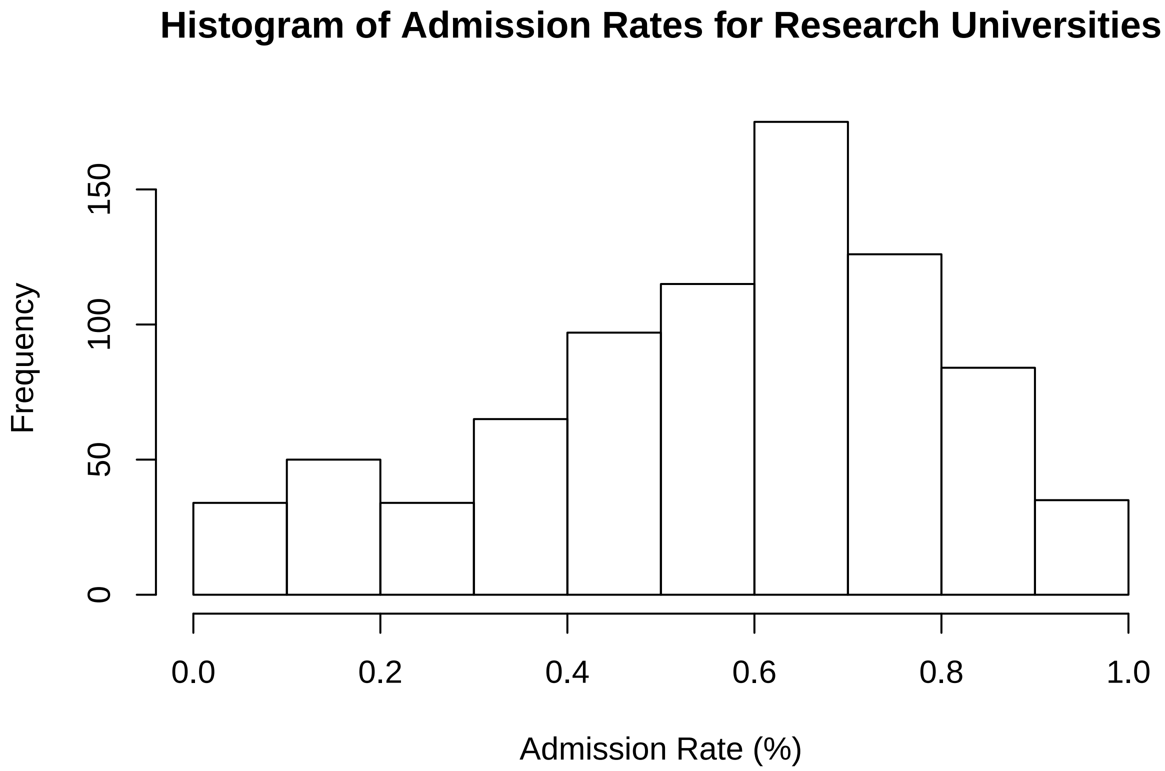
```
# Histogram of SAT Averages for US Colleges and Universities
hist(usunivfilter$SAT_AVG_ALL, main = "Histogram of SAT Averages for US Colleges and Universities", xlab
```



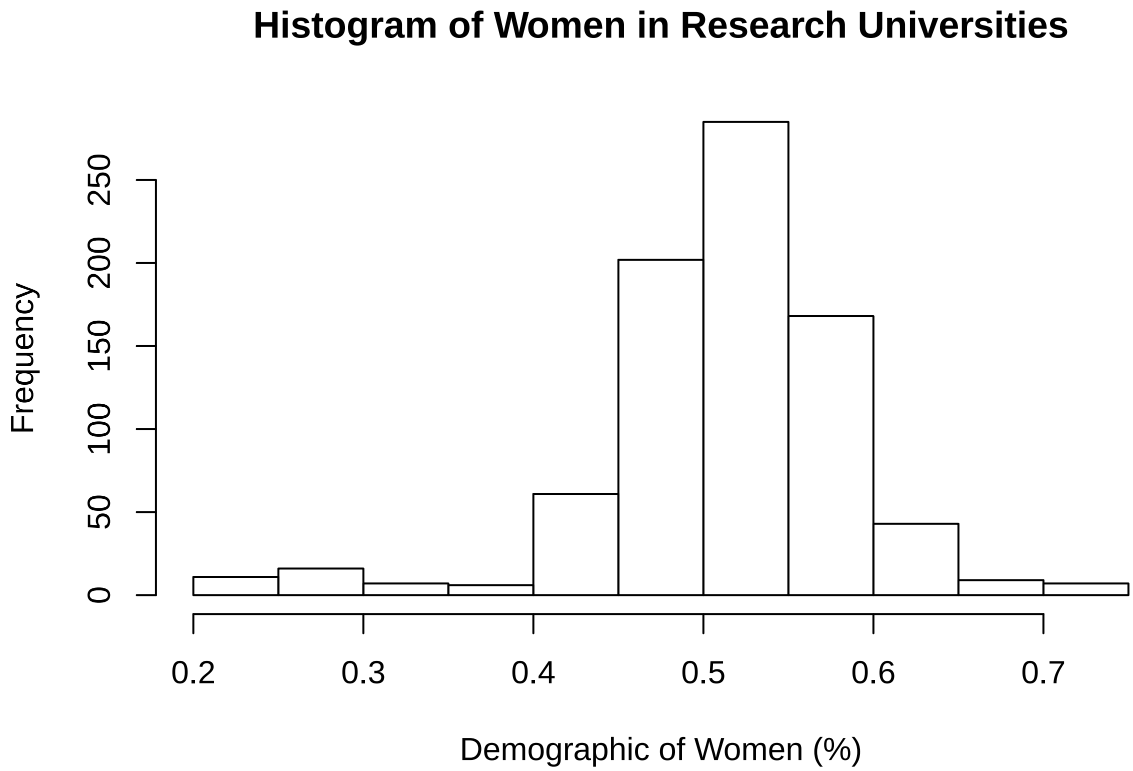
```
# Histogram of SAT Averages for US Research Universities
hist(usresearchuniv$SAT_AVG_ALL, main = "Histogram of SAT Averages for US Colleges and Universities", xl
```



