Tuition Variation Research

Rory Quinlan

2024-01-25

Set up

```
library(readxl)
library(dplyr)
library(viridis)
require(caret)
require(MASS)

# Most complete data use as a base
Base<- read_excel("C:\\Users\\roryq\\Downloads\\Stat 1223\\CollegeDatEdited.xlsx")
BaseData<- as.data.frame(Base)


# Second data to be joined
Rank<-  read_excel("C:\\Users\\roryq\\Downloads\\Stat 1223\\collegeRank.xlsx")
RankData<- as.data.frame(Rank)

# Change column name to match other data set
colnames(RankData)[which(names(RankData) == "College Name")] <- "University"

# Third data to be joined
CityData<- read_excel("C:\\Users\\roryq\\Downloads\\Stat 1223\\schoolInfo.xlsx")

# Crime data to merge
CrimeData<- read_excel("C:\\Users\\roryq\\Downloads\\Stat 1223\\Crime.xlsx")
```

Merge 1 and 2

```
# Merge 1 with Rank and Base


M1 <- merge(x=BaseData,y=RankData,
      by.x=c("University","University"),
      by.y=c("University","University"))


# Create new column in data frame that calculates that acceptance rate for the school, and a col
umn that calculates total number of undergrads
M1$number_Undergrads<- M1$F.Undergrad+ M1$P.Undergrad
M1$Acceptance_rate<- (M1$Accept/M1$Apps)*100
```

```
# Merge 2 with City and M1
M2<- left_join(x= M1, y= CityData, by = "University")
```

```
# Export csv for Assignment checking submission and read that cvs back (requirement for assignme
nt)

    # write.csv(M2, "C:\\Users\\roryq\\Downloads\\M2.csv",           row.names=FALSE)

M2.0<- read_excel("C:\\Users\\roryq\\Downloads\\Stat 1223\\M2.xlsx")
```

```
# Remove excess/duplicate columns (by names)
M2.1 <- subset(M2.0, select = -c(P.Undergrad,F.Undergrad,Enroll,Top10perc,Top25perc,Personal,App
s,Accept,ranking,Terminal,PhD,primaryKey,Private,businessRepScore,engineeringRepScore,enrollmen
t, state,Books))


# Or remove columns (by index)
M2.2 <- M2.1[-c(22,19,6,21,20)]
```

Merge 3

```
# Merge 3;  Crime and M2.2



# Change column name to match M3 for join
colnames(CrimeData)[which(names(CrimeData) == "NMCNTY")] <- "city"
colnames(CrimeData)[which(names(CrimeData) == "Overall Rank")] <- "City Rank"



# Merge 3
M3<- inner_join(x= M2.2, y= CrimeData, by = "city")


# Remove excess/duplicate columns (by name)
M3.1 <- subset(M3, select = -c(NtnlPrkCnt,FIPS,LSTATE))


#  Or remove columns (by index)
M3.2 <- M3.1[-c(35,37:45,19,31,20,18,27,30,22,25,28,29,26,47)]
```

```
# Export csv for Assignment checking submission and read that cvs back (requirement for assignme
nt)

  #write.csv(M3.2, "C:\\Users\\roryq\\Downloads\\M3.2.csv",        row.names=FALSE)

obs_60_final<- read_excel("C:\\Users\\roryq\\Downloads\\Stat 1223\\M3.2.xlsx")

# Change column names
colnames(obs_60_final)[which(names(obs_60_final) == "2016 Crime Rate")] <- "Crime_Rate"
colnames(obs_60_final)[which(names(obs_60_final) == "2022 Median Income")] <- "Median_Income"
colnames(obs_60_final)[which(names(obs_60_final) == "Cost of Living")] <- "Cost_of_Living"
```

```
# convert crime rate to be adjusted by 1000
vec<- c(18,24,16,19,21,19,13,48,36,31,42,21,27,15,16,42,52,37,8,44,33,33,50,3,34,13,18,21,50,24,
10,21,45,29/22,26,15,11,42,24,15,36,31,15,8, 0, 42,22, 68,32,51,16,32,51,35,13,51,19,27, NA,NA)

vec2<-numeric()
for(i in vec){
  vec2<- c(vec2,i/1000)
}

obs_60_final$Crime.Rate<-vec2
```

```
# Display number of observations and column names for final dataset to model
colnames(obs_60_final)
```

```
##  [1] "University"           "city"                 "Rank"
##  [4] "institutionalControl" "Tuition"              "S.F.Ratio"
##  [7] "perc.alumni"          "Expend"               "Grad.Rate"
## [10] "number_Undergrads"    "Acceptance_rate"      "Crime_Rate"
## [13] "Unemployment"         "Cost_of_Living"       "Median_Income"
## [16] "AVG_C_two_I"          "1p1c"                 "Diversity_Rank_Race"
## [19] "Crime.Rate"
```

```
nrow(obs_60_final)
```

```
## [1] 60
```

