STA141AHW1

Sam Tsoi

Due: 10/17/2017

1.

1. How many observations are recorded in the dataset? How many colleges are recorded?

```
## [1] 3312   51
```

```
## [1] 3312
```

```
## [1] 2431
```

There are 3312 observations recorded in the dataset. There are 2431 college campuses in the data.

2.

How many features are there? How many of these are categorical? How many are discrete? Are there any other kinds of features in this dataset?

```
## [1] 51
```

```
## var_class
## character    factor   integer   logical   numeric
##         4         4        15         3        25
```

There are 51 features in this data. There are 8 features that are categorical (since 4 features are 'factors' and 4 features are are 'characters'), 40 that are discrete (since 25 are 'numeric' and 15 are 'integers'), and 3 features that are 'logical' (TRUE or FALSE). Some examples of the features in the character class are ope_id, name, city, zip.

3.

How many missing values are in the dataset? Which feature has the most missing values? Are there any patterns?

```
## [1] 23197
```

```
## avg_sat
##      14
```

```
## unit_id
##       1
```

```
## [1] 1923
```

```
## avg_sat
##      14
```

```
## [1] 3127
```

This is a summary table of how many missing values are in each feature: 0, 0, 0, 0, 0, 0, 0, 0, 0, 199, 0, 0, 0, 1923, 490, 1149, 717, 472, 635, 201, 201, 317, 558, 1416, 966, 864, 510, 510, 267, 267, 480, 480, 480, 487, 489, 183, 538, 792, 267, 1735, 505, 490, 490, 490, 490, 490, 490, 490, 490, 689, 490. There's a total of 23197 missing values in the dataset. The feature that had the most missing values is at index 14, which is the average SAT score of the college. The feature that had the least missing value is at index 1, which is the unit ID of the college. But, we found that when we list the missing values of each feature, there are 11 others (12 total) features that also had no missing values. Some patterns I observe are that IDs, names, zip code, etc. (probably features that are mandated by law) have less missing values. These data probably does not require one to obtain data, as these features are just identifications of a campus. On the other hand, things that change every year and features that are discrete have more missing values, such as average SAT, graduation population, veteran, etc.

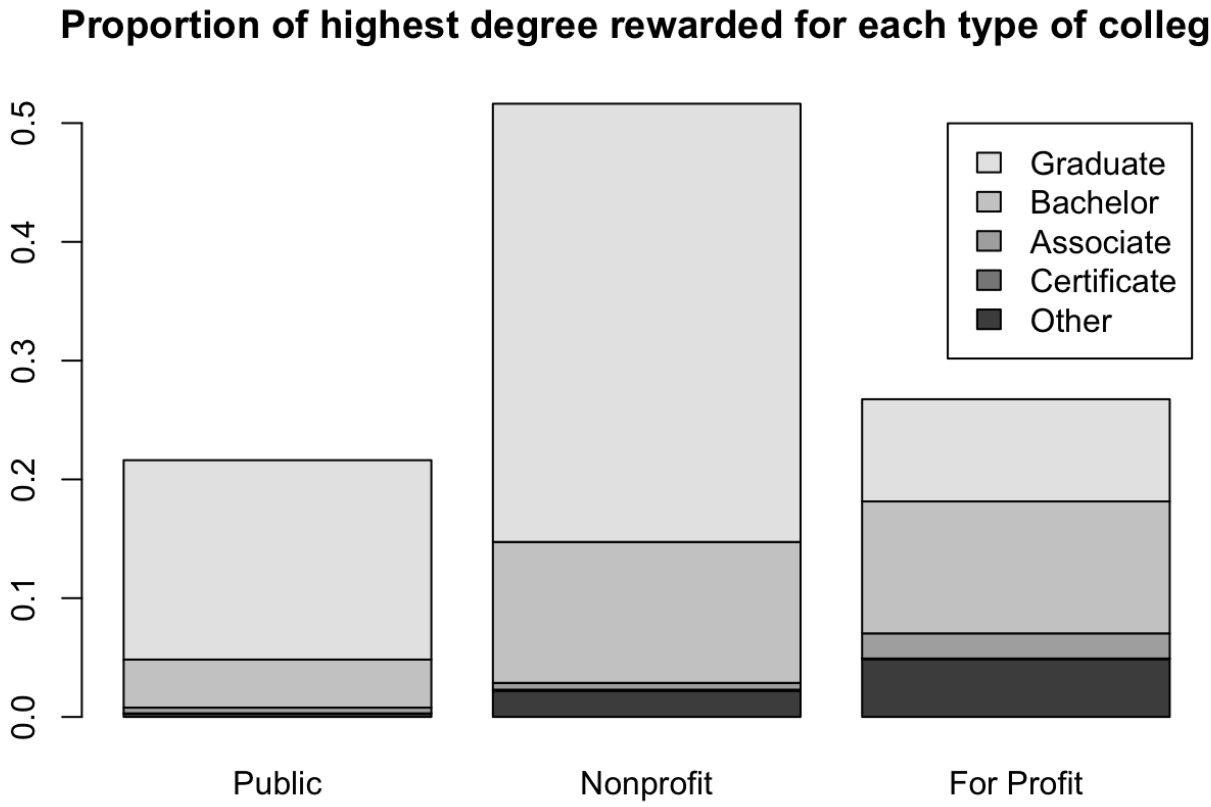
4.

Are there more public colleges or private colleges recorded? For each of these, what are the proportions of highest degree awarded? Display this information in one graph and comment on what you see.

```
##
##     Public  Nonprofit For Profit
##        716       1710        886
```

There are more private colleges recorded, as in the data, there are 716 public colleges, 1710 nonprofit colleges, and 886 for profit colleges. The proportion of the highest degrees are as follows, separated for public, nonprofit, and for profit schools:

```
##               ownership
## highest_degree       Public    Nonprofit   For Profit
##    Other       0.0030193237 0.0220410628 0.0486111111
##    Certificate 0.0000000000 0.0009057971 0.0006038647
##    Associate   0.0048309179 0.0057367150 0.0211352657
##    Bachelor    0.0404589372 0.1186594203 0.1111111111
##    Graduate    0.1678743961 0.3689613527 0.0860507246
```



