Textübung: Korrelation versus Kausalität

Korrelation impliziert nicht Kausalität - diesem Merksatz bist du vielleicht schon mal begegnet. Trotzdem gibt es eine Menge Beispiele, in denen Medien, politische Entscheidungsträger oder Manager Kausalität behaupten, wo nur Korrelation festgestellt werden kann. In dieser Übung werden wir den Korrelations-Kausalitäts-Zusammenhang etwas genauer betrachten. Dazu schauen wir uns ein Beispiel an.

Korrelation impliziert nicht Kausalität

Betrachte die folgende Situation: In einem (fiktiven) Landkreis kommt es immer wieder zu Badeunfällen. Die Verwaltung möchte nun untersuchen, wie die Anzahl der Badeunfälle an lokalen Gewässern reduziert werden kann. Dazu beauftragt sie einen Data Analyst. Dieser präsentiert der Verwaltung eine Abbildung, die den Zusammenhang zwischen der Anzahl der täglichen Rettungseinsätze infolge von Badeunfällen und der Anzahl der verkauften Eiskugeln lokaler Eisverkäufer am selben Tag darstellt.

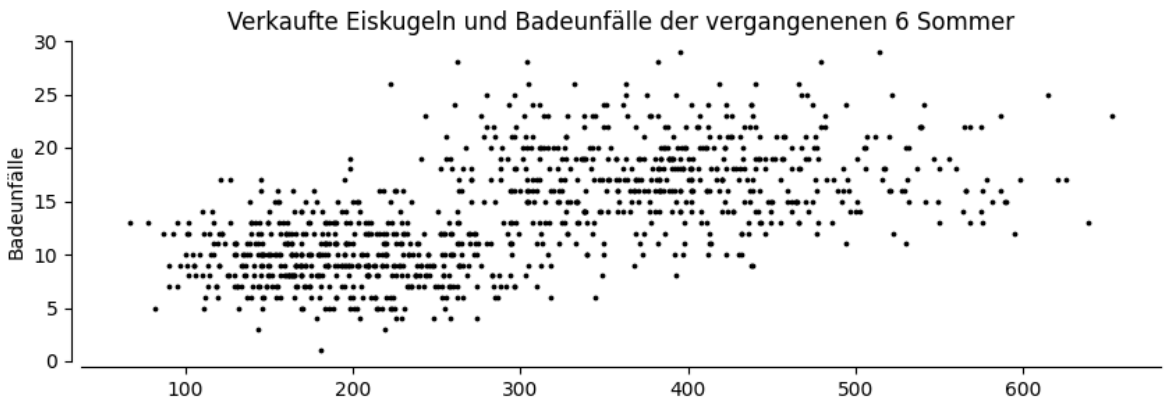

Abbildung 1: Streudiagramm der Daten der vergangenen sechs Sommer (fiktive Daten).

Die Abbildung zeigt einen *eindeutigen positiven linearen Zusammenhang* zwischen Badeunfällen und verkauften Eiskugeln. Würden wir eine Gerade so durch die Punkte legen, dass die Abstände zwischen Punkten und Geraden möglichst klein werden, so würde die Gerade nach oben zeigen und die Abstände wären recht klein. Damit liegt eine **hohe positive Korrelation** vor. Der Korrelationsindex gibt einerseits Auskunft über die Richtung des Zusammenhangs (positiv oder negativ), andererseits über dessen Stärke.

Wir sehen also: je höher der Eisabsatz, desto mehr Badeunfälle haben sich in der Vergangenheit zugetragen. Was sollte die Landkreisverwaltung deiner Meinung nach tun?

A) Den Verkauf von Speiseeis untersagen.
B) Vorsorgende Maßnahmen ergreifen, zum Beispiel könnte verstärkt vor Gefahren gewarnt werden, sobald sich morgens hohe Eisverkäufe abzeichnen.

C) Nichts. Eine hohe Korrelation hat keine praktischen Implikationen.

Die Vorstellung, dass der Eisabsatz Ursache für Badeunfälle sein könnte, erscheint abwegig. Es ist offensichtlich, dass hier **kein kausaler Zusammenhang besteht.** Daher wird ein Verbot von Eisverkäufen nicht den gewünschten Effekt haben. Die Korrelation alleine gibt uns also keine Hinweise darauf, *wie* wir tätig werden sollen. Sie hilft uns aber dabei, eine **Vorhersage** zu treffen**.** Oder anders ausgedrückt: Sie gibt einen Hinweis darauf, *wann* wir tätig werden sollen.

