Final Project:

Math Educational Disparity

Aaron Drexler

1/16/2022

DSC 630

Abstract

For my project, I wanted to focus on the equity aspect of education as well as optimizing student performance given external factors. First, let's focus on equity. The educational grant that I am participating in is studying the disparity in math performance. Minority communities, particularly as African American and Hispanic, consistently score lower on standardized testing in mathematics when compared to White and Asian students. While this grant is specifically studying methods of instruction that would improve minority scores, I want to see if I can duplicate their initial data, highlighting the concern present and score disparity, potentially refining and improving on their conclusions.

The next aspect of this how to improve and maximize math performance. I want to find the optimal student. I want to look beyond individual or racial data and look at a wide range of descriptors and data points about the students. I want to gather as many data points as possible in relation to the students. I am not looking at maximizing the student totals, but rather maximizing the number of comparable attributes and tendencies about each student that I can gather, then using them to find useful and actionable conclusions about education and mathematical testing.

Intro/background of the problem

While I am getting my master's in data science, I currently am working as a math teacher working at a high school in Florida. I teach two subjects entitled Liberal Arts

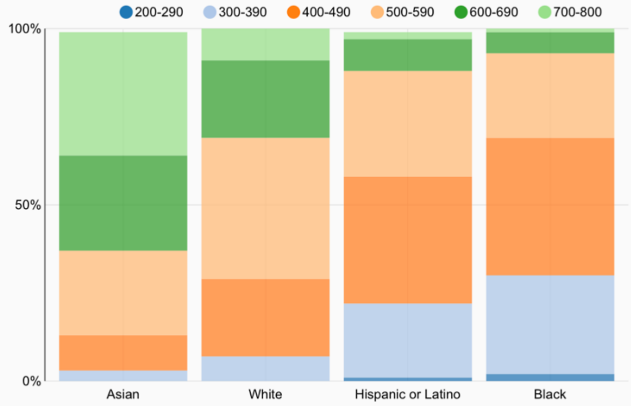
Mathematics and Advanced Topics in Mathematics. While most do not know these subjects, they consist of a very wide range of topic coverage. One area of focus in teaching math is equity, in ensuring every student is receiving the best opportunities to succeed. The reason why I am bringing up both the subjects that I teach and the focus on equity is because I both personally use data in my instructional plans, and I am also participating in an educational grant to study equity in mathematics.

There is a significant gap in mathematics education for minorities. Several studies have shown the minority students test and score consistently lower. Below are a few visualizations that demonstrate this gap as found by a study of SAT test results. As is clearly shown, Asian and White students consistently score high on their SATs and focus in college on STEM degrees. Comparatively, Hispanic and Latino students score substantially lower, and black students scored even lower than that both in general studies, but especially in terms of math achievements. These visualizations are found in the study report in this link:

https://www.brookings.edu/blog/up-front/2020/12/01/sat-math-scores-mirror-and-maintain-racial-inequity/

Wide race gaps in SAT math scores

Math score distribution by race or ethnicity

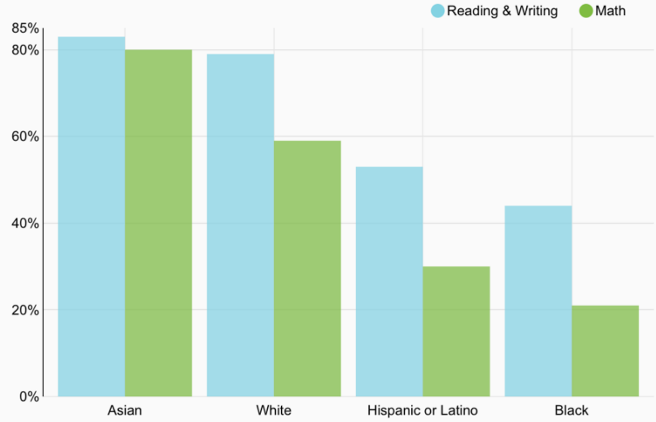


College Board, "SAT Suite of Assessments Annual Report," 2020.

