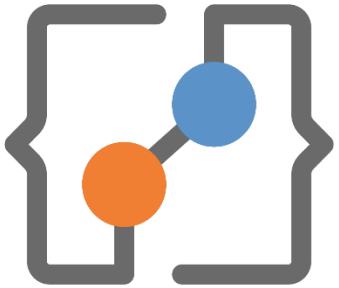
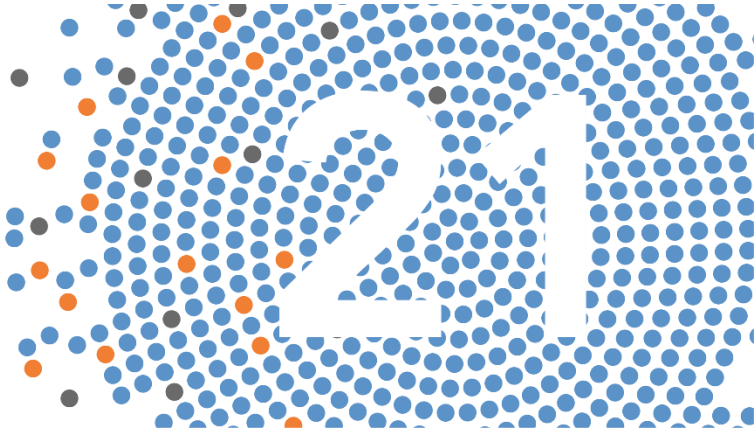


digiGEBF



ggplotting: Datenvisualisierung in R

Workshop im Rahmen der digiGEBF 2021

Prof. Dr. Martin Schultze (schultze@psych.uni-frankfurt.de)
Goethe Universität Frankfurt

Dr. Janine Buchholz (buchholz@dipf.de)
DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation

Kurze Vorstellungsrunde

- Wer sind wir?

Martin Schultze

- Juniorprofessor an der Goethe Universität Frankfurt
- Jahrelang Statistik- und Methodenberatung in Erziehungswissenschaft und Psychologie

Janine Buchholz

- Post-Doc am DIPF
- Jahrelange Mitarbeit in PISA und Lehrbeauftragte in Psychologie

- Wer sind Sie?
  ahaslides.com/ggplotting

  - Aus welchem Fach kommen Sie?
  - In welchem Bundesland sind Sie im Moment?
  - In welcher Phase sind Sie?
  - Wofür benutzen Sie R?

BLOCK 1:
EINFÜHRUNG

Warum Grafiken?

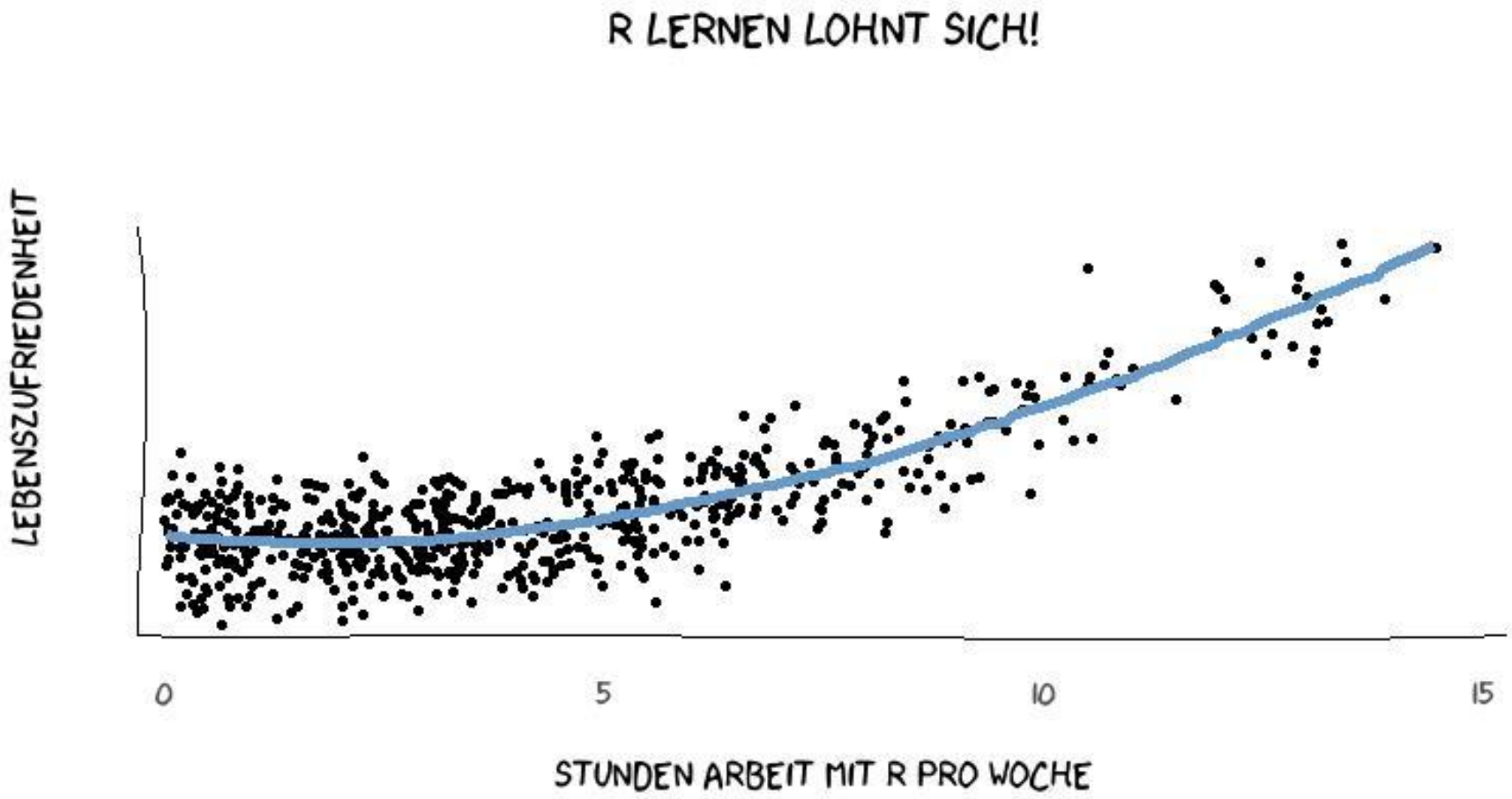
„Ein Bild sagt mehr als tausend… Zahlen!“

- Buchholz (2018)

Warum Grafiken?

Stunden_r	Happiness	Stunden_r	Happiness	Stunden_r	Happiness	Stunden_r	Happiness
8.79354157	5.926531	2.19365868	3.647024	4.77026893	4.124653	5.24821523	4.276749
1.64643868	5.09765	7.80326224	5.054431	2.07222655	3.830272	1.8740135	4.070819
3.30458103	3.364463	11.936665	8.675308	1.47291764	3.440652	3.14083383	4.15761
7.12040899	5.388203	3.59703602	4.39944	4.00354412	3.488141	2.93013284	3.947415
3.00385725	4.194416	1.52838097	3.603432	11.9940517	8.575037	1.59072256	4.095585
3.48925239	2.94819	1.58823299	3.411485	9.25453734	6.076989	2.10509734	5.075601
4.50677951	4.082005	10.21931	6.221165	1.59622618	4.165136	6.84191937	5.724141
5.18893522	4.640681	2.17693286	4.092225	7.16237939	4.850222	6.72631555	4.65076
3.13232111	3.954258	8.22275429	5.901928	6.5492407	5.401638	11.0124137	7.148269
5.47562831	5.176026	1.20616461	3.734106	5.4232525	3.963887	10.3342592	5.845189
5.10790541	4.958384	10.5455178	5.909956	10.7764698	6.904295	1.10491793	4.790856
3.76958827	4.198509	4.50892168	4.215662	2.07269413	3.568566	7.18337523	6.473316
7.03980999	4.419954	11.5005447	6.602336	0.08537306	4.290907	7.50587996	5.790064
4.89982559	4.418495	4.93869424	4.967047	2.62152172	3.58718	0.47164333	3.21261
2.26659257	4.031652	2.6610129	3.44773	8.48030335	5.014406	9.42217854	6.774196
0.78564771	4.867004	0.40273697	3.833715	0.10474203	5.227992	2.18715136	4.814066
1.61927122	4.148456	3.81494767	4.009963	0.49512557	3.147859	0.69589023	4.403185
2.89819741	3.793942	5.88486693	4.502429	8.99254386	5.923259	7.67758808	5.386187
3.24104889	4.189133	6.26355733	4.194851	6.40399748	5.303752	6.08779319	4.050406
5.61133734	5.955442	4.18284837	4.007913	0.43180052	3.040511	2.88693987	2.81238
6.02869262	4.334974	3.58431022	4.440284	1.23577216	4.82404	5.32889702	4.516181
1.9065174	3.937619	0.82966373	5.34705	12.4653949	9.051769	4.12093736	4.908319
1.02882093	4.243698	1.188942	4.197471	0.33424518	2.902807	9.35527414	6.198431
6.48419507	5.207116	3.1826174	3.857706	5.22506502	5.161855	0.74009047	2.893248
5.96751928	5.111402	6.54960145	5.007836	4.42154715	3.480279	1.49663797	3.645972

Warum Grafiken?

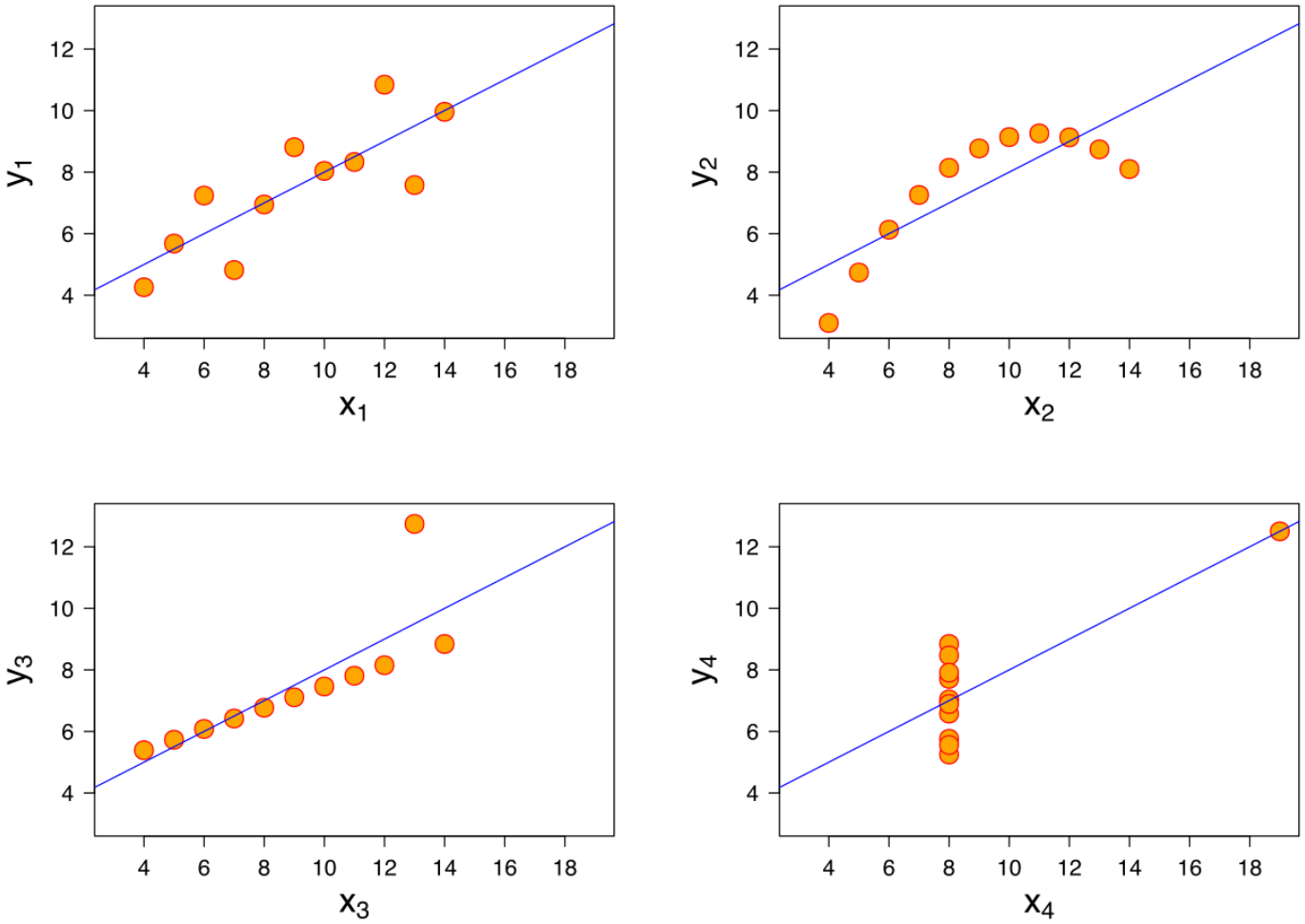


Warum Grafiken?

Gründe für Grafiken

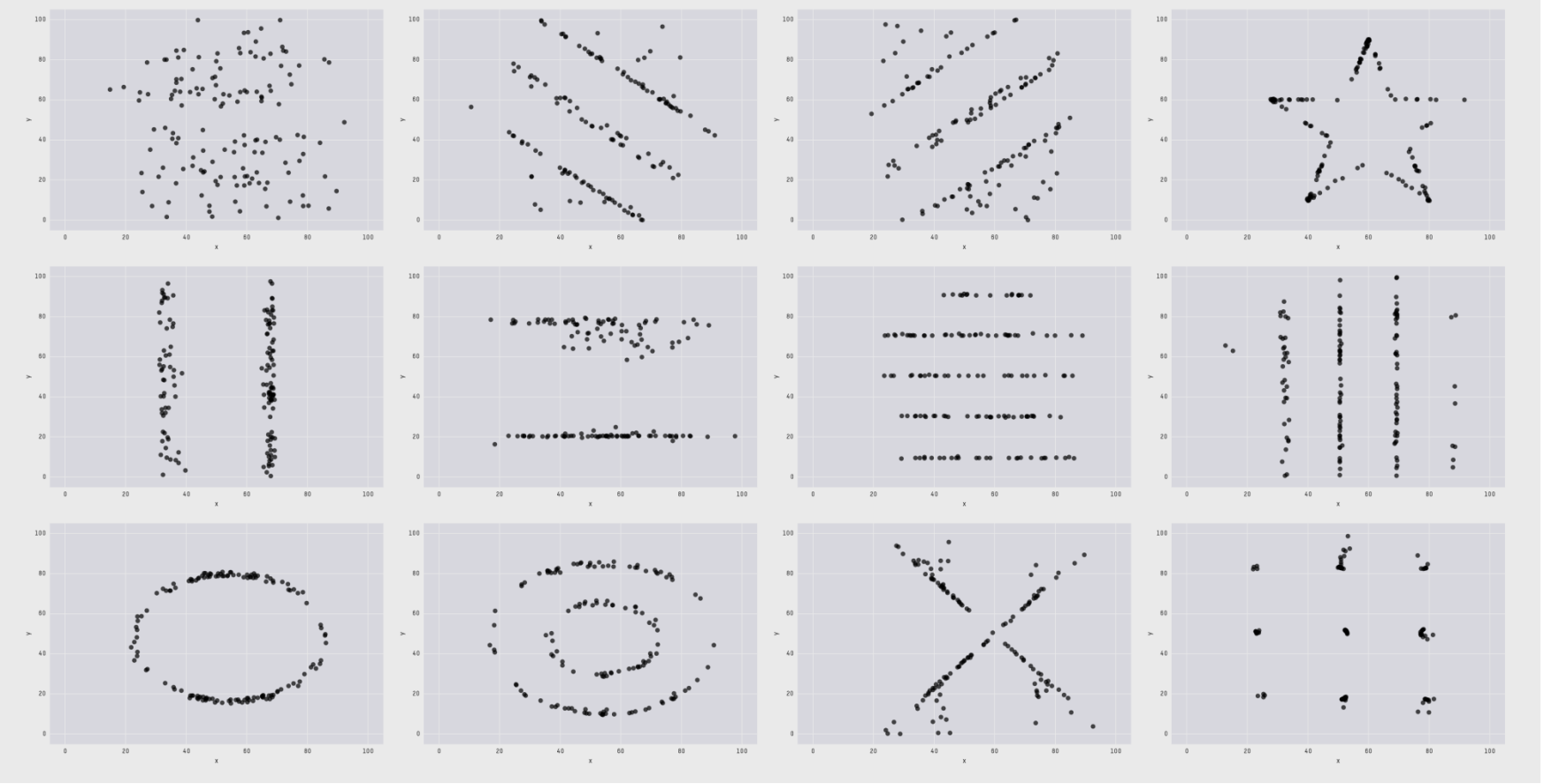
- Grafiken sind gut geeignet um
  - Muster und Informationen **anderen** zu kommunizieren.
  - Muster und Informationen **selbst** zu erkennen.
- Grafisch aufbereitete Inhalte können schnell verarbeitet werden
- Grafiken können auflockern im Gegensatz zu viel Text oder vielen Zahlen
- Grafiken können mehr Informationen vermitteln als andere statistische Verfahren

Anscombe-Quartett

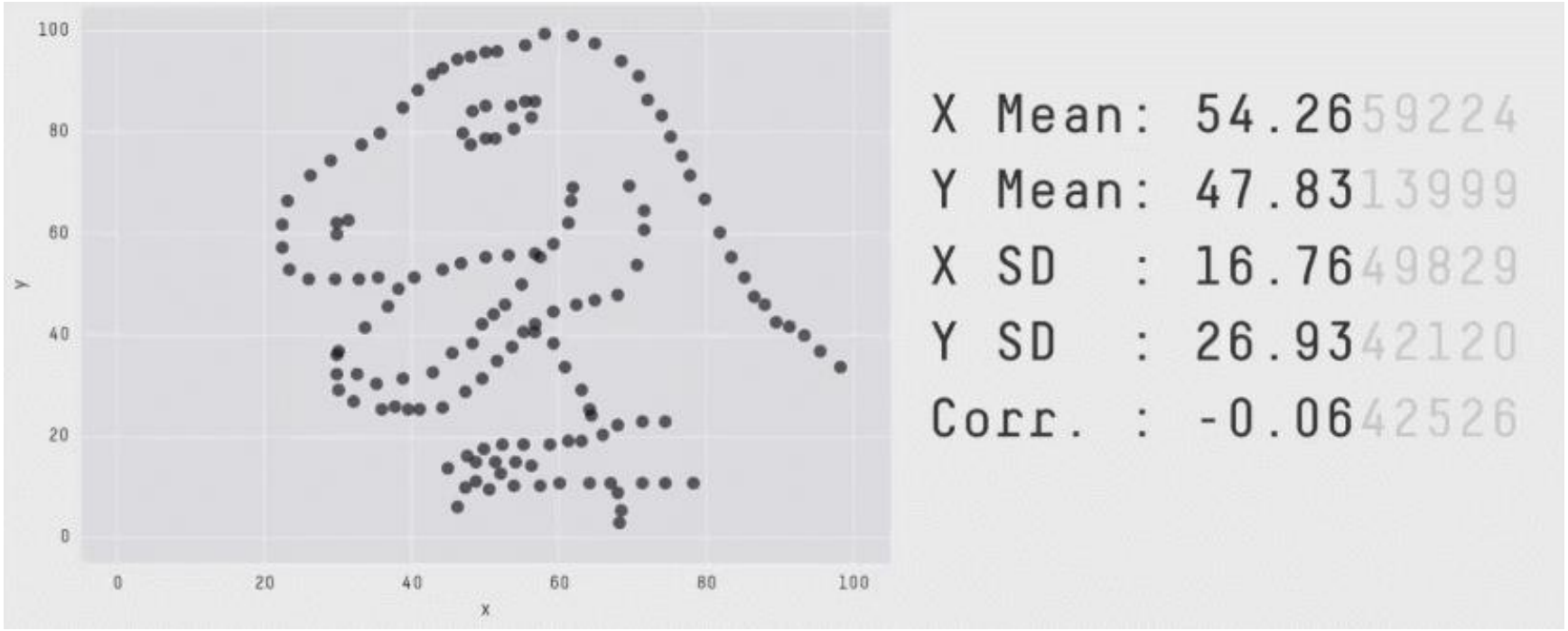


https://de.wikipedia.org/wiki/Anscombe-Quartett#/media/Datei:Anscombe's_quartet_3.svg

Anscombe-Quartett…erweitert



Datasaurus Dozen



https://www.autodeskresearch.com/publications/samestats

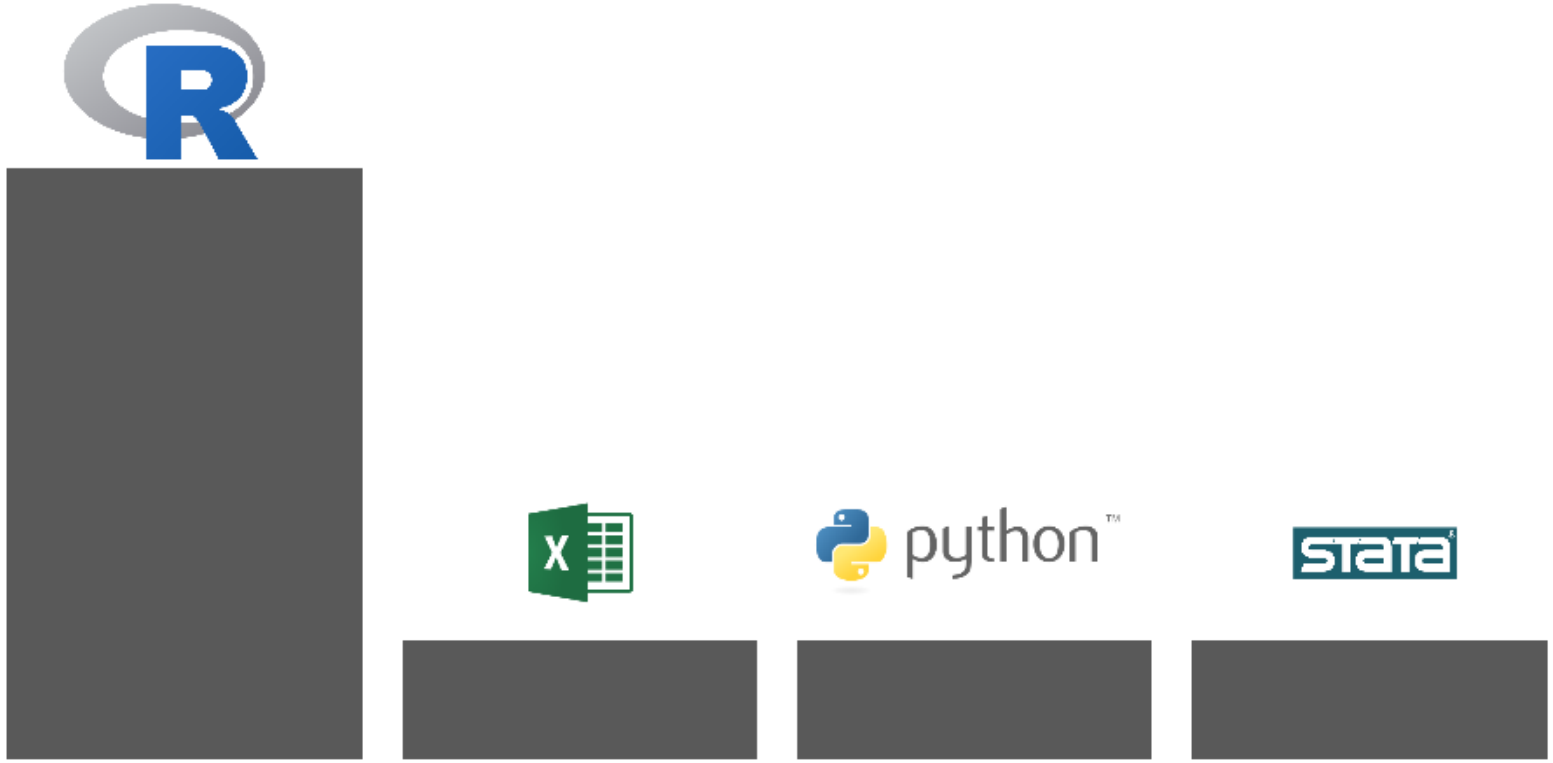
Warum Grafiken mit R?

- R ist open source
- R ist sehr flexibel und so können die verschiedensten Arten von Grafiken realisiert werden
- (Im Gegensatz zu Excel) Grafiken sind reproduzierbar
- (Im Gegensatz zu SPSS) Man ist nicht auf „Schaltflächen“ (vordefinierte Grafikypten) angewiesen, sondern ist deutlich flexibler
- Auch sehr komplexe Grafiken können erstellt werden
- Es gibt eine große, aktive Community, die bei Fragen hilft und neue Möglichkeiten entwickelt
- Grafiken und statistische Analysen (und auch beschreibender Text) kommen "aus einem Guss"
- An jeder Stelle in der Grafikerstellung kann R-Syntax verwendet werden
- Grafiken können präzise maßgeschneidert werden, z.B. um einem Corporate Design (Uni) oder allgemeinem Grafikthema (Tageszeitung, Lehrbuch) zu entsprechen
- Alles, was in R eingelesen werden kann, kann auch in eine Grafik gebracht werden
- Man kann sich kreativ austoben ☺

Warum Grafiken mit R?

DATA CHALLENGE
- ERGEBNISSE -

- digiGEBF Data Challenge
  https://www.digigebf21.de/frontend/index.php?page_id=17723



Warum dieser Kurs?

