The Correlation Between Parent Education and Child Academics

Brendan Collins, Andrew Zheng, Sid Su, Eshita Maheshwari Section 0133

Dec. 5, 2021

Contents

Introduction

Standardized testing is commonplace for students across the world. It is said to be a fair means of analysis for student aptitude and is used in various countries for university admissions, job selection, and other equally important decisions. This project will focus on the correlations between student test scores for the math, writing, and reading sections of a standardized test given various demographic factors. The primary demographic explored was parental education and how it correlates with student scores across all sections of the exam. We expect to see a strong positive correlation between student test scores and parent level of education, with higher levels of education correlating to higher overall test scores. This information is important as a strong correlation can provide a basis for research into the effects of parental education on their children's academic success. Additionally, it can help to level the playing field for test takers from different backgrounds. In this project, we explored a dataset of student test scores, which includes demographic information about the test takers' backgrounds taken from a school in the United States. Personal information about the students, school and exams were omitted for privacy reasons. The dataset can be found here: https://www.kaggle.com/spscientist/students-performance-in-exams.

Analysis

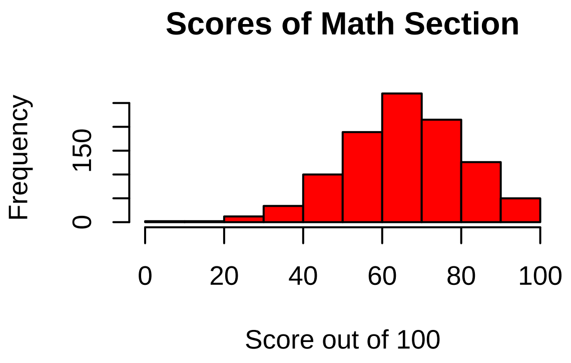
In order to access the data, first download the .csv file from the link provided above. Then, change the directory in the read.csv function below to wherever the .csv was downloaded.

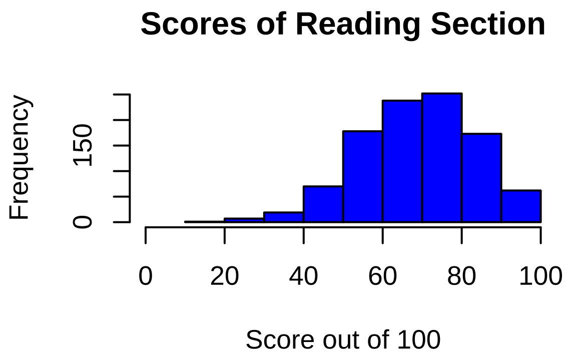
```r
grades_data <- read.csv("StudentsPerformance.csv")
```

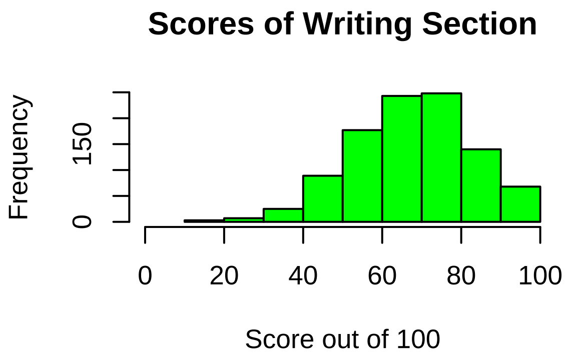
Exam Score Histograms

```r
par(mfcol=c(2,2))
math_scores <- sort(grades_data$math.score)
reading_scores <- sort(grades_data$reading.score)
```

```r
writing_scores <- sort(grades_data$writing.score)
hist(math_scores,
     main="Scores of Math Section",
     xlab="Score out of 100",
     xlim=c(0,100),
     col="red")
hist(writing_scores,
     main="Scores of Writing Section",
     xlab="Score out of 100",
     xlim=c(0,100),
     col="green")
hist(reading_scores,
     main="Scores of Reading Section",
     xlab="Score out of 100",
     xlim=c(0,100),
     col="blue")
```







