Final Project - Case Study

Hadley Dixon, Haleigh Cole, Daniel Baltrons

2022-12-15

Students Performance in Exams

The write-up - a knit.pdf file

“Components” Section Analysis

Statement of question/topic that we want to answer and what motivated you to study the question/topic

The guiding question for our final project is, “How do students’ backgrounds affect academic performance?” Our group imagined we were speaking to the President of USF, who instructed us to look into the disparities in test scores among students of different demographics. This research will help to identify and provide institutional support for students of different identities.

If applicable, what data will you analyze? Identify at least one source.

We will analyze the dataset titled, “Students Performance in Exams.” This dataset is from Kaggle. We chose this dataset because as students, much of our academic lives are centered around midterm and final exams. It’s important as a diverse Jesuit school that we understand how certain socioeconomic, racial, and gender identities impact academic performance so we can better address inequalities.

What challenges did you face in analyzing this data?

Before analyzing the data, the main challenge we encountered was finding context about the variables. There were many questions left unanswered. What is the grade level of the students? What kind of exam was taken? Which races/ethnicities belong to Groups A through E? Was the test prep course free? Knowing the grade level of the exam would have helped us understand what institutional resources the students had access to, which may account for the lack of/ prevalence of prep course completion and the need for standard or reduced/ free lunches. In addition, knowing the kind of exam would have informed us whether the population sampled is representative of a standard school population or primarily those in higher education. Contextualizing the potential financial burden of the test prep course could give insight into results involving parental level of education or race/ethnicity variables

In doing the analysis, the main challenge we faced was in learning the balance between aesthetics and comprehensibility of visualizations. There are lots of components that go into graphing data, especially when analyzing multiple variables as we did. There is a certain level of aesthetics needed to differentiate between variables or provide accurate labels, but on the other hand, creating too flashy of a graph can distract from the actual content. Secondly, there are important requirements of visualizations needed to ensure that the viewer is not being misled by the graph. This includes standardized y-axis, to ensure easy compatibility between variables, as well as units of measurement and labeled legends. This was a challenge we ran into, specifically when there were large variations from the average.

What packages were needed for this case study?

We worked with the functions and methods available in RStudio. We applied knowledge of importing and reading csv files, data frame creation and manipulation, subsetting data by criteria, and variance & standard deviations. Our analysis was conducted using the summary(), max(), apply(), mean(), median(), and sd() functions. Visualizations were developed with the barplot() function, specifically when graphing multiple barplots on one graph.

What did you learn from this experience?

Please refer to our Case Study below for our group's data analysi conclusions. Individual conclusions are broken down by question.

What more could you do with this project in the future?

In conclusion, what more could we do with this project in the future? Further analysis of this data could provide essential information to school administration, especially in higher level education as it is well established that there are many barriers to entry. This information would be incredibly important at a school like USF whose mission is toward social justice and equity. Understanding how one's race, socioeconomic status, and gender identities impact academic performance is the first step in leveling the playing field. Additionally, unrelated to this project, Daniel was interested in using the skils he learned in this project for a possible capstone in which he will analyze the performances of Latinx queens on the show RuPaul's Drag Race in relation to their type casting.

Case Study Analysis

Load the dataset

```
exams <- read.csv("/Users/hadleydixon/Desktop/StudentsPerformance.csv")
```

1. Take the mean, median, and standard deviation of students' reading, writing, and math scores.

The mean, median, and standard deviation of math scores are 66.089, 66, and 15.16308, respectively.

The mean, median, and standard deviation of reading scores are 69.169, 70, and 14.60019, respectively.

The mean, median, and standard deviation of writing scores are 68.054, 69, and 15.19566, respectively.

```
summary(exams[6:8])
```

```
##    math.score     reading.score    writing.score   
##  Min.   :  0.00   Min.   : 17.00   Min.   : 10.00  
##  1st Qu.: 57.00   1st Qu.: 59.00   1st Qu.: 57.75  
##  Median : 66.00   Median : 70.00   Median : 69.00  
##  Mean   : 66.09   Mean   : 69.17   Mean   : 68.05  
##  3rd Qu.: 77.00   3rd Qu.: 79.00   3rd Qu.: 79.00  
##  Max.   :100.00   Max.   :100.00   Max.   :100.00
```

```
apply(exams[,6:8], 2, sd)
```

```
##    math.score reading.score writing.score 
##      15.16308      14.60019      15.19566
```

2. Which subject (math, science, reading) had the highest performing scores?

The subject with the highest performing scores is reading.

```
means <- c(mean(exams$math.score), mean(exams$reading.score), mean(exams$writing.score))
medians <- c(median(exams$math.score), median(exams$reading.score), median(exams$writin
g.score))
standard.devations <- c(sd(exams$math.score), sd(exams$reading.score),sd(exams$writing.s
core))

score.stats <- data.frame("Subject" = c("Math", "Reading", "Writing"),
                          "Mean" = means,
                          "Median" = medians,
                          "Standard Deviation" = standard.devations)
```

```
max(score.stats$Subject[which.max(score.stats$Mean)]) # Reading
max(score.stats$Subject[which.max(score.stats$Median)]) # Reading
```

