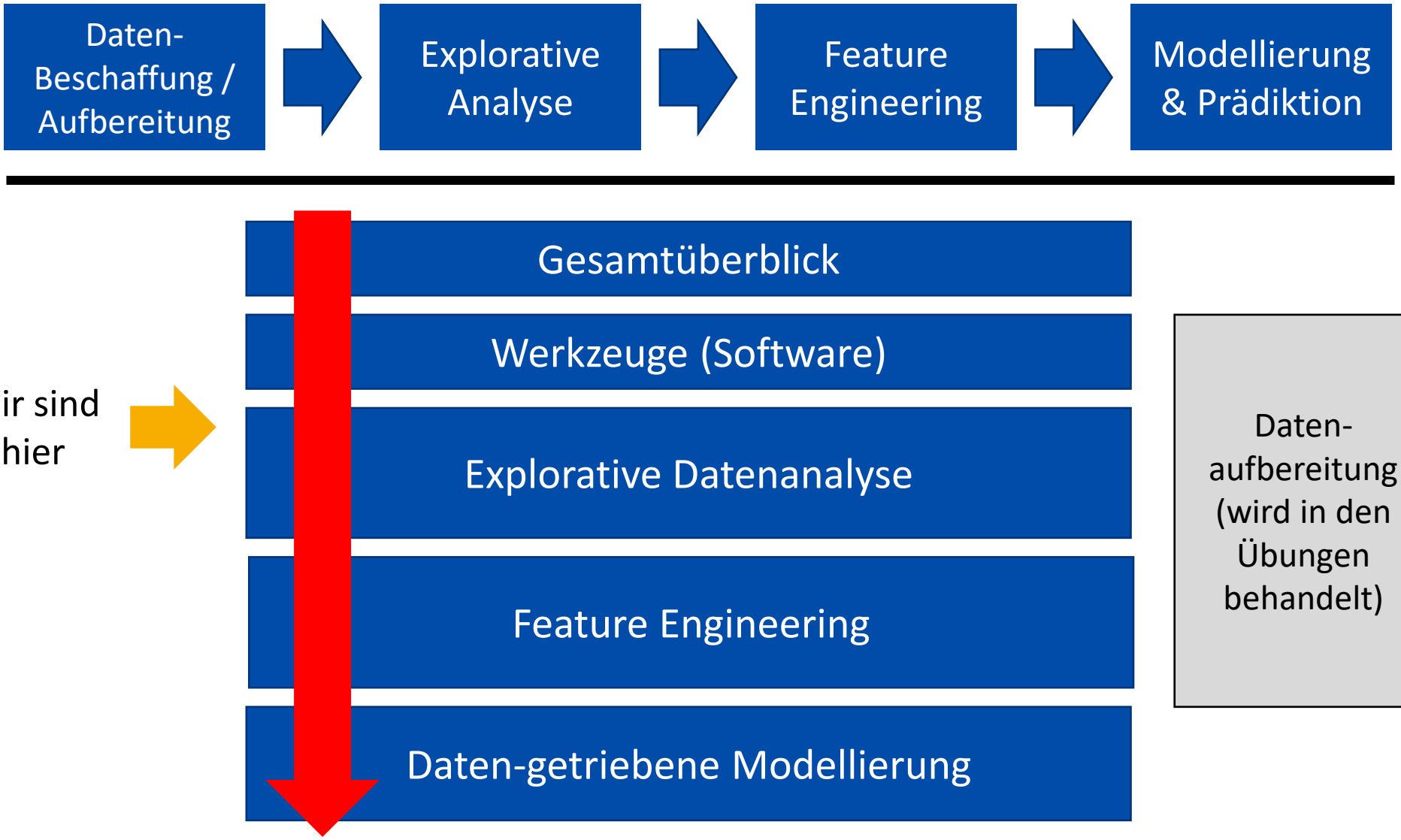
Einführung in Data Science

Unser Plan für heute:

1. Wiederholung
2. Explorative Datenanalyse (EDA)
3. Visualisierung

Data Science



Curriculum

1. Grundbegriffe / Überblick
2. Zentrale Softwarebibliotheken

3. Univariate explorative Analyse (EDA) Visualisierung (Teil 1)
4. Visualisierung (Teil 2), Multivariate explorative Analyse (Teil 1)
5. Multivariate explorative Analyse (Teil 2)
6. Dimensionsreduktion (Teil 1): PCA
7. Dimensionsreduktion (Teil 2): MDS, Isomap
8. Clustering: K-Means, HCA
9. Clustervalidierung
10. Probeklausur
11. Feature Engineering, Datengetriebene Modellierung (Teil 1)
12. Datengetriebene Modellierung (Teil 2)

Überblick / Begriffe

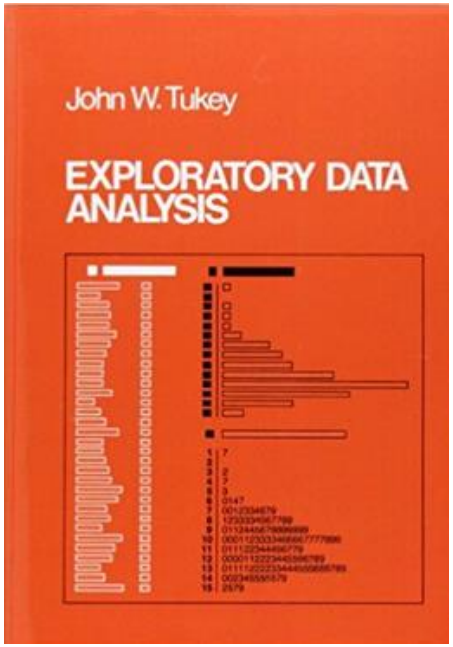
Explorative Analyse (EDA)

Feature Engineering & Modellierung

Explorative Datenanalyse

- basiert auf Einsichten der Statistiker am Bell Laboratories (60er Jahre)
- Begriff stammt von John W. Tukey, amerikanischer Statistiker
- Haltung: „Was können die Daten uns erzählen“ (daten-getriebenes Arbeiten ergänzt Hypothesen-getriebenes Arbeiten)
- Einführung verschiedener Techniken wie z.B. 5-Number Summary, Box Plots, …

John W. Tukey
(1915—2000)



Pearson Verlag (1977)

Explorative Datenanalyse

… ist die Erkundung von Daten mit folgenden Zielen:

1. Identifikation von Problemen im Datensatz
2. Prüfung, ob initiale Fragen beantwortbar sind
3. Erzeugung erster Antwortskizzen
4. **Erzeugung neuer Fragen / Hypothesen**

Typische Werkzeuge:

- Erwartungshaltung (Fragen formulieren)
- Deskriptive Statistik
- Visualisierung
- Dokumentation des Erkenntnisweges
  - zur Steuerung der eigenen Analyse
  - zur Kommunikation/Diskussion der Ergebnisse mit Dritten
  - zur Sicherung der Reproduzierbarkeit

F

Explorative Analyse | Dokumentation

- Dokumentationswege unterscheiden sich je nach Arbeitsumfeld / Gruppe

Beispiele für Dokumentationen

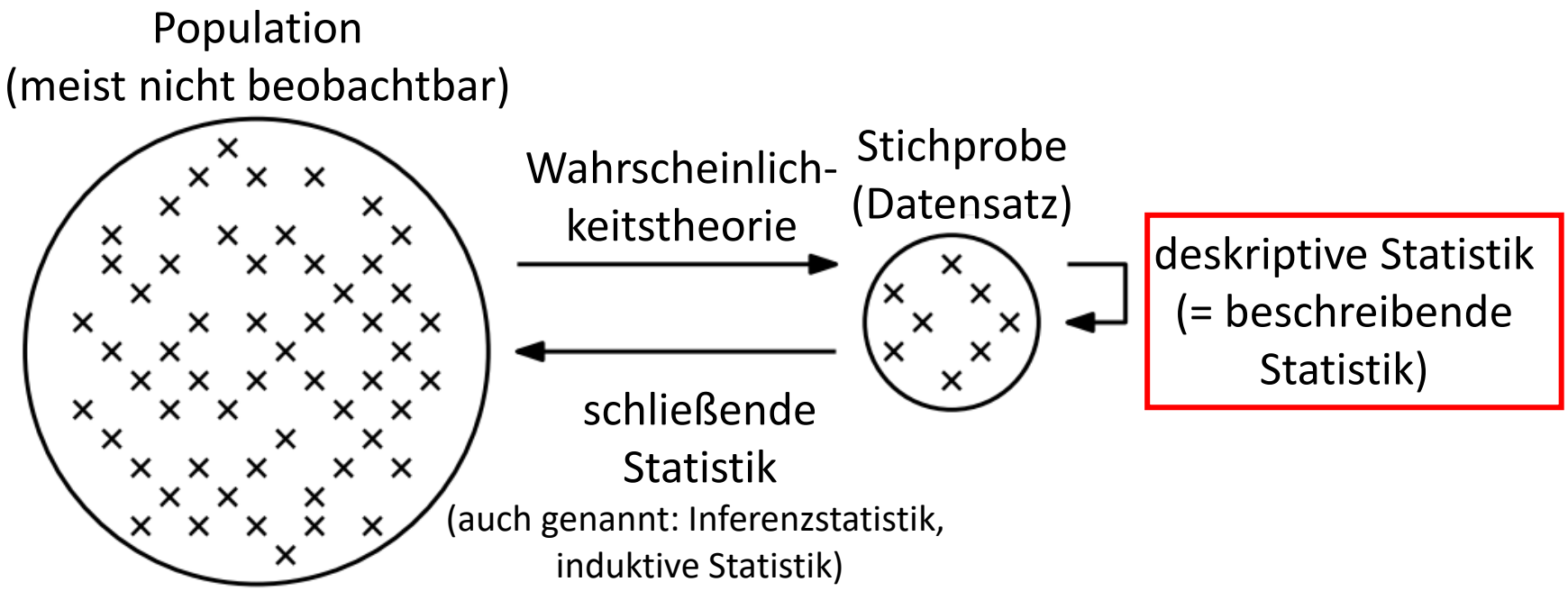
- Jupyter Notebooks
  (typisch im Bereich Data Science seit wenigen Jahren)
- „Labor-Buch“ (handschriftliche oder elektronische Notizen) angefertigte Abbildungen
- kommentierter Code

Beispiel

Explorative Datenanalyse | Deskriptive Statistik

Typische Werkzeuge:

- Erwartungshaltung (Fragen formulieren)
- Deskriptive Statistik
- Visualisierung
- Dokumentation des Erkenntnisweges



Explorative Datenanalyse | Deskriptive Statistik

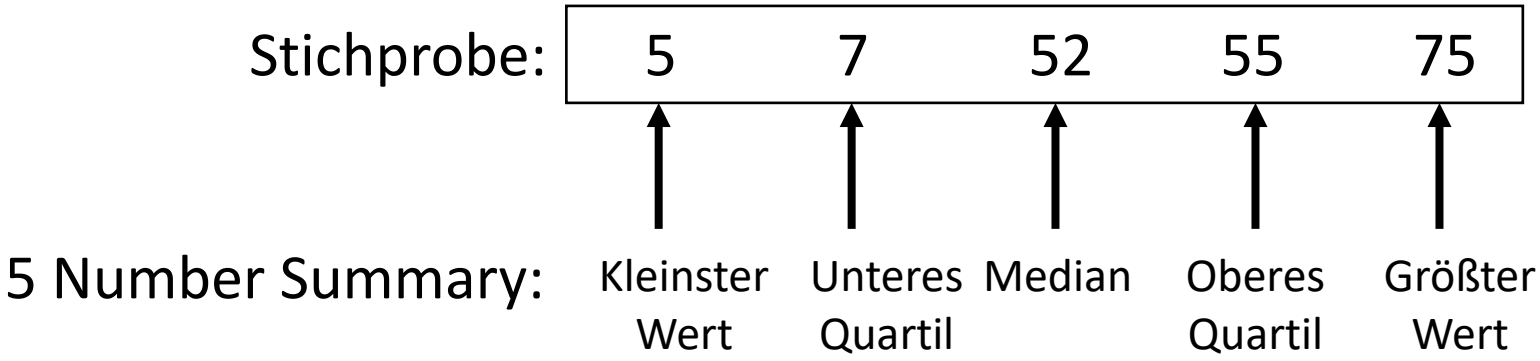
Kennzahlen (englisch: Summary Statistics)
(auf Deutsch auch genannt: aggregierende Parameter, Maßzahlen)