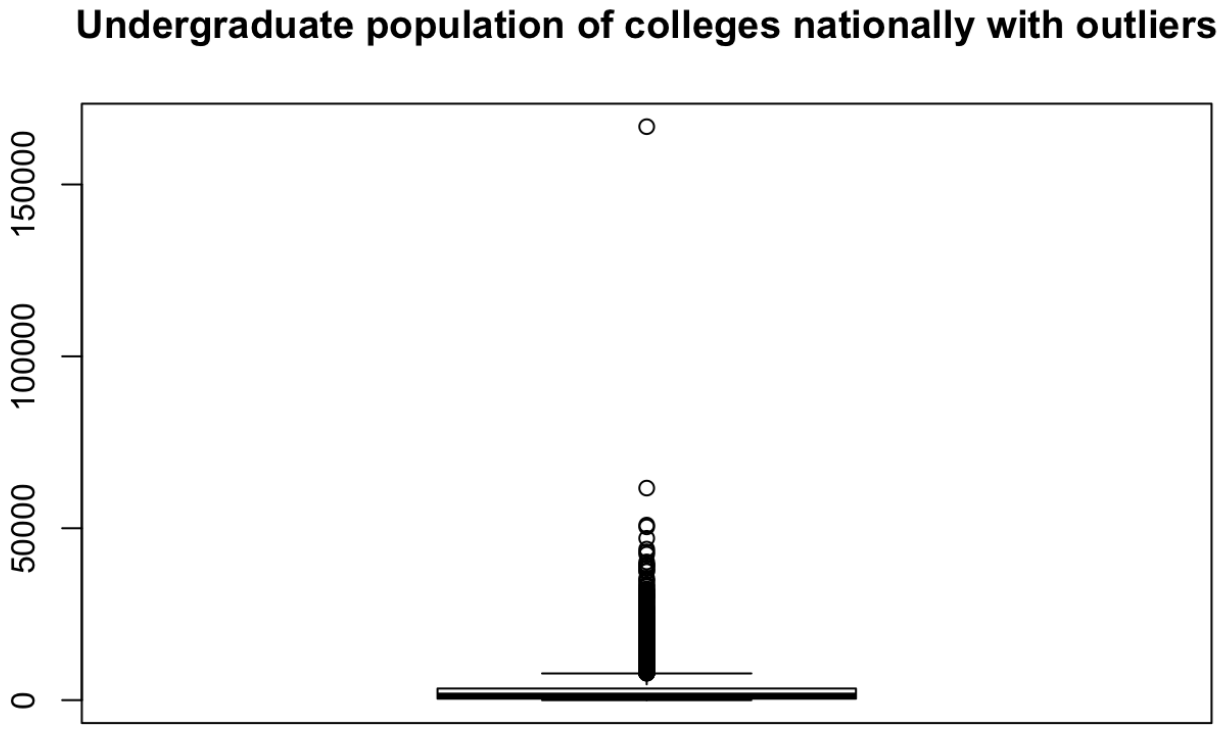
As shown, the most highest degree awarded for public and nonprofit schools are graduate degrees, whereas for for-profit schools is bachelor degree. Interestingly, public schools do not award certificates and only private schools do. For-profit schools award associate degrees the most out of the 3 school, and public schools seem to only award bachelor, graduate, and associate degrees the most. The greatest proportion of degrees awarded are nonprofit private schools, then for profit private schools, then public schools. This might suggest how majority of the education system is still owned by business owners, and aren't necessarily for the "good of the people" (as suggested by the large proportion of degree awarded by for profit schools), but is still for making money. I presume that the larger amount of certificate, associate, and other degree awarded by for-profit schools are to propel individuals to the working force of the economy. This includes technicians and other jobs that require licensing and other technical skills.

5.

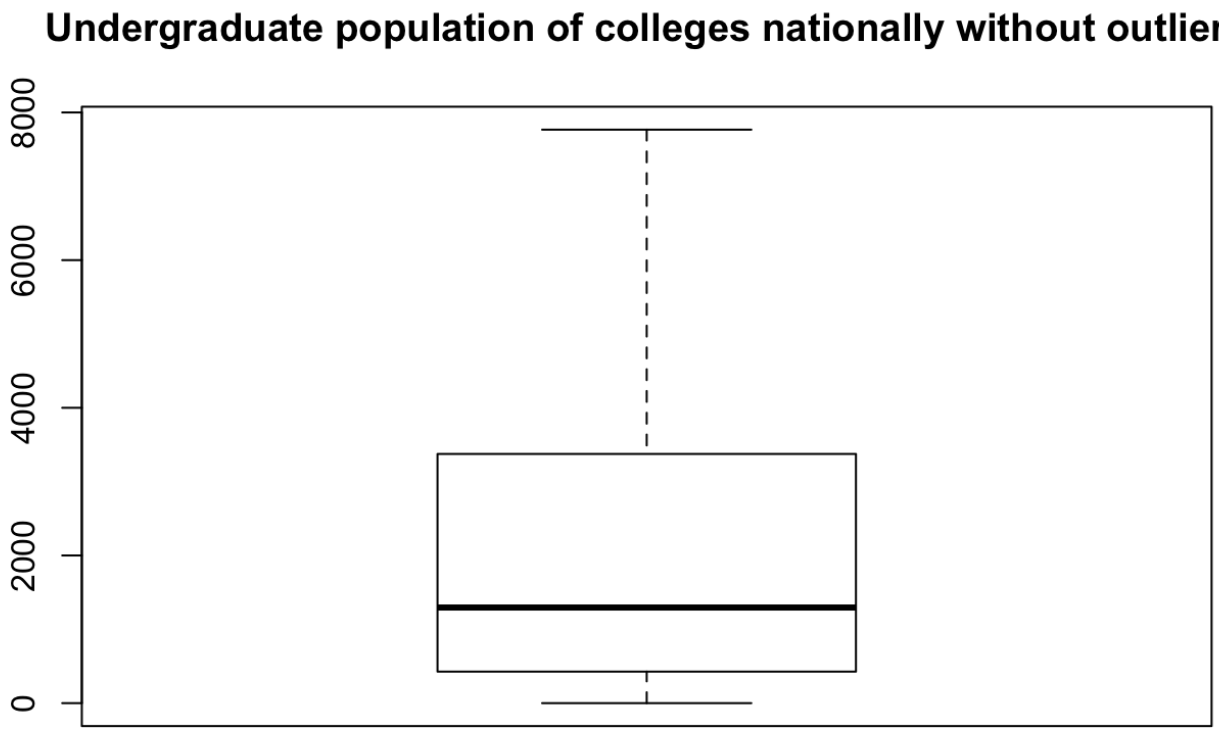
What is the average undergraduate population? What is the median? What are the deciles? Display these statistics and the distribution graphically. Do you notice anything unusual?

```
## [1] 3599.502
```

```
## [1] 1295
```



