KI für Alle 2: Verstehen, Bewerten, Reflektieren

Themenblock Datenbeschaffung und -aufbereitung: 04_03Aufbereitung_Ausreisser_AS

Ausreißerbehandlung

Erarbeitet von
Dr. Katarina Boland und Dr. Ann-Kathrin Selker

Lernziele

- Du kannst erklären, was man unter Ausreißern versteht
- Du kannst Beispiele nennen, wie einfache Fehler in Daten automatisch erkannt werden können
- Du kannst die IQR-Methode zum Erkennen von Ausreißern anwenden

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

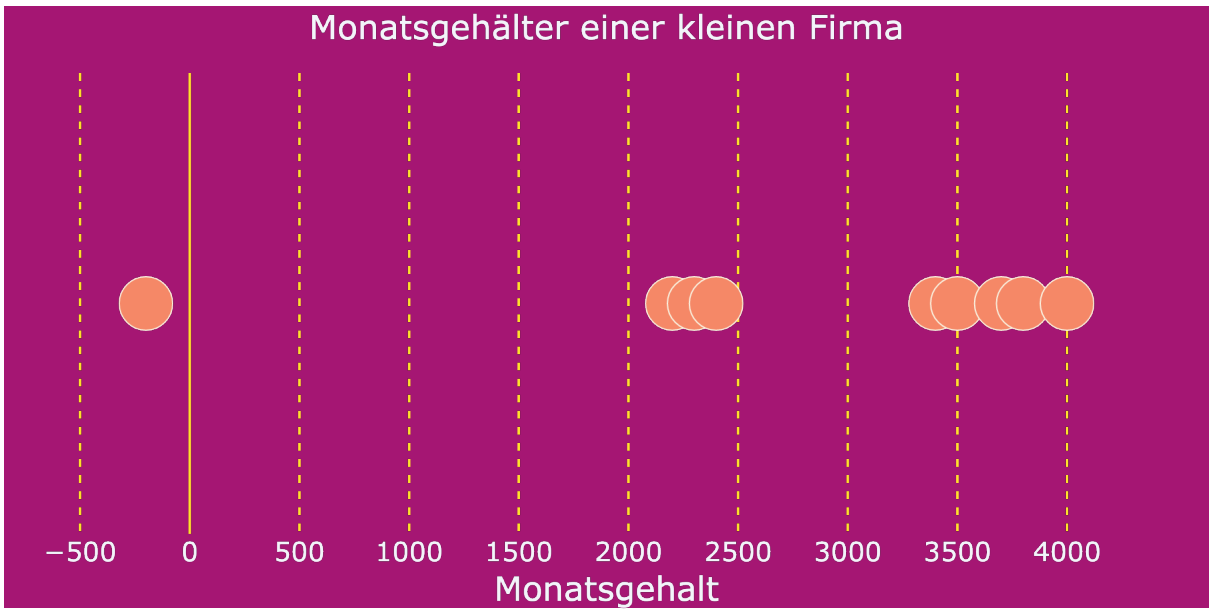
Inhalt

Einstieg

Was passiert eigentlich, wenn einzelne Datenpunkte stark von den anderen abweichen? Woran erkenne ich solche Ausreißer und wie gehe ich mit ihnen um?

