Chapter 2: Data Visualization

- Data visualization through **graphics** is crucial for learning patterns in data.
- Example: Federal Election spending in 2012.
- Different graphical elements can allow information about several variables to be imparted in the graph.
- Compare Figures 2.1 and 2.2: An additional categorical variable (type of spending) can be introduced through the use of *color* in the graph.
- Figure 2.4 shows how still another categorical variable (office sought) can be incorporated through use of *panels* in the graphic.

Figure 2.1

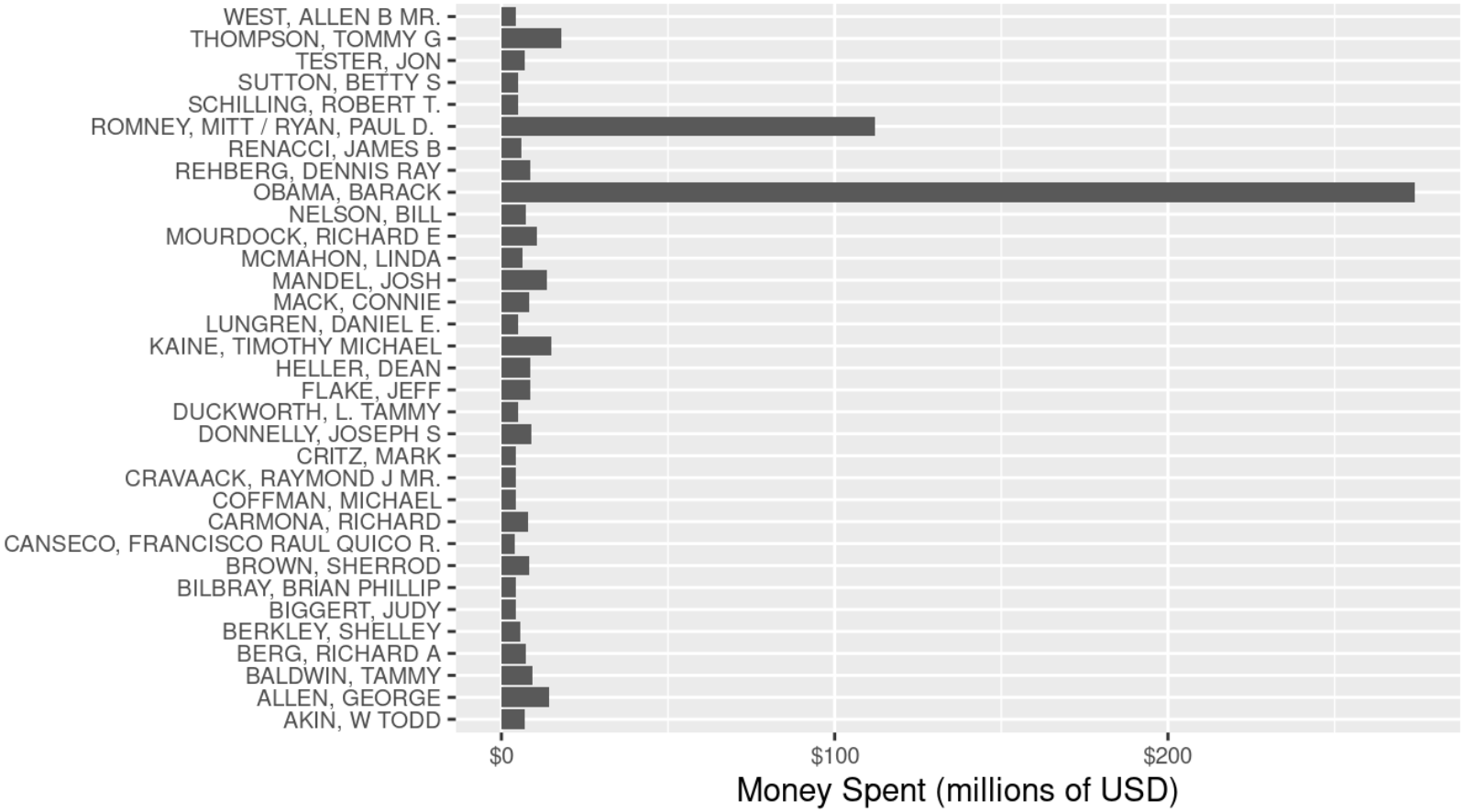


Figure 2.1 from MDSR textbook

Figure 2.2

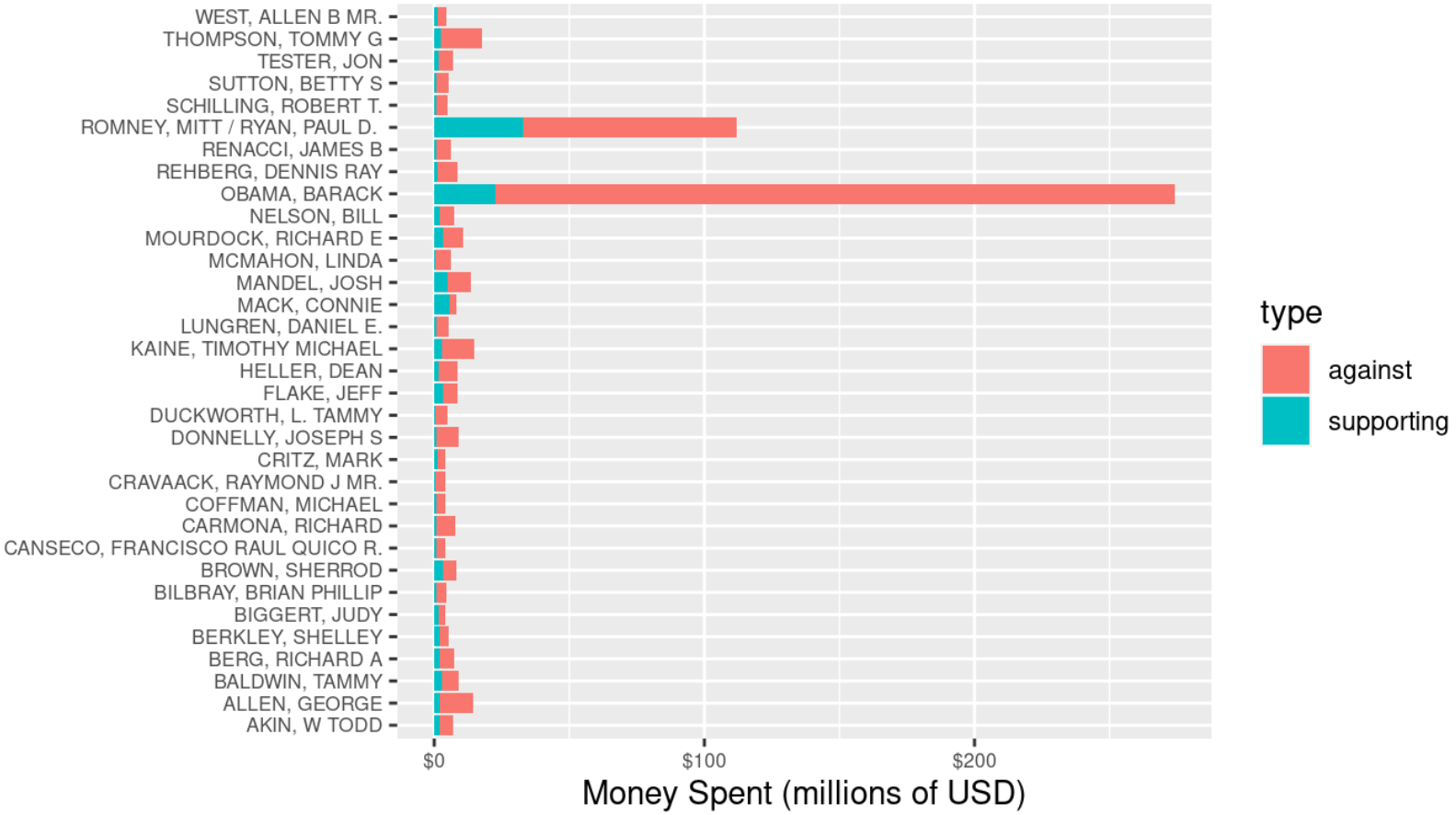


