Final Project

Funny Guys

2024-04-29

INTRODUCTION

This project delves into an extensive data set, capturing a range of educational metrics across various U.S. states from the years 1992 and 2015, focusing on financial, demographic, and academic performance indicators. By comparing changes over these years, the project aims to shed light on trends and shifts in education funding, resource allocation, and student performance. Critical data points such as student enrollment figures, total revenue and expenditures for education, instructional expenditures, and standardized test scores in mathematics are analyzed to understand their interrelations and impact on educational outcomes. The introduction of calculated variables like proportional instruction expenditure and gender-specific performance differences further enriches the analysis, providing deeper insights into the dynamics of educational equity and efficacy over the observed period.

DATA

The data we used for this project can be found here. This data set was created by aggregating financial data from the US Census, enrollment data from the National Center for Education Statistics, and academic achievement data from the National Assessment of Educational Progress.

VARIABLES

Continuous Variables:

- **ENROLLMENT**: The maximum count of students from two different sources, namely "ENROLL" and "A_A_A." It consolidates data from these sources to provide a more comprehensive picture of student enrollment in the state.
- **TOTAL_REVENUE**: The total amount of revenue generated by the state's educational system, in dollars. It encompasses federal, state, and local contributions.
- **TOTAL_EXPENDITURE**: The total expenditure from the state's educational system, covering expenses such as salaries, maintenance, and program funding.
- **INSTRUCTION_EXPENDITURE**: The portion of total expenditure dedicated to instructional activities, including teacher salaries, classroom materials, and curriculum development. It highlights the investment made in direct educational services.
- **PROP_INSTR_EXP**: This variable is derived by dividing the instruction expenditure by the total expenditure, resulting in the proportion of total expenditure allocated to instructional activities. It provides insight into the budget allocation, emphasizing the relative importance of instruction within the overall education budget.
- **AVG_MATH**: The average math score, out of 500, of eighth-grade students within the state.
- **AVG_MATH_M**: The average math score, out of 500, for eighth-grade students classified as "male."

- **AVG_MATH_F**: The average math score, out of 500, for eighth-grade students classified as "female."
- **MATH_GENDER_DIF**: The difference between the average math scores of male and female eighth-grade students.
- **SCORE_ASIAN**: The average math score, out of 500, of eighth-grade students who identify as Asian.
- **SCORE_NATIVE_AM**: The average math score, out of 500, of eighth-grade students who identify as Native American.
- **SCORE_HISPANIC_LATINO**: The average math score, out of 500, of eighth-grade students who identify as Hispanic or Latino.
- **SCORE_BLACK**: The average math score, out of 500, of eighth-grade students who identify as Black or African American.
- **SCORE_WHITE**: The average math score, out of 500, of eighth-grade students who identify as White.
- **SCORE_HAWAIIAN_PI**: The average math score, out of 500, of eighth-grade students who identify as Hawaiian or Pacific Islander.
- **SCORE_MULTIPLE_RACES**: The average math score, out of 500, of eighth-grade students who identify as two or more races.

Categorical Variables

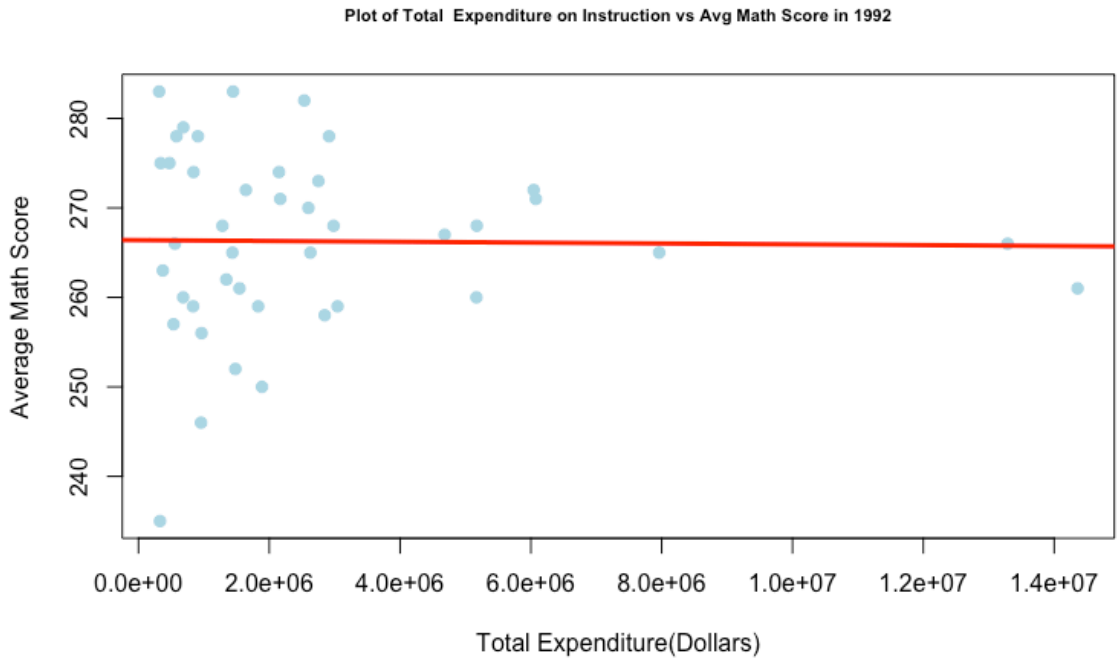
- **YEAR**: The year to which the data corresponds. It serves as a categorical identifier for different time periods within the dataset.
- **STATE**: The state from which the data is sourced.
- **Region**: This variable categorizes geographic areas, such as states or territories, into broader regions like Northeast, Midwest, South, and West.