Das bedeutet: Die Eisverkäufe sind ein guter Prädiktor für die Gefahr von Badeunfällen. Zeichnet sich im Laufe eines Vormittags ein hoher Eisabsatz ab, sollten die Verantwortlichen zumindest gewarnt sein. Ob das verstärkte Warnen vor Gefahren das Problem mindern kann, ist eine Frage, die mit den vorliegenden Daten nicht zu beantworten ist. B) ist also die korrekte Antwort.

Merke: Korrelation bedeutet nicht automatisch Kausalität.

In vielen Anwendungsbeispielen können wir die Implikationen von Korrelation und Kausalität so zusammenfassen:

- Korrelation → Vorhersage (*wann* sollen wir reagieren?)
- Kausalität → Aktion (*wie* sollen wir reagieren?)

Korrelation kann durch eine gemeinsame Ursache entstehen

Woher rührt die hohe Korrelation? Eine naheliegende Hypothese ist, dass die Temperatur ein entscheidender Faktor ist. Hohe Temperaturen verleiten Menschen sowohl dazu, baden zu gehen, als auch dazu, Eis zu kaufen. Die folgende Abbildung veranschaulicht den Zusammenhang:

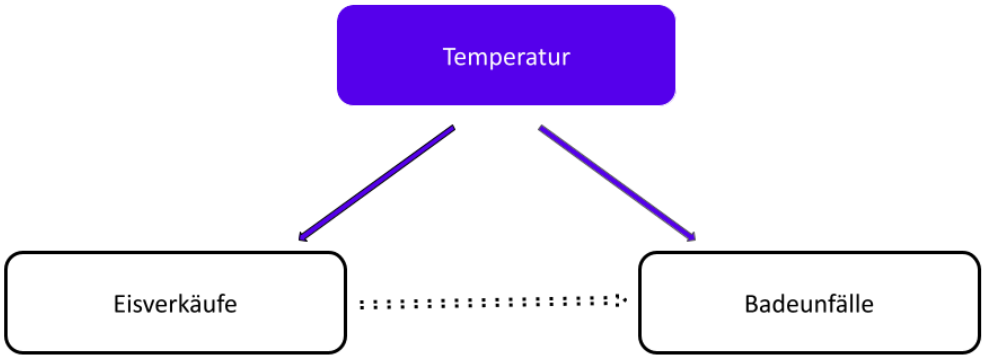

Abbildung 2: Darstellung der kausalen Zusammenhänge im Beispiel.

Schauen wir uns noch einmal das Streudiagramm von vorhin an. Diesmal machen wir zusätzlich die Temperaturen sichtbar, indem wir die Datenpunkte einfärben - rosa für hohe Tageshöchsttemperaturen, blau für niedrige Tageshöchsttemperaturen.

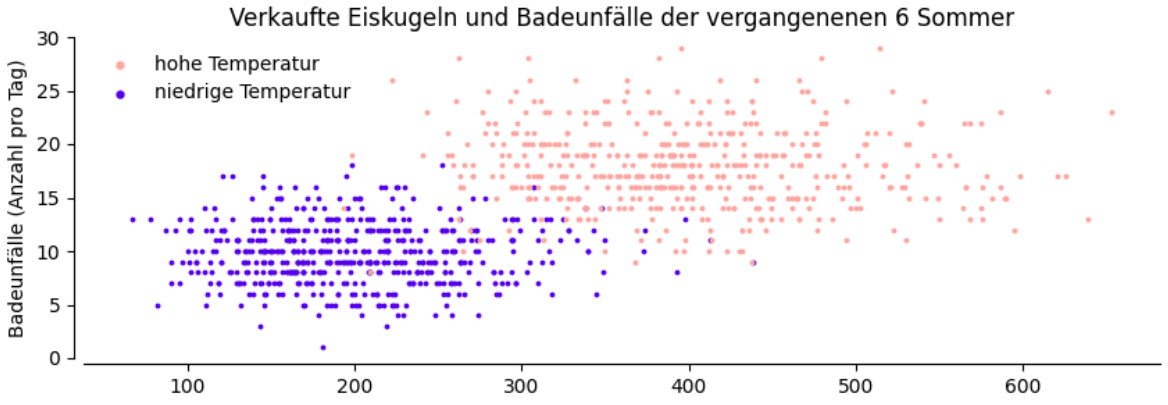

Abbildung 3: Streudiagramm nach Tageshöchsttemperaturen eingefärbt.

