Assignment 1

Jie Gu

SID: 913953707

1.The purpose of this data set is to increase transparency, putting the power in the hands of students and families to compare how well individual postsecondary institutions are preparing their students to be successful. (according to data_document)

The U.S. Department of Education created this data set.

The sources for the data are federal reporting from institutions, data on federal financial aid, and tax information. Also, many data elements are from data reported to the IPEDS.

2.There are 38068 rows in this data set. These rows represent each college recorded in each year.

3.There are 142 columns in this data set. These columns represent different information, like city, state, etc., for colleges.

4.The data set span of 5 years: from 2012 to 2016.

In 2012, the data set records 7793 colleges;

In 2013, the data set records 7804 colleges;

In 2014, the data set records 7703 colleges;

In 2015, the data set records 7593 colleges;

In 2016, the data set records 7175 colleges;

Year	Number of colleges
2012	7793
2013	7804
2014	7703
2015	7593
2016	7175

5.Situation for 2012

State with the most colleges	Number of colleges	State with the fewest colleges	Number of colleges
CA	797	AS	1
TX	476	FM	1
NY	464	MH	1
FL	435	MP	1
PA	416	PW	1

Situation for 2013

State with the most colleges	Number of colleges	State with the fewest colleges	Number of colleges
CA	804	AS	1
TX	485	FM	1
NY	464	MH	1
FL	437	MP	1
PA	414	PW	1

Situation for 2014

State with the most colleges	Number of colleges	State with the fewest colleges	Number of colleges
CA	795	AS	1
TX	485	FM	1
NY	467	MH	1
FL	446	MP	1
PA	405	PW	1

Situation for 2015

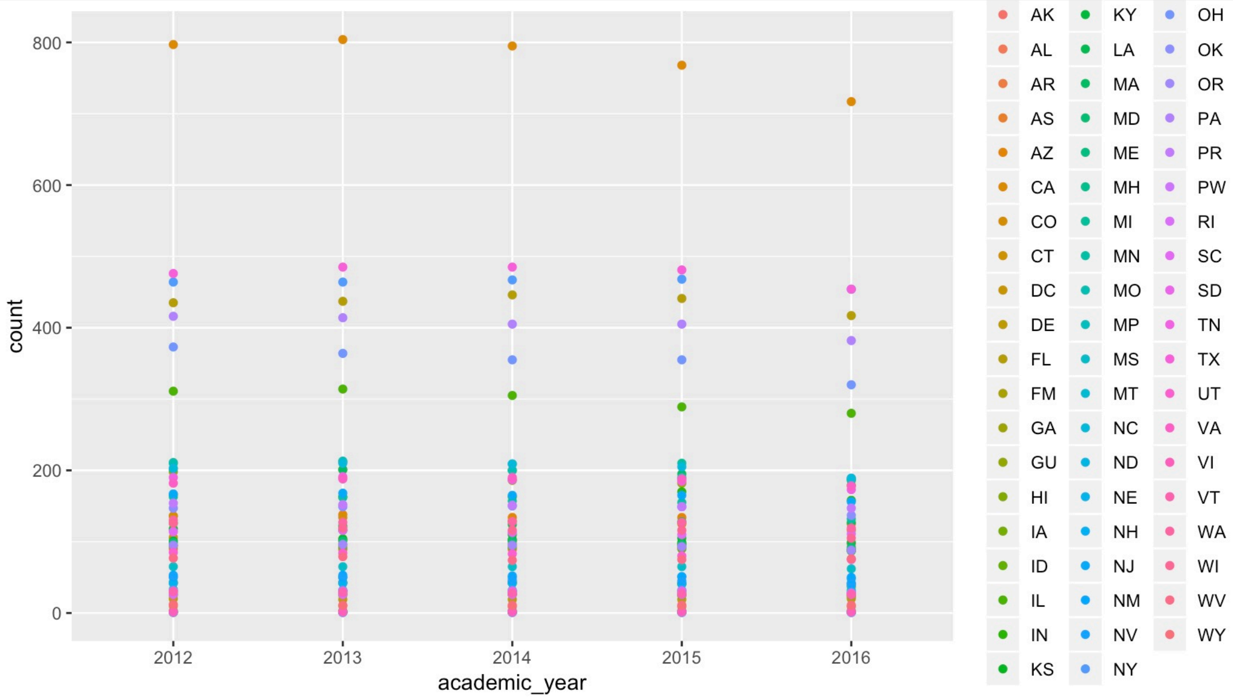
State with the most colleges	Number of colleges	State with the fewest colleges	Number of colleges
CA	768	AS	1
TX	481	FM	1
NY	468	MH	1
FL	441	MP	1
PA	405	PW	1

