Chapter 5: Exploratory Data Analysis

Alex Chi

Exercises

5.1 (Exploring percentages of small classes).

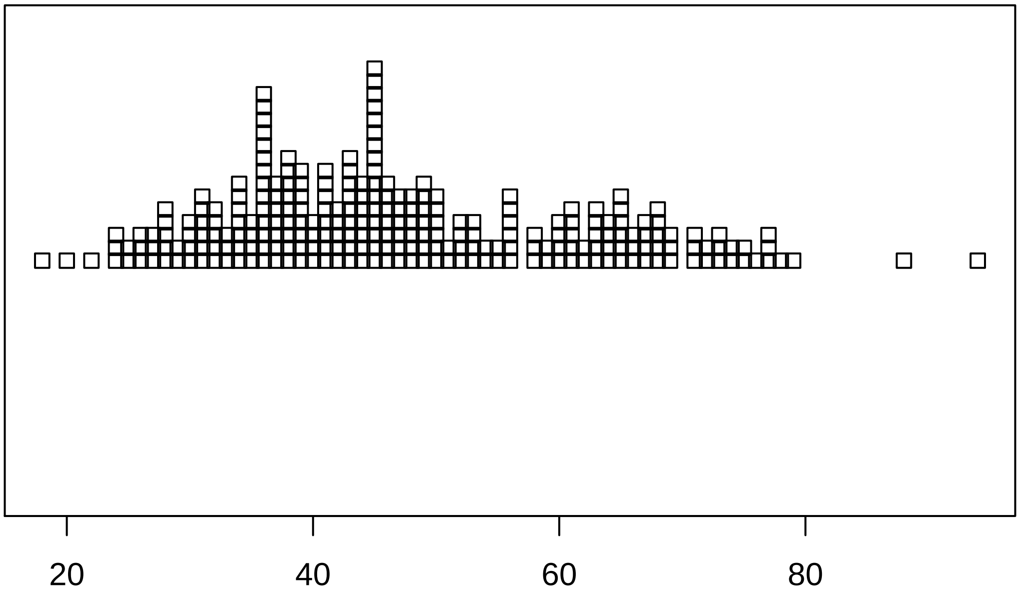
The variable Pct.20 in the college dataset contains the percentage of “small classes” (defined as 20 or fewer students) in the National Universities.

a. Construct a dotplot of the small-class percentages using the stripchart function. To see the density of points, it is helpful to use either the method=stack or method=jitter arguments. What is the shape of this data?

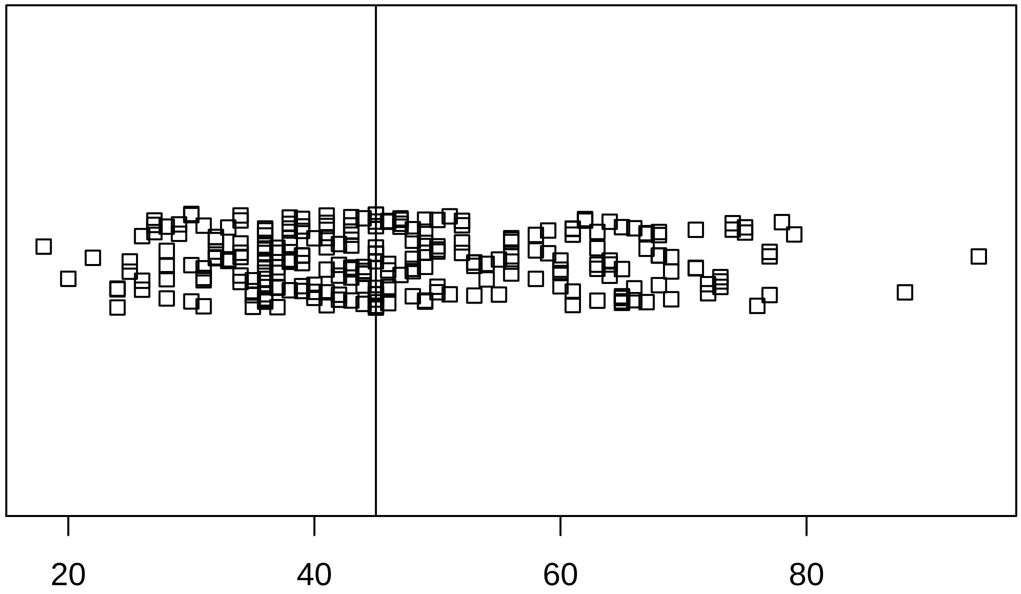
b. There is a single school with an unusually large small-class percentage. Use the identify function to find the name of this unusual school.

c. Find the median small-class percentage and draw a vertical line (using the abline function) on the dotplot at the location of the median.

```
college = read.table("Rx-data/college.txt",
                     header=TRUE, sep="\t")
stripchart(college$Pct.20, method="stack")
```



```
stripchart(college$Pct.20, method="jitter")
abline(v=median(college$Pct.20, na.rm=TRUE))
```



```r
college[order(college$Pct.20,decreasing=TRUE),][1:2,]
```

```
##           School Rank Tier Retention Grad.rate Pct.20 Pct.50 Full.time
## 210 Golden Gate   NA    4        NA        20     94      0        NA
## 151  New School   NA    3        81        61     88      1        37
##     Top.10 Accept.rate Alumni.giving
## 210     NA         100            NA
## 151     18          51             8
```

5.2 (Relationship between the percentages of small classes and large classes).

The variables Pct.20 and Pct.50 in the college dataset contain respectively the percentage of“small classes”(defined as 20 or fewer students) and the percentage of “large classes” (defined as 50 or more students) in the National Universities.

a. Use the plot function to construct a scatterplot of Pct.20 (horizontal) against Pct.50 (vertical).

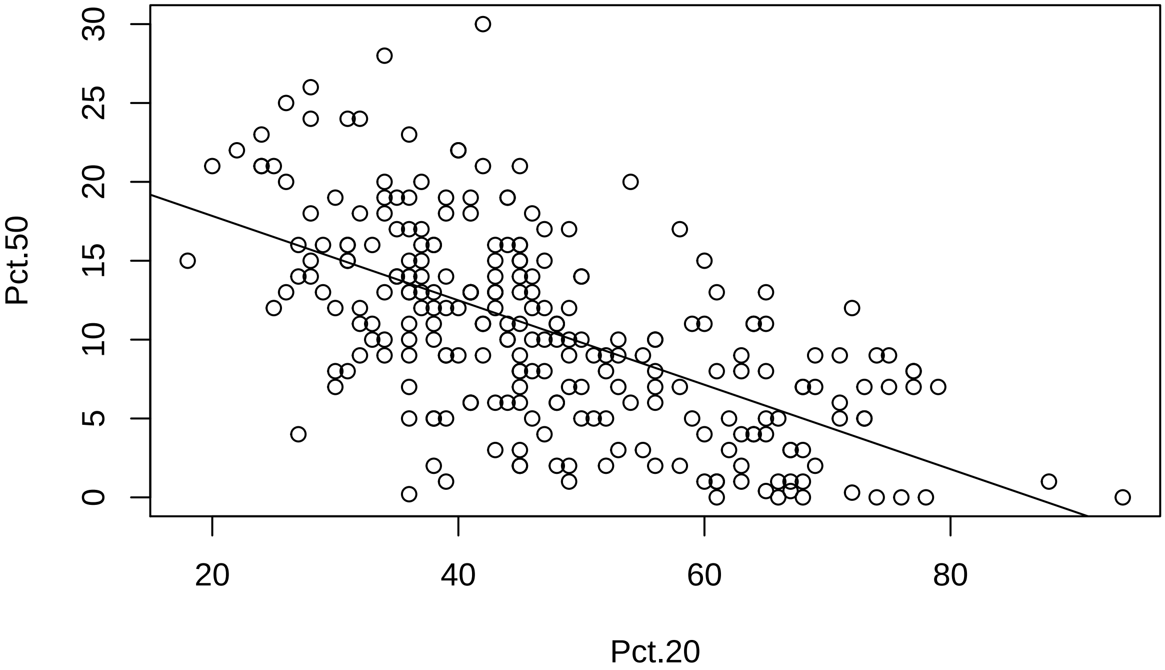
b. Use the line function to find a resistant line to these data. Add this resistant line to the scatterplot constructed in part a.

c. If 60% of the classes at a particular college have 20 or fewer students, use the fitted line to predict the percentage of classes that have 50 or more students.