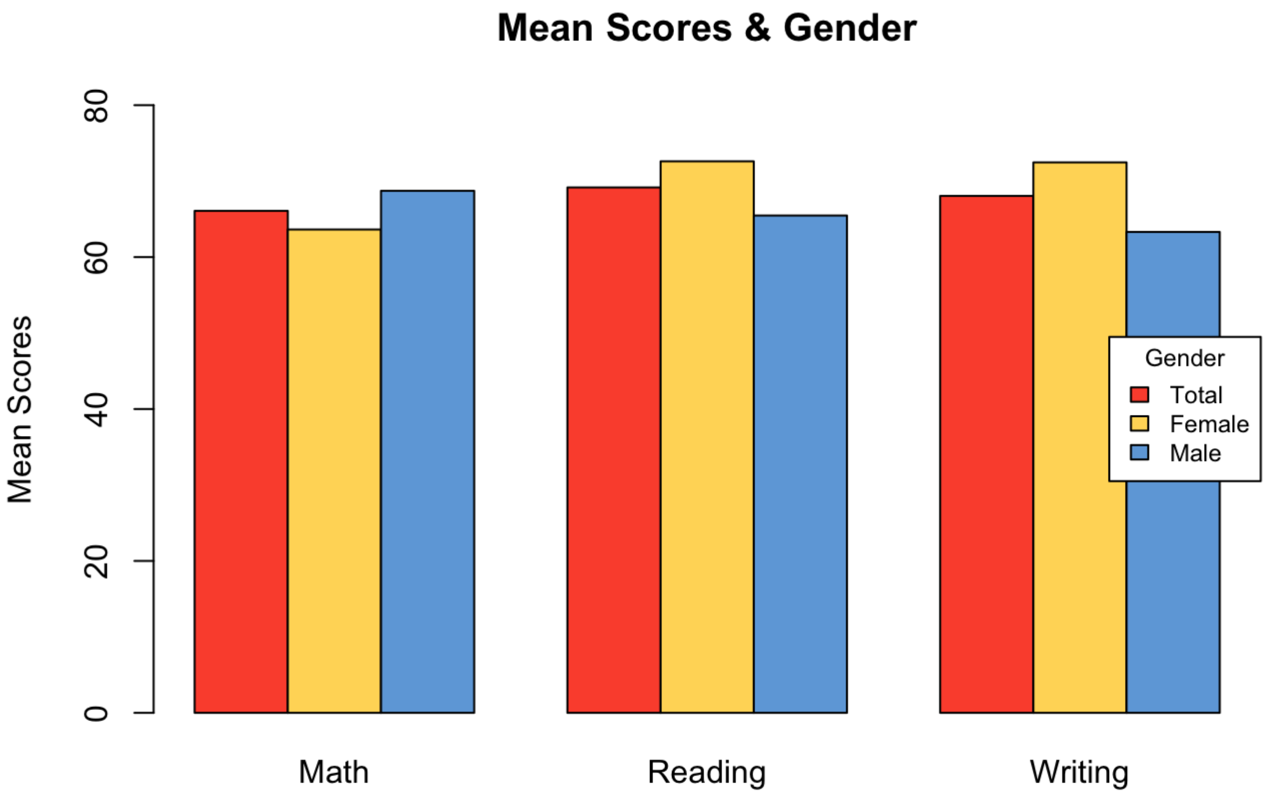
3. (a) How does (i) gender, (ii) race/ethnicity, and (iii) parental level of education impact the students' scores?

(i) Gender

On average males perform worse on reading and writing compared to females, yet outperform females in math. The degree of which females outperform males in reading and writing is greater than the difference that males outperform females in math. The score most impacted by gender is writing. Overall, the variance across genders is no more than 5 points from the total mean score, for all subjects.

If we were to infer from this data, one conclusion we could make is that careers in STEM have been historically dominated by men, while subjects like education and literature historically dominated by women. If children are socialized in accordance to this historical trend, cognitive biases such as confirmation bias come into play and where students may seek to conform to the identities society has given them.

```
female.data <- exams[exams$gender == "female",]
male.data <- exams[exams$gender == "male",]

gender.means <- data.frame("Gender" = c("Total", "Female", "Male"),
                           "Math" = c(mean(exams$math.score), mean(female.data$math.score), mean(male.data$math.score)),
                           "Reading" = c(mean(exams$reading.score), mean(female.data$reading.score), mean(male.data$reading.score)),
                           "Writing" = c(mean(exams$writing.score), mean(female.data$writing.score), mean(male.data$writing.score)))

gender.data <- data.frame(gender.means$Math, gender.means$Reading, gender.means$Writing)

barplot(as.matrix(gender.data),
        main="Mean Scores & Gender",
        ylab="Mean Scores",
        names.arg = c("Math", "Reading", "Writing"),
        col = c("#fc543a", "#ffd966", "#6fa8dc"),
        beside=TRUE,
        ylim = c(0,80),
        legend.text = c("Total", "Female", "Male"),
        args.legend = list(title = "Gender", x = "right", bty = "o", cex = .75))
```

