Displaying Distributions:
Quantitative Variables

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

<u>Example</u>: 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?

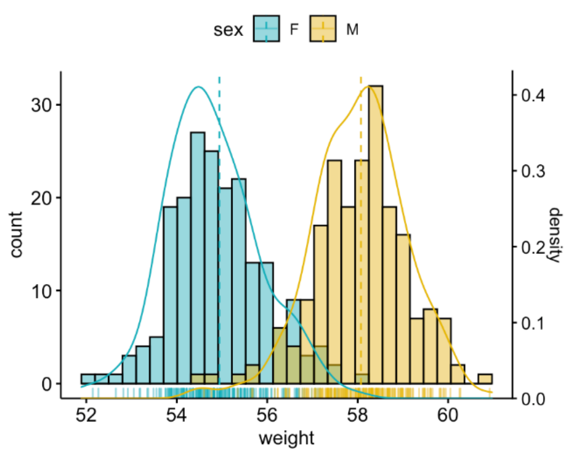
Graphical Methods: Quantitative Variables

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

One of the most widely used graphic to visualize numerical data:

Histogram



Displaying Distributions:
Histogram

Displaying Distributions:
Histogram

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

<u>Example</u>: 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**

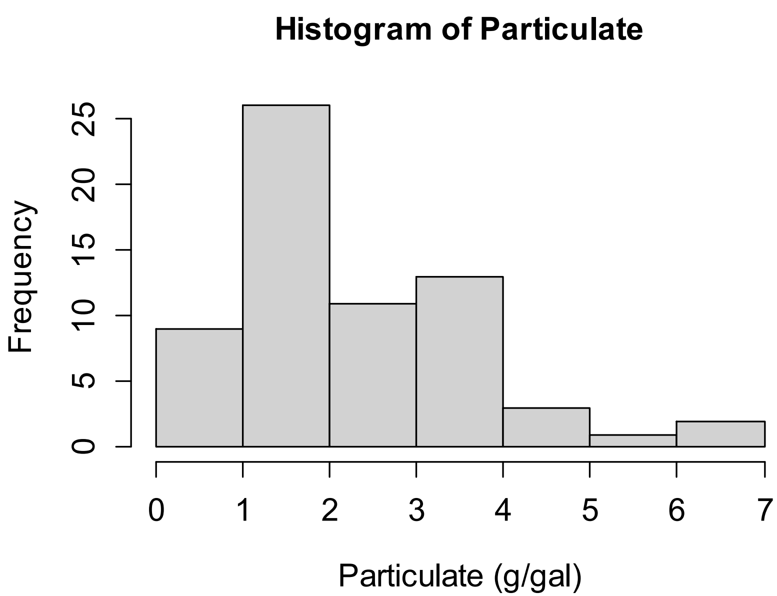
Relative Frequency Distribution of Particulate Data

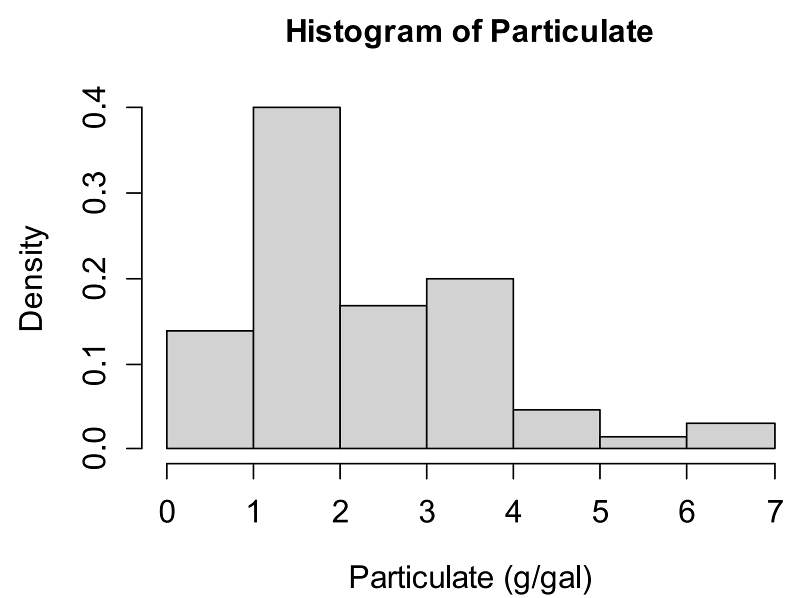
Class		**Frequency**	**Relative Frequency**
0.00-0.99	[0, 1)	9	0.138
1.00-1.99	[1, 2)	26	0.400
2.00-2.99	[2, 3)	11	0.169
3.00-3.99	[3, 4)	13	0.200
4.00-4.99	[4, 5)	3	0.046
5.00-5.99	[5, 6)	1	0.015
6.00-6.99	[6, 7)	2	0.031

