Data Fallacies

Lennart Dobs (340472), Niklas Dunkel (340715)

July 24, 2025

Gliederung

- Definition Data Fallacies
- Survivorship Bias
- Base Rate Fallacy
- Simpson's Paradox
- Danger of Summary Metrics
- False Causality
- Quiz
- Fazit und Abschluss

Definition von Data Fallacies[1]

- Data Fallacies sind häufige Ursachen für Fehlinterpretationen von (statistischen) Daten
- Diese Fehler können verschiedene Ursprünge haben und werden deshalb mit verschiedenen Fallacy-Arten beschrieben
- Data Fallacies sind problematisch, da Fehlinterpretationen von Daten, je nach Anwendungsfall der Daten, weitreichende negative Folgen haben können.
- Data Fallacies können auch absichtlich zur Manipulation der Dateninterpretation angewendet werden.

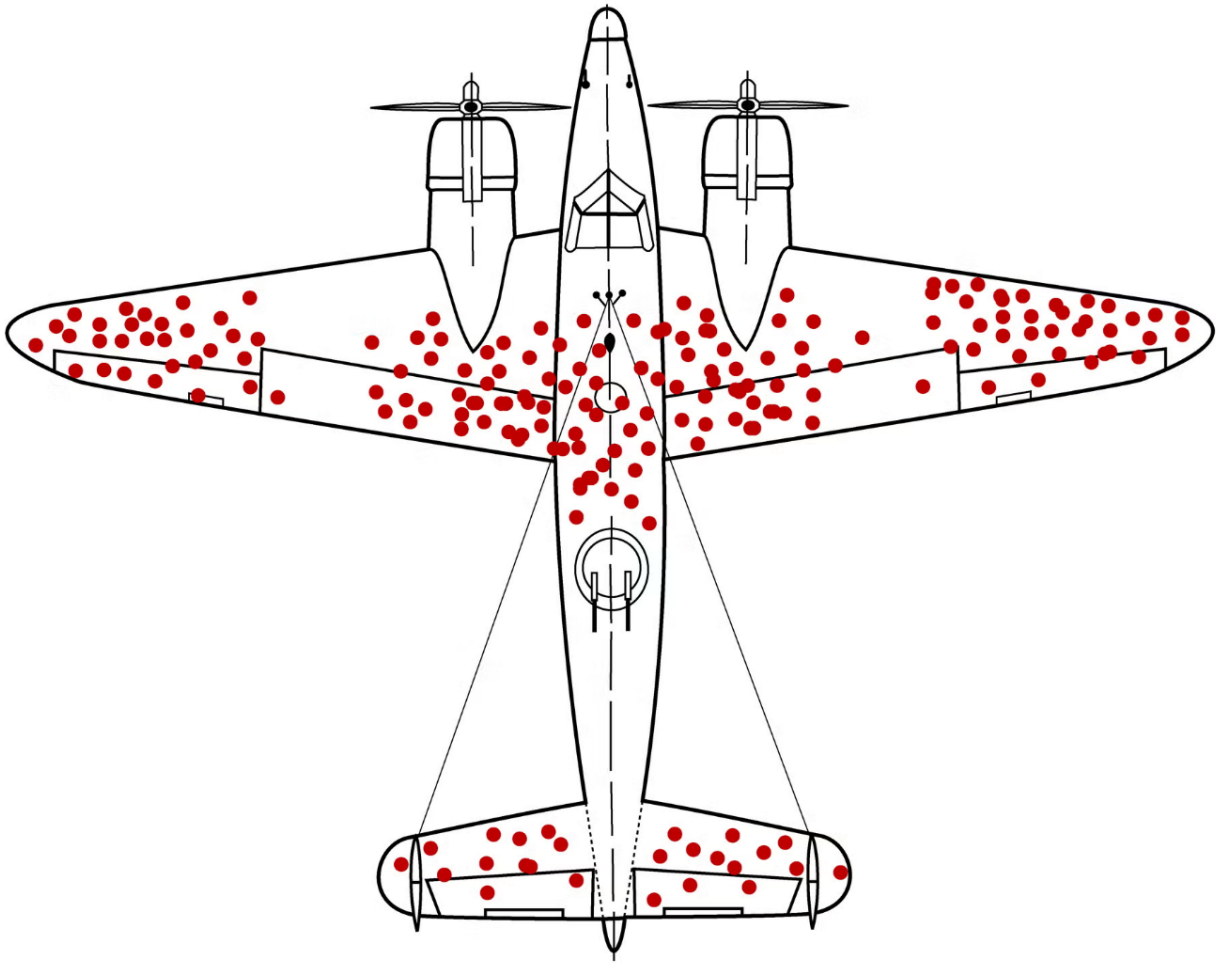
[1] Quelle: Ogbonnaya et al. (2019). S. 297.

Survivorship Bias[2]

- Basiert auf dem menschlichen Instinkt, von „Überlebenden“ bzw. „Gewinnern“ zu lernen
- Beschreibt die Verzerrung, die entsteht, wenn bei der Analyse von Daten nur die „Gewinner“ betrachtet werden und die Gesamtheit nicht korrekt abgebildet wird
- Fälschliche Schlussfolgerung: Eigenschaften oder gewissen Handlungen, die bei allen „Survivors“ vorliegen, haben zum „Überleben“ geführt
- Dass „Verlierer“ diese Eigenschaften möglicherweise auch gehabt haben, welche aber in den Daten nicht auftauchen, wird vernachlässigt
- Somit können falsche Schlussfolgerungen und Entscheidungen entstehen

[2] Quelle: Miller (2020).

Survivorship Bias - Beispiel



3

[3] Quelle: © Martin Grandjean (vector) McGeddon (picture). US Air Force (hit plot concept) / Survivorship Bias / CC BY-SA 4.0 (Ausschnitt). (o.J.).

Survivorship Bias[4]

- Weiteres Beispiel: Annahme, dass Dinge aus älteren Generationen langlebiger sind:
- Nur Dinge, die wenig genutzt wurden oder von besonderer Qualität waren, können noch vorgefunden werden – Gegenstände, die nicht „überlebt“ haben, können nicht gesehen werden
- Handlungsempfehlung: Vollständigkeit des Datensatzes hinterfragen

[4] Quelle: Elston (2021).

Survivorship Bias - Datensatz

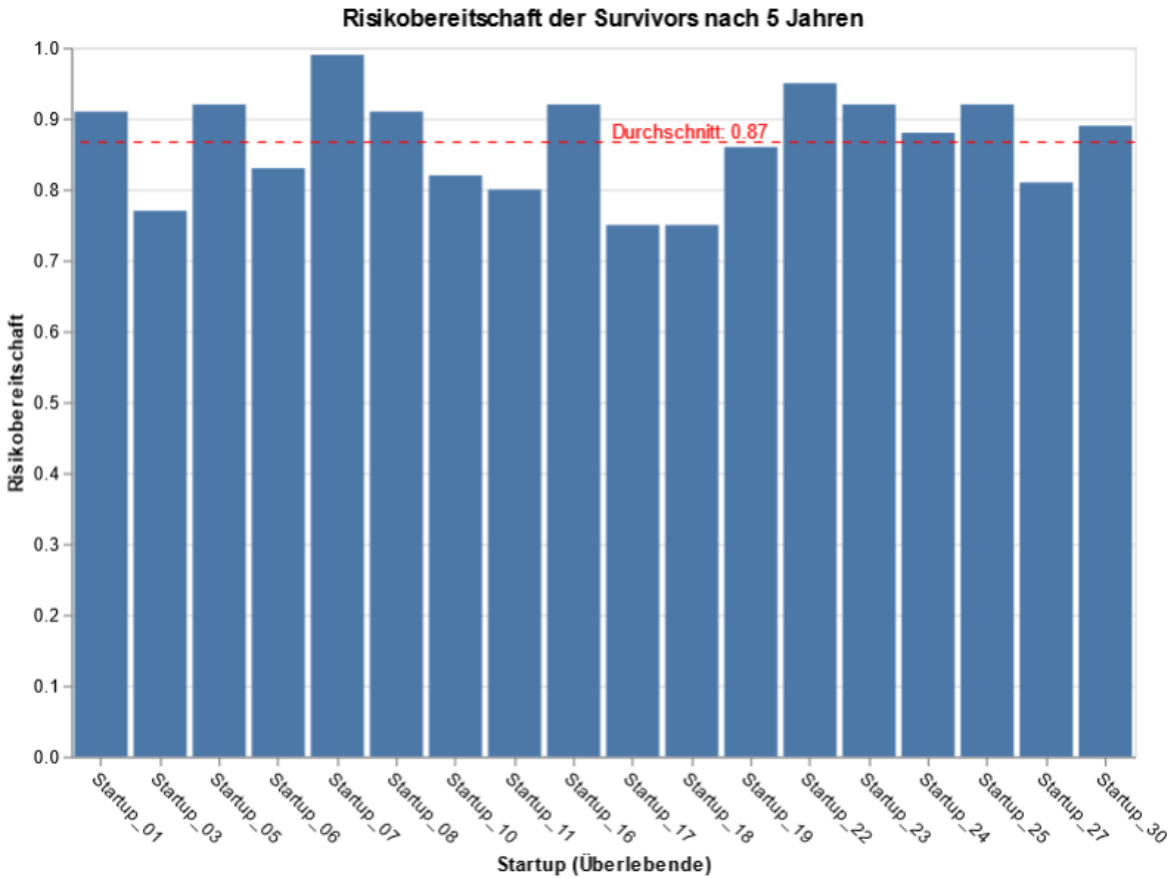
[3]:

	Startup	RiskTolerance	Survived5Years
0	Startup_01	0.91	True
1	Startup_02	0.79	False
2	Startup_03	0.77	True
3	Startup_04	0.87	False
4	Startup_05	0.92	True
5	Startup_06	0.83	True
6	Startup_07	0.99	True
7	Startup_08	0.91	True
8	Startup_09	0.84	False
9	Startup_10	0.82	True
10	Startup_11	0.80	True

[5]

[5] Quelle: Selbsterstellter Datensatz

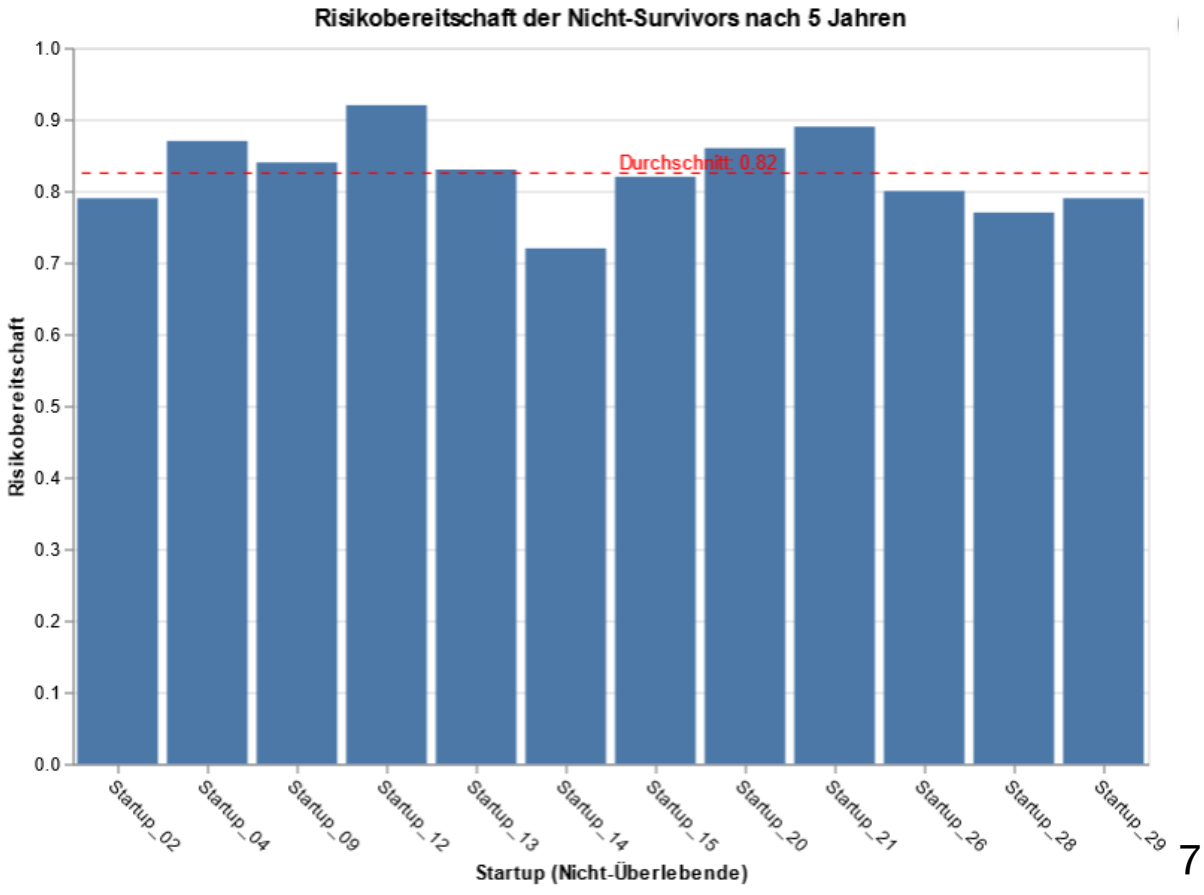
Survivorship Bias



[6]

[6] Quelle: Eigene Darstellung

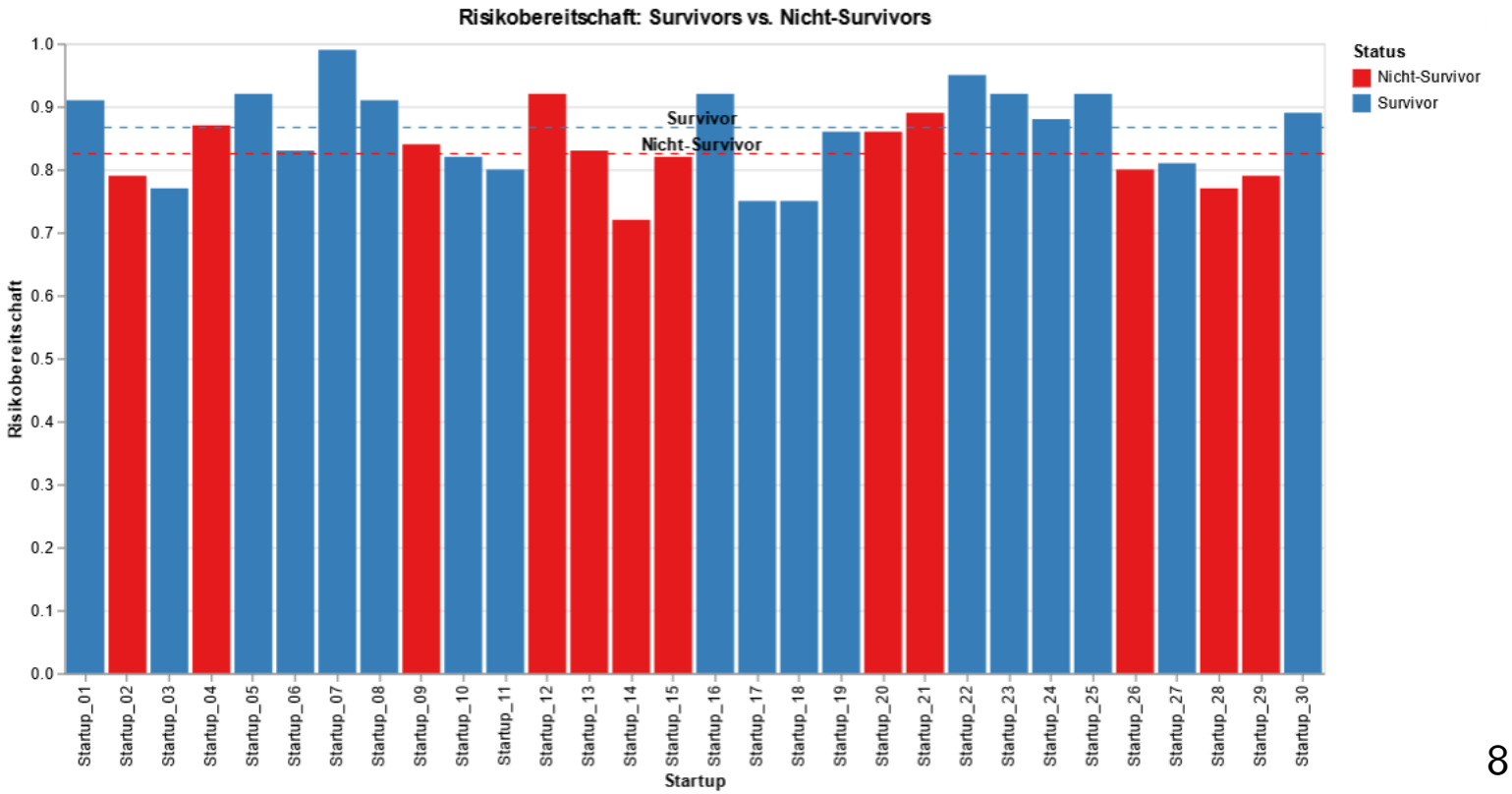
Survivorship Bias



[7]

[7] Quelle: Eigene Darstellung

Survivorship Bias



8

[8] Quelle: Eigene Darstellung

Base Rate Fallacy[9]

- Aufgabe:
- In einer Stadt gibt es nur Autos in zwei Farben: 85% der Autos sind blau und 15% sind grün.
- Eine Person beobachtet einen Autounfall mit Fahrerflucht und behauptet, dass das Auto grün war.
- Zeugen identifizieren die Farbe von Autos in 80% der Fälle korrekt.
- Wie groß ist die Wahrscheinlichkeit, dass das Auto tatsächlich grün war?

[9] Quelle: Flyvbjerg (2022).

Base Rate Fallacy[10]

- Beschreibt einen Fehler in der Beurteilung von Wahrscheinlichkeiten, der durch die Vernachlässigung von Basiswahrscheinlichkeiten entsteht, wenn eine Person spezifischere Informationen erhält
- Menschen neigen dazu, unterbewusst vorhandene Informationen nach ihrer subjektiven Relevanz zu sortieren
- Relevanz wird dabei meist dadurch bestimmt, wie spezifisch eine Information ist
- In Entscheidungen und Beurteilungen werden meistens nur die subjektiv relevantesten Informationen berücksichtigt

[10] Quelle: Bar-Hillel (1979). S. 211.

