Unit 2 Live Session

W203 Instructional Team

Exploratory Data Analysis

![title](data.png)

Class Announcements

1. Lab 1 Assignment 2. Announcement 4. Announcement

1 Pre Class Exercise Responses

** 1.1 ** Sample Student response from PCE 2 Pasted Here

For Example: "I think that assuming women above the age of 35 are finished with both having kids and with education is flawed. In my graduate school studies (MIDS and MBA), I have encountered women over the age of 35. Moreover, due to technologies that exist and are used frequently (freezing eggs/embryos), women are having kids in their late 30s and early 40s."

Follow-up question:

For Example: "If it's true that many women in the study have not finished having children or have not finished their education, how might this be reflected in the observed relationship between fertility and education?"

In [ ]:

```

```

** 1.2 ** Sample student response from PCE 2 pasted here

For Example: "I found the binning approach for Number of Children by Educational Attainment to be the most interesting. My initial reaction was that I've typically encountered the education question on surveys phrased as "highest level of education completed". I think that is to get clarity/specificity on the educational attainment rather than inferring the level of attainment from a continuous data set. For example, it's not clear to me that all of the participants in this survey were educated in the United States and that our binning assumptions are accurate. How many of the respondents had completed a GED instead of a high school diploma? What about an Associate's degree? Also, I think that socio-economic status is a major factor that isn't adequately addressed here."

Follow-up question:

For Example: "How do you decide when to bin levels of a variable together?"

In [ ]:

** 1.3 ** Sample student response from PCE 2 pasted here

James Bauserman: There is a lot of commentary on how we have relatively few observations at lower levels of education and whether these observations might be erroneous. However, no attempt is made to look at the data excluding these observations, for example to see if those observations are clustered at any particular age groups or have other distinguishing features.

Follow-up question:

For Example: "Can you trust the data points you see with 0-5 years of education?"

In [ ]:

Notes on Preclass

> Way in which schooling is related to fertility (# of children)
>
> - Do women delay child bearing in order to gain education?
> - Do women delay education to bare children?

> Upper and Lower Limits to age range

> - Uniform Distribution or ages between 35 and 44

> - Lower limit of 35 chosen on the assumption that women have had all their children by 35. (Suppose that women are not done having children by 35)

- If women delay child bearing for education, there will be some women who have high education with a smaller than final count of children, this will reduce the averge number of children in the highly educated cohort and lead to a more negative correlation in the data than the true effect.

- If women delay education for child bearing (and suppose that the true correlation is negative) you will have women in the lower education group (who have finished having their reduced number of children) who will reduce the average number of children in that group which leads to a more positive correlation than the true effect.

- Upper limit of 44 to limit analysis to a given generational cohort.
  - Social norms may change; cultural and social attitude toward education may be different for different generations. We don't want to include this effect in our analysis since we will then be measuring two effects without a way to isolate either.

- If women were are less likely to pursue post grad education in older cohorts regardless of the number of children, we may find no effect, when one is present.

- Could maybe do a generational by cohort analysis for robustness.

Missing Values/Low Education Values

- Is it appropriate to remove these, does it make sense that anyone would have less than 5 year of education.

- Is it possible that they thought that were giving year of post secondary education? From the means plot is it more plausible?

- bins that paul uses,

```
+(0,11): Some Primary School
```

+(12,15): High School Grad

+(15, ∞∞): College Grad

2 Exploratory Data Analysis Review

Things to Consider

- Examine each variable's characteristics and distribution. Are there any strange features of the data?
- Consider transforming the variable, why and how would you do this?
- Tell a story about each variable. In words and in context, what is this graph telling you? Is it suprising/interesting or not in some way ?
- Are there any outliers to the data? Could it be an error or just some rare event?
- Anyone with a reasonable level of programming skill can write a program which pumps out figures and sample characteristics with no context, your job is to provide both!
- No data dumps
- Practice forming a research question about the data population, with your EDA, i.e. this feature in the data is not what I expect, everyone else is ignoring it as an error or uninteresting phenomenon, I want to explore it further and this is why. This is seriously how Nobel prizes get awarded.
- Always look for missing data and data with wrong types

3 Data Exercise

You are to begin an exploratory analysis with the objective of understanding how the price of a home relates to neighborhood characteristics, with an emphasis on crime.

In [1]:

```
Boston = read.csv("Boston_w203.csv")
library(car)
```

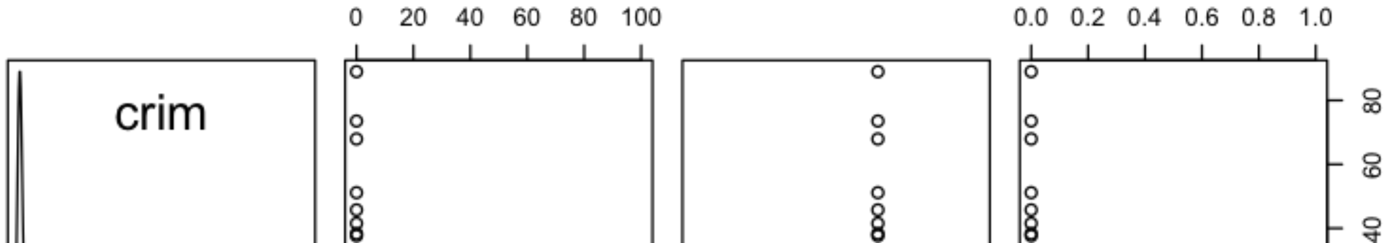
Variable Name	Description
crim	per capita crime rate by town
zn	proportion of residential land zoned for lots over 25,000 sq.ft.
indus	proportion of non-retail business acres per town
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
nox	nitrogen oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted mean of distances to five Boston employment centres
rad	index of accessibility to radial highways
tax	full-value property-tax rate per $10,000
ptratio	pupil-teacher ratio by town
black	$1000(Bk - 0.63)^2$ $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
lstat	lower status of the population (percent)
medv	median value of owner-occupied homes in $1000

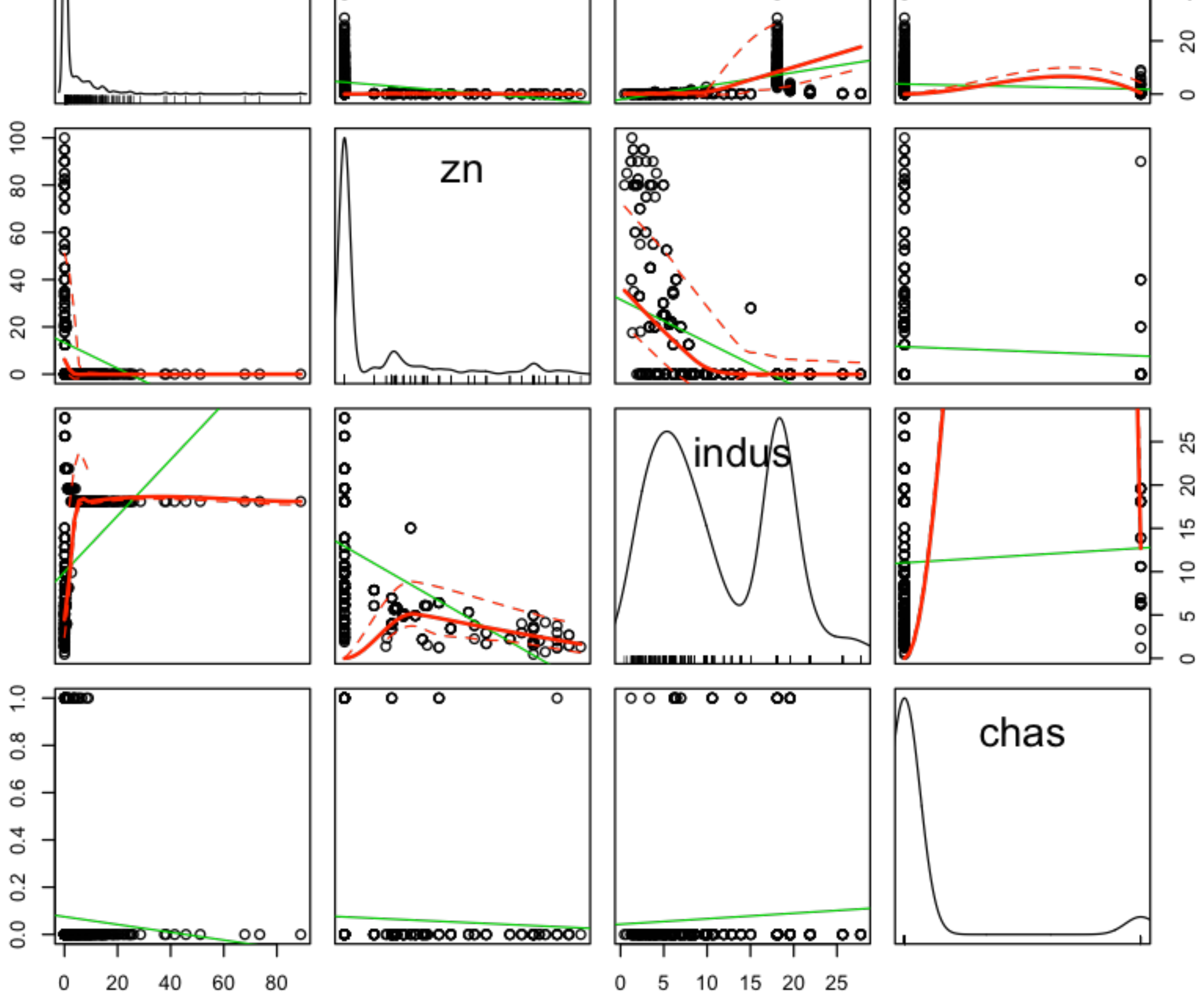
3.1 Generate a scatterplot matrix for all metric variables. Take a few minutes to draw as many insights as you can about the relationships in the data.

