**Business Problem**

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

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

My focus is on education, more specifically what personal life metrics has the largest impact on a student’s grade. For the sake of educational equity and excellence, I want to investigate more into student’s personal life and how it compares to their grade. This may be what is causing the educational gap we see in the education system today. ***I want to predict what aspects of students’ personal home life has the greatest impact on their school performance to help better student performance.***

**Background/History**

While I am getting my master’s in Data Science right now, I am also a high school math teacher during the day. One thing that I deal with daily are poorly performing students in my class. There are many students in high school who regularly get low grades on a regular basis. Now, the cause of their bad grades usually has two different factors, classroom performance and home life. As a teacher, one of the main things that I do every day is analyze in class performance and then either create or adjust the educational plans or methods accordingly to compensate and try to improve student scores.

Then again, let’s consider the second factor on a student’s grade, the student’s personal life. As a teacher, you find that there are some students that do not want or will not learn, no matter what the educational plan happens to be, due to either lack of desire or motivation. One of the biggest things experience teachers tell their younger counterparts (aka me) is that you do not know what is going on in students’ personal life, and how it is affecting their education.

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.

**Data Explanation (Data Prep/Data Dictionary/etc)**

There was a decent amount of data prep and data cleaning. Below is the code with the R code that I used to clean my data base, along with commentary about what the code is doing.

library(ggplot2)

getwd()

setwd("~/Documents/DSC 680")

##SET DIRECTORY

library(readr)

df <-read\_csv("student-por.csv")

View(df)

##IMPORT DATA

library(dplyr)

df1 <- df[, c('G1','G2','G3','studytime','activities','romantic','freetime',

'goout','absences','Medu','Fedu','Pstatus','traveltime',

'schoolsup','famsup','paid','nursery','higher','internet','famrel'

,'health')]

View(df1)

##REDUCED VARIABLES TO THOSE WE WANT

famedu <- df1$Medu + df1$Fedu

famedu

##CREATED A NEW VARIABLE, JOINT EDUCATION LEVEL

df1$Famedu <- famedu

View(df1)

##INSERT BACK INTO DF

df1 <- df1[, c('G1','G2','G3','studytime','activities','romantic','freetime',

'goout','absences','Medu','Fedu','Famedu','Pstatus','traveltime',

'schoolsup','famsup','paid','nursery','higher','internet','famrel'

,'health')]

View(df1)

##REORDERED TO PUT FAM EDU NEXT TO INDIVIDUAL PARENT ED

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

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

df1$Pstatus<-ifelse(df1$Pstatus=="T",1,0)

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

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

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

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

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

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

View(df1)

##CONVERTED CATEGORICAL DATA INTO NUMERICAL

**Methods**

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:

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

View(corr)

##FINDS CORRELATIONS OF VARIABLES

library(reshape2)

melted\_corr <- melt(corr)

View(melted\_corr)

##MELTS DATA TO BE ABLE TO PUT INTO HEAT MAP

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(),

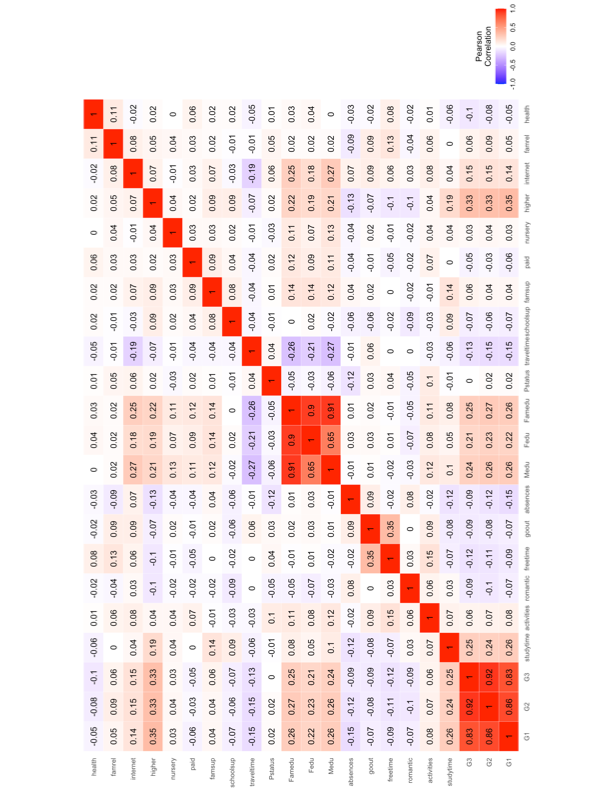
legend.justification = c(1, 0),

legend.direction = "horizontal")+

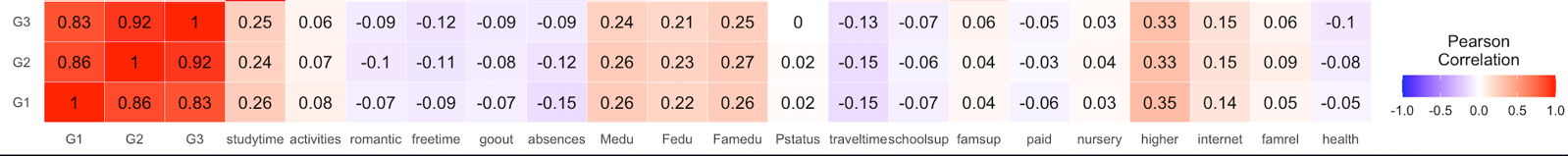
guides(fill = guide\_colorbar(barwidth = 7, barheight = 1,

title.position = "top", title.hjust = 0.5))

##GRAPHED MELTED DF INTO HEAT MAP



I am focusing how these external factors affect the student grades, so I looked at just the bottom three rows.



