ST517 Note Outline 2: Summarizing Data

Notes for Lecture 2.1: Types of Data

Types of Variables (Data)

- Categorical Variable:
    - Example: Do you own a car? (Yes, No)
    - Also called **qualitative**
    - Breaking this down further, a categorical variable can have values that are:
        - Nominal:
        - Ordinal:
- Quantitative Variable:
    - Example: What is your age in years?
    - Breaking this down further, a quantitative variable can have values that are:
        - Discrete:
        - Continuous:
- Key idea: Different summaries are appropriate for the different types of variables

Exploratory Data Analysis (EDA)

- Visualize, summarize, and examine data
    - Visualize via graphical display
    - Also use numeric summaries to summarize and examine data
- Critical first step in the data analysis process (i.e. before you get to formal statistical inference)

Graphical Displays

- Quickly tells us the story behind the data
- Key idea:
    - Good data visualizations tell the story of the data in a way that is informative, easy to get, and visually appealing
    - Poor data visualizations misrepresent the story of the data, either inadvertently or intentionally
- Graphical Displays for Categorical Variables: Bar (or Mosaic) charts, Pie Charts
- Graphical Displays for Numeric Variables: Histograms, Boxplots, and Scatterplots, Time-series plots, Heat maps

Numeric Summaries

- Allow us to make comparisons

- Simplest numeric summaries for categorical variables: Count or percent in each category, tables
- Some numeric summaries for numeric variables:
    - Measures of Central Tendency: Mean, median

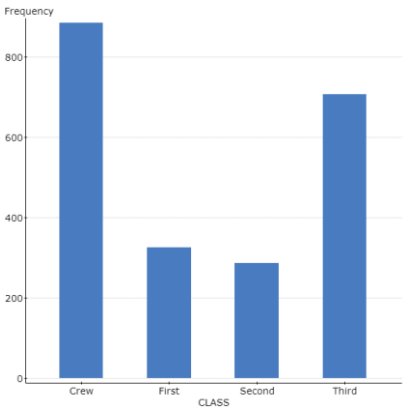
    - Measures of Variability: Variation, standard deviation, range, IQR

Numeric Summaries for Categorical Data

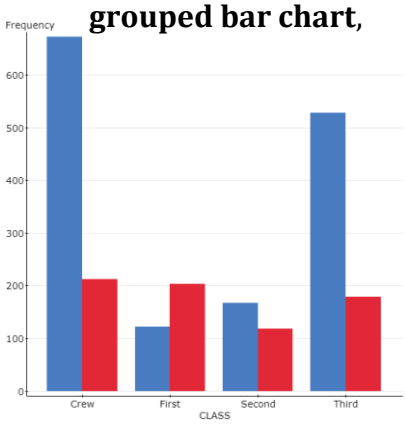
- Count in each category
- Proportion (or percent) in each category
  - Sample proportion: $\hat{p} = \frac{Count}{Sample\ size} = \frac{y}{n}$
- Tables display counts or percent for one or more categorical variables

Bar Charts

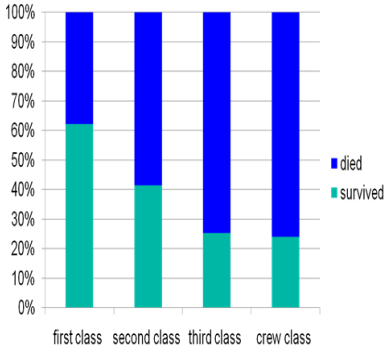
- Useful for displaying one or more categorical variables
- One bar for each category a variable
- Height of bar indicates how many [frequency] *or* percent of individuals in each category
- Ex (at right): Variable = class on the Titanic; there were over 800 crew members and 300 passengers in 2nd class
- Represent multiple categorical variables with color, using either a…



