DHBW Lörrach

Business Intelligence

Vert. Prof. Dr. Aikaterini Nakou

Datenvisualisierung

Bei der Datenvisualisierung geht es darum:

- Informationen in Form von Zahlen und Daten mit Hilfe von grafischen Mitteln aufzubereiten
- um rasch und mit wenig Vorkenntnissen Muster, Trends, Beziehungen und Ausreißer erkennbar zu machen.

Die gängigsten grafischen Hilfsmittel für Datenvisualisierung sind u.a. **Diagramme, Graphen, Karten, Tabellen**, **Infografiken** und **Dashboards**.

DHBW Lörrach

Warum Datenvisualisierung

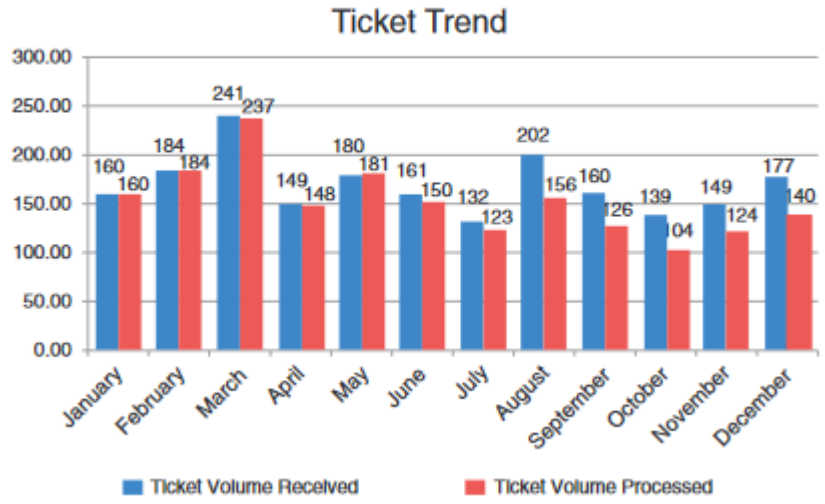
- Menschen sind in der Lage, extrem schnell aus komplexen visuellen Szenen wichtige Informationen herauszulesen
- Im Gegensatz zu verbalen Informationen, verarbeitet der Gehirn visuelle Informationen parallel
- Der Mensch erlernt visuelle Informationen schneller und erinnert sich an diese besser als verbale.

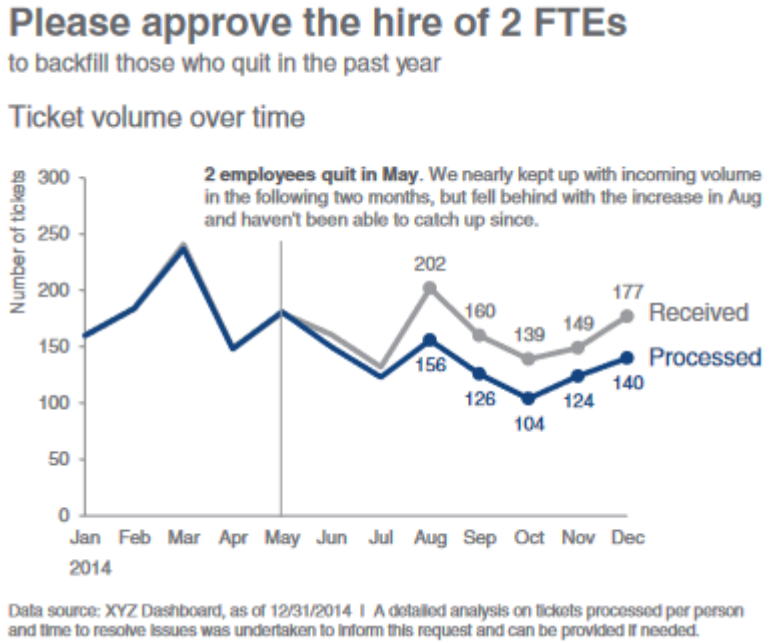
Arten der Datenvisualisierung

- Visualisierungen, die die Exploration der Daten bzw. die Verifikation der Datenanalyse dienen

- Visualisierungen, die die Präsentationen der Daten dienen → „Storytelling“ mit Daten

Schlechte vs Gute Grafiken





Prinzipien der Datenvisualisierung

- Den Kontext verstehen
- Auswahl der geeignete visuelle Darstellung
- Beseitigung der Unordnung
- Richten der Aufmerksamkeit dorthin, wo es gewollt ist
- Richtige Designing
- Storytelling

Kontext verstehen

Frage1: Welche Geschichte möchten Sie erzählen?

Frage 2: Was ist Ihre Zielgruppe?

Frage 3: Möchten Sie bestimmte Trends analysieren?

Frage 4: Soll die Zusammensetzung von Daten präsentiert werden?

Frage 5: Sollen Daten miteinander verglichen werden?

Frage 6: Ist ein bestimmter Zeitrahmen zu betrachten?

