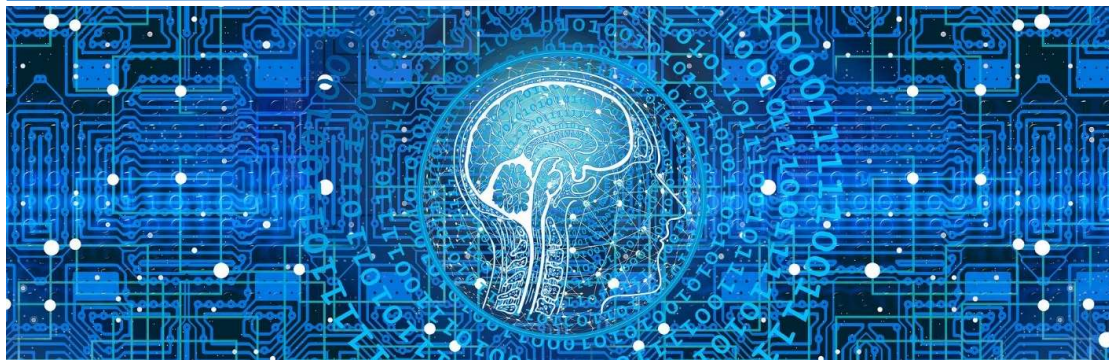


Data Science

3. Teil – Visualisierung und Plotten

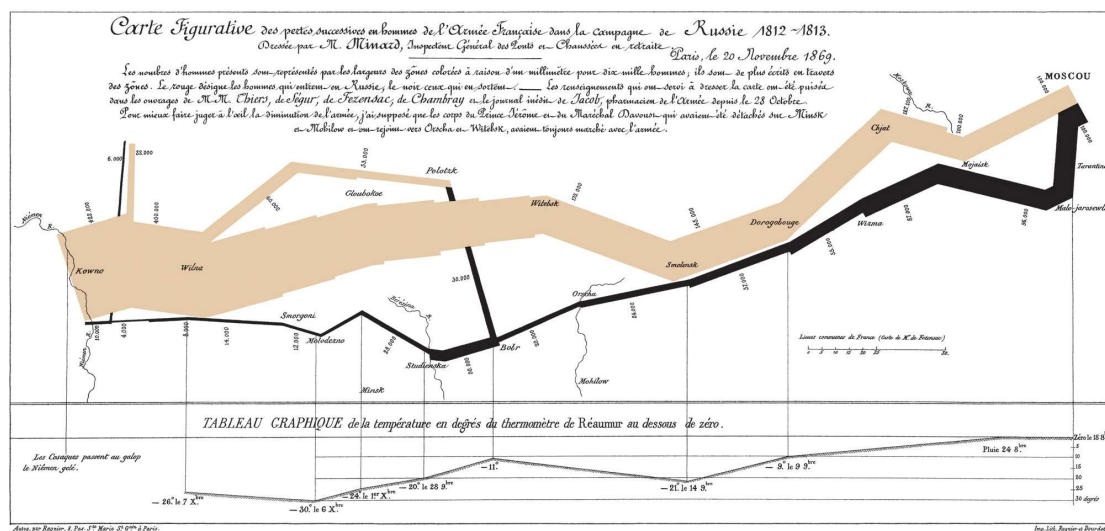
Vorlesung an der DHBW Stuttgart, Prof. Dr. Monika Kochanowski



Data Science Vorlesung | DHBW Stuttgart | Prof. Dr. Monika Kochanowski | 1

Charles Joseph Minard

Infografik von 1869



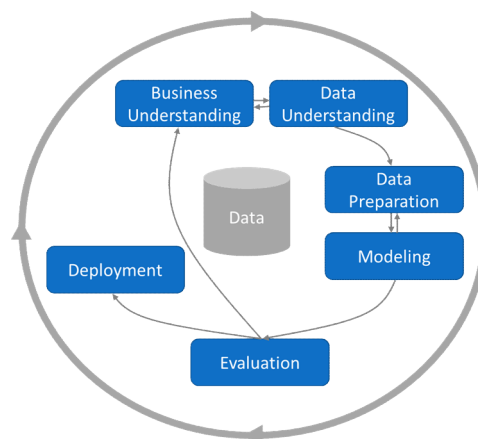
Data Science Vorlesung | DHBW Stuttgart | Prof. Dr. Monika Kochanowski | 2

Inhalte der heutigen Vorlesung

- Grundlagen visueller Exploration
- Informationsvisualisierung
- Einsatzzwecke
- Richtlinien
- Plotten
- Python – Übung Plotten



Wiederholung



Gedanken zu Informationsvisualisierung

- Zwei Hauptanwendungsfälle
 - **Exploration und Bestätigung**
 - Visuelle Exploration der Daten
 - **Erklärung**
 - Vermittlung von Informationen (und Wissen) anhand der Daten
- **Informationsvisualisierung** beschäftigt sich mit der Darstellung von abstrakten Daten (wie z. B. Betrugsquoten)
- **Wissenschaftliche Visualisierung** beschäftigt sich mit physischen Daten (z. B. Wellen visualisieren im Meer)

Diskussion: welche Zielgruppen haben die verschiedenen Anwendungen? Welche Erwartungshaltung haben diese? Welche Vorkenntnisse liegen vor?



