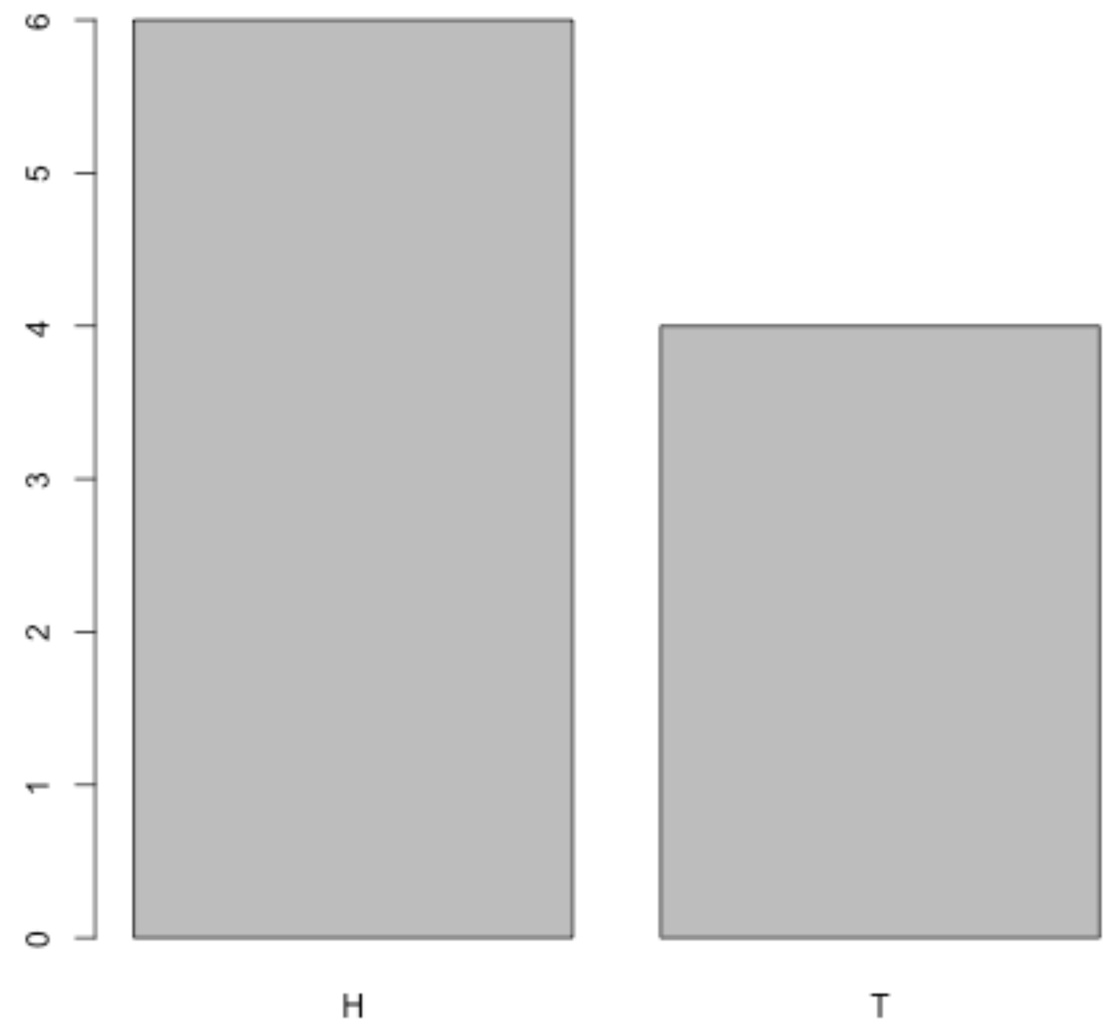
Basic numerical and graphical summaries

Categorical data

- We can create a vector of characters as follows:

```
> tosses=scan(what="character")
1: H T H H T T H T H H
11:
Read 10 items
> table(tosses)
tosses
H T
6 4
>barplot(table(tosses))
```



Basic numerical and graphical summaries

- As seen before, factors are useful to represent character vectors:

```
> as.factor(tosses)
 [1] H T H H T T H T H H
Levels: H T
```

Multinomial experiments: Chi-square goodness-of-fit test

Example: Do car crashes occur on different days with the same frequency?

Day	Mon	Tue	Wed	Thur	Fri	Sat	Sun
# of fatalities	20	20	22	22	29	36	31

Basic numerical and graphical summaries

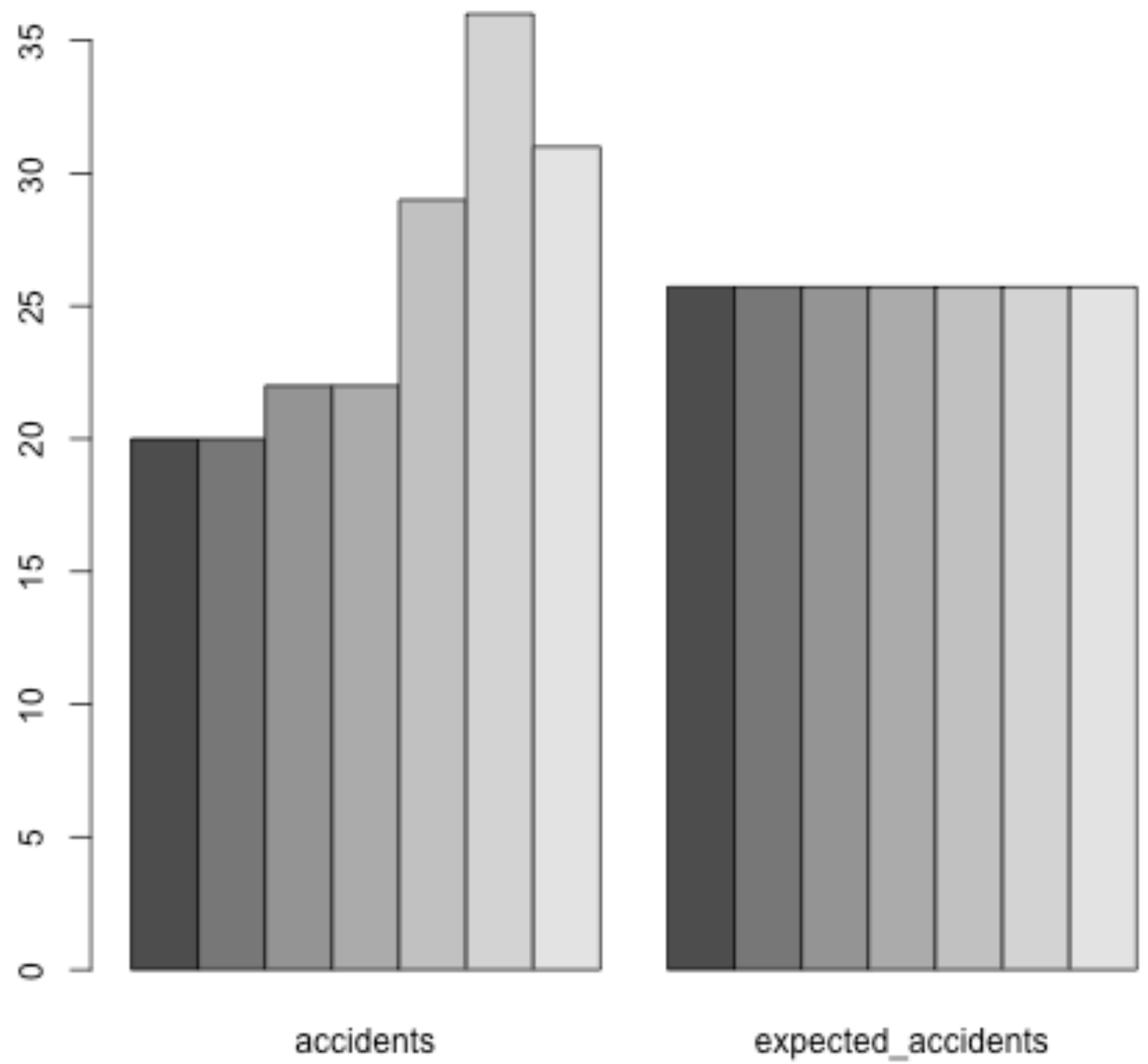
```
> accidents=c(20,20,22,22,29,36,31)
> sum(accidents)
[1] 180
> expected_accidents=rep(180/7,7)
> chi_statistic=sum((accidents-
expected_accidents)^2/expected_accidents)
> chi_statistic # Compare with chi-square k-1 df
[1] 9.233333
> 1-pchisq(chi_statistic,6) # p-value
[1] 0.1608746
> chisq.test(accidents)

    Chi-squared test for given probabilities

data:  accidents
X-squared = 9.2333, df = 6, p-value = 0.1609
```

Syntax: `chisq.test(x,p,…)`

Expected probabilities (uniform by default)

Basic numerical and graphical summaries

```
>names(accidents)=days
>names(expected_accidents)=days
>barplot(cbind(accidents,expected_accidents),beside=TRUE)
```



Relating Two Categorical Variables

Observational Study: How much will an additional year of schooling raise one's income?

Challenges in observational study: (a) many variables relate to a person's income; (b) it can be difficult to obtain truthful information (people are more likely to report a higher level).

Researchers interviewed monozygotic twins (identical family backgrounds). Information on schooling was obtained from a twin (self-reporting) and from his/her twin (cross-reported).

183 pairs of twins; one randomly assigned to "twin1" other one "twin2"; EDUCL and EDUCH: self-reported education from twin1 and twin2.

Basic numerical and graphical summaries

```
> table(twn$EDUCL)
 8 10 11 12 13 14 15 16 17 18 19 20
 1  4  1 61 21 30 11 37  1 10  3  3
```

high school　　college

```
> table(twn$EDUCH)
 8  9 10 11 12 13 14 15 16 17 18 19 20
 2  1  2  1 65 22 22 15 33  2 11  2  5
```