In [2]:

```
scatterplotMatrix(Boston[,2:5])
```

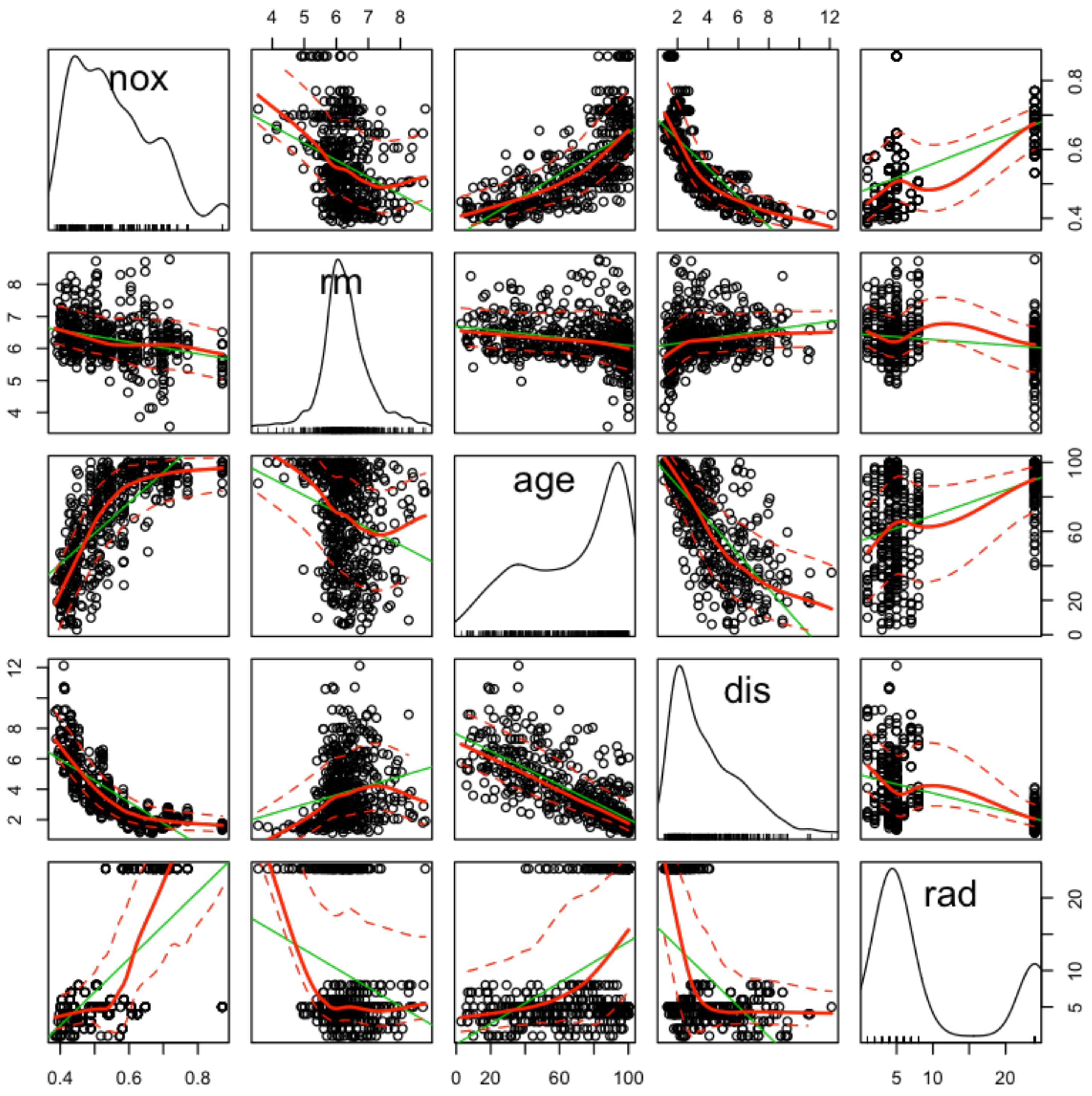
```
Warning message in smoother(x, y, col = col[2], log.x = FALSE, log.y =
FALSE, spread = spread, :
“could not fit smooth”Warning message in smoother(x, y, col = col[2],
log.x = FALSE, log.y = FALSE, spread = spread, :
“could not fit smooth”Warning message in smoother(x, y, col = col[2],
log.x = FALSE, log.y = FALSE, spread = spread, :
“could not fit smooth”Warning message in smoother(x, y, col = col[2],
log.x = FALSE, log.y = FALSE, spread = spread, :
“could not fit smooth”
```




zn
indus
chas
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
0
20
0
5
10
15
20
25
0
20
40
60
80
0
5
10
15
20
25

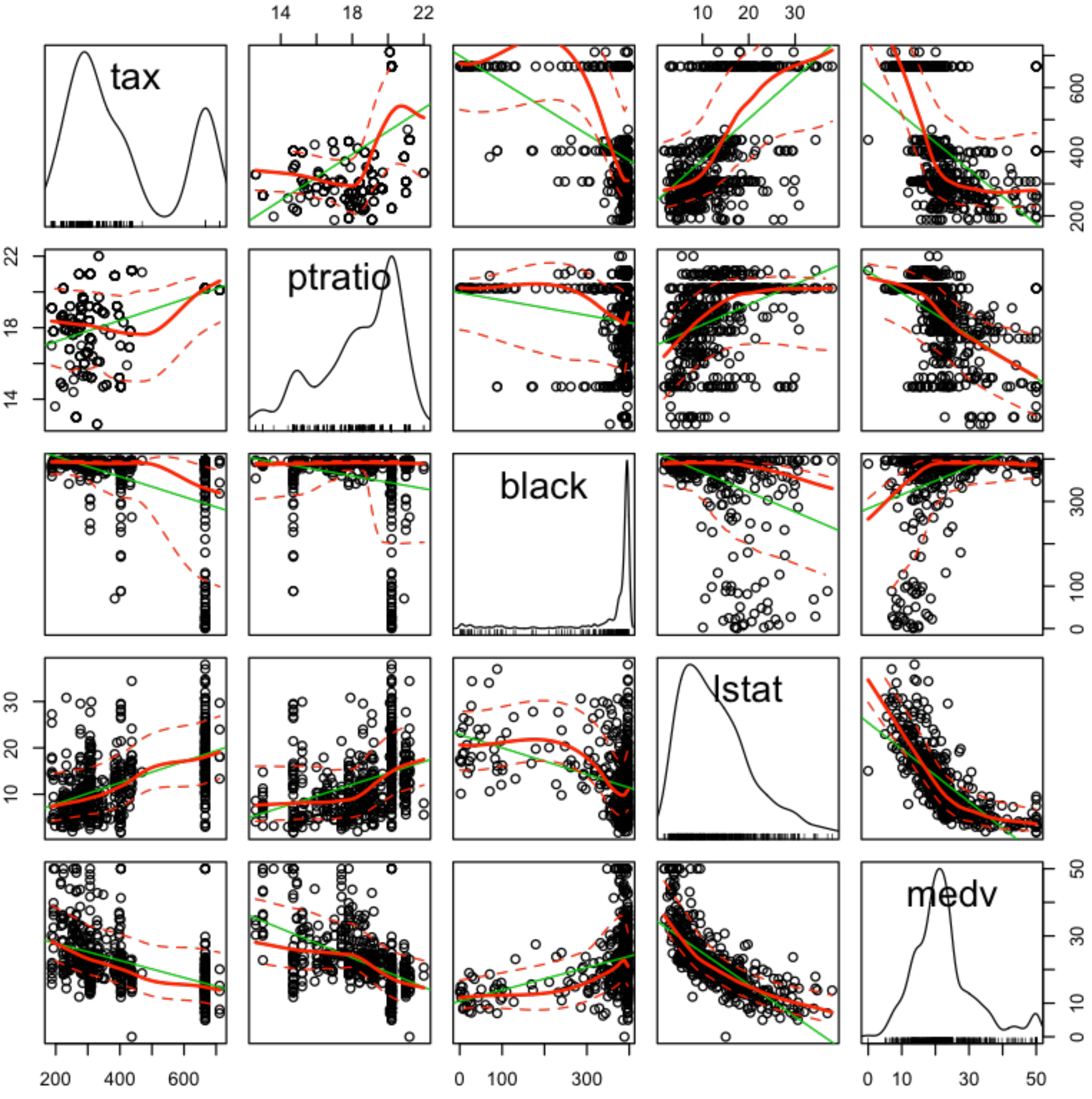
In [3]:

```
#options(repr.plot.height = 15, repr.plot.width = 15, repr.plot.pointsize = 22)
scatterplotMatrix(Boston[,6:10])
```



In [4]:

```
#options(repr.plot.height = 15, repr.plot.width = 15, repr.plot.pointsize = 22)
scatterplotMatrix(Boston[,11:15])
```

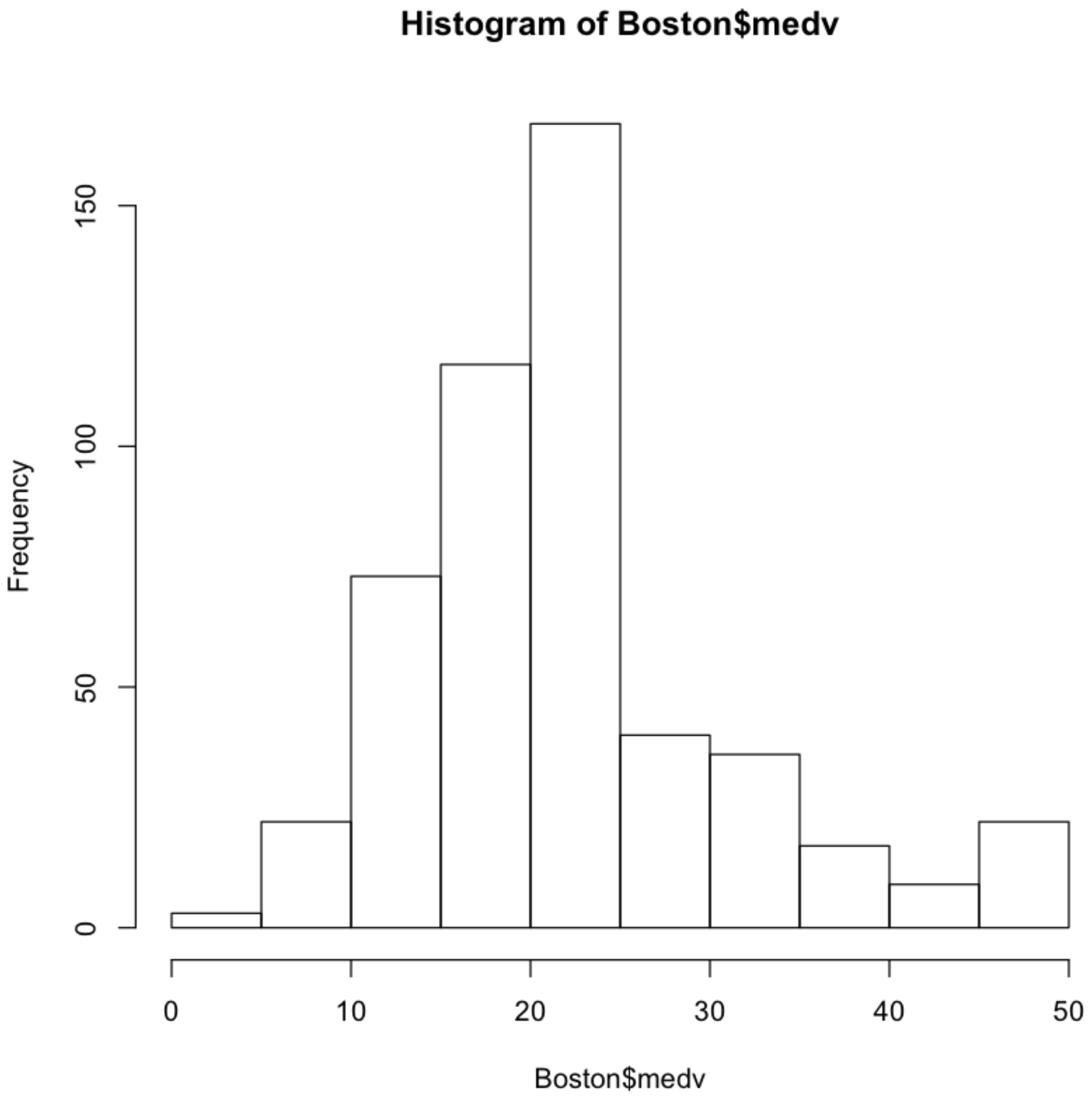


** 3.2 ** Examine the main output variable, medv. Comment on any unusual values you find, and any features that might be important for statistical modeling.

In [5]:

```
hist(Boston$medv)
summary(Boston$medv)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00   16.85   21.20   22.50   25.00   50.00 
```



Having a minimum value median home price of zero is nonsensical, lets take a look at the data points there are with this value

In [6]:

```
Boston[Boston$medv == 0,]
```

	X	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat
134	134	0.32982	0	21.89	0	0.624	5.822	95.4	2.4699	4	437	21.2	388.69	15.03

Looks like this is actually a missing value since, if we have a look at the summary of the data set the other variable values are reasonable.