Auswahl der geeignete visuelle Darstellung

Simple text

Scatterplot

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Table

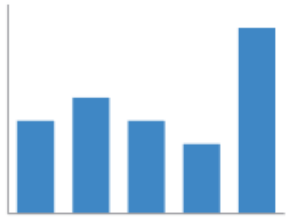
Line

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Heatmap

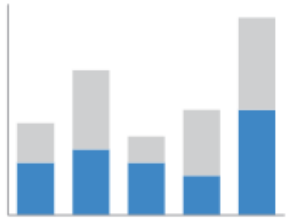
Slopegraph

FIGURE 2.1 The visuals I use most

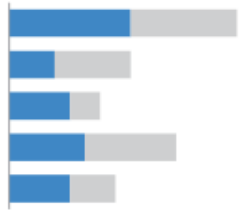


Vertical bar

Horizontal bar

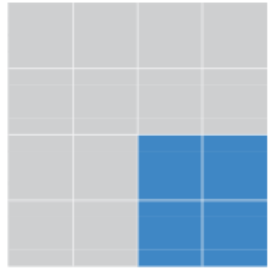


Stacked vertical bar

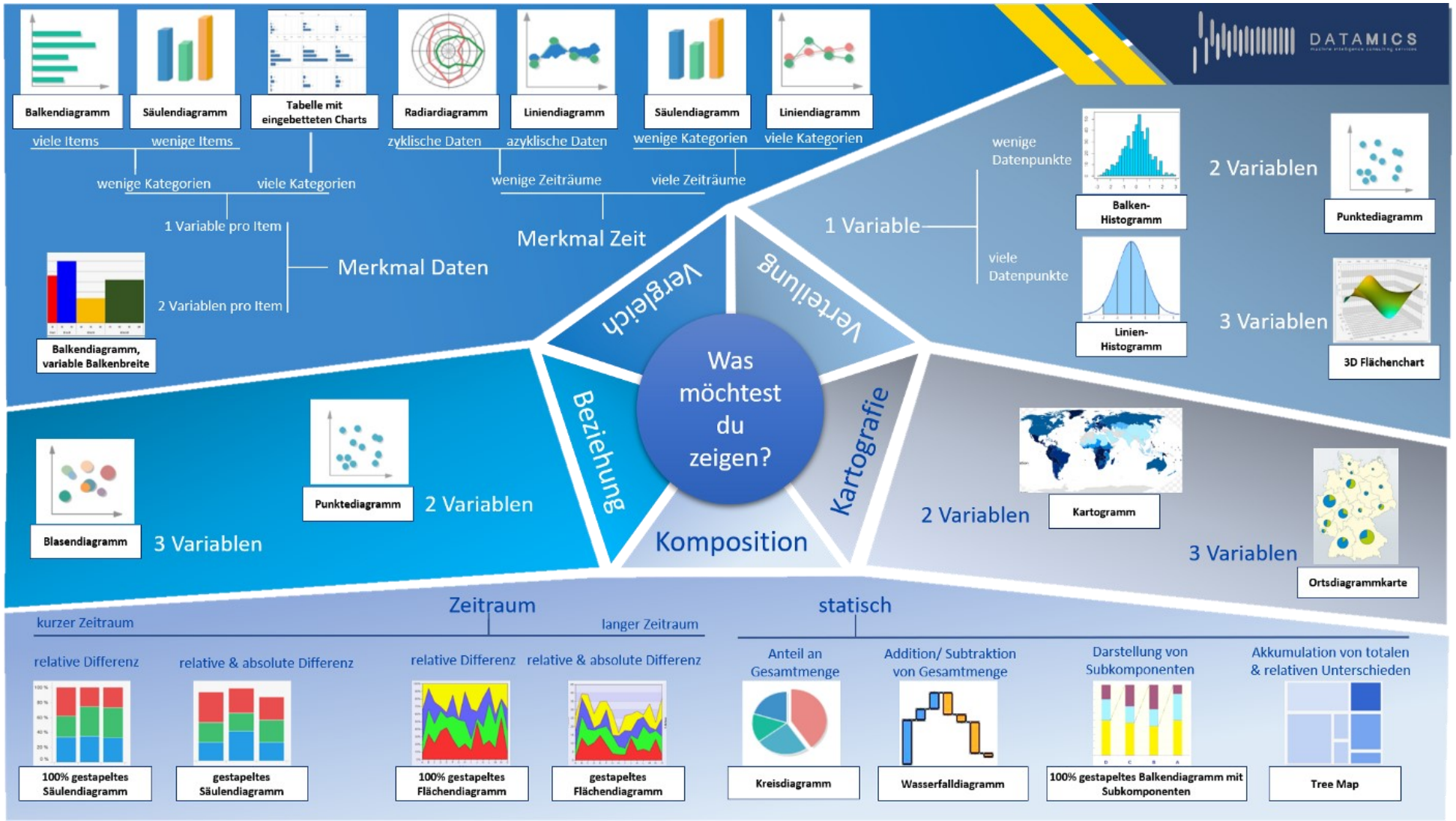


Stacked horizontal bar

Waterfall



Square area


DATAMICS
Was möchtest du zeigen?
Merkmal Daten
Merkmal Zeit
Vergleich
Verteilung
Beziehung
Kartografie
Komposition
Balkendiagramm
Säulendiagramm
Tabelle mit eingebetteten Charts
Radiardiagramm
Liniendiagramm
Säulendiagramm
Liniendiagramm
viele Items
wenige Items
wenige Kategorien
viele Kategorien
zyklische Daten
azyklische Daten
wenige Kategorien
viele Kategorien
1 Variable pro Item
2 Variablen pro Item
wenige Zeiträume
viele Zeiträume
Balkendiagramm, variable Balkenbreite
1 Variable
wenige Datenpunkte
viele Datenpunkte
Balken-Histogramm
Linien-Histogramm
2 Variablen
Punktediagramm
3 Variablen
3D Flächenchart
Blasendiagramm
3 Variablen
Punktediagramm
2 Variablen
2 Variablen
Kartogramm
3 Variablen
Ortsdiagrammkarte
Zeitraum
kurzer Zeitraum
langer Zeitraum
relative Differenz
relative & absolute Differenz
relative Differenz
relative & absolute Differenz
100% gestapeltes Säulendiagramm
gestapeltes Säulendiagramm
100% gestapeltes Flächendiagramm
gestapeltes Flächendiagramm
statisch
Anteil an Gesamtmenge
Addition/ Subtraktion von Gesamtmenge
Darstellung von Subkomponenten
Akkumulation von totalen & relativen Unterschieden
Kreisdiagramm
Wasserfalldiagramm
100% gestapeltes Balkendiagramm mit Subkomponenten
Tree Map

Beseitigung der Unordnung

- **Kognitive Belastung** ist die geistige Anstrengung, die erforderlich ist, um neue Informationen zu lernen.

- Als Designer von Informationen müssen wir wie wir die **Gehirnleistung** unseres Publikums nutzen und **nicht ausnutzen!**

Beseitigung der Unordnung

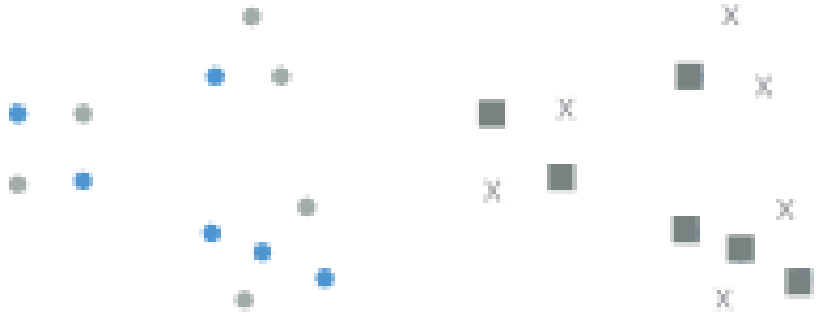
Gestaltprinzipien der visuellen Wahrnehmung:

- Nähe
- Ähnlichkeit
- Einschluss
- Abgeschlossenheit
- Kontinuität
- Verbindung

Beseitigung der Unordnung

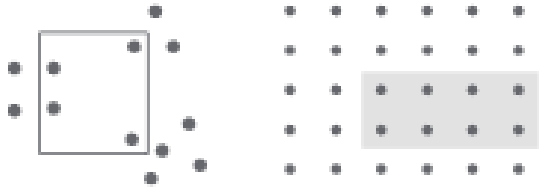
Nähe

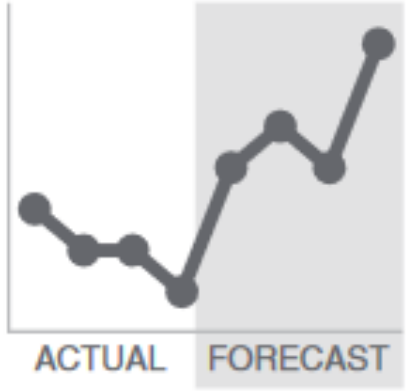
Ähnlichkeit



Beseitigung der Unordnung

Einschluss

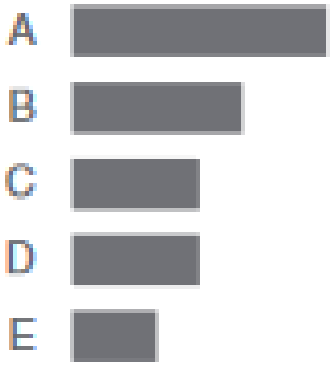




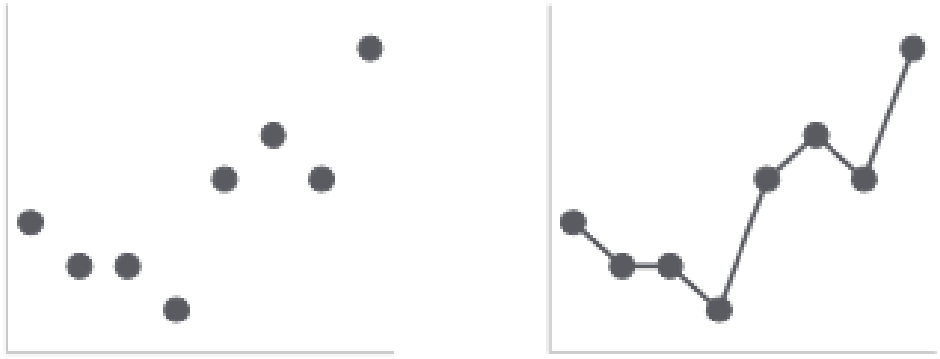
Abgeschlossenheit

Beseitigung der Unordnung

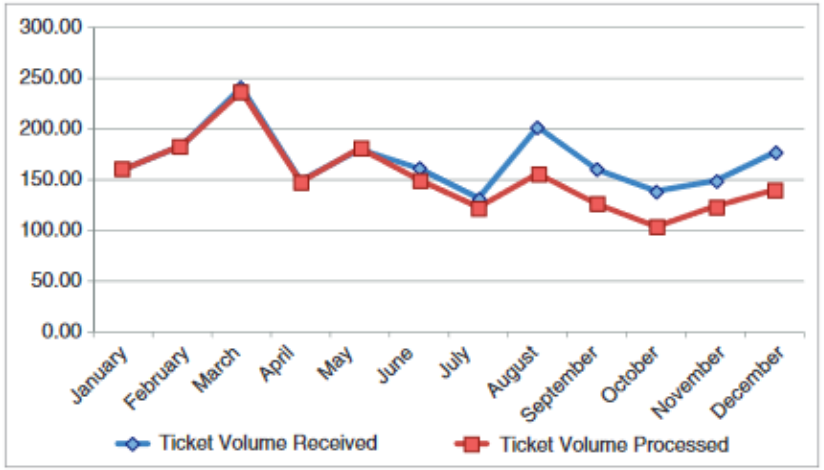
Kontinuität

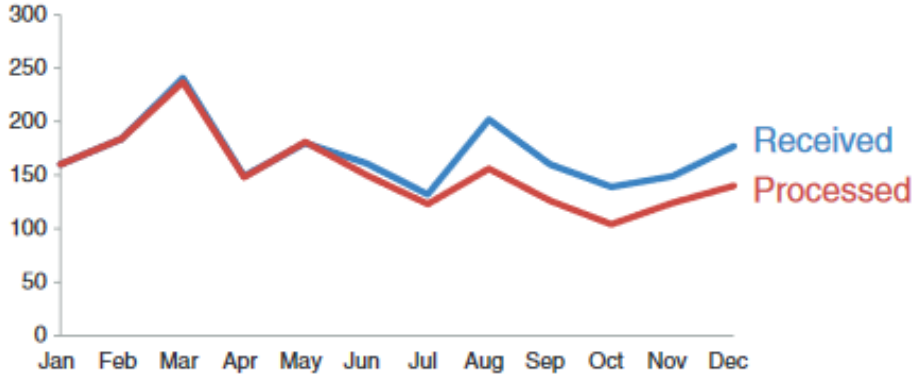


Verbindung



Beseitigung der Unordnung





Fokussierung der Aufmerksamkeit

Sehen und Gedächtnis → präattentive Attribute → Lenkung der Aufmerksamkeit des Publikums, wo wir sie haben möchten!

