Data Story

Varun Nadgir

August 23, 2017

Introduction

The dataset I will be exploring is from **The College Scorecard**, a service meant to help prospective students with their college decision. By recording data regarding admissions, education, and financial activity (to name a few), a new student can try to find a best-fit school based on previous paths of success for students that match their demographic, field of interest, socio-economic status, and so on. In this project, I will be looking at standardized test scoring and popular majors. Do high SAT scorers only go to the very exclusive schools? Where are those schools? What degrees are popular there? Which schools award the most degrees in what I'm interested in? These are the sort of questions that prospective students are asking when considering their college options.

Preparing the Data

The download contains 19 .csv files - one for each academic year from 1996-'97 to 2014-'15, with approximately 7,500 schools (rows) per file and 1,744 recorded data points (columns). I began by creating a merged .csv from these 19 files, so that the data could be easily found for plotting. Before combining the .csv files, I had to add a new column that would indicate the data collection year (taken from the filename).

This process gives me a 2 GB file that contains all of the data, called **fulldata.csv**. The next step is to read **the data dictionary** to deciper the column names and try to find what is needed for plotting.

Data Analysis

This dataset holds a wide variety of information. It has *school identification data* (name, ID, location, level of institution), *admission data* (student demographics, SAT scores), *education data* (majors, completion rates), *financial data* (cost, aid, debt, repayment), and more. To begin plotting, it is a good idea to create subsets of fulldata.csv that contain just the fields necessary. This makes data manipulation and running calculations easier and safer.

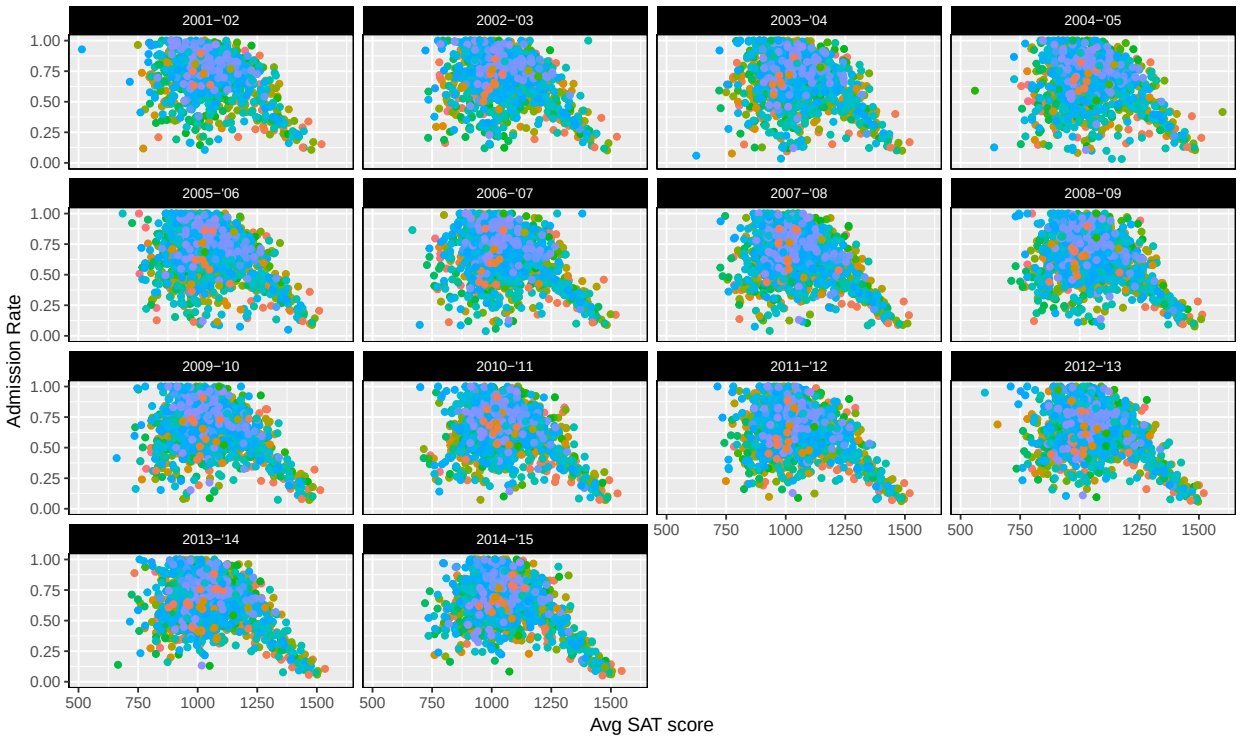
Admission Plot

```
# load libraries
library(ggplot2)
library(readr)

# import data for plot
sat_adm_data <- read_csv("~/Springboard/Foundations of Data Science/plotdata/sat_adm_data.csv")


# scatter plot with X as SAT AVG and Y as ADM RATE, facet by DATAYEAR, coloured by STABBR
ggplot(sat_adm_data,
       aes(x = as.numeric(sat_adm_data$SAT_AVG),
           y = as.numeric(sat_adm_data$ADM_RATE),
           col = factor(sat_adm_data$STABBR))) +
```

```
  geom_point(alpha = 0.5, size = 1.25) +
  labs(x = "Avg SAT score",
       y = "Admission Rate") +
  facet_wrap( ~ sat_adm_data$DATAYEAR) +
  scale_colour_discrete(name = "State/Territory", h = c(0, 270)) +
  theme(legend.position = "none",
        panel.border = element_rect(linetype = "solid",
                                    fill = "NA"),
        strip.background = element_rect(fill = "black"),
        strip.text = element_text(size = 8, colour = "white")) +
  geom_jitter()
```



