Man Pan

Student ID: 914656278
Email: manpan@ucdavis.edu

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

This data is to help students and families compare college costs and outcomes as they weigh the tradeoffs of different colleges, accounting for their own needs and educational goals.
These data are provided through federal reporting from institutions, data on federal financial aid, and tax information. The data is created by US department of education.
Many data elements are drawn directly from, or derived from, data reported to the IPEDS.IPEDS is the primary source of data on post-secondary education institutions in the United States.

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

The data has 38068 rows. Rows represent observation of different colleges in multiple years.

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

The data has 142 columns. Columns represent basic information (features) of each college and some criteria for judging colleges' value.

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

The range of years : 2012 - 2016;
2012 : 7793 colleges
2013 : 7804 colleges
2014 : 7703 colleges
2015 : 7593 colleges
2016 : 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 5 states with the most colleges：
"PA"　：　2022　colleges
"FL"　：　2176　colleges
"NY"　：　2317　colleges
"TX"　：　2381　colleges
"CA"　：　3881　colleges

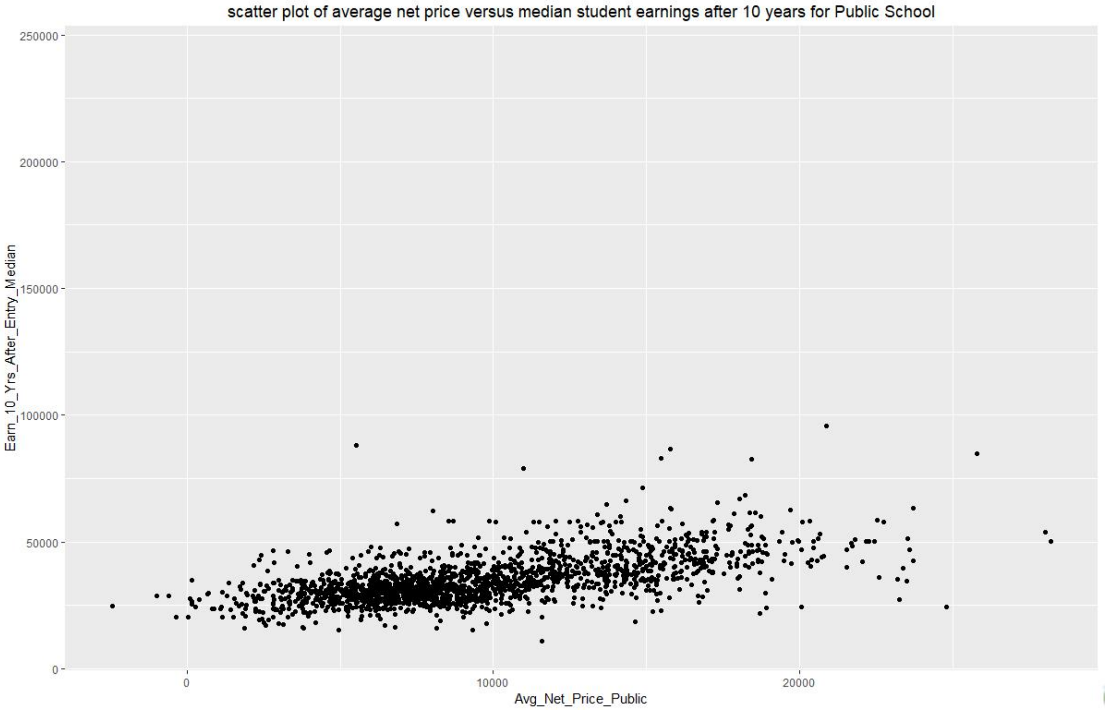
the states with the fewest colleges:
"AS" "FM" "MH" "MP" "PW". They all have 5 colleges.

Hypothesis: States that have a higher GDP and population will have more colleges.
Outside sources:

https://www.brookings.edu/research/what-colleges-do-for-local-economies-a-direct-measure-based-on-consumption/

