Chapter 1

Introduction to data

Boston housing data

```
library(mlbench)
data(BostonHousing2)
```

Consider the following dataset:

The Boston house-price data of Harrison, D. and Rubinfeld, D.L

	crim	chas	...	age	town	rad	medv
1	0.00632	0	...	65.2	Nahant	1	24.0
2	0.02731	0	...	78.9	Swampscott	2	21.6
3	0.02729	1	...	61.1	Swampscott	2	34.7
4	0.03237	0	...	45.8	Marblehead	6	33.4
...	...	...	...	...	...	...	...
505	0.10959	1	...	89.3	Winthrop	1	22.0
506	0.04741	0	...	80.8	Winthrop	1	11.9

Boston housing data

1
2
3
4
...
505
506

506
observations

506 census tracts

Boston housing data

	crim	chas	...	age	town	rad	medv
1	0.00632	0	...	65.2	Nahant	1	24.0
2	0.02731	0	...	78.9	Swampscott	2	21.6
3	0.02729	1	...	61.1	Swampscott	2	34.7
4	0.03237	0	...	45.8	Marblehead	6	33.4
...	...	...	...	...	...	...	...
505	0.10959	1	...	89.3	Winthrop	1	22.0
506	0.04741	0	...	80.8	Winthrop	1	11.9

Boston housing data

	crim	chas	...	age	town	rad	medv

We see 6 **variables**

Types of variables

	crim	chas	...	age	town	rad	medv
1	0.00632	0	...	65.2	Nahant	1	24.0
2	0.02731	0	...	78.9	Swampscott	2	21.6
3	0.02729	1	...	61.1	Swampscott	2	34.7
4	0.03237	0	...	45.8	Marblehead	6	33.4
...	...	...	...	...	...	...	...
505	0.10959	1	...	89.3	Winthrop	1	22.0
506	0.04741	0	...	80.8	Winthrop	1	11.9

Types of variables

crim
0.00632
0.02731
0.02729
0.03237
...
0.10959
0.04741

variable crim:

per capita crime rate

Types of variables

crim
0.00632
0.02731
0.02729
0.03237
...
0.10959
0.04741

variable crim:

per capita crime rate

`?BostonHousing2`

Types of variables

crim
0.00632
0.02731
0.02729
0.03237
...
0.10959
0.04741

variable crim:

per capita crime rate

`?BostonHousing2`

- Numerical (it's a number)
- Continuous (can take potentially a continuum of values)

Types of variables

	crim	chas	...	age	town	rad	medv
1	0.00632	0	...	65.2	Nahant	1	24.0
2	0.02731	0	...	78.9	Swampscott	2	21.6
3	0.02729	1	...	61.1	Swampscott	2	34.7
4	0.03237	0	...	45.8	Marblehead	6	33.4
...	...	...	...	...	...	...	...
505	0.10959	1	...	89.3	Winthrop	1	22.0
506	0.04741	0	...	80.8	Winthrop	1	11.9

Types of variables

rad
1
2
2
6
...
1
1

variable rad

index of accessibility to radial highways

Types of variables

rad
1
2
2
6
...
1
1

variable rad

index of accessibility to radial highways

- Numerical (it's a number)
- Discrete (can take only a discrete set of values. Here: 6 or more)

Types of variables

	crim	chas	...	age	town	rad	medv
1	0.00632	0	...	65.2	Nahant	1	24.0
2	0.02731	0	...	78.9	Swampscott	2	21.6
3	0.02729	1	...	61.1	Swampscott	2	34.7
4	0.03237	0	...	45.8	Marblehead	6	33.4
...	...	...	...	...	...	...	...
505	0.10959	1	...	89.3	Winthrop	1	22.0
506	0.04741	0	...	80.8	Winthrop	1	11.9

Types of variables

town
Nahant
Swampscott
Swampscott
Marblehead
...
Winthrop
Winthrop

variable town

location of house

Types of variables

town
Nahant
Swampscott
Swampscott
Marblehead
...
Winthrop
Winthrop

variable town

location of house

- Categorical (it’s a label)

Types of variables

town
Nahant
Swampscott
Swampscott
Marblehead
...
Winthrop
Winthrop

variable town

location of house

- Categorical (it's a label)

Categorical variables are ALWAYS discrete.
The discrete/continuous distinction is made
only for numerical variables

Types of variables

	crim	chas	...	age	town	rad	medv
1	0.00632	0	...	65.2	Nahant	1	24.0
2	0.02731	0	...	78.9	Swampscott	2	21.6
3	0.02729	1	...	61.1	Swampscott	2	34.7
4	0.03237	0	...	45.8	Marblehead	6	33.4
...	...	...	...	...	...	...	...
505	0.10959	1	...	89.3	Winthrop	1	22.0
506	0.04741	0	...	80.8	Winthrop	1	11.9

Types of variables

chas
0
0
1
0
...
1
0

variable chas:

Charles River (= 1 if tract bounds river; 0 otherwise)

Types of variables

chas
0
0
1
0
...
1
0

variable chas:

Charles River (= 1 if tract bounds river; 0 otherwise)

- Numerical (it's a number)
- Discrete (can take only a discrete set of values. Here: 2)

Types of variables

chas
0
0
1
0
...
1
0

variable chas:

Charles River (= 1 if tract bounds river; 0 otherwise)

- Numerical (it's a number)
- Discrete (can take only a discrete set of values. Here: 2)
- Could also be considered categorical
- Categories: YES or NO

Types of variables

	crim	chas	...	age	town	rad	medv
1	0.00632	0	...	65.2	Nahant	1	24.0
2	0.02731	0	...	78.9	Swampscott	2	21.6
3	0.02729	1	...	61.1	Swampscott	2	34.7
4	0.03237	0	...	45.8	Marblehead	6	33.4
...	...	...	...	...	...	...	...
505	0.10959	1	...	89.3	Winthrop	1	22.0
506	0.04741	0	...	80.8	Winthrop	1	11.9

Types of variables

medv
24.0
21.6
34.7
33.4
...
22.0
11.9

variable medv:

median value of owner-occupied homes in USD 1000's

What type of variable is it?

Variables relationships

What is the variable of interest here?

Questions we may ask:

- Are houses on the river more expensive?
- Does the crime rate impact the value?
- Is the crime rate affected by the index rad?
- Is the town of winthrop more expensive than nahant?

```
attach(BostonHousing2)
mean(medv[town=="Winthrop"])
mean(medv[town=="Nahant"])
```

Variables relationships

What is the variable of interest here?

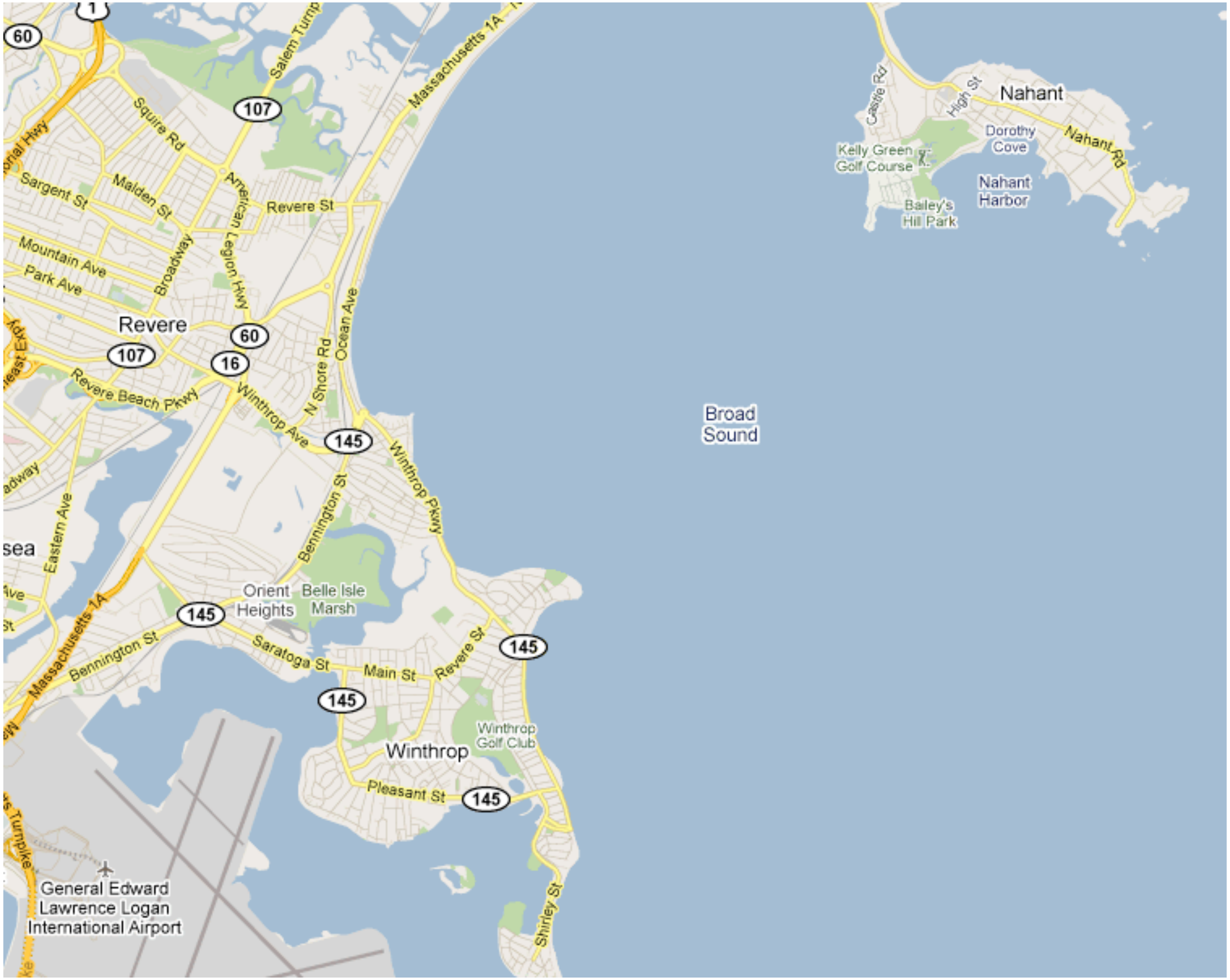
medv

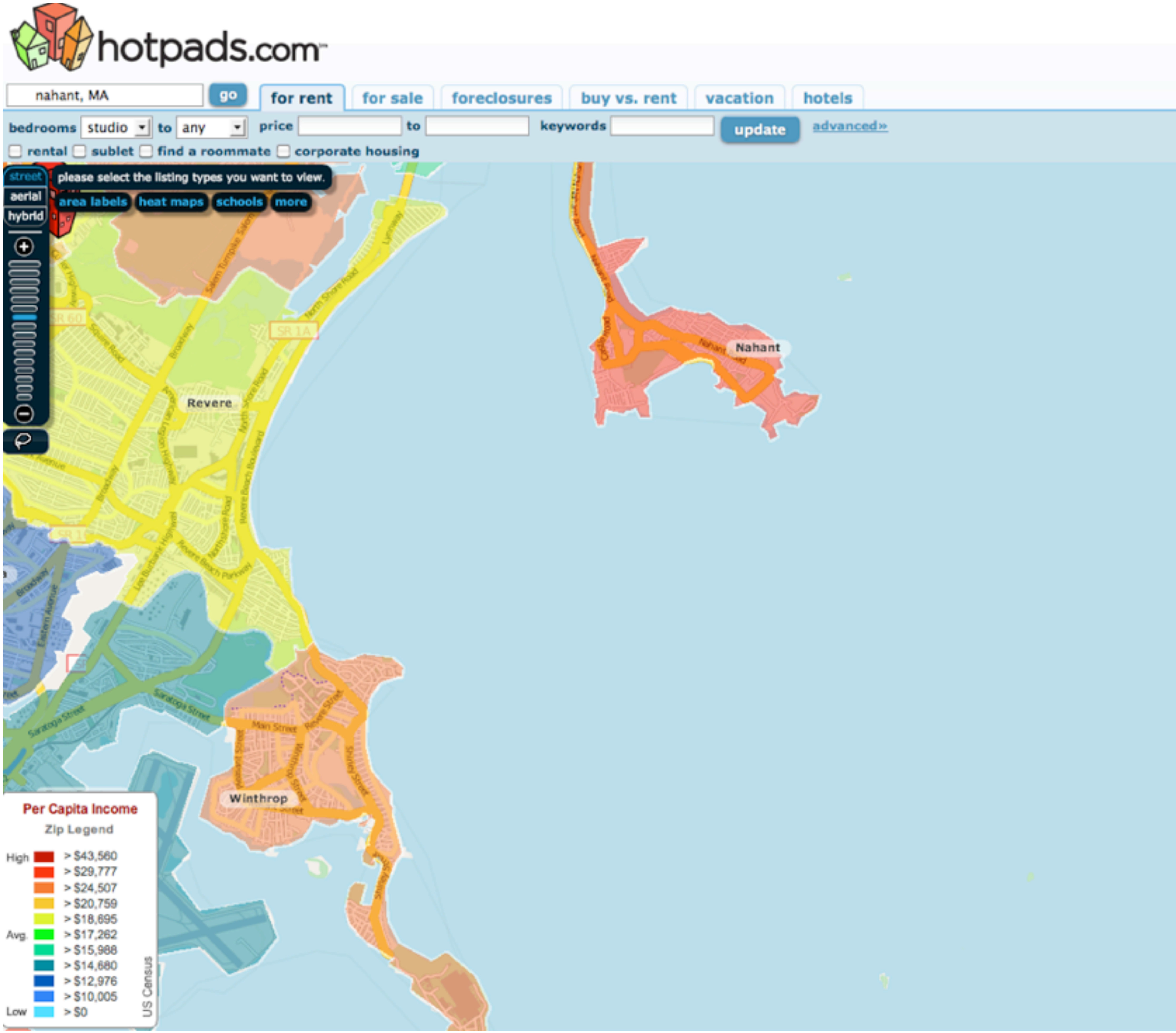
Questions we may ask:

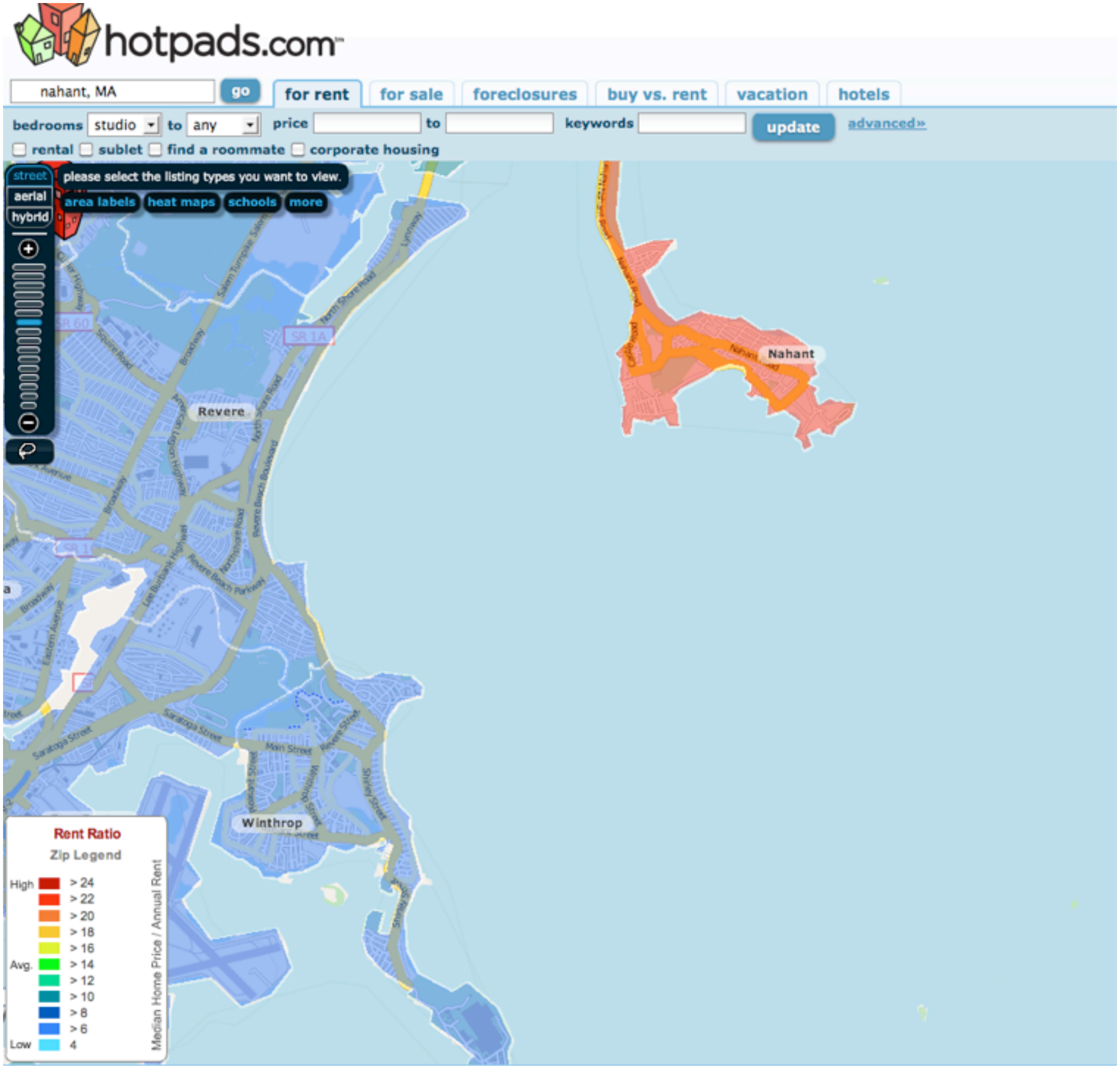
- Are houses on the river more expensive?
- Does the crime rate impact the value?
- Is the crime rate affected by the index rad?
- Is the town of winthrop more expensive than nahant?

```
attach(BostonHousing2)
mean(medv[town=="Winthrop"])
mean(medv[town=="Nahant"])
```