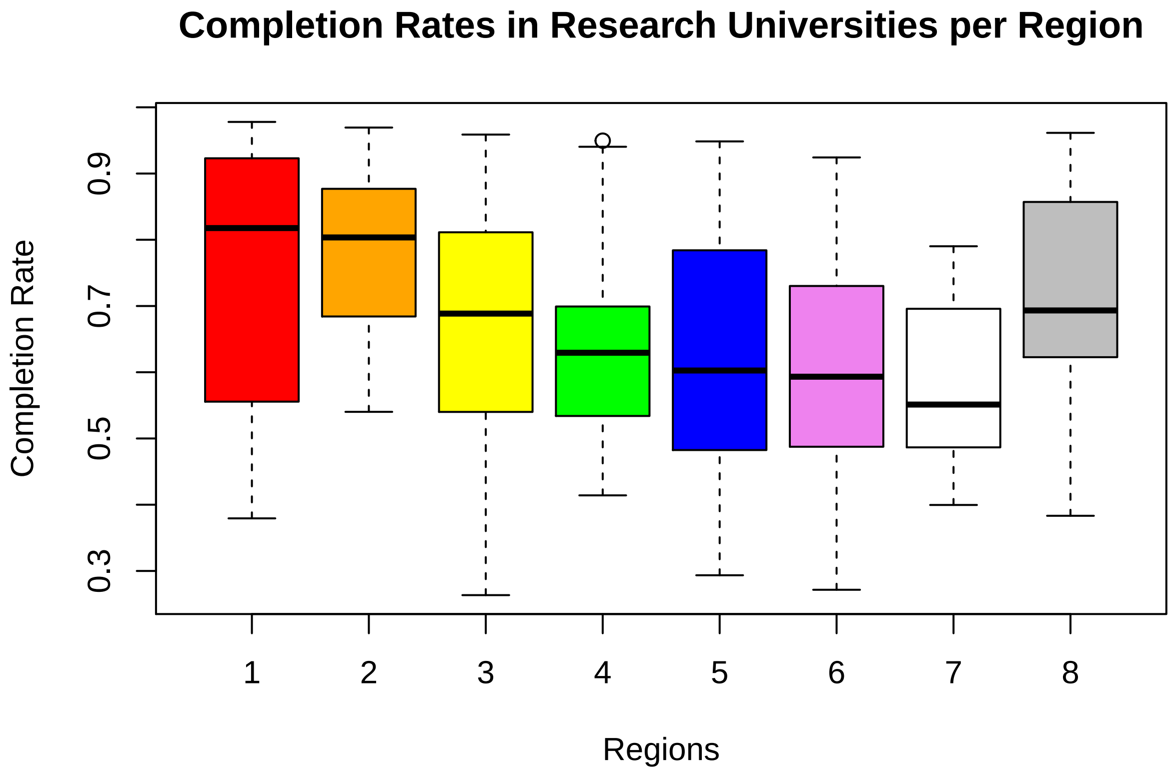
```
# Histogram of Admission Rates for US Research Universities
hist(usresearchuniv$ADM_RATE_ALL, main = "Histogram of Admission Rates for Research Universities", xlab
```



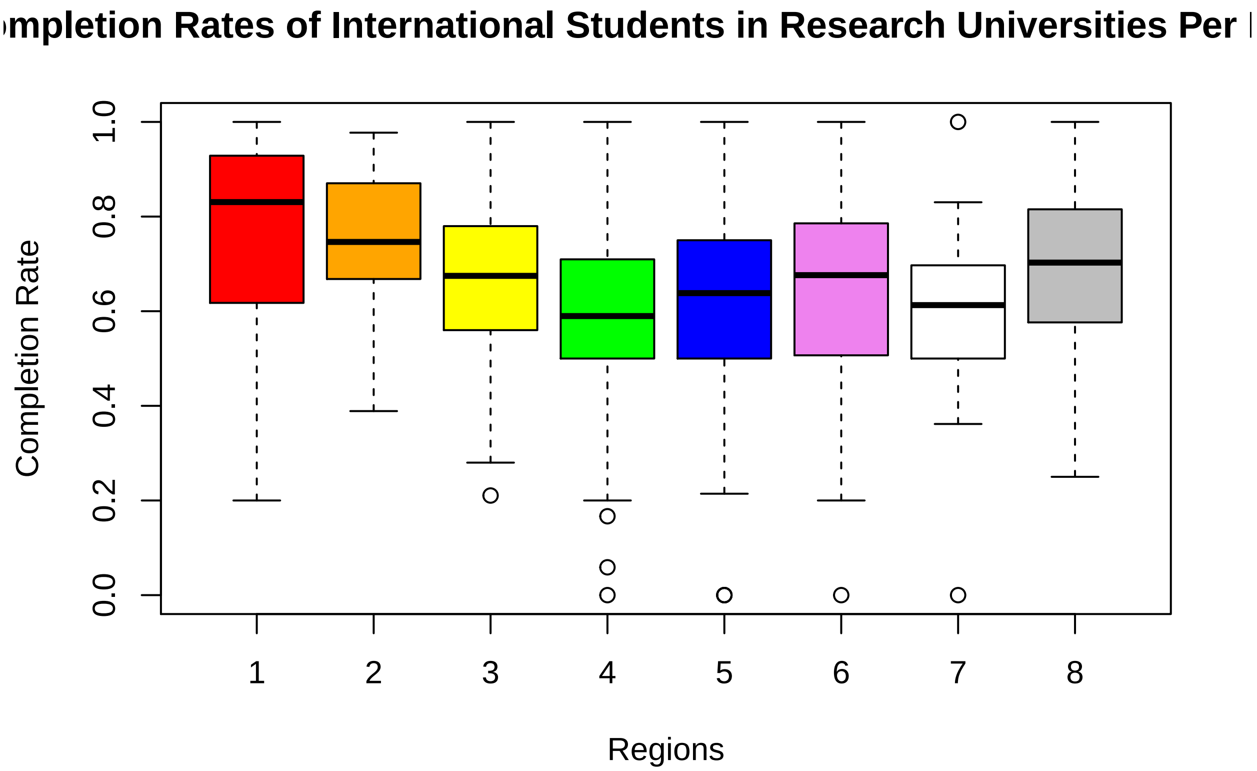
```
# Histogram of Women in US Research Universitie
hist(usresearchuniv$UGDS_WOMEN, main = "Histogram of Women in Research Universities", xlab = "Demographi
```