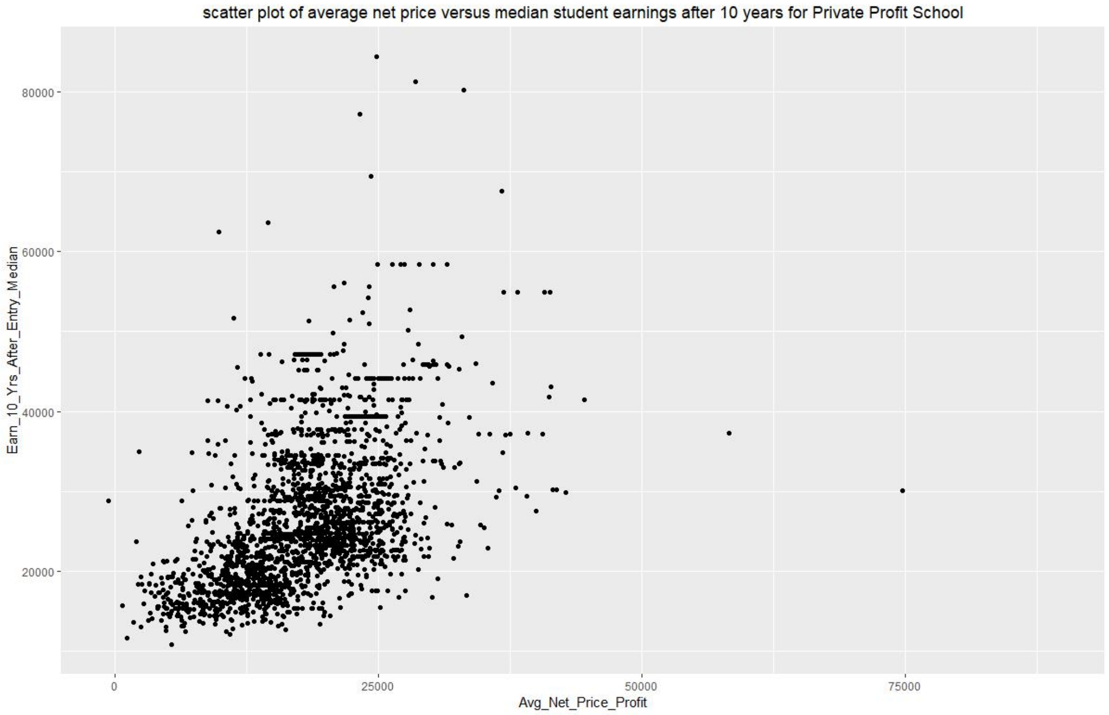
https://www.forbes.com/sites/petesaunders1/2017/11/29/state-capitals-and-college-towns-a-recipe-for-success/#de4e118781af

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.



In this plot, we can conclude that average net price and median student earnings after 10 years are slightly positive related. The higher average net price is, the higher median student earnings after 10 years is. The median student earnings after 10 years mainly concentrates between $2000-$10000.

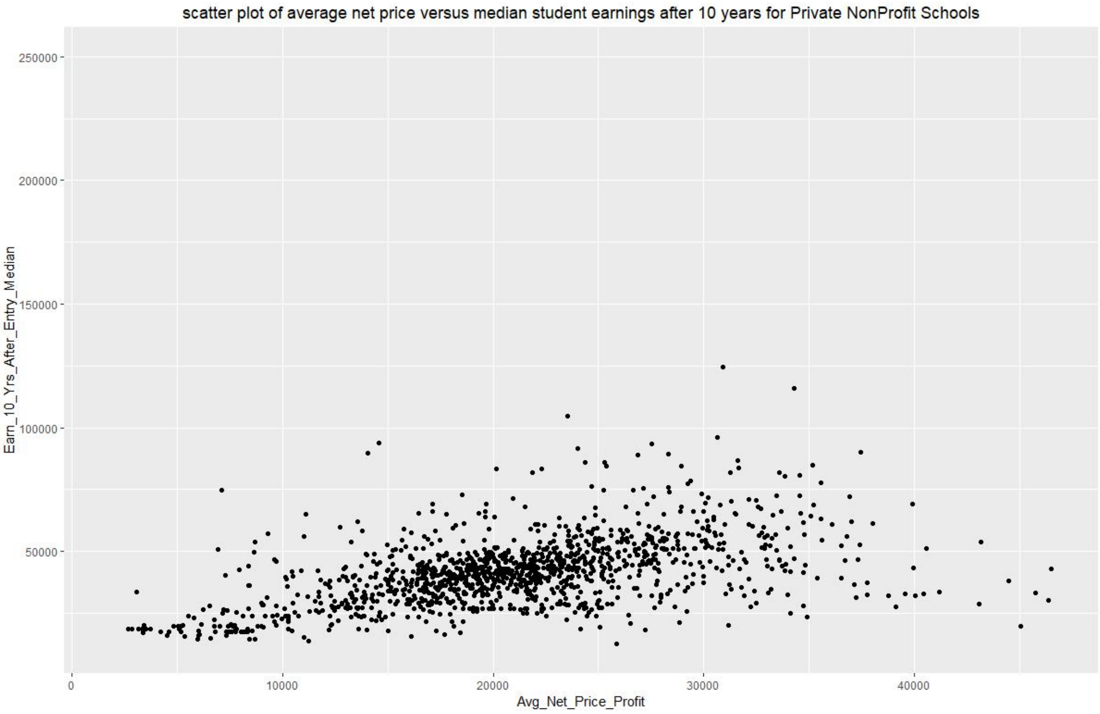
7. Create the plot from the previous question for private for-profit schools. How do the two plots compare? Whether you see similarities or differences, discuss what your results imply about public schools and private for-profit schools.

