Section A01
CDS 101 *Introduction to Computational and Data Sciences*

Class 4: Introduction to data and visualization

May 24, 2018

General

Annoucements

- Reading for next class: *R for Data Science* - chapter 3, section 3.1 through to the end of section 3.6

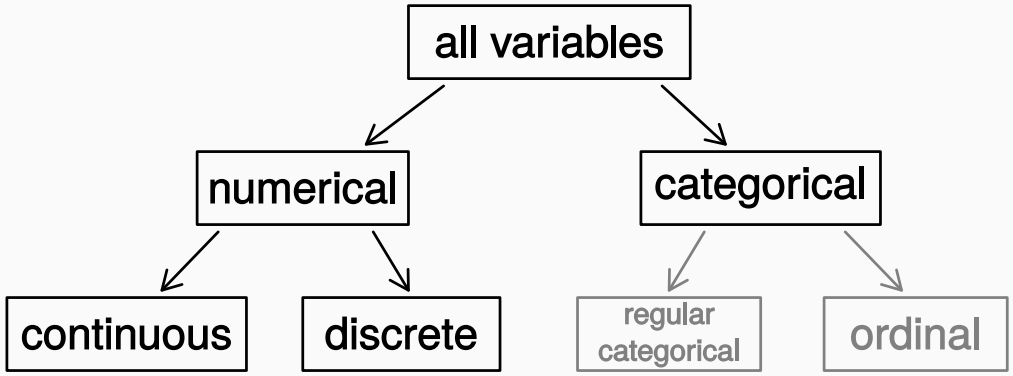
Data basics

Data matrix

Data collected on students in a data science class on a variety of variables:

Stu.	sex	intro_extra	⋯	dread
1	male	extravert	⋯	3
2	female	extravert	⋯	2
3	female	extravert	⋯	4
4	female	extravert	⋯	2
⋮	⋮	⋮	⋮	⋮
21	male	extravert	⋯	3

Types of variables



Types of variables

	sex	sleep	bedtime	countries	dread
1	male	5	12 – 2	13	3
2	female	7	10 – 12	7	2
3	female	5.5	12 – 2	1	4
4	female	7	12 – 2		2
5	female	3	12 – 2	1	3
6	female	3	12 – 2	9	4

Types of variables

	sex	sleep	bedtime	countries	dread
1	male	5	12 – 2	13	3
2	female	7	10 – 12	7	2
3	female	5.5	12 – 2	1	4
4	female	7	12 – 2		2
5	female	3	12 – 2	1	3
6	female	3	12 – 2	9	4

- *sex:* categorical

Types of variables

	sex	sleep	bedtime	countries	dread
1	male	5	12 – 2	13	3
2	female	7	10 – 12	7	2
3	female	5.5	12 – 2	1	4
4	female	7	12 – 2		2
5	female	3	12 – 2	1	3
6	female	3	12 – 2	9	4

- *sex:* categorical
- *sleep:* numerical, continuous

Types of variables

	sex	sleep	bedtime	countries	dread
1	male	5	12 – 2	13	3
2	female	7	10 – 12	7	2
3	female	5.5	12 – 2	1	4
4	female	7	12 – 2		2
5	female	3	12 – 2	1	3
6	female	3	12 – 2	9	4

- *sex:* categorical
- *sleep:* numerical, continuous
- *bedtime:* categorical, ordinal

Types of variables

	sex	sleep	bedtime	countries	dread
1	male	5	12 – 2	13	3
2	female	7	10 – 12	7	2
3	female	5.5	12 – 2	1	4
4	female	7	12 – 2		2
5	female	3	12 – 2	1	3
6	female	3	12 – 2	9	4

- *sex:* categorical
- *sleep:* numerical, continuous
- *bedtime:* categorical, ordinal
- *countries:* numerical, discrete

Types of variables

	sex	sleep	bedtime	countries	dread
1	male	5	12 – 2	13	3
2	female	7	10 – 12	7	2
3	female	5.5	12 – 2	1	4
4	female	7	12 – 2		2
5	female	3	12 – 2	1	3
6	female	3	12 – 2	9	4

- *sex:* categorical
- *sleep:* numerical, continuous
- *bedtime:* categorical, ordinal
- *countries:* numerical, discrete
- *dread:* categorical, ordinal (or numerical)

Practice

What type of variable is a telephone area code?

1. numerical, continuous
2. numerical, discrete
3. categorical
4. categorical, ordinal

Practice

What type of variable is a telephone area code?

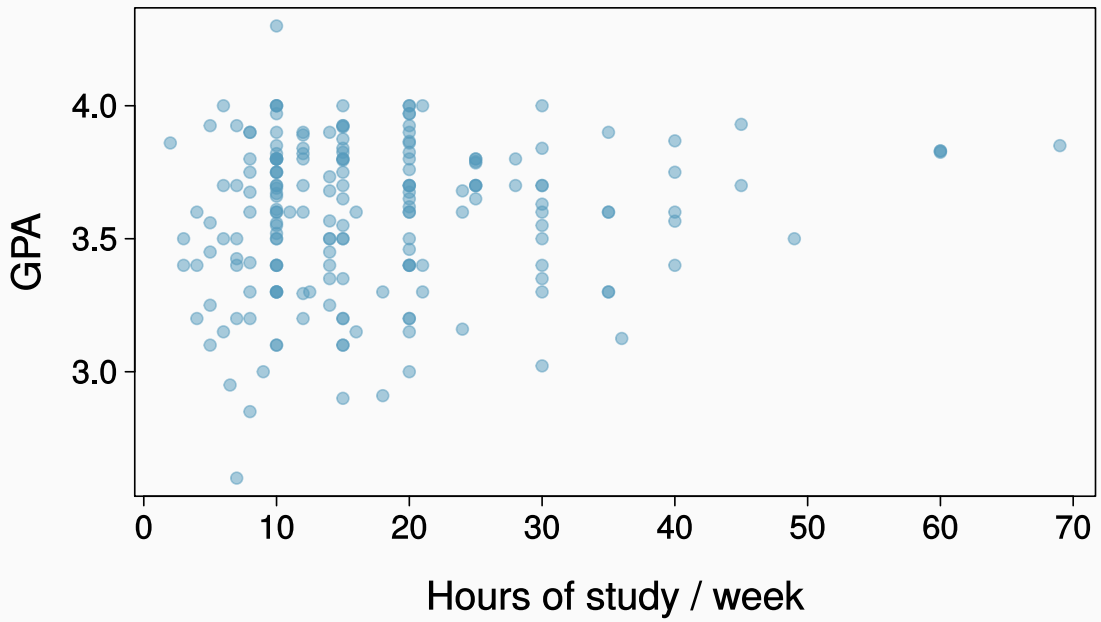
1. numerical, continuous
2. numerical, discrete
3. categorical
4. categorical, ordinal

categorical

Relationships among variables

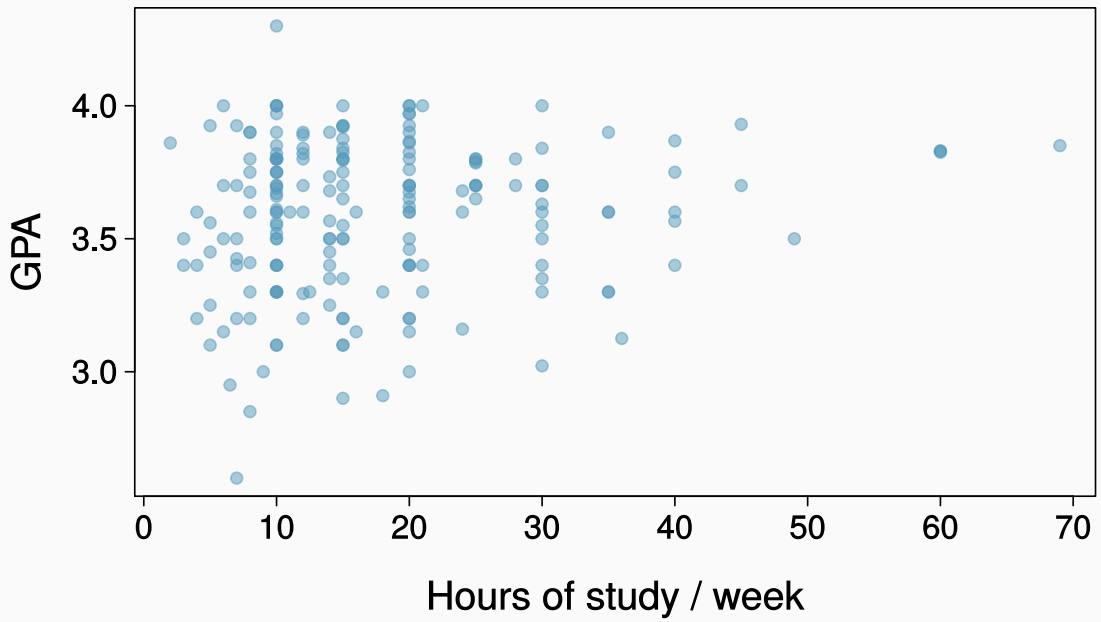
Relationships among variables

Does there appear to be a relationship between GPA and number of hours students study per week?



Relationships among variables

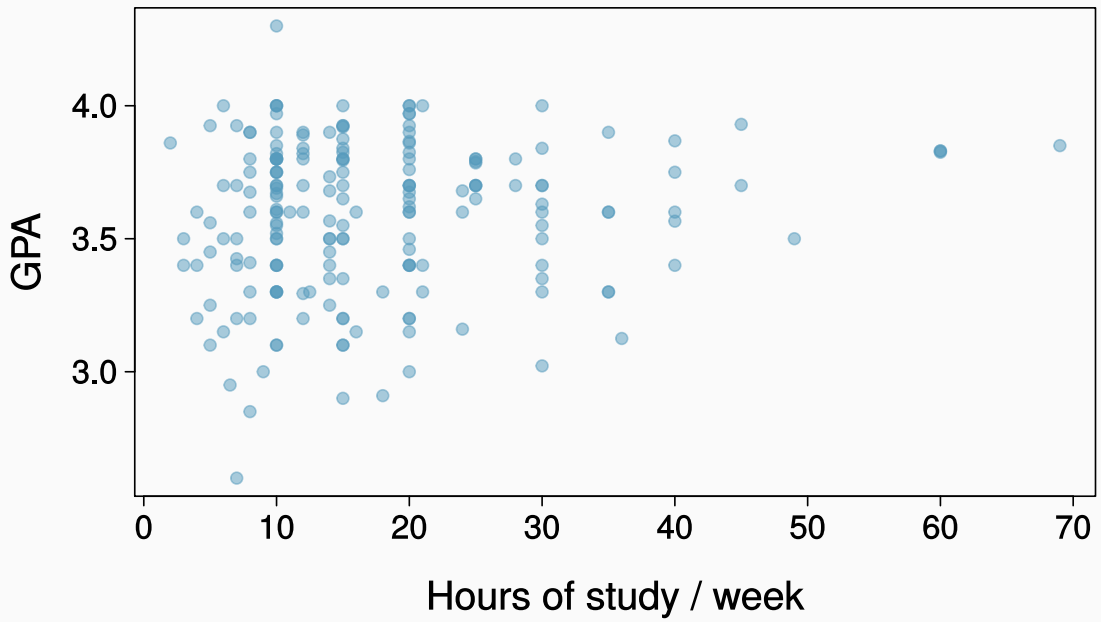
Does there appear to be a relationship between GPA and number of hours students study per week?



