Themenblock Datenbeschaffung und -aufbereitung: 04_03Aufbereitung_Datenexploration

Explorative Datenanalyse

Erarbeitet von
Dr. Ann-Kathrin Selker

Lernziele

- Du kannst beispielhaft eine explorative Datenanalyse durchführen
- Du kannst Beispiele für Erkenntnisse nennen, die mit explorativer Datenanalyse gewonnen werden können

Projekt ai4all „KI für Alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Inhalt

Deine Daten sind gesammelt. Zeit für die Bereinigung und Aufbereitung… Aber wie erkennst du eigentlich, was alles gemacht werden muss?

Einstieg

Ein elementares Mittel, um sich in den gesammelten Daten auszukennen, ist die explorative Datenanalyse. Dabei benutzt du Visualisierungen und statistische Berechnungen, um deine Daten besser zu verstehen. In diesem Video liegt der Fokus auf der Visualisierung. Visualisierung erlaubt es dir, dir schnell einen Überblick über Verteilungen zu verschaffen, erste Zusammenhänge festzustellen und anderen deine Erkenntnisse mitzuteilen.

Am besten gucken wir uns zusammen mal einige Datensätze an, um typische Erkenntnisse einer Datenexploration nachvollziehen zu können.

Als ersten Schritt bietet es sich an, einmal alle Features durchzugehen und ihre Bedeutung und Intention zu verstehen. Wenn wir uns das California Housing Dataset ansehen, fällt auf, dass beim Feature „Durchschnittliche Anzahl Schlafzimmer" erstaunlich große Werte auftreten. So liegt der maximale Wert zum Beispiel bei über 34, bei „Durchschnittliche Anzahl Zimmer" sogar bei über 141. Das liegt an der Berechnung dieser Werte: Die Durchschnitte wurden pro „Block" und pro Haushalt berechnet. Ein Block ist dabei die kleinste geographische Einheit, für die das U.S. Census Bureau Daten veröffentlicht, und umfasst etwa 600-3000 Personen.

Quelle [1]

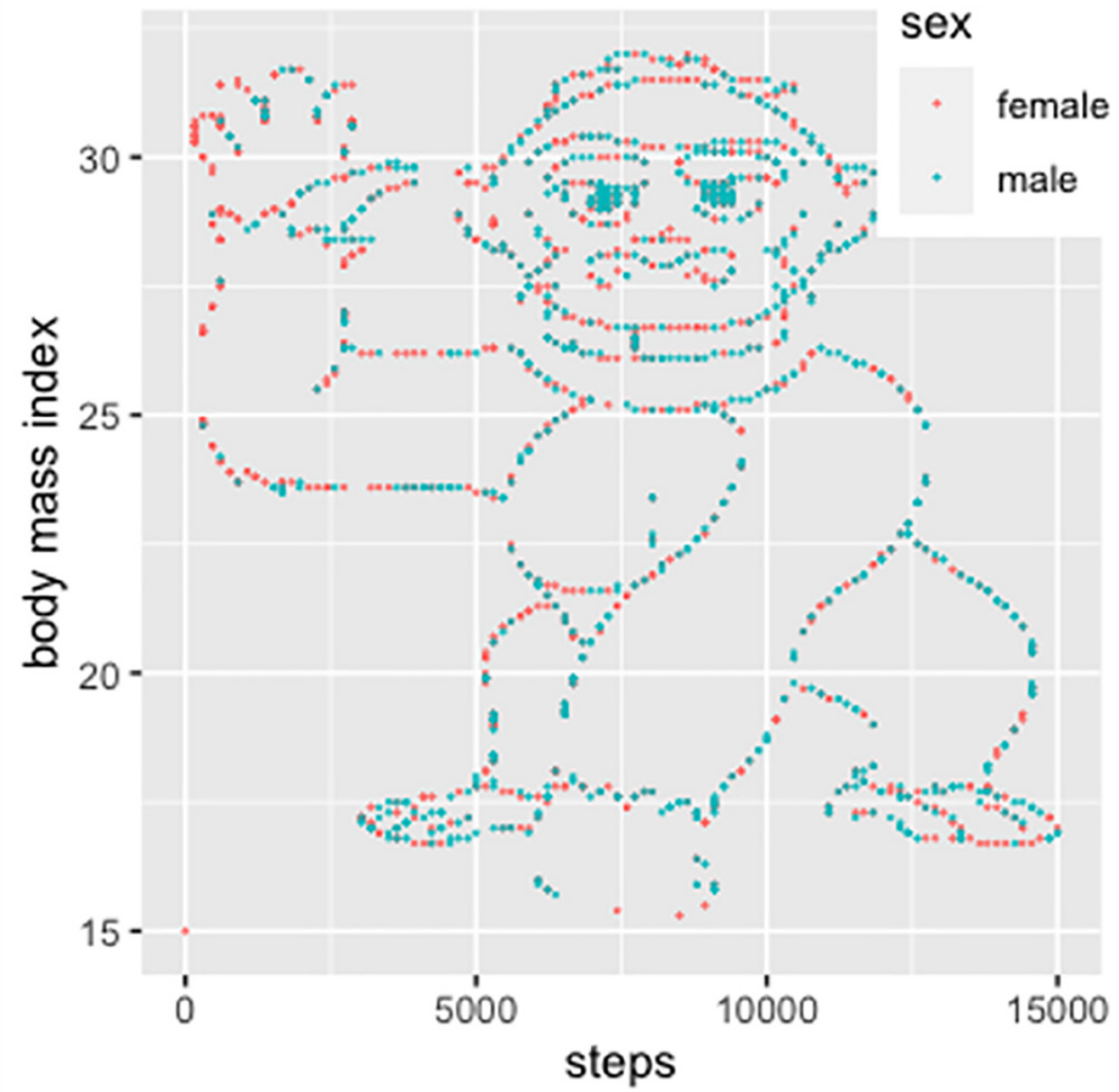
Falls ein Block leere Häuser enthält, werden deren Zimmer bei der Gesamtanzahl der Zimmer im Block berücksichtigt, geteilt wird am Ende aber durch die Anzahl an Haushalten in dem Block. In Blöcken mit vielen Ferienhäusern oder aus anderen Gründen leerstehenden Häusern wird die durchschnittliche Zimmeranzahl pro Haushalt somit sehr hoch ausfallen.

