Digital Research Toolkit for Linguists

Week 8: Data visualization and exporting

Anna Pryslopska

May 7, 2024

Psycholinguistics and Cognitive Modeling Lab

Document (write down what, who, where, and when) and get help:

Susan Völkel (Ansprechperson für Antidiskriminierung)
antidiskriminierung@uni-stuttgart.de
+49 711 685 82274

Homework


GOOD JOB
YOU GET A GOLD STAR FOR TODAY
memegenerator.net

Main goal: Get acquainted with `ggplot2` and make different types of plots (bars, lines, points)

✖ Didn't finish the assignment (<8 plots) without explanation

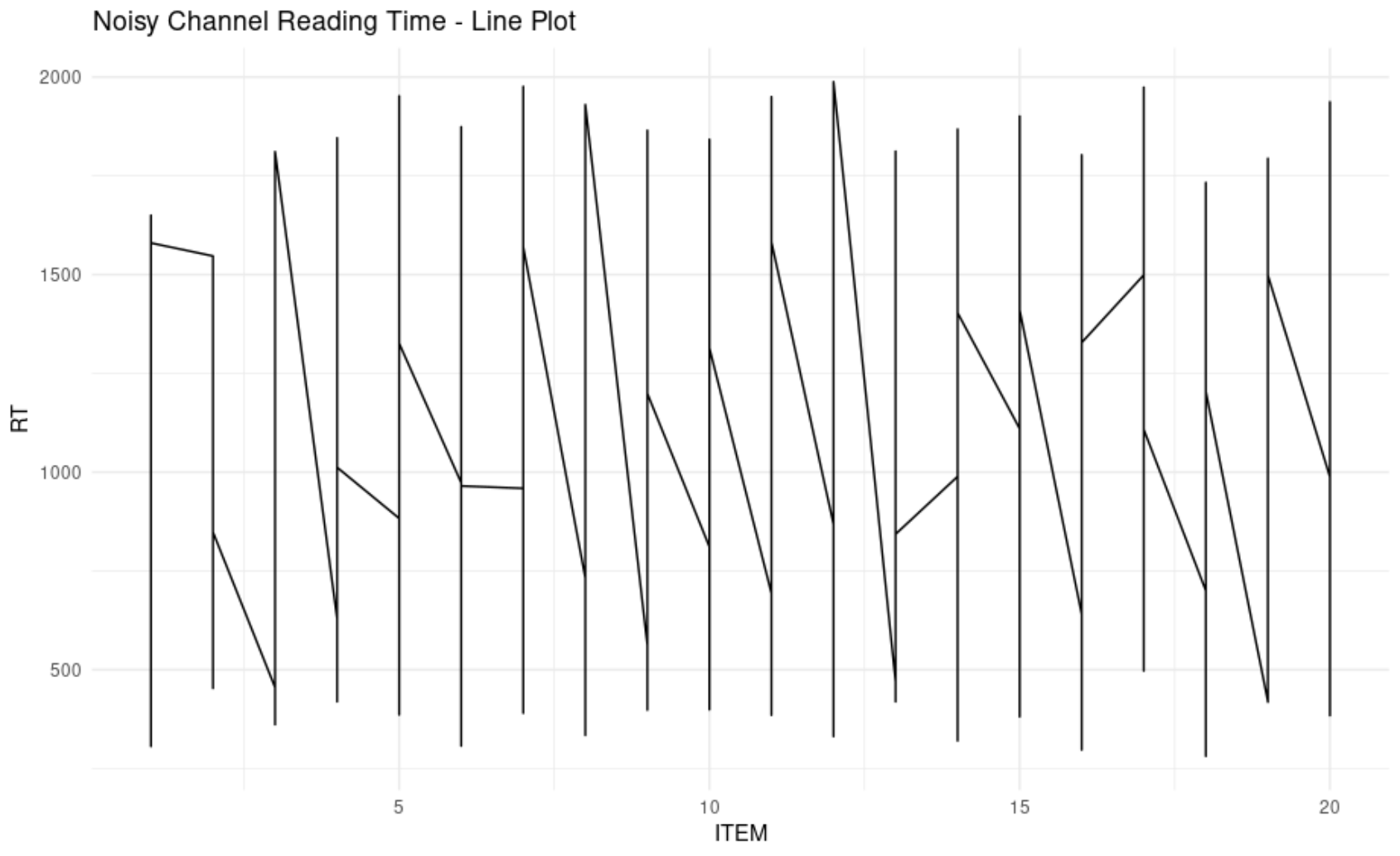
✖ Made only 1 kind of plot (e.g. all bars)

✖ Did not run your code (you can copy & paste from `esquisse`)

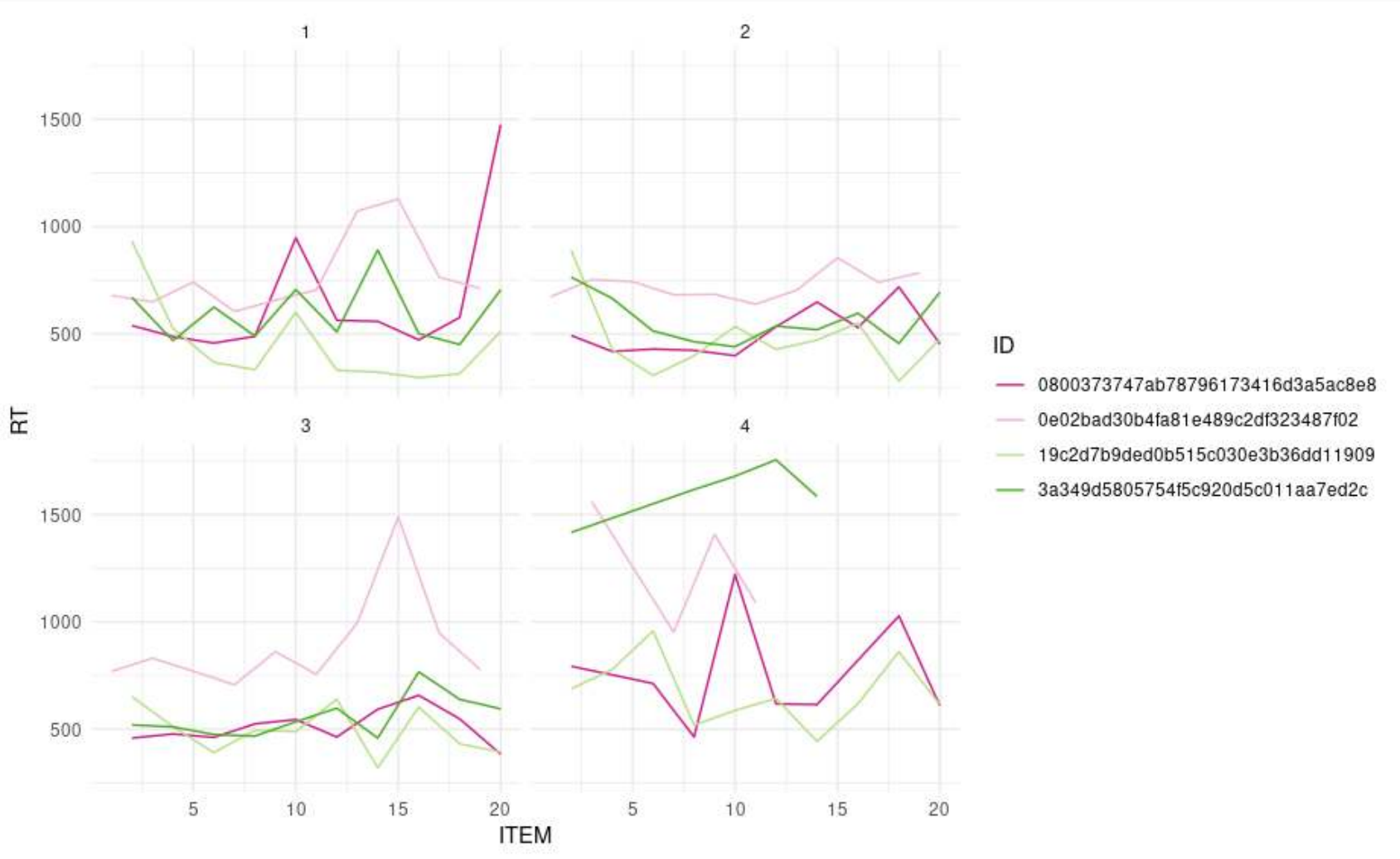
✖ Missed the essence of `ggplot2` and how layers work.

Recommended watching:
`https://www.youtube.com/watch?v=HPJn1CMvtmI`

⊘ `ggplot(noisy_rt.csv)` Assign your data to a variable.



Typically, you want **1 observation per row**.



If you have more than 1 observation per row, `ggplot2` will assume you have a good reason.

```
Error in `geom_bar()`:
! Problem while computing stat.
i Error occurred in the 1st layer.
Caused by error in `setup_params()`:
! `stat_count()` must only have an x or y
aesthetic.
```

You specified both x and y in the aesthetics.

> That was the whole point!

The `stat` argument within `geom_bar()` determines how the data should be summarized before plotting.

`stat = "count"`
Default setting. It counts the observations, e.g. for counting occurrences for each category (this will be 1 if you have 1 observation per row).

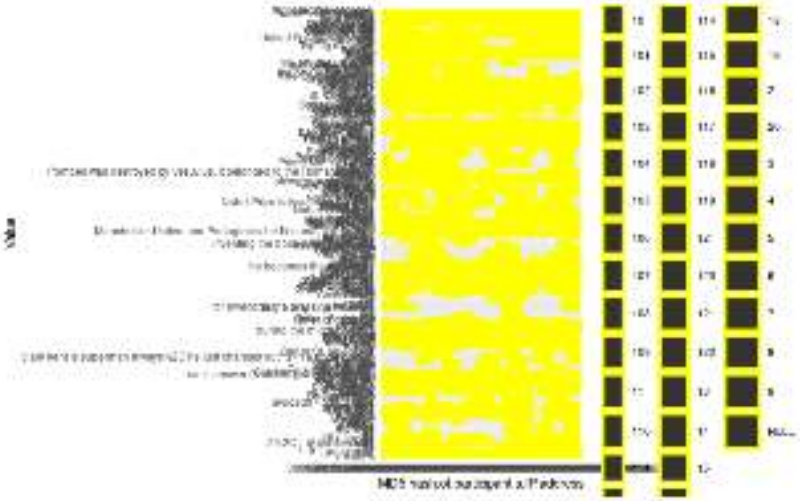
`stat = "identity"`
Uses data as is, without transformations, e.g. when you have the data in the form you want.

`stat = "summary", fun = "mean"`
Summarizes the data using a summary function `fun`, e.g. to calculate the mean.

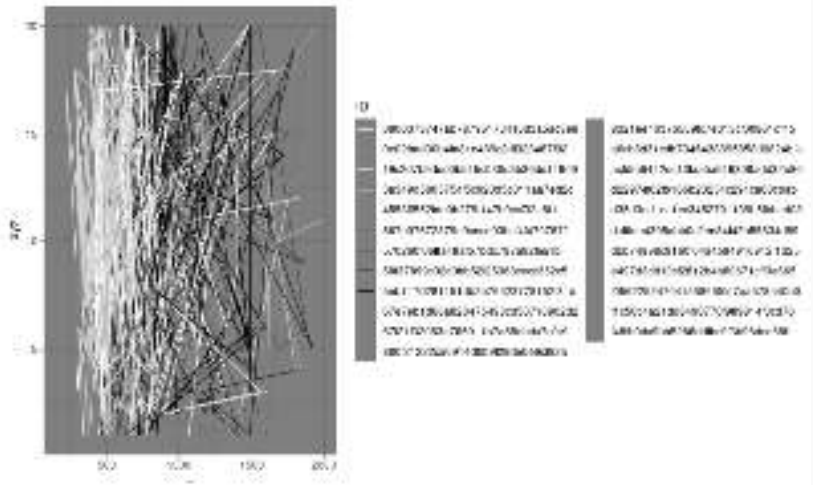
`bin`, `smooth`, `density`, ...

Ugly plots

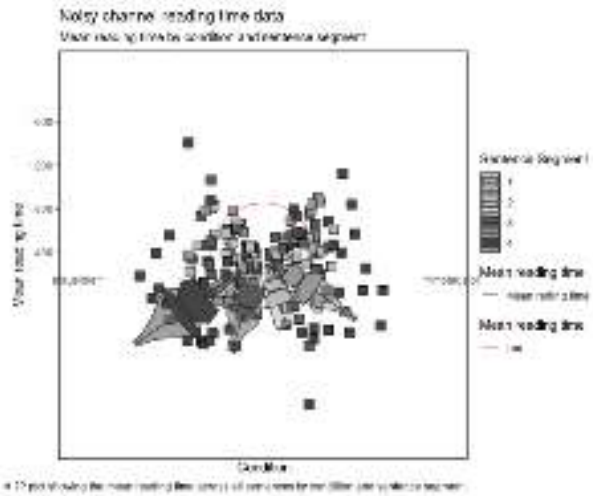
PLOT A



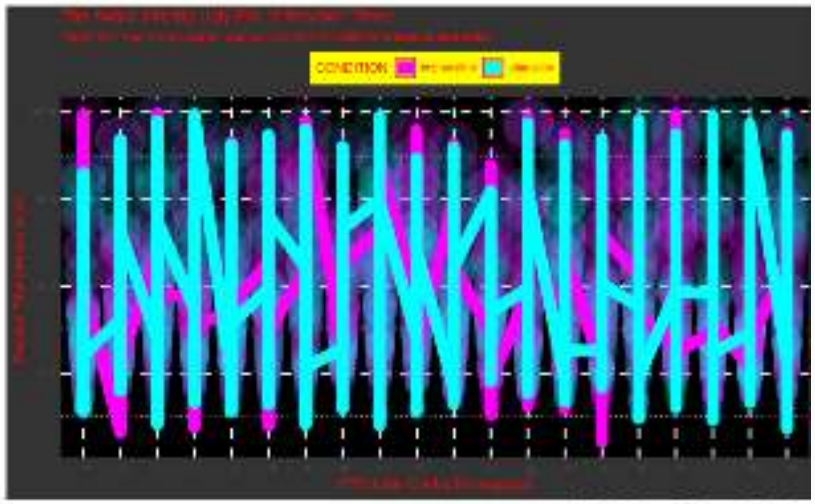
PLOT B



PLOT C



PLOT D



Questions?

Course credit and exam

3 credit points

✔ complete $n - 2$ homeworks.

✖ complete $< n - 2$ homeworks.

6 credit points

✔ complete $n - 2$ homeworks + written exam (multiple choice*).

✖ complete $< n - 2$ homeworks.

9 credit points

✔ complete $n - 2$ homeworks + term paper on **new** data.

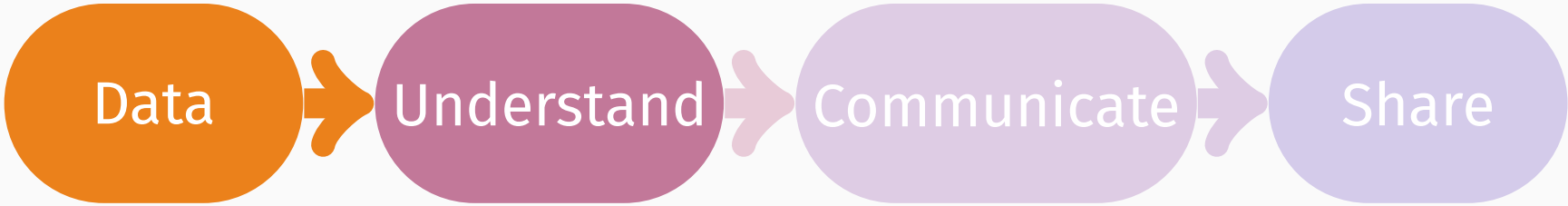
✖ complete $< n - 2$ homeworks.

Tentative exam data: Monday, July 29th at 14:00

Questions?

Table of contents

Where are we this week?



R & RStudio, packages, data types, formats, encoding

import from workspace, assign values, operations, clean, filter, arrange, select, merge, group, summarize, export, visualize

document, create clean and beautiful reports

connect, collaborate, backup

Data visualization

Data visualization

is an interdisciplinary field. It's the graphical representation of data & information in order efficiently communicate communicate complex relationships & insights in a comprehensible way.