- beschreiben eine Stichprobe (Häufigkeitsverteilung) in wenigen Zahlen

Typ	**Verwendung**	**Beispiele**
Lageparameter	beschreiben zentrale Tendenz der Stichprobe (z.B. wo die meisten Werte der Stichprobe sind)	▪ Modus (= häufigster Wert) ▪ Median ▪ Quartile ▪ p-Quantile ▪ arithmetisches Mittel
Streuungsparameter	beschreiben Streubreiten von Stichprobe	▪ Varianz ▪ Standardabweichung ▪ Interquartilsabstand

Explorative Datenanalyse | Deskriptive Statistik

„5 Number Summary“ (*5-Punkte-Zusammenfassung)*
nach J. Tukey



Explorative Datenanalyse | Deskriptive Statistik

Empirisches p-Quantil

Bezeichne $\lfloor x \rfloor$ die Abrundungsfunktion. Beispiel: $\lfloor 5.7 \rfloor = 5$

Sei $(x_1, \ldots, x_n)$ eine Stichprobe der Größe *n*. Seien diese *n* Elemente geordnet, so dass gilt: $x_1 \leq x_2 \leq \ldots \leq x_n$

Dann heißt für eine Zahl $p \in (0, 1)$

$$x_p = \begin{cases} \frac{1}{2}(x_{n \cdot p} + x_{(n \cdot p)+1}), & \text{wenn } n \cdot p \text{ ganzzahlig,} \\ x_{\lfloor n \cdot p + 1 \rfloor}, & \text{wenn } n \cdot p \text{ nicht ganzzahlig.} \end{cases}$$

das *empirische p-Quantil*.

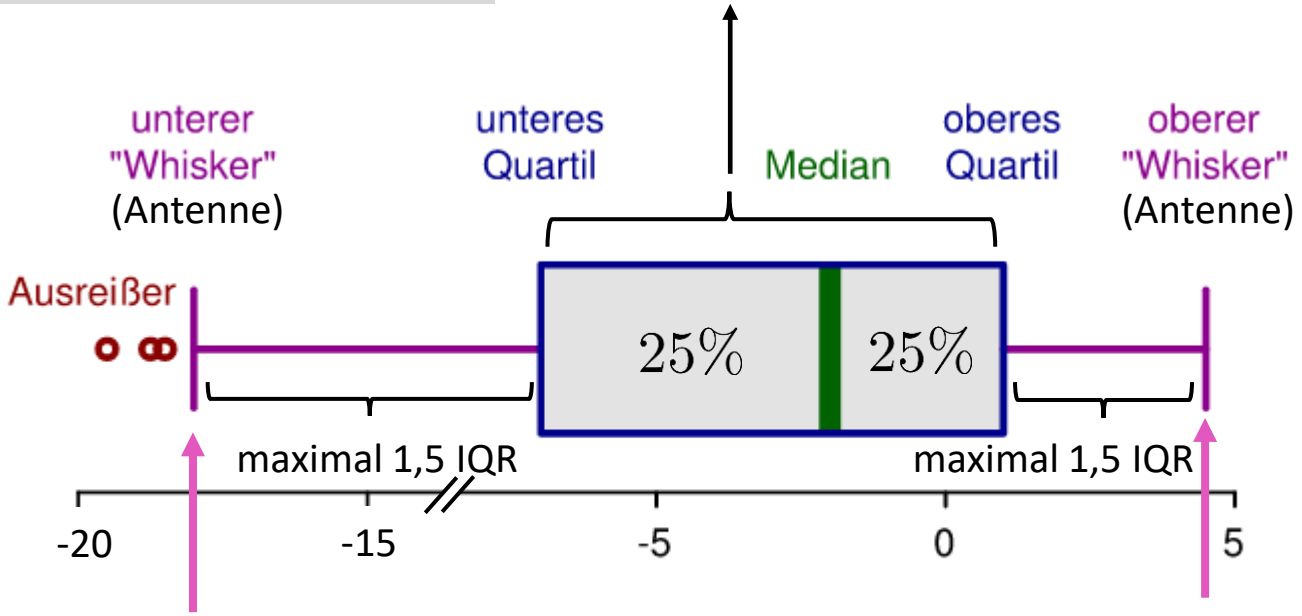
Median: $x_{0.5}$ Unteres *Quartil*: $x_{0.25}$ Oberes *Quartil*: $x_{0.75}$

Explorative Datenanalyse | Deskriptive Statistik

Tukey Boxplot – Variante nach J. W. Tukey
(auch *Kastengrafik* oder einfach „Boxplot“ genannt)

50% aller Werte der Stichprobe liegen im IQR

Interquartilsabstand
interquartile range IQR = $x_{0.75} - x_{0.25}$



Datenpunkte jenseits von 1,5 IQR vom Quartil entfernt werden oft „Ausreißer“ genannt (markiert als Einzelpunkte)

Datenpunkt, der gerade noch innerhalb 1,5 IQR liegt

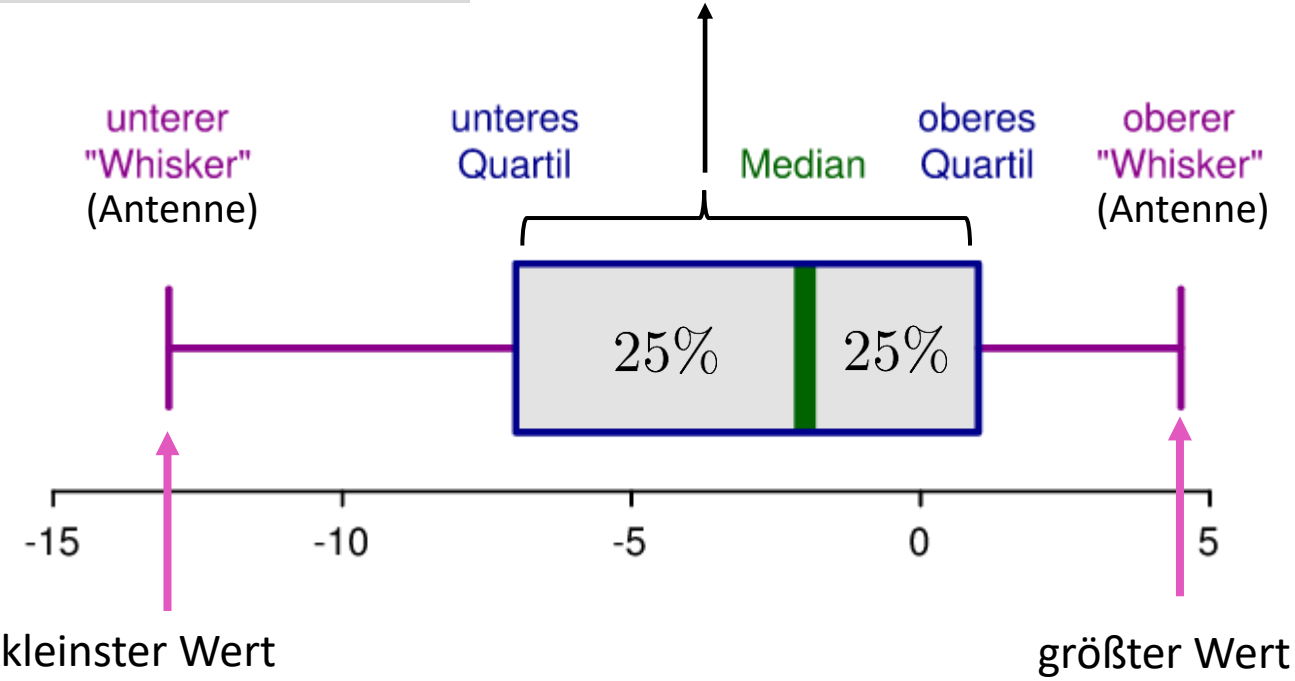
Datenpunkt, der gerade noch innerhalb 1,5 IQR liegt

Explorative Datenanalyse | Deskriptive Statistik

Boxplot – alternative Variante (seltener genutzt)

50% aller Werte der Stichprobe liegen im IQR

Interquartilsabstand
interquartile range IQR = $x_{0.75} - x_{0.25}$



Unterschied zur Tukey Variante:

- Definition der Antennen (Whiskers): Antennen erfassen die beiden extremsten Werte der Stichprobe

Explorative Datenanalyse | Deskriptive Statistik

5 Number Summary – Beispiel[1]

	Min	25%	Median	75%	Max
Age	241	418	584	748	959
Weight	32.4	67.2	78.8	92.6	218.2
Height	140	160	167	175	204
Leg Length	23.7	35.7	38.4	41	55.5
Arm Length	29.5	35.5	37.4	39.4	47.7
Arm Circumference	19.5	29.7	32.8	36.1	141.1
Waist	59.1	87.5	97.95	108.3	172

Frage
Welche Auffälligkeiten sehen Sie in den Daten?

F

Beispiele
- Median-Alter der Teilnehmer ist 584. *Jahre?* Nein, Monate!
- Variabilität in der Beinlänge ist größer als in der Armlänge.

1) Daten des *National Health and Nutrition Examination Survey (NHANES), USA.*
Langzeitstudie seit 1971 zu Gesundheitsstatus und Ernährungsgewohnheiten von Menschen in den USA.

