

Digital Research Toolkit for Linguists

Week 8: Data visualization and exporting

Anna Pryslopska

May 27, 2024

Psycholinguistics and Cognitive Modeling Lab

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.

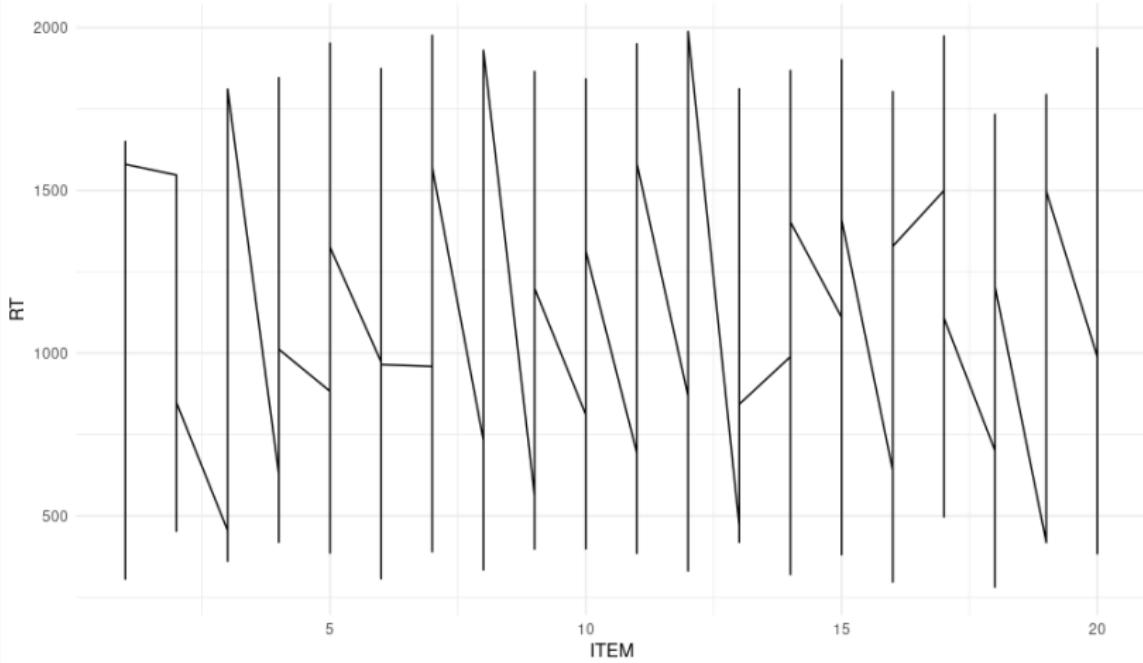
Recommended watching:

<https://www.youtube.com/watch?v=HPJn1CMvtmI>

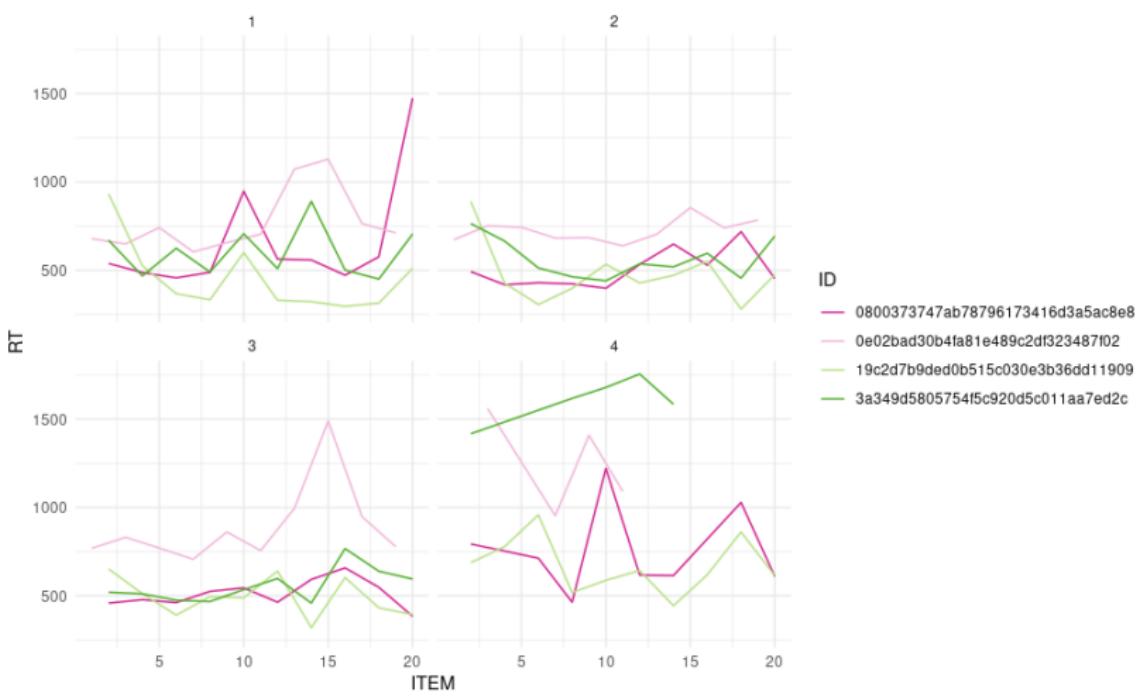
Ø `ggplot(noisy_rt.csv)`

Assign your data to a variable.

Noisy Channel Reading Time - Line Plot



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.

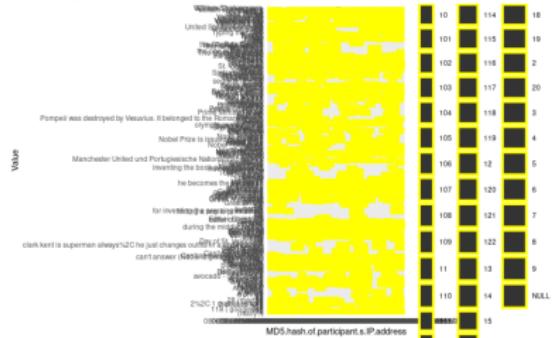
`stat = "summary", fun = "mean"`

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

`bin, smooth, density, ...`

Ugly plots

PLOT A



Noisy channel reading time data

Mean reading time by condition and sentence segment

Condition

Mean reading time

Sentence Segment

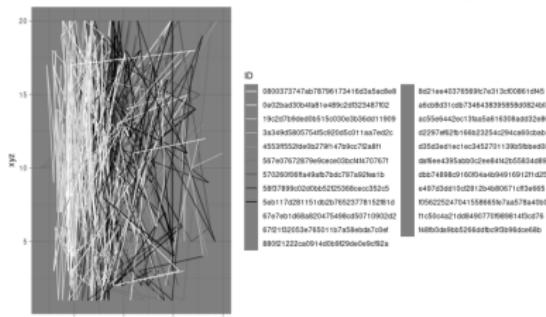
*implausibility

A T1 plot showing the mean reading time across all sentence segments by condition and sentence segment.

A TTT plot showing the mean reading time across all sentences by condition and sentence segment.

PLOT C

PLOT B



PLOT D

Questions?

Table of contents

1. Where are we this week?

2. Data visualization

3. Colors

4. Color use guidelines

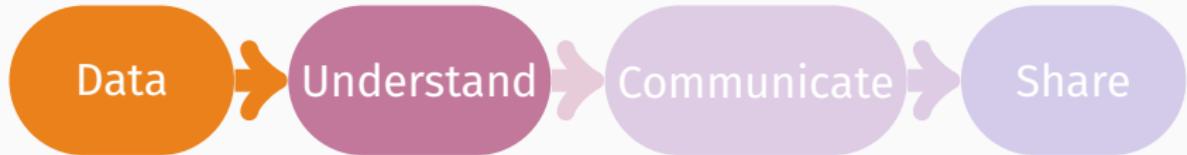
5. Lying with plots

6. Generative art

7. Moving on from R

8. Wrap-up

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 to efficiently communicate complex relationships & insights in a comprehensible way.