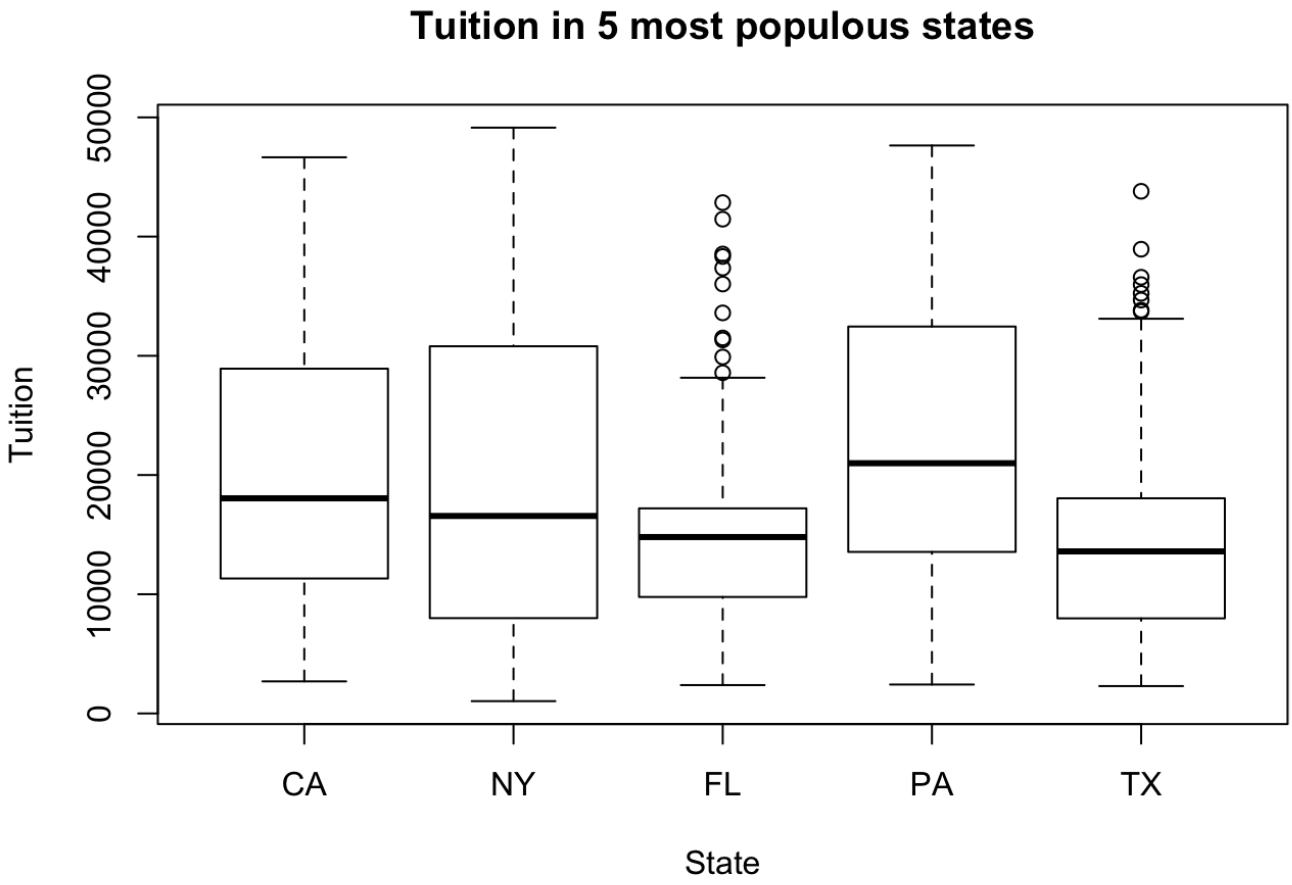
Zooming into the quartiles of the graph, I will remove outliers:



The mean (average) undergraduate population is 3599.5021262, and the median is 1295. The quantiles and deciles are as follows 0, 428, 1295, 3372, 1.6681610^{5}. This demonstrates that the data is skewed right, so median is lower than the mean. This might mean that there are outliers that have a really large population of undergraduate students that would shift the mean to larger than the median. With both the boxplots with and without the outliers, it seems like there are a few schools that have really really large populations of undergraduate students, compared to those of colleges nationally. Additionally, there is one school that has an undergraduate population of 166816, which I found is the University of Pheonix-Online Campus. Online campuses data might differ from the traditional in-class college data.

6.

Compare tuition graphically in the 5 most populous states. Discuss conclusions you can draw from your results.

