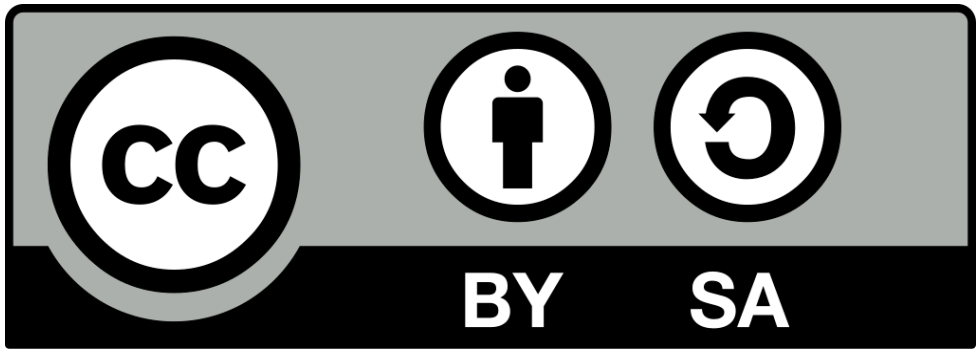
Data Visualization with RStudio and Jupyter

Infra4NextGen Webinar

Franz Eder
franz.eder@uibk.ac.at
Department of Political Science | University of Innsbruck

Housekeeping rules

- Please mute your microphone when you are not speaking
- Please turn off your camera during presentations – the session is being recorded
- Post your questions in the chat or raise your hand using the “raise your hand” function in Zoom (“Reactions” button)
- Tell us how we did in the survey (more about that later)
- The presentation will be shared afterwards



About me

Dr. Franz Eder

Assoc. Prof. for International Relations

University of Innsbruck

Research focus: Foreign and Security Policy; (Counter-)Terrorism; USA, Europe, Austria; social science research methods (v.a. QTA, DNA); academic writing and presentation; open and reproducible science

www.franz-eder.info | franz.eder@uibk.ac.at

FranzEder | franzeder.bsky.social

0000-0003-2625-3631 | 89MVg38AAAAJ

Foreign Policy Lab

FoP Lab

HOME MISSION RESEARCH AFP3 SURVEY EDUCATION OUTREACH MEMBERS PARTNERS

HOME

Welcome to the Foreign Policy Lab at the **Department of Political Science** of the **University of Innsbruck**. Our Lab is committed to addressing the growing complexities of foreign policy and world politics. As a hub for innovative ideas it seeks to promote research, education, and public debate.

News

Wer soll dieses Land verteidigen?

Martin Senn spricht mit Der Standard über Neutralität und Solidarität

OCT 26, 2024

Wie sieht die 'multipolare Welt' aus

Martin Senn im Gespräch mit Der Standard

OCT 22, 2024

New FWF Project!

Congratulations to Stefanie Kirchweger and Franz Eder

OCT 8, 2024

AFP3

AF P3

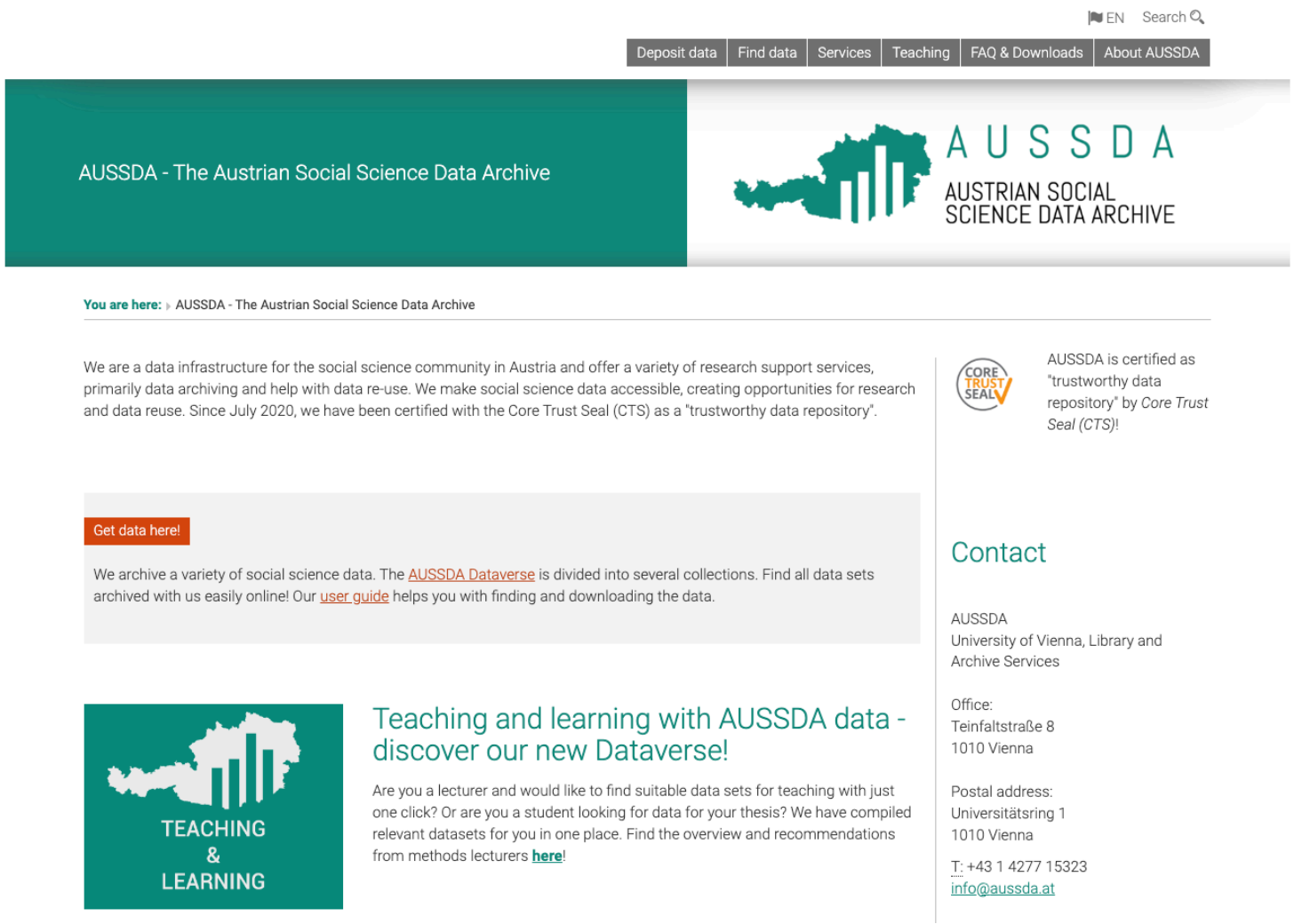
HOME DASHBOARD METHODIK DATENSATZ KONTAKT & PARTNER

HOME

Willkommen auf dem Dashboard des Austrian Foreign Policy Panel Projects (AFP3)

Welche Einstellungen haben Österreicher:innen zur Außen- und Sicherheitspolitik ihres Landes? Was wissen sie über die Vereinten Nationen? Woher beziehen sie ihr außenpolitisches Wissen? Welche Parteien sehen sie als kompetent in der Außenpolitik an? Welchen Stellenwert hat die Neutralität für sie und inwiefern wären sie bereit, einen Beitrag zur militärischen Landesverteidigung zu leisten? All diese Fragen und noch viele mehr werden in unserem **Austrian Foreign Policy Panel Project (AFP3)** untersucht.

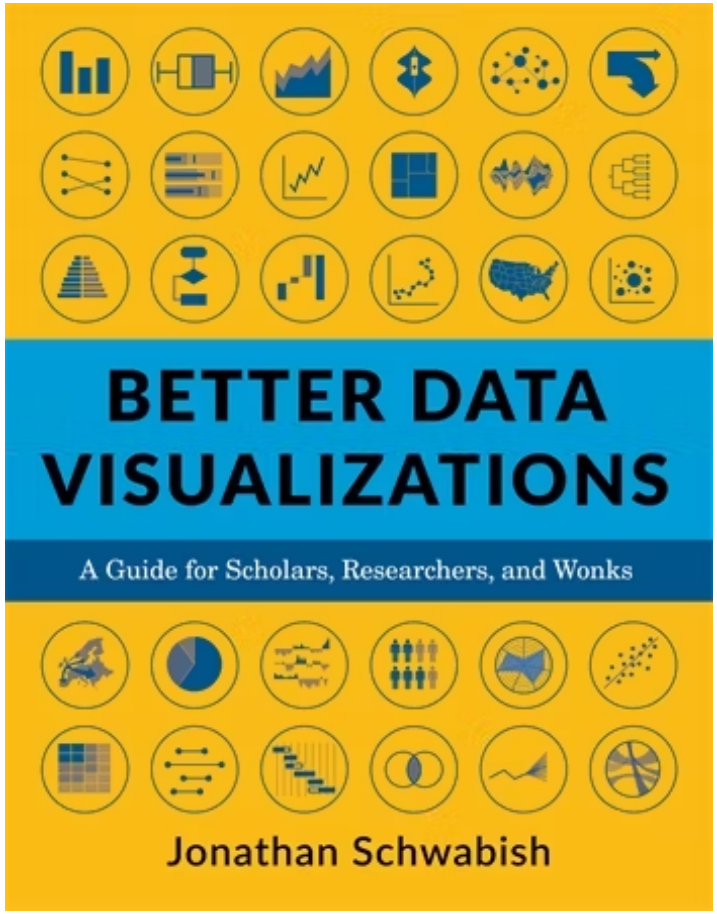
AUSSDA


EN
Search
Deposit data
Find data
Services
Teaching
FAQ & Downloads
About AUSSDA
AUSSDA - The Austrian Social Science Data Archive
AUSSDA
AUSTRIAN SOCIAL SCIENCE DATA ARCHIVE
You are here: AUSSDA - The Austrian Social Science Data Archive
We are a data infrastructure for the social science community in Austria and offer a variety of research support services, primarily data archiving and help with data re-use. We make social science data accessible, creating opportunities for research and data reuse. Since July 2020, we have been certified with the Core Trust Seal (CTS) as a "trustworthy data repository".
AUSSDA is certified as "trustworthy data repository" by Core Trust Seal (CTS)!
Get data here!
We archive a variety of social science data. The AUSSDA Dataverse is divided into several collections. Find all data sets archived with us easily online! Our user guide helps you with finding and downloading the data.
Contact
AUSSDA
University of Vienna, Library and Archive Services
Office:
Teinfaltstraße 8
1010 Vienna
Postal address:
Universitätsring 1
1010 Vienna
T: +43 1 4277 15323
info@aussda.at
TEACHING & LEARNING
Teaching and learning with AUSSDA data - discover our new Dataverse!
Are you a lecturer and would like to find suitable data sets for teaching with just one click? Or are you a student looking for data for your thesis? We have compiled relevant datasets for you in one place. Find the overview and recommendations from methods lecturers here!