Base Rate Fallacy[11]

- Grund für diese Art der Entscheidungsfindung sind Heuristiken
- Eine Heuristik, die zu dem beschriebenen Ablauf führt, ist die Repräsentativitäts-Heuristik
- Bei der Representativitäts-Heuristik bewerten Menschen die Relevanz einer Information basierend auf ihrer Ähnlichkeit zu dem untersuchten Objekt und blenden somit häufig Basiswahrscheinlichkeiten aus
- Handlungsempfehlung: Hinterfragen, welche Basiswahrscheinlichkeit einer spezifischen Wahrscheinlichkeit zu Grunde liegt und diese einbinden oder WK mit Bayes-Regel berechnen

[11] Quelle: Kahneman & Tversky (1974). S. 1124.

Base Rate Fallacy - Datensatz

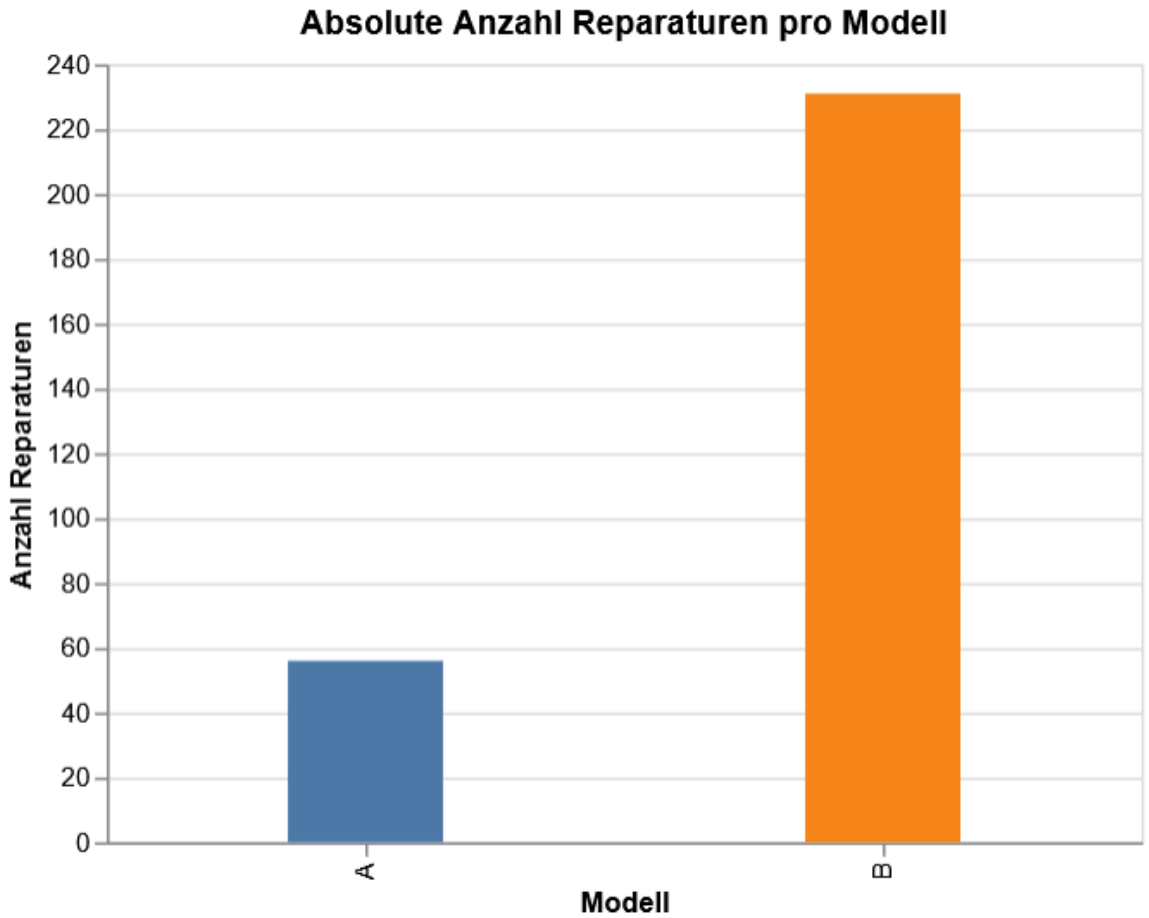
[5]:

	Model	Repaired
0	A	False
1	A	False
2	A	False
3	A	False
4	A	False
...	...	...
5995	B	False
5996	B	False
5997	B	False
5998	B	False
5999	B	False

6000 rows × 2 columns [12]