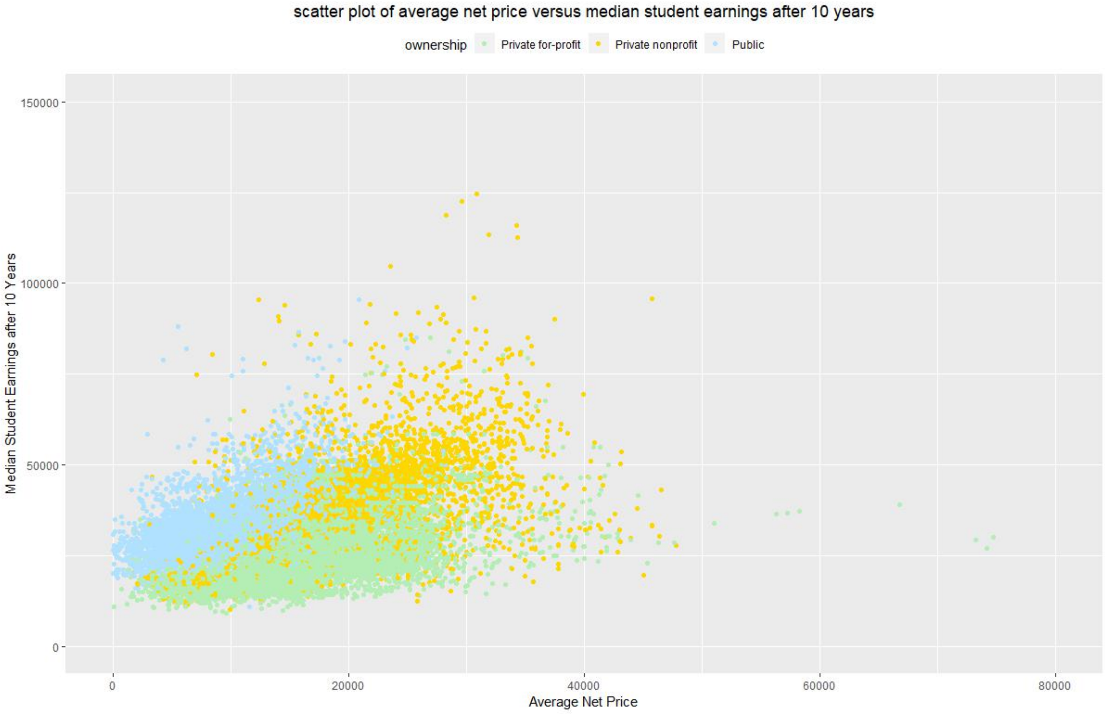


For the two plots, we can find they are all positive related. However, they have different correlation coefficient. For plot1, two variables are sightly related. If we increase average net price in public schools, the median student earnings after 10 years just increase a little. For plot2, two variables are highly related. If we increase average net price in profit schools, we would get a much higher median student earnings after 10 years.

For plot 1, the median student earnings after 10 years mainly concentrates between $10000-$75000. For plot 2, the median student earnings after 10 years mainly concentrates between $10000-$60000. Thus, for students, they are more likely to get more earnings in Public school.