Structure

1. What is a “visualization”?
2. Process of visualizing data
3. RStudio/Jupyter and ggplot2
4. Example from European Values Study

Books



Schwabish (2021)

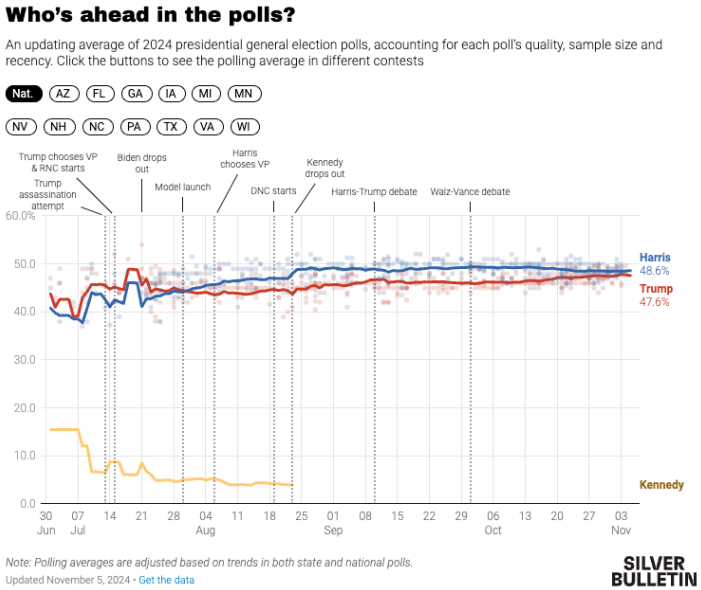
Wickham (2024)

What is “visualization”?

Definitions

Visualization

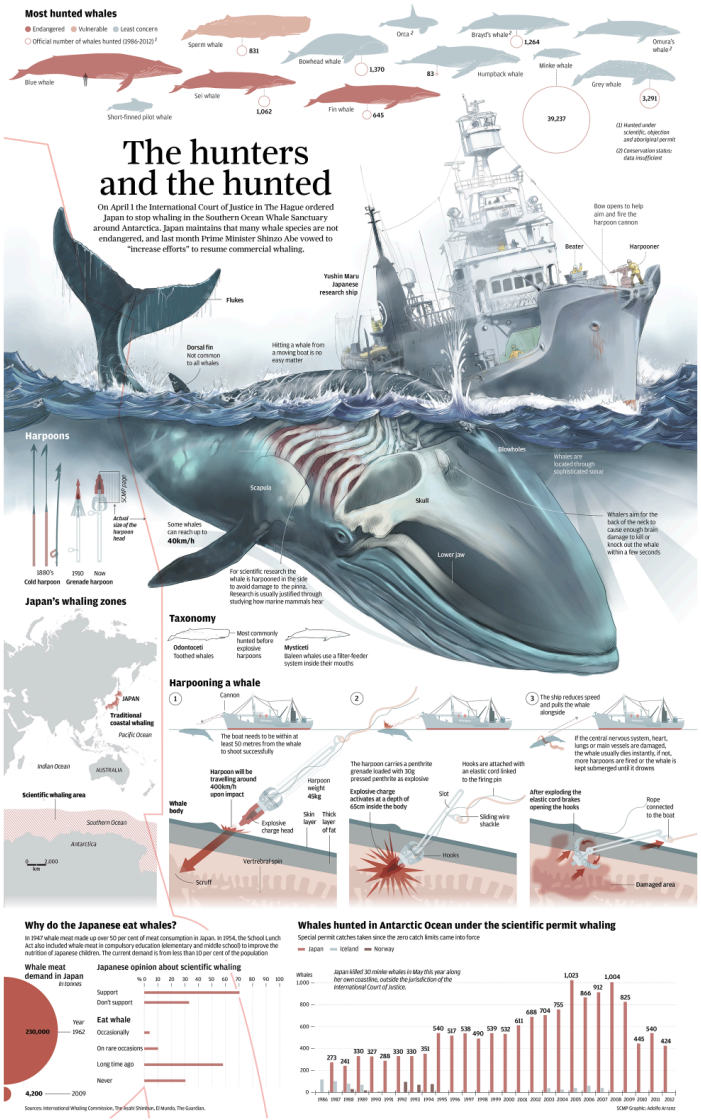
"A visualization is any kind of `visual representation of information` designed to `enable communication`, analysis, discovery, exploration, etc" [emphasis by FE] (Cairo 2016, 28).



Source: Final Silver Bulletin 2024 presidential election forecast

Infographics

"An infographic is a `multi-section visual representation` of information intended to communicate one or more specific messages. Infographics are made of a mix of charts, maps, illustrations, and text (or sound) that provides explanation and context." [emphasis by FE] (Cairo 2016, 31).



Source: South China Morning Post

Purpose of visualizations

> “The purpose of infographics and data visualizations is to enlighten people–not to entertain them, not to sell them products, services, or ideas, but to `inform` them.” [Hervorhebung durch FE] (Cairo 2016, 13).

Communication

“[It, FE] is about drawing and organizing lines and shapes to communicate a specific bit of science-related information to another person… [It, FE] is about using imagery in the service of communication” (Christiansen 2023, 13).

Understanding

“data visualisation aims to facilitate understanding” (Kirk 2019, 20).

Three phases of understanding (Kirk 2019, 20)

perceiving interpreting comprehending

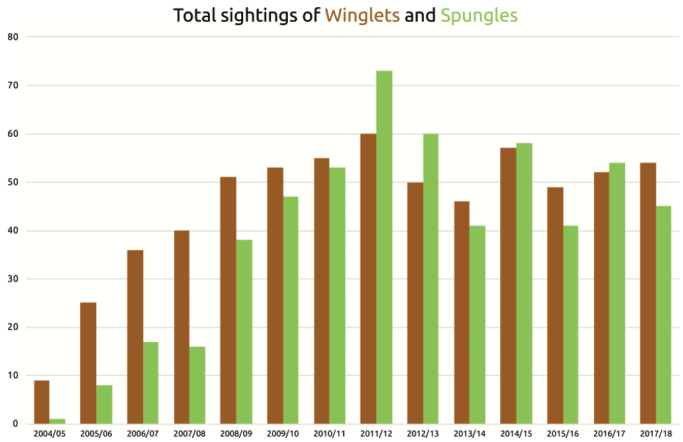


Figure 1.6 Total Sightings of Winglets and Spungles

source: Kirk (2019), p. 23

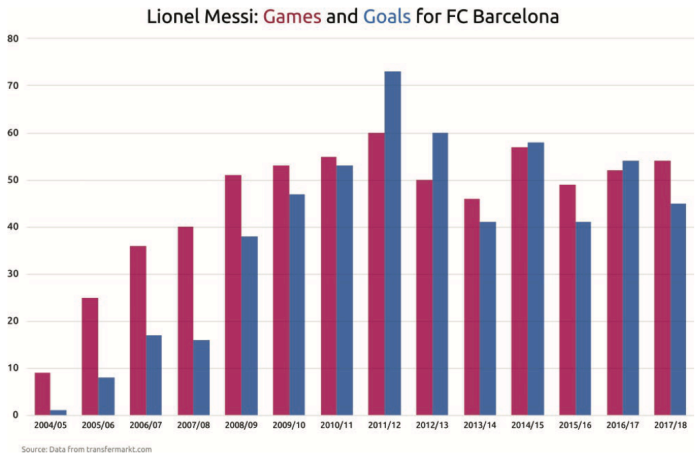


Figure 1.5 Lionel Messi: Games and Goals for FC Barcelona

Source: Data from transfermarkt.com

source: Kirk (2019), p. 21

Criteria for good visualizations

Principles of Graphical Excellence (Tufte 2007, 51)

1. **“Graphical excellence is the well-designed presentation of interesting data–a matter of substance, of statistics, and of design.”**
2. **“Graphical excellence consists of complex ideas communicated with clarity, precision, and efficiency.”**
3. **“Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.”**

Cairo (2016), p. 12 und Kirk (2019), p. 38

- truthful (see also D’Ignazio and Klein (2020))
- functional and accessible
- insightful and enlightening
- elegant and beautiful

Simplicity vs complexity

> "Any visualization is a model" (Cairo 2016, 69)

- "Good models abstract reality while keeping its essence at the same time… The more adequately a model fits whatever it stands for without being needlessly complex, and the easier it is for its intended audience to interpret it correctly, the better it will be." (Cairo 2016, 70)
- "Simplicity is about subtracting the obvious and adding the meaningful." (Cairo 2016, 97)
- "Good visualizations shouldn't oversimplify information. They need to clarify it. In many cases, clarifying a subject requires increasing the amount of information, not reducing it." (Cairo 2016, 78)
- "Simplicity isn't just about reduction. It can (and should) also be about augmentation. It consists of removing what isn't relevant from our models but also of bringing in those elements that are essential to making those models truer." (Cairo 2016, 97)

Preattentative Processing

Use `preattentive attributes` to direct observer's focus.