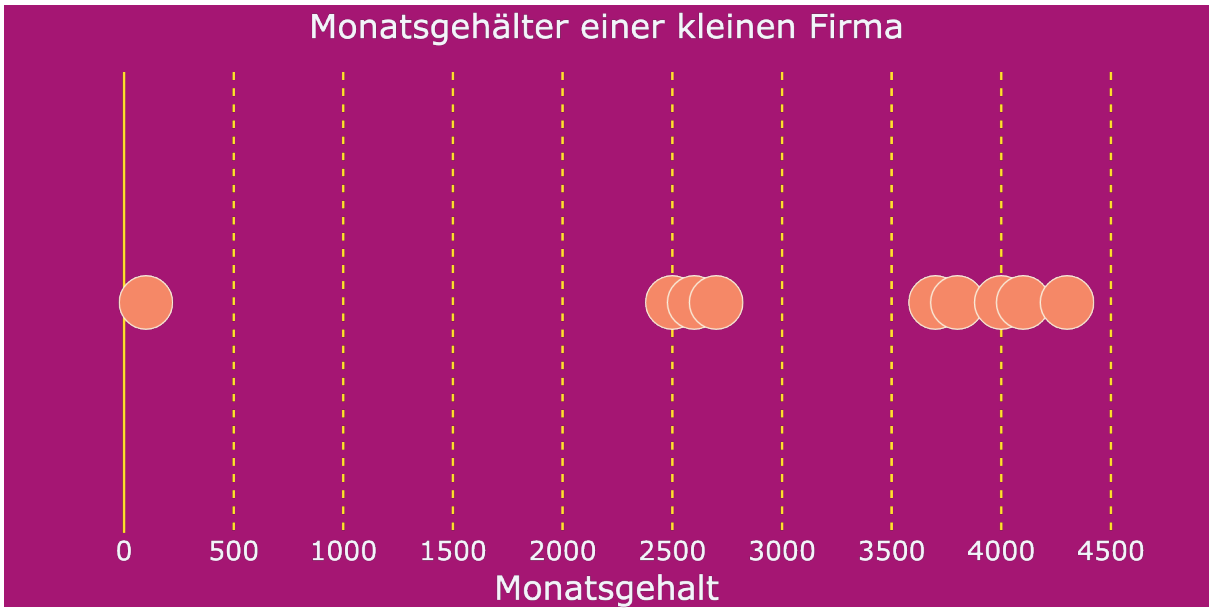
Bei der Generierung und Weiterverarbeitung von Daten können Fehler entstehen, z. B. durch fehlerhafte Messinstrumente, Fehler bei der Datenübertragung oder bei der Dateneingabe oder -konvertierung. Aber auch zufällige Schwankungen der Daten sind möglich. Diese Fehler oder Schwankungen bezeichnet man auch als „Rauschen" in den Daten.

Manche Datenpunkte können bei direkter Betrachtung als unplausibel erkannt werden, wie in diesem Diagramm der Gehälter einer kleinen Firma, in dem ein Monatsgehalt negativ ist.



Diese Art von Fehlern kann über einfache Prüfmechanismen erkannt werden, die auf Allgemein- oder Fachwissen beruhen. So können beispielsweise Regeln erstellt werden, dass Gehälter immer im positiven Bereich sein oder unterschiedliche Personen unterschiedliche Steuernummern haben müssen. Die Gültigkeit von Postleitzahlen kann über den Abgleich mit Tabellen geprüft werden usw.

Was aber, wenn die Verteilung so aussieht?