Can you spot anything unusual about any of the data points?

Relationships among variables

Does there appear to be a relationship between GPA and number of hours students study per week?



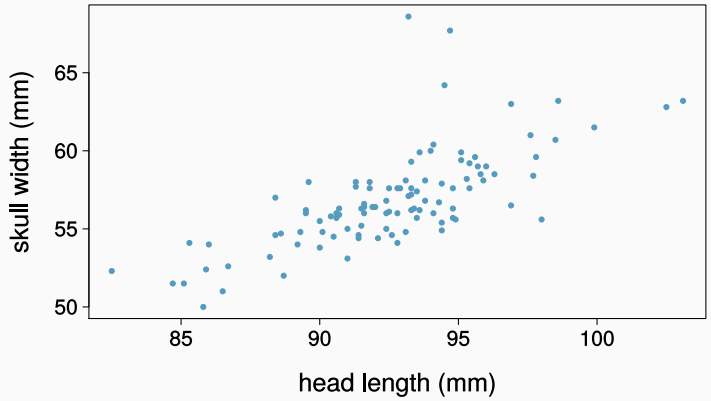
Can you spot anything unusual about any of the data points?

There is one student with GPA (>) 4.0, this is likely a data error.

Associated and independent variables

Practice

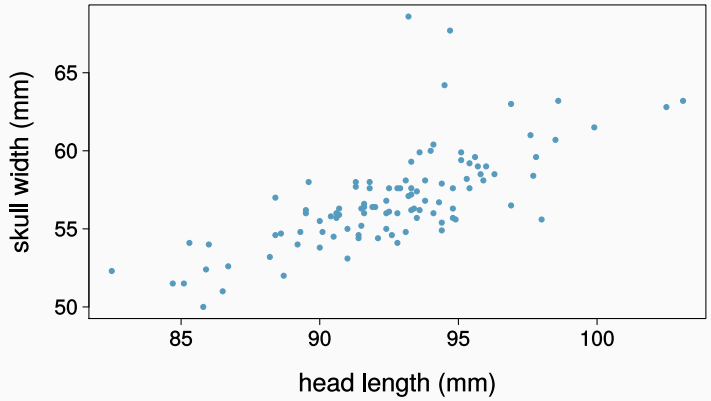
Based on the following scatterplot, which of the following statements is correct about the head and skull lengths of possums?



1. There is no relationship between head length and skull width, i.e. the variables are independent.
2. Head length and skull width are positively associated.}
3. Skull width and head length are negatively associated.
4. A longer head causes the skull to be wider.
5. A wider skull causes the head to be longer.

Practice

Based on the following scatterplot, which of the following statements is correct about the head and skull lengths of possums?



1. There is no relationship between head length and skull width, i.e. the variables are independent.
2. Head length and skull width are positively associated.}
3. Skull width and head length are negatively associated.
4. A longer head causes the skull to be wider.
5. A wider skull causes the head to be longer.

Head length and skull width are positively associated.

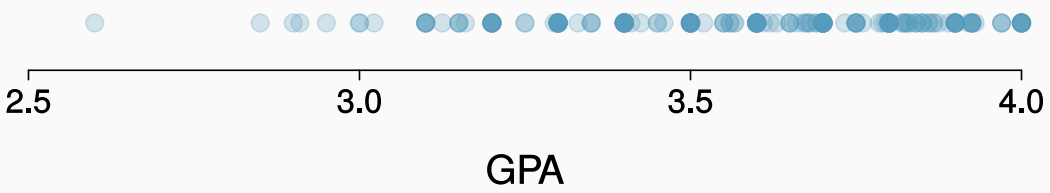
Associated vs. independent

- When two variables show some connection with one another, they are called **associated** variables.
    - Associated variables can also be called **dependent** variables and vice-versa.
- If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be **independent**.

Examining numerical data

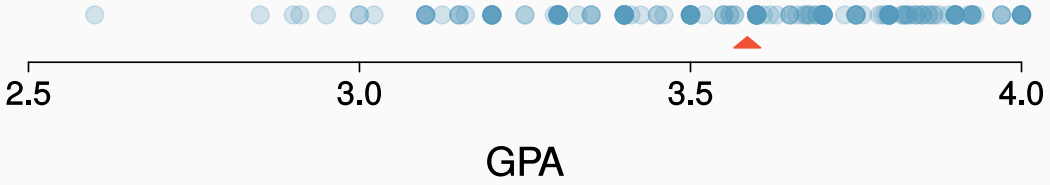
Dot plots

Useful for visualizing one numerical variable. Darker colors represent areas where there are more observations.



How would you describe the distribution of GPAs in this data set? Make sure to say something about the center, shape, and spread of the distribution.

Dot plots & mean



- The **mean**, also called the **average** (marked with a triangle in the above plot), is one way to measure the center of a **distribution** of data.
- The mean GPA is 3.59.

Mean

- The **sample mean**, denoted as $\bar{x}$, can be calculated as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

  where $x_1, x_2, \cdots, x_n$ represent the **n** observed values.

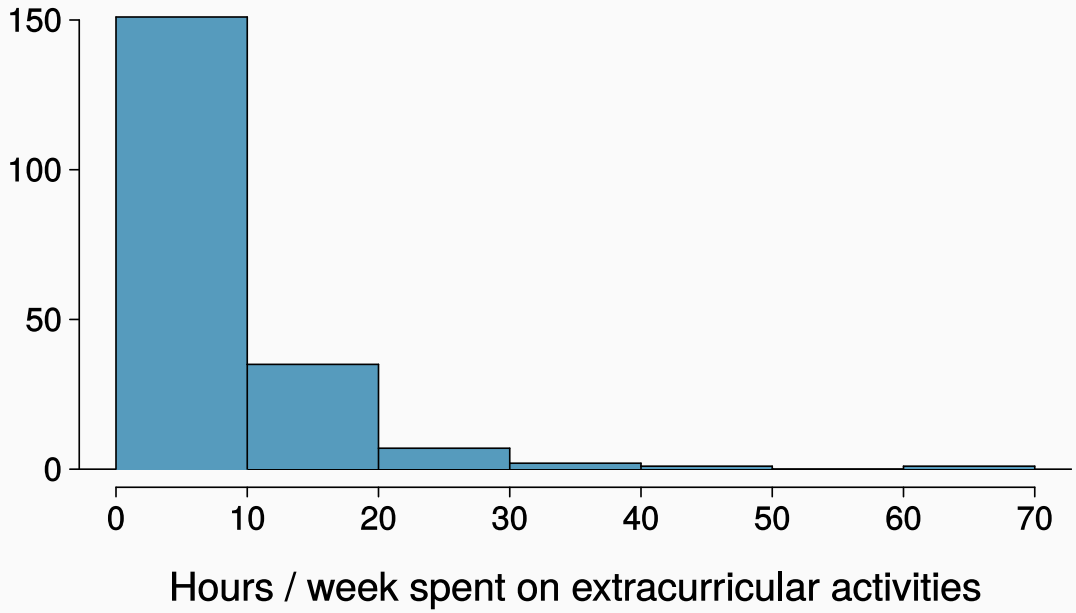
- The **population mean** is also computed the same way but is denoted as μ. It is often not possible to calculate μ since population data are rarely available.

- The sample mean is a **sample statistic**, and serves as a **point estimate** of the population mean. This estimate may not be perfect, but if the sample is good (representative of the population), it is usually a pretty good estimate.

Histograms and shape

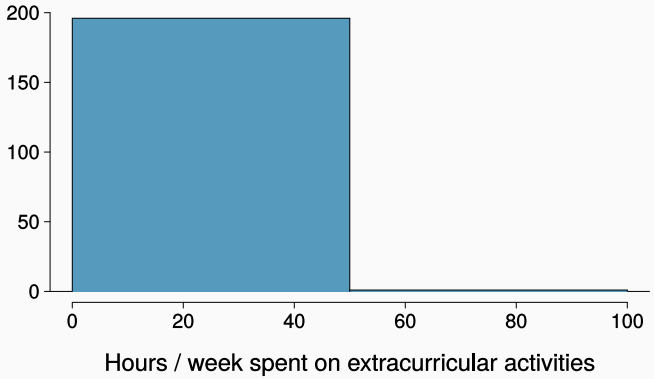
Histograms — Extracurricular hours

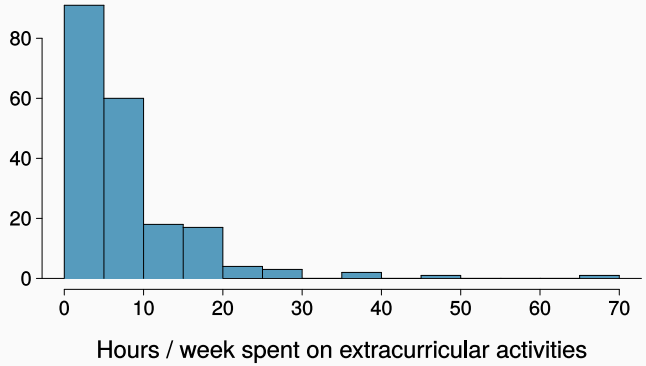
- Histograms provide a view of the **data density**. Higher bars represent where the data are relatively more common.
- Histograms are especially convenient for describing the **shape** of the data distribution.
- The chosen **bin width** can alter the story the histogram is telling.

