Education Across the United States of America

Pavel Olariu po3787

INTRODUCTION

```
library(tidyverse)
library(mosaicData)
library(carData)
#install packages that allow library use and opened datasets for viewing and situating
```

```
states1992 <- States[-c(9),]
#deletes row describing Washington D.C. as this is not really a state and not included in the other dataset.
states1992renamed <- states1992%>%
  rename(
    verbal1992 = SATV,
    math1992 = SATM,
    percent1992 = percent,
    pupilexpense1992 = dollars,
    salary1992 = pay
  )
#renamed variables for first set to represent the associated year in the variable name for easier comparison
states1995renamed <- SAT%>%
  rename(
    pupilexpense1995 = expend,
    salary1995 = salary,
    percent1995 = frac,
    verbal1995 = verbal,
    math1995 = math,
    sat1995 = sat
  )
#renamed variables for other dataset
states1992renamedv2 <- states1992renamed%>%
  mutate(sat1992 = math1992 + verbal1992)
#created a new variable for total sat score by adding verbal and math scores together
states1992renamedv3 <- states1992renamedv2%>%
  mutate(state = states1995renamed$state)
#added a variable of states in 1992 set because the state were being used as the row names instead of its variable, had
 to do it for easier joining.
combinedstates <- states1995renamed%>%
  full_join(states1992renamedv3, by = "state")
#joined both sets using full join to ensure all columns and rows were kept, joined with the key variable designated as
 'states'
combinedstates <- combinedstates%>%
  mutate(averagesat = (sat1992 + sat1995)/2)%>%
  mutate(averagepupilexpense = (pupilexpense1992 + pupilexpense1995)/2)%>%
  mutate(averagesalary = (salary1992 + salary1995)/2)%>%
  mutate(averagepercent = (percent1992 + percent1995)/2)%>%
  mutate(averagesatverbal = (verbal1992 + verbal1995)/2)%>%
  mutate(averagesatmath = (math1992 + math1995)/2)
#created averages for all the variables that were common across both years
levels <- levels(combinedstates$region)
levels[length(levels) + 1] <- "West"
combinedstates$region <- factor(combinedstates$region, levels = levels)
levels <- levels(combinedstates$region)
levels[length(levels) + 1] <- "Central"
combinedstates$region <- factor(combinedstates$region, levels = levels)
levels <- levels(combinedstates$region)
levels[length(levels) + 1] <- "East Central"
combinedstates$region <- factor(combinedstates$region, levels = levels)
levels <- levels(combinedstates$region)
levels[length(levels) + 1] <- "East"
combinedstates$region <- factor(combinedstates$region, levels = levels)
#created new factors in order to put regions into only four new categories

combinedstates <- combinedstates%>%
  mutate(region=replace(region, region=="MTN", "West"))%>%
  as.data.frame()
combinedstates <- combinedstates%>%
  mutate(region=replace(region, region=="PAC", "West"))%>%
  as.data.frame()
combinedstates <- combinedstates%>%
  mutate(region=replace(region, region=="WNC", "Central"))%>%
  as.data.frame()
combinedstates <- combinedstates%>%
  mutate(region=replace(region, region=="WSC", "Central"))%>%
  as.data.frame()
```

```
combinedstates <- combinedstates%>%
  mutate(region=replace(region, region=="ESC", "East Central"))%>%
  as.data.frame()
combinedstates <- combinedstates%>%
  mutate(region=replace(region, region=="ENC", "East Central"))%>%
  as.data.frame()
combinedstates <- combinedstates%>%
  mutate(region=replace(region, region=="MA", "East"))%>%
  as.data.frame()
combinedstates <- combinedstates%>%
  mutate(region=replace(region, region=="NE", "East"))%>%
  as.data.frame()
combinedstates <- combinedstates%>%
  mutate(region=replace(region, region=="SA", "East"))%>%
  as.data.frame()
#changed the regions to only have four categories, mostly just consolidated two regions that are neighbors.
view(combinedstates)
```

In order to obtain the dataset I will be using for further ananlyses, I took two pre-existing datasets from the R packages “mosaicData” and “carData”. These datasets deal with education statistics across each individual US state (50 observations) from two different years (1992,1995). In order to tidy up and combine the datasets, I renamed shared variables to reflect the original year they were recorded, as well as creating an average of these two values from different years to get one common value for each shared variable. I also changed the 9 existing categories of regions into only four in order to simplify analyses and ease interpretations. The main variables in this dataset are the region of the state (East, East Central, Central, West), average SAT score between the two years (averagesat), the population in thousands (pop), average teacher’s salary in thousands (averagesalary), average amount of moeny spent on each individual student in thousands (averagepupilexpense), and the average percent between the two years of graduating seniors who had taken the SAT (averagepercent). I am interested in seeing how different states and regions across the United States compare in terms of education, and what factors can help predict the success and increase the SAT scores of students. I would expect to see that increased funding in the form of higher teacher’s salaries and pupil expenses would lead to increased SAT scores, as well as lower pupil to teacher ratios.