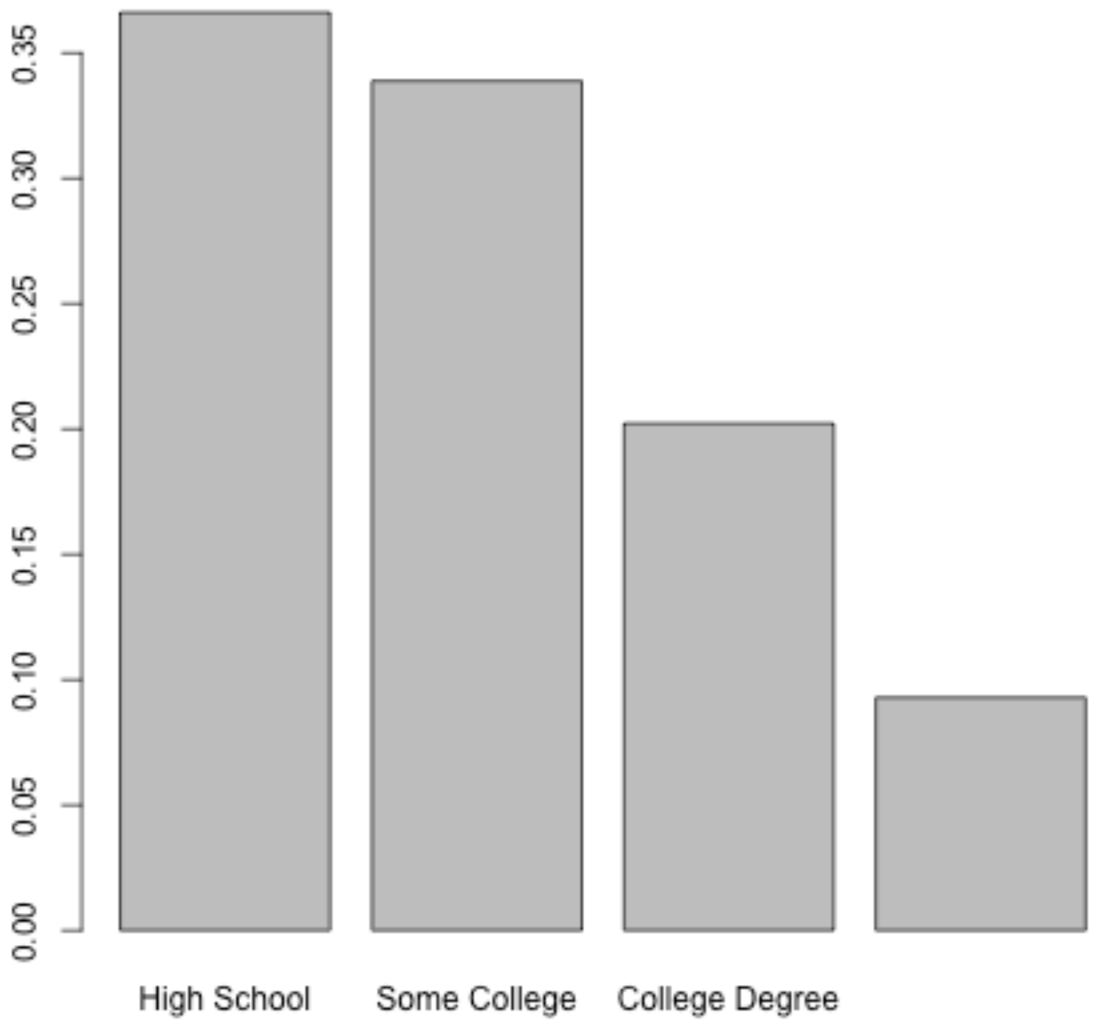
high school　　college

It is useful to categorize this variable into a smaller number of levels: “high-school” (12 years), “some college” (13-15), “college degree” (16 years), “graduate school” (above 16)…

Basic numerical and graphical summaries

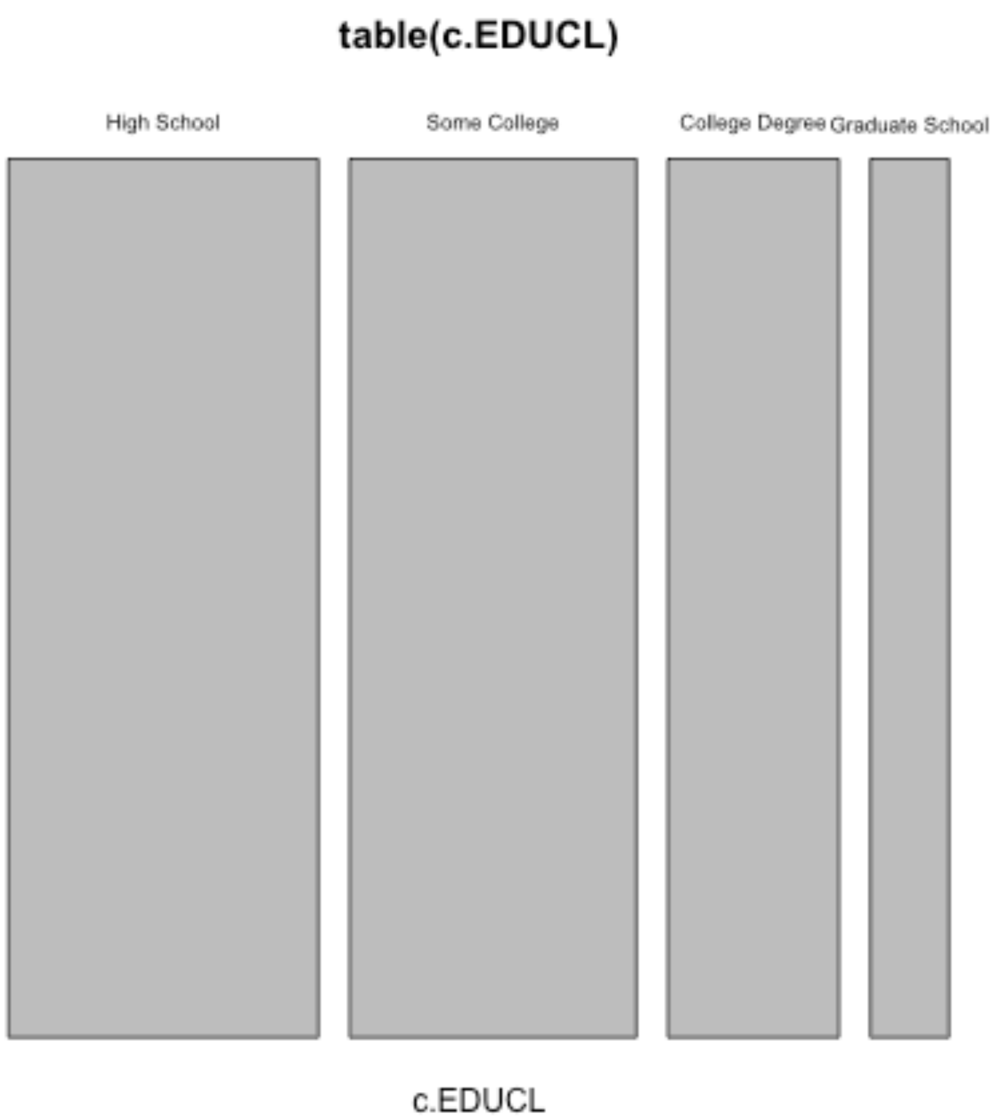
```
>c.EDUCL = cut(twn$EDUCL, breaks=c(0, 12, 15, 16, 24),
  labels=c("High School", "Some College", "College Degree",
  "Graduate School"))
>c.EDUCH = cut(twn$EDUCH, breaks=c(0, 12, 15, 16, 24),
  labels=c("High School", "Some College", "College Degree",
  "Graduate School"))

>table(c.EDUCL)
c.EDUCL
High School     Some College   College Degree Graduate School
67               62               37               17
>prop.table(c.EDUCL)
c.EDUCL
    High School  Some College  College Degree  Graduate School
         0.3661        0.3388          0.2022           0.0929
```

Basic numerical and graphical summaries



Barplot

```
>barplot(c.EDUCL)
```



Mosaic Plot

```
>mosaicplot(table(c.EDUCL))
```

Creating contingency tables

```
> table(c.EDUCL, c.EDUCH)
                 c.EDUCH
c.EDUCL           High School Some College College Degree Graduate School
  High School              47           16              2               2
  Some College             18           32              8               4
  College Degree            5           10             18               4
  Graduate School           1            1              5              10
```

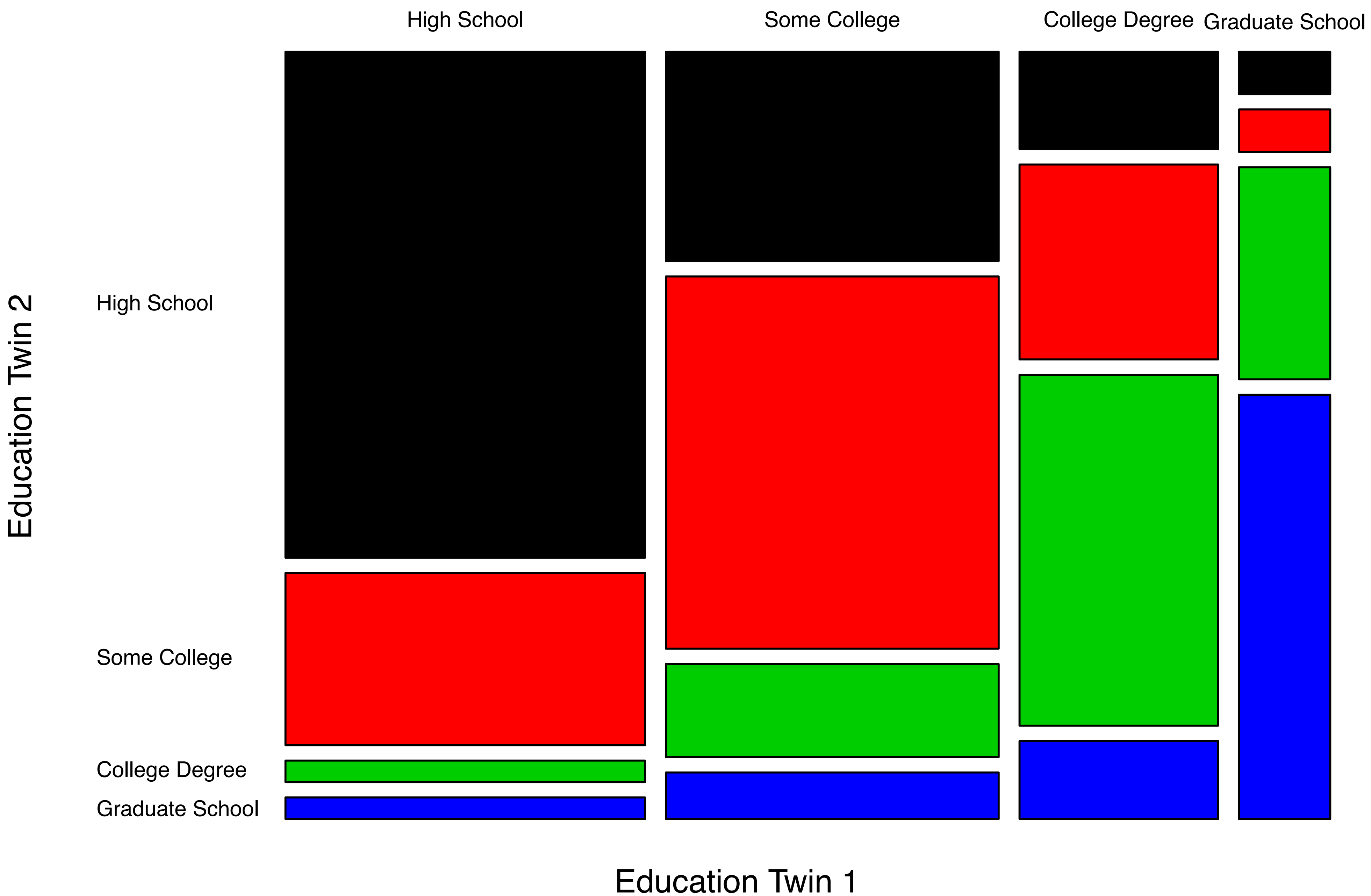
Diagonal has largest counts: twins with the same reported educational levels

What proportion of twins have the same level?

```
> T1=table(c.EDUCL, c.EDUCH)
> diag(T1)
    High School    Some College  College Degree Graduate School
             47              32              18              10
>
> sum(diag(T1)) / sum(T1)
[1] 0.5846995
```

