Final Project

Zachary Dougherty & Adam Villarreal

November 18, 2019

USF's Class of 2018

Data Cleaning

We need to make a few adjustments to our data to allow for proper analysis. We factorize our character columns and set the GPA of observations equal to 0 to 0.001. This is because a Box Cox transformation requires all response variables to be positive and non-zero.

```
students$PELL[is.na(students$PELL)] <- "N"
students$RESIDENCE_HALL[is.na(students$RESIDENCE_HALL)] <- "Off-Campus"
students$RESIDENCE_HALL <- as.factor(students$RESIDENCE_HALL)
sum(is.na(students)) # 14
```

```
## [1] 14
```

```
students <- na.omit(students)
sum(is.na(students))
```

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

```
factor.cols <- c("GENDER", "IN_STATE", "PELL", "MAJOR", "ETHNICITY")
students$GENDER <- as.factor(students$GENDER)
students$IN_STATE <- as.factor(students$IN_STATE)
students$PELL <- as.factor(students$PELL)
students$MAJOR <- as.factor(students$MAJOR)
students$ETHNICITY <- as.factor(students$ETHNICITY)

# Changing GPA where it is equal to 0 to 0.001 instead, so preserves emptiness of value but also allows
students[students$GPA == 0, ]$GPA <- 0.001
nrow(students[students$GPA == 0, ]) # 0
```

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

```
str(students)
```

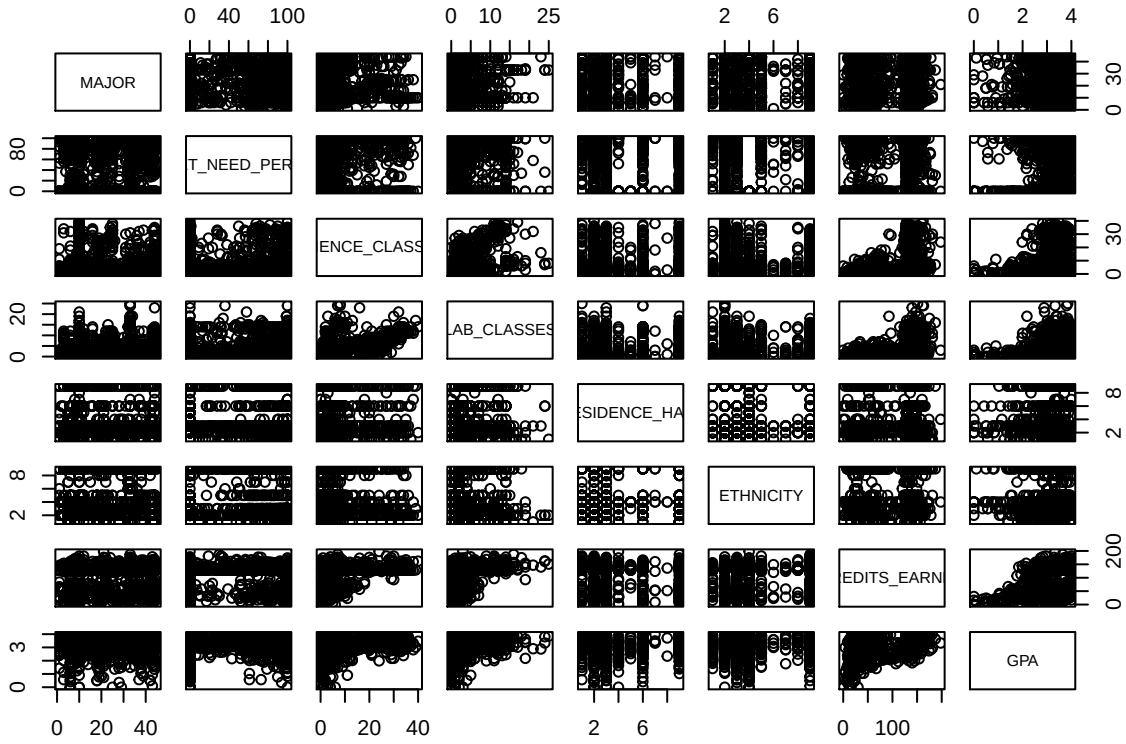
```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1495 obs. of  14 variables:
##  $ RANDOM_ID         : num  2462 31327 35938 39172 42202 ...
##  $ GENDER            : Factor w/ 2 levels "F","M": 1 1 2 2 1 1 1 2 2 1 ...
##  $ IN_STATE          : Factor w/ 2 levels "N","Y": 1 2 2 2 2 2 2 2 2 2 ...
##  $ UNMET_NEED_PERCENT: num  0 100 0 0 100 83 0 0 0 0 ...
##  $ PELL              : Factor w/ 2 levels "N","Y": 1 2 1 1 2 1 1 1 1 1 ...
##  $ MAJOR             : Factor w/ 45 levels "ADVT","ANST",..: 25 32 39 8 38 6 25 6 6 24 ...
##  $ MAJOR_DESC        : chr  "Exercise and Sport Science" "Media Studies" "Sociology" "Entrepreneursh
##  $ SCIENCE_CLASSES   : num  20 4 0 4 0 6 25 3 1 7 ...
##  $ LAB_CLASSES       : num  5 10 0 1 0 1 5 1 0 10 ...
##  $ RESIDENCE_HALL    : Factor w/ 9 levels "Fromm","Gillson",..: 3 9 2 2 3 3 3 5 3 2 ...
##  $ ETHNICITY         : Factor w/ 9 levels "African American",..: 5 3 4 4 3 2 9 4 9 2 ...
##  $ CREDITS_EARNED    : num  139 135 16 139 56 131 134 129 32 134 ...
##  $ GPA_CREDITS       : num  138 129 32 163 60 134 118 128 32 132 ...
```

```
##  $ GPA               : num  3.512 3.191 0.838 2.587 2.993 ...
##  - attr(*, "na.action")= 'omit' Named int  14 262 470 615 705 957 1497
##   ..- attr(*, "names")= chr  "14" "262" "470" "615" ...
```

Data Exploration

We will not perform some exploratory analysis and gather simple statistics on our various predictors and responses.

```
getPalette <- colorRampPalette(brewer.pal(8, "Paired"))
pairs(students[, c("MAJOR", "UNMET_NEED_PERCENT", "SCIENCE_CLASSES", "LAB_CLASSES", "RESIDENCE_HALL", "
```




```
num_male <- sum(students$GENDER == "M") #525 males
num_female <- sum(students$GENDER == "F") #970 females
num_male/nrow(students)
```

```
## [1] 0.3511706
```

```
num_female/nrow(students)
```

```
## [1] 0.6488294
```

Around 65% of the class of 2018 were females, while the other 35% were male

```
in_state <- sum(students$IN_STATE == "Y") #1116 students were from California
out_state <- nrow(students) - in_state #379 students were from outside California
in_state/nrow(students)
```

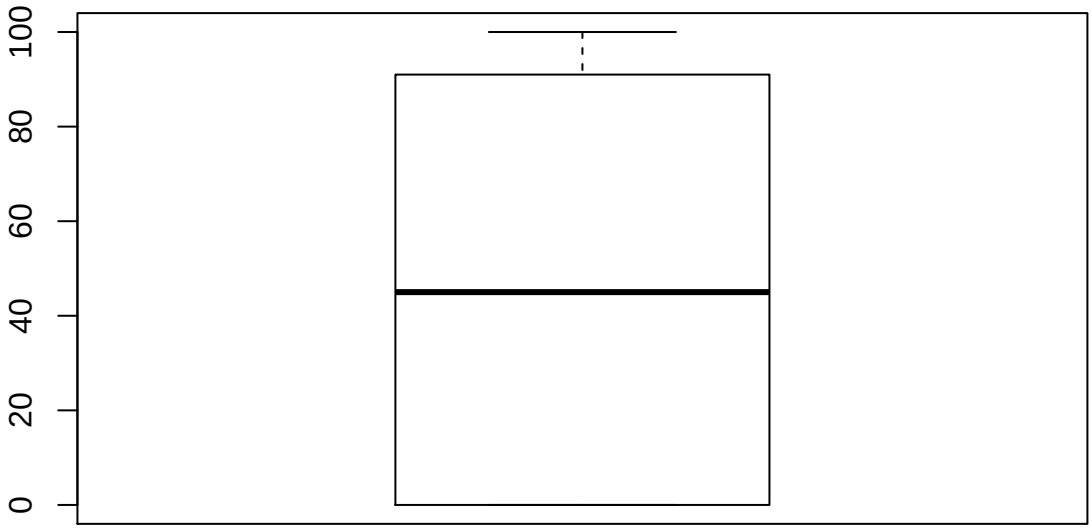
```
## [1] 0.7464883
```

```
out_state/nrow(students)
```

```
## [1] 0.2535117
```

Most students are from the state of California, which there were 1116 students from California that attended USF. There were 379 students from outside California. 75% of the student body is from California while the 25% are outside of California. The 25% can include both US students or Foreign students.

```
boxplot(students$UNMET_NEED_PERCENT)
```




```
summary(students$UNMET_NEED_PERCENT)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00   45.00   44.54   91.00  100.00
```

UNMET_NEED_PERCENT column contains continuous values that tell at what percentage was aid not provided to fill the cost of tuition. Someone that has 100 for their data point states that they had to fully pay the tuition cost at the time without aid from the university or Federal Government. If someone has 0 for their data point, that means that their tuition was fully aided by either or both by the university and Federal Government. The mean unmet need percentage was 44.54% while the median was 45%. The minimum was 0% and the maximum was 100%.

```
qualified <- sum(students$PELL == "Y") #357 students qualified for Federal Pell Grants
not_qualified <- nrow(students) - qualified #1128 students did not qualify for Federal Pell Grants
qualified/nrow(students)
```

```
## [1] 0.2454849
```

```
not_qualified/nrow(students)
```

```
## [1] 0.7545151
```

The PELL column states whether the student qualified for the Federal Pell Grant given by the United States Government. 357 students qualified for Federal Pell Grants while the other 1128 did not. Therefore, around 25% of students qualified while 75% did not.

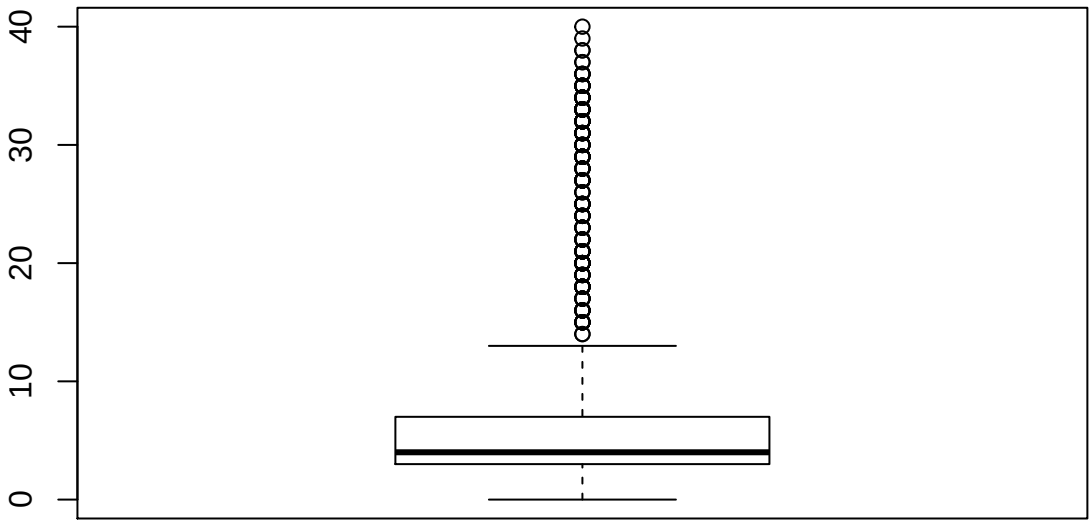
```
summary(dataset$MAJOR_DESC)
```

```
##                       Accounting                       Advertising
##                               25                                20
## Architecture & Community Design     Art History/Arts Management
##                               27                                 4
##                     Asian Studies                          Biology
##                                 4                               142
##          Business Administration                        Chemistry
```

```
##                             174                              19
##           Communication Studies      Comparative Lit. & Culture
##                              66                               4
##                Computer Science      Critical Diversity Studies
##                              48                               3
##                    Data Science                          Design
##                               5                              31
##                        Economics                         English
##                              32                              40
##   Entrepreneurship & Innovation           Environmental Science
##                              19                              10
##           Environmental Studies      Exercise and Sport Science
##                              14                              37
##                         Finance       Financial Economics (4+1)
##                              44                               2
##                       Fine Arts                         History
##                              11                              19
##          Hospitality Management          International Business
##                              22                              36
##           International Studies                Japanese Studies
##                              39                               3
##                       Marketing                     Mathematics
##                              47                              10
##                   Media Studies                         Nursing
##                              25                             120
##  Organizational Behav.& Ldrship       Perf. Arts & Soc. Justice
##                               4                              17
##                      Philosophy                         Physics
##                               7                              11
##                        Politics                       Psychology
##                              33                             103
##                       Sociology                         Spanish
##                              27                               3
##               Theology  Studies                 Undeclared Arts
##                               1                              59
##             Undeclared Business              Undeclared Sciences
##                              77                              56
##                   Urban Studies
##                               2
```

These are the Majors with the amount of students within those majors. The most popular majors are Business Administration, Biology, Nursing, and Psychology. The least popular majors are Theology with 1 student, Urban Studies with 2 students, Japanese Studies and Spanish with 3 students. The Data Science undergraduate program gained 5 students from the class of 2018!

```
boxplot(students$SCIENCE_CLASSES)
```

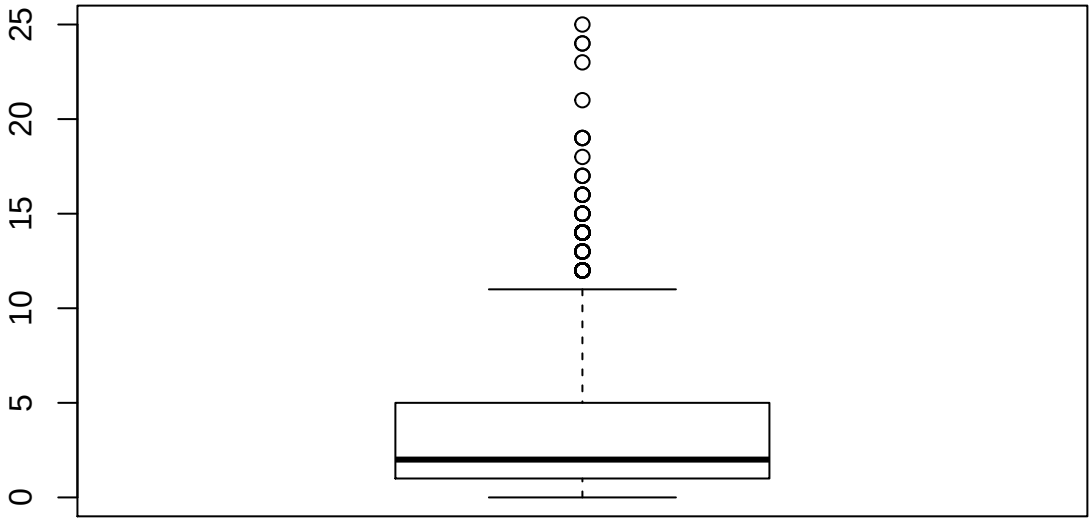



```
summary(students$SCIENCE_CLASSES)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.000   4.000   7.127   7.000  40.000
```

The average student would take at least 7 science classes as a USF student. The maximum of science classes taken per student is 40, while the minimum is 0 science classes.

```
boxplot(students$LAB_CLASSES)
```




```
summary(students$LAB_CLASSES)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   3.584   5.000  25.000
```

LAB_CLASSES column provides the number of classes with labs taken by each student. The average student took at least 3 classes with labs. The maximum amount of lab classes per student is 14 and the minimum is 0.

```
summary(students$RESIDENCE_HALL)
```

```
##         Fromm       Gillson   Hayes Healy Lone Mountain  Lone Village
##           122           431           443            46            17
##    Off-Campus  Pacific Wing  Pedro Arrupe        Phelan
```

```
##           113             5             2           316
```

```
sum(students[students$RESIDENCE_HALL == "Off-Campus", ]$IN_STATE == "Y")
```

```
## [1] 99
```

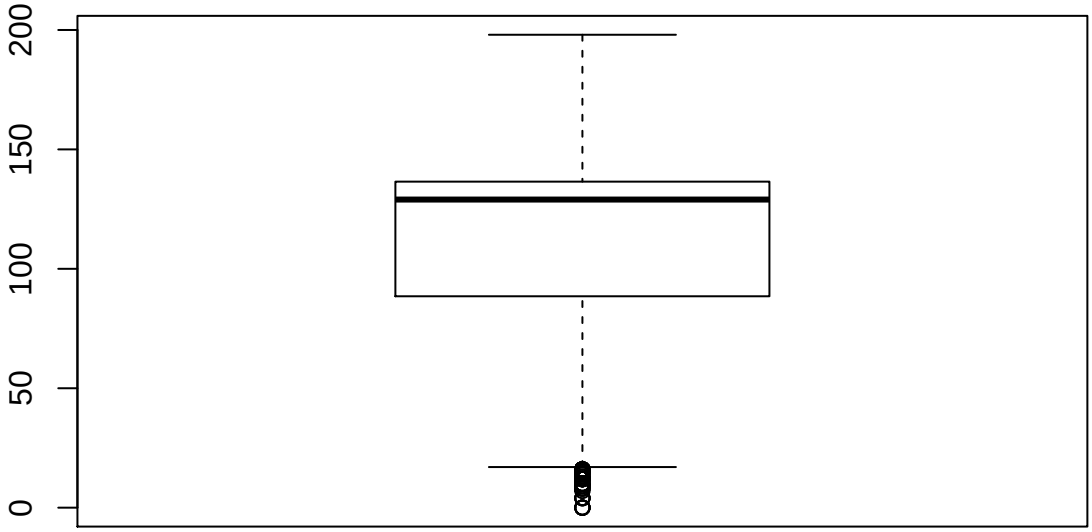
Most of 2019 graduating class lived in Hayes Healy their freshmen year. However, one important factor is that a total of 113 students lived Off-Campus. 99 of 113 students are from the state of California, which might indicate that these students are oringially from San Francisco or that they commute from neighboring cities. I am not quite sure of the 14 students who live on campus who are not originally from California. My best guess would be that they might have family connections here in the city to have a place to stay.

```
summary(students$ETHNICITY)/nrow(students)
```

```
##   African American              Asian Hispanic or Latino
##         0.034782609        0.223411371        0.208026756
##      International         Multi Race    Native American
##         0.171906355        0.058862876        0.004013378
##   Pacific Islander            Unknown              White
##         0.008026756        0.011371237        0.279598662
```

These are the students ethnicities for the class of 2018 by percentage. The majority of the class identify as white at 28%. The smallest ethnicity group are those who idenitify as Native American which is less that 1% or 0.4% exactly. What is interesting is that 17% of the student population are international students which is the 4th largest ethnicity group.

```
boxplot(students$CREDITS_EARNED)
```




```
summary(students$CREDITS_EARNED)
```

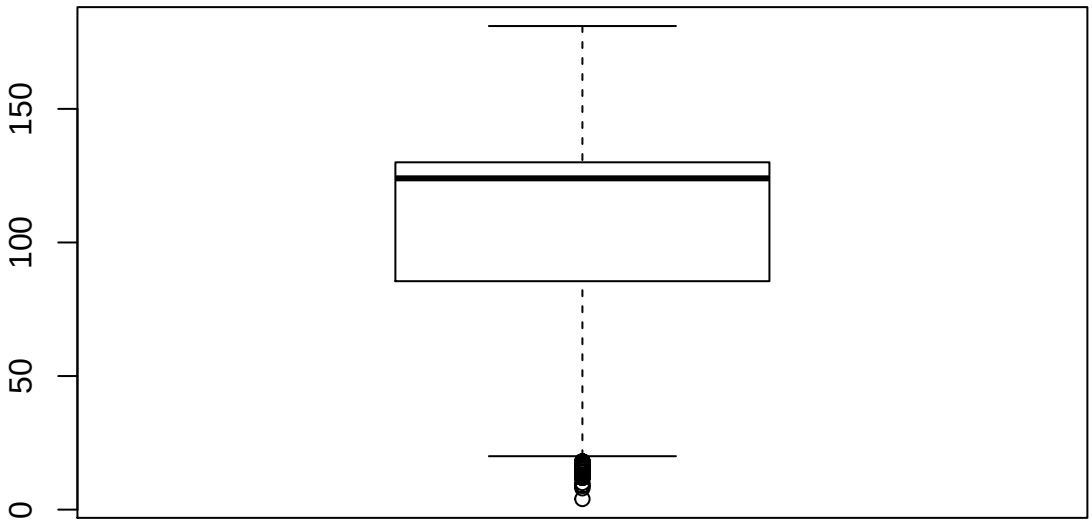
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    88.5   129.0   111.3   136.5   198.0
```

```
over.128 <- subset(students, CREDITS_EARNED >= 128)
nrow(over.128) / nrow(students)
```

```
## [1] 0.7210702
```

The average student completed at least 111 credits. The maximum amount of credits earned by a student was 198 credits, while the minimum was 0 credits. However, we calculated that only 72% achieved the graduation requirement of attaining 128 credits. Therefore, around 28% of the class of 2018 did not graduate. This can entail those students either transferred, dropped-out or did not complete their USF education in 4 years.