In [7]:

```
summary(Boston)
```

```
       X              crim                zn             indus
 Min.   :  1.0   Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46
 1st Qu.:127.2   1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19
 Median :253.5   Median : 0.25651   Median :  0.00   Median : 9.69
 Mean   :253.5   Mean   : 3.61352   Mean   : 11.36   Mean   :11.14
 3rd Qu.:379.8   3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10
 Max.   :506.0   Max.   :88.97620   Max.   :100.00   Max.   :27.74
      chas              nox               rm             age
 Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90
 1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02
 Median :0.00000   Median :0.5380   Median :6.208   Median : 77.50
 Mean   :0.06917   Mean   :0.5547   Mean   :6.285   Mean   : 68.57
 3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08
 Max.   :1.00000   Max.   :0.8710   Max.   :8.780   Max.   :100.00
      dis              rad              tax           ptratio
 Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60
 1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40
 Median : 3.207   Median : 5.000   Median :330.0   Median :19.05
 Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46
 3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20
 Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00
     black            lstat            medv
 Min.   :  0.32   Min.   : 1.73   Min.   : 0.00
 1st Qu.:375.38   1st Qu.: 6.95   1st Qu.:16.85
 Median :391.44   Median :11.36   Median :21.20
 Mean   :356.67   Mean   :12.65   Mean   :22.50
 3rd Qu.:396.23   3rd Qu.:16.95   3rd Qu.:25.00
 Max.   :396.90   Max.   :37.97   Max.   :50.00
```

As a result we will recode this value as NA

In [8]:

```
Boston$medv[Boston$medv==0] = NA
Boston[134,]
```

	X	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat
134	134	0.32982	0	21.89	0	0.624	5.822	95.4	2.4699	4	437	21.2	388.69	15.03

Next, Having a median value of exactly 50 is weird, we may be concerned that this is a top code lets have a look at how many data points have them

In [9]:

```
Boston[Boston$medv == 50,]
```

	X	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lsta
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
162	162	1.46336	0	19.58	0	0.6050	7.489	90.8	1.9709	5	403	14.7	374.43	1.73
163	163	1.83377	0	19.58	1	0.6050	7.802	98.2	2.0407	5	403	14.7	389.61	1.92
164	164	1.51902	0	19.58	1	0.6050	8.375	93.9	2.1620	5	403	14.7	388.45	3.32
167	167	2.01019	0	19.58	0	0.6050	7.929	96.2	2.0459	5	403	14.7	369.30	3.70
187	187	0.05602	0	2.46	0	0.4880	7.831	53.6	3.1992	3	193	17.8	392.63	4.45
196	196	0.01381	80	0.46	0	0.4220	7.875	32.0	5.6484	4	255	14.4	394.23	2.97
205	205	0.02009	95	2.68	0	0.4161	8.034	31.9	5.1180	4	224	14.7	390.55	2.88
226	226	0.52693	0	6.20	0	0.5040	8.725	83.0	2.8944	8	307	17.4	382.00	4.63
258	258	0.61154	20	3.97	0	0.6470	8.704	86.9	1.8010	5	264	13.0	389.70	5.12
268	268	0.57834	20	3.97	0	0.5750	8.297	67.0	2.4216	5	264	13.0	384.54	7.44
284	284	0.01501	90	1.21	1	0.4010	7.923	24.8	5.8850	1	198	13.6	395.52	3.16
369	369	4.89822	0	18.10	0	0.6310	4.970	100.0	1.3325	24	666	20.2	375.52	3.26
370	370	5.66998	0	18.10	1	0.6310	6.683	96.8	1.3567	24	666	20.2	375.33	3.73
371	371	6.53876	0	18.10	1	0.6310	7.016	97.5	1.2024	24	666	20.2	392.05	2.96
372	372	9.23230	0	18.10	0	0.6310	6.216	100.0	1.1691	24	666	20.2	366.15	9.53
373	373	8.26725	0	18.10	1	0.6680	5.875	89.6	1.1296	24	666	20.2	347.88	8.88

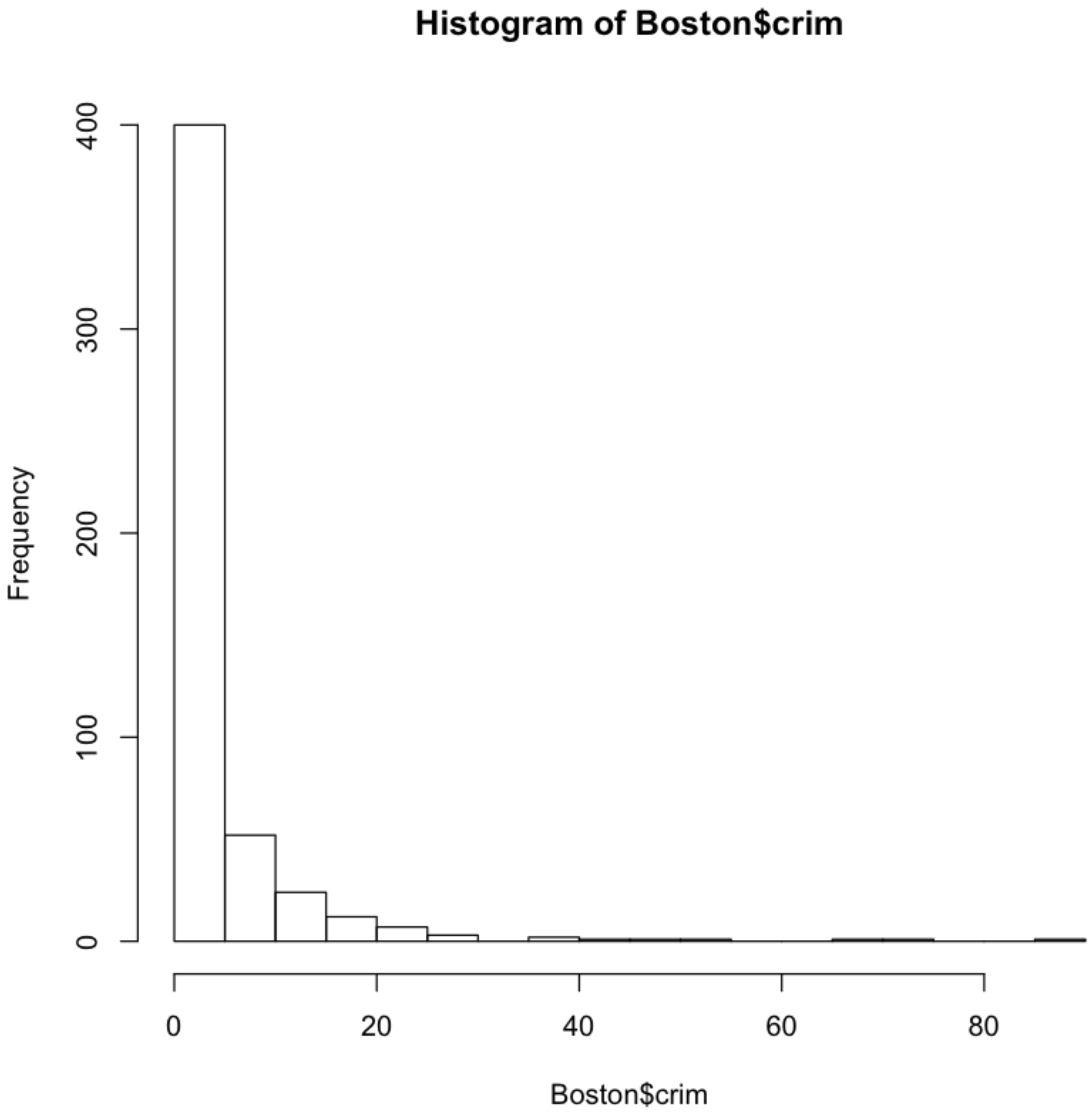
Looks like the value of 50 is a top code meaning that median home value in these towns is greater than or equal to 50

There is nothing more to do here other than just keep this fact in mind.

3.3 Examine the main independent variable of interest, crim. What transformation could you apply to this variable to aid in visualizing it? Comment on any unusual features you find.

In [10]:

```
#options(repr.plot.height = 8.5, repr.plot.width = 15, repr.plot.pointsize = 22)
hist(Boston$crim,breaks = 20)
```



In [11]:

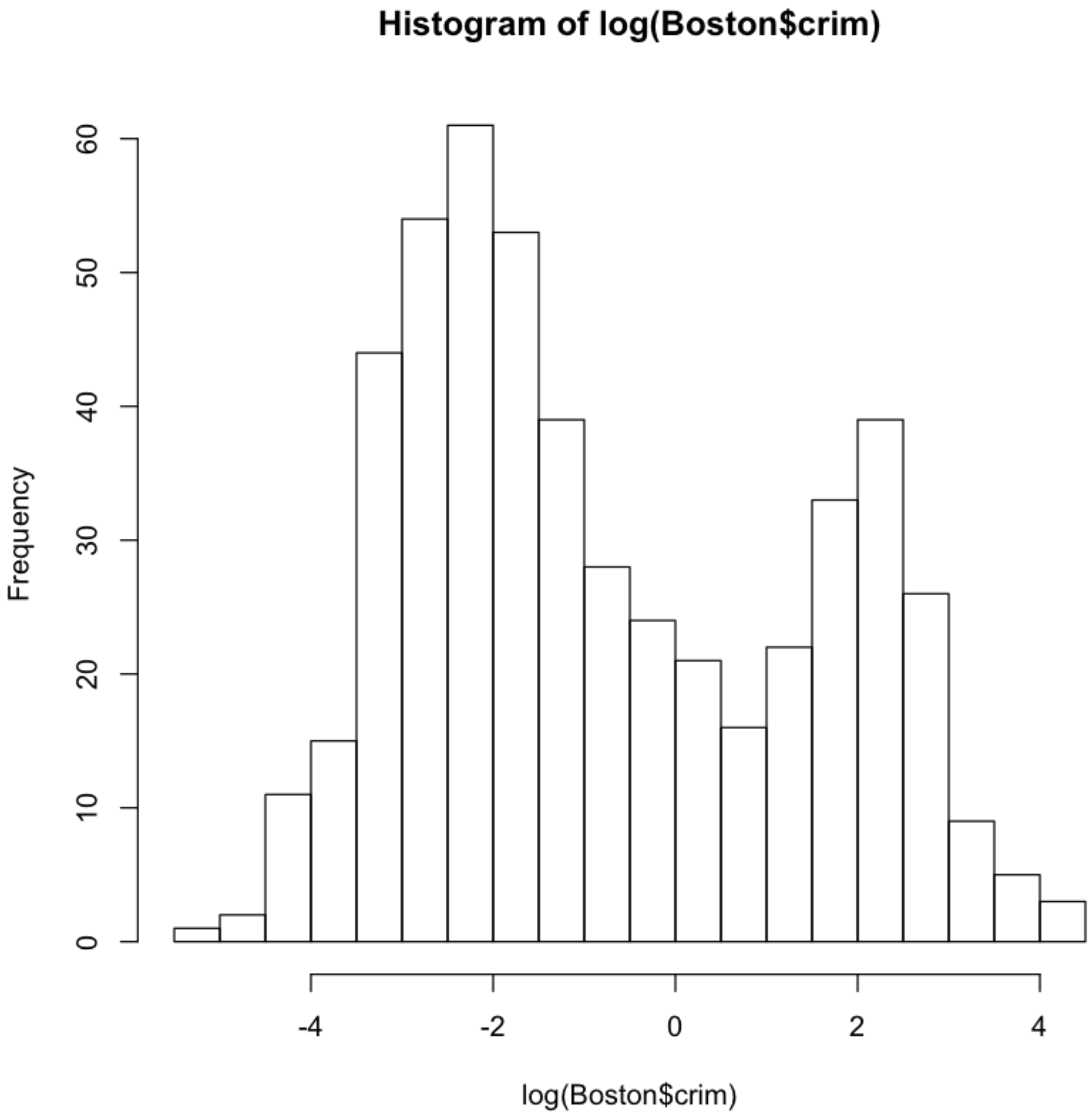
```
summary(Boston$crim)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
 0.00632  0.08204  0.25651  3.61352  3.67708 88.97620 
```

> We can see that the crim variable has is strictly positive with a heavy right skew, these are the kind of variables which are strong candidates for a log transformation