EDA

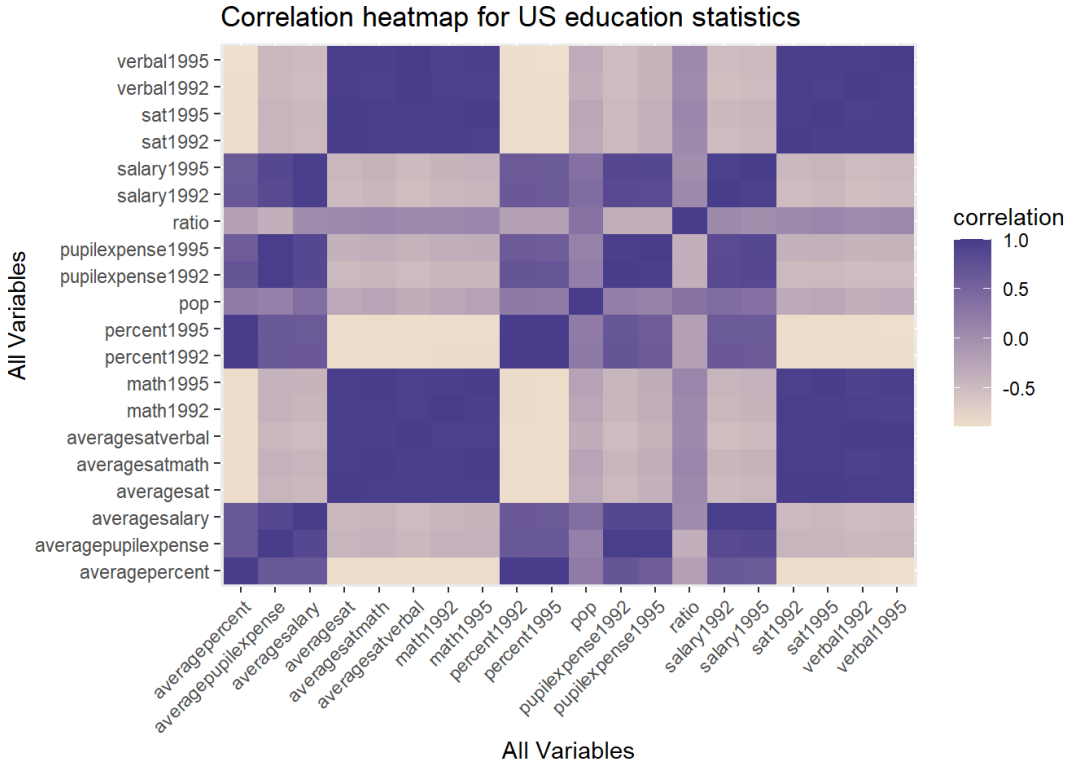
```
library(kableExtra)
#a useful package that can help with tables

combinedstates %>%
  group_by(region)%>%
  summarise(mean_ratio = mean(ratio),mean_averagesat = mean(averagesat),mean_averagepupilexpense = mean(averagepupilexpe
nse),mean_averagesalary = mean(averagesalary),mean_averagepercent = mean(averagepercent), mean_averagesat = mean(average
sat), mean_pop = mean(pop))%>%
  kbl%>%
  kable_styling()
```

region	mean_ratio	mean_averagesat	mean_averagepupilexpense	mean_averagesalary	mean_averagepercent	mean_pop
West	19.06154	956.9231	5.184577	32.85627	29.65385	4060.462
Central	15.74545	1027.7727	4.731818	28.24241	11.36364	4033.091
East Central	17.52222	994.1111	5.120000	32.90861	16.55556	6353.889
East	15.54118	890.7059	6.468941	35.66162	61.64706	5515.706

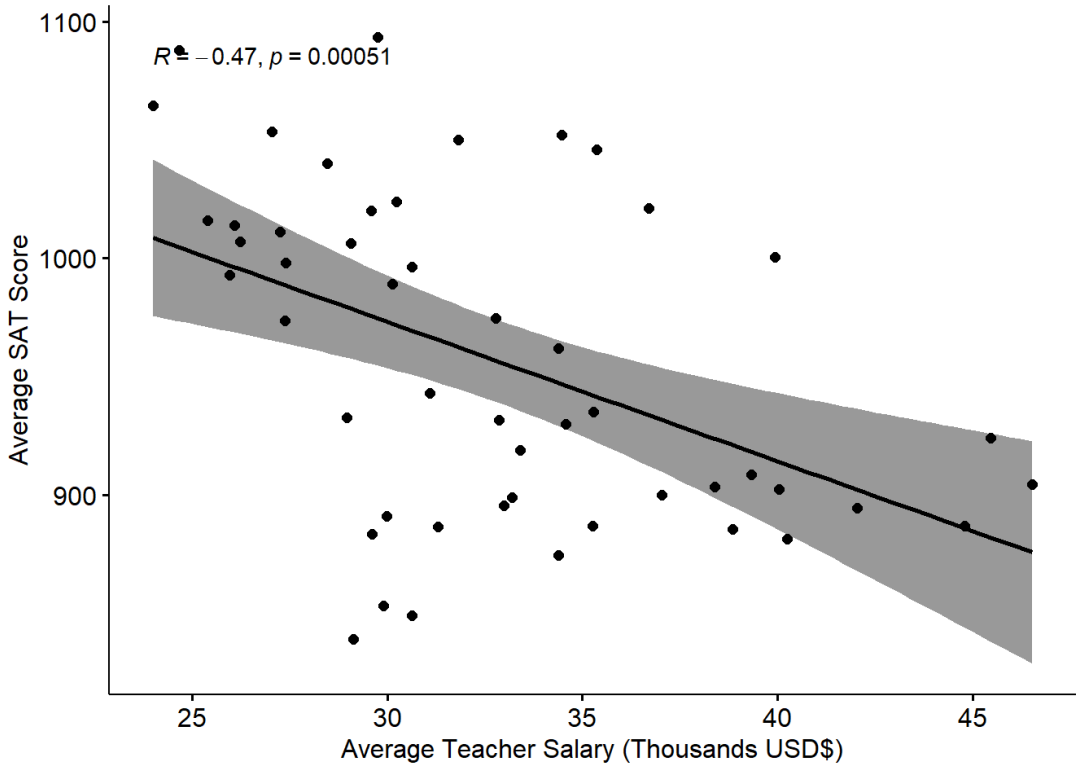
```
#calculated the mean for each variable within the dataset, including only averages and not both renditions for each y ea
r of duplicate variables.
combinedstates %>%
  group_by(region)%>%
  summarise(sd_ratio = sd(ratio),sd_averagesat = sd(averagesat),sd_averagepupilexpense = sd(averagepupilexpense),sd_aver
agesalary = sd(averagesalary),sd_averagepercent = sd(averagepercent), sd_averagesat = sd(averagesat), sd_pop = sd(pop))%
>%
  kbl%>%
  kable_styling()
```

region	sd_ratio	sd_averagesat	sd_averagepupilexpense	sd_averagesalary	sd_averagepercent	sd_pop
West	2.682517	47.71785	1.2333564	5.486687	16.67881	7839.361
Central	1.006344	57.64736	0.5617127	3.200345	11.22963	4532.399
East Central	1.210142	50.40572	1.0092045	4.441746	15.62139	3285.518
East	1.442679	25.31308	1.4173641	5.670610	13.95838	4928.842