```r
boxplot(students$GPA_CREDITS)
```

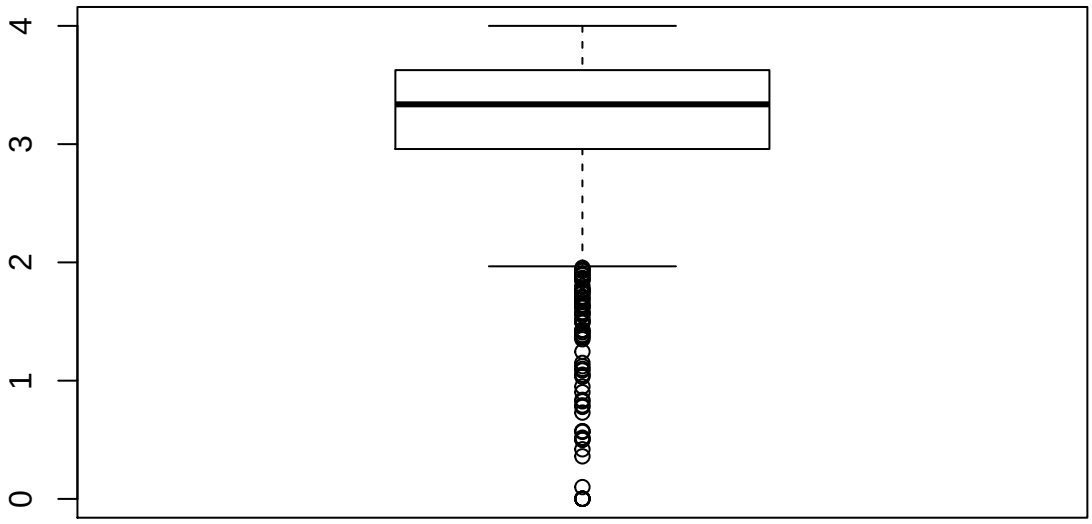



```r
summary(students$GPA_CREDITS)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     4.0    85.5   124.0   103.9   130.0   181.0
```

The average student gained 104 GPA credits during their career as a student. The max GPA credits earned by a student was 181.00. The min GPA credits earned by a student was 4.00.

```r
boxplot(students$GPA)
```




```r
summary(students$GPA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.001   2.959   3.337   3.194   3.626   4.000
```

```r
nrow(subset(students, GPA==4.00))
```

```
## [1] 10
```

```r
nrow(subset(students, GPA==0.00))
```

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

The average student GPA is a 3.19. The max GPA is 4.00 and the min GPA is 0.00. Only 10 students were able to earn a GPA of 4.00. Only 7 students were able to get a 0.00 GPA.

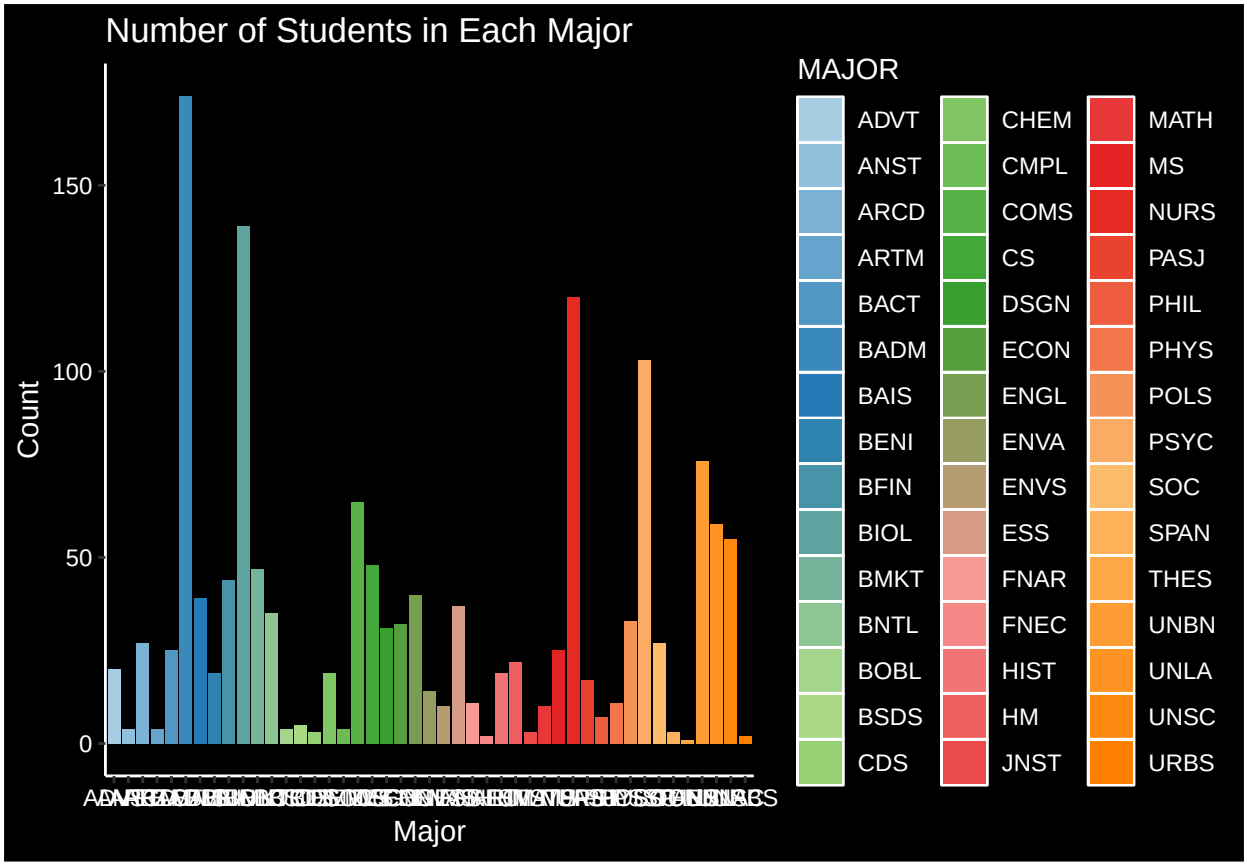
Zach's Contribution

WE CAN ADJUST THESE PLOTS BASED ON IMPORTANT PREDICTORS

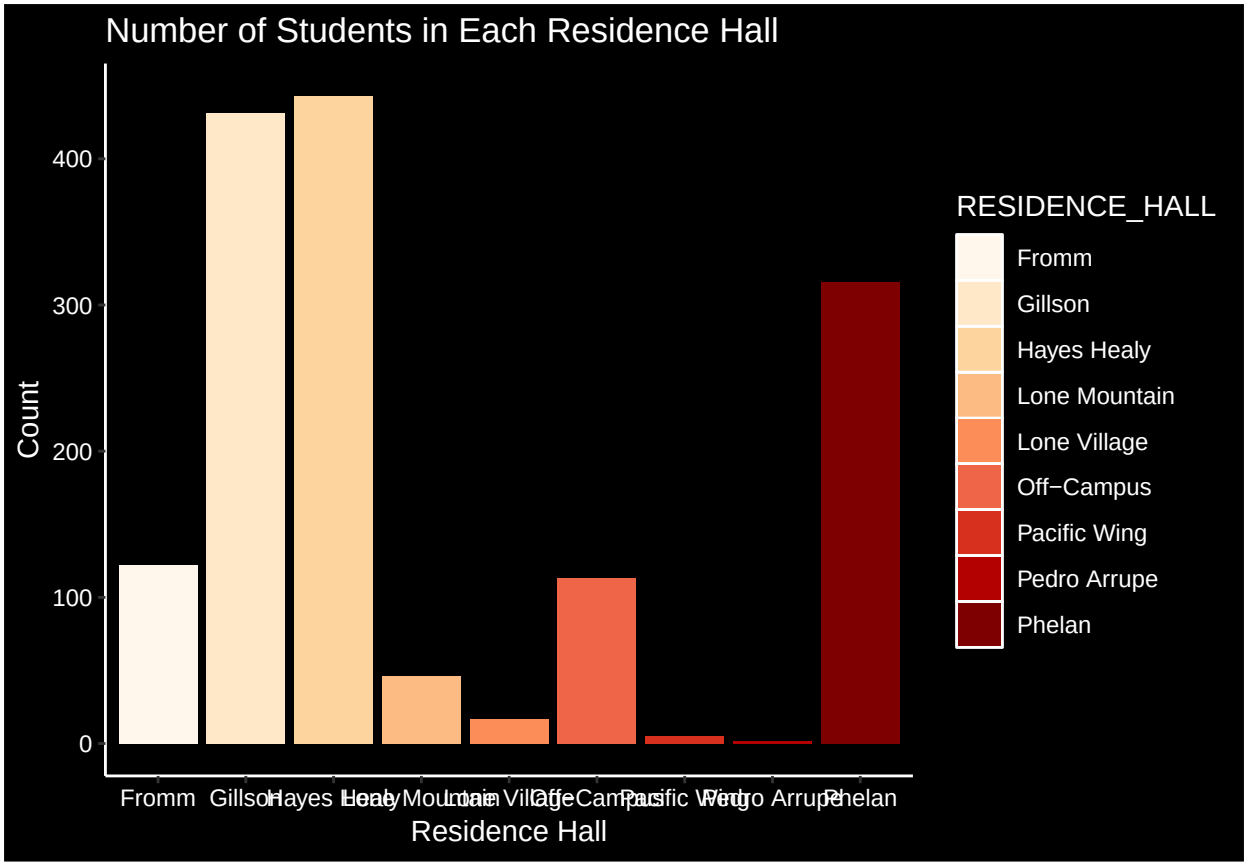
```
over.128 <- subset(students, CREDITS_EARNED >= 128)
finished.at.usf <- nrow(over.128) / nrow(students) # 0.508
avg.gpa <- mean(students$GPA) # 3.19
avg.creds.earned <- mean(students$CREDITS_EARNED) # 111.276

# How many students earned at least 128 credits
did.graduate <- students$CREDITS_EARNED >= 128

# How many of each major are there?
majors <- students %>% group_by(MAJOR) %>% tally()
majors$prop <- majors$n / nrow(students)
major.plot <- ggplot(aes(x = MAJOR, y = n, fill = MAJOR), data = majors, fill = MAJOR)
major.plot + geom_bar(stat = "identity") +
  scale_fill_manual(values = getPalette(nrow(majors))) +
  xlab("Major") +
  ylab("Count") +
  ggtitle("Number of Students in Each Major") +
  my.theme
```




```
# How many residence halls?
halls <- students %>% group_by(RESIDENCE_HALL) %>% tally()
halls$prop <- halls$n / nrow(students)
halls.plot <- ggplot(data = halls, aes(x = RESIDENCE_HALL, y = n, fill = RESIDENCE_HALL))
halls.plot + geom_bar(stat = "identity") +
  scale_fill_brewer(palette = "OrRd") +
  xlab("Residence Hall") +
  ylab("Count") +
  ggtitle("Number of Students in Each Residence Hall") +
  my.theme
```



Based on the various response variables we have, GPA, CREDITS_EARNED, and GPA_CREDITS_EARNED, we have decided to build 4 different models, 2 explanatory models for both GPA and CREDITS_EARNED and 2 predictive models for both GPA and CREDITS_EARNED. We will perform LASSO subset selection before building a simple OLS model for our explanatory purposes followed by a test of many different models to attain the most accurate model for predictive purposes.

LASSO for GPA Explanatory Model

After some data exploration, we see that we have over 40 majors, and so these different majors may have an impact on how likely a student is to get a good GPA or graduate at USF. We will attempt to first see which variables provide the most value to our model by using LASSO to select a subset of all our predictors.

We first isolate our independent variable from our predictors

```
creds <- students$CREDITS_EARNED
gpa.creds <- students$GPA_CREDITS
gpa <- students$GPA

students.vs.creds <- students[ , -c(1, 7, 13, 14)]
students.vs.gpa.cred <- students[ , -c(1, 7, 12, 14)]
students.vs.gpa <- students[, -c(1, 7, 12, 13)]
```

We will run a LASSO penalty to obtain a subset of important predictors for GPA. First, we are going to build an explanatory model for the GPA against our selected features.

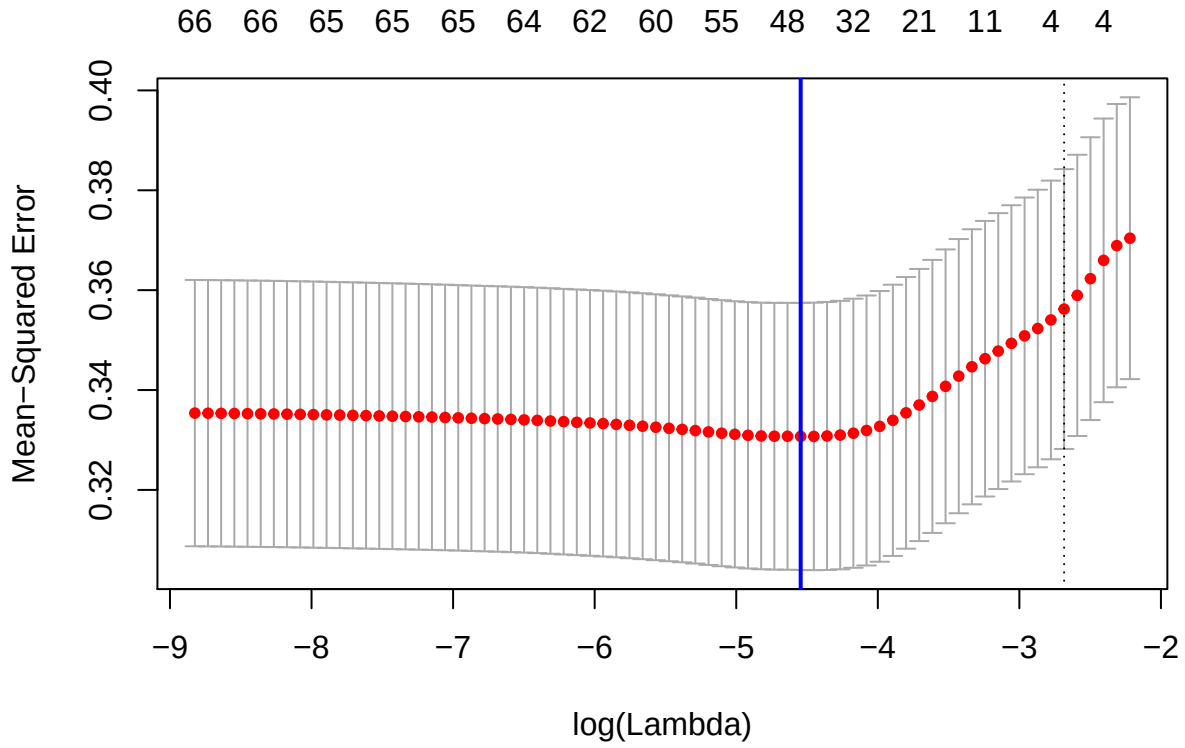
```
train.test <- students[, -c(1, 7, 12, 13, 14)]
train.test <- cbind(gpa, train.test)
```

```
train.test <- model.matrix(gpa ~ ., train.test)
train.test <- train.test[, -1]

train <- sample(1:nrow(train.test), (nrow(train.test) * 0.75 ))
test <- (-train)

y.train <- gpa[train]
y.test <- gpa[test]


lasso.selection <- cv.glmnet(train.test[train,], y.train, alpha = 1)
plot(lasso.selection)
abline(v = log(lasso.selection$lambda.min), col = "blue", lwd = 2)
```




```
coef(lasso.selection, lasso.selection$lambda.min)
```

```
## 67 x 1 sparse Matrix of class "dgCMatrix"
##                                       1
## (Intercept)                3.232978793
## GENDERM                   -0.177779142
## IN_STATEY                  .
## UNMET_NEED_PERCENT         .
## PELLY                     -0.029464325
## MAJORANST                  .
## MAJORARCD                  .
## MAJORARTM                  .
## MAJORBACT                  0.148703716
## MAJORBADM                 -0.055612543
## MAJORBAIS                  0.206641571
## MAJORBENI                  .
```

```
## MAJORBFIN                    0.130604329
## MAJORBIOL                   -0.194018521
## MAJORBMKT                    0.029853300
## MAJORBNTL                    0.080080783
## MAJORBOBL                    .
## MAJORBSDS                    0.428227524
## MAJORCDS                     0.049341861
## MAJORCHEM                   -0.147027862
## MAJORCMPL                    .
## MAJORCOMS                    0.013558148
## MAJORCS                     -0.112424542
## MAJORDSGN                    0.059394864
## MAJORECON                    .
## MAJORENGL                    0.009641324
## MAJORENVA                    .
## MAJORENVS                   -0.450937340
## MAJORESS                    -0.038420750
## MAJORFNAR                   -0.010557186
## MAJORFNEC                    .
## MAJORHIST                   -0.002153868
## MAJORHM                     -0.033608225
## MAJORJNST                    .
## MAJORMATH                    .
## MAJORMS                      0.019496869
## MAJORNURS                   -0.023884892
## MAJORPASJ                    0.013299424
## MAJORPHIL                   -0.081406107
## MAJORPHYS                   -0.120956515
## MAJORPOLS                    0.038191117
## MAJORPSYC                    .
## MAJORSOC                     .
## MAJORSPAN                    0.198251837
## MAJORTHES                    0.054865999
## MAJORUNBN                    0.155482233
## MAJORUNLA                   -0.219027801
## MAJORUNSC                   -0.233621120
## MAJORURBS                    0.254984779
## SCIENCE_CLASSES              0.007798873
## LAB_CLASSES                  0.013128429
## RESIDENCE_HALLGillson        .
## RESIDENCE_HALLHayes Healy   -0.060395556
## RESIDENCE_HALLLone Mountain -0.172315724
## RESIDENCE_HALLLone Village  -0.014595165
## RESIDENCE_HALLOff-Campus     0.011743847
## RESIDENCE_HALLPacific Wing   0.300039163
## RESIDENCE_HALLPedro Arrupe   .
## RESIDENCE_HALLPhelan         0.045480051
## ETHNICITYAsian               .
## ETHNICITYHispanic or Latino -0.103561856
## ETHNICITYInternational      -0.180603225
## ETHNICITYMulti Race          0.014635787
## ETHNICITYNative American     .
## ETHNICITYPacific Islander    0.009814291
## ETHNICITYUnknown             .
```

```
## ETHNICITYWhite                0.129124062
```

```
# Removing Lasso variables
train.test <- train.test[, -c(2, 3)]
```

Our LASSO penalty exlcluded many variables from our model, but we will not be excluding all of them. Some majors were exlcuded while others were not and it does not make sense, in the context of our study, to only include some majors since we want to see the effect of any particular major on a student's GPA. A similar argument is used to include all Residence Hallsas predictors, despite our LASSO indicating that Gillson and Lone Village students do not have a relatively large contribution to the model. However, we will exclude the predictors IN_STATE and UNMET_NEED_PERCENT according to our optimal LASSO model.

GPA OLS Explanatory Model

We will now build an explanatory Linear Regression model. Then we will perform diagnostics to understand better the data and its implications.

GPA Diagnostics

1) Model Creation, Homoskedasticity, and Normality

```
train.test.gpa <- cbind(train.test, gpa)

gpa.ols <- lm(gpa ~ ., as.data.frame(train.test.gpa[train,]))
summary(gpa.ols)
```

```
##
## Call:
## lm(formula = gpa ~ ., data = as.data.frame(train.test.gpa[train,
##     ]))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.95731 -0.24461  0.07174  0.34107  1.16727
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.083362   0.181596  16.979  < 2e-16 ***
## GENDERM                      -0.206059   0.040561  -5.080 4.46e-07 ***
## PELLY                        -0.043946   0.043595  -1.008 0.313662
## MAJORANST                     0.273386   0.587151   0.466 0.641588
## MAJORARCD                    -0.140311   0.213535  -0.657 0.511269
## MAJORARTM                    -0.023624   0.319204  -0.074 0.941018
## MAJORBACT                     0.225061   0.200475   1.123 0.261845
## MAJORBADM                    -0.086148   0.151893  -0.567 0.570721
## MAJORBAIS                     0.250651   0.175176   1.431 0.152768
## MAJORBENI                     0.065769   0.200827   0.327 0.743364
## MAJORBFIN                     0.221481   0.175588   1.261 0.207456
## MAJORBIOL                    -0.347648   0.159296  -2.182 0.029300 *
## MAJORBMKT                     0.067154   0.169712   0.396 0.692411
## MAJORBNTL                     0.125752   0.178280   0.705 0.480738
## MAJORBOBL                     0.208608   0.426358   0.489 0.624745
## MAJORBSDS                     0.678139   0.435399   1.558 0.119649
```