- Grafiken: Mächtiges Instrument der Datenanalyse und Ergebniskommunikation
- Es gibt Fallstricke und Dinge, die sich lohnen zu wissen
- Grafikerstellung ist Teil aller datenbasierten, wissenschaftlichen Disziplinen…
- …aber wird kaum bis gar nicht im Studium behandelt

Ablauf 8./9. Juli 2021

Tag	Block	Thema	Asynchrone Inhalte auf pandaR
Donnerstag	1	Intro	
	2	ggplot-Intro	https://pandar.netlify.app/post/ggplotting-intro/
	3	Hübsche Grafiken: Theorie	
	4	Hübsche Grafiken: Praxis	https://pandar.netlify.app/post/ggplotting-themes/
	5	Übung 1	
	6	ggplotpourri	https://pandar.netlify.app/post/ggplotting-ggplotpourri/
Freitag	7	gganimate	https://pandar.netlify.app/post/ggplotting-gganimate/
	8	plotly	https://pandar.netlify.app/post/ggplotting-plotly/
	9	Explorative Grafiken	https://pandar.netlify.app/post/ggplotting-exploration/
	10	Übung 2	
	11	Outro	

Material

- https://pandar.netlify.app/extras/#ggplotting
  - Folien
  - R-Skripte
  - Inhalte zur asynchronen (Nach-) Bearbeitung

- Datensatz:
  - Download:
    `load(url('https://pandar.netlify.com/post/edu_exp.rda'))`
  - Hinweise zur Datenaufbereitung:
    https://pandar.netlify.app/post/ggplotting-daten/

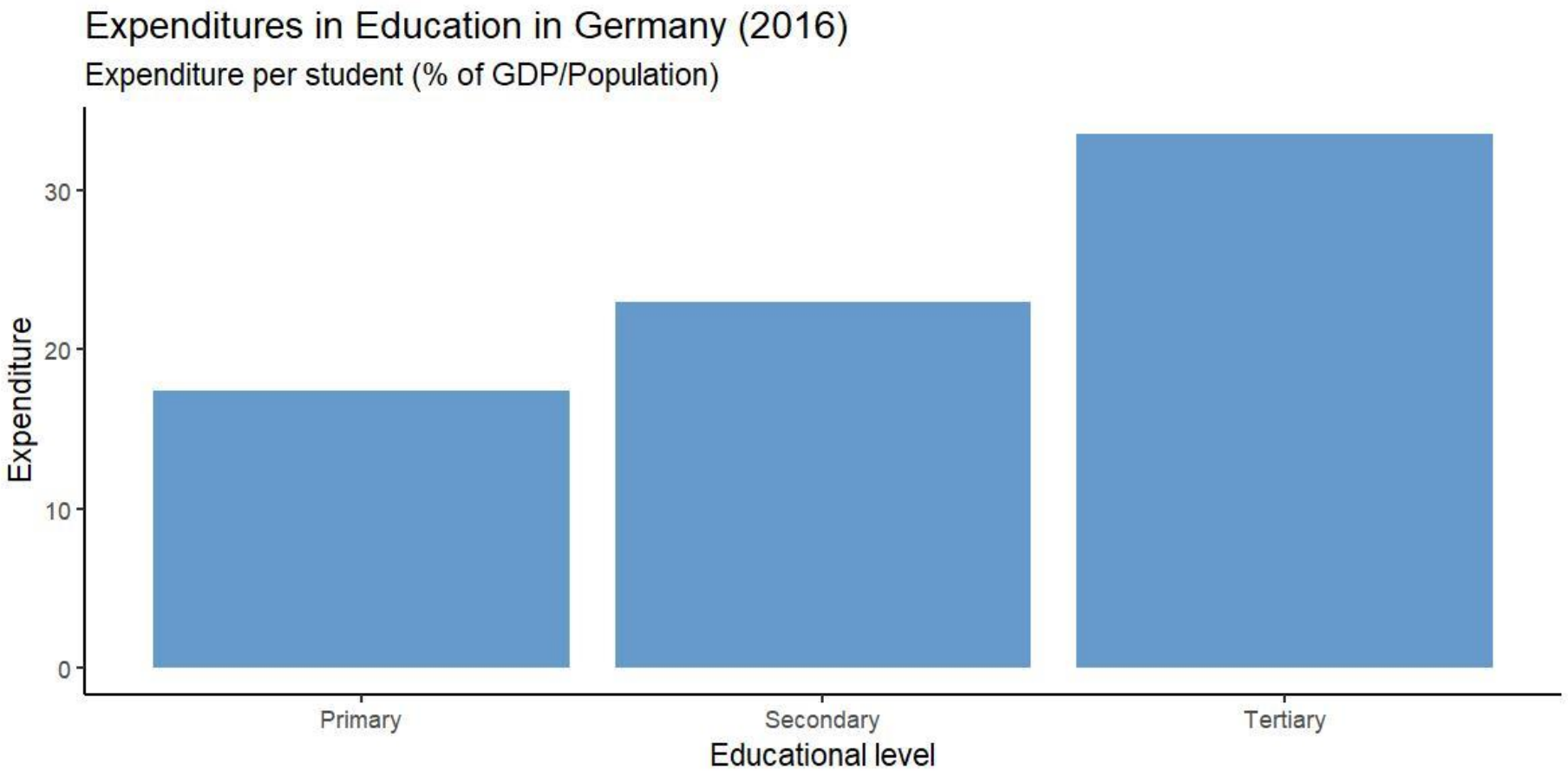
- Fragenspeicher:
  ahaslides.com/ggplotting

Datensatz `edu_exp`

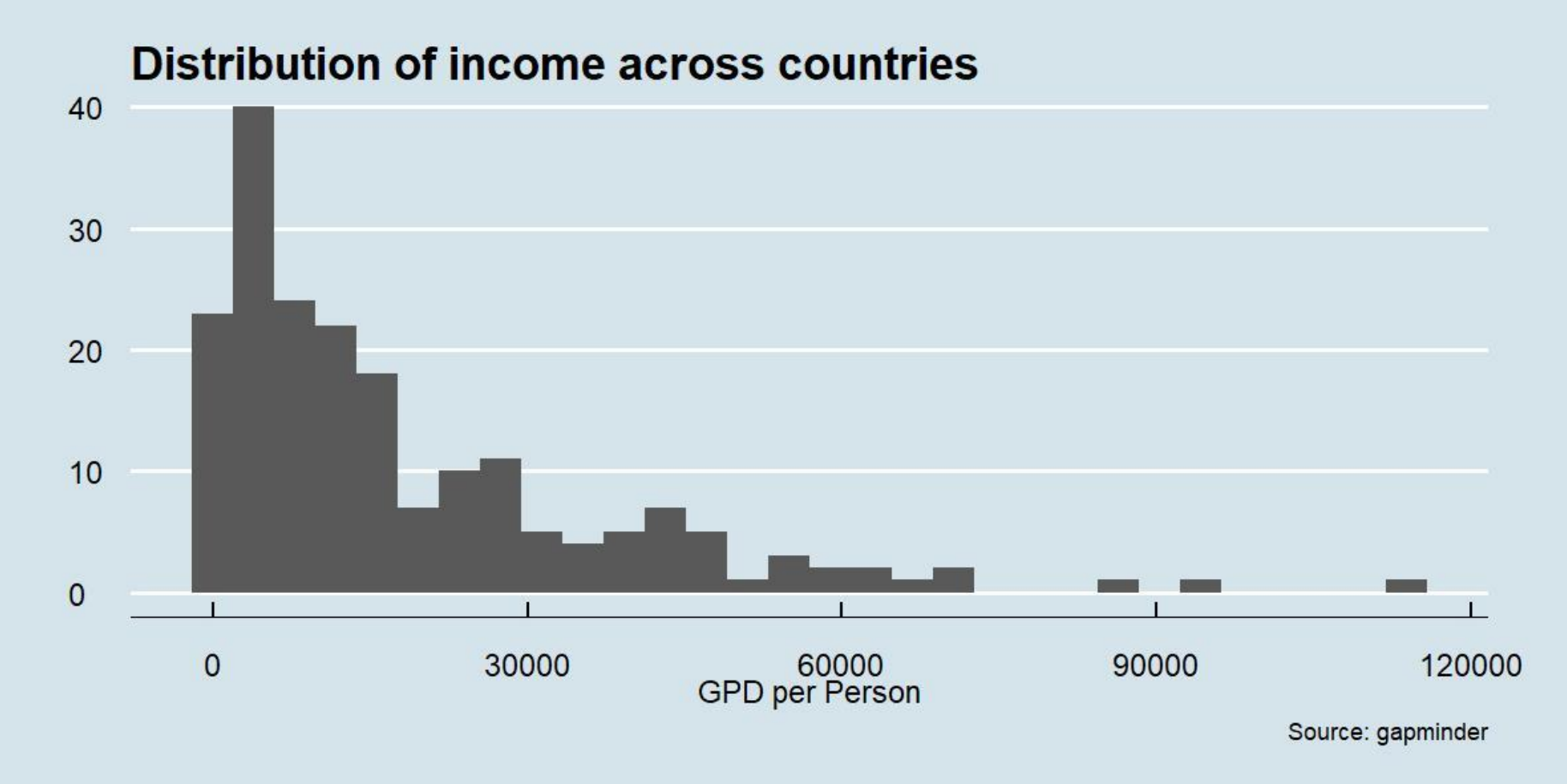
Variable	Bedeutung
geo	Länderkürzel, das zur Identifikation der Länder über verschiedene Datenquellen hinweg genutzt wird
Country	der Ländername im Englischen
Wealth	Wohlstandseinschätzung des Landes, unterteilt in fünf Gruppen
Region	Einteilung der Länder in die vier groben Regionen africa, americas, asia und europe
Year	Jahreszahl
Population	Bevölkerung
Expectancy	Lebenserwartung eines Neugeborenen, sollten die Lebensumstände stabil bleiben.
Income	Stetiger Wohlstandsindikator für das Land (GDP pro Person)
Primary	Staatliche Ausgaben pro Schüler*in in der primären Bildung als Prozent des income (GDP pro Person)
Secondary	Staatliche Ausgaben pro Schüler*in in der sekundären Bildung als Prozent des income (GDP pro Person)
Tertiary	Staatliche Ausgaben pro Schüler*in oder Student*in in der tertiären Bildung als Prozent des income (GDP pro Person)
Index	Education Index des United Nations Development Programme

GRAFIKBEISPIELE

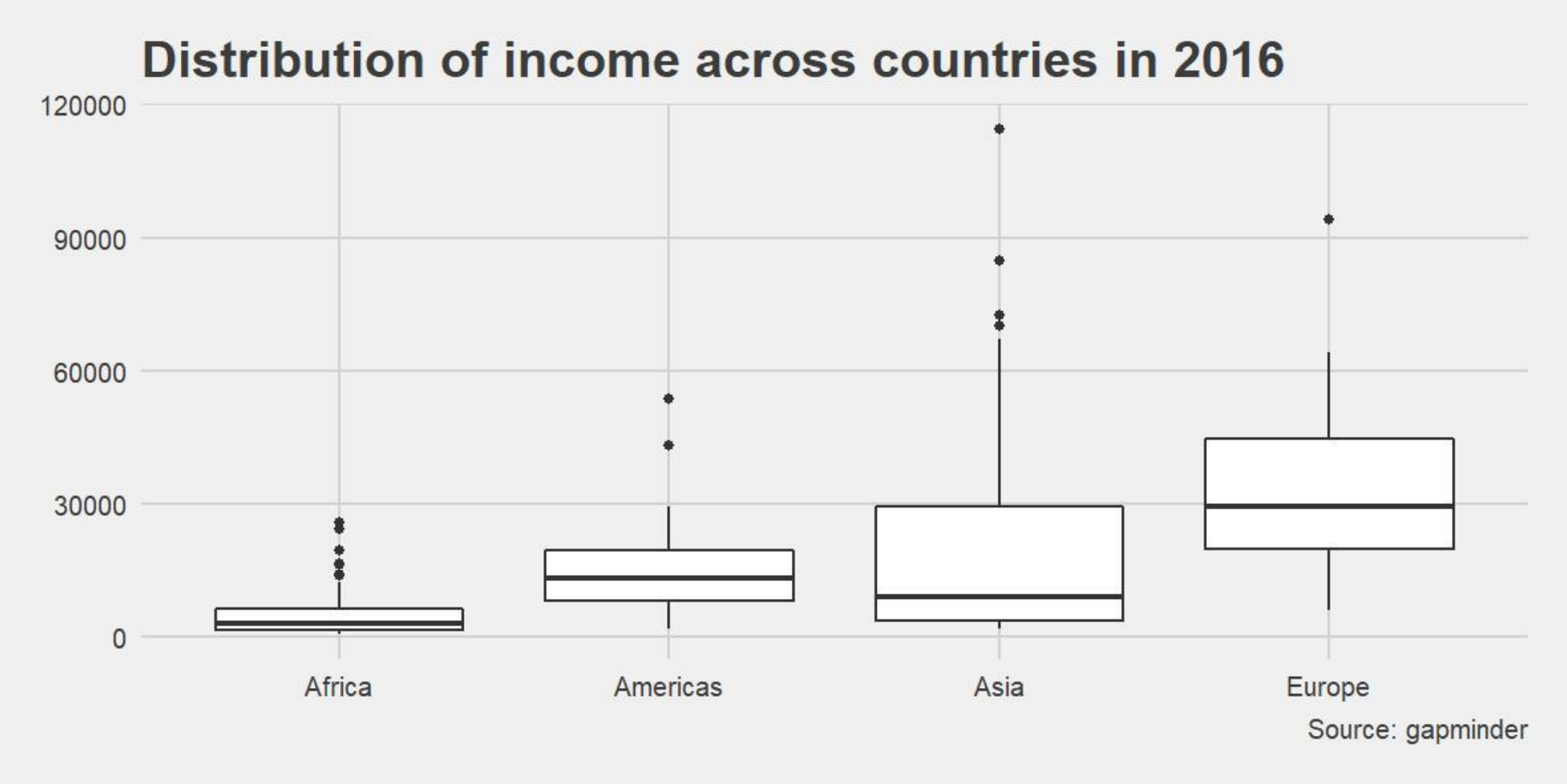
Grafikbeispiele: Balkendiagramm



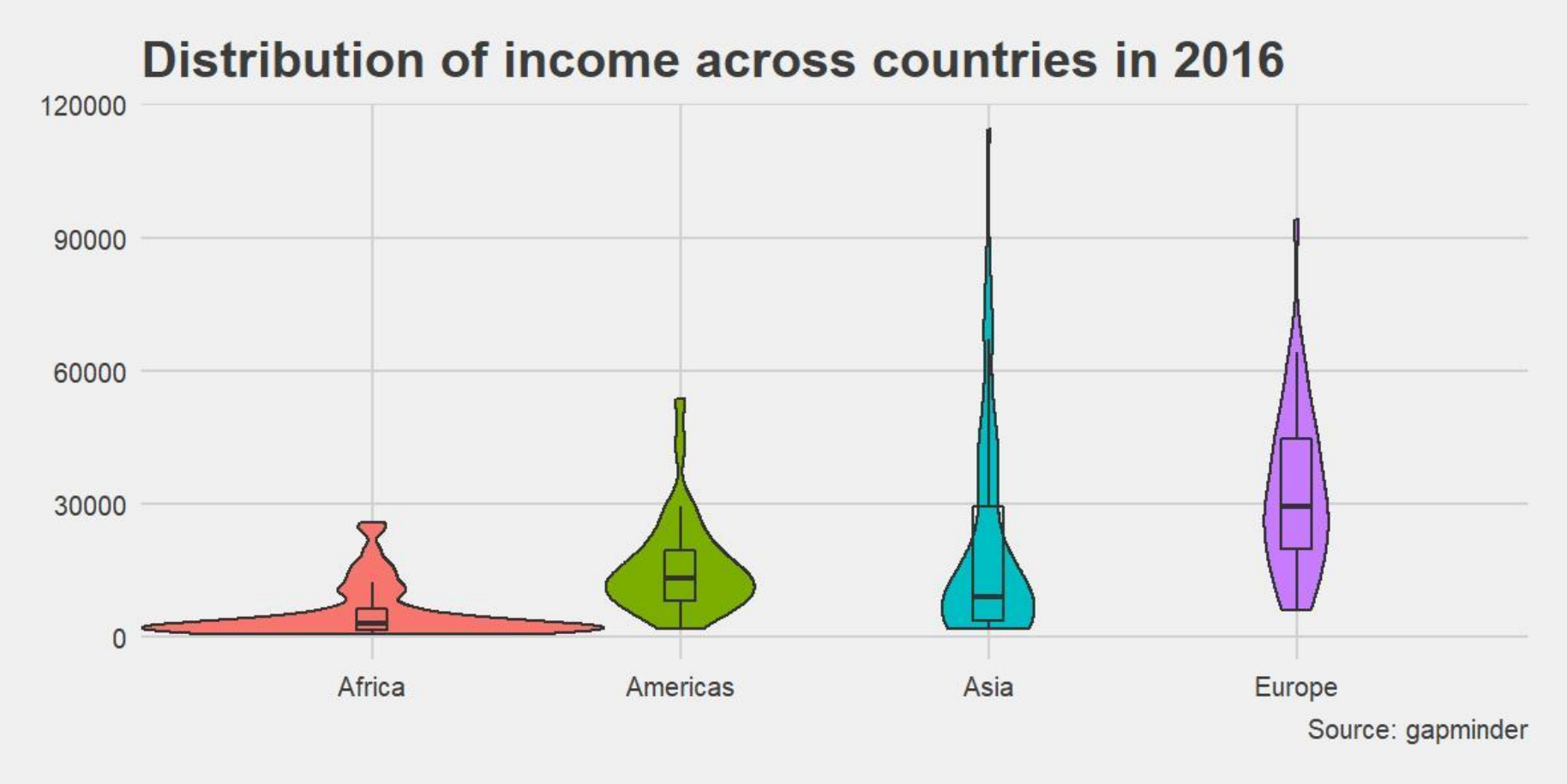
Grafikbeispiele: Histogramm



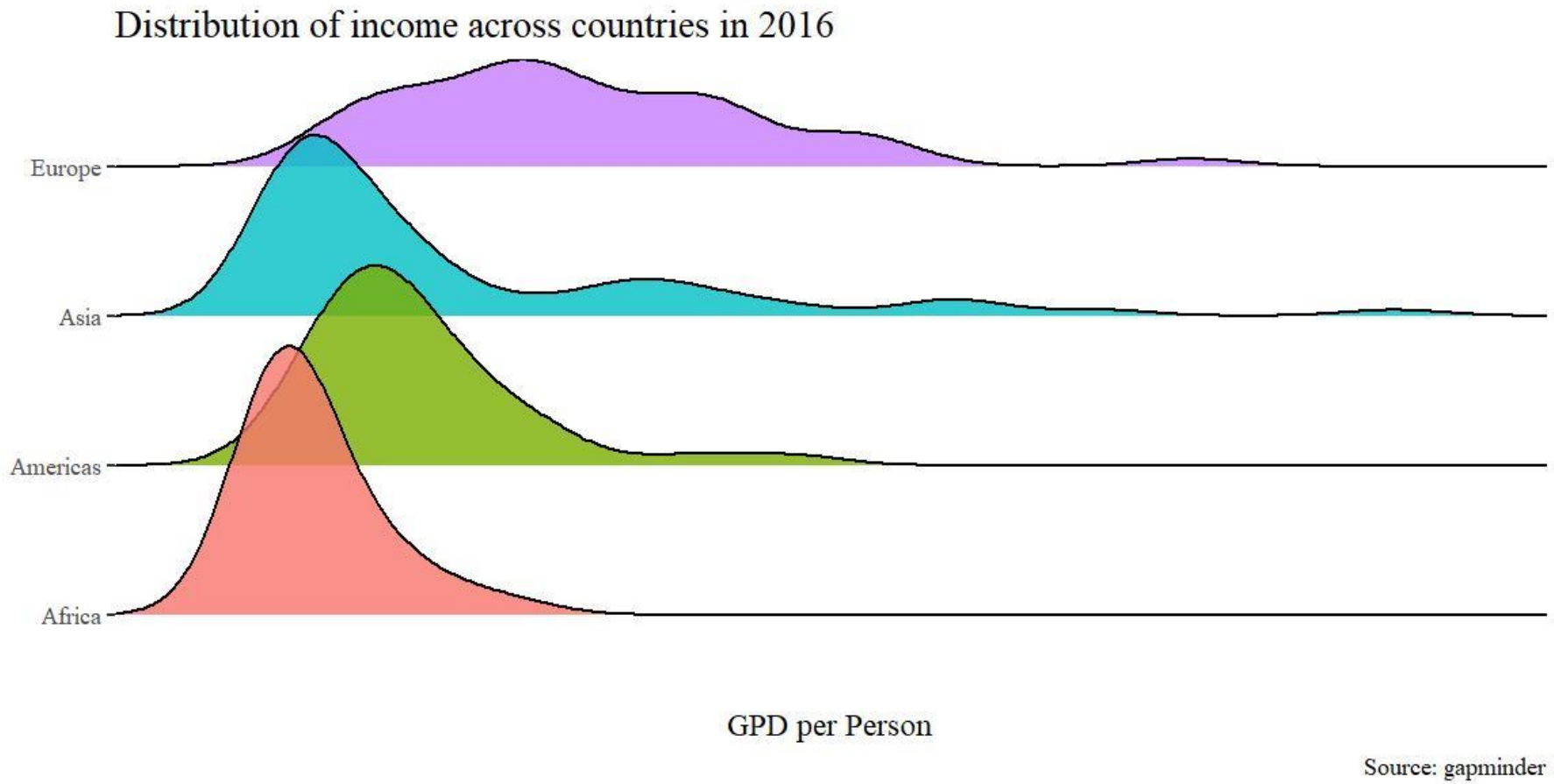
Grafikbeispiele: Boxplot



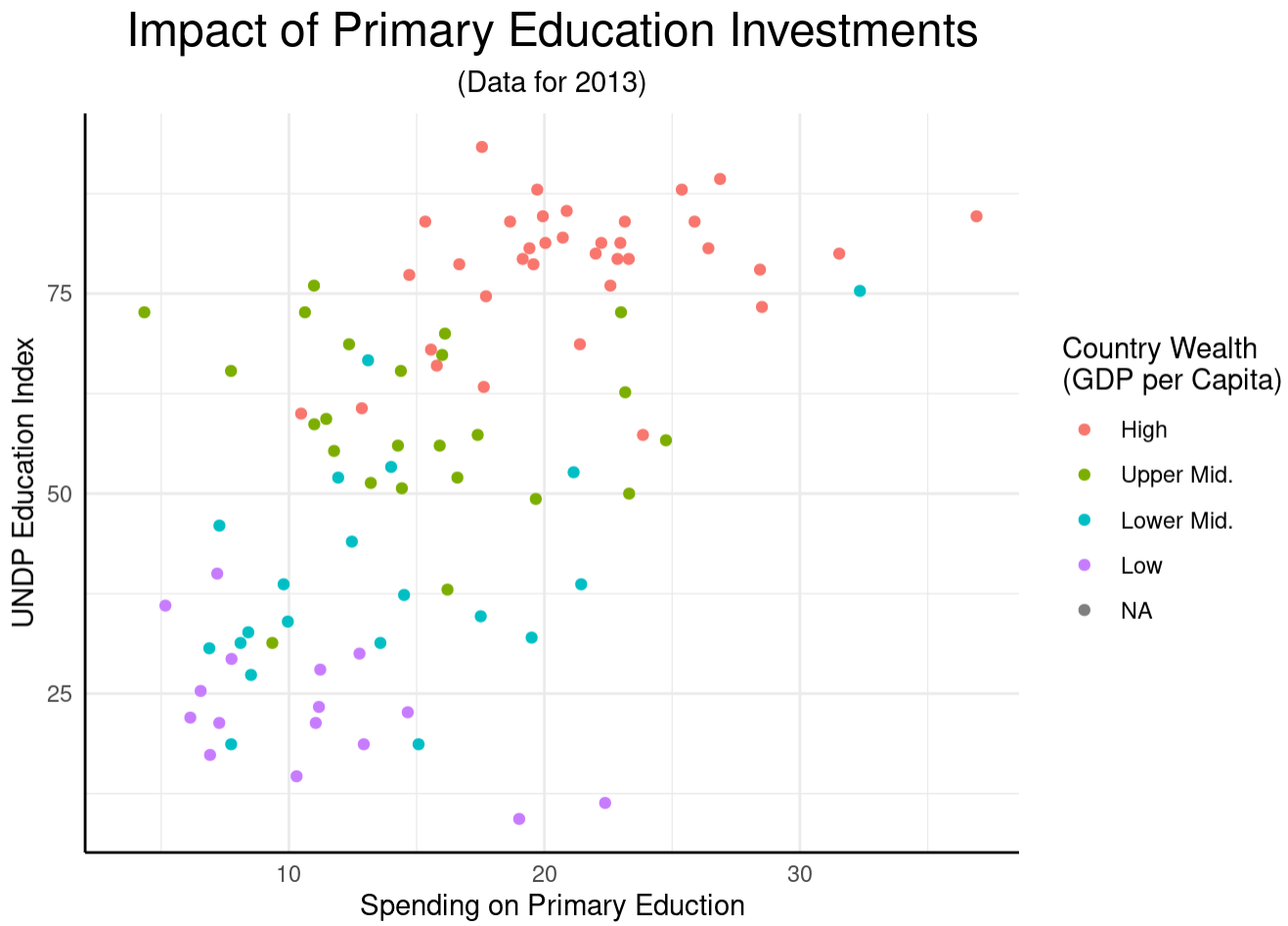
Grafikbeispiele: Pirate (Violin) Plot



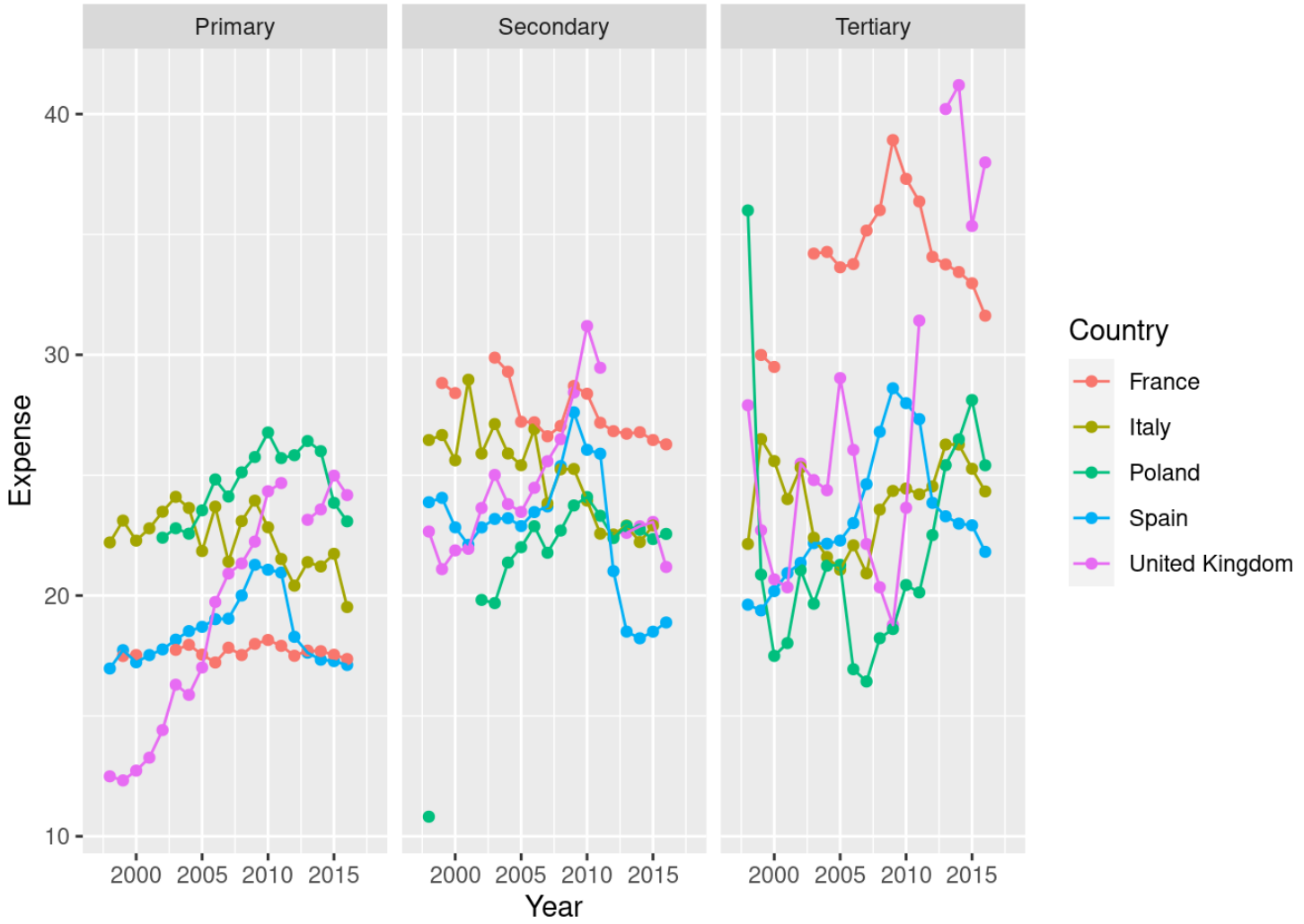
Grafikbeispiele: Ridge Line



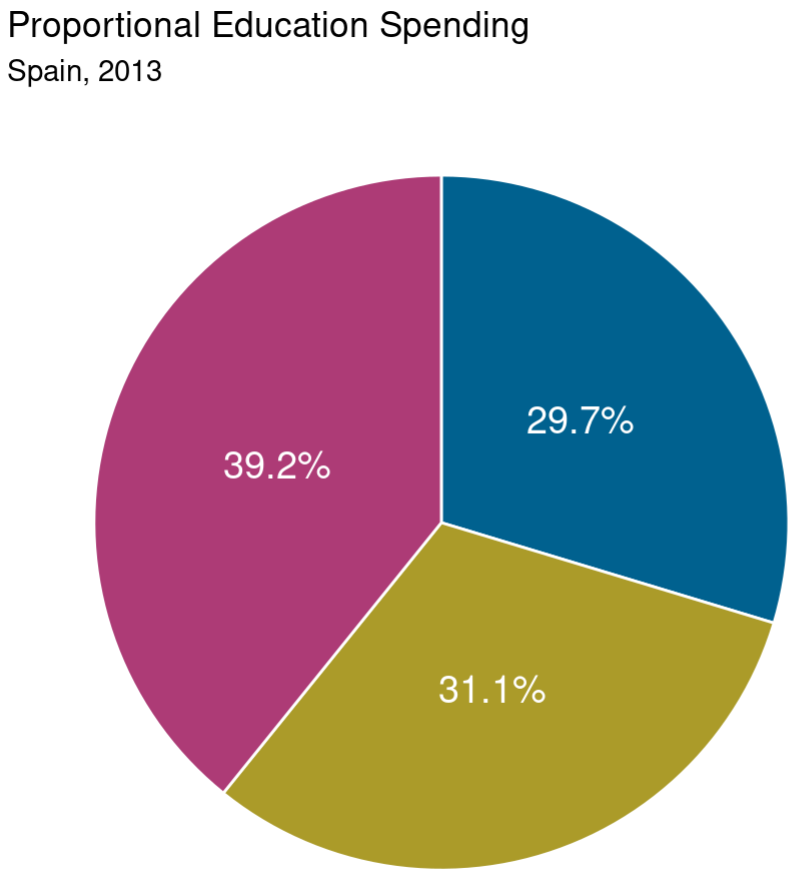
Grafikbeispiele: Scatterplot

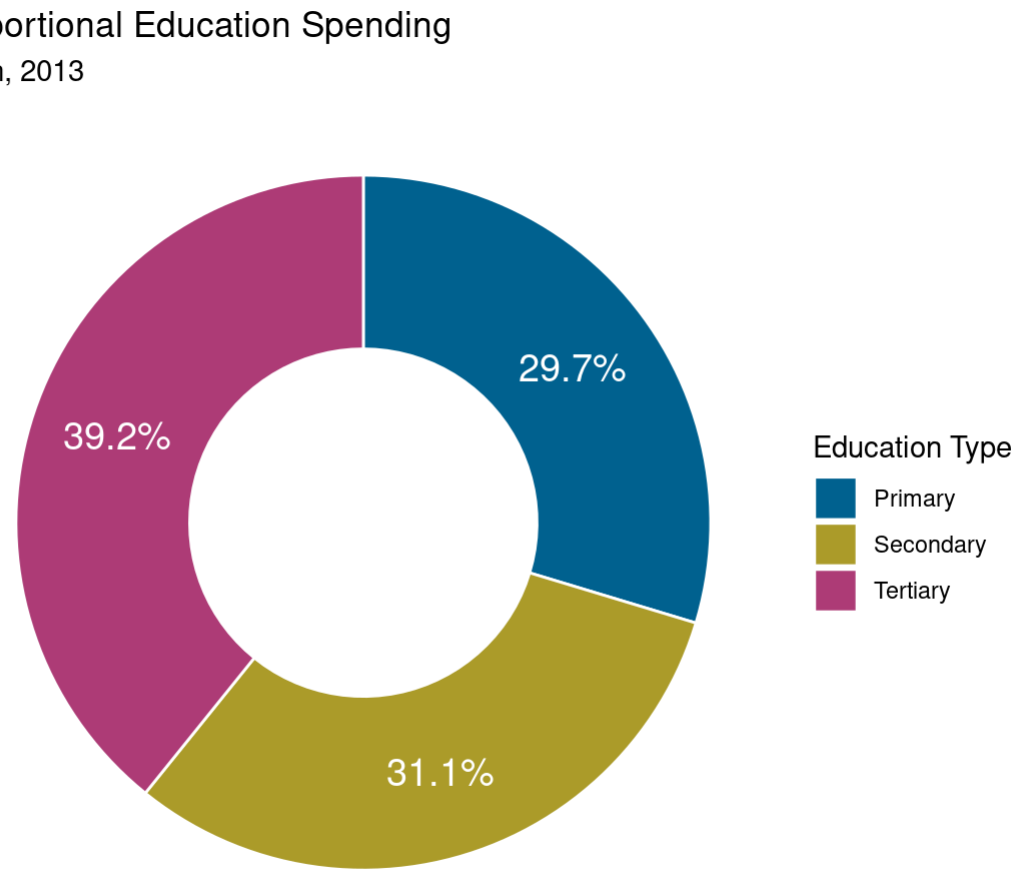


Grafikbeispiele: Liniendiagramm (Zeitreihe)