Präattentive Attribute

- Farbe
- Größe
- Position

Fokussierung der Aufmerksamkeit

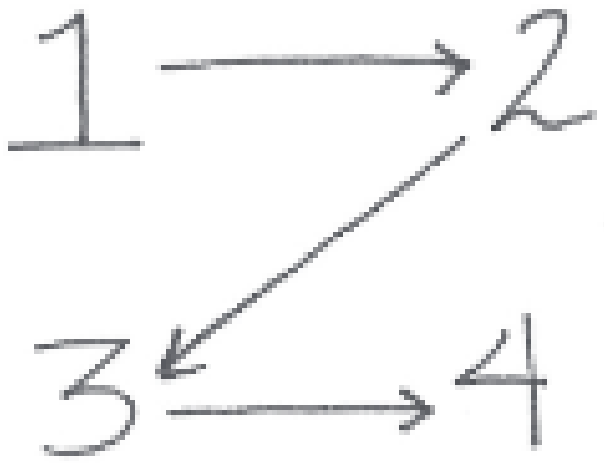
Präattentives Attribute: Farbe

Fokussierung der Aufmerksamkeit

Präattentives Attribute:Farbe

Fokussierung der Aufmerksamkeit

Präattentives Attribute: Position



Richtige Designe

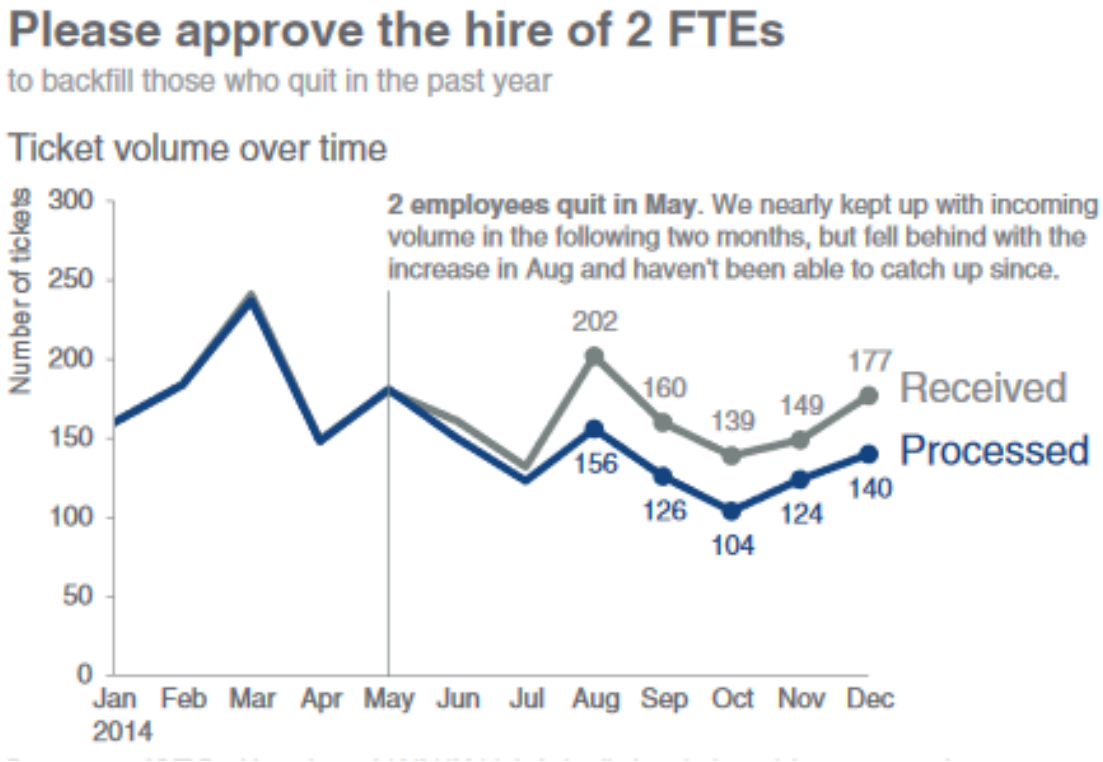
Folge das Sprichwort der Produktdesign:

Die Form folgt der Funktion

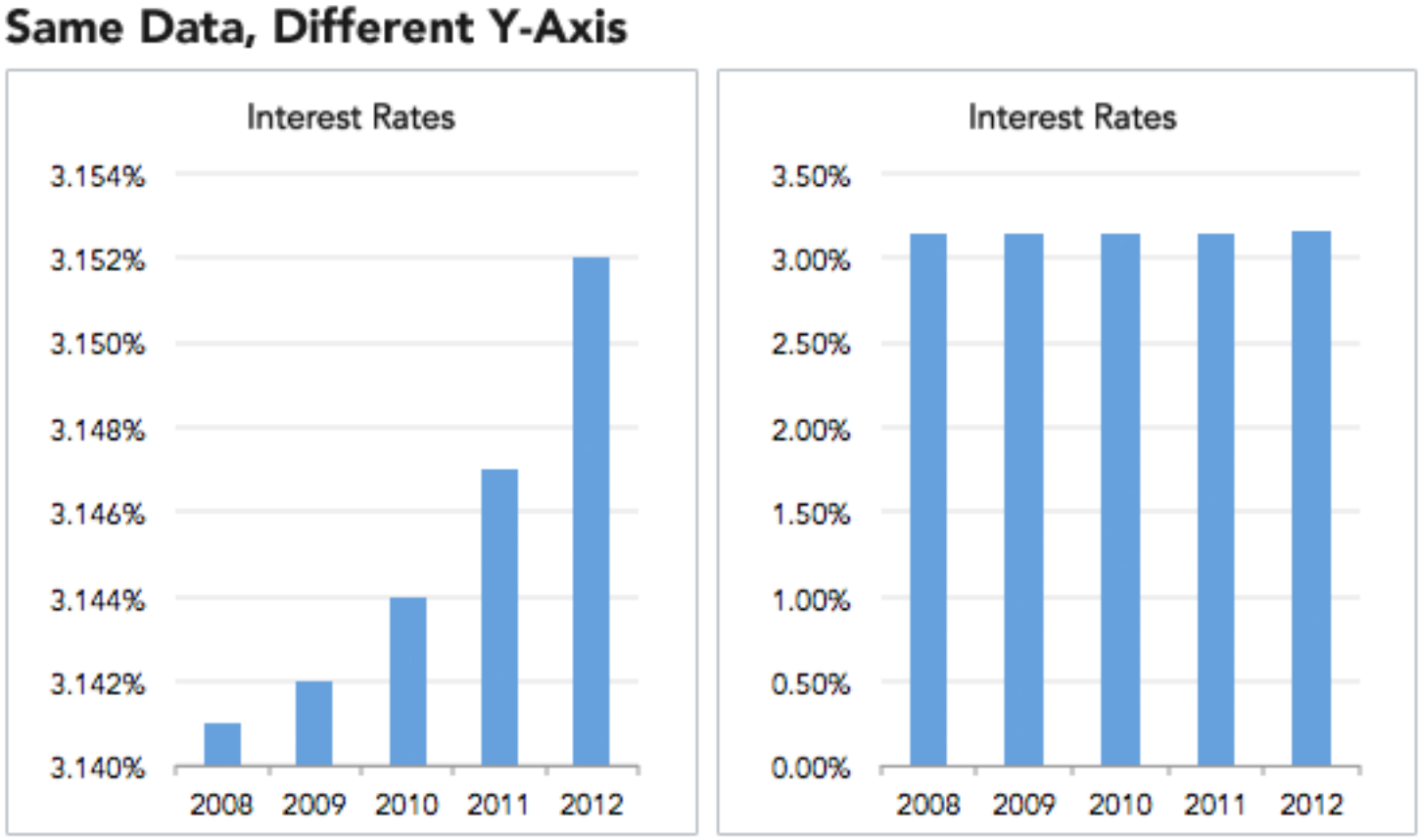
Also überlegen was unser Publikum mit den Daten machen soll (Funktion)
und dann eine Visualisierung (Form) erstellen!

Richtige Designe

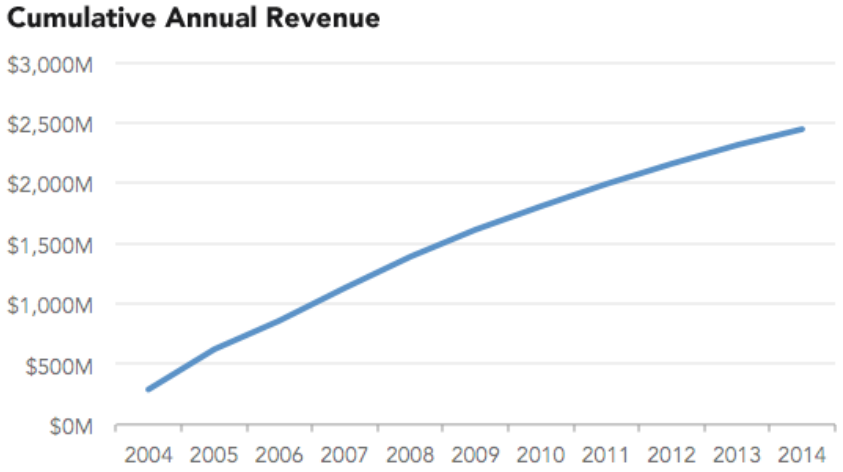
Zugänglichkeit der Informationen!

