

Displaying Distributions: Quantitative Variables

When a variable is quantitative, there are no natural categories.

Example: Consider the data

1.5	0.87	1.12	1.25	3.46	1.11	1.12	0.88	1.29	0.94	0.64	1.31	2.49
1.48	1.06	1.11	2.15	0.86	1.81	1.47	1.24	1.63	2.14	6.64	4.04	2.48
1.4	1.37	1.81	1.14	1.63	3.67	0.55	2.67	2.63	3.03	1.23	1.04	1.63
3.12	2.37	2.12	2.68	1.17	3.34	3.79	1.28	2.1	6.55	1.18	3.06	0.48
0.25	0.53	3.36	3.47	2.74	1.88	5.94	4.24	3.52	3.59	3.1	3.33	4.5

giving the particulate emissions for 65 vehicles in grams per gallon.

The maximum value is 6.64, while the minimum value is 0.25.

How are the values distributed?

Graphical Methods: Quantitative Variables

Graphical Methods often convey important information about the structure of the data set including:

- Overall shape of the data set (e.g., symmetric or skewed)
- Presence of gaps in the data set
- Number and location of peaks (modes) in the data set
- Presence of outliers
- Identification of a representative value and the extent of spread

Some single variable graphical methods include:

- Stem-and-Leaf Displays
- Dotplots
- Histograms

Example of Stem-and-Leaf Plot

Raw Data (Test Grades): 67, 72, 85, 75, 89, 89, 88, 99, 90, 100, 120

Stem	Leaves
6	7
7	2 5
8	5 9 9 8
9	9 0
10	0
11	
12	0

Unordered Leaves

Stem	Leaves
6	7
7	2 5
8	5 8 9 9
9	0 9
10	0
11	
12	0

Ordered Leaves

6 | 7 represents
67

Stem-and-Leaf Plot

It is sometimes helpful to round the data to a convenient number of figures before constructing the plot.

Example: Recall the particulate emissions data

1.5	0.87	1.12	1.25	3.46	1.11	1.12	0.88	1.29	0.94	0.64	1.31	2.49
1.48	1.06	1.11	2.15	0.86	1.81	1.47	1.24	1.63	2.14	6.64	4.04	2.48
1.4	1.37	1.81	1.14	1.63	3.67	0.55	2.67	2.63	3.03	1.23	1.04	1.63
3.12	2.37	2.12	2.68	1.17	3.34	3.79	1.28	2.1	6.55	1.18	3.06	0.48
0.25	0.53	3.36	3.47	2.74	1.88	5.94	4.24	3.52	3.59	3.1	3.33	4.5

This data is rounded to tenths

1.5	0.9	1.1	1.3	3.5	1.1	1.1	0.9	1.3	0.9	0.6	1.3	2.5
1.5	1.1	1.1	2.2	0.9	1.8	1.5	1.2	1.6	2.1	6.6	4	2.5
1.4	1.4	1.8	1.1	1.6	3.7	0.6	2.7	2.6	3	1.2	1	1.6
3.1	2.4	2.1	2.7	1.2	3.3	3.8	1.3	2.1	6.6	1.2	3.1	0.5
0.3	0.5	3.4	3.5	2.7	1.9	5.9	4.2	3.5	3.6	3.1	3.3	4.5

Producing the Stem-and-Leaf plot in R

```
> stem(Particulate)
```

```
The decimal point is at the |
```

```
0 | 355669999
```

```
1 | 01111112222333344555666889
```

```
2 | 11124556777
```

```
3 | 0111334555678
```

```
4 | 025
```

```
5 | 9
```

```
6 | 66
```

Key: 4|5 equals 4.5 g/gal

Stem-and-Leaf Plot



You did WHAT
with my leaves???

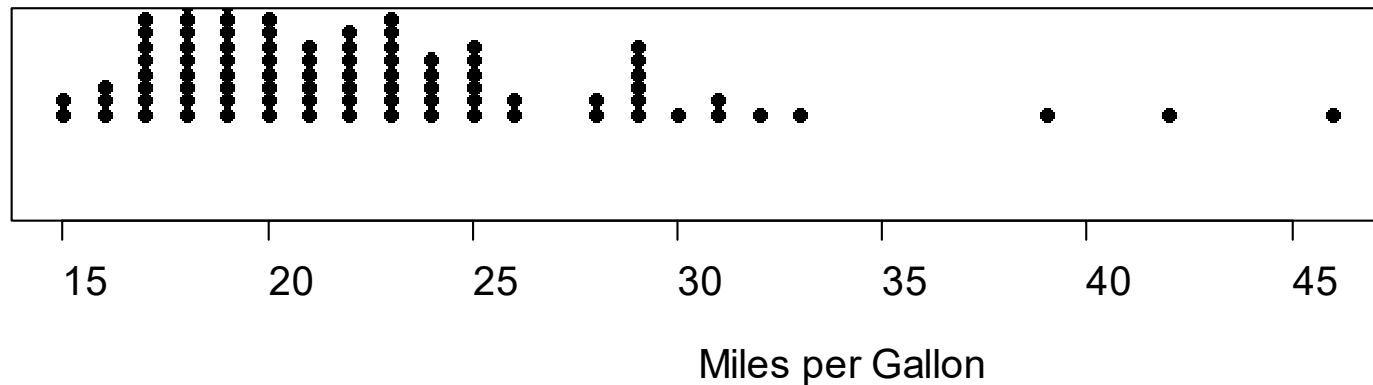
Dot Plots

- Dot Plots are another simple way of displaying a single quantitative variable.
- To construct the dot plot:
 - Draw an axis and mark off units along this axis that span the range of the data.
 - Place a dot (symbol) above the axis for each value in the data set.
 - If there are replicate values, they can be “stacked” above existing points. Alternatively, they can be “jittered” which adds some noise to their vertical position.
- Also called a one-dimensional scatter plot.
- Most useful for small and medium sized data sets.

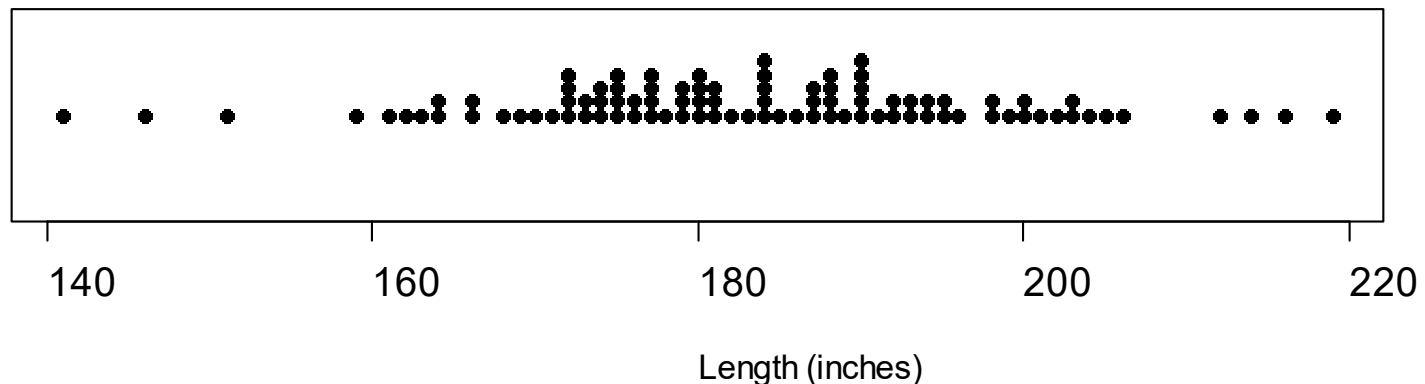
Dot Plots from the Cars93 Data Frame

A pair of Dot Plots from the Cars93 data frame using the stack method:

```
> stripchart(Cars93$MPG.city,method="stack",pch=16,  
+ cex.axis=1.2,cex.lab=1.2,xlab="Miles per Gallon")
```



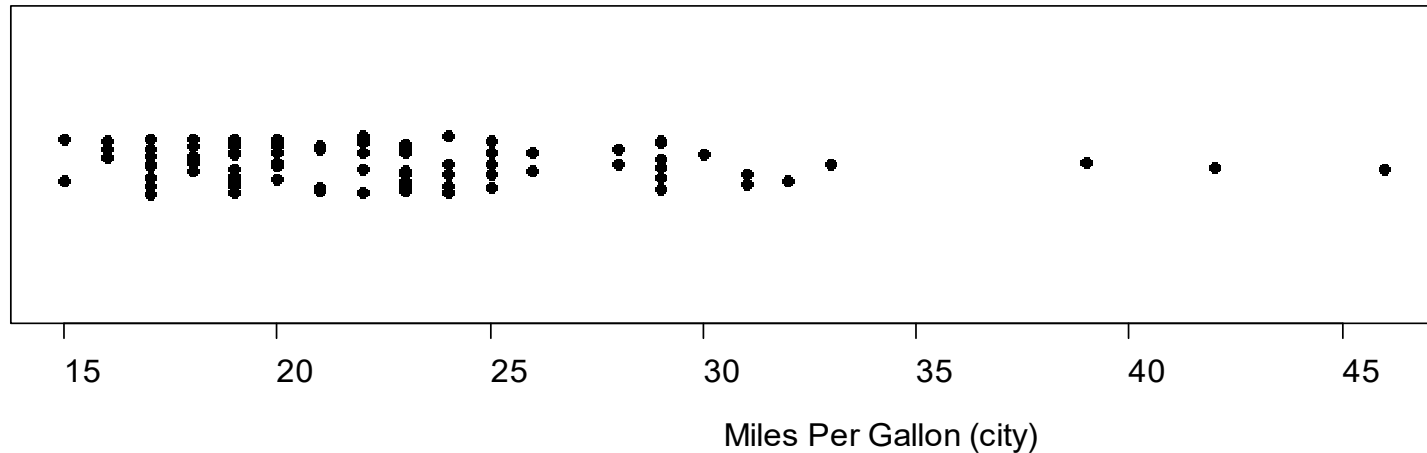
```
> stripchart(Cars93$Length,method="stack",pch=16,  
+ cex.axis=1.2,cex.lab=1.2,xlab="Length (inches)")
```



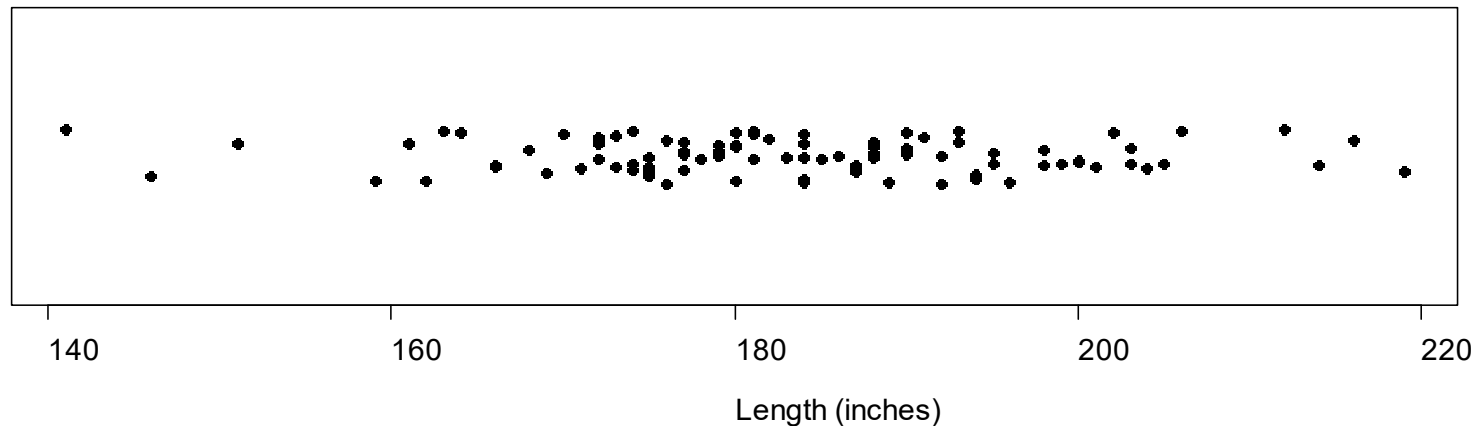
Dot Plots from the Cars93 Data Frame

A pair of Dot Plots from the Cars93 data frame using the `jitter` method:

```
> stripchart(Cars93$MPG.city,method="jitter",pch=16,  
+ cex.axis=1.2,cex.lab=1.2,xlab="Miles Per Gallon (city)")
```



```
> stripchart(Cars93$Length,method="jitter",pch=16,  
+ cex.axis=1.2,cex.lab=1.2,xlab="Length (inches)")
```



Displaying Distributions: Histogram



Displaying Distributions:

Histogram

When a variable is quantitative, there are no natural categories. One strategy is to divide the range of the data into **classes** that cover all the values that are observed.

