Nathan Chan
10/11/18
STA 141A
Professor Ulle

Assignment 2

1. ***Are there any features with no missing values? Which features have the most missing values? Explore the missing values and report any patterns you find.***

There are 128 features that contain a missing value, out of the total 142 features. There are ten features that lack data completely, which are the following:

Ten Features with Data Completely Missing

Feature in Dataset	Description: Is the college a...
"minority_serving.historically_black"	Historically black college
"minority_serving.predominantly_black"	Predominantly black college
"minority_serving.annh"	Alaska Native/Native Hawaiian-serving college
"minority_serving.tribal"	Tribal college
"minority_serving.aanipi"	Asian American/Native American-Pacific Islander-serving college
"minority_serving.hispanic"	Hispanic-serving college
"minority_serving.nant"	Native American Non-Tribal college
"men_only"	Men-only college
"women_only"	Women-only college
"operating"	Currently operating college

Looking at the missing values for each column over all five years, it appears that there is more missing data for the features that have less importance and the features that are likely not reported on federal reporting, federal financial aid information, and tax information (where the data is from). Here are five different levels of missing data:

1) There is no missing data on features like the name of the school, the city, the state, the year, or the number of degrees rewarded.
2) There is not as much missing data on features like student demographics, percentage of degrees awarded in each program, or information on loans.
3) There is a bit more missing data on features like salary and tuition amounts.
4) There is a ton of missing data on features like price of college by income level and test scores.
5) There is data completely missing in the features named in the table.

Going down the levels, the features become less and less relevant to source of the data. Basic geographical information is easily obtainable, but data that relates to the extent that the colleges serves minorities, which are completely empty, is harder to obtain, especially from financial information. Another reason for this missing data is that it is missing on accident; otherwise, there is no reason for the feature to be there.

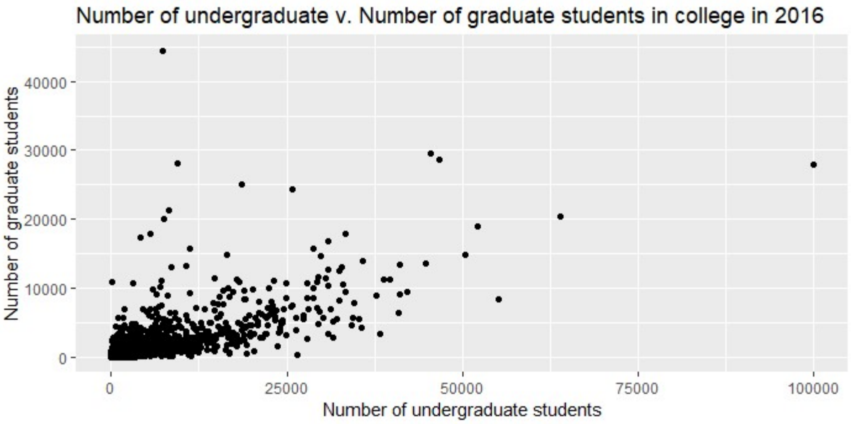
Some colleges may not even report their information, or none of the students going there require financial aid, which is why some data is missing; for example, the demographics information and program percentages information is consistently missing, probably from the colleges who simply don't report that information. (Only the colleges who participate in the specific federal aid programs need to report the information.)

Another interesting trend is that as the years go by, the number of total missing data goes down. This may be caused by better reporting with increased access to the Internet. This could also be because of the decrease in private for-profit colleges over the years. Private for-profit colleges are the main source of the missing data, likely because of their lack of federal reporting, and less private for-profit colleges means less missing data.

2. ***Explore student populations for the universities. Are there any schools with unusual populations? What is the relationship between undergraduate and graduate populations? Are there exceptions to the relationship?***

Note: I will use data only from 2016 to have a consistent analysis of population.

In 2016, the School of the Museum of Fine Arts at Tufts University had 0 students. While strange, this is likely because the School of the Museum of Fine Arts became part of Tufts University in 2016 (https://en.wikipedia.org/wiki/School_of_the_Museum_of_Fine_Arts_at_Tufts) so the number of students would be included under Tufts University. There are also a few specialized colleges with a small number of students. The biggest school was the University of Phoenix, with 100,011 undergrad students and 27,918 grade students. Out of 7,175 colleges in 2016, there are 728 missing values.



There is a very clear positive relationship between the number of undergraduate and number of graduate students. If the undergraduate population is larger, the graduate population will probably also be large. Most schools follow the general trend of a milder slope, with there being more undergraduate students than graduate students. There are a few exceptions from schools that have more grad students than undergraduate students, a few notable ones being Walden University, University of Southern California, Columbia University, Harvard University, Johns Hopkins University, and University of Pennsylvania.

3. ***Explore the program percentages for the universities. What programs are the most popular? What programs are the least popular? Are there any program percentages that show patterns different from the others?***

The largest average program percentage in 2016 was the health program, with an average of 26.79% across all the colleges. That means, at the average college, 26.79% of the degrees awarded at that college were from the college's health programs. Another program that comes close is the personal and culinary services programs, with an average of 20.71% of the degrees awarded at the average school. No other program has comparatively high percentages, but these high percentages are likely because of a high number of specialized schools that have 100% of the degrees awarded are from health or culinary programs, therefore inflating the average.

Many of the schools have decent average percentages because there are a lot of schools that specialize in those programs, where 100% of the degrees awarded are for a particular program. But the two programs with the lowest percentage are the military and library programs. For the military technologies and applied science programs, at the average school, 0.01% of the degrees awarded were from this program. For the library science program, 0.00% of the degrees awarded were from this program at the average school. At one school, the University of Maine at Augusta, had 4% of their degrees awarded belonging to the library program, which was the highest percentage out of all the schools in the US. The library program is by far the least popular program.