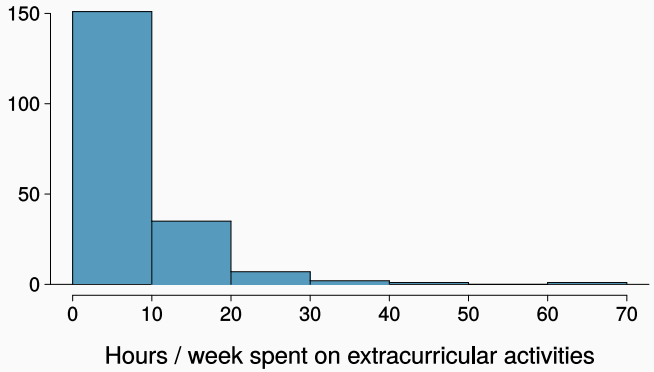


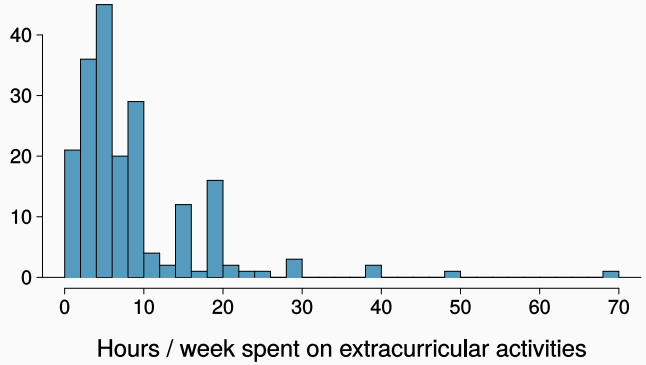
Bin width

Which one(s) of these histograms are useful? Which reveal too much about the data? Which hide too much?



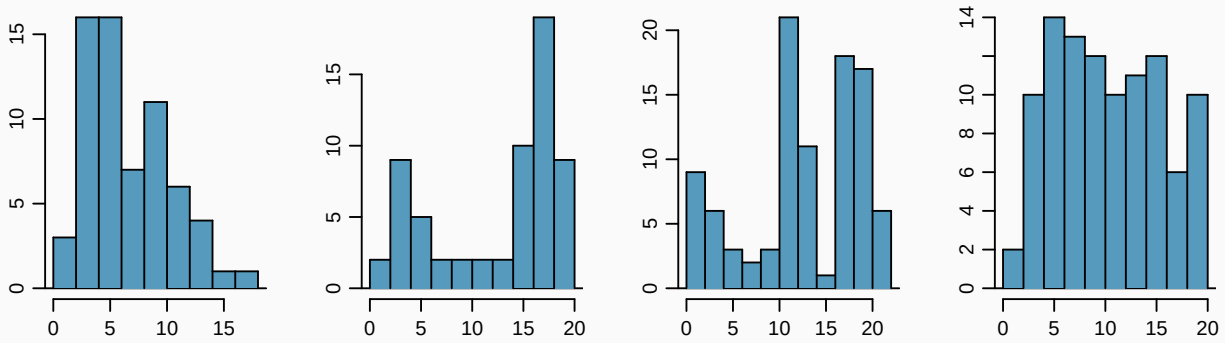






Shape of a distribution: modality

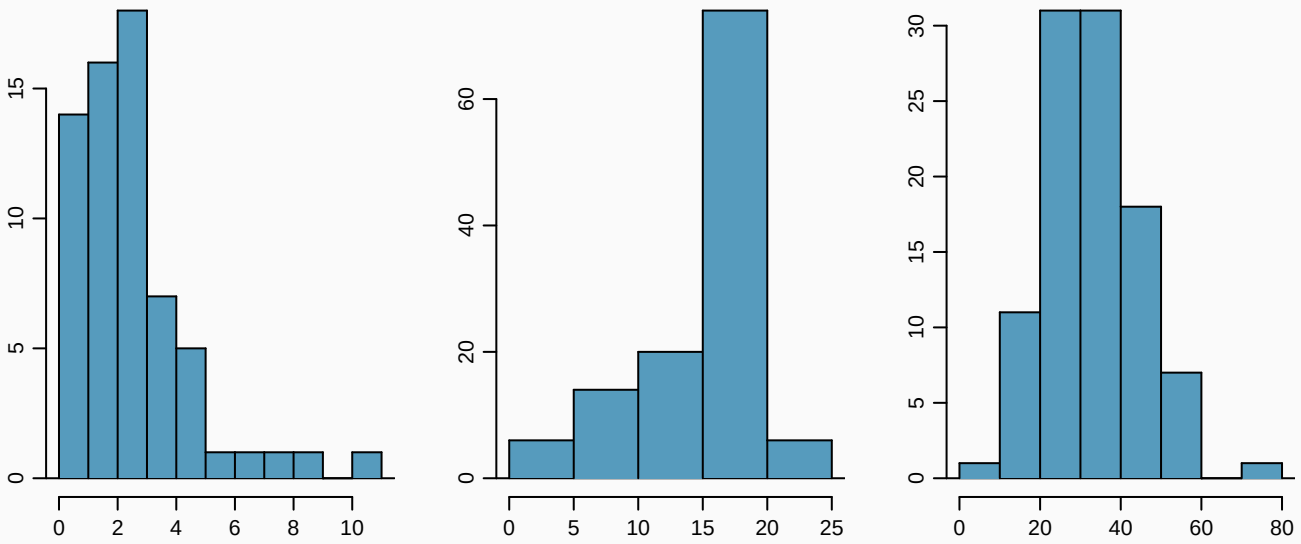
Does the histogram have a single prominent peak (*unimodal*), several prominent peaks (*bimodal/multimodal*), or no apparent peaks (*uniform*)?



Note: In order to determine modality, step back and imagine a smooth curve over the histogram – imagine that the bars are wooden blocks and you drop a limp spaghetti over them, the shape the spaghetti would take could be viewed as a smooth curve.

Shape of a distribution: skewness

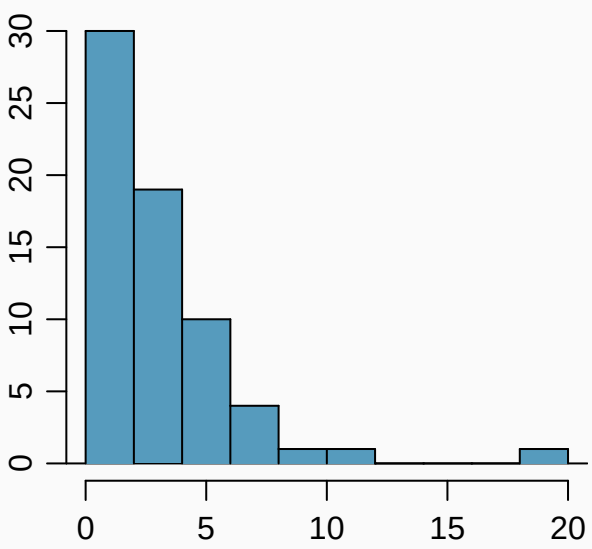
Is the histogram *right skewed*, *left skewed*, or *symmetric*?

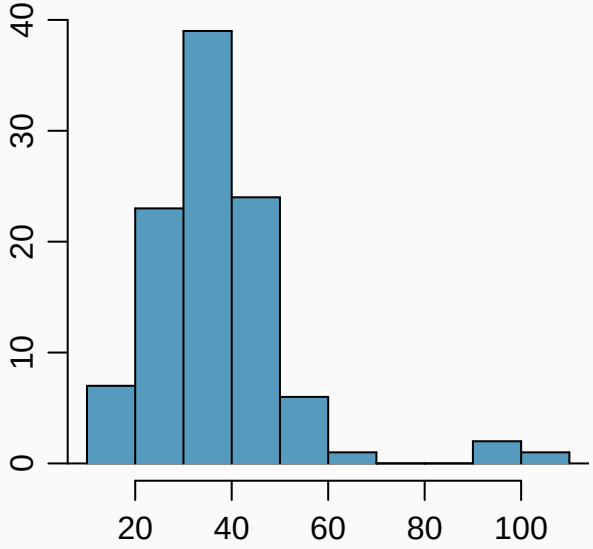


Note: Histograms are said to be skewed to the side of the long tail.

Shape of a distribution: unusual observations

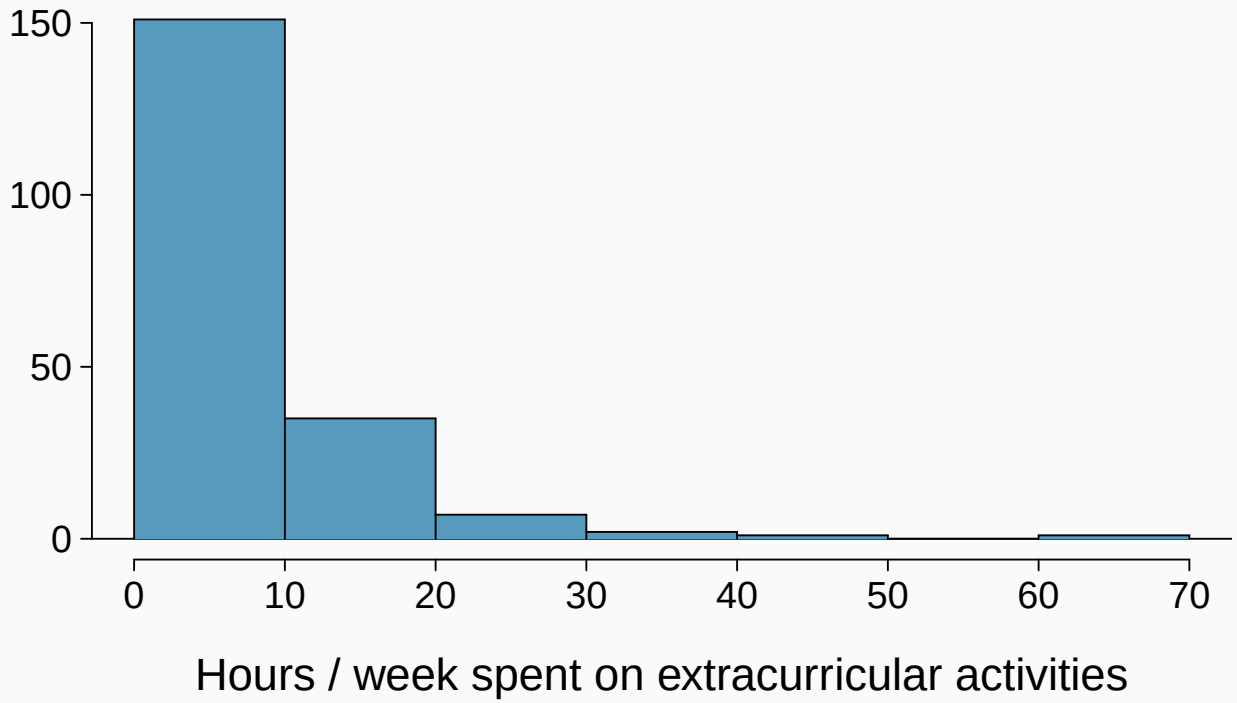
Are there any unusual observations or potential *outliers*?





