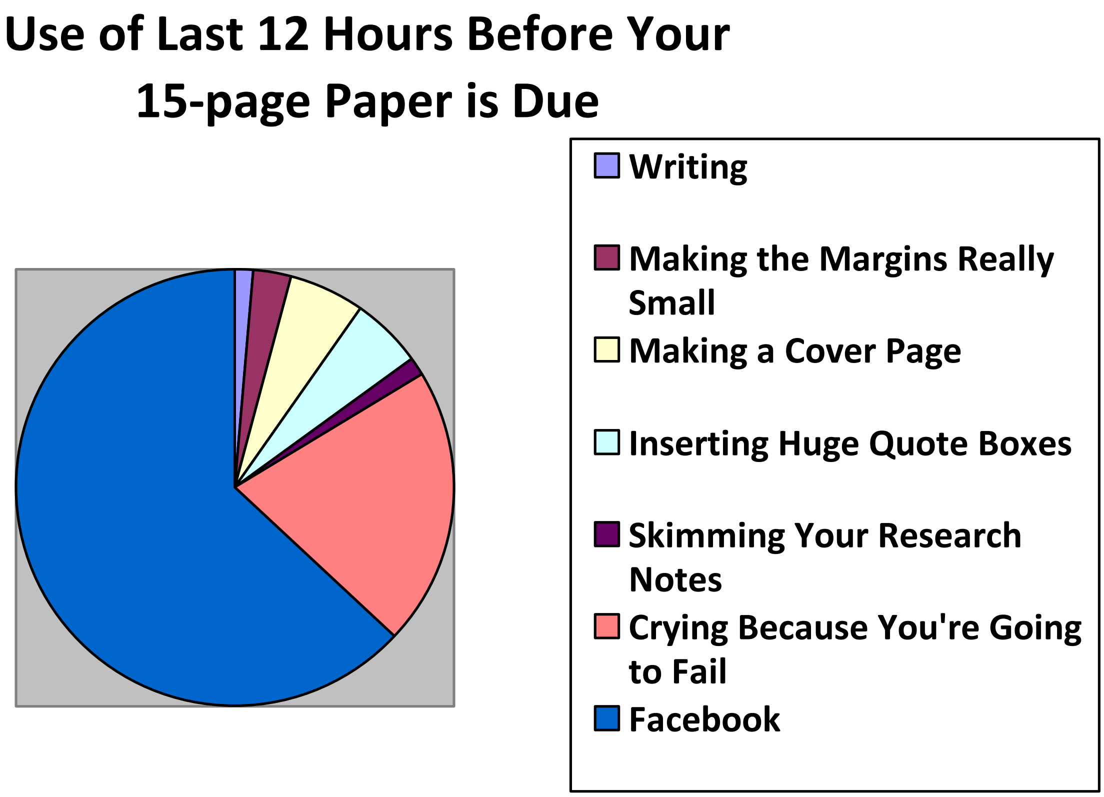
Chapter 2:
Summarizing & Graphing Data



Name ____

Period ______________

2.1 Review and Preview

Important characteristics of data:

- → **center:** a representative or average value that indicates where the middle of the data set is located

- → **variation:** the measure of the amount that the data values vary among themselves

- → **distribution:** the nature or shape of the distribution of the data

- → **outliers:** sample values that lie very far from the vast majority of the other sample values

- → **time:** changing characteristics of the data over time

*Helpful pneumonic device:
CVDOT (Computer Viruses Destroy or Terminate)

It's not as important to "crunch the numbers," so we will use technology instead & make practical use of the data through critical thinking. (However, you will learn the manual calculations so you gain deeper understanding & a better appreciation for the technology used.)