Variables relationships

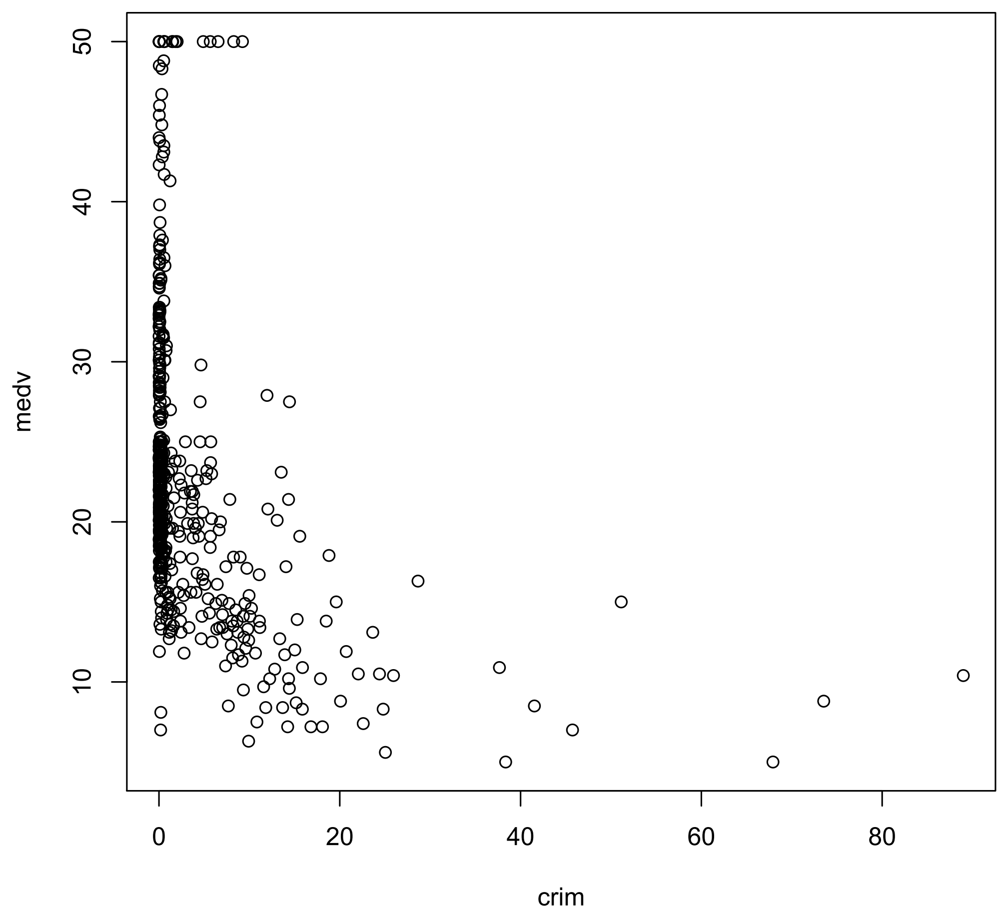



hotpads.com
nahant, MA
go
for rent
for sale
foreclosures
buy vs. rent
vacation
hotels
bedrooms studio to any price to keywords update advanced»
rental sublet find a roommate corporate housing
street
aerial
hybrid
please select the listing types you want to view.
area labels
heat maps
schools
more
Nahant
Revere
Winthrop
Per Capita Income
Zip Legend
High > $43,560
> $29,777
> $24,507
> $20,759
> $18,695
Avg. > $17,262
> $15,988
> $14,680
> $12,976
> $10,005
Low > $0
US Census


hotpads.com
nahant, MA
go
for rent
for sale
foreclosures
buy vs. rent
vacation
hotels
bedrooms studio to any
price
to
keywords
update
advanced»
rental
sublet
find a roommate
corporate housing
street
aerial
hybrid
please select the listing types you want to view.
area labels
heat maps
schools
more
Revere
Nahant
Winthrop
Rent Ratio
Zip Legend
High
Avg.
Low
> 24
> 22
> 20
> 18
> 16
> 14
> 12
> 10
> 8
> 6
4
Median Home Price / Annual Rent

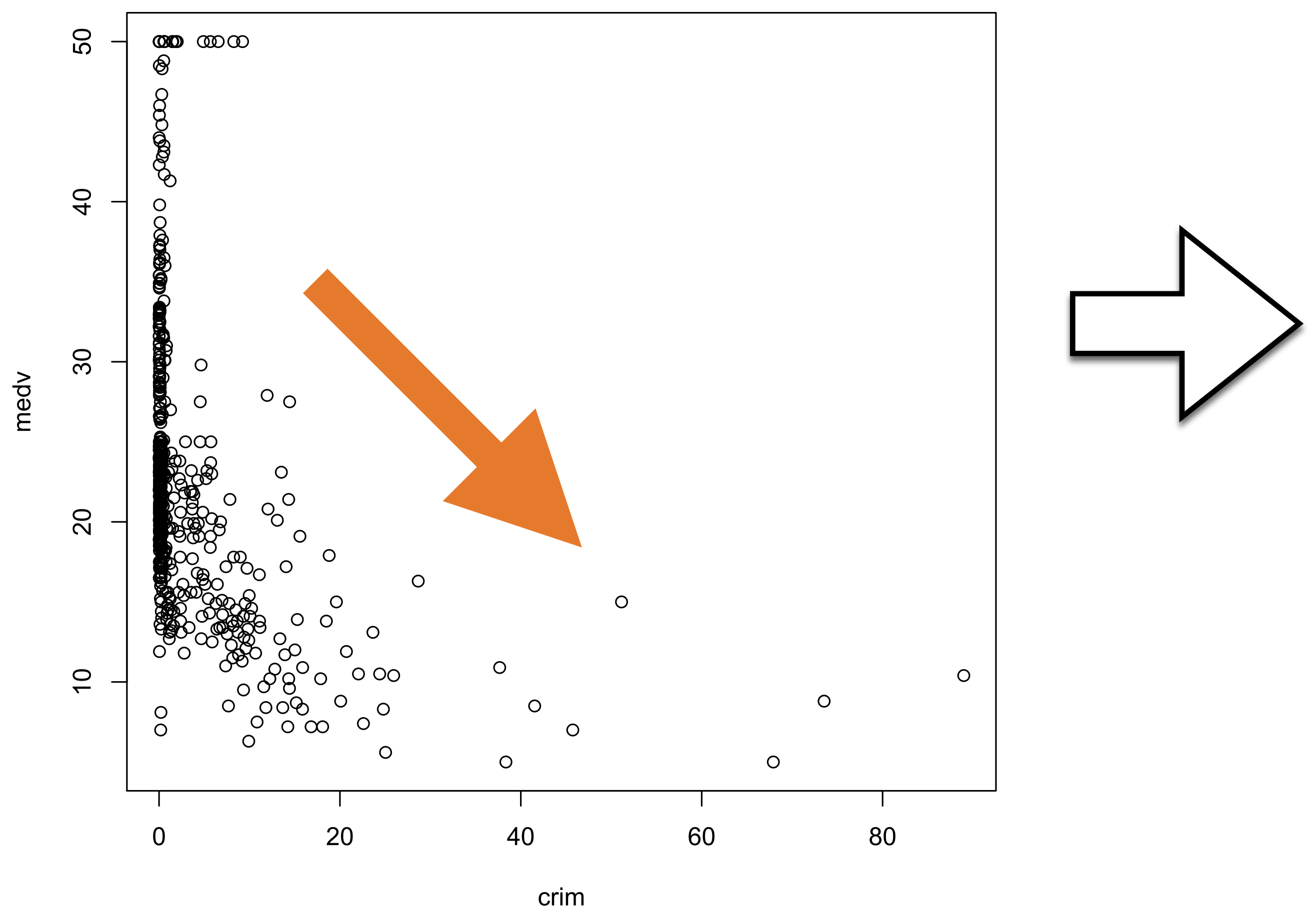
Associated and independent variables

Two variables are associated if one affects the other (which affects which does not matter)



Associated and independent variables

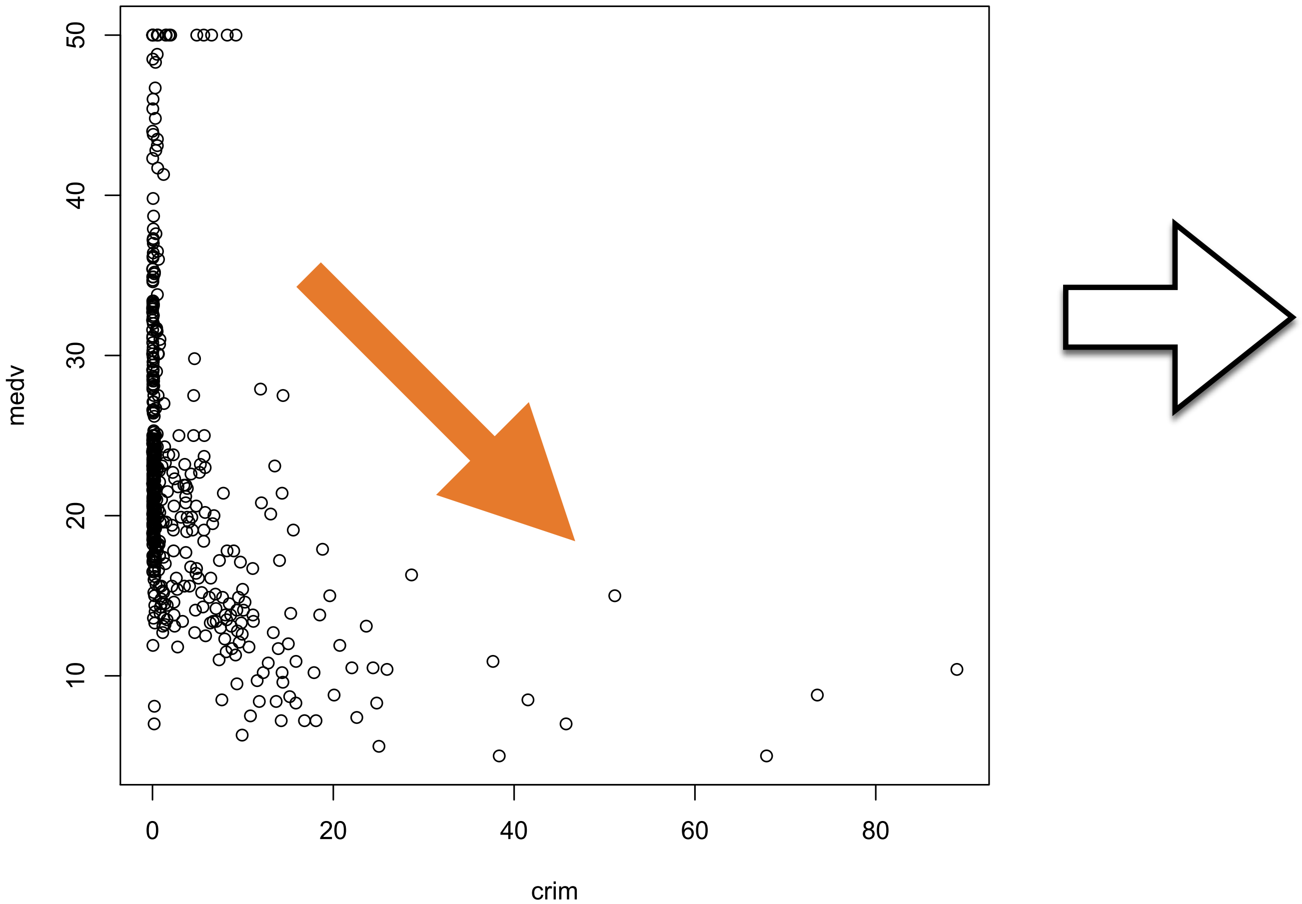
Two variables are associated if one affects the other (which affects which does not matter)



Negative association

Associated and independent variables

Two variables are associated if one affects the other (which affects which does not matter)

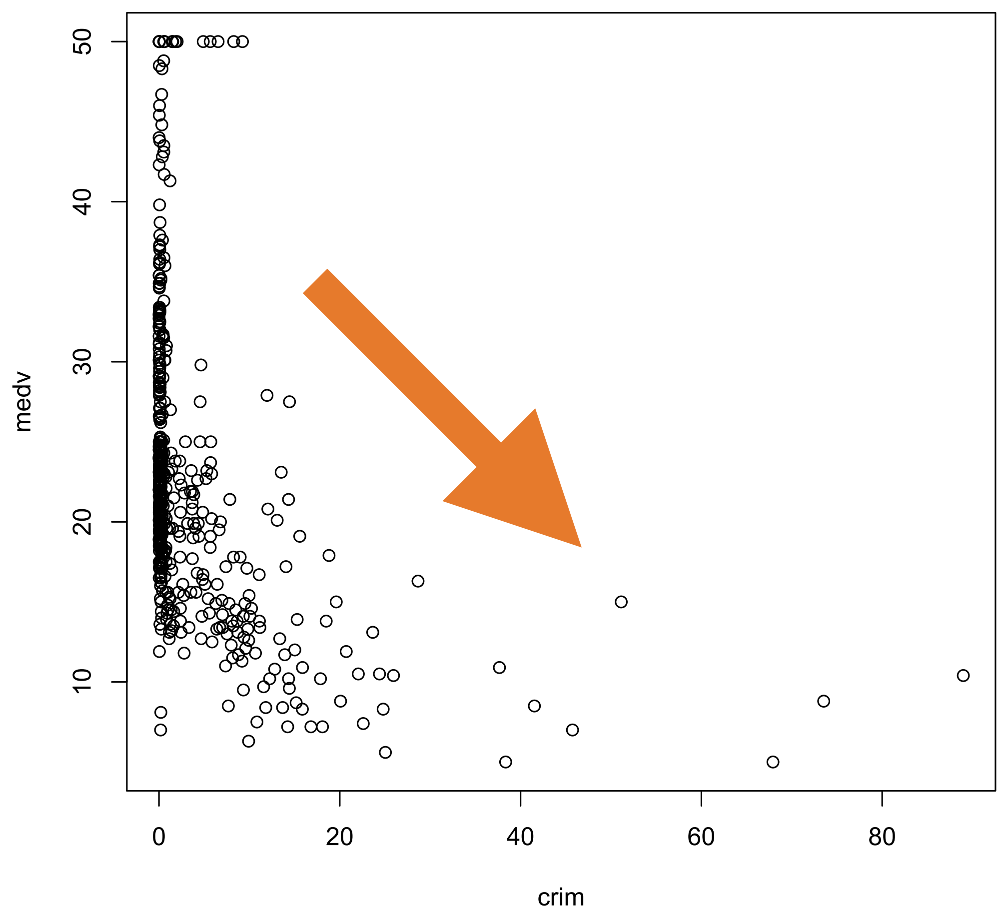


Negative association

If not associated: independent

Associated and independent variables

Two variables are associated if one affects the other (which affects which does not matter)



Negative association

This is a scatterplot

`plot(crim, medv)`

If not associated: independent

Summarizing numerical data

The nice things about numerical data:

1. We can add (average) them
2. There is a natural order

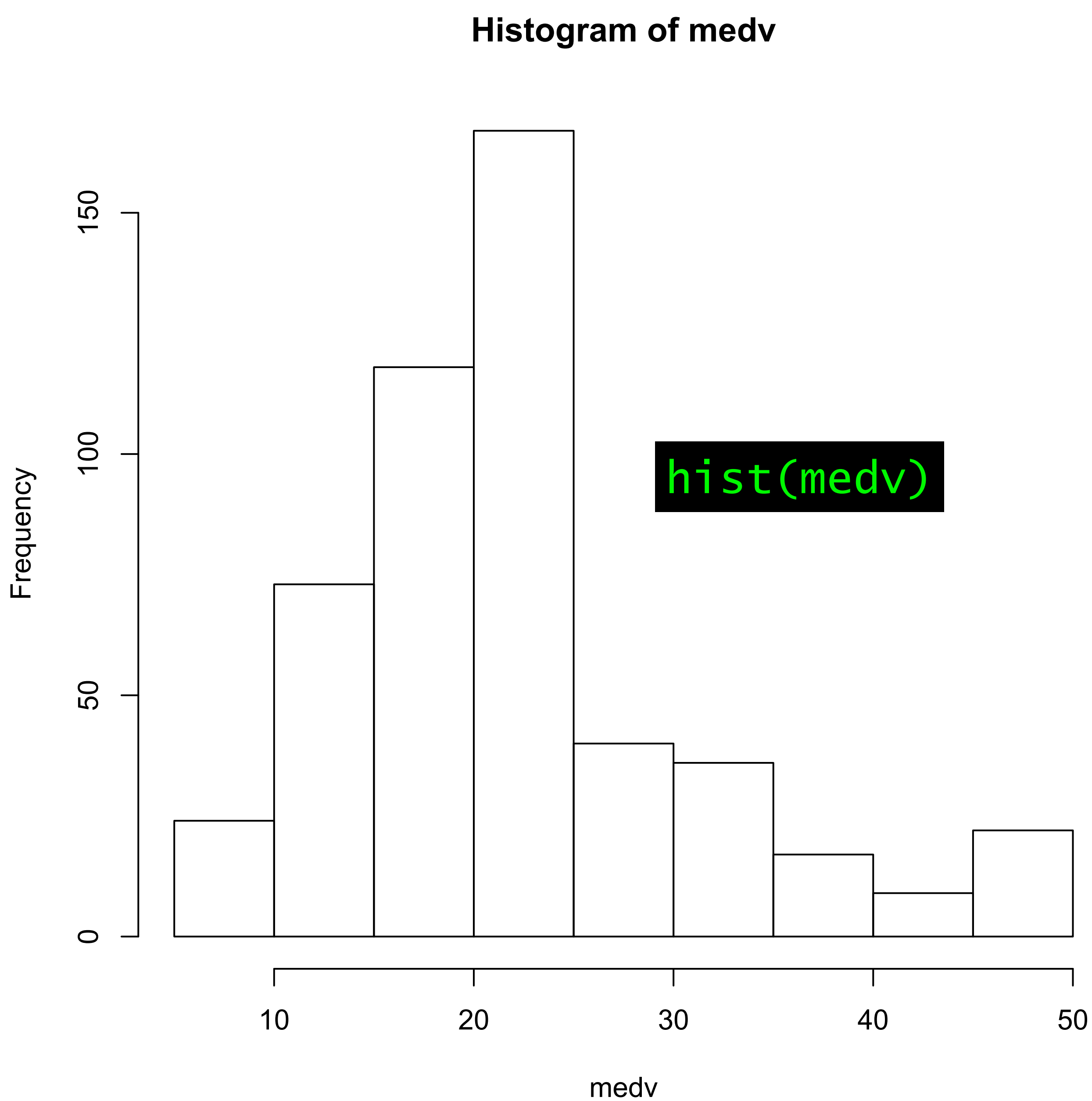
We will consider two types of summaries

- Numerical: mean, median, standard deviation, quantiles
- Graphical: scatter plots, boxplots.