Explorative Datenanalyse | Deskriptive Statistik

Weitere typische Kennzahlen (*summary statistics*):

Lageparameter

Arithmetisches Mittel $\mu = \frac{1}{n} \sum_{i=1}^{N} x_i$

Streuungsparameter

Standardabweichung $\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n-1}}$

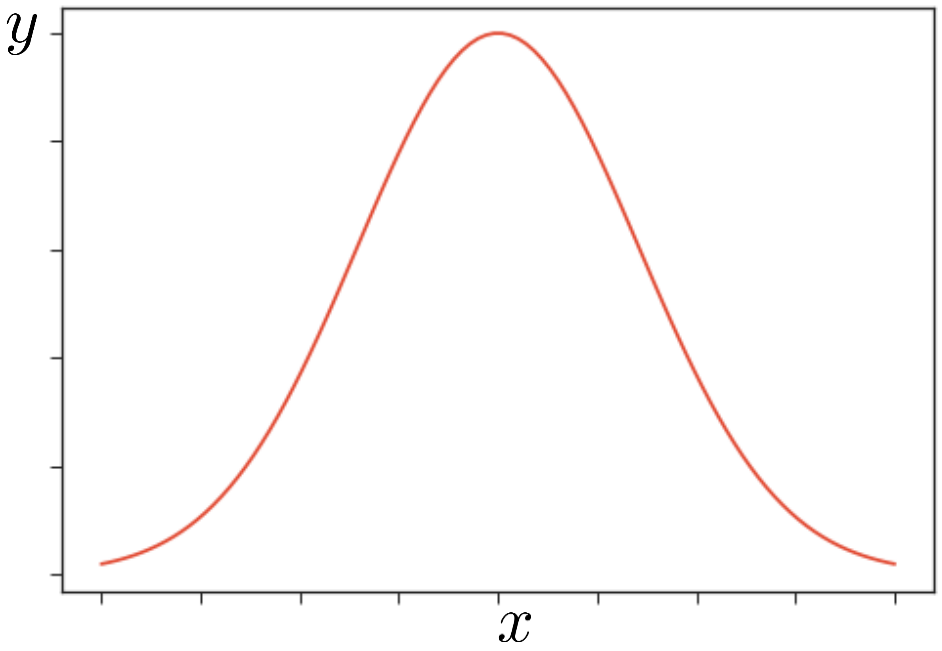
Varianz $V = \sigma^2$

Explorative Datenanalyse | Deskriptive Statistik

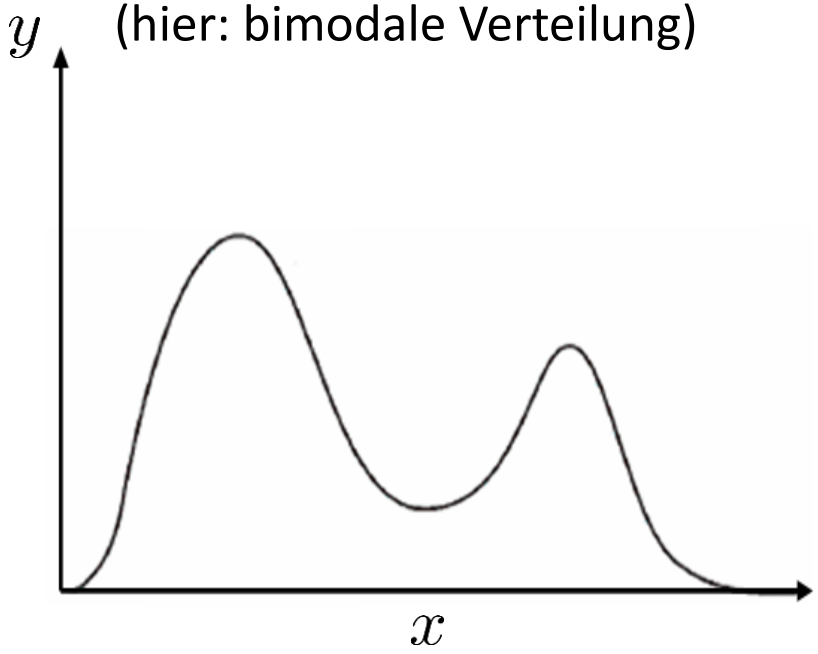
Frage
Wie schätzen Sie die Aussagekraft von Mittelwert und Standardabweichung bei folgenden zwei Häufigkeitsverteilungen ein?

F

Unimodale Verteilung ✓



Multimodale Verteilung (hier: bimodale Verteilung) ✗



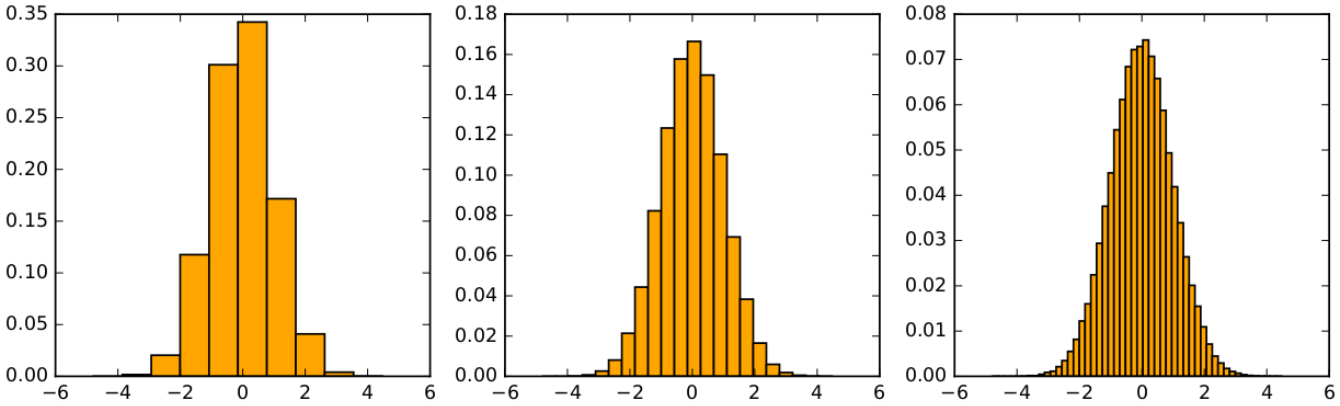
Mittelwert ist keine gute Beschreibung multimodaler Verteilungen.
Für diese besser Histogramme verwenden

Explorative Datenanalyse | Deskriptive Statistik

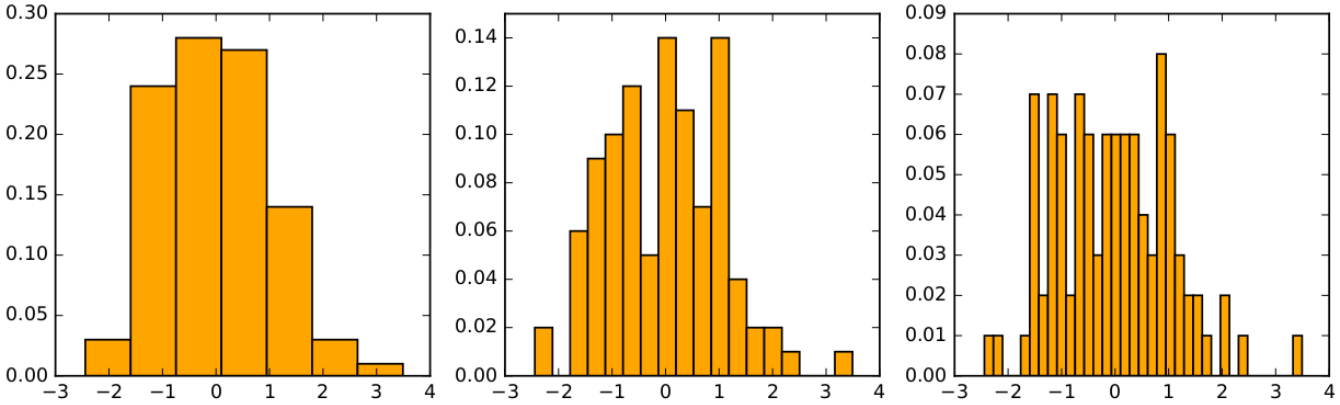
Histogramm

graphische Darstellung einer Häufigkeitsverteilung durch Einteilen der Daten in Klassen (englisch: *Bins*).

100000 Datenpunkte einer Normalverteilung



100 Datenpunkte einer Normalverteilung



Bingrößen der Datenanzahl angemessen wählen

Bins werden kleiner

Explorative Datenanalyse | Deskriptive Statistik

Kennzahlen (summary statistics) für 4 Datensätze:

Anscombes Quartett

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.31	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Var.	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Corr.	0.816		0.816		0.816		0.816	

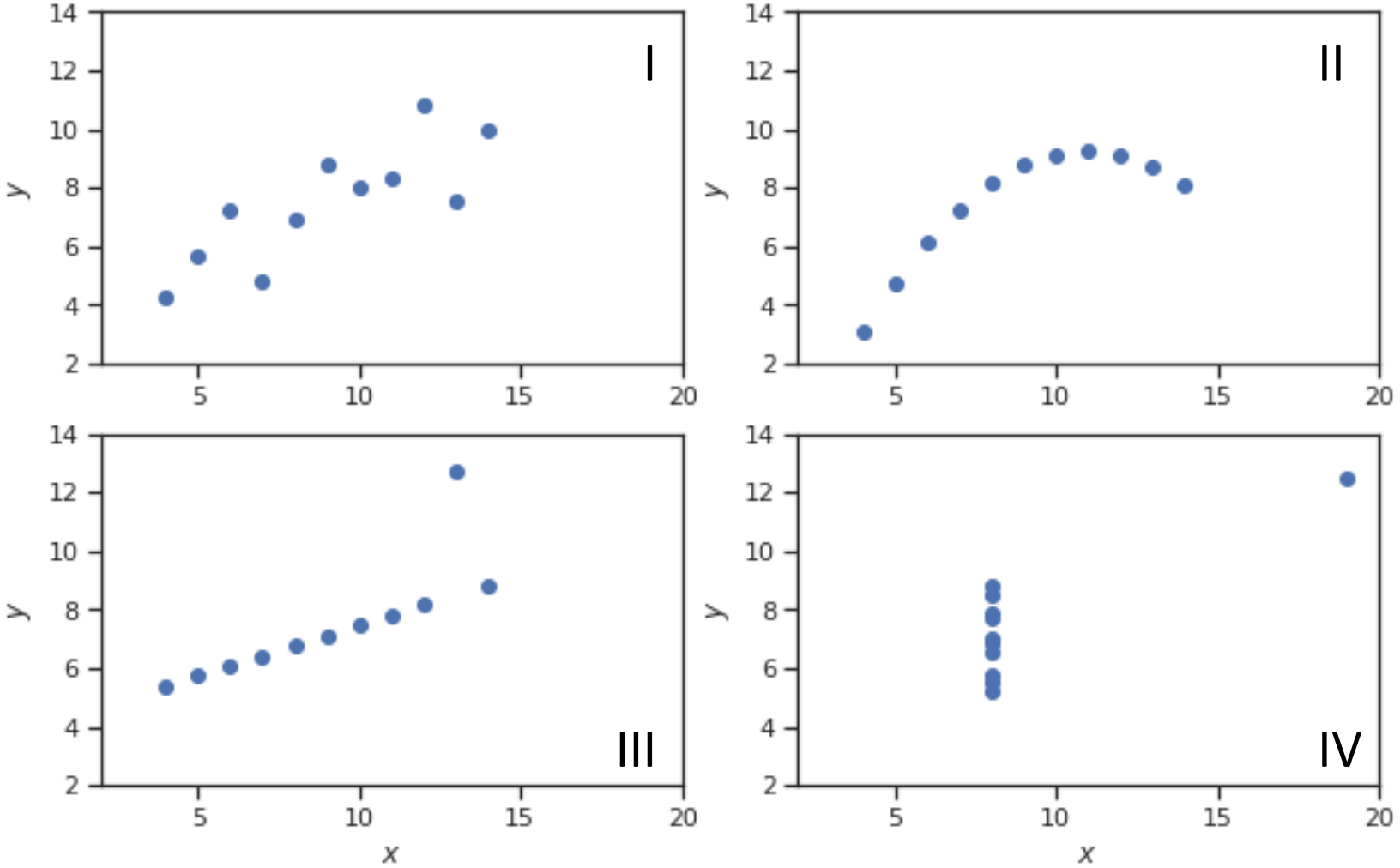
Unterschiedliche Datensätze **aber** gleiche Summary Statistics

→ Summary Statistics hier wenig hilfreich. Stattdessen: Visualisierung

Summary Statistics

Explorative Datenanalyse | Deskriptive Statistik

Visualisierung des Anscombe Quartetts[1]:



1) Daten wurden 1971 vom englischen Statistiker Francis Anscombe erzeugt, um die Wichtigkeit von Datenvisualisierung zu demonstrieren.