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

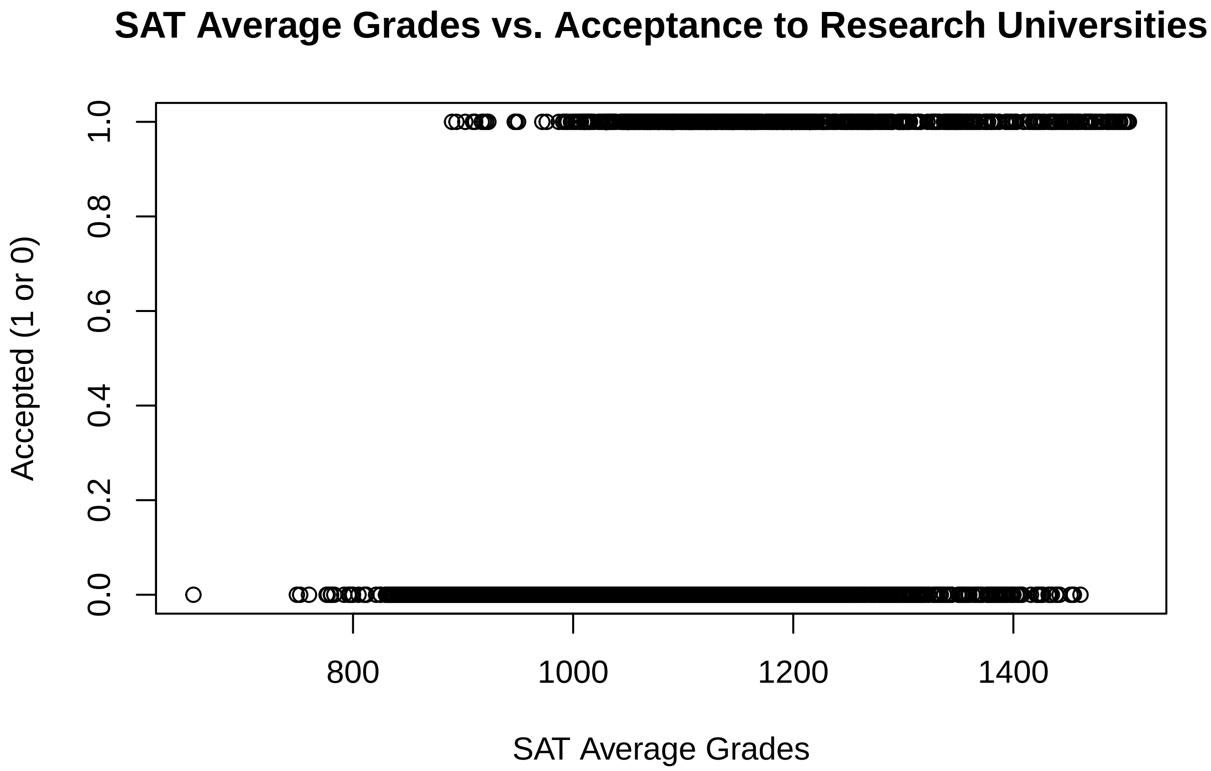


```
# Boxplot of COmpletion Rates of International Students per Region in US Research Universities
boxplot(C150_4_NRA ~ REGION, usresearchuniv, main = "Completion Rates of International Students in Rese
```

