SummerProgramExample

Dr. Ami Gates

Week 2 SYR – Data Cleaning and Preparation

```
#####################################################################
## Tutorial for Cleaning, Prepping, and Evaluating Data
## using Decision Trees, Naive Bayes
## and clustering - k-means & hier, mclust with EM
## Cosine sime...
## Goals, clustering and Prediction

## NOTICE: The data used here is intended to simulate real
## student application data for a special summer study
## abroad program.

## *** IMPOTRANT ** THIS DATA IS PRETEND :) (of course)

## IT IS NOT REAL and is not associated with any humans,
## organizations, opinions, venues, or institutions.
## It was created to feel like real data so as to simulate
## a real data analysis experience.

#####################################################################
## Gates, 2018

## NOTE: You will notice that I comment out many lines of code.
## I do this so that the lines can be added back in for testing
## and review. I also do this to sometimes show options for
## doing the same thing.
#####################################################################

## LIBRARIES
library(stringr)
#install.packages("e1071")
library(e1071)
#install.packages("mlr")
library(mlr)

## Loading required package: ParamHelpers

##
## Attaching package: 'mlr'

## The following object is masked from 'package:e1071':
##
##     impute
```

```r
# install.packages("caret")
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.5.1

##
## Attaching package: 'caret'

## The following object is masked from 'package:mlr':
##
##     train

#install.packages("naivebayes")
library(naivebayes)
#install.packages("e1071")
library(e1071)
#install.packages("mlr")
library(mlr)
# install.packages("caret")
library(caret)
#install.packages("naivebayes")
library(naivebayes)
library(mclust)

## Package 'mclust' version 5.4
## Type 'citation("mclust")' for citing this R package in publications.

library(cluster)
library(tm)

## Loading required package: NLP

##
## Attaching package: 'NLP'

## The following object is masked from 'package:ggplot2':
##
##     annotate

## install.packages("rpart")
## install.packages('rattle')
## install.packages('rpart.plot')
## install.packages('RColorBrewer')
## install.packages("Cairo")
library(rpart)
library(rattle)
```

```
## Rattle: A free graphical interface for data science with R.
## Version 5.1.0 Copyright (c) 2006-2017 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.

library(rpart.plot)
library(RColorBrewer)
library(Cairo)
# install.packages("philentropy")
library(philentropy)
# install.packages("forcats")
library(forcats)
# install.packages("lsa")
library(lsa) #for cosine similarity

## Loading required package: SnowballC

# install.packages("igraph")
library(igraph)  #to create network of cos sim matrix

##
## Attaching package: 'igraph'

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union

# install.packages("ggplot2")
library(ggplot2)
# install.packages("corrplot")
library(corrplot)

## corrplot 0.84 loaded

## install.packages("pastecs") ## for stats
library(pastecs)
##install.packages("dplyr")
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:pastecs':
##
##     first, last

## The following objects are masked from 'package:igraph':
##
##     as_data_frame, groups, union
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## install.packages("ggpubr")
library(ggpubr)

## Loading required package: magrittr

##
## Attaching package: 'magrittr'

## The following object is masked from 'package:pastecs':
##
##     extract

## !!! YOU must set this to YOUR working dir and your filename !!!
setwd("C:\\Users\\profa\\Documents\\R\\RStudioFolder_1\\DrGExamples")
StudentFile="StudentSummerProgramData.csv"
#StudentFileDF <- read.csv(StudentFile, header = TRUE, sep = ",", encoding =
"UTF-8",
 #                          stringsAsFactors = FALSE)
StudentFileDF <- read.csv(StudentFile, header = TRUE, sep = ",", encoding = "
latin1",
                          stringsAsFactors = FALSE)

##  Whenever you do anything - always check to see if it did what you think i
t did
(head(StudentFileDF))

##   N_ID    DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 1    3  9/13/2017 Female California 3.90     6.7      962    NA    Admit
## 2    4  9/20/2017 Female    Florida 3.80     1.4      969    97    Admit
## 3    7  10/4/2017   Male California 3.80     2.3      970    NA    Admit
## 4   10  10/7/2017   Male   Colorado 3.60     0.9      969    NA    Admit
## 5   13  11/3/2017   Male   Colorado 3.92     1.2      969    95    Admit
## 6   18 11/18/2017   Male California 3.80     1.2      967    NA    Admit

#(StudentFileDF)
#dim(StudentFileDF)
#(colnames(StudentFileDF))
########################################## Data Cleaning and Prep ------------
-----------
## Start the prep process by investigating the data attributes
## Once I investigate - I comment out these code lines
(table(StudentFileDF$GPA))
```

```
##
## 2.34 2.77 2.81 2.85  2.9 2.91 2.98    3 3.01  3.1 3.11 3.12 3.18 3.21 3.22
##    1    1    1    1    1    1    1    1    1    1    2    1    1    2    1
## 3.24 3.29 3.32 3.33 3.34 3.35 3.37 3.38 3.39  3.4 3.41 3.42 3.43 3.44 3.45
##    1    1    1    2    2    1    1    1    3    2    3    2    1    3    5
## 3.46 3.47 3.48 3.49  3.5 3.51 3.52 3.53 3.54 3.55 3.56 3.57 3.58 3.59  3.6
##    1    1    1    3    3    2    2    3    5    3    7    2    2    5    2
## 3.61 3.62 3.64 3.65 3.66 3.67 3.69  3.7 3.71 3.74 3.75 3.77 3.78 3.79  3.8
##    3    1    1    2    3    1    3    4    2    1    1    4    6    1    4
## 3.84 3.86 3.87 3.88 3.89  3.9 3.91 3.92 3.93 3.94 3.97
##    1    1    3    3    3    4    1    3    1    1    1
```

```
(table(StudentFileDF$State))
```

```
##
##     Alabama  California    Colorado     Florida     Georgia       Maine
##           1          27          32          49           1           1
## mississippi    New York      Oregon        Utah     Vermont     Virgina
##           1           1           4          12           1           1
##    virginia    Virginia   virginia
##           3          14           1
```

```
str(StudentFileDF)
```

```
## 'data.frame':    149 obs. of  9 variables:
##  $ N_ID    : int  3 4 7 10 13 18 19 20 22 23 ...
##  $ DateSub : chr  "9/13/2017" "9/20/2017" "10/4/2017" "10/7/2017" ...
##  $ Gender  : chr  "Female" "Female" "Male" "Male" ...
##  $ State   : chr  "California" "Florida" "California" "Colorado" ...
##  $ GPA     : num  3.9 3.8 3.8 3.6 3.92 3.8 3.88 3.7 3.9 3.7 ...
##  $ WorkExp : num  6.7 1.4 2.3 0.9 1.2 1.2 NA 1.2 4.7 1.4 ...
##  $ MathTest: int  962 969 970 969 969 967 967 969 961 966 ...
##  $ Essay   : num  NA 97 NA NA 95 NA NA NA NA 94 ...
##  $ Decision: chr  "Admit" "Admit" "Admit" "Admit" ...
```

```
## Change date format ------------------------------------------
```

```
## Change N_ID into a factor
StudentFileDF$N_ID<- as.factor(StudentFileDF$N_ID)
```

```
## Update and categorize the dates
str(StudentFileDF$DateSub)
```

```
##  chr [1:149] "9/13/2017" "9/20/2017" "10/4/2017" "10/7/2017" ...
```

```
(dates <- as.Date(StudentFileDF$DateSub, "%m/%d/%Y") )
```

```
##   [1] "2017-09-13" "2017-09-20" "2017-10-04" "2017-10-07" "2017-11-03"
##   [6] "2017-11-18" "2017-11-19" "2017-12-08" "2017-10-25" "2017-12-26"
##  [11] "2017-11-27" "2018-11-02" "2017-10-21" "2018-12-03" "2018-11-07"
##  [16] "2017-12-25" "2018-01-06" "2018-11-07" "2017-12-30" "2018-01-10"
```

```
##  [21] "2018-01-10" "2017-12-24" "2017-10-31" "2018-01-10" "2018-01-11"
##  [26] "2017-12-15" "2017-12-28" "2018-01-11" "2018-01-12" "2017-11-03"
##  [31] "2017-12-30" "2018-01-14" "2018-01-14" "2018-01-15" "2018-01-15"
##  [36] "2018-01-15" "2018-01-15" "2017-12-04" "2017-12-24" "2018-01-12"
##  [41] "2018-01-12" "2018-01-12" "2018-01-14" "2018-01-14" "2018-01-14"
##  [46] "2018-01-14" "2018-01-14" "2018-01-15" "2018-01-15" "2018-01-15"
##  [51] "2018-01-15" "2018-01-09" "2018-01-10" "2018-01-13" "2018-01-14"
##  [56] "2018-01-14" "2018-01-14" "2018-01-14" "2018-01-14" "2018-01-15"
##  [61] "2018-01-15" "2018-01-15" "2018-01-15" "2018-01-15" "2018-01-15"
##  [66] "2018-01-15" "2018-01-14" "2018-01-14" "2018-01-15" "2018-01-06"
##  [71] "2018-01-15" "2018-01-23" "2018-01-13" "2018-01-14" "2018-01-14"
##  [76] "2018-01-03" "2018-01-14" "2017-09-02" "2017-11-09" "2017-11-15"
##  [81] "2017-11-17" "2017-12-11" "2017-12-14" "2017-12-09" "2017-12-05"
##  [86] "2018-01-07" "2018-01-07" "2018-01-05" "2017-11-20" "2017-12-22"
##  [91] "2018-01-13" "2018-01-15" "2018-01-13" "2018-01-14" "2018-01-15"
##  [96] "2017-12-23" "2018-03-15" "2018-02-16" "2017-12-21" "2018-03-10"
## [101] "2018-02-14" "2018-01-15" "2018-02-03" "2018-01-13" "2018-03-15"
## [106] "2018-01-29" "2017-11-12" "2017-11-07" "2017-12-01" "2017-11-23"
## [111] "2017-12-21" "2017-12-26" "2017-12-29" "2018-01-06" "2018-01-10"
## [116] "2017-11-06" "2018-01-11" "2018-01-12" "2018-01-15" "2018-01-15"
## [121] "2018-01-12" "2018-01-13" "2018-01-13" "2018-01-14" "2018-01-14"
## [126] "2018-01-14" "2018-01-14" "2018-01-14" "2018-01-15" "2018-01-12"
## [131] "2018-01-14" "2018-01-14" "2018-01-14" "2018-01-15" "2018-01-17"
## [136] "2018-01-13" "2018-01-12" "2018-01-15" "2018-02-16" "2018-01-14"
## [141] "2017-12-16" "2018-01-10" "2018-01-11" "2018-01-12" "2018-01-15"
## [146] "2018-01-09" "2018-01-12" "2018-01-15" "2018-01-15"
```

```
StudentFileDF$DateSub <- dates
(head(StudentFileDF))
```

```
##   N_ID    DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 1    3 2017-09-13 Female California 3.90     6.7      962    NA    Admit
## 2    4 2017-09-20 Female    Florida 3.80     1.4      969    97    Admit
## 3    7 2017-10-04   Male California 3.80     2.3      970    NA    Admit
## 4   10 2017-10-07   Male   Colorado 3.60     0.9      969    NA    Admit
## 5   13 2017-11-03   Male   Colorado 3.92     1.2      969    95    Admit
## 6   18 2017-11-18   Male California 3.80     1.2      967    NA    Admit
```

```
#Change to month names
StudentFileDF$DateSub <- months(as.Date(StudentFileDF$DateSub))
#StudentFileDF$DateSub <- as.factor(StudentFileDF$DateSub)
(table(StudentFileDF$DateSub))
```

```
##
##  December  February   January     March  November   October September
##        23         4        95         3        16         5         3
```

```
## Here, you can see the month(s) with most/more summer program app submissio
ns
```

```
(table(StudentFileDF$Essay))
```

```
##
##  -2 0.3  11  56  65  69  70  71  73  74  75  78  79  80  81  82  83  84
##   1   1   1   2   1   3   1   2   1   2   1   3   2   2   4   2   3   4
##  85  86  87  88  89  90  91  93  94  95  96  97  98  99 100
##   2   1   3   2   2   1   5   5   3   4   1   5   2   4   2
```

```
## From looking at the table, we can see that there are some
## Essay scores that are odd. There is a -2, a .3, and an 11.
## Given that all other score are 56 and above, these three
## are likely errors....we will correct them first with NA and
## then later with another option.
StudentFileDF$Essay[StudentFileDF$Essay < 40] <- NA
```

```
## Let's look at the States from which students mostly apply....
(table(StudentFileDF$State))
```

```
##
##     Alabama  California    Colorado     Florida     Georgia       Maine
##           1          27          32          49           1           1
## mississippi    New York      Oregon        Utah     Vermont     Virgina
##           1           1           4          12           1           1
##    virginia    Virginia   virginia
##           3          14           1
```

```
## From this table, we can see that we have "Virginia" in three versions.
## We need to combine these. We also have a mis-spell...
StudentFileDF$State[StudentFileDF$State == "virginia"] <- "Virginia"
#(StudentFileDF$State)
StudentFileDF$State[StudentFileDF$State == "Virgina"] <- "Virginia"
StudentFileDF$State[StudentFileDF$State == "virginia "] <- "Virginia"
```

```
## Check it now...
table(StudentFileDF$State)
```

```
##
##     Alabama  California    Colorado     Florida     Georgia       Maine
##           1          27          32          49           1           1
## mississippi    New York      Oregon        Utah     Vermont    Virginia
##           1           1           4          12           1          19
```

```
## Our next goal is to create Groups. We can see that most apps are
## from California, Colorado, Florida, Utah and  Virginia
## Let's keep these 5 and create an "Other"
(MyList <- unique(StudentFileDF$State))
```

```
##  [1] "California"  "Florida"     "Colorado"    "Utah"        "Virginia"
##  [6] "Oregon"      "mississippi" "New York"    "Alabama"     "Georgia"
## [11] "Vermont"     "Maine"
```

```
MyList <-MyList[-c(1,2,3,4,5)]  ## Remove Cali, Colordao, FL, Utah, VA
#(MyList)
## Now, re-label all remaining states as "Other"
StudentFileDF$State[StudentFileDF$State %in% MyList] <- "Other"
(table(StudentFileDF$State))

##
## California   Colorado    Florida      Other       Utah   Virginia
##         27         32         49         10         12         19

## Remove Student ID ## NOTICE - these IDs are NOT REAL and
## are not associated with any students in real life. They are
## invented and pretend numbers.
#(head(StudentFileDF))
StudentFileDF = StudentFileDF[,-c(1)]
(head(StudentFileDF))

##     DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 1 September Female California 3.90     6.7      962    NA    Admit
## 2 September Female    Florida 3.80     1.4      969    97    Admit
## 3   October   Male California 3.80     2.3      970    NA    Admit
## 4   October   Male   Colorado 3.60     0.9      969    NA    Admit
## 5  November   Male   Colorado 3.92     1.2      969    95    Admit
## 6  November   Male California 3.80     1.2      967    NA    Admit

## CHange "Wait List" to "Waitlist"
StudentFileDF$Decision[StudentFileDF$Decision == "Wait List"] <- "Waitlist"
#(StudentFileDF)

## CHange Work experience to 0 if NA
str(StudentFileDF$WorkExp)

##  num [1:149] 6.7 1.4 2.3 0.9 1.2 1.2 NA 1.2 4.7 1.4 ...

StudentFileDF$WorkExp[is.na(StudentFileDF$WorkExp)] <- 0
(sum(is.na(StudentFileDF$WorkExp)))   ## No NA's should be left here

## [1] 0

## It is OK to set WorkExp to 0 if its NA as this is as likely as
## using the average.

## Let's look at the MathTest Scores next...
(table(StudentFileDF$MathTest))

##
## 742 751 753 754 757 758 761 762 763 764 765 766 767 768 769 799 853 855
##   1   1   1   1   1   1   1   1   2   2   1   3   1   5   6   1   1   1
## 859 862 863 864 865 866 867 868 869 952 956 957 959 960 961 962 963 964
##   1   1   1   2   3   9   8   6  11   1   1   2   1   1   2   2   2   1
## 965 966 967 968 969 970
##   4  10  12   9  27   1
```

```r
(sum(is.na(StudentFileDF$MathTest)))
```

```
## [1] 0
```

```r
## So - we have no NAs - this is good.
## We also have seemingly normal scores - nothing far outside

## Look at GPA next.
(table(StudentFileDF$GPA))
```

```
##
## 2.34 2.77 2.81 2.85  2.9 2.91 2.98    3 3.01  3.1 3.11 3.12 3.18 3.21 3.22
##    1    1    1    1    1    1    1    1    1    1    2    1    1    2    1
## 3.24 3.29 3.32 3.33 3.34 3.35 3.37 3.38 3.39  3.4 3.41 3.42 3.43 3.44 3.45
##    1    1    1    2    2    1    1    1    3    2    3    2    1    3    5
## 3.46 3.47 3.48 3.49  3.5 3.51 3.52 3.53 3.54 3.55 3.56 3.57 3.58 3.59  3.6
##    1    1    1    3    3    2    2    3    5    3    7    2    2    5    2
## 3.61 3.62 3.64 3.65 3.66 3.67 3.69  3.7 3.71 3.74 3.75 3.77 3.78 3.79  3.8
##    3    1    1    2    3    1    3    4    2    1    1    4    6    1    4
## 3.84 3.86 3.87 3.88 3.89  3.9 3.91 3.92 3.93 3.94 3.97
##    1    1    3    3    3    4    1    3    1    1    1
```

```r
## Any NAs?  If so - how many...
(sum(is.na(StudentFileDF$GPA)))
```

```
## [1] 2
```

```r
## There are 2 NA's. Let's replace them with the median GPA
(MedGPA <- median(StudentFileDF$GPA, na.rm=TRUE))
```

```
## [1] 3.56
```

```r
StudentFileDF$GPA[is.na(StudentFileDF$GPA)] <- MedGPA
## It is not an easy choice to replace the missing GPA with the median.
## GPA may play an important role. Here, we only have two NAs. We
## could also just remove the rows.

## IMPORTANT ##
## Notice that I am NOT replacing with 0 in most cases.
## Using "0" can have a critical affect and so be careful when using it.

## Change all "WL to Admit" into "Waitlist"
StudentFileDF$Decision[StudentFileDF$Decision == "WL to Admit"] <- "Waitlist"
(StudentFileDF)
```

```
##       DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 1   September Female California 3.90     6.7      962    NA    Admit
## 2   September Female    Florida 3.80     1.4      969    97    Admit
## 3     October   Male California 3.80     2.3      970    NA    Admit
## 4     October   Male   Colorado 3.60     0.9      969    NA    Admit
## 5    November   Male   Colorado 3.92     1.2      969    95    Admit
## 6    November   Male California 3.80     1.2      967    NA    Admit
```

```
## 7    November Female California 3.88     0.0      967    NA    Admit
## 8    December Female California 3.70     1.2      969    NA    Admit
## 9     October Female    Florida 3.90     4.7      961    NA    Admit
## 10   December Female California 3.70     1.4      966    94    Admit
## 11   November Female    Florida 3.56     1.7      968    91    Admit
## 12   November Female    Florida 3.93     0.8      969    NA    Admit
## 13    October Female   Colorado 3.60     1.2      967    94    Admit
## 14   December   Male California 3.69     3.2      967    93    Admit
## 15   November   Male    Florida 3.70     3.7      969    99    Admit
## 16   December Female   Colorado 3.90     0.0      967    NA    Admit
## 17    January   Male   Colorado 3.78     1.2      966   100    Admit
## 18   November   Male California 3.70     2.7      799    97    Admit
## 19   December   Male    Florida 3.50     0.7      965    NA    Admit
## 20    January   Male   Colorado 3.65     1.7      963    NA    Admit
## 21    January Female   Colorado 3.75     1.1      969    NA    Admit
## 22   December Female   Colorado 3.58     0.8      969    93    Admit
## 23    October   Male California 3.78     8.7      966    91    Admit
## 24    January Female California 3.92     2.8      967    95    Admit
## 25    January Female    Florida 3.54     0.7      965    NA    Admit
## 26   December   Male    Florida 3.66     2.2      967    91    Admit
## 27   December Female    Florida 3.90     0.0      967    88    Admit
## 28    January   Male    Florida 3.55     0.0      962    97    Admit
## 29    January Female   Colorado 3.59     1.7      969    93    Admit
## 30   November   Male California 3.66     0.9      956    89    Admit
## 31   December Female       Utah 3.78     1.2      968    87    Admit
## 32    January   Male   Colorado 3.88     1.0      969    93    Admit
## 33    January Female    Florida 3.80     1.9      965    94    Admit
## 34    January   Male    Florida 3.77     1.4      969    99    Admit
## 35    January Female California 3.87     1.7      966    97    Admit
## 36    January Female   Virginia 3.65     1.0      966    NA    Admit
## 37    January   Male   Virginia 3.89     0.7      966    NA    Admit
## 38   December Female   Virginia 3.59     1.8      966    NA    Admit
## 39   December Female California 3.78     4.7      952    99    Admit
## 40    January   Male   Virginia 3.87     8.2      957    NA    Admit
## 41    January   Male   Colorado 3.56     1.7      969    91    Admit
## 42    January   Male California 3.66     1.0      963    NA    Admit
## 43    January Female    Florida 3.87     1.8      968   100    Admit
## 44    January Female      Other 3.77     1.7      969    NA    Admit
## 45    January   Male    Florida 3.56     1.1      968    95    Admit
## 46    January   Male    Florida 3.51     2.9      964    95    Admit
## 47    January Female   Virginia 3.61     0.0      967    NA    Admit
## 48    January Female California 3.77     1.4      969    96    Admit
## 49    January Female   Virginia 3.59     0.7      967    NA    Admit
## 50    January Female   Colorado 3.71     1.1      969    NA    Admit
## 51    January Female California 3.64     3.2      969    NA    Admit
## 52    January   Male California 3.91     1.2      969    NA    Admit
## 53    January Female California 3.88     0.0      969    NA    Admit
## 54    January   Male    Florida 3.56     3.7      966    NA  Decline
## 55    January Female   Virginia 3.78     2.7      957    NA    Admit
## 56    January   Male    Florida 3.59     1.2      968    NA    Admit
```

```
## 57    January Female    Florida 3.79     1.4      969    NA  Decline
## 58    January   Male California 3.97     3.7      968    91    Admit
## 59    January   Male   Colorado 3.89     1.7      969    NA    Admit
## 60    January Female   Colorado 3.77     1.2      967    NA    Admit
## 61    January Female   Colorado 3.78     1.7      969    90    Admit
## 62    January   Male   Virginia 3.56     1.2      969    NA    Admit
## 63    January Female   Virginia 3.67     2.7      969    97    Admit
## 64    January   Male    Florida 3.71     0.7      968    98  Decline
## 65    January Female California 3.62     2.7      961    NA    Admit
## 66    January Female      Other 3.69     1.2      968    NA    Admit
## 67    January   Male   Virginia 3.92     1.3      966    99    Admit
## 68    January Female    Florida 3.74     1.3      966    NA  Decline
## 69    January Female   Colorado 3.86     1.2      967    98    Admit
## 70    January   Male   Virginia 3.69     5.7      968    89    Admit
## 71    January   Male   Virginia 3.45     0.7      969    NA    Admit
## 72    January Female   Virginia 3.89     1.5      969    87    Admit
## 73    January Female    Florida 3.61     1.3      959    NA  Decline
## 74    January Female    Florida 3.54     0.9      969    NA  Decline
## 75    January Female      Other 3.94     0.9      965    NA    Admit
## 76    January   Male   Colorado 3.59     1.4      969    93    Admit
## 77    January Female    Florida 3.84     2.7      960    NA  Decline
## 78  September   Male    Florida 2.81     9.2      764    NA  Decline
## 79   November Female California 2.34     0.8      754    75  Decline
## 80   November Female      Other 2.90     0.9      769    56  Decline
## 81   November Female    Florida 3.33     1.6      766    NA  Decline
## 82   December   Male   Virginia 3.37     0.9      766    NA  Decline
## 83   December Female   Colorado 3.00     1.2      768    56  Decline
## 84   December   Male    Florida 3.10     1.9      751    NA  Decline
## 85   December   Male   Virginia 3.22     3.2      769    78  Decline
## 86    January   Male    Florida 3.54     1.1      767    65  Decline
## 87    January Female   Colorado 3.44     3.2      757    NA  Decline
## 88    January   Male   Colorado 2.98     0.7      763    71  Decline
## 89   November Female   Virginia 2.77     3.7      763    NA  Decline
## 90   December   Male    Florida 3.18     1.4      768    NA  Decline
## 91    January Female   Colorado 3.11     1.7      758    69  Decline
## 92    January   Male   Colorado 3.32     1.7      768    78  Decline
## 93    January   Male       Utah 3.01     1.4      769    69  Decline
## 94    January   Male       Utah 3.33     0.8      768    NA  Decline
## 95    January Female      Other 2.91     6.2      753    NA  Decline
## 96   December Female    Florida 3.56     1.7      769    81  Decline
## 97      March Female   Colorado 2.85     4.6      762    NA  Decline
## 98   February Female   Virginia 3.21     1.7      766    79  Decline
## 99   December Female    Florida 3.38     0.7      768    NA  Decline
## 100     March Female    Florida 3.35     4.2      764    69  Decline
## 101  February Female   Virginia 3.11     8.9      742    NA  Decline
## 102   January Female   Virginia 3.12     0.0      761    NA  Decline
## 103  February   Male   Colorado 3.21     0.0      765    NA  Decline
## 104   January   Male   Colorado 3.34     3.7      769    NA  Decline
## 105     March Female    Florida 3.24     1.7      769    74  Decline
## 106   January   Male    Florida 3.40     1.9      859    NA Waitlist
```

```
## 107  November   Male    Florida 3.45     4.7      867    71 Waitlist
## 108  November   Male    Florida 3.50     1.7      869    73 Waitlist
## 109  December Female      Other 3.55     2.2      866    74 Waitlist
## 110  November   Male      Other 3.41     1.2      868    85 Waitlist
## 111  December   Male      Other 3.56     0.9      866    NA Waitlist
## 112  December Female    Florida 3.53     1.7      869    NA Waitlist
## 113  December   Male California 3.42     0.7      869    84 Waitlist
## 114   January   Male   Colorado 3.50     3.5      869    83 Waitlist
## 115   January Female       Utah 3.39     1.8      866    82 Waitlist
## 116  November Female California 3.52     2.7      855    NA Waitlist
## 117   January   Male   Colorado 3.49     1.3      866    NA Waitlist
## 118   January Female   Colorado 3.43     1.5      869    NA Waitlist
## 119   January   Male    Florida 3.44     7.2      865    NA Waitlist
## 120   January Female    Florida 3.29     1.2      869    NA Waitlist
## 121   January Female       Utah 3.58     0.9      864    81 Waitlist
## 122   January Female       Utah 3.57     1.4      869    80 Waitlist
## 123   January Female    Florida 3.56     1.3      869    84 Waitlist
## 124   January   Male    Florida 3.55     2.0      853    NA Waitlist
## 125   January   Male   Colorado 3.54     1.2      868    83 Waitlist
## 126   January Female      Other 3.53     3.3      862    NA Waitlist
## 127   January   Male    Florida 3.52     0.7      868    81 Waitlist
## 128   January Female    Florida 3.51     3.4      865    88 Waitlist
## 129   January Female California 3.47     2.2      867    NA Waitlist
## 130   January Female    Florida 3.46     1.9      869    NA Waitlist
## 131   January   Male    Florida 3.45     0.7      866    NA Waitlist
## 132   January Female California 3.44     0.7      867    83 Waitlist
## 133   January   Male California 3.42     1.7      866    NA Waitlist
## 134   January   Male       Utah 3.41     1.4      869    81 Waitlist
## 135   January   Male       Utah 3.40     1.2      868    80 Waitlist
## 136   January   Male    Florida 3.39     1.2      866    NA Waitlist
## 137   January Female   Colorado 3.41     2.7      866    79 Waitlist
## 138   January   Male   Colorado 3.45     1.7      869    78 Waitlist
## 139  February Female   Colorado 3.49     2.7      866    NA Waitlist
## 140   January   Male    Florida 3.53     0.7      864    70 Waitlist
## 141  December Female    Florida 3.45     1.7      867    87 Waitlist
## 142   January   Male California 3.54     4.2      865    NA Waitlist
## 143   January Female       Utah 3.34     2.3      867    NA Waitlist
## 144   January Female      Other 3.61     0.8      867    84 Waitlist
## 145   January   Male       Utah 3.48     2.3      867    85 Waitlist
## 146   January   Male    Florida 3.39     4.2      863    86 Waitlist
## 147   January Female       Utah 3.49     0.7      868    82 Waitlist
## 148   January Female       Utah 3.56     1.4      867    84 Waitlist
## 149   January   Male    Florida 3.57     1.0      868    NA Waitlist

## Let's look at the Essay scores next...
(table(StudentFileDF$Essay))

##
##  56  65  69  70  71  73  74  75  78  79  80  81  82  83  84  85  86  87
##   2   1   3   1   2   1   2   1   3   2   2   4   2   3   4   2   1   3
```