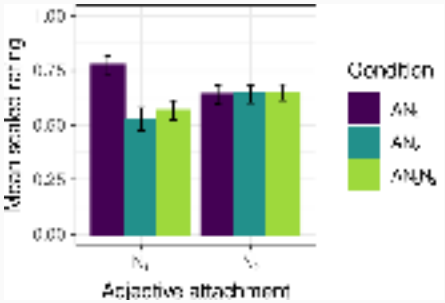
Goals

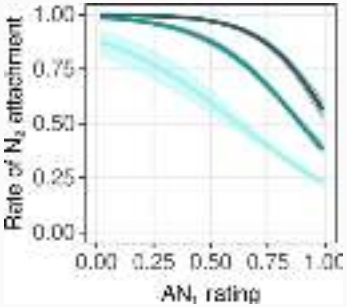
- Communication & understanding
- Analysis & exploration
- Decision making

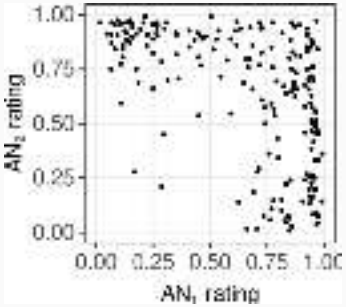
Visualization types

boil down to tables, bars, lines, and points, and a mix of all four.

item	sentence	N1	N2	unclear
3	sommerliche Arbeitskleidung			
11	eiserne Ostentenuhr			
12	amtliche Baugenehmigung			
16	roher Bienenstock			
19	goldenes Brillenetui			
22	leckeres Rückenspeck			
23	evangelischer Bürgermeister			







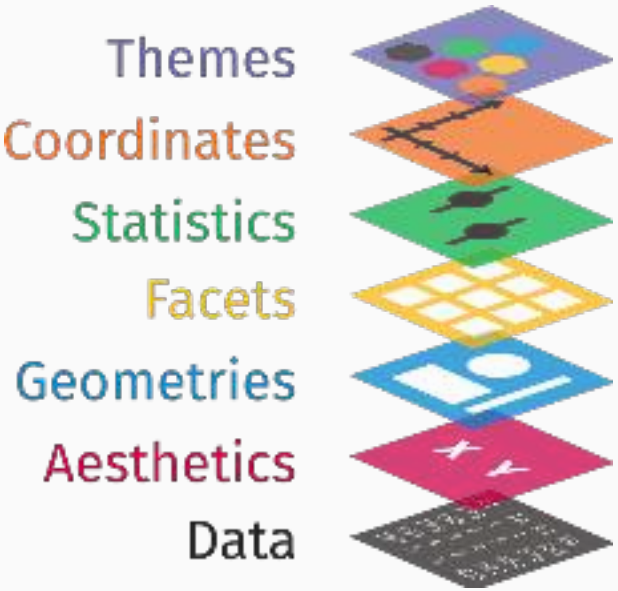
The 4 Principles of Accessibility

Web Content Accessibility Guidelines

- **P** Perceivable
- **O** Operable
- **U** Understandable
- **R** Robust

`ggplot2`

a layered grammar of graphics

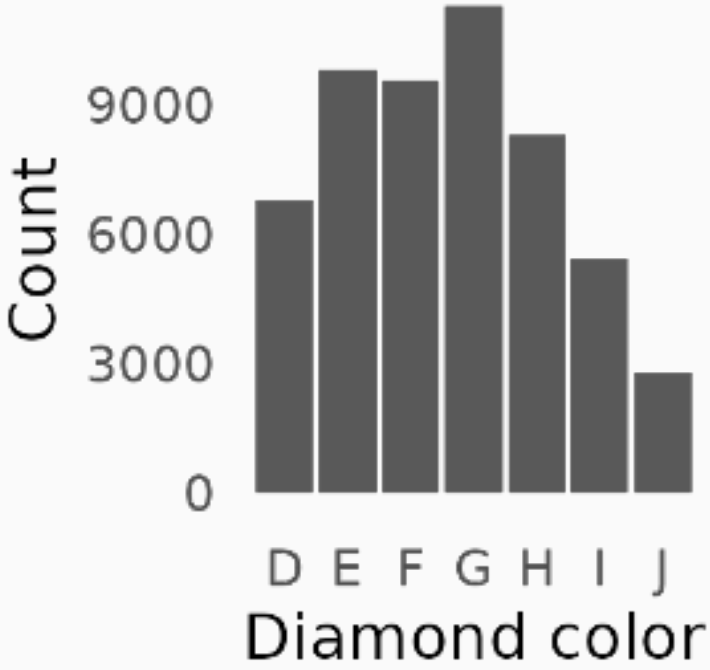


`esquisse`

GUI for exploring data based on `ggplot2`

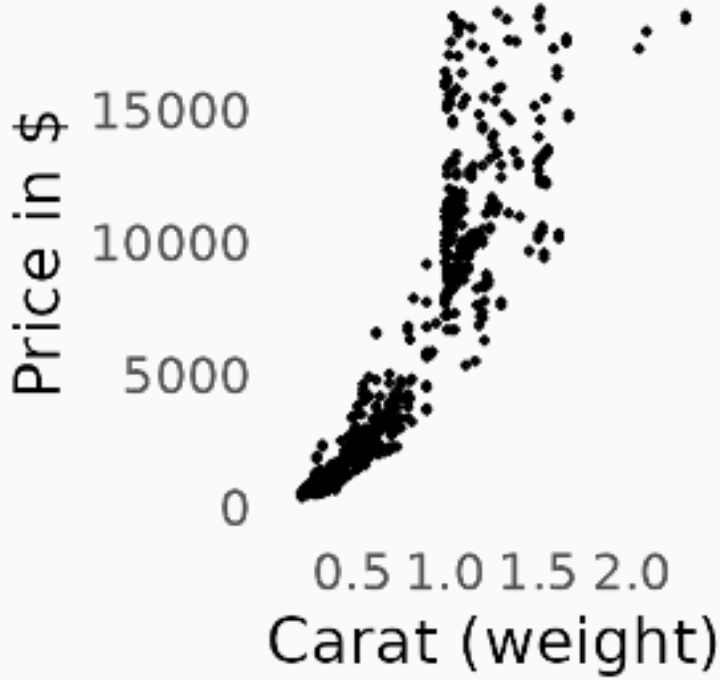
Histogram

Counts observations of categories



Scatterplot

Shows the relationship between two numeric variables



Colors

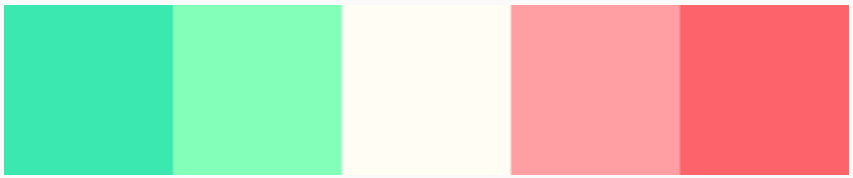
Color palette types

Discrete

Uses distinct, unrelated colors for categorical data.

Divergent

Shows data with a central midpoint, using contrasting colors to highlight deviations.



Sequential

Represents data with a gradient of colors to indicate ordered values.

DIY

Adobe color.adobe.com

Coolors coolors.co

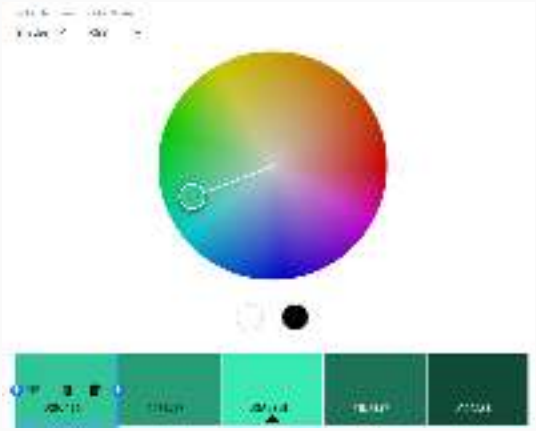
LearnUI www.learnui.design

Test the colors for accessibility and BW print:
webaim.org/resources/contrastchecker

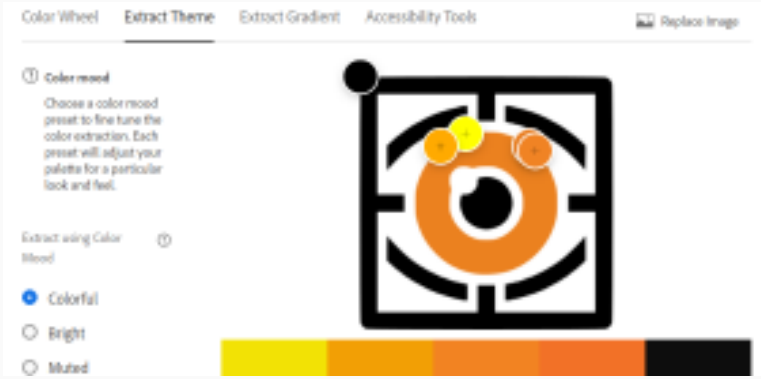
```
my.palette <- c("#8c510a","#d8b365","#f6e8c3","#f5f5f5")
plot + scale_color_manual(values = my.palette)

plot + scale_fill_manual(values =
c("#8c510a","#d8b365","#f6e8c3","#f5f5f5"))
```

Adobe color

Assemble colors

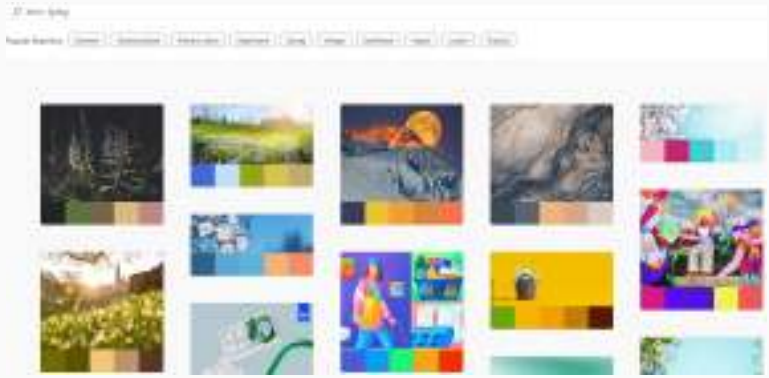


Extract from image



Check contrast

Browse trends



Coolors

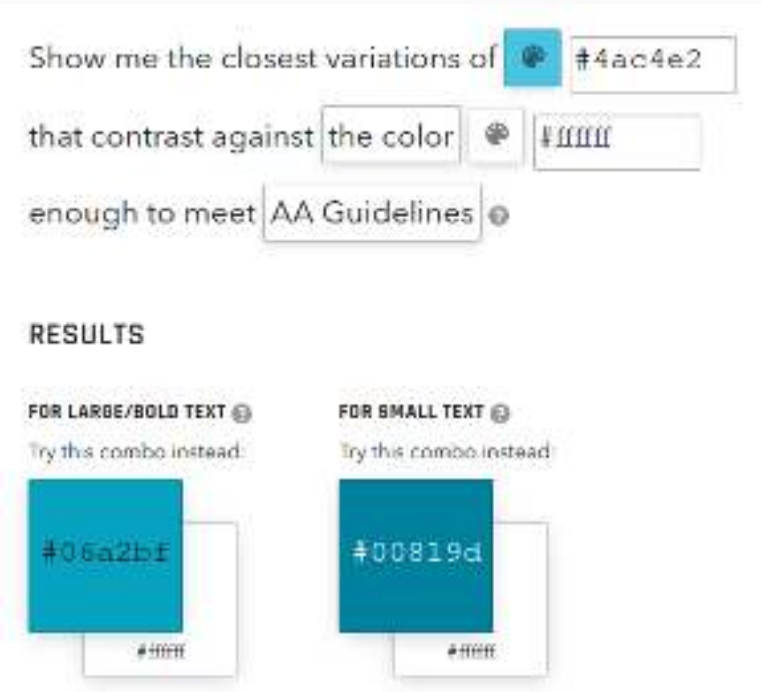
Assemble colors and check for color blindness

Visualize the palette

Browse trends

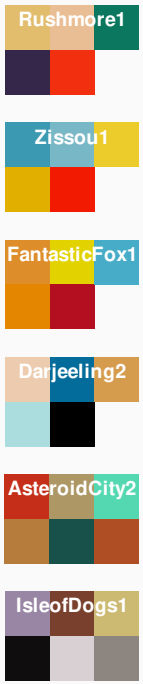
Generate color palettes

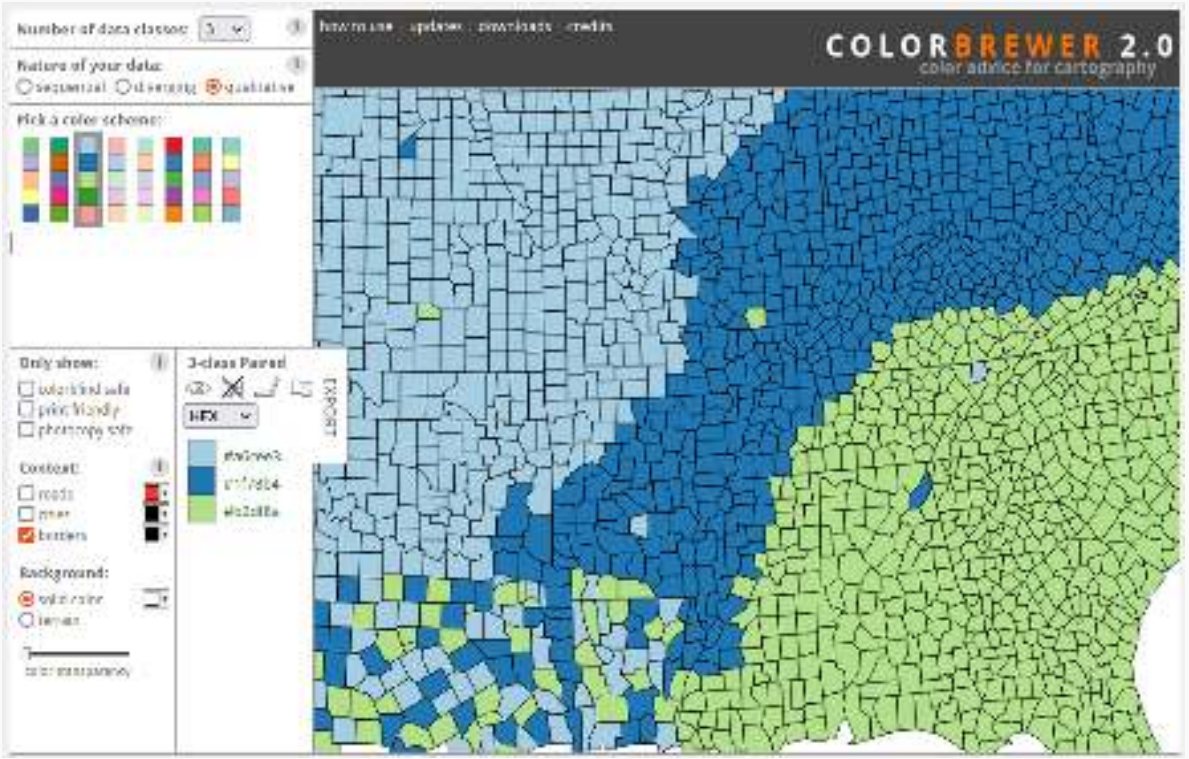
Check contrast



CBA


viridis
magma
plasma
inferno
cividis


Rushmore1
Zissou1
FantasticFox1
Darjeeling2
AsteroidCity2
IsleofDogs1


COLORBREWER 2.0
color advice for cartography

There's a package for that

`RColorBrewer`

```
scale_fill_brewer(…, type="div", palette=1)
```

`viridis`

```
scale_colour_viridis_c(…, option="inferno")
```

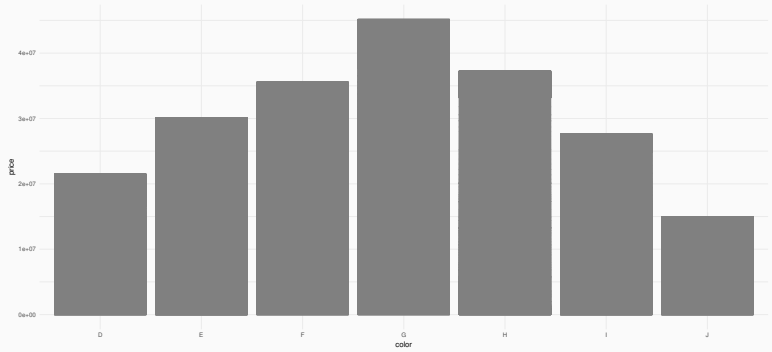
`wesanderson`

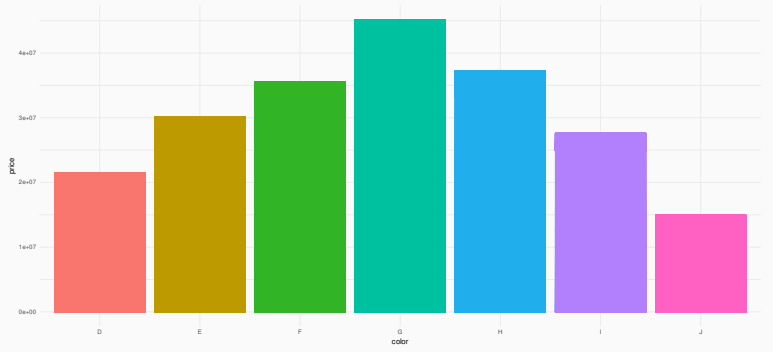
```
scale_fill_manual(values=wes_palette("BottleRocket1"))
```