The histograms of exam scores reveal many things about the data set. Overall, scores tended to be the highest in the writing section with the lowest scores in the math section. The math section had the largest range of scores due to the existence of some outlier scores. Additionally, the histogram is unimodal and evenly distributed. This shape allows for sample mean and median to be useful statistics for interpreting the data, as well as allowing a normal distribution to be used to analyze the data further.

```r
math_mean <- mean(math_scores)
writing_mean <- mean(writing_scores)
reading_mean <- mean(reading_scores)
```

```
math_median <- median(math_scores)
reading_median <- median(reading_scores)
writing_median <- median(writing_scores)
print(paste("Math Section Mean: ",math_mean, " Median: ", math_median))
```

```
## [1] "Math Section Mean:  66.089  Median:  66"
```

```
print(paste("Math Section Writing: ",writing_mean, " Median: ", writing_median))
```

```
## [1] "Math Section Writing:  68.054  Median:  69"
```

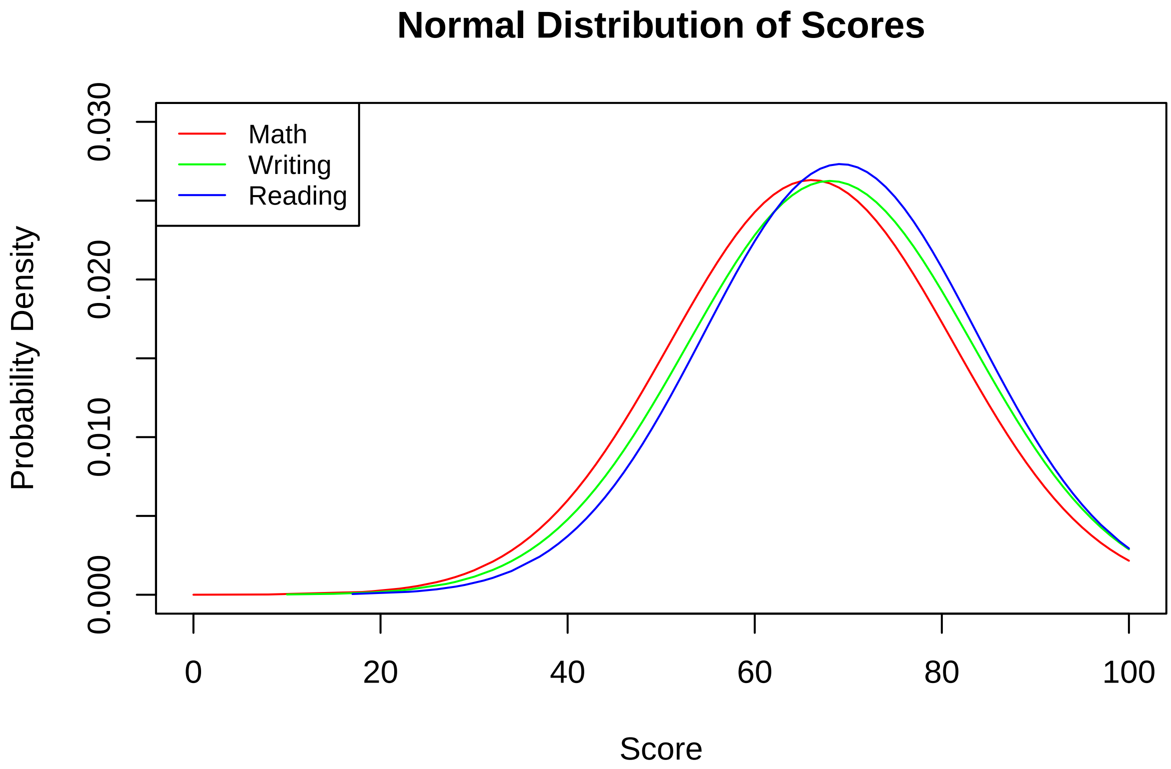
```
print(paste("Math Section Reading: ",reading_mean, " Median: ", reading_median))
```

```
## [1] "Math Section Reading:  69.169  Median:  70"
```

These statistics corroborate the information demonstrated by the histograms. Overall the mean and median of each section tended to be very similar.The reading section had the highest average score along with the highest median score. It was noted that the outliers in the math section do not affect the mean significantly due to the large size of the data. This is evident from the similarities between the mean and median for the section.

