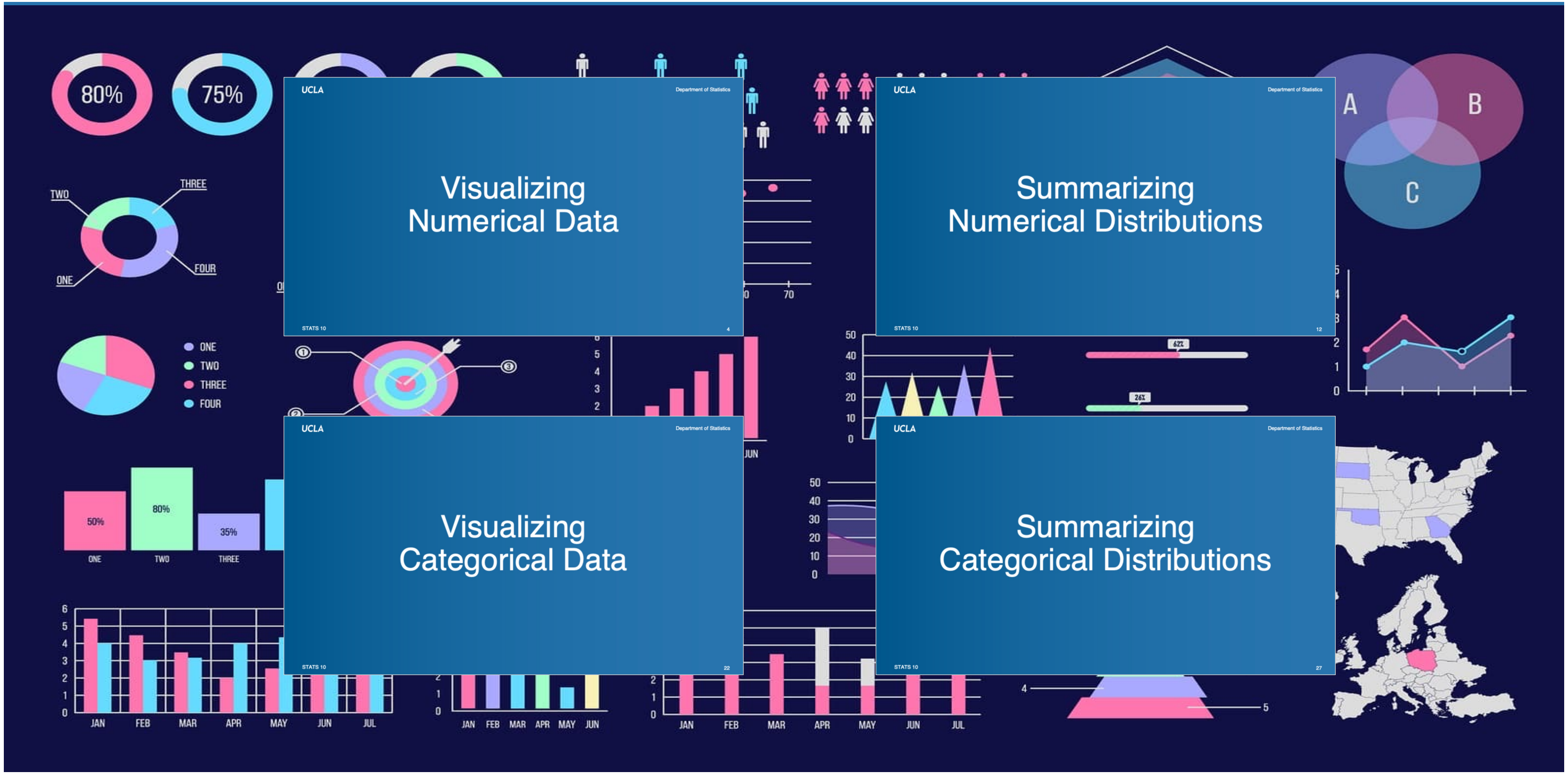
UCLA College | Physical Sciences
Statistics

STATS 10 - Chapter 2
Descriptive Statistics and Visualizing Data

Bingling Wang
Department of Statistics

Topics



UCLA Department of Statistics

Visualizing Numerical Data

STATS 10 4

UCLA Department of Statistics

Summarizing Numerical Distributions

STATS 10 12

UCLA Department of Statistics

Visualizing Categorical Data

STATS 10 22

UCLA Department of Statistics

Summarizing Categorical Distributions

STATS 10 27

Examining Distributions

Distribution -- The most important tool for organizing the variation in data.

what values the variable takes, and how often the variable takes those values.

Distributions are important because:

- Make comparisons between groups
- Examine data for errors
- Learn about real-world processes

Graphics can be extraordinarily powerful ways of organizing data, detecting patterns and trends, and communicating findings.

Visualizing Numerical Data

Dot Plot

Construct a dot plot:

Put a dot above a number line for each value occurs in the data.

If a value occurs more than once, we stack dots on top of each other.

Example:

Here are the launch-temperatures of the first 25 shuttle missions (in degrees F) in the U.S.

66,70,69,80,68,67,72,70,70,57,63,70,78,67,53,67,75,70,81,76,79,75,76,58,29

Dot Plot

Construct a dot plot:

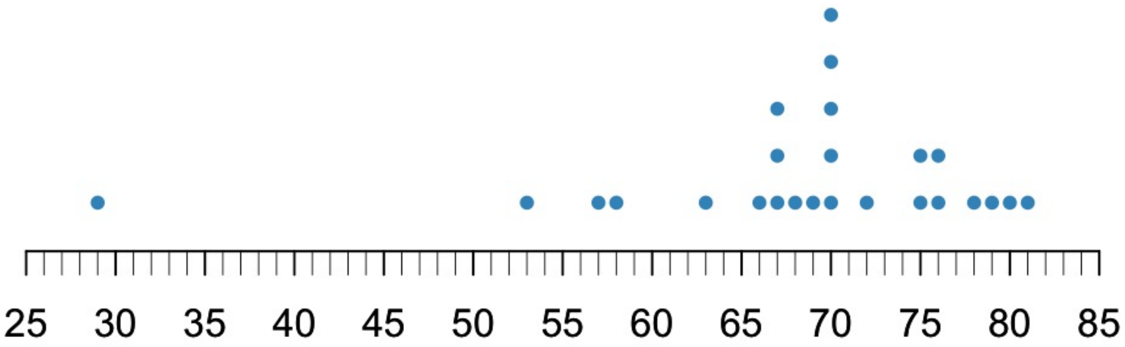
Put a dot above a number line for each value occurs in the data.