Table 1. Our sales grew to $600 million this year

	Q1	Q2	Q3	Q4
Bob	26	35	72	84
Ellie	22	15	61	35
Gerrie	19	20	71	55
Jack	22	95	13	64
Jon	83	62	46	48
Karen	30	65	98	82
Ken	38	28	45	71
Lauren	98	81	41	63
Steve	16	50	23	41
Valerie	46	24	30	57
Total	**$400**	**$475**	**$500**	**$600**

Table 1. Our sales grew to $600 million this year

	Q1	Q2	Q3	Q4
Bob	26	35	72	84
Ellie	22	15	61	35
Gerrie	19	20	71	55
Jack	22	95	13	64
Jon	83	62	46	48
Karen	30	65	98	82
Ken	38	28	45	71
Lauren	98	81	41	63
Steve	16	50	23	41
Valerie	46	24	30	57
Total	**$400**	**$475**	**$500**	**$600**

Table 1. Our sales grew to $600 million this year

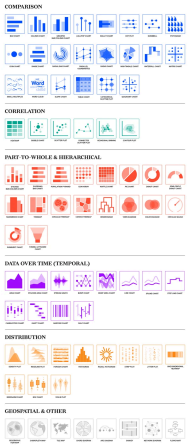
	Q1	Q2	Q3	Q4
Bob	26	35	72	**84**
Ellie	22	15	61	35
Gerrie	19	20	71	55
Jack	22	**95**	13	64
Jon	83	62	46	48
Karen	30	65	**98**	82
Ken	38	28	45	71
Lauren	**98**	81	41	63
Steve	16	50	23	41
Valerie	46	24	30	57
Total	**$400**	**$475**	**$500**	**$600**

source: Schwabish (2021), p. 26

Process of visualizing data

Building blocks

> Visual “[r]epresentation involves making decisions about how you are going to portray your data visually so that the subject understanding it offers can be made accessible to your audience. In simple terms, this is all about charts and the act of selecting the right chart to show the features of your data that you think are most relevant.” (Kirk 2019, 17)



Source: https://www.datylon.com/

Building blocks of any visualization (Kirk 2019, 17–18):

- **`marks`**: Elements used to represent items of data (i.e. points, columns, lines, etc.)
- **`attributes`**: visual variations of marks to represent the values associated with each (text, color, shape, etc.)

4 phases of the visualization design process

see Kirk (2019, 32)

`Phase 1: concept`

planing and defining project

`Phase 2: data`

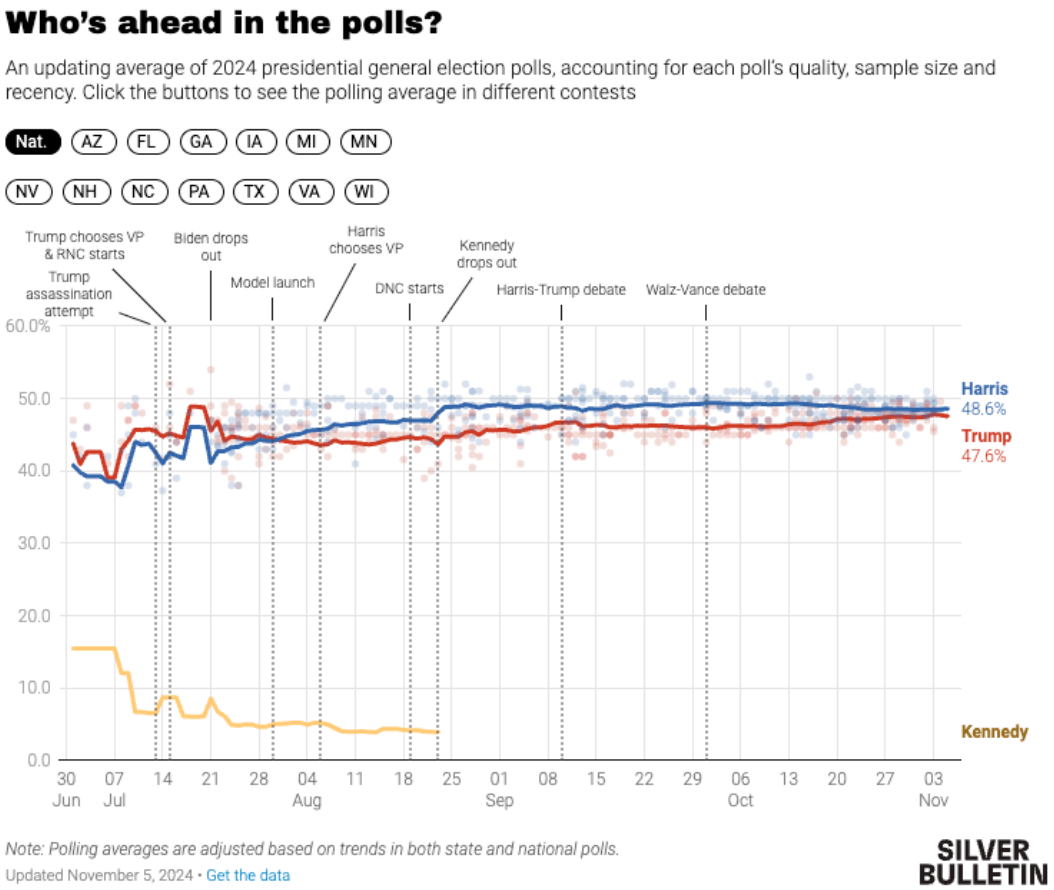
gathering, handling and preparing your data; getting to know the data

`Phase 3: "editorial thinking"`

defining what you will show your audience; what do we want to communicate (main message)

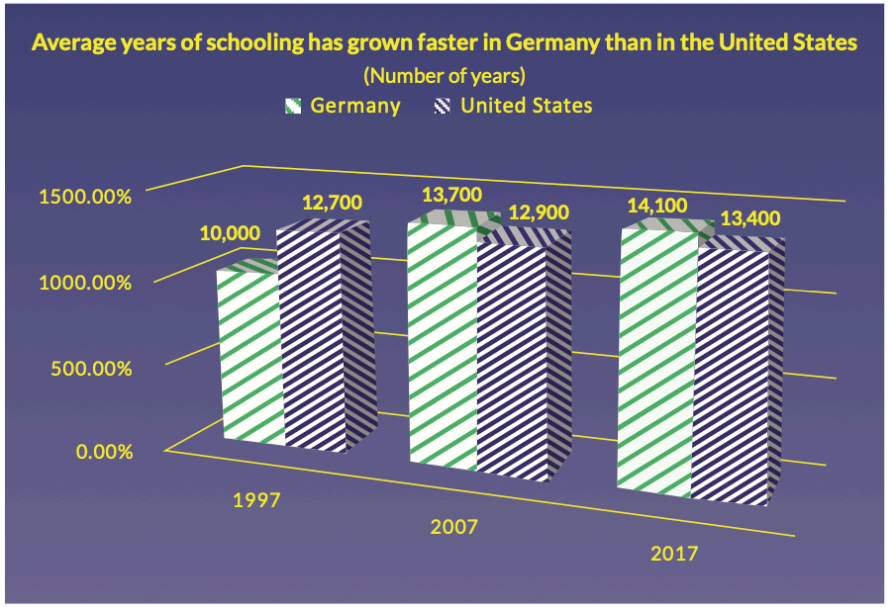
Phase 4: "Design" (see Schwabish 2021, 29–45)