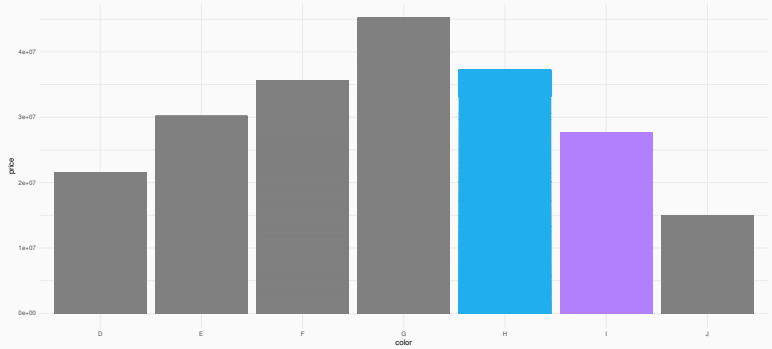
Others, e.g. `ggsci` (for journals)

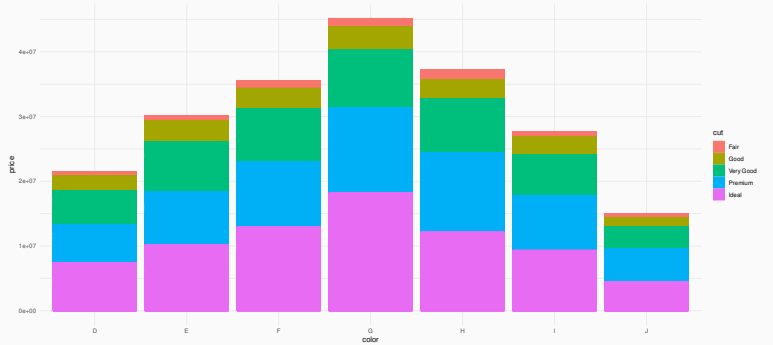
Color use guidelines

Avoid unnecessary usage of color

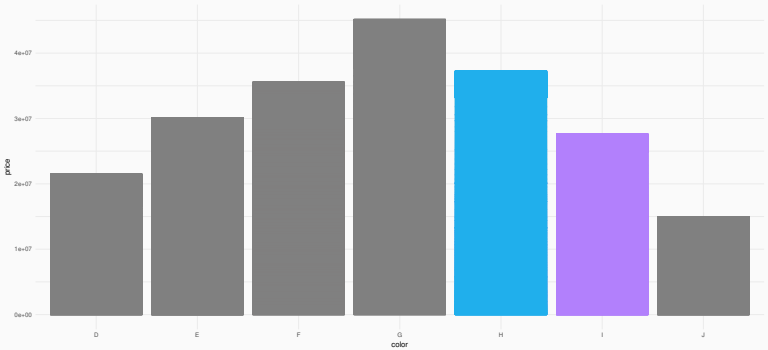


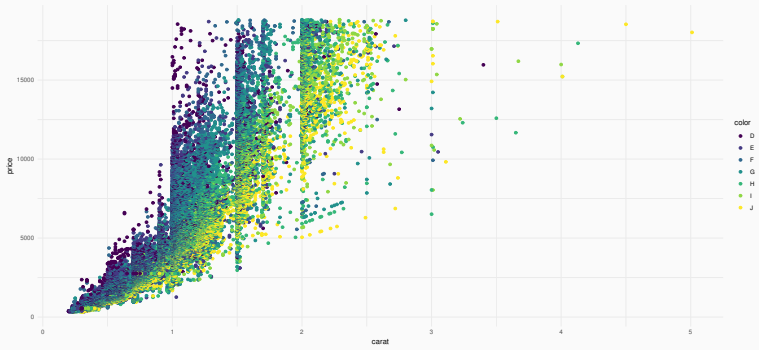


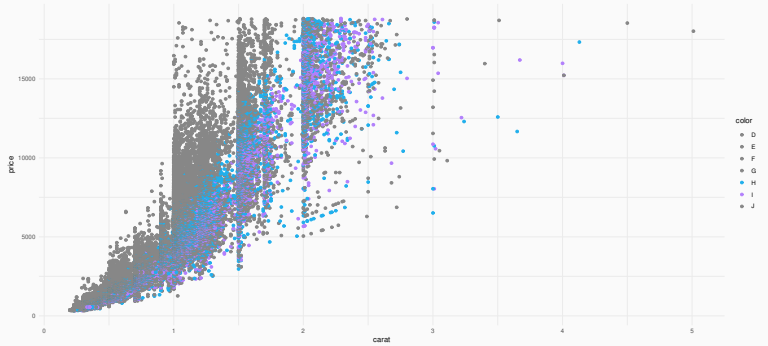


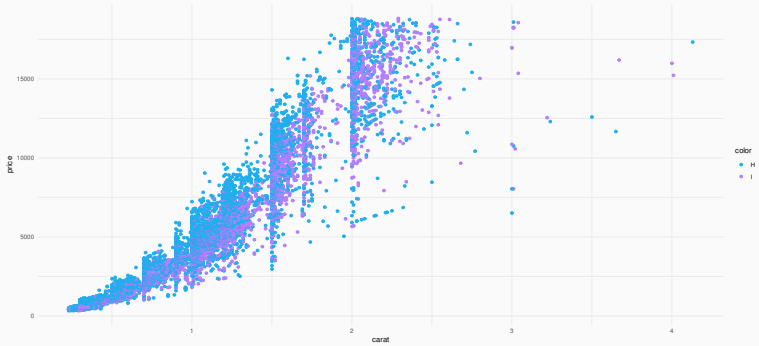


Be consistent

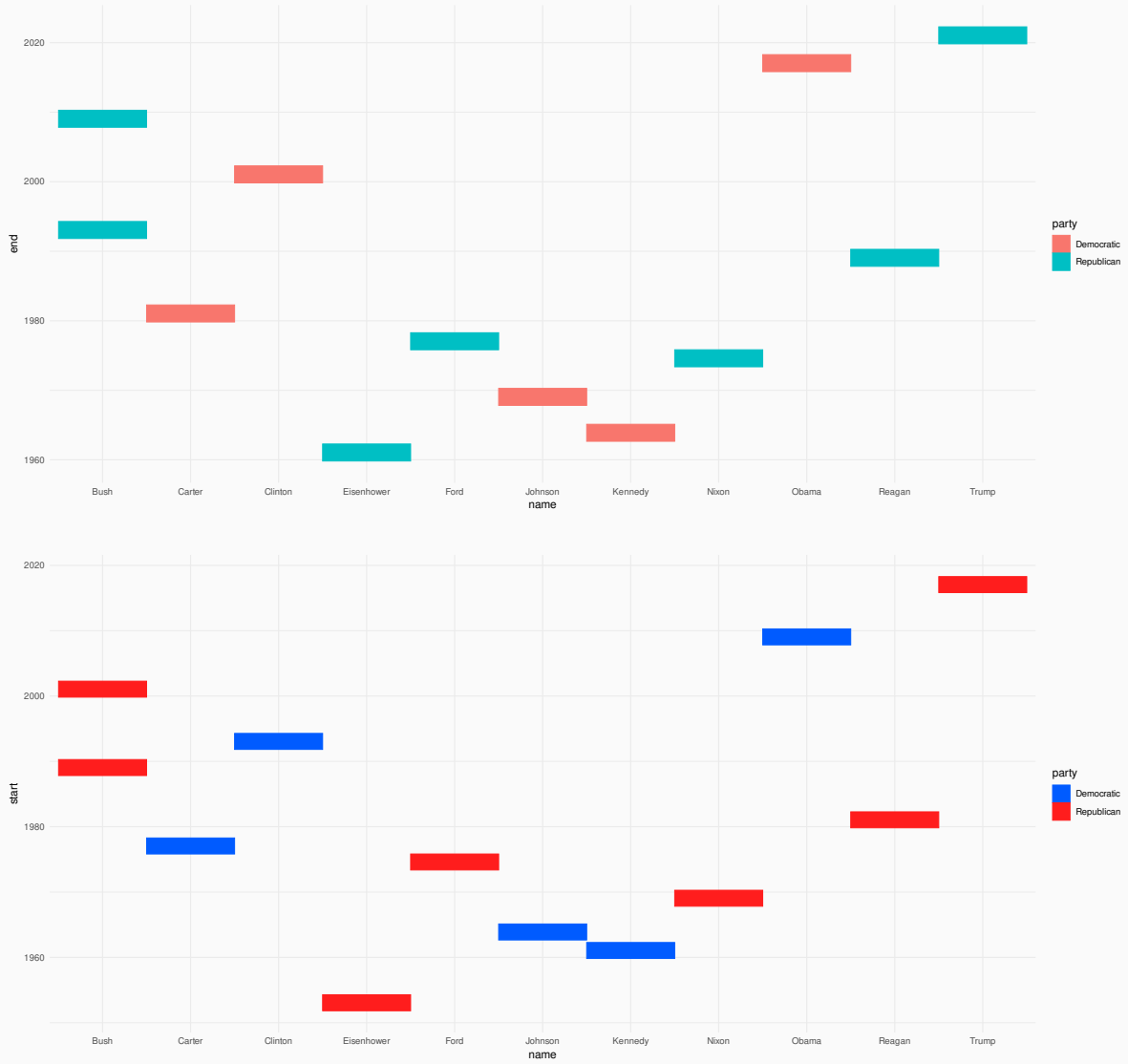




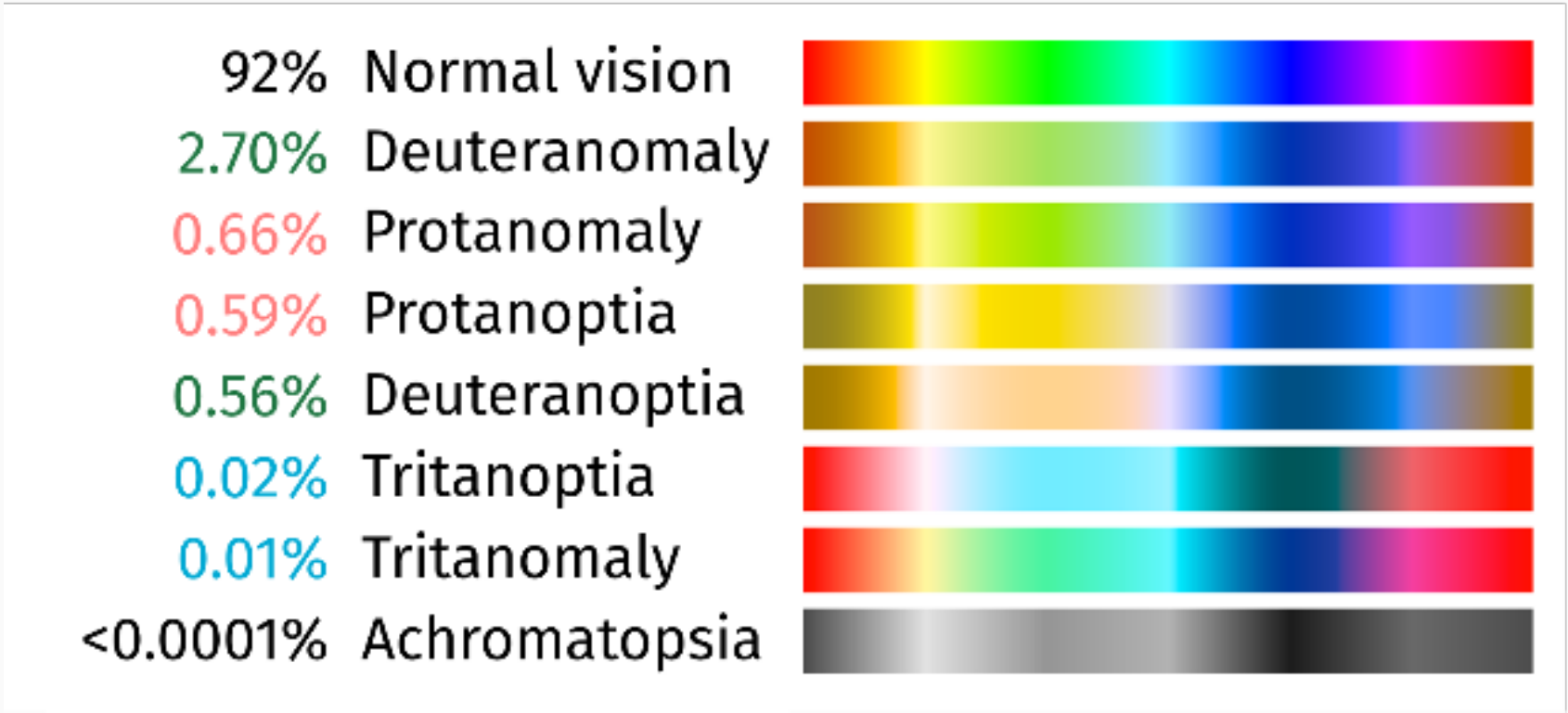




Use in a meaningful way

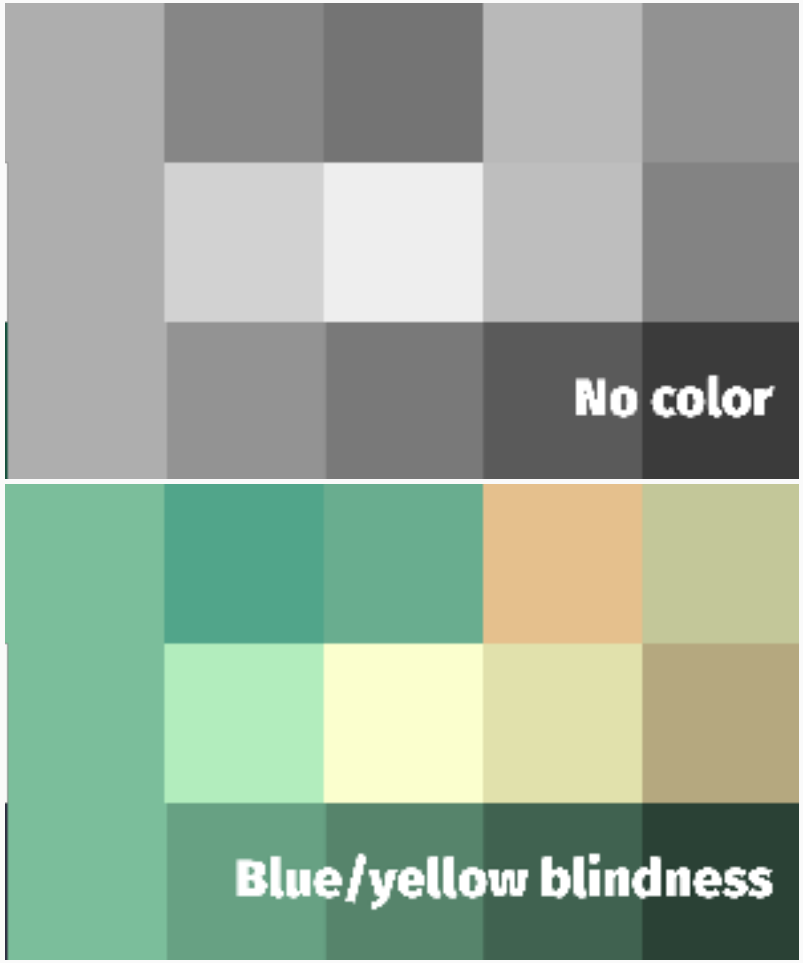