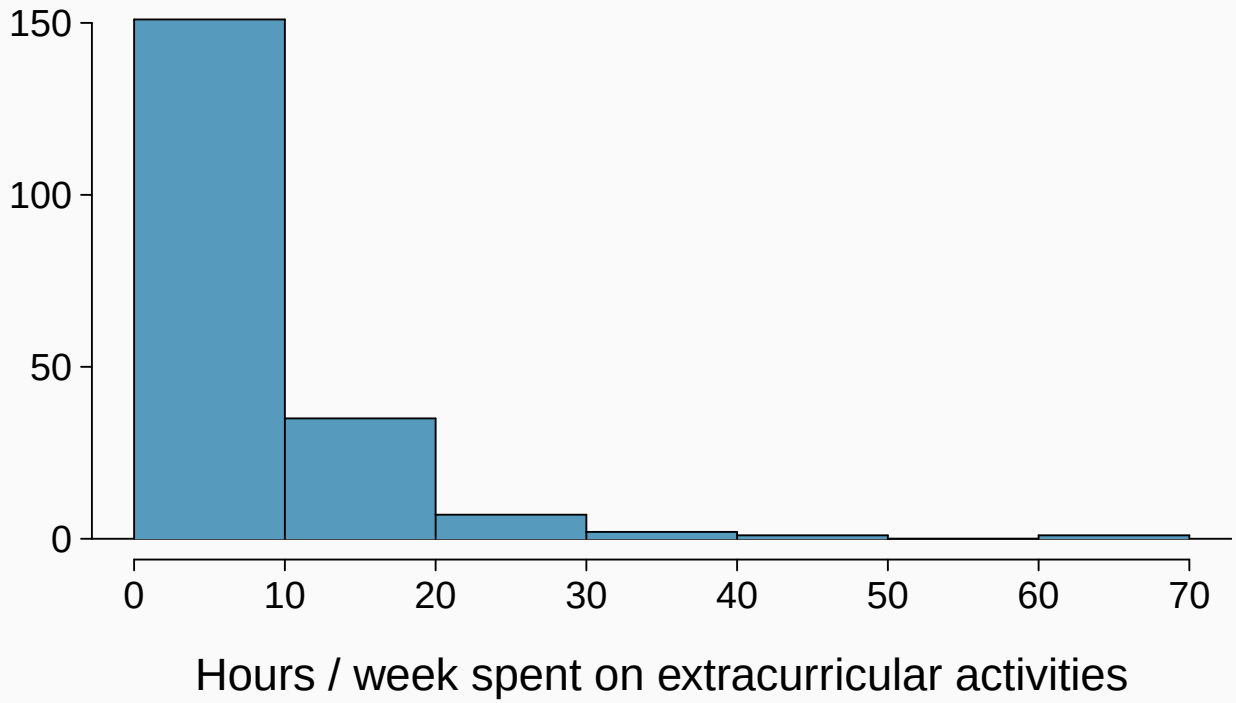
Extracurricular activities

How would you describe the shape of the distribution of hours per week students spend on extracurricular activities?



Extracurricular activities

How would you describe the shape of the distribution of hours per week students spend on extracurricular activities?



Unimodal and right skewed, with a potentially unusual observation at 60 hours/week.

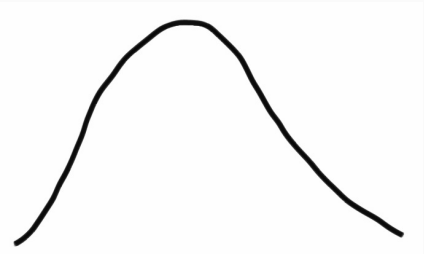
Commonly observed shapes of distributions

- modality

Commonly observed shapes of distributions

- modality

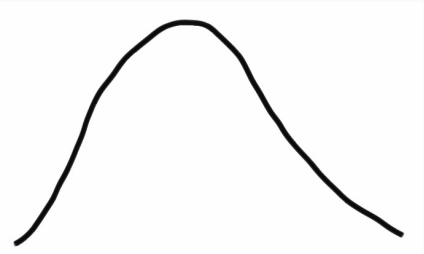
unimodal



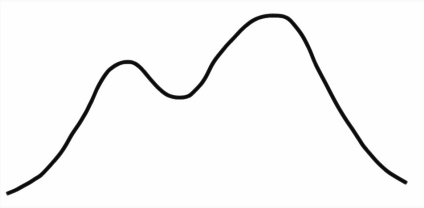
Commonly observed shapes of distributions

- modality

unimodal



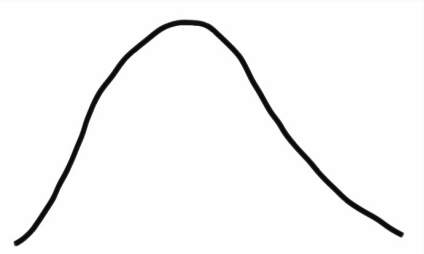
bimodal



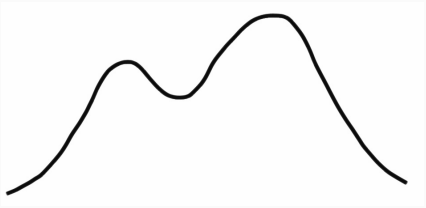
Commonly observed shapes of distributions

- modality

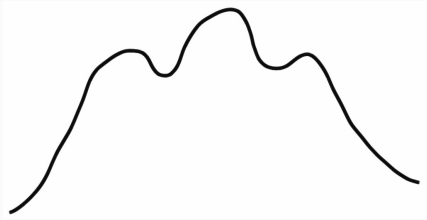
unimodal



bimodal



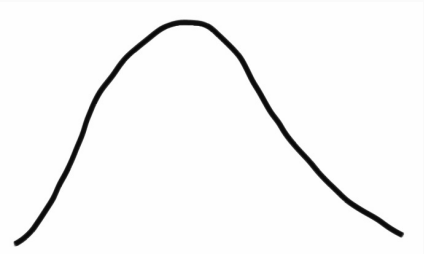
multimodal



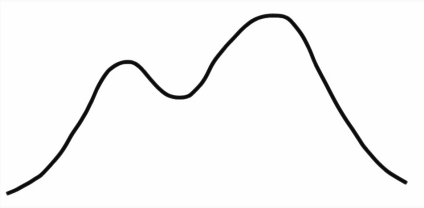
Commonly observed shapes of distributions

- modality

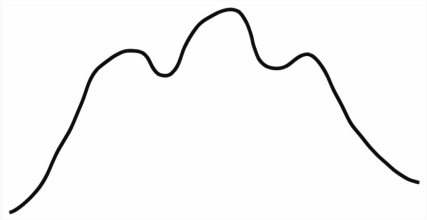
unimodal



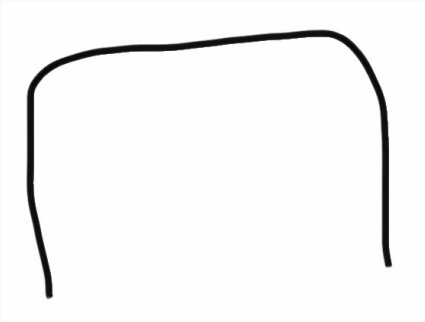
bimodal



multimodal



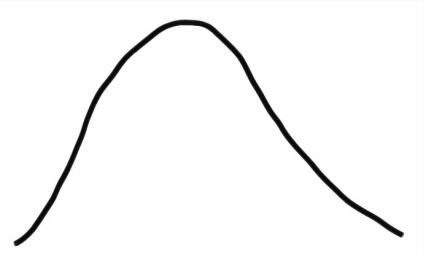
uniform



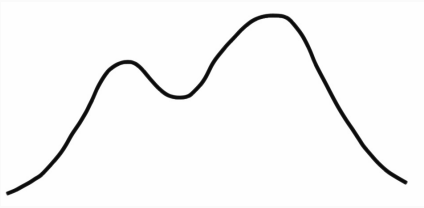
Commonly observed shapes of distributions

- modality

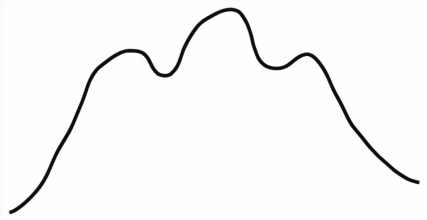
unimodal



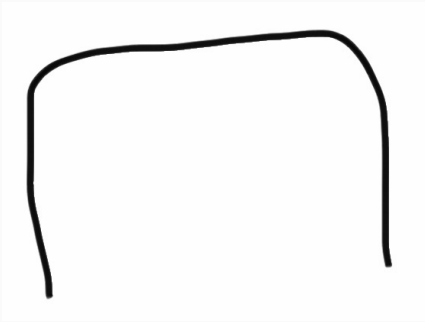
bimodal



multimodal



uniform

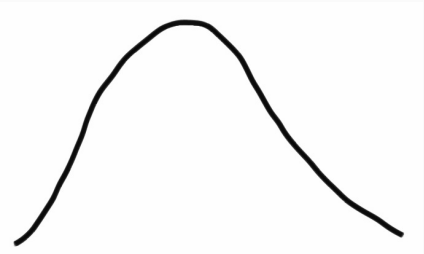


- skewness

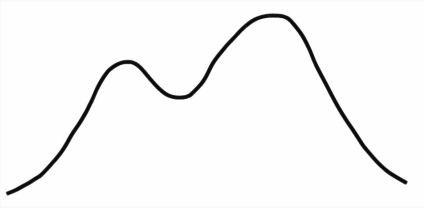
Commonly observed shapes of distributions