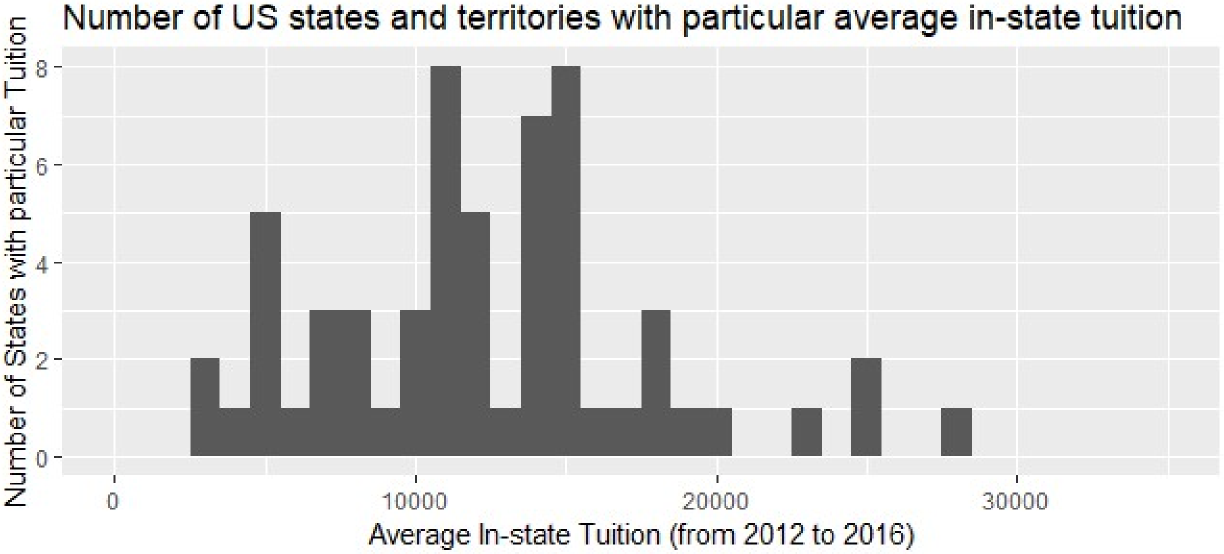
It is interesting that the highest program percentage for mathematics and statistics was 14.06%, at the California Institute of Technology. Nearly all of the programs have at least one school that specializes in that program (where the maximum percentage would be 100%) but that is not the case for mathematics. While mathematics is a fundamental field taught nearly everywhere, there is no school in the US that specializes in it. This situation extends to history programs, language programs, and science and technology programs. While taught universally, there are no highly specialized schools for them. This may be because teaching these programs inherently belongs together with other programs; for example, teaching a math and engineering together makes sense, and a school might choose to specialize in math and engineering and related fields in general, compared to only specializing in math. Compare this to culinary school, where other programs are not necessarily relevant and don't utilize the same resources.

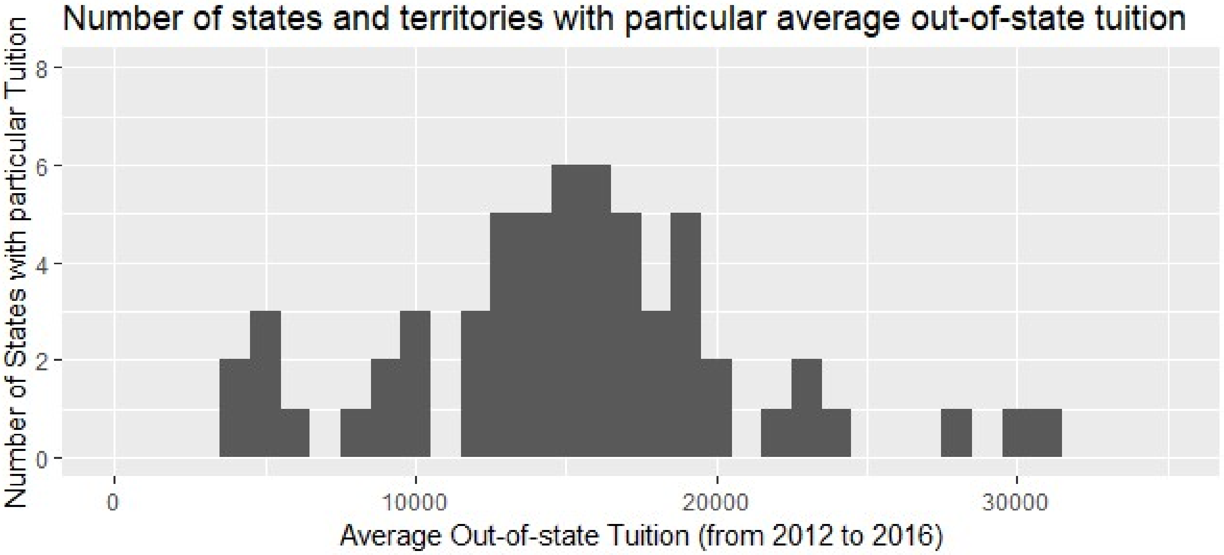
4. ***How does tuition vary across different states? Is there a relationship between the number of universities in a state and tuition? Do these characteristics differ for in-state tuition and out-of-state tuition?***

The average tuition varies a lot, depending on the state. In-state tuition is always less than out-of-state tuition in all the states and territories, except for the Federated States of Micronesia, where the tuition is the same for in-state and out-of-state. Here is a table with the range of tuition costs, and observe how large it is:

Range of tuition costs (average across all colleges per state from 2012 to 2016)

