STATS415 hw1 Cai,Yunguo 38349078

1.(a) Categorical Variable: a binary variable whether the student is legal to drink.
Ordinal Variable: the year the student is in (freshman, sophomore, junior, senior, ..).
Interval Variable: the birthday of the student.
Ratio Variable: the age of the student.

(b) The students who took STATS415 before.

(c) All the students in University of Michigan now.

2.(a) The effect of this transformation is: If a term occurs in one document, it has maximum weight log(n) since $g_j = 1$. If a term occurs in every document, it has weight 0 since $g_j = n$ and $\log(\frac{n}{n}) = 0$.

(b) The purpose of this transformation might be normalization to reflect the observation that the terms occur in every document can't be used to distinguish one document from another, while the fewer times a term occurs in documents, the more importance it has in distinguishing documents.

3. **Read the data into R, call it and make sure that it's in the right directory and form.**

```
> setwd("/Users/cyunguo/Desktop/2018WN/STATS415/hw/hw1")
> college <- read.csv("College.csv")
> rownames(college)=college[,1]
> college=college[,-1]
> dim(college)
[1] 777  18
> head(college)
                             Private Apps Accept Enroll Top10perc Top25perc
Abilene Christian University     Yes 1660   1232    721        23        52
Adelphi University               Yes 2186   1924    512        16        29
Adrian College                   Yes 1428   1097    336        22        50
Agnes Scott College              Yes  417    349    137        60        89
Alaska Pacific University        Yes  193    146     55        16        44
Albertson College                Yes  587    479    158        38        62
                             F.Undergrad P.Undergrad Outstate Room.Board Books
Abilene Christian University        2885         537     7440       3300   450
Adelphi University                  2683        1227    12280       6450   750
Adrian College                      1036          99    11250       3750   400
Agnes Scott College                  510          63    12960       5450   450
Alaska Pacific University            249         869     7560       4120   800
Albertson College                    678          41    13500       3335   500
                             Personal PhD Terminal S.F.Ratio perc.alumni
Abilene Christian University     2200  70       78      18.1          12
Adelphi University               1500  29       30      12.2          16
Adrian College                   1165  53       66      12.9          30
Agnes Scott College               875  92       97       7.7          37
Alaska Pacific University        1500  76       72      11.9           2
Albertson College                 675  67       73       9.4          11
                             Expend Grad.Rate
Abilene Christian University   7041        60
Adelphi University            10527        56
Adrian College                 8735        54
Agnes Scott College           19016        59
Alaska Pacific University     10922        15
Albertson College              9727        55
```

Numeric summaries for each variable.

```
> college$Private=as.factor(college$Private)
> summary(college)
 Private        Apps           Accept          Enroll       Top10perc
 No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
 Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
           Median : 1558   Median : 1110   Median : 434   Median :23.00
           Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
           3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
           Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
   Top25perc      F.Undergrad     P.Undergrad         Outstate
 Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
 1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
 Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
 Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
 3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
 Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
   Room.Board       Books           Personal         PhD
 Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
 1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
 Median :4200   Median : 500.0   Median :1200   Median : 75.00
 Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
 3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
 Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
    Terminal       S.F.Ratio      perc.alumni        Expend
 Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
 1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
 Median : 82.0   Median :13.60   Median :21.00   Median : 8377
 Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
 3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
 Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
   Grad.Rate
 Min.   : 10.00
 1st Qu.: 53.00
 Median : 65.00
 Mean   : 65.46
 3rd Qu.: 78.00
 Max.   :118.00
```

Multivariate numerical summaries

A correlation matrix of the variables Apps, Accept and Enroll is built.

```
> corrs <- round(cor(college[,2:4]),4)
> corrs
         Apps Accept Enroll
Apps   1.0000 0.9435 0.8468
Accept 0.9435 1.0000 0.9116
Enroll 0.8468 0.9116 1.0000
```