- modality

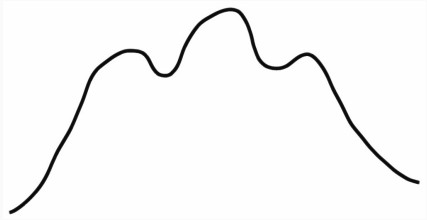
unimodal



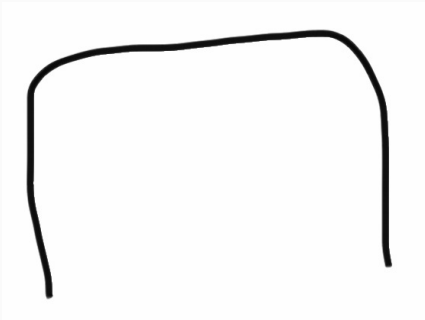
bimodal



multimodal

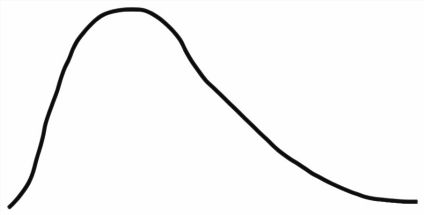


uniform



- skewness

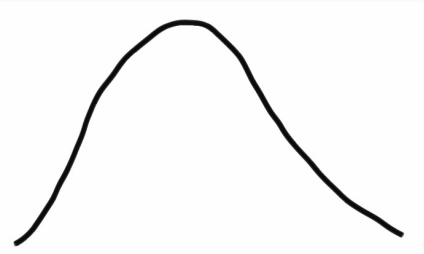
right skew



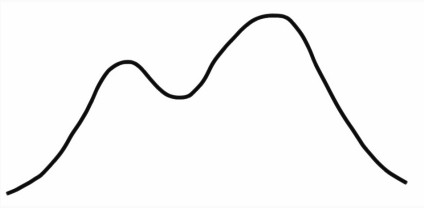
Commonly observed shapes of distributions

- modality

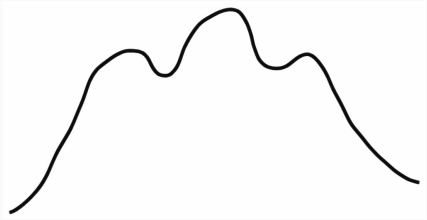
unimodal



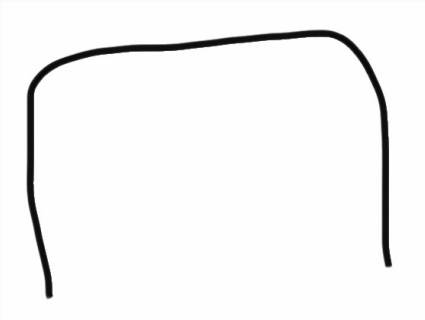
bimodal



multimodal

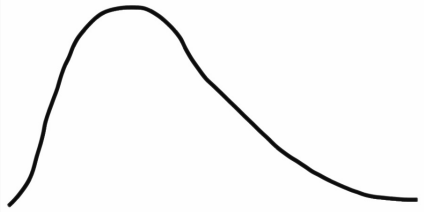


uniform

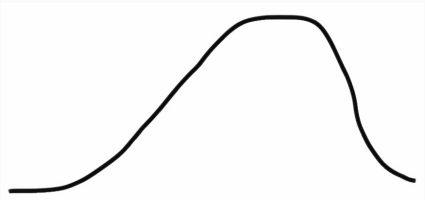


- skewness

right skew



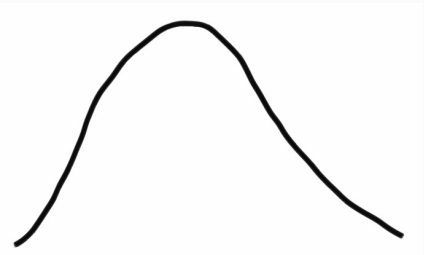
left skew



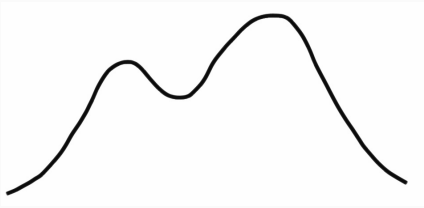
Commonly observed shapes of distributions

- modality

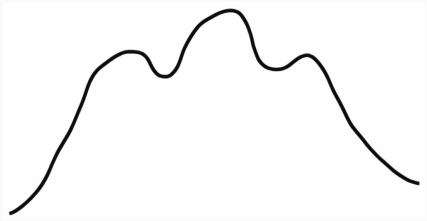
unimodal



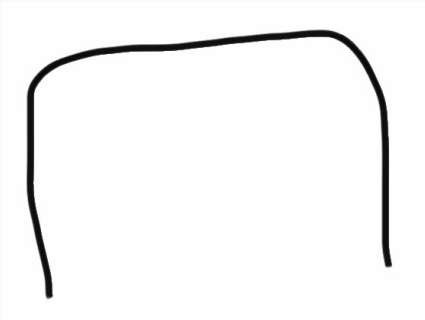
bimodal



multimodal

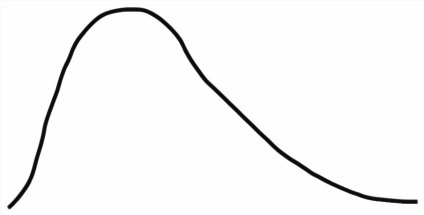


uniform

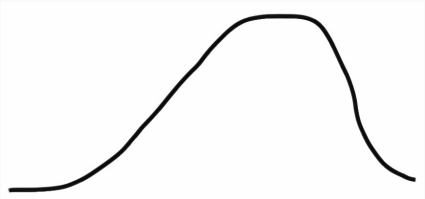


- skewness

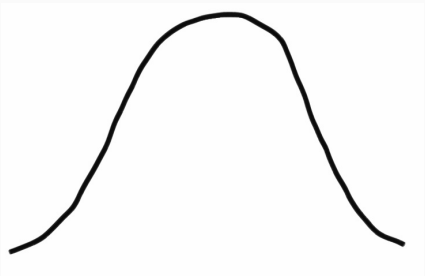
right skew



left skew



symmetric



Practice

Which of these variables do you expect to be uniformly distributed?

(a) weights of adult females

(b) salaries of a random sample of people from North Carolina

(c) house prices

(d) birthdays of classmates (day of the month)

Practice

Which of these variables do you expect to be uniformly distributed?

(a) weights of adult females

(b) salaries of a random sample of people from North Carolina

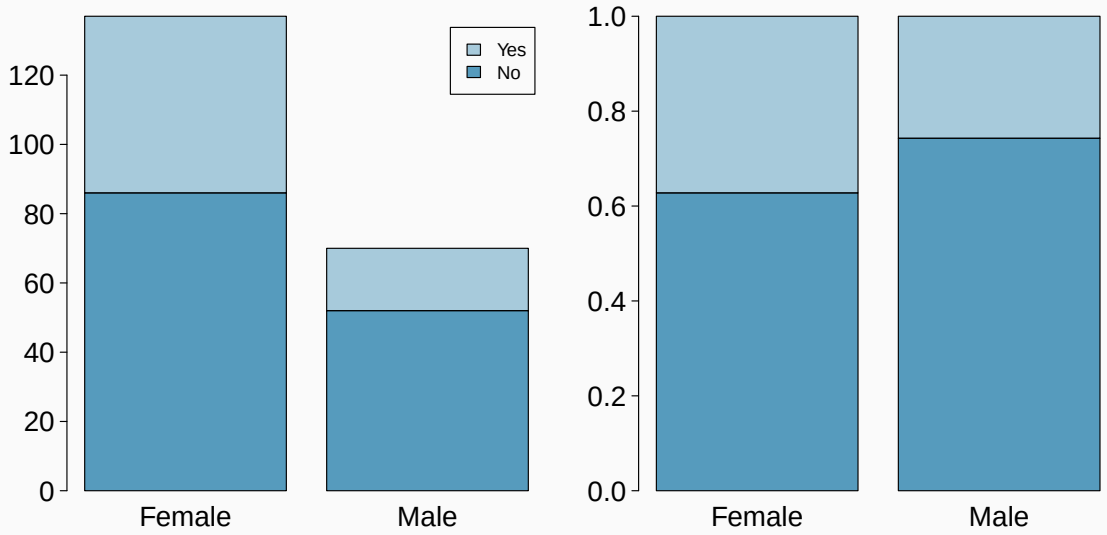
(c) house prices

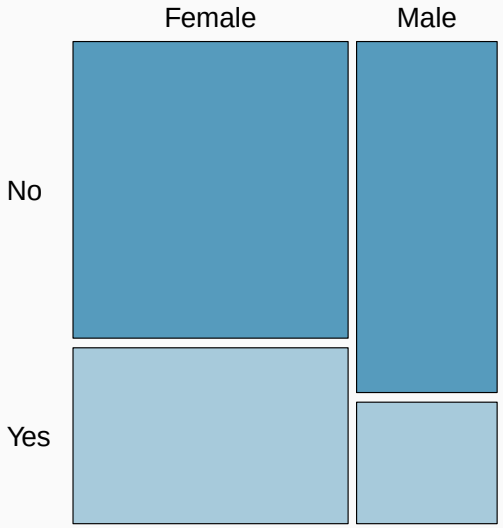
(d) *birthdays of classmates (day of the month)*

Considering categorical data

Segmented bar and mosaic plots

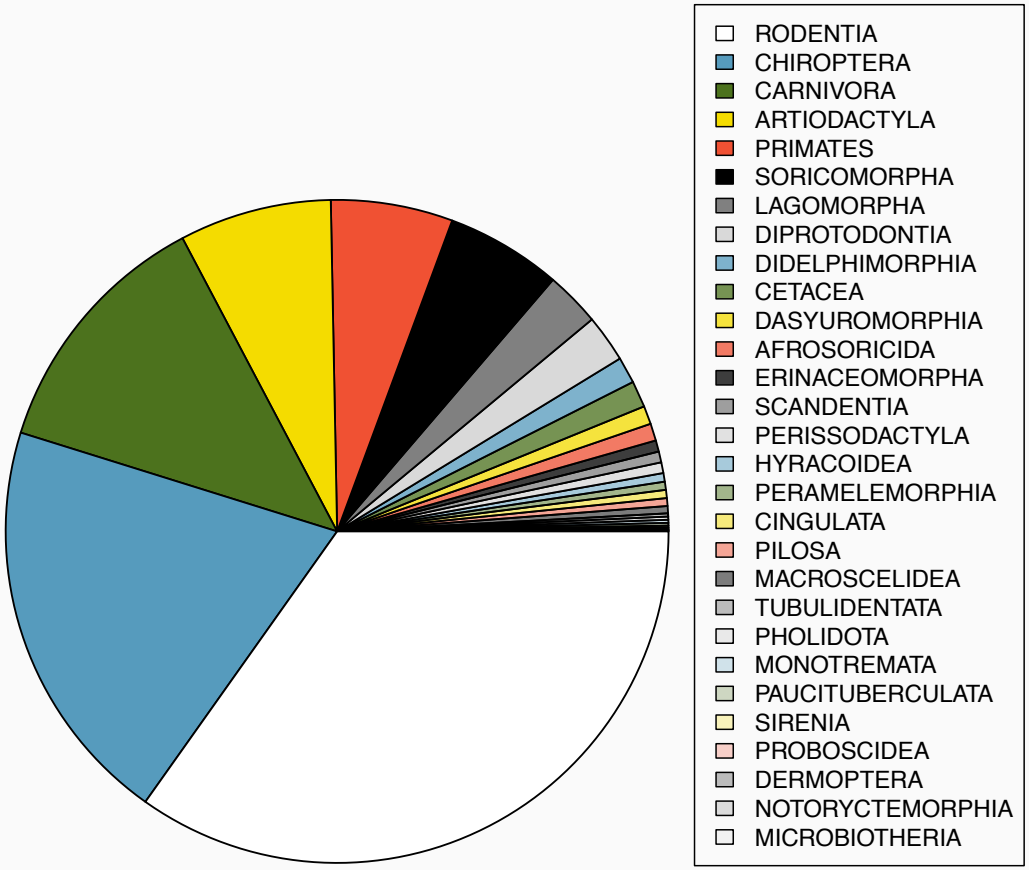
What are the differences between the three visualizations shown below?