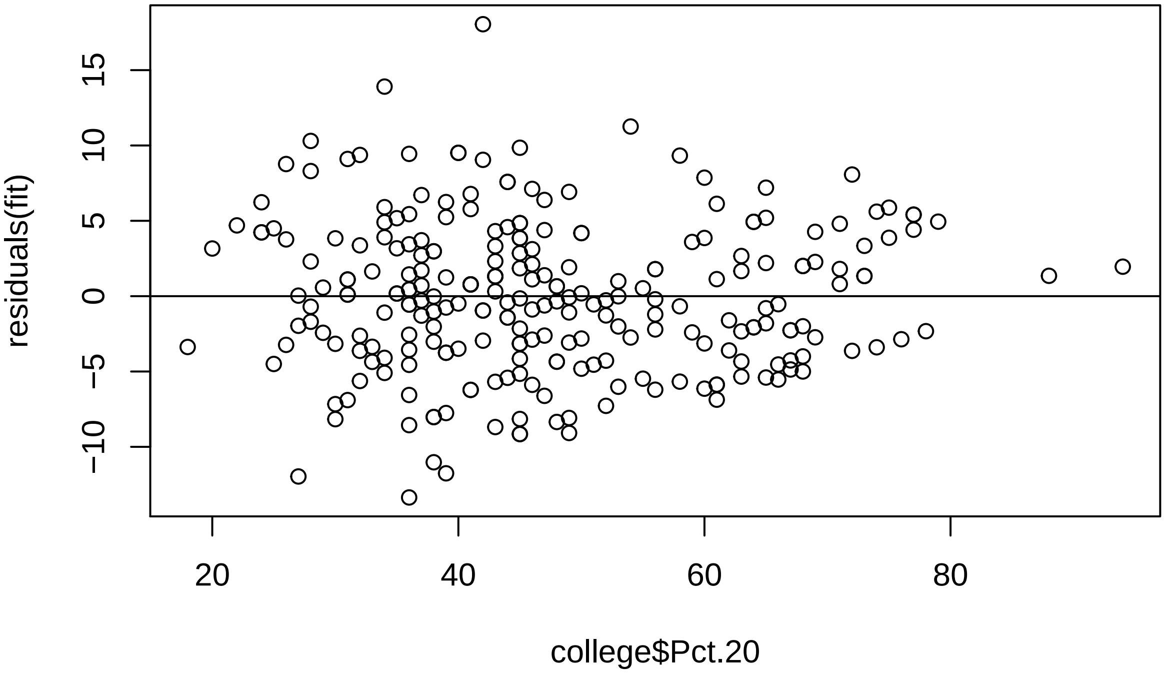
d. Construct a graph of the residuals (vertical) against Pct.20 (horizontal) and add a horizontal line at zero (using the abline function).

e. Is there a distinctive pattern to the residuals? (Compare the sizes of the residuals for small Pct.20 and the sizes of the residuals for large Pct.50.) f. Use the identify function to identify the schools that have residuals that exceed 10 in absolute value. Interpret these large residuals in the context of the problem.

```r
plot(Pct.50~Pct.20, college)
fit = lm(Pct.50 ~ Pct.20, college,
         na.action=na.exclude)
abline(fit)
```



```r
plot(college$Pct.20, residuals(fit))
abline(h=0)
```



```r
college[na.omit(abs(residuals(fit)) > 10), ]
```

```
##                    School Rank Tier Retention Grad.rate Pct.20 Pct.50
## 24                   UCLA   24    1        97        89     54     20
## 37  Univ Calif San Diego   35    1        94        85     42     30
## 43      Univ Calif Davis   42    1        90        81     34     28
## 112          New Hamsire  110    2        87        73     44     16
## 139        Bowling Green   NA    3        76        58     41      6
## 190        South Florida   NA    3        83        48     27     14
## 191           St. Thomas   NA    3        88        72     38      2
## 213        Indiana State   NA    4        67        43     50      5
##     Full.time Top.10 Accept.rate Alumni.giving
## 24         90     97          23            14
## 37         93    100          42             7
## 43         94     98          53            11
## 112        84     24          65            10
## 139        94     12          87             9
## 190        97     25          46            15
## 191        73     20          81            15
## 213        87     10          66            13
```

5.3 (Relationship between acceptance rate and “top-ten” percent- age).

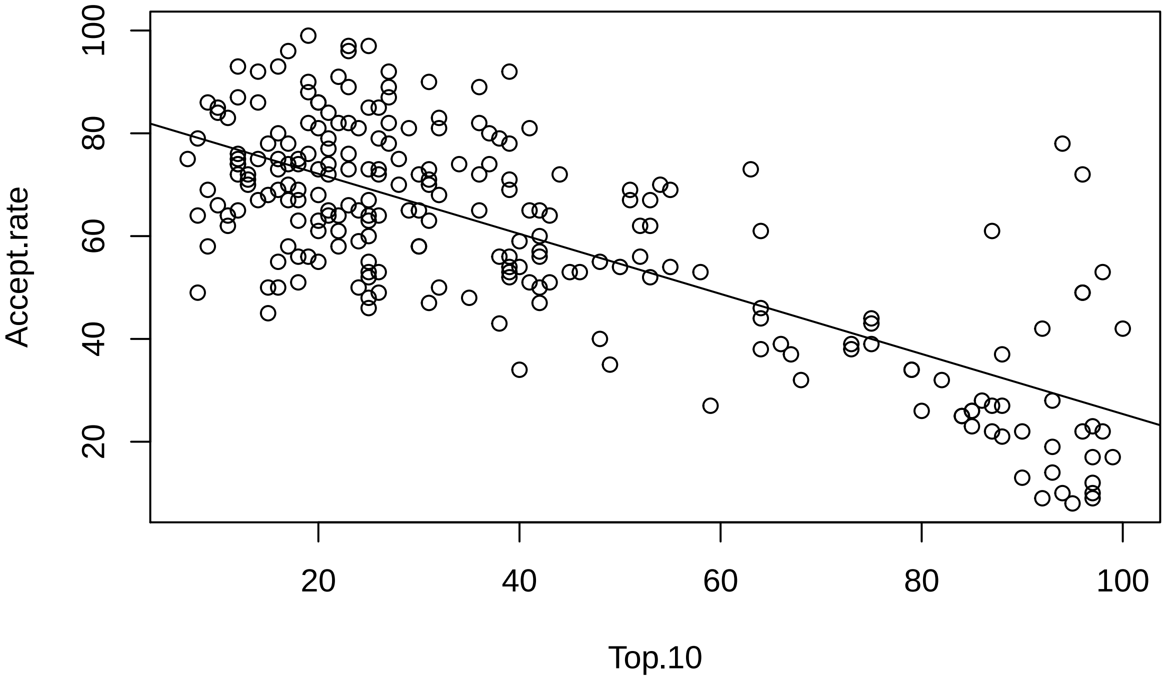
The variables Accept.rate and Top.10 in the college dataset contain respectively the acceptance rate and the percentage of incoming students in the top 10 percent of their high school class in the National Universities.