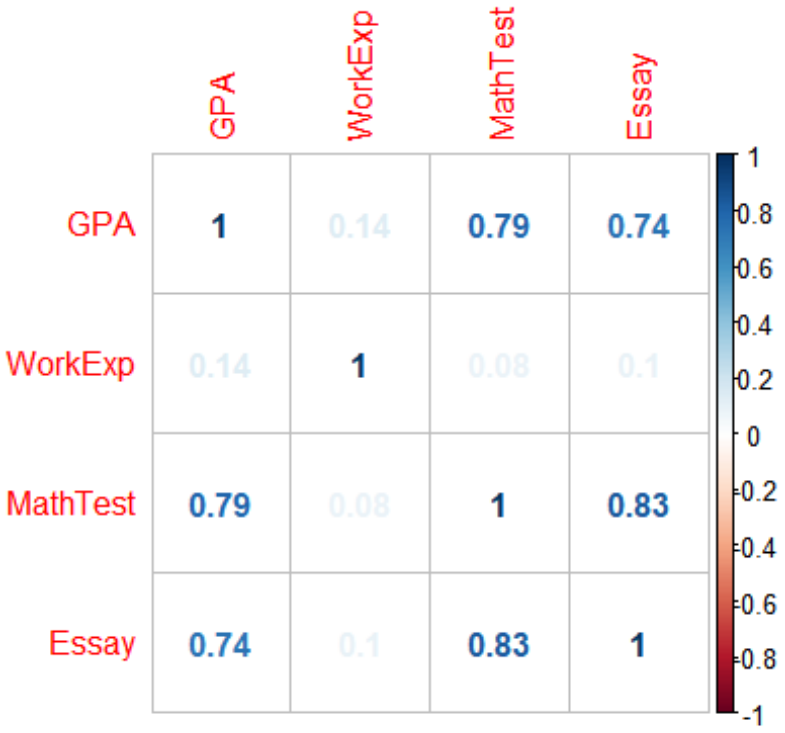
```
##  88  89  90  91  93  94  95  96  97  98  99 100
##   2   2   1   5   5   3   4   1   5   2   4   2

(sum(is.na(StudentFileDF$Essay)))

## [1] 74

## Yikes! We have 74 NAs in the Essay column. This suggests that the Essay
## may have been optional. For this reason, we have several choices. We can r
emove
## the column. We can fill in with the mean. We can create a new category "No
tIncluded".
## Below, I have chosen to create a new category called "NotIncluded". This w
ay
## we do not lose the rows and we do not alter the dataset.

## Make sure that the Decision class is a factor
StudentFileDF$Decision <- as.factor(StudentFileDF$Decision)

## Check it current state and create a new DF as a copy
CleanStudentDF <- StudentFileDF
(head(CleanStudentDF, n = 20))

##      DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 1  September Female California 3.90     6.7      962    NA    Admit
## 2  September Female    Florida 3.80     1.4      969    97    Admit
## 3    October   Male California 3.80     2.3      970    NA    Admit
## 4    October   Male   Colorado 3.60     0.9      969    NA    Admit
## 5   November   Male   Colorado 3.92     1.2      969    95    Admit
## 6   November   Male California 3.80     1.2      967    NA    Admit
## 7   November Female California 3.88     0.0      967    NA    Admit
## 8   December Female California 3.70     1.2      969    NA    Admit
## 9    October Female    Florida 3.90     4.7      961    NA    Admit
## 10  December Female California 3.70     1.4      966    94    Admit
## 11  November Female    Florida 3.56     1.7      968    91    Admit
## 12  November Female    Florida 3.93     0.8      969    NA    Admit
## 13   October Female   Colorado 3.60     1.2      967    94    Admit
## 14  December   Male California 3.69     3.2      967    93    Admit
## 15  November   Male    Florida 3.70     3.7      969    99    Admit
## 16  December Female   Colorado 3.90     0.0      967    NA    Admit
## 17   January   Male   Colorado 3.78     1.2      966   100    Admit
## 18  November   Male California 3.70     2.7      799    97    Admit
## 19  December   Male    Florida 3.50     0.7      965    NA    Admit
## 20   January   Male   Colorado 3.65     1.7      963    NA    Admit

str(CleanStudentDF)

## 'data.frame':    149 obs. of  8 variables:
##  $ DateSub : chr  "September" "September" "October" "October" ...
##  $ Gender  : chr  "Female" "Female" "Male" "Male" ...
##  $ State   : chr  "California" "Florida" "California" "Colorado" ...
```

```
##  $ GPA     : num  3.9 3.8 3.8 3.6 3.92 3.8 3.88 3.7 3.9 3.7 ...
##  $ WorkExp : num  6.7 1.4 2.3 0.9 1.2 1.2 0 1.2 4.7 1.4 ...
##  $ MathTest: int  962 969 970 969 969 967 967 969 961 966 ...
##  $ Essay   : num  NA 97 NA NA 95 NA NA NA NA 94 ...
##  $ Decision: Factor w/ 3 levels "Admit","Decline",..: 1 1 1 1 1 1 1 1 1 1
...

## Change Gender and State to factors
CleanStudentDF$Gender <- as.factor(CleanStudentDF$Gender)
CleanStudentDF$State <- as.factor(CleanStudentDF$State)
str(CleanStudentDF)

## 'data.frame':    149 obs. of  8 variables:
##  $ DateSub : chr  "September" "September" "October" "October" ...
##  $ Gender  : Factor w/ 2 levels "Female","Male": 1 1 2 2 2 2 1 1 1 1 ...
##  $ State   : Factor w/ 6 levels "California","Colorado",..: 1 3 1 2 2 1 1
1 3 1 ...
##  $ GPA     : num  3.9 3.8 3.8 3.6 3.92 3.8 3.88 3.7 3.9 3.7 ...
##  $ WorkExp : num  6.7 1.4 2.3 0.9 1.2 1.2 0 1.2 4.7 1.4 ...
##  $ MathTest: int  962 969 970 969 969 967 967 969 961 966 ...
##  $ Essay   : num  NA 97 NA NA 95 NA NA NA NA 94 ...
##  $ Decision: Factor w/ 3 levels "Admit","Decline",..: 1 1 1 1 1 1 1 1 1 1
...

###################### Visual and Summary EDA ########################
## Exploratory Data Analysis
##
######################################################################
(head(CleanStudentDF, n = 10))

##      DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 1  September Female California 3.90     6.7      962    NA    Admit
## 2  September Female    Florida 3.80     1.4      969    97    Admit
## 3    October   Male California 3.80     2.3      970    NA    Admit
## 4    October   Male   Colorado 3.60     0.9      969    NA    Admit
## 5   November   Male   Colorado 3.92     1.2      969    95    Admit
## 6   November   Male California 3.80     1.2      967    NA    Admit
## 7   November Female California 3.88     0.0      967    NA    Admit
## 8   December Female California 3.70     1.2      969    NA    Admit
## 9    October Female    Florida 3.90     4.7      961    NA    Admit
## 10  December Female California 3.70     1.4      966    94    Admit

(names(CleanStudentDF))

## [1] "DateSub"  "Gender"   "State"    "GPA"      "WorkExp"  "MathTest"
## [7] "Essay"    "Decision"

str(CleanStudentDF)

## 'data.frame':    149 obs. of  8 variables:
##  $ DateSub : chr  "September" "September" "October" "October" ...
##  $ Gender  : Factor w/ 2 levels "Female","Male": 1 1 2 2 2 2 1 1 1 1 ...
```

```
##  $ State   : Factor w/ 6 levels "California","Colorado",..: 1 3 1 2 2 1 1
1 3 1 ...
##  $ GPA     : num  3.9 3.8 3.8 3.6 3.92 3.8 3.88 3.7 3.9 3.7 ...
##  $ WorkExp : num  6.7 1.4 2.3 0.9 1.2 1.2 0 1.2 4.7 1.4 ...
##  $ MathTest: int  962 969 970 969 969 967 967 969 961 966 ...
##  $ Essay   : num  NA 97 NA NA 95 NA NA NA NA 94 ...
##  $ Decision: Factor w/ 3 levels "Admit","Decline",..: 1 1 1 1 1 1 1 1 1 1
...

### TABLES

(table(CleanStudentDF$Gender, CleanStudentDF$Decision))

##
##          Admit Decline Waitlist
##   Female    39      21       21
##   Male      31      14       23

(table(CleanStudentDF$State, CleanStudentDF$Decision))

##
##              Admit Decline Waitlist
##   California    20       1        6
##   Colorado      17       8        7
##   Florida       16      16       17
##   Other          3       2        5
##   Utah           1       2        9
##   Virginia      13       6        0

## The above tables are interesting. There does not appear to be
## any sig diff between genders.
## There does appear to be a larger Decline for FL.

## Create a small dataframe to view months of application for summer
## program and decision to join
MonthsDec <- data.frame(Months=CleanStudentDF$DateSub, DecisionMade=CleanStud
entDF$Decision)
MonthsDec$Months = factor(MonthsDec$Months, levels = month.name)
(table(MonthsDec$Months, MonthsDec$DecisionMade))

##
##             Admit Decline Waitlist
##   January      44      17       34
##   February      0       3        1
##   March         0       3        0
##   April         0       0        0
##   May           0       0        0
##   June          0       0        0
##   July          0       0        0
##   August        0       0        0
##   September     2       1        0
```

```
##   October       5       0        0
##   November      8       4        4
##   December     11       7        5

## Notice that Dec and Jan seem the most active for Admit.

EssayCategories <- data.frame(EssayRank=CleanStudentDF$Essay, Decision=CleanS
tudentDF$Decision)
(head(EssayCategories))

##   EssayRank Decision
## 1        NA    Admit
## 2        97    Admit
## 3        NA    Admit
## 4        NA    Admit
## 5        95    Admit
## 6        NA    Admit

EssayCategories$EssayRank <-
  cut(EssayCategories$EssayRank, breaks=c(-Inf, 60, 75, 88, Inf),
      labels=c("VeryLow","Low", "Med","High"))
EssayCategories$EssayRank <-fct_explicit_na(EssayCategories$EssayRank, "NotIn
cluded")
(table(CleanStudentDF$Decision,EssayCategories$EssayRank))

##
##            VeryLow Low Med High NotIncluded
##   Admit          0   0   3   33          34
##   Decline        2   7   4    1          21
##   Waitlist       0   4  21    0          19

## That looks good and makes sense.



## CORRELATION
## Correlation Matrix
(head(CleanStudentDF))

##     DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 1 September Female California 3.90     6.7      962    NA    Admit
## 2 September Female    Florida 3.80     1.4      969    97    Admit
## 3   October   Male California 3.80     2.3      970    NA    Admit
## 4   October   Male   Colorado 3.60     0.9      969    NA    Admit
## 5  November   Male   Colorado 3.92     1.2      969    95    Admit
## 6  November   Male California 3.80     1.2      967    NA    Admit

correlationMatrix <- cor(CleanStudentDF[,-c(1,2,3,8)],use="complete.obs")
(correlationMatrix)

##                GPA    WorkExp   MathTest      Essay
## GPA      1.0000000 0.13750313 0.79414644 0.73699644
```

```
## WorkExp  0.1375031 1.00000000 0.08285564 0.09535277
## MathTest 0.7941464 0.08285564 1.00000000 0.83490788
## Essay    0.7369964 0.09535277 0.83490788 1.00000000

#corrplot(correlationMatrix, method="circle")
#corrplot(correlationMatrix, method="color")
corrplot(correlationMatrix, method="number")
```




```
## Note: We can see that MathTest and GPA are strongly
## correlated. This makes sense. When we do Naive Bayes -
## because we assume independence, we should not use
## both measures.
## Essay is also highly correlated with GPA and with MathTest.

## SUMMARY STATS
(summary(CleanStudentDF))

##    DateSub             Gender          State         GPA
##  Length:149         Female:81   California:27   Min.   :2.340
##  Class :character   Male  :68   Colorado  :32   1st Qu.:3.430
##  Mode  :character               Florida   :49   Median :3.560
##                                 Other     :10   Mean   :3.546
##                                 Utah      :12   3rd Qu.:3.750
##                                 Virginia  :19   Max.   :3.970
##
##     WorkExp         MathTest         Essay            Decision
##  Min.   :0.000   Min.   :742.0   Min.   : 56.00   Admit   :70
```

```
##  1st Qu.:1.000   1st Qu.:865.0   1st Qu.: 80.00   Decline :35
##  Median :1.400   Median :956.0   Median : 87.00   Waitlist:44
##  Mean   :1.997   Mean   :897.5   Mean   : 85.77
##  3rd Qu.:2.300   3rd Qu.:967.0   3rd Qu.: 94.00
##  Max.   :9.200   Max.   :970.0   Max.   :100.00
##                                  NA's   :74
```

