

GRD 610A Data Visualization II

Labels

Jenn Schilling

February 24, 2021

Today

- Data Visualization of the Week
- Discussion of Chapter 4: Of Conjectures and Uncertainty of *The Truthful Art* (Cairo)
- Lab on adding axis labels, titles, legends, labeling points, and annotating plots - Chapter 5 of *Data Visualization* (Healy)
- Homework Assignment #3
- Midterm

Of Conjectures and Uncertainty

Conjecture

Definition: "a guess about something based on how it seems and not on proof" (Cambridge Dictionary)

Requirements (Cairo, 2016, pp. 102-104)

- Makes sense
- Testable
- Made of several components
- If any component changes, the conjecture becomes useless

Of Conjectures and Uncertainty

Hypothesis

Definition: "a conjecture that is formalized to be tested empirically" (Cairo, 2016, p. 105)

Variables

- Independent (predictor/explanatory)
- Dependent (outcome/response)

Variable Scales

- Nominal or Categorical
- Ordinal
- Interval
- Ratio

Variable Types

- Discrete
- Continuous

Of Conjectures and Uncertainty

Study

"Test it [the hypothesis] against reality" (Cairo, 2016, p. 107)

Types of Studies

- Observational Study
- Cross-Sectional Study
- Longitudinal Study

Population versus Sample

Extraneous Variables

- Confounding variable
- Lurking variable

A **controlled experiment** goes beyond an observational study and tries to account for confounding variables.

Of Conjectures and Uncertainty

Uncertainty

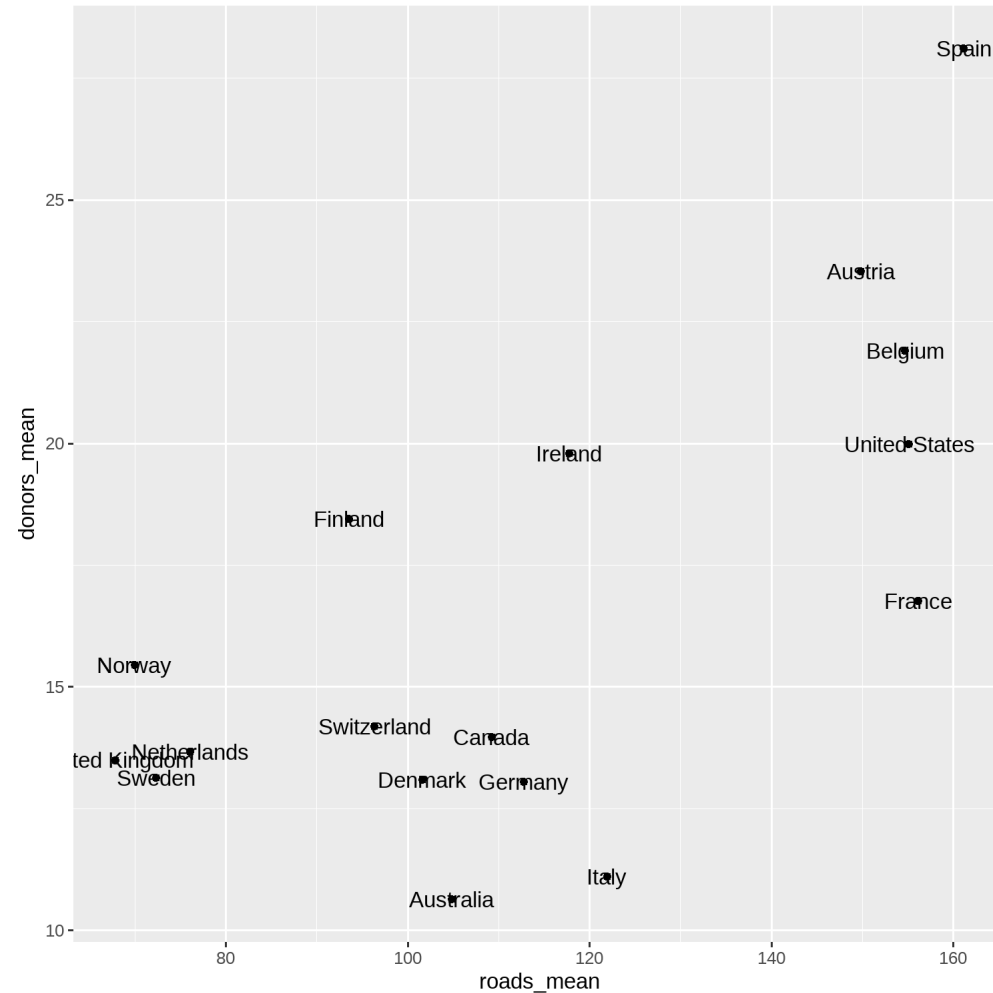
"Data always vary randomly because the object of our inquiries, nature itself, is also random."
(Cairo, 2016, p. 114)

Sample Variation is the variation or change in a statistic or measurement from sample to sample.

15 Minute Break

15:00

Chapter 5: Graph Tables, Add Labels, Make Notes



Building a Labeled Plot

```
library(ggrepel)

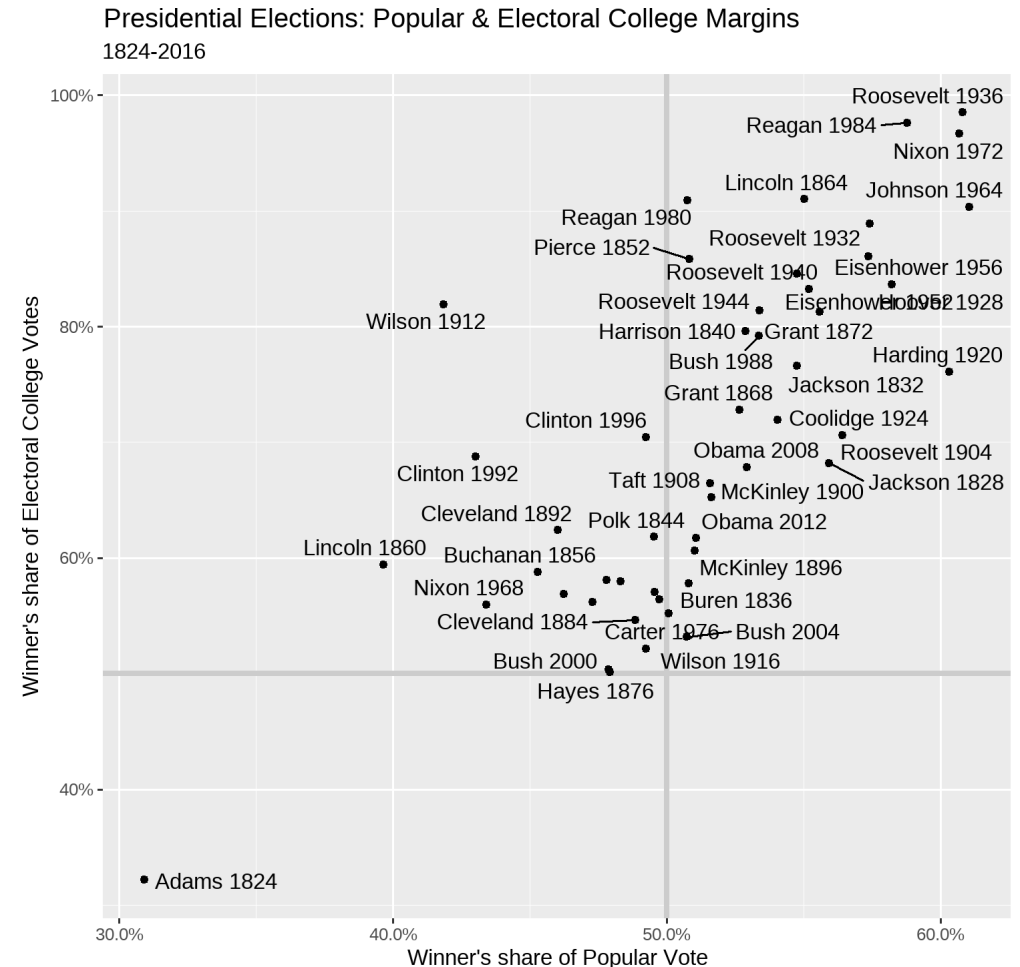
elections_historic %>% select(2:7)
```

```
## # A tibble: 49 x 6
##   year winner      win_party ec_pct popular_pct popular_margin
##   <int> <chr>      <chr>      <dbl>      <dbl>      <dbl>
## 1  1824 John Quincy Adams D.-R.      0.322      0.309      -0.104
## 2  1828 Andrew Jackson  Dem.      0.682      0.559      0.122
## 3  1832 Andrew Jackson  Dem.      0.766      0.547      0.178
## 4  1836 Martin Van Buren Dem.      0.578      0.508      0.142
## 5  1840 William Henry Harrison Whig      0.796      0.529      0.0605
## 6  1844 James Polk      Dem.      0.618      0.495      0.0145
## 7  1848 Zachary Taylor  Whig      0.562      0.473      0.0479
## 8  1852 Franklin Pierce Dem.      0.858      0.508      0.0695
## 9  1856 James Buchanan  Dem.      0.588      0.453      0.122
## 10 1860 Abraham Lincoln  Rep.      0.594      0.396      0.101
## # ... with 39 more rows
```

Building a Labeled Plot

```
p <- ggplot(data = elections_historic,
  mapping = aes(x = popular_pct,
    y = ec_pct,
    label = winner_label))

p +
  geom_hline(yintercept = 0.5,
    size = 1.4,
    color = "gray80") +
  geom_vline(xintercept = 0.5,
    size = 1.4,
    color = "gray80") +
  geom_point() +
  geom_text_repel(seed = 1234) +
  scale_x_continuous(labels = scales::percent) +
  scale_y_continuous(labels = scales::percent) +
  labs(x = x_label,
    y = y_label,
    title = p_title,
    subtitle = p_subtitle,
    caption = p_caption)
```



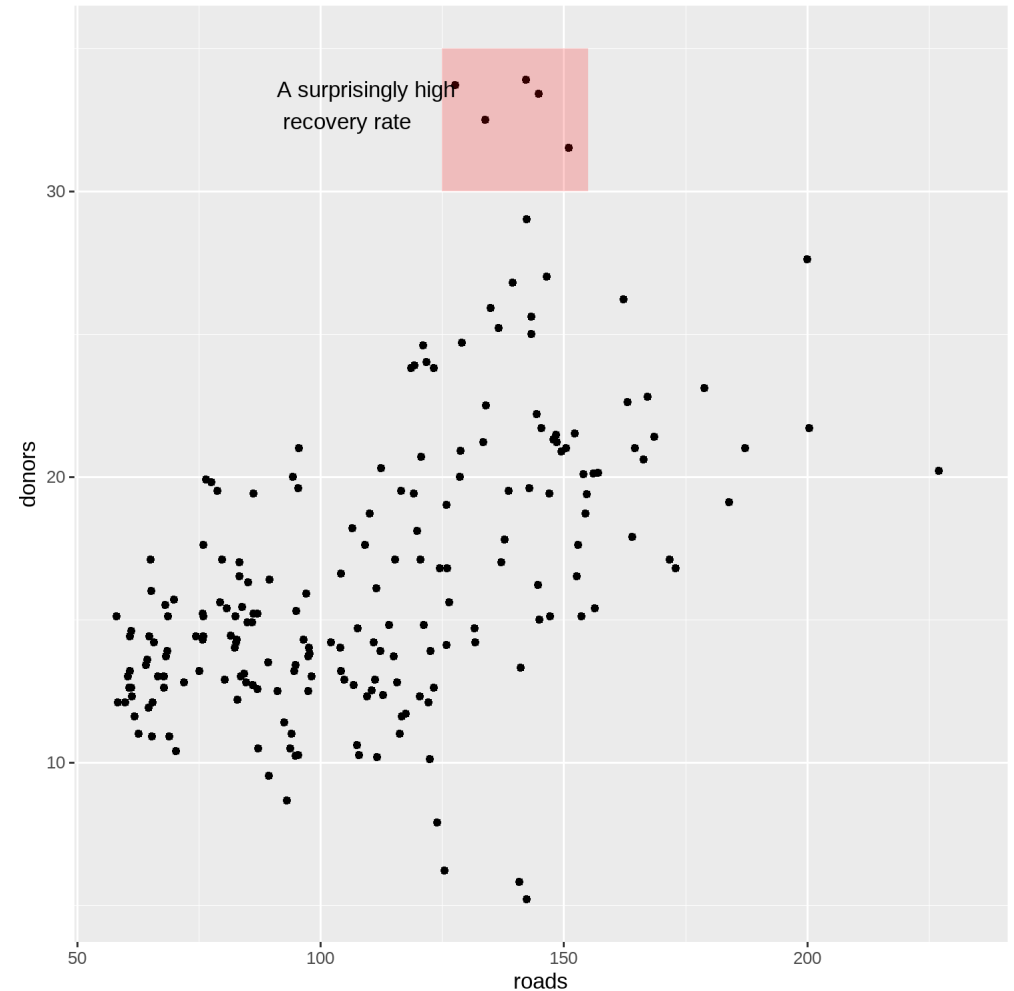
Label Outliers

Only label points with mean GDP greater than 25,000 OR mean health less than 1,500 OR the country Belgium

```
p <- ggplot(data = by_country,  
           mapping = aes(x = gdp_mean,  
                         y = health_mean))  
  
p +  
  geom_point() +  
  geom_text_repel(data = subset(by_country,  
                               gdp_mean > 25000 |  
                               health_mean < 1500 |  
                               country %in% "Belgium"),  
                 mapping = aes(label = country))
```

Annotations

```
p <- ggplot(data = organdata,  
            mapping = aes(x = roads,  
                          y = donors))  
  
p +  
  geom_point() +  
  annotate(geom = "rect",  
          xmin = 125, xmax = 155,  
          ymin = 30, ymax = 35,  
          fill = "red", alpha = 0.2) +  
  annotate(geom = "text",  
          x = 91,  
          y = 33,  
          label = "A surprisingly high \n recovery rate",  
          hjust = 0)
```



Lab Time

Pages 115 - 131 , 132 (bullet point 1)

Homework Assignment

Task: Modify 3 previous plots with additional labels/annotations.

Due: March 3, 2021

Rubric

Notes

- Use plots from labs or homework assignments
- Use your resources: Healy, Google, Student Community BUT cite where you get code from if you copy it directly
- See Blackboard assignment for a template `.Rmd` file

Midterm

Task: Using R and an approved dataset, create a polished infographic that contains at least 3 charts.

Due: March 3, 2021

Rubric

- Create summary statistics and 3-5 polished visuals based on the dataset using **only R**.
- The visuals created in R must include: titles, axis labels, data labels and legends (as needed), your name and the data source (as footnotes/captions)
- Output the visuals to **.svg** and import them into Adobe Illustrator to create an infographic that includes a narrative, the visuals from R, and other graphics.
- Turn in your R code as an **.Rmd** file, including comments about what you did and why, **and** the final infographic.
- In class on **March 3**, be prepared to share information about your dataset, your visuals, what they show, and the choices you made in their design. You will share both your code and the final infographic.

Tasks to Complete

- Homework #3
- Midterm