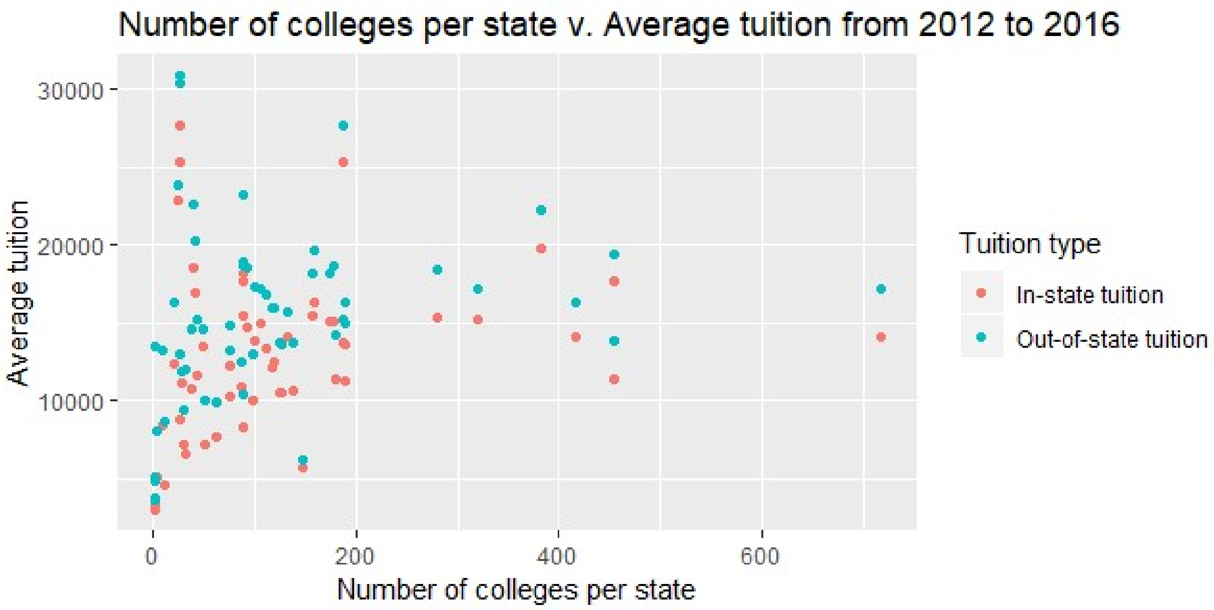
	Lowest in all states and territories	Lowest in a state	Highest in all states and territories
In-state tuition (per year)	$3,063.60 Northern Mariana Islands	$4,595.28 Wyoming	$27,627.92 Rhode Island
Out-of-state tuition (per year)	$3610.00 Palau	$8691.46 Wyoming	$30,844.09 Vermont





As seen in the two histograms above, the distribution of in-state tuition is lower than the distribution of out-of-state tuition.

Here is a plot showing the relationship between the number of colleges per state and the average tuition. (Note: number of colleges is based on counts from 2016.)



Again, in-state tuition is always lesser (or equal) to out-of-state tuition. Regarding the relationship tuition has with number of colleges, it seems that if a state has a small number of colleges, there is no way of predicting the average tuition costs: the costs are all over the board. But when the number of colleges in the state increases, the average cost heads towards the middle of the range. This relationship is seen in both in-state and out-of-state tuition. This may be because with more colleges in a state, there is a more diverse selection, from high to low end colleges, leaving the average tuition somewhere in the middle.

5. ***Which colleges have the most diverse demographics? Make sure to explain how you measured "diversity" for this problem, in addition to discussing your conclusions.***

I will consider "diversity" meaning there are three ethnic groups that make up 30% or more of the student population. I chose this measurement because I interpreted "diversity" to mean a high number of ethnic groups, each making up a high percentage of the student population. There are too many schools that have three ethnic groups that make up 20% or more of the student population. There are no schools that have four ethnic groups that make up 20% of the population.

I only used data from 2016. Here are the most diverse schools:

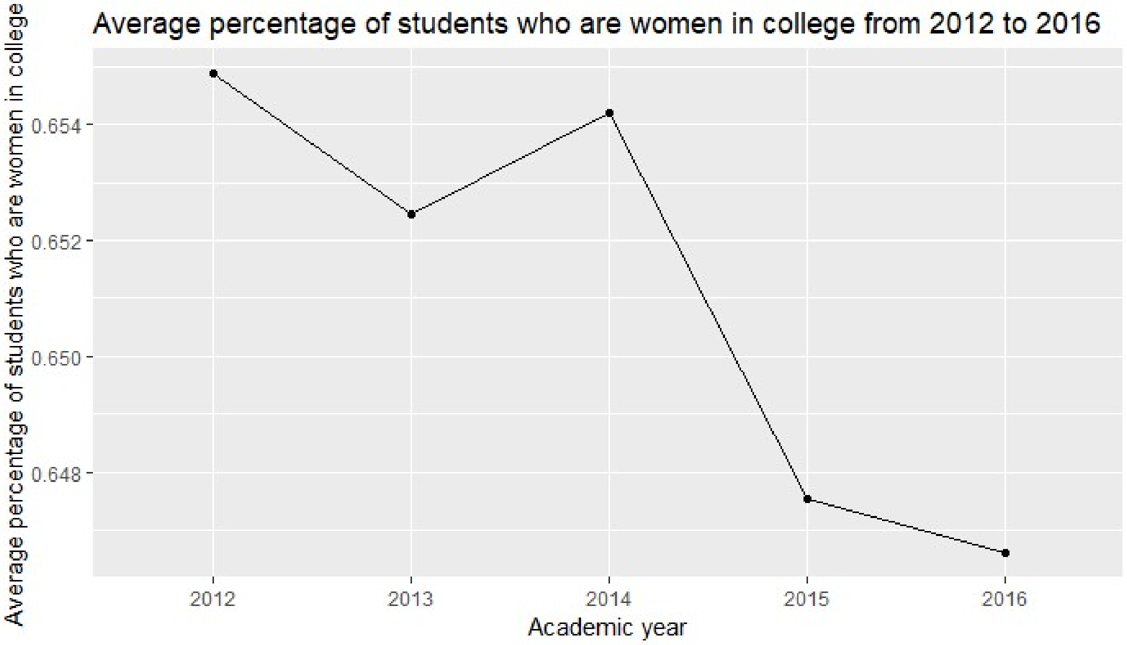
Old Town Barber College-Wichita (33.33% white, 35.42% black, 31.25% Hispanic)
Arlington Career Institute (30.54% white, 32.22% black, 36.82% Hispanic, 0.42% Asian)
Remington College-Fort Worth Campus (31.39% white, 31.58% black, 32.14% Hispanic, 1.32% Asian, 0.56% American Indian/Alaska Native, 2.82% two or more, 0.19% unknown)
Institute of Professional Careers (33.33% black, 33.33% Hispanic, 33.33% two or more)
Of course, these are only the most diverse schools under my definition of diversity. There are many schools that are diverse, but not under my definition.