One would believe that these two variables are strongly associated, since, for example, “exclusive” colleges with small acceptance rates would be expected to have a large percentage of “top-ten” students.

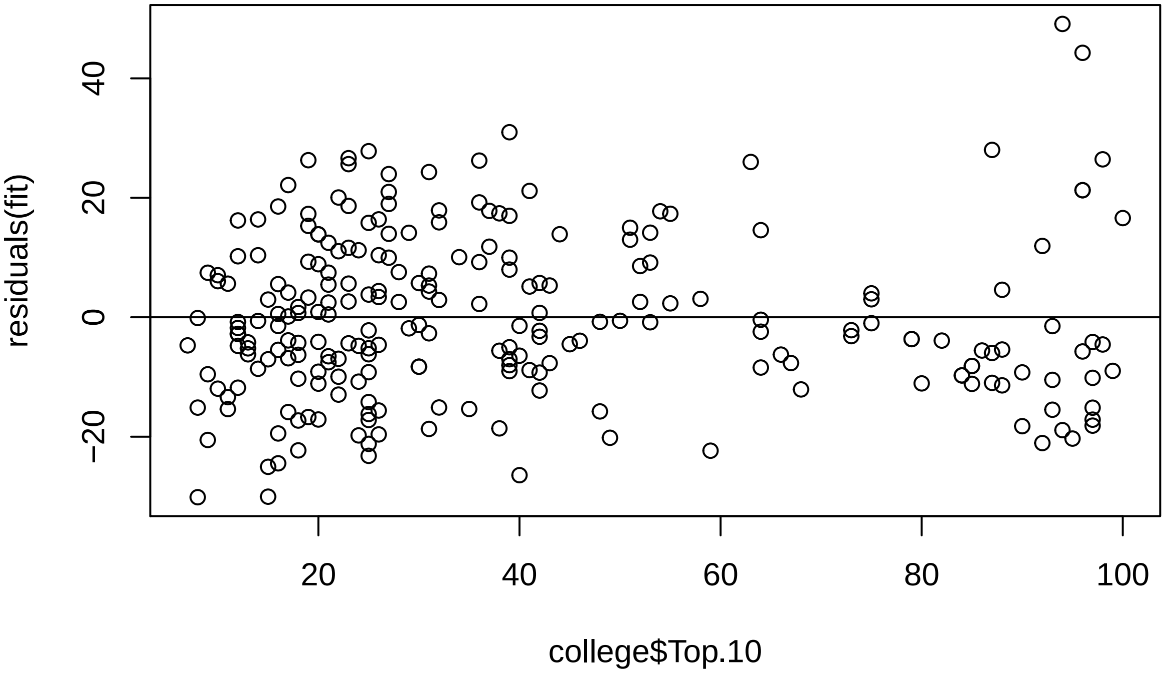
a. Explore the relationship between Accept.rate and Top.10. This explo- ration should include a graph and linear fit that describe the basic pattern in the relationship and a residual graph that shows how schools differ from the basic pattern.

b. Schools are often classified into “elite” and “non-elite” colleges depending on the type of students they admit. Based on your work in part a, is there any evidence from Accept.rate and Top.10 that schools do indeed cluster into “elite” and “non-elite” groups? Explain.

```r
plot(Accept.rate ~ Top.10, college)
fit = lm(Accept.rate ~ Top.10, college,
         na.action=na.exclude)
abline(fit)
```



```
plot(college$Top.10, residuals(fit))
abline(h=0)
```



5.4 (Exploring the pattern of college enrollment in the United States).

The U.S. National Center for Education Statistics lists the total enrollment at Institutions of Higher Education for years 1900-1985 at their website http://nces.ed.gov. Define the ordered pair (x,y), where y is the total enrollment in thousands in year x. Then we observe the data (1955, 2653), (1956, 2918), (1957, 3324), (1959, 3640), (1961, 4145), (1963, 4780), (1964, 5280), (1965, 5921), (1966, 6390), (1967, 6912), (1968, 7513), (1969, 8005), (1970, 8581).

a. Enter this data into R.

b. Use the lm function to fit a line to the pattern of enrollment growth in the period 1955 to 1970. By inspecting a graph of the residuals, decide if a line is a reasonable model of the change in enrollment.

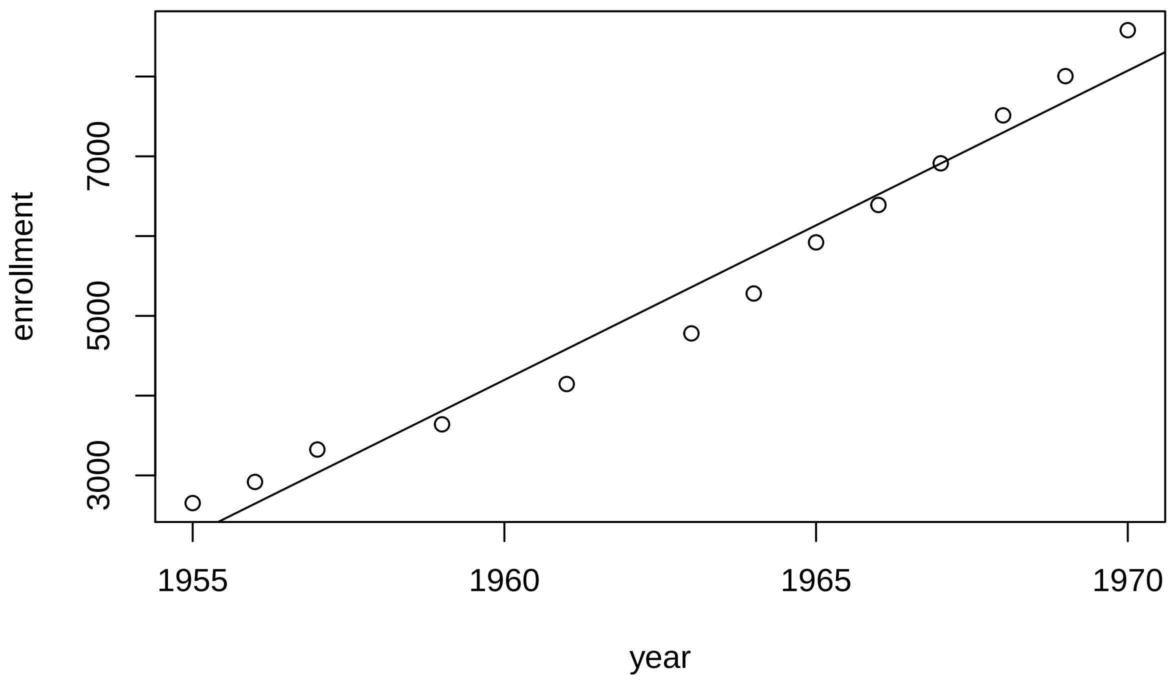
c. Transform the enrollment by a logarithm, and fit a line to the (year, log enrollment) data. Inspect the pattern of residuals and explain why a line is a better fit to the log enrollment data.

d. By interpreting the fit to the log enrollment data, explain how the college enrollment is changing in this time period. How does this growth compare to the growth of the BGSU enrollment in Section 5?

