week2

Hakan Mehmetcik

Data visualization

well-designed data graphics are important

the analogy that creating data graphics is like cooking: Anyone can learn to type graphical commands and generate plots on the computer. Similarly, anyone can heat up food in a microwave. What separates a high-quality visualization from a plain one are the same elements that separate great chefs from novices: mastery of their tools, knowledge of their ingredients, insight, and creativity.

```
spent_tidy <- spent2 %>%
  group_by(party) %>%
  summarize(supporting = sum(supporting), against = sum(against)) %>%
  pivot_longer(-party, names_to = "type", values_to = "spent") %>%
  filter(spent > 1000)

ggplot(data = spent_tidy, aes(x = party, y = spent / 1e6, fill = type)) +
  scale_x_discrete(name = NULL) +
  scale_y_continuous(name = "Money Spent (millions of USD)", labels = scales::dollar) +
  geom_col()
```

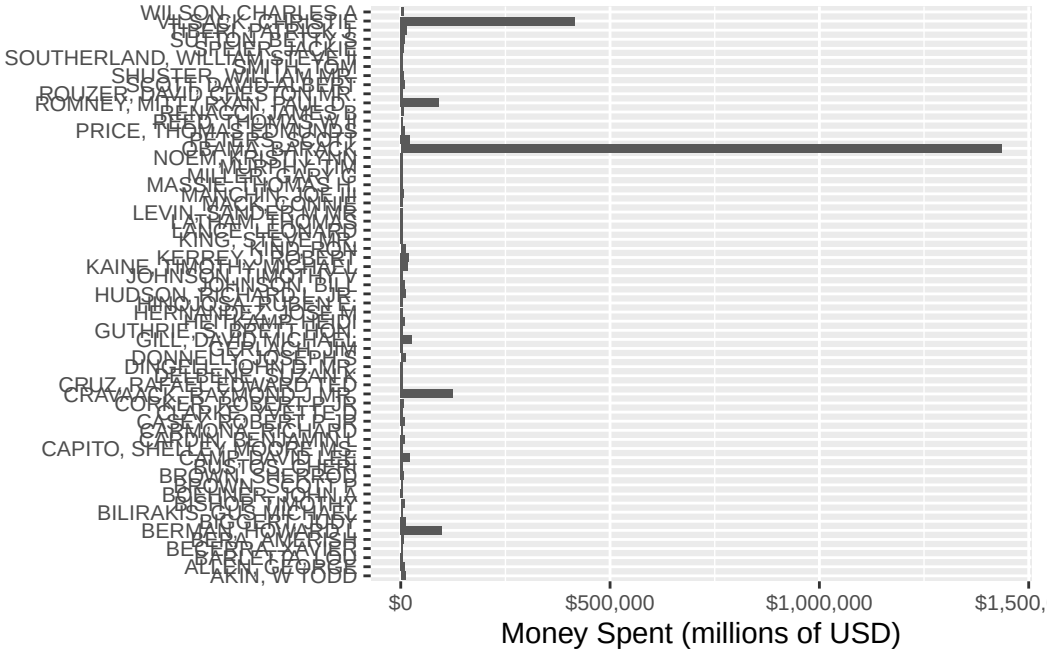
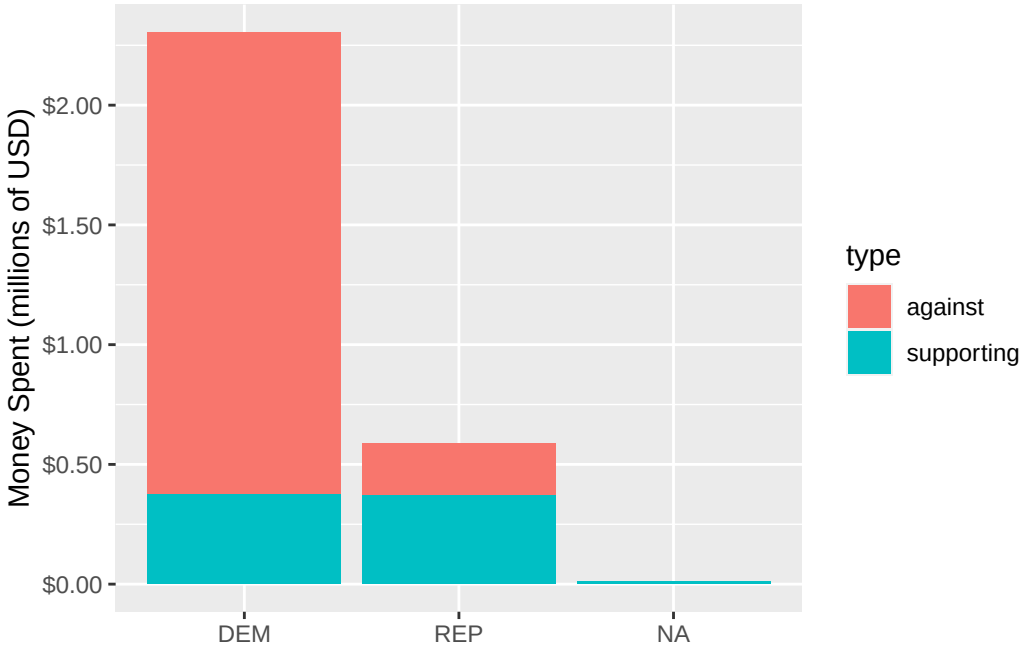
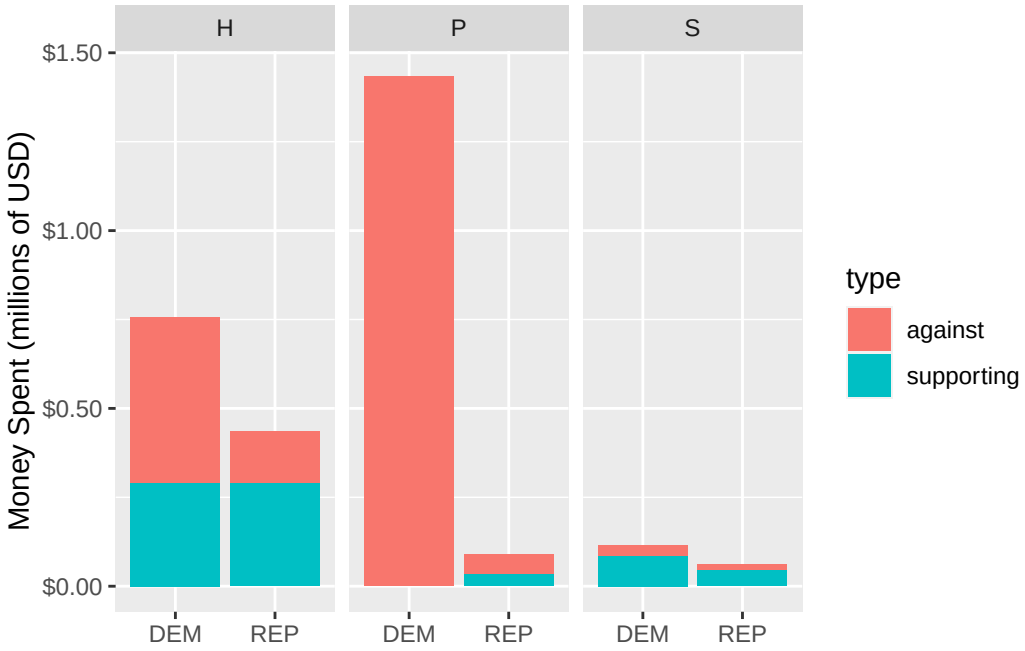


Figure 1: Amount of money spent on individual candidates in the general election phase of the 2012 federal election cycle, in millions of dollars. Candidacies with at least $4 million in spending are depicted.




```
spent2 %>%
  filter(!is.na(office)) %>%
  group_by(party, office) %>%
  summarize(supporting = sum(supporting), against = sum(against)) %>%
  pivot_longer(-c(party, office), names_to = "type", values_to = "spent") %>%
  filter(spent > 1000) %>%
ggplot(aes(x = party, y = spent / 1e6, fill = type)) +
  scale_x_discrete(name = NULL) +
  scale_y_continuous(name = "Money Spent (millions of USD)", labels = scales::dollar) +
  geom_col() +
  facet_wrap(~office)
```

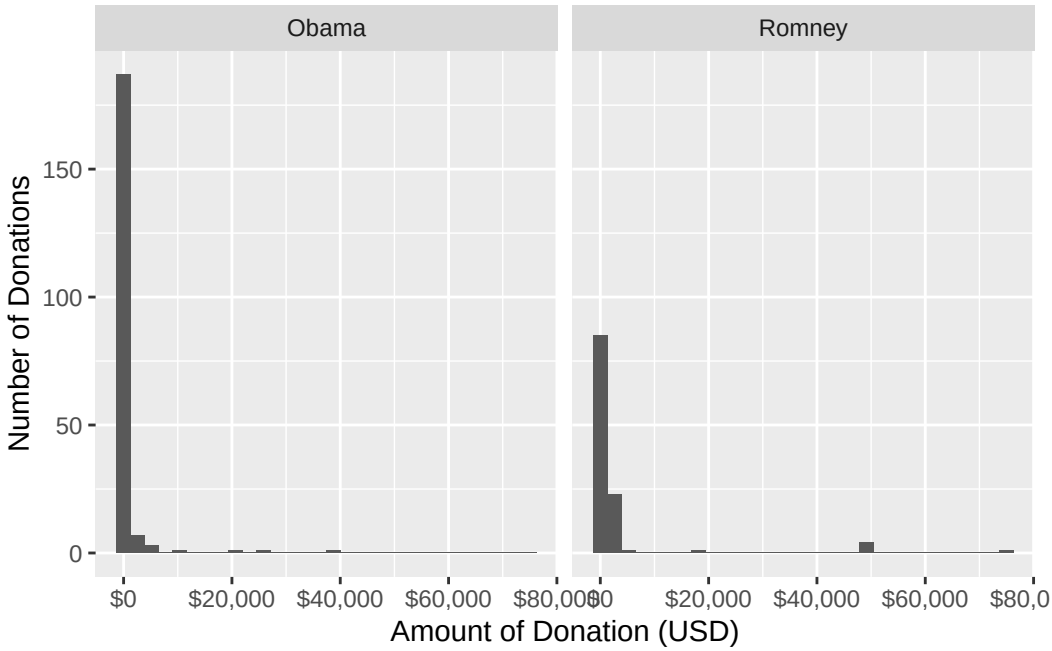
```
`summarise()` has grouped output by 'party'. You can override using the
`.groups` argument.
```



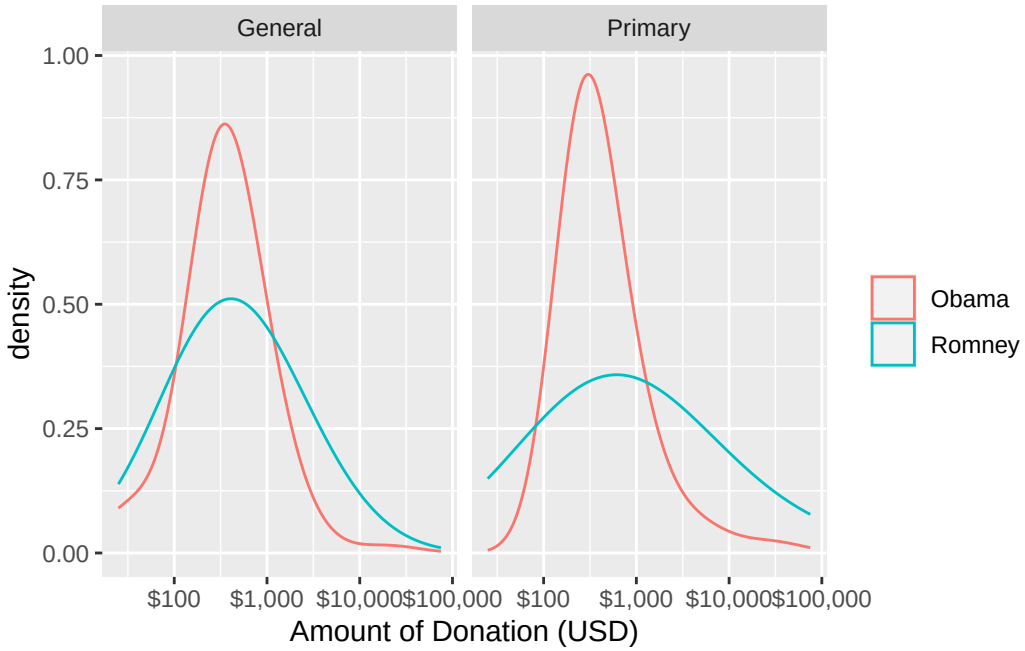

```
load(here::here("data", "fec12_donations.rda"))

ggplot(data = donations, aes(x = transaction_amt)) +
  geom_histogram() +
  scale_x_continuous(name = "Amount of Donation (USD)", labels = scales::dollar) +
  scale_color_discrete(name = NULL) +
  ylab("Number of Donations") +
  facet_wrap(~candidate)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(data = donations, aes(x = transaction_amt)) +
  geom_density(aes(color = candidate), adjust = 4) +
  scale_x_log10(name = "Amount of Donation (USD)", labels = scales::dollar) +
  scale_color_discrete(name = NULL) +
#  coord_trans(x = "log10") +
  facet_wrap(~phase)
```



In ggplot framework, data graphics can be understood in terms of four basic elements: visual cues, coordinate systems, scale, and context.

- **Visual Cues:** Visual cues are graphical elements that draw the eye to what you want your audience to focus upon

Table 2.1: Visual cues and what they signify.

Visual Cue	Variable Type	Question
Position	numerical	where in relation to other things?
Length	numerical	how big (in one dimension)?
Angle	numerical	how wide? parallel to something else?
Direction	numerical	at what slope? in a time series, going up or down?
Shape	categorical	belonging to which group?
Area	numerical	how big (in two dimensions)?
Volume	numerical	how big (in three dimensions)?
Shade	either	to what extent? how severely?
Color	either	to what extent? how severely?

Research into graphical perception (dating back to the mid-1980s) has shown that human beings' ability to perceive differences in magnitude accurately descends in this order (Cleve-

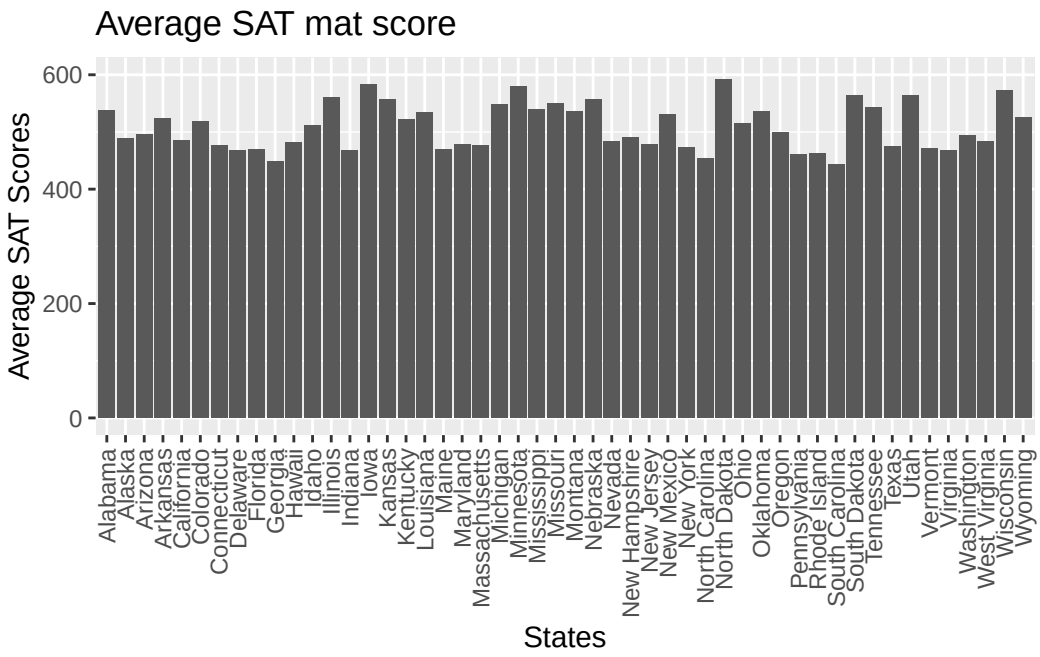
land and McGill 1984). That is, humans are quite good at accurately perceiving differences in position (e.g., how much taller one bar is than another), but not as good at perceiving differences in angles. This is one reason why many people prefer bar charts to *pie charts*. Our relatively poor ability to perceive differences in color is a major factor in the relatively low opinion of *heat maps* that many data scientists have.

- **Coordinate systems:** How are the data points organized? While any number of coordinate systems are possible, three are most common:
  - **Cartesian**: The familiar (x,y)-rectangular coordinate system with two perpendicular axes.
  - **Polar**: The radial analog of the Cartesian system with points identified by their radius and angle .
  - **Geographic**: The increasingly important system in which we have locations on the curved surface of the Earth, but we are trying to represent these locations in a flat two-dimensional plane.
- **Scale** translate values into visual cues. The choice of scale is often crucial. The central question is *how*does distance in the data graphic translate into meaningful differences in quantity? Each coordinate axis can have its own scale, for which we have three different choices:
  - **Numeric**: A numeric quantity is most commonly set on a *linear*, *logarithmic*, or *percentage* scale. Note that a logarithmic scale does not have the property that, say, a one-centimeter difference in position corresponds to an equal difference in quantity anywhere on the scale.
  - **Categorical**: A categorical variable may have no ordering (e.g., Democrat, Republican, or Independent), or it may be *ordinal* (e.g., never, former, or current smoker).
  - **Time**: A numeric quantity that has some special properties. First, because of the calendar, it can be demarcated by a series of different units (e.g., year, month, day, etc.). Second, it can be considered periodically (or cyclically) as a "wrap-around" scale. Time is also so commonly used and misused that it warrants careful consideration.
- **Context:** Context can be added to data graphics in the form of titles or subtitles that explain what is being shown, axis labels that make it clear how units and scale are depicted, or reference points or lines that contribute relevant external information. While one should avoid cluttering up a data graphic with excessive annotations, it is necessary to provide proper context.

Color is one of the flashiest, but most misperceived and misused visual cues. In making color choices, there are a few key ideas that are important for any data scientist to understand. First, while color can be visually appealing to humans, it often isn't as informative as we might hope. Second, approximately 8% of the population—most of whom are men—have some form of color blindness. To prevent issues with color blindness, avoid contrasting red with green in data graphics. As a bonus, your plots won't seem Christmas-y! Sidenote: The **RColorBrewer** package provides functionality to use these palettes directly in **R**

With a little practice, one can learn to dissect data graphics in terms of the taxonomy outlined above. For example, your basic scatterplot uses *position* in the *Cartesian* plane with *linear* scales to show the relationship between two variables. In what follows, we identify the visual cues, coordinate system, and scale in a series of simple data graphics.

```
# use SAT score for visualization
ggplot(data=SAT, aes(x=state, y=math))+
  geom_col()+
  scale_x_discrete("States")+
  scale_y_continuous(name="Average SAT Scores", limits=c(0, 600))+
  ggtitle("Average SAT mat score")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



This plot uses the visual cue of *length* to represent the math SAT score on the vertical axis with

a *linear*scale. The *categorical* variable of `state` is arrayed on the horizontal axis. Although the states are ordered alphabetically, it would not be appropriate to consider the `state` variable to be ordinal, since the ordering is not meaningful in the context of math SAT scores. The coordinate system is *Cartesian*, although as noted previously, the horizontal coordinate is meaningless. Context is provided by the axis labels and title.