ECKERT, Jakim; SCHÖNBRODT, Sarah & FRANK, Martin
Karlsruhe, Salzburg, Karlsruhe

Mathematische Grundlagen des Data Cleanings

Data Cleaning spielt eine wichtige Rolle in Data Science (DS) Prozessen. Wie mit fehlerhaften oder fehlenden Daten umgegangen wird, kann Erkenntnisse, Vorhersagen und abgeleitete Entscheidungen aus DS-Prozessen maßgeblich beeinflussen. Gleichzeitig sind diese Methoden mathematisch äußerst reichhaltig. DS-Bildung und die frühe Förderung von Data Literacy haben in den letzten Jahren in der Mathematik- und Informatikdidaktik zunehmend an Bedeutung gewonnen (Fleischer et al., 2022; Schüller et al., 2021). Das vorgestellte Design-Based Research-Projekt untersucht, wie die mathematischen Grundlagen von Methoden des Data Cleanings für Schüler*innen der Sekundarstufen zugänglich gemacht werden können.

Data Cleaning als Unterrichtsgegenstand

In früheren Arbeiten wurde Data Cleaning als notwendiger Bestandteil von DS-Projekten für Schüler*innen beurteilt und aufgegriffen (Fleischer et al., 2022; Markulin et al., 2021). Im Vergleich zu diesen Studien setzen wir einen neuen Schwerpunkt auf die mathematischen Grundlagen. Zentrale Aufgaben des Data Cleanings sind die Ausreißeridentifikation, der Umgang mit fehlenden Daten und die Datenanreicherung. Es können verschiedene Imputationsverfahren (z. B. Mittelwerts- oder Regressionsimputationen) verwendet werden, um mit fehlenden Daten umzugehen und Datenlücken zu füllen.

In diesem Forschungsprojekt fokussieren wir uns auf statistische Verfahren zur Ausreißeridentifikation, wie den Dixon- oder Grubbs-Test, die auf schulmathematische Inhalte (wie Lage- und Streumaße) zurückgreifen und darüber hinausgehen. Es soll herausgearbeitet werden, welche Methoden und mathematischen Grundlagen der Ausreißeridentifikation Lernenden anhand realer Daten zugänglich gemacht werden können und welche didaktischen Ansätze dafür geeignet sind. Darüber hinaus soll betrachtet werden, wie der Einfluss von Data Cleaning auf DS-Prozesse und damit einhergehende Auswirkungen auf reale Fragestellungen vermittelt werden kann.

Literatur

Fleischer, Y., Biehler, R. & Schulte, C. (2022). Teaching and Learning data-driven Machine Learning with educationally designed Jupyter Notebooks. *SERJ, 21*(2), 1-12. https://doi.org/10.52041/serj.v21i2.61

Markulin, K., Bosch, M. & Florensa, I. (2021). Project-based learning in Statistics: A critical analysis. *Caminhos da Educação Matemática em Revista, 11*(1), 200-220.

Schüller, K., Koch, H. & Rampelt, F. (2021). *Data-Literacy-Charta*. https://www.stifterverband.org/sites/default/files/ data-literacy-charter.pdf

In: P. Ebers, F. Rösken, B. Barzel, A. Büchter, F. Schacht & P. Scherer (Hrsg.),
Beiträge zum Mathematikunterricht 2024.
57. Jahrestagung der Gesellschaft für Didaktik der Mathematik. WTM.
https://doi.org/10.37626/GA9783959872782.0

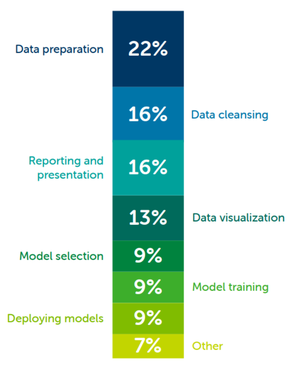
Mathematische Grundlagen des Data Cleanings

Design-Based-Research-Projekt zu mathematischen Methoden der Ausreißeridentifikation

Jakim Eckert, Sarah Schönbrodt, Martin Frank

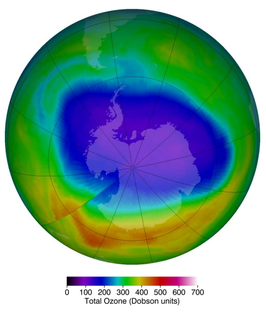
Kontext

Bei der Entwicklung von zahlreichen Anwendungen aus Alltag, Technik und Forschung (bspw. KI-Systeme) nutzt man häufig große Datensätze zum Trainieren. Allerdings sollten diese Datensätze eine gewisse Qualität besitzen. Während der Data Preparation und dem Data Cleaning wird der Datensatz daher auf potenzielle Fehler und Unstimmigkeiten untersucht.

