Group Data Visualization Project

The Bobo Dolls

Gemma Graziosi, Tia Kleiner and Aliya Aiyub

Department of Psychology, York University

HH/PSYC 3031: Intermediate Statistics Laboratory

Dr. Monique Herbert

November 4, 2020

**A General Description of the Data.**

Our dataset includes data from 55 students from a midwestern liberal arts college. It is unknown how the data was collected.

There are four variables that we are focusing on for our visualization question:

1) Cumulative GPA: the total grade point average a student has achieved across all courses. Cumulative GPA is a continuous variable that can be a minimum of 0 and a maximum of 4.

2) Total MCAT Score: the total score for the Medical College Admission Test (MCAT), which is the sum of students’ scores in Verbal Reasoning, Physical Sciences, and Biological Sciences. Total MCAT Score is a continuous variable that can be a minimum of 3 and a maximum of 45.

3) Number of Medical Schools Applied to is a categorical variable with 2 levels: accepted into medical school or denied admission into medical school.

4) Medical School Admission Status: whether students were accepted into medical school or denied admission into medical school (after submitting their application(s)). Number of Medical Schools Applied to is a categorical variable with levels from 1 to 24. The data for the Number of Medical Schools Applied to has been divided into eight bins with a width of three. The first bin (1-3) is considered a very low number of schools applied to, the second bin (4-6) is considered a low number of schools applied to, the third bin (7-9) is considered a somewhat low number of schools applied to, the fourth bin (10-12) is considered a regular number of schools applied to (but on the lower end), the fifth bin (13-15) is considered a regular number of schools applied to (but on the higher end), the sixth bin (16-18) is considered a somewhat high number of schools applied to, the seventh bin (19-21) is considered a high number of schools applied to, and the eighth bin (22-24) is considered a very high number of school applied to.

**The Visualization Questions**

Our primary visualization question is the following: Is there a relationship between Cumulative GPA, Total MCAT score, and Number of Medical Schools Applied to?

Our follow-up question is the following: Does the relationship between Cumulative GPA, Total MCAT score, and Number of Medical Schools Applied to differ when measured across Medical School Admission Status?

**Goals and Outcomes**

As per our visualization questions, our primary goals are to 1) demonstrate the relationship between Cumulative GPA, Total MCAT score, and Number of Medical Schools Applied to, and 2) demonstrate any differences in the aforementioned relationship when students who were accepted into medical school and students who were denied admission into medical school are measured separately.

As we are answering our visualization questions, there is additional information we wish to communicate. Specifically, we wish to demonstrate any additional, pairwise relationships between Cumulative GPA, Total MCAT Score, Number of Medical Schools Applied to, and Medical School Admission Status.

Our hypothesis for the first visualization question is that as Cumulative GPA increases, Total MCAT score and Number of Medical Schools Applied to will also increase.

Our hypothesis for the follow-up question is that the aforementioned relationship between the three variables will be the same (it will be positive) in both groups, except that all three variables in the accepted into medical school group will be higher than those same variables in the denied admission into medical school group.

**R Code**

#Load necessary packages.  
library(tidyverse)

## -- Attaching packages --------------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.3 v dplyr 1.0.2  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(RColorBrewer)

## Warning: package 'RColorBrewer' was built under R version 4.0.3

#Import and rename data.  
OriginalDataset <- read\_csv("GPA and Medical School Admission.csv")

## Warning: Missing column names filled in: 'X1' [1]

## Parsed with column specification:  
## cols(  
## X1 = col\_double(),  
## Accept = col\_character(),  
## Acceptance = col\_double(),  
## Sex = col\_character(),  
## BCPM = col\_double(),  
## GPA = col\_double(),  
## VR = col\_double(),  
## PS = col\_double(),  
## WS = col\_double(),  
## BS = col\_double(),  
## MCAT = col\_double(),  
## Apps = col\_double()  
## )

#Keep only the variables for acceptance (Accept), GPA, verbal reasoning (VR), physical sciences (PS), biological sciences (BS), and number of applications (Apps).  
RelevantVariables <- OriginalDataset %>%  
 select(Accept, GPA, VR, PS, BS, Apps)  
  
#Create a new column that calculates the sum of scores for verbal reasoning, physical sciences and biological sciences.  
CompleteDataset <- RelevantVariables %>%  
 mutate(MCATScoreWithoutWritten = VR+PS+BS)  
  
#Rename levels of acceptance variable to make it easier to understand (no need to convert acceptance into character data, it is character data by default).

CompleteDataset <- CompleteDataset %>%  
 mutate(Accept = fct\_recode(Accept, "Accepted into Medical School" = "A", "Denied Admission into Medical School" = "D"))  
  
#Convert acceptance variable from character data into factor data.  
CompleteDataset <- CompleteDataset %>%  
 mutate(Accept = as.factor(Accept))  
  
#Convert Apps variable from numerical data into factor data, separate values into labeled intervals of 3, and rename them.  
Labelapps <- cut(CompleteDataset$Apps, breaks=c(0,3,6,9,12,15,18,21,24), labels=c("1 to 3", "4 to 6", "7 to 9", "10 to 12", "13 to 15", "16 to 18", "19 to 21", "22 to 24"), right = TRUE)  
  