Quelle [1]

Für eine richtige Interpretation der Ergebnisse ist es also wichtig, genau zu verstehen, was die gemessenen Werte eines Features eigentlich aussagen.

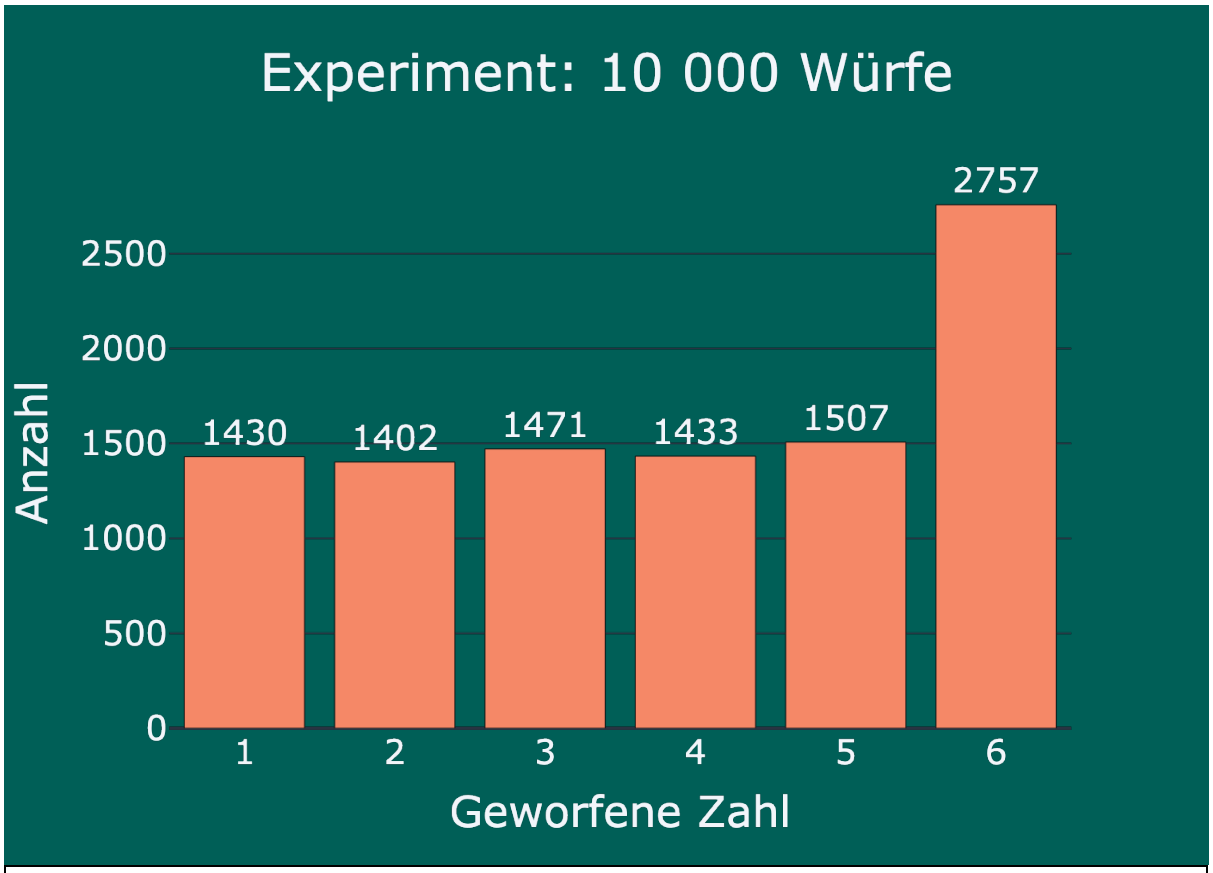
Verteilung der Daten

Sobald wir die Bedeutung unserer Features verstanden haben, können wir dazu übergehen, die einzelnen Features und ihre Beziehungen untereinander