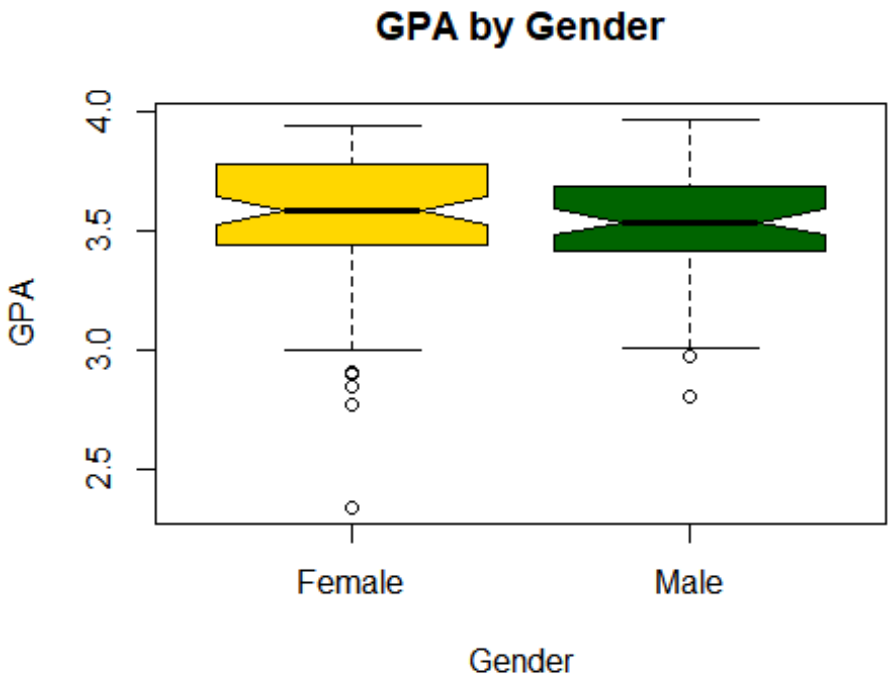
```
Stats <- round(stat.desc(CleanStudentDF[,-c(1,2,3,8)]))  #numerical only
Stats
```

```
##              GPA WorkExp MathTest Essay
## nbr.val      149     149      149    75
## nbr.null       0       8        0     0
## nbr.na         0       0        0    74
## min            2       0      742    56
## max            4       9      970   100
## range          2       9      228    44
## sum          528     298   133733  6433
## median         4       1      956    87
## mean           4       2      898    86
## SE.mean        0       0        6     1
## CI.mean.0.95   0       0       13     2
## var            0       3     6213   107
## std.dev        0       2       79    10
## coef.var       0       1        0     0
```

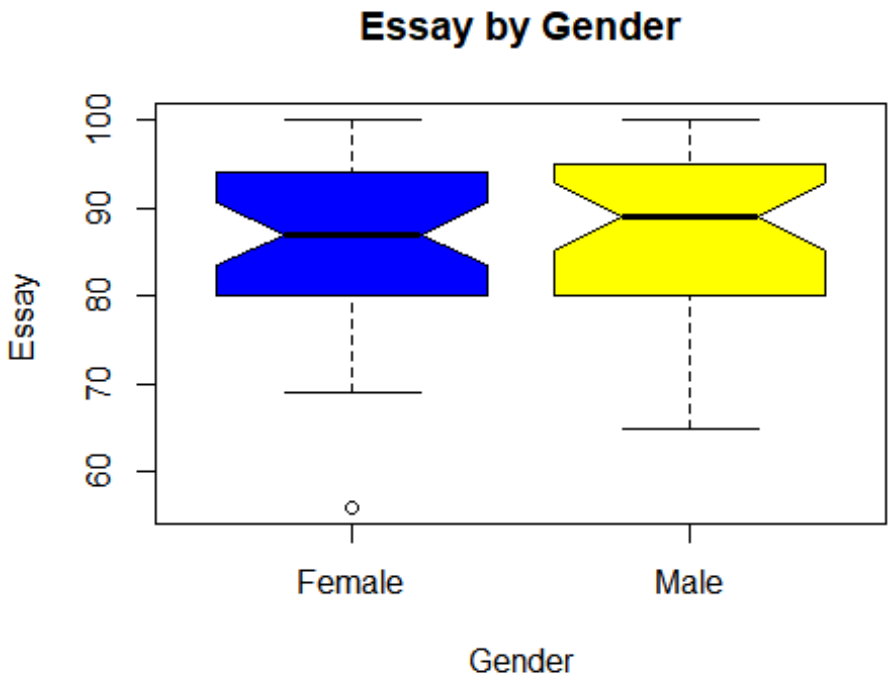
```
## Differences by Groups
str(CleanStudentDF$Gender)
```

```
##  Factor w/ 2 levels "Female","Male": 1 1 2 2 2 2 1 1 1 1 ...
```

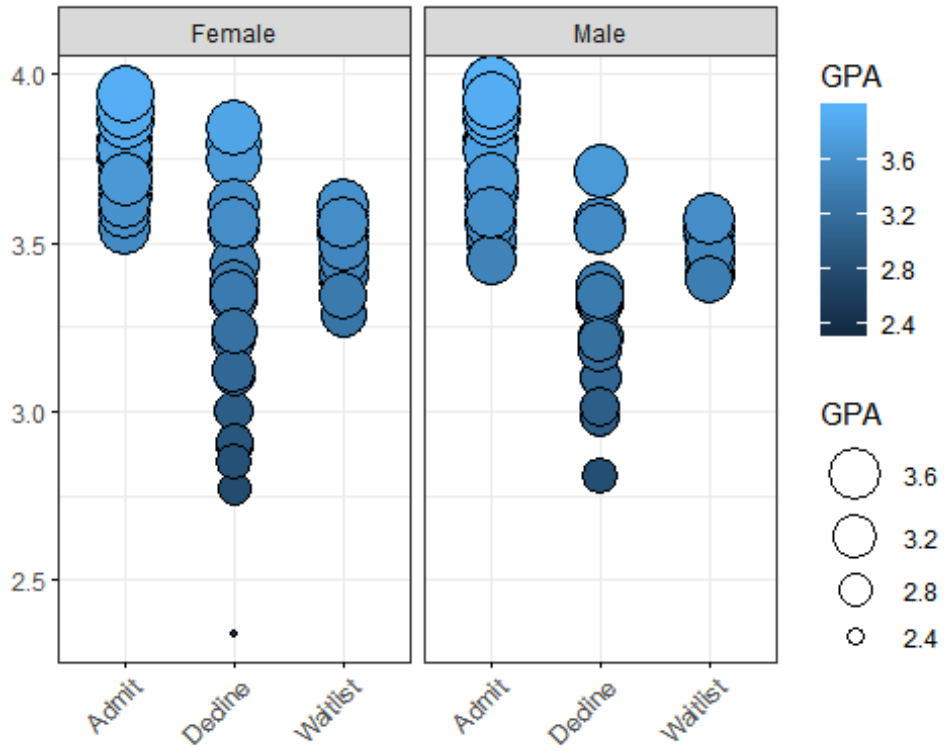
```
boxplot(GPA~Gender,data=CleanStudentDF, main="GPA by Gender",
        xlab="Gender", ylab="GPA",notch=TRUE,
        col=(c("gold","darkgreen")))
```




```
#(head(CleanStudentDF))
boxplot(Essay~Gender,data=CleanStudentDF, main="Essay by Gender",
        xlab="Gender", ylab="Essay ",notch=TRUE,
        col=(c("blue","yellow")))
```

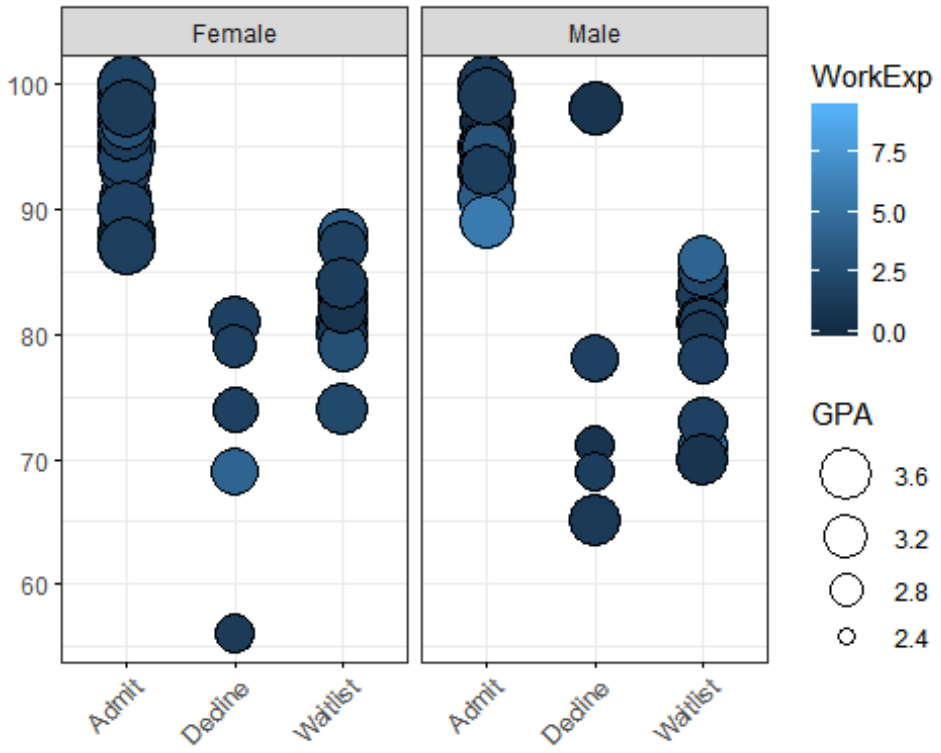



```
ggballoonplot(CleanStudentDF, x = "Decision", y = "GPA", size = "GPA",
              fill = "GPA", facet.by = "Gender",
              ggtheme = theme_bw())
```



```
ggballoonplot(CleanStudentDF, x = "Decision", y = "Essay", size = "GPA",
              fill = "WorkExp", facet.by = "Gender",
              ggtheme = theme_bw())
```

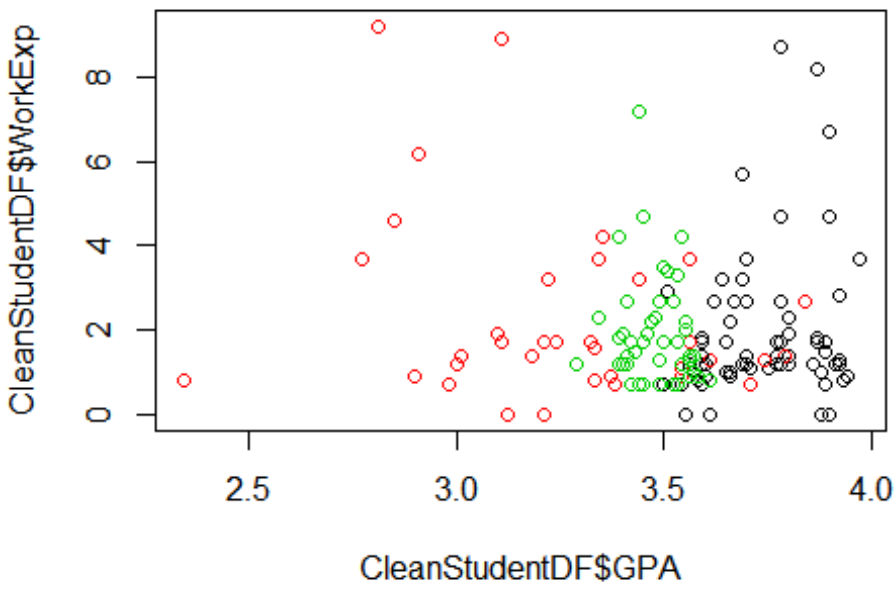
```
## Warning: Removed 74 rows containing missing values (geom_point).
```



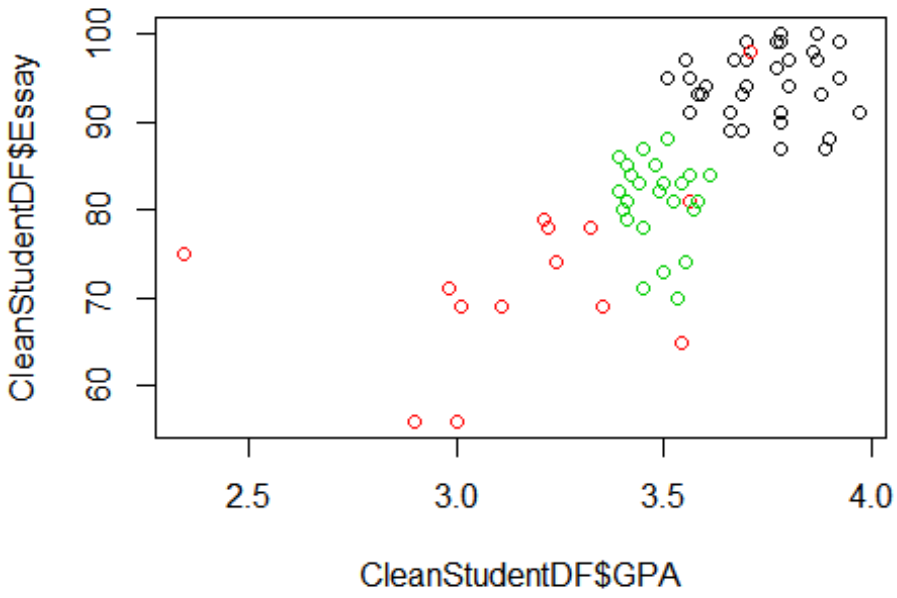

```
(head(CleanStudentDF))
```

```
##     DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 1 September Female California 3.90     6.7      962    NA    Admit
## 2 September Female    Florida 3.80     1.4      969    97    Admit
## 3   October   Male California 3.80     2.3      970    NA    Admit
## 4   October   Male   Colorado 3.60     0.9      969    NA    Admit
## 5  November   Male   Colorado 3.92     1.2      969    95    Admit
## 6  November   Male California 3.80     1.2      967    NA    Admit
```