If a value occurs more than once, we stack dots on top of each other.

Example:

Here are the launch-temperatures of the first 25 shuttle missions (in degrees F) in the U.S.

66,70,69,80,68,67,72,70,70,57,63,70,78,67,53,67,75,70,81,76,79,75,76,58,29

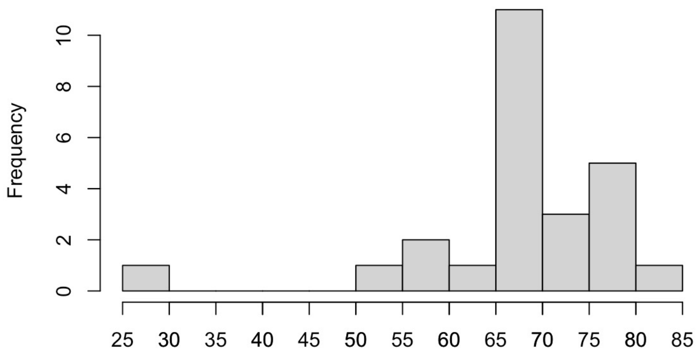


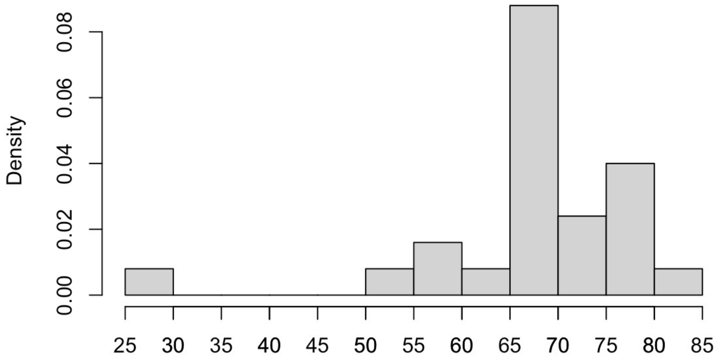
Histogram

Construct a histogram:

1. Group data into **intervals (bins)** of equal width.
2. Count the number of observations that fall into each bin.
3. Draw a vertical bar over the bins, the height is the proportional to the frequency in each interval.

Changing the vertical scale (**frequency/density)** does not change the shape





Example

Exam grades

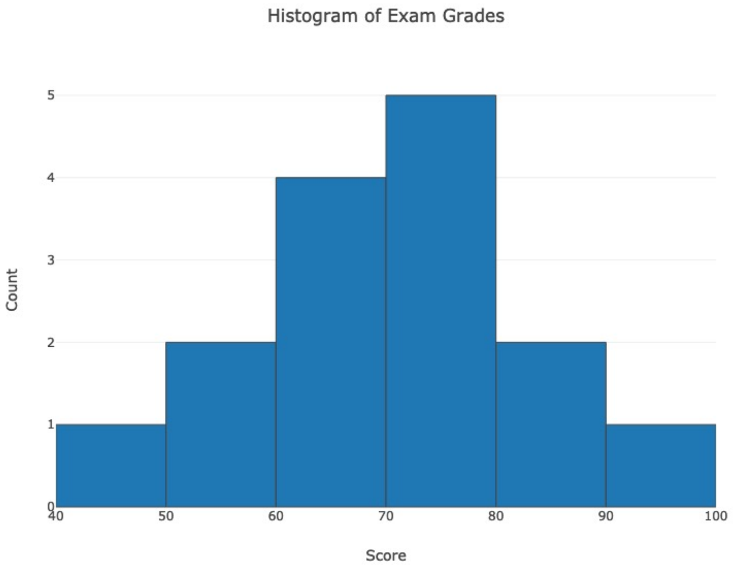
88, 48, 60, 51, 57, 85, 69, 75, 97, 72, 71, 79, 65, 63, 73

1. Choose intervals: e.g., 10 points wide, [40-50), [50-60), …, [90-100]
2. Count:

Score	**[40, 50)**	**[50, 60)**	**[60, 70)**	**[70, 80)**	**[80, 90)**	**[90, 100]**
Count						

Example

Score	[40, 50)	[50, 60)	[60, 70)	[70, 80)	[80, 90)	[90, 100]
Count	1	2	4	5	2	1

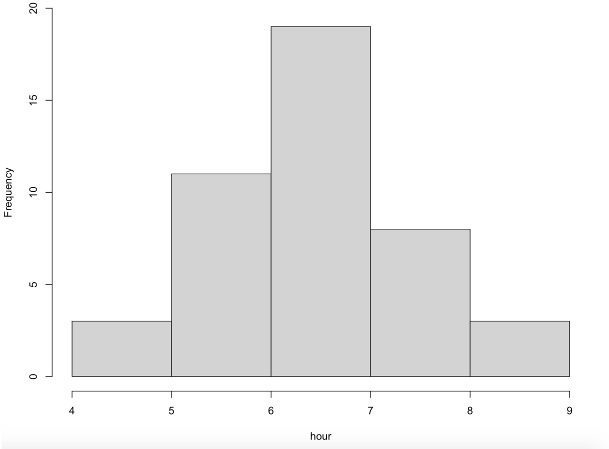


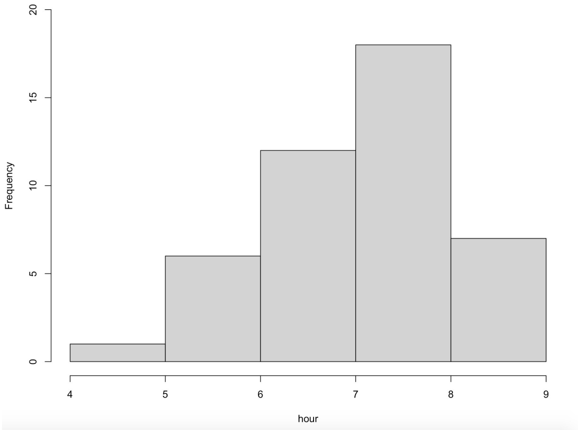
Boundary points

Observations may land right on the edge (or boundary) of two bins. We need to decide which bin these edge cases would fall into.