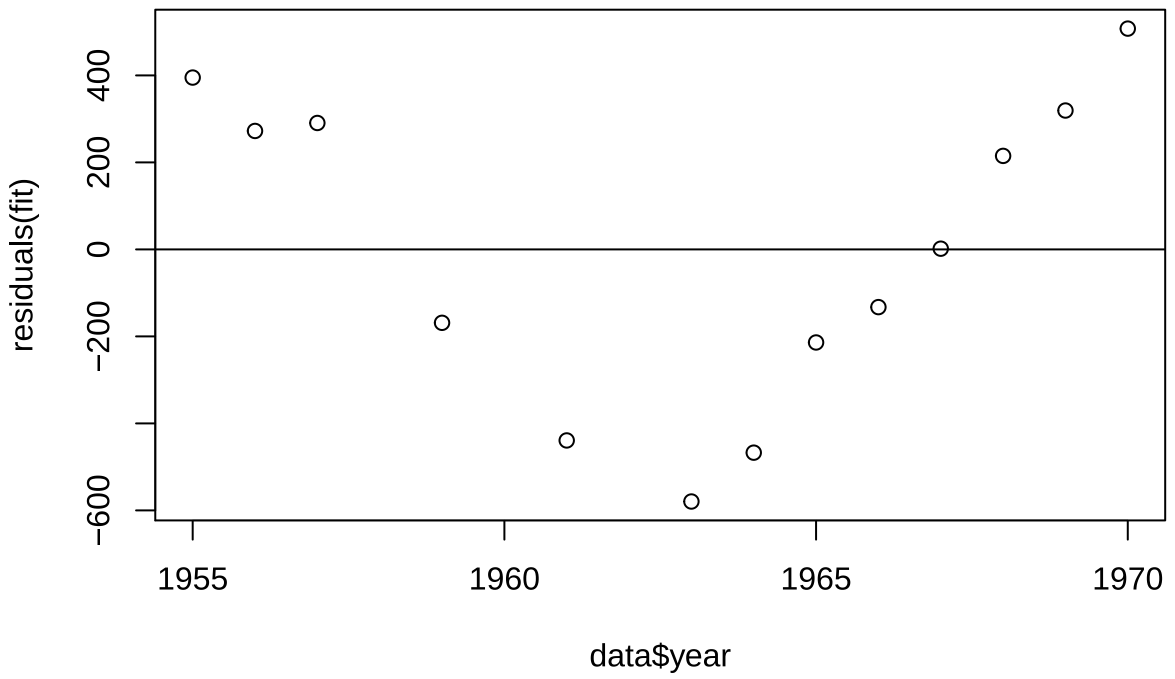
```
year =       c(1955, 1956, 1957, 1959, 1961, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970)
enrollment = c(2653, 2918, 3324, 3640, 4145, 4780, 5280, 5921, 6390, 6912, 7513, 8005, 8581)
data = data.frame(cbind(year, enrollment))
data
```

```
##    year enrollment
## 1  1955       2653
## 2  1956       2918
## 3  1957       3324
## 4  1959       3640
## 5  1961       4145
## 6  1963       4780
## 7  1964       5280
## 8  1965       5921
## 9  1966       6390
## 10 1967       6912
## 11 1968       7513
## 12 1969       8005
## 13 1970       8581
```

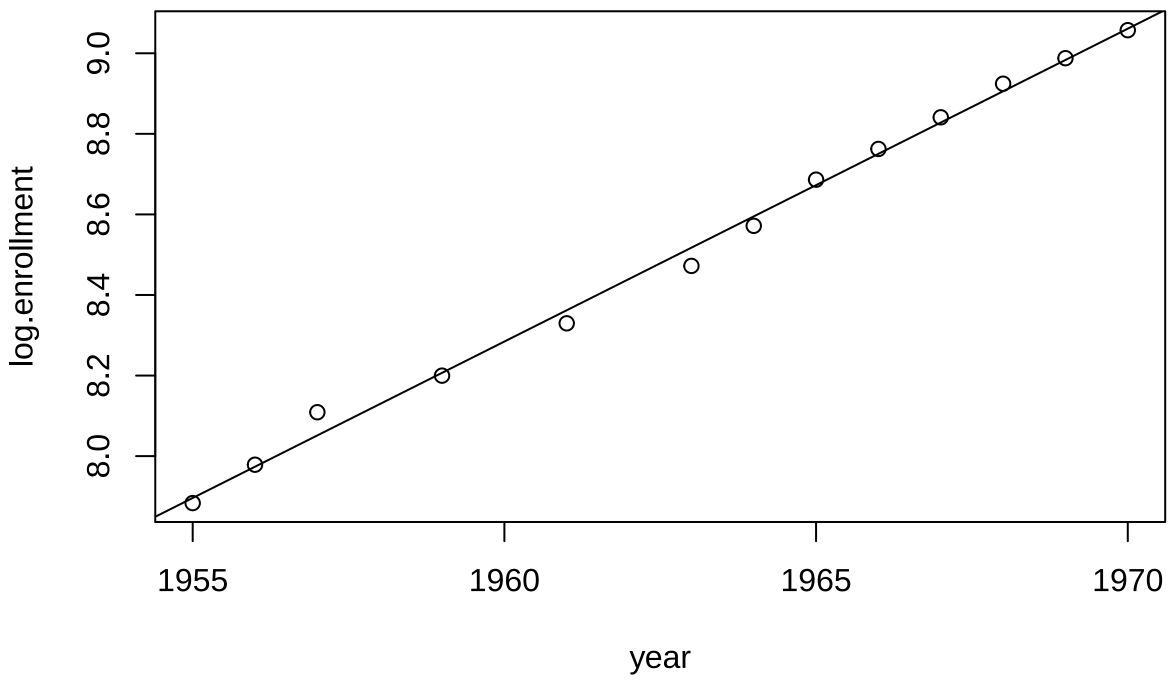
```
fit = lm(enrollment ~ year, data)
plot(enrollment ~ year, data)
abline(fit)
```