They give the "big picture" about the dataset.
Hereafter, we work on the variable medv

Histograms

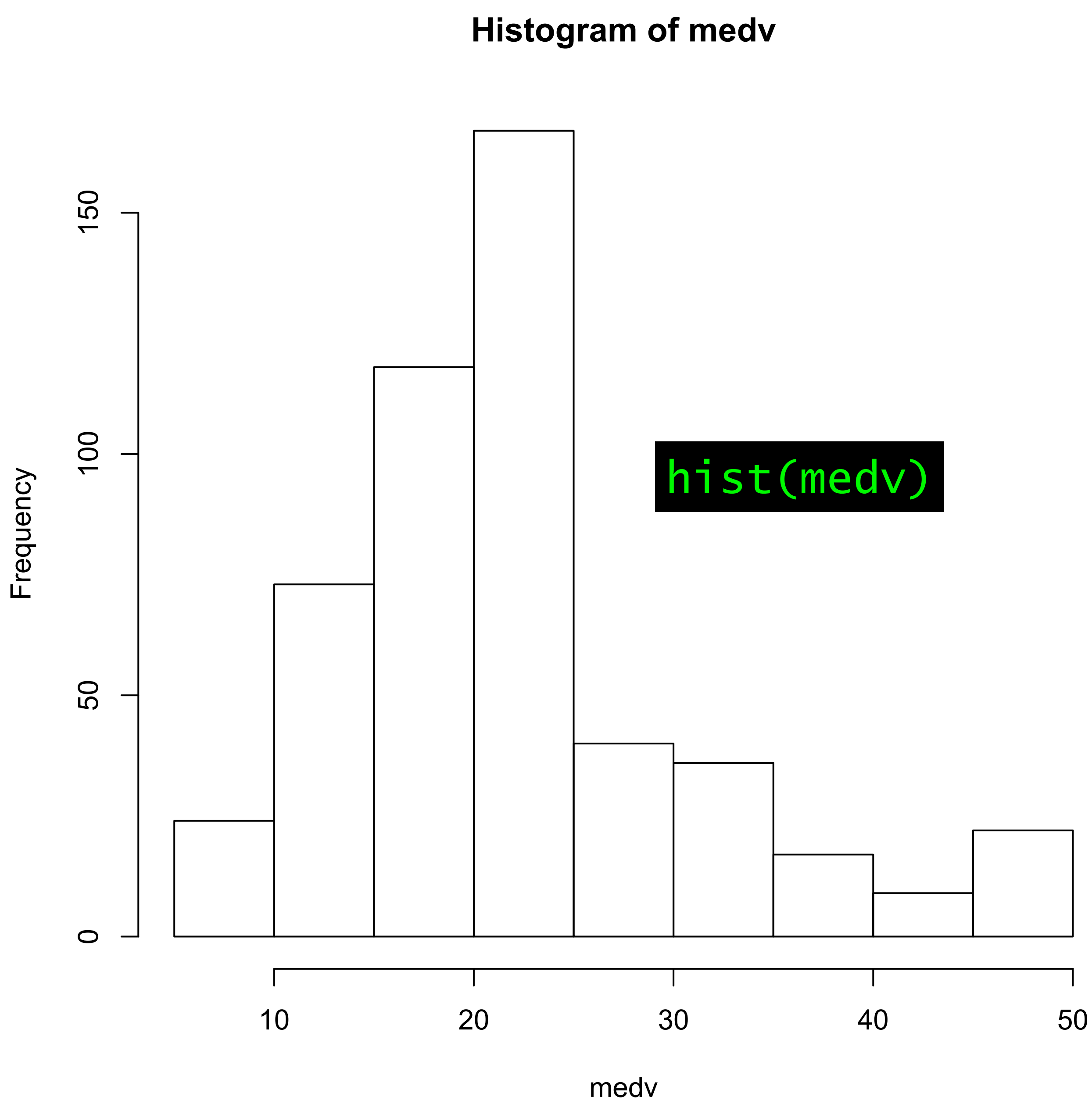
- A histogram is probably the most informative graphical summary about a variable.




```
hist(medv)
```

Histograms

- A histogram is probably the most informative graphical summary about a variable.

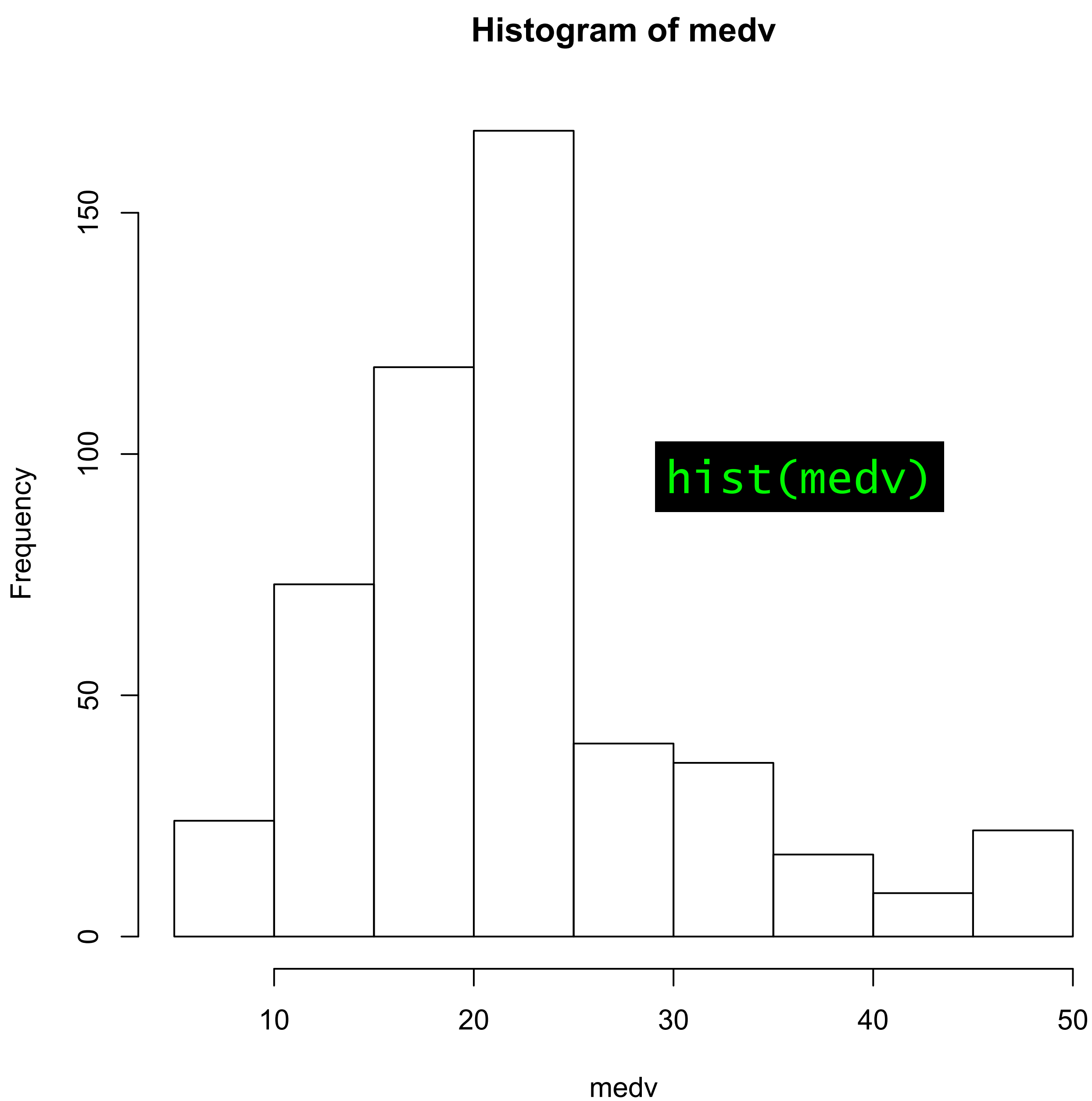


Comments:

- Peak in 20-25
- 25 seems to be a cut-off
- Few houses below 10

Histograms

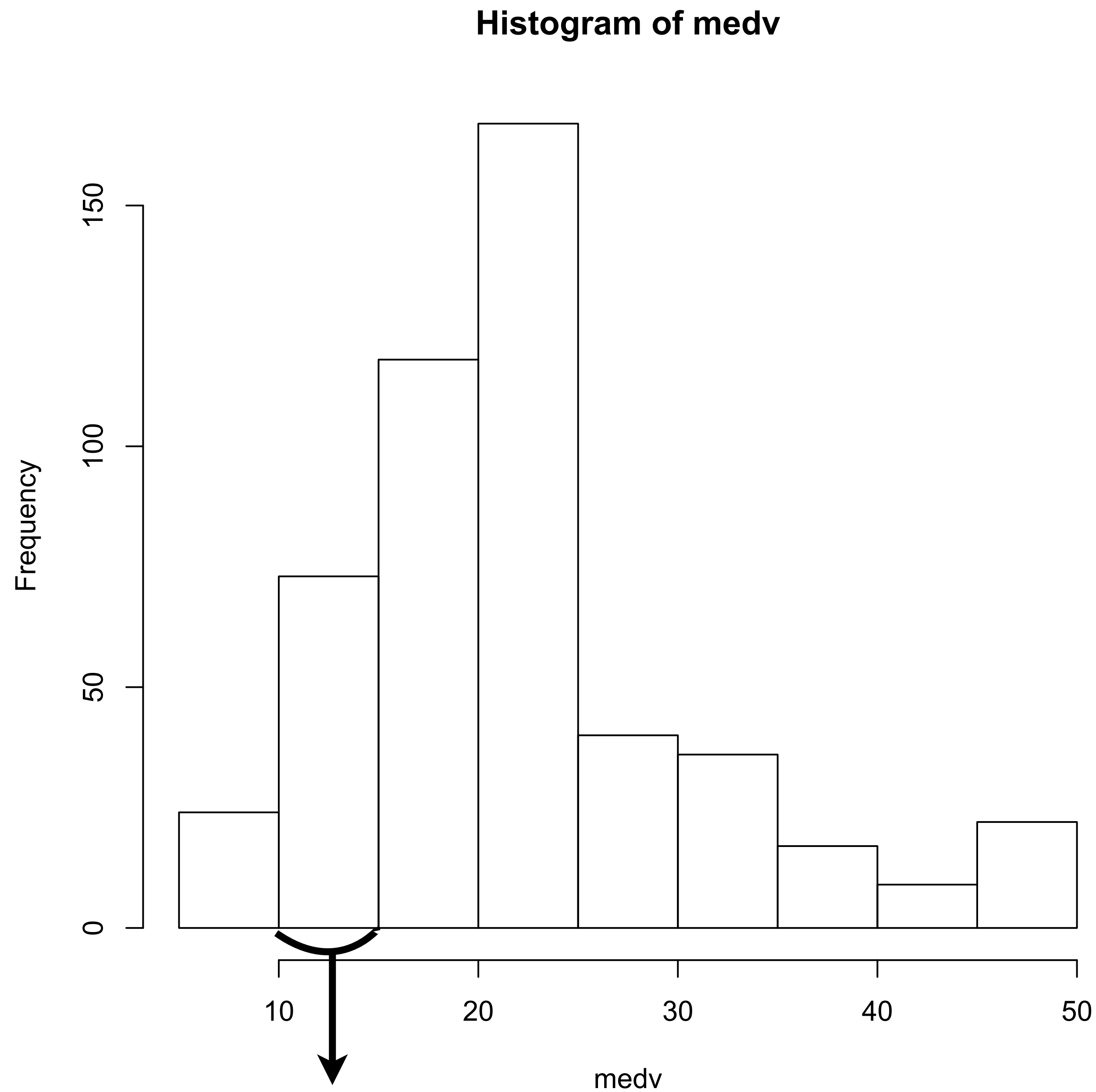
- A histogram is probably the most informative graphical summary about a variable.



Comments:
- Peak in 20-25
- 25 seems to be a cut-off
- Few houses below 10

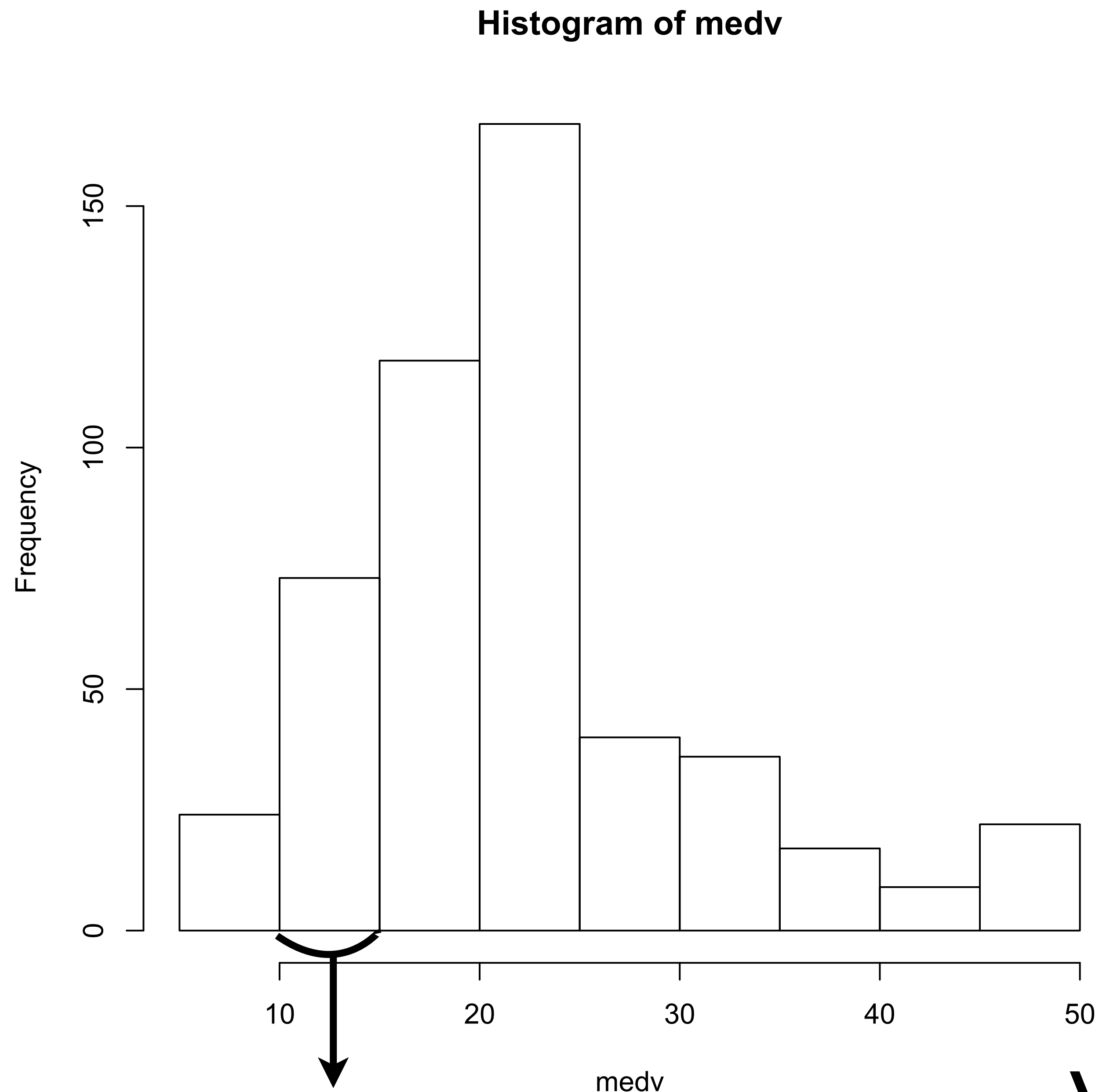
To read it better, we need to know how it's made.