```
## MAJORCDS                        0.221998   0.363967   0.610 0.542034
## MAJORCHEM                      -0.332974   0.219193  -1.519 0.129039
## MAJORCMPL                       0.055919   0.318628   0.175 0.860721
## MAJORCOMS                       0.051580   0.165977   0.311 0.756042
## MAJORCS                        -0.204711   0.174029  -1.176 0.239740
## MAJORDSGN                       0.099314   0.187151   0.531 0.595765
## MAJORECON                      -0.012216   0.181661  -0.067 0.946399
## MAJORENGL                       0.039258   0.173024   0.227 0.820552
## MAJORENVA                      -0.073200   0.218812  -0.335 0.738043
## MAJORENVS                      -0.691166   0.249015  -2.776 0.005607 **
## MAJORESS                       -0.200198   0.181349  -1.104 0.269871
## MAJORFNAR                      -0.175822   0.247691  -0.710 0.477959
## MAJORFNEC                      -0.053126   0.427601  -0.124 0.901147
## MAJORHIST                      -0.100739   0.213203  -0.473 0.636666
## MAJORHM                        -0.129650   0.209675  -0.618 0.536485
## MAJORJNST                       0.257383   0.585726   0.439 0.660443
## MAJORMATH                      -0.110428   0.260431  -0.424 0.671637
## MAJORMS                         0.066888   0.189710   0.353 0.724474
## MAJORNURS                      -0.185110   0.170088  -1.088 0.276702
## MAJORPASJ                       0.067821   0.214910   0.316 0.752386
## MAJORPHIL                      -0.221124   0.292516  -0.756 0.449855
## MAJORPHYS                      -0.258749   0.231123  -1.120 0.263170
## MAJORPOLS                       0.073715   0.186456   0.395 0.692666
## MAJORPSYC                       0.011335   0.158198   0.072 0.942895
## MAJORSOC                       -0.029138   0.193550  -0.151 0.880362
## MAJORSPAN                       0.392247   0.358431   1.094 0.274054
## MAJORTHES                       0.311805   0.589961   0.529 0.597251
## MAJORUNBN                       0.205891   0.163796   1.257 0.209033
## MAJORUNLA                      -0.303432   0.172691  -1.757 0.079194 .
## MAJORUNSC                      -0.370493   0.168564  -2.198 0.028170 *
## MAJORURBS                       0.465645   0.426222   1.092 0.274865
## SCIENCE_CLASSES                 0.011233   0.003378   3.325 0.000914 ***
## LAB_CLASSES                     0.021100   0.007224   2.921 0.003563 **
## RESIDENCE_HALLGillson           0.032046   0.068879   0.465 0.641851
## `RESIDENCE_HALLHayes Healy`    -0.050703   0.068542  -0.740 0.459624
## `RESIDENCE_HALLLone Mountain` -0.176772   0.111174  -1.590 0.112123
## `RESIDENCE_HALLLone Village`  -0.095423   0.187201  -0.510 0.610344
## `RESIDENCE_HALLOff-Campus`     0.079194   0.087444   0.906 0.365327
## `RESIDENCE_HALLPacific Wing`   0.528405   0.267869   1.973 0.048799 *
## `RESIDENCE_HALLPedro Arrupe`   0.364001   0.412005   0.883 0.377175
## RESIDENCE_HALLPhelan            0.094962   0.070323   1.350 0.177185
## ETHNICITYAsian                  0.138694   0.100117   1.385 0.166248
## `ETHNICITYHispanic or Latino`  0.002581   0.098616   0.026 0.979128
## ETHNICITYInternational         -0.098101   0.107066  -0.916 0.359736
## `ETHNICITYMulti Race`           0.177660   0.115624   1.537 0.124707
## `ETHNICITYNative American`      0.199828   0.307593   0.650 0.516058
## `ETHNICITYPacific Islander`     0.222839   0.207507   1.074 0.283120
## ETHNICITYUnknown                0.077042   0.186777   0.412 0.680071
## ETHNICITYWhite                  0.273279   0.098355   2.779 0.005558 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5657 on 1056 degrees of freedom
## Multiple R-squared:  0.1851, Adjusted R-squared:  0.1357
```

```
## F-statistic: 3.748 on 64 and 1056 DF,  p-value: < 2.2e-16
```

```
bptest(gpa.ols)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  gpa.ols
## BP = 99.088, df = 64, p-value = 0.00323
```

```
shapiro.test(residuals(gpa.ols))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(gpa.ols)
## W = 0.89569, p-value < 2.2e-16
```

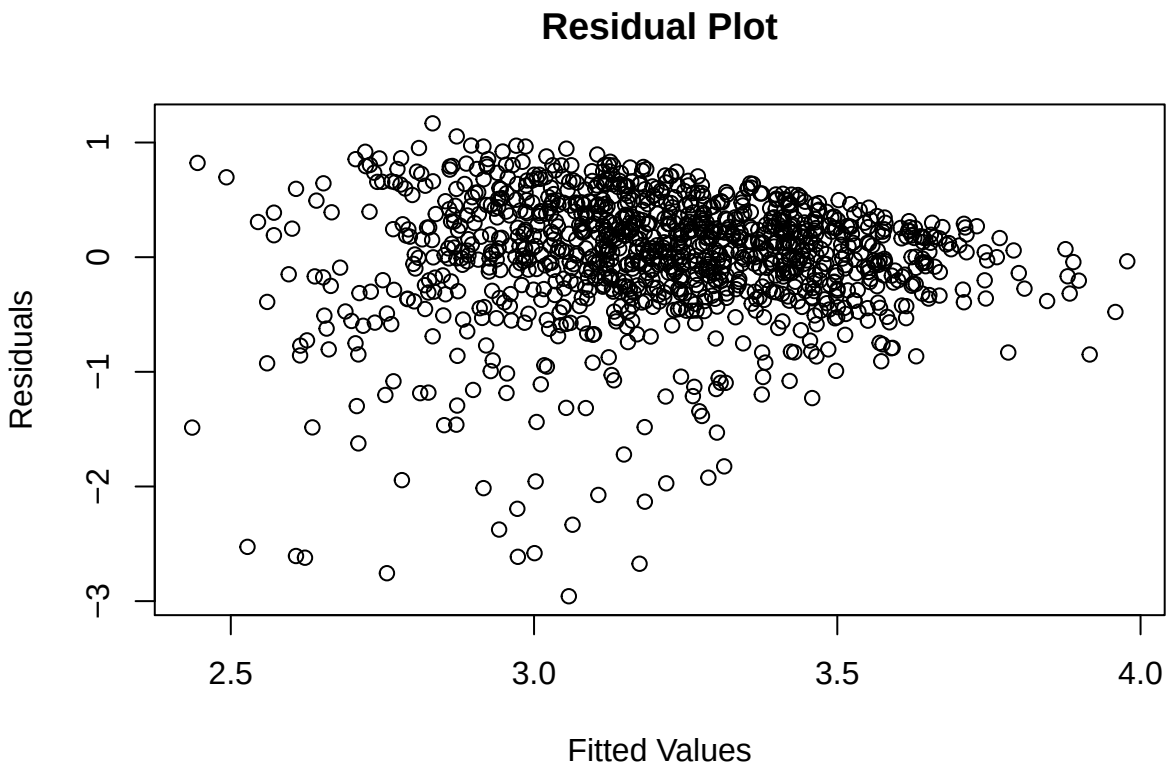
A few initial diagnostics show that we apparently have some heteroskedasticity and normality errors. Our initial model fails both the Shapiro Wilkes and Breusch Pagan tests and has an adjusted R^2 value of only 0.14. We will now conduct more tests and perform transforms in an attempt to correct these issues.

2) Extreme Values

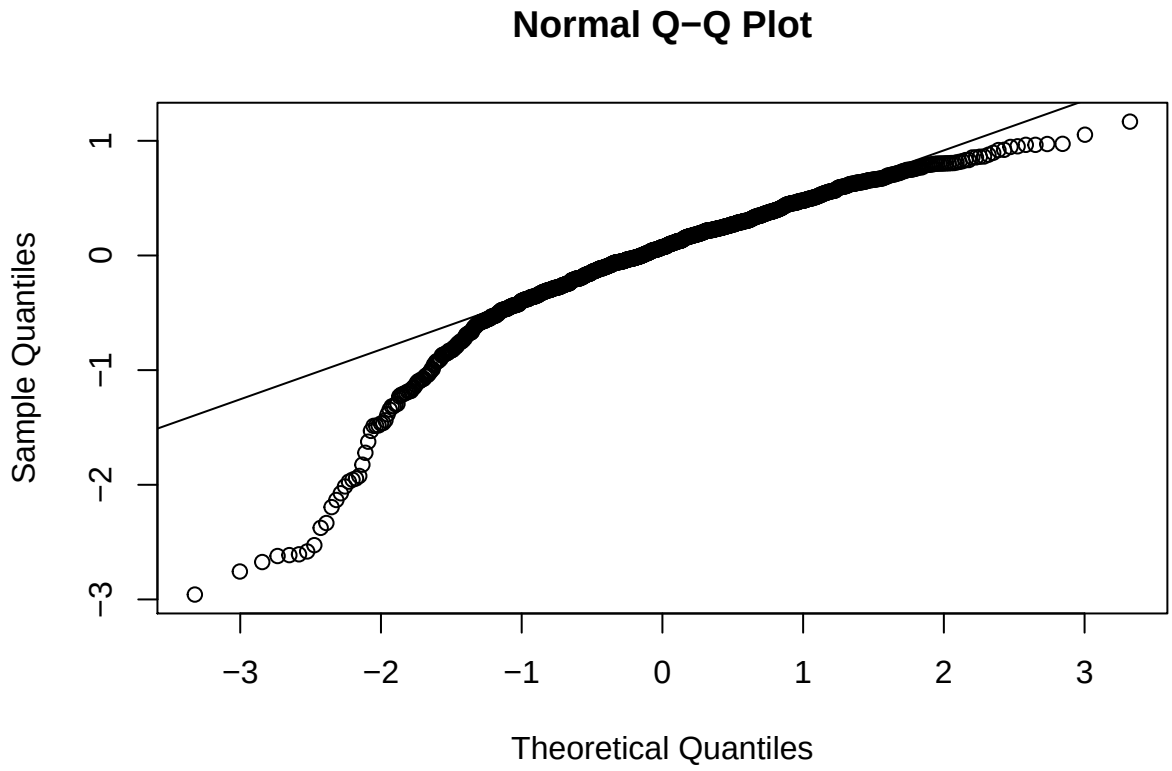
I have created a function which gives many important diagnistic plots for quick analysis.

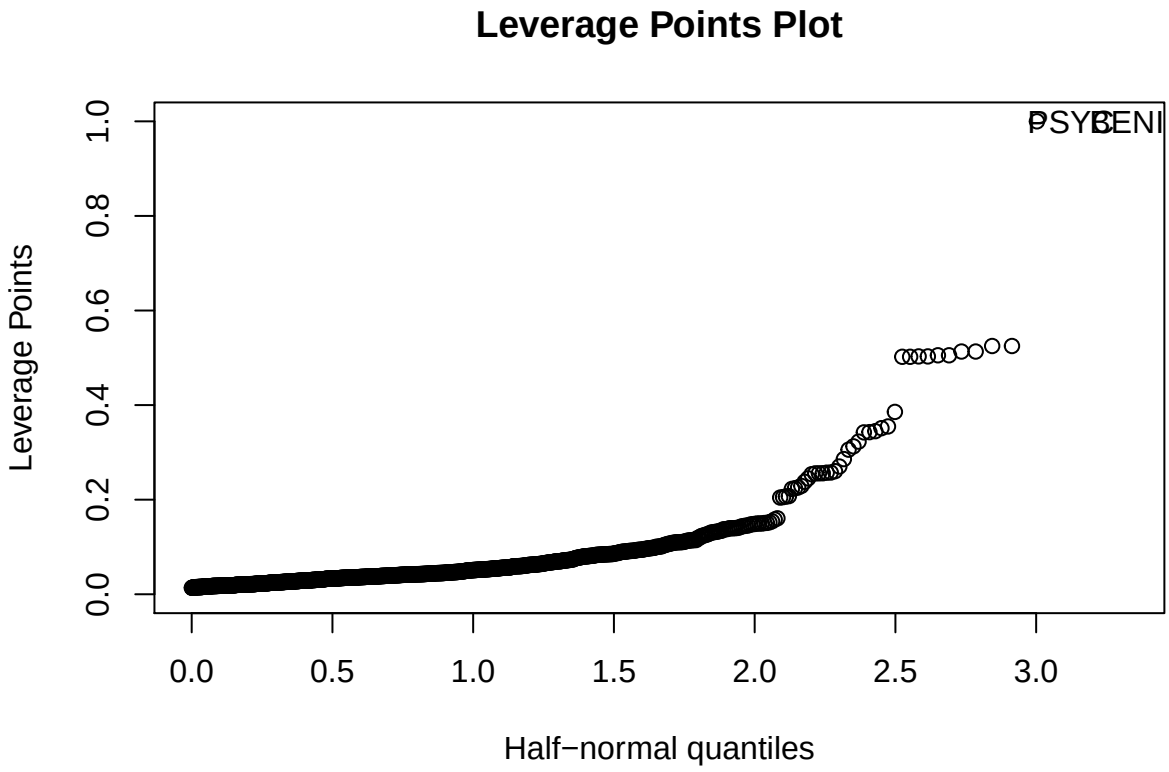
```
plot.extreme <- function(x, labs=NA) {
  hats <- hatvalues(x)
  plot(fitted(x), residuals(x), xlab = "Fitted Values", ylab = "Residuals", main = "Residual Plot")
  qqnorm(residuals(x))
  qqline(residuals(x))
  if (!is.na(labs)) {
    halfnorm(hats, labs = labs, ylab = "Leverage Points", main = "Leverage Points Plot")
  } else {
    halfnorm(hats, labs = rownames(x), ylab = "Leverage Points", main = "Leverage Points Plot")
  }
  std.resids <- rstudent(x)
  plot(fitted(x), std.resids, xlab = "Fitted Values", ylab = "Studentized Residuals", main = "Outliers"
  abline(h = 3, col = "purple", lwd = 2)
  abline(h = -3, col = "purple", lwd = 2)
  if (!is.na(labs)) {
    halfnorm(cooks.distance(gpa.ols), labs = labs, ylab = "Cook's Distance", main = "Influential Points
  } else {
    halfnorm(cooks.distance(gpa.ols), labs = rownames(x), ylab = "Cook's Distance", main = "Influential
  }
}
```

```
plot.extreme(gpa.ols, labs = students$MAJOR)
```

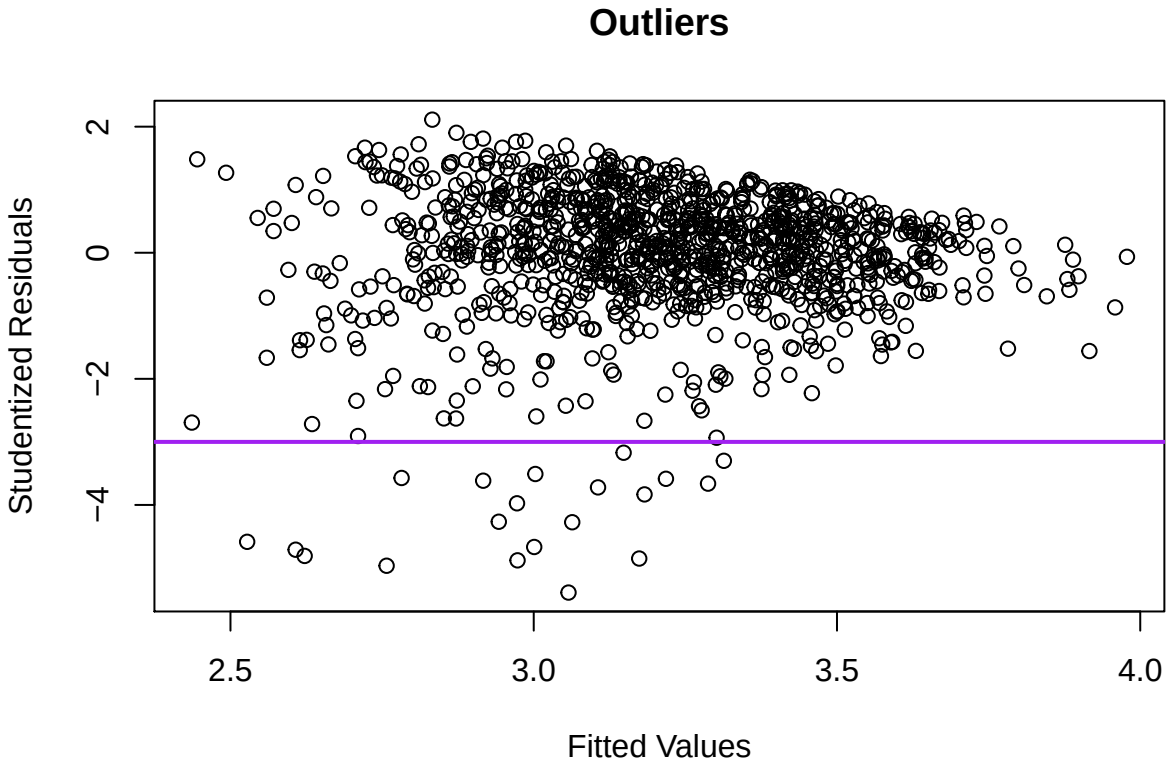


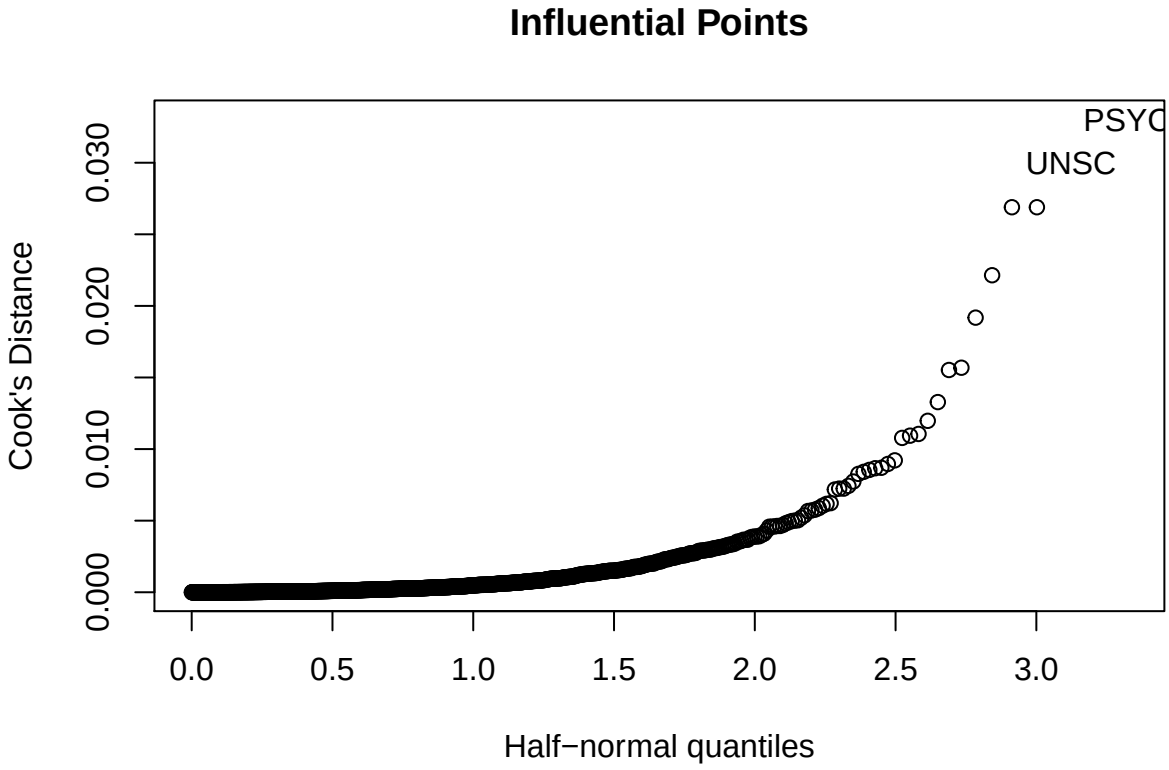

```
## Warning in if (!is.na(labs)) {: the condition has length > 1 and only the
## first element will be used
```






```
## Warning in if (!is.na(labs)) {: the condition has length > 1 and only the
## first element will be used
```

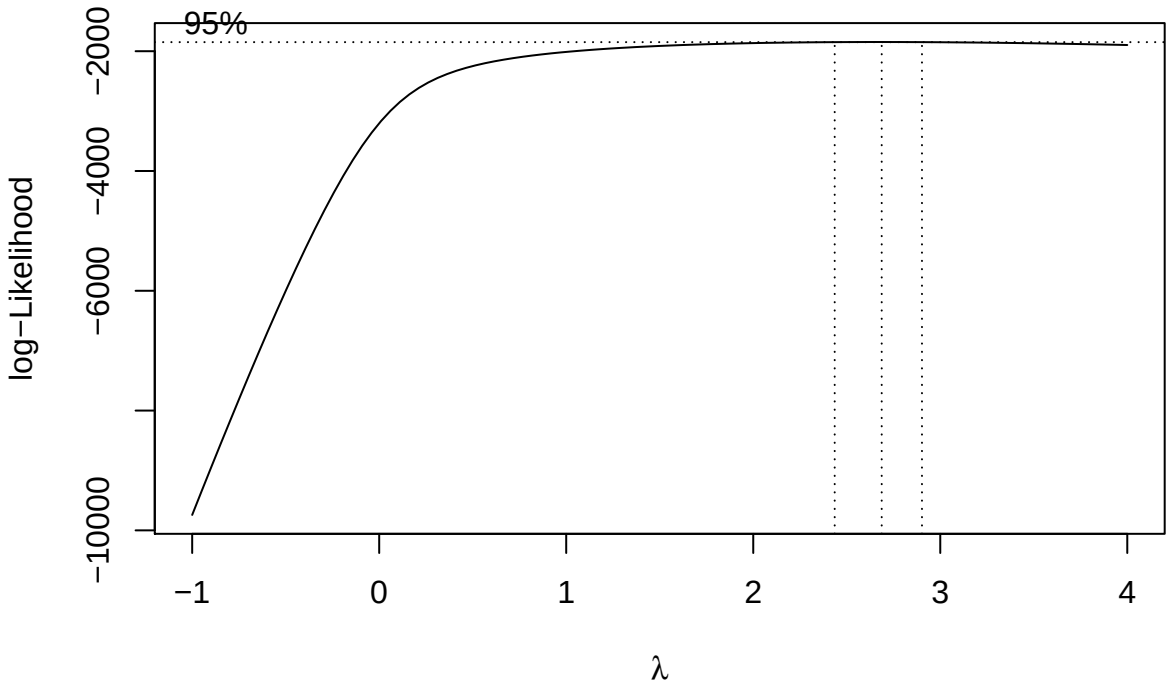




From a plot of the *h-values*, we can see that there are many observations which have high leverage, including an UNSC (undeclared sciences) and the only THES (theology studies) major. A plot of the studentized residuals also gives the indication that we have many extreme values, as we see many observations with studentized residuals greater than 3 standard deviations. We find a trend in our Cook's Distance plot, with many influential points. Before removing any observations, we will examine the effect of a Box-Cox transform to see if we can correct some issues in our data, and then re-examine our extreme values.

3) Box-Cox Transformation

