Nathan Chan
10/3/18
STA 141A
Assignment 1

1. ***What is the purpose of this data set? Who created it? What are the sources for the data?***

The US Department of Education created the College Scorecard data set so students and families could compare data from different colleges to determine what school is best for them. According to the Data Documentation for College Scorecard, the data is from "federal reporting from institutions, data on federal financial aid, and tax information."

2. ***How many rows are there? What do rows represent in this data set?***

There are 38068 rows. Each row represents the data from one college for one year. One row will have all the variable values that belong to one specific college for one specific year. A college may appear in multiple rows, each row representing a different year.

3. ***How many columns are there? What do columns represent?***

There are 142 columns. Each column represents a different variable. The variables describe a certain characteristic of the college. One column will have all the values of all the colleges for one specific variable.

4. ***What range of years does the data set span? How many colleges are recorded for each year?***

The range of years is from 2012 to 2016.
For 2012, there are 7793 colleges.
For 2013, there are 7804 colleges.
For 2014, there are 7703 colleges.
For 2015, there are 7593 colleges.
For 2016, there are 7175 colleges.

5. ***What are the 5 states with the most colleges? How many colleges do they have? What are the states with the fewest colleges? Make a hypothesis about why some states have a lot of colleges. Can you confirm your hypothesis (possibly using outside sources)?***

The states with the most colleges are

California with 717 colleges,
Texas with 454 colleges,
New York with 454 colleges,
Florida with 417 colleges, and
Pennsylvania with 382 colleges.

The states (territories, rather) with the fewest colleges are

American Samoa with 1 college,
Federated States of Micronesia with 1 college,
Marshall Islands with 1 college,
Northern Mariana Islands with 1 college, and
Palau with 1 college.

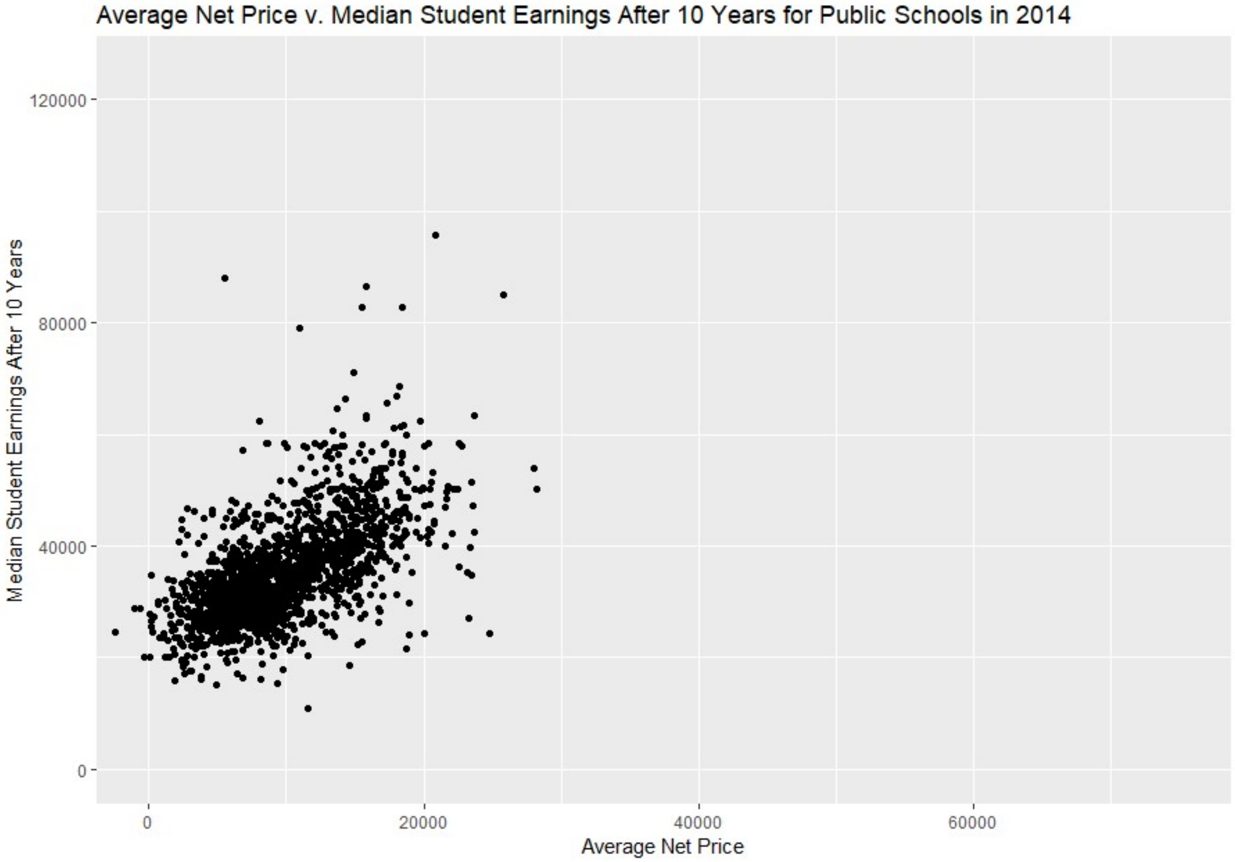
A hypothesis on why some states have a lot of colleges is the following: a state with a larger population will have more colleges than a state or territory with a smaller population.