In [12]:

```
hist(log(Boston$crim), breaks = 15)
```

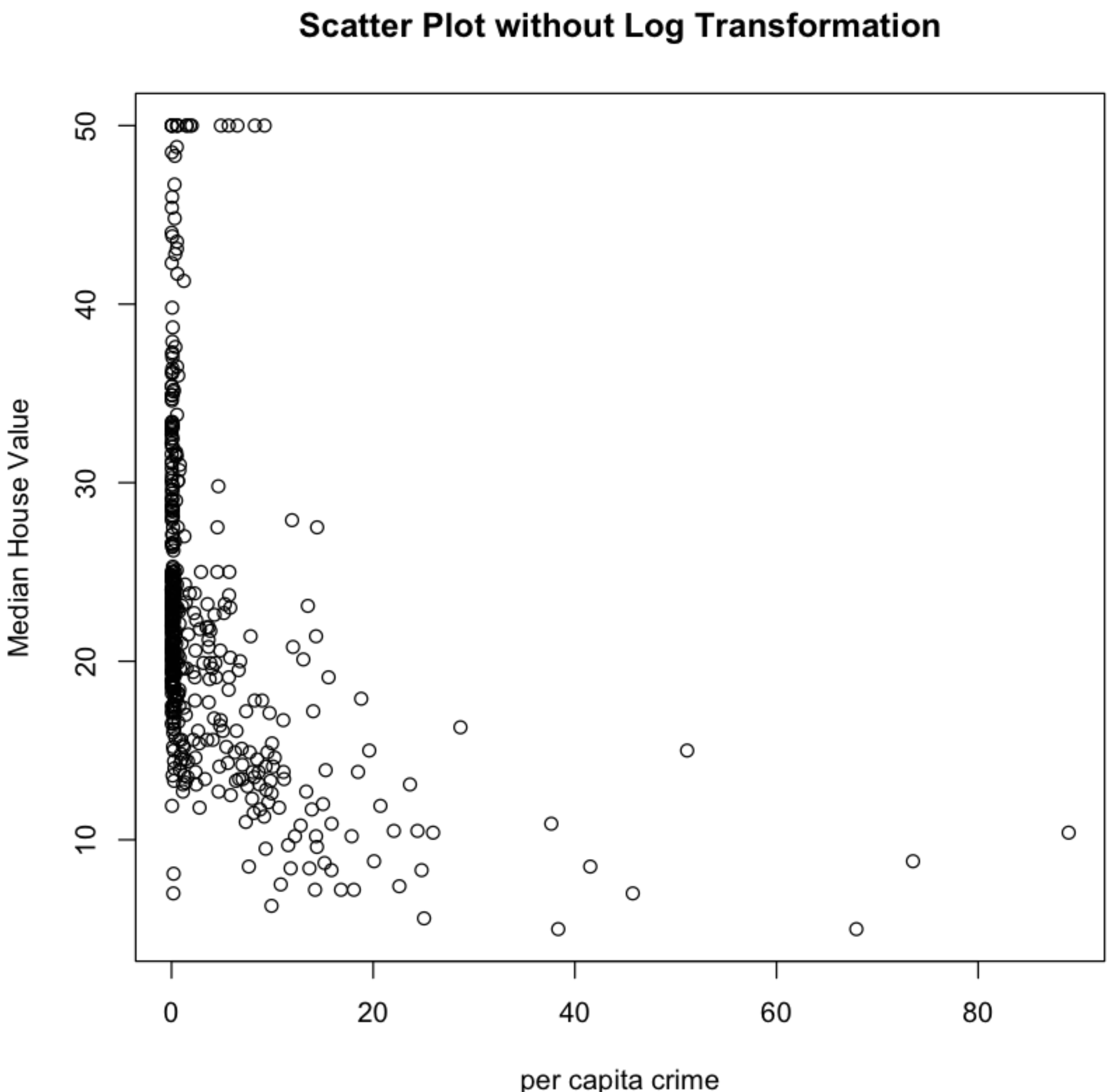


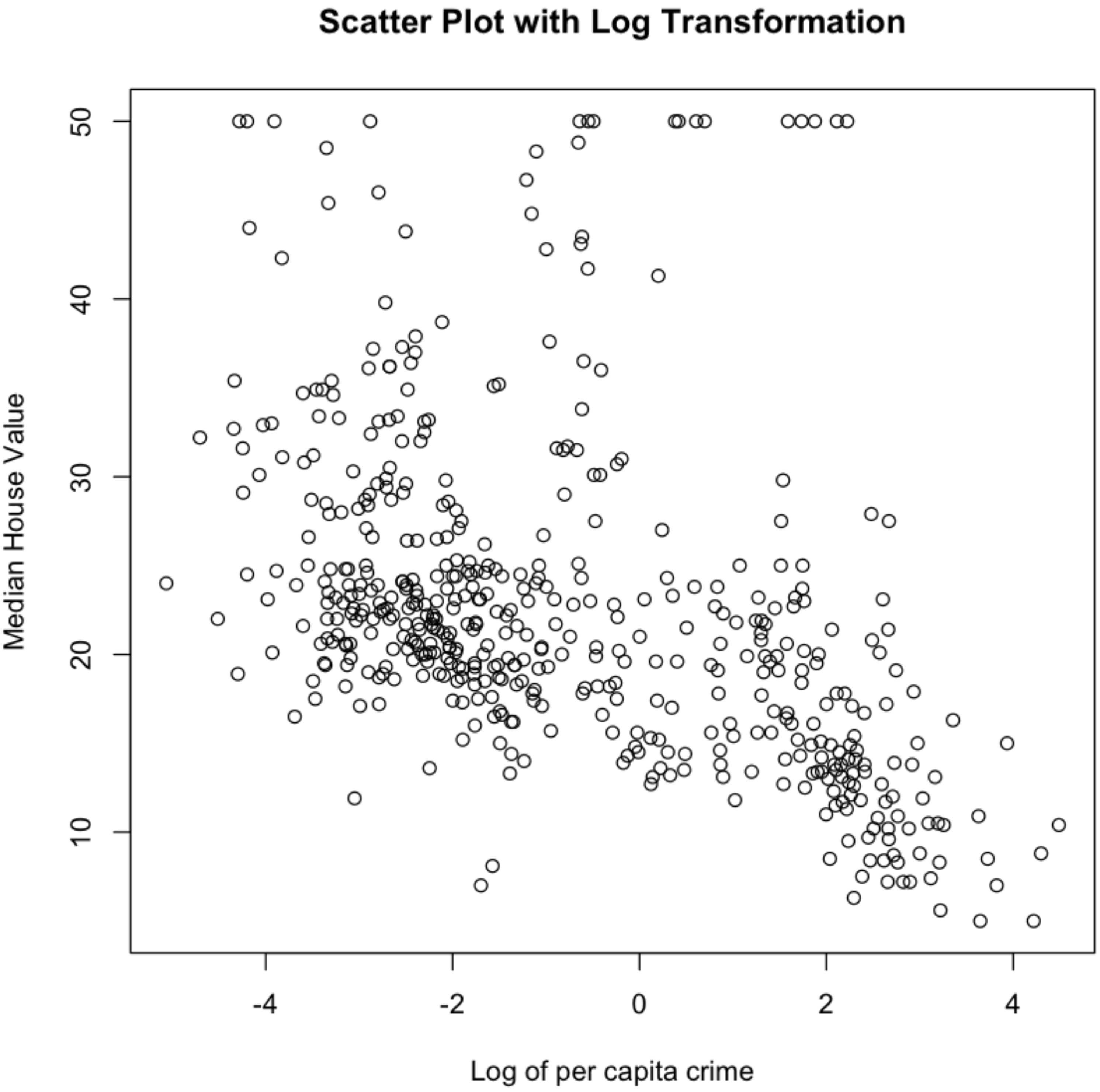
> We can see that the log transformation gives us a nice compact bi-modal distribution which as we will see much later in the class will help up identify linear relationships between variables.

** 3.4 ** Examine the bivariate relationship between medv and crime. What type of relationship do these variables have?

In [13]:

```
plot(Boston$crim, Boston$medv, main = "Scatter Plot without Log Transformation",
     col = 'black', xlab = 'per capita crime', ylab = 'Median House Value')
plot(log(Boston$crim),Boston$medv, main = "Scatter Plot with Log Transformation",
     col = 'black', xlab = 'Log of per capita crime', ylab = 'Median House Value')
```





We can see that there is a generally negative relationship between crime and median home value

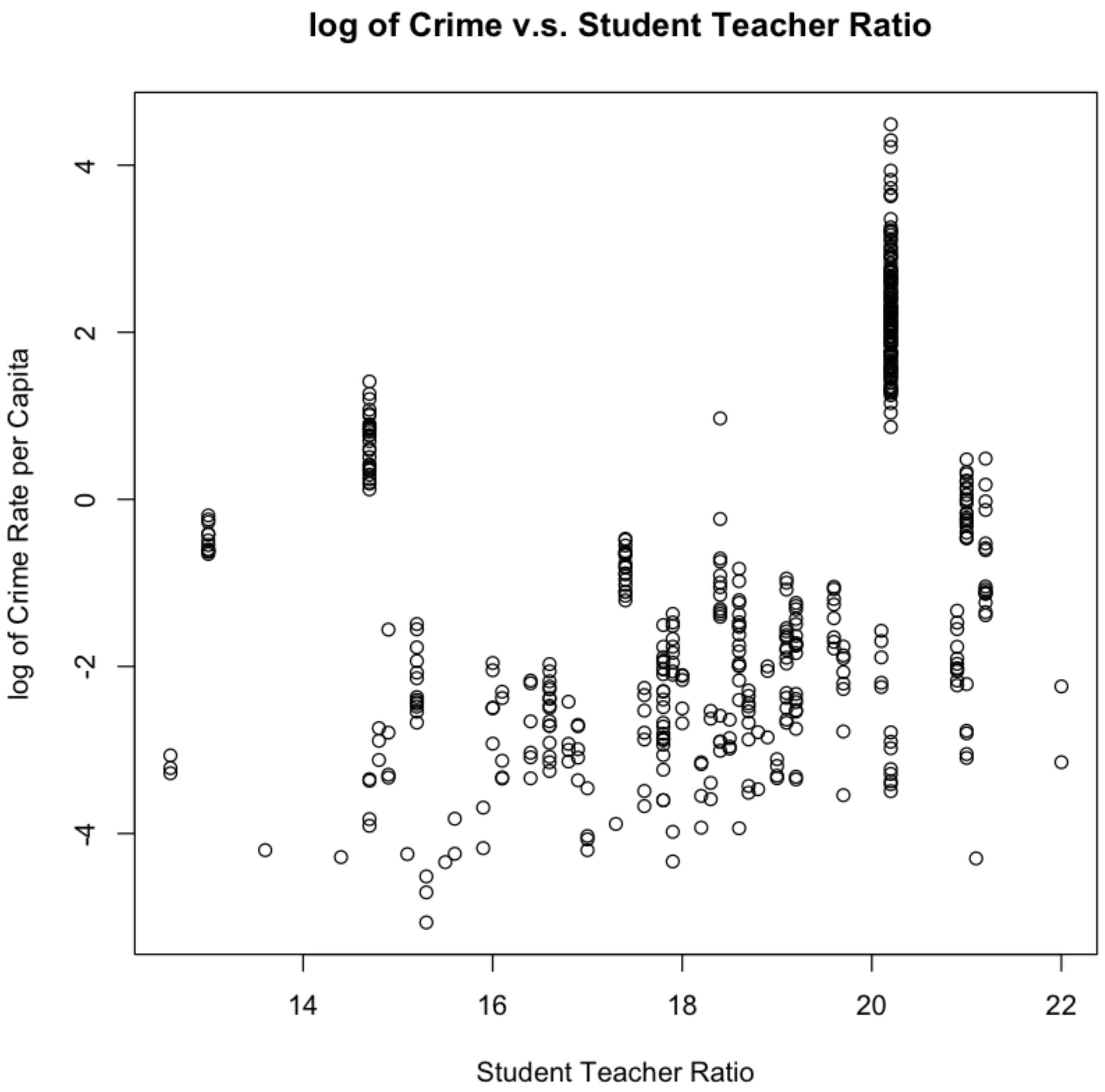
We might be a little critical of this because of a couple values with medv = 50 and crim value of between -1 and 3 which would throw of any linear fit we might want to do.