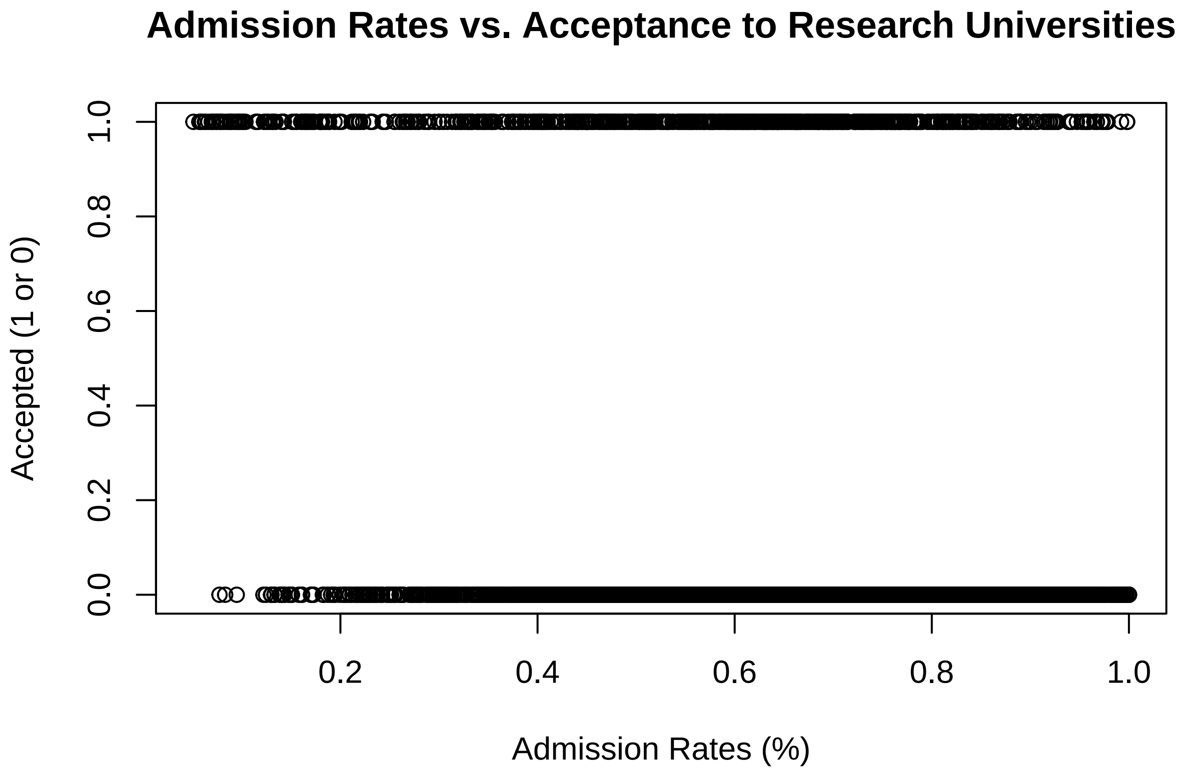


Correlations

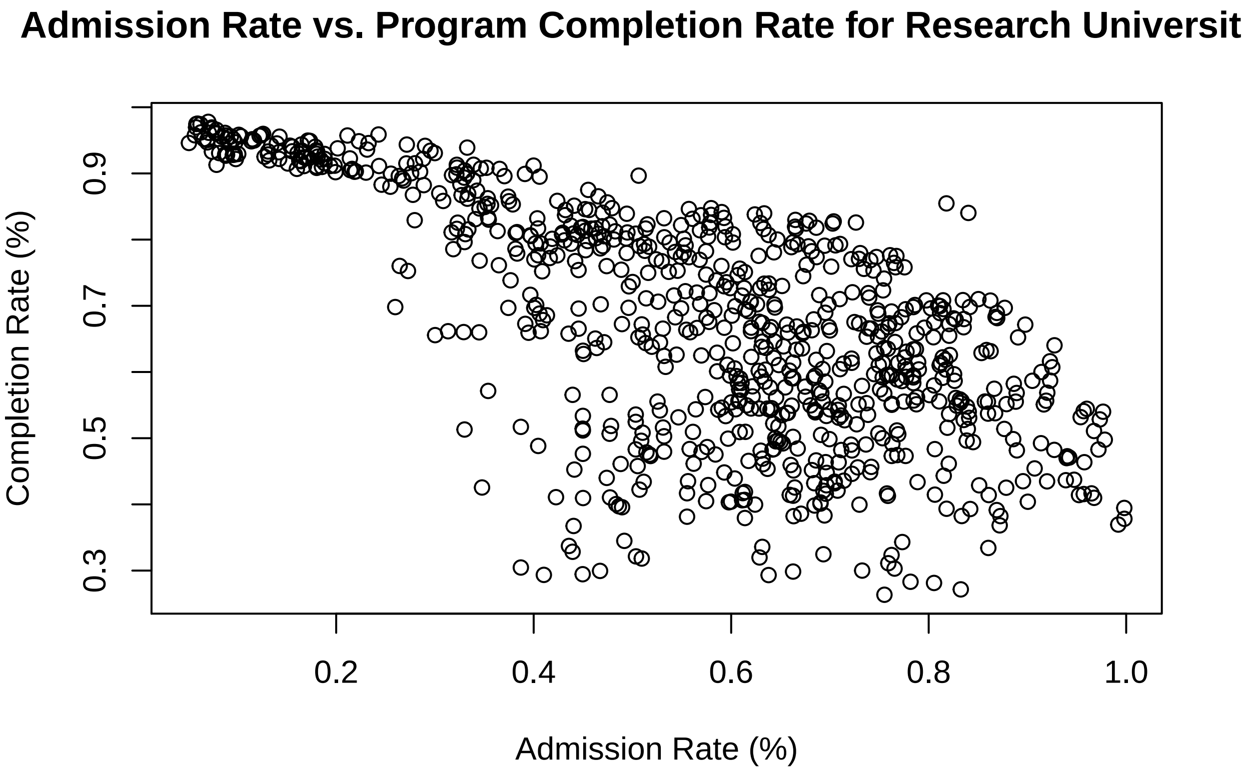
```
#Correlation between the SAT grades and the acceptance for the research universities
plot(usunivfilter$SAT_AVG_ALL, usunivfilter$ACCEPTED, main="SAT Average Grades vs. Acceptance to Researc
```



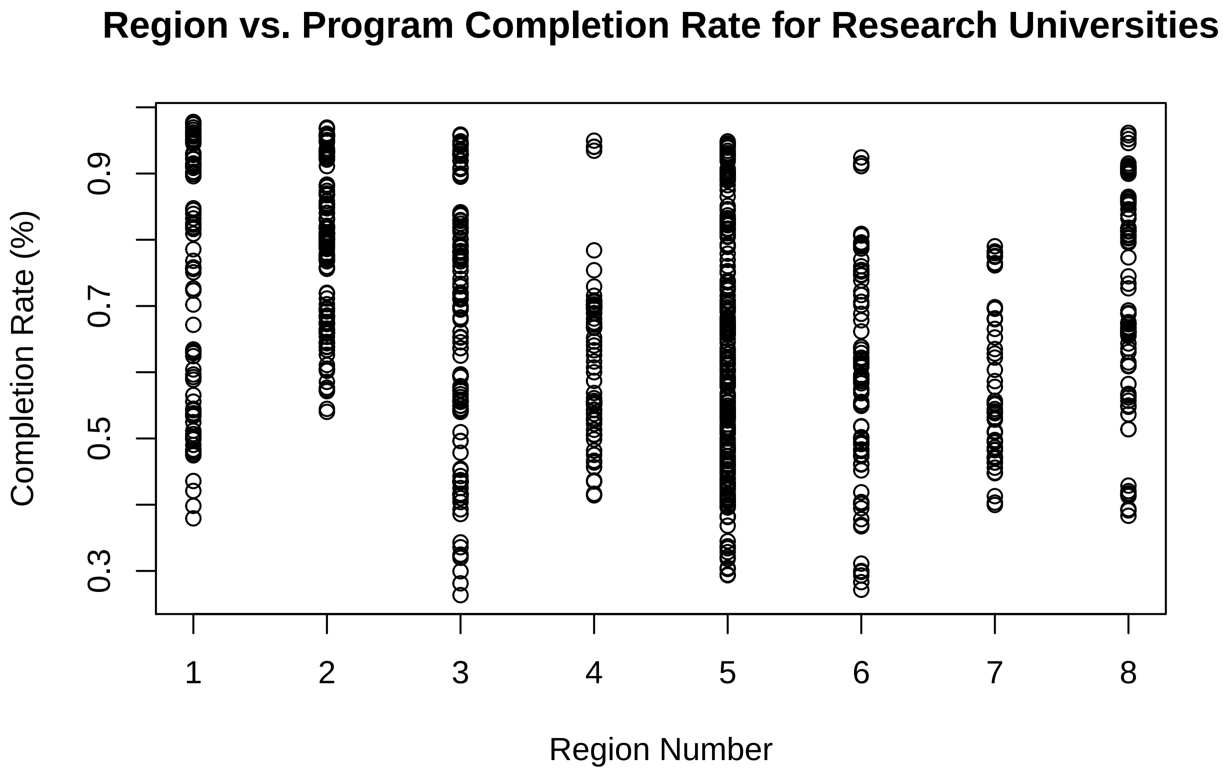
```
#Correlation between the admission rates and the acceptance for the research universities
plot(usunivfilter$ADM_RATE_ALL, usunivfilter$ACCEPTED, main="Admission Rates vs. Acceptance to Research
```