Normal Distribution Of Exam Scores

```
math_sd <- sd(math_scores)
plot(math_scores, dnorm(math_scores, mean = math_mean, math_sd),
     col = "red", type = "l",
     main = "Normal Distribution of Scores", ylim = c(0,.03),
     xlab = "Score", ylab = "Probability Density")
reading_sd <- sd(reading_scores)
writing_sd <- sd(writing_scores)
writing_norm <- dnorm(writing_scores, mean = writing_mean, writing_sd)
lines(writing_scores, writing_norm, col = "green")
reading_norm <- dnorm(reading_scores, mean = reading_mean, reading_sd)
lines(reading_scores, reading_norm, col = "blue")
legend(x = "topleft", legend=c("Math", "Writing", "Reading"),
       col=c("red", "green", "blue"), lty=1, cex=0.8)
```



The normal distributions of scores across the sections were similar. This was anticipated from the histograms. The normal distributions corroborate that the math section has a significantly higher range than the reading and writing sections. Additionally, the probability density of scoring the mean was the highest for the reading section while the probability density of scoring the mean for the math and writing sections was lower but very similar. This information indicates that students are more likely to score closer to the mean for the reading section than for other sections.

Normal Distributions By Parental Education

There are six levels of parental education provided in the demographics of the data. The six levels are "master's degree", "bachelor's degree", "associate's degree", "some college", "high school", and "some high school." By finding the normal distributions of student scores across sections, the correlation between parent education and child academic performance can be found.

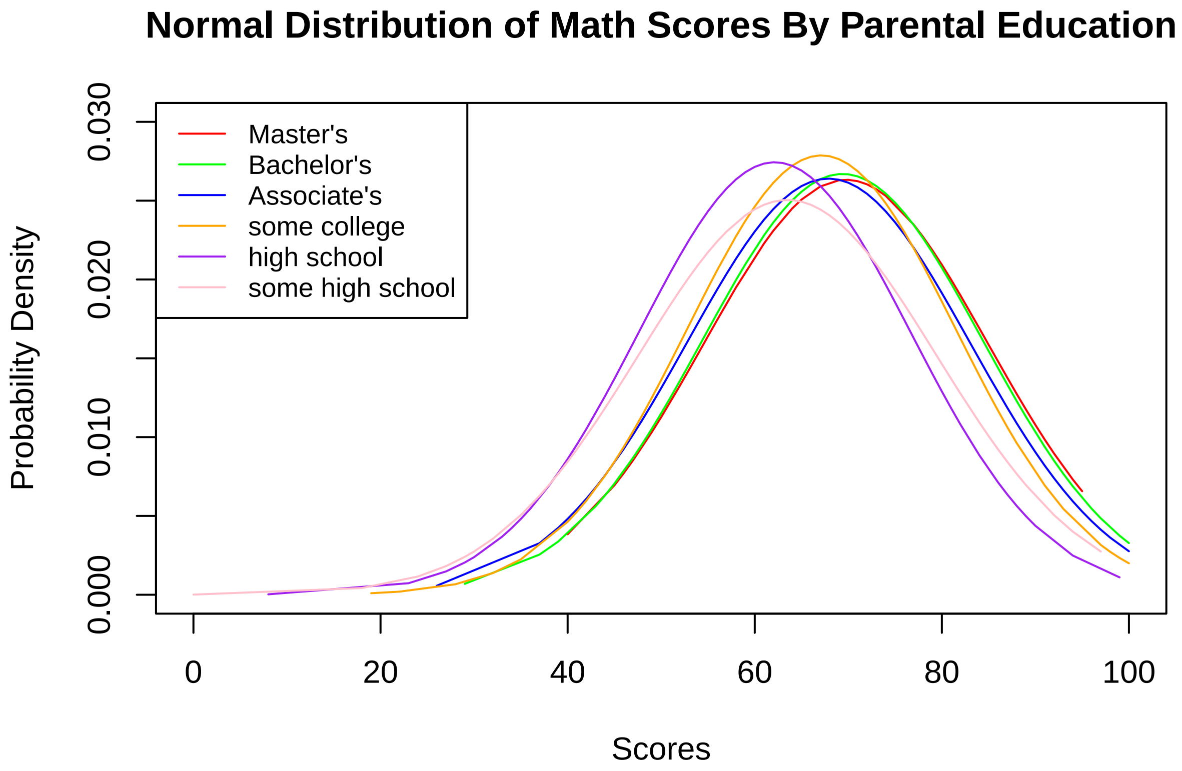
```
masters <- dplyr::filter(grades_data, grades_data$parental.level.of.education == "master's degree")
masters_math <- sort(masters$math.score)
masters_writing <- sort(masters$writing.score)
masters_reading <- sort(masters$reading.score)
bachelors <- dplyr::filter(grades_data, grades_data$parental.level.of.education == "bachelor's degree")
bachelors_math <- sort(bachelors$math.score)
bachelors_writing <- sort(bachelors$writing.score)
bachelors_reading <- sort(bachelors$reading.score)
associates <- dplyr::filter(grades_data, grades_data$parental.level.of.education == "associate's degree
associates_math <- sort(associates$math.score)
associates_writing <- sort(associates$writing.score)
associates_reading <- sort(associates$reading.score)
some_college <- dplyr::filter(grades_data, grades_data$parental.level.of.education == "some college")
```

```r
some_college_math <- sort(some_college$math.score)
some_college_writing <- sort(some_college$writing.score)
some_college_reading <- sort(some_college$reading.score)
high_school <- dplyr::filter(grades_data, grades_data$parental.level.of.education == "high school")
high_school_math <- sort(high_school$math.score)
high_school_writing <- sort(high_school$writing.score)
high_school_reading <- sort(high_school$reading.score)
some_high_school <- dplyr::filter(grades_data, grades_data$parental.level.of.education == "some high sc
some_high_school_math <- sort(some_high_school$math.score)
some_high_school_writing <- sort(some_high_school$writing.score)
some_high_school_reading <- sort(some_high_school$reading.score)
```

