Capstone Project Stats Analysis

Varun Nadgir

August 8, 2017

The College Scorecard

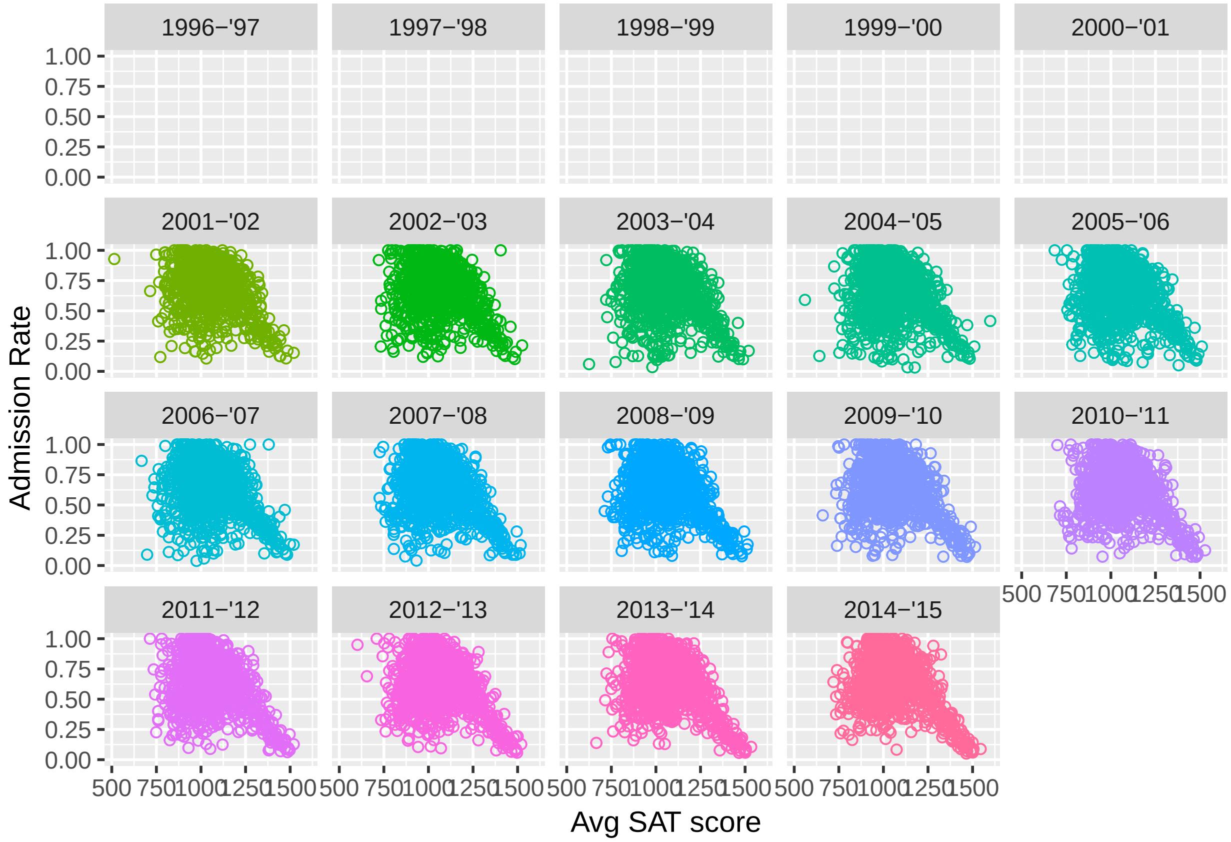
Statistical Analysis

By reading the data dictionary, I tried to determine which data points would be interesting to plot together. By now, the data is split into categorical data frames - **Admission**, **Location**, **Financial**, **Education**, and a fifth called **minidata** (containing a combination of elements from the other 4). First, I wanted to see the relationship between the average SAT score (or equivalent) and the admission rate of the school. My theory would be that the higher scores would get into the more exclusive schools, and we can check if that is roughly true.

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

# read admission data csv
admission_data <- read_csv("~/Springboard/Foundations of Data Science/data/admission_data.csv")