#Create graph for main visualization question where the x and y axes encompass the entire possible range of scores.   
CompleteDataset %>%  
#Plot GPA and MCATScoreWithoutWritten, with points colored according to Number of Applications.  
 ggplot(aes(x = GPA, y = MCATScoreWithoutWritten, color = Labelapps)) +  
#Display chosen variables in a scatterplot.  
 geom\_point() +  
#Add title to the plot and rename the x-axis, y-axis, and legend.   
 labs(title = "Cumulative GPA vs Total MCAT Score", x = "Cumulative GPA",  
 y = "Total MCAT Score", color = "Number of Medical Schools Applied To") +  
#Change the limits of the x-axis to 0-4, and the y-axis to 0-45 with some extra expansion.  
 coord\_cartesian(xlim = c(0,4), ylim = c(0,45), expand = TRUE) +  
#Apply custom colour palette to Number of Applications variable.  
 scale\_colour\_brewer(type = "seq", palette = "Paired") +  
#Make the background of the graph panel white, make the major/primary grid lines light gray, make the minor/secondary grid lines light gray, and give the graph panel a black outline (without filling in the inside).  
 theme(panel.background = element\_rect(fill = "white"),  
 panel.grid.major = element\_line(colour = "grey90", size = 0.2),  
 panel.grid.minor = element\_line(colour = "grey90", size = 0.2),  
 panel.border = element\_rect(colour = "black", fill = NA))

Chart, scatter chart

Description automatically generated

#Create graph for main visualization question.  
CompleteDataset %>%  
#Plot GPA and MCATScoreWithoutWritten, with points colored according to Number of Applications.  
 ggplot(aes(x = GPA, y = MCATScoreWithoutWritten, color = Labelapps)) +  
#Display chosen variables in a scatterplot.  
 geom\_point() +  
#Add title to the plot and rename the x-axis, y-axis, and legend.  
 labs(title = "Cumulative GPA vs Total MCAT Score", x = "Cumulative GPA",  
 y = "Total MCAT Score", color = "Number of Medical Schools Applied To") +  
#Create regression line (without a shaded area to represent the confidence interval) and change the regression line colour to black.  
 geom\_smooth(method=lm, se=FALSE, col = "black") +  
#Change the limits of the x-axis to 2.5-4.  
 xlim(c(2.5,4)) +  
#Apply custom colour palette to Number of Applications variable.  
 scale\_colour\_brewer(type = "seq", palette = "Paired") +  
#Make the background of the graph panel white, make the major/primary grid lines light gray, make the minor/secondary grid lines light gray, and give the graph panel a black outline (without filling in the inside).  
 theme(panel.background = element\_rect(fill = "white"),  
 panel.grid.major = element\_line(colour = "grey90", size = 0.2),  
 panel.grid.minor = element\_line(colour = "grey90", size = 0.2),  
 panel.border = element\_rect(colour = "black", fill = NA))

## `geom\_smooth()` using formula 'y ~ x'

Chart, scatter chart

Description automatically generated

#Create graphs for follow-up visualization question.  
CompleteDataset %>%  
#Plot GPA and MCATScoreWithoutWritten, with points colored according to Number of Applications.  
 ggplot(aes(x = GPA, y = MCATScoreWithoutWritten, color = Labelapps)) +  
#Display chosen variables in a scatterplot.  
 geom\_point() +  
#Create two panels, one for each level of the variable Acceptance.  
 facet\_grid(~Accept) +  
#Add title to the plot and rename the x-axis, y-axis, and legend.  
 labs(title = "Cumulative GPA vs Total MCAT Score (across Medical School Admission Status)",  
 x = "Cumulative GPA", y = "Total MCAT Score",  
 color = "Number of Medical Schools Applied To") +  
#Create regression lines for both panels (without a shaded area to represent the confidence interval) and change the regression line colour to black.  
 geom\_smooth(method=lm, se=FALSE, col = "black") +  
#Change the limits of the x-axis to 2.5-4.  
 xlim(c(2.5,4)) +  
#Apply custom colour palette to Number of Applications variable.  
 scale\_colour\_brewer(type = "seq", palette = "Paired") +  
#Make the background of the graph panel white, make the major/primary grid lines light gray, make the minor/secondary grid lines light gray,   
#give the graph panel a black outline (without filling in the inside), and position the legend at the bottom of the plot and orient it horizontally.  
 theme(panel.background = element\_rect(fill = "white"),  
 panel.grid.major = element\_line(colour = "grey90", size = 0.2),  
 panel.grid.minor = element\_line(colour = "grey90", size = 0.2),  
 panel.border = element\_rect(colour = "black", fill = NA),  
 legend.position = "bottom", legend.direction = "horizontal")

## `geom\_smooth()` using formula 'y ~ xChart, scatter chart

Description automatically generated

**Limitations­**

Our data set has several limitations that must be addressed in order to properly understand the data. The first major limitation is our small sample size. While the entire sample size is above 30 (which is sufficient to assume that the data is normally distributed), it becomes significantly reduced when grouping participants based on Medical School Admission Status, with 30 students in the accepted group, and 25 students in the denied admission group. By increasing the sample size, the sample data should become more valid, as its variability will be more representative of the population. In terms of the visualization of our data, a limitation is that the relationships based on our visualization questions are difficult to see because they are very weak. This could be due to our small sample size, but we cannot be certain unless we had a larger sample.

Another limitation is regarding the characteristics of our sample. The data for this visualization was collected from a single liberal arts college in the Midwest, therefore, these participants may not be representative of the entire population of medical school applicants. Moreover, the results of this data visualization are not representative of the entire student population, as our sample consists entirely of students who are applying to medical school (who generally require a higher GPA than the average student).

Finally, while our graph does visualize the data in a manner that allows us to answer our visualization question, it may not be immediately clear what we are trying to prove through our visualization due to the large number of variables shown.