

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.

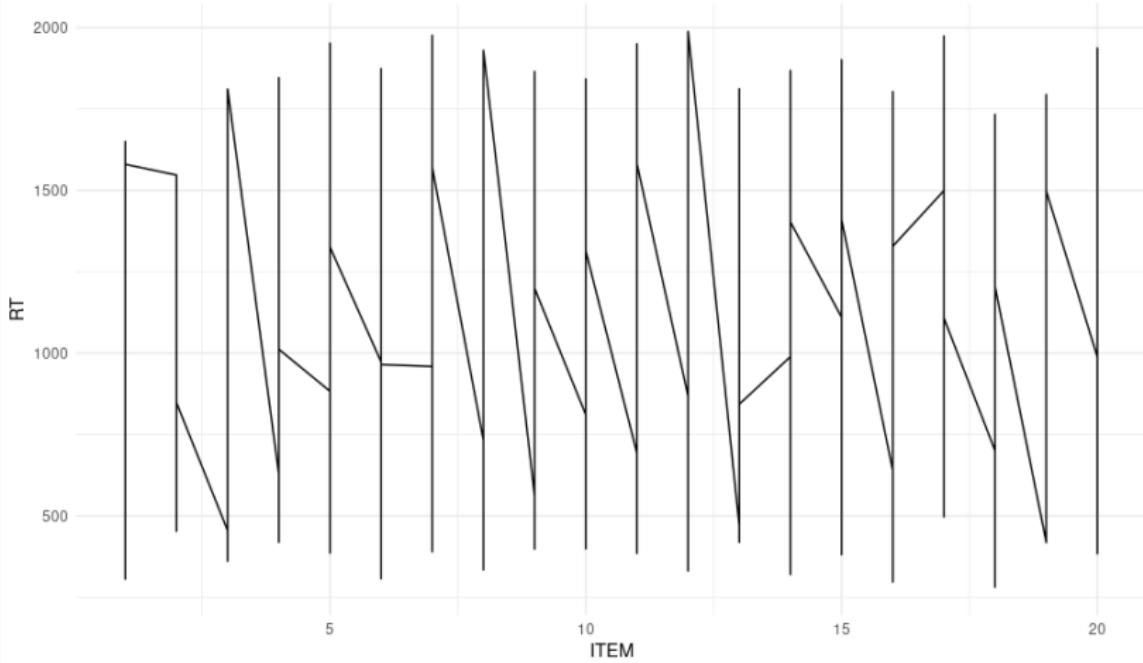
Recommended watching:

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

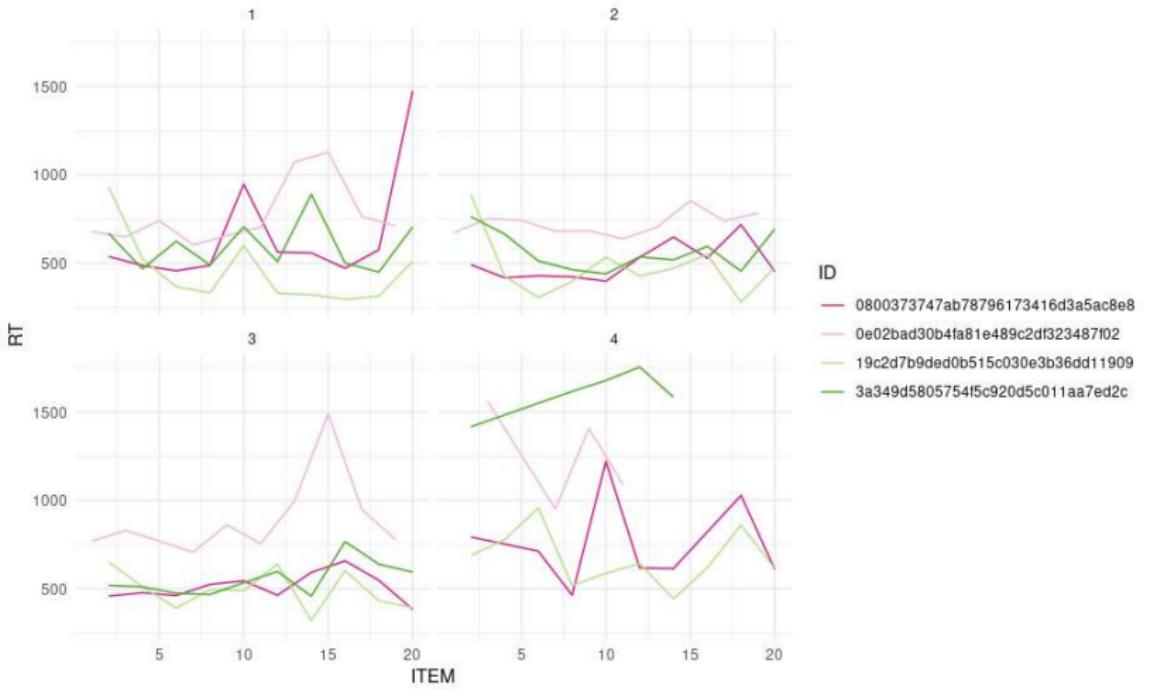
Ø `ggplot(noisy_rt.csv)`

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.

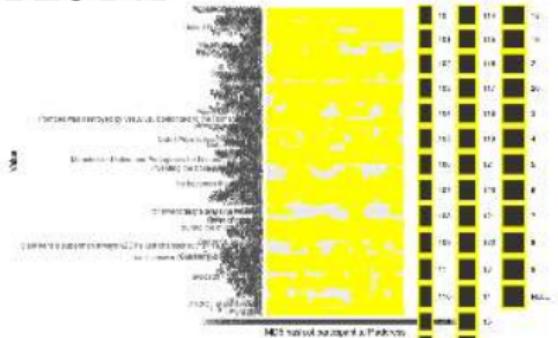
`stat = "summary", fun = "mean"`

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

`bin, smooth, density, ...`

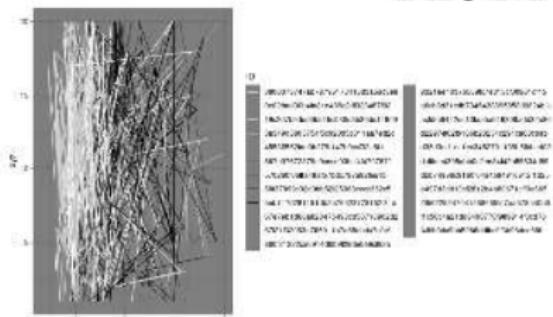
Ugly plots

PLOT A



PLOT C

PLOT B



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

1. Where are we this week?

2. Data visualization

3. Colors

4. Color use guidelines

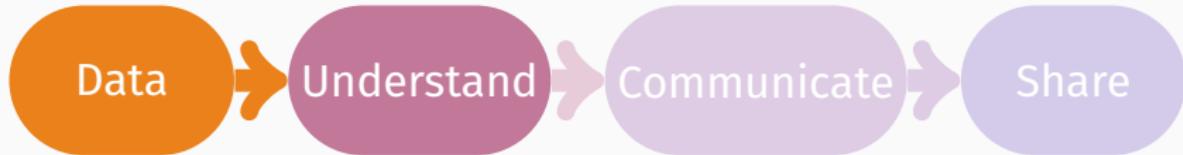
5. Lying with plots

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

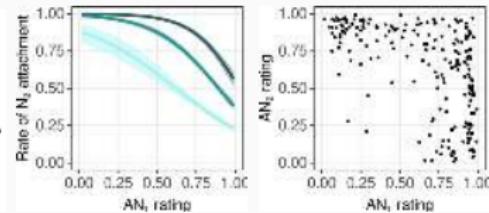
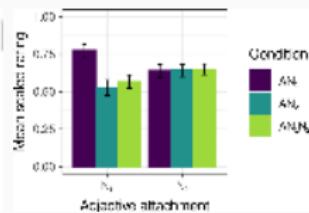
- 🗣 Communication & understanding
- 📊 Analysis & exploration
- ⚖️ Decision making



Visualization types

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

Item	adjective	N1	N2	unclear
9	sommerliche Arbeitsatmosphäre	+	+	0
11	stilreiche Lehrmethoden	+	+	0
12	amtliche Bezugsschreibung	+	+	+
16	sozialer Lernraum	+	+	0
19	geöffnetes Bildungsumfeld	+	0	0
20	ökologische Bildungsumgebung	+	0	0
23	evangelischer Bildungszentrale	+	0	0