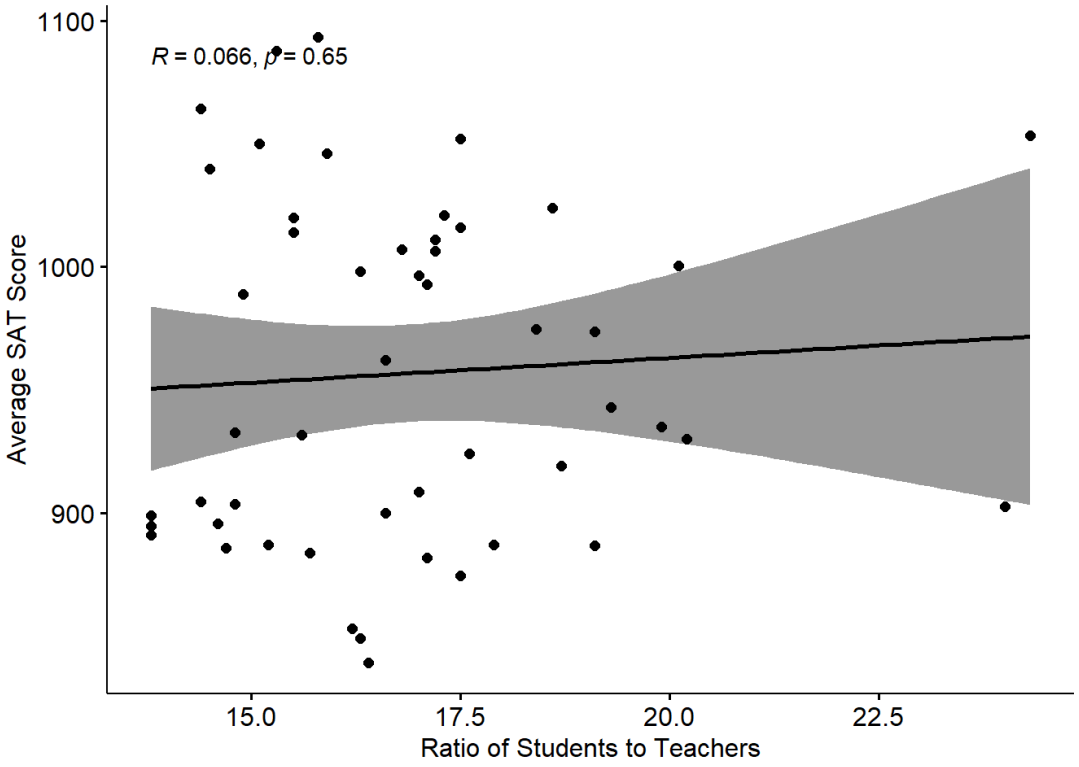
```
#created a table with a summary of standard deviations for all variables grouped by region

combinedstates_num <- combinedstates %>%
  select_if(is.numeric)
library(ggplot2)
combinedstates123 <- cor(combinedstates_num, use = "pairwise.complete.obs") %>%
  as.data.frame %>%
  rownames_to_column %>%
  pivot_longer(-1, names_to = "other_var", values_to = "correlation")
ggplot(combinedstates123, aes(rowname, other_var, fill= correlation)) +
  geom_tile()+
  scale_fill_gradient(low = "antiquewhite2", high = "darkslateblue")+
  labs(title = "Correlation heatmap for US education statistics")+
  xlab("All Variables")+
  ylab("All Variables")+
  theme(axis.text.x = element_text(angle=45, hjust=1))
```



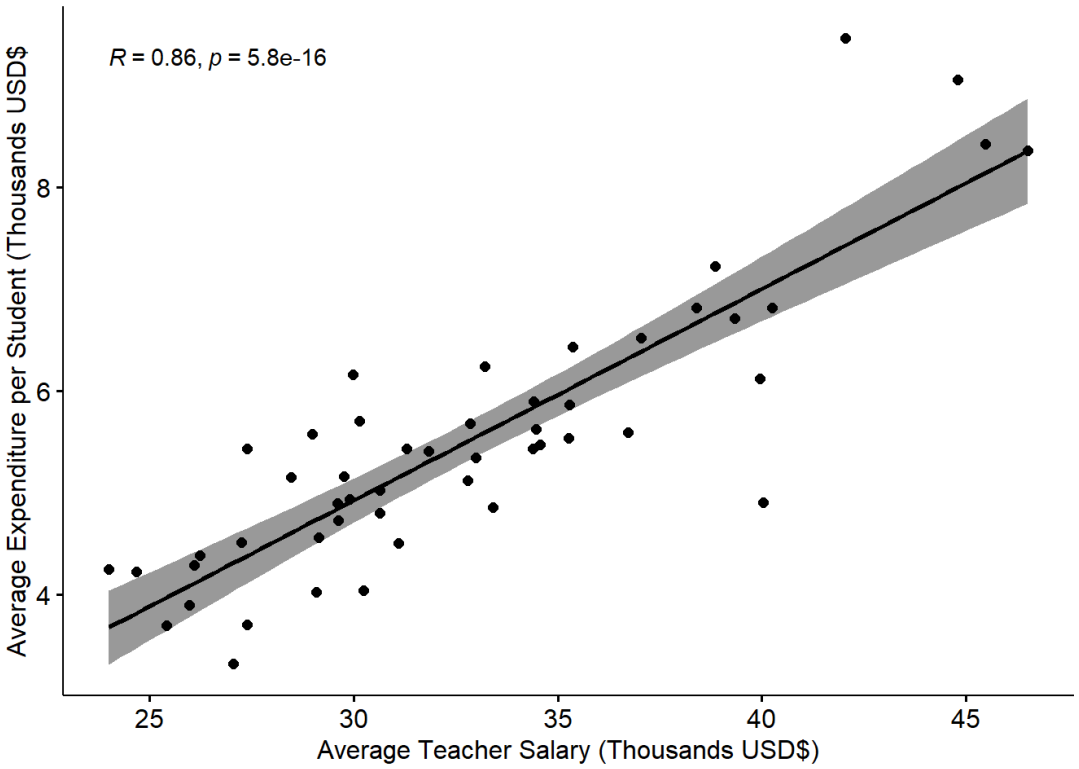
```
#correlation heatmap matrix of all variables with
library("ggpubr")
ggscatter(combinedstates, x = "averagesalary", y = "averagesat",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Average Teacher Salary (Thousands USD$)", ylab = "Average SAT Score")
```




```
ggscatter(combinedstates, x = "ratio", y = "averagesat",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Ratio of Students to Teachers", ylab = "Average SAT Score")
```




```
ggscatter(combinedstates, x = "averagesalary", y = "averagepupilexpense",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Average Teacher Salary (Thousands USD$)", ylab = "Average Expenditure per Student (Thousands USD$")
```




```
#scatterplots created with correlation coefficient
```