Distribution of Math Scores by Parental Education

```r
masters_math_mu <- mean(masters_math)
bachelors_math_mu <- mean(bachelors_math)
associates_math_mu <- mean(associates_math)
some_college_math_mu <- mean(some_college_math)
high_school_math_mu <- mean(high_school_math)
some_high_school_math_mu <- mean(some_high_school_math)
masters_math_sd <- sd(masters_math)
bachelors_math_sd <- sd(bachelors_math)
associates_math_sd <- sd(associates_math)
some_college_math_sd <- sd(some_college_math)
high_school_math_sd <- sd(high_school_math)
some_high_school_math_sd <- sd(some_high_school_math)


plot(masters_math, dnorm(masters_math, mean = masters_math_mu, masters_math_sd),
     col = "red", type = "l", xlim = c(0,100), ylim = c(0,.03),
     main = "Normal Distribution of Math Scores By Parental Education",
     xlab = "Scores", ylab = "Probability Density")
bachelors_norm <- dnorm(bachelors_math, mean = bachelors_math_mu, bachelors_math_sd)
lines(bachelors_math, bachelors_norm, col = "green")
associates_norm <- dnorm(associates_math, mean = associates_math_mu, associates_math_sd)
lines(associates_math, associates_norm, col = "blue")
some_college_norm <- dnorm(some_college_math, mean = some_college_math_mu, some_college_math_sd)
lines(some_college_math, some_college_norm, col = "orange")
high_school_norm <- dnorm(high_school_math, mean = high_school_math_mu, high_school_math_sd)
lines(high_school_math, high_school_norm, col = "purple")
some_high_school_norm <- dnorm(some_high_school_math, mean = some_high_school_math_mu, some_high_school_
lines(some_high_school_math, some_high_school_norm, col = "pink")
legend(x = "topleft", legend=c("Master's", "Bachelor's", "Associate's","some college","high school", "s
       col=c("red", "green", "blue", "orange", "purple", "pink"), lty=1, cex=0.8)
```