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

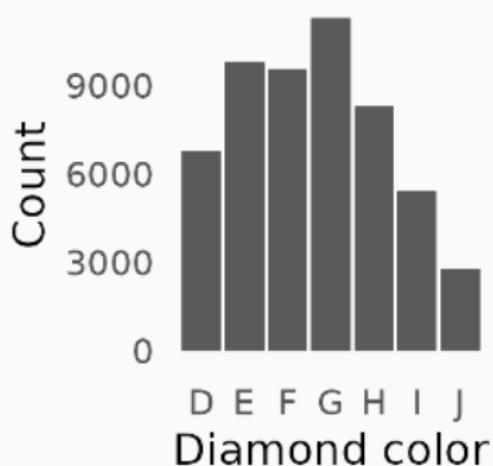
esquisse

GUI for exploring data based on ggplot2



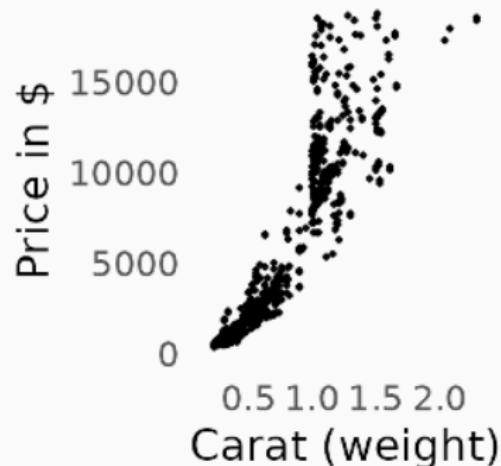
Histogram

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)