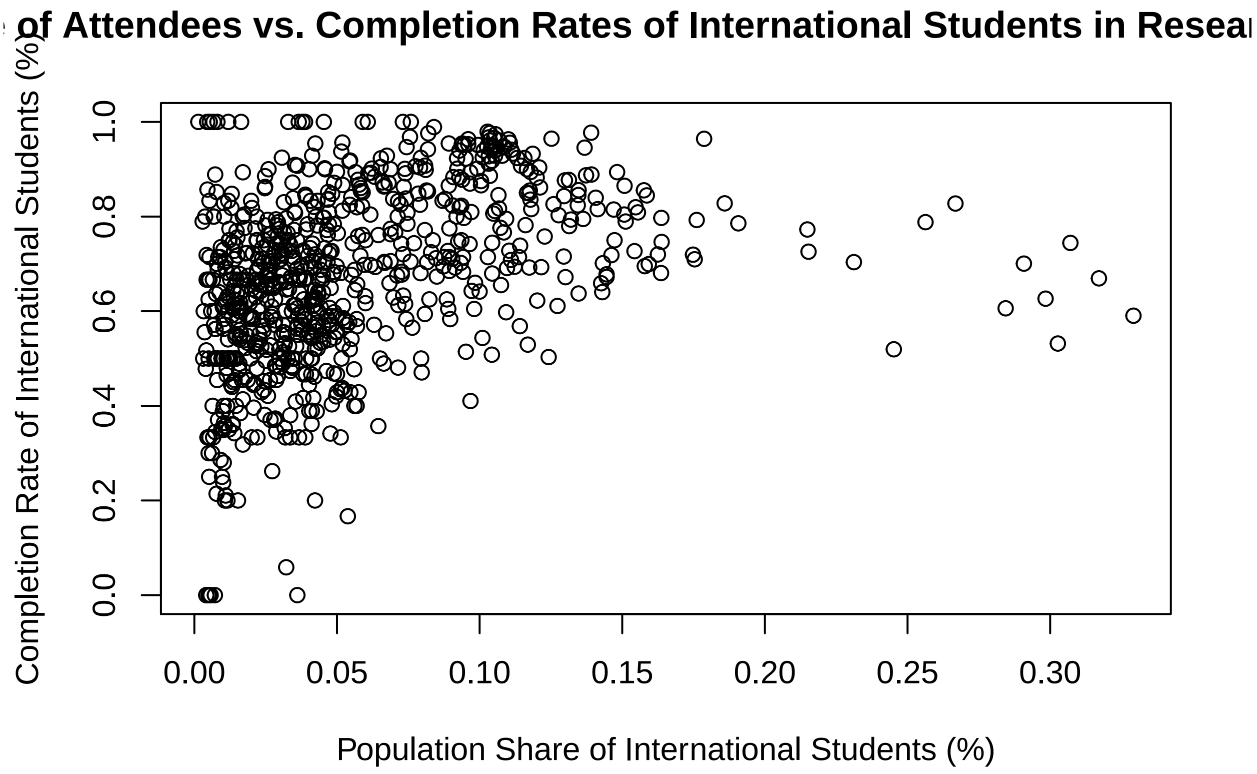
```
#Correlation between admission rate for research universities and program completion rate
plot(usresearchuniv$ADM_RATE_ALL, usresearchuniv$C150_4, main="Admission Rate vs. Program Completion Ra
```



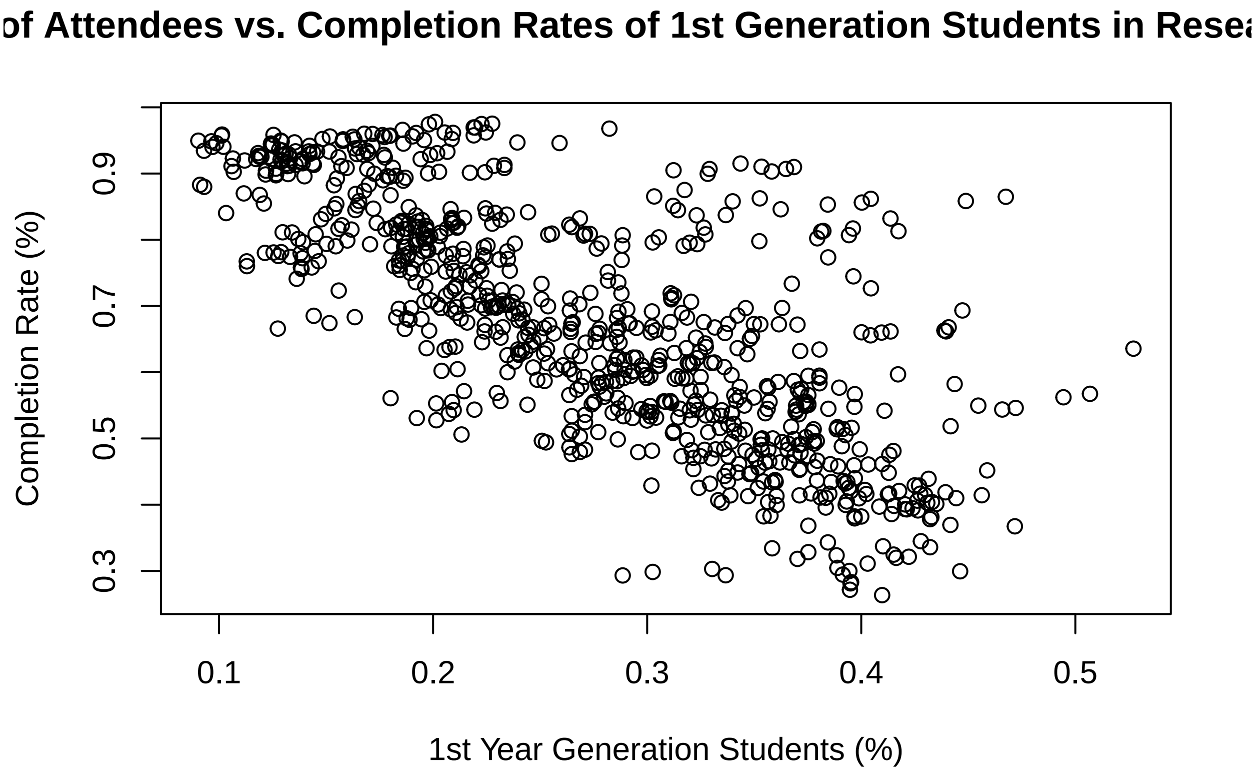
```
#Correlation between admission rate for research universities and program completion rate
plot(usresearchuniv$REGION, usresearchuniv$C150_4, main="Region vs. Program Completion Rate for Researc
```



```
#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
```



```
#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
```