Data from the United States Census Bureau supports my hypothesis. The top seven states with the highest population in 2016 are

California with 39,250,017,
Texas with 27,862,596,
Florida with 20,612,439 ,
New York with 19,745,289,
Illinois with 12,801,539,
Pennsylvania with 12,784,227, and
Ohio with 11,614,373

have the highest populations of the US states. Illinois has 280 colleges and Ohio has 320 colleges, making their college count sixth and seventh, so the top seven population states have the top seven number of colleges. The population and college count don't line up exactly probably because the population constantly changes over the years.

6. ***For public schools in the 2014 academic year, create a scatter plot of average net price versus median student earnings after 10 years (earn_10_yrs_after_entry.median). Comment on any patterns you see, interpreting what they mean for college students.***

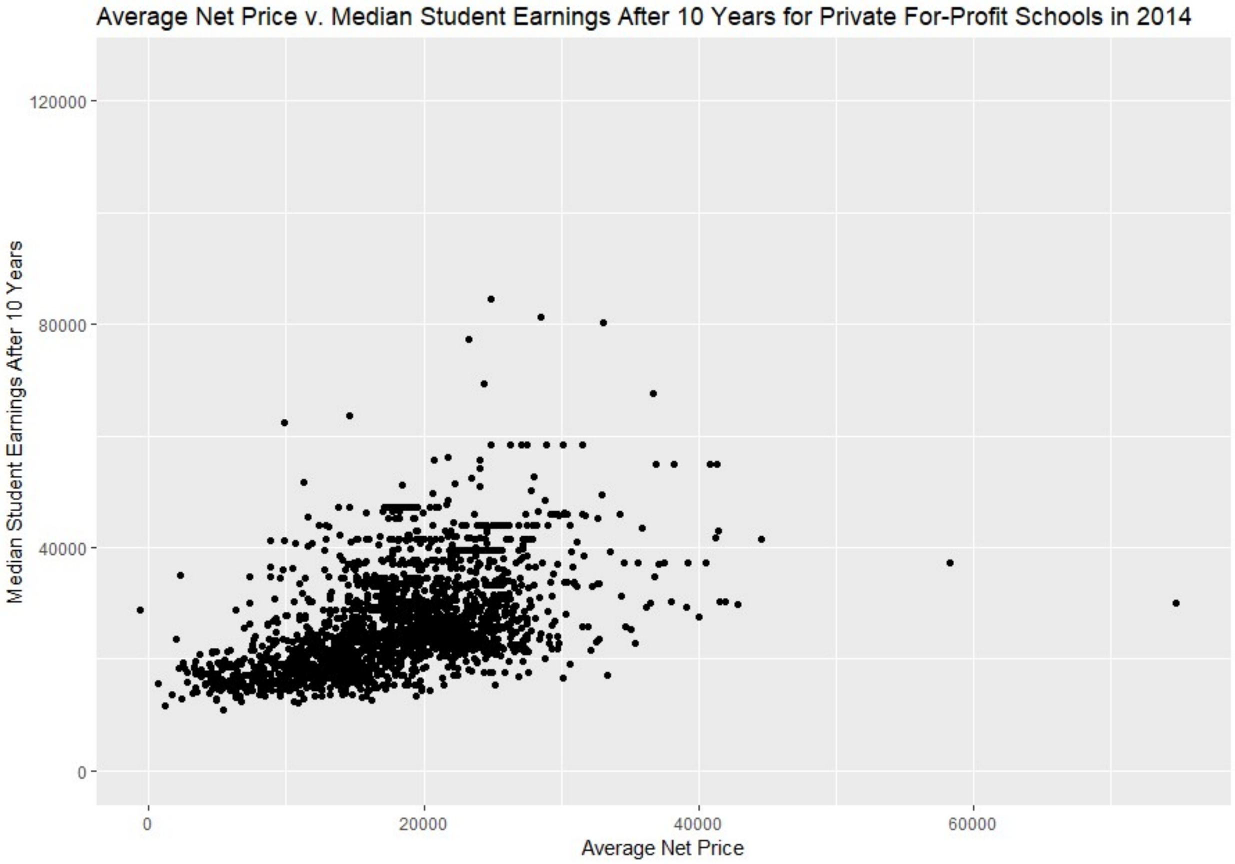


There is a clear positive relationship between Average Net Price and Median Earnings After 10 Years. Students who pay more for public school tend to make more money after 10 years. Those who pay less tend to make less.

There are a few outliers from students who do not conform to the general trend, so it is possible to pay less for public school and still make more than someone who paid more for public school, although it is very unlikely. Public school students should expect to pay more for college if they want to earn more in the future, and moving away from this norm is very unlikely.

7. ***Create the plot from the previous question for private for-profit schools. How do the two plots compare? Whether you see similarities or differences, discuss what your results imply about public schools and private for-profit schools.***

Note: All the scatter plots have been put on the same axis scale to make the comparison easier.



The private for-profit school plot follows the same general positive relationship as the public school plot: the more a student pays for college, the more they will, on average, make; but, of course, there are outliers and there is a lot of spread as private for-profit school price increases. Paying a lot for college will not guarantee higher earnings. Paying little for college, however, will tend to guarantee lower earnings.

