Stat 201: Statistics I
Chapter 2

January 22, 2018

Chapter 2
Summarizing and Graphing Data

Section 2.1
Frequency Distributions for Organizing and Summarizing Data

Frequency distributions

A **frequency** is the number of times a particular value occurs in a set of data, i.e. the count.

A **frequency distribution** (or **frequency table**) summarizes a set of data by listing the frequencies of data in categories or classes (groups).

- For categorical data, the categories are simply the possible values of the data.
- For quantitative data, the classes are usually ranges of possible values.

Frequency distribution for categorical data

Example

Favorite kind of taco = {Chicken, Fish, Fish, Veggie, Chicken, Beef }

Kind of taco	Frequency
Beef	1
Chicken	2
Pork	0
Fish	2
Veggie	1

Frequency distribution for quantitative data

Example

Tacos eaten $= \{3, 0, 17, 6, 4, 3, 5\}$

Number of tacos eaten	Frequency
0 - 4	4
5 - 9	2
10 - 14	0
15 -20	1

Relative frequency

Relative frequency is the proportion (fraction) of the whole data set that resides in each category or class. When expressed as a percent it is called **percentage frequency**.

To calculate: For each class,

$$\text{Relative frequency} = \frac{\text{class frequency}}{\text{total count}}$$

$$\text{Percentage frequency} = \frac{\text{class frequency}}{\text{total count}} \times 100$$

Relative frequency example

Example

Tacos eaten	Frequency	Relative	Percentage
0 - 4	4	0.5714	57.14 %
5 - 9	2	0.2857	28.57 %
10 - 14	0	0	0 %
15 -20	1	0.1428	14.28 %
Total	7	1	100 %

Cumulative frequency

Cumulative frequency is the frequency for a class and *all previous classes*.

Example

Tacos eaten	Frequency	Cumulative
0 - 4	4	4
5 - 9	2	6
10 - 14	0	6
15 -20	1	7

Outliers

An **outlier** is a data point that is distant from other data or that deviates from an established pattern.

- Outliers can result from chance, an unusual subject, or error.

Example

Tacos eaten $= \{3, 0, 17, 6, 4, 3, 5\}$

Number of tacos eaten	Frequency
0 - 4	4
5 - 9	2
10 - 14	0
15 -20	1

17 tacos eaten in a month is likely an outlier.

Normal distributions

A **normal distribution** can be identified from a frequency table that has the following characteristics:

- The frequencies start low, increase to a high point and then decrease to low frequencies at the end
- The frequencies are approximately symmetric around the high point.

Normal distributions, example

Example

Normal

IQ	Frequency
80 - 89	1
90 - 99	5
100 - 109	11
110 - 119	10
120 - 129	4
130 - 139	2

Not normal

IQ	Frequency
80 - 89	2
90 - 99	13
100 - 109	7
110 - 119	4
120 - 129	3
130 - 139	1

Gaps in frequency tables

A **gap** of frequencies in a table indicates that the data probably come from two different populations.

The converse is not necessarily true. Data from two different populations might not display a gap.

Example

- Pennies made before 1983 are 95% copper and 5% zinc.
- Pennies made after 1983 are 2.5% copper and 97.5% zinc.

Gaps in frequency tables, example

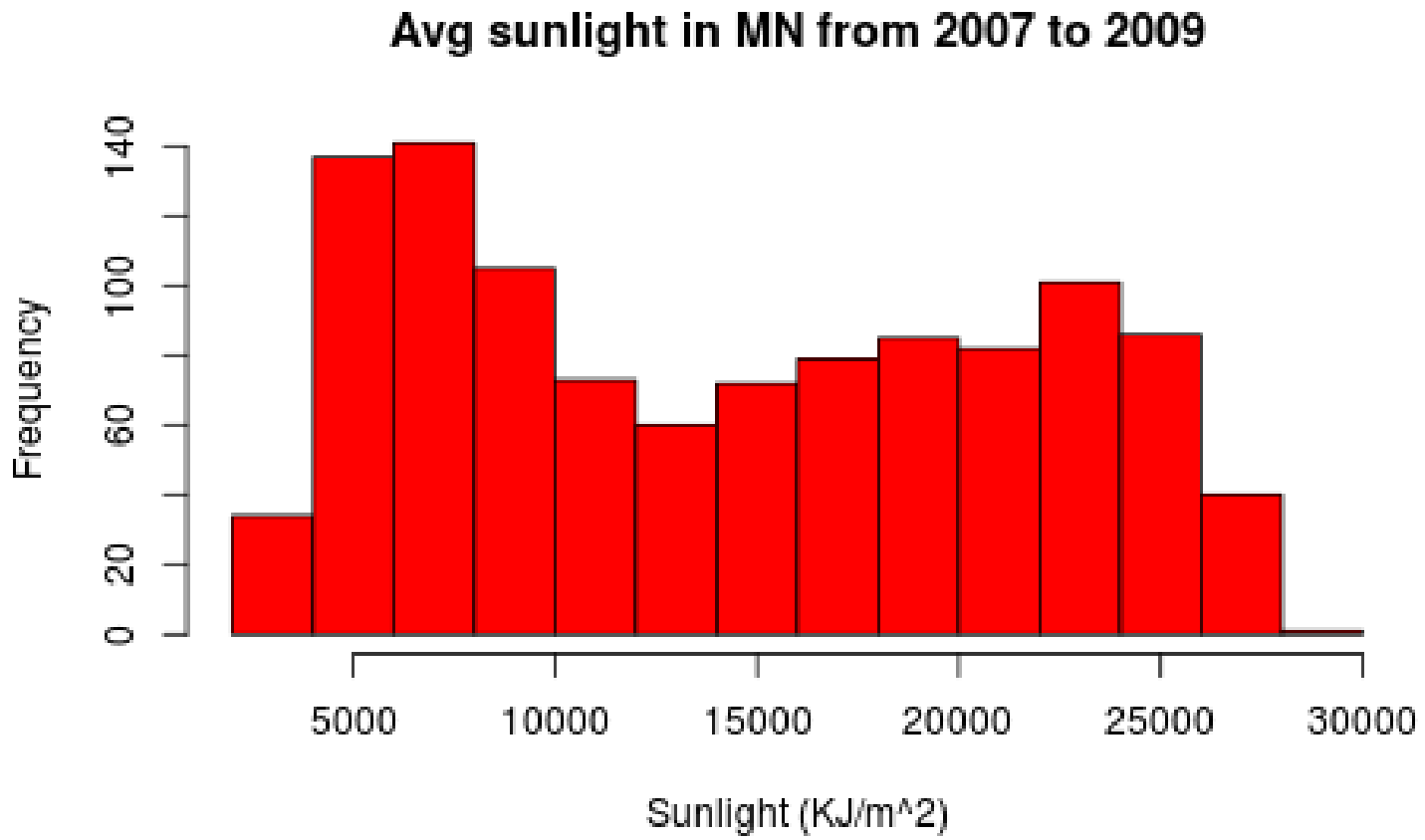
Example, cont.

Weight (g) 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

Section 2.2
Histograms

Histograms

A **histogram** is a graphical representation of a frequency distribution of quantitative data. This allows the distribution of the data to be more easily visualized.



Properties of histograms

- A graph of bars of equal width drawn adjacent to each other.
- The horizontal scale (x-axis) represents values of the quantitative data. Each bar represents a class, or range of values, from a frequency table.
- The vertical scale (y-axis) represents frequency (counts), or proportions (relative frequency) or percentages (percentage frequency).
- The number of bars is largely an aesthetic choice. There should be enough bars to adequately show the shape of the distribution, but too many can make a "busy" graph that's hard to read. Most software will automatically choose the number of bars.