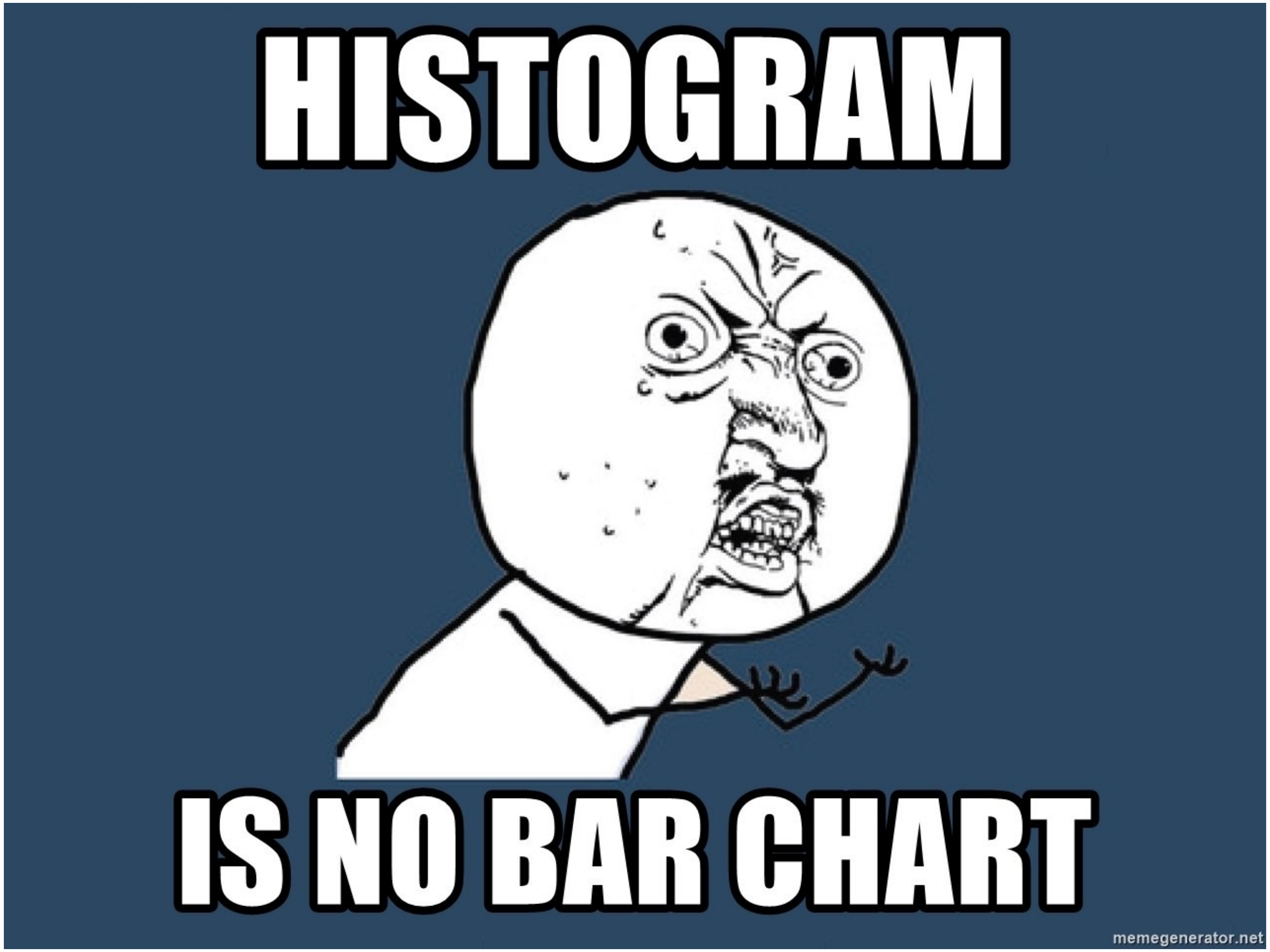
Histogram of Particulate Data

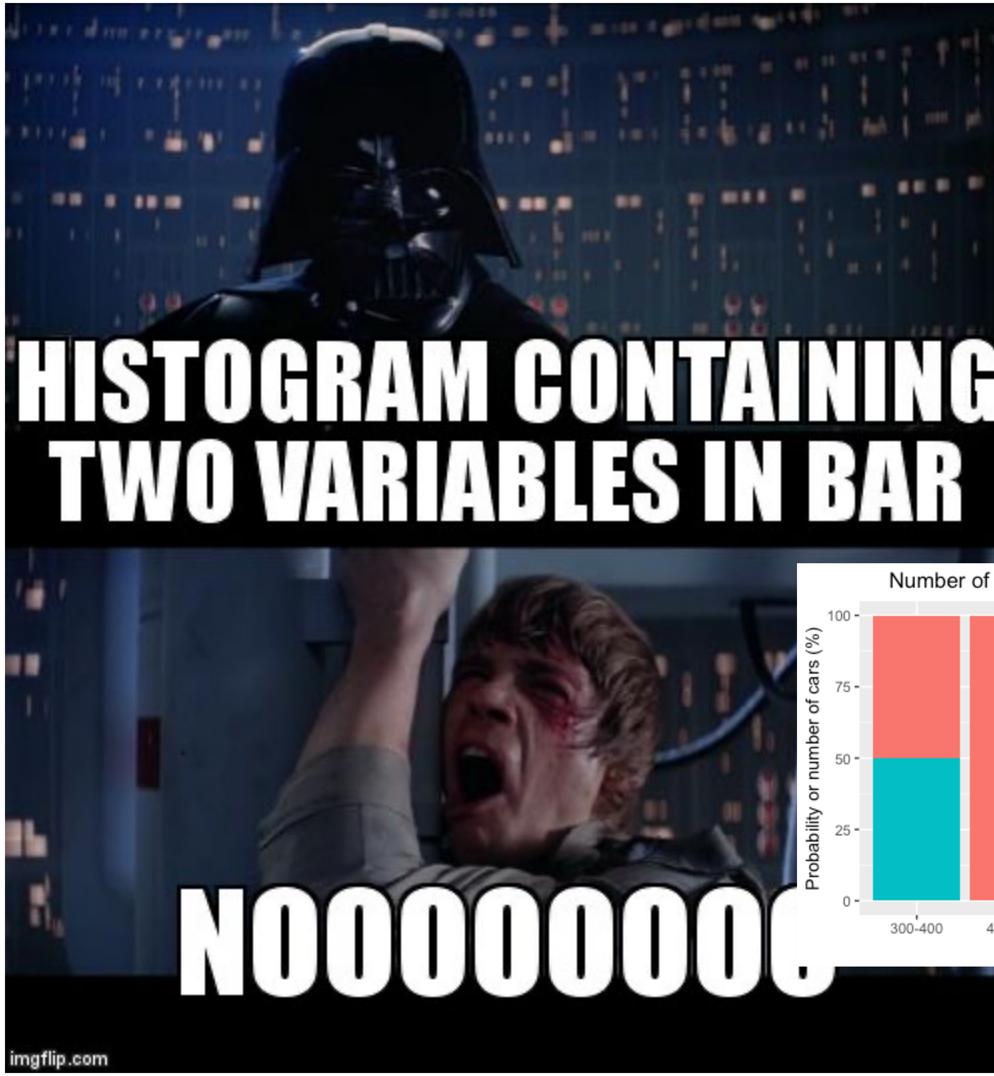
Class		Frequency	Relative Frequency
0.00-0.99	[0, 1)	9	0.138
1.00-1.99	[1, 2)	26	0.400
2.00-2.99	[2, 3)	11	0.169
3.00-3.99	[3, 4)	13	0.200
4.00-4.99	[4, 5)	3	0.046
5.00-5.99	[5, 6)	1	0.015
6.00-6.99	[6, 7)	2	0.031

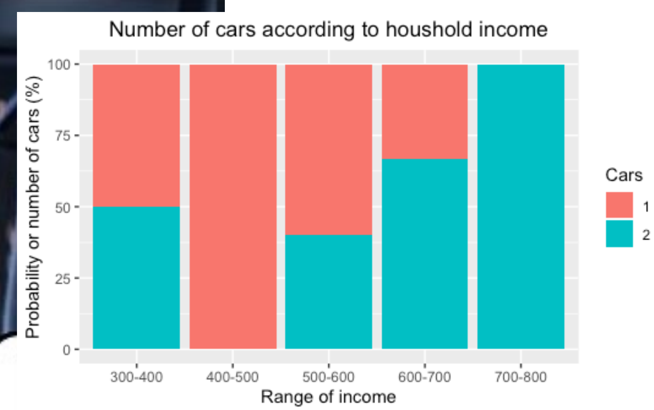
```
> hist(Particulate,cex.lab=1.2,cex.axis=1.2,col="lightgray",
+ xlab="Particulate (g/gal)")
```



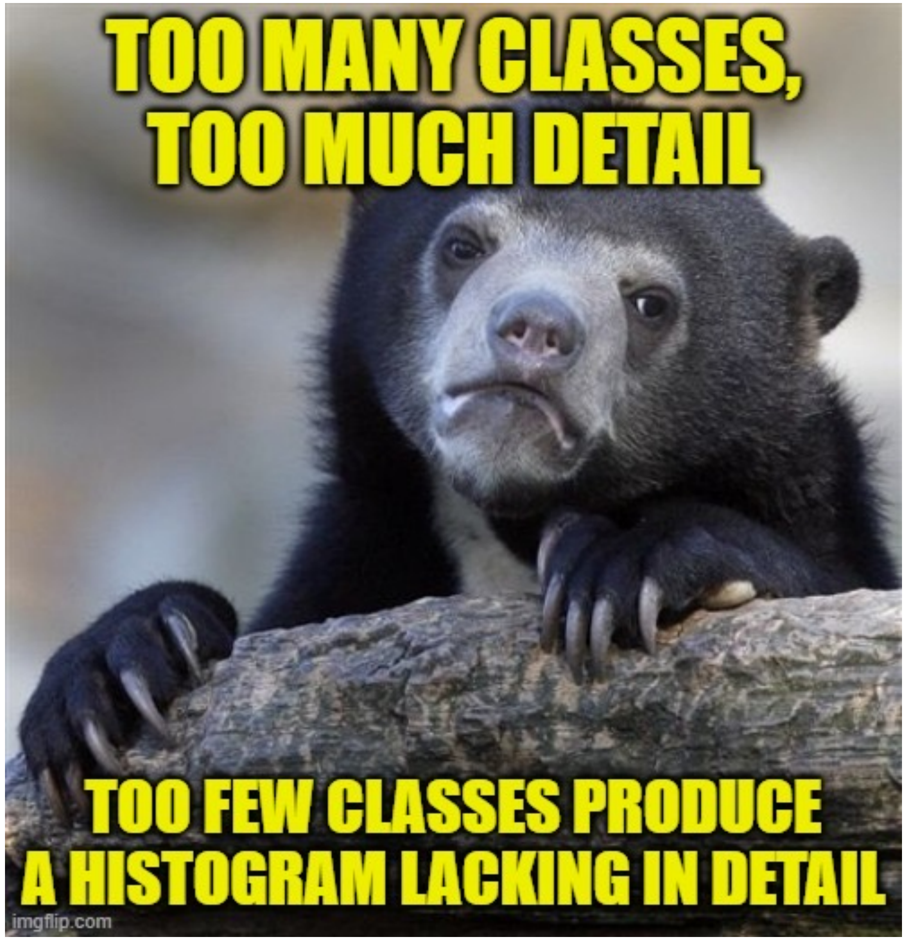



HISTOGRAM
IS NO BAR CHART
memegenerator.net


HISTOGRAM CONTAINING TWO VARIABLES IN BAR
NOOOOOOO
imgflip.com


Number of cars according to houshold income
Probability or number of cars (%)
100
75
50
25
0
300-400
400-500
500-600
600-700
700-800
Range of income
Cars
1
2

Selecting the Number of Classes?