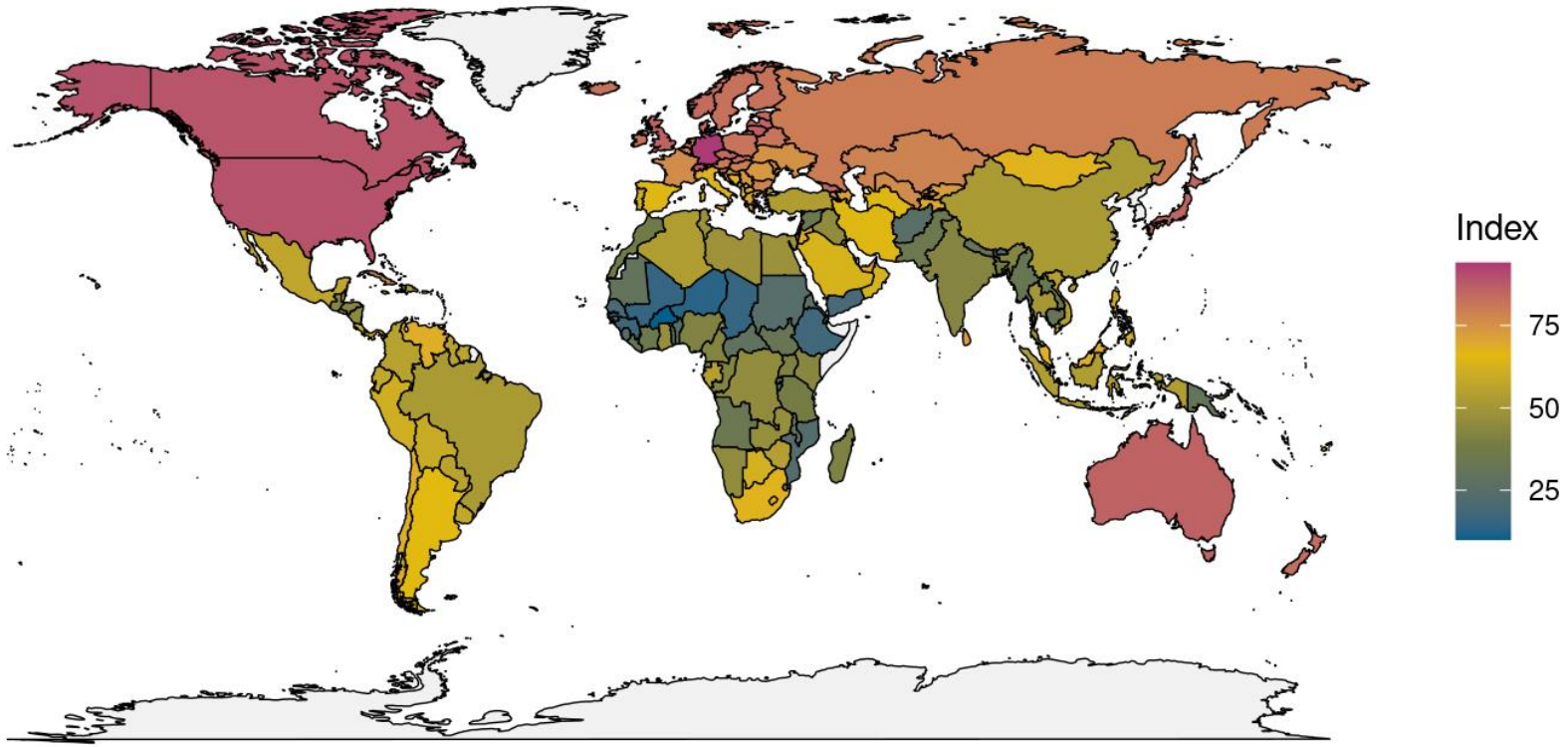


Grafikbeispiele: Pie und Donut Chart

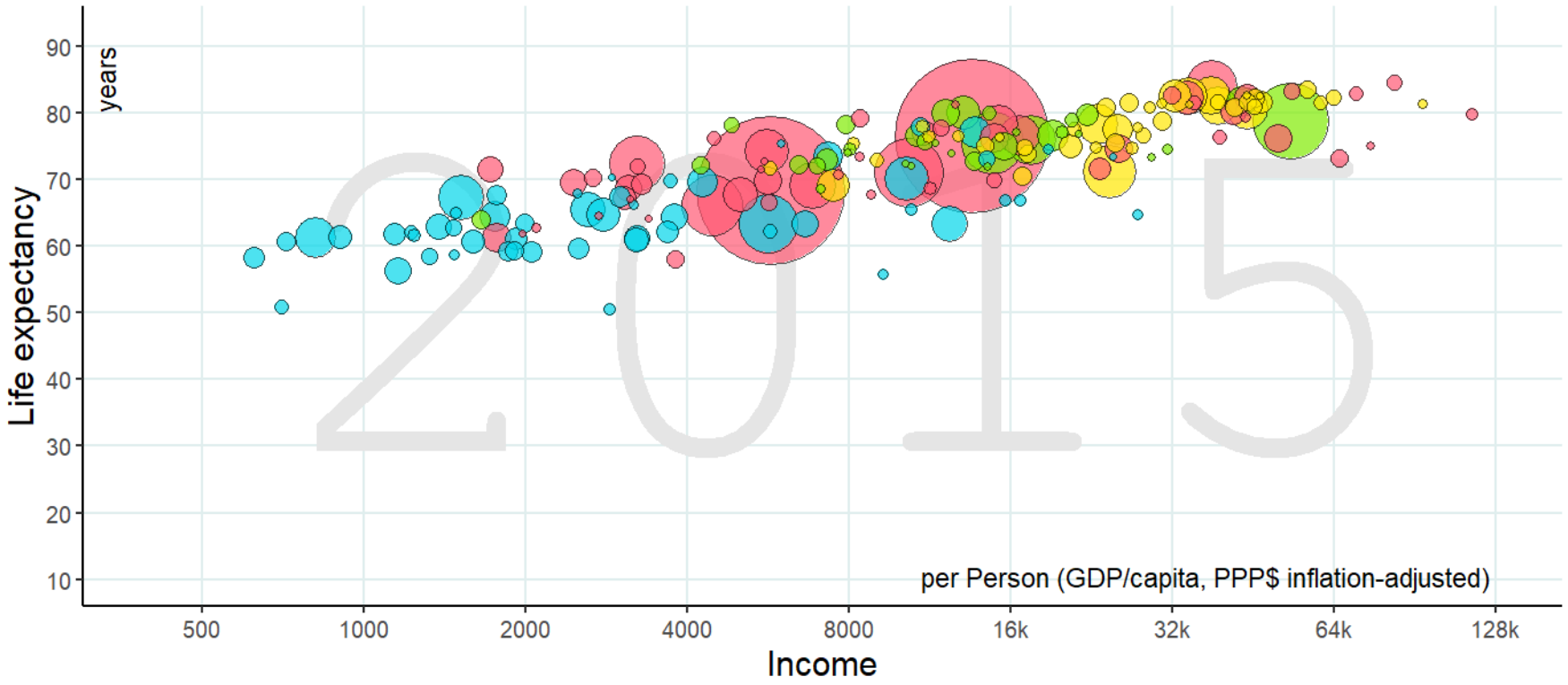




Grafikbeispiele: Map

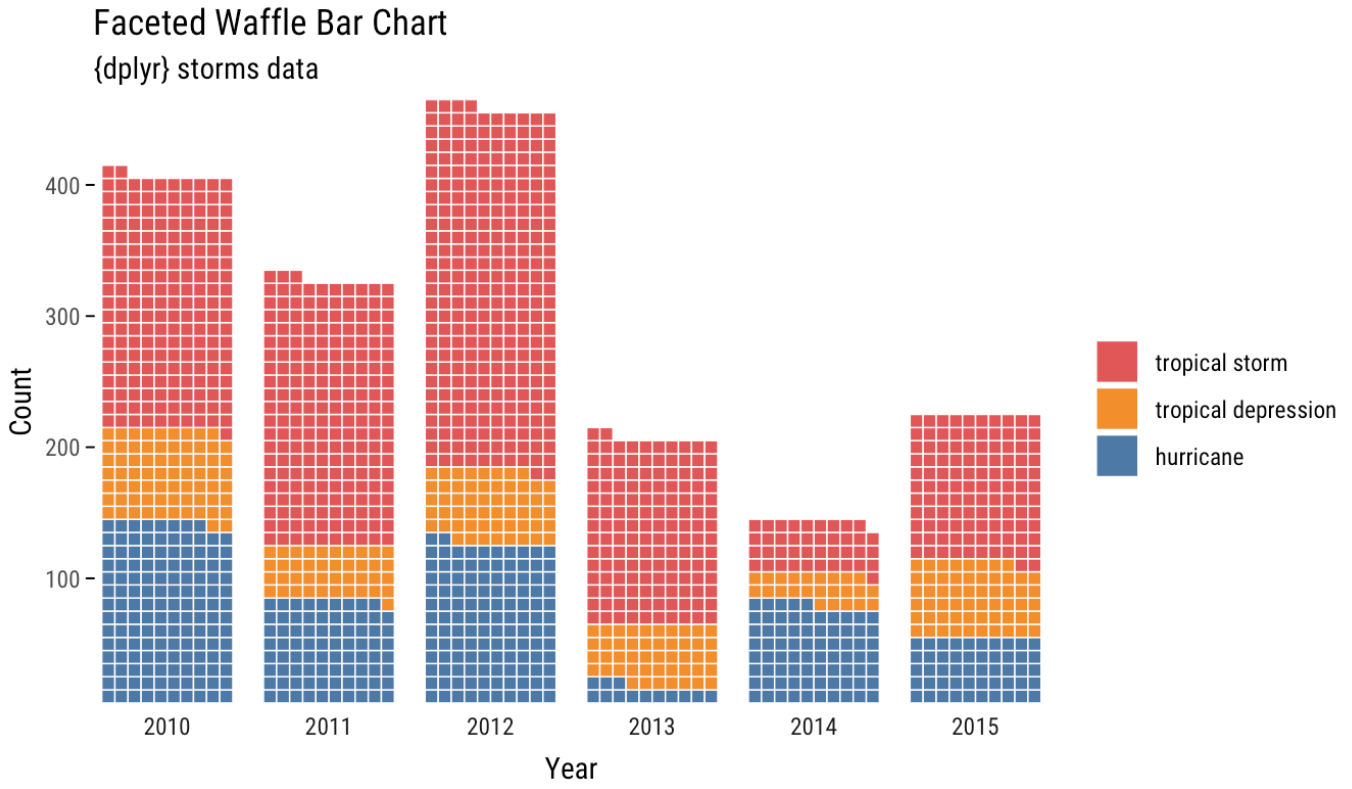


Grafikbeispiele: Bubble Chart



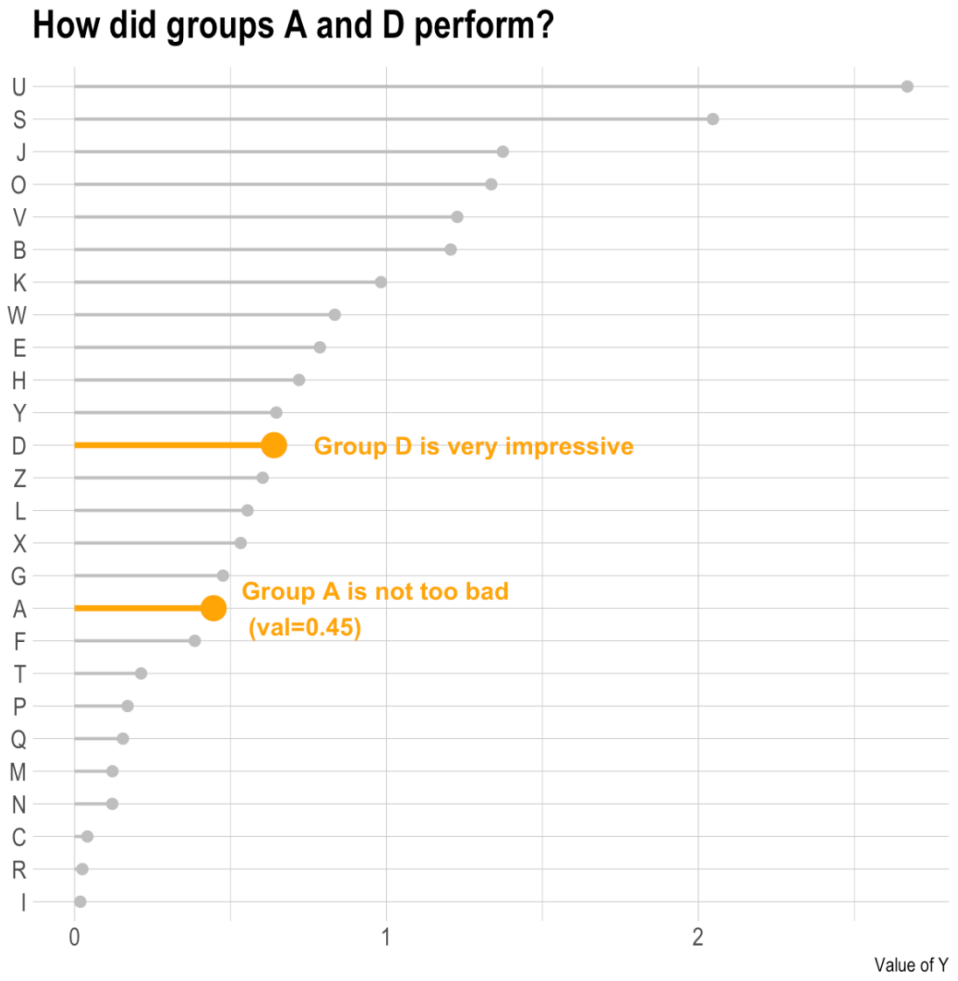
Grafikbeispiele: Wordcloud

Grafikbeispiele: Waffle Plot



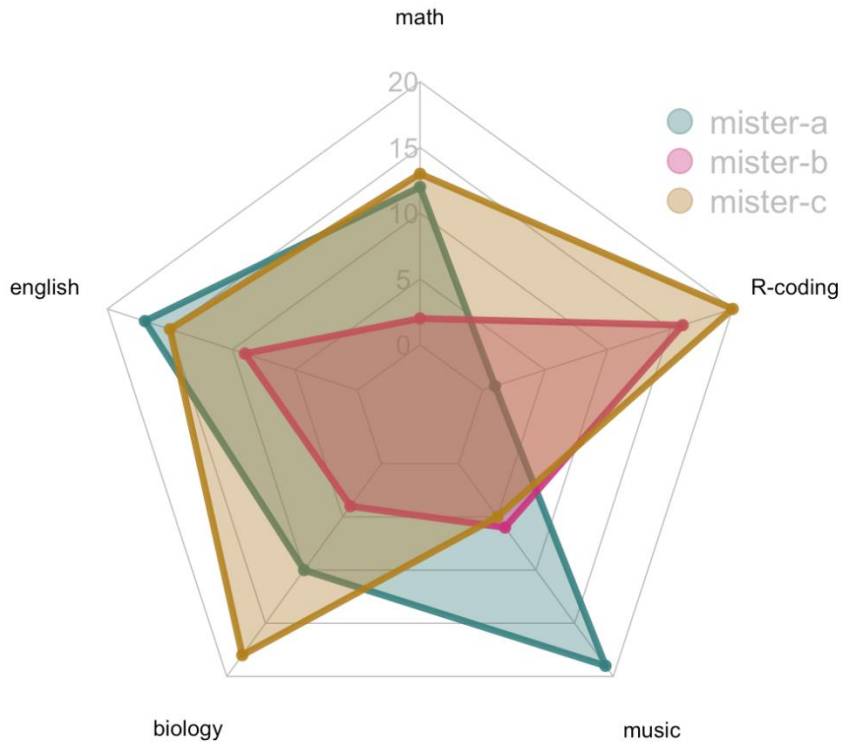
https://github.com/hrbrmstr/waffle

Grafikbeispiele: Lollipop Chart



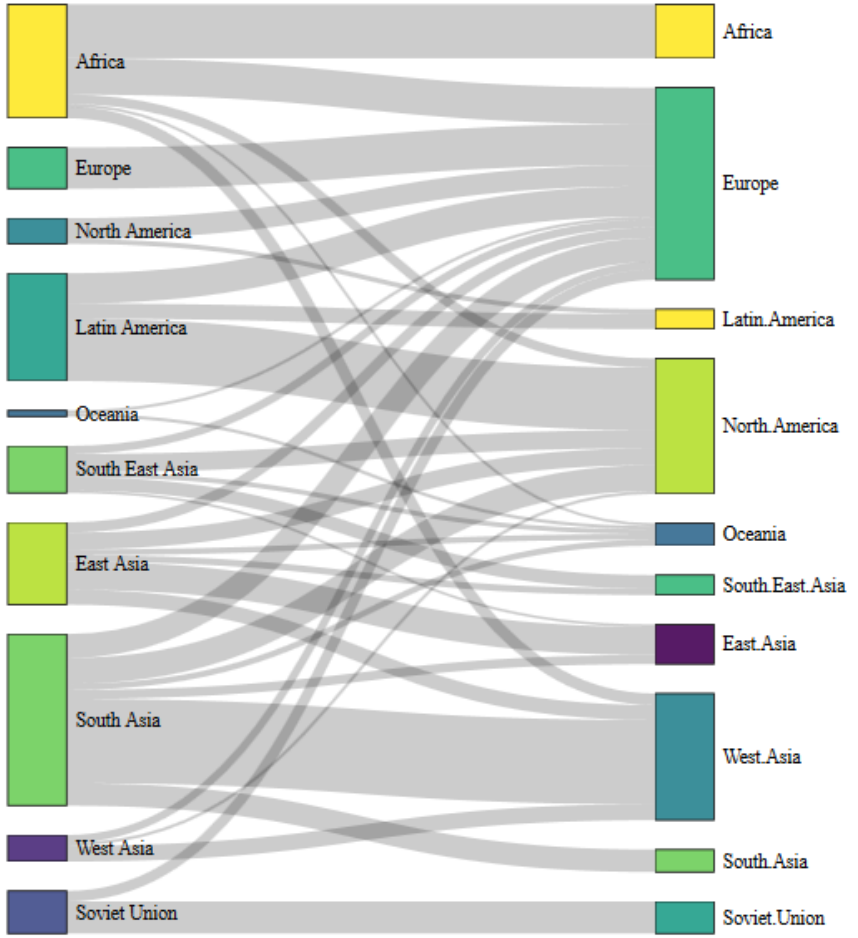
https://www.r-graph-gallery.com/304-highlight-a-group-in-lollipop.html

Grafikbeispiele: Radar Chart



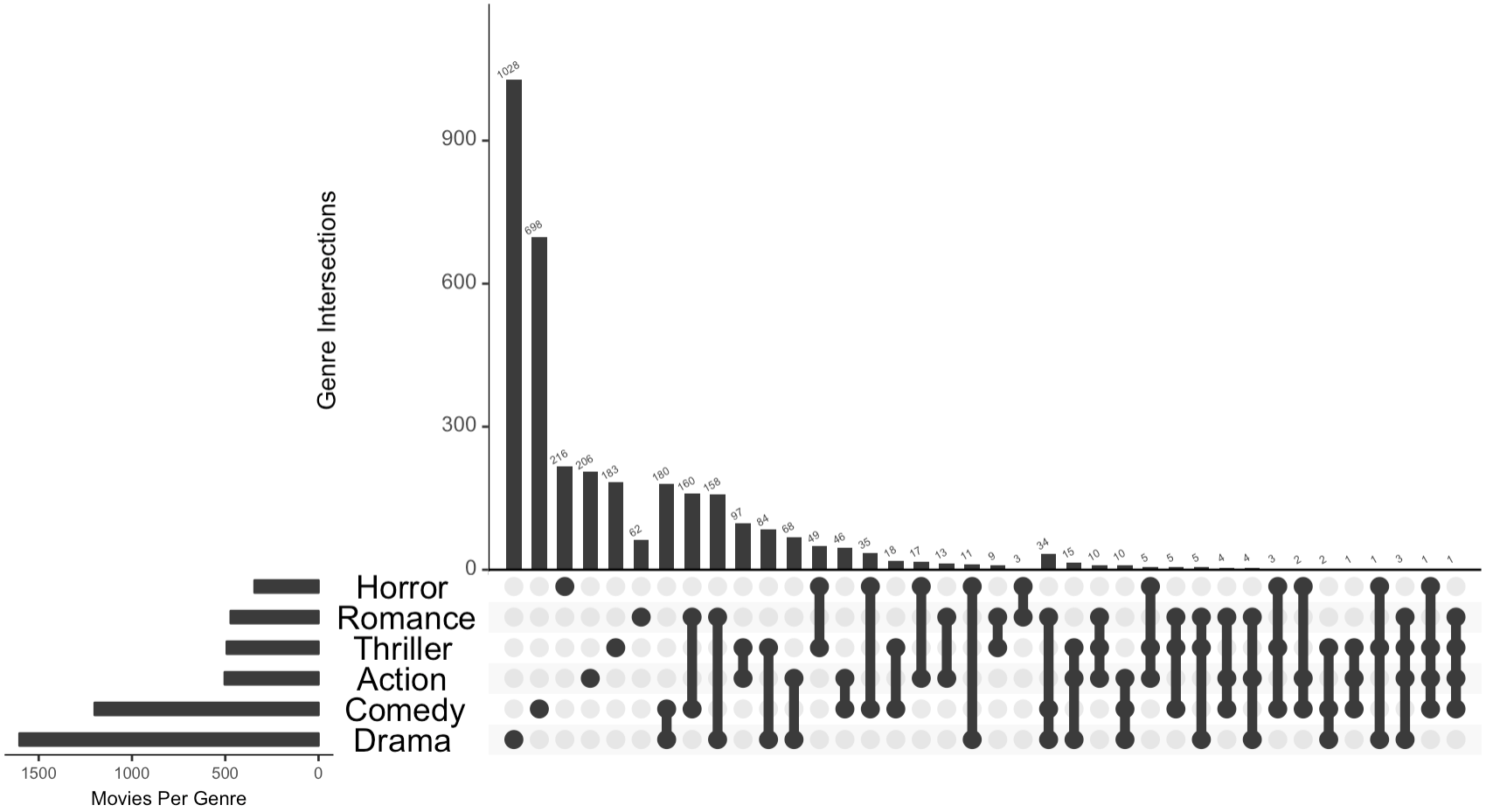
https://www.r-graph-gallery.com/143-spider-chart-with-saveral-individuals_files/figure-html/thecode2-1.png

Grafikbeispiele: Sankey Diagram



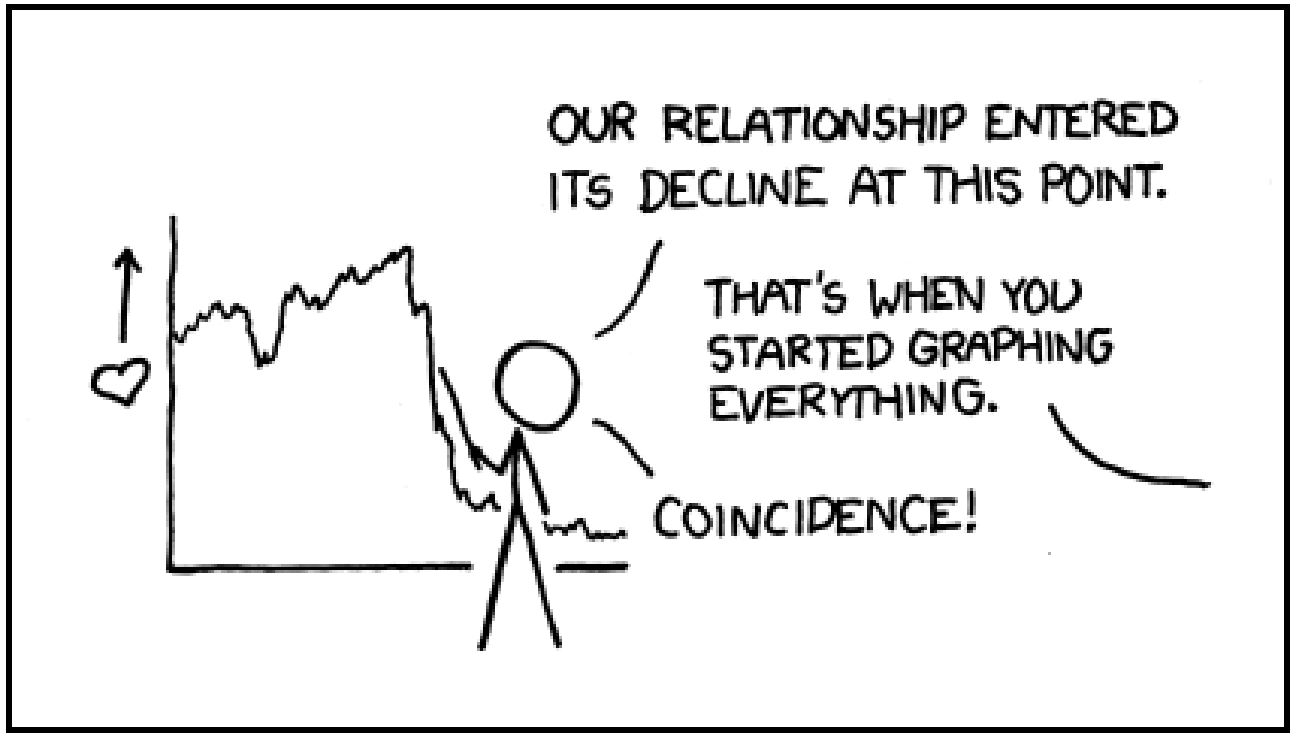
https://www.data-to-viz.com/graph/sankey.html

Grafikbeispiele: Upset Plot



https://cran.r-project.org/web/packages/UpSetR/vignettes/basic.usage.html

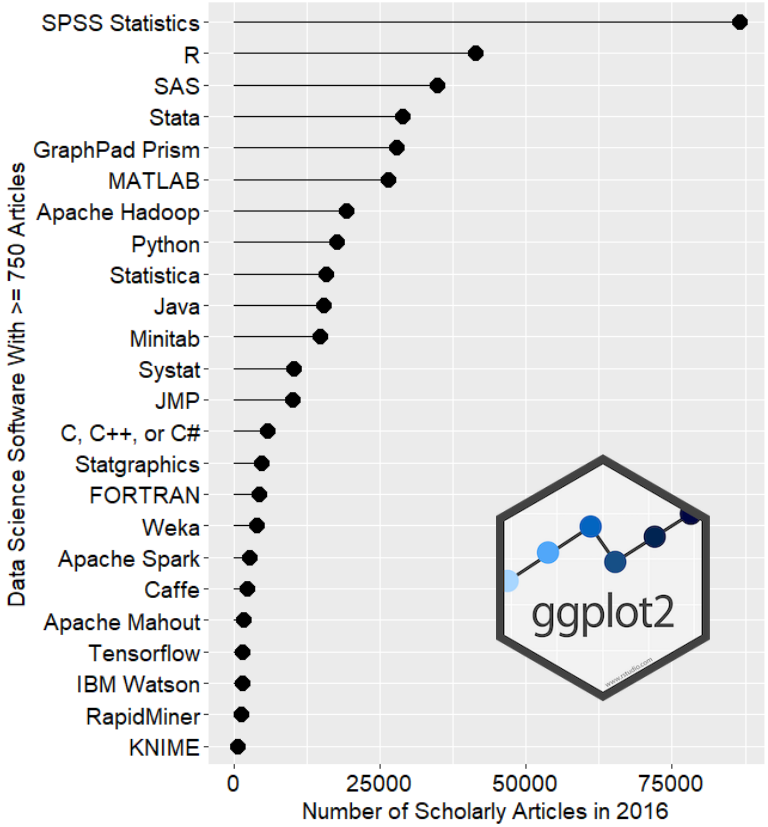
Grafikbeispiele: Liniendiagramm

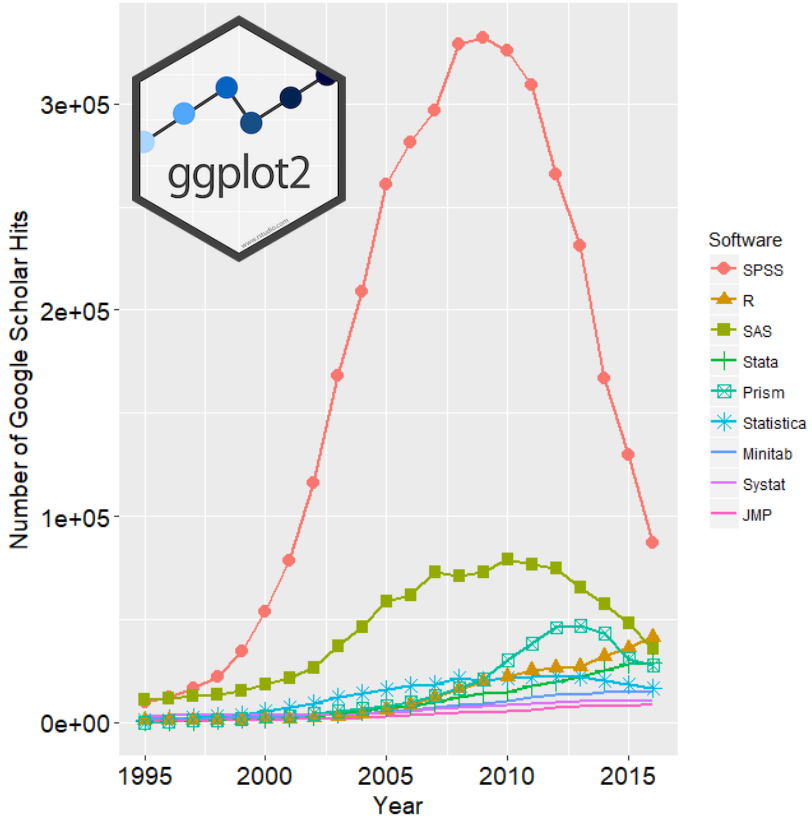


https://xkcd.com/523/

BLOCK 2:
EINFÜHRUNG IN DAS R-PAKET GGPLOT2

Grafiken mit ggplot2? – Aber warum?