Gedanken zu Informationsvisualisierung

- Zwei Hauptanwendungsfälle
 - **Exploration und Bestätigung**
 - Visuelle Exploration der Daten
 - Zielgruppe: DATA SCIENTIST
 - **Erklärung**
 - Vermittlung von Informationen (und Wissen) anhand der Daten
 - Zielgruppe: KUNDE
- **Informationsvisualisierung** beschäftigt sich mit der Darstellung von abstrakten Daten (wie z. B. Betrugsquoten)
- **Wissenschaftliche Visualisierung** beschäftigt sich mit physischen Daten (z. B. Wellen visualisieren im Meer)
 - Zielgruppe: WISSENSCHAFTLER

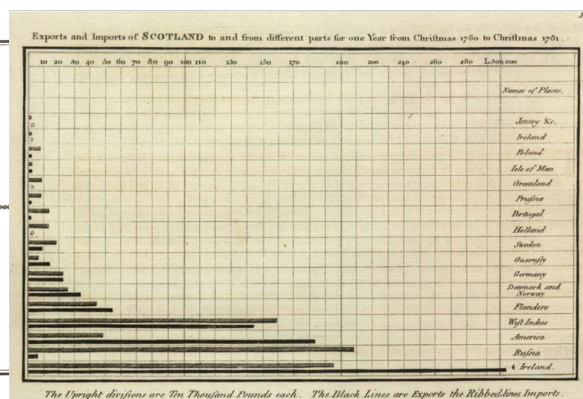
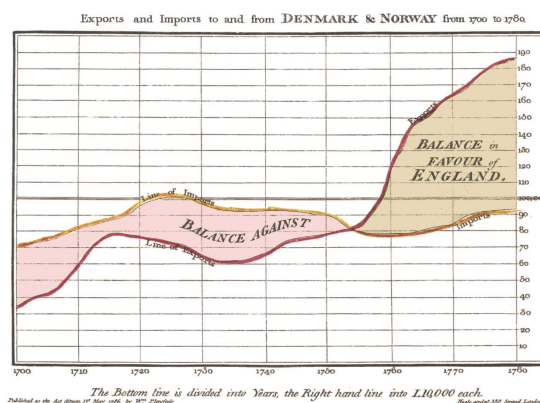


Erste Karte (9000 Jahre alt) Çatalhöyük

























Bildquelle: <https://www.zeit.de/wissen/geschichte/2014-01/karte-catalhoeyuek-vulkan-tuerkel>

Älteste Balkendiagramme Schottland 18. Jahrhundert



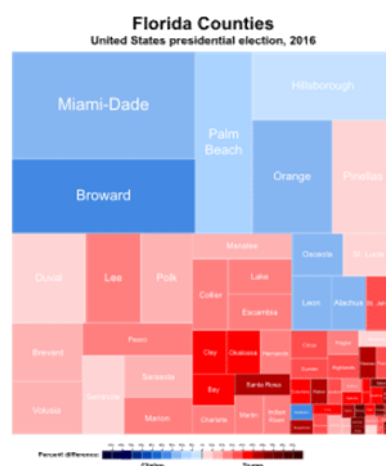
Bildquelle: https://en.wikipedia.org/wiki/William_Playfair

Moderne Ansichten Sparklines

17,172.48	11,760.20	9,951.90		5.57		1.89
1,290.77	1,440.00	639.05		-30.13		14.14
225.66	7,200.00	4,500.00		22.79		11.74
1,069.53	677.95	481.25		16.65		3.85
465.00	1,490.60	1,090.10		23.11		6.70
67.32	383.45	242.25		-26.65		0.94
765.39	2,994.00	1,511.20		47.54		9.79
781.69	332.75	232.35		-6.08		3.94
405.38	4,599.90	3,260.45		-4.99		1.56
131.91	542.50	238.55		-31.52		11.63
261.63	993.00	694.00		-48.35		0.58

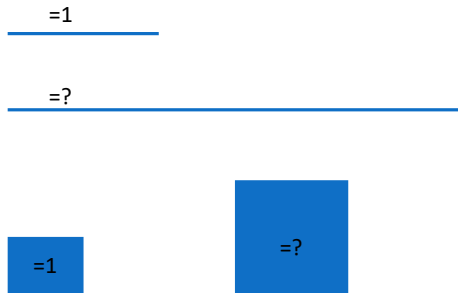
Begriff Sparkline: Edward Tufte (2006). Beautiful Evidence. Graphics Press. [ISBN 0-9613921-7-7](https://www.amazon.de/dp/0961392177).
Bildquelle: <https://stackoverflow.com/questions/53852739/creating-sparkline-chart-inside-table-in-data-studio>

Moderne Ansichten Treemaps



<https://en.wikipedia.org/wiki/Treemapping>

Kleine Übung



- Menschen können generell Längen besser schätzen als Flächen
- Diagramme sollten meistens **Länge statt Fläche** nutzen
- Winkel sind ebenfalls schwer zu schätzen (Stichwort Kreisdiagramme)
- Balken- und Liniendiagramme bieten hier Vorteile

Gedanken zu Informationsvisualisierung

- Zwei Hauptanwendungsfälle
 - **Exploration und Bestätigung**
 - Visuelle Exploration der Daten
 - **Erklärung**
 - Vermittlung von Informationen (und Wissen) anhand der Daten
- **Informationsvisualisierung** beschäftigt sich mit der Darstellung von abstrakten Daten (wie z. B. Betrugsquoten)
- **Wissenschaftliche Visualisierung** beschäftigt sich mit physischen Daten (z. B. Wellen visualisieren im Meer)