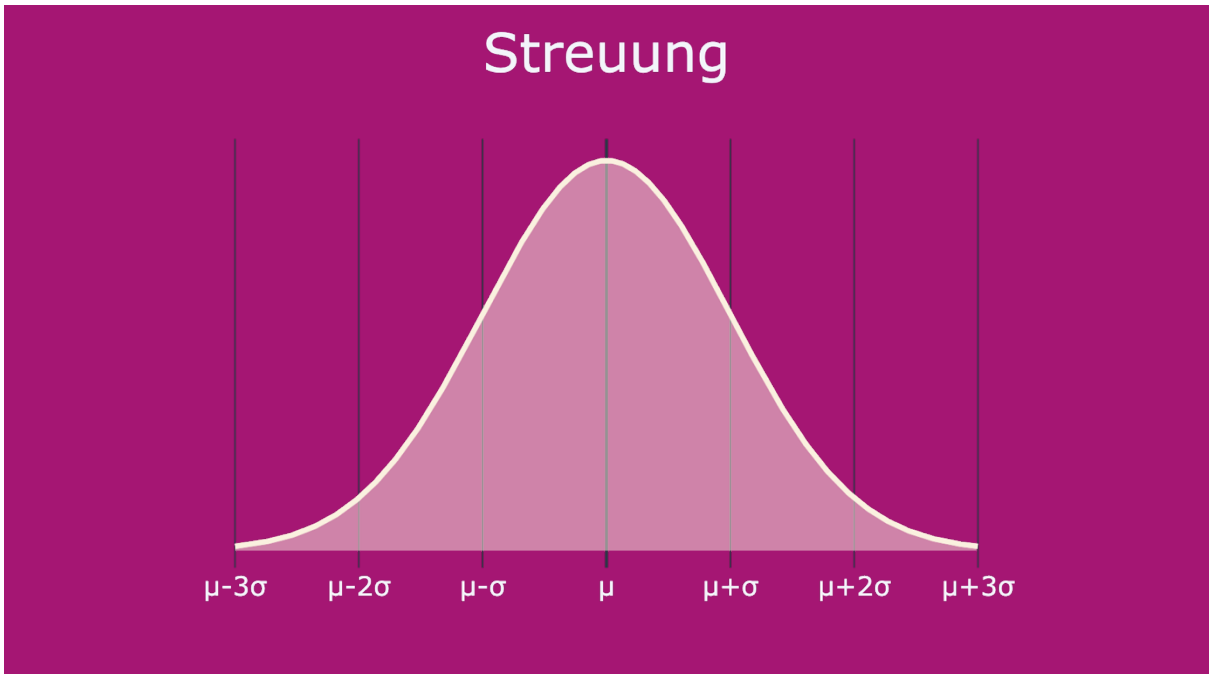


Die Gehälter sind nun alle im positiven Bereich, aber sieht der linke Wert nicht trotzdem verdächtig aus? Wie detektiert man so etwas nun automatisch und vor allem: Wie definiert man, ab wann ein Wert „verdächtig" ist? Und was bedeutet dies in Bezug auf die Daten selbst und für das Training von Modellen?

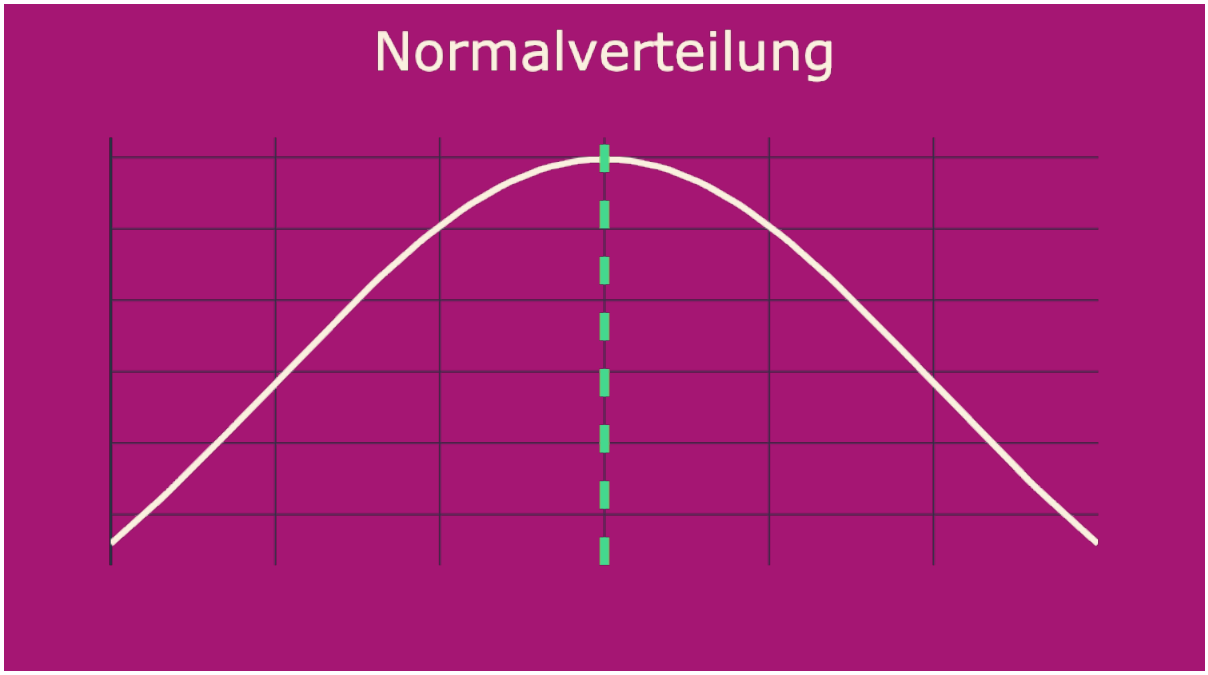
Hierfür müssen wir uns die Verteilung der Datenpunkte ansehen. Was sind erwartbare Werte und was sind Ausreißer, also Werte, die außerhalb des typischen oder vielmehr erwartbaren Bereichs liegen?

