Themenblock Datenbeschaffung und -aufbereitung: 04_03Aufbereitung_Korrelation

Korrelation

Erarbeitet von
Dr. Ann-Kathrin Selker

Lernziele

- Du kannst Korrelation erklären
- Du kannst anhand von Beispielen erläutern, wie Korrelation erkannt werden kann

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

Inhalt

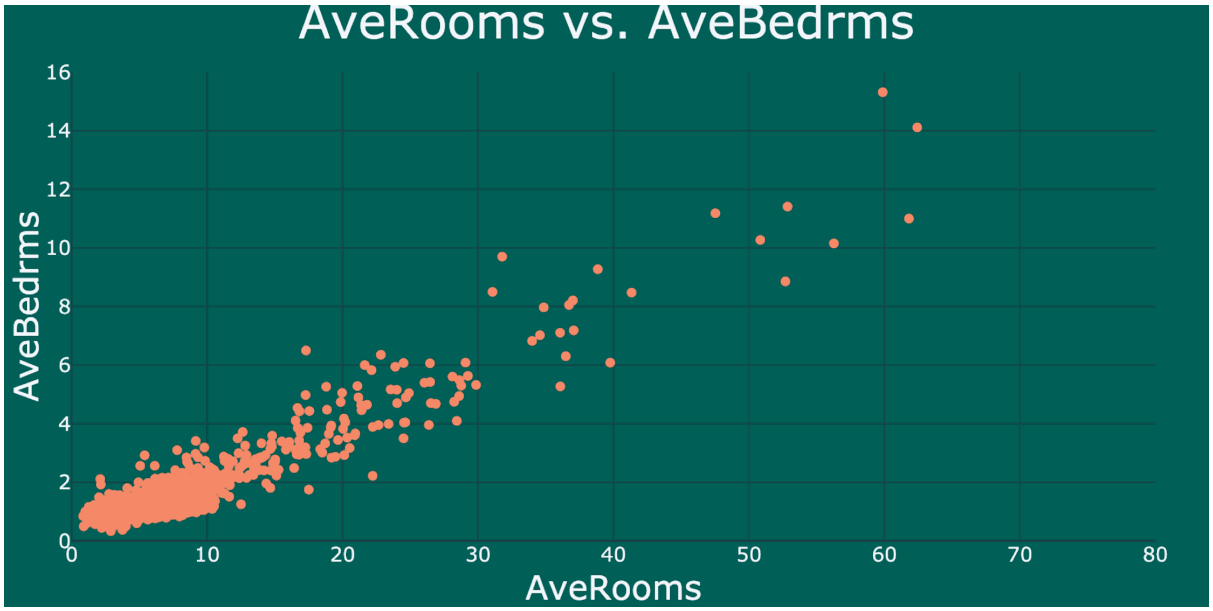
Einstieg

Der Großteil der Zeit beim Machine Learning wird damit verbracht, die Daten und ihre Eigenarten kennenzulernen. Dazu gehören auch Zusammenhänge zwischen den einzelnen Features: Gibt es Gemeinsamkeiten? Beeinflussen sich die Werte mancher Features sogar gegenseitig? Und wie erkennen wir solche Fälle?