zu untersuchen. Wenn wir Features plotten, können wir nicht nur versteckte Easter Eggs wie diesen Gorilla finden,



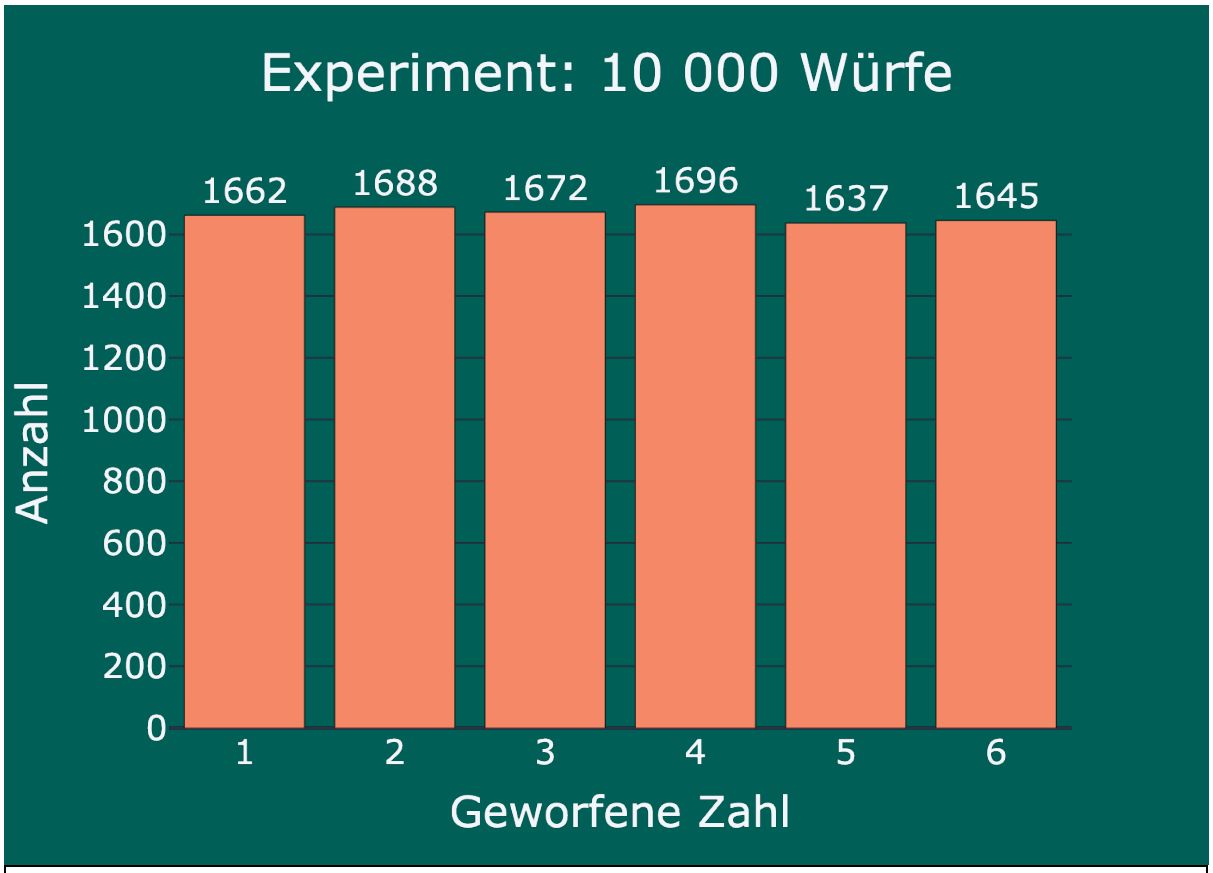
Einblendung Gorilla-Diagramm (Quelle [2])

sondern auch ohne großen Zeitaufwand schon einmal sehen, wie unsere Daten verteilt sind. Vielleicht erleben wir dabei Überraschungen, wie die Ergebnisse dieser Würfe.