[12] Quelle: Selbsterstellter Datensatz

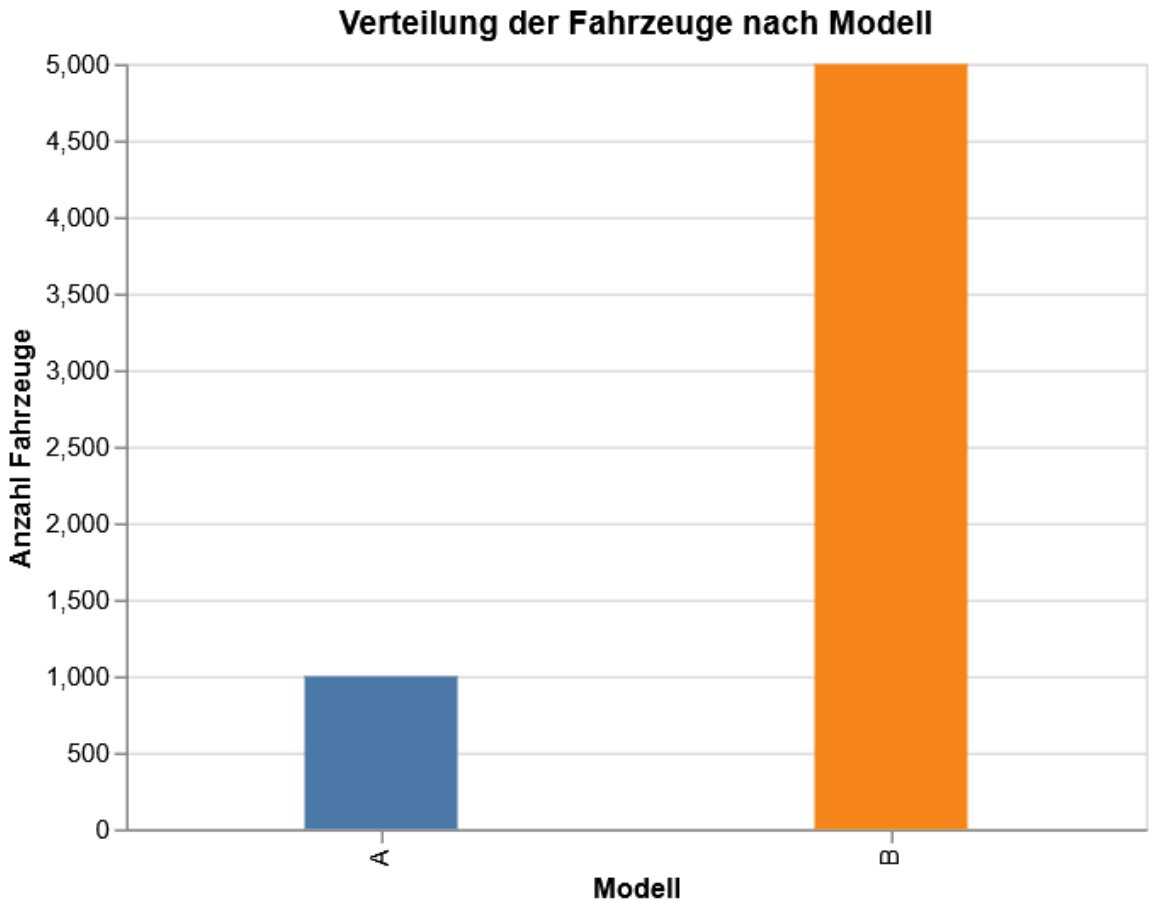
Base Rate Fallacy



13

[13] Quelle: Eigene Darstellung

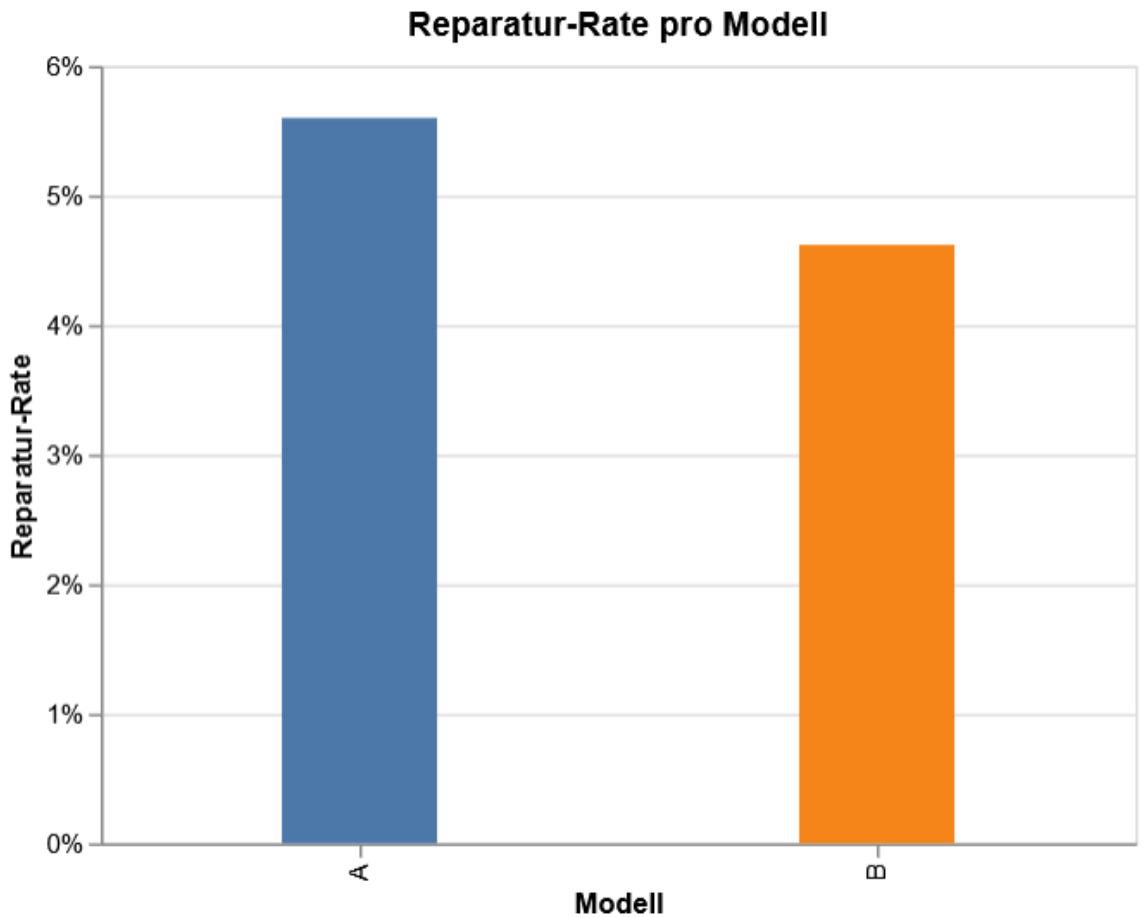
Base Rate Fallacy



14

[14] Quelle: Eigene Darstellung

Base Rate Fallacy



[15]

[15] Quelle: Eigene Darstellung

Simpson's Paradox[16]

- Beschreibt Verzerrung, bei der ein Zusammenhang zwischen 2 Variablen erkennbar ist, welcher sich umkehrt, wenn die Daten in Gruppen geteilt werden
- Dabei kehrt sich der Effekt in beiden Gruppen um oder ist nicht mehr zu erkennen
- Simpson Paradoxon entsteht dabei durch eine verdeckte Störvariable (confounding variable)
- Die Störvariable hängt dabei mit der Gruppenzugehörigkeit und dem Ergebnis zusammen.

[16] Quelle: Ameringer et al. (2009). S. 2.

Simpson's Paradox[17]

- Der wahre Effekt zwischen den beiden betrachteten Variablen wird damit erst aufgedeckt, wenn die Daten nach der Störvariable gruppiert werden
- Erforderlich für Simpson Paradoxon: Störvariable muss erkennbaren Effekt auf die Variablen haben und ungleichmäßig auf Gruppen verteilt sein
- Handlungsvorschlag: Korrelationen immer hinterfragen und nicht voreilig auf Kausalität schließen

[17] Quelle: Ameringer et al. (2009). S. 2.; Hintzman (1980) zitiert nach Ameringer et al. (2009). S. 2.; Hsu (1989) zitiert nach Ameringer et al. (2009). S. 2.

Simpson's Paradox

- Der wahre Effekt zwischen den beiden betrachteten Variablen wird damit erst aufgedeckt, wenn die Daten nach der Störvariable gruppiert werden[18]
- Erforderlich für Simpson Paradoxon: Störvariable muss erkennbaren Effekt auf die Variablen haben und ungleichmäßig auf Gruppen verteilt sein[19] [20]
- Handlungsvorschlag: Korrelationen immer hinterfragen und nicht voreilig auf Kausalität schließen

[18] Quelle: Ameringer et al. (2009). S. 2.

[19] Quelle: Hintzman (1980) zitiert nach Ameringer et al. (2009). S. 2.

[20] Quelle: Hsu (1989) zitiert nach Ameringer et al. (2009). S. 2.

Simpson's Paradox - Datensatz

[14]:

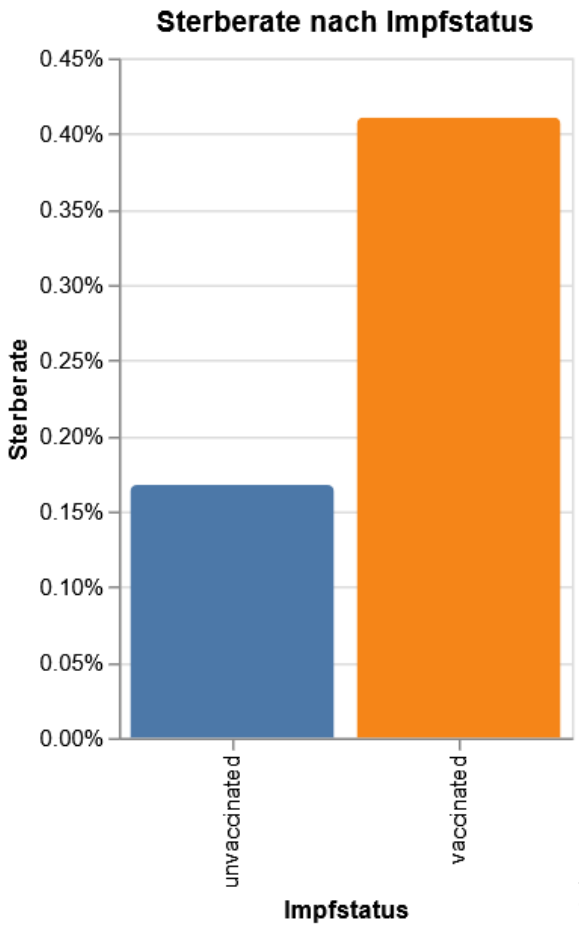
	age_group	vaccine_status	outcome
0	under 50	vaccinated	death
1	under 50	vaccinated	death
2	under 50	vaccinated	death
3	under 50	vaccinated	death
4	under 50	vaccinated	death
...	...	...	...
268161	50 +	unvaccinated	survived
268162	50 +	unvaccinated	survived
268163	50 +	unvaccinated	survived
268164	50 +	unvaccinated	survived
268165	50 +	unvaccinated	survived

268166 rows × 3 columns

21

[21] Quelle: `https://www.openintro.org/data/index.php?data=simpsons_paradox_covid`

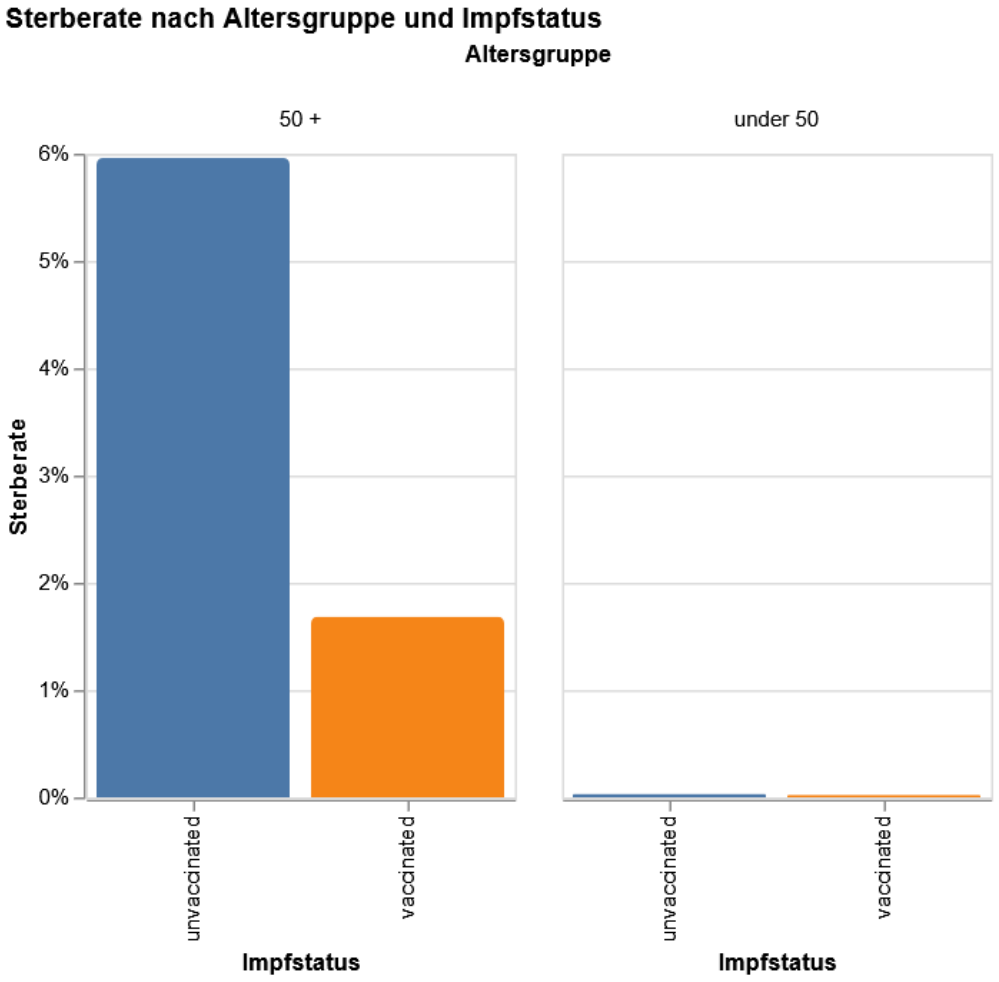
Simpson's Paradox



22

[22] Quelle: Eigene Darstellung

Simpson's Paradox



[23]