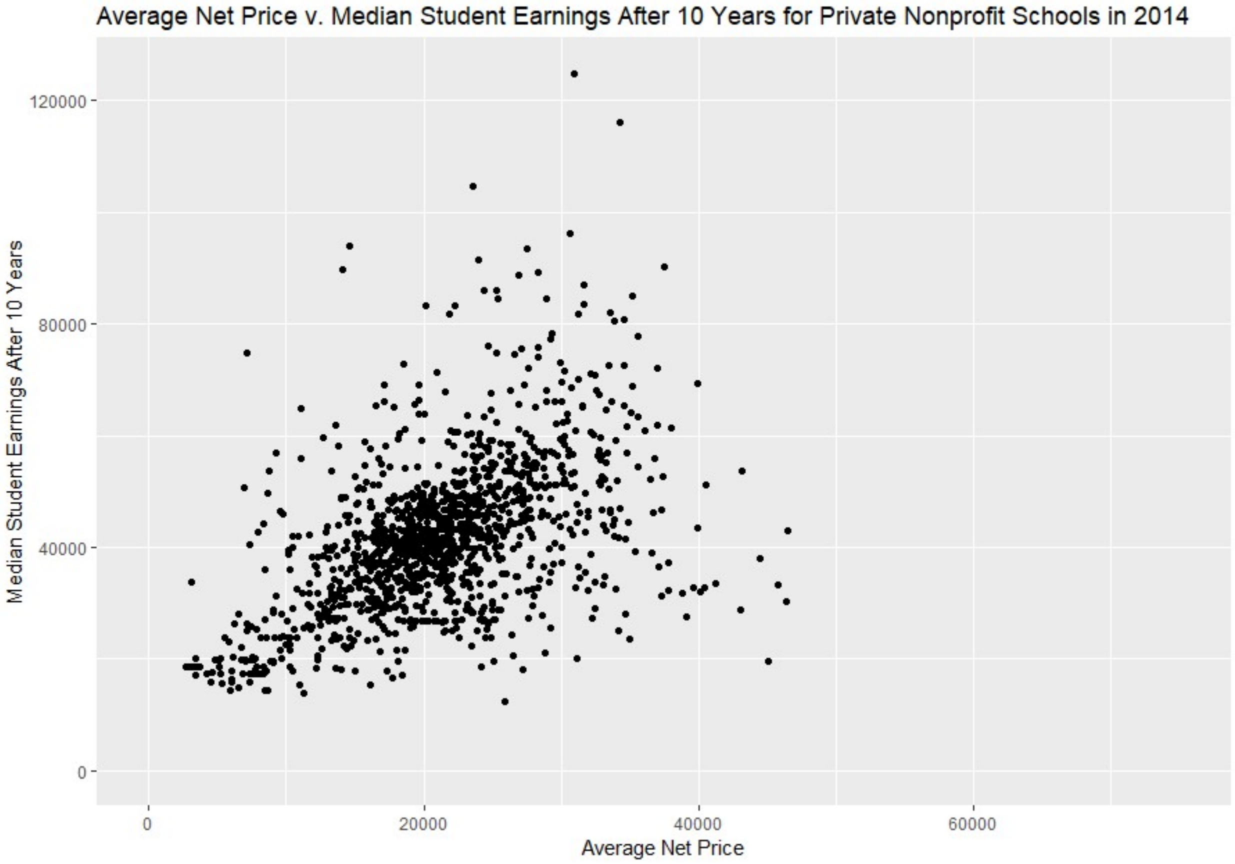
The slope of this private school plot is also lower than the slope of the public school plot, so public school may be better in terms of price and future earnings.

For private for-profit schools, the costs can go much higher. While the majority of the prices for both public and private are in a similar range, there are many private for-profit schools that cost more than the most expensive public school.

There is more variance in the private for-profit plot, especially as the cost goes up. The bottom of the scatter is almost flat, indicating a change in price of private for-profit school may not affect earnings at all. Paying less for private for-profit school will likely lead to a lower salary, but paying more for college may lead to either a lower or higher salary. This variance with an increase in private for-profit school price is very different from the public school plot: generally, paying more for public school will lead to higher earnings than paying less for public school. In private for-profit college, paying more may or may not lead to higher earnings.

There is no easy distinction between the average earnings after going to a public college compared to a private college. Deciding on a public or private college depends on many factors that are not in this analysis, the most important factor probably being the major. Of course, private for-profit schools will charge more than public schools, and students must make sure they are getting a quality education for the higher price they pay. Looking at the plot, it looks like many students pay a lot for little return.