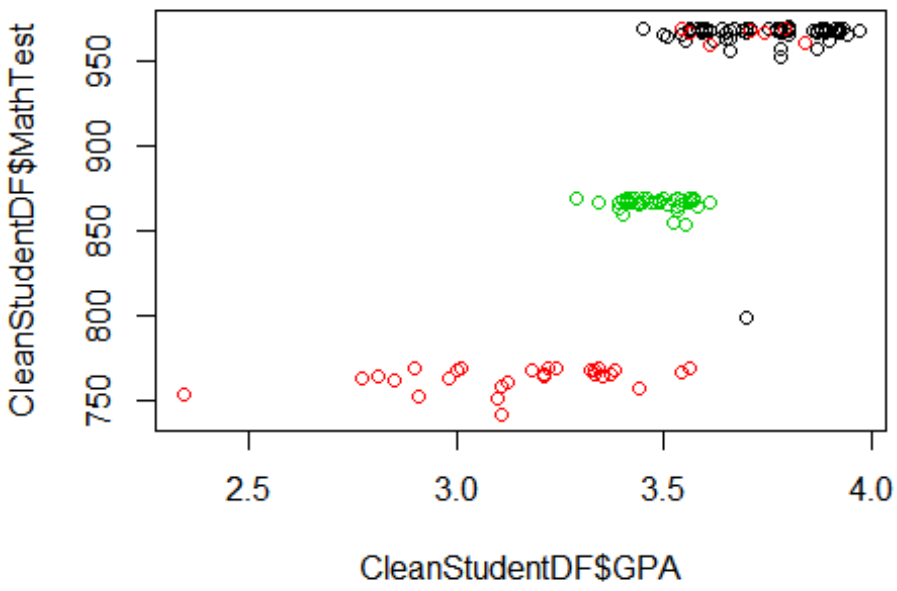
```
## Plots of Decision and Attributes
plot(CleanStudentDF$GPA,CleanStudentDF$WorkExp,
     col=CleanStudentDF$Decision)
```




```
## We are starting to see some clustering here.

## GPA, Essay, and Decision
plot(CleanStudentDF$GPA,CleanStudentDF$Essay,
     col=CleanStudentDF$Decision)
```




```
## This is very clustered and will be great for k-means!

## GPA, MathTest,and Decision
plot(CleanStudentDF$GPA,CleanStudentDF$MathTest,
     col=CleanStudentDF$Decision)
```




```
## Here, we see mostly what we expect - but also something *odd*
## We see some Declines that have high scores!
## Can this be based on the State?


######### Analysis ------------------------------------------------------
##
## Naive Bayes
## For NB - we can only use numerical data
## Create a NB_DF to use for Naive Bayes
## Also take away the Decision - which is the class
######### ------------------------------------------------------------------
-

## For Naive Bayes, we will work with the Cleaned Dataset
(head(CleanStudentDF,n=10))

##      DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 1  September Female California 3.90     6.7      962    NA    Admit
## 2  September Female    Florida 3.80     1.4      969    97    Admit
## 3    October   Male California 3.80     2.3      970    NA    Admit
## 4    October   Male   Colorado 3.60     0.9      969    NA    Admit
## 5   November   Male   Colorado 3.92     1.2      969    95    Admit
## 6   November   Male California 3.80     1.2      967    NA    Admit
## 7   November Female California 3.88     0.0      967    NA    Admit
## 8   December Female California 3.70     1.2      969    NA    Admit
```

```
## 9    October Female    Florida 3.90     4.7      961    NA    Admit
## 10  December Female California 3.70     1.4      966    94    Admit

## First, create a trainng and testing set and give them new names
## DO NOT change the CleanStudentDF as you may need it later.
## Always use copies or create new DFs
## To create the Testing Set, I will use very 7 values
## The Training Set will be all the remaining values
(every7_indexes<-seq(1,nrow(CleanStudentDF),7))

##  [1]   1   8  15  22  29  36  43  50  57  64  71  78  85  92  99 106 113
## [18] 120 127 134 141 148

NB_DF_Test=CleanStudentDF[every7_indexes, ]
NB_DF_Train=CleanStudentDF[-every7_indexes, ]
## View the created Test and Train sets
(head(NB_DF_Train,n=10))

##      DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 2  September Female    Florida 3.80     1.4      969    97    Admit
## 3    October   Male California 3.80     2.3      970    NA    Admit
## 4    October   Male   Colorado 3.60     0.9      969    NA    Admit
## 5   November   Male   Colorado 3.92     1.2      969    95    Admit
## 6   November   Male California 3.80     1.2      967    NA    Admit
## 7   November Female California 3.88     0.0      967    NA    Admit
## 9    October Female    Florida 3.90     4.7      961    NA    Admit
## 10  December Female California 3.70     1.4      966    94    Admit
## 11  November Female    Florida 3.56     1.7      968    91    Admit
## 12  November Female    Florida 3.93     0.8      969    NA    Admit

(head(NB_DF_Test,n=10))

##      DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 1  September Female California 3.90     6.7      962    NA    Admit
## 8   December Female California 3.70     1.2      969    NA    Admit
## 15  November   Male    Florida 3.70     3.7      969    99    Admit
## 22  December Female   Colorado 3.58     0.8      969    93    Admit
## 29   January Female   Colorado 3.59     1.7      969    93    Admit
## 36   January Female   Virginia 3.65     1.0      966    NA    Admit
## 43   January Female    Florida 3.87     1.8      968   100    Admit
## 50   January Female   Colorado 3.71     1.1      969    NA    Admit
## 57   January Female    Florida 3.79     1.4      969    NA  Decline
## 64   January   Male    Florida 3.71     0.7      968    98  Decline

## Notice that there are still some NAs in Essay
## We will have to deal with these....

## Naive Bayes works on numerical data ONLY
## Remove labels and nominal variables, etc.
NB_DF_Train_onlynums <- NB_DF_Train[-c(1,2,3)]
NB_DF_Test_onlynums <- NB_DF_Test[-c(1,2,3,8)]
```

```
NB_Student_TrainLABELS <-NB_DF_Train$Decision
NB_Student_TestLABELS <- NB_DF_Test$Decision

## Check what you have now...
(head(NB_DF_Train_onlynums, n=5))

##    GPA WorkExp MathTest Essay Decision
## 2 3.80     1.4      969    97    Admit
## 3 3.80     2.3      970    NA    Admit
## 4 3.60     0.9      969    NA    Admit
## 5 3.92     1.2      969    95    Admit
## 6 3.80     1.2      967    NA    Admit

(head(NB_DF_Test_onlynums, n=5))

##     GPA WorkExp MathTest Essay
## 1  3.90     6.7      962    NA
## 8  3.70     1.2      969    NA
## 15 3.70     3.7      969    99
## 22 3.58     0.8      969    93
## 29 3.59     1.7      969    93

## Now we will run the Naive Bayes (NB) classifier. We can do this two ways
## The first way will retain the Essay column and will skip (pass) the NAs
## The next way will not use the Essay column at all and so will have more da
ta
## but one less variable. Since the Essay is correlated to GPA - this is OK.

## WAY 1 NB
NBStudentclassfier <- naiveBayes(Decision ~.,data=NB_DF_Train_onlynums, na.ac
tion = na.pass)
NBStudentClassifier_Prediction <- predict(NBStudentclassfier, NB_DF_Test_only
nums)
NBStudentclassfier

##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##     Admit   Decline  Waitlist
## 0.4803150 0.2283465 0.2913386
##
## Conditional probabilities:
##           GPA
## Y             [,1]       [,2]
##   Admit    3.743770 0.13309350
##   Decline  3.227241 0.32076801
```

```
##   Waitlist 3.484865 0.06576987
##
##           WorkExp
## Y              [,1]     [,2]
##   Admit    1.839344 1.678519
##   Decline  2.203448 1.946881
##   Waitlist 2.075676 1.370402
##
##           MathTest
## Y              [,1]      [,2]
##   Admit    963.5738 21.720543
##   Decline  797.4828 77.628060
##   Waitlist 866.0541  3.431439
##
##           Essay
## Y              [,1]     [,2]
##   Admit    93.81250 3.710817
##   Decline  69.45455 8.128513
##   Waitlist 80.55000 5.041668

print(NBStudentClassifier_Prediction)

##  [1] Admit    Admit    Admit    Admit    Admit    Admit    Admit
##  [8] Admit    Admit    Admit    Admit    Decline  Decline  Decline
## [15] Decline  Waitlist Waitlist Waitlist Waitlist Waitlist Waitlist
## [22] Waitlist
## Levels: Admit Decline Waitlist

table(NBStudentClassifier_Prediction,NB_DF_Test$Decision)

##
## NBStudentClassifier_Prediction Admit Decline Waitlist
##                        Admit        9       2        0
##                        Decline      0       4        0
##                        Waitlist     0       0        7

plot(NBStudentClassifier_Prediction)
## This gave excellent results!

## WAY 2 NB
(head(NB_DF_Train_onlynums, n=5))

##    GPA WorkExp MathTest Essay Decision
## 2 3.80     1.4      969    97    Admit
## 3 3.80     2.3      970    NA    Admit
## 4 3.60     0.9      969    NA    Admit
## 5 3.92     1.2      969    95    Admit
## 6 3.80     1.2      967    NA    Admit

(head(NB_DF_Test_onlynums, n=5))
```

```
##     GPA WorkExp MathTest Essay
## 1  3.90     6.7      962    NA
## 8  3.70     1.2      969    NA
## 15 3.70     3.7      969    99
## 22 3.58     0.8      969    93
## 29 3.59     1.7      969    93
```