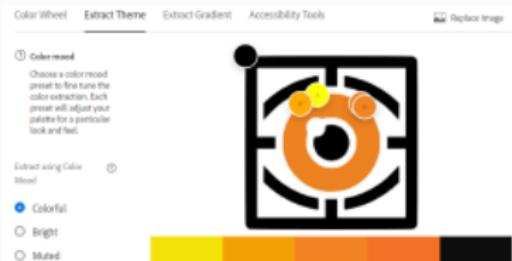
plot + scale_fill_manual(values =
c("#8c510a", "#d8b365", "#f6e8c3", "#f5f5f5"))
```

Adobe color

Assemble colors



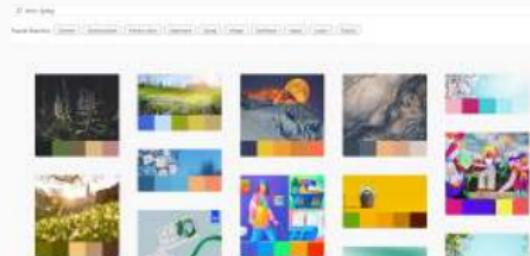
Extract from image



Check contrast



Browse trends



Coolors

Assemble colors and check for color blindness



Visualize the palette



Browse trends

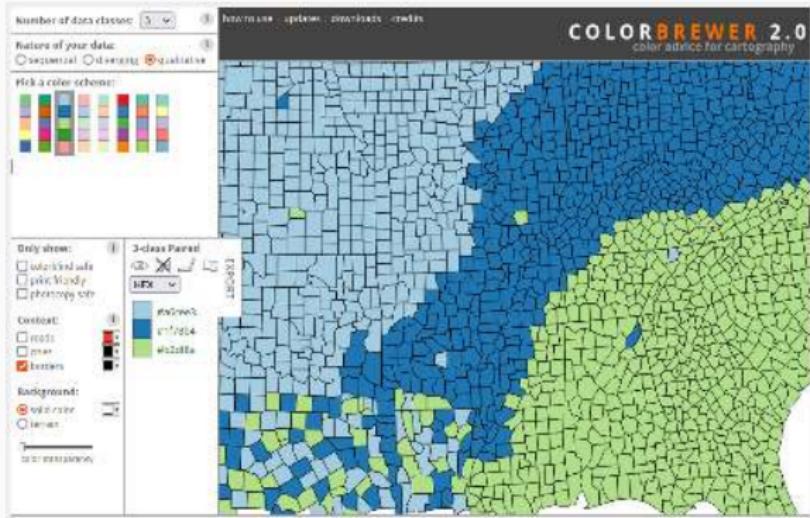
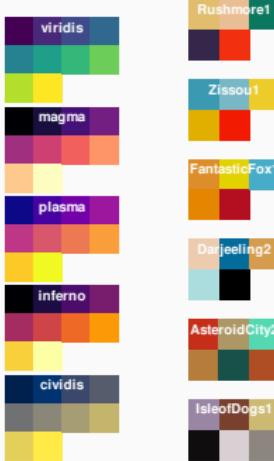


Generate color palettes

The screenshot shows the 'PALETTE GENERATOR' interface. At the top, there are tabs for 'PALETTE', 'SWATCH FAB', and 'DEMOVIEW'. Below the tabs, there's a 'NUMBER OF COLORS' slider set to 7, and a 'BACKGROUND COLOR' switch set to 'DARK'. A row of seven color swatches is displayed, each with its corresponding hex code: #003399, #374ac8, #7a5195, #bc5090, #e35373, #ff7043, and #ffad00. Below the swatches is a 'COPY HEX VALUES' button and an 'EXPORT AS SWC' button. Under the heading 'IN CONTEXT', there's a pie chart and a small map of Washington state where different regions are colored according to the palette.

Check contrast

