Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Datenmanagement & -analyse

Explorative Datenanalyse

Prof. Dr. Christoph M. Flath
Lehrstuhl für WI & BA
Julius-Maximilians-Universität Würzburg

Sommersemester 2021

Agenda

1. Einführung
2. Beschreibende Statistik
3. Datenvisualisierung eindimensionaler Daten
4. Datentransformationen
5. Datenvisualisierung mehrdimensionaler Daten

Beispiele für geschäftliche Fragen – wir brauchen Antworten

- „Wer sind die profitabelsten Kunden?"
- „Gibt es einen Unterschied im Wert für das Unternehmen bei diesen Kunden?"
- „Welche Kundensegmente bedienen wir?“
- „Wird dieser neue Kunde ein profitabler Kunde werden? Wenn ja, wie profitabel?“

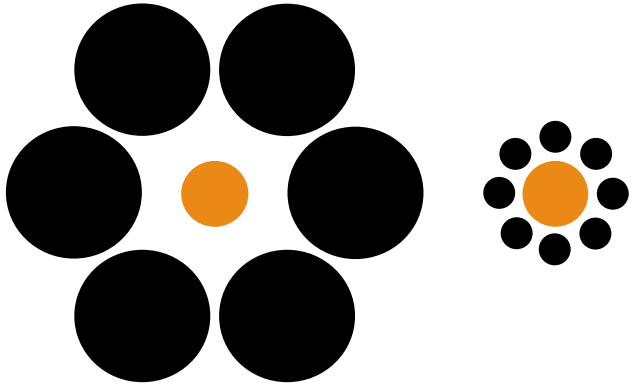
- Deskriptive Statistik
- Hypothesentest
- Segmentierung / Clusteranalyse
- Klassifikation / Regression

Heutiger Fokus: Explorative Datenanalyse
- Datendiagnose
- Datenvisualisierung
- Datentransformation

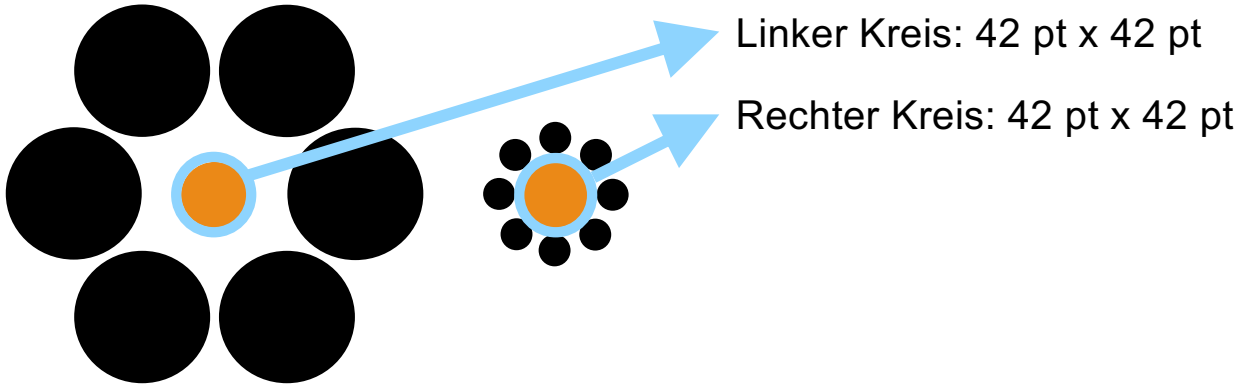
Quelle: Provost und Fawcett, "Data Science for Business"

Intuitive Entscheidungsfindung

- Viele Geschäftsentscheidungen werden aus dem Bauch heraus getroffen

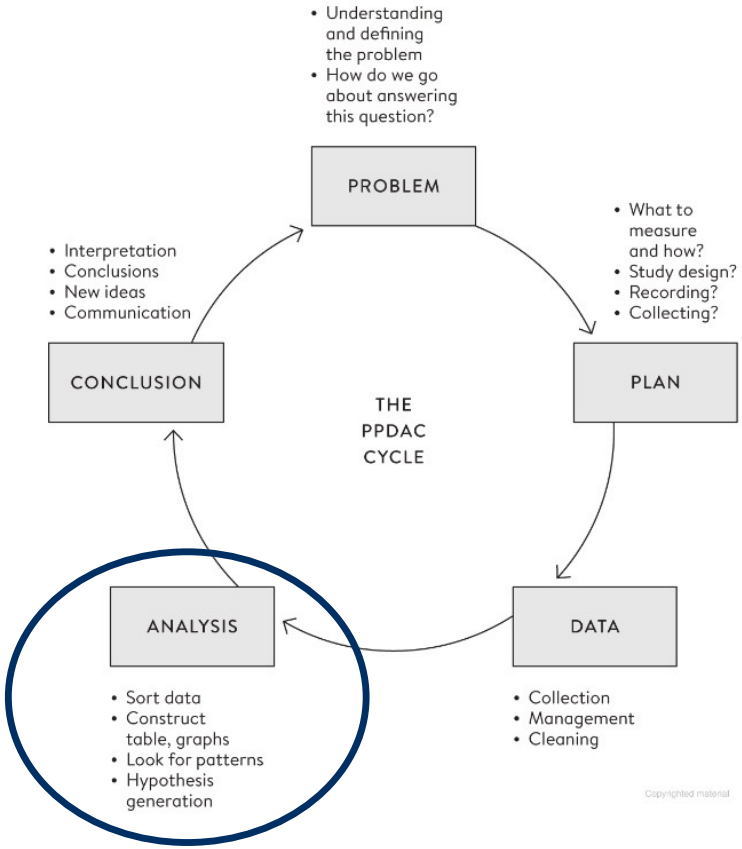


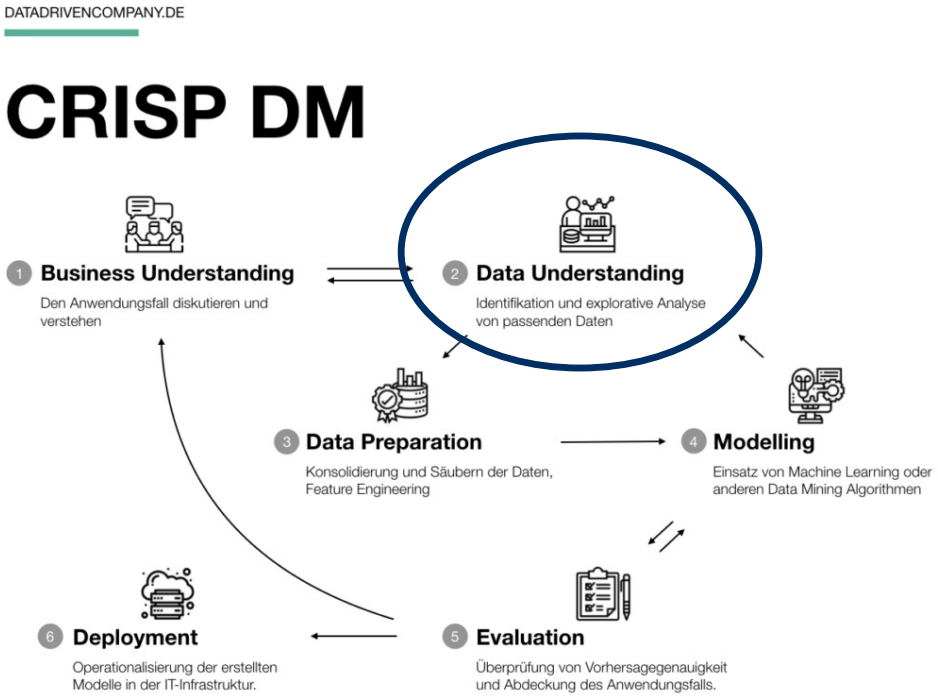
Faktenbasierte Entscheidungsfindung



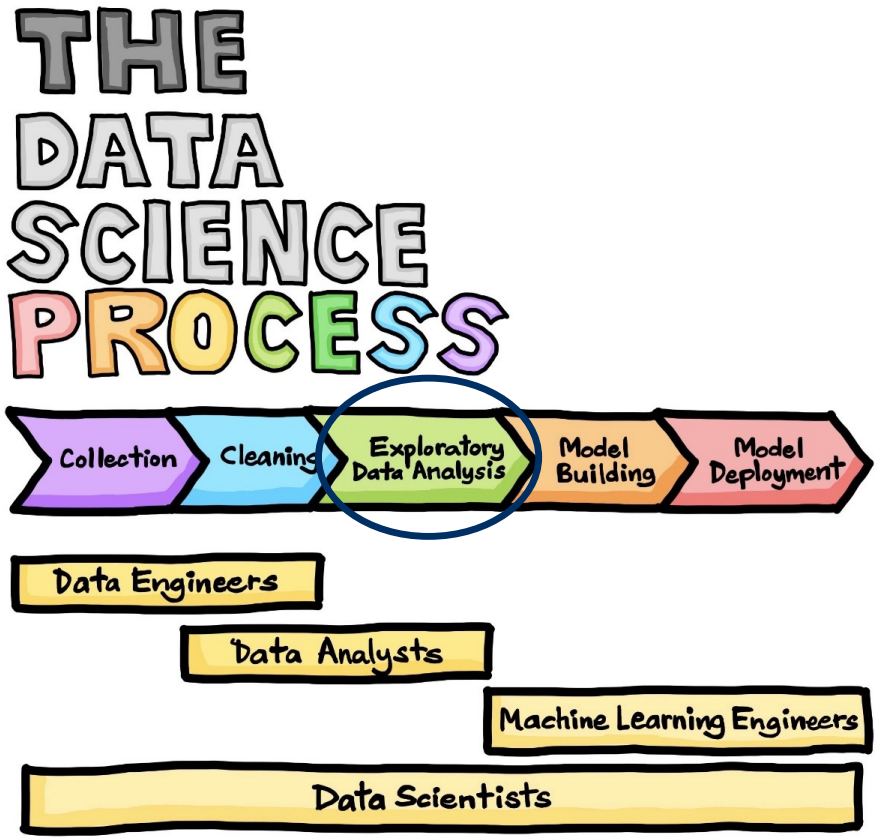
- Datenanalyse ist ein Weg um die Lücke zwischen intuitivem und faktenbasiertem Handeln zu schließen

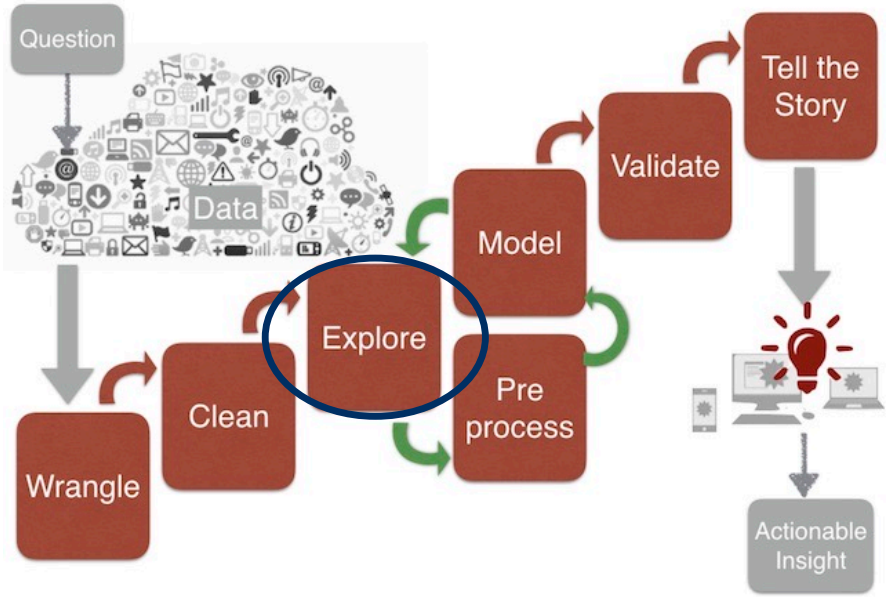
Datenanalyse, Datenverständnis, Datenexploration als Ausgangspunkt





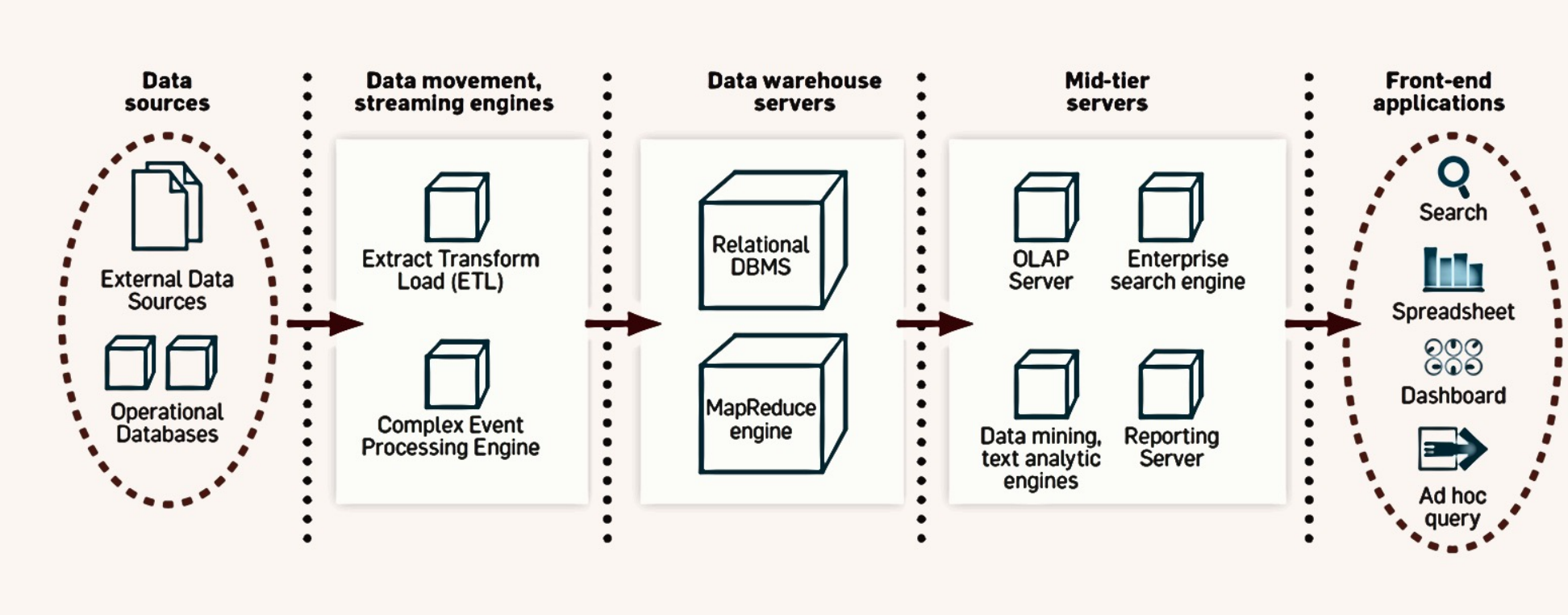
Datenanalyse, Datenverständnis, Datenexploration als Ausgangspunkt





Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

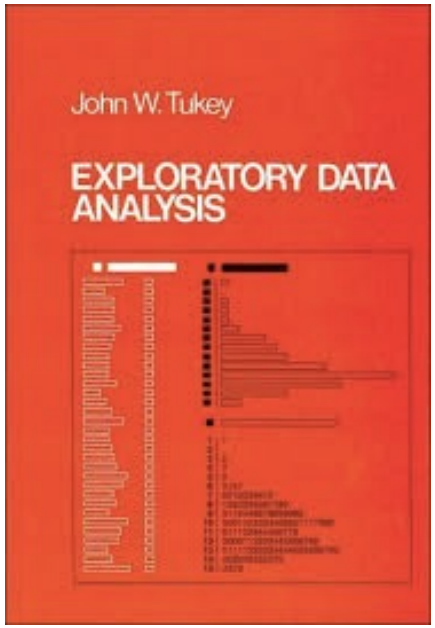
Business Intelligence Stack



Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

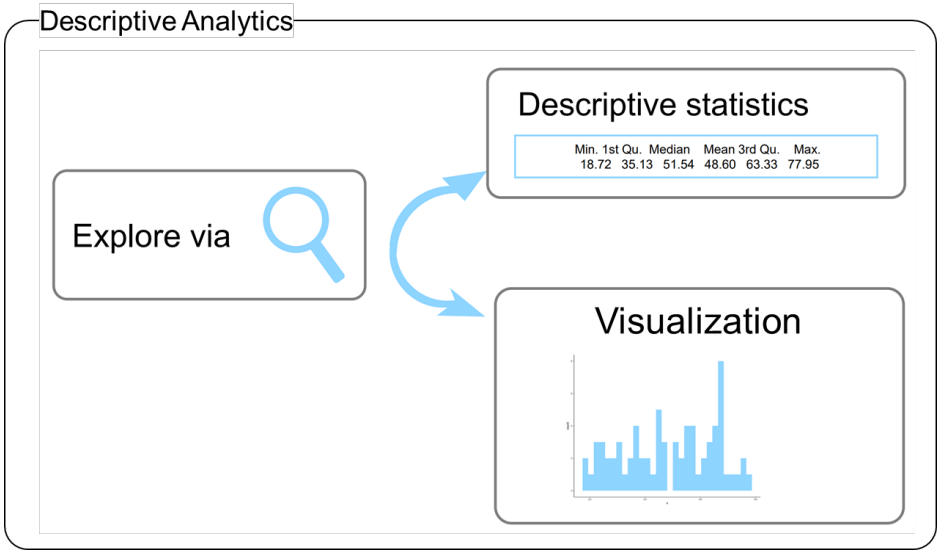
John Tukey 1977: Explorative Datenanalyse

- Basierend auf Erkenntnissen aus den Bell Labs schrieb John W. Tukey 1977 das Buch „Exploratory Data Analysis“
- Tukey vertrat die Ansicht, dass in der Statistik zu viel Wert auf statistische Hypothesentests (konfirmatorische Datenanalyse) gelegt wird
- Die Ziele der EDA sind:
  - Hypothesen über die Ursachen der beobachteten Phänomene aufzustellen
  - Bewertung der Annahmen, auf denen die statistische Inferenz basiert
  - Unterstützung bei der Auswahl geeigneter statistischer Werkzeuge und Techniken
  - Eine Grundlage für weitere Datenerhebungen durch Umfragen oder Experimente zu schaffen
  - Viele EDA-Techniken sind in das Data Mining übernommen worden.



Explorative Datenanalyse

- Einer der wichtigsten und am meisten übersehenen Teile der Statistik ist die Explorative Datenanalyse (EDA)
- Die Ziele der EDA sind:
  - Hypothesen vorschlagen
  - Annahmen zu bewerten
  - Unterstützung bei der Auswahl geeigneter statistischer Verfahren
  - Eine Grundlage für die weitere Datenerhebung zu schaffen
- Außerdem hilft es, ein Gefühl für die Zahlen zu bekommen
  - Leichter, Fehler zu finden
  - Leichter zu erraten, was tatsächlich passiert ist
  - Leichter, ungerade Ausreißer zu finden



Agenda

1. Einführung
2. **Beschreibende Statistik**
3. Datenvisualisierung eindimensionaler Daten
4. Datentransformationen
5. Datenvisualisierung mehrdimensionaler Daten

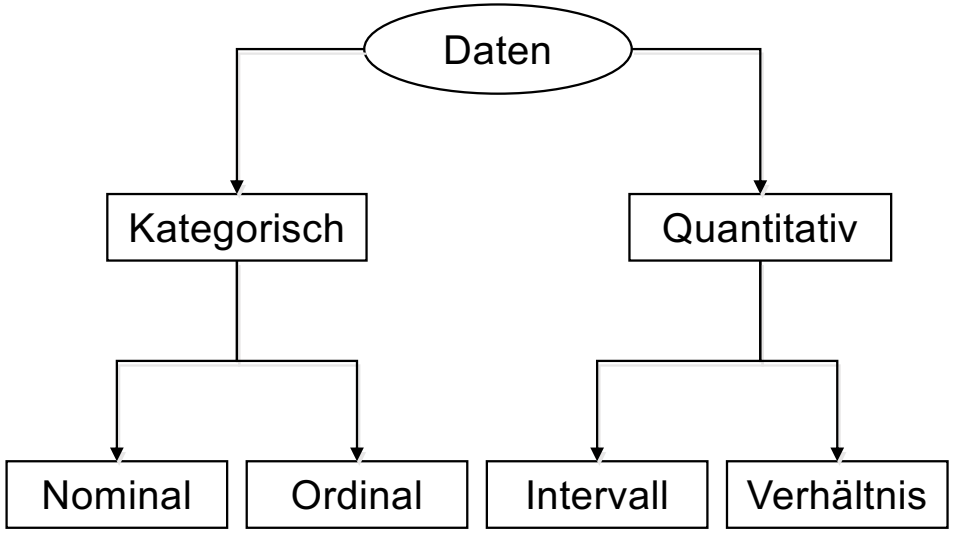
Skalen

- Messung: "Das Zuordnen von Zahlen zu Objekten, Ereignissen oder abstrakten Konzepten nach einem bekannten Satz von Regeln" (Stevens 1946).
- Die Skala bestimmt die Menge der in den Daten enthaltenen Informationen.
- Die Skala gibt an, welche Daten-zusammenfassung und statistischen Analysen am besten geeignet sind.
- So können die Daten kategorisiert, quantifiziert und/ oder analysiert werden, um sinnvolle Schlussfolgerungen zu ziehen.

- Nominalskala
  - Ein Maß für die Identität oder Kategorie
- Ordinalskala
  - Ein Maß für die Ordnung oder den Rang
- Intervall-Skala
  - Ein Maß für Ordnung und Menge
  - Differenzen können berechnet werden
  - Kann keine 'x-fache' Steigerung feststellen
- Verhältnisskala
  - Intervallskala mit einem absoluten Nullpunkt

Kategorische und quantitative Daten

- Kategorische (Qualitative) Daten
  - Bezeichnungen oder Namen, die zur Identifizierung eines Attributs eines jeden Elements verwendet werden
  - Kann entweder numerisch oder nicht numerisch sein
  - Geeignete statistische Analysen sind eher begrenzt
- Quantitative Daten
  - geben an, wie viele oder wie viel
  - Immer numerisch
  - Gewöhnliche arithmetische Operationen sind sinnvoll



Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Skalen und Verarbeitungsmöglichkeiten

- Welche statistische Analyse geeignet ist, hängt davon ab, ob die Daten für die Variable kategorisch oder quantitativ sind
- Im Allgemeinen gibt es mehr Alternativen für die statistische Analyse, wenn die Daten quantitativer Natur sind

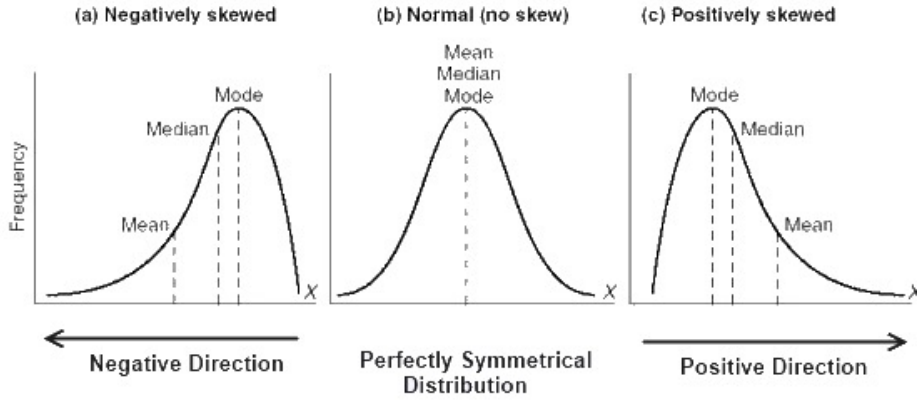
Provides:	Nominal	Ordinal	Interval	Ratio
The “order” of values is known		✔	✔	✔
“Counts,” aka “Frequency of Distribution”	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has “true zero”				✔

Nominal-, ordinal-, intervall- oder verhältnisskaliert?

- Blutlaktat-Konzentration (mmol.l-1)
- Profil der Stimmungszustände (Skala 1-7)
- Herzfrequenz (Schläge.min-1)
- Blutgruppe
- Bankdrücken 1RM (kg)
- Geburtsjahr (AD)
- Atmosphärischer Druck (mmHg)

Zusammenfassende deskriptive Statistik

- Zentrale Tendenz: Was sind die typischsten Werte?
- Variabilität: Wie variieren die Werte?
- Form: Sind die Werte symmetrisch oder asymmetrisch verteilt?



Zentrale Tendenz

- Es existieren drei gängige Maße der zentralen Tendenz, die alle versuchen, die grundlegende Frage zu beantworten, welcher Wert der "typischste" ist
  - Mittelwert (Durchschnitt aller Beobachtungen)
  - Median (mittlere Beobachtung)
  - Modus (tritt am häufigsten auf)
- Jedes dieser Maße kann für eine einzelne Variable oder über alle Variablen in einem Datensatz berechnet werden

- Die Variabilität kann auf verschiedene Arten zusammengefasst werden, die jeweils ein einzigartiges Verständnis für die Verteilung der Werte liefern.
  - Spannweite
  - Perzentile
  - Standardabweichung
  - Varianz
  - Variationskoeffizient

- Die Form wird typischerweise in Bezug auf Schiefe (Symmetrie) und Wölbung (wie spitz die Verteilung ist) beurteilt

5-Number-Summary

- Minimum
- 1. Quartil
- 2. Quartil (Median)
- 3. Quartil
- Maximum

```
results %>%
 filter(tournament=="UEFA Euro") %>%
 na.omit() %>%
 select(home_score, away_score) %>%
 summary()
```

```
 home_score      away_score
 Min.   :0.000   Min.   :0.000
 1st Qu.:0.000   1st Qu.:0.000
 Median :1.000   Median :1.000
 Mean   :1.307   Mean   :1.095
 3rd Qu.:2.000   3rd Qu.:2.000
 Max.   :6.000   Max.   :5.000
```

https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

5-Number-Summary

- Minimum
- 1. Quartil
- 2. Quartil (Median)
- 3. Quartil
- Maximum

```
results %>%
 filter(tournament=="UEFA Euro") %>%
 filter(home_team == "Germany") %>%
 na.omit() %>%
 select(home_score, away_score) %>%
 summary()
```

```
 home_score      away_score
 Min.   :0.000   Min.   :0.0000
 1st Qu.:1.000   1st Qu.:0.0000
 Median :1.500   Median :1.0000
 Mean   :1.583   Mean   :0.8333
 3rd Qu.:2.000   3rd Qu.:1.2500
 Max.   :4.000   Max.   :2.0000
```

5-Number-Summary

- Minimum
- 1. Quartil
- 2. Quartil (Median)
- 3. Quartil
- Maximum

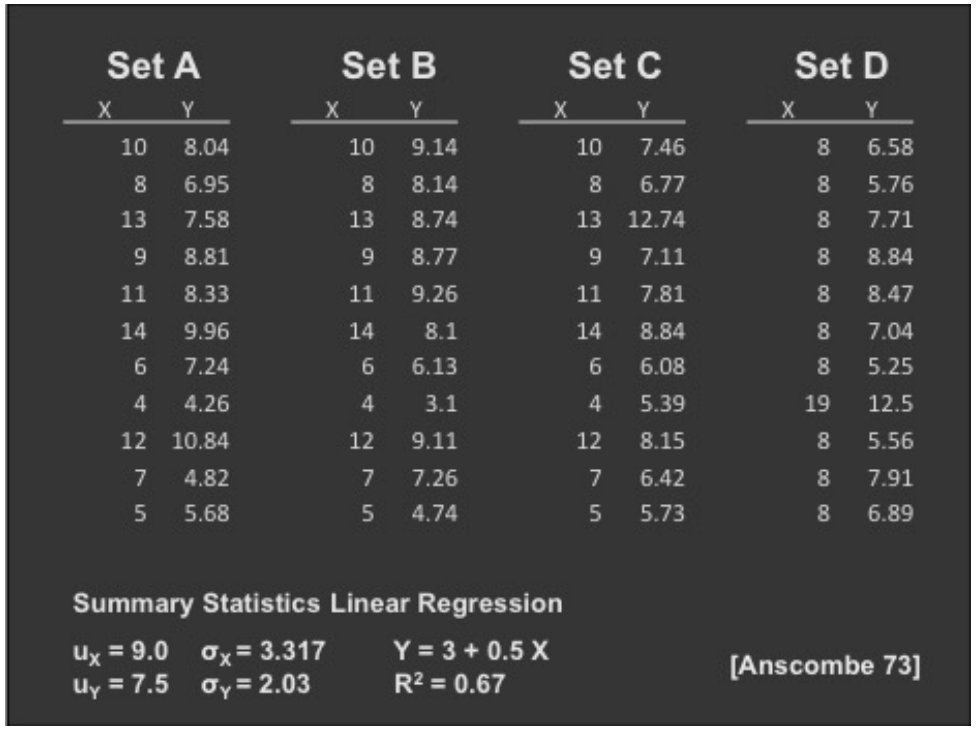
```
results %>%
 filter(tournament=="UEFA Euro") %>%
 filter(away_team == "Germany") %>%
 na.omit() %>%
 select(home_score, away_score) %>%
 summary()
```

```
 home_score     away_score
 Min.   :0.00   Min.   :0.00
 1st Qu.:0.00   1st Qu.:0.00
 Median :1.00   Median :1.00
 Mean   :1.12   Mean   :1.36
 3rd Qu.:2.00   3rd Qu.:2.00
 Max.   :3.00   Max.   :4.00
```

