Project 2 - Exam Scores

Zachary Palmore

10/2/2020

Objective

In this part of project 2, we will be analyzing education statistics to see if there is a correlation between a student's performance on exams and their parent's highest level of education. It was compiled from scores given to high school students in the United States. It is only a small subset of the data from every student, but with it we try to find:

- If there is a correlation between exam scores and parental education
- If parental education level is a proxy for student exam performance

To get those answers, the data needs to be cleaned and tidied before we can make use of it. We will start by importing the data.

```
require(tidyverse)
require(magick)
require(tesseract)
require(reshape2)
```

Importing

In this case the data was provided in a portable network graphic format (.png). To extract the data from the image we can first read the image into an external pointer of class "magick." This is a file format that calls specifically for the *magick* package. It will allow us to process the image into a readable format for another function to pull information out of as a string of characters.

```
# Importing the image with magick
scores <- image_read("examscores.png") %>%
# Create a transparent background without any blur
  image_transparent("transparent") %>%
# Reinforce the background transparency
  image_background("transparent") %>%
# Reassure that the image is in black and white only
  image_negate() %>%
# Use a rectangle to 'thin' the image and simultaneously
# define edges in, on, and around the image
  image_morphology(method = "Thinning",
                   kernel = "Rectangle") %>%
# reinforce clarity for reading through contrast by negating # any variation in background colors
  image_negate()
# Show the image stats
scores
```

gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading sco
female	group B	bachelor's degree	standard	none	72	
female	group C	some college	standard	completed	69	
female	group B	master's degree	standard	none	90	
male	group A	associate's degree	free/reduced	none	47	
male	group C	some college	standard	none	76	
female	group B	associate's degree	standard	none	71	
female	group B	some college	standard	completed	88	
male	group B	some college	free/reduced	none	40	
male	group D	high school	free/reduced	completed	64	
female	group B	high school	free/reduced	none	38	
male	group C	associate's degree	standard	none	58	
male	group D	associate's degree	standard	none	40	
female	group B	high school	standard	none	65	

Like Magick, this image has become much more readable and the data did not go through any changes. From here, we need a new function. Something that can read the information we can see into more functional form.

Tidying

Our end goal is the create a data frame that we can export as a spreadsheet (.csv) for sharing and also perform analysis on to answer those questions. To read the numbers and writing in the image, another package called *tesseract* specializes in optical character recognition (OCR). This is exactly what we need.

Importantly, for the *tesseract* functions to perform properly, the clarity of the image needs to be as good as possible. The clearer the writing in the image, the more accurate the output. This is why we enhanced the edges of all the characters and changed the background of the image in the importing process. Way to work that *magick*!

```
# Uses the power of the tesseract we process the image
scores <- scores %>%
  image_ocr()
# After it pulls out all the characters we process the string
# We use the readr read_delim function here
# to separate the strings by their spaces
scores <- read_delim(scores, " ", comment = "")
```

```
## Warning: Duplicated column names deduplicated: 'score' => 'score_1' [14],
## 'score' => 'score_2' [16]
```

```
## Warning: 13 parsing failures.
## row col   expected     actual         file
##   1  -- 16 columns 10 columns literal data
##   2  -- 16 columns 10 columns literal data
##   3  -- 16 columns 10 columns literal data
##   4  -- 16 columns 10 columns literal data
##   5  -- 16 columns 10 columns literal data
## ... ... ......... ......... ...........
## See problems(...) for more details.
```

```
# Selecting the useful columns
scores <- scores[,c(1,3,4,6,7,8,9,10)]
# Renaming the columns
colnames(scores) <- c("Sex",
                      "EthnicGroup",
                      "Parent_Edu",
                      "LunchCost",
                      "TestPrep",
                      "Math",
                      "Reading",
                      "Writing")
# Convert the variables of interest into
# data types we can work with downstream
as.numeric(scores$Math)
```

```
##  [1] 72 69 90 47 76 71 88 40 64 38 58 40 65
```

```
as.numeric(scores$Reading)
```

```
##  [1] 72 90 95 57 78 83 95 43 64 60 54 52 81
```

```r
as.numeric(scores$Writing)
```

```
##  [1] 74 88 93 44 75 78 92 39 67 50 52 43 73
```

```r
# Converting the wide data into a longer format
# Each student only has one parental education but has 3
# scores - math, reading, writing
# We will compare all scores at once
scores <- melt(scores, na.rm = TRUE)
```

```
## Using Sex, EthnicGroup, Parent_Edu, LunchCost, TestPrep as id variables
```

```r
# Rename the combined math, reading, and writing values
# column as 'scores' and its variable as 'subject' using
# dply rename function then assigning to the df
scores <- scores %>%
  dplyr::rename(Score = value,
         Subject = variable)
```