Descriptive Statistics

```
# Create Descriptive Stats


# Filter by private or public schools
Private_60 = obs_60_final[which(obs_60_final$institutionalControl == "private"),]
Public_60 = obs_60_final[which(obs_60_final$institutionalControl == "public"),]


labels<- c("Mean Cost Private",
"Mean Cost Public",
"Mean Expend Public",
"Mean Expend Private",
"sd Cost Public",
"sd Cost Public",
"sd Cost Private",
"sd Expend Private",
"sd Expend Public",
"sd income Public",
"Mean income Public",
"sd income Private",
"Mean income Private")

values<-c(mean(Private_60$Tuition)
,mean(Public_60$Tuition)
,mean(Public_60$Expend)
,mean(Private_60$Expend)
,sd(Public_60$Tuition)
,sd(Public_60$Tuition)
,sd(Private_60$Tuition)
,sd(Private_60$Expend)
,sd(Public_60$Expend)
,sd(Public_60$Median_Income)
,mean(Public_60$Median_Income)
,sd(Private_60$Median_Income)
,mean(Private_60$Median_Income))


# View descriptive stats as a data frame
as.data.frame(cbind(labels,values))
```

```
##                 labels           values
## 1    Mean Cost Private 47879.7567567568
## 2     Mean Cost Public 26162.9130434783
## 3   Mean Expend Public 7652.78260869565
## 4  Mean Expend Private 15012.9189189189
## 5       sd Cost Public 8922.40503101879
## 6       sd Cost Public 8922.40503101879
## 7      sd Cost Private 12793.9760052347
## 8    sd Expend Private 10726.2550350333
## 9     sd Expend Public 2150.88890623308
## 10    sd income Public 28239.0114816075
## 11  Mean income Public 83063.5756521739
## 12   sd income Private 24296.9830063389
## 13 Mean income Private 85944.4245945946
```

```
# Create Descriptive Stats

values2<-
c(sd(as.numeric(Public_60$number_Undergrads))
,mean(as.numeric(Public_60$number_Undergrads))
,sd(as.numeric(Private_60$number_Undergrads))
,mean(as.numeric(Private_60$number_Undergrads)))


labels2<-
c("sd undergrad Public",
"Mean undergrad Public",
"sd undergrad Private",
"Mean undergrad Private")
```

```
# View descriptive stats as a data frame
as.data.frame(cbind(labels,values))
```

```
##                 labels           values
## 1    Mean Cost Private 47879.7567567568
## 2     Mean Cost Public 26162.9130434783
## 3   Mean Expend Public 7652.78260869565
## 4  Mean Expend Private 15012.9189189189
## 5       sd Cost Public 8922.40503101879
## 6       sd Cost Public 8922.40503101879
## 7      sd Cost Private 12793.9760052347
## 8    sd Expend Private 10726.2550350333
## 9     sd Expend Public 2150.88890623308
## 10    sd income Public 28239.0114816075
## 11  Mean income Public 83063.5756521739
## 12   sd income Private 24296.9830063389
## 13 Mean income Private 85944.4245945946
```

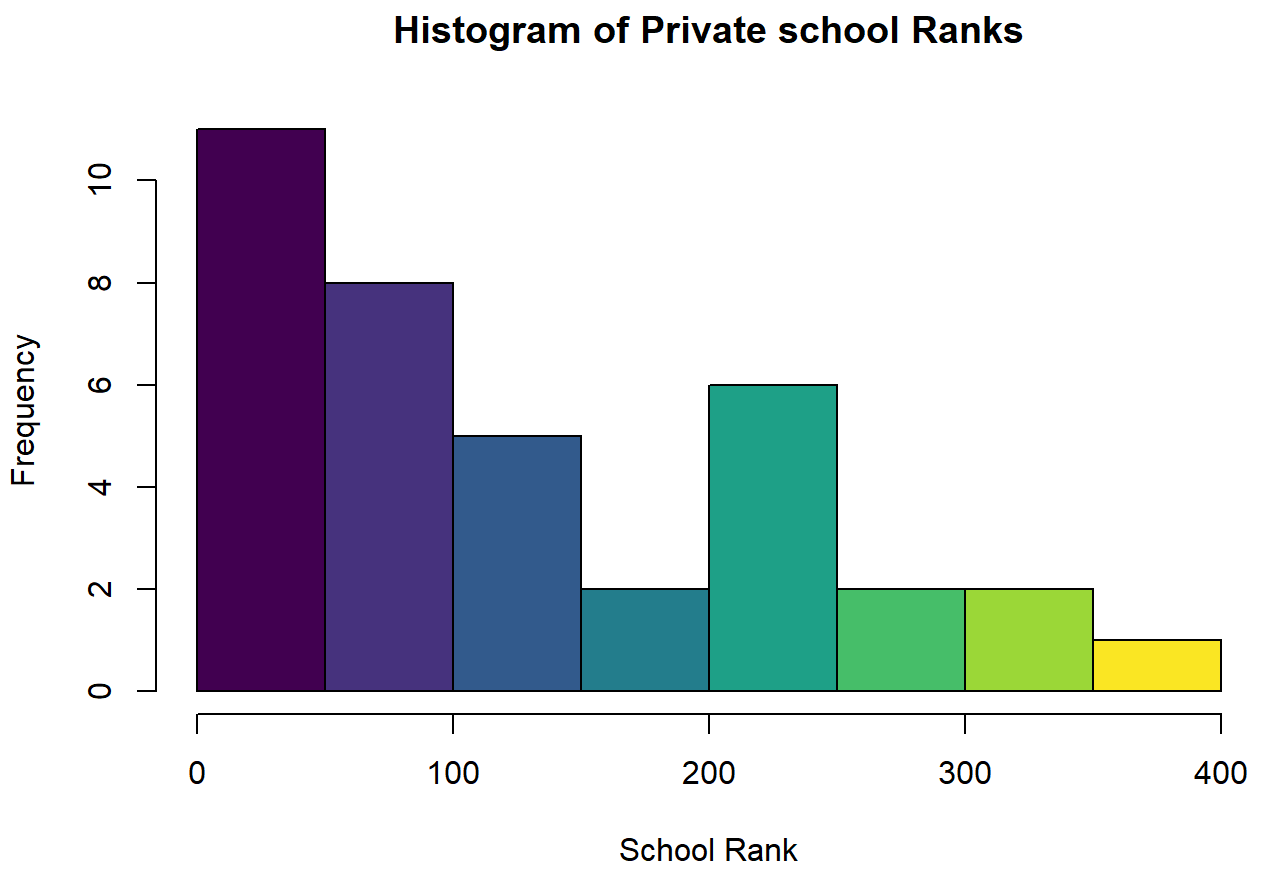
```
# Number of observations of private schools
nrow(Private_60)
```

```
## [1] 37
```