The schools that I chose as the most diverse schools are not the typical four-year universities. These four schools are highly specialized, with only a few programs. There are tiny number of students, from 6 students to 532 students, which is a fraction of the average 2,424 students at a college. As expected, these four schools are private schools.

So why are these schools the most diverse? It's likely because of how small the schools are and how focused they are. The giant applicant pools of large public universities do not bog down these four schools, and they are given a certain amount of selectively. It could also simply be change. There are thousands of small, specialized colleges who are not as diverse as these four, and that may be due purely to probability.

6. ***Answer 2 of the questions you invented for Assignment 1, Problem 12. Use statistical summaries (including graphics) as evidence to support your conclusions. Also make sure to clearly state each question before your answer!***

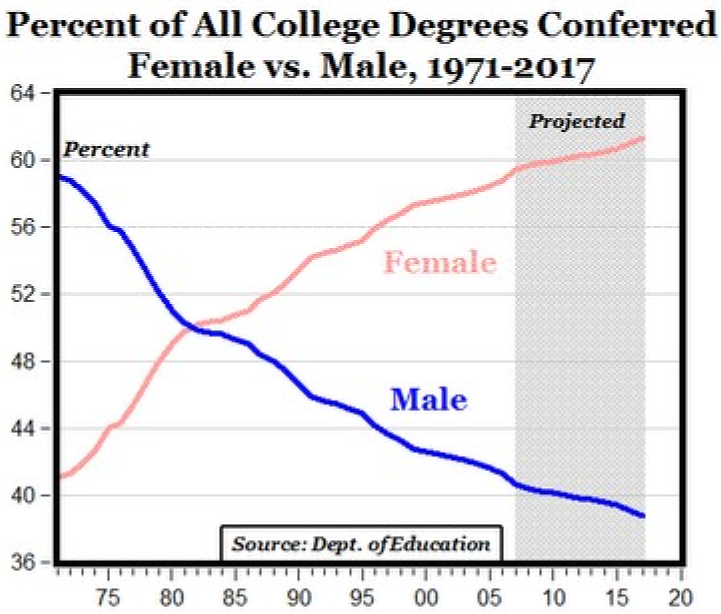
1) <u>How does the enrollment of women in college change over time?</u>



Note: the average percentage of students who are women was calculated by taking the average percentage across all the colleges in the US.

Interestingly, over the last few years, enrollment of women in college has dropped very slightly. While there is still a majority of women in college compared to men, their majority has gotten slightly slimmer. However, this data only encompasses five years, and is not representative of the larger picture going back several decades.

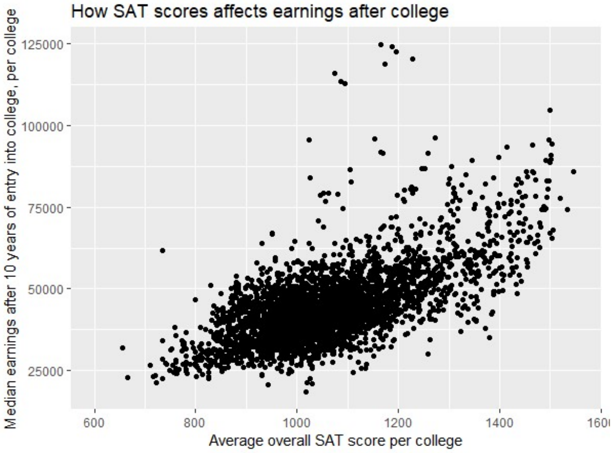
The following is a graph of similar data (percentage of college degrees conferred to women, rather than percentage of women enrolled in college), but it gives a broader perspective on the data, going back to 1971.

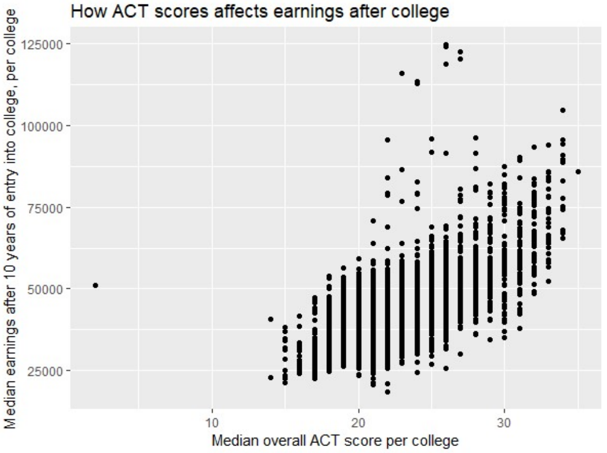

Percent of All College Degrees Conferred
Female vs. Male, 1971-2017

Source: Dept. of Education

Clearly, the percentage of women in college has increased dramatically since 1971, so this small dip of a few percentages observed from 2012 to 2016 does not indicate that women enrollment is dropping.