- show the data

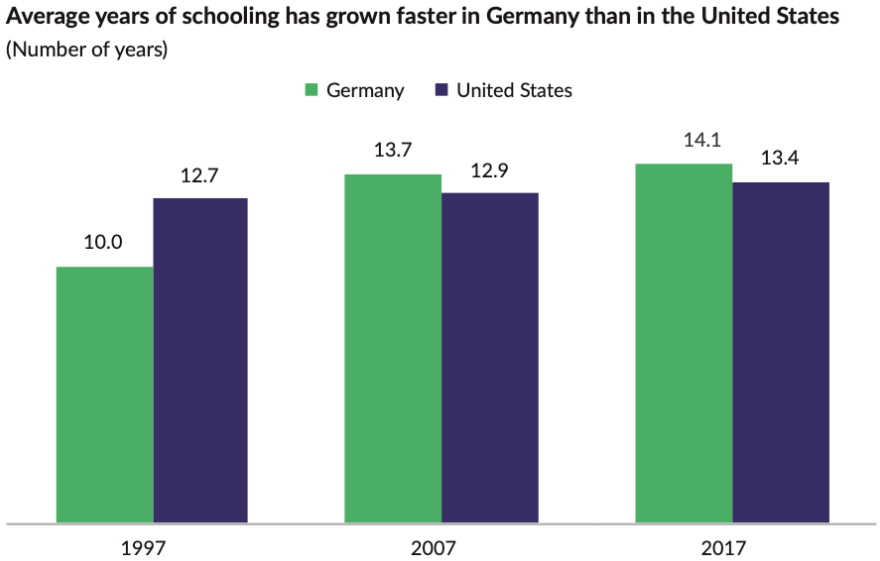


Source: Final Silver Bulletin 2024 presidential election forecast

- reduce the clutter

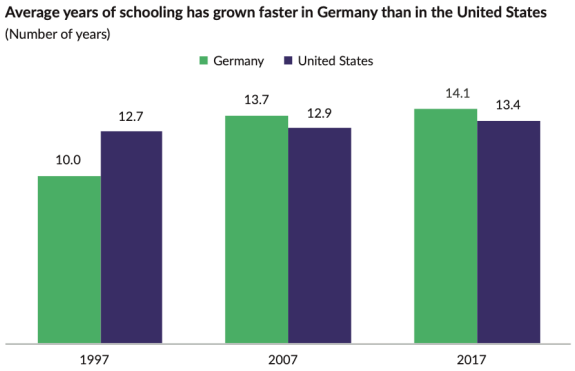


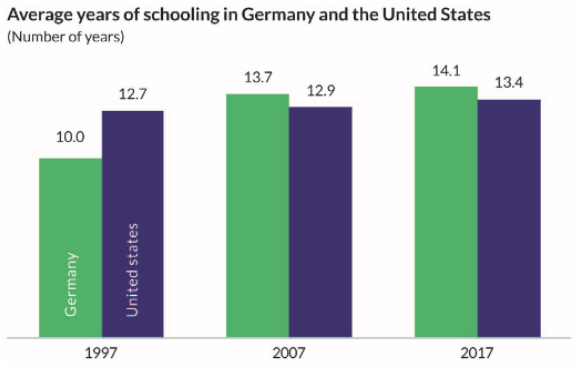
source: Schwabish (2021), p. 32

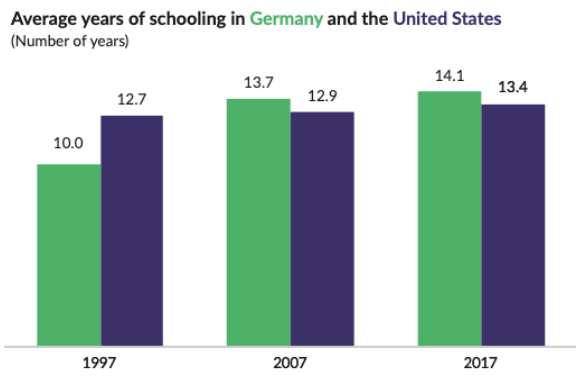


source: Schwabish (2021), p. 33

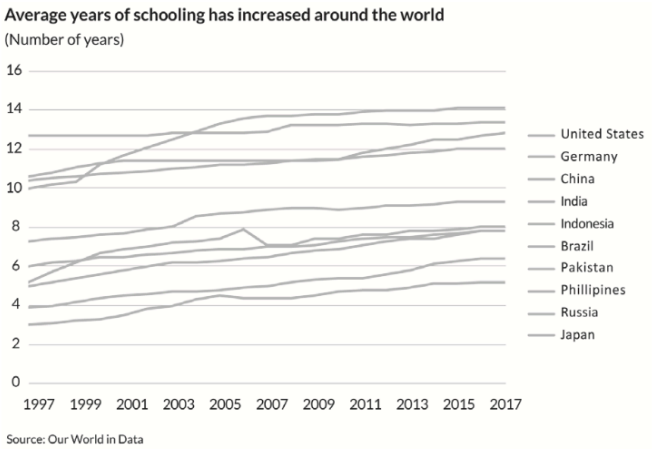
- integrate the graphics and text
  - remove legend (if possible) and label data directly
  - titles like newspaper headlines
  - add explainers

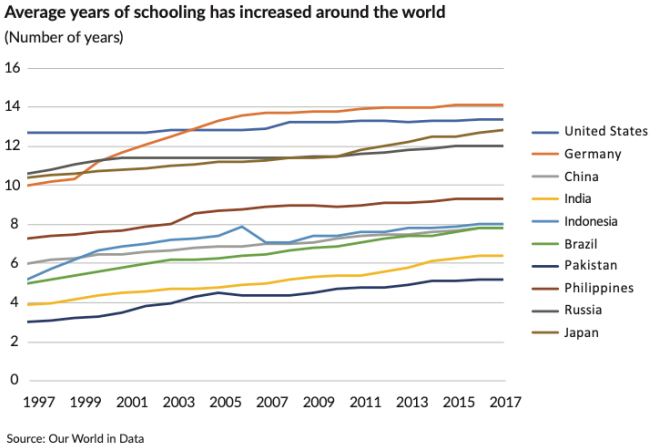


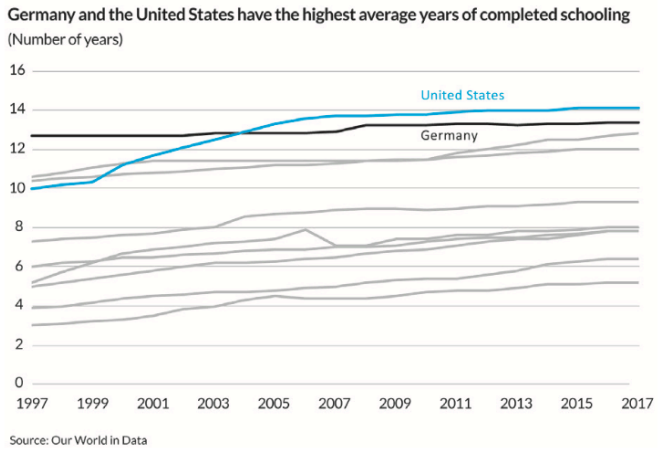




- avoid the “spaghetti chart”
- start with gray







RStudio/Jupyter and ggplot2

Grammar of Graphics

- graphic packages in R (base, lattice/trellis, grid graphics)
- ggplot2 part of tidyverse
- based on the idea of “Grammar of Graphics” (Wilkinson 2005; Wickham 2010)

> “A grammar of graphics is a tool that enables us to concisely describe the components of a graphic. Such a grammar allows us to move beyond named graphics (e.g., the”scatterplot”) and gain insight into the deep structure that underlies statistical graphics.” (Wickham 2010, 3)

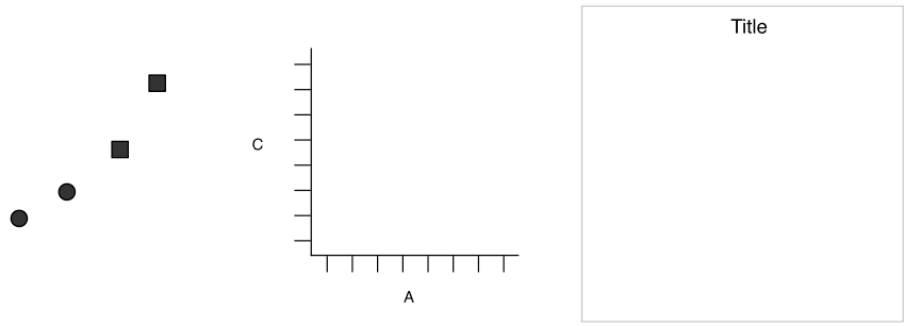


Figure 1. Graphics objects produced by (from left to right): geometric objects, scales and coordinate system, plot annotations.

Wickham (2010), p. 6

ggplot2 – layered graphics

`plot = data ∩ mapping`

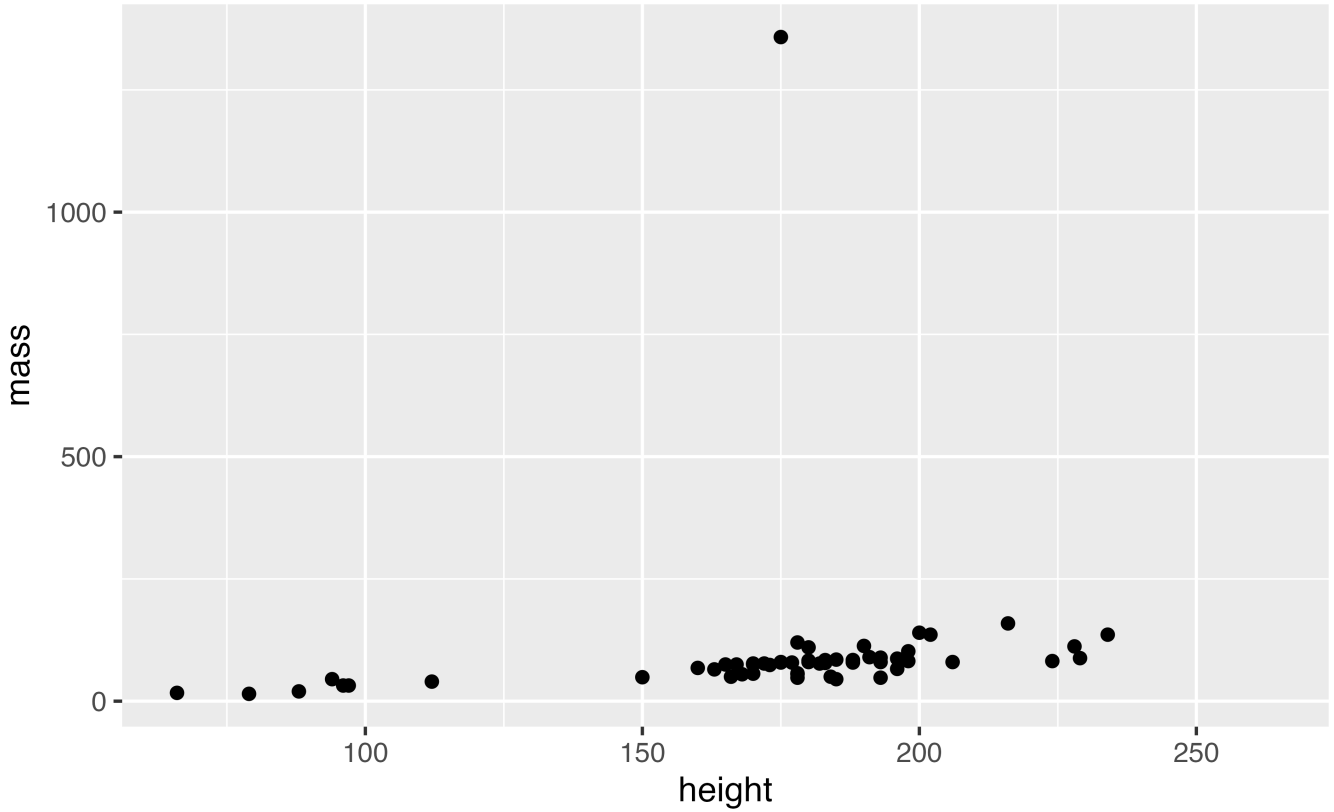
mappings comprise 5 elements

- **layer**: collection of geometric elements (**geom**s) und statistical transformations (**stat**s)
- **scale**s: display of values (colors, shapes, size) and axis
- **coord**: coordination system
- **facet**: splitting data into subsets
- **theme**: “design” of the plots (background color, fonts, etc.)