Explorative Datenanalyse | Visualisierung

Typische Werkzeuge der explorativen Analyse:

- Erwartungshaltung (Fragen formulieren)
- Deskriptive Statistik
- Visualisierung
- Dokumentation des Erkenntnisweges

Visualisierung

- erlaubt schnelle Exploration von Mustern durch unser visuelles System
- mächtiges Instrument zur Kommunikation von Ergebnissen

Explorative Datenanalyse | Visualisierung

Aktivität

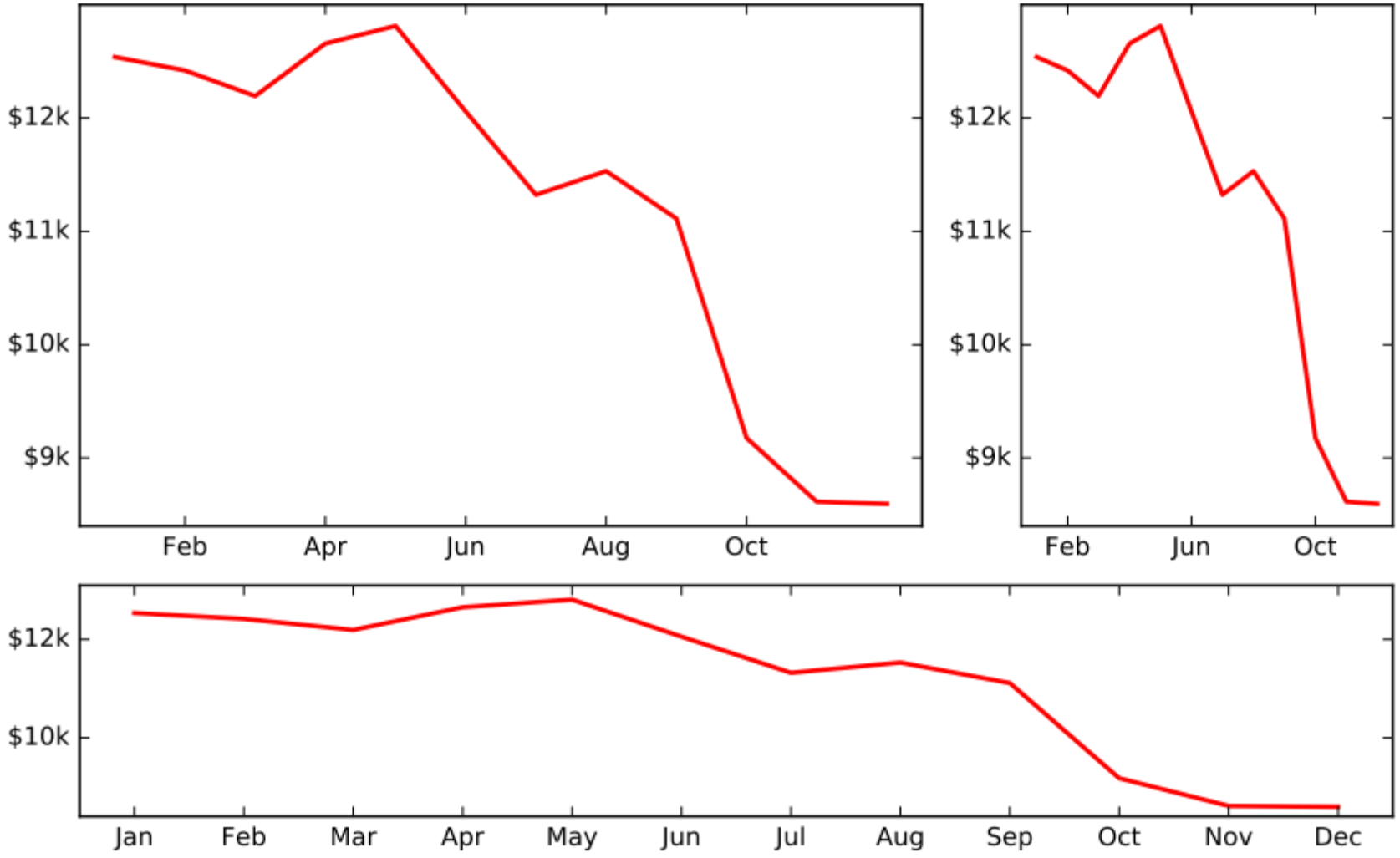
Sie erhalten von mir eine Graphik.

Ihre Aufgaben

1. Betrachten Sie die Graphik.

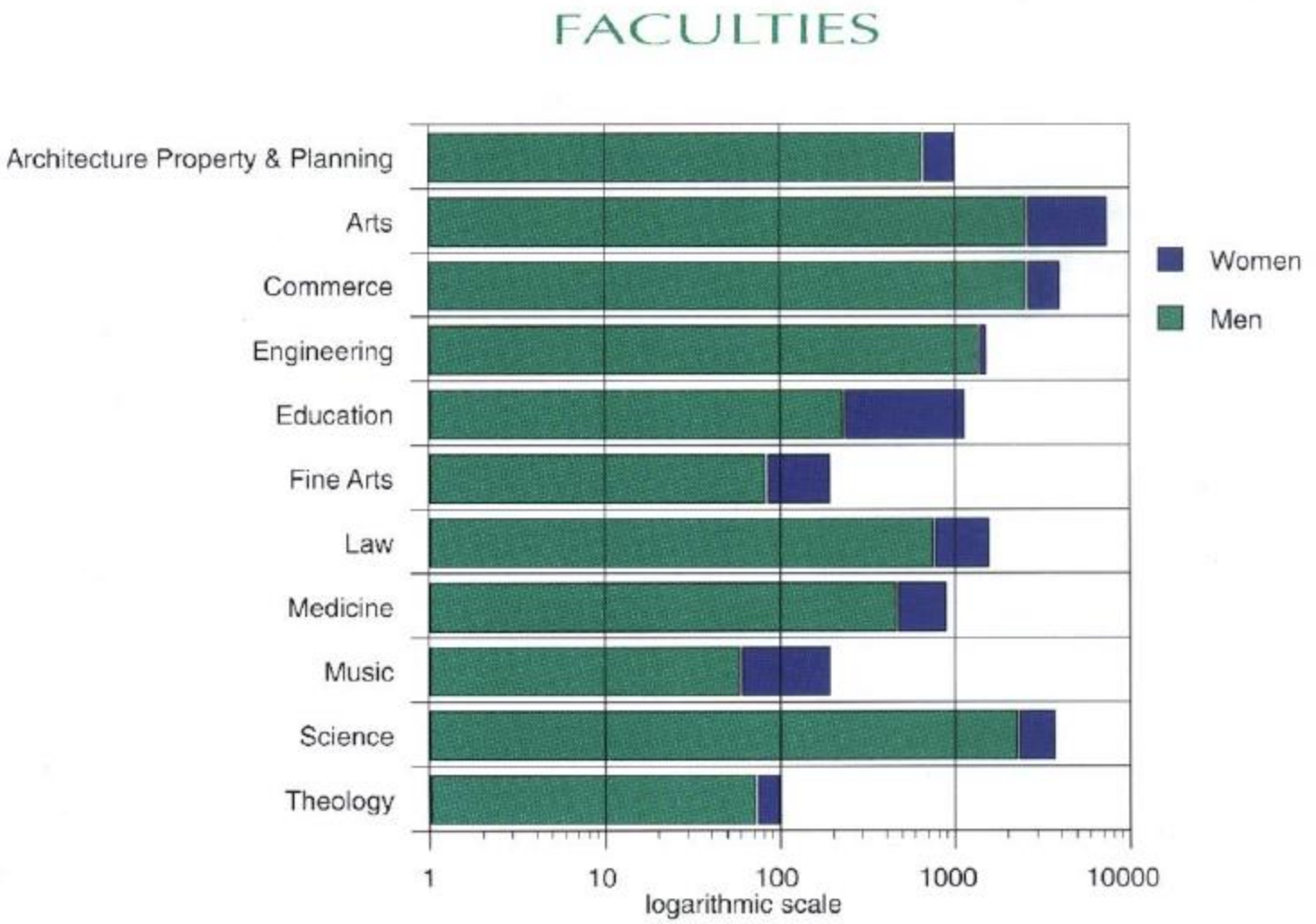
2. Notieren Sie alle Aspekte, die Sie an der Graphik problematisch finden.

Visualisierung | Beispiel 1



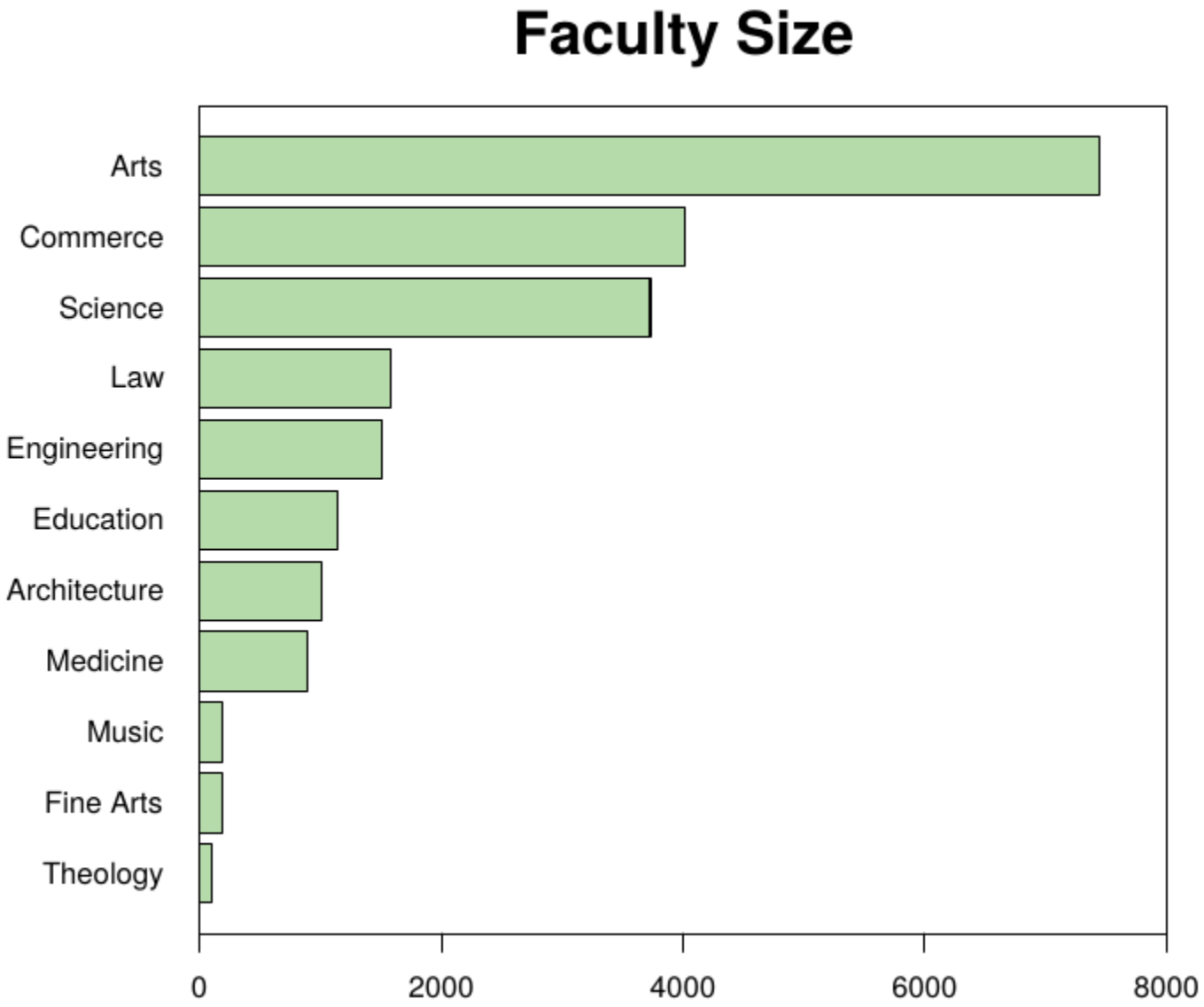
- Skalierung von x- und y-Achsen können Preisentwicklung dramatisieren oder entdramatisieren