2) Do SAT and ACT scores affect future earnings?


How SAT scores affects earnings after college


How ACT scores affects earnings after college

As visible in the scatter plots, SAT and ACT score has a pretty clear positive relationship with earnings after college. On average, students with high test scores earned more in the future. There is, however, a lot of variance, so there is no guarantee. There are many other factors that affect earnings, such as major and school. There are a few colleges that have middling scores but much higher earnings than higher scores. This might be because of these colleges do a better job of preparing their students for a career. Overall, lower scores leads to lower earnings. A higher score leads to higher earnings, at least, higher than a lower score.

7. ***Reflect on the questions you answered in Problem 6. Did they lead to interesting conclusions? Why or why not? Did they raise new questions? Is it the question that makes a result interesting, the data, or both? Explain.***

The first question's conclusion was very interesting, because it was not what I expected. I kind of knew the answer to it already. I knew that over the years female enrollment in college has increased over the last few decades. What I didn't expect was a slight drop in percentage of women; I would have expected a steady increase. If data from previous years were available, we could check the trend that goes back for more than five years, and I would be able to use the exact same method (of taking the mean of the demographics of all the colleges for each year) to get an accurate reading on the trend over many years.

I was thinking that there was another way to approach the problem: I could multiply the percentage of women by the size of the college, so I can get the number of women. From there, I could find the total number of women in college in the entire country, and divide that by the sum of all the colleges' sizes to get the percentage. I wonder if it would have gotten the same result as taking an average.

The results in the first question raised a few new questions, such as, will the trend of the slowly dropping percentage of women in college continue? Or is it because we don't have enough data?

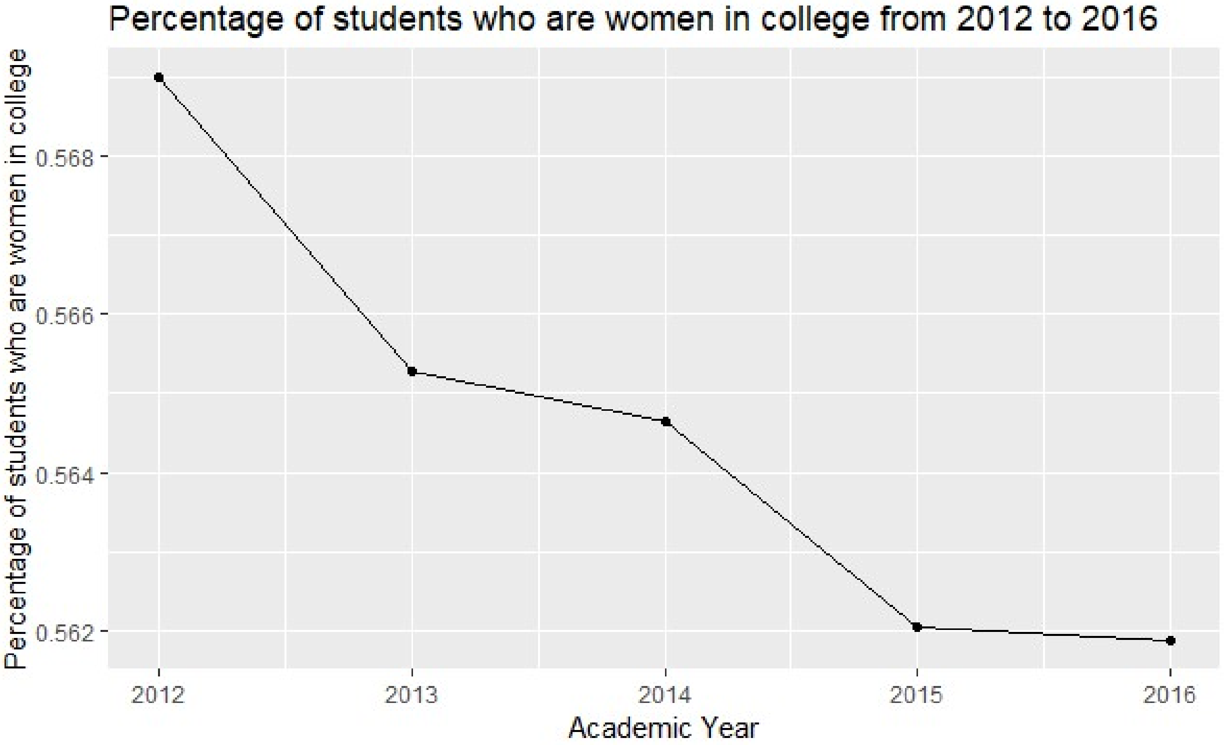
The second question's conclusion was also quite interesting. I wasn't too sure what to expect. It seems obvious that higher test score would mean more success, a thus, higher earnings, but it's also not a guarantee that test scores from high school would mean a higher salary ten years down the road. The positive relationship between the two variables was clearer than I expected, and it conveys how much test scores, even from high school, matters a lot. It's hard to draw a strict conclusion for this simple graph, but with studies questioning the effectiveness of SAT and ACT scores, (see https://www.pbs.org/newshour/education/nail-biting-standardized-testing-may-miss-mark-college-students) this information may be worth looking into.