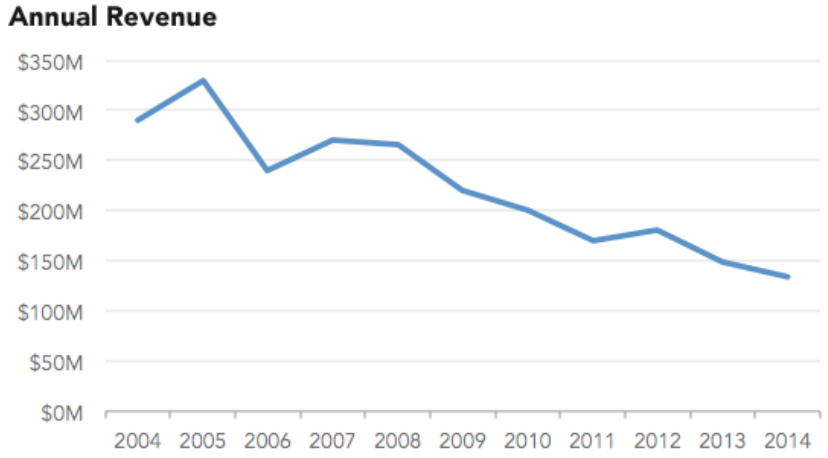


Manipulation durch Datenvisualisierung



Manipulation durch Datenvisualisierung

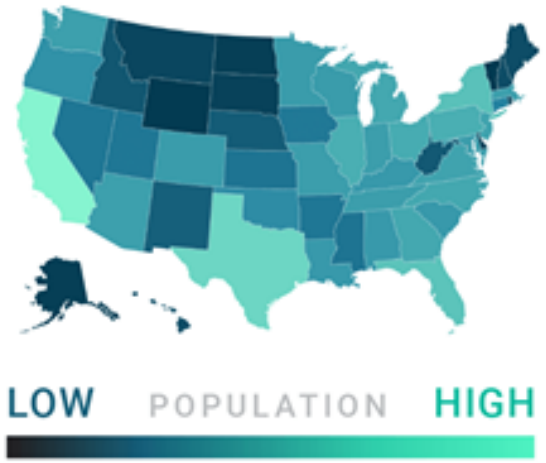




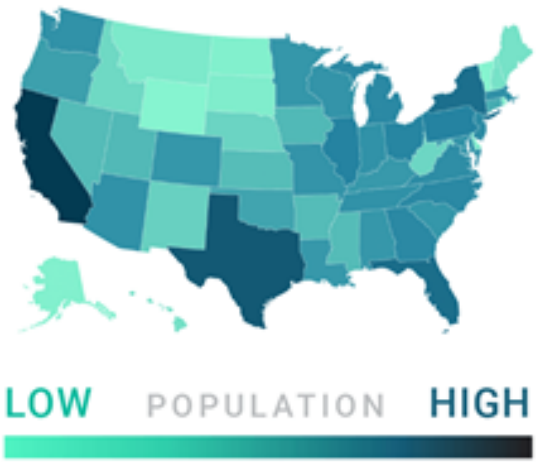
DHBW Lörrach

Manipulation durch Datenvisualisierung

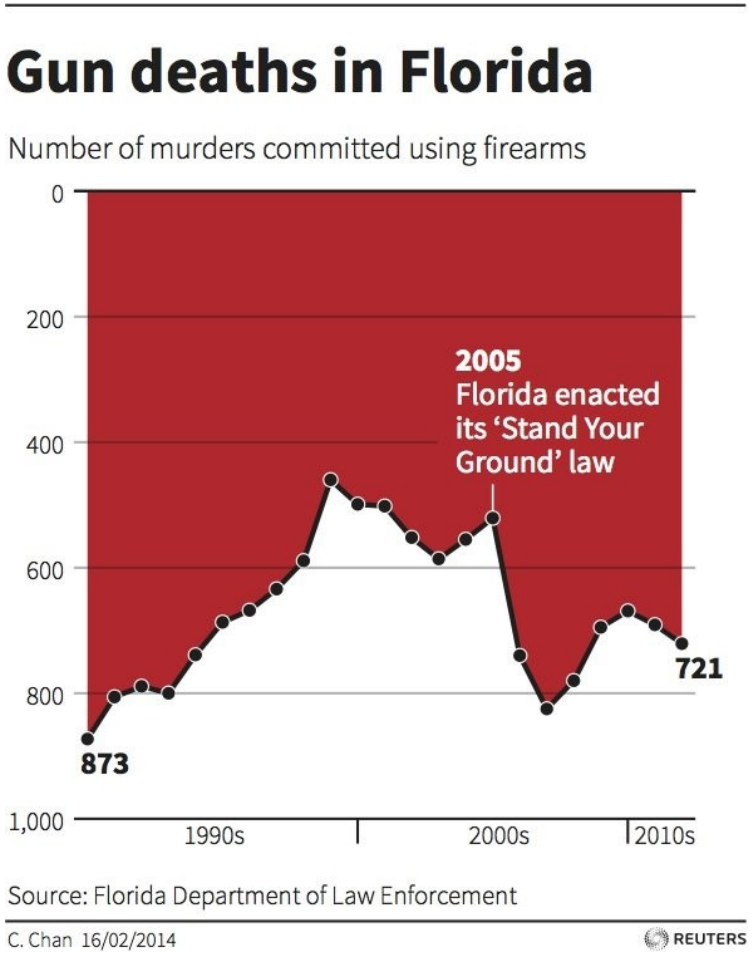
NOT IDEAL



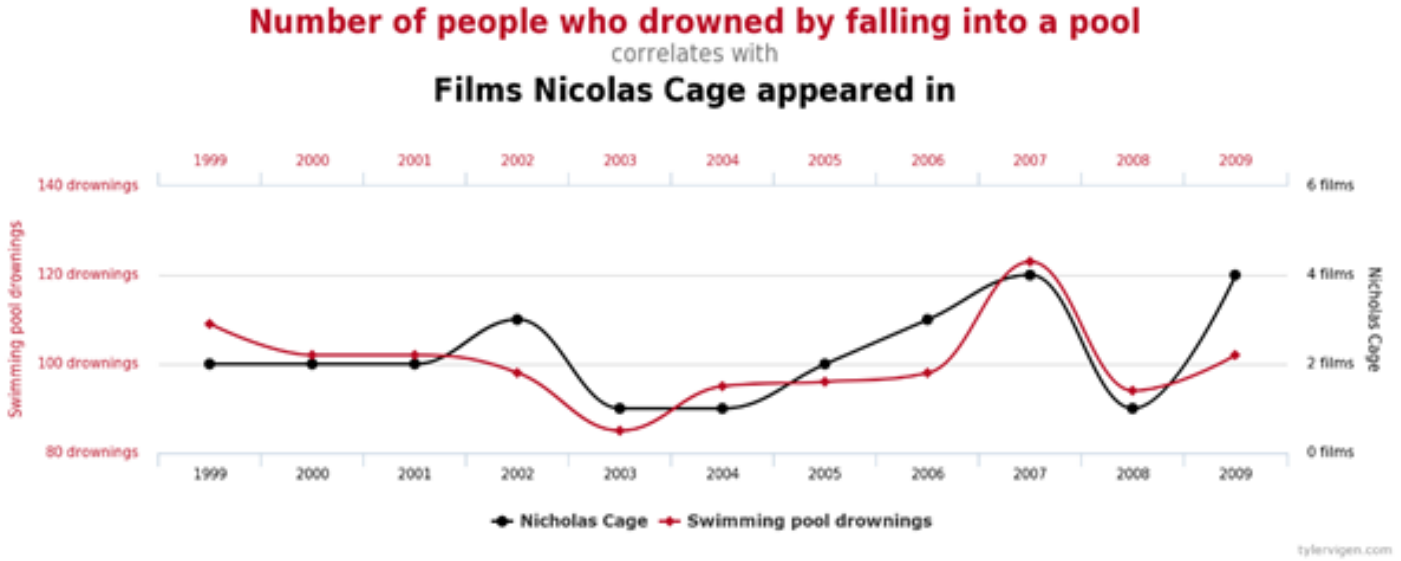
BETTER



Manipulation durch Datenvisualisierung



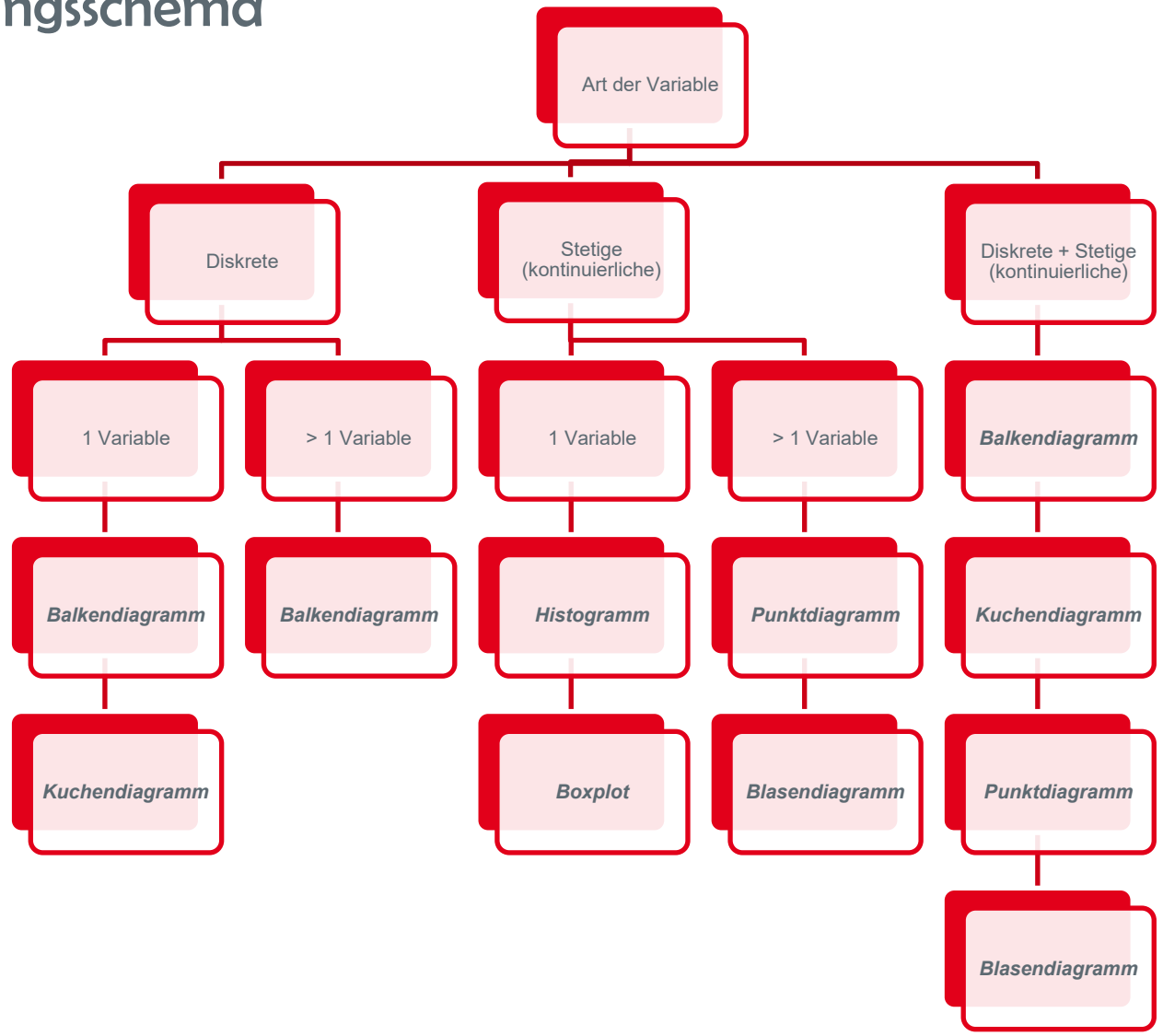
Manipulation durch Datenvisualisierung



DHBW Lörrach

Pipeline der Datenvisualisierung

Datenvisualisierungsschema



Plotly

Warum Plotly

- Interaktive ope-source Framework für Datenvisualisierung
- kann verschiedene Arten von Diagrammen und Grafiken leicht (low-code) erstellen
- optisch tolles Design
- unendliche Anpassungsmöglichkeiten der Diagramme
- Großes Community

Infos und Beispiele

R: *https://plotly.com/r*

DHBW Lörrach

Syntax für alle Plotly Grafiken

```
plot_ly(data = <Daten>,
        x = ~<x-Achsenvariable>,
        y = ~<y-Achsenvariable>,
        z = ~<z-Achsenvariable>,
        type = '<Diagrammtyp>',
        mode = '<Modus>',
        color = ~<Farbvariable>,
        colors = <Farbschema>,
        symbol = ~<Symbolvariable>,
        size = <Punktgröße>,
        text = ~<Tooltiptext>,
        hoverinfo = '<Hover-Info>',
        marker = list(<Marker-Einstellungen>),
        line = list(<Linien-Einstellungen>),
        name = '<Legendeneintrag>',
        opacity = <Transparenz>,
        showlegend = <TRUE/FALSE>)
```

DHBW Lörrach

Syntax für alle Plotly Grafiken

- `data`: Das Datenset, das zur Erstellung des Diagramms verwendet wird.
- `x` und `y`: Variablen für die x- und y-Achsen. Sie sind in fast allen Diagrammtypen erforderlich.
- `z`: Wird für 3D-Diagramme benötigt, um die z-Achsenwerte anzugeben.
- `type`: Bestimmt den Typ des Diagramms. Einige gängige Typen sind:
  - `'scatter'` (Streudiagramm oder Linienplot)
  - `'bar'` (Balkendiagramm)
  - `'box'` (Boxplot)
  - `'histogram'` (Histogramm)
  - `'scatter3d'` (3D-Streudiagramm)
  - `'surface'` (3D-Oberflächendiagramm)
  - `'pie'` (Tortendiagramm)
  - `'heatmap'` (Heatmap)
  - `'line'` (Linienplot)
  - `'area'` (Flächendiagramm)
  - `'bubble'` (Blasendiagramm)

DHBW Lörrach

Syntax für alle Plotly Grafiken

- `mode` : Definiert die Darstellungsweise bei Streudiagrammen:
  - `'markers'` für Punkte
  - `'lines'` für Linien
  - `'lines+markers'` für eine Kombination aus beidem
- `color` : Variable, die verwendet wird, um die Farbe der Datenpunkte oder Balken zu bestimmen.
- `colors` : Gibt das Farbschema an, das für das Diagramm verwendet wird, wie z. B. `'Viridis'` oder `'Blues'` .
- `symbol` : Variable, die zur Festlegung der Symbolform in Streudiagrammen verwendet wird.
- `size` : Bestimmt die Größe der Marker in Streudiagrammen.
- `text` : Text, der angezeigt wird, wenn der Benutzer über einen Punkt fährt (Tooltip).
- `hoverinfo` : Steuerung, welche Informationen im Tooltip angezeigt werden, z. B. `'x+y+text'` .

DHBW Lörrach

Syntax für alle Plotly Grafiken

- `marker` : Liste zur Anpassung der Marker (Datenpunkte), wie:
  - `color` : Farbe der Marker
  - `size` : Größe der Marker
  - `line` : Randfarbe und -breite der Marker
- `line` : Liste zur Anpassung der Linien, wie:
  - `color` : Farbe der Linie