The screenshot shows the 'Check contrast' interface. At the top, there's a search bar containing the text 'Show me the closest variations of #4ac4e2 that contrast against the color #f0f0f0 enough to meet AA Guidelines'. Below the search bar, there's a 'RESULTS' section. It contains two rows. The first row is for 'FOR LARGE/BOLD TEXT' and shows a blue square with the hex code #06a2bf and a white square with the hex code #f0f0f0. The second row is for 'FOR SMALL TEXT' and shows a dark blue square with the hex code #00819d and a white square with the hex code #f0f0f0. Both rows have a 'Try this combo instead.' link below them.



There's a package for that

RColorBrewer

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

viridis

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

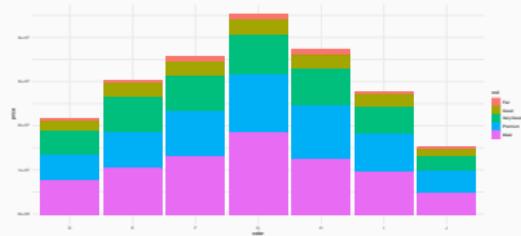
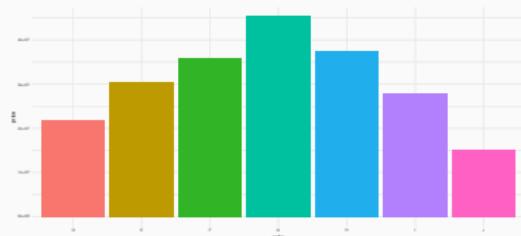
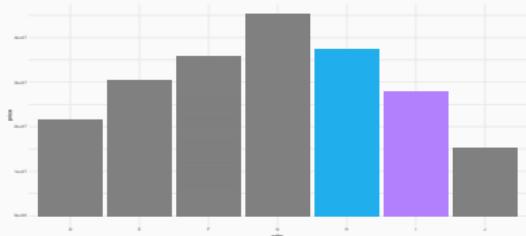
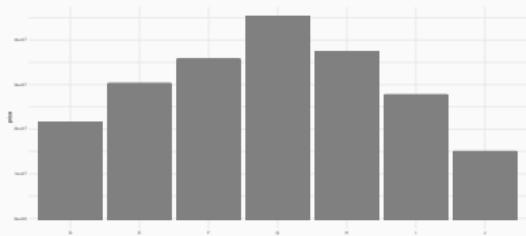
wesanderson

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

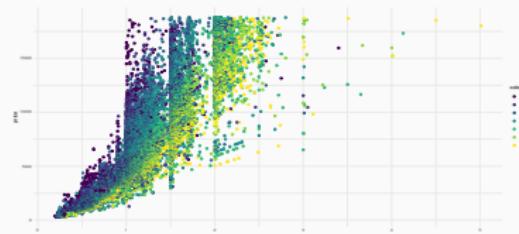
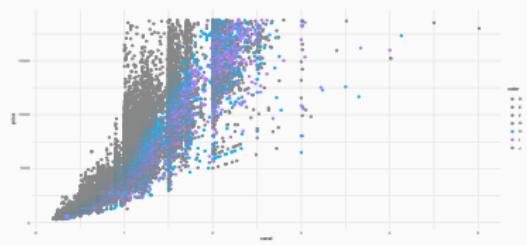
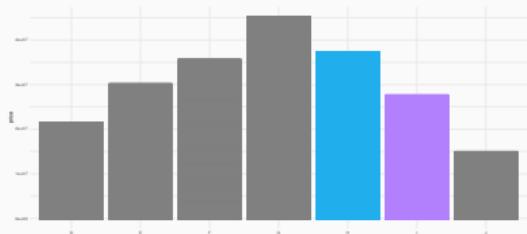
Others, e.g. ggsci (for journals)

Color use guidelines

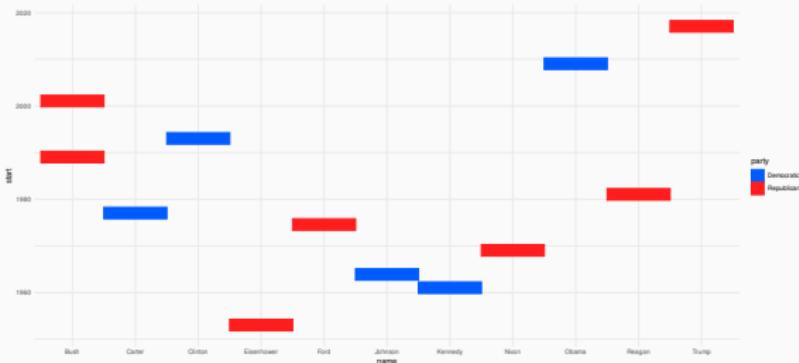
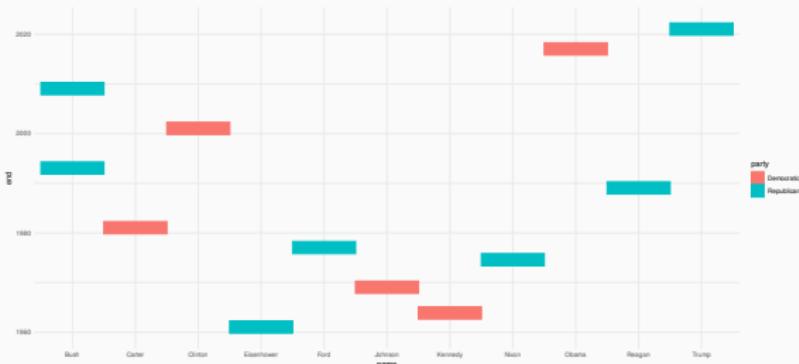
Avoid unnecessary usage of color



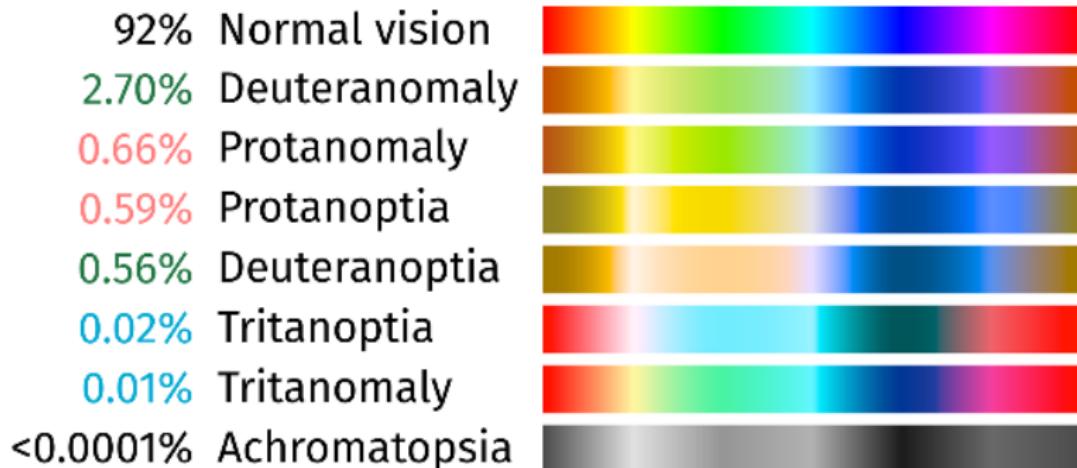
Be consistent



Use in a meaningful way

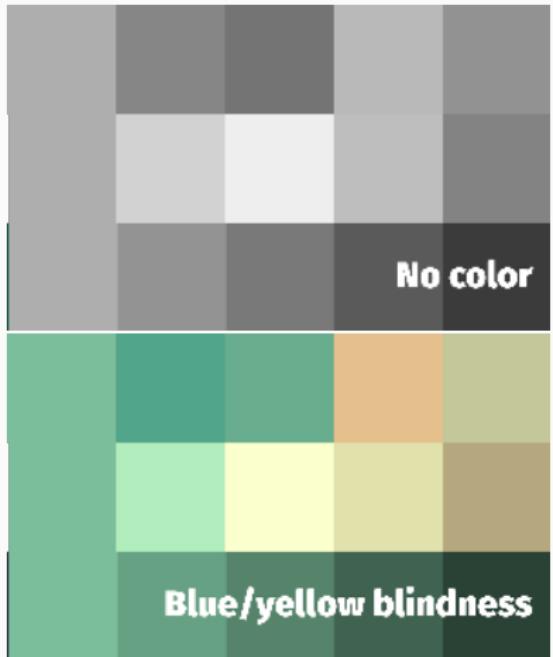