Pie charts

Can you tell which order encompasses the lowest percentage of mammal species?



Data from *http://www.bucknell.edu/msw3*.

Data visualization as communication

Why is data visualization important?

> *Nothing in science has any value to society if it is not communicated, and scientists are beginning to learn their social obligations.*
>
> — Anne Roe, *The Making of a Scientist* (1953)

Why is data visualization important?

> *Nothing in science has any value to society if it is not communicated, and scientists are beginning to learn their social obligations.*
>
> — Anne Roe, *The Making of a Scientist* (1953)

> *If you cannot - in the long run - tell everyone what you have been doing, your doing has been worthless.*
>
> — Erwin Schrodinger (Nobel Prize winner in physics)

Why is data visualization important?

> *Nothing in science has any value to society if it is not communicated, and scientists are beginning to learn their social obligations.*
>
> — Anne Roe, *The Making of a Scientist* (1953)

> *If you cannot - in the long run - tell everyone what you have been doing, your doing has been worthless.*
>
> — Erwin Schrodinger (Nobel Prize winner in physics)

> *The greatest value of a picture is when it forces us to notice what we never expected to see.*
>
> — John Tukey (Mathematician, recipient of National Medal of Science)

Why is data visualization important?

> *Nothing in science has any value to society if it is not communicated, and scientists are beginning to learn their social obligations.*
>
> — Anne Roe, *The Making of a Scientist* (1953)

> *If you cannot - in the long run - tell everyone what you have been doing, your doing has been worthless.*
>
> — Erwin Schrodinger (Nobel Prize winner in physics)

> *The greatest value of a picture is when it forces us to notice what we never expected to see.*
>
> — John Tukey (Mathematician, recipient of National Medal of Science)

> *Numbers have an important story to tell. They rely on you to give them a clear and convincing voice.*
>
> — Stephen Few (Founder of Perceptual Edge, author of *Show Me the Numbers*)

Why is data visualization important?

> *Nothing in science has any value to society if it is not communicated, and scientists are beginning to learn their social obligations.*
>
> — Anne Roe, *The Making of a Scientist* (1953)

> *If you cannot - in the long run - tell everyone what you have been doing, your doing has been worthless.*
>
> — Erwin Schrodinger (Nobel Prize winner in physics)

> *The greatest value of a picture is when it forces us to notice what we never expected to see.*
>
> — John Tukey (Mathematician, recipient of National Medal of Science)

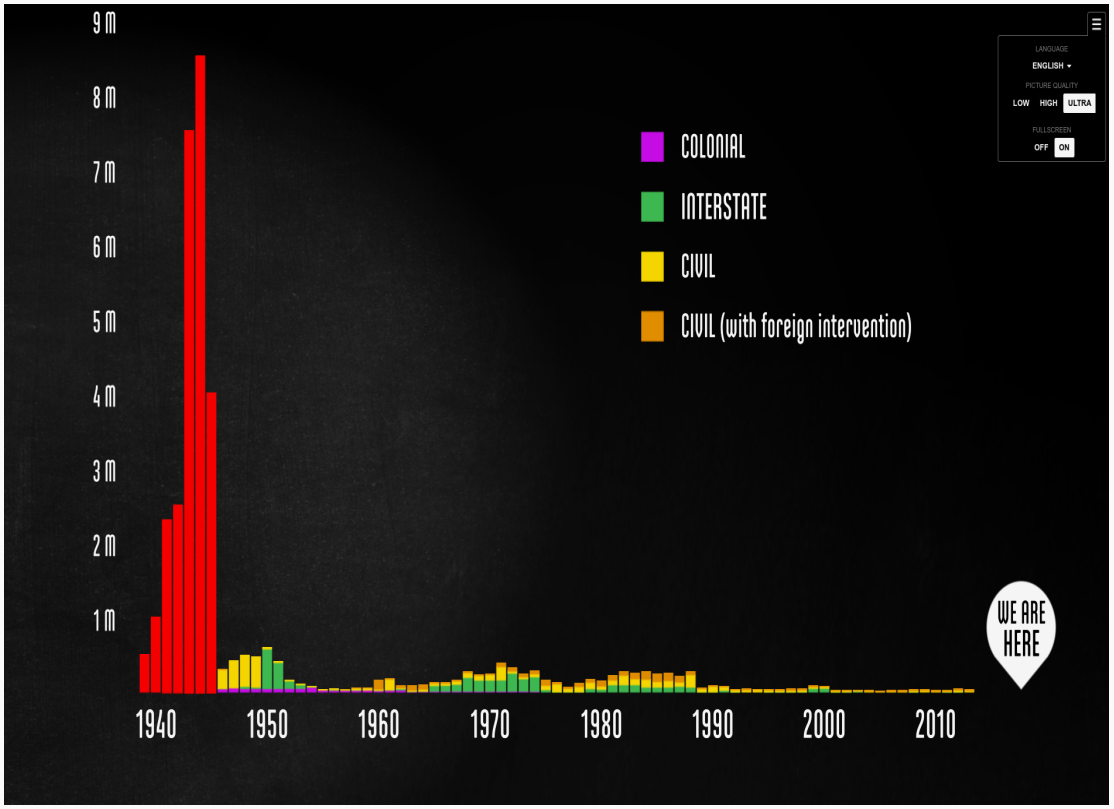
> *Numbers have an important story to tell. They rely on you to give them a clear and convincing voice.*
>
> — Stephen Few (Founder of Perceptual Edge, author of *Show Me the Numbers*)

> *Visualizations act as a campfire around which we gather to tell stories.*
>
> — Al Shalloway (Founder and CEO of Net Objectives)

Effective presentations ↔ effective visuals

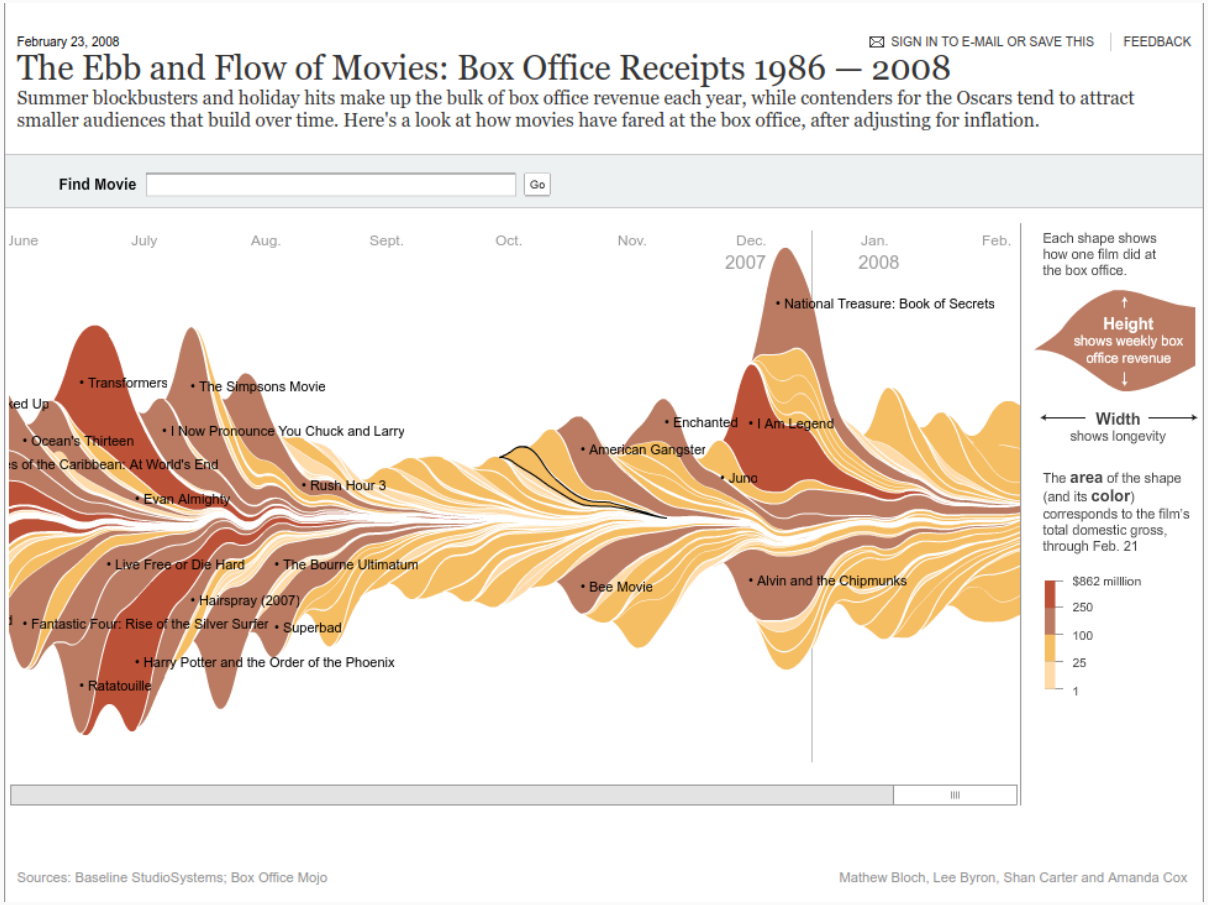
Source: Digital Image, *AP* photo used on *Business Insider*, Accessed September 10, 2017, http://www.businessinsider.com/the-first-iphone-2013-12

Visualizations can lead to comprehension...