Figure 2.2 from MDSR textbook

Figure 2.4

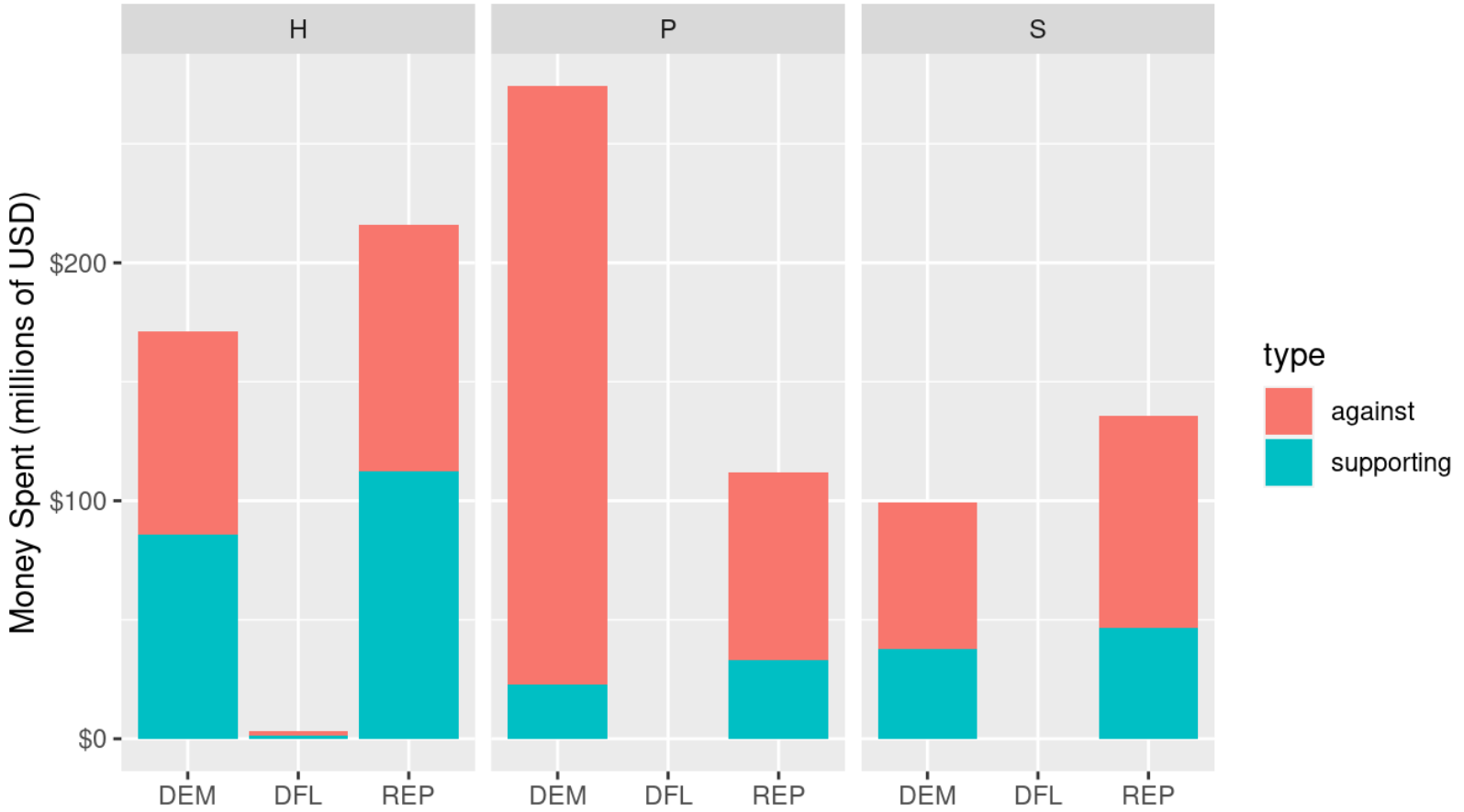


Figure 2.4 from MDSR textbook

Another example: Comparing Distributions using Graphs

- A distribution of a numerical variable shows the set of values the variable takes and how often it takes them.
- A *histogram* is a traditional way to show a distribution for a set of data.
- If we increase the number of classes in a histogram, the resulting plot begins to look smoother.
- In the limit, we get a completely smooth version of a histogram: a *density plot*.
- Figures 2.5 and 2.6 meant to display and compare two distributions: What are the differences?