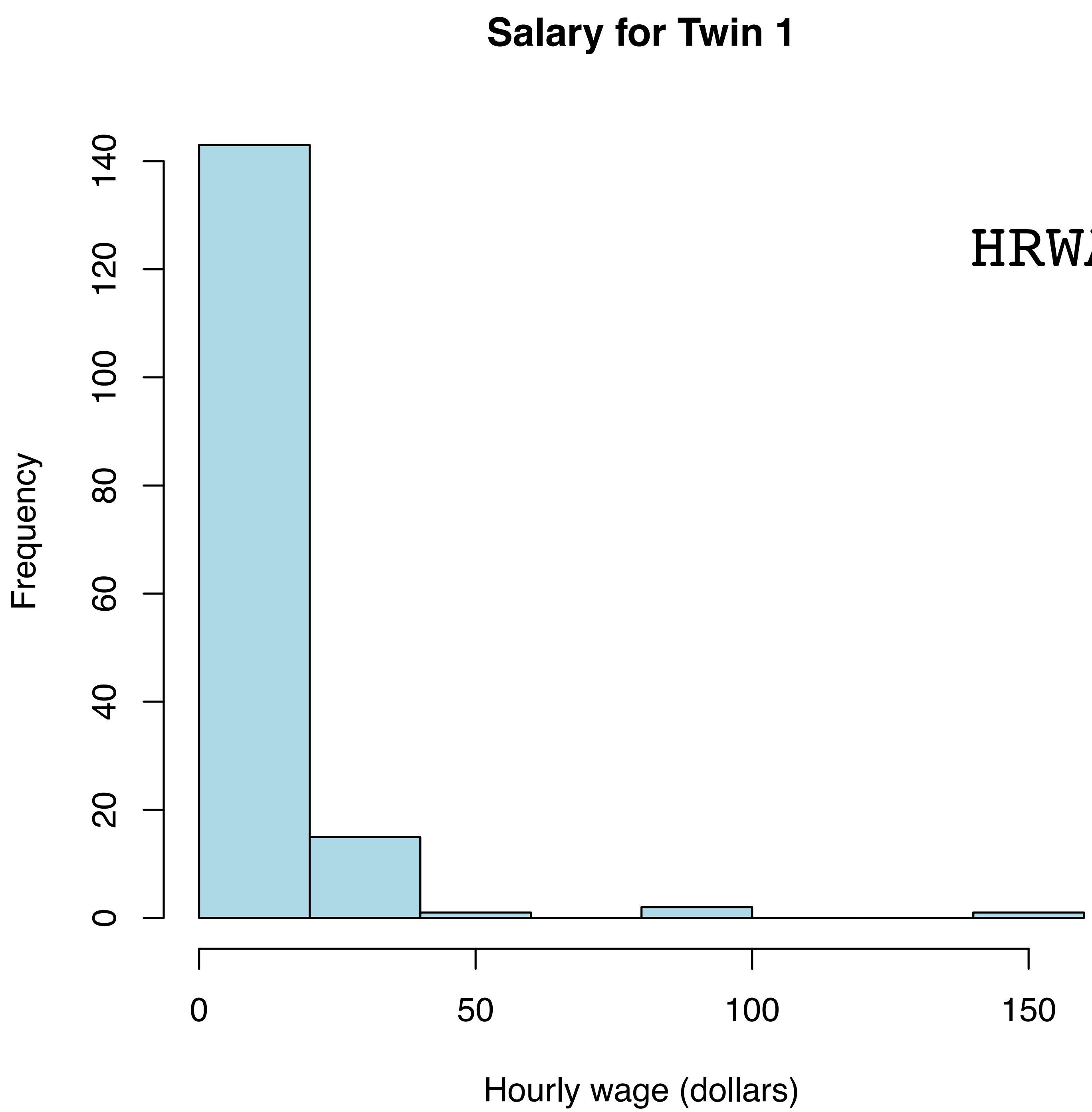
Basic numerical and graphical summaries

```
>mosaicplot(T1,color=1:4,las=1,main="",xlab="Education Twin 1”,
ylab="Education Twin 2”)
```



Basic numerical and graphical summaries

Goal: explore the relationship between educational level and salary.



`HRWAGEL`: hourly wage for Twin 1

right skewed

Basic numerical and graphical summaries

```
> c.wage = cut(twn$HRWAGEL, c(0, 7, 13, 20, 150))
> table(c.wage)
c.wage
   (0,7]   (7,13]  (13,20] (20,150]
      47       58       38       19
```

Note that there were 21 twins who did not respond the wage question so we have 183-21=162 recorded wages.

```
> table(c.EDUCL,c.wage)
                   c.wage
c.EDUCL             (0,7] (7,13] (13,20] (20,150]
  High School          23     21      10        1
  Some College         15     23      12        5
  College Degree        7     12      14        3
  Graduate School       2      2       2       10
```

Basic numerical and graphical summaries

Compute the proportions of different wages for each educational level:

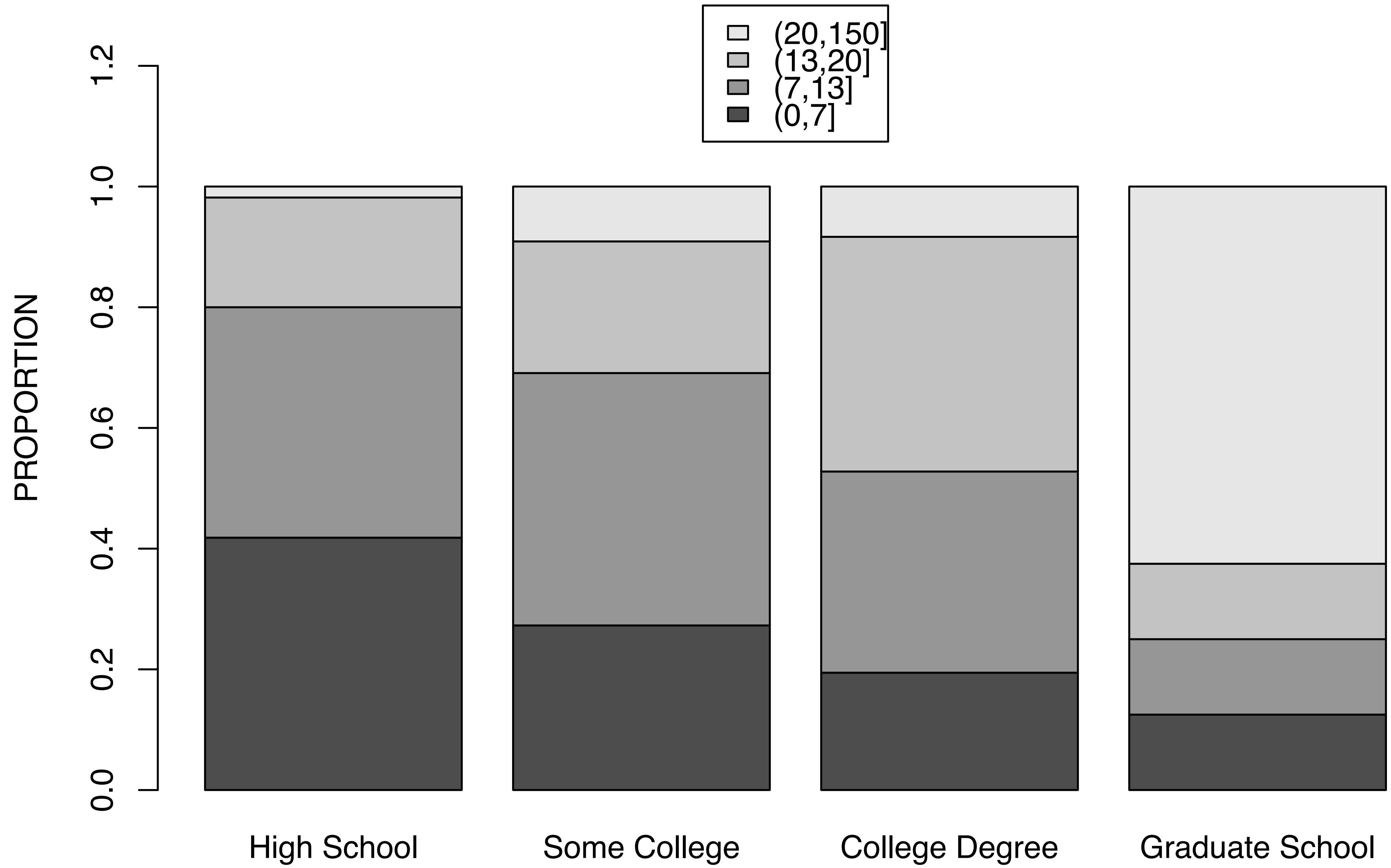
```
> round(prop.table(T2, margin=1),4)
                    c.wage
c.EDUCL               (0,7] (7,13] (13,20] (20,150]
  High School        0.4182 0.3818  0.1818   0.0182
  Some College       0.2727 0.4182  0.2182   0.0909
  College Degree     0.1944 0.3333  0.3889   0.0833
  Graduate School    0.1250 0.1250  0.1250   0.6250
```

Is there an association between educational level and the salary?

Some visualization tools before formal testing…

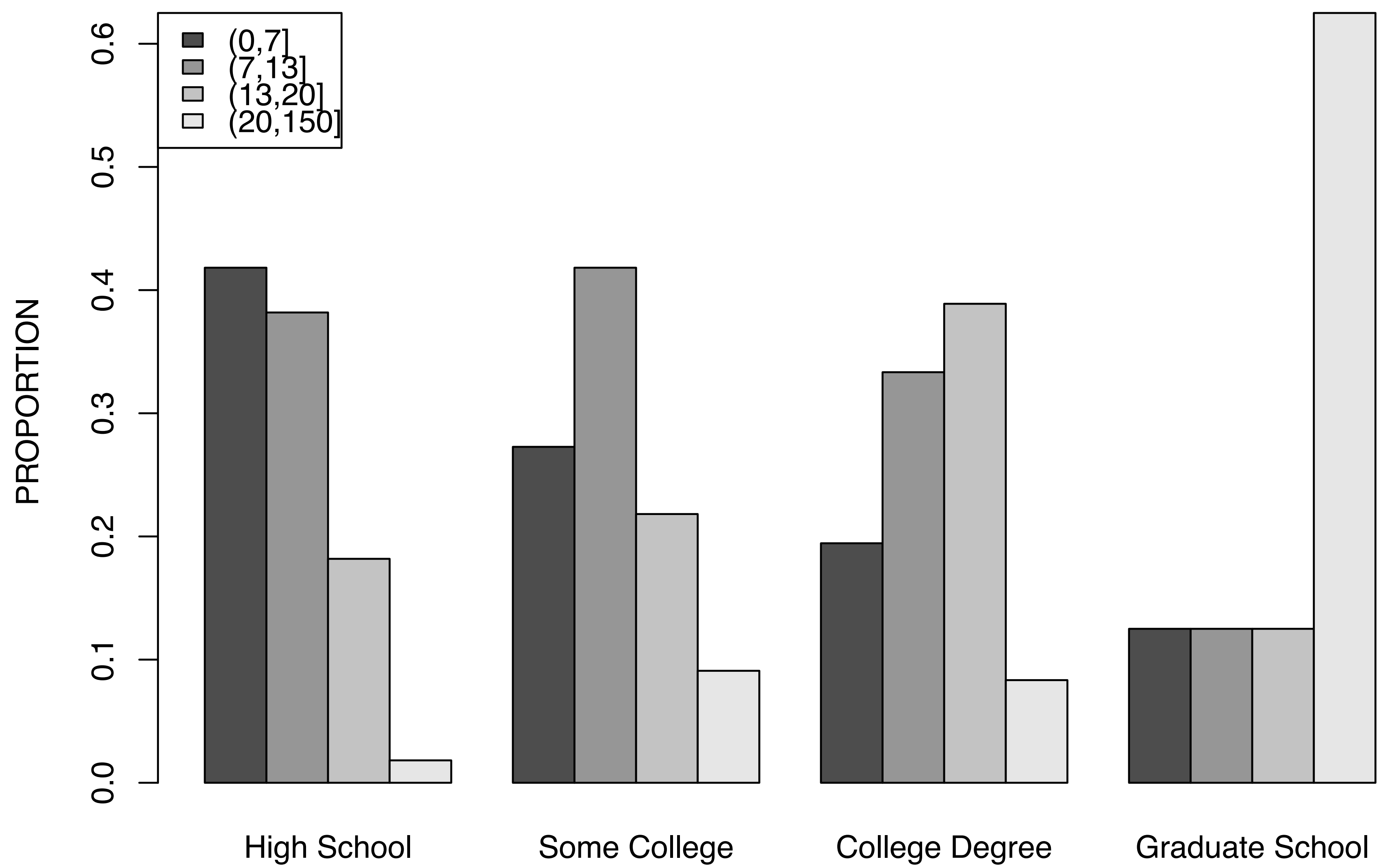
Basic numerical and graphical summaries

```
> P = prop.table(T2, 1)
> barplot(t(P), ylim=c(0, 1.3), ylab="PROPORTION",
+   legend.text=dimnames(P)$c.wage,
+   args.legend=list(x = "top"))
```