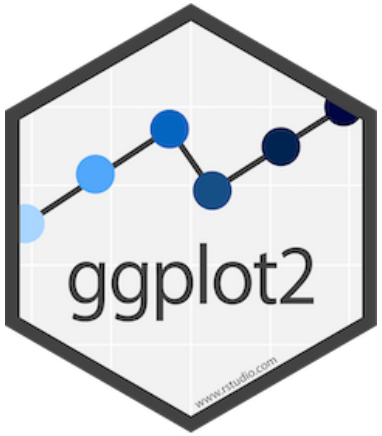




http://r4stats.com/articles/popularity/

DAS PAKET GGPLOT2

Das Paket ggplot2



- R-Paket für Datenvisualisierung aus dem tidyverse-"Universum"
- Entwickelt von Hadley Wickham seit 2005
- aktuelle Version: 3.3.5 (Stand: 7.7.2021)
- Eins der beliebtesten R-Pakete überhaupt:
  >55 Millionen Downloads seit 2005! (Stand: 7.7.2021)

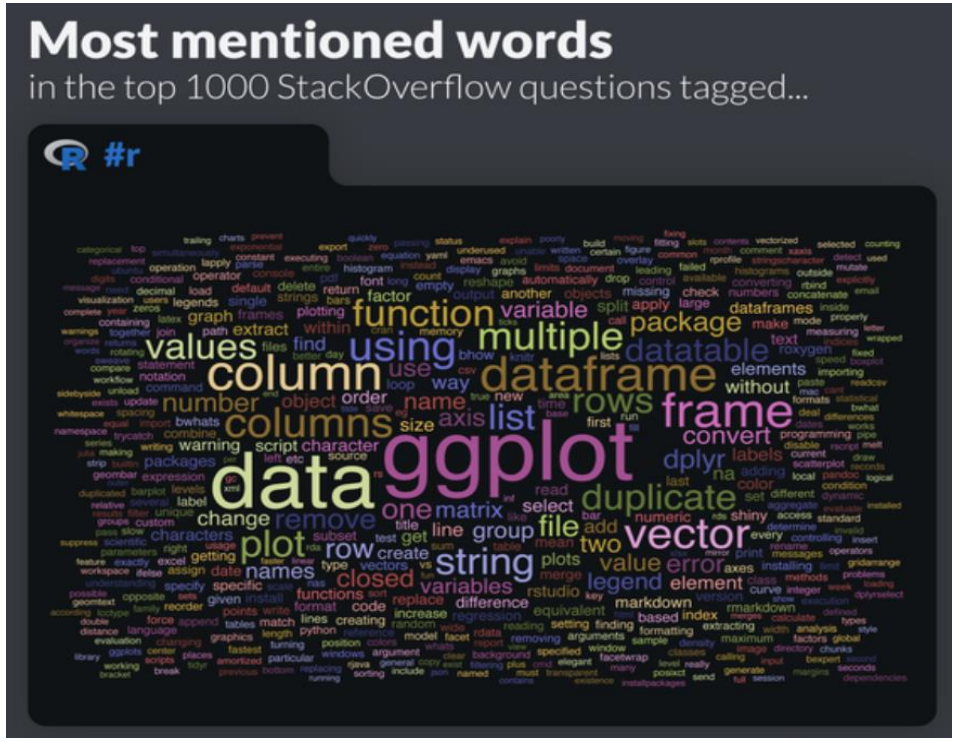
```
> library(cranlogs)
> sum(cran_downloads("ggplot2", from = "2005-01-01", to = "2021-07-07")$count)
[1] 55381775
```

Ressourcen

- Offizielle Seite: https://ggplot2.tidyverse.org/
- Vollständige Darstellung aller Funktionen: https://ggplot2.tidyverse.org/reference/
- Cheat Sheet: https://github.com/rstudio/cheatsheets/blob/master/data-visualization-2.1.pdf
- Beispiele (inkl. Syntax):
  - The R Graph Gallery: https://www.r-graph-gallery.com/
  - Top 50 ggplot2 Visualizations: 50 http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html
- Hilfe:
  - RStudio Forum: https://community.rstudio.com/
  - StackOverflow: https://stackoverflow.com/questions/tagged/ggplot2

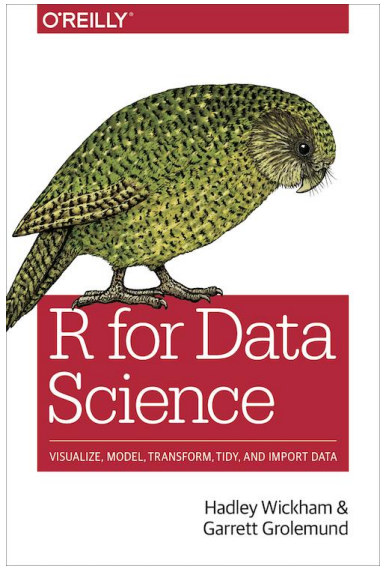
Ressourcen

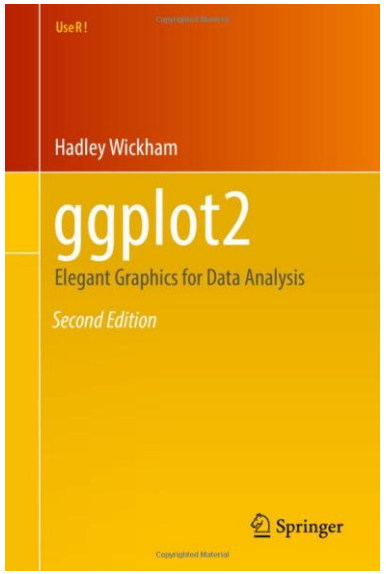
- Stichwort StackOverflow:
  - https://www.reddit.com/r/rprogramming/comments/av2e8b/most_frequently_mentioned_words_in_the_top_1000/

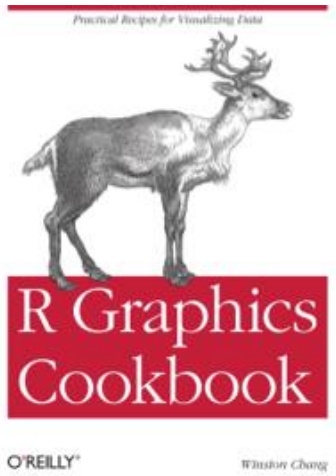


Ressourcen

- Theoretische Grundlage/die Logik hinter ggplot2:
  https://vita.had.co.nz/papers/layered-grammar.html
- Online-Lehrbuch „R for Data Science“:
  https://r4ds.had.co.nz/
  - Kap. 3, Einführung in ggplot2:
    https://r4ds.had.co.nz/data-visualisation.html
  - Kap. 28, Feinheiten, ggplot-Werkzeuge für „gute“ Grafiken:
    https://r4ds.had.co.nz/graphics-for-communication.html
- Online-Buch „ggplot2: elegant graphics for data analysis“:
  https://ggplot2-book.org/
- The R Graphics Cookbook (verwendet ggplot2):
  http://www.cookbook-r.com/Graphs/







Die Grundidee

Hadley Wickham's

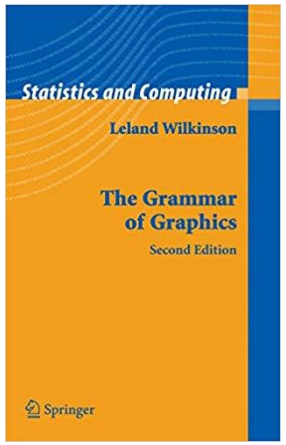
A layered grammar of graphics

Hadley Wickham.
A layered grammar of graphics.
Journal of Computational and Graphical Statistics, vol. 19, no. 1, pp. 3–28, 2010.
Download: pre-print | via doi

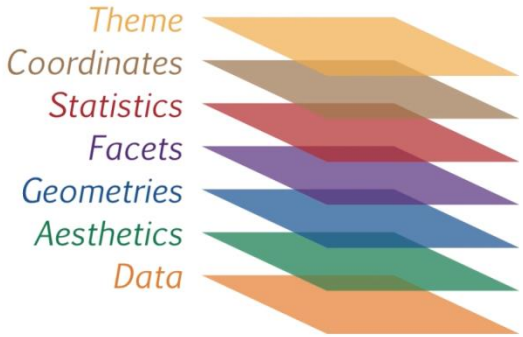
A grammar of graphics is a tool that enables us to concisely describe the components of a graphic. Such a grammar allows us to move beyond named graphics (e.g., the ``scatterplot'') and gain insight into the deep structure that underlies statistical graphics. This paper builds on Wilkinson (2006), describing extensions and refinements developed while building an open source implementation of the grammar of graphics for R, ggplot2.

https://vita.had.co.nz/papers/layered-grammar.html

Die Grundidee

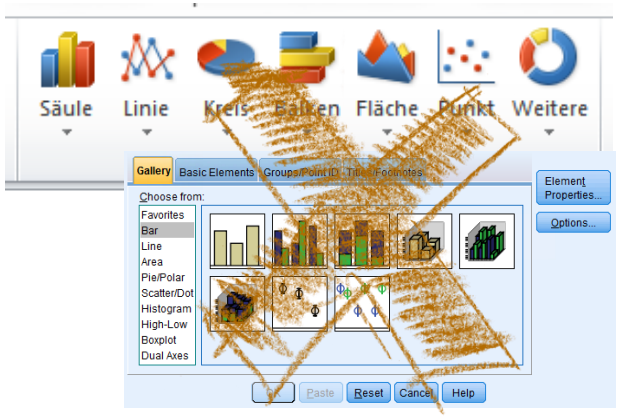


- *builds on Wilkinson (2006)*
  - beruht auf Leland Wilkinson‘s Idee der „***g**rammar of **g**raphics*“ (→ **gg**plot)



- *concisely describe the components of a graphic*
  - eine Grafik kann zerlegt oder beschrieben werden durch ein Set von Bestandteilen

- *move beyond named graphics (e.g., the "scatterplot")*
  - Eine Grafik ist nicht schon vorab festgelegt (wie z.B. in Excel, SPSS), sondern wird anhand von *layer*s „Schicht für Schicht“ nach bestimmten Regeln (wie Wörter in einem Satz) zusammengesetzt
  - Der Prozess der Grafikerstellung ist somit extrem flexibel



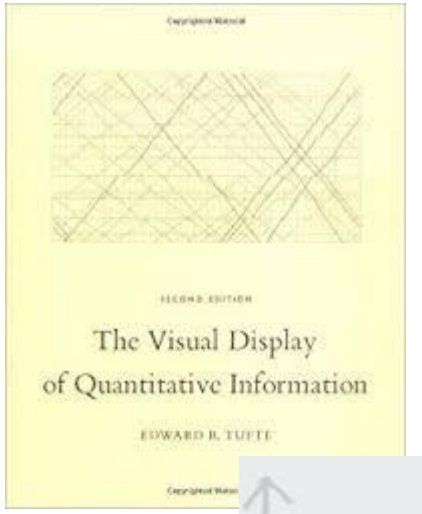
BLOCK 3:
SCHÖNE GRAFIKEN: THEORIE

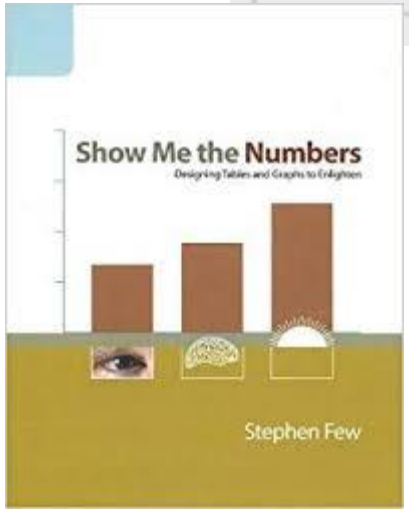
Schöne Grafiken: Theorie

- Ressourcen
- Ziele von Grafiken
- Schlechte Grafiken
- Gute Grafiken (Prinzipien von Tufte)
- Farben und Formen
- Grafiken im wissenschaftlichen Kontext (DGPs/APA)

Ressourcen

- Edward R Tufte (2001): The visual display of quantitative information
  - Papst und Vorreiter der Datenvisualisierung
  - Teilweise vielleicht etwas dogmatisch und angestaubt
- Scott Berinato (2016): Good charts
  - Verkäufer, der beibringt, wie man mit Grafiken Personen überzeugt
  - Sehr wirtschaftsorientiert und anwendungsbezogen
- Stephen C. Few (2012): Show me the numbers
  - Umfassender
  - Behandelt auch Tabellen

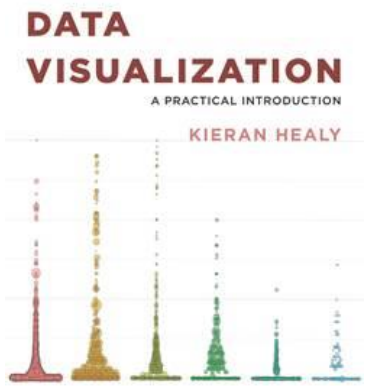




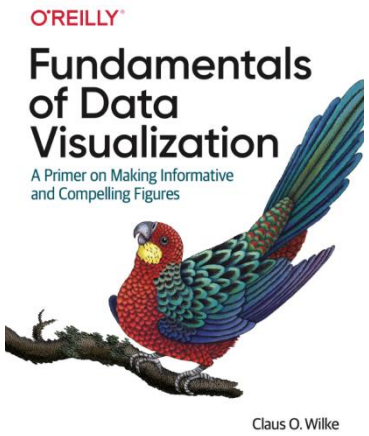
Ressourcen

Frei verfügbare Online-Bücher mit R-Syntax

- Kieran Healy (2018): Data visualization: A practical introduction
  - http://socviz.co/index.html
  - Sehr ausführlich und differenziert, teilweise eher abstrakt



- Claus O. Wilke (2017): Fundamentals of data visualization: A primer on making informative and compelling figures
  - https://serialmentor.com/dataviz/
  - Sehr zügig und auf den Punkt, praktischer, etwas knapper im Umfang, etwas weniger differenziert



Ziele von Grafiken

- Bezieht sich auf **Zweck**, **Publikum**, **Situation** und **Kontext** der Präsentation

- Explorativ vs. Explanativ
  - Dient die Grafik dazu, einen bestimmten Sachverhalt zu **demonstrieren**,…
  - oder soll sie dabei helfen, bestimmte Sachverhalte zu **erkennen**?

- Fachpublikum oder Laien
  - Wie sind **Vorerfahrungen, Gewohnheiten, Wissen** etc. des Zielpublikums? Wie ist deren Kenntnis des Inhalts und der grafischen Gestaltungsmöglichkeiten?

- Technische Aspekte
  - Ist die **Lesbarkeit** unter den Bedingungen, in denen die Grafik präsentiert wird, gewährleistet? (Auflösung, Schriftgröße, Schwarz-Weiß-Druck, …)

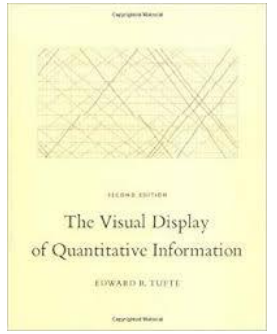
Ziele von Grafiken

- Wissenschaftliche Publikation oder Twitter-Feed
  - Soll die Grafik in einem wissenschaftlichen Format (Konferenzpräsentation, Abschlussarbeit, Publikation)…
  - oder auf Social Media in einem Feed gezeigt werden?
  - D.h. kann man sich der **Aufmerksamkeit** des Publikums sicher sein?

- Besonderheiten der Situation
  - Ist die Darstellungssituation durch andere Besonderheiten gekennzeichnet?
  - Beispiel: Präsentation am Pie-Day (14.03.)
  - Beispiel: Farben im Corporate Design

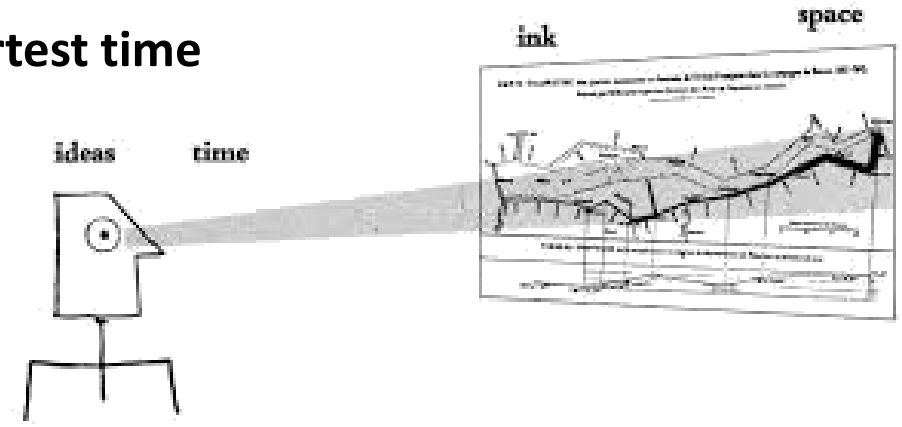
GUTE GRAFIKEN

Gute Grafiken



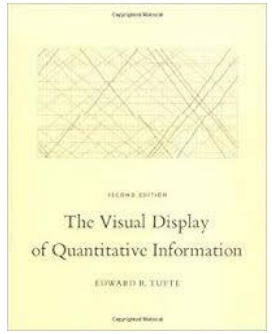
Graphical Excellence nach Tufte (2001, S. 51)

- … is the well-designed presentation of interesting data—a matter of **substance**, of **statistics**, and of **design**
- … consists of **complex ideas** communicated with **clarity, precision, and efficiency**
- … gives to the viewer the greatest **number of ideas** in the **shortest time** with the **least ink** in the **smallest space**



- … is nearly always **multivariate**
- … requires telling the **truth** about the data

Gute Grafiken: Design-Prinzipien nach Tufte (2001)

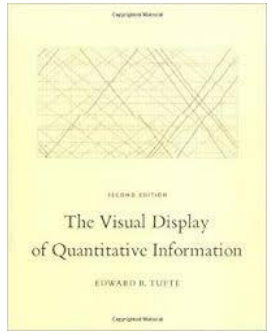


Prinzipien nach Tufte (2001)

- *Graphical integrity*
  - Die Wiedergabe von Zahlen sollte – physikalisch auf der Oberfläche der jeweiligen Grafik gemessen – proportional zu den dargestellten numerischen Größen sein (siehe "**Principle of Proportional Ink**")
  - Verwendung von **klaren, detaillierten und sorgfältigen Beschriftungen**, um grafische Verzerrungen und Doppeldeutigkeiten zu vermeiden
  - „Zeigen Sie die **Variation in den Daten**, nicht die Variation des Designs"
  - Grafiken sollten Daten nicht außerhalb des **Kontexts** wiedergeben

- *Above all else, show the data!*
- *Revise and edit!*

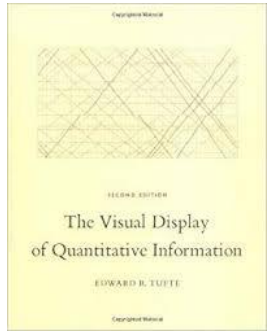
Gute Grafiken: Design-Prinzipien nach Tufte (2001)



Prinzipien nach Tufte (2001)

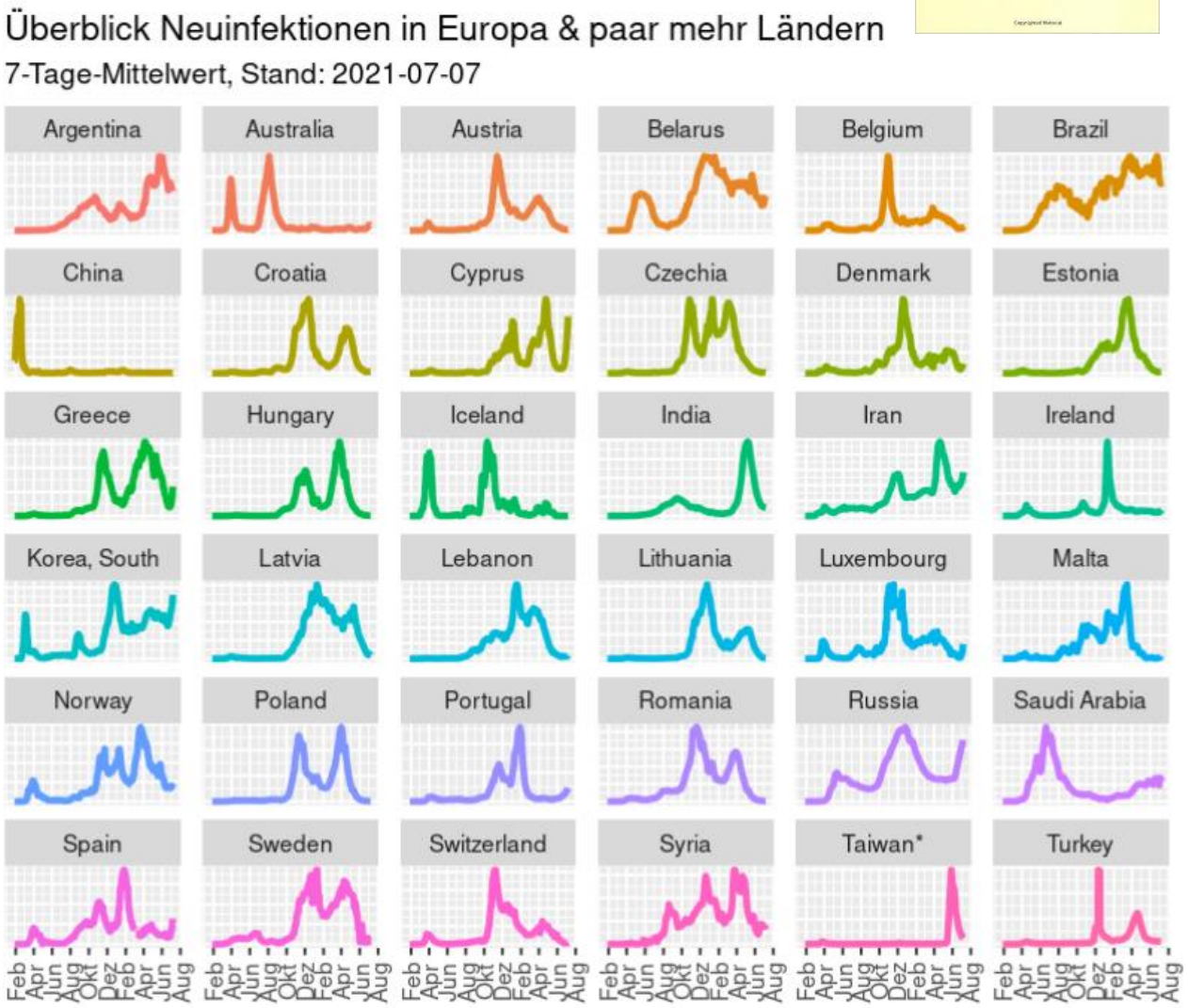
- Maximierung der Data-Ink-Ratio
  - $Data - ink - ratio = \frac{data{-}ink}{total\ ink\ used\ to\ print\ the\ graphic}$
  - Anteil der in der Grafik verwendeten „Tinte“ für die Darstellung nicht-redundanter Dateninformation
  - Redundante Anteile sind „**chartjunk**“
- Datendichte
  - $Data\ density\ of\ a\ display = \frac{number\ of\ entries\ in\ data\ matrix}{area\ of\ data\ display}$
  - Maximieren der Datendichte und der Größe der Datenmatrix in einem sinnvollen Maße, unter Ausnutzung der maximalen Auflösung der Anzeigetechnologie

Gute Grafiken: Design-Prinzipien nach Tufte (2001)



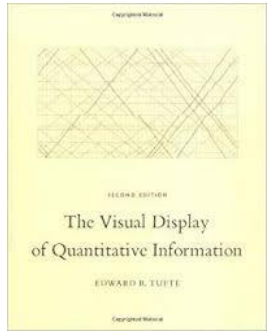
Prinzipien nach Tufte (2001)

- Small Multiples
  - Grafik-Serie, die die gleiche Kombination von Variablen für jede Ausprägung einer anderen
    Variable (z.B. Gruppe) zeigt
  - Aufmerksamkeit liegt auf Variation in den Daten
  - Erlaubt das schnelle Erkennen von Mustern
  - *„Small multiples are an excellent architecture for showing large quantities of multivariate data“* (S. 169)



https://queenjanine.shinyapps.io/COVID/

Gute Grafiken: Design-Prinzipien nach Tufte (2001)



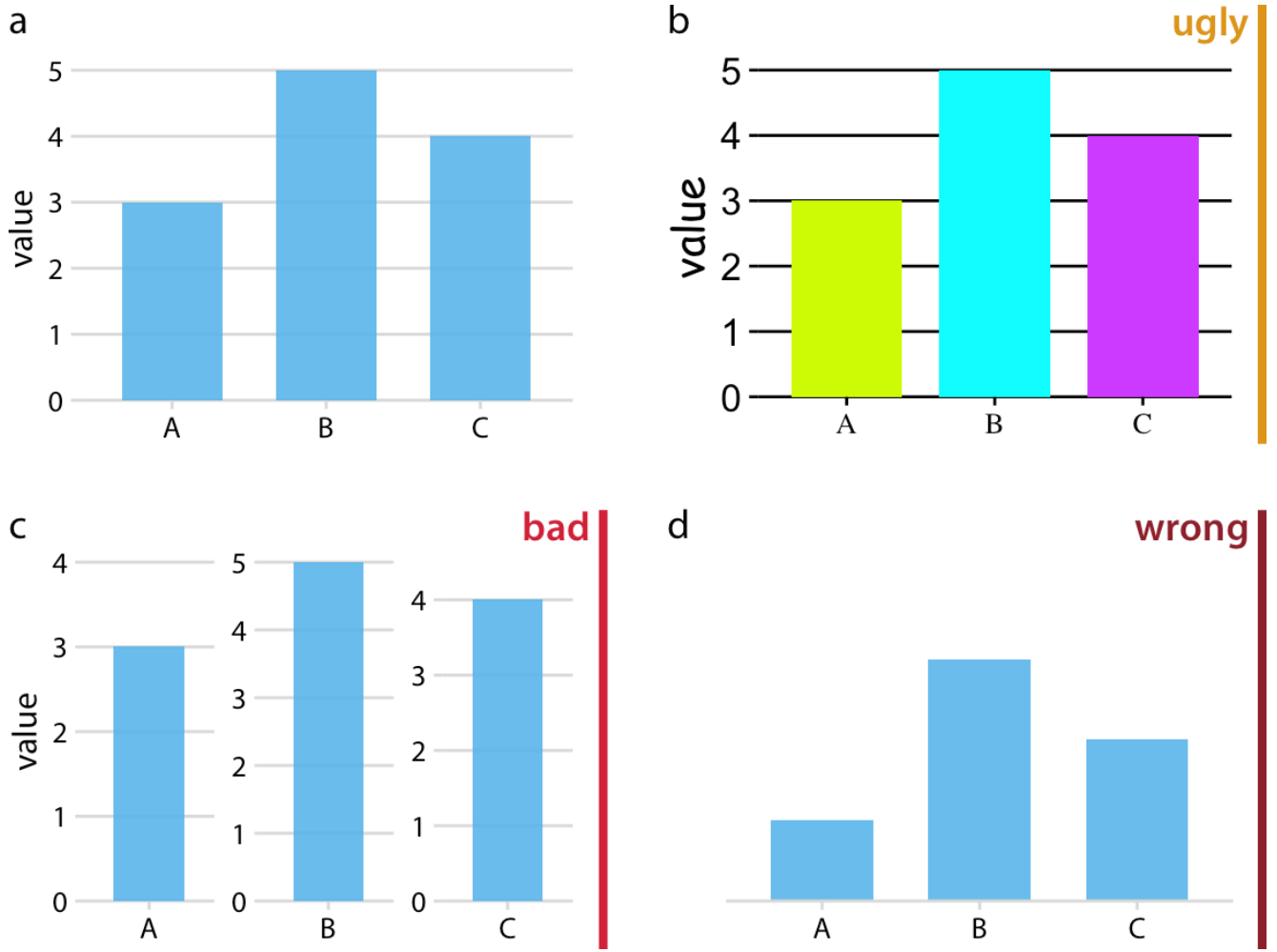
Prinzipien nach Tufte (2001)

- The friendly data graphic (S. 183)

Friendly	Unfriendly
Words are spelled out, mysterious and elaborate encoding avoided	Abbreviations around, requiring the viewer to sort through text to decode abbreviations
Words run from left to right, the usual direction for reading occidental languages	Words run vertically, particularly along the Y-axis; words run in several different directions
Little messages help explain data	Graphic is cryptic, requires repeated references to scattered text
Elaborately encoded shadings, cross-hatching, and colors are avoided; instead, labels are placed on the graphic itself; no legend is required	Obscure codings require going back and forth between legend and graphic
Graphic attracts viewer, provokes curiosity	Graphic is repellent, filled with chartjunk
Colors, if used, are chosen so that the color-deficient and color-blind (5-10% of viewers) can make sense of the graphic (blue can be distinguished from other colors by most color-deficient people)	Design insensitive to color-deficient viewers; red and green used for essential contrasts
Type is clear, precise, modest; lettering may be done by hand	Type is clotted, overbearing
Type is upper-and-lower case, with serifs	Type is all capitals, sans serif

SCHLECHTE GRAFIKEN

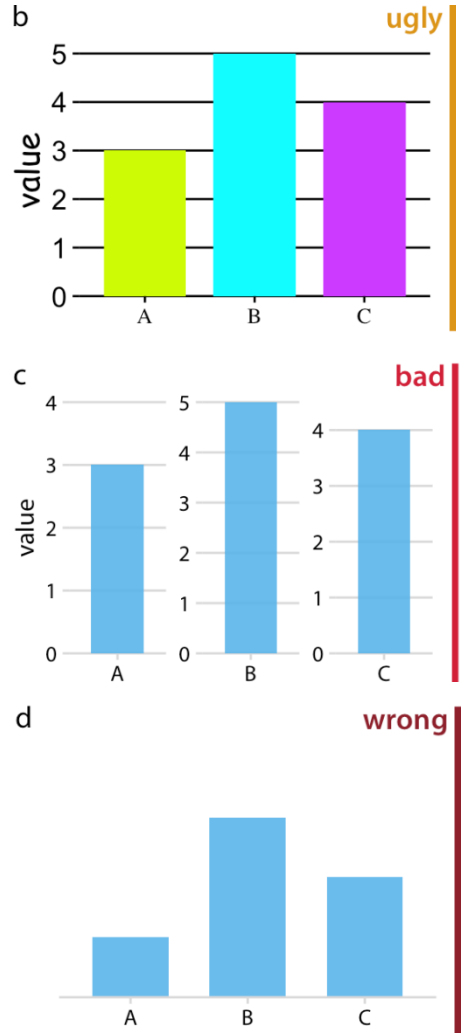
Warum sind Grafiken „schlecht“?