Pay attention to color blindness



1–2 people in this class

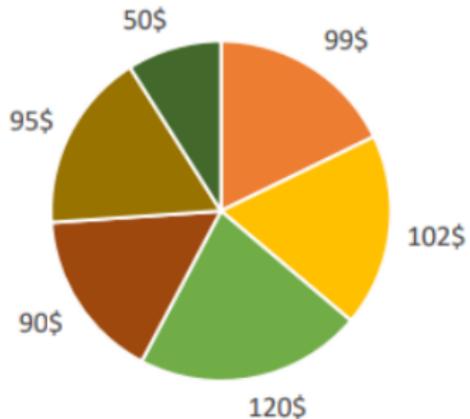
Perceivable colors



Lying with plots

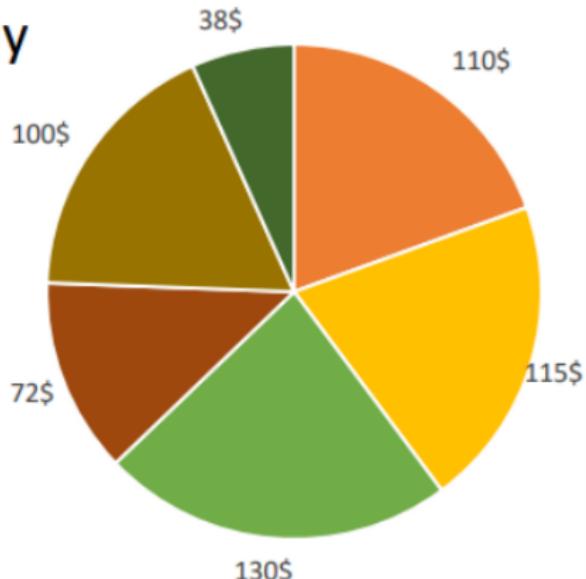
Pie plots: Manipulate the size

Total Sales by Company (Millions)



Last year

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



This year

Fallenbüchel (2019)

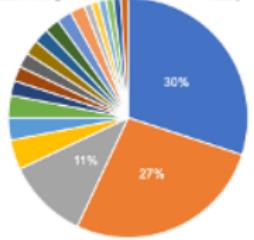
Pie plots: Manipulate categories

Too many slices

Declared majors

% OF ANDERSON HALL RESIDENTS

- Undecided
- Education
- Accounting
- Management
- Mathematics
- Creative writing
- Biology
- Political science
- Psychology
- Biochemistry
- Forensics
- History

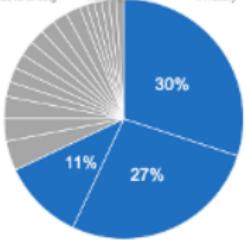


De-emphasize

Declared majors

% OF ANDERSON HALL RESIDENTS

- Undecided
- Education
- Accounting
- Management
- Mathematics
- Creative writing
- Biology
- Political science
- Psychology
- Biochemistry
- Forensics
- History

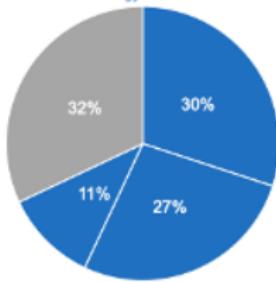


Aggregate

Declared majors

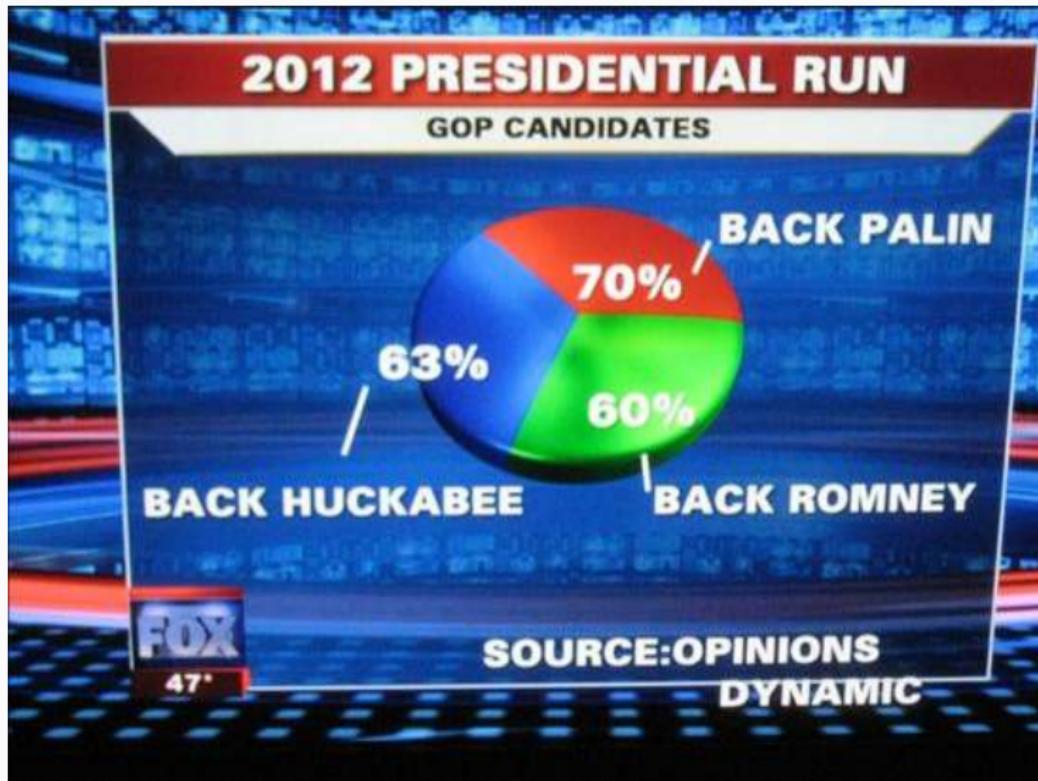
% OF ANDERSON HALL RESIDENTS

- Undecided
- Biology
- Finance
- All other (17)



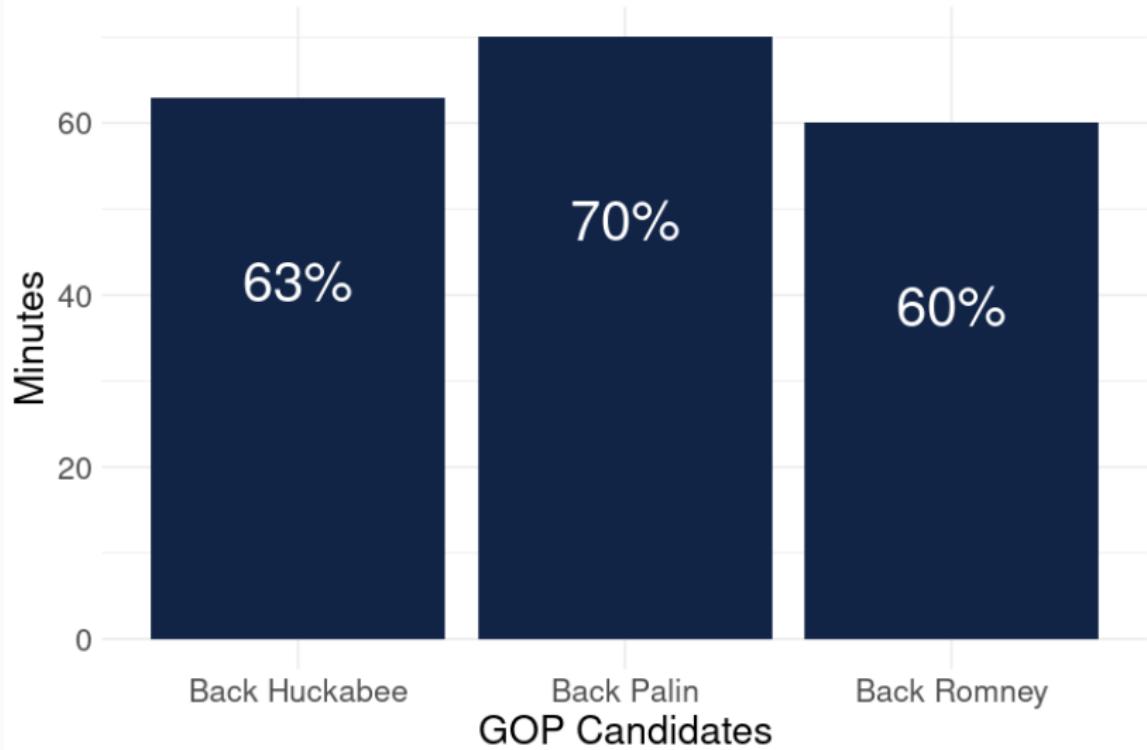
Ricks (2020)

Pie plots: Wrong plot type



Pie plots: Wrong plot type

2012 Presidential Run



Bar plots: Stack (+ color)