But we have to remember that these are top codes which might lead us to exclude them from a linear fitting procedure.

3.5 (As time permits) Continue your exploratory data analysis. Be prepared to share interesting findings with the class.

In [14]:

```
plot(Boston$ptratio,log(Boston$crim),
     main = ' log of Crime v.s. Student Teacher Ratio',
     xlab ='Student Teacher Ratio',
     ylab = 'log of Crime Rate per Capita' )
```



The above scatter plot makes us wonder why almost all the towns which have a log crim rate above 1 have a student teacher ratio of either 14.7 or 20.2

In [15]:

```
summary(Boston[log(Boston$crim) > 1,]$ptratio)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  14.70   20.20   20.20   19.96   20.20   20.20 
```

We know that crime and income are negatively correllated meaning that crime rates in lower income areas tend to be higher for various reasons

We also know that many places implement caps on student teacher ratios and areas which are low income are more likely to run up against these caps due to tight fiscal situations, thus this might be the story which explains this feature of the data

This particular story may not be true but it is important that when you find strange features in your data set to try to use out of sample (domain) information to understand/explain it.

You must be a story teller in data science.

Unit 2 Live Session

W203 Instructional Team

Exploratory Data Analysis

![title](data.png)

Class Announcements

1. Lab 1 Assignment 2. Announcement 4. Announcement

1 Pre Class Exercise Responses

** 1.1 ** Sample Student response from PCE 2 Pasted Here

For Example: "I think that assuming women above the age of 35 are finished with both having kids and with education is flawed. In my graduate school studies (MIDS and MBA), I have encountered women over the age of 35. Moreover, due to technologies that exist and are used frequently (freezing eggs/embryos), women are having kids in their late 30s and early 40s."

Follow-up question:

For Example: "If it's true that many women in the study have not finished having children or have not finished their education, how might this be reflected in the observed relationship between fertility and education?"

In [ ]:

** 1.2 ** Sample student response from PCE 2 pasted here

For Example: "I found the binning approach for Number of Children by Educational Attainment to be the most interesting. My initial reaction was that I've typically encountered the education question on surveys phrased as "highest level of education completed". I think that is to get clarity/specificity on the educational attainment rather than inferring the level of attainment from a continuous data set. For example, it's not clear to me that all of the participants in this survey were educated in the United States and that our binning assumptions are accurate. How many of the respondents had completed a GED instead of a high school diploma? What about an Associate's degree? Also, I think that socio-economic status is a major factor that isn't adequately addressed here."

Follow-up question:

For Example: "How do you decide when to bin levels of a variable together?"

In [ ]:

** 1.3 ** Sample student response from PCE 2 pasted here

James Bauserman: There is a lot of commentary on how we have relatively few observations at lower levels of education and whether these observations might be erroneous. However, no attempt is made to look at the data excluding these observations, for example to see if those observations are clustered at any particular age groups or have other distinguishing features.

Follow-up question:

For Example: "Can you trust the data points you see with 0-5 years of education?"

In [ ]:

Notes on Preclass

> Way in which schooling is related to fertility (# of children)
>
> - Do women delay child bearing in order to gain education?
> - Do women delay education to bare children?

> Upper and Lower Limits to age range

> - Uniform Distribution or ages between 35 and 44

> - Lower limit of 35 chosen on the assumption that women have had all their children by 35. (Suppose that women are not done having children by 35)

> - If women delay child bearing for education, there will be some women who have high education with a smaller than final count of children, this will reduce the averge number of children in the highly educated cohort and lead to a more negative correlation in the data than the true effect.

- If women delay education for child bearing (and suppose that the true correlation is negative) you will have women in the lower education group (who have finished having their reduced number of children) who will reduce the average number of children in that group which leads to a more positive correlation than the true effect.

- Upper limit of 44 to limit analysis to a given generational cohort.
  - Social norms may change; cultural and social attitude toward education may be different for different generations. We don't want to include this effect in our analysis since we will then be measuring two effects without a way to isolate either.

- If women were are less likely to pursue post grad education in older cohorts regardless of the number of children, we may find no effect, when one is present.

- Could maybe do a generational by cohort analysis for robustness.

Missing Values/Low Education Values

- Is it appropriate to remove these, does it make sense that anyone would have less than 5 year of education.

- Is it possible that they thought that were giving year of post secondary education? From the means plot is it more plausible?

- bins that paul uses,

```
+(0,11): Some Primary School
```

+(12,15): High School Grad

+(15, ∞): College Grad

2 Exploratory Data Analysis Review

Things to Consider

- Examine each variable's characteristics and distribution. Are there any strange features of the data?
- Consider transforming the variable, why and how would you do this?
- Tell a story about each variable. In words and in context, what is this graph telling you? Is it suprising/interesting or not in some way ?
- Are there any outliers to the data? Could it be an error or just some rare event?
- Anyone with a reasonable level of programming skill can write a program which pumps out figures and sample characteristics with no context, your job is to provide both!
- No data dumps
- Practice forming a research question about the data population, with your EDA, i.e. this feature in the data is not what I expect, everyone else is ignoring it as an error or uninteresting phenomenon, I want to explore it further and this is why. This is seriously how Nobel prizes get awarded.
- Always look for missing data and data with wrong types

3 Data Exercise

You are to begin an exploratory analysis with the objective of understanding how the price of a home relates to neighborhood characteristics, with an emphasis on crime.

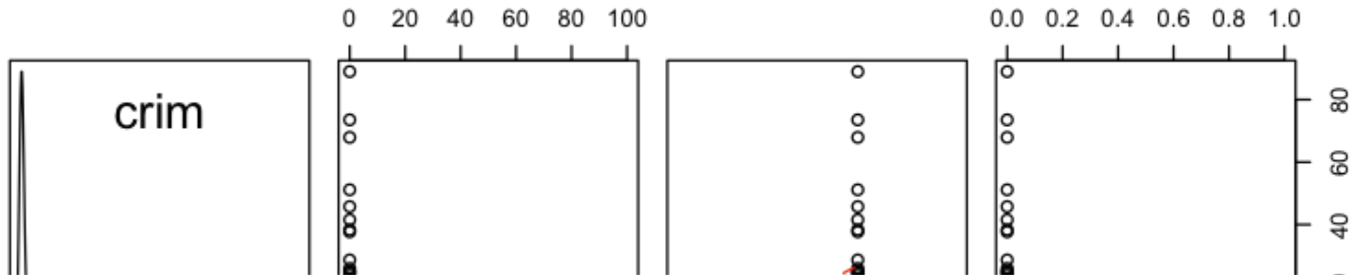
In [1]:

Variable Name	Description
crim	per capita crime rate by town
zn	proportion of residential land zoned for lots over 25,000 sq.ft.
indus	proportion of non-retail business acres per town
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
nox	nitrogen oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted mean of distances to five Boston employment centres
rad	index of accessibility to radial highways
tax	full-value property-tax rate per $10,000
ptratio	pupil-teacher ratio by town
black	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
lstat	lower status of the population (percent)
medv	median value of owner-occupied homes in $1000

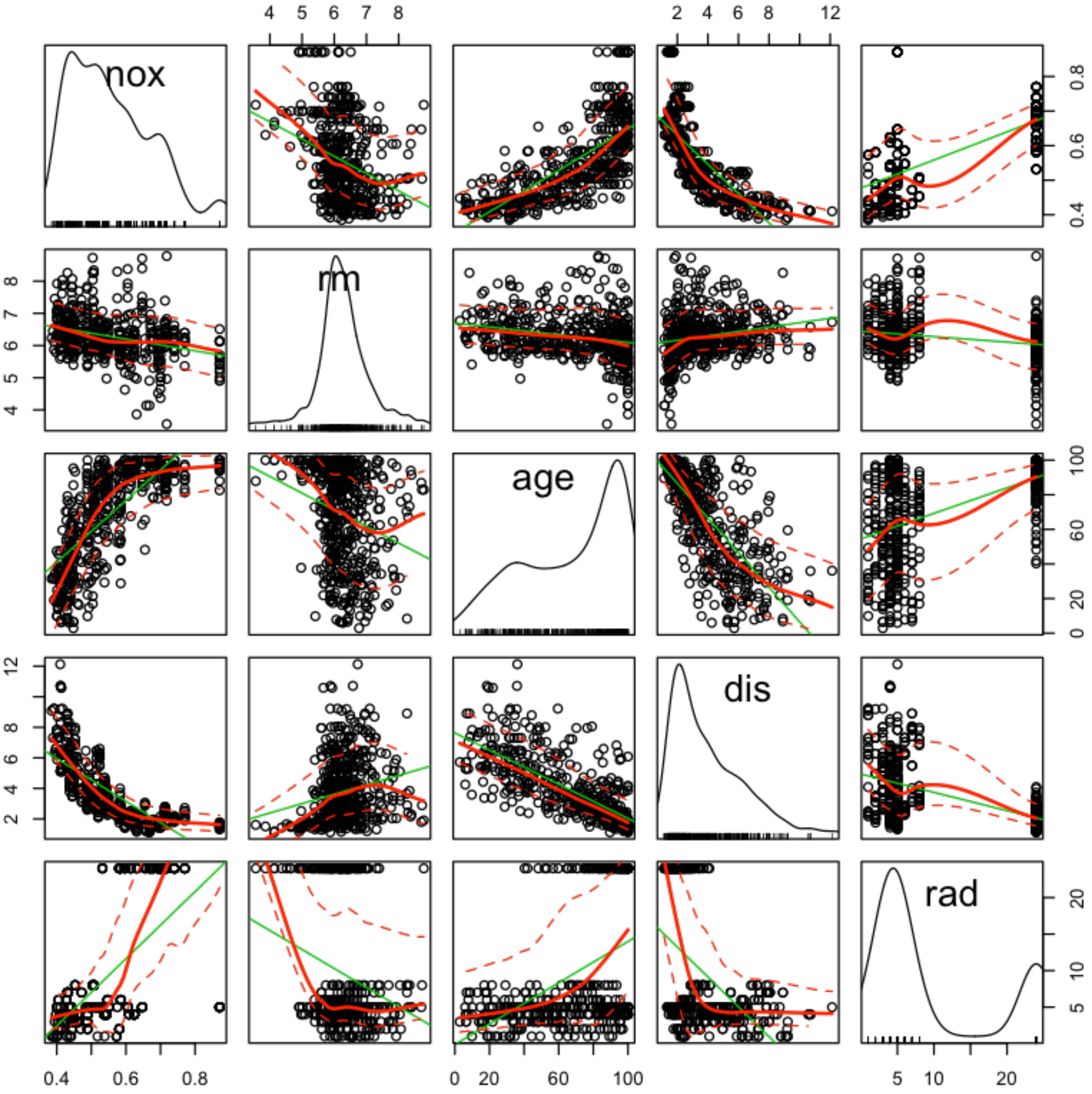
3.1 Generate a scatterplot matrix for all metric variables. Take a few minutes to draw as many insights as you can about the relationships in the data.

