

Projektbericht zum Modul Information Retrieval und
Visualisierung Sommersemester 2022

Visualisierung des Datensatzes Student Behaviour

Paul Tebbe, Matrikelnummer: 216237588

21.12.2022

1 Einleitung

..... Viele verschiedene Faktoren haben einen Einfluss auf die Note der Studenten. Dies geht bereits damit los in welche Familie das Kind hineingeboren wird, welchen akademischen Standart die Eltern haben, welche finanziellen Mittel usw. Weiter spielt die Erziehung und der spätere soziale Umgang eine Rolle. [Earthman] Diese Einflüsse sind schwer zu messen, vor allem schwer genau zu messen. Diese Faktoren sind kaum von alleine beeinflussbar, es lässt sich nicht bestimmen, in welche Familie ich geboren werde und welche Voraussetzungen ich also für später mit auf den Weg gegeben bekomme. Doch auf welche Faktoren kann der Student selber später Einfluss nehmen. Die Zeit, die täglich in oder eben nicht in die schulische Weiterbildung investiert wird, sollte in der Theorie so einen selbst beeinflussbaren Faktor darstellen. In dieser Arbeit werden sich also folgende Fragen und folgende gestellt Zielstellungen gesetzt:

- Wie korrelieren Noten und Zeitgestaltung und lassen sich daraus Rückschlüsse ziehen, wenn ja welche?
- Wie stehen die Noten im Zusammenhang mit anderen Variablen?
- Wie können die verschiedenen Daten übersichtlich veranschaulicht werden um dem Leser einen schnellen Überblick zu geben

Tipps zu Latex und Koma-Script für Hausarbeiten sind im LaTeX Reference Sheet for a thesis with KOMA-Script von Marion Lammarsch und Elke Schubert zusammengefasst. Der Bericht fällt in die Kategorie von InfoVis-Paper, die Tamara Munzner Design Study nennt [Munzner2008]: In der Einleitung sollen sie zuerst das Zielproblem beschrieben. Daraus sollen sie Fragestellungen motivieren, die mittels Techniken der Informationsvisualisierung beantwortet werden können. In dem Abschnitt direkt unter der Überschrift Einleitung sollen Sie nach einer kurzen Einleitung Fragestellungen und das Zielproblem motivieren und beschreiben. ...

1.1 Anwendungshintergrund

Wer in der Schule und im College gute Noten hat und sich somit einen besseren Abschluss erarbeitet hat im späteren Leben deutlich bessere Chancen auf „individuelle Lebenschancen, Selbstverwirklichung, beruflichen Erfolg sowie soziale, politische und kulturelle Teilhabe.“

1.2 Zielgruppen

Die Zielgruppe für die Daten umfasst ein weites Spektrum an Personen. Die Visualisierung sind sowohl für Schüler und Studenten, als auch für Lehrer und Eltern interessant, also jegliche Personengruppe, die Berührungspunkte mit der Notengebung und deren Einflussfaktoren aufweist. Die Informationen, die aus den Visualisierungen gewonnen werden können, können sehr gut dafür genutzt werden zu sehen, wie grade der Einflussfaktor Lernzeit und auch die vorangegangen

Noten aus der Schulzeit einen Einfluss auf die Noten im College nehmen. Differenziert werden kann hierbei auch nochmal unter den Geschlechtern. Auch interessant könnte dieses Projekt aber auch für Forschende im Bereich soziale Ungleichheiten in der Bildung sein. Wenn keine eindeutigen Anzeichen zu sehen sind für eine Korrelation zwischen investierter Lernzeit und besseren Noten, so könnte dies ein Anzeichen dafür sein, dass doch andere Einflussgrößen mehr Auswirkung auf die Noten haben als der reine Fleiß. Dies wäre ein starker Indikator für soziale Ungleichheiten bei der Bildung, wie es bereits in der Einleitung angedeutet wurde.

Beschreiben sie die Personengruppe oder Personengruppen, die das von ihnen benannte Anwendungsproblem lösen möchte. Auf welches Vorwissen können sie in dieser Gruppen von Anwenderinnen aufbauen? Welche Informationsbedürfnisse werden durch die Visualisierungen adressiert?

1.3 Überblick und Beiträge

Der in dieser Arbeit verwendete Datensatz differenzieren die Studenten nach verschiedenen Merkmalen. Diese gehen von allgemeinen Unterscheidungen, wie Geschlecht, Alter und Gewicht, über die Aufteilung ihrer Freizeit und die aufgewendete Lernzeit bis hinzu dem körperlichen Wohlbefinden ausgedrückt mittels der Variable ‚stress level‘.