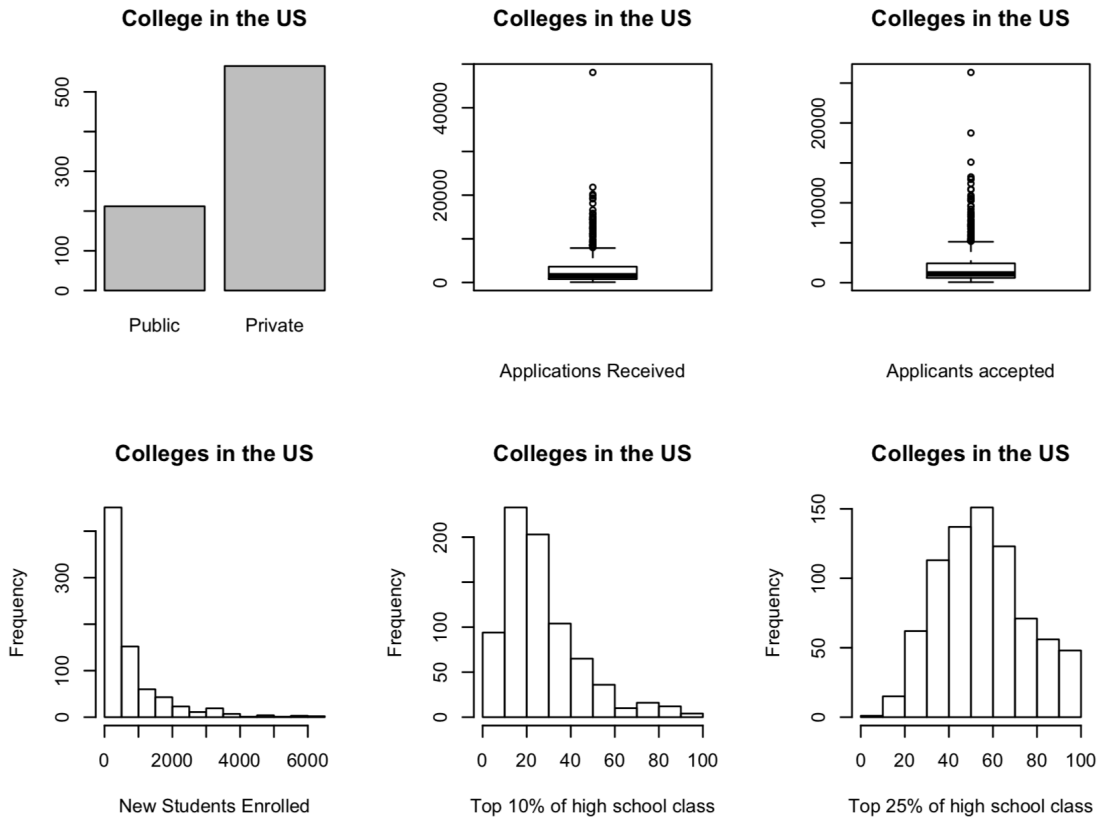
cor(Apps, Accept) = 0.9435, which shows that the number of applications received has a strong linear relation with the number of applications accepted since 0.9435 is close to 1.

cor(Accept, Enroll) = 0.9116, which shows that the number of students enrolled has a strong linear relation with the number of new students enrolled since 0.9116 is close to 1.

cor(Apps, Enroll) = 0.8468, which shows that it has weaker relation than cor(Apps, Accept) and cor(Accept, Enroll). It might be caused by the fact that the applications accepted by college include original students of the college except the new students enrolled.

Graphical summaries for each variable

```
> par(mfrow=c(2,3))
> barplot(table(college$Private),main="College in the
US",names.arg=c("Public","Private"),ylim=c(0,570))
> boxplot(college$Apps, main="Colleges in the US",xlab="Applications Received")
> boxplot(college$Accept, main="Colleges in the US",xlab="Applicants accepted")
> hist(college$Enroll, main="Colleges in the US",xlab="New Students Enrolled")
> hist(college$Top10perc, main="Colleges in the US",xlab="Top 10% of high school
class")
> hist(college$Top1p25erc, main="Colleges in the US",xlab="Top 25% of high school
class")
> hist(college$Top25perc, main="Colleges in the US",xlab="Top 25% of high school
class")
```



The six plots above show the recruitment statistics of the colleges in the US.