Korrelation

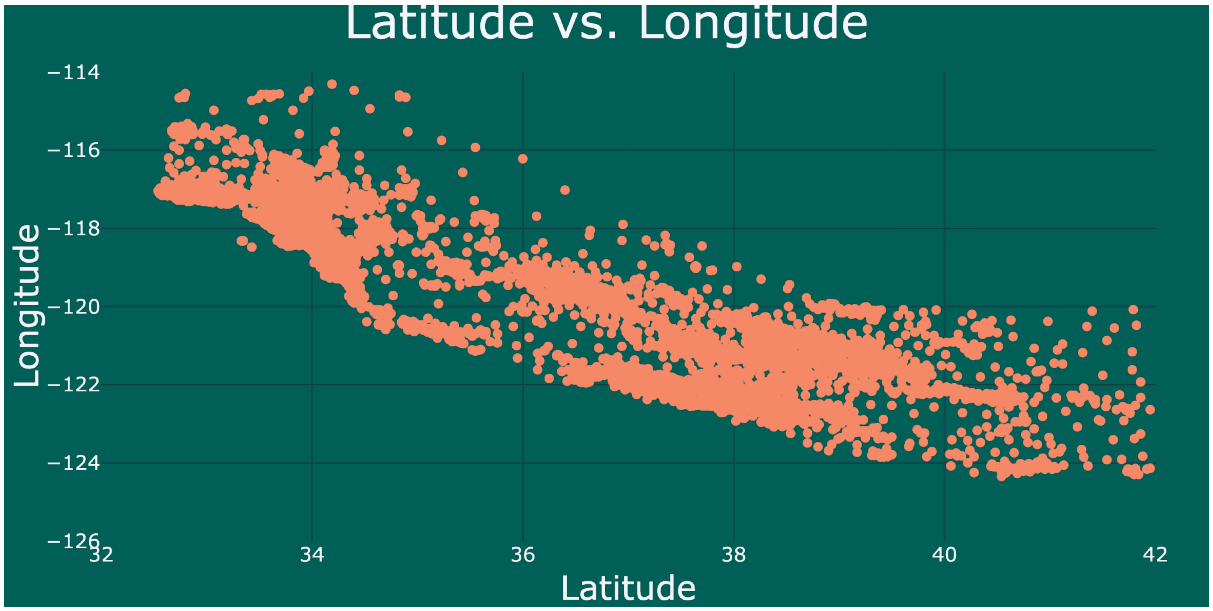
In diesem Video beschäftigen wir uns mit der sogenannten Korrelation. Das Wort Korrelation steht dabei für Zusammenhang. Wir sagen, dass zwei Features zusammenhängen, wenn sich ihre Werte auch ähnlich entwickeln. Am einfachsten kannst du dies vielleicht an einem Beispiel sehen. Betrachten wir einmal den Datensatz California Housing, der dir schon bekannt sein sollte. Zur Erinnerung: Der Datensatz beschreibt die Immobilienpreise in Kalifornien laut einem Zensus von 1990.

Wenn wir uns die Werte der Features „Durchschnittliche Anzahl Zimmer" und „Durchschnittliche Anzahl Schlafzimmer" ansehen, fällt auf, dass sich diese Werte ähnlich entwickeln.



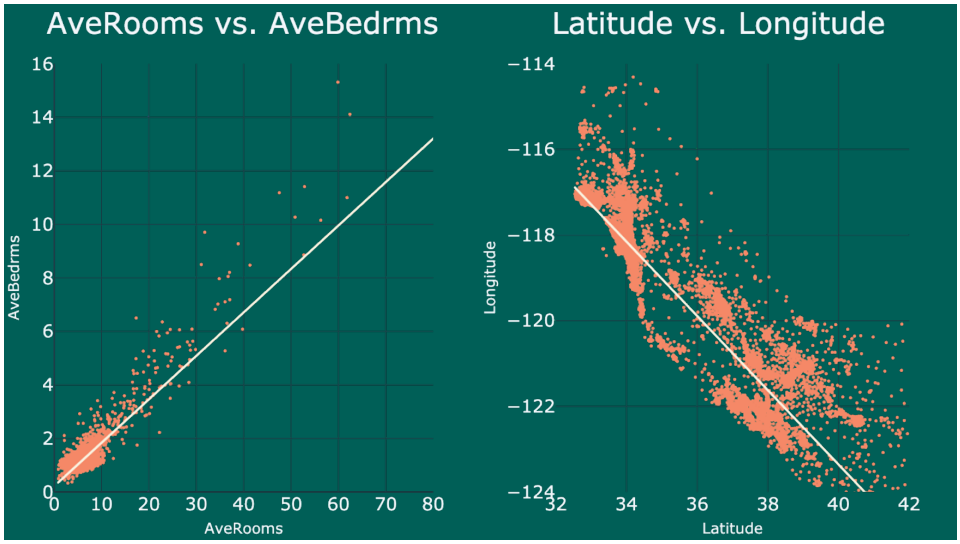
Einblendung Diagramm California Housing AveRooms vs AveBedrms

