**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.

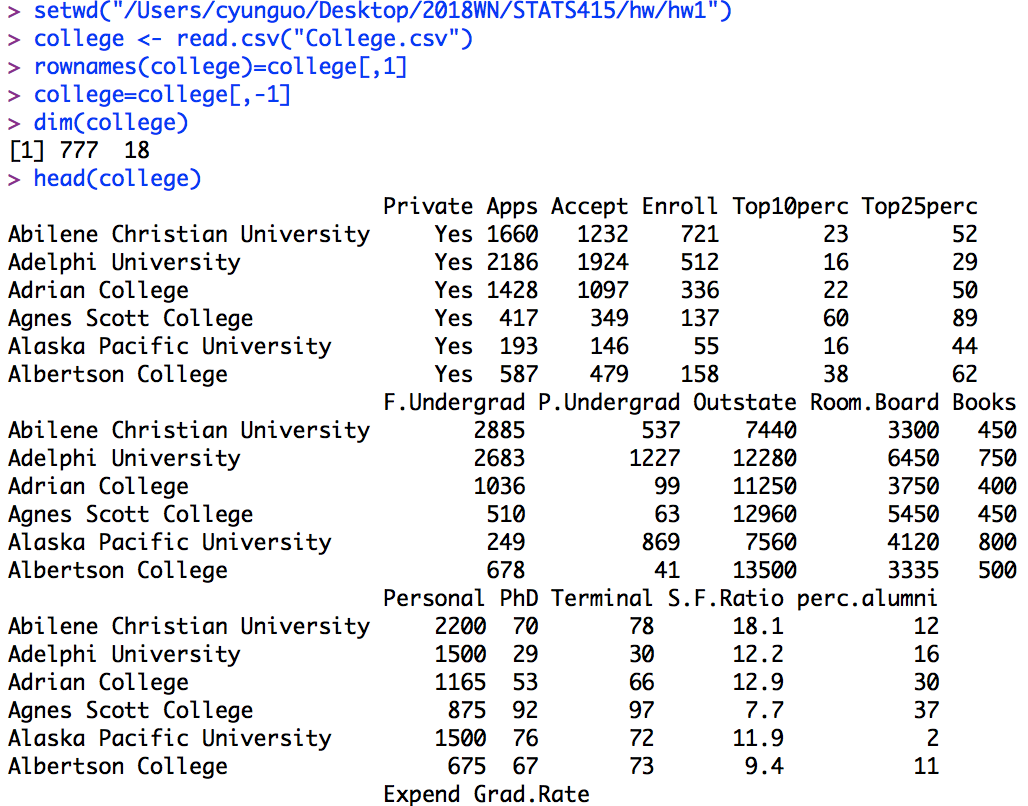
(b) The students who took STATS415 before.

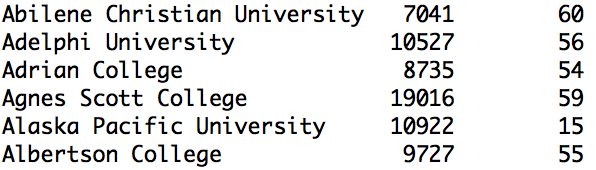
(c) All the students in University of Michigan now.