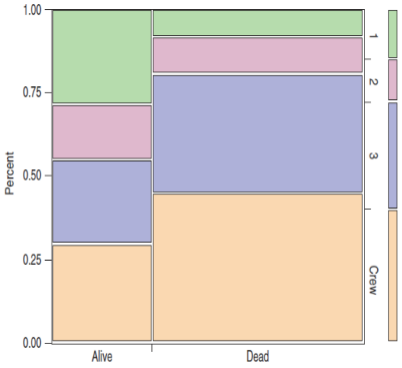
grouped bar chart,



stacked bar char, or

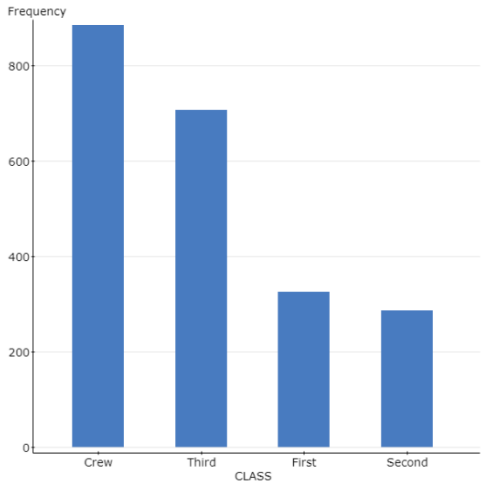


mosaic plot

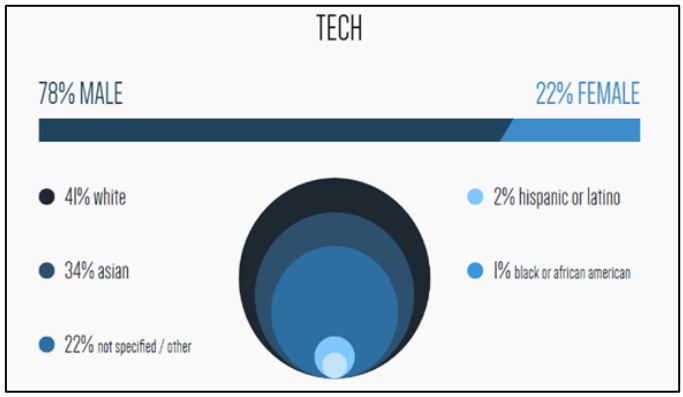


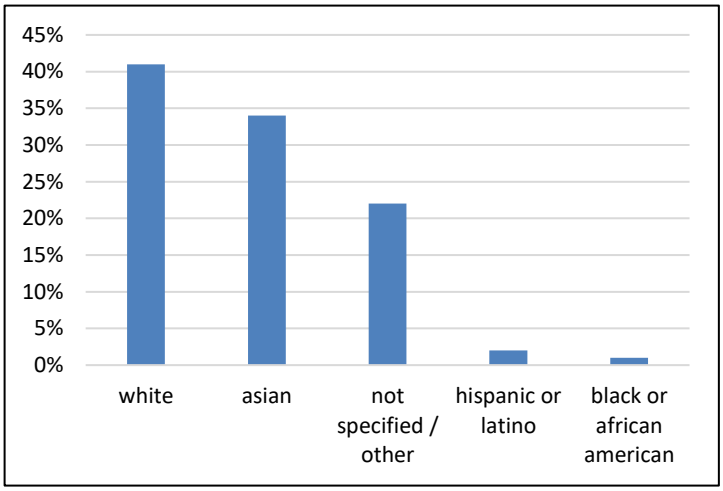
Bar Charts (continued)

- Flexibility in how you arrange bars
  - E.g. alphabetically; by frequency
  - Ordinal variable: arrange bars in order (e.g. S, M,L or L,M,S)

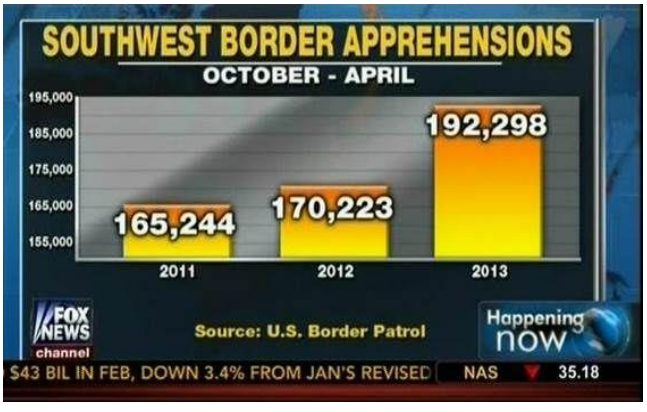


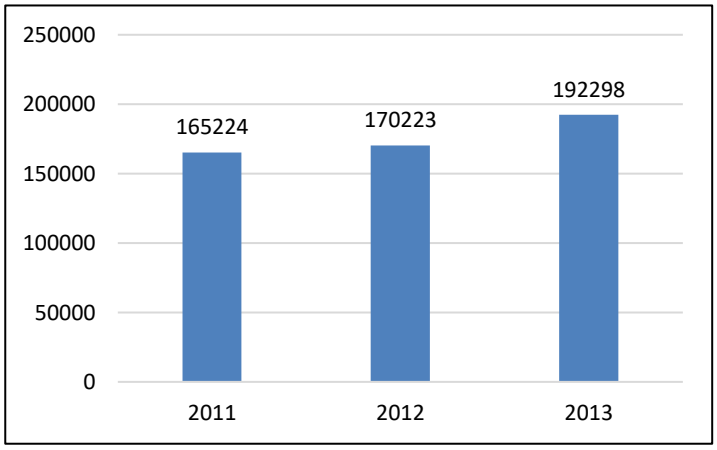
- Caution! Bar charts with other shapes can distort volume or scale, and thus distort the story of the data