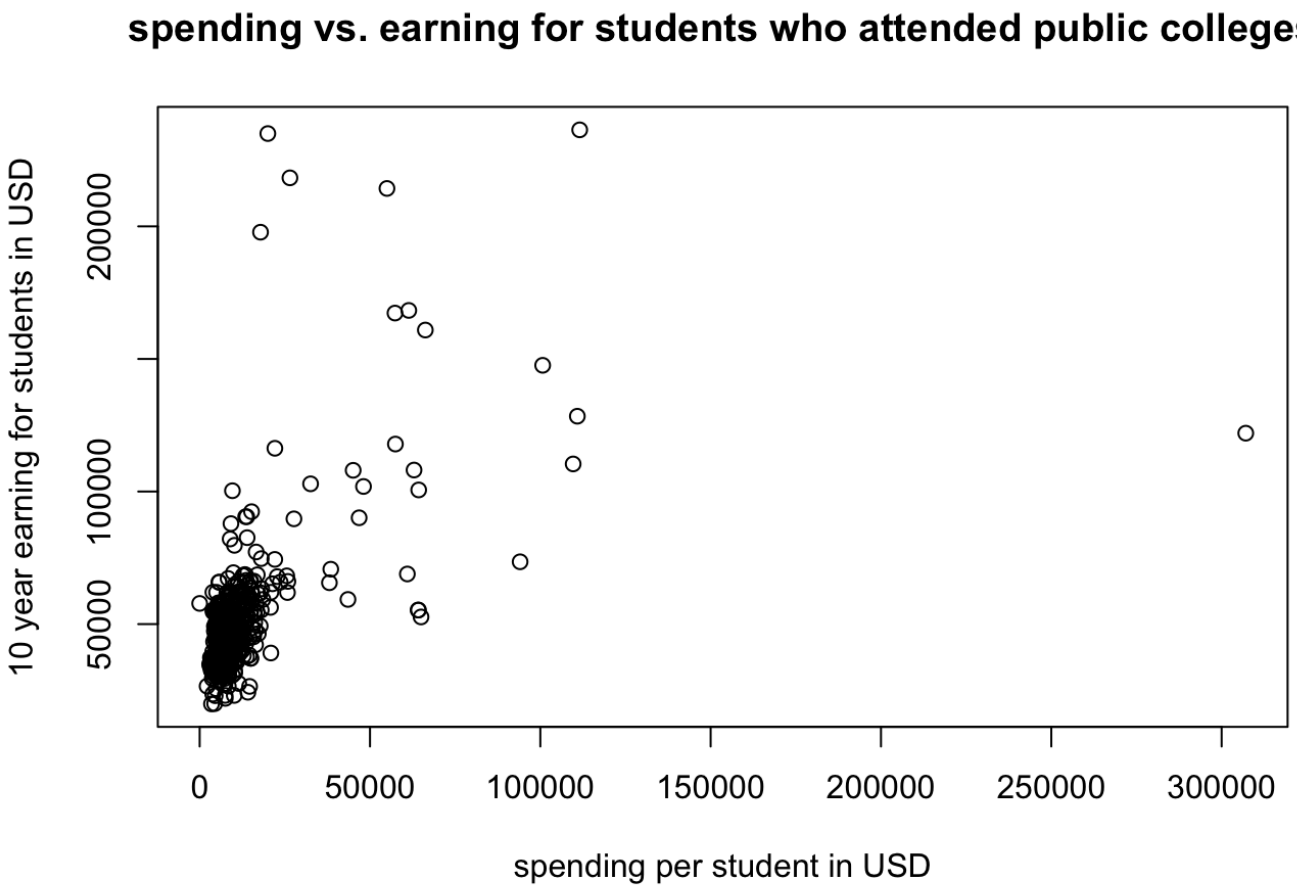


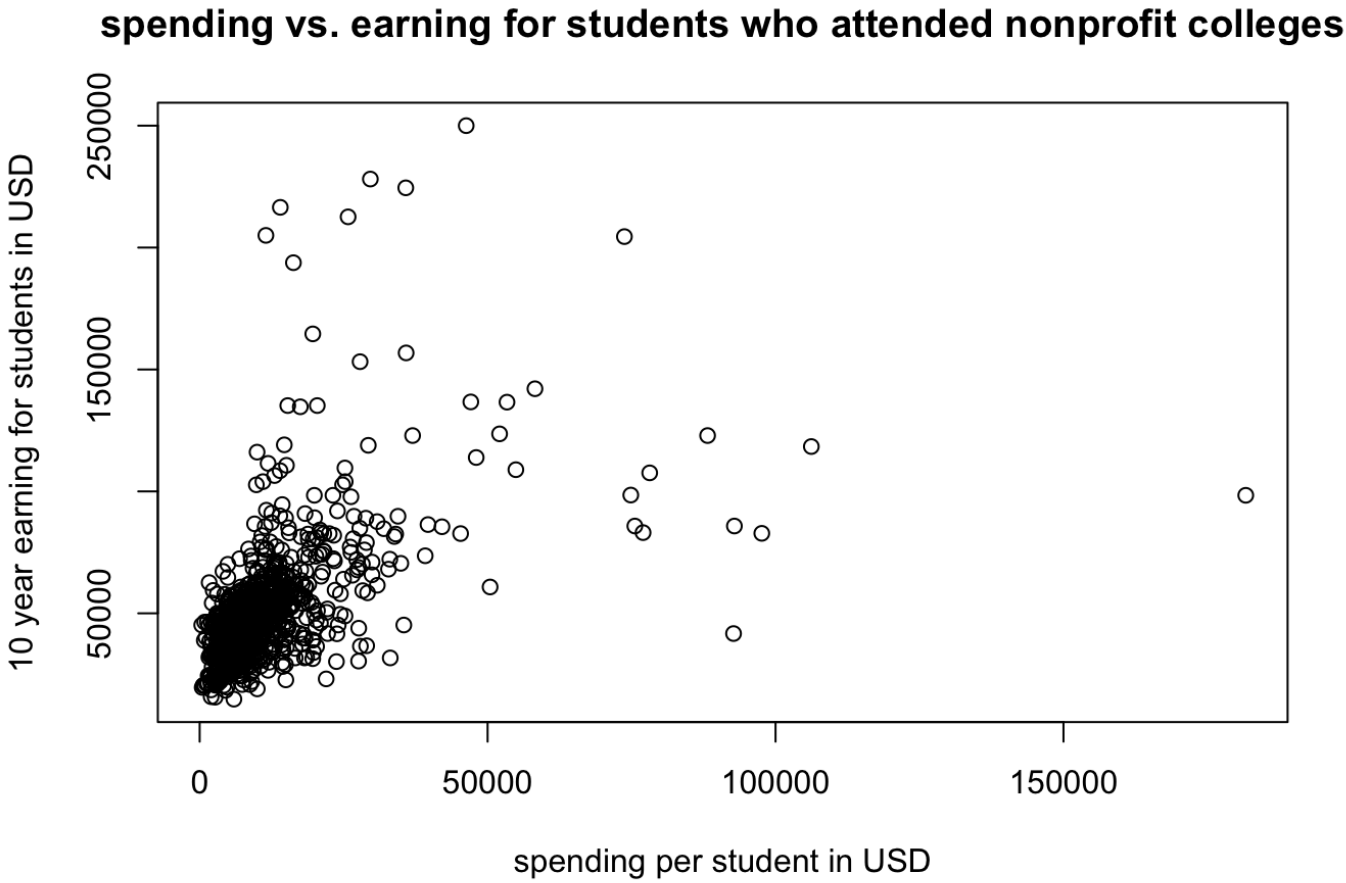
Theres a wider range of the 1st quantile and 3rd quantile in New York than that of any of the other 4 states. There are outliers in colleges in Florida and Texas, while there are no outliers in California, New York, and Pennsylvania. Colleges in Pennsylvania has the highest median tuition compared to that of the other 4 states, and New York has colleges the highest tuition overall compared to that of other 4 states. It is interesting that these two states are both in the Northeast, while Florida and Texas are more so in the south of the US, and has lower median tuitions in college. At the same time, though, New York also has the lowest tuition overall compared to that of other 4 states. Tuition median range between $1.359410^{4} (Texas) and $2.09910^{4} (Pennsylvania) yearly in the 5 most populous states, and the mean tuition of all colleges is 1.761052610^{4}.