- Put "boundary" observations in the bin on the left.
  Then 5 would go into the bin from 4 to 5.
- Put "boundary" observations in the bin on the right.
  Then 5 would go into the bin from 5 to 6.

Be consistent!

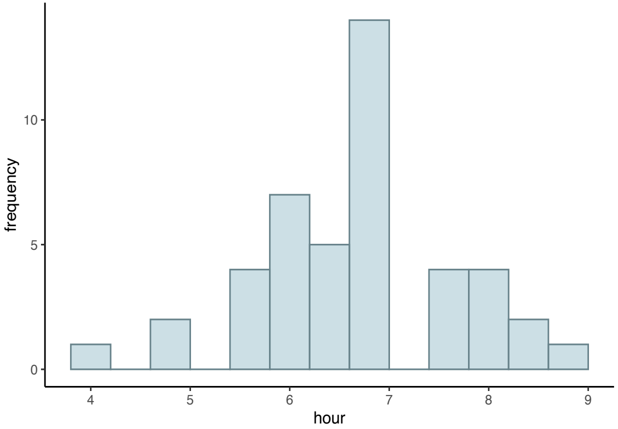


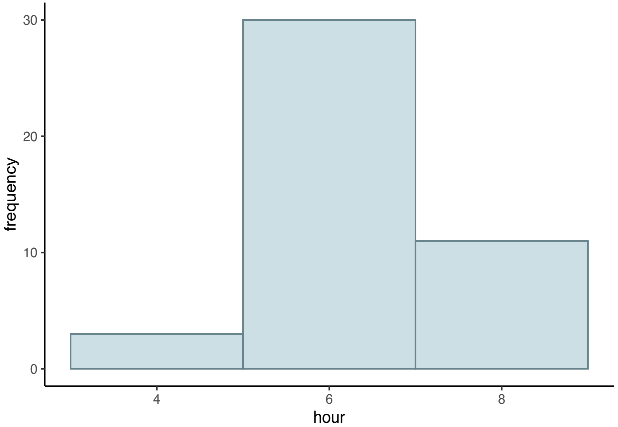


Histogram Bin Widths

- Bin width is the width of the interval.
- Changing the width of the bins in a histogram changes its shape.
- Too many bins show too much detail while too little bin shows too little.

[Data source] [Plot tool]





Summarizing Numerical Distributions

Important Features of a Numerical Distribution

When describing a numerical distribution, we should consider the following features:

Shape

Center (Typical Value)

Spread (Variability)

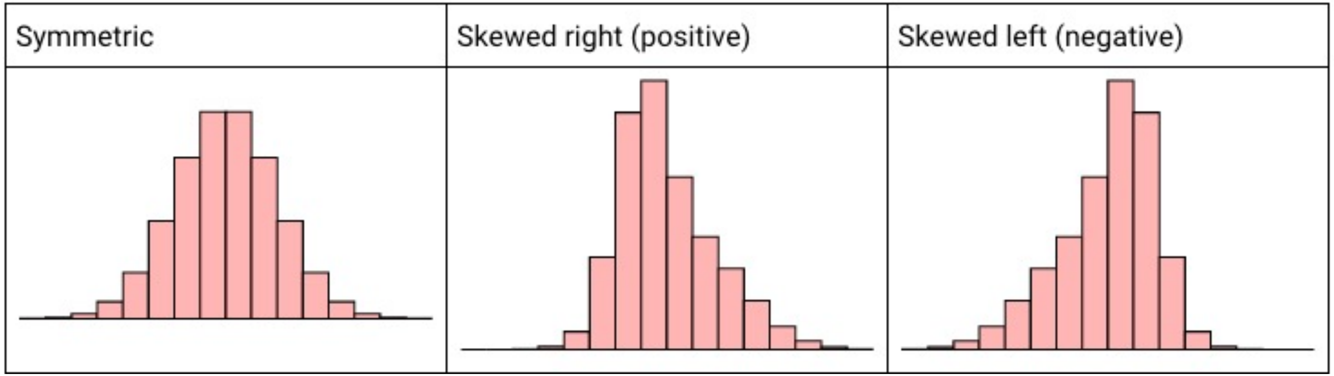
Outliers

Shape I – Symmetry/Skewness

Symmetric: left and right side roughly the same

Skewed:

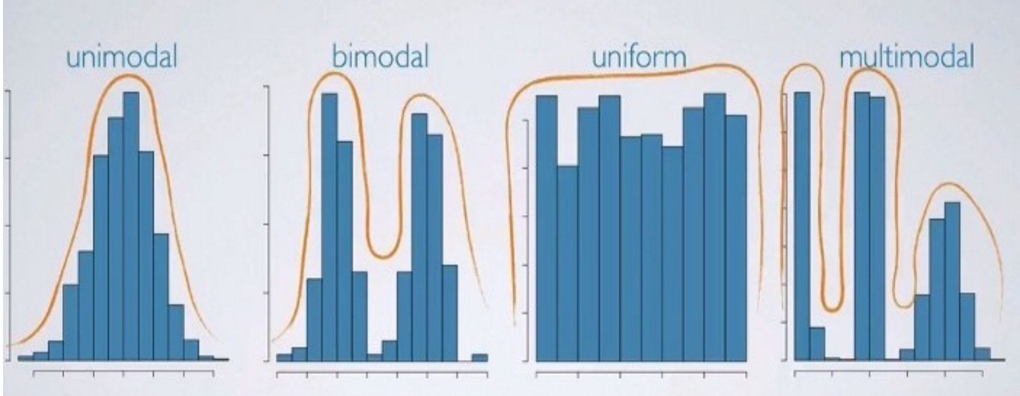
- Right-skewed/positively skewed: tail goes to the right
- Left-skewed/negatively skewed: tail goes to the left
- The right- and left-skewness refers to the direction of the tail, not to where the bulk of the data is.



