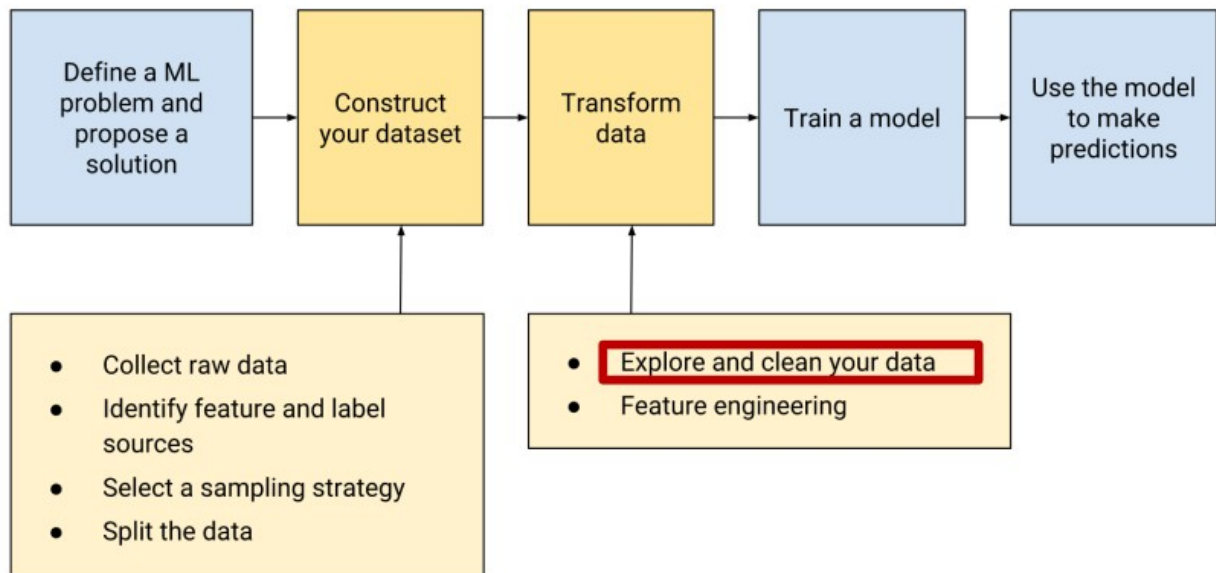
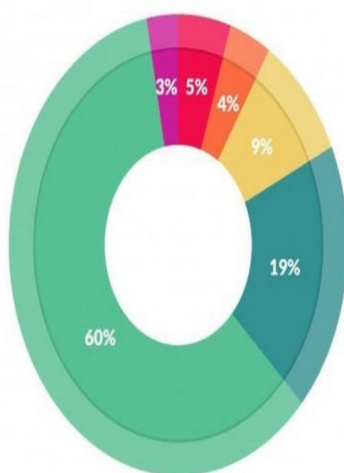


Datenaufbereitung

Datenaufbereitung beinhaltet die Anpassung eines strukturierten Datensatzes, um sicherzustellen, dass er nahtlos in unser Machine Learning-Modell integriert werden kann. Dies ist von entscheidender Bedeutung, um sicherzustellen, dass unsere Modelle nicht nur funktionieren, sondern auch optimale Leistung und präzise Vorhersagen liefern können. Wir haben also einen strukturierten Datensatz vorliegen und müssen diesen gemäß unseres Machine Learning Modells anpassen.

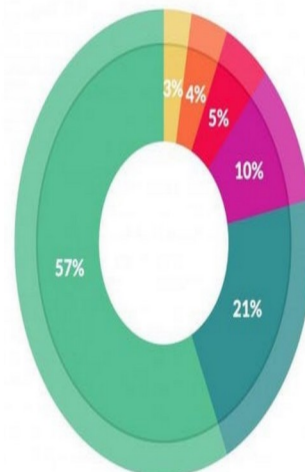


Datenaufbereitung ist **Teil der Datentransformation**. Hier verändert man die Daten, damit Sie für unser ML-Modell geeignet sind und gut performed. Manche Schritte sind notwendig, damit die Modelle überhaupt funktionieren. Heute geht es um die Datenexploration und Datensäuberung (Feature Engineering kommt später in einer anderen VL).



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%



What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

Frage: Welcher Prozess nimmt die meiste Zeit ein und macht laut Data Scientists am wenigsten Spaß?


Antwort: Datensäuberung und Datenorganisation – Datenvorverarbeitung.

Siehe Abbildung.

Laut einer Forbes Umfrage von 2016.

Thema 1: Datenstrukturierung

Zuerst sollten wir analysieren, **wie** strukturierte Daten für Ihr Modell **am besten strukturiert werden können** und welche **wichtigen Aspekte** dabei berücksichtigt werden müssen.

 Australian Bureau of Statistics		Australian Bureau of Statistics										Table junk									
1800.0 Australian Marriage Law Postal Survey, 2017																					
Released on 15 November 2017																					
Table 5 Participation by Federal Electoral Division(a), Males and Age		Gender apartheid																			
Yeah NA																					
		18-19 years	20-24 years	25-29 years	30-34 years	35-39 years	40-44 years	45-49 years	50-54 years	55-59 years	60-64 years										
Lingia(c)	Total participants	292	1,058	1,465	1,653	1,515	1,516	1,710	1,730	1,753	1,574										
	Eligible participants	572	2,910	3,789	3,996	3,607	3,506	3,645	3,331	2,960	2,456										
	Participation rate (%)	51.0	36.4	38.7	41.4	42.0	43.2	46.9	51.9	59.2	64.1										
Primary keynotes		Comma on																			
Merged cells Solomon	Total participants	442	1,461	2,066	2,357	2,188	2,057	2,224	2,108	2,134	1,772										
	Eligible participants	750	2,991	3,994	4,155	3,634	3,398	3,427	3,066	2,931	2,355										
	Participation rate (%)	58.9	48.8	51.7	56.7	60.2	60.5	64.9	68.8	72.8	75.2										
Northern Territory (Total)	Total participants	734	2,519	3,531	4,010	3,703	3,573	3,934	3,838	3,887	3,346										
	Eligible participants	1,322	5,901	7,783	8,151	7,241	6,904	7,072	6,397	5,891	4,811										
	Participation rate (%)	55.5	42.7	45.4	49.2	51.1	51.8	55.6	60.0	66.0	69.5										
Australian Capital Territory Divisions		Covariate as Subheading										Summary of data inside data									
Canberra(d)	Total participants	1,764	4,789	4,817	4,973	4,626	4,453	5,074	4,826	5,169	4,394										
	Eligible participants	2,260	6,471	6,448	6,509	5,983	5,805	6,302	5,902	6,044	5,057										
	Participation rate (%)	78.1	74.0	74.7	76.4	77.3	76.7	80.5	81.8	85.5	86.9										
Fenner(e)	Total participants	1,477	4,687	5,178	5,786	6,025	5,463	5,191	4,208	3,948	3,465										
	Eligible participants	1,904	6,354	7,121	7,822	7,960	7,155	6,480	5,206	4,692	3,945										
	Participation rate (%)	77.6	73.8	72.7	74.0	75.7	76.4	80.1	80.8	84.1	87.8										
		NA Yeah																			
Australian Capital Territory (Total)	Total participants	3,241	9,470	9,995	10,739	10,051	9,910	10,205	9,034	9,117	7,639										
	Eligible participants	4,164	12,825	13,569	14,331	13,943	12,960	12,782	11,108	10,736	9,002										
	Participation rate (%)	77.8	73.9	73.7	75.1	76.4	76.5	80.3	81.3	84.9	87.3										
Australia																					
Total	Total participants	151,297	438,166	441,658	460,548	462,206	479,360	524,620	517,693	543,449	506,799										
	Eligible participants	201,439	635,909	646,916	665,250	656,446	660,841	693,850	659,150	664,720	597,386										
	Participation rate (%)	75.1	68.9	68.3	69.2	70.4	72.5	75.6	78.5	81.8	84.8										
a) The Federal Electoral Divisions are current as at 24 August 2017																					
b) Includes those whose age is unknown												Return of the table junk									
c) Includes Christmas Island and the Cocos (Keeling) Islands																					
d) Includes Norfolk Island																					
e) Includes Jervis Bay																					
												MS Excel or Die									

Häufig wird es der Fall sein, dass ihre Daten nicht strukturiert sind. Oft werden Daten für andere Zwecke erhoben und erst dann für eine Analyse oder ML angewendet. Deshalb liegen die Daten nicht für ML-Modelle strukturiert/ geeignet vor. In der Abbildung sieht man ein Beispiel. Hier handelt es sich um eine Umfrage, wo Australier gefragt wurden, was ihre Einstellung gegenüber same sex marriage ist. Die Teilnahme war freiwillig. Das Australian Bureau of Statistics hat die Ergebnisse dann so veröffentlicht.

Frage: Welche Probleme gibt es im Datensatz des Bureaus of Statistics?

Table Junk: Einige Elemente wie Header oder Footnotes sind nicht gut für ML. Diese müssten vorher entfernt werden.

Yeah NA: Einige Felder haben eine Feldbezeichnung (Spaltenname z.B.) und andere nicht.

Sonderzeichen sind oft schwer zu importieren.

Komma in csv: **Englische csv-Dateien nutzen das Komma (,) als Separator und sollten deshalb nicht für Float-Nummern verwendet werden.**

Zwischenüberschriften entfernen.

Summary of data inside data: Summary-Statistik sollte nicht in der selben Spalte wie die dazugehörigen Daten stehen. Trenne Sie in einer eigenen Spalte.

NA Yeah: Leerzeilen machen Probleme beim importieren von Daten in R oder Python.

Der erste Schritt der Datenvorverarbeitung besteht deshalb darin, die Daten erstmals vernünftig zu strukturieren – den ganzen Müll von oben entfernen.

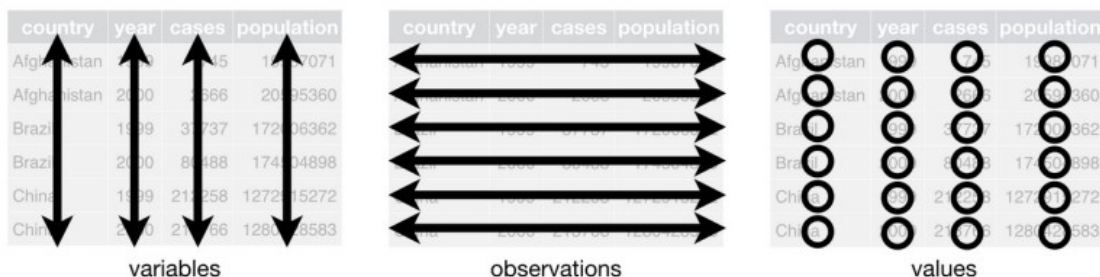
Miles McBain. Tidying the Australian Same Sex Marriage Postal Survey Data with R. medium.com zeigt wie man es mit R machen kann. Man kann es aber auch schneller händisch in Excel oder co machen.

"Happy families are all alike; every unhappy family is unhappy in its own way." — Leo Tolstoy

"Tidy datasets are all alike, but every messy dataset is messy in its own way." — Hadley Wickham

Was bedeutet das Zitat? Gut aufgeräumte Datensätze sind gleich, aber alle nicht gut aufgeräumten Datensätze sind auf ihre eigene Art und Weise messy. Jeder Datensatz braucht also eine individuelle Behandlung, weil nicht jeder Fehler in der Struktur in jedem Datensatz am selben Ort vorliegt. Das macht die Datenaufbereitung so kompliziert.

Frage: Was sind Variablen, Observationen und Values in einer Matrix/ Datensatz? (Begriffe kennen und verstehen)



Die Prinzipien, wie ein vernünftig strukturierter Datensatz vorzuliegen hat, scheinen offensichtlich zu sein. Leider ist es sehr oft (immer) so, dass die Daten, mit denen Sie arbeiten werden, nicht aufgeräumt sein werden. Das liegt daran, dass die meisten Menschen, die nicht regelmäßig und wissenschaftlich mit Daten arbeiten, die Prinzipien nicht verinnerlicht haben. Oft werden Daten aber auch nicht für Maschinelle Analysen aufbereitet. Oft versuchen die Leute ihre Daten so zu gestalten, dass Sie möglichst schnell eingetippt werden können oder Sie bereiten Sie für eine Präsentation vor (siehe Abbildung oben). Sie versuchen ihre Daten also hübsch zu gestalten und bauen dabei viele Fußnoten, Header, Subheader, blank spaces und andere Elemente in ihre Daten ein, die eine maschinelle Analyse erschweren.

Es gibt gewisse Regeln, an die man sich halten kann, damit ein Datensatz strukturiert vorliegt.

1. Man kann jede Observation in die Zeilen/Reihen eintragen – 1 Observation = eine eigene Reihe.
2. Die Variablen/ Features sollten in den Spalten stehen – 1 Spalte = 1 Variable.
3. In den Zellen der Kreuztabelle sollten die Werte/Ausprägungen der einzelnen Variablen für jede einzelne Observation stehen.

Man muss oft seinen Datensatz umstrukturieren, um diese Form (Kreuztabelle) zu erreichen.

Oft muss man erstmals herausfinden, was die einzelnen Observationen sind. Dafür muss man Domainexperten und die Forscher, die mit dem Projekt verbunden sind, reden.