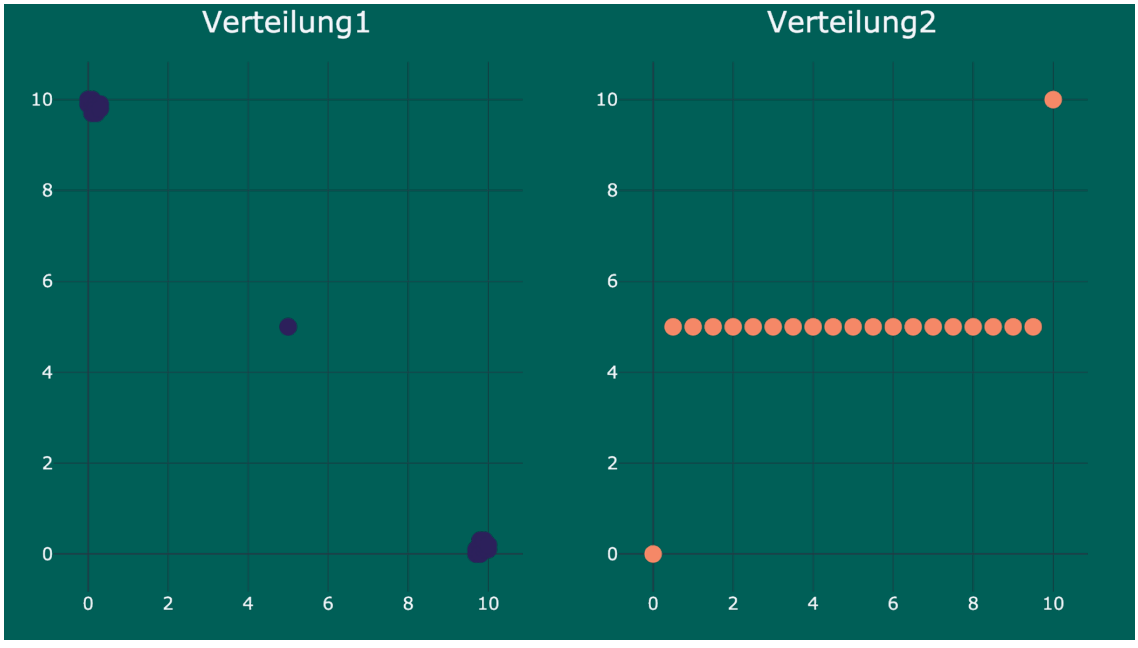
Einblendung Diagramm 10 000 Würfe (1)

Eigentlich sollte das Bild doch ungefähr so aussehen, oder? Vielleicht handelt es sich also um einen gezinkten Würfel. Durch Visualisierung können wir also auch grob überprüfen, ob unsere Daten plausibel sind.



Einblendung Diagramm 10 000 Würfe (2)

Wenn wir die Verteilung der Datenpunkte grob einschätzen können, hilft uns das schon, in der späteren statistischen Analyse die besten Verfahren auszuwählen. Rein durch statistische Kennzahlen kann man nicht immer gut sehen, wie sich die Daten tatsächlich verteilen. So haben zum Beispiel diese beiden Datensätze denselben arithmetischen Mittelwert, Minimal- und Maximalwert, sowie Median, sind aber grundlegend verschieden.