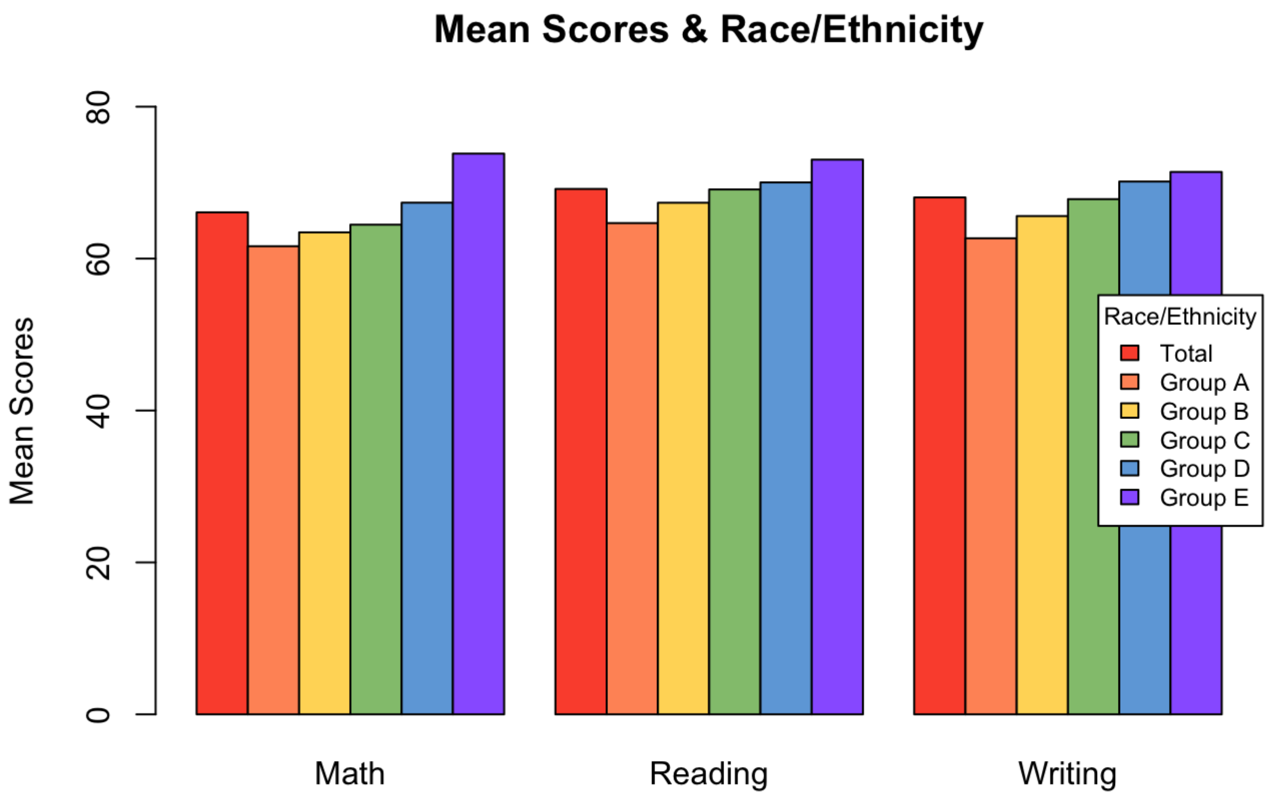


(ii) Race/Ethnicity

On average, Groups A to E performs sequentially higher than the previous group, with Group A performing the worse and Group E perfoming the best. Group E outperforms all other races/ethnicities as well as the total mean score in every subject. Conversely, Group A performs the worst of all groups in every subject, as well as under performs compared to the total mean scores. The score most impacted by race/ethnicity is math. Overall, the variance across race/ethnicity is no more than 7 points from the total mean score, for all subjects.

It is difficult to draw from this data because the dataset had no attribute as to which race/ethnicity belongs to each Group. However, what we can say is that more support of students from Group A is needed.

```
groupA <- exams[exams$race.ethnicity == "group A",]
groupB <- exams[exams$race.ethnicity == "group B",]
groupC <- exams[exams$race.ethnicity == "group C",]
groupD <- exams[exams$race.ethnicity == "group D",]
groupE <- exams[exams$race.ethnicity == "group E",]

race.means <- data.frame("Race/Ethncity" = c("Total", "Group A", "Group B", "Group C",
"Group D", "Group E"),
                           "Math" = c(mean(exams$math.score), mean(groupA$math.score), m
ean(groupB$math.score), mean(groupC$math.score), mean(groupD$math.score), mean(groupE$ma
th.score)),
                           "Reading" = c(mean(exams$reading.score), mean(groupA$reading.
score), mean(groupB$reading.score), mean(groupC$reading.score), mean(groupD$reading.scor
e), mean(groupE$reading.score)),
                           "Writing" = c(mean(exams$writing.score), mean(groupA$writing.
score), mean(groupB$writing.score), mean(groupC$writing.score), mean(groupD$writing.scor
e), mean(groupE$writing.score)))

race.data <- data.frame(race.means$Math, race.means$Reading, race.means$Writing)

barplot(as.matrix(race.data),
        main="Mean Scores & Race/Ethnicity",
        ylab="Mean Scores",
        names.arg = c("Math", "Reading", "Writing"),
        col = c("#fc543a", "#ff9566", "#ffd966", "#93c47d", "#6fa8dc", "#9966ff"),
        beside = TRUE,
        ylim = c(0,80),
        legend.text = c("Total", "Group A", "Group B", "Group C", "Group D", "Group E"),
        args.legend = list(title = "Race/Ethnicity", x = "right", bty = "o", cex = .75))
```

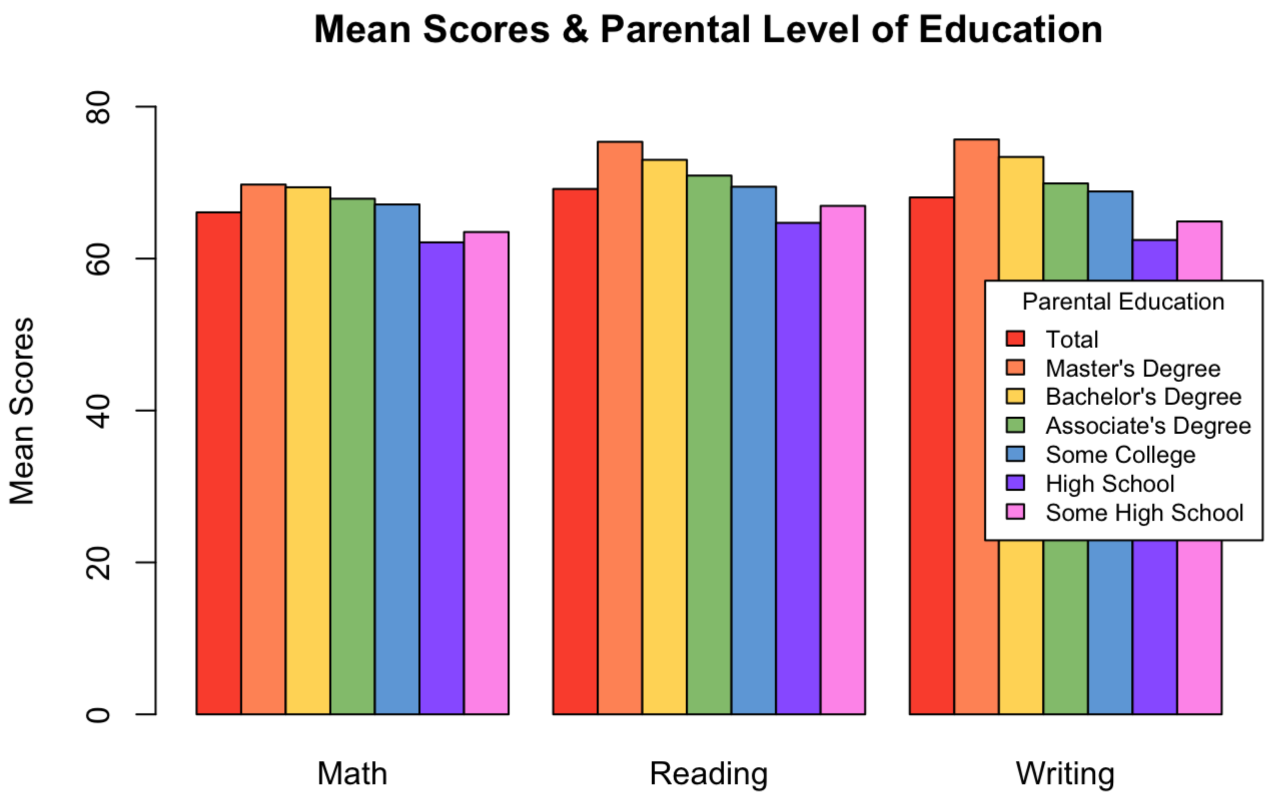


(iii) Parental level of education

On average, students whose parents have a master's degree outperform students with parents of other education levels. This difference is most significant in students' performance on writing, and least significant in students' performance on math. Students who have parents who completed high school perform the worst in all subjects compared to students with parents of other education levels. The score most impacted by parental level of education is writing. Overall, the variance across parental level of education is no more than 7 points from the total mean score, for all subjects.

