

Telling stories with R: Data Visualization

Jan Zilinsky

Table of contents

Telling stories with R: Data Visualization	3
Topics:	3
A ggplot deep dive	3
Other lessons	4
1 Summary	5
2 Toplines and crosstabs	6
3 Standard charts	8
4 Advanced ggplot	9
4.1 Heatmaps	9
5 Visualizing statistical models	10
References	11
Useful resources include:	11

Telling stories with R: Data Visualization

Instructor: Jan Zilinsky

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:

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

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.

Topics:

- What works and what to avoid even if it works?
- Principles of visual perception and effective 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

Other lessons

- 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 (e.g. time dimension)
 - Easy example of #1: rating things you like.
 - * Suppose you decide go the “quantified self” route and create a dataset with evaluations the favorite TV shows of the members of your group.
 - * What’s the outcome variable? (Multiple attributes can be assessed.)
 - * What levels can be measured?
 - If you are rating a TV show, the natural components would be seasons, and episodes. Within episodes there might be themes or actors. Unpacking these attributes opens possibilities for creating stories, making interactive visualizations, etc.
- Final project (70%)
 - Form a group of 3 (max. 4) classmates
 - Start thinking about topic after lecture 3
 - Prepare a compelling data visualization
 - Some elements in R are expected, you could also use [D3](#) or another language if you wish.

1 Summary

In summary, this book has no content whatsoever.

`1 + 1`

[1] 2

2 Toplines and crosstabs

In summary, this book has no content whatsoever.

```
-- Attaching packages ----- tidyverse 1.3.2 --
v ggplot2 3.4.2      v purrr   1.0.1
v tibble  3.2.1      v dplyr   1.1.1
v tidyr   1.3.0      v stringr 1.5.0
v readr   2.1.3      v forcats 0.5.2
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

d1

```
# A tibble: 2,000 x 61
  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 50 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, would 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>