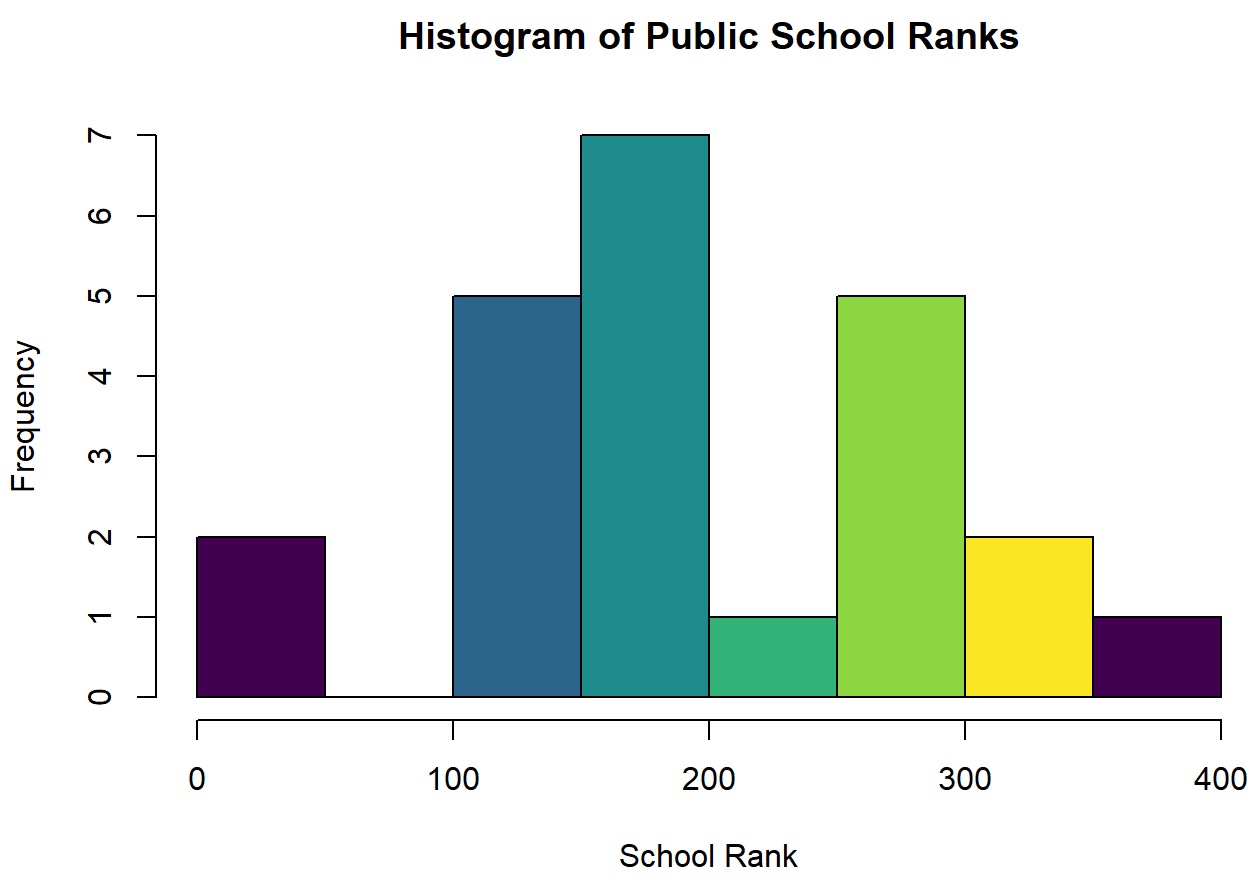
```
# Number of observations of public schools
nrow(Public_60)
```

```
## [1] 23
```

```
hist(Private_60$Rank, xlab= "School Rank", main= "Histogram of Private school Ranks", col= viridis(8))
```




```
hist(Public_60$Rank, xlab= "School Rank", main= "Histogram of Public School Ranks" , col= viridis(7) )
```