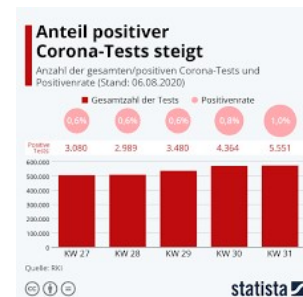
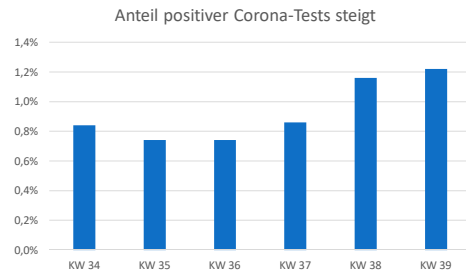
Daten genau anschauen
Details können eine Rolle spielen
Vorwissen vorhanden

„Message“ als wichtigster Inhalt
Verständlichkeit
Übersichtlichkeit

Ziel: Erklärung Ein Anwendungsfall der Visualisierung

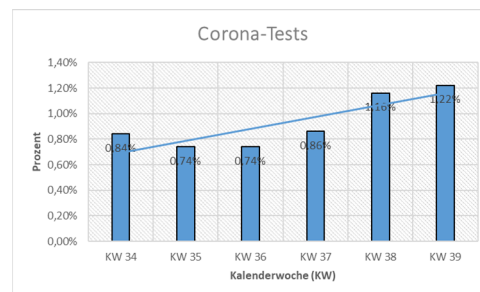
- Titel – soll eine Aussage transportieren (wir erklären etwas)
- Achsenbeschriftungen & Intervalle klug wählen
- Einheiten angeben
- Keine 3D Effekte oder sonstige Hervorhebungen
- Einfache Grafiken sind besser als komplexe
- Achsen
 - Korrekte Intervalle
 - Einheiten müssen erkennbar sein
 - Überflüssige Information vermeiden
- Auch in Grautönen zu lesen
 - Noch besser: Barrierefrei (rot-grün vermeiden)
 - Wenn Farbe dann muss diese eine Bedeutung haben!

Bild unten: <https://de.statista.com/infografik/22496/anzahl-der-gesamten-positiven-corona-tests-und-positivenrate/>



Data Science Vorlesung | DHBW Stuttgart | Prof. Dr. Monika Kochanowski | 14

Anti-Beispiel

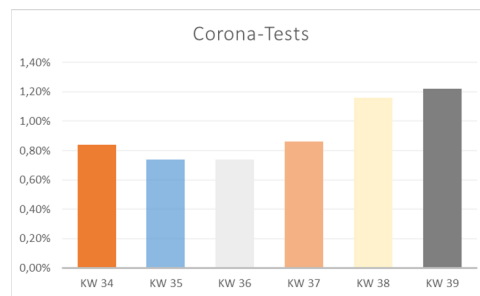


Diskussion

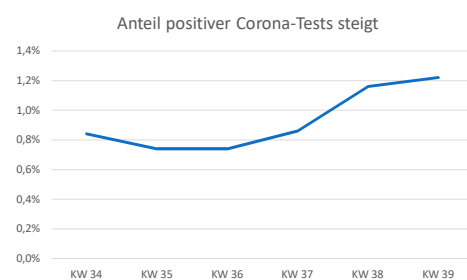
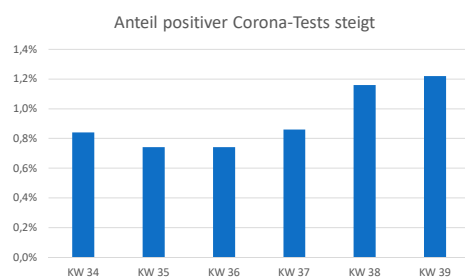


Data Science Vorlesung | DHBW Stuttgart | Prof. Dr. Monika Kochanowski | 15

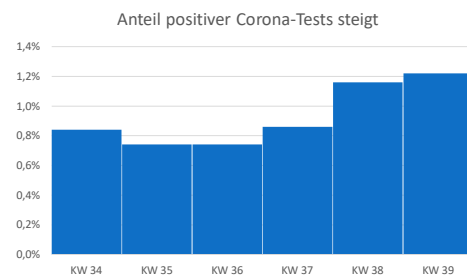
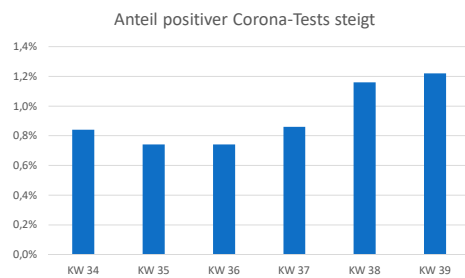
Anti-Beispiel 2



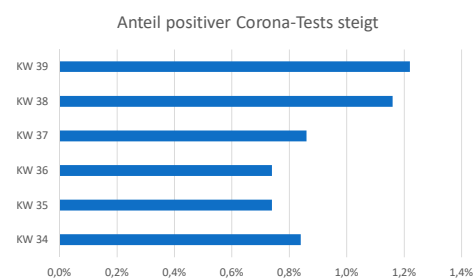
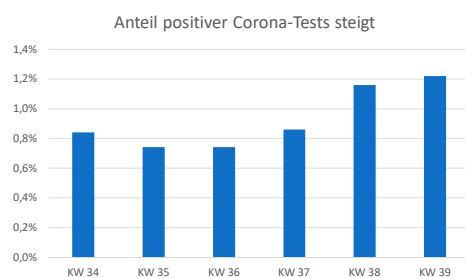
Balken oder Linien?



Abstand in Balkendiagrammen



Horizontal oder vertikal?