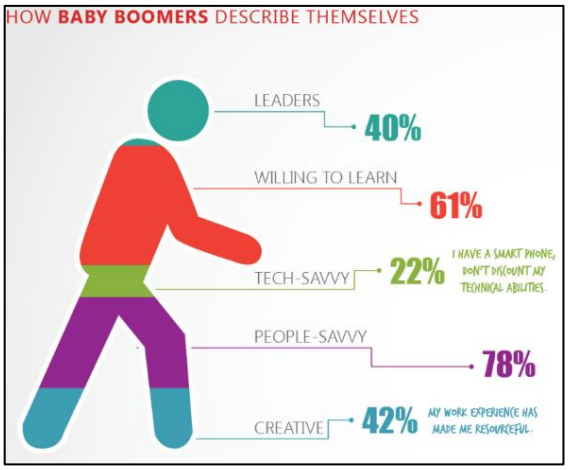
- Caution! Watch for bar charts with the baseline omitted (y-axis truncated)—meaning the y-axis does not start at zero!

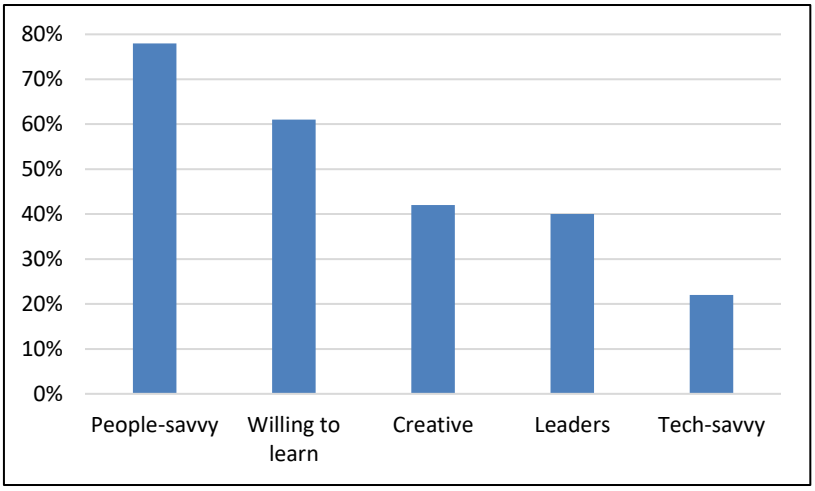




Pie Charts

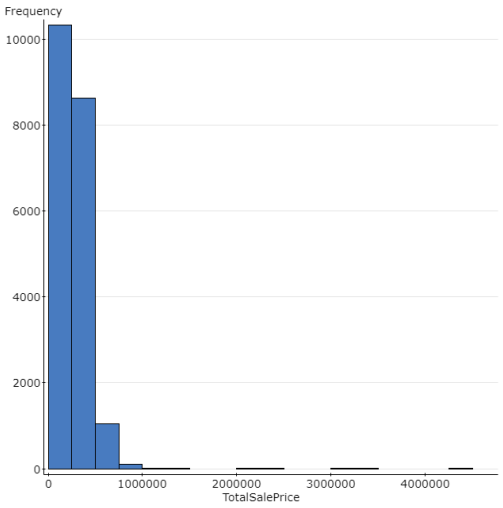
- Useful for displaying a single categorical variable
- One “wedge” for each category of a variable
- Size of wedge shows percent in each category
- Additional variables cannot be added via color, but you could compare pie charts across different levels of a second categorical variable
- Caution! Pie charts are often misused! They are used when not appropriate (e.g. when categories add to more than 100%) or visually distorted (e.g. 3-d pie charts, unusual shapes)



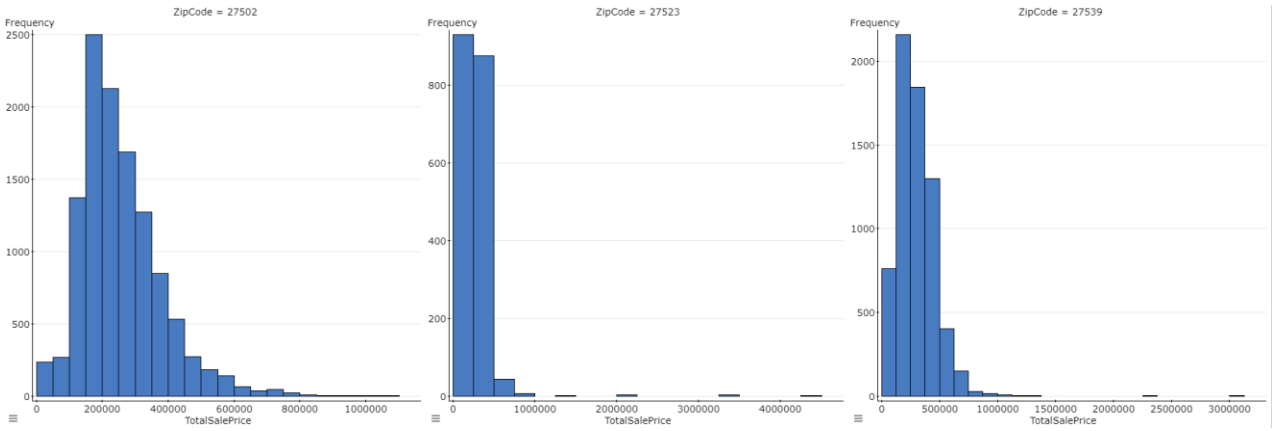


Notes for Lecture 2.3: Graphical Displays for Quantitative Data
Histograms and Distribution