  - `width` : Breite der Linie
  - `dash` : Stil der Linie (z. B. `'solid'` , `'dash'` , `'dot'`)
- `name` : Text, der als Bezeichnung in der Legende des Diagramms angezeigt wird.
- `opacity` : Wert zwischen 0 und 1, der die Transparenz des Diagramms steuert (1 = vollständig sichtbar, 0 = vollständig transparent).
- `showlegend` : Steuert, ob die Legende im Diagramm angezeigt wird (`TRUE` oder `FALSE`).

Syntax für alle Plotly Grafiken

DHBW Lörrach

```
layout(
  title = '<Diagrammtitel>',
  xaxis = list(
    title = '<x-Achsentitel>',
    showgrid = <TRUE/FALSE>,
    zeroline = <TRUE/FALSE>,
    showline = <TRUE/FALSE>
  ),
  yaxis = list(
    title = '<y-Achsentitel>',
    showgrid = <TRUE/FALSE>,
    zeroline = <TRUE/FALSE>,
    showline = <TRUE/FALSE>
  ),
  legend = list(
    title = list(text = '<Legendentitel>'),
    x = <x-Position>,
    y = <y-Position>,
    orientation = '<horizontal/vertical>'
  ),
  plot_bgcolor = '<Diagrammhintergrundfarbe>',
  paper_bgcolor = '<Hintergrundfarbe>',
  showlegend = <TRUE/FALSE>,
  barmode = '<stack/group>'
)
```

Syntax für alle Plotly Grafiken

DHBW Lörrach

- `title`: Legt den Titel des Diagramms fest. Sie können auch zusätzliche Formatierungen wie Schriftgröße und Farbe definieren.
- `xaxis` und `yaxis`: Listen von Parametern zur Anpassung der Achsen.
  - `title`: Titel der Achse.
  - `showgrid`: Steuerung, ob das Gitter angezeigt wird (`TRUE` oder `FALSE`).
  - `zeroline`: Anzeige der Nulllinie, die die Nullposition markiert.
  - `showline`: Anzeige einer Linie entlang der Achse.
- `legend`: Anpassung der Legende des Diagramms.
  - `title`: Titel der Legende.
  - `x` und `y`: Position der Legende im Diagramm (Koordinaten zwischen 0 und 1).
  - `orientation`: Ausrichtung der Legende (`horizontal` oder `vertical`).

Syntax für alle Plotly Grafiken

DHBW Lörrach

- `plot_bgcolor`: Hintergrundfarbe des Diagrammbereichs.
- `paper_bgcolor`: Hintergrundfarbe des gesamten Diagrammpapiers.
- `showlegend`: Steuerung, ob die Legende angezeigt wird (`TRUE` oder `FALSE`).
- `barmode`: Definiert die Art der Anordnung der Balken in einem Balkendiagramm.
  - Werte: `'stack'` für gestapelte Balken oder `'group'` für gruppierte Balken.

DHBW Lörrach

Statistische Tests

- Testen von Hypothesen (Annahmen, Behauptungen)
- Statistischer Test: Verfahren, mit dessen Hilfe sich bestimmte Hypothesen auf ihre Richtigkeit hin überprüfen lassen.
- Statistische Testverfahren basieren auf Stichprobentheorie
- **Ziel:** Richtigkeit von Aussagen über die Verteilung einer Zufallsvariablen überprüfen.

Statistische Tests: Hypothesen

… für Unterschiede

Nullhypothese (H_0):

- besagt, dass es keinen Mittelwertunterschied zwischen den Gruppen gibt

$$\mu_1 = \mu_2$$

Alternativhypothese (H_1):

- besagt, dass es einen Mittelwertunterschied zwischen den Gruppen gibt

$$\mu_1 \neq \mu_2$$

… für Zusammenhänge

Nullhypothese (H_0):

- besagt, dass es keinen Zusammenhang zwischen den Merkmalen gibt

$$\rho = 0$$

Alternativhypothese (H_1):

- besagt, dass es einen Zusammenhang zwischen den Merkmalen gibt

$$\rho \neq 0$$

Hypothesenbildung - Inferenzstatistik

Hypothese:
Unterscheidet sich die durchschnittliche Kompetenz von Schülern, die an dem Training teilnehmen von derjenigen von Schülern, die nicht an diesem Training teilnehmen?