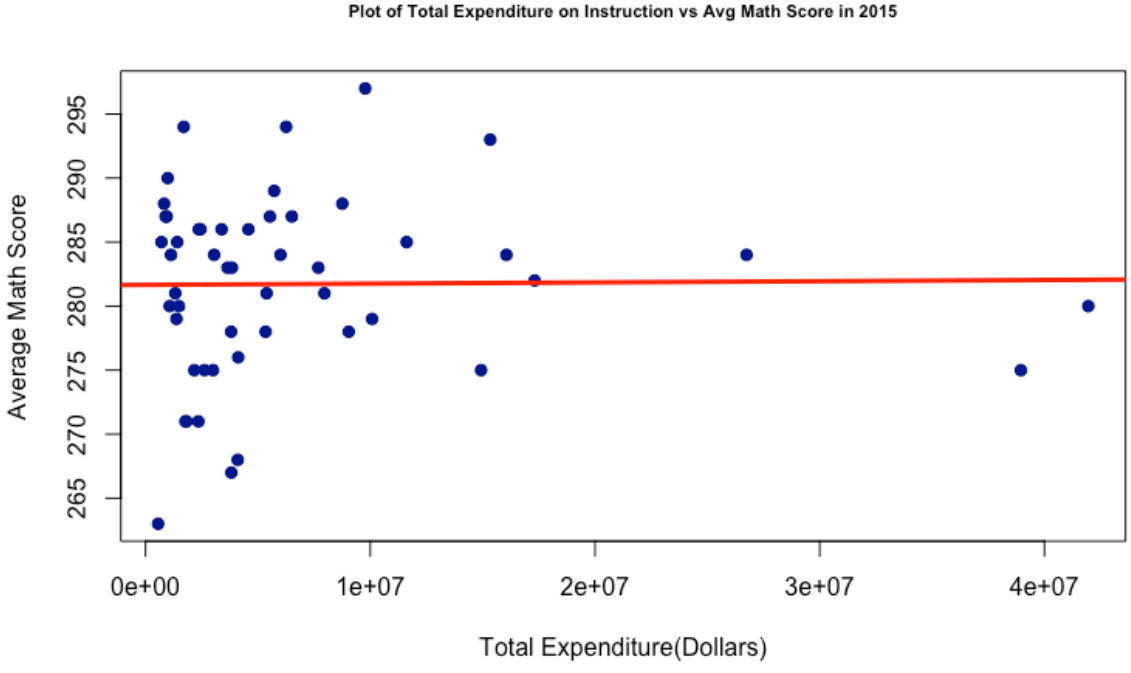
Data Cleaning

The cleaning process for the dataset began by creating a unified enrollment column, *ENROLLMENT*, by selecting the maximum value between two existing columns, *ENROLL* and *A_A_A*, which represented student counts from different sources. This step ensured consistency and completeness in student enrollment data. Next, a new variable, *PROP_INSTRUCTION_EXPENDITURE*, was calculated by dividing the *INSTRUCTION_EXPENDITURE* by the *TOTAL_EXPENDITURE*, providing insight into the proportion of total expenditure allocated to instruction. Another new variable, *MATH_GENDER_DIF*, was created to capture the difference between male and female math scores in grade 8, derived from existing columns. Additionally, variable names were standardized and simplified for clarity and consistency. The process also involved renaming several variables for clarity and consistency. The new names were more descriptive names compared to their original counterparts, facilitating easier interpretation and analysis of the dataset. To get an in depth look at the renaming process please look into the rmd file. Two separate data frames, *States1992* and *States2015*, were then created to isolate data from 1992 and 2015, respectively, for focused analysis. The rows in these data frames were labeled with state names to facilitate identification and interpretation. Furthermore, columns containing all NA values were removed to ensure data quality and reduce redundancy. Additionally, we created a new column called *Region* where we categorized each of the states into one of four regions(Northeast, South, Midwest, West). This is used later on in our ANOVA testing. Overall, these cleaning steps aimed to prepare the

dataset for subsequent analysis by addressing missing values, standardizing variable names, and organizing the data for comparative analysis between the two selected years, 1992 and 2015.

Analyzing State Expenditure and Average Math Scores



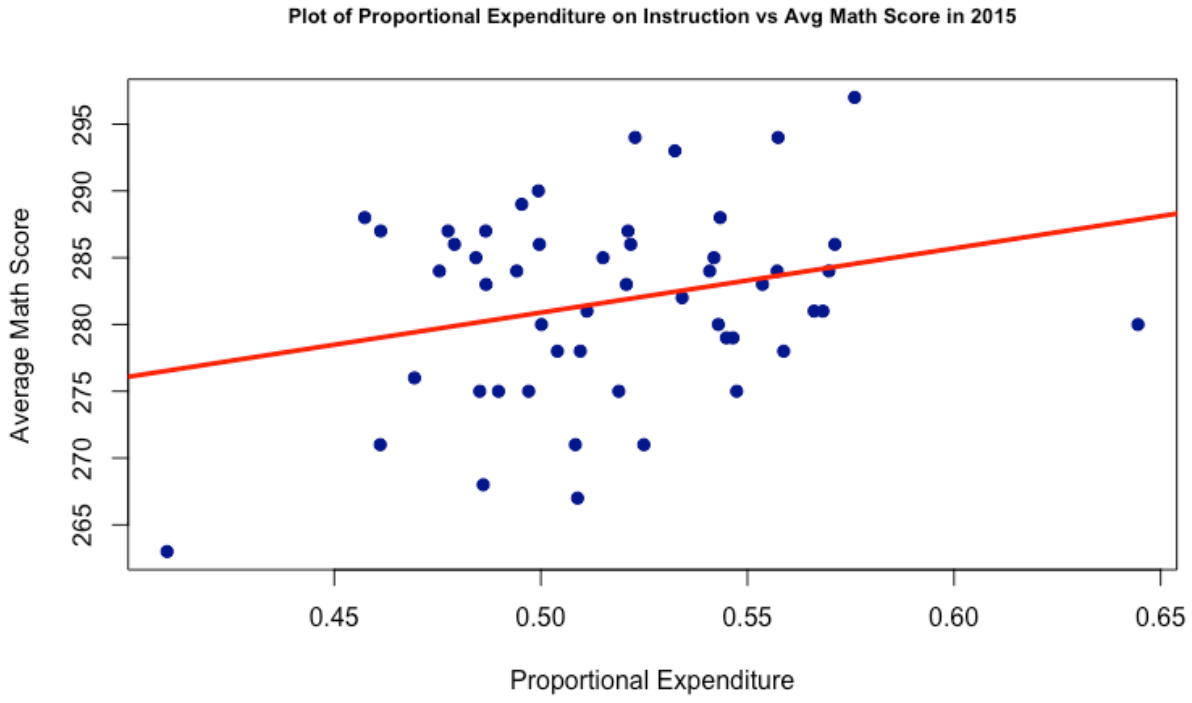


The first scatter plot shows the relationship between total expenditure on instruction and average math score in 1992. There appears to be a positive correlation, with higher expenditures generally associated with higher average math scores, though the relationship is not perfectly linear. Specifically, most of the data points are concentrated in the upper-left of the plot, further hinting at the fact that this relationship is not linear.

The second scatter plot displays the same variables as the previous one, but for the year 2015. Here, there is a similar issue where the data does not appear linear and is concentrated in the upper left of the plot. To get some further clarification on this information, we run correlations tests to see how strong the correlation is between these two values between the two years.

```
## [1] "Correlation between Avg Math Scores and Instruction Expenditure in 1992 and 2015 respectively:"

## [1] -0.01446896

## [1] 0.01160543
```

Based on the correlations, it seems like the correlation coefficients for both 1992 and 2015 are very close to zero, indicating a very weak linear relationship between average math scores and instruction expenditure in both years. This suggests that there may not be a strong linear association between these variables, at least not one that is captured by the correlation coefficient alone. This suggests that we should use the proportional expenditure on instruction rather than just pure expenditure on instruction.



This first scatter plot displays the relationship between the proportional expenditure on instruction and average math score in 1992. Unlike the plot displaying the total expenditure, there is a positive correlation. In this graph then correlation is much more pronounced and is more linear. Here we can see that, generally when a higher proportion of the budget goes to instruction, we would expect that the average math score would increase.

The variables in the second plot are the same as the previous plot, this plot just focuses on 2015. Just like the previous plot, this one demonstrates how we should take into account the proportion of the total expenditure that goes to instruction rather than just the total expenditure on instruction. The different sizes of different states will skew the data if we only focus on the total expenditure on instruction. We see a slight positive correlation in this plot.

```
## [1] "Correlation between Avg Math Scores and Proportional Instruction Expenditure in 1992 and 2015 respectively"

## [1] 0.1828254

## [1] 0.2694149
```

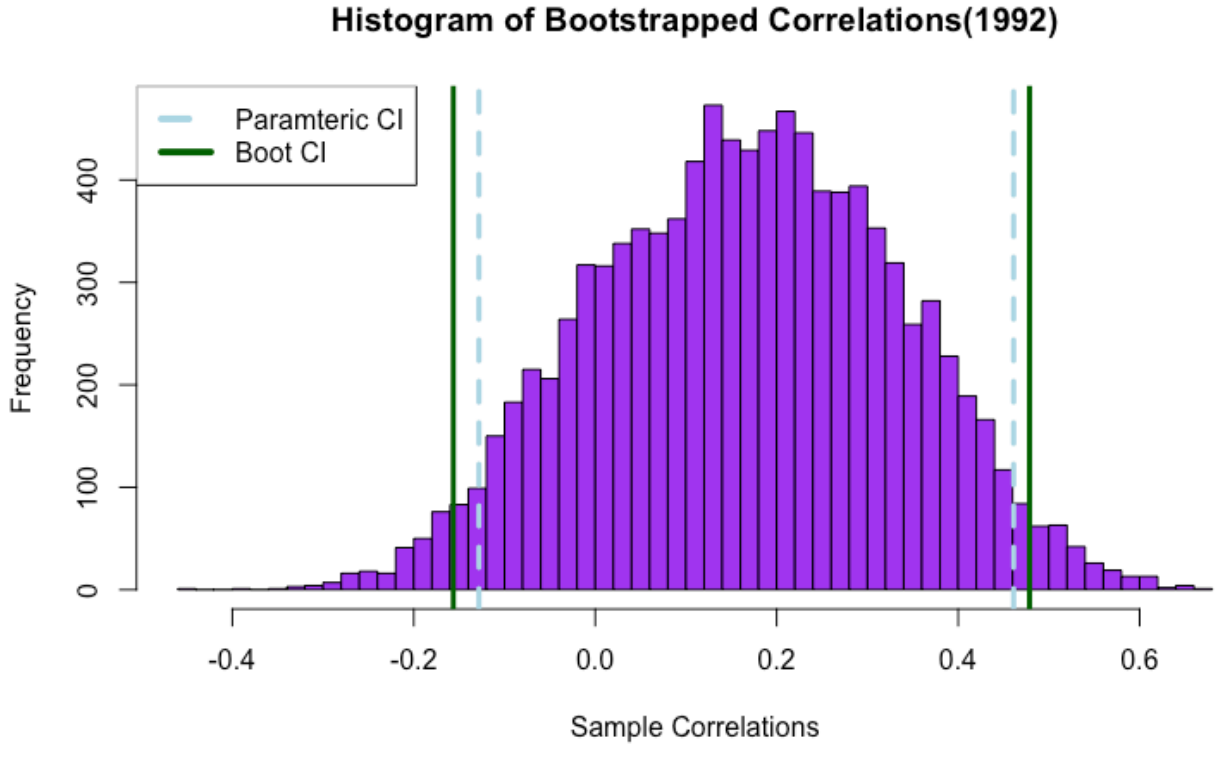
The correlation coefficients between average math scores and proportional instruction expenditure for 1992 and 2015 are 0.1828254 and 0.2694149, respectively, indicating a positive but relatively weak linear relationship in both years. States allocating a higher proportion of expenditure to instruction tended to have slightly higher average math scores, with a slightly stronger association observed in 2015. However, it's important to also note that correlation doesn't imply causation, and other factors not captured by this analysis may also influence average math scores. Let's look at the Bootstrapped