U.S. Research University Acceptance Model

```
# 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 formula for the US research university acceptance model for International Students taking up
test_formulagen <- formula(ACCEPTED ~ REGION + ADM_RATE_ALL + SAT_AVG_ALL + PCIP11 + PCIP12 + PCIP14 + 

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

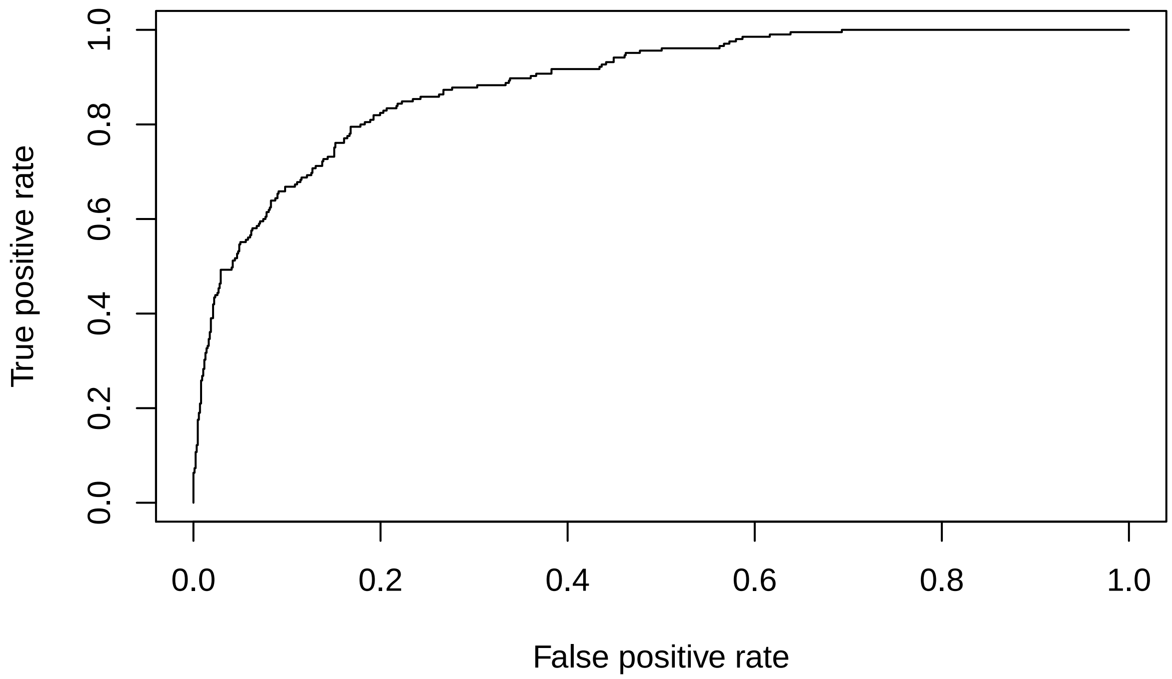
```
##
## Call:
## glm(formula = test_formulagen, family = binomial(), data = univ_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5894  -0.4736  -0.2334  -0.0755   3.1683
##
## Coefficients:
```

```
##                  Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  -1.844e+01  1.477e+00 -12.481  < 2e-16 ***
## REGION        1.502e-01  3.229e-02   4.651 3.31e-06 ***
## ADM_RATE_ALL  9.035e-01  4.235e-01   2.133  0.03291 *  
## SAT_AVG_ALL   1.611e-02  1.044e-03  15.434  < 2e-16 ***
## PCIP11        2.678e+00  2.221e+00   1.206  0.22791    
## PCIP12        3.367e+00  1.879e+01   0.179  0.85783    
## PCIP14        5.685e+00  7.813e-01   7.276 3.43e-13 ***
## PCIP15       -4.148e-01  2.337e+00  -0.178  0.85909    
## PCIP24       -5.704e+00  1.312e+00  -4.348 1.37e-05 ***
## PCIP26        7.600e+00  1.802e+00   4.218 2.46e-05 ***
## PCIP27       -2.792e+01  7.130e+00  -3.916 9.01e-05 ***
## PCIP40       -3.330e+01  4.977e+00  -6.691 2.21e-11 ***
## PCIP45        8.596e+00  1.223e+00   7.028 2.09e-12 ***
## PCIP51        1.894e+00  6.150e-01   3.080  0.00207 ** 
## PCIP52        7.409e-01  6.779e-01   1.093  0.27444    
## UGDS_NRA      8.244e+00  1.507e+00   5.469 4.53e-08 ***
## UGDS_UNKN    -4.899e-01  1.602e+00  -0.306  0.75977    
## COSTT4_A     -1.144e-04  7.502e-06 -15.247  < 2e-16 ***
## PCTFLOAN     -2.850e-01  5.751e-01  -0.496  0.62022    
## UGDS_WOMEN    5.852e-01  8.381e-01   0.698  0.48504    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 3111.3  on 3184  degrees of freedom
## Residual deviance: 1864.6  on 3165  degrees of freedom
## AIC: 1904.6
## 
## Number of Fisher Scoring iterations: 6
```

```r
# do the testing with the prediction model
univ_test$scores <- predict(model_glm, type="response", newdata = univ_test)
pred <- prediction(univ_test$scores, univ_test$ACCEPTED)

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

```
##           Reference
## Prediction   0   1
##          0 815  96
##          1  42 109
```