Example: Recall the data

1.5	0.87	1.12	1.25	3.46	1.11	1.12	0.88	1.29	0.94	0.64	1.31	2.49
1.48	1.06	1.11	2.15	0.86	1.81	1.47	1.24	1.63	2.14	6.64	4.04	2.48
1.4	1.37	1.81	1.14	1.63	3.67	0.55	2.67	2.63	3.03	1.23	1.04	1.63
3.12	2.37	2.12	2.68	1.17	3.34	3.79	1.28	2.1	6.55	1.18	3.06	0.48
0.25	0.53	3.36	3.47	2.74	1.88	5.94	4.24	3.52	3.59	3.1	3.33	4.5

giving the particulate emissions for 65 vehicles in grams per gallon.

The maximum value is 6.64, while the minimum value is 0.25.

How are the values distributed?

Choosing the Classes

Requirements for Choosing Classes

- Every observation must fall into one of the classes
- The classes must not overlap
- Classes of equal width
- There must be no gaps between classes, **even if there are no observations in a class, it is included in the frequency distribution**

Relative Frequency Distribution of Particulate Data

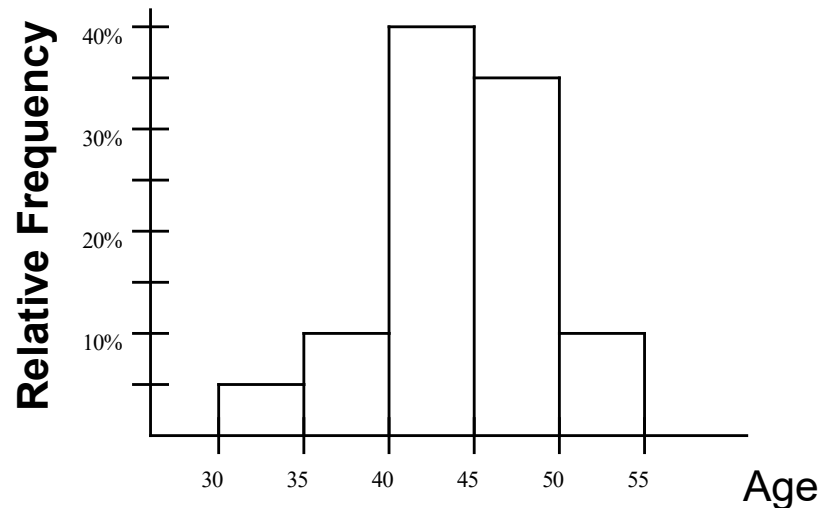
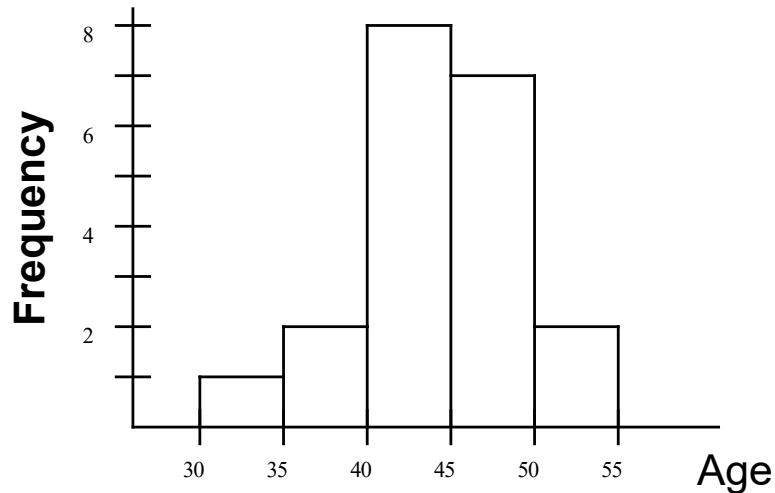
Class		Frequency	Relative Frequency
0.00-0.99	[0, 1)	9	0.138
1.00-1.99	[1, 2)	26	0.400
2.00-2.99	[2, 3)	11	0.169
3.00-3.99	[3, 4)	13	0.200
4.00-4.99	[4, 5)	3	0.046
5.00-5.99	[5, 6)	1	0.015
6.00-6.99	[6, 7)	2	0.031

Histogram of Age

The following is a histogram of the 20 ages of the attendees of an academic conference.

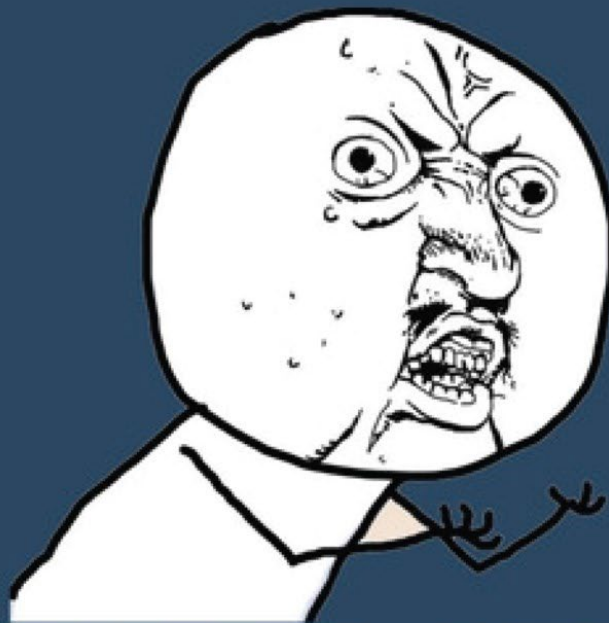
45, 41, 51, 46, 47, 42, 43, 50, 39, 32, 41, 44, 47, 49, 45, 42, 41, 40, 45, 37

Class	Tally	# Observations	Proportion/Percentage
[30, 35)	/	1	$1/20 = 0.05 \rightarrow 5\%$
[35, 40)	//	2	$2/20 = 0.10 \rightarrow 10\%$
[40, 45)	////////	8	$8/20 = 0.40 \rightarrow 40\%$
[45, 50)	////////	7	$7/20 = 0.35 \rightarrow 35\%$
[50, 55)	//	2	$2/20 = 0.10 \rightarrow 10\%$



Both the Frequency and Relative Frequency histogram reveal the same shape of the distribution.

HISTOGRAM



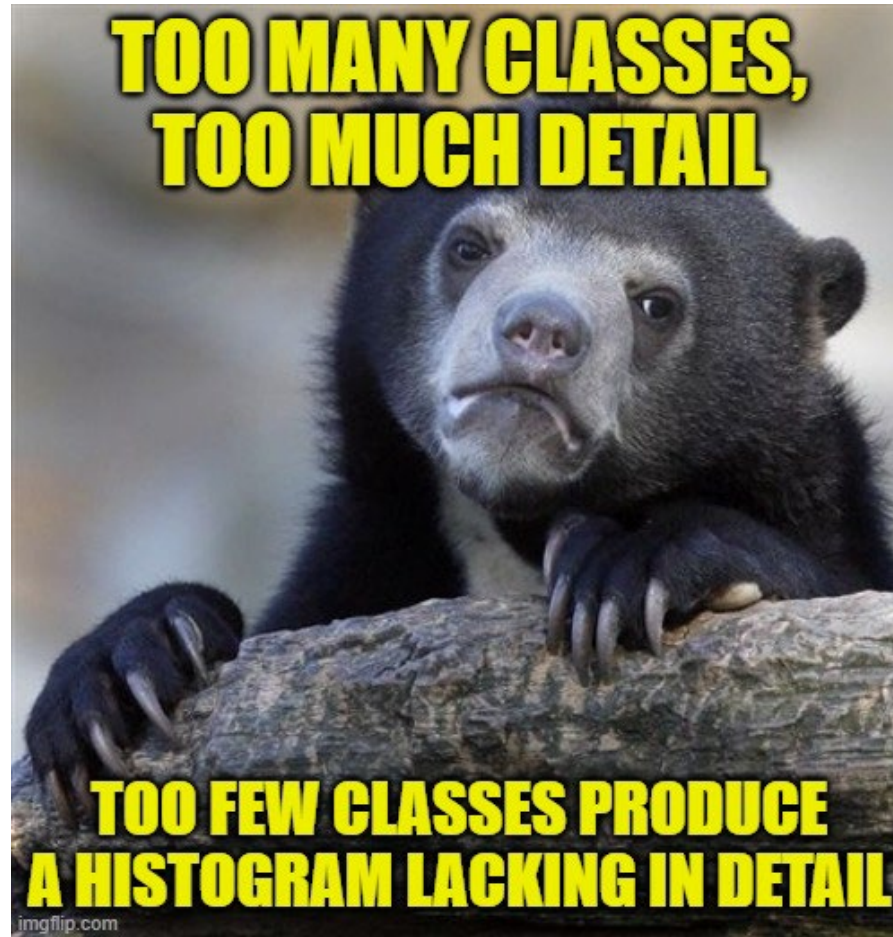
IS NO BAR CHART

A meme featuring two panels. The top panel shows Darth Vader from the chest up, standing in a control room with a wall of buttons and lights. The bottom panel shows Luke Skywalker in the same control room, looking distressed with a bloody forehead and his mouth open in a scream. The text "HISTOGRAM CONTAINING TWO VARIABLES IN BAR" is overlaid on the top panel, and "NOOOOOOOOO" is overlaid on the bottom panel.

**HISTOGRAM CONTAINING
TWO VARIABLES IN BAR**

NOOOOOOOOO

Selecting the Number of Classes?