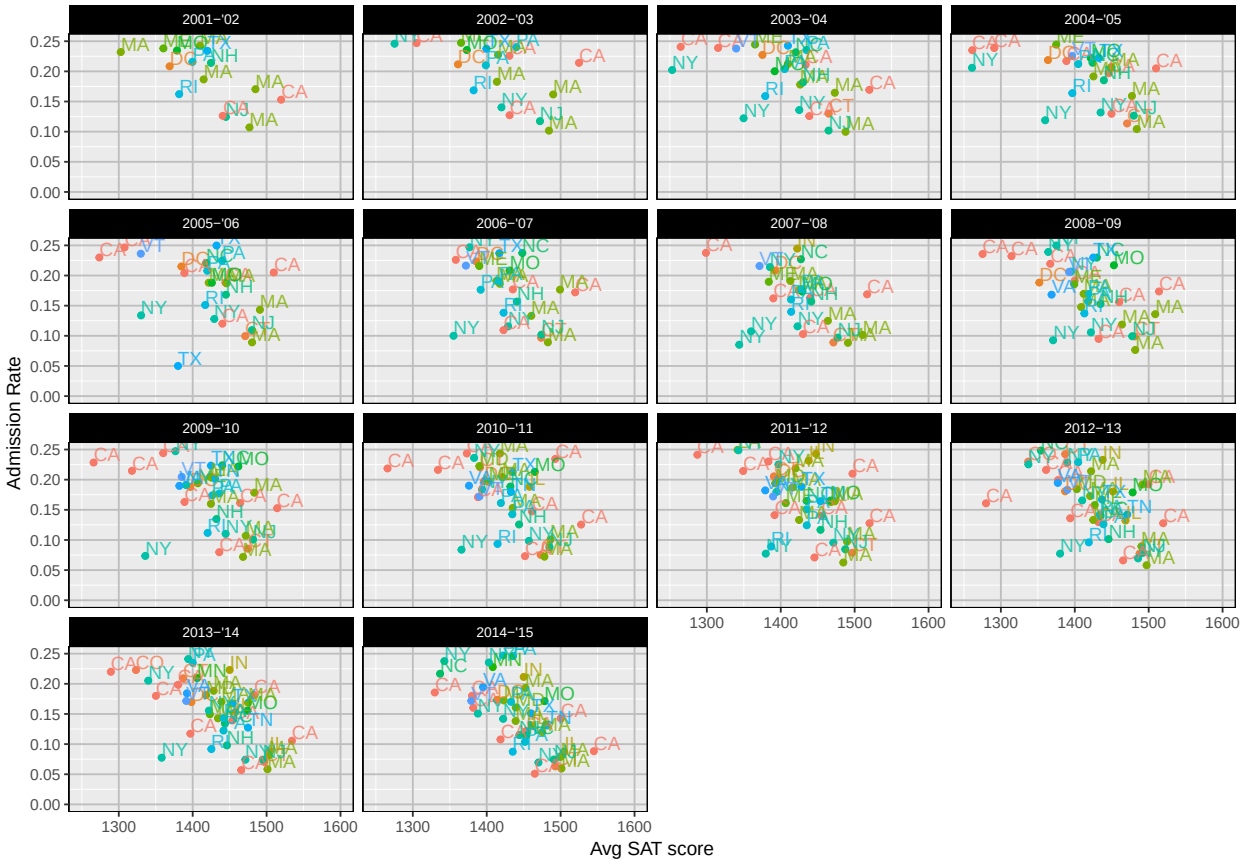
As we can see, students that attend schools with the lower admission rates generally have very high SAT score averages. The center of the distribution suggests that it is possible for low-mid SAT scores to get accepted to low admission rate schools as well - possibly due to extracurriculars or personal recommendations. The top-middle of the plot is a very dense cluster, which is reasonable since students with mid-range SAT scores apply to a lot of high admission rate schools. This is roughly what one might expect of this data. For a closer look, we can also see which states/territories contain the higher-performing, more exclusive schools by subsetting our dataset to just the scores > 1250 and admission rates < 0.25.

```
# load libraries
library(ggplot2)
library(readr)

# import data for plot
sat_adm_data_high <- read_csv("~/Springboard/Foundations of Data Science/data/sat_adm_data_high.csv")

# shrink plot range
ggplot(sat_adm_data_high,
```

```
      aes(x = as.numeric(sat_adm_data_high$SAT_AVG),
          y = as.numeric(sat_adm_data_high$ADM_RATE),
          col = factor(sat_adm_data_high$STABBR),
          label = sat_adm_data_high$STABBR)) +
geom_point(alpha = 0.5, size = 1.25) +
labs(x = "Avg SAT score",
     y = "Admission Rate") +
facet_wrap( ~ sat_adm_data_high$DATAYEAR) +
scale_colour_discrete(name = "State/Territory", h = c(0, 270)) +
theme(legend.position = "none",
      panel.grid.major = element_line(colour = "grey"),
      panel.border = element_rect(linetype = "solid",
                                  fill = "NA"),
      strip.background = element_rect(fill = "black"),
      strip.text = element_text(size = 8, colour = "white")) +
geom_jitter() +
coord_cartesian(xlim=c(1250, 1600)) +
geom_text(size = 4, alpha = 0.8, hjust = 0, nudge_x = 0.05, vjust = -0.1)
```