The power of the tesseract optical character recognition is impressive. It parsed the image into strings of text, and since it was clear enough, we extracted all the information. We also separated out the data into categories well enough that we can now process the information using statistics. Time to begin the analysis!

Analysis

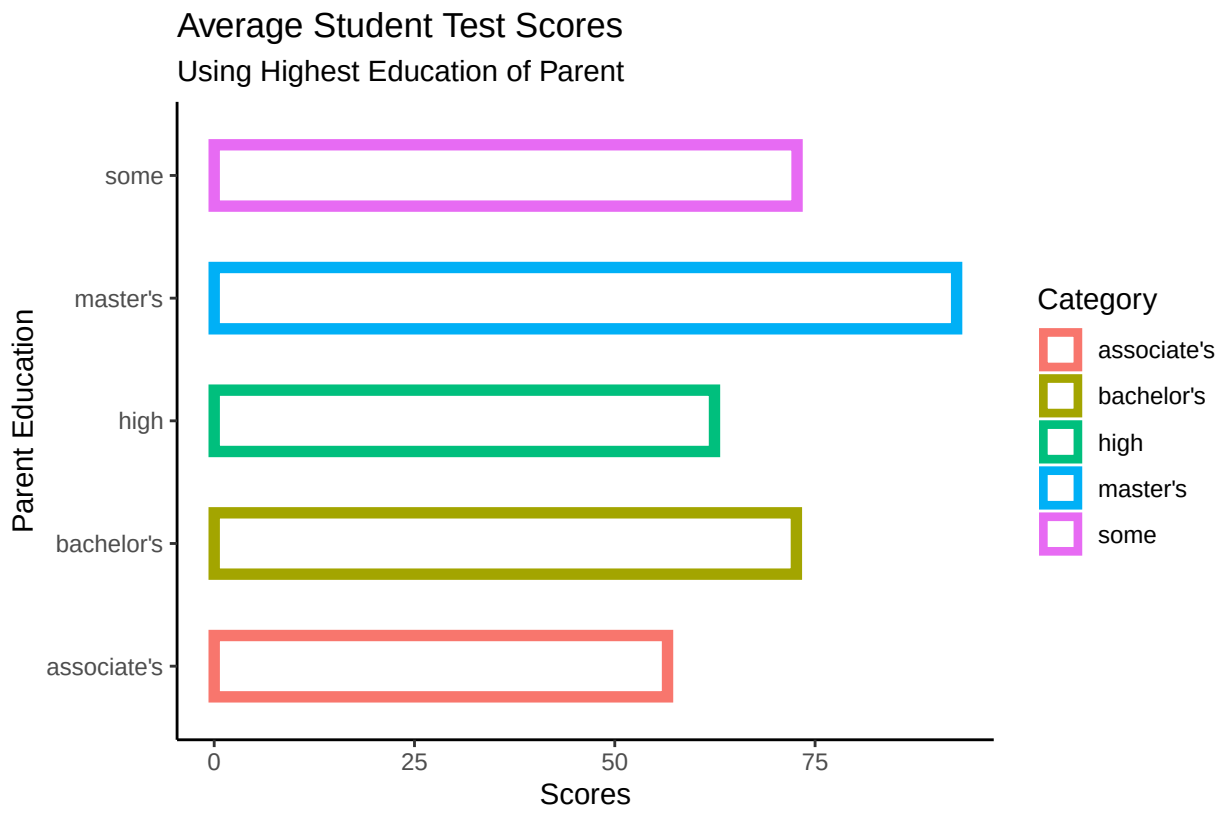
As a reminder, we wanted to find out if there is a correlation between the highest obtained education level of a student's parent and that student's performance on exams. Let's review what we have so far.

```
head(scores, 3)
```

```
##      Sex EthnicGroup Parent_Edu LunchCost  TestPrep Subject Score
## 1 female           B bachelor's  standard      none    Math    72
## 2 female           C       some  standard completed    Math    69
## 3 female           B   master's  standard      none    Math    90
```