Anscombe-Quartett

- Das Anscombe-Quartett besteht aus vier Datensätzen, die nahezu identische einfache deskriptive Statistiken aufweisen
- Jeder Datensatz besteht aus elf (x,y) Punkten

Eigenschaft	Wert
Mittelwert von *x*	9 (exakt)
Varianz von *x*	11 (exakt)
Mittelwert von *y*	7,50 (auf 2 Stellen)
Varianz von *y*	4,122 oder 4,127 (auf 3 Stellen)
Korrelation zwischen *x* und *y*	0,816 (auf 3 Stellen)
Lineare Regression	*y* = 3,00 + 0,500*x* (auf 2 bzw. 3 Stellen)



Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Summary Statistics Linear Regression

$u_X = 9.0$ $\sigma_X = 3.317$ $Y = 3 + 0.5\ X$

$u_Y = 7.5$ $\sigma_Y = 2.03$ $R^2 = 0.67$

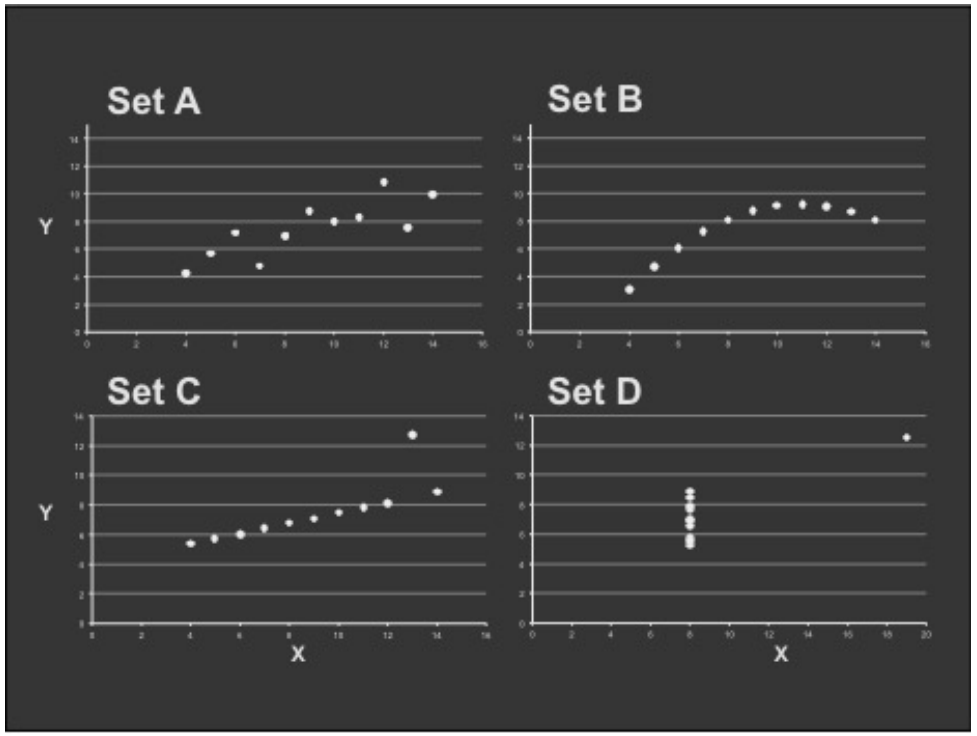
[Anscombe 73]

!

Die Grenzen der Statistik

Vier Datensätze mit zwei Variablen x und y liegen vor. Beschreibende Statistiken:

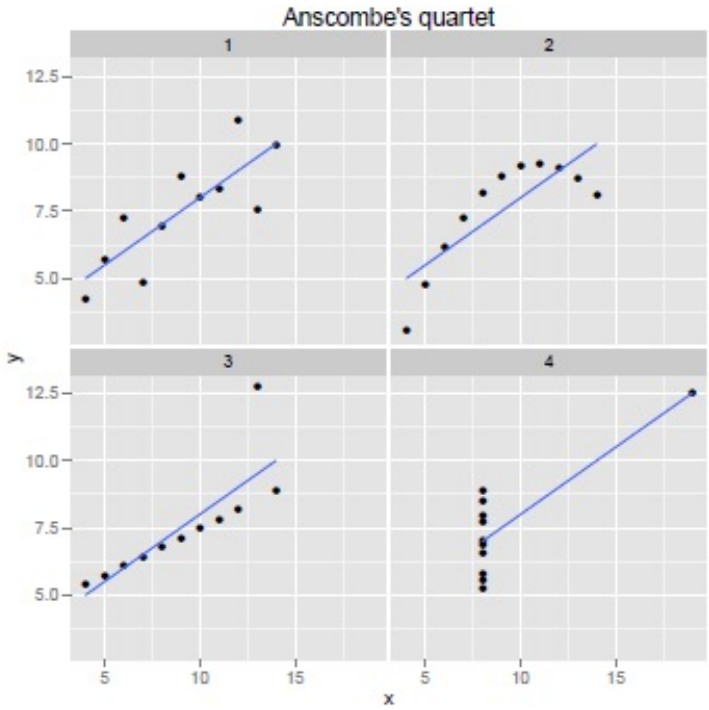
Eigenschaft	Wert
Mittelwert von *x*	9 (exakt)
Varianz von *x*	11 (exakt)
Mittelwert von *y*	7,50 (auf 2 Stellen)
Varianz von *y*	4,122 oder 4,127 (auf 3 Stellen)
Korrelation zwischen *x* und *y*	0,816 (auf 3 Stellen)
Lineare Regression	$y = 3{,}00 + 0{,}500x$ (auf 2 bzw. 3 Stellen)



!!

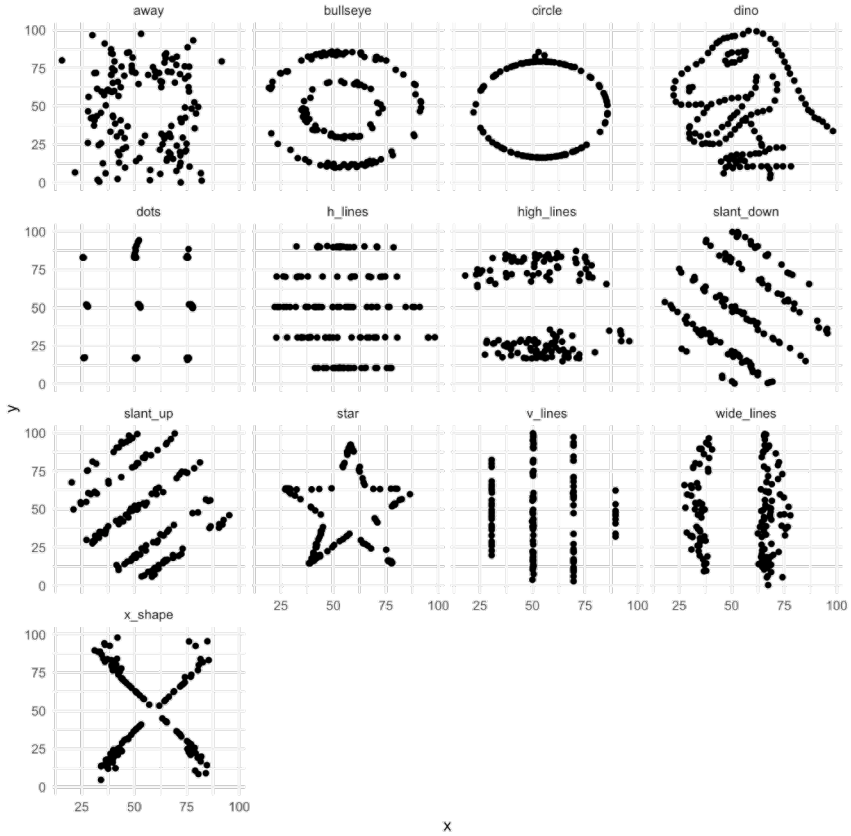
Die Grenzen der Statistik (1)

- Das Quartett wurden 1973 von dem Statistiker Francis Anscombe konstruiert
- Ziel ist es, die Wichtigkeit der grafischen Darstellung vor der Analyse von Daten zu demonstrieren
- Hervorhebung des Effekts von Ausreißern und anderen einflussreichen Beobachtungen auf statistische Eigenschaften



Die Grenzen der Statistik (2)

ID	N	$\bar{X}$	$\bar{Y}$	σ_X	σ_Y	R
1	142	54.3	47.8	16.8	26.9	-0.064
2	142	54.3	47.8	16.8	26.9	-0.069
3	142	54.3	47.8	16.8	26.9	-0.068
4	142	54.3	47.8	16.8	26.9	-0.064
5	142	54.3	47.8	16.8	26.9	-0.060
6	142	54.3	47.8	16.8	26.9	-0.062
7	142	54.3	47.8	16.8	26.9	-0.069
8	142	54.3	47.8	16.8	26.9	-0.069
9	142	54.3	47.8	16.8	26.9	-0.069
10	142	54.3	47.8	16.8	26.9	-0.063
11	142	54.3	47.8	16.8	26.9	-0.069
12	142	54.3	47.8	16.8	26.9	-0.067
13	142	54.3	47.8	16.8	26.9	-0.066



Die Grenzen der Statistik (3)

“Never try to walk across a river just because it has an average depth of four feet.”

Milton Friedman

Die Grenzen der Statistik (3) – “Flaw of Averages"

Die Familie mit 1½ Kindern

Oft ist das durchschnittliche Szenario, wie die durchschnittliche Familie mit 1½ Kindern, nicht existent. Zum Beispiel kann eine Bank zwei Hauptgruppen junger Kunden haben: Studenten mit einem Durchschnittseinkommen von 10.000 $ und junge Berufstätige mit einem Durchschnittseinkommen von 70.000 $. Wäre es für die Bank sinnvoll, Produkte oder Dienstleistungen für Kunden mit einem Durchschnittseinkommen von $40.000 zu entwickeln?

Warum alles hinter dem Zeitplan liegt

Betrachten Sie ein typisches Projektmanagement-Problem: Wenn Sie jede Aufgabe auf ihren Durchschnitt setzen, wird das Projekt in sechs Monaten abgeschlossen, aber die Chance, dass alle zehn Aufgaben ihren Durchschnitt oder früher erreichen, ist dieselbe wie das Umdrehen von zehn Köpfen in einer Reihe, so dass die Chance, in sechs Monaten fertig zu werden, weniger als eins zu tausend beträgt.

Der Eierkorb

Überlegen Sie, ob Sie zehn Eier in denselben Korb legen oder jeweils eines in getrennte Körbe. Wenn es eine 10-prozentige Chance gibt, dass ein bestimmter Korb herunterfällt, dann ergeben beide Strategien durchschnittlich neun unversehrte Eier. Allerdings hat die erste Strategie eine 10-prozentige Chance, alle Eier zu verlieren, während die zweite nur eine Chance von 1 zu 10.000.000.000 hat, alle Eier zu verlieren.

Ignorieren von (Kapazitäts-)Beschränkungen.

Betrachten Sie eine Firma mit einer Kapazität, die dem Durchschnitt der unsicheren zukünftigen Nachfrage entspricht. Wenn die tatsächliche Nachfrage geringer ist als der Durchschnitt, sinkt der Gewinn. Aber wenn die Nachfrage größer als der Durchschnitt ist, sind die Verkäufe durch die Kapazität eingeschränkt. Somit gibt es einen Nachteil ohne einen damit verbundenen Vorteil, und der durchschnittliche Gewinn ist geringer als der mit der durchschnittlichen Nachfrage verbundene Gewinn.

flawofaverages.com/book/index.html

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Agenda

1. Einführung
2. Beschreibende Statistik
3. Datenvisualisierung eindimensionaler Daten
4. Datentransformationen
5. Datenvisualisierung mehrdimensionaler Daten

Was ist Datenvisualisierung?

„Transformation des Symbolischen ins Geometrische“

(McCormick et al., 1987)

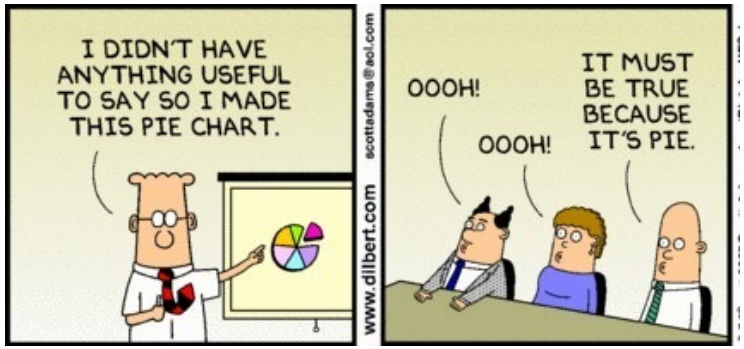
"Grafische Exzellenz ist etwas, das dem Betrachter die größte Anzahl von Ideen in der kürzesten Zeit mit der wenigsten Tinte auf dem kleinsten Raum vermittelt."|
(Edward Tufte)

„Die Darstellung von Informationen mit Hilfe von räumlichen oder grafischen Darstellungen, um Vergleiche, Mustererkennung, Erkennung von Veränderungen und andere kognitive Fähigkeiten durch den Einsatz des visuellen Systems zu erleichtern.“

Ziele der Datenvisualisierung?