RUSSIA ISSUES VS. NON-RUSSIA ISSUES

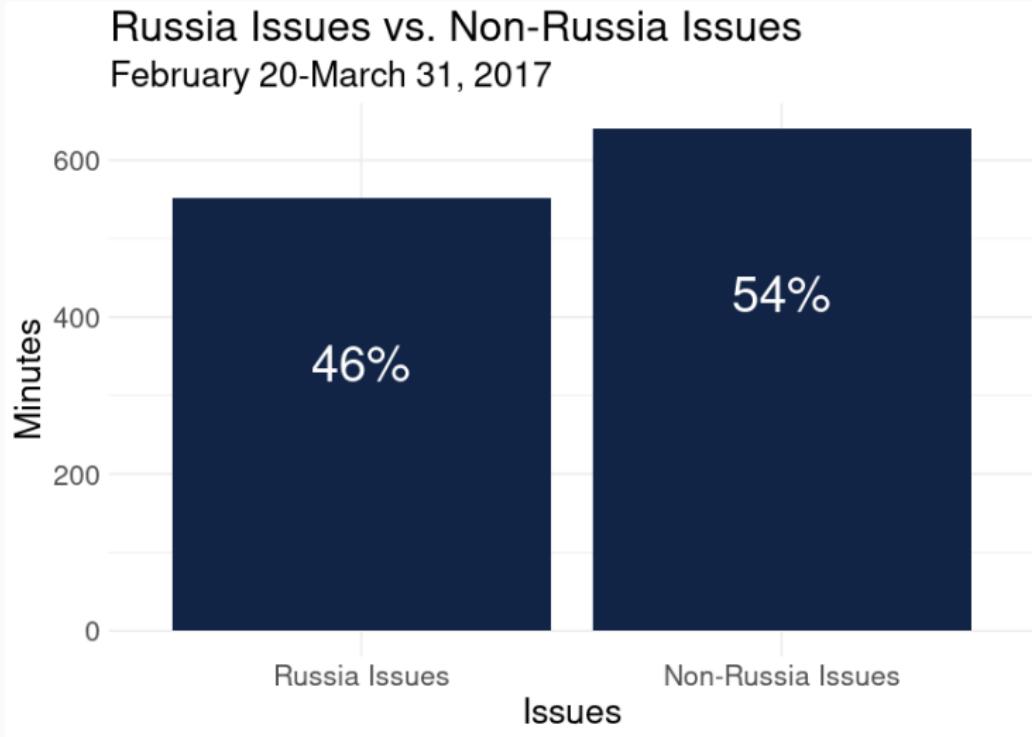
February 20–March 31, 2017

1191:58 (min:sec) Total Show Minutes

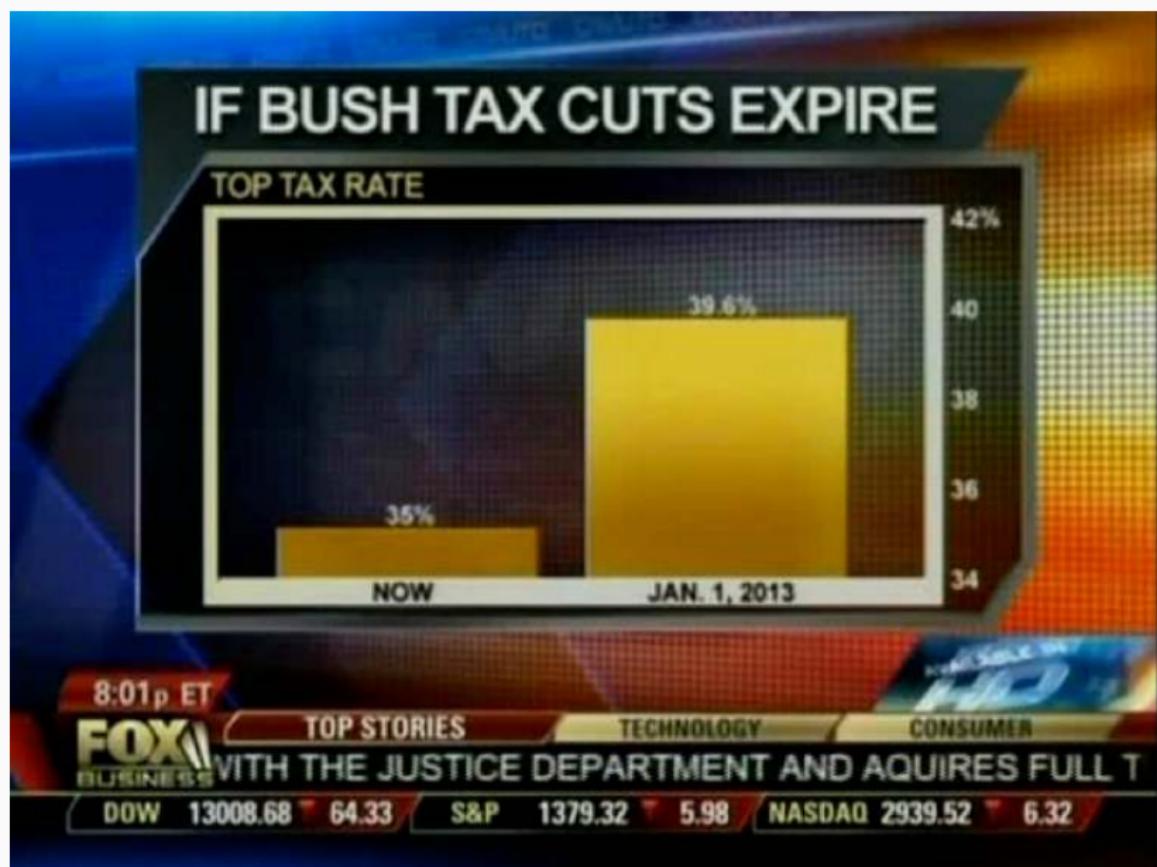


Maté (2017)

Bar plots: Stack (+ color)

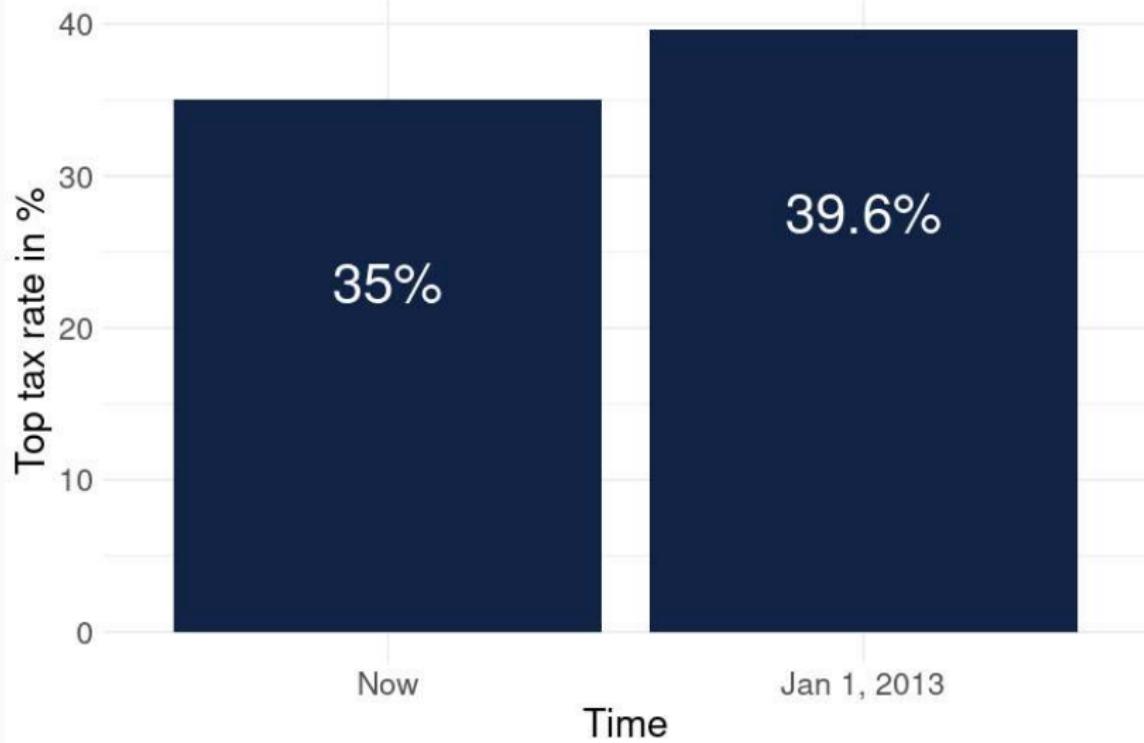


X/Y plots: Omit the baseline

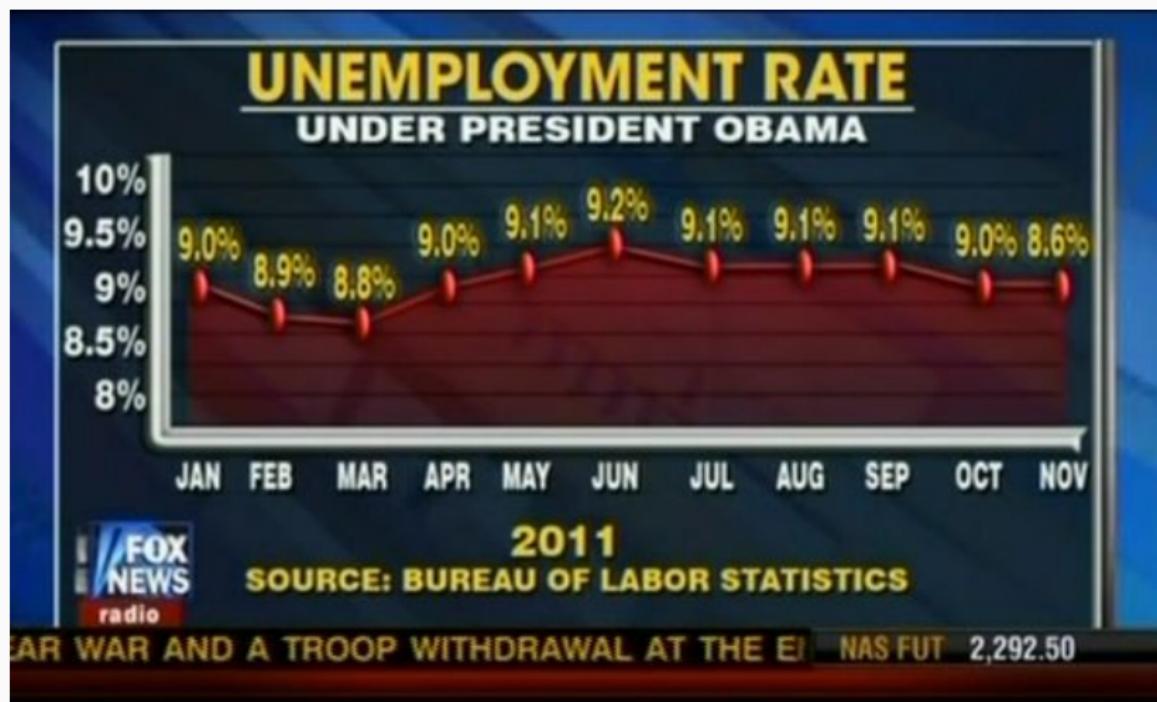


X/Y plots: Omit the baseline

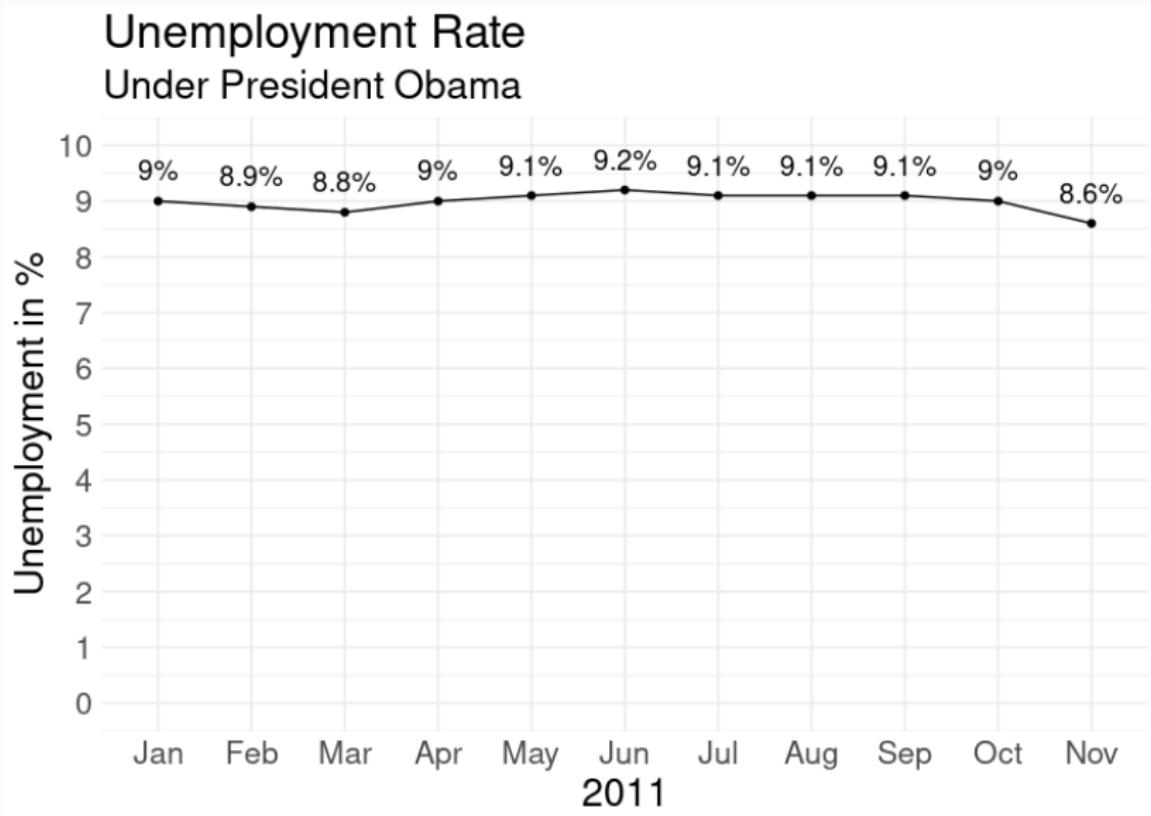
If Bush Tax Rates Expire



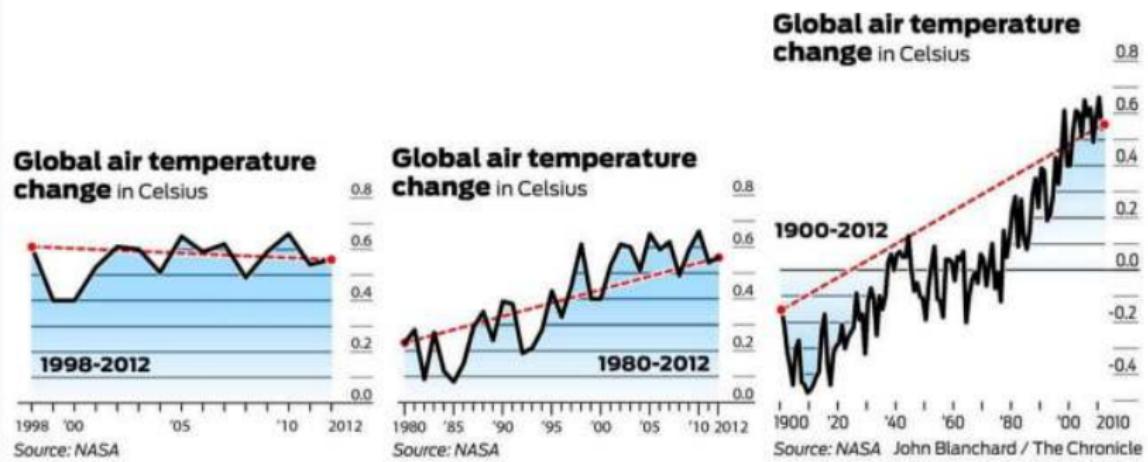
X/Y plots: Manipulate the Y-axis



X/Y plots: Manipulate the Y-axis



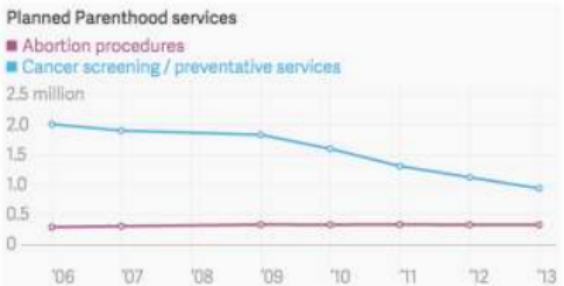
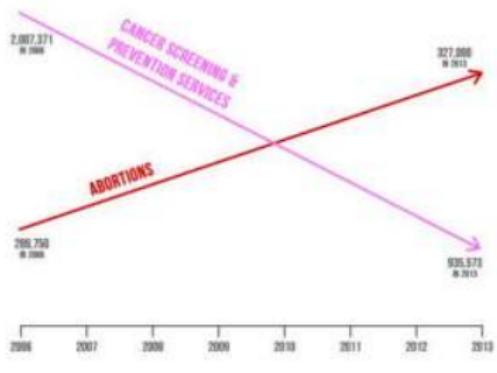
X/Y plots: Pick the range



Fallenbüchel (2019)

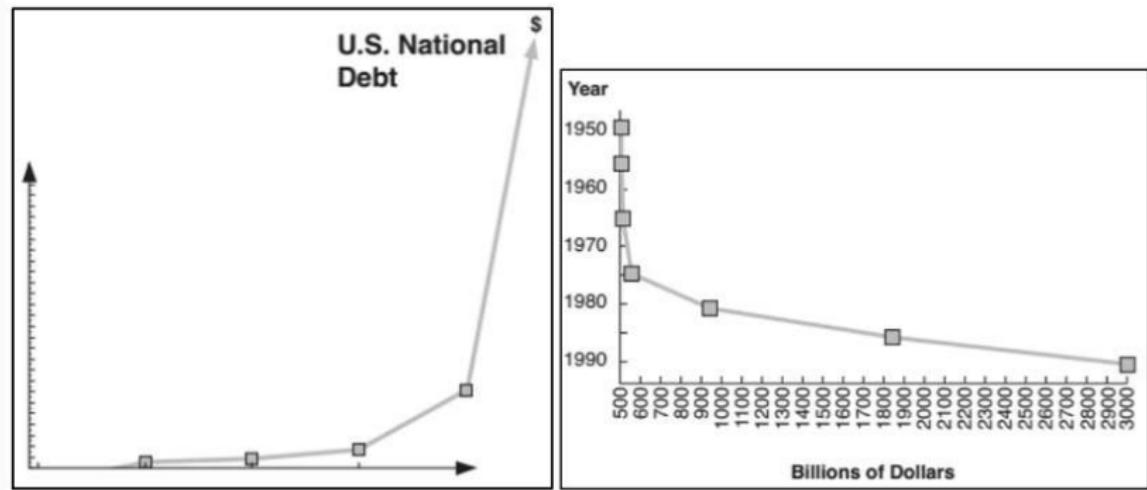
X/Y plots: Fuck the y-axis entirely

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



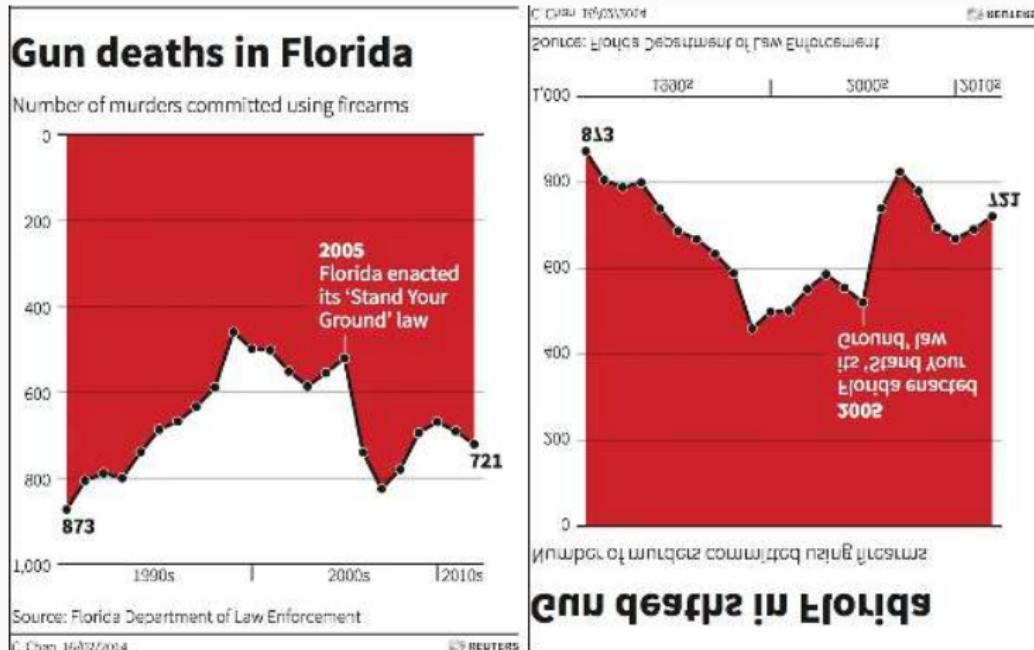
Fallenbüchel (2019)

X/Y plots: Flip plot



Fallenbüchel (2019)

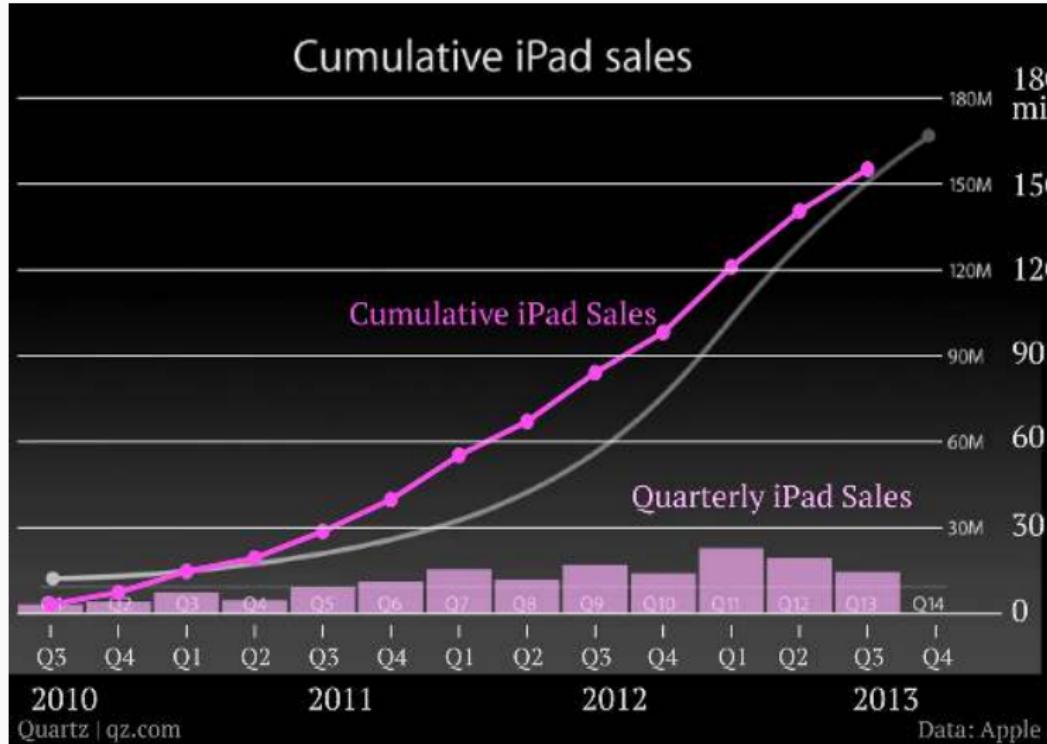
X/Y plots: Flip plot



X/Y plots: Cumulative growth



X/Y plots: Cumulative growth



Yanofsky (2013)

X/Y plots: Cherrypick data

PRESIDENT TRUMP'S JOB APPROVAL
AMONG REPUBLICANS

APPROVE	88%
DISAPPROVE	9%

NBC NEWS/WALL STREET JOURNAL POLL
JULY 18, 2018
MORI 3.2/PTS

DEVELOPING