A logical conclusion from this data is that students whose parents have a high school degree or less need extra support compared to students whose parents have some form of higher eductaion.

```
masters <- exams[exams$parental.level.of.education == "master's degree",]
bachelors <- exams[exams$parental.level.of.education == "bachelor's degree",]
associates <- exams[exams$parental.level.of.education == "associate's degree",]
somecollege <- exams[exams$parental.level.of.education == "some college",]
hs <- exams[exams$parental.level.of.education == "high school",]
somehs <- exams[exams$parental.level.of.education == "some high school",]

education.means <- data.frame("Education" = c("Total", "Master's Degree", "Bachelor's Degree", "Associate's Degree", "Some College", "High School", "Some High School"),
                           "Math" = c(mean(exams$math.score), mean(masters$math.score), mean(bachelors$math.score), mean(associates$math.score), mean(somecollege$math.score), mean(hs$math.score), mean(somehs$math.score)),
                           "Reading" = c(mean(exams$reading.score), mean(masters$reading.score), mean(bachelors$reading.score), mean(associates$reading.score), mean(somecollege$reading.score), mean(hs$reading.score), mean(somehs$reading.score)),
                           "Writing" = c(mean(exams$writing.score), mean(masters$writing.score), mean(bachelors$writing.score), mean(associates$writing.score), mean(somecollege$writing.score), mean(hs$writing.score), mean(somehs$writing.score)))

education.data <- data.frame(education.means$Math, education.means$Reading, education.means$Writing)

barplot(as.matrix(education.data),
        main="Mean Scores & Parental Level of Education",
        ylab="Mean Scores",
        names.arg = c("Math", "Reading", "Writing"),
        col = c("#fc543a", "#ff9566", "#ffd966", "#93c47d", "#6fa8dc", "#9966ff", "#ff9bec"),
        beside = TRUE,
        ylim = c(0,80),
        legend.text = c("Total", "Master's Degree", "Bachelor's Degree", "Associate's Degree", "Some College", "High School", "Some High School"),
        args.legend = list(title = "Parental Education", x = "right", bty = "o", cex = .75))
```



3. (b) For each subject, which variable has the greatest impact on the total mean score?

For math, race/ethnicity is the most significant variable that impacts on the total mean score.

For reading, parental level of education is the most significant variable that impacts on the total mean score.

For writing, gender is the most significant variable that impacts on the total mean score.

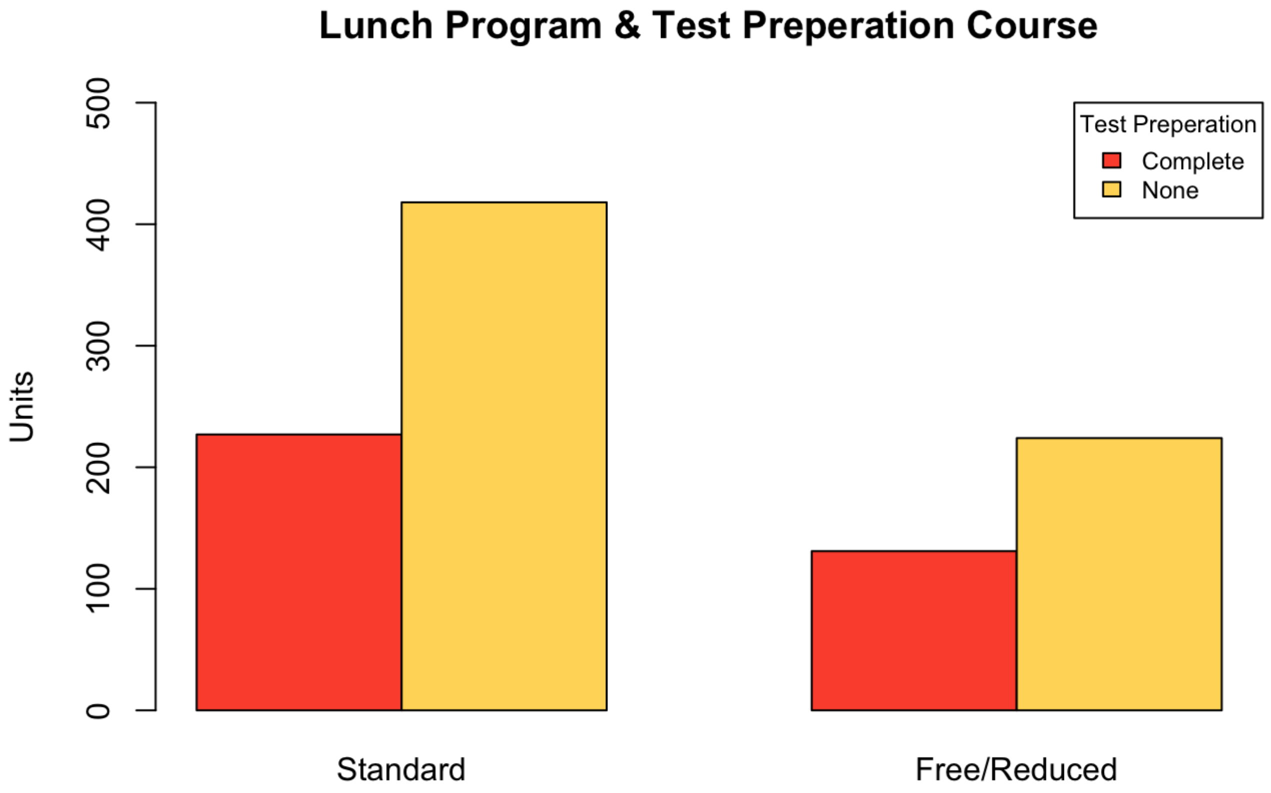
```
# Math
sd(c(mean(exams$math.score), mean(female.data$math.score), mean(male.data$math.score)))
# 2.548056
sd(c(mean(exams$math.score), mean(groupA$math.score), mean(groupB$math.score), mean(grou
pC$math.score), mean(groupD$math.score), mean(groupE$math.score))) # 4.26333
sd(c(mean(exams$math.score), mean(masters$math.score), mean(bachelors$math.score), mean
(associates$math.score), mean(somecollege$math.score), mean(hs$math.score), mean(somehs
$math.score))) # 2.869044

# Reading
sd(c(mean(exams$reading.score), mean(female.data$reading.score), mean(male.data$reading.
score))) # 3.56831
sd(c(mean(exams$reading.score), mean(groupA$reading.score), mean(groupB$reading.score),
 mean(groupC$reading.score), mean(groupD$reading.score), mean(groupE$reading.score))) #
 2.781682
sd(c(mean(exams$reading.score), mean(masters$reading.score), mean(bachelors$reading.scor
e), mean(associates$reading.score), mean(somecollege$reading.score), mean(hs$reading.sco
re), mean(somehs$reading.score))) # 3.584783

# Writing
sd(c(mean(exams$writing.score), mean(female.data$writing.score), mean(male.data$writing.
score))) # 4.578978
sd(c(mean(exams$reading.score), mean(groupA$reading.score), mean(groupB$reading.score),
 mean(groupC$reading.score), mean(groupD$reading.score), mean(groupE$reading.score))) #
 2.781682
sd(c(mean(exams$writing.score), mean(masters$writing.score), mean(bachelors$writing.scor
e), mean(associates$writing.score), mean(somecollege$writing.score), mean(hs$writing.sco
re), mean(somehs$writing.score))) # 4.570421
```