Features können auf unterschiedliche Weise zusammenhängen. Im Beispiel hast du die positive Korrelation gesehen. Das bedeutet, dass die Werte eines Features steigen, wenn auch die Werte des anderen Features steigen. Im Gegensatz dazu liegt eine negative Korrelation zweier Features vor, wenn die Werte eines Features steigen, wenn die des anderen Features fallen. Hier ist der Vergleich zwischen dem Längen- und dem Breitengrad der jeweiligen Häuser.



Einblendung Diagramm California Housing Latitude vs. Longitude

In beiden Diagrammen können wir das Wachstum beider Variablen im Verhältnis zueinander mit einer Geraden beschreiben. Es handelt sich also um einen linearen Zusammenhang. Es gibt aber auch quadratische, exponentielle und andere Zusammenhänge. Auch kategoriale Daten können korrelieren.



Einblendung Diagramm California Housing AveRooms vs. AveBedrms und Latitude vs. Longitude mit eingezeichneten Trendlines

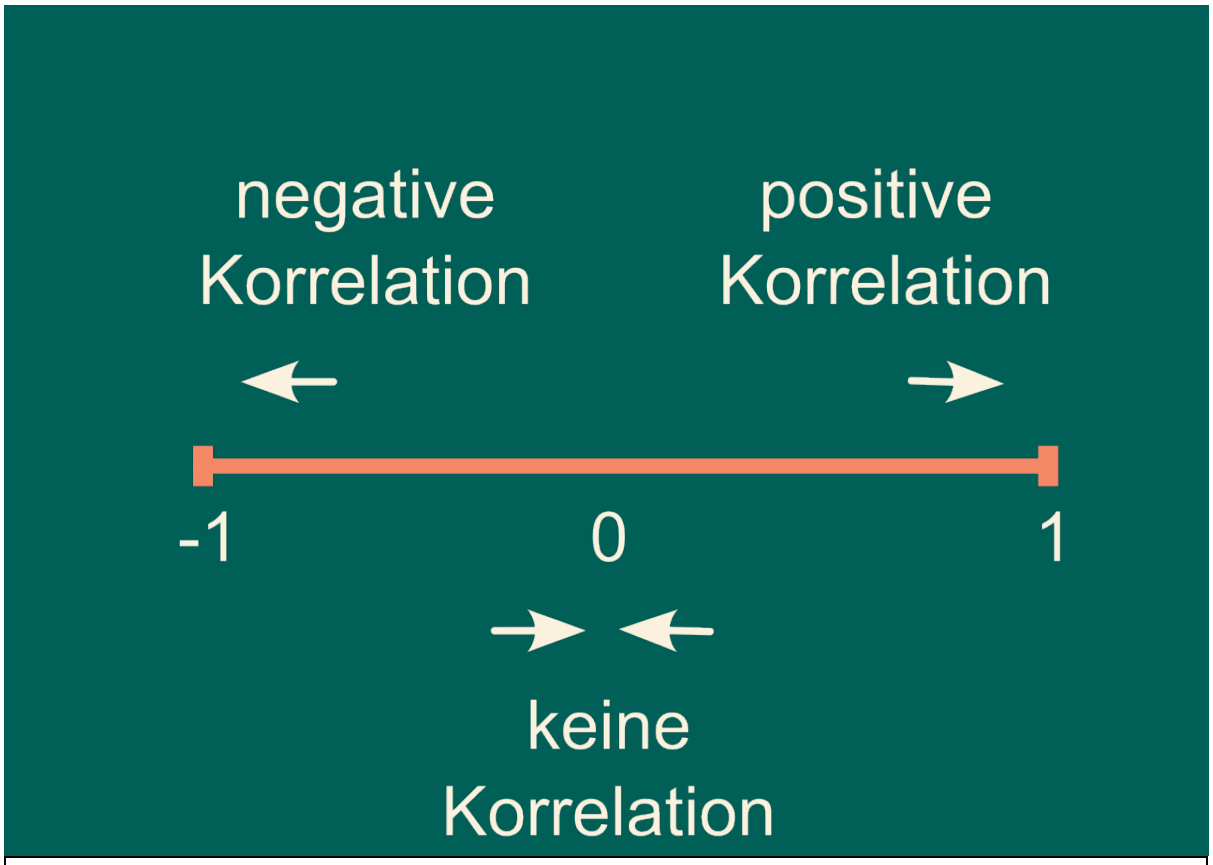