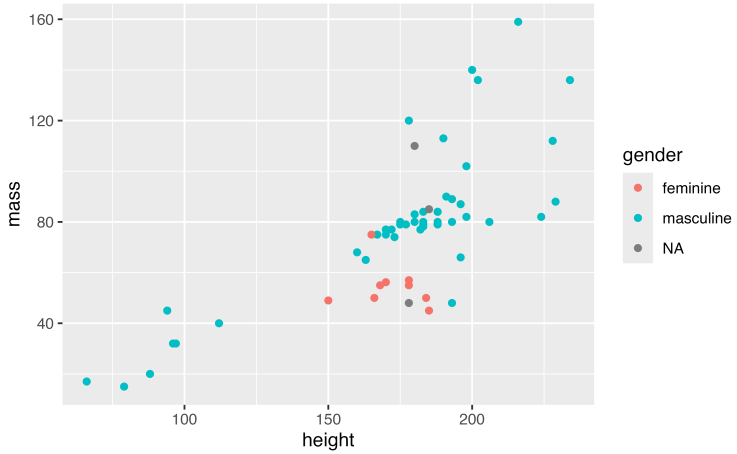
Plotting with ggplot2

- ggplot2 code consists of three components: (1) data, (2) aesthetic mapping, (3) geom function

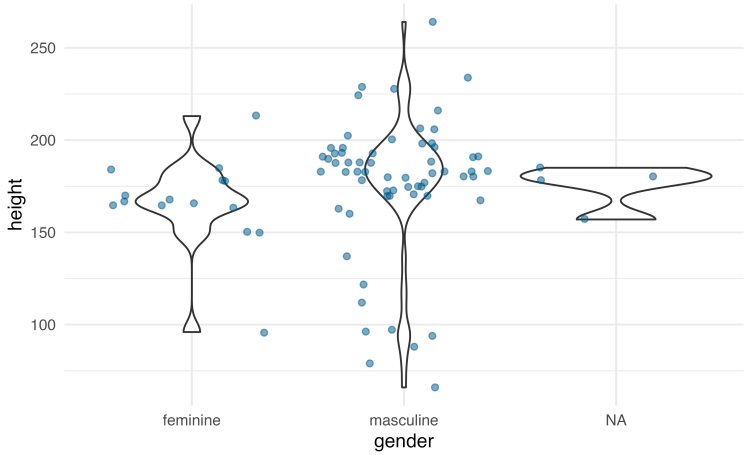
```
require(tidyverse) # loading the tidyverse package

ggplot(data = starwars, aes(x = height, y = mass)) +
    geom_point()
```



```
starwars |> filter(mass < 500) |>
    ggplot(aes(height, mass, color = gender)) +
    geom_point()
```




```
ggplot(starwars, aes(gender, height)) +
    geom_violin() +
    geom_jitter(color = "#005c8b", alpha = 0.5) +
    theme_minimal()
```



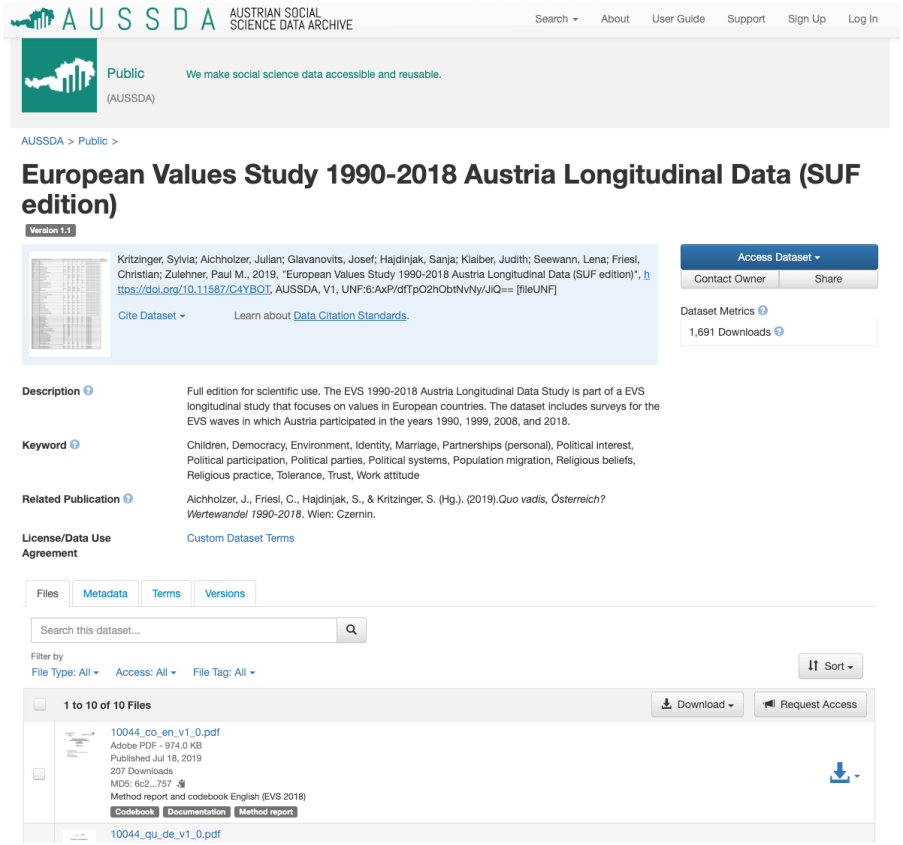
Example from European Values Study

Phase 1: Concept

`Research question:` “How tolerant has Austrian society become over time?”

`How to measure` tolerance?

Phase 2: Data



Kritzinger, Sylvia; Aichholzer, Julian; Glavanovits, Josef; Hajdinjak, Sanja; Klaiber, Judith; Seewann, Lena; Friesl, Christian; Zulehner, Paul M., 2019, "European Values Study 1990-2018 Austria Longitudinal Data (SUF edition)", https://doi.org/10.11587/C4YBOT, AUSSDA, V1.

Questionnaire: 10044_qu_en_v1_0.pdf

Q44 Please tell me for each of the following whether you think it can always be justified, never be justified, or something in between, using this card.

		never									always	DK	NA
v149	Claiming state benefits which you are not entitled to	1	2	3	4	5	6	7	8	9	10	88	99
v150	Cheating on tax if you have the chance	1	2	3	4	5	6	7	8	9	10	88	99
v151	Taking the drugs marijuana or hashish	1	2	3	4	5	6	7	8	9	10	88	99
v152	Someone accepting a bribe in the course of their duties	1	2	3	4	5	6	7	8	9	10	88	99
v153	Homosexuality	1	2	3	4	5	6	7	8	9	10	88	99
v154	Abortion	1	2	3	4	5	6	7	8	9	10	88	99
v155	Divorce	1	2	3	4	5	6	7	8	9	10	88	99
v156	Euthanasia (terminating the life of the incurably sick)	1	2	3	4	5	6	7	8	9	10	88	99
v157	Suicide	1	2	3	4	5	6	7	8	9	10	88	99
v158	Having casual sex	1	2	3	4	5	6	7	8	9	10	88	99
v159	Avoiding a fare on public transport	1	2	3	4	5	6	7	8	9	10	88	99
v160	Prostitution	1	2	3	4	5	6	7	8	9	10	88	99
v161	Artificial insemination or in-vitro fertilization	1	2	3	4	5	6	7	8	9	10	88	99
v162	Political violence	1	2	3	4	5	6	7	8	9	10	88	99
v163	Death penalty	1	2	3	4	5	6	7	8	9	10	88	99

Codebook/Method report

Variable: Justifiable: homosexuality

F117	Justifiable: someone accepting a bribe
F118	Justifiable: homosexuality
F119	Justifiable: prostitution