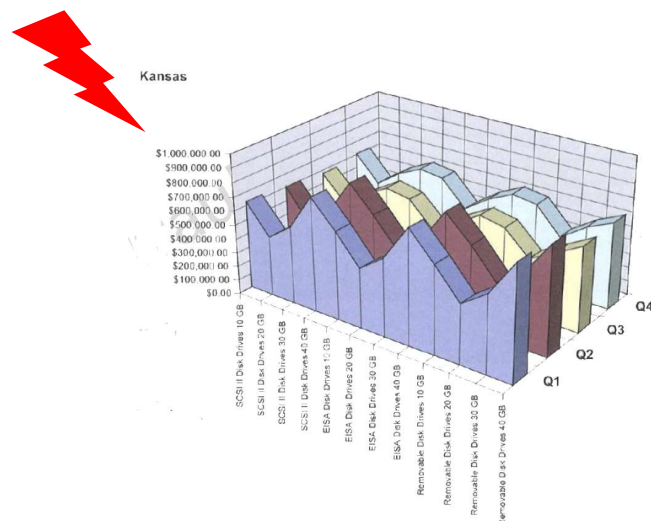


Die Sache mit den Farben

- Idealerweise ist Ihre Grafik in Graustufen lesbar
- Wenn Farben dann mit Bedeutung!
- Rot-Grün-Kontraste vermeiden
 - Einfarbig wo möglich/sinnvoll
 - Andere Skalen verwenden
 - Form
 - Schraffierung

0,0084	0,0084
0,0074	0,0074
0,0074	0,0074
0,0086	0,0086
0,0116	0,0116
0,0122	0,0122


Beispielgrafik zur Verbesserung




Abgabe über Chat

Formatierung von Tabellen

- Schattierung ist besser als Linien
- Abstand ist besser als Schattierung
- Vergleichbarkeit
- Symbole zentriert
- Text linksbündig
- Zahlen
 - Format gleich
 - Rechtsbündig
 - Überschrift analog
- Inhalt IMMER links nach rechts



Name	Anton	Max	Miriam
Matrikelnr.	123	234	342
Anwesend	x		x
Durchschnitt	1,2	1,0	2,0



Name	Matrikelnr.	Anwesend	Durchschnitt
Anton	123	x	1,2
Max	234		1
Miriam	342	x	2
Jonas	1234	x	3,1
Julia	1212	x	3,2
Sabine	1212		1

Name	Matrikelnr.	Anwesend	Durchschnitt
Anton	123	x	1,2
Max	234		1,0
Miriam	342	x	2,0
Jonas	1234	x	3,1
Julia	1212	x	3,2
Sabine	1212		1,0

Name	Matrikelnr.	Anwesend	Durchschnitt
Anton	123	x	1,2
Max	234		1,0
Miriam	342	x	2,0
Jonas	1234	x	3,1
Julia	1212	x	3,2
Sabine	1212		1,0

Beispiele für weitere Visualisierung Animation, Verdichtung, Interaktion, ..

- Interaktive Datenexploration
 - <https://sense-demo.qlik.com/sso/sense/app/1413a7f0-a1cb-4009-9395-c3b4ae0d0c62/sheet/VXmZj/state/analysis>
- Animation und Visualisierung von Zusammenhängen
 - <http://www.randalolson.com/2015/08/23/small-multiples-vs-animated-gifs-for-showing-changes-in-fertility-rates-over-time/>
- Verdichtung von Information über parallele Koordinaten
 - <https://www.csc.kth.se/~weinkauf/gallery/catpalmas14a.html>
- Visualisierung und Big Data
 - Transformieren, Reduzieren, Animieren, Verdichten, Interaktion
- 1. Overview 2. Zoom & Filter 3. Details-on-Demand
 - Am Beispiel Google Maps

Ziel: Exploration Ein Anwendungsfall der Visualisierung

- Titel – soll eine Aussage transportieren (wir erklären etwas)
- Achsenbeschriftungen & Intervalle klug wählen
- Einheiten angeben
- Keine 3D Effekte oder sonstige Hervorhebungen
- Einfache Grafiken sind besser als komplexe
- Achsen
 - Korrekte Intervalle
 - Einheiten müssen erkennbar sein
 - Überflüssige Information vermeiden
- Auch in Grautönen zu lesen
 - Noch besser: Barrierefrei (rot-grün vermeiden)

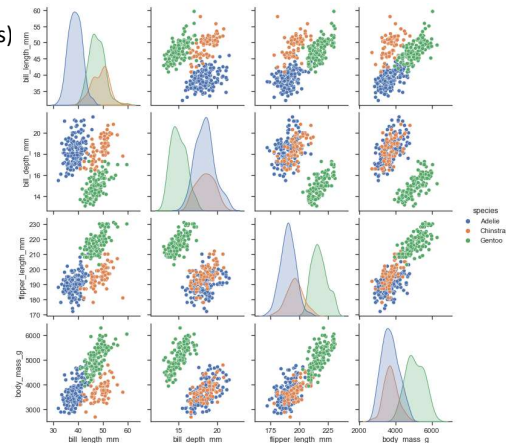
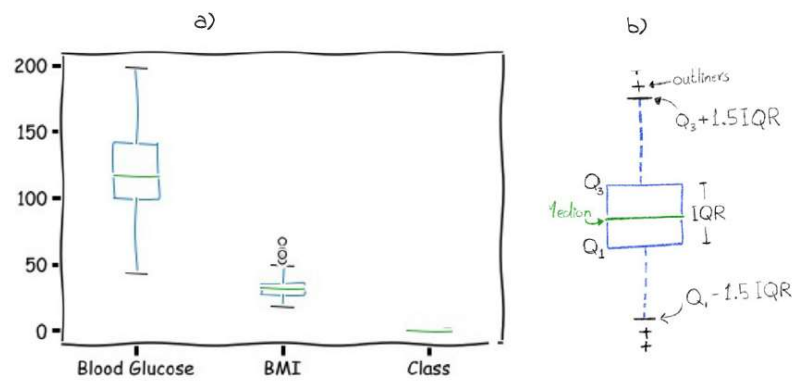