Situation for 2016

State with the most colleges	Number of colleges	State with the fewest colleges	Number of colleges
CA	717	AS	1
TX	454	FM	1
NY	454	MH	1
FL	417	MP	1
PA	382	PW	1

From five tables, we can see that CA, TX, NY, FL, PA are always the five states with the most colleges, and AS, FM, MH, MP, PW are always the five states with the fewest colleges.

In addition, it is easier for us to observe the trend of numbers of colleges in the 5 states with most colleges and the 5 states with the least colleges in the following dot plot.



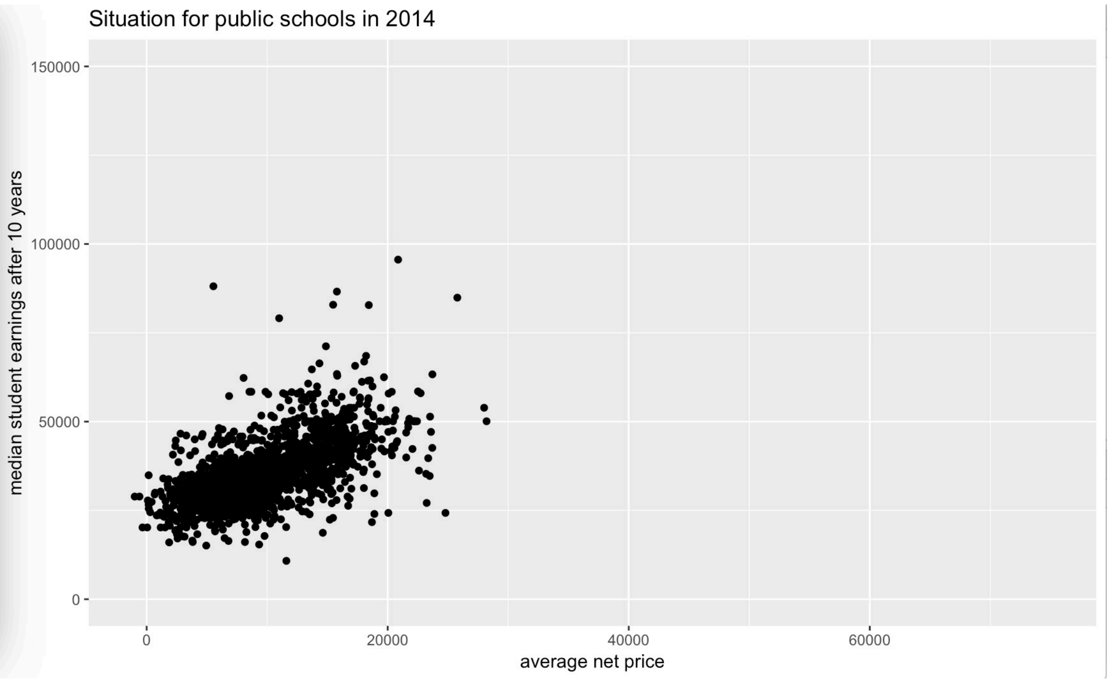
From this plot, we can see, from 2012 to 2016, the numbers of colleges in CA, TX, NY, FL, PA all increase at first and then decline; the numbers of colleges in AS, FM, MH, MP, PW all keep the same from 2012 to 2016.

My hypothesis about why some states have a lot of colleges is that these states have a higher population than other states, and the economy develops better in these states. To meet the needs of a high population of people, these states have to open more colleges. Also, the great economy situation in these states enables these states have enough money to open colleges.