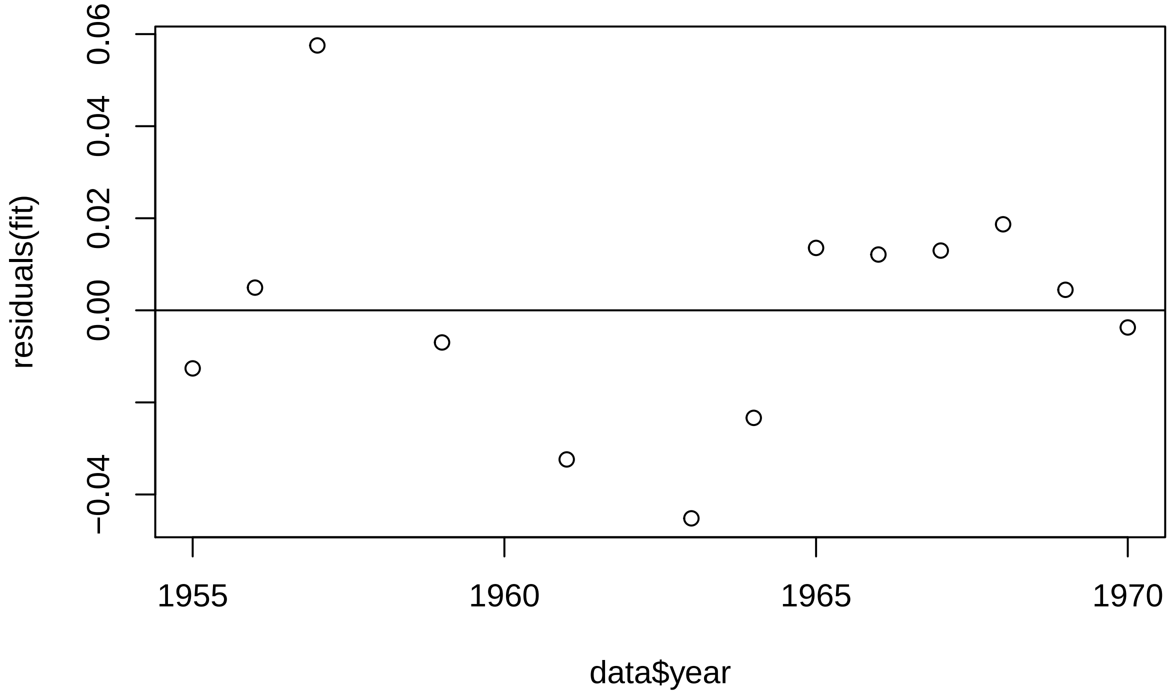
```
plot(data$year, residuals(fit))
abline(h=0)
```



```
data$log.enrollment = log(data$enrollment)
fit = lm(log.enrollment ~ year, data)
plot(log.enrollment ~ year, data)
abline(fit)
```



```r
plot(data$year, residuals(fit))
abline(h = 0)
```



5.5 (Exploring percentages of full-time faculty).

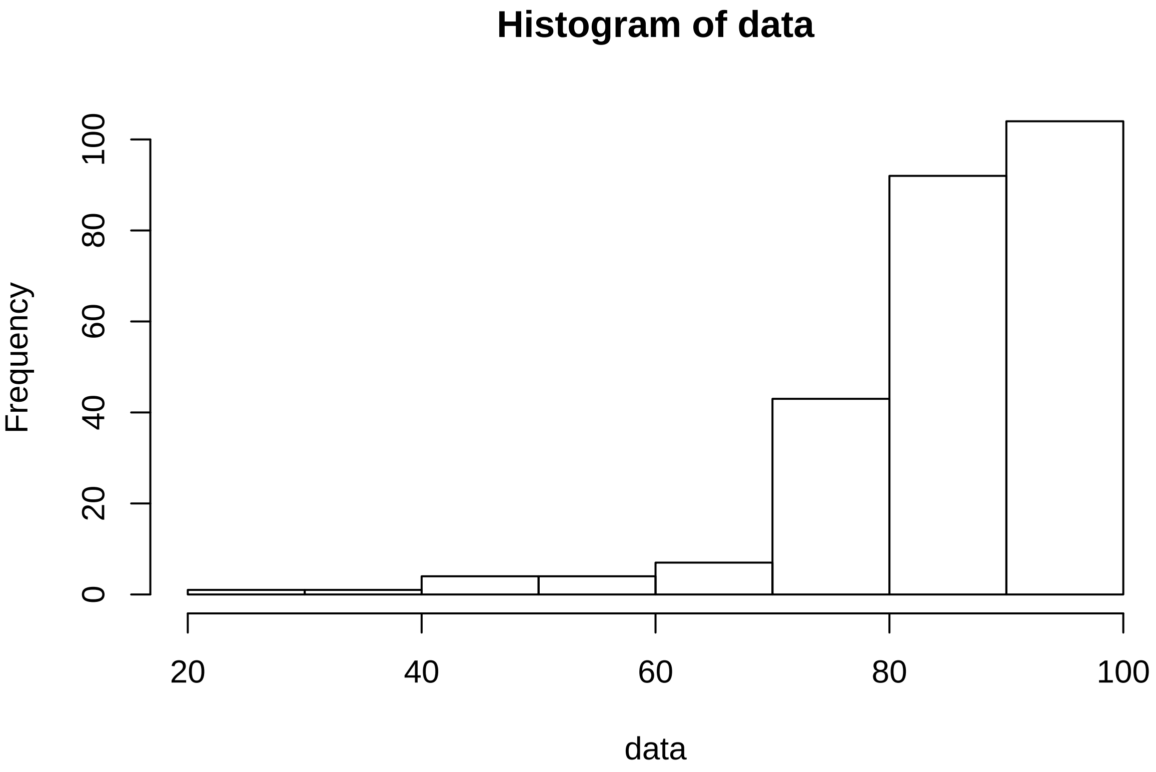
The variable Full.time in the college dataset (see Example 5.3) contains the percentage of faculty who are hired full-time in the group of National Universities.

a. Using the hist function, construct a histogram of the full-time percentages and comment on the shape of the distribution.

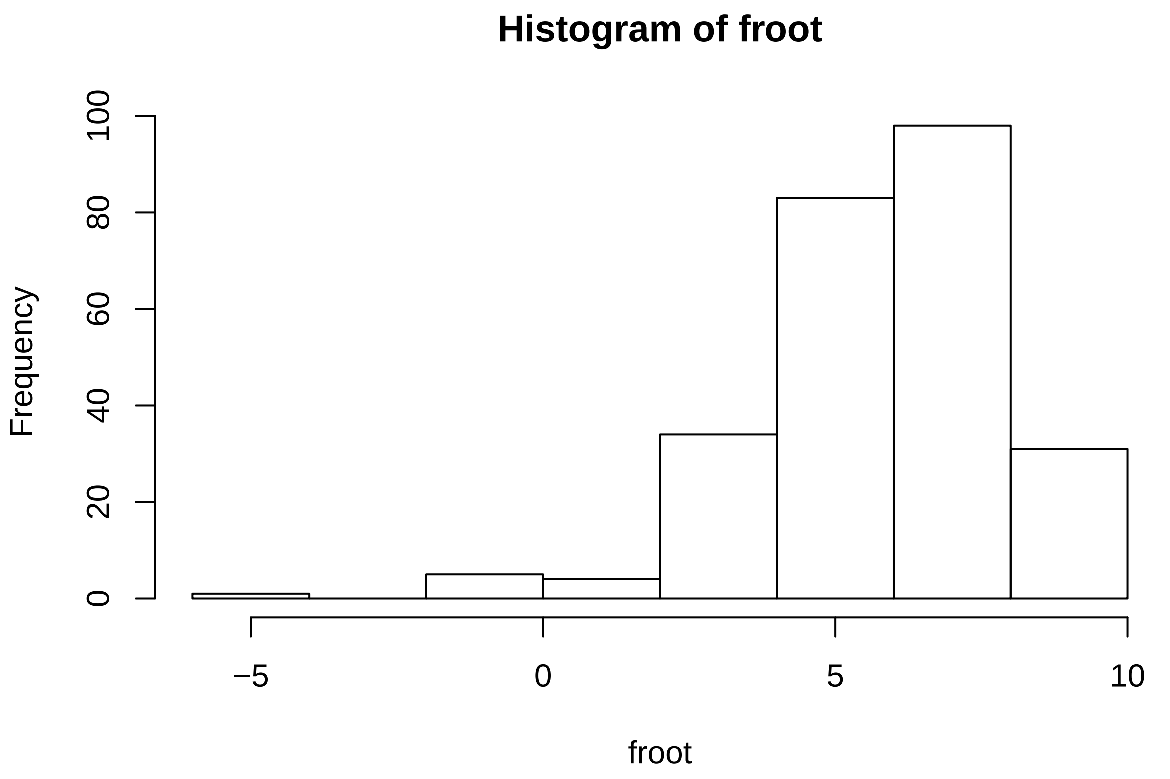
b. Use the froot and flog transformations to reexpress the full-time percent- ages. Construct histograms of the collection of froots and the collection of flogs. Is either transformation successful in making the full-time percent- ages approximately symmetric?