# scatter plot with X as SAT AVG and Y as ADM RATE, facet by DATAYEAR
ggplot(admission_data, aes(x = as.numeric(admission_data$SAT_AVG), y = as.numeric(admission_data$ADM_RAT
```



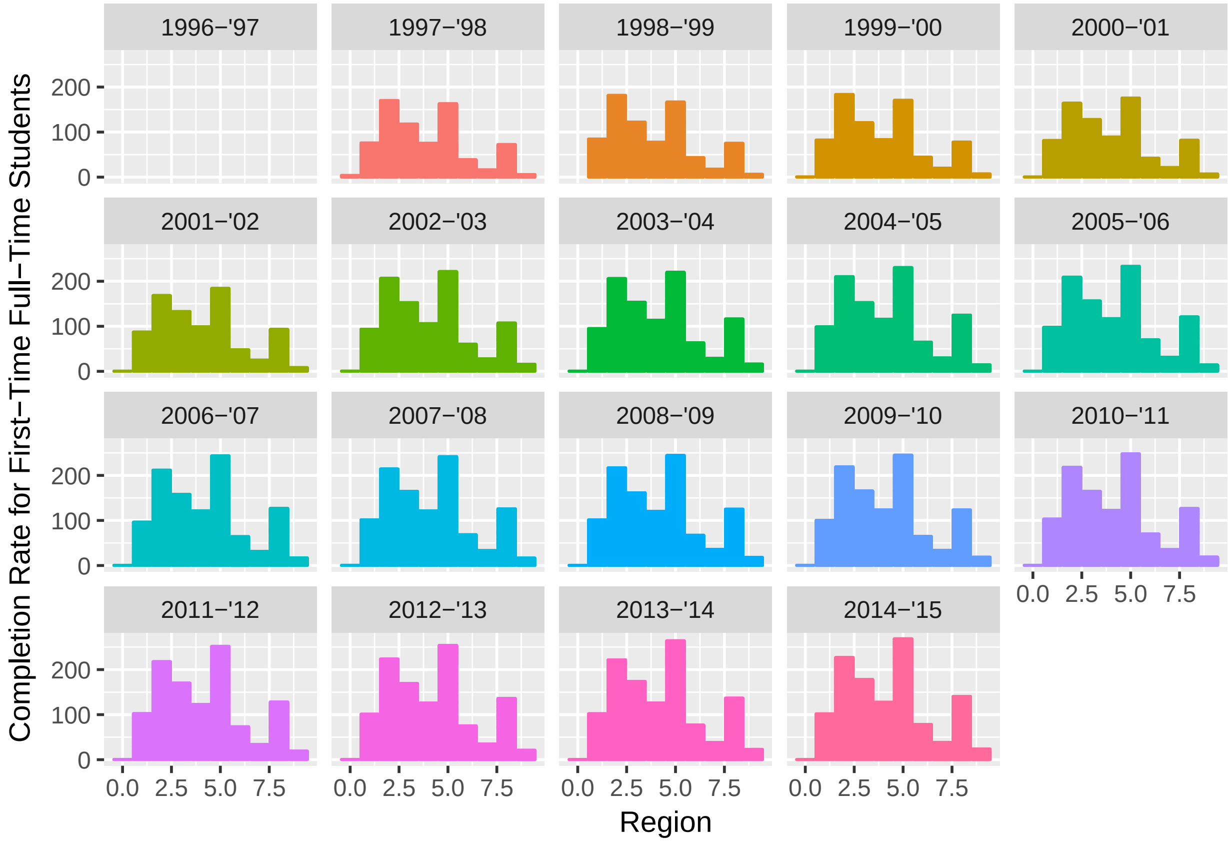
Apparently, there was no data collected between 1996-'97 and 2001-'01, but we can still see a pattern across the other years. There is a large bubble of data in the upper middle, and it falls steadily as the SAT scores get higher. This is pretty much what we expected - the students who have the higher SAT scores (~1500) regularly choose to go to the low admission rate schools (highly exclusive). Other schools with low admission rates sometimes accept low-mid range SAT scores, maybe because of some extracurriculars or other recommendations, but the majority of low-mid scores go to the higher admission rate schools.

Another relationship I explored was between completion rate (of first-time, full-time students at 4-year schools) and region of the school (New England, Rocky Mountains, South East, etc.). I'm not sure what to expect from this data, but I am personally curious to see how New England fares.

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

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

# bar plot with X as REGION and Y as C150_4, facet by DATAYEAR
ggplot(minidata, aes(x = minidata$REGION, y = as.numeric(minidata$C150_4), col = minidata$DATAYEAR)) +
```



Though it's not immediately obvious, the REGION in this dataset is stored as a number from 0 - 9. They are:

- 0: U.S. Service Schools
- 1: New England (CT, ME, MA, NH, RI, VT)
- 2: Mid East (DE, DC, MD, NJ, NY, PA)
- 3: Great Lakes (IL, IN, MI, OH, WI)
- 4: Plains (IA, KS, MN, MO, NE, ND, SD)
- 5: Southeast (AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, VA, WV)
- 6: Southwest (AZ, NM, OK, TX)
- 7: Rocky Mountains (CO, ID, MT, UT, WY)
- 8: Far West (AK, CA, HI, NV, OR, WA)
- 9: Outlying Areas (AS, FM, GU, MH, MP, PR, PW, VI)

This means that the US Service Schools, the left-most bar in the plot, have the lowest completion rate for first-time, full-time students. It also seems like the Southeast, while about equal with the Mid East at the beginning of data collection, has slowly climbed above the rest in recent years. The basic shape seems to stay consistent otherwise.