From this link: https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_population_growth_rate, we can see the five states with the highest population are CA, TX, FL, NY, PA, which are the five states with most colleges.

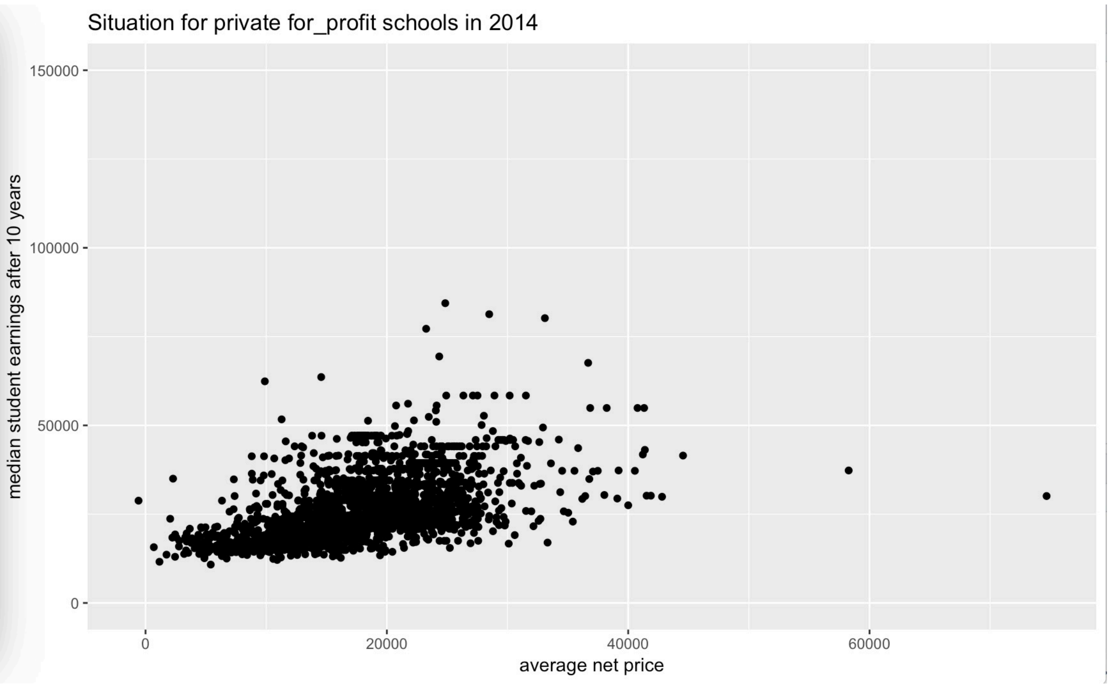
From this link:
https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_GDP,
we can see CA, TX, FL, NY, PA have a high GDP, and they all belong to the 6 states with the highest GDP.

6.



From this scatter plot, we see that for most points with higher ‘average net price’, they also have a higher ‘the median student earnings after 10 years’. These points look like a line with a positive slope. In other words, for most public school in 2014, with the increase in average net price, the median student earnings after 10 years also increases. Some schools have very low average net price, but the median student earnings after 10 years for these schools is high. For most college students in public schools in 2014, this pattern means that their earnings after 10 years may be better with a higher cost of studying in their schools. However, for a minority of schools, students can still have a great earning even though the cost of studying in these schools is low.

7.