2.2 Frequency Distributions

Frequency: the number of times a data value or group of intervals occurs

Frequency Distribution : Ages of Best Actresses	
Age of Actress	Frequency
21-30	28
31-40	30
41-50	12
51-60	2
61-70	2
71-80	2

Frequency distribution (or frequency table): lists data values along with their corresponding frequencies (or counts) for the following purposes:

→ so large data sets can be organized and summarized
→ so we can gain some insight into the nature of the data
→ so we can construct important graphs (i.e.: histograms)

Lower class limits: the smallest numbers that can belong to the different classes

Upper class limits: the largest numbers that can belong to the different classes

Class boundaries: the numbers used to separate classes but without the gaps created by class limits

Class Midpoints: the values in the middle of the classes found by adding the lower class limit to the upper class limit & dividing by 2

Class width: the difference between two consecutive lower class limits or two consecutive lower class boundaries

→ Do not make the class width using the lower class limit & the upper class limit

Constructing a Frequency Distribution:

1. Decide on the number of classes you want (should be between 5 and 20)
2. Calculate class width = $\frac{\text{max. value} - \text{minimum value}}{\#\text{ of classes}}$

   → Round this (usually up) to get a convenient number
3. Begin by choosing a number for the lower limit of the 1st class
4. List the other lower limits by adding the class width to the lower limit of the 1st class then to the new lower limits
5. Enter the upper class limits
6. Use tally marks to find the total frequency for each class

7. Count the tally marks to express the total frequency for each class

- Be sure that classes do not overlap (each of the original values must belong to exactly one class)
- Try to use the same width for all classes (although it's sometimes impossible to avoid open-ended intervals i.e.: 65 and older)

Ex: Use the following depths of 50 earthquarkes (Data Set 8 in Appendix B) to construct a frequency table with a lower class limit of 2.0 and a class width of 2.0.

6.6	2.2	18.5	7.0	13.7	5.4	5.3	5.9	4.7	14.5
2.0	14.8	8.1	18.6	4.5	17.7	15.9	15.1	8.6	5.2
15.3	5.6	10.0	8.2	8.3	9.9	13.7	8.5	8.2	7.9
17.2	6.1	13.7	5.7	6.0	17.3	4.2	14.7	15.2	3.3
3.2	9.1	8.0	18.9	14.2	5.1	5.7	16.4	10.1	6.4

Earthquake Depths (km)	**Tally**	**Frequency (number of earthquakes)**

Relative Frequency: Found by dividing each class frequency by the total of all frequencies

Relative Frequency = $\frac{\text{class frequency}}{\text{sum of all frequencies}}$

Relative Frequency Distribution: includes the same class limits as a frequency distribution, but relative frequencies are used instead of actual frequencies & are often expressed as percents

Earthquake Depths (km)	Relative Frequency

* If constructed correctly, the sum of the relative frequencies should total 1 or 100% (may be a little off due to rounding errors)

* Percents make it easier for us to understand the distribution of the data & compare it to other data sets

Cumulative Frequency: the sum of the frequencies for that class & all previous classes

Earthquake Depths	Cumulative Frequency

→ Lower class limits are replaced by "*less than*" expressions

→ *Upper class boundaries* are used instead of upper class limits

Interpreting Frequency Distributions:

Characteristics of a Normal Distribution:

1. *"bell shape"*—frequencies start low, increase to some maximum frequency, & then decrease to a low frequency

2. *approximately symmetric*—frequencies are evenly distributed on both sides of the maximum frequency

Ex #1: Would you say that the 50 earthquakes from the previous example, has a normal distribution?

Ex #2: Complete the following frequency distribution so that there are 21 sample values that make up a normal distribution.

Interval	Frequency
20-24	2
25-29	5
30-34	7
35-39	?
40-44	2

Ex #3: Complete the following frequency distribution so that there are 30 sample values that make up a normal distribution.

Interval	Frequency
20-29	3
30-39	5
40-49	?
50-59	?
60-69	5
70-79	3

Gaps: gaps in the data can suggest that we have data from 2 or more populations

Weight (grams) of Penny	Frequency
2.40–2.49	18
2.50–2.59	19
2.60–2.69	0
2.70–2.79	0
2.80–2.89	0
2.90–2.99	2
3.00–3.09	25
3.10–3.19	8

Ex: This shows a frequency table of the weights (in grams) of randomly selected pennies but there is a large gap between the lightest & heaviest pennies. Later it was discovered that pennies made before 1983 are 97% copper and 3% zinc. After 1983, they are 3% copper and 97% zinc.