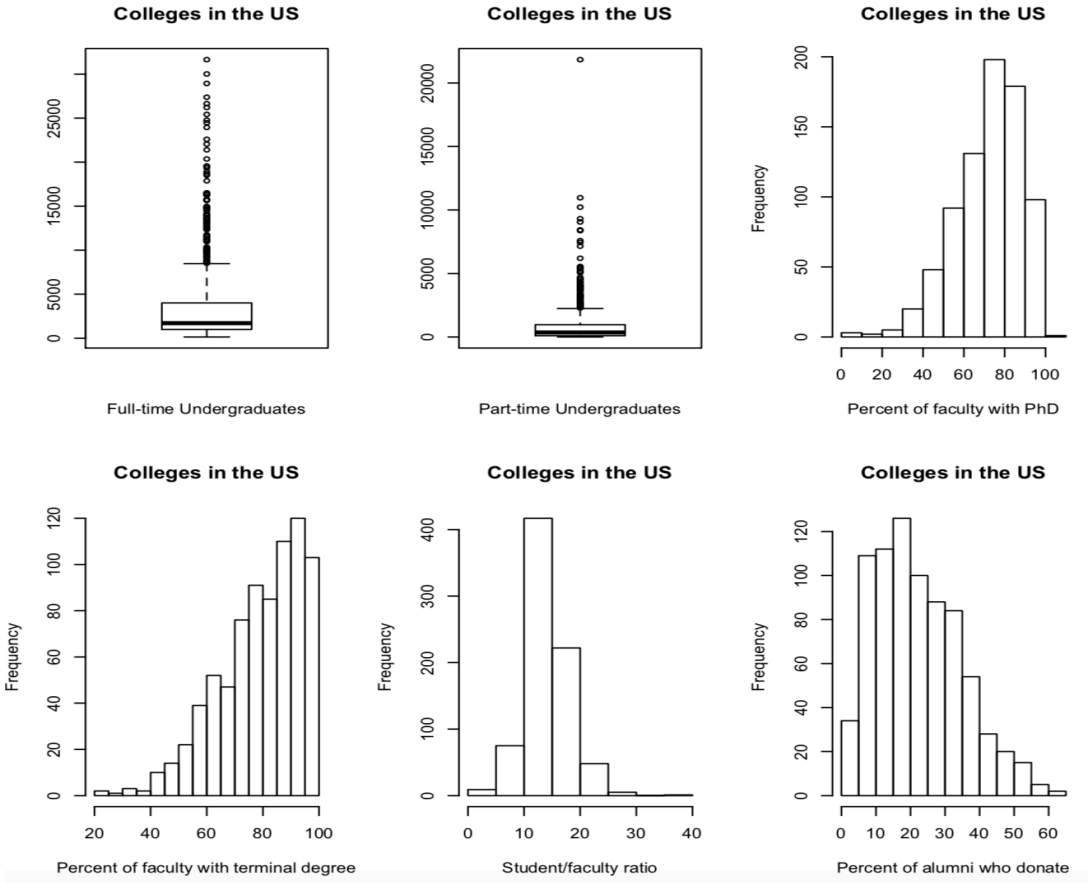
In terms of quantity, some colleges are more popular and contribute larger capacity for students. The bar plot shows that the number of private colleges is more than twice as much as the public colleges, indicating that private college dominates in college education. Boxplots of applications received, applicants accepted and histogram of new students enrolled all show a non-normal distribution. There exist many outliers in the boxplots, which indicates that the number of applications received by different colleges and the number of students accepted by different colleges vary a lot. The numerical statistics summary tells us the mean of applications is 3002 but the median is only 1558, while the maximum is 48094, which confirms that the standard deviation of the variable-application received is large, namely couples of colleges receive far more applications than the rest. The reason for this might be the comprehensive strength of these colleges attract more students or the scale of the campus with more majors can accept more students. The variable applicants accepted also shows the big deviation between colleges. The mean of applications is 2019 but the median is only 1110, while the maximum is 26330. The histogram of the number of new students enrolled skewed to the right significantly also reflects this imbalance.