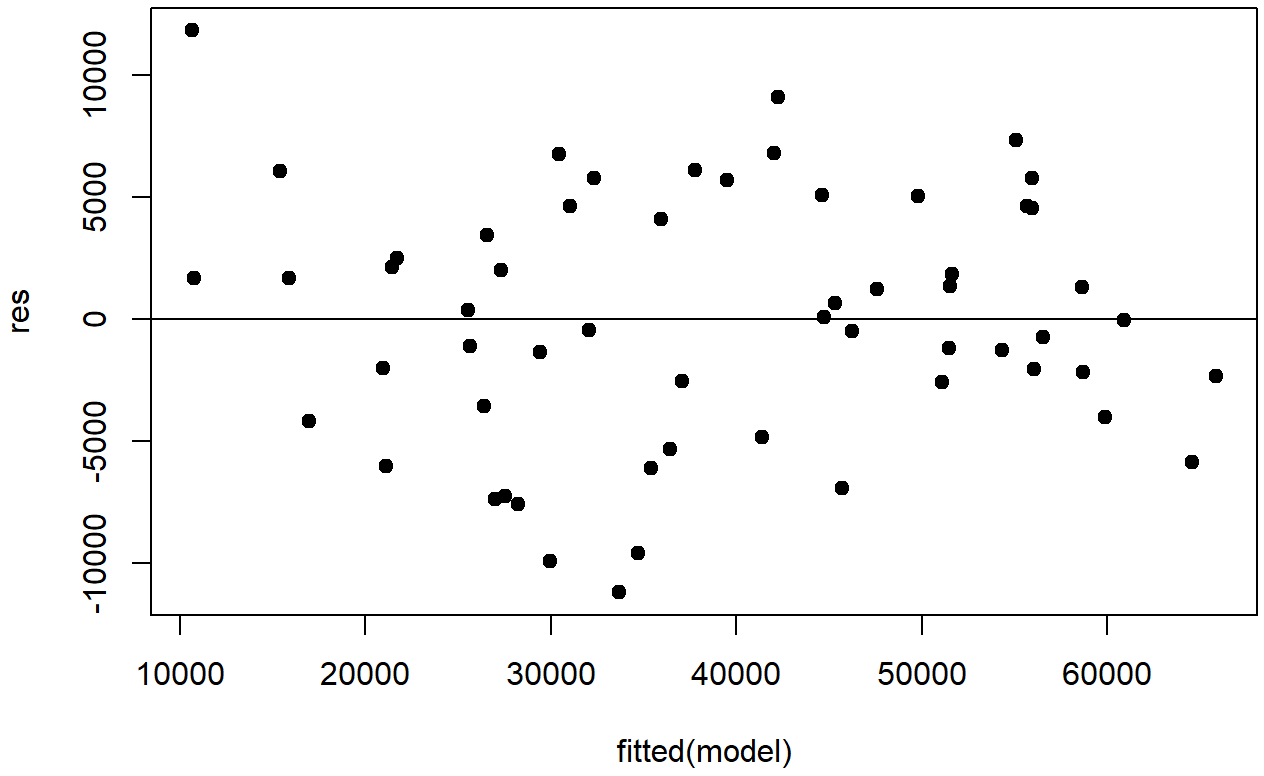


Model (Data Set Complete, 60 Observations)

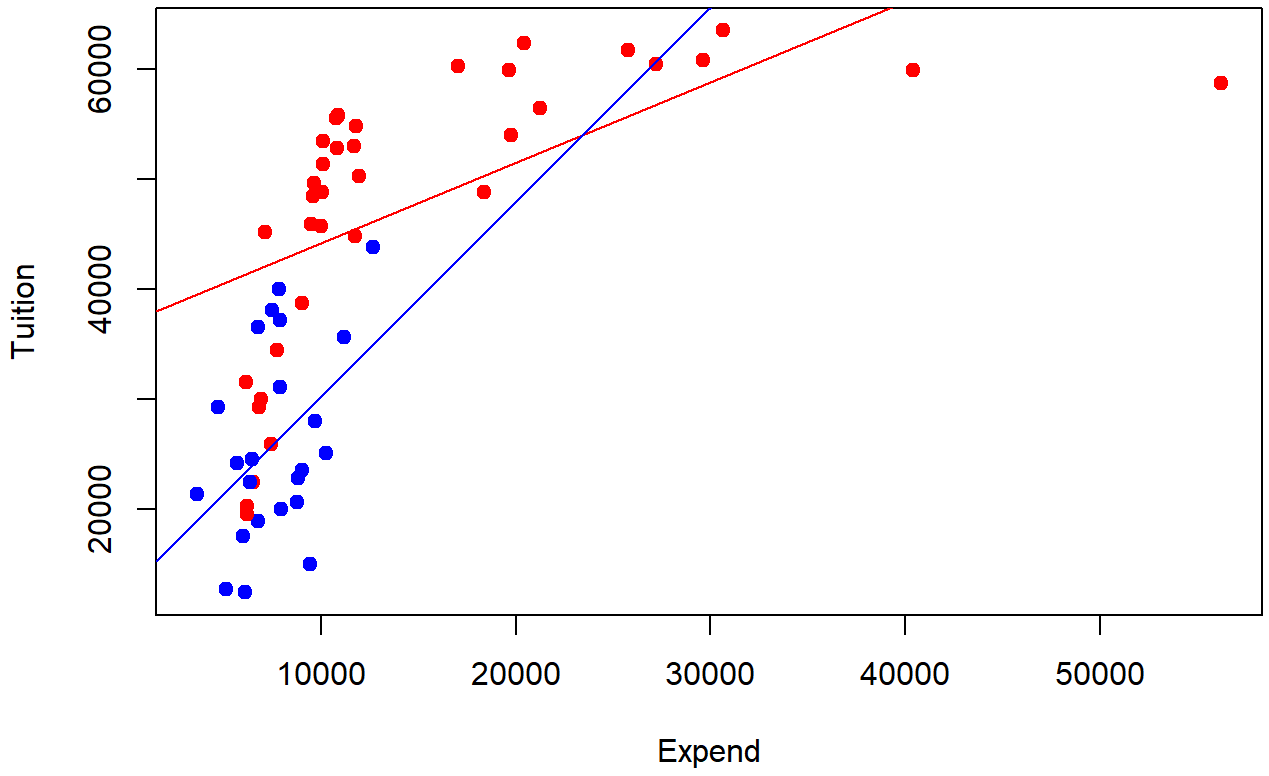
```
# Create linear regression model with all factors we are interested
 model = lm(Tuition ~ Rank+S.F.Ratio+Unemployment+Diversity_Rank_Race+ Expend+perc.alumni +insti
tutionalControl+number_Undergrads+Median_Income+Grad.Rate+ Crime.Rate+Cost_of_Living+AVG_C_two_I
, data = obs_60_final)

# Print model summary
summary(model)
```

```
##
## Call:
## lm(formula = Tuition ~ Rank + S.F.Ratio + Unemployment + Diversity_Rank_Race +
##     Expend + perc.alumni + institutionalControl + number_Undergrads +
##     Median_Income + Grad.Rate + Crime.Rate + Cost_of_Living +
##     AVG_C_two_I, data = obs_60_final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11192.0  -3320.3     23.6   4458.7  11842.6
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.877e+04  1.557e+04   2.490 0.016628 *