Histograms



- Useful for displaying a single numeric variable
- Horizontal axis shows values of variable
- Bars represent ranges ("bins") of values
- Height of bar indicates how many [frequency] *or* percent of individuals in each bin
- Ex (at right): variable = sale price for homes in Apex, NC; looking at 1st bar, we see that there were over 10,000 homes that sold for somewhere between $0 and $250,000
- Additional variables cannot be added via color, but you could compare histograms across different levels of a second categorical variable



- Histograms allow us to understand the **<u>distribution</u>** of the data
  - 3 major elements of a distribution:

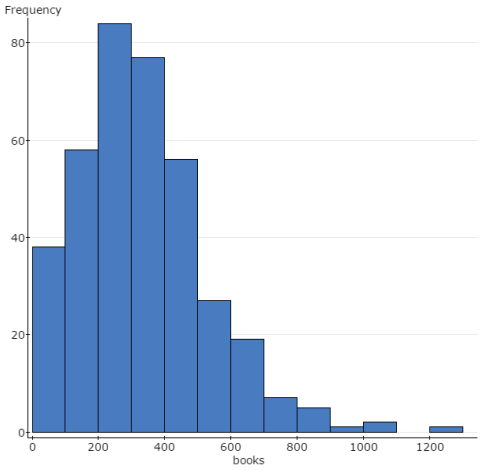
    1.

    2.

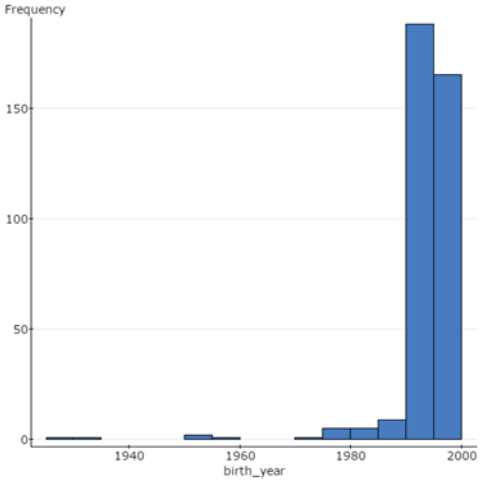
    3.

Shape of Distributions

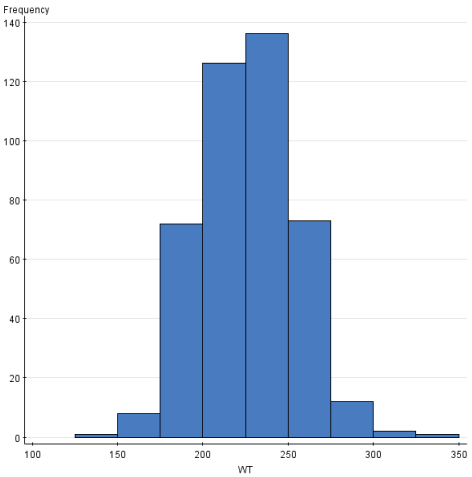
- Skewed Right
  - Long tail to the right
  - Generally because individuals are stacked up near a lower limit and unlimited on the upper end



- Skewed Left
  - Long tail to the left
  - Generally because individuals are stacked up near an upper limit and unlimited on the lower end

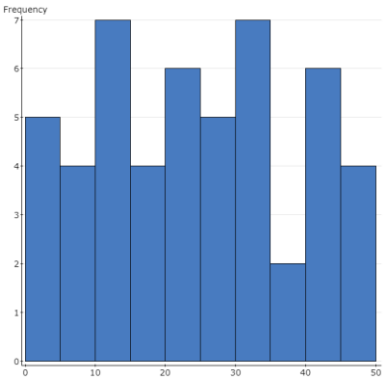


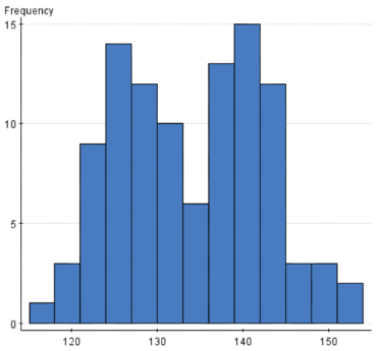
- Symmetric
  - Tails approximately equal in both directions
  - Major cluster far from limits on both ends