BROOKINGS

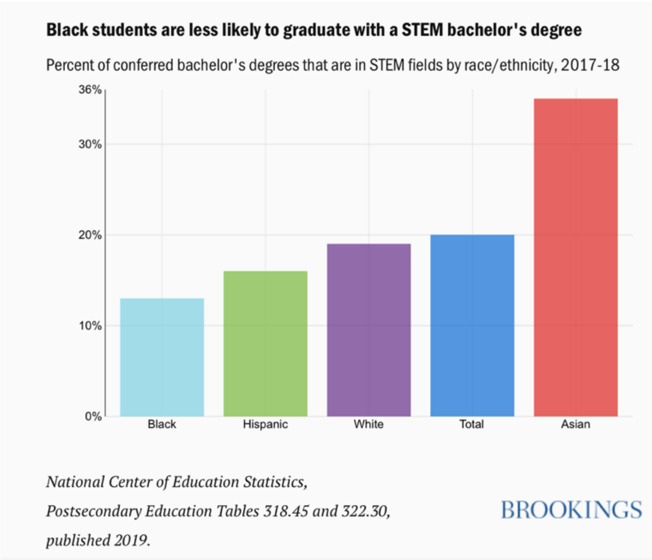
Race gaps among college ready students

Portion of test takers meeting college readiness benchmarks



College Board, "SAT Suite of Assessments Annual Report," 2020.

BROOKINGS


Black students are less likely to graduate with a STEM bachelor's degree

Percent of conferred bachelor's degrees that are in STEM fields by race/ethnicity, 2017-18

National Center of Education Statistics, Postsecondary Education Tables 318.45 and 322.30, published 2019.

Methods

The focus of my project is going to be to study student performance and cross reference that with the attributes and descriptors about each student. So much of education is about bridging the educational divide, and with this study I want to find the specific areas where that bridge is pronounced. The outcomes that I will be looking at is to find patterns in the data to find descriptors that have a high correlation to poor math performance while also searching for an optimum educational situation. The reason to search for the optimal situation is to create the goal to reach for those who are educationally deprived.

The first and foremost concern with this project is data collection. When dealing with these large-scale comparisons, having a large enough population as well as enough data elements to compare is a struggle. Initially, I have one good database that I found. This is a database that I found on Kaggle that gathered a large amount of secondary information about students' personal lives, habits, and backgrounds, while also sharing their mathematical performance. The data points in this database are:

- student’s school
- student’s sex
- student’s age
- student’s family size
- parental cohabitation status
- mother’s education
- father’s education
- mother’s job
- father’s job
- guardian
- home to school travel time
- weekly study time
- number of past class failures
- extra educational support
- family educational support
- extra-curricular activities
- wants to take higher education
- home internet access
- in a romantic relationship
- quality of family relationships
- free time after school
- going out with friends

- health status
- absences
- math scores.

Using this data, I want to build an optimal profile, by determining, via correlation, which of these attributes have the greatest effect on their scores, while also determining which group of students are in the most need of assistance. I plan to run various models such as K-means, linear regression, correlation matrices, and other correlation-based visualizations.

I chose to narrow down to the key data elements that are relevant to the math scores and are something that can be controlled and altered by the student. In other words, what can a student change about his/her behavior to adjust math scores. So here are the following topics that I narrowed my selections down to,