[23] Quelle: Eigene Darstellung

Simpson's Paradox

outcome	age_group	vaccine_status	death	survived	death_rate
0	50 +	unvaccinated	205	3235	0.059593
1	50 +	vaccinated	460	26847	0.016845
2	under 50	unvaccinated	48	147564	0.000325
3	under 50	vaccinated	21	89786	0.000234

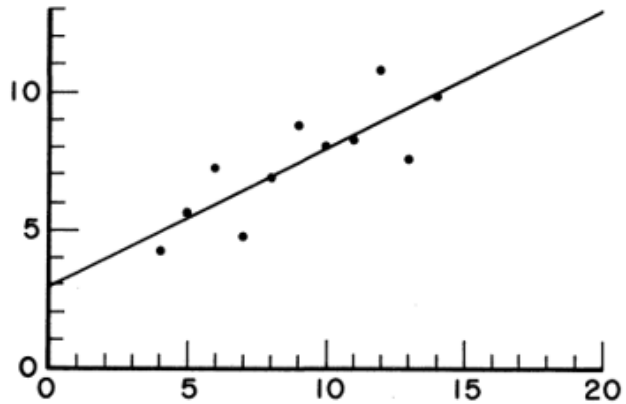
[24]

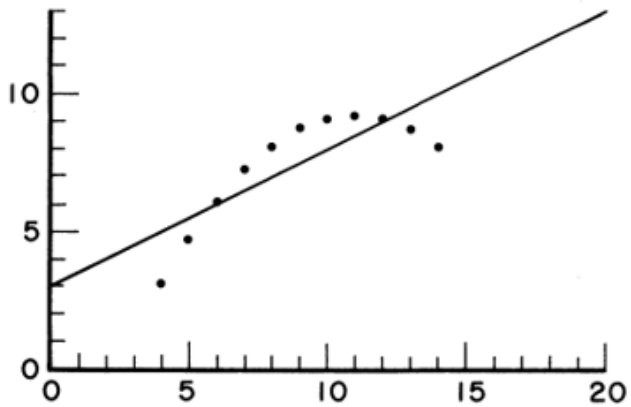
[24] Quelle: Eigene Darstellung

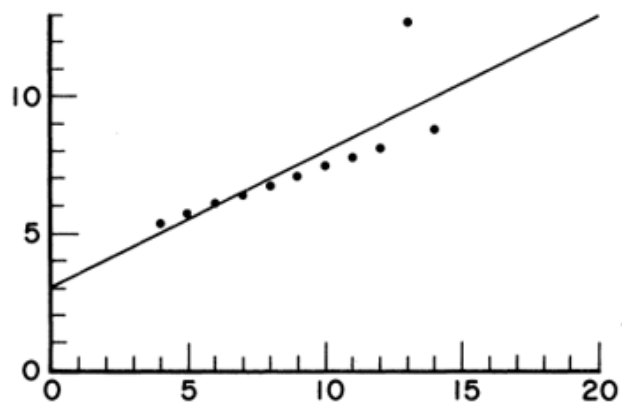
Danger of Summary Metrics

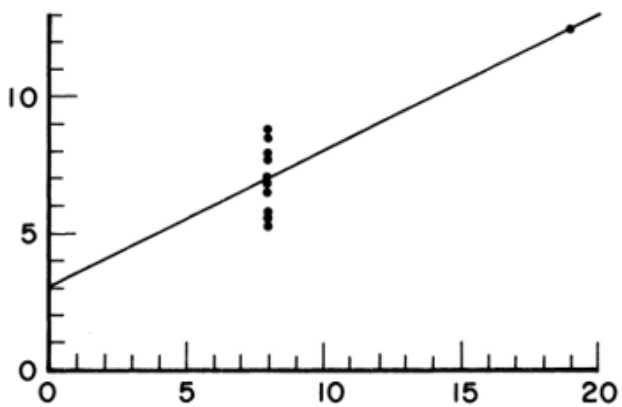
- Summary Metrics sind Aggregationsmaße von Daten bzw. zusammenfassende Metriken
- Beispiele für Summary Metrics sind: arithmetisches Mittel, Median, Standardabweichung, Prozentwerte, KPIs, ...
- Das Problem von diesen ist, dass bei alleiniger Betrachtung dieser viele Informationen aus dem Datensatz nicht abgebildet sind, da bei der Zusammenfassung Details verloren gehen
- Ausreißer, Unterschiede zwischen Gruppen, Verteilungen, etc. gehen verloren

Danger of Summary Metrics - Anscombe's Quartett









25

[25] Quelle: Anscombe (1973). S.19-20.

Danger of Summary Metrics[26]

- Die vernachlässigten Informationen können für die Entscheidungsfindung von großer Bedeutung sein
- Beispiele: durchschnittliches Einkommen, prozentuale Wirksamkeit von Medikamenten ohne Betrachtung von Gruppen, Erfolgsrate von Suchanfragen in einer App über alle Sprachen hinweg
- Handlungsempfehlung: Entscheidungen nicht nur auf zusammenfassenden Metriken aufbauen; Graphiken (z.B. Streudiagramme) anschauen

[26] Quelle: Anscombe (1973). S. 17.

Danger of Summary Metrics - Datensatz

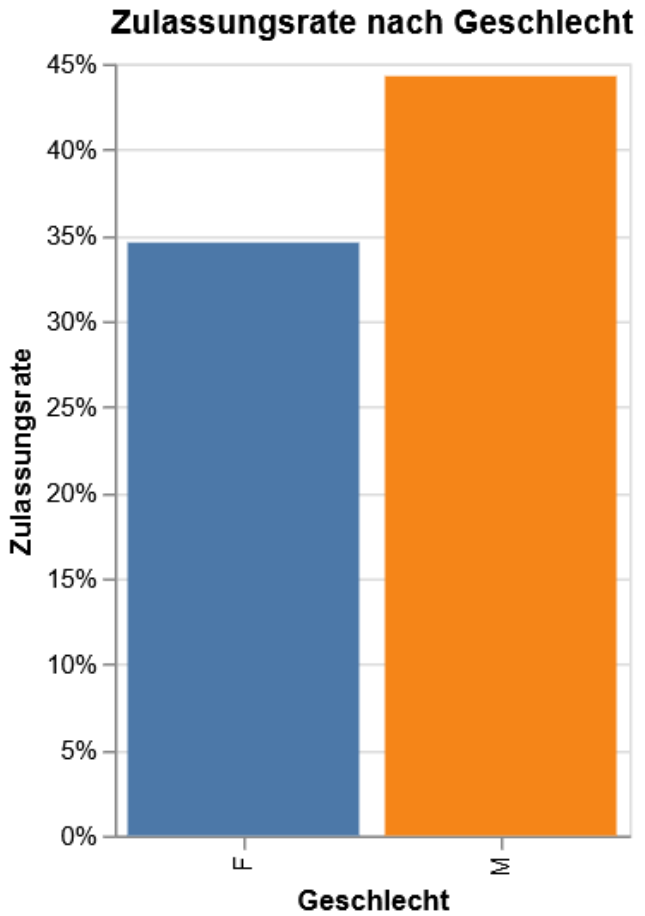
	Year	Major	Gender	Admission
0	1973	C	F	Rejected
1	1973	B	M	Accepted
2	1973	Other	F	Accepted
3	1973	Other	M	Accepted
4	1973	Other	M	Rejected
...	...	...	...	...
12758	1973	Other	M	Accepted
12759	1973	D	M	Accepted
12760	1973	Other	F	Rejected
12761	1973	Other	M	Rejected
12762	1973	Other	M	Accepted

12763 rows × 4 columns

27

[27] Quelle: `https://waf.cs.illinois.edu/discovery/berkeley.csv`

Danger of Summary Metrics



[28]