Variable: Wave

Indicator for the four waves

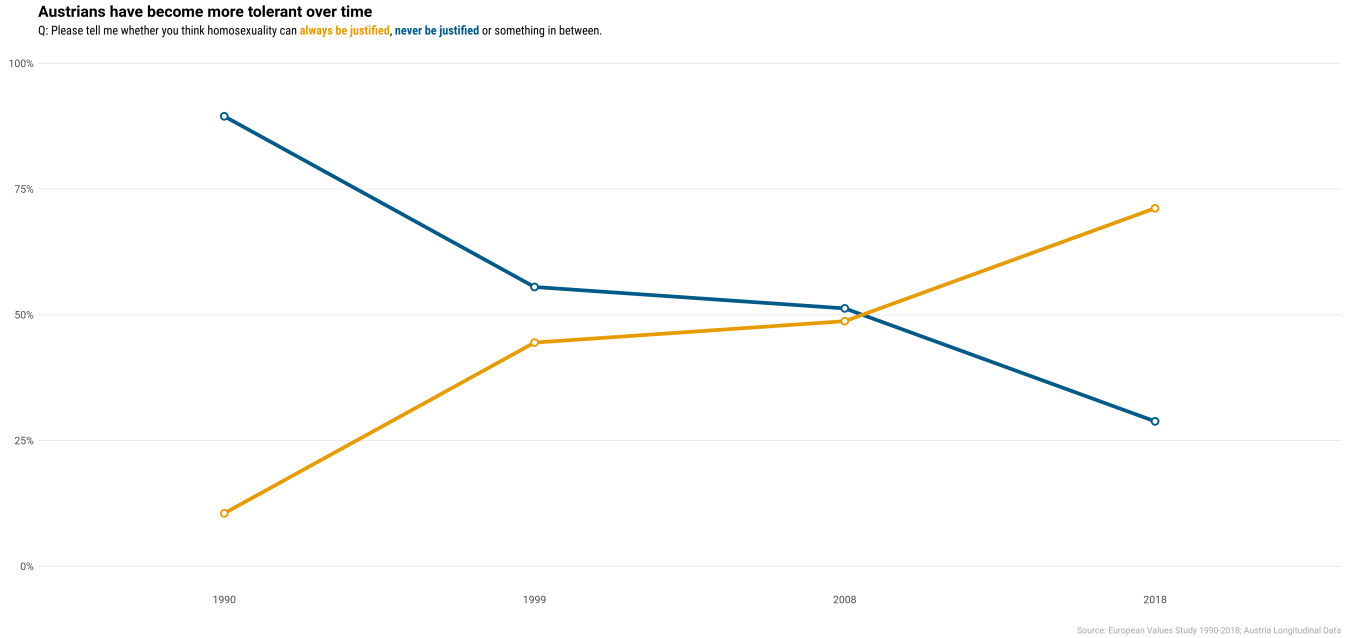
The variable **S002EVS** is the indicator for which wave respondents were asked. The value labels are structured as follows:

- 2 „1990“
- 3 „1999“
- 4 „2008“
- 5 „2018“

Variable: Sex

W011	Dependency on social security during last 5 years spouse/partner
X001	Sex
X002	Year of birth

Phase 3: “Editorial thinking”



see Section 6.1 for code

- only 2 out of 10 categories
- variance by gender?
- danger of spaghetti chart (2 x 10 categories)
- show the data?

Phase 4: Design

Configuration

```
require(pacman) # R package management tool

p_load(tidyverse,
       showtext, #  using non-standard fonts in R graphs (extrafonts)
       Cairo, # embed fonts in graphs
       ggtext, # for coloring title in plots
       sjlabelled, # for using SPSS labels
       dataverse # for API access to AUSSDA/Dataverse
)
```

Load and tidy data

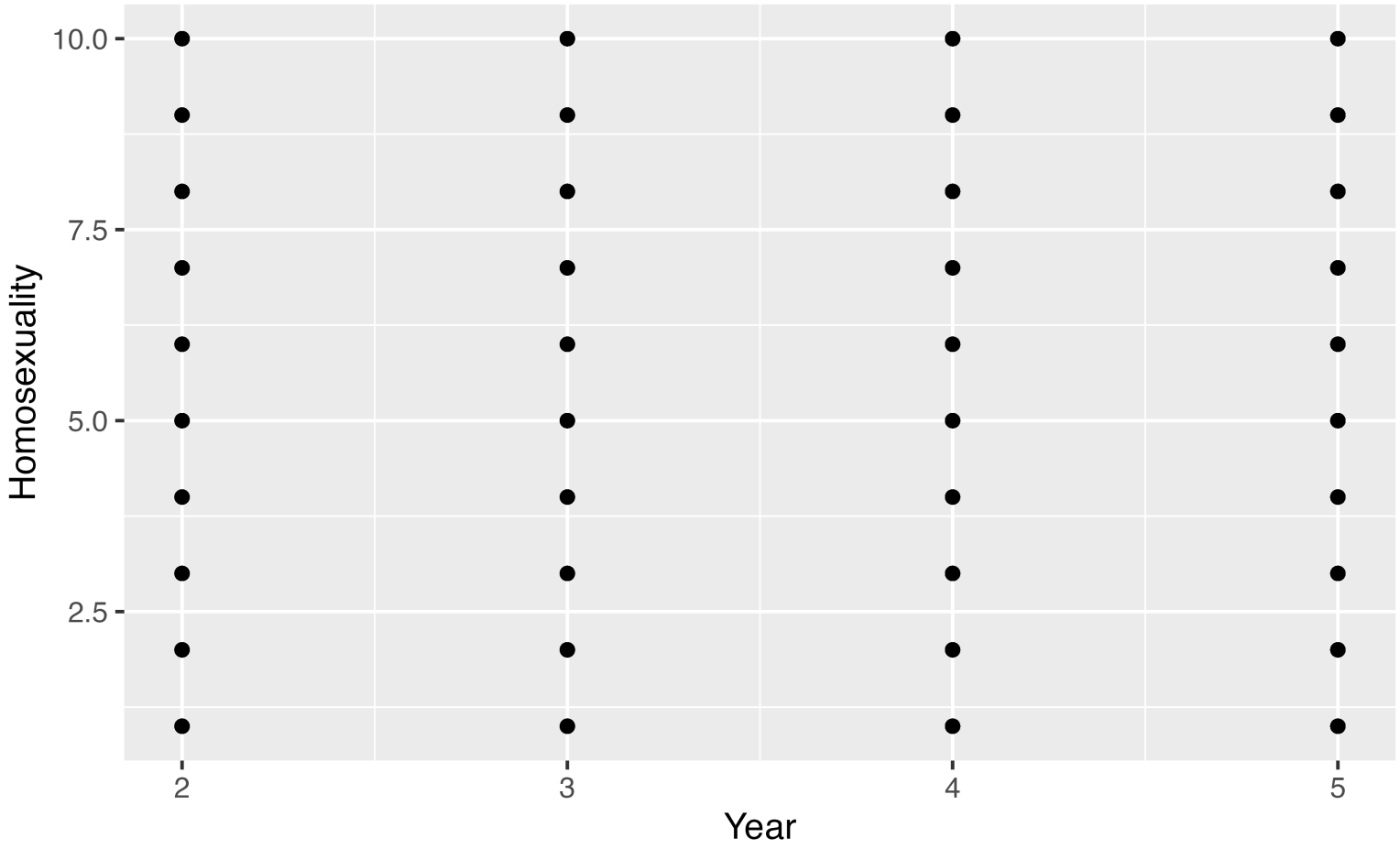
```r
## Specifying the API Token we received from AUSSDA
Sys.setenv("DATAVERSE_KEY" = "xyz")

df_evs <-
  get_dataframe_by_name(
    filename    = "10048_da_en_v1_0-1.tab",
    dataset     = "10.11587/C4YBOT",
    .f          = haven::read_dta, # for reading SPSS tab file
    original    = TRUE,
    server      = "data.aussda.at")

df <- df_evs |> select(Year = S002EVS,
                       Sex = X001,
                       Homosexuality = F118) # select variables and rename them
```

```r
head(df)
```

```
# A tibble: 6 × 3
  Year      Sex         Homosexuality
  <dbl+lbl> <dbl+lbl>   <dbl+lbl>
1 3 [1999]  1 [Male]    10 [Always justifiable]
2 5 [2018]  2 [Female]   6 [6]
3 3 [1999]  1 [Male]    10 [Always justifiable]
4 3 [1999]  2 [Female] 10 [Always justifiable]
5 2 [1990]  2 [Female]   3 [3]
6 4 [2008]  2 [Female]   3 [3]
```

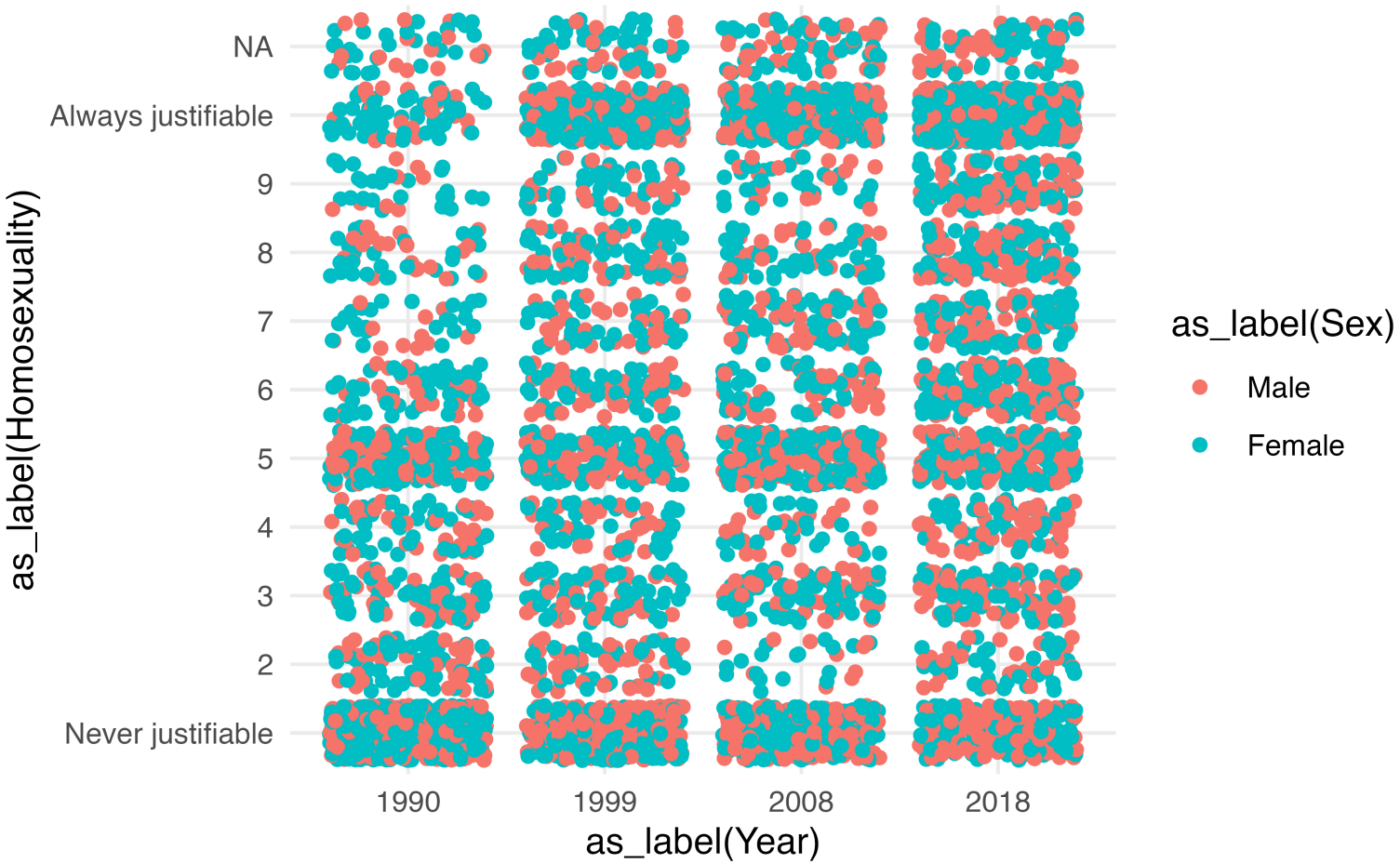
Step 1

```
df |> group_by(Sex, Year) |>
    ggplot(aes(Year, Homosexuality)) +
    geom_point()
```