Detektion von Ausreißern

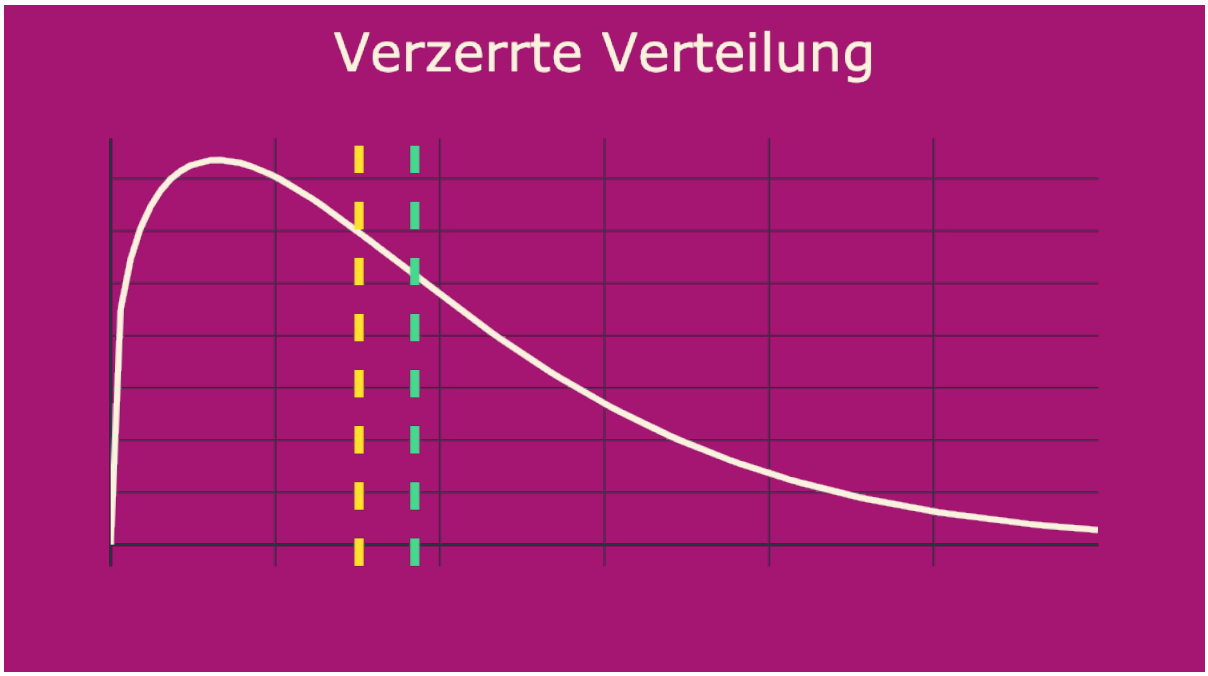
Zur Erinnerung: In einer Normalverteilung befinden sich rund 99.7 % aller Messwerte im Bereich mit bis zu dreifacher Standardabweichung vom Mittelwert.



Wir können nun also sagen, dass alles, was außerhalb dieses Bereichs liegt, sehr stark von den anderen Daten abweicht und diese Messwerte als Ausreißer betrachten. Doch was ist, wenn unsere Verteilung verzerrt ist?
Bei einer perfekt symmetrischen Normalverteilung entsprechen arithmetisches Mittel und Median demselben Wert.

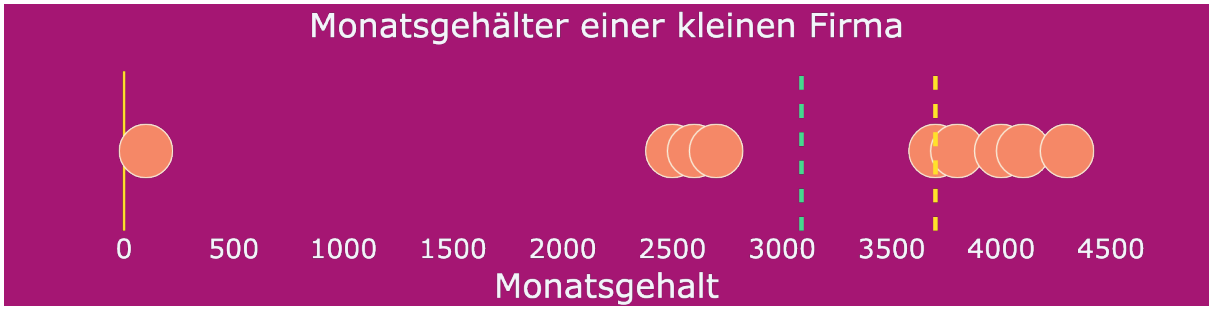


Ist die Verteilung verzerrt, ist dies nicht der Fall.