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

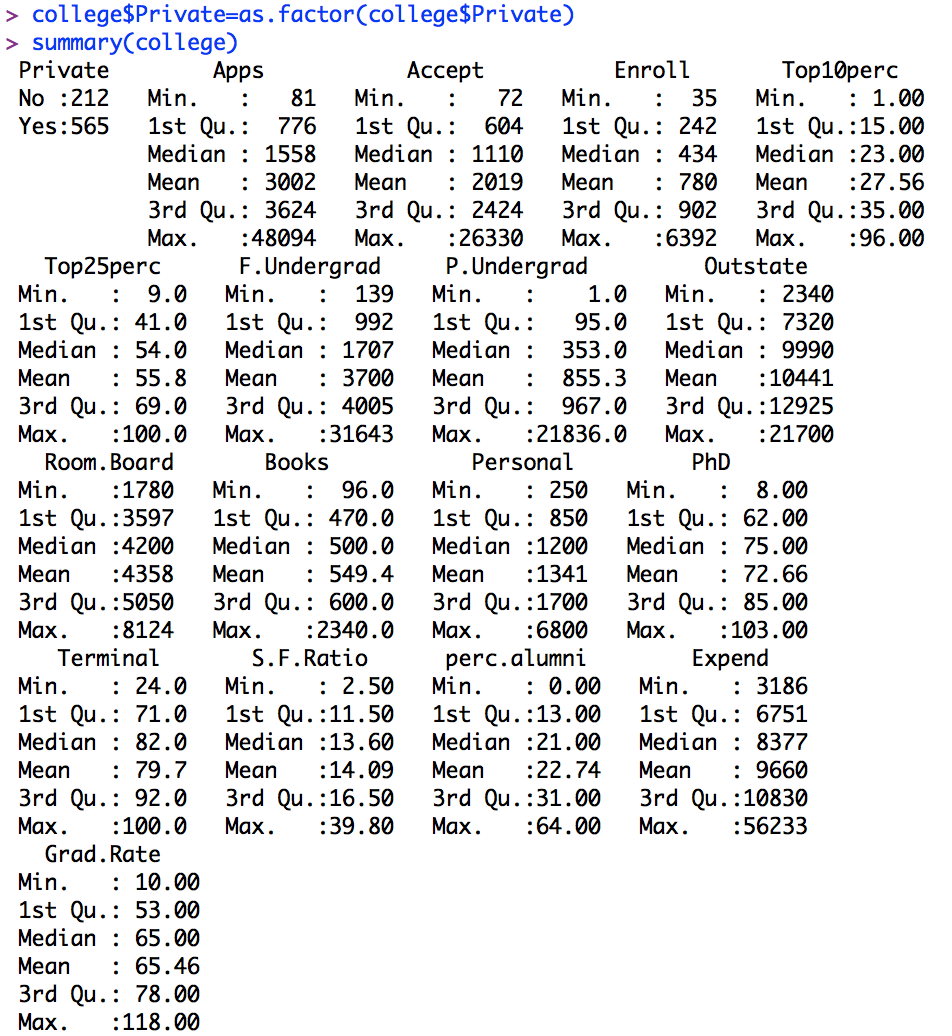
(b) The purpose of this transformation might be normalization to reflect the observation that the terms occur in every document can’t be used to distinguish one document from another, while the fewer times a term occurs in documents, the more importance it has in distinguishing documents.

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



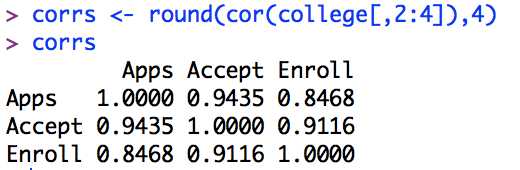


**Numeric summaries for each variable.**



**Multivariate numerical summaries**

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

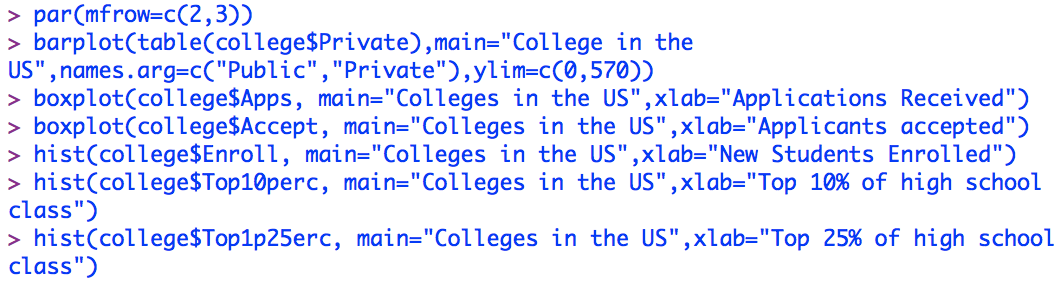


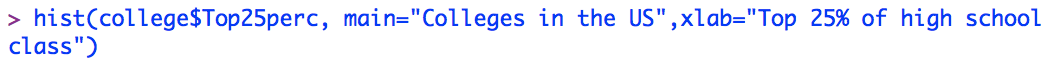
cor(Apps, Accept) = 0.9435, which shows that the number of applications received has a strong linear relation with the number of applications accepted since 0.9435 is close to 1.