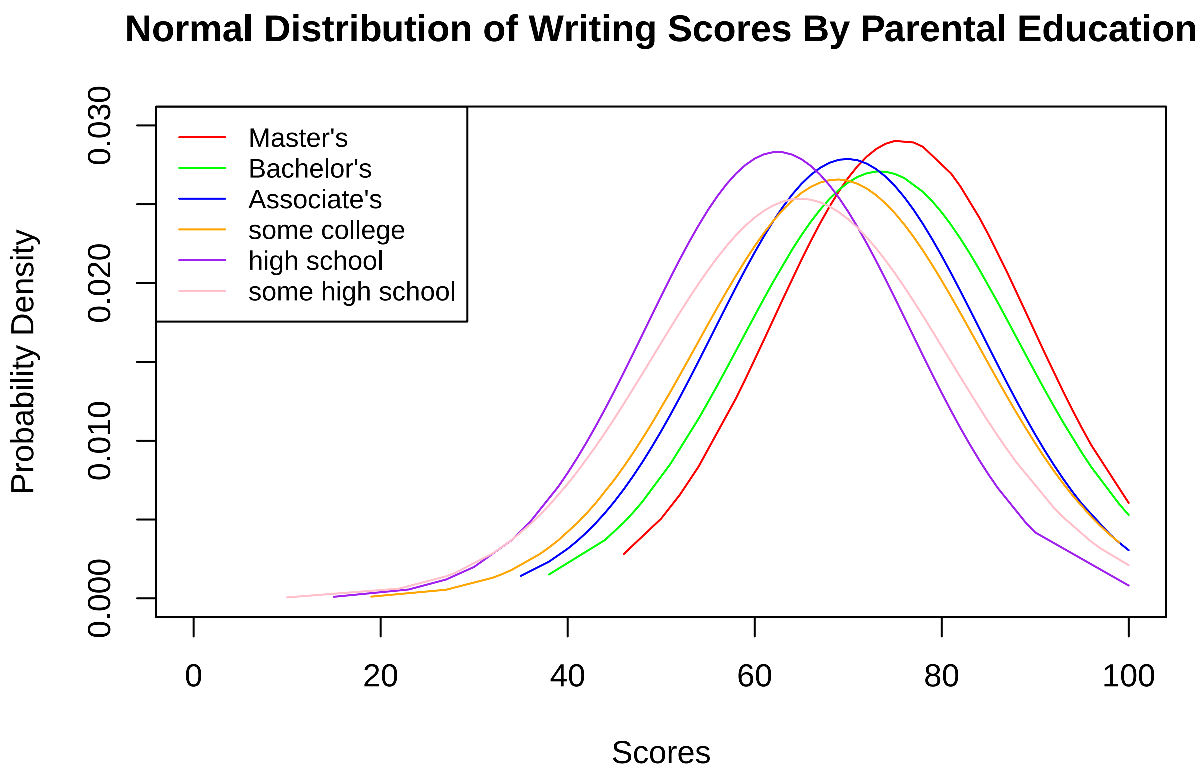


The normal distributions for math scores by parental education show a disparity in scores. Students with parental education levels “some high school” and “high school” tended to score 6-7% lower than students whose parents had reached collegiate education. The mean score for those with parental education level “high school” was the lowest at 62.14 while “master’s degree” was the highest at 69.75. Additionally, there is very little difference between the normal distribution for “associate’s degree”, “bachelor’s degree”, and “master’s degree” indicating that these levels of education correlate with student math knowledge somewhat. Overall, it was observed that a higher level of parental education was correlated with better on average scores for the math section of the exam.

Distribution of Writing Scores by Parental Education

```r
masters_writing_mu <- mean(masters_writing)
bachelors_writing_mu <- mean(bachelors_writing)
associates_writing_mu <- mean(associates_writing)
some_college_writing_mu <- mean(some_college_writing)
high_school_writing_mu <- mean(high_school_writing)
some_high_school_writing_mu <- mean(some_high_school_writing)
masters_writing_sd <- sd(masters_writing)
bachelors_writing_sd <- sd(bachelors_writing)
associates_writing_sd <- sd(associates_writing)
some_college_writing_sd <- sd(some_college_writing)
high_school_writing_sd <- sd(high_school_writing)
some_high_school_writing_sd <- sd(some_high_school_writing)


plot(masters_writing, dnorm(masters_writing, mean = masters_writing_mu, masters_writing_sd),
     col = "red", type = "l", xlim = c(0,100), ylim = c(0,.03),
     main = "Normal Distribution of Writing Scores By Parental Education",
     xlab = "Scores", ylab = "Probability Density")
bachelors_norm <- dnorm(bachelors_writing, mean = bachelors_writing_mu, bachelors_writing_sd)
lines(bachelors_writing, bachelors_norm, col = "green")
associates_norm <- dnorm(associates_writing, mean = associates_writing_mu, associates_writing_sd)
lines(associates_writing, associates_norm, col = "blue")
some_college_norm <- dnorm(some_college_writing, mean = some_college_writing_mu, some_college_writing_sd
lines(some_college_writing, some_college_norm, col = "orange")
high_school_norm <- dnorm(high_school_writing, mean = high_school_writing_mu, high_school_writing_sd)
lines(high_school_writing, high_school_norm, col = "purple")
some_high_school_norm <- dnorm(some_high_school_writing, mean = some_high_school_writing_mu, some_high_s
lines(some_high_school_writing, some_high_school_norm, col = "pink")
legend(x = "topleft", legend=c("Master's", "Bachelor's", "Associate's","some college","high school", "s
       col=c("red", "green", "blue", "orange", "purple", "pink"), lty=1, cex=0.8)
```

