Homwork 1 - Stats 4620

Tyler Poelking and Karen Somes

9/4/2017

Question 1:

a) Flexible statistical learning methods favor a dataset with a large number of observations, because larger n's minimize the likelihood of overfitting the data.

b) An less flexible statistical learning method is preferred here, since, a small number of observations increases the likelihood of overfitting the data. A less flexible model will also increase bias, which will protect the model from adhering to meaningless noise in the data.

c) If the relationship between the predictors and the response is highly non-linear, a flexible statistical learning method is better because more intricate functions of the predictors are required to properly estimate the response, and flexible methods allow for such functions.

d) An inflexible method is better, because a flexible method will capture much of the useless noise in the data, thus causing overfit and poor performance on non-training data.

```
#Part A
library(readr)
college <- read_csv("~/Desktop/All Stuff/School Stuff/STATS/Data/College.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
fix (college )
```

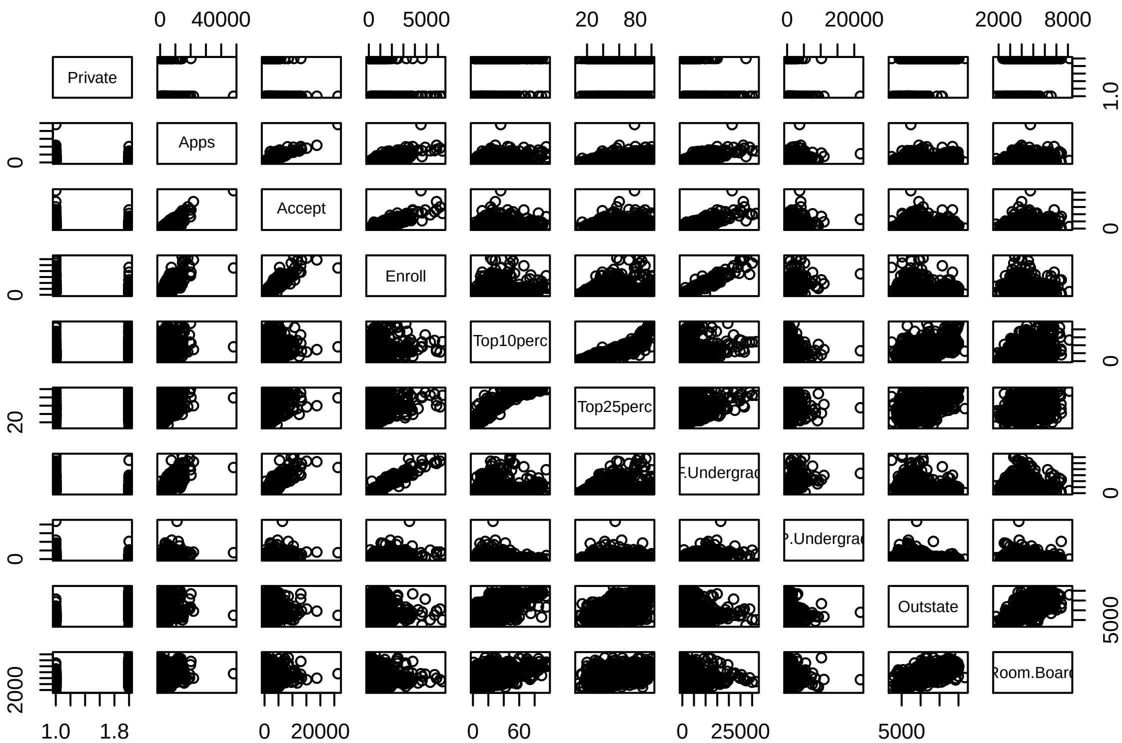
```
#Part B
rownames(college)=college[,1]

college =college [,-1]
#fix (college )
```

```
#Part C

#i.
summary(college)
```

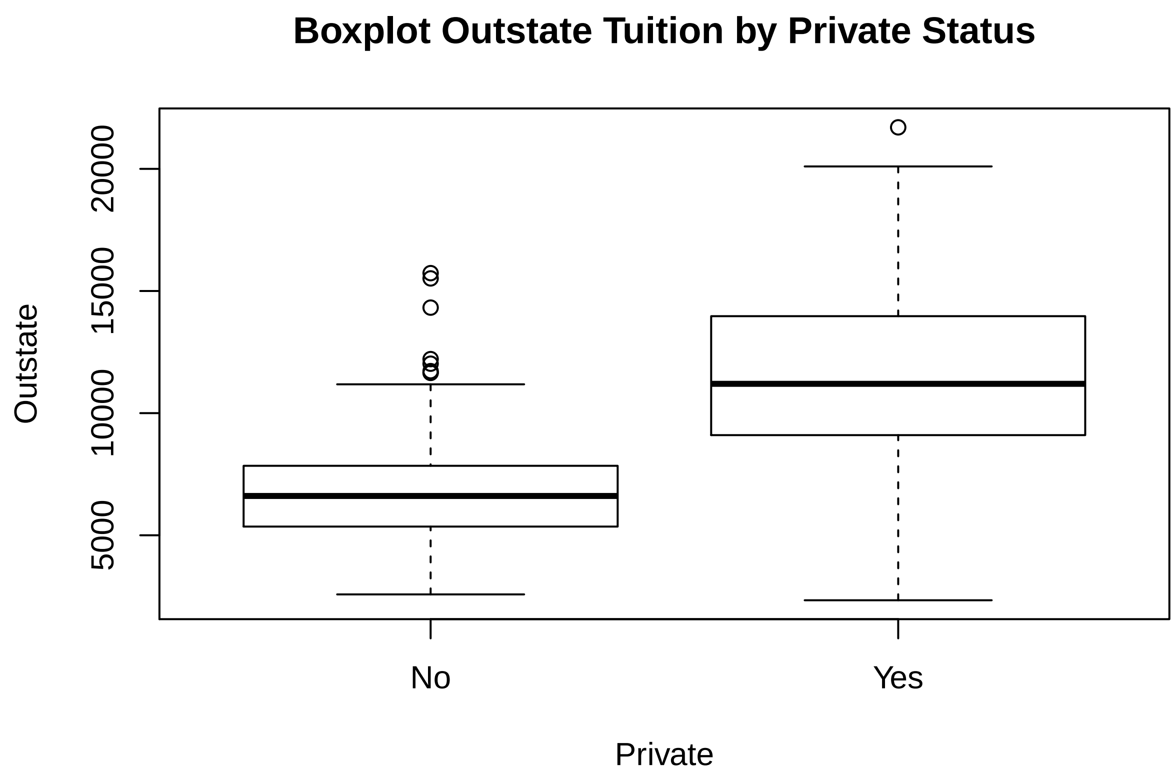
```
##    Private              Apps           Accept          Enroll
##  Length:777         Min.   :   81   Min.   :   72   Min.   :  35
##  Class :character   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242
##  Mode  :character   Median : 1558   Median : 1110   Median : 434
##                     Mean   : 3002   Mean   : 2019   Mean   : 780
##                     3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902
##                     Max.   :48094   Max.   :26330   Max.   :6392
##    Top10perc       Top25perc      F.Undergrad     P.Undergrad
##  Min.   : 1.00   Min.   :  9.0   Min.   :  139   Min.   :    1.0
##  1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0
##  Median :23.00   Median : 54.0   Median : 1707   Median :  353.0
##  Mean   :27.56   Mean   : 55.8   Mean   : 3700   Mean   :  855.3
##  3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0
##  Max.   :96.00   Max.   :100.0   Max.   :31643   Max.   :21836.0
##     Outstate       Room.Board       Books           Personal
##  Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250
```

```
##  1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850
##  Median : 9990   Median :4200   Median : 500.0   Median :1200
##  Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341
##  3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700
##  Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800
##       PhD            Terminal       S.F.Ratio      perc.alumni
##  Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00
##  1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00
##  Median : 75.00   Median : 82.0   Median :13.60   Median :21.00
##  Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74
##  3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00
##  Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00
##      Expend         Grad.Rate
##  Min.   : 3186   Min.   : 10.00
##  1st Qu.: 6751   1st Qu.: 53.00
##  Median : 8377   Median : 65.00
##  Mean   : 9660   Mean   : 65.46
##  3rd Qu.:10830   3rd Qu.: 78.00
##  Max.   :56233   Max.   :118.00
```

```r
#ii.
college$Private =as.factor(college$Private)
attach(college)
A = college[,1:10]
pairs(A)
```




```r
#iii.
#WORKS
plot(Private, Outstate, main="Boxplot Outstate Tuition by Private Status",
    xlab="Private", ylab="Outstate")
```