Shape II -- Modality

Number of modes (peaks)

- **Unimodal**: single mode
- **Bimodal**: two modes
- **Multimodal**: more than two modes
- **Uniform**: no apparent peaks

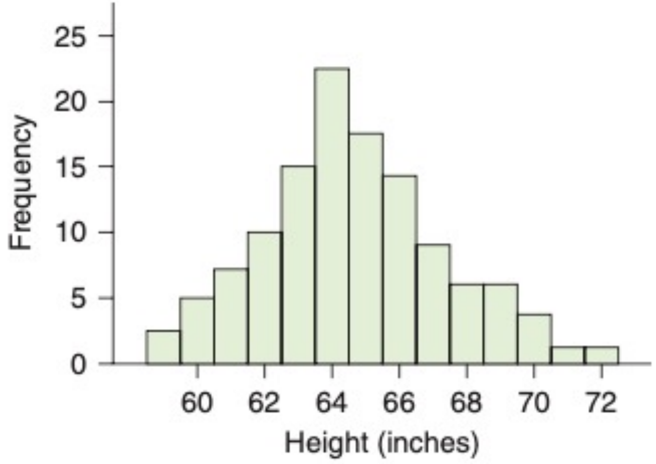


Center I

The typical data value

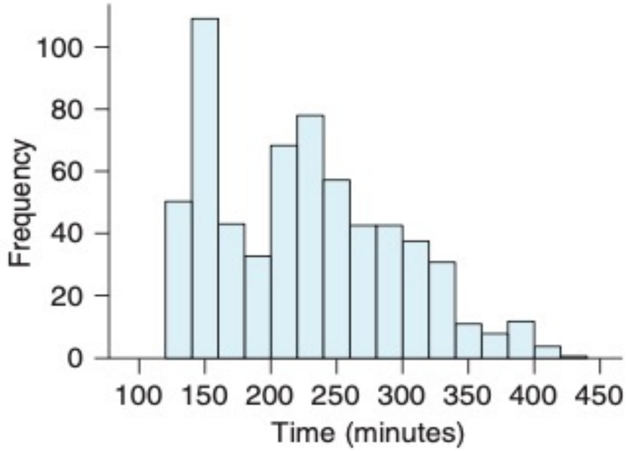
What are the typical values for the following examples?

123 college women’s heights



The finishing times for two different marathons.

- Marathon in the 2012 Olympic Games
- A marathon in Portland, Oregon



Center II

Numerical variables are often summarized using numbers to communicate their central tendency.

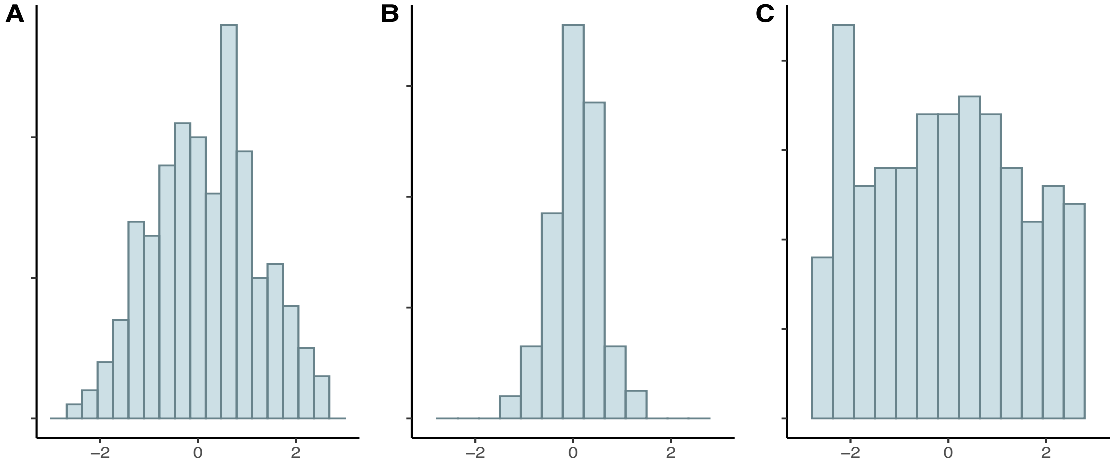
Different measures for the center

- **Mean:** the average of observed data values
- **Median:** the middle value that divides the ordered data into half
- **Mode:** the most frequent value

Spread / Variability

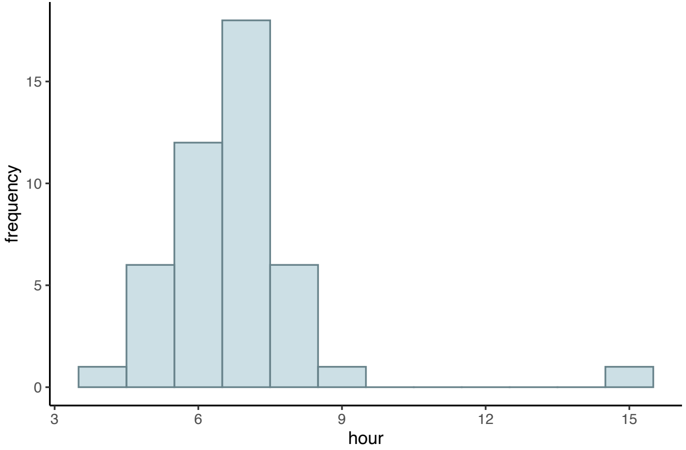
How spread out the data is from the center