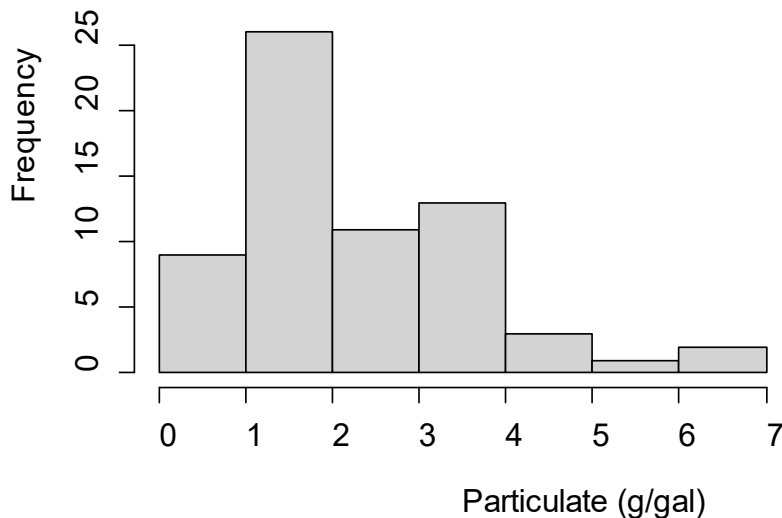
- A good rule of thumb is number of classes $\approx \sqrt{n}$

Histogram of Particulate Data

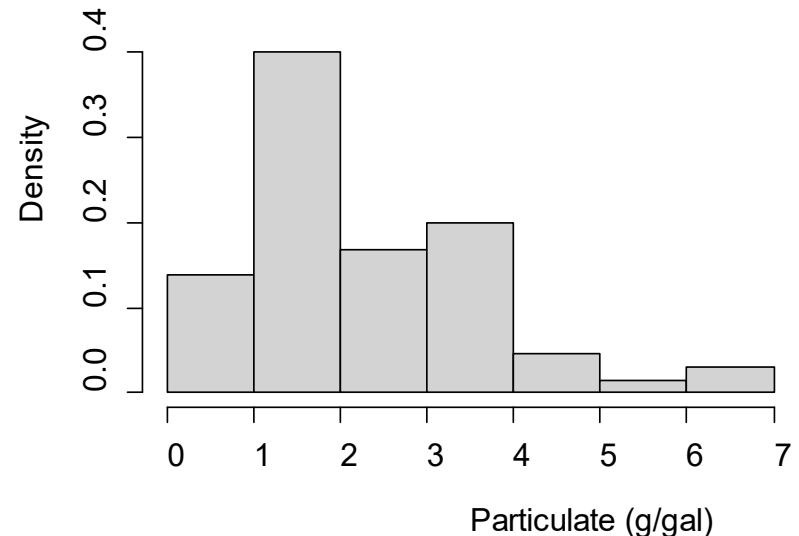
Class		Frequency	Relative Frequency
0.00-0.99	[0, 1)	9	0.138
1.00-1.99	[1, 2)	26	0.400
2.00-2.99	[2, 3)	11	0.169
3.00-3.99	[3, 4)	13	0.200
4.00-4.99	[4, 5)	3	0.046
5.00-5.99	[5, 6)	1	0.015
6.00-6.99	[6, 7)	2	0.031

```
> hist(Particulate,cex.lab=1.2,cex.axis=1.2,col="lightgray",  
+ xlab="Particulate (g/gal)")
```

Histogram of Particulate



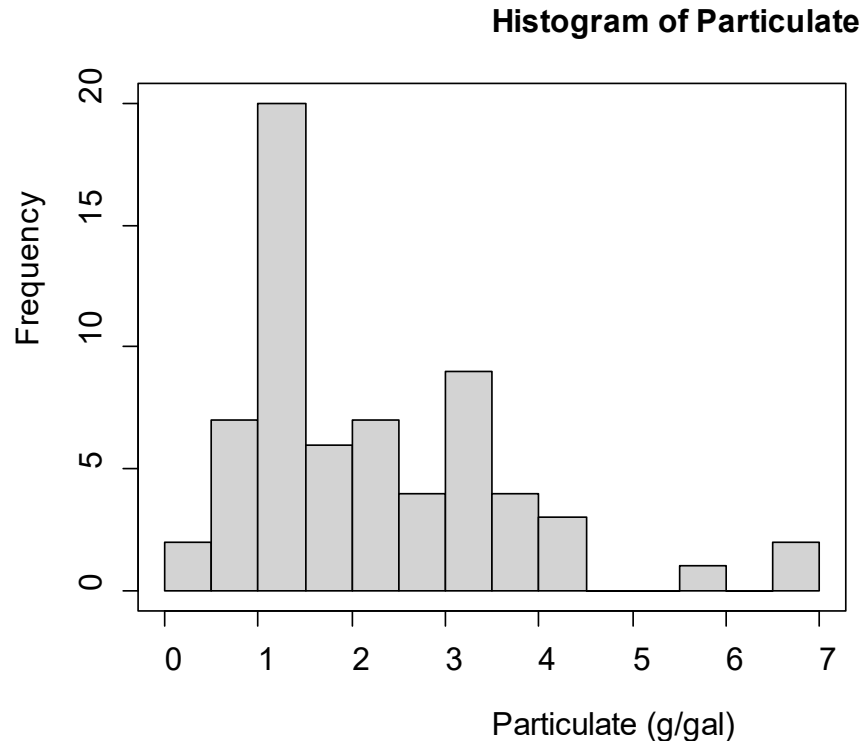
Histogram of Particulate



Histogram of Particulate Data

Setting the classes using `breaks`

```
> hist(Particulate, cex.lab=1.2, cex.axis=1.2, col="lightgray", breaks=seq(0, 7, .5),  
+ xlab="Particulate (g/gal)")  
> box()
```



Above, the command `seq(0, 7, .5)` creates the vector with the sequence

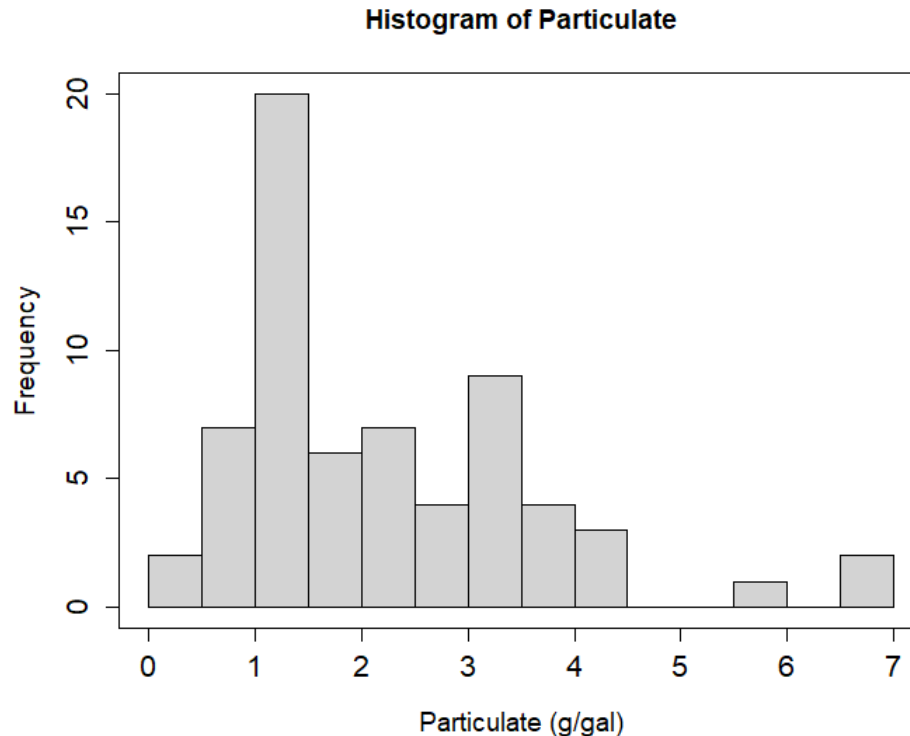
(0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7)

which is used for the endpoints of the histogram classes.

Histogram of Particulate Data

Setting the classes using `breaks`

```
> hist(Particulate,breaks=14,col="lightgray",xlab="Particulate (g/gal)",  
+ cex.lab=1.2,cex.axis=1.3)  
> box()
```



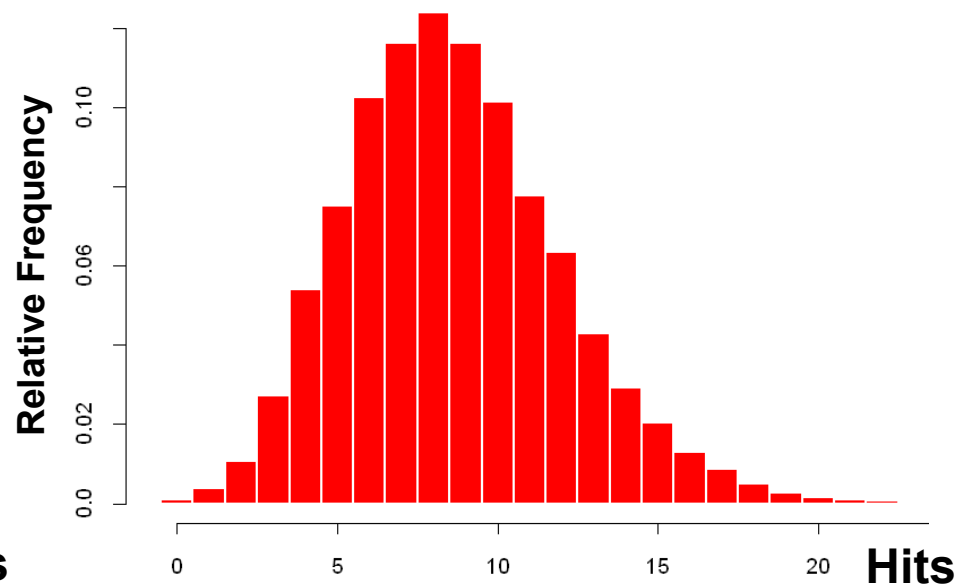
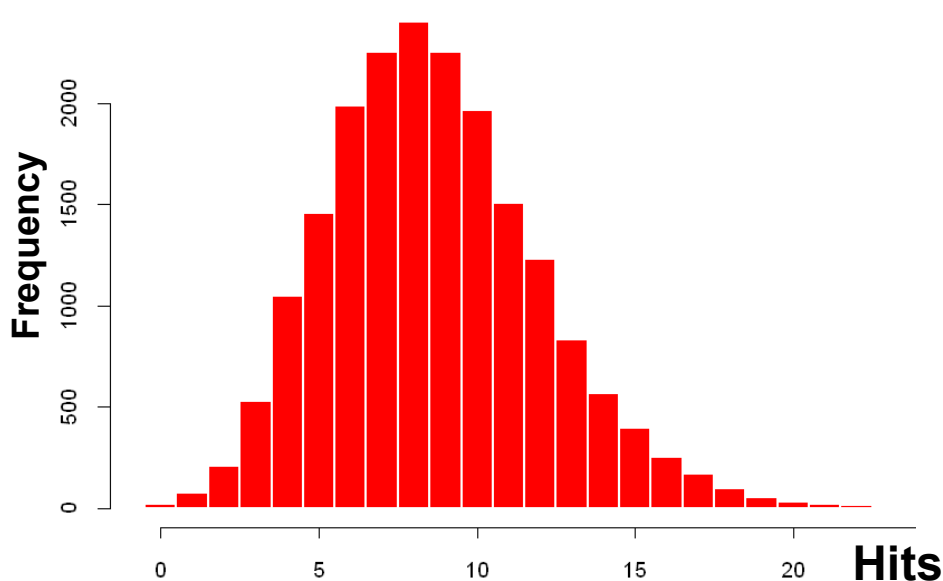
Above, the command setting `breaks` to 14 tells R we would like 14 classes. Note: when an integer is used in the `breaks` argument, the number is taken as a suggestion only, and the breakpoints will be set to “pretty” values.