The mean of each significant vairable (including only averages for variables present in both years) was calculated and displayed by region, as well as the standard deviation, in order to get a good idea of where each region stands and how they compare. A correlation heatmap of all variables showed that there are heavy correlations between variables seen across both years as expected, but other correlations between unrelated variables were harder to pick out. Thus, some of the

variables that might have a correlation were plotted on a scatterplot showing the correlation coefficient in order to get a better idea. Some interesting points found was that salary of teachers had a negative correlation as compared to SAT scores, the ratio of students to teachers had a fairly week correlation with an R-value of 0.066, although still a positive correlation, and salary and expenditure per pupil were one of the strongest of correlations, with an R-value of 0.86, showing a very strong positive correlation.

MANOVA

```
manova_states <- manova(cbind(ratio,averagesat,averagesalary,averagepupilexpense,averagepercent,pop) ~ region, data = co
mbinedstates)
# OUtput of MANOVA
summary(manova_states)
```

```
##           Df Pillai approx F num Df den Df   Pr(>F)
## region     3 1.4673   6.8611     18    129 7.31e-12 ***
## Residuals 46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#one way anova performed
summary.aov(manova_states)
```

```
##  Response ratio :
##             Df Sum Sq Mean Sq F value  Pr(>F)
## region       3 110.19  36.729  11.941 6.6e-06 ***
## Residuals   46 141.50   3.076
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response averagesat :
##             Df Sum Sq Mean Sq F value    Pr(>F)
## region       3 142200   47400  23.925 1.768e-09 ***
## Residuals   46  91134    1981
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response averagesalary :
##             Df  Sum Sq Mean Sq F value   Pr(>F)
## region       3  367.84 122.615  4.9651 0.004552 **
## Residuals   46 1135.99  24.695
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response averagepupilexpense :
##             Df Sum Sq Mean Sq F value   Pr(>F)
## region       3 25.039  8.3465  6.2227 0.001227 **
## Residuals   46 61.700  1.3413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response averagepercent :
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## region       3 21613.8  7204.6  34.276 8.623e-12 ***
## Residuals   46  9668.8   210.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response pop :
##             Df     Sum Sq  Mean Sq F value Pr(>F)
## region       3   42705849 14235283  0.4618 0.7103
## Residuals   46 1417945963 30824912
```

```
pairwise.t.test(combinedstates$ratio,combinedstates$region, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  combinedstates$ratio and combinedstates$region
##
##              West    Central East Central
## Central      3.2e-05 -       -
## East Central 0.0488  0.0290  -
## East         1.9e-06 0.7648  0.0087
##
## P value adjustment method: none
```

```
#t.tests performed for variables showing significant difference to see between which regions it is.
pairwise.t.test(combinedstates$averagesat,combinedstates$region, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  combinedstates$averagesat and combinedstates$region
##
##              West    Central East Central
## Central      0.00033 -       -
## East Central 0.06020 0.09923 -
## East         0.00020 3.4e-10 1.0e-06
##
## P value adjustment method: none
```

```
pairwise.t.test(combinedstates$averagepercent,combinedstates$region, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  combinedstates$averagepercent and combinedstates$region
##
##              West    Central East Central
## Central      0.0035  -       -
## East Central 0.0428  0.4297  -
## East         3.0e-07 1.2e-11 1.4e-09
##
## P value adjustment method: none
```

```
bonferronicorrection <- 0.05/25
bonferronicorrection
```

```
## [1] 0.002
```

```
#bonferroni correction is calculated by taking p value over number of tests (edited bonferroni correction as the number
 of tests was incorrect)
type1 <- 1-(0.95^25)
type1
```

```
## [1] 0.7226104
```

```
#bonferroni correction made after adding anova, manova, and each interaction tested across the pariwise posthoc t tests.
Type 1 error is calculated by subracting .95 to the power of the number of tests undergone from 1.(edited again due to i
ncorrect number of tests being calculated)
```

The MANOVA test showed that at least one of the tested numeric variables was significantly different across the four regions of states (Pillai=1.4673, p<.0001). A one way ANOVA test showed that the variables that had the most significant change

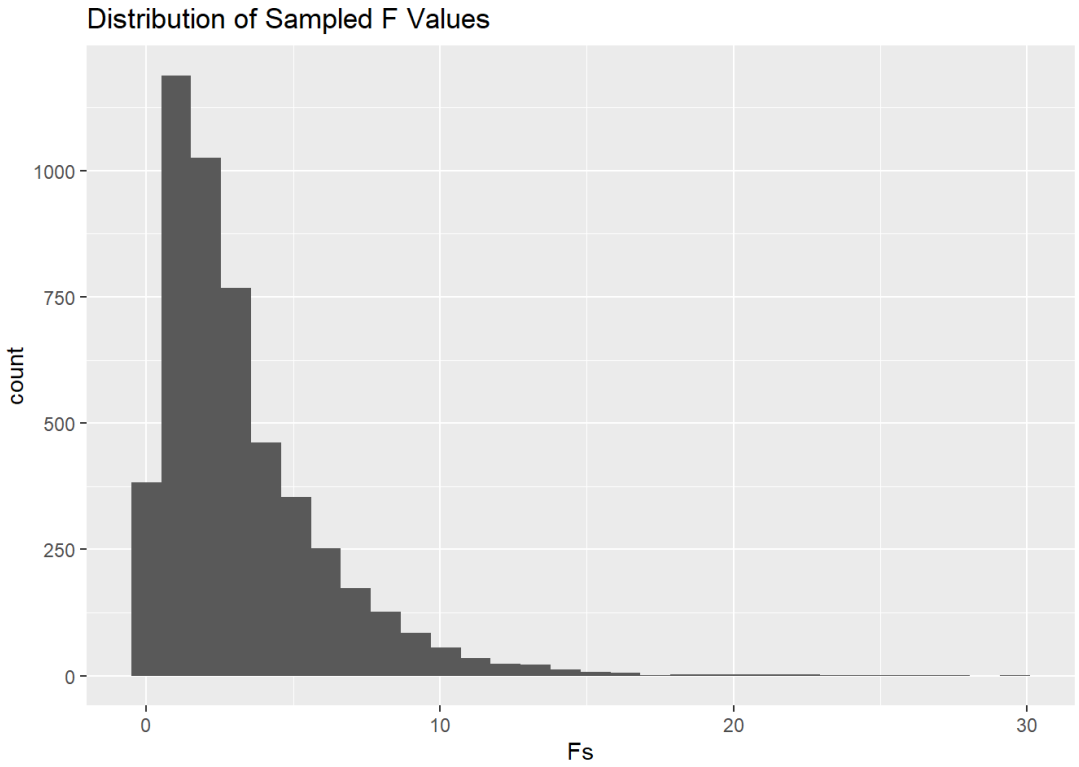
across the four regions were average SAT score, ratio of students to teachers, and average percent of graduating students who had taken the SAT. The assumptions for this MANOVA test may not have all been met. Random and independent sample assumption is met, but the assumptions of of homogeneity and normal distribution may not have been met due to there being a few outliers across all the variables and a fair amount of variance in the variables. After the Bonferroni correction which accounted for the 20 tests performed, the new p-value is now 0.002 (which means the probability of a type 1 error is now about 72%). With this in mind, the ratio of students to teachers is significantly different between the West and Central regions and the West and East regions. The average SAT score is significantly different between West and Central, West and East, Central and East, and East Central and East. This may be because of the East's high scores compared to the others. Finally, we have the average percent of graduating students taking the SAT be significantly different between the East region and every single other region (East Central, Central, and West).