Wilke (2017): https://serialmentor.com/dataviz/introduction_files/figure-html/ugly-bad-wrong-examples-1.png

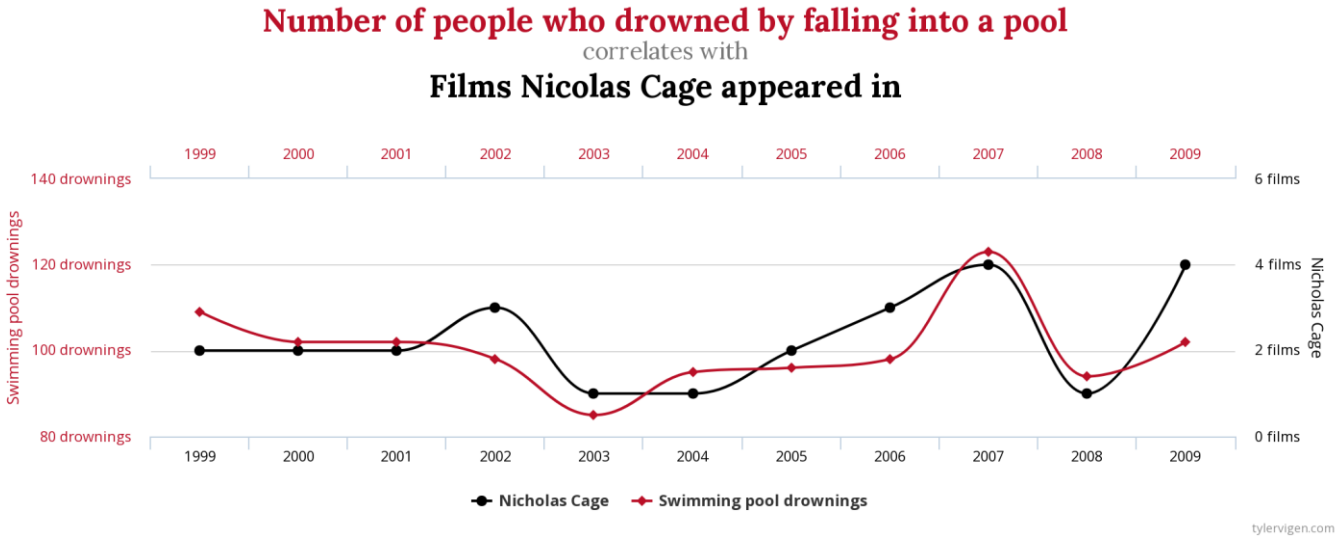
Warum sind Grafiken „schlecht“?

- **Hässlich:** Ästhetisch abschreckend, es macht keinen Spaß anzusehen, erregt negative Aufmerksamkeit
- **Schlecht**: Ungünstig designt, Verständnis wird unnötig erschwert, Fehler beim Lesen werden provoziert
- **Falsch:** Statistisch-mathematische Fehler beim Umgang mit Daten, zentrale Elemente zum Verständnis der Daten fehlen, Verständnis wird unmöglich gemacht



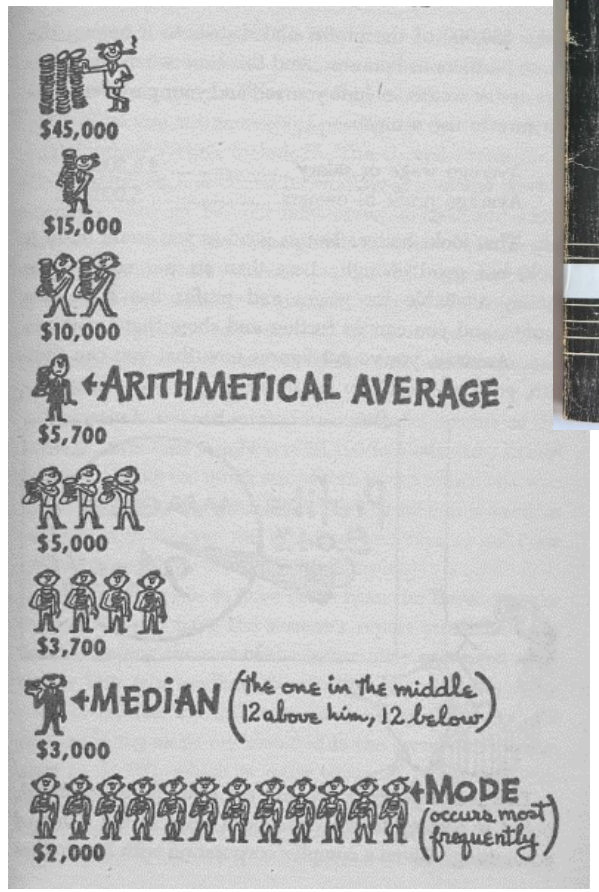
Schlechte Grafiken: Datengrundlage

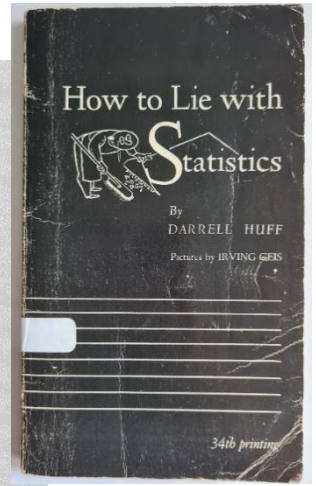
- **Scheinzusammenhänge**



http://www.tylervigen.com/spurious-correlations

- **Verwendung des Mittelwerts bei schiefen Verteilungen**





Huff (1954, S. 33)

Schlechte Grafiken: The Principle of Proportional Ink

- **Principle of proportional ink**
  - when a shaded region is used to represent a numerical value, the area of that shaded region should be directly proportional to the corresponding value

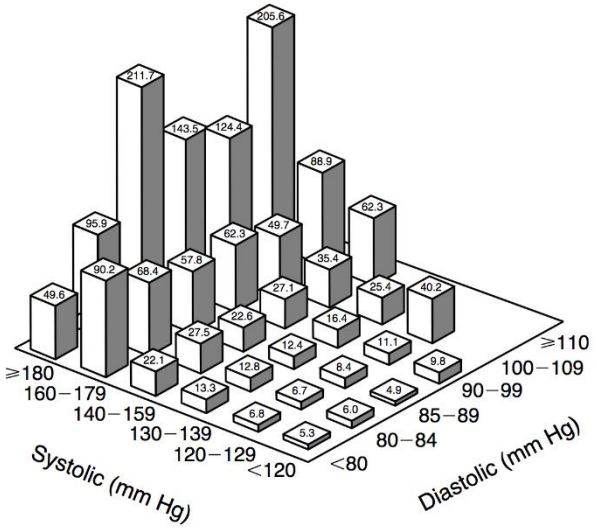
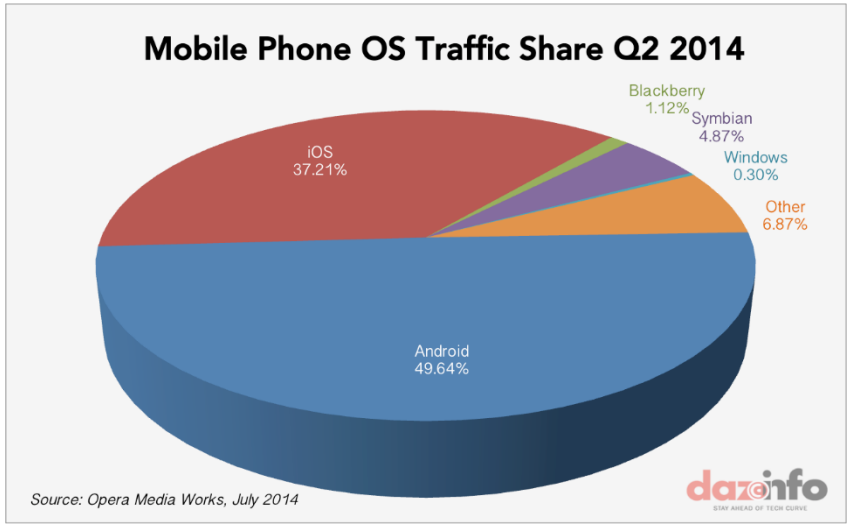
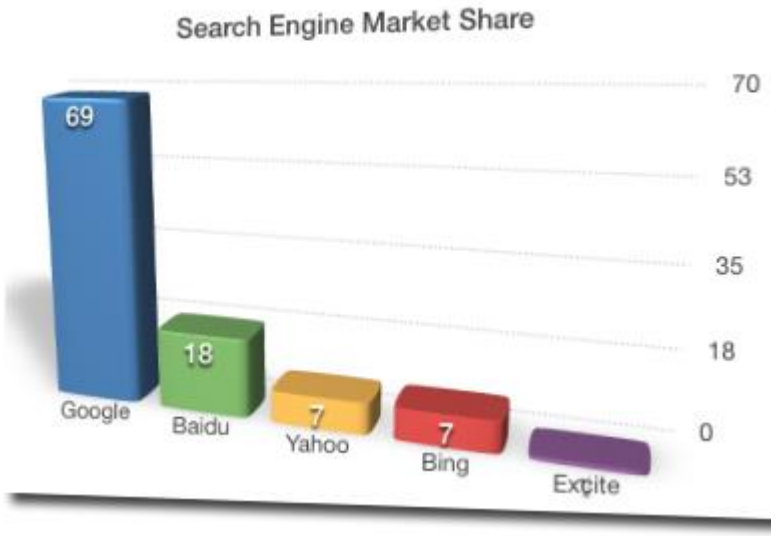


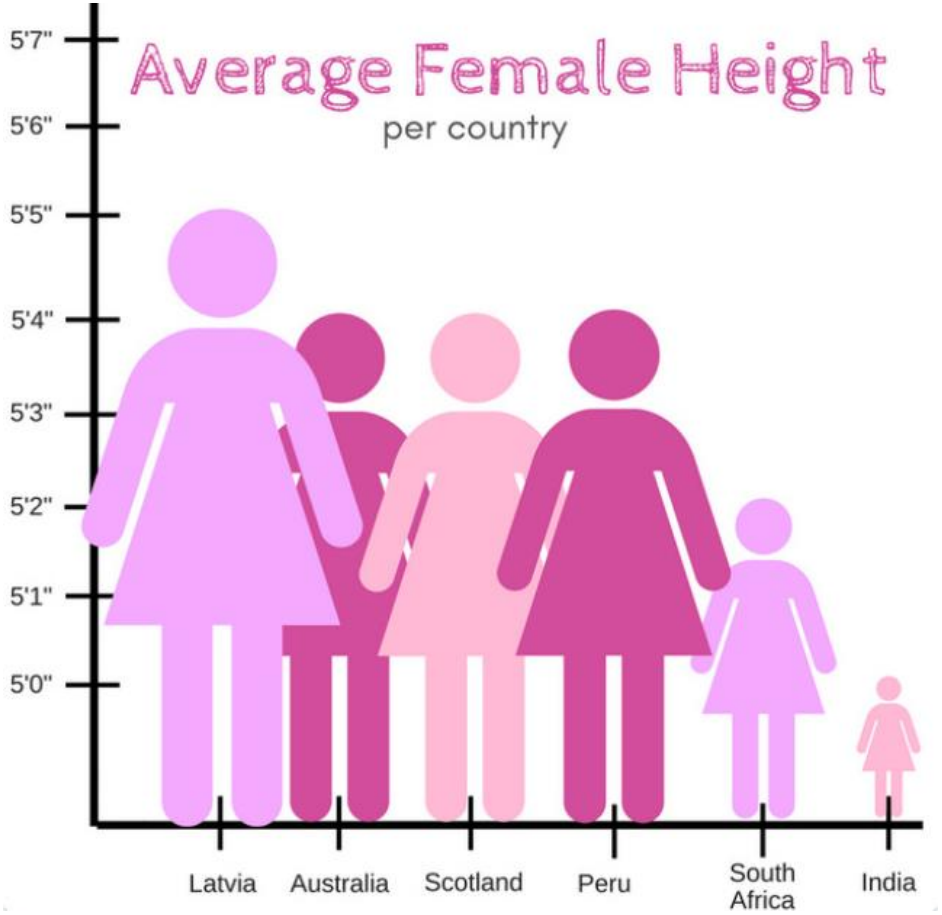
Figure 2. Age-Adjusted Rate of End-Stage Renal Disease Due to Any Cause per 100,000 Person-Years, According to Systolic and Diastolic Blood Pressure in 332,544 Men Screened for MRFIT.



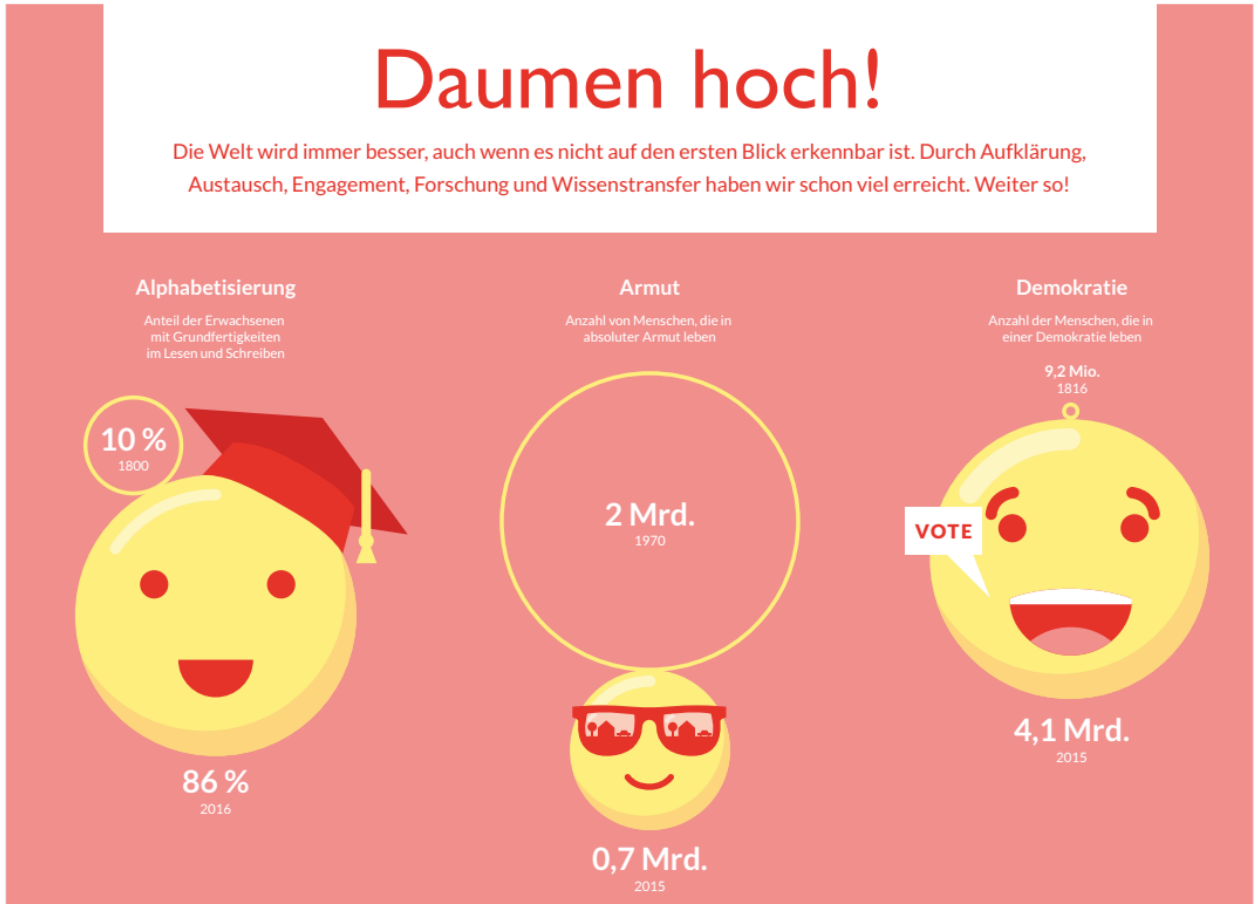


https://www.callingbullshit.org/tools/tools_proportional_ink.html

Schlechte Grafiken: The Principle of Proportional Ink



https://www.boredpanda.com/average-women-height-data-chart-latvian-indian-women/

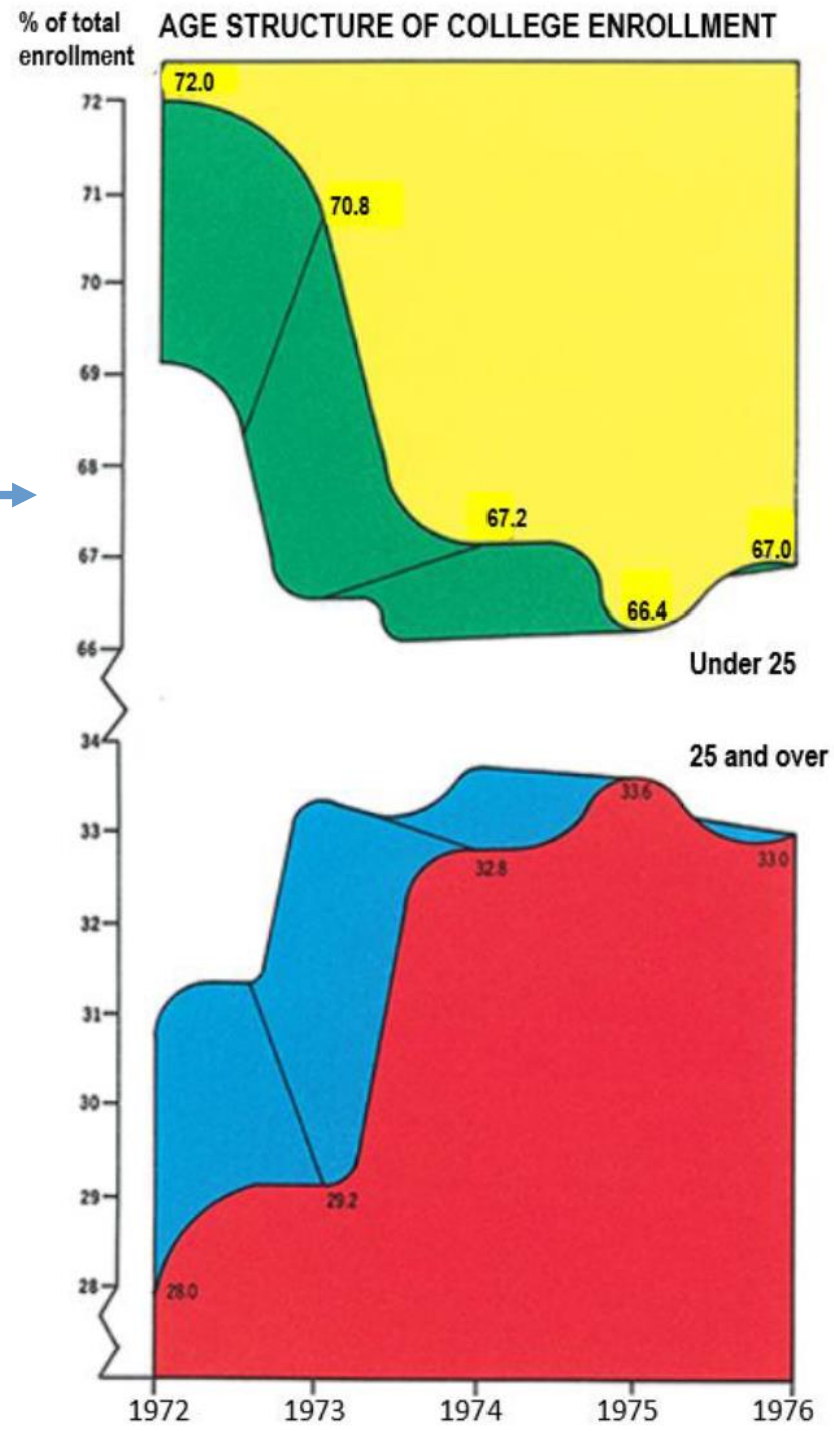


https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/change-22018-themenposter

Schlechte Grafiken: Data Density & Chartjunk

- **“This may well be the worst graphic ever to find its way into print”** (Tufte, 2001, S. 118)

(Tufte, 2001, S. 161)



Schlechte Grafiken: Lesbarkeit, Proportional Ink, Chartjunk…

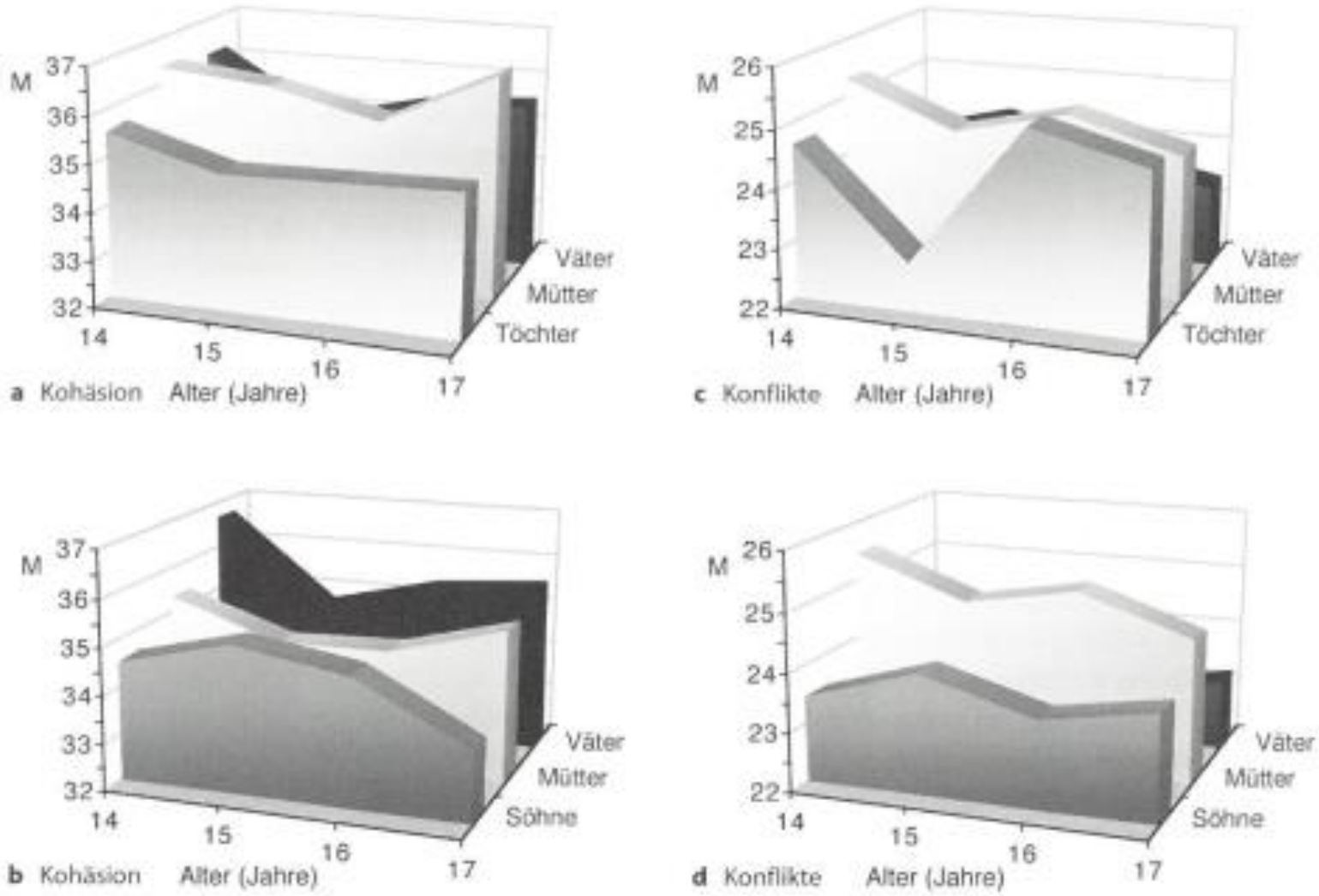
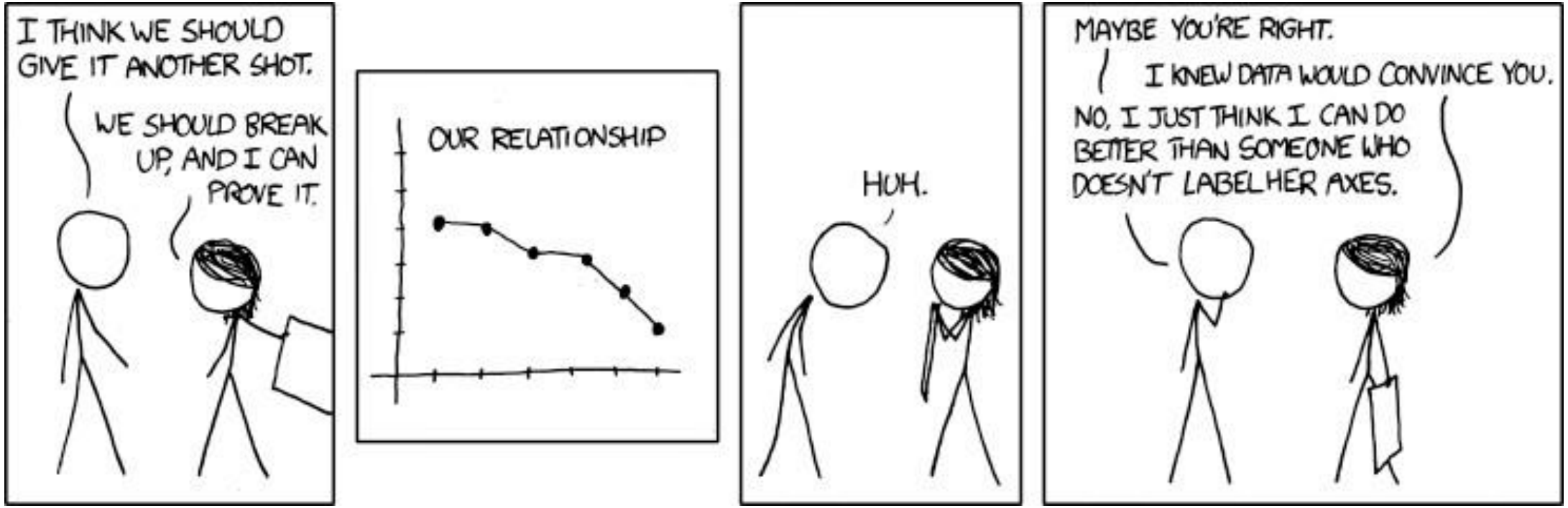


Abb. 6.10a-d. Kohäsion in Familien mit Töchtern und Söhnen (a,b) und Konflikthäufigkeit in Familien mit Töchtern und Söhnen (c,d)