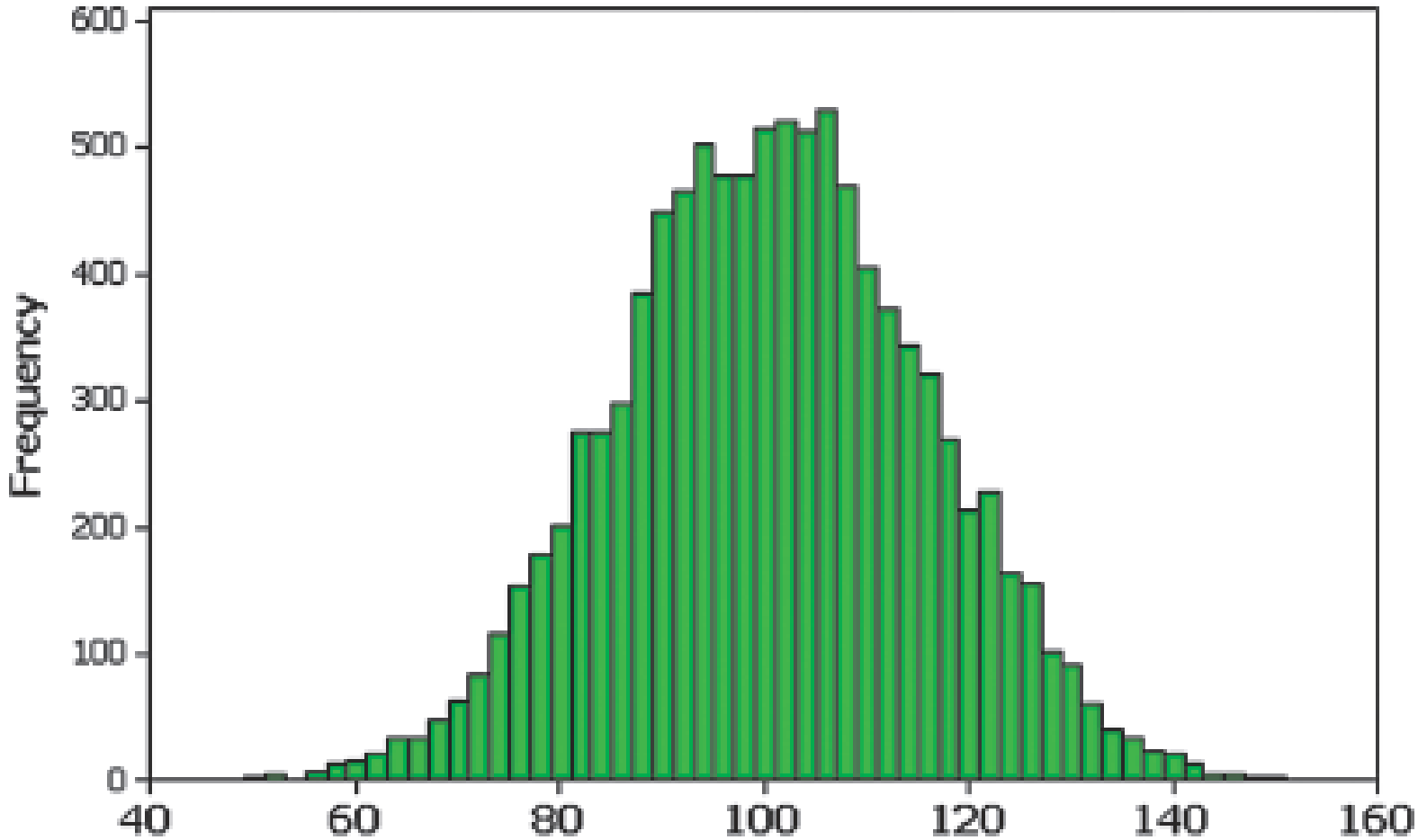
Histograms and normal distributions

Recall, a **normal distribution** can be identified from a frequency table that has the following characteristics:

- The frequencies start low, increase to a high point and then decrease to low frequencies at the end
- The frequencies are approximately symmetric around the high point.

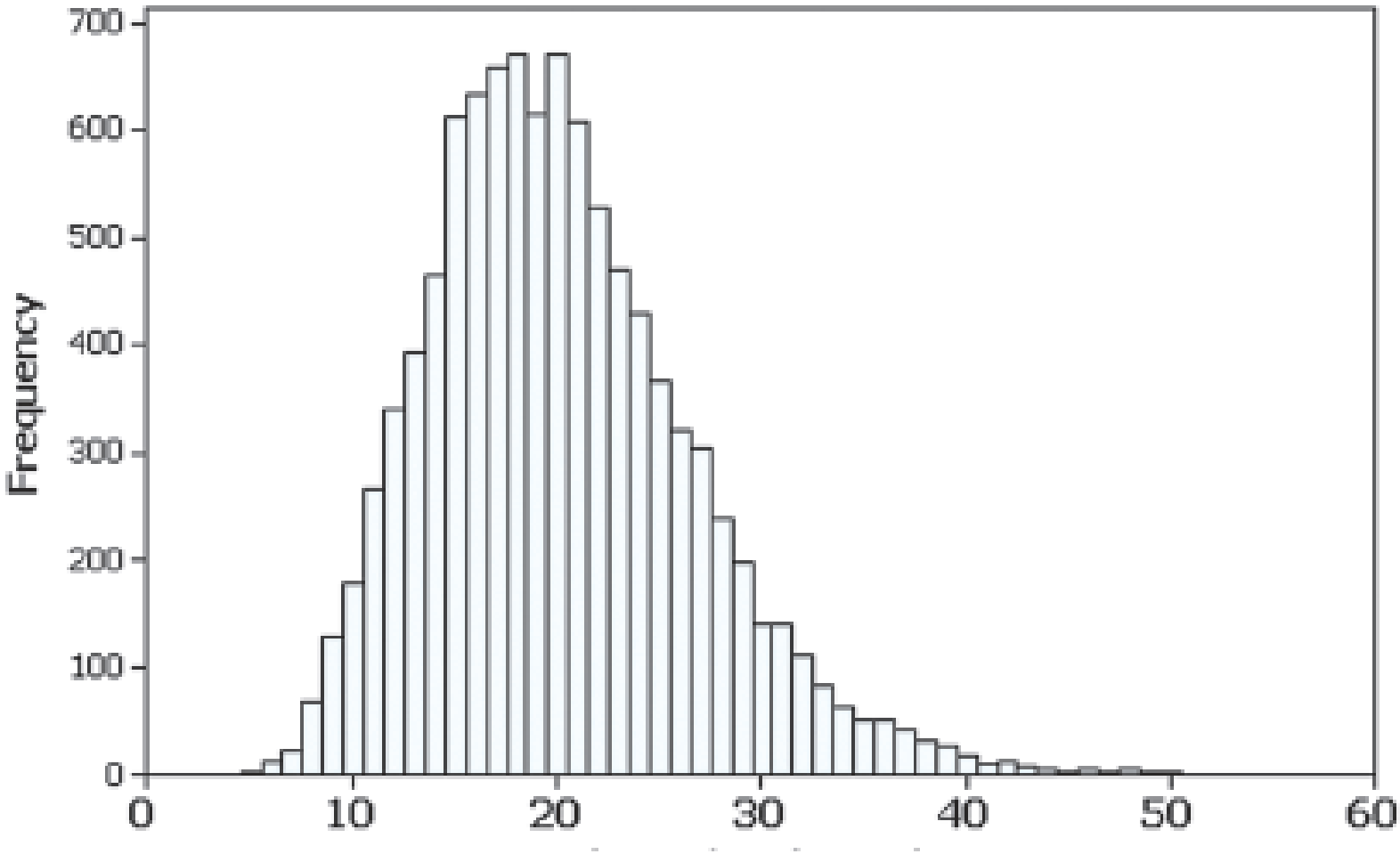
Graphically, normal distributions are commonly known as "bell curves". Histograms can be used to recognize when data follows a normal distribution.