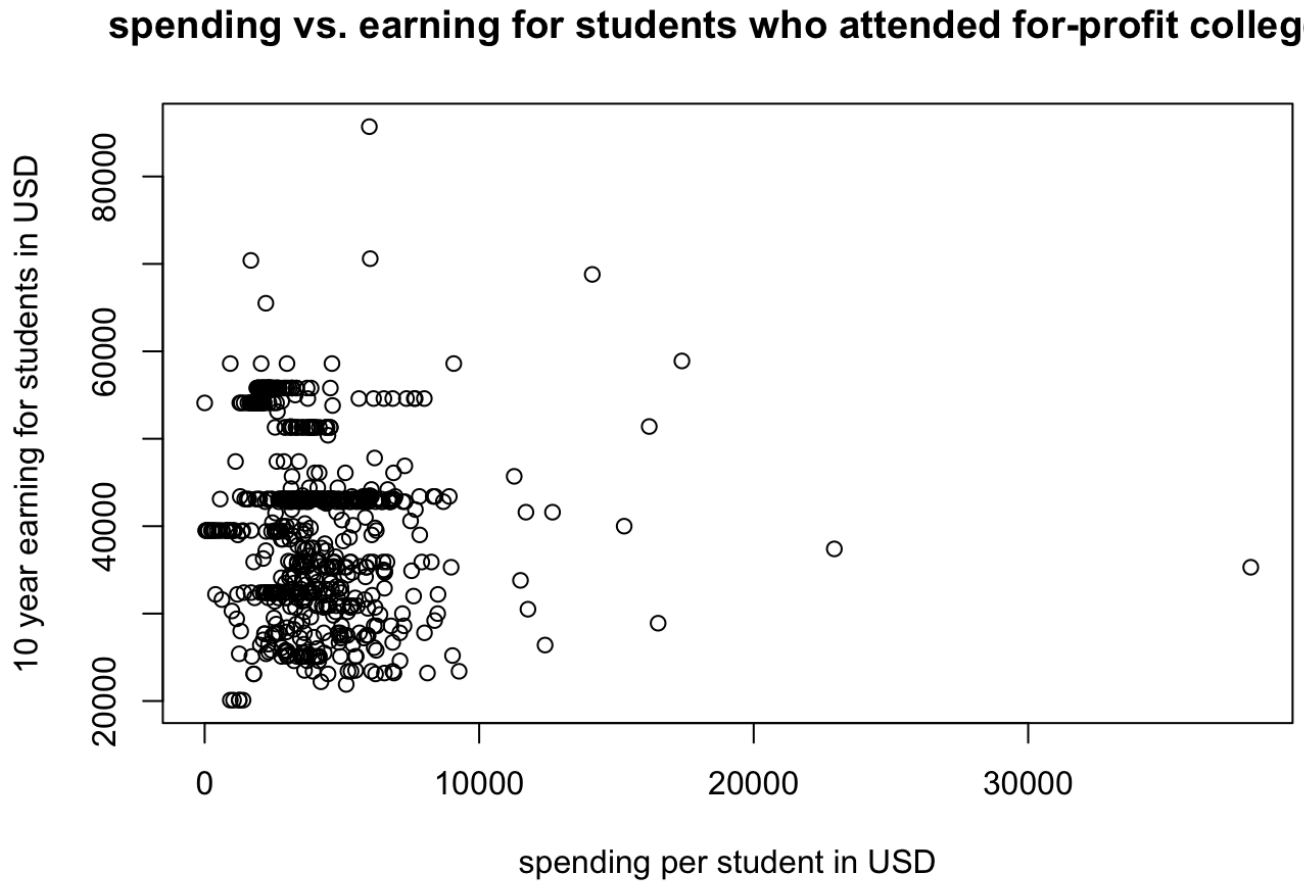
7.

Display and comment on how spending per student (by the college) and students' 10-year earnings are

related. Is this relationship affected by whether a college is public, nonprofit, or for profit?







There doesn't seem to be strong correlation in any of the different types of colleges (public, nonprofit, for profit). There is a slight positive correlation of spending per student to 10-year earning for public and nonprofit colleges, but there is no correlation for spending per student to 1-year earning for for-profit colleges. Spending seems to differ no matter how much people earned over the course of 1-0 years for for-profit colleges. However, in non-profit colleges, there seems to be greater spending vs. 10-year earning, but this relationship isn't very strong.

8.

Which colleges give the best earnings for the cost? Explain how you determined this. Discuss limitations of your result and features2 you did not examine that could confound your result.

I determined the best earning for the cost by dividing the average 10-year salary of each college by the cost of the college. Then, I sorted the earnings for the cost from largest to smallest to see that the top ten best earning for the cost are 13.9936393, 8.2047959, 5.3337492, 4.7891241, 4.7589807, 4.7331857, 4.7100909, 4.613019, 4.5042126, 4.4740372. This means that for each dollar spent on the cost of college, you earn $13.9936393, 8.2047959, 5.3337492, 4.7891241, 4.7589807, 4.7331857, 4.7100909, 4.613019, 4.5042126, 4.4740372 back over the course of 10 years. The colleges that correspond with the best earning for the cost are United States Merchant Marine Academy, Augusta University, South Texas College, University of Connecticut-Avery Point, University of Connecticut-Stamford, University of Connecticut-Tri-Campus, Indian River State College, Palm Beach State College. What I find interesting is that

```
## [1] Public Public Public Public Public Public Public Public
## Levels: Public Nonprofit For Profit
```

all of these schools are public schools.

Some limitations… We are not taking into account of student debt. For those students who had to take out a

loan, an interest is not included into this calculation. So, for some students, they may have used some of their earnings throughout the 10-year period to pay back all the college tuition + interest. Other factors, such as majors, can impact which college has greater earnings. For example, there may be schools that have greater number of STEM students, such as technical schools, and these students may have greater earnings than those who are not STEM. While this may not be true, this is an example of how different colleges may produce different earnings and disregards the “prestige” of certain colleges.

9. Which colleges are the most racially diverse? Explain the strategy you used to determine this.

```
##  [1] "California State University-East Bay"
##  [2] "Golden Gate University-San Francisco"
##  [3] "Holy Names University"
##  [4] "Andrews University"
##  [5] "Vaughn College of Aeronautics and Technology"
##  [6] "LIU Brooklyn"
##  [7] "Houston Baptist University"
##  [8] "The University of Texas MD Anderson Cancer Center"
##  [9] "University of Phoenix-Hawaii"
## [10] "Pacific Rim Christian University"
```