Bild unten: <https://de.statista.com/infografik/22496/anzahl-der-gesamten-positiven-corona-tests-und-positivenrate/>

Datentypen und -skalen

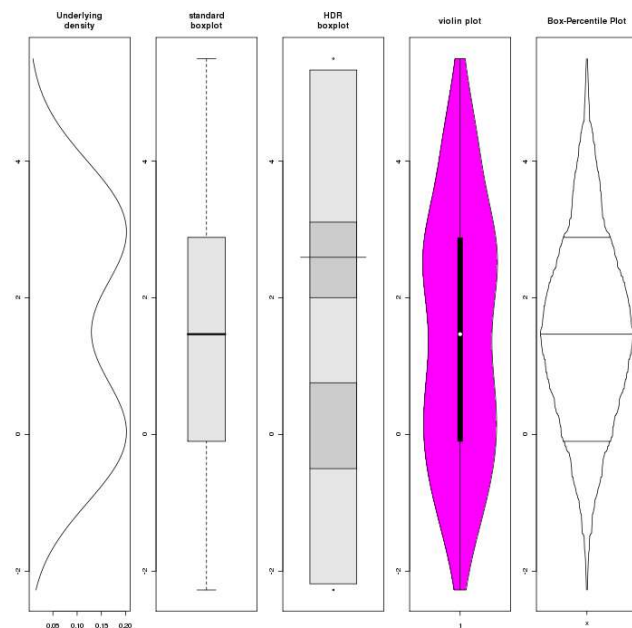
Skala	Alias	Mögliche Operationen	Beispiele und Aussagen
Nominalskala	Kategoriale Daten, qualitatives Merkmal	Gleichheit, Ungleichheit ($=$ / \neq), Häufigkeit (Modus)	Zweitstimme bei der Bundestagswahl, Geschlecht (gleiche Wahl wie ..)
Ordinalskala	Rangordnung	Ordnen möglich ($=$, \neq , $>$, $<$), Häufigkeit, Reihenfolge, Median	Likert-Skala (stimme zu, stimme eher zu, teils / teils, stimme eher nicht zu, lehne ab), Schulnoten (mehr / weniger als..)
Intervallskala (Kardinalskala)	Quantitative Merkmale, metrische Daten, numerical data	Abstände (Intervalle) besitzen eine Bedeutung ($=$, \neq , $>$, $<$, $+$, $-$, $*$, $\%$), Häufigkeit, Reihenfolge, Abstand, arith. Mittel	Temperatur (Celsius), Geburtsjahr (Unterschied ist..), IQ
Verhältnisskala (Kardinalskala)		Mit absolutem Nullpunkt, Häufigkeit, Reihenfolge, Abstand	Einkommen, Alter (doppelt so viel), Geschwindigkeit, Längen, Zeiten, ..

Boxplots Ausreißer und Verteilung visualisieren

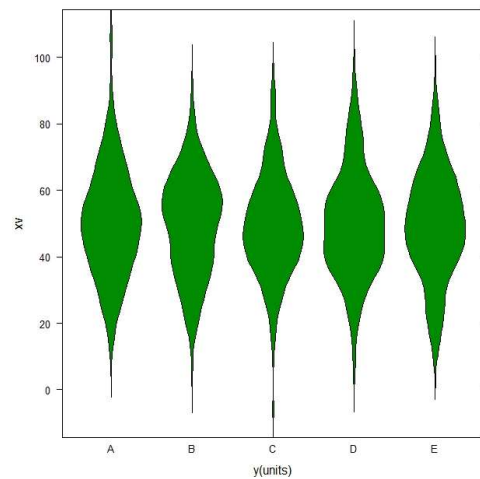


Bildquelle: *How Machine Learning Works*, M. S. Abd El-Fattah, Manning, 2020

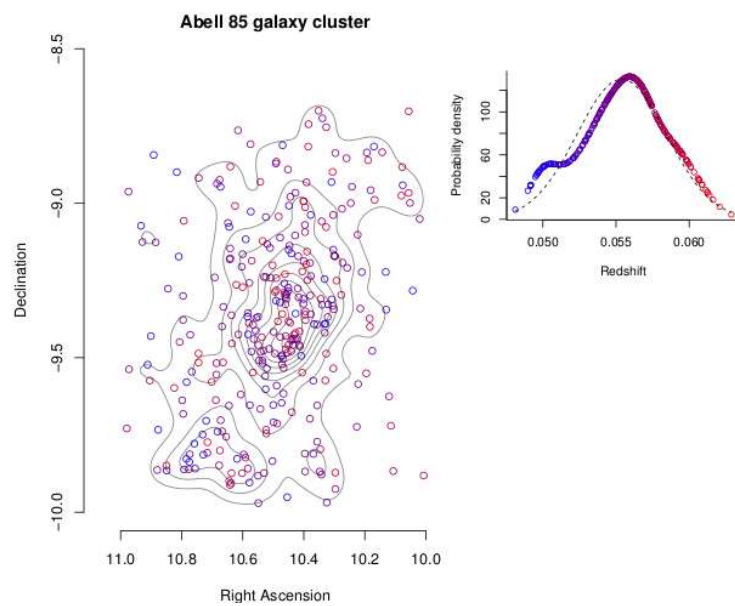
Wenn Boxplots nicht ausreichen...