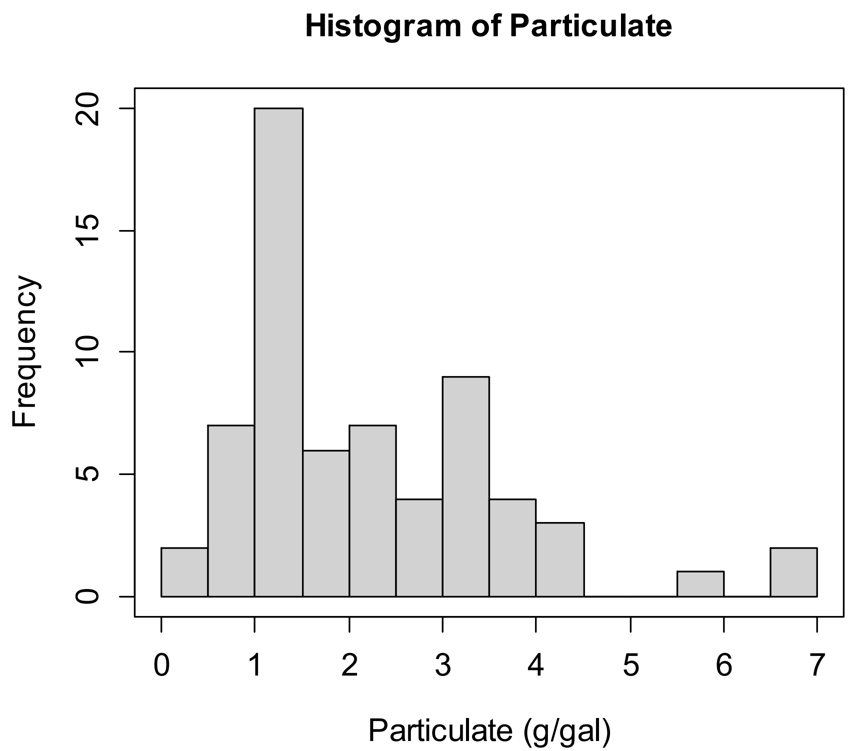


- A good rule of thumb is number of classes ≈ $\sqrt{n}$

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

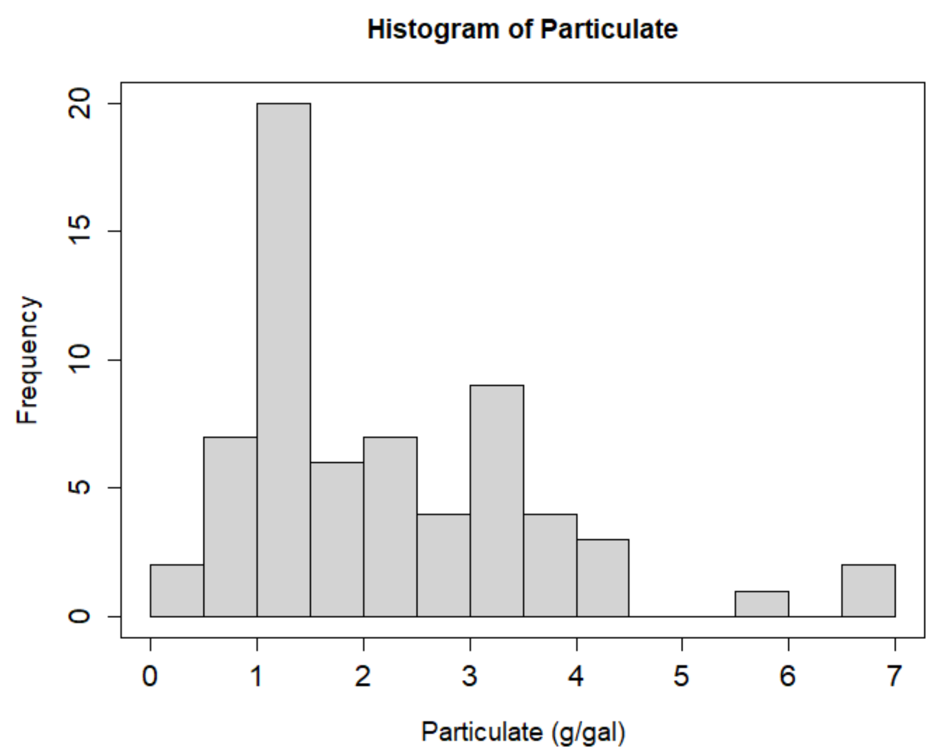
(0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7)