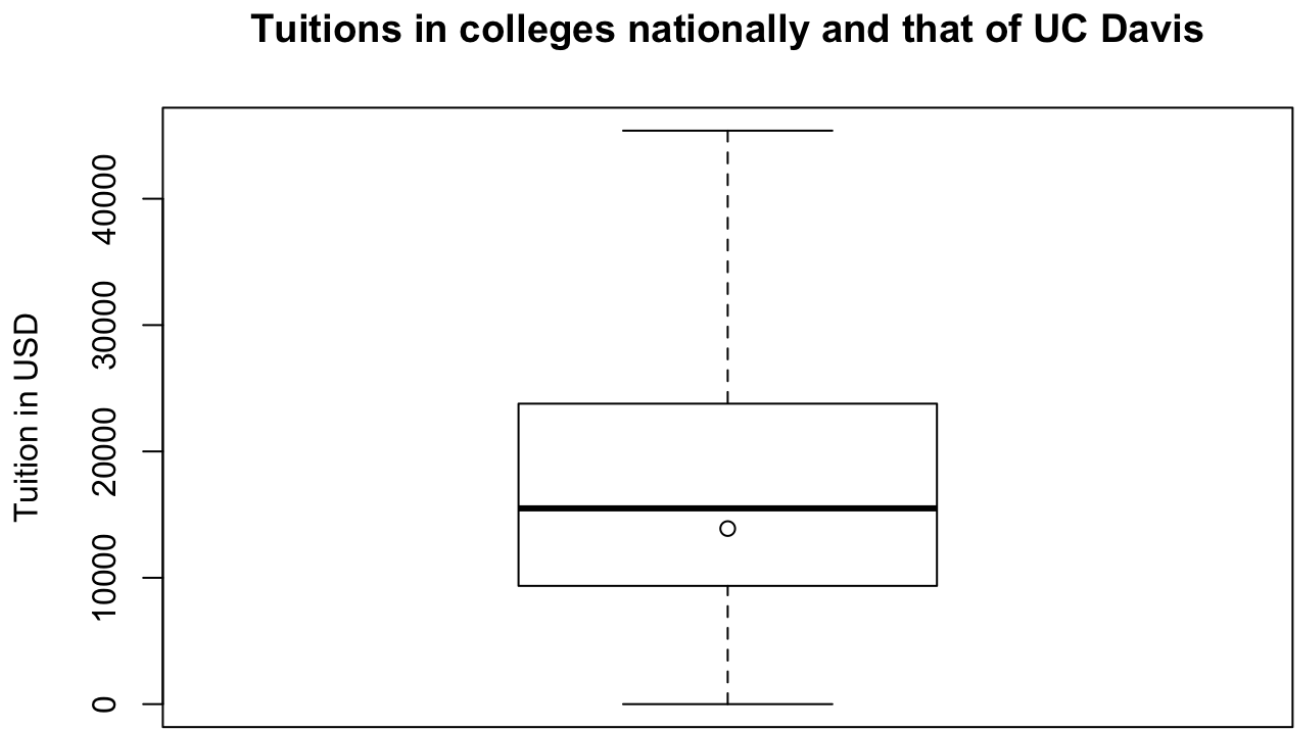
```
##  [1] 3.8650 3.9251 4.0212 4.0764 4.1828 4.2068 4.2220 4.2235 4.2879 4.3067
```

I found that To see which colleges are the most racially diverse, I created a diversity “score” in which the closest the score is to 0, the greater the racial diversity. I found that the “ideal diversity” of each race should be around 0.1428571, as there are 7 categories of race, which I divided equally for each race. Then, for each race, I subtracted the proportion to the ideal diversity and then divided the ideal diversity, so we can see how much each race differed from the “ideal diversity”. I then added each of these numbers for each race together for the college to compute the “diversity score”. Thus, the closer each diversity is to 0.1428571, the closer the diversity score would be to 0. The larger the diversity score is, the less racially diverse the college would be.
I found that California State University-East Bay, Golden Gate University-San Francisco, Holy Names University, Andrews University, Vaughn College of Aeronautics and Technology, LIU Brooklyn, Houston Baptist University, The University of Texas MD Anderson Cancer Center, University of Phoenix-Hawaii, Pacific Rim Christian University are the top 10 most diverse, according to my “diversity score”, with Holy Names University being the most diverse with a score of 3.8650, and LIU Brooklyn being second with a score of 3.9251. The diversity scores are as follows: 3.865, 3.9251, 4.0212, 4.0764, 4.1828, 4.2068, 4.222, 4.2235, 4.2879, 4.3067. The least diverse colleges have a diversity score of 12.

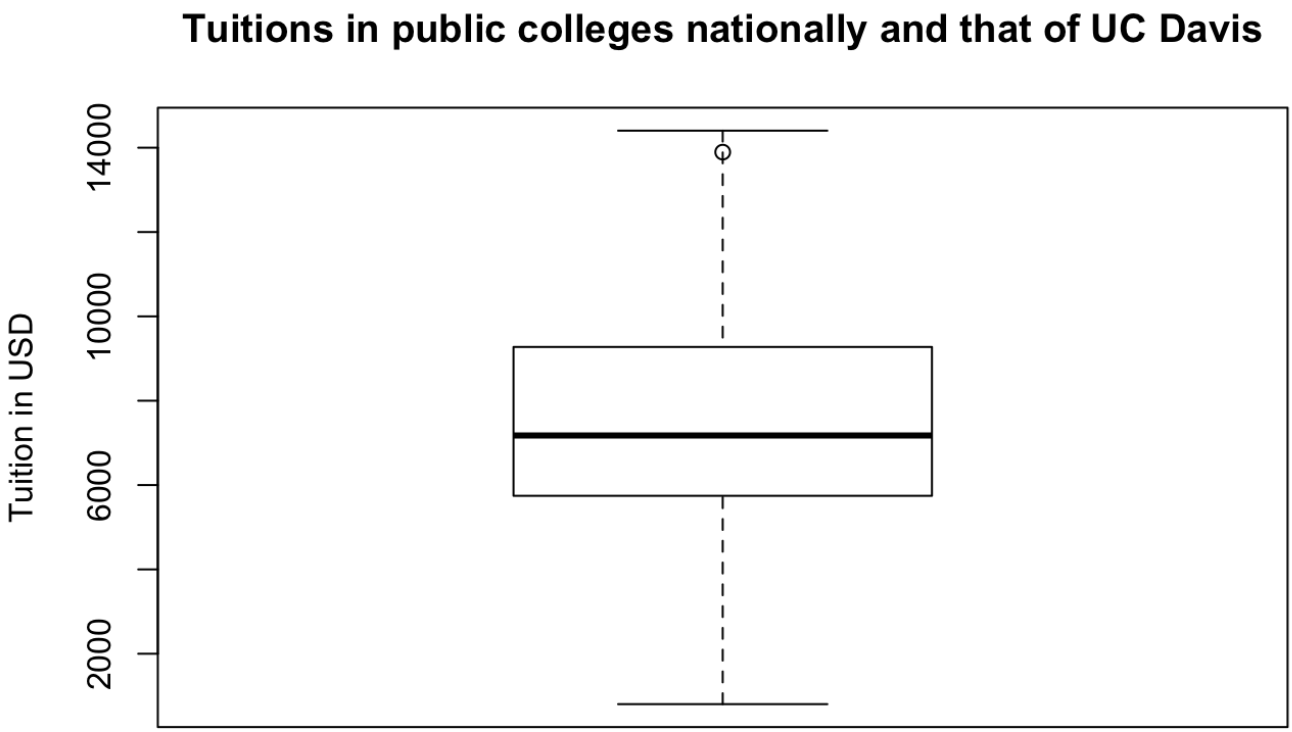
10.