Goals

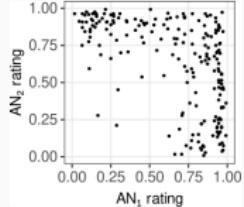
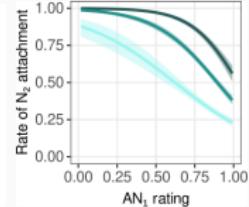
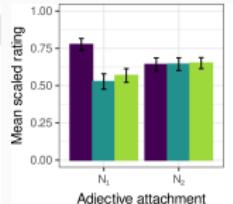
- 🗣 Communication & understanding
- 📊 Analysis & exploration
- ⚖️ Decision making



Visualization types

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

item	sentence	N1	N2	unclear
3	sommerliche Arbeitskleidung	p	■	p
11	silberne Bahnhofsuhr	p	■	p
12	amtliche Baugenehmigung	p	■	■
16	hohler Bienennstock	p	■	p
19	gefüttertes Brillenetui	p	■	p
22	lackiertes Bücherregal	p	■	p
23	evangelischer Büroangestellter	p	■	■



The 4 Principles of Accessibility

Web Content Accessibility Guidelines

P Perceivable

O Operable

U Understandable

R Robust

ggplot2

a layered grammar of graphics

Themes
Coordinates
Statistics
Facets
Geometries
Aesthetics
Data



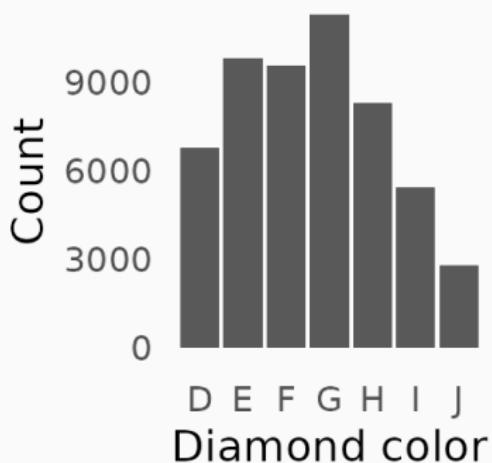
esquisse

GUI for exploring data based on ggplot2



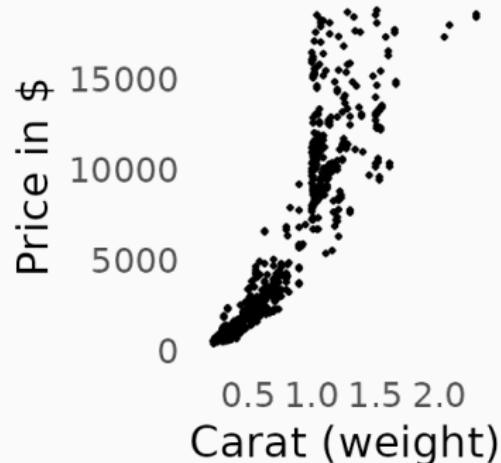
Histogram

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.



Adobe	color.adobe.com
Coolors	coolors.co
LearnUI	www.learnui.design

Test the colors for accessibility and BW print:

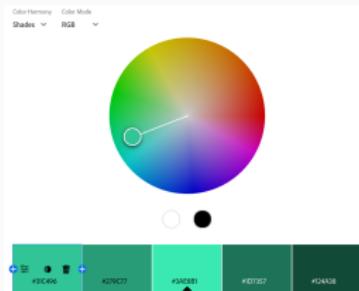
webaim.org/resources/contrastchecker

```
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

A screenshot of the Extract from image feature in Adobe Color. At the top, there are tabs for "Color Wheel", "Extract Theme", "Extract Gradient", and "Accessibility Tools". The "Extract Theme" tab is selected. Below it is a section titled "Color mood" with a descriptive text and a "Choose a color mood" button. Underneath are three radio buttons: "Colorful" (selected), "Bright", and "Muted". To the right is a square target icon with concentric circles and colored dots (black, orange, yellow, red). Below the target is a horizontal color bar with segments in black, orange, yellow, and red.

Check contrast

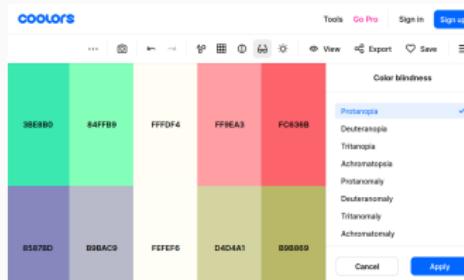
A screenshot of the Contrast checker tool in Adobe Color. It shows two color swatches: a dark blue "#001F3F" and a light blue "#A9C9E8". Below the swatches is a text box stating "Contrast: 3.84 : 1". There are also sections for "Text Color" and "Background Color" with sliders and dropdown menus. At the bottom, there are three cards: "A high color contrast makes anything easier to read", "A high color contrast makes anything easier to read", and "Create contrast".

Browse trends

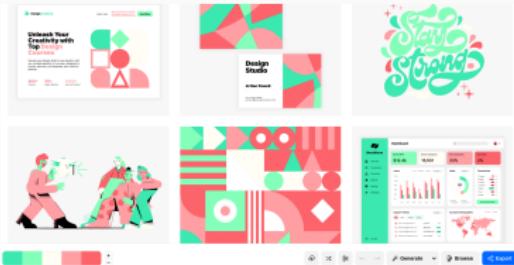
A screenshot of the Trends section in Adobe Color. At the top, there is a search bar with the text "Jen Spring" and a list of popular searches: "Summer", "Neutral palette", "Primary colors", "Neon", "Spring", "Orange", "Yellow", "Autumn", and "Neon". Below the search bar is a grid of nine images. Each image is a thumbnail of a design or scene with a distinct color palette. Some thumbnails include small text labels like "Summer", "Neon", and "Autumn".