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()
```



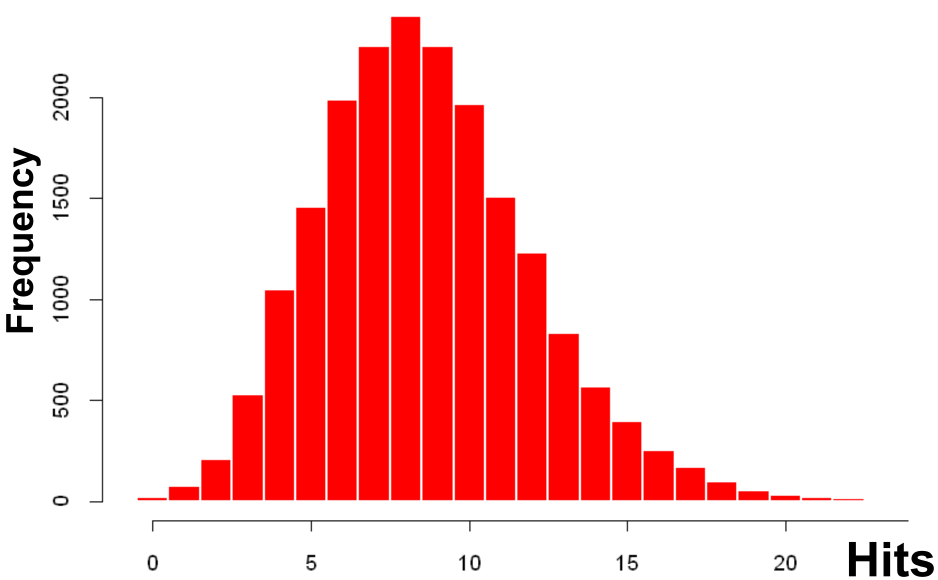
Above, the command setting breaks to 14 tells R we would like 14 classes. Note: when an integer is used in the `breaks` argument, the number is taken as a suggestion only, and the breakpoints will be set to "pretty" values.

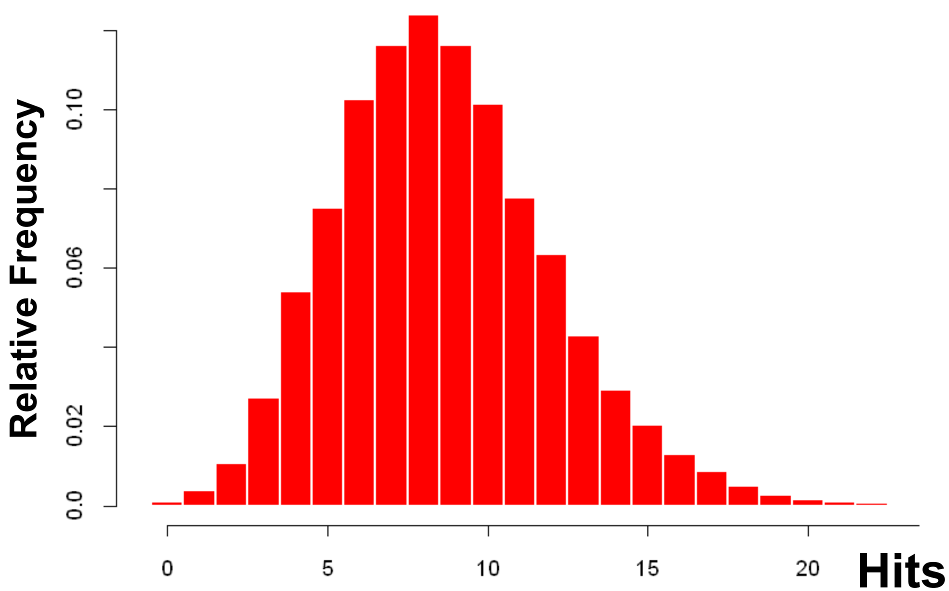
Histograms for Discrete Data

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.

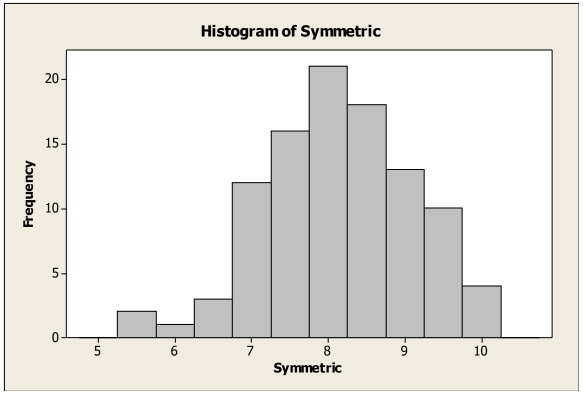




Practice time! Go to `Cars93` dataset and

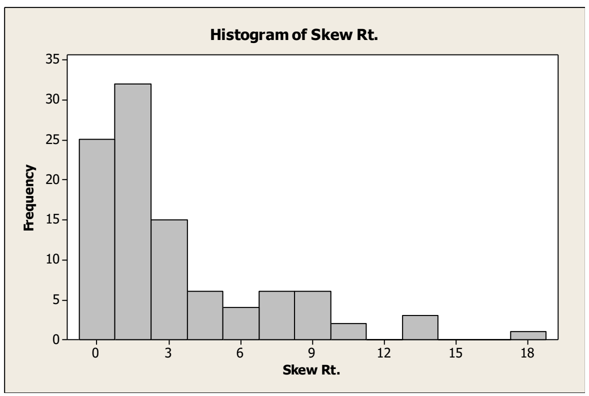
1. Pick a continuous numerical variable, create a histogram.
2. Pick a discrete numerical variable, create a barplot where each value forms its own category.

Shapes of Distributions



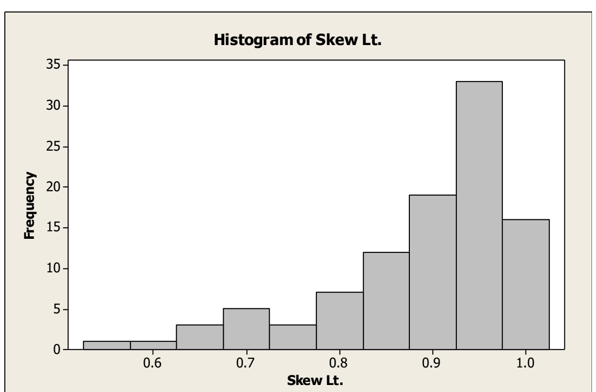
A histogram is ***symmetric*** if its right half is a mirror image of its left half.

- 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*.



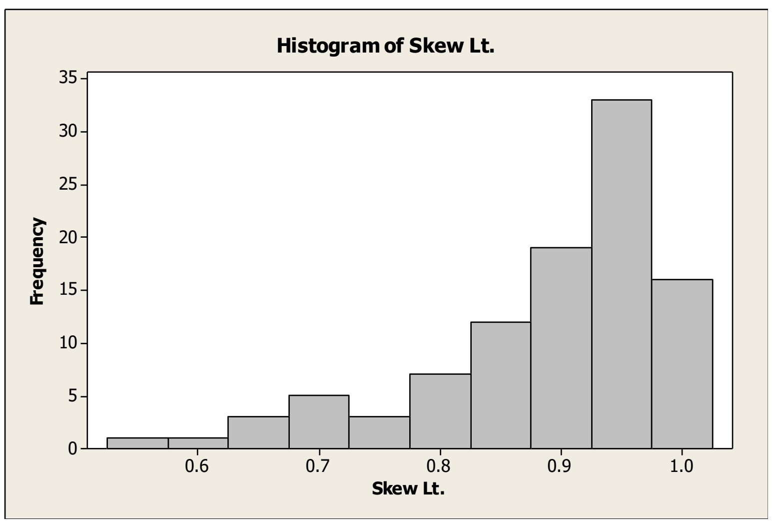
A histogram is ***skewed to the left*** if it has a long left-hand tail.

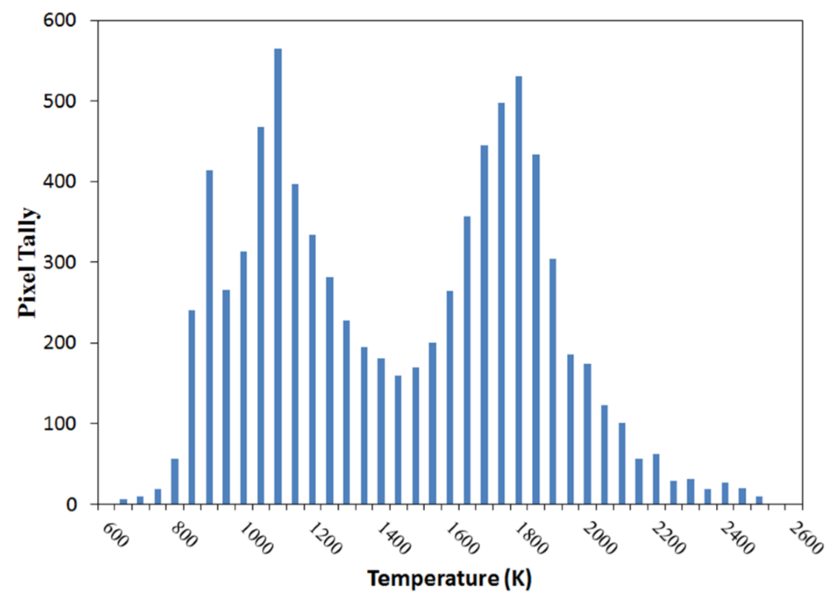
- Also called *negatively skewed*.

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



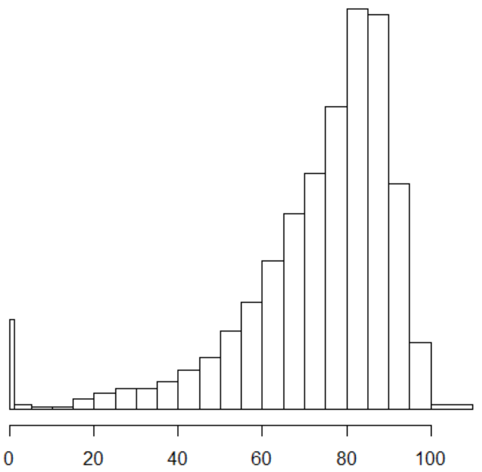


Elvidge CD, Zhizhin M, Baugh K, Hsu F-C, Ghosh T. Methods for Global Survey of Natural Gas Flaring from Visible Infrared Imaging Radiometer Suite Data. *Energies*. 2016; 9(1):14.