Die für diese Visualisierung mit am interessantesten Daten sind die der verschiedenen Noten in der Schule und im college. Die Noten, gerade mit der von den Studenten angegebenen täglich Lernzeit sollte in einem positiven Zusammenhang stehen. Also steigt die Lernzeit am Tag, steigt auch die Note. Das gleiche sollte auch für die Beziehung der Noten untereinander gelten, also welcher Student schon in der 10.ten und 12.Klasse gute Noten hatte, sollte nun auch im College gut abschneiden. Die körperlichen Voraussetzungen sowie das Department sollten eigentlich keine größeren Einflüssen auf die Noten haben. Die Zeit verbracht auf sozialen Medien könnte eine negative Korrelation mit den Noten aufweisen, genauso wie der finanzielle Status und das Stress Level bei den Studenten.

In diesem Abschnitt geben sie einen kurzen Überblick über die Daten und verwendeten Visualisierungen. Dann benennen sie die Beiträge ihres Projekts. Diese Beiträge müssen sie in den hinteren Teilen des Berichts genauer ausführen und belegen.

2 Daten

Wie bereits in den Abschnitten vorher erwähnt, handelt es sich bei den vorliegenden Daten um den Datensatz *Student Behaviour* von *kaggle*. Die darin aufgeführten und gesammelten Daten geben nähere Auskunft über das Verhalten der Studenten und deren Noten. Die Verwendung des Datensatzes zu wissenschaftlichen Zwecken ist abzuraten, da der Verfasser angibt die Daten selber von Universitäten gesammelt zu haben und bis auf diesen Satz keine näheren Informationen zur Datenbeschaffung liefert. Seiner Meinung nach kann der Datensatz dazu genutzt werden das Verhalten der Studenten in Zukunft besser zu deuten.

Im Datensatz auf *kaggle* sind ein paar Rechtschreibfehler bei der Variablenbenennung enthalten, wie z.B. „social medai time“, diese werden daher der Leserlichkeit halber im Folgenden grammatikalisch korrekt geschrieben)

Er enthält Informationen von 235 Studenten in Form einer CSV-Datei. Diese CSV-Datei hat 19 Spalten mit den Informationen der Studenten über *Certification* als Boolean, der ausgibt ob der Student einen certification Kurs besucht hat oder nicht, *Gender* mit Boolean über das Geschlecht, *Department*, welcher Student welchem Department angehört, *Height(CM)* und *Weight(KG)* als Float, die Größe und Gewicht aufzeigen, *10th Mark*, *12th Mark* und *College Mark*, die als Float die Note der Studenten zeigen, *Hobbies*, sowie *stress level* und *financial status* als String, *part-time job* und *Do you like your degree* als Boolean, *salary expectation* als integer (in Elm allerdings als Float deklariert und verwendet, damit es passend visualisiert und gemapt werden kann), *willingness to pursue a career based on their degree* als String in Form einer Prozentangabe (möglich sind hier nur die Angaben: 0,25,50,75 und 100 Prozent) und *travelling time*, *social media time* und *daily studying time* als String in Form einer Klassenangabe (also z.B. 30-60 Minuten ist eine Angabe).

Während sich grade die Noten, also 10th Mark, 12th Mark und College Mark sehr gut dazu eignen die Fragestellung zu beantworten und sich zusätzlich dazu noch gut visualisieren lassen, da sie als Float sich relativ einfach veranschaulichen lassen ist dies bei anderen Variablen weniger der Fall. Bei ein paar der Daten ist nichteinmal genau klar, was damit überhaupt ausgesagt werden soll, wie z.B. bei Certification Course. Was für ein Kurs ist das, was zertifiziert man damit, bzw. welche Art von Zertifikation erwirbt man damit. Aufgrund des fehlenden Hintergrundwissen, welches leider im Datensatz nicht näher erläutert wird, kann bei diesen Visualisierungen nicht näher darauf eingegangen werden. Ähnlich verhält es sich bei Department. Auch hier fehlen die nötigen zusätzlichen Informationen, um diese Variable in geeigneter Weise einbauen zu können. Auch die meisten Boolean (part-time job, Do you like your degree, werden bei den Visualisierungen nicht berücksichtigt (bis auf gender), da es sich, bei den hier verwendeten Darstellungen nicht anbietet boolsche Werte zu integrieren. Diese bieten schlichtweg zu wenige Möglichkeiten. Vielseitig genutzt wurde auch die daily studying time, die sich vor allem als *drop-down* zur Sortierung der Noten eignete, sodass schnell ersichtlich wurde mit welcher Lernzeit welche Noten erreicht wurden. Hier wurde also eine der oben gestellten Zielstellungen sinnvoll und gut umgesetzt.