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

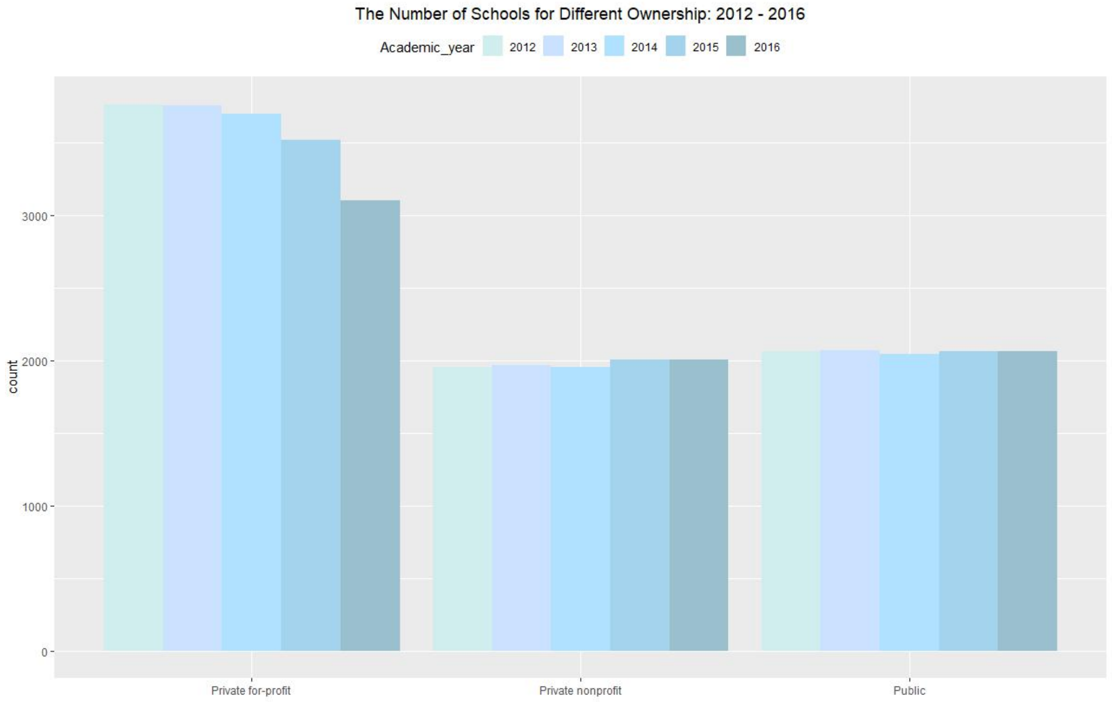




If we plot the three scatter plot into one scale. We can find that the three plot have similar trend - positive related. But if we increase the same average net price, the median student earnings after 10 years would increase a little for private and profit schools (compared with public and private nonprofit schools). The median student earnings after 10 years would increase much more for public schools.

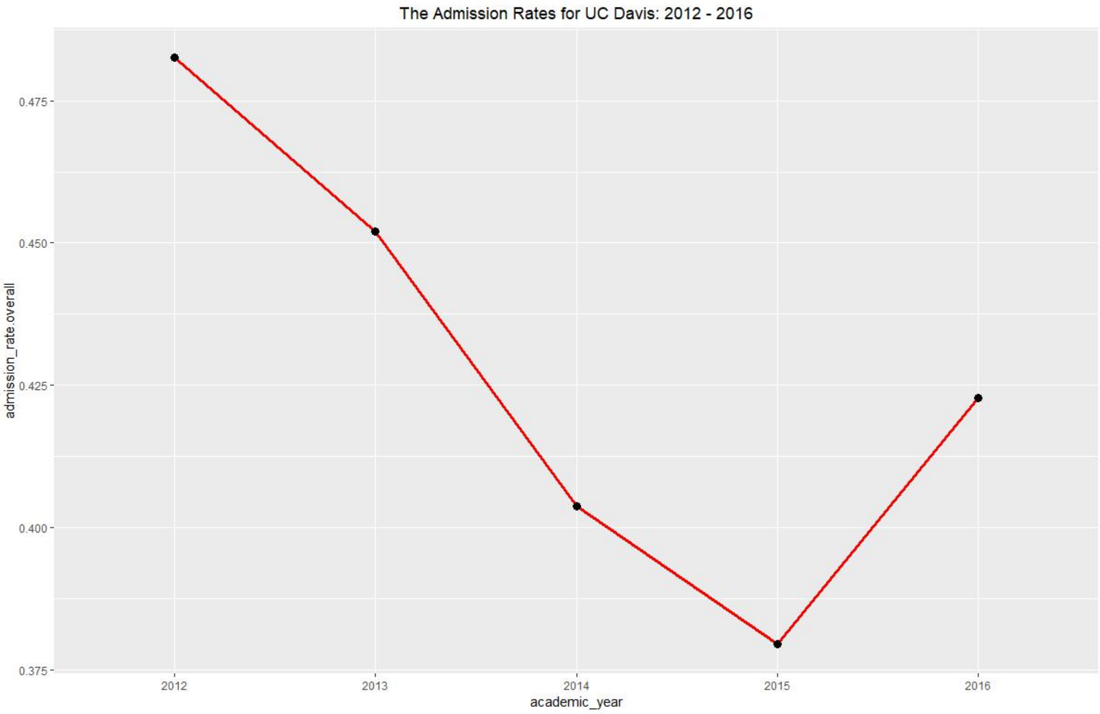
Also, for public school students, the median student earnings after 10 years mainly concentrates between \$10000-\$75000; for private profit school students, the median student earnings after 10 years mainly concentrates between \$10000-\$60000; for private nonprofit school students, the median student earnings after 10 years mainly concentrates between \$10000-\$75000. Thus, students are more likely to get more earnings after 10 years in public and private nonprofit schools.