Oft sind Variablen und Observationen über mehrere Spalten verteilt. Dann muss man den Domainexperten fragen, was jetzt zusammengehört und was eine Observation sein soll.

```
table1
#> # A tibble: 6 x 4
#>   country      year cases population
#>   <chr>      <int> <int>      <int>
#> 1 Afghanistan 1999     745  19987071
#> 2 Afghanistan 2000    2666  20595360
#> 3 Brazil      1999   37737  172006362
#> 4 Brazil      2000   80488  174504898
#> 5 China       1999  212258 1272915272
#> 6 China       2000  213766 1280428583
```



In der Abbildung sehen Sie ein gutes Beispiel für einen gut strukturierten Datensatz.

```
table2
#> # A tibble: 12 x 4
#>   country      year type      count
#>   <chr>      <int> <chr>      <int>
#> 1 Afghanistan 1999 cases         745
#> 2 Afghanistan 1999 population 19987071
#> 3 Afghanistan 2000 cases         2666
#> 4 Afghanistan 2000 population 20595360
#> 5 Brazil      1999 cases         37737
#> 6 Brazil      1999 population 172006362
```



Table 2 ist hingegen ein schlechtes Beispiel. Es sind dieselben Daten wie Table 1, aber nicht mehr aufgeräumt. Wie weiter oben erwähnt, kann es sein, dass dieselbe Observation über mehrere Spalten oder Zeilen aufgeteilt sein kann. Das sehen Sie in der tabelle, weil Afghanistan über mehrere Reihen aufgeteilt vorliegt. Das müssten Sie dann verändern.


```
table3
#> # A tibble: 6 x 3
#>   country      year rate
#> * <chr>      <int> <chr>
#> 1 Afghanistan 1999 745/19987071
#> 2 Afghanistan 2000 2666/20595360
#> 3 Brazil      1999 37737/172006362
#> 4 Brazil      2000 80488/174504898
#> 5 China       1999 212258/1272915272
#> 6 China       2000 213766/1280428583
```



In Tabelle 3 sind Fallzahl und Population in einer Spalte eingetragen wurden und als falscher Datentyp abgespeichert wurden. Das müsste korrigiert werden.

```
table4a # cases
#> # A tibble: 3 x 3
#>   country      `1999` `2000`
#> * <chr>      <int> <int>
#> 1 Afghanistan    745    2666
#> 2 Brazil         37737   80488
#> 3 China          212258  213766
table4b # population
#> # A tibble: 3 x 3
#>   country      `1999`      `2000`
#> * <chr>      <int>      <int>
#> 1 Afghanistan 19987071  20595360
#> 2 Brazil      172006362 174504898
#> 3 China       1272915272 1280428583
```



Tabelle 4 hat 2 Observationen in einer Zeile zusammengefasst, was eigentlich nicht passieren sollt (1 Observation = eigene Reihe). Die Merkmale wurden weiterhin auf 2 Tabellen verteilt.

Oft hat man das Problem, dass Spaltenüberschriften nicht die Namen einzelner Variablen enthalten. Numerische Spaltenüberschriften machen auch Probleme. Das sehen wir in der Abbildung, wo 1999 und 2000 als Character genutzt werden, um als Spaltennamen zu dienen. 1999 und 2000 sind Ausprägungen einer anderen Variable und sollten nicht die Spaltenüberschrift sein.

> Tipp: **Numerische Spaltenüberschriften vermeiden!**

Frage: Was ist die Unterscheidung zwischen den Wide Format und dem Long Format?

Wide-Format

	Probandennummer	Geschlecht	Gewicht.1	Gewicht.2	Gewicht.3
1	1	w	163,7	153,0	131,9
2	2	w	144,4	162,9	117,2
3	3	w	153,6	126,1	133,3
4	4	w	156,3	137,2	147,1
5	5	w	154,0	138,7	154,0
6	6	m	148,9	146,4	130,7
7	7	m	165,1	137,2	132,4
8	8	m	149,1	113,4	117,4
9	9	m	170,2	115,6	139,6
10	10	m	149,4	153,2	128,6

Long-Format

	Probandennummer	Geschlecht	Zeitpunkt	Gewicht
1.1	1	w	1	163,7
2.1	2	w	1	144,4
3.1	3	w	1	153,6
4.1	4	w	1	156,3
5.1	5	w	1	154,0
6.1	6	m	1	148,9
7.1	7	m	1	165,1
8.1	8	m	1	149,1
9.1	9	m	1	170,2
10.1	10	m	1	149,4
1.2	1	w	2	153,0
2.2	2	w	2	162,9

Eine wichtiger Aspekt in der Datenaufbereitung ist die Unterscheidung zwischen Wide und Long Format. Hier geht es ganz einfach um die **Form der Tabellen**. Im Wide Format sind die Daten in einer breiten, spaltenorientierten Form angeordnet, während das Long Format die Daten in einer schmalen, zeilenorientierten Struktur präsentiert. Die Wahl zwischen diesen Formaten kann erheblichen Einfluss auf die Analyse und Modellierung haben, abhängig von den spezifischen Anforderungen Ihres Projekts.

> wider = spaltenorientiert.

> long = Zeilenorientiert.

Thema 2: Datensichtung

Hier geht es um das Thema der vernünftigen Datensichtung. Es geht also nicht mehr um die Struktur der Daten (wie gehen davon aus, dass Sie jetzt strukturiert wurden), sondern es geht um den Inhalt der Daten.

Datensichtung: Plot your data!

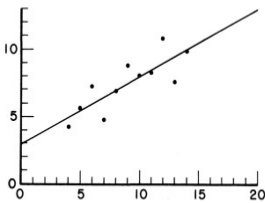


Figure 1

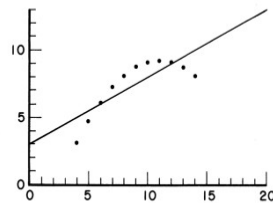


Figure 2

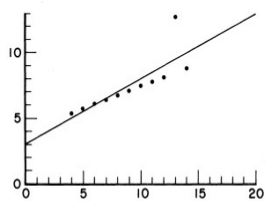


Figure 3

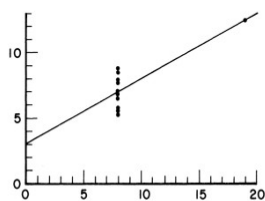
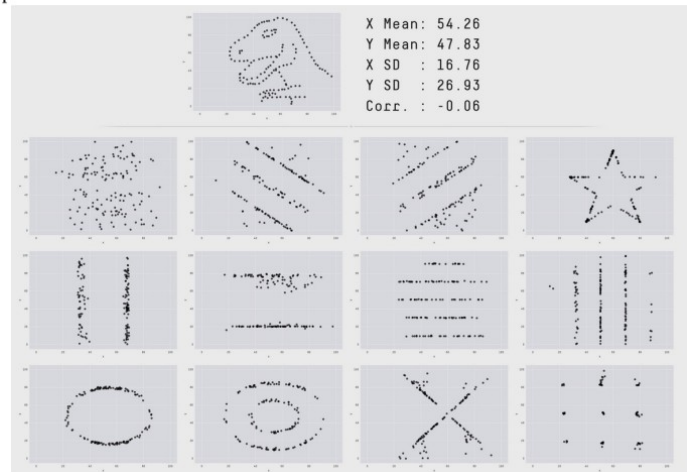


Figure 4

Each of the four data sets yields the same standard output from a typical regression program, namely

Number of observations (n) = 11
Mean of the x 's (\bar{x}) = 9.0
Mean of the y 's (\bar{y}) = 7.5
Regression coefficient (b_1) of y on x = 0.5
Equation of regression line: $y = 3 + 0.5x$
Sum of squares of $x - \bar{x}$ = 110.0
Regression sum of squares = 27.50 (1 d.f.)
Residual sum of squares of y = 13.75 (9 d.f.)
Estimated standard error of b_1 = 0.118
Multiple R^2 = 0.667



Frage: Reicht es nicht aus, wenn ich die Summary Statistik alleine angebe? Wieso muss man Daten visualisieren?

Eine wichtige Anweisung zum Thema Sichtung, die oft in Blogs, Tweets zu Data Science oder Forumeinträgen angegeben wird, ist folgende: **Plot your Data.**

Stell deine Daten graphisch da und schau euch die Verteilung der Daten an und sucht nach Besonderheiten/ Auffälligem im Datensatz. Wichtige Summary Statistics (siehe Abbildung) für die Beschreibung der Daten können bei völlig verschiedenen Datensätzen ähnlich oder identisch sein kann. Das hat die berühmte Arbeit von Anscombe FJ von 1973 (1) gezeigt. In der Abbildung sieht man 4 Datensätze, die völlig verschieden sind, wenn man Sie plottet, aber ihre Summary Statistiken sind identisch. Wichtige Informationen über die Form der Daten gehen verloren, wenn man nur Summary Statistiken angibt. Der DataSaurier in der anderen Abbildung generiert auch mehrere verschiedene Plots bei gleicher Summary Statistik (2).

Lerne: Summary Statistik alleine reicht nicht aus. Visualisiere die Daten.

1: Anscombe FJ. Graphs in Statistical Analysis The American Statistician, Vol. 27, No. 1. (Feb., 1973), pp. 17-21

2: Matejka J & Fitzmaurice G. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing, Autodesk Research, Toronto Ontario Canada

Thema 3: Missing Data

Hier geht es um das große Thema Missing Data. Sie lernen hier, wie man mit Missing Data umgehen kann, was man zu tun hat und was man ignorieren kann. Es ist ein extrem prävalentes Thema in allen Projekten, die mit Daten zu tun haben.

Methodology Editorial	R H H Groenwold and O M Dekkers	Missing data: the impact of what isn't there	183:4	E8
-----------------------	------------------------------------	---	-------	----

Table 1 Numerical examples of the possible impact of missing data in a hypothetical trial of levothyroxine.

Scenario	Percentage missing data	Levothyroxine treatment	Placebo	RR (95% CI)
		n = 120	n = 120	
A	0%	24/120 (20%)	40/120 (33%)	0.60 (0.93; 0.93)
B	2%	19/115 (17%)	40/120 (33%)	0.50 (0.31; 0.80)
C	2%	24/120 (20%)	35/115 (30%)	0.66 (0.42; 1.03)
D	50%	12/60 (20%)	20/60 (33%)	0.60 (0.32; 1.12)
E	17%	15/100 (15%)	25/100 (25%)	0.60 (0.34; 1.07)

RR, risk ratio.

European Journal of
Endocrinology
(2020) 183, E7–E9



In der Abbildung sehen Sie, wie das Fehlen von Daten (im fiktiven Datensatz) die Ergebnisse einer Studie verändern kann. Im Datensatz wird die Zugabe von Levothyroxin bei Schilddrüsenunterfunktion dokumentiert. Es handelt sich um eine Placebokontrollierte Studie und es wird die weitere Entwicklung der Schilddrüse nach Medikamentverabreichung beobachtet. In Datensatz A gibt es keine fehlenden Daten (siehe Spalte „Percentage Missing Data“). Bei B und C fehlen 2 %, bei D fehlen die Hälfte der Daten und bei E 17 % der Daten. In allen Szenarien hat man jeweils 120 Probanden in Treatment und Placebo-Gruppe. Bei A beobachtete man eine 20 %-ige Verschlechterung der Drüsenfunktion bei Treatment und eine 33 % Verschlechterung bei der Placebo-Gruppe. Das ergibt ein relatives Risiko von 0,6 durch die Zugabe von Levothyroxin. In Szenario B fehlen Daten aus der Treatment-Gruppe von Patienten, die eine Verschlechterung aufwiesen. Dadurch ist das Risiko geringer in der Treatmentgruppe und man beobachtet einen größeren klinischen Effekt, der aber auf dem Fehlen von Daten basiert und nicht auf einem echten Effekt. Nur 2 % der Daten fehlten und doch beobachten wir einen großen Effekt. In Szenario C fehlen nur Daten in der Kontrollgruppe. Dadurch haben wir 30 % statt 33 % für die Kontrollgruppe und das relative Risiko ist größer (spricht schlechter für den Einsatz von Levothyroxin). In D fehlen viele Daten, aber Sie fehlen gleich verteilt über beide Gruppen. Die Ergebnisse ändern sich deshalb nicht. Aber: durch das Fehlen von Daten ist der klinische Effekt nicht mehr signifikant.

Fazit: Fehlende Daten können Ergebnisse verzehren. Man darf Sie nicht ignorieren. Man muss Sie behandeln.


```
stocks <- tibble(
  year   = c(2015, 2015, 2015, 2015, 2016, 2016, 2016),
  qtr    = c( 1,    2,    3,    4,    2,    3,    4),
  return = c(1.88, 0.59, 0.35, NA, 0.92, 0.17, 2.66)
)
```

- **Explicitly**, i.e. flagged with NA .
- **Implicitly**, i.e. simply not present in the data.



Frage: Welche 2 Arten von Missing Data gibt es?

Es gibt zwei verschiedene Arten von Missing Data:

1. **Explizit** fehlende Daten: Das ist dadurch ausgeprägt, dass man im Datensatz Werte hat, die als fehlend gekennzeichnet werden – also über NA oder NaN zum Beispiel. Daten sind schon als fehlend gekennzeichnet.

> Explizit = NA/ NaN Werte

2. **Implizit** fehlende Daten: Das sind fehlende Daten, die einfach nicht vorhanden sind im Datensatz. Implizit bedeutet also einfach, dass ganze Zeilen zum Beispiel nicht dokumentiert wurden.

> Implizit = Nicht im Datensatz gekennzeichnet.

In der Abbildung sehen wir unter return ein NA. Das ist explizit. Wir sehen aber auch, dass wir für 2015 und 2016 4 Quartale angegeben haben sollten, aber Quartal 1 für 2016 fehlt. Das ist implizit. Implizierte Missing Data muss man raus suchen. Man erkennt Sie nicht so einfach.

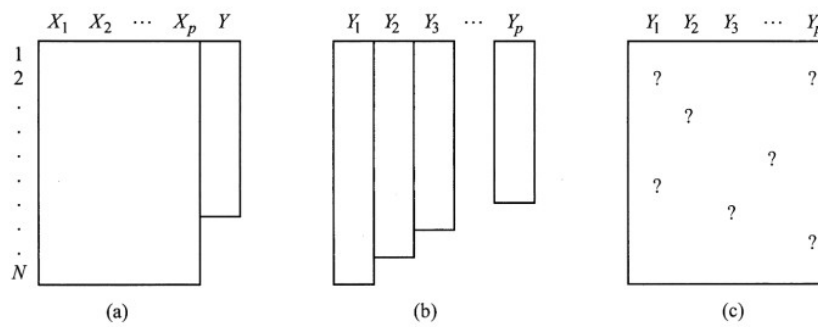


Figure 1. Patterns of nonresponse in rectangular data sets: (a) univariate pattern, (b) monotone pattern, and (c) arbitrary pattern. In each case, rows correspond to observational units and columns correspond to variables.

Frage: Was sind die Muster/ pattern bei Missing Data und welche Muster gibt es? Definiere die 3 Muster. Was ist das häufigste Pattern?

Eine weitere wichtige Unterscheidung bei fehlenden Daten sind die **Muster/ pattern**. Es ist wichtig, dass man die Muster kennt und unterscheiden kann, um zu entscheiden, wie man mit Missing Data richtig umzugehen hat.