2.1 Technische Bereitstellung der Daten

Die Daten wurden als einzelne CSV-Datei in *kaggle* bereitgestellt und auch also solche weiterverarbeitet. Zudem wurden sie im eigenen Github hochgeladen und dort im Ordner Daten gespeichert. Mithilfe dieser Datei wurden der Scatterplot und der Paralleplot erstellt. zur Visualisierung des Baumdiagramms wurde zusätzlich eine json-Datei erstellt *"json.Name"* und auch im Ordner Daten hinterlegt. Zusätzlich dazu finden sich im Github-Repository auch noch Python-Code, mit dessen Hilfe die Daten aufbereitet werden sollen, genaueres in 2.2

2.2 Datenvorverarbeitung

Im obigen Abschnitt wurde bereits kurz auf den Datensatz eingegangen und erklärt, dass es sich um eine einzelne CSV-Datei handelte. Aufgrunddessen musste keinerlei größere Vorverarbeitung an den Daten vorgenommen werden, zumindest für den Paralleplot und den Scatterplot. Für das Baumdiagramm wurde zuerst versucht mithilfe von Pythoncode die csv-Datei in eine json-Datei umzuwandeln. Leider gelang dies nicht, sodass die csv erst in eine Excel-Datei umgewandelt wurde. Anschließend wurde mithilfe der Excel-Datei manuell eine json-Datei erstellt um anhand dieser die Baumdarstellung zu visualisieren. In der json wurden dann nur die für das Baumdiagramm relevanten Daten übernommen. Aufgrund anfänglicher Überlegungen die verschiedenen klassierten *Zeiten*, wie *daily studying time*, *social media time* und *travelling time* in einzelne Klassen in Form von Integers festzuhalten, also *daily studying time* von 0-30 Minuten wäre 1, 30-60 Minuten 2 usw. wurde die csv-Datei in R geladen um die Daten somit zu manipulieren. Aufgrund der späteren Nutzung der *daily studying time* als Drop-down Tabelle wurde dieser Gedanke wieder verworfen.

3 Visualisierungen

In den nachfolgenden Abschnitten wird näher auf die oben genannten Zielstellungen eingegangen.

3.1 Analyse der Anwendungsaufgaben

Wie bereits in 1 erwähnt, sollen die folgenden Visualisierungen der Zielgruppe dabei helfen sich einen schnellen und übersichtlichen Eindruck über den Zusammenhang der unterschiedlichen Noten und anderen Variablen machen. Also welche Variablen nehmen möglicherweise Einflüsse auf die Note oder stehen in einem Zusammenhang mit diesen. Die Leser:innen sollen sehen können, wie sich die Noten verändern wenn sich gewisse Variablen verändern, also wie ändert sich die Notenverteilung, wenn sich die Lernzeit erhöht oder verringert, welche Individuen der Ausgewählten sind Männer, welche Frauen. Auch auf einen möglichen Zusammenhang der sich im Laufe der Zeit verändernden Noten soll mithilfe der Visualisierungen eingegangen werden. Sind Studenten, die in der 10. Klasse gute Noten hatten auch in der 12. Klasse und im College noch so gut und umgekehrt genauso: War ein Student mit schlechten Noten schon in der 10. und der 12. Klasse schlecht und hängen diese Dinge möglicherweise mit dem Lernaufwand zusammen. All diese Fragen können in anschaulicher und so auch schneller Art und Weise von den Beobachtern auf den unterschiedlichen Visualisierungen überblickt werden, mit Unterstützung von verschiedenen Drop-Down-Möglichkeiten und der Hover-Funktion (hier werden den Lesern in den Plots immer die Auskunft über das Geschlecht und in manchen Fällen noch weitere allgemeine Informationen geliefert) über den Punkten oder den Parallellinien. Das normale mentale Modell, dass bei einem Menschen entstehen sollte, wenn er sich diese Fragen und Anwendungsaufgaben