## Rank                       -9.464e+01  1.397e+01  -6.773 2.47e-08 ***
## S.F.Ratio                   7.649e+01  2.988e+02   0.256 0.799158
## Unemployment                1.618e+05  9.703e+04   1.668 0.102447
## Diversity_Rank_Race        -1.214e+00  1.201e+00  -1.011 0.317495
## Expend                      1.945e-01  1.577e-01   1.234 0.223908
## perc.alumni                -9.637e+01  1.050e+02  -0.918 0.363653
## institutionalControlpublic -1.136e+04  2.751e+03  -4.128 0.000161 ***
## number_Undergrads          -3.797e-01  2.012e-01  -1.888 0.065679 .
## Median_Income               2.769e-01  1.392e-01   1.989 0.052977 .
## Grad.Rate                  -2.934e+01  7.147e+01  -0.411 0.683419
## Crime.Rate                 -4.815e+04  5.943e+04  -0.810 0.422237
## Cost_of_Living             -1.480e-01  1.459e-01  -1.014 0.315949
## AVG_C_two_I                 1.075e+04  1.269e+04   0.847 0.401521
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5816 on 44 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.8909, Adjusted R-squared:  0.8586
## F-statistic: 27.63 on 13 and 44 DF,  p-value: < 2.2e-16
```