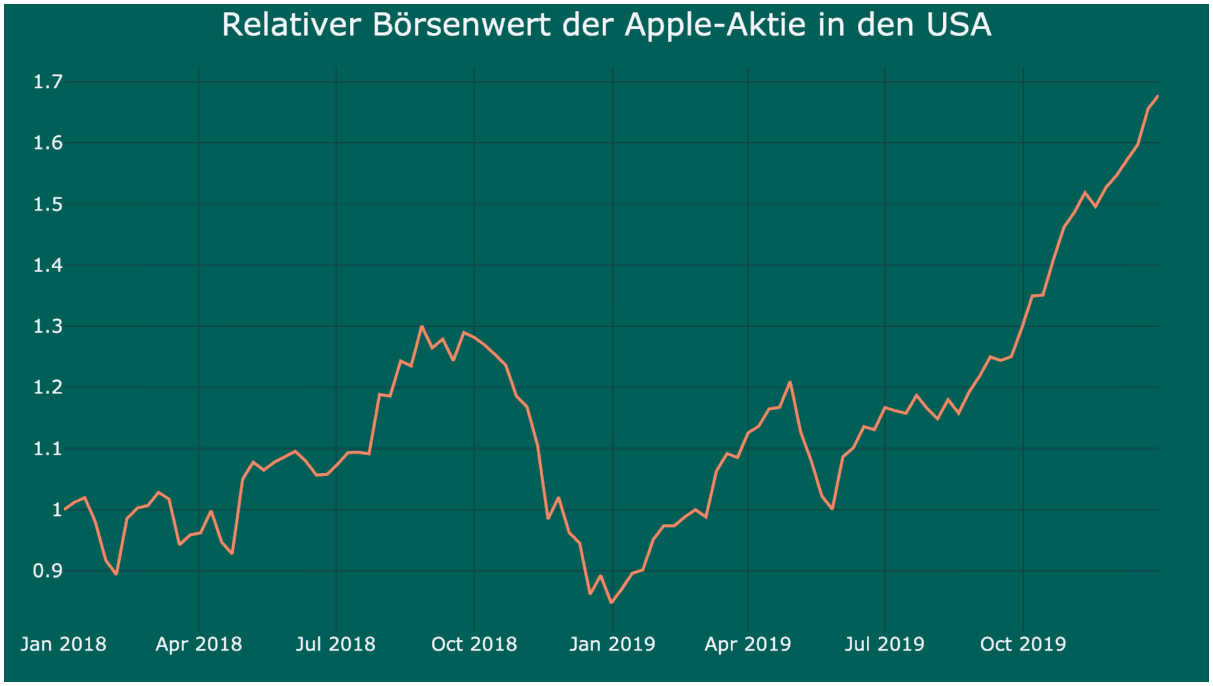


Einblendung Diagramme

Dieses Beispiel zeigt auch, dass du dich nicht auf reine Aggregationen deiner Daten verlassen kannst. Es liegt in der Natur der Aggregationen, dass durch ihre Anwendung Informationen verloren gehen. Daher ist es so wichtig, dass du deine Daten erst einmal kennenlernst, bevor du anfängst sie zu transformieren.