RANDOMIZATION TEST

```
#observed F statistic from ANOVA performed on averagesat score based on region
obs_F <- 23.925
set.seed(348)
#set seed for replicability
Fs <- replicate(5000,{
  new <- combinedstates %>%
    mutate(averagesat = sample(averagesat))
  SSW <- new %>%
    group_by(region) %>%
    summarize(SSW= sum((averagesat - mean(averagesat))^2)) %>%
    summarize(sum(SSW)) %>%
    pull
  SSB <- new %>%
    mutate(mean = mean(averagesat)) %>%
    group_by(region) %>%
    mutate(groupmean = mean(averagesat))%>%
    summarize(SSB = sum((mean - groupmean)^2))%>%
    summarize(sum(SSB))%>%
    pull
  (SSB/1)/(SSW/48)
})
#created a loop that calculates the F statistic 5000 times
mean(Fs>obs_F)
```

```
## [1] 0.001
```

```
#compared the randomization tests f statistic to the observed one before
library(ggplot2)
Fs1 <- as.data.frame(Fs)
ggplot(Fs1, aes(x=Fs))+
  geom_histogram()+
  ggtitle("Distribution of Sampled F Values")
```




```
#created a histogram of the randomization test f values, the abline kept getting error so I had to let it go.
```

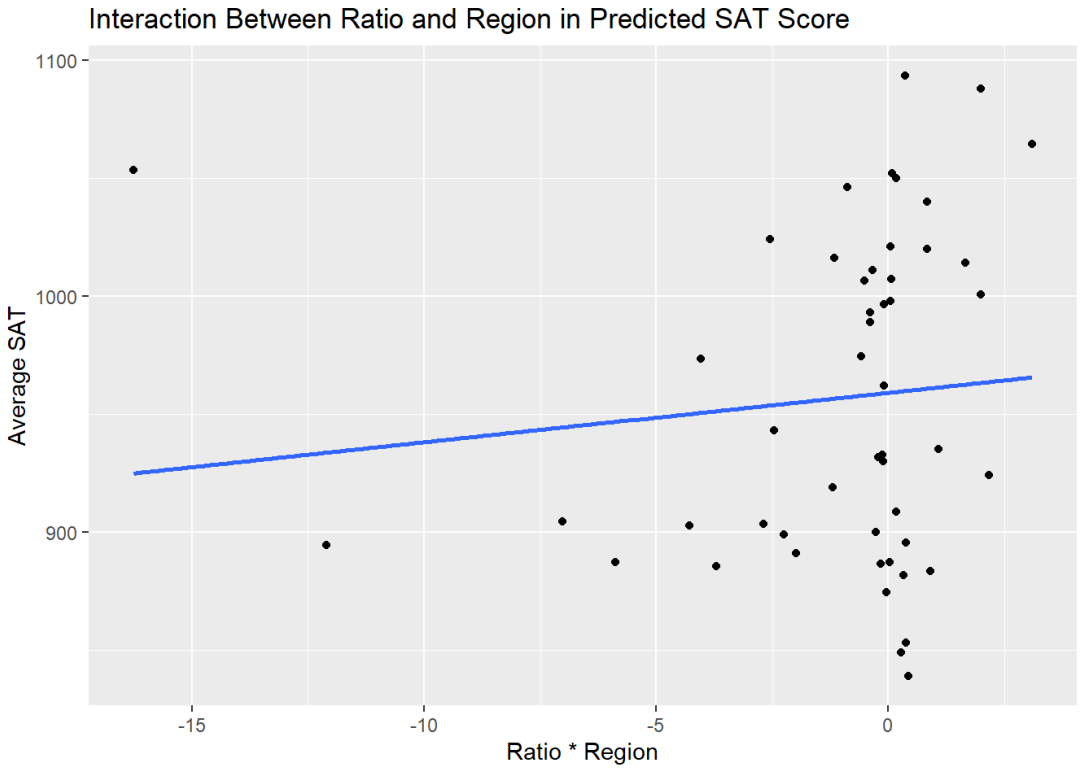
The null hypothesis is that the average SAT score does not significantly differ across any of the four regions of states. The alternative hypothesis is that the average SAT score does significantly differ across any of the four regions. The test statistic found through a one-way ANOVA was 23.925, and after the randomization test, which showed F statistics being higher than 23.925 only 0.02% of the time, which allows for rejection of the null hypothesis (without Bonferonni correction), meaning there is a significant difference in average SAT score across at least some of the regions of states. The abline function was not working for me but you can see in the histogram of the randomization test that a very very small percentage of F values were greater than 20.