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.



For private nonprofit and public schools, the number of schools recorded has a similar trend: 2015 and 2016 have a little more recorded schools, compared to 2012,2013 and 2014’s. However, private for profit schools have a different trend. The number of schools recorded decreases form 2012 to 2016.

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 admission rates change much from year to year. The admission rates change a lot from 2013 to 2014. Also, from 2012 to 2015, the admission rate decreases, while from 2015 to 2016, the admission rate increases.

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 : integer, character, logical, double
Statistical data types : nominal, ordinal, discrete and continuous.
We can use typeof() to get R data types. R data type does not map to just one statistical data type. For example, State’s R data type is “integer”, while its statistical data type is “nominal”. Degrees_awarded.highests R data type is also “integer”, while its statistical data type is “ordinal”. Also, name’s R data type is “character”, while its statistical data type is “nominal”. Academic_year’s R data type is “character”, while its statistical data type is “discrete”.

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.

(a) Compare program percentage in California colleges.

Each student has their favorite program. If we can present the program percentage graphically, it will be helpful for students to make a decision which college is much more suitable in California.
Variables I would use:

$ program_percentage.agriculture $ program_percentage.resources
$ program_percentage.architecture $ program_percentage.ethnic_cultural_gender
$ program_percentage.communication $ program_percentage.communications_technology
$ program_percentage.computer $ program_percentage.personal_culinary
$ program_percentage.education $ program_percentage.engineering
$ program_percentage.engineering_technology $ program_percentage.language
$ program_percentage.family_consumer_science $ program_percentage.legal
$ program_percentage.english $ program_percentage.humanities
$ program_percentage.library $ program_percentage.biological
$ program_percentage.mathematics $ program_percentage.military
$ program_percentage.multidiscipline $ program_percentage.parks_recreation_fitness
$ program_percentage.philosophy_religious
$ program_percentage.theology_religiou $ program_percentage.psychology
$ program_percentage.security_law_enforcement $ program_percentage.construction
$ program_percentage.public_administration_social_service
$ program_percentage.social_science $ program_percentage.mechanic_repair_technology
We can get the percentage of several programs for each college, then pie chart is helpful to show which program is the most popular in one college.

(b) **Compare tuition cost in California states.**
Having a good understanding of tuition will be helpful for students and families to choose the most suitable college, combined with the understanding of the outcomes of question 12(a).
Variable I should use:
$ tuition.in_state
We can use bar chart to compare which college has the higher tuition.

(c)Which colleges are the most racially diverse?
From the outcome of that question, it will be useful for students who want to attend college that have a national diversity.
Variables I should use:
$ demographics.race_ethnicity.white
$ demographics.race_ethnicity.black
$ demographics.race_ethnicity.hispanic
$ demographics.race_ethnicity.asian
$ demographics.race_ethnicity.aian
$ demographics.race_ethnicity.nhpi
$ demographics.race_ethnicity.two_or_more
$ demographics.race_ethnicity.non_resident_alien
$ demographics.race_ethnicity.unknown

We can find the college which has the most variance of the percentage of race to get the college has the most diverse.