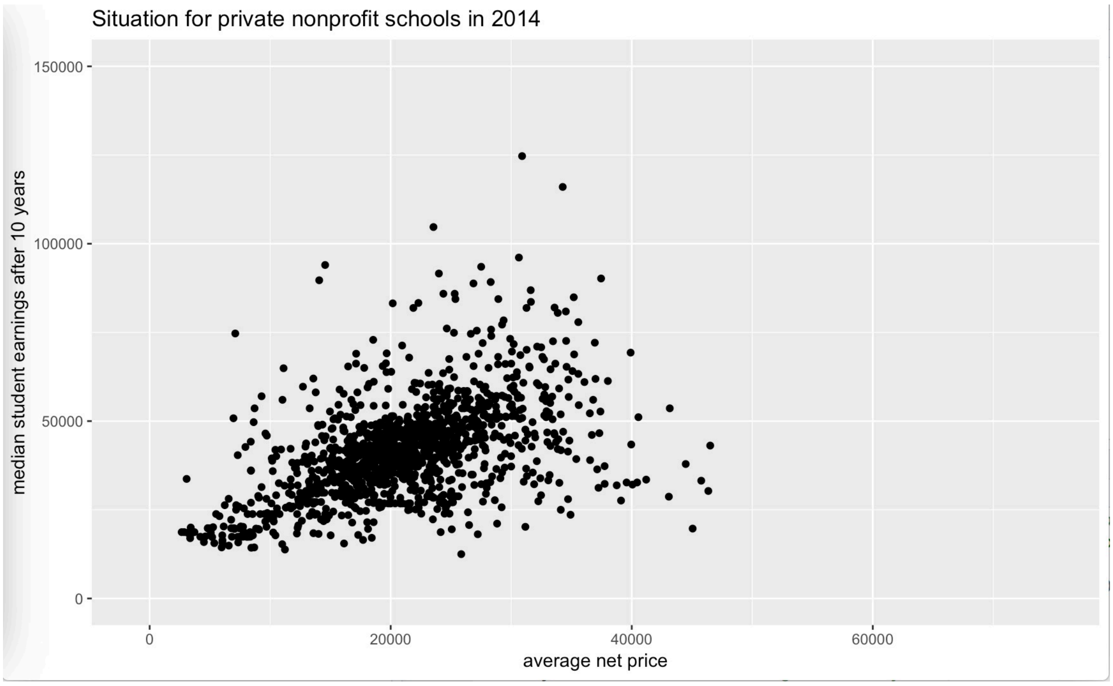
Compared with the scatter plot of public school, most points with a higher 'average net price' also have a higher 'the median student earnings after 10 years'. These points also look like a line with positive slope. That is to say, this plot for private for-profit school also shows that, for most schools, with the increase in average net price, the median student earnings after 10 years also increases.

However, the median student earnings after 10 years increases much slower compared with that of public school. Also, for public schools, the 'average net price' for most points is smaller than 20000, but in this plot, the 'average net price' for many points is larger than 20000. This means many private for-profit school cost more for students compared with public schools. Therefore, the cost performance for private for-profit school is lower that for public school. Usually, it is a better choice for students to choose public school instead of private for-profit school.

In addition, for some private for-profit school, the average net price is quietly high, but the median student earnings after 10 years is relatively low. These schools are Hallmark Institute of Photography, The International Culinary Center, L3 Commercial Training Solutions Airline Academy, and Aviator College of

Aeronautical Science and Technology. This means the cost performance of these schools is bad.

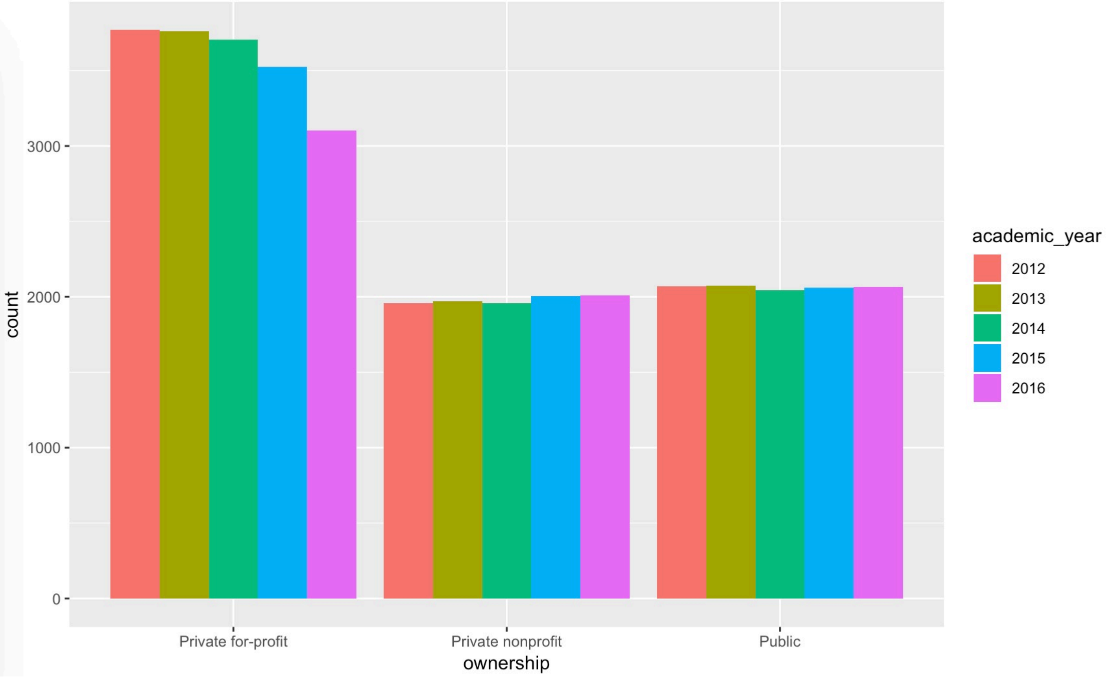
8.To evaluate private non-profit schools, we can also use a scatterplot.