Histograms for Discrete Data

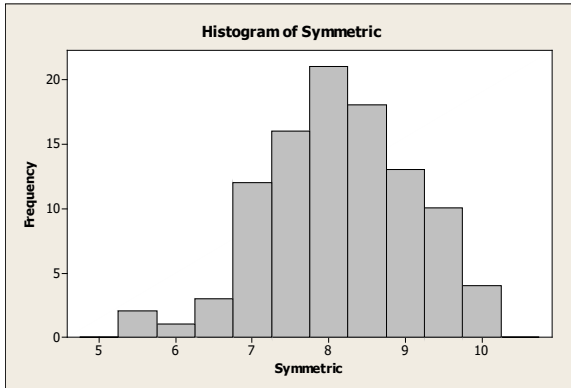
When the data are discrete, we can construct a frequency distribution in which each possible value of the variable forms a class.

- Each rectangle represents one value of the variable
- Rectangles are just wide enough to touch
 - This tells you on the spot that what you have is in fact a barplot!!
 - You can create such graphs as barplots in R.

Example: Below are the frequency and relative frequency histograms for the number of hits in a 9-inning game for 19,383 baseball games.

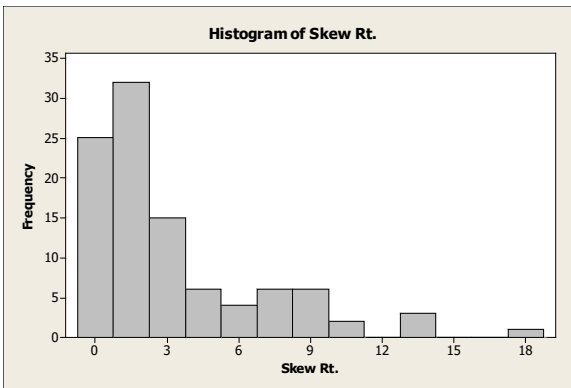


Shapes of Distributions



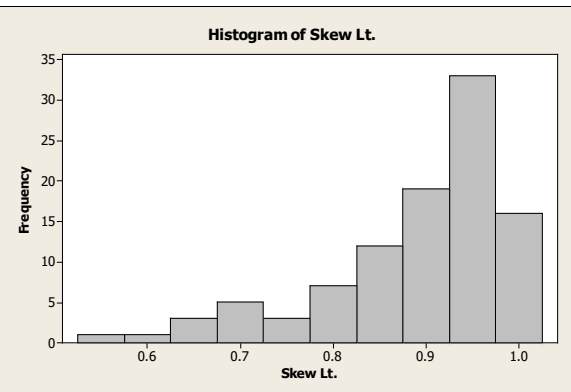
A histogram is ***symmetric*** if its right half is a mirror image of its left half.

- Rarely perfectly symmetric, many are *approximately* symmetric



A histogram is ***skewed to the right*** if it has a long right-hand tail.

- Also called *positively skewed*.



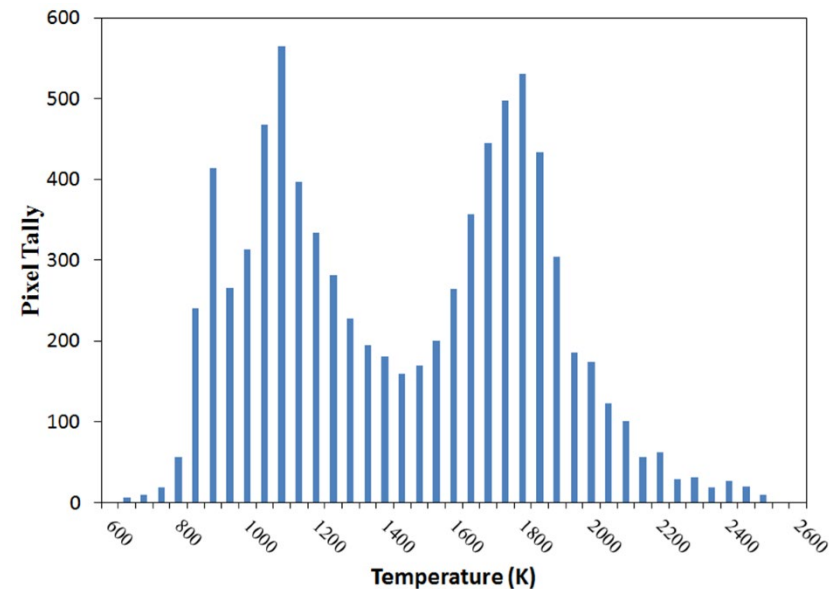
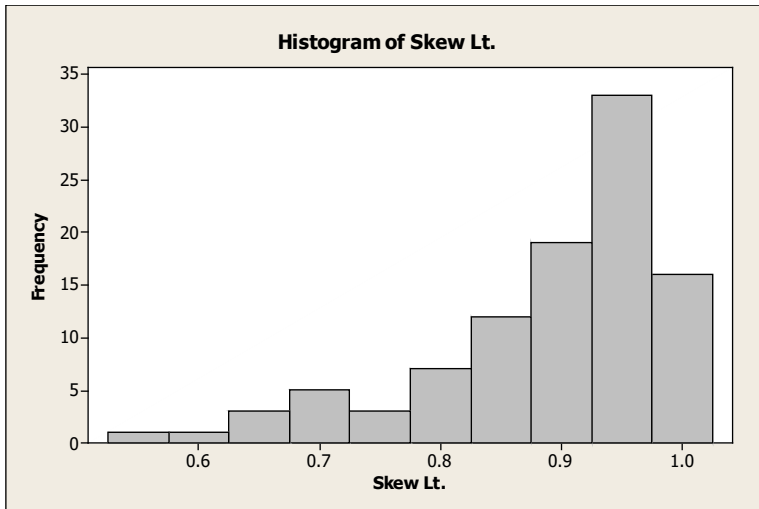
A histogram is ***skewed to the left*** if it has a long left-hand tail.

- Also called *negatively skewed*.

Modes

A peak, or high point, of a histogram is referred to as a ***mode***.

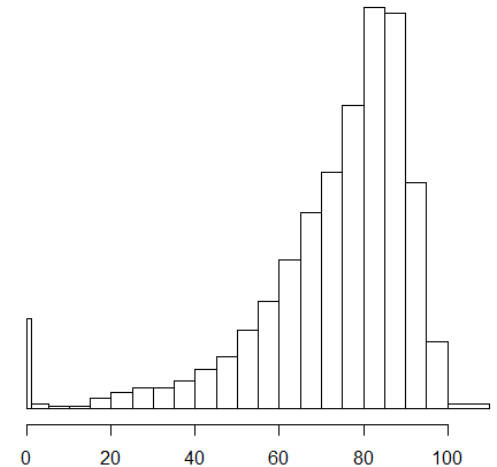
- A histogram is ***unimodal*** if it has only one mode and ***bimodal*** if it has two distinct modes.
- A histogram can have three or more modes, but this is rarely seen in practice
- Modes often represent different populations within the group



Shapes of Distributions

What shape would you expect for the following distributions?

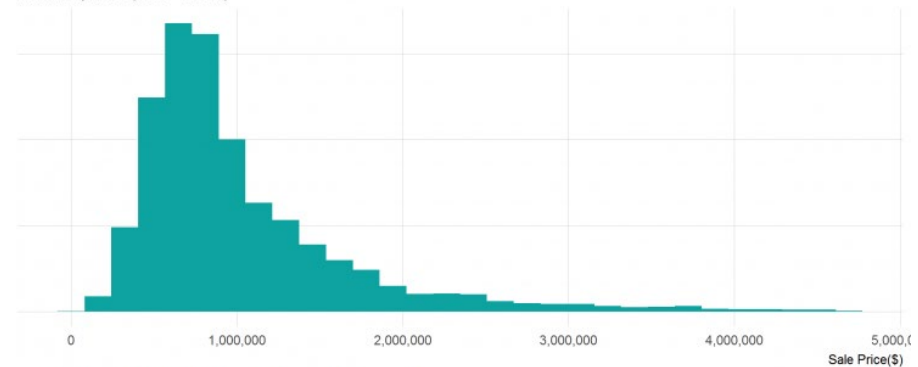
- Age at death of American citizens



- Home prices in a metropolitan area?

Distribution of San Francisco home prices

Nominal prices (2009 - 2015)



Source: San Francisco Office of the Assessor-Recorder
@KenSteif & @SimonKassel

How about numerical summaries? Measures of Center

Definition: ***Measures of center*** are numerical values that attempt to report in some sense the “middle” of a set of data.

Notation:

Σ denotes the *addition* (or *sum*) of a set of values

x is the *variable* used to represent individual data values

n represents the number of data values in a *sample*

N represents the number of data values in a *population*

A list of n numbers is denoted x_1, x_2, \dots, x_n

The sum of these numbers is $\Sigma x_i = x_1 + x_2 + \dots + x_n$

Measures of Center

Suppose we want to add up n numbers $a_1 + a_2 + \dots + a_n$.

We write this compactly as

A diagram showing the components of the summation formula $\sum_{i=1}^n a_i$. Arrows point from descriptive text to parts of the formula: 'upper limit of the summation' points to the n above the sigma; 'terms' points to the a_i ; 'lower limit of the summation' points to the $i=1$ below the sigma; 'index of summation' points to the i in the denominator; and 'means sum' points to the sigma symbol.

$$\text{means sum} \rightarrow \sum_{i=1}^n a_i \leftarrow \text{terms}$$

upper limit of the summation

lower limit of the summation

index of summation

Find

$$\sum_{i=3}^7 i = 3 + 4 + 5 + 6 + 7 = 25$$

Find

$$\sum_{k=1}^3 (2^k - k) = (2^1 - 1) + (2^2 - 2) + (2^3 - 3) = 1 + 2 + 5 = 8$$

When you need to find the
center of a dataset



stats

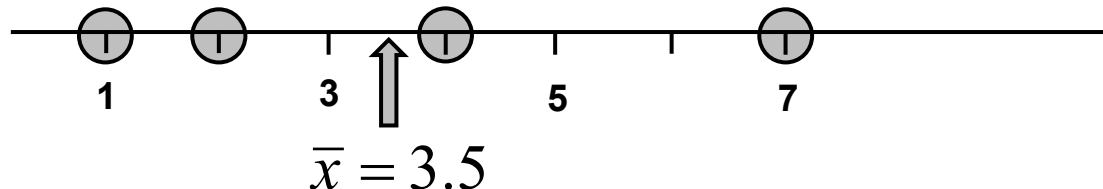
The Mean

Definition: The **mean** is the ordinary arithmetic average, found by adding up the values and dividing by the number of observations.

- The mean of a sample is denoted by \bar{x} (pronounced “x-bar”)
- The population mean is denoted by the Greek letter μ . (“mu”)

- Example: 4 7 2 1

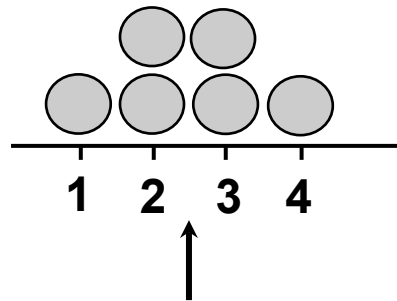
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{4 + 7 + 2 + 1}{4} = 3.5$$



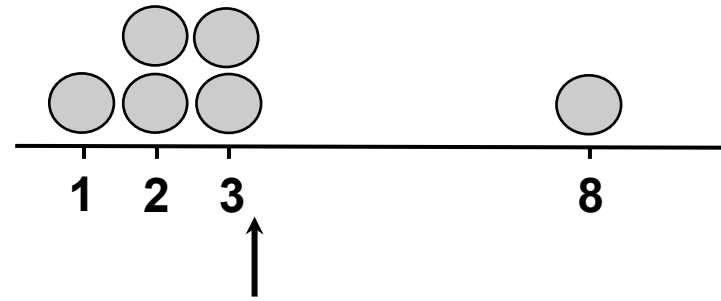
Is the Mean Always the Center?