Visualisierung | Beispiel 2



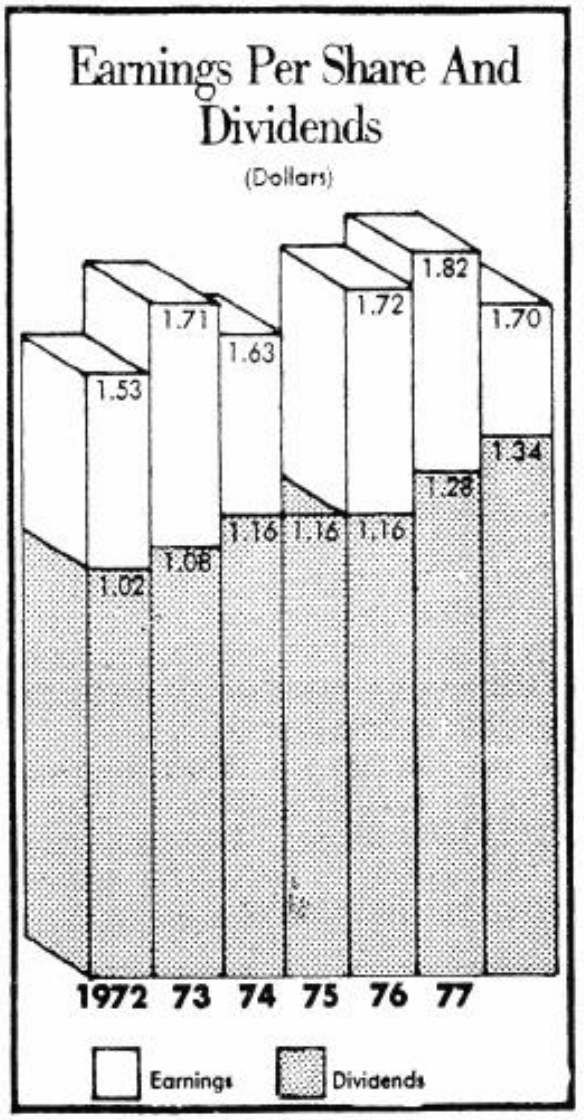
- Manipulative x-Achsen Skalierung suggeriert auf den ersten Blick geringe Frauenanteile; schwierig, Fakultätsgrößen in absoluten Zahlen abzulesen

Visualisierung | Beispiel 2 (besser)



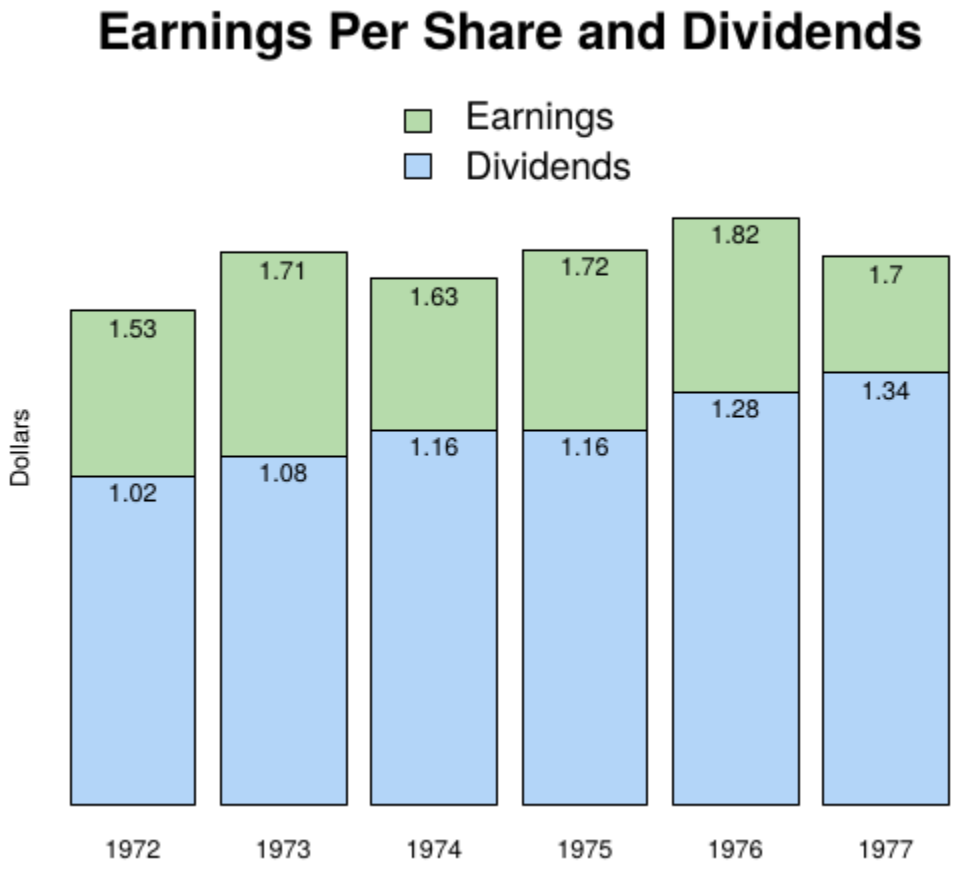
Visualisierung | Beispiel 3

- Dritte Dimension überflüssig und verwirrend
- Rätselhafte Darstellung von Daten (Jahr 1975); zusätzlicher Balken rechts ohne Jahresdatum

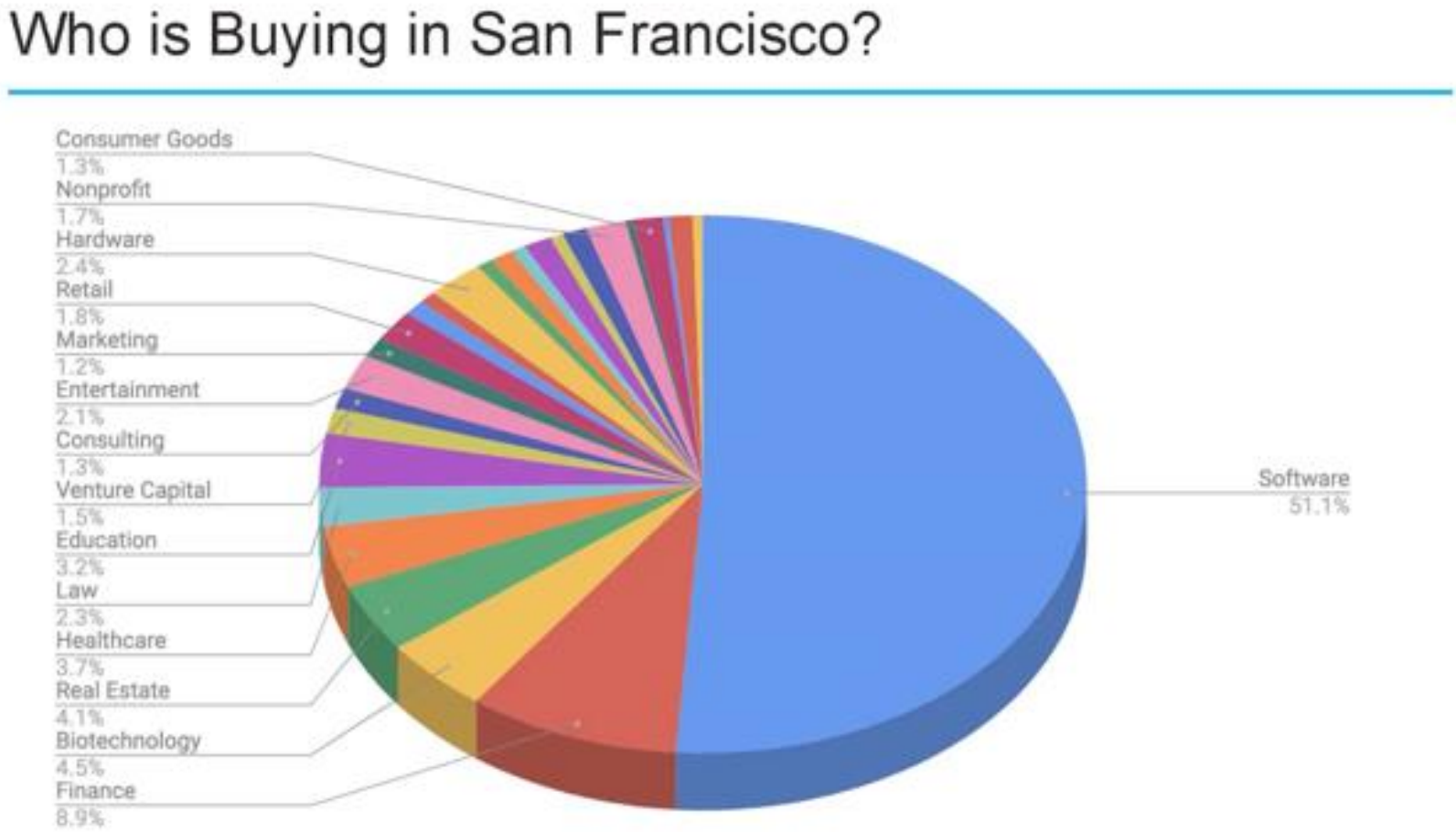


Washington Post, 1979

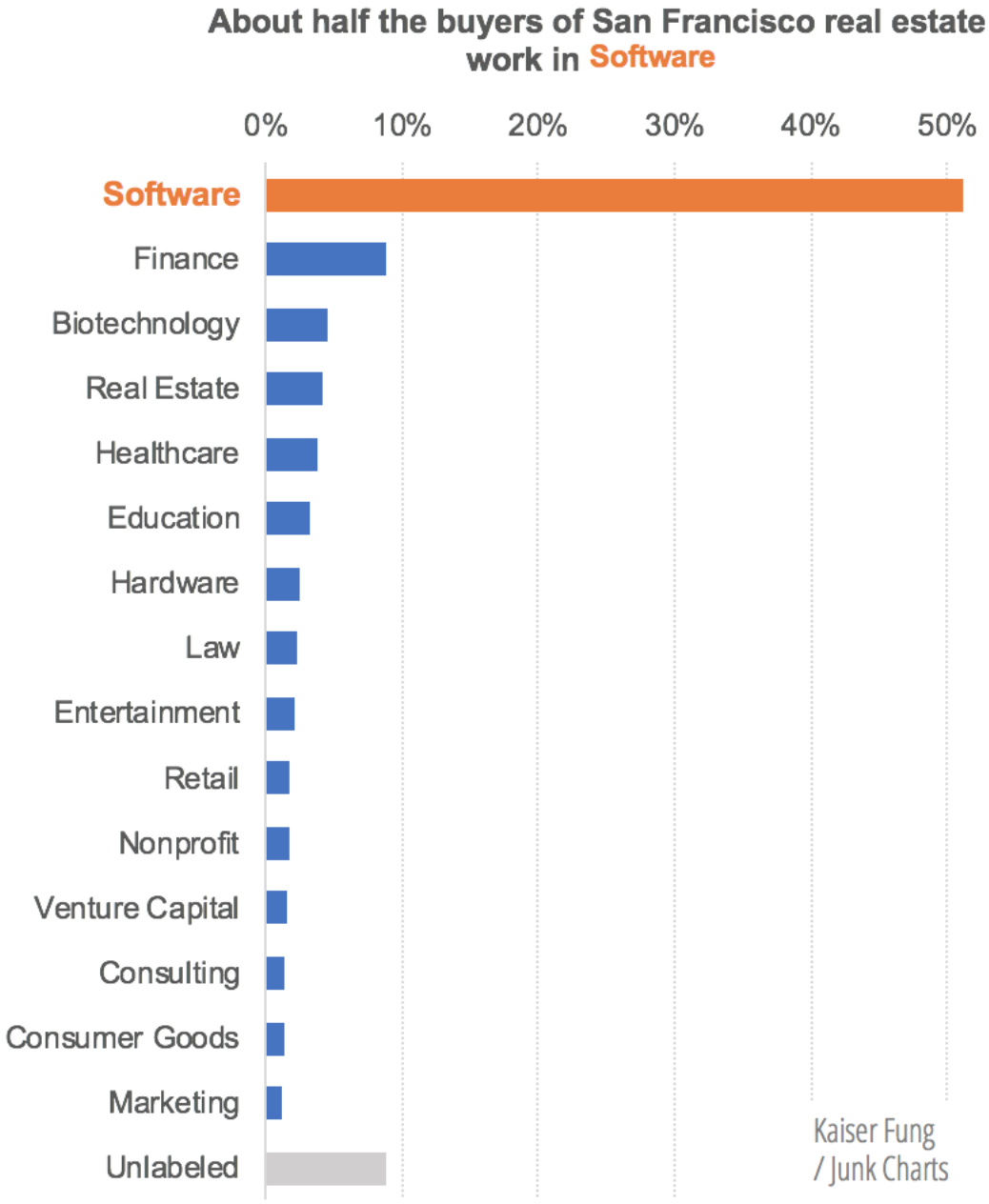
Visualisierung | Beispiel 3 (besser)