From this plot, most points with a higher 'average net price' also have a higher 'the median student earnings after 10 years', which is the same as the public school and private for-profit school. So these points also look like a line with a positive slope. The slope of this 'line' is similar to the slope of 'line' for public school, which is larger than the slope of 'line' for private for-profit school. This means that as 'average net price' increases, the velocity of increase of 'median student earnings after 10 years' is similar to that for public schools. Additionally, in this plot, the 'average net price' of many points is higher than 20000, and the 'median student earnings after 10 years' of many points is higher than 50000.

In conclusion, the cost performance of private non-profit school is similar to public school, which is higher than private for-profit school. Also, the cost of most private non-profit school is similar to private for-profit school, which is higher than public school. More importantly, many students from private non-profit school can earn more than students from other two kinds of school after 10 years. Therefore, for students who have a relatively high family income, it is a good choice to choose private non-profit school.

9.



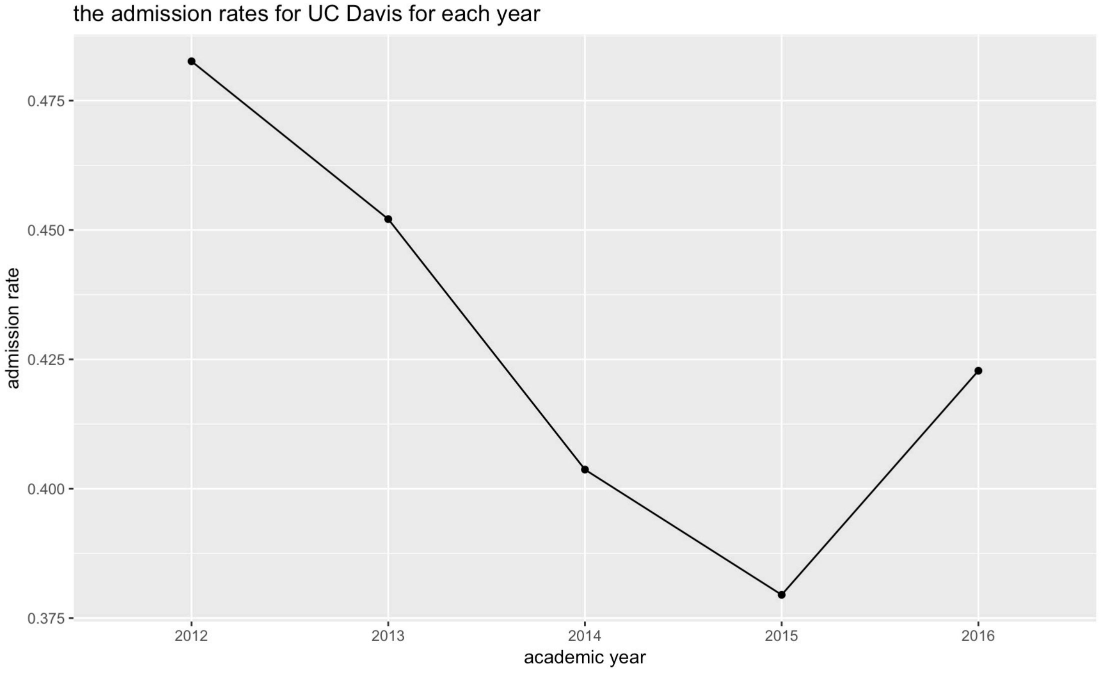
From the bar plot, we can clearly observe that the amount of private for-profit schools for each year is larger than the amount of private non-profit schools and the amount public schools for each year.