Histograms



This a bin (or class).
Here: tracts where
10<medv<15

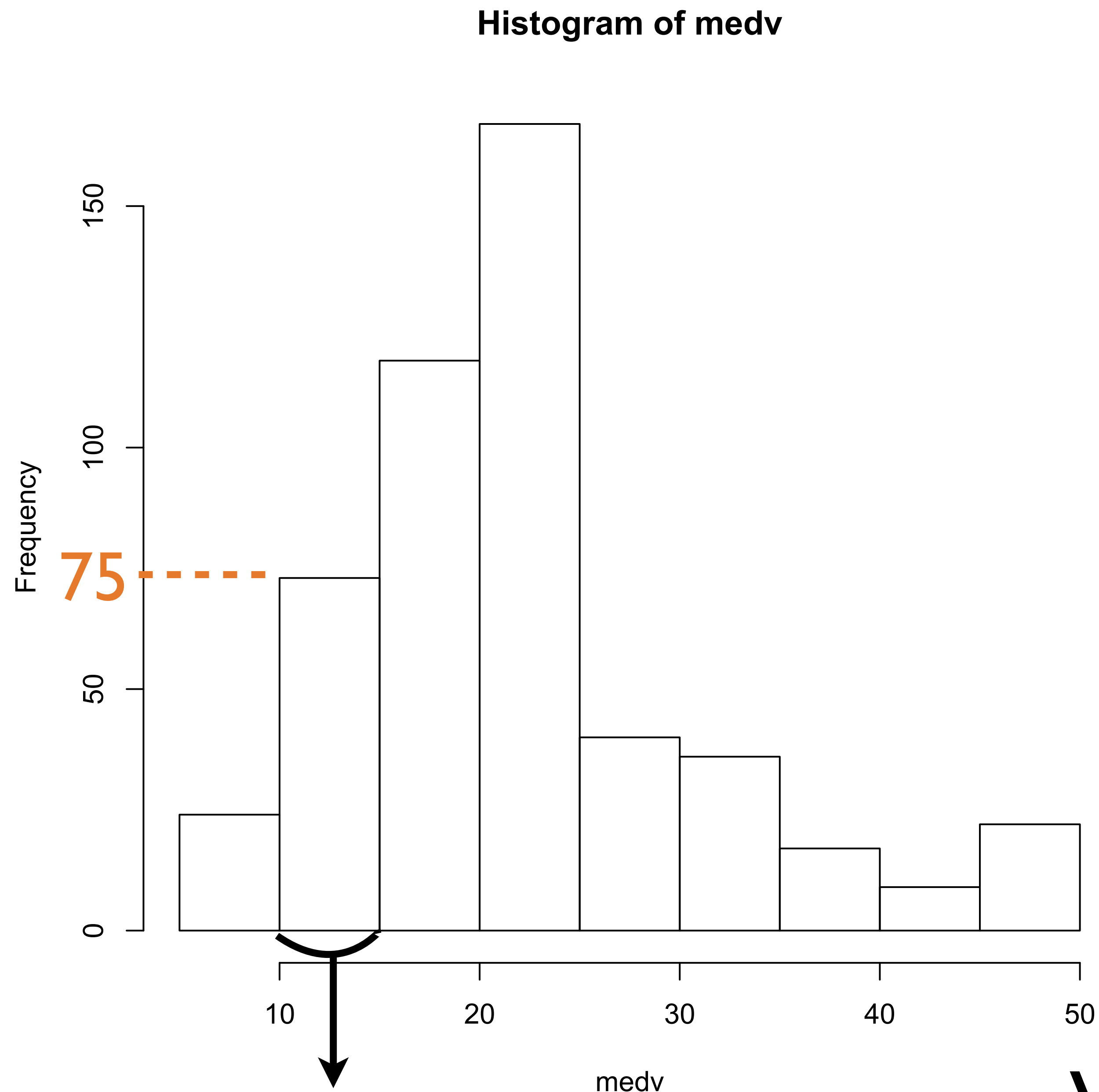
Histograms



This a bin (or class).
Here: tracts where
10<medv<15

We can read on the histogram the number of such census tracts

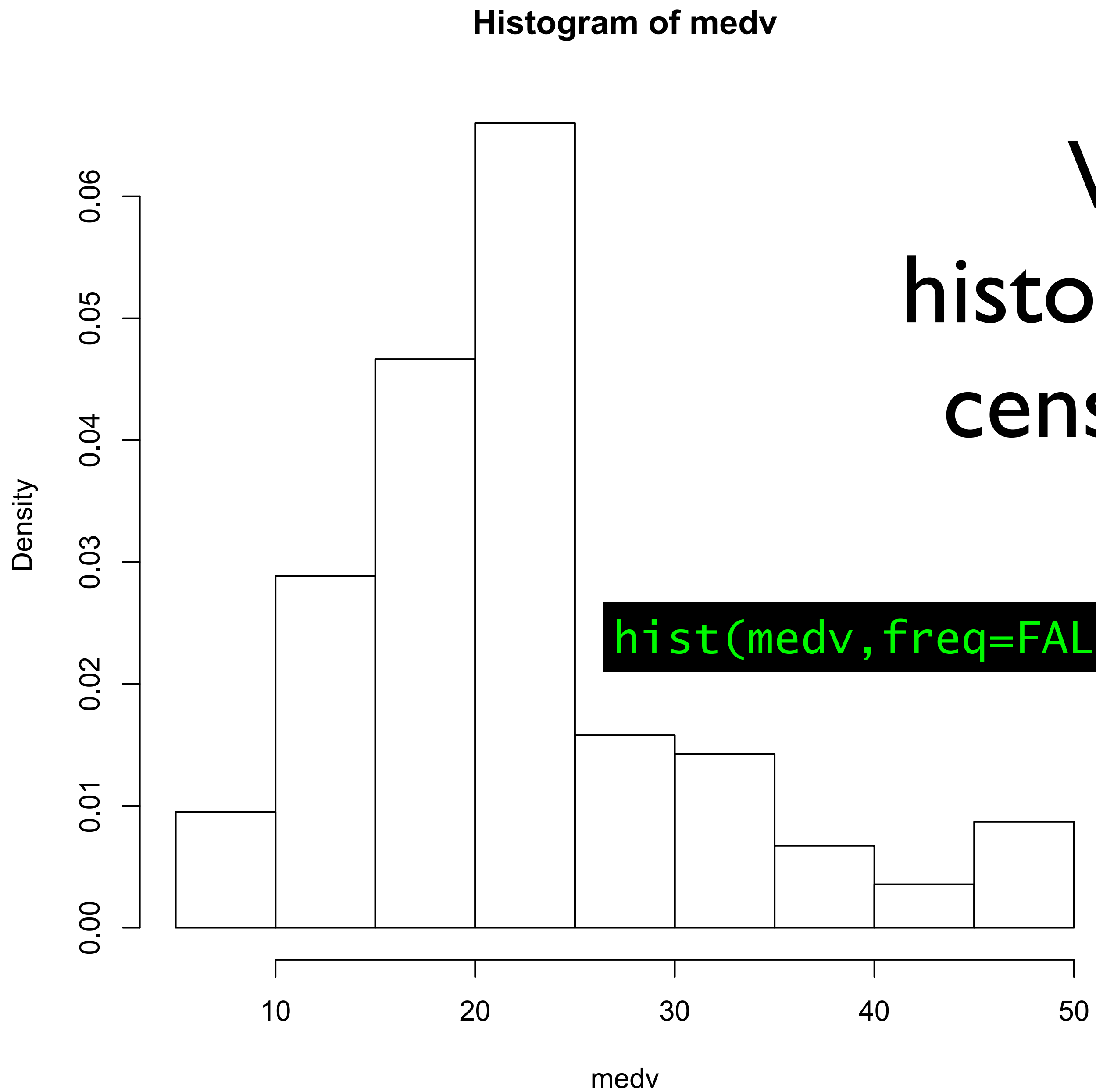
Histograms



This a bin (or class).
Here: tracts where
10<medv<15

We can read on the histogram the number of such census tracts

Histograms with frequencies

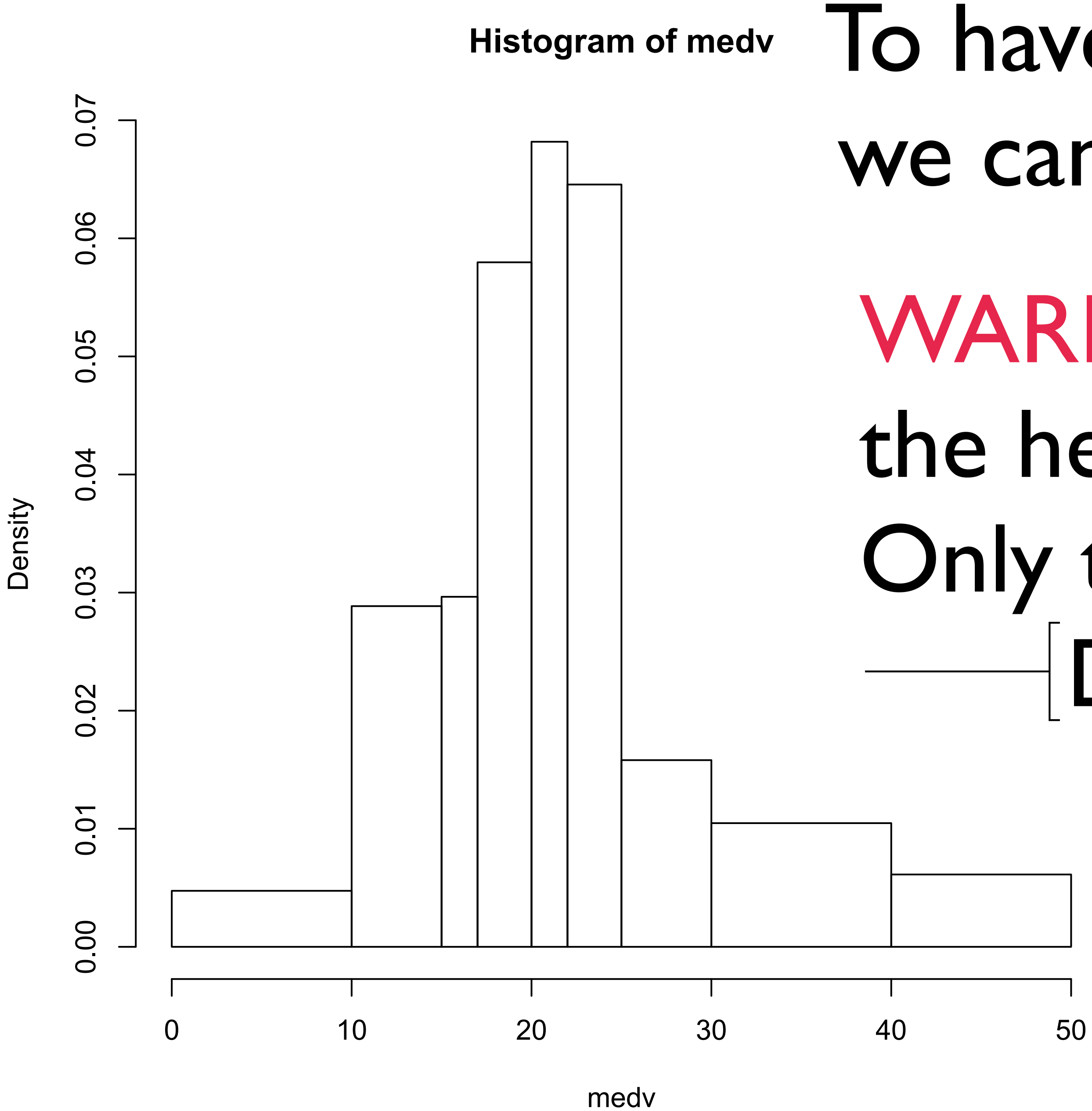



```
hist(medv,freq=FALSE)
```

We can read on the histogram the proportion of census tracts in a bin/class

Histograms with different bin size

MTH 183 - Prof. Bradic Winter 2013

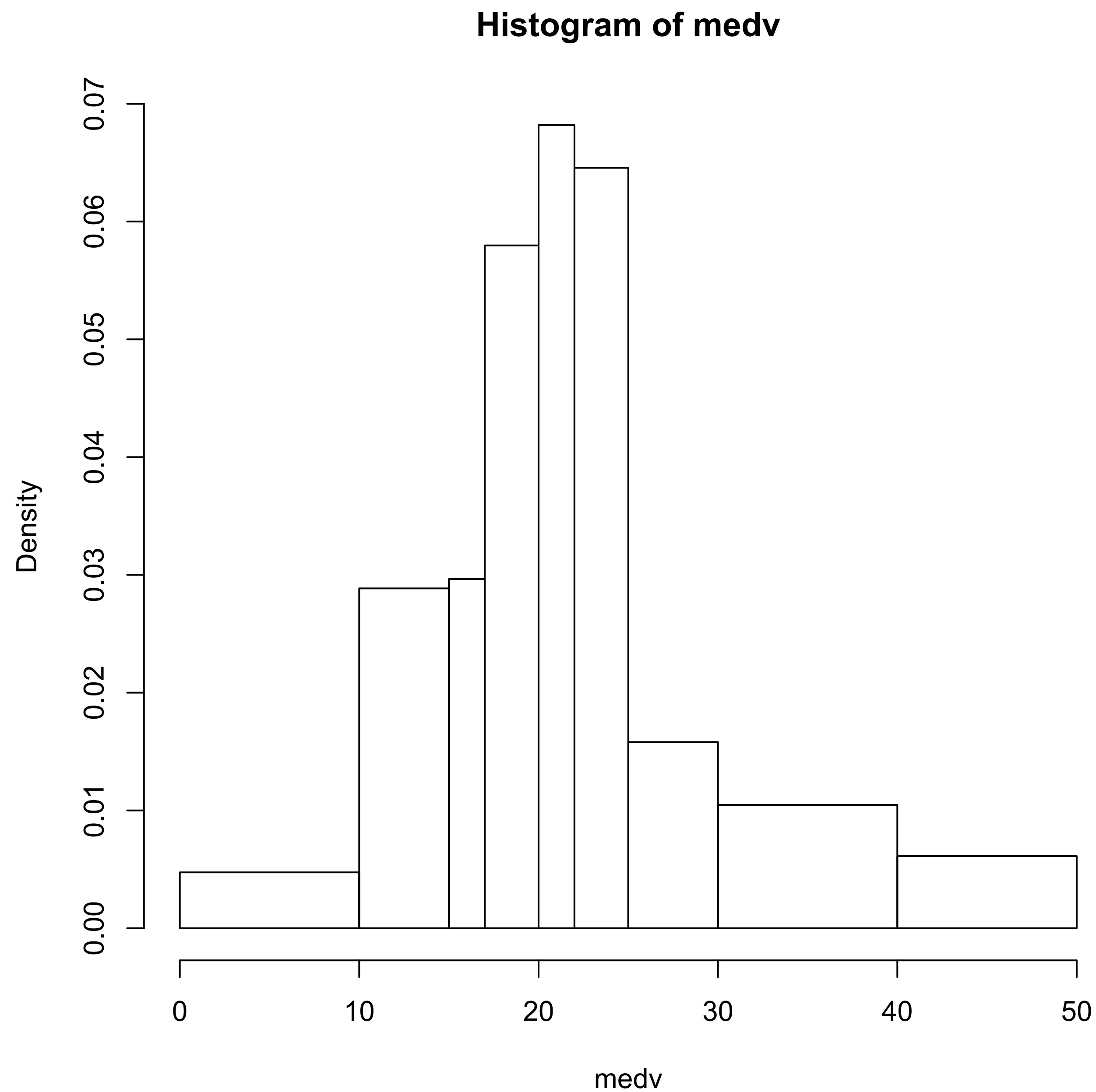


To have a better picture, we can change the bins.

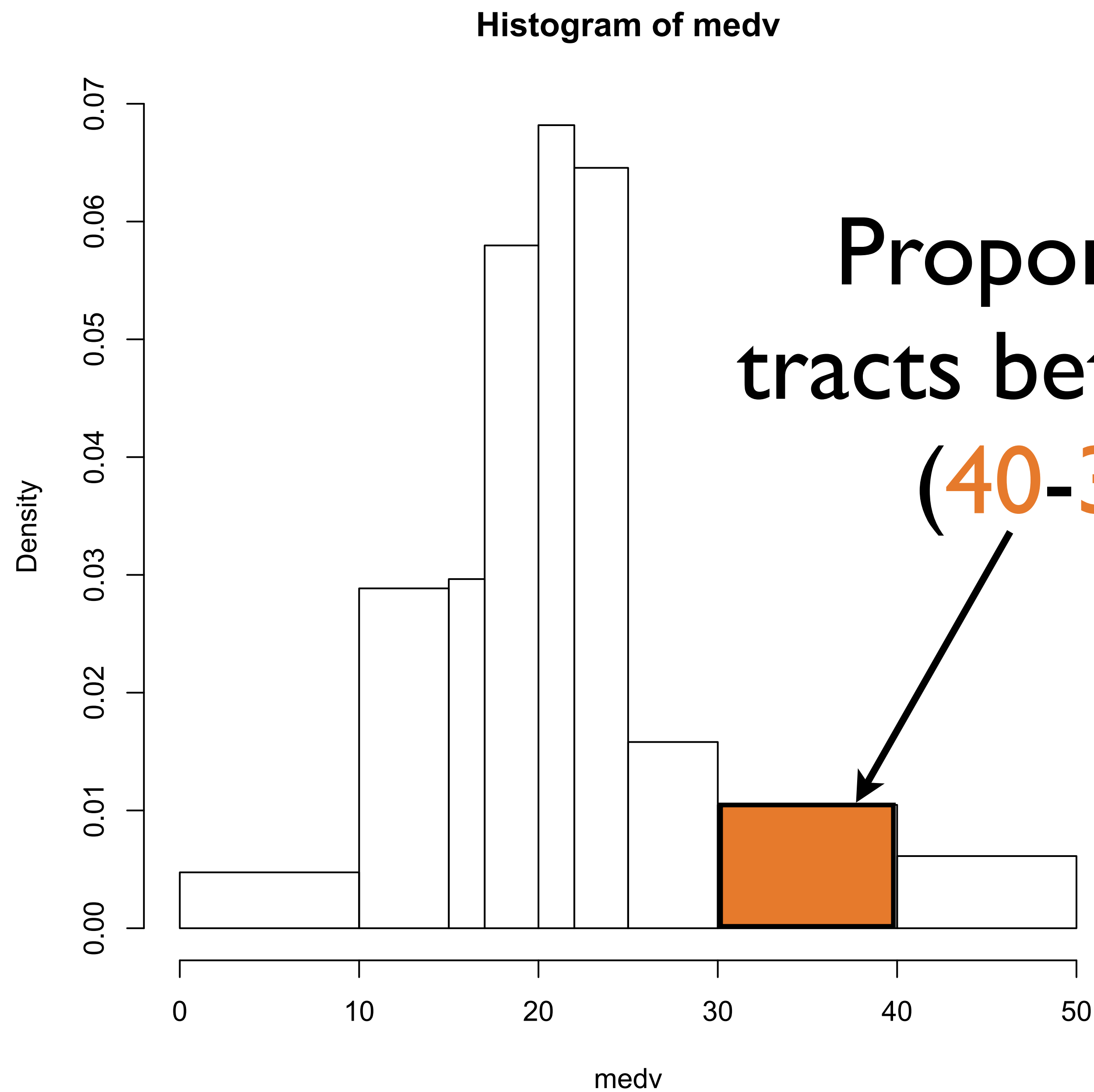
WARNING:
the height of the bar is meaningless
Only the area matters
Data density!

```
hist(medv, freq=FALSE, breaks=c(0, 10, 15, 17, 20, 22, 25, 30, 40, 50))
```