2.3 Histograms

<u>Histogram:</u> a bar graph in which the horizontal scale represents classes of data values & the vertical scale represents frequencies

→ the heights of the bars correspond to the frequency values

→ the bars are drawn next to each other (no gaps)

<u>Constructing a Histogram:</u>

1. Construct a *frequency distribution table*

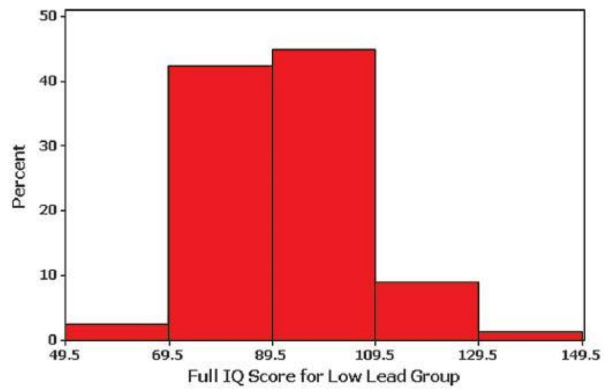
2. *Horizontal scale of the histogram:* Mark with the *<u>class boundaries.</u>*
   --Scale of the horizontal axis: subdivide in a way that allows all classes to fit well

3. *Vertical scale of the histogram:* Use the *<u>class frequencies</u>*
   --Scale of the vertical axis: the maximum frequency (or next highest convenient #) should be at the top of the vertical scale & 0 should be at the bottom
   --rule of thumb: The vertical height of the histogram should be about $\frac{3}{4}$ of the total width

4. Both axes should be *clearly labeled* & give the histogram a title

   Ex: Use the frequency table about the 50 earthquakes to construct a histogram:

Relative Frequency Histogram: has the same shape & horizontal scale as a histogram, but the vertical scale is marked with relative frequencies instead of actual frequencies

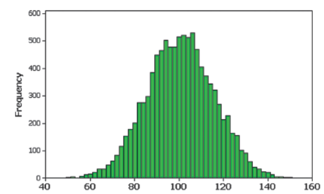


Analyze this histogram using the pneumonic device "CVDOT":

Interpreting the Distribution of Histograms:

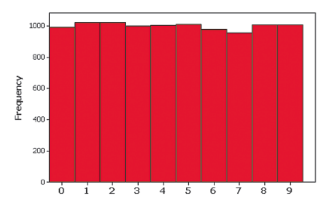
Normal Distribution:

- "bell" shape
- Frequencies increase to a maximum and then decrease
- Symmetric where the left half is roughly a mirror image of the right half



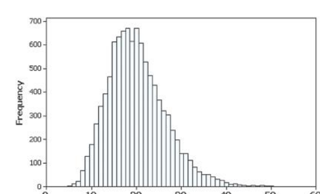
Uniform Distribution:

- The different possible values occur with approximately the same frequency
- The heights of the bars are approximately uniform

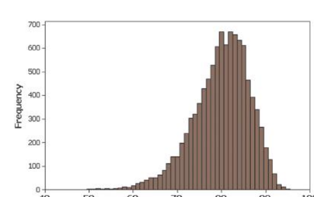


Skewed:

- Not symmetric & extends more to one side than the other
- Skewed to the right (or positively skewed): has a longer right tail
  - More common because it's easier to get exceptionally large values rather than very small (ie: incomes)