Shape of Distributions—Other Things to Consider

- Number of peaks (**<u>modes</u>**)





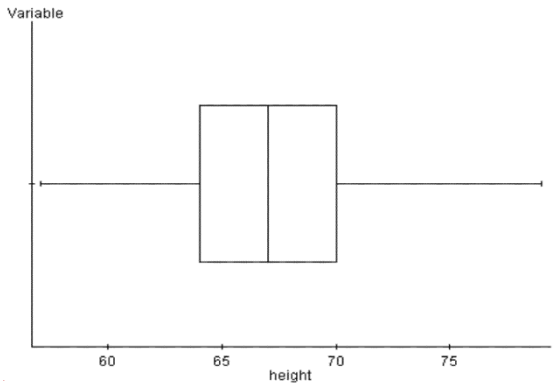
- Outliers—Unusual values that do not fit with the rest of the pattern
  - E.g. Total sale prices above $2,000,000 or Birth years before 1960
  - Why are they outliers?
    - Data entry errors
    - Invalid data points
    - Actual unusual values
  - How to deal with outliers (if you cannot remove them)?

Notes for Lecture 2.4: Graphical Displays for Quantitative Data
Boxplots and Other Graphs

Boxplots

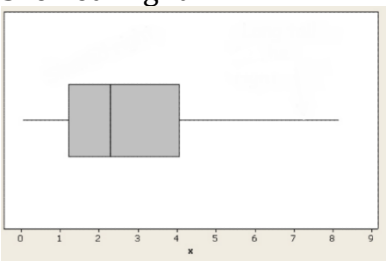
- Good for a first look at the data
- Visual display of the **5 number summary:**

1.
2.
3.
4.
5.

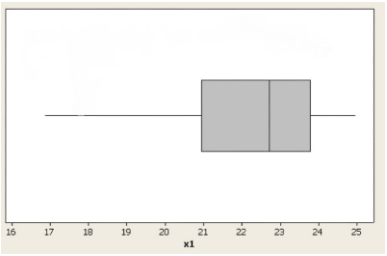


- The middle _____% of the data is located inside of the box
- Can help determine the shape of a distribution

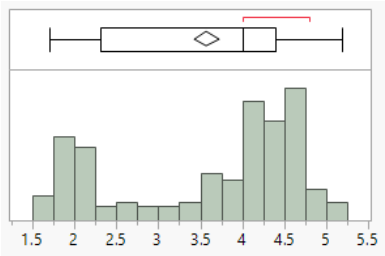
Skewed Right



Skewed Left



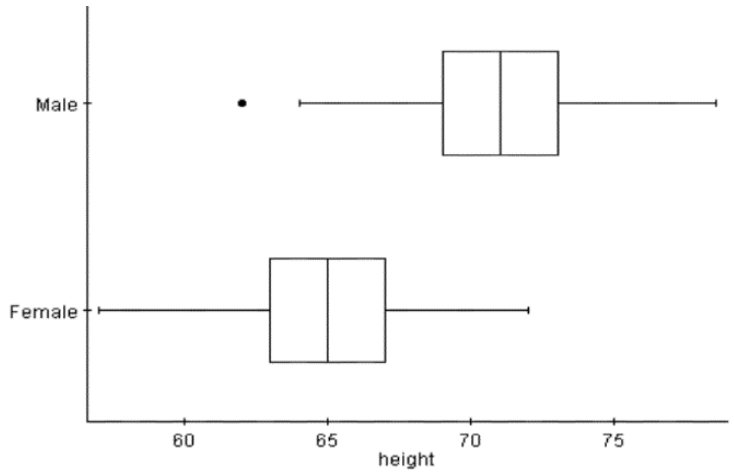
- We cannot determine if a distribution is multimodal from a boxplot



- Computer programs identify outliers
  - Box is not subject to outliers
  - Whiskers extend to largest/smallest non-outliers
  - Uses asterisk or dots to mark outliers