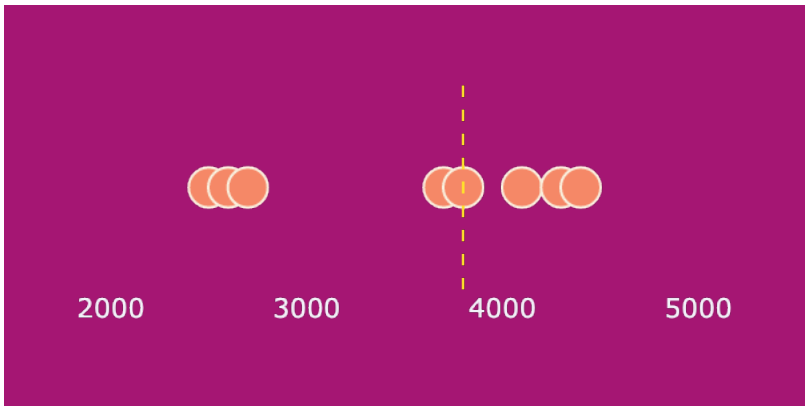


Wie man hier sieht, weicht das arithmetische Mittel dann weiter vom Höhepunkt der Kurve ab, als der Median.

Als Beispiel können wir uns noch einmal unsere Gehälter anschauen.