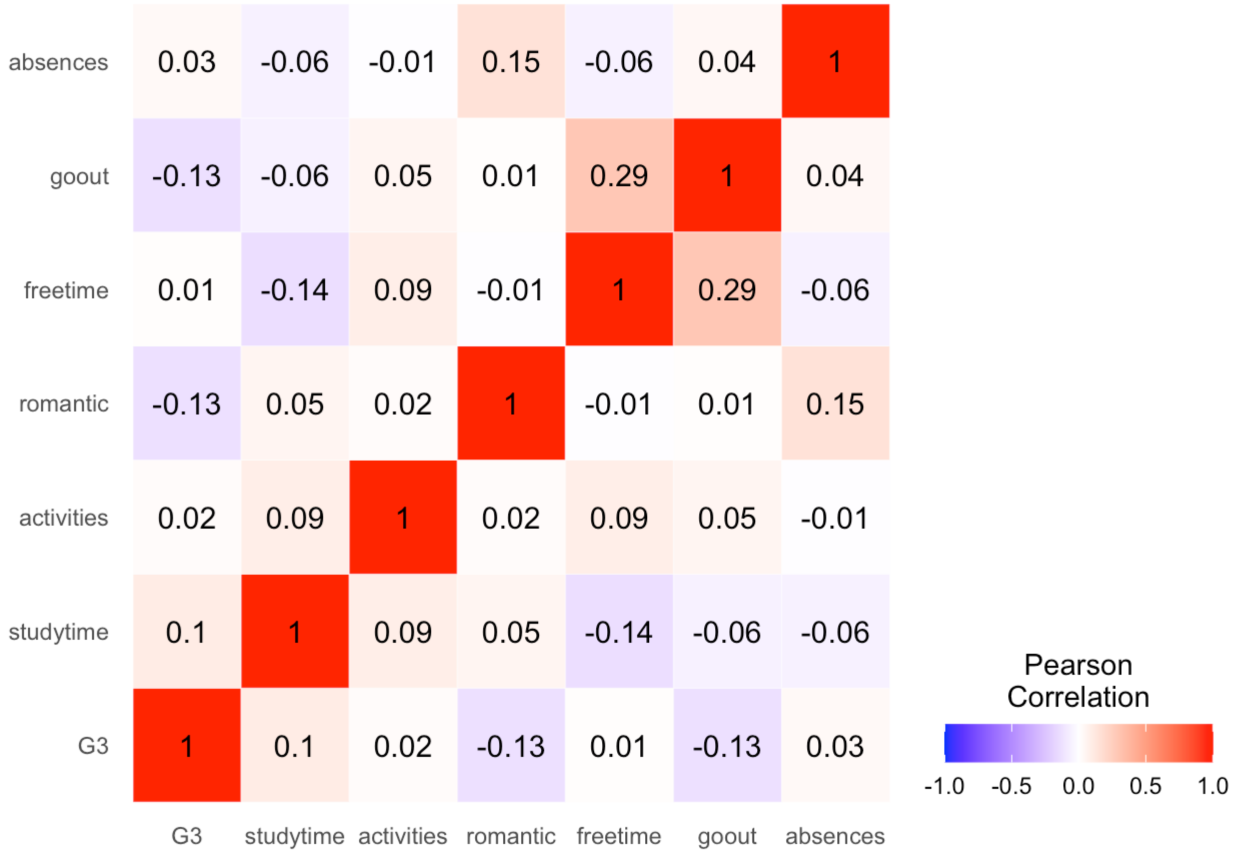
- math scores (G3 is the final grade)
- weekly study time
- extra-curricular activities
- in a romantic relationship
- free time after school
- going out with friends
- absences

Results

The main focus of the model was looking at the Pearson Correlation Heat map to determine correlation between the previously selected features of the data. I used R for the visualization and coding, because I felt that ggplot2 was the best and most forward way to find the relevant data and best picture my data. Below is the R code used:

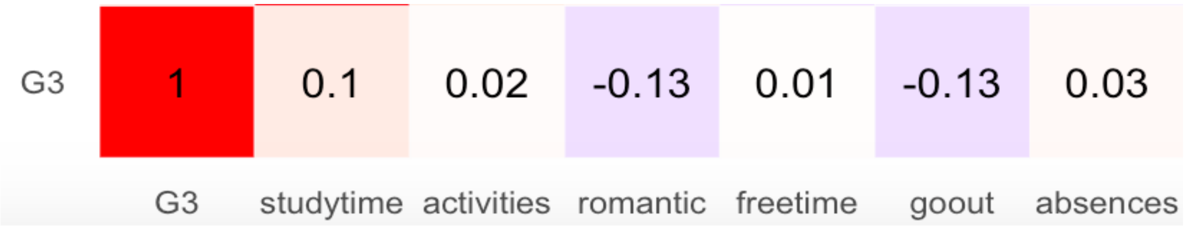
```
library(ggplot2)

getwd()

setwd("~/Documents/DSC 630")

library(readr)

df <-read_csv("student-mat.csv")

View(df)

library(dplyr)

df1 <- df[, c('G3','studytime','activities','romantic','freetime','goout','absences')]

View(df1)

df1$activities<-ifelse(df1$activities=="yes",1,0)

df1$romantic<-ifelse(df1$romantic=="yes",1,0)

View(df1)

corr <- round(cor(df1),2)

View(corr)

library(reshape2)

melted_corr <- melt(corr)

View(melted_corr)

ggplot(data = melted_corr, aes(x=Var1, y=Var2, fill=value)) + geom_tile(color = "white")+
 scale_fill_gradient2(low = "blue", high = "red", mid = "white",
           midpoint = 0, limit = c(-1,1), space = "Lab",
           name="Pearson\nCorrelation") + geom_text(aes(Var2, Var1, label =
value), color = "black", size = 4) +
 theme(
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  panel.grid.major = element_blank(),
  panel.border = element_blank(),
  panel.background = element_blank(),
  axis.ticks = element_blank(),
  legend.justification = c(1, 0),
  legend.direction = "horizontal")+
 guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
               title.position = "top", title.hjust = 0.5))
```