Can you think of a situation where the mean is not as “informative”?

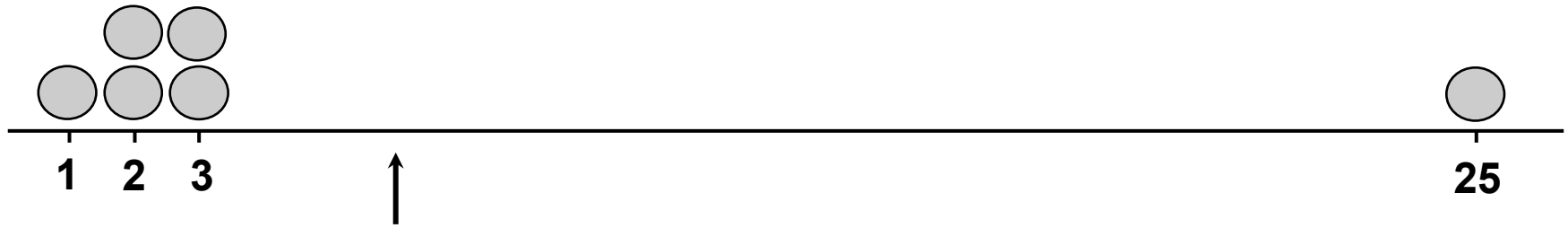
Is the Mean Always the Center?



Mean = 2.5



Mean = 3.17



Mean = 6

- The mean shifts towards an extreme observation.
- If a distribution appears skewed, we may wish to also report a more *resistant* measure of center.

Note: The mean is not necessarily a “typical” value for the data set.

The Median

Definition: The ***median*** is the “middle” observation (once the values are arranged in order)

To calculate the median M , sort all observations from smallest to largest:

1. If n is odd, M is the middle number in the list
 2. If n is even, M is the mean of the middle two numbers in the list
- The symbol \tilde{x} is also commonly used for the median.
 - The median is the 50th ***percentile***.

Median Example

Odd number: 6.72 3.46 3.60 6.44 26.70

First arrange the values in order, then pick the middle value

3.46 3.60 6.44 6.72 26.70

↑
M

Even number: 6.72 3.46 3.60 6.44

First arrange the values in order, then compute the mean of the two middle values

3.46 3.60 6.44 6.72

$$M = \frac{3.60 + 6.44}{2} = 5.02$$

Median is Resistant

Definition: A statistic is ***resistant*** if its value is not affected much by extreme values (in either direction) in the data set.

Recall the median of the following data set:

3.46 3.60 6.44 6.72 26.70

 ↑

M

Suppose the last number was incorrectly recorded as 2670.
Would the median change?

The median would stay the same.

The median is resistant, but the mean is not.
--

The Mode

Definition: The ***mode*** of a data set is the value that occurs most frequently.

- When two or more values occur with the same frequency, each one is a mode.
- If no value appears more than once, we say the data set has no mode

Examples:

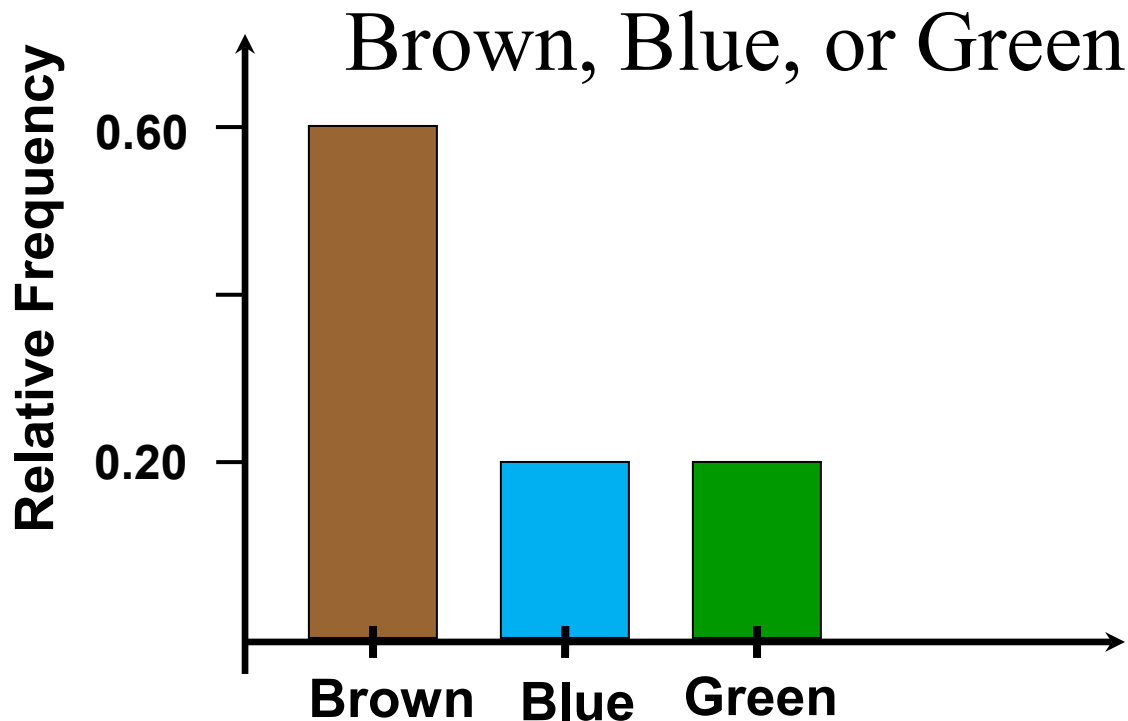
{ 0, 0, 0, 0, 1, 1, 2, 2, 3, 4 }

{ 0, 0, 0, 1, 1, 2, 2, 2, 3, 4 }

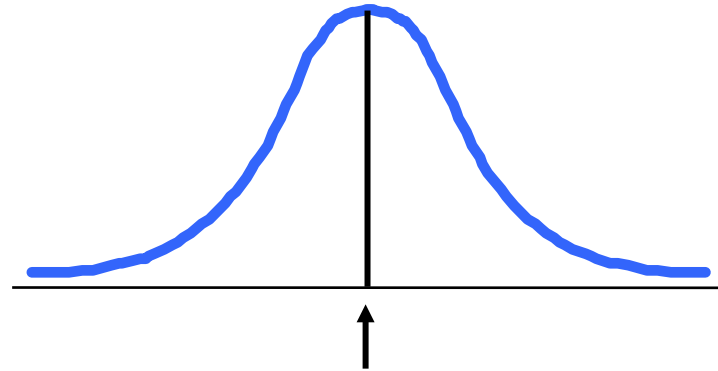
Mode Continued

The mode can be computed for qualitative data.

Example: Suppose that eye color was categorized as follows:

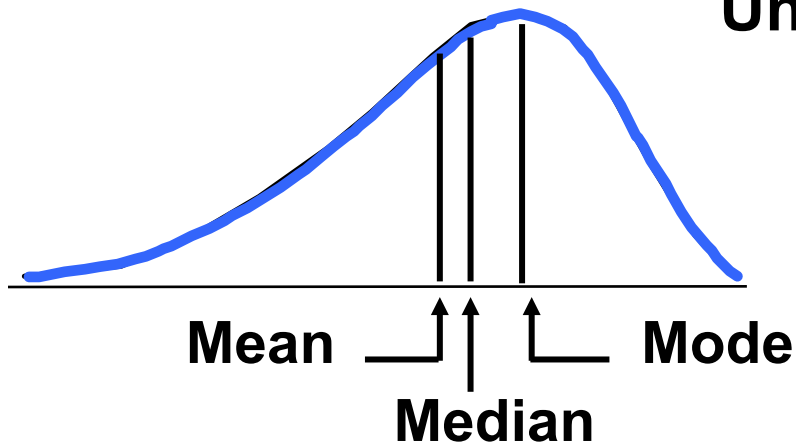


Relationships of Measures of Center

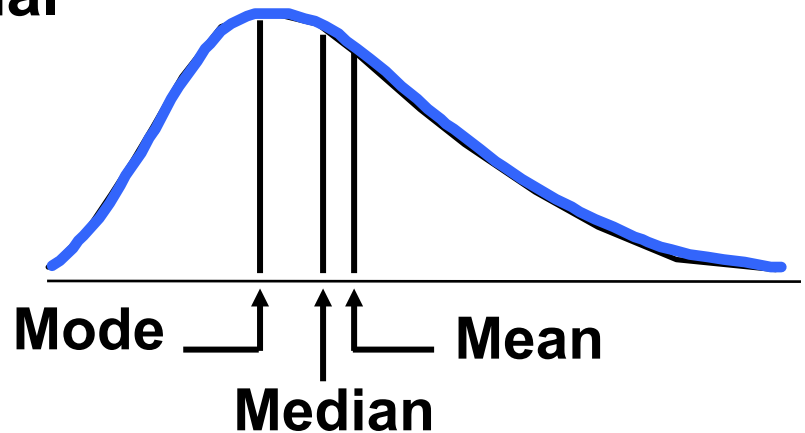


Mode = Mean = Median

**Symmetric &
Unimodal**



Left Skewed



Right Skewed

Examples from the Cars93 Data Frame

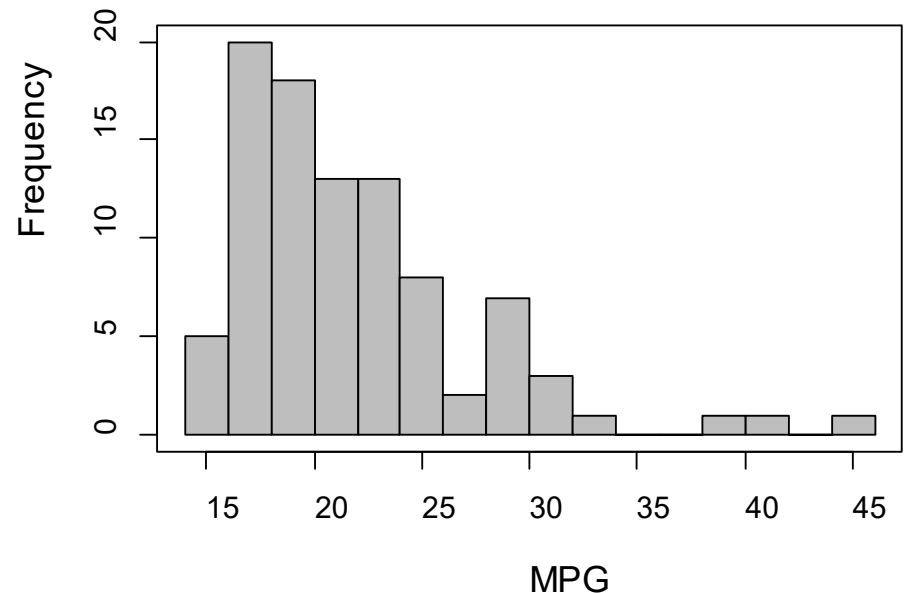
	Manufacturer	Model	Type	Min.Price	Price	Max.Price	MPG.city	MPG.highway	AirBags
1	Acura	Integra	Small	12.9	15.9	18.8	25	31	None
2	Acura	Legend	Midsize	29.2	33.9	38.7	18	25	Driver & Passenger
3	Audi	90	Compact	25.9	29.1	32.3	20	26	Driver only
4	Audi	100	Midsize	30.8	37.7	44.6	19	26	Driver & Passenger
5	BMW	535i	Midsize	23.7	30.0	36.2	22	30	Driver only
6	Buick	Century	Midsize	14.2	15.7	17.3	22	31	Driver only

Get the mean and median of MPG.city

```
> mean()  
> median()
```

Mode of # of cylinders? # of airbags?

```
> table()
```



What if you want to visualize more than one variable, aka relationships?