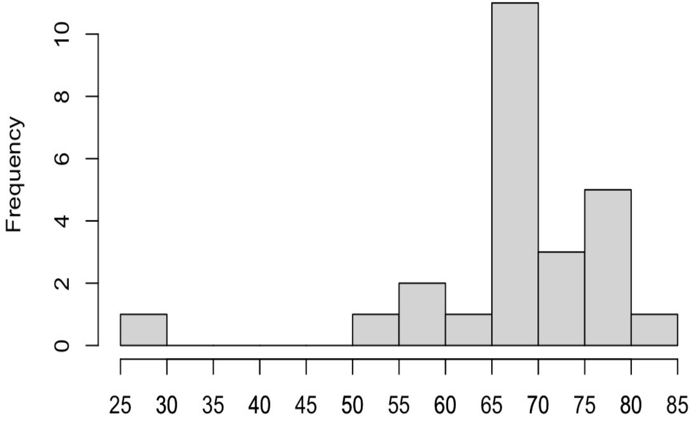
- The data values are tightly clustered around the center, little variability
- Data values are scattered far from the center, high variability



Outliers

- Extremely small or large values
- Data values that don’t fit into the pattern of the distribution





Any potential outliers should be identified and investigated.

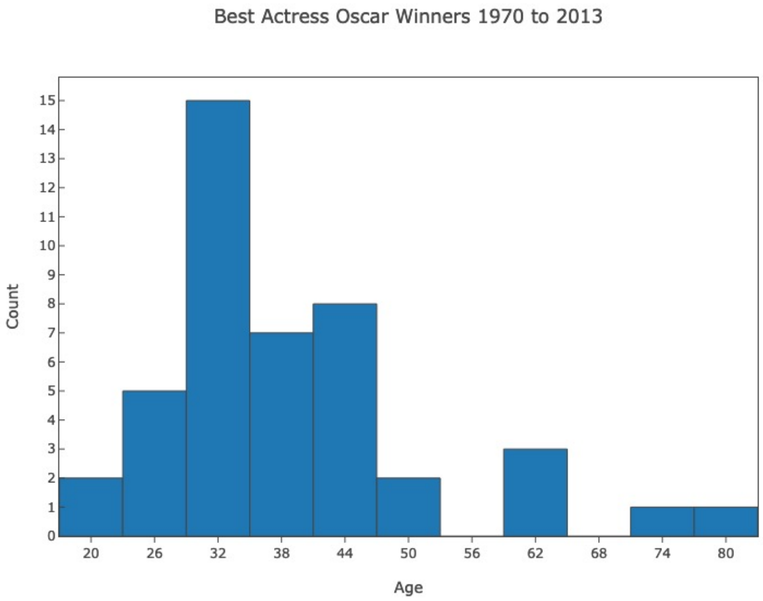
Summarizing a Numerical Distribution

Checklist:

- **Shape**
  - Is the distribution symmetric or skewed?
  - How many peaks are there?
- **Center**
  - Where do most of the values lie?
  - What is the typical value(s)?
- **Variability**
  - How much variability is there?
  - Are the data values clustered closely together, or spread far apart?
- **Outliers**
  - Any extreme/unusual value?

Exercise

Best Actress Oscar Winners



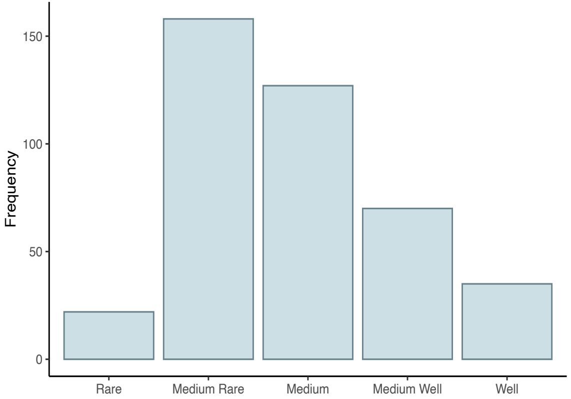
Visualizing Categorical Data

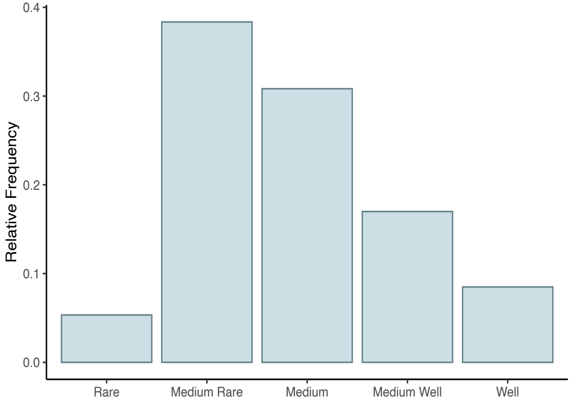
Bar Chart

- Horizontal axis – data categories
- Vertical axis – (relative) frequency for each category

Example: Steak preference of 412 American steak eaters [StatCrunch]

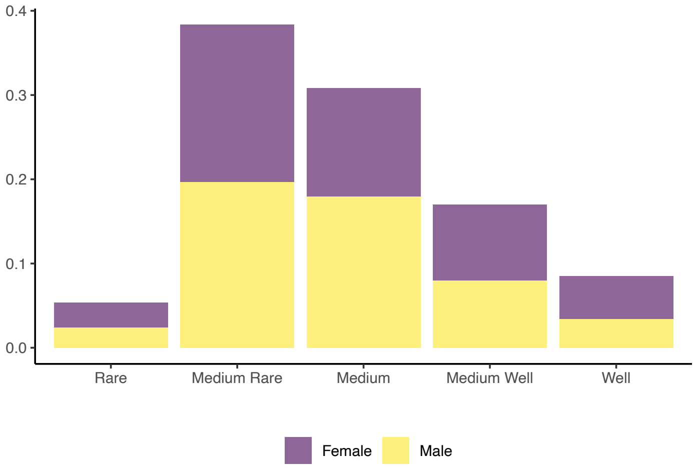
Preference	Rare	Medium Rare	Medium	Medium Well	Well
Frequency	22	158	127	70	35