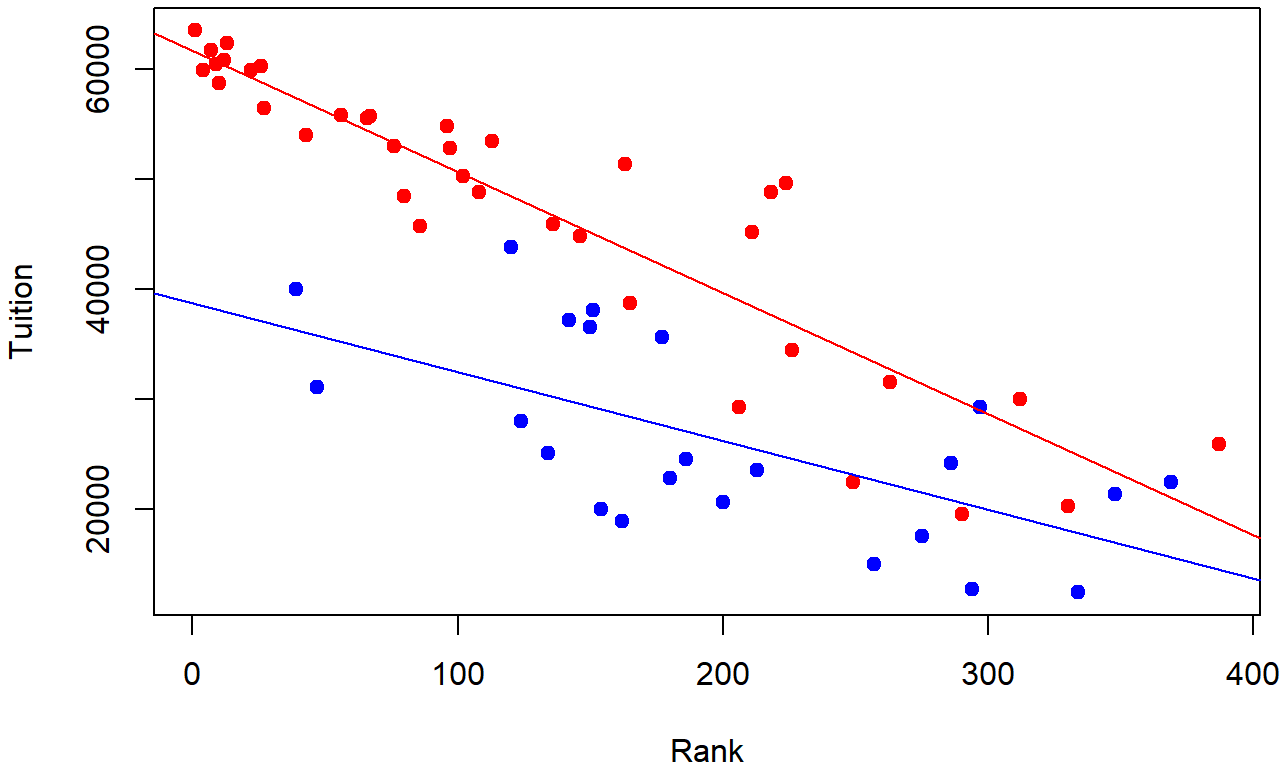
```
# View residuals to check heteroskewdasticity
res=resid(model)
plot(fitted(model), res, pch=19)
abline(0,0)
```




```
# Check for interaction terms



# Create 2 linear regression models one with private and one with public to compare expenditure
per student and tuition levels
model_pri60E = lm(Tuition ~ Expend, data = Private_60)
model_pub60E = lm(Tuition ~ Expend, data = Public_60)

# Scatterplot with groups
# Specify colors to be used in scatterplot

colors = c("red", "blue")
plot(obs_60_final$Expend, obs_60_final$Tuition, pch = 19, col = colors[factor(obs_60_final$insti
tutionalControl)], xlab = "Expend", ylab = "Tuition")

abline(model_pri60E, col = "red")  # Plot the regression line for private colleges

abline(model_pub60E, col = "blue")   # Plot the regression line for public colleges
```




```
# Create 2 linear regression models one with private and one with public to compare expenditure
per student and tuition levels
model_pri60 = lm(Tuition ~ Rank, data = Private_60)
model_pub60 = lm(Tuition ~ Rank, data = Public_60)

# Scatterplot with groups
colors = c("red", "blue")  # Specify colors to be used in scatterplot
plot(obs_60_final$Rank, obs_60_final$Tuition, pch = 19, col = colors[factor(obs_60_final$institu
tionalControl)], xlab = "Rank", ylab = "Tuition")
abline(model_pri60, col = "red")  # Plot the regression line for private colleges
abline(model_pub60, col = "blue")   # Plot the regression line for public colleges
```