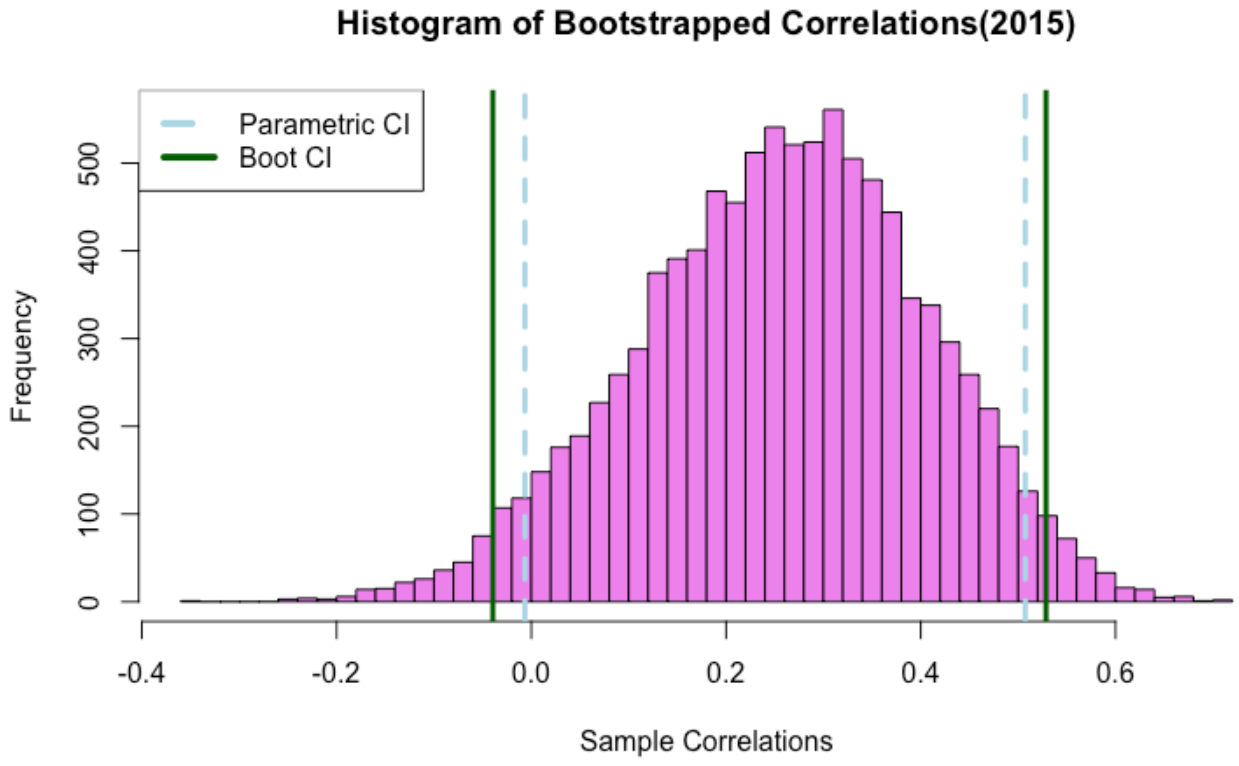
Confidence Intervals Between the Proportional Expenditures and average math scores to get a more complete pictures.

Bootstrapped Confidence Intervals

```
## [1] "1992 BootStrapped Correlations"

##      2.5%     97.5%
## -0.1564281  0.4786892

## [1] "2015 BootStrapped Correlations"

##       2.5%      97.5%
## -0.03953775  0.52860831
```

These are the results from our bootstrapped confidence intervals for correlations between proportional expenditures and average math scores in 1992 and 2015. For 1992, the bootstrapped correlation is estimated to lie between -0.1590 and 0.4828 with a 95% confidence level. In 2015, the bootstrapped correlation is estimated to range from -0.0352 to 0.5304 with the same confidence level. These intervals signify the range of plausible values for the correlation coefficients based on resampling from the data. The wider interval in 1992 suggests greater uncertainty in the correlation estimate compared to the narrower interval in 2015, indicating more variability in the relationship between proportional expenditures and math average scores in the earlier period.