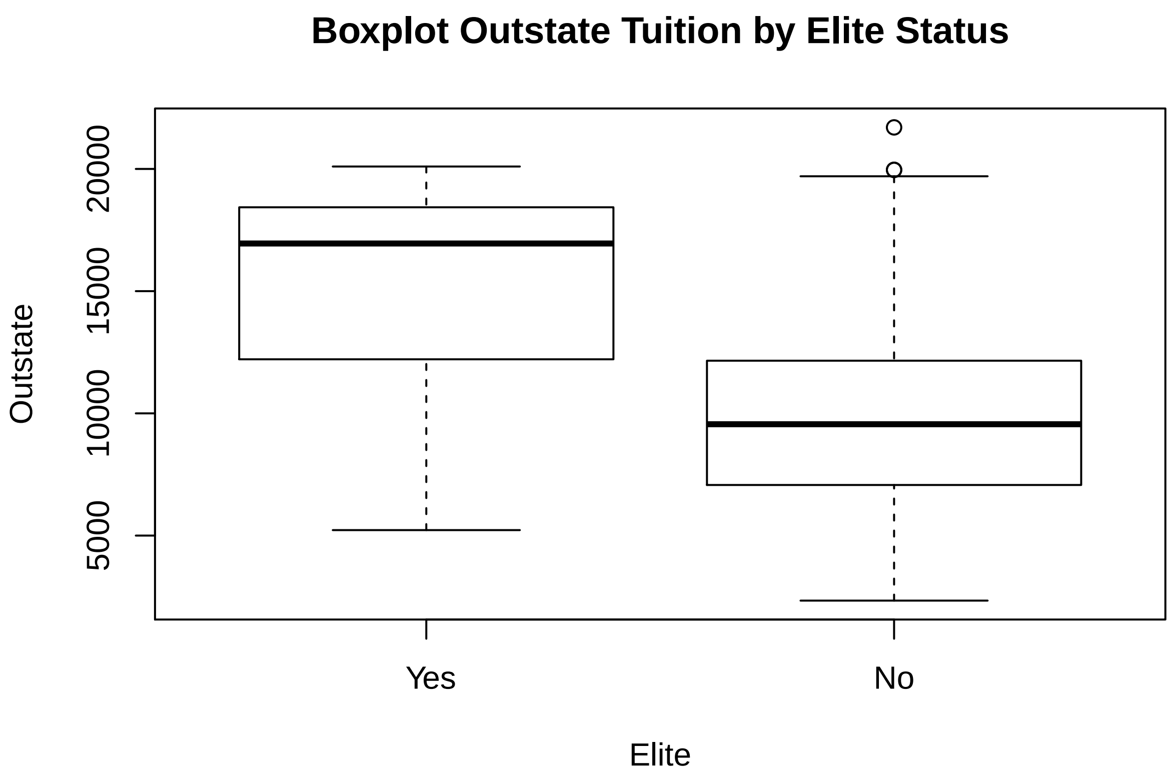



```
#iv.
Elite =rep("No",nrow(college ))
Elite [college$Top10perc >50]=" Yes"
college =data.frame(college ,Elite)
college$Elite =as.factor(college$Elite)
summary(college)
```