Coolors

Assemble colors and check for color blindness



Visualize the palette



Browse trends



Generate color palettes

PALETTE SINGLE HUE DIVERGENT

PALETTE GENERATOR

NUMBER OF COLORS: 7

BACKGROUND COLOR: LIGHT DARK

#003f5c #374c80 #7a5195 #bc5090 #e15675 #ff764a #ffaa00

ACTIONS: COPY HEX VALUES EXPORT AS SVG

IN CONTEXT:

Check contrast

Show me the closest variations of  #4ac4e2
that contrast against  #ffffff
enough to meet  AA Guidelines ?

RESULTS

FOR LARGE/BOLD TEXT ?

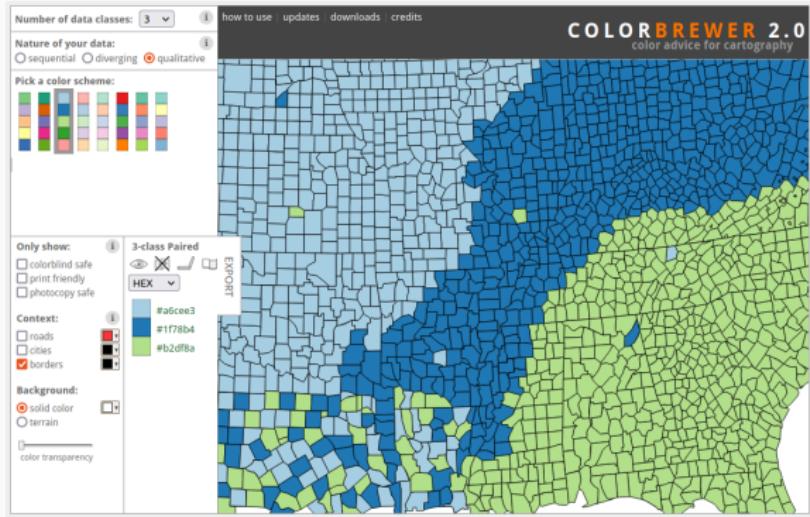
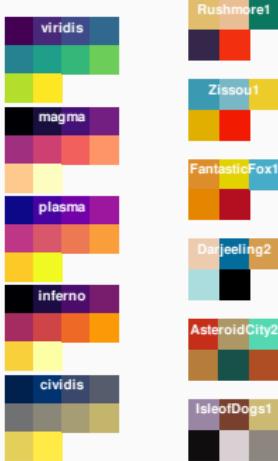
Try this combo instead:



FOR SMALL TEXT ?

Try this combo instead:





There's a package for that

RColorBrewer

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

viridis

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

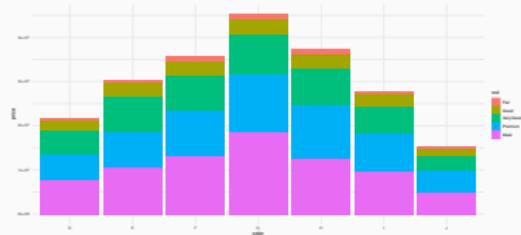
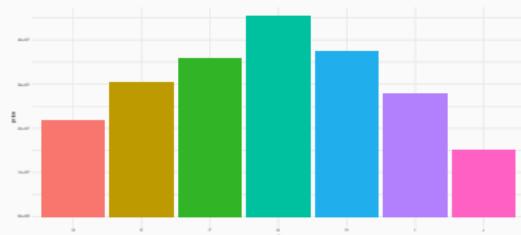
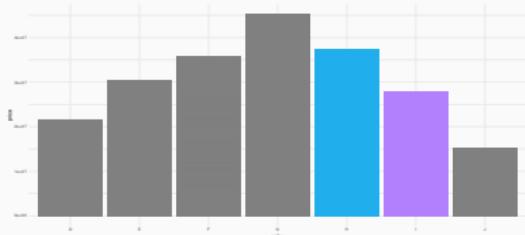
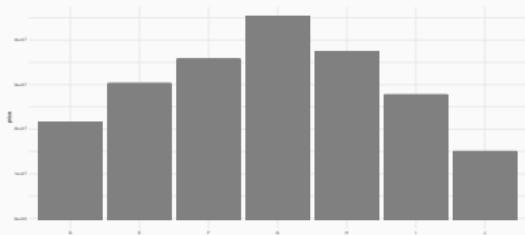
wesanderson

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

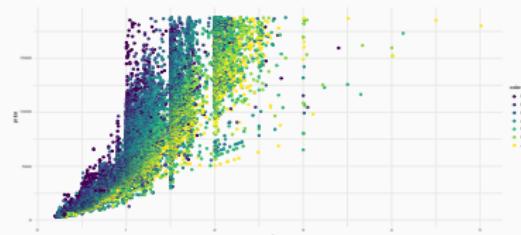
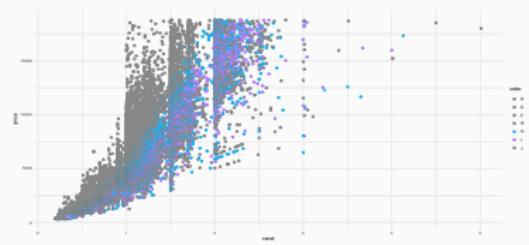
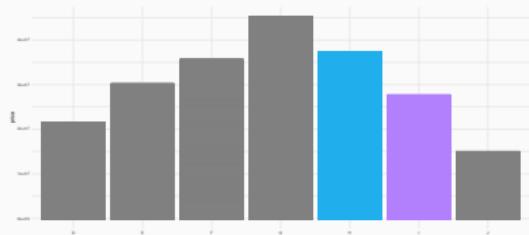
Others, e.g. ggsci (for journals)

Color use guidelines

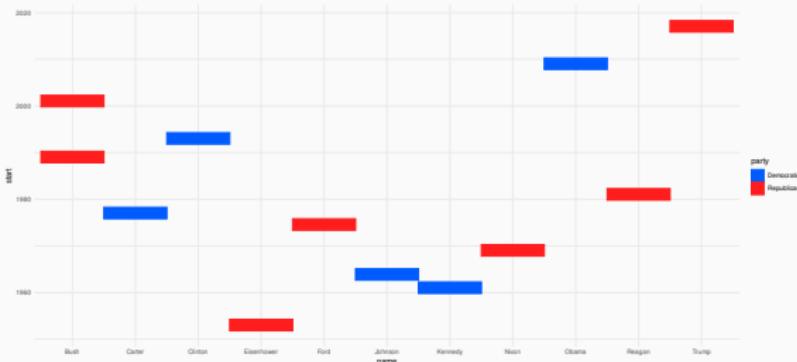
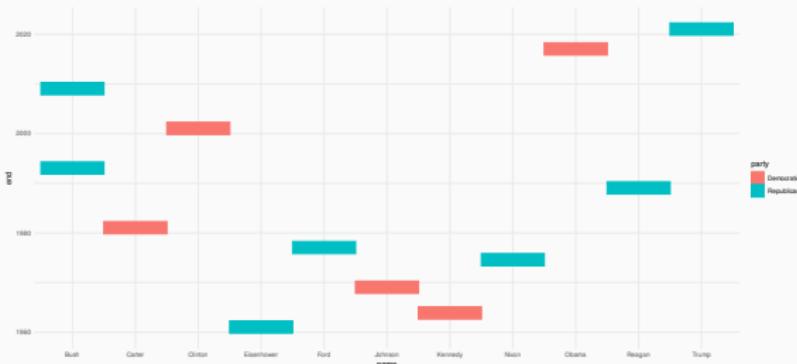
Avoid unnecessary usage of color