In Bezug auf fehlende Daten (Missing Data) bezieht sich das Konzept von "Mustern" auf die Art und Weise, **wie das Fehlen von Daten in einem Datensatz organisiert oder verteilt ist**. Es geht darum, zu verstehen, ob es eine systematische oder zufällige Struktur im Fehlen von Daten gibt. Im Kontext der fehlenden Daten werden "Muster" verwendet, um zu beschreiben, ob das Fehlen von Daten auf vorhersehbaren Regeln oder Bedingungen basiert.

Abbildung 1 zeigt die 3 wichtigen Muster, wonach die Daten fehlen können.

A zeigt das **Univariate** Muster. B zeigt das **Monotone** Muster. und C zeigt das **willkürliche** Muster.

Univariate Muster des fehlenden Datensatzes beziehen sich auf die Anzahl der Variablen, die fehlende Daten aufweisen. Ein univariates Muster tritt auf, wenn nur eine einzelne Variable in einem Datensatz fehlende Daten aufweist. Das bedeutet, dass das Fehlen von Daten ausschließlich auf diese eine Variable zurückzuführen ist und nicht auf andere Variablen im Datensatz. In der Abbildung sehen wir z.B., dass nur Spalte Y fehlende Werte hat.

Ein fehlendes Datenmuster wird als **monoton** bezeichnet, wenn die Variablen Y_j so geordnet werden können, dass, wenn Y_j fehlt, alle Variablen Y_k mit $k > j$ ebenfalls fehlen. Das tritt beispielsweise in **Langzeitstudien** mit Teilnehmerabbruch (Drop-Out) auf. Wenn ich also die Variablen so sortiere, dass ich mit den vollständigsten Spalten(Variablen) anfangen und dann alle Variablen nach ihrer absteigenden Länge links nach rechts sortiere, dass die Spalten nach rechts hin kürzer werden sollten, weil mehr Daten fehlen. Wenn eine Variable fehlt, dann würden alle folgenden Variablen auch Fehlende Daten enthalten.

Das häufigste Pattern: Willkürliche Pattern. Bereitet auch am meisten Probleme. Hier kann man kein genaues Muster beobachten. Es fehlt einfach irgendwo irgendetwas im Datensatz. Die fehlenden Werte sind zufällig im Datensatz verteilt.

Das Wissen über das Muster ist wichtig, um zu wissen, wie man dagegen vorzugehen hat. Multiple Imputation zum Beispiel (und andere Methoden) setzt voraus, dass man es auf das richtige Muster anwendet.

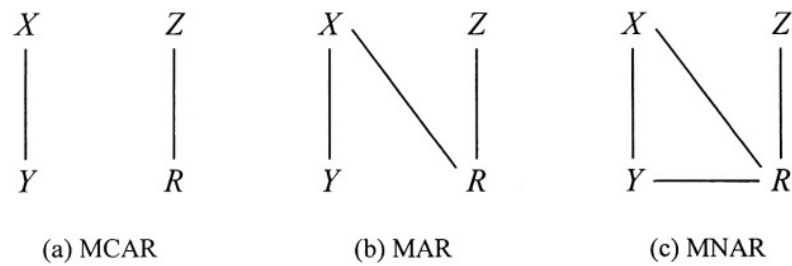


Figure 2. Graphical representations of (a) missing completely at random (MCAR), (b) missing at random (MAR), and (c) missing not at random (MNAR) in a univariate missing-data pattern. X represents variables that are completely observed, Y represents a variable that is partly missing, Z represents the component of the causes of missingness unrelated to X and Y , and R represents the missingness.

Frage: Welche 3 Unterscheidungen macht man bei zufälligen und nicht-zufällig fehlenden Daten? Kenne alle 3 Begriffe.

Eine weitere sehr wichtige Unterscheidung ist, ob die Daten **zufällig fehlen** oder ob Sie **nicht zufällig fehlen**.

Man unterscheidet zwischen

MCAR → Missing completely at random

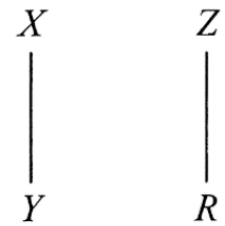
MAR → Missing at random

MNAR → Missing not at random.

Im Kontext von MCAR, MAR und MNAR ist es von entscheidender Bedeutung, die Ursache für das Fehlen der Daten zu verstehen, festzustellen, ob es sich um zufälliges oder systematisches Fehlen handelt, und herauszufinden, ob es eine erkennbare Struktur hinter diesem Muster gibt. Diese Erkenntnisse sind unerlässlich, um festzulegen, wie die Daten am besten behandelt werden sollten und welche Analysemethoden angewendet werden müssen.

MCAR: Missing completely at random

- Definition: Das Fehlen der Daten in einer Variable ist weder mit der (fehlenden) Ausprägung der Variable verbunden noch mit den anderen Variablen (Rubin, 1976)
- Es gibt statistische Tests, die MCAR ablehnen (Little, 1988), aber keine, um MCAR abschließend zu bestätigen.
- MCAR kann angenommen werden, wenn es einen eindeutig zufälligen Anlass im Datengenerierungsprozess gibt, der in keinem Zusammenhang mit den betrachteten Variablen stand, wie z. B. ein technisches Problem in der elektronischen Krankenakten, die eine einzelne Variable einer bestimmten Beobachtung löschen



(a) MCAR

Kommen wir zuerst zu MCAR. Das bedeutet, dass Fehlen der Daten in einer Variable weder mit der fehlenden Ausprägung der Variable verbunden ist, noch mit den anderen Variablen in Verbindung steht.

Simple Erklärung:

Wenn Daten nach dem MCAR-Muster fehlen, bedeutet das, dass das Fehlen der Daten in einer Variable nichts mit den Werten in dieser Variable oder mit den Werten in anderen Variablen im Datensatz zu tun hat. Das Fehlen tritt einfach völlig zufällig auf, ohne dass es eine erkennbare Verbindung zu den Daten gibt.

Ein einfaches Beispiel: Stell dir vor, du hast einen Datensatz mit Informationen über Menschen, einschließlich ihres Geschlechts, Alters und Einkommens. Wenn das Fehlen der Daten im Datensatz MCAR ist, bedeutet das, dass beispielsweise das Fehlen des Einkommenswerts nicht von Geschlecht oder Alter abhängt. Es kann genauso wahrscheinlich bei Männern wie bei Frauen, jungen oder alten Menschen auftreten. Wenn die Wahrscheinlichkeit, dass Daten fehlen, für alle Fälle gleich ist, dann sagt man, dass die Daten vollständig zufällig fehlen (MCAR). Das bedeutet im Grunde, dass die Ursachen für das Fehlen der Daten keine Verbindung zu den Daten haben. Ein Beispiel für MCAR wäre eine Waage, deren Batterien leer sind. Einige der Daten fehlen einfach aufgrund von Pech.

Kurz gesagt, MCAR bedeutet, dass das Fehlen der Daten rein zufällig ist und keine Beziehung zu den beobachteten oder nicht beobachteten Variablen besteht.

> Das Fehlen von Daten hat keinen bestimmten Grund und kann nicht durch andere Variablen erklärt werden.

- **Rein zufällig fehlend**

- **Fehlen ist weder von der eigentlichen Variablen noch von anderen Variablen abhängig**

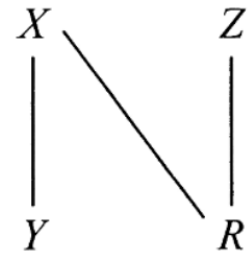
Es existieren statistische Tests, mit denen man MCAR ablehnen kann (Referenz von Little 1). Diese Tests können Hinweise darauf liefern, ob das Fehlen von Daten zufällig oder systematisch ist. Es gibt aber, um MCAR abschließend zu bestätigen.

1. Little RJA. A Test of Missing Completely at Random for Multivariate Data with Missing Values. Journal of the American Statistical Association 1988;83:1198–202. doi:10.2307/2290157

MCAR deutet auf zufällig auftretende Missing Data hin. Das passiert zum Beispiel, wenn es bei der Dokumentation zu Fehlern gekommen ist. Wenn zum Beispiel die Maschine, die die Daten dokumentiert hat, zufällig Fehler gemacht hat.

MAR: Missing at random

- Definition: Das Fehlen der Daten einer Variable steht in keinem Zusammenhang mit ihren (fehlenden) Werten, nachdem für die anderen Variablen im Datensatz kontrolliert wurde (Rubin, 1976)
- MAR ist eine „notwendige“ Voraussetzung für Verfahren der Behandlung von Missing Data, siehe nächste Folien zu Multiple Imputation
- In einer Befragung könnten zum Beispiel junge Teilnehmer eher bereit gewesen sein eine Antwort auf die Frage nach der sexuellen Orientierung zu geben als ältere.



(b) MAR

MAR bedeutet, dass das Fehlen der Daten einer Variable in keinem Zusammenhang mit ihren fehlenden Werten steht, nachdem für die anderen Variablen im Datensatz kontrolliert wurde.

Simple Erklärung:

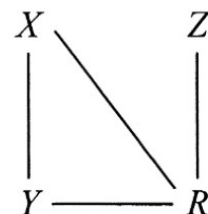
MAR bedeutet, dass, wenn wir Daten vermissen, dies von den bereits vorhandenen Daten abhängt. Zum Beispiel, wenn wir wissen, dass Menschen mit höherem Einkommen eher ihre Einkommensdaten angeben, dann sind die fehlenden Einkommensdaten MAR. Die fehlenden Daten hängen also von den bereits vorhandenen Daten ab.

- **Fehlen ist unabhängig von der Variable selbst, bei der der Wert fehlt**
- **Fehlen ist abhängig von anderen Variablen im Datensatz**

MAR ist eine „notwendige“ Voraussetzung für Verfahren der Behandlung von Missing Data, siehe Multiple Imputation.

MNAR: Missing not at random

- Definition: Die Daten sind weder MCAR noch MAR, d.h. das Fehlen der Daten ist auch nach Kontrolle für die übrigen Variablen noch mit dem (fehlenden) Wert der Missing Data verbunden



(c) MNAR

Die Daten sind weder MCAR noch MAR, d.h. das Fehlen der Daten ist auch nach Kontrolle für die übrigen Variablen noch mit dem (fehlenden) Wert der Missing Data verbunden.

Simple Erklärung:

Mit anderen Worten, selbst wenn du alle anderen Informationen und Faktoren im Datensatz berücksichtigst, bleibt immer noch eine Verbindung zwischen dem Fehlen der Daten und dem Wert der fehlenden Daten bestehen. Das Fehlen ist also nicht zufällig oder unabhängig von den Daten selbst.

Wenn Männer sich weigern, an einer Umfrage zur Depression teilzunehmen, weil sie tatsächlich depressiv sind, dann sind die fehlenden Daten MNAR. Das Fehlen der Daten ist hier nicht zufällig, sondern wird von der tatsächlichen Depression beeinflusst.

- Fehlen hängt von der fehlenden Variable selbst ab

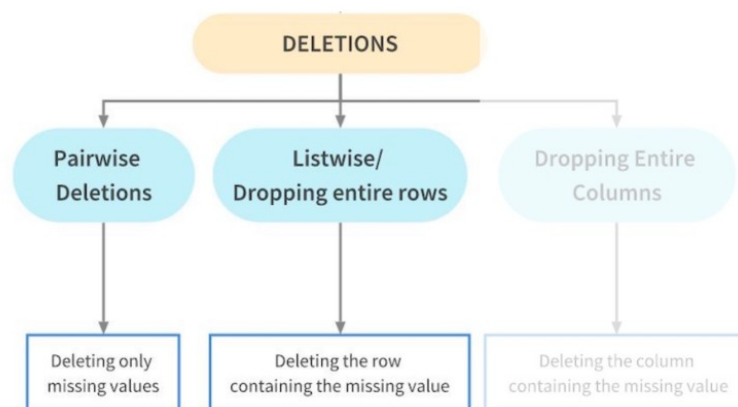
Methoden zur Behandlung von Missing Data

Nachdem wir jetzt wissen, auf welche Art und Weise unsere Daten fehlen können, stellen wir uns jetzt die Frage, was wir dagegen unternehmen können. Es gibt drei Möglichkeiten, wie man mit Missing Data umgehen kann.

Achtung: Viele Machine Learning Algorithmen haben bereits Methoden zur Behandlung von Missing Data integriert. Oft wird nur gelöscht. Manche haben aber auch Indicator variables implementiert.

1. Deletion der Missing Data

Deletions



- Vorteil: Extrem einfach
- Nachteil: Nur bei MCAR, sonst Biasgefahr. Selbst bei MCAR Verlust von Datenpunkten.

Zuerst hat man die Option Missing Data aus seinem Datensatz zu löschen. Fehlende Daten werden also auf unterschiedliche Arten und Weisen gelöscht.

Frage: Aus welche Art und Weise kann man Missing Data löschen? - 3 Möglichkeiten

1. Man kann **ganze Spalten löschen**. Wir entfernen also ganze Variablen, wo zum Beispiel zu viele Daten fehlen. Wenn in einer Variable deutlich mehr Daten fehlen als in allen anderen Variablen, dann macht es Sinn die Variable zu entfernen.

2. Der Klassiker und am häufigsten benutzte Verfahren ist die **Listwise deletion**. In Papern wird auch manchmal auch von Casewise Analysis gesprochen. Das bedeutet, dass man die Observationen/ **Reihen entfernt**, wo die fehlenden Werte für unsere Variable auftreten. Man entfernt hier also nicht die ganze Spalte/ Variable, sondern nur die Patienten/Samples in den Reihen, die keine Werte aufweisen. Man nutzt also nur die Observationen, die vollständig sind.