Basic numerical and graphical summaries

```
> barplot(t(P), beside=T,  legend.text=dimnames(P)$c.wage,
+ args.legend=list(x="topleft"), ylab="PROPORTION")
```



Testing independence using a chi-square test

$$H_0: \quad \text{education background and wage are independent}$$

The Pearson statistic is defined by:

$$X^2 = \sum_{\text{all cells}} \frac{(observed - expected)^2}{expected}$$

Under the null hypothesis and for large samples this statistic will be distributed as

$$\chi^2_{(n_r-1)\times(n_c-1)}$$

Basic numerical and graphical summaries

```
> T2 = table(c.EDUCL, c.wage)
> T2
                   c.wage
c.EDUCL             (0,7] (7,13] (13,20] (20,150]
  High School          23     21      10        1
  Some College         15     23      12        5
  College Degree        7     12      14        3
  Graduate School       2      2       2       10
> S = chisq.test(T2)
```

Warning message:
In chisq.test(T2) : Chi-squared approximation may be incorrect

```
> print(S)

	Pearson's Chi-squared test

data:  T2
X-squared = 54.578, df = 9, p-value = 1.466e-08
```

Basic numerical and graphical summaries

Checking our calculations:

- Expected frequency for a given cell:

$$E = \frac{(\text{row total}) \times (\text{column total})}{(\text{grand total})}$$

```
> A=matrix(rep(rowSums(T2),4),4,4,byrow=T)
> B=matrix(rep(colSums(T2),4),4,4)
> Expected=t(A*B/sum(T2))
> Expected
          [,1]      [,2]      [,3]     [,4]
[1,] 15.956790 19.691358 12.901235 6.450617
[2,] 15.956790 19.691358 12.901235 6.450617
[3,] 10.444444 12.888889  8.444444 4.222222
[4,]  4.641975  5.728395  3.753086 1.876543
```

Basic numerical and graphical summaries

Same as:

```
> S$expected
                    c.wage
c.EDUCL                 (0,7]    (7,13]   (13,20] (20,150]
  High School       15.956790 19.691358 12.901235 6.450617
  Some College      15.956790 19.691358 12.901235 6.450617
  College Degree    10.444444 12.888889  8.444444 4.222222
  Graduate School    4.641975  5.728395  3.753086 1.876543
```

Then,

```
> sum((T2 - S$expected)^2 / S$expected)
[1] 54.57759
>
> 1 - pchisq(54.57759, df=9)
[1] 1.465839e-08
```

Based on this p-value we reject the null hypothesis

Basic numerical and graphical summaries

One useful component is `residuals`, defined as:

$$\frac{observed - expected}{\sqrt{expected}}$$

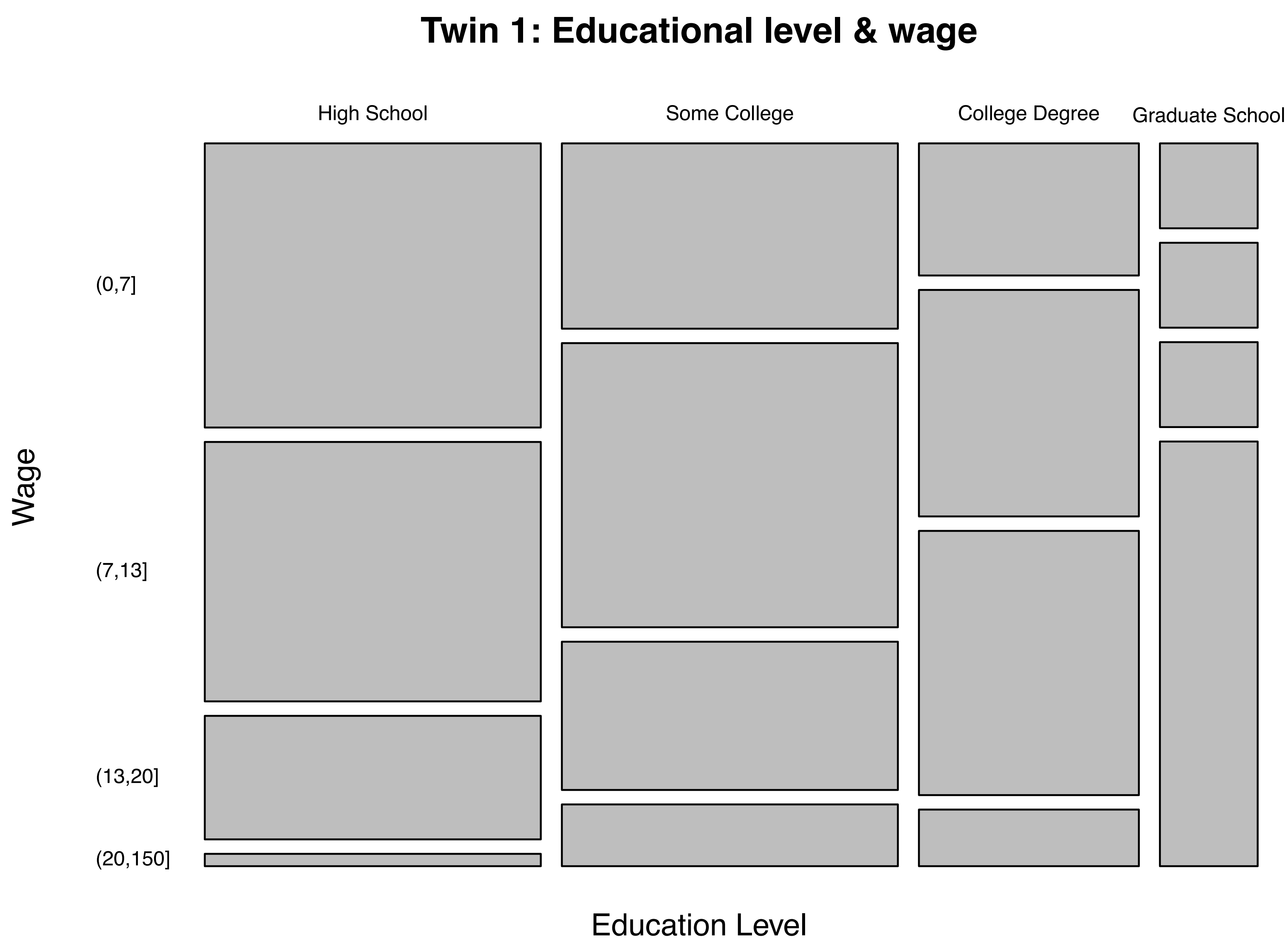
```
> S$residuals
                 c.wage
c.EDUCL                (0,7]     (7,13]    (13,20]   (20,150]
  High School      1.7631849  0.2949056 -0.8077318 -2.1460758
  Some College    -0.2395212  0.7456104 -0.2509124 -0.5711527
  College Degree  -1.0658020 -0.2475938  1.9117978 -0.5948119
  Graduate School -1.2262453 -1.5577776 -0.9049176  5.9300942
```

- Fewer High School people earning wages over $20 than expected under the independence model

- More Graduate School people earning wages over $20 than anticipated under the independence model

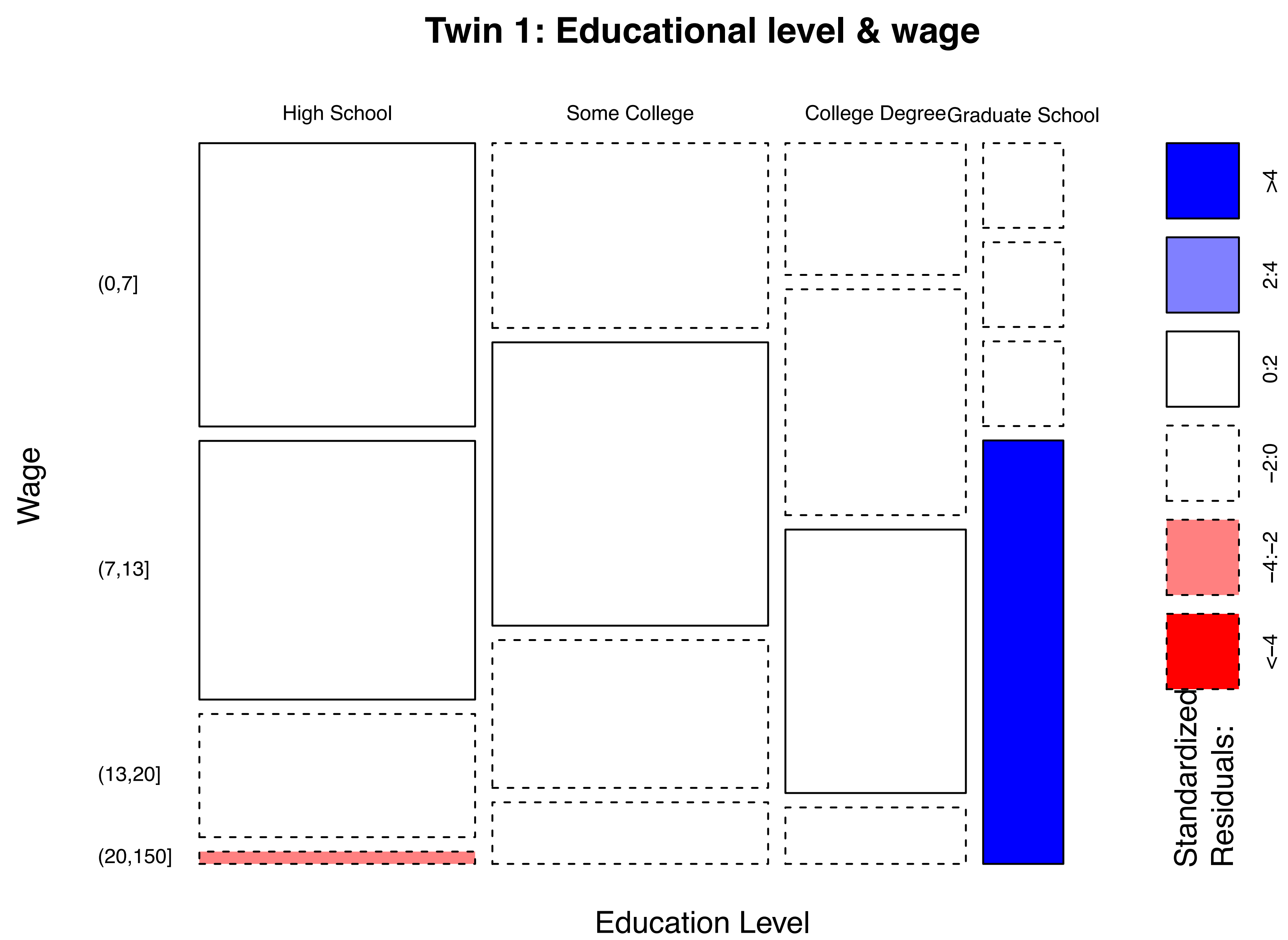
Basic numerical and graphical summaries

```
> mosaicplot(T2, shade=FALSE,main="Twin 1: Educational level &
wage",las=1,xlab="Education Level",ylab="Wage")
```



Basic numerical and graphical summaries

```
> mosaicplot(T2, shade=TRUE,main="Twin 1: Educational
level & wage",las=1,xlab="Education Level",ylab="Wage")
```



EDA

- John Tukey and other statisticians devised a collection of methods for exploratory data analysis (**EDA**). Tukey makes a distinction between *confirmatory analysis* (drawing inferential conclusions) and *exploratory methods* (few assumptions about distributions, only looking for patterns)

- General themes:
  - Revelation
  - Resistance
  - Residuals
  - Reexpression

- **Revelation:** graphical displays; discovering patterns
- **Resistant methods:** methods insensitive to extreme observations

- **Residuals:** focus is not on the fitted model (e.g., fitted regression line) but on the deviations from that model
- **Reexpress:** focus is transforming the data to see patterns that cannot be seen in the original scale.