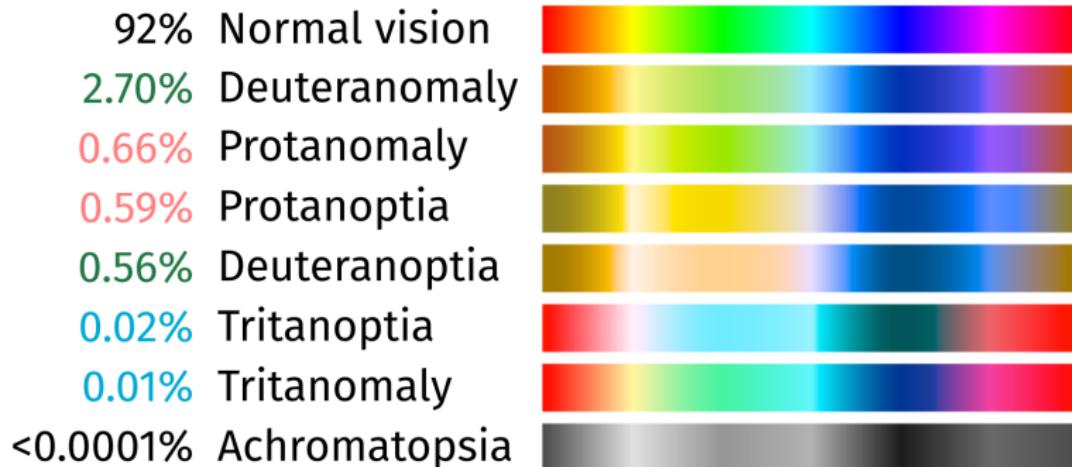
Be consistent



Use in a meaningful way

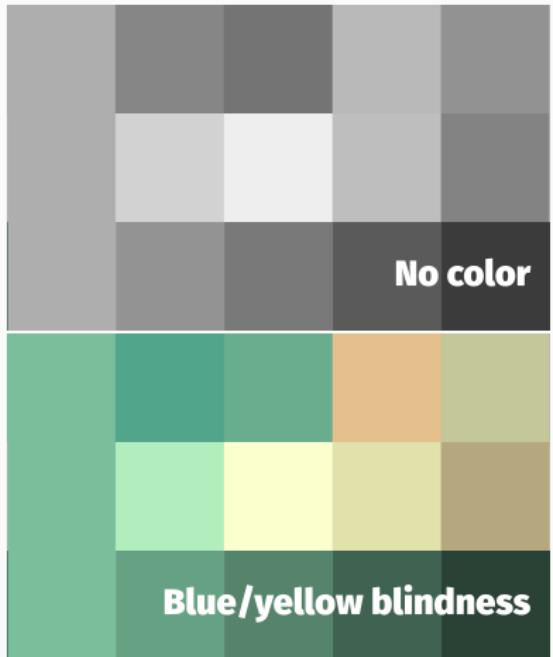


Pay attention to color blindness



1–2 people in this class

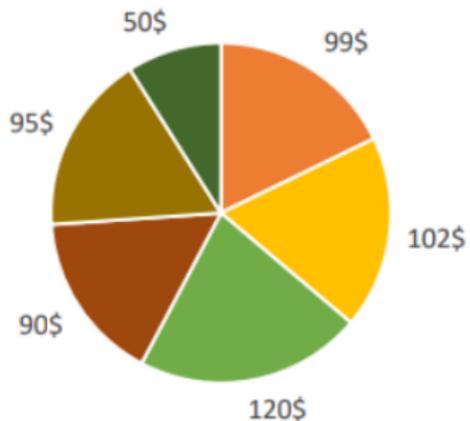
Perceivable colors



Lying with plots

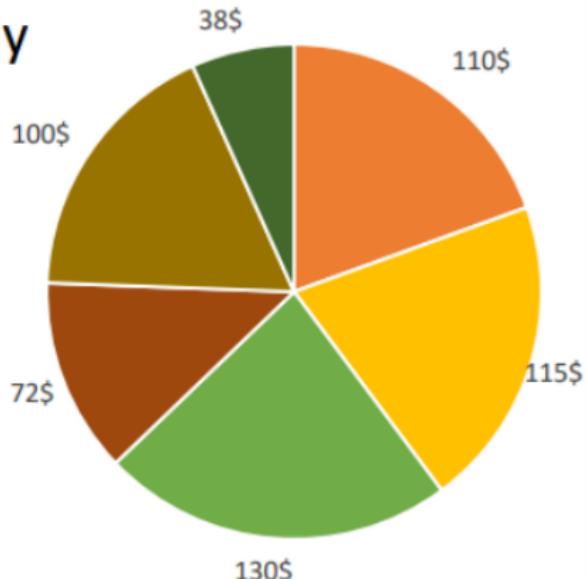
Pie plots: Manipulate the size

Total Sales by Company
(Millions)



Last year

■ Apple ■ Amazon ■ Google ■ Microsoft ■ Facebook ■ Volkswagen

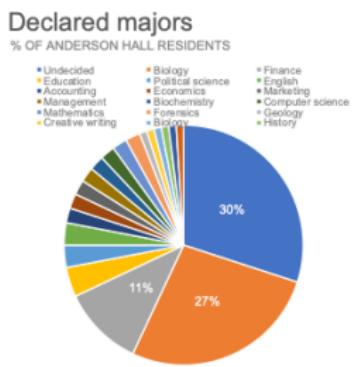


This year

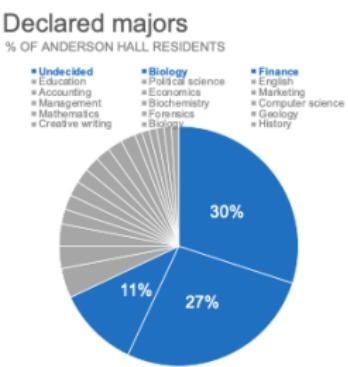
Fallenbüchel (2019)

Pie plots: Manipulate categories

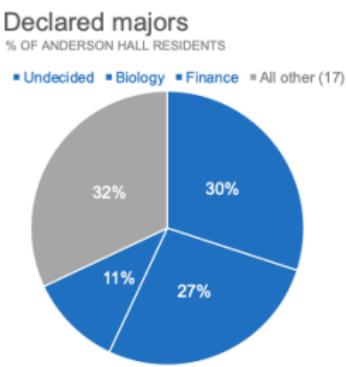
Too many slices



De-emphasize

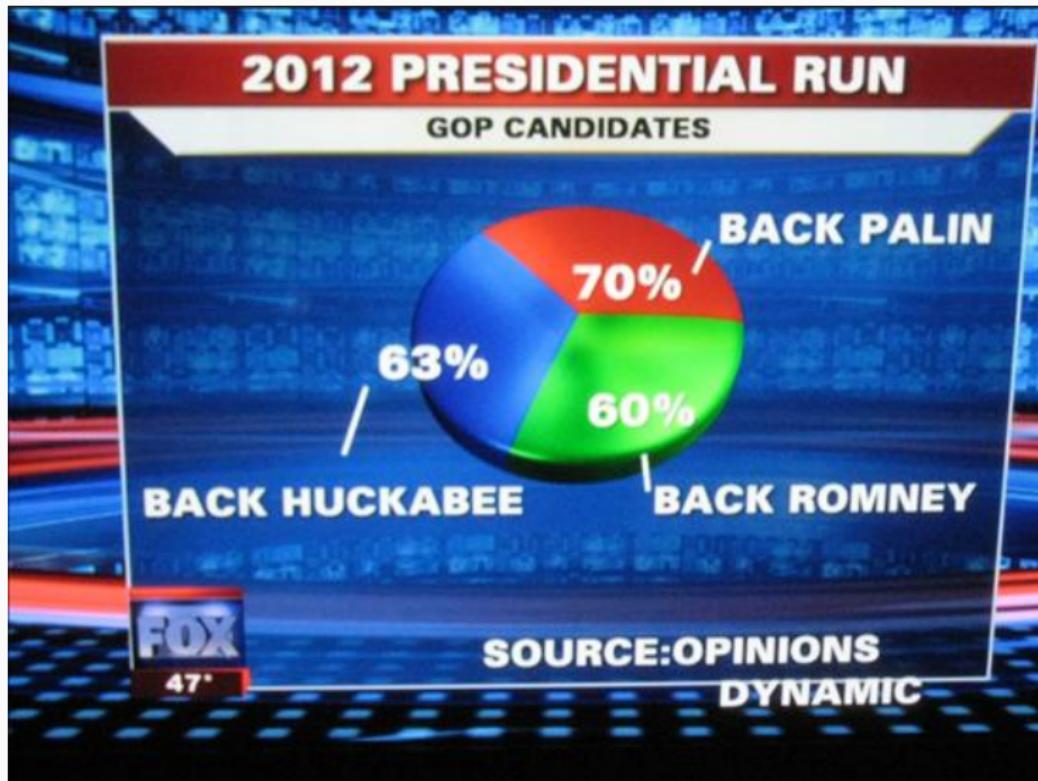


Aggregate



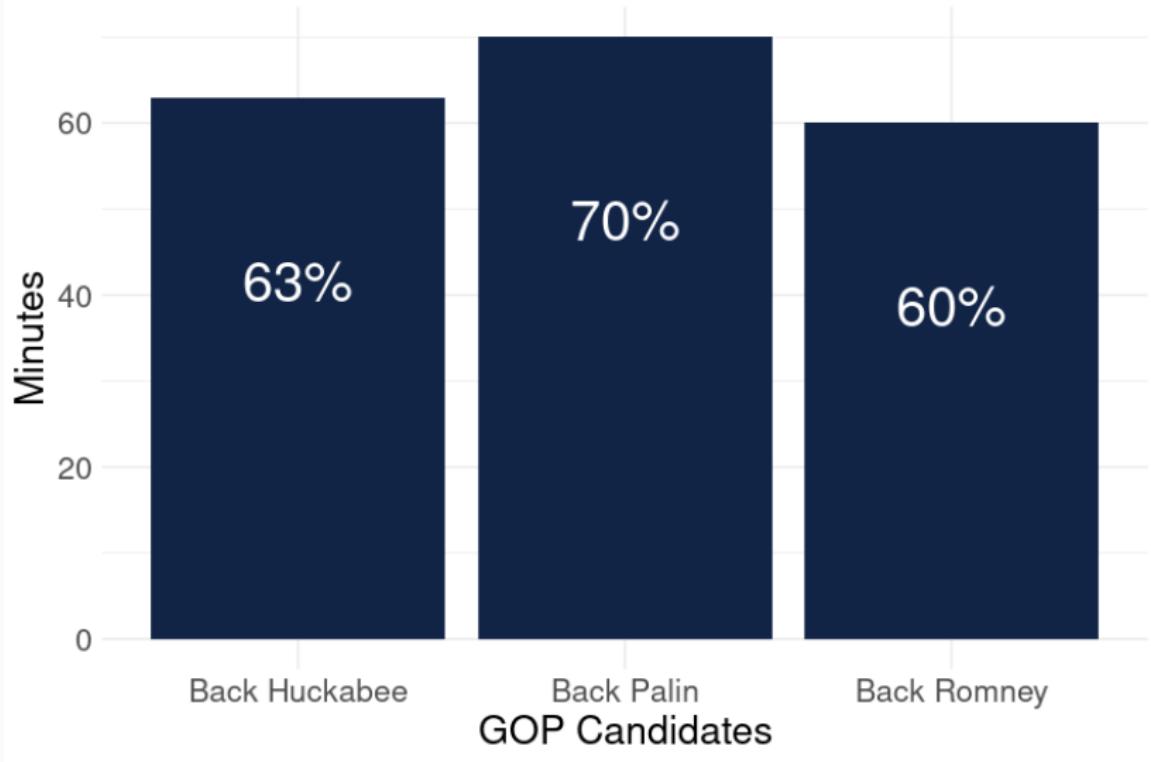
Ricks (2020)

Pie plots: Wrong plot type



Pie plots: Wrong plot type

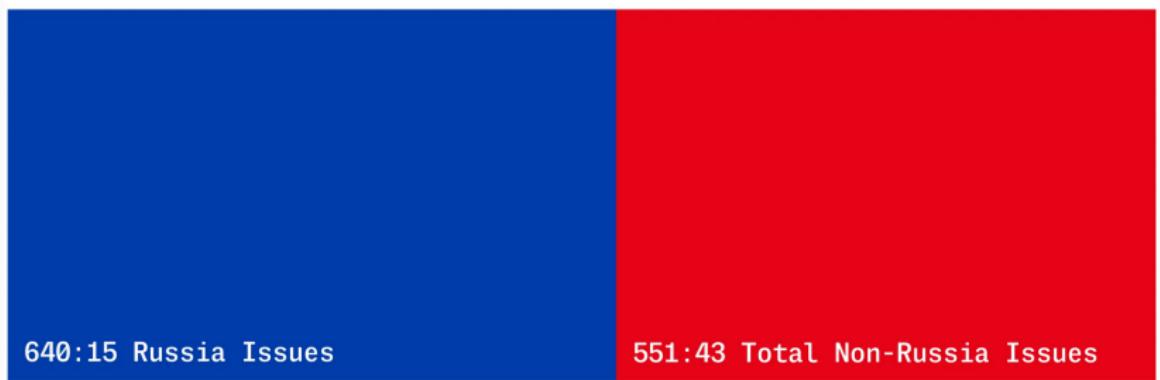
2012 Presidential Run



RUSSIA ISSUES VS. NON-RUSSIA ISSUES

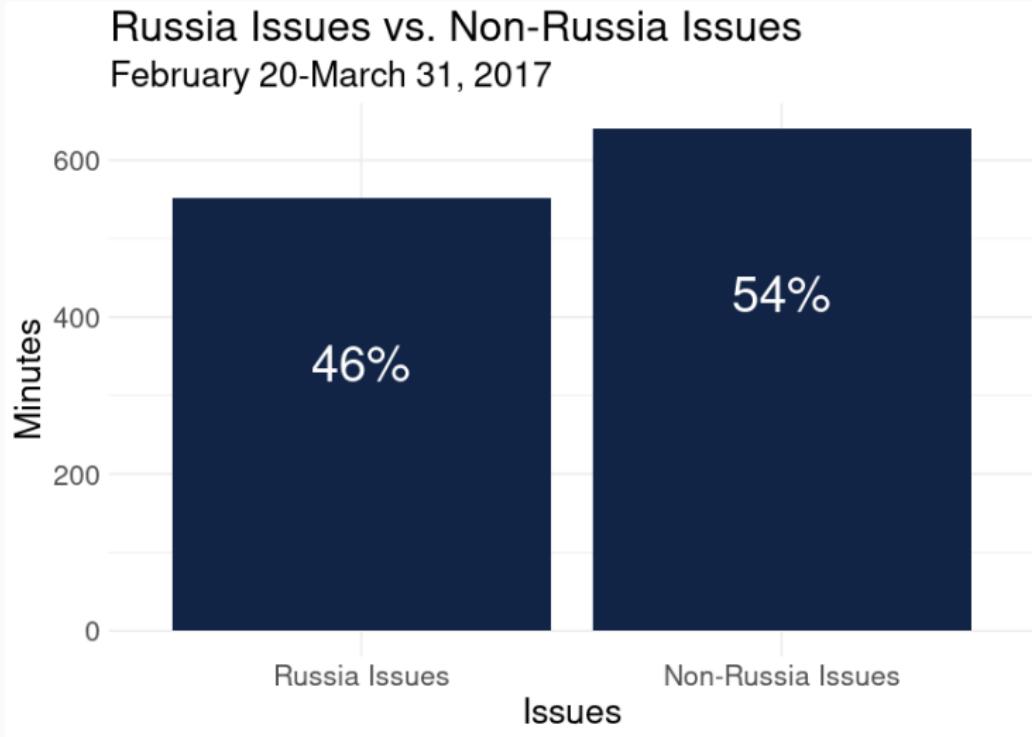
February 20–March 31, 2017

1191:58 (min:sec) Total Show Minutes

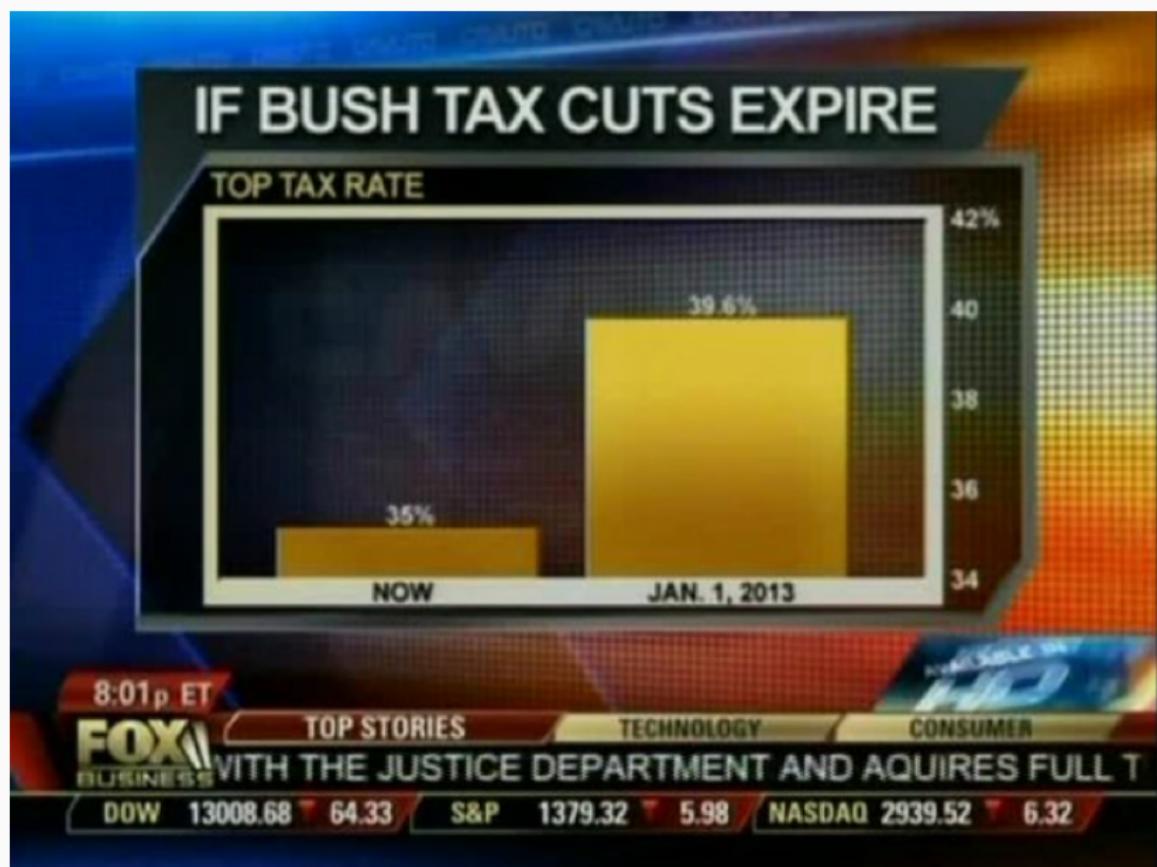


Maté (2017)

Bar plots: Stack (+ color)

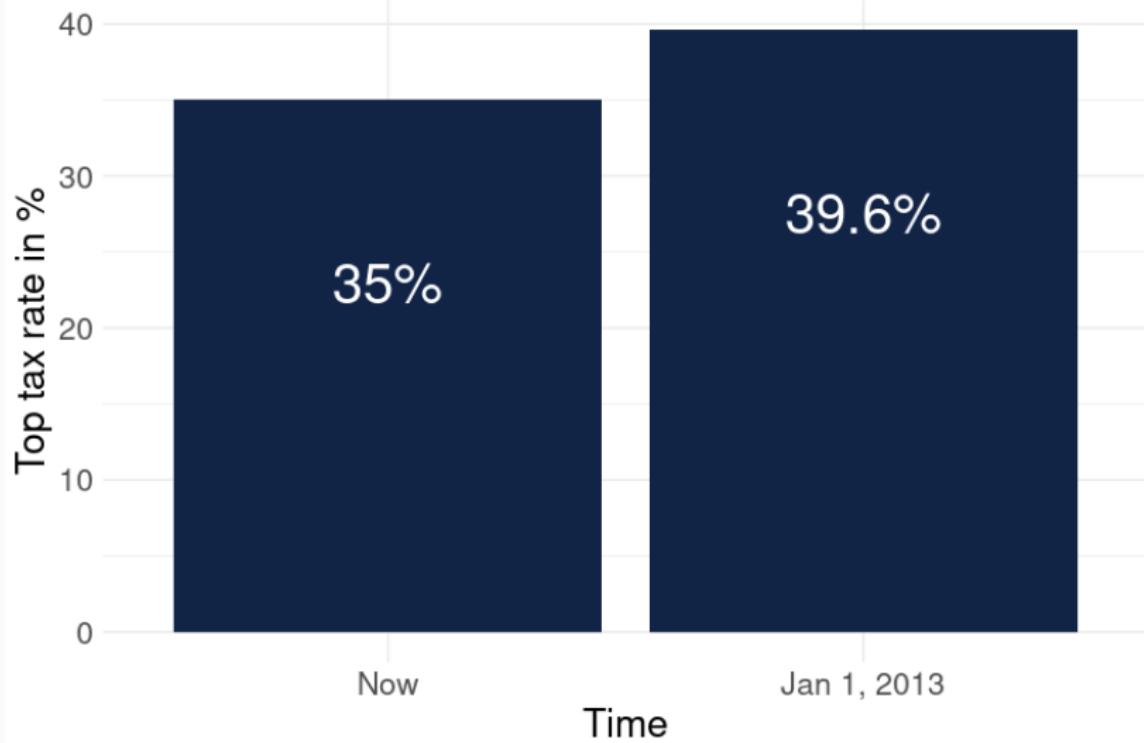


X/Y plots: Omit the baseline

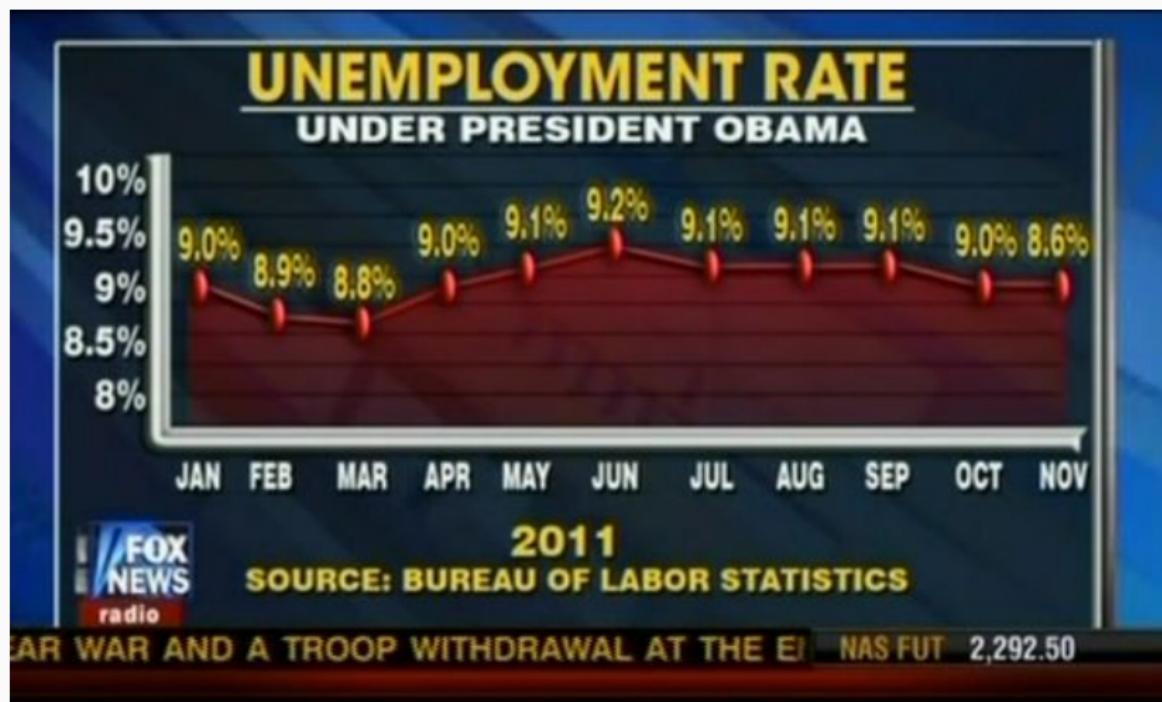


X/Y plots: Omit the baseline

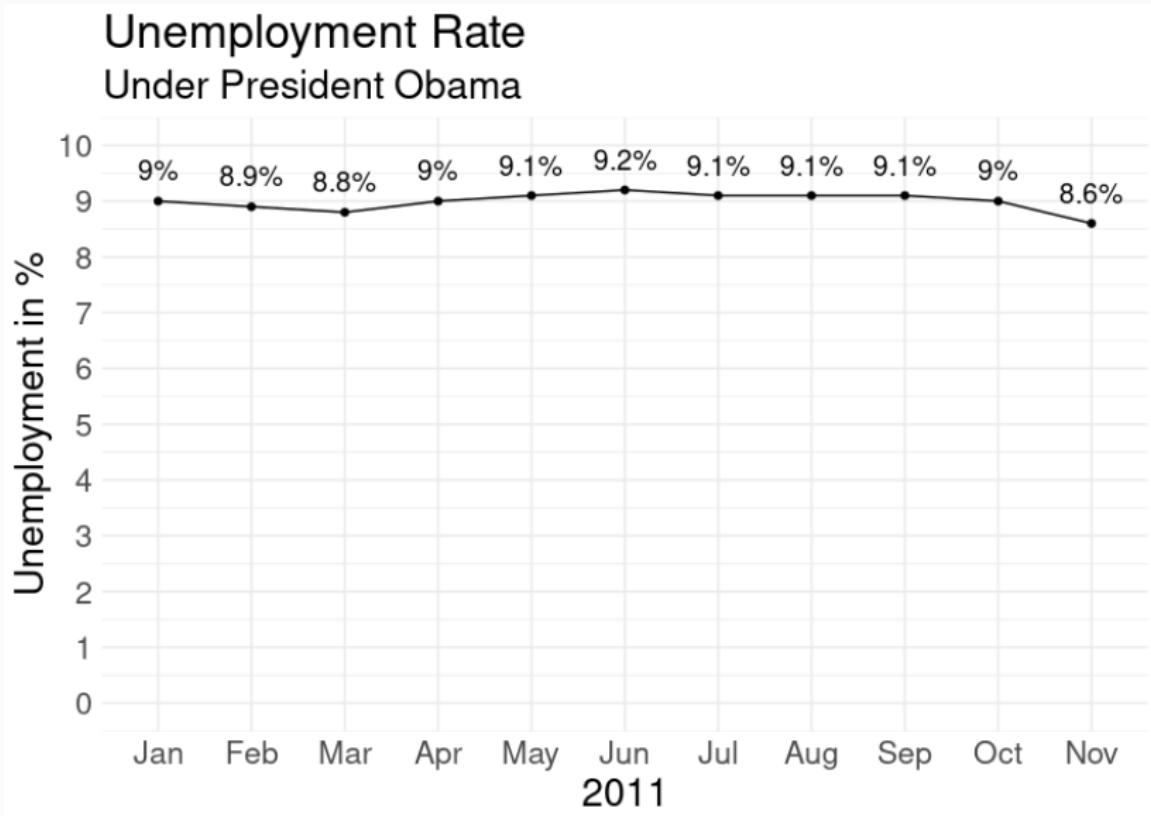
If Bush Tax Rates Expire



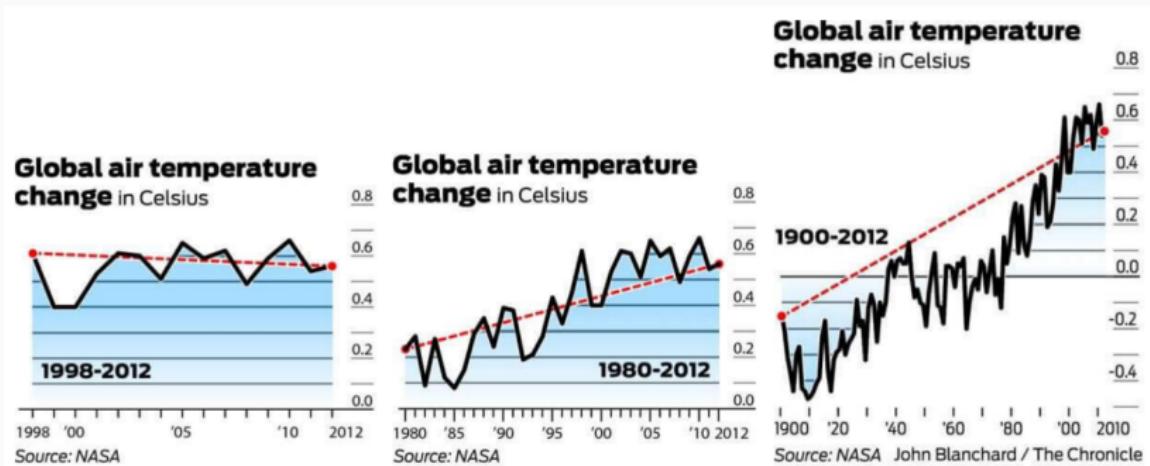
X/Y plots: Manipulate the Y-axis



X/Y plots: Manipulate the Y-axis



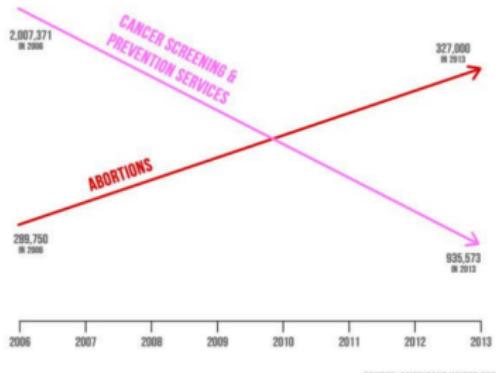
X/Y plots: Pick the range



Fallenbüchel (2019)

X/Y plots: Fuck the y-axis entirely