The conclusion from the second question raises several new questions. How do individuals' test scores (not the colleges' average test scores) affect individuals' earnings after ten years? This may be a better indicator than averages based on college. How do test scores play a role in admissions? The way a college uses the test score may be different for each college, and it may affect the selection of students, thereby affecting future earnings. How much of students' future success is due to the college, and due to the individual student? Some colleges with average test scores had students' with higher earnings than everyone else. Maybe the college plays a large part in preparing a student for college, and maybe some colleges are not fulfilling that role. Or maybe, despite their test scores, students at that college were more active in preparing for career, and the college had nothing to do with it. Answering these questions would help us know how test scores impacts future success.

The conclusions are interesting because of a combination of a question and the data. We need an interesting question to get somewhere, and we need the data to give the answer. The answer is not always what we expect, but we need to always start by asking the question. Both the quality of the question and the data matter: interesting questions leads to interesting answers.

8. ***List and answer 2 follow-up questions raised by any of the work you did for this assignment. Along with each question and answer, make sure to explain what raised the question for you.***

1) Does enrollment of women look the same when looking at average percentage of women in college per college, and total number of women enrolled in college in the entire country divided by the total number of students in college?



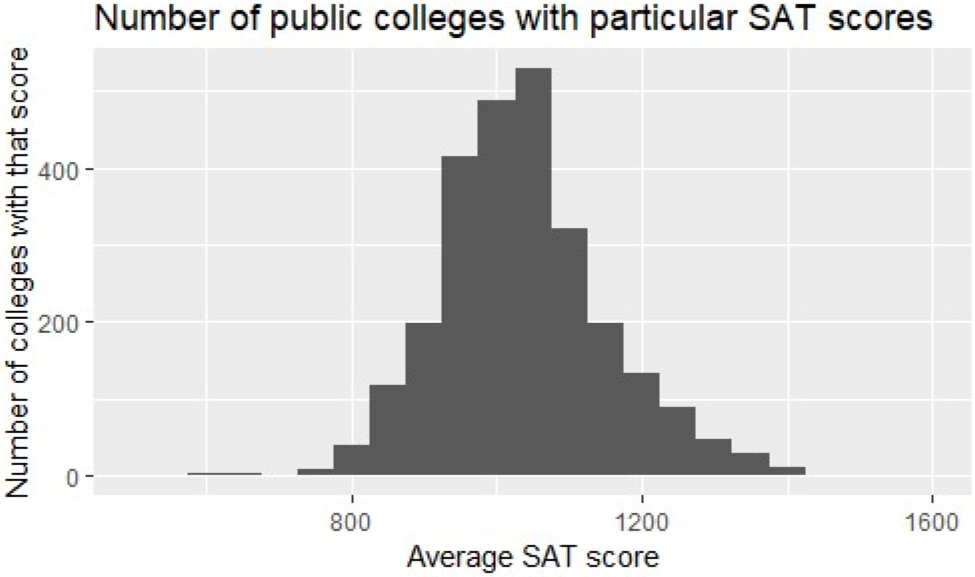
Note: these percentages were calculated by taking the total number of women in college (percentage of students who are women in a school times the number of total students in the school) and dividing that by the total number of students in the whole country.

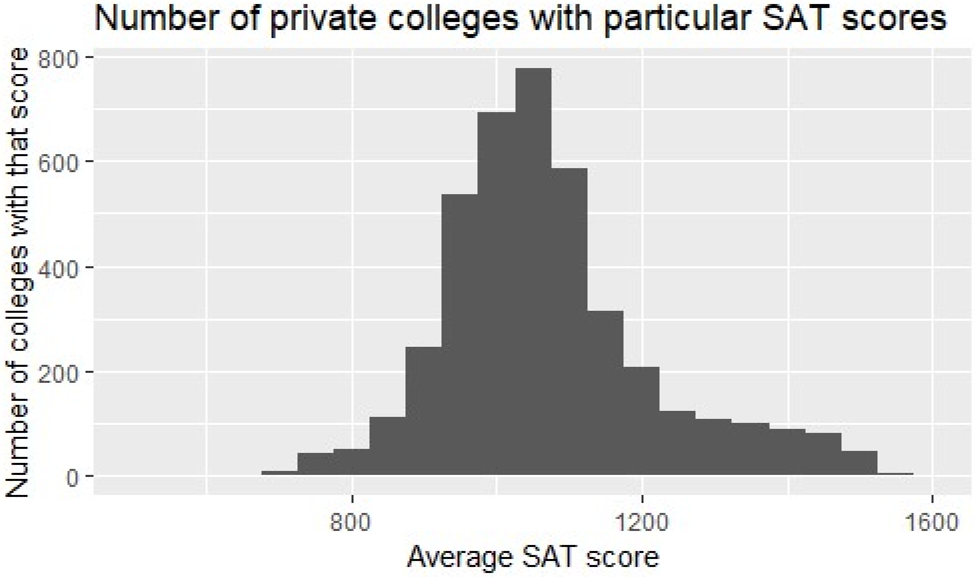
Comparing this graph with the graph in Question 6 Part 1, the downward trend is the same, but the percentages are different. This graph starts at 56.90% of students in college are women to 56.19%. The previous graph ranged from 65.49% down to 64.66%.

It is interesting how these two methods of taking the same measure led to different results. There is a huge 10% difference between the two, and I was attempting to trying to measure the same thing each time. This could be because of a variety of reasons: perhaps the high number of all women's schools inflated the average, or including the large amount of missing data may lead to different results.

I initially thought of this question when thinking about how I would answer the original question of the enrollment of women in college. I wondered if different methods would be easier to calculate or would lead to different results. I guess I was right.

2) What are the average SAT scores for public compared to private colleges?