8. ***Continuing from the previous two questions, what can you say about private non-profit schools? Use evidence to support your claims.***

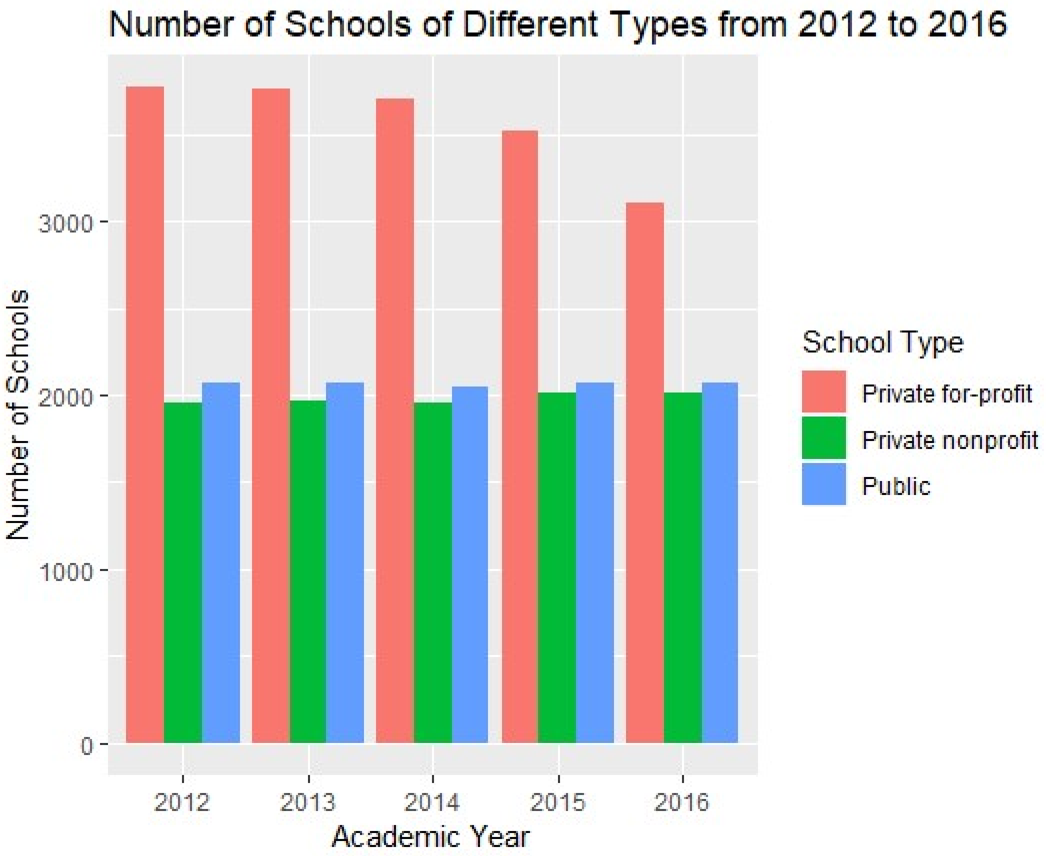


The general trend of this plot is the same as the previous two. Higher price tends to lead to higher earnings. The slope of this private nonprofit plot is similar to the slope of the public school plot. It seems that both private nonprofit schools and public schools are better than private for-profit schools in terms of price and future earnings.

The private nonprofit school plot differs from the public school in terms of spread. This spread is very similar to the spread in the private for-profit school plot. Paying less likely leads to earning less, but paying more may or may not lead to earning more. Paying more for private school instead of public school may lead to higher earnings, but it also may not.

This trend in private schools of paying more, but not necessarily making more money than paying less, conveys one important thing (out of many important things): paying more for private school is not always worth it. Students must make sure they are getting a proper education in preparation for a job in the future, and know that high earnings are not guaranteed by paying more for college. An important consideration is one’s major, the skills that are taught, and the opportunities different schools give.