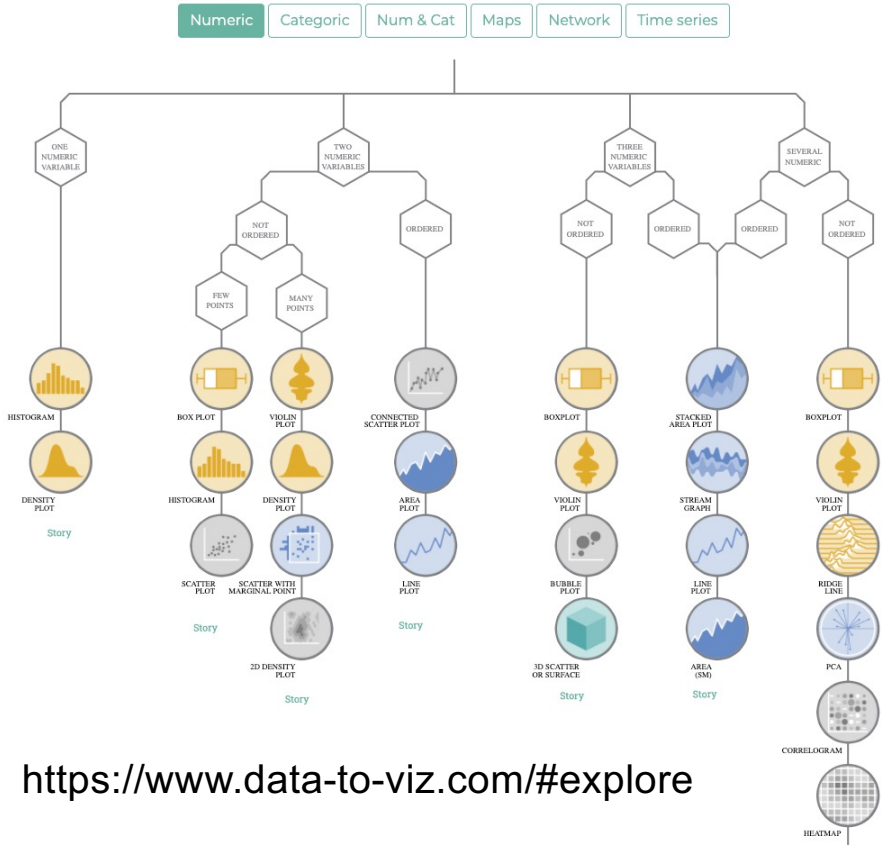
- Erforschen/Berechnen
  - Analysieren
  - Über Informationen nachdenken

- Kommunizieren
  - Erläutern
  - Entscheidungen treffen
  - Über Informationen urteilen



Nützliche Diagramme für eindimensionale Daten

- Punktdiagramm
- Jitter-Plot
- Box-Plot
- Histogramm
- Kumulative Verteilungsfunktion



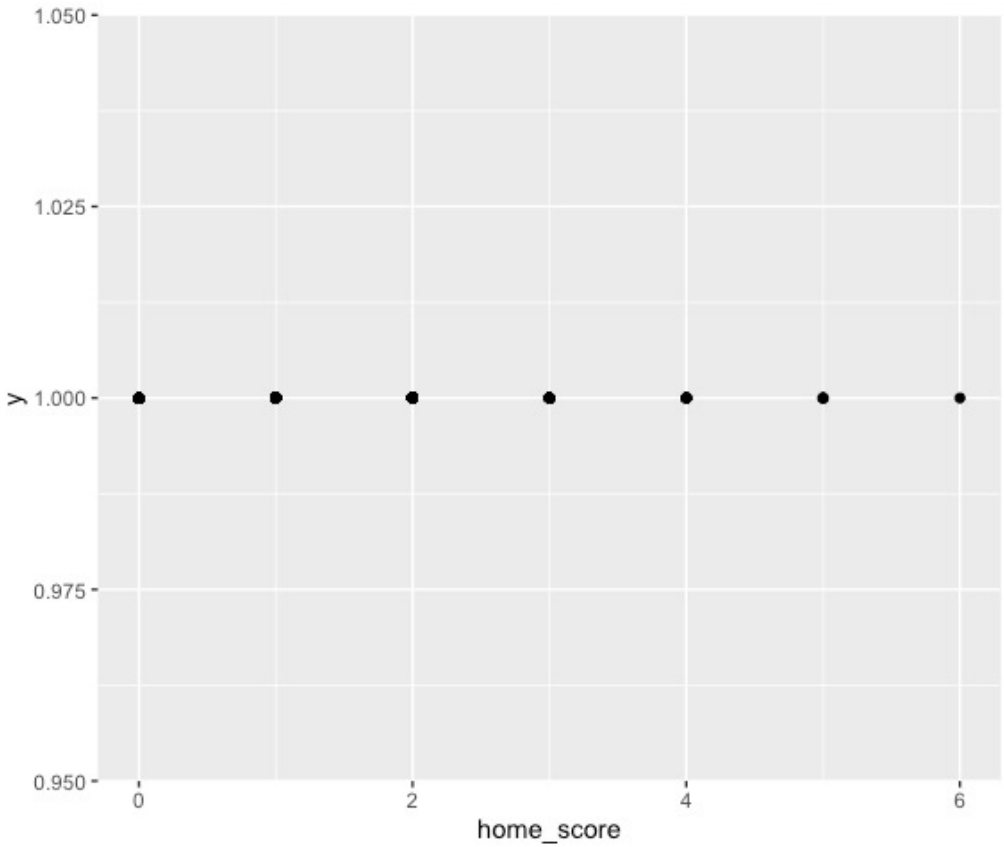
https://www.data-to-viz.com/#explore

Nützliche Diagramme für eindimensionale Daten

- **Punktdiagramm**
- Jitter-Plot
- Box-Plot
- Histogramm
- Kumulative Verteilungsfunktion

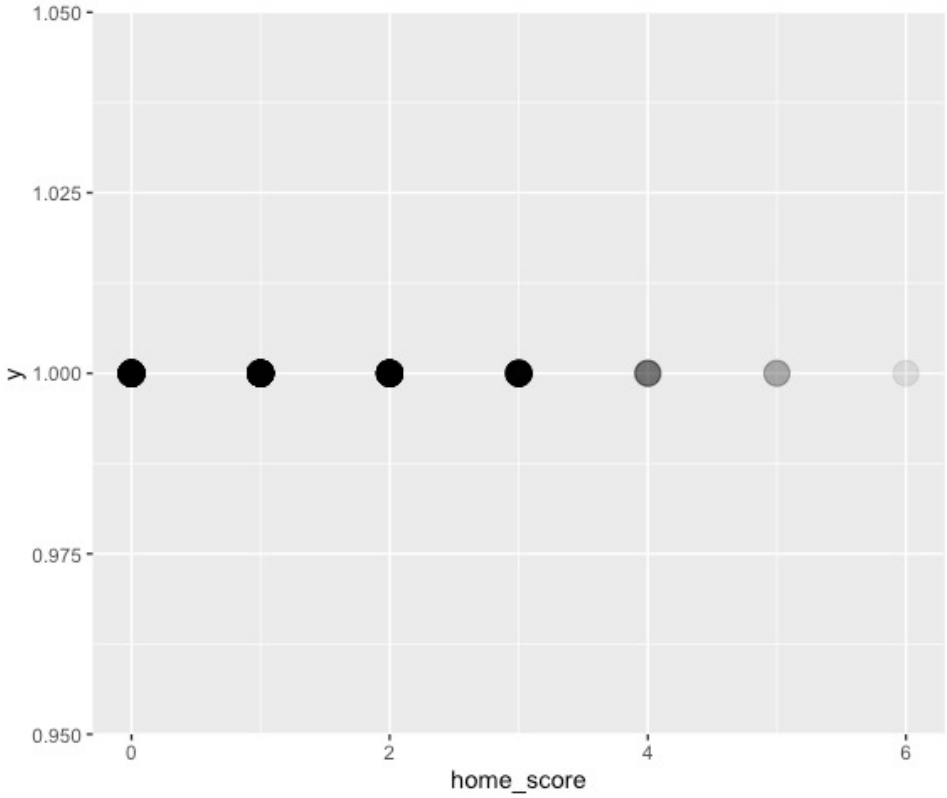
```
results %>%
 filter(tournament=="UEFA Euro") %>%
 na.omit() %>%
 select(home_score) %>%
 ggplot(aes(x=1, y=home_score)) +
 geom_point()
```

Oha!



Ok…

```
results %>%
  filter(tournament=="UEFA Euro") %>%
  na.omit() %>%
  select(home_score) %>%
  ggplot(aes(x=home_score, y=1)) +
  geom_point(size = 5, alpha=0.1)
```



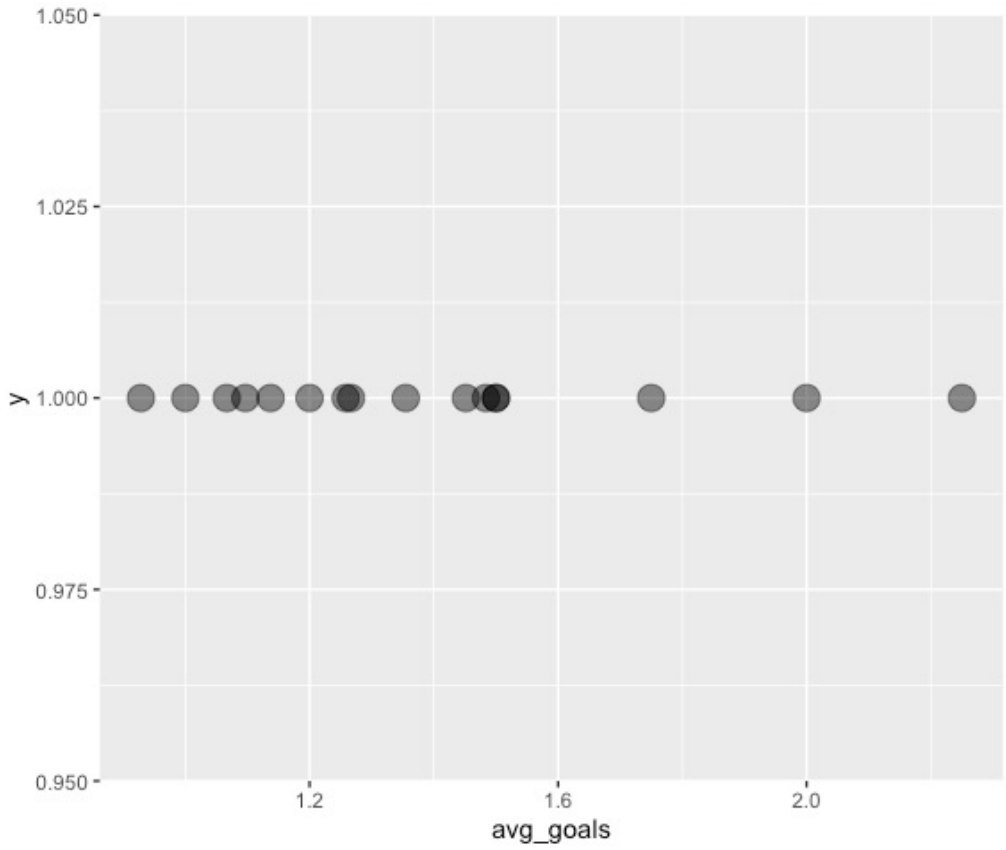
Zweiter Versuch

- **Punktdiagramm**
- Jitter-Plot
- Box-Plot
- Histogramm
- Kumulative Verteilungsfunktion

```
results %>%
 filter(tournament=="UEFA Euro") %>%
 mutate(year = lubridate::year(date)) %>%
 na.omit() %>%
 select(year, home_score) %>%
 group_by(year) %>%
 summarise(avg_goals = mean(home_score)) %>%
 ggplot(aes(x=1, y=avg_goals)) +
 geom_point()
```

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Besser!

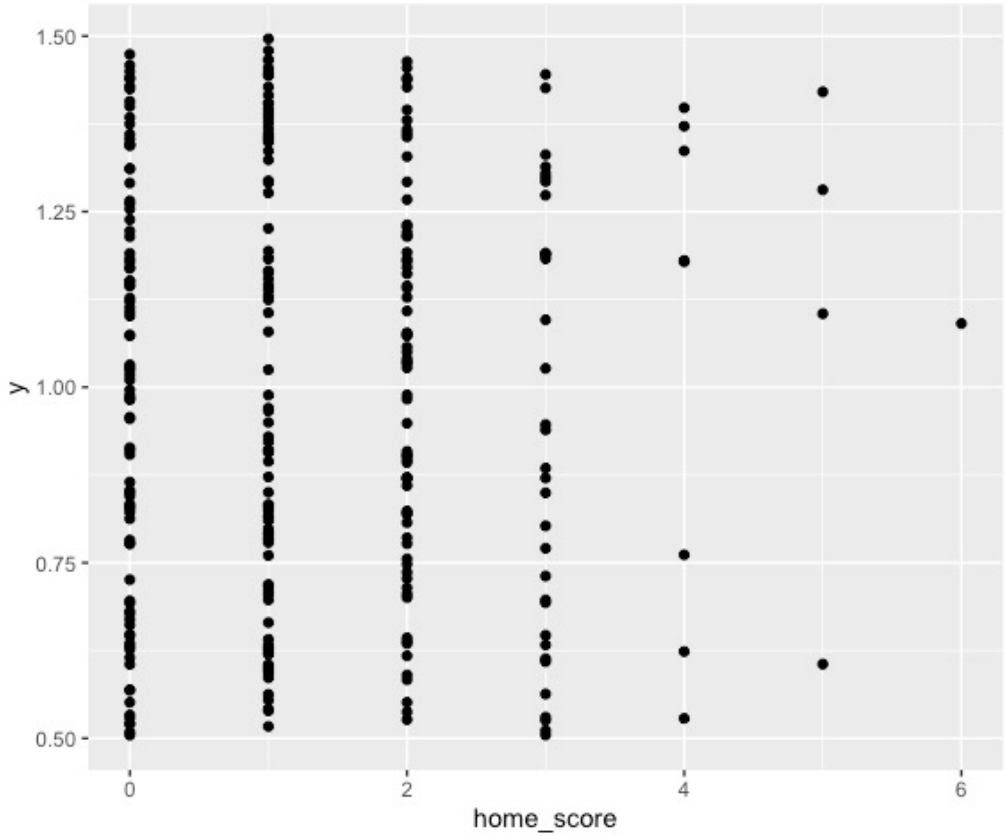


Nützliche Diagramme für eindimensionale Daten

- Punktdiagramm
- **Jitter-Plot**
- Box-Plot
- Histogramm
- Kumulative Verteilungsfunktion

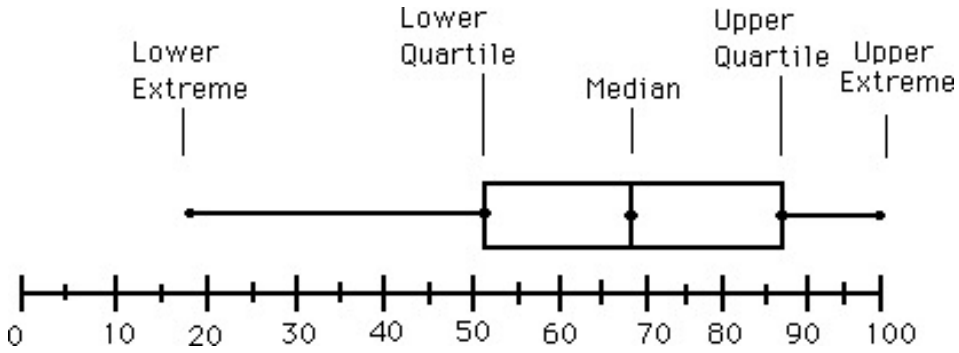
```
results %>%
 filter(tournament=="UEFA Euro") %>%
 na.omit() %>%
 select(home_score) %>%
 ggplot(aes(x=home_score, y=1)) +
 geom_jitter(width = 0, height = 0.5)
```

Viel besser!