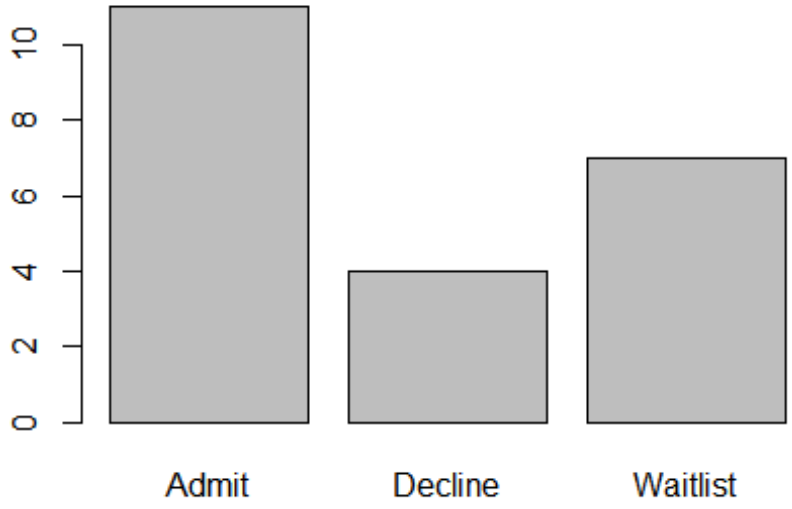
```
NB_DF_Train_onlynums_noEssay <- NB_DF_Train_onlynums[-4]
NB_DF_Test_onlynums_noEssay <- NB_DF_Test_onlynums[-4]
NBStudentclassfier2 <- naiveBayes(Decision ~.,data=NB_DF_Train_onlynums_noEss
ay, na.action = na.pass)
NBStudentClassifier_Prediction2 <- predict(NBStudentclassfier2, NB_DF_Test_on
lynums_noEssay)
NBStudentclassfier2
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##     Admit   Decline  Waitlist
## 0.4803150 0.2283465 0.2913386
##
## Conditional probabilities:
##           GPA
## Y              [,1]       [,2]
##   Admit    3.743770 0.13309350
##   Decline  3.227241 0.32076801
##   Waitlist 3.484865 0.06576987
##
##           WorkExp
## Y              [,1]     [,2]
##   Admit    1.839344 1.678519
##   Decline  2.203448 1.946881
##   Waitlist 2.075676 1.370402
##
##           MathTest
## Y              [,1]      [,2]
##   Admit    963.5738 21.720543
##   Decline  797.4828 77.628060
##   Waitlist 866.0541  3.431439
```

```
print(NBStudentClassifier_Prediction2)
```

```
##  [1] Admit    Admit    Admit    Admit    Admit    Admit    Admit
##  [8] Admit    Admit    Admit    Admit    Decline  Decline  Decline
## [15] Decline  Waitlist Waitlist Waitlist Waitlist Waitlist Waitlist
```

```
## [22] Waitlist
## Levels: Admit Decline Waitlist

table(NBStudentClassifier_Prediction2,NB_DF_Test$Decision)

##
## NBStudentClassifier_Prediction2 Admit Decline Waitlist
##                         Admit        9       2        0
##                         Decline      0       4        0
##                         Waitlist     0       0        7

plot(NBStudentClassifier_Prediction2)
```




```
## ------------------------------------------------------------------------
-
## Decision Tree Classification
## ------------------------------------------------------------------------
-
## Next, we will use Decision Trees to see if we can classify data
## by acceptance type : Admit, Decline, Waitlist
## Notice that I am always resorting back to my original(ish) dataframes
## so that I can adjust them as needed for each application.
## Recall that students who apply for this summer study abroad program
## can be accepted (Admitted), Declined, or placed on a waitlist.
## Unlike Naive Bayes, Decision Trees can run on discretized/categorical data
.
```

```
(head(NB_DF_Train,n=5))
```

```
##     DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 2 September Female    Florida 3.80     1.4      969    97    Admit
## 3   October   Male California 3.80     2.3      970    NA    Admit
## 4   October   Male   Colorado 3.60     0.9      969    NA    Admit
## 5  November   Male   Colorado 3.92     1.2      969    95    Admit
## 6  November   Male California 3.80     1.2      967    NA    Admit
```

```
(head(NB_DF_Test,n=5))
```

```
##      DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 1  September Female California 3.90     6.7      962    NA    Admit
## 8   December Female California 3.70     1.2      969    NA    Admit
## 15  November   Male    Florida 3.70     3.7      969    99    Admit
## 22  December Female   Colorado 3.58     0.8      969    93    Admit
## 29   January Female   Colorado 3.59     1.7      969    93    Admit
```

```
DT_Test <- NB_DF_Test[-c(8)] ## remove the Decision Class
DT_Test_Labels <- NB_DF_Test$Decision
DT_Train <- NB_DF_Train
(head(DT_Train,n=5))
```

```
##     DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 2 September Female    Florida 3.80     1.4      969    97    Admit
## 3   October   Male California 3.80     2.3      970    NA    Admit
## 4   October   Male   Colorado 3.60     0.9      969    NA    Admit
## 5  November   Male   Colorado 3.92     1.2      969    95    Admit
## 6  November   Male California 3.80     1.2      967    NA    Admit
```

```
(head(DT_Test,n=5))
```

```
##      DateSub Gender      State  GPA WorkExp MathTest Essay
## 1  September Female California 3.90     6.7      962    NA
## 8   December Female California 3.70     1.2      969    NA
## 15  November   Male    Florida 3.70     3.7      969    99
## 22  December Female   Colorado 3.58     0.8      969    93
## 29   January Female   Colorado 3.59     1.7      969    93
```

```
## -----------------------
## Step 1 - categorize GPA
DT_Train$GPAcategory <-
      cut(DT_Train$GPA, breaks=c(-Inf, 3.4, 3.7, Inf),
          labels=c("LowGPA","MedGPA","HighGPA"))

DT_Train$GPAcategory[is.na(DT_Train$GPA)] <- "MedGPA"
(DT_Train)
```

```
##       DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 2   September Female    Florida 3.80     1.4      969    97    Admit
## 3     October   Male California 3.80     2.3      970    NA    Admit
```

```
## 4     October   Male   Colorado 3.60     0.9      969    NA    Admit
## 5    November   Male   Colorado 3.92     1.2      969    95    Admit
## 6    November   Male California 3.80     1.2      967    NA    Admit
## 7    November Female California 3.88     0.0      967    NA    Admit
## 9     October Female    Florida 3.90     4.7      961    NA    Admit
## 10   December Female California 3.70     1.4      966    94    Admit
## 11   November Female    Florida 3.56     1.7      968    91    Admit
## 12   November Female    Florida 3.93     0.8      969    NA    Admit
## 13    October Female   Colorado 3.60     1.2      967    94    Admit
## 14   December   Male California 3.69     3.2      967    93    Admit
## 16   December Female   Colorado 3.90     0.0      967    NA    Admit
## 17    January   Male   Colorado 3.78     1.2      966   100    Admit
## 18   November   Male California 3.70     2.7      799    97    Admit
## 19   December   Male    Florida 3.50     0.7      965    NA    Admit
## 20    January   Male   Colorado 3.65     1.7      963    NA    Admit
## 21    January Female   Colorado 3.75     1.1      969    NA    Admit
## 23    October   Male California 3.78     8.7      966    91    Admit
## 24    January Female California 3.92     2.8      967    95    Admit
## 25    January Female    Florida 3.54     0.7      965    NA    Admit
## 26   December   Male    Florida 3.66     2.2      967    91    Admit
## 27   December Female    Florida 3.90     0.0      967    88    Admit
## 28    January   Male    Florida 3.55     0.0      962    97    Admit
## 30   November   Male California 3.66     0.9      956    89    Admit
## 31   December Female       Utah 3.78     1.2      968    87    Admit
## 32    January   Male   Colorado 3.88     1.0      969    93    Admit
## 33    January Female    Florida 3.80     1.9      965    94    Admit
## 34    January   Male    Florida 3.77     1.4      969    99    Admit
## 35    January Female California 3.87     1.7      966    97    Admit
## 37    January   Male   Virginia 3.89     0.7      966    NA    Admit
## 38   December Female   Virginia 3.59     1.8      966    NA    Admit
## 39   December Female California 3.78     4.7      952    99    Admit
## 40    January   Male   Virginia 3.87     8.2      957    NA    Admit
## 41    January   Male   Colorado 3.56     1.7      969    91    Admit
## 42    January   Male California 3.66     1.0      963    NA    Admit
## 44    January Female      Other 3.77     1.7      969    NA    Admit
## 45    January   Male    Florida 3.56     1.1      968    95    Admit
## 46    January   Male    Florida 3.51     2.9      964    95    Admit
## 47    January Female   Virginia 3.61     0.0      967    NA    Admit
## 48    January Female California 3.77     1.4      969    96    Admit
## 49    January Female   Virginia 3.59     0.7      967    NA    Admit
## 51    January Female California 3.64     3.2      969    NA    Admit
## 52    January   Male California 3.91     1.2      969    NA    Admit
## 53    January Female California 3.88     0.0      969    NA    Admit
## 54    January   Male    Florida 3.56     3.7      966    NA  Decline
## 55    January Female   Virginia 3.78     2.7      957    NA    Admit
## 56    January   Male    Florida 3.59     1.2      968    NA    Admit
## 58    January   Male California 3.97     3.7      968    91    Admit
## 59    January   Male   Colorado 3.89     1.7      969    NA    Admit
## 60    January Female   Colorado 3.77     1.2      967    NA    Admit
## 61    January Female   Colorado 3.78     1.7      969    90    Admit
```

```
## 62    January   Male   Virginia 3.56     1.2      969    NA    Admit
## 63    January Female   Virginia 3.67     2.7      969    97    Admit
## 65    January Female California 3.62     2.7      961    NA    Admit
## 66    January Female      Other 3.69     1.2      968    NA    Admit
## 67    January   Male   Virginia 3.92     1.3      966    99    Admit
## 68    January Female    Florida 3.74     1.3      966    NA  Decline
## 69    January Female   Colorado 3.86     1.2      967    98    Admit
## 70    January   Male   Virginia 3.69     5.7      968    89    Admit
## 72    January Female   Virginia 3.89     1.5      969    87    Admit
## 73    January Female    Florida 3.61     1.3      959    NA  Decline
## 74    January Female    Florida 3.54     0.9      969    NA  Decline
## 75    January Female      Other 3.94     0.9      965    NA    Admit
## 76    January   Male   Colorado 3.59     1.4      969    93    Admit
## 77    January Female    Florida 3.84     2.7      960    NA  Decline
## 79   November Female California 2.34     0.8      754    75  Decline
## 80   November Female      Other 2.90     0.9      769    56  Decline
## 81   November Female    Florida 3.33     1.6      766    NA  Decline
## 82   December   Male   Virginia 3.37     0.9      766    NA  Decline
## 83   December Female   Colorado 3.00     1.2      768    56  Decline
## 84   December   Male    Florida 3.10     1.9      751    NA  Decline
## 86    January   Male    Florida 3.54     1.1      767    65  Decline
## 87    January Female   Colorado 3.44     3.2      757    NA  Decline
## 88    January   Male   Colorado 2.98     0.7      763    71  Decline
## 89   November Female   Virginia 2.77     3.7      763    NA  Decline
## 90   December   Male    Florida 3.18     1.4      768    NA  Decline
## 91    January Female   Colorado 3.11     1.7      758    69  Decline
## 93    January   Male       Utah 3.01     1.4      769    69  Decline
## 94    January   Male       Utah 3.33     0.8      768    NA  Decline
## 95    January Female      Other 2.91     6.2      753    NA  Decline
## 96   December Female    Florida 3.56     1.7      769    81  Decline
## 97      March Female   Colorado 2.85     4.6      762    NA  Decline
## 98   February Female   Virginia 3.21     1.7      766    79  Decline
## 100     March Female    Florida 3.35     4.2      764    69  Decline
## 101  February Female   Virginia 3.11     8.9      742    NA  Decline
## 102   January Female   Virginia 3.12     0.0      761    NA  Decline
## 103  February   Male   Colorado 3.21     0.0      765    NA  Decline
## 104   January   Male   Colorado 3.34     3.7      769    NA  Decline
## 105     March Female    Florida 3.24     1.7      769    74  Decline
## 107  November   Male    Florida 3.45     4.7      867    71 Waitlist
## 108  November   Male    Florida 3.50     1.7      869    73 Waitlist
## 109  December Female      Other 3.55     2.2      866    74 Waitlist
## 110  November   Male      Other 3.41     1.2      868    85 Waitlist
## 111  December   Male      Other 3.56     0.9      866    NA Waitlist
## 112  December Female    Florida 3.53     1.7      869    NA Waitlist
## 114   January   Male   Colorado 3.50     3.5      869    83 Waitlist
## 115   January Female       Utah 3.39     1.8      866    82 Waitlist
## 116  November Female California 3.52     2.7      855    NA Waitlist
## 117   January   Male   Colorado 3.49     1.3      866    NA Waitlist
## 118   January Female   Colorado 3.43     1.5      869    NA Waitlist
## 119   January   Male    Florida 3.44     7.2      865    NA Waitlist
```

```
## 121   January Female       Utah 3.58     0.9      864    81 Waitlist
## 122   January Female       Utah 3.57     1.4      869    80 Waitlist
## 123   January Female    Florida 3.56     1.3      869    84 Waitlist
## 124   January   Male    Florida 3.55     2.0      853    NA Waitlist
## 125   January   Male   Colorado 3.54     1.2      868    83 Waitlist
## 126   January Female      Other 3.53     3.3      862    NA Waitlist
## 128   January Female    Florida 3.51     3.4      865    88 Waitlist
## 129   January Female California 3.47     2.2      867    NA Waitlist
## 130   January Female    Florida 3.46     1.9      869    NA Waitlist
## 131   January   Male    Florida 3.45     0.7      866    NA Waitlist
## 132   January Female California 3.44     0.7      867    83 Waitlist
## 133   January   Male California 3.42     1.7      866    NA Waitlist
## 135   January   Male       Utah 3.40     1.2      868    80 Waitlist
## 136   January   Male    Florida 3.39     1.2      866    NA Waitlist
## 137   January Female   Colorado 3.41     2.7      866    79 Waitlist
## 138   January   Male   Colorado 3.45     1.7      869    78 Waitlist
## 139  February Female   Colorado 3.49     2.7      866    NA Waitlist
## 140   January   Male    Florida 3.53     0.7      864    70 Waitlist
## 142   January   Male California 3.54     4.2      865    NA Waitlist
## 143   January Female       Utah 3.34     2.3      867    NA Waitlist
## 144   January Female      Other 3.61     0.8      867    84 Waitlist
## 145   January   Male       Utah 3.48     2.3      867    85 Waitlist
## 146   January   Male    Florida 3.39     4.2      863    86 Waitlist
## 147   January Female       Utah 3.49     0.7      868    82 Waitlist
## 149   January   Male    Florida 3.57     1.0      868    NA Waitlist
##     GPAcategory
## 2       HighGPA
## 3       HighGPA
## 4        MedGPA
## 5       HighGPA
## 6       HighGPA
## 7       HighGPA
## 9       HighGPA
## 10       MedGPA
## 11       MedGPA
## 12      HighGPA
## 13       MedGPA
## 14       MedGPA
## 16      HighGPA
## 17      HighGPA
## 18       MedGPA
## 19       MedGPA
## 20       MedGPA
## 21      HighGPA
## 23      HighGPA
## 24      HighGPA
## 25       MedGPA
## 26       MedGPA
## 27      HighGPA
## 28       MedGPA
```

```
## 30       MedGPA
## 31      HighGPA
## 32      HighGPA
## 33      HighGPA
## 34      HighGPA
## 35      HighGPA
## 37      HighGPA
## 38       MedGPA
## 39      HighGPA
## 40      HighGPA
## 41       MedGPA
## 42       MedGPA
## 44      HighGPA
## 45       MedGPA
## 46       MedGPA
## 47       MedGPA
## 48      HighGPA
## 49       MedGPA
## 51       MedGPA
## 52      HighGPA
## 53      HighGPA
## 54       MedGPA
## 55      HighGPA
## 56       MedGPA
## 58      HighGPA
## 59      HighGPA
## 60      HighGPA
## 61      HighGPA
## 62       MedGPA
## 63       MedGPA
## 65       MedGPA
## 66       MedGPA
## 67      HighGPA
## 68      HighGPA
## 69      HighGPA
## 70       MedGPA
## 72      HighGPA
## 73       MedGPA
## 74       MedGPA
## 75      HighGPA
## 76       MedGPA
## 77      HighGPA
## 79       LowGPA
## 80       LowGPA
## 81       LowGPA
## 82       LowGPA
## 83       LowGPA
## 84       LowGPA
## 86       MedGPA
## 87       MedGPA
```

```
## 88       LowGPA
## 89       LowGPA
## 90       LowGPA
## 91       LowGPA
## 93       LowGPA
## 94       LowGPA
## 95       LowGPA
## 96       MedGPA
## 97       LowGPA
## 98       LowGPA
## 100      LowGPA
## 101      LowGPA
## 102      LowGPA
## 103      LowGPA
## 104      LowGPA
## 105      LowGPA
## 107      MedGPA
## 108      MedGPA
## 109      MedGPA
## 110      MedGPA
## 111      MedGPA
## 112      MedGPA
## 114      MedGPA
## 115      LowGPA
## 116      MedGPA
## 117      MedGPA
## 118      MedGPA
## 119      MedGPA
## 121      MedGPA
## 122      MedGPA
## 123      MedGPA
## 124      MedGPA
## 125      MedGPA
## 126      MedGPA
## 128      MedGPA
## 129      MedGPA
## 130      MedGPA
## 131      MedGPA
## 132      MedGPA
## 133      MedGPA
## 135      LowGPA
## 136      LowGPA
## 137      MedGPA
## 138      MedGPA
## 139      MedGPA
## 140      MedGPA
## 142      MedGPA
## 143      LowGPA
## 144      MedGPA
## 145      MedGPA
```

```
## 146      LowGPA
## 147      MedGPA
## 149      MedGPA

(table(DT_Train$GPAcategory))

##
##  LowGPA  MedGPA HighGPA
##      26      65      36

DT_Test$GPAcategory <-
  cut(DT_Test$GPA, breaks=c(-Inf, 3.4, 3.7, Inf),
      labels=c("LowGPA","MedGPA","HighGPA"))
## Change any NAs to MedGPA
DT_Test$GPAcategory[is.na(DT_Test$GPA)] <- "MedGPA"
(DT_Test)

##       DateSub Gender      State  GPA WorkExp MathTest Essay GPAcategory
## 1   September Female California 3.90     6.7      962    NA     HighGPA
## 8    December Female California 3.70     1.2      969    NA      MedGPA
## 15   November   Male    Florida 3.70     3.7      969    99      MedGPA
## 22   December Female   Colorado 3.58     0.8      969    93      MedGPA
## 29    January Female   Colorado 3.59     1.7      969    93      MedGPA
## 36    January Female   Virginia 3.65     1.0      966    NA      MedGPA
## 43    January Female    Florida 3.87     1.8      968   100     HighGPA
## 50    January Female   Colorado 3.71     1.1      969    NA     HighGPA
## 57    January Female    Florida 3.79     1.4      969    NA     HighGPA
## 64    January   Male    Florida 3.71     0.7      968    98     HighGPA
## 71    January   Male   Virginia 3.45     0.7      969    NA      MedGPA
## 78  September   Male    Florida 2.81     9.2      764    NA      LowGPA
## 85   December   Male   Virginia 3.22     3.2      769    78      LowGPA
## 92    January   Male   Colorado 3.32     1.7      768    78      LowGPA
## 99   December Female    Florida 3.38     0.7      768    NA      LowGPA
## 106   January   Male    Florida 3.40     1.9      859    NA      LowGPA
## 113  December   Male California 3.42     0.7      869    84      MedGPA
## 120   January Female    Florida 3.29     1.2      869    NA      LowGPA
## 127   January   Male    Florida 3.52     0.7      868    81      MedGPA
## 134   January   Male       Utah 3.41     1.4      869    81      MedGPA
## 141  December Female    Florida 3.45     1.7      867    87      MedGPA
## 148   January Female       Utah 3.56     1.4      867    84      MedGPA

(table(DT_Test$GPAcategory))

##
##  LowGPA  MedGPA HighGPA
##       6      11       5

## Step 2 -----------------
### Categorize the Essay scores AND deal with the NAs
## I will group the Essays as shown and will change NA to a new group
## called NotIncluded.
```

```
DT_Test$Essaycategory <-
  cut(DT_Test$Essay, breaks=c(-Inf, 60, 85, Inf),
      labels=c("Low","Med","High"))
DT_Test$Essaycategory <-fct_explicit_na(DT_Test$Essaycategory, "NotIncluded")
str(DT_Test$Essaycategory)

##  Factor w/ 4 levels "Low","Med","High",..: 4 4 3 3 3 4 3 4 4 3 ...

## I need DT_Test$Essaycategory to be a string not factor
## so I can add a category of "NotIncluded"
(table(DT_Test$Essaycategory))

##
##         Low         Med        High NotIncluded
##           0           6           6          10

(DT_Test$Essaycategory)

##  [1] NotIncluded NotIncluded High        High        High
##  [6] NotIncluded High        NotIncluded NotIncluded High
## [11] NotIncluded NotIncluded Med         Med         NotIncluded
## [16] NotIncluded Med         NotIncluded Med         Med
## [21] High        Med
## Levels: Low Med High NotIncluded

DT_Train$Essaycategory <-
  cut(DT_Train$Essay, breaks=c(-Inf, 60, 85, Inf),
      labels=c("Low","Med","High"))
DT_Train$Essaycategory <-fct_explicit_na(DT_Train$Essaycategory, "NotIncluded
")
(table(DT_Train$Essaycategory))

##
##         Low         Med        High NotIncluded
##           2          27          34          64

## As we can see, the Essay column is not very helpful overall. The NotIncludd
ed
## is large and the Low is very small.


## Step 3----------------------
## Categorize the submission time as Early, Ontime, or Later
DT_Train$DateSub[DT_Train$DateSub %in% c("September", "October")] <- "Early"
DT_Train$DateSub[DT_Train$DateSub %in% c("November","December")] <- "Ontime"
DT_Train$DateSub[DT_Train$DateSub %in% c("January","February","March","April"
)] <- "Later"
(head(DT_Train,n=15))

##    DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 2    Early Female    Florida 3.80     1.4      969    97    Admit
## 3    Early   Male California 3.80     2.3      970    NA    Admit
```

```
## 4    Early   Male   Colorado 3.60     0.9      969    NA    Admit
## 5   Ontime   Male   Colorado 3.92     1.2      969    95    Admit
## 6   Ontime   Male California 3.80     1.2      967    NA    Admit
## 7   Ontime Female California 3.88     0.0      967    NA    Admit
## 9    Early Female    Florida 3.90     4.7      961    NA    Admit
## 10  Ontime Female California 3.70     1.4      966    94    Admit
## 11  Ontime Female    Florida 3.56     1.7      968    91    Admit
## 12  Ontime Female    Florida 3.93     0.8      969    NA    Admit
## 13   Early Female   Colorado 3.60     1.2      967    94    Admit
## 14  Ontime   Male California 3.69     3.2      967    93    Admit
## 16  Ontime Female   Colorado 3.90     0.0      967    NA    Admit
## 17   Later   Male   Colorado 3.78     1.2      966   100    Admit
## 18  Ontime   Male California 3.70     2.7      799    97    Admit
##    GPAcategory Essaycategory
## 2      HighGPA          High
## 3      HighGPA   NotIncluded
## 4       MedGPA   NotIncluded
## 5      HighGPA          High
## 6      HighGPA   NotIncluded
## 7      HighGPA   NotIncluded
## 9      HighGPA   NotIncluded
## 10      MedGPA          High
## 11      MedGPA          High
## 12     HighGPA   NotIncluded
## 13      MedGPA          High
## 14      MedGPA          High
## 16     HighGPA   NotIncluded
## 17     HighGPA          High
## 18      MedGPA          High
```

```
(table(DT_Train$DateSub))
```

```
##
##  Early  Later Ontime
##      6     89     32
```

```
DT_Test$DateSub[DT_Test$DateSub %in% c("September", "October")] <- "Early"
DT_Test$DateSub[DT_Test$DateSub %in% c( "November", "December")] <- "Ontime"
DT_Test$DateSub[DT_Test$DateSub %in% c("January", "February","March","April")
] <- "Later"
(head(DT_Test,n=5))
```

```
##    DateSub Gender      State  GPA WorkExp MathTest Essay GPAcategory
## 1    Early Female California 3.90     6.7      962    NA     HighGPA
## 8   Ontime Female California 3.70     1.2      969    NA      MedGPA
## 15  Ontime   Male    Florida 3.70     3.7      969    99      MedGPA
## 22  Ontime Female   Colorado 3.58     0.8      969    93      MedGPA
## 29   Later Female   Colorado 3.59     1.7      969    93      MedGPA
##    Essaycategory
## 1    NotIncluded
## 8    NotIncluded
```

```
## 15          High
## 22          High
## 29          High

(table(DT_Test$DateSub))

##
##  Early  Later Ontime
##      2     13      7

##---------------------------

## Step 4 -------------------
## Let's see where we are....
(head(DT_Train,n=5))

##   DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 2   Early Female    Florida 3.80     1.4      969    97    Admit
## 3   Early   Male California 3.80     2.3      970    NA    Admit
## 4   Early   Male   Colorado 3.60     0.9      969    NA    Admit
## 5  Ontime   Male   Colorado 3.92     1.2      969    95    Admit
## 6  Ontime   Male California 3.80     1.2      967    NA    Admit
##   GPAcategory Essaycategory
## 2     HighGPA          High
## 3     HighGPA   NotIncluded
## 4      MedGPA   NotIncluded
## 5     HighGPA          High
## 6     HighGPA   NotIncluded

(head(DT_Test,n=5))

##    DateSub Gender      State  GPA WorkExp MathTest Essay GPAcategory
## 1    Early Female California 3.90     6.7      962    NA     HighGPA
## 8   Ontime Female California 3.70     1.2      969    NA      MedGPA
## 15  Ontime   Male    Florida 3.70     3.7      969    99      MedGPA
## 22  Ontime Female   Colorado 3.58     0.8      969    93      MedGPA
## 29   Later Female   Colorado 3.59     1.7      969    93      MedGPA
##    Essaycategory
## 1    NotIncluded
## 8    NotIncluded
## 15          High
## 22          High
## 29          High

ReadyDT_Train<-DT_Train[-c(4,5,6,7)]
ReadyDT_Test<-DT_Test[-c(4,5,6,7)]
TestLabels <- DT_Test_Labels

(head(ReadyDT_Test,n=20))
```

```
##     DateSub Gender      State GPAcategory Essaycategory
## 1     Early Female California     HighGPA   NotIncluded
## 8    Ontime Female California      MedGPA   NotIncluded
## 15   Ontime   Male    Florida      MedGPA          High
## 22   Ontime Female   Colorado      MedGPA          High
## 29    Later Female   Colorado      MedGPA          High
## 36    Later Female   Virginia      MedGPA   NotIncluded
## 43    Later Female    Florida     HighGPA          High
## 50    Later Female   Colorado     HighGPA   NotIncluded
## 57    Later Female    Florida     HighGPA   NotIncluded
## 64    Later   Male    Florida     HighGPA          High
## 71    Later   Male   Virginia      MedGPA   NotIncluded
## 78    Early   Male    Florida      LowGPA   NotIncluded
## 85   Ontime   Male   Virginia      LowGPA           Med
## 92    Later   Male   Colorado      LowGPA           Med
## 99   Ontime Female    Florida      LowGPA   NotIncluded
## 106   Later   Male    Florida      LowGPA   NotIncluded
## 113  Ontime   Male California      MedGPA           Med
## 120   Later Female    Florida      LowGPA   NotIncluded
## 127   Later   Male    Florida      MedGPA           Med
## 134   Later   Male       Utah      MedGPA           Med
```

```
(head(ReadyDT_Train,n=20))
```

```
##    DateSub Gender      State Decision GPAcategory Essaycategory
## 2    Early Female    Florida    Admit     HighGPA          High
## 3    Early   Male California    Admit     HighGPA   NotIncluded
## 4    Early   Male   Colorado    Admit      MedGPA   NotIncluded
## 5   Ontime   Male   Colorado    Admit     HighGPA          High
## 6   Ontime   Male California    Admit     HighGPA   NotIncluded
## 7   Ontime Female California    Admit     HighGPA   NotIncluded
## 9    Early Female    Florida    Admit     HighGPA   NotIncluded
## 10  Ontime Female California    Admit      MedGPA          High
## 11  Ontime Female    Florida    Admit      MedGPA          High
## 12  Ontime Female    Florida    Admit     HighGPA   NotIncluded
## 13   Early Female   Colorado    Admit      MedGPA          High
## 14  Ontime   Male California    Admit      MedGPA          High
## 16  Ontime Female   Colorado    Admit     HighGPA   NotIncluded
## 17   Later   Male   Colorado    Admit     HighGPA          High
## 18  Ontime   Male California    Admit      MedGPA          High
## 19  Ontime   Male    Florida    Admit      MedGPA   NotIncluded
## 20   Later   Male   Colorado    Admit      MedGPA   NotIncluded
## 21   Later Female   Colorado    Admit     HighGPA   NotIncluded
## 23   Early   Male California    Admit     HighGPA          High
## 24   Later Female California    Admit     HighGPA          High
```

```
(table(ReadyDT_Train$DateSub))
```

```
##
##  Early  Later Ontime
##      6     89     32
```

```
## Step 5 ----------------------------------------
## Now, we can train and test the decision tree
Treefit <- rpart(ReadyDT_Train$Decision ~ ., data = ReadyDT_Train, method="cl
ass")
summary(Treefit)

## Call:
## rpart(formula = ReadyDT_Train$Decision ~ ., data = ReadyDT_Train,
##     method = "class")
##   n= 127
##
##           CP nsplit rel error    xerror       xstd
## 1 0.31818182      0 1.0000000 1.0000000 0.08530826
## 2 0.24242424      1 0.6818182 0.6818182 0.08167091
## 3 0.02525253      2 0.4393939 0.4393939 0.07167476
## 4 0.01000000      5 0.3636364 0.4848485 0.07412949
##
## Variable importance
##   GPAcategory Essaycategory         State       DateSub
##            48            43             8             1
##
## Node number 1: 127 observations,    complexity param=0.3181818
##   predicted class=Admit     expected loss=0.519685  P(node) =1
##     class counts:    61    29    37
##    probabilities: 0.480 0.228 0.291
##   left son=2 (101 obs) right son=3 (26 obs)
##   Primary splits:
##       GPAcategory   splits as  RLL,    improve=18.8361500, (0 missing)
##       Essaycategory splits as  RRLR,   improve=14.7925700, (0 missing)
##       State         splits as  LLRRRL, improve= 5.5170590, (0 missing)
##       DateSub       splits as  LRL,    improve= 1.1537370, (0 missing)
##       Gender        splits as  LR,     improve= 0.5156044, (0 missing)
##   Surrogate splits:
##       Essaycategory splits as  RLLL, agree=0.811, adj=0.077, (0 split)
##
## Node number 2: 101 observations,    complexity param=0.2424242
##   predicted class=Admit     expected loss=0.3960396  P(node) =0.7952756
##     class counts:    61     8    32
##    probabilities: 0.604 0.079 0.317
##   left son=4 (83 obs) right son=5 (18 obs)
##   Primary splits:
##       Essaycategory splits as  -RLL,   improve=15.1799800, (0 missing)
##       GPAcategory   splits as  -RL,    improve=12.1314400, (0 missing)
##       State         splits as  LLRRRL, improve= 5.1710590, (0 missing)
##       DateSub       splits as  LRL,    improve= 1.1016320, (0 missing)
##       Gender        splits as  LR,     improve= 0.1820409, (0 missing)
##   Surrogate splits:
##       State splits as  LLLLRL, agree=0.851, adj=0.167, (0 split)
##
## Node number 3: 26 observations
```

```
##   predicted class=Decline   expected loss=0.1923077  P(node) =0.2047244
##     class counts:     0    21     5
##    probabilities: 0.000 0.808 0.192
##
## Node number 4: 83 observations,    complexity param=0.02525253
##   predicted class=Admit     expected loss=0.2650602  P(node) =0.6535433
##     class counts:    61     6    16
##    probabilities: 0.735 0.072 0.193
##   left son=8 (36 obs) right son=9 (47 obs)
##   Primary splits:
##       GPAcategory   splits as  -RL,    improve=5.170697, (0 missing)
##       Essaycategory splits as  --LR,   improve=4.751208, (0 missing)
##       State         splits as  LLRRLL, improve=2.584846, (0 missing)
##       DateSub       splits as  LRL,    improve=1.129223, (0 missing)
##       Gender        splits as  RL,     improve=0.226041, (0 missing)
##   Surrogate splits:
##       Essaycategory splits as  --LR,   agree=0.602, adj=0.083, (0 split)
##       DateSub       splits as  LRR,    agree=0.590, adj=0.056, (0 split)
##       Gender        splits as  LR,     agree=0.578, adj=0.028, (0 split)
##       State         splits as  RRRRLR, agree=0.578, adj=0.028, (0 split)
##
## Node number 5: 18 observations
##   predicted class=Waitlist  expected loss=0.1111111  P(node) =0.1417323
##     class counts:     0     2    16
##    probabilities: 0.000 0.111 0.889
##
## Node number 8: 36 observations
##   predicted class=Admit     expected loss=0.05555556  P(node) =0.2834646
##     class counts:    34     2     0
##    probabilities: 0.944 0.056 0.000
##
## Node number 9: 47 observations,    complexity param=0.02525253
##   predicted class=Admit     expected loss=0.4255319  P(node) =0.3700787
##     class counts:    27     4    16
##    probabilities: 0.574 0.085 0.340
##   left son=18 (15 obs) right son=19 (32 obs)
##   Primary splits:
##       Essaycategory splits as  --LR,   improve=4.6479610, (0 missing)
##       State         splits as  LRRR-L, improve=1.4511470, (0 missing)
##       DateSub       splits as  LRL,    improve=1.0279200, (0 missing)
##       Gender        splits as  RL,     improve=0.3603095, (0 missing)
##   Surrogate splits:
##       DateSub splits as  RRL, agree=0.702, adj=0.067, (0 split)
##
## Node number 18: 15 observations
##   predicted class=Admit     expected loss=0.06666667  P(node) =0.1181102
##     class counts:    14     0     1
##    probabilities: 0.933 0.000 0.067
##
## Node number 19: 32 observations,    complexity param=0.02525253
```

```
##   predicted class=Waitlist  expected loss=0.53125  P(node) =0.2519685
##     class counts:    13     4    15
##    probabilities: 0.406 0.125 0.469
##   left son=38 (11 obs) right son=39 (21 obs)
##   Primary splits:
##       State  splits as  LRRR-L, improve=1.334686, (0 missing)
##       Gender splits as  LR,     improve=0.214951, (0 missing)
##
## Node number 38: 11 observations
##   predicted class=Admit     expected loss=0.3636364  P(node) =0.08661417
##     class counts:     7     0     4
##    probabilities: 0.636 0.000 0.364
##
## Node number 39: 21 observations
##   predicted class=Waitlist  expected loss=0.4761905  P(node) =0.1653543
##     class counts:     6     4    11
##    probabilities: 0.286 0.190 0.524
```

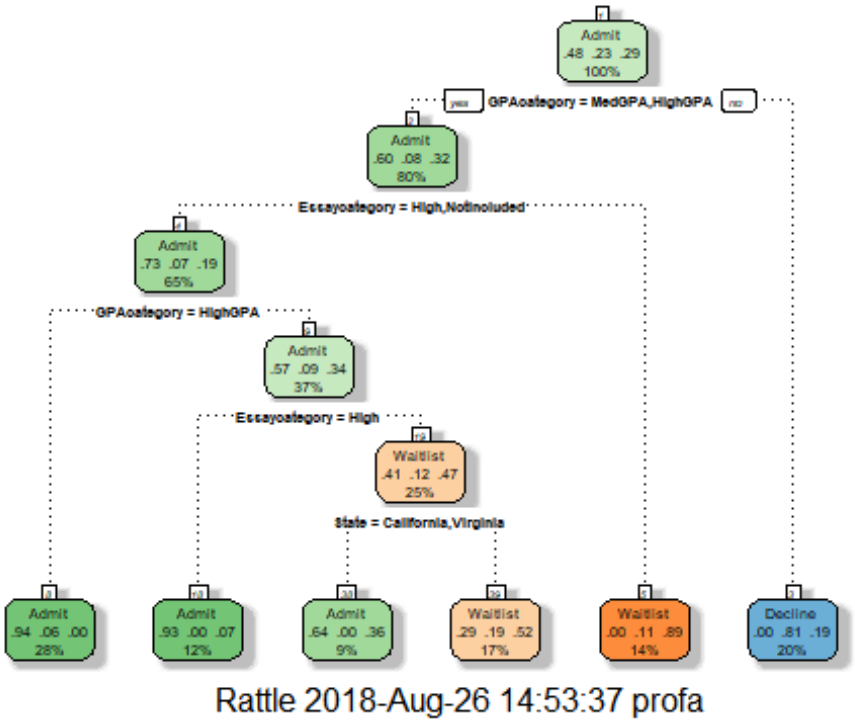
```
predicted= predict(Treefit,ReadyDT_Test, type="class")
(Results <- data.frame(Predicted=predicted,Actual=TestLabels))
```

```
##     Predicted   Actual
## 1       Admit    Admit
## 8       Admit    Admit
## 15      Admit    Admit
## 22      Admit    Admit
## 29      Admit    Admit
## 36      Admit    Admit
## 43      Admit    Admit
## 50      Admit    Admit
## 57      Admit  Decline
## 64      Admit  Decline
## 71      Admit    Admit
## 78    Decline  Decline
## 85    Decline  Decline
## 92    Decline  Decline
## 99    Decline  Decline
## 106   Decline Waitlist
## 113  Waitlist Waitlist
## 120   Decline Waitlist
## 127  Waitlist Waitlist
## 134  Waitlist Waitlist
## 141     Admit Waitlist
## 148  Waitlist Waitlist
```

```
(table(Results))
```

```
##           Actual
## Predicted  Admit Decline Waitlist
##   Admit        9       2        1
```

```
##   Decline      0       4        2
##   Waitlist     0       0        4

fancyRpartPlot(Treefit)
```




```
###  DT 2----------------------
## Sometimes it is interesting to remove variables to see the
## affect this will have on the Decision Tree.
## Suppose I remove GPA as this seems to be critical.
## Let's remove that variable and see what we get

(head(DT_Train,n=5))

##   DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 2   Early Female    Florida 3.80     1.4      969    97    Admit
## 3   Early   Male California 3.80     2.3      970    NA    Admit
## 4   Early   Male   Colorado 3.60     0.9      969    NA    Admit
## 5  Ontime   Male   Colorado 3.92     1.2      969    95    Admit
## 6  Ontime   Male California 3.80     1.2      967    NA    Admit
##   GPAcategory Essaycategory
## 2     HighGPA          High
## 3     HighGPA   NotIncluded
## 4      MedGPA   NotIncluded
## 5     HighGPA          High
## 6     HighGPA   NotIncluded

(head(DT_Test,n=5))
```

```
##    DateSub Gender      State  GPA WorkExp MathTest Essay GPAcategory
## 1    Early Female California 3.90     6.7      962    NA     HighGPA
## 8   Ontime Female California 3.70     1.2      969    NA      MedGPA
## 15  Ontime   Male    Florida 3.70     3.7      969    99      MedGPA
## 22  Ontime Female   Colorado 3.58     0.8      969    93      MedGPA
## 29   Later Female   Colorado 3.59     1.7      969    93      MedGPA
##    Essaycategory
## 1    NotIncluded
## 8    NotIncluded
## 15          High
## 22          High
## 29          High

DT_Train_NoGPA<-DT_Train[-c(4,5,6,7,9)]
DT_Test_NoGPA<-DT_Test[-c(4,5,6,7,8)]
DT_Test_Labels <- DT_Test_Labels
(head(DT_Train_NoGPA,n=5))

##   DateSub Gender      State Decision Essaycategory
## 2   Early Female    Florida    Admit          High
## 3   Early   Male California    Admit   NotIncluded
## 4   Early   Male   Colorado    Admit   NotIncluded
## 5  Ontime   Male   Colorado    Admit          High
## 6  Ontime   Male California    Admit   NotIncluded

(head(DT_Test_NoGPA,n=5))

##    DateSub Gender      State Essaycategory
## 1    Early Female California   NotIncluded
## 8   Ontime Female California   NotIncluded
## 15  Ontime   Male    Florida          High
## 22  Ontime Female   Colorado          High
## 29   Later Female   Colorado          High

Treefit2 <- rpart(DT_Train_NoGPA$Decision ~ ., data = DT_Train_NoGPA, method=
"class")
summary(Treefit2)

## Call:
## rpart(formula = DT_Train_NoGPA$Decision ~ ., data = DT_Train_NoGPA,
##     method = "class")
##   n= 127
##
##           CP nsplit rel error    xerror       xstd
## 1 0.13636364      0 1.0000000 1.0000000 0.08530826
## 2 0.03030303      2 0.7272727 0.8181818 0.08441362
## 3 0.01515152      4 0.6666667 0.9848485 0.08535063
## 4 0.01000000      5 0.6515152 1.0000000 0.08530826
##
## Variable importance
## Essaycategory         State       DateSub
```

```
##            80            16             3
##
## Node number 1: 127 observations,    complexity param=0.1363636
##   predicted class=Admit     expected loss=0.519685  P(node) =1
##     class counts:    61    29    37
##    probabilities: 0.480 0.228 0.291
##   left son=2 (34 obs) right son=3 (93 obs)
##   Primary splits:
##       Essaycategory splits as  RRLR,   improve=14.7925700, (0 missing)
##       State         splits as  LLRRRL, improve= 5.5170590, (0 missing)
##       DateSub       splits as  LRL,    improve= 1.1537370, (0 missing)
##       Gender        splits as  LR,     improve= 0.5156044, (0 missing)
##
## Node number 2: 34 observations
##   predicted class=Admit     expected loss=0.05882353  P(node) =0.2677165
##     class counts:    32     0     2
##    probabilities: 0.941 0.000 0.059
##
## Node number 3: 93 observations,    complexity param=0.1363636
##   predicted class=Waitlist  expected loss=0.6236559  P(node) =0.7322835
##     class counts:    29    29    35
##    probabilities: 0.312 0.312 0.376
##   left son=6 (66 obs) right son=7 (27 obs)
##   Primary splits:
##       Essaycategory splits as  LR-L,   improve=6.9237540, (0 missing)
##       State         splits as  LRRRRL, improve=3.9956670, (0 missing)
##       Gender        splits as  LR,     improve=0.6589166, (0 missing)
##       DateSub       splits as  LRL,    improve=0.4760531, (0 missing)
##   Surrogate splits:
##       State splits as  LLLLRL, agree=0.763, adj=0.185, (0 split)
##
## Node number 6: 66 observations,    complexity param=0.03030303
##   predicted class=Admit     expected loss=0.5606061  P(node) =0.519685
##     class counts:    29    20    17
##    probabilities: 0.439 0.303 0.258
##   left son=12 (23 obs) right son=13 (43 obs)
##   Primary splits:
##       State   splits as  LRRRRL, improve=2.4339550, (0 missing)
##       DateSub splits as  RRL,    improve=0.5731818, (0 missing)
##       Gender  splits as  LR,     improve=0.3407382, (0 missing)
##
## Node number 7: 27 observations
##   predicted class=Waitlist  expected loss=0.3333333  P(node) =0.2125984
##     class counts:     0     9    18
##    probabilities: 0.000 0.333 0.667
##
## Node number 12: 23 observations
##   predicted class=Admit     expected loss=0.3478261  P(node) =0.1811024
##     class counts:    15     4     4
##    probabilities: 0.652 0.174 0.174
```

```
##
## Node number 13: 43 observations,    complexity param=0.03030303
##   predicted class=Decline   expected loss=0.627907  P(node) =0.3385827
##     class counts:    14    16    13
##    probabilities: 0.326 0.372 0.302
##   left son=26 (21 obs) right son=27 (22 obs)
##   Primary splits:
##       State   splits as  -LRLR-, improve=0.6663646, (0 missing)
##       DateSub splits as  LRL,    improve=0.4774944, (0 missing)
##       Gender  splits as  LR,     improve=0.3492506, (0 missing)
##   Surrogate splits:
##       Gender        splits as  LR,   agree=0.581, adj=0.143, (0 split)
##       Essaycategory splits as  L--R, agree=0.558, adj=0.095, (0 split)
##       DateSub       splits as  RLR,  agree=0.535, adj=0.048, (0 split)
##
## Node number 26: 21 observations
##   predicted class=Admit     expected loss=0.5714286  P(node) =0.1653543
##     class counts:     9     7     5
##    probabilities: 0.429 0.333 0.238
##
## Node number 27: 22 observations,    complexity param=0.01515152
##   predicted class=Decline   expected loss=0.5909091  P(node) =0.1732283
##     class counts:     5     9     8
##    probabilities: 0.227 0.409 0.364
##   left son=54 (7 obs) right son=55 (15 obs)
##   Primary splits:
##       DateSub splits as  LRL, improve=0.9203463, (0 missing)
##       Gender  splits as  LR,  improve=0.2727273, (0 missing)
##
## Node number 54: 7 observations
##   predicted class=Admit     expected loss=0.5714286  P(node) =0.05511811
##     class counts:     3     3     1
##    probabilities: 0.429 0.429 0.143
##
## Node number 55: 15 observations
##   predicted class=Waitlist  expected loss=0.5333333  P(node) =0.1181102
##     class counts:     2     6     7
##    probabilities: 0.133 0.400 0.467
```

```
predicted2= predict(Treefit2,DT_Test_NoGPA, type="class")
(Results2 <- data.frame(Predicted=predicted2,Actual=DT_Test_Labels))
```