Since the data is much less dense, we can add labels. It seems like Massachusetts, California, and New York are some of the more recurring data points. This makes sense, since these states are home to multiple prestigious schools.

Majors Plot

This dataset also contains the percentage of degrees awarded for certain majors at each school. By taking averages, can we see if students from some states favour certain subjects? As a mathematician, I have a personal interest in seeing trends for Maths degrees. Unlike the last plot, this one will require some calculations.

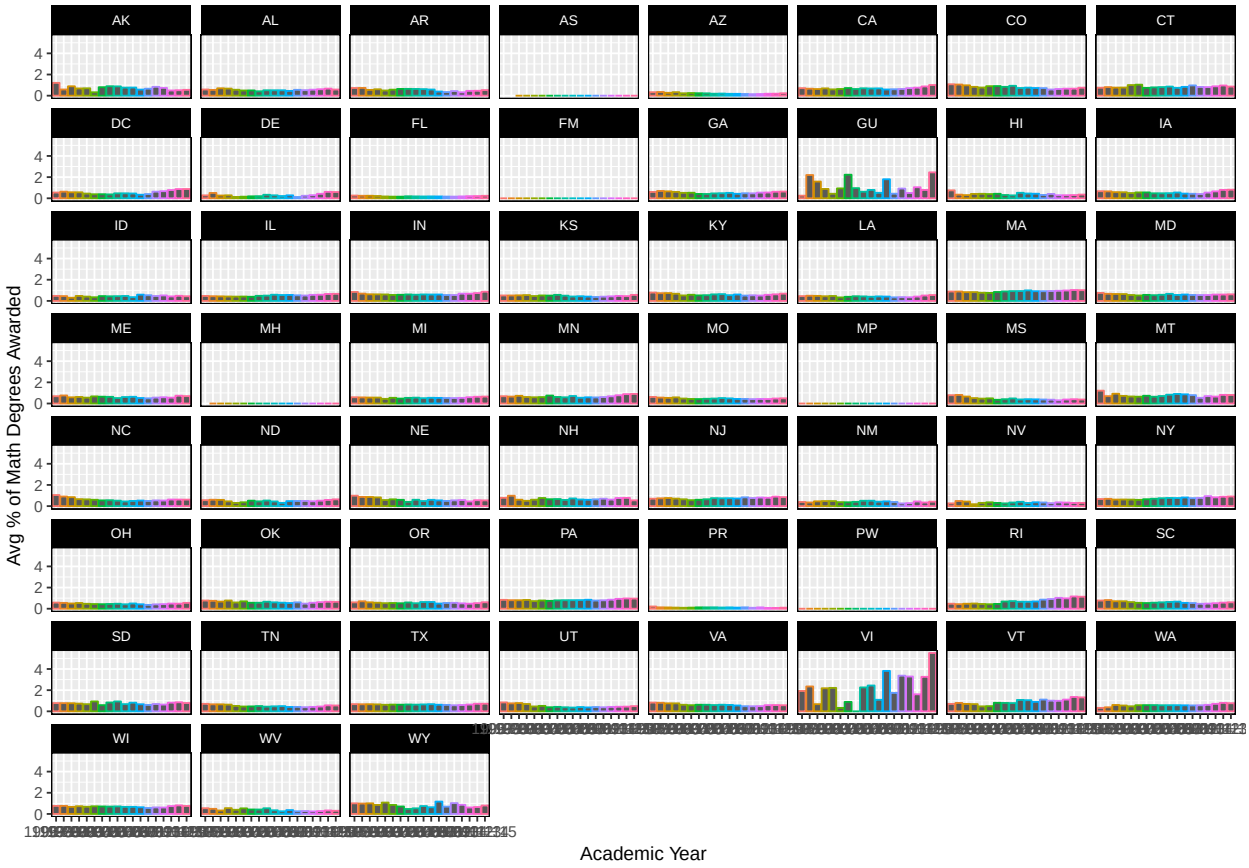
After noticing that many PCIP scores were recorded as "1" (meaning 100% of degrees were awarded in that major at that school in that year), it was clear that taking an average of this would cause some problems - we would end up with greater-than-100% averages in many cases. This presented issues when plotting because the data range would be inconsistent and a few outliers could ruin the plot even with lots of good data available. To account for this, I filtered the data a few different ways. I removed fields where UGDS = "NULL": records where the enrollment size was not recorded. I also removed ICLEVEL = 3 schools, which are "less than 2-year"" schools. Finally, I removed records with PREDDEG = 0, schools with "not classified" as their predominant degree type. This shrank the number of observations from 130,000 to 84,000, but the quality of the data was much better for taking averages and sums.