In terms of quality, the situation that minority of the colleges get the majority of the most top students. The histogram of the number of new students from top 10% of high school class is right-skewed with skewness at value of 1.407765, which is quite large. Only 78 out of 777 colleges get more than 50 top 10% students from high school class. In comparison, the histogram of the number of new students from top 25% of high school class is almost symmetric with skewness at value of 0.25834, which is quite close to 0. 449 out of 777 (more than half) get more than 50 top 25% students from high school class. This indicates the distribution of college education resources for students is generally balanced.

```
>  sum(college$Top10perc >50)
[1] 78
> sum(college$Top25perc >50)
[1] 449
```

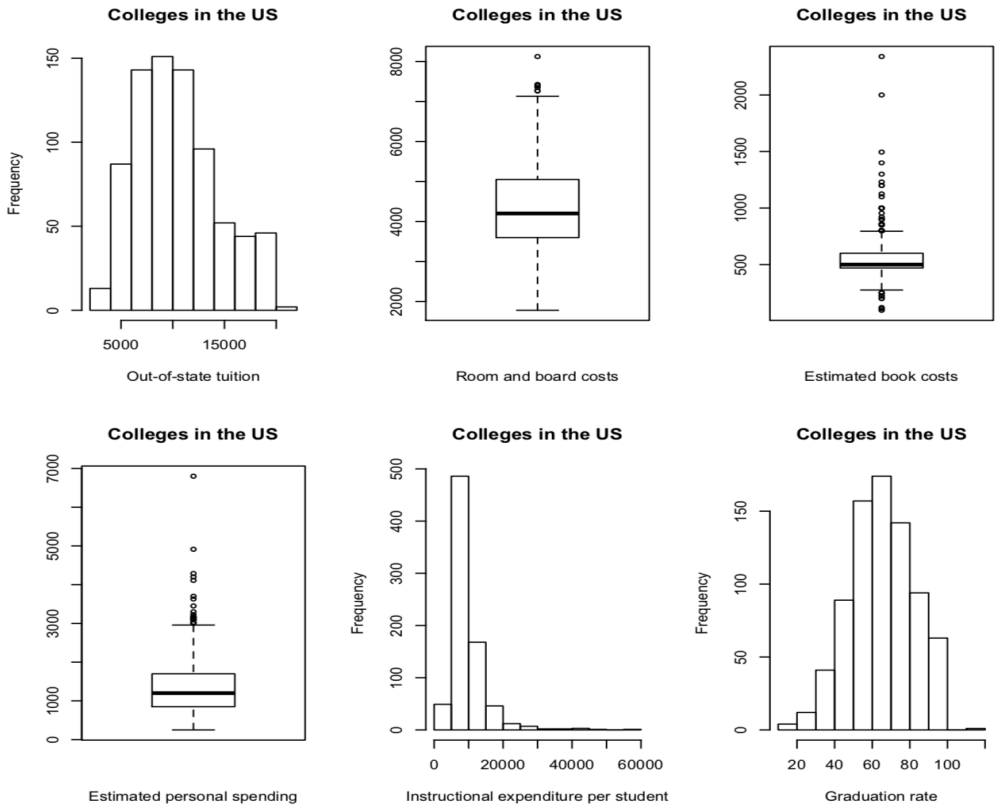
```
> skewness(college$Top25perc)
[1] 0.2583399
attr(,"method")
[1] "moment"
> skewness(college$Top10perc)
[1] 1.407765
attr(,"method")
[1] "moment"
```

```
> par(mfrow=c(2,3))
> boxplot(college$F.Undergrad,main="Colleges in the US",xlab="Full-time
Undergraduates")
> boxplot(college$P.Undergrad,main="Colleges in the US",xlab="Part-time
Undergraduates")
> hist(college$PhD,main="Colleges in the US",xlab="Percent of faculty with PhD")
> hist(college$Terminal,main="Colleges in the US",xlab="Percent of faculty with
terminal degree")
> hist(college$S.F.Ratio,main="Colleges in the US",xlab="Student/faculty ratio")
> hist(college$perc.alumni,main="Colleges in the US",xlab="Percent of alumni who
donate")
```



The outliers in full-time undergraduates and part-time undergraduates can be explained with the same reason as the number of applications, applicants accepted and students enrolled because they are all variables related to capacity. In terms of faculty, the histogram of percent of faculty with PhD and percent of faculty with terminal degree are significantly left-skewed while percent of alumni who donate are right-skewed. This shows that the general degree level of faculty is high and the outstanding alumni of each college is not so much. The mean of percent of alumni who donate is 22.74% and the median is 21%. The distinguished alumni are 10%-40% percent for most of the colleges and only a few colleges can get more than 50% of alumni who donate. This is also related to the comprehensive strength and capacity of colleges.