Pay attention to color blindness



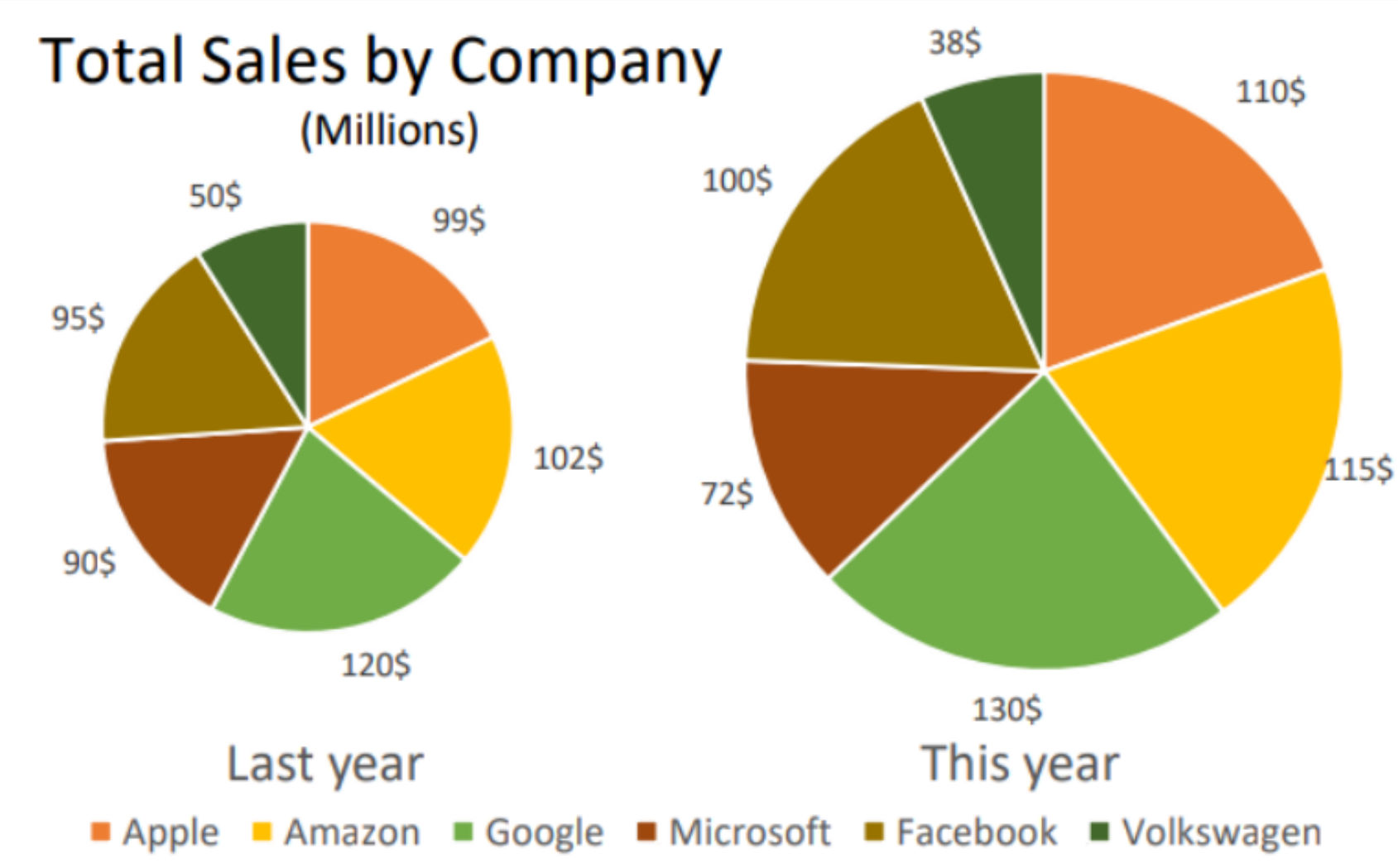
1–2 people in this class

Perceivable colors



Lying with plots

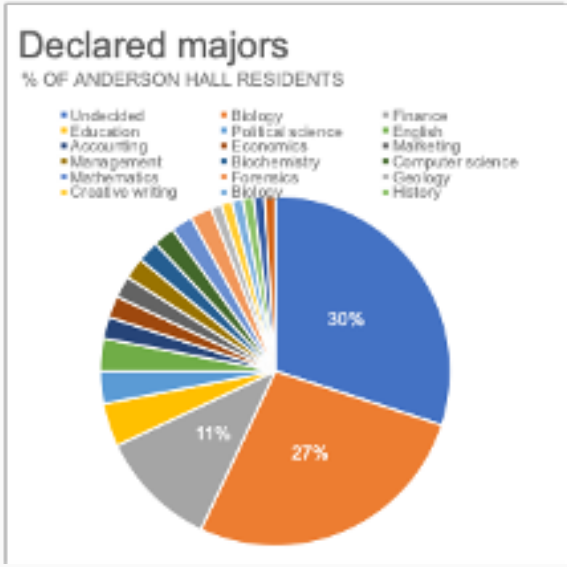
Pie plots: Manipulate the size



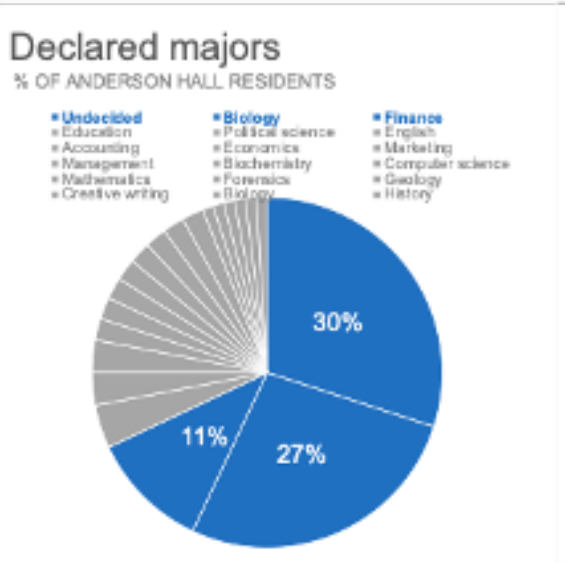
Fallenbüchel (2019)

Pie plots: Manipulate categories

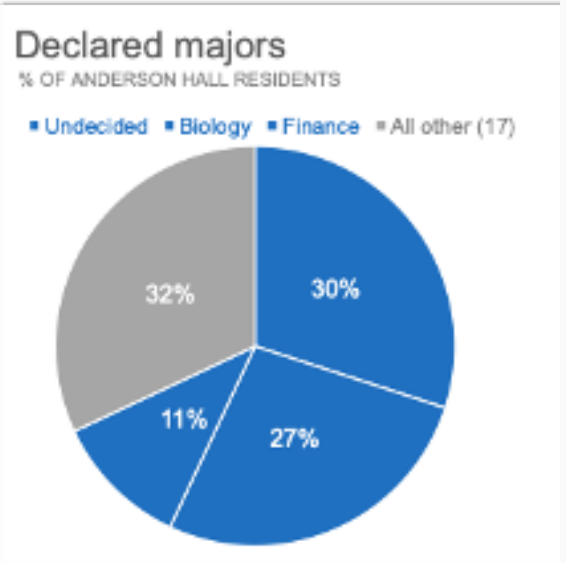
Too many slices



De-emphasize

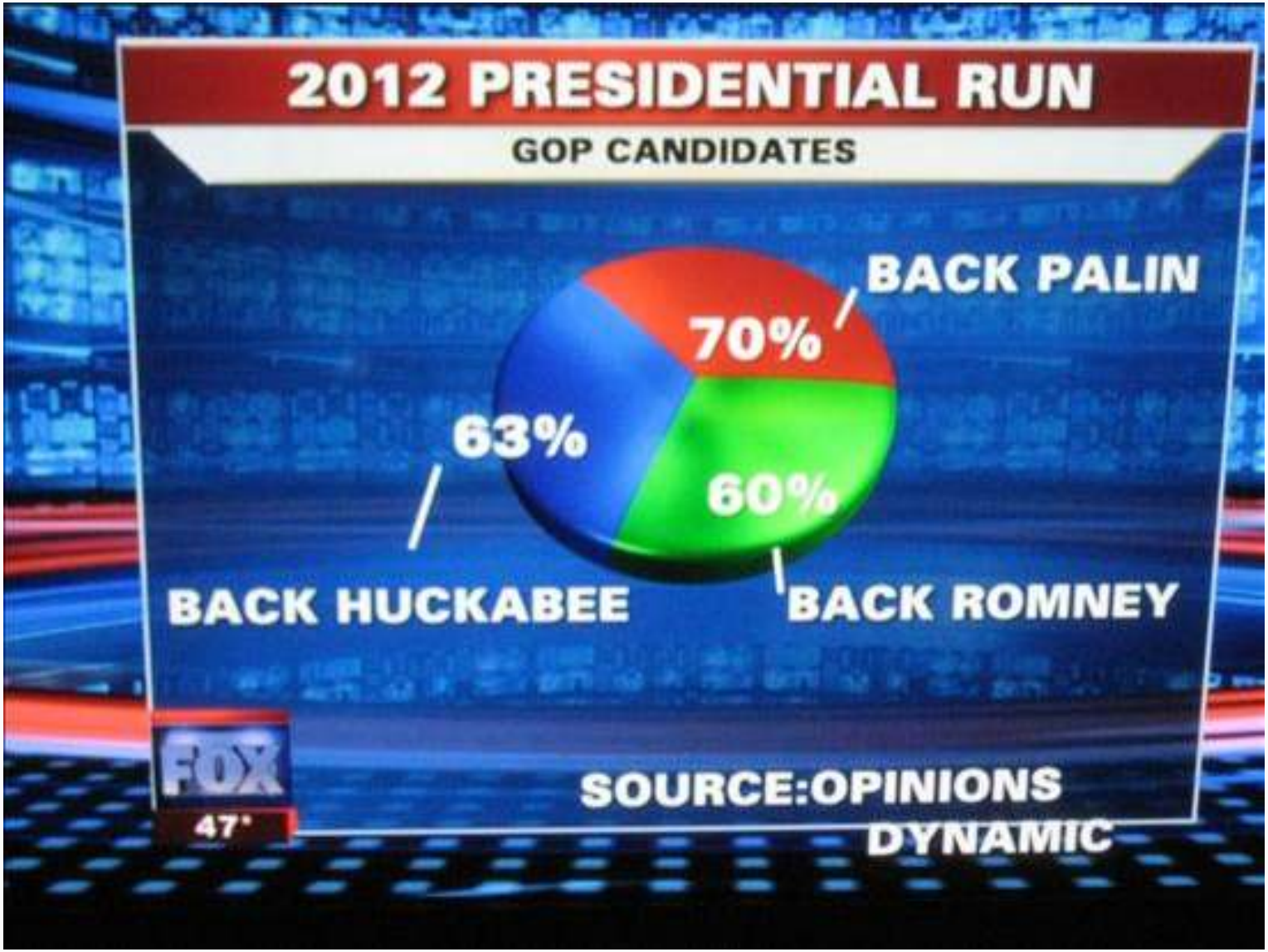


Aggregate

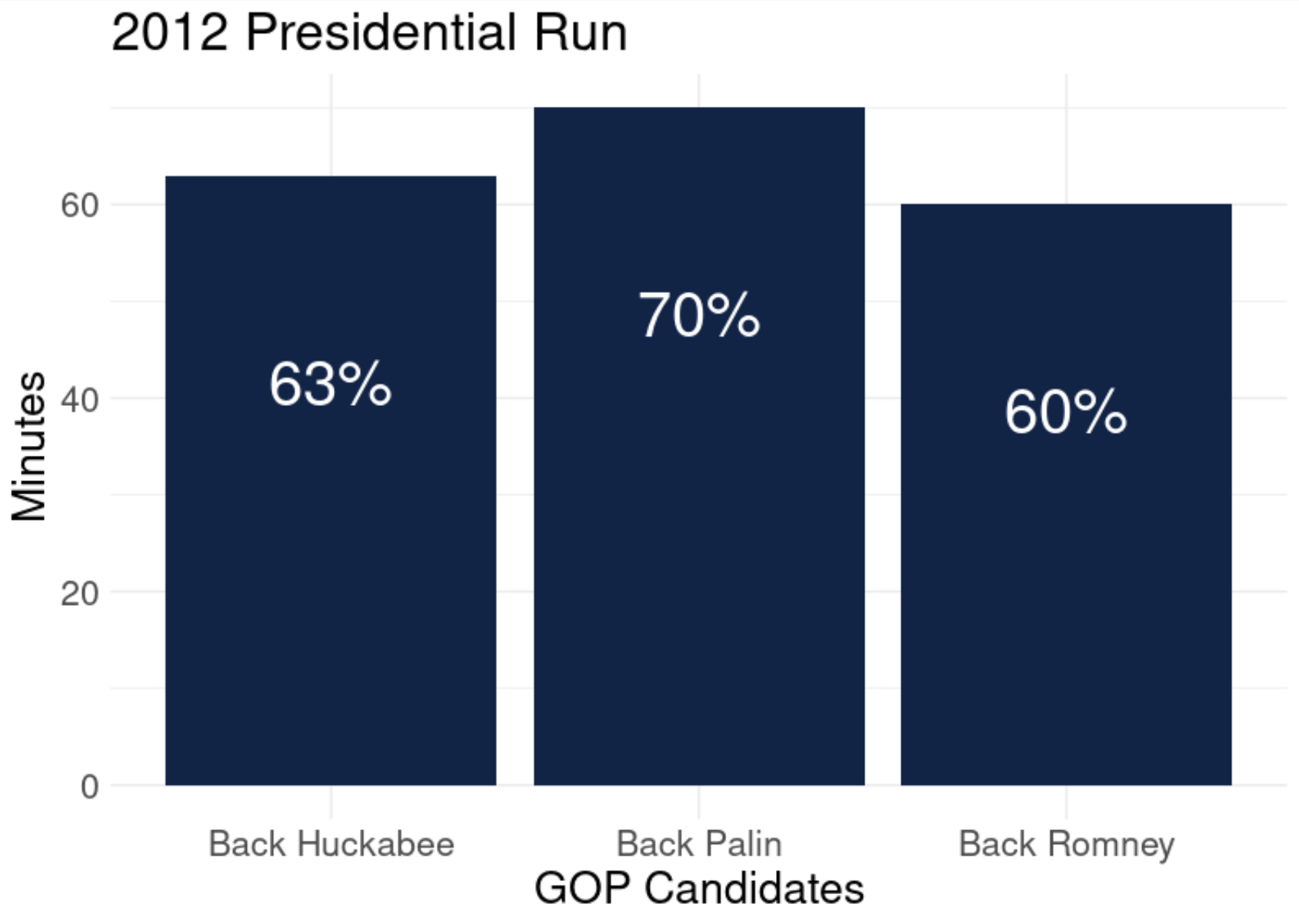


Ricks (2020)