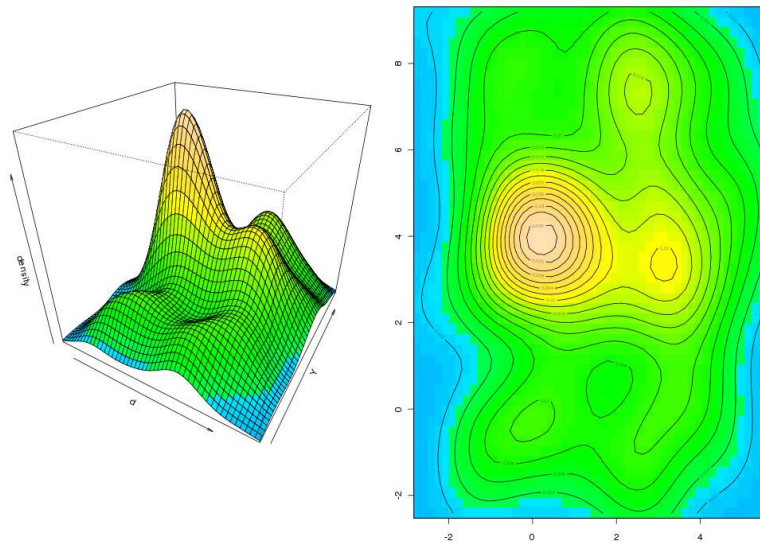
Violin Plots



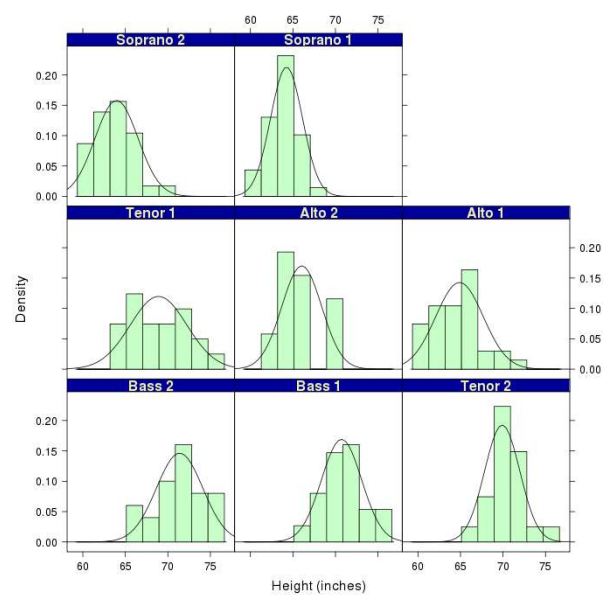
Mehrdimensionale Informationen



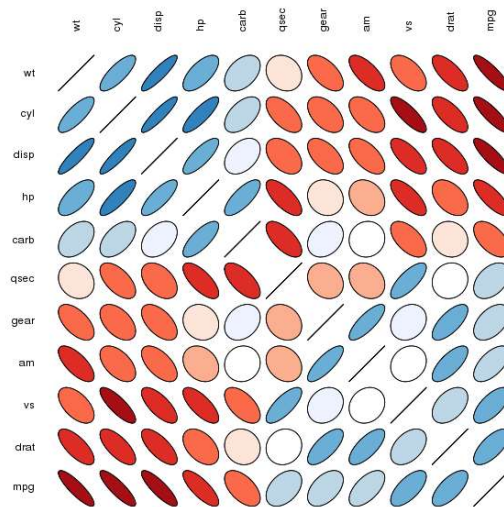
Mehrdimensionale Plots



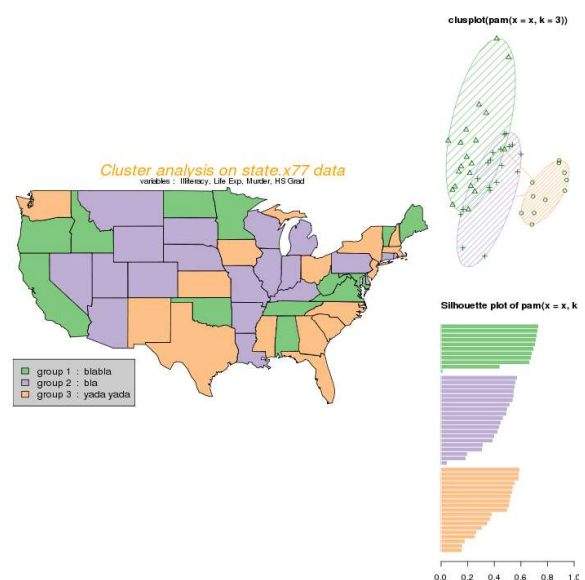
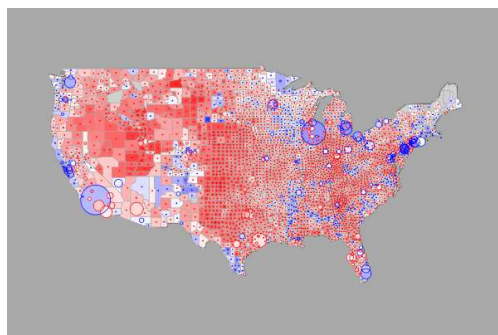
Dichte anzeigen