Given *paired data*, we may wish to determine if there is a relationship between the two variables and, if so, identify what the relationship is.

Does blood pressure predict life expectancy?

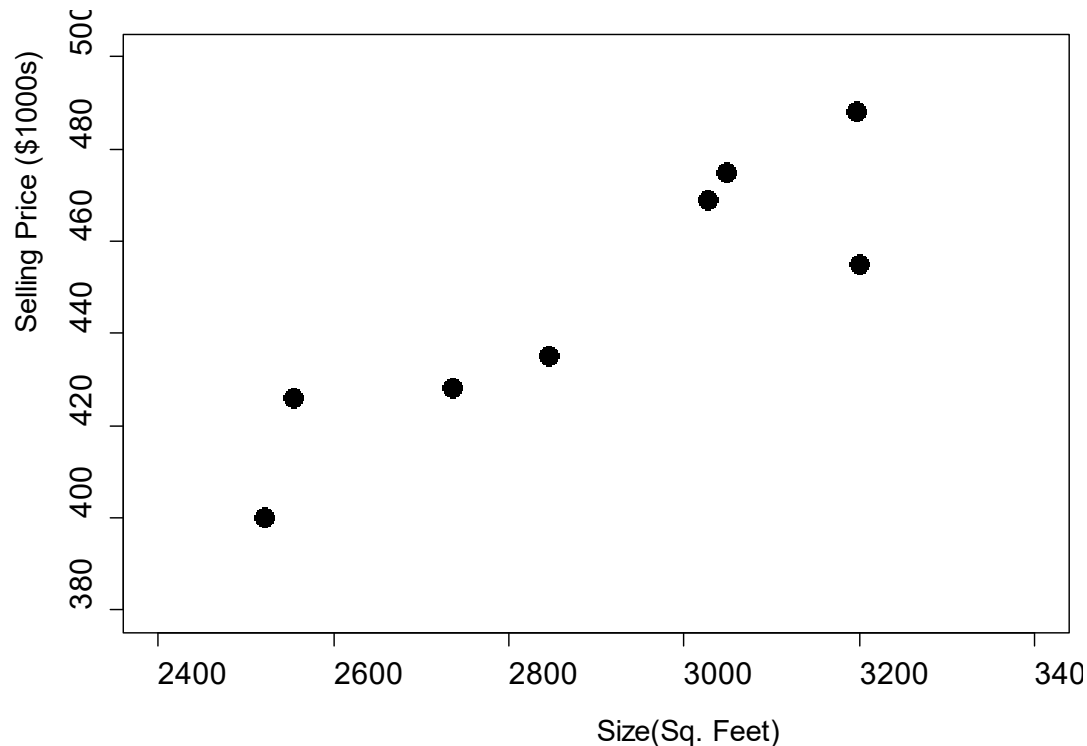
Do SAT scores predict college performance?

If such a relationship exists, perhaps we can find an equation describing it, then we could use the equation to make predictions, aka regression. (more on this later in the course!)

Visualizing Bivariate Data with a Scatterplot

Size (Sq. ft.)	2521	2555	2735	2846	3028	3049	3198	3198
Selling Price (\$1000s)	400	426	428	435	469	475	488	455

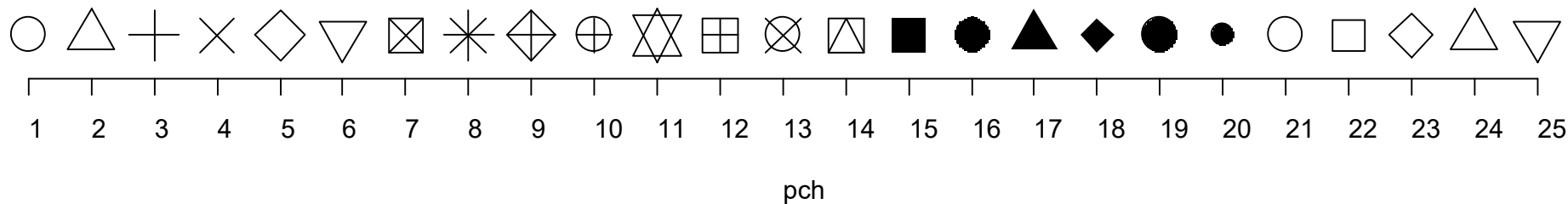
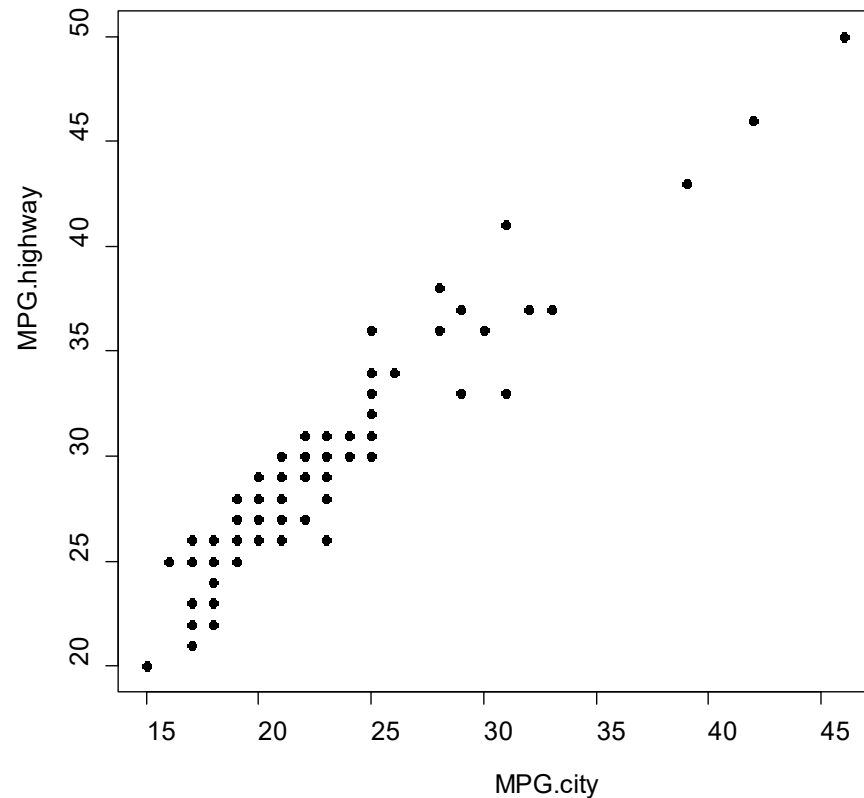
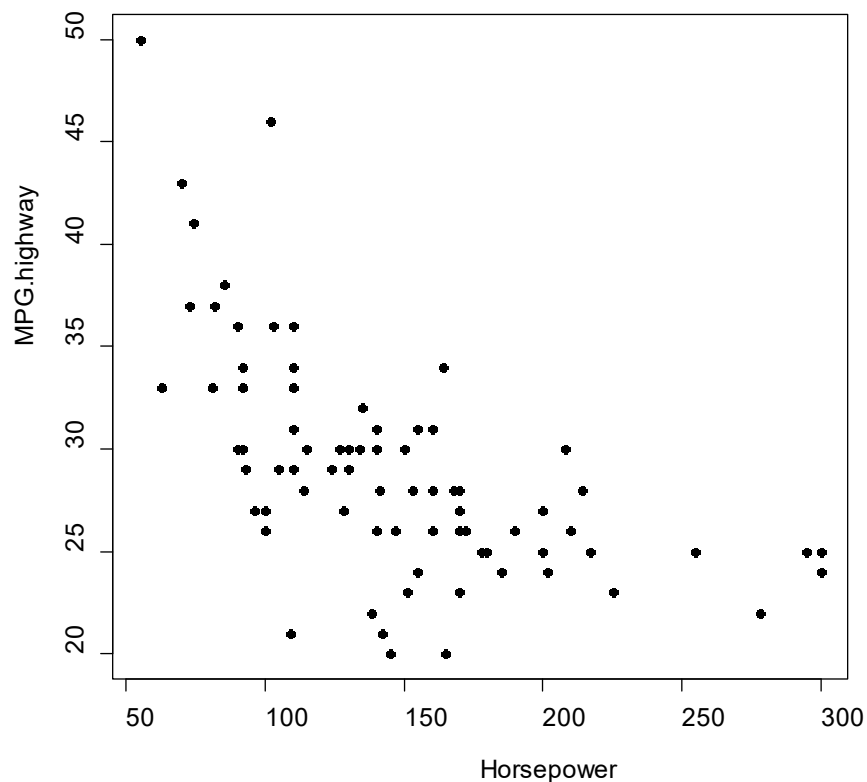
The table gives the size in square feet and the selling price in 1000s of dollars, for a sample of houses in a suburban Denver neighborhood. Here, each house is a unit and contributes an ordered pair of numbers: (Size, Selling Price) **NOTE: Correlation does not mean causation!!**



Examples from the Cars93 Data Frame

```
> plot(MPG.highway~Horsepower,data=Cars93,pch=16,cex.lab=1.2,cex.axis=1.2)
```

