Class 5: Visualizing data

API-201
Quantitative Analysis and Empirical Methods I

Profs. Goel and Taylor
Harvard Kennedy School

Agenda

1. Telling stories with graphs
2. Principles of data visualization

Objectives of visualization

Some principles of visualization

Graphs are typically about **comparisons**; make it easy for the viewer to make them.

Graphs should be made as **simple** as possible, but no simpler.
[Maximize the data-to-ink ratio.]

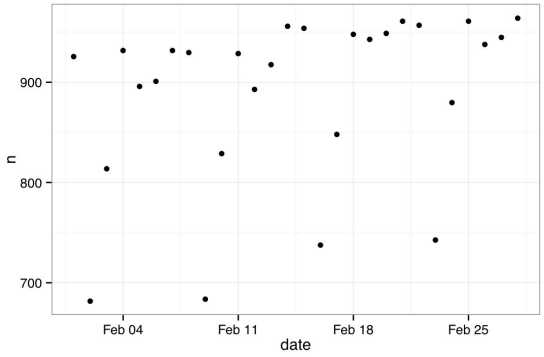
All graphs tell a **story**; don't tell a misleading one.

Plotting parameters

Plot type
Orientation
Scale & order
Color
Annotations & labels
Size & aspect ratio

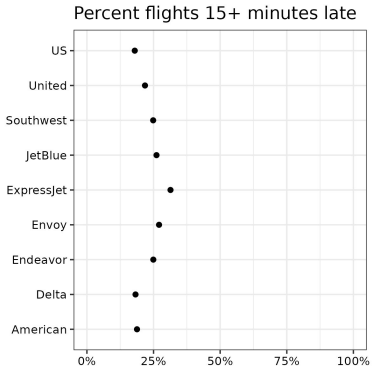
Daily volume of flights

[library(nycflights13)]



Flight delays

[library(nycflights13)]



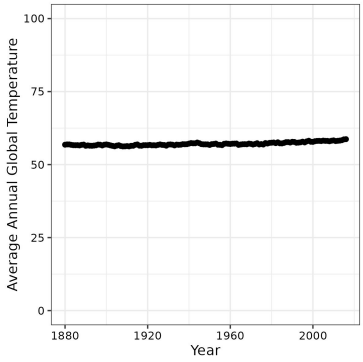
Sketch out a new plot
[but don't change the underlying data!]
to tell a new story below.
[Discuss with your neighbor.]

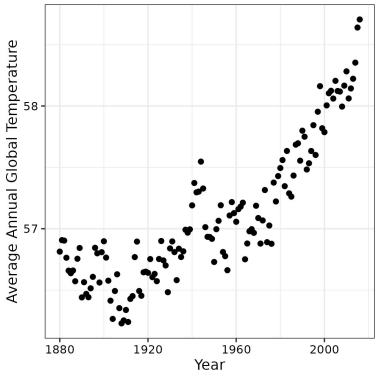
Beware of misleading plots

Plots can be used to tell misleading stories, even when the underlying data are correct.

Global temperatures over time

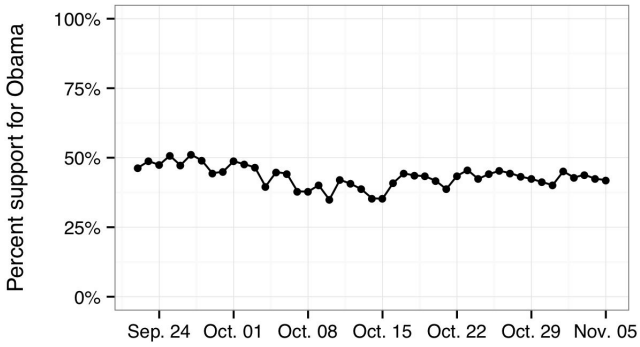
Plots can be used to tell misleading stories, even when the underlying data are correct.





Voter intent over time

2012 presidential campaign



Returning to stop-and-frisk

Hit rate by race and precinct

```
data %>%
  group_by(race, precinct) %>%
  summarize(hit_rate = mean(weapon))
```

This command would produce 100+ numbers, corresponding to…

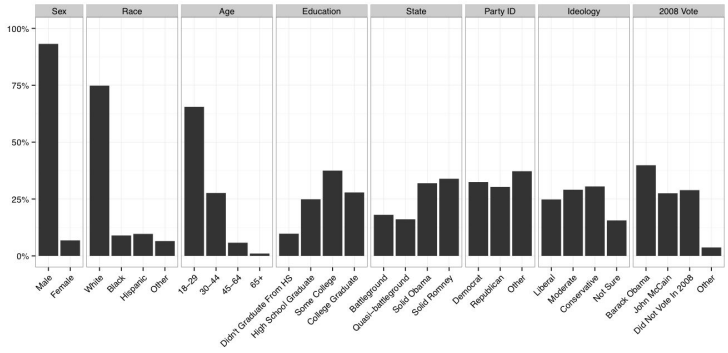
How would you visualize the resulting 100+ numbers to investigate the question of discrimination and communicate the results? [Discuss with your neighbor.]

Plotting distributions

Histogram
Density
Cumulative distribution function (CDF)
Box plot

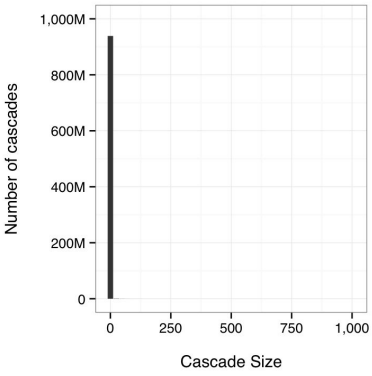
Demographic distribution

Xbox election survey participants



Cascade size distribution

Twitter

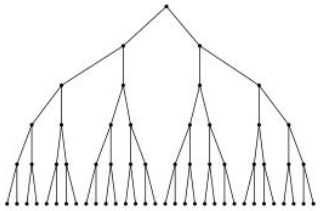


A structural definition of virality

Definition. The *structural virality* of a diffusion tree is the average distance between all pairs of nodes in the cascade.

Low "virality"

High "virality"



Popularity vs. virality

Twitter

What type of plot would you make to examine the relationship between popularity and virality?

[Discuss with your neighbor.]

Key takeaways

- Visualization is about exploration and communication.
- Graphs are typically about comparisons; make it easy for the viewer to make them.
- All graphs tell a story; be thoughtful about your design choices.