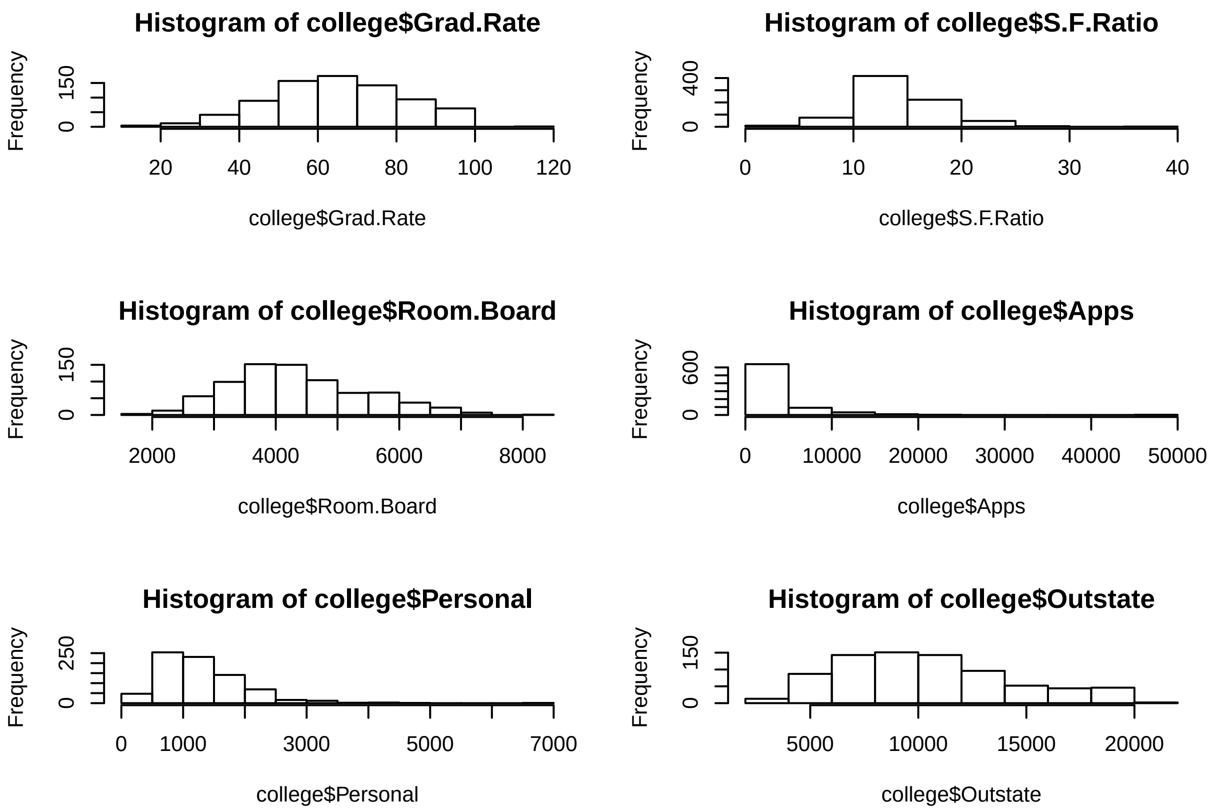
```
##  Private        Apps            Accept          Enroll        Top10perc
##  No :212    Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565    1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##              Median : 1558   Median : 1110   Median : 434   Median :23.00
##              Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
##              3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
##              Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
##    Top25perc      F.Undergrad     P.Undergrad          Outstate
##  Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books           Personal          PhD
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
##  Median :4200   Median : 500.0   Median :1200   Median : 75.00
##  Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
##  3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
##  Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
##     Terminal        S.F.Ratio      perc.alumni        Expend
##  Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
```

```
##  1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
##  Median : 82.0   Median :13.60   Median :21.00   Median : 8377
##  Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
##  3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
##  Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
##    Grad.Rate       Elite
##  Min.   : 10.00   Yes: 78
##  1st Qu.: 53.00   No  :699
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

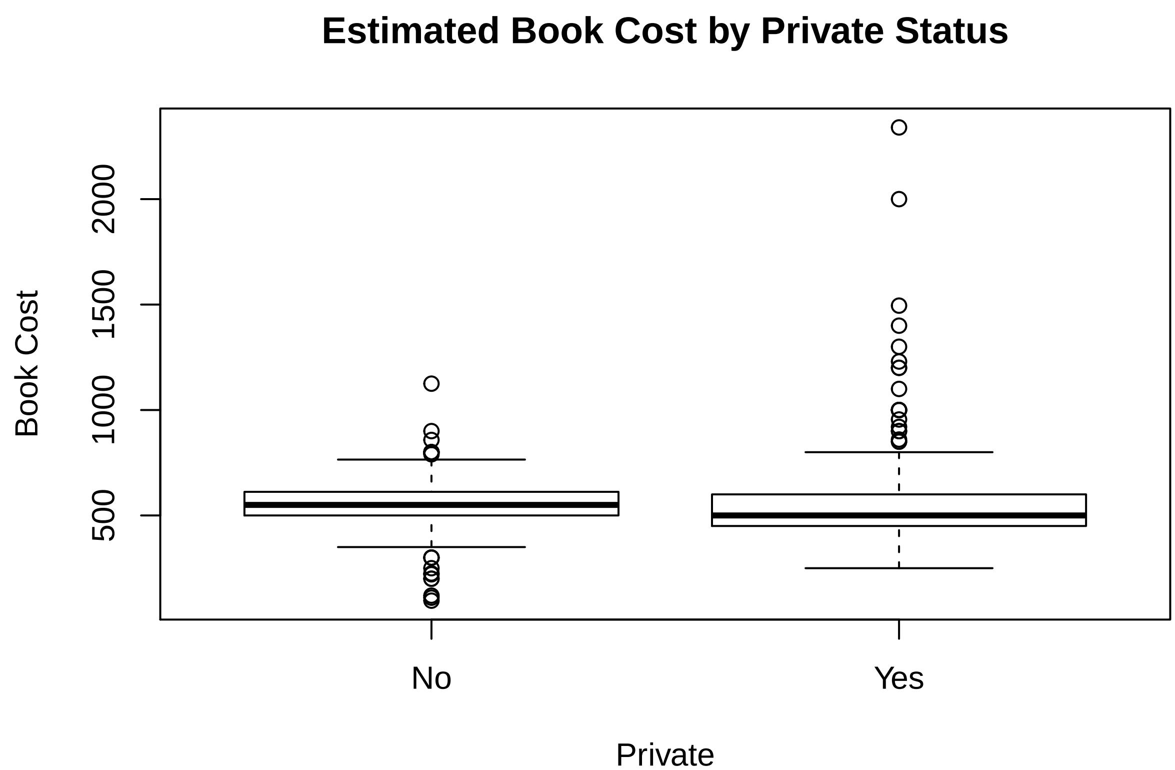
```
boxplot(Outstate~Elite, main="Boxplot Outstate Tuition by Elite Status",
    xlab="Elite", ylab="Outstate")
```



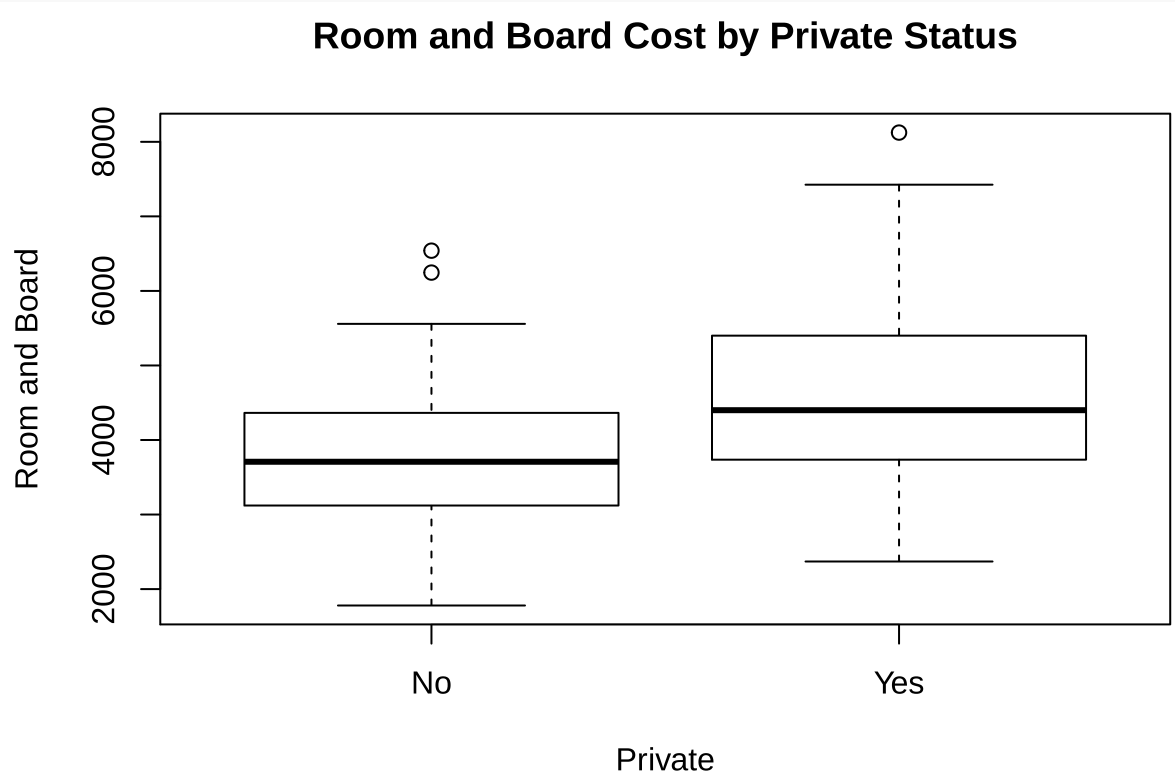

```
#v.
par(mfrow=c(3,2))
hist(college$Grad.Rate)
hist(college$S.F.Ratio)
hist(college$Room.Board)
hist(college$Apps)
hist(college$Personal)
hist(college$Outstate)
```



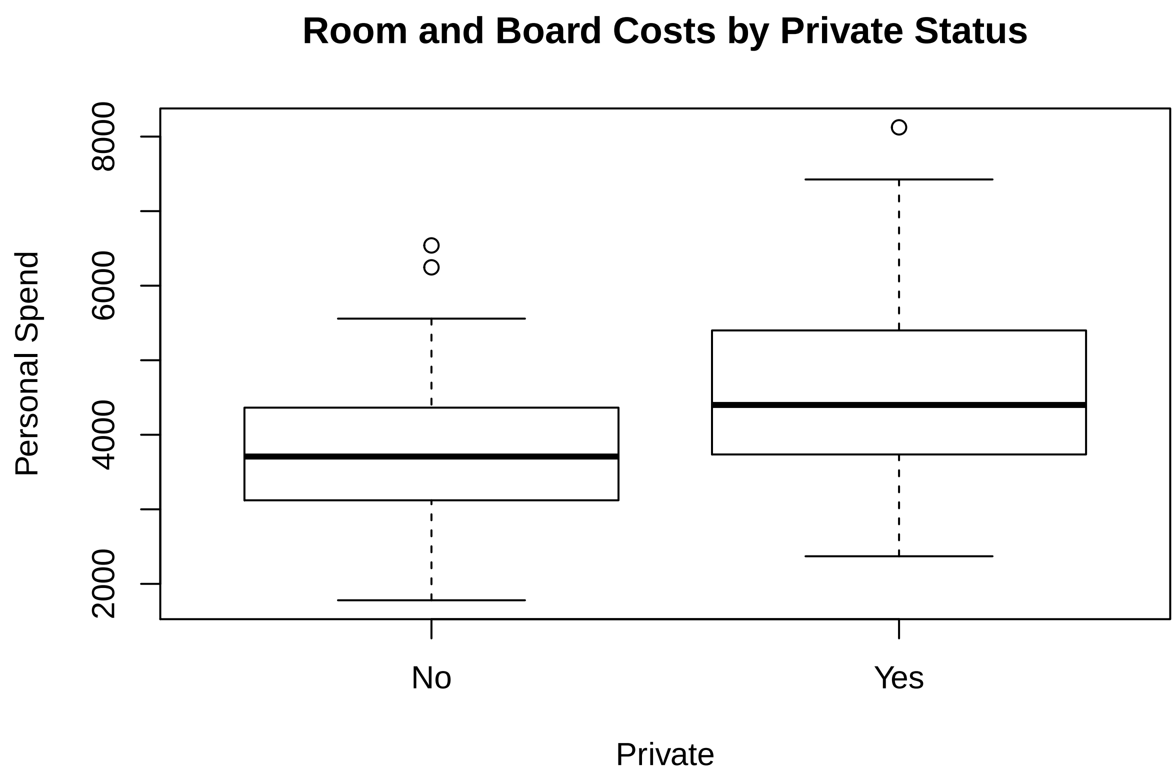

```r
#vi.

#Look into other cost variables and how Private status affects
par(mfrow=c(1,1))
boxplot(Books~Private, main="Estimated Book Cost by Private Status",
    xlab="Private", ylab="Book Cost")
```



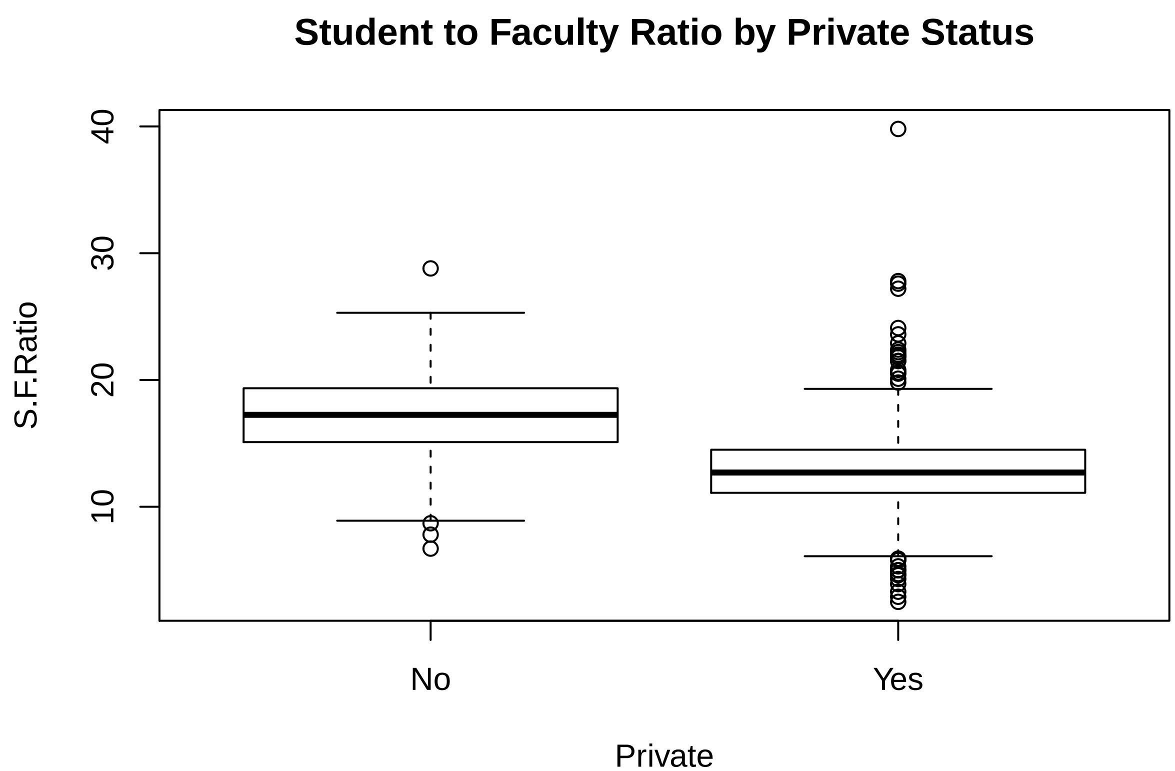

```
boxplot(Room.Board~Private, main="Room and Board Cost by Private Status",
    xlab="Private", ylab="Room and Board")
```



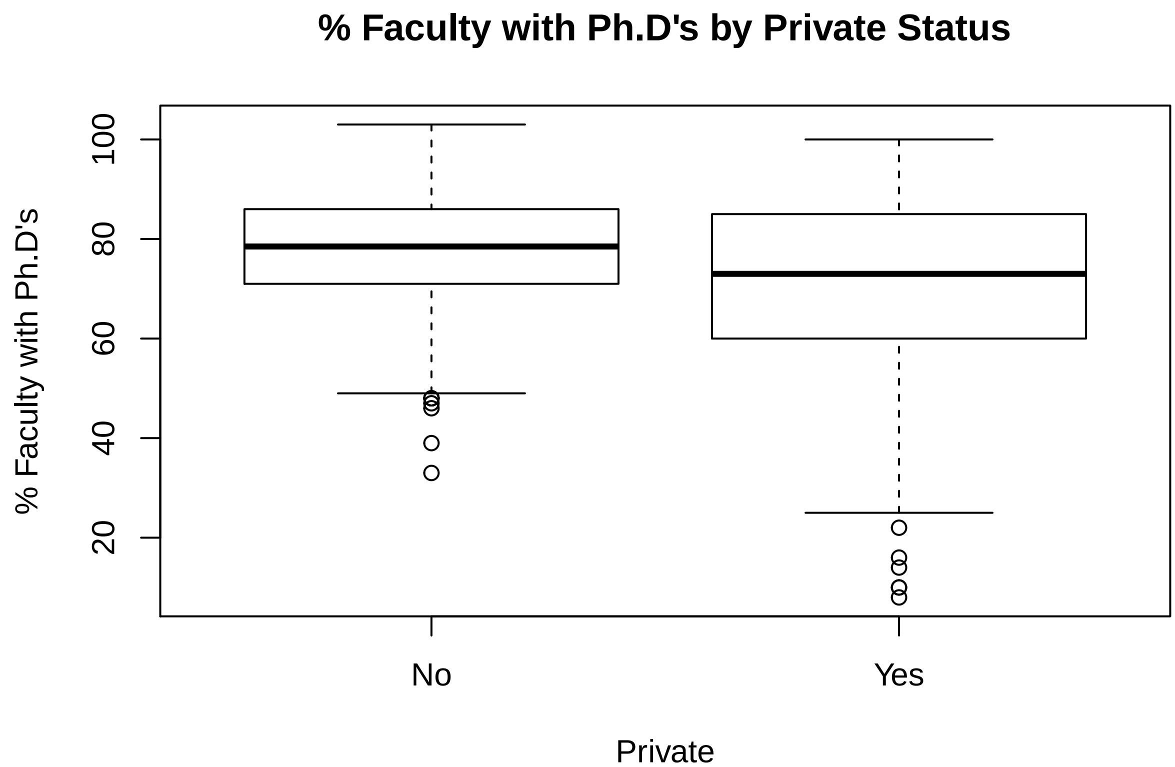

```
boxplot(Room.Board~Private, main="Room and Board Costs by Private Status",
    xlab="Private", ylab="Personal Spend")
```



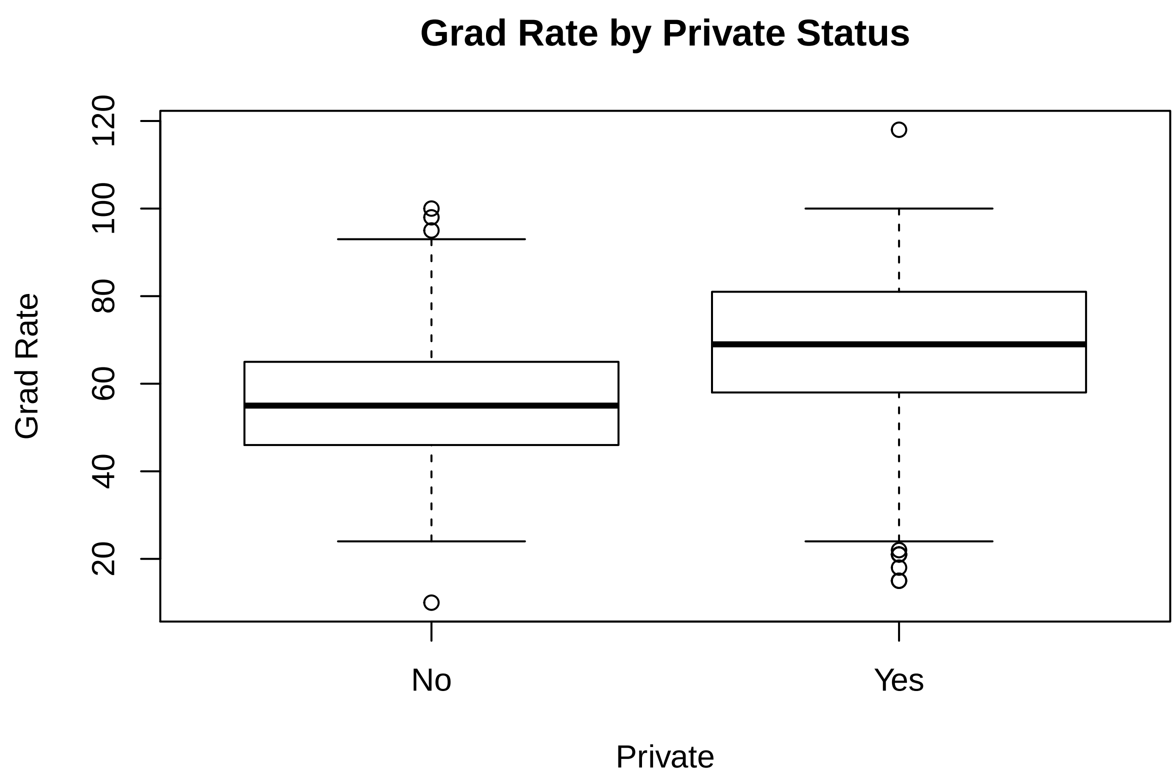

```r
#Are private school students getting more for their $?
boxplot(S.F.Ratio~Private, main="Student to Faculty Ratio by Private Status",
    xlab="Private", ylab="S.F.Ratio")
```




```r
boxplot(PhD~Private, main="% Faculty with Ph.D's by Private Status",
    xlab="Private", ylab="% Faculty with Ph.D's")
```




```r
boxplot(Grad.Rate~Private, main="Grad Rate by Private Status",
    xlab="Private", ylab="Grad Rate")
```



```r
boxplot(Expend~Private, main="Instructional Expenditure per Student by Private Status",
    xlab="Private", ylab="Instructional Expenditure per Student")
```

Of the initial 10 features forming the scatterplot matrix, continuous features that had moderate to high positive correlation were: Enroll+F.Undergrad, Accept+Enroll, Accept+F.Undergrad, Accept+Apps, Apps+Enroll, Top10perc+Top25perc, and Outstate+Room.Board. Top10perc+F.Undergrad as well as Top25perc+F.Undergrad also seems to have positive correlation, but not as strong.

My analysis included exploring how whether or not a college is private affects the various types of costs associated with it. Based on the above boxplots, private colleges have higher state tuition than non-private colleges. Student Book Costs have greater variance for private schools but on average students spend slightly less on books. Room and board costs more on average for private schools. And lastly, personal spend is estimated to be higher on average for private schools.

Private school have a smaller Student to Faculty Ratio, a higher Graduation Rate and a higher Instructional Expenditure per Student amount. However, the average Percent Faculty with Ph.D's is smaller for private schools, which comes as a suprise, since one might assume more elite professors with higher creditials come with a school that costs more money.

```r
load('~/Desktop/All Stuff/School Stuff/STATS/Data/credit.Rdata')
print(length(newcredit))
```

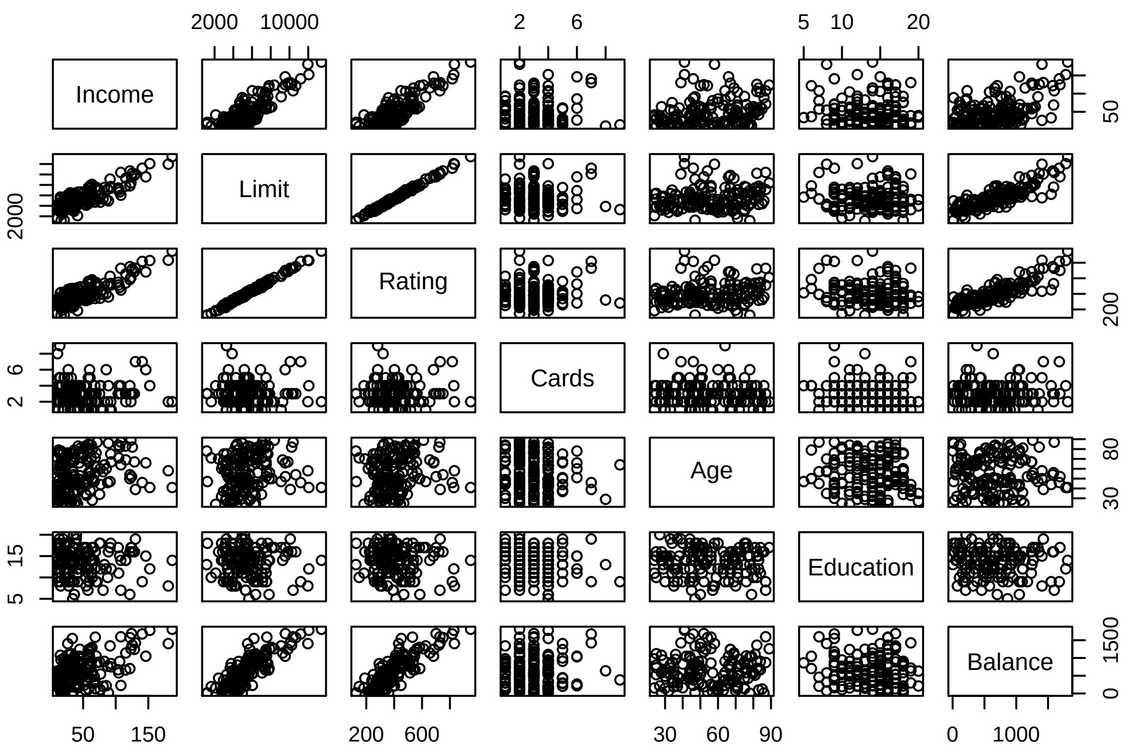
```
## [1] 11
```

```r
#Summary for initial analysis
summary(newcredit)
```

```
##      Income           Limit           Rating          Cards
##  Min.   : 10.63   Min.   : 1160   Min.   :126.0   Min.   :1
##  1st Qu.: 23.73   1st Qu.: 3914   1st Qu.:301.0   1st Qu.:2
##  Median : 39.42   Median : 5198   Median :383.0   Median :3
##  Mean   : 50.35   Mean   : 5499   Mean   :406.6   Mean   :3
```

```
## 3rd Qu.: 63.73   3rd Qu.: 6438   3rd Qu.:465.5   3rd Qu.:4
## Max.   :186.63   Max.   :13414   Max.   :949.0   Max.   :9
##       Age          Education        Gender    Student   Married
## Min.   :24.00   Min.   : 5.00    Male :73   No :137   No :58
## 1st Qu.:41.50   1st Qu.:11.00   Female:82   Yes: 18   Yes:97
## Median :53.00   Median :14.00
## Mean   :55.23   Mean   :13.61
## 3rd Qu.:70.00   3rd Qu.:16.00
## Max.   :89.00   Max.   :20.00
##              Ethnicity     Balance
## African American:39   Min.   :   5.0
## Asian           :42   1st Qu.: 332.0
## Caucasian       :74   Median : 606.0
##                       Mean   : 666.3
##                       3rd Qu.: 916.5
##                       Max.   :1809.0
```

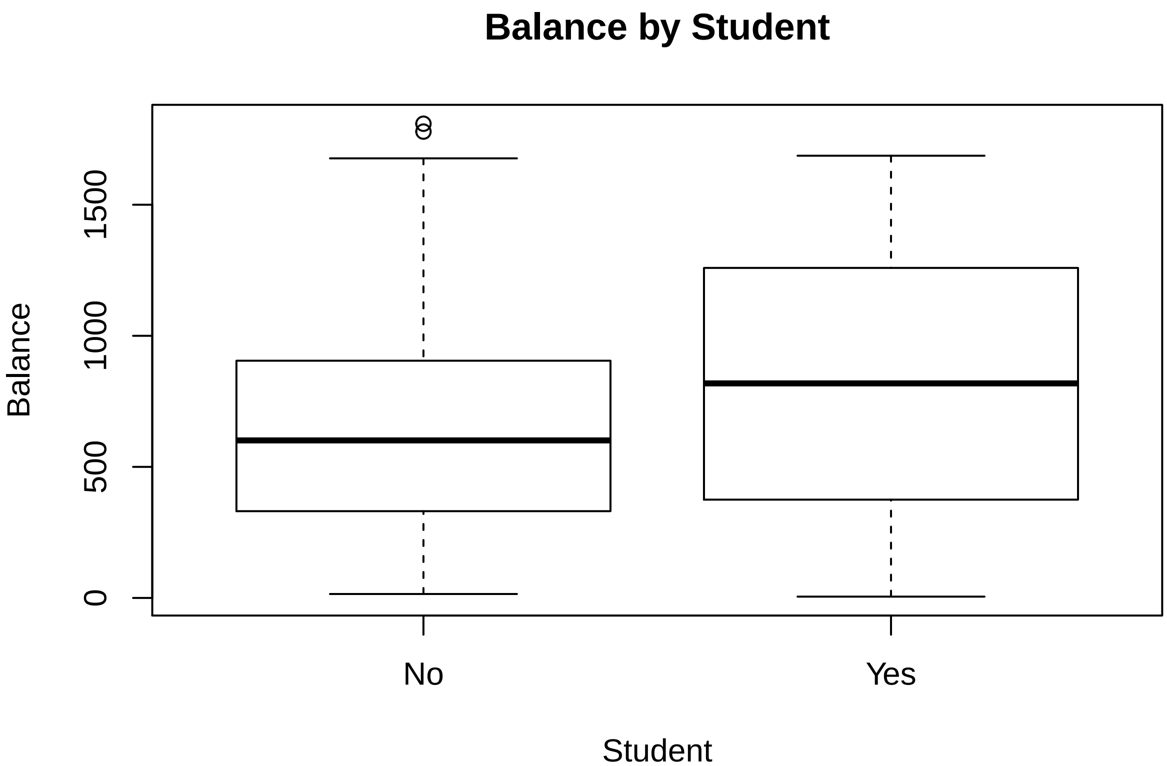
```
keeps <- c("Income", "Limit", "Rating", "Cards", "Age", "Education", "Balance")
newcreditCont = newcredit[keeps]

#newcredit$Private =as.factor(college$Private)
attach(newcredit)

#Scatterplot variables
pairs(newcreditCont)
```



Of all the continuous variables, Limit and Rating appear to have the strongest (positive) correlation with Balance. Both correlations seem almost equal, which is not a suprise, since Limit and Rating themselves have an extremely high correlation between each other. Income is also highly correlated. My assumption is that we will only need one of these in the model as to avoid multicollinearity. The reltionship appears linear but it may be of a higher or lower degree, we will have to test this. Cards, Age, and Education don't have high correlation, so the degree of information each of these would add to our model stands questionable.

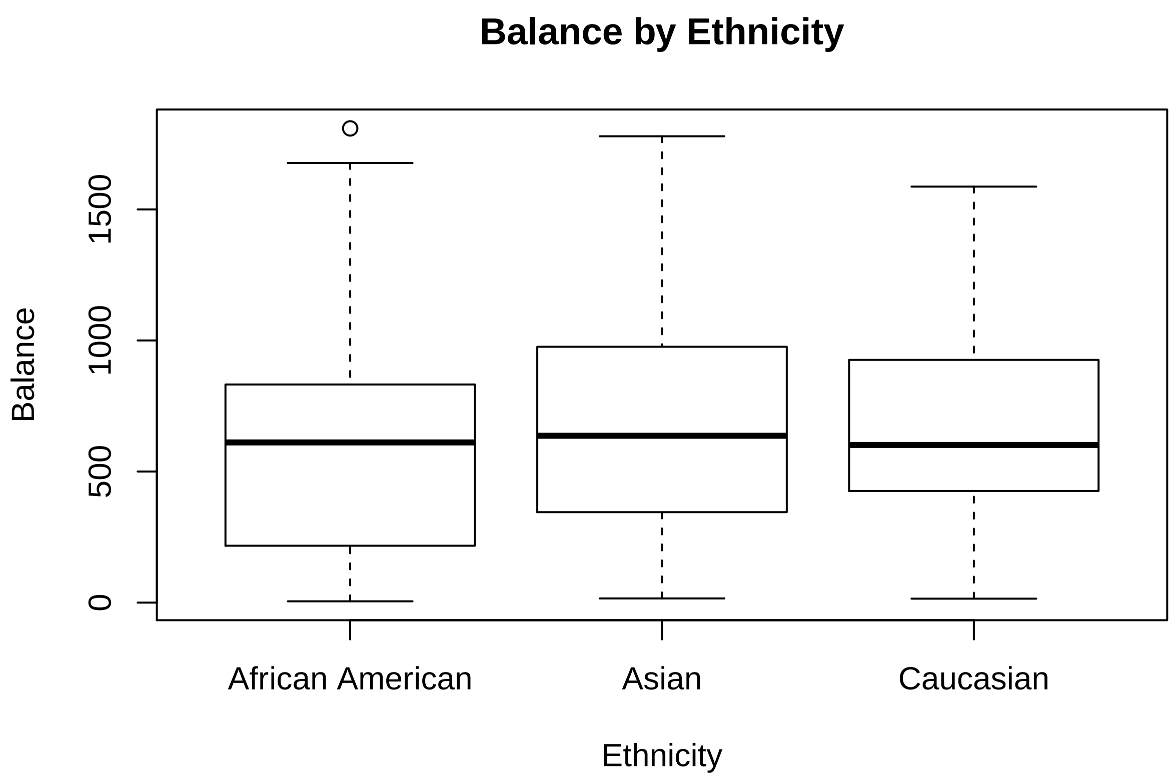
Categorical variables and their affects on Balance are not easily analyized in a scatterplot such as the one above. We are going to construct boxplots, plotting Balance against these each categorical variable, as seen below. These charts are much more interpretable and make analysis easier.

```
#Student
boxplot(Balance~Student, main="Balance by Student",
    xlab="Student", ylab="Balance")
```




```
#Married
boxplot(Balance~Married, main="Balance by Married",
    xlab="Married", ylab="Balance")
```


```r
#Ethnicity
boxplot(Balance~Ethnicity, main="Balance by Ethnicity",
    xlab="Ethnicity", ylab="Balance")
```



The scale of the y-axis on each of these plots is the same, which allows us to compare between plots. Of the three categorical variables charted, the Student has the most prominent difference in meen balance and thus might contribute the most to our model. The mean Balance between ethnicities varries some too, so we may still use Ethnicity. The two box and whiskers in the 'Balance by Married' chart, however, are quite similar, indicating marrital status does not impact Balance.