Schlechte Grafiken: Lesbarkeit



https://xkcd.com/833/

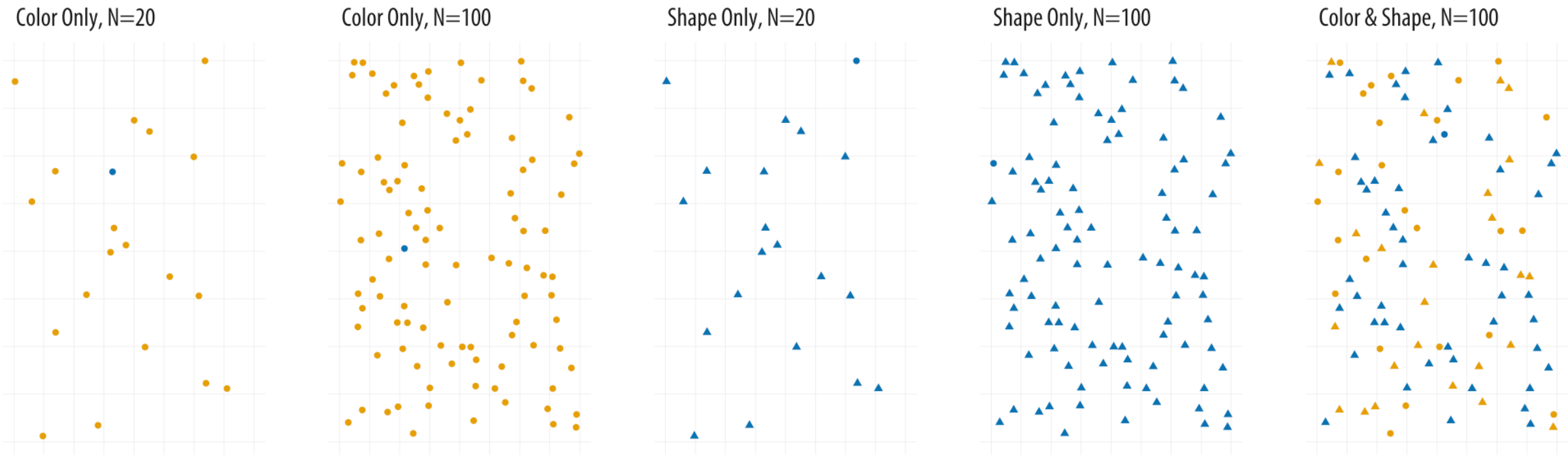
FARBEN UND FORMEN

Farben und Formen

- Farben und Formen können zur Enkodierung von Dateninformation in Grafiken verwendet werden
- Wir nehmen Muster nach gewissen Prinzipien wahr (*Gestalt-Prinzipien*), beispielsweise:
  - **Nähe**: Einander nahe Elemente werden als zusammengehörig wahrgenommen
  - **Ähnlichkeit**: Einander ähnliche Elemente werden als zusammengehörig wahrgenommen
  - **Verbundene Elemente**: Verbundene Elemente werden als zusammengehörig wahrgenommen
  - **Gemeinsames Schicksal**: Elemente mit ähnlicher Bewegungsrichtung werden als zusammengehörig wahrgenommen
  - …
- Diese können genutzt werden, um Elemente einer Grafik als zusammengehörig bzw. verschieden zu kennzeichnen

Visuelle Suche

- Wo ist der blaue Punkt?



http://socviz.co/dataviz-pdfl_files/figure-html4/ch-01-dual-search-1.png

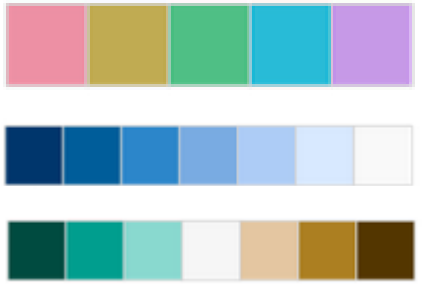
Visuelle Suche

- Wie gut zentrale Elemente erkannt werden können, hängt davon ab…
  - wie viele Elemente zu sehen sind
    - mehr Elemente ist schwieriger
  - auf welchen Kanälen die Elemente variieren
    - Form ist schwieriger als Farbe
  - wie ähnlich die Variationen in den einzelnen Kanälen sind
    - größere Ähnlichkeit ist schwieriger
  - in wie vielen Kanälen die Informationen variieren
    - mehr Kanäle wie Farbe *und* Form ist schwieriger
- Je mehr Elemente und je mehr Kanäle zur Darstellung genutzt werden, desto schwieriger wird es, Grafiken zu lesen
- Farben, insbesondere deutlich unterschiedliche, eignen sich gut, um Dinge hervorzuheben

Farben

- Farben können in R direkt als RGB oder HEX-Wert angesprochen werden, z.B. **DIPF-Blau**:
  - `"#6699CC"`
  - `rgb( 102 , 153 , 204, max=255 )`
- Farben können jedoch auch aus „Farbpaletten“ gezogen werden, z.B. `rainbow()`

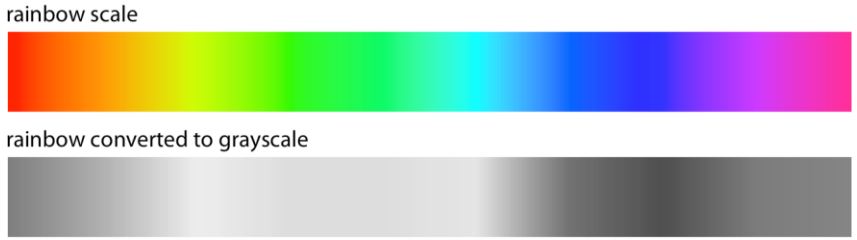
- Es gibt Farbpaletten für verschiedene Zwecke
  - *Qualitative*: Kodierung qualitativer Information ohne Rangreihe, gleiche Gewichtung aller Farben
  - *Sequential*: Kodierung ordinaler/metrischer Information, von klein zu groß (bzw. groß zu klein)
  - *Diverging*: Kodierung ordinaler/metrischer Information um einen neutralen Punkt, mit Abweichungen in beide Extreme

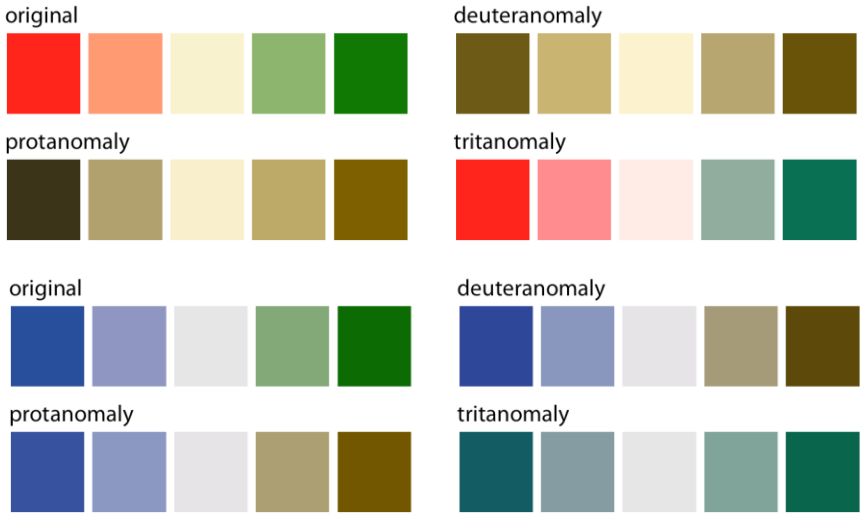


http://colorspace.r-forge.r-project.org/articles/hcl_palettes.html

Farben

- Potentielle Probleme
  - Verwendung nicht-monotoner Farbpaletten (z.B. rainbow) für sequentielle Daten → ungleichmäßige Helligkeit, impliziert nicht monotone Gewichtung
  - Schwarz-/Weiß-Druck
    - Sichtbar ist nur noch die Farbhelligkeit → Farben können nicht unterschieden werden oder implizieren nicht-monotone Gewichtung (z.B. Rainbow)
  - Farbfehlsichtigkeit
    - Je nach Fehlsichtigkeit können bestimmte Farben nicht voneinander unterschieden werden





https://clauswilke.com/dataviz/color-pitfalls.html

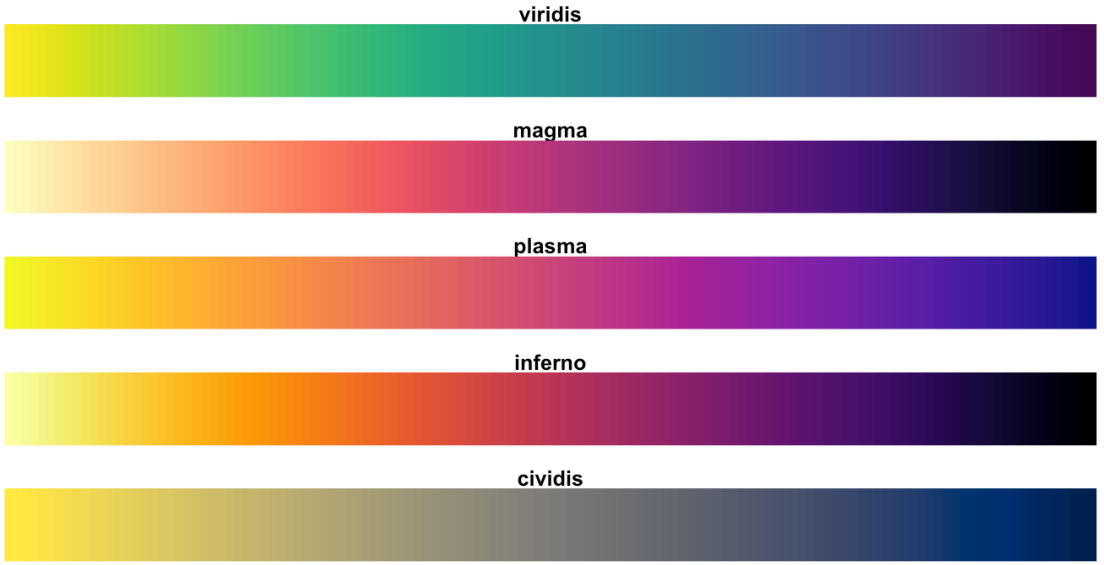
Farben

- Lösung für potentielle Probleme:
  Gezielte Auswahl von Farbpaletten für den spezifischen Zweck
  - Für qualitative Daten, z.B. Okabe & Ito (2008)

#E69F00 #56B4E9 #009E73 #F0E442 #0072B2 #D55E00 #CC79A7 #000000

https://clauswilke.com/dataviz/color-pitfalls.html

  - Für sequentielle Daten, z.B. viridis



https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html

  - ColorBrewer
    `library(RColorBrewer)`
    - Erstellung der eigenen Palette gegeben des Zwecks und der Art der Daten (Anzahl Gruppen)

GUTE GRAFIKEN: WISSENSCHAFTLICHER KONTEXT

DGPs Richtlinien zur Manuskriptgestaltung (2019)

- Die Richtlinien zur Manuskriptgestaltung der Deutschen Gesellschaft für Psychologie (DGPs) (2019) enthalten Regeln für die Erstellung von Manuskripten im Bereich der psychologischen Forschung, die oft auch in der Bildungsforschung herangezogen werden
- Neben formalen Hinweisen in Bezug auf statistische und mathematische Textteile, Tabellen, Quellenangaben im Text, wörtliche Zitate und die Erstellung des Literaturverzeichnisses sind auch Hinweise zur **Erstellung von Abbildungen** enthalten
- Kapitel 6, „Abbildungen“
  - „Jede Art nicht textlicher Darstellung von Material, die nicht in Tabellenform geschieht, wird unter dem Begriff Abbildung zusammengefasst. Abbildungen beinhalten u.a. Fotografien, Grafiken, Diagramme, Schemata.“

DGPs Richtlinien zur Manuskriptgestaltung (2019)

- **6.1 Arten von Abbildungen**
  - Grafische Darstellungen statistischer Ergebnisse zeigen in der Regel **Vergleiche oder Verteilungen** und können z.B. absolute Werte, Prozentwerte oder Maßzahlen illustrieren. Die **Linien sollten sauber und klar** dargestellt sein, **überflüssige Details sind zu vermeiden**. Abszisse und Ordinate sollen von **klein nach groß** skaliert sein und **vergleichbare Maßeinheiten** enthalten
  - Details zu einzelnen Grafiktypen (Liniendiagramme, Balkendiagramme, Kreis- oder Kuchendiagramme, Streudiagramme, Fluss- und Strukturdiagramme), besonders relevant:
    - Farben sollen im Falle eines einfarbigen Drucks (schwarz-weiß) noch unterscheidbar sein (Liniendiagramm: bspw. eine gestrichelte und eine durchgezogene Linie)
    - 3D-Darstellungen oder Schattierungen der Balken sind zu vermeiden
    - Bedeutung der Farben muss durch eine Legende erläutert werden

DGPs Richtlinien zur Manuskriptgestaltung (2019)

- Kreis- oder Kuchendiagramme (S. 92)

> *Kreis- oder Kuchendiagramme* werden verwendet, um Prozentsätze oder Größenverhältnisse darzustellen. Dabei sollte die Anzahl der Segmente fünf nicht übersteigen. Die Segmente sind nach Größe geordnet zu reihen (vom größten zum kleinsten), beginnend mit der Position, die 12 Uhr entspricht. Zur Kennzeichnung der Segmente können entweder unterschiedliche Grauschattierungen von weiß bis schwarz (für das kleinste Segment) oder unterschiedliche Schraffuren oder Punktmuster eingesetzt werden. Dabei ist darauf zu achten, dass die Farben auch noch im Falle eines einfarbigen Drucks (schwarz-weiß) eindeutig unterscheidbar sind. Die Bedeutung der Farben muss durch eine Legende oder eine eindeutige Beschriftung der Kreissegmente erläutert werden.

DGPs Richtlinien zur Manuskriptgestaltung (2019)

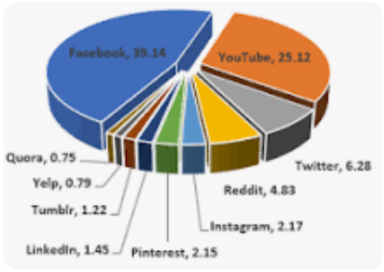
- Kreis- oder Kuchendiagramme

Alle | Bilder | News | Videos | Shopping | Mehr | Suchfilter

Ungefähr 72.700.000 Ergebnisse (0,60 Sekunden)

Why you shouldn't use pie charts

- Quantity is represented by slices; humans aren't particularly good at estimating quantity from angles, which is the skill needed.
- Matching the labels and the slices can be hard **work**.
- Small percentages (which might be important) are tricky to show.



https://scc.ms.unimelb.edu.au › resources-list › no_pie-ch...

Why you shouldn't use pie charts - Statistical Consulting Centre

https://scc.ms.unimelb.edu.au/resources-list/data-visualisation-and-exploration/no_pie-charts

Bad by definition

A pie chart is a circle divided into sectors that each represent a proportion of the whole. It is often used to show percentage, where the sum of the sectors equals 100%.

The problem is that humans are pretty bad at reading angles. In the adjacent pie chart, try to figure out which group is the biggest one and try to order them by value. You will probably struggle to do so and this is why pie charts must be avoided.

“The issue with pie chart”:
https://www.data-to-viz.com/caveat/pie.html

DGPs Richtlinien zur Manuskriptgestaltung (2019)

- **6.2 Formale Gestaltung von Abbildungen im Manuskript**
  - **Dateiformat**: gängige Dateiformate wie .tif, .jpg, .bmp, .gif oder .eps

    ```
    ggsave("filename.xxx")
    ```
    . xxx = .tiff, .jpg, .bmp, .eps
  - Für Abszisse und Ordinate sollten mitteldicke Linien verwendet werden

    ```
    geom_line(size=<value>
    ```
  - Passende, unmissverständliche **Skalierung der Achsen**; bei Ratingskalen:
    - Abszisse geht vom kleinstmöglichen zum größtmöglichen Wert

      ```
      xlim=c(min, max)
      ```
    - Skalierung der Ordinate wird an den Wertebereich der abhängigen Variablen angepasst
  - Bei der Darstellung von Mittelwerten sollten **Fehlerbalken** für Standardabweichungen, Standardfehler oder Konfidenzintervalle eingefügt werden

    ```
    geom_errorbar()
    ```
  - Achsen sind stets zu **beschriften**: Abszisse (unterhalb), Ordinate (links von der Ordinate, 90 Grad gedreht)

    ```
    title(xlab="...", ylab="...")
    ```

DGPs Richtlinien zur Manuskriptgestaltung (2019)

- **6.2 Formale Gestaltung von Abbildungen im Manuskript (Fortsetzung)**
  - Die **Größe der Beschriftung, Zeichen, Symbole und der Legende** muss mindestens einer 8-Punkt-Schrift entsprechen, sie soll andererseits nicht größer als eine 14-Punkt-Schrift sein. Die Schriftgrößen sollten innerhalb einer Abbildung um nicht mehr als 4 Punkte variieren. Als Faustregel gilt, dass die in einer grafischen Darstellung verwendeten Symbole die Größe eines durchschnittlichen kleinen Buchstabens haben sollten. **Als bevorzugter Schrifttyp für Abbildungen sind serifenlose Schriften** einzusetzen, z.B. die Schriftarten Arial oder Calibri
  - Die Verlaufskurven sollten durch einfache geometrische Formen an den Messpunkten **unterscheidbar** sein. Offene und gefüllte Kreise und Dreiecke sind gut unterscheidbar; weniger gut unterscheidbar sind Kombinationen aus Quadraten und Kreisen oder Quadraten und Rauten. Ein Diagramm sollte nicht mehr als vier Verlaufskurven enthalten; die Abstände zwischen den Verlaufskurven sollten auch nach der Reproduktion (Verkleinerung) noch gut erkennbar sein.
  - Bei der Variation der Größe der Elemente ist auch deren **Wichtigkeit** zu berücksichtigen, d.h. wichtigere Elemente sollten hervorstechen. Zum Beispiel sollten Verlaufskurven oder Balkendiagramme dicker sein als die Achsenbezeichnungen und diese wieder dicker als die Achsen selber

DGPs Richtlinien zur Manuskriptgestaltung (2019)

- **6.4 Titel und Legenden zur Abbildung**
  - Die Abbildung soll jedenfalls **für sich allein verständlich** sein, ohne dass die Leserinnen und Leser auf den Text des Artikels angewiesen sind
  - **Legenden** stellen einen Bestandteil der Abbildung dar, sie erklären die in der Abbildung verwendeten Zeichen und Symbole. Sie werden innerhalb der Abbildung angebracht und die Art ihrer Schriftgestaltung soll daher der der übrigen Abbildung entsprechen.

DGPs Richtlinien zur Manuskriptgestaltung (2019)

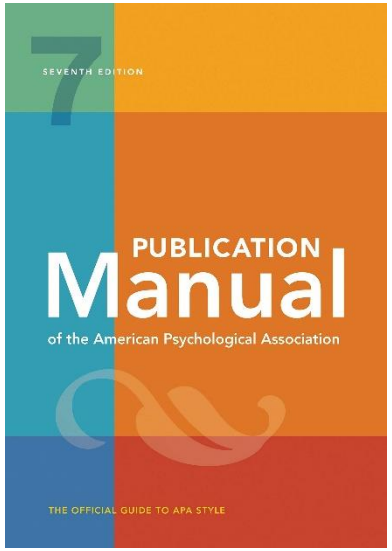
- **6.5 Herstellung der Abbildungen für das Druckverfahren**
  - Bei der Erstellung einer Abbildung ist darauf zu achten, dass sie in ihren **Größenverhältnissen in eine Spalte der Zeitschrift passt**, in der sie abgedruckt werden soll. Eine Ausnahme von dieser Regel sollte nur gemacht werden, wenn die Darstellung feinster Details oder zahlreicher Einzelheiten die Nutzung der vollen Breite der Zeitschriftenseite erforderlich macht.
  - **Teilabbildungen** sind mit Großbuchstaben zu kennzeichnen (bspw. „A“ und „B“)
  - Nicht mehr als zwei bis drei verschiedene Formen der Schattierung oder Musterung verwenden
  - Die **Druckqualität** sollte zumindest 300 dpi betragen (600 bis 1200 dpi werden empfohlen)
  - Bei der Reproduktion sind grundsätzlich **reine Schwarz-Weiß-Grafiken**, Halbtongrafiken (mit Grauschattierungen) und Farbabbildungen zu unterscheiden. Farbabbildungen erfordern einen speziellen Druckvorgang und sind deshalb in der Reproduktion teurer.

DGPs Richtlinien zur Manuskriptgestaltung (2019)

- **6.7 Checkliste für die Erstellung von Abbildungen** (Auszug)
  - Ist die Abbildung notwendig und sinnvoll?
  - Ist die Abbildung einfach, klar und frei von unwesentlichen Details?
  - Wurden alle Werte korrekt eingezeichnet?
  - Ist die Skaleneinteilung korrekt proportioniert?
  - Ist die Beschriftung hinsichtlich ihrer Größe, Farbe und ihrem Verhältnis zum Rest der Abbildung einwandfrei?
  - Sind gleiche Abbildungen oder Abbildungen gleicher Wichtigkeit alle mit der gleichen Skaleneinteilung versehen?
  - Sind alle Ausdrücke richtig geschrieben?
  - Werden alle verwendeten Abkürzungen und Symbole im Titel oder in der Legende zu der Abbildung erläutert? Sind die in der Abbildung verwendeten Symbole, Abkürzungen und Terminologien konsistent zu denen in der Beschriftung zu dieser Abbildung, zu anderen Abbildungen und zum Text?
  - Sind digitale Dateien von Abbildungen in entsprechenden Dateiformaten abgespeichert (z.B. .tif, ."pg, .bmp, .gif oder .eps* oder wurden als Originaldatei des Grafikprogramms, in der die Abbildung erstellt wurde gespeichert?

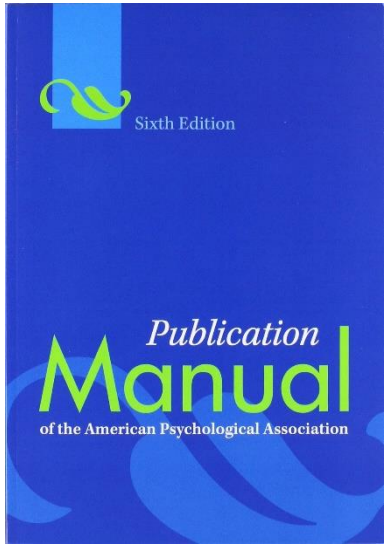
APA Publication Manual

- Das Publication Manual of the American Psychological Association (APA) liegt in der 7. Auflage vor (2020)
- Die Hinweise zur Manuskripterstellung bilden den Standard, den Verlage an Manuskripte in unserem Bereich stellen – sind denen der DGPs jedoch sehr ähnlich
- Hinweise zur Erstellung von Abbildungen in der 6. Auflage (2010) sind in Kapitel 5 enthalten: „Displaying Results“



- Potentiell ebenfalls relevant:
  - 4.37 Commas in Numbers
    - Use Commas between groups of three digits in most figures of 1,000 or more

```
library(scales)
scale_x_continuous(labels = scales::comma)
```



APA Publication Manual (2010)

- **5.02 Design and Preparation of a Data Display**
  - Design your graphical display with the reader in mind; that is, remember the communicative function of the display.
    - Place items that are to be compared next to each other.
    - Place labels so that they clearly abut the elements they are labeling.
    - Use fonts that are large enough to be read without the use of magnification.
    - Include all of the information needed to understand it within the graphical image – avoid novel abbreviations, use table notes, and label graphical elements
    - Keep graphical displays free of extraneous materials, no matter how decorative those materials may make the graphic look.

  - **Communication is the primary purpose of the graphic. This does not mean, however, that well-designed, aesthetically pleasing graphics are not important. An attractive graphical display makes a scientific article a more effective communication device.**

APA Publication Manual (2010)

- **5.22 Standards for Figures**
  - The standards for good figures are simplicity, clarity, continuity, and (of course) information value.
    A good figure
    - augments rather than duplicates the text,
    - conveys only essential facts,
    - omits visually distracting detail,
    - is easy to read-its elements (type, lines, Iabels, symbols, etc.) are large enough to be read with ease,
    - is easy to understand-its purpose is readily apparent,
    - is consistent with and in the same style as similar figures in the same article, and
    - is carefully planned and prepared.

APA Publication Manual (2010)

- **5.22 Standards for Figures** (Fortsetzung)
  - Be certain in figures of all types that
    - lines are smooth and sharp,
    - typeface is simple (sans serif) and legible,
    - units of measure are provided,
    - axes are clearly labeled, and
    - elements within the figure are labeled or explained.

  - In addition, be sure in all figures that
    - sufficient information is given in the legend to make the figure understandable on its own,
    - symbols are easy to differentiate, and
    - the graphic is large enough for its elements to be discernible

APA Publication Manual (2010)

- **5.30 Figure Checklist**
  - The following checklist may be helpful in ensuring that your figure communicates most effectively and conforms to APA Style and formatting conventions.
    - Is the figure necessary?
    - Is the figure simple, clear, and free of extraneous detail?
    - Is the figure title descriptive of the content of the figure?
    - Are all elements of the figure clearly labeled?
    - Are the magnitude, scale, and direction of grid elements clearly labeled?
    - Are figures of equally important concepts prepared according to the same scale?
    - Are all figures numbered consecutively with Arabic numerals?
    - Are all figures mentioned in the text?
    - Has written permission for print and electronic reuse been obtained? Is proper credit given in the figure caption?
    - Have all substantive modifications to photographic images been disclosed?
    - Are the figures being submitted in a file format acceptable to the publisher?
    - Have the files been produced at a sufficiently high resolution to allow for accurate reproduction?

Fazit

- Die Qualität einer Grafik hängt ab von der Qualität der **Daten**, dem **Design** und dem **Ziel der Grafik**
- Zur Bemessung der Qualität einer Grafik spielen subjektiv-bewertende Aspekte, aber auch handwerklich-objektive Faktoren eine Rolle
- Es gibt keine allgemeingültigen Kriterien für eine „gute“ Grafik, aber es gibt viele Möglichkeiten, eine schlechte Grafik zu erzeugen
- Die Erstellung von Grafiken ist **Wissenschaft – und ein bisschen Kunst**
- Für den Bereich wissenschaftlicher Manuskripterstellung gibt es teils sehr präzise Vorgaben

BLOCK 4:
SCHÖNE GRAFIKEN: PRAXIS

BLOCK 5:
ÜBUNG 1

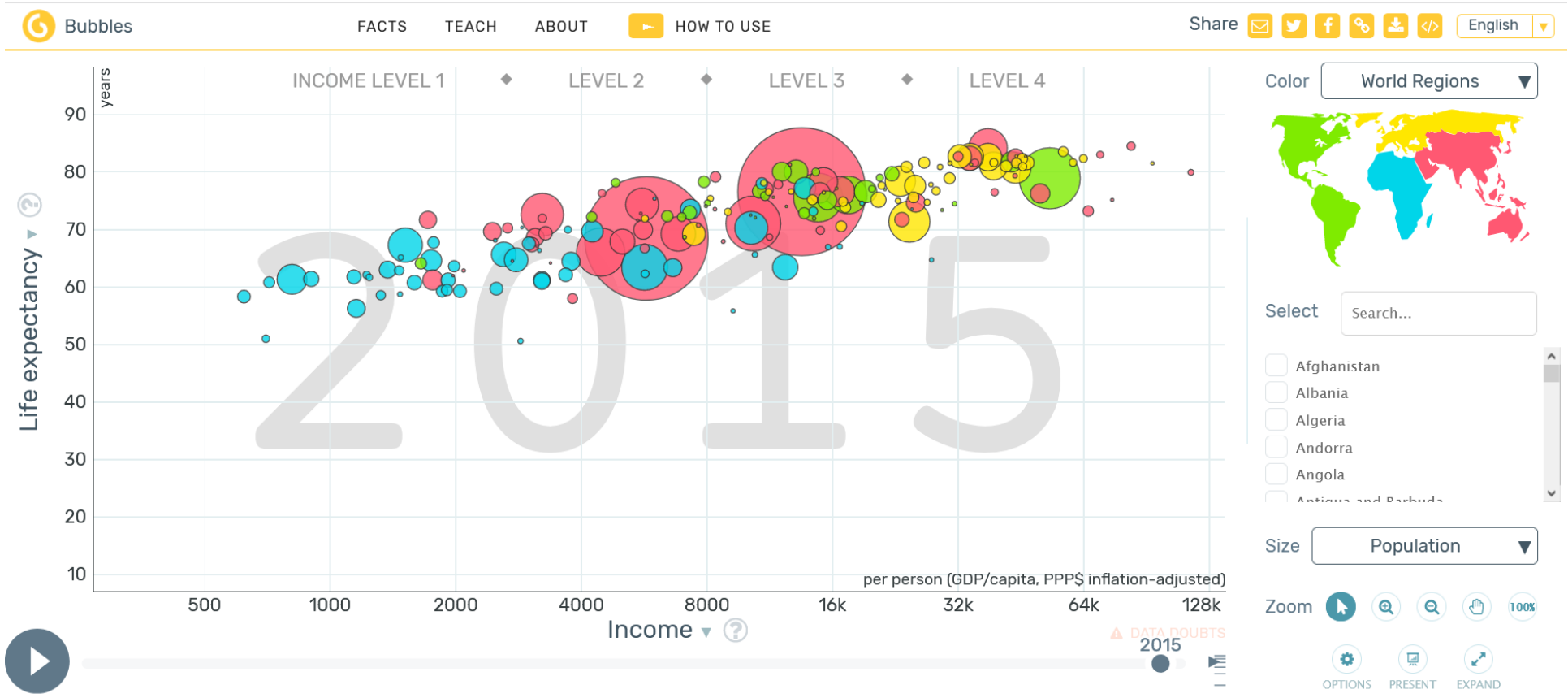
Übung 1: Grafik „nachbauen“

Auf der nachfolgenden Folie sehen Sie ein auf gapminder.org erzeugtes bubble chart.