4. Does lunch program have an effect on whether a student takes the test prep course?

No, lunch program does not have an effect on whether a student takes the test prep course. There were more students who has a standard lunch program than students who had a free/reduced lunch program, but the distribution across test preperation options were relatively similar.

We asked this question to gain insight into if the test prep course was free or required purchase. A correlation between students on a standard lunch program and completion of a test prep course would indicate that there might be a price associated with the course, however the data shows this is not the case.

```
completed.data <- exams[exams$test.preparation.course == "completed",]
none.data <- exams[exams$test.preparation.course == "none",]

lunch.testprep <- data.frame("Test Preperation Course" = c("Completed", "None"),
                        "Standard" = c(nrow(completed.data[completed.data$lunch == "stan
dard",]), nrow(none.data[none.data$lunch == "standard",])),
                        "Free/Reduced" = c(nrow(completed.data[completed.data$lunch ==
"free/reduced",]), nrow(none.data[none.data$lunch == "free/reduced",])))

lunch.testprep.data <- data.frame(lunch.testprep$Standard, lunch.testprep$Free.Reduced)

barplot(as.matrix(lunch.testprep.data),
        main="Lunch Program & Test Preperation Course",
        ylab="Units",
        names.arg = c("Standard", "Free/Reduced"),
        col = c("#fc543a", "#ffd966"),
        beside=TRUE,
        ylim = c(0,500),
        legend.text = c("Complete", "None"),
        args.legend = list(title = "Test Preperation", x = "topright", bty = "o", cex =
.75))
```

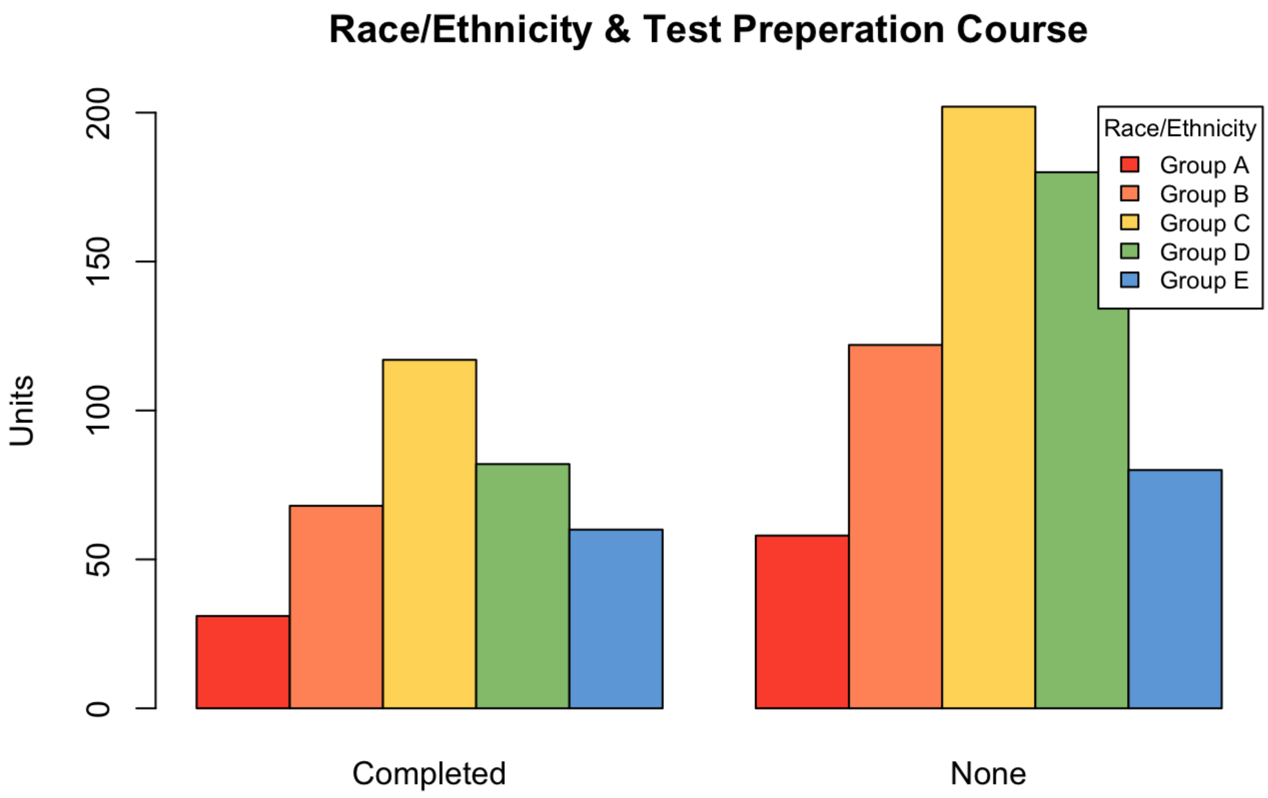


5. Does race/ethnicity have an effect on whether a student takes the test prep course?

No, race/ethnicity does not have an effect on whether a student takes the test prep course. There were more students who did not complete the test preperation course than students who did, but the distribution across race/ethnicity groups were relatively similar.

Despite this data, one cannot draw broader conclusions without first knowing the race/ethnicity belonging to each group.