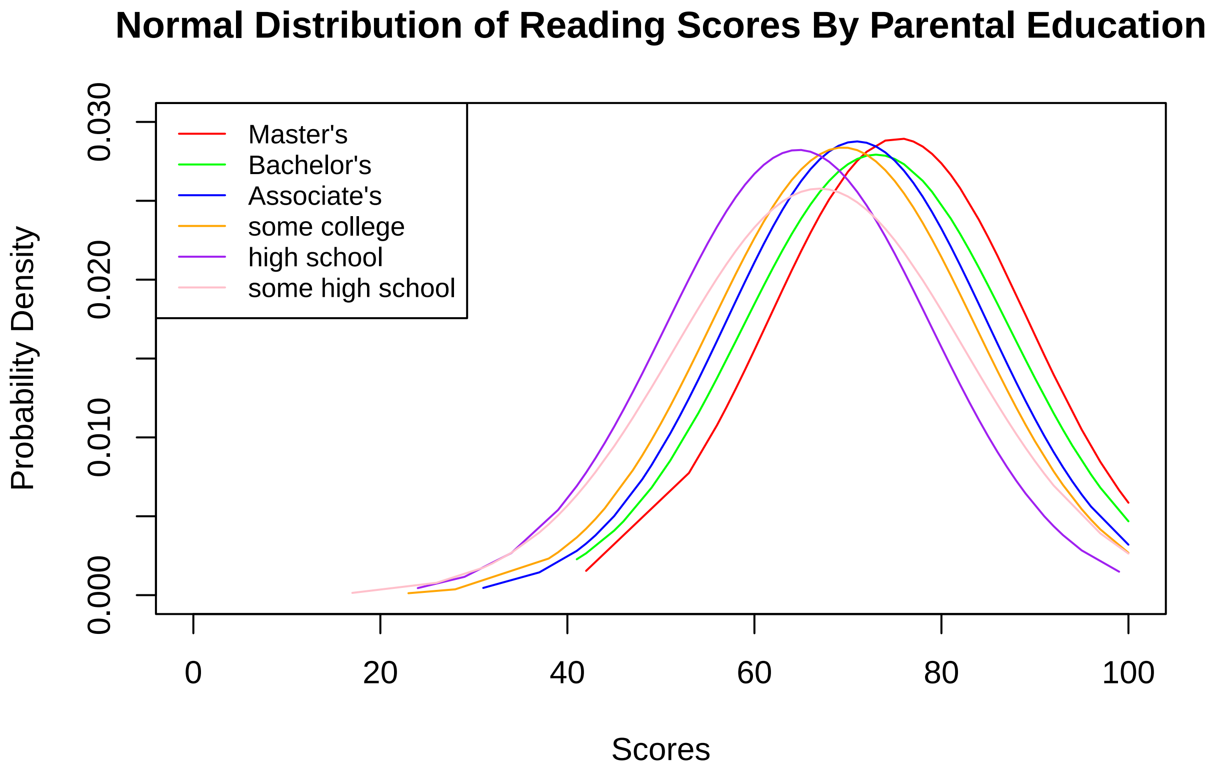


The normal distributions for the writing scores by parental education reveal the greatest disparity in scores. Students with parental education levels “some high school” and “high school” tended to score 12-13% lower than students whose parents had reached collegiate education. The mean score for those with parental education levels “high school” was the lowest at 62.45 while “master’s degree” was the highest at 75.68. Additionally, there is a greater difference between the normal distribution for “associate’s” degree, “bachelor’s degree”, and “master’s degree” than there was for the math section. This indicates that parental level of education correlates more strongly with students’ writing ability than their math ability. Overall, it was observed that a higher level of parental education was strongly correlated with better scores on the writing section of the exam.