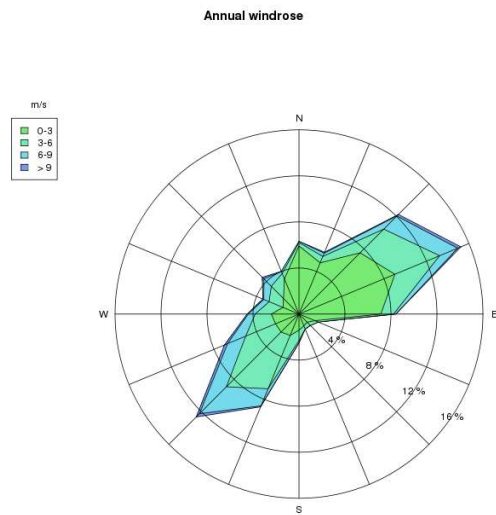
Korrelation



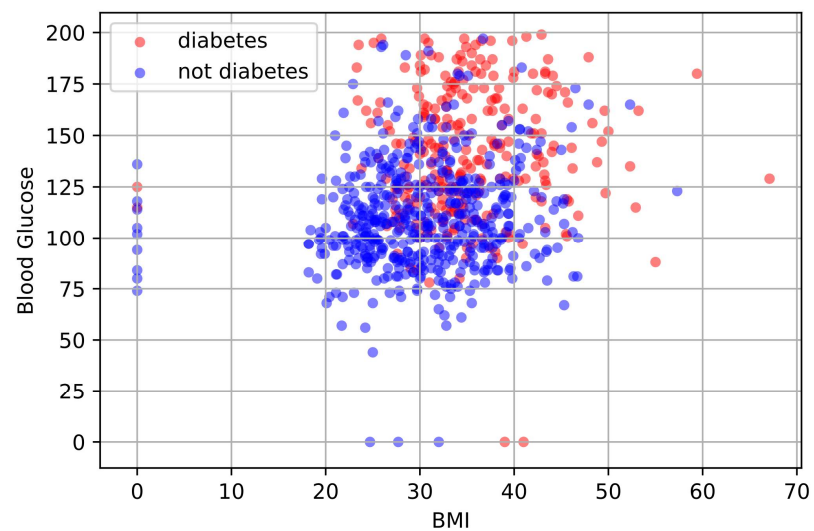
Kartendarstellungen



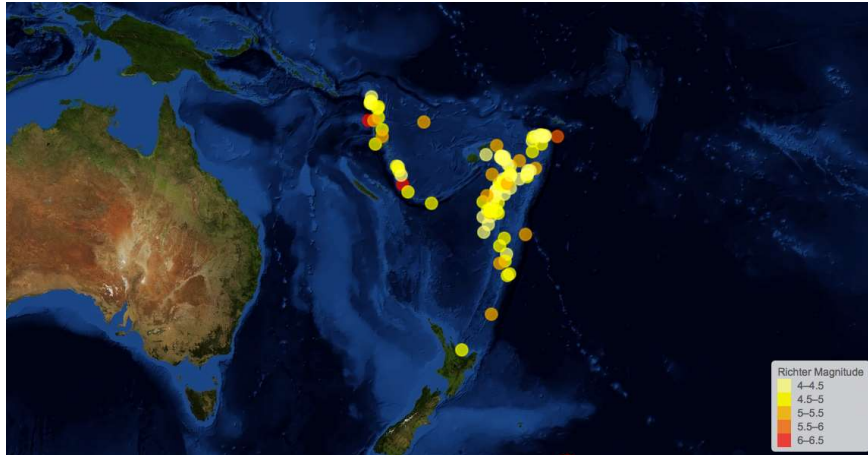
Mehr Dimensionen darstellen



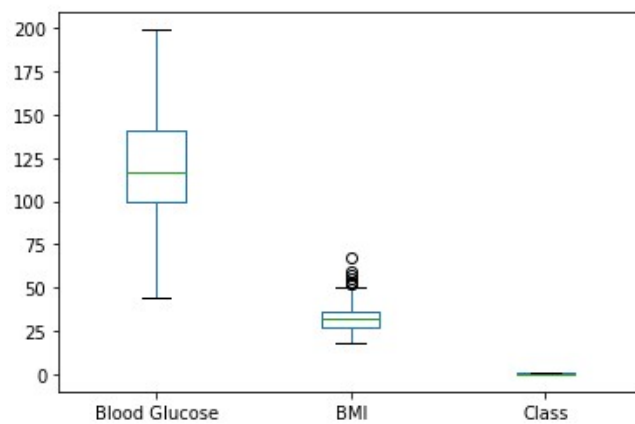
Diabetes



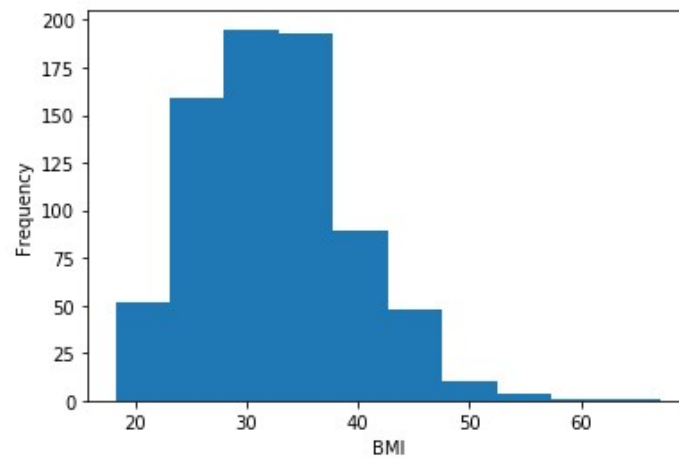
Wissenschaftliche Daten darstellen



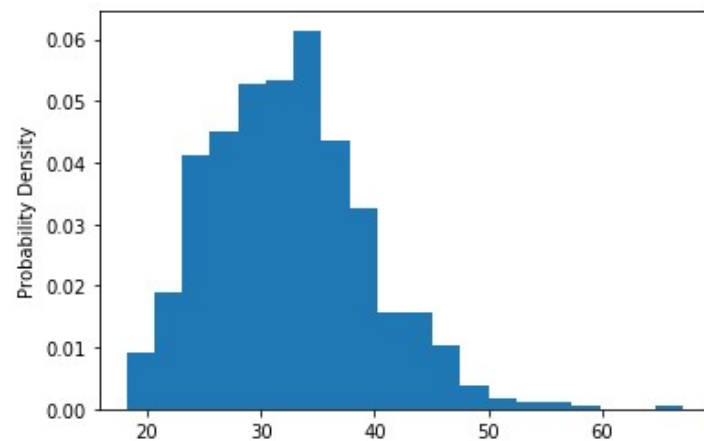
Diabetes



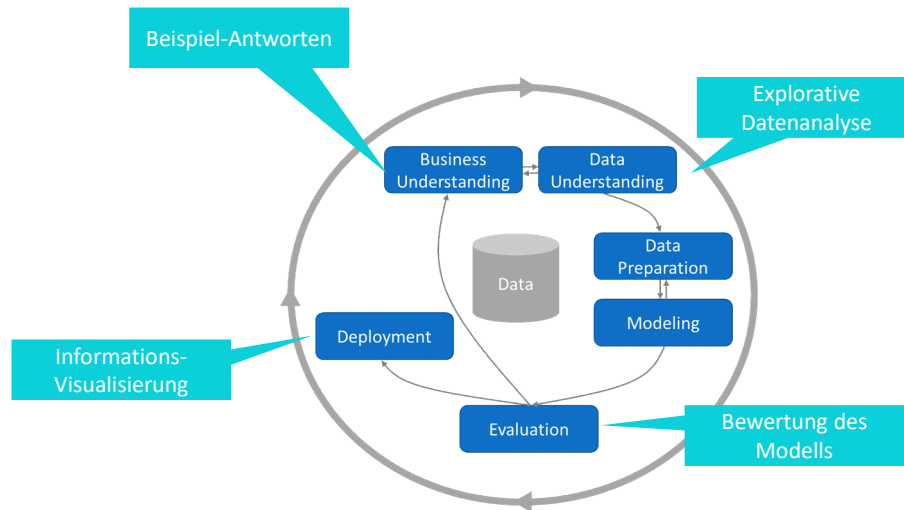
Diabetes



Diabetes



Wiederholung



Plotten in Python

- <https://seaborn.pydata.org/examples/index.html>
- <https://matplotlib.org/3.1.1/gallery/index.html>

Literaturquellen zu Visualisierung

- Online-Ressource zu Visualisierung
 - <https://www.visualisingdata.com/>
- Storytelling with Data [Buch]: Klassiker für Überzeugungsarbeit in Präsentationen von Ergebnissen
 - <http://www.bdbanalytics.ir/media/1123/storytelling-with-data-cole-nussbaumer-knaflac.pdf>
- Show Me the Numbers [Buch]: Ganz konkrete Tipps für die Praxis
 - https://courses.washington.edu/info424/2007/readings/Show_Me_the_Numbers_v2.pdf
- Now you see it [Buch]: Ebenfalls ganz konkrete Inhalte

Literaturliste

- [James et al. 2013] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani: An introduction to statistical learning
 - Favorit: Sehr gut gemachte Einführung, jedoch Beispiele in R, verständlich mit Mathematik, als pdf frei erhältlich