Histograms and normal distributions, examples



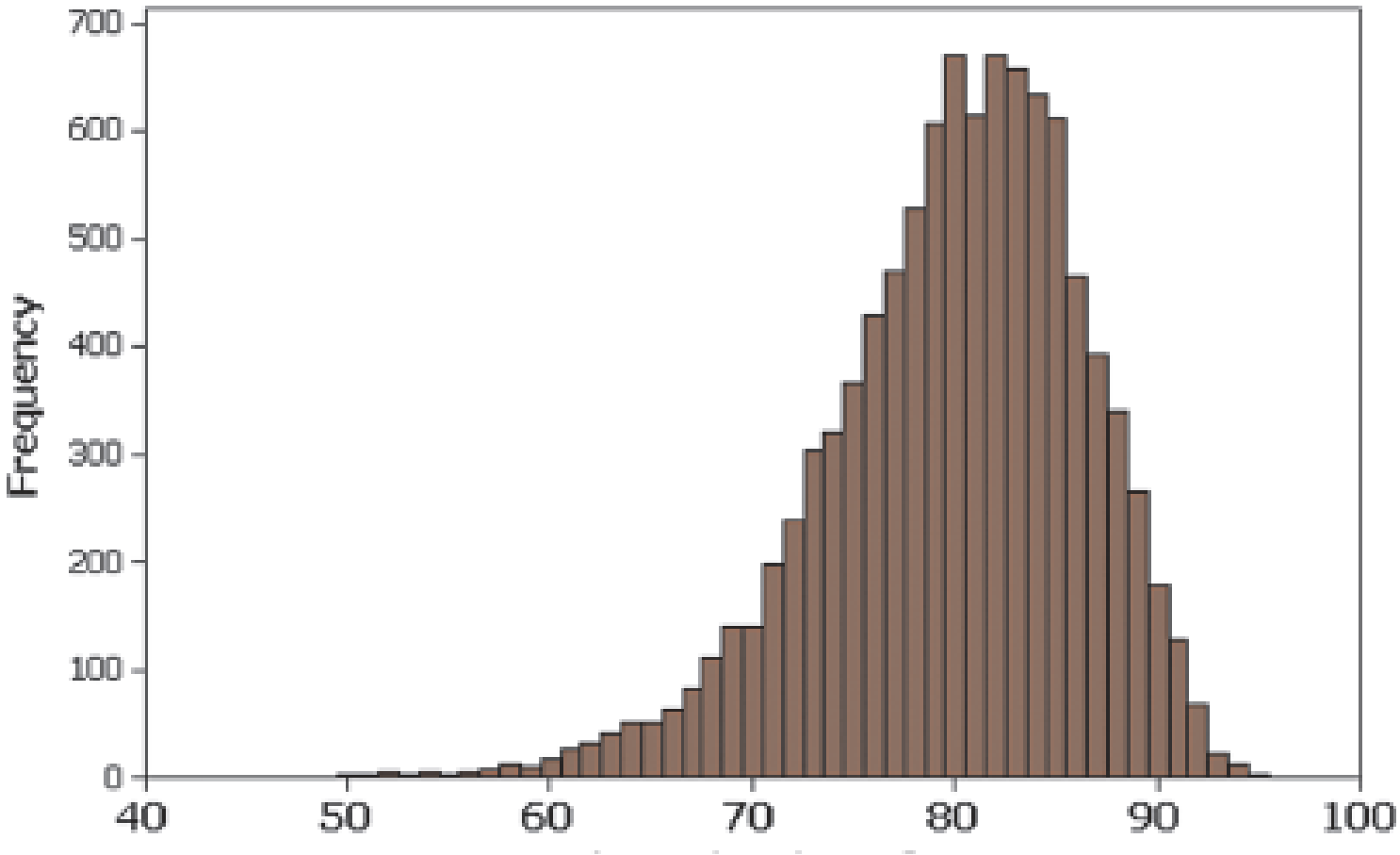
Normal

Histograms and normal distributions, examples



Right skewed

Histograms and normal distributions, examples



Left skewed

Histograms in StatCrunch

- Graph $\rightarrow$ Histogram
- Select column that contains data for histogram
- Optional: Select type of histogram. This will adjust y-axis scale.
- Optional: Set "Bin: Width". This will determine number of bars displayed.
- Click "Compute!"

Note

StatCrunch expects raw data for generating histograms. It won't work with data in frequency tables. To approximate a histogram using a frequency table, use a bar graph (see next section).

Section 2.3
Graphs that Enlighten and Graphs that Deceive

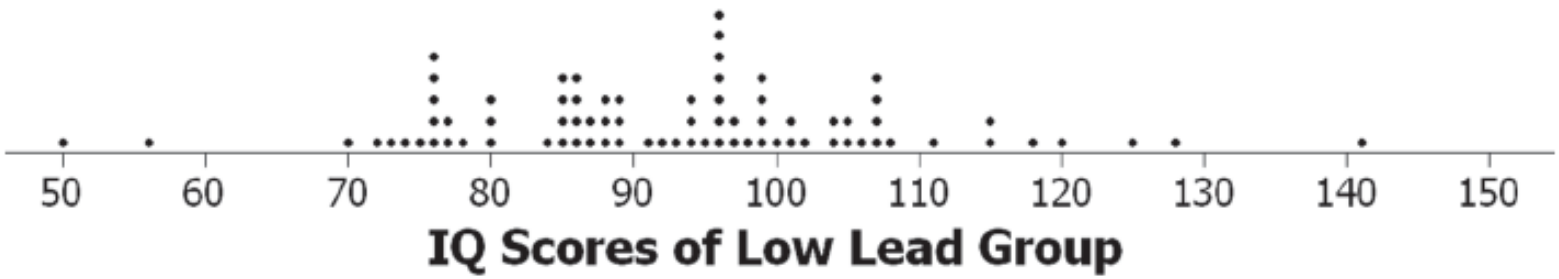
Types of graphs

There are many types of graphs. Deciding which to use depends on the type of data involved and the message to be delivered.

Types of graphs: dotplots

A **dotplot** is similar to a histogram.

- The x-axis represents values of the quantitative data
- Instead of bars, a dot is placed for each instances of a value



Types of graphs: stem-and-leaf plots

A **stem-and-leaf plot** is also used display frequencies of quantitative data

- Each numeric value is separated into two parts, the leftmost digits (the stem) and the last digit (the leaf). For example, 142 $\Rightarrow$ 14 and 2.
- Each stem is arranged vertically on the left side of the graph.
- Every leaf belonging to a stem is listed to the right, in numeric order.