PLANNED PARENTHOOD FEDERATION OF AMERICA: ABORTIONS UP – LIFE-SAVING PROCEDURES DOWN



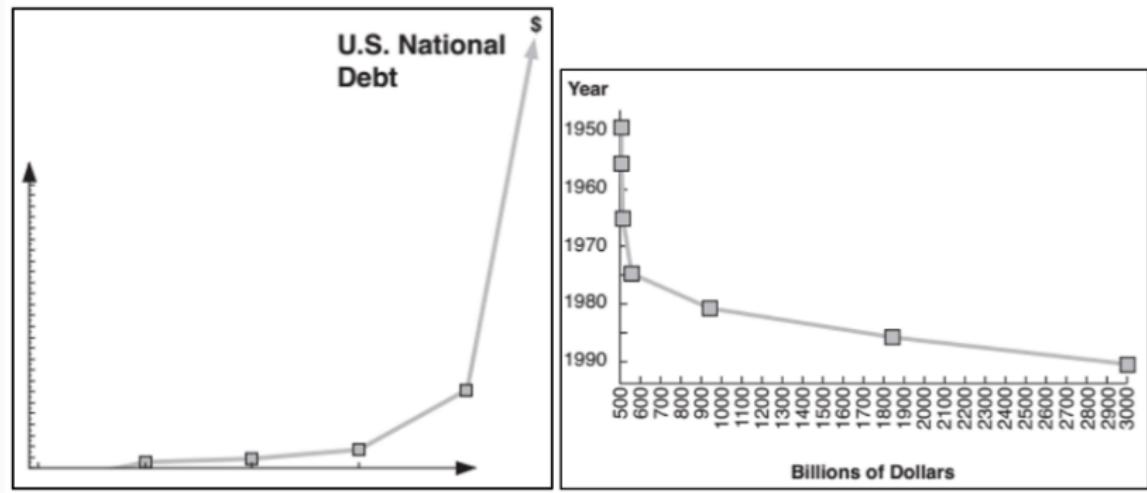
Planned Parenthood services

- Abortion procedures
- Cancer screening / preventative services



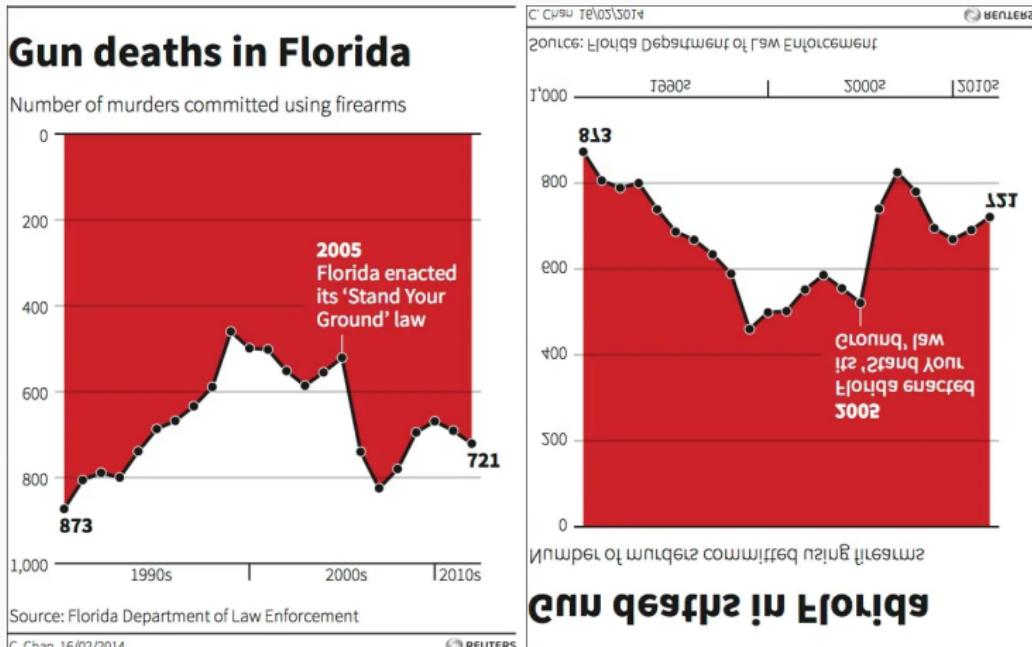
Fallenbüchel (2019)

X/Y plots: Flip plot



Fallenbüchel (2019)

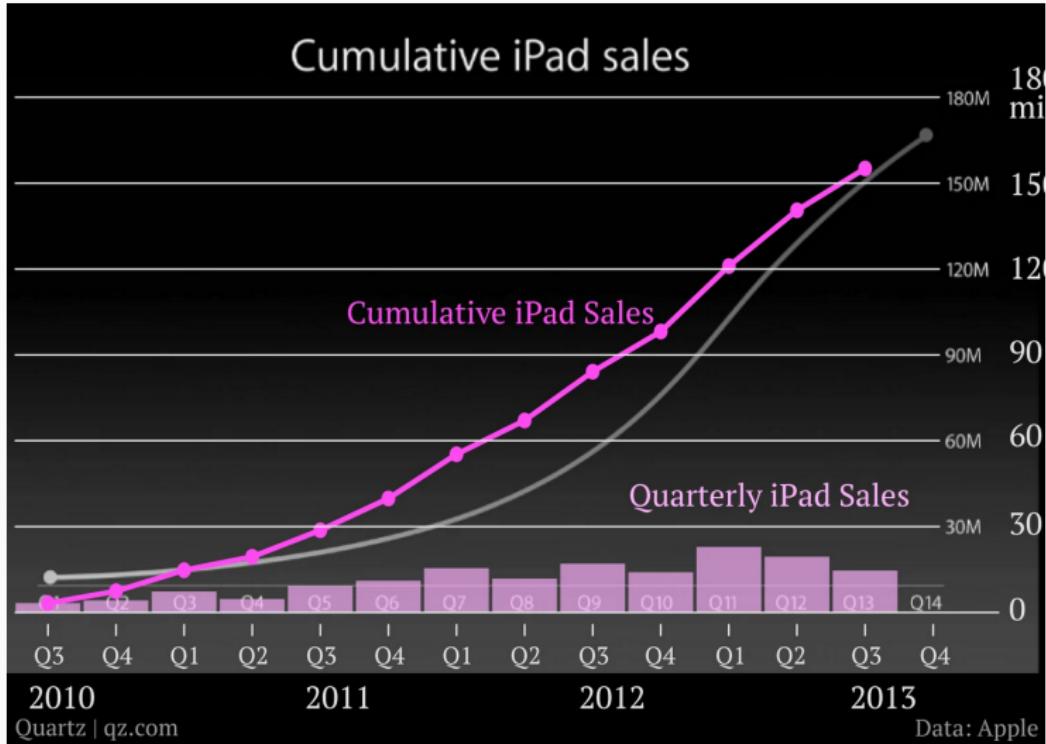
X/Y plots: Flip plot



X/Y plots: Cumulative growth



X/Y plots: Cumulative growth



Yanofsky (2013)

X/Y plots: Cherrypick data

The image shows a poll graphic from NBC News/Wall Street Journal. The title is "PRESIDENT TRUMP'S JOB APPROVAL AMONG REPUBLICANS". It features a portrait of Donald Trump on the left. The main results are displayed in a red-bordered box: "APPROVE" at 88% and "DISAPPROVE" at 9%. Below the main chart, there is a smaller text box that reads: "NBC NEWS/WALL STREET JOURNAL POLL JULY 15-18 MOE +/- 3.27-9.15". At the bottom of the graphic, it says "DEVELOPING".

realdonaldtrump • Follow

realdonaldtrump Thank you very much, working hard!

Load more comments

riot_racer @figboot31977 I know your a Donald trump supporter. Kind of obvious

figboot31977 @riot_racer Here's a thought for ya.... You don't know crap!!! GET a LIFE and STAY the #@## OUT OF MINE!!!

riot_racer @figboot31977 I'm not in yours. If I was then I'd have a bigger understanding of who you are. Like I said if you support

Heart icon Comment icon Share icon Bookmark icon

625,072 likes

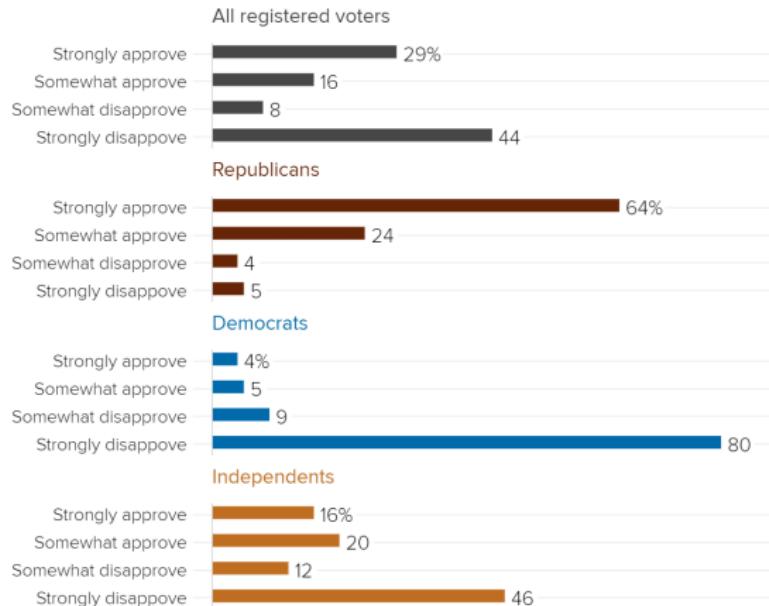
JULY 26

Log in to like or comment.

...

