Data Story

Varun Nadgir

August 14, 2017

Introduction

The dataset I will explore in this project is from The College Scorecard, a service meant to help prospective students with their college decision. By recording data regarding admissions, predominant majors, 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. Ideally, it would be possible to suggest a school to a student based on their qualities, or suggest what qualities might be necessary to succeed at a more exclusive or personal reach school.

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.

```r
# load libraries
library(dplyr)
library(readr)

# import 19 .csv as individual data frames
MERGED1996_97_PP.csv <- read_csv("~/Springboard/Foundations of Data Science/project files/MERGED1996_97_

# after importing the .csv files, add the DATAYEAR column to all
MERGED1996_97_PP.csv <- MERGED1996_97_PP.csv %>% mutate(DATAYEAR = "1996-'97") # +18 more

# create and write 19 new .csv files which will be merged
write_csv(MERGED1996_97_PP.csv,
          path = "~./Springboard/Foundations of Data Science/data/MERGED1996_97_PP.csv",
          col_names = TRUE) # +18 more

# path to new folder that holds multiple .csv files
data_folder <- "~./Springboard/Foundations of Data Science/data/"

# create new list of all .csv files in folder
data_list <- list.files(path=data_folder, pattern="*.csv")

# read in each .csv file in file_list and rbind them into a data frame called fulldata
fulldata <-
  do.call("rbind",
          lapply(data_list,
                 function(x)
                 read.csv(paste(data_folder, x, sep=''),
                 stringsAsFactors = FALSE)))
```

```
# save fulldata as .csv locally
write_csv(fulldata,
          path = "~./Springboard/Foundations of Data Science/data/fulldata.csv",
          col_names = TRUE)
```

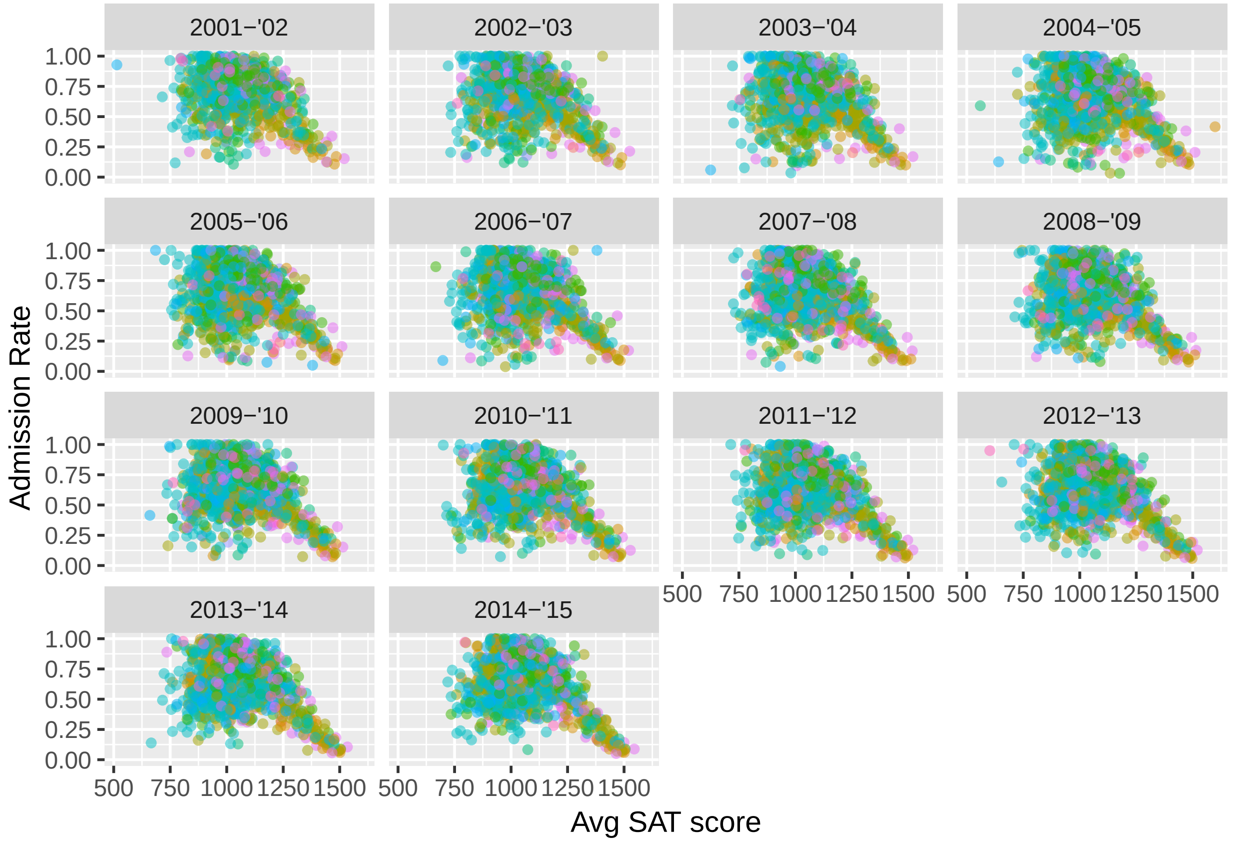
This process lets me create and store a merged 2 GB file that contains all of the data, along with the new column that I added as an identifier. The next step is to read the data dictionary to decipher the column names and try to find what is worth plotting and exploring deeper. Initially, I had thought that creating subsets of **fulldata.csv** would be a convenient way to organize data that I considered important; however, I've realized that it is more economic to create these subsets on the fly as needed for plotting.