To get an average of degrees awarded for majors at the state level, I'll be using (0.25)*UGDS, which is the undergraduate enrollment size. This will be used as an approximation for the number of students in the graduating class. Then I will multiply that number by the PCIP fields, which are the percentage of degrees awarded. After grouping by STATE and DATAYEAR to find the sum of the graduating size and degrees awarded, we can divide to find a new percent for each major: PCIP_AVG.

```
# load libraries
library(ggplot2)
library(readr)

majors_data <- read_csv("~/Springboard/Foundations of Data Science/data/majors_data.csv")

# plot Avg % of Maths Degrees by STATE over YEARS
ggplot(majors_data,
       aes(x = majors_data$DATAYEAR,
           y = majors_data$PCIP27_AVG,
           col = majors_data$DATAYEAR)) +
  geom_bar(stat = "identity") +
  labs(x = "Academic Year",
       y = "Avg % of Math Degrees Awarded") +
  facet_wrap( ~ STABBR) +
  scale_colour_discrete(name = "DataYear") +
  theme(legend.position = "none",
        panel.border = element_rect(linetype = "solid",
                                    fill = "NA"),
        strip.background = element_rect(fill = "black"),
        strip.text = element_text(size = 8, colour = "white"))
```



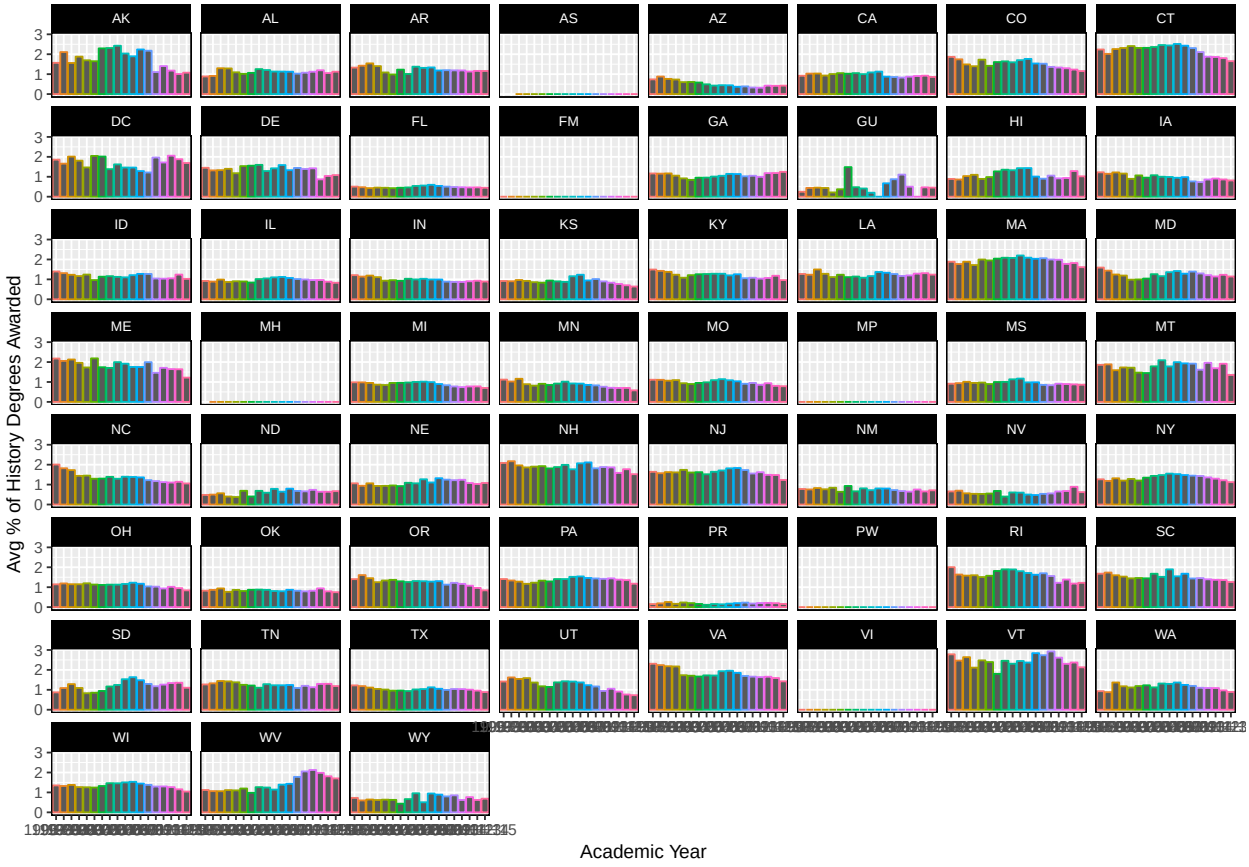
Now we are able to see, approximately, how many students from the collective graduating class of a state get a degree in mathematics. To my disappointment, Maths representation is so low that the Y-scale only goes from 0% to 6%. However, two of the US territories, Guam and the Virgin Islands, show some interesting trends. It is also very easy to recreate this plot for another major. To check trends for, say, History degrees, PCIP27 just needs to be changed to PCIP54.

```
# load libraries
library(ggplot2)
library(readr)

majors_data <- read_csv("~/Springboard/Foundations of Data Science/data/majors_data.csv")

# plot Avg % of History Degrees by STATE over YEARS
ggplot(majors_data,
       aes(x = majors_data$DATAYEAR,
           y = majors_data$PCIP54_AVG,
