101A Project EDA

```
## Loading the data

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
data <- read.csv('revised_dataset.csv')
data <- tibble(data) %>% select(acceptance_rate, name, sat_avg, hs_gpa_avg, enrollment, public_private)
attach(data)
data
```

```
## # A tibble: 216 x 6
##    acceptance_rate name                           sat_avg hs_gp~1 enrol~2 publi~3
##              <int> <chr>                            <int>   <dbl>   <int>   <dbl>
##  1              67 Seton Hall University             1070    3.6     5956       1
##  2              69 University of Vermont             1100    3.6    11159       0
##  3              66 Texas A&M University--Colleg~     1070    3.68   50735       0
##  4              71 University of Oklahoma            1060    3.6    22436       0
##  5              66 Michigan State University          990    3.7    39090       0
##  6              66 University of California--Ri~      940    3.7    19799       0
##  7              68 University of South Carolina      1120    3.69   25556       0
##  8              76 University of Utah                1050    3.6    23789       0
##  9              76 University of Cincinnati          1030    3.6    25860       0
## 10              78 University of Oregon               980    3.6    20049       0
## # ... with 206 more rows, and abbreviated variable names 1: hs_gpa_avg,
## #   2: enrollment, 3: public_private
```

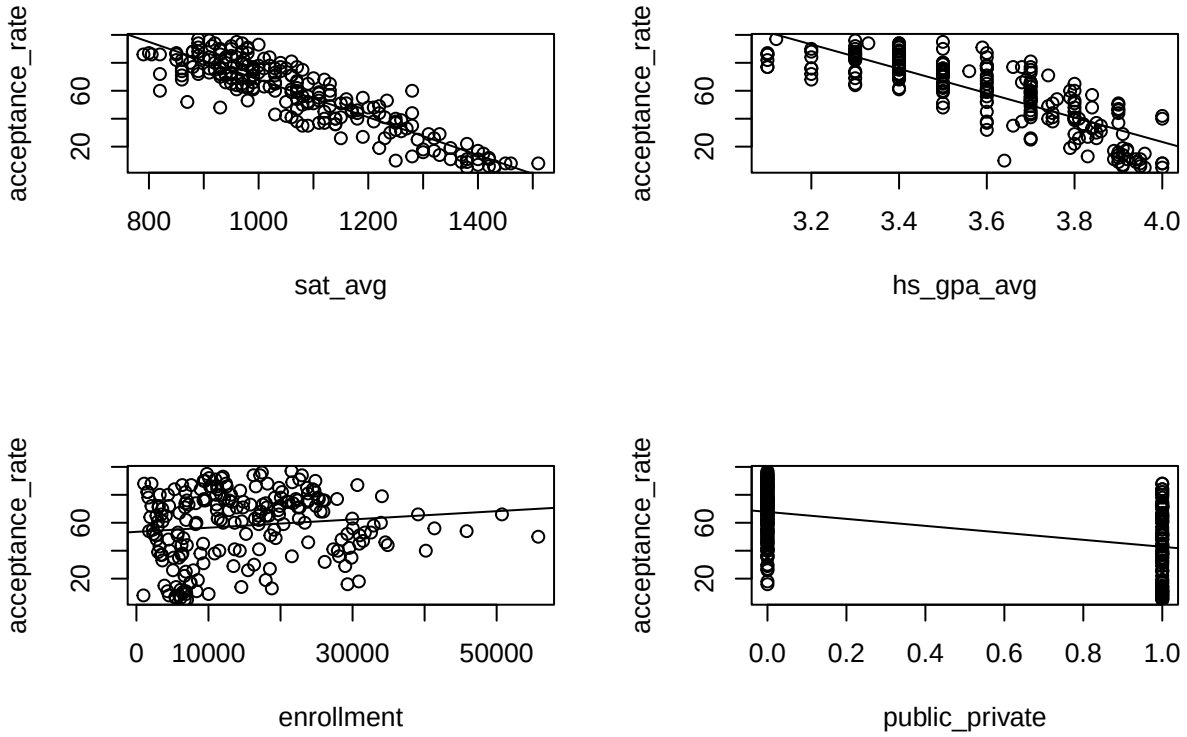