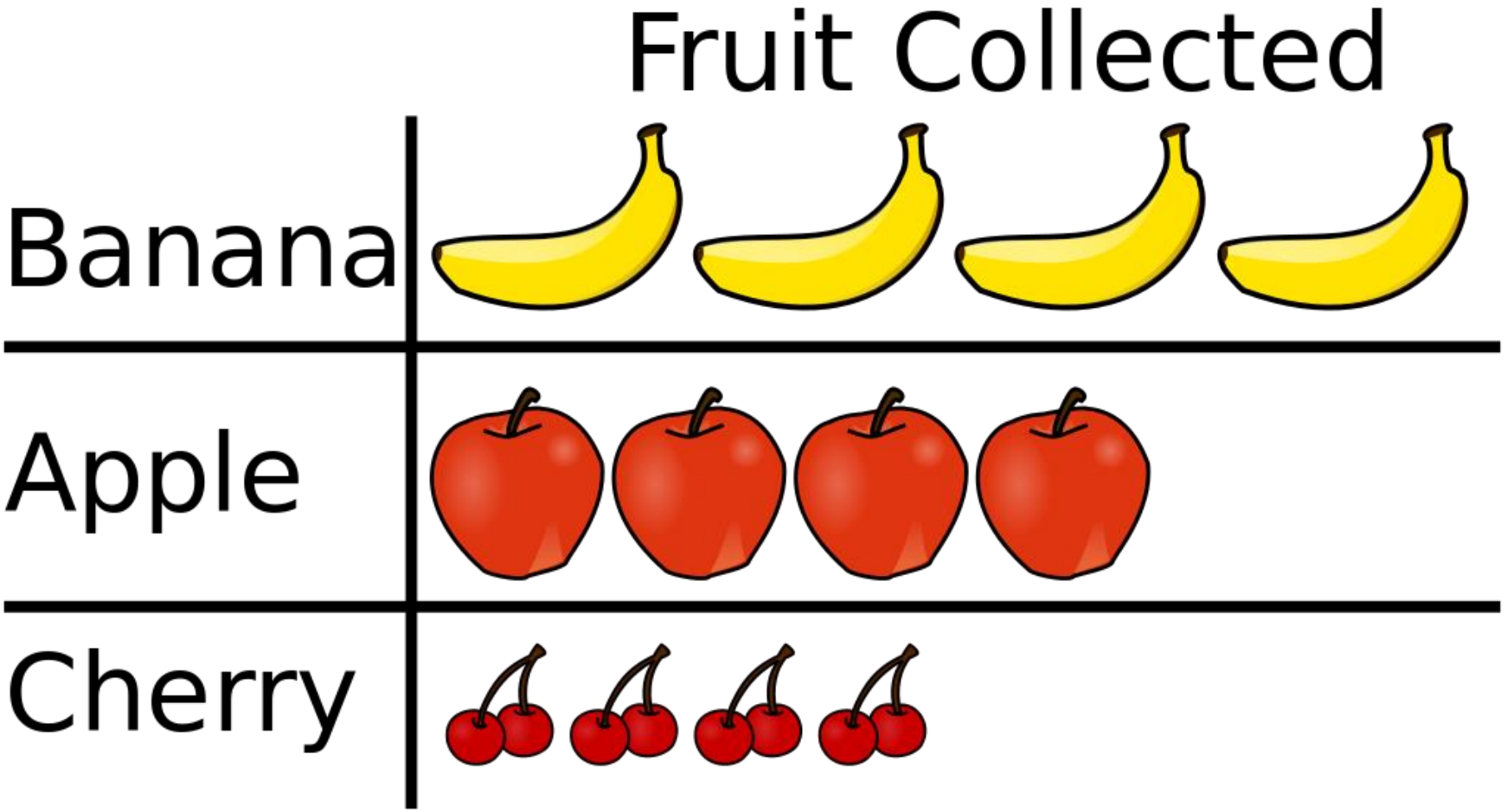
Visualisierung | Beispiel 4



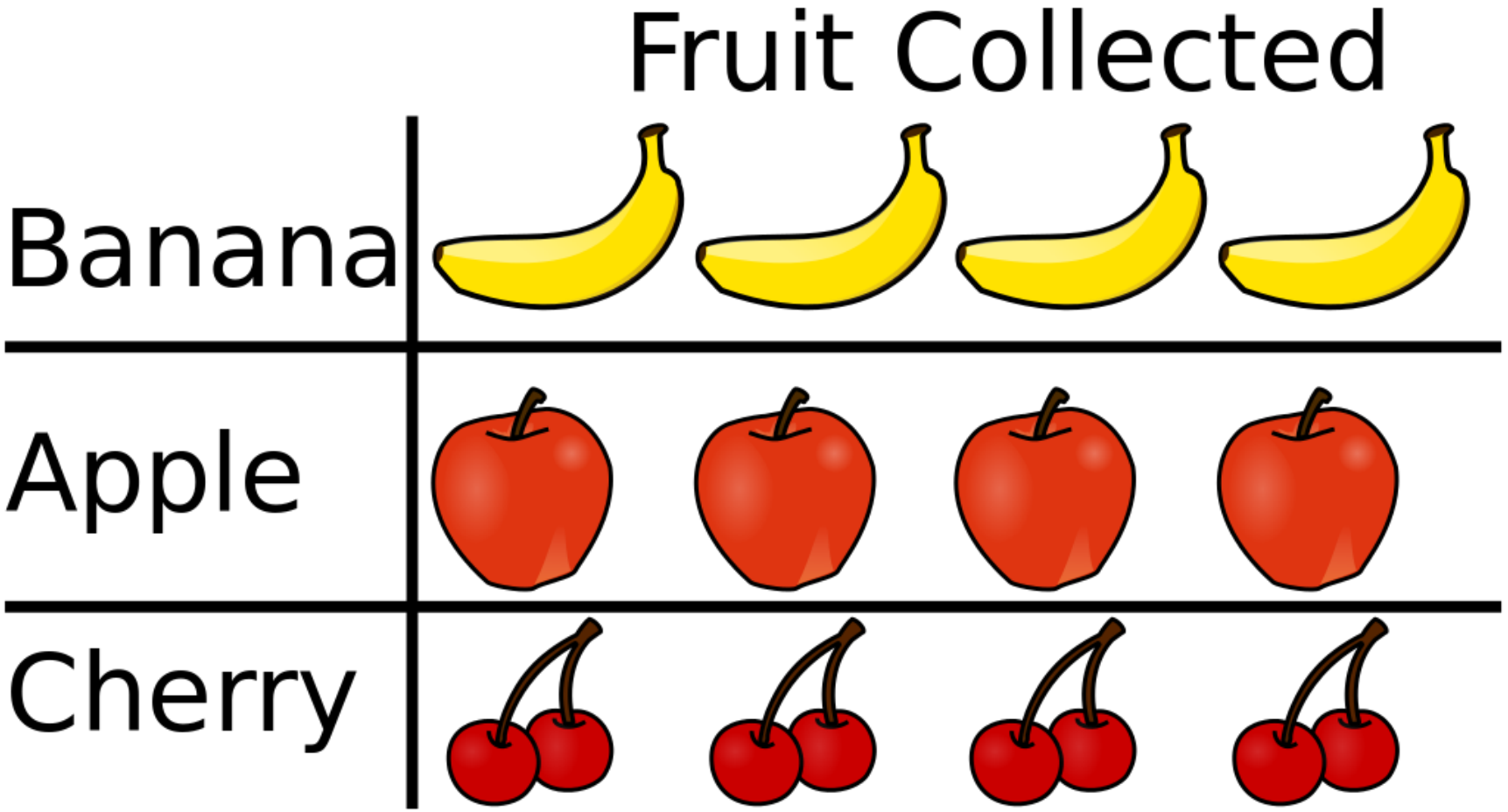
Beispiel 4 (besser)



Visualisierung | Beispiel 5

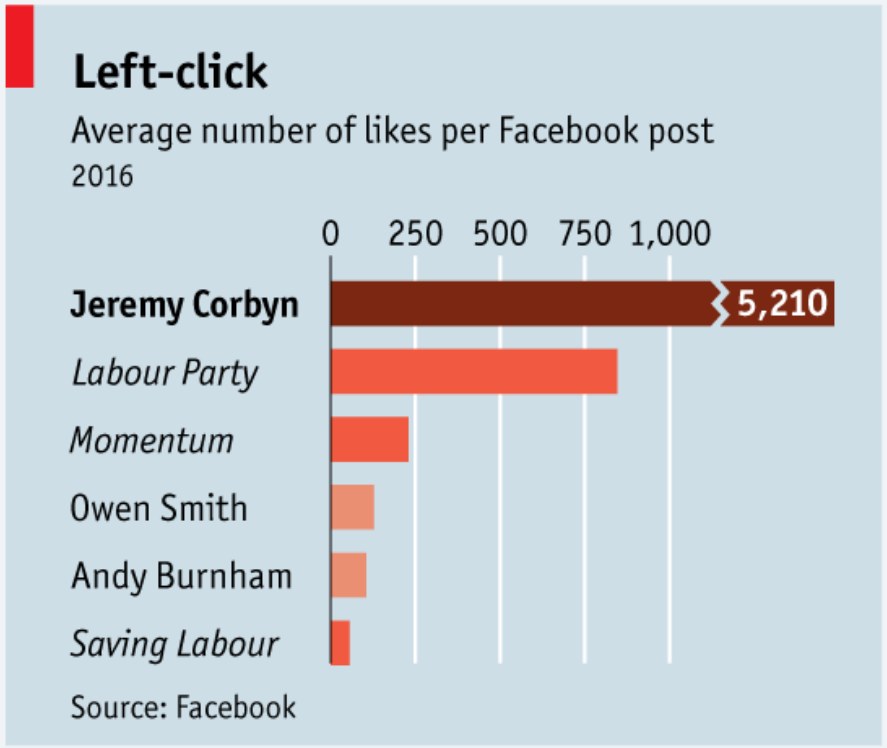


Visualisierung | Beispiel 5 (besser)