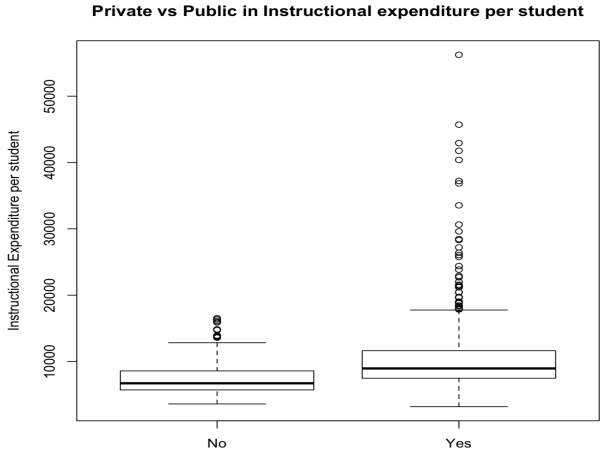
```
> par(mfrow=c(2,3))
> hist(college$Outstate,main="Colleges in the US",xlab="Out-of-state tuition")
> boxplot(college$Room.Board,main="Colleges in the US",xlab="Room and board costs")
> boxplot(college$Books,main="Colleges in the US",xlab="Estimated book costs")
> boxplot(college$Personal,main="Colleges in the US",xlab="Estimated personal
spending")
> hist(college$Expend,main="Colleges in the US",xlab="Instructional expenditure per
student")
> hist(college$Grad.Rate,main="Colleges in the US",xlab="Graduation rate")
```



The plot shows that out-state tuition mostly cumulates in the range of 5000 to 15000. Estimated book and personal spending have a lot of outliers, which shows that the textbooks of different colleges vary a lot and the free book resources that colleges can provide for students distinct. The personal spending varies from person to person. The room and board costs are relatively stable, which indicates the general living conditions of college students. It's interesting about the instructional expenditure per student. The mean is 9660 and the median is 8337, while the maximum is 56233. The maximum seems unbelievable because it's far more than the tuition fee and the result of such a high expenditure is worthy of researching. It's also ridiculous that in the graduation rate plot, there's a college that its graduation rate is over 100%. This might be an error or caused by the fact that some students postpone their graduation and thus make the number of graduation students greater than the estimated graduation students in the corresponding year. And the colleges with graduation rate lower than 30% is such a low rate.