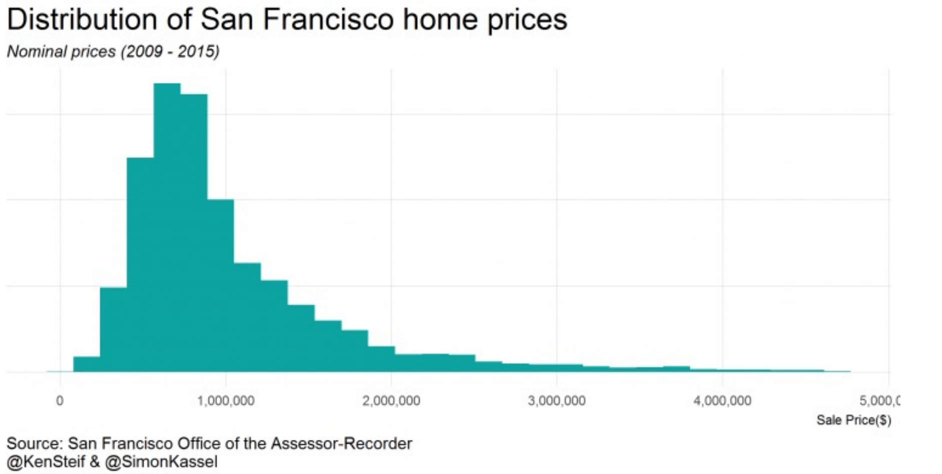
Shapes of Distributions

What shape would you expect for the following distributions?

– Age at death of American citizens



– Home prices in a metropolitan area?



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 $x_1, x_2, \ldots, x_n$

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

Measures of Center

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

We write this compactly as

upper limit of the summation

means sum $\longrightarrow \sum_{i=1}^{n} a_i \longleftarrow$ **terms**

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

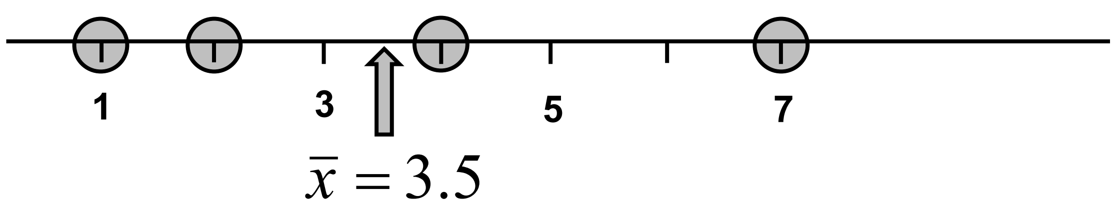
The Mean

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

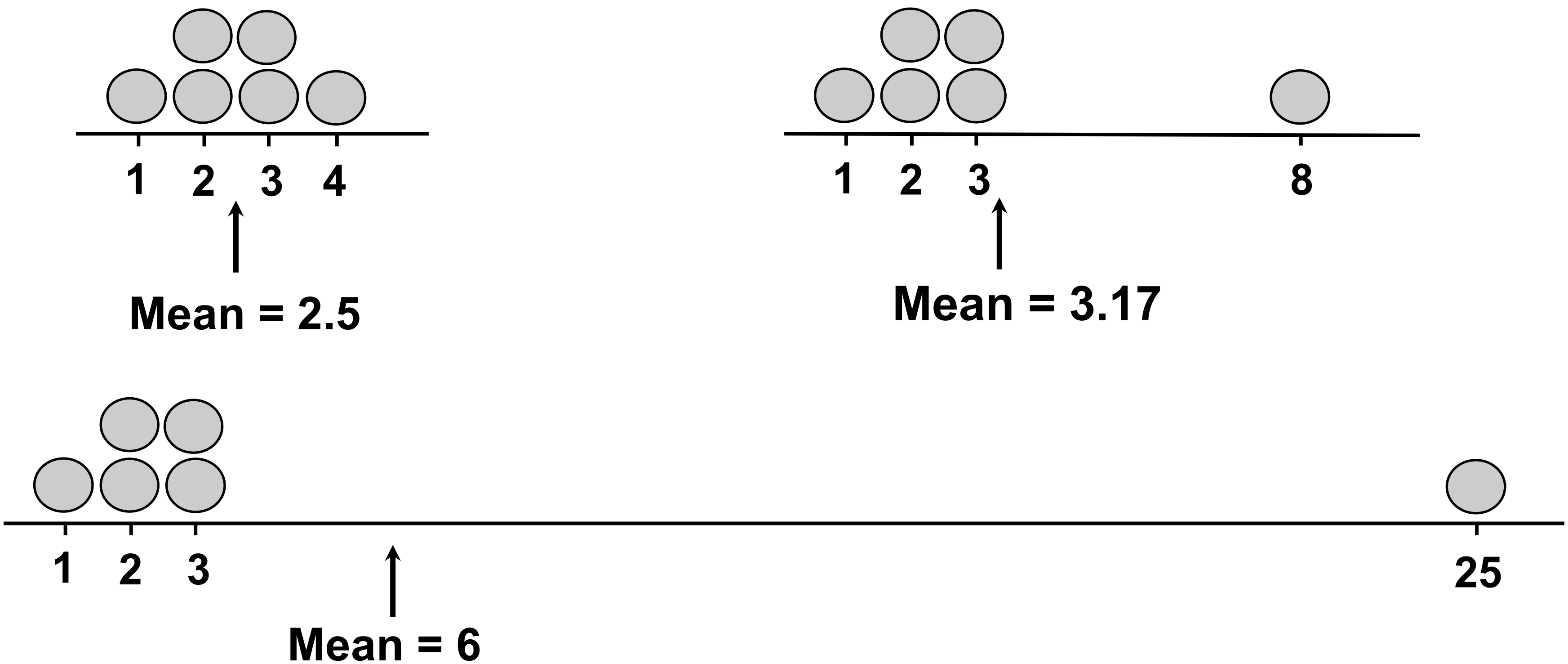
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{4+7+2+1}{4} = 3.5$$



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?



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

<u>Note</u>: The mean is not necessarily a "typical" value for the data set.

The Median

<u>Definition</u>: The ***median*** is the "middle" observation (once the values are arranged in order)

To calculate the median M, <u>sort all observations from smallest to largest:</u>

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