From this data and code, we got the following correlation Heat map:



	G3	studytime	activities	romantic	freetime	goout	absences
absences	0.03	-0.06	-0.01	0.15	-0.06	0.04	1
goout	-0.13	-0.06	0.05	0.01	0.29	1	0.04
freetime	0.01	-0.14	0.09	-0.01	1	0.29	-0.06
romantic	-0.13	0.05	0.02	1	-0.01	0.01	0.15
activities	0.02	0.09	1	0.02	0.09	0.05	-0.01
studytime	0.1	1	0.09	0.05	-0.14	-0.06	-0.06
G3	1	0.1	0.02	-0.13	0.01	-0.13	0.03

The focus of this correlation map should be in relation to G#, or in other words the student's math final scores:



	G3	studytime	activities	romantic	freetime	goout	absences
G3	1	0.1	0.02	-0.13	0.01	-0.13	0.03

Discussion/conclusion

There are a few main take-aways from this data and correlations. First, as a general note, none of the variables had or strong, or even moderately sized relationship with one another. The strongest relationship between two variables is the relationship between going out and free time, which should be considered strong, since more free time usually means more time to go out. However, even that had only a 0.29 correlation factor, indicating not even mid-level correlation. The reason for these lower-than-expected correlation score can be attributed to small sample size and sample bias. This data was collected of just 150 students and was collected in just a single classroom in a single, non-US location. Right now, as a math teacher, I have almost 30 more students than that. After that 0.29, every other correlation was below 0.15. I believe given more unbiased data, these correlations would have grown in strength, but for now we look at the correlation numbers with a perspective comparing it to a 0.15 correlation factor to compare and determine the individual strength of each variable on the students' grades.

Next thing that I am going to look at is each individual variable in relationship to the final G3 grade of the students. First, let's start with the obvious take-aways from the data. There is a relatively (in terms of the small population of the data, so out of 0.15) strong relationship between study time and grade scores. With a .10 out of .15 comparatively, studying had an obviously positive effect on grades. On the reverse side, both going out and being in a romantic relationship had a relatively strong negative relationship with the grades, meaning the more a person is in a relationship or/and the more a student goes out after school, the worse the student's grades will become.

What is interesting is that with both variables having a correlation factor score of -0.13, these factors have a stronger impact than studying does in the opposite direction. That means that being in a relationship or going out often hurts the grade more than studying helps the grade. I am a little surprised at first glance at the negative strength of a romantic relationship. However, we must consider that these are high school students, and romantic relationships can be more distracting and can have a greater effect at that age.

The other three variables, amount of free time, having after school activities, and having absences have a relatively (and not relatively) weak positive relationship with the grade. The biggest surprise of these semi-neutral variables is absences. Originally, I would have thought that absences would have a strong negative impact on a grade, but according to the data, being absent helps the grade a tiny amount. I might consider this a result of small data bias. Over a larger sample, I am sure that it might shift a bit, but the main take away is that absences do not have nearly as big an impact on grades as previously predicted.

References

1. https://www.kaggle.com/janiobachmann/math-students/version/1
2. https://www.brookings.edu/blog/up-front/2020/12/01/sat-math-scores-mirror-and-maintain-racial-inequity/