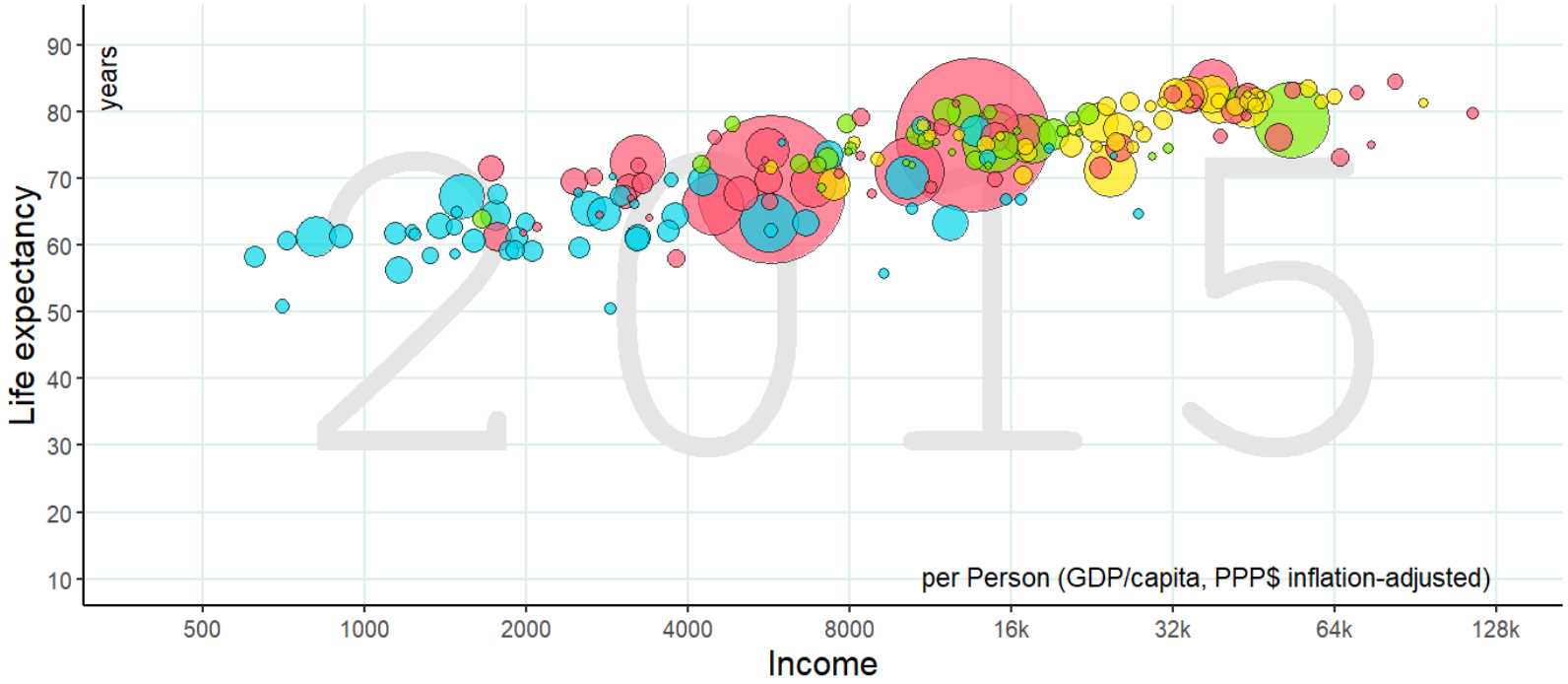
1. Wie viele und welche Variablen sind darin enkodiert?

2. Versuchen Sie, diese Grafik anhand des `edu_exp`-Datensatzes so gut wie möglich mit ggplot2 nachzubasteln.

Übung 1: Grafik „nachbauen“



Übung 1: Grafik „nachbauen“



BLOCK 6:
GGPLOTPOURRI

BLOCK 7:
GGANIMATE

BLOCK 8:
PLOTLY

BLOCK 9:
EXPLORATIVE GRAFIKEN

Explanative vs. explorative Grafiken

- Unterschiedliche Ziele von Grafiken
  - *Explanativ*: Die Grafik dient dazu, anderen einen bestimmten Sachverhalt (ein Muster in den Daten) zu **kommunizieren**
  - *Explorativ*: Die Grafik soll dabei helfen, bestimmte Sachverhalte (Muster in den Daten) zu **erkennen**

- Der Fokus des Workshops lag bisher stark auf *explanativen* Grafiken

- Nachfolgend werden Techniken für *explorative* Grafiken vorgestellt:
  - Prinzip des „small multiple“ anhand von *faceting*
  - Eine Serie gleicher Grafiken mit Funktion und Schleife erzeugen
  - Viele bivariate Zusammenhänge mit **ggpairs()**
  - Multivariate Daten mit **tableplot()** durchkämmen

GEOM_FACET
PRINZIP DES „SMALL MULTIPLE“

Faceting

- Die Technik des *faceting* erlaubt es, Grafiken gemäß Edward Tufte‘s Prinzip des ***small multiple*** (siehe Block 3) zu erzeugen

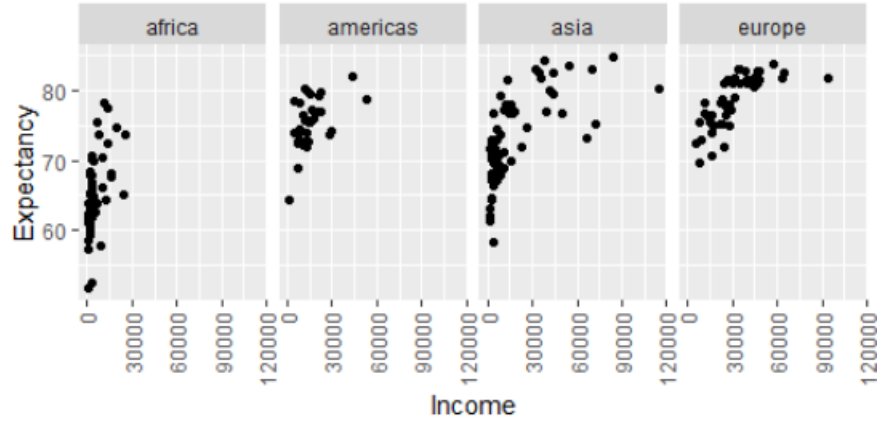
> *At the heart of quantitative reasoning is a single question: Compared to what? Small multiple designs, multivariate and data bountiful, answer directly by visually enforcing comparisons of changes, of the differences among objects, of the scope of alternatives. For a wide range of problems in data presentation, small multiples are the best design solution.*
>
> Tufte (1990). Envisioning Information. Graphics Press (p. 67).

- Grafiken, die diesem Prinzip entsprechen, sind stets gleich aufgebaut und ermöglichen so einen schnellen Vergleich der dargestellten Information zwischen den einzelnen Grafiken.

Faceting

- Zwei Funktionen für *faceting* in gglot2:
  - **facet_grid()**: Darstellung der Ergebnisse von Gruppen in
    - Zeilen oder
    - Spalten oder
    - einem Raster (*grid*) aus Zeilen und Spalten
  - **facet_wrap()**: Darstellung der Ergebnisse von Gruppen mit Zeilenumbruch (*wrap*)
- Wichtig: Gruppe muss als *diskrete* Variable im Datensatz enthalten sein
- Beispiel: Darstellung der Gruppen in Spalten `facet_grid(. ~ Gruppenvariable)`

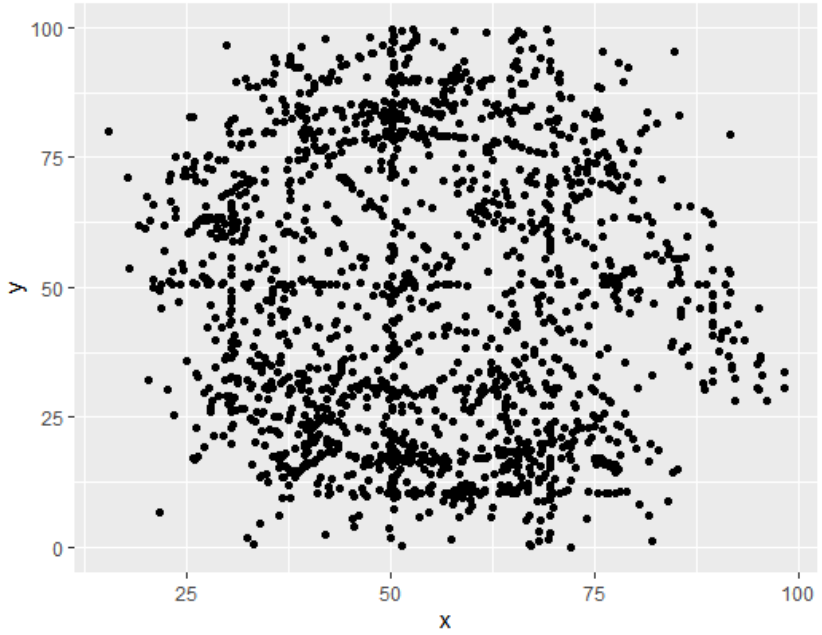
```
> edu_exp ▷
+   subset(Year == 2016) ▷
+   ggplot(aes(x=Income, y=Expectancy)) +
+   geom_point() +
+   theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = .5)) +
+   facet_grid(. ~ Region)
```



Faceting: Der Datasaurus

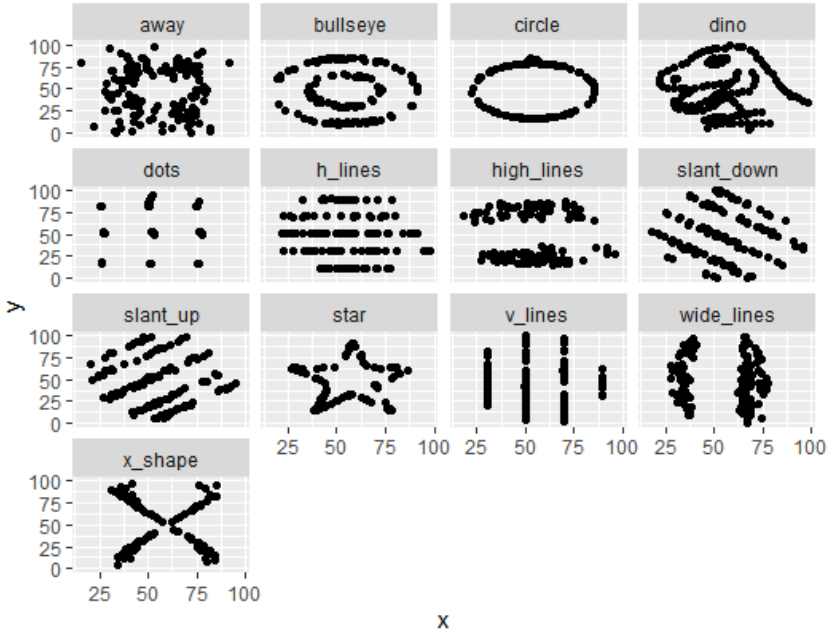
- Ohne *faceting*

```
> library(datasauRus)
> ggplot(datasaurus_dozen, aes(x=x, y=y)) +
+   geom_point()
```



- Mit *faceting* bedingt auf Variable „dataset“

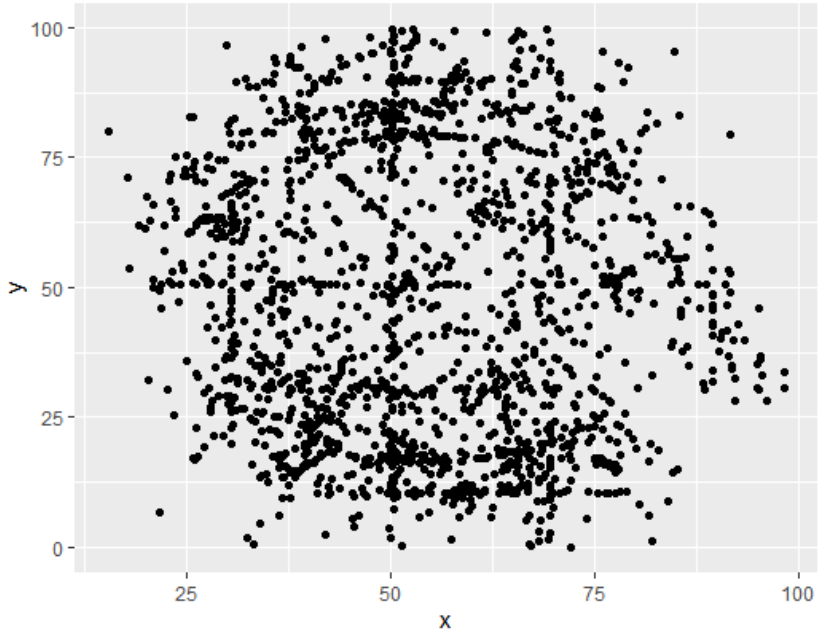
```
> ggplot(datasaurus_dozen, aes(x=x, y=y)) +
+   geom_point() +
+   facet_wrap(~dataset)
```



Faceting: Der Datasaurus

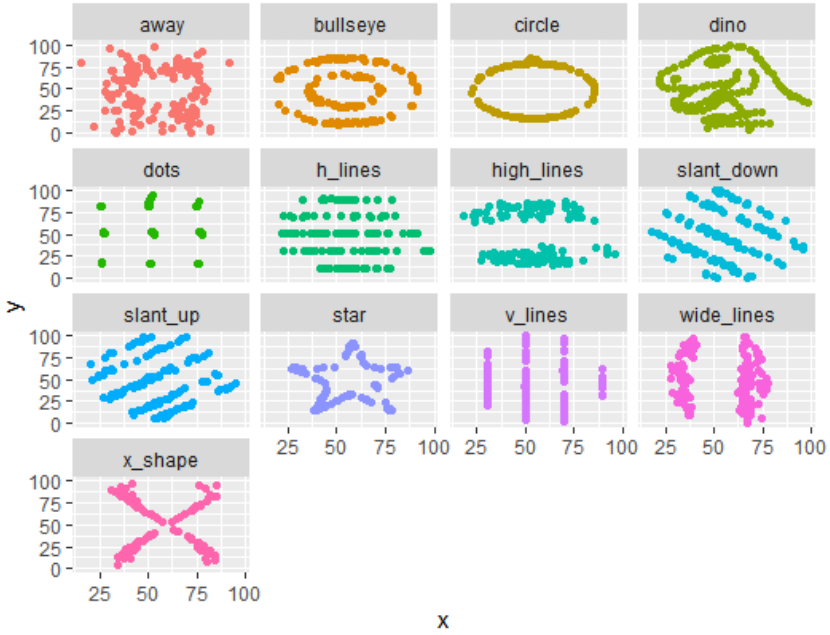
- Ohne *faceting*

```
> library(datasauRus)
> ggplot(datasaurus_dozen, aes(x=x, y=y)) +
+   geom_point()
```



- Mit *faceting* bedingt auf Variable „dataset“, und Farbe

```
> ggplot(datasaurus_dozen, aes(x=x, y=y)) +
+   geom_point(aes(colour=dataset), show.legend=F) +
+   facet_wrap(~dataset)
```



EINE SERIE GLEICHER GRAFIKEN MIT FUNKTION UND SCHLEIFE ERZEUGEN

Eine Serie gleicher Grafiken mit Funktion und Schleife erzeugen

- Bisher haben wir stets Funktionen benutzt, die im Basis-R oder in den aktuell geladenen Paketen enthalten sind
- **Funktionen** können jedoch auch selbst geschrieben und an die eigenen Bedürfnisse (z.B. eine bestimmte Grafik) angepasst werden
- Mithilfe einer **`for()`-Schleife** kann diese Funktion dann auf alle Variablen eines Datensatzes, auf alle Länder eines Datensatzes etc. angewendet werden
  → so entsteht eine Serie von Grafiken, die man schnell durchscrollen und auf Auffälligkeiten überprüfen kann

```
> my_fun <- function(argument){
+   ergebnis <- argument + 2
+   return(ergebnis)
+ }
> my_fun(1)
[1] 3
```

```
> for(z in 1:5){
+   print(z*2)
+ }
[1] 2
[1] 4
[1] 6
[1] 8
[1] 10
```

Eine Serie gleicher Grafiken mit Funktion und Schleife erzeugen

- Beispiel: es soll ein Diagramm für die Lebenserwartung im Zeitverlauf für jedes Land im Datensatz `edu_exp` erzeugt und gespeichert werden
  - Teil 1: Funktion zur Erstellung des Plots

```
> gm.plot ← function(which.country, show.mean=FALSE){
+
+   dipfblau ← rgb(102,153,204, max=255)
+   tmp.data ← subset(edu_exp, Country == which.country)
+   tmp.mw   ← mean(tmp.data$Expectancy)
+
+   tmp.plot ← ggplot(tmp.data, aes(x=Year, y=Expectancy)) +
+     geom_line(size=1.5, show.legend=F) +
+     geom_point(size=2, show.legend=F) +
+     xlim(1997, 2017) +
+     ylim(20, 85) +
+     labs(x="Jahr", y="Lebenserwartung",
+          title=paste0("Lebenserwartung in ", which.country, " im Zeitverlauf"),
+          subtitle=paste0("Kontinent: ", tmp.data$Region[1]))
+
+   if(show.mean==TRUE){
+     tmp.plot ← tmp.plot +
+       geom_hline(aes(yintercept=tmp.mw), lty=2, col=dipfblau) +
+       annotate("text", 1960, tmp.mw+3, fontface='italic',
+                label=paste0("Mittelwert = ", round(tmp.mw, 2)), col=dipfblau)
+   }
+   return(tmp.plot)
+ }
```

- Notwendiges Argument: Angabe des Landes
- Optionales Argument: ob Mittelwert eingezeichnet werden soll (Default=nein)

Nur falls Argument *show.mean=TRUE*

Eine Serie gleicher Grafiken mit Funktion und Schleife erzeugen

- Beispiel: es soll ein Diagramm für die Lebenserwartung im Zeitverlauf für jedes Land im Datensatz `edu_exp` erzeugt und gespeichert werden
  - Teil 2: Einrichten eines neuen Ordners, Erstellen eines Vektors aller Länder:

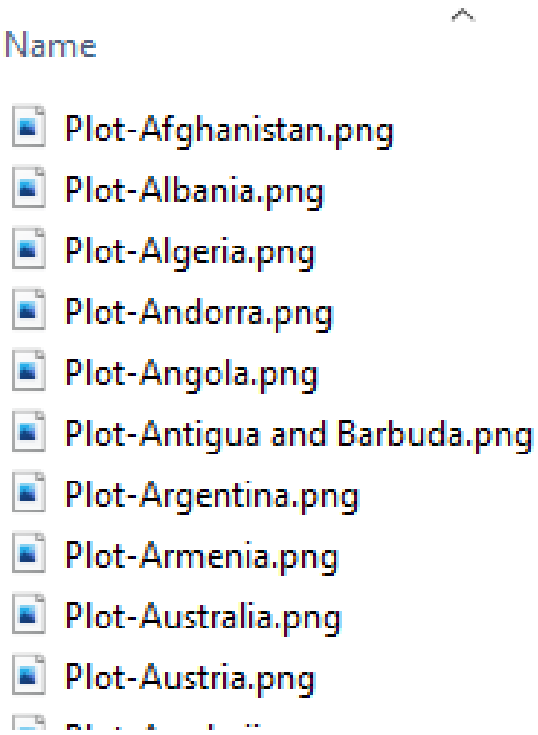
```
> dir.create("./gapminder-Laender-Plots")
> countries ← unique(edu_exp$Country)
> countries
  [1] "Afghanistan"                "Angola"                     "Albania"
  [5] "United Arab Emirates"       "Argentina"                  "Armenia"
  [9] "Australia"                  "Austria"                    "Azerbaijan"
 [13] "Belgium"                    "Benin"                      "Burkina Faso"
 [17] "Bulgaria"                   "Bahrain"                    "Bahamas"
```

  - Teil 3: Anwenden der Funktion per `for()`-Schleife auf alle Länder (Elemente des Vektors „countries“), dabei den Mittelwert einzeichnen lassen, speichern

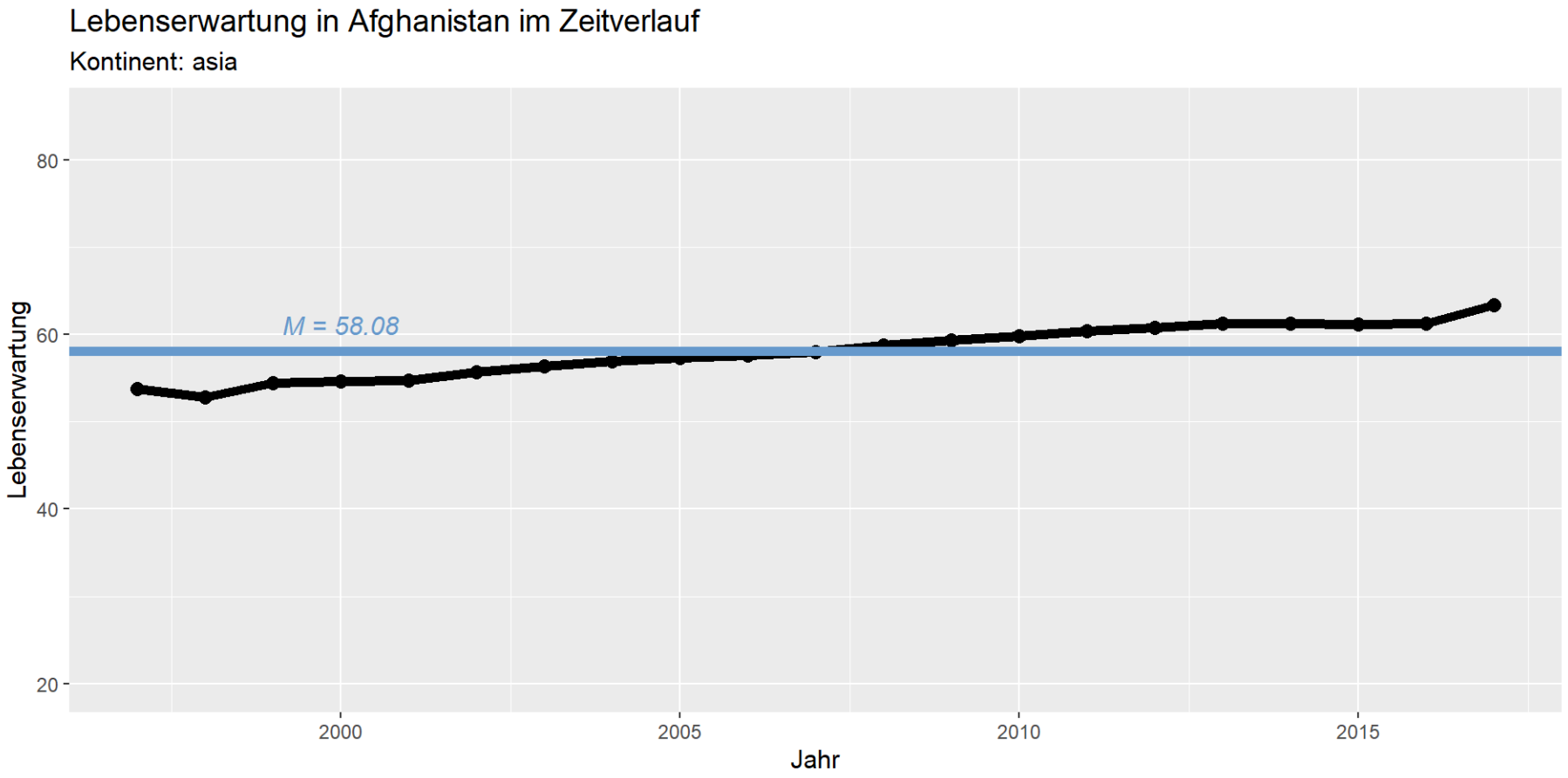
```
> for(c in 1:length(countries)){
+   gm.plot(countries[c], show.mean = TRUE)
+   ggsave(paste0("./gapminder-Laender-Plots/Plot-", countries[c], ".png"),
+                 width=24, height=12, units="cm", dpi=200)
+   print(paste0("Grafik erstellt fuer: ", countries[c], "(", c, "/", length(countries), ")"))
+ }
[1] "Grafik erstellt fuer: Afghanistan(1/142)"
[1] "Grafik erstellt fuer: Albania(2/142)"
[1] "Grafik erstellt fuer: Algeria(3/142)"
[1] "Grafik erstellt fuer: Angola(4/142)"
[1] "Grafik erstellt fuer: Argentina(5/142)"
[1] "Grafik erstellt fuer: Australia(6/142)"
```

Eine Serie gleicher Grafiken mit Funktion und Schleife erzeugen

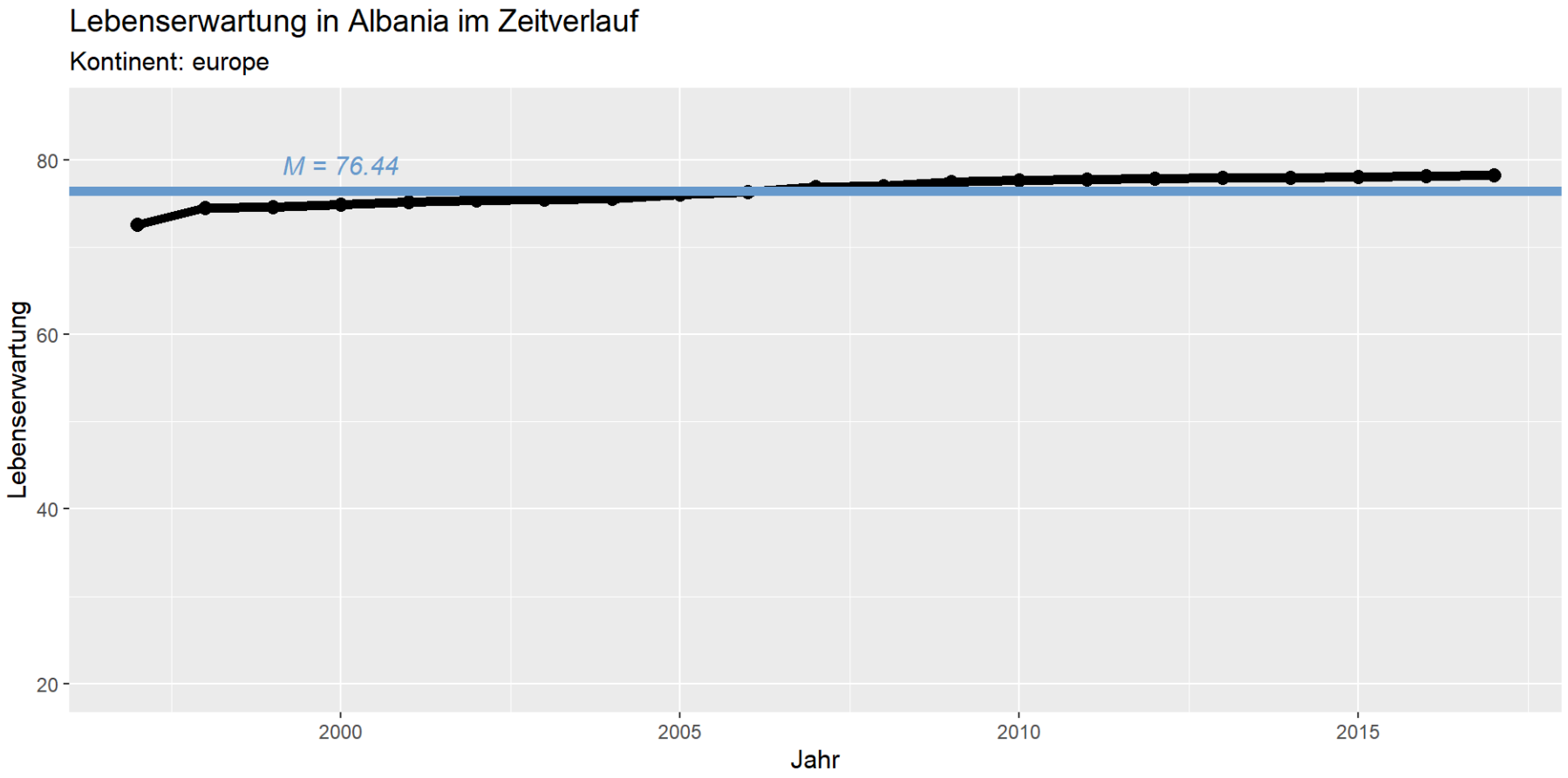
- Ergebnis:
  - Es wurde ein Ordner erzeugt
    (Funktion **`dir.create( '`***`Pfad/Name`***`' )`**),
  - Dieser enthält 197 Dateien
    (eine Grafik für jedes Element im Vektor „countries“)