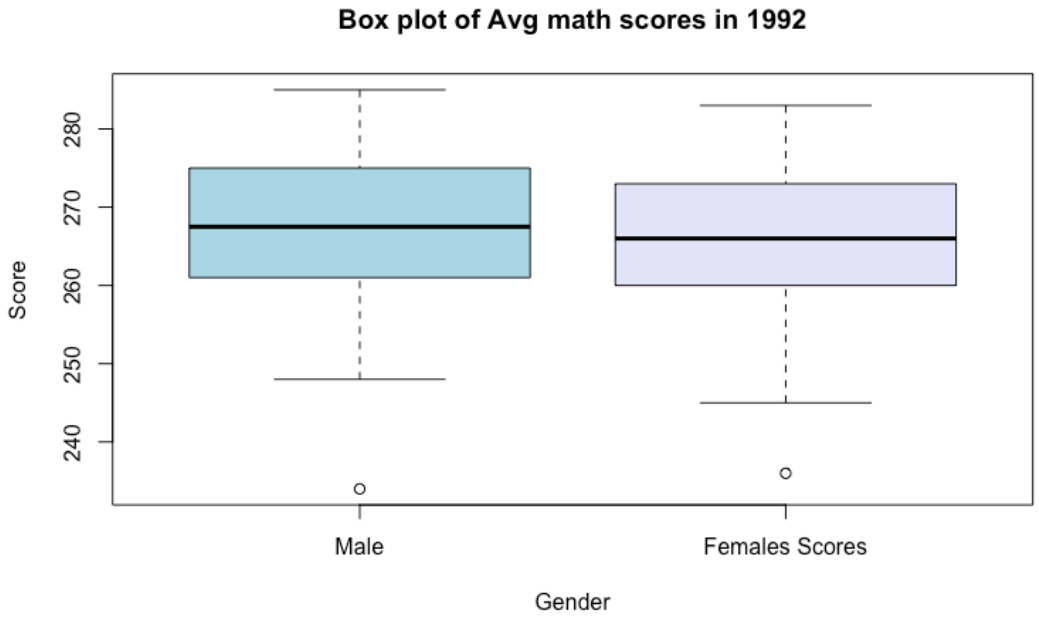


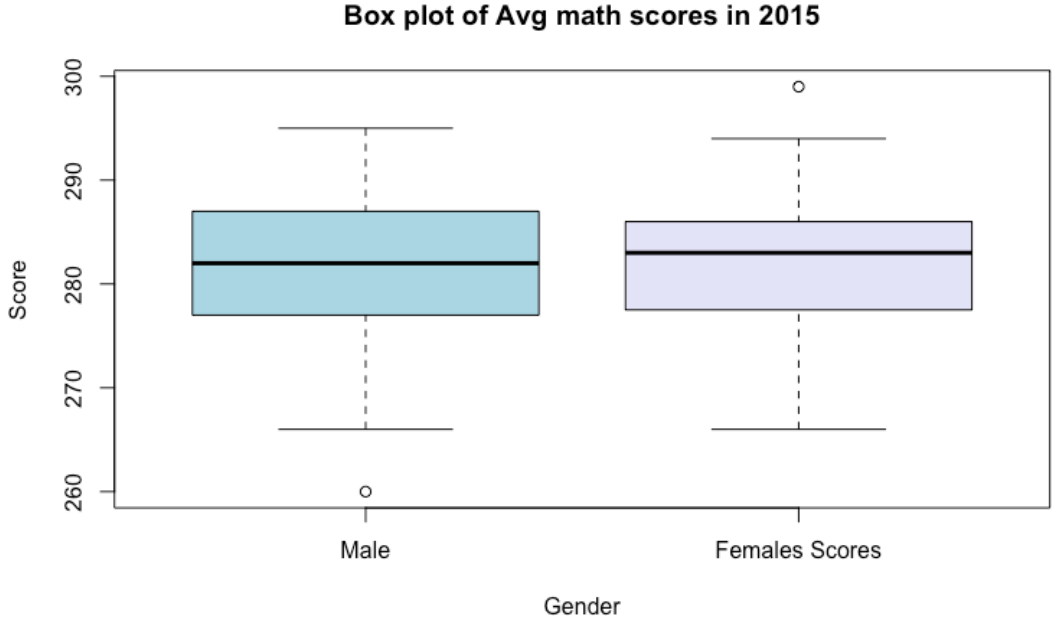


The first histogram displays a distinct peak just below 0.2, indicating a concentration of bootstrapped correlation values around that point. The parametric confidence interval (teal line) appears narrower compared to the bootstrap confidence interval (green line). This suggests the data may not follow a normal distribution, and the bootstrap approach provides a more robust estimate of the confidence interval. Furthermore, the value of 0 is included within the confidence interval which may hint at the possibility that the correlation is not as strong as expected. Still the interval goes from just below -.1 to about .5.

In the second histogram, the distribution of bootstrapped correlations in 2015 exhibits a similar shape to the 1992 data, with a pronounced peak near 0.3. The peak appears slightly shifted to the right compared to 1992. The parametric confidence interval remains narrower than the bootstrap interval, consistent with the earlier observation. Furthermore, the confidence interval itself seems to be slightly smaller, and we see it go from slightly less than 0 to over .5. We are more confident of a higher correlation here than in 1992.

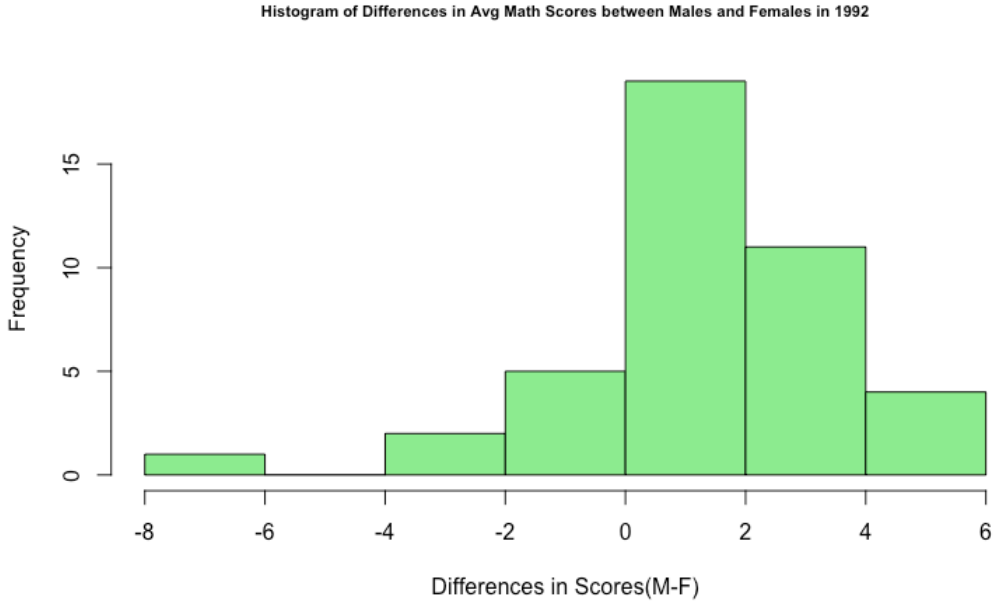
Analyzing Scores between Males and Females

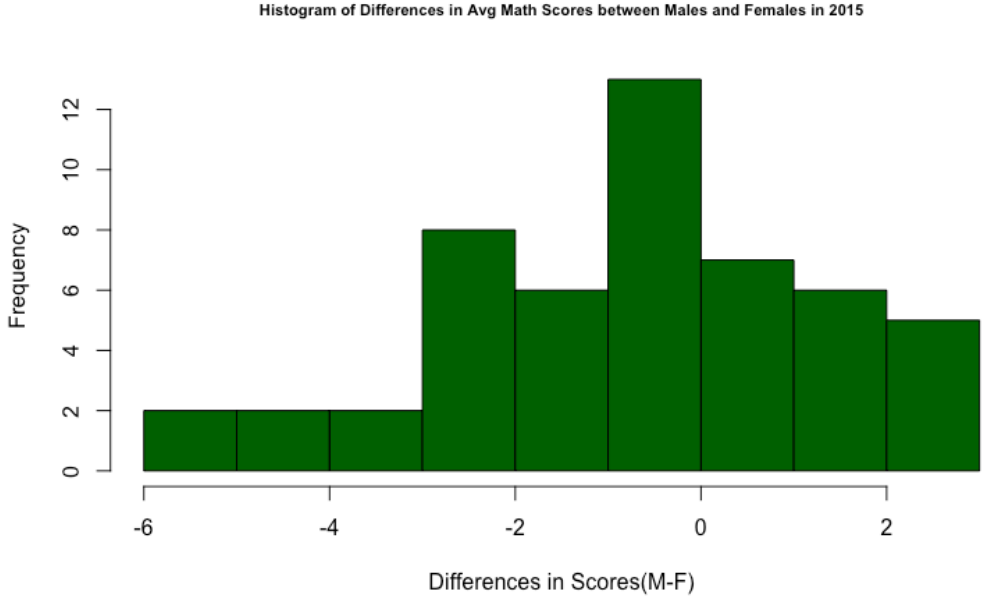




The first box plot displays the average math scores across all states in 1992, divided by gender. In this plot, both men and women have one outlier far below the other states with average scores below 240. It seems as though men have generally higher scores than women in math. Besides the outlier, the median, quartiles, min, and max are all higher for men then they are for women, indicating that men are performing better than women in math.