Histograms with different bin size

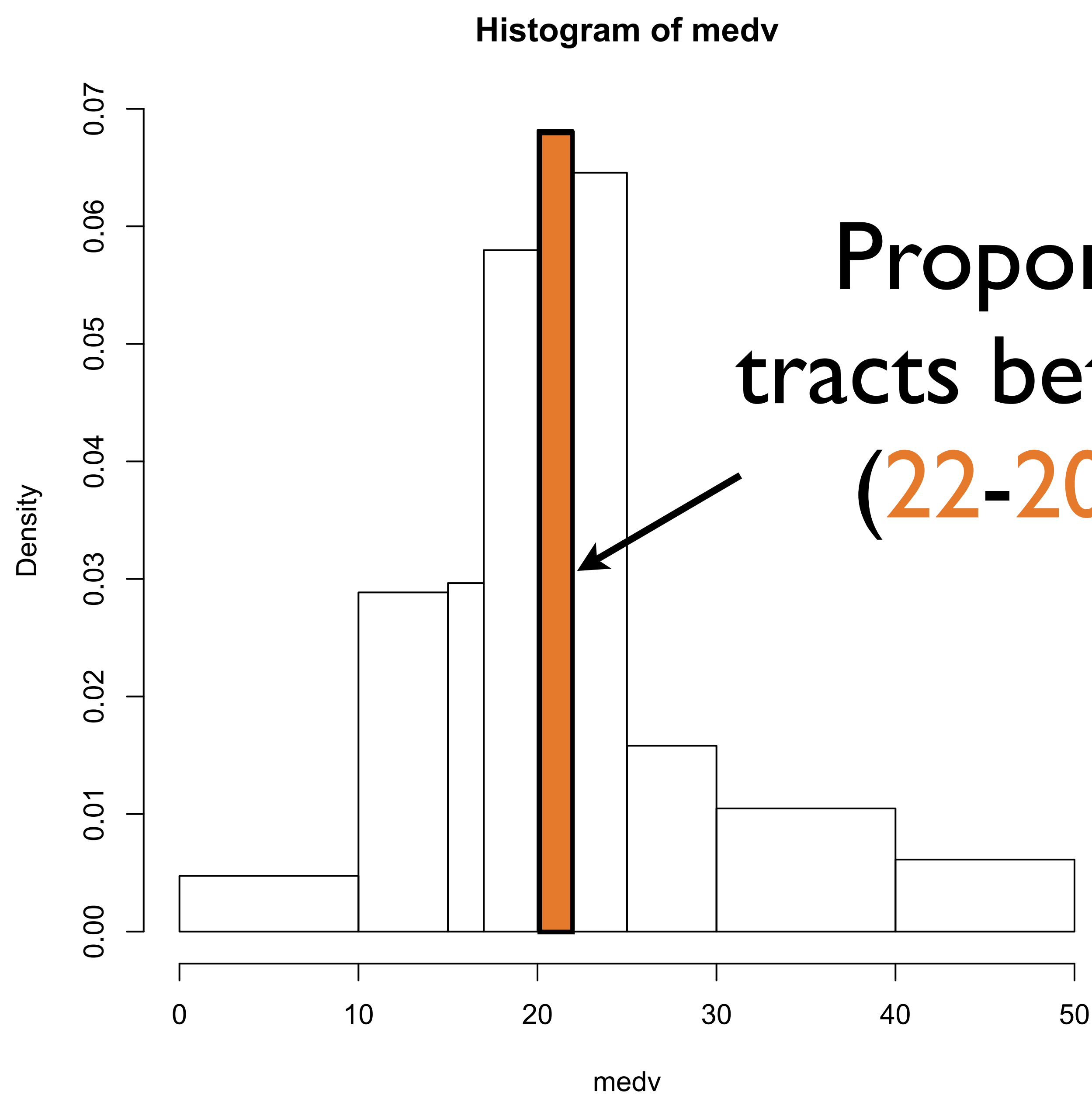


Histograms with different bin size



Proportion of census tracts between 30 and 40: (40-30)0.01=10%

Histograms with different bin size

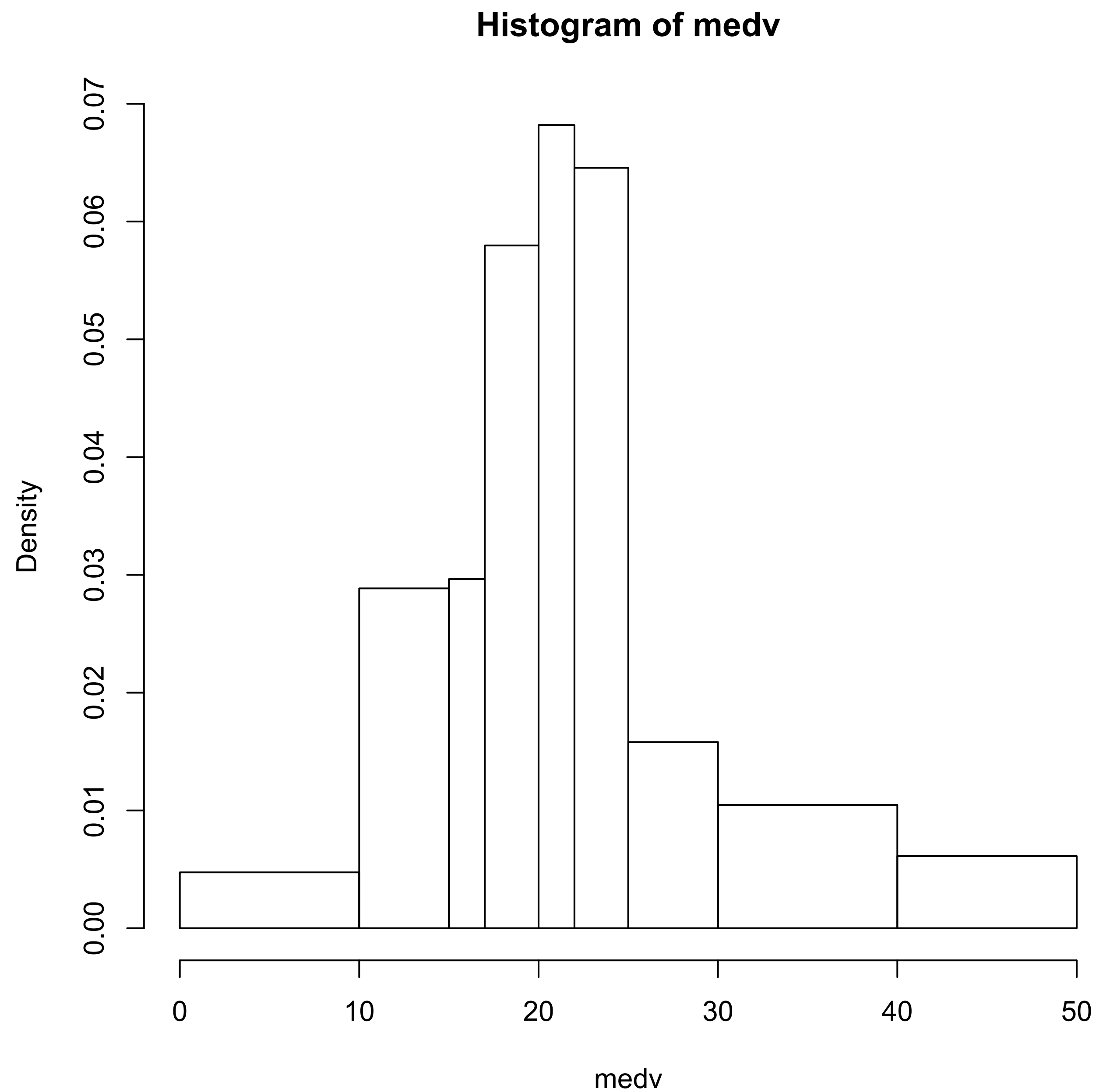


Proportion of census tracts between 20 and 22:
(22-20)0.068=13.6%

Questions:

1. where is the largest proportion? 20-22 or 22-25
2. What is the proportion of medv>30?

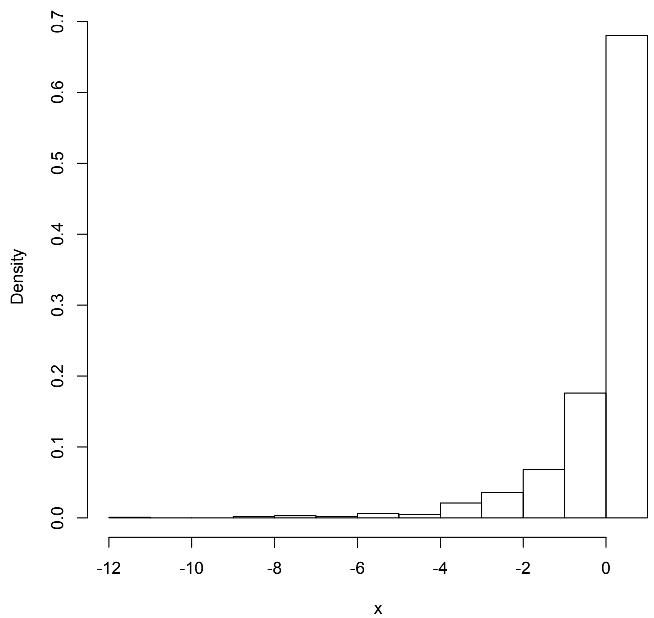
Shape



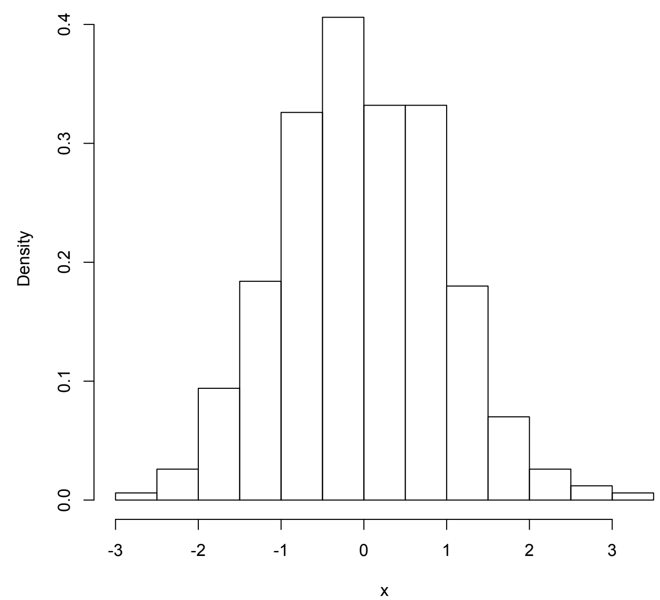
This histogram is said to be

- Skewed to the right (or right skewed)
- The data tend to trail off to the right

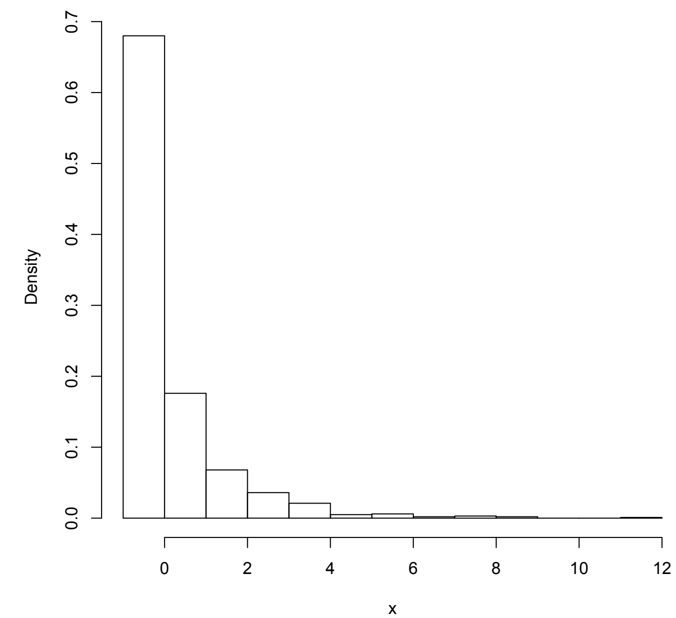
Shape



Left skewed



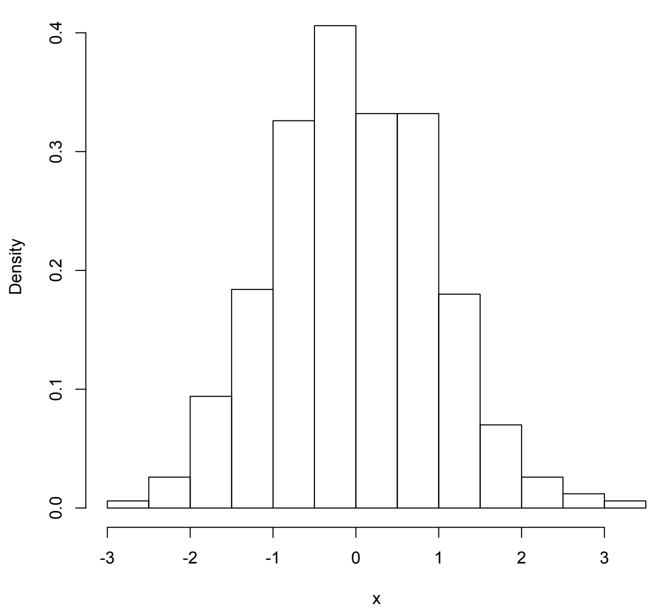
Symmetric



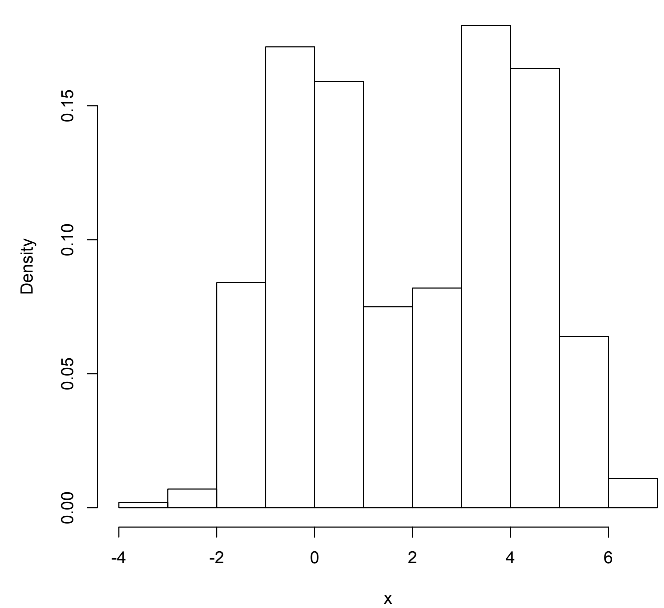
Right skewed

Modes

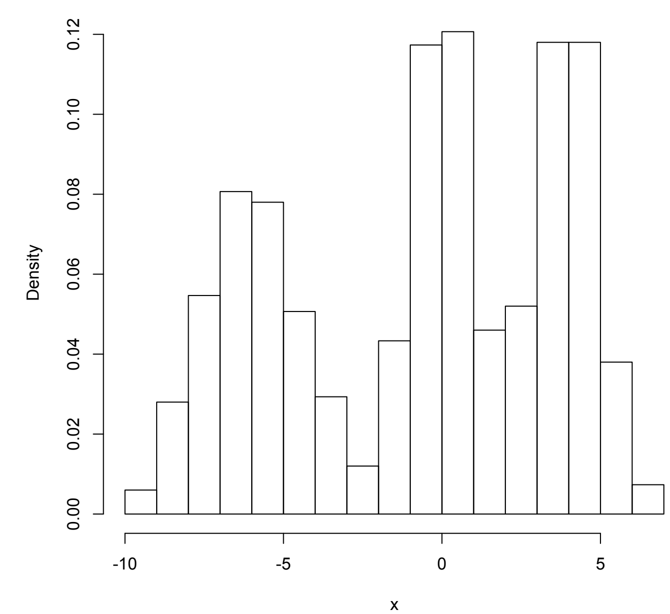
A mode is a peak in the distribution. It indicates a highly populated bin.
We can have **one**, **two**, or **more than two** modes indicating that there may be one, two or more sub-populations in our data.