Model Selection

```
# Model selection
# Use Forward and Backward Stepwise Regression Selection

min_model = lm(Tuition ~ 1, data = obs_60_final)
max_model = formula(lm(Tuition ~ Rank + S.F.Ratio + Unemployment + Diversity_Rank_Race + Expend+
institutionalControl+number_Undergrads+Median_Income+Grad.Rate+Crime.Rate+Cost_of_Living, data =
obs_60_final))
best_model = step(min_model, direction = "both", scope = max_model)
```

```
## Start:  AIC=1159.49
## Tuition ~ 1
```

```
## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 58/60 rows from a combined fit
```

```
##                         Df  Sum of Sq        RSS    AIC
## + Rank                   1 8786615224 4.8526e+09 1062.1
## + Grad.Rate              1 6543660673 7.0956e+09 1084.1
## + institutionalControl   1 6210508037 7.4287e+09 1086.8
## + S.F.Ratio              1 5841327150 7.7979e+09 1089.6
## + Expend                 1 5730376759 7.9089e+09 1090.4
## + Cost_of_Living         1 2080042695 1.1559e+10 1112.4
## + number_Undergrads      1 2048877599 1.1590e+10 1112.5
## + Unemployment           1 1265208252 1.2374e+10 1116.3
## + Diversity_Rank_Race    1 1053308676 1.2586e+10 1117.3
## + Median_Income          1  896111172 1.2743e+10 1118.0
## <none>                                1.3639e+10 1120.0
## + Crime.Rate             1   55039386 1.3584e+10 1121.8
##
## Step:  AIC=1096.98
## Tuition ~ Rank
```

```
## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 58/60 rows from a combined fit
```

```
##                         Df  Sum of Sq        RSS    AIC
## + institutionalControl   1 2566878018 2.2858e+09 1020.4
## + number_Undergrads      1 1993859369 2.8588e+09 1033.4
## + S.F.Ratio              1  599239985 4.2534e+09 1056.4
## + Cost_of_Living         1  516535540 4.3361e+09 1057.5
## + Grad.Rate              1  499040782 4.3536e+09 1057.8
## + Median_Income          1  442716657 4.4099e+09 1058.5
## + Expend                 1  380271112 4.4724e+09 1059.3
## + Diversity_Rank_Race    1  160368441 4.6923e+09 1062.1
## + Unemployment           1    2279317 4.8504e+09 1064.0
## + Crime.Rate             1     852788 4.8518e+09 1064.0
## <none>                                4.8908e+09 1097.0
## - Rank                   1 9442427717 1.4333e+10 1159.5
##
## Step:  AIC=1053.45
## Tuition ~ Rank + institutionalControl
```

```
## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 58/60 rows from a combined fit
```

```
##                         Df  Sum of Sq        RSS    AIC
## + Median_Income          1  415807153 1869957307 1010.8
## + Cost_of_Living         1  242401872 2043362587 1015.9
## + number_Undergrads      1   81995697 2203768762 1020.3
## + Expend                 1   55795938 2229968521 1021.0
## + Crime.Rate             1   18813816 2266950643 1021.9
## + Diversity_Rank_Race    1   13692198 2272072262 1022.0
## + Grad.Rate              1   11002050 2274762410 1022.1
## + S.F.Ratio              1    3866790 2281897670 1022.3
## + Unemployment           1     580703 2285183756 1022.4
## <none>                               2290077189 1053.5
## - institutionalControl  1 2600751657 4890828847 1097.0
## - Rank                  1 5354017257 7644094447 1123.8
##
## Step:  AIC=1043.31
## Tuition ~ Rank + institutionalControl + Median_Income
```

```
## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 58/60 rows from a combined fit
```

```
##                         Df  Sum of Sq        RSS    AIC
## + Diversity_Rank_Race    1   88696636 1781260671 1009.9
## + number_Undergrads      1   80471866 1789485441 1010.2
## + Unemployment           1   62708007 1807249300 1010.8
## + Expend                 1   36644244 1833313063 1011.6
## + Crime.Rate             1   33689900 1836267407 1011.7
## + S.F.Ratio              1   26231231 1843726076 1011.9
## + Grad.Rate              1    4266839 1865690468 1012.6
## + Cost_of_Living         1    1584200 1868373106 1012.7
## <none>                               1870561072 1043.3
## - Median_Income          1  419516118 2290077189 1053.5
## - institutionalControl  1 2562373052 4432934123 1093.1
## - Rank                  1 5021704222 6892265293 1119.6
##
## Step:  AIC=1042.43
## Tuition ~ Rank + institutionalControl + Median_Income + Diversity_Rank_Race
```

```
## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 58/60 rows from a combined fit
```

```
##                         Df  Sum of Sq        RSS    AIC
## + Unemployment           1   90042221 1691218450 1008.9
## + number_Undergrads      1   68996034 1712264637 1009.6
## + Expend                 1   32649244 1748611427 1010.9
## + S.F.Ratio              1   16625754 1764634917 1011.4
## + Crime.Rate             1   11781928 1769478743 1011.5
## + Grad.Rate              1      59081 1781201590 1011.9
## + Cost_of_Living         1      30303 1781230368 1011.9
## <none>                                1782767956 1042.4
## - Diversity_Rank_Race    1   87793115 1870561072 1043.3
## - Median_Income          1  492521434 2275289390 1055.1
## - institutionalControl   1 2647980574 4430748530 1095.0
## - Rank                   1 5109484417 6892252373 1121.6
##
## Step:  AIC=1041.29
## Tuition ~ Rank + institutionalControl + Median_Income + Diversity_Rank_Race +
##     Unemployment
```

```
## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 58/60 rows from a combined fit
```

```
##                         Df  Sum of Sq        RSS    AIC
## + number_Undergrads      1   96771778 1594446672 1007.5
## + Expend                 1   49741762 1641476688 1009.2
## + S.F.Ratio              1   33380839 1657837611 1009.8
## + Cost_of_Living         1    9440520 1681777929 1010.6
## + Crime.Rate             1    3585586 1687632864 1010.8
## + Grad.Rate              1      10108 1691208342 1010.9
## <none>                                1691963378 1041.3
## - Unemployment           1   90804578 1782767956 1042.4
## - Diversity_Rank_Race    1  115638270 1807601648 1043.3
## - Median_Income          1  583214309 2275177687 1057.1
## - institutionalControl   1 2622018821 4313982199 1095.5
## - Rank                   1 3874111872 5566075250 1110.7
##
## Step:  AIC=1039.75
## Tuition ~ Rank + institutionalControl + Median_Income + Diversity_Rank_Race +
##     Unemployment + number_Undergrads
```

```
## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, : using
## the 58/60 rows from a combined fit
```

```
##                         Df  Sum of Sq        RSS    AIC
## + Expend                 1   33830118 1560616555 1008.3
## + S.F.Ratio              1    8250107 1586196566 1009.2
## + Grad.Rate              1    6958332 1587488340 1009.2
## + Crime.Rate             1    6263077 1588183595 1009.3
## + Cost_of_Living         1    5024640 1589422032 1009.3
## <none>                                1594947456 1039.8
## - number_Undergrads      1   97015922 1691963378 1041.3
## - Diversity_Rank_Race    1  106456269 1701403725 1041.6
## - Unemployment           1  117399840 1712347296 1042.0
## - Median_Income          1  590406576 2185354033 1056.6
## - institutionalControl   1  632111004 2227058460 1057.8
## - Rank                   1 3918980495 5513927952 1112.2
##
## Step:  AIC=1040.76
## Tuition ~ Rank + institutionalControl + Median_Income + Diversity_Rank_Race +
##     Unemployment + number_Undergrads + Expend
```