Appendix

R code:

```
# 2
data = readRDS("college_scorecard.rds")
nrow(data)
head(data)

# 3
ncol(data)
colnames(data)

# 4
data$academic_year = factor(data$academic_year)
levels(data$academic_year)
nrow(data[data[,'academic_year'] == '2012',])
nrow(data[data[,'academic_year'] == '2013',])
nrow(data[data[,'academic_year'] == '2014',])
nrow(data[data[,'academic_year'] == '2015',])
nrow(data[data[,'academic_year'] == '2016',])

# or
table(data$academic_year)

# 5
data$state = factor(data$state)
a=levels(data$state)
list = rep(0,length(a))
for (i in 1:length(list)) {
  list[i] = nrow(data[data[,'state'] == a[i],])
}
sort(list)
a[which(list == 5)]

a[which(list == 2022)]
a[which(list == 2176)]
a[which(list == 2317)]
a[which(list == 2381)]
a[which(list == 3881)]

# 6
data_2014 = data[data[,'academic_year'] == '2014',]
data$ownership = factor(data$ownership)
levels(data$ownership)
```

```
data_public_2014 = data_2014[data_2014[,'ownership'] == 'Public',]

library(ggplot2)
# Basic scatter plot
ggplot(data_public_2014, aes(x = data_public_2014$avg_net_price.public, y =
data_public_2014$earn_10_yrs_after_entry.median
)) + geom_point()+ theme(plot.title=element_text(hjust=0.5))+
   labs(title = "scatter plot of average net price versus median student earnings after 10 years for
Public School", x= "Avg_Net_Price_Public", y = "Earn_10_Yrs_After_Entry_Median")

# 7
data_profit_2014 = data_2014[data_2014[,'ownership'] == 'Private for-profit',]
ggplot(data_profit_2014, aes(x = data_profit_2014$avg_net_price.private, y =
data_profit_2014$earn_10_yrs_after_entry.median)) +
   geom_point()+ theme(plot.title=element_text(hjust=0.5))+
   labs(title = "scatter plot of average net price versus median student earnings after 10 years for
Private Profit School", x= "Avg_Net_Price_Profit", y = "Earn_10_Yrs_After_Entry_Median")

# 8
data_nonprofit_2014 = data_2014[data_2014[,'ownership'] == 'Private nonprofit',]
ggplot(data_nonprofit_2014, aes(x = data_nonprofit_2014$avg_net_price.private, y =
data_nonprofit_2014$earn_10_yrs_after_entry.median)) +
   geom_point()+ theme(plot.title=element_text(hjust=0.5))+
   labs(title = "scatter plot of average net price versus median student earnings after 10 years for
Private NonProfit Schools", x= "Avg_Net_Price_Profit", y = "Earn_10_Yrs_After_Entry_Median")

# combine the three plots in one plot
a = data$avg_net_price.public
b = data$avg_net_price.private
a[is.na(a)] = 0
b[is.na(b)] = 0

avg_net_price = a+b
avg_net_price[avg_net_price == 0] = NA

data$avg_net_price = avg_net_price

ggplot(data, aes(x = data$avg_net_price, y = data$earn_10_yrs_after_entry.median, color =
ownership)) +
   geom_point()+ theme(plot.title=element_text(hjust=0.5))+
   labs(title = "scatter plot of average net price versus median student earnings after 10 years", x=
"Average Net Price", y = "Median Student Earnings after 10 Years")+
   scale_color_manual(values = c("Private for-profit"="darkseagreen2", "Private
nonprofit"="gold1","Public"="lightskyblue1"))+
   theme(legend.position="top")+ ylim(0, 150000) + xlim(0,80000)

# 9
Academic_year = data$academic_year
ggplot(data,aes(x = data$ownership, fill = Academic_year)) + geom_bar(position = "dodge")+
   scale_fill_manual(values = c("2012"="lightcyan2",
"2013"="lightsteelblue1","2014"="lightskyblue1","2015"="lightskyblue2","2016"="lightblue3"))+
   theme(legend.position="top")+
   labs(title = "The Number of Schools for Different Ownership: 2012 - 2016", x= "")+
theme(plot.title=element_text(hjust=0.5))

# 10
data_UCD = data[data[,'id'] == '110644',]
ggplot(data=data_UCD, aes(x=academic_year, y=admission_rate.overall, group=1)) +
   geom_line(color="red", size=1.2)+geom_point(size=3)+
   labs(title = "The Admission Rates for UC Davis: 2012 - 2016")+
theme(plot.title=element_text(hjust=0.5))

# 11
for(a in names(data)){
   print(c(a, typeof(data[, a])))
}
typeof(data$state)
typeof(data$degrees_awarded.highest)
typeof(data$name)
typeof(data$academic_year)
```