It looks like both public and private colleges have a peak of the distribution at around the same SAT score. The private colleges distribution has a small tail, so there are a few private schools that have a high average SAT score. This is likely because some private schools are highly selective, and will choose a small number of incredibly high-achieving students in their admissions. Public schools are large and cannot afford to be hyper-selective, so the averages are a bit lower, and the distribution does not go as high (although it does go a bit lower).

I thought about this question when I was writing about how different schools use SAT and ACT scores in their admissions in the reflection in Question 7. Different colleges may use the test scores differently, depending on the objective of the college; for example, a highly focused culinary school may not look at test scores, while a highly desired elite college will.

R Code Appendix:

```
# STA 141A Assignment 2
# Nathan Chan

scorecard = readRDS("college_scorecard.rds")

# 1. Exploring missing values

check_na = function(data) {
  # Prints out the column names that have an NA in them
  # Prints out number of columns
  count = 0
  for (i in 1:ncol(data)) {
    if (any(is.na(data[, i])) == TRUE) {
      print(colnames(data)[i])
      count = count + 1
    }
  }
  cat("Number of columns with NA:", count)
}

explore_na = function(data, number) {
  # See how many columns have more NA's than the given number
  # Prints out number of columns
  count = 0
  for (i in 1:ncol(data)) {
    if (sum(is.na(data[, i])) >= number) {
      print(colnames(data)[i])
      count = count + 1
    }
  }
  cat("Number of columns with specific number of NA:", count)
}

number_of_na = function(data) {
  # Returns ordered dataframe of number of NA's per column
  col1 = c()
  col2 = c()
  for (i in 1:ncol(data)) {
    col1[i] = colnames(data)[i]
    col2[i] = sum(is.na(data[, i]))
  }
  NA_number = data.frame("Feature" = col1, "Number_of_NA" = col2)
  sorted_NA_number = NA_number[order(NA_number$Number_of_NA),]
  return(sorted_NA_number)
}

check_na(scorecard) # which columns have NA in them
explore_na(scorecard, 38068) # how many columns are empty
sorted_NA_number = number_of_na(scorecard)

# As in Nick's suggestion, look at NA's over the years.
sum(is.na(subset(scorecard, scorecard$academic_year == 2012)))
sum(is.na(subset(scorecard, scorecard$academic_year == 2013)))
sum(is.na(subset(scorecard, scorecard$academic_year == 2014)))
sum(is.na(subset(scorecard, scorecard$academic_year == 2015)))
sum(is.na(subset(scorecard, scorecard$academic_year == 2016)))

sum(is.na(subset(scorecard, scorecard$ownership == "Public")))
sum(is.na(subset(scorecard, scorecard$ownership == "Private nonprofit")))
sum(is.na(subset(scorecard, scorecard$ownership == "Private for-profit")))

# 2. Exploring student populations

create_list = function(data, variable1, variable2) {
  # Returns ordered dataframe of two given variables
  col1 = c()
  col2 = c()
  for (i in 1:nrow(data)) {
    col1[i] = data[i, grep("name", colnames(data))]
    col2[i] = data[i, grep("size", colnames(data))]
  }
  df = data.frame(variable1 = col1, variable2 = col2)
  ordered_df = df[order(df$variable2),]
  return(ordered_df)
}

scorecard2016 = subset(scorecard, scorecard$academic_year == 2016)
name_size = create_list(scorecard, scorecard$name, scorecard$size) # create list of data with name and size
name_size_2016 = create_list(scorecard2016, scorecard$name, scorecard$size) # create list with only 2016 data

library(ggplot2)
ggplot(scorecard2016, aes(x = size)) + geom_density() + xlim(0, 70000) # see the distribution of sizes
ggplot(scorecard2016, aes(x = size)) + geom_bar() + xlim(0, 70000)

install.packages("dplyr")
library(dplyr)
```

```
size2016 = select(scorecard2016, name, size, grad_students) # does same thing as create_list function

ggplot(scorecard2016, aes(x = size, y = grad_students)) + geom_point() +
  labs(title = "Number of undergraduate v. Number of graduate students in college in 2016",
       x = "Number of undergraduate students", y = "Number of graduate students")
length(name_size_2016[,1])
NA_size = sum(is.na(scorecard2016$size))

more_grad = scorecard2016$name[scorecard2016$grad_students > scorecard2016$size] # which schools have more grads

# 3. Looking at program percentages

summary(scorecard2016) # look at the distribution for program percentages

# 4. Tuition and states

names(scorecard2016)

instate = aggregate(tuition.in_state ~ state, scorecard, mean, na.rm = TRUE)
outstate = aggregate(tuition.out_of_state ~ state, scorecard, mean, na.rm = TRUE)

tuition2016 = select(scorecard2016, name, tuition.in_state, tuition.out_of_state)

sorted_instate = instate[order(instate$tuition.in_state),]
sorted_outstate = outstate[order(outstate$tuition.out_of_state),]

in_vs_out = function(data1, data2) {
  # Prints out whichever states or territories have larger or equal in-state tuition
  for (i in 1:nrow(data1)) {
    if (data1[i, 2] >= data2[i, 2]) {
      print(data1[i, 1])
    } else {
      print("Nope.")
    }
  }
}

in_vs_out(instate, outstate)

ggplot(instate, aes(x = tuition.in_state)) + geom_histogram(binwidth = 1000) +
  labs(x = "Average In-state Tuition (from 2012 to 2016)", y = "Number of States with particular Tuition",
       title = "Number of US states and territories with particular average in-state tuition") +
  xlim(0, 35000) + ylim(0, 8)
```

```
ggplot(outstate, aes(x = tuition.out_of_state)) + geom_histogram(binwidth = 1000) +
  labs(x = "Average Out-of-state Tuition (from 2012 to 2016)", y = "Number of States with
particular Tuition",