Case Study: 2009 ratings of colleges. Data from U.S. News and World Report (America's Best Colleges):

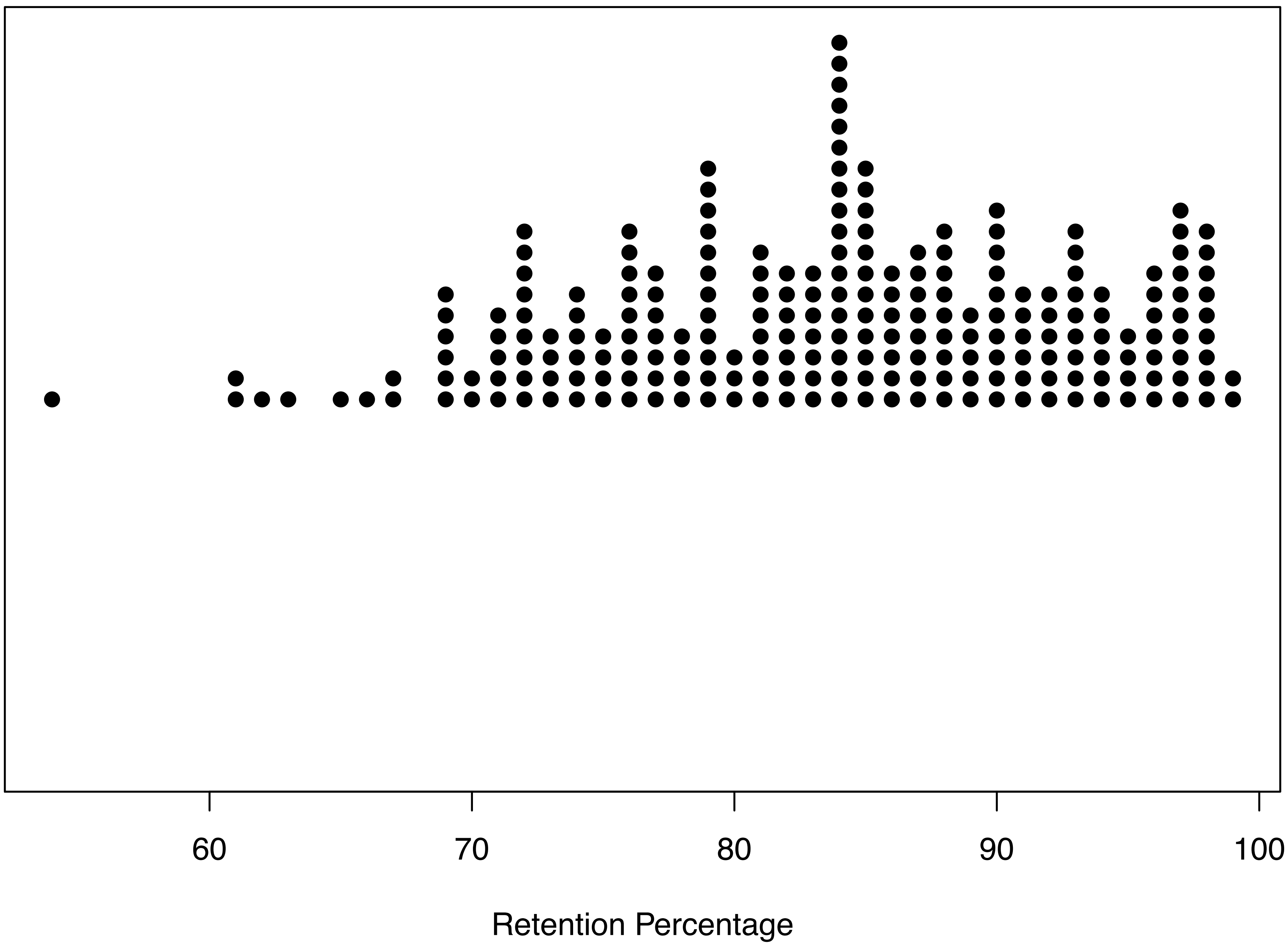
a. School – the name of the college
b. Tier – the rank of the college into one of four tiers
c. Retention – the percentage of freshmen who return to the school the following year
d. Grad.rate – the percentage of freshman who graduate in a period of six years
e. Pct.20 – the percentage of classes with 20 or fewer students
f. Pct.50 – the percentage of classes with 50 or more students
g. Full.time – the percentage of faculty who are hired full-time
h. Top.10 – the percentage of incoming students who were in the top ten percent of their high school class
i. Accept.rate – the acceptance rate of students who apply to the college
j. Alumni.giving – the percentage of alumni from the college who contribute financially

EDA

```
> dat = read.table("college.txt", header=TRUE, sep="\t")
> college = subset(dat, complete.cases(dat))
> head(college)
     School Enrollment Tier Retention Grad.rate Pct.20 Pct.50 Full.time Top.10
1   Harvard      19230    1        97        98     77      8        93     95
2 Princeton       7497    1        98        96     75      9        92     97
3      Yale      11446    1        99        97     79      7        88     97
4  Cal Tech       2126    1        98        88     71      6        97     97
5       MIT      10299    1        98        94     65     13        90     97
6  Stanford      17833    1        98        94     72     12        99     92
  Accept.rate Alumni.giving
1           8            40
2          10            61
3           9            41
4          17            31
5          12            37
6           9            35

> names(college)
 [1] "School"        "Enrollment"    "Tier"          "Retention"
 [5] "Grad.rate"     "Pct.20"        "Pct.50"        "Full.time"
 [9] "Top.10"        "Accept.rate"   "Alumni.giving"
>attach(college)
```

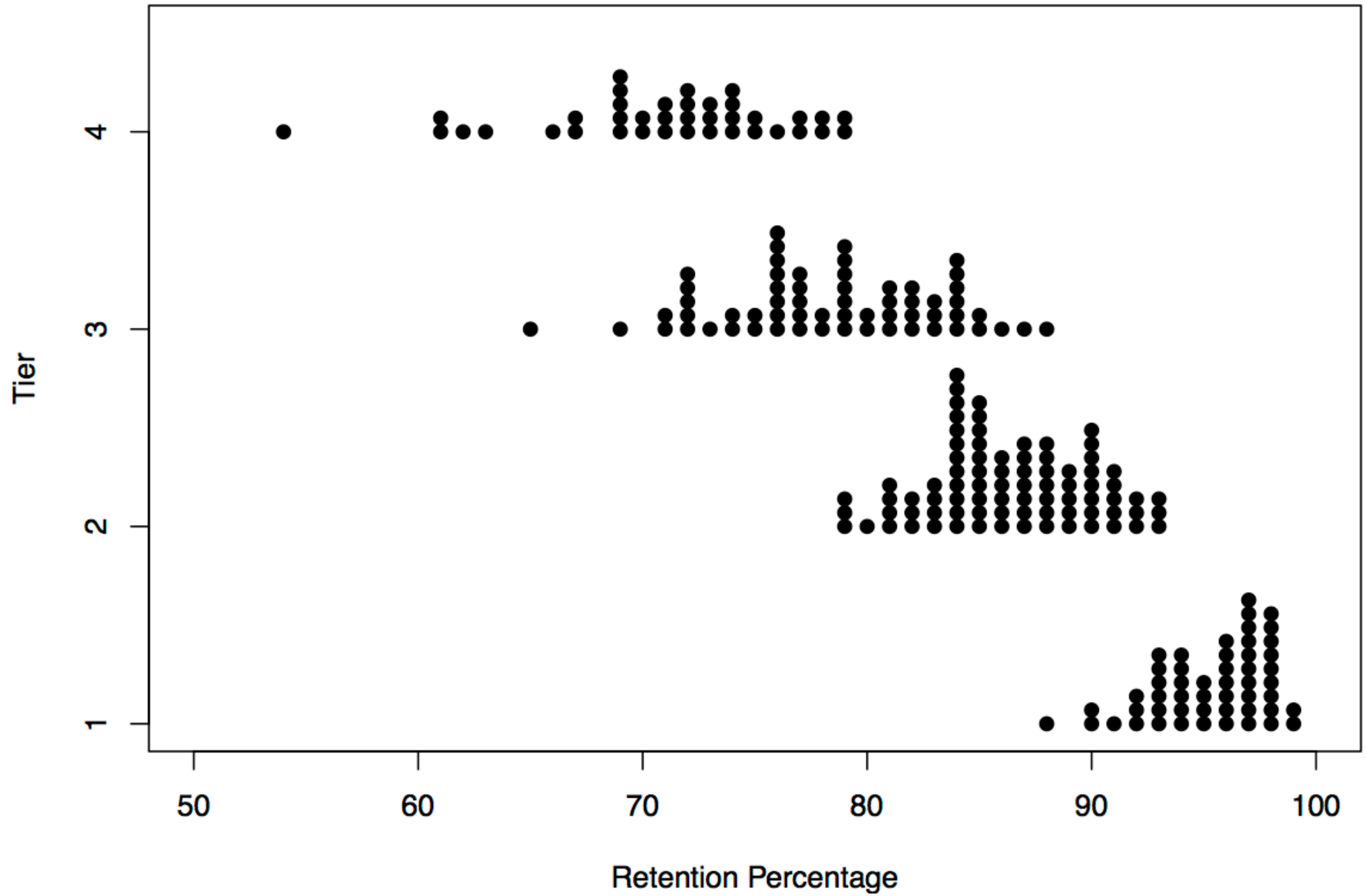
- Let's look at the retention variable:

```
>stripchart(Retention, method="stack", pch=19,
xlab="Retention Percentage")
```



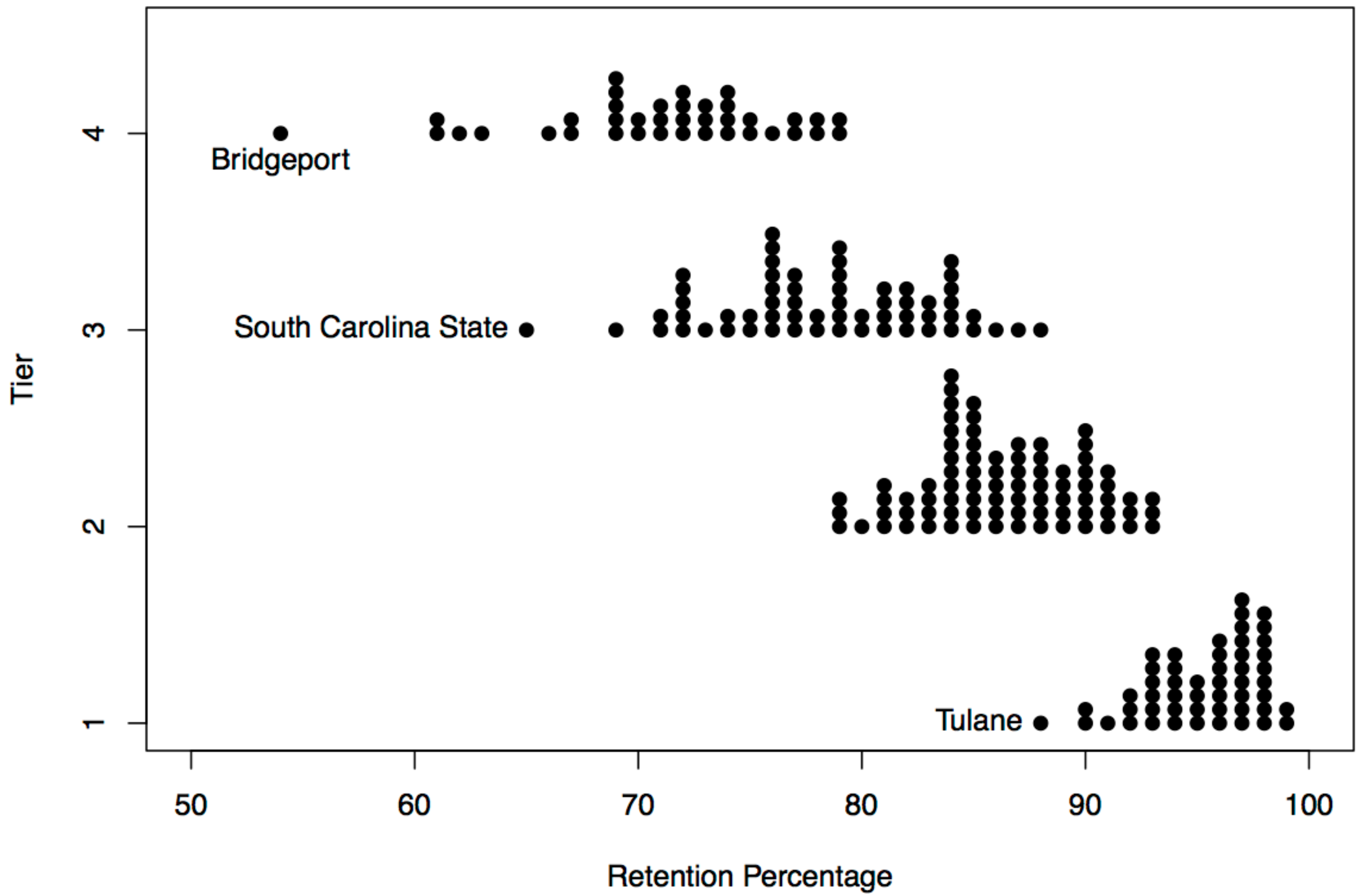
EDA

```
> stripchart(Retention ~ Tier, method="stack", pch=19,
xlab="Retention Percentage",ylab="Tier", xlim=c(50, 100))
```