```
## Single Variable Relationships, Correlations

attach(data)
```

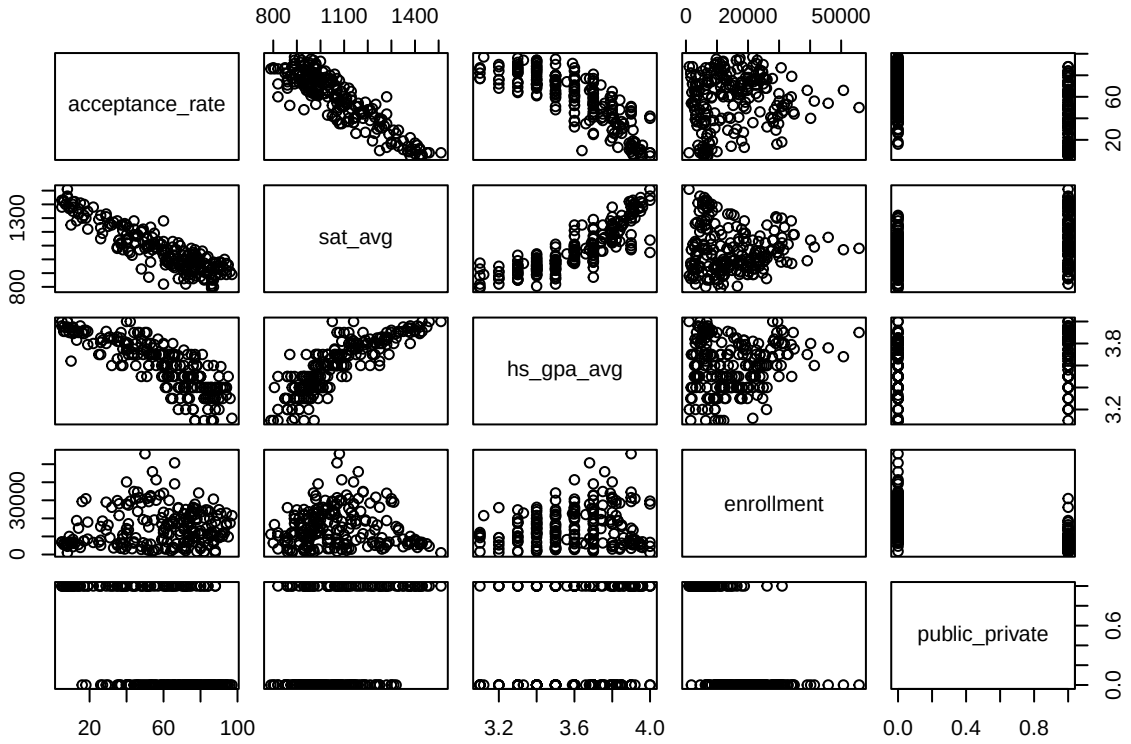
```
## The following objects are masked from data (pos = 3):
##
##     acceptance_rate, enrollment, hs_gpa_avg, name, public_private,
##     sat_avg
```

```
par(mfrow = c(2, 2))
plot(sat_avg, acceptance_rate)
abline(lm(acceptance_rate ~ sat_avg))
plot(hs_gpa_avg, acceptance_rate)
abline(lm(acceptance_rate ~ hs_gpa_avg))
plot(enrollment, acceptance_rate)
```

```
abline(lm(acceptance_rate ~ enrollment))
plot(public_private, acceptance_rate)
abline(lm(acceptance_rate ~ public_private))
```




```
pairs(data %>% select(acceptance_rate, sat_avg, hs_gpa_avg, enrollment, public_private))
```




```
cor(data %>% select(acceptance_rate, sat_avg, hs_gpa_avg, enrollment, public_private))
```

```
##                 acceptance_rate     sat_avg hs_gpa_avg  enrollment
## acceptance_rate       1.0000000 -0.89475093 -0.8125129  0.12438424
## sat_avg              -0.8947509  1.00000000  0.8434822 -0.06562565
## hs_gpa_avg           -0.8125129  0.84348217  1.0000000  0.12047118
## enrollment            0.1243842 -0.06562565  0.1204712  1.00000000
## public_private       -0.4957500  0.43636409  0.3092901 -0.63270822
##                 public_private
## acceptance_rate     -0.4957500
## sat_avg              0.4363641
## hs_gpa_avg           0.3092901
## enrollment          -0.6327082
## public_private       1.0000000
```

```
##VIF
```

