educX

# Anfertigen eines CAT bzw. PIKE

**Kann … // Algorithmus x // exact … // (Kunden)-Problem … berechnen / lösen?**

1. Ja, man kann in diesen Daten durch Histogramme einen guten Eindruck darüber gewinnen, wie viele Schüler pro Schule an der Umfrage teilgenommen haben, in welchem Alter die Schüler waren, ob in etwa gleich viele Mädchen und Jungen teilgenommen haben und ob die Teilnehmer gute oder schlechte Noten haben.
2. Ja, man kann via Korrelation zeigen, welche Daten sich miteinander, neutral oder gegeneinander verändern und so erkennen, dass es einen Zusammenhang zwischen den Noten, der wöchentlichen Lernzeit und dem Wunsch nach einer Hochschulausbildung gibt, sowie, dass die Noten und der Alkoholkonsum in einer entgegengesetzten Weise korrelieren.

**Data Science Kernaussage:**

**(P) roblem):**
**[Welcher Frage hat für die Lösung des Kunden / Auftraggeber die größte Bedeutung?]:**

1. Können Daten einer Umfrage so visualisiert werden, dass man auf einen Blick ein Verständnis für die Daten entwickeln kann?
2. Kann man in den Daten einer Umfrage Zusammenhänge zwischen dem Umfeld der Schüler, ihren Noten und Alkoholkonsum erkennen?

**(I) ntervention:**
**(Bibliotheken- und Algorithmen-Auswahl, … z.B. pandas für Finanzdaten … )**
**[Welche Berechnung erwäge ich vornehmlich?]:**

1. Pandas zum laden der CSV Daten in einen Dataframe, sowie zur Korrelation der Daten miteinander.
    1. Dabei wird die Korrelation auf alle möglichen Spalten-Paare angewendet damit man heraus finden kann, welche Spalten eine Beziehung zueinander haben könnten.
2. Seaborn und Matplotlib zur Visualisierung der Daten in Diagrammen.
3. Counter um ein paar Textuelle Zusammenfassungen der Anzahl an bestimmten Antworten anzuzeigen.
    1. Hauptsächlich um zu berechnen, wie viele Prozent der Befragten X oder Y geantwortet haben.

**(K) ontrollintervention**
**(falls erforderlich: Bliotheken- und Algorithmen-Auswahl … z.B. scikit-learn für Finanzdaten … )**
**[Was ist die andere Möglichkeit?]:**
Man könnte anstelle einer Korrelation des gesamten Datensatzes auch ein paar Vermutungen aufstellen und die Daten nach mehreren Bedingungen filtern und die Ergebnisse dann miteinander Vergleichen. Zum Beispiel könnte man einen Zusammenhang zwischen Noten und Alkoholkonsum vermuten und dann alle Zeilen finden, in denen viel und alle, in denen wenig Alkoholkonsum angegeben wurde und in diesen zwei Teil-Datensätzen dann die Notenverteilung plotten um zu sehen, ob eine der beiden Gruppen bessere oder schlechtere Noten hat. Dies könnte man dann für weitere Vermutungen, wie zum Beispiel „Bildung der Eltern und Noten der Schüler“ machen.

**(E) rgebnismaß (Zielgröße(n)) – Die Evidence**
**[Was möchte ich / der Kunde erreichen? Z.B. Prädiktor oder Klassifikator erstellen … ]:**

1. In ein paar Diagrammen mehr über die Umstände der Befragten erfahren.
2. Aufzeigen welche dieser Umstände den größten Einfluss auf die schulischen Leistungen und den Alkoholkonsum haben, bzw. wie stark sich Alkoholkonsum auf die Noten auswirkt oder ob andere Faktoren relevanter sind.

**Anmerkungen**
**Literaturhinweise**
https://www.kaggle.com/datasets/gabrielluizone/high-school-alcoholism-and-academic-performance

# Die Suche nach der besten Evidenz

## 1. Problem

Die Korrelation mit der Tabelle, in denen die Antworten als Text enthalten waren, war nicht möglich, auch wenn diese Tabelle für die Erstellung der Diagramme sehr gut war, da man so gleich die Achsen-Beschriftungen mit bekommt. Daher wurde für die Korrelation auf den klassifizierten Datensatz zurück gegriffen. Dieser hat jedoch keine Legende, wodurch nicht offensichtlich ist, ob bei der positiven Korrelation von zum Beispiel Geschlecht und Lernzeit damit nun gemeint ist, ob Frauen oder Männer mehr lernen.
Man müsste also noch die beiden Datensätze vergleichen um zu erfahren, welche der beiden Gruppen nun mehr lernt.

## 2. Definition einer wichtigen suchbaren Frage

Kann ein Zusammenhang zwischen Alkoholkonsum und schulischer Leistung hergestellt werden oder sind andere Einflüsse bedeutender?

## 3. Auswahl der wahrscheinlichsten Quelle für diese Evidenz

1. Diagramme und textuelle Zusammenfassungen erstellen und begutachten.
2. Daten miteinander korrelieren, in einem Diagramm darstellen, die Paare mit der Höchsten positiven und negativen Korrelation ausgeben und begutachten. (Bei der Textuellen Ausgabe vorher die Diagonale/Selbstkorrelation herausfiltern, da diese keinen Mehrwert hat.)

## 4. Erstellung einer Suchstrategie

1. Datensätze mit Pandas laden und mit Seaborn und Matplotlib darstellen
2. Erstellung einer Korrelation des gesamten Datensatzes, Darstellung der Korrelation, sowie Ausgabe der größten positiven und negativen Korrelation.

## 5.0 Zusammenstellung der Evidenzausbeute

1. Die grafische Darstellung bietet einen sehr guten Überblick darüber, welche Umstände sich am stärksten auf die Noten auswirken.
2. Die textuelle Darstellung bietet einen sehr guten Überblick darüber, welche die größten positiven und negativen Einflüsse sind.
    1. Positiv:
        1. Desire Graduate Education
        2. Weekly Study Time
        3. Mother Education
        4. Father Education
    2. Negativ:
        1. School
        2. Alcohol Weekdays
        3. Mother Work
        4. Age
        5. Housing Type
        6. Commute Time

## 6. Anwendung der Evidenz

1. Die Visualisierung der Daten in Histogrammen erlaubt einen guten Überblick und kann auch bei anderen Datensätzen angewendet werden.
2. Die Korrelation der Daten miteinander, sowie das herausfiltern relevanter Teile dieser Korrelation, erlaubt das einfache identifizieren der relevantesten Einflüsse auf ausgewählte Spalten dieser Daten. Diese Methode kann auch bei anderen ähnlichen Daten angewendet werden.