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



```
# Now we check on what acceptable ways we could do for regression
#doing single decision tree
model_tree <- rpart(test_formulagen, method="anova",data = univ_train)
pred_tree <- predict(model_tree, newdata = univ_test)
accu = abs(pred_tree - univ_test$ACCEPTED) < 0.25
frac = sum(accu)/length(accu)
print(frac)
```

```
## [1] 0.8625235
```

```
#doing random forest
model_forest <- randomForest(test_formulagen, 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_forest <- predict(model_forest, newdata = univ_test)
accu2 <- abs(pred_forest - univ_test$ACCEPTED) < 0.25
frac2 <- sum(accu2)/length(accu2)
print(frac2)
```

```
## [1] 0.8709981
```

```
#doing support vector machine
model_svm <- svm(test_formulagen, data = univ_train)
pred_svm <- predict(model_svm, newdata = univ_test)
accu3 <- abs(pred_svm - univ_test$ACCEPTED) < 0.25
frac3 <- sum(accu3)/length(accu3)
print(frac3)
```

```
## [1] 0.8436911
```

```
# We will consider all variables, and use Boruta to use what variables we could use for doing a better

# 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,doTrace = 2)
```

```
##  1. run of importance source...

##  2. run of importance source...

##  3. run of importance source...

##  4. run of importance source...

##  5. run of importance source...

##  6. run of importance source...

##  7. run of importance source...

##  8. run of importance source...

##  9. run of importance source...

##  10. run of importance source...

##  11. run of importance source...

##  12. run of importance source...

##  13. run of importance source...
## Confirmed 45 attributes: ADM_RATE, ADM_RATE_ALL, C150_4, C150_4_AIAN, C150_4_ASIAN and 40 more.

## Rejected 6 attributes: PCIP12, PCIP25, PCIP29, PCIP46, PCIP47 and 1 more.

##  14. run of importance source...

##  15. run of importance source...

##  16. run of importance source...

##  17. run of importance source...

## Confirmed 6 attributes: C150_4_NRA, PCIP09, PCIP30, PCIP31, PCIP40 and 1 more.

##  18. run of importance source...

##  19. run of importance source...

##  20. run of importance source...

##  21. run of importance source...

## Confirmed 3 attributes: PCIP42, UGDS_2MOR, UGDS_NHPI.

##  22. run of importance source...

##  23. run of importance source...

##  24. run of importance source...

## Confirmed 1 attributes: PCIP38.

## Rejected 1 attributes: C150_4_NHPI.

##  25. run of importance source...

##  26. run of importance source...

##  27. run of importance source...

##  28. run of importance source...

##  29. run of importance source...

##  30. run of importance source...

##  31. run of importance source...
## Confirmed 1 attributes: UGDS_AIAN.

##  32. run of importance source...

##  33. run of importance source...

##  34. run of importance source...

## Confirmed 1 attributes: PCIP51.

##  35. run of importance source...

##  36. run of importance source...

##  37. run of importance source...

## Confirmed 1 attributes: PCIP54.

##  38. run of importance source...

##  39. run of importance source...

##  40. run of importance source...

##  41. run of importance source...

##  42. run of importance source...

##  43. run of importance source...

##  44. run of importance source...

##  45. run of importance source...

##  46. run of importance source...

##  47. run of importance source...

##  48. run of importance source...

##  49. run of importance source...

##  50. run of importance source...

##  51. run of importance source...

##  52. run of importance source...
##  53. run of importance source...

## Confirmed 1 attributes: PCIP49.

##  54. run of importance source...

##  55. run of importance source...

##  56. run of importance source...

##  57. run of importance source...

##  58. run of importance source...

##  59. run of importance source...

##  60. run of importance source...

##  61. run of importance source...

##  62. run of importance source...

##  63. run of importance source...

##  64. run of importance source...

##  65. run of importance source...

##  66. run of importance source...

##  67. run of importance source...

##  68. run of importance source...

##  69. run of importance source...

##  70. run of importance source...

##  71. run of importance source...

##  72. run of importance source...

##  73. run of importance source...

##  74. run of importance source...

##  75. run of importance source...
##  76. run of importance source...

##  77. run of importance source...

##  78. run of importance source...

##  79. run of importance source...

##  80. run of importance source...

##  81. run of importance source...

##  82. run of importance source...

##  83. run of importance source...

##  84. run of importance source...

##  85. run of importance source...

##  86. run of importance source...

##  87. run of importance source...

## Confirmed 1 attributes: C150_4_2MOR.

##  88. run of importance source...

##  89. run of importance source...

##  90. run of importance source...

##  91. run of importance source...

##  92. run of importance source...

##  93. run of importance source...

##  94. run of importance source...

##  95. run of importance source...

##  96. run of importance source...

##  97. run of importance source...

##  98. run of importance source...

##  99. run of importance source...
```

```r
print(boruta.train)
```

```
## Boruta performed 99 iterations in 2.383894 mins.
##  60 attributes confirmed important: ADM_RATE, ADM_RATE_ALL,
## C150_4, C150_4_2MOR, C150_4_AIAN and 55 more.
##  7 attributes confirmed unimportant: C150_4_NHPI, PCIP12, PCIP25,
## PCIP29, PCIP46 and 2 more.
##  3 tentative attributes left: PCIP10, PCIP22, PCIP41.
```

```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             5.4744373  5.437121412  4.1756023  6.807213 0.98989899