Figure 2.5

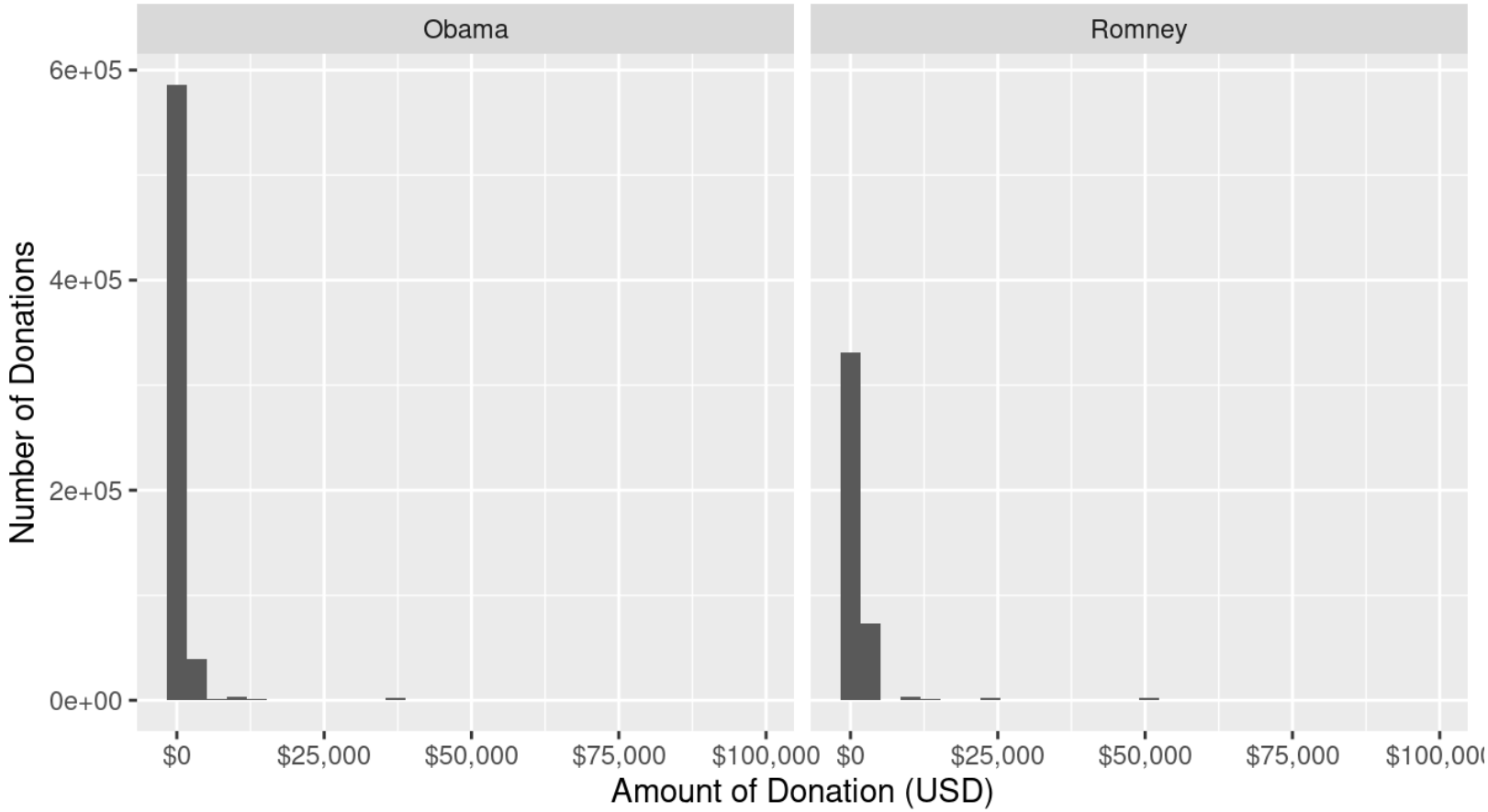


Figure 2.5 from MDSR textbook

Figure 2.6

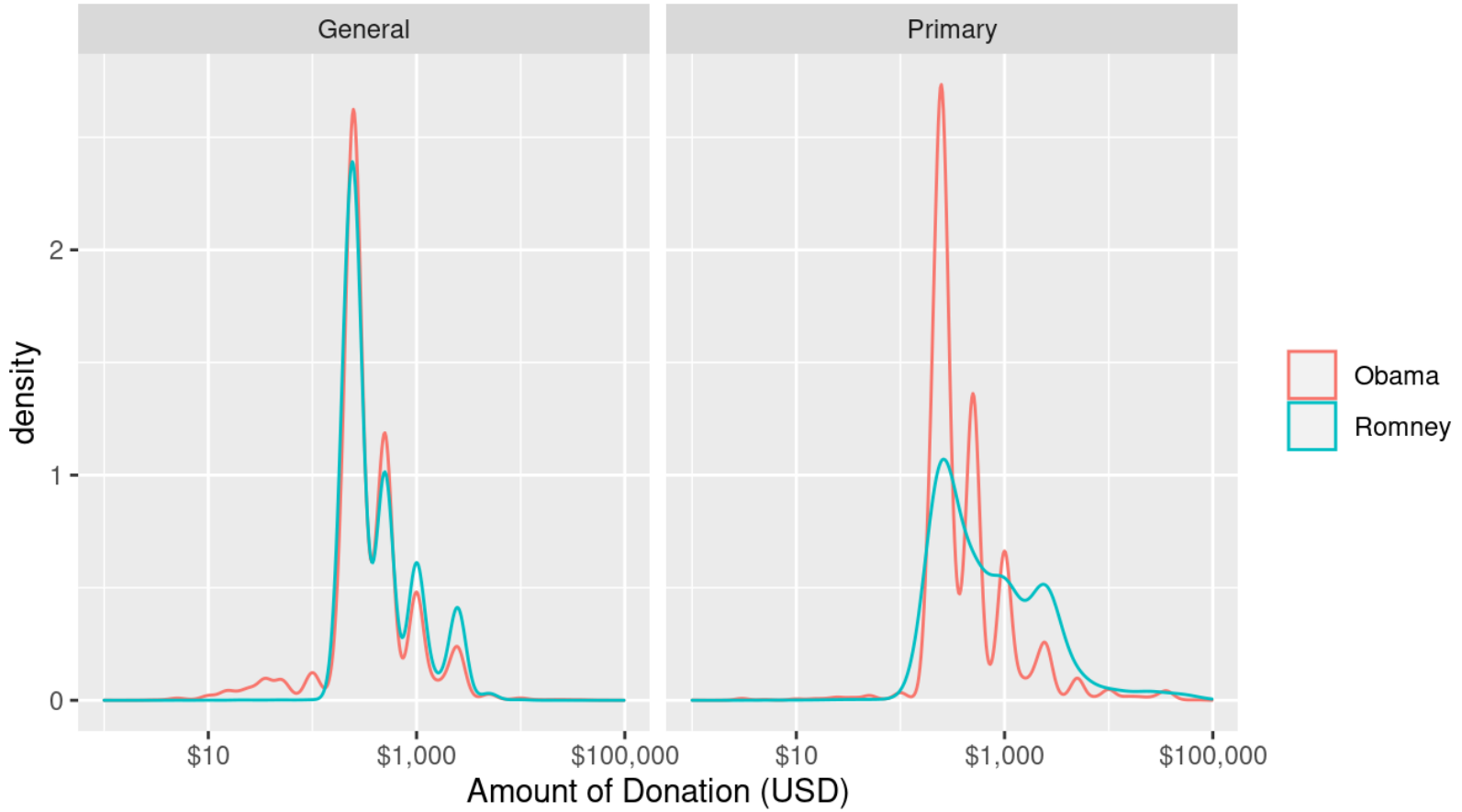


Figure 2.6 from MDSR textbook

Useful Tips for Displaying Distributions

- Density curves plotted instead of histograms: Easier to overlay density curves.
- Could put bars side-by-side, but it's not as natural.
- Two Panels: Distributions are conditional on a categorical variable (here, phase of campaign).
- Conditioning on a categorical variable via panels can be used on both histograms and density plots (and other graphs, such as boxplots).
- The numerical axis in Figure 2.6 is on a logarithmic scale.
- Plotting on a log scale can be very useful when a distribution is severely right-skewed (mostly small values, with very few larger values).

Relationships between Numerical Variables

- A *scatterplot* is a classic graph for showing the relationship between two numerical variables.
- But what variables should be plotted on the vertical and horizontal axes?