In [2]:

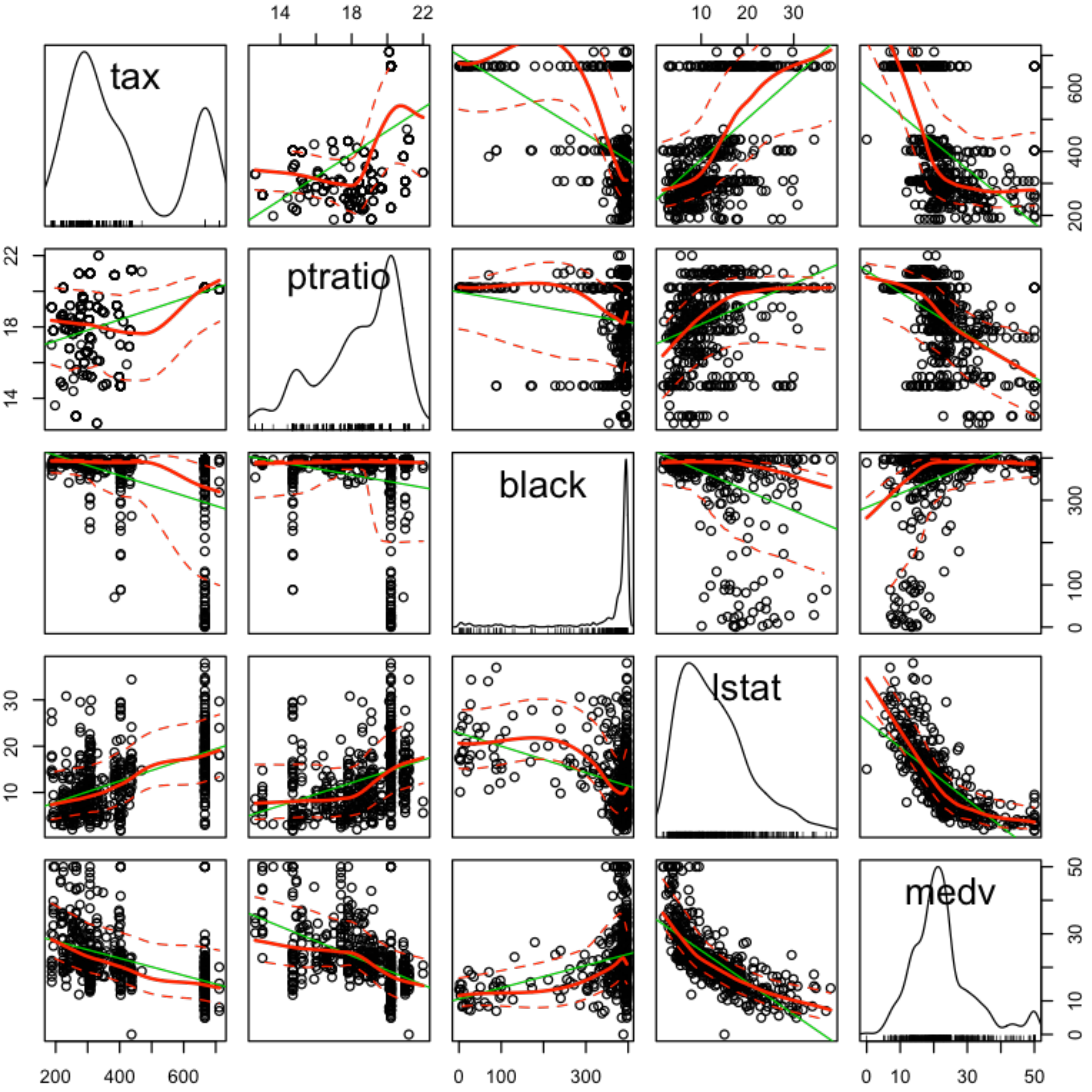
```
Warning message in smoother(x, y, col = col[2], log.x = FALSE, log.y
= FALSE, spread = spread, :
“could not fit smooth”Warning message in smoother(x, y, col = col[2],
log.x = FALSE, log.y = FALSE, spread = spread, :
“could not fit smooth”Warning message in smoother(x, y, col = col[2],
log.x = FALSE, log.y = FALSE, spread = spread, :
“could not fit smooth”Warning message in smoother(x, y, col = col[2],
log.x = FALSE, log.y = FALSE, spread = spread, :
“could not fit smooth”
```



In [3]:


nox
rm
age
dis
rad
4
5
6
7
8
2
4
6
8
12
0.8
0.6
0.4
8
7
6
5
4
100
60
20
0
12
8
6
4
2
20
10
5
0.4
0.6
0.8
0
20
60
100
5
10
20

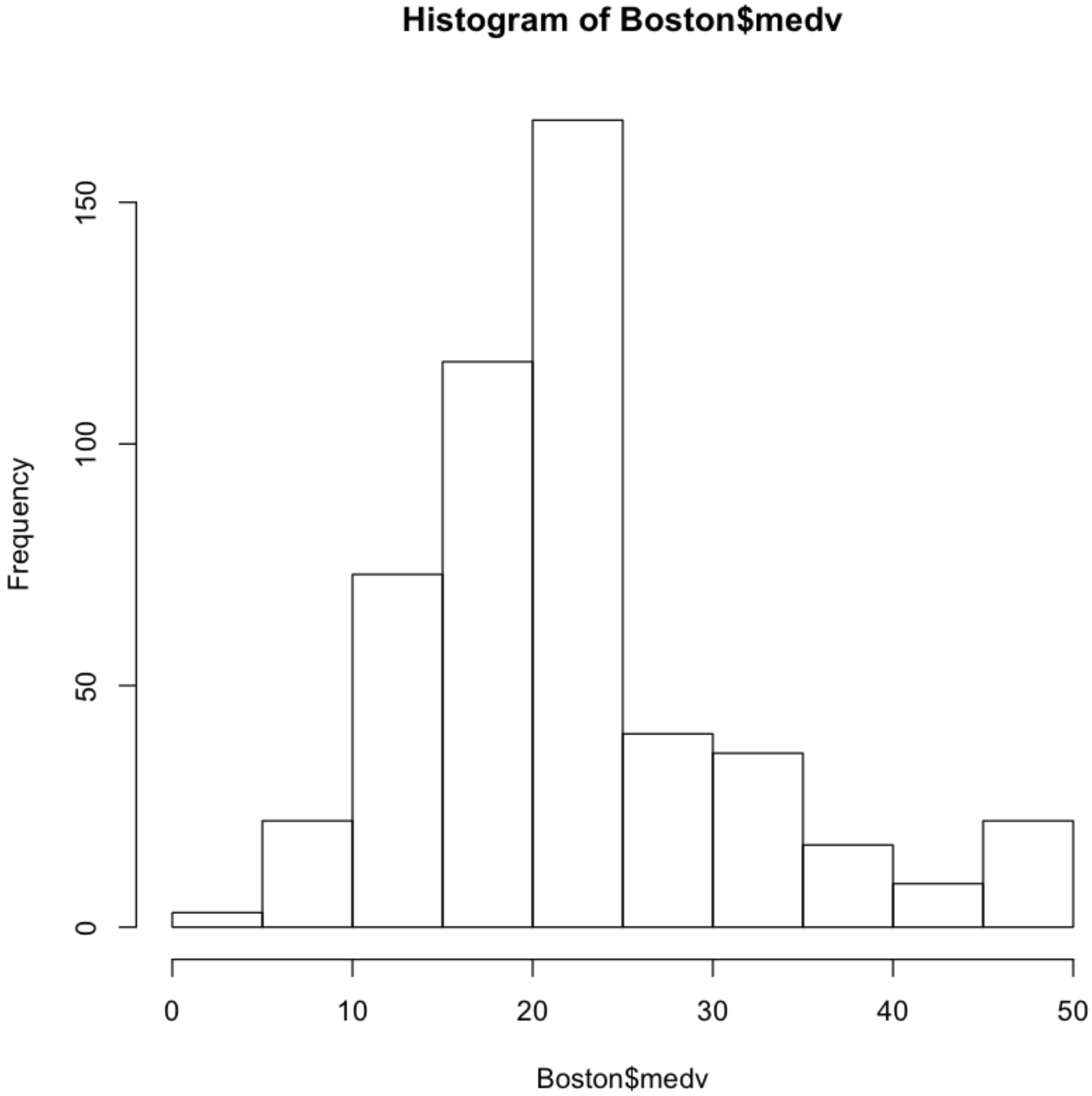
In [4]:



** 3.2 ** Examine the main output variable, medv. Comment on any unusual values you find, and any features that might be important for statistical modeling.

In [5]:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00   16.85   21.20   22.50   25.00   50.00 
```



Having a minimum value median home price of zero is nonsensical, lets take a look at the data points there are with this value

In [6]:

	X	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat
134	134	0.32982	0	21.89	0	0.624	5.822	95.4	2.4699	4	437	21.2	388.69	15.03

Looks like this is actually a missing value since, if we have a look at the summary of the data set the other variable values are reasonable.

In [7]:

```
       X               crim                zn             indus      
 Min.   :  1.0   Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46  
 1st Qu.:127.2   1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19  
 Median :253.5   Median : 0.25651   Median :  0.00   Median : 9.69  
 Mean   :253.5   Mean   : 3.61352   Mean   : 11.36   Mean   :11.14  
 3rd Qu.:379.8   3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10  
 Max.   :506.0   Max.   :88.97620   Max.   :100.00   Max.   :27.74  
      chas              nox               rm             age        
 Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90  
 1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02  
 Median :0.00000   Median :0.5380   Median :6.208   Median : 77.50  
 Mean   :0.06917   Mean   :0.5547   Mean   :6.285   Mean   : 68.57  
 3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08  
 Max.   :1.00000   Max.   :0.8710   Max.   :8.780   Max.   :100.00  
      dis              rad              tax           ptratio     
 Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60  
 1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40  
 Median : 3.207   Median : 5.000   Median :330.0   Median :19.05  
 Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46  
```

As a result we will recode this value as NA

In [8]:

	X	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat
134	134	0.32982	0	21.89	0	0.624	5.822	95.4	2.4699	4	437	21.2	388.69	15.03

Next, Having a median value of exactly 50 is weird, we may be concerned that this is a top code lets have a look at how many data points have them