[28] Quelle: Eigene Darstellung

Danger of Summary Metrics

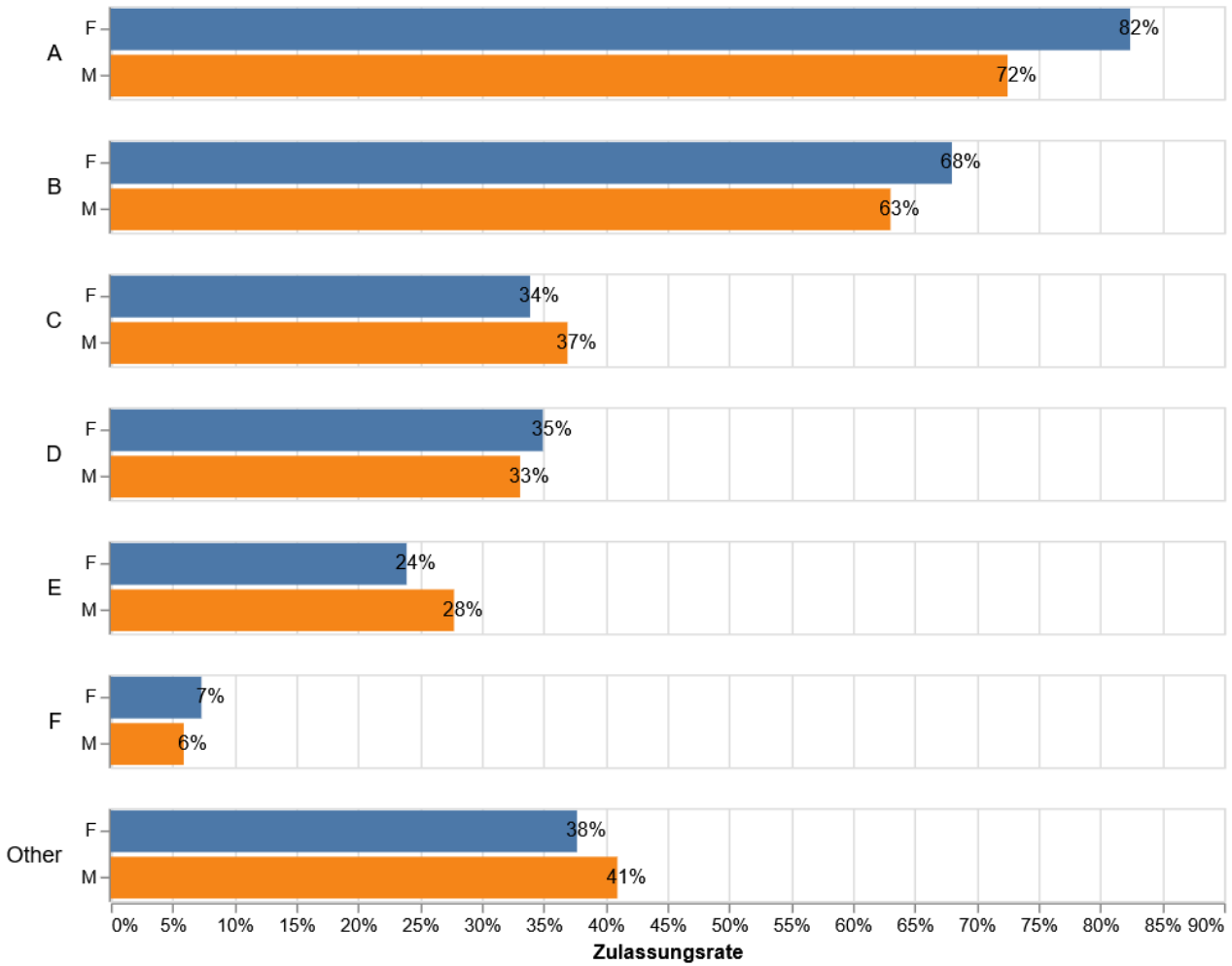
Aggregierte Tabelle nach Gender:

	Gender	Total_Applications	Total_Admissions	Admission_Rate
0	F	4321	1494	0.345753
1	M	8442	3738	0.442786

[29]

[29] Quelle: Eigene Darstellung

Danger of Summary Metrics



[30]

[30] Quelle: Eigene Darstellung

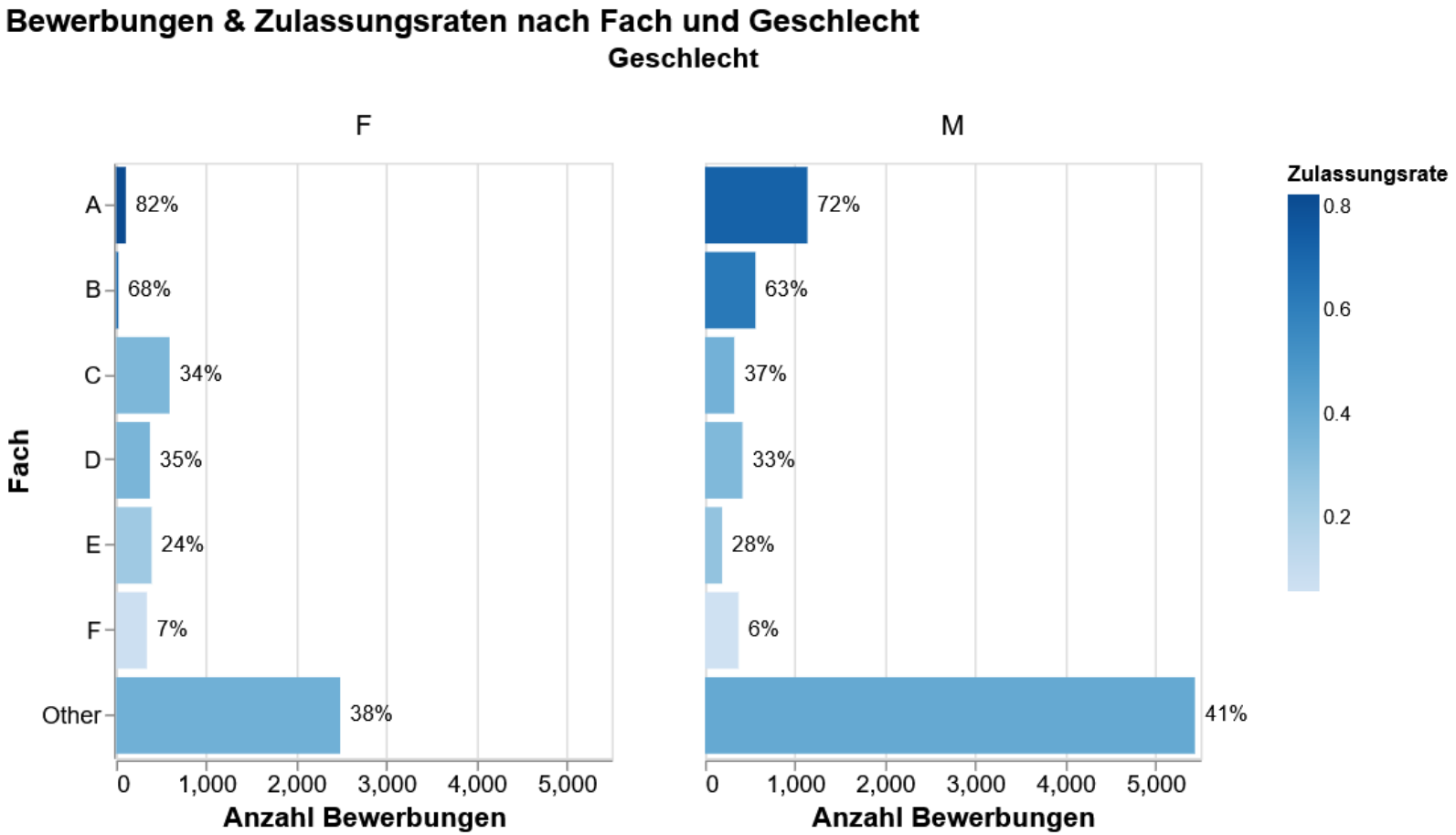
Danger of Summary Metrics

	Major	Gender	Total_Applications	Total_Admissions	Admission_Rate
0	A	F	108	89	0.824074
1	A	M	1138	825	0.724956
2	B	F	25	17	0.680000
3	B	M	560	353	0.630357
4	C	F	593	201	0.338954
5	C	M	325	120	0.369231
6	D	F	375	131	0.349333
7	D	M	417	138	0.330935
8	E	F	393	94	0.239186
9	E	M	191	53	0.277487
10	F	F	341	25	0.073314
11	F	M	373	22	0.058981
12	Other	F	2486	937	0.376911
13	Other	M	5438	2227	0.409526