Best Model and model summary below

```
# View best model
best_model
```

```
##
## Call:
## lm(formula = Tuition ~ Rank + institutionalControl + Median_Income +
##     Diversity_Rank_Race + Unemployment + number_Undergrads +
##     Expend, data = obs_60_final)
##
## Coefficients:
##                 (Intercept)                         Rank
##                   4.303e+04                   -8.751e+01
## institutionalControlpublic               Median_Income
##                  -1.129e+04                    1.639e-01
##         Diversity_Rank_Race                 Unemployment
##                  -1.909e+00                    1.731e+05
##           number_Undergrads                       Expend
##                  -2.744e-01                    9.929e-02
```

```
# print summary of best model
# Table 2 mentioned in the Inferential Statistics section of readme file
summary(best_model)
```

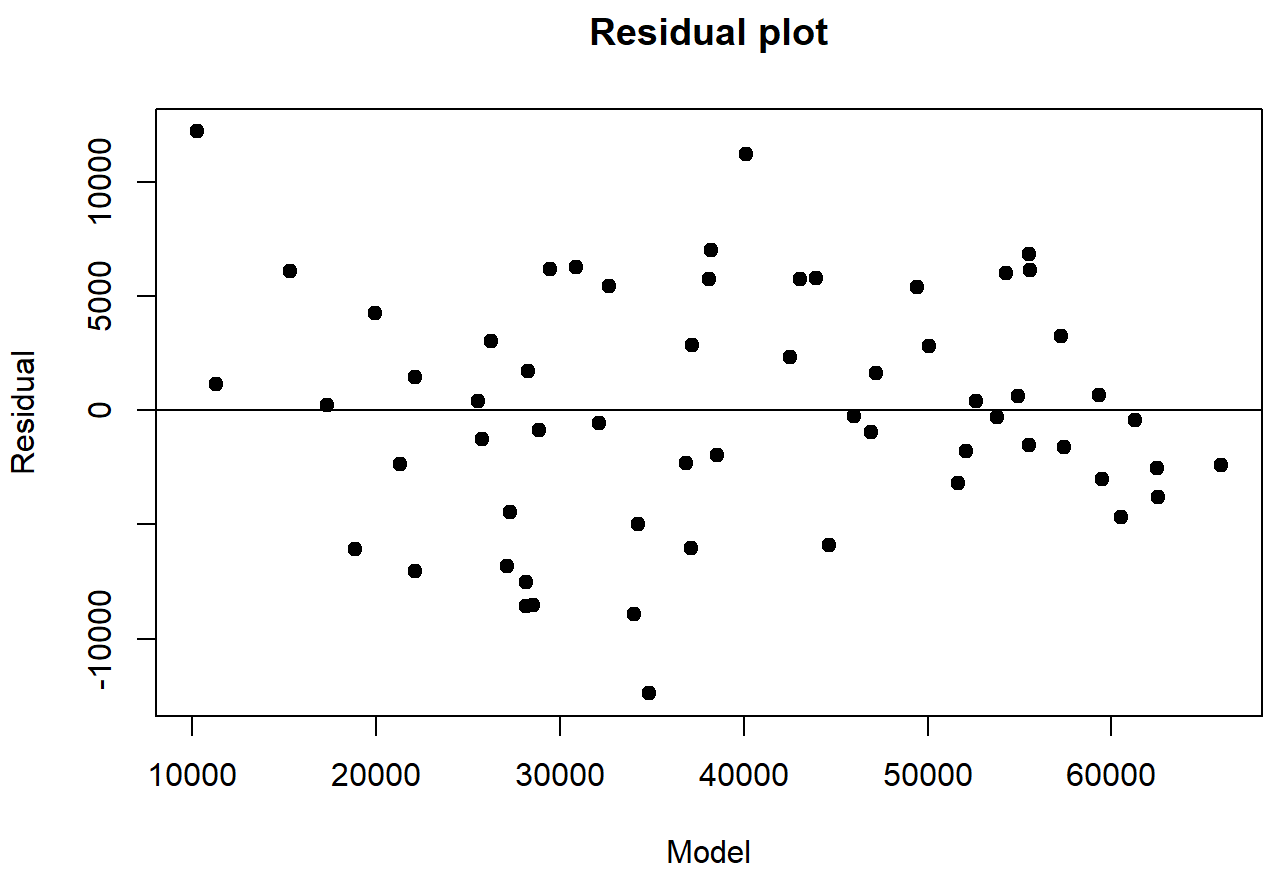
```
##
## Call:
## lm(formula = Tuition ~ Rank + institutionalControl + Median_Income +
##     Diversity_Rank_Race + Unemployment + number_Undergrads +
##     Expend, data = obs_60_final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12394.8  -3047.2   -265.4   3491.5  12214.3
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.303e+04  6.028e+03   7.139 2.97e-09 ***
## Rank                        -8.751e+01  1.017e+01  -8.602 1.44e-11 ***
## institutionalControlpublic -1.129e+04  2.497e+03  -4.522 3.58e-05 ***
## Median_Income                1.639e-01  3.720e-02   4.406 5.29e-05 ***
## Diversity_Rank_Race         -1.909e+00  1.017e+00  -1.878   0.0660 .
## Unemployment                 1.731e+05  8.446e+04   2.049   0.0455 *
## number_Undergrads           -2.744e-01  1.656e-01  -1.656   0.1036
## Expend                       9.929e-02  1.071e-01   0.927   0.3583
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5493 on 52 degrees of freedom
## Multiple R-squared:  0.8905, Adjusted R-squared:  0.8758
## F-statistic: 60.43 on 7 and 52 DF,  p-value: < 2.2e-16
```

Model Validation

```
# Model validation
# Use Leave One Our Cross Validation
# Output is table 3 referenced in the inferential statistics section in the readme file

ctrl = trainControl(method = "LOOCV")
model1 = train(Tuition ~ Rank + institutionalControl + Median_Income +
Diversity_Rank_Race + Unemployment + number_Undergrads + Expend, data = obs_60_final, method =
"lm", trControl = ctrl)
model1$results
```

```
##   intercept     RMSE  Rsquared      MAE
## 1      TRUE 5988.835 0.8513063 4797.676
```

```
# Plot residuals and check heteroskewdasticity
res=resid(best_model)
plot(fitted(best_model), res, pch=19, main="Residual plot", ylab="Residual", xlab="Model")
abline(0,0)
```



Model With only Rank

```
summary(lm(Tuition ~ Rank, data = obs_60_final))
```

```
##
## Call:
## lm(formula = Tuition ~ Rank, data = obs_60_final)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21226  -4591   2242   5438  18368
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 57924.08    2102.08   27.56  < 2e-16 ***
## Rank         -118.69      11.22  -10.58 3.66e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9183 on 58 degrees of freedom
## Multiple R-squared:  0.6588, Adjusted R-squared:  0.6529
## F-statistic:   112 on 1 and 58 DF,  p-value: 3.659e-15
```