c. For data that is approximately normally distributed, about 68% of the data fall within one standard deviation of the mean. Assuming you have found a transformation in part (b) that makes the full-time percentages approximately normal, find an interval that contains roughly 68% of the data on the new scale.

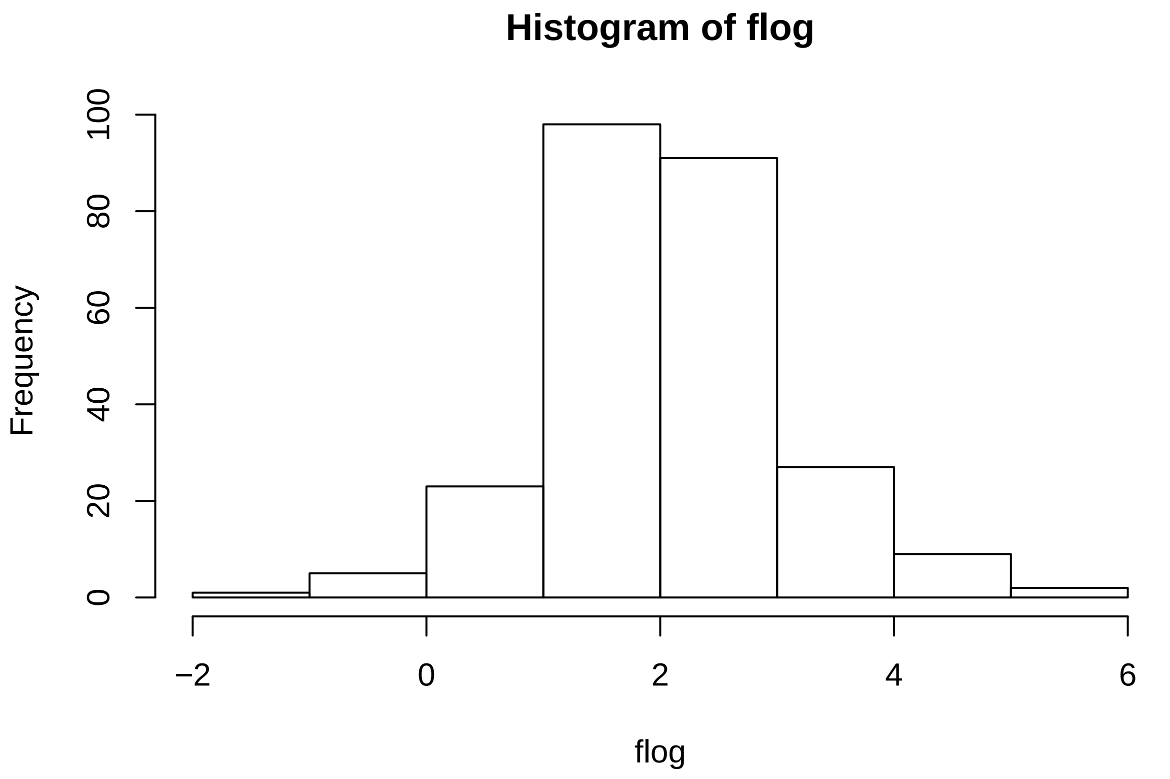
```
data = college$Full.time
hist(data)
```



```
froot = sqrt(data) - sqrt(100 - data)
flog = log(data + 0.5) - log(100 - data + 0.5)
hist(froot)
```


Histogram of froot
Frequency
0
20
40
60
80
100
−5
0
5
10
froot

```r
hist(flog)
```


Histogram of flog
Frequency
0
20
40
60
80
100
−2
0
2
4
6
flog

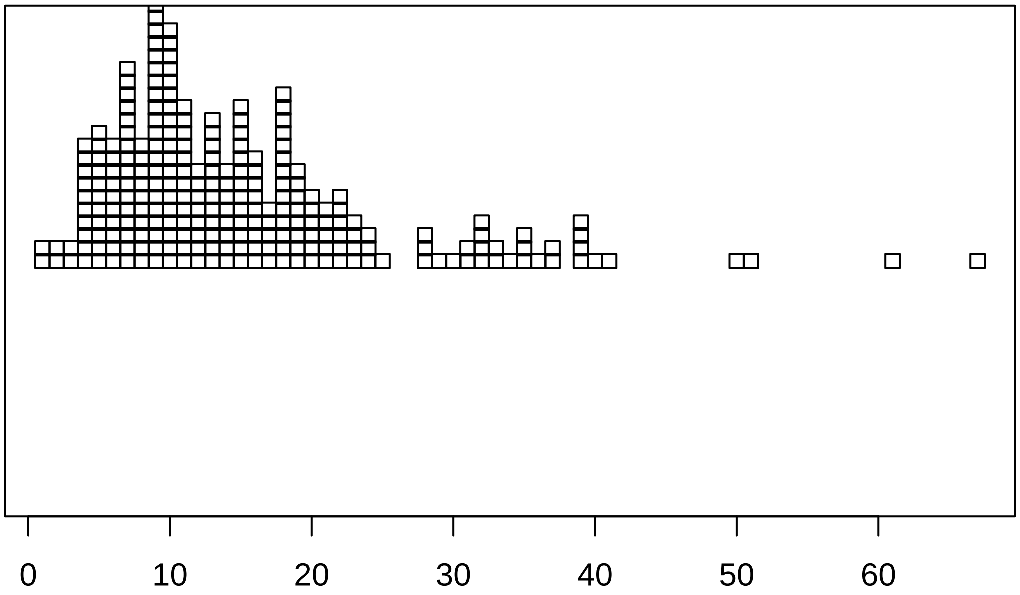
5.6 (Exploring alumni giving rates).

The variable Alumni.giving con- tains the percentage of alumni from the college who make financial contribu- tions.

a. Construct a "stacked" dotplot of the alumni giving percentages using the stripchart function.
b. Identify the names of the three schools with unusually large giving per- centages.
c. It can be difficult to summarize these giving percentages since the dis- tribution is right-skewed. One can make the dataset more symmetric by applying either a square root transformation or a log transformation.

Apply both square root and log transformations. Which transformation makes the alumni giving rates approximately symmetric?