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

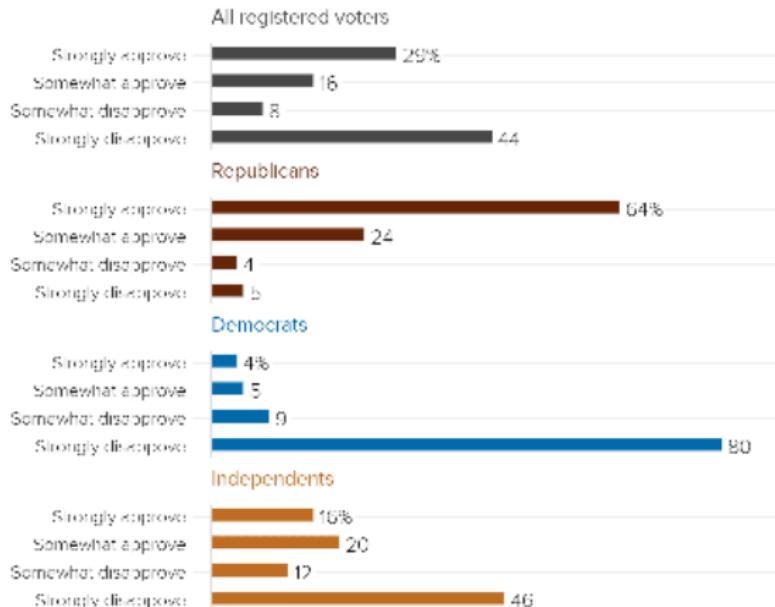
625,072 likes

JULY 26

Log in to like or comment.

X/Y plots: Cherrypick data

Strength of Trump approval/disapproval by party



NBC NEWS

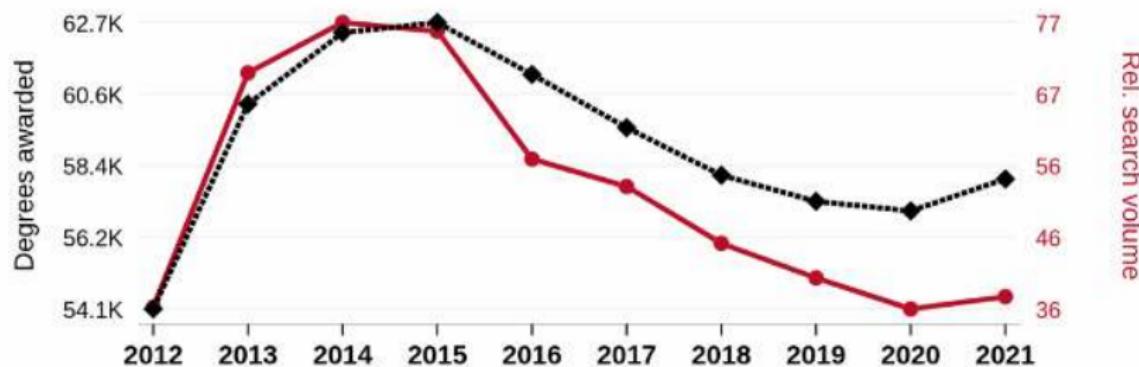
Data: NBC News/Well Street Journal poll, July 15-16, 2016.

X/Y plots: Correlate axes

Bachelor's degrees awarded in law enforcement

correlates with

Google searches for 'sleepwalking'



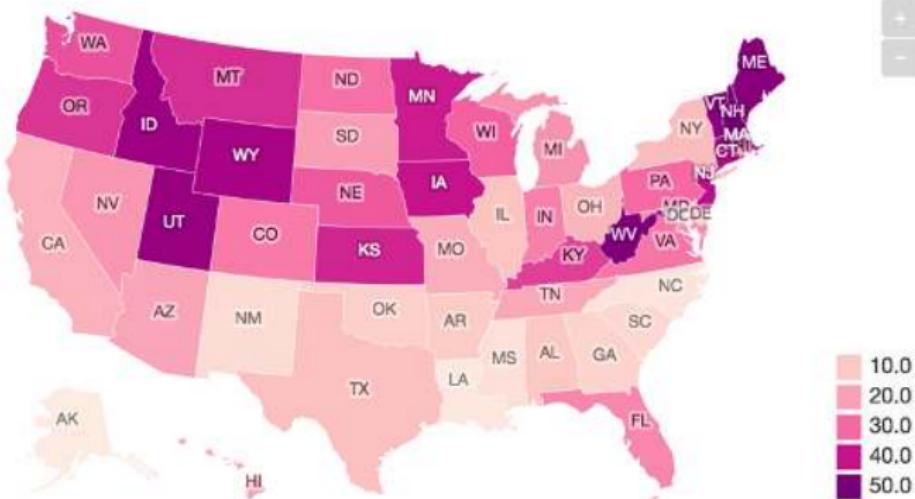
◆ Bachelor's degrees conferred by postsecondary institutions, in field of study: Homeland security, law enforcement, and firefighting · Source: National Center for Education Statistics

● 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



Individuals per km



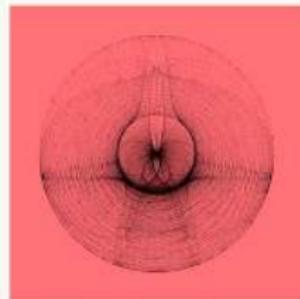
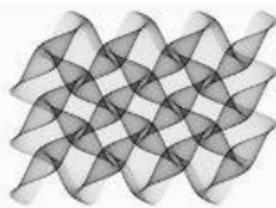
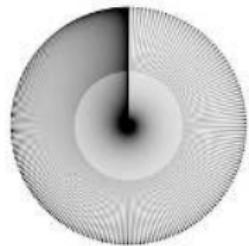
Individuals per km

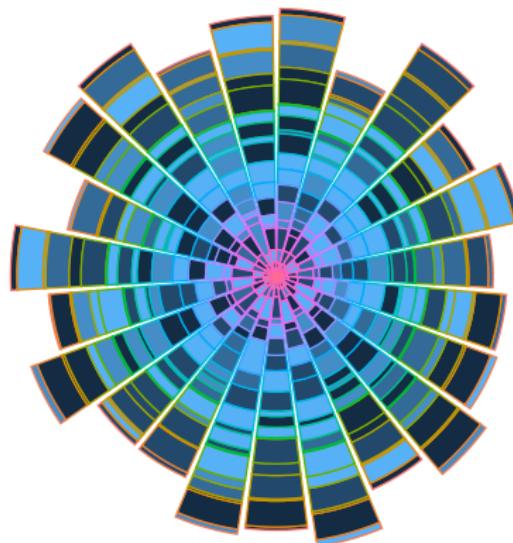


McCready (2020)

Generative art

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



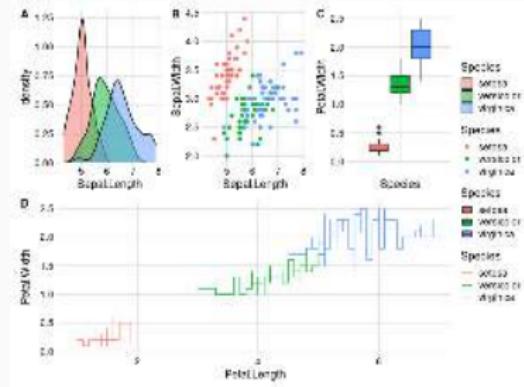


Moving on from R

Arranging plots

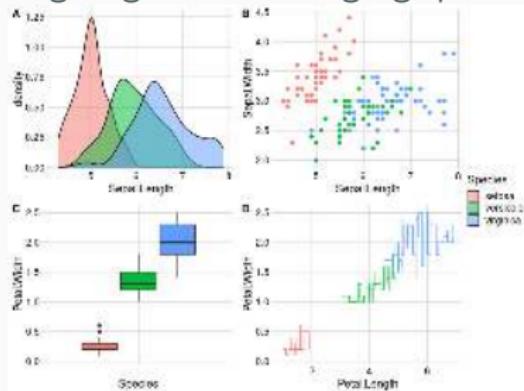
patchwork

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.



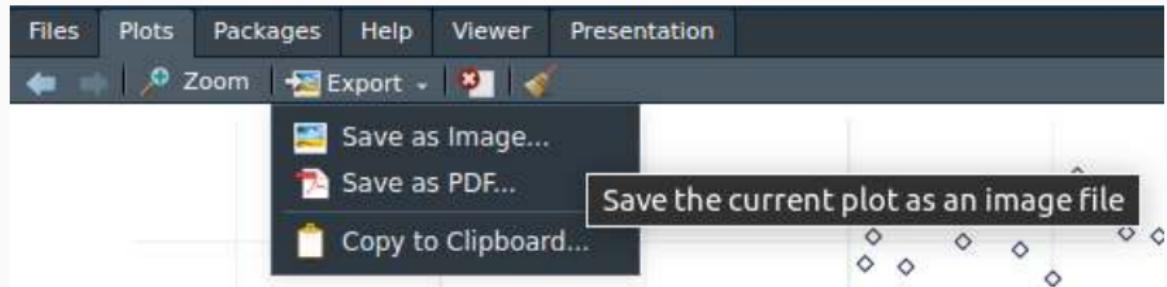
Exporting plots