- [Hastie et al. 2008] Trevor Hastie, Robert Tibshirani, Jerome Friedman: The elements of statistical learning
 - DIE Referenz, für Mathematiker geschrieben, als pdf frei erhältlich
- [O'Neil and Schutt 2013] Cathy O'Neil and Rachel Schutt: Doing Data Science
 - Spannend zu lesen, teilweise Erfahrungsberichte (durch Drittautoren)
- [Mueller and Guido 2017] Andreas C. Müller & Sasha Guido: An Introduction to Machine Learning with Python
 - Interessant da Python 3 tatsächlich genutzt wird für die Einführung inklusive der üblichen Bibliotheken
- [Grues 2016] Joel Grues (übersetzt von Kristian Rother): Einführung in Data Science
 - Auf deutsch gut übersetzt, nutzt Python für grundlegendes Verständnis ohne die üblichen Bibliotheken, extrem leicht lesbar
- [Alpaydin 2008]: Ethem Alpaydin (übersetzt von Simone Linke): Maschinelles Lernen
 - Auf deutsch gut übersetzt, relativ viel Mathematik, in Deutschland scheint das weit verbreitet zu sein
- [Bruce et al. 2020]: Peter Bruce, Andrew Bruce, Peter Gedeck: Practical Statistics for Data Scientists
 - Das einzig wahre Statistikbuch was keines ist
- [Reinhart 2016]: Alex Reinhart (übersetzt von Knut Lorenzen): Statistics done wrong
 - Bevor man wirklich Konfidenzintervalle oder p-Werte angibt und über „Signifikanz“ spricht, sollte man das gelesen haben