3. ein sehr ähnliches Verfahren zur Listwise Deletion ist das **Paarweise Entfernen**. Hier löschen wir nicht die ganze Observation/Reihe, sondern wir behalten die Observationen und löschen nur die Daten, die fehlen. Man behält also so viele Daten, wie es möglich ist. Im Gegensatz zur Listwise Deletion, bei der gesamte Beobachtungen gelöscht werden, wenn auch nur ein einziger Wert fehlt, werden beim Paarweisen Entfernen nur die fehlenden Werte gelöscht, während die übrigen Datenpunkte in einer Beobachtung beibehalten werden. Beim Paarweisen Entfernen wird jede Beobachtung oder Zeile im Datensatz für sich betrachtet. Wenn in einer Beobachtung fehlende Werte vorkommen, werden nur diese spezifischen fehlenden Werte gelöscht. Die restlichen Datenpunkte in dieser Beobachtung bleiben unverändert und werden in der Analyse berücksichtigt. Dies bedeutet, dass unterschiedliche Beobachtungen unterschiedlich viele Datenpunkte enthalten können, abhängig davon, welche Werte fehlen.

Frage: Wann kann man eine Deletion von Missing Data in Betracht ziehen? Was sind die Vor und Nachteile der Deletion?

Ein Problem bei dieser ganzen Sache ist, dass man Deletion nur nutzen kann, wenn die Missing Data **MCAR** vorliegen. Sonst hat man die Gefahr, dass es einen Bias gibt. Das bedeutet, dass das Fehlen der Daten vollständig zufällig und unabhängig von anderen Variablen oder der zugrunde liegenden Population ist. In diesem Fall ist die Wahrscheinlichkeit, dass eine Beobachtung fehlt, unabhängig von den Werten dieser Beobachtung oder anderen Faktoren.

Selbst bei MCAR ist es so, dass man Datenpunkte verliert. Wir verlieren also auch an statistischer Power (das haben wir bei dem Schilddrüsenbeispiel oben gesehen).

Vorteil von Deletion: Es ist sehr **einfach umzusetzen**.

2. Arbeite mit Indicator Variables

x	x^*	r
-1	-1	0
1	→ 1	0
3	3	0
—	0	1

Missing Indicator Verfahren:

Nutz eine Dummy variable zur Repräsentation des Fehlens

Vorteil: Schnell und einfach

Nachteil: Selbst bei MCAR Biasrisiko (Donders et al., 2006)

Frage: Was sind Indicator Variables?

Eine weitere Methode für den Umgang mit Missing Data sind die Indicator Variablen. Die Verwendung von Indikatorvariablen, auch als Dummy-Variablen bezeichnet, ist eine gängige Methode, um mit fehlenden Daten umzugehen. Diese Methode beinhaltet die Schaffung einer zusätzlichen binären (0 oder 1) Variable, die anzeigt, ob in einer bestimmten Beobachtung Daten fehlen oder nicht. Diese Indikatorvariablen werden

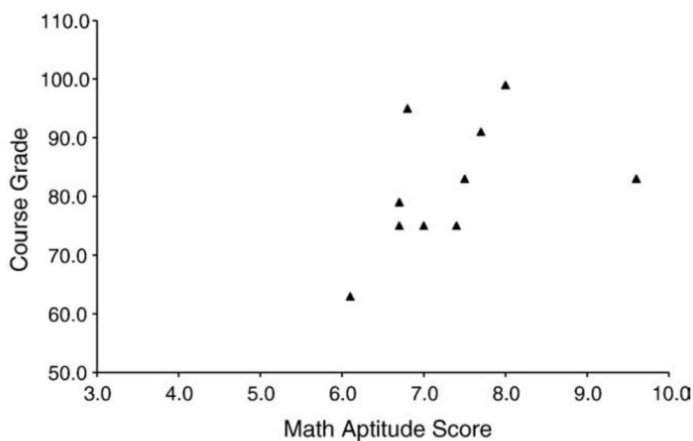
oft in statistischen Analysen verwendet, um sicherzustellen, dass die Information über das Fehlen der Daten in den Analysen berücksichtigt wird. Angenommen, Sie haben eine Variable x , bei der einige Werte fehlen. Statt die fehlenden Werte einfach zu löschen, erstellen Sie eine zusätzliche Spalte r (die Indikatorvariable), die für jede Beobachtung den Wert 1 hat, wenn der Wert von x fehlt, und den Wert 0, wenn x vorhanden ist. Das sehen Sie auch in der Abbildung. Diese Indikatorvariablen sind hilfreich, da sie es ermöglichen, das Fehlen der Daten als eigenständige Variable in den Analysen zu berücksichtigen (Sie können beispielsweise in Regressionsanalysen als Prädiktoren verwendet werden, um den Einfluss des Fehlens auf andere Variablen zu bewerten).

Frage: Was sind die Vor- und Nachteile von Indikatorvariablen?

+ Schnell und einfach.

- Selbst bei MCAR hat man ein *Biasrisiko*. In der klassischen statistischen Datenmodellierung, wo man häufig den Zusammenhang zwischen Variable X und dem Outcome darstellen will, sollte man Indikatorvariablen nicht nutzen. Hier wären Indikatoren schlechter als listwises löschen.

3. Imputationsmethoden



Imputation methods

1. Mean imputation
2. Regression imputation
3. Stochastic regression imputation
4. Multiple imputation

Baraldi AN, Enders CK. An introduction to modern missing data analyses. *Journal of School Psychology* 2010;48:5–37. doi:10.1016/j.jsp.2009.10.001



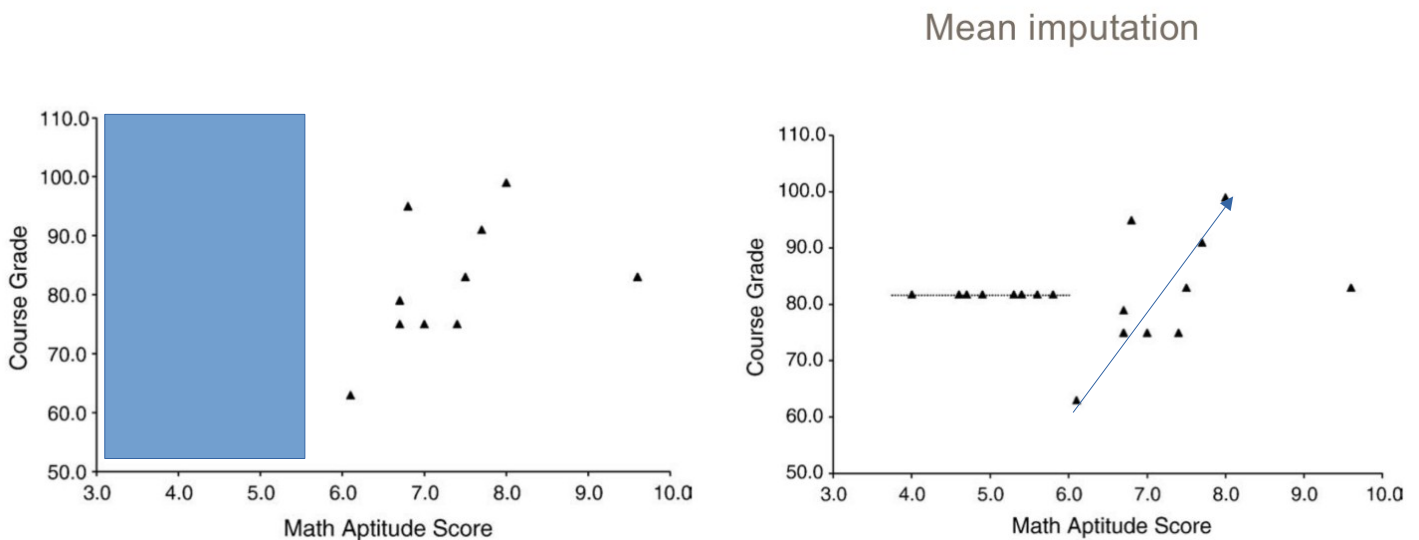
Frage: Welche Imputationsmethode ist der goldene Standard?

Jetzt kommen wir zu den Imputationsmethoden. Es existieren verschiedene Imputationsmethoden. Der goldene Standard in der Statistik ist die **Multiple Imputation**. Aber auch die anderen 3 Methoden (Abbildung) werden häufig genutzt, weil Sie schneller zu implementieren sind. Beim ML kann man auch die anderen Methoden nutzen.

Frage: Welche 4 Imputationsmethoden haben wir in dieser VL kennengelernt (siehe Abbildung)?

Gehen wir durch die 4 Imputationsmethoden.

1. Mean Imputation



Reden wir zuerst über Mean Imputation. In der Abbildung links hat man einen Datensatz visualisiert und man sieht, dass die Datenpunkte auf der X-Achse erst bei 6.0 anfangen. Die Bereiche davor (blauer Kasten) haben keine Punkte. Mean Imputation ist eine gängige Methode zur Behandlung fehlender Daten, bei der der Durchschnitt (Mittelwert) der verfügbaren Daten für die betreffende Variable berechnet wird. Anschließend werden die leeren oder fehlenden Werte in dieser Variable durch den ermittelten Mittelwert ersetzt. Das sieht man in der Abbildung rechts, wo die leeren Bereiche mit demselben Mittelwert gefüllt wurden (siehe Linie aus Punkten).

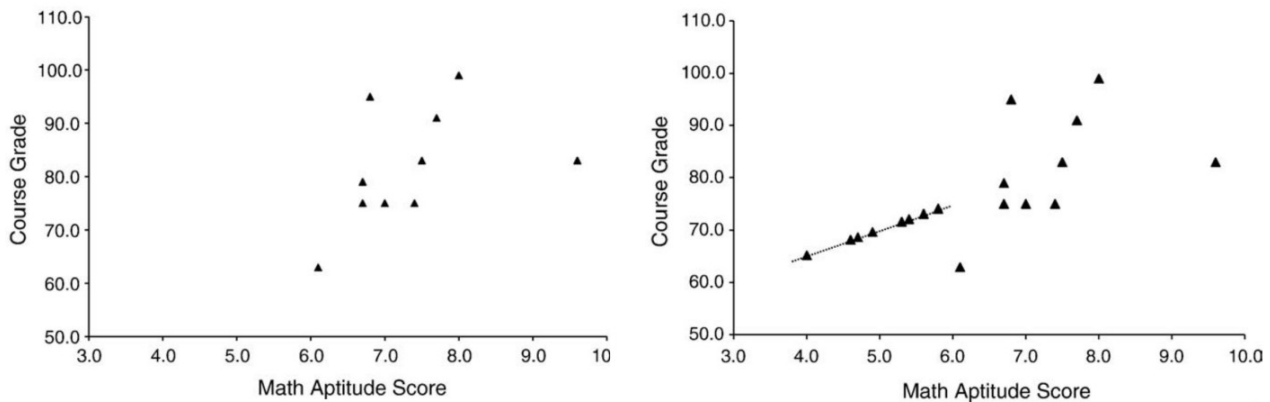
Frage: Was sind die Nachteile der Mean Imputation?

Aus der Abbildung rechts wird einem klar, welche Nachteile die Mean Imputation hat. Wenn wir zum Beispiel wissen wollen, welchen Einfluss der Math Aptitude Score auf die Course Grades hat, dann würden wir durch die Mean Imputation einen Bias bekommen. Dies liegt daran, dass die Mean Imputation den Durchschnittswert für den Math Aptitude Score in den leeren Bereichen einfügt, unabhängig davon, welche individuellen Unterschiede zwischen den Studierenden bestehen. Die resultierenden Daten könnten daher dazu führen, dass der Math Aptitude Score als homogener und weniger variabler Parameter betrachtet wird, als er tatsächlich ist.

In der Realität können die Math Aptitude Scores jedoch erheblich variieren, und diese Variabilität kann wichtige Informationen darüber liefern, wie dieser Faktor die Course Grades beeinflusst. Die Werte links, wo jetzt die Mean-Werte eingetragen sind, könnten zum Beispiel geringere Werte haben auf der Y-Achse. Dann wäre der Mean zu groß. Der natürliche Verlauf der Daten wäre damit zerstört (man erkennt an den Daten, dass es einen leichten Verlauf der Punkte von unten links nach oben rechts gibt (siehe Pfeil)). Durch die Mean Imputation gehen diese individuellen Unterschiede verloren, und die wahre Beziehung zwischen Math Aptitude und Course Grades wird verzerrt.

2. Regression Imputation

Regression imputation



ine weitere Möglichkeit zur Behandlung fehlender Daten ist die Regression Imputation. Im Gegensatz zur Mean Imputation, bei der einfach der Durchschnittswert verwendet wird, um fehlende Daten zu ersetzen, versucht die Regression Imputation, die fehlenden Werte auf Grundlage der Beziehungen zu anderen verfügbaren Variablen genauer zu schätzen. Dieser Ansatz mildert das Problem der Mean Imputation, bei dem der Durchschnittswert zu hoch sein kann und somit potenziell verzerrte Ergebnisse liefert.

Regressionsimputation bedeutet, dass man in einer bivariaten Analyse die Werte, die man hat, nutzt, um ein Regressionsmodell zu bauen. Darüber werden die fehlenden Werte über eine Regression vorhergesagt und mit den Regressionswerten aufgefüllt.

Konkret läuft der Prozess wie folgt ab:

Eine Regressionsanalyse wird durchgeführt, bei der die Variable mit fehlenden Werten (die abhängige Variable) in Beziehung zu anderen relevanten Variablen (unabhängige Variablen) gesetzt wird.

Das Regressionsmodell wird basierend auf den vorhandenen Daten geschätzt, wodurch die Beziehung zwischen den Variablen aufgezeichnet wird.

Mit diesem geschätzten Modell können dann die fehlenden Werte der abhängigen Variable vorhergesagt werden, indem die Werte der unabhängigen Variablen in das Modell eingesetzt werden.

Die vorhergesagten Werte werden verwendet, um die fehlenden Werte in der ursprünglichen Variable zu ersetzen.

Es ist jedoch wichtig zu beachten, dass die Qualität der Regressionsimputation von der Gültigkeit des Modells und der Auswahl der unabhängigen Variablen abhängt.

Frage: Was sind die Vor- und Nachteile?