Pie plots: Wrong plot type

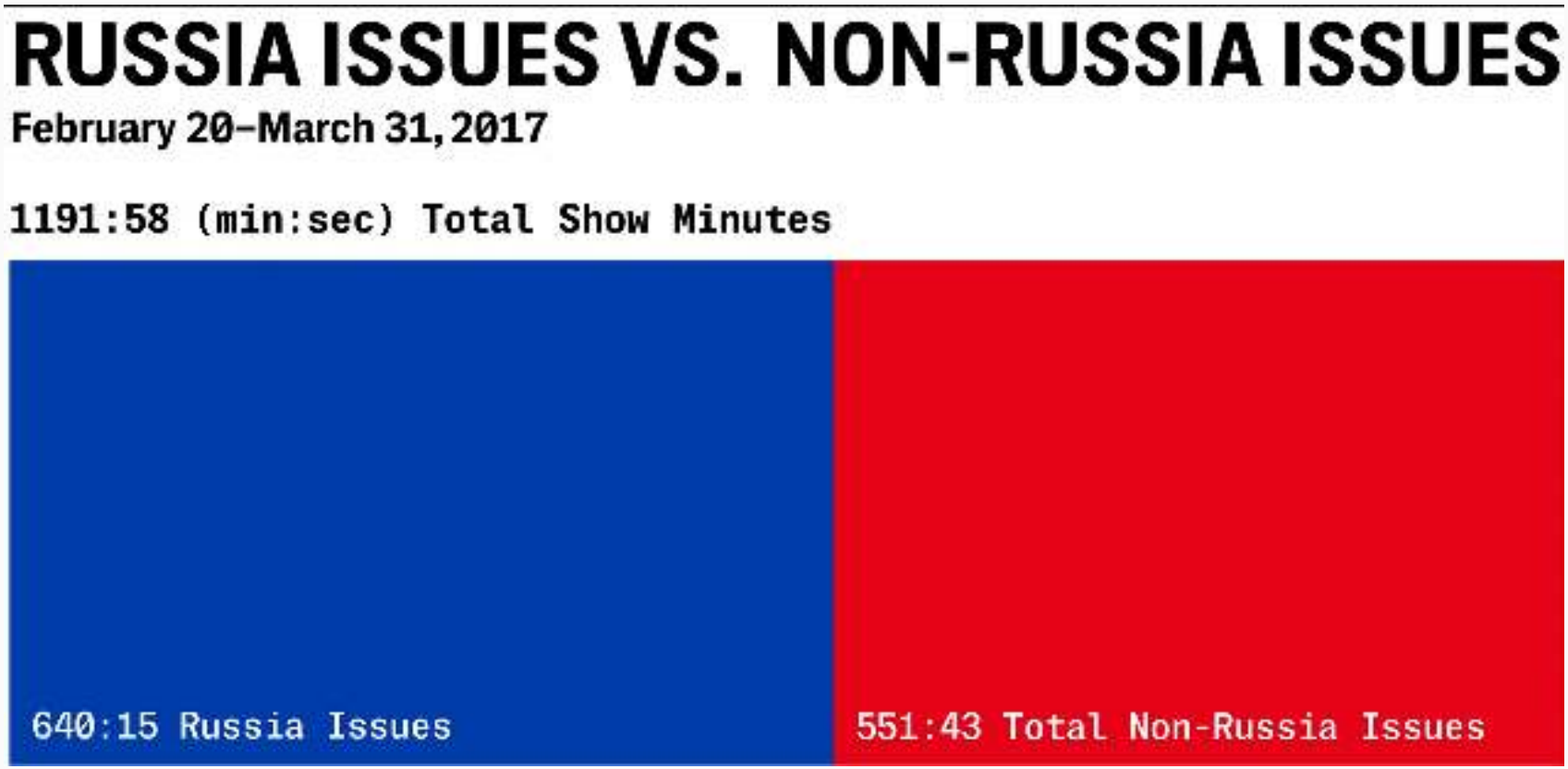


Pie plots: Wrong plot type



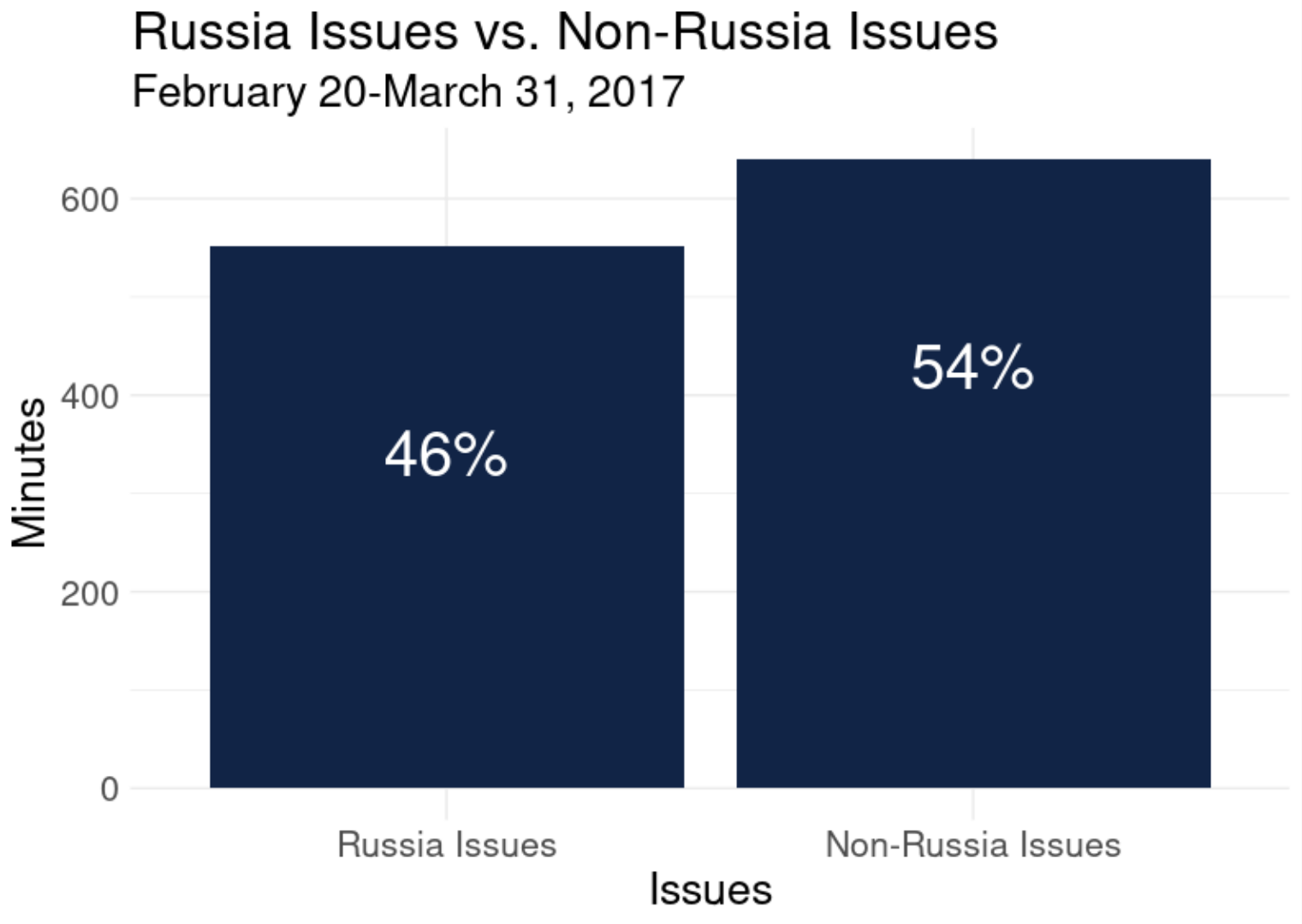
McCready (2020)

Bar plots: Stack (+ color)

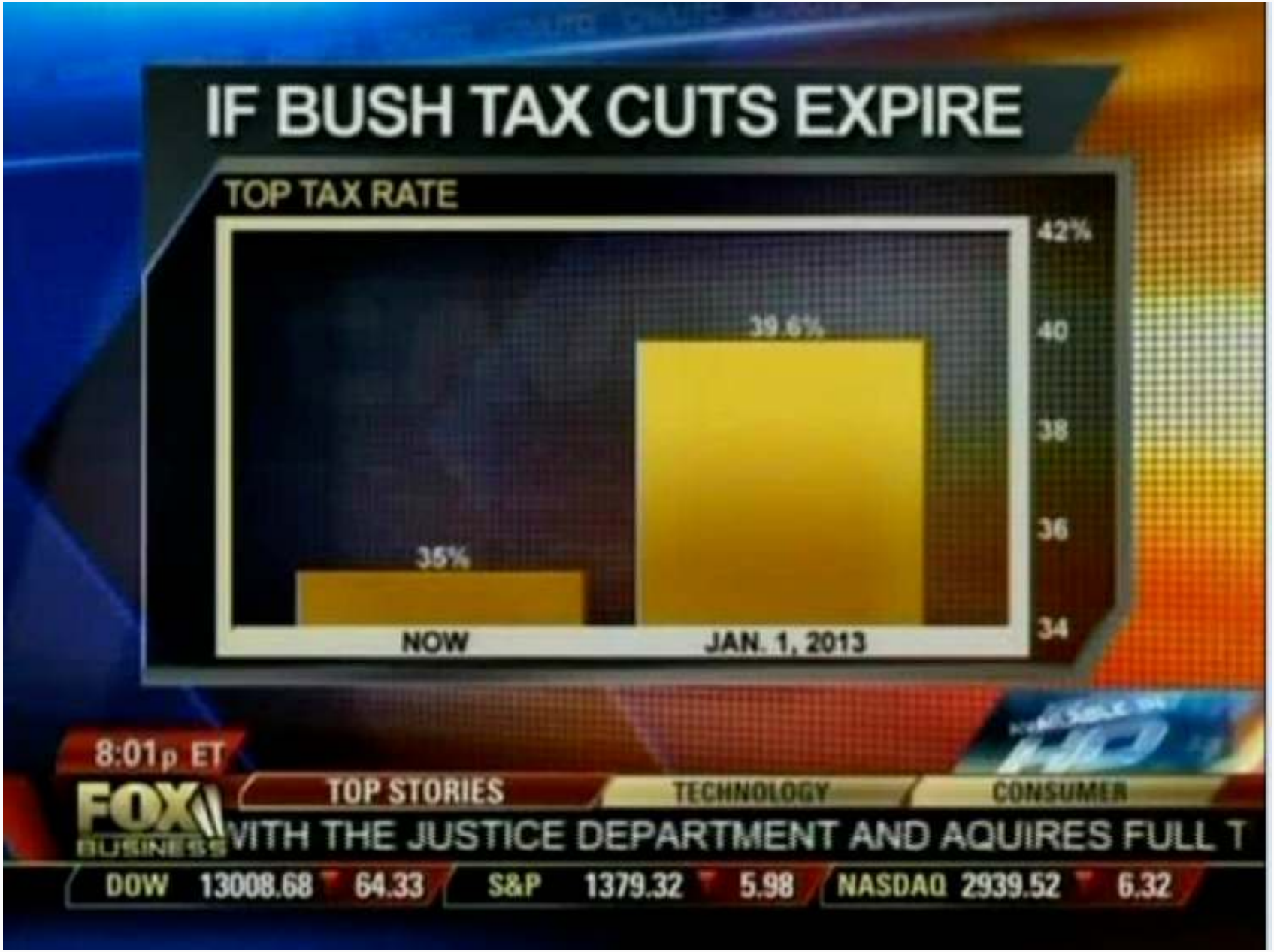


Maté (2017)

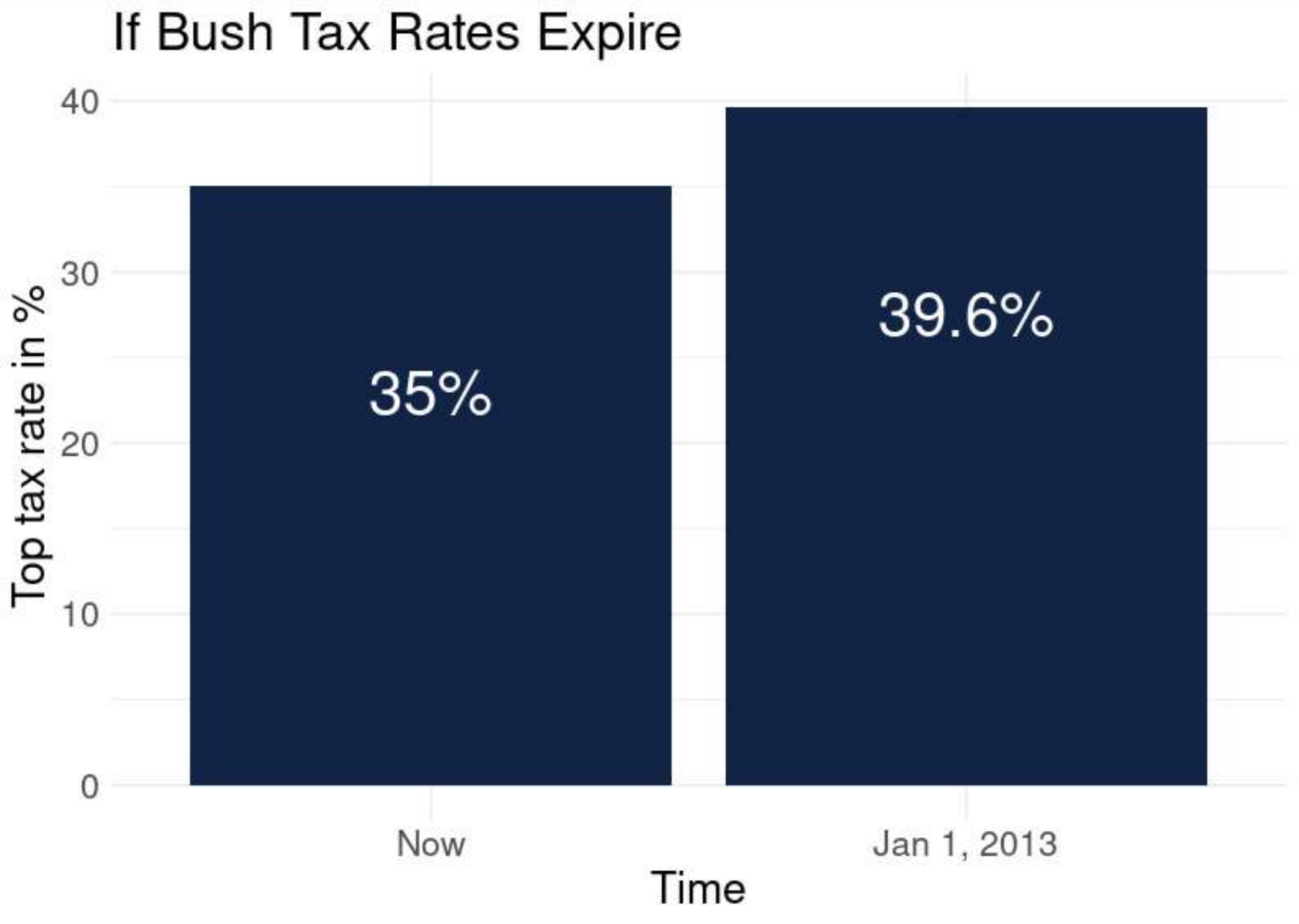
Bar plots: Stack (+ color)