Source: The Fallen of World War II

...or to confusion



Source: The Ebb and Flow of Movies - Box Office Receipts 1986--2008 - Interactive Graphic - NYTimes.com

Poor visualizations may lead to tragedy

- The *Challenger* disaster, January 28th, 1986

Poor visualizations may lead to tragedy

- The *Challenger* disaster, January 28th, 1986
- The Space Shuttle Challenger broke apart 73 seconds into flight, all seven crew members died

Poor visualizations may lead to tragedy

- The *Challenger* disaster, January 28th, 1986
- The Space Shuttle Challenger broke apart 73 seconds into flight, all seven crew members died
- The rubber O-rings, which held the rockets together, had failed due to the low temperatures (below 30° F)

Poor visualizations may lead to tragedy

- The *Challenger* disaster, January 28th, 1986
- The Space Shuttle Challenger broke apart 73 seconds into flight, all seven crew members died
- The rubber O-rings, which held the rockets together, had failed due to the low temperatures (below 30° F)
- Engineers at Morton Thiokol, who supplied solid rocket motors to NASA, warned about this on January 27th, 1986 in a conference call

Poor visualizations may lead to tragedy

- The *Challenger* disaster, January 28th, 1986
- The Space Shuttle Challenger broke apart 73 seconds into flight, all seven crew members died
- The rubber O-rings, which held the rockets together, had failed due to the low temperatures (below 30° F)
- Engineers at Morton Thiokol, who supplied solid rocket motors to NASA, warned about this on January 27th, 1986 in a conference call
- NASA and the managers at Morton Thiokol overruled their concerns, unpersuaded by the engineers

The engineers presented tables like this one

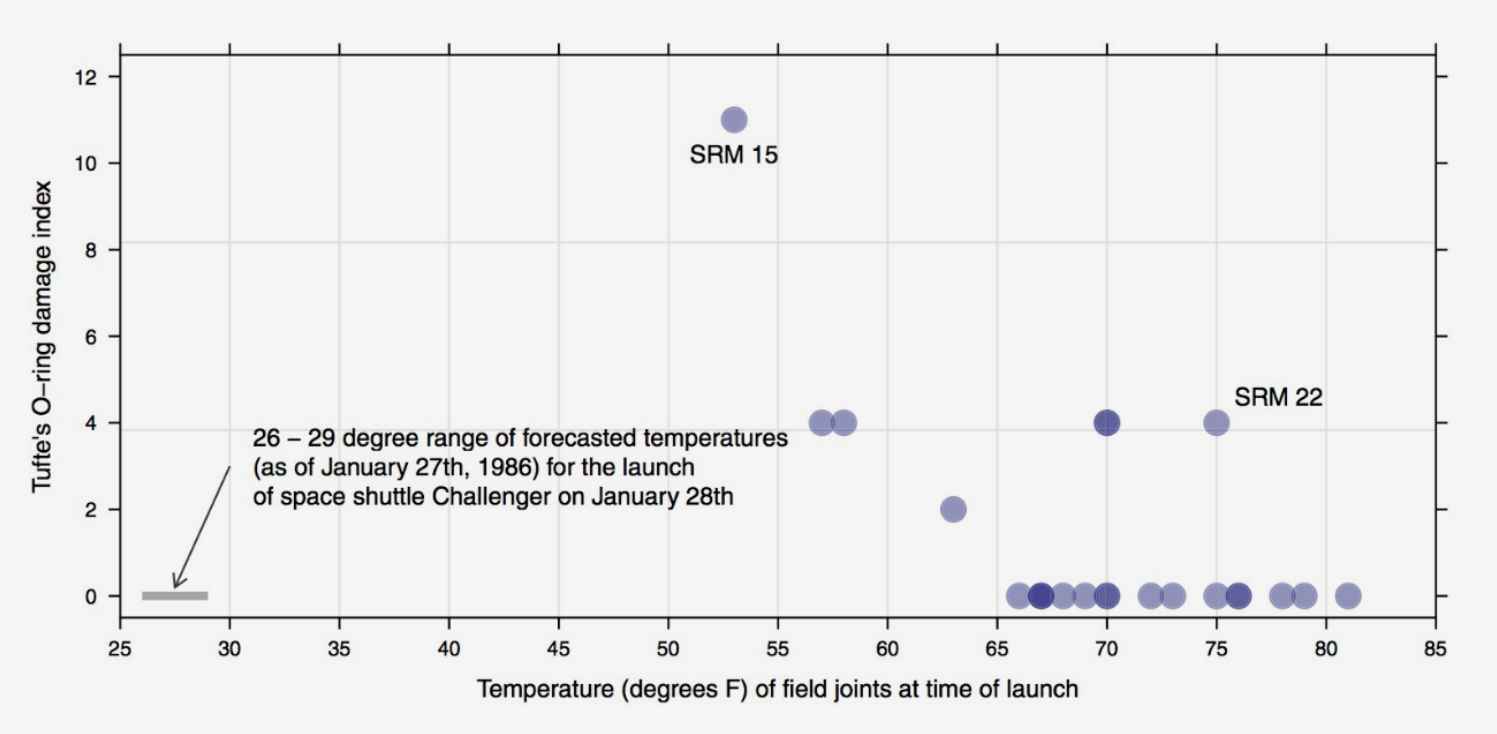
HISTORY OF O-RING TEMPERATURES
(DEGREES - F)

MOTOR	MBT	AMB	O-RING	WIND
QM-1	68	36	47	10 MPH
QM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH

Source: Figure 2.18(a) in *Modern Data Science with R* by Benjamin Baumer, Daniel Kaplan, and Nicholas Horton

Edward Tufte's critique of the Challenger disaster

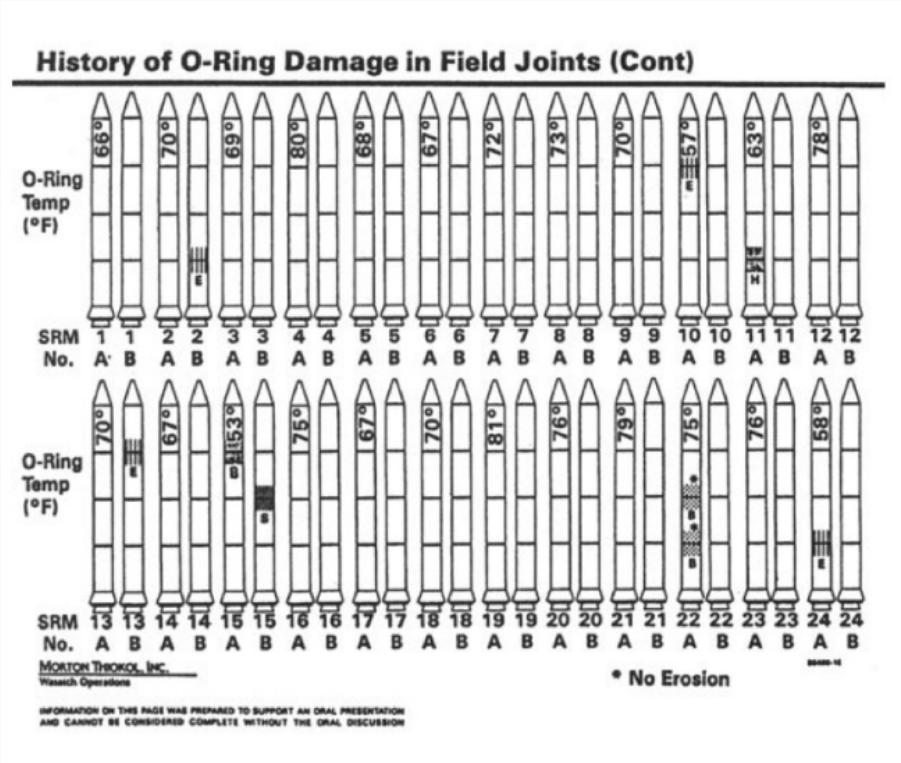
Mathematician Edward Tufte issued a critique and argued that the data should have been presented this way:



Source: Figure 2.17 in *Modern Data Science with R* by Benjamin Baumer, Daniel Kaplan, and Nicholas Horton

"Chartjunk" in Challenger Congressional Hearings

This information was presented in Congressional Hearings about the incident in this format:



Source: Figure 2.18(b) in *Modern Data Science with R* by Benjamin Baumer, Daniel Kaplan, and Nicholas Horton

How to Lie with Statistics

- Book by Darrell Huff, published in 1954

How to Lie with Statistics

- Book by Darrell Huff, published in 1954
- Aside: The title is tongue-in-cheek and is usually misunderstood. The book is not about "fudging the numbers" with statistics.

How to Lie with Statistics

- Book by Darrell Huff, published in 1954
- Aside: The title is tongue-in-cheek and is usually misunderstood. The book is not about "fudging the numbers" with statistics.
- Illustrates ways that visualizations can be manipulated such that they are misleading, but technically show accurate information

How to Lie with Statistics

- Book by Darrell Huff, published in 1954
- Aside: The title is tongue-in-cheek and is usually misunderstood. The book is not about "fudging the numbers" with statistics.
- Illustrates ways that visualizations can be manipulated such that they are misleading, but technically show accurate information
- **General method:** Violate conventions and expectations

Example 1: gun deaths in Florida over time

- Context: Florida passed a "Stand Your Ground" law in 2005

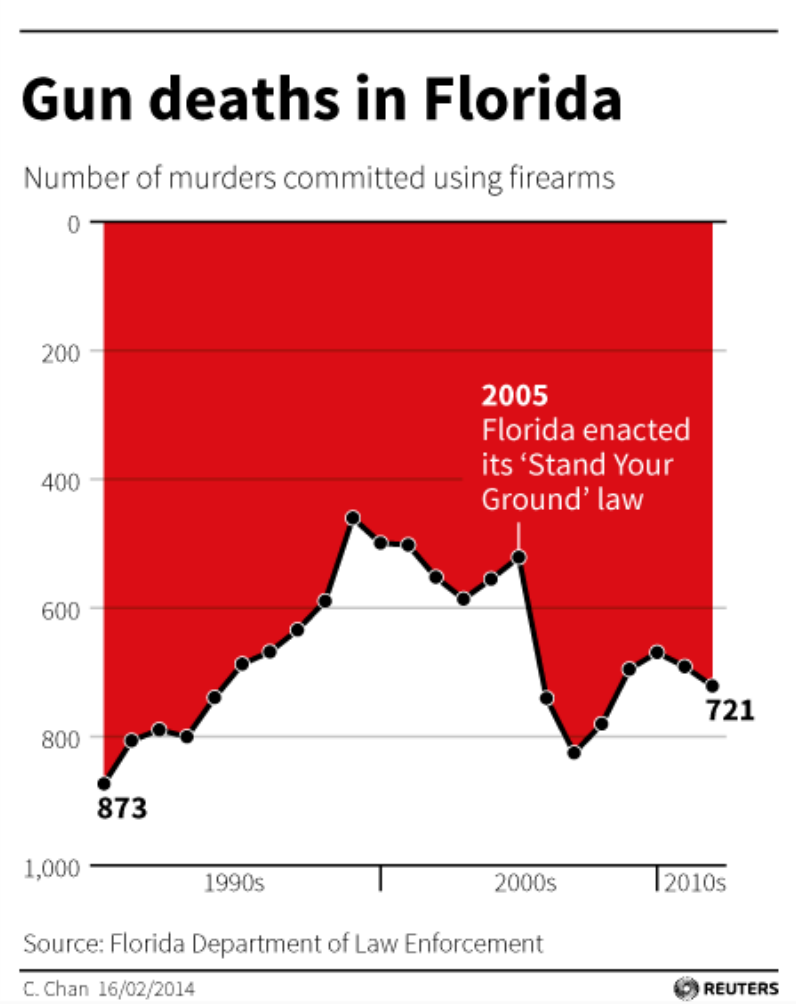
Example 1: gun deaths in Florida over time