Unimodal



Bimodal



Multimodal

Summary statistics

	crim	chas	...	age	town	rad	medv
1	0.00632	0	...	65.2	Nahant	1	24.0
2	0.02731	0	...	78.9	Swampscott	2	21.6
3	0.02729	1	...	61.1	Swampscott	2	34.7
4	0.03237	0	...	45.8	Marblehead	6	33.4
...	...	...	...	...	...	...	...
505	0.10959	1	...	89.3	Winthrop	1	22.0
506	0.04741	0	...	80.8	Winthrop	1	11.9

Summary statistics

crim
0.00632
0.02731
0.02729
0.03237
...
0.10959
0.04741

Numbers that summarize crim:

Mean	3.61	`summary(crim)`
Median	0.26	
Min	0.01	`mean(crim)` `median(crim)` `min(crim)` `max(crim)` `quantile(crim,.25)` `quantile(crim,.75)`
Max	88.98	
1st quartile	0.08	
3rd quartile	3.68	
Standard deviation	8.60	`sd(crim)`
Variance	73.99	`var(crim)`

A little reminder

Consider a list of numbers: $x_1, x_2, \ldots, x_n$

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

$$\sum_{i=1}^{4} i = 1 + 2 + 3 + 4 = 10$$

$$\left(\sum_{i=1}^{4} i\right)^2 = (1 + 2 + 3 + 4)^2 = 100$$

$$\sum_{i=1}^{4} i^2 = 1^2 + 2^2 + 3^2 + 4^4 = 30$$

A little reminder

$$\sum_{i=1}^{4}(2i) = \qquad\qquad = 20$$

$$\sum_{i=1}^{4}(2+i) = \qquad\qquad = 18$$

Measures of location

Mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Median: the number in the middle

Order the list $x_1, x_2, \ldots, x_n$ into $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$

Then median: $x_{med} = \begin{cases} x_{(\frac{n+1}{2})} & \text{if n odd} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & \text{if n even} \end{cases}$

Example: variable age

65.2 78.9 61.1 45.8 89.3 80.3

Median: n even

45.8 61.1 65.2 78.9 80.3 89.3

Median: n even

$$45.8 \leq 61.1 \leq 65.2 \leq 78.9 \leq 80.3 \leq 89.3$$

Median: n even

$$\begin{array}{ccccccccccc} x_{(1)} & & x_{(2)} & & x_{(3)} & & x_{(4)} & & x_{(5)} & & x_{(6)} \\ 45.8 & \leq & 61.1 & \leq & 65.2 & \leq & 78.9 & \leq & 80.3 & \leq & 89.3 \end{array}$$

Median: n even

$x_{(1)}$		$x_{(2)}$		$x_{(3)}$		$x_{(4)}$		$x_{(5)}$		$x_{(6)}$
45.8	$\leq$	61.1	$\leq$	65.2	$\leq$	78.9	$\leq$	80.3	$\leq$	89.3

Median: n even

$x_{(1)}$		$x_{(2)}$		$x_{(3)}$		$x_{(4)}$		$x_{(5)}$		$x_{(6)}$
45.8	$\leq$	61.1	$\leq$	65.2	$\leq$	78.9	$\leq$	80.3	$\leq$	89.3

n=6 even: $x_{med} = \frac{x_{(3)} + x_{(4)}}{2} = \frac{65.2 + 78.9}{2} = 72.05$

Median: n odd

$$\begin{array}{ccccccccc} x_{(1)} & & x_{(2)} & & x_{(3)} & & x_{(4)} & & x_{(5)} \\ 45.8 & \leq & 61.1 & \leq & 65.2 & \leq & 78.9 & \leq & 80.3 \end{array}$$

n=5 odd: $x_{med} = x_{(3)} = 62.5$

Quartiles

Median splits the list in half. Quartiles split it into 1/4 - 3/4 and 3/4 - 1/4

$Q_1 \simeq x_{(\frac{n}{4})}$

First quartile

$Q_3 \simeq x_{(\frac{3n}{4})}$

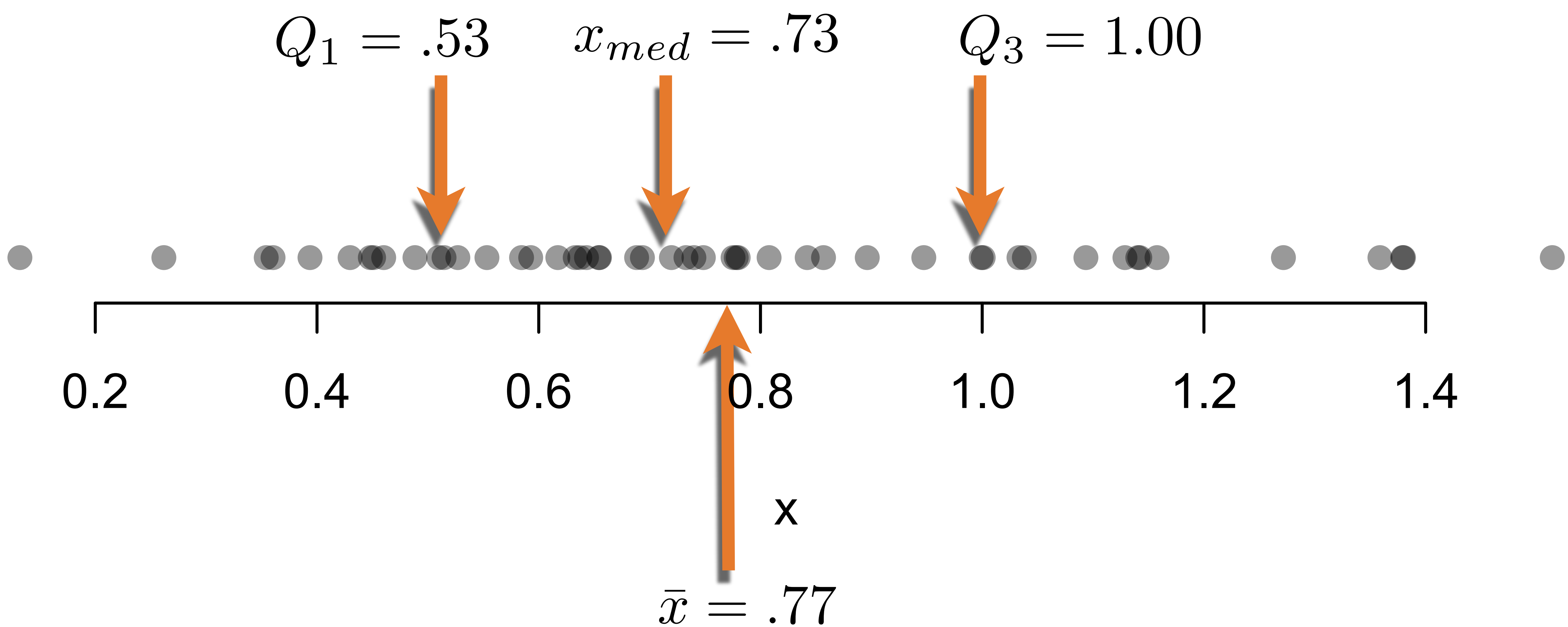
Third quartile

What is the second quartile?

A fourth of the data is smaller than Q_1

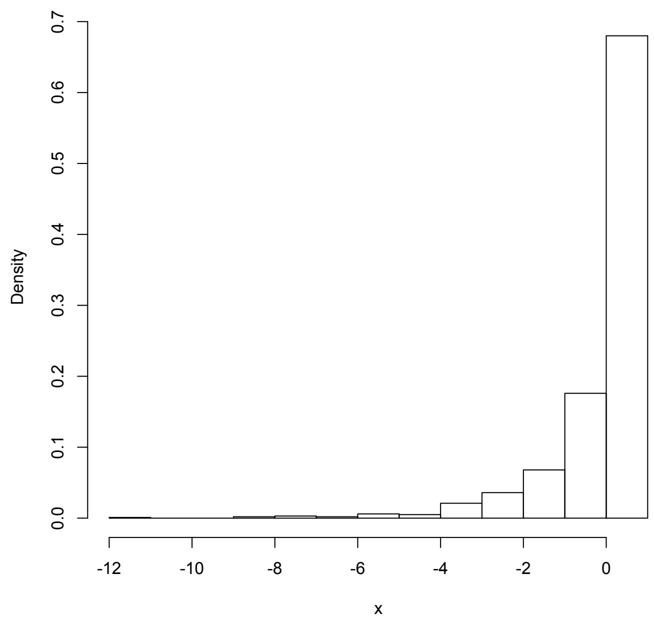
A fourth of the data is larger than Q_3

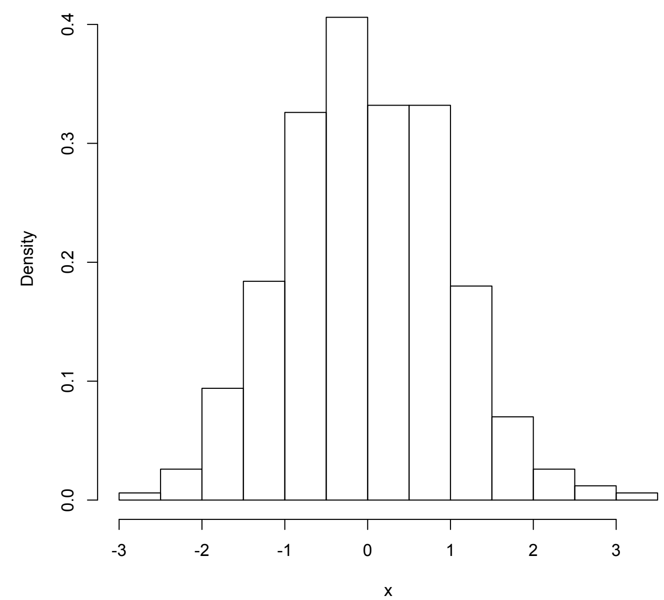
On a dot plot

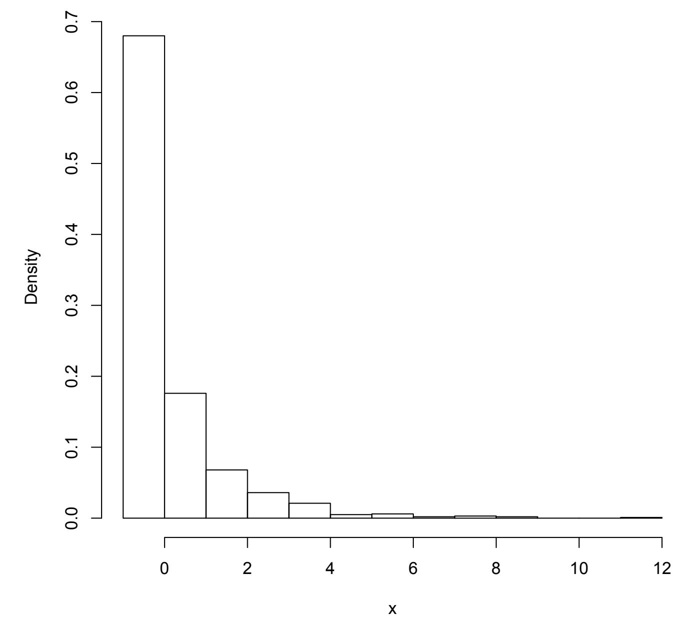



```
library(openintro)
dotPlot(x, ylab="")
```

Shape, mean and median







Left skewed — Symmetric — Right skewed

Mean $<$ median — Mean $\simeq$ median — Mean $>$ median

Measures of dispersion

Variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Standard deviation: $s = \sqrt{s^2}$ same units as x_i

Measures of dispersion

Variance: $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Standard deviation: $s = \sqrt{s^2}$ same units as x_i

If all the numbers in the list are close to each other, they are close to their mean $\bar{x}$ and $(x_i - \bar{x})^2$ is close to zero. The variance and (standard deviation) measure how the data is "clustered" around its mean $\bar{x}$

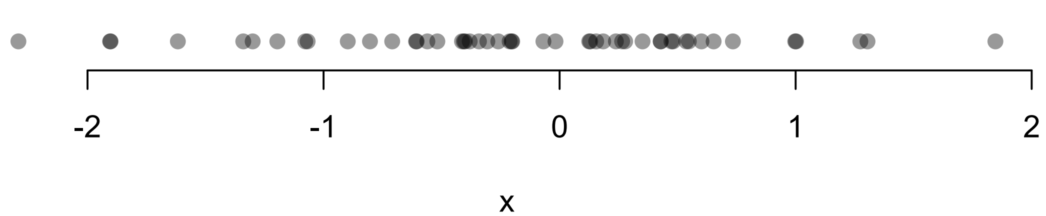
Measures of dispersion

Variance: $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

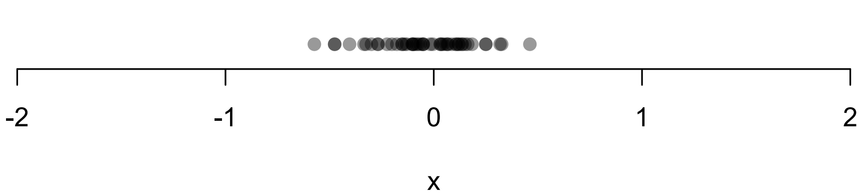
Standard deviation: $s = \sqrt{s^2}$ same units as x_i

If all the numbers in the list are close to each other, they are close to their mean $\bar{x}$ and $(x_i - \bar{x})^2$ is close to zero. The variance and (standard deviation) measure how the data is "clustered" around its mean $\bar{x}$

$\bar{x} = 0 \quad s = 0.90$



$\bar{x} = 0 \quad s = 0.22$

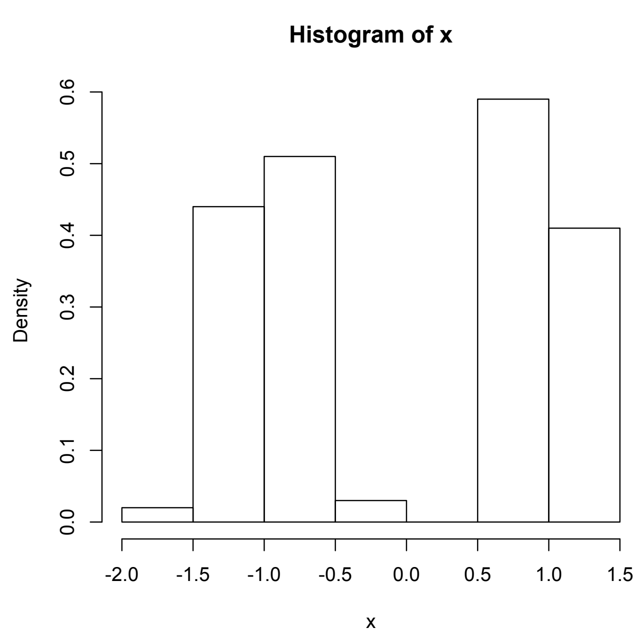


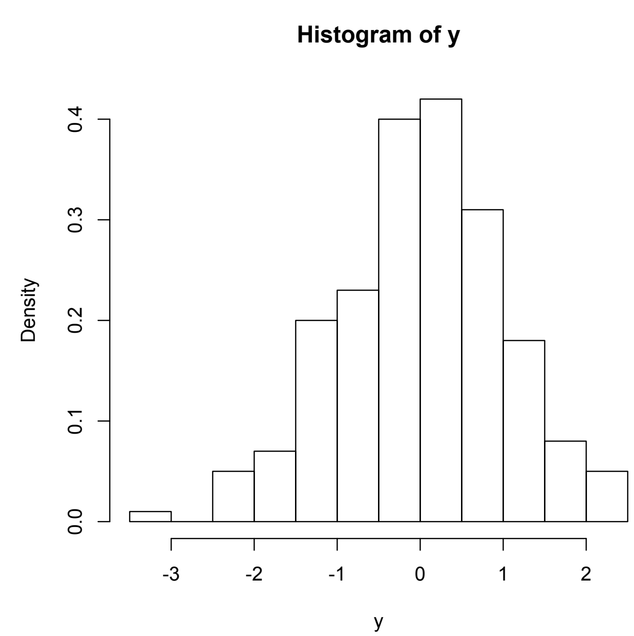
Measures of dispersion

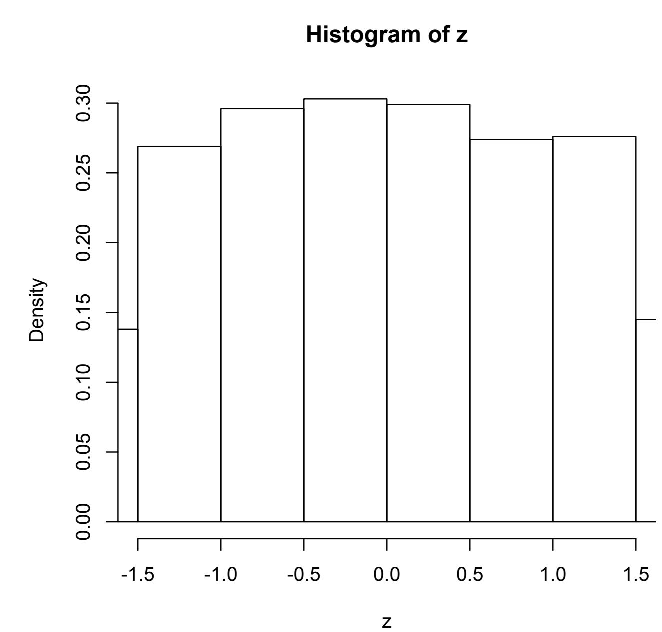
Interquartile range (IQR) = $Q_3 - Q_1$

Is also a measure a dispersion.

Measure of location and dispersion say nothing about shape. Data x, y and z all have mean 0 and SD 1.