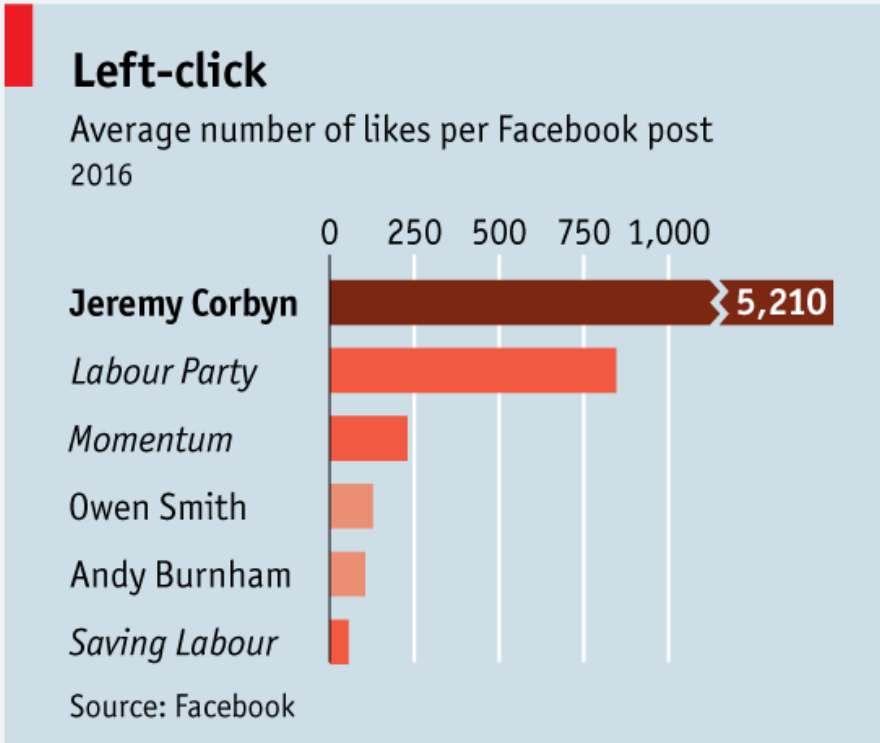
Visualisierung | Beispiel 6

Original

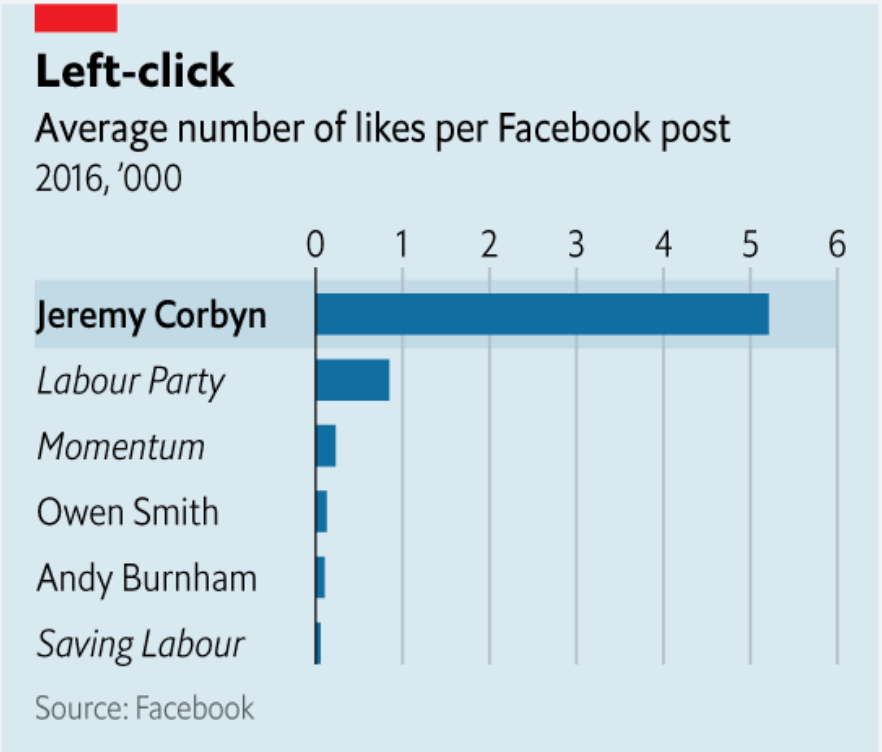


Visualisierung | Beispiel 6 (besser)