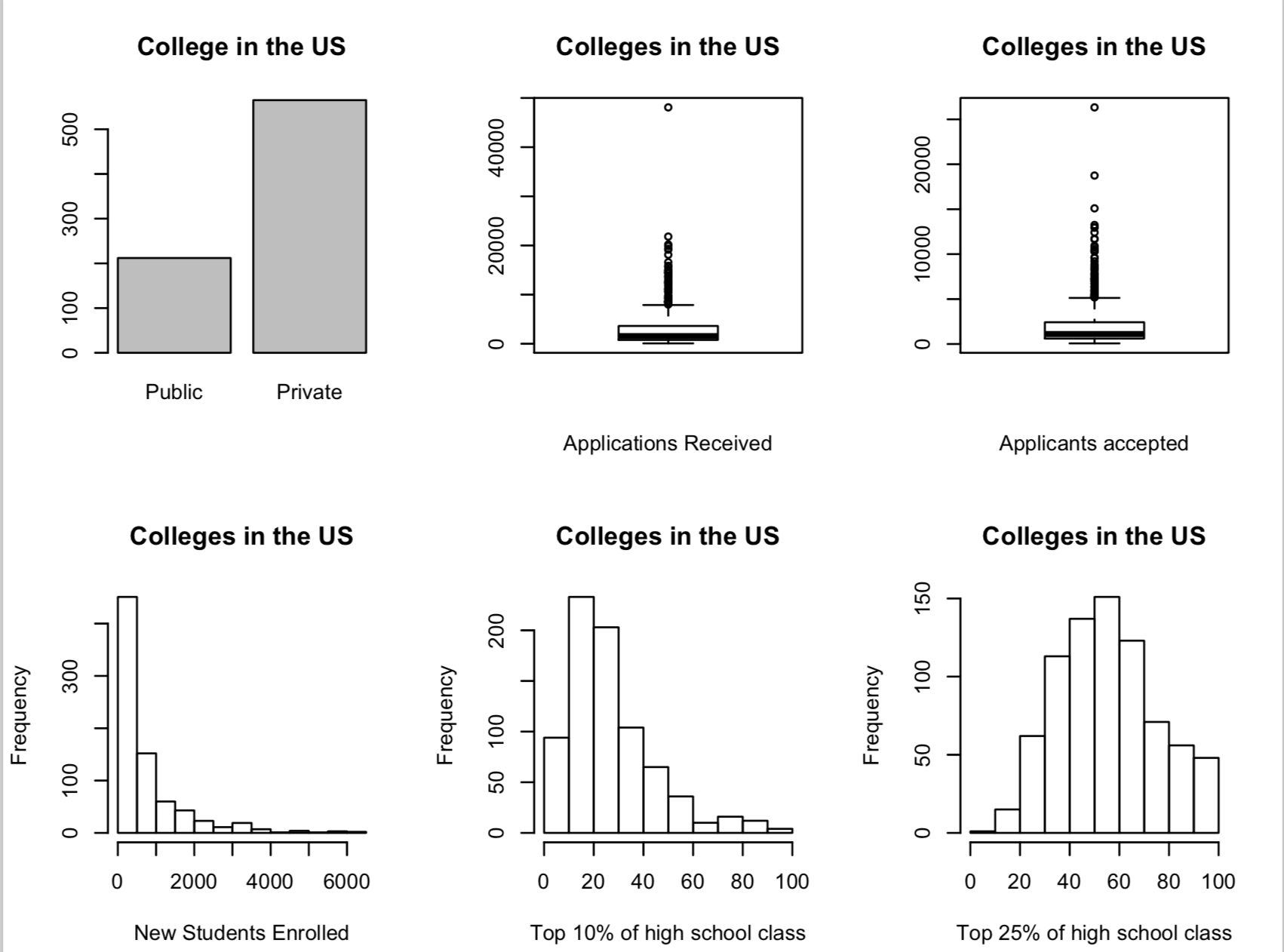
cor(Accept, Enroll) = 0.9116, which shows that the number of students enrolled has a strong linear relation with the number of new students enrolled since 0.9116 is close to 1.