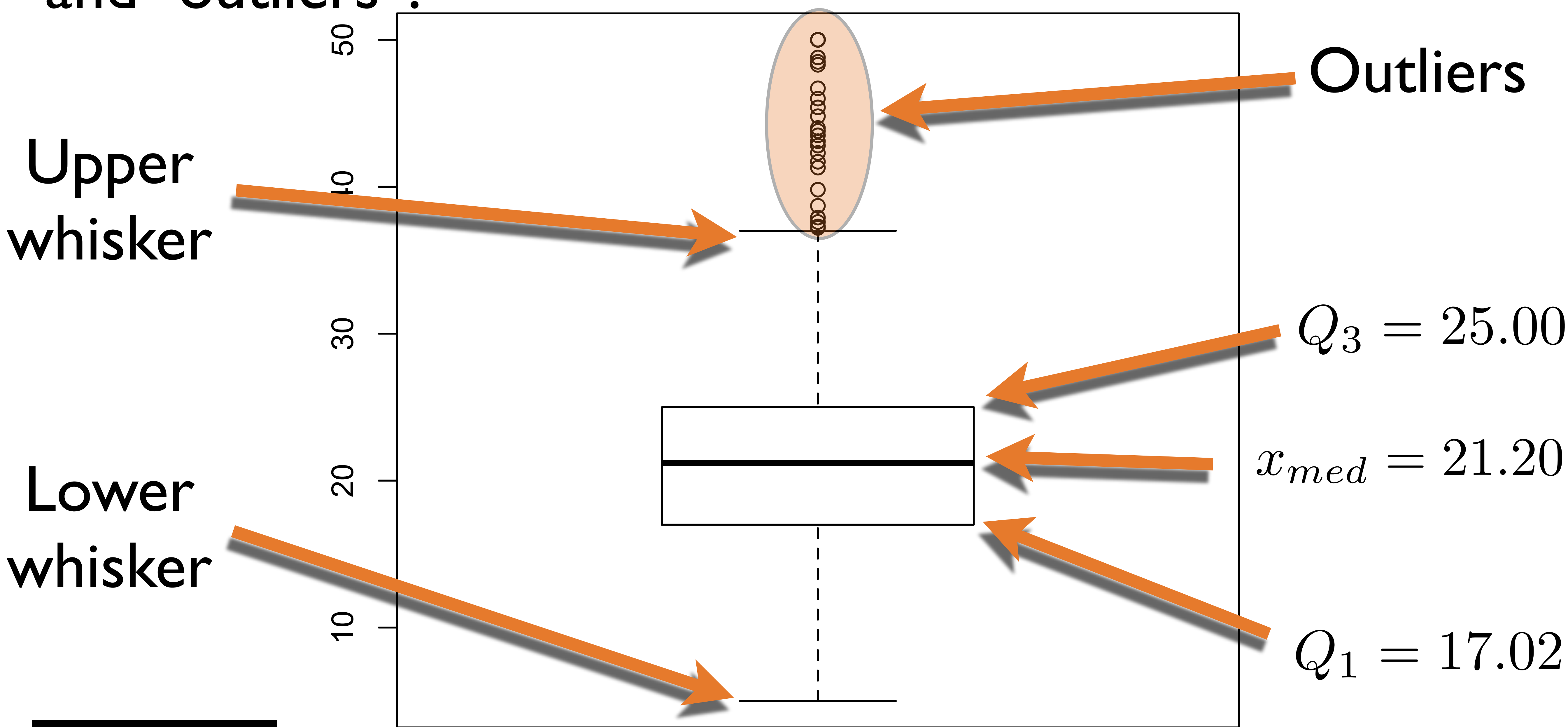






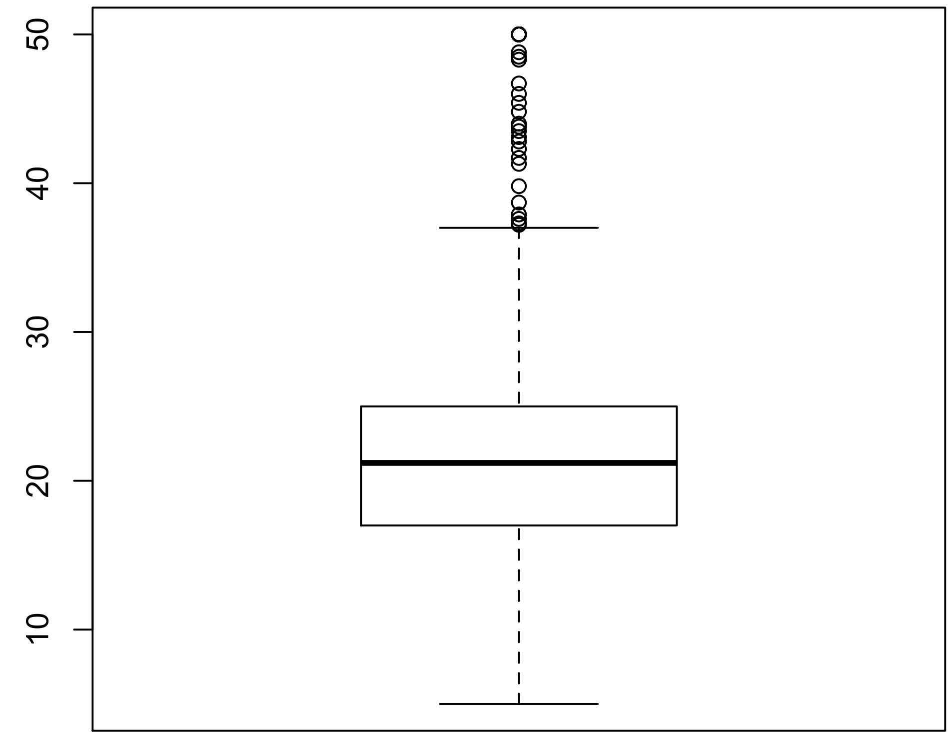
Boxplots and outliers

A **boxplot** is a plot that summarizes quartiles, median and "outliers".




```
boxplot(medv)
```

Boxplots and outliers



Upper whisker $Q_3 + 1.5 \cdot IQR$

Lower whisker $Q_1 - 1.5 \cdot IQR$

Outliers Observations outside of the whiskers

Note: The whiskers draw at the last observation that is not an outlier. This may differ from the exact value $Q_3 + 1.5 \cdot IQR$ or $Q_1 - 1.5 \cdot IQR$ significantly

Outliers

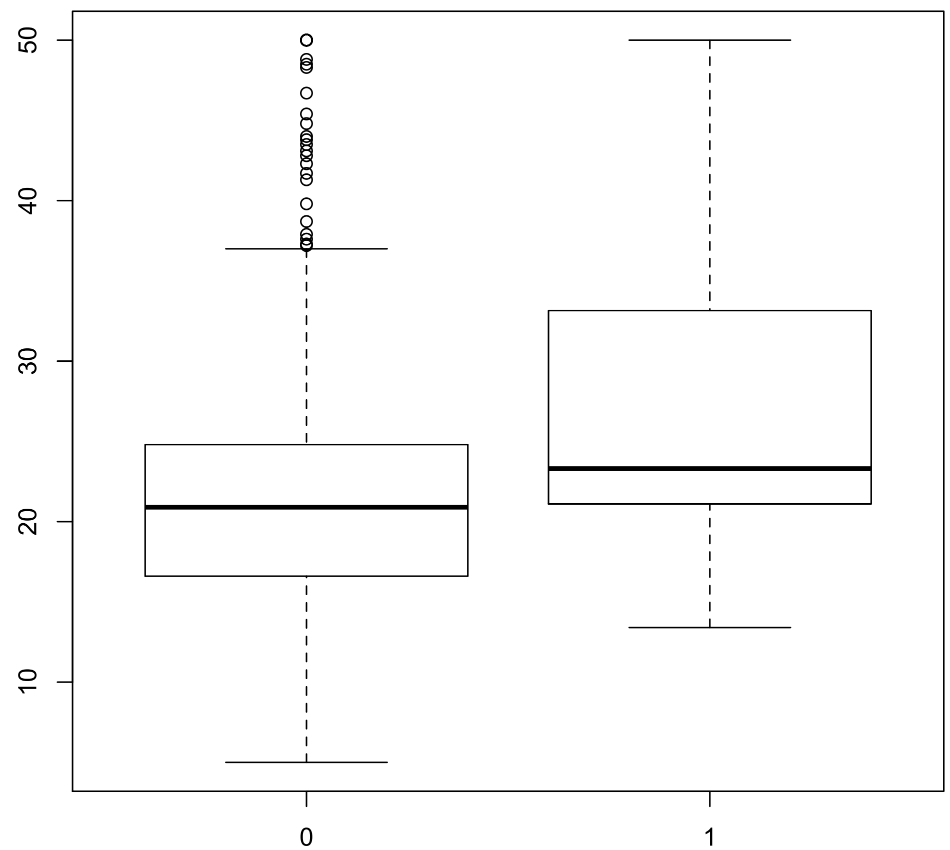
It is important to understand the data to identify why outliers are outliers. Possible sources:

- Number incorrectly reported
- Heterogeneous data (several populations)

The median and IQR are robust to outliers.
The mean and SD are not.

Comparative boxplots

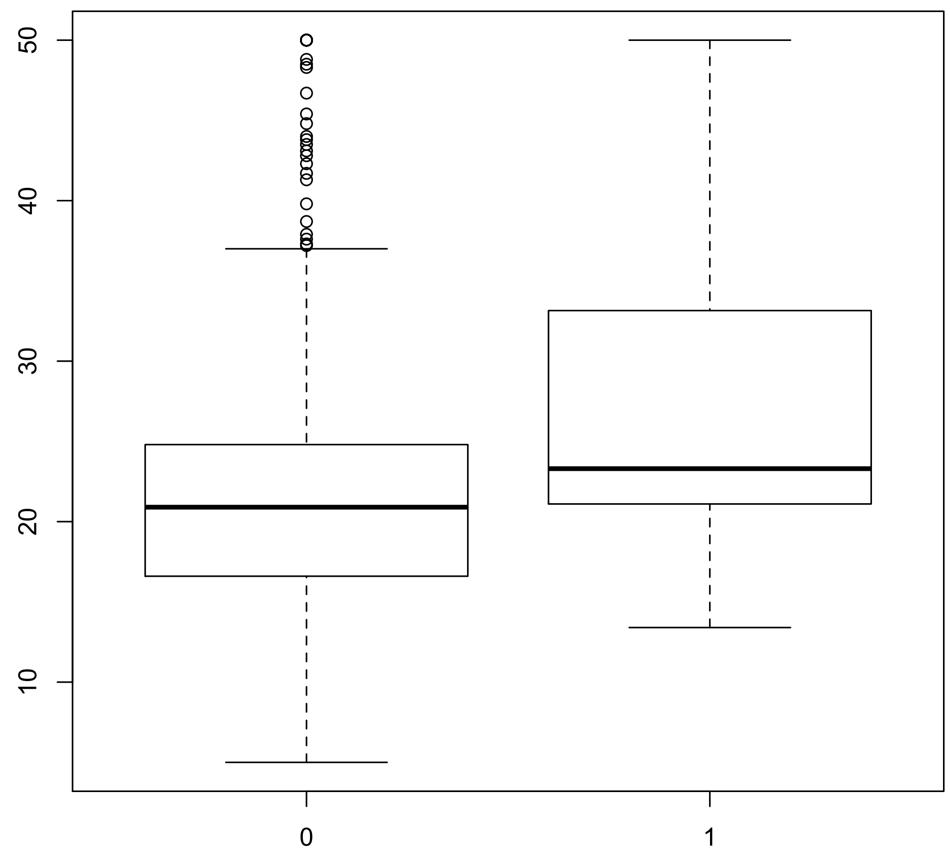
Boxplots become very useful when we want to compare one variable across two categories. For example, we want to know if census tracts along the Charles river have more expansive houses.




```
boxplot(medv[chas==0], medv[chas==1])
```

Comparative boxplots

Boxplots become very useful when we want to compare one variable across two categories. For example, we want to know if census tracts along the Charles river have more expansive houses.




```
boxplot(medv[chas==0], medv[chas==1])
```

```
plot(chas, medv)
```

This should be a scatter plot but R understands that comparative boxplots carry more information

Contingency tables

- For categorical data we cannot compute the mean, variance, median, ...

town
Nahant
Swampscott
Swampscott
Marblehead
...
Winthrop
Winthrop

One way to summarize is to count!

→ Contingency tables

Arlington	Ashland	Bedford	...	Winchester	Winthrop
7	2	2	...	5	5

Counts the number of census tracts in each town

```
table(town)
```

Two-way contingency tables

For two variables, we can cross them.
Consider this dataset about 2,201 passengers of the RMS Titanic.

Survived

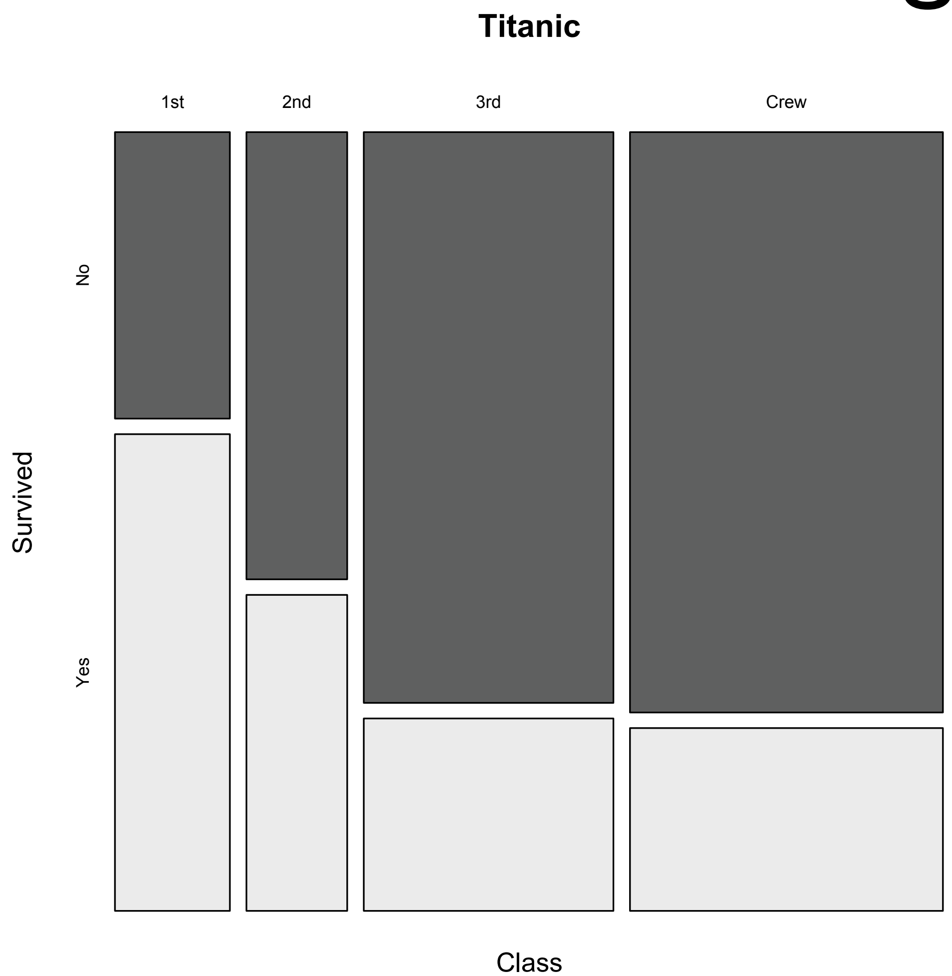
Class

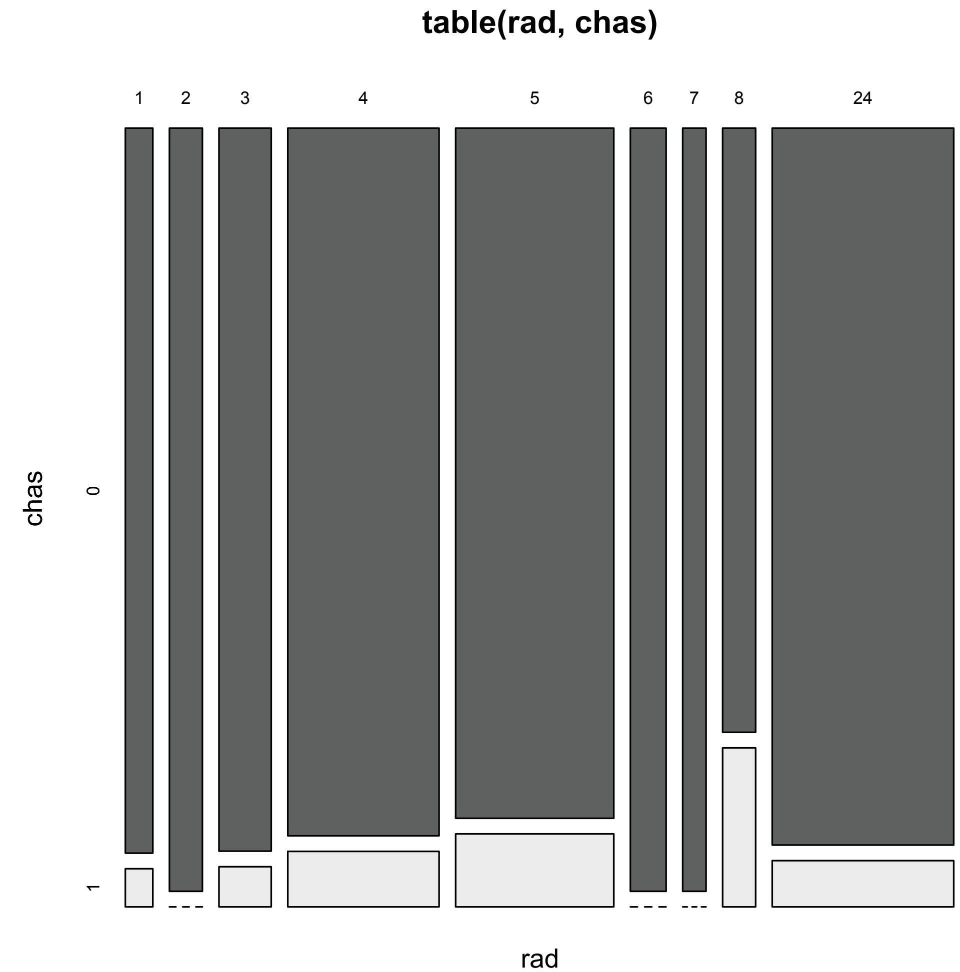
	No	Yes	Total
1st class	122	203	325
2nd class	167	118	285
3rd class	528	178	706
crew	673	212	885
Total	1,490	711	2,201

```
ftable(Titanic, row.vars="Class", col.vars="Survived")
```

Mosaic plots

Useful represent categorical data. Displays the content of a contingency table.