Vorteile der Regression Imputation:

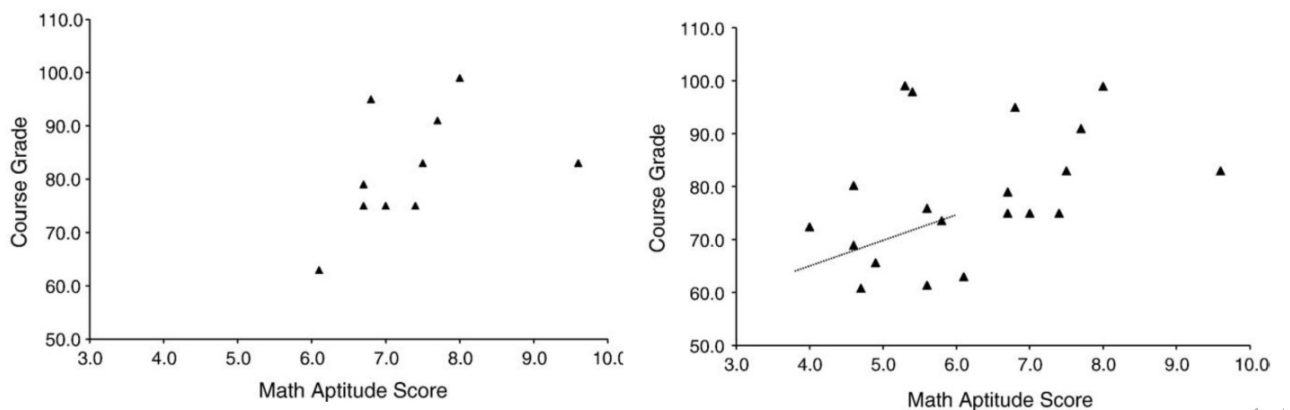
Anders als der Mean erhält man einen Verlauf der Daten und nicht nur einen Datenwert für alle Missing Values.

Probleme/ Nachteile:

1. Wenn unser Interesse darauf gerichtet ist, die Korrelation zwischen den beiden Variablen zu untersuchen, dann kann die Verwendung der Regressionsimputation zu einer Überschätzung führen. Dies liegt daran, dass die Regressionsimputation auf der Annahme basiert, dass es eine kausale Beziehung zwischen den Variablen gibt, wobei eine als unabhängige und die andere als abhängige Variable betrachtet wird. Durch diese Annahme wird die Beziehung zwischen den Variablen verstärkt, da die imputierten Werte in der abhängigen Variable stark von den unabhängigen Variablen beeinflusst werden. Dies kann dazu führen, dass die beobachtete Korrelation zwischen den Variablen künstlich erhöht wird, da die Imputation die Annahme einer stärkeren Verbindung zwischen ihnen impliziert, als es möglicherweise der Fall ist.
 - > Regressionsimputation sorgt für perfekt korrelierte Datenpunkte, dort wo Missing Values sind.
 - > **Regressionsimputation nicht bei Korrelationsbestimmungen nutzen!**
2. Die **Variabilität der Daten geht verloren**, da der Verlauf linear verläuft. Variabilität des gesamten Datensatzes wird unterschätzt durch den künstlichen Zusammenhang der 2 Variablen.

3. Stochastic regression imputation

Stochastic regression imputation



Frage: Welchen Vorteil hat die stochastische Regression gegenüber der „normalen“ Regressionsimputation?

Die stochastische Regressionsimputation kann das *Problem mit der fehlenden Variabilität, die in der normalen Regressionsimputation beobachtet wird, mildern*. Im Kontext der stochastischen Regressionsimputation wird erneut eine Regressionsanalyse durchgeführt. Allerdings wird hierbei eine wichtige Ergänzung vorgenommen, indem ein zufällig generierter Fehler (random noise) in das Modell einbezogen wird. Dieser Fehler bewirkt, dass die geschätzten Werte nicht nur auf der Regressionsgeraden liegen, sondern um diese Linie herum verteilt werden.

Durch die Einbeziehung des zufälligen Fehlers wird die Simulation von Unsicherheit in die Imputation integriert. Dies bedeutet, dass die geschätzten Werte nicht mehr nur genau auf der Regressionslinie liegen, sondern eine gewisse Streuung aufweisen, die die natürliche Variation in den Daten besser widerspiegelt.

Die Unterschätzung der Varianz in den Daten ist dadurch gemildert und die **Korrelation ist nicht mehr so stark überschätzt**.

Man sieht diese Methode aber fast nie.

Frage: Was sind die Nachteile?

Problem: Die stochastische Regressionsimputation unterschätzt den **tatsächlichen Standardfehler**.

Der Standardfehler wird auch als Stichprobenfehler oder „standard error of the mean (SEM)“ bezeichnet. Er gibt an, wie stark der Mittelwert einer Stichprobe vom Mittelwert der Grundgesamtheit abweicht

Standardfehler = Standardabweichung der Stichprobe / Stichprobengröße

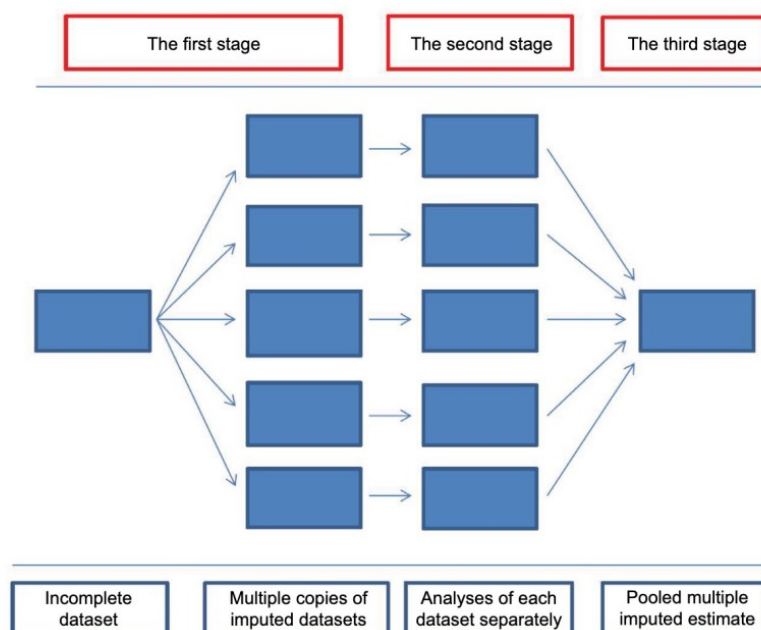
Dies geschieht, weil sie nicht in Betracht zieht, dass die Punkte aufgrund von Schätzungen generiert wurden. Es ist anzunehmen, dass diese Schätzungen nicht exakt den wahren Werten entsprechen. Die geschätzten Werte repräsentieren somit nicht die exakten, realen Werte. Der Standardfehler wird damit oft unterschätzt, weil die Unsicherheit in den geschätzten Werten nicht berücksichtigt wird. Um dieser zusätzlichen Variabilität und Unsicherheit angemessen Rechnung zu tragen, greifen wir auf die Multiple Imputation zurück.

4. Multiple Imputation

Multiple Imputation bedeutet, dass der Schritt, den wir bei der stochastischen Regressionsimputation gesehen haben, nicht nur einmal durchgeführt wird, sondern wiederholt und mehrmals. Dabei werden stochastisch generierte Punkte um eine Regressionslinie herum platziert. Dieser Vorgang wird nicht nur einmal wiederholt, sondern viele Male, um wiederholt Punkte im Raum zu setzen.

Mit anderen Worten, bei Multiple Imputation werden die fehlenden Werte nicht nur einmal geschätzt, sondern es werden mehrere Schätzungen unter Berücksichtigung von Unsicherheit und Variation erstellt. Dies ermöglicht eine umfassendere Erfassung der Unsicherheit in den Daten, da die geschätzten Werte in verschiedenen Kombinationen und mit zufälligen Faktoren mehrmals erzeugt werden.

Es muss nicht nur stochastische Regression verwendet werden. Es gibt auch andere Möglichkeiten.



In der Abbildung sehen Sie schematisch, wie die multiple Imputation funktioniert.

Frage: Was sind die 3 Hauptstadien der multiplen Imputation?

Die Methode der Multiplen Imputation (Multiple Imputation, MI) besteht aus drei Hauptstadien:

Imputationsstadium (Imputation Stage): In diesem ersten Stadium werden die fehlenden Werte im Datensatz mithilfe eines Imputationsmodells geschätzt. Hierbei wird für jede fehlende Beobachtung ein Satz von mehreren (üblicherweise 3 bis 5) geschätzten Werten erstellt. Diese Schätzungen basieren auf statistischen Techniken wie linearen Regressionen, Random Forests, oder anderen Modellen, die auf den verfügbaren Daten beruhen. Bei diesem Schritt wird Unsicherheit in die Schätzungen eingeführt, indem stochastische oder zufällige Fehler in die Imputation einbezogen werden.

In der Abbildung sehen wir, dass wir mit einem unvollständigem Datensatz anfangen. Im ersten Schritt erzeugen wir mehrere vollständige Datensätze (5 in der Abbildung) durch die multiple Imputation. **Wir haben so 5 vollständige Datensätze, aber die Werte, die anstelle der Missing Data eingesetzt wurden, sind zufällig.**

Analysenstadium (Analysis Stage): Im zweiten Stadium werden separate statistische Analysen für jede der imputierten Datensätze durchgeführt. Das bedeutet, dass für jeden der erstellten Imputationsdatensätze eine vollständige Analyse ausgeführt wird. Dies kann beispielsweise die Anwendung von Regressionsmodellen, Hypothesentests oder anderen statistischen Analysen umfassen. Die Ergebnisse dieser Analysen werden für jeden Datensatz separat erfasst.

> **Wir analysieren jeden imputierten Datensatz einzeln.**

Poolingstadium (Pooling Stage): Im dritten Stadium erfolgt das Zusammenführen der Ergebnisse aus den separaten Analysen, um einen aggregierten Output zu generieren. Dieser Schritt wird als "Pooling" bezeichnet. Die Ergebnisse, wie Schätzwerte, Konfidenzintervalle und Standardfehler, aus den separaten Analysen werden miteinander kombiniert, um einen konsolidierten Ausgang zu erzeugen. Typischerweise werden Durchschnittswerte und Varianzen der geschätzten Parameter aus den verschiedenen Imputationsdatensätzen berechnet.

Die Verwendung von Multiple Imputation zielt darauf ab, die Unsicherheit, die durch fehlende Daten entsteht, in den analytischen Ergebnissen zu berücksichtigen. Die Methode ermöglicht realistischere Schätzungen und eine genauere Quantifizierung der Unsicherheit in den Ergebnissen, indem sie mehrere Schätzungen für die fehlenden Daten erstellt und diese miteinander kombiniert.

Der komplexeste Schritt ist der erste Schritt: Imputationsstadium.

Es existieren verschiedene Algorithmen für die Generierung von vollständigen imputierten Datensätzen.

Welche Methode ich benutze, hängt von der Art der Daten ab. Es gibt manche Verfahren, die nur bei univariaten und monotonen Missing Data funktionieren. Es gibt aber auch Einige, die bei willkürlichen Missing Data Mustern funktionieren

Frage: Wieso muss man die Art und Muster der Missing Data kennen?

Merke: Muster und Art der Missing Data ist entscheidend für die Wahl des Algorithmus für die multiple Imputation im ersten Stadium.

Frage: Was ist der Vorteil von Multiple Imputation?

Damit können wir die Variabilität und Unsicherheit in unseren Daten besser darstellen.

Frage: Welche Methode eignet sich gut für willkürliche Muster bei missing Data?

Multiple Imputation by Chained Equations (MICE) funktioniert bei willkürlichen Mustern. Es handelt sich also um Multiple Imputation durch aneinander gekettete Gleichungen.

Hier soll die Methode erklärt werden (aber nicht klausurrelevant). MICE (Multiple Imputation by Chained Equations) ist ein Verfahren zur Schätzung von fehlenden Daten in einem Datensatz. Der Prozess beginnt damit, dass fehlende Werte einer Variable zufällig aus den vorhandenen Werten derselben Variable ersetzt werden. Wenn also in einer Beobachtung Daten fehlen, werden diese durch zufällig aus den vorhandenen Daten ausgewählte Werte ersetzt. Dies stellt sicher, dass die Imputation realistische Werte verwendet, die bereits im Datensatz vorhanden sind.

In einem weiteren Schritt wird eine Variable ausgewählt, die im Vergleich zu den anderen Variablen die geringste Anzahl von fehlenden Werten aufweist. Diese ausgewählte Variable wird dann in Beziehung zu den anderen Variablen mit fehlenden Werten gesetzt und mithilfe einer Regressionsanalyse modelliert. Das Hauptziel dieses Schritts besteht darin, die Beziehungen zwischen den fehlenden und vorhandenen Variablen zu erfassen und die verfügbaren Daten bestmöglich zu nutzen.

Die Grundidee hinter MICE ist, dass die verbleibenden fehlenden Werte in einer Variable wahrscheinlich von den Werten anderer Variablen beeinflusst werden. Durch die Regression können diese Abhängigkeiten erfasst werden, was zu genaueren Schätzungen der fehlenden Werte führt, basierend auf den Beziehungen zwischen den Variablen.

Es ist wichtig zu beachten, dass die Regression auf Beobachtungen beschränkt ist, bei denen keine fehlenden Werte in der Variable mit den wenigsten fehlenden Daten vorhanden sind. **Wir nehmen also die Variable mit den wenigsten Fehlern und wir entfernen die zufällig eingesetzten Werte aus Schritt 1 wieder. Unsere beobachtete Variable hat wieder Missing Data.** Die anderen Variablen werden jetzt zur Feature-Matrix, während die Variable mit den wenigsten Fehlern zum Target/ Response wird. Ich sage durch die Regressionsanalyse voraus, welcher Wert in die beobachtete Spalte eingesetzt werden muss. Ich muss dafür aber im ersten Schritt zufällige Zahlen einsetzen, weil ansonsten die Feature-Matrizen keine Werte für die Vorhersage haben würden (man muss die NA-Werte für die Regressionsanalyse entfernen). Dieser Prozess wird dann für alle Variablen im Datensatz durchgeführt. Wenn man dann eine Variable – nennen wir Sie x_1 – durch die Regression mit neuen Werten imputiert hat, macht man mit der nächsten Variable x_2 weiter. Für x_2 macht man wieder eine Regressionsanalyse, wobei man diese auf alle anderen Variablen regressiert – einschließlich der imputierten Werte aus x_1 . Man nutzt bei x_2 aber nur die Beobachtungen, die nicht fehlen. Das führt man für alle Variablen durch. Wenn man alle Variablen regressiert hat gegen alle anderen Variablen, dann hat man einen Zyklus.