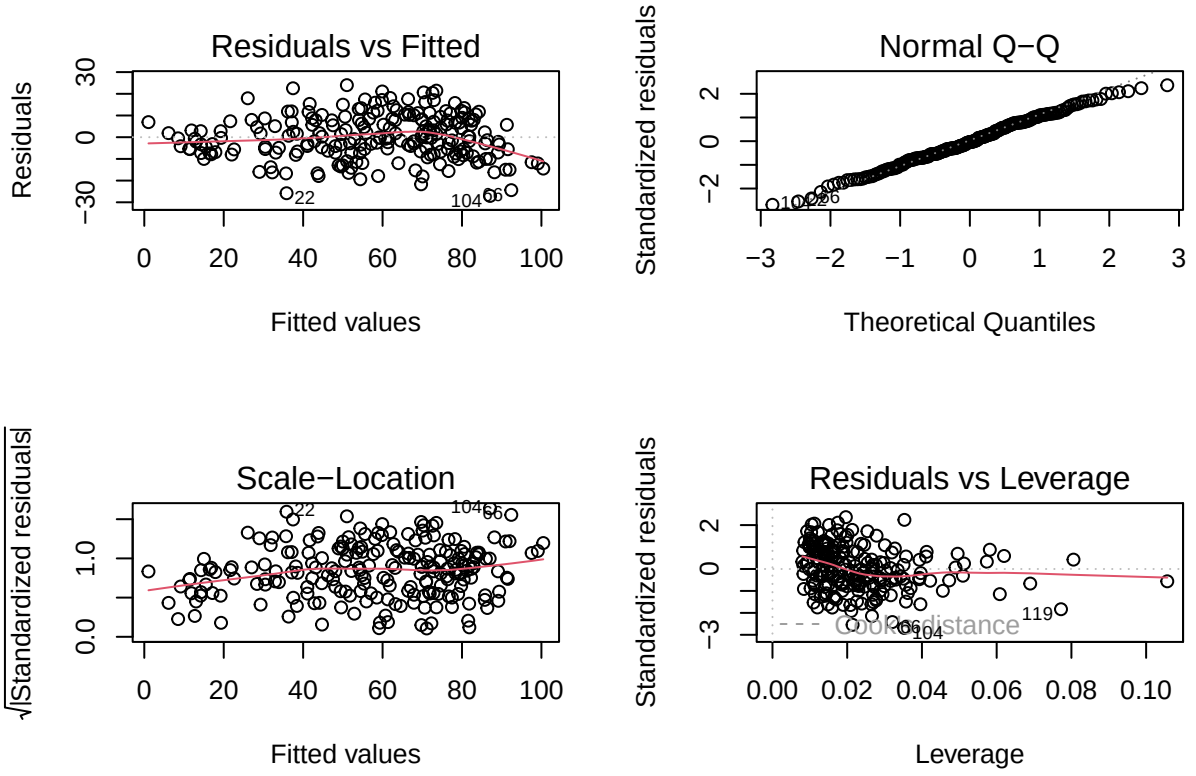
```
## Untransformed and Unreduced Model

m1 <- lm(acceptance_rate ~ sat_avg + hs_gpa_avg + enrollment + public_private)
summary(m1)
```

```
##
## Call:
## lm(formula = acceptance_rate ~ sat_avg + hs_gpa_avg + enrollment +
##     public_private)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -27.067  -7.079  -0.342   7.830  23.988
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.570e+02  1.468e+01  17.506  < 2e-16 ***
## sat_avg        -9.523e-02  8.537e-03 -11.155  < 2e-16 ***
## hs_gpa_avg     -2.653e+01  6.070e+00  -4.371 1.94e-05 ***
## enrollment      8.583e-05  9.698e-05   0.885  0.37716
## public_private -6.093e+00  2.168e+00  -2.810  0.00542 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.26 on 211 degrees of freedom
## Multiple R-squared:   0.83,  Adjusted R-squared:  0.8268
## F-statistic: 257.5 on 4 and 211 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(m1)
```




```
## Transformed Model (don't include categorical variable public/private)

library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
tranxy <- powerTransform(cbind(acceptance_rate, sat_avg, hs_gpa_avg, enrollment) ~ 1)

summary(tranxy)
```

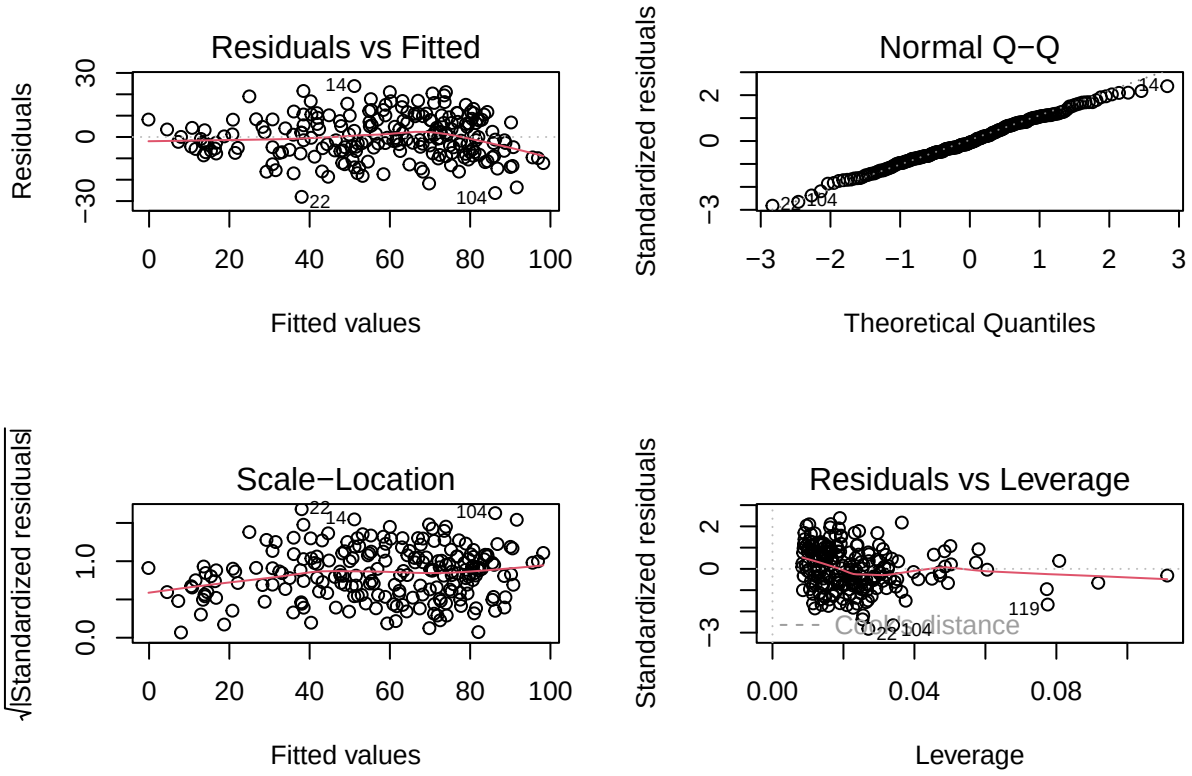
```
## bcPower Transformations to Multinormality
##                 Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## acceptance_rate    1.0017        1.00       0.8275       1.1758
## sat_avg            0.2323        0.00      -0.3938       0.8585
## hs_gpa_avg         5.1274        5.13       3.5475       6.7073
## enrollment         0.3559        0.50       0.1988       0.5131
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                   LRT df       pval
## LR test, lambda = (0 0 0 0) 249.9157  4 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                  LRT df       pval
## LR test, lambda = (1 1 1 1) 96.6896  4 < 2.22e-16
```

```
m2 <- lm(acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private + enrollment)

summary(m2)
```

```
##
## Call:
## lm(formula = acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private +
##     enrollment)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.0167  -6.7305  -0.6803   7.6850  23.8480
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.766e+02  6.143e+00  28.750  < 2e-16 ***
## sat_avg         -8.618e-02  8.916e-03  -9.666  < 2e-16 ***
## I(hs_gpa_avg^5) -3.964e-02  7.508e-03  -5.279 3.22e-07 ***
## public_private  -6.125e+00  2.121e+00  -2.888  0.00429 **
## enrollment       9.329e-05  9.404e-05   0.992  0.32232
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.07 on 211 degrees of freedom
## Multiple R-squared:  0.8362, Adjusted R-squared:  0.8331
## F-statistic: 269.3 on 4 and 211 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(m2)
```




```
## Transforming sat_avg to log(sat_avg) makes our R^2 and diagnostics a bit worse, so the only variable
```

```
## Variable Reduction