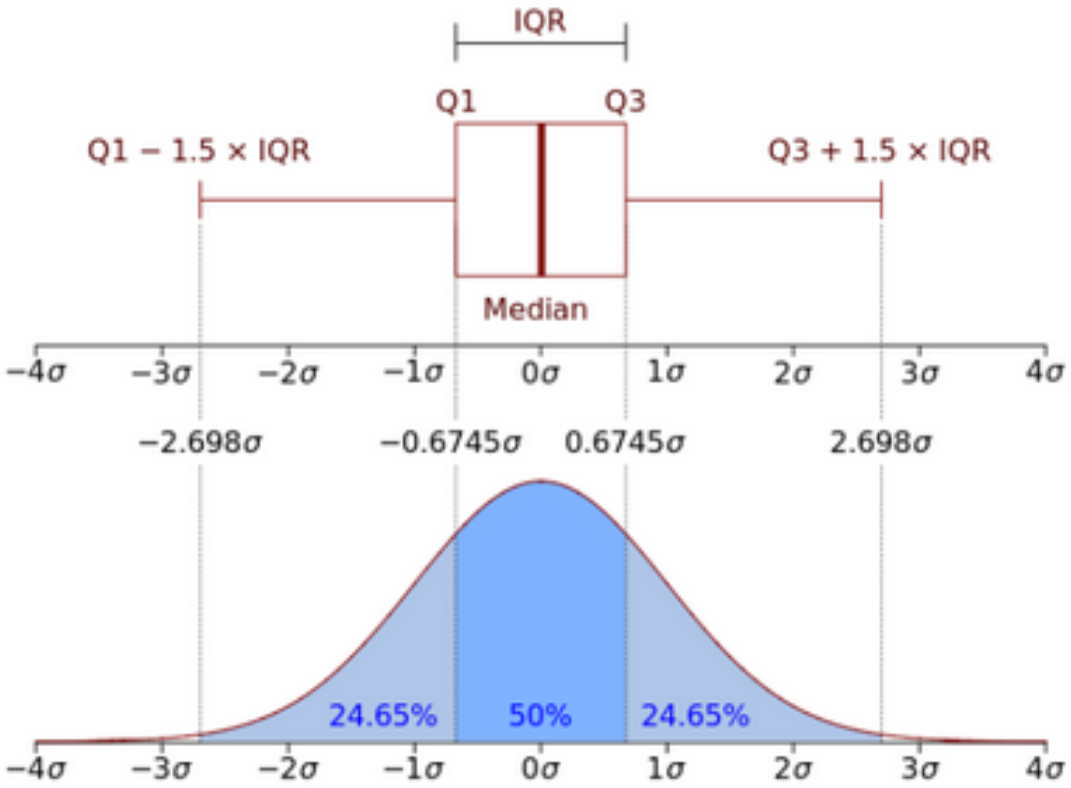


Der arithmetische Mittelwert liegt hier bei etwa 3.089 Euro und der Median bei 3.700 Euro, die Gehälter sind also nicht normalverteilt. Wenn wir jetzt unseren letzten Datenpunkt durch das Gehalt des Geschäftsführers ersetzen, der unglaubliche 10 Millionen Euro verdient, verändert sich der Median nicht, der Mittelwert liegt aber plötzlich bei über einer Million Euro.



Wenn wir nun Verteilungen untersuchen wollen, bei denen wir davon ausgehen, dass sie Verzerrungen durch Ausreißer enthalten könnten, dann wäre es vorteilhaft, weniger verzerrungsanfällige Maße einzusetzen. Das erreichen wir, indem wir statt dem Prinzip der Standardabweichung, das den Mittelwert benutzt, stattdessen ein Maß unter Berücksichtigung des Medianprinzips verwenden, die sogenannte Inter-Quartile Range Regel (IQR).

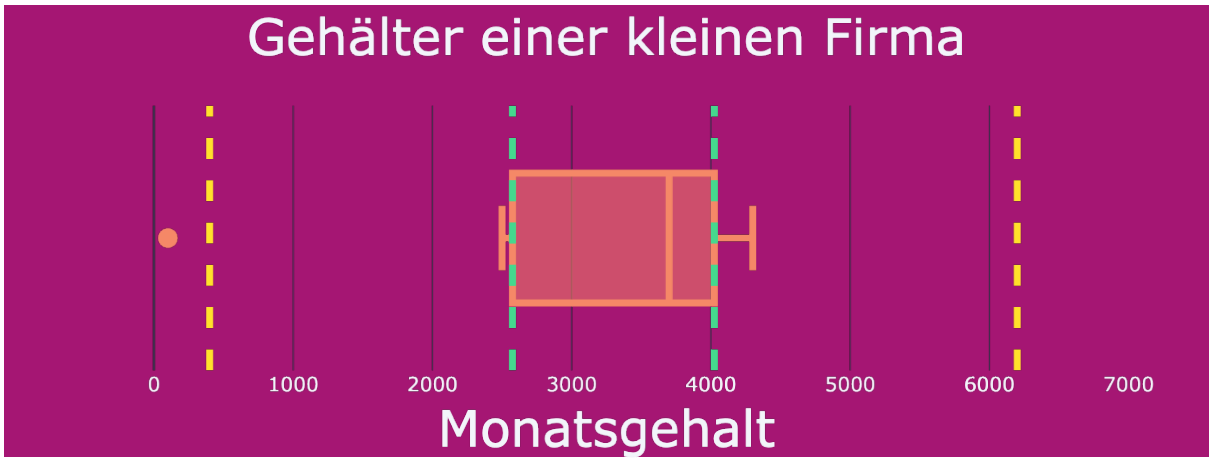
Das Prinzip der IQR lässt sich gut durch einen Boxplot veranschaulichen.



Quelle [1]

Zur Erinnerung: Der Bereich, der in der Box des Boxplots liegt, enthält 50 % aller Ergebnisse, nämlich alle Werte, die zwischen dem 25 %-Quantil und dem 75 %-Quantil liegen. Diese Quantile nennt man übrigens auch Q1 (Quartil 1) bzw. Q3 (Quartil 3). Die Breite dieser Box, also Q3 minus Q1, bezeichnen wir dann als Inter-Quartile-Range. Alle Werte, die nicht mehr als das 1.5-fache der Inter-Quartile-Range von unserer Box entfernt sind, sind dann erwartbare Werte. Die 1.5 sind dabei so gewählt, dass mit der IQR-Regel bei einer Normalverteilung in etwa der gleiche Bereich als erwartbare Werte bezeichnet werden, wie wenn wir das alte Maß der dreifachen Standardabweichung verwenden würden.