Rates vs. Counts

- Often we have a choice of whether to plot raw counts of a variable, or to scale the variable by dividing by some total amount.
- In many cases, *rates*, (i.e., proportions) are more meaningful than counts.

Rates and Counts

- Figure 2.7 shows two counts plotted against each other.
- For 2012 elections for the House of Representatives:
  - number of dollars spent supporting Democrats (on y-axis)
  - number of votes earned by Democrats (on x-axis)
- What is the conclusion about the association between spending and votes earned based on this plot?

Figure 2.7

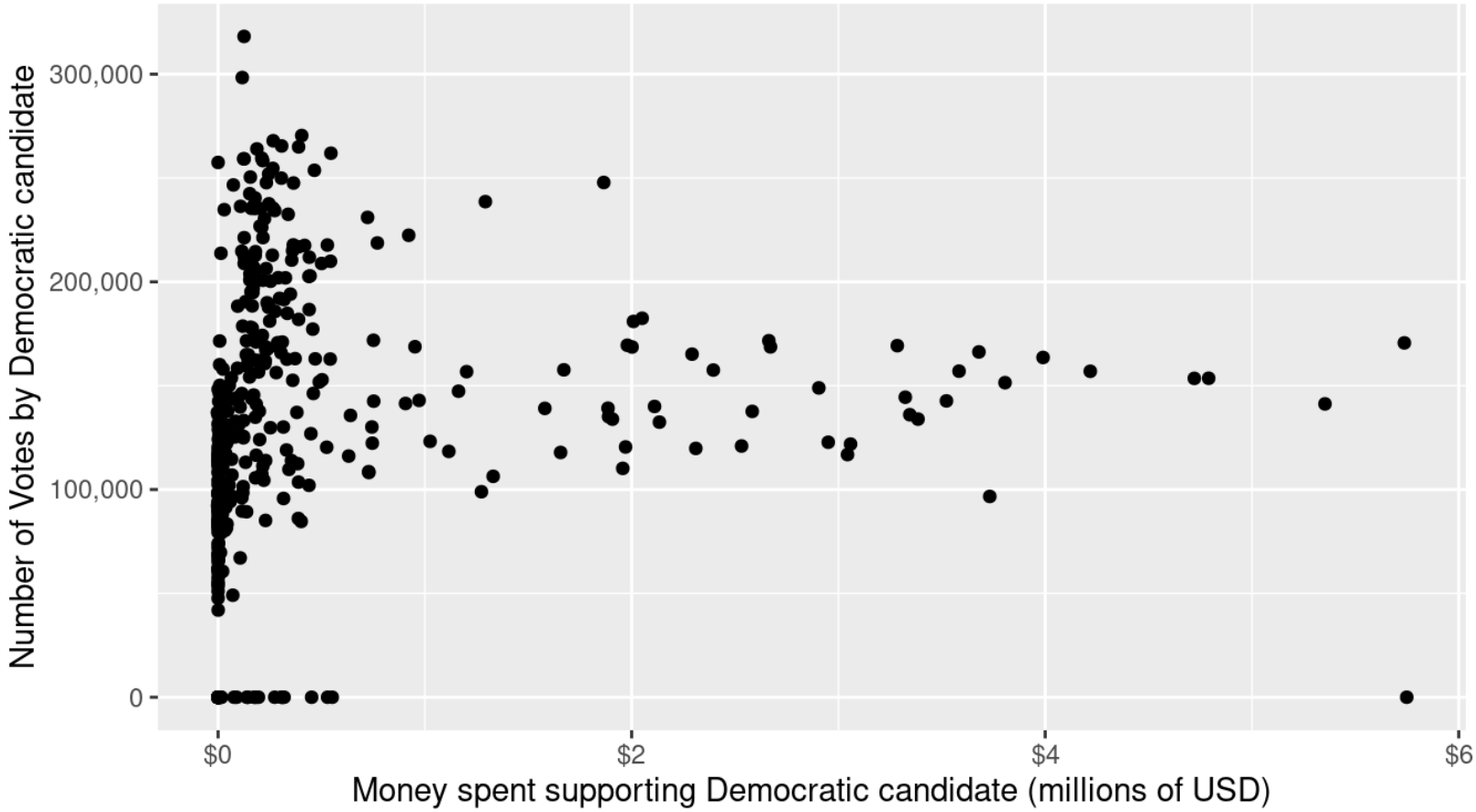


Figure 2.7 from MDSR textbook

Rates and Counts (Continued)

- Figure 2.8 shows two proportions plotted against each other.
- For 2012 elections for the House of Representatives:
  - proportion of dollars spent supporting Democrats (on y-axis)
  - proportion of votes earned by Democrats (on x-axis)
- What is the conclusion about the association between spending and votes earned based on this plot?