Below is a table of these values and their subsequent correlations.

|  |  |  |
| --- | --- | --- |
| G1 | studytime | 0.26 |
| G1 | activities | 0.08 |
| G1 | romantic | -0.07 |
| G1 | freetime | -0.09 |
| G1 | goout | -0.07 |
| G1 | absences | -0.15 |
| G1 | Medu | 0.26 |
| G1 | Fedu | 0.22 |
| G1 | Famedu | 0.26 |
| G1 | Pstatus | 0.02 |
| G1 | traveltime | -0.15 |
| G1 | schoolsup | -0.07 |
| G1 | famsup | 0.04 |
| G1 | paid | -0.06 |
| G1 | nursery | 0.03 |
| G1 | higher | 0.35 |
| G1 | internet | 0.14 |
| G1 | famrel | 0.05 |
| G1 | health | -0.05 |
| G2 | studytime | 0.24 |
| G2 | activities | 0.07 |
| G2 | romantic | -0.10 |
| G2 | freetime | -0.11 |
| G2 | goout | -0.08 |
| G2 | absences | -0.12 |
| G2 | Medu | 0.26 |
| G2 | Fedu | 0.23 |
| G2 | Famedu | 0.27 |
| G2 | Pstatus | 0.02 |
| G2 | traveltime | -0.15 |
| G2 | schoolsup | -0.06 |
| G2 | famsup | 0.04 |
| G2 | paid | -0.03 |
| G2 | nursery | 0.04 |
| G2 | higher | 0.33 |
| G2 | internet | 0.15 |
| G2 | famrel | 0.09 |
| G2 | health | -0.08 |
| G3 | studytime | 0.25 |
| G3 | activities | 0.06 |
| G3 | romantic | -0.09 |
| G3 | freetime | -0.12 |
| G3 | goout | -0.09 |
| G3 | absences | -0.09 |
| G3 | Medu | 0.24 |
| G3 | Fedu | 0.21 |
| G3 | Famedu | 0.25 |
| G3 | Pstatus | 0.00 |
| G3 | traveltime | -0.13 |
| G3 | schoolsup | -0.07 |
| G3 | famsup | 0.06 |
| G3 | paid | -0.05 |
| G3 | nursery | 0.03 |
| G3 | higher | 0.33 |
| G3 | internet | 0.15 |
| G3 | famrel | 0.06 |
| G3 | health | -0.10 |

**Analysis**

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 to the grades themselves was the desire for higher education. However, even that had only an, at most, 0.35 correlation factor, indicating barely 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 649 students and was collected in just in a select number of high schools, also in a non-US location. 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 of relative strength to the rest of the correlations.

Next thing that I am going to look at is each individual variable in relationship to the final G3 grade of the students. The G3, or final grade, will probably reflect student long term performance in school. 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) strong relationship between grades and the following variables: desire for a higher education, all parental Education level data, study time, and Internet access. Each averaged a correlation between 0.15-0.35, a significant positive correlation, meaning these have the strongest impact on student performance(Internet access is the lowest correlation with 15%, while Higher educational desire averaged a correlation of 0.34. On the reverse side, the largest negative impacts on student grades are amount of student free time, school absences, and amount of travel time, all averaging a correlation factor of just over 0.10.

**Conclusion**

To help students succeed, here are some takeaways that may boost long term educational scores. First, promote higher education in lower-level schools. Often students see college or higher education as unrealistic and therefore unattainable. They are told that they cannot achieve this, so their desire to succeed in school diminishes. The next topic is obvious. Ensure that student study times remain high and that students have access to the internet has a positive impact on the student, as more studying often equates to higher scores in grades. Next, education is absolutely a generational factor. Investing in community adult education as well as educating parents on their student’s education will impact student performance. Finally, ensuring close (location wise) quality education is available to everyone is important as well.

**Assumptions/limitations/Challenges/Ethical Assessment**

There are a few ethical dilemmas when it comes to this data base. First, the accuracy of the data might not be reliable, as we are relying on the truthfulness of high school students, which is very suspect at best. Second, this study was conducted with Portuguese high school students. This may cause a drift or bias in the data that is unique to Portugal and does not apply to American Students. Third, the study has only 648 students. While that may appear to be a lot, I believe that it still will present as a small population creating bias in the data. Finally, this data is taken from just two schools, and bias occurs with specific schools. In short, the data base is lacking diversity and size, while relying on student honesty. So, while it may not affect the results, we must take the outcomes with a grain of salt.

**References**

https://www.kaggle.com/datasets/larsen0966/student-performance-data-set