```
boxcox(gpa.ols,
       lambda = seq(-1, 4, 1/10))
```




```
bx.1.gpa <- train.test.gpa
bx.1.gpa[, "gpa"] <- (bx.1.gpa[, "gpa"] ** 3) # transformation on GPA
head(bx.1.gpa[, "gpa"])
```

```
##          1          2          3          4          5          6
## 43.3293127 32.4830625  0.5874277 17.3160164 26.8203997 37.0453295
```

```
#head(train.test.gpa[, "gpa"])

# After y ^ 3 transform...
bx.1.gpa.ols <- lm(gpa ~ ., as.data.frame(bx.1.gpa[train, ]))

shapiro.test(residuals(bx.1.gpa.ols))
```

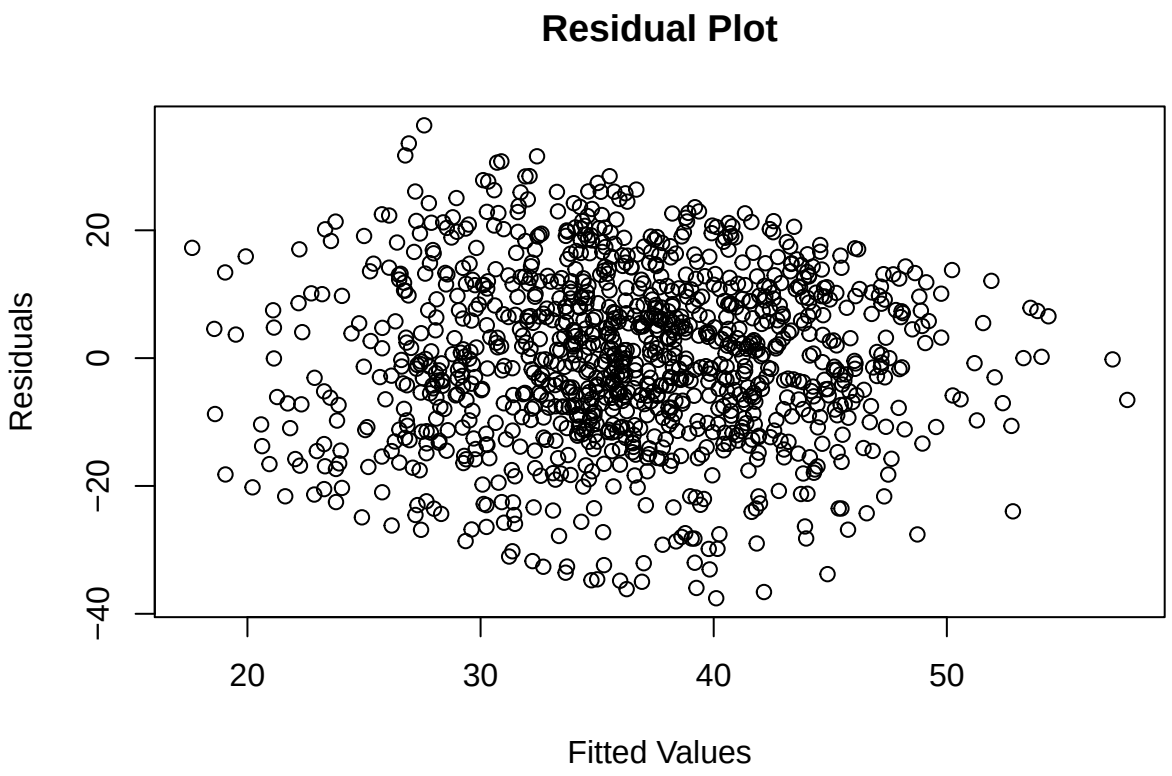
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(bx.1.gpa.ols)
## W = 0.9949, p-value = 0.0007736
```

```
bptest(bx.1.gpa.ols)
```

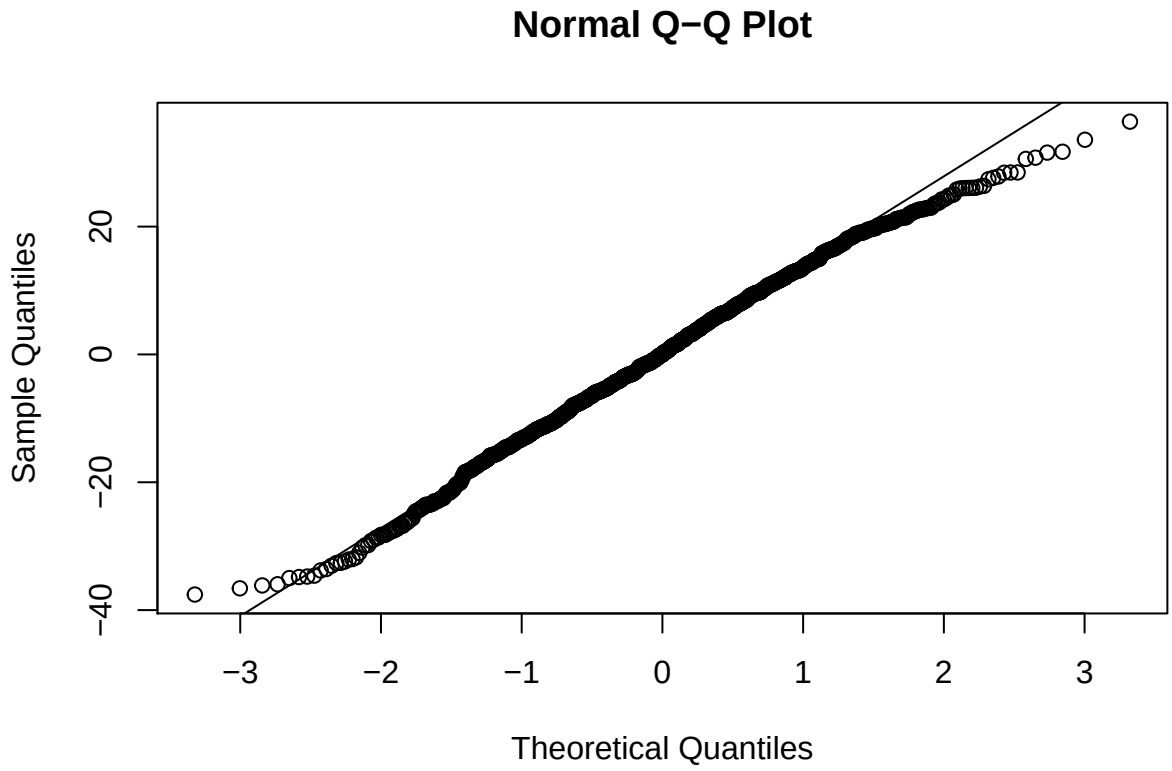
```
##
##  studentized Breusch-Pagan test
##
## data:  bx.1.gpa.ols
## BP = 108.66, df = 64, p-value = 0.000419
```

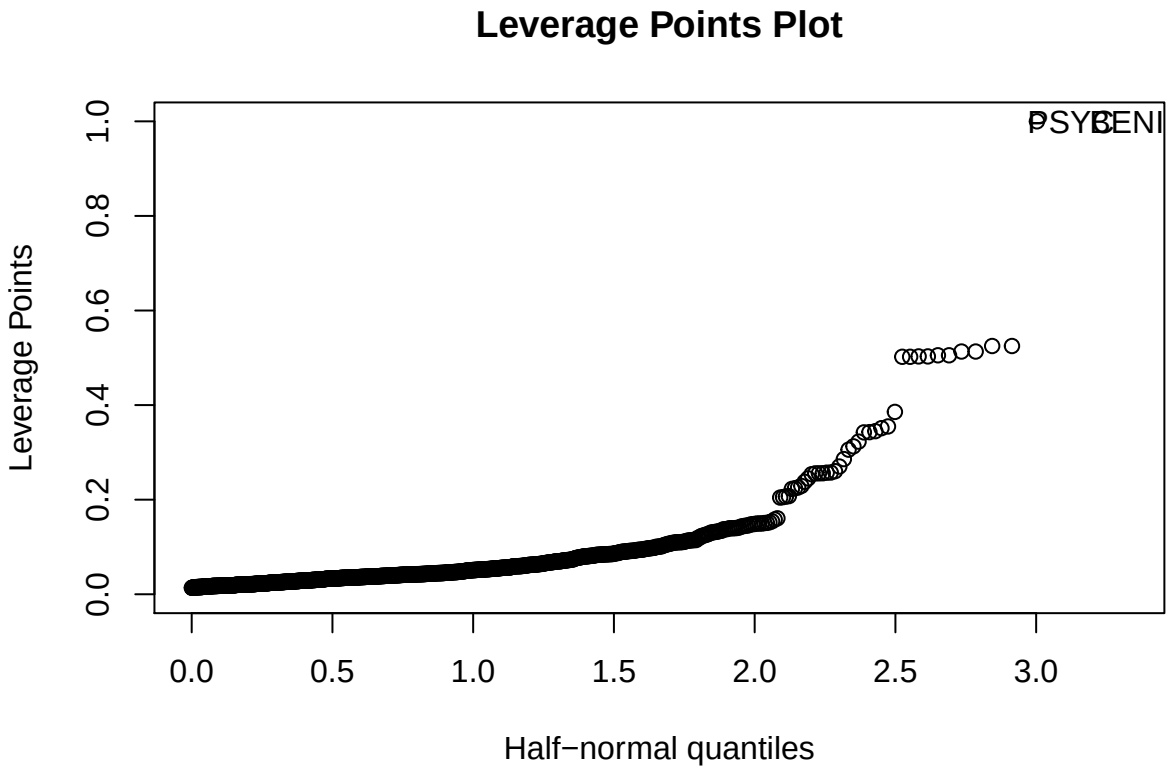
The Box-Cox suggests that the best integer transformation on Y involves a λ of 3. Replotting the residuals and QQ plots shows much better adherence to normality and homoskedasticity, though we do still have a somewhat predictable structure in the residual plot. However, we still find that our Shapiro Wilkes and Breusch Pagan tests fail. Let's re-examine our extreme values.

```
plot.extreme(bx.1.gpa.ols, students$MAJOR)
```

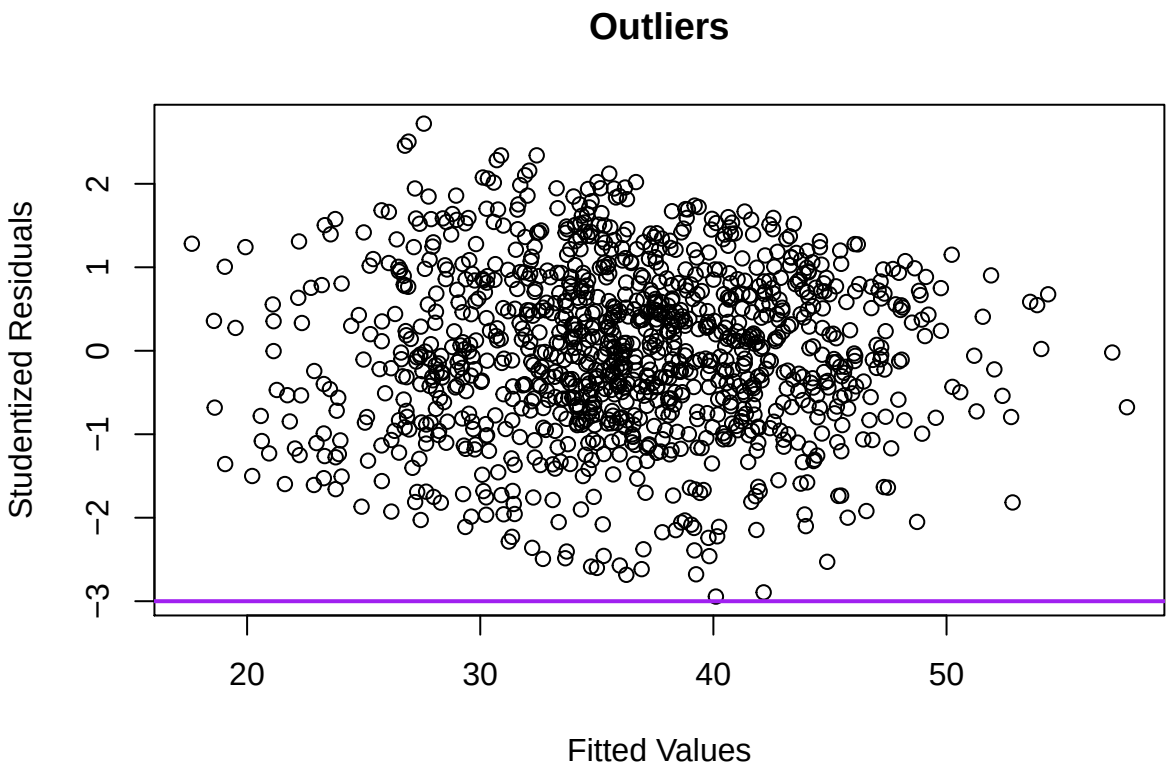


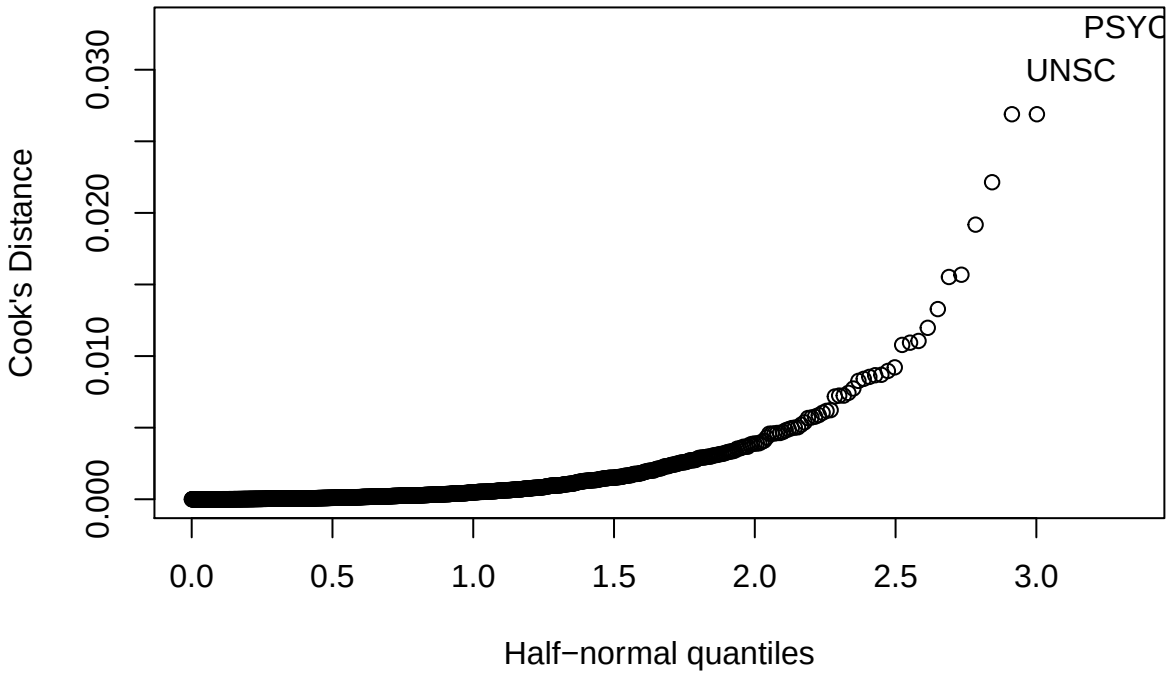

```
## Warning in if (!is.na(labs)) {: the condition has length > 1 and only the
## first element will be used
```






```
## Warning in if (!is.na(labs)) {: the condition has length > 1 and only the
## first element will be used
```





We are still left with many Leverage points and influential points which could be hurting the reliability of our model, but our studentized residual plot shows that we did fit nearly all of our previous outliers within 3 standard deviations. Let us remove all observations with a Cook's distance greater than 0.010 and re-examine our plots.

4) Removing Influential Points

We now need to find all influential observations from our data. We need to remove the observations and store new dataset, then create new train and test splits. However, we will examine our influential points to see if there are any noticeable trends or patterns in our extreme data.

```
influential <- as.numeric(names(na.omit(cooks.distance(bx.1.gpa.ols)))[(na.omit(cooks.distance(bx.1.gpa

length(cooks.distance(bx.1.gpa.ols)[cooks.distance(bx.1.gpa.ols) > 4/(nrow(students))]) # 103
```

```
## [1] 104
```

```
print("Proportion of Majors in Data with Influential Points")
```

```
## [1] "Proportion of Majors in Data with Influential Points"
```

```
summary(students[influential, ]$MAJOR)/nrow(students[influential,])
```

```
##       ADVT       ANST       ARCD       ARTM       BACT       BADM
## 0.04950495 0.00000000 0.01980198 0.02970297 0.01980198 0.00990099
##       BAIS       BENI       BFIN       BIOL       BMKT       BNTL
## 0.00000000 0.03960396 0.01980198 0.05940594 0.00990099 0.03960396
##       BOBL       BSDS        CDS       CHEM       CMPL       COMS
## 0.00000000 0.00000000 0.02970297 0.00990099 0.02970297 0.01980198
##         CS       DSGN       ECON       ENGL       ENVA       ENVS
## 0.01980198 0.00990099 0.04950495 0.05940594 0.01980198 0.00990099
```

```
##         ESS        FNAR        FNEC        HIST          HM        JNST
## 0.00990099 0.05940594 0.01980198 0.02970297 0.03960396 0.00000000
##        MATH          MS        NURS        PASJ        PHIL        PHYS
## 0.01980198 0.01980198 0.00990099 0.02970297 0.01980198 0.01980198
##        POLS        PSYC         SOC        SPAN        THES        UNBN
## 0.02970297 0.00000000 0.03960396 0.02970297 0.00000000 0.00000000
##        UNLA        UNSC        URBS
## 0.03960396 0.00990099 0.01980198
```

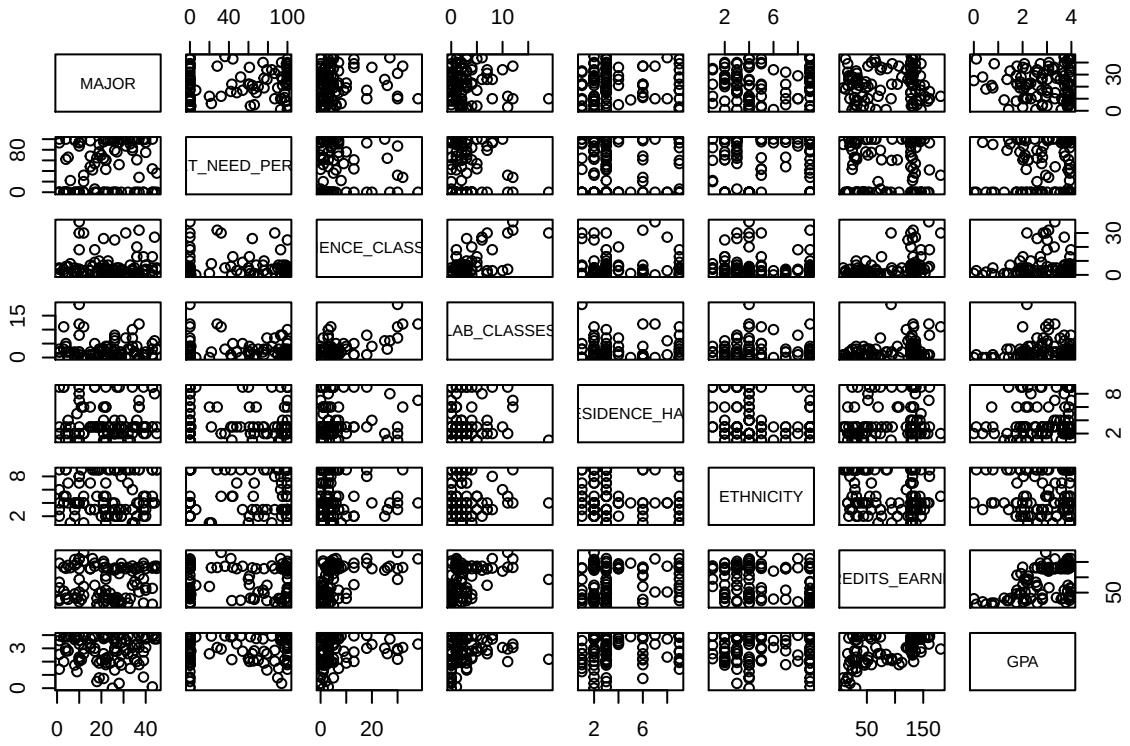
```
print("Proportion of Majors in Data without Influential Points")
```

```
## [1] "Proportion of Majors in Data without Influential Points"
```

```
summary(students$MAJOR)/nrow(students)
```

```
##          ADVT          ANST          ARCD          ARTM          BACT
## 0.0133779264 0.0026755853 0.0180602007 0.0026755853 0.0167224080
##          BADM          BAIS          BENI          BFIN          BIOL
## 0.1163879599 0.0260869565 0.0127090301 0.0294314381 0.0929765886
##          BMKT          BNTL          BOBL          BSDS           CDS
## 0.0314381271 0.0234113712 0.0026755853 0.0033444816 0.0020066890
##          CHEM          CMPL          COMS            CS          DSGN
## 0.0127090301 0.0026755853 0.0434782609 0.0321070234 0.0207357860
##          ECON          ENGL          ENVA          ENVS           ESS
## 0.0214046823 0.0267558528 0.0093645485 0.0066889632 0.0247491639
##          FNAR          FNEC          HIST            HM          JNST
## 0.0073578595 0.0013377926 0.0127090301 0.0147157191 0.0020066890
##          MATH            MS          NURS          PASJ          PHIL
## 0.0066889632 0.0167224080 0.0802675585 0.0113712375 0.0046822742
##          PHYS          POLS          PSYC           SOC          SPAN
## 0.0073578595 0.0220735786 0.0688963211 0.0180602007 0.0020066890
##          THES          UNBN          UNLA          UNSC          URBS
## 0.0006688963 0.0508361204 0.0394648829 0.0367892977 0.0013377926
```

```
pairs(students[influential, c("MAJOR", "UNMET_NEED_PERCENT", "SCIENCE_CLASSES", "LAB_CLASSES", "RESIDEN
```