```
In [9]:
```

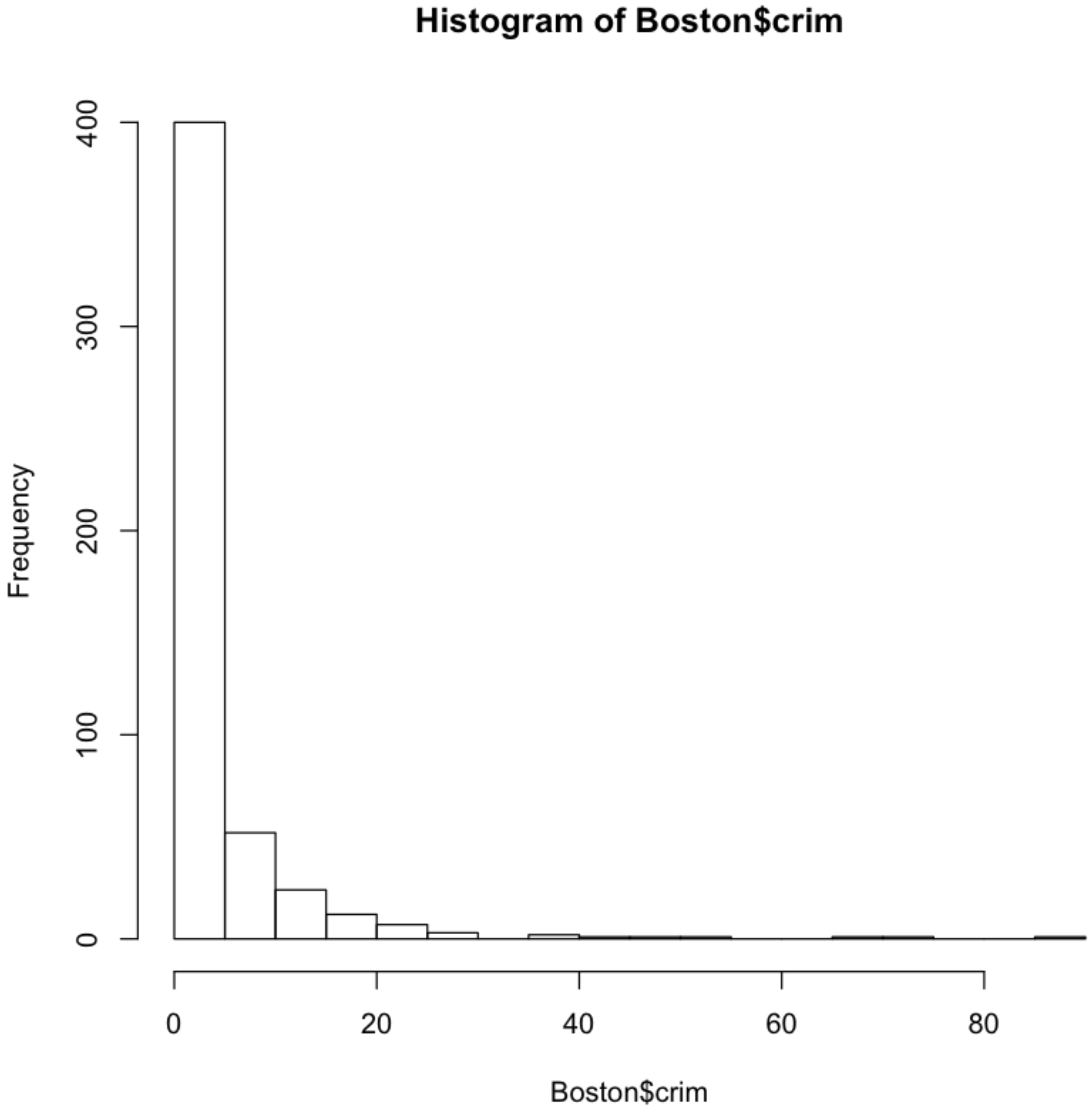
	X	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lsta
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
162	162	1.46336	0	19.58	0	0.6050	7.489	90.8	1.9709	5	403	14.7	374.43	1.73
163	163	1.83377	0	19.58	1	0.6050	7.802	98.2	2.0407	5	403	14.7	389.61	1.92
164	164	1.51902	0	19.58	1	0.6050	8.375	93.9	2.1620	5	403	14.7	388.45	3.32
167	167	2.01019	0	19.58	0	0.6050	7.929	96.2	2.0459	5	403	14.7	369.30	3.70
187	187	0.05602	0	2.46	0	0.4880	7.831	53.6	3.1992	3	193	17.8	392.63	4.45
196	196	0.01381	80	0.46	0	0.4220	7.875	32.0	5.6484	4	255	14.4	394.23	2.97
205	205	0.02009	95	2.68	0	0.4161	8.034	31.9	5.1180	4	224	14.7	390.55	2.88
226	226	0.52693	0	6.20	0	0.5040	8.725	83.0	2.8944	8	307	17.4	382.00	4.63
258	258	0.61154	20	3.97	0	0.6470	8.704	86.9	1.8010	5	264	13.0	389.70	5.12
268	268	0.57834	20	3.97	0	0.5750	8.297	67.0	2.4216	5	264	13.0	384.54	7.44
284	284	0.01501	90	1.21	1	0.4010	7.923	24.8	5.8850	1	198	13.6	395.52	3.16
369	369	4.89822	0	18.10	0	0.6310	4.970	100.0	1.3325	24	666	20.2	375.52	3.26
370	370	5.66998	0	18.10	1	0.6310	6.683	96.8	1.3567	24	666	20.2	375.33	3.73
371	371	6.53876	0	18.10	1	0.6310	7.016	97.5	1.2024	24	666	20.2	392.05	2.96
372	372	9.23230	0	18.10	0	0.6310	6.216	100.0	1.1691	24	666	20.2	366.15	9.53
373	373	8.26725	0	18.10	1	0.6680	5.875	89.6	1.1296	24	666	20.2	347.88	8.88

Looks like the value of 50 is a top code meaning that median home value in these towns is greater than or equal to 50

There is nothing more to do here other than just keep this fact in mind.

3.3 Examine the main independent variable of interest, crim. What transformation could you apply to this variable to aid in visualizing it? Comment on any unusual features you find.

In [10]:

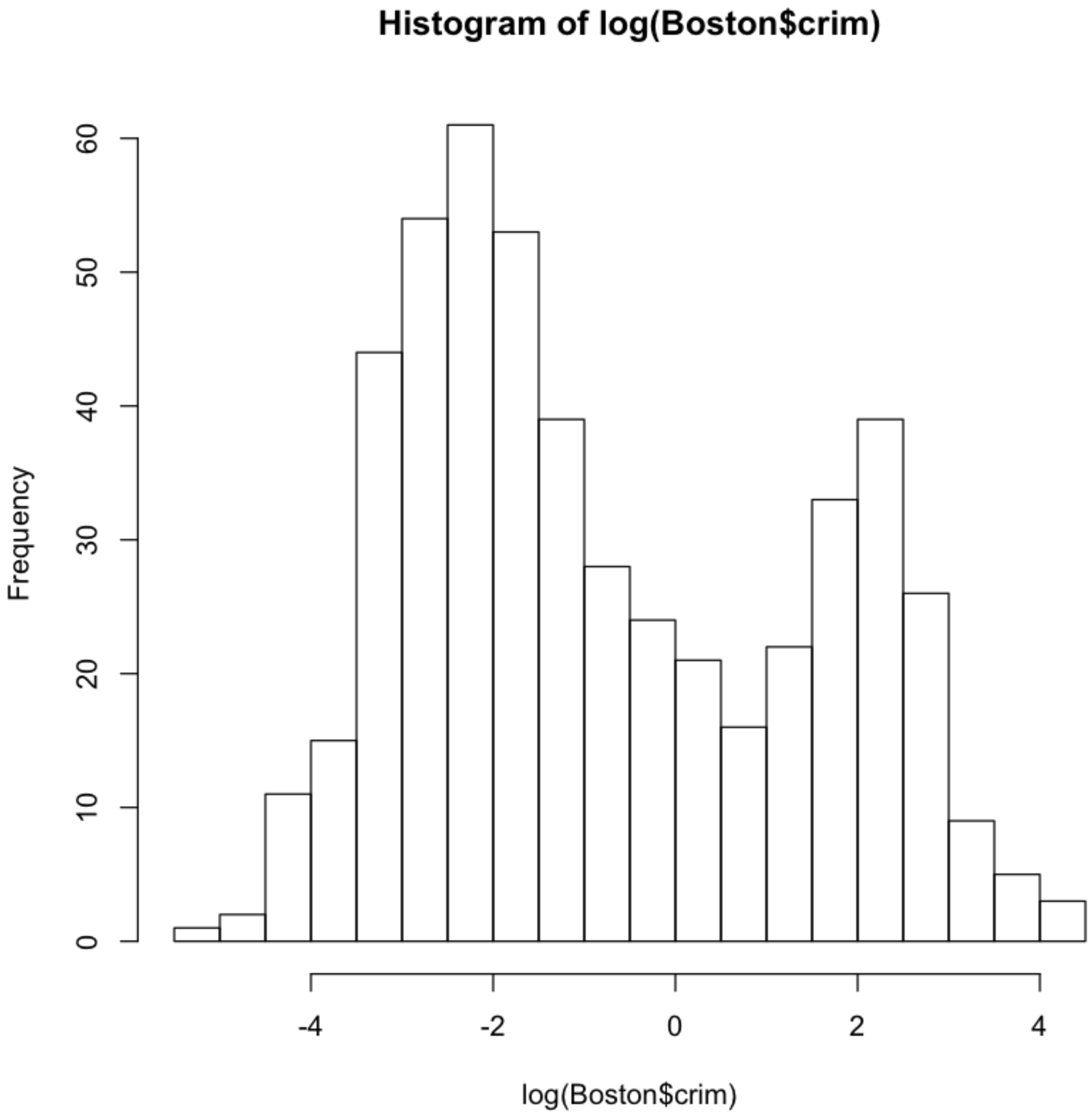

Histogram of Boston$crim
Frequency
0
100
200
300
400
0
20
40
60
80
Boston$crim