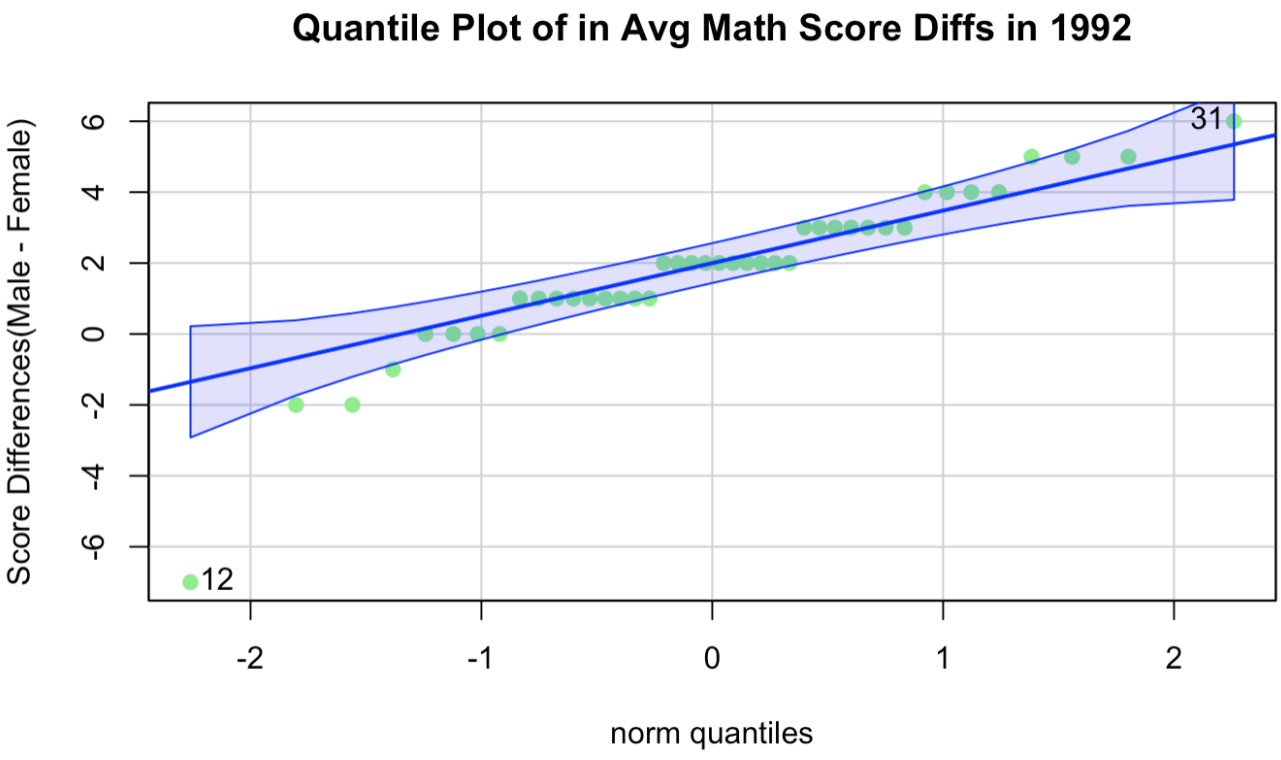
The second box plot displays the average math scores across all states in 2015, divided by gender. In this plot, unlike in 1992, men have an outlier further below, but it is only at a score of 260. Furthermore, women actually have an outlier further ABOVE, which did not occur in 1992. Additionally, it actually seems as though men and women have similar math scores, but women have a higher mean. Based on this plot and the former, it seems as though while scores for both men and women increased, the scores for women certainly increased the most. We can look into this further by utilizing the MATH_GENDER_DIF variable that we created to visualize the differences in scores between the genders.



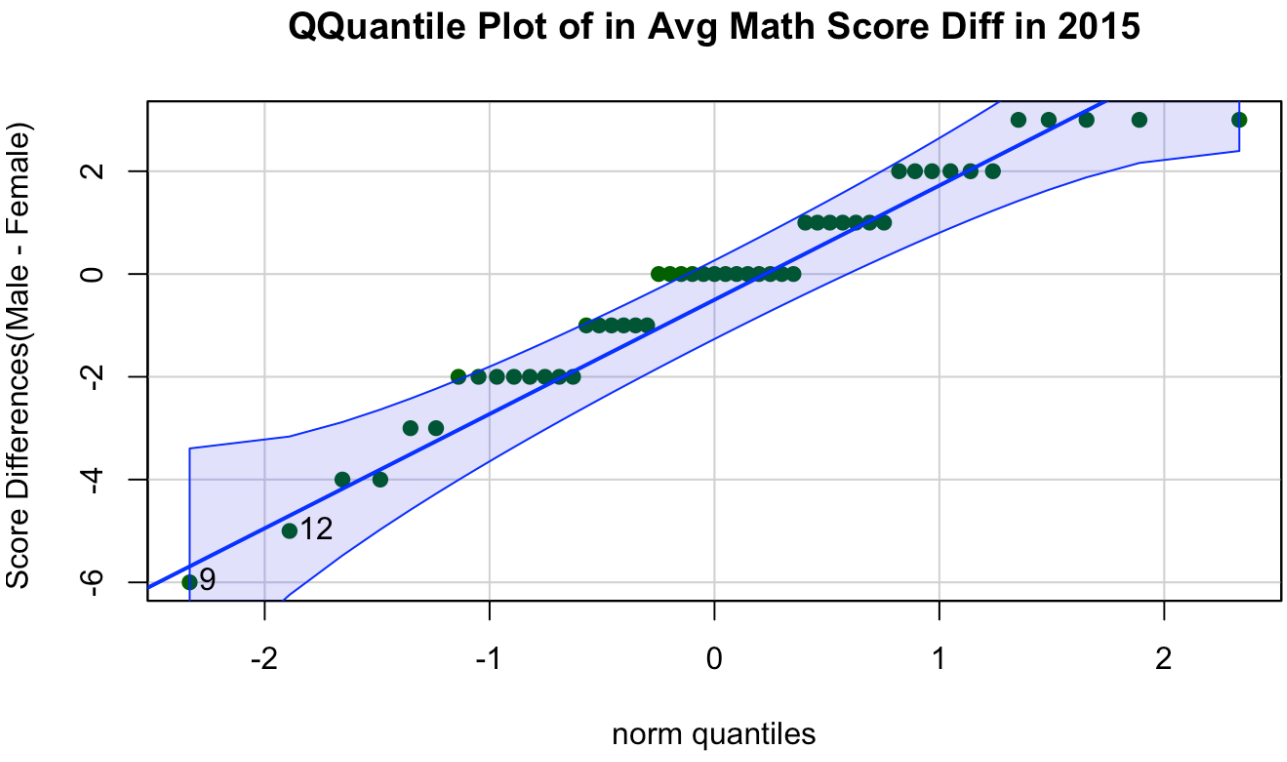


The first histogram displays the frequency distribution of the states' math gender difference scores in 1992. The distribution looks somewhat normal with a center at 0, however there's still a decent left skew. It still seems as though male scores tend to be higher since most states fall with a difference above 0 (meaning men have a higher score). Still, it seems more even than we would have believed from the box plots.

The second histogram displays the frequency distribution of the states' math gender difference scores in 2015. The distribution appears to be slightly skewed to the left, with the large majority of states falling below a difference of 0 (meaning that women tended to have a higher average score than men in. most states). This is certainly a change from 1992 where men had higher scores, but the data was more centered around 0 and normally distributed. This seems to be less normally distributed, as we will see in the qq plot.



```
## [1] 12 31
```




```
## [1]  9 12
```

This first normal quantile plot depicts the quantiles of a standard normal distribution, with the x-axis representing the norm quantiles and the y-axis showing the states' 1992 math gender difference scores. The points closely follow the diagonal line, suggesting that the gender difference scores are approximately normally distributed.*

Similar to the previous plot, this one shows the quantiles of the states' 2015 math gender difference scores plotted against the norm quantiles. The points appear to deviate slightly from the diagonal line, indicating a potential departure from normality, though the deviation is relatively small.

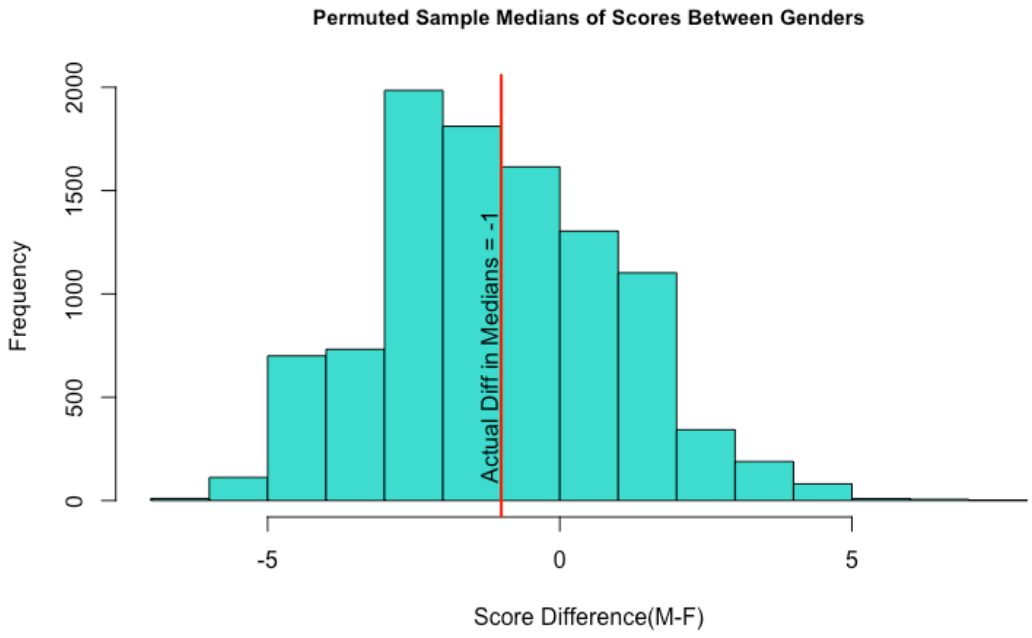
T-Test

```
##  Welch Two Sample t-test
##
## data:  States1992$AVG_MATH_M and States1992$AVG_MATH_F
## t = 0.79624, df = 81.902, p-value = 0.4282
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.675791  6.247219
## sample estimates:
```

```
## mean of x mean of y
##  267.2857  265.5000

##  Welch Two Sample t-test
##
## data:  States2015$AVG_MATH_M and States2015$AVG_MATH_F
## t = -0.17853, df = 99.178, p-value = 0.8587
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.087821  2.578017
## sample estimates:
## mean of x mean of y
##  281.6471  281.9020

##  Welch Two Sample t-test
##
## data:  States1992$AVG_MATH and States2015$AVG_MATH
## t = -8.2789, df = 71.068, p-value = 5.049e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19.15833 -11.72122
## sample estimates:
## mean of x mean of y
##  266.2857  281.7255
```