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**





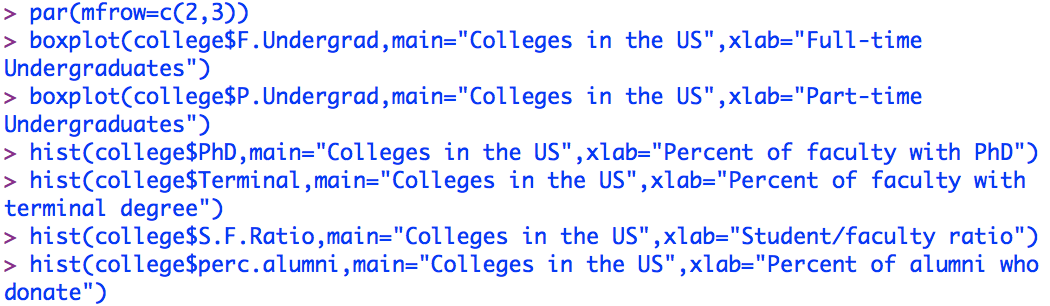


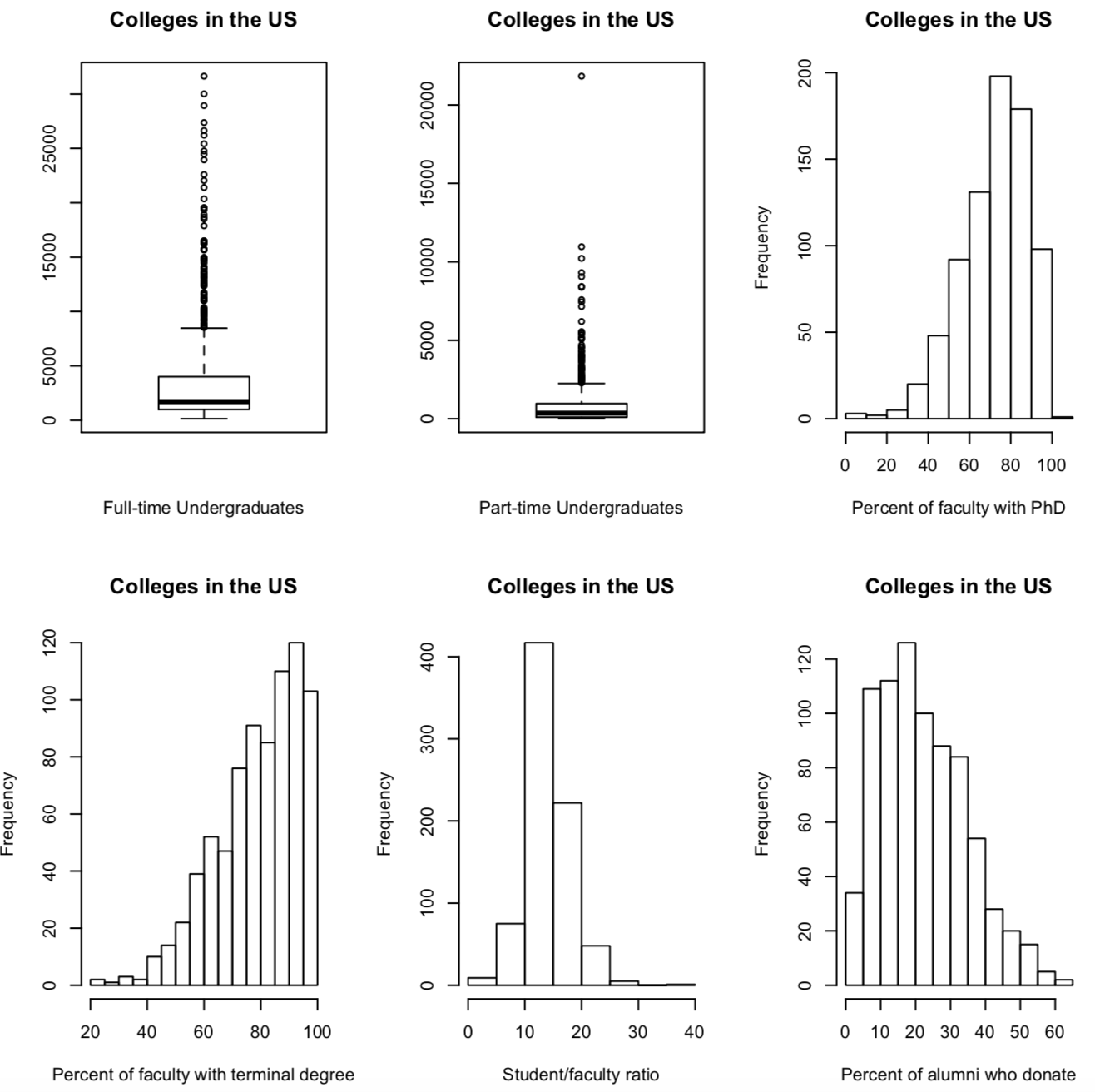
The six plots above show the recruitment statistics of the colleges in the US.