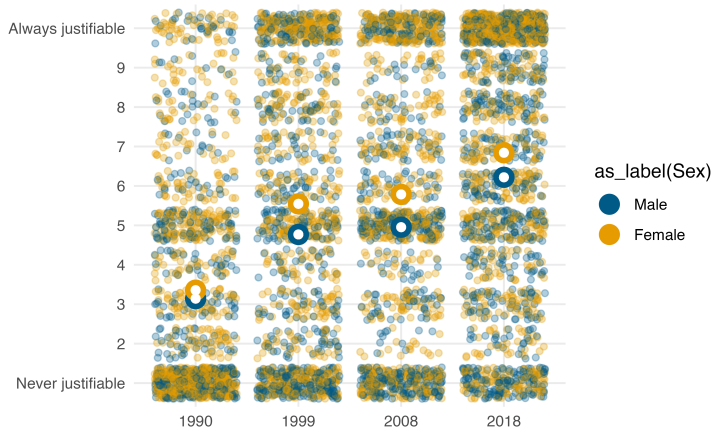
Step 2

```
df |> group_by(Sex, Year) |>
    ggplot(aes(as_label(Year), as_label(Homosexuality), color = as_label(Sex))) +
    geom_jitter() +
    theme_minimal()
```



Step 3

```
p <- df |>
    filter(!is.na(Homosexuality)) |>
    group_by(Sex, Year) |>

    mutate(mean_Homosexuality =
               mean(Homosexuality,
                    na.rm = TRUE)) |>

    ggplot(aes(as_label(Year),
               as_label(Homosexuality),
               color = as_label(Sex))) +

    scale_color_manual(values =
                           c("#005c8b",
                             "#E69F00")) +

    geom_jitter(alpha = .3) +

    geom_point(aes(y = mean_Homosexuality,
                   color = as_label(Sex)),
               size = 5) +

    geom_point(aes(y = mean_Homosexuality),
               size = 2, color = "white") +

    labs(x = "", y = "") +
    theme_minimal()

p
```

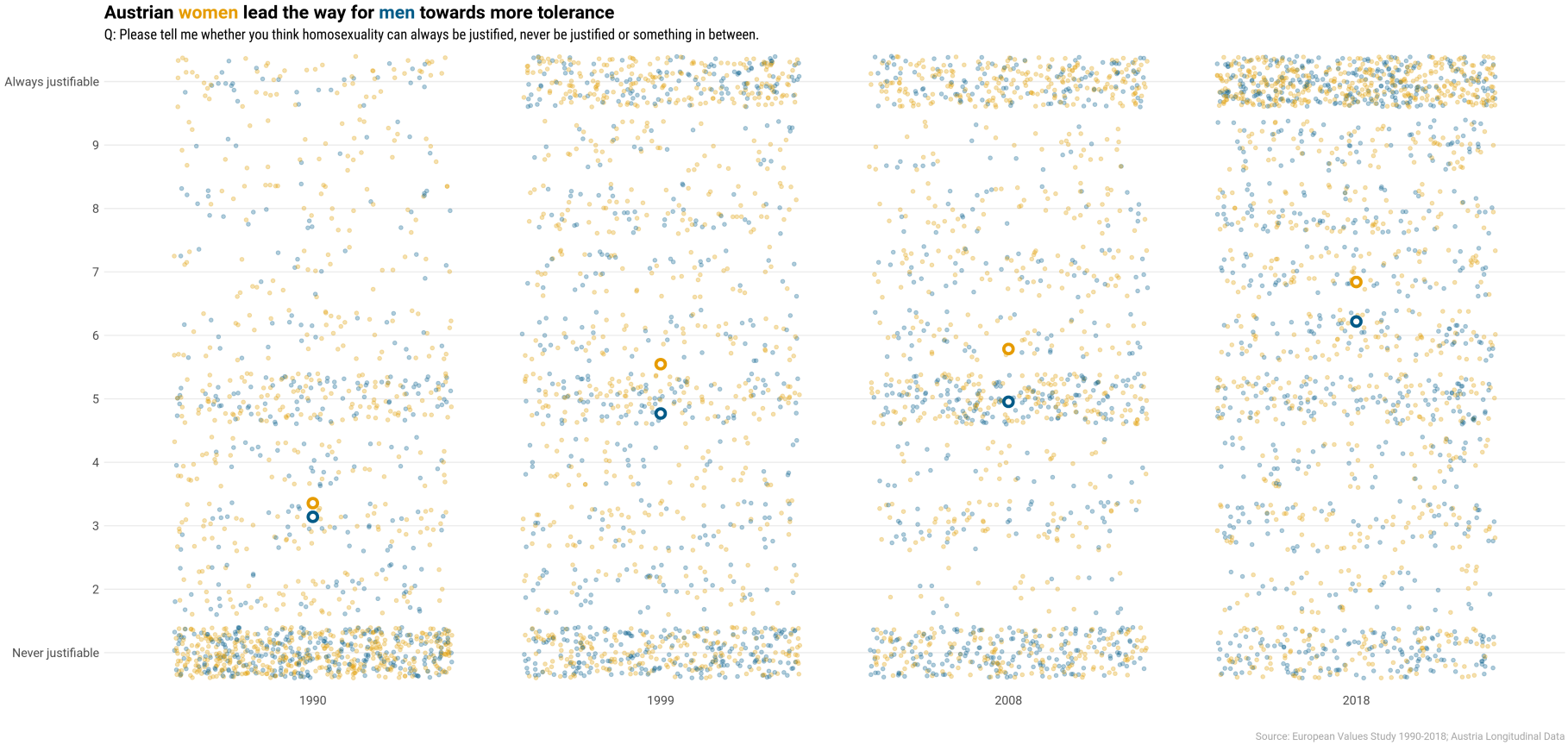


- Also check COLORBREWER 2.0 for “colorblind safe” colors.

Step 4

```
fontfamily1 <- "Roboto" # fonts have to be installed on the computer
fontfamily2 <- "Roboto Condensed"

p <- p + labs(title = "<b>Austrian <span style = 'color: #E69F00;'>women</span>
              lead the way for <span style = 'color: #005c8b;'>men</span> towards more
              tolerance</b>") +
    labs(subtitle = "Q: Please tell me whether you think homosexuality can always be justified,
         never be justified or something in between.") +
    labs(caption = "Source: European Values Study 1990-2018; Austria Longitudinal Data") +
    theme(text = element_text(size = 14, family = fontfamily1),
          title = element_text(size = 18, family = fontfamily1),
          plot.title = element_text(size = 18, family = fontfamily1),
          plot.subtitle =  element_markdown(size = 14, family = fontfamily2,
                                            margin = ggplot2::margin(1, 0, 1, 0)),
          axis.text.x = element_text(size = 12, family = fontfamily1),
          axis.text.y = element_text(size = 12, family = fontfamily1),
          plot.caption = element_text(size = 10, family = fontfamily1, color = "darkgrey")) +
    theme(plot.title = element_markdown(),
          plot.subtitle = element_markdown(),
          plot.caption = element_markdown(),
          panel.grid.major.x = element_blank(),
          panel.grid.minor.y = element_blank(),
          legend.position="none")
```

Final Plot


Austrian women lead the way for men towards more tolerance
Q: Please tell me whether you think homosexuality can always be justified, never be justified or something in between.
Always justifiable
9
8
7
6
5
4
3
2
Never justifiable
1990
1999
2008
2018
Source: European Values Study 1990-2018; Austria Longitudinal Data

Final code (including saving plot)

```
png(filename = "plots/plot_homosexuality-final.png",
    width = 21.7,
    height = 10.2,
    units = "in",
    res = 300,
    bg = "#ffffff",
    type = "cairo-png"
)

df |> filter(!is.na(Homosexuality)) |>
    group_by(Sex, Year) |>
    mutate(mean_Homosexuality = mean(Homosexuality, na.rm = TRUE)) |>
    ggplot(aes(as_label(Year), as_label(Homosexuality), color = as_label(Sex))) +
    scale_color_manual(values = c("#005c8b", "#E69F00")) +
    geom_jitter(alpha = .3) +
    geom_point(aes(y = mean_Homosexuality, color = as_label(Sex)), size = 5) +
    geom_point(aes(y = mean_Homosexuality), size = 2, color = "white") +
    labs(x = "", y = "") +
    labs(title = "<b>Austrian <span style = 'color: #E69F00;'>women</span> lead the way for <span style = 'color: #005c8b;'>men
    labs(subtitle = "Q: Please tell me whether you think homosexuality can always be justified, never be justified</span> or so
    labs(caption = "Source: European Values Study 1990-2018; Austria Longitudinal Data") +
    theme_minimal() +
    theme(text = element_text(size = 14, family = fontfamily1),
          title = element_text(size = 18, family = fontfamily1),
          plot.title = element_text(size = 18, family = fontfamily1),
          plot.subtitle =  element_markdown(size = 14, family = fontfamily2, margin = ggplot2::margin(1, 0, 1, 0)),
          axis.text.x = element_text(size = 12, family = fontfamily1),
          axis.text.y = element_text(size = 12, family = fontfamily1),
          plot.caption = element_text(size = 10, family = fontfamily1, color = "darkgrey")) +
    theme(plot.title = element_markdown(),
          plot.subtitle = element_markdown(),
```

Bibliography

Cairo, Alberto. 2016. *The Truthful Art: Data, Charts, and Maps for Communication*. Indianapolis, IN: New Riders.