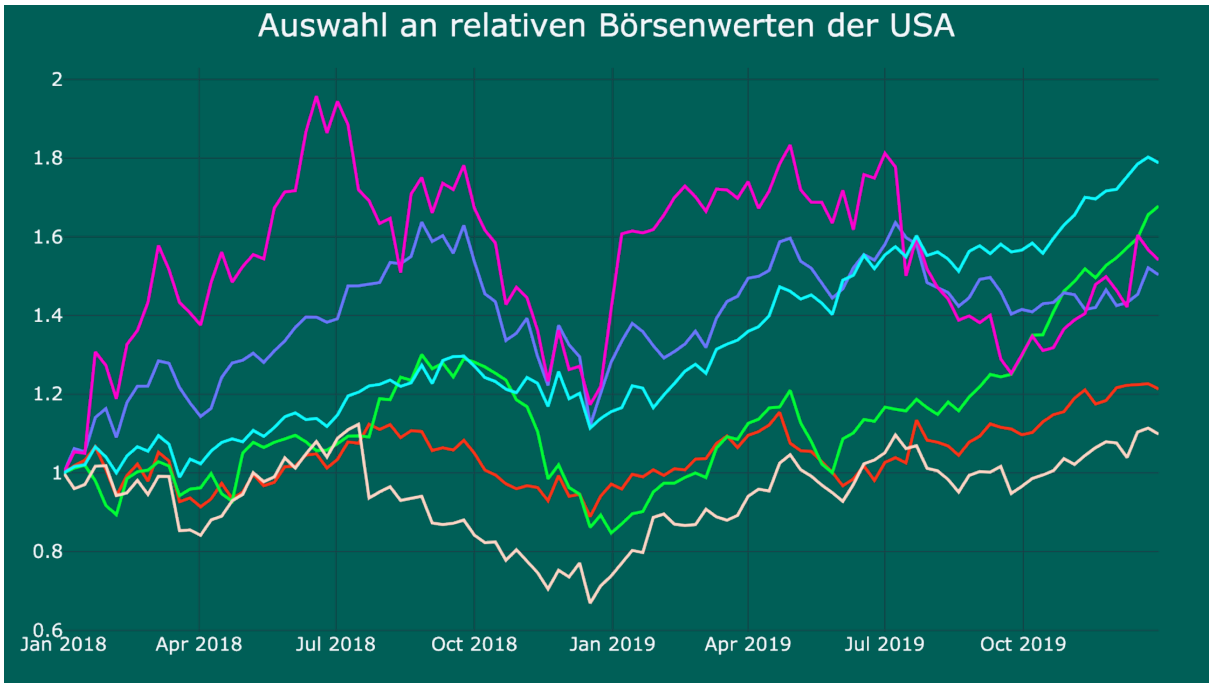
Zusammenhänge und Muster

Eine Visualisierung hilft uns auch dabei, schnell Ausreißer und Muster zu erkennen. Wenn wir relative Börsenwerte plotten, wie hier den relativen Aktienwert von Apple, können wir sofort sehen, ob die Kurve längerfristig steigt oder fällt, auch wenn die täglichen Schwankungen das vielleicht erst nicht vermuten lassen.



Einblendung Diagramm Plotly Stocks

Und wenn wir nicht nur den Börsenwert einer Aktie, sondern vieler Aktien in einem Diagramm plotten, erkennen wir auch schnell Gemeinsamkeiten, wie hier den kollektiven Wertverlust Ende 2018.

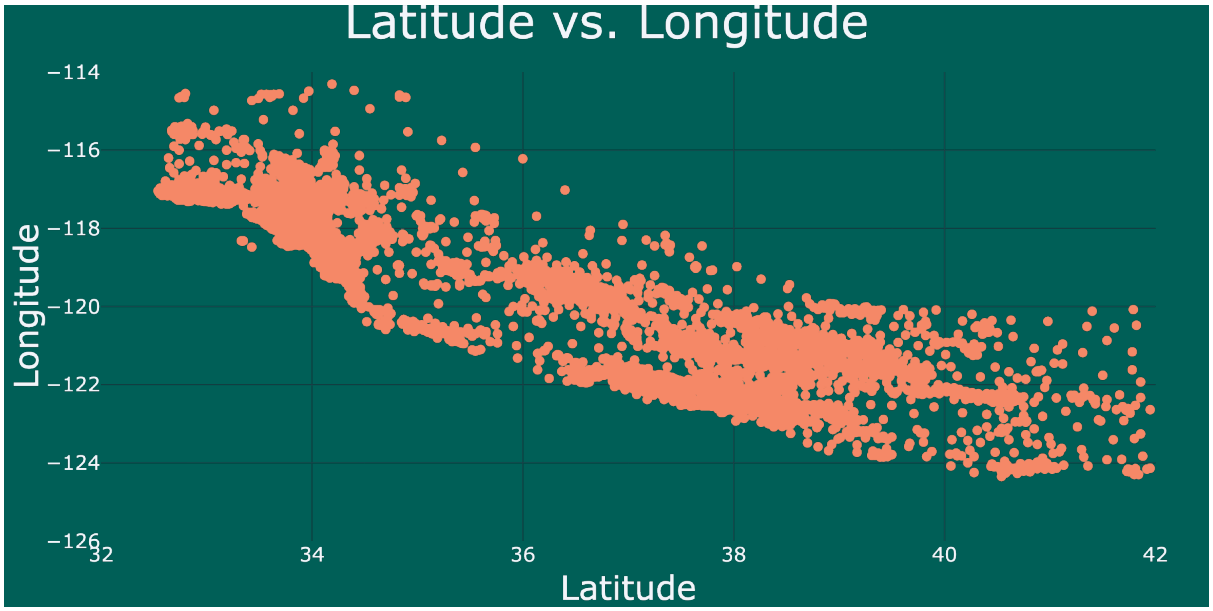


Einblendung Diagramm Plotly Stocks

Nach einer kurzen Nachforschung wird klar, dass es Ende 2018 in den USA einen großen Aufruhr um einige Entscheidungen des damaligen Präsidenten Donald Trump gab und es dementsprechend zu Unsicherheiten auf dem Aktienmarkt kam.