Figure 2.8

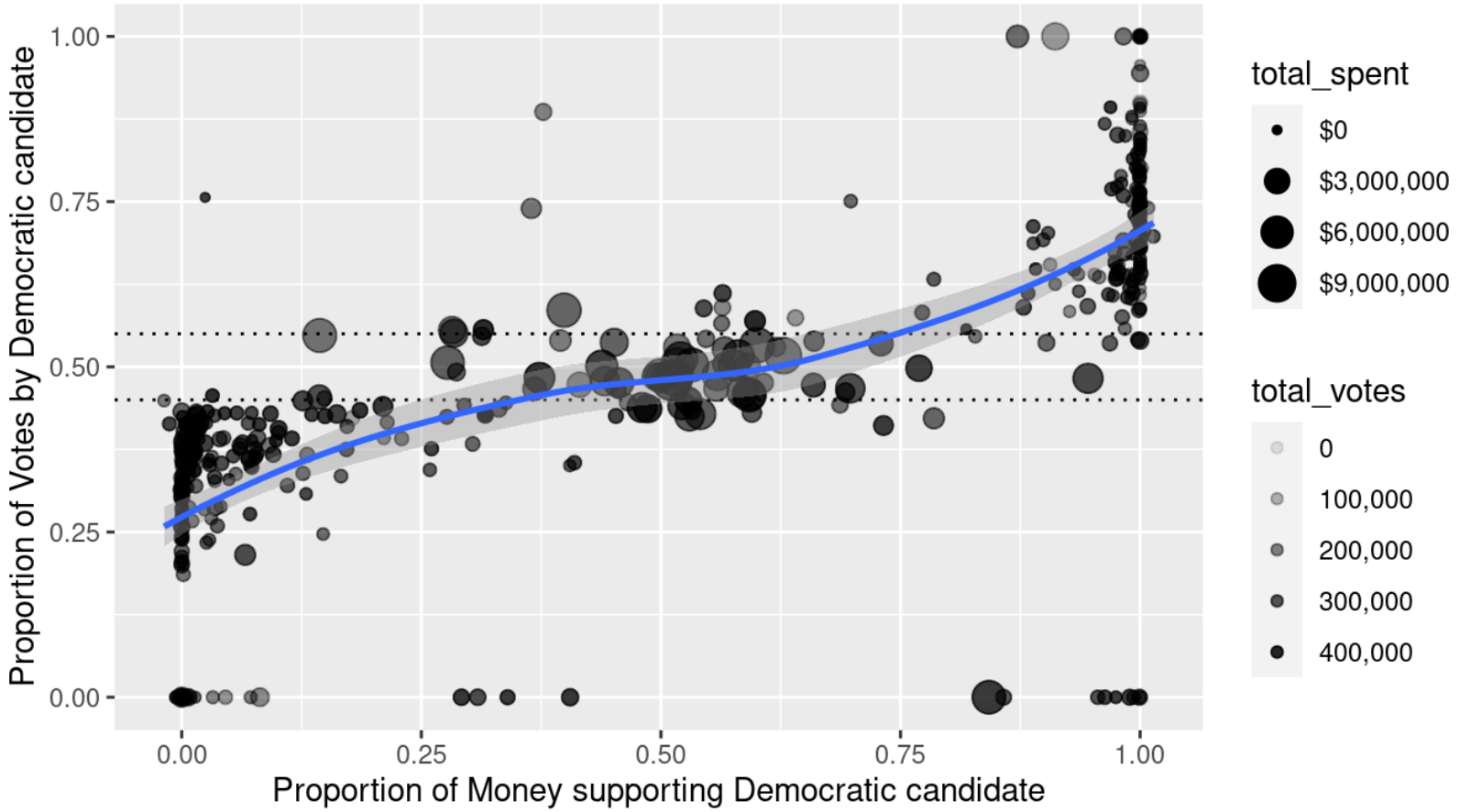


Figure 2.8 from MDSR textbook

Rates better than Counts Here?

- By using rates, we allow large and small districts to be treated equally.
- And will parties spend a lot of money on "sure-thing" races?
- Note that we can still incorporate information about the counts in Figure 2.8.
- The size and shading of the bubbles give information on the counts.

Composition of Data Graphics

- Variables can be represented on graphs via several different visual cues.
- Numerical variables can be represented on a graph by:
  1. Position
  2. Length
  3. Direction
  4. Shape
  5. Area
  6. Volume
- Note that it is easier for the eye to distinguish between values based of some of these cues (like Length) than other (like Area).
- Reason to prefer bar charts over pie charts?

Composition of Data Graphics (continued)

- Categorical variables can be represented on a graph by Shape (i.e., via different shapes of plotting characters (circles, triangles, boxes, etc.)) on a symbolic scatter plot.
- Visual cues such as Color and Shading might represent either categorical variables (e.g., via completely separate colors for the categories) or numerical variables (i.e., via a gradation of brightness on a scale of colors).
- People can have difficulty (or a complete inability, for colorblind people) judging differences in colors — possible reason to avoid *heat maps*.
- Also, for paper publications, printing in color is an extra expense, which could be a consideration in which graphic to use.

Choice of Coordinate System

- *Cartesian*: Familiar flat coordinate system with two axes (x-axis, y-axis).
- A 3-D coordinate system can show another variable on a third axis (not as easy for our eyes to fully comprehend 3 dimensions on a flat surface)
- *Polar*: Points identified by the radius and angle (useful for specialized variables such as directional data)
- *Geographic*: Shows locations on the curved surface of the Earth.
- Geographic displays are useful for global data where using a flat coordinate system can lead to distorted displays.
- *Radial*: Where the values are marked at positions on a circle, such as clock times or as pie charts.

Scales on Graphics

- Relates to how differences in data quantities are represented by differences in the visual cue.
- *Numeric* data could be plotted on a *linear*, *logarithmic*, or *percentage* scale.
- Logarithmic scale very useful for showing right-skewed data.
- Percentage scale valuable when multiplicative change is more meaningful than additive change.
- *Categorical* data could be *nominal* (no ordering to categories), or *ordinal* (ordered categories).
- *Time* is a numeric variable with special properties.
- Time is often measured using a combination of several units (year, month, day, hour, etc.)
- These units are on a "wrap around" scale that can show the *periodic* (cyclical) nature of the variable.