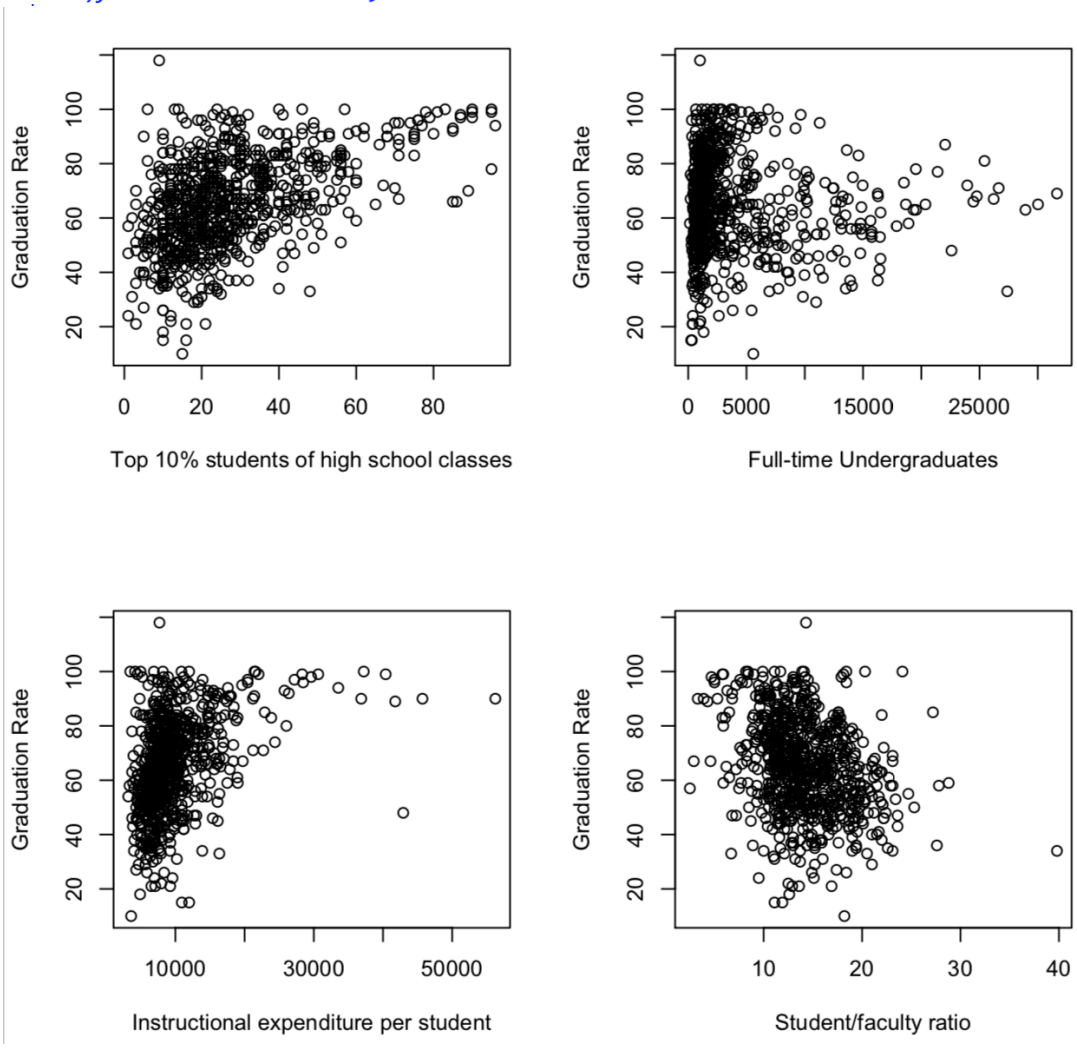
Multivariate Graphical Summaries

```
> plot(college$Private,college$Expend,main="Private vs Public in Instructional
expenditure per student",xlab="Private or not",ylab="Instructional Expenditure per
student")
```



From the side-by-side boxplot, we can see that most of private colleges have higher instructional expenditure per student than public colleges. This is mainly caused by the higher tuition of private colleges.

```
> par(mfrow=c(2,2))
> plot(college$Top10perc,college$Grad.Rate,xlab="Top 10% students of high school
classes",ylab="Graduation Rate")
> plot(college$F.Undergrad,college$Grad.Rate,xlab="Full-time
Undergraduates",ylab="Graduation Rate")
> plot(college$Expend,college$Grad.Rate,xlab="Instructional expenditure per
student",ylab="Graduation Rate")
> plot(college$S.F.Ratio,college$Grad.Rate,xlab="Student/faculty
ratio",ylab="Graduation Rate")
```



The scatter plots show how the number of top 10% students of high school class, the number of full-time undergraduates, the instructional expenditure per student and student/faculty ratio impact the graduation rate. There seems a positive relation between the student quality and graduation rate, but when the number of top 10% new students is smaller than 50, the influence is not obvious. The number of full-time graduates and instructional expenditure per student also reveals a positive relation and when the number of full-time graduates is smaller than 15000 and instructional expenditure is less than 20000, the influence is not obvious. The reason for this might be the student number smaller than 15000 and the instructional expenditure less than 20000 is not enough to show a distinction for education achievements. However, though the graduation rate is estimated to have a negative relationship with the student/faculty ratio and somehow is shown in the plot, the colleges with same student/faculty ratio can have graduation rate ranging from 20% to 100%, which indicates that the student/faculty ratio is not an important factor for graduation rate.