Nützliche Diagramme für eindimensionale Daten

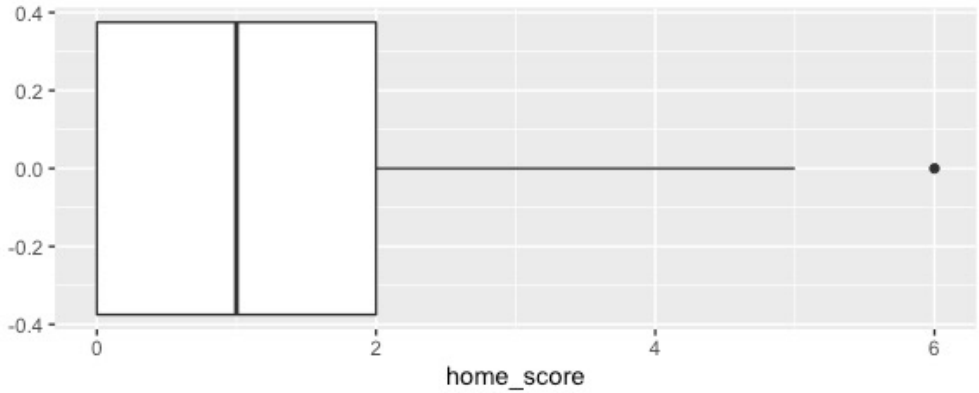
- Punktdiagramm
- Jitter-Plot
- **Box-Plot**
- Histogramm
- Kumulative Verteilungsfunktion



Nützliche Diagramme für eindimensionale Daten

- Punktdiagramm
- Jitter-Plot
- **Box-Plot**
- Histogramm
- Kumulative Verteilungsfunktion

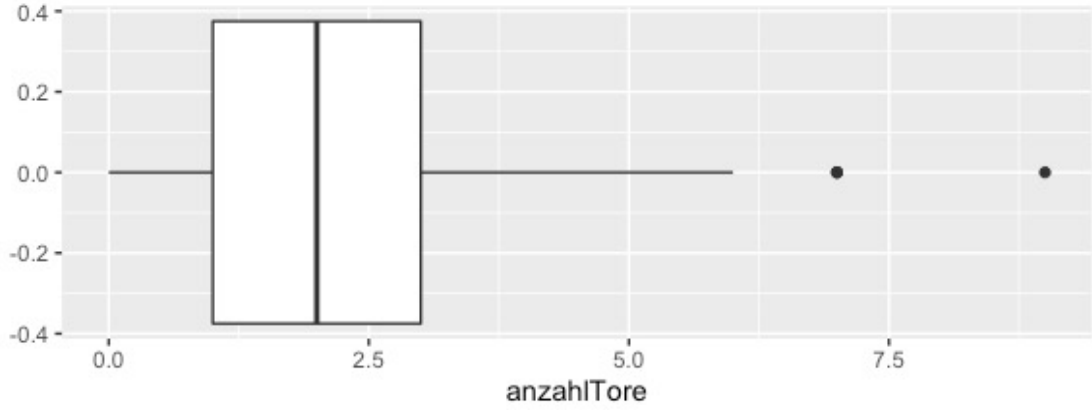
```
results %>%
 filter(tournament=="UEFA Euro") %>%
 na.omit() %>%
 select(home_score) %>%
 ggplot(aes(x=home_score, y=1)) +
 geom_jitter(width = 0, height = 0.5)
```



Nützliche Diagramme für eindimensionale Daten

- Punktdiagramm
- Jitter-Plot
- **Box-Plot**
- Histogramm
- Kumulative Verteilungsfunktion

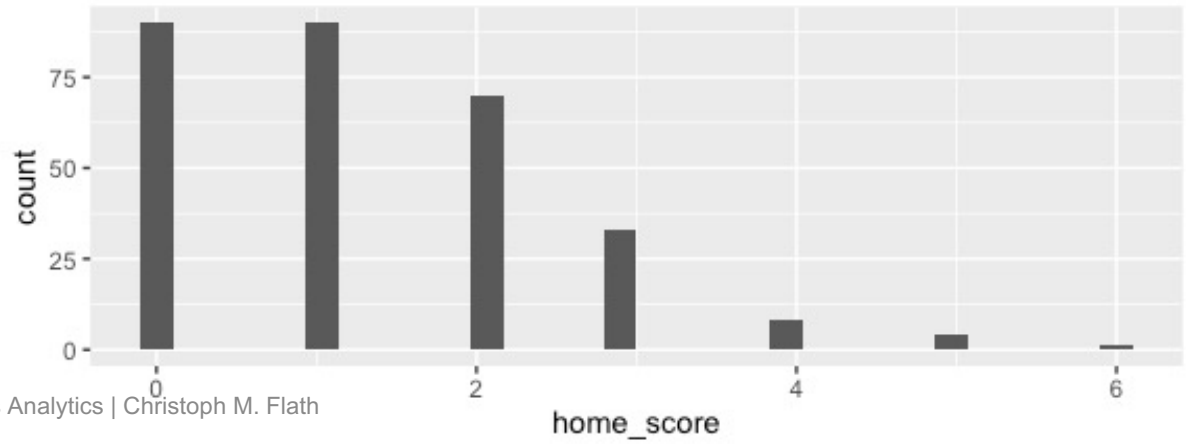
```
results %>%
 filter(tournament=="UEFA Euro") %>%
 na.omit() %>%
 mutate(anzahlTore = home_score + away_score) %>%
 ggplot(aes(x=anzahlTore)) +
 geom_boxplot()
```



Nützliche Diagramme für eindimensionale Daten

- Punktdiagramm
- Jitter-Plot
- Box-Plot
- **Histogramm**
- Kumulative Verteilungsfunktion

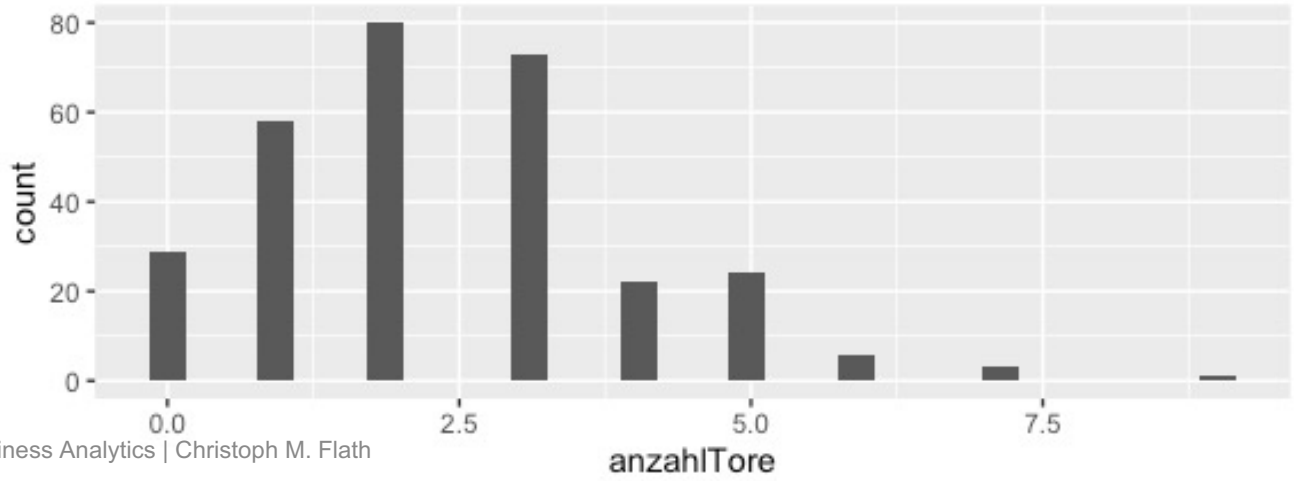
```
results %>%
  filter(tournament=="UEFA Euro") %>%
  na.omit() %>%
  ggplot(aes(x=home_score)) +
  geom_histogram()
```



Nützliche Diagramme für eindimensionale Daten

- Punktdiagramm
- Jitter-Plot
- Box-Plot
- **Histogramm**
- Kumulative Verteilungsfunktion

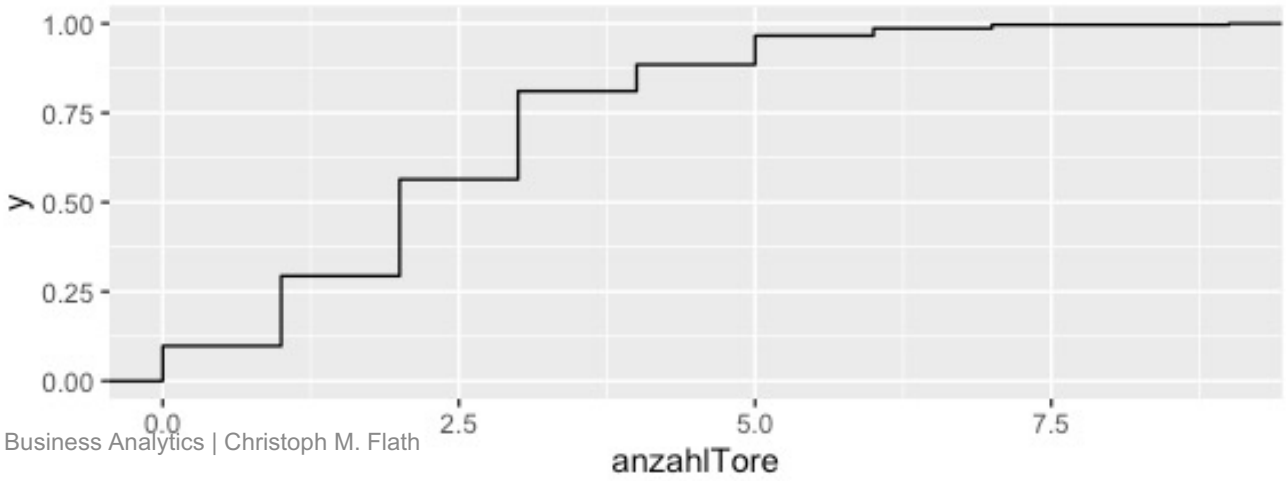
```
results %>%
 filter(tournament=="UEFA Euro") %>%
 na.omit() %>%
 mutate(anzahlTore = home_score + away_score) %>%
 ggplot(aes(x=anzahlTore)) +
 geom_histogram()
```



Nützliche Diagramme für eindimensionale Daten

- Punktdiagramm
- Jitter-Plot
- Box-Plot
- Histogramm
- **Kumulative Verteilungsfunktion**

```
results %>%
 filter(tournament=="UEFA Euro") %>%
 na.omit() %>%
 mutate(anzahlTore = home_score + away_score) %>%
 ggplot(aes(x=anzahlTore)) +
 stat_ecdf(geom = "step")
```



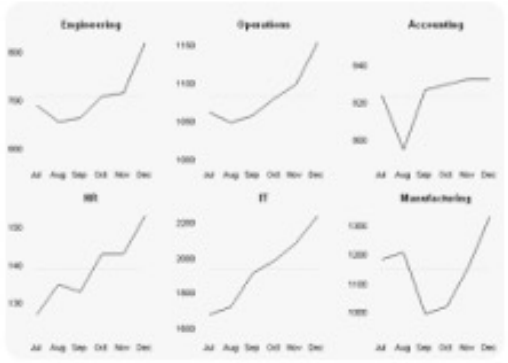
Was mache ich wenn ich mehrere Variablen habe?

Boston ist eine Datenbank mit Informationen über Gebiete rund um die Stadt Boston und die mittleren Hauspreise.

crim Pro-Kopf-Verbrechensrate nach Stadt.
zn Anteil der Wohnbauflächen, die für Grundstücke über 25.000 sq.ft.
indus Anteil der Nicht-Einzelhandelsgeschäftsflächen pro Stadt.
chas Charles River Dummy-Variable (= 1, wenn Trakt an Fluss grenzt; 0 sonst).
nox Konzentration von Stickoxiden (Teile pro 10 Millionen).
rm durchschnittliche Anzahl der Zimmer pro Wohnung.
age Anteil der Eigentumswohnungen, die vor 1940 gebaut wurden.
dis gewichteter Mittelwert der Entfernungen zu fünf Bostoner Beschäftigungszentren.
rad Index der Erreichbarkeit von Radialautobahnen.
tax Vollwertiger Grundsteuersatz pro 10.000 $.
ptratio Schüler-Lehrer-Verhältnis nach Stadt.
black 1000(Bk-0.63)21000(Bk-0.63)2 wobei BkBk der Anteil der Schwarzen nach Stadt ist.
lstat unterer Status der Bevölkerung (Prozent).
medv medianer Wert der Eigenheime in 1000er Dollar.