- Context: Florida passed a "Stand Your Ground" law in 2005
- Advocates claimed it would reduce crime, opponents argued it would increase use of lethal force

Example 1: gun deaths in Florida over time

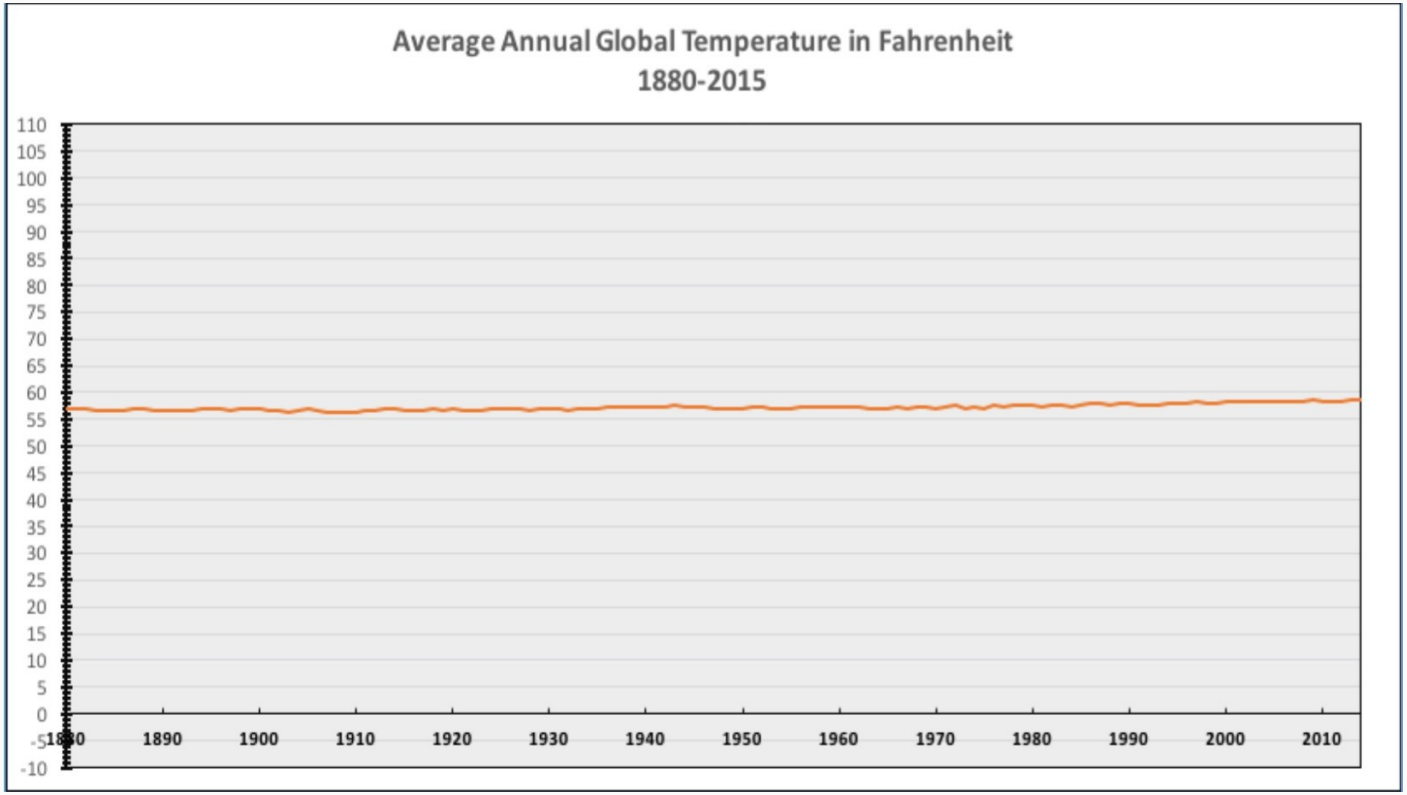
- Context: Florida passed a "Stand Your Ground" law in 2005
- Advocates claimed it would reduce crime, opponents argued it would increase use of lethal force
- If you wanted to use data to answer this question, and you came across this graphic published by the news organization Reuters, what would you conclude?

Example 1: gun deaths in Florida over time



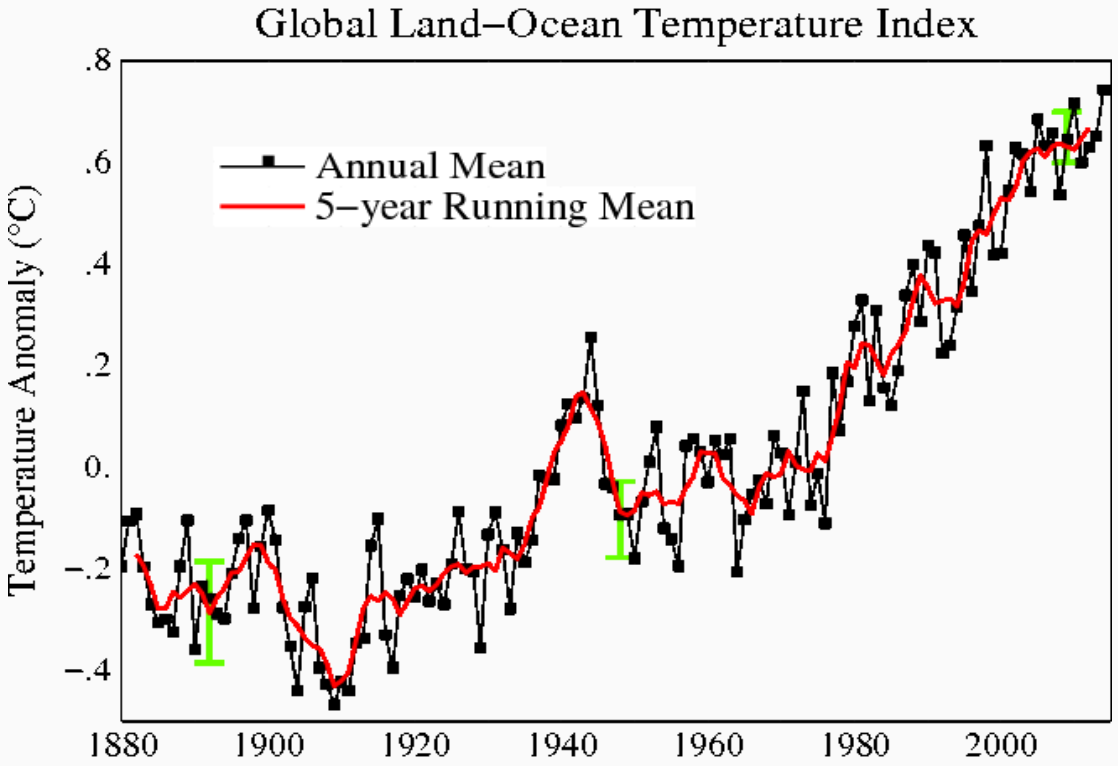
Example 2: average global temperature over time

- The political/editorial magazine National Review tweeted the following visualization on December 14, 2015. The visualization originates from a post on a political blog called Power Line.



Example 2: average global temperature over time

- Here's a conventional version of the same data:



Source: Nasa Goddard Institute for Space Studies

Side note: How do we have a record going back to the 1880s?

Temperatures from the 1800s and onward were recorded using thermometers at various locations around the globe, and by the 1880s thermometers had become precise. Systematic measurements began around the mid-1800s at various army posts, and in 1891 the National Weather Service was formed to continue the effort.

Source: National Oceanic and Atmospheric Administration, "How do we observe today's climate?"

Principles and ethics for scientific visualizations

1. Present your results transparently and honestly

Principles and ethics for scientific visualizations

1. Present your results transparently and honestly
2. Show all data, including outliers, that are valid measurements

Principles and ethics for scientific visualizations

1. Present your results transparently and honestly
2. Show all data, including outliers, that are valid measurements
3. Use graph layouts that show trends and lets readers easily read quantitative values

Principles and ethics for scientific visualizations

1. Present your results transparently and honestly
2. Show all data, including outliers, that are valid measurements
3. Use graph layouts that show trends and lets readers easily read quantitative values
4. Do not break conventions regarding scaling, axis orientation, the type of plot to use, etc.

Principles and ethics for scientific visualizations

1. Present your results transparently and honestly
2. Show all data, including outliers, that are valid measurements
3. Use graph layouts that show trends and lets readers easily read quantitative values
4. Do not break conventions regarding scaling, axis orientation, the type of plot to use, etc.
5. If you leave something out of a visualization, say so and justify it

Principles and ethics for scientific visualizations

1. Present your results transparently and honestly
2. Show all data, including outliers, that are valid measurements
3. Use graph layouts that show trends and lets readers easily read quantitative values
4. Do not break conventions regarding scaling, axis orientation, the type of plot to use, etc.
5. If you leave something out of a visualization, say so and justify it
6. Strongly consider including your datasets and any scripts used to create figures with your reports or journal articles

Credits

- Examples from slides before **Data visualization as communication**, as well as the slides with the blue headers, were adapted from the Chapter 1 OpenIntro Statistics Slides developed by Mine Cetinkaya-Rundel and made available under the CC BY-SA 3.0 license.
- Ideas and examples in the section **Data visualization as communication** were adapted from *Modern Data Science with R* by Benjamin Baumer, Daniel Kaplan, and Nicholas Horton, chapters 2 and 6.