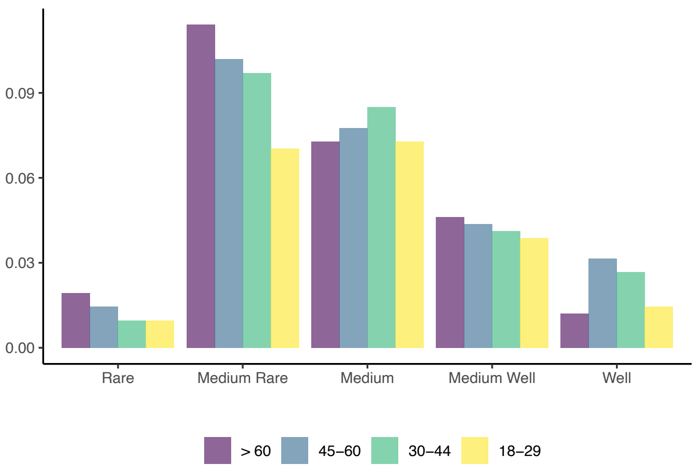


Grouped Bar Chart

Stacked Bar Chart



Side-by-Side Bar Chart



Bar Chart vs. Histogram

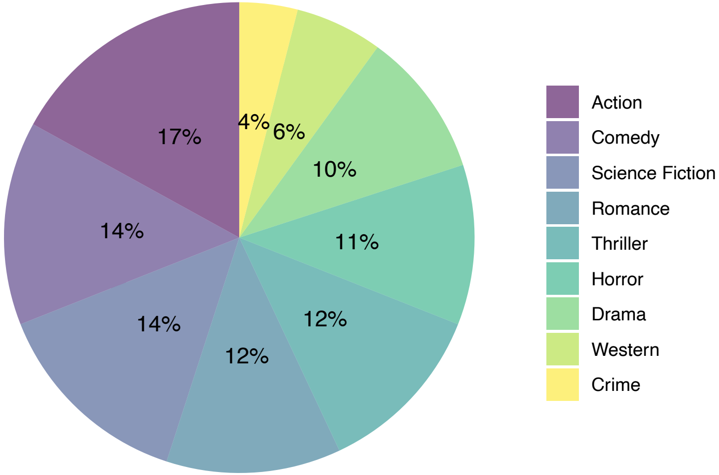
	Bar Chart	Histogram
Data	Categorical	Numerical
Bars	Usually do not touch; Gaps in between	Usually touch; Gap indicates no values
Bar width	Does not matter; No meaning	Width matters; Width same for all bars
Order	Order can change	X-axis values sorted in ascending order

Pie Chart

A circle divided into pieces. Each piece represents a **category** in the data, and the area of each piece is proportional to the **relative frequency/percentage** of the subjects in each category.

The percentages should sum up to 1.

Example: favorite type of movie [StatCrunch]



Summarizing Categorical Distributions

Describing a Categorical Distribution

Two main components:

- **Mode**: typical outcome, category of the highest frequency
  - There may be more than one mode if more than one value is tied for occurring most frequently
- **Variability/Diversity**
  - If many observations spread across many different categories, then the variability is high
  - If many observations fall into the same categories, then the variability is low

Choose a Graph

Which graph would you choose to visualize the data below?

Cell Phone Use	0-4 hours	4-8 hours	9-12 hours	12+ hours
Female	7	9	5	4
Male	10	5	4	1

A. Histogram
B. Dot plot
C. Pie Chart
D. Side-by-Side Bar Chart

Misleading Graphs

Misleading Graphs

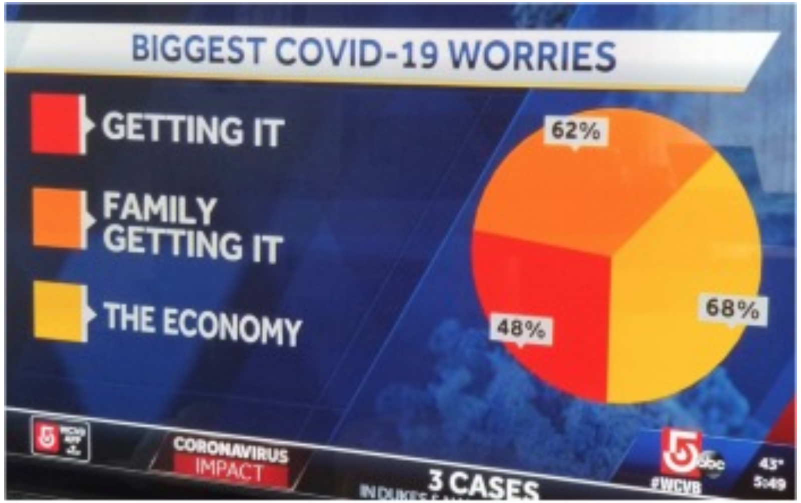
Caution!

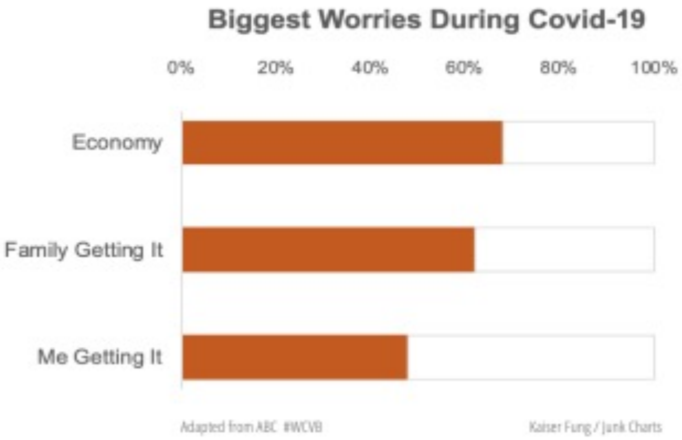
Misleading graphs play tricks with our eyes and lead to wrong conclusions.