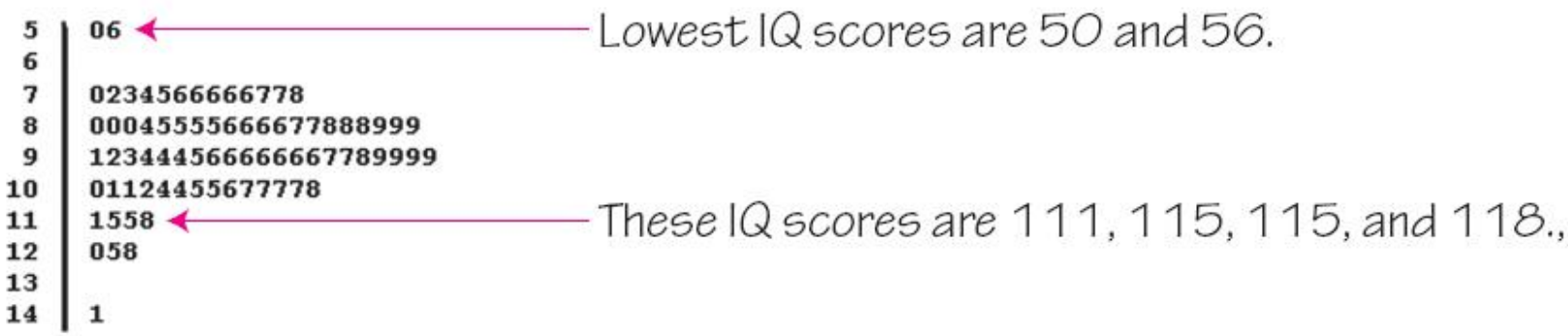
Example

Value	$\Rightarrow$	Stem	Leaf
142		14	2
146		14	6
138		13	8
143		14	3

Stem-and-leaf plot

13	8
14	2 3 6

Stem-and-leaf plot, example

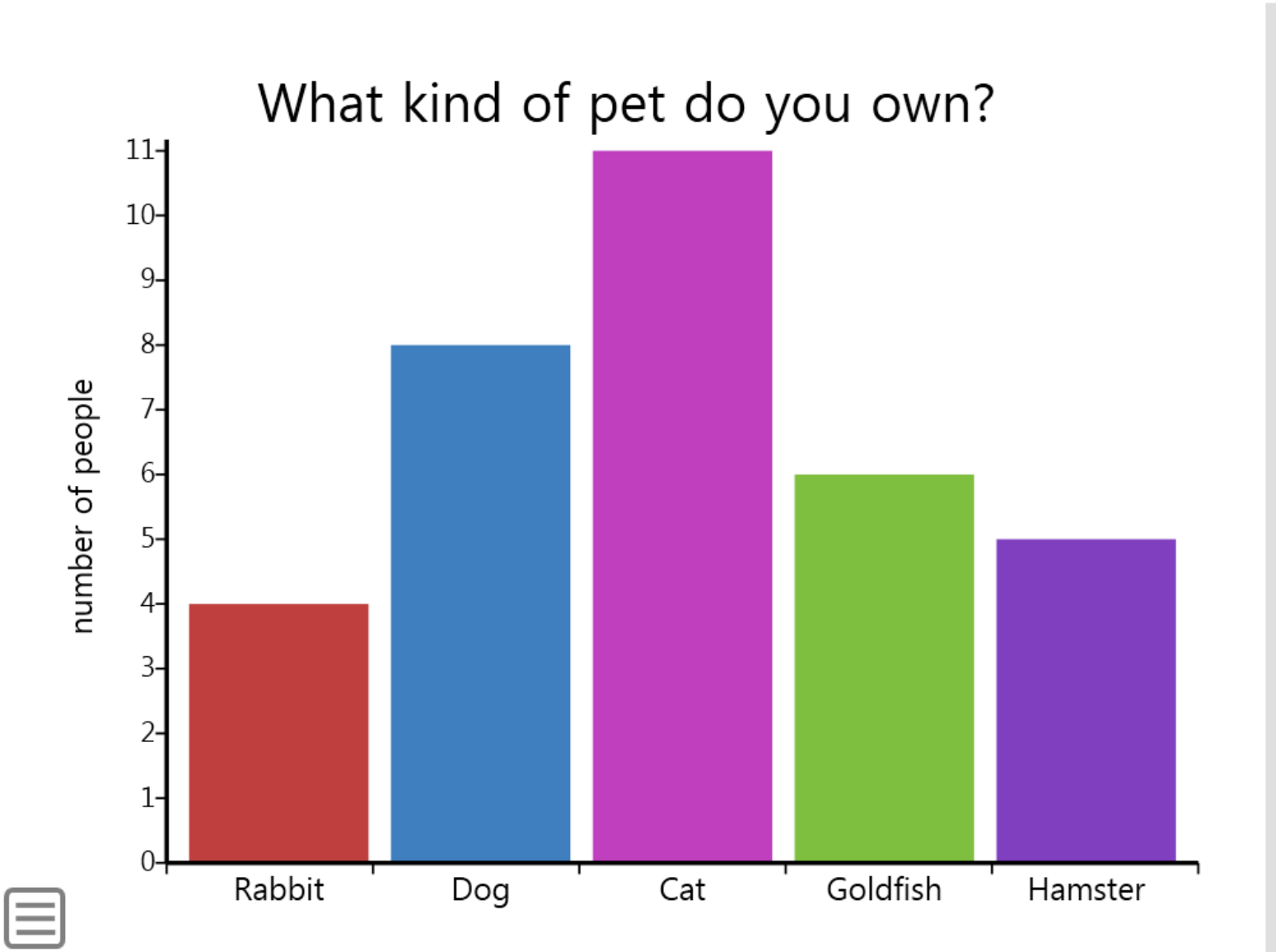


Types of graphs: bar graph

A **bar graph** displays frequencies of categorical data.

- The horizontal scale (x-axis) represents values of the categorical data.
- The vertical scale (y-axis) represents frequencies (or proportions or percentages).
- Often, but not always, bars are drawn with a gap between values.

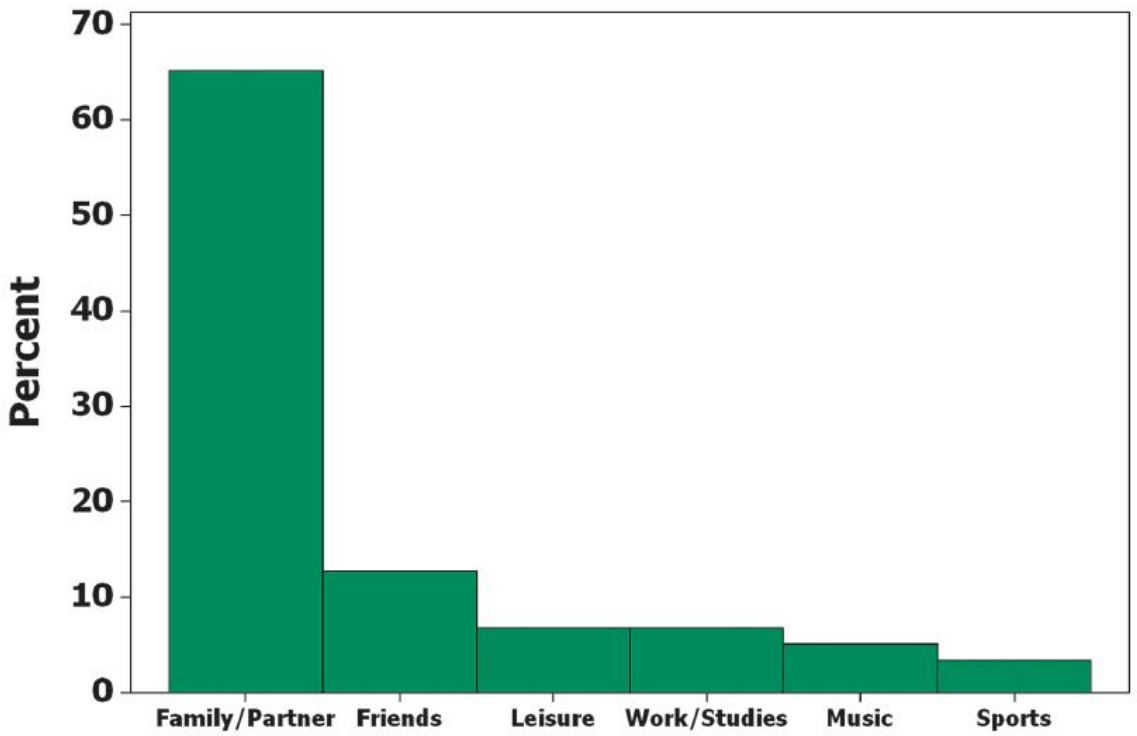
Bar graph, example



Types of graphs: Pareto charts

A **Pareto chart** is very similar to a bar graph, except the bars are arranged from most frequent to least, left to right.

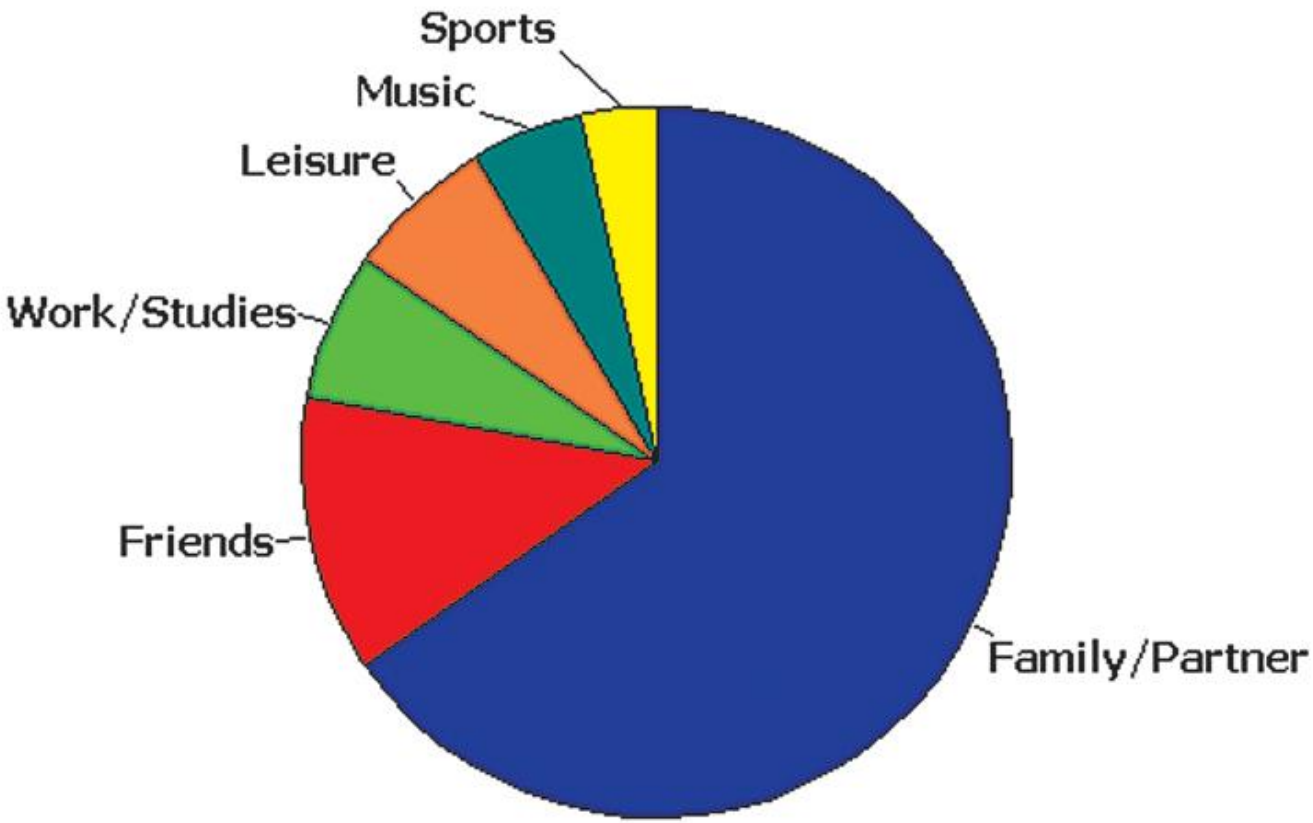
- Can be confusing if used with ordinal data.