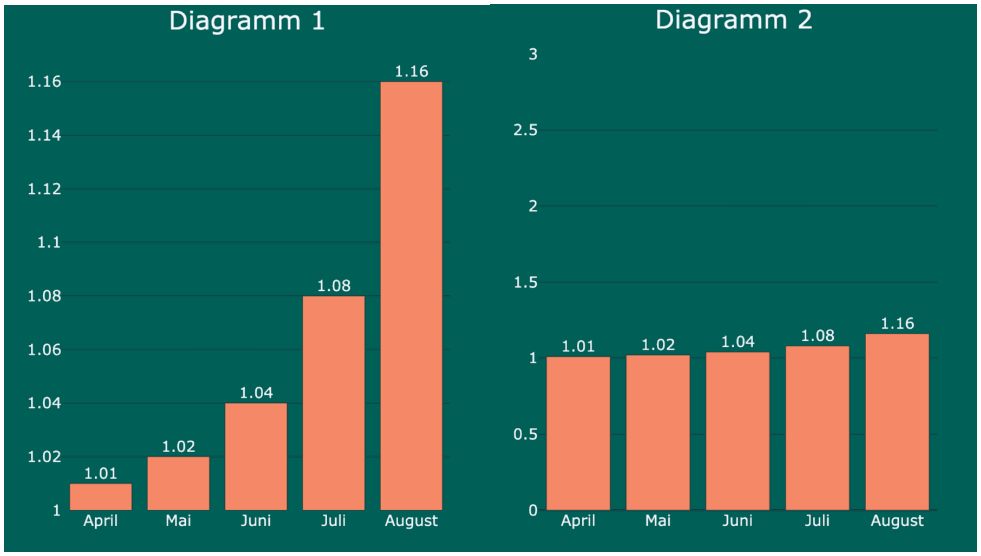
Quelle [3]

Aber auch erste Zusammenhänge zwischen Features können visuell erkannt werden. Diese sogenannten Korrelationen können bei der späteren Aufbereitung deiner Daten noch eine wichtige Rolle spielen. So kannst du zum Beispiel nach dem Betrachten des Diagramms des California Housing Datasets vermuten, dass es einen Zusammenhang zwischen dem Längen- und dem Breitengrad der Häuser geben könnte.



Einblendung Diagramm California Housing Latitude vs. Longitude

Doch genau in diesem Bereich befinden sich auch große Stolperfallen: Visualisierung hilft zwar dabei, erste Ansätze für die Untersuchung der Daten zu liefern, aber eine statistische Untersuchung gehört zur Datenanalyse immer dazu. Denn je nachdem, wie du deine Daten darstellst, kannst du unterschiedliche Schlüsse ziehen. Betrachte doch mal die folgenden beiden Diagramme.