- **Using the wrong chart**
- **Inappropriate scaling**
  - Omitting the baseline, starting at a value other than 0
  - Manipulating y-axis
- **Using symbols of different sizes rather than bars of equal width**
- **Lack of labels**

Example I

What's wrong with this chart?

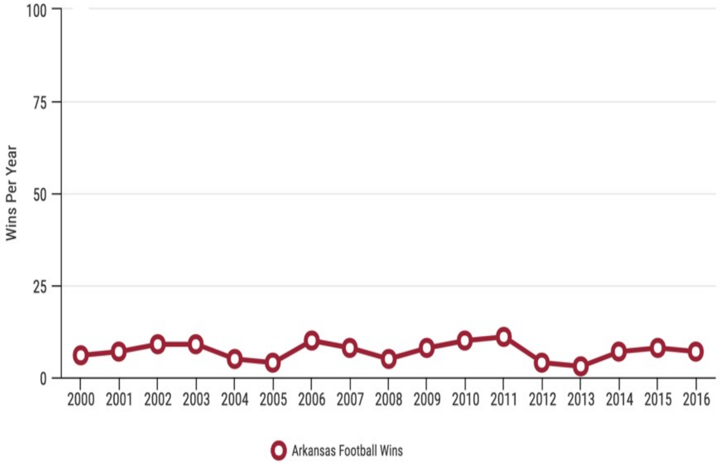


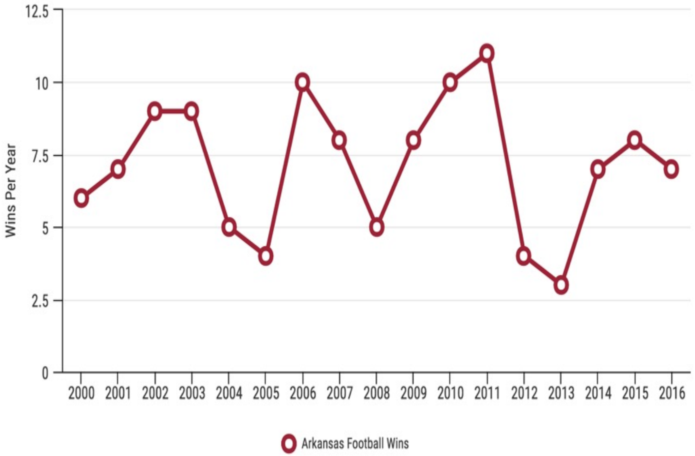


$$62\% + 48\% + 68\% \; = \; 178\% \neq 100\%$$

Example II

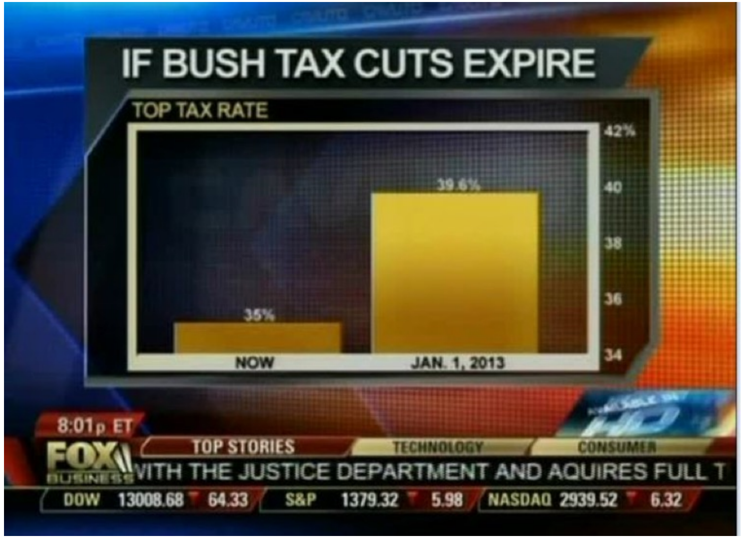
What's wrong with this chart?