How does UC Davis compare to other colleges in the nation? Use statistical summaries and graphics to examine at least 3 characteristics that students might be interested in.

```
## [1] 13895
```

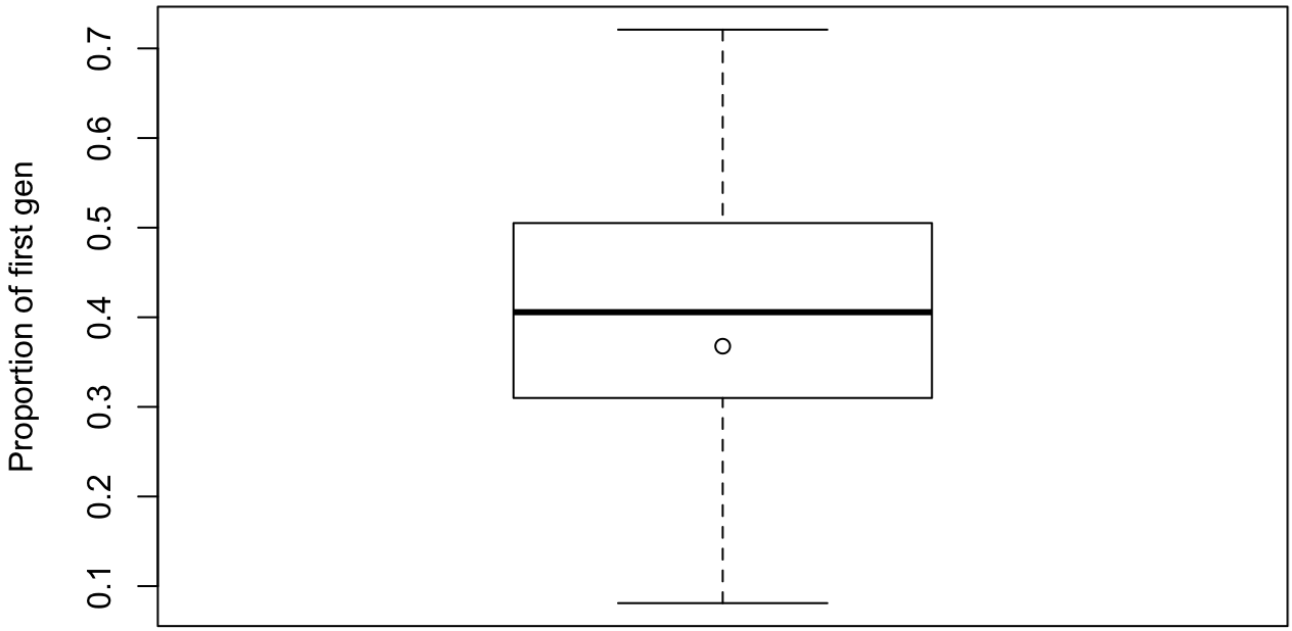


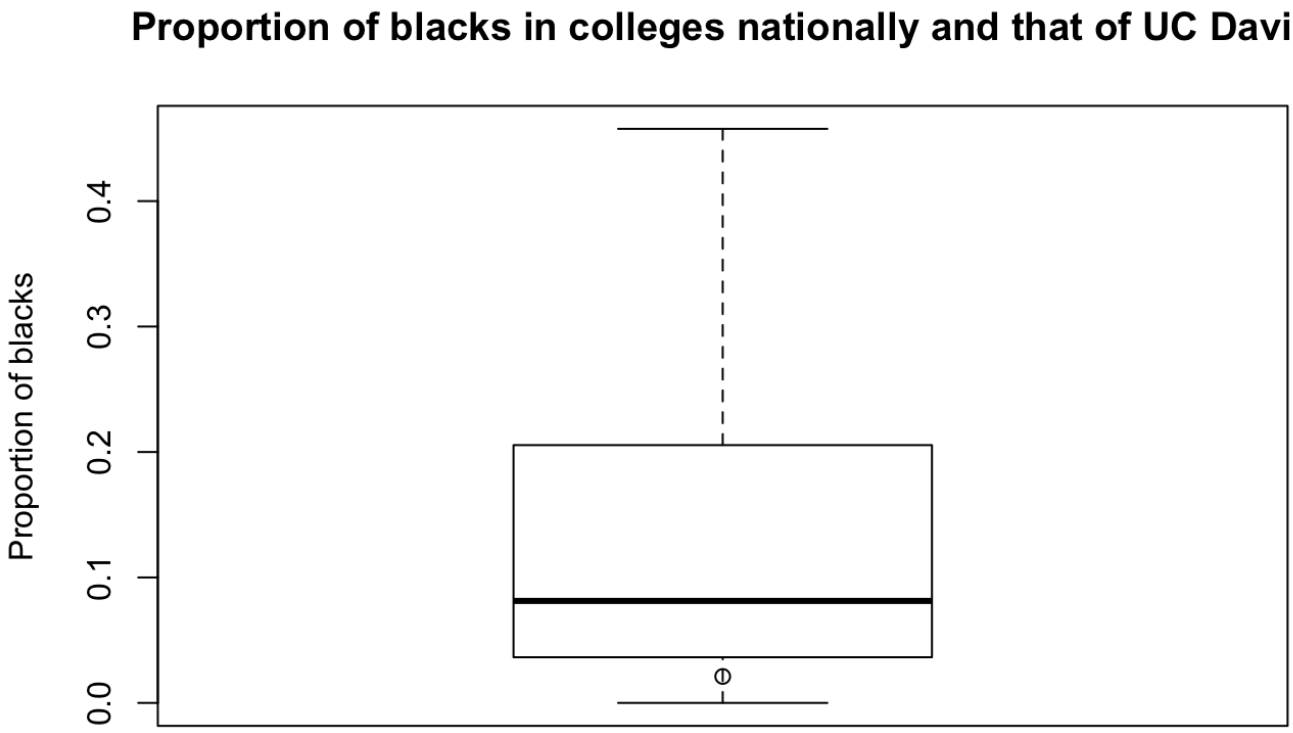
The tuition recorded at UC Davis is $13895, which seems to be lower than the median of tuition in colleges nationally. This might have to do with the fact that UC Davis is a public school, and the data includes both public and private schools. So, I looked at how tuition of UC Davis compares to those of other public colleges:



This box plot demonstrates that the tuition of UC Davis is at the latter end of the spectrum, and is much higher than the median tuition of public colleges nationally.

Looking at more the diversity of UC Davis and that of other colleges,





UC Davis has a lower proportion of first generations compared to that of colleges nationally, as the proportion is lower than the median proportion of all colleges. Similarly, UC Davis has a lower proportion of

blacks compared to that of colleges nationally, as the proportion is lower than the 1st quartile proportion of all colleges.

Citations:

Colloborated with Brody Lowry

Looked on Piazza for tips on how to do diversity score and using %in% instead of ==

Consulted lecture notes for #1 to #4.