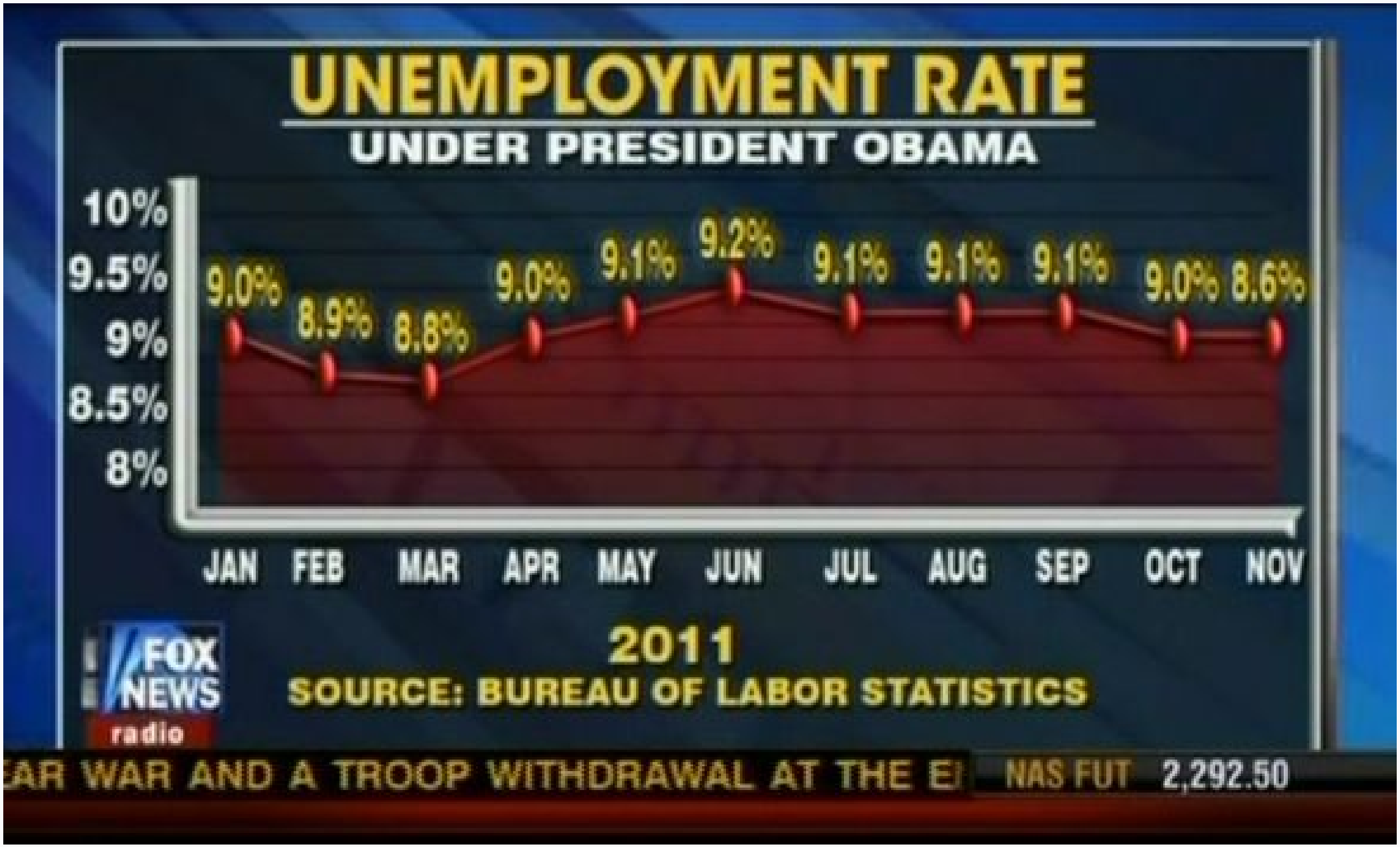
X/Y plots: Omit the baseline



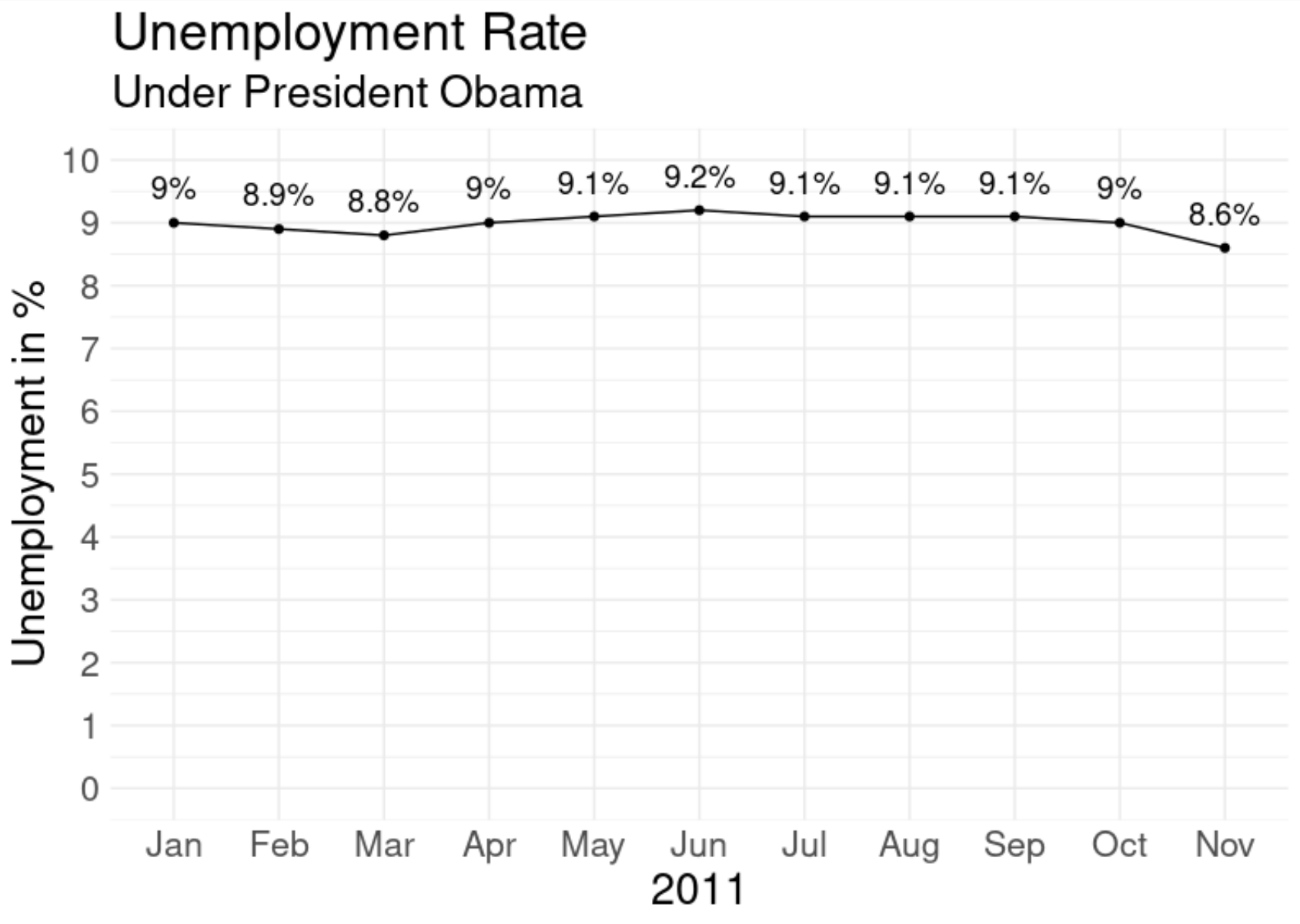
X/Y plots: Omit the baseline



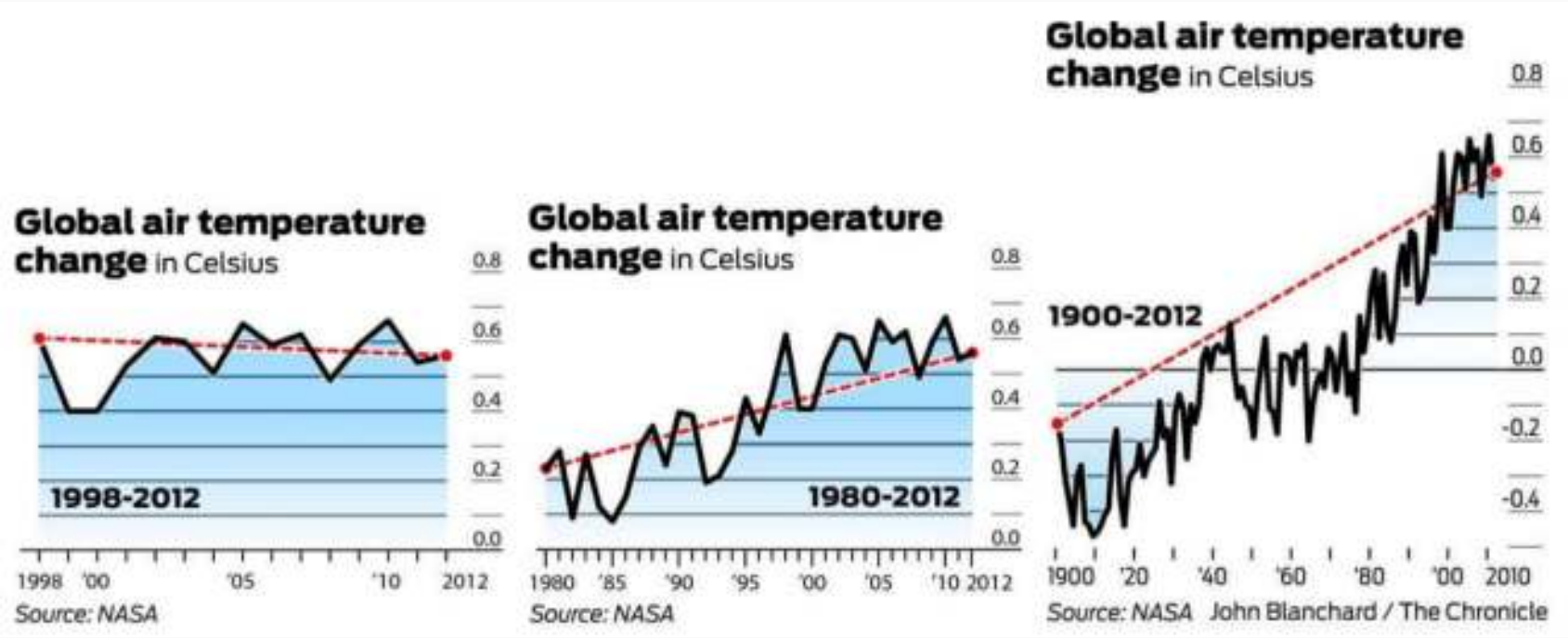
X/Y plots: Manipulate the Y-axis



X/Y plots: Manipulate the Y-axis

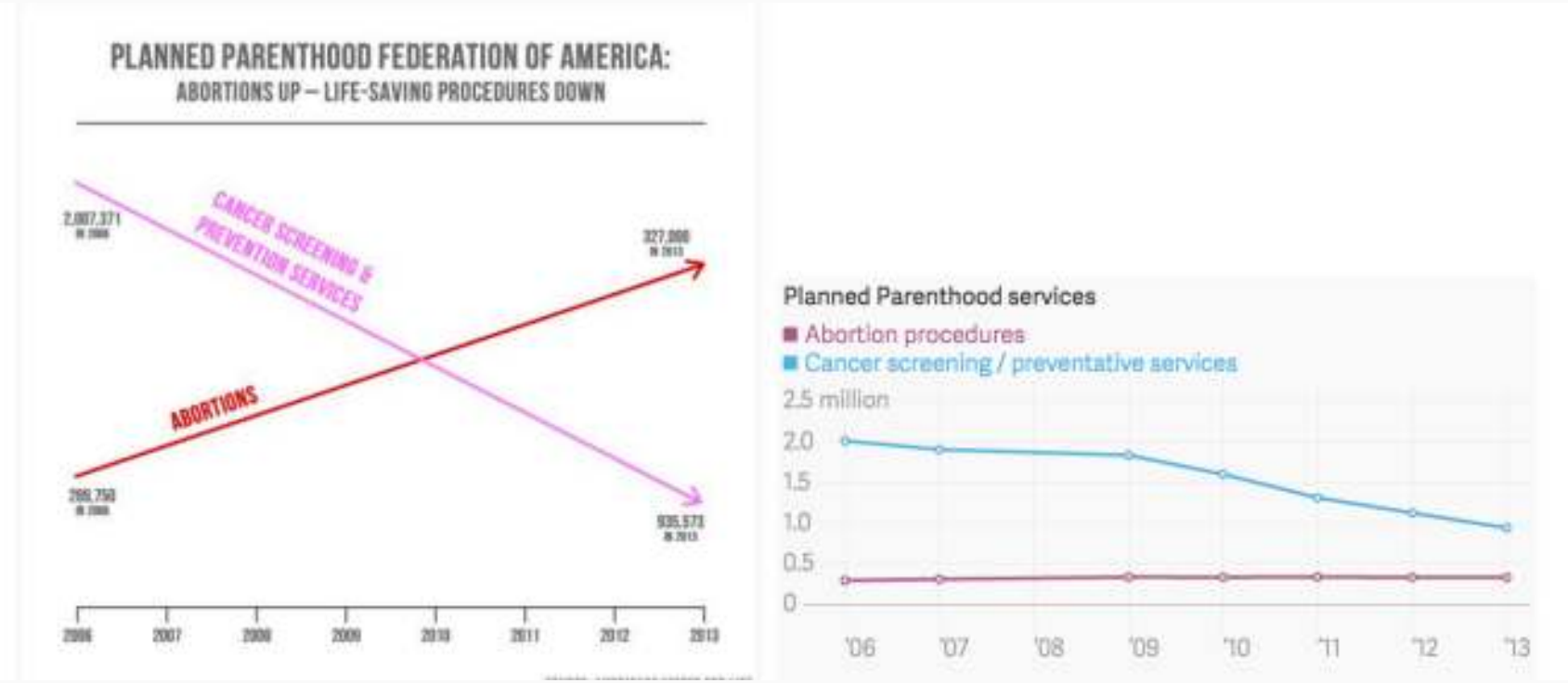


X/Y plots: Pick the range



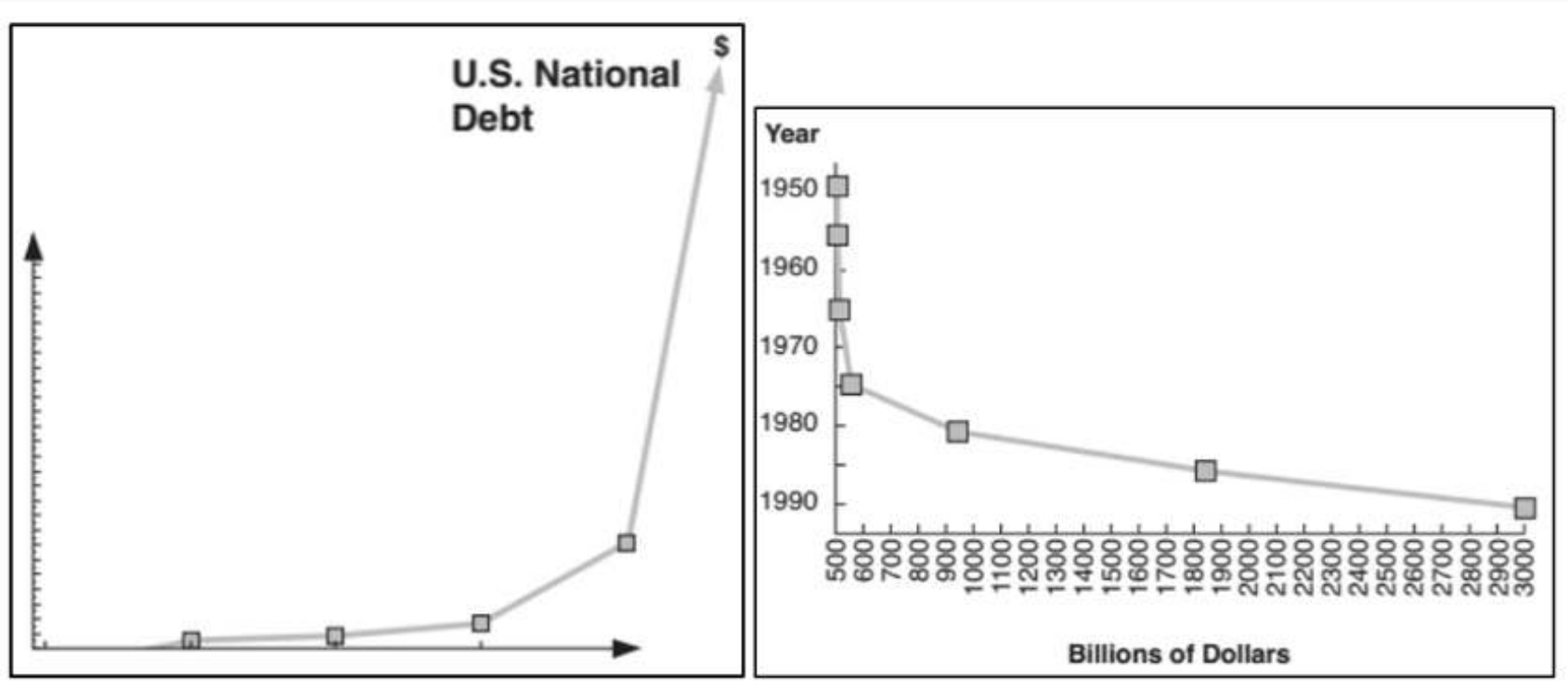
Fallenbüchel (2019)

X/Y plots: Fuck the y-axis entirely



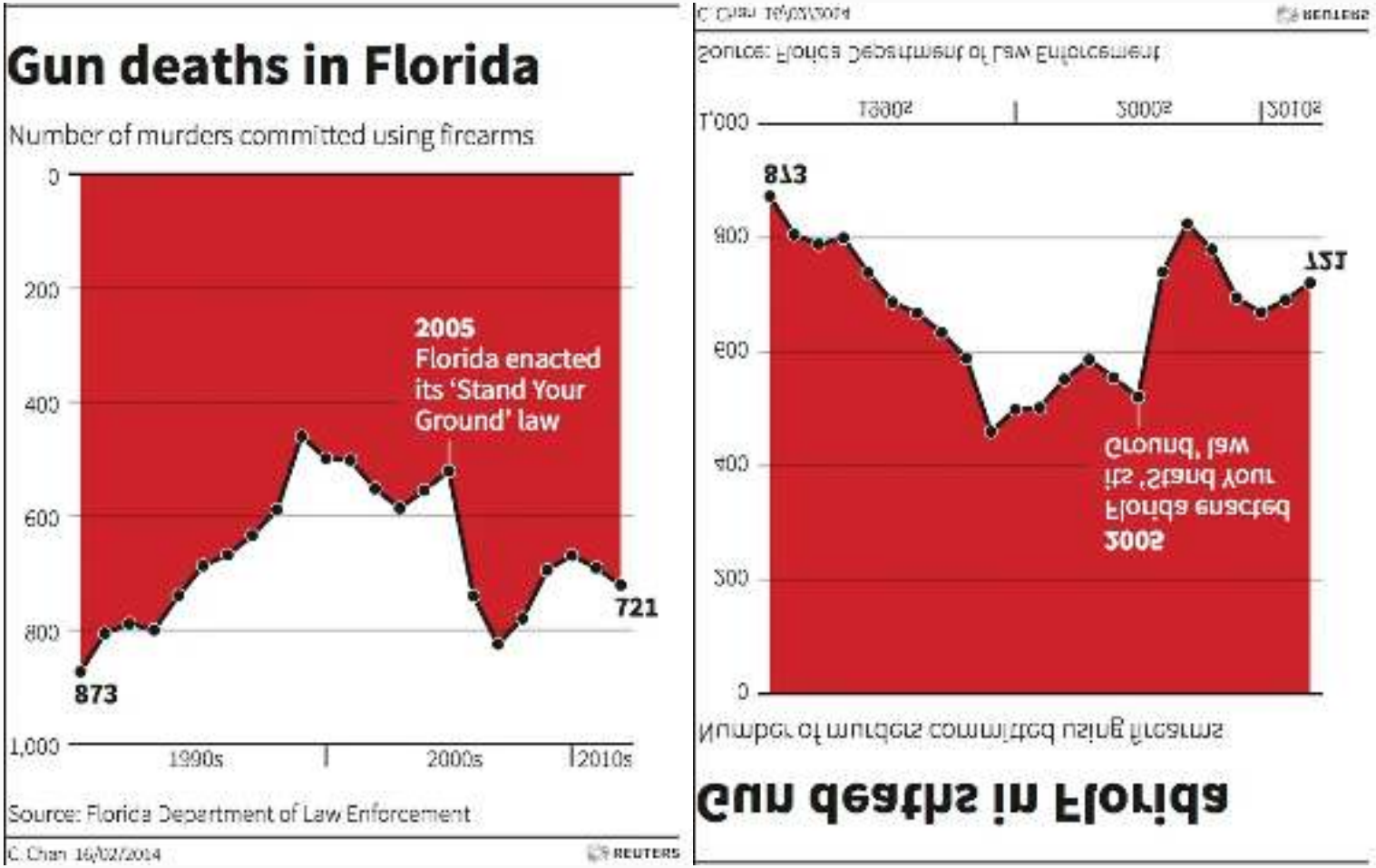
Fallenbüchel (2019)