Data Analysis

Fortunately, 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* (predominant majors, completion rates), and *financial data* (cost, aid, debt, repayment). At first, I had thought to create data frames for these 4 subsets - however, I would run into the problem of having difficulty comparing factors such as REGION and COMPLETION RATE, since they would be contained in two separate data frames when it came time to plot data. For this reason, it made more sense to create data frames for the purpose of plotting, and name them accordingly.

Admission Plot

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

# identify fields for this plot, export as .csv
sat_adm_fields <- c("UNITID", "OPEID", "OPEID6", "DATAYEAR", "INSTNM", "ALIAS",
                    "CITY", "STABBR", "REGION", "ADM_RATE", "ADM_RATE_ALL",
                    "SAT_AVG", "SAT_AVG_ALL")

# 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 REGION
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$REGION))) +
  geom_point(alpha = 0.5,
             size = 1.25) +
  labs(x = "Avg SAT score",
       y = "Admission Rate") +
  facet_wrap( ~ sat_adm_data$DATAYEAR) +
  scale_colour_discrete() +
  theme(legend.position = "none")
```



As we can see, students with the higher SAT scores ultimately go to schools with the lower admission rates (more exclusive). 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. This is roughly what one might expect of this data. Additionally, we can also see which regions tend to represent the higher-performing, more exclusive schools - the bottom right corner contains some schools from orange (New England), olive-green (Mid East), and pink (Outlying Areas) regions, while not having as many from blue (Southwest).

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.

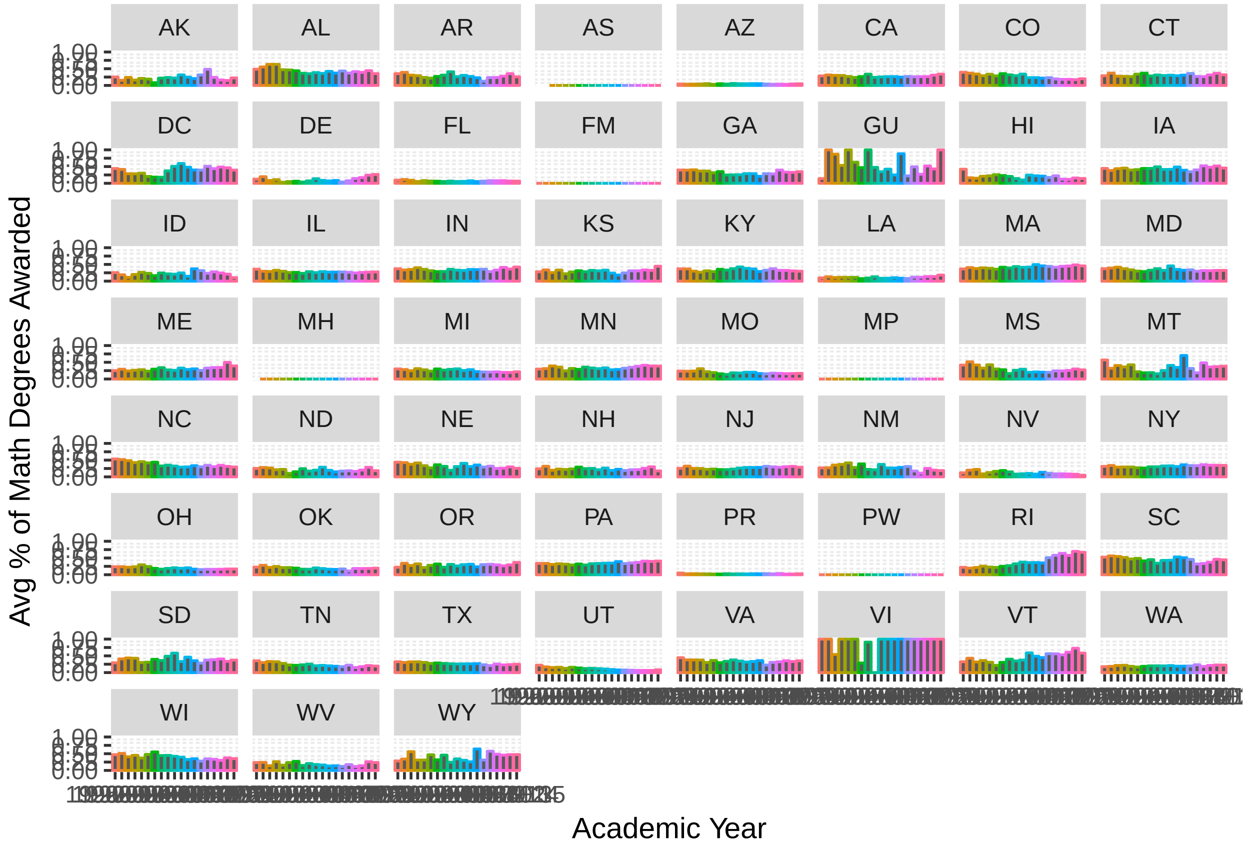
```
# identify fields for this plot
majors_fields <- names(fulldata) %in% c("UNITID", "OPEID", "OPEID6", "INSTNM", "REGION", "STABBR",
"PCIP01", "PCIP03", "PCIP04", "PCIP05", "PCIP09", "PCIP10", "PCIP11", "PCIP12", "PCIP13", "PCIP14",
"PCIP15", "PCIP16", "PCIP19", "PCIP22", "PCIP23", "PCIP24", "PCIP25", "PCIP26", "PCIP27", "PCIP29",
"PCIP30", "PCIP31", "PCIP38", "PCIP39", "PCIP40", "PCIP41", "PCIP42", "PCIP43", "PCIP44", "PCIP45",
"PCIP46", "PCIP47", "PCIP48", "PCIP49", "PCIP50", "PCIP51", "PCIP52", "PCIP54", "DATAYEAR")