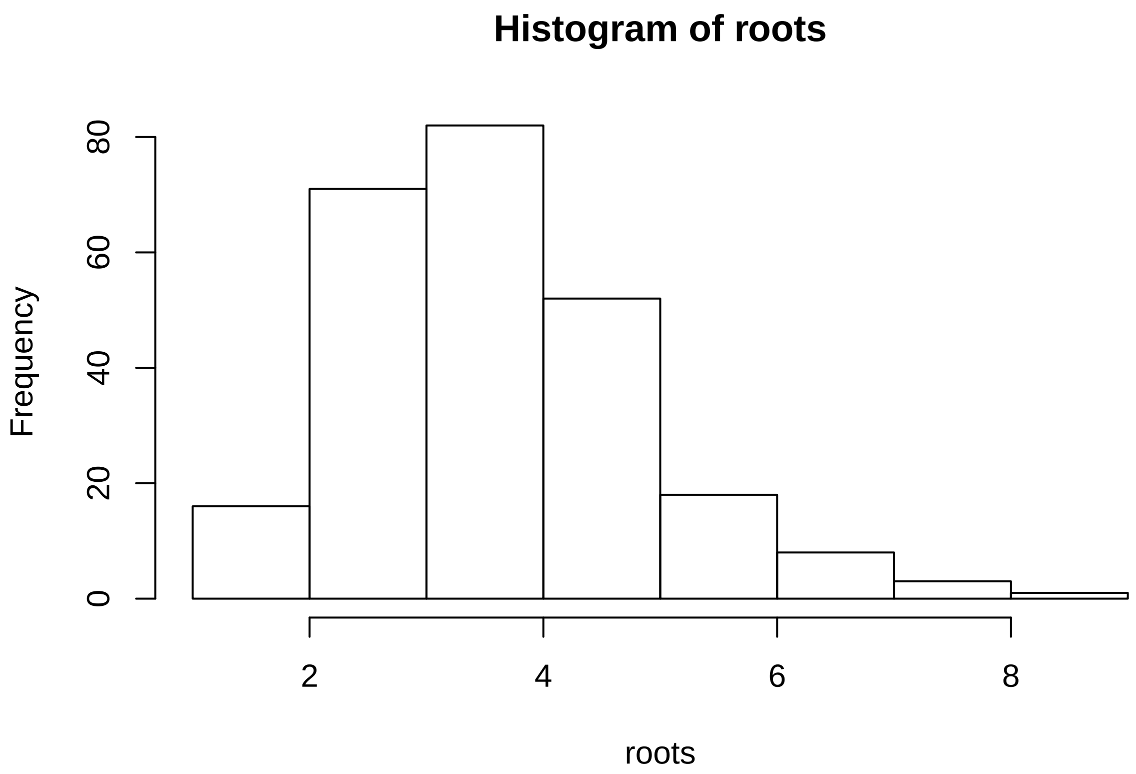
```
roots = sqrt(college$Alumni.giving)
logs = log(college$Alumni.giving)
stripchart(college$Alumni.giving, method="stack")
```



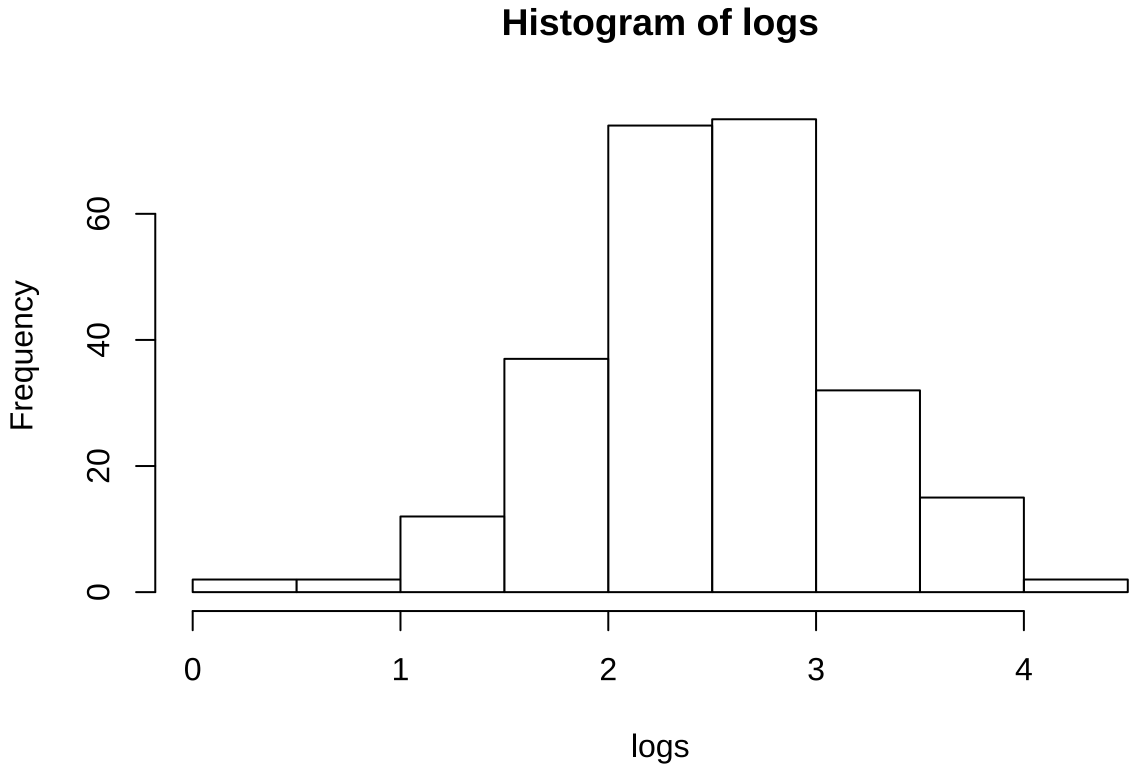
```
college[order(college$Alumni.giving,decreasing=TRUE),][1:3,]
```

```
##           School Rank Tier Retention Grad.rate Pct.20 Pct.50 Full.time
## 211 Idaho State   NA    4        56        27     NA     NA        NA
## 2      Princeton    1    1        98        96     75      9        92
## 11     Dartmouth   11    1        98        95     63      9        93
##     Top.10 Accept.rate Alumni.giving
## 211     NA          NA            67
## 2       97          10            61
## 11      90          13            51
```

```
hist(roots)
```


Histogram of roots
Frequency
0
20
40
60
80
2
4
6
8
roots

```r
hist(logs)
```


Histogram of logs
Frequency
0
20
40
60
0
1
2
3
4
logs

5.7 (Exploring alumni giving rates (continued)).

In this exercise, we focus on the comparison of the alumni giving percentages between the four tiers of colleges.

a. Using the stripchart function with the stacked option, construct parallel dotplots of alumni giving by tier.