In the data we have scores for exams called 'scores' and we have a category of education for the parents of each of those students. One way to observe the differences is to combine the scores by each parent's highest level of education. Then we can use a horizontal bar chart to compare levels of student scores based on their parent's highest education level.

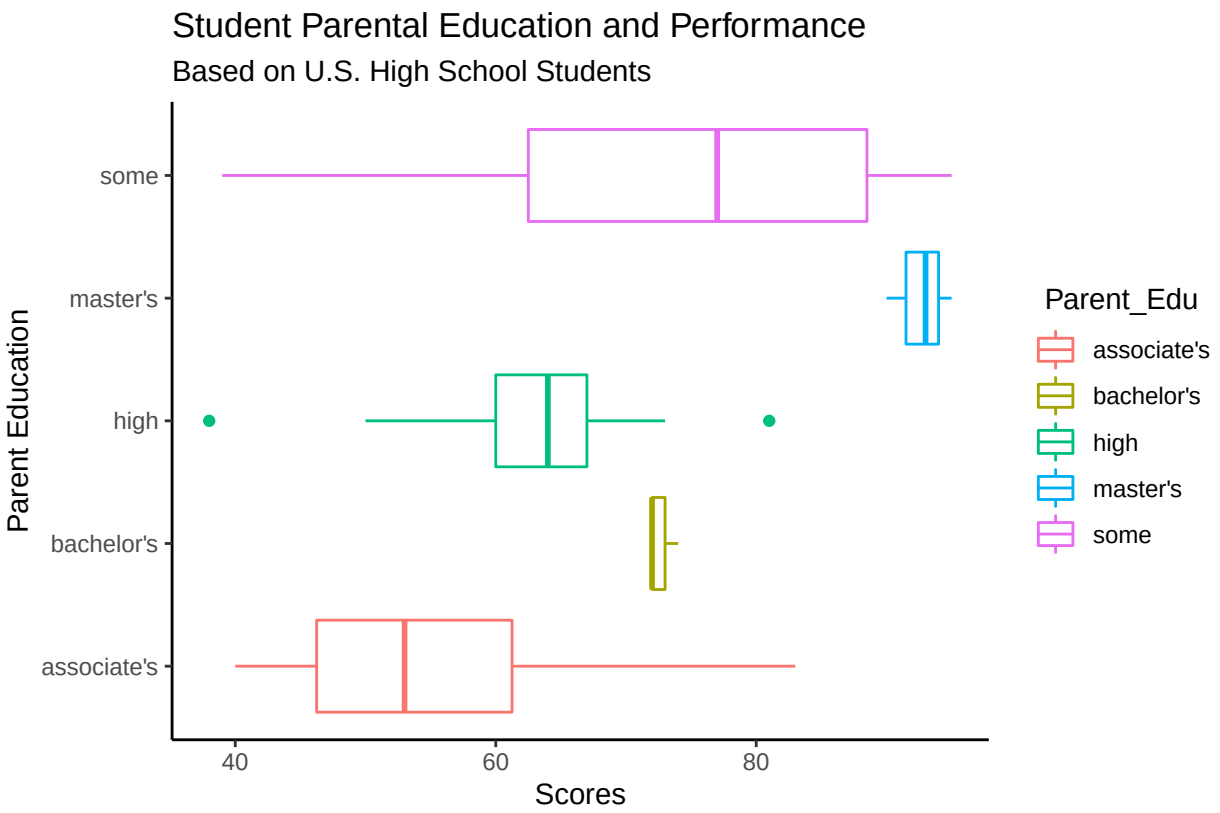
```
# Sum the scores by parental education
stats <- aggregate(scores$Score, by = list(Category = scores$Parent_Edu), FUN = mean)
# plot it
ggplot(stats, aes(x= Category, y = x, col = Category)) + geom_bar(stat="identity", fill = "transparent"
  labs(title = "Average Student Test Scores",
       subtitle = "Using Highest Education of Parent",
       x = "Parent Education",
       y = "Scores",
       caption = "Data based on marks of U.S. high school students") +
  theme_classic() +
  theme(plot.caption = element_text(hjust = 0.45, face = "italic"),
        plot.caption.position =  "plot") +
  coord_flip()
```