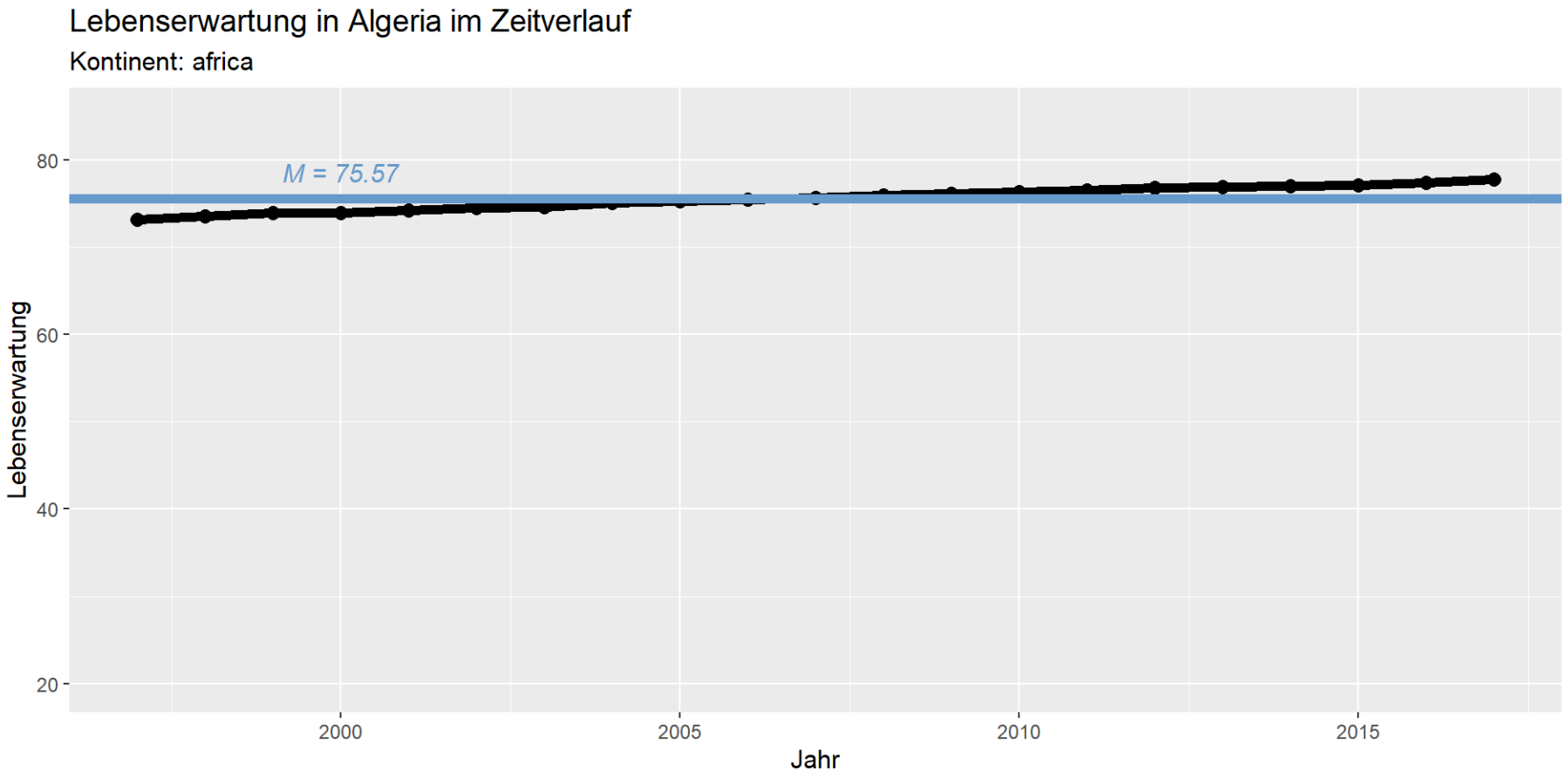
Eine Serie gleicher Grafiken mit Funktion und Schleife erzeugen



Eine Serie gleicher Grafiken mit Funktion und Schleife erzeugen



Eine Serie gleicher Grafiken mit Funktion und Schleife erzeugen

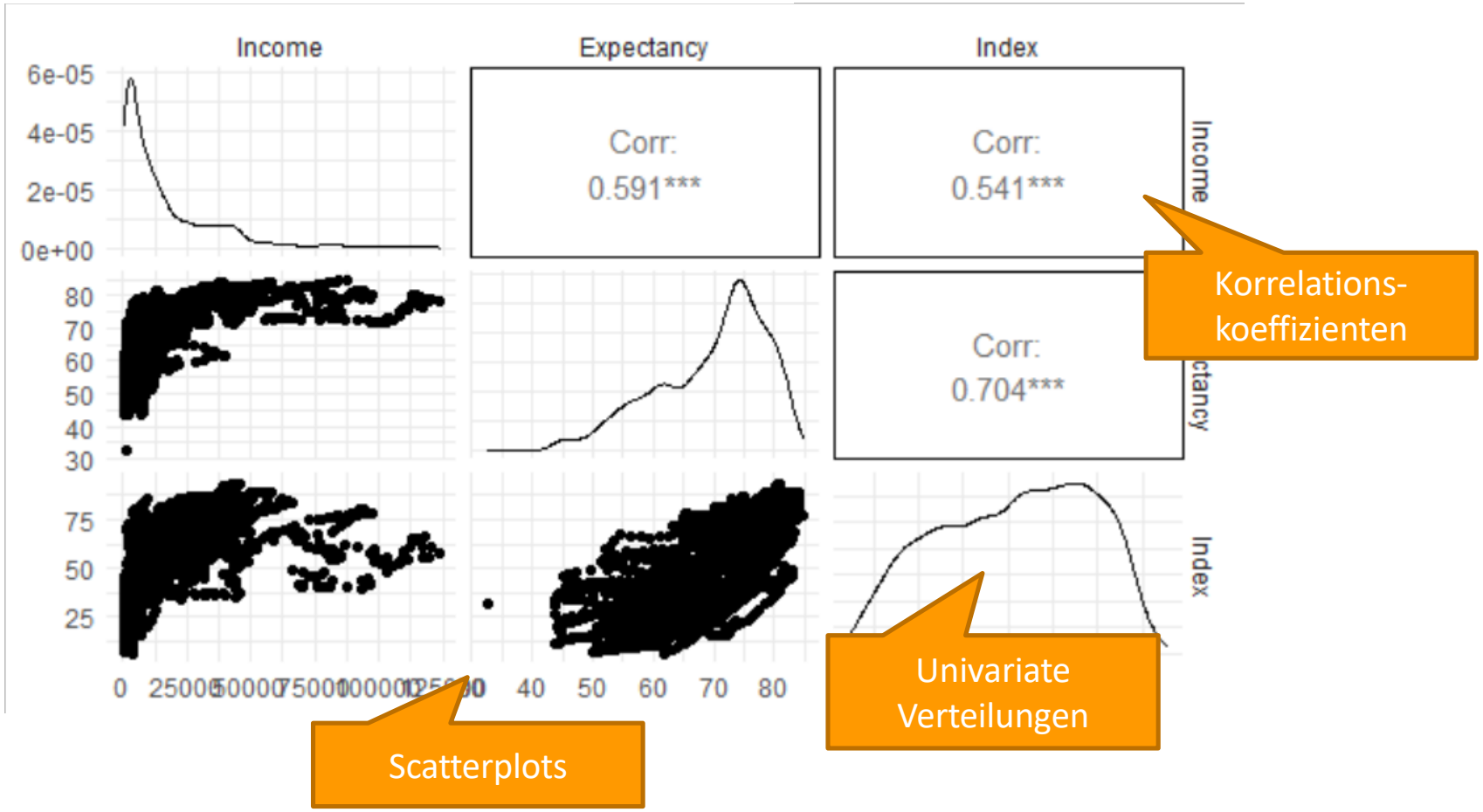


GGPAIRS:
VIELE BIVARIATE ZUSAMMENHÄNGE

Viele bivariate Zusammenhänge

- Funktion **ggpairs(** *Datensatz* **)** aus dem Paket "GGally", Default-Darstellung:

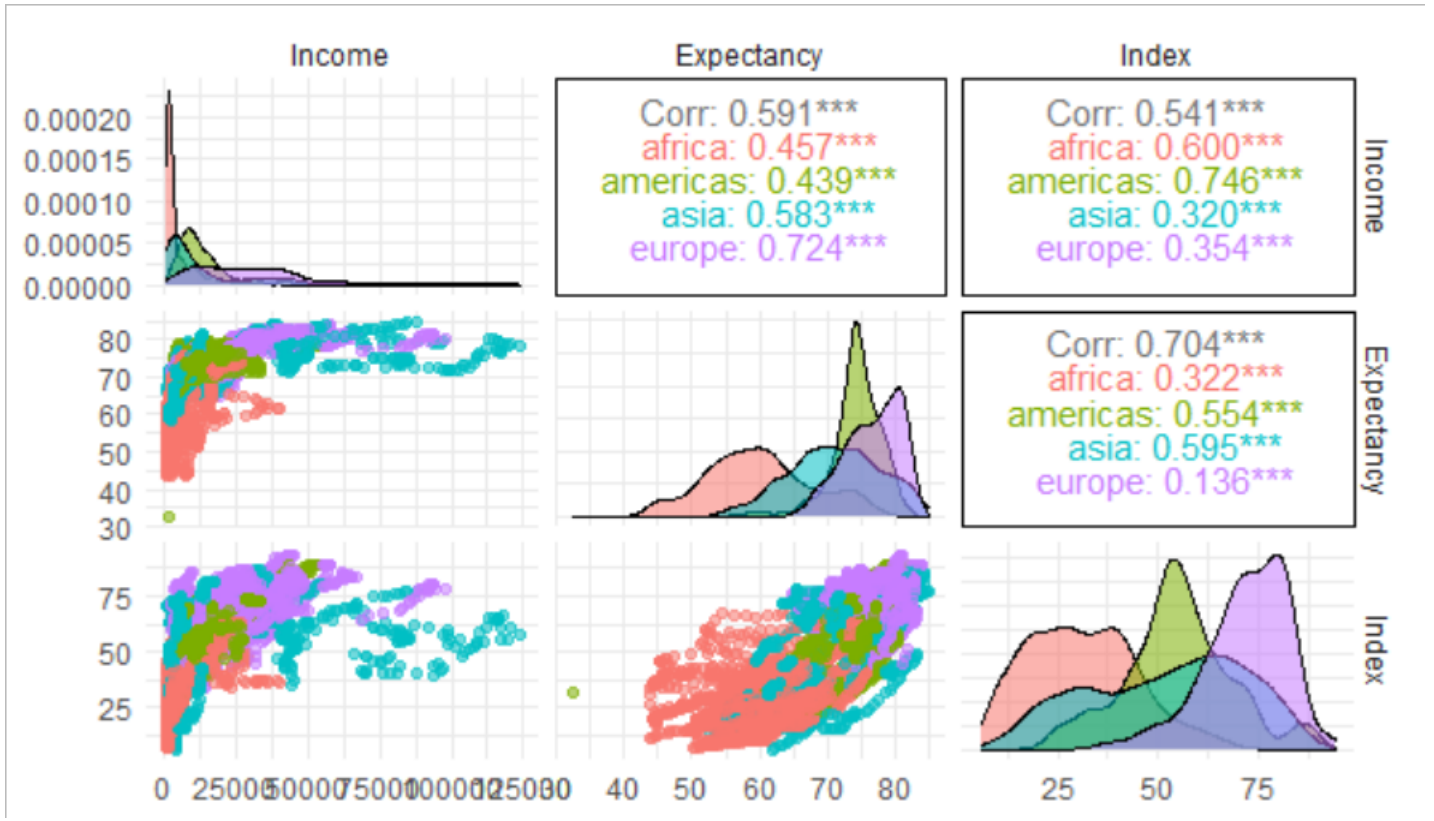
```
> ggpairs(edu_exp_sel, columns = 1:3)
```



Viele bivariate Zusammenhänge

- Bedingt auf die Gruppenzugehörigkeit (Variable „Species“)

```
> ggpairs(edu_exp_sel, columns = 1:3, aes(color = Region, alpha = .5))
```



Transparenz

TABLEPLOT:
MULTIVARIATE DATEN DURCHKÄMMEN

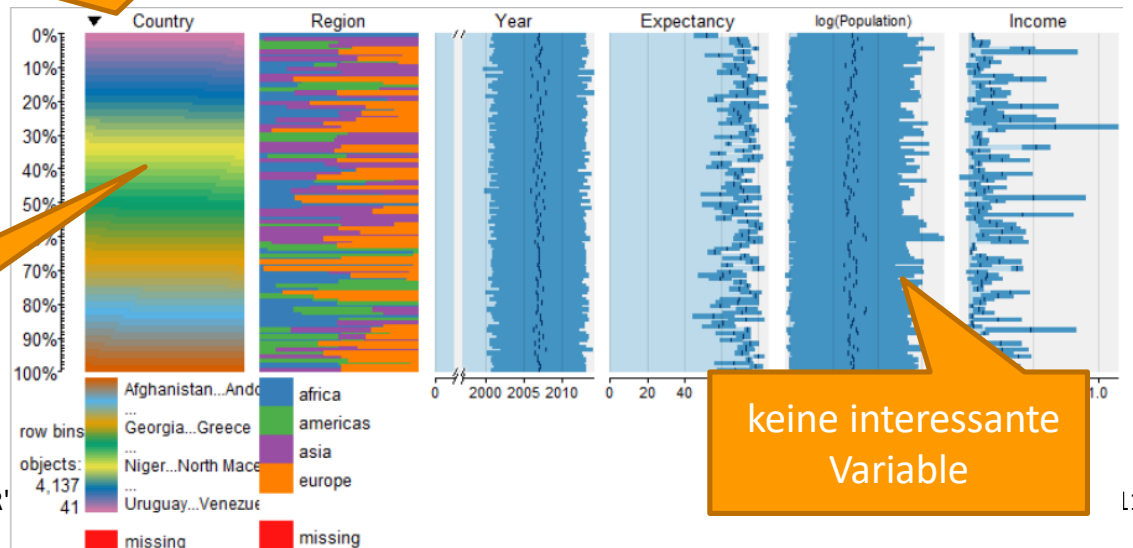
Tableplot

- Funktion **`tableplot(`** *`datensatz`* **`)`** aus dem Paket "tabplot"
- Visualisierung großer multivariater Datensätze
- Für jede Subgruppe (per Default: für jedes Quantil) auf einer Variablen werden Statistiken auf anderen Variablen dargestellt:
  - für kontinuierliche Variablen: Balkendiagramm der Mittelwerte
  - für kategoriale Variablen: Gestapeltes Balkendiagramm der Häufigkeiten einzelner Kategorien

```
> library(tabplot)
> tableplot(edu_exp_2)
```

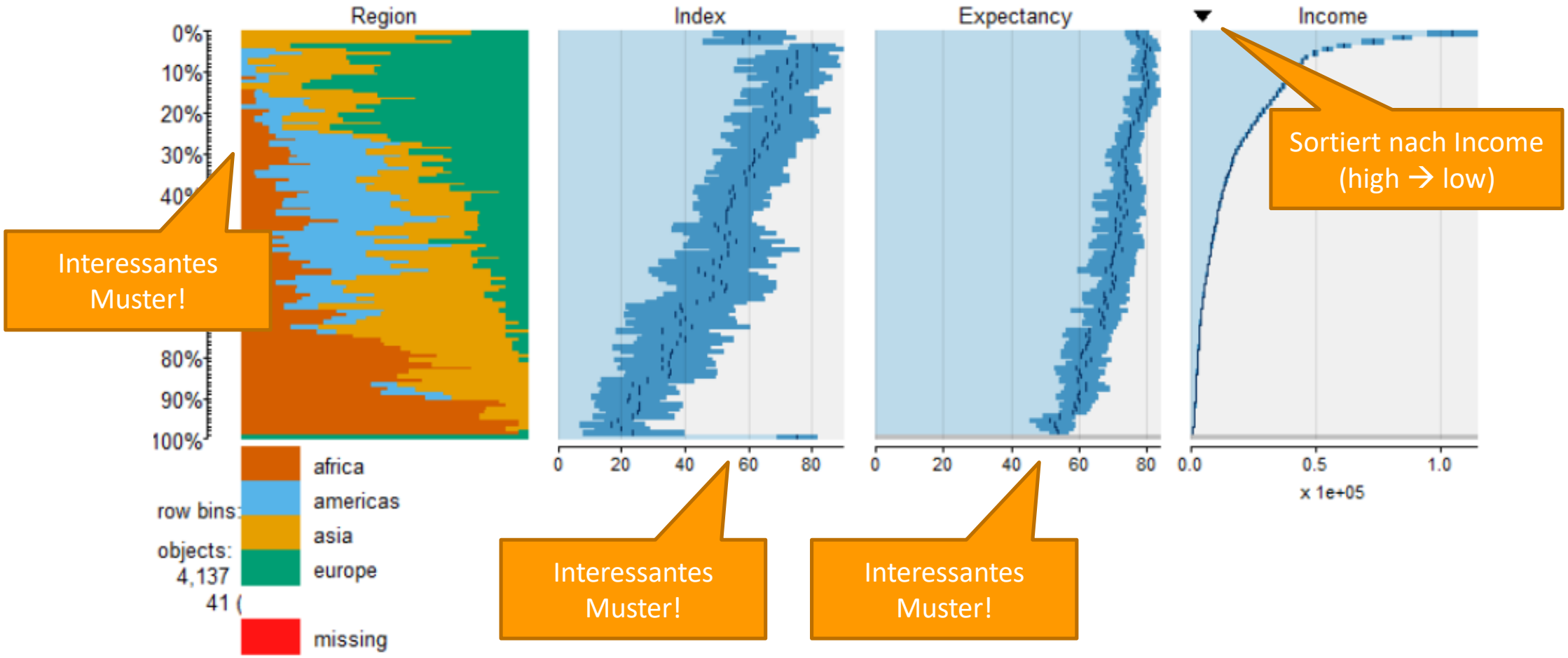
keine sinnvolle Sortierung

keine interessante Variable



keine interessante Variable

```
> tableplot(edu_exp_2,
+           select=c(Region, Index, Expectancy, Income),
+           sortCol = Income)
```



BLOCK 10:
ÜBUNG 2

Übung 2: Eigene Grafik erzeugen

- Nutzen Sie den `edu_gap`-Datensatz und erzeugen Sie eine Grafik eines zufällig ausgewählten Typs, die ein zufällig ausgewähltes Ziel bestmöglich erfüllt.
- Im Skript „12_Übung2.R“ finden Sie eine Funktion, die die Zufallsauswahl für Sie vornimmt.

Ziel der Grafik
Explorativ
Wissenschaftlicher Vortrag
Twitter-Feed
Manuskript (Schwarz-/Weiß-Druck)
Wissenschaftliches Poster

Grafik-Typ
Balkendiagramm
Histogramm
Violin Plot
Boxplot
Scatterplot
Liniendiagramm/Zeitreihe

BLOCK 11:
OUTRO: TIPPS, TRICKS UND ABSCHLUSS

Der wichtigste Tipp überhaupt

Ruhe bewahren! ☺

- Alles geht – man muss nur die richtige Funktion oder das richtige Argument finden. (Überblick über diverse Ressourcen: siehe Block 3)

- Die Community ist groß – insbesondere mit englischen Schlagworten ist Google sehr hilfreich.

Linktipps

- Extrem (!) gute Sammlung an Links:
  https://cogtales.wordpress.com/2018/07/27/data-visualization-the-why-and-how/

- Umfassende Sammlung an Grafiken und R-Code:
  https://www.data-to-viz.com/

- Reddit: Data is beautiful
  https://www.reddit.com/r/dataisbeautiful/

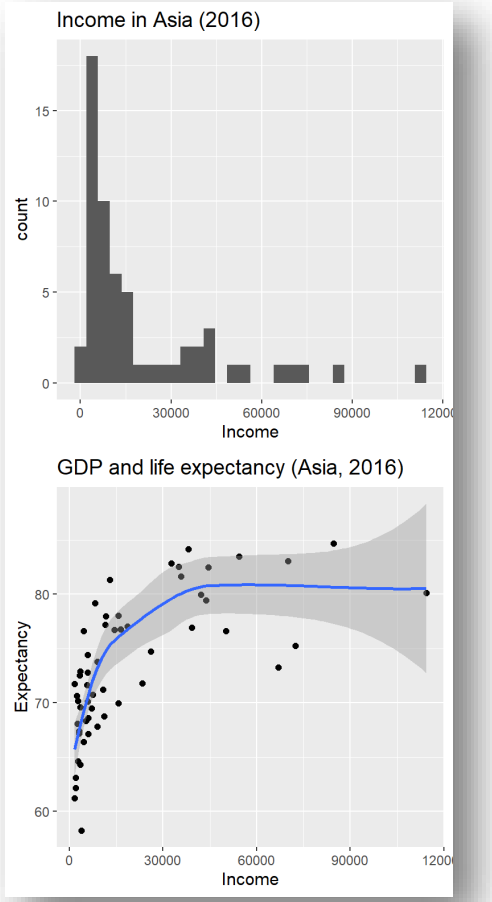
- Sehr spannende Visualisierungen:
  https://pudding.cool/

GRAFIKEN ANORDNEN

Grafiken anordnen

- **`grid.arrange(`** *`plot.objekt1 , plot.objekt2 , … , nrow = … , ncol = …`* **`)`**
  - aus dem Paket „gridExtra“
  - eine einzige Datei bestehend aus mehreren Grafiken
  - dafür ein Raster (*grid*) aus Zeilen (*nrow*) und Spalten (*ncol*) definieren:

```
> p.histo ← edu_exp ▷
+   subset(Region=="asia" & Year == 2016) ▷
+   ggplot(aes(x=Income)) +
+   geom_histogram() +
+   labs(title = "Income in Asia (2016)")
> p.scatter ← edu_exp ▷
+   subset(Region=="asia" & Year == 2016) ▷
+   ggplot(aes(x=Income, y=Expectancy)) +
+   geom_point() +
+   geom_smooth() +
+   labs(title = "GDP and life expectancy (Asia, 2016)")
> grid.arrange(p.histo, p.scatter, ncol=2)
```



  - Speichern: **`ggsave(`** *`"name.png" , plot = grid.arrange ()`* **`)`**

```
> ggsave("C:/Workshops/R/zwei-scatterplots.png",
+        plot=grid.arrange(p.histo, p.scatter, ncol=1))
```

ÜBERLAPPENDE DATENPUNKTE VERMEIDEN

Überlappende Datenpunkte vermeiden

- In Scatterplots für diskrete Variablen oder mit sehr großen *N* kann es dazu kommen, dass Datenpunkte *übereinander* liegen

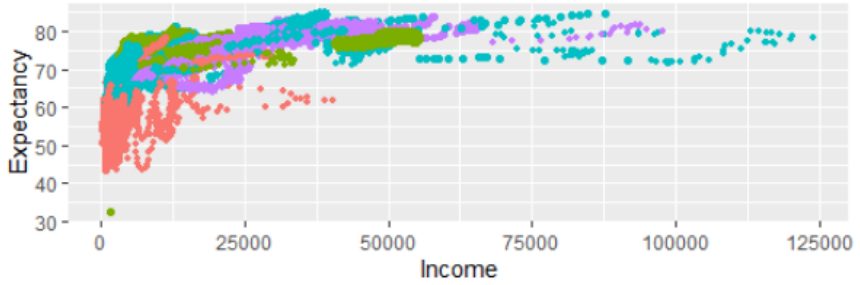
- Die Punktewolke kann somit nur schwer erfasst werden

- Zwei *technische* Möglichkeiten, damit in ggplot umzugehen
  - Transparenz der Datenpunkte: Argument `alpha`
  - Hinzufügen von „*noise*“: Funktion `geom_jitter()`

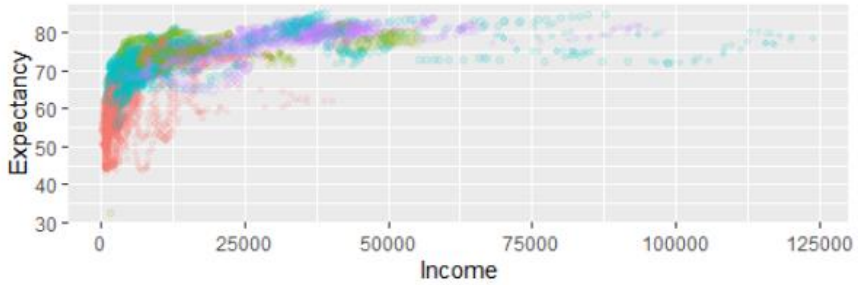
Überlappende Datenpunkte vermeiden

- Transparenz der Datenpunkte: Argument `alpha`

```
> ggplot(edu_exp, aes(x=Income, y=Expectancy)) +
+   geom_point(aes(colour = Region, size = Population), show.legend = FALSE)
```



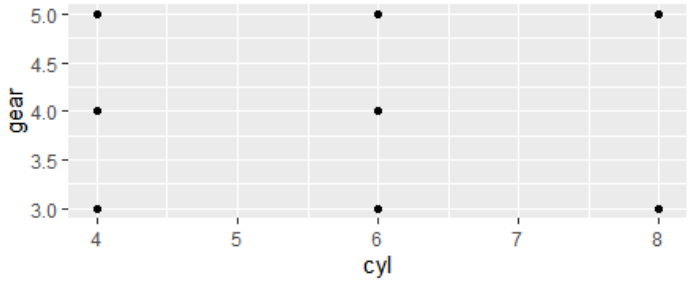

```
> ggplot(edu_exp, aes(x=Income, y=Expectancy)) +
+   geom_point(aes(colour = Region, size = Population), show.legend = FALSE, alpha = 0.1)
```



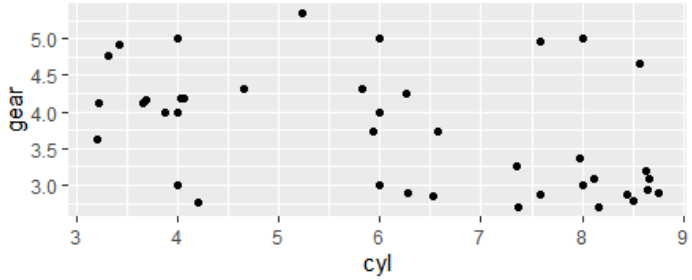
Überlappende Datenpunkte vermeiden

- Hinzufügen von „*noise*“: Funktion **`geom_jitter()`**

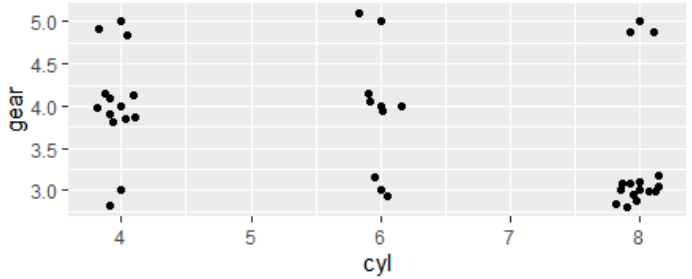
```
> ggplot(mtcars, aes(x=cyl, y=gear)) +
+   geom_point()
```




```
> ggplot(mtcars, aes(x=cyl, y=gear)) +
+   geom_point() +
+   geom_jitter()
```



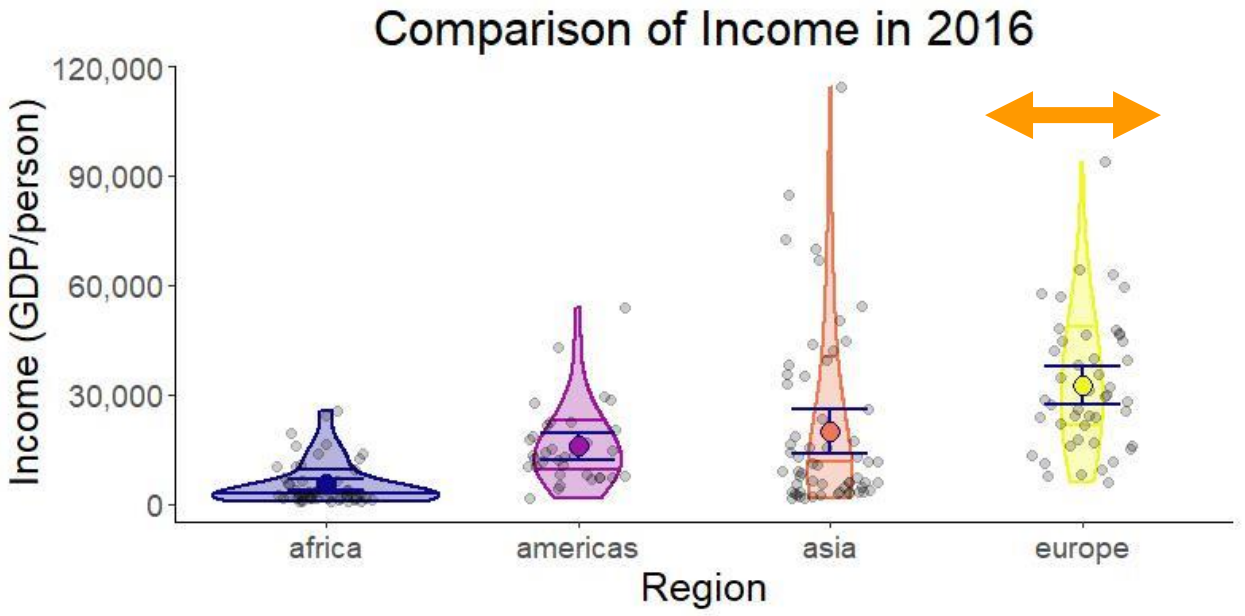

```
> ggplot(mtcars, aes(x=cyl, y=gear)) +
+   geom_point() +
+   geom_jitter(width=.2, height=.2)
```



Überlappende Datenpunkte vermeiden

- Hinzufügen von „*noise*“:
  Funktion **`geom_jitter()`**

```
+   geom_jitter(
+     data = usedat,
+     aes(x = order, y = Income),
+     height = 0, width = .2,
+     size = 2, alpha=.2, fill= plasma(1)) +
```



jitter bezieht sich nur auf width (=x-Achse),
Variable auf der y-Achse ist unverändert

EFFIZIENTES ARBEITEN IM GGPLOT-BEFEHL

Effizientes Arbeiten im `ggplot`-Befehl

- Eine komplexe ggplot-Grafik erfordert oft Code über viele Zeilen
- Während des Überarbeitens Ihrer Grafik werden Sie wahrscheinlich einzelne Zeilen per `# Kommentar` „an- und ausschalten“ wollen
- Problem: Wird die letzte Zeile auskommentiert, kann der Befehl nicht abgeschlossen werden, weil das letzte `+` ins Leere führt

```
ggplot(edu_exp, aes(x=Income, y=Primary)) +
  geom_point() +
  ggtitle("Interesting Plot") +
  # theme_classic() +
```

- Lösung: Ergänzen Sie standardmäßig folgenden Befehl: `NULL`

```
ggplot(edu_exp, aes(x=Income, y=Primary)) +
  geom_point() +
  ggtitle("Interesting Plot") +
  # theme_classic() +
  NULL
```

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

DIPF Bildungsforschung und Bildungsinformation