EDA

```
> identify(Retention, Tier, n=3, labels=School)
[1] 50 158 212
```



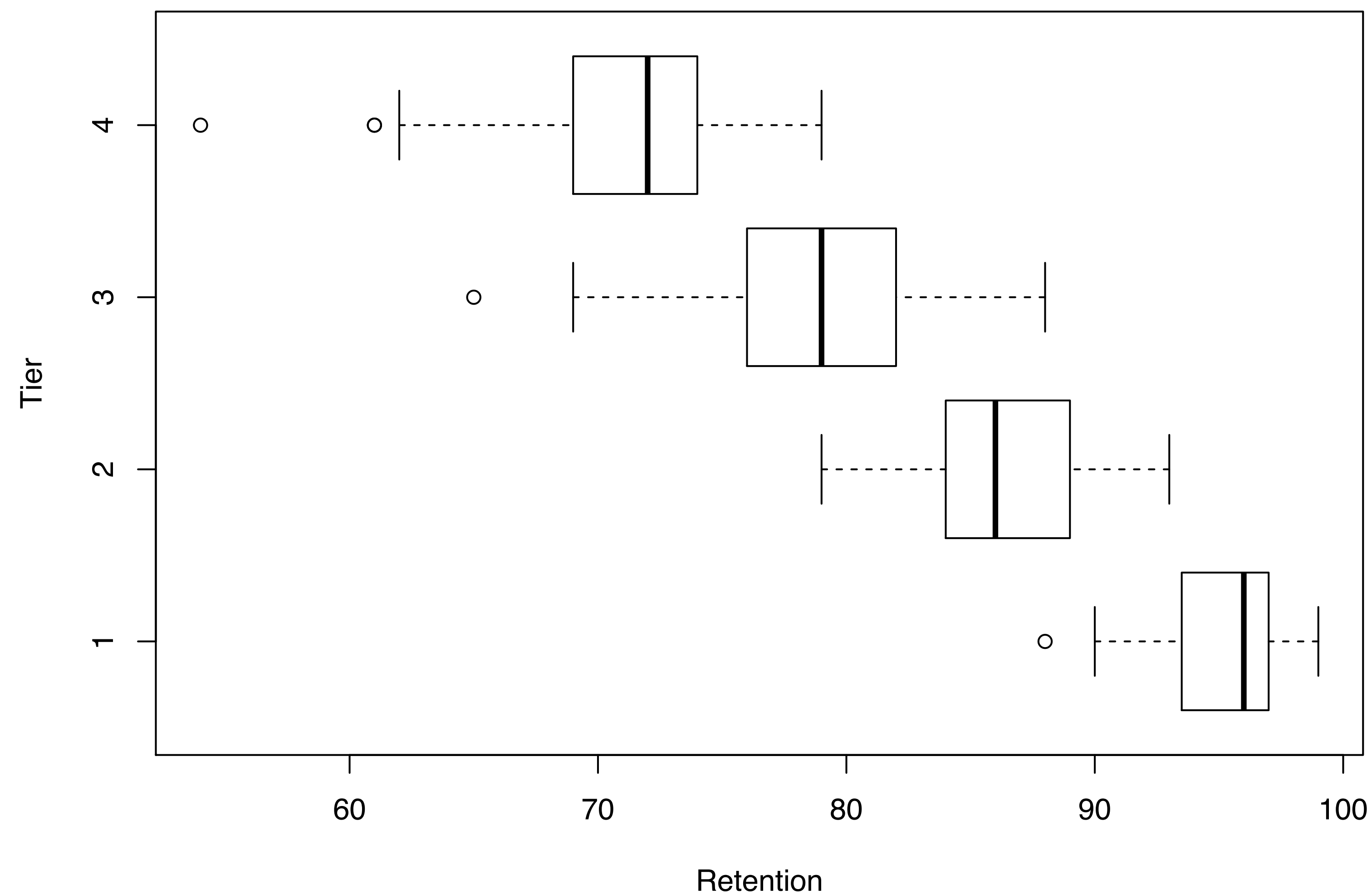
EDA

```
> b.output = boxplot(Retention ~ Tier, data=college,
horizontal=TRUE, ylab="Tier", xlab="Retention")
> b.output$stats
     [,1] [,2] [,3] [,4]
[1,] 90.0   79   69   62
[2,] 93.5   84   76   69
[3,] 96.0   86   79   72
[4,] 97.0   89   82   74
[5,] 99.0   93   88   79
attr(,"class")
        1
"integer"
```

1st quartiles

← **medians**

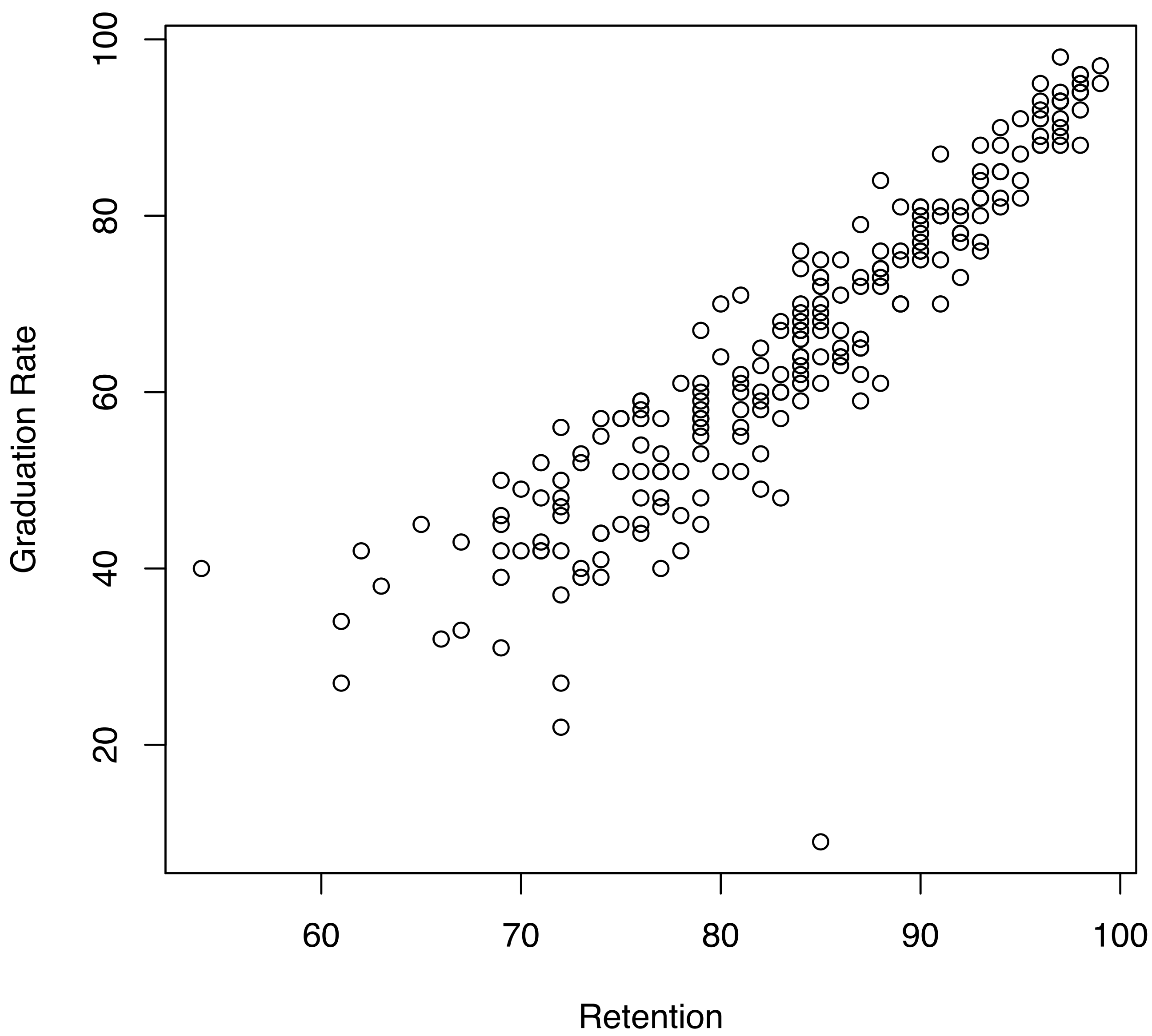
3rd quartiles



```
> b.output$out
[1] 88 65 61 61 54
> b.output$group
[1] 1 3 4 4 4
```

Info about outliers

- Relationships between 2 variables: Resistant line (robust regression)