Distribution of Reading Scores by Parental Education

```
masters_reading_mu <- mean(masters_reading)
bachelors_reading_mu <- mean(bachelors_reading)
associates_reading_mu <- mean(associates_reading)
some_college_reading_mu <- mean(some_college_reading)
high_school_reading_mu <- mean(high_school_reading)
some_high_school_reading_mu <- mean(some_high_school_reading)
masters_reading_sd <- sd(masters_reading)
bachelors_reading_sd <- sd(bachelors_reading)
associates_reading_sd <- sd(associates_reading)
some_college_reading_sd <- sd(some_college_reading)
high_school_reading_sd <- sd(high_school_reading)
some_high_school_reading_sd <- sd(some_high_school_reading)


plot(masters_reading, dnorm(masters_reading, mean = masters_reading_mu, masters_reading_sd),
     col = "red", type = "l", xlim = c(0,100), ylim = c(0,.03),
     main = "Normal Distribution of Reading Scores By Parental Education",
     xlab = "Scores", ylab = "Probability Density")
bachelors_norm <- dnorm(bachelors_reading, mean = bachelors_reading_mu, bachelors_reading_sd)
lines(bachelors_reading, bachelors_norm, col = "green")
associates_norm <- dnorm(associates_reading, mean = associates_reading_mu, associates_reading_sd)
lines(associates_reading, associates_norm, col = "blue")
some_college_norm <- dnorm(some_college_reading, mean = some_college_reading_mu, some_college_reading_s
lines(some_college_reading, some_college_norm, col = "orange")
high_school_norm <- dnorm(high_school_reading, mean = high_school_reading_mu, high_school_reading_sd)
lines(high_school_reading, high_school_norm, col = "purple")
some_high_school_norm <- dnorm(some_high_school_reading, mean = some_high_school_reading_mu, some_high_
lines(some_high_school_reading, some_high_school_norm, col = "pink")
legend(x = "topleft", legend=c("Master's", "Bachelor's", "Associate's","some college","high school", "s
       col=c("red", "green", "blue", "orange", "purple", "pink"), lty=1, cex=0.8)
```



The normal distributions for the reading scores given different levels of parental education were similar to that of the math section. Students with parental education levels “some high school” and “high school” tended to score 10-11% lower than students whose parents had reached collegiate education. The mean score for those with parental education level “high school” was the lowest at 64.08 while “Master’s” was the highest at 75.37. Again, there was a more significant difference between the normal distribution for “associate’s degree”, “bachelor’s degree”, and “master’s degree” than there was for the math section, though not as much as the writing section. This indicates that parental level of education correlates more strongly with students’ reading ability than their math ability, but less strongly than with their writing ability. Overall, it was observed that a higher level of parental education was correlated with better scores on the writing section of the exam.

Conclusion

Our analysis of the data showed that students whose parents have higher education tended to score better. It’s important to note that we only analysed categorical aspects of the dataset, there are many other parameters which have been proven to have more of an impact on student exam performance that are hard to express or quantify, such as the student’s situation at home or students’ engagement in activities outside of school.

It’s also important to note that it is hard to make conclusions about our analysis without making assumptions. For example, it was found that students whose parents had higher education generally scored better than students whose parents had lower education. However, there is nothing that we can conclude from this statistic without making assumptions such as “parents with higher education are able to help the students more with their schoolwork” or “parents with higher education emphasize good education more compared to those with a lower degree of education” which fall outside of the scope of our data.

Future Work

If we had more time and tools at our disposal, we would try to expand on our findings qualitatively. In this study, we have identified correlations between different parameters. However, we cannot deduce anything further from this without making assumptions about the students.

We can expand the project to be more oriented towards the student. For example, from our findings, we were able to correlate that the students whose parents had a higher degree of education performed better on the exam compared to students whose parents achieved a lower degree of education. We can base our future work around discovering why students whose parents with higher education perform better than students whose parents have lower education.One of such explorations could be what parents of higher education do differently than parents with lower education regarding academics and standardized testing for their children? One way we can test this is by giving parents surveys where we ask questions on topics where we suspect there may be a difference, such as hours spent studying, money spent on students' education, emphasis on students' success in education, or families' role in their community.

Future work should primarily focus on understanding the factors at play during standardized testing. Identifying what causes disparities in standardized test scores is the best way to begin this process. Eliminating these disparities can help to make the standardized testing process more refined and a better means of aptitude analysis.