<u>Definition</u>: 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

$\uparrow$

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:

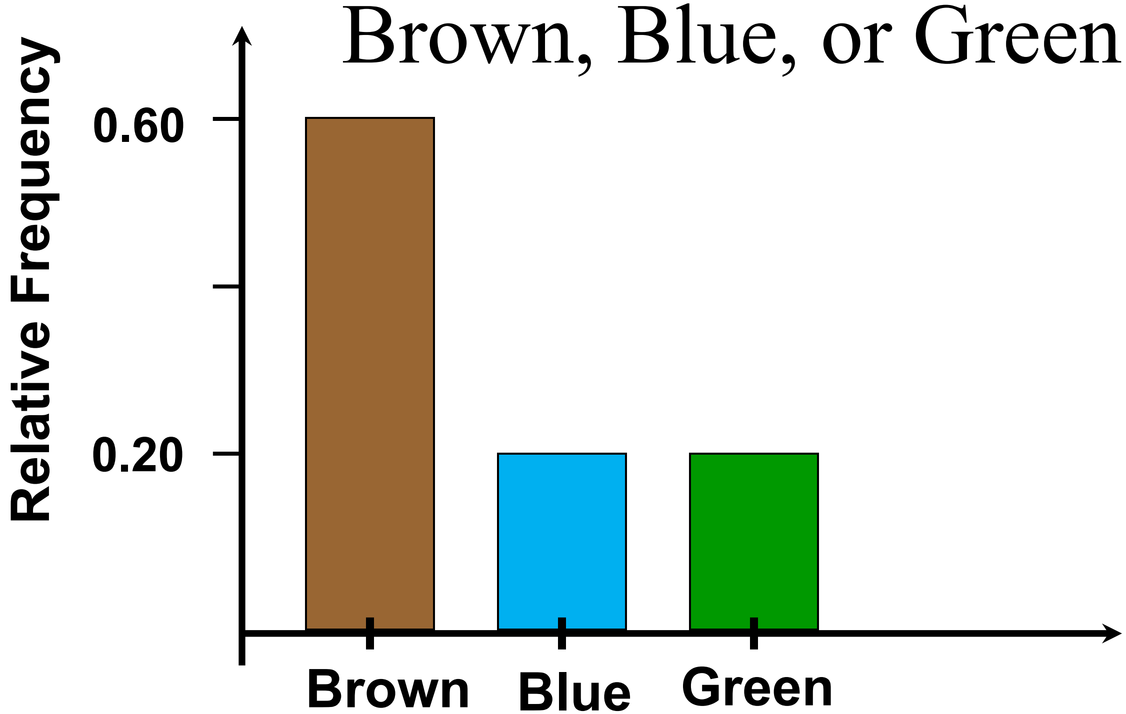
$\{ 0, \; 0, \; 0, \; 0, \; 1, \; 1, \; 2, \; 2, \; 3, \; 4 \}$

$\{ 0, \; 0, \; 0, \; 1, \; 1, \; 2, \; 2, \; 2, \; 3, \; 4 \}$

Mode Continued

The mode can be computed for qualitative data.

Example: Suppose that eye color was categorized as follows:



Practice time! Go to `majors` dataset posted in Github and

1. Make a barplot showing the different majors in our Data 180 class.
2. Report the mode.

Hints:

Remember, you can read a csv file in R directly from Github by clicking on “Raw” and using that link when calling `read.csv()`

Use `names=df$Major` option when using `barplot()` where `df` is a dataframe object

Variance and Standard Deviation

- Most commonly used numeric measures of variability.
- Both measure how far away data points are, on average, from their mean.
- Consider waiting times in minutes for the Carlisle brach of M&T Bank:

2, 3, 5, 7, 8

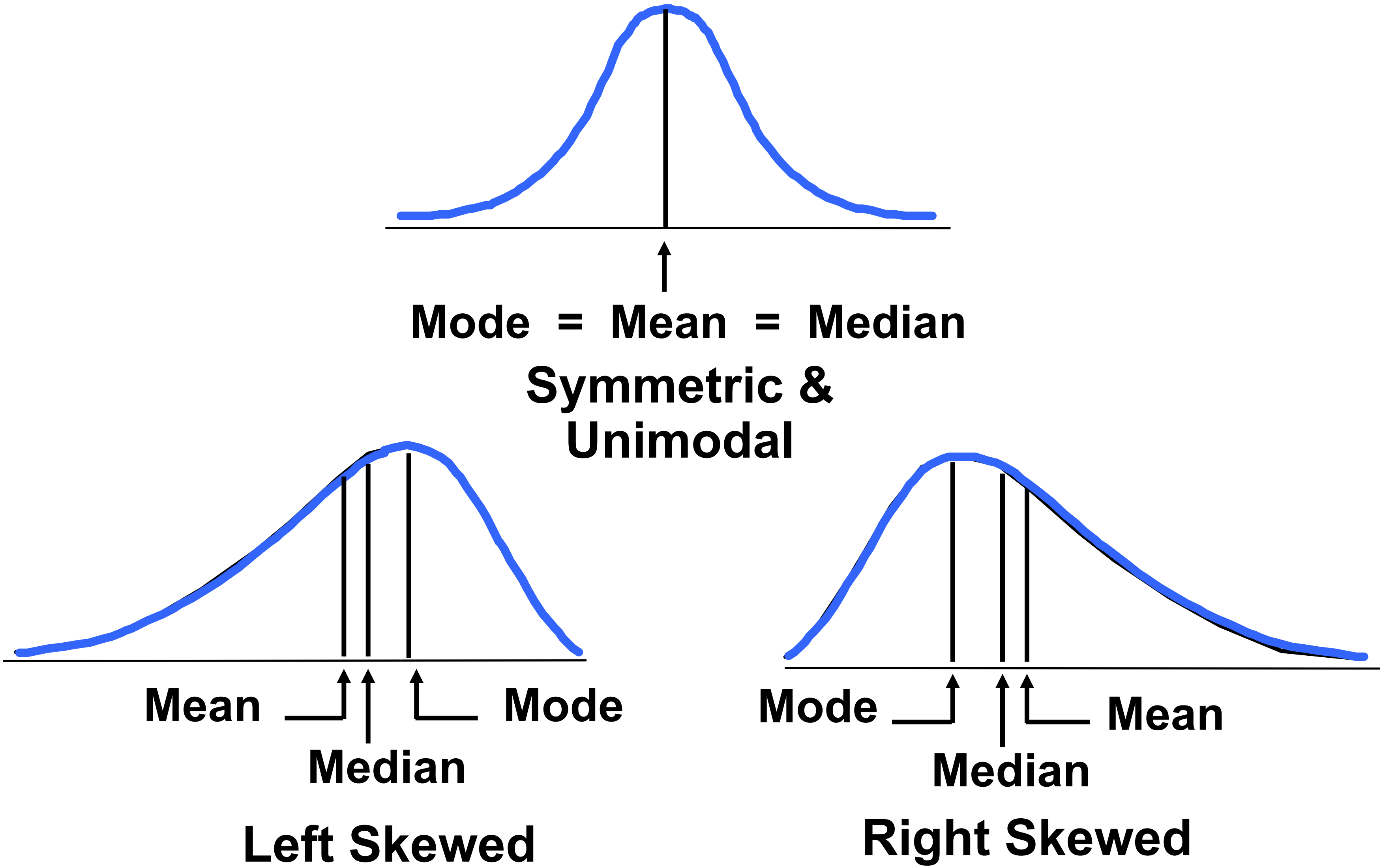
Definition: The ***sample variance***, denoted by s^2, is computed as follows:

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

Definition: The **sample standard deviation**, denoted by s, is the square root of the sample variance:

$$s = \sqrt{s^2}$$

Relationships of Measures of Center



Examples from the `Cars93` Data Frame

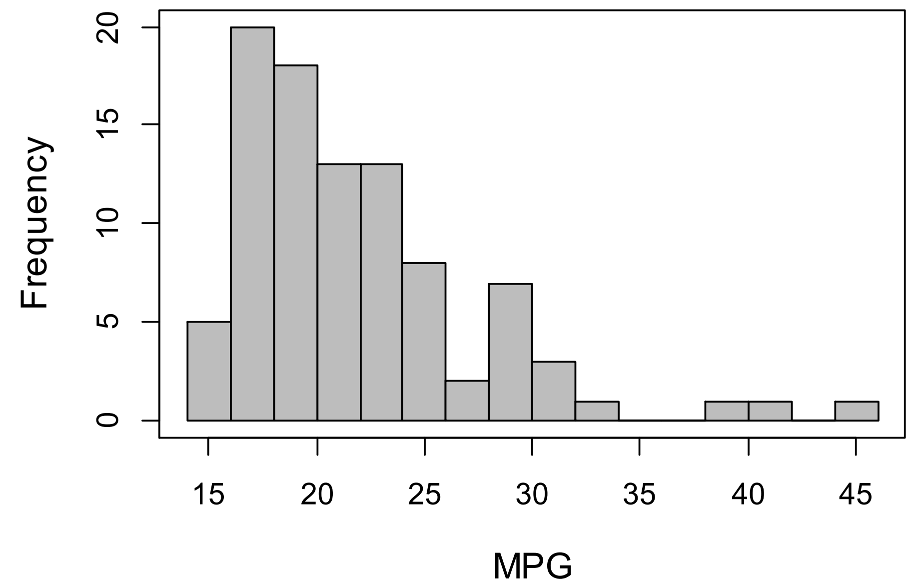
	Manufacturer	Model	Type	Min.Price	Price	Max.Price	MPG.city	MPG.highway	AirBags
1	Acura	Integra	Small	12.9	15.9	18.8	25	31	None
2	Acura	Legend	Midsize	29.2	33.9	38.7	18	25	Driver & Passenger
3	Audi	90	Compact	25.9	29.1	32.3	20	26	Driver only
4	Audi	100	Midsize	30.8	37.7	44.6	19	26	Driver & Passenger
5	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, median, sd of MPG.city

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

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?

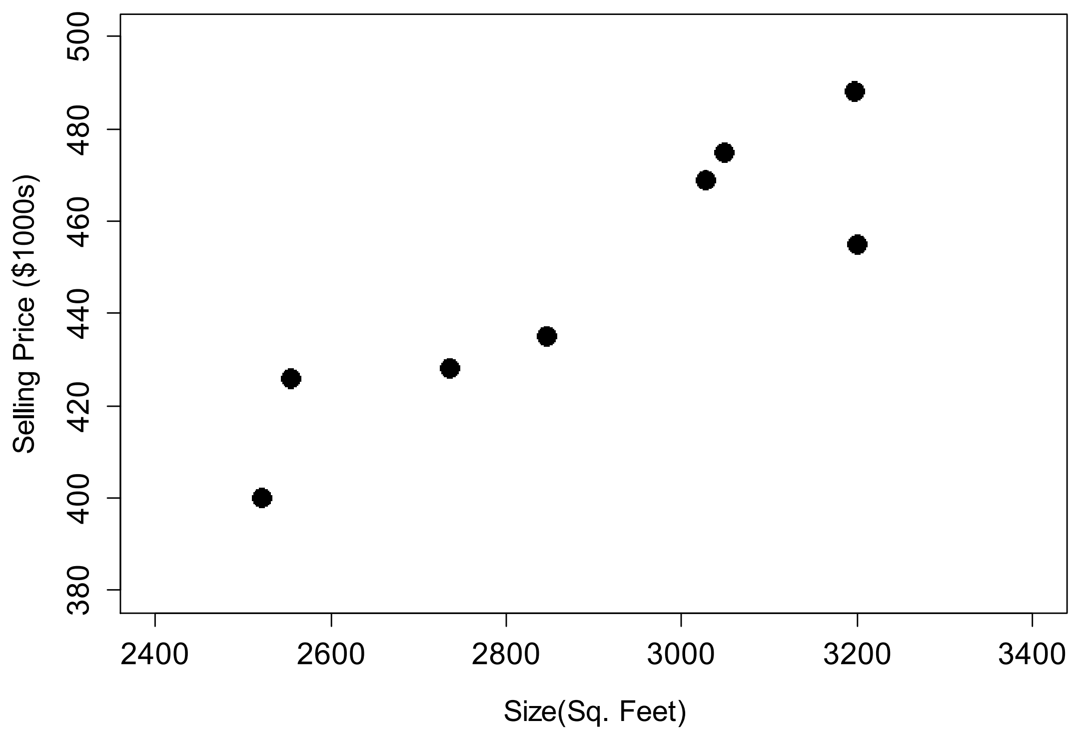
Do SAT scores predict college performance?

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

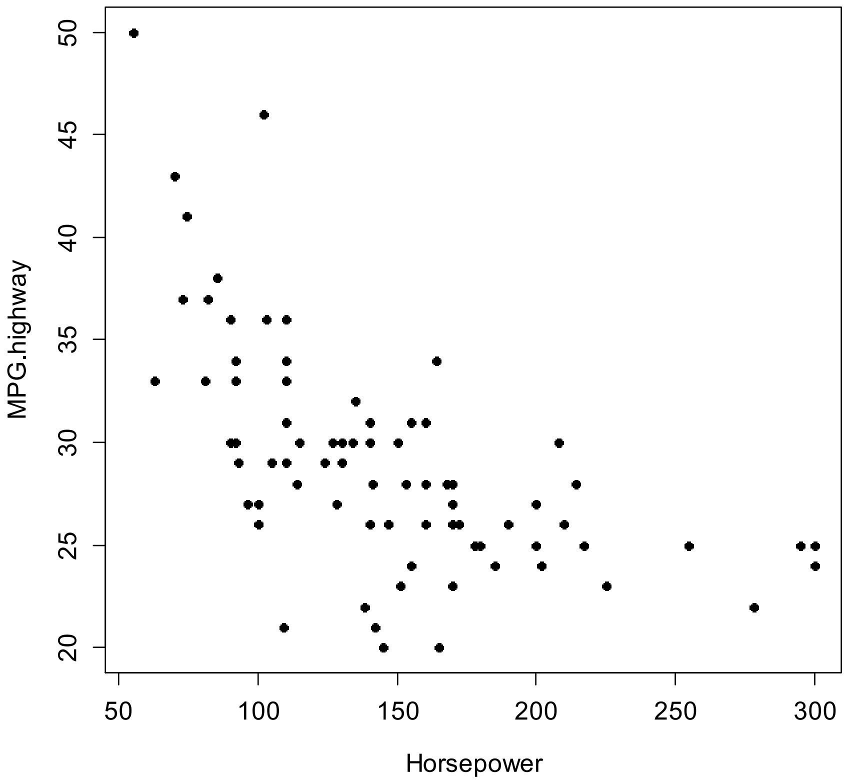
Size (Sq. ft.)	2521	2555	2735	2846	3028	3049	3198	3198
Selling Price ($1000s)	400	426	428	435	469	475	488	455

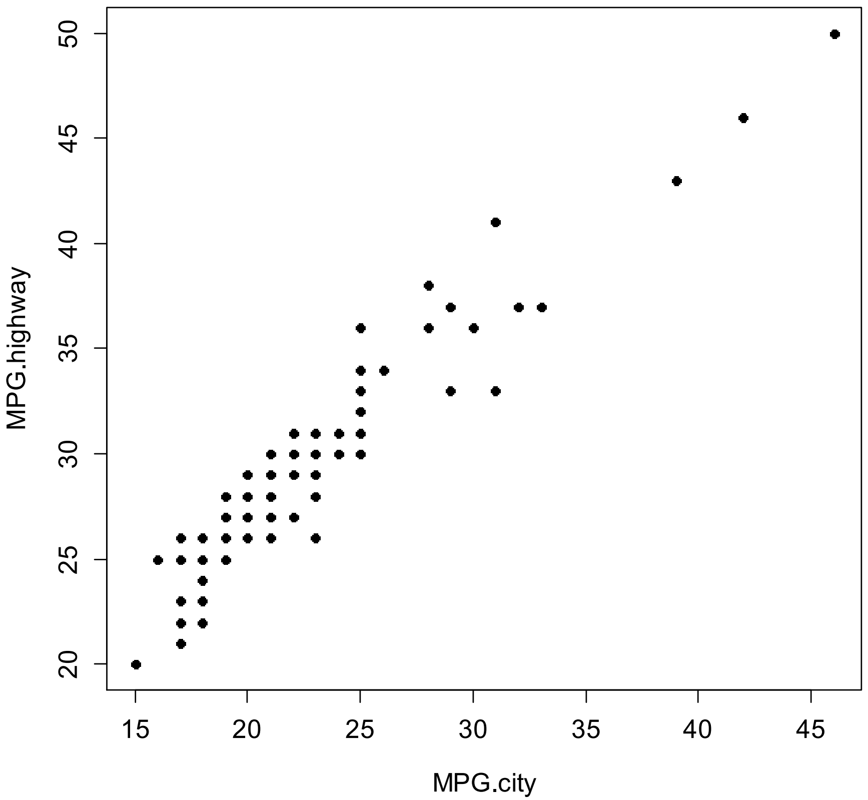
The table gives the size in square feet and the selling price in 1000s of dollars, for a sample of houses in a suburban Denver neighborhood. Here, each house is a unit and contributes an ordered pair of numbers: (Size, Selling Price) **NOTE: Correlation does not mean causation!!**

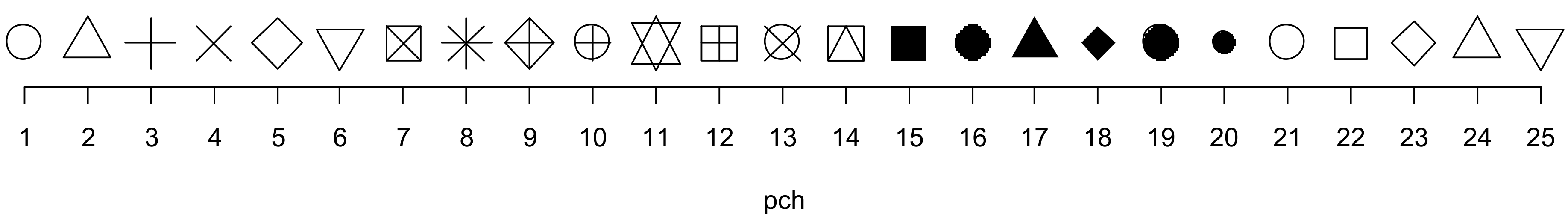


Examples from the `Cars93` Data Frame