The Two Sample t-tests were conducted to assess the differences in average math scores between male and female students in 1992 and 2015, as well as the difference in average math scores between the two years. In 1992, the difference in means between male (mean = 267.2857) and female (mean = 265.5000) average math scores was not statistically significant (t = 0.79624, p = 0.4282, df = 81.902), indicating that there was no significant gender gap in math performance that year. Similarly, in 2015, there was no significant difference in mean math scores between males (mean = 281.6471) and females (mean = 281.9020) (t = -0.17853, p = 0.8587, df = 99.178), suggesting that gender parity in math achievement was maintained over time. However, comparing average math scores between 1992 (mean = 266.2857) and 2015 (mean = 281.7255) revealed a statistically significant difference (t = -8.2789, p < 0.000000000005049, df = 71.068), with a substantial increase in mean scores over the years. This suggests an overall improvement in math performance from 1992 to 2015.

Permutation Test




```
## [1] "p-Value:"

## [1] 0.771
```

The histogram displays the distribution of permuted sample medians for the difference in math scores between males and females across states in 2015. The distribution is somewhat skewed to the right, with the peak around -3, indicating females tended to outperform males in most permuted samples. The vertical red line, representing the observed median difference in the original data, falls in the upper tail of this distribution. However, the p-value of 0.771 from the permutation test suggests the observed gender difference in median math scores is not statistically significant at a typical 0.05 level. The permutation test does not provide strong evidence against the null hypothesis of no gender difference in medians.

Scores by Race/Group

Correlation Tests

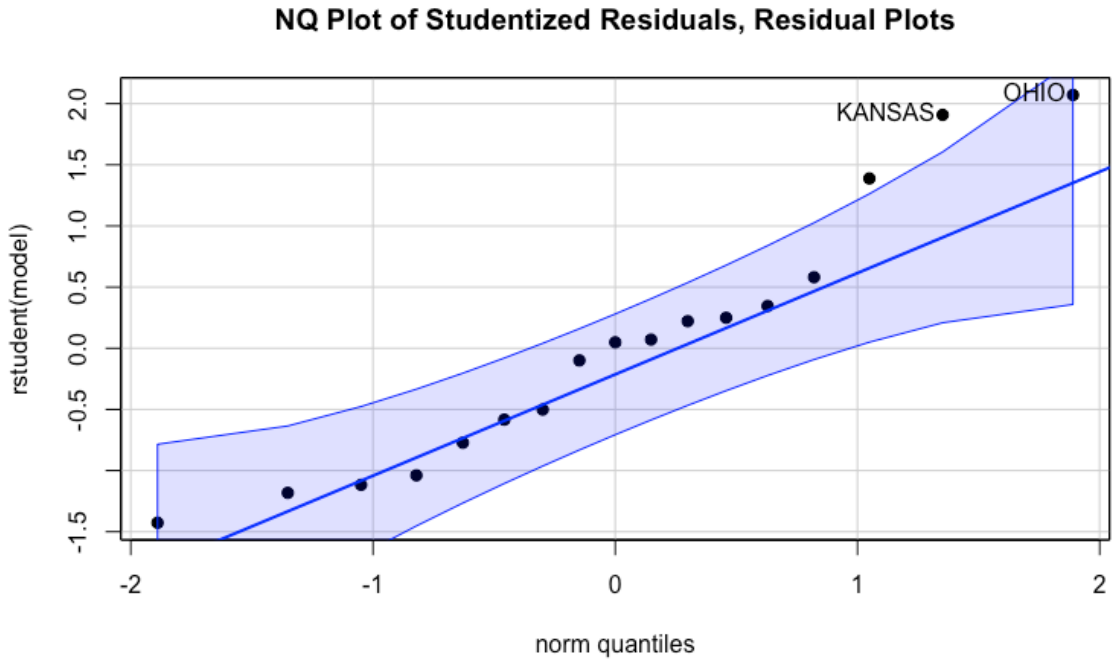
To look at average scores by race or groups, we also performed Correlation Tests between the average Math score in 2015 and the Average math scores for the Individual Races to see how the correlations might vary between different races. We haven't included the evaluation of this code in this document to keep it short, but the code is available in the rmd file. The following is an outline of the statistics from the correlation tests. For Native American students (SCORE_NATIVE_AM), a positive correlation was found (cor = 0.3513), although it was not statistically significant at the 0.05 significance level (p-value = 0.06679). Hispanic/Latino students (SCORE_HISPANIC_LATINO) displayed a positive correlation as

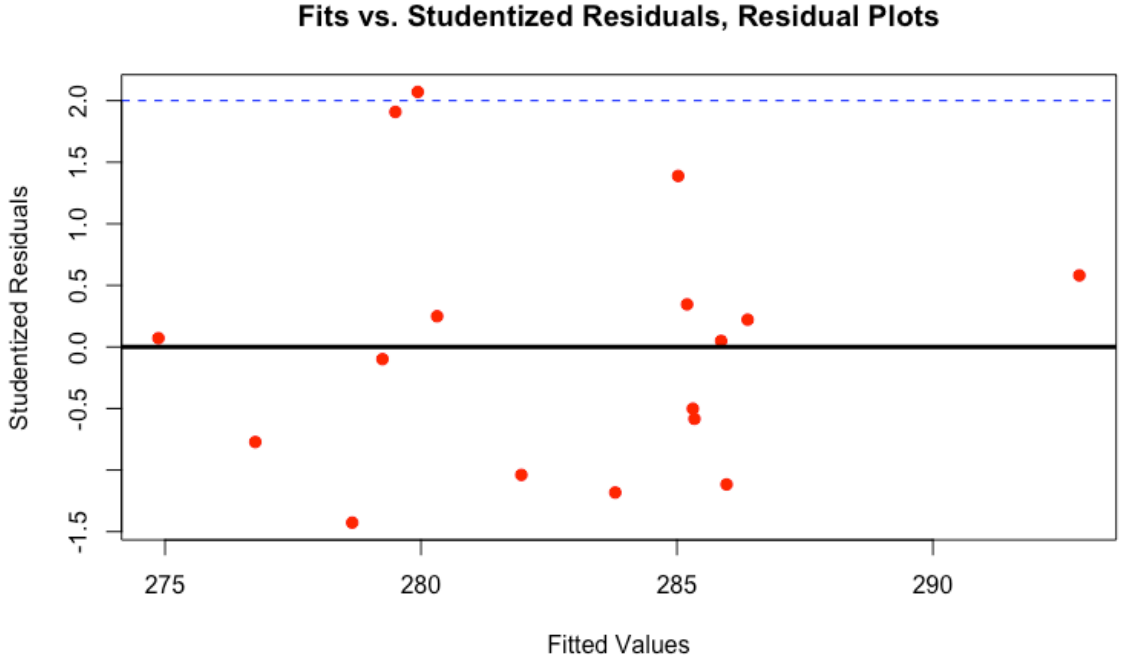
well (cor = 0.2962), which was statistically significant (p-value = 0.04091). Similarly, Black students (SCORE_BLACK) exhibited a positive correlation (cor = 0.4104), with statistical significance (p-value = 0.008523). White students (SCORE_WHITE) showed a stronger positive correlation (cor = 0.4674), which was highly statistically significant (p-value = 0.0005439). However, for students of multiple races (SCORE_MULTIPLE_RACES), the correlation was positive (cor = 0.3545) but not statistically significant at the 0.05 significance level (p-value = 0.08921). These results suggest varying degrees of correlation between average math scores and scores for different racial groups in 2015. Notably, statistically significant positive correlations are present for Hispanic/Latino, Black, and White racial groups, indicating that as average math scores increase, the scores for these racial groups tend to increase as well. However, the strength of these correlations varies among the different racial groups.