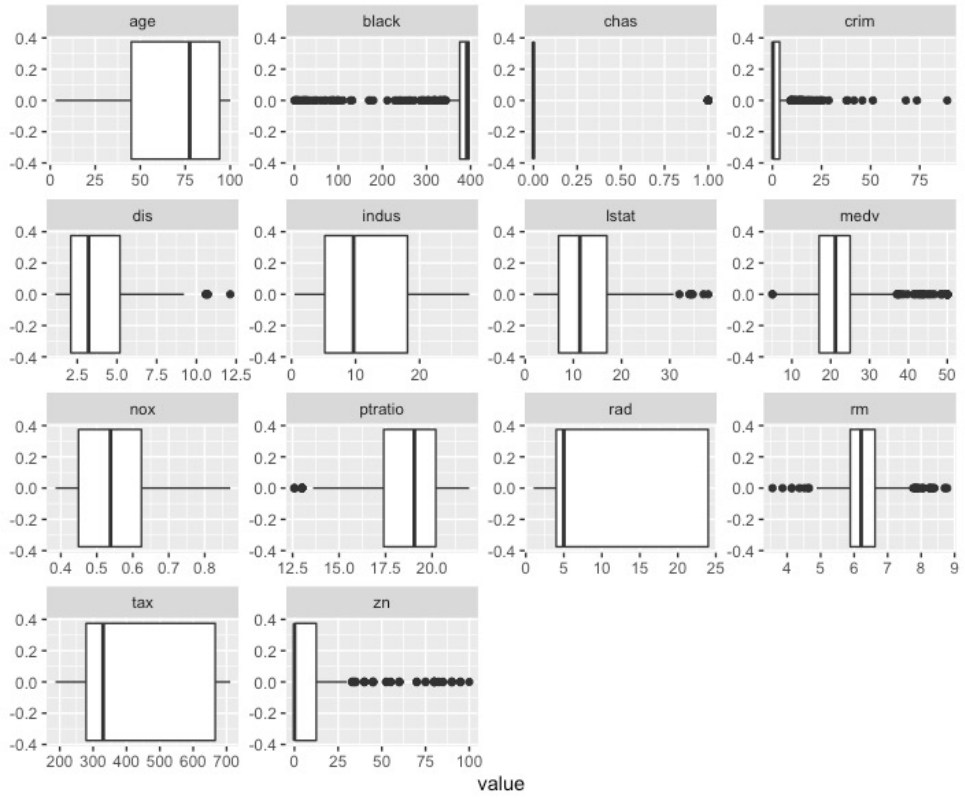
Trellis Plots

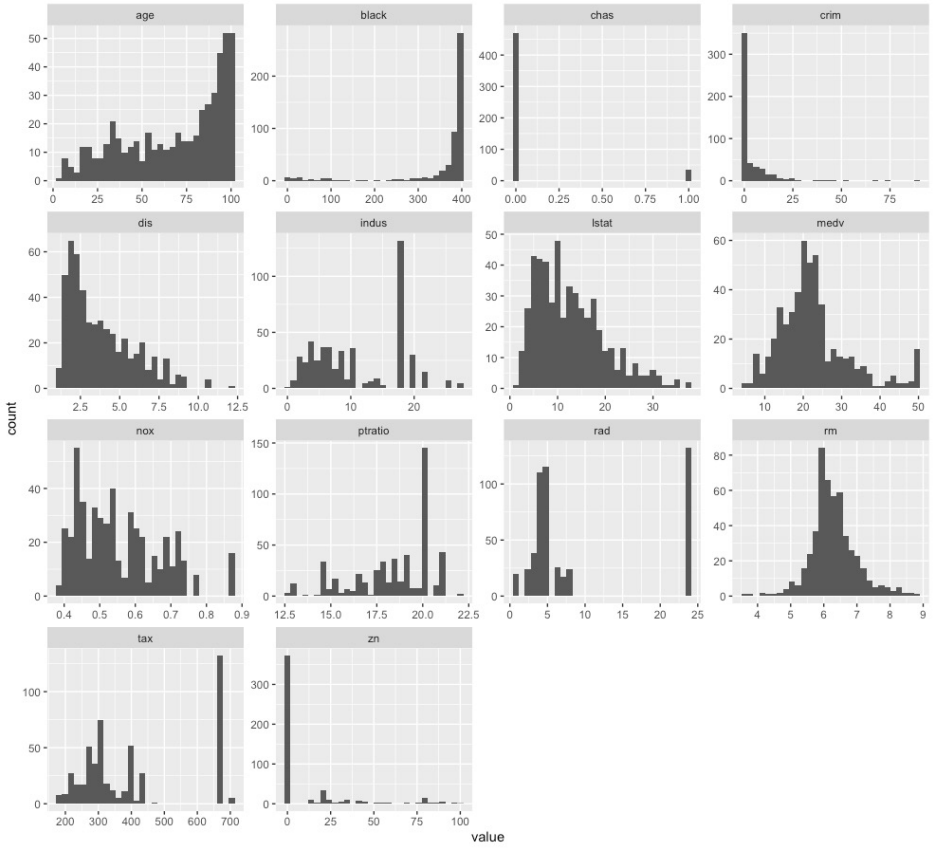
A **small multiple** (sometimes called **trellis** chart, lattice chart, grid chart, or panel chart) is a series of similar graphs or charts using the same scale and axes, allowing them to be easily compared. It uses **multiple** views to show different partitions of a dataset. The term was popularized by Edward Tufte.



Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Es funktioniert!





Agenda

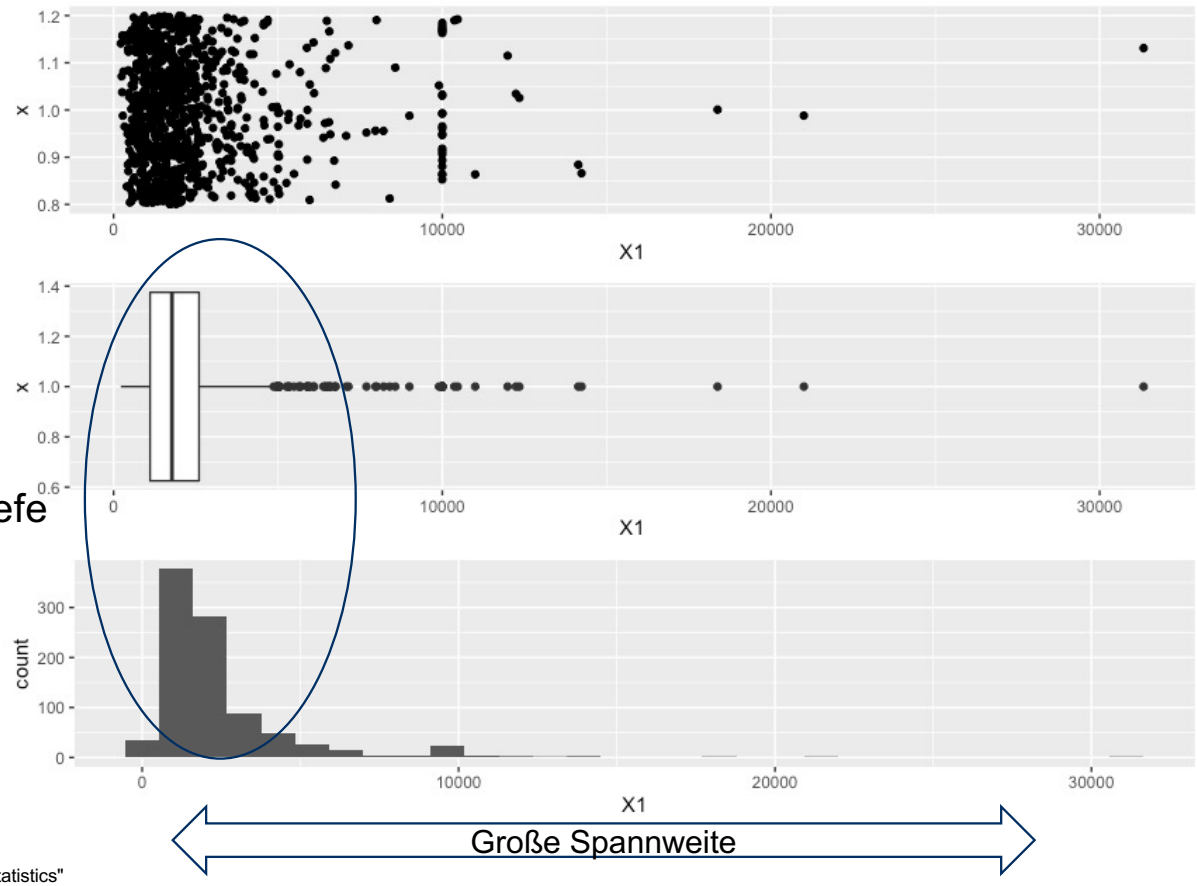
1. Einführung
2. Beschreibende Statistik
3. Datenvisualisierung eindimensionaler Daten
4. Datentransformationen
5. Datenvisualisierung mehrdimensionaler Daten

Die Weisheit der Vielen

Wie viele Jelly Beans sind in diesem Glas?

(wir sparen uns heute das Befragen denn es ist nicht das Thema dieser VL)

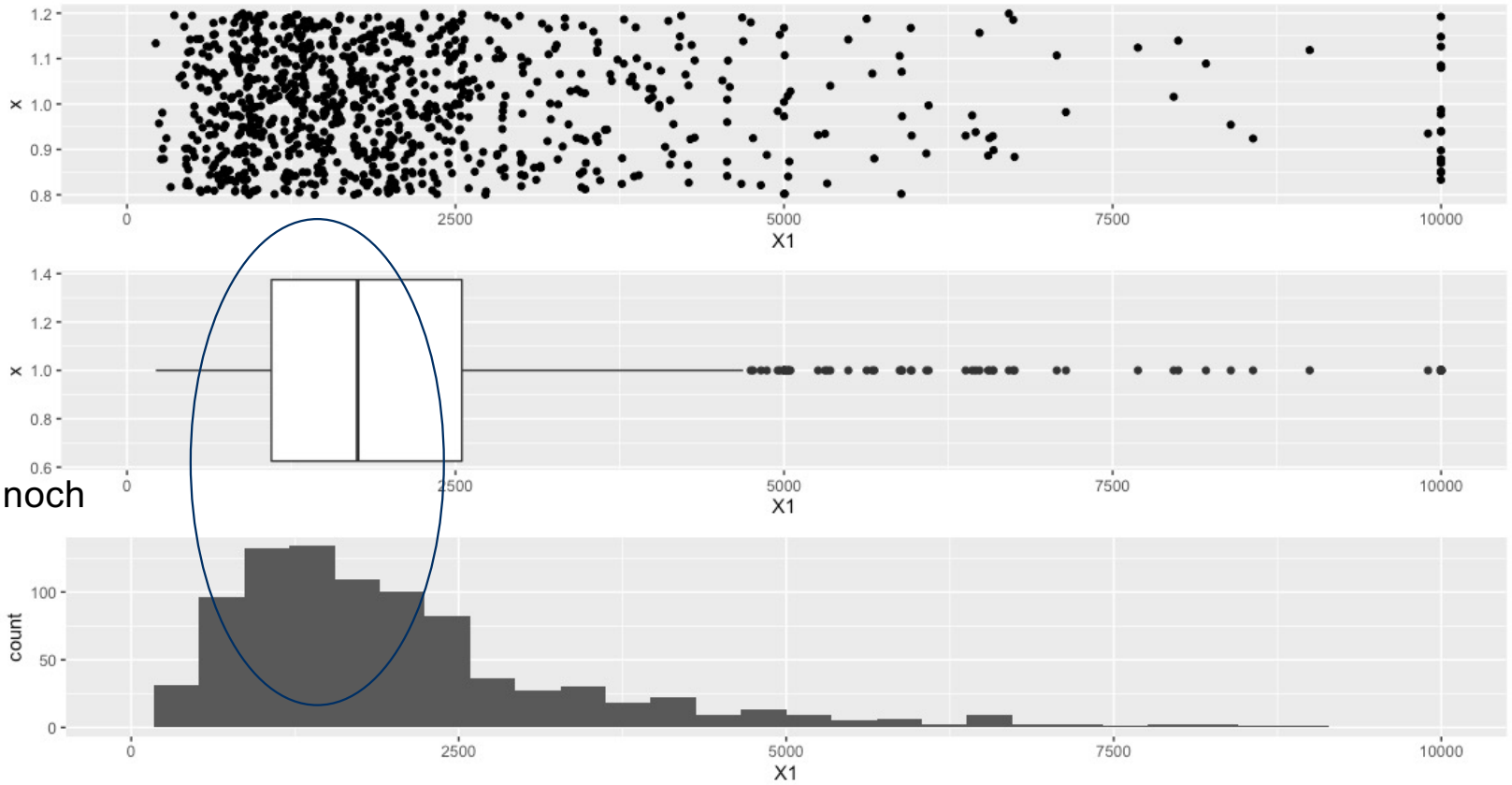
Die Plots helfen nur bedingt weiter



Sehr schiefe Verteilung

Quelle: Spiegelhalter, "The Art of Statistics"

Ausreißer entfernen (x<=10000) – Verbesserung aber um welchen Preis?



Immer noch schief

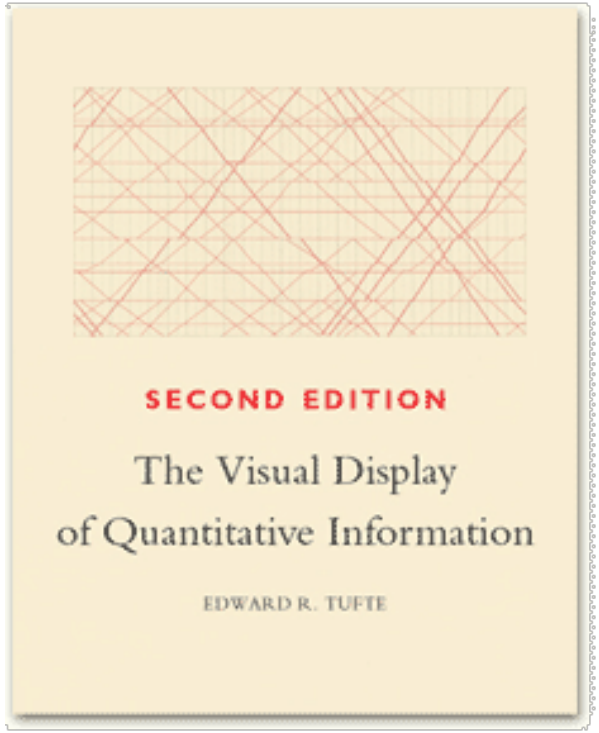
Quelle: Spiegelhalter, "The Art of Statistics"

Ein Manifest für gute Datenvisualisierung

- Wo möglich müssen alle Daten gezeigt werden!