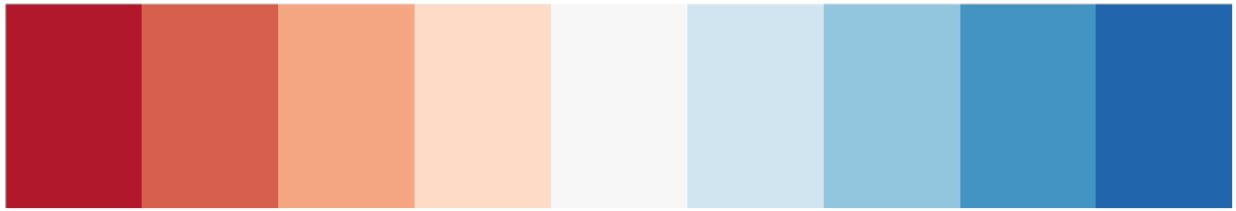
Multiples and Layers of Graphs

- *Facets* are several small multiples of the same plot.
- The different facets usually correspond to different values of some discrete or categorical variable.
- *Layers* are extra information plotted on top of an existing graphic.
- These can add extra information, but can become overdone and clutter up the display.
- *Animation* is when a sequence of related plots are shown quickly over time.
- This is useful when the data are gathered over time and can show temporal changes in data patterns.
- This is only useful on a screen display, not on a printed page display.

Color

- Color on a plot has its place sometimes, but can be overused.
- Avoid mixing red and green colors on plots to help red-green colorblind viewers.
- The R package `RColorBrewer` offers colorblind-safe color palettes to represent values that are:
  1. Sequential (varying in one direction), such as values starting at 0 and going up
  2. Diverging (varying in both directions), such as where there is one neutral category and other categories both above and below the neutral one
  3. Qualitative, where there is no ordering to the categories, so we just need several distinct colors
- See Figures 2.10 and 2.11 for a few color palettes available in `RColorBrewer`.

Figure 2.10



RdBu (divergent)

Figure 2.10 from MDSR textbook

Figure 2.11

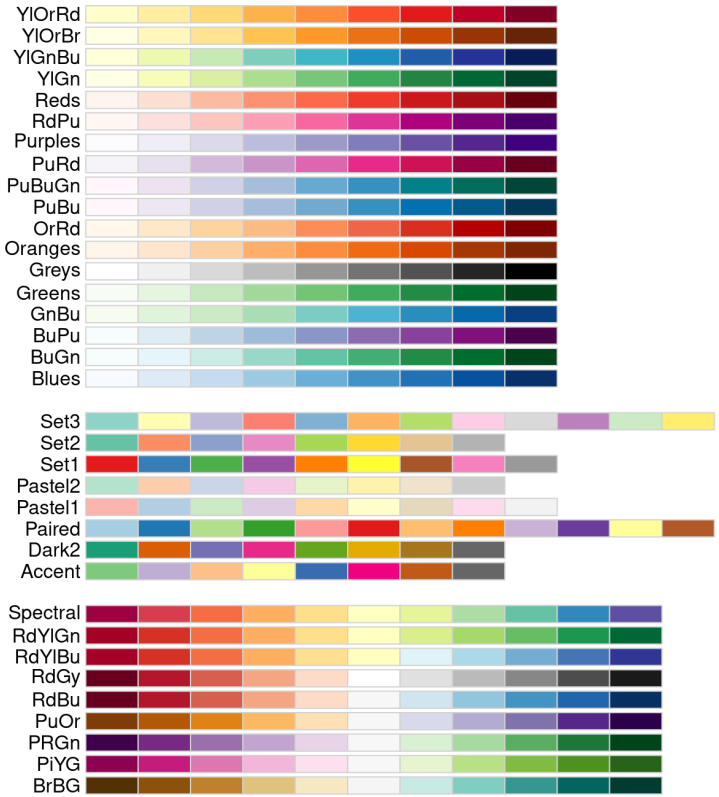


Figure 2.11 from MDSR textbook

Identifying Visual Cues and Other Aspects of Graphics

- Look at Figures 2.12, 2.13, 2.14, and 2.15.
- What are the visual cues in each plot?