```
In [11]:
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
 0.00632  0.08204  0.25651  3.61352  3.67708 88.97620 
```

We can see that the crim variable has is strictly positive with a heavy right skew, these are the kind of variables which are strong candidates for a log transformation

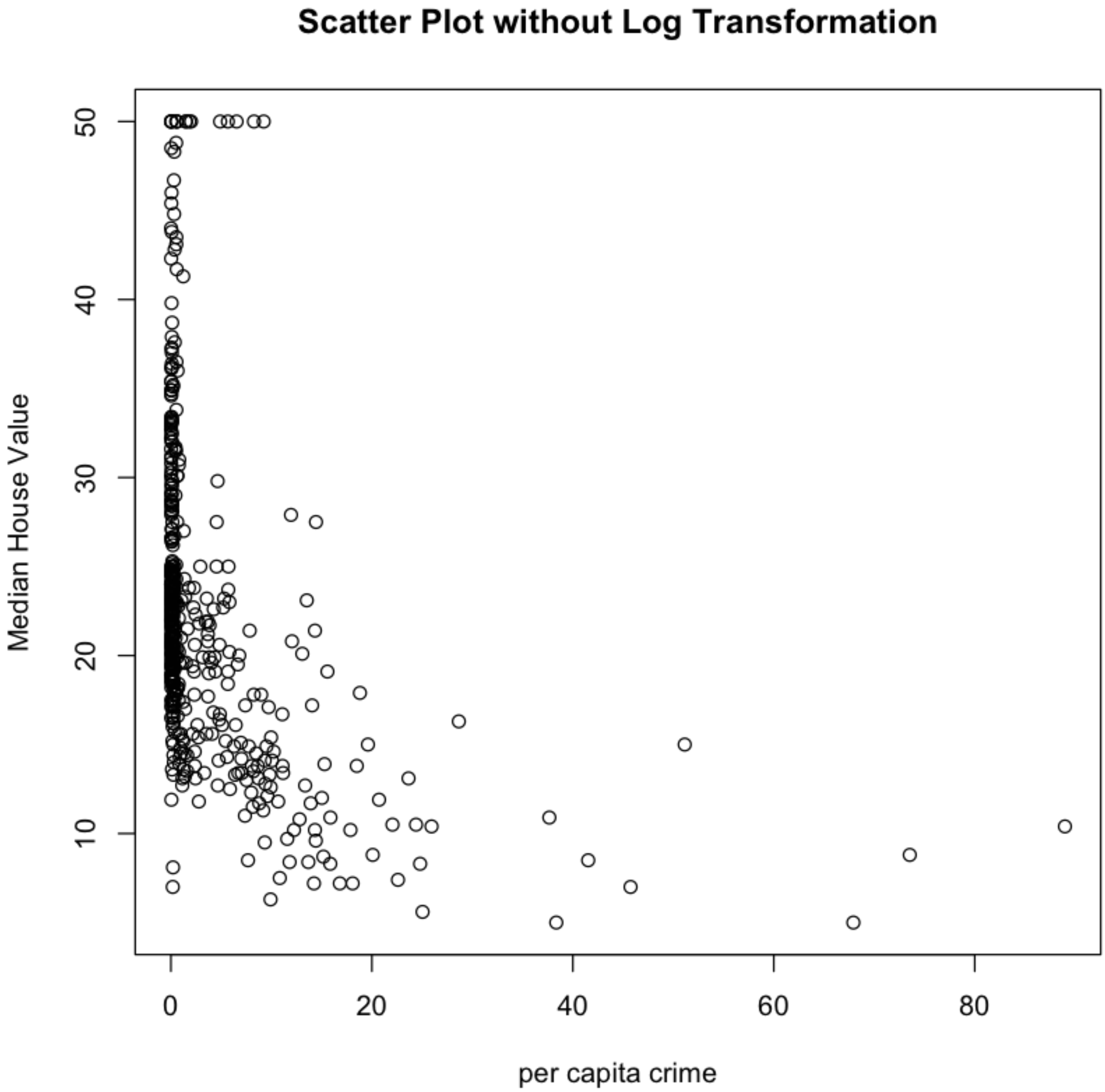
```
In [12]:
```

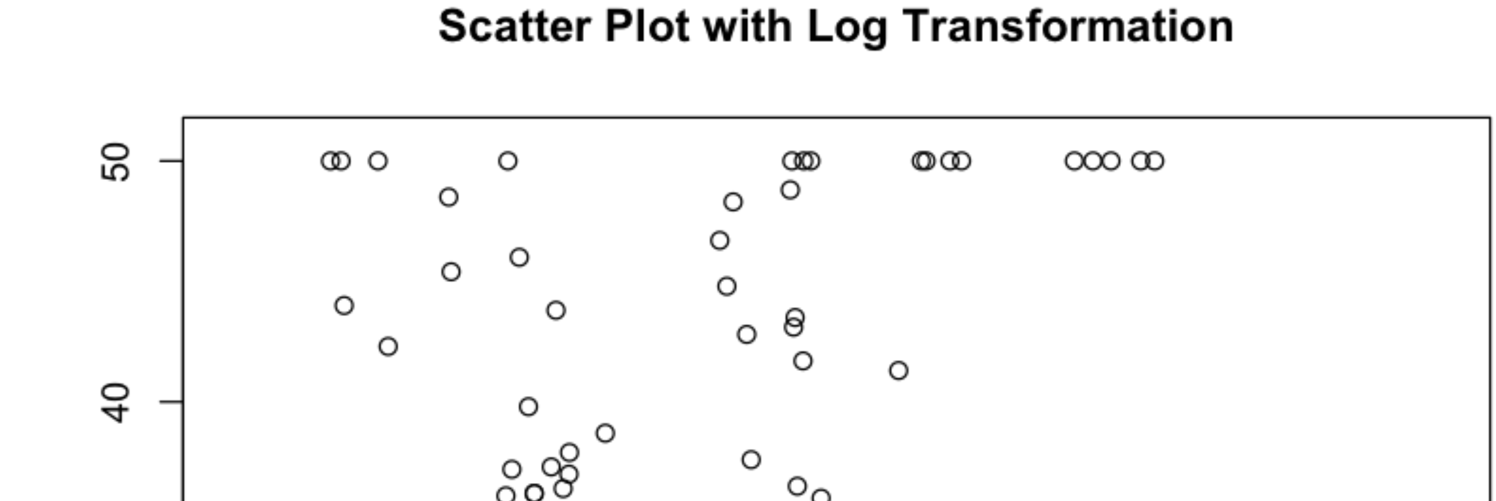


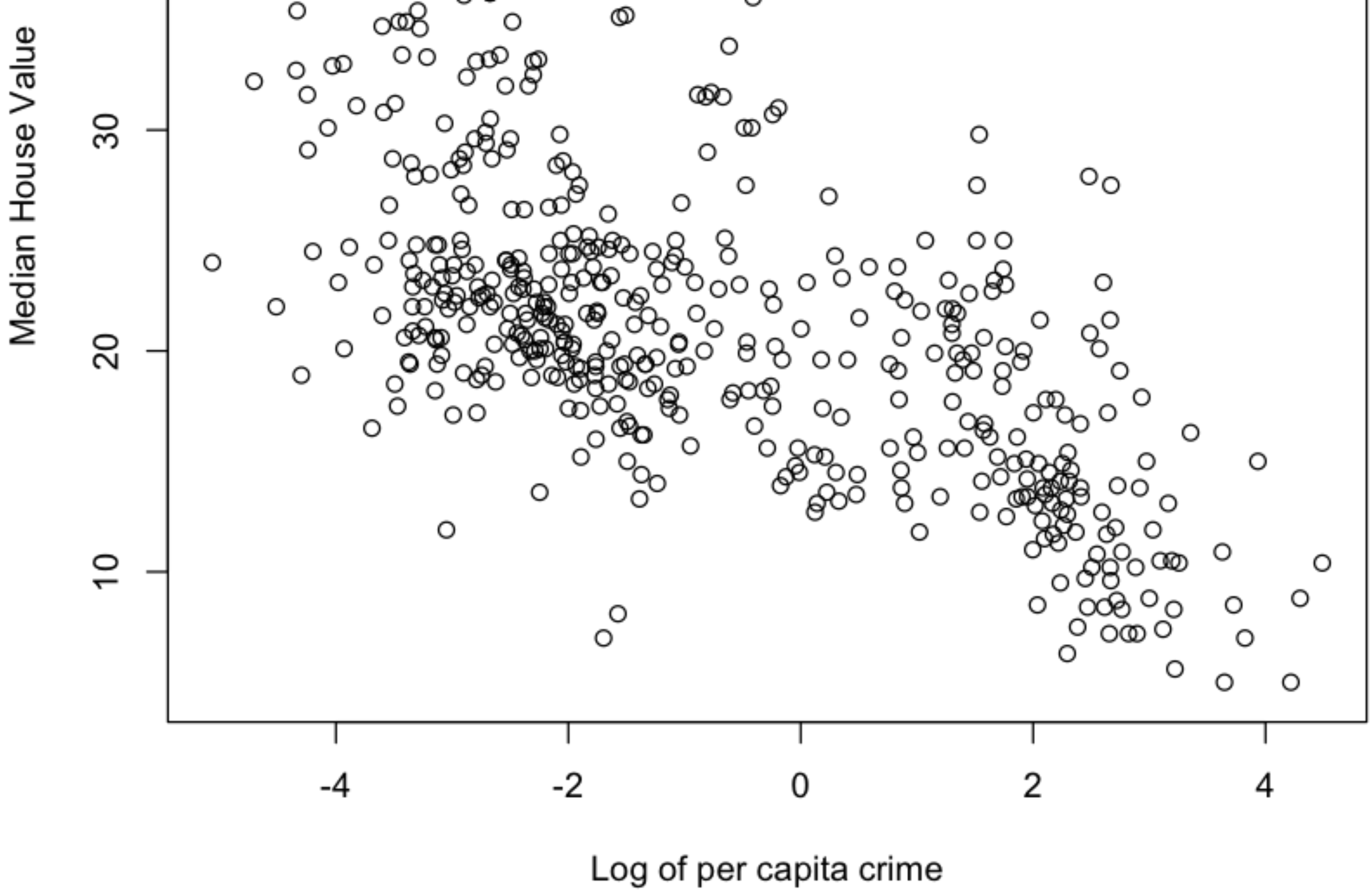
> We can see that the log transformation gives us a nice compact bi-modal distribution which as we will see much later in the class will help up identify linear relationships between variables.

** 3.4 ** Examine the bivariate relationship between medv and crime. What type of relationship do these variables have?

In [13]:


Scatter Plot without Log Transformation
Median House Value
50
40
30
20
10
0
20
40
60
80
per capita crime


Scatter Plot with Log Transformation
50
40



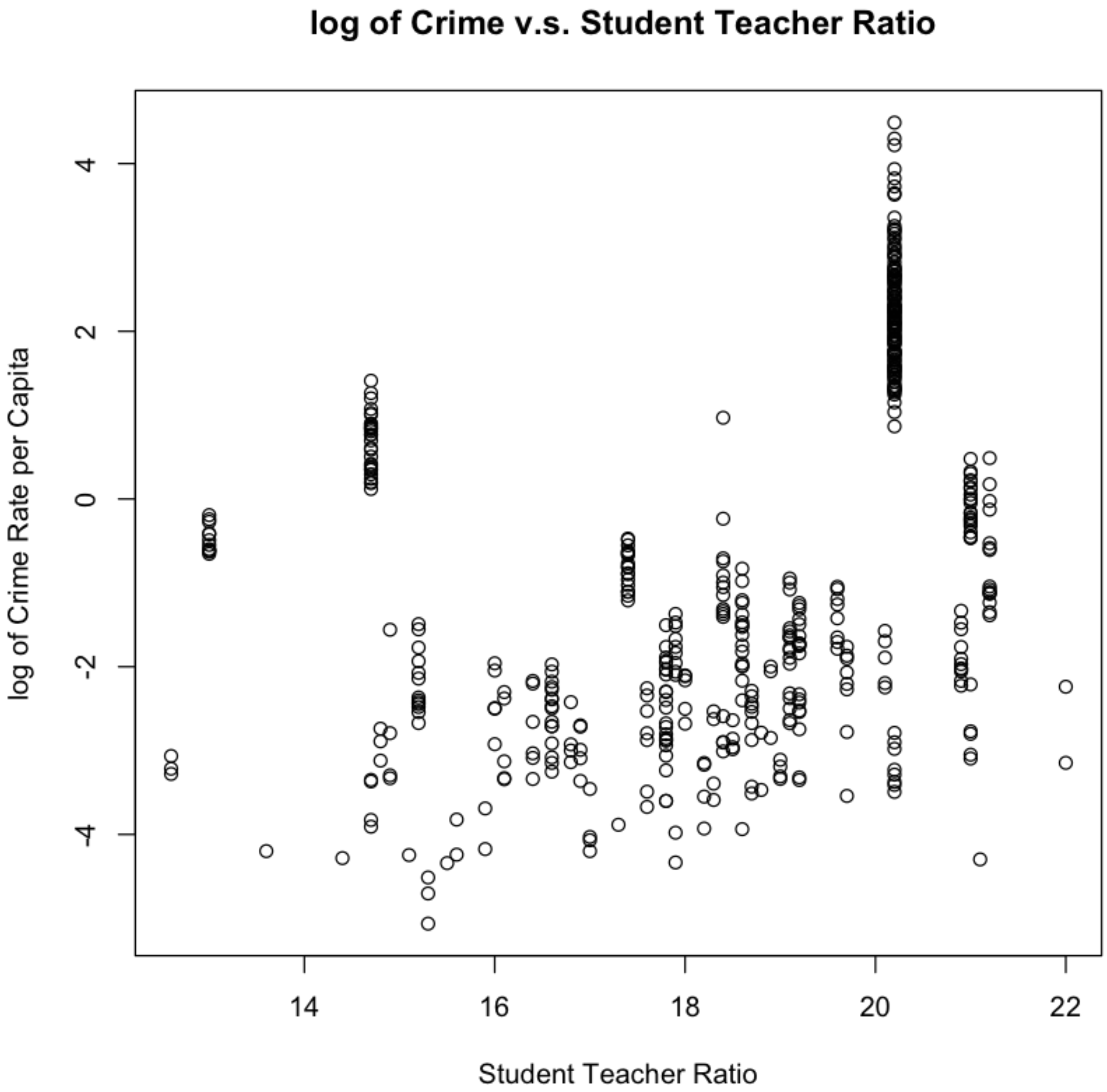
We can see that there is a generally negative relationship between crime and median home value

We might be a little critical of this because of a couple values with medv = 50 and crim value of between -1 and 3 which would throw of any linear fit we might want to do.

But we have to remember that these are top codes which might lead us to exclude them from a linear fitting procedure.

3.5 (As time permits) Continue your exploratory data analysis. Be prepared to share interesting findings with the class.

In [14]:



The above scatter plot makes us wonder why almost all the towns which have a log crim rate above 1 have a student teacher ratio of either 14.7 or 20.2

```
In [15]:
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  14.70   20.20   20.20   19.96   20.20   20.20 
```

We know that crime and income are negatively correllated meaning that crime rates in lower income areas tend to be higher for various reasons

We also know that many places implement caps on student teacher ratios and areas which are low income are more likely to run up against these caps due to tight fiscal situations, thus this might be the story which explains this feature of the data

This particular story may not be true but it is important that when you find strange features in your data set to try to use out of sample (domain) information to understand/explain it.

You must be a story teller in data science.