We can see some trends in our influential pairs plot, such as a positive linear trend between LAB_CLASSES and GPA and SCIENCE_CLASSES and GPA. The most noticeable difference here is that we do not see the thick bands which represent students who get 128 credits. There is no large difference between the summary statistics for the predictors of the entire dataset and those of the influential points, so we will proceed with removing them.

```
bx.1.gpa.screened <- bx.1.gpa[-(influential), ]

train.screened <- sample(1:nrow(bx.1.gpa.screened), (nrow(bx.1.gpa.screened) * 0.75 ))
test.screened <- (-train.screened)

y.train.screened <- gpa[train.screened]
y.test.screened <- gpa[test.screened]

bx.screened.ols <- lm(gpa ~ ., as.data.frame(bx.1.gpa.screened[train.screened, ]))
summary(bx.screened.ols)
```

```
##
## Call:
## lm(formula = gpa ~ ., data = as.data.frame(bx.1.gpa.screened[train.screened,
##     ]))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.858  -8.718   0.345   9.490  38.042
##
## Coefficients: (5 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    31.19980    4.49632   6.939 7.15e-12 ***
## GENDERM                        -5.51971    0.96464  -5.722 1.40e-08 ***
## PELLY                          -2.72812    1.08879  -2.506 0.012383 *
## MAJORANST                       9.67528    8.44992   1.145 0.252481
```

```
## MAJORARCD                       -5.03751    4.85671  -1.037 0.299884
## MAJORARTM                        7.93499   13.60425   0.583 0.559843
## MAJORBACT                        3.11010    4.66524   0.667 0.505150
## MAJORBADM                        0.10323    3.72853   0.028 0.977918
## MAJORBAIS                        4.99469    4.44861   1.123 0.261816
## MAJORBENI                        1.00493    5.27526   0.190 0.848957
## MAJORBFIN                        3.21177    4.36871   0.735 0.462408
## MAJORBIOL                       -4.78675    3.93727  -1.216 0.224370
## MAJORBMKT                        1.58415    4.15182   0.382 0.702873
## MAJORBNTL                        7.45332    4.65644   1.601 0.109776
## MAJORBOBL                        9.53658    8.38203   1.138 0.255505
## MAJORBSDS                       10.07255    8.50298   1.185 0.236465
## MAJORCDS                            NA         NA      NA       NA
## MAJORCHEM                      -11.15766    5.04454  -2.212 0.027207 *
## MAJORCMPL                        4.36769   13.64833   0.320 0.749024
## MAJORCOMS                        2.41258    4.04155   0.597 0.550682
## MAJORCS                         -1.53240    4.23558  -0.362 0.717584
## MAJORDSGN                        2.04358    4.49296   0.455 0.649324
## MAJORECON                       -2.09446    4.67300  -0.448 0.654104
## MAJORENGL                        9.94626    4.38982   2.266 0.023683 *
## MAJORENVA                        2.70017    5.87131   0.460 0.645694
## MAJORENVS                      -14.28371    6.19746  -2.305 0.021387 *
## MAJORESS                        -1.17480    4.39476  -0.267 0.789280
## MAJORFNAR                       -4.16860    6.91842  -0.603 0.546956
## MAJORFNEC                           NA         NA      NA       NA
## MAJORHIST                        5.08683    5.10632   0.996 0.319407
## MAJORHM                         -6.02029    4.95074  -1.216 0.224262
## MAJORJNST                        6.21369    9.96862   0.623 0.533215
## MAJORMATH                        3.10993    6.90835   0.450 0.652688
## MAJORMS                          3.03899    4.85681   0.626 0.531646
## MAJORNURS                       -0.69496    4.17227  -0.167 0.867745
## MAJORPASJ                        0.97798    5.53186   0.177 0.859710
## MAJORPHIL                      -10.94913    6.85683  -1.597 0.110626
## MAJORPHYS                       -7.72572    5.80365  -1.331 0.183437
## MAJORPOLS                        4.39019    4.47282   0.982 0.326574
## MAJORPSYC                        3.93390    3.84821   1.022 0.306905
## MAJORSOC                         5.36545    4.88902   1.097 0.272713
## MAJORSPAN                           NA         NA      NA       NA
## MAJORTHES                       16.78588   13.74972   1.221 0.222448
## MAJORUNBN                        1.64832    3.95140   0.417 0.676660
## MAJORUNLA                        1.94005    4.19583   0.462 0.643915
## MAJORUNSC                       -5.38837    4.15766  -1.296 0.195276
## MAJORURBS                           NA         NA      NA       NA
## SCIENCE_CLASSES                  0.23013    0.08322   2.765 0.005793 **
## LAB_CLASSES                      0.41596    0.17602   2.363 0.018313 *
## RESIDENCE_HALLGillson           -3.31501    1.69905  -1.951 0.051328 .
## `RESIDENCE_HALLHayes Healy`     -4.58890    1.69985  -2.700 0.007061 **
## `RESIDENCE_HALLLone Mountain` -13.12404    3.00477  -4.368 1.39e-05 ***
## `RESIDENCE_HALLLone Village`    -6.08824    4.55826  -1.336 0.181973
## `RESIDENCE_HALLOff-Campus`      -0.88490    2.18480  -0.405 0.685546
## `RESIDENCE_HALLPacific Wing`     8.79255    8.01668   1.097 0.273005
## `RESIDENCE_HALLPedro Arrupe`        NA         NA      NA       NA
## RESIDENCE_HALLPhelan            -2.79866    1.73639  -1.612 0.107331
## ETHNICITYAsian                   8.21834    2.44571   3.360 0.000808 ***
```

```
## `ETHNICITYHispanic or Latino`   5.22416    2.38718   2.188 0.028873 *
## ETHNICITYInternational          2.66024    2.64032   1.008 0.313921
## `ETHNICITYMulti Race`           8.05744    2.87235   2.805 0.005128 **
## `ETHNICITYNative American`     12.14096    9.88379   1.228 0.219601
## `ETHNICITYPacific Islander`    13.69567    5.88860   2.326 0.020232 *
## ETHNICITYUnknown               12.08166    4.49660   2.687 0.007335 **
## ETHNICITYWhite                 12.25256    2.40228   5.100 4.06e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.1 on 985 degrees of freedom
## Multiple R-squared:  0.2392, Adjusted R-squared:  0.1937
## F-statistic:  5.25 on 59 and 985 DF,  p-value: < 2.2e-16
```

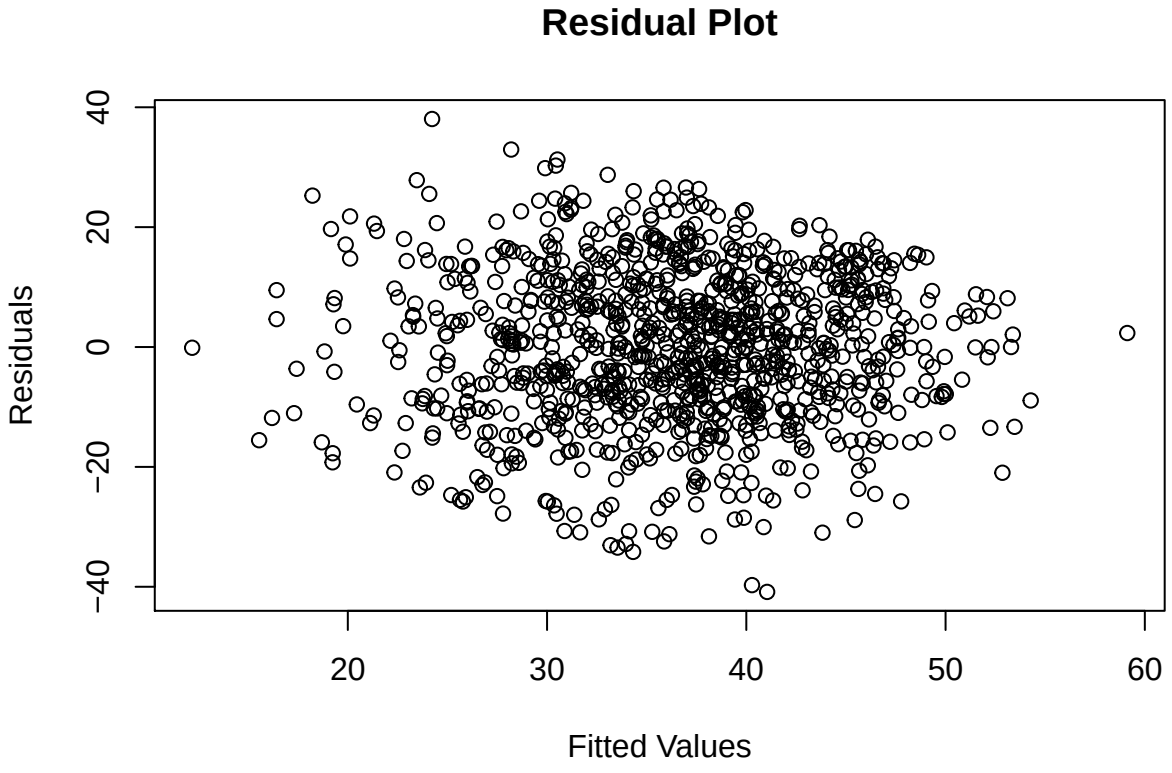
```
shapiro.test(residuals(bx.screened.ols))
```

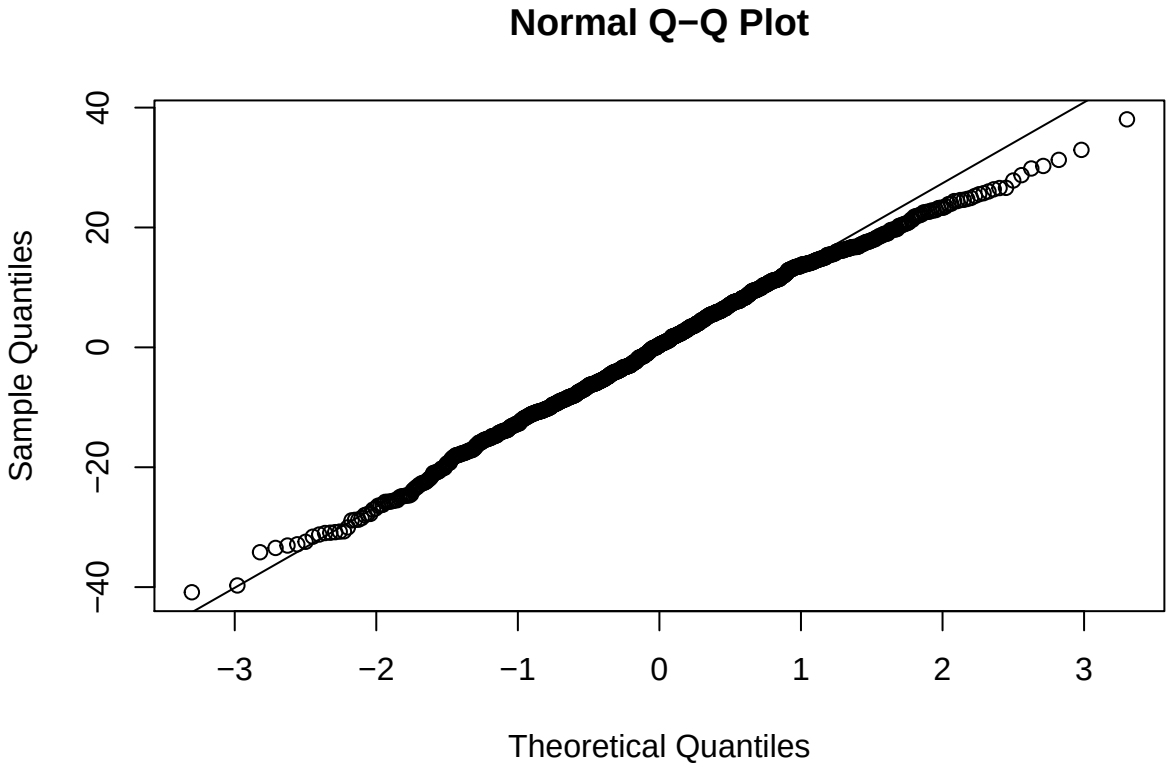
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(bx.screened.ols)
## W = 0.99539, p-value = 0.003015
```

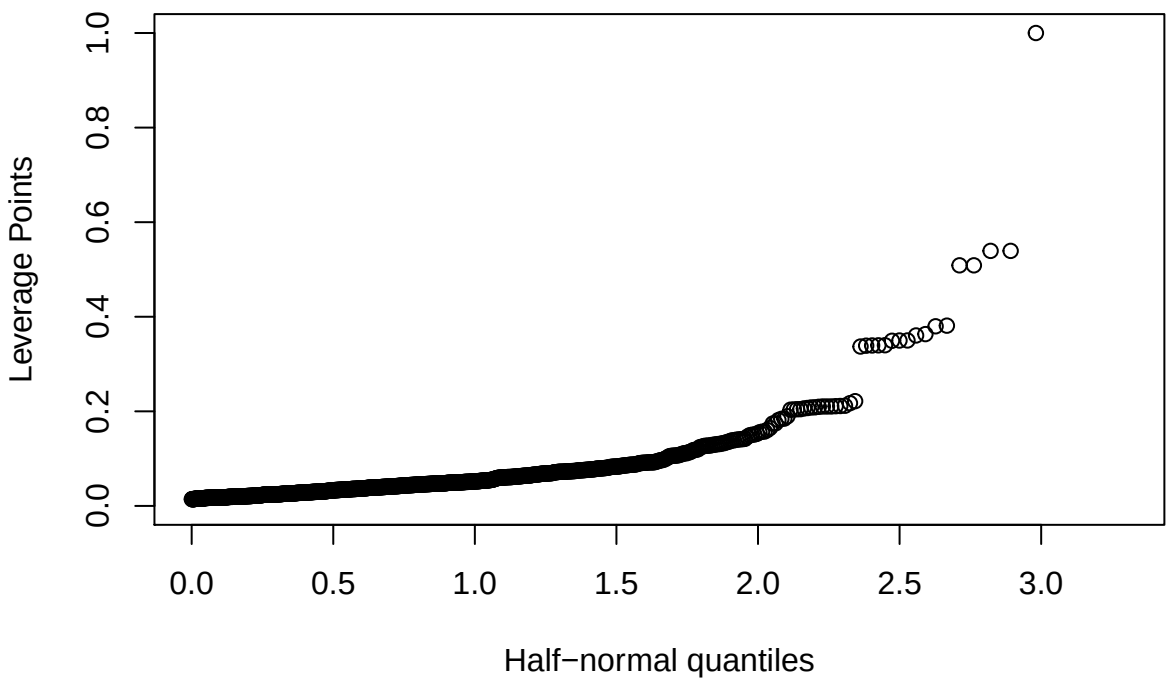
```
bptest(bx.screened.ols)
```

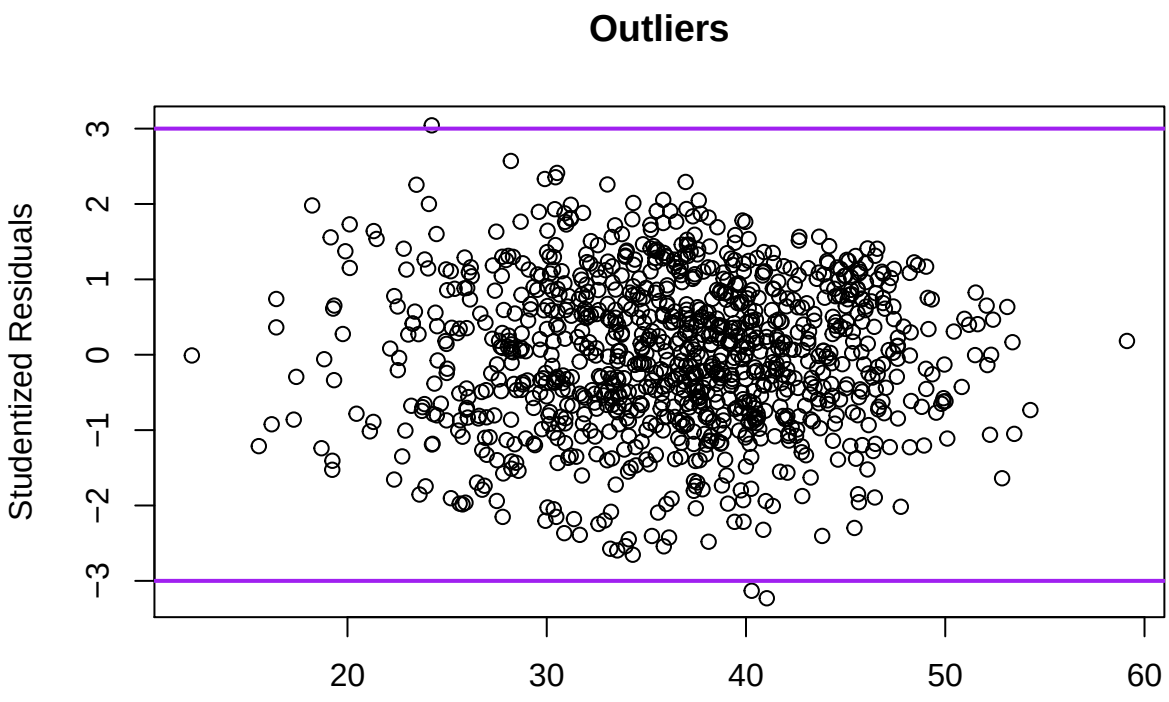
```
##
##  studentized Breusch-Pagan test
##
## data:  bx.screened.ols
## BP = 84.944, df = 59, p-value = 0.01512
```

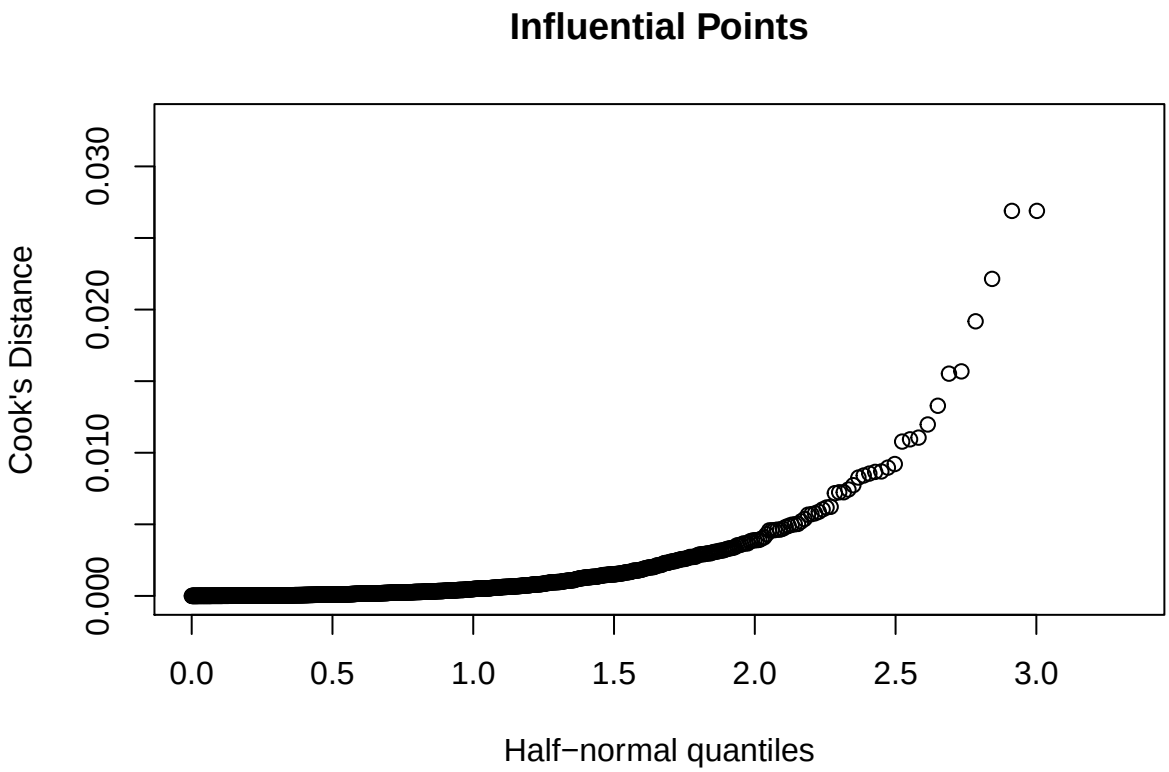
```
plot.extreme(bx.screened.ols)
```




Normal Q−Q Plot
Sample Quantiles
40
20
0
−20
−40
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Leverage Points Plot
Leverage Points
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Half−normal quantiles