## No multicollinearity issues, since all VIFs are less than 5.
vif(m2)
```

```
##         sat_avg I(hs_gpa_avg^5)  public_private      enrollment
##        4.498801        4.432735        2.288202        2.018236
```

```
## Forward Step and Backward Elmination with AIC or BIC: Include all but enrollment (as we expected)

step(lm(acceptance_rate ~ 1), acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private + enrollment
```

```
## Start:  AIC=1388.75
## acceptance_rate ~ 1
##
##                   Df Sum of Sq    RSS    AIC
## + sat_avg          1    104547  26042 1045.9
## + I(hs_gpa_avg^5)  1     91991  38598 1130.9
## + public_private   1     32095  98495 1333.2
## <none>                         130589 1388.8
## + enrollment       1      2020 128569 1390.8
##
```

```
## Step:  AIC=1045.86
## acceptance_rate ~ sat_avg
##
##                   Df Sum of Sq    RSS    AIC
## + I(hs_gpa_avg^5)  1      2216  23826 1032.0
## + public_private   1      1789  24253 1035.9
## <none>                          26042 1045.9
## + enrollment       1       566  25477 1046.5
## - sat_avg          1    104547 130589 1388.8
##
## Step:  AIC=1032.03
## acceptance_rate ~ sat_avg + I(hs_gpa_avg^5)
##
##                   Df Sum of Sq   RSS    AIC
## + public_private   1    2338.0 21488 1015.1
## + enrollment       1    1592.5 22234 1022.5
## <none>                         23826 1032.0
## - I(hs_gpa_avg^5)  1    2215.9 26042 1045.9
## - sat_avg          1   14771.5 38598 1130.9
##
## Step:  AIC=1015.1
## acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private
##
##                   Df Sum of Sq   RSS    AIC
## <none>                         21488 1015.1
## + enrollment       1      99.8 21389 1019.5
## - public_private   1    2338.0 23826 1032.0
## - I(hs_gpa_avg^5)  1    2765.0 24253 1035.9
## - sat_avg          1    9739.2 31228 1090.5
```

```
##
## Call:
## lm(formula = acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private)
##
## Coefficients:
##     (Intercept)          sat_avg  I(hs_gpa_avg^5)   public_private
##       178.02956         -0.08701         -0.03739         -7.54023
```

```
#step(lm(acceptance_rate ~ 1), acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private + enrollmen

step(lm(acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private + enrollment), acceptance_rate ~ sa
```

```
## Start:  AIC=1002.59
## acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private +
##     enrollment
##
##                   Df Sum of Sq   RSS    AIC
## - enrollment       1      99.8 21488 1001.6
## <none>                         21389 1002.6
## - public_private   1     845.3 22234 1009.0
## - I(hs_gpa_avg^5)  1    2824.8 24213 1027.4
## - sat_avg          1    9471.8 30860 1079.8
##
## Step:  AIC=1001.6
## acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private
```