However, the amount of private for-profit school decreases from 2012 to 2016, while the amount of private non-profit schools and the amount public schools almost keep the same from 2012 to 2016.

The reason why the amount of private for-profit school decreases from 2012 to 2016 is that are that more and more adults tend to work more instead of studying from 2012 to 2016. Also, those for-profit colleges find it hard to "attract students at an ever-rising price point with a relatively strong economy, combined with political and social pressure to restrain tuition growth". (***"This Is the Way the College 'Bubble' Ends",*** Derek Thompson**)**

Here is a link for the article ***"This Is the Way the College 'Bubble' Ends"***:
https://www.theatlantic.com/business/archive/2017/07/college-bubble-ends/534915/

10.



The admission rates change much from year to year. More precisely, from 2012 to 2015, the admission rate decreases more than 10%, and from 2015 to 2016, the admission rate increases roughly 5%.

11. R data type: double, character, list, logical, etc.

“double” are numbers:1, 1.2, 333, 333.33, etc.

“character”: ‘a’, ‘asd’, ‘123’, etc.

“list” has rows and columns

“NULL always means the missing data.

“logical” always means TRUE or FALSE.

Statistical data type:

Categorical	Numerical
nominal (no order): color, etc.	discrete--years, people, etc.
ordinal (have order)): size, etc.	continuous--years,height,weight,etc.

Other type:

Images

Spatial

Text

For the question: Does each R data type map to just one statistical data type?

No, it doesn’t. For example, R data type “character” can map to both categorical date type and numerical data type. More precisely, when we type color = ‘red’ in R studio, it is a “character” data type in R, and it is a categorical date type as for statistical data type. Also, when we type height = ‘180’ in R studio, the height is also a “character” data type in R, but it is a numerical data type as for statistical data type.

12.Question 1: What is the average median family income for public school, private non-profit school, and private for-profit school?

This question can benefit students and their family by letting them know what type of school is a popular choice for people having a similar family income with them. In other words, they can know which type of school is suitable for a family with a similar income with them.

I can use variables ‘demographic.median_family_income’ and ‘ownership’. First I subset the colleges by ‘ownership’. Then R studio can help me calculate the mean of ‘demographic.median_family_income’ for each subset.

Question 2: What colleges have a relatively high percentage of program for computer?

This question can benefit students who are interested in learning computer since a high percentage of program for computer means that the colleges devote much money and resource on computer major. Also, the atmosphere for learning computer is good in those colleges. Therefore, students who are interested in computer can make a selection between those schools.

I can use variable ‘program_percentage. computer’. I can subset the colleges by set the variable ‘program_percentage. computer’ larger than 0.1, and then I can just view the subset.

Question 3: For colleges in 2016, what is the relation between admission rate and the median student earnings after 10 years?

This question will benefit all the students when applying for colleges. More precisely, if the median student earnings after 10 years increases as the admission rate decrease, it is worthy of spending much time and energy on applying those colleges with a low admission rate. Also, students can avoid applying for some colleges with a low admission rate but with a low “median student earnings after 10 years”.

For this problem, I can use variables ‘academic_year’, ‘admission_rate.overall’ and ‘median student earnings after 10 years’. First I can subset school with ‘academic_year’ equals ‘2016’. Then I can use ggplot to plot a scatter plot with ‘admission_rate.overall’ as x-axis and ‘median student earnings after 10 years’ as y-axis. The scatter plot will show the trend.