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.

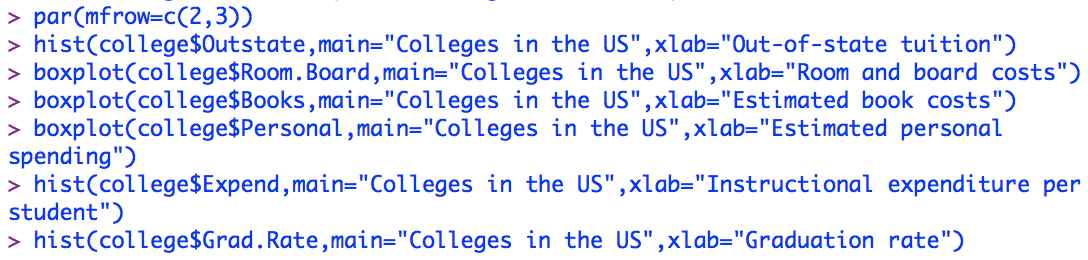
In terms of quality, the situation that minority of the colleges get the majority of the most top students. The histogram of the number of new students from top 10% of high school class is right-skewed with skewness at value of 1.407765, which is quite large. Only 78 out of 777 colleges get more than 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.

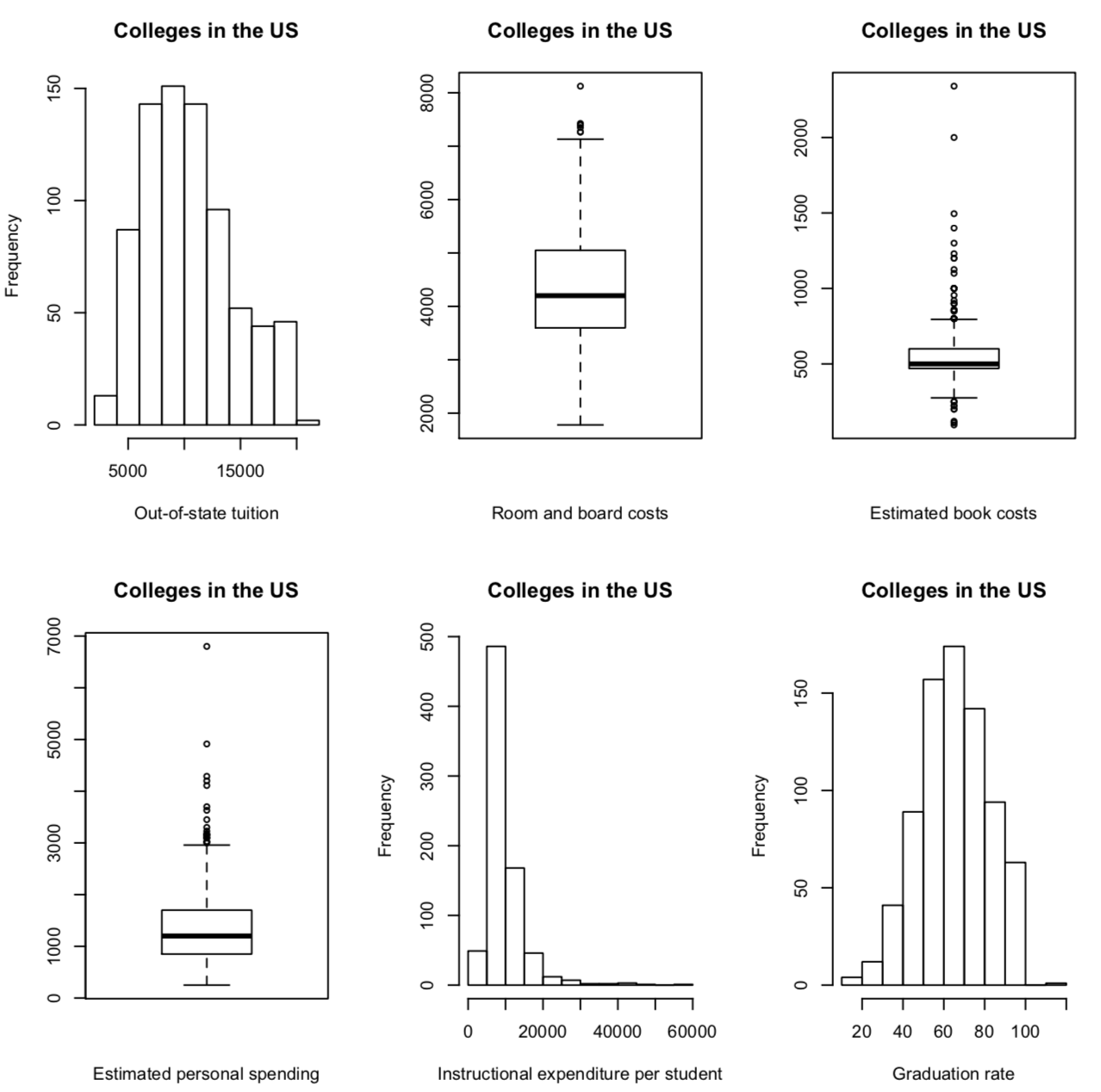
|  |  |
| --- | --- |
|  |  |





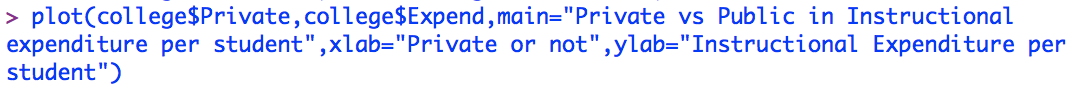
The outliers in full-time undergraduates and part-time undergraduates can be explained with the same reason as the number of applications, applicants accepted and students enrolled because they are all variables related to capacity. In terms of faculty, the histogram of percent of faculty with PhD and 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.