X/Y plots: Flip plot



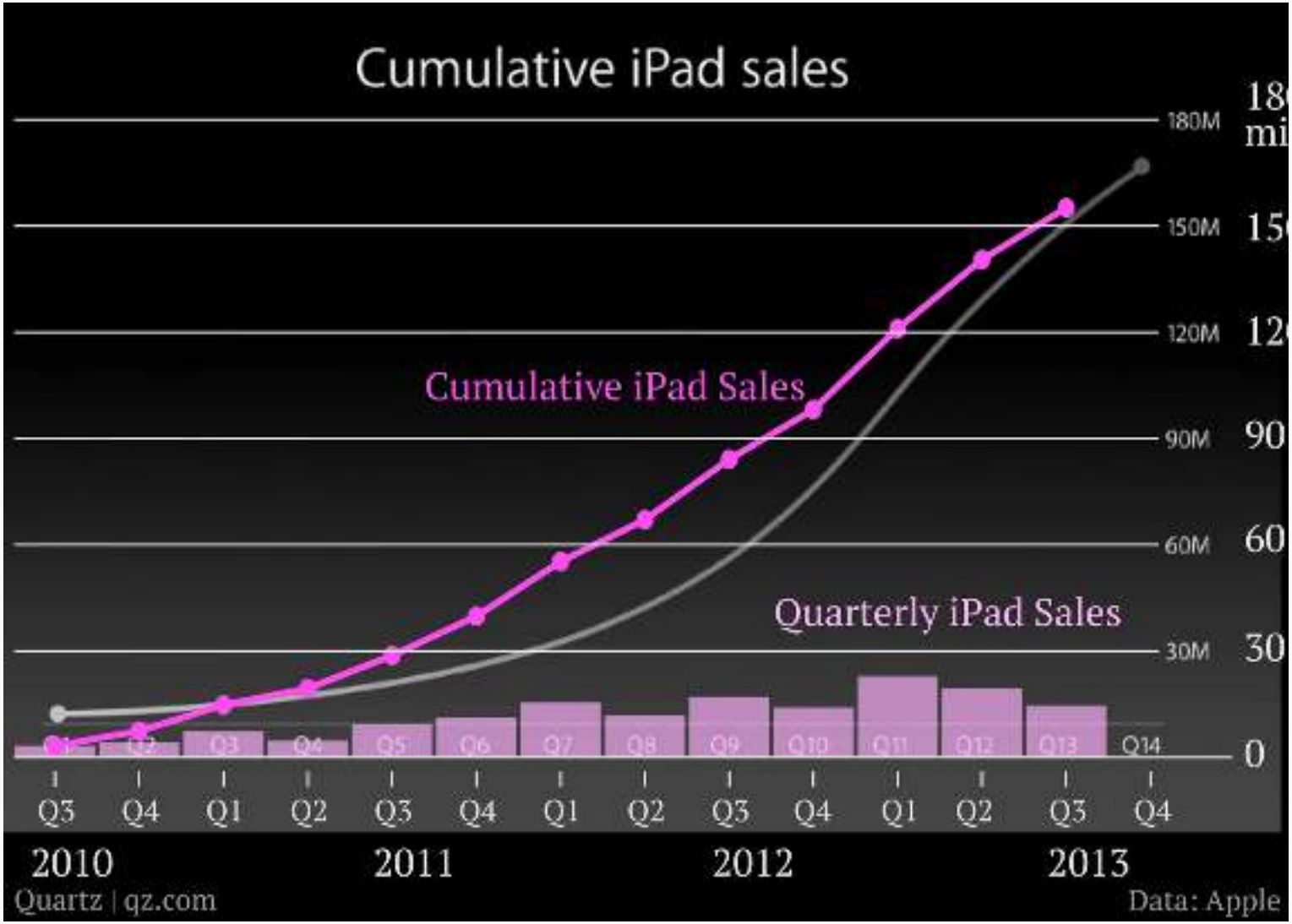
Fallenbüchel (2019)

X/Y plots: Flip plot



X/Y plots: Cumulative growth

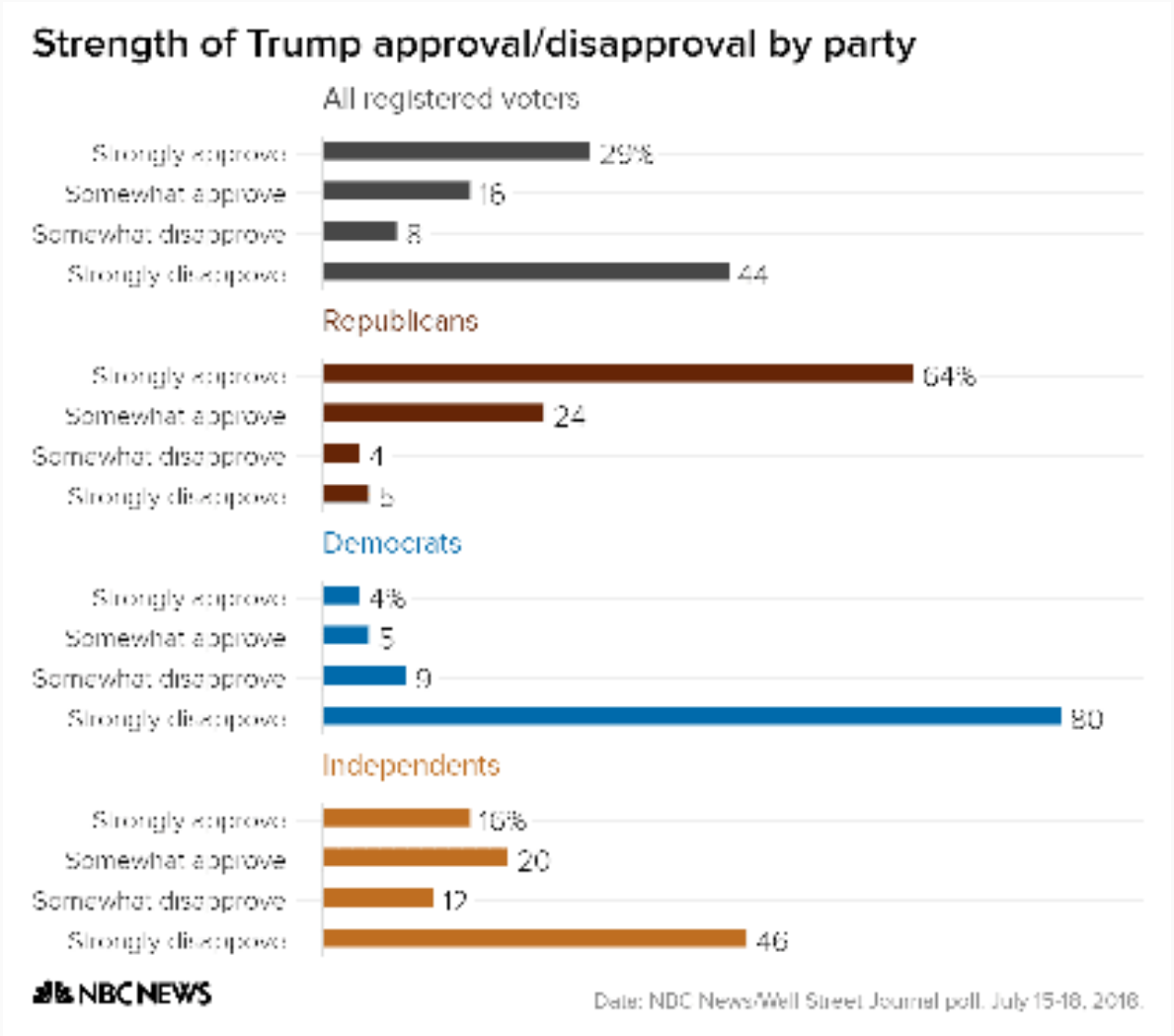
X/Y plots: Cumulative growth



Yanofsky (2013)

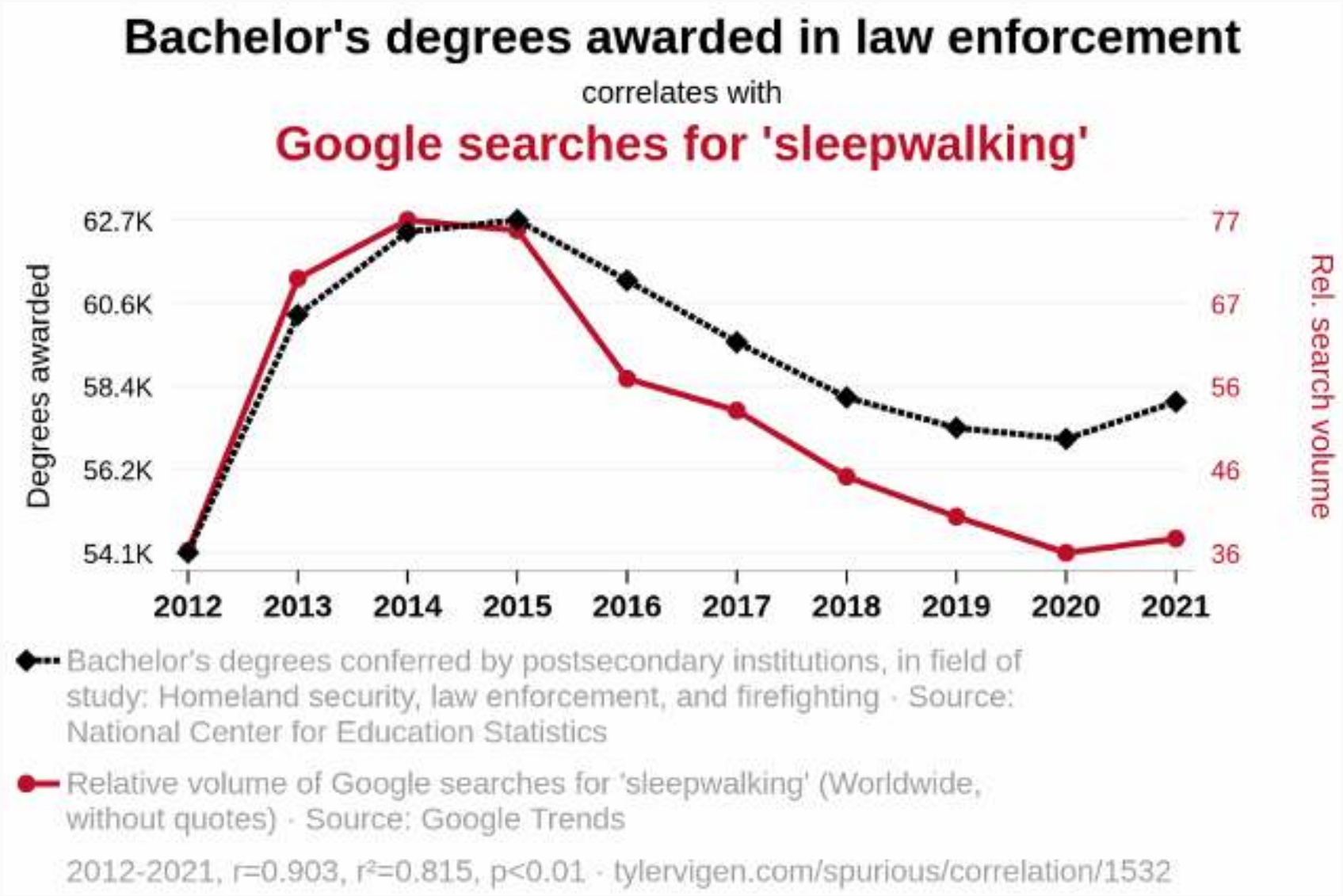
X/Y plots: Cherrypick data

X/Y plots: Cherrypick data



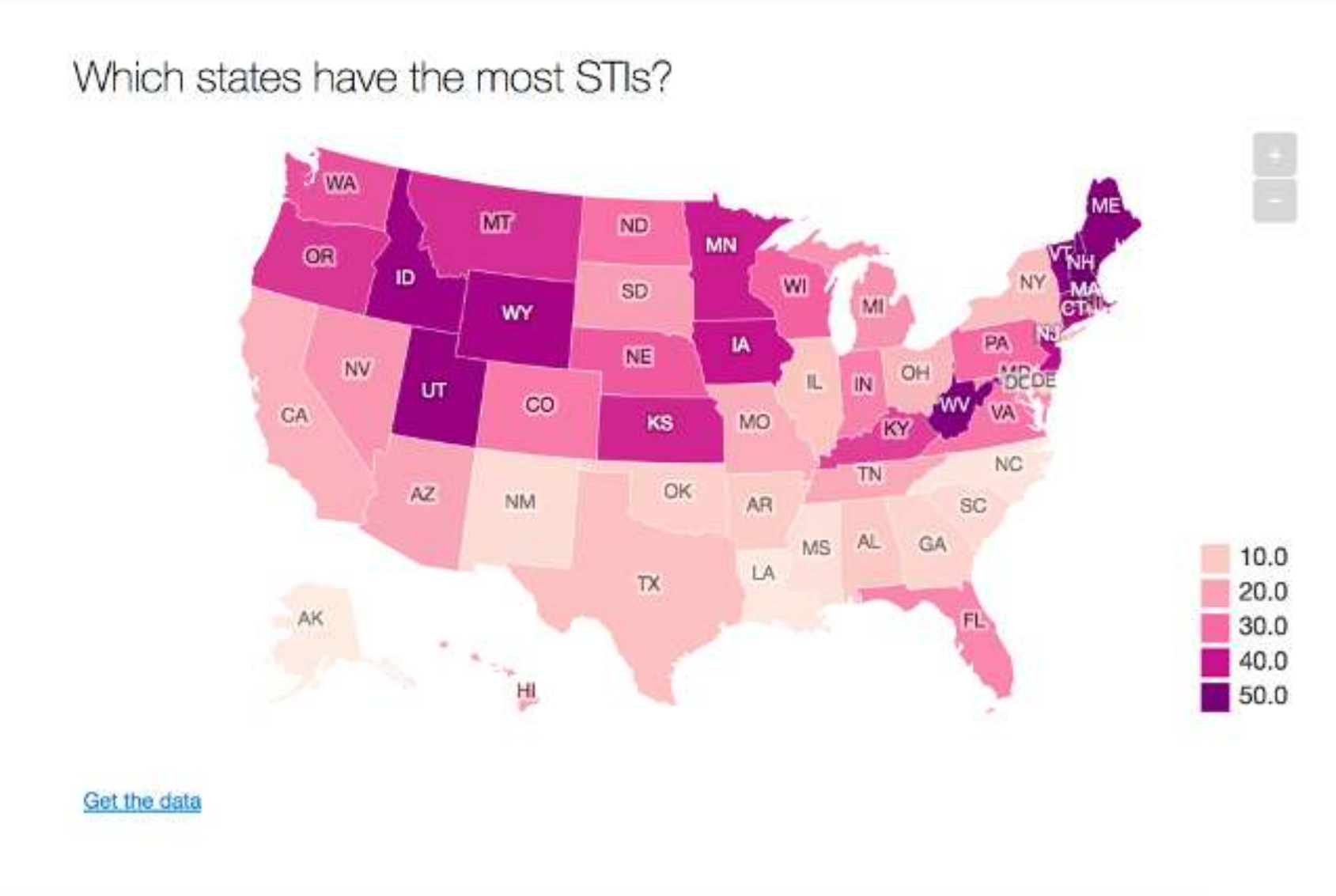
Murray (2018)