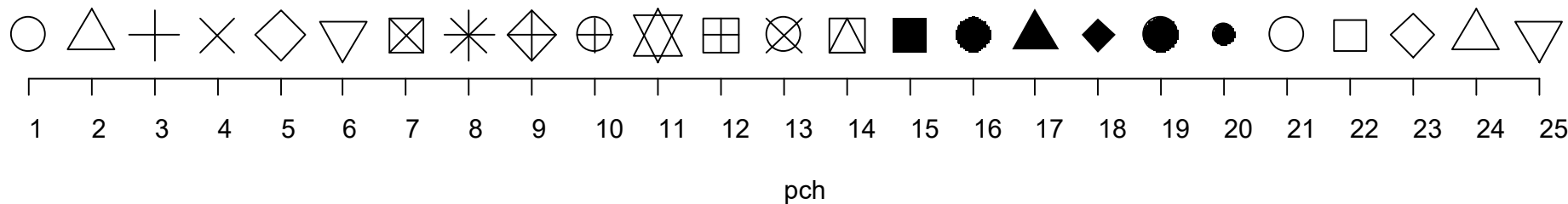
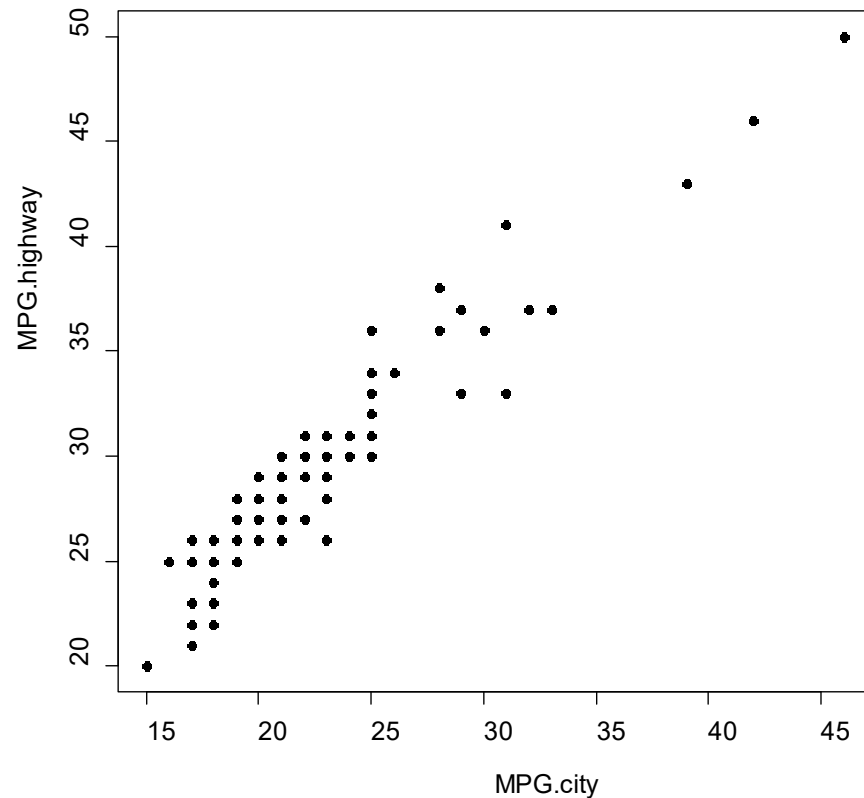
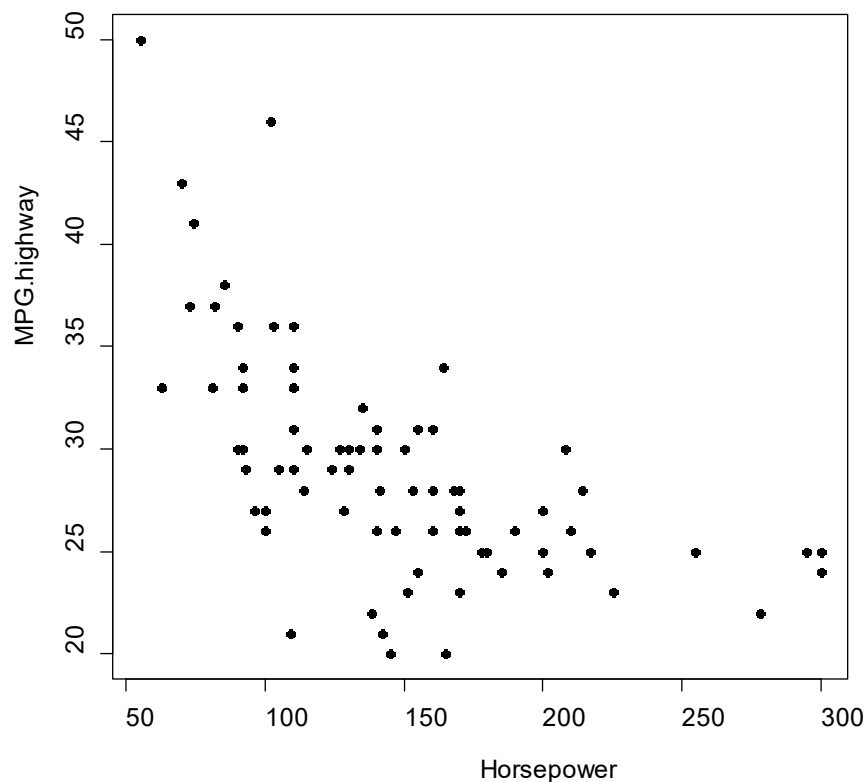
```
> plot(MPG.highway~MPG.city,data=Cars93,pch=16,cex.lab=1.2,cex.axis=1.2)
```



Examples from the Cars93 Data Frame

Or,

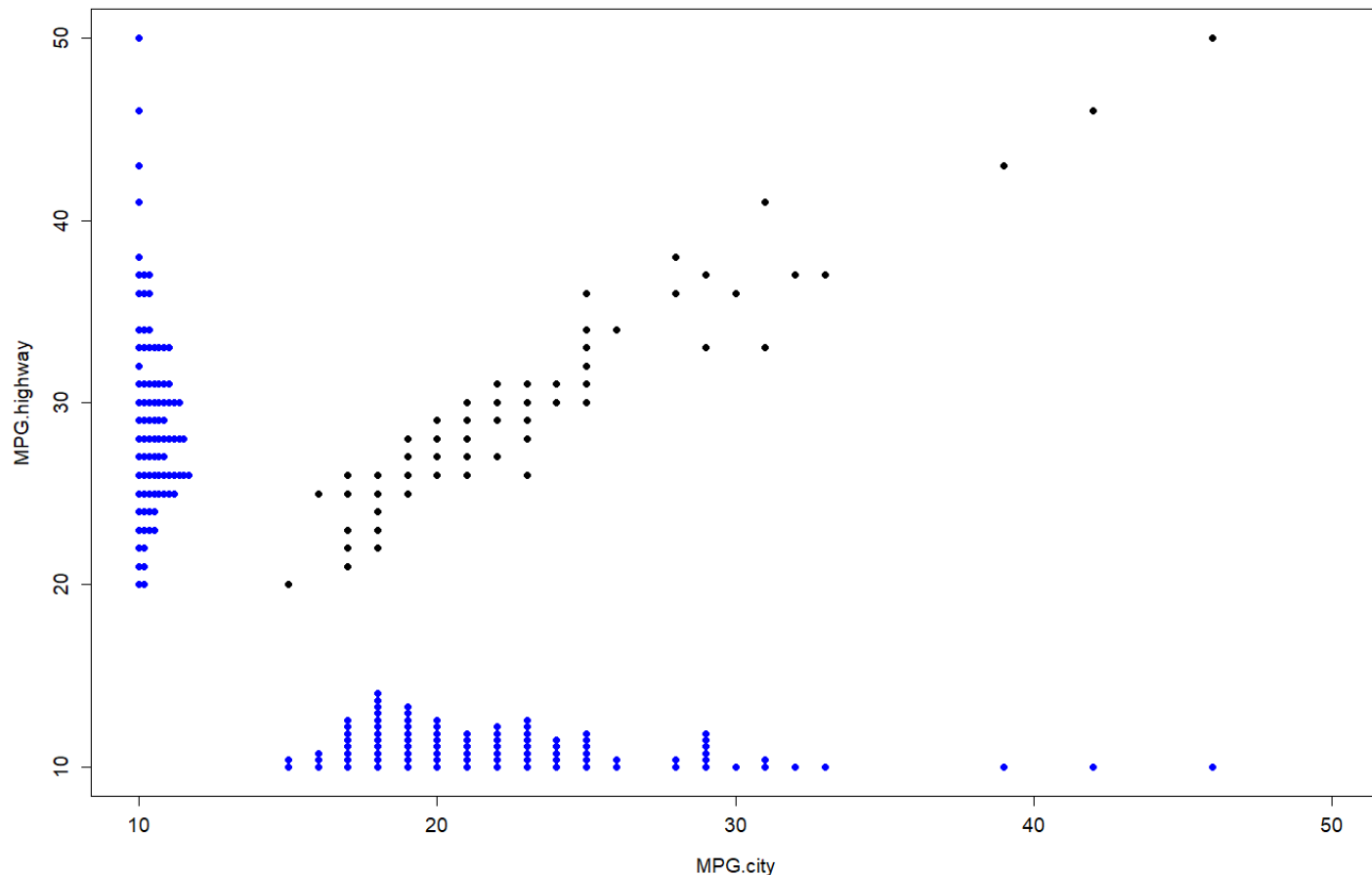
```
> plot(Cars93$Horsepower, Cars93$MPG.highway, pch=16, cex.lab=1.2, cex.axis=1.2)  
> plot(Cars93$MPG.city, Cars93$MPG.highway, pch=16, cex.lab=1.2, cex.axis=1.2)
```



Adding a dotplot to a Scatterplot

You can also add a dotplot to a scatterplot using the `stripchart` function with the `add=T` argument as follows:

```
> plot(MPG.highway~MPG.city,data=Cars93,pch=16,cex.lab=1.2,cex.axis=1.2,  
+ ylim=c(10,50),xlim=c(10,50))  
> stripchart(Cars93$MPG.city,add=T,method="stack",at=10,col="blue",pch=16)  
> stripchart(Cars93$MPG.highway,add=T,vertical=T,method="stack",at=10,col="blue",pch=16)
```



Layout option

We can use the `layout` command to place boxplots or histograms in the margins of a scatterplot to provide the distribution of the variables in the axes.

Usage

```
layout(mat, widths = rep.int(1, ncol(mat)), heights = rep.int(1, nrow(mat)),  
respect = FALSE)
```

Numbers on the right follow the order which you run your graph lines:

```
> matrix(c(1,2,3,4),nrow=2,byrow=T)  
      [,1] [,2]  
[1,]    1    2  
[2,]    3    4
```

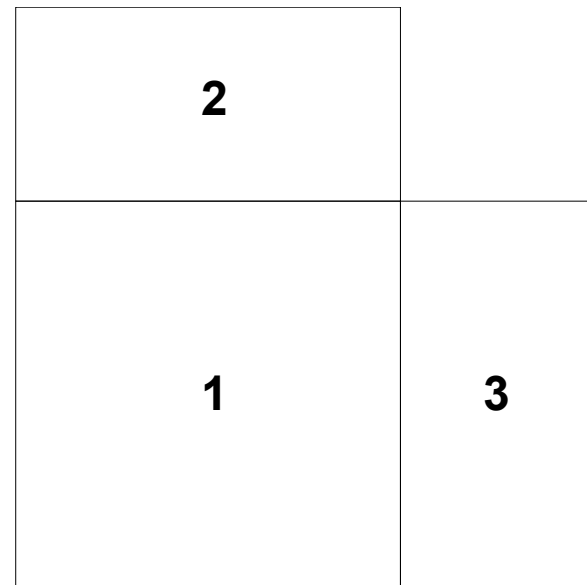
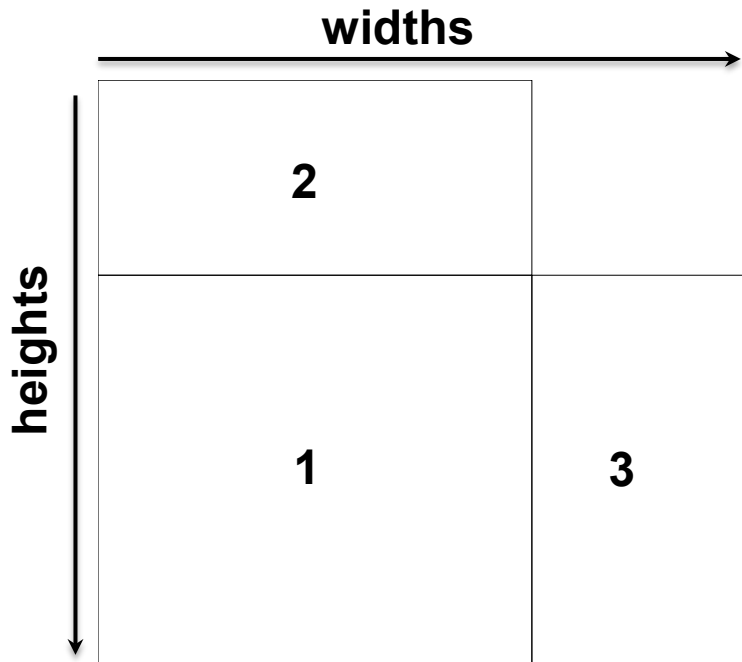
1	2
3	4