Diese Zyklen von Imputation und Regression werden iterativ wiederholt. Jede Iteration erzeugt unterschiedliche Ergebnisse aufgrund der zufälligen Auswahl von Werten und der Verwendung der zuvor imputierten Daten in den Regressionen. Mit zunehmenden Iterationen konvergieren die Imputationen schließlich zu stabilen Werten.

Man macht es also so, dass man die vorhandenen Daten jeweils nutzt, um die fehlenden Daten zu modellieren. Dann werden in mehreren Iterationen auch die imputierten Daten genutzt, um die anderen Variablen vorherzusagen.

Multiple Imputation for Nonresponse in Surveys

JOHN WILEY & SONS
New York • Chichester • Brisbane • Toronto • Singapore

DONALD B. RUBIN

Department of Statistics
Harvard University

$$Var_{within} = \frac{\sum_{i=1}^M SE_i^2}{M}$$

$$Var_{between} = \frac{\sum_{i=1}^M (\beta_i - \bar{\beta})^2}{M - 1}$$

$$Var_{total} = Var_{within} + Var_{between} + \frac{Var_{between}}{M}$$

Im dritten Schritt, bei dem die Daten aus den verschiedenen imputierten Datensätzen zusammengeführt werden, basiert dies auf den sogenannten **Rubin-Regeln**. In einfachen Worten bedeutet das, dass Sie den Durchschnitt des gewünschten Parameters über alle imputierten Datensätze ziehen. Die Varianz des Parameters wird dann über die dritte Formel in der Abbildung berechnet. Die Varianz von Interesse ist die within Varianz. M ist die Anzahl der Imputation.

Thema 4: Outlier

Zum Schluss reden wir über das Thema Outlier – Ausreißer in den Daten. Was kann man machen, was sollte man machen. Wie werden Outlier definiert? Das wird hier behandelt.

Table 1. Outlier Definitions Based on a Review of Methodological and Substantive Organizational Science Sources.

1. Single construct outliers	Data values that are unusually large or small compared to the other values of the same construct. These points typically fall in the tails of a data distribution.	8. Influential meta-analysis sample size outliers	In the context of a meta-analysis, these are single construct outliers in terms of their sample size compared to the other studies' sample sizes.
2. Error outliers	Data points that lie at a distance from other data points because they are the result of inaccuracies. More specifically, error outliers include outlying observations that are caused by not being part of the population of interest (i.e., an error in the sampling procedure), lying outside the possible range of values, errors in observation, errors in recording, errors in preparing data, errors in computation, errors in coding, or errors in data manipulation.	9. Influential meta-analysis effect and sample size outliers	Primary-level studies that, via a combination of unusually large or small effect sizes and unusually large sample sizes, exert a large influence on the meta-analytic results.
3. Interesting outliers	Accurate (i.e., nonerror) data points that lie at a distance from other data points and may contain valuable or unexpected knowledge.	10. Cluster analysis outliers	Outliers that exist as a result of conducting cluster analysis.
4. Discrepancy outliers	Data points with large residual values, with possibly (but not necessarily) large influence on model fit and/or parameter estimates.	11. Influential time series additive outlier	An observation that markedly deviates from surrounding others in a time series analysis. A time series additive outlier may exist in isolation, such that connecting the surrounding data points and the outlier with a continuous line would yield a spike shape at the time point where the outlier exists. Alternatively, a group of time series additive outliers may exist as a patch within a range of time points.
5. Model fit outlier	An influential outlier whose presence influences the fit of the model.	12. Influential time series innovation outlier	An observation that not only has a large absolute value compared to surrounding others in a time series analysis, but also affects the values of subsequent observations in unequal amounts.
6. Prediction outlier	An influential outlier whose presence affects the parameter estimates of the model.	13. Influential level shift outliers	A data point causing an abrupt and permanent step change (i.e., jump) in the values of subsequent observations in a series.
7. Influential meta-analysis effect size outlier	A data point that is unusually large or small compared to others in a meta-analytic database, specifically regarding the size of the effect or relationship.	14. Influential temporary changes outliers	A data point causing an abrupt step change (i.e., jump) in the values of subsequent observations in a series, but this differs from a level shift outlier in that this change eventually dies out with time. That is, the step change is not permanent.

Bildquelle: Aguinis et al. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers Organizational Research Methods 16(2) 270-301

Frage: Nennen Sie 3 Arten von Outliern und definieren Sie diese?

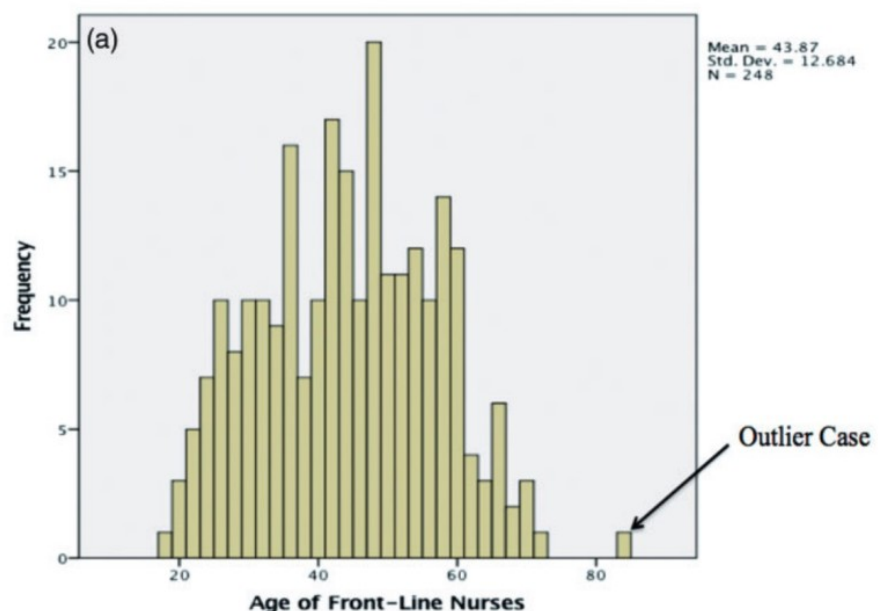
Kommen wir zuerst zur Definition von Outliern. Was verstehen wir unter Outlier? Wir werden uns primär mit den ersten 3 Arten von Outliern beschäftigen. Es gibt die folgenden 3:

1. Single Construct Outliers: In diesem Kontext beziehen sich diese auf Werte, die sich signifikant von der üblichen Datenverteilung abheben, indem sie entweder erheblich größer oder kleiner sind. Diese Ausreißer tendieren dazu, an den Extremen der Verteilung zu liegen und sind durch zufällige Ziehung zu erwarten. (Die Annahme ist, dass diese Ausreißer durch Zufall in die Datensammlung gelangt sind, anstatt ein tatsächliches Muster oder eine systematische Abweichung darzustellen. Mit anderen Worten, sie sind ungewöhnlich, aber es gibt keine klare Erklärung für ihr Auftreten, außer dass sie durch zufällige Stichprobenvariabilität verursacht wurden). Angenommen, die Forscher haben die Körpergröße von 1.000 Menschen in einer bestimmten Region gemessen und die Ergebnisse in einem Datensatz erfasst. Die meisten der gemessenen Körpergrößen liegen innerhalb eines bestimmten Bereichs, der der allgemeinen Verteilung in dieser Region entspricht. Allerdings gibt es auch einige Ausreißer in den Daten, bei denen die Körpergröße signifikant von dieser Verteilung abweicht. Zum Beispiel könnte es einen Datensatzpunkt geben, der eine extrem geringe Körpergröße aufweist, während ein anderer Punkt eine außergewöhnlich große Körpergröße darstellt. In diesem Fall könnten diese Ausreißer als "Single Construct Outliers" betrachtet werden, da sie an den Extremen der Körpergrößenverteilung liegen und aufgrund der zufälligen Variation in einer Stichprobe erwartet werden. Die Annahme ist, dass diese ungewöhnlichen Körpergrößen nicht auf spezifische genetische Merkmale oder Umweltfaktoren zurückzuführen sind, sondern das Ergebnis von zufälligen Schwankungen in der Stichprobenziehung sind.

2. Error Outliers: Das sind die Outlier, die durch Fehler entstanden sind – also durch Ungenauigkeiten in der Messung. Angenommen, ein Unternehmen produziert Flaschen, und die Füllmenge wird regelmäßig überprüft, um sicherzustellen, dass sie den vorgeschriebenen Standards entspricht. In einem bestimmten Durchgang werden die meisten Flaschen mit einem konsistenten Füllvolumen hergestellt. Allerdings treten einige Flaschen mit einem extremen Füllvolumen auf, das weit über oder unter dem Zielwert liegt. Diese extremen Füllmengen können auf Fehler oder Ungenauigkeiten im Füllprozess hinweisen, wie beispielsweise eine fehlerhafte Maschineneinstellung oder ein Versagen des Füllmechanismus. Die "Error

Outliers" in diesem Fall sind Ausreißer, die durch Fehler oder Ungenauigkeiten im Produktionsprozess entstehen. Sie sind nicht das Ergebnis einer gezielten Variation oder eines interessanten Phänomens, sondern vielmehr unerwünschte Abweichungen, die korrigiert werden müssen, um die Produktqualität sicherzustellen.

3. Interesting Outliers: Bei dieser Art von Ausreißern handelt es sich um Werte, die signifikant von den übrigen Datenpunkten abweichen, jedoch keine Fehler darstellen. Sie können auch nicht auf einfache Zufallsvariationen zurückgeführt werden. Stattdessen handelt es sich um Ausreißer, die auf ein tiefgreifendes Muster oder eine systematische Abweichung im Datensatz hinweisen. Diese Ausreißer liefern wertvolle Einblicke und deuten auf ein wichtiges Phänomen innerhalb der Daten hin. Angenommen, ein Team von Wissenschaftlern untersucht den Blutdruck von Patienten in einer klinischen Studie. Die meisten Patienten haben Blutdruckwerte im normalen Bereich, aber es gibt auch einige, bei denen der Blutdruck extrem hoch ist. Diese extremen Blutdruckwerte könnten als bedeutsame Ausreißer betrachtet werden. Anstatt auf zufällige Messfehler oder ungewöhnliche Bedingungen zurückzuführen zu sein, könnten diese Ausreißer tatsächlich auf eine spezifische Behandlung oder eine genetische Veranlagung hinweisen, die den Blutdruck beeinflusst.



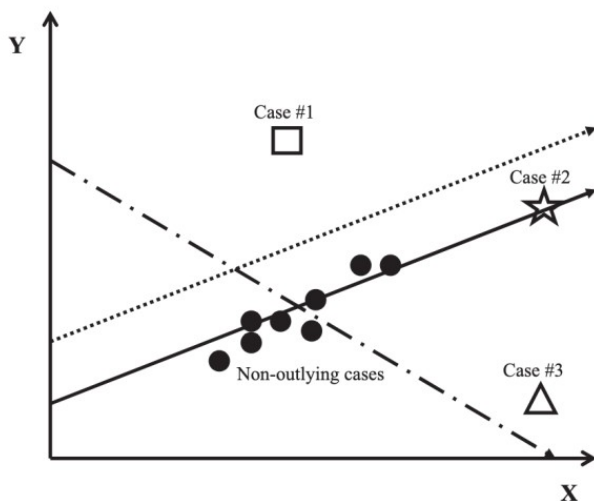
Mowbray F et al. Univariate Outliers: A Conceptual Overview for the Nurse Researcher Canadian Journal of Nursing Research 2019, Vol. 51(1)

Frage: Was sind univariate Outlier? Was versteht man darunter?

Hier sehen Sie einen Univariaten Outlier. In der Abbildung sehen Sie ein Histogramm. Auf der X-Achse ist das Alter von Front-Line Nurses dargestellt und auf der Y-Achse sehen wir, wie häufig Sie vorkommen. Wir sehen eine Person, die mit über 80 Jahren deutlich älter ist als die anderen Personen. Es handelt sich um einen Univariaten Outlier.

Dieser Outlier ist schon dadurch zu erkennen, indem man sich das Alter der Person anschaut. Man muss sich also nur die eine Variable anschauen.

Ein univariater Ausreißer (auch als "univariate Outlier" bezeichnet) ist ein Datenpunkt oder eine Beobachtung in einem Datensatz, der sich signifikant von den anderen Datenpunkten in einer einzigen Variablen oder Dimension unterscheidet. Dies bedeutet, dass der Ausreißer entweder deutlich größer oder kleiner ist als die anderen Werte in dieser speziellen Variable, ohne Berücksichtigung von anderen Variablen im Datensatz.

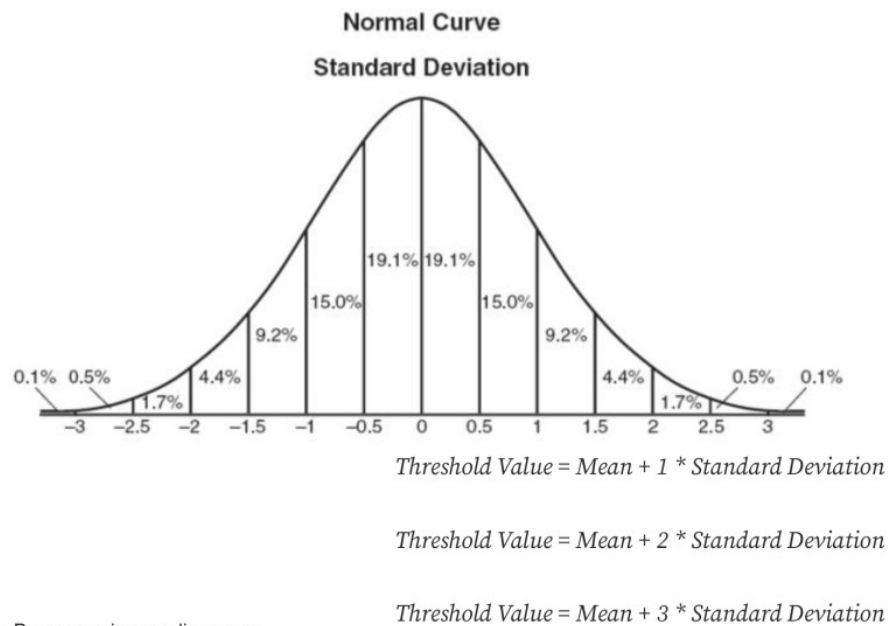


	Regression line	R^2	Slope	Intercept
With ● only	—————→	.73	.83	5.34
With ● and □→	.11	.83	9.13
With ● and ☆	—————→	.95	.83	5.34
With ● and △	- . - . - . →	.17	-.22	30.34

Frage: Was ist das Problem an Outliern?