Korrelationsanalyse

Wie bereits gesehen, sind Streudiagramme eine gute Methode, um Zusammenhänge zwischen zwei Features zu visualisieren. Dies kannst du gut einsetzen, wenn du bereits Zusammenhänge zwischen diesen Features vermutest. Kompliziert wird es aber natürlich, sobald du viele Features in deinem Datensatz vorliegen hast. Es ist extrem ineffizient, alle möglichen Kombinationen von Features zu plotten, in der Hoffnung, Zusammenhänge zu finden.

Hier kommt die Statistik ins Spiel. Um die Korrelation zwischen Variablen zu berechnen, kannst du eine sogenannte Korrelationsanalyse durchführen. Je nach Art des Zusammenhangs (linear vs. nicht-linear, metrisch vs. kategorial) gibt es andere Verfahren, mit denen du diese Analyse durchführst. Wir benutzen hier als Beispiel die sogenannte Pearson-Korrelationsanalyse, die jeweils den linearen Zusammenhang zwischen zwei Features berechnet.

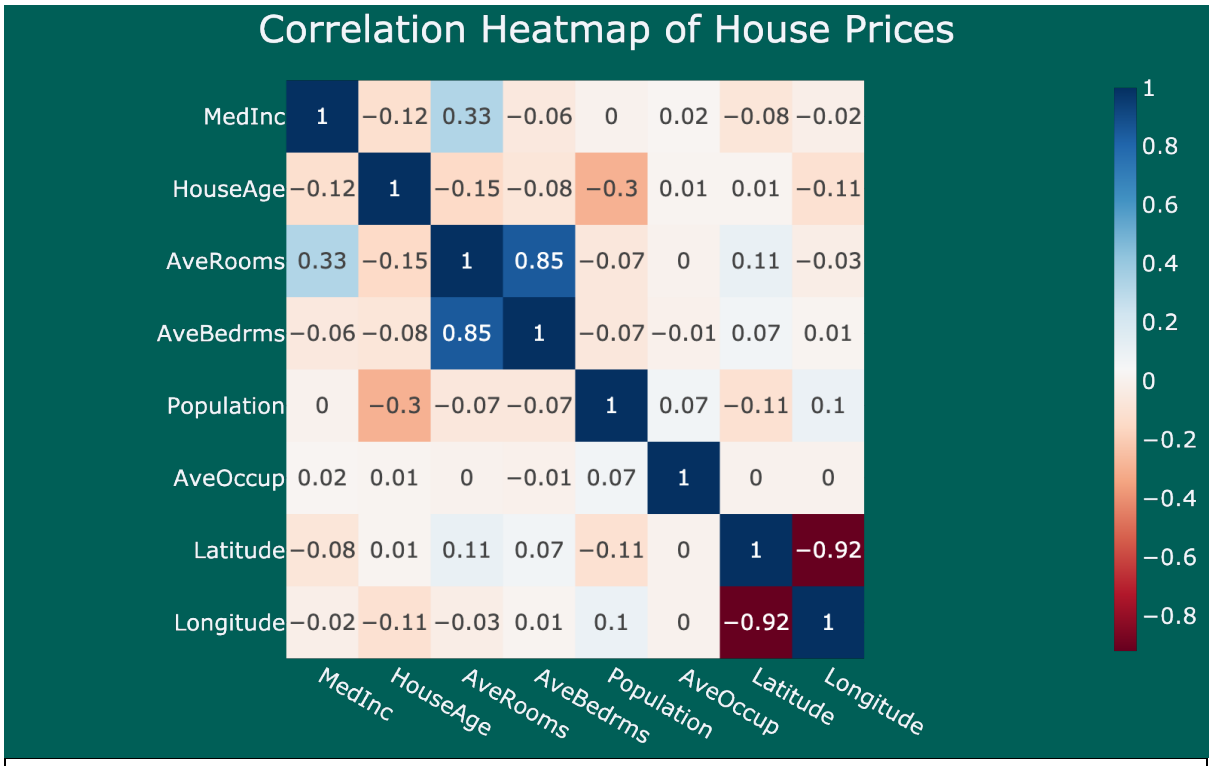
Quelle [1]