           col = majors_data$DATAYEAR)) +
  geom_bar(stat = "identity") +
  labs(x = "Academic Year",
       y = "Avg % of History Degrees Awarded") +
  facet_wrap( ~ STABBR) +
  scale_colour_discrete(name = "DataYear") +
  theme(legend.position = "none",
        panel.border = element_rect(linetype = "solid",
                                    fill = "NA"),
        strip.background = element_rect(fill = "black"),
```

```
        strip.text = element_text(size = 8, colour = "white"))
```



Immediately, the plot can be seen to be drastically different. The US territories are almost all at 0%, while the states have pretty healthy numbers. Though the automatic Y-scale is smaller in this plot, the overall distribution is just a little bit higher. A lot of the higher bars live in the 2-3% range, while in the Maths plot they barely reached 2%. The difference is that the Maths plot had an individual data point that was high enough to require expanding the scale, while the History plot is pretty consistently at a mid level. Does this mean History is more popular than Maths? It's possible, but we have to remember that we did use an approximation to plot this.

Conclusion

After preparing the data and trying a few initial plots, my approach feels good but has some room for improvement. The plots look nice and are legible, but it is possible to draw the wrong conclusions. In the case of the Majors data, I would like to set a Y-scale that makes it easier to compare the trends across multiple Majors. I would also like to be able to hover some points to see plot details and adding some general interactivity would make this dataset much more useful. I will be looking into Shiny and how to incorporate that into my next projects. Since my goal is to ultimately have some sort of web service or app that helps in giving school suggestions, the visualizations I provide will have to present the information in such a way that a layperson will be able to make reasonable conclusions from what they see. This means cleanliness of the data and clarity in the plotting are a must.