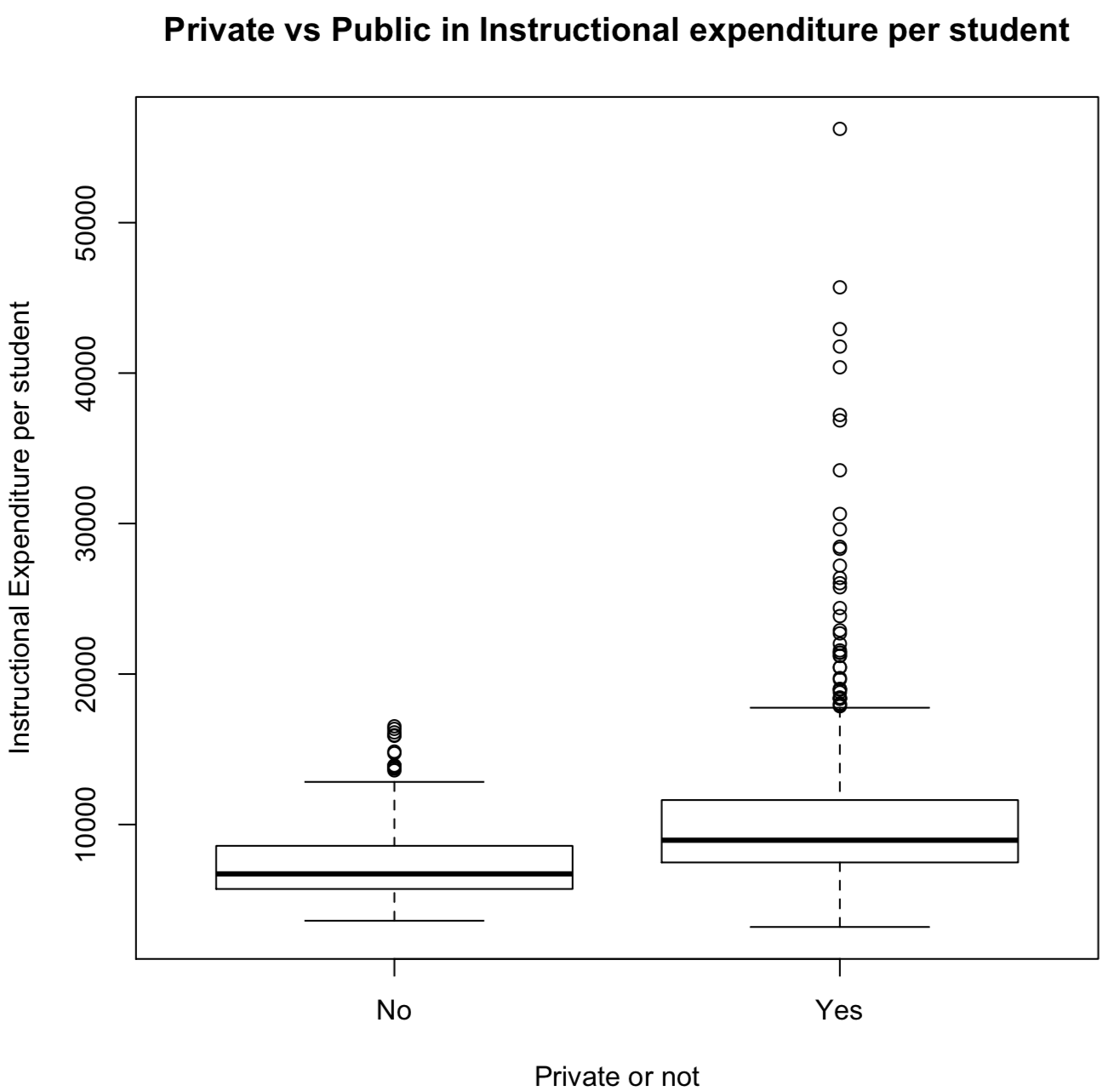




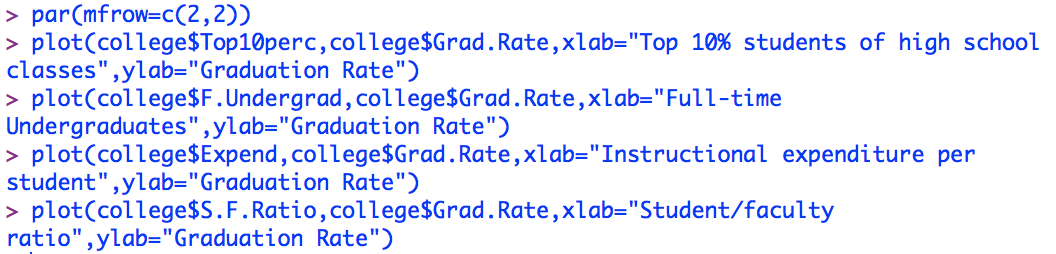
The plot shows that out-state tuition mostly cumulates in the range of 5000 to 15000. Estimated book and personal spending have a lot of outliers, which shows that the textbooks of different colleges vary a lot and the free book resources that colleges can provide for students distinct. The personal 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.