X/Y plots: Correlate axes



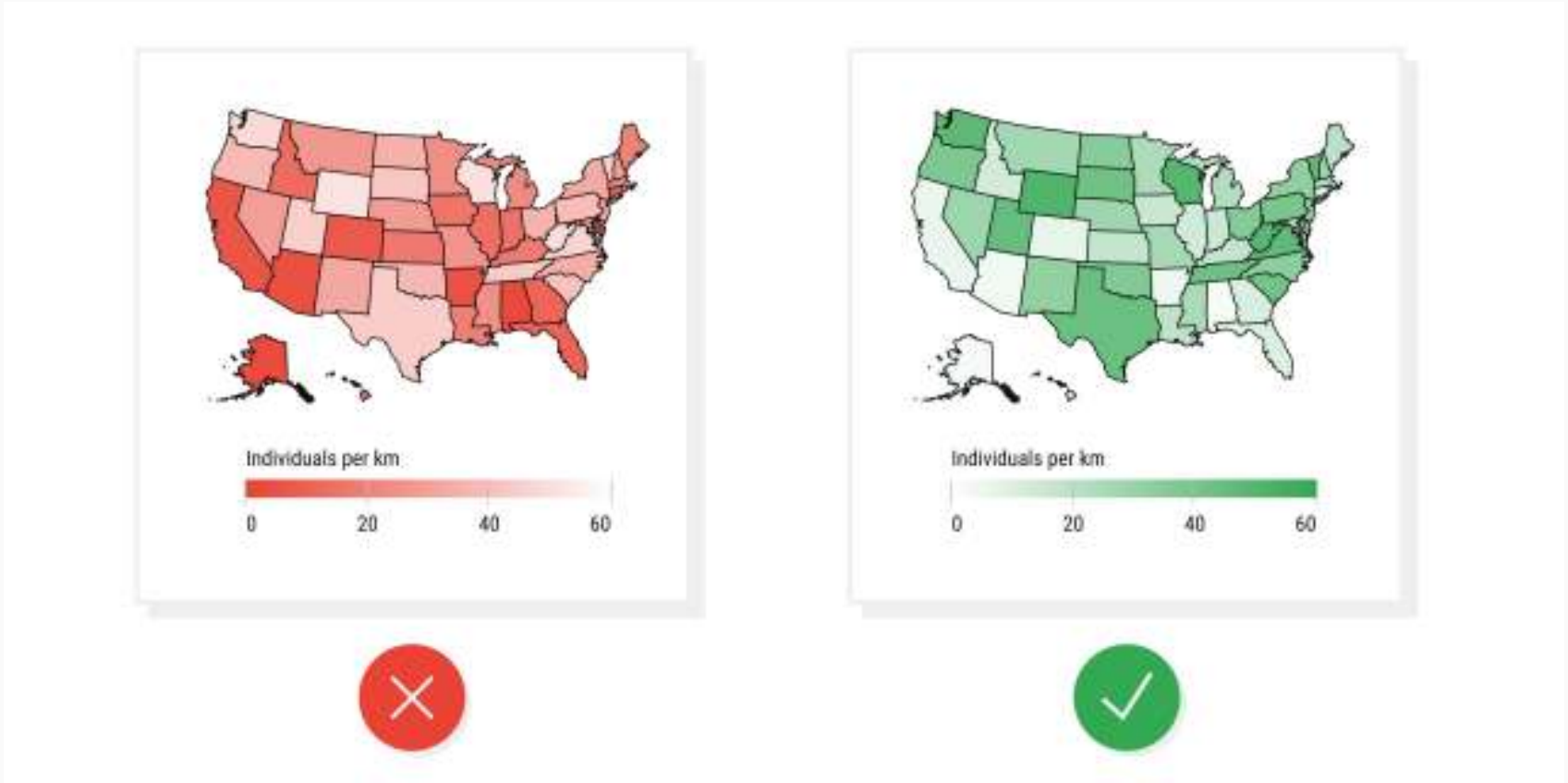
Vigen (2024)

Colors: go against convention



McCready (2020)

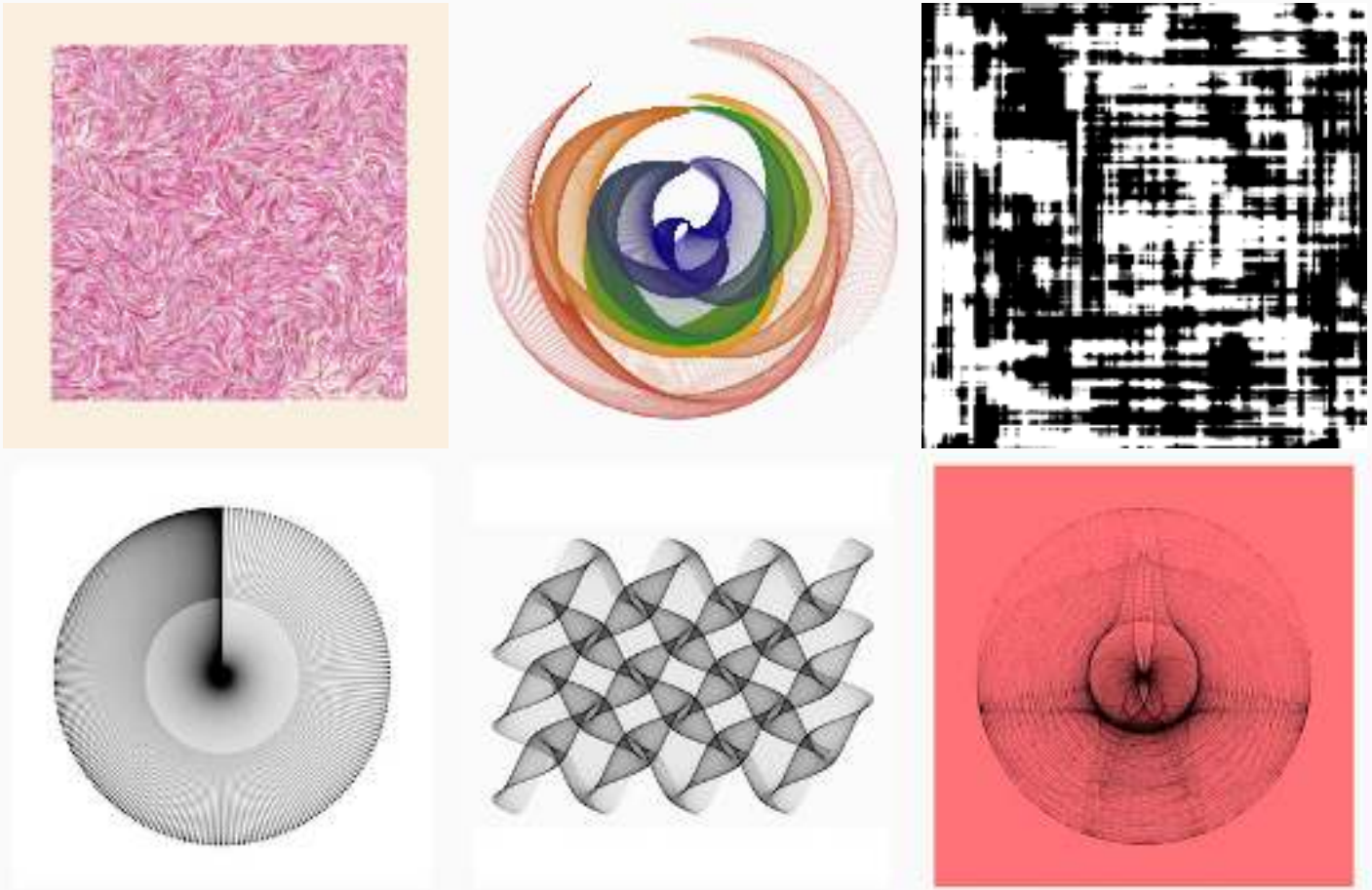
Colors: go against convention

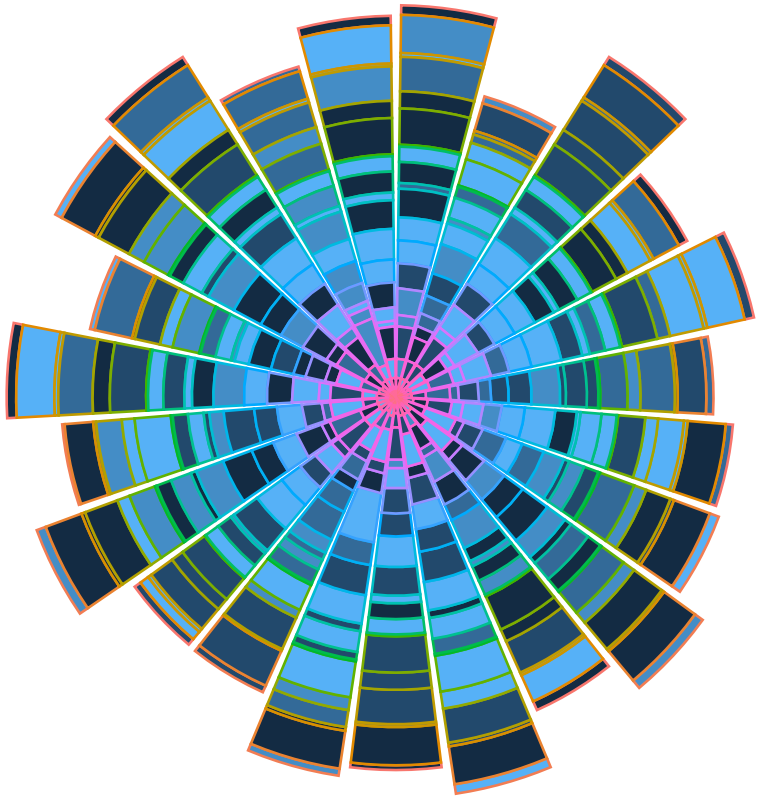


McCready (2020)

Generative art

Start out with `aRtsy` and `generativeart`, but also a mix of `ggplot2, ggforce, rayshader, ggpattern, gridExtra`



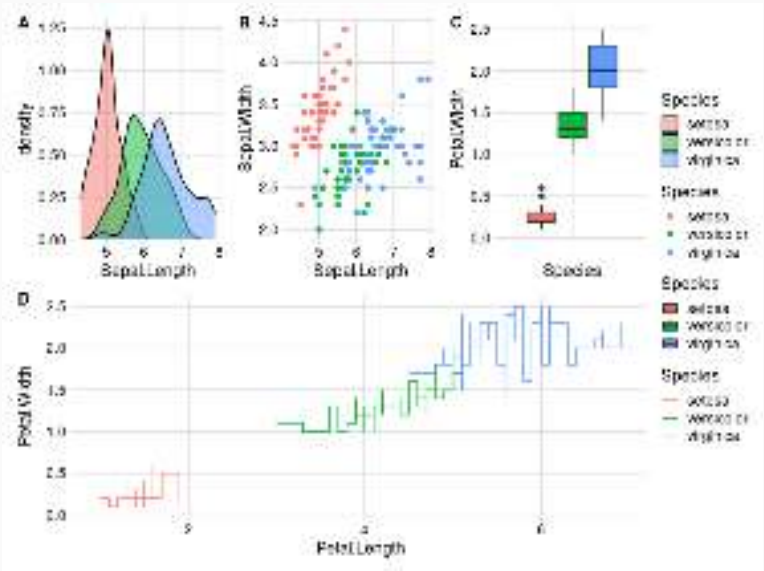


Moving on from R

Arranging plots

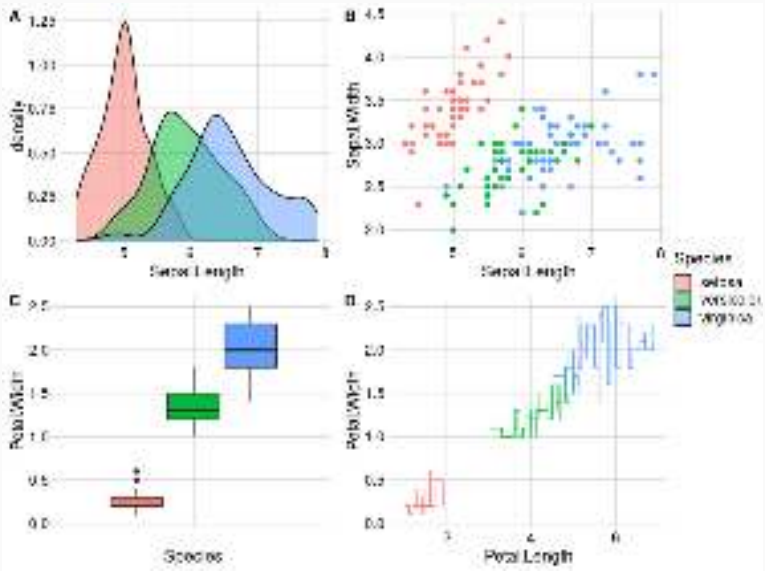
`patchwork`

makes it ridiculously simple to combine separate ggplots into the same graphic.



`cowplot`

provides various features to make plots beautiful, including aligning and arranging plots.



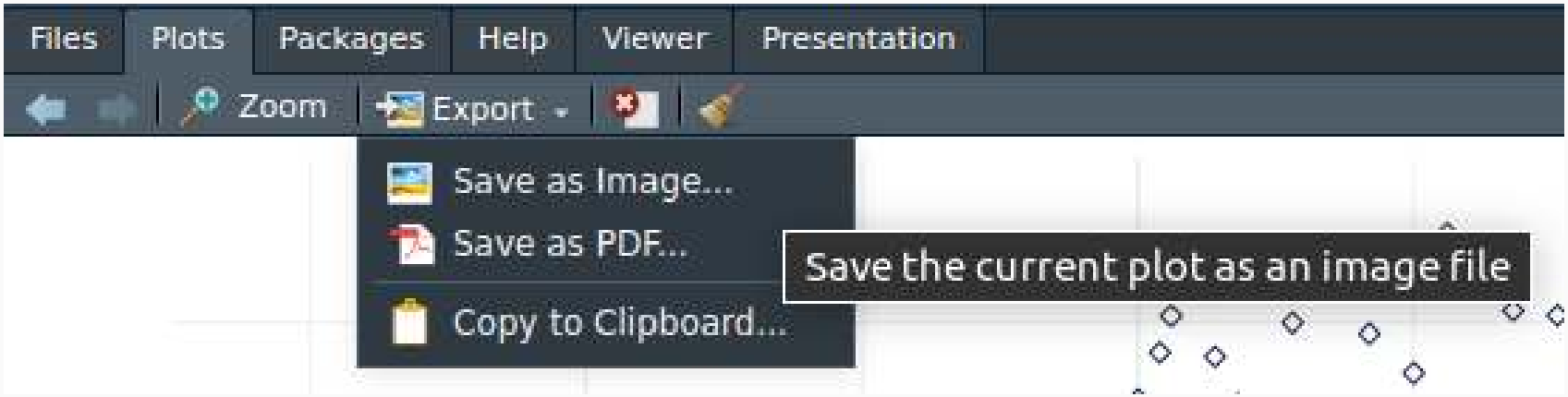
Exporting plots

`ggsave("FILENAME", width=NR, height=NR, dpi =300)`

Saves the last plot by default, but you can specify `plot = my.plot`

A common standard for high-quality prints is 300 DPI (dots per inch)

Choose PDF or PNG file extension for printing, SVG for more editing.



Exporting dataframes

Save a data frame to the **current working directory**.

```
write_csv(WHAT, "PATH TO WHERE", row.names=FALSE)   comma
write_tsv()                                          tab
write_excel_csv()                                    CSV for Excel
write_delim()                                        specify how to separate
```

```
write_csv(moses_accuracy, "Moses accuracy.csv",
          row.names = FALSE)

write_delim(moses_accuracy, "Moses accuracy.txt",
            row.names = FALSE, delim = ":")
```

Questions?

Wrap-up

Summary

- ✔ R programming basics and RStudio IDE
- ✔ write scripts
- ✔ file encoding, variable naming, and tidy code with pipes
- ✔ install and load packages
- ✔ import/export data from/to the working directory
- ✔ save and remove objects in the environment
- ✔ preprocess raw data (filtering, renaming, arranging, mutating, selecting, if else)
- ✔ make sense of data (merging, grouping, summarizing)
- ✔ print and visualize the results
- ✔ find help

Homework assignment due May 31st 15:30

- Complete assignment 6 ($\rightarrow$ ILIAS)
- Vote for the ugliest plot.
- Install Quarto: `https://quarto.org/docs/get-started/`
- Watch the Quarto introductory video: `https://youtu.be/_f3latmOhew?si=xxovQvYkUosC_4uB`

Next time: Reporting on data

References


Fallenbüchel, Florian (2019). *How To Lie With Charts: Elaboration on the presentation for the seminar "How do I lie with statistics?"* Report. URL: https://hci.iwr.uni-heidelberg.de/system/files/private/downloads/121410023/florian-fallenbuechel_report.pdf.


Maté, Aaron (Apr. 2017). *MSNBC's Rachel Maddow Sees a "Russia Connection" Lurking Around Every Corner*. URL: https://theintercept.com/2017/04/12/msnbcs-rachel-maddow-sees-a-russia-connection-lurking-around-every-corner/ (visited on 05/23/2024).

 McCready, Ryan (2020). *How to Avoid Misleading Graphs: Practical Tips and Examples*. URL: https://venngage.com/blog/misleading-graphs/ (visited on 05/23/2024).

 Murray, Mark (2018). *NBC/WSJ poll: Public gives Trump thumbs down on Russia, thumbs up on economy*. URL: https://www.nbcnews.com/politics/first-read/nbc-wsj-poll-public-gives-trump-thumbs-down-russia-thumbs-n893266 (visited on 05/23/2024).

 Ricks, Elizabeth (May 2020). *What Is a Pie Chart?* URL: https://www.storytellingwithdata.com/blog/2020/5/14/what-is-a-pie-chart (visited on 05/23/2024).

 Vigen, Tyler (2024). *Spurious Correlations*. URL: https://tylervigen.com/spurious-correlations (visited on 05/23/2024).

 Yanofsky, David (2013). *Apple is either terrible at designing charts, or thinks you won't notice the difference*. URL: https://qz.com/138458/apple-is-either-terrible-at-designing-charts-or-thinks-you-wont-notice-the-difference (visited on 05/23/2024).