X/Y plots: Cherrypick data

Strength of Trump approval/disapproval by party



NBC NEWS

Data: NBC News/Wall Street Journal poll. July 15-18, 2018.

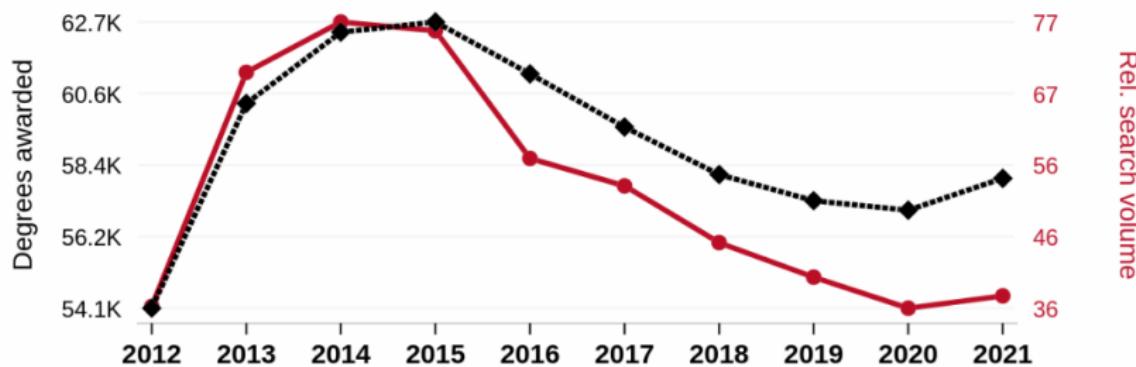
Murray (2018)

X/Y plots: Correlate axes

Bachelor's degrees awarded in law enforcement

correlates with

Google searches for 'sleepwalking'



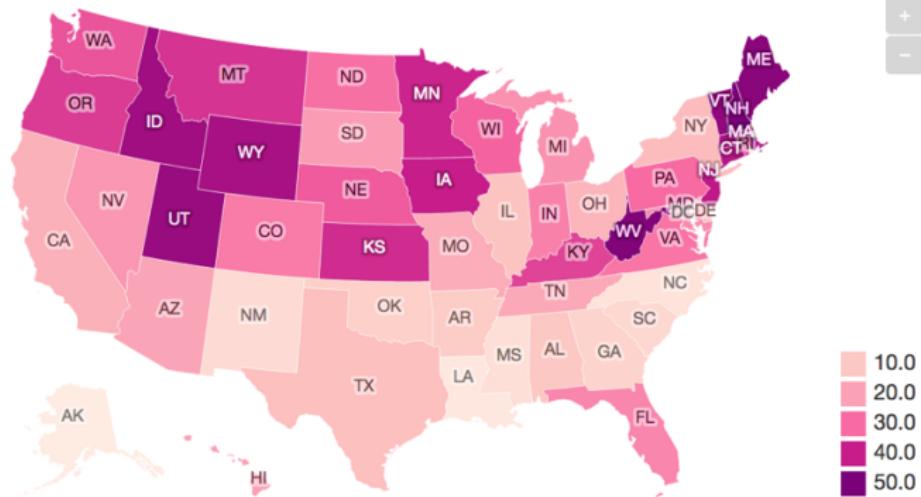
◆··· Bachelor's degrees conferred by postsecondary institutions, in field of study: Homeland security, law enforcement, and firefighting · Source: National Center for Education Statistics

●— Relative volume of Google searches for 'sleepwalking' (Worldwide, without quotes) · Source: Google Trends

2012-2021, $r=0.903$, $r^2=0.815$, $p<0.01$ · tylervigen.com/spurious/correlation/1532

Colors: go against convention

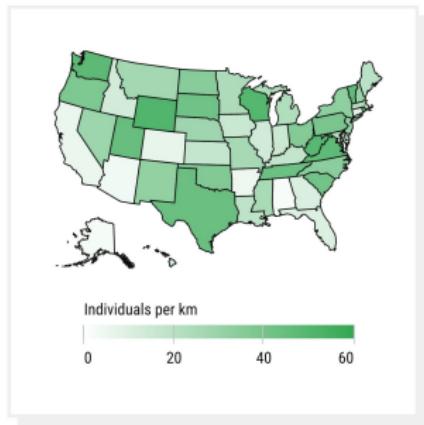
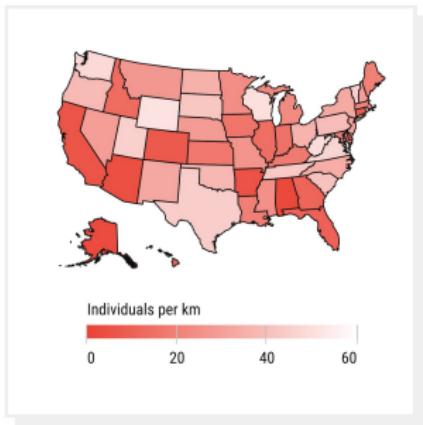
Which states have the most STIs?



[Get the data](#)

McCready (2020)

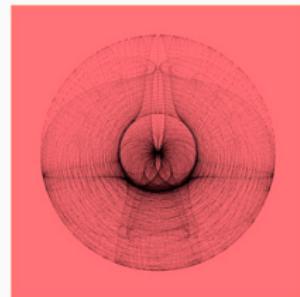
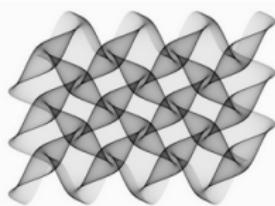
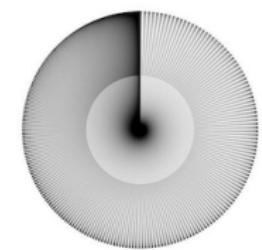
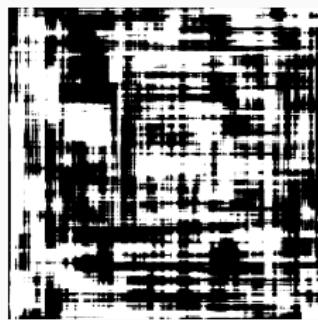
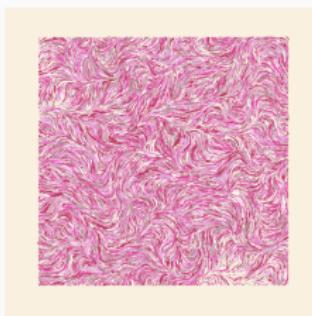
Colors: go against convention

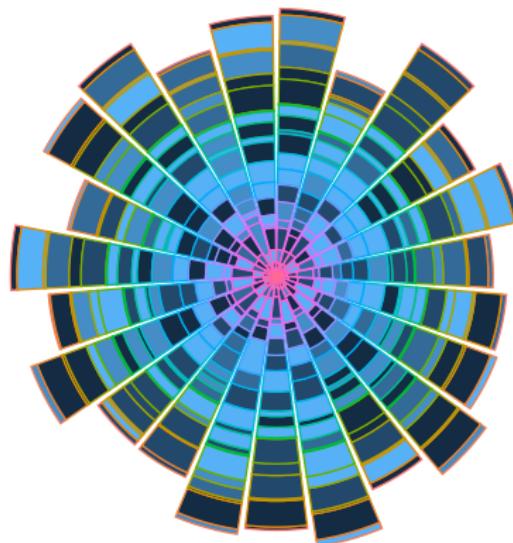


McCready (2020)

Generative art

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



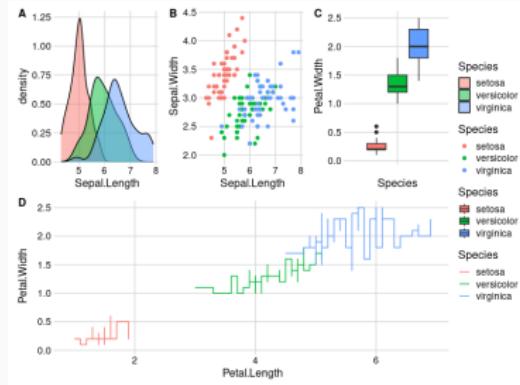


Moving on from R

Arranging plots

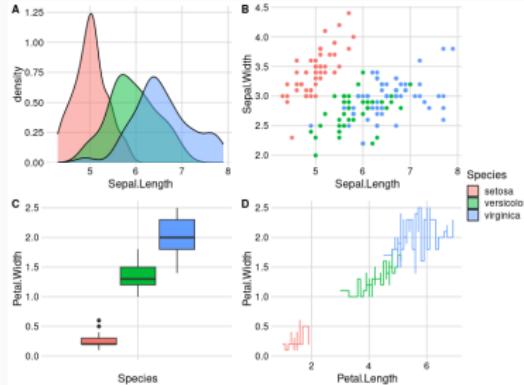
patchwork

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.



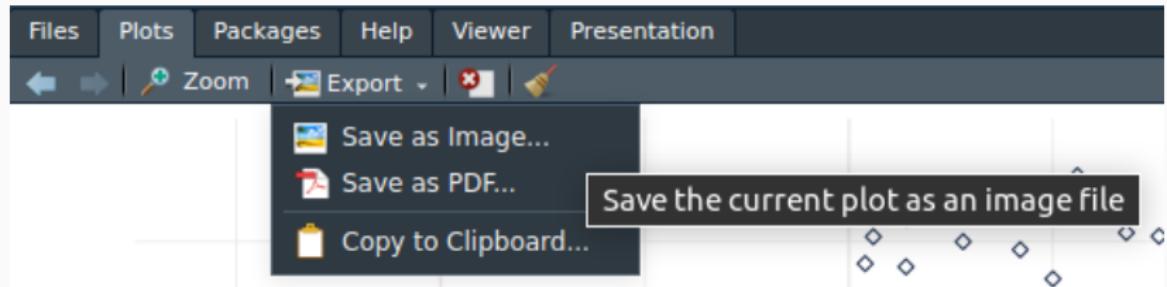
Exporting plots