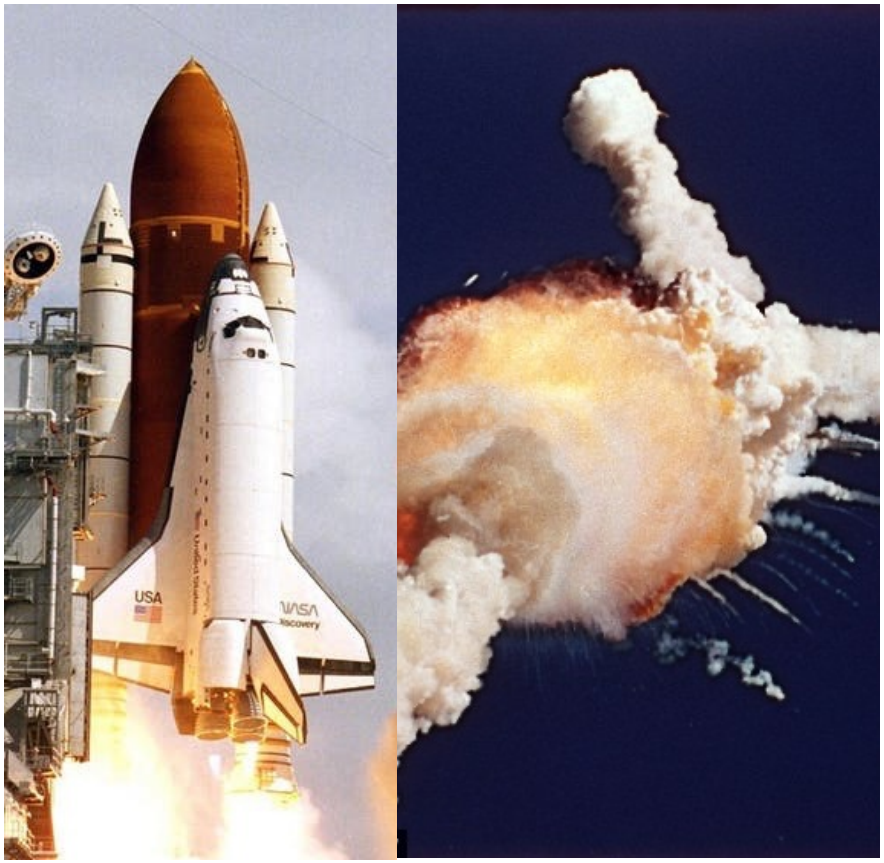
- Maximieren Sie das Verhältnis von Daten zu Tinte
  - Löschen von Nicht-Daten-Tinte
  - Überflüssige Tinte löschen

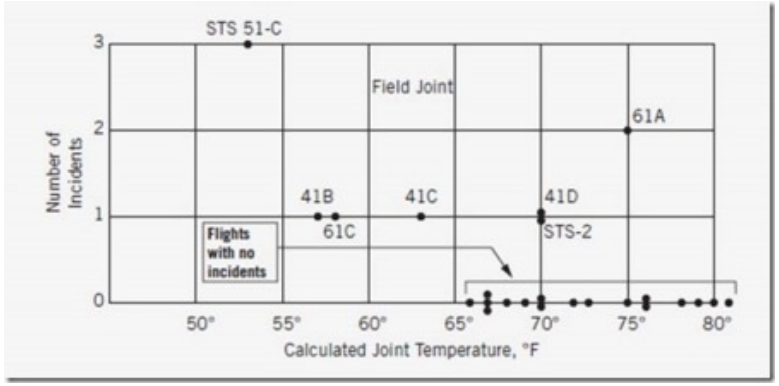
- Überarbeiten und editieren



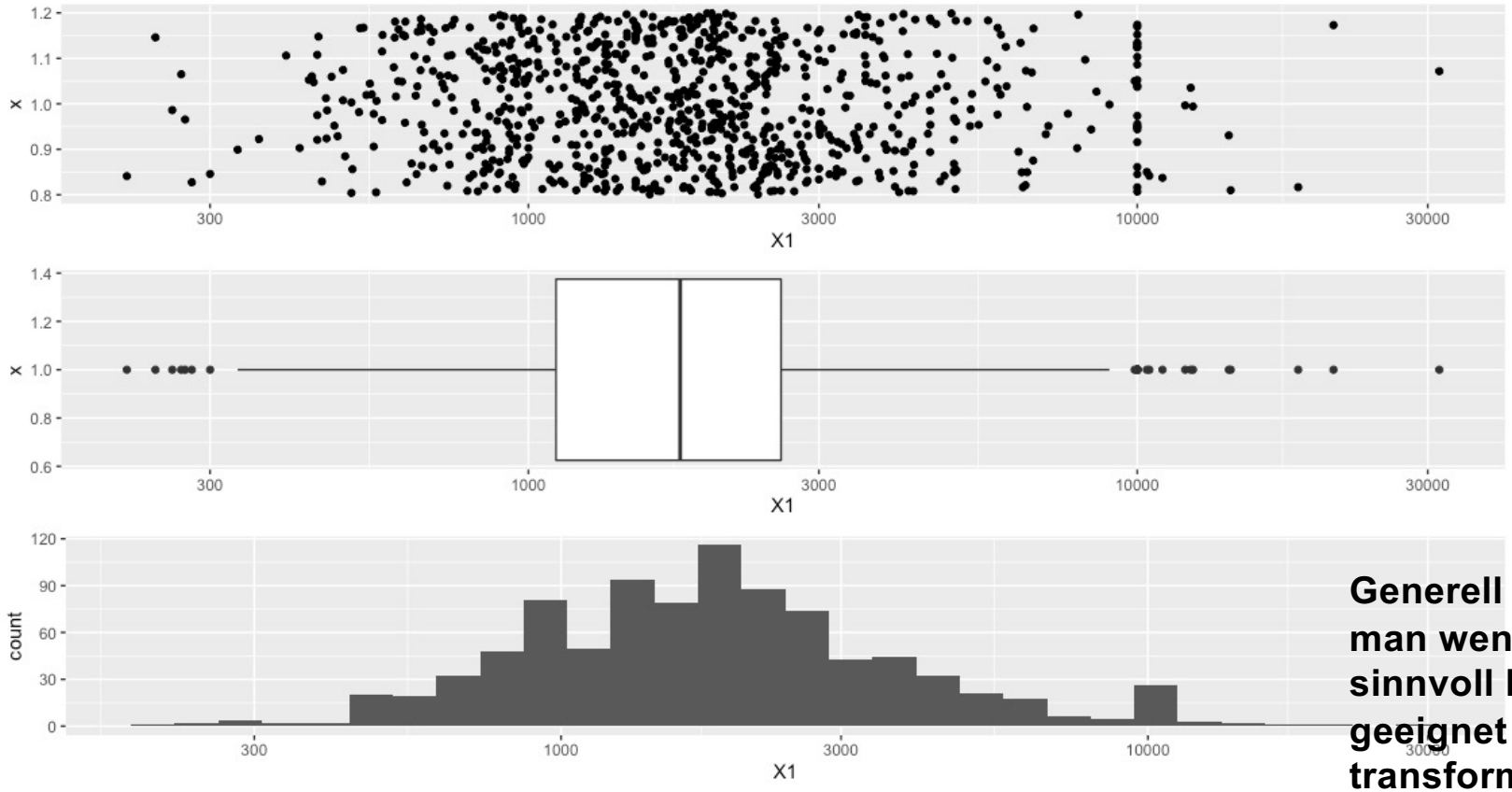
Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Wo möglich müssen alle Daten gezeigt werden!





Achse logarithmieren – viel besser!



Generell solle man wenn sinnvoll Daten geeignet transformieren

Quelle: Spiegelhalter, "The Art of Statistics"

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Agenda

1. Einführung
2. Beschreibende Statistik
3. Datenvisualisierung eindimensionaler Daten
4. Datentransformationen
5. Datenvisualisierung mehrdimensionaler Daten

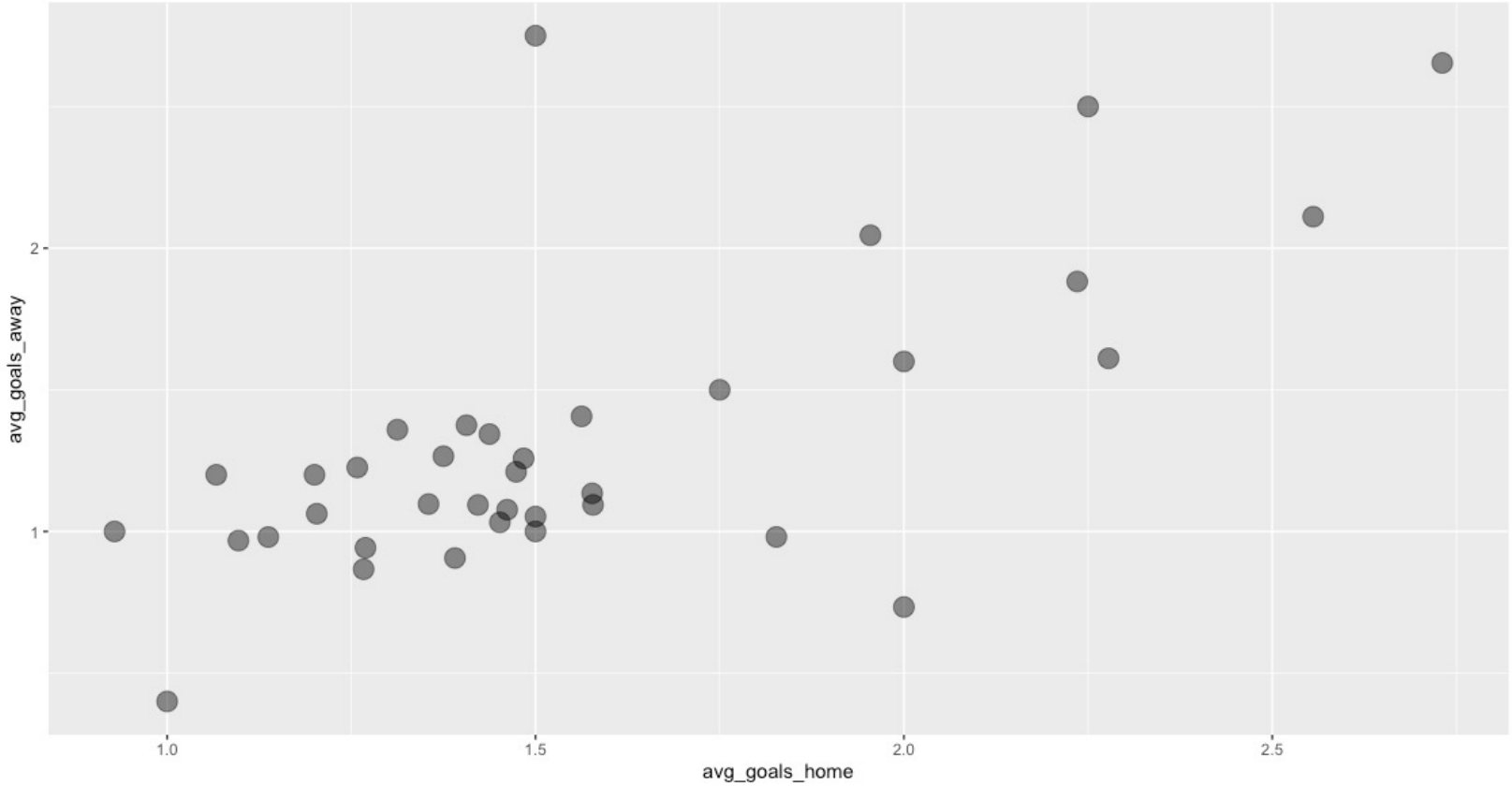
Nützliche Diagramme für zweidimensionale Daten

- **Punktwolke**
- Linienplot
- Paarplots
- Heatmaps

```
read_csv("results.csv") -> results
results %>%
 filter(tournament=="UEFA Euro" |
      tournament == "FIFA World Cup") %>%
 mutate(year = lubridate::year(date)) %>%
 na.omit() %>%
 select(year, home_score,away_score) %>%
 group_by(year) %>%
 summarise(avg_goals_home = mean(home_score),
        avg_goals_away = mean(away_score)) %>%
 ggplot(aes(x=avg_goals_home, y=avg_goals_away)) +
 geom_point(size = 5, alpha=0.5)
```

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Es könnte einen Zusammenhang geben