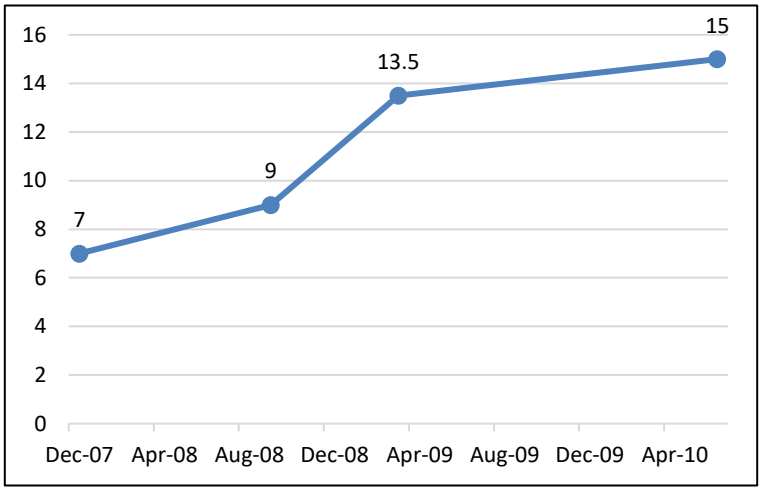
Side-by-side Boxplots

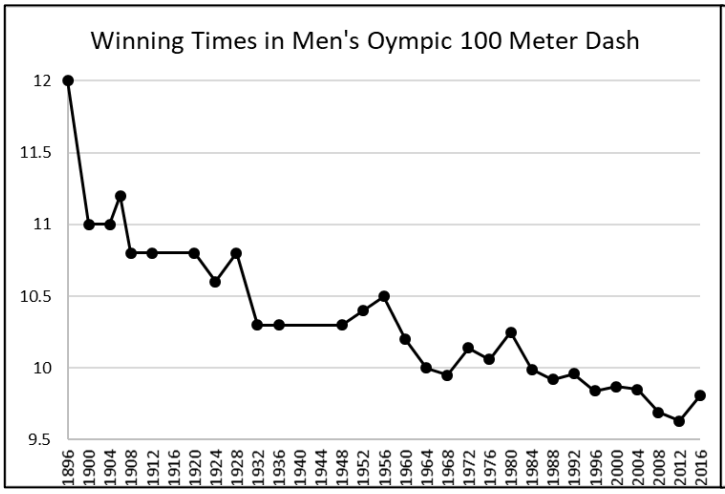
- Way to summarize a quantitative variable within levels of a categorical variable
- Useful for comparing distributions

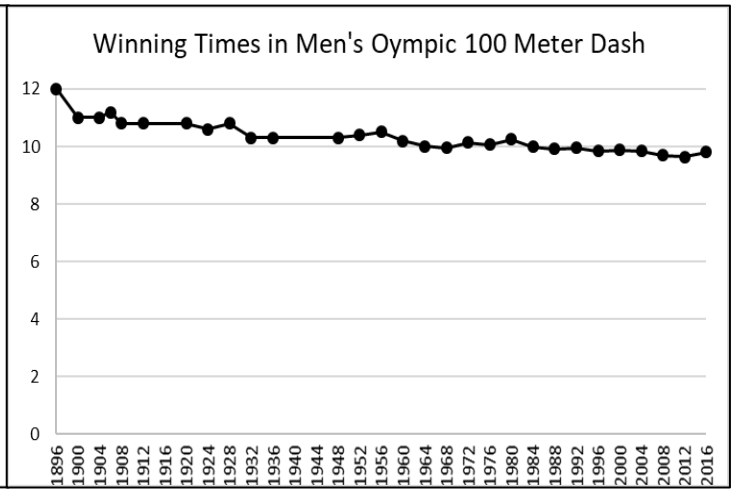


Time Plots

- Special type of scatterplot with *time* on the horizontal access
- **<u>Time series data</u>:** Measurements of a variable taken at regular intervals over time
  - Ex: monthly unemployment, daily market performance, progression of symptoms
  - Plot variable over time to observe trends
- Caution! Watch for time plots with the baseline omitted or otherwise distorted axes

