DATENSCHULE

DATENANALYSE IN GOOGLE SHEETS

OPEN KNOWLEDGE FOUNDATION DEUTSCHLAND

Mehr zu unseren Projekten & Workshops:
datenschule.de
E-Mail: info@datenschule.de |
Telefon: 030-57703666-2

DATENANALYSE IN GOOGLE SHEETS

INHALTSVERZEICHNIS

Durch die Analyse von Daten erhalten wir Antworten auf bestimmte Fragen, testen unsere Hypothesen oder bekommen neue Anhaltspunkte für weitere Forschungen. Es gibt eine Fülle an verschiedenen Methoden zur Datenanalyse, die von deskriptiven und inferentiellen Statistiken bis zur Wahrscheinlichkeitsanalyse reichen. Dieses Lernmaterial bietet allgemeine Ansätze zur Analyse eines Datensatzes sowie Links zu nützlichen Programmen und Ressourcen, die sich für dein Projekt als hilfreich erweisen könnten.

1. Datenanalyse

Ein Gefühl für die Daten bekommen:

Wenn du mit einer Datenanalyse beginnst, ist es wichtig, zunächst ein Gefühl für die Informationen innerhalb des Datensatzes zu bekommen. Dabei kannst du dir z. B. folgende Fragen stellen:

- Wie viele Beobachtungen befinden sich in meinem Datensatz (Anzahl der Zeilen), und wie viele Spalten sind enthalten?
- Enthält der Datensatz Zeitreihendaten?
- Erkennst du allgemeine Trends in den Daten (siehe hierzu auch: Maße der zentralen Tendenz, S. 3)
- Gibt es Ausreißer, deren Werte besonders hoch oder niedrig sind; treten sie am häufigsten oder am wenigsten auf?

Wichtig sind auch folgende Kontextinformationen:

- Wer hat den Datensatz erstellt?
- Wann wurde er erstellt?
- Warum wurde er erstellt?

Diese Fragen sind auch hilfreich zur Überprüfung der Daten.

Weitere Fragen könnten sein:

- Wie hängen Spalten zusammen?
- Können wir die Hauptergebnisse mit dem Thema des Datensatzes in Beziehung setzen?
- Beweisen oder widerlegen unsere Ergebnisse etablierte Fakten oder Theorien?

2. Maße der zentralen Tendenz

Maße der zentralen Tendenz sind zusammenfassende Statistiken, die darauf abzielen, eine Menge von Daten in einer bestimmten Anzahl zu beschreiben. Die wichtigsten Statistiken werden im Folgenden beschrieben.

Mittelwert - der Durchschnitt:

- Was ist der durchschnittliche EU-Zuschuss, den ein Mitgliedstaat erhält?
  =AVERAGE(A1:A29)
- Alle Datenpunkte dividiert durch die Anzahl der Beobachtungen

Median - der Wert in der Mitte:

- der Wert, der den Datensatz in zwei gleiche Hälften teilt (50/50)
- Welcher Wert liegt genau in der Mitte der Verteilung?
  = MEDIAN(A1: A29)
- Sehr nützlich für Einkommensdatensätze - hilft, mögliche Verzerrungen beim Einkommen zu identifizieren

Modus - der häufigste Wert:

- Datenpunkt, der im Datensatz am häufigsten vorkommt
  = MODE(A1: A29)
- Dies ist nur sinnvoll bei absoluten Zahlen oder geringen Nachkommastellen
- Zum Beispiel Schulnoten: Welches ist die häufigste Note?

2. Maße der zentralen Tendenz

Streuungsmaße:

Minimum: niedrigster Wert im Dataset

= MIN(A1: A29)

Maximum: höchster Wert im Dataset

= MAX(A1: A29)

Standardabweichung: Misst, was "normal" oder erwartet ist

- "Standardabweichung ist die durchschnittliche Entfernung zum Durchschnitt"

  = STABW (A1: A29)

- Die Standardabweichung gibt an, um wie viel sich ein bestimmter Wert innerhalb des Datensatzes voraussichtlich vom Mittelwert entfernt

3. Korrelationen

Eine Korrelation ist eine statistische Beziehung zwischen zwei Variablen. Zum Beispiel: Wir beobachten eine hohes Vorkommen von Störchen und Babys, die in derselben Region geboren werden. Bedeutet dies, dass der Storch die Babys bringt (und somit eine "kausale" Beziehung)? Unser gesunder Menschenverstand sagt uns: Das kann nicht stimmen! Wir haben vergessen die Variable der “Region” zu berücksichtigen, denn: Es gibt eine höhere Storchpopulation auf dem Land und es werden mehr Babys auf dem Land geboren als in der Stadt. Deshalb gilt: "Korrelation bedeutet keine Kausalität!"