Code Appendix

```
hwdata1 <- readRDS("~/Desktop/college_scorecard_2013.rds")

dim(hwdata1)
nrow(hwdata1)
sum(hwdata1$main)
ncol(hwdata1)
var_class<-sapply(hwdata1, class)

table(var_class)
sum(var_class=="factor")
sum(var_class=="factor")
sum(var_class=="numeric")
sum(var_class=="integer")
names(var_class[var_class=="character"])
sum(is.na(hwdata1))


which.max(colSums(is.na(hwdata1)))
which.min(colSums(is.na(hwdata1)))

sum(is.na(hwdata1$avg_sat))

#length(colSums(is.na(hwdata)))

#length(sapply(as.data.frame(is.na(hwdata)), sum))

which.max(sapply(as.data.frame(is.na(hwdata1)), sum))

which.max(apply(as.matrix(is.na(hwdata1)),1, sum))
table(hwdata1$ownership)
degree_dat<-hwdata1[c("primary_degree","highest_degree","ownership")]

degree_dat_tab<-table(degree_dat)
dimnames(degree_dat_tab)

prop.table(degree_dat_tab,margin=1)
prop.table(degree_dat_tab,margin=2)
highAndOwn <-hwdata1[c("highest_degree","ownership")]
highestDeg <-prop.table(table(highAndOwn))
highestDeg
#mosaicplot(prop.table(degree_dat_tab,margin=1),color=TRUE, shade=TRUE)

barplot(highestDeg, legend=T, main = "Proportion of highest degree rewarded for eac
h type of college")
mean(hwdata1$undergrad_pop, na.rm=TRUE)
median(hwdata1$undergrad_pop, na.rm=TRUE)
boxplot(hwdata1$undergrad_pop, na.rm=TRUE, main="Undergraduate population of colleg
es nationally with outliers")
quantile(hwdata1$undergrad_pop, na.rm=TRUE)
hwdata1$name[hwdata1$undergrad_pop == 166816]
boxplot(hwdata1$undergrad_pop, na.rm=TRUE,outline=FALSE, main="Undergraduate popula
tion of colleges nationally without outliers")
meantut <- mean(hwdata1$tuition, na.rm = TRUE)
tuitionCA <- hwdata1$tuition[hwdata1$state %in% "CA"]
```

```
tuitionNY <- hwdata1$tuition[hwdata1$state %in% "NY"]
tuitionFL <- hwdata1$tuition[hwdata1$state %in% "FL"]
tuitionPA <- hwdata1$tuition[hwdata1$state %in% "PA"]
tuitionTX <- hwdata1$tuition[hwdata1$state %in% "TX"]
medTX <- median(tuitionTX, na.rm=TRUE)
medPA <- median(tuitionPA, na.rm=TRUE)
boxplot(tuitionCA,tuitionNY,tuitionFL,tuitionPA,tuitionTX, names = c("CA","NY","FL"
,"PA","TX"), xlab= "State",ylab="Tuition", main="Tuition in 5 most populous states"
)

spendingPub <- hwdata1$spend_per_student[hwdata1$ownership %in% "Public"]
spendingPub2 <- spendingPub[!is.na(spendingPub)]
spendingNon <- hwdata1$spend_per_student[hwdata1$ownership %in% "Nonprofit"]
spendingNon2 <- spendingNon[!is.na(spendingNon)]
spendingFor <- hwdata1$spend_per_student[hwdata1$ownership %in% "For Profit"]
spendingFor2 <- spendingFor[!is.na(spendingFor)]
earningPub <- hwdata1$avg_10yr_salary[hwdata1$ownership %in% "Public"]
earningPub2 <- earningPub[!is.na(earningPub)]
earningNon <- hwdata1$avg_10yr_salary[hwdata1$ownership %in% "Nonprofit"]
earningNon2 <- earningNon[!is.na(earningNon)]
earningFor <- hwdata1$avg_10yr_salary[hwdata1$ownership %in% "For Profit"]
earningFor2 <- earningPub[!is.na(earningFor)]
plot(spendingPub,earningPub, main = "spending vs. earning for students who attended
public colleges", xlab = "spending per student in USD", ylab ="10 year earning for
students in USD")
plot(spendingNon,earningNon, main = "spending vs. earning for students who attended
nonprofit colleges", xlab = "spending per student in USD", ylab ="10 year earning f
or students in USD")
plot(spendingFor,earningFor, main = "spending vs. earning for students who attended
for-profit colleges", xlab = "spending per student in USD", ylab ="10 year earning
for students in USD")
hwdata1$best <- hwdata1$avg_10yr_salary/hwdata1$cost
hwdata1$name[hwdata1$avg_10yr_salary]
hwdata1$best
best <- sort(hwdata1$best,decreasing=T)[1:10]

hwdata1$best[hwdata1$name == "Auburn University at Montgomery"]
hwdata1$name[best]
hwdata1$ownership[hwdata1$name %in% c("United States Merchant Marine Academy", "Aug
usta University", "South Texas College", "University of Connecticut-Avery Point", "
University of Connecticut-Stamford", "University of Connecticut-Tri-Campus", "India
n River State College", "Palm Beach State College")]
idealDiv <- 1/7

hwdata1$diversity <- abs((hwdata1$race_white - idealDiv)/idealDiv) + abs((hwdata1$
race_black - idealDiv)/idealDiv) + abs((hwdata1$race_hispanic - idealDiv)/idealDiv)
+ abs((hwdata1$race_asian - idealDiv)/idealDiv) + abs((hwdata1$race_native - ideal
Div)/idealDiv) + abs((hwdata1$race_pacific - idealDiv)/idealDiv) + abs((hwdata1$rac
e_other - idealDiv)/idealDiv)
top10scores <- sort(hwdata1$diversity,decreasing=F)[1:10]
mostDivColleges <- hwdata1$name[hwdata1$diversity %in% top10scores]
bot10scores <- sort(hwdata1$diversity,decreasing=T)[1]
botDivColleges <- hwdata1$name[hwdata1$diversity %in% bot10scores]
mostDivColleges
top10scores
hwdata1$tuition[hwdata1$name %in% "University of California-Davis"]
boxplot(hwdata1$tuition,outline=FALSE, main = "Tuitions in colleges nationally and
```

```
that of UC Davis", ylab = "Tuition in USD")
points(hwdata1$tuition[hwdata1$name %in% "University of California-Davis"])
boxplot(hwdata1$tuition[hwdata1$ownership %in% "Public"],outline=FALSE, main = "Tu
itions in public colleges nationally and that of UC Davis", ylab = "Tuition in USD"
)
points(hwdata1$tuition[hwdata1$name %in% "University of California-Davis"])

boxplot(hwdata1$first_gen,outline=FALSE, main="Proportion of first generations in c
olleges nationally and that of UC Davis",ylab="Proportion of first gen")
points(hwdata1$first_gen[hwdata1$name %in% "University of California-Davis"])

boxplot(hwdata1$race_black,outline=FALSE, main="Proportion of blacks in colleges na
tionally and that of UC Davis", ylab = "Proportion of blacks")
points(hwdata1$race_black[hwdata1$name %in% "University of California-Davis"])
```