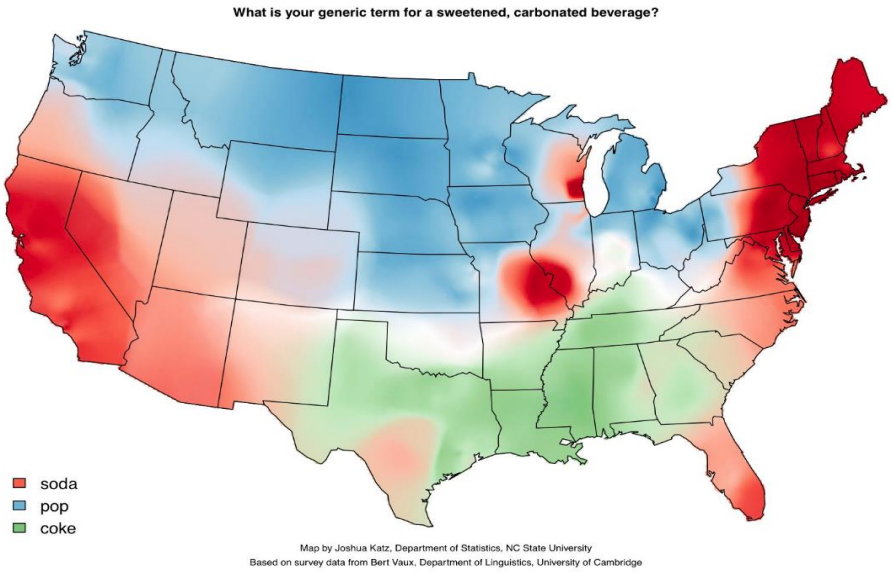




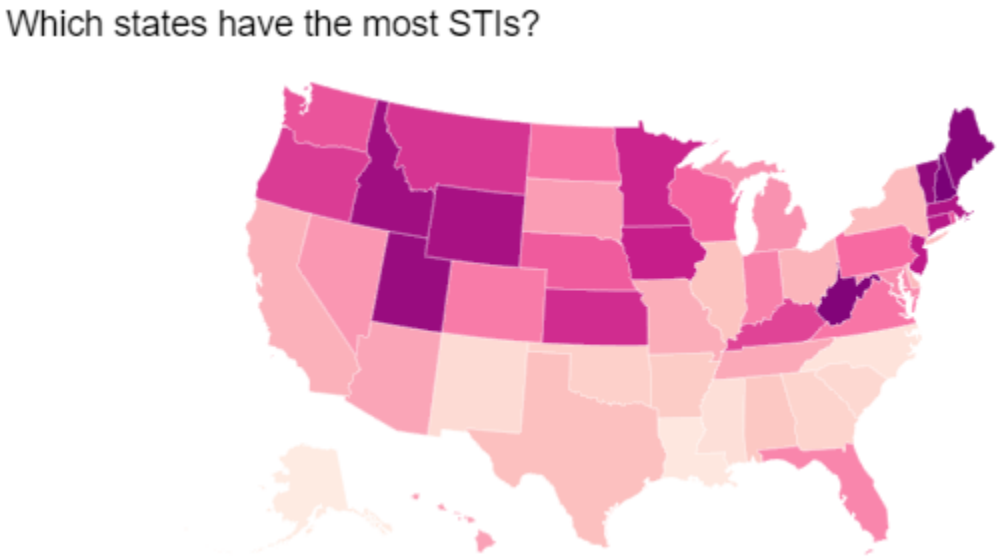


Heat Maps

- Useful for representing a variable that has a spatial element to it
- **<u>Spatial data</u> (<u>Geospatial data</u>)**: Data that involves physical space (e.g. size or shape) or geography (e.g. location)
- Uses color to represent different values of the variable
    - Darker colors indicate a higher or larger values

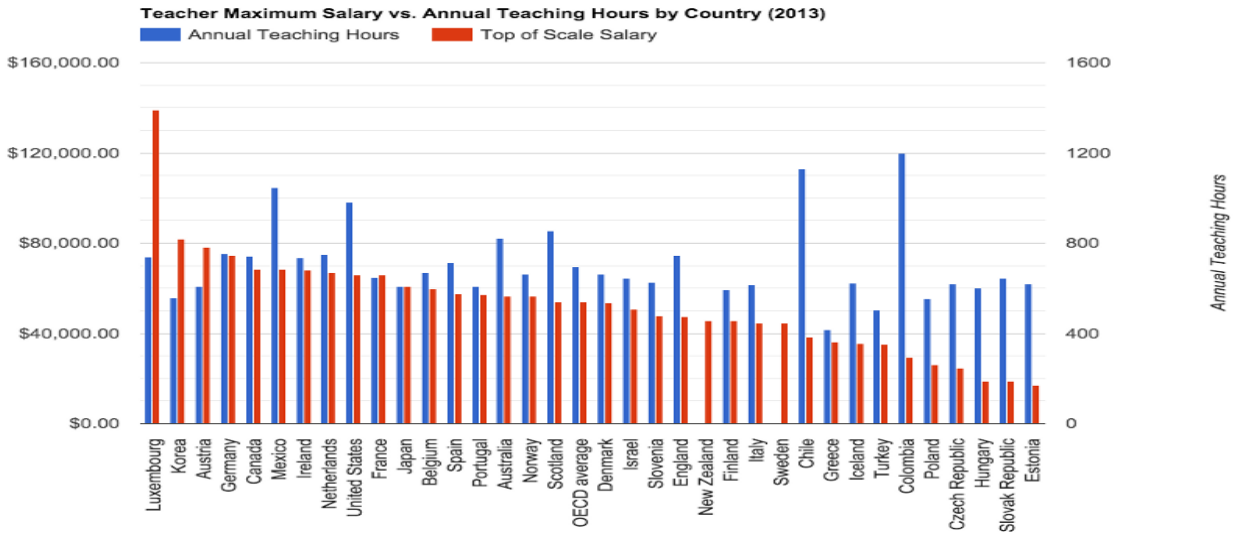


- Caution! Beware of heat maps that go against color convention; these can be confusing (e.g. blue = hot & red = cold) or misleading (e.g. using lighter color to indicate larger values)