31

[31] Quelle: Eigene Darstellung

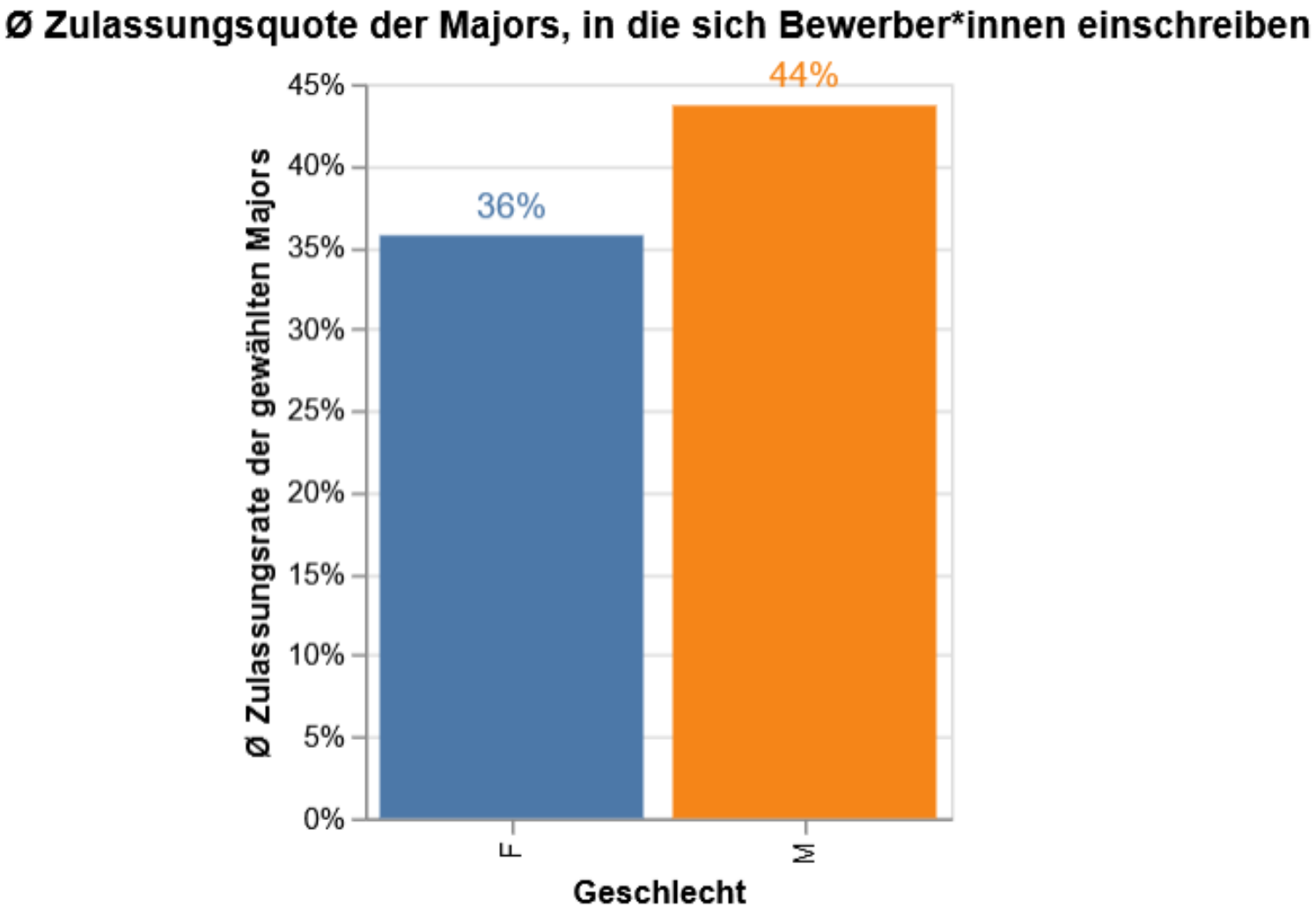
Danger of Summary Metrics



32

[32] Quelle: Eigene Darstellung

Danger of Summary Metrics



33

[33] Quelle: Eigene Darstellung

False Causality[34]

- Korrelation ist nicht gleich Kausalität!
- Das Auffinden kausaler Zusammenhänge ist eine wichtige Aufgabe in der Datenanalyse und ist vor allem für Prognosen und datengetriebene Entscheidungen relevant
- Ein häufiger Fehler ist es, von einer Korrelation auf einen kausalen Zusammenhang zu schließen
- Korrelationen ohne Kausalität nennt man Scheinkorrelationen
- Scheinkorrelationen entstehen i.d.R. durch eine unbekannte Variable, die die Variablen X und Y beeinflusst

[34] Quelle: Backhaus et al. (2022). S. 46.

False Causality[35]

- Beispiele für Korrelationen ohne Kausalität:
- Geburtenrate und Storchpopulation
- Schuhgröße von Schulkindern und deren Lesekompetenz
- Hopfenerträge und Konsum von Bier
- Kennt ihr noch weitere Korrelationen ohne Kausalität? Welche verdeckten Variablen könnten die beschriebenen Fälle verursachen?

[35] Quelle: Backhaus et al. (2022). S. 46.

False Causality[36]

- Handlungsempfehlung:
- Korrelationskoeffizienten betrachten: passt die Richtung des Zusammenhangs; ist Zusammenhang stark genug?
- Plausibilität prüfen: ist ein kausaler Zusammenhang logisch; gibt es einen zeitlichen Verzug in der Zeitreihe bei Ursache und Wirkung?
- Alternative Hypothesen aufstellen: welche verdeckten Variablen könnte es geben, die X und Y beeinflussen?

[36] Quelle: Backhaus et al. (2022). S. 46-47.

False Causality - Datensatz

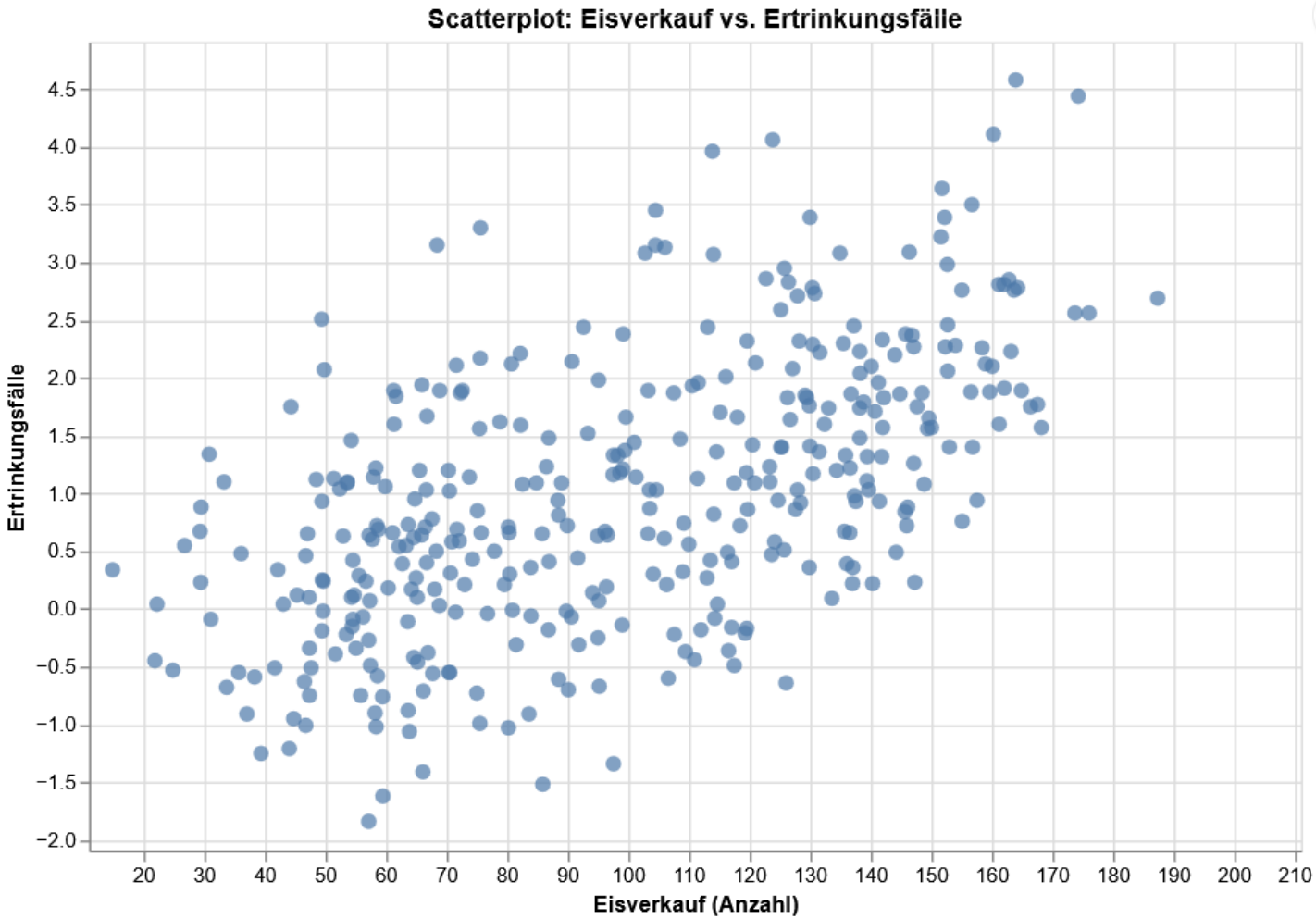
	date	temperature	ice_cream_sales	drownings
0	2023-01-01	13.70	119.79	0.86
1	2023-01-02	11.14	104.20	0.30
2	2023-01-03	12.47	109.21	0.74
3	2023-01-04	15.17	123.71	0.47
4	2023-01-05	14.59	113.95	3.96
...	...	...	...	...
360	2023-12-27	10.71	96.45	0.19
361	2023-12-28	9.49	89.99	0.72
362	2023-12-29	11.52	117.54	-0.49
363	2023-12-30	10.51	109.04	0.32
364	2023-12-31	9.97	104.54	3.15

365 rows × 4 columns

37

[37] Quelle: Selbsterstellter Datensatz

False Causality



[38]