```
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

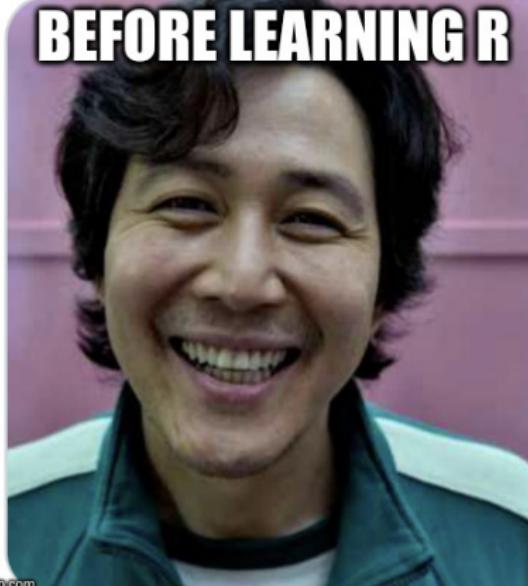
- ✓ R programming basics and RStudio IDE
- ✓ write scripts
- ✓ file encoding, variable naming, and tidy code with pipes
- ✓ install and load packages
- ✓ import/export data from/to the working directory
- ✓ save and remove objects in the environment
- ✓ preprocess raw data (filtering, renaming, arranging, 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

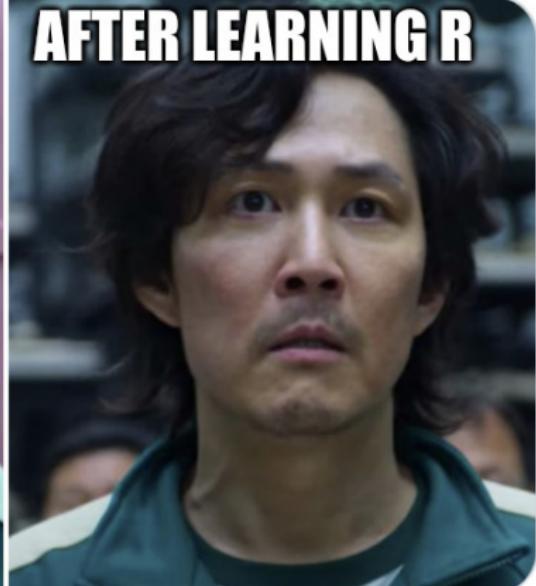
- ② Complete assignment 6 (\rightarrow ILIAS)
- ② Vote for the ugliest plot.
- ② Install Quarto: <https://quarto.org/docs/get-started/>
- ② Watch the Quarto introductory video: https://youtu.be/_f3latm0hew?si=xxovQvYkUosC_4uB

Next time: Reporting on data

BEFORE LEARNING R



AFTER LEARNING R



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/msnbc-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).