Figure 2.12

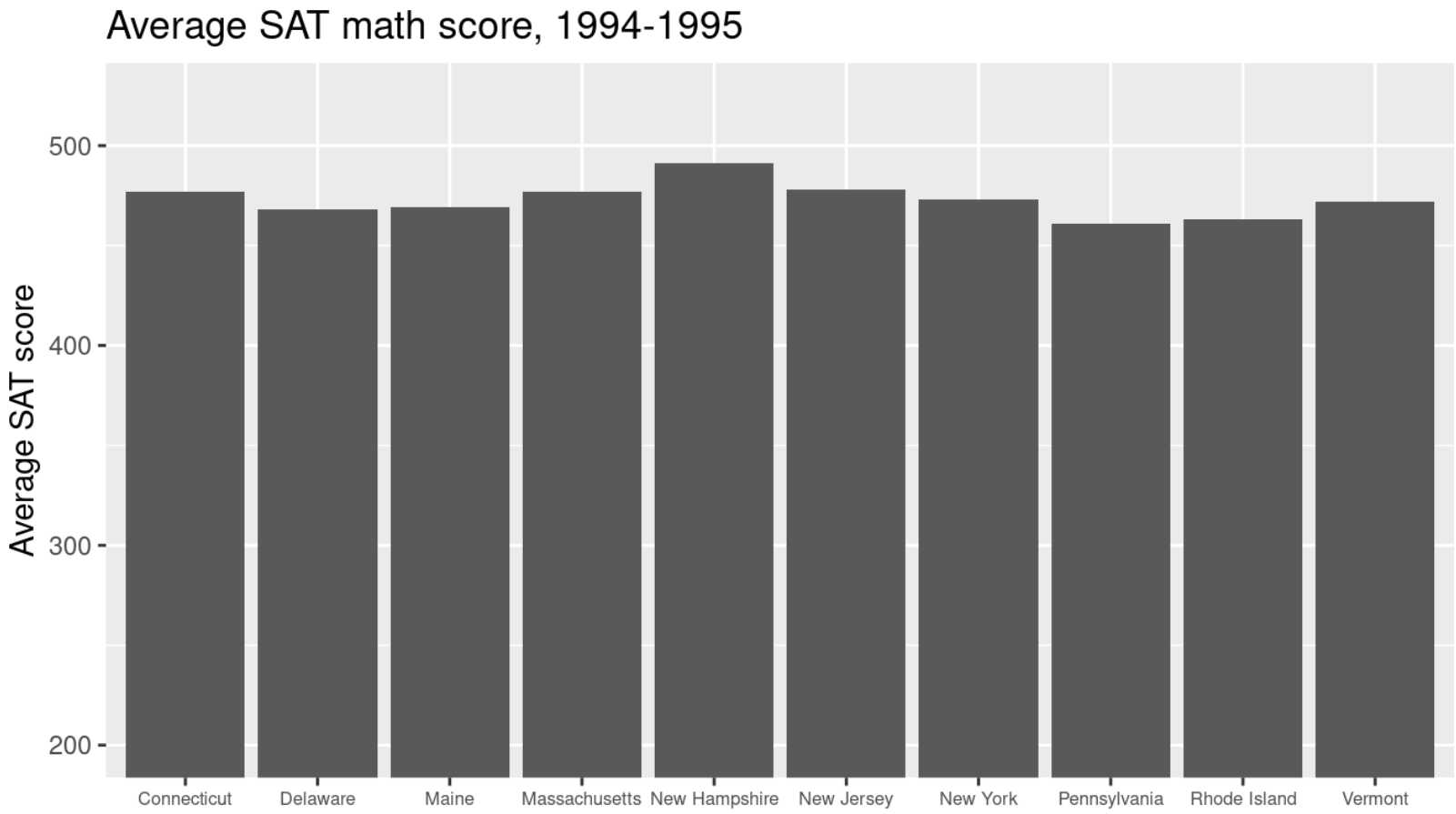


Figure 2.12 from MDSR textbook

Figure 2.13

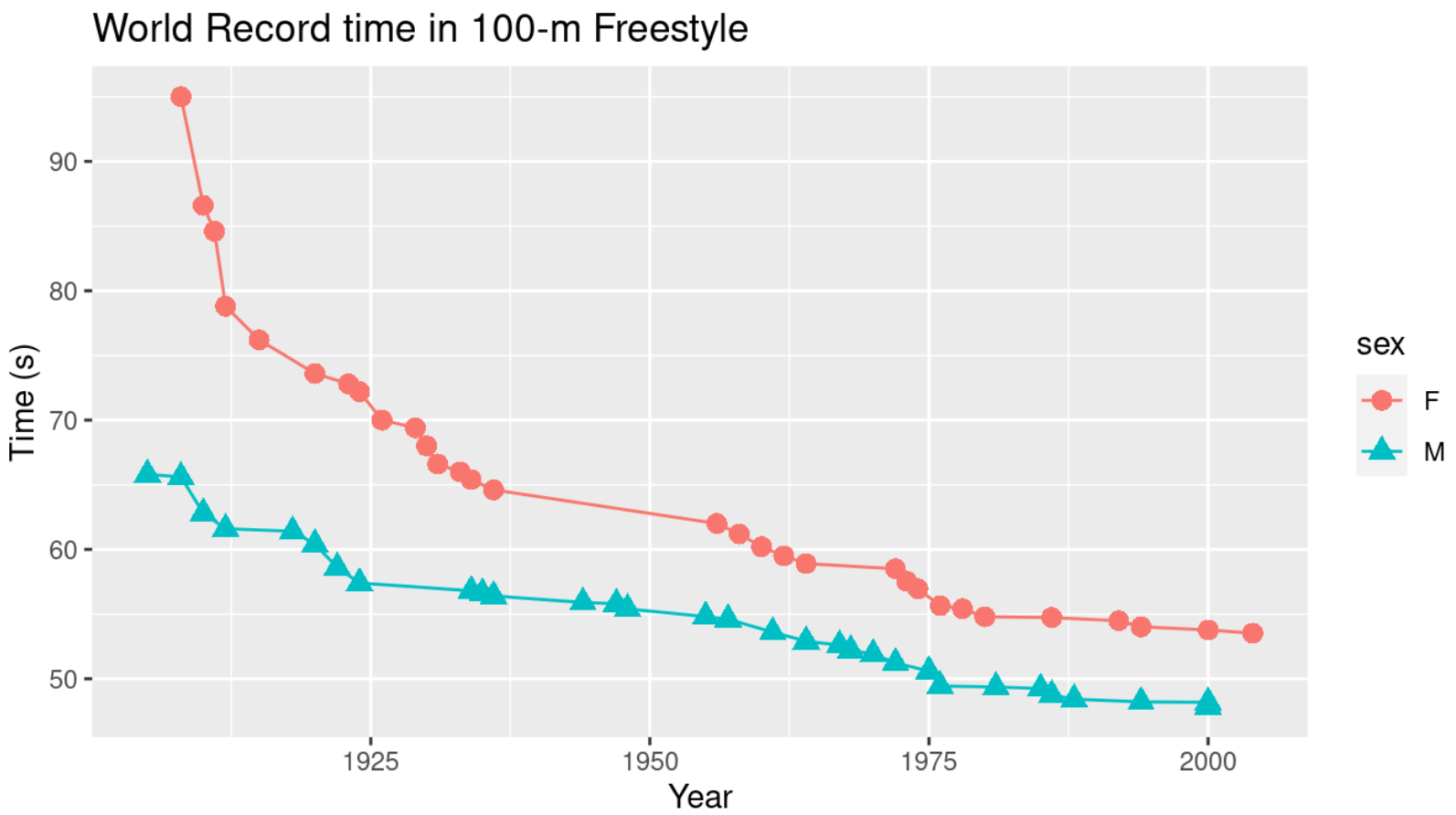


Figure 2.13 from MDSR textbook

Figure 2.14

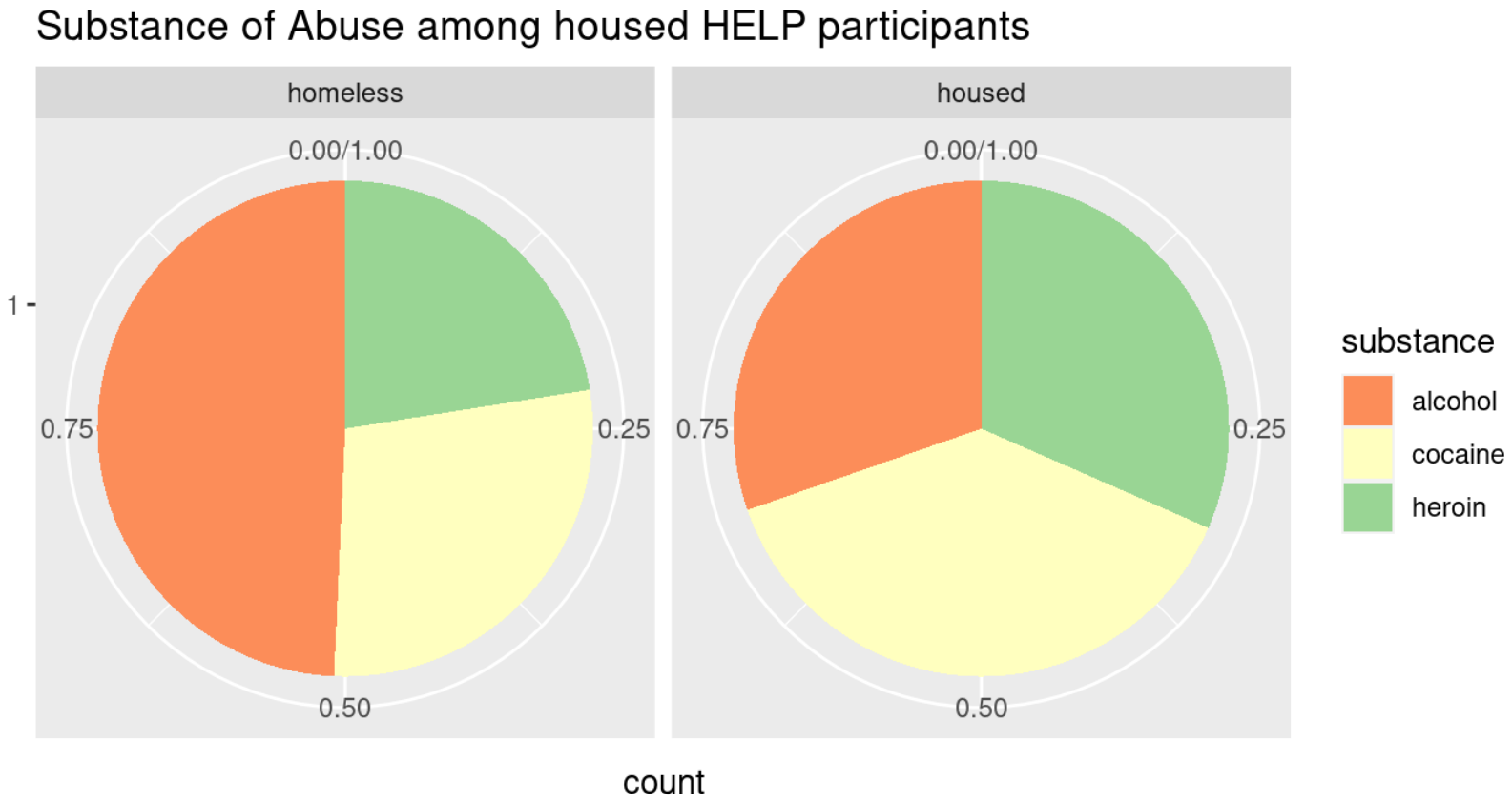


Figure 2.14 from MDSR textbook