```
# What is error of the model?
gpa.final.pred <- predict(bx.screened.ols, newx = y.test.screened)
gpa.final.mspe <- mean((gpa.final.pred - y.test.screened) ** 2)
```

```
## Warning in gpa.final.pred - y.test.screened: longer object length is not a
## multiple of shorter object length
```

```
print("MSPE of Final Explanatory GPA Model")
```

```
## [1] "MSPE of Final Explanatory GPA Model"
```

```
gpa.final.mspe
```

```
## [1] 1157.556
```

We have decided to use the standard rule of removing influential values with Cook's Distance greater than $4/n$, f which there are 103. We have obtained a slightly better adjusted R^2 and see more predictors as statistically significant which is an improvement. Another problem that we ran into is that some in the process of removing influential points from our model, we have incidentally excluded some majors from consideration. This may be why these students were considered influential, because they represented a very large fraction of their respective major, and in some cases, entire majors were excluded because of the small number of students in them.

Interpretation of GPA OLS Explanatory Model

The goal from this model was to form an understanding of how a freshman student's various attributes contribute to their cumulative GPA after 4 years of education at USF. In order to achieve a better model fit, we performed a transformation on the response, GPA, in particular, we used a $\lambda = 3$ transformation. So, we can say that for each coefficient value, a single unit change in X has the effect of increasing GPA by the cubed root of each coefficient. This is very rough and is far from what ideal results would be. Regardless, we will describe some of the most significant or interesting coefficients (cubed root of) and their standard errors: - Gender had a very significant impact, with a value of -6.44 and std err of 0.97. This corresponds to a value of -1.86 per unit change in X on y. This means that, approximately, being a male negatively impacts your GPA by 1.86 points compared to females. - Having a Pell grant negatively impacted a person's GPA by 1.28 as opposed to those who did not. Coef = -2.13 with std err = 1.03. Note: All dummy coded Major variables are compared to Advertising majors. - Being an Environmental Science major has a -2.60 impact on one's GPA compared to Advertising majors. Coef = -17.60 with std err = -5.74 - Though Data Science majors did not have a significant result, we did have a -2.29 impact on GPA compared to Advertising majors, but again, not statistically significant. Coef = -12.04 with std err = 13.38. Note: All dummy coded residence halls are compared to Fromm. - Those who live in Hayes had a -1.65 impact on GPA compared to Fromm. Coef = -4.5 with std err = 1.66. - Those who lived in Lomo had a -2.00 impact on GPA compared to those in Fromm. Note: All dummy encoded ethnicities are compared to African American. - Being Asian American had a 1.91 impact on GPA compared to African American students. Coef = 6.99 with std err = 2.49. - Being white had a 2.20 impact on GPA compared to African American students. Coef = 10.74 with std err = 2.47.

These were just a few of the coefficient estimates gathered for this explanatory model. There are many issues with this model, the glaring problem being an abysmal R^2 value of 0.19. After examining the coefficient estimates and their standard errors, we found that most were not statistically significant and had extremely high standard errors relative to their estimates such as the Data Science Major. I believe that there is most likely some multicollinearity between dummy variables, even though that is supposed to not happen. I believe that this multicollinearity is leading to high variance and low interpretability. We could have used Ridge Regression to solve this issue but we would have lost all interpretability in the process.

We also have the problem of failing both the Shapiro Wilkes and Breusch Pagan tests. The Shapiro Wilkes test is shown to be very sensitive to small deviations from normality, and our QQ plot still demonstrated a very near approximation to normality after our Box Cox transformation. We could have used a more optimal λ value, but our interpretability would have been even worse. Also, our residual plot is very peculiar in that there is randomness, but there appears to be almost be a random scattering of points within certain bounds. This is most likely due to the many students who received exactly 128 credits or very near that number. This is because it is the requirement to graduate at USF. This also explains the very distinguishable lines in our pairs plot of the data.

To conclude, the purposes of our study prevented us from excluding many predictors which seem to have harmed the accuracy and meaningness of our model. However, we can derive a few statistcially significant results regarding the impact that a student's ethnicity or residence hall may have on their final, cumulative GPA at USF.

GPA Predictive Model

Now that we don't have to try and fix normality assumptions, let's try and build a predictive model for a student's final, cumulative GPA.

LASSO

We actually already built a LASSO model for GPA in the Explanatory step, so let's revisit that.

```
gpa.pred.lasso <- glmnet(train.test[train, ], y.train, alpha = 1,
                         lambda = lasso.selection$lambda.min)
gpa.lasso.preds <- predict(gpa.pred.lasso, s = lasso.selection$lambda.min,
                           newx = train.test[test, ])
gpa.lasso.mspe <- mean((gpa.lasso.preds - y.test) ^ 2)
coef(gpa.pred.lasso)
```

```
## 65 x 1 sparse Matrix of class "dgCMatrix"
##                                       s0
## (Intercept)                 3.233306751
## GENDERM                    -0.177777775
## PELLY                      -0.029441064
## MAJORANST                   .
## MAJORARCD                   .
## MAJORARTM                   .
## MAJORBACT                   0.148321619
## MAJORBADM                  -0.055955945
## MAJORBAIS                   0.206294083
## MAJORBENI                   .
## MAJORBFIN                   0.130287365
## MAJORBIOL                  -0.194404945
## MAJORBMKT                   0.029500657
## MAJORBNTL                   0.079751828
## MAJORBOBL                   .
## MAJORBSDS                   0.427751421
## MAJORCDS                    0.049115800
## MAJORCHEM                  -0.147385428
## MAJORCMPL                   .
## MAJORCOMS                   0.013208197
## MAJORCS                    -0.112782382
## MAJORDSGN                   0.059083289
## MAJORECON                   .
## MAJORENGL                   0.009300161
## MAJORENVA                   .
## MAJORENVS                  -0.451279144
## MAJORESS                   -0.038766436
## MAJORFNAR                  -0.010922035
## MAJORFNEC                   .
## MAJORHIST                  -0.002491273
```

```
## MAJORHM                      -0.033973769
## MAJORJNST                     .
## MAJORMATH                     .
## MAJORMS                       0.019156372
## MAJORNURS                    -0.024238115
## MAJORPASJ                     0.013005282
## MAJORPHIL                    -0.081748598
## MAJORPHYS                    -0.121240569
## MAJORPOLS                     0.037908380
## MAJORPSYC                     .
## MAJORSOC                      .
## MAJORSPAN                     0.197953531
## MAJORTHES                     0.054516898
## MAJORUNBN                     0.155223376
## MAJORUNLA                    -0.219293502
## MAJORUNSC                    -0.233928676
## MAJORURBS                     0.254724147
## SCIENCE_CLASSES               0.007800547
## LAB_CLASSES                   0.013129435
## RESIDENCE_HALLGillson         .
## RESIDENCE_HALLHayes Healy    -0.060417282
## RESIDENCE_HALLLone Mountain -0.172272657
## RESIDENCE_HALLLone Village   -0.014547182
## RESIDENCE_HALLOff-Campus      0.011730440
## RESIDENCE_HALLPacific Wing    0.300033895
## RESIDENCE_HALLPedro Arrupe    .
## RESIDENCE_HALLPhelan          0.045484173
## ETHNICITYAsian                .
## ETHNICITYHispanic or Latino -0.103693057
## ETHNICITYInternational      -0.180710494
## ETHNICITYMulti Race           0.014562542
## ETHNICITYNative American      .
## ETHNICITYPacific Islander     0.009769720
## ETHNICITYUnknown              .
## ETHNICITYWhite                0.129055012
```

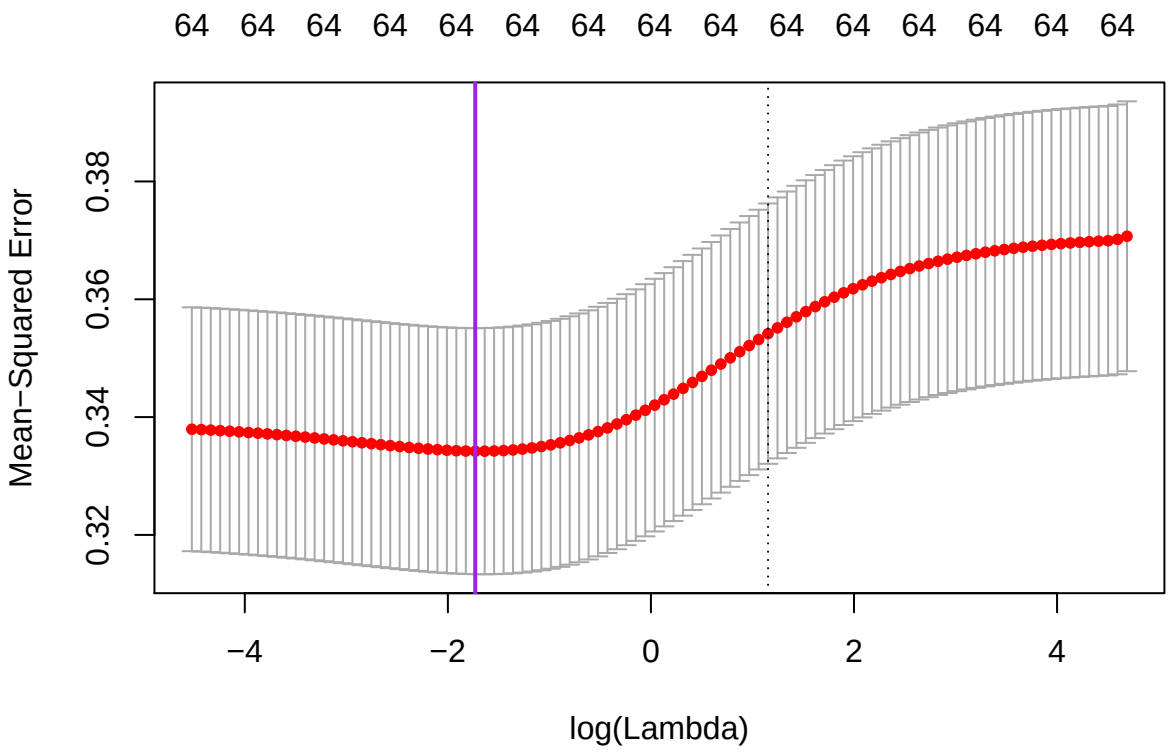
```
gpa.lasso.mspe
```

```
## [1] 0.5050713
```

Ridge

We find that our optimal LASSO model gives an MSPE of 0.405 which seems spectacular, but given the small range of values for GPA, may not be too great. Let's check cross-validated Ridge Regression model now.

```
ridge.gpa.cv <- cv.glmnet(train.test[train,], y.train, alpha = 0)
plot(ridge.gpa.cv)
abline(v = log(ridge.gpa.cv$lambda.min), col = "purple", lwd = 2)
```




```
ridge.gpa.preds <- predict(ridge.gpa.cv, s = ridge.gpa.cv$lambda.min,
                           newx = train.test[test, ])
ridge.gpa.mspe <- mean((ridge.gpa.preds - y.test) ^ 2)
ridge.gpa.mspe
```

```
## [1] 0.5056697
```

The Ridge Regression model actually gives a slightly worse test error of 0.407, however, it has the advantage of being able to predict a student's GPA no matter their major.

Final

We will use the Ridge Regression model for prediction because even though it has a slightly worse testing error, the LASSO model is not able to predict the GPA of student's from certain majors or residence halls and so gives value to less students overall.

Adam's Contribution

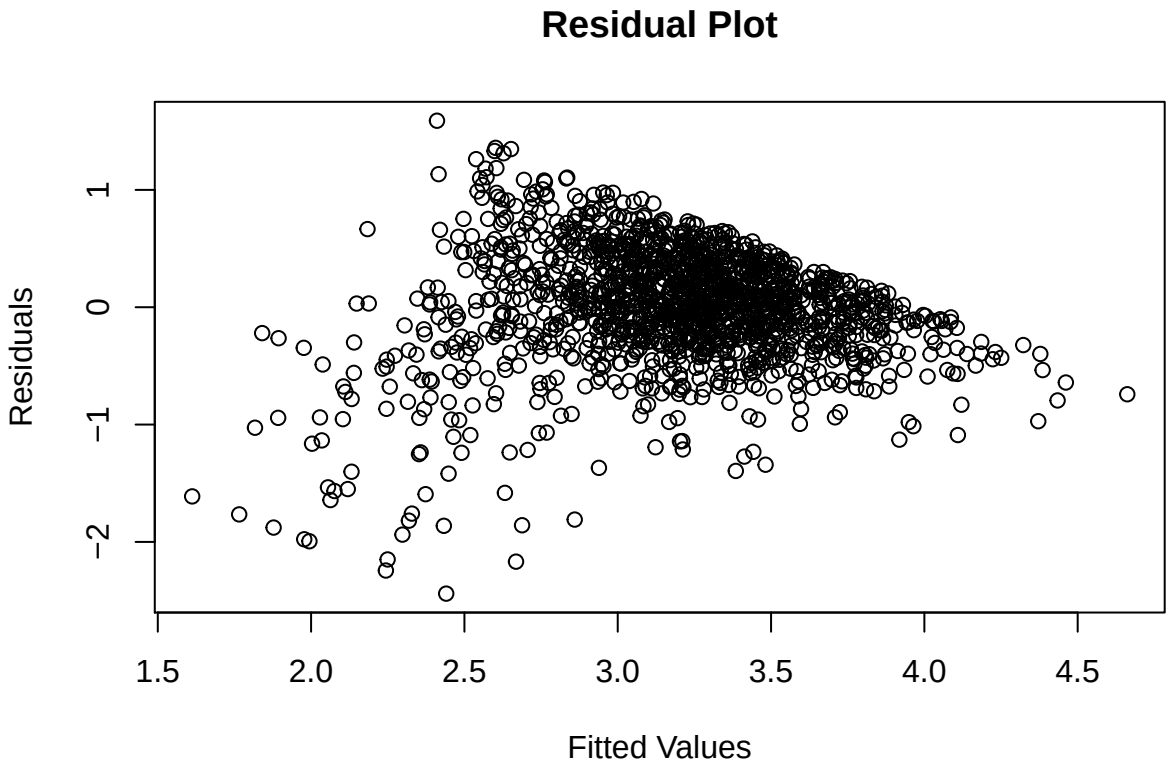
```
levels(dataset$PELL)[dataset$PELL != "Y"] <- "N"
levels(dataset$RESIDENCE_HALL)[1] <- "Off-Campus"
sum(is.na(dataset))
```

```
## [1] 14
```

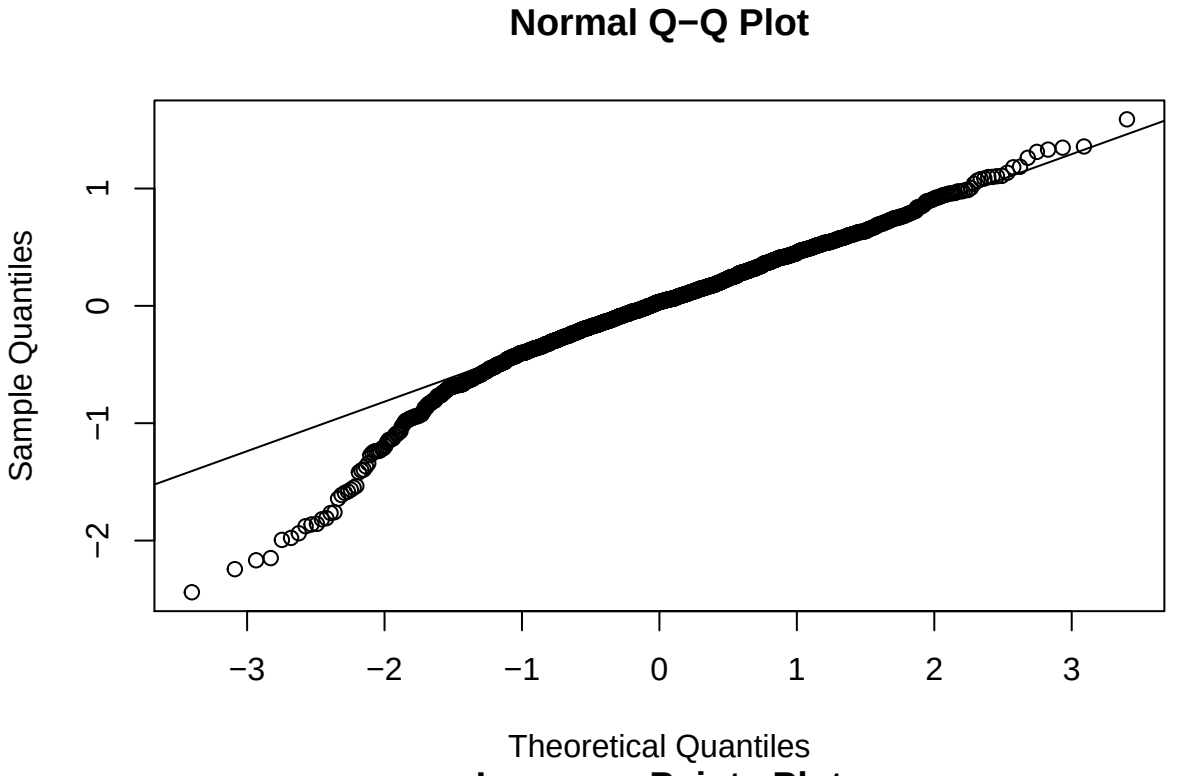
```
na.dataset <- dataset[is.na(dataset), ]
dataset <- na.omit(dataset)
```

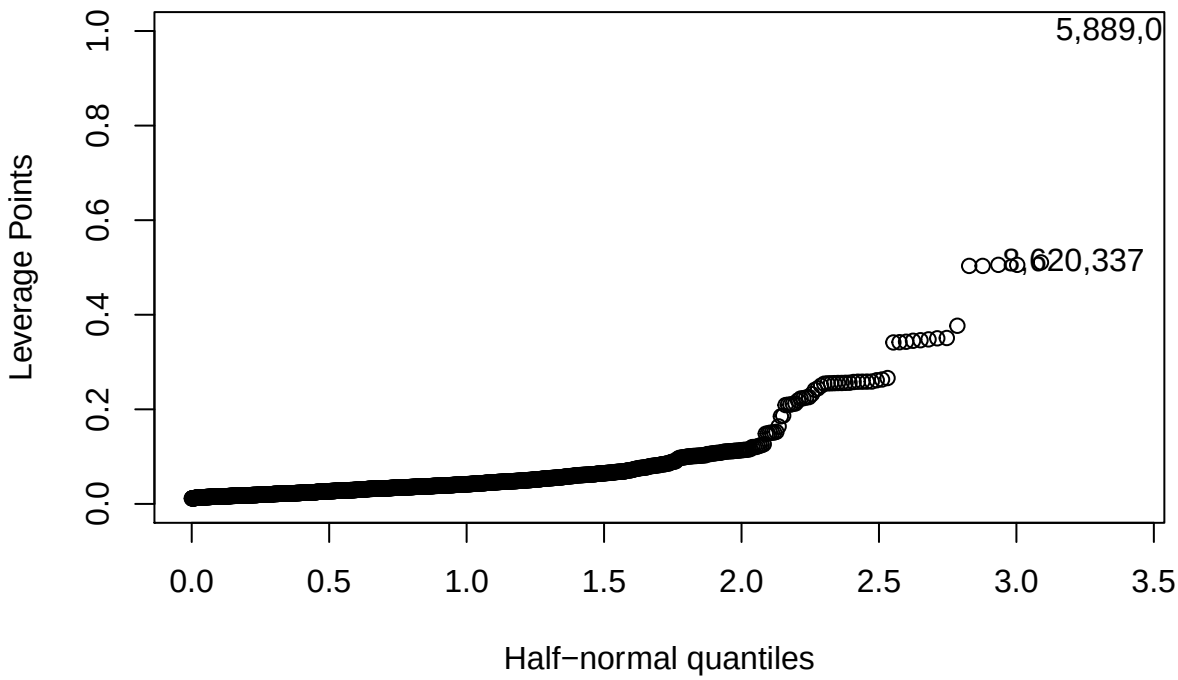
Diagnostics

```
first.model <- lm(GPA ~ . - RANDOM_ID, data=dataset)
plot.extreme(first.model, dataset$RANDOM_ID)
```

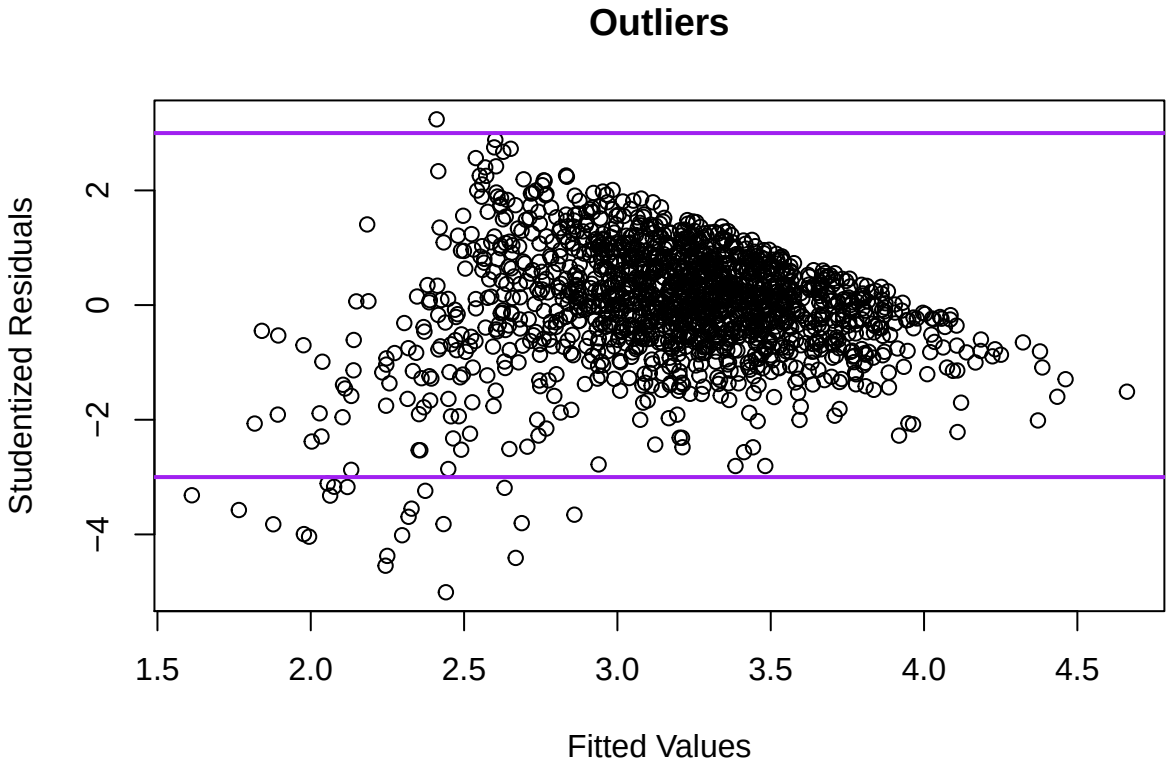


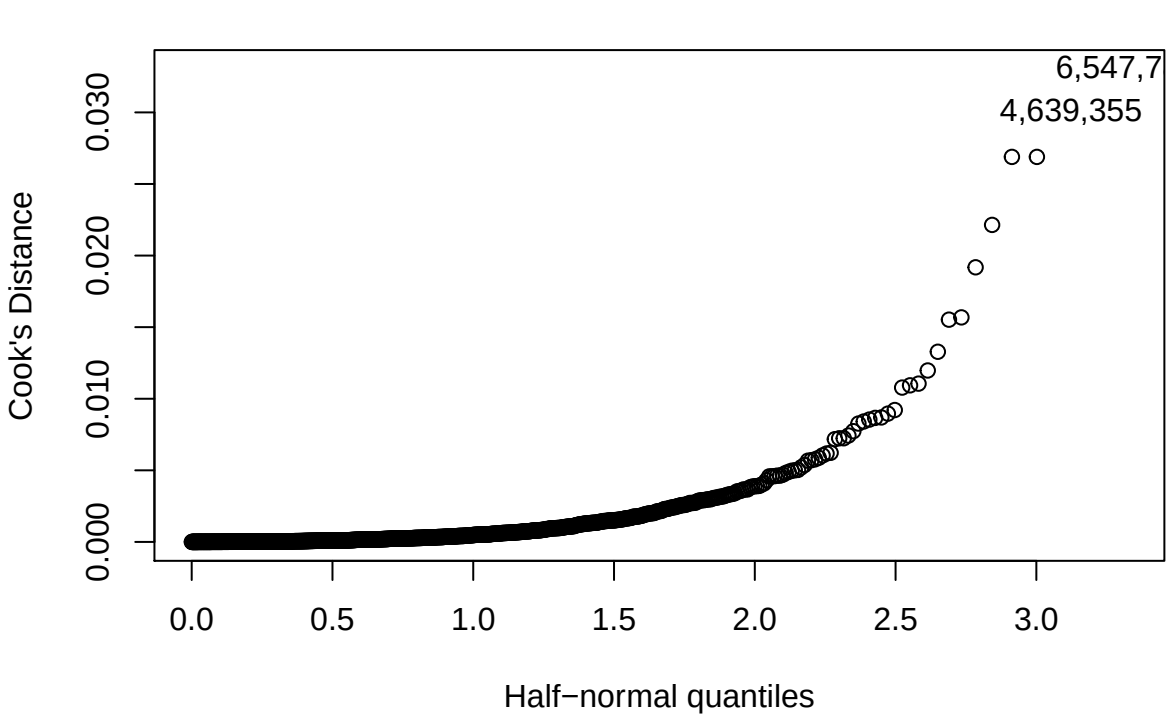

```
## Warning in if (!is.na(labs)) {: the condition has length > 1 and only the
## first element will be used
```






```
## Warning in if (!is.na(labs)) {: the condition has length > 1 and only the
## first element will be used
```





Based on Zach's plots, I will change the logistic dataset to remove the outliers, leverage and influential points to create a logistic model.

Logistic Model

I will first create a new dataset just for the logistic model. I will also delete the RANDOM_ID column and the CREDITS_EARNED. RANDOM_ID is not needed and CREDITS_EARNED determines whether the students graduate so we need to delete it.

```
logistic.col <- dataset$CREDITS_EARNED >= 128
logistic.col[logistic.col == TRUE] <- 1
logistic.dataset <- cbind(dataset, logistic.col)
logistic.dataset$RANDOM_ID <- NULL
logistic.dataset$CREDITS_EARNED <- NULL
```

I will now clean the inluential points based on the graph above.

```
#This is to clean the influential point
cooks <- cooks.distance(first.model)
threshold <- 4/nrow(logistic.dataset)
cooks.delete <- which(cooks>threshold)
#logistic.dataset <- logistic.dataset[-c(cooks.delete), ]

#This is to clean the outliers
student.residuals <- rstudent(first.model)
student.residuals.delete <- which(student.residuals > 3 | student.residuals < -3)

#Unionize both vectors to recieve all indices to delete at once
delete.vector <- union(cooks.delete, student.residuals.delete)
logistic.dataset <- logistic.dataset[-c(delete.vector), ]
total_num <- nrow(logistic.dataset)
num_graduated <- sum(logistic.dataset$logistic.col == 1)
num_notgraduated <- total_num - num_graduated

total_num
```

```
## [1] 1393
```

```
num_graduated
```

```
## [1] 1052
```

```
num_notgraduated
```

```
## [1] 341
```

After deleting the influential and the outlier points, there are 1393 students in total in the dataset. Out of the 1393 students, 1052 graduated while 341 did not. One issue of thinking about running a logistic model over the data is that it is very unbalanced. So the best decision is to keep this dataset and create a new balanced dataset. I will keep 341 students that graduated and the 341 students that did not graduate and combine them to create a balanced dataset of 682 students.

```
#I set the random number generator state to be able to split the training and testing sets to create th
set.seed(150)
unbalanced.train.indices <- sample(seq_len(nrow(logistic.dataset)), size = floor(0.75 * nrow(logistic.da

unbalanced.train <- logistic.dataset[unbalanced.train.indices, ]
unbalanced.test <- logistic.dataset[-unbalanced.train.indices, ]
unbalanced.y.train <- unbalanced.train$logistic.col
unbalanced.y.test <- unbalanced.test$logistic.col
```

```r
graduated <- logistic.dataset[logistic.dataset$logistic.col==1, ]
balanced.graduated <- graduated[1: 341,]
balanced.notgraduated <- logistic.dataset[logistic.dataset$logistic.col==0, ]
balanced.logistic.dataset <- rbind(balanced.graduated, balanced.notgraduated)

#Need to randomize the observations to create a train and test set
balanced.logistic.dataset <- balanced.logistic.dataset[sample(nrow(balanced.logistic.dataset)), ]

train_indices <- sample(seq_len(nrow(balanced.logistic.dataset)), size = floor(0.75 * nrow(balanced.logi

train <- balanced.logistic.dataset[train_indices, ]
test <- balanced.logistic.dataset[-train_indices, ]

#I will find the difference of data
diff.majors2 <- setdiff(test$MAJOR, train$MAJOR)
delete.majors2 <- which(test$MAJOR %in% diff.majors2)
train <- train[-c(delete.majors2), ]
test <- test[-c(delete.majors2), ]

y.test <- test$logistic.col
y.train <- train$logistic.col
```

Here above, I created the training and testing dataset of the balanced dataset. The training dataset has 75% of the data in the balanced dataset while the testing dataset has 25%.

First logistic model

This model will run on unbalanced data

```r
logistic.model1 <- glm(logistic.col ~ ., family = "binomial", data = unbalanced.train)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
#You must round
model1preds <- round(predict(logistic.model1, unbalanced.test, type = "response"))
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading
```

```r
summary(logistic.model1)
```

```
##
## Call:
## glm(formula = logistic.col ~ ., family = "binomial", data = unbalanced.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2931   0.0001   0.0166   0.0650   2.2648
##
## Coefficients: (41 not defined because of singularities)
##                                         Estimate Std. Error z value
## (Intercept)                           -37.32553    6.49018  -5.751
## GENDERM                                -0.53315    0.71585  -0.745
## IN_STATEY                              -0.07218    0.78678  -0.092
## UNMET_NEED_PERCENT                     -0.01315    0.01513  -0.869
```

```
## PELLY                                  -0.31941    1.12785  -0.283
## MAJORANST                              11.70010 4348.27594   0.003
## MAJORARCD                               3.94616   13.86442   0.285
## MAJORARTM                               3.28726  123.43306   0.027
## MAJORBACT                               6.11907   19.31216   0.317
## MAJORBADM                               0.47296    3.92105   0.121
## MAJORBAIS                               5.94969    7.14290   0.833
## MAJORBENI                               4.76371   24.15243   0.197
## MAJORBFIN                               2.08627    4.15774   0.502
## MAJORBIOL                               0.82868    4.01531   0.206
## MAJORBMKT                               5.41322    6.98774   0.775
## MAJORBNTL                               4.97713   12.57748   0.396
## MAJORBOBL                              12.73918 2913.65190   0.004
## MAJORBSDS                              -7.44513 6522.63990  -0.001
## MAJORCHEM                               2.74273    4.27817   0.641
## MAJORCMPL                               9.36925 6522.63995   0.001
## MAJORCOMS                               0.51009    4.17755   0.122
## MAJORCS                                 3.92238    4.25302   0.922
## MAJORDSGN                               0.78218    4.39271   0.178
## MAJORECON                               6.22909   10.87000   0.573
## MAJORENGL                              -1.08501    4.03268  -0.269
## MAJORENVA                               3.31554   10.24052   0.324
## MAJORENVS                               6.87968   26.21439   0.262
## MAJORESS                                1.55477    4.16318   0.373
## MAJORFNAR                              -5.31869    5.51460  -0.964
## MAJORHIST                              -1.75307    4.17951  -0.419
## MAJORHM                                 3.38600    5.45643   0.621
## MAJORJNST                              16.18295 6522.63982   0.002
## MAJORMATH                               4.68645   12.19576   0.384
## MAJORMS                                 6.24367   17.21388   0.363
## MAJORNURS                               1.21620    4.14905   0.293
## MAJORPASJ                               2.83031   12.66497   0.223
## MAJORPHIL                              14.43520 4540.99293   0.003
## MAJORPHYS                               1.53989    4.60105   0.335
## MAJORPOLS                              -0.53012    4.07519  -0.130
## MAJORPSYC                               0.39310    3.94556   0.100
## MAJORSOC                                1.32167    4.57304   0.289
## MAJORSPAN                               9.86558 6522.63996   0.002
## MAJORUNBN                               1.62032    4.05208   0.400
## MAJORUNLA                               0.87418    4.00496   0.218
## MAJORUNSC                               0.91782    4.40985   0.208
## MAJORURBS                               9.78686 6522.63978   0.002
## MAJOR_DESCAdvertising                        NA         NA      NA
## MAJOR_DESCArchitecture & Community Design    NA         NA      NA
## MAJOR_DESCArt History/Arts Management        NA         NA      NA
## MAJOR_DESCAsian Studies                      NA         NA      NA
## MAJOR_DESCBiology                            NA         NA      NA
## MAJOR_DESCBusiness Administration            NA         NA      NA
## MAJOR_DESCChemistry                          NA         NA      NA
## MAJOR_DESCCommunication Studies              NA         NA      NA
## MAJOR_DESCComparative Lit. & Culture         NA         NA      NA
## MAJOR_DESCComputer Science                   NA         NA      NA
## MAJOR_DESCData Science                       NA         NA      NA
## MAJOR_DESCDesign                             NA         NA      NA
```

```
## MAJOR_DESCEconomics                                NA         NA      NA
## MAJOR_DESCEnglish                                  NA         NA      NA
## MAJOR_DESCEntrepreneurship & Innovation            NA         NA      NA
## MAJOR_DESCEnvironmental Science                    NA         NA      NA
## MAJOR_DESCEnvironmental Studies                    NA         NA      NA
## MAJOR_DESCExercise and Sport Science               NA         NA      NA
## MAJOR_DESCFinance                                  NA         NA      NA
## MAJOR_DESCFine Arts                                NA         NA      NA
## MAJOR_DESCHistory                                  NA         NA      NA
## MAJOR_DESCHospitality Management                   NA         NA      NA
## MAJOR_DESCInternational Business                   NA         NA      NA
## MAJOR_DESCInternational Studies                    NA         NA      NA
## MAJOR_DESCJapanese Studies                         NA         NA      NA
## MAJOR_DESCMarketing                                NA         NA      NA
## MAJOR_DESCMathematics                              NA         NA      NA
## MAJOR_DESCMedia Studies                            NA         NA      NA
## MAJOR_DESCNursing                                  NA         NA      NA
## MAJOR_DESCOrganizational Behav.& Ldrship           NA         NA      NA
## MAJOR_DESCPerf. Arts & Soc. Justice                NA         NA      NA
## MAJOR_DESCPhilosophy                               NA         NA      NA
## MAJOR_DESCPhysics                                  NA         NA      NA
## MAJOR_DESCPolitics                                 NA         NA      NA
## MAJOR_DESCPsychology                               NA         NA      NA
## MAJOR_DESCSociology                                NA         NA      NA
## MAJOR_DESCSpanish                                  NA         NA      NA
## MAJOR_DESCUndeclared Arts                          NA         NA      NA
## MAJOR_DESCUndeclared Business                      NA         NA      NA
## MAJOR_DESCUndeclared Sciences                      NA         NA      NA
## MAJOR_DESCUrban Studies                            NA         NA      NA
## SCIENCE_CLASSES                               -0.13761    0.06389  -2.154
## LAB_CLASSES                                    0.25485    0.14884   1.712
## RESIDENCE_HALLFromm                           -1.80756    1.59872  -1.131
## RESIDENCE_HALLGillson                          0.12185    1.39404   0.087
## RESIDENCE_HALLHayes Healy                     -0.91910    1.28923  -0.713
## RESIDENCE_HALLLone Mountain                    1.89798    3.90020   0.487
## RESIDENCE_HALLLone Village                     2.77001   14.67608   0.189
## RESIDENCE_HALLPacific Wing                    -3.98028 6522.63931  -0.001
## RESIDENCE_HALLPhelan                          -0.04159    1.43294  -0.029
## ETHNICITYAsian                                -0.46107    1.84216  -0.250
## ETHNICITYHispanic or Latino                    0.18795    1.63363   0.115
## ETHNICITYInternational                        -2.28374    2.13299  -1.071
## ETHNICITYMulti Race                           -1.43438    1.86602  -0.769
## ETHNICITYNative American                      -1.15751 1113.92248  -0.001
## ETHNICITYPacific Islander                      3.05670   26.27499   0.116
## ETHNICITYUnknown                               2.38258   17.45725   0.136
## ETHNICITYWhite                                -0.98489    1.85235  -0.532
## GPA_CREDITS                                    0.20341    0.02808   7.244
## GPA                                            5.80996    0.98652   5.889
##                                              Pr(>|z|)
## (Intercept)                                  8.87e-09 ***
## GENDERM                                       0.4564
## IN_STATEY                                     0.9269
## UNMET_NEED_PERCENT                            0.3849
## PELLY                                         0.7770
```

```
## MAJORANST                                0.9979
## MAJORARCD                                0.7759
## MAJORARTM                                0.9788
## MAJORBACT                                0.7514
## MAJORBADM                                0.9040
## MAJORBAIS                                0.4049
## MAJORBENI                                0.8436
## MAJORBFIN                                0.6158
## MAJORBIOL                                0.8365
## MAJORBMKT                                0.4385
## MAJORBNTL                                0.6923
## MAJORBOBL                                0.9965
## MAJORBSDS                                0.9991
## MAJORCHEM                                0.5215
## MAJORCMPL                                0.9989
## MAJORCOMS                                0.9028
## MAJORCS                                  0.3564
## MAJORDSGN                                0.8587
## MAJORECON                                0.5666
## MAJORENGL                                0.7879
## MAJORENVA                                0.7461
## MAJORENVS                                0.7930
## MAJORESS                                 0.7088
## MAJORFNAR                                0.3348
## MAJORHIST                                0.6749
## MAJORHM                                  0.5349
## MAJORJNST                                0.9980
## MAJORMATH                                0.7008
## MAJORMS                                  0.7168
## MAJORNURS                                0.7694
## MAJORPASJ                                0.8232
## MAJORPHIL                                0.9975
## MAJORPHYS                                0.7379
## MAJORPOLS                                0.8965
## MAJORPSYC                                0.9206
## MAJORSOC                                 0.7726
## MAJORSPAN                                0.9988
## MAJORUNBN                                0.6893
## MAJORUNLA                                0.8272
## MAJORUNSC                                0.8351
## MAJORURBS                                0.9988
## MAJOR_DESCAdvertising                       NA
## MAJOR_DESCArchitecture & Community Design   NA
## MAJOR_DESCArt History/Arts Management       NA
## MAJOR_DESCAsian Studies                     NA
## MAJOR_DESCBiology                           NA
## MAJOR_DESCBusiness Administration           NA
## MAJOR_DESCChemistry                         NA
## MAJOR_DESCCommunication Studies             NA
## MAJOR_DESCComparative Lit. & Culture        NA
## MAJOR_DESCComputer Science                  NA
## MAJOR_DESCData Science                      NA
## MAJOR_DESCDesign                            NA
## MAJOR_DESCEconomics                         NA
```

```
## MAJOR_DESCEnglish                                NA
## MAJOR_DESCEntrepreneurship & Innovation          NA
## MAJOR_DESCEnvironmental Science                  NA
## MAJOR_DESCEnvironmental Studies                  NA
## MAJOR_DESCExercise and Sport Science             NA
## MAJOR_DESCFinance                                NA
## MAJOR_DESCFine Arts                              NA
## MAJOR_DESCHistory                                NA
## MAJOR_DESCHospitality Management                 NA
## MAJOR_DESCInternational Business                 NA
## MAJOR_DESCInternational Studies                  NA
## MAJOR_DESCJapanese Studies                       NA
## MAJOR_DESCMarketing                              NA
## MAJOR_DESCMathematics                            NA
## MAJOR_DESCMedia Studies                          NA
## MAJOR_DESCNursing                                NA
## MAJOR_DESCOrganizational Behav.& Ldrship         NA
## MAJOR_DESCPerf. Arts & Soc. Justice              NA
## MAJOR_DESCPhilosophy                             NA
## MAJOR_DESCPhysics                                NA
## MAJOR_DESCPolitics                               NA
## MAJOR_DESCPsychology                             NA
## MAJOR_DESCSociology                              NA
## MAJOR_DESCSpanish                                NA
## MAJOR_DESCUndeclared Arts                        NA
## MAJOR_DESCUndeclared Business                    NA
## MAJOR_DESCUndeclared Sciences                    NA
## MAJOR_DESCUrban Studies                          NA
## SCIENCE_CLASSES                             0.0313 *
## LAB_CLASSES                                 0.0868 .
## RESIDENCE_HALLFromm                         0.2582
## RESIDENCE_HALLGillson                       0.9303
## RESIDENCE_HALLHayes Healy                   0.4759
## RESIDENCE_HALLLone Mountain                 0.6265
## RESIDENCE_HALLLone Village                  0.8503
## RESIDENCE_HALLPacific Wing                  0.9995
## RESIDENCE_HALLPhelan                        0.9768
## ETHNICITYAsian                              0.8024
## ETHNICITYHispanic or Latino                 0.9084
## ETHNICITYInternational                      0.2843
## ETHNICITYMulti Race                         0.4421
## ETHNICITYNative American                    0.9992
## ETHNICITYPacific Islander                   0.9074
## ETHNICITYUnknown                            0.8914
## ETHNICITYWhite                              0.5949
## GPA_CREDITS                              4.36e-13 ***
## GPA                                      3.88e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1156.249  on 1043  degrees of freedom
## Residual deviance:   99.467  on  979  degrees of freedom
```

```
## AIC: 229.47
##
## Number of Fisher Scoring iterations: 17
```

```
which(logistic.model1$coefficients>0)
```

```
##                   MAJORANST                   MAJORARCD
##                           6                           7
##                   MAJORARTM                   MAJORBACT
##                           8                           9
##                   MAJORBADM                   MAJORBAIS
##                          10                          11
##                   MAJORBENI                   MAJORBFIN
##                          12                          13
##                   MAJORBIOL                   MAJORBMKT
##                          14                          15
##                   MAJORBNTL                   MAJORBOBL
##                          16                          17
##                   MAJORCHEM                   MAJORCMPL
##                          19                          20
##                   MAJORCOMS                     MAJORCS
##                          21                          22
##                   MAJORDSGN                   MAJORECON
##                          23                          24
##                   MAJORENVA                   MAJORENVS
##                          26                          27
##                    MAJORESS                     MAJORHM
##                          28                          31
##                   MAJORJNST                   MAJORMATH
##                          32                          33
##                     MAJORMS                   MAJORNURS
##                          34                          35
##                   MAJORPASJ                   MAJORPHIL
##                          36                          37
##                   MAJORPHYS                   MAJORPSYC
##                          38                          40
##                    MAJORSOC                   MAJORSPAN
##                          41                          42
##                   MAJORUNBN                   MAJORUNLA
##                          43                          44
##                   MAJORUNSC                   MAJORURBS
##                          45                          46
##                 LAB_CLASSES      RESIDENCE_HALLGillson
##                          89                          91
## RESIDENCE_HALLLone Mountain  RESIDENCE_HALLLone Village
##                          93                          94
## ETHNICITYHispanic or Latino   ETHNICITYPacific Islander
##                          98                         102
##            ETHNICITYUnknown                GPA_CREDITS
##                         103                         105
##                         GPA
##                         106
```

After creating the logistic model with the unbalanced data, we can see what are the positive coeffecient values that lead to whether a person graduated. Based on this random sample of the data, it seems that

most majors contribute positively for a student to graduate. The coeffiecent for the Japanese Studies major is the highest, but only one student graduated with Japanese Studies as a major. The number of Lab Classes does not contribute much to whether that student graduates or not. The best dorm building is Loyola Village with the highest coefficient of all the dorm buildings. Most ethnicities contribute badly, except if you are Pacific Islander, Hispanic or Unknown. However, this does not insinuate that some ethnicities perform better than others since it is the number of observations between ethnicities are not equal. Of course, the GPA_CREDITS and the GPA have the most significance of determing whether a person graduates. This should not be surprising that since if the GPA of a student is higher than most, it establishes that student is achieving academically enough to graduate.

Disclaimer: I set the random number generator so that the model and its coefficients can be repeated. I ran the model multiple times and recieved mixed results, this is just an example of interpreting the model.