Das Ergebnis der Korrelationsanalyse ist der sogenannte Korrelationskoeffizient. Bei Pearson bewegt sich dieser zwischen -1 und 1. Je näher der Wert an 1 liegt, desto stärker sind die Features positiv korreliert, je näher an -1, desto stärker sind sie negativ korreliert. Ein Wert von 0 steht für keine Korrelation.



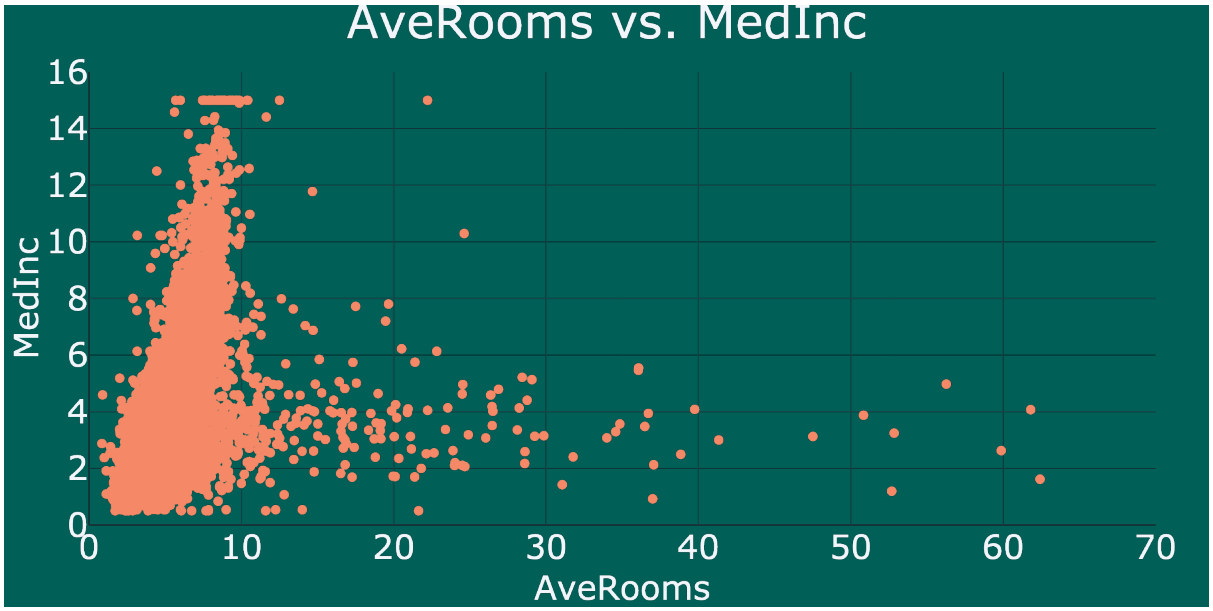
Einblendung Korrelationskoeffizient

Das Ergebnis der Analyse ist eine Korrelationsmatrix. Bei dieser Grafik handelt es sich übrigens um eine sogenannte Heatmap, bei der die Höhe des jeweiligen Korrelationskoeffizienten farbig dargestellt werden. Zuerst fällt auf, dass die Diagonale der Matrix jeweils einen Koeffizienten von 1 hat. Dies liegt daran, dass wir hier ein Feature mit sich selber vergleichen und dabei natürlich immer eine perfekte Korrelation besteht. Die Beispiele „Durchschnittliche Anzahl Räume vs. durchschnittliche Anzahl Schlafzimmer" und „Breiten- vs. Längengrad" stechen auch durch einen sehr hohen bzw. einen sehr niedrigen Koeffizienten heraus. Je weiter der Koeffizient von 0 entfernt ist, desto stärker ist die Korrelation.



Einblendung Diagramm California Housing Korrelationsanalyse

Ein Beispiel für eine schwache Korrelation ist hingegen die durchschnittliche Raumanzahl und das Medianeinkommen der Gegend, in der das Haus steht. Hier haben wir einen Korrelationskoeffizienten von 0.33, was in diesem Diagramm visualisiert wird.

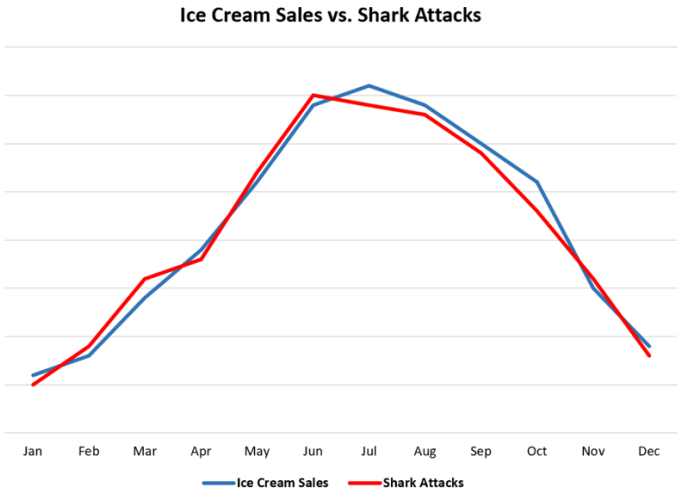