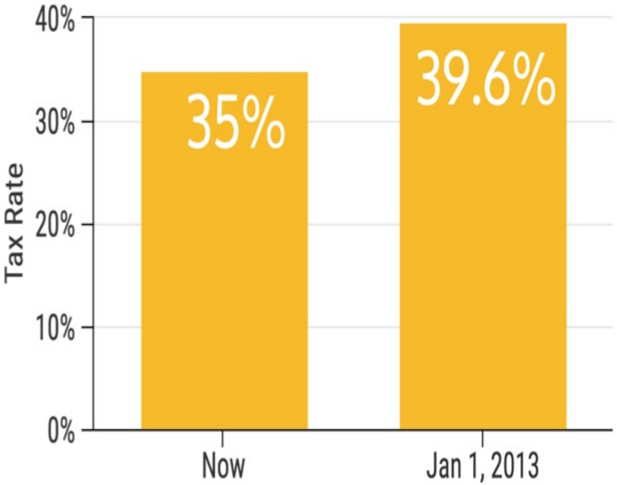




Example III

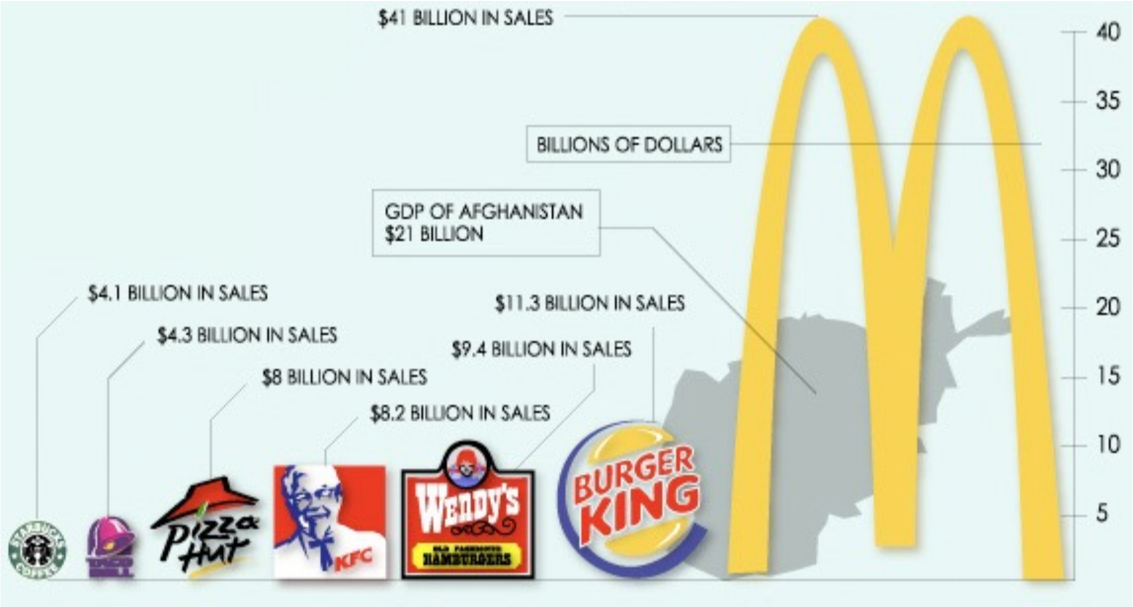
What’s wrong with this chart?



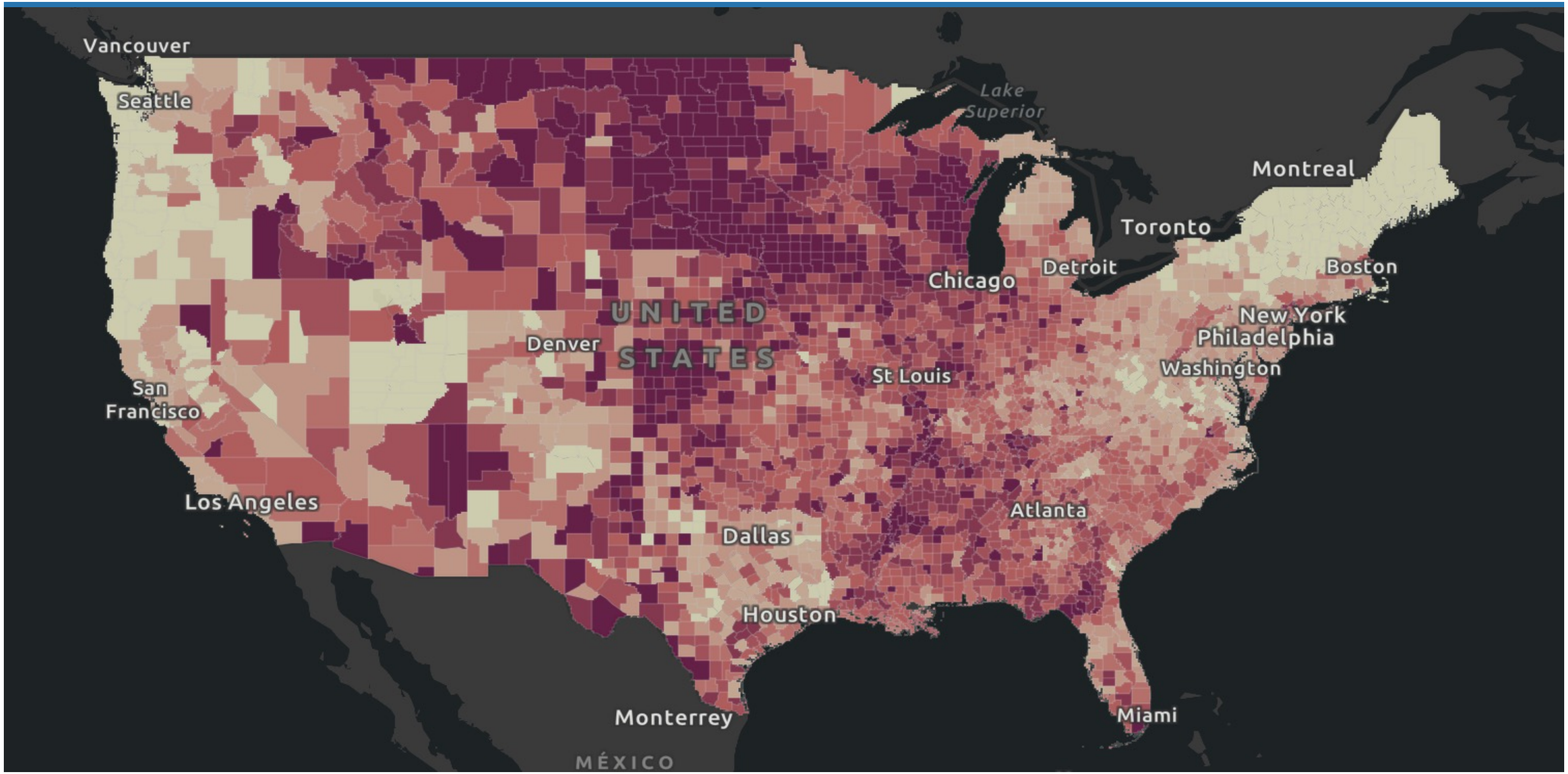


Example IV

What's wrong with this chart?



Other Visualizations


Vancouver
Seattle
Lake Superior
Montreal
Toronto
Chicago
Detroit
Boston
UNITED STATES
New York
Philadelphia
Denver
St Louis
Washington
San Francisco
Los Angeles
Atlanta
Dallas
Houston
Monterrey
Miami
MÉXICO