Figure 2.15

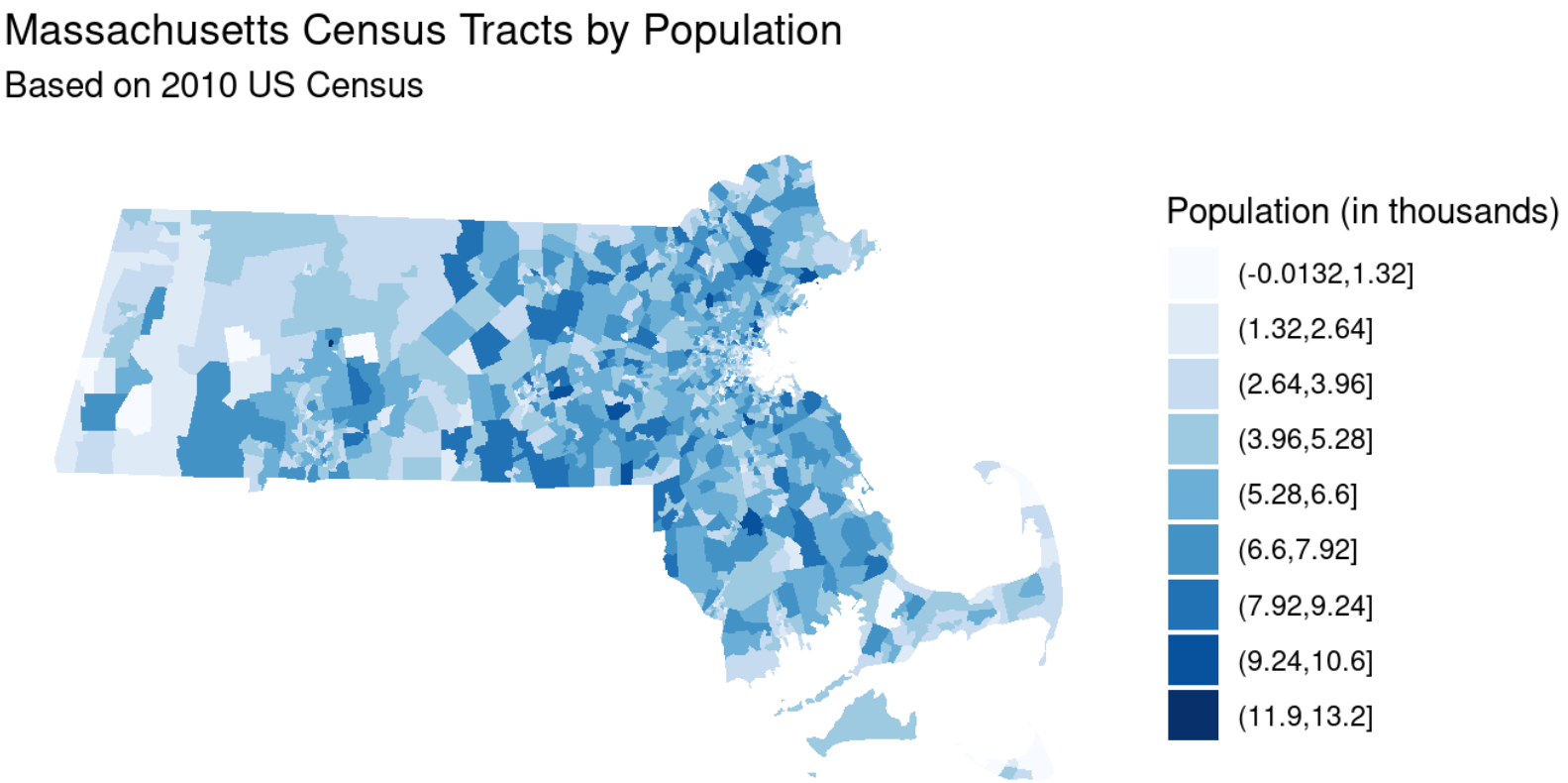


Figure 2.15 from MDSR textbook

Other Sections

- Read Section 2.3 for a famous example of a tragedy that might be been prevented with better data graphics.
- Read Section 2.4 for tips about effective presentations.