```r
#Rating? or Limit?
lmfit = lm( Balance ~ Rating)
summary(lmfit)
```

```
##
## Call:
## lm(formula = Balance ~ Rating)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -546.67 -149.83   14.66  143.87  770.27
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -363.3467    54.2676  -6.695 3.84e-10 ***
## Rating         2.5323     0.1259  20.111  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 224.1 on 153 degrees of freedom
## Multiple R-squared:  0.7255, Adjusted R-squared:  0.7237
## F-statistic: 404.5 on 1 and 153 DF,  p-value: < 2.2e-16
```

```r
lmfit2 = lm( Balance ~ Limit)
summary(lmfit2)
```

```
##
## Call:
## lm(formula = Balance ~ Limit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -559.20 -153.44    7.14  134.55  763.76
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.585e+02  5.049e+01   -5.12 9.08e-07 ***
## Limit        1.682e-01  8.557e-03   19.65  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 227.8 on 153 degrees of freedom
## Multiple R-squared:  0.7162, Adjusted R-squared:  0.7144
## F-statistic: 386.2 on 1 and 153 DF,  p-value: < 2.2e-16
```

The model using Rating resulted in a smaller Residual standard error and larger Multiple R-squared (though they were both close, of course), so we will stick with that.

```r
#Seeing how addition of student affects fit
lmfit3 = lm( Balance ~ Rating+ Student)
summary(lmfit3)
```

```
##
```

```
## Call:
## lm(formula = Balance ~ Rating + Student)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -510.9 -123.5   11.3  145.1  451.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -456.3793    48.7691  -9.358  < 2e-16 ***
## Rating         2.6598     0.1105  24.062  < 2e-16 ***
## StudentYes   354.8117    49.3132   7.195 2.66e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 194.2 on 152 degrees of freedom
## Multiple R-squared:  0.7953, Adjusted R-squared:  0.7926
## F-statistic: 295.2 on 2 and 152 DF,  p-value: < 2.2e-16
```

Adding student decreased our RSE and increased our R-squared, awesome, definitely going to add to our model. Now that we have the essentials, lets see how the remaining variables effect the simple linear model that models Balance.

```r
#Income
model = lm( Balance ~ Rating+ Student + Income)
print(sprintf('%s : %s' ,'Income', summary(model)$sigma))
```

```
## [1] "Income : 53.1069024261126"
```

```r
#Married
model = lm( Balance ~ Rating+ Student + Married)
print(sprintf('%s : %s' ,'Married', summary(model)$sigma))
```

```
## [1] "Married : 194.715559567576"
```

```r
#Ethnicity
model = lm( Balance ~ Rating+ Student + Ethnicity)
print(sprintf('%s : %s' ,'Ethnicity', summary(model)$sigma))
```

```
## [1] "Ethnicity : 189.824327630236"
```

```r
#Cards
model = lm( Balance ~ Rating+ Student + Cards)
print(sprintf('%s : %s' ,'Cards', summary(model)$sigma))
```

```
## [1] "Cards : 194.377242212607"
```

```r
#Age
model = lm( Balance ~ Rating+ Student + Age)
print(sprintf('%s : %s' ,'Age', summary(model)$sigma))
```

```
## [1] "Age : 184.517947231624"
```

```r
#Education
model = lm( Balance ~ Rating+ Education + Age)
print(sprintf('%s : %s' ,'Education', summary(model)$sigma))
```

```
## [1] "Education : 218.416374283862"
```

Income is certainly a must add. From here, we will use an anova table to explore the addition of any other variables. I'll form the anova table in order of the variables above that corresponded to the smallest residual standard error first.

```
#lmfit5 = lm( Balance ~ I(Rating^.8)+ Student + Income + Age + Ethnicity + Cards + Married + Education)
lmfit4 = lm( Balance ~ Rating+ Student + Income + Age + Ethnicity + Cards + Married + Education)
anova(lmfit4)
```

```
## Analysis of Variance Table
##
## Response: Balance
##            Df   Sum Sq  Mean Sq   F value    Pr(>F)
## Rating      1 20304749 20304749 8560.6710 < 2.2e-16 ***
## Student     1  1951440  1951440  822.7453 < 2.2e-16 ***
## Income      1  5303792  5303792 2236.1279 < 2.2e-16 ***
## Age         1    66888    66888   28.2006 4.027e-07 ***
## Ethnicity   2    11001     5500    2.3190    0.1020
## Cards       1      624      624    0.2632    0.6087
## Married     1     1601     1601    0.6751    0.4126
## Education   1     1837     1837    0.7747    0.3802
## Residuals 145   343920     2372
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

None of the other variables after Income add significant information to the model according to the F-statistic.

```
finalModel1 =lm( Balance ~ Rating+ Student + Income + Age)
summary(finalModel1)
```

```
##
## Call:
## lm(formula = Balance ~ Rating + Student + Income + Age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -120.863  -32.930    0.652   32.576  105.925
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -781.26325   19.46918 -40.128  < 2e-16 ***
## Rating         4.75957    0.05346  89.035  < 2e-16 ***
## StudentYes   467.33399   12.65081  36.941  < 2e-16 ***
## Income        -9.43534    0.21108 -44.701  < 2e-16 ***
## Age           -1.20988    0.22885  -5.287 4.32e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.92 on 150 degrees of freedom
## Multiple R-squared:  0.9872, Adjusted R-squared:  0.9868
## F-statistic:  2886 on 4 and 150 DF,  p-value: < 2.2e-16
```

```
finalModel2 =lm( Balance ~ I(Rating^2)+ Student + Income + Age)
summary(finalModel2)
```

```
##
## Call:
## lm(formula = Balance ~ I(Rating^2) + Student + Income + Age)
##
```

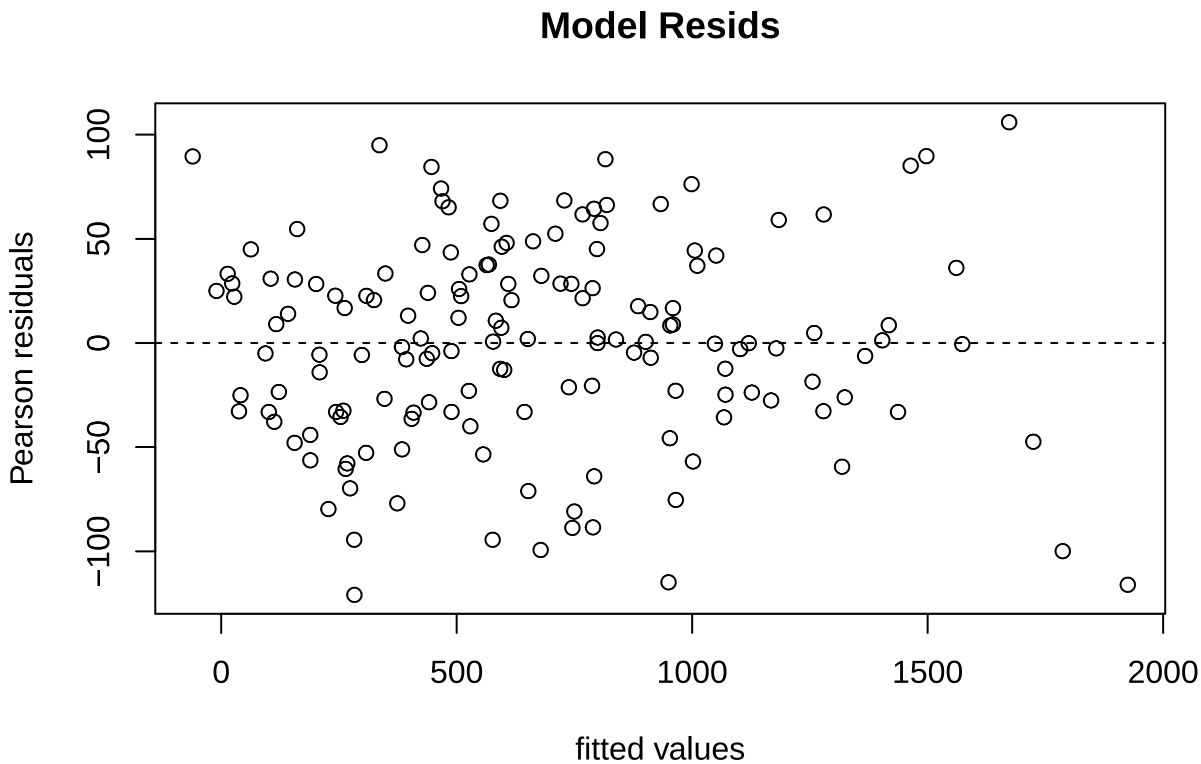
```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -830.24  -86.39   21.65  113.51  353.46
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.467e+02  4.460e+01   5.533 1.37e-07 ***
## I(Rating^2)  4.528e-03  1.872e-04  24.191  < 2e-16 ***
## StudentYes   3.360e+02  4.111e+01   8.172 1.17e-13 ***
## Income      -8.999e+00  7.405e-01 -12.153  < 2e-16 ***
## Age         -1.320e-01  7.663e-01  -0.172    0.863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 162.2 on 150 degrees of freedom
## Multiple R-squared:  0.8591, Adjusted R-squared:  0.8553
## F-statistic: 228.6 on 4 and 150 DF,  p-value: < 2.2e-16
```

```
pwrs = c(.4,.5,.6,.7,.8,.9,1.0,1.1,1.2,1.3,1.4,1.5,1.6,1.7,1.8,1.9,2.0)
#Quotient?
for (i in pwrs){
    #model = lm(Balance ~ poly(Rating, i))
    model = lm(Balance ~ I(Rating^i)+ Student + Income + Age)

    print(sprintf('%s : %s' ,i, summary(model)$sigma))
}
```