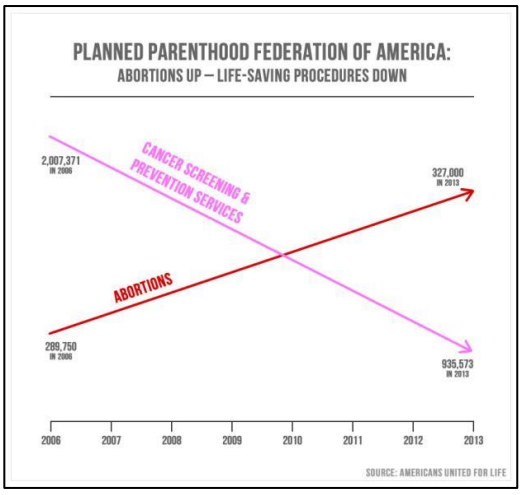


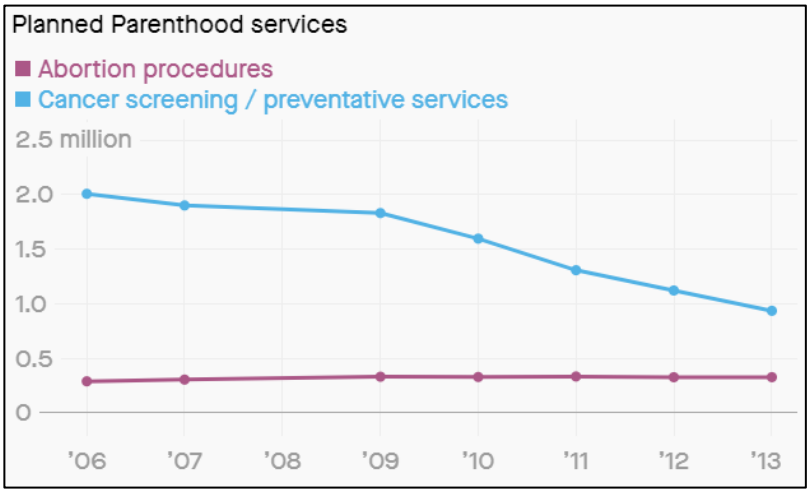
Caution! Graphs with Two Y-axes

- Sometimes variables measured on different scales will be put on the same graph
  - Done to pack a lot of information into a single graph
- Use caution when comparing the variables shown
- Look out for misleading or distorted axes!



- Ex (above): Can't say US has higher working hours than salary—that doesn't make sense! Can say US has longer working hours and lower max salaries than Luxemburg





Measures of Central Tendency

- **Mean**
  - *Population mean*: μ
  - *Sample mean*: $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- **Median**: Middle value in a data set when values are put in increasing order
- Benefits of the Mean:
- Problems with the Mean:
  - Sometimes misunderstood
  - Sensitive to skewed data

    Skewed Right:

    Skewed Left:

    Symmetric:
  - Sensitive to unusual values:

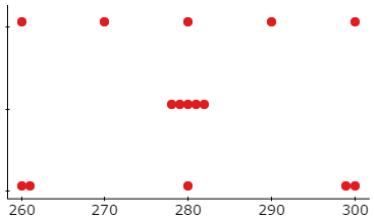
Measures of Variability

- Once we have an idea of a "typical" value, it is good to know about how much the individual values vary around this central value
- Example: 3 distributions with same mean (μ = 280) that look very different

260, 270, 280, 290, 300

278, 279, 280, 281, 282

260, 261, 280, 299, 300



Range	IQR	Std.Dev.

- **Range** = maximum – minimum
  - Spread of entire dataset
- **Interquartile range**: IQR = Q3 – Q1
  - Spread of middle 50%
- **Variance**: Summarizes distance between each individual and the mean
  - *Population Variance:* σ^2
  - *Sample Variance*: $s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$
- **Standard deviation**: s = square root of the variance
- Each measure of variability tells us how inconsistent the data values are

- Measures of variability are most useful for comparing distributions
- Benefits of using variance or standard deviation:

- Problems with using variance:
  - Variance (and standard deviation) are sensitive to unusual values or skew
  - Variance is measured in units squared (e.g. dollars^2); standard deviation is measured in the original units of the problem (e.g. dollars)

What does Standard Deviation Measure?

- Essentially represents the average distance from each point to the mean
- Simple example: A group of employees at a local company are paid by the hour. The amount they are paid for the six workers is $7, $8, $9, $10, $12, and $14.

7 8 9 10 12 14

When to Use Each Numeric Summary

- Mean (average value)

- Median (middle value)

- Range

- Standard deviation

- IQR

Notes for Lecture 2.6: Transformations of Numerical Summaries

Recall the wage example: A group of employees at a local company are paid by the hour. The amount they are paid for the six workers is $7, $8, $9, $10, $12, and $14

- What would happen if we gave everyone a $3 raise?

7 8 9 10 12 14

10 11 12 13 15 17

- What would happen if we doubled everyone's pay (multiplied by 2)?

7 8 9 10 12 14

14 16 18 20 24 28

Summary: Transformations

- Data is often transformed (adjusted, rescaled, standardized) to better represent the values or compare variables
- How will the measures change?
- Measures of variability and center respond differently

- Adding or subtracting:

- Multiplying or dividing: