Telling stories with R: Data Visualization

Jan Zilinsky

Table of contents

Overview of this class

Creating effective visualizations of social and political data can help you discover and communicate new insights. This is a course designed to help students become better communicators with R. The focus is on graphing various types of evidence including:

```
- public opinion data
- macro-economic data
- summaries of statistical models
- quantitative representations of text (e.g. content of social media post and the accompanyi

Students are encouraged to think creatively about visualizing different types of information
```

After taking this course, students will be expected to be able to present real data clearly and to identify strengths and weakness of existing data displays and dashboards.

Introductory topics

- What works and what to avoid even if it works?
- Principles of visual perception and effect communication
- Getting familiar with ggplot

A ggplot deep dive

```
- Toplines, cross-tabs
- Geometries, statistics and coordinates
- Facets, themes
- Refining plots
- 3-way cross-tabs
- Heatmaps
```

Visualizing output from statistical models

```
- Coefficients and uncertainty
- Predicted probabilities, marginal effects, and interactions
- Model performance (in-sample and out-of-sample comparisons)
- Machine learning output (regression trees, most important variables, etc.)
```

Assignments:

- Create your own dataset (30%).
  - Create your own dataset. It needs to have at least one of these 3 attributes
    1. Multiple levels (at least 2).
    2. Original topic, subject or angle.
    3. Impressive scope
- Final project (70%)
  - Form a group of 2-3 students
  - Prepare a compelling data visualization
  - Some elements in R are expected, you could also use D3 or another language if you wish.

1 Principles

1.1 There are always tradeoffs

The central tradeoff often is **trutfulness** vs.

- Readability vs. “completeness”
- Concise vs. “attention-gabbing”
- Simplicity vs. other goals

If you drop outliers, for example, your chart’s readability will almost surely improve, but it could be less truthful.

2 Toplines and crosstabs

In summary, this book has no content whatsoever.

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.1     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.1     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
d1
```

```
# A tibble: 2,000 x 94
   caseid   female   edu black hispanic   age income   pid  ideo interest attend
   <chr>     <dbl> <dbl> <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>    <dbl>  <dbl>
 1 R_24COU~      1     5     0        0    23      2     7     3        5      3
 2 R_2B2nP~      0     6     0        0    39      7     4     6        3      2
 3 R_p5eQb~      0     3     0        0    43      4     4     1        3      2
 4 R_2dYYB~      0     2     1        0    22      2     1     7        4      3
 5 R_3sgIL~      0     3     1        0    40      5     1     1        4      3
 6 R_31Ab1~      0     6     0        0    28      4     4     4        3      1
 7 R_2f36X~      0     6     0        0    41      7     4     2        4      2
 8 R_2XcYI~      0     2     1        0    21      4     1     3        4      4
 9 R_339E8~      1     6     0        0    58      6     3     3        4      4
10 R_3mlfI~      0     5     0        0    43      6     1     1        5      4
# i 1,990 more rows
# i 83 more variables: facebook <dbl+lbl>, twitter <dbl+lbl>, reddit <dbl+lbl>,
#   chans <dbl>, con1 <dbl>, con2 <dbl>, con3 <dbl>, con4 <dbl>, conwis <dbl>,
#   msm <dbl>, onepercent <dbl>, deepstate <dbl>, goodevil <dbl+lbl>,
#   vio1 <dbl>, vio2 <dbl>, violence <dbl>, argue1 <dbl>, argue2 <dbl>,
#   argue3 <dbl>, argument <dbl>, pop1 <dbl>, pop2 <dbl>, official <dbl>,
#   manip1 <dbl>, manip2 <dbl>, manip3 <dbl>, manip4 <dbl>, ...
```

```
table(d2$climatechange)
```

```

  1   2   3   4   5 
733 454 395 233 206 
```

```
table(d2$climatechangeBIN)
```

```

   0    1 
1582  439 
```

```
d2 %>% count(climatechangeBIN)
```

```
# A tibble: 3 x 2
  climatechangeBIN     n
             <dbl> <int>
1                0  1582
2                1   439
3               NA     2
```

Are the missing observations the same for the original and the recoded variable? (If not, we would want to check whether earlier code did something unintended.)

```
d2 %>% count(climatechangeBIN,climatechange)
```

```
# A tibble: 6 x 3
  climatechangeBIN climatechange     n
             <dbl>         <dbl> <int>
1                0             1   733
2                0             2   454
3                0             3   395
4                1             4   233
5                1             5   206
6               NA            NA     2
```

3 Standard charts

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

4 Advanced ggplot

4.1 Heatmaps

```
1 + 1
```

```
[1] 2
```

5 Visualizing statistical models

A more accurate title, of course, woudl be “visualizing *outputs* from statistical models”.

References

Useful resources include:

Gestalt Principles

Gestalt Principles (Part 2)

https://socviz.co/

https://ggplot2-book.org/index.html

https://cssbook.net/content/chapter06.html

https://storymaps.arcgis.com/stories/1e7f582d478a4b99bd0c70fffeac4c8b

https://cup.columbia.edu/book/better-data-visualizations/9780231193115

https://journals.sagepub.com/doi/pdf/10.1177/15291006211057899