This gives a clear indication that those students who have at least one parent with a master's degree score higher on exams. In general, it appears as though a student's exam score is improved by having a parent with at least some post-secondary education. However, there is an exception within this sample. Those students who had at least one parent obtain an associates degree, did slightly poorer on exams than students with parents who only finished high school.

If we wanted to dig a little further, we could compare the mean, inter-quartile ranges, and if there are any anomalies in this small sample of the data. To perform this all at once, we can create a boxplot based on the Parent's Education and student scores overall.

```
ggplot(data = scores, aes(x = Score, y = Parent_Edu, col = Parent_Edu)) + geom_boxplot() + theme_classi
  labs(title = "Student Parental Education and Performance",
       subtitle = "Based on U.S. High School Students",
       x = "Scores",
       y = "Parent Education",
       caption = "Data is small sample of U.S. student population") +
theme(legend.title = element_text(hjust = 0.5),
      plot.caption = element_text(hjust = 0.5, face = "italic"))
```

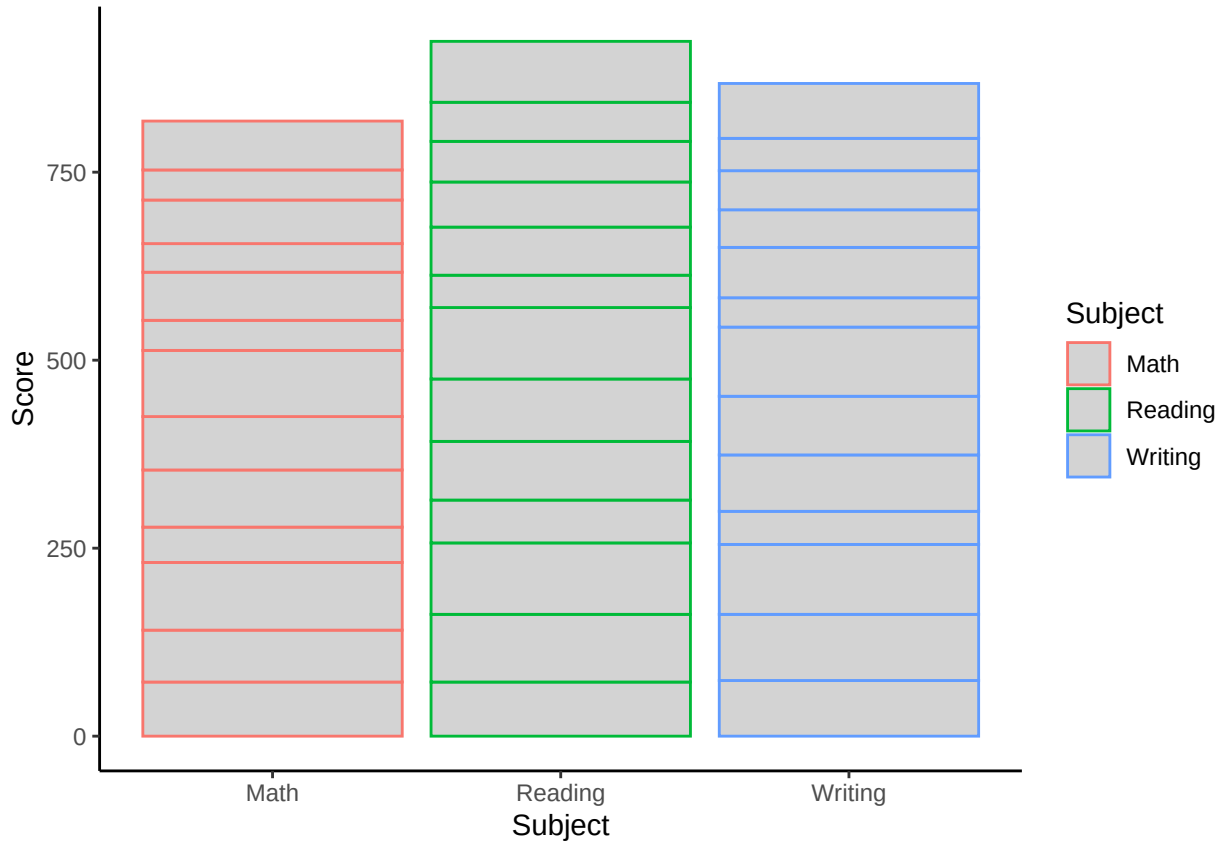


This mostly confirms the bar chart's assumptions as we can see the order of scores based on parental education more clearly. It start with students who have at least one parent with a master's degree as those who are given the highest scores. Then in sequence in the boxplot, it follows that the average scores for students with parents with some post-secondary education is higher than those students whose parent had at least a bachelor's degree. High school diplomas were next, followed by those whose parents had earned at least an associates degree.

However, using the boxplot we can see that the data for students who had at least one parent that earned a bachelor's is very tight in proportion to the other education levels. The same applies to those who had parents with at least a master's. High school earners also had students with outliers, with one scoring higher on exams than some students whose parent earned at least one bachelor degree and the other as the lowest score of all distributions.