In unserem Gehälterbeispiel liegt das linke Ende der Box bei 2525, das rechte Ende bei 4025. Unsere Inter-Quartile-Range beträgt also 1.5 * (4025 - 2525) = 2250. Alle Werte, die mehr als 2.250 Euro unter dem linken Ende der Box bzw. über dem rechten Ende der Box liegen, werden dann als Ausreißer deklariert. Im Beispiel wird klar, dass unser Ausreißergehalt von 100 Euro pro Monat unter dieser Grenze liegt und damit auch mit der IQR-Regel als Ausreißer markiert wird.



Du kannst jetzt Datenpunkte als Ausreißer erkennen, aber was bedeutet dies nun und wie gehst Du am besten mit den Ausreißern um?

Bereinigung von verrauschten Daten

Während Fehler und Ungenauigkeiten in Daten grundsätzlich die Qualität von Modellen beeinträchtigen können, die auf ihnen trainiert werden, können Ausreißer als besonders problematisch angesehen werden: Durch die extreme Abweichung von den restlichen Daten kann ihr Einfluss besonders groß sein, zum Beispiel dadurch, dass sie den arithmetischen Mittelwert stark verzerren. Methoden, die sich auf diesen Mittelwert verlassen, funktionieren unter Umständen bei Vorhandensein von Ausreißern deutlich schlechter. Dies trifft beispielsweise auf den k-means-Algorithmus zu.

Es kann also sinnvoll sein, Ausreißer aus den Daten zu entfernen. Aber Achtung: Ausreißer können durch Fehler entstehen, sie können aber auch korrekte Beobachtungen darstellen. Es kann trotzdem sinnvoll sein, sie für Training und Evaluation zu ignorieren, da selten

auftretende und für den Anwendungsfall irrelevante Spezialfälle das Erlernen von allgemeingültigen Regeln behindern können. Um dies beurteilen zu können, ist es wichtig, die Entstehung und Bedeutung der Ausreißer für den individuellen Datensatz und Einsatzzweck zu verstehen.

Dass Ausreißer auch korrekte Beobachtungen darstellen können, zeigt das Beispiel der Entdeckung des Lochs in der Ozonschicht über der Antarktis in den 1980er Jahren. Die NASA-Software, die die von Satelliten gesammelten atmosphärischen Daten auswertete, soll wohl Daten bezüglich der stark fallenden Ozonwerte der Ozonschicht als Ausreißer deklariert und herausgefiltert haben.

Quelle [2]

Daher tauchte diese drastische Änderung nicht in den veröffentlichten Berichten auf und das Loch in der Ozonschicht konnte erst Jahre später nachgewiesen werden, als Joe Farman zusammen mit Brian Gardiner und Jon Shanklin einen entsprechenden Artikel veröffentlichte.

Quelle [3]

Ihre Ergebnisse bezogen sie von einem eigenen Sensor in der Antarktis.

Über die Behandlung von Ausreißern hinaus gibt es viele verschiedene Techniken, um Rauschen in Daten zu minimieren. Zu diesen Techniken gehören z. B. Normalisierung und Diskretisierung. Bevor diese Techniken allerdings angewendet werden, solltest du dich informiert entscheiden, wie du mit vorhandenen Ausreißern umgehst. Die Wahl der Ausreißerbehandlung beeinflusst nämlich auch die Ergebnisse der weiteren Schritte.

Quellen

Quelle [1] Jhguch at en.wikipedia (9. März 2011). Boxplot vs PDF. CC BY-SA 2.5. Image cropped https://commons.wikimedia.org/wiki/File:Boxplot_vs_PDF.svg

Quelle [2] Brain, T. (2018). The Environment is not a system. *A Peer-Reviewed Journal About*, *7*, 152-165. https://doi.org/10.7146/aprja.v7i1.116062

Quelle [3] Farman, J. C., Gardiner, B. G., & Shanklin, J. D. (1985). Large losses of total ozone in Antarctica reveal seasonal ClO x/NO x interaction. *Nature*, *315*(6016), 207-210. https://doi.org/10.1038/315207a0

Weiterführendes Material

https://scikit-learn.org/stable/modules/outlier_detection.html

https://www.geeksforgeeks.org/detect-and-remove-the-outliers-using-python/

https://www.kaggle.com/discussions/general/241610

Disclaimer

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