Multiple Regression

In this multiple regression analysis, we aim to explore the relationship between average math scores (AVG_MATH) and scores for various racial/ethnic groups in the 2015 dataset. The variables considered include scores for Native American, Hispanic/Latino, Black, White, and students of multiple races. By employing multiple linear regression, we can assess how each racial/ethnic group's scores contribute to the overall average math performance while controlling for other variables.

```
##
## Call:
## lm(formula = AVG_MATH ~ SCORE_NATIVE_AM + SCORE_HISPANIC_LATINO +
##     SCORE_BLACK + SCORE_WHITE + SCORE_MULTIPLE_RACES, data = States2015)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6571 -1.7633  0.1281  0.8017  5.0627
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         50.67927   60.97970   0.831 0.423599
## SCORE_NATIVE_AM      0.04005    0.07631   0.525 0.610153
## SCORE_HISPANIC_LATINO  0.64739    0.25953   2.494 0.029797 *
## SCORE_BLACK         -0.90759    0.30345  -2.991 0.012279 *
## SCORE_WHITE          1.19887    0.22401   5.352 0.000233 ***
## SCORE_MULTIPLE_RACES  -0.24617    0.14678  -1.677 0.121665
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.986 on 11 degrees of freedom
##   (34 observations deleted due to missingness)
## Multiple R-squared:  0.7613, Adjusted R-squared:  0.6528
## F-statistic: 7.018 on 5 and 11 DF,  p-value: 0.003554
```



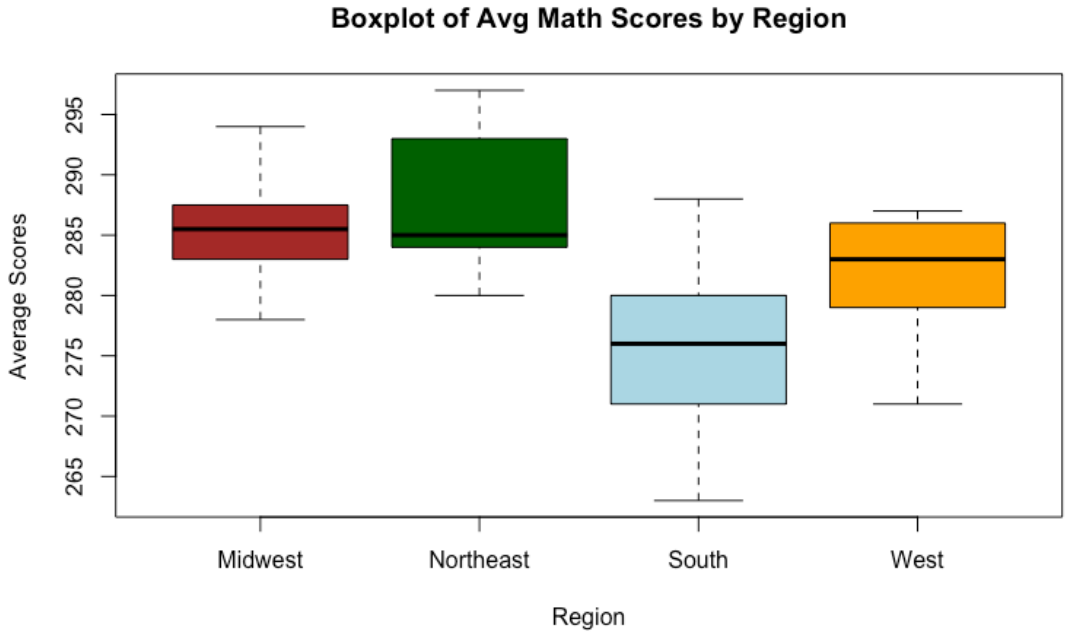


This multiple linear regression analysis examines the relationship between average math scores and scores for various racial/ethnic groups in 2015. The model demonstrates a strong explanatory power, with a multiple R-squared of 0.7613, indicating that approximately 76% of the variation in average math scores can be attributed to the predictor variables. Among the racial/ethnic groups, scores for Hispanic/Latino, Black, and White students emerge as significant predictors of average math scores,

with coefficients indicating their respective impacts. Notably, Hispanic/Latino scores positively influence average math scores, while Black scores negatively affect them. However, scores for Native American and multiple race students do not show significant predictive power. These findings underscore the significance of racial/ethnic disparities in educational outcomes and emphasize the importance of targeted interventions to address them. There is no clear indication of heteroskedasticity in the Fits vs Studentized Residuals Plot.

Scores by Region

Next, we wanted to look into if there was the region of the state has any significance on the average scores of the test. Before we got into that, lets look at boxplots of the scores by region to give us a general idea of the data.