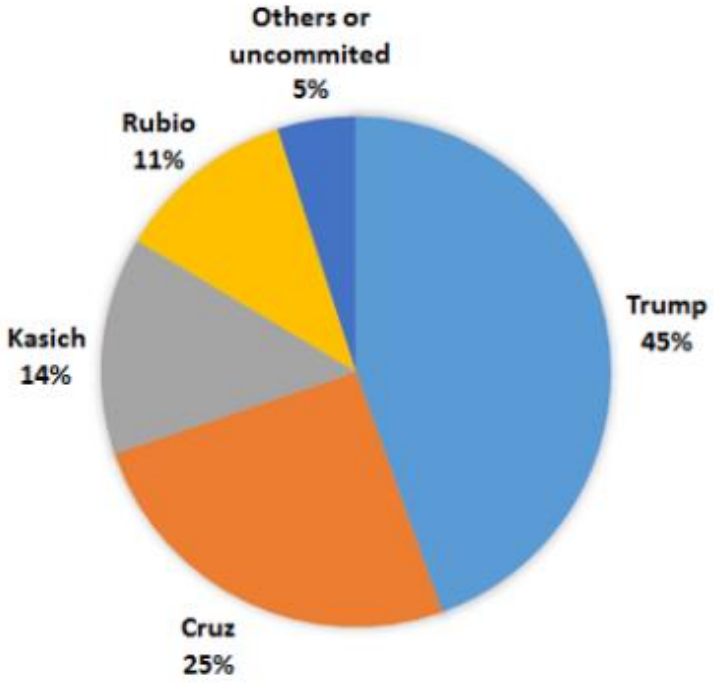
Original

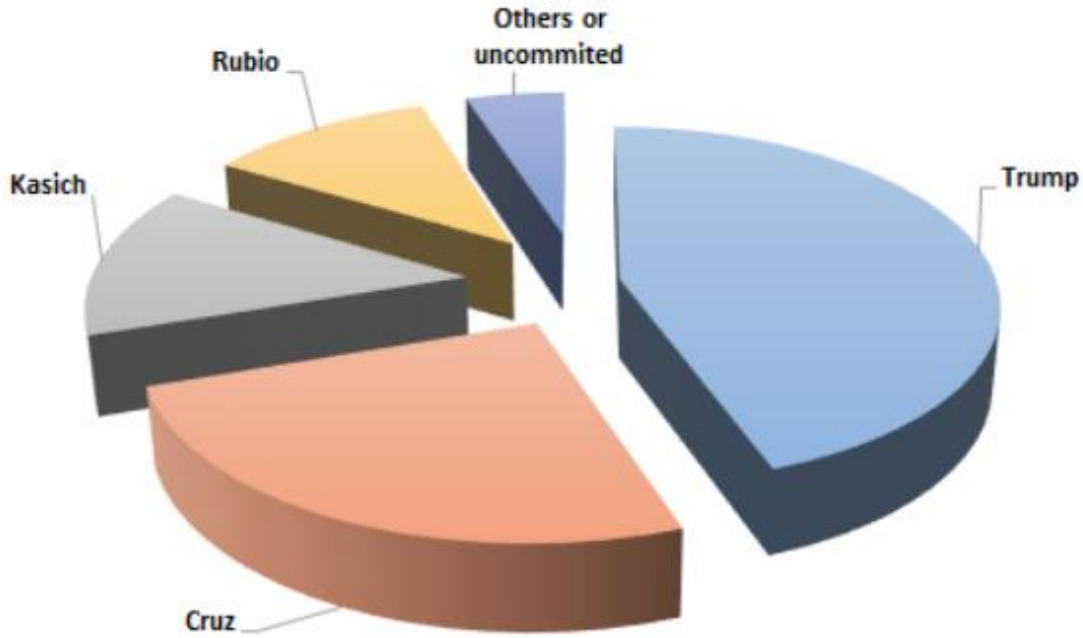


Better

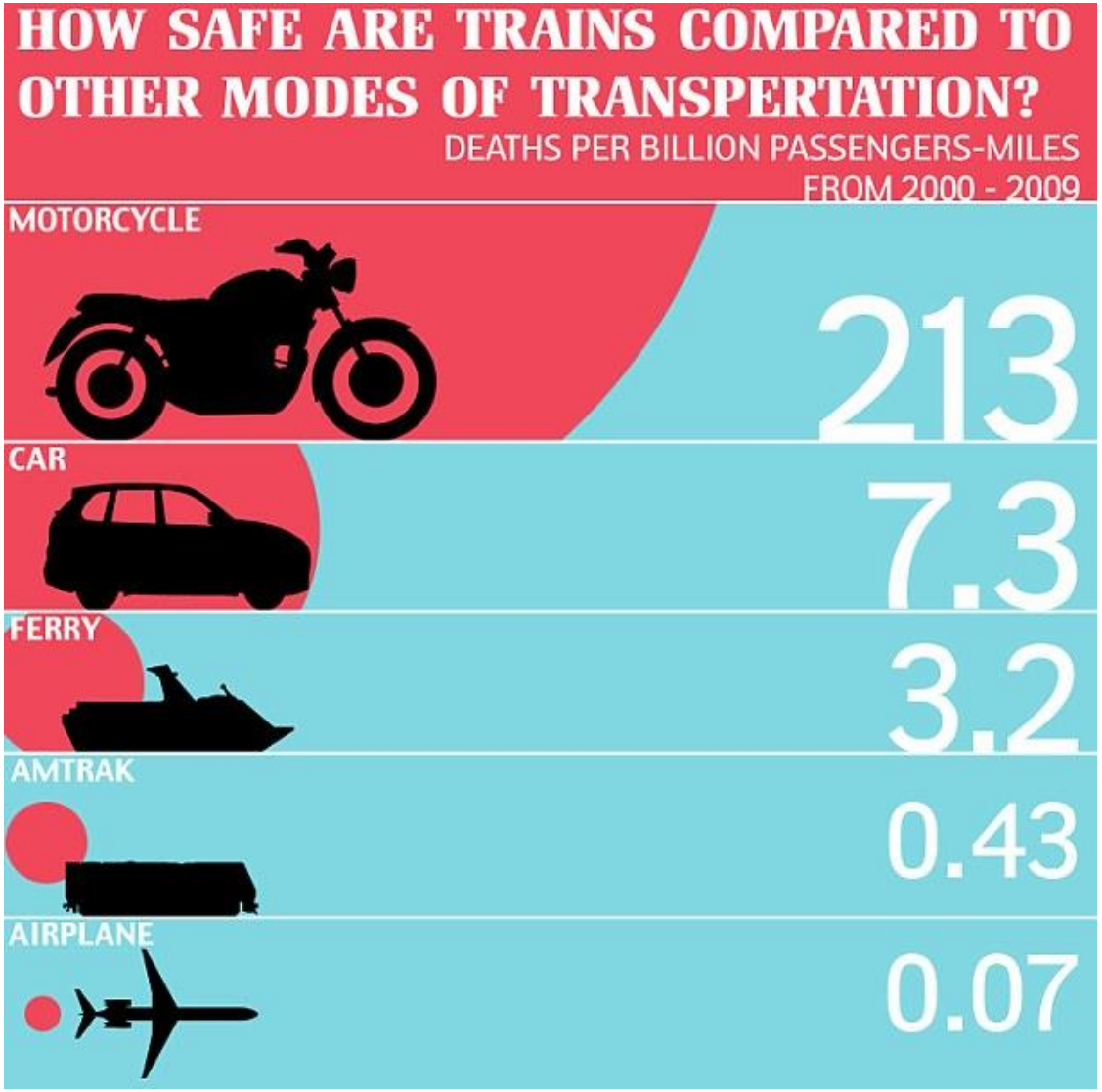


Visualisierung | Beispiel 7



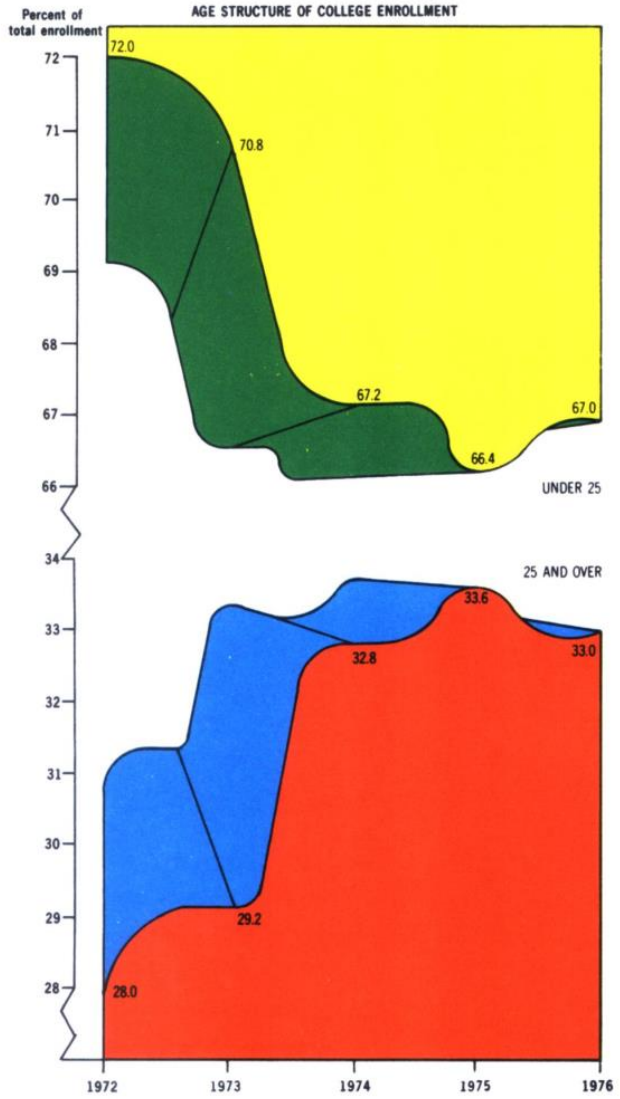


Visualisierung | Beispiel 8

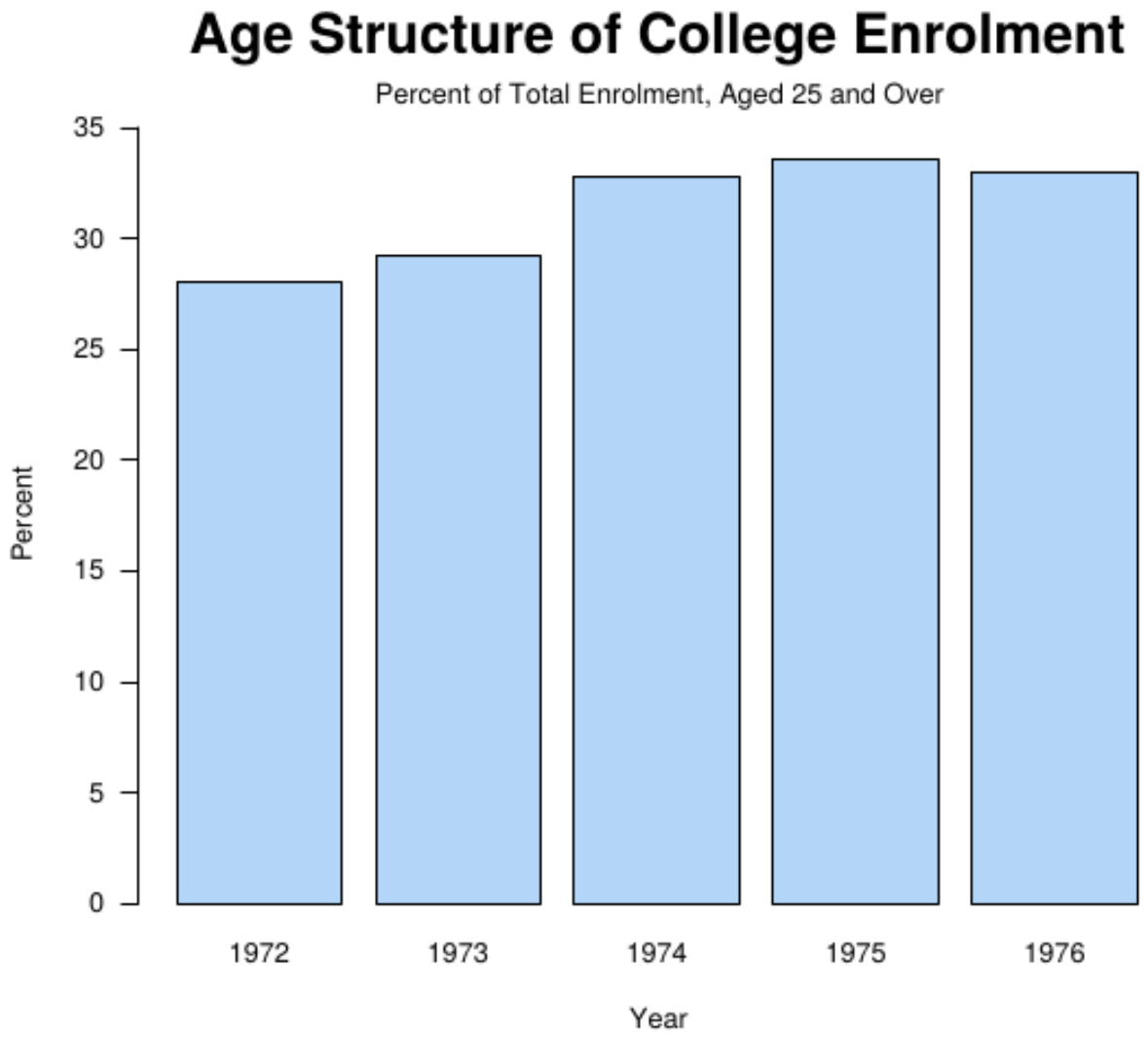


Visualisierung | Beispiel 9

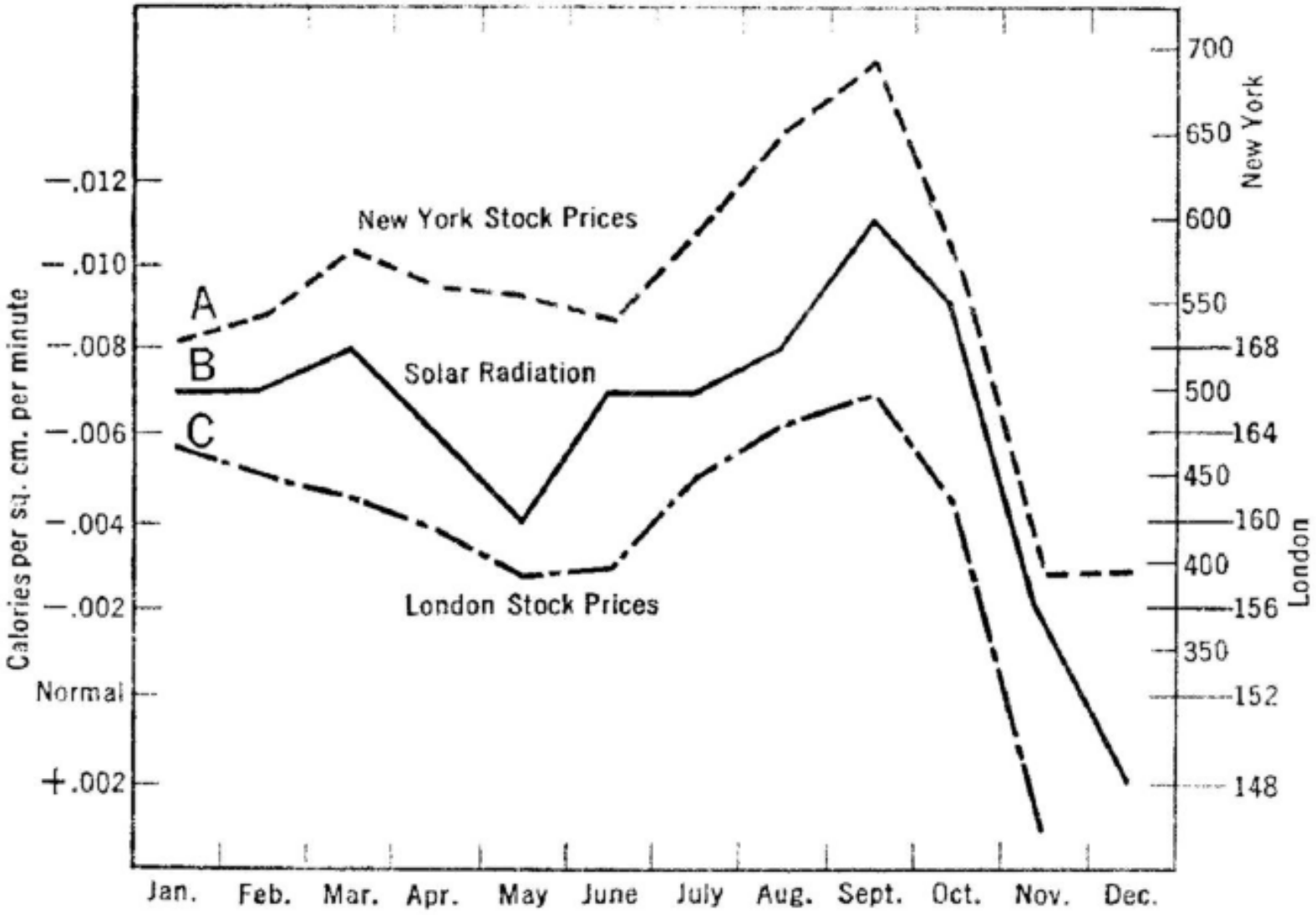
- Abbildung zeigt lediglich 5 Werte, nutzt 6 Farben, eine geteilte y-Achse mit redundanter Information sowie unnötige drei-dimensionale Darstellung



Visualisierung | Beispiel 9 (besser)



Visualisierung | Beispiel 10



Explorative Analyse | Visualisierung

Regeln nach Edward Tufte
(US amerikanischer Informationswissenschaftler; bis 2004 Yale University)

1. Maximieren Sie das Daten-Druckerschwärze Verhältnis
   (auf Deutsch: so wenig Druckerschwärze für soviele Daten wie möglich)
2. Minimieren Sie den Lügenfaktor
3. Minimieren Sie „Chartjunk“
   (keine visuellen Spielereien; lassen Sie die Daten für sich selber sprechen)
4. Nutzen Sie angemessene Skalen; beschriften Sie Ihre Achsen