We can also calculate the total score for everyone in each subject, just for fun.

```
ggplot(scores, aes(x = Subject, y = Score, col = Subject)) + geom_col(fill = "light grey") + theme_clas
```



It looks like, based on the subjects all students were tested on, their best score overall was in reading. Writing was a close second, followed by math as the lowest score in this sample of students.

Conclusion

From this we can conclude that, in general, students who have at least one parent that has gone through some college, or earned a bachelor's or a master's degree will do better on exams than those whose parents do not have a degree or just graduated high school. However, there is an exception.When a parent has earned at least an associate's degree, the students in this study tend to do worse than those who had parents with a high school diploma. It follows contrary to the notion that the higher a parent's education level the better the student will do. In other words, the education level of a parent could not be used as a proxy at all levels of education because there is not a clear positive or negative correlation.

Given that our sample is based on 13 total students with an aggregated score from each, it is best not to interpret these as factual evidence. The sample size is too small to even closely estimate the population. Increasing the sample size could allow us to validate or disprove the notion that having a parent with a higher post-secondary education will result in students that perform better on exams.

We also noticed that students performed best in the subject of reading is this study while math was their worst. It would be interesting to compare these students' parents education level on each subject by using the score of each exam. This might let us determine further if a certain level of education from a parent improves the scores of a student in certain subjects. Although, doing so with this data would be moot. Additional data should be gathered. As a result, more research is needed.