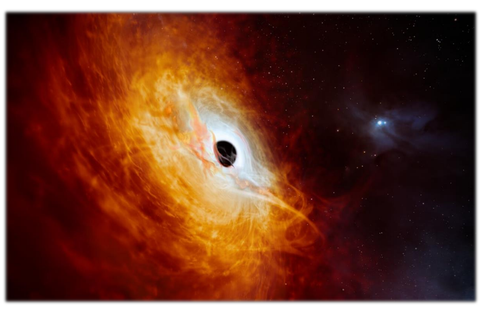


How do data scientists spend their time?
(Anaconda Report)

- Data Cleaning und Data Preparation machen 37,75% der Arbeitszeit von Data Scientists aus.
- Immer wieder werden Ausreißer fälschlicherweise aus Datensätzen herausgefiltert (bspw. Entdeckung Ozonloch, und Quasar).



Ozonschicht über der Antarktis
(NASA Ozone Watch)



Künstlerische Darstellung des Quasars J0529-4351
(Pressemitteilung der Europäischen Südsternwarte - ESO)

Mathematische Aspekte des Data Cleanings

- *Umgang mit fehlenden Daten*
  - Mittelwert Imputationen
  - Regressionsimputationen
  - Hot deck imputation
- *Ausreißeridentifikation*
  - Boxplots, Lage- und Streumaße
  - Statistische Tests (bspw. Grubbs und Dixon)
  - Teststatistiken (bspw. Mahalanobisdistanz)

Data Cleaning in der Mathematikdidaktik

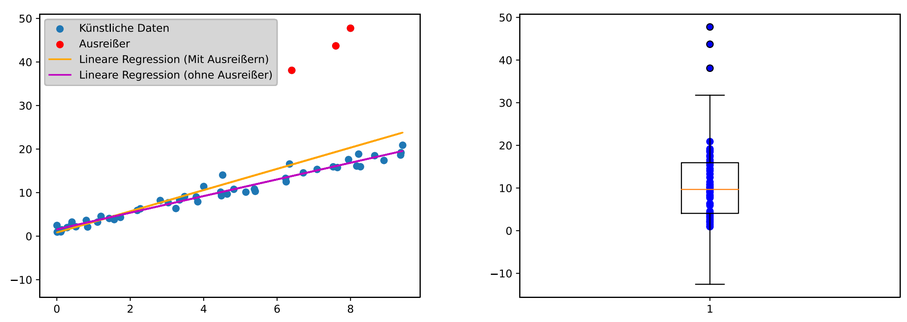
- Kaum deutschsprachigen Materialien zu mathematischen Grundlagen des Data Cleanings für den Schulgebrauch vorhanden
- Teilaspekt in vereinzelten Data Science Projekten thematisiert - ProDaBi[1], Markulin et al.[2] und CAMMP[3]

Ziel des Forschungsvorhabens

- Entwicklung von Lehr- und Lernmaterialien mit Fokus auf den mathematischen Aspekten des Data Cleanings
- Methoden und mathematische Grundlagen des Data Cleanings anhand realer Daten zugänglich machen und geeignete didaktische Ansätze identifizieren
- Vermittlung des Einflusses von Data Cleaning auf Data Science-Prozesse

Schulische Anknüpfungen

- Lineare Regression und Boxplots



- Prozessbezogen
  - Reflektierter Umgang mit mathematischen Verfahren und Hilfsmitteln
  - Einschätzung der Eignung von mathematischen Verfahren und Modellen hinsichtlich der Realität
  - Probleme analysieren, Strategien auswählen, anwenden und daraus einen Plan zur Lösung entwickeln
- Außercurricular[4]
  - Evaluation der Qualität von Datensätzen (data literacy)
  - Reflektierte statistische Datenanalyse (statistical literacy & data literacy)

Ausblick

- Aktuelle Materialentwicklungen:
  - Digitale Arbeitsmaterialien in Jupyter Notebooks
  - Betrachtung von Vorhersagen zu dem Datensatz mit und ohne Ausreißer
  - Vergleich verschiedener Identifikationsmethoden
  - Auf Grundlage eines realen Datensatzes
- Erste Materialerprobungen zur Ausreißeridentifikation im April

Jakim Eckert
jakim.eckert@kit.edu
Scientific Computing Center (SCC),
KIT, Karlsruhe

cammp.online/108.php

Referenzen:

1) Fleischer, Y., Biehler, R. & Schulte, C. (2022). Teaching and Learning data-driven Machine Learning with educationally designed Jupyter Notebooks. Statistics Education Research Journal, 21(2), 7. https://doi.org/10.52041/serj.v21i2.61
2) Markulin, K., Bosch, M., & Florensa, I. (2021). Project-based learning in Statistics: A critical analysis. Caminhos da Educação Matemática em Revista, 11(1), 200-220.
3) Steffen, N.: Sicherheit der Privatsphäre in sozialen Netzwerken - Wie Mathematik die Nutzer ausspioniert. Ein Lehr-Lern-Modul im Rahmen eines mathematischen Modellierungstages für Schülerinnen und Schüler der Sekundarstufe I, Masterarbeit, RWTH Aachen, 2019.
4) Biehler, R. et al. (2018): Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts. Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374