- Skewed to the left (or negatively skewed): has a longer left tail



2.4 Graphs that Enlighten and Graphs that Deceive

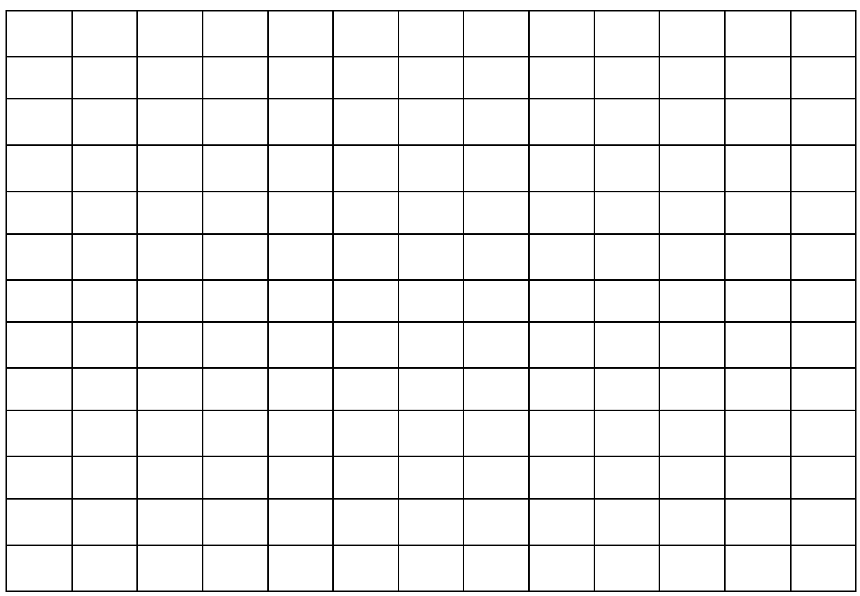
<u>Scatterplot</u> (or scatter diagram): is a plot of paired (x, y) data with a horizontal x-axis and a vertical y-axis

→ *To Construct a scatterplot:*

1. Construct a horizontal axis for the values of the first variable
2. Construct a vertical axis for the values of the second variable
3. Plot the points
4. Label the axes

→ Ex: The following data represents the numbers of cricket chirps per minute paired with temperatures in °F:

temperature (°F)	72	84	76	87	80	92	70	83
chirps per minute	960	1150	860	1200	900	1190	875	1040



- Does there appear to be a relationship between chirps & temperature as is shown by the pattern of the points?

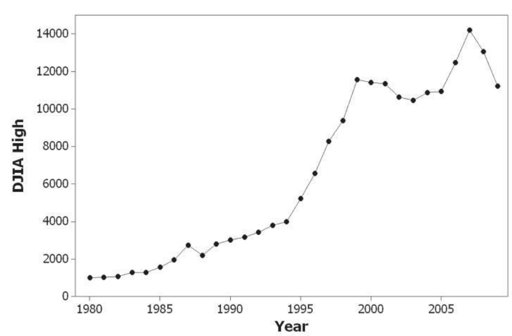
→ Correlation Coefficient: the r value (between -1 and +1) that shows how closely the points on a scatter plot fit the pattern of a straight line.

- If r is close to +1: there is a strong positive correlation between variables
- If r is close to -1: there is a strong negative correlation between variables
- If r is close to 0: there is no relationship between variables
- Example from above: ____ ____
- We will learn to find the actual correlation coefficient in unit 10

Time series graph: a graph of time-series data, which are data that have been collected at different points in time

→ It is often important to know when population values change over time

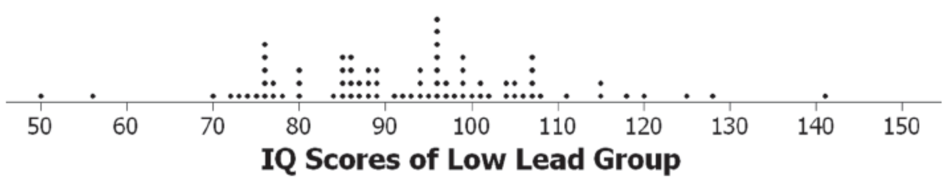
→ Ex: A graph that shows the yearly high values of the Dow Jones Industrial Average for the New York Stock Exchange



Dotplot: a graph in which each data value is plotted as a point (or dot) along a horizontal scale of values

→ *Dots* representing equal values are stacked

→ *Be sure to label the number line with equal increments*



Stemplots (or stem-and-leaf plot): represents data by separating each value into two parts: the *stem* (the leftmost digit(s)) and the *leaf* (the rightmost digit)