Pareto Chart: What Contributes Most to Happiness?

Types of graphs: pie charts

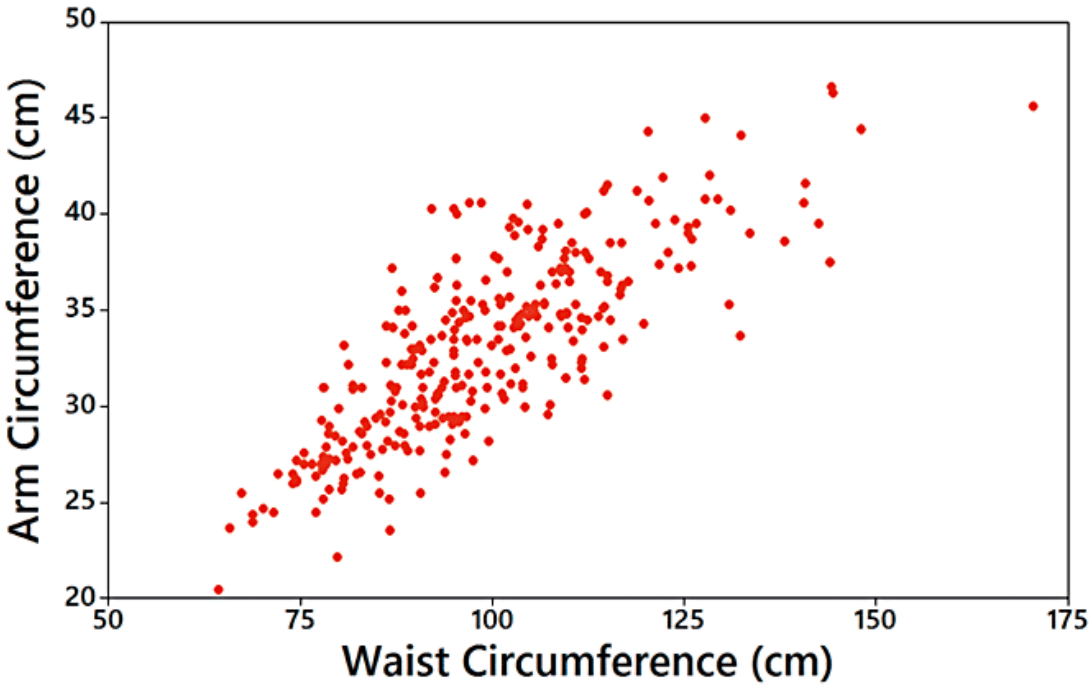
A **pie chart** displays relative frequencies of categorical data as "slices" of a whole circle. The "slices" must be labelled or distinguished by color.



Types of graphs: scatterplots

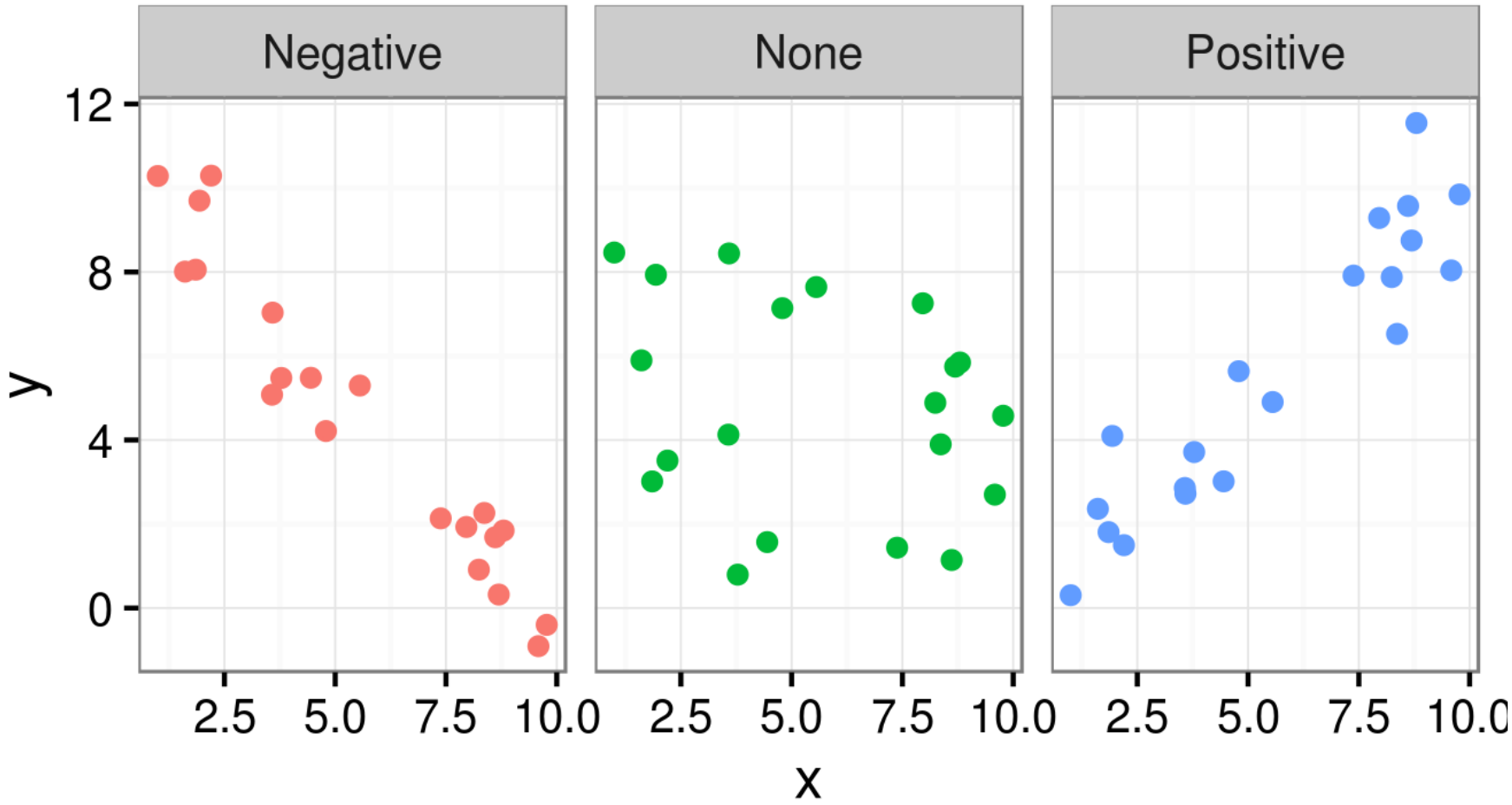
A **scatterplot** displays the relationship between paired quantitative variables.