[38] Quelle: Eigene Darstellung

False Causality

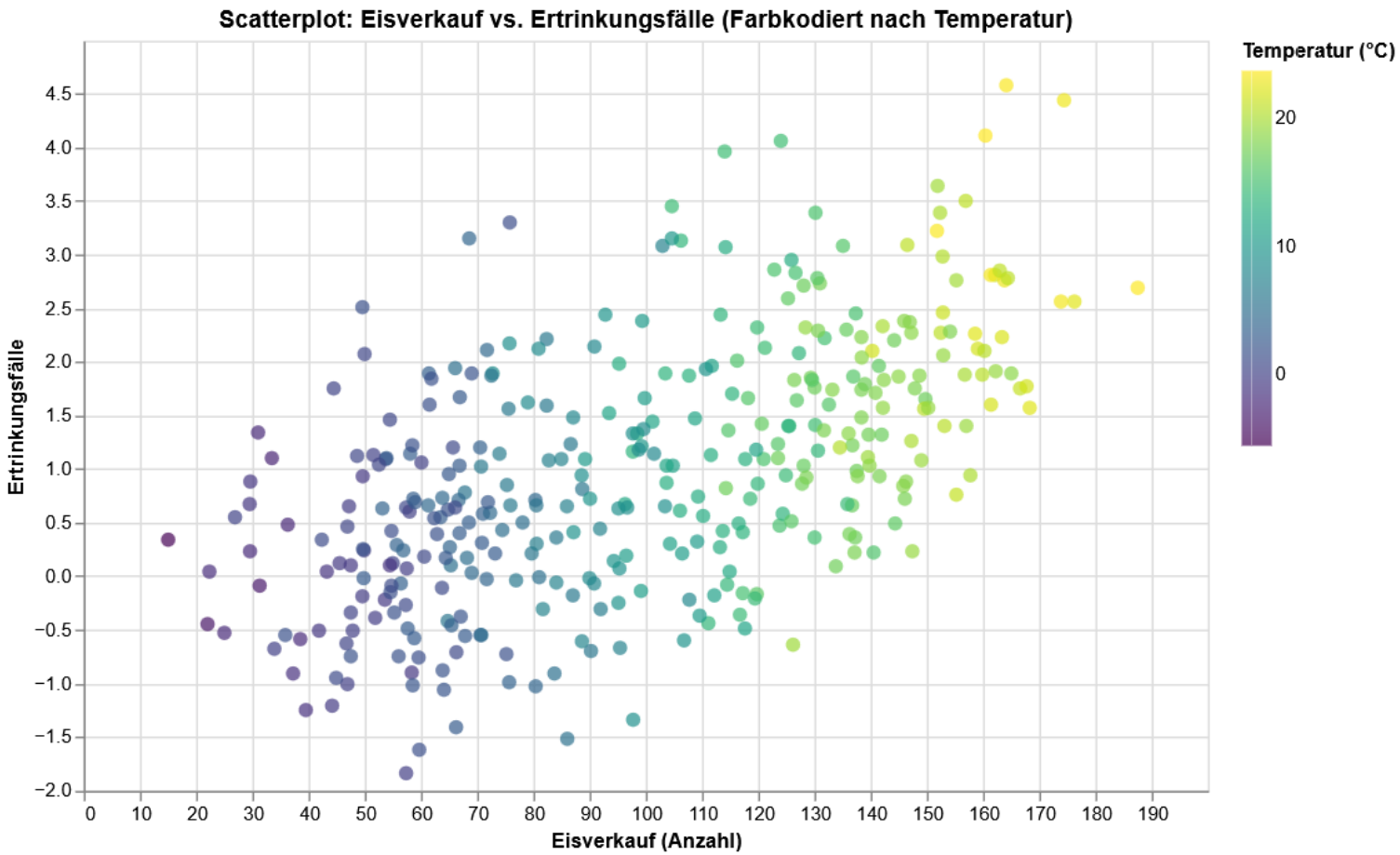
```
                            OLS Regression Results
==============================================================================
Dep. Variable:              drownings   R-squared:                       0.328
Model:                            OLS   Adj. R-squared:                  0.327
Method:                 Least Squares   F-statistic:                     177.6
Date:                Mon, 09 Jun 2025   Prob (F-statistic):           2.97e-33
Time:                        17:52:45   Log-Likelihood:                -501.06
No. Observations:                 365   AIC:                             1006.
Df Residuals:                     363   BIC:                             1014.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.7126      0.138     -5.176      0.000      -0.983      -0.442
ice_cream_sales  0.0173      0.001     13.325      0.000       0.015       0.020
==============================================================================
Omnibus:                        5.263   Durbin-Watson:                   2.126
Prob(Omnibus):                  0.072   Jarque-Bera (JB):                5.219
Skew:                           0.293   Prob(JB):                       0.0736
Kurtosis:                       3.006   Cond. No.                         292.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

[39]

[39] Quelle: Eigene Darstellung

False Causality



40

[40] Quelle: Eigene Darstellung

False Causality

OLS Regression Results

Dep. Variable:	drownings	R-squared:	0.348
Model:	OLS	Adj. R-squared:	0.345
Method:	Least Squares	F-statistic:	96.78
Date:	Mon, 09 Jun 2025	Prob (F-statistic):	2.13e-34
Time:	17:52:45	Log-Likelihood:	-495.55
No. Observations:	365	AIC:	997.1
Df Residuals:	362	BIC:	1009.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2111	0.309	0.683	0.495	-0.396	0.819
ice_cream_sales	-0.0027	0.006	-0.437	0.662	-0.015	0.009
temperature	0.1052	0.032	3.328	0.001	0.043	0.167

Omnibus:	7.814	Durbin-Watson:	2.096
Prob(Omnibus):	0.020	Jarque-Bera (JB):	7.702
Skew:	0.347	Prob(JB):	0.0213
Kurtosis:	3.159	Cond. No.	671.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.[41]

[41] Quelle: Eigene Darstellung

Quiz

Join at menti.com | Use vote code 2409 3670

Instructions

Go to

www.menti.com

Enter the code

2409 3670

Or use QR code [42]

[42] Quelle: Eigene Darstellung

Fazit und Abschluss

- Data Fallacies können zu datenbasierten Fehlentscheidungen führen, welche weitreichende Konsequenzen haben können
- Um die Risiken von Data Fallacies zu vermeiden, ist es sinnvoll, diese zu kennen und stets zu prüfen, ob eine Fallacy oder mehrere vorliegen

Vielen Dank für eure Aufmerksamkeit und Mitarbeit

Literaturverzeichnis

- Ameringer, S., Serlin, R. C., & Ward, S. (2009). *Simpson's Paradox and Experimental Research*. Nursing Research, 58(2), 123–127. https://doi.org/10.1097/NNR.0b013e318199b517
- Anscombe, F. J. (1973). *Graphs in Statistical Analysis*. The American Statistician, 27(1), 17–21. https://www.sjsu.edu/faculty/gerstman/StatPrimer/anscombe1973.pdf
- Backhaus, K., Erichson, B., Gensler, S., Weiber, R., & Weiber, T. (2022). *Multivariate Analysemethoden* (17. Aufl.). Springer Gabler.
- Bar-Hillel, M. (1980). *The Base-Rate Fallacy in Probability Judgements*. Acta Psychologica, 44, 211–233. https://bear.warrington.ufl.edu/brenner/mar7588/Papers/barhillel-acta1980.pdf
- Elston, D. (2021). *Survivorship Bias*. Journal of the American Academy of Dermatology, 1–2. https://www.sciencedirect.com/science/article/abs/pii/S0190962221019861?fr=RR-2&ref=pdf_download&rr=94e38a7b2e1b4534
- Flyvbjerg, B. (2022, Februar 1). *The Base-Rate Fallacy*. Towards Data Science. https://towardsdatascience.com/the-base-rate-fallacy-b94c0a1b9938/
- Grandjean, M., & McGeddon. (o.J.). *US Air Force (Hit Plot Concept)* [Graphic].
- Hintzman, D. L. (1980). *Simpson's Paradox and the Analysis of Memory Retrieval*. Psychological Review, 87(4), 398–410. https://www.researchgate.net/publication/20508020_Random_Sampling_Randomization_and_Equivalence_of_Contrasted_Groups_in_Psychotherapy_Outcome_Research

Literaturverzeichnis

- Hsu, L. M. (1989). *Random Sampling, Randomization, and Equivalence of Contrasted Groups in Psychotherapy Outcome Research*. Journal of Consulting and Clinical Psychology, 57(1), 131–137. https://www.researchgate.net/publication/20508020_Random_Sampling_Randomization_and_Equivalence_of_Contrasted_Groups_in_Psychotherapy_Outcome_Research
- Kahneman, D., & Tversky, A. (1974). *Judgment under Uncertainty: Heuristics and Biases*. Science, New Series, 185(4157), 1124–1131. http://www.jstor.org/stable/1738360
- Miller, B. (2020, August 29). *How 'survivorship bias' can cause you to make mistakes*. BBC. https://www.bbc.com/worklife/article/20200827-how-survivorship-bias-can-cause-you-to-make-mistakes
- Ogbonnaya, K. E., Okechi, B. C., & Nwankwo, B. C. (2019). *Statistical Fallacy: A Menace to the Field of Science*. International Journal of Scientific and Research Publications (IJSRP), 9(6), p9048. https://doi.org/10.29322/IJSRP.9.06.2019.p9048
- OpenIntro. (n.d.). *Simpson's paradox and COVID-19 vaccine effectiveness*. OpenIntro. https://www.openintro.org/data/index.php?data=simpsons_paradox_covid
- University of Illinois Urbana-Champaign. (n.d.). *UC Berkeley Admissions Data* [Berkeley Gender Bias Dataset]. Discovery: Illinois Data Science Initiative. https://discovery.cs.illinois.edu/dataset/berkeley/