Bei jedem Test sind wieder zwei Hypothesen zu formulieren:

Nullhypothese H_0	Alternativhypothese H_1
Der vermutete Unterschied besteht nicht: $\mu_1 = \mu_2$ bzw. $\mu_1 - \mu_2 = 0$	Der vermutete Unterschied besteht : $\mu_1 \neq \mu_2$ bzw. $\mu_1 - \mu_2 \neq 0$

μ_1 = Mittelwert der Schülerpopulation, die an dem Training teilgenommen hat
μ_2 = Mittelwert der Schülerpopulation, die nicht an dem Training teilgenommen hat

DHBW Lörrach

Signifikanztest

- Um entscheiden zu können, ob das Stichprobenergebnis mit der Nullhypothese übereinstimmt, muss eine Grenze für die beschriebene Wahrscheinlichkeit festgelegt werden.

- Dieser Wahrscheinlichkeitsgrenze wird fest vorgeben. Die **Wahrscheinlichkeitsgrenze** wird als **Irrtumswahrscheinlichkeit** α bezeichnet.

- Normalerweise Irrtumswahrscheinlichkeit $\alpha=0.05$

DHBW Lörrach

Schritte- Statistische Tests

1. Formulierung von Null- und Alternativhypothese
2. Bestimmung von Skalenniveau und Verteilungstyp der Merkmalswerte
3. Wahl der entsprechenden Signifikanztests
4. Festlegung der Irrtumswahrscheinlichkeit α, z.B. 0,05
5. Berechnung der Signifikanztests
6. Vergleichen mit der Irrtumswahrscheinlichkeit
7. Ablehnung oder Behalten der Null-Hypothese

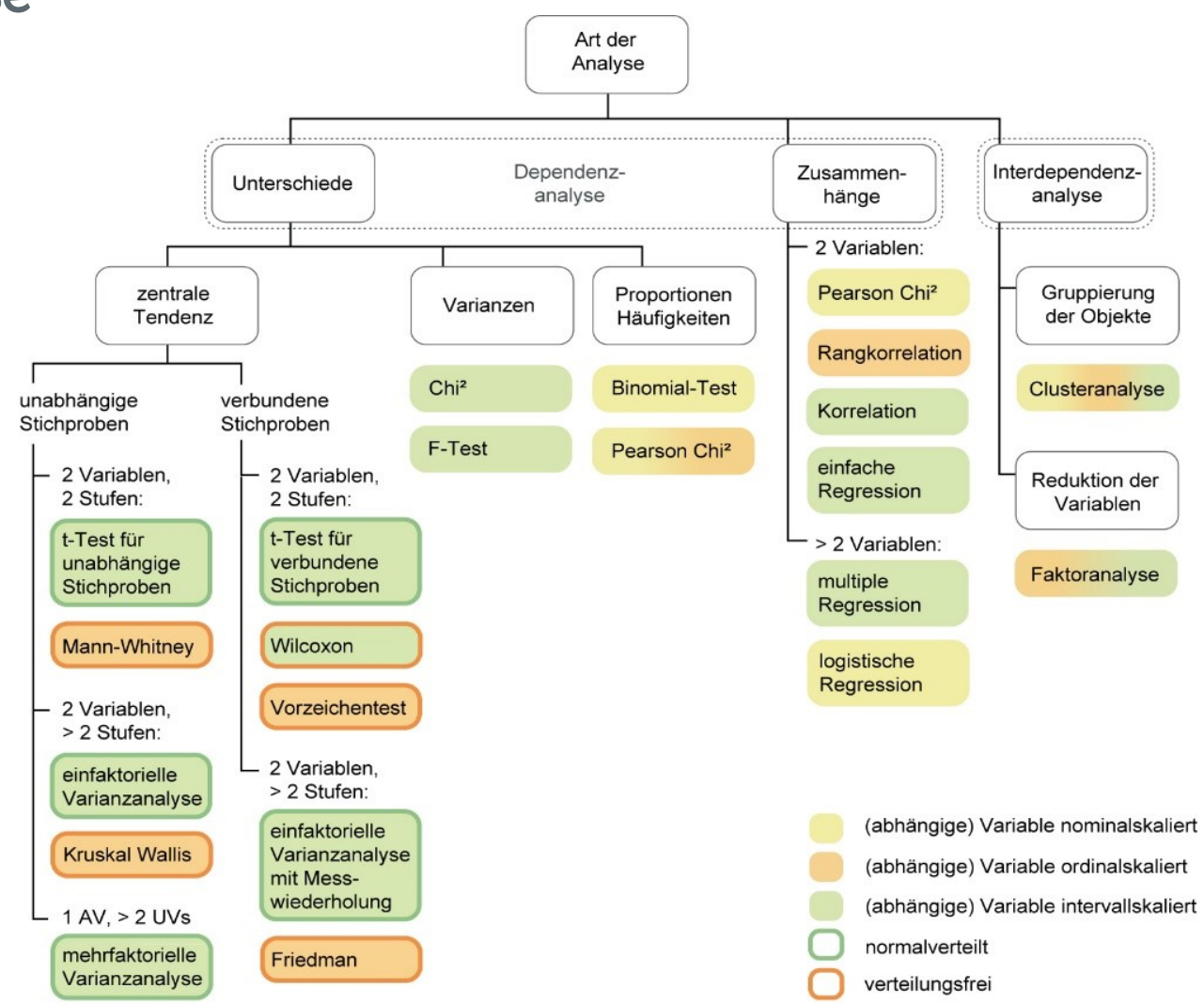
DHBW Lörrach

Signifikanztests

- Bei **parametrischen Signifikanztests** werden angewendet, wenn die die Daten metrisch skaliert sind, da sonst Operationen wie die Berechnung des arithmetischen Mittels oder der Standardabweichung nicht zulässig sind.

- **Nichtparametrische Tests** werden angewendet, wenn die Daten lediglich nominal oder ordinal skaliert sind, oder zwar metrische Daten vorliegen, aber sonstige Voraussetzungen der entsprechenden parametrischen Verfahren nicht erfüllt werden.

Statistische Analyse



DHBW Lörrach

Shapiro-Wilk-Test

Frage:

Stammt eine Stichprobe aus einer normalverteilten Population?

Hypothesen:

H0: Die Daten sind normalverteilt.

HA: Die Daten sind nicht normalverteilt

Beispiele:

Ist der Intelligenzquotient in einer Schulklasse normalverteilt?

Ist die Risikobereitschaft von jungen Männern normalverteilt?

Folgen die Mathematiknoten einer Lehrabschlussprüfung einer Normalverteilung?

DHBW Lörrach

t-Test: zwei unabhängige Stichproben

Frage:

Weisen die Grundgesamtheiten zweier Stichproben gleiche arithmetische Mittelwerte auf?

Hypothesen:

$$H_0: \mu_1 = \mu_2 \text{ mit } H_A: \mu_1 \neq \mu_2$$

Beispiele:

Unterscheiden sich Personen mit selbstständiger oder unselbständiger Tätigkeit bezüglich ihrer Zufriedenheit mit ihrer beruflichen Situation?

Gibt es einen Unterschied in der durchschnittlichen Anzahl Einbrüche in Häuser mit und ohne Alarmanlage?

t-Test: zwei abhängige Stichproben

Frage:

Unterscheiden sich die Mittelwerte zweier abhängiger Beobachtungsreihen signifikant?

Hypothesen:

H_0: $\mu_d = 0$ (oder H_0: $\mu_1 = \mu_2$ mit $\mu_d = \mu_1 - \mu_2$)

H_A: $\mu_d \neq 0$ (oder H_A: $\mu_1 \neq \mu_2$)

Beispiele:

Gibt es Unterschiede zwischen den Testwerten von Probanden vor und nach einem Gedächtnistraining?

Nichtparametrische Tests

Diese Verfahren kommen in Betracht, wenn:

- Stichproben nominaler oder ordinaler Daten vorliegen
- Stichproben metrischer Daten vorliegen, die die Voraussetzungen der parametrischen Tests nicht erfüllen (z. B. keine Normalverteilung oder nicht gleiche Varianzen)

DHBW Lörrach

MannWhitney-Test: Voraussetzungen

- Bedingungen für t-Test sind nicht gegeben, z. B. eine Stichprobe ist nicht normalverteilt oder die Intervallskala ist fraglich oder die Homogenität der Varianzen ist nicht gegeben

- Unabhängige Stichproben

- Metrische Skalenniveau

Beispiele:

Unterscheiden sich die Schulnoten von Mädchen und Jungen?

Zeigen sich Unterschiede in der Darlehenshöhe zwischen Kreditnehmern, die den Kredit zurückgezahlt haben, und solchen, die dies versäumt haben?

Wilcoxon-Test

- Bedingungen für t-Test (abhängige Stichprobe) sind nicht gegeben, z. B. eine Stichprobe ist nicht normalverteilt oder die Intervallskala ist fraglich oder die Homogenität der Varianzen ist nicht gegeben

- Abhängige Stichproben

- Metrische Skalenniveau

Beispiele:

Verkauft sich ein Produkt nach der Vergabe eines Rabatts bei Onlinekauf besser als zuvor?

Zeigen Alzheimer-Patienten auf einer Emotionsskala nach einem Besuch von Verwandten andere Werte als vor deren Besuch?

DHBW Lörrach

Weitere Nichtparametrische Tests

H-Test von Kruskal und Wallis:

Das Verfahren vergleicht mehrere unabhängige Stichproben. Es stellt eine Verallgemeinerung des MannWhitney-Tests dar. Angewendet wird der H-Test bei mehr als 2 Stichproben, die stetig, aber nicht normalverteilt sind.

Kolmogorov-Smirnow-Anpassungstest (K-S-Test):

Der Test dient zum Prüfen der Anpassung einer empirischen Verteilung an eine theoretische Verteilungsfunktion (nicht nur die Normalverteilung). Wie der X^2-Test ist er ein verteilungsunabhängiger bzw. parameterfreier Test.