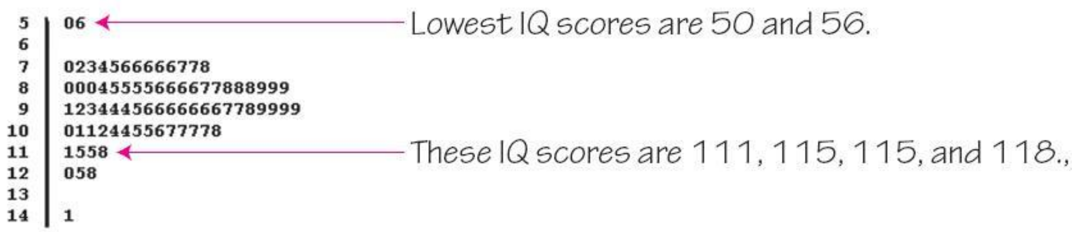
→ Leaves are arranged in increasing order, not the order in which they occur in the original list

→ If we turn the stemplot on its side, we can see a distribution of the data

→ *Advantages of a stemplot:* shows the distribution of the data, retains all the information from the original list, & the construction is a quick and easy way to sort data

→ Better stemplots are often obtained by first rounding the original data values

→ Stemplots can be expanded to create more rows and condensed to include fewer rows



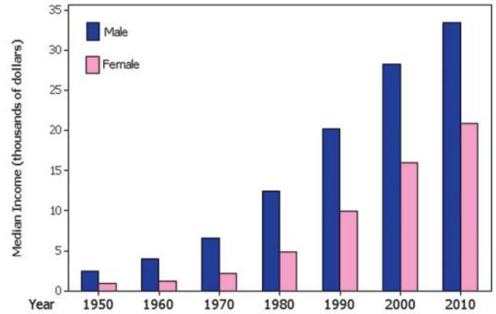
→ Ex: Use the following data to create a stemplot.

Minutes Spent Watching TV Over the Last 18 Days					
48	125	98	147	45	94
92	101	75	90	120	61
60	60	63	65	136	84

Stem | Leaf

Bar Graph: uses bars of equal width to show frequencies of categories of categorical (or qualitative) data

- ➔ Vertical scale represents frequencies or relative frequencies
- ➔ Horizontal scale identifies the different categories
- ➔ Because it is based on categorical data, the bars may or may not be separated by small gaps



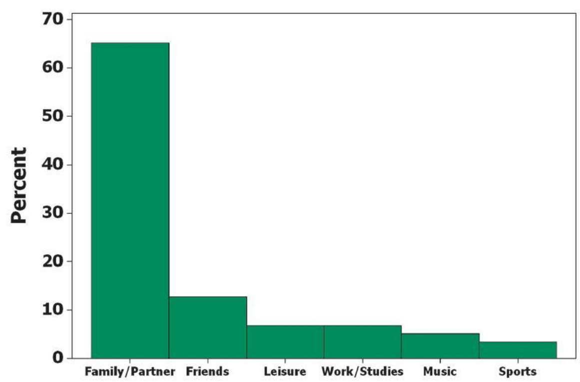
Multiple Bar Graph: Median Income by Gender

Multiple bar graph: has two or more sets of bars and is used to compare two or more data sets

Pareto chart: a bar graph for qualitative data, with the bars arranged in descending order according to frequencies