Outlier können die Ergebnisse verzehren. Das ist das wichtigste Problem. Die Abbildung zeigt es beispielhaft. Wir sehen Datenpunkte in der Mitte des Graphen. Es wurden 4 Fälle untersucht und bei 3 von 4 Fällen wurden Ausreißer eingefügt. Es wurden dann beobachtet, wie sich die Regressionslinie bei einer Regressionsanalyse verhält, abhängig vom eingefügten Ausreißer. Ohne Ausreißer kriegt man ein R^2 von 0,73, eine Steigung von 0,83 und ein Intercept = 5,34. In Case 1, wo ich den Ausreißer oben einführe, bleibt die Steigung gleich, aber R^2 sinkt und der Schnittpunkt mit der Y-Achse steigt. In Fall 2 hat man den Ausreißer auf dieselbe Höhe wie die restlichen Datenpunkte gepackt, entlang der Regressionslinie. Hier ändert sich nur R^2 . Dadurch, dass ich einen Punkt auf die Regressionslinie hinzufüge, glaubt das Modell, es habe eine viel bessere Varianzaufklärung. Den krassesten Fall beobachten wir, wenn wir einen Ausreißer unter den anderen Punkt hinzufügt. Dann ändern sich alle Werte. Ich habe damit die Richtung der Regressionslinie verändert und der Zusammenhang zwischen den 2 beobachteten Variablen wird nicht mehr richtig dargestellt.

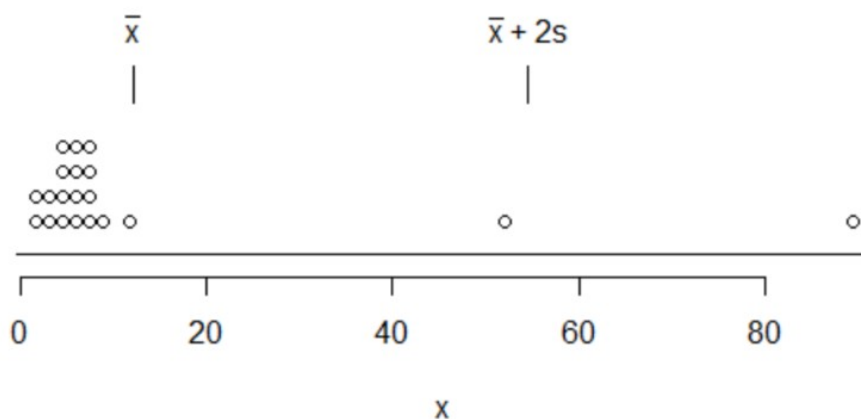
Outlier können den Zusammenhang (Regressionslinie) gleich lassen, aber den Effekt verstärken oder Sie können den Zusammenhang völlig verändern.



Colak U. Machine Learning — Data Preprocessing. medium.com

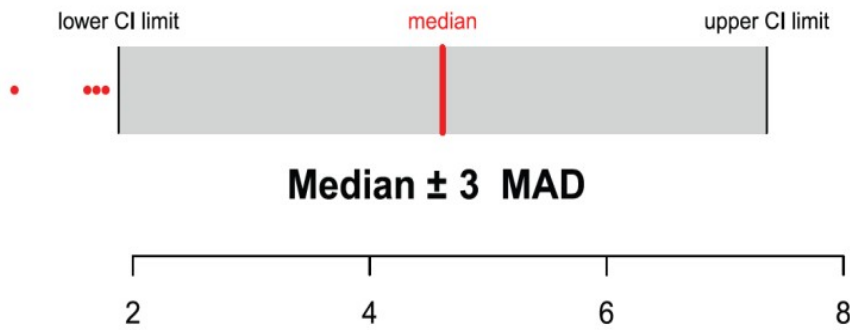
Frage: Wie kann man univariate Outlier systematisch identifizieren?

Man muss univariate Ausreißer nicht nur graphisch identifizieren. Man kann sie auch systematisch erkennen, indem man die Annahme macht, dass die Daten einer Normalverteilung folgen. Dann identifiziert man einen univariaten Ausreißer dadurch, dass er sich um eine bestimmte Anzahl von Standardabweichungen vom Mittelwert entfernt. Wenn eine Observation **mehr als 3 Standardabweichungen (andere Thresholds möglich)** vom Mittelwert entfernt liegt, dann behandeln wir es wie einen Outlier. Dann handelt es sich nämlich um einen Punkt, der mit einer Wahrscheinlichkeit von 0.1 % vorkommen kann.



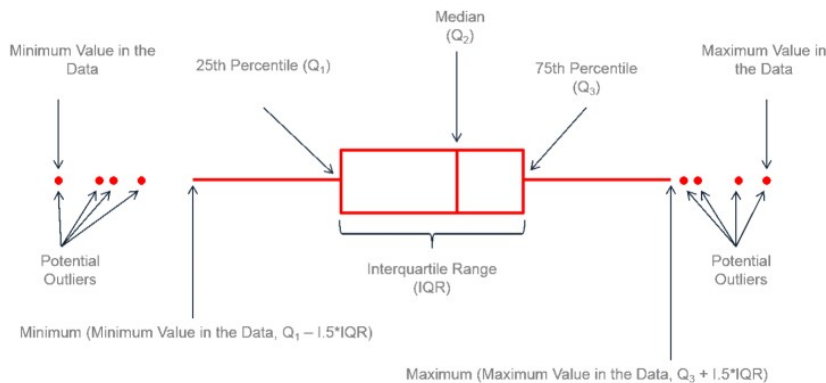
Frage: Was ist das Problem an dem Ansatz mit der Identifizierung der univariaten Outlier anhand des means und Standardabweichung (SD)?

Das Ganze wird schwerer gemacht, weil Outlier einen Einfluss auf den Mittelwert haben, als auch auf die Standardabweichung. In der Abbildung sehen wir, dass der extreme Werte 90 den Mittelwert und die SD verzerrt. Die Methode, wonach wir Observationen mit einem bestimmten Threshold entfernt vom Mean als Outlier definieren, funktioniert also nicht mehr richtig. Varianz wird quadriert, bevor für SD die Wurzel gezogen wird. Dadurch haben Outlier einen großen Einfluss auf die Standardabweichung.



$$\text{MAD} = \text{Median}(|X_i - \tilde{X}|)$$

Leys, C., et al. (2019). How to Classify, Detect, and Manage Univariate and Multivariate Outliers, With Emphasis on Pre-Registration. International Review of Social Psychology, 32(1):



$$\text{IQR} = 1.5 * (Q_3 - Q_1)$$

$$\text{Lower Threshold} = Q_1 - \text{IQR}$$

$$\text{Upper Threshold} = Q_3 + \text{IQR}$$

Frage: Welche Lösung gibt es für das Problem?

Man nutzt statt des Means und der SD eine Messgröße verwenden, die robust gegenüber Outliern ist: **Median und Quantile**. Man kann Outlier über den MAD bestimmen.

Die Erkennung von Ausreißern mithilfe des 3-MAD (3-Median Absolute Deviation) und des Medians ist ein robustes Verfahren. Um Ausreißer zu identifizieren, geht man wie folgt vor:

Berechnen Sie den Median (M) Ihrer Daten. Der Median ist der Wert, der in der Mitte Ihrer sortierten Daten liegt.

Berechnen Sie zuerst den Median Ihrer Daten. Der Median ist der Wert, der in der Mitte Ihrer sortierten Daten liegt. Wenn Ihre Daten noch nicht sortiert sind, ordnen Sie sie zuerst in aufsteigender oder absteigender Reihenfolge an.

Berechnen Sie den Betrag der Differenz zwischen jedem Datenpunkt und dem Median. Das bedeutet, für jeden Datenpunkt subtrahieren Sie den Median von diesem Datenpunkt und nehmen den Betrag (d.h., machen Sie negative Werte positiv).

Sammeln Sie alle berechneten absoluten Abweichungen.

Berechnen Sie den Median dieser absoluten Abweichungen, um den MAD zu erhalten.

Multiplizieren Sie den MAD mit 3, um eine Toleranzschwelle festzulegen. Dieser Schwellenwert entspricht dem 3-fachen MAD.

Identifizieren Sie Ausreißer, indem Sie jeden Datenpunkt überprüfen, um festzustellen, ob er mehr als das 3-fache des MAD vom Median entfernt ist. Wenn ein Datenpunkt diese Schwelle überschreitet, gilt er als Ausreißer.

Alternativ kann man univariate Ausreißer über einen Boxplot identifizieren.

Sie können Ausreißer mithilfe von Boxplots (auch als Box-Whisker-Plots bezeichnet) erkennen, indem Sie die folgenden Schritte ausführen:

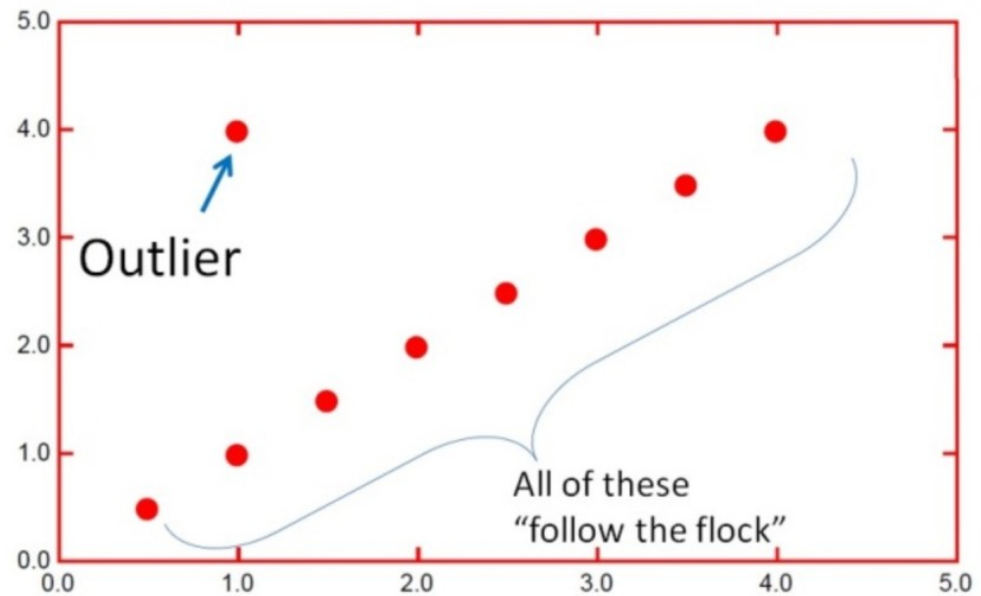
Zeichnen Sie den Boxplot: Zunächst erstellen Sie einen Boxplot für Ihre Daten. Ein Boxplot zeigt die Verteilung der Daten, indem es Quartile, Median und mögliche Ausreißer visualisiert.

Identifizieren Sie die Quartile: Im Boxplot gibt es eine Box, die den Interquartilbereich (IQR) darstellt. Dieser Bereich enthält 50 % der Daten. Die untere Kante der Box entspricht dem 25. Perzentil (Q1), während die obere Kante dem 75. Perzentil (Q3) entspricht.

Berechnen Sie den IQR: Subtrahieren Sie Q1 von Q3, um den Interquartilbereich (IQR) zu erhalten. Dieser IQR repräsentiert den mittleren 50 % der Daten.

Bestimmen Sie die Grenzen für Ausreißer: Multiplizieren Sie den IQR mit 1,5. Die untere Grenze für potenzielle Ausreißer wird durch Subtrahieren von $1,5 * \text{IQR}$ von Q1 berechnet, während die obere Grenze durch Hinzufügen von $1,5 * \text{IQR}$ zu Q3 berechnet wird.

Identifizieren Sie Ausreißer: Datenpunkte, die unterhalb der unteren Grenze oder oberhalb der oberen Grenze liegen, werden als potenzielle Ausreißer betrachtet.



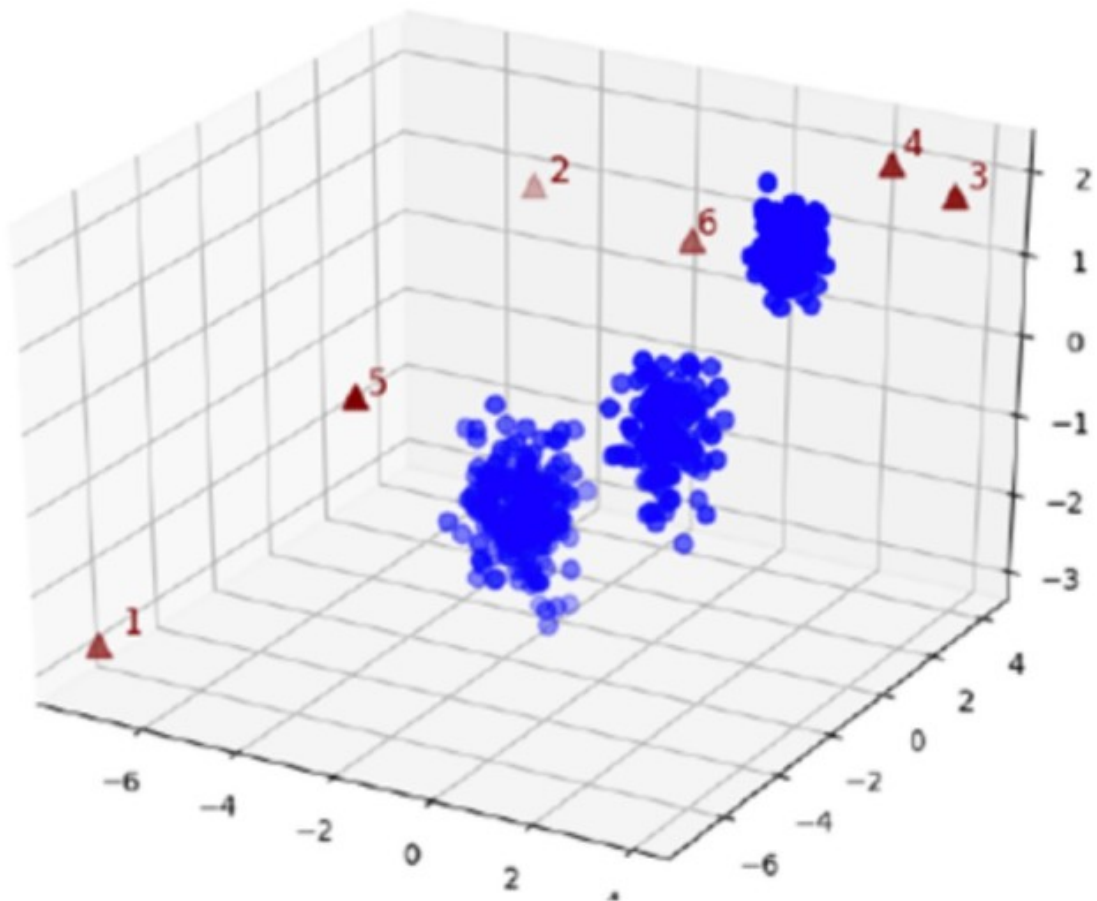
vatsalparsaniya.github.io/

Frage: Was sind bivariate Outlier?