Levene-Test:

Der Test wird dazu verwendet, die Homogenität der Varianzen zwischen zwei oder mehr Gruppen zu überprüfen. In anderen Worten: Der Test prüft, ob die Varianzen in den verschiedenen Gruppen, die in einer Studie verglichen werden, signifikant voneinander abweichen.

Varianzanalyse

- Die Varianzanalyse stellt einen Mittelwertvergleich zwischen mehr als 2 Gruppen dar (bis zu 2 Gruppen t-Test). z. B. Belastung von Lehrer/innen von 3 Schultypen (AHS, HS, VS)

- Mittelwertsunterschiede werden mit Hilfe von Varianzschätzungen untersucht. Es passiert ein Vergleich zweier Varianzen über einen F-Test.

- Warum rechnet man nicht 3 t-Tests oder U-Tests?

→Weil Fehler wird kumuliert!!!!

Varianzanalyse

Varianzanalysen werden danach klassifiziert **wie viele unabhängige Variablen (UV)** in ihrer Bedeutung für eine abhängige Variable (AV) untersucht werden.

Einfaktorielle V. untersucht den Einfluss einer UV auf eine AV

Beispiele:

Unterscheiden sich stark, moderat und geringfügig emotionale Menschen bezüglich der Menge des Stresshormons Cortisol im Blut?

Werden bei Konsumenten nach einer Degustation von fünf Traubensorten unterschiedlich hohe Zahlungsbereitschaften für die verschiedenen Sorten gemessen?

Varianzanalyse: Voraussetzungen

Abhängige Variable:

- metrische Skalenniveau
- Homogenität der Varianzen
- Normalverteilung der Messwerte

Unabhängige Variable:

- nominal oder ordinal Skalenniveau
- mehr als 2 Stuffen

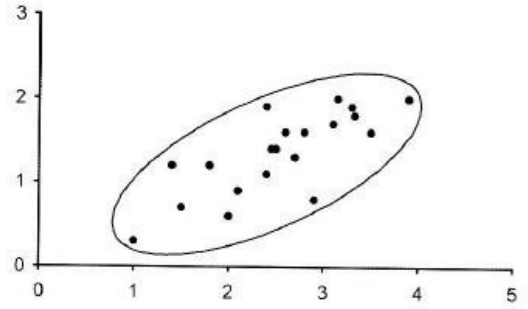
Korrelation

Streu(ungs)diagramme sind grafische Hilfsmittel, die die Anordnung der Beobachtungspunkte veranschaulichen

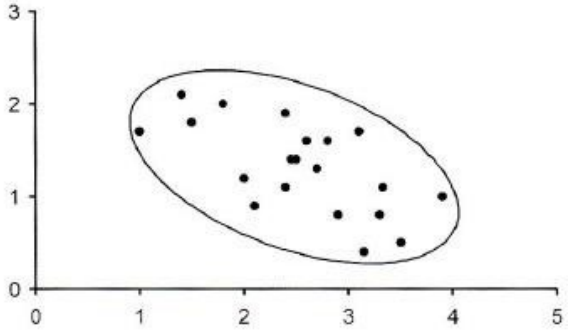
Jedes xi/yi - Beobachtungspaar wird in ein x/y-Koordinatensystem eingetragen

es lässt sich ein erster Eindruck gewinnen, ob und wie stark zwei Merkmale zusammenhängen

positiver Zusammenhang



negativer Zusammenhang



Korrelation

- als Korrelation bezeichnet man den wechselseitigen Zusammenhang zwischen Größen

- Korrelation bedeutet nicht das Vorhandensein von Kausalität. Besteht eine Korrelation zwischen X und Y, so gibt es mindestens drei alternative Möglichkeiten einer Kausalitätsbeziehung:

  - X bewirkt Y
  - Y bewirkt X und
  - X und Y werden durch Z bewirkt (Scheinkorrelation bzw. partielle Korrelation).

- die Korrelationsanalyse liefert ein Maß für die Stärke des Zusammenhangs; erfasst jedoch nur monotone bzw. lineare Zusammenhänge

Korrelation

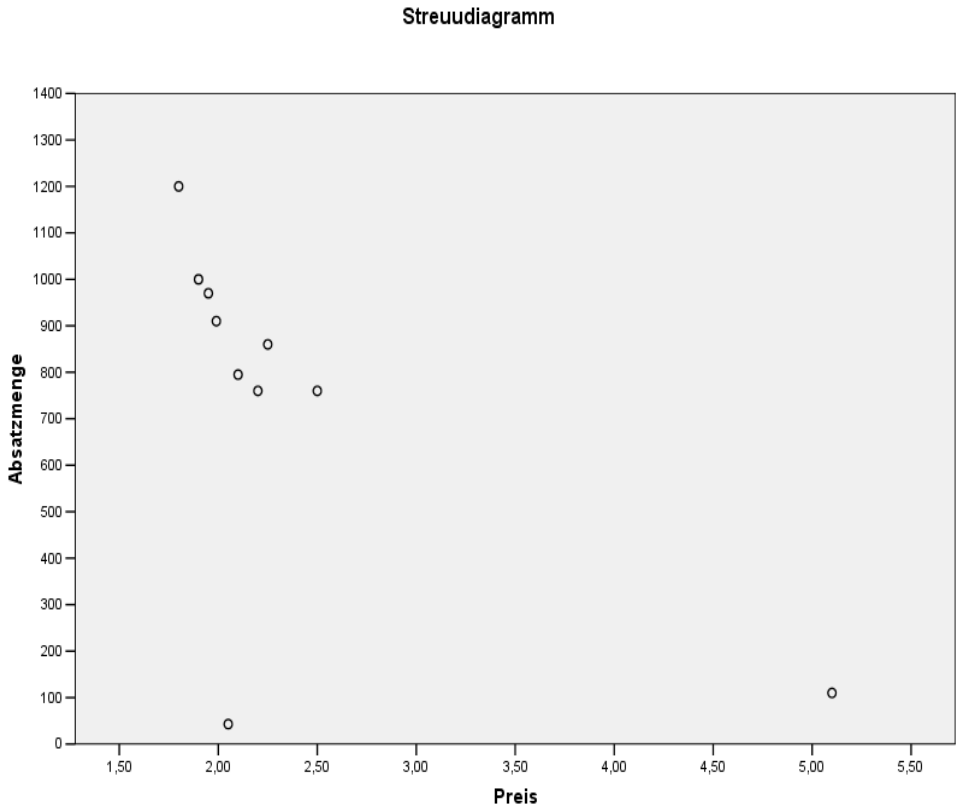
- die Stärke des Zusammenhangs wird durch den Korrelationskoeffizienten r gemessen
- r liegt stets in den Grenzen von -1 bis +1

- für die Stärke des Zusammenhangs ist allein der Betrag des Korrelationskoeffizienten maßgebend

- das Vorzeichen gibt an, ob der Zusammenhang gleichläufig (+) oder gegenläufig (–) ist

Korrelationskoeffizient	Einstufung
$\lvert r \rvert \leq 0.25$	schwache Korrelation
$0.25 < \lvert r \rvert \leq 0.66$	mittlere Korrelation
$0.66 < \lvert r \rvert < 1$	starke Korrelation
$\lvert r \rvert = 1$	perfekte Korrelation

Korrelationskoeffizient nach Pearson



Vermutung:

Zwischen den Variablen Preis und Verkaufsmenge besteht ein linearer und gegenläufiger Zusammenhang;

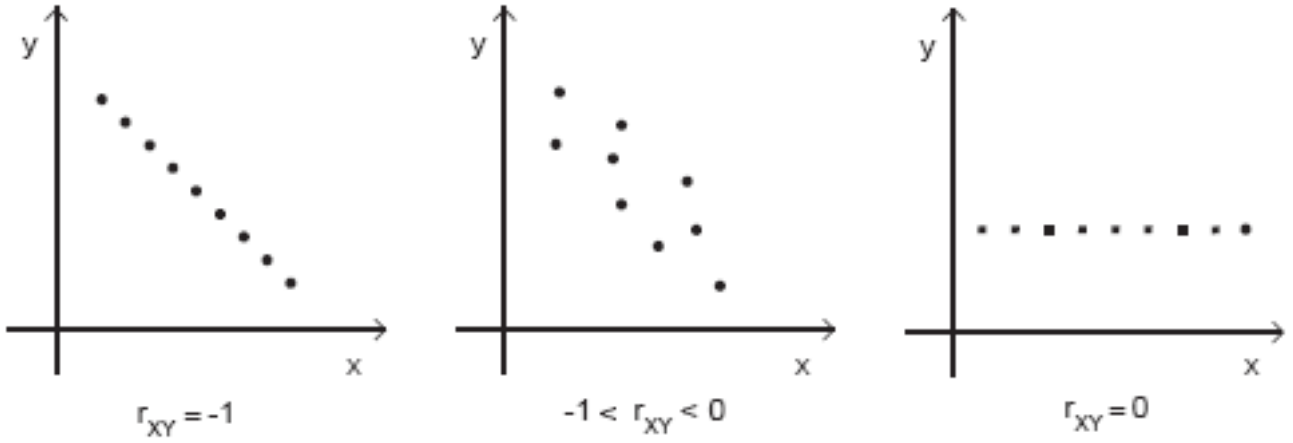
je höher der Verkaufspreis umso geringer die Absatzmenge.