stellt ist, dass es eine positive Korrelation zwischen guten Noten in der Schule und guten Noten im College, als auch zwischen Lernzeit und guten Noten existiert. In den Köpfen der Anwender entsteht also ein stetiger Anstieg der durchschnittlichen Notenverteilung, je mehr Lernzeit dazu kommt und dies sollte auch in den Abbildungen zu sehen sein. Die Mentalen Modelle sind für die Verständlichkeit der Visualisierungen für Leser allerdings kaum notwendig, da sie relativ simpel gehalten sind und es auch eine der oben gestellten Anforderungen an das Projekt war, eine gewisse Komplexität nicht zu überschreiten. Die mentalen Modelle unterstützen also mehr die Visualisierungen, als dass sie vorausgesetzt werden um sie zu verstehen. Allerdings ist, wie schon im oberen Bereich der Hausarbeit erwähnt die Gleichung, die Noten beeinflusst ein hochkomplexes Konstrukt, welche von reinen Zahlen kaum abgebildet werden kann, dies muss den Beobachtern bewusst bleiben und verzehrt so möglicherweise das eigene mentale Modell, gerade wenn bedacht wird, dass es sich bei *Student Behaviour* um einen sehr kleinen Datensatz von 236 Studenten und um keine wissenschaftliche Studie, sondern eigens erhobene Daten handelt. Verwirrend wäre dies vor allem, wenn Studenten mit einer geringen Lernzeit sehr gute Noten haben und Studenten mit höher Lernzeit sehr schlechte, die könnte also durchaus dazu führen, dass der Leser mit einer Diskrepanz zwischen seinem mentalen Modell und der abgebildeten Visualisierung konfrontiert wird.

3.2 Anforderungen an die Visualisierungen

Da die Leser einschätzen sollen können, ob und wenn ja inwieweit eine Korrelation zwischen den verschiedenen Parametern mit Noten und Zeitaufwand besteht, war es wichtig diese Daten aufeinander abgestimmt, in jeweils einer Visualisierung darstellen zu können.

Der Leser sollte bei allen Visualisierungen die Möglichkeit haben nach dem Parameter Zeit zu differenzieren und sich am Ende die Noten von den unterschiedlichen Studenten aus den unterschiedlichen Klassen/College ausgeben zu lassen. Es sollte also erst eine grobe Differenzierung getroffen werden können, um sich nach dieser ersten Einteilung genauer mit den verschiedenen Noten auseinandersetzen zu können.

Weiterhin war es wichtig die Visualisierungen recht einfach zu halten und schnell erkennbar werden zu lassen, ob die verschiedenen Parameter korrelieren oder nicht. Die einzelnen Visualisierungen wurden daher auf den ersten Blick relativ simple gehalten. Die Plots bieten die Möglichkeit des Drop-Downs und stellen am Anfang daher eine einfache Methode für den Betrachter dar, sich durch die verschiedenen Optionen zu klicken, ohne ihn direkt am Anfang mit Darstellungen und Informationen zu überhäufen. Durch die Anzahl von drei Drop-Downs ist es ihm aber wiederum möglich eine große Anzahl an Auswahlmöglichkeiten zu treffen und die Daten sich in vielen verschiedenen Kombinationen anzeigen zu lassen. Bei den Plots bietet sich die zusätzliche Optionen über die Punkte bzw. Linien zu hovern und so noch mehr Informationen über den jeweilig ausgewählten Studenten zu erlangen.

Die Visualisierung sollte in conclusio also anfänglich einfach und übersichtlich gehalten sein und bei mehr Interesse auch weitere Möglichkeiten bieten.

Daher wurde sich für den Anfang für den Scatterplot entschieden, der eine Visualisierung darstellt, die die meisten Menschen kennen und der eine schnelle Übersicht über die Daten liefert, als zweites für einen Parallelplot, der die ersten Anfänge des Scatterplots erweitert, indem alle Noten auf einmal zu sehen sind. Als letztes das Baumdiagramm, dass alle im bisher im Fokus stehende Daten abbildet.