Bei der Betrachtung wird deutlich, dass der positive lineare Zusammenhang zwischen Eisverkäufen und Badeunfällen verschwindet, wenn wir die Daten auf Temperaturen bedingen. Die Korrelationen innerhalb der Gruppen 'hohe Temperatur' und 'niedrige Temperatur' sind nahezu Null.
Die Korrelation kommt nur über die **gemeinsame Ursache** hohe bzw. niedrige Temperatur zustande. Die gemeinsame Ursache ist ein möglicher Grund, wie Korrelation entstehen kann, ohne dass ein kausaler Zusammenhang vorliegt.

Wenige Menschen würden **eine Kausalität** zwischen Badeunfällen und Eisverkäufen vermuten. Unsere Intuition hat uns hier nicht getäuscht.
Manchmal tut sie das aber. Daher sollten wir vorsichtig sein, wenn wir Daten interpretieren.
Nimm dir doch ein wenig Zeit, um über die folgenden, weniger offensichtlichen Zusammenhänge nachzudenken. Entwerfe auch gerne eine Abbildung, wie du sie oben gesehen hast. Welche Schlüsse kannst du ziehen? Mögliche Lösungsansätze findest du ganz am Ende des Texts.

1) Bei der Analyse der Retouren in deinem eCommerce-Shop ist einem deiner Mitarbeitenden aufgefallen, dass diejenigen Kund:innen, die mit Paypal zahlen, dazu tendieren, Artikel eher zurückzusenden. Sollte deine Firma einen Rabatt auf andere Zahlungsmethoden einführen, um damit die Anzahl der Retouren zu reduzieren?
2) Deine Kollegin visualisiert den Zusammenhang zwischen der Nachfrage nach Übernachtungen in deinem Hotel und dem Übernachtungspreis, wobei jeder Datenpunkt einen Tag darstellt. Ihr stellt fest, dass es einen positiven linearen Zusammenhang, also eine starke Korrelation gibt. Hohe Preise implizieren also eine hohe Nachfrage. Sollten die Preise angehoben werden, um damit auch die Nachfrage zu steigern?
3) Die Google Anfragen 'Wein kaufen' in einem Monat korrelieren stark mit der allgemeinen Nachfrage nach Wein im Folgemonat. Macht es Sinn, auf dieser Basis die Einführung eines neuen Weins zu planen?

Praktisches Wissen: Actionable Insights

Oft möchten Entscheidungstragende in Unternehmen ihre Daten dazu nutzen, *actionable insights* zu produzieren, also ganz konkrete Vorschläge, wie der Gewinn gesteigert, die Kosten reduziert oder Prozesse optimiert werden können. Wir haben oben gesehen, dass solche Fragestellungen in der Regel kausaler Natur sind. Hier ist ein kleiner Leitfaden zu kausalen Analysen:

- Werden kausale Effekte anhand der bereits vorliegenden Daten abgeschätzt, sprechen wir von **kausaler Inferenz**. Data Analysten machen Annahmen über *Richtung und Natur* aller relevanten kausalen Zusammenhänge.
  In der Betrachtung Preis/Gewinn würde beispielsweise angenommen, dass der Preis ursächlich für Gewinn (und nicht andersherum) ist - das ist die Richtung.
  Ist der Zusammenhang linear? Wohl kaum. Das ist die Natur. Außerdem müssen Richtung und Natur aller anderen relevanten Variablen (wie den gemeinsamen Ursachen) korrekt angenommen werden. *Nur wenn alle Annahmen korrekt sind, ist auch die Schlussfolgerung korrekt.* Neuere Methoden zielen darauf ab, die Richtung von kausalen Zusammenhängen aus den Daten herauszufiltern. Für kausale Inferenz ist in der Regel ein sehr umfassender Datensatz vonnöten, der alle wichtigen Variablen enthält.
- Wird aus der kausalen Analyse eine Kausalität abgeleitet, oder geben die Daten eine tiefergehende Analyse gar nicht erst her, kommen oft **randomisierte Experimente** ins Spiel. Sie sind meistens teuer, aber sie sind die einzige Möglichkeit, kausale Effekte sicher identifizieren zu können. Oft sind sie nicht machbar, sei es weil sie ethisch nicht vertretbar sind, gesetzlich unzulässig, eine Randomisierung nicht möglich ist, oder sich schlicht der Aufwand nicht lohnt.