9. ***Create a bar plot that shows the number of schools recorded for each year. Use a separate color for each year and a separate group of bars for each kind of ownership. Are there any trends? Comment on what you see.***



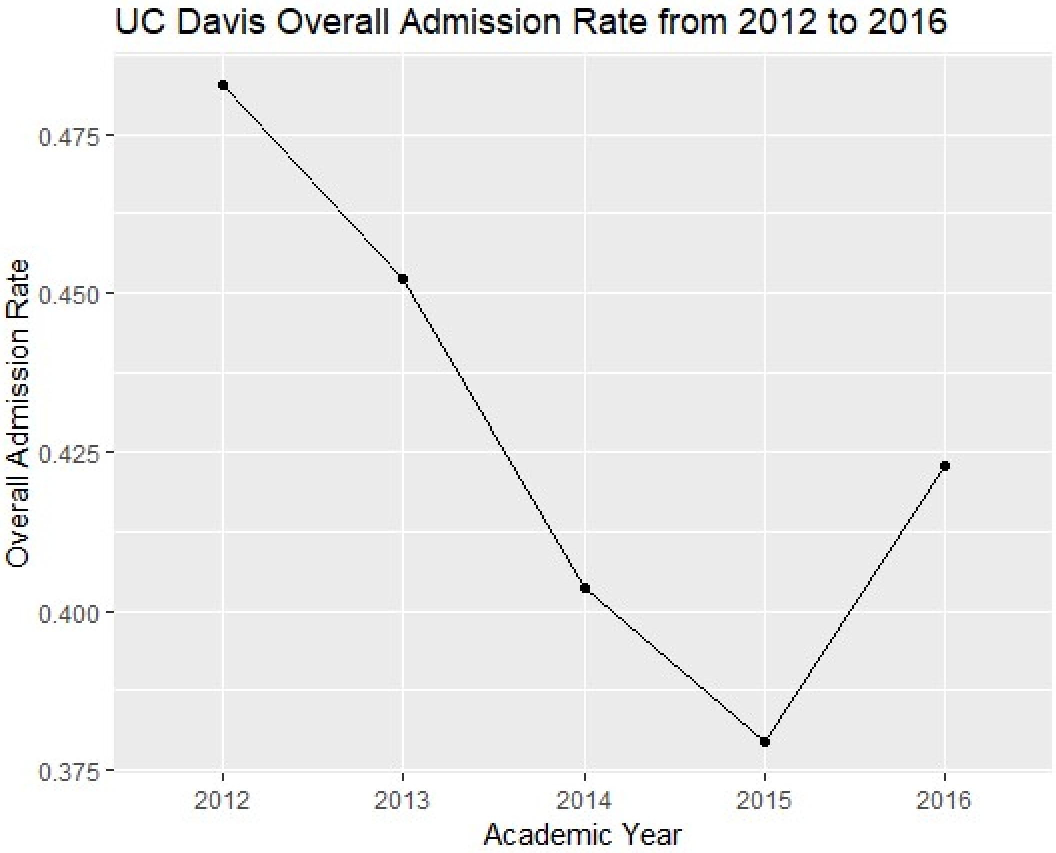
From 2012 to 2016, the number of public schools has stayed the same, the number of private nonprofit schools has risen slightly, and the number of private for-profit schools has dropped significantly.

The reason for the drop in number if private for-profit schools is possibly the access to information for students: people see that perhaps going to a private for-profit school is may not be worth it, considering there is no guarantee of higher earnings.

A question may be, if there are less private for-profit schools, then why have the number of other types of schools not risen? Where are the students going? This is due to the increase of enrollment of the private nonprofit and public schools. While the number of schools stays the same, the number of students in the school increases.

Also, this bar plot does not mean that there are more students who go to private for-profit schools. Those types of schools may be smaller that larger public and private nonprofit schools. The environment of these smaller schools may be a reason that students still want to pay more for school, without the guarantee of higher earnings.