LINEAR REGRESSION

```
combinedstates$ratio_c <- (combinedstates$ratio - mean(combinedstates$ratio, na.rm = TRUE))
combinedstates$averagepupilexpense_c <- (combinedstates$averagepupilexpense - mean(combinedstates$averagepupilexpense, n
a.rm = TRUE))
fit1 <- lm(averagesat ~ ratio_c + averagepupilexpense_c + ratio_c*averagepupilexpense_c, data = combinedstates)
summary(fit1)
```

```
##
## Call:
## lm(formula = averagesat ~ ratio_c + averagepupilexpense_c + ratio_c *
##     averagepupilexpense_c, data = combinedstates)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -142.361  -39.190    8.505   37.571  125.070
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   955.7971     9.5724  99.849  < 2e-16 ***
## ratio_c                        -3.8140     4.6307  -0.824  0.41440
## averagepupilexpense_c         -25.4972     7.8458  -3.250  0.00216 **
## ratio_c:averagepupilexpense_c  -0.8179     2.9964  -0.273  0.78611
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.61 on 46 degrees of freedom
## Multiple R-squared:  0.2023, Adjusted R-squared:  0.1502
## F-statistic: 3.888 on 3 and 46 DF,  p-value: 0.01473
```

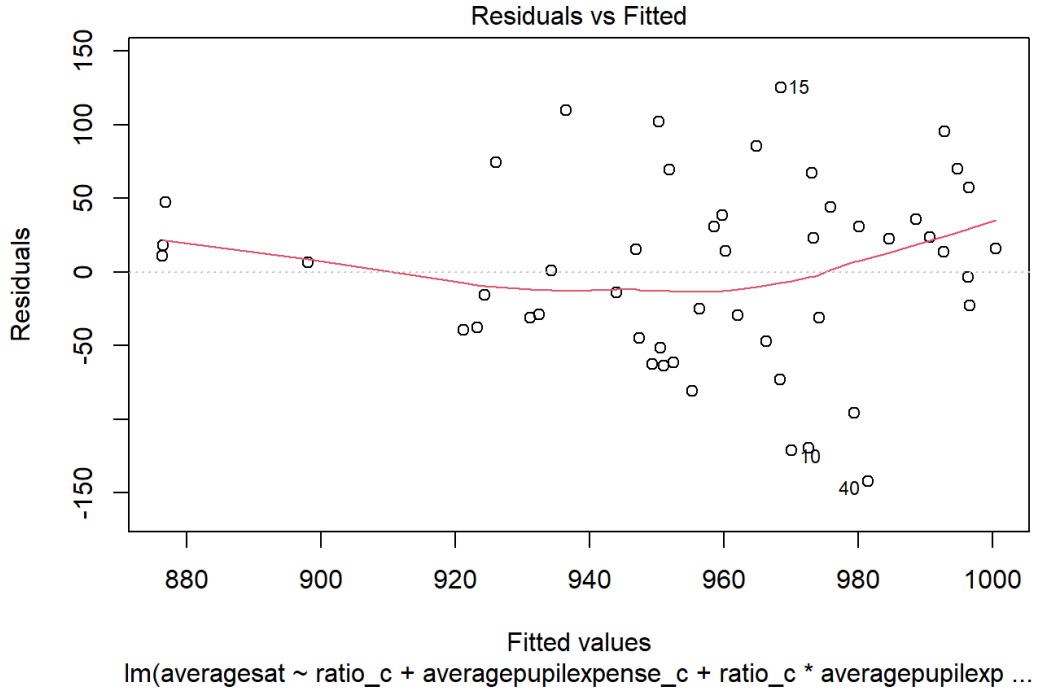
```
#created simple linear regression
#mean centered both of the predicting/explanatory variables
ggplot(combinedstates, aes( x = ratio_c*averagepupilexpense_c, y = averagesat)) +
  geom_point()+
  geom_smooth(method = 'lm', se = FALSE)+
  ggtitle("Interaction Between Ratio and Region in Predicted SAT Score")+
  xlab("Ratio * Region")+
  ylab("Average SAT")
```



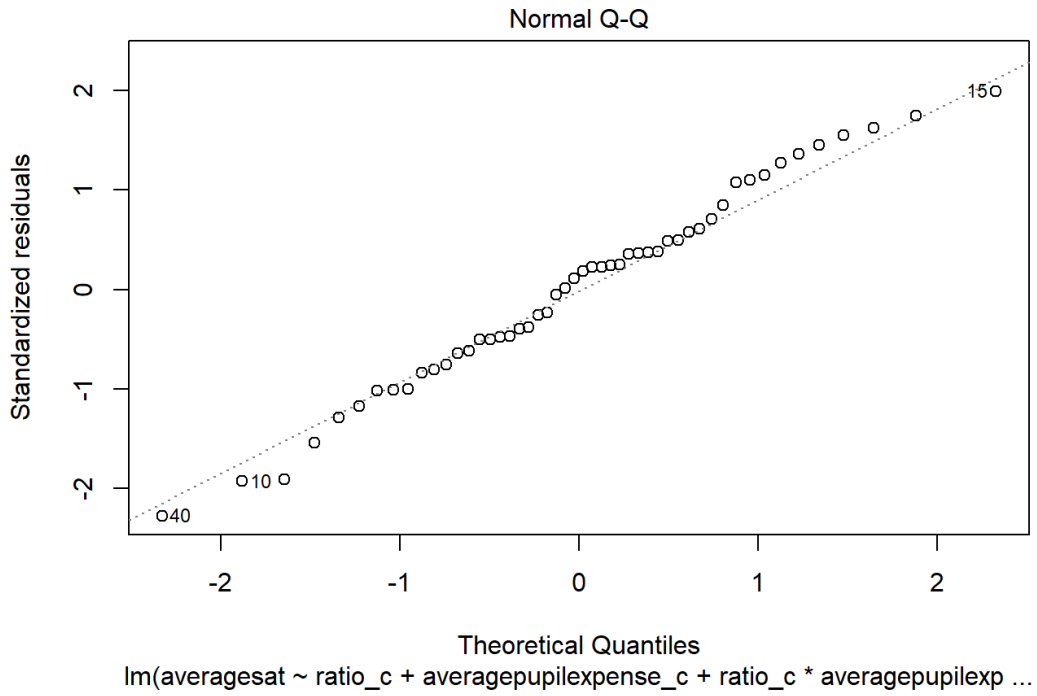

```
#added a regression line to the plot in order to better visualize the interaction
#plotted the interaction of these two variables in predicting the average SAT score.
data.frame(coef(fit1))
```

```
##                                coef.fit1.
## (Intercept)                   955.7971377
## ratio_c                        -3.8139681
## averagepupilexpense_c         -25.4972447
## ratio_c:averagepupilexpense_c  -0.8178778
```

```
# Residuals against fitted values plot to check for any problematic patterns (nonlinear, equal variance)
plot(fit1, which = 1)
```




```
# Q-Q plot to check for normality of the residuals
plot(fit1, which = 2)
```




```
#create robust standard errors
library(lmtest)
library(sandwich)
coeftest(fit1, vcov = vcovHC(fit1))
```

```
##
## t test of coefficients:
##
##                               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                  955.79714   10.22350 93.4902 < 2.2e-16 ***
## ratio_c                       -3.81397    4.68323 -0.8144  0.419619
## averagepupilexpense_c        -25.49724    8.20436 -3.1078  0.003228 **
## ratio_c:averagepupilexpense_c  -0.81788    4.69460 -0.1742  0.862460
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
samp_SEs <- replicate(5000, {
  # Bootstraped the data and fit the same linear regression model but with the new parameters
  boot_data <- sample_frac(combinedstates, replace = TRUE)
  fitboot <- lm(averagesat ~ ratio_c + averagepupilexpense_c + ratio_c*averagepupilexpense_c, data = boot_data)
  coef(fitboot)
})
samp_SEs %>%
  t %>%
  as.data.frame %>%
  # Compute the standard error (standard deviation of the sampling distribution)
  summarize_all(sd)
```

```
##   (Intercept)  ratio_c averagepupilexpense_c ratio_c:averagepupilexpense_c
## 1    9.944477 4.132509             6.876716                      3.653315
```