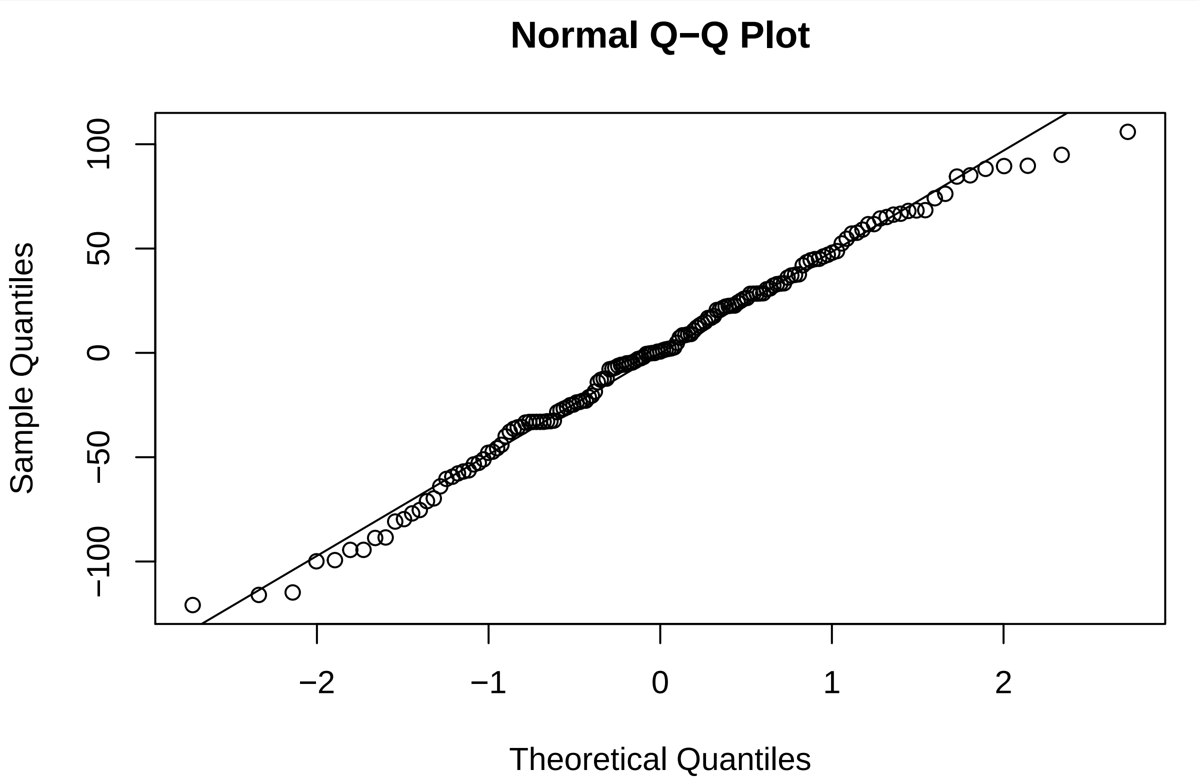
```
## [1] "0.4 : 94.7883694773213"
## [1] "0.5 : 83.2714941151627"
## [1] "0.6 : 72.246806932736"
## [1] "0.7 : 62.2665284227645"
## [1] "0.8 : 54.2218569084508"
## [1] "0.9 : 49.3710757325043"
## [1] "1 : 48.9205999616659"
## [1] "1.1 : 53.1385879325552"
## [1] "1.2 : 61.1318270532806"
## [1] "1.3 : 71.639123599256"
## [1] "1.4 : 83.6524430350291"
## [1] "1.5 : 96.5020696520658"
## [1] "1.6 : 109.755575511632"
## [1] "1.7 : 123.124180704386"
## [1] "1.8 : 136.405461104815"
## [1] "1.9 : 149.451620608373"
## [1] "2 : 162.152117421008"
```

Adding a second degree polynomial to the continuous variable Rating did not add much information at all. 1.0 also has the lowest residual sum of squares when testing all possible models with the Rating quotient ranging from 0.4-2.0. For sake of simplicity and interpretability, we will keep the degree equal to 1.0.

```
fits <- fitted(finalModel1)
## calculate the deviance residuals
dev.resids  <- resid(finalModel1)
plot(fits, dev.resids,
     xlab="fitted values", ylab="Pearson residuals" ,main = "Model Resids")
abline(h=0, lty=2)
```




```r
#QQ to see if residuals follow normal
qqnorm(finalModel1$residuals)
qqline(finalModel1$residuals)
```



The redisuals look good. They follow the normal qqline well, they experience the same variance across all fitted values (homoscedasticity), and they are centered about zero. This indicates a strong linear model that will predict credit balance accurately.

Question 4 The reducible error can be broken down into the variance of the function $f(x)$ and the squared bias of $f(x)$ as follows:
MSE $= E[(y_0 - \hat{f}(x))^2]$
$= E[(y_0 - E[\hat{f}(x)] + E[\hat{f}(x)] - \hat{f}(x))^2]$
$= E[(y_0 - E[\hat{f}(x)])^2 + 2((y_0 - E[\hat{f}(x)])(E[\hat{f}(x)] - \hat{f}(x))) + (E[\hat{f}(x)] - \hat{f}(x))^2]$
$= E[(y_0 - E[\hat{f}(x)])^2] + 2E[y_0 - E(\hat{f}(x)))(E(\hat{f}(x)) - y_0)] + E[(E(\hat{f}(x)) - \hat{f}(x)^2]$
$= E[(y_0 - E[\hat{f}(x)])^2] + 2E(y_0 - (E(\hat{f}(x)))E(\hat{f}(x) - E(\hat{f}(x))) + E[(E(\hat{f}(x)) - \hat{f}(x))^2]$
$= E[(y_0 - E[\hat{f}(x)])^2] + E[(E(\hat{f}(x)) - \hat{f}(x))^2]$
The irreducible error is the variance of the error terms for $f(x)$ and cannot be accounted for in the estimated function. Therefore the MSE is comprised of the variance of the estimated function, the squared bias, and the variance of the error terms.