→ *Vertical scales* can represent frequencies or relative frequencies

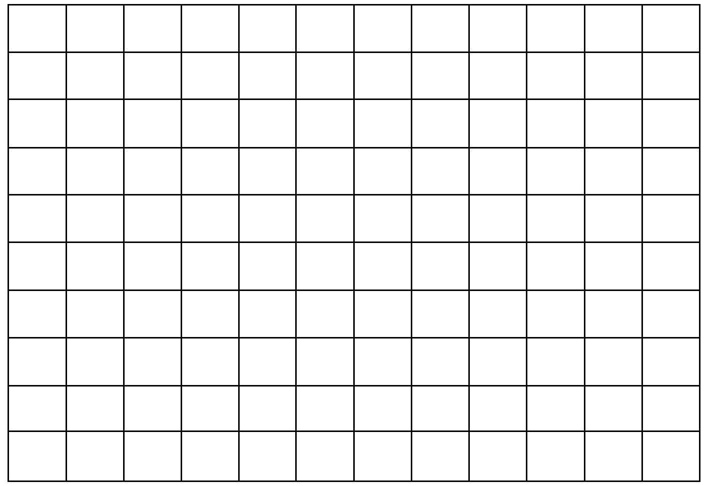
→ *Tallest bar* is at the left and smaller bars are to the right which focuses the attention to the more important categories



Pareto Chart: What Contributes Most to Happiness?

→ Let's create a pareto chart with the types of pets that we have:

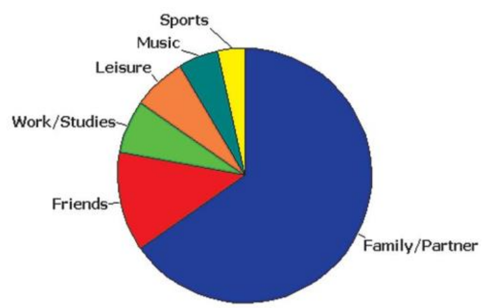
Type of Pet	**Tally**
Dog	
Cat	
Bird	
Fish	
Gerbil	
Horse	
Snake	



Pie Chart: a graph that depicts categorical (or qualitative) data as slices of a circle

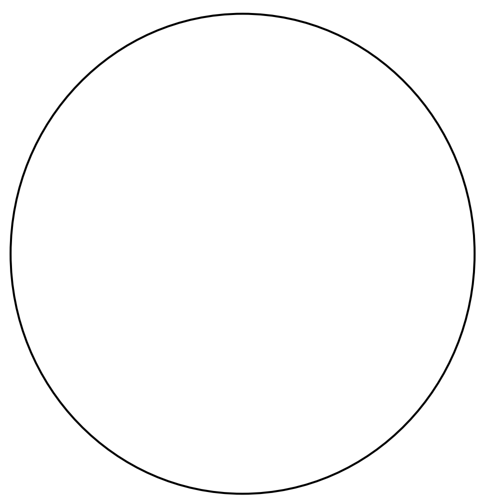
→ *Construction involves slicing the pie into the proper proportions:*

1. Find the % of the total for each category
2. Change the % to a decimal & multiply it by 360 to get the # of degrees in each slice of the pie



→ Let's use the pet example to create a pie chart:

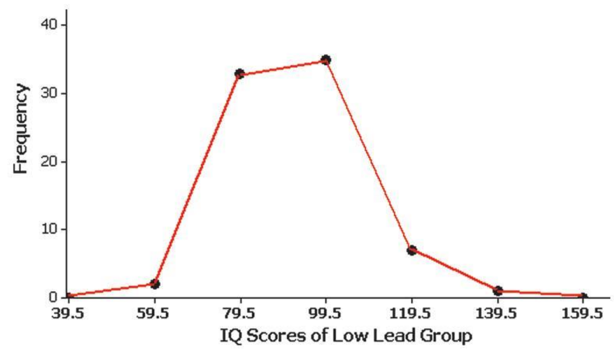
Type of Pet	**Percent of Total**	**# of Degrees**
Dog		
Cat		
Bird		
Fish		
Gerbil		
Horse		
Snake		



* The pareto chart does a better job of showing us the relative sizes of the different categories than the pie chart.