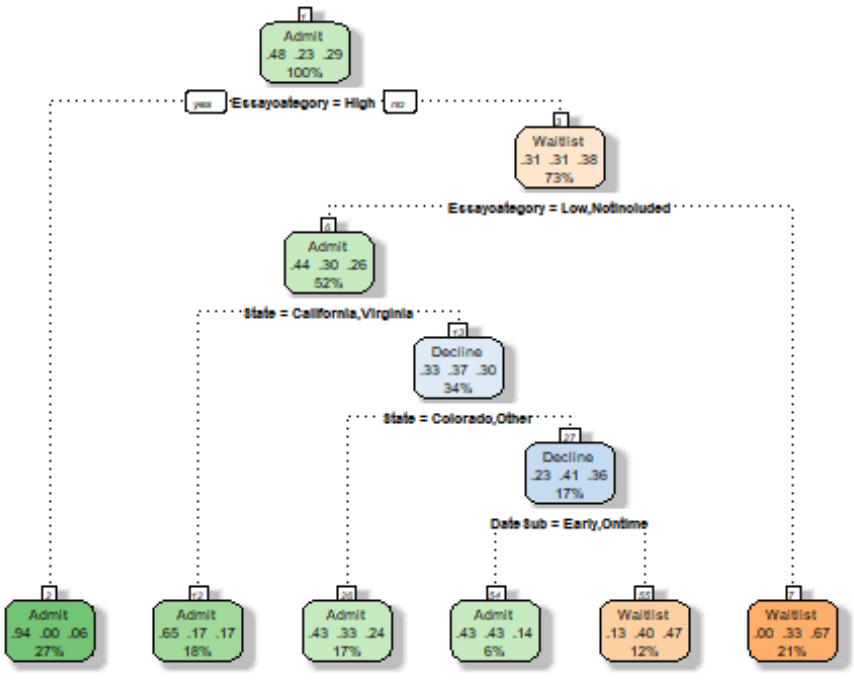
```
##     Predicted   Actual
## 1       Admit    Admit
## 8       Admit    Admit
## 15      Admit    Admit
## 22      Admit    Admit
## 29      Admit    Admit
## 36      Admit    Admit
## 43      Admit    Admit
```

```
## 50      Admit    Admit
## 57   Waitlist  Decline
## 64      Admit  Decline
## 71      Admit    Admit
## 78      Admit  Decline
## 85   Waitlist  Decline
## 92   Waitlist  Decline
## 99      Admit  Decline
## 106  Waitlist Waitlist
## 113  Waitlist Waitlist
## 120  Waitlist Waitlist
## 127  Waitlist Waitlist
## 134  Waitlist Waitlist
## 141     Admit Waitlist
## 148  Waitlist Waitlist
```

```
(table(Results2))
```

```
##           Actual
## Predicted  Admit Decline Waitlist
##   Admit        9       3        1
##   Decline      0       0        0
##   Waitlist     0       3        6
```

```
fancyRpartPlot(Treefit2)
```



Rattle 2018-Aug-26 14:53:37 profa

```
## Here we can see that the State plays a role in the Decision.

### DT 3 ------------------------------
## Let's see what happens if we leave in the numerical values
(head(DT_Train,n=5))

##   DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 2   Early Female    Florida 3.80     1.4      969    97    Admit
## 3   Early   Male California 3.80     2.3      970    NA    Admit
## 4   Early   Male   Colorado 3.60     0.9      969    NA    Admit
## 5  Ontime   Male   Colorado 3.92     1.2      969    95    Admit
## 6  Ontime   Male California 3.80     1.2      967    NA    Admit
##   GPAcategory Essaycategory
## 2     HighGPA          High
## 3     HighGPA   NotIncluded
## 4      MedGPA   NotIncluded
## 5     HighGPA          High
## 6     HighGPA   NotIncluded

(head(DT_Test,n=5))

##    DateSub Gender      State  GPA WorkExp MathTest Essay GPAcategory
## 1    Early Female California 3.90     6.7      962    NA     HighGPA
## 8   Ontime Female California 3.70     1.2      969    NA      MedGPA
## 15  Ontime   Male    Florida 3.70     3.7      969    99      MedGPA
## 22  Ontime Female   Colorado 3.58     0.8      969    93      MedGPA
## 29   Later Female   Colorado 3.59     1.7      969    93      MedGPA
##    Essaycategory
## 1    NotIncluded
## 8    NotIncluded
## 15          High
## 22          High
## 29          High

DT_Train_WithQuantData<-DT_Train[-c(9,10)]
DT_Test_WithQuantData<-DT_Test[-c(8,9)]
DT_Test_Labs <- DT_Test_Labels
(head(DT_Train_WithQuantData,n=5))

##   DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 2   Early Female    Florida 3.80     1.4      969    97    Admit
## 3   Early   Male California 3.80     2.3      970    NA    Admit
## 4   Early   Male   Colorado 3.60     0.9      969    NA    Admit
## 5  Ontime   Male   Colorado 3.92     1.2      969    95    Admit
## 6  Ontime   Male California 3.80     1.2      967    NA    Admit

(head(DT_Test_WithQuantData,n=5))

##    DateSub Gender      State  GPA WorkExp MathTest Essay
## 1    Early Female California 3.90     6.7      962    NA
## 8   Ontime Female California 3.70     1.2      969    NA
```

```
## 15  Ontime   Male    Florida 3.70     3.7      969    99
## 22  Ontime Female   Colorado 3.58     0.8      969    93
## 29   Later Female   Colorado 3.59     1.7      969    93
```

```
Treefit3 <- rpart(DT_Train_WithQuantData$Decision ~ ., data = DT_Train_WithQu
antData, method="class")
summary(Treefit3)
```

```
## Call:
## rpart(formula = DT_Train_WithQuantData$Decision ~ ., data = DT_Train_WithQ
uantData,
##     method = "class")
##   n= 127
##
##          CP nsplit  rel error     xerror       xstd
## 1 0.5454545      0 1.00000000 1.00000000 0.08530826
## 2 0.3636364      1 0.45454545 0.45454545 0.07252716
## 3 0.0100000      2 0.09090909 0.09090909 0.03622618
##
## Variable importance
## MathTest      GPA    State  WorkExp  DateSub
##       45       36        9        5        4
##
## Node number 1: 127 observations,    complexity param=0.5454545
##   predicted class=Admit     expected loss=0.519685  P(node) =1
##     class counts:    61    29    37
##    probabilities: 0.480 0.228 0.291
##   left son=2 (65 obs) right son=3 (62 obs)
##   Primary splits:
##       MathTest < 910.5 to the right, improve=40.455540, (0 missing)
##       GPA      < 3.585 to the right, improve=31.841070, (0 missing)
##       Essay    < 86.5  to the right, improve=22.603460, (64 missing)
##       State    splits as  LLRRRL,    improve= 5.517059, (0 missing)
##       WorkExp  < 0.35  to the left,  improve= 1.349633, (0 missing)
##   Surrogate splits:
##       GPA     < 3.585 to the right, agree=0.906, adj=0.806, (0 split)
##       State   splits as  LLRRRL,    agree=0.614, adj=0.210, (0 split)
##       WorkExp < 1.45  to the left,  agree=0.583, adj=0.145, (0 split)
##       DateSub splits as  LRR,       agree=0.535, adj=0.048, (0 split)
##
## Node number 2: 65 observations
##   predicted class=Admit     expected loss=0.07692308  P(node) =0.511811
##     class counts:    60     5     0
##    probabilities: 0.923 0.077 0.000
##
## Node number 3: 62 observations,    complexity param=0.3636364
##   predicted class=Waitlist  expected loss=0.4032258  P(node) =0.488189
##     class counts:     1    24    37
##    probabilities: 0.016 0.387 0.597
##   left son=6 (25 obs) right son=7 (37 obs)
```

```
##   Primary splits:
##       MathTest < 826   to the left,  improve=28.692900, (0 missing)
##       GPA      < 3.38  to the left,  improve=21.353810, (0 missing)
##       Essay    < 76.5  to the left,  improve= 4.885881, (30 missing)
##       State    splits as  RLRRRL,    improve= 3.073124, (0 missing)
##       DateSub  splits as  -RL,       improve= 1.318786, (0 missing)
##   Surrogate splits:
##       GPA     < 3.38  to the left,  agree=0.919, adj=0.80, (0 split)
##       State   splits as  RRRRRL,    agree=0.677, adj=0.20, (0 split)
##       DateSub splits as  -RL,       agree=0.645, adj=0.12, (0 split)
##       WorkExp < 0.35  to the left,  agree=0.629, adj=0.08, (0 split)
##
## Node number 6: 25 observations
##   predicted class=Decline   expected loss=0.04  P(node) =0.1968504
##     class counts:     1    24     0
##    probabilities: 0.040 0.960 0.000
##
## Node number 7: 37 observations
##   predicted class=Waitlist  expected loss=0  P(node) =0.2913386
##     class counts:     0     0    37
##    probabilities: 0.000 0.000 1.000
```

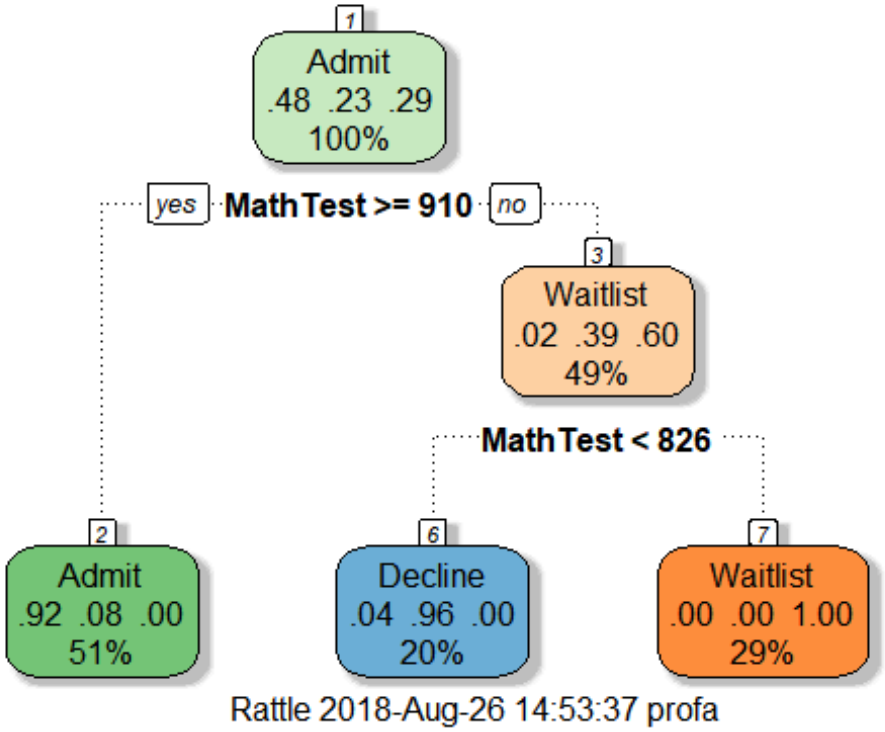
```
predicted3= predict(Treefit3,DT_Test_WithQuantData, type="class")
(Results3 <- data.frame(Predicted=predicted3,Actual=DT_Test_Labs))
```

```
##     Predicted   Actual
## 1       Admit    Admit
## 8       Admit    Admit
## 15      Admit    Admit
## 22      Admit    Admit
## 29      Admit    Admit
## 36      Admit    Admit
## 43      Admit    Admit
## 50      Admit    Admit
## 57      Admit  Decline
## 64      Admit  Decline
## 71      Admit    Admit
## 78    Decline  Decline
## 85    Decline  Decline
## 92    Decline  Decline
## 99    Decline  Decline
## 106  Waitlist Waitlist
## 113  Waitlist Waitlist
## 120  Waitlist Waitlist
## 127  Waitlist Waitlist
## 134  Waitlist Waitlist
## 141  Waitlist Waitlist
## 148  Waitlist Waitlist
```

```
(table(Results3))
```

```
##           Actual
## Predicted  Admit Decline Waitlist
##   Admit        9       2        0
##   Decline      0       4        0
##   Waitlist     0       0        7
```

```
fancyRpartPlot(Treefit3)
```




```
## By doing the above, we can see that the MathTest plays a large role.
## We can also see the natural cut-off value.

#### DT  4 ------------------------------------
### Let's see if State has any real affect.
### to do this, we will need to remove some of the more
## active elements such as GPA and MathTest...

(head(DT_Train,n=5))

##   DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 2   Early Female    Florida 3.80     1.4      969    97    Admit
## 3   Early   Male California 3.80     2.3      970    NA    Admit
## 4   Early   Male   Colorado 3.60     0.9      969    NA    Admit
## 5  Ontime   Male   Colorado 3.92     1.2      969    95    Admit
## 6  Ontime   Male California 3.80     1.2      967    NA    Admit
##   GPAcategory Essaycategory
## 2     HighGPA          High
## 3     HighGPA   NotIncluded
```

```
## 4      MedGPA   NotIncluded
## 5     HighGPA          High
## 6     HighGPA   NotIncluded

(head(DT_Test,n=5))

##    DateSub Gender      State  GPA WorkExp MathTest Essay GPAcategory
## 1    Early Female California 3.90     6.7      962    NA     HighGPA
## 8   Ontime Female California 3.70     1.2      969    NA      MedGPA
## 15  Ontime   Male    Florida 3.70     3.7      969    99      MedGPA
## 22  Ontime Female   Colorado 3.58     0.8      969    93      MedGPA
## 29   Later Female   Colorado 3.59     1.7      969    93      MedGPA
##    Essaycategory
## 1    NotIncluded
## 8    NotIncluded
## 15          High
## 22          High
## 29          High

DT_Test_GLabs<-DT_Test_Labels
DT_Train_Focus<-DT_Train[-c(4,5,6,7,9,10)]
DT_Test_Focus<-DT_Test[-c(4,5,6,7,8,9)]
(head(DT_Train_Focus,n=5))

##   DateSub Gender      State Decision
## 2   Early Female    Florida    Admit
## 3   Early   Male California    Admit
## 4   Early   Male   Colorado    Admit
## 5  Ontime   Male   Colorado    Admit
## 6  Ontime   Male California    Admit

(head(DT_Test_Focus,n=5))

##    DateSub Gender      State
## 1    Early Female California
## 8   Ontime Female California
## 15  Ontime   Male    Florida
## 22  Ontime Female   Colorado
## 29   Later Female   Colorado

Treefit4 <- rpart(DT_Train_Focus$Decision ~ ., data = DT_Train_Focus, method=
"class")
summary(Treefit4)

## Call:
## rpart(formula = DT_Train_Focus$Decision ~ ., data = DT_Train_Focus,
##     method = "class")
##   n= 127
##
##           CP nsplit rel error    xerror       xstd
## 1 0.10606061      0 1.0000000 1.0000000 0.08530826
## 2 0.02272727      1 0.8939394 1.0151515 0.08524467
```

```
## 3 0.01000000      4 0.8181818 0.9393939 0.08535063
##
## Variable importance
##   State  Gender DateSub
##      80      13       7
##
## Node number 1: 127 observations,    complexity param=0.1060606
##   predicted class=Admit     expected loss=0.519685  P(node) =1
##     class counts:    61    29    37
##    probabilities: 0.480 0.228 0.291
##   left son=2 (68 obs) right son=3 (59 obs)
##   Primary splits:
##       State   splits as  LLRRRL, improve=5.5170590, (0 missing)
##       DateSub splits as  LRL,    improve=1.1537370, (0 missing)
##       Gender  splits as  LR,     improve=0.5156044, (0 missing)
##   Surrogate splits:
##       DateSub splits as  LLR, agree=0.551, adj=0.034, (0 split)
##
## Node number 2: 68 observations
##   predicted class=Admit     expected loss=0.3676471  P(node) =0.5354331
##     class counts:    43    13    12
##    probabilities: 0.632 0.191 0.176
##
## Node number 3: 59 observations,    complexity param=0.02272727
##   predicted class=Waitlist  expected loss=0.5762712  P(node) =0.4645669
##     class counts:    18    16    25
##    probabilities: 0.305 0.271 0.424
##   left son=6 (49 obs) right son=7 (10 obs)
##   Primary splits:
##       State   splits as  --LLR-, improve=1.4864750, (0 missing)
##       DateSub splits as  LRL,    improve=0.7052186, (0 missing)
##       Gender  splits as  LR,     improve=0.4061080, (0 missing)
##
## Node number 6: 49 observations,    complexity param=0.02272727
##   predicted class=Waitlist  expected loss=0.6326531  P(node) =0.3858268
##     class counts:    17    14    18
##    probabilities: 0.347 0.286 0.367
##   left son=12 (27 obs) right son=13 (22 obs)
##   Primary splits:
##       Gender  splits as  LR,     improve=1.166564, (0 missing)
##       State   splits as  --LR--, improve=0.341078, (0 missing)
##       DateSub splits as  LRL,    improve=0.084778, (0 missing)
##   Surrogate splits:
##       State splits as  --RL--, agree=0.571, adj=0.045, (0 split)
##
## Node number 7: 10 observations
##   predicted class=Waitlist  expected loss=0.3  P(node) =0.07874016
##     class counts:     1     2     7
##    probabilities: 0.100 0.200 0.700
##
```

```
## Node number 12: 27 observations,    complexity param=0.02272727
##   predicted class=Admit     expected loss=0.6296296  P(node) =0.2125984
##     class counts:    10    10     7
##    probabilities: 0.370 0.370 0.259
##   left son=24 (10 obs) right son=25 (17 obs)
##   Primary splits:
##       DateSub splits as  LRL,    improve=0.4013072, (0 missing)
##       State   splits as  --LR--, improve=0.3172515, (0 missing)
##
## Node number 13: 22 observations
##   predicted class=Waitlist  expected loss=0.5  P(node) =0.1732283
##     class counts:     7     4    11
##    probabilities: 0.318 0.182 0.500
##
## Node number 24: 10 observations
##   predicted class=Admit     expected loss=0.5  P(node) =0.07874016
##     class counts:     5     3     2
##    probabilities: 0.500 0.300 0.200
##
## Node number 25: 17 observations
##   predicted class=Decline   expected loss=0.5882353  P(node) =0.1338583
##     class counts:     5     7     5
##    probabilities: 0.294 0.412 0.294
```