- The x-axis represents one variable and the y-axis the other.
- A dot (or other symbol) for each data pair is placed at the appropriate x and y values.



Scatterplots and correlation

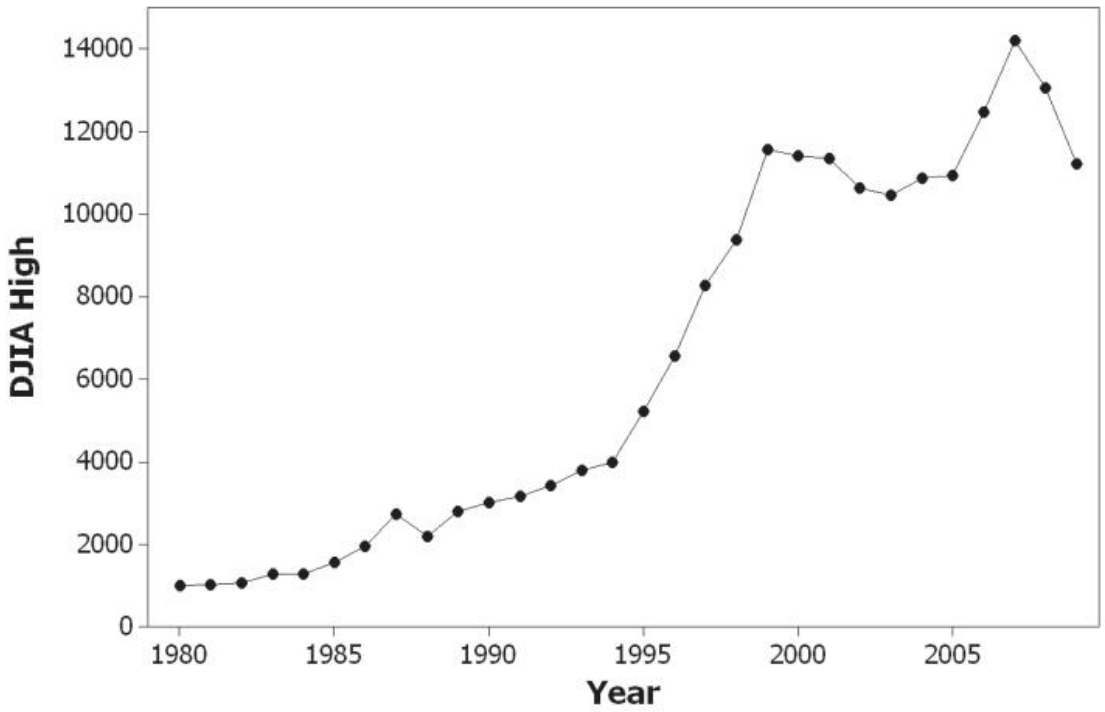
A question that can be answered with a scatterplot is whether there is an association or correlation between variables.



Types of graphs: time series

A graph of paired quantitative data where one variable represents time is called a **time series**. It is much like a scatterplot, except...

- The x-axis always represents the time variable.
- Often a line is drawn between the points.



Graphs in StatCrunch

Dotplots and stem-and-leaf plots

- Graph $\rightarrow$ Dotplot or Stem and Leaf
- Select column that contains the data to be graphed
- Click "Compute!"

Bar plots and pie charts

- Graph $\rightarrow$ Bar Plot or Pie Chart $\rightarrow$ With Summary
- Select the column containing category names
- Select the column containing category counts
- Click "Compute!"

Pareto charts

Follow steps for bar chart, except...

- Under "Order by" select "Count descending"

Graphs in StatCrunch

Scatterplots and time series plots

- Graph $\rightarrow$ Scatter plot
- Select column that contains the data for the x-axis (time variable for time series plots)
- Select column that contains the data for the y-axis
- For time series plots, under "Display" select lines (shift-click to select both points and lines)
- Click "Compute!"

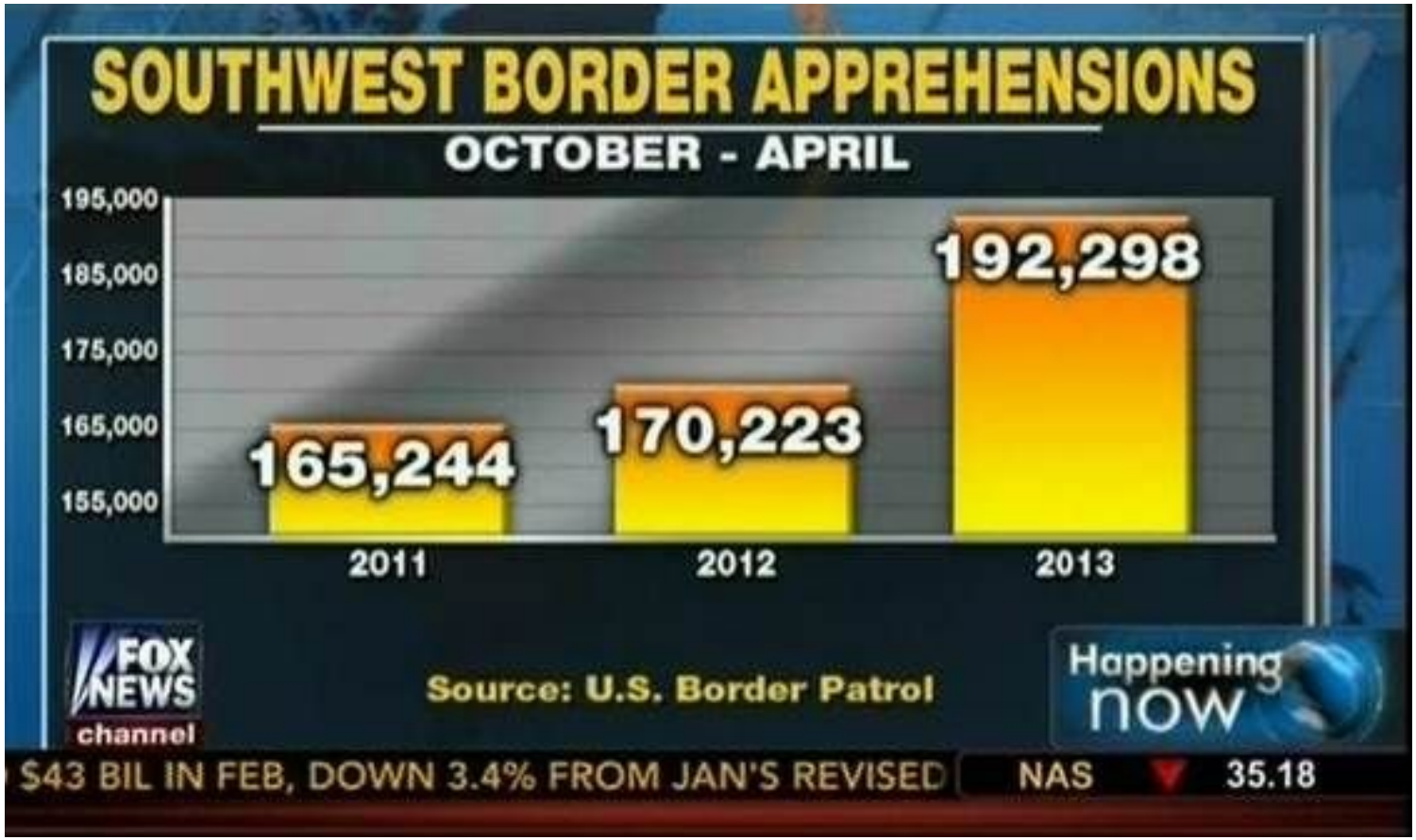
Graphs that deceive

There are two types of bad graphs:

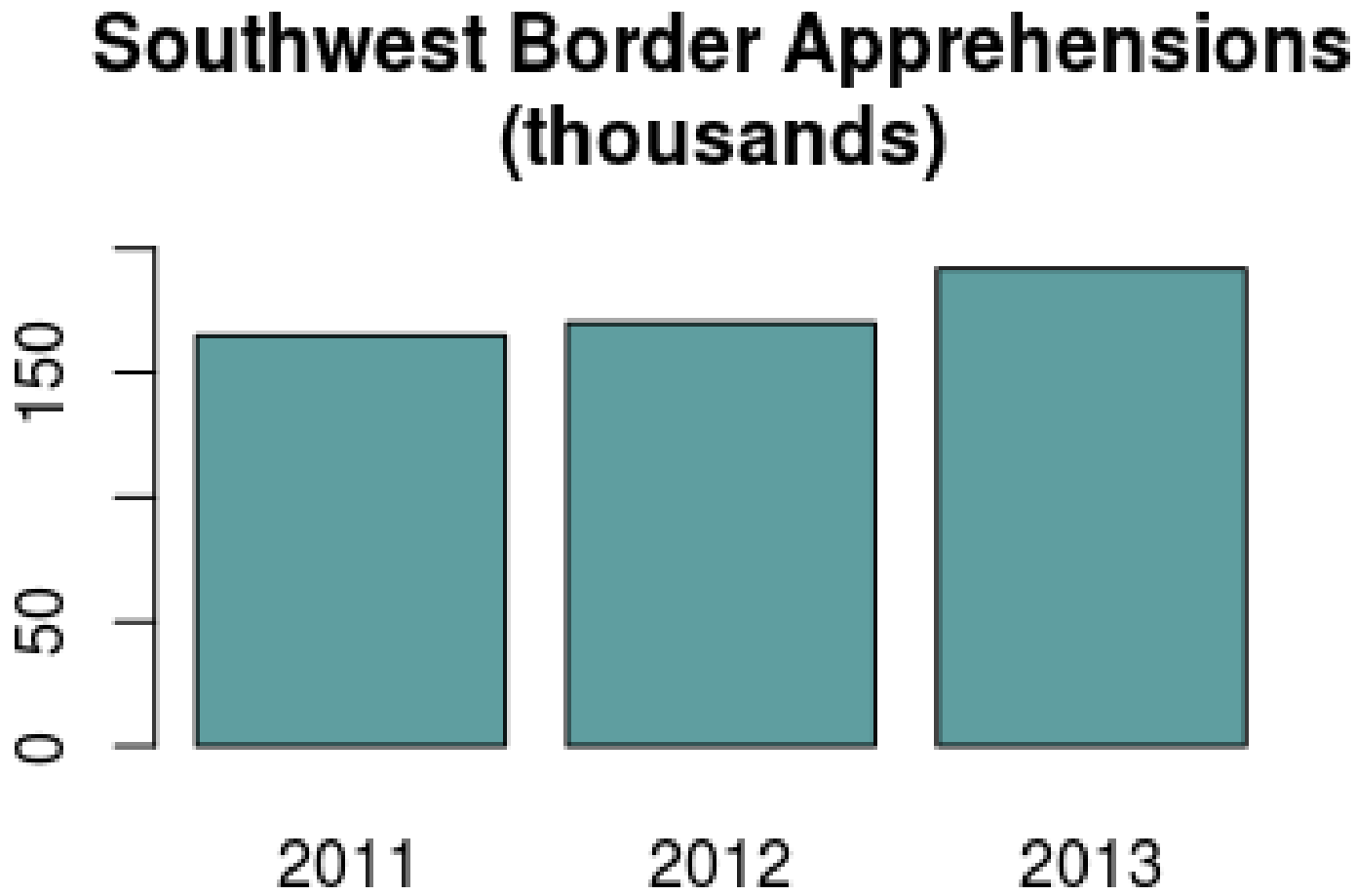
- Sometimes a graph is factually incorrect, whether because of errors in the data or a mistake in creating the graph. This is often difficult to detect without access to the original data.
- Sometimes graphs are technically correct, but designed to give a false impression of the data. Part of being a critical consumer of statistics is learning to recognize these misleading graphs.

Misleading graphs: non-zero axis

A **non-zero axis** is when one of the axis has a scale which does not include zero. This can make the relative sizes of the graph items to be distorted, especially in histograms or bar graphs.



Non-zero axis, fixed



Misleading graphs: pictographs

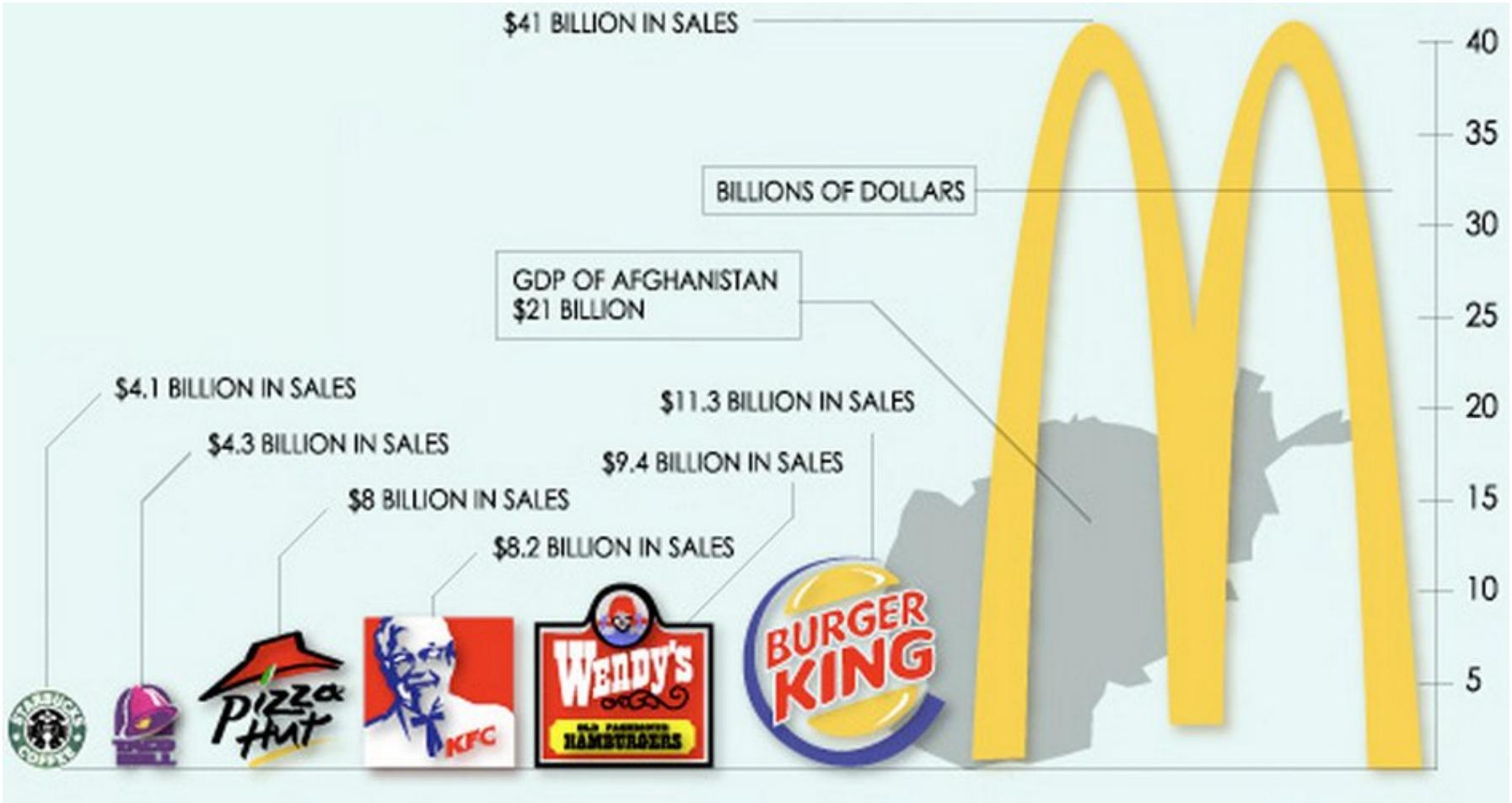
A **pictograph** uses pictures or 3D objects to represent size, rather than simple bars or points. This can also distort relative sizes.

Example

Suppose we wanted to graph the difference in sales between two oil companies, one of which is has twice the sales as the other. If we created a pictograph, we would draw the height of the larger sales twice as tall as the other.