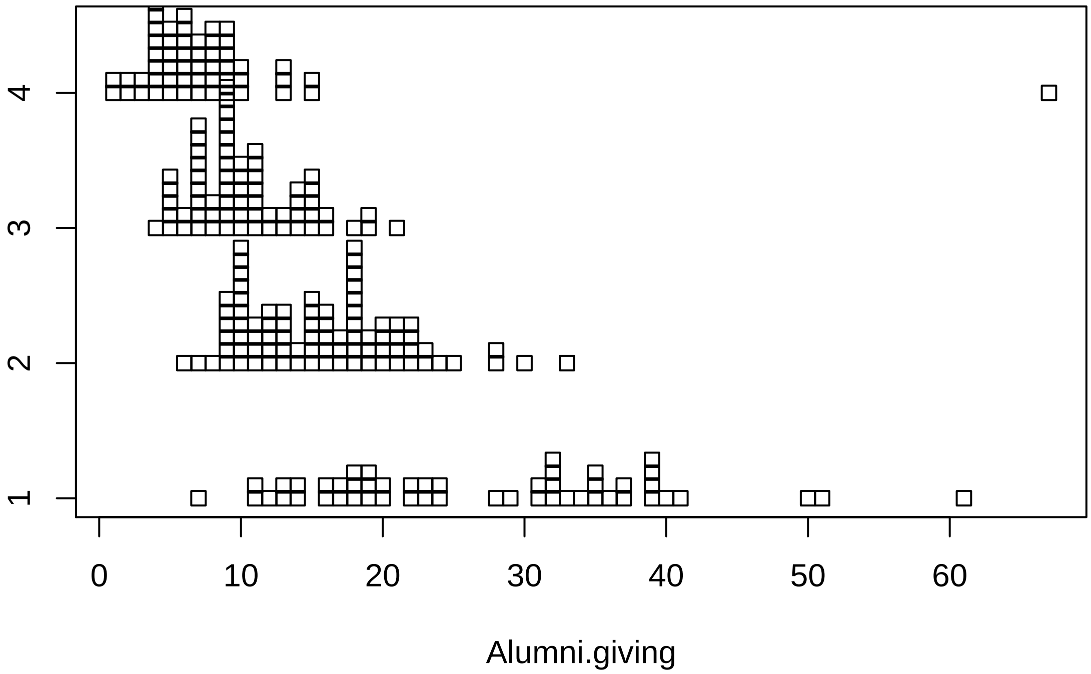
b. As one moves from Tier 4 to Tier 1, how does the average giving change?

c. As one moves from Tier 4 to Tier 1, how does the spread of the giving rates change?

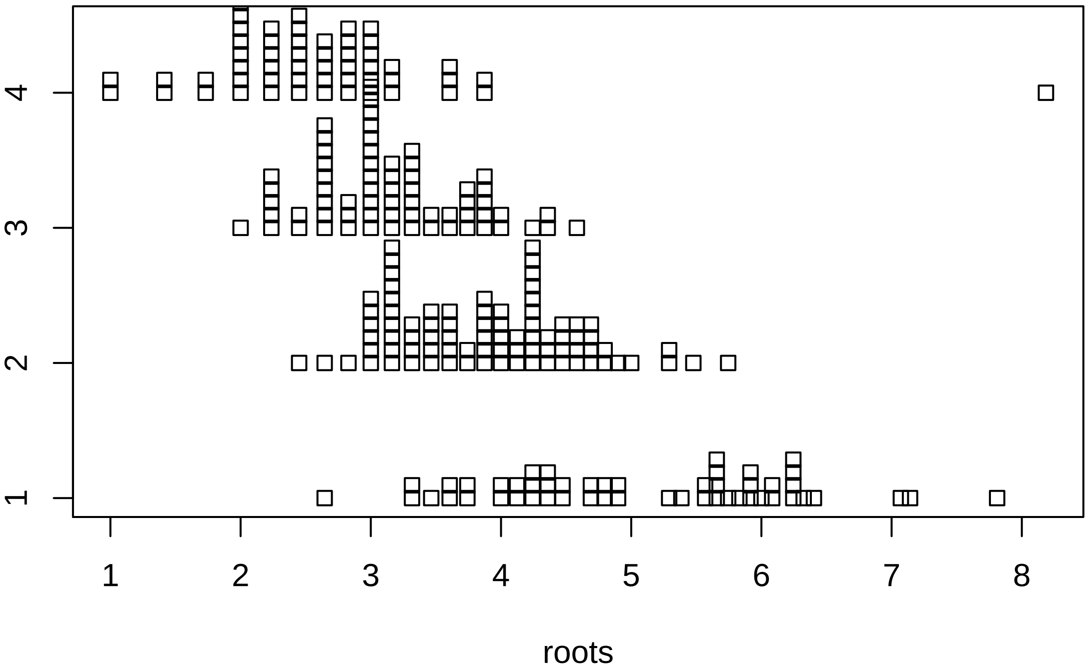
d. We note from parts (b) and (c), that small giving rates tend to have small variation, and large giving rates tend to have large variation. One way of removing the dependence of average with spread is to apply a power transformation such as a square root or a log. Construct parallel stripcharts of the square roots of the giving rates, and parallel boxplots of the log giving rates.

e. Looking at the two sets of parallel stripcharts in part (d), were the square root rates or the log rates successful in making the spreads approximately the same between groups?

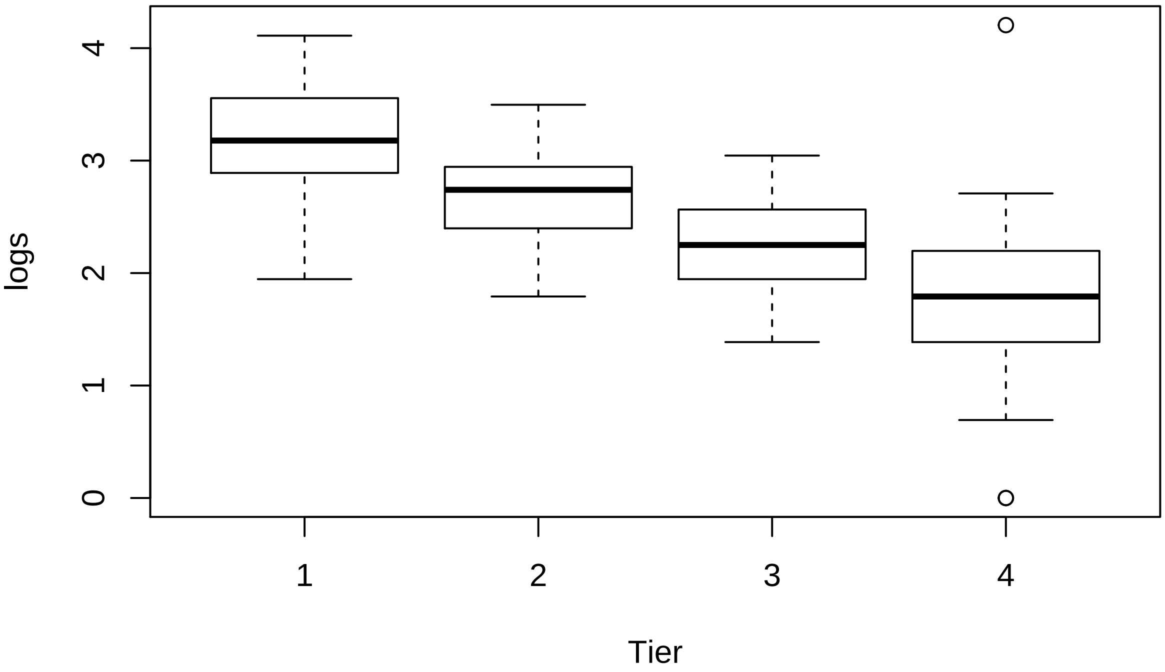
```
stripchart(Alumni.giving ~ Tier, college, method="stack")
```



```
stripchart(roots ~ Tier, college, method="stack")
```



```
boxplot(logs ~ Tier, college, method="stack")
```


logs
Tier