Einblendung Diagramm California Housing AveRooms vs. MedInc

Wenn du eine Korrelationsanalyse für deine eigenen Daten einsetzen möchtest, kannst du das mit Python erledigen. Viele Module, die statistische Funktionen beinhalten, bieten auch diverse Korrelationsanalysen für unterschiedliche Zusammenhangsarten und Anwendungen. In diesem Video ging es zum Beispiel nur um Korrelation zwischen zwei Features. Es kann aber auch Zusammenhänge mit mehreren beteiligten Features geben.

Abschluss

Denk aber bitte immer daran: Selbst, wenn Features zusammenhängen, lässt das noch keinen Rückschluss auf eine eventuell vorhandene Kausalität zu!



Einblendung Diagramm Ice Cream Sales vs. Shark Attacks (**Quelle [2]**)

Nur weil diese Graphen ähnlich aussehen, heißt das nicht, dass Eisverkäufe schuld an Haiangriffen sind.

In diesem Video haben wir uns noch einmal mit der Korrelation beschäftigt. Du kennst jetzt Korrelationsanalysen und kannst die entstandene Korrelationsmatrix interpretieren.

Quellen

Quelle [1] Kosfeld, R., Eckey, H. F. & Türck, M. (2016). *Deskriptive Statistik: Grundlagen - Methoden - Beispiele - Aufgaben*. Springer-Verlag.

Quelle [2] Bobbitt, Z. (2021). *Correlation does not imply causation: 5 Real-World examples*. Statology. https://www.statology.org/correlation-does-not-imply-causation-examples/

Weiterführendes Material

https://studyflix.de/suche?query=korrelation

https://realpython.com/numpy-scipy-pandas-correlation-python/

Disclaimer

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