## ADM_RATE           7.1771165  7.228901906  6.0390532  8.093531 1.00000000
## ADM_RATE_ALL       7.3687573  7.380258604  5.9532257  8.656572 1.00000000
## SAT_AVG_ALL       12.6677144 12.655786014 11.6084892 13.971439 1.00000000
## PCIP01             6.2168168  6.247229035  5.1109627  7.426304 1.00000000
## PCIP03             6.7048878  6.685731151  5.0923366  8.024387 1.00000000
## PCIP04            11.6633329 11.596998822  9.9733377 13.261449 1.00000000
## PCIP05             8.3050793  8.325927761  7.2773996  9.422381 1.00000000
## PCIP09             4.6951336  4.673520095  2.8108687  6.934877 0.94949495
## PCIP10             2.6528575  2.658804922  0.4454767  4.911361 0.49494949
## PCIP11             6.5059567  6.554636949  4.5918813  8.262611 1.00000000
## PCIP12             0.5509474  0.353325251 -1.0669728  2.299341 0.00000000
## PCIP13             6.0977431  6.156386451  4.2227750  7.508339 1.00000000
## PCIP14            18.6134104 18.713594604 16.1691332 20.922420 1.00000000
## PCIP15             4.9772239  4.925653746  3.1198039  7.393903 0.96969697
## PCIP16             7.5962002  7.598226993  5.8462783  9.097015 1.00000000
## PCIP19             7.5415378  7.556854663  5.7850539  9.049480 1.00000000
## PCIP22             2.4022802  2.414647838  0.8317286  4.172157 0.36363636
## PCIP23             8.5266827  8.520035328  7.0071059 10.040757 1.00000000
## PCIP24             5.8874692  5.890028738  4.3067697  7.250042 0.98989899
## PCIP25            -0.9726754 -1.001001503 -1.7369486  1.001002 0.00000000
## PCIP26             5.8827098  5.777586159  4.3339598  7.955142 0.98989899
## PCIP27             5.1316515  5.250736385  2.3844392  6.402645 0.95959596
## PCIP29             0.1540002  0.000000000  0.0000000  1.001002 0.00000000
## PCIP30             4.1722850  4.237703974  1.6429832  5.822066 0.89898990
## PCIP31             4.7894844  4.736011553  2.2515575  6.657689 0.93939394
## PCIP38             4.1650553  4.370945518  2.2346836  5.991752 0.80808081
## PCIP39             5.3934307  5.401710185  4.1512923  6.717005 0.96969697
## PCIP40             5.6629928  5.688726617  3.1500225  7.181317 0.97979798
## PCIP41             3.1565885  3.153017243  0.9984023  5.040465 0.62626263
## PCIP42             4.8658299  4.840201699  2.4722604  6.780566 0.94949495
## PCIP43             7.2982438  7.254221812  5.2427446  9.017156 1.00000000
## PCIP44             4.3931083  4.527186968  2.6444481  6.011138 0.93939394
## PCIP45             7.6312425  7.596119228  6.0330702  8.898818 1.00000000
## PCIP46             0.3327630  0.000000000 -1.0010015  1.339068 0.00000000
## PCIP47             0.3004310  0.007103357 -1.3732711  1.076954 0.00000000
## PCIP48             0.1567115  0.000000000 -1.0010015  1.261645 0.00000000
## PCIP49             3.3721866  3.425373692  1.2469717  4.818251 0.73737374
## PCIP50             5.9150943  5.955009806  4.0734608  7.214410 1.00000000
## PCIP51             4.0822881  4.081832035  1.2697965  5.603184 0.85858586
## PCIP52             9.7189220  9.743214118  8.5082383 11.175685 1.00000000
## PCIP54             3.7015048  3.626973660  2.1451159  5.362206 0.77777778
## UGDS_WHITE         8.0886688  8.096749563  7.0342571  9.463207 1.00000000
## UGDS_BLACK        10.7646501 10.813176742  9.0897299 12.168616 1.00000000
## UGDS_HISP          6.2726291  6.327185547  3.6820898  8.890406 1.00000000
## UGDS_ASIAN         9.2120393  9.192869734  7.8613224 10.991136 1.00000000
## UGDS_AIAN          4.3360892  4.392166772  2.4115052  5.899400 0.87878788
## UGDS_NHPI          3.9400272  3.982382090  1.8455958  6.749607 0.87878788
## UGDS_2MOR          4.4293440  4.441380500  2.4132079  6.312217 0.91919192
## UGDS_NRA           7.1848596  7.186749790  5.5381773  8.685459 1.00000000
## UGDS_UNKN          6.1249276  6.107241375  4.4570236  7.595158 1.00000000
## PPTUG_EF           6.9252571  6.893766476  5.6649436  8.637785 1.00000000
## COSTT4_A           9.8124440  9.802573729  8.3765787 10.794314 1.00000000
## TUITIONFEE_IN      9.4133056  9.487930915  8.3784801 10.459977 1.00000000
## TUITIONFEE_OUT     5.5935655  5.609239419  3.9129053  6.839641 0.98989899
## C150_4             7.9945622  8.013674503  6.1732411  9.563041 1.00000000
## C150_4_WHITE       6.6786801  6.801966891  4.9819788  7.782207 1.00000000
## C150_4_BLACK       7.1035898  7.067923943  6.1021741  8.072265 1.00000000
## C150_4_HISP        5.7879240  5.889406092  4.0645787  6.842054 1.00000000
## C150_4_ASIAN       5.9856984  6.038513643  4.7191157  7.196764 0.98989899
## C150_4_AIAN        7.3337452  7.360580787  5.7408871  8.572327 1.00000000
## C150_4_NHPI        0.7407442  0.864998508 -1.5404945  2.805573 0.03030303
## C150_4_2MOR        3.3702221  3.425099272  0.3401175  5.013446 0.68686869
## C150_4_NRA         4.4750571  4.561964159  2.8771250  6.122983 0.94949495
## C150_4_UNKN        7.2538256  7.275432811  6.0680943  8.668261 1.00000000
## RET_FT4           10.6247681 10.619179715  8.9615765 11.747290 1.00000000
## PCTFLOAN          14.1502247 14.147321246 13.0886824 15.547501 1.00000000
## PAR_ED_PCT_1STGEN  5.9685891  6.045787778  4.2756905  7.347704 1.00000000
## UGDS_MEN          12.4057365 12.392528202 11.0346205 13.789792 1.00000000
## UGDS_WOMEN        12.4467286 12.374689074 10.8065370 14.316389 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            Tentative
## 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       Confirmed
## C150_4_NRA        Confirmed
## C150_4_UNKN       Confirmed
## RET_FT4           Confirmed
## PCTFLOAN          Confirmed
## PAR_ED_PCT_1STGEN Confirmed
## UGDS_MEN          Confirmed
## UGDS_WOMEN        Confirmed
```

US Research University Completion Rate Prediction Model

```
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 ~ REGION + ADM_RATE_ALL + UGDS_NRA + PPTUG_EF + COSTT4_A + PCT

model_tree2 <- rpart(formula_completionrate, method="anova",data = univ_train2)
pred_tree2 <- predict(model_tree2, newdata = univ_test2)
accu4 = abs(pred_tree2 - univ_test2$C150_4_NRA) < 0.25
frac4 = sum(accu4)/length(accu4)
print(frac4)
```

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

```
model_forest2 <- randomForest(formula_completionrate, data = univ_train2)
pred_forest2 <- predict(model_forest2, newdata = univ_test2)
accu5 <- abs(pred_forest2 - univ_test2$ACCEPTED) < 0.25
frac5 <- sum(accu5)/length(accu5)
print(frac5)
```

```
## [1] 0.3823529
```

```
model_svm2 <- svm(formula_completionrate, data = univ_train2)
pred_svm2 <- predict(model_svm2, newdata = univ_test2)
accu6 <- abs(pred_svm2 - univ_test2$ACCEPTED) < 0.25
frac6 <- sum(accu6)/length(accu6)
print(frac6)
```

```
## [1] 0.3627451
```