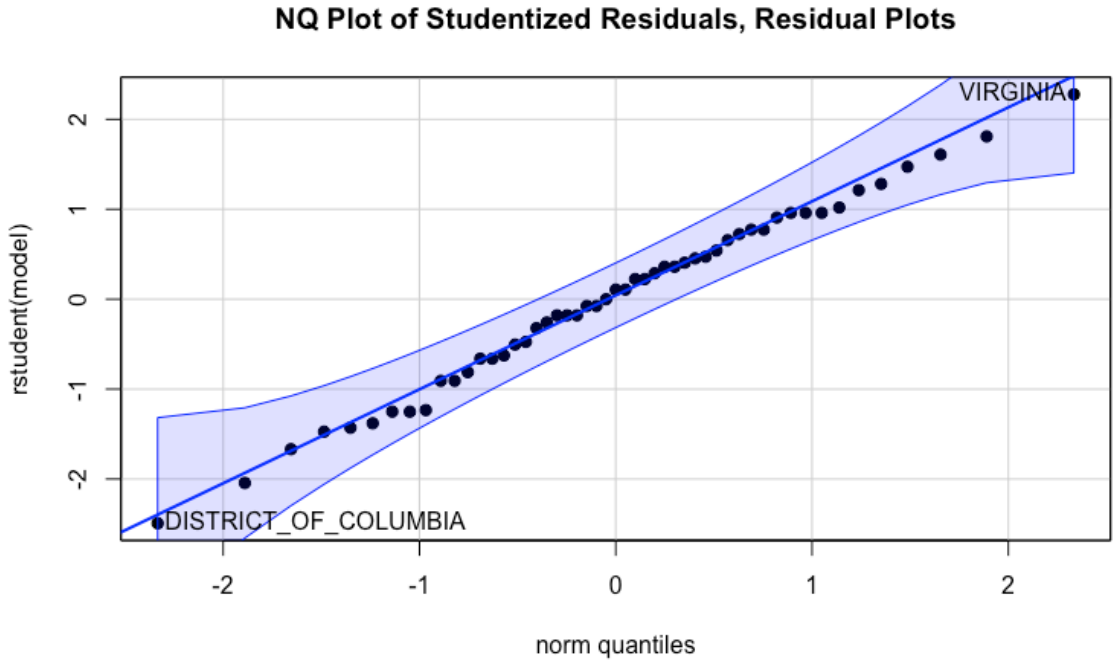


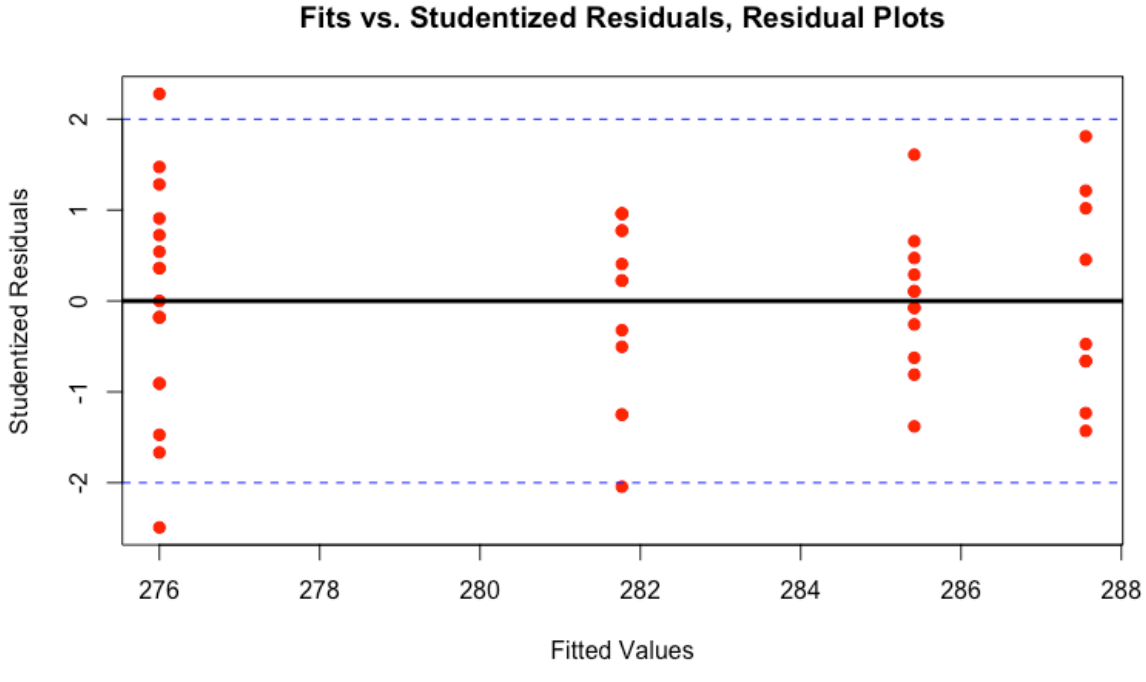
This box plot illustrates the 2015 average math score distribution across US regions. The Midwest exhibits the highest median score and the narrowest interquartile range (IQR), suggesting consistent performance. Conversely, the South displays the lowest median score and the widest IQR, indicating more variability. The Northeast falls between these extremes, while the West aligns closely with the Midwest. These differences hint at varying math performance, with further analysis needed to confirm and address potential factors.

ANOVA Test

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Region       3   1027   342.2   10.66 1.84e-05 ***
## Residuals   47   1509    32.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

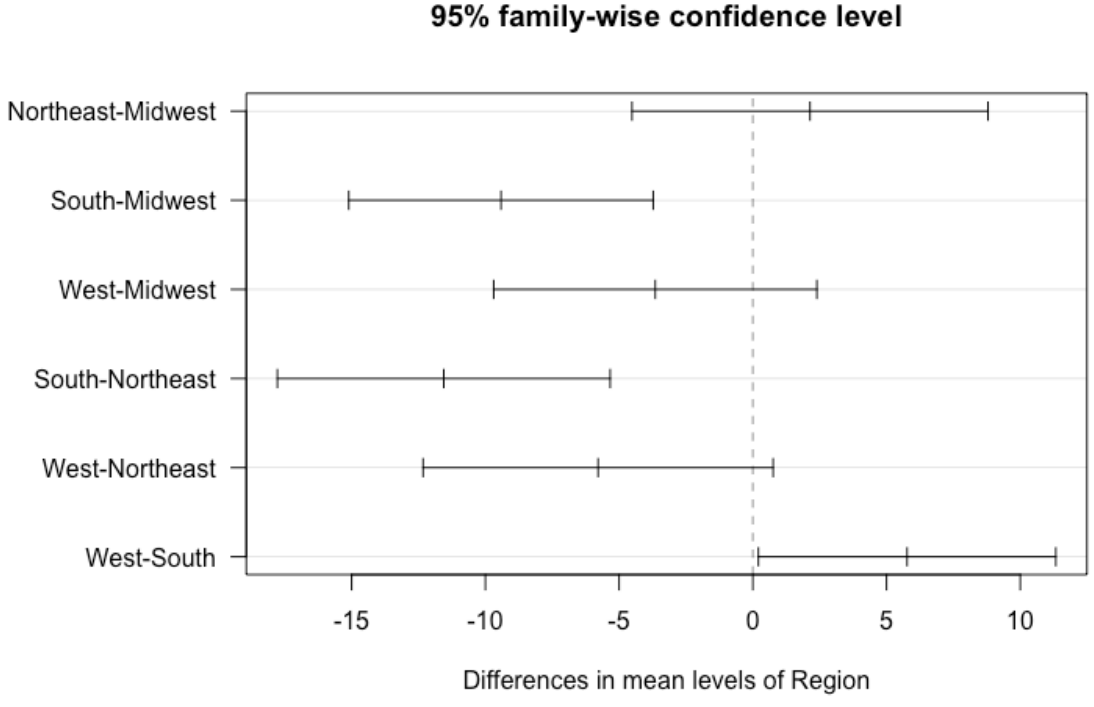
The ANOVA table (Image 1) reveals a significant difference in average math scores among the four regions (Midwest, Northeast, South, and West), as indicated by the F-value of 10.66 (p = 0.0000184). This suggests that the observed variations in scores are unlikely to have occurred by chance alone, supporting the rejection of the null hypothesis of no difference between regions. We'll look at this further using residual plots to get a more clear idea.





The normal Q-Q plot of studentized residuals shows a reasonably linear pattern, suggesting that the assumption of normality is generally met, albeit with a few potentially influential points. The residuals vs. fitted values plot shows a relatively random scatter around the horizontal line, indicating that the assumptions of linearity is reasonably satisfied, and there isn't any clear evidence of heteroskedasticity However, there are some mild deviations from these assumptions, as shown by a few points with larger residuals. Further investigation might be needed to account for these potential violations and to improve the model's fit. Let's use Tukey's HSD test to get a better idea.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = AVG_MATH ~ Region, data = States2015)
##
## $Region
##                       diff       lwr       upr     p adj
## Northeast-Midwest   2.138889  -4.5167900  8.7945678 0.8273003
## South-Midwest      -9.416667 -15.1075346 -3.7257987 0.0003425
## West-Midwest       -3.647436  -9.6897361  2.3948643 0.3842198
## South-Northeast   -11.555556 -17.7776350 -5.3334761 0.0000582
## West-Northeast     -5.786325 -12.3313753  0.7587257 0.1003165
## West-South          5.769231   0.2081516 11.3303100 0.0393329
```



The Tukey's multiple comparisons test further reveals specific pairs of regions with significantly different mean scores. The comparisons that included the South were all statistically significant as 0

doesn’t fall within any of those intervals. This can help us conclude that the average scores from the South were different(lower) than the other regions.

Conclusions and Summary

This project's thorough analysis of educational data from 1992 to 2015 revealed key insights into the U.S. education system, particularly the nuanced relationship between instructional spending and student math performance. Despite varying degrees of correlation in different years, our findings suggest that a higher proportion of budget allocated to instruction correlates with better math outcomes, a trend that became slightly stronger in 2015. Gender-specific analysis indicated a notable shift towards parity in math performance, with women showing increased scores over time. Additionally, regional disparities highlighted through ANOVA tests showed that students in the South performed consistently lower compared to those in the Midwest, which had the highest scores. Racial disparities were also evident, as regression analyses revealed that while Hispanic/Latino and Black students' scores were significant predictors of average math scores, the impact was varied—positively for Hispanic/Latino and negatively for Black students. These statistical insights provide a deeper understanding of the dynamics affecting educational efficacy and equity. While there have been improvements in scores from 1992 to 2015, there are still disparities that come from state expenditures, race, gender, and even region that need to be addressed. Overall, this study underscores the critical role of sustained and equitable investment in education to enhance student outcomes and reduce disparities, setting a foundation for more comprehensive policy reforms aimed at fostering an educational system that's inclusive and effective.