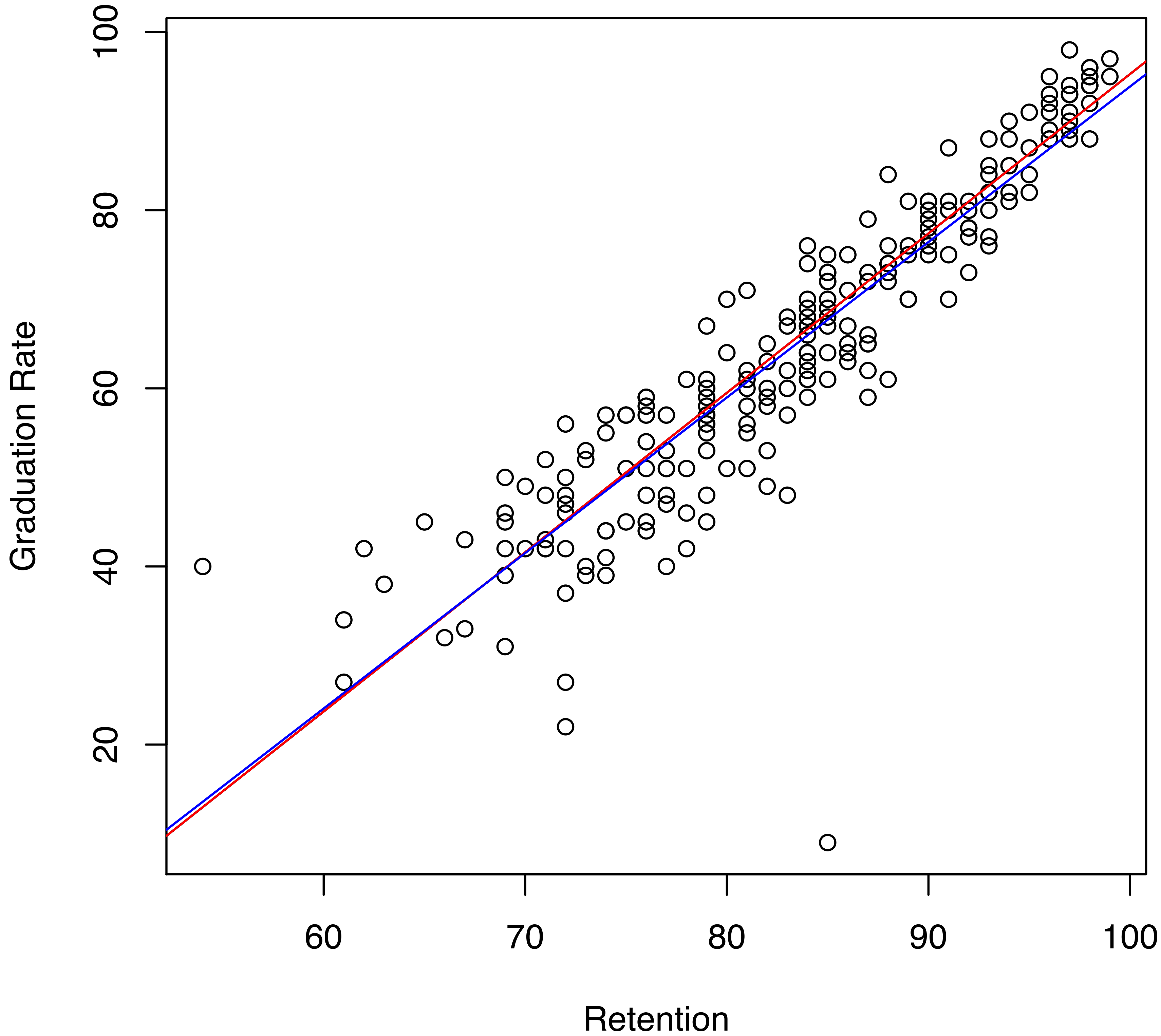
- Tukey's resistant line is implemented in `line`. Robust to outliers. It divides plot into 3 regions (left, middle, right), computes "resistant" summary points for each region, and finds a line from summary points.

```
> fit = line(Retention, Grad.rate)
> coef(fit)
[1] -83.657895   1.789474
```

For every 1% increase in retention the average graduation rate increases by 1.79%

```
> abline(coef(fit),col='red')
> coef(lm(Grad.rate~Retention))
(Intercept)   Retention
 -80.702851    1.745892
> abline(lm(Grad.rate~Retention),col='blue')
```

EDA



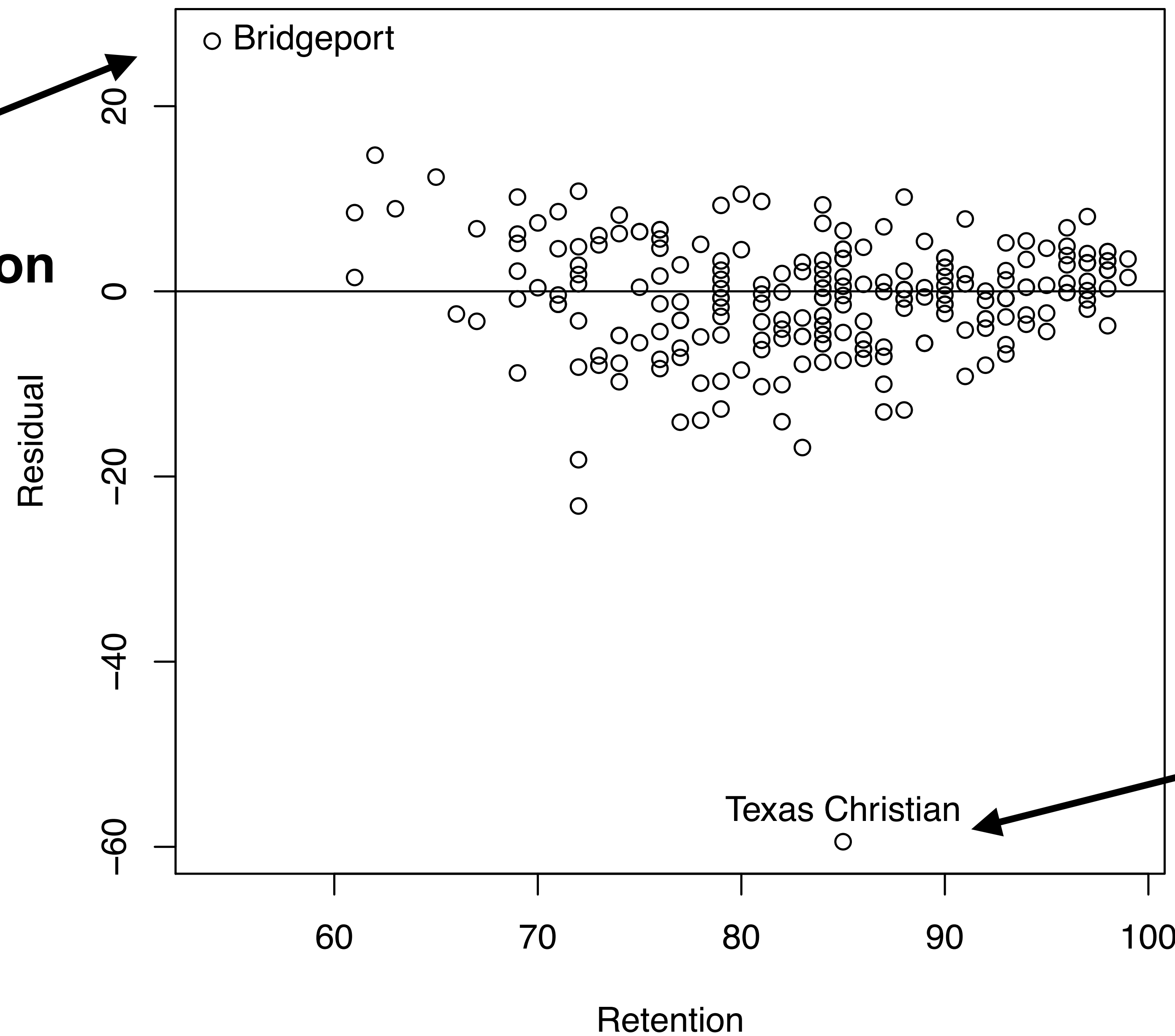
red: robust
blue: lse

EDA

- Residuals: identifying patterns & outliers

```
> plot(Retention, fit$residuals, xlab="Retention",
ylab="Residual")
> abline(h=0)
> identify(Retention, fit$residuals, n=2, labels=School)
[1] 109 212
```

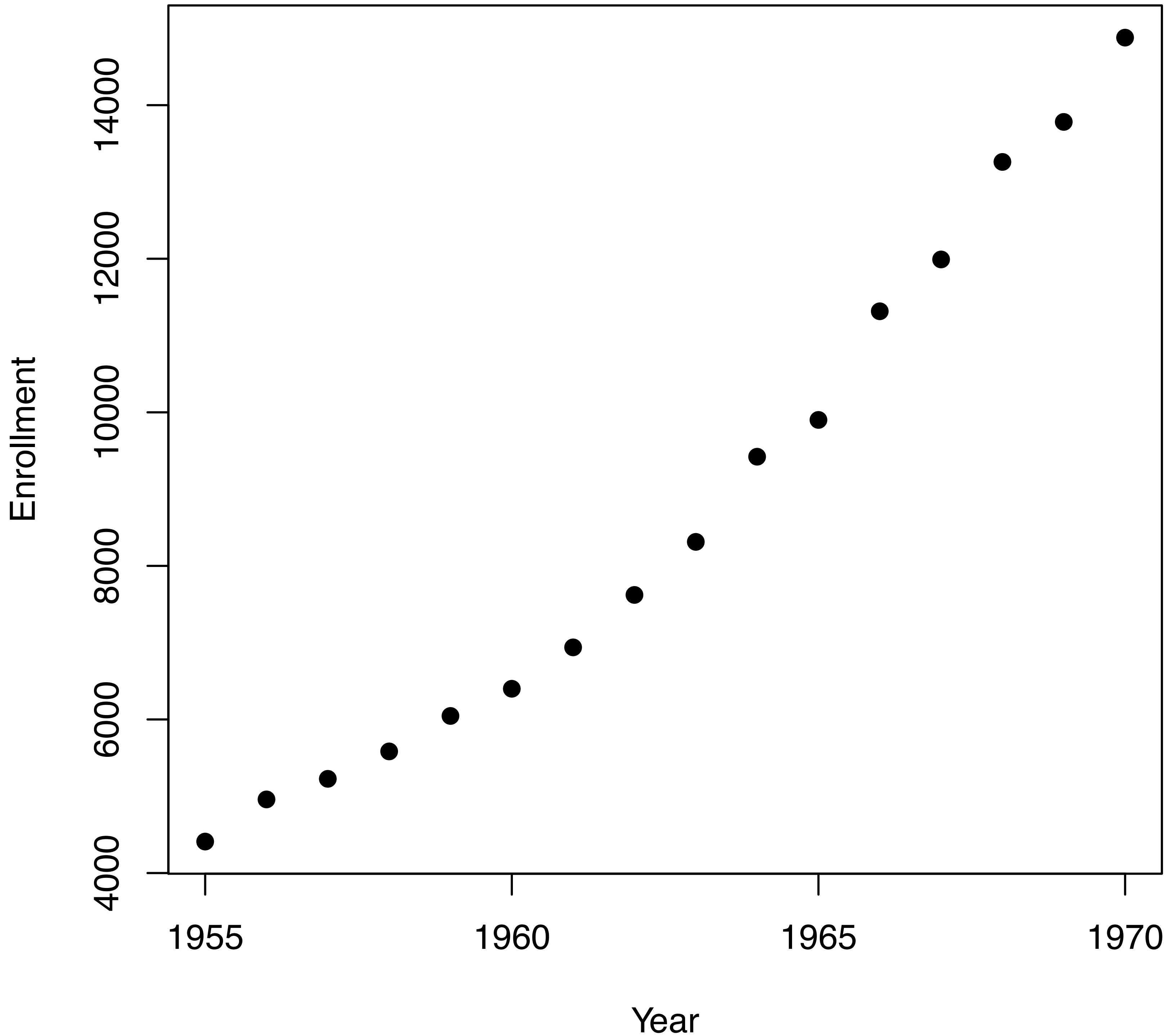
Graduation rate high given retention percentage