# create raw data frame containing total data
majors_data_raw <- fulldata[majors_fields]

# isolate fields for majors only
```

```r
pcip_fields <- names(majors_data_raw) %in% c("PCIP01", "PCIP03", "PCIP04", "PCIP05", "PCIP09",
"PCIP10", "PCIP11", "PCIP12", "PCIP13", "PCIP14", "PCIP15", "PCIP16", "PCIP19", "PCIP22",
"PCIP23", "PCIP24", "PCIP25", "PCIP26", "PCIP27", "PCIP29", "PCIP30", "PCIP31", "PCIP38",
"PCIP39", "PCIP40", "PCIP41", "PCIP42", "PCIP43", "PCIP44", "PCIP45", "PCIP46", "PCIP47",
"PCIP48", "PCIP49", "PCIP50", "PCIP51", "PCIP52", "PCIP54")

# loop that creates dataframe called majors_data_PCIP for each major using SQL
# each table contains STABBR, DATAYEAR, and PCIP_AVG that has been calculated
# GROUP BY
for (p in 1:length(pcip_fields)) {
print(paste("Creating majors_data_",pcip_fields[p],"...", sep = ''))
major <- pcip_fields[p]
assign(paste("majors_data_",major, sep = ''),
            sqldf(paste("SELECT STABBR,
            (CASE WHEN (avg(",major,")*100)>=1 THEN 1 ELSE avg(",major,")*100 END) as ",major,"_AVG,
            DATAYEAR from majors_data_raw
            WHERE ",major," <1
            GROUP BY STABBR, DATAYEAR",
            sep = '')))
}

# list of data frames
df_list <- list(majors_data_PCIP01, majors_data_PCIP03, majors_data_PCIP04, majors_data_PCIP05,
majors_data_PCIP09, majors_data_PCIP10, majors_data_PCIP11, majors_data_PCIP12, majors_data_PCIP13,
majors_data_PCIP14, majors_data_PCIP15, majors_data_PCIP16, majors_data_PCIP19, majors_data_PCIP22,
majors_data_PCIP23, majors_data_PCIP24, majors_data_PCIP25, majors_data_PCIP26, majors_data_PCIP27,
majors_data_PCIP29, majors_data_PCIP30, majors_data_PCIP31, majors_data_PCIP38, majors_data_PCIP39,
majors_data_PCIP40, majors_data_PCIP41, majors_data_PCIP42, majors_data_PCIP43, majors_data_PCIP44,
majors_data_PCIP45, majors_data_PCIP46, majors_data_PCIP47, majors_data_PCIP48, majors_data_PCIP49,
majors_data_PCIP50, majors_data_PCIP51, majors_data_PCIP52, majors_data_PCIP54)