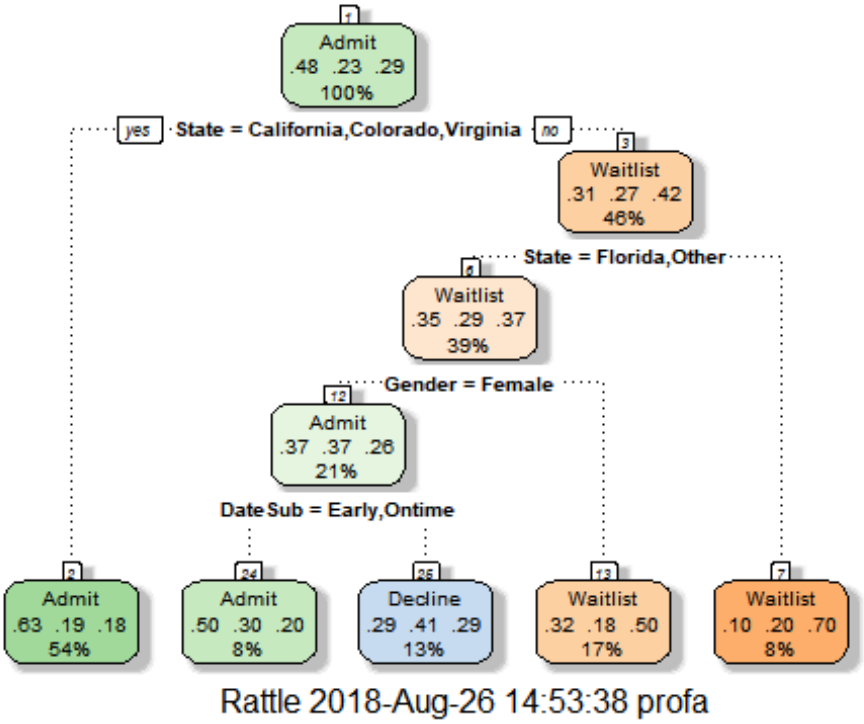
```
predicted4= predict(Treefit4,DT_Test_Focus, type="class")
(Results4 <- data.frame(Predicted=predicted4,Actual=DT_Test_GLabs))
```

```
##     Predicted   Actual
## 1       Admit    Admit
## 8       Admit    Admit
## 15   Waitlist    Admit
## 22      Admit    Admit
## 29      Admit    Admit
## 36      Admit    Admit
## 43    Decline    Admit
## 50      Admit    Admit
## 57    Decline  Decline
## 64   Waitlist  Decline
## 71      Admit    Admit
## 78   Waitlist  Decline
## 85      Admit  Decline
## 92      Admit  Decline
## 99      Admit  Decline
## 106  Waitlist Waitlist
## 113     Admit Waitlist
## 120   Decline Waitlist
## 127  Waitlist Waitlist
## 134  Waitlist Waitlist
## 141     Admit Waitlist
## 148  Waitlist Waitlist
```

```
(table(Results4))
```

```
##           Actual
## Predicted  Admit Decline Waitlist
##   Admit        7       3        2
##   Decline      1       1        1
##   Waitlist     1       2        4
```

```
fancyRpartPlot(Treefit4)
```




```
## This offers some insight....
## Florida has a higher Decline rate


## ------------------------------------------------------------------------
-
##  Clustering
## ------------------------------------------------------------------------
(head(CleanStudentDF,n=5))
```

```
##     DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 1 September Female California 3.90     6.7      962    NA    Admit
## 2 September Female    Florida 3.80     1.4      969    97    Admit
## 3   October   Male California 3.80     2.3      970    NA    Admit
## 4   October   Male   Colorado 3.60     0.9      969    NA    Admit
## 5  November   Male   Colorado 3.92     1.2      969    95    Admit
```

```r
X <- CleanStudentDF
## Make sure there are no NAs
(sum(is.na(CleanStudentDF$GPA)))
```

```
## [1] 0
```

```r
(sum(is.na(CleanStudentDF$MathTest)))
```

```
## [1] 0
```

```r
(sum(is.na(CleanStudentDF$WorkExp)))
```

```
## [1] 0
```

```r
(sum(is.na(CleanStudentDF$Essay)))
```

```
## [1] 74
```

```r
## Essay has 74 NA values....
## Remember, we cluster with numerical data and so my next step
## is to remove the non-numerical columns
(head(X,n=10))
```

```
##      DateSub Gender      State  GPA WorkExp MathTest Essay Decision
## 1  September Female California 3.90     6.7      962    NA    Admit
## 2  September Female    Florida 3.80     1.4      969    97    Admit
## 3    October   Male California 3.80     2.3      970    NA    Admit
## 4    October   Male   Colorado 3.60     0.9      969    NA    Admit
## 5   November   Male   Colorado 3.92     1.2      969    95    Admit
## 6   November   Male California 3.80     1.2      967    NA    Admit
## 7   November Female California 3.88     0.0      967    NA    Admit
## 8   December Female California 3.70     1.2      969    NA    Admit
## 9    October Female    Florida 3.90     4.7      961    NA    Admit
## 10  December Female California 3.70     1.4      966    94    Admit
```

```r
X <- X[,-c(1,2,3,7,8)] ## Here I have removed Essay
(head(X,n=5))
```

```
##    GPA WorkExp MathTest
## 1 3.90     6.7      962
## 2 3.80     1.4      969
## 3 3.80     2.3      970
## 4 3.60     0.9      969
## 5 3.92     1.2      969
```

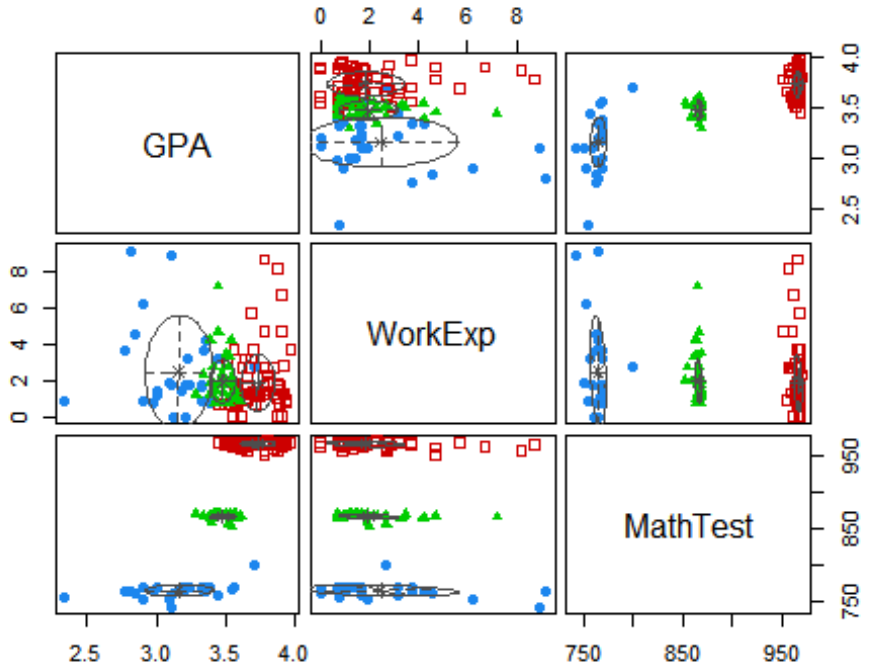
```r
## When Clustering, there are many options.
## OPtion 1 - I will choose the number of clusters as 3
ClusFIT1 <- Mclust(X,G=3)
(ClusFIT1)
```

```
## 'Mclust' model object:
##  best model: ellipsoidal, equal shape and orientation (VEE) with 3 compone
nts
```

```
summary(ClusFIT1)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VEE (ellipsoidal, equal shape and orientation) model with 3 compone
nts:
##
##  log.likelihood   n df       BIC       ICL
##       -766.6072 149 19 -1628.289 -1628.289
##
## Clustering table:
##  1  2  3
## 29 76 44
```

```
plot(ClusFIT1, what = "classification")
```



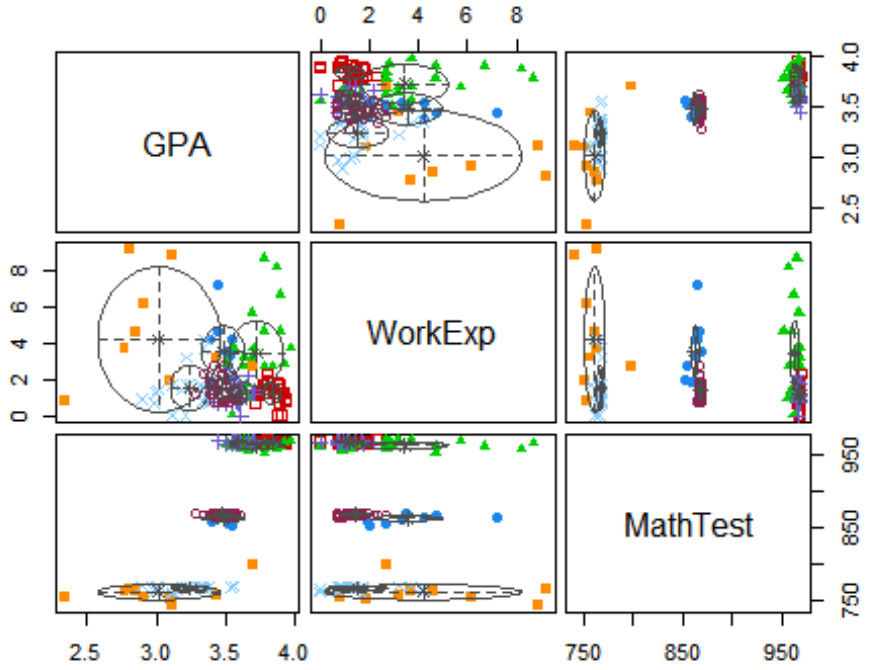

```
## OPtion 2 - cluster with no k selected
ClusFIT2 <- Mclust(X)
(ClusFIT2)
```

```
## 'Mclust' model object:
##  best model: diagonal, equal shape (VEI) with 7 components
```

```
summary(ClusFIT2)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VEI (diagonal, equal shape) model with 7 components:
##
##  log.likelihood   n df       BIC      ICL
##       -666.9693 149 36 -1514.081 -1532.74
##
## Clustering table:
##  1  2  3  4  5  6  7
## 10 32 22 22  9 20 34
```

```
plot(ClusFIT2, what = "classification")
```




```
######### Now - let's look at k - means

######################################################################
##   Mclust vs. k-means
##
## mclust above presents normal-model-based EM clustering,
## wherease k-means is not dependent on any type of the distribution,
## it is not model-based. Mclust will use a soft assignment,
## whereas k-means uses a hard assignment. The mclust is a contributed
## R package for model-based clustering, classification, and density
## estimation based on finite normal mixture modelling. It provides
## functions for parameter estimation via the EM algorithm for normal
```

```
## mixture models with a variety of covariance structures.
## Details: https://cran.r-project.org/web/packages/mclust/vignettes/mclust.h
tml
##
##  ALternatively, K means will start with the assumption that a given data
## point belongs to one cluster.
#####################################################################

#########  k means ---------------------------------
## Recall - we have:
(head(X))

##    GPA WorkExp MathTest
## 1 3.90     6.7      962
## 2 3.80     1.4      969
## 3 3.80     2.3      970
## 4 3.60     0.9      969
## 5 3.92     1.2      969
## 6 3.80     1.2      967

kmeansFIT1 <- kmeans(X,3)
(kmeansFIT1)

## K-means clustering with 3 clusters of sizes 12, 64, 73
##
## Cluster means:
##        GPA  WorkExp MathTest
## 1 3.735000 3.108333 959.4167
## 2 3.731094 1.607813 967.7031
## 3 3.352329 2.154795 825.8493
##
## Clustering vector:
##   [1] 1 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 3 2 1 2 2 2 2 2 2 2 1 2 1 2 2 2 2
2
##  [36] 2 2 2 1 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 1 2 2 2 2 2
2
##  [71] 2 2 1 2 2 2 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3
## [106] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3
## [141] 3 3 3 3 3 3 3 3 3
##
## Within cluster sum of squares by cluster:
## [1]    191.7151    245.7625 183347.7000
##  (between_SS / total_SS =  80.0 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
```

```
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```
summary(kmeansFIT1)
```

```
##              Length Class  Mode
## cluster      149    -none- numeric
## centers        9    -none- numeric
## totss          1    -none- numeric
## withinss       3    -none- numeric
## tot.withinss   1    -none- numeric
## betweenss      1    -none- numeric
## size           3    -none- numeric
## iter           1    -none- numeric
## ifault         1    -none- numeric
```

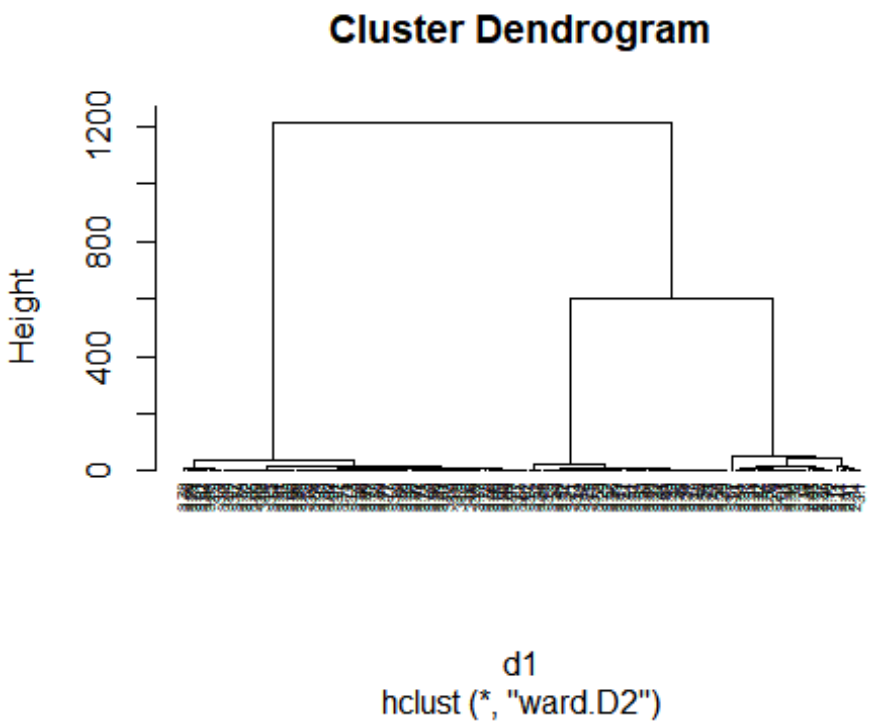
```
# get cluster means
aggregate(X,by=list(kmeansFIT1$cluster),FUN=mean)
```

```
##   Group.1      GPA  WorkExp MathTest
## 1       1 3.735000 3.108333 959.4167
## 2       2 3.731094 1.607813 967.7031
## 3       3 3.352329 2.154795 825.8493
```

```
X2 <- data.frame(X, kmeansFIT1$cluster)
(head(X2,n=10))
```

```
##     GPA WorkExp MathTest kmeansFIT1.cluster
## 1  3.90     6.7      962                  1
## 2  3.80     1.4      969                  2
## 3  3.80     2.3      970                  2
## 4  3.60     0.9      969                  2
## 5  3.92     1.2      969                  2
## 6  3.80     1.2      967                  2
## 7  3.88     0.0      967                  2
## 8  3.70     1.2      969                  2
## 9  3.90     4.7      961                  1
## 10 3.70     1.4      966                  2
```

```
## The aggregate really shows the cut-offs for the groups.

## Hierarchical Cluster - this one is ugly because there are too many GPAs
d1 <- dist(X, method = "euclidean") # distance matrix
fit1 <- hclust(d1, method="ward.D2")
plot(fit1, labels = X$GPA,hang = -.2, cex = 0.5) # display dendogram
```




```
######### Cosine similarity
## To use Cos Sim and library lsa, you will need
## to place your data into a matrix
## Let's look at our data:
(head(X,n=10))
```

```
##     GPA WorkExp MathTest
## 1  3.90     6.7      962
## 2  3.80     1.4      969
## 3  3.80     2.3      970
## 4  3.60     0.9      969
## 5  3.92     1.2      969
## 6  3.80     1.2      967
## 7  3.88     0.0      967
## 8  3.70     1.2      969
## 9  3.90     4.7      961
## 10 3.70     1.4      966
```

```
## Now, convert this to a matrix
X_Matrix <- as.matrix(X)
Cos_SimMatrix <- cosine(X_Matrix)
diag(Cos_SimMatrix) <- 0 # Remove relationship with self
Cos_SimMatrix
```

```
##               GPA   WorkExp  MathTest
## GPA      0.000000 0.7501660 0.9987190
```

```
## WorkExp  0.750166 0.0000000 0.7473097
## MathTest 0.998719 0.7473097 0.0000000

# Prune edges of the tree
edgeLimit <- .70
Cos_SimMatrix[(Cos_SimMatrix < edgeLimit)] <- 0

## Make the network
(Cos_Sim_Network <- graph_from_adjacency_matrix(Cos_SimMatrix,
                                                mode = 'undirected',
                                                weighted = T))

## IGRAPH 5841bff UNW- 3 3 --
## + attr: name (v/c), weight (e/n)
## + edges from 5841bff (vertex names):
## [1] GPA    --WorkExp  GPA    --MathTest WorkExp--MathTest

##plot the network
plot(Cos_Sim_Network)
```


```
tkplot(Cos_Sim_Network, vertex.color="yellow")

## [1] 1

## This shows that the cosine Sim is high between GPA, WOrkExp
## and MathTest
```

```
## Network Refs
## http://www.business-science.io/business/2016/10/01/CustomerSegmentationPt3
.html
## http://igraph.org/r/doc/plot.common.html
```