Korrelationskoeffizient nach Pearson

- zur Beschreibung des Zusammenhangs zwischen metrisch skalierten
- misst die Stärke des linearen Zusammenhangs

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Korrelationskoeffizient nach Pearson



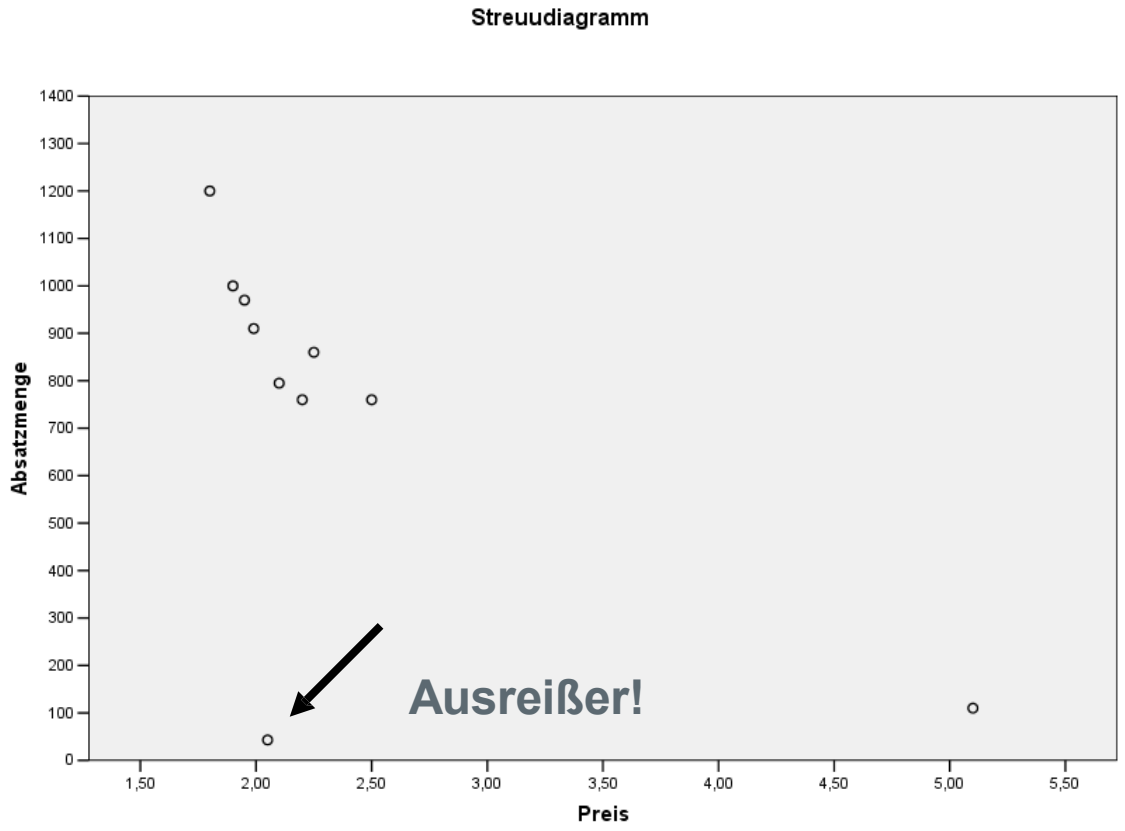
Korrelationskoeffizient nach Pearson

r misst den linearen Zusammenhang zweier Variablen

- **Konsequenz**: einzelne Ausreißer, d.h. einzelne extreme Datenpunkte, können einen starken, unerwünschten Effekt auf den numerischen Wert von r haben; hohe Korrelationen können als gering erscheinen und umgekehrt.

- **Lösung**: Ermittlung von Rangkorrelationskoeffizienten, die von Ausreißern wesentlich weniger beeinflusst werden, da ihre Ermittlung auf den Rängen der Beobachtungen basiert.

Rangkorrelationskoeffizient nach Spearman



Rangkorrelationskoeffizient nach Spearman

- wird zwischen zwei Variablen berechnet, die mindestens ordinalskaliert sind; für metrisch skalierte Variablen, bei Unsicherheit hinsichtlich der Normalverteilungsanahme misst die Stärke des monotonen Zusammenhangs

- basiert auf Rangzahlen, die den Messwerten zugeordnet sind für beide Variablen wird eine Rangreihe der Werte erstellt, dem höchsten Wert wird der Rangplatz 1 verliehen; bei gleichen Werten werden gemittelte Rangplätze vergeben

- die Differenz d der zugehörigen Rangplatzpaare wird bestimmt

Rangkorrelationskoeffizient nach Spearman

- Wertebereich von rs: -1 bis +1
- gehen mit steigenden x-Werten auch steigende y-Werte einher, so nimmt rs tendenziell einen großen Wert an
- sind die Rangzahlen bei den Merkmalen beider Variablen völlig gleich, so nimmt rs den Wert 1 an (die Rangpaare liegen auf einer Geraden mit positiver Steigung liegen)
- bei entgegengesetzt laufenden Rangzahlen wird rs = -1 (die Rangpaare liegen auf einer Geraden mit negativer Steigung)

Rangkorrelationskoeffizient nach Spearman

Rechenschritte zur Rangkorrelation nach Spearman rs :

Tiefkühlpizza	A	B	C	D	E	F	G	H	I	J
Preis in Euro	5,10	1,80	2,10	2,05	1,99	1,90	2,20	1,95	2,50	2,25
Absatzmenge im Monat	110	1200	100	43	910	1000	760	970	685	860
Rang Preis	1	10	5	6	7	9	4	8	2	3
Rang Absatzmenge	8	1	9	10	4	2	6	3	7	5
d_i	-7	9	-4	-4	3	7	-2	5	-5	-2
d^2_i	49	81	16	16	9	49	4	25	25	4

Rangkorrelationskoeffizient nach Spearman

Es ergibt sich

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d^2}{n(n^2 - 1)} \qquad r_s = 1 - \frac{6 * 278}{10 * (100 - 1)} = -0{,}685$$

Rangkorrelationskoeffizient nach Spearman

Interpretation des Ergebnisses

rs= -0,685

=> starker Zusammenhang

rs< 0 => gegenläufiger monotoner Zusammenhang

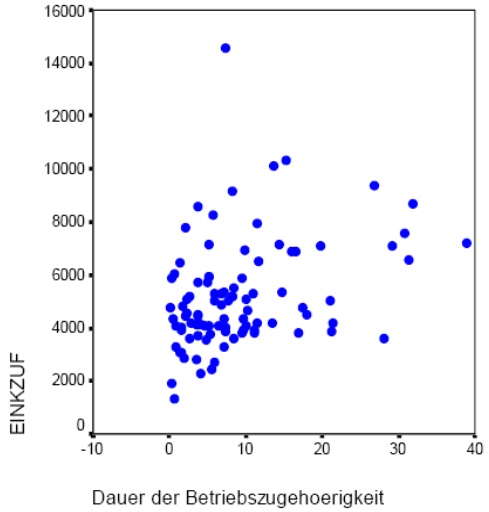
Es zeigt sich ein mittlerer gegenläufiger Zusammenhang zwischen Preis und Absatzmenge: Je höher der Preis einer Tiefkühlpizza, umso niedriger ist die verkaufte Menge an Tiefkühlpizzen.

Kontigenzkoeffizient

- Maß zur Charakterisierung des Zusammenhangs zweier nominal skalierter Variablen
- Da hier nur Häufigkeiten untersucht werden spricht man von Kontingenz statt Korrelation.
- C kann nur positive Werte annehmen mit den theoretischen Grenzen 0 und 1.
- Der Maximale Kontingentkoeffizient ergibt sich aus der Anzahl Spalten und Zeilen einer Kontingenztabelle.

Probleme

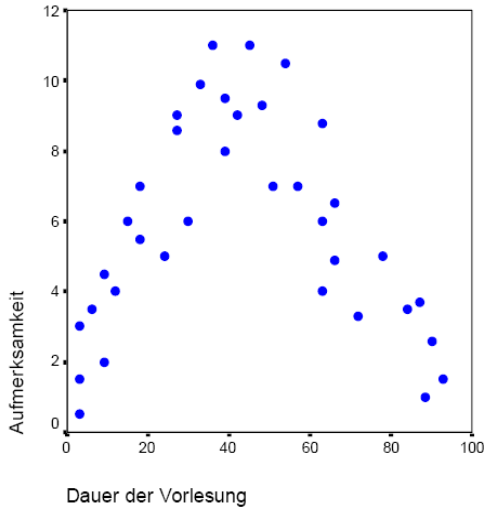
für die Korrelation i.e.S gilt: Einzelne Fälle können einen starken Einfluss auf den Korrelationskoeffizienten ausüben.



Korrelation über alle Fälle: r=0,35.

Korrelation ohne Einkommen über 14.000: r=0,39.

Korrelationen lassen sich für alle Funktionstypen berechnen; allerdings werden nur monotone bzw. lineare Zusammenhänge erfasst.



Korrelation: -0,05, d.h. praktisch gleich null.

Das Beispiel ist fiktiv, Sie sind natürlich ständig aufmerksam! ☺

DHBW Lörrach

Probleme

- Kausalzusammenhänge können nicht erfast werden

- Scheinkorrelationen (Korrelation zwischen Merkmalen, die inhaltlich nicht gerechtfertigt ist) können auftreten; Zusammenhänge ergeben sich dann, wenn ein mit beiden beobachtbaren Merkmalen hochkorreliertes drittes Merkmal übersehen wird und unberücksichtigt bleibt.

DHBW Lörrach

Vielen Dank!