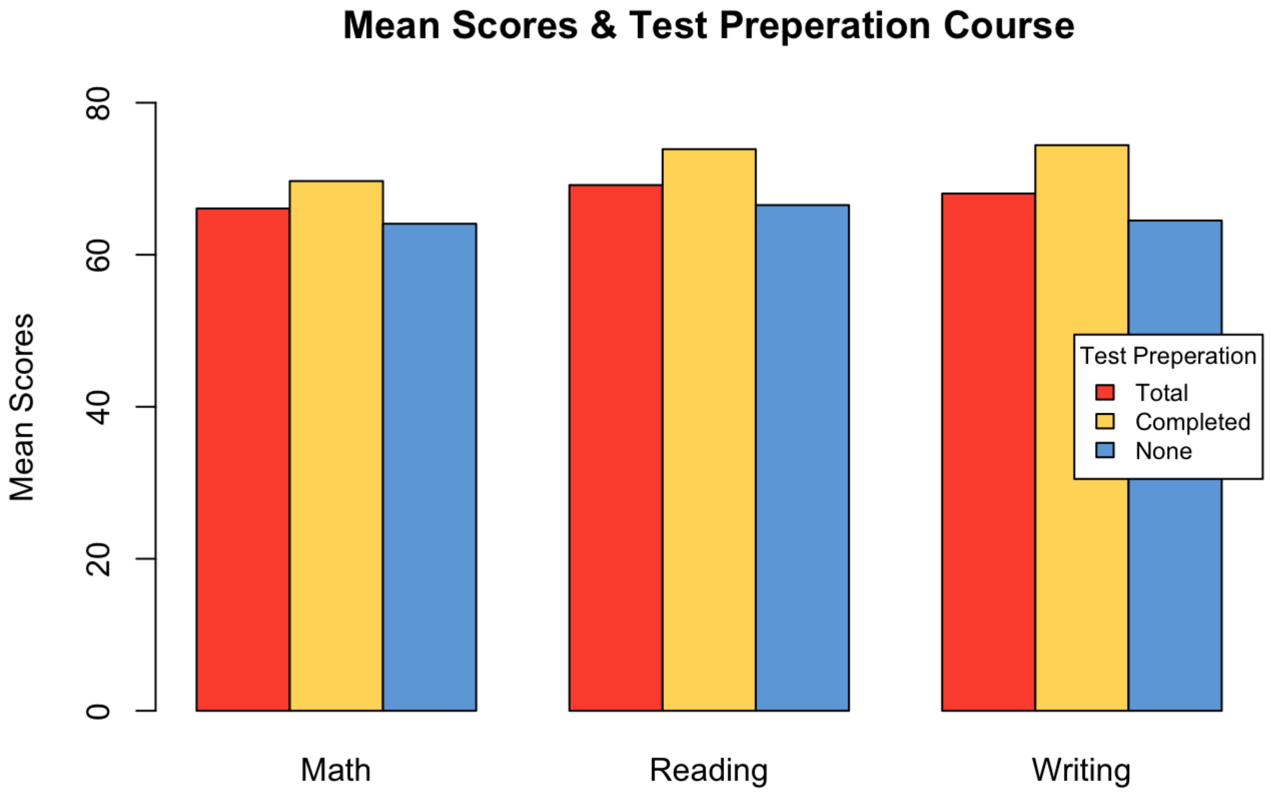
```
race.testprep <- data.frame("Race/Ethnicity" = c("Group A", "Group B", "Group C", "Group
D", "Group E"),
                        "Completed" = c(nrow(groupA[groupA$test.preparation.course == "c
ompleted",]), nrow(groupB[groupB$test.preparation.course == "completed",]), nrow(groupC
[groupC$test.preparation.course == "completed",]), nrow(groupD[groupD$test.preparation.c
ourse == "completed",]), nrow(groupE[groupE$test.preparation.course == "completed",])),
                        "None" = c(nrow(groupA[groupA$test.preparation.course == "none"
,]), nrow(groupB[groupB$test.preparation.course == "none",]), nrow(groupC[groupC$test.pr
eparation.course == "none",]), nrow(groupD[groupD$test.preparation.course == "none",]),
 nrow(groupE[groupE$test.preparation.course == "none",])))

race.testprep.data <- data.frame(race.testprep$Completed, race.testprep$None)

barplot(as.matrix(race.testprep.data),
        main="Race/Ethnicity & Test Preperation Course",
        ylab="Units",
        names.arg = c("Completed", "None"),
        col = c("#fc543a", "#ff9566", "#ffd966", "#93c47d", "#6fa8dc"),
        beside=TRUE,
        #ylim = c(0,500),
        legend.text = c("Group A", "Group B", "Group C", "Group D", "Group E"),
        args.legend = list(title = "Race/Ethnicity", x = "topright", bty = "o", cex = .7
5))
```



6. How does completion of a test prep course impact the average score?

From this analysis, we see that the students who complete the test preparation course perform above average in all subjects. Similar, students who do not complete the test preparation course perform below average in all subjects. The subject most impacted by the completion of the test preparation course is writing.

```
testprep.means <- data.frame("Gender" = c("Total", "Completed", "None"),
                           "Math" = c(mean(exams$math.score), mean(completed.data$math.s
core), mean(none.data$math.score)),
                           "Reading" = c(mean(exams$reading.score), mean(completed.data
$reading.score), mean(none.data$reading.score)),
                           "Writing" = c(mean(exams$writing.score), mean(completed.data
$writing.score), mean(none.data$writing.score)))

testprep.means.data <- data.frame(testprep.means$Math, testprep.means$Reading, testprep.
means$Writing)

barplot(as.matrix(testprep.means.data),
        main="Mean Scores & Test Preperation Course",
        ylab="Mean Scores",
        names.arg = c("Math", "Reading", "Writing"),
        col = c("#fc543a", "#ffd966", "#6fa8dc"),
        beside=TRUE,
        ylim = c(0,80),
        legend.text = c("Total", "Completed", "None"),
        args.legend = list(title = "Test Preperation", x = "right", bty = "o", cex = .75
))
```