Using the `layout` Function

You can control the widths of the columns and the heights of the rows using the `widths` and `heights` optional arguments. (note: 0 gives no plot)

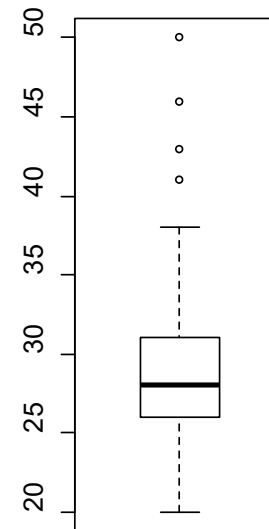
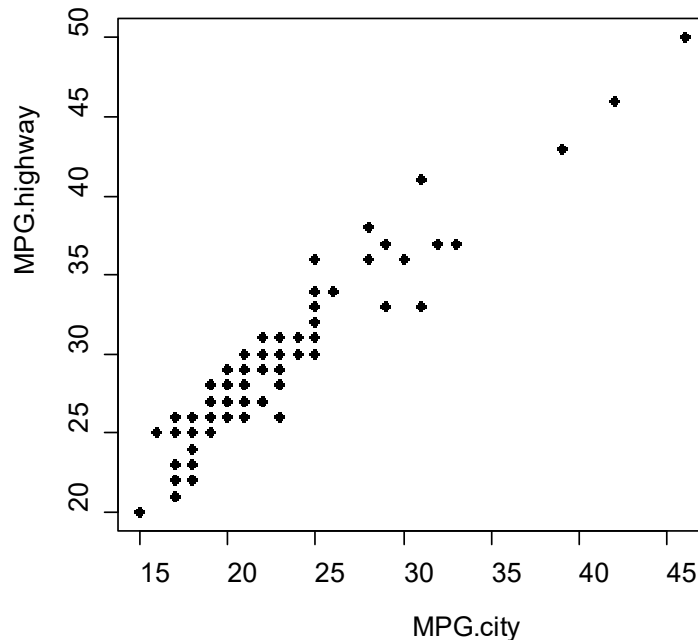
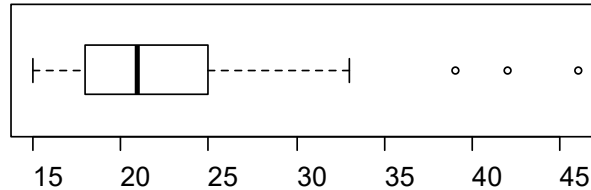
The function `layout.show(n)` plots (part of) the current layout, namely the outlines of the next `n` figures.

```
> layout(matrix(c(2,0,1,3),nrow=2,byrow=T),widths=c(2,1),heights=c(1,2))  
> layout.show(n=3)  
  
> layout(matrix(c(2,2,0,1,1,3,1,1,3),nrow=3,byrow=T))  
> layout.show(n=3)
```

$$\begin{bmatrix} 2 & 0 \\ 1 & 3 \end{bmatrix}$$
$$\begin{bmatrix} 2 & 2 & 0 \\ 1 & 1 & 3 \\ 1 & 1 & 3 \end{bmatrix}$$


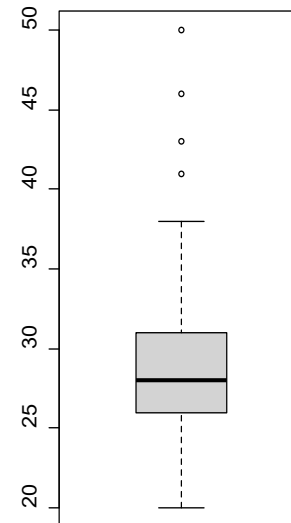
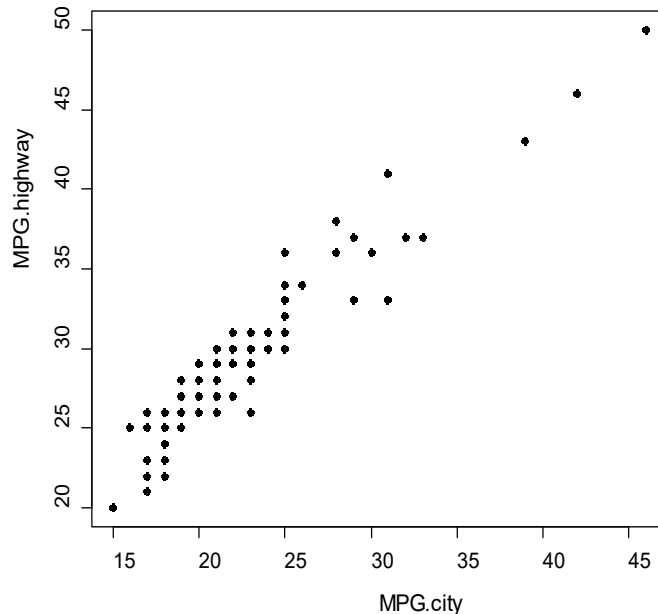
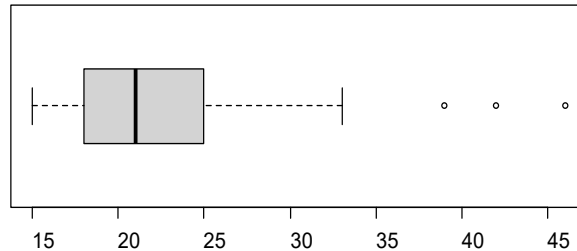
Including Boxplots with a Scatterplot

```
> layout(matrix(c(2,0,1,3),nrow=2,byrow=T),widths=c(2,1),heights=c(1,2))  
> plot(MPG.highway~MPG.city,data=Cars93,pch=16,cex.lab=1.2,cex.axis=1.2)  
> boxplot(Cars93$MPG.city,horizontal=T,cex.axis=1.2)  
> boxplot(Cars93$MPG.highway,cex.axis=1.2)
```



Including Boxplots with a Scatterplot

```
> par()$mar  
[1] 5.1 4.1 4.1 2.1  
> layout(matrix(c(2,0,1,3),nrow=2,byrow=T),widths=c(2,1),heights=c(1,2))  
> par(mar=c(5.1,4.1,0.1,0.1))  
> plot(MPG.highway~MPG.city,data=Cars93,pch=16,cex.lab=1.3,cex.axis=1.2)  
> boxplot(Cars93$MPG.city,horizontal=T,cex.axis=1.2,col="lightgray")  
> boxplot(Cars93$MPG.highway,cex.axis=1.2,col="lightgray")
```



Adding Text to a Plot Using the `text` Function

Description

`text` draws the strings given in the vector `labels` at the coordinates given by `x` and `y`. `y` may be missing since `xy.coords(x, y)` is used for construction of the coordinates.

Usage

```
text(x, y = NULL, labels = seq_along(x$x), adj = NULL, pos = NULL, offset = 0.5, vfont = NULL, cex = 1, col = NULL, font = NULL, ...)
```

Arguments

- | | |
|----------------------------|---|
| <code>x,y</code> | numeric vectors of coordinates where the text labels should be written. If the length of <code>x</code> and <code>y</code> differs, the shorter one is recycled. |
| <code>labels</code> | a character vector or expression specifying the <i>text</i> to be written. An attempt is made to coerce other language objects (names and calls) to expressions, and vectors and other classed objects to character vectors by <u><code>as.character</code></u>. If <code>labels</code> is longer than <code>x</code> and <code>y</code>, the coordinates are recycled to the length of <code>labels</code>. |
| <code>pos</code> | a position specifier for the text. If specified this overrides any <code>adj</code> value given. Values of 1, 2, 3 and 4, respectively indicate positions below, to the left of, above and to the right of the specified (x,y) coordinate |

Example Using the `text` Function

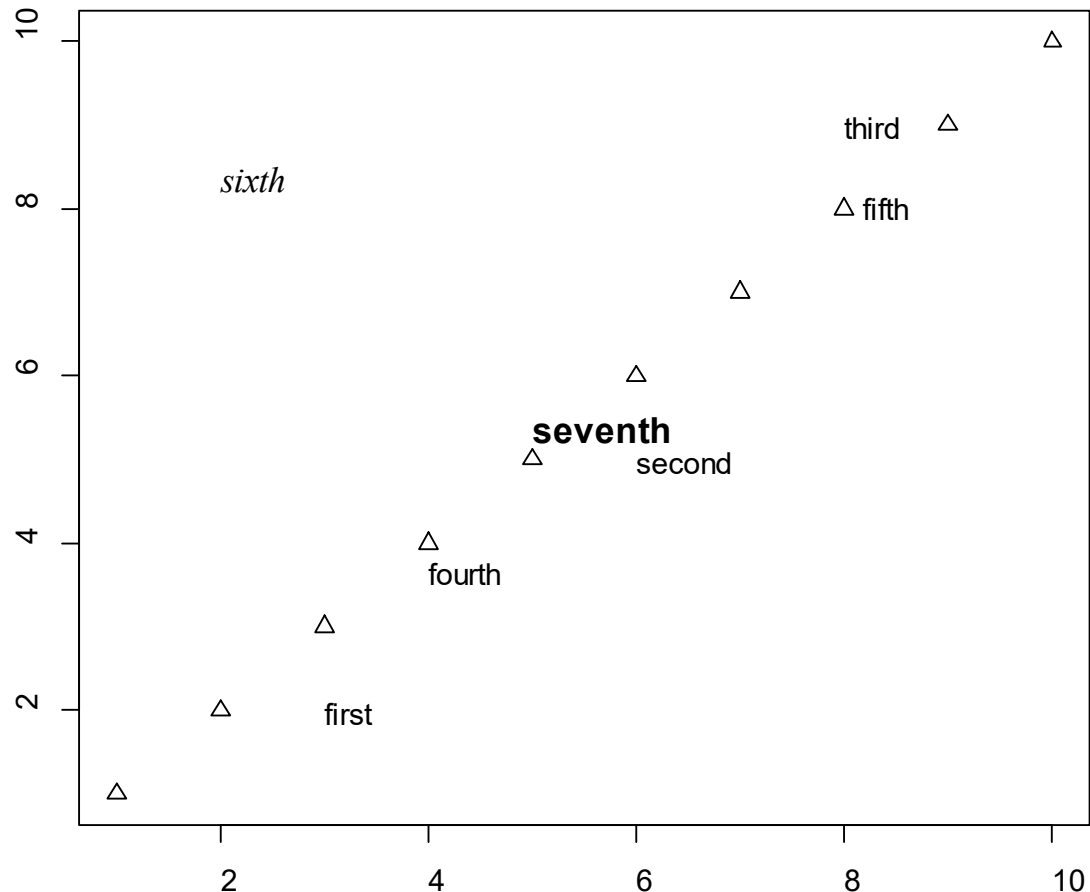
```
> plot(1:10,1:10,xlab="",ylab="",pch=2)
> text(c(3,6,8),c(2,5,9),c("first","second","third"))
> text(c(4,8),c(4,8),c("fourth","fifth"),pos=c(1,4))
> text(c(2),c(8),c("sixth"),pos=c(3),cex=1.2,font=3,family="serif")
> text(c(5),c(5),c("seventh"),pos=c(3),cex=1.2,font=2)
```

family: font family for drawing text. Standard values are "serif", "sans", "mono", "symbol". You can add your own.

font: Integer specifying font to use for text.

1=plain, 2=bold, 3=italic, 4=bold italic, 5=symbol

Serif = "Times New Roman"



ggplot?

- R built-in functions for plots
- ggplot