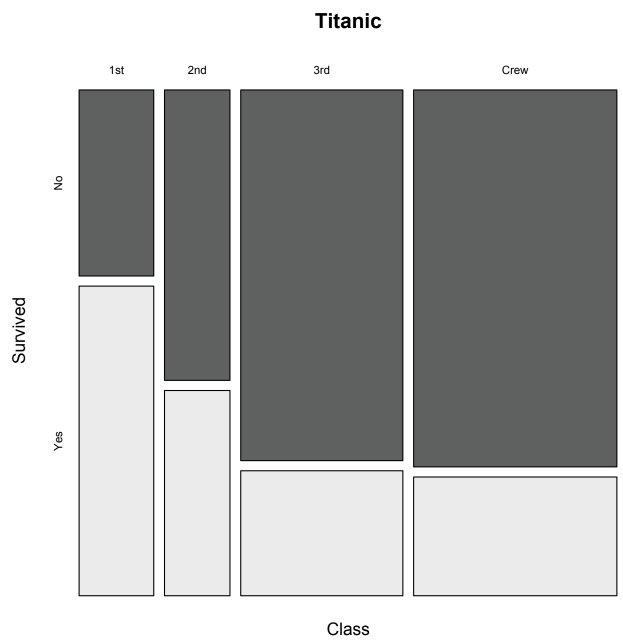




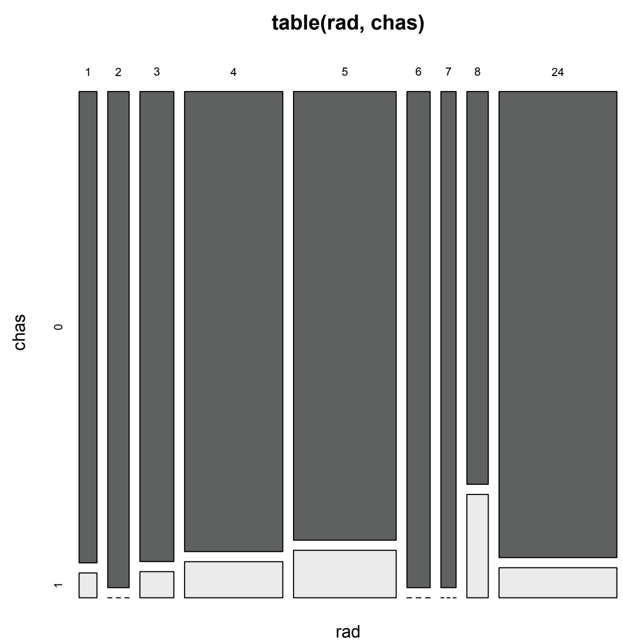

```
mosaicplot(~ Class + Survived, data = Titanic, color = TRUE)
```

```
mosaicplot(table(rad, chas), color = TRUE)
```

Mosaic plots

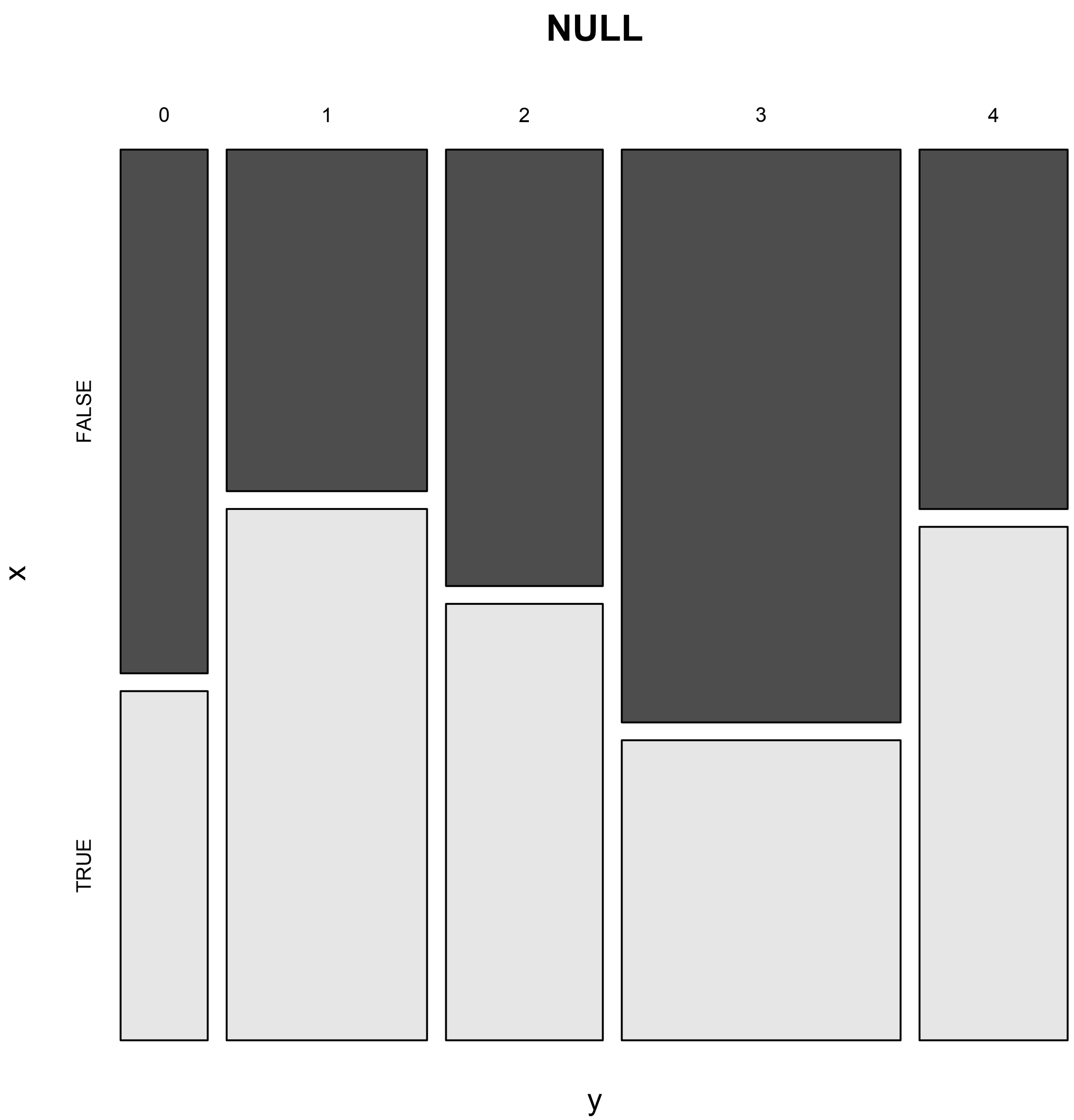


We can see that Class and Survived are associated variables.



chas and rad are associated variables

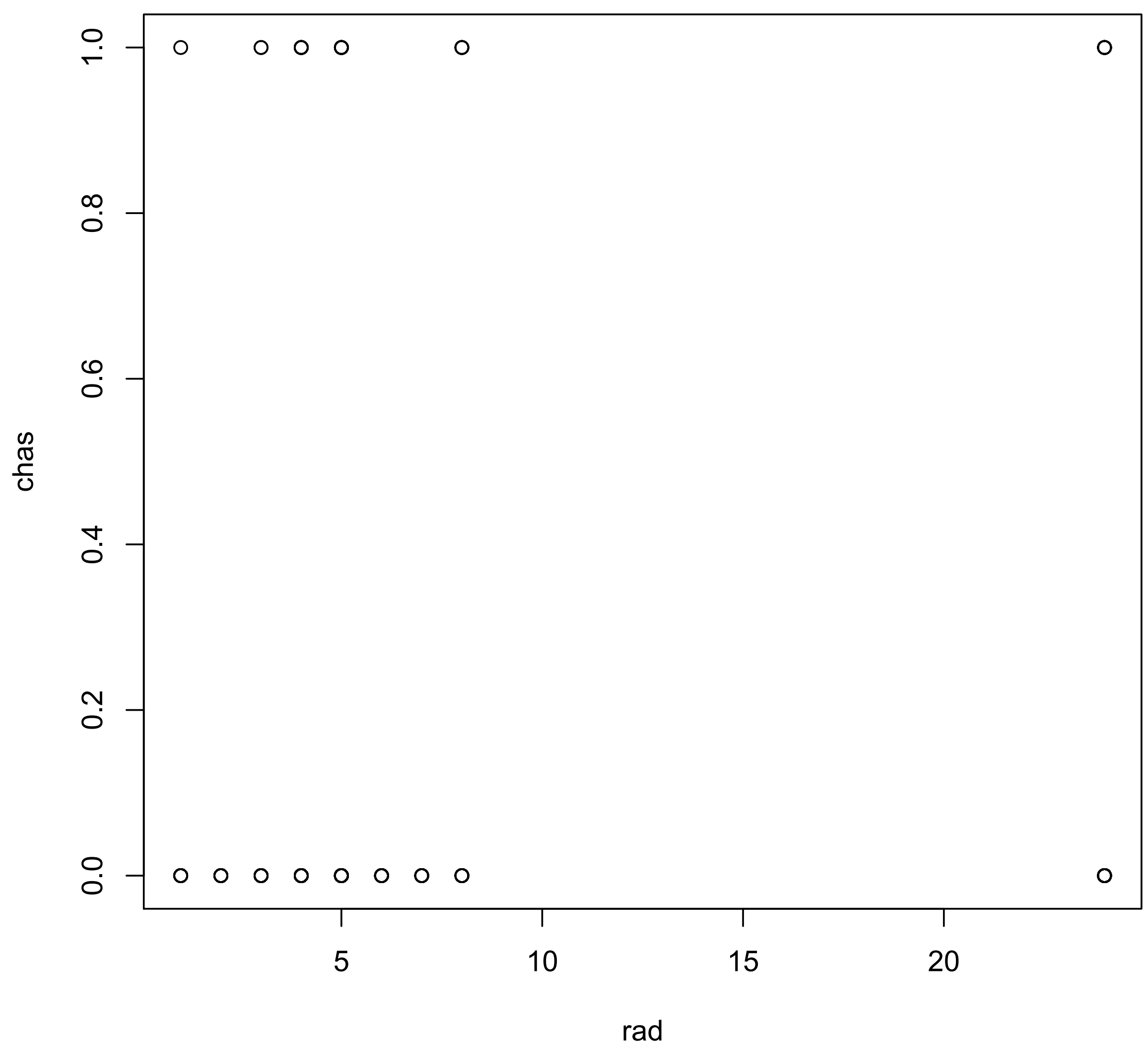
Mosaic plots



These variables seem to be independent

Why mosaic plots?

Arguably, mosaic plots are not the most accurate plot to assess independence. But compare this to the following scatter plot of chas vs rad



Population and samples

- Our Boston Housing dataset is a sample of census tracts in the Boston area.
- All census tracts in the greater Boston area represent the population.
- The census tracts were selected randomly from the population.

Population and samples

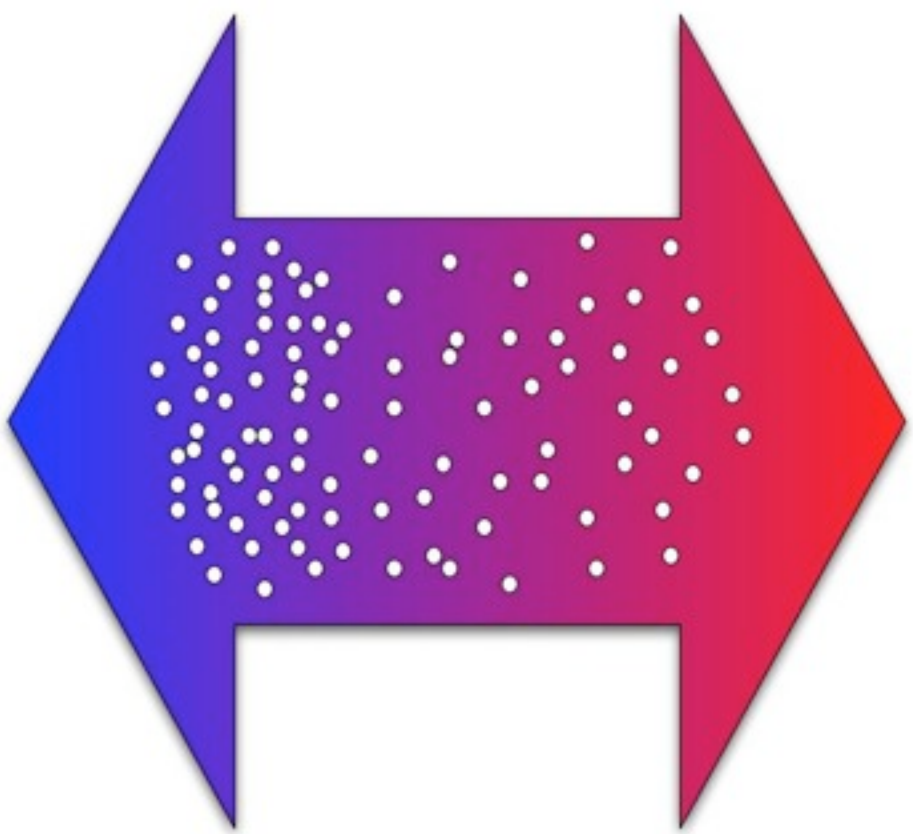
- Our Boston Housing dataset is a sample of census tracts in the Boston area.
- All census tracts in the greater Boston area represent the population.
- The census tracts were selected randomly from the population.

It is important to trust the data collection process: no census tract should be preferred over another in the selection.
This way we can extrapolate to other census tracts.

Population and samples

Without this uniform sampling at random things could go wrong.
For instance, if we selected US census tracts that vote republican, we could end up with mostly houses with high median value.

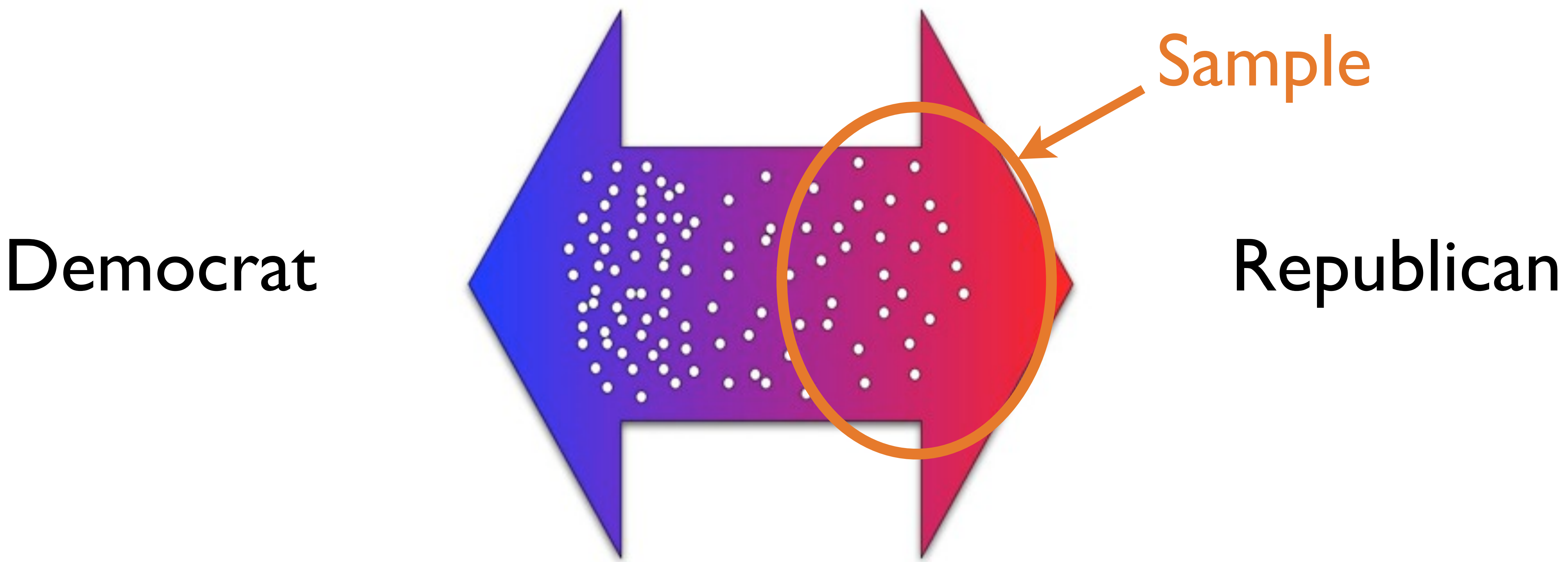
Democrat



Republican

Population and samples

Without this uniform sampling at random things could go wrong.
For instance, if we selected US census tracts that vote republican, we could end up with mostly houses with high median value.



Representative sample

Thus, the sample should be representative of the population.
In this course, we will always assume that it is but if you design your own statistical experiment our if you use data that has been collected by someone else, you should make sure that you have a sample representative of the population that you are interested in.
Note: you can always consider a smaller population. For example, the population of republican voting census tracts.

Random sapling and probability

Assuming uniform sampling at random allows us to use probability.
For example, if we know that 10% of the census tracts are on the Charles river, probability tells us how many census tracts among the 506 sampled are on the Charles river (in average): 50.6
The idea of statistics is to replace this 10% by a variable p to be estimated. In our sample, there are 35 census tracts on the Charles river so we estimate p to be 35/506=6.9%.
Probability will allow us to do much more.