Christiansen, Jen. 2023. *Building Science Graphics: An Illustrated Guide to Communicating Science Through Diagrams and Visualizations*. Boca Raton, FL: CRC Press. https://doi.org/10.1201/9781003217817.

D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism*. Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/11805.001.0001.

Kirk, Andy. 2019. *Data Visualisation: A Handbook for Data Driven Design*. 2nd ed. London et al.: Sage.

Schwabish, Jonathan. 2021. *Better Data Visualizations: A Guide for Scholars, Researchers, and Wonks*. New York, NY: Columbia University Press.

Tufte, Edward R. 2007. *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, CT: Graphic Press.

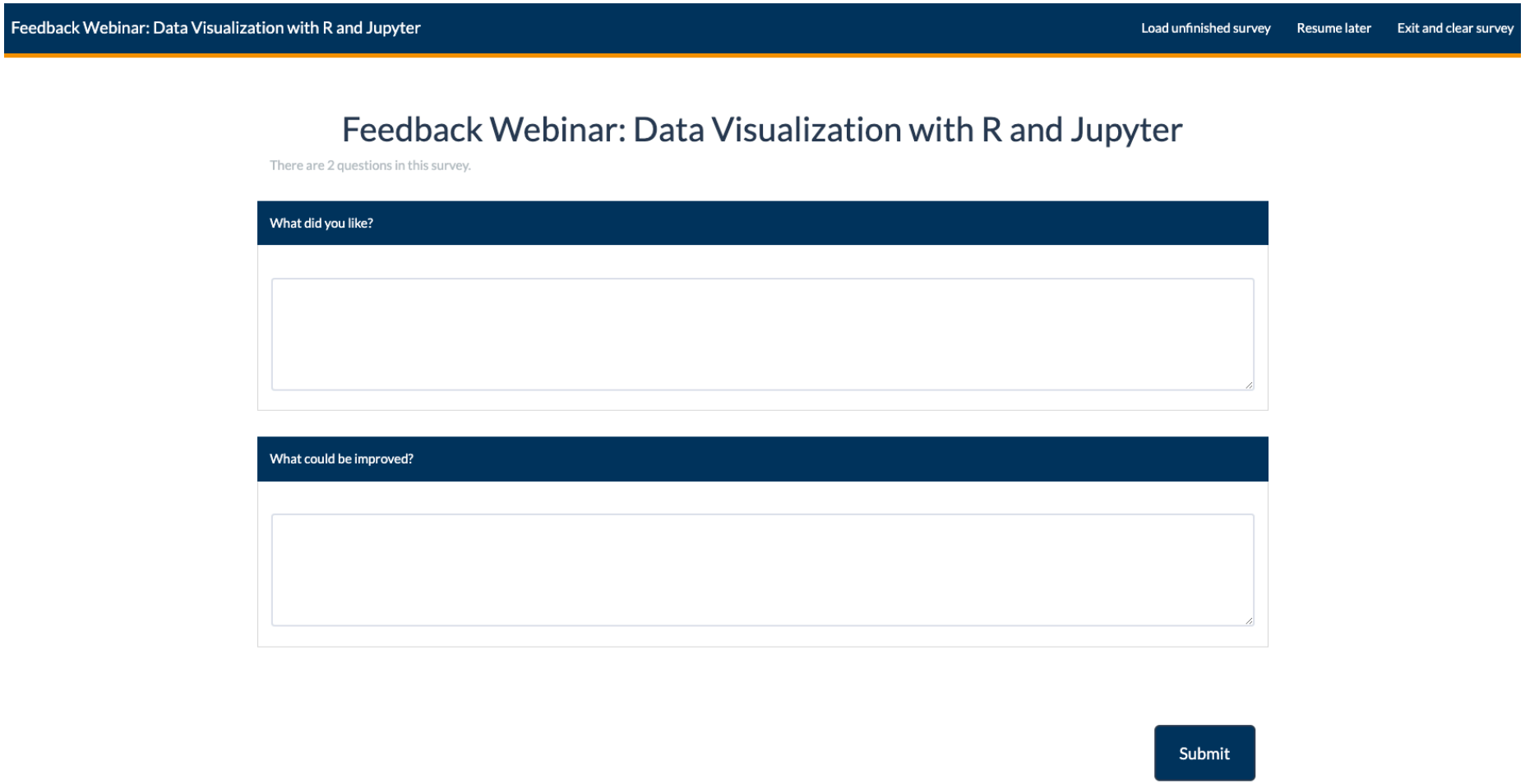
Wickham, Hadley. 2010. "A Layered Grammar of Graphics." *Journal of Computational and Graphical Statistics* 19 (1): 3–28. https://doi.org/10.1198/jcgs.2009.07098.

———. 2024. *Ggplot2: Elegant Graphics for Data Analysis*. 3rd ed. Heidelberg: Springer. https://ggplot2-book.org/.

Wilkinson, Leland. 2005. *The Grammar of Graphics*. 2nd ed. New York, NY: Springer. https://doi.org/10.1007/0-387-28695-0.

Q & A

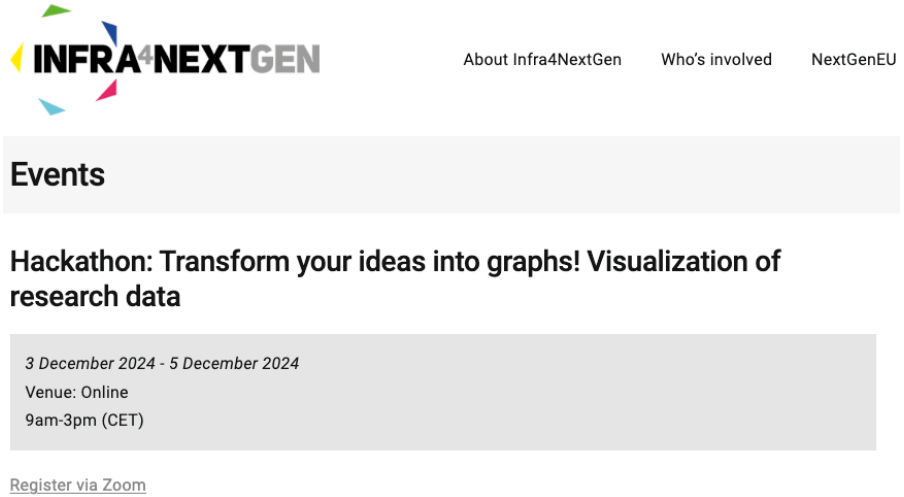
Feedback



https://umfrage.uibk.ac.at/limesurvey/allgemein/index.php/693916?lang=en

Upcoming Events

Register



Register

Appendix

geom_line plot

```r
# define font families for title, subtitel and annotations
fontfamily1 <- "Roboto"
fontfamily2 <- "Roboto Condensed"

df |> filter(Homosexuality == 10 | Homosexuality == 1) |>
    pivot_longer(cols = c(Homosexuality)) |>
    group_by(Year, value) |>
    summarise(n = n()) |>
    mutate(N = max(cumsum(n)), freq = n/N) |>
    ggplot(aes(x = as_label(Year), y = freq, group = as_label(value), color = as_label(value))) +
    scale_color_manual(values = c("#005c8b", "#E69F00")) +
    geom_line(linewidth = 2) +
    geom_point(size = 4) +
    geom_point(size = 2, color = "white") +
    scale_y_continuous(labels = scales::percent, limits = c(0,1)) +
    labs(x = "", y = "") +
    labs(title = "<b>Austrians have become more tolerant over time</b>") +
    labs(subtitle = "Q: Please tell me whether you think homosexuality can <b><span style = 'color: #E69F00;'>always be justifi
    labs(caption = "Source: European Values Study 1990-2018; Austria Longitudinal Data") +
    theme_minimal() +
    theme(text = element_text(size = 14, family = fontfamily1),
          title = element_text(size = 18, family = fontfamily1),
          plot.title = element_text(size = 18, family = fontfamily1),
          plot.subtitle =  element_markdown(size = 14, family = fontfamily2, margin = ggplot2::margin(1, 0, 1, 0)),
          axis.text.x = element_text(size = 12, family = fontfamily1),
          axis.text.y = element_text(size = 12, family = fontfamily1),
          plot.caption = element_text(size = 10, family = fontfamily1, color = "darkgrey")) +
    theme(plot.title = element_markdown(),
          plot.subtitle = element_markdown(),
          panel.grid.major.x = element_blank(),
          panel.grid.minor.y = element_blank(),
```