Ein bivariater Ausreißer ist nicht offensichtlich, wenn man nur eine einzelne Variable betrachtet. Erst wenn man diesen in Beziehung zu einer zweiten Variable setzt, wird er als Ausreißer erkennbar.

Ein bivariater Ausreißer (auch als "bivariater Outlier" bezeichnet) ist eine Beobachtung oder ein Datenpunkt in einem Datensatz, der in Bezug auf zwei miteinander verknüpfte Variablen oder Dimensionen ungewöhnlich ist. Anders ausgedrückt: Dieser Ausreißer wird erst dann als abnorm erkannt, wenn man sowohl die Werte der ersten Variable als auch die Werte der zweiten Variable in Betracht zieht und die Beziehung zwischen ihnen berücksichtigt.

Man muss sich die kombinierte Verteilung zweier Variablen anschauen.



Frage: Was sind multivariate Outlier?

Es existieren auch multivariate Outlier. Multivariate Ausreißer (auch als "multivariate Outliers" bezeichnet) sind Datenpunkte oder Beobachtungen in einem Datensatz, die in Bezug auf mehrere miteinander verknüpfte Variablen oder Dimensionen ungewöhnlich sind. Im Gegensatz zu univariaten Ausreißern, die sich auf eine einzelne Variable beziehen, und bivariaten Ausreißern, die sich auf zwei Variablen beziehen, sind multivariate Ausreißer das Ergebnis von Abweichungen, die sich in einem multidimensionalen Raum zeigen.

Ab einem bestimmten Punkt hat man zu viele Dimensionen und man kann die Outlier nicht mehr graphisch darstellen. Deshalb gibt es dann viele statistische Werkzeuge, um Outlier zu erkennen.

Multiple-construct (i.e., “distance”) techniques

6. Scatter plot	A plot of the values of two variables, with one variable on the x-axis (usually the independent variable) and the other variable on the y-axis (usually the dependent variable). A potential outlier can be identified by a data point lying far away from the centroid of data.	13. Euclidean distance	Length of the line segment between two specified points in a one-, two-, or n-dimensional space. A large Euclidean distance between two data points may mean that one of the two data points is an outlier.
7. q-q plot	A plot (q stands for quantile) that compares two probability distributions by charting their quantiles against each other. A nonlinear trend indicates the possible presence of outlier(s).	14. Mahalanobis distance	Similar to Euclidean distance, but different in that Mahalanobis distance is the length of the line segment between a data point and the centroid (instead of another observation) of the remaining cases, where the centroid is the point created at the intersection of the means of all the predictor variables. A large Mahalanobis distance may mean that the corresponding observation is an outlier.
8. p-p plot	A plot (p stands for probability) that assesses the degree of similarity of two data sets (usually the observed and expected) by plotting their two cumulative distribution functions against each other. A nonlinear trend indicates the possible presence of outlier(s).	15. K-clustering (with or without modified hat matrix) or other similar cluster analysis techniques	Yields different candidate subsets that then have to be evaluated by one or more multiple-case diagnostics.
9. Standardized residual	A residual value that is calculated by dividing the <i>i</i> th observation's residual value by a standard deviation term. Observations with high standardized residual values are likely to be outliers. However, an observation's standardized residual value does not measure an observation's outlyingness on the predictor variables.	16. 2- or 3-dimensional plots of the original and the principal component variables	A two- or three-dimensional plot of variables produced as a result of a principal component analysis. An isolated data point denotes a potential outlier.
10. Studentized residual	A residual value that measures both the outlyingness of the observation in terms of its standardized residual value (i.e., one type of distance) and the outlyingness of the observation on the predictor variables (i.e., another type of distance), such that a data point that is outlying in terms of both types of distance would have a studentized residual value that is greater than its standardized residual value. Observations with high studentized residual values are likely to be outliers.	17. Autocorrelation function plot	A plot created by computing autocorrelations for data values at varying time lags. Potential outliers can be identified by data points that lie at a distance from other data points.
11. Standardized deleted residual	A residual value that is identical to a standardized residual, except that the predicted value for the focal observation is calculated without the observation itself. This exclusion prevents the focal observation from deflating the residual value and inflating the standard deviation term, where such deflation and inflation mask the existence of any outlyingness of the observation. Observations with high standardized deleted residual values are likely to be outliers.	18. Time plot	A plot of the relationship between a certain variable and time. Potential outliers can be identified by data points that lie at a distance from other data points.
12. Studentized deleted residual (i.e., externally studentized residual, jackknife residual)	A residual value that is identical to a studentized residual, except that the predicted value for the focal observation is calculated without the observation itself. This exclusion prevents the focal observation from deflating the residual value and inflating the standard deviation term, where such deflation and inflation mask the existence of any outlyingness of the observation. Observations with high studentized deleted residual values are likely to be outliers.	19. Extreme studentized deviate (i.e., Grubbs method)	Difference between a variable's mean and query value, divided by a standard deviation value.
		20. Hosmer and Lemeshow goodness-of-fit test	A Pearson chi-square statistic from a table of observed and expected (i.e., implied) frequencies.
		21. Leverage values	Also known as the diagonal elements of the hat matrix, leverage values measure the extent to which observations are outliers in the space of predictors.
		22. Centered leverage values	A centered index of leverage values. Certain statistical packages (e.g., SPSS) report centered leverage values instead of regular leverage values.
		23. Deletion standardized multivariate residual	A standardized residual term in the context of multilevel modeling. This allows for an assessment of the effect that a higher level outlier has on model fit. If an outlier is found at the higher level, lower level units should be investigated.
<hr/>			
Influence techniques			
24. Cook's D_i	Assesses the influence that a data point <i>i</i> has on all regression coefficients as a whole.	34. Nearest neighbor techniques	Calculation of the closest value to the query value using various types of distance metrics such as Euclidean or Mahalanobis distance. Techniques include K-nearest neighbor, optimized nearest neighbor, efficient Type 1 nearest neighbor, Type 2 nearest neighbor, nearest neighbor with reduced features, dragon method, PAM (partitioning around medoids), CLARANS (clustering large applications based on randomized search), and graph connectivity method.
25. Modified Cook's D_i	Similar to Cook's D_i , but it uses standardized deleted residuals rather than standardized residuals.	35. Nonparametric methods	Consist of fitting a smoothed curve without making any constraining assumptions about the data. A lack of a linear trend in the relationship signals the presence of outliers.
26. Generalized Cook's D_i	Similar to Cook's D_i , but applied to structural equation modeling to assess the influence that a data point has on the parameter estimates.	36. Parametric methods	Unlike nonparametric methods, parametric methods make certain assumptions about the nature of the data. One such assumption is that the data come from a particular type of probability distribution (e.g., normal distribution). Outliers are identified by these techniques as data points that fall outside the expectations about the nature of the data. Parametric methods include convex peeling, ellipsoidal peeling, iterative deletion, iterative trimming, depth trimming, least median of squares, least trimmed squares, and M-estimation.
27. Difference in fits, standardized (DFFITS)	Just like Cook's D_i , this technique also assesses the influence that a data point <i>i</i> has on all regression coefficients as a whole. A large difference between the two techniques is that they produce information that exists on different scales.	37. Semiparametric methods	These methods combine the speed and complexity of parametric methods with the flexibility of nonparametric methods to investigate local clusters or kernels rather than a single global distribution model. Outliers are identified as lying in regions of low density.
28. Difference in beta, standardized (DFBETAS _j)	Indicates whether the inclusion of a case <i>i</i> leads to an increase or decrease in a single regression coefficient <i>j</i> (i.e., a slope or intercept).	38. Iterative outlier identification procedure	In a sequence of steps, this procedure allows for the estimation of the residual standard deviation to identify data points that are sensitive to the estimation procedure that is used when conducting a time series analysis. Such data points are subsequently identified as outliers.
29. Chi-squared difference test	This method allows a researcher conducting SEM to assess the difference in the model fit between two models, one with the outlier included and the other without the outlier.	39. Independent component analysis	A computation method used to separate independent components by maximizing the statistical independence among them. The separate independent components, when found in a time series analysis, are identified as outliers.
30. Single parameter influence	Similar to DFBETAS _j , this identification technique is used in SEM to assess the effect of an outlier on a specific parameter estimate, as opposed to the overall influence of an outlier on all parameter estimates in the model.		
<hr/>			
31. Average squared deviation technique	When conducting multilevel modeling, this method, a direct analog of Cook's D_i , investigates the effect that each group has on the fixed and/or random parameters, allowing for the identification of higher level prediction outliers. If an outlier is found at the higher level, lower level units should be investigated.		
32. Sample-adjusted meta-analytic deviancy (SAMD)	In meta-analysis, this test statistic takes the difference between the value of each primary-level effect size estimate and the mean sample-weighted coefficient computed without that effect size in the analysis, and then adjusts the difference value based on the sample size of the primary-level study. Outliers are identified by their extreme SAMD values.		
33. Conduct analysis with and without outliers	This technique refers to conducting the statistical analysis with and without a particular data point. If results differ across the two analyses, the data point is identified as an outlier.		

Frage: Wie kann man Multivariate Outlier detektieren? Welche 2 Gruppen an Methoden gibt es und was messen diese?

Es gibt sehr viele verschiedene Möglichkeiten und die Abbildung gibt einige Beispiele. Es gibt Methoden, die nur die Distanz identifizieren – wie weit sind die Outlier entfernt von den typischen Werten. Dafür kann man verschiedene Plots nutzen wie z.B. pp oder qq-Plots. Als zweite Gruppe gibt es die **Influence Techniken**. Diese Methoden haben die Gemeinsamkeit, dass die Größe des Outliers daran gemessen wird, wie stark der Einfluss auf die Ergebnisse des Modells ist. Wenn ein Outlier einen sehr großen Effekt auf zum Beispiel die Koeffizienten eines Regressionsmodells hat, dann wird er durch die Methoden als ein sehr starker Outlier erkannt.

> Man misst also den Einfluss/ **Influence** des Outliers oder die **Distanz** zu den herkömmlichen Punkten.

Beispiel: Cook's distance

- der jeweilige Datenpunkt wird aus dem Modell entfernt und die Regression neu berechnet
- fast zusammen, wie stark sich alle Werte im Regressionsmodell ändern, wenn die jeweilige Beobachtung entfernt wird
- Konsens scheint zu sein, dass $D > 1$ = Outlier

$$D_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{p \text{ MSE}},$$

- \hat{y}_j is the j th fitted response value.
- $\hat{y}_{j(i)}$ is the j th fitted response value, where the fit does not include observation i .
- MSE is the mean squared error.
- p is the number of coefficients in the regression model.

Cook, R. Dennis (February 1977). "Detection of Influential Observations in Linear Regression". *Technometrics*

Frage: Welche klassische Methode, die zu den Influence Methoden gehört, kennen Sie und wie funktioniert Sie?

Ein gängiges Beispiel für Influence-Methoden ist der **Cook's Distance**. Hierbei wird das bestehende Regressionsmodell mehrmals berechnet, wobei in jedem Durchlauf ein bestimmter Datenpunkt aus dem Modell entfernt wird. Die Cook's Distance gibt an, in welchem Maße sich alle Parameter und Vorhersagen des Regressionsmodells ändern, wenn dieser Punkt aus der Analyse ausgeschlossen wird.

Für jede dieser Berechnungen wird die Distanz gemessen, die angibt, wie stark die Vorhersagen und Parameter des Modells beeinflusst werden, wenn man den Punkt rauslässt. Man schaut sich dann für jede Berechnung an, wie groß die Distanz ist. Die Werte mit der größten Distanz sind Outlier.

Durch die Analyse der Cook's Distance kann man Ausreißer in den Daten erkennen, die das Modell stark beeinflussen und somit die Zuverlässigkeit der Regressionsanalyse gefährden könnten.



Error	<i>e.g., coding error</i>
Interesting	<i>e.g., moderator underlying a potentially interesting psychological process</i>
Random	<i>e.g., a very large value of a given distribution</i>

Frage: Was macht mit Outliern, nachdem man Sie gefunden hat? Welche Optionen hat man?

Zuerst sollte ich schauen, um welche Art von Outliern es sich handelt. Wenn es Fehler sind, dann kann ich versuchen Sie zu beheben. Wenn ich zum Beispiel einen BMI Wert von 199 habe im Datensatz, dann kann ich einfach durch die Daten gehen, mir die Größe und Gewicht des Patienten anschauen und mir einen neuen BMI-Wert berechnen lassen. Das wird aber nicht immer gehen, weil Daten oft nicht vorhanden sind.

Wenn es zufällige Outlier (**Single Construct Outliers**) sind, dann kann es Sinn machen, Sie zu behalten. Diese Punkte wurden nämlich per Zufall aus einer Stichprobe gezogen und sind keine Fehler. Vor allem wenn Sie viele Observationen haben (1 Million z.B.), dann werden Sie viele solcher Outlier bekommen.

Outlier zu entfernen oder anzupassen (Über Imputation wie bei Missing Data) macht Sinn, wenn man den Verdacht hat, dass die Outlier die Modellierung stark behindern oder wenn die Ergebnisse stark verzerrt werden. Wenn also die Erklärung über die Daten oder die Performance des Modells behindert wird.