# initiate majors_data with first dataframe
majors_data <- majors_data_PCIP01

# loop through df_list and bind cols
for (d in 2:length(df_list)){
print(paste("Binding majors_data_",pcip_fields[d]," to majors_data...", sep = ''))
majors_data <- bind_cols(majors_data, df_list[d])
majors_data <- select(majors_data, -c(STABBR1, DATAYEAR1))
}
```

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 added a line that would set the average to 1 if the calculated result was greater than or equal to 1. While this doesn’t completely remove the outliers, it allows the data range to be consistent (0% to 100%) and I could mitigate the effect of outliers by looking at only a handful of majors.

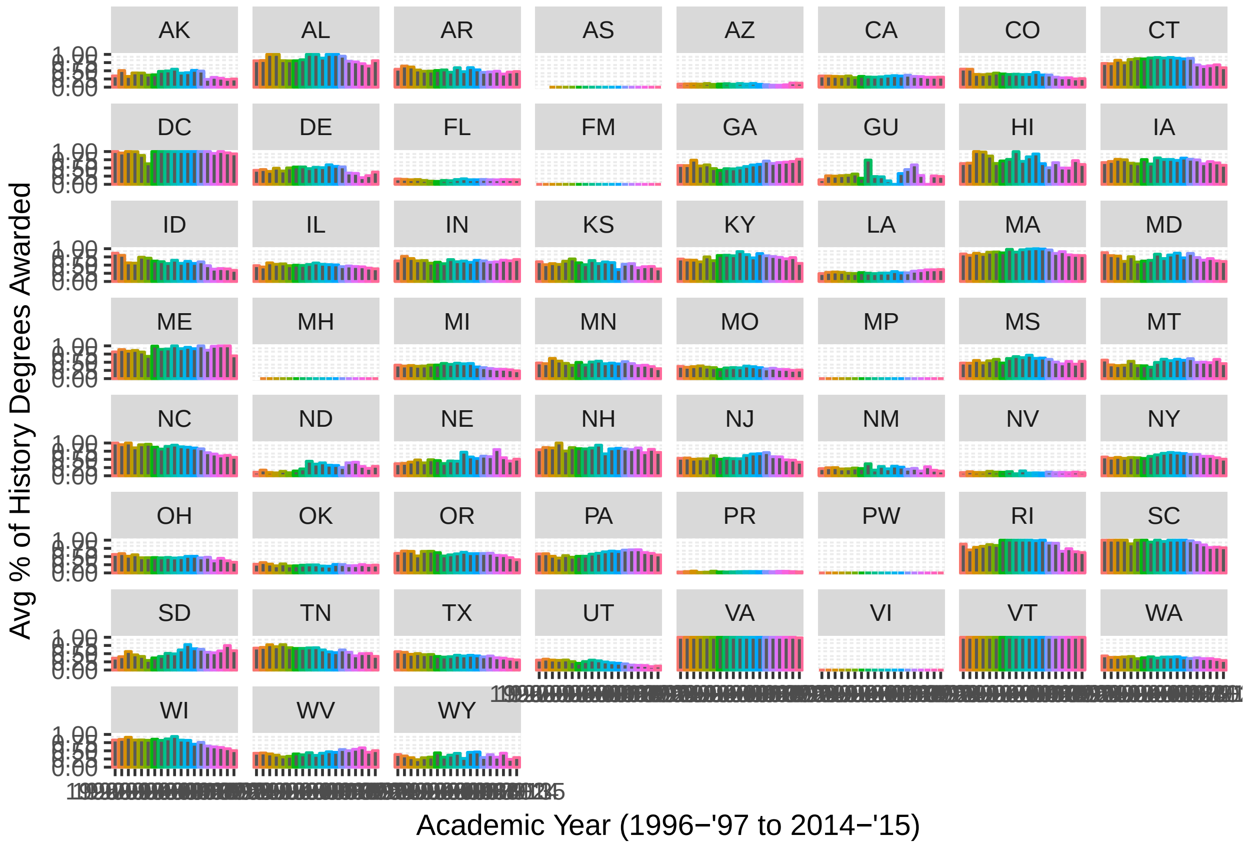
```r
# load libraries
library(ggplot2)
library(readr)

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() +
  theme(legend.position = "none")
```



Now we are able to see, approximately, how many students from the collective graduating class of a state get a degree in mathematics. Students from states like New York, New Jersey, and Illinois are represented about the same across the years, while Rhode Island has grown continually. Checking trends for other majors is also extremely simple. 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 (1996-'97 to 2014-'15)",
       y = "Avg % of History Degrees Awarded") +
  facet_wrap( ~ STABBR) +
  scale_colour_discrete() +
  theme(legend.position = "none")
```



There are some odd results in this plot, caused by the outliers mentioned earlier. One potential reason for this could be related to the programs available at certain schools. If History is one of the only subjects offered, a number close to 100% might be believable. This type of outlier would be legitimate and may need to be handled differently to avoid confusions when plotting, perhaps by only including schools with a minimum of ~5 programs offered.

Conclusion

After preparing the data and trying a few initial plots, my approach feels good but has obvious rooms for improvement. The plots look nice and are legible, but it can be easy to draw the wrong conclusions. Without being aware of the outliers in the Majors data, one might think that Vermont and Virginia only give History degrees. Since my hope is to provide a service to prospective students and their parents, the data presentation should be such that a specialist is not needed to explain results to them. In other words, clarity is of great importance and will be one of my main goals moving forward.