```
first.loss <- mean((model1preds - unbalanced.y.test)^2)
first.accuracy <- 1 - first.loss
first.accuracy
```

```
## [1] 0.9713467
```

The accuracy of the first model is about 0.97 so 97% correct which is pretty good. However, the dataset is unblananced, so the model has seen mostly observations that show students that graduated, and not as much observations that show students that did not graduate. However, we can assess the model more.

```
unbalanced.y.test.equals1 <- unbalanced.y.test == 1
model1preds.equals1 <- model1preds == 1
unbalanced.y.test.equals0 <- unbalanced.y.test == 0
model1preds.equals0 <- model1preds == 0

true.positives1 <- sum(unbalanced.y.test.equals1 == T & model1preds.equals1==T)
false.positives1 <- sum(model1preds.equals1 == T & unbalanced.y.test.equals0==T)
true.negatives1 <- sum(unbalanced.y.test.equals0 == T &  model1preds.equals0 == T)
false.negatives1 <- sum(model1preds.equals0 == T & unbalanced.y.test.equals1==T)
true.positives1
```

```
## [1] 257
```

```
false.positives1
```

```
## [1] 6
```

```
true.negatives1
```

```
## [1] 82
```

```
false.negatives1
```

```
## [1] 4
```

For a logistic model to be better, we want to decrease the number of False-Positives and False-Negatives. We would also like to increase the number of True-Positives and True-Negatives.

```
first.accuracy <- (true.positives1+true.negatives1)/length(model1preds)
first.sensitity <- true.positives1/(true.positives1 + false.negatives1)
first.specifity <- true.negatives1/(true.negatives1 + false.positives1)
first.precision <- true.positives1/(true.positives1 + false.positives1)

first.accuracy
```

```
## [1] 0.9713467
```

```
first.sensitity
```

```
## [1] 0.9846743
```

```
first.specifity
```

```
## [1] 0.9318182
```

```
first.precision
```

```
## [1] 0.9771863
```

The accuracy of the first model is 97%, which is really good. The sensitivity of the model is 98% which is the true positive rate. The specifity of the model is 93% which is the true negative rate. The precision of the model is 98% which is the positive predictive value. On face value, it seems that this model has done well.

Second logistic model

This will used the balanced data

Note: I deleted the some majors in the variable MAJOR, because there were factors that were being tested that the model was not trained on, so I deleted the majors that were unique to the testing dataset. There is only one way to get around this, and it is to oversample the data, so that those unique majors would have not been deleted from our dataset.

```
#If there is an issue, please run the chunk where train is initialized
logistic.model2 <- glm(logistic.col ~ ., family = "binomial", data = train)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
#You must round
model2preds <- round(predict(logistic.model2, test, type = "response"))
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading
```

```
summary(logistic.model2)
```

```
##
## Call:
## glm(formula = logistic.col ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.93029  -0.00003   0.00000   0.00944   1.74655
##
## Coefficients: (37 not defined because of singularities)
##                                        Estimate Std. Error z value
## (Intercept)                          -6.771e+01  5.119e+01  -1.323
## GENDERM                               8.693e-01  1.106e+00   0.786
## IN_STATEY                             3.173e+00  1.456e+00   2.179
## UNMET_NEED_PERCENT                    1.443e-02  2.517e-02   0.573
## PELLY                                -4.053e+00  2.041e+00  -1.986
## MAJORANST                             5.726e-01  1.075e+04   0.000
## MAJORARCD                            -1.050e+01  4.882e+01  -0.215
## MAJORARTM                             4.171e+00  2.937e+03   0.001
## MAJORBACT                             1.011e+01  1.123e+02   0.090
```

```
## MAJORBADM                                8.256e-01  4.862e+01   0.017
## MAJORBAIS                                9.042e+00  6.724e+01   0.134
## MAJORBENI                                1.606e+00  4.681e+03   0.000
## MAJORBFIN                                6.904e-03  4.946e+01   0.000
## MAJORBIOL                               -2.434e+00  4.865e+01  -0.050
## MAJORBMKT                                4.030e+00  4.872e+01   0.083
## MAJORBNTL                                8.784e+00  9.461e+01   0.093
## MAJORBOBL                                8.271e+00  1.075e+04   0.001
## MAJORBSDS                               -2.769e+00  1.075e+04   0.000
## MAJORCHEM                                1.652e+00  4.877e+01   0.034
## MAJORCOMS                               -1.284e+00  4.866e+01  -0.026
## MAJORCS                                  8.474e+00  4.871e+01   0.174
## MAJORDSGN                                7.418e+00  6.366e+01   0.117
## MAJORECON                                1.060e+01  5.118e+01   0.207
## MAJORENGL                               -8.575e+00  4.868e+01  -0.176
## MAJORENVA                                2.788e+00  3.318e+02   0.008
## MAJORENVS                                1.987e+01  3.989e+03   0.005
## MAJORESS                                 1.260e+01  1.602e+02   0.079
## MAJORFNAR                               -1.881e+01  5.460e+03  -0.003
## MAJORHIST                               -7.985e-01  4.872e+01  -0.016
## MAJORHM                                  7.584e-01  6.162e+01   0.012
## MAJORJNST                                1.732e+01  1.075e+04   0.002
## MAJORMATH                                1.633e+01  1.075e+04   0.002
## MAJORMS                                  4.494e+00  5.152e+01   0.087
## MAJORNURS                               -3.195e+00  4.870e+01  -0.066
## MAJORPASJ                               -4.739e+00  2.170e+02  -0.022
## MAJORPHYS                               -1.228e+01  5.079e+03  -0.002
## MAJORPOLS                                1.473e+00  4.863e+01   0.030
## MAJORPSYC                               -1.354e+00  4.865e+01  -0.028
## MAJORSOC                                 3.924e+00  5.055e+01   0.078
## MAJORUNBN                                3.419e+00  4.879e+01   0.070
## MAJORUNLA                                7.461e-01  4.864e+01   0.015
## MAJORUNSC                                5.071e-01  4.868e+01   0.010
## MAJOR_DESCAdvertising                           NA         NA      NA
## MAJOR_DESCArchitecture & Community Design       NA         NA      NA
## MAJOR_DESCArt History/Arts Management           NA         NA      NA
## MAJOR_DESCAsian Studies                         NA         NA      NA
## MAJOR_DESCBiology                               NA         NA      NA
## MAJOR_DESCBusiness Administration               NA         NA      NA
## MAJOR_DESCChemistry                             NA         NA      NA
## MAJOR_DESCCommunication Studies                 NA         NA      NA
## MAJOR_DESCComputer Science                      NA         NA      NA
## MAJOR_DESCData Science                          NA         NA      NA
## MAJOR_DESCDesign                                NA         NA      NA
## MAJOR_DESCEconomics                             NA         NA      NA
## MAJOR_DESCEnglish                               NA         NA      NA
## MAJOR_DESCEntrepreneurship & Innovation         NA         NA      NA
## MAJOR_DESCEnvironmental Science                 NA         NA      NA
## MAJOR_DESCEnvironmental Studies                 NA         NA      NA
## MAJOR_DESCExercise and Sport Science            NA         NA      NA
## MAJOR_DESCFinance                               NA         NA      NA
## MAJOR_DESCFine Arts                             NA         NA      NA
## MAJOR_DESCHistory                               NA         NA      NA
## MAJOR_DESCHospitality Management                NA         NA      NA
```

```
## MAJOR_DESCInternational Business                    NA         NA      NA
## MAJOR_DESCInternational Studies                     NA         NA      NA
## MAJOR_DESCJapanese Studies                          NA         NA      NA
## MAJOR_DESCMarketing                                 NA         NA      NA
## MAJOR_DESCMathematics                               NA         NA      NA
## MAJOR_DESCMedia Studies                             NA         NA      NA
## MAJOR_DESCNursing                                   NA         NA      NA
## MAJOR_DESCOrganizational Behav.& Ldrship            NA         NA      NA
## MAJOR_DESCPerf. Arts & Soc. Justice                 NA         NA      NA
## MAJOR_DESCPhysics                                   NA         NA      NA
## MAJOR_DESCPolitics                                  NA         NA      NA
## MAJOR_DESCPsychology                                NA         NA      NA
## MAJOR_DESCSociology                                 NA         NA      NA
## MAJOR_DESCUndeclared Arts                           NA         NA      NA
## MAJOR_DESCUndeclared Business                       NA         NA      NA
## MAJOR_DESCUndeclared Sciences                       NA         NA      NA
## SCIENCE_CLASSES                           -2.422e-01  1.308e-01  -1.853
## LAB_CLASSES                                5.175e-01  3.087e-01   1.677
## RESIDENCE_HALLFromm                       -4.474e-01  2.901e+00  -0.154
## RESIDENCE_HALLGillson                     -2.539e+00  2.465e+00  -1.030
## RESIDENCE_HALLHayes Healy                 -3.165e+00  2.289e+00  -1.383
## RESIDENCE_HALLLone Mountain               -9.680e-01  8.516e+01  -0.011
## RESIDENCE_HALLLone Village                 3.677e-01  8.811e+01   0.004
## RESIDENCE_HALLPacific Wing                 5.519e+00  1.075e+04   0.001
## RESIDENCE_HALLPhelan                      -2.120e+00  2.020e+00  -1.049
## ETHNICITYAsian                            -6.138e-02  2.262e+00  -0.027
## ETHNICITYHispanic or Latino                2.029e-01  1.840e+00   0.110
## ETHNICITYInternational                    -1.790e+00  2.571e+00  -0.696
## ETHNICITYMulti Race                       -1.659e+00  2.255e+00  -0.736
## ETHNICITYNative American                  -5.021e+00  1.075e+04   0.000
## ETHNICITYPacific Islander                  6.853e+00  1.027e+02   0.067
## ETHNICITYUnknown                           3.984e+00  2.644e+02   0.015
## ETHNICITYWhite                            -7.732e-01  1.980e+00  -0.391
## GPA_CREDITS                                3.169e-01  7.386e-02   4.291
## GPA                                        1.098e+01  2.831e+00   3.878
##                                           Pr(>|z|)
## (Intercept)                               0.185905
## GENDERM                                   0.432060
## IN_STATEY                                 0.029343 *
## UNMET_NEED_PERCENT                        0.566516
## PELLY                                     0.047033 *
## MAJORANST                                 0.999958
## MAJORARCD                                 0.829754
## MAJORARTM                                 0.998867
## MAJORBACT                                 0.928258
## MAJORBADM                                 0.986451
## MAJORBAIS                                 0.893025
## MAJORBENI                                 0.999726
## MAJORBFIN                                 0.999889
## MAJORBIOL                                 0.960103
## MAJORBMKT                                 0.934075
## MAJORBNTL                                 0.926029
## MAJORBOBL                                 0.999386
## MAJORBSDS                                 0.999795
```

```
## MAJORCHEM                                0.972988
## MAJORCOMS                                0.978951
## MAJORCS                                  0.861903
## MAJORDSGN                                0.907239
## MAJORECON                                0.835907
## MAJORENGL                                0.860177
## MAJORENVA                                0.993296
## MAJORENVS                                0.996026
## MAJORESS                                 0.937301
## MAJORFNAR                                0.997251
## MAJORHIST                                0.986924
## MAJORHM                                  0.990180
## MAJORJNST                                0.998715
## MAJORMATH                                0.998788
## MAJORMS                                  0.930484
## MAJORNURS                                0.947699
## MAJORPASJ                                0.982576
## MAJORPHYS                                0.998071
## MAJORPOLS                                0.975840
## MAJORPSYC                                0.977800
## MAJORSOC                                 0.938132
## MAJORUNBN                                0.944132
## MAJORUNLA                                0.987762
## MAJORUNSC                                0.991688
## MAJOR_DESCAdvertising                          NA
## MAJOR_DESCArchitecture & Community Design      NA
## MAJOR_DESCArt History/Arts Management          NA
## MAJOR_DESCAsian Studies                        NA
## MAJOR_DESCBiology                              NA
## MAJOR_DESCBusiness Administration              NA
## MAJOR_DESCChemistry                            NA
## MAJOR_DESCCommunication Studies                NA
## MAJOR_DESCComputer Science                     NA
## MAJOR_DESCData Science                         NA
## MAJOR_DESCDesign                               NA
## MAJOR_DESCEconomics                            NA
## MAJOR_DESCEnglish                              NA
## MAJOR_DESCEntrepreneurship & Innovation        NA
## MAJOR_DESCEnvironmental Science                NA
## MAJOR_DESCEnvironmental Studies                NA
## MAJOR_DESCExercise and Sport Science           NA
## MAJOR_DESCFinance                              NA
## MAJOR_DESCFine Arts                            NA
## MAJOR_DESCHistory                              NA
## MAJOR_DESCHospitality Management               NA
## MAJOR_DESCInternational Business               NA
## MAJOR_DESCInternational Studies                NA
## MAJOR_DESCJapanese Studies                     NA
## MAJOR_DESCMarketing                            NA
## MAJOR_DESCMathematics                          NA
## MAJOR_DESCMedia Studies                        NA
## MAJOR_DESCNursing                              NA
## MAJOR_DESCOrganizational Behav.& Ldrship       NA
## MAJOR_DESCPerf. Arts & Soc. Justice            NA
```

```
## MAJOR_DESCPhysics                              NA
## MAJOR_DESCPolitics                             NA
## MAJOR_DESCPsychology                           NA
## MAJOR_DESCSociology                            NA
## MAJOR_DESCUndeclared Arts                      NA
## MAJOR_DESCUndeclared Business                  NA
## MAJOR_DESCUndeclared Sciences                  NA
## SCIENCE_CLASSES                          0.063954 .
## LAB_CLASSES                              0.093629 .
## RESIDENCE_HALLFromm                      0.877447
## RESIDENCE_HALLGillson                    0.302985
## RESIDENCE_HALLHayes Healy                0.166794
## RESIDENCE_HALLLone Mountain              0.990930
## RESIDENCE_HALLLone Village               0.996670
## RESIDENCE_HALLPacific Wing               0.999591
## RESIDENCE_HALLPhelan                     0.293955
## ETHNICITYAsian                           0.978350
## ETHNICITYHispanic or Latino              0.912193
## ETHNICITYInternational                   0.486202
## ETHNICITYMulti Race                      0.461898
## ETHNICITYNative American                 0.999627
## ETHNICITYPacific Islander                0.946785
## ETHNICITYUnknown                         0.987976
## ETHNICITYWhite                           0.696161
## GPA_CREDITS                              1.78e-05 ***
## GPA                                      0.000105 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 706.814  on 509  degrees of freedom
## Residual deviance:  49.922  on 449  degrees of freedom
## AIC: 171.92
##
## Number of Fisher Scoring iterations: 18
```

```
which(logistic.model2$coefficients>0)
```

```
##                     GENDERM                    IN_STATEY
##                           2                            3
##          UNMET_NEED_PERCENT                    MAJORANST
##                           4                            6
##                   MAJORARTM                    MAJORBACT
##                           8                            9
##                   MAJORBADM                    MAJORBAIS
##                          10                           11
##                   MAJORBENI                    MAJORBFIN
##                          12                           13
##                   MAJORBMKT                    MAJORBNTL
##                          15                           16
##                   MAJORBOBL                    MAJORCHEM
##                          17                           19
##                     MAJORCS                    MAJORDSGN
##                          21                           22
```

```
##                     MAJORECON                     MAJORENVA
##                            23                            25
##                     MAJORENVS                      MAJORESS
##                            26                            27
##                       MAJORHM                     MAJORJNST
##                            30                            31
##                     MAJORMATH                      MAJORMS
##                            32                            33
##                     MAJORPOLS                      MAJORSOC
##                            37                            39
##                     MAJORUNBN                     MAJORUNLA
##                            40                            41
##                     MAJORUNSC                  LAB_CLASSES
##                            42                            81
##  RESIDENCE_HALLLone Village  RESIDENCE_HALLPacific Wing
##                            86                            87
## ETHNICITYHispanic or Latino   ETHNICITYPacific Islander
##                            90                            94
##             ETHNICITYUnknown                  GPA_CREDITS
##                            95                            97
##                           GPA
##                            98
```

The results from the logistic model ran on the balanced data is radically different from the previous. There are 36 positive coeffecients compared to the 45 in the previous model. One noticing difference from the previous is that the second model says that if the student is male, the student will be more likely to graduate. Also, it is the same for if the student is from the State of California, which makes sense if the student is from outside California, it is more likely for a student to transfer out. Another difference is that unmet need percentage of students also contribute positively for a student to graduate, so those who have to pay more will be more likely to graduate from USF. If a student has been given Pell grants, the student is less likely to graduate from USF. Students who take more Lab Classes are more likely to graduate as well. International Studies majors are more likely to graduate than any other major. All dorms are contribute negatively except Pac-Wing and Loyola Village. GPA_CREDITS and GPA remain as before to be variables that greatly contribute positively to whether a person will graduate or not.

```r
second.loss <- mean((model2preds - y.test)^2)
second.accuracy <- 1 - second.loss
second.accuracy
```

```
## [1] 0.9411765
```

The accuracy of the mode is 94%, which is overall good since the data we have is very close to being perfectly balanced.

```r
y.test.equals1 <- y.test == 1
model2preds.equals1 <- model2preds == 1
y.test.equals0 <- y.test == 0
model2preds.equals0 <- model2preds == 0

true.positives2 <- sum(y.test.equals1 == T & model2preds.equals1==T)
false.positives2 <- sum(model2preds.equals1 == T & y.test.equals0==T)
true.negatives2 <- sum(y.test.equals0 == T &  model2preds.equals0 == T)
false.negatives2 <- sum(model2preds.equals0 == T & y.test.equals1==T)
true.positives2
```

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

```
false.positives2
```

```
## [1] 5
```

```
true.negatives2
```

```
## [1] 86
```

```
false.negatives2
```

```
## [1] 5
```

Just as before, we would like to minimize the False-Positives and the False-Negatives, but we would like to maximize the True-Positives and the False-Negatives. The following ratios will allow us to assess which model is better.

```
second.accuracy <- (true.positives2+true.negatives2)/length(model2preds)
second.sensitity <- true.positives2/(true.positives2 + false.negatives2)
second.specifity <- true.negatives2/(true.negatives2 + false.positives2)
second.precision <- true.positives2/(true.positives2 + false.positives2)

second.accuracy
```

```
## [1] 0.9411765
```

```
second.sensitity
```

```
## [1] 0.9367089
```

```
second.specifity
```

```
## [1] 0.9450549
```

```
second.precision
```

```
## [1] 0.9367089
```

The accuracy of the first model is 94%, which is really good. The sensitivity of the model is 94% which is the true positive rate. The specifity of the model is 95% which is the true negative rate. The precision of the model is 94% which is the positive predictive value.

Conclusion

Which Logistic Model is Better?

I will compare the various ratios that we calculated to measure the performance of each models.

```
first.accuracy > second.accuracy
```

```
## [1] TRUE
```

```
first.sensitity > second.sensitity
```

```
## [1] TRUE
```

```
first.specifity > second.specifity
```

```
## [1] FALSE
```

```
first.precision > second.precision
```

```
## [1] TRUE
```

This clearly states that the first model performed better than the second model which was somewhat suprising to me, excpet for the specifity of the model. The second model has higher specificity rate, which is basically the ability to rightly predict if the student would not graduate. In other words, the second model was better at guessing which students were going to not graduate. It seems that the number of observations matters more rather than the whether the data being used is balanced. My first guess was that the first model overfitted, but it returned a greater testing accuracy than the second model. For predicitive purposes, I would proceed using the first model. But for inference, I would work with the second model, especially if we had more balanced data.