10. ***The ID for UC Davis is 110644. Create a line plot (with points marked) that shows the admission rates for UC Davis for each year. Do the admission rates change much from year to year?***



The UC Davis admission rate changed dramatically from 2012 to 2016. From 2012 to 2015, the admission rate dropped by about 10%, which is very significant. In 2016, the admission rate increased. The admission rate at UC Davis does not seem to be fixed and appears to fluctuate a lot as the years go on, likely due to changing demographics of student applicants (an increase in international students or out-of-state students, for example) or changes in funding for UC Davis.

11. ***Discuss the R data types and statistical data types of the features. Does each R data type map to just one statistical data type? Give examples.***

R data types are built-in methods of holding data in R, and each data type holds the data in different ways. Some popular R data types include strings, scalars, vectors, matrices, data frames, and lists.

Statistical data types are ways we organize data in the real world. Data types include categorical data (divided into nominal and ordinal data), numerical data (divided into discrete and continuous data), and other data, like image, spatial, and text data.

R data types do not coincide with statistical data types. The two are very different in terms of their representation and how we use them and how they interact with the data. R data types are for people to organize the data in the program to use it in the computer. Statistical data types are for people to organize the data conceptually, and in their heads; we group data in the world with different categories, depending on their properties.

One of the challenges of programming in R is how to represent the real-life, statistical data type data we have in R data types so we can use the real-life data in the computer. We can represent categorial data as a string, such as "red" or "blue" or "small" or "large." We can represent numerical data as a scalar, such as a simple 1, 10, or 40. The other R data types, like vectors, matrices, data frames, and lists, are tools in R for us to use to manipulate the data within he program, and the closest thing in it to real-life would be a table or chart of data.

12. ***List 3 questions you think can be answered with this data set. For each question, explain why the question is compelling (Who would benefit from knowing the answer, and why?), which variables you would use to answer the question, and how these variables help you answer the question. You do not need to write any code for this problem.***

1) What are the average racial demographics for the private schools and public schools?

Different schools have different policies on their admissions. Public schools are legally not allowed to make any admissions decisions based on race, but private schools can have their own policies regarding how they factor race into their admissions decisions. There is a lawsuit right now against Harvard, saying that Harvard discriminates against Asians. It would be interesting to see how demographics differ between public and private schools, and how affirmative action or other diversity policies affects students of different backgrounds. (Note that we don't have any data on the demographics of applicants, so we can't make too concrete conclusions.) We would use the "ownership" variable to determine whether a school is public or private, and then use

“demographics.race_ethnicity.white”
“demographics.race_ethnicity.black”
“demographics.race_ethnicity.asian”
“demographics.race_ethnicity.nhpi”
“demographics.race_ethnicity.two_or_more”
“demographics.race_ethnicity.non_resident_alien” and
“demographics.race_ethnicity.unkown”

to determine race percentages for the colleges. We could create a bar plot with two bars: one for private and one for public. We would then fill the bars with the average demographics for private schools and public schools to determine the average difference between the two.

2) How does the enrollment of women in college change over time?

Today, women are being encouraged to pursue fields outside of the “traditional norm” more than ever. This increase of women in college will inherently lead to a decrease of men in college. Answering this question would be helpful to social scientists who study trends relating to gender. The variables we would use are “academic_year” (giving the year) and “demographics.women” (giving the percentage of women at the college) to chart the possible changing enrollment of women in college over the years. We would then plot the two variables on a scatter plot, with year being the independent variable and percentage of women the dependent variable.

3) Do SAT and ACT scores affect future earnings?

It would be interesting to see whether a college’s average SAT score and median ACT score are a predictor for earnings after ten years. This can help answer questions on how useful these tests actually are in the long run (not just as a filter for college admissions), and whether being good at taking tests in high school is an indication of future success. We would use “earn_10_yrs_after_entry.median” to determine future earnings and use “sat_scores.average.overall” and “act_scores.midpoint.cumulative” to get the median SAT and median ACT scores for the colleges. We would make two plots, one for SAT scores and one for ACT scores, and then plot the score as the independent variable and the earnings and the dependent variable. We’ll disregard the year, because it is not relevant in our analysis, and every year, the college will have the scores and the future earnings available, so it is not necessary to make a distinction for the year.