```
##
##                   Df Sum of Sq   RSS    AIC
## <none>                         21488 1001.6
## + enrollment       1      99.8 21389 1002.6
## - public_private   1    2338.0 23826 1021.9
## - I(hs_gpa_avg^5)  1    2765.0 24253 1025.7
## - sat_avg          1    9739.2 31228 1080.3
```

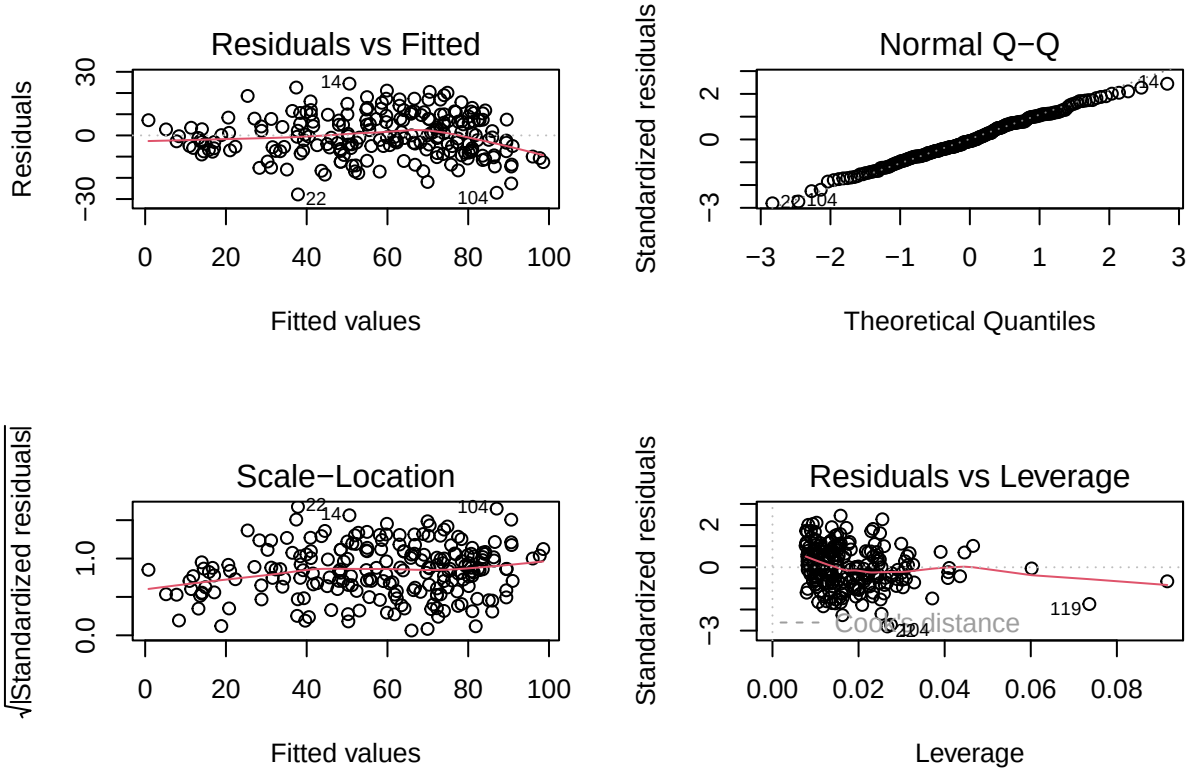
```
##
## Call:
## lm(formula = acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private)
##
## Coefficients:
##     (Intercept)          sat_avg  I(hs_gpa_avg^5)   public_private
##       178.02956         -0.08701         -0.03739         -7.54023
```

```
#step(lm(acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private + enrollment), acceptance_rate ~ .
```

```
#Final Model
final_model <- lm(acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private)
summary(final_model)
```

```
##
## Call:
## lm(formula = acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.8329  -6.9711  -0.6892   7.4151  24.4105
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     178.029556   5.972222  29.810  < 2e-16 ***
## sat_avg          -0.087011   0.008877  -9.802  < 2e-16 ***
## I(hs_gpa_avg^5)  -0.037391   0.007159  -5.223 4.20e-07 ***
## public_private   -7.540226   1.569971  -4.803 2.96e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.07 on 212 degrees of freedom
## Multiple R-squared:  0.8355, Adjusted R-squared:  0.8331
## F-statistic: 358.8 on 3 and 212 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(final_model)
```




```
## Model Assessment

## R^2 and p-values are all great, and coefficients make intuitive sense
## Diagnostic plots are pretty good - appears to be normally distributed
## Only one bad leverage point (119) which is a small private school (Gardner-Webb University) with a pi
## There is a bit of nonlinearity present that we can see on the far right of the residuals vs fitted pl
## This is likely because we don't have many colleges in our dataset that have near a 100% acceptance ra


## Interpretations:

# SAT SCORE

## For every one point increase in SAT score, the acceptance rate tends to decrease by 0.087%.
#### This makes sense because higher average test scores should be associated with a lower acceptance ra
#### Additionally, this number is very small because a one point change in an SAT score (which is out of

# HIGH SCHOOL GPA

## Difficult to interpret directly, since we transformed the variable.
#### However, the negative coefficient shows that an increase in average high school gpa also leads to a

# PUBLIC/PRIVATE

## With all else held equal, a private school tends to have a 7.54% lower acceptance rate than a public
#### This could be partially affected by the fact that our dataset has a bunch of public state schools t
```