Kausalität impliziert auch nicht Korrelation

Wir haben gesehen, dass Korrelation nicht unbedingt Kausalität impliziert. Doch **Kausalität bedeutet auch nicht automatisch Korrelation.** Betrachte beispielsweise den Zusammenhang zwischen dem Preis des einzigen Produkts einer fiktiven Firma und deren Gewinn. Auch wenn dieser Zusammenhang in der Regel nicht bekannt sein dürfte, wird er im Normalfall eine hügelige Form aufweisen:

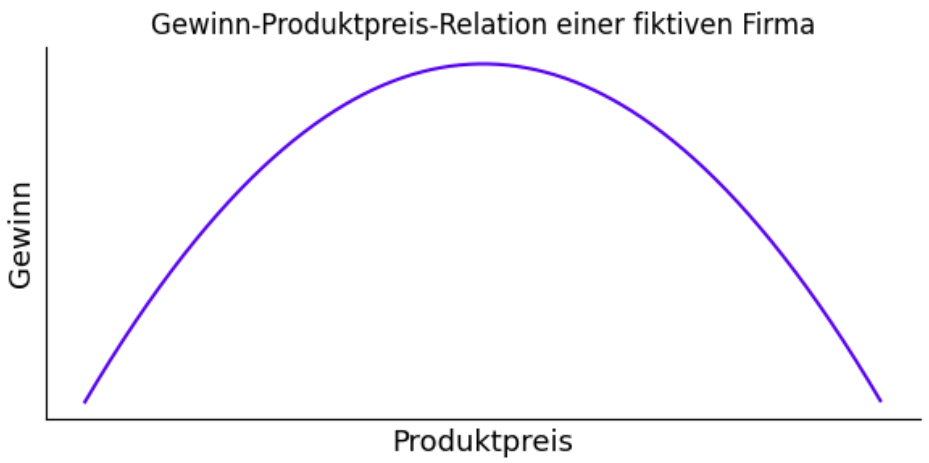

Abbildung 4: Zusammenhang zwischen Produktpreis und Firmengewinn.

Im allgemeinen Sprachgebrauch könnte die Korrelation als statistischer Zusammenhang jeglicher Form verstanden werden. Die statistische Korrelation hingegen ist ein Maß für den *linearen* Zusammenhang. Produktpreis und Gewinn stehen nicht in linearem Zusammenhang und sind daher auch nicht (statistisch) korreliert. Eine kausale Beziehung besteht aber sehr wohl. Eine Vorhersage ist auch möglich. Kausalität impliziert zwar nicht Korrelation, allerdings impliziert sie die Möglichkeit der Vorhersage. Oder anders ausgedrückt: Kausalität bedingt einen statistischen Zusammenhang.

Merke:

- Sobald die erklärende Variable (im Beispiel: Eisverkäufe) manuell geändert werden soll, um die zu erklärende Variable (Badeunfälle) zu beeinflussen, ist Vorsicht geboten. Um die Folgen absehen zu können, brauchen wir kausale Inferenz und/oder randomisierte Experimente.
- Eine Korrelationsanalyse ist im Vergleich hierzu unaufwändig. Eine hohe Korrelation bedeutet, dass eine Vorhersage möglich ist. Oft hilft uns das, zu entscheiden, wann gehandelt werden sollte.
- Ist die Korrelation zwischen zwei Variablen gering, bedeutet das nicht automatisch, dass eine Vorhersage nicht möglich ist, oder dass keine kausale Beziehung besteht.

Lösungen:

1) (Kauf auf Rechnung) Möglicherweise tendieren beispielsweise jüngere Menschen dazu, sowohl per Paypal zu zahlen, als auch mehr zu bestellen, als sie eigentlich brauchen. Ein Anreiz auf andere Zahlungsmethoden wird dann nicht den erwünschten Effekt (weniger Retouren) haben, sondern einfach nur den Gewinn schmälern.

2) (Preise) Das ist unwahrscheinlich. Wahrscheinlicher ist, dass sich Saisonalitäten hinter diesem Effekt verbergen. Eine hohe antizipierte Nachfrage an Feiertagen könnte zum Beispiel dafür gesorgt haben, dass die Mitarbeiter höhere Preise festgelegt haben, und die Gesamtnachfrage trotz der höheren Preise gestiegen ist. Der direkte kausale Zusammenhang zwischen Preis und Nachfrage ist jedoch wie üblich negativ.

3) (Wein) Absolut. Die Googlesuche ist ein guter Prädiktor für die Industrienachfrage, und je höher die Nachfrage, desto wahrscheinlicher ist der Erfolg für einen neuen Wein.