Wann ist die Kausalität angebracht?

- Folge deinem gesunden Menschenverstand: Wenn ein Zusammenhang zu gut ist, um wahr zu sein, stelle ihn in Frage
- Denke nur dann an Kausalität, wenn du alle Informationen in deine Analyse einbezogen hast und keine Informationen fehlen

Beispiel für einen Kausalzusammenhang:

- Korrelation zwischen dem Verkauf von Chicken Wings am ersten Sonntag im Feburar in den USA
- Kann der Ansturm auf Chicken Wings durch den sonntags stattfindenden Super Bowl plausibel erklärt werden?
- Ja! Zumindest kennen wir keinen anderen Grund, warum der Verkauf von Chicken Wings so drastisch zunimmt!

4. EU-Finanzdaten

Die Analyse von Finanzdaten kann aufgrund ihrer Komplexität schwierig sein!

Beim Betrachten von Finanzdaten ist es entscheidend:

- die Art der Daten zu verstehen → Kontext
- Woher kommen diese Daten?
- Was repräsentieren die Zahlen?
- Um welchen Zeitraum handelt es sich?

Wie kannst du große Zahlen in Beziehung setzen?

- Die Höhe der EU-Subventionen pro Land liegt oft in Milliardenhöhe
- Gibt es eine Möglichkeit, die Anzahl zu relativieren? Zusammenhang zu anderen großen Zahlen!
- Wie hoch ist der Anteil der Subventionen am nationalen Bruttoinlandsprodukt?
- Vorsicht: Vergleiche nur Gleiches mit Gleichem (Subventionen werden über 7 Jahre gezahlt)

Weitere Daten und Informationen dazu findest du unter:
https://storyhunt.de

5. Nützliche Tools und weiterführende Links

Online-Ressourcen:

1. Die DataScience Academy hat eine umfangreiche Liste von kostenlosen Ressourcen erstellt: http://datascienceacademy.com/free-data-science-courses/

2. Das DataCamp bietet viele Data Science-Kurse für Python, R und SQL. Grundlegende sind in der Regel kostenlos: https://www.datacamp.com/

3. Errate die Korrelation - ein lustiges Spiel zur Einführung in das Konzept: http://guesthecorrelation.com/

5. Nützliche Tools und weiterführende Links

Tools:

1. Stata http://www.stata.com/: Das wahrscheinlich am meisten benutzte Statistikanalysetool im Wissenschaftsbereich, aber leider nicht Open Source.

2. R https://www.r-project.org/ & R Studio https://www.rstudio.com/: Eine weitere sehr beliebte Programmiersprache für statistische Analysen. Das tool R-Studio stellt ein Interface für verschiedene Analysen bereit. Das Tool ist außerdem Open Source.

3. Python Libraries: Die Programmiersprache Python stellt ebenfalls sehr nützliche Libraries für statistische Analysen bereit, so wie pandas, numpy and scikit. Das ipython-notebook ist ein gutes Interface dafür.Es hilft dabei deine Analysen zu strukturieren.

DATENSCHULE

Die Datenschule vermittelt gemeinnützigen Organisationen die nötigen Fähigkeiten, Daten und Technologien verstehen, um sie zielgerichtet für ihre gesellschaftlichen Aufgaben einzusetzen.

OPEN KNOWLEDGE FOUNDATION DEUTSCHLAND

Die Open Knowledge Foundation Deutschland ist ein gemeinnütziger Verein, der sich für offenes Wissen, offene Daten, Transparenz und Beteiligung einsetzt.

Mehr zu unseren Projekten & Workshops: datenschule.de
E-Mail: info@datenschule.de |
Telefon: 030-57703666-2