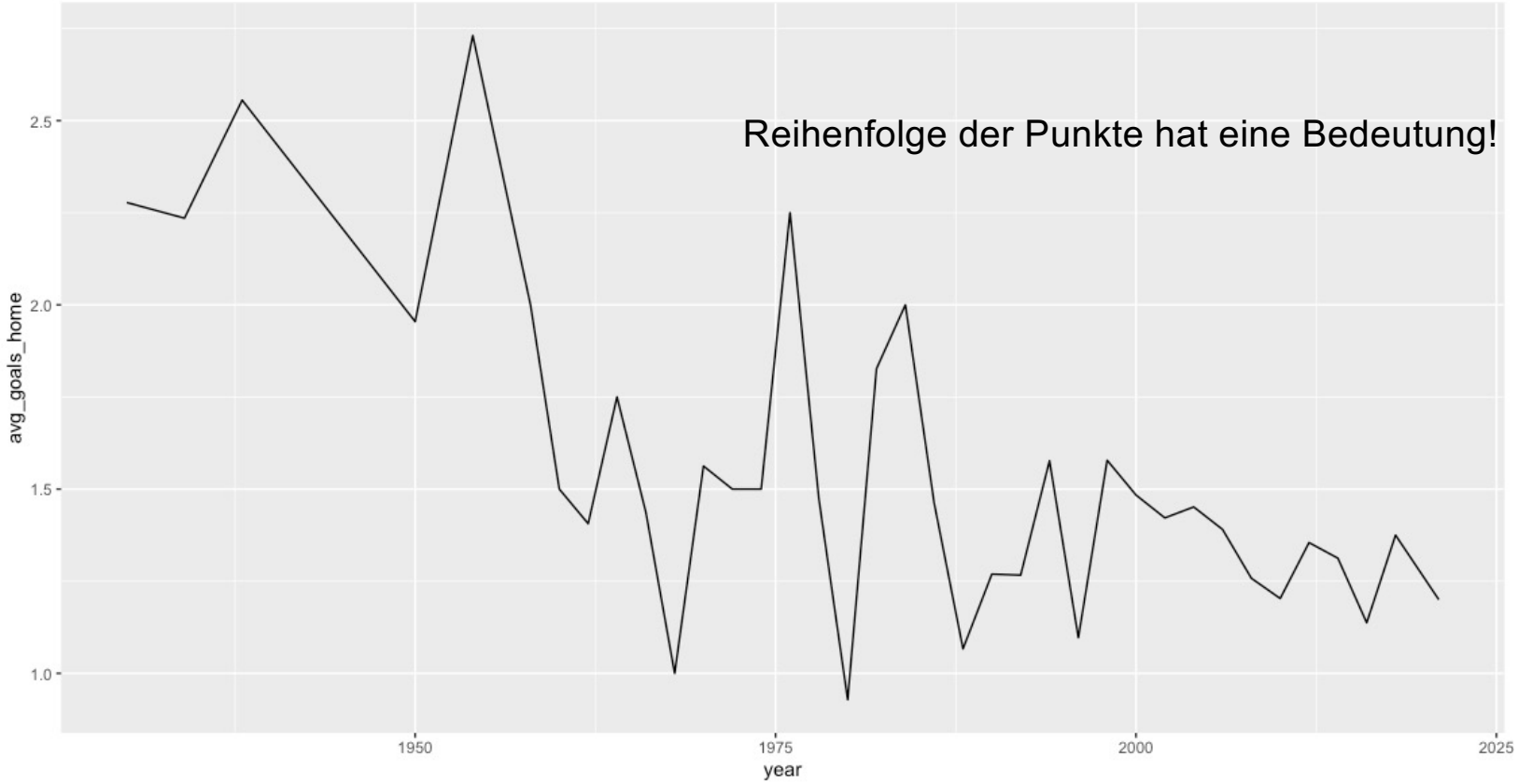


Nützliche Diagramme für zweidimensionale Daten

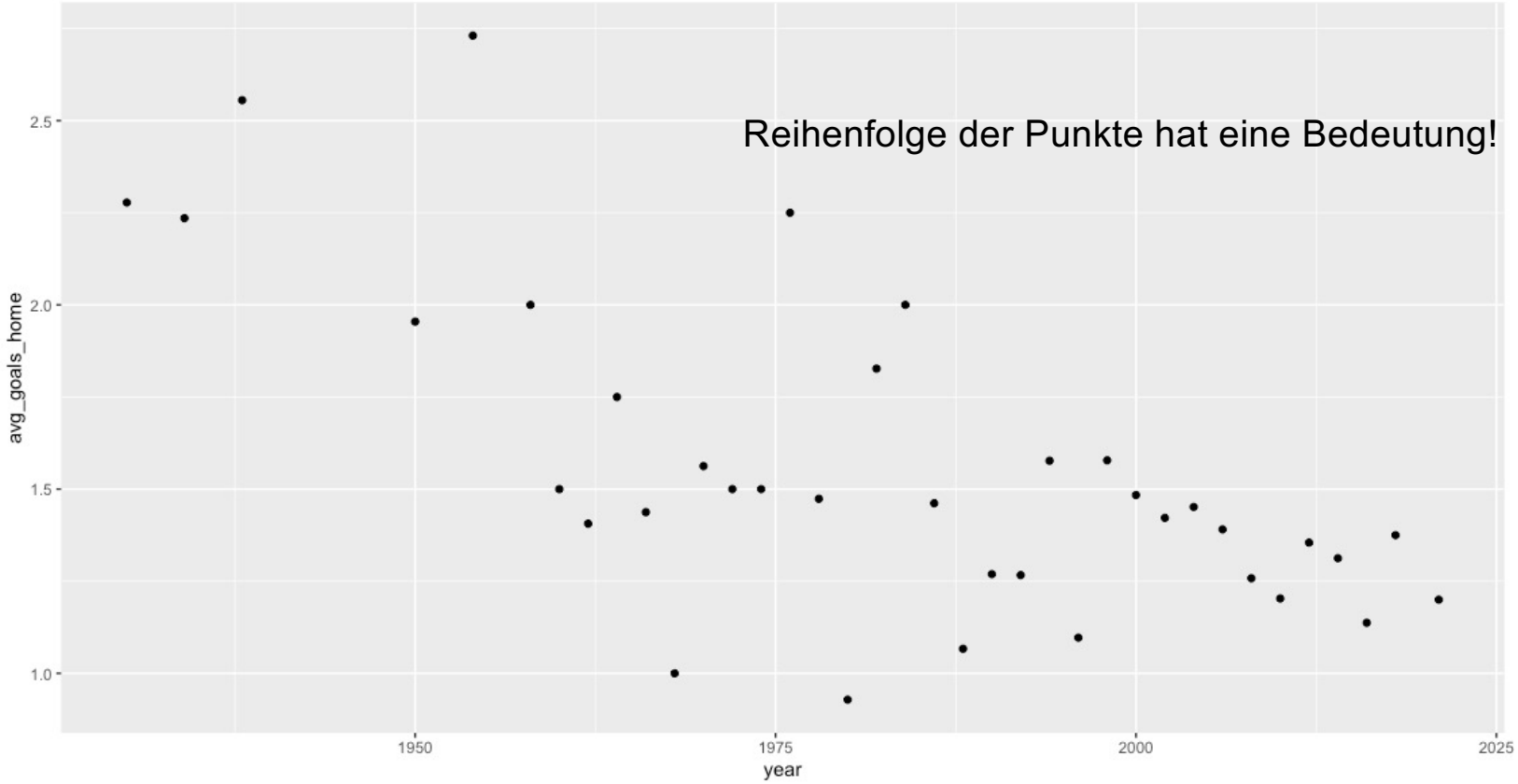
- Punktwolke
- **Linienplot**
- Paarplots
- Heatmaps

```
read_csv("results.csv") -> results
results %>%
 filter(tournament=="UEFA Euro" |
        tournament == "FIFA World Cup") %>%
 mutate(year = lubridate::year(date)) %>%
 na.omit() %>%
 select(year, home_score) %>%
 group_by(year) %>%
 summarise(avg_goals_home = mean(home_score)) %>%
 ggplot(aes(x=year, y=avg_goals_home)) +
 geom_line()
```

Ist Fußball über die Jahre defensiver geworden?



Diese Variante funktioniert viel schlechter obwohl prinzipiell das Gleiche zu sehen ist



Nützliche Diagramme für hochdimensionale Daten

- Punktwolke
- Linienplot
- **Paarplots**
- Heatmaps

Boston ist eine Datenbank mit Informationen über Gebiete rund um die Stadt Boston und die mittleren Hauspreise. Wir werden die lineare Regression verwenden, um die Hauspreise vorherzusagen.

crim Pro-Kopf-Verbrechensrate nach Stadt.

zn Anteil der Wohnbauflächen, die für Grundstücke über 25.000 sq.ft.

indus Anteil der Nicht-Einzelhandelsgeschäftsflächen pro Stadt.

chas Charles River Dummy-Variable (= 1, wenn Trakt an Fluss grenzt; 0 sonst).

nox Konzentration von Stickoxiden (Teile pro 10 Millionen).

rm durchschnittliche Anzahl der Zimmer pro Wohnung.

age Anteil der Eigentumswohnungen, die vor 1940 gebaut wurden.

dis gewichteter Mittelwert der Entfernungen zu fünf Bostoner Beschäftigungszentren.

rad Index der Erreichbarkeit von Radialautobahnen.

tax Vollwertiger Grundsteuersatz pro 10.000 $.

ptratio Schüler-Lehrer-Verhältnis nach Stadt.

black 1000(Bk-0.63)21000(Bk-0.63)2 wobei BkBk der Anteil der Schwarzen nach Stadt ist.

lstat unterer Status der Bevölkerung (Prozent).

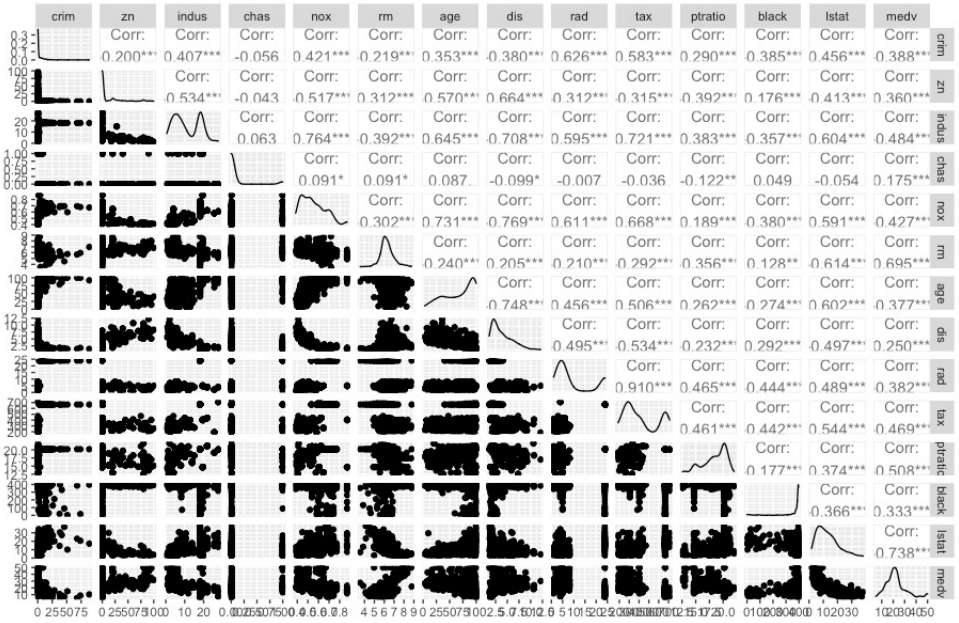
medv medianer Wert der Eigenheime in 1000er Dollar.

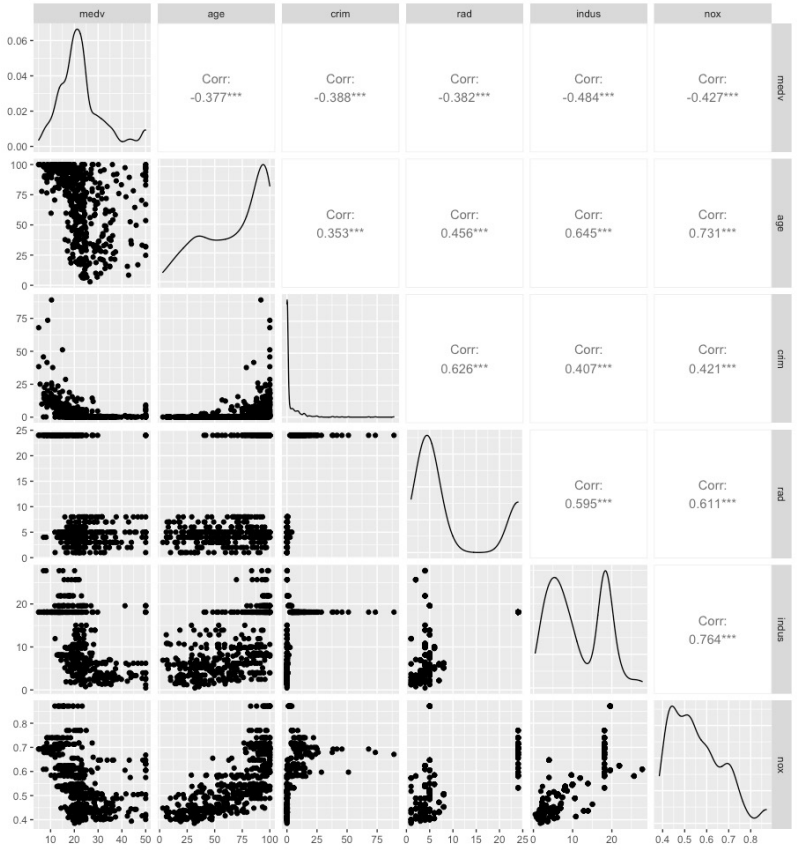
Was treibt den Wert?

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

ggpairs(Boston)

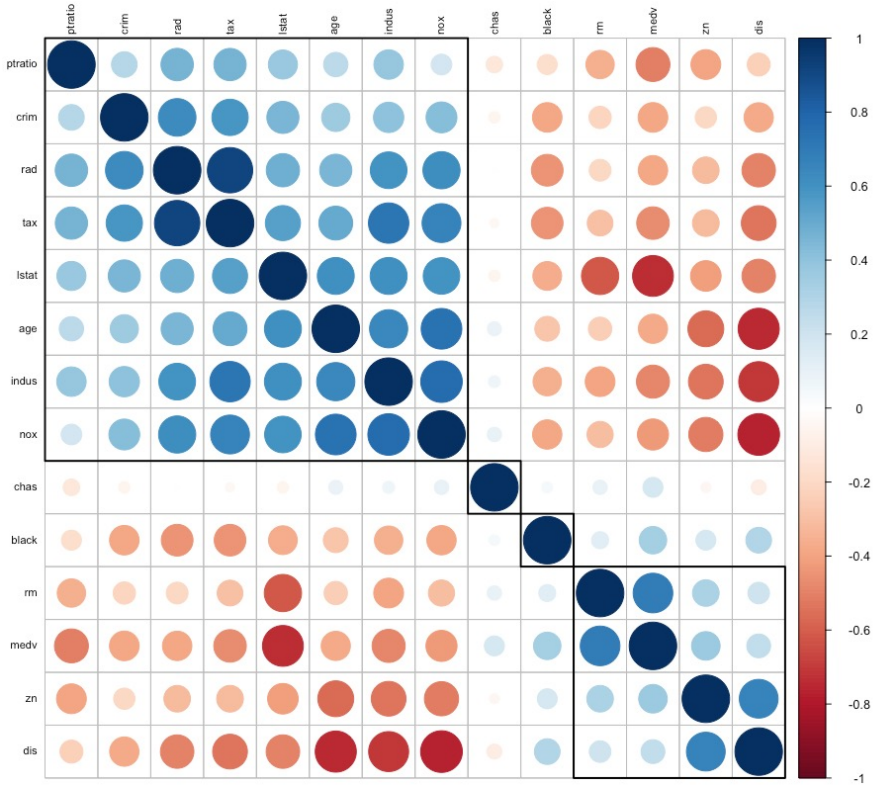
Information overflow?





Heatmaps können die Nadel im Heuhaufen zeigen

- Punktwolke
- Linienplot
- Paarplots
- **Heatmaps**




```
corrplot(cor(Boston), tl.col = "black", order = "hclust",
hclust.method = "average", addrect = 4, tl.cex = 0.7)
```