Graduation rate low given retention percentage

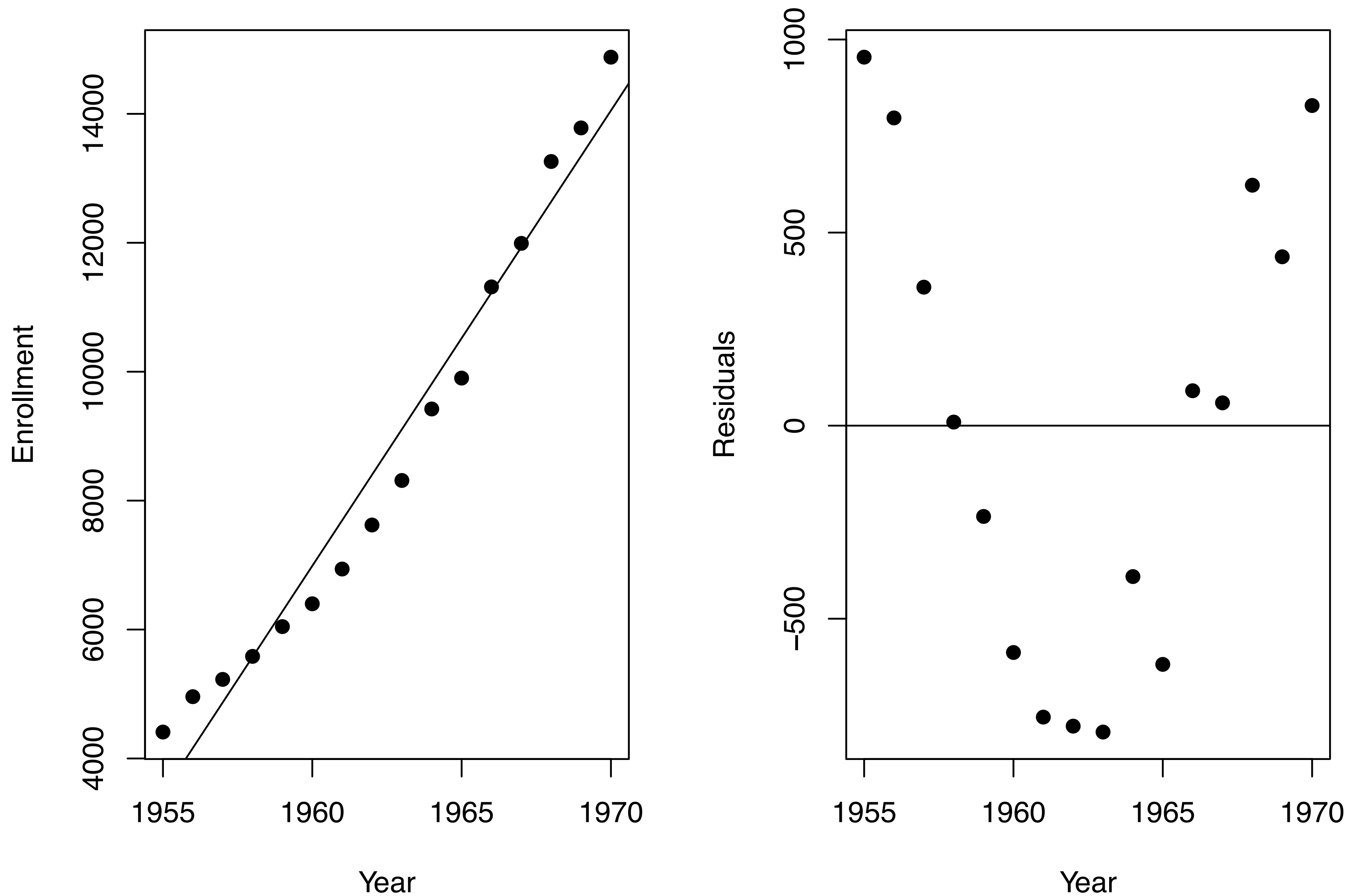
Reexpression

Example: The dataset bgsu.txt contains the enrollment counts for Bowling Green State University from 1955 to 1970.



EDA

- We fit a linear model and look at the residuals plot:




```
> par(mfrow=c(1,2))
> fit = lm(Enrollment ~ Year, data=bgsu)
> plot(Year, Enrollment, pch=19)
> abline(fit)
> plot(Year, fit$residuals,xlab="Year",ylab="Residuals",pch=19)
> abline(h=0)
```

- We now consider a model of the form:

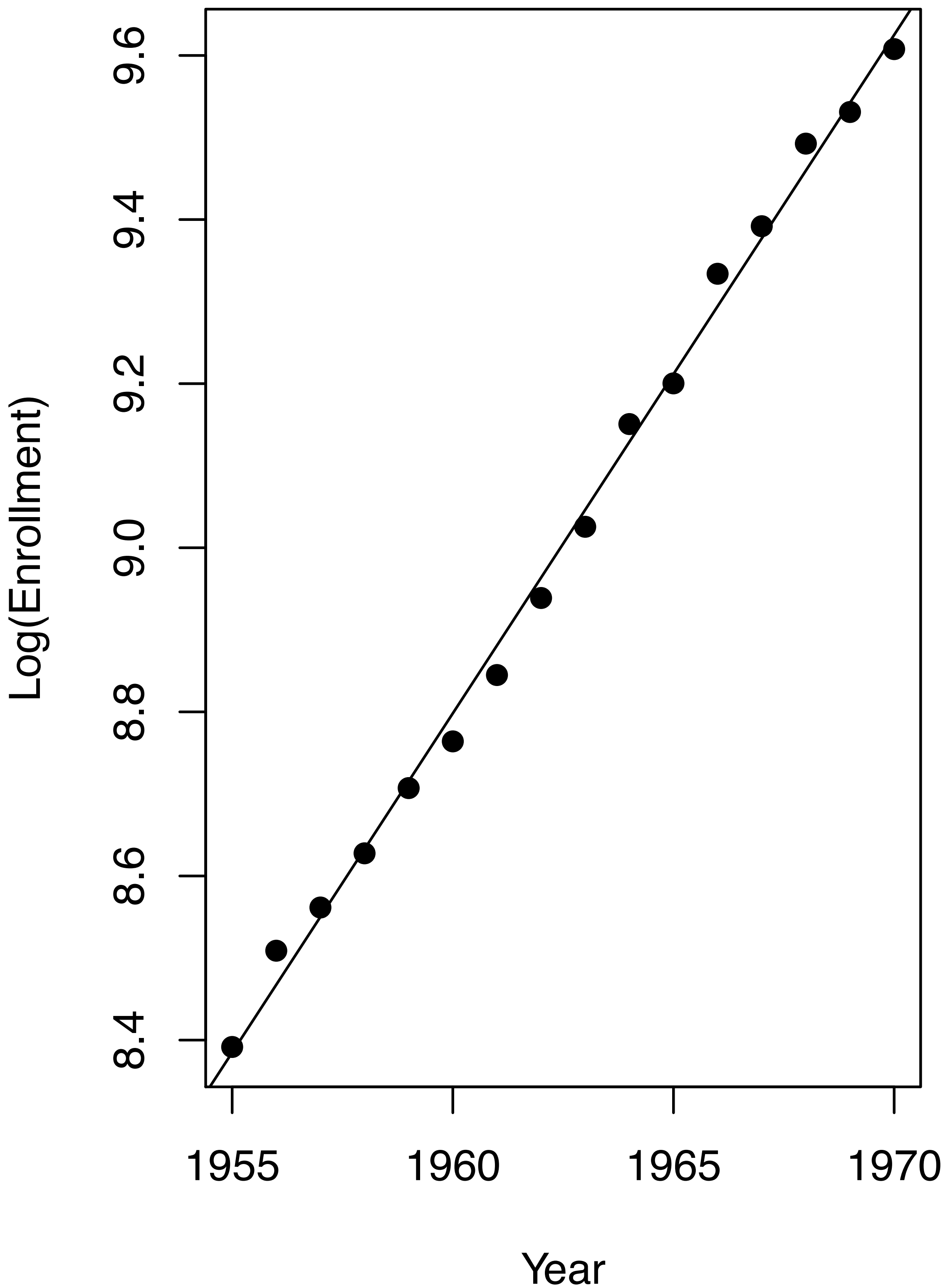
$$\text{Enrollment} = a \exp(b\ \text{Year})$$

Taking the log:

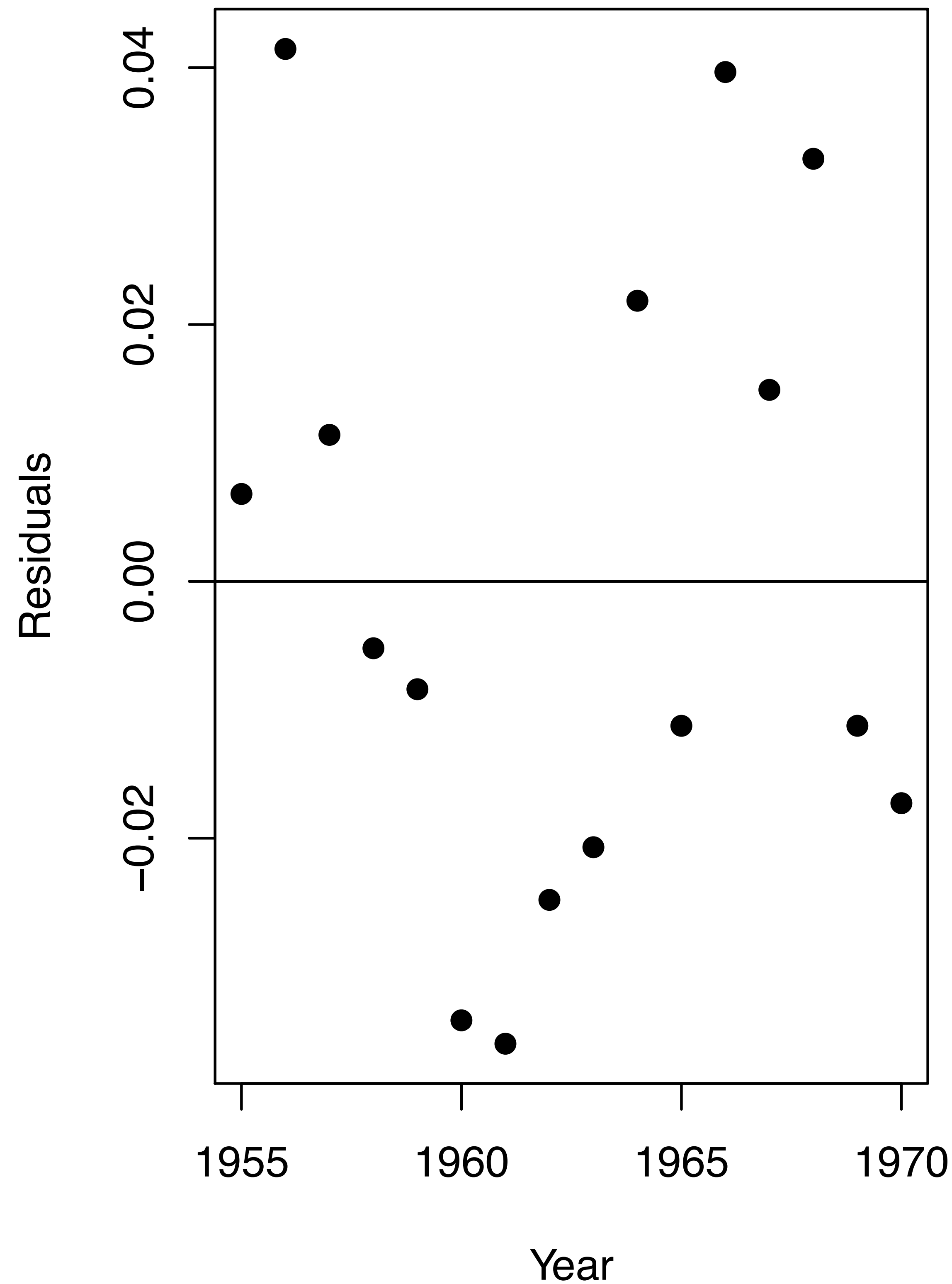
$$\log(\text{Enrollment}) = \log a + b\ \text{Year}$$

```
> bgsu$log.Enrollment = log(bgsu$Enrollment)
> attach(bgsu)
> par(mfrow=c(1,2))
> plot(Year, log.Enrollment, ylab="Log(Enrollment)",pch=19)
> fit2 = lm(log.Enrollment ~ Year, data=bgsu)
> fit2$coef
  (Intercept)          Year
-153.25703366    0.08268126
> abline(fit2)
> plot(Year, fit2$residuals,ylab="Residuals",pch=19)
> abline(h=0)
```

EDA



Log(Enrollment)
9.6
9.4
9.2
9.0
8.8
8.6
8.4
1955
1960
1965
1970
Year



Residuals
0.04
0.02
0.00
−0.02
1955
1960
1965
1970
Year