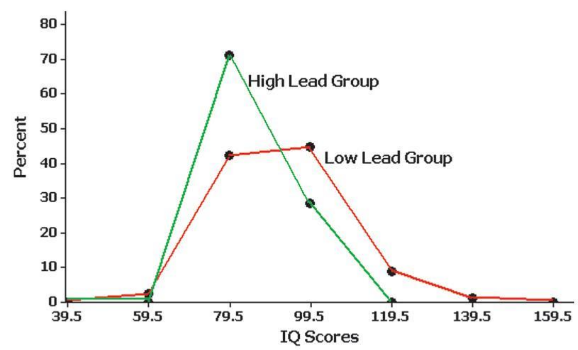
Frequency Polygon: uses line segments connected to points located directly above class midpoint values

- → Similar to a histogram but a frequency polygon uses line segments instead of bars
- → heights of the points correspond to the class frequencies
- → line segments are extended to the left and right so that the graph begins & ends on the horizontal axis



Relative Frequency Polygon: similar to a frequency polygon, but uses relative frequencies for the vertical scale

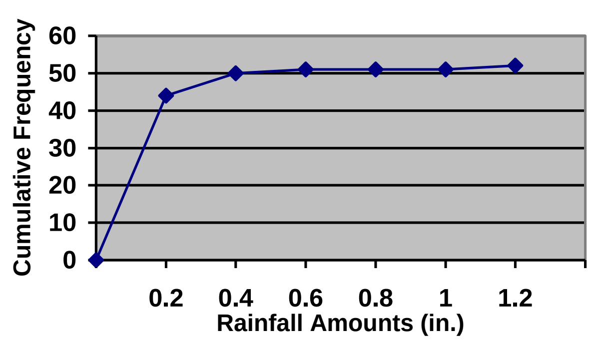
→ When comparing 2 data sets, it's often helpful to graph two relative frequency graphs on the same axes



Ogive (pronounced "oh-jive"): a line graph that depicts cumulative frequencies

- → Uses class boundaries along the horizontal scale
- → Uses cumulative frequency for the vertical scale
- → Graph begins with the lower boundary of the 1st class & ends with the upper boundary of the last class
- → Useful for determining the number of values below some particular value

Sunday Rainfall in Boston



--How many Sundays in Boston had rainfall amounts less than 0.395 inches?

Example: The following data describes the number of wins by 20 major league baseball teams in the 2014 season.

64	79	66	71	73	73	76	85	66	90
70	89	98	94	77	82	77	79	84	88

a. Create a relative frequency distribution with 5 classes

Number of Games Won	**Tally**	**Frequency**	**Relative Frequency**

b. Using the relative frequency distribution from part a, construct a relative frequency polygon:

c. Create a cumulative frequency distribution.

Number of Games Won	**Frequency**	**Cumulative Frequency**

d. Using the cumulative frequency distribution from part c, construct an ogive:

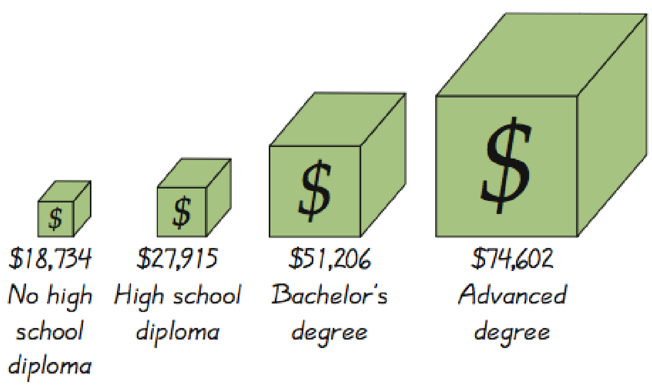
Graphs that Deceive:

1. Non-zero Axis:
   - by using a vertical axis that starts with a value other than 0, small differences can be exaggerated
2. Pictographs:
   - Data that are one-dimensional in nature are often depicted with two- or three-dimensional objects
   - By using pictographs, artists can create false impressions that grossly distort differences by using basic geometric principles

     --When you double each side of a square, the area increases by a factor of 4 (not 2)

     --When you double each side of a cube, the volume increases by a factor of 8 (not 2)

Ex: Look at the following pictograph that compares income and educational attainment. It depicts one-dimensional data with 3-dimensional boxes. The last box is actually 64 times as large as the first box but the income is only 4 times as large.



3. Missing Data

   * Make sure all data is included in the graph, otherwise accurate conclusions cannot be made.