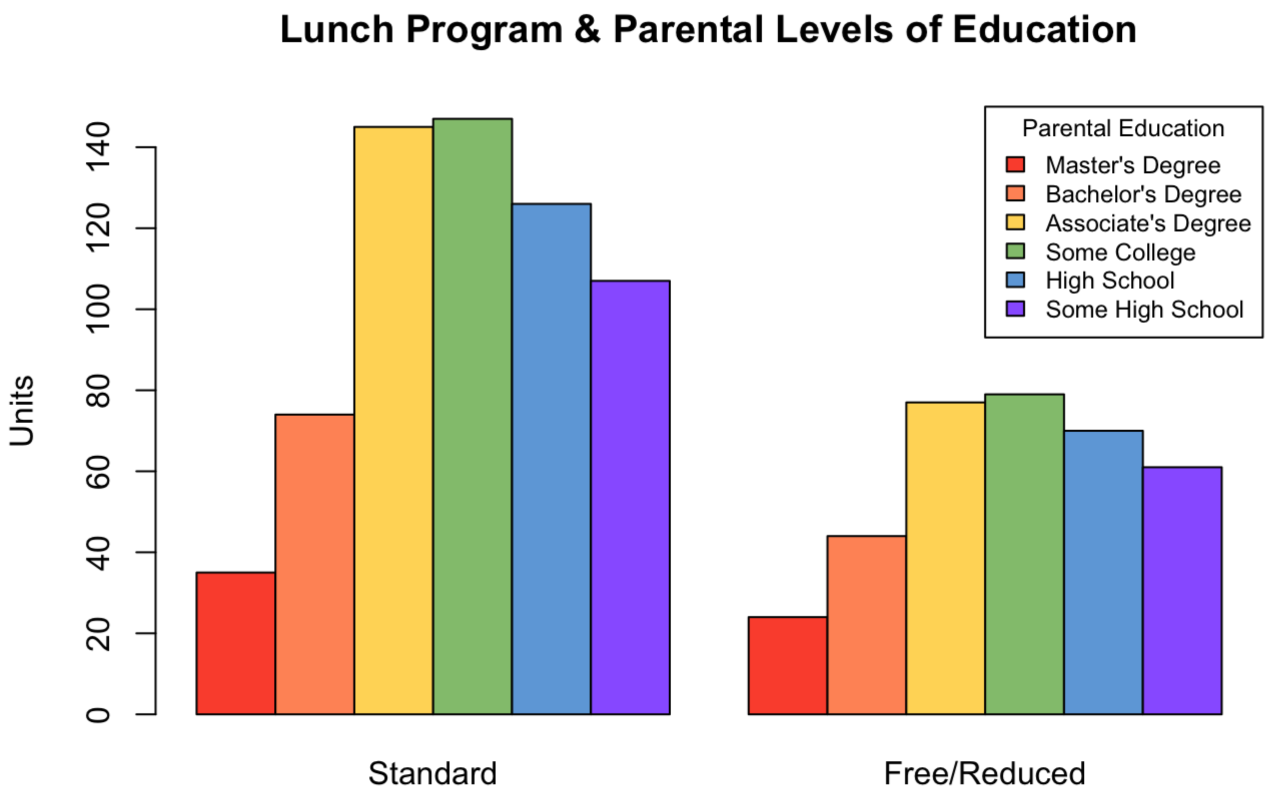


7. Does the students’ parental level of education affect whether the student has a standard or free/reduced lunch?

Yes, students’ parental level of education affects whether the student has a standard or free/reduced lunch. Generally, more students receive a standard lunch than a free/reduced lunch. Parental level of education on lunch program affects standard lunch programs more than free/reduced lunch programs. Across both programs, students whose parents have master’s degrees are far less likely to be on any lunch program, while students whose parents have associate’s degrees or some college experience are most likely to be on any lunch program.

A reasonable conclusion from this data is that students whose parents earn master’s degrees might chose to prepare homemade lunches over a standard or free/reduced lunch program. These families are more affluent and have more resources to support their child academically, and a higher familial income would allow for one parent to stay at home and prepare lunches.

```
edu.lunch <- data.frame("Education" = c("Total", "Master's Degree", "Bachelor's Degree",
"Associate's Degree", "Some College", "High School", "Some High School"),
                        "Standard" = c(nrow(exams[exams$lunch == "standard",]), nrow(mas
ters[masters$lunch == "standard",]), nrow(bachelors[bachelors$lunch == "standard",]), nr
ow(associates[associates$lunch == "standard",]), nrow(somecollege[somecollege$lunch ==
"standard",]), nrow(hs[hs$lunch == "standard",]), nrow(somehs[masters$lunch == "standar
d",])),
                        "Free/Reduced" = c(nrow(exams[exams$lunch == "free/reduced",]),
 nrow(masters[masters$lunch == "free/reduced",]), nrow(bachelors[bachelors$lunch == "fre
e/reduced",]), nrow(associates[associates$lunch == "free/reduced",]), nrow(somecollege[s
omecollege$lunch == "free/reduced",]), nrow(hs[hs$lunch == "free/reduced",]), nrow(someh
s[somehs$lunch == "free/reduced",])))

edu.lunch.data <- data.frame(edu.lunch$Standard[2:7], edu.lunch$Free.Reduced[2:7])

barplot(as.matrix(edu.lunch.data),
        main="Lunch Program & Parental Levels of Education",
        ylab="Units",
        names.arg = c("Standard", "Free/Reduced"),
        col = c("#fc543a", "#ff9566", "#ffd966", "#93c47d", "#6fa8dc", "#9966ff"),
        beside = TRUE,
        ylim = c(0,150),
        legend.text = c("Master's Degree", "Bachelor's Degree", "Associate's Degree", "S
ome College", "High School", "Some High School"),
        args.legend = list(title = "Parental Education", x = "topright", bty = "o", cex
 = .75))
```