The coefficents show that interestingly enough, for every 1,000 USD$ more that is spent per pupils who are already recieving the average expenditure, SAT scores actually go down by 25.50 units, whereas for every 1 unit increase in the ratio between students and teachers (more students per teacher) that are at the average of ratios, the SAT score goes down 3.81 units. When looking at their interaction, the higher the ratio of students to teachers, the lower the effect of expense per pupil was on SAT score. The model explains 15.02% of the variation in the average SAT score as seen by the adjusted R-squared value. The QQ plot shows a not completely normal but fairly normal distribution. However, the residuals versus

fitted values scatterplot shows that there is not equal variance and nonlinearity due to the funnel and nonlinear nature of the scatterplot. After running the same linear regression but with robust standard errors, the coefficients remained more or less the same, with changes less than one thousandth. The centered ratio p value went from 0.4144 to 0.4196, the centered average pupil expense went from 0.0022 to 0.0032, and their interaction's p value went from 0.786 to 0.862. This means that we are now less confident in rejecting the null hypothesis and this is due to failing assumptions such as equal variance. The bootstrapped standard errors came out to be lower for both variables and was lower for the interaction on the robust version, but on the original version the interaction had a lower standard deviation. The changes were not very large, but were interesting to see as they relate again to the failing of assumptions such as linearity and normality.

LOGISTIC REGRESSION

```
combinedstates<- combinedstates%>%
  mutate(combinedstates, goodSAT = ifelse(averagesat > 1000, 0, 1))
#switched 1 and 0 so that the good SAT is represented as a 1 instead of a 0
fit4 <- glm(goodSAT ~ ratio + averagesalary, data = combinedstates, family = "binomial")
summary(fit4)
```

```
##
## Call:
## glm(formula = goodSAT ~ ratio + averagesalary, family = "binomial",
##     data = combinedstates)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1231  -0.9527   0.4955   0.9428   1.4317
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.75218    3.34518  -1.421  0.15543
## ratio         -0.07689    0.15508  -0.496  0.62006
## averagesalary  0.21128    0.08177   2.584  0.00977 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 64.104  on 49  degrees of freedom
## Residual deviance: 54.607  on 47  degrees of freedom
## AIC: 60.607
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coef(fit4))
```

```
##   (Intercept)         ratio averagesalary
##   0.008632836   0.925996042   1.235264151
```

```
#created a logit logistic regression, and created variables for the predicted outcome of the sat score
fit5 <- glm(goodSAT ~ ratio + averagesalary, data = combinedstates, family = binomial(link="logit"))
summary(fit5)
```

```
##
## Call:
## glm(formula = goodSAT ~ ratio + averagesalary, family = binomial(link = "logit"),
##     data = combinedstates)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1231  -0.9527   0.4955   0.9428   1.4317
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.75218    3.34518  -1.421  0.15543
## ratio         -0.07689    0.15508  -0.496  0.62006
## averagesalary  0.21128    0.08177   2.584  0.00977 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 64.104  on 49  degrees of freedom
## Residual deviance: 54.607  on 47  degrees of freedom
## AIC: 60.607
##
## Number of Fisher Scoring iterations: 5
```

```
combinedstates$prob <- predict(fit5, type = "response")
combinedstates$predicted <- ifelse(combinedstates$prob > .5, "good", "not good")

#confusion matrix created in order to discuss the accuracies and the shortcomings of the predictions.
table(truth = combinedstates$goodSAT, prediction = combinedstates$predicted)
```

```
##      prediction
## truth good not good
##     0   10        7
##     1   30        3
```

```
#accuracy
13/50
```

```
## [1] 0.26
```

```
#sensitivity
10/17
```

```
## [1] 0.5882353
```