```
> 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)
```



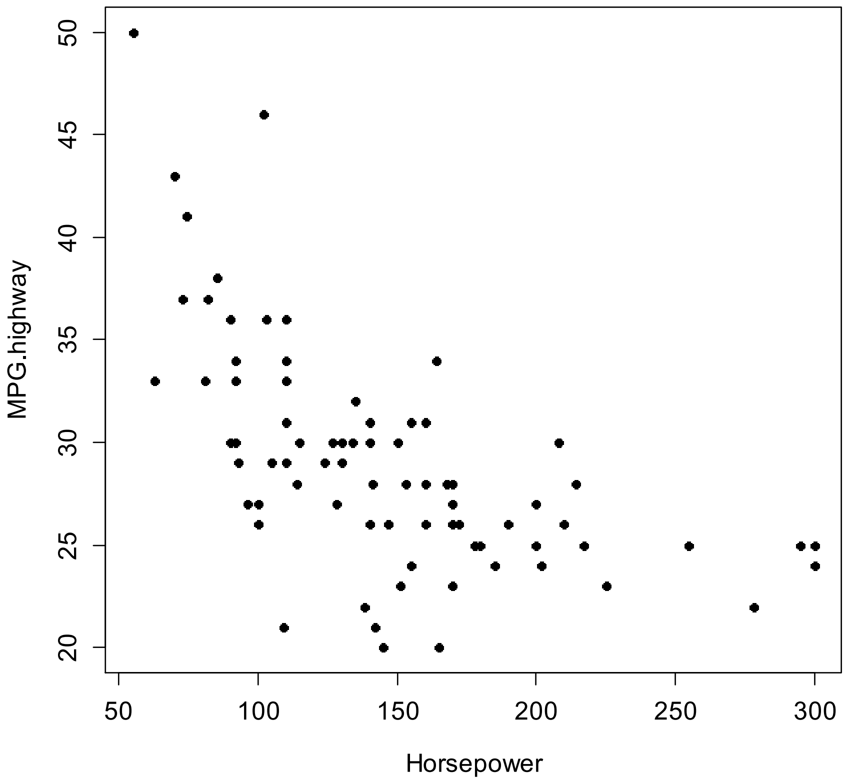


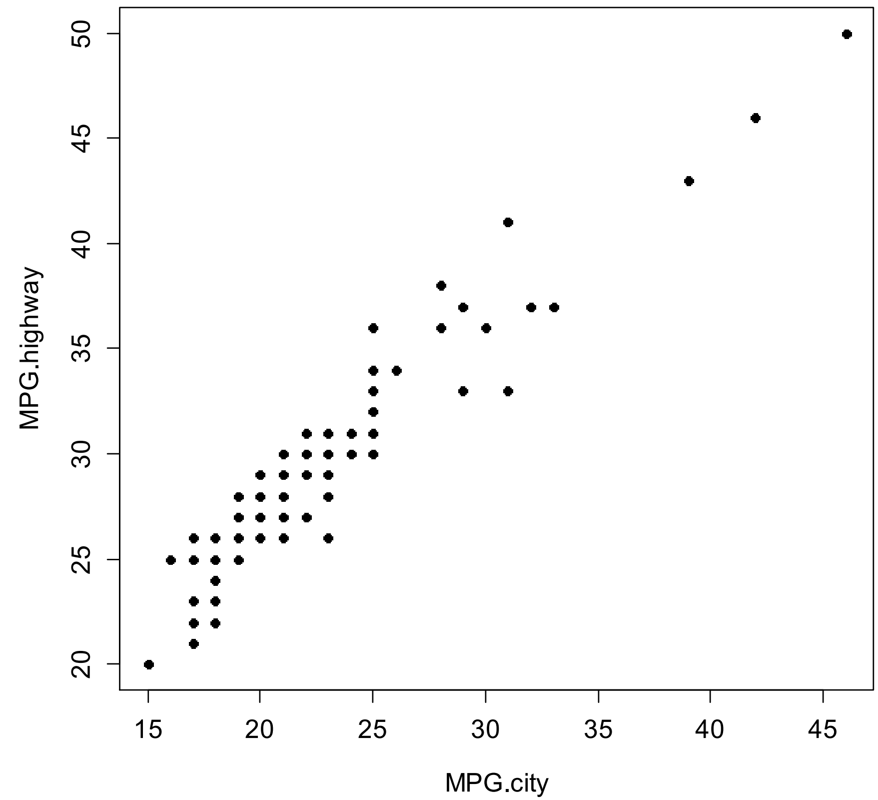


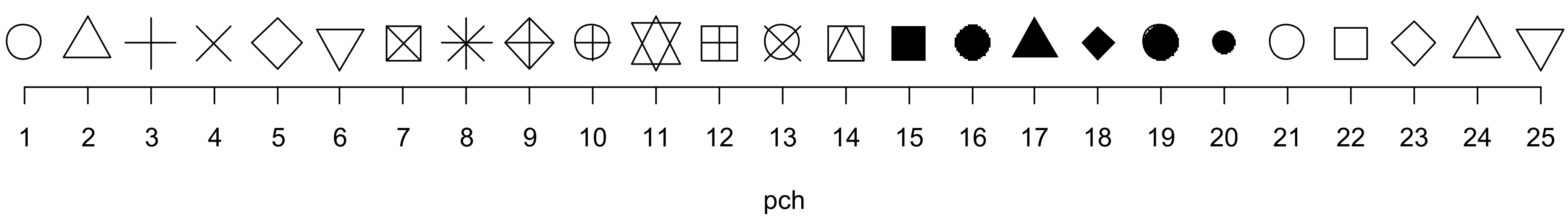
Examples from the `Cars93` Data Frame

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)
```







Practice time! Go to `Cars93` dataset and

1. Pick two continuous numerical variables, create a scatterplot.
2. Report the relationship between the two variables you chose.

ggplot?

- R built-in functions for plots
- ggplot