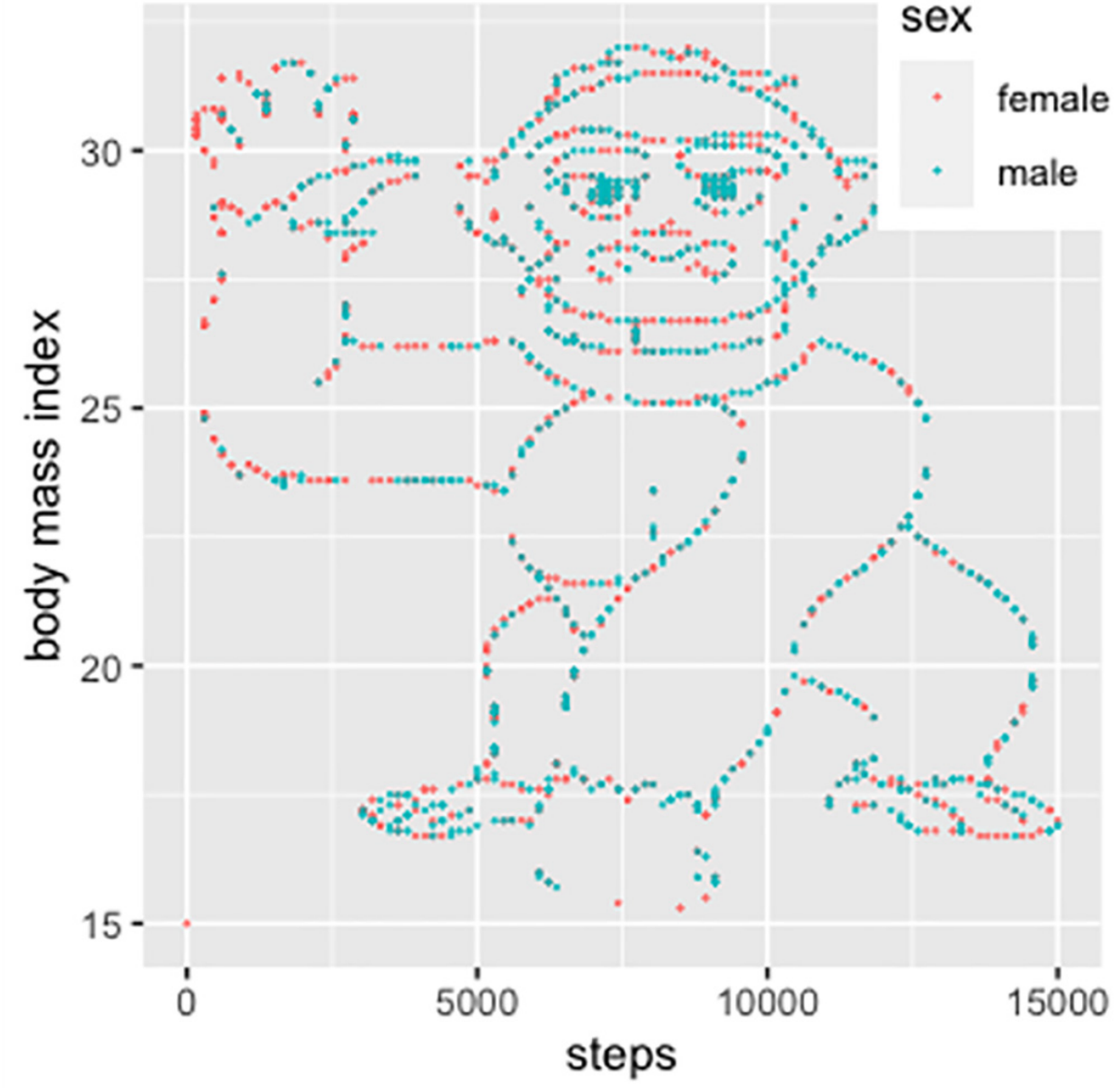


Einblendung Diagramme

In beiden Diagrammen ist derselbe Datensatz dargestellt, die Interpretationen wären bei mir aber verschiedene. Im ersten Diagramm würde ich einen starken Anstieg der Werte

bemerken, im zweiten dagegen eher eine Stagnation. Welche dieser Interpretationen die richtige ist, hängt stark davon ab, was die betrachteten Werte genau aussagen.

Bei der explorativen Datenanalyse ist es besonders wichtig, unvoreingenommen an die Erkundung der Daten heranzugehen. Wenn du von Anfang an Hypothesen darüber hast, welche Erkenntnisse in deinen Daten versteckt sind, prüfst du eventuell nur auf diese und lässt dir dadurch andere Erkenntnisse entgehen. Das eben gezeigte Gorilla-Bild ist dafür ein gutes Beispiel.



Einblendung Gorilla-Diagramm (Quelle [2])

In einer Studie stellte sich heraus, dass Studierende ohne vorgegebene Analyseaufgabe fünfmal häufiger den Gorilla im Datensatz fanden als ihre Kommiliton*innen, denen eine Hypothese zum Überprüfen mitgegeben wurde.

Quelle [2,4]

Lass dir also am Anfang wirklich Zeit, um deine Daten frei zu erkunden. Wer weiß, was in deinen Daten alles versteckt ist?

Ausblick

Du hast jetzt einen kleinen Überblick darüber enthalten, was für Informationen in deinen Daten auf dich warten können. Der Schwerpunkt dieses Videos lag auf strukturierten Daten, aber auch bei unstrukturierten Daten gehört eine explorative Datenanalyse natürlich zu den allerersten Schritten auf der To-Do-Liste. Für eine erfolgreiche Datenexploration benötigt man in der Regel gute Statistikkenntnisse und ein Auge dafür, Muster in Visualisierungen zu erkennen. Wie immer gilt also: Übung macht den Meister!

Quellen

Quelle [1] *7.2. Real world datasets.* (o. D.). Scikit-learn. https://scikit-learn.org/stable/datasets/real_world.html#california-housing-dataset

Quelle [2] Yanai, I. & Lercher, M. (2020). A hypothesis is a liability. *Genome Biology, 21(1).* https://doi.org/10.1186/s13059-020-02133-w

Quelle [3] Heath, T. & Rucker, P. (2018). Markets stage one of worst Christmas Eves ever, closing down more than 600 points as Trump blames Fed for stock losses in a tweet. *Washington Post*. https://www.washingtonpost.com/business/economy/us-markets-continue-sharp-sell-off-ignoring-efforts-by-the-trump-administration-to-stabilize-stock-prices/2018/12/24/59d4eae8-078e-11e9-85b6-41c0fe0c5b8f_story.html

Quelle [4] Claussen, A. (2021). Den Gorilla vor lauter Hypothesen nicht sehen. *hhu.* https://www.hhu.de/news-einzelansicht/den-gorilla-vor-lauter-hypothesen-nicht-sehen

Weiterführendes Material

https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/

https://www.geeksforgeeks.org/exploratory-data-analysis-in-python/

Disclaimer

Transkript zu dem Video „04 Datenbeschaffung und -aufbereitung: Explorative Datenanalyse“, Ann-Kathrin Selker.