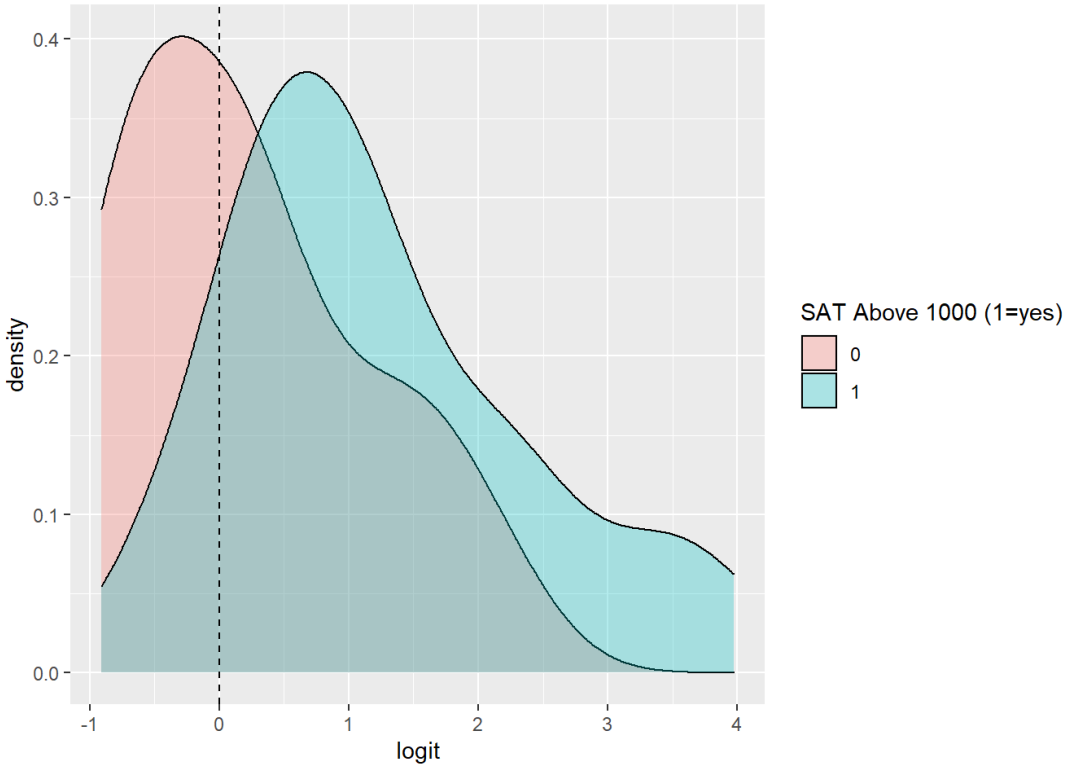
```
#specificity
3/33
```

```
## [1] 0.09090909
```

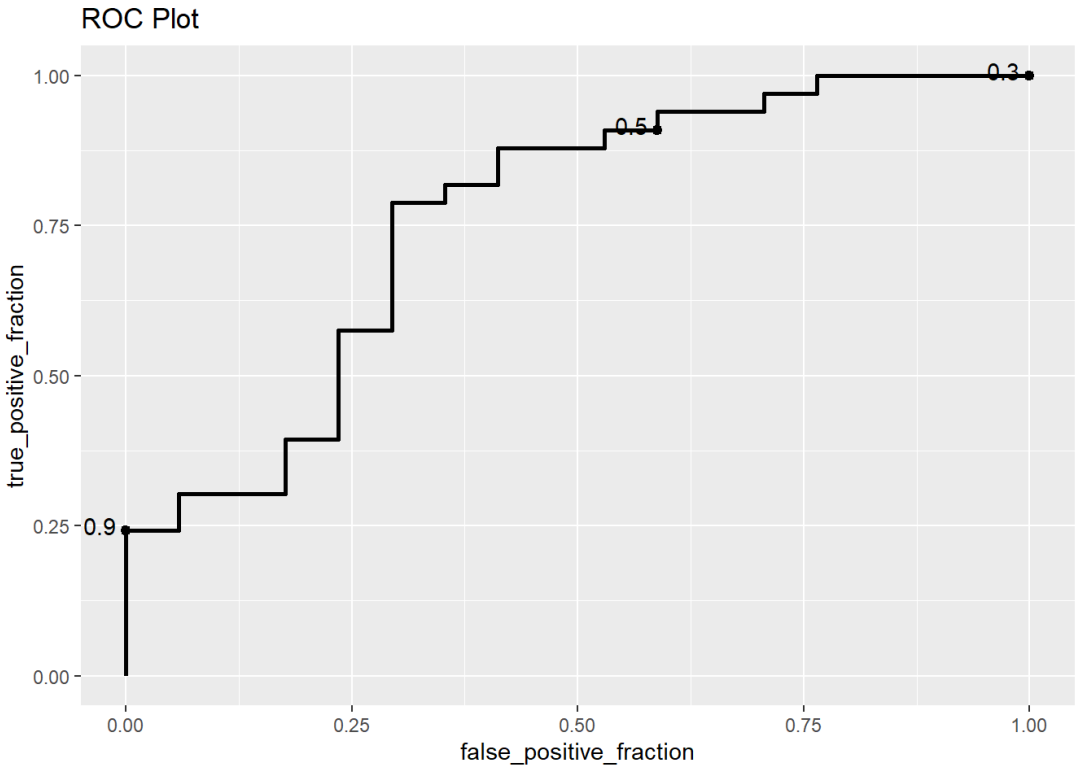
```
#precision
10/40
```

```
## [1] 0.25
```

```
#density plot
combinedstates$logit <- predict(fit4)
ggplot(combinedstates, aes(logit, fill = as.factor(goodSAT))) +
  geom_density(alpha = .3) +
  geom_vline(xintercept = 0, lty = 2) +
  labs(fill = "SAT Above 1000 (1=yes)")
```




```
#graph the ROC curve and then used calc_auc in order to calculate the AUC or the area under the line.
library(plotROC)
combinedstates$prob <- predict(fit4, type = "response")
ROC <- ggplot(combinedstates)+
  geom_roc(aes(d = goodSAT, m = prob), cutoffs.at = list(0.1,0.5,0.9))+
  ggtitle("ROC Plot")
ROC
```




```
calc_auc(ROC)
```

```
##   PANEL group       AUC
## 1     1    -1 0.7611408
```

Controlling for the average teachers salary, for every 1 unit increase in the ratio of students to teachers , the odds of getting a “good” SAT score (above 1000), increase by 1.0799. Controlling for the ratio of students to teachers, for every 1 unit increase (1,000$) in the average teacher salary, the odds of getting a “good” SAT score decrease by 0.8095. The accuracy of the model, which is them proportion of correctly classified cases is 0.26. The sensitivity which is the proportion of true positive cases is 0.59. The specificity which is the proportion of true negatives is .091, and the precision which is the proportion of true positive predictions is 0.25. The calculated AUC for the ROC graph is 0.7611, which means the model has a “fair” prediction power.

```
##           sysname           release           version          nodename
##         "Windows"          "10 x64"     "build 19042" "LAPTOP-VDO7L3JC"
##           machine             login              user    effective_user
##          "x86-64"              "HP"              "HP"              "HP"
```