- If we used a pictures, such as a company logos, the larger would have 4 times the area.
- If we used a 3D object, such as an oil barrel, the larger would have 8 times the volume.

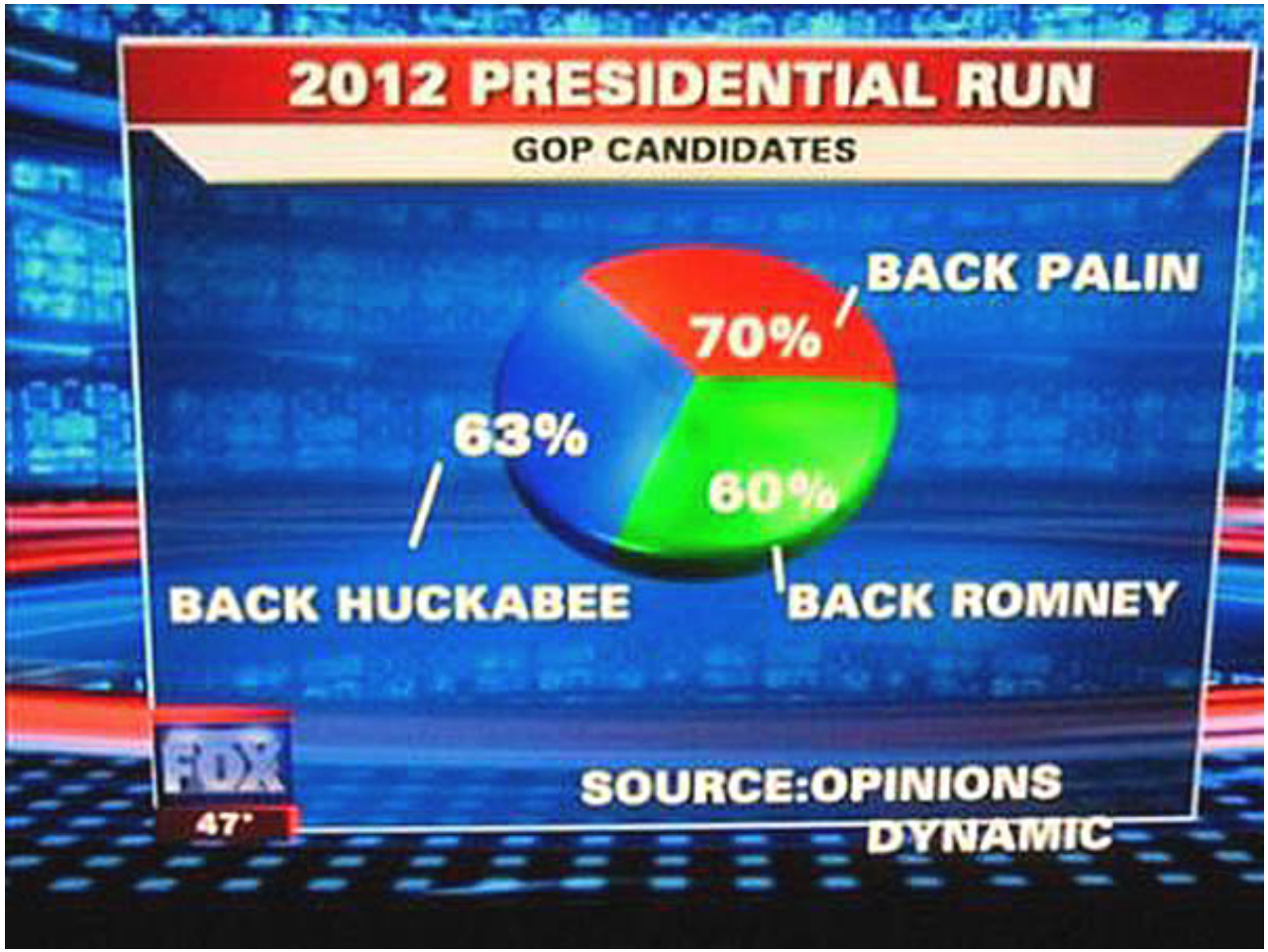
Pictograph, example



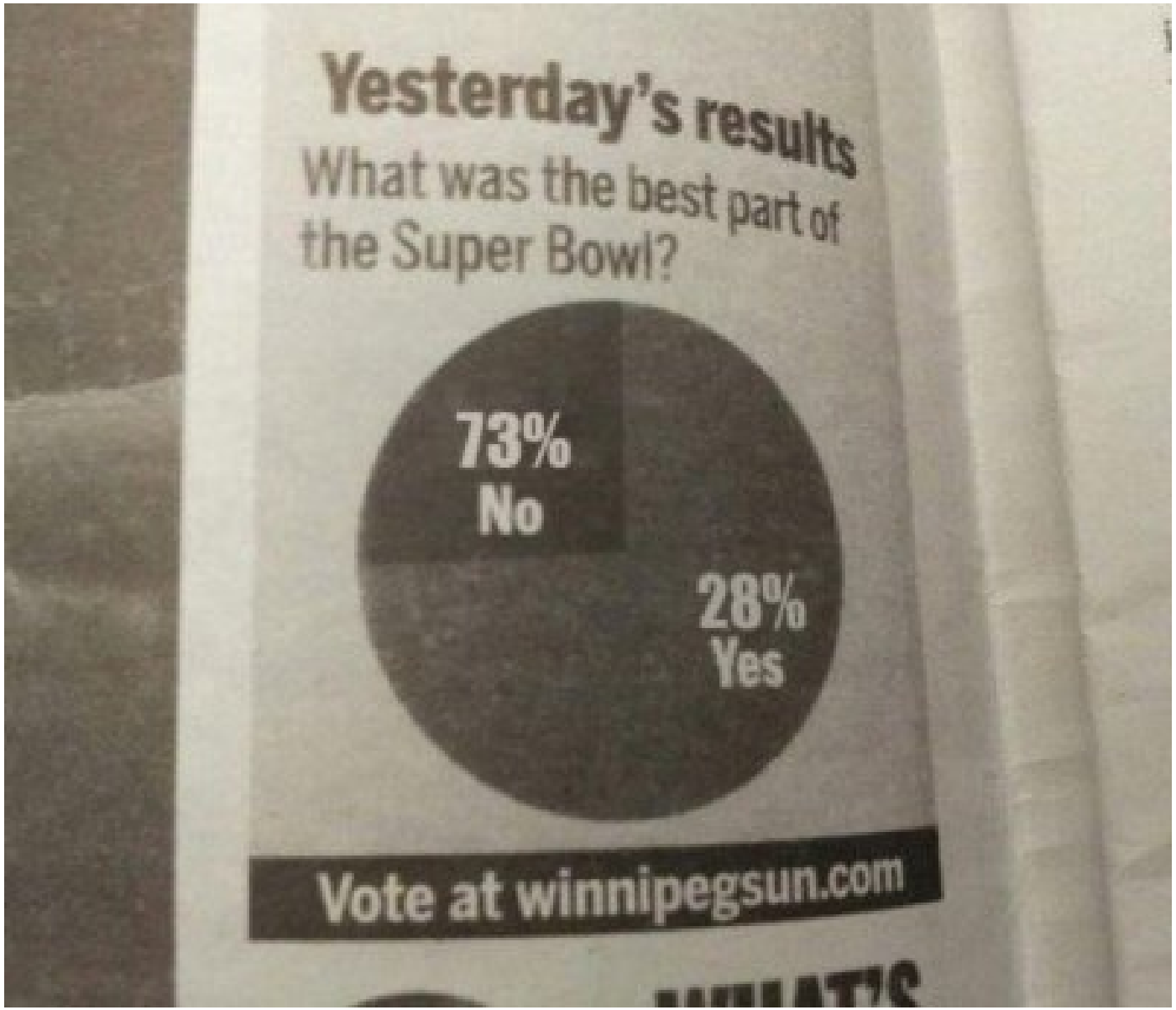
Note that KFC has twice the sales of Starbucks and McDonald's is about 4 times Burger King, but both differences appear much greater.

Misleading graphs: pie chart abuse

Since pie charts represent portions of a whole, the slices should always add up to 100%.



No. Just no.


Yesterday's results
What was the best part of the Super Bowl?
73% No
28% Yes
Vote at winnipegsun.com