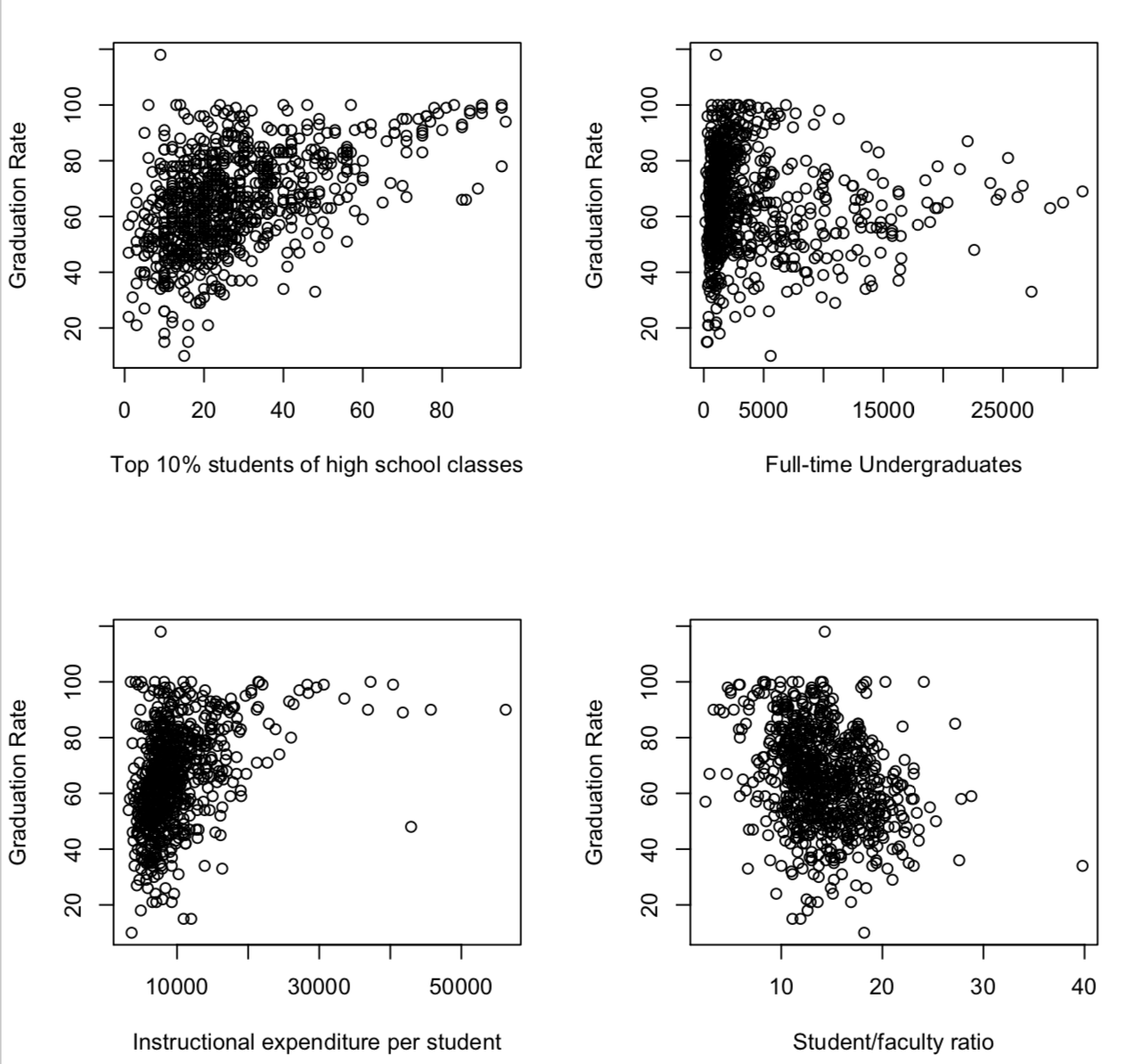
**Multivariate Graphical Summaries**





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.





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.