R Code Appendix:

```
# Assignment 1

scorecard = readRDS("college_scorecard.rds")

#2
nrow(scorecard)

#3
ncol(scorecard)

#4
range(scorecard$academic_year)
table(scorecard$academic_year == 2012)
table(scorecard$academic_year == 2013)
table(scorecard$academic_year == 2014)
table(scorecard$academic_year == 2015)
table(scorecard$academic_year == 2016)

#5
install.packages("plyr")
library("plyr")

scorecard2016 = subset(scorecard, scorecard$academic_year == 2016)

state_freq = count(scorecard2016, "state")
state_freq
sort(state_freq$freq)

# largest college counts: 717 454 454 417 382
# fewest college counts: 1 1 1 1 1

table(scorecard$state == "CA" & scorecard$academic_year == 2016)
table(scorecard$state == "TX" & scorecard$academic_year == 2016)
table(scorecard$state == "NY" & scorecard$academic_year == 2016)
table(scorecard$state == "FL" & scorecard$academic_year == 2016)
table(scorecard$state == "PA" & scorecard$academic_year == 2016)

table(scorecard$state == "AS" & scorecard$academic_year == 2016)
table(scorecard$state == "FM" & scorecard$academic_year == 2016)
table(scorecard$state == "MH" & scorecard$academic_year == 2016)
table(scorecard$state == "MP" & scorecard$academic_year == 2016)
table(scorecard$state == "PW" & scorecard$academic_year == 2016)

#6

library(ggplot2)
scorecard2014 = subset(scorecard, scorecard$academic_year == 2014)

# public schools
ggplot(scorecard2014, aes(x = avg_net_price.public, y =
earn_10_yrs_after_entry.median)) +
  geom_point() +
  labs(title = "Average Net Price v. Median Student Earnings After 10 Years for
Public Schools in 2014",
       x = "Average Net Price", y = "Median Student Earnings After 10 Years") +
  coord_cartesian(xlim = c(0, 75000), ylim = c(0, 125000))

#7 private, for-profit schools
ggplot(scorecard2014[scorecard2014$ownership == "Private for-profit", ], aes(x =
avg_net_price.private,
                                                                          y =
earn_10_yrs_after_entry.median)) +
  geom_point() +
  labs(title = "Average Net Price v. Median Student Earnings After 10 Years for
Private For-Profit Schools in 2014",
       x = "Average Net Price", y = "Median Student Earnings After 10 Years") +
  coord_cartesian(xlim = c(0, 75000), ylim = c(0, 125000))

#8 private, non-profit schools
ggplot(scorecard2014[scorecard2014$ownership == "Private nonprofit", ], aes(x =
avg_net_price.private,
                                                                            y =
earn_10_yrs_after_entry.median)) +
  geom_point() +
  labs(title = "Average Net Price v. Median Student Earnings After 10 Years for
Private Nonprofit Schools in 2014",
       x = "Average Net Price", y = "Median Student Earnings After 10 Years") +
  coord_cartesian(xlim = c(0, 75000), ylim = c(0, 125000))

#9

ggplot(scorecard, aes(x = academic_year, fill = ownership)) +
  geom_bar(position = "dodge") +
  labs(title = "Number of Schools of Different Types from 2012 to 2016", x =
"Academic Year", y = "Number of Schools") +
  guides(fill = guide_legend(title = "School Type"))

#10

# UC Davis scorecard
davis_scorecard = subset(scorecard, scorecard$id == 110644)

ggplot(davis_scorecard, aes(x = academic_year, y = admission_rate.overall, group =
1)) +
  geom_line() +
  geom_point() +
  labs(title = "UC Davis Overall Admission Rate from 2012 to 2016", x = "Academic
Year", y = "Overall Admission Rate")
```