3.3 Präsentation der Visualisierungen

Präsentieren sie die visuelle Abbildungen und Kodierungen der Daten und Interaktionsmöglichkeiten. Sie müssen begründen, warum und wie gut ihre Designentscheidungen die erstellten Anforderungen erfüllen. Weiterhin müssen sie begründen, warum die gewählte visuelle Kodierung der Daten für das zu lösende Problem passend ist. Typische Argumente würden hier auf Wahrnehmungsprinzipien und Theorie über Informationsvisualisierung verweisen. Die besten Begründungen diskutieren explizit die konkrete Auswahl der Visualisierungen im Kontext von mehreren verschiedenen Alternativen. Machen sie hier nicht den Fehler, einfach nur Visualisierung aus den vorgegebenen Bereichen zu diskutieren, weil das in der Regel nicht sinnvoll ist. Wenn sie sich für einen Scatterplot entschieden haben, ist ein Zeitreihendiagramm in der Regel keine Alternative. Diskutieren sie also nicht einfach Zeitreihendiagramme, weil sie in den Anforderungen an das Projekt neben Scatterplots stehen, sondern suchen sie nach echten alternativen Visualisierungen, die zum Aufbau eines vergleichbaren mentalen Modells führen. Diskutieren sie die Expressivität und die Effektivität der einzelnen Visualisierungen.

Die eben beschriebenen Präsentationen und Begründungen sollen für jede der drei folgenden Visualisierungen durchgeführt werden.

3.3.1 Visualisierung Eins

Wie bereits in 3.2 erwähnt wurde sich als erste Visualisierung für den Scatterplot entschieden, abgebildet in Abb.1.

Warum sich für den Scatterplot entschieden wurde, wurde bereits in 3.2 erläutert, zur genaueren Definition des Scatterplots und wie die *Student Behaviour*-Daten darin dargestellt wurden wird in den folgenden Abschnitten näher darauf eingegangen. Der Scatterplot an sich besitzt

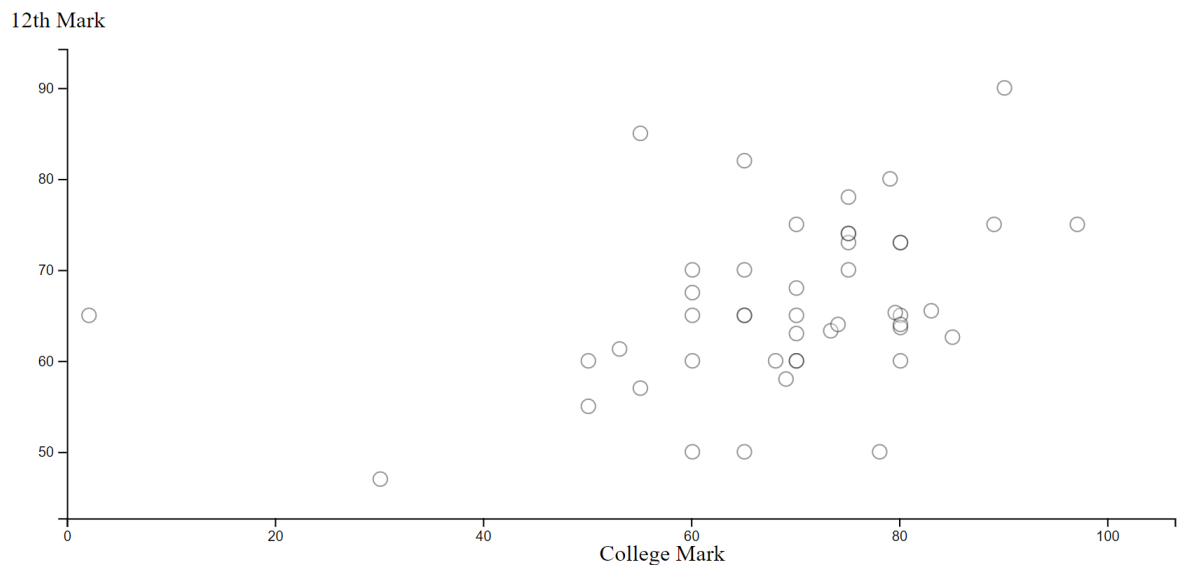


Abbildung 1: Test12

zwei Achsen, die X-Achse und die Y-Achse. Deren Größe kann entweder manuell bestimmt werden oder hat vier weitere Möglichkeiten gleiche Daten verschieden darzustellen([**Hinneburg**])

- X- und Y-Achse haben einen gleich großen Wertebereich
- Nur die Y-Achse hat einen großen Wertebereich
- Nur die X-Achse hat einen großen Wertebereich
- Der Wertebereich in X und Y wird durch die Daten bestimmt

Die manuelle Einstellung der Achsen wäre sehr großer Aufwand und auch durchaus ungeeignet, weswegen sich hier für eine dynamische Option entschieden wurde. Die beiden Achsen wurden dem Wertebereich