       title = "Number of states and territories with particular average out-of-state tuition") +
  xlim(0, 35000) + ylim(0, 8)

# Looking at relationship between tuition and number of colleges
library("plyr")

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

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

# Individual scatter plots
ggplot(state_freq, aes(x = state_freq$freq, y = instate$tuition.in_state)) + geom_point() +
  labs(x = "Number of colleges per state", y = "Average in-state tuition",
       title = "Number of colleges per state v. Average in-state tuition") + ylim(0, 32000)

ggplot(state_freq, aes(x = state_freq$freq, y = outstate$tuition.out_of_state)) + geom_point() +
  labs(x = "Number of colleges per state", y = "Averate out-of-state tuition",
       title = "Number of colleges per state v. Averate out-of-state tutition") + ylim(0, 32000)

# Combined scatter plot
ggplot(state_freq) + geom_point(aes(x = state_freq$freq, y = instate$tuition.in_state,
                                    color = "In-state tuition")) +
  geom_point(aes(x = state_freq$freq, y = outstate$tuition.out_of_state, color = "Out-of-state
tuition")) +
  labs(x = "Number of colleges per state", y = "Average tuition",
       title = "Number of colleges per state v. Average tuition from 2012 to 2016") +
  guides(color = guide_legend(title = "Tuition type"))

# 5. Diversity

find_diversity = function(data, ethnic_group_number, percentage) {
  # Prints name of colleges that have three ethnic groups, with each group being 30% of
population
  for (i in 1:nrow(data)) {
    count = 0
    for (j in 87:95) {
      if (is.na(data[i, j]) == TRUE) {
        next
      } else if (data[i, j] > percentage) {
        count = count + 1
        if (count >= ethnic_group_number) {
          print(data[i, 4])
        }
      }
    }
  }
}

find_diversity(scorecard2016, 3, .3) # three ethnic groups make up 30% of student population
find_diversity(scorecard2016, 3, .2) # three ethnic groups make up 20% of student population
find_diversity(scorecard2016, 4, .2) # four ethnic groups make up 20% of student population
find_diversity(scorecard2016, 3, .3)

diverse_schools = subset(scorecard2016, scorecard2016$name == "Old Town Barber College-
Wichita" |
                         scorecard2016$name == "Arlington Career Institute" |
                         scorecard2016$name == "Remington College-Fort Worth Campus" |
                         scorecard2016$name == "Institute of Professional Careers")

diverse_schools$size
mean(scorecard2016$size, na.rm = TRUE)
diverse_schools$ownership

# 6. My own questions

# 1) Enrollment of women over time
women_year = aggregate(demographics.women ~ academic_year, scorecard, mean, na.rm =
TRUE)

ggplot(women_year, aes(x = academic_year, y = demographics.women, group = 1)) +
geom_point() + geom_line() +
  labs(x = "Academic year", y = "Average percentage of students who are women in college",
       title = "Average percentage of students who are women in college from 2012 to 2016")

# 2) SAT and ACT scores and future earnings
ggplot(subset(scorecard, !is.na(sat_scores.average.overall)),
       aes(y = earn_10_yrs_after_entry.median, x = sat_scores.average.overall, na.rm == TRUE))
+ geom_point() +
  labs(y = "Median earnings after 10 years of entry into college, per college",
       x = "Average overall SAT score per college",
       title = "How SAT scores affects earnings after college")

ggplot(subset(scorecard, !is.na(act_scores.midpoint.cumulative)),
       aes(y = earn_10_yrs_after_entry.median, x = act_scores.midpoint.cumulative, na.rm ==
TRUE)) + geom_point() +
  labs(y = "Median earnings after 10 years of entry into college, per college",
      x = "Median overall ACT score per college",
      title = "How ACT scores affects earnings after college")

# 8. Follow-up questions

# 1) Women enrollment, again

percentage_of_women_per_year = function(data) {
  # Prints out percentage of women enrolled in college per year
  col1 = c(2012, 2013, 2014, 2015, 2016)
  col2 = c()
  for (i in 1:5) {
    year = i + 2011
    remove_na = subset(data, !is.na(data$demographics.women))
    yearly_data = subset(remove_na, remove_na$academic_year == year)
    number_of_women = 0
    total_people = 0
    for (j in 1:nrow(yearly_data)) {
      number_of_women = number_of_women + (yearly_data[j, 86] * yearly_data[j, 141])
      total_people = total_people + yearly_data[j, 86]
    }
    col2[i] = number_of_women / total_people
  }
    percentage = data.frame("Year" = col1, "Percentage of women in college" = col2)
    return(percentage)
}

percentage = percentage_of_women_per_year(scorecard)

ggplot(percentage, aes(x = Year, y = Percentage.of.women.in.college, group = 1)) +
geom_point() + geom_line() +
  labs(x = "Academic Year", y = "Percentage of students who are women in college",
      title = "Percentage of students who are women in college from 2012 to 2016")

# 2) SAT and ACT scores for private vs. public schools

private = subset(scorecard, ownership == "Private nonprofit" | ownership == "Private for-profit")
public = subset(scorecard, ownership == "Public")

ggplot(private, aes(x = sat_scores.average.overall)) + geom_histogram(binwidth = 50) +
  labs(x = "Average SAT score", y = "Number of colleges with that score",
      title = "Number of private colleges with particular SAT scores") + xlim(500, 1600)

ggplot(public, aes(x = sat_scores.average.overall)) + geom_histogram(binwidth = 50) +
  labs(x = "Average SAT score", y = "Number of colleges with that score",
      title = "Number of public colleges with particular SAT scores") + xlim(500, 1600)
```