```
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

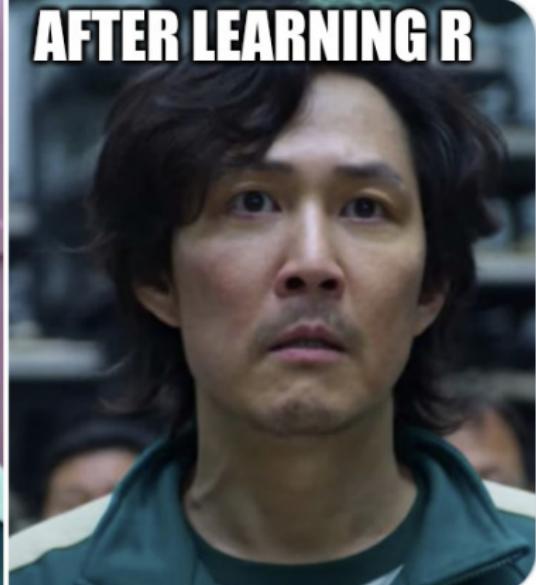
- ② Complete assignment 6 (\rightarrow ILIAS)
- ② Vote for the ugliest plot.
- ② Install Quarto: <https://quarto.org/docs/get-started/>
- ② Watch the Quarto introductory video: https://youtu.be/_f3latm0hew?si=xxovQvYkUosC_4uB

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).