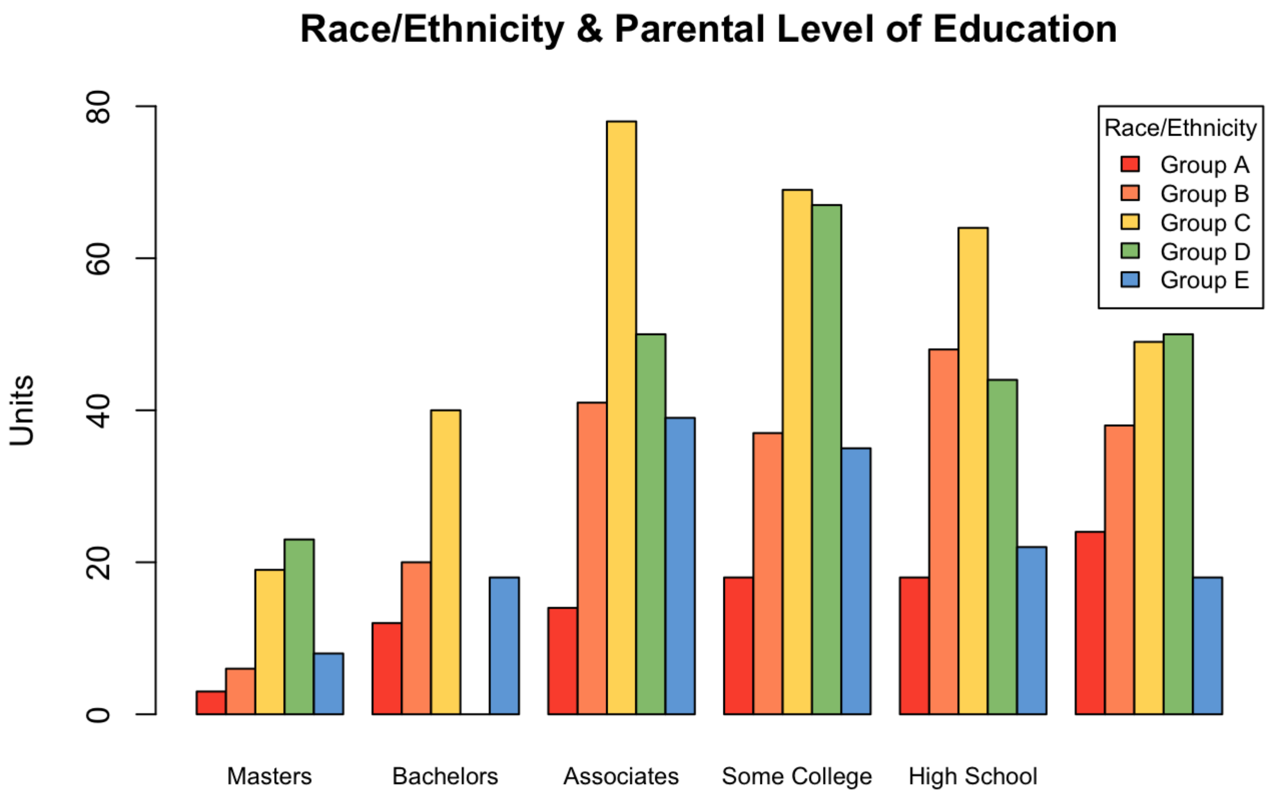


8. How are parental education and race/ethnicity related?

On both sides of the educational range, students who identify as Group A are least likely to have parents with a Master’s degree, while Group D is most likely. Conversely, students who identify as Group E are least likely to have parents with some high school education, while Group D is most likely. We also see that students who identify as Group C are most likely to have parents with an Associate’s degree.

Despite this data, one cannot draw broader conclusions without first knowing the race/ethnicity belonging to each group.

```
edu.race <- data.frame("Race/Ethnicity" = c("Group A", "Group B", "Group C", "Group D",
"Group E"),

                       "Masters" = c(nrow(groupA[groupA$parental.level.of.education ==
"master's degree",]), nrow(groupB[groupB$parental.level.of.education == "master's degre
e",]), nrow(groupC[groupC$parental.level.of.education == "master's degree",]), nrow(grou
pD[groupD$parental.level.of.education == "master's degree",]), nrow(groupE[groupE$parent
al.level.of.education == "master's degree",])),

                       "Bachelors" = c(nrow(groupA[groupA$parental.level.of.education ==
"bachelor's degree",]), nrow(groupB[groupB$parental.level.of.education == "bachelor's de
gree",]), nrow(groupC[groupC$parental.level.of.education == "bachelor's degree",]), nrow
(groupD[groupD$lunch == "bachelor's degree",]), nrow(groupE[groupE$parental.level.of.edu
cation == "bachelor's degree",])),

                       "Associates" = c(nrow(groupA[groupA$parental.level.of.education =
= "associate's degree",]), nrow(groupB[groupB$parental.level.of.education == "associat
e's degree",]), nrow(groupC[groupC$parental.level.of.education == "associate's degree"
,]), nrow(groupD[groupD$parental.level.of.education == "associate's degree",]), nrow(gro
upE[groupE$parental.level.of.education == "associate's degree",])),

                       "Some College" = c(nrow(groupA[groupA$parental.level.of.education
== "some college",]), nrow(groupB[groupB$parental.level.of.education == "some college"
,]), nrow(groupC[groupC$parental.level.of.education == "some college",]), nrow(groupD[gr
oupD$parental.level.of.education == "some college",]), nrow(groupE[groupE$parental.leve
l.of.education == "some college",])),

                       "High School" = c(nrow(groupA[groupA$parental.level.of.education
 == "high school",]), nrow(groupB[groupB$parental.level.of.education == "high school"
,]), nrow(groupC[groupC$parental.level.of.education == "high school",]), nrow(groupD[gro
upD$parental.level.of.education == "high school",]), nrow(groupE[groupE$parental.level.o
f.education == "high school",])),

                       "Some High School" = c(nrow(groupA[groupA$parental.level.of.educa
tion == "some high school",]), nrow(groupB[groupB$parental.level.of.education == "some h
igh school",]), nrow(groupC[groupC$parental.level.of.education == "some high school",]),
nrow(groupD[groupD$parental.level.of.education == "some high school",]), nrow(groupE[gro
upE$parental.level.of.education == "some high school",])))

edu.race.data <- data.frame(edu.race$Masters, edu.race$Bachelors, edu.race$Associates, e
du.race$Some.College, edu.race$High.School, edu.race$Some.High.School)

barplot(as.matrix(edu.race.data),
        main="Race/Ethnicity & Parental Level of Education",
        ylab="Units",
        names.arg = c("Masters", "Bachelors", "Associates", "Some College", "High Schoo
l", "Some High School"),
        col = c("#fc543a", "#ff9566", "#ffd966", "#93c47d", "#6fa8dc"),
        beside = TRUE,
        ylim = c(0,80),
        cex.names = .75,
        legend.text = c("Group A", "Group B", "Group C", "Group D", "Group E"),
        args.legend = list(title = "Race/Ethnicity", x = "topright", bty = "o", cex = .7
5))
```


Race/Ethnicity & Parental Level of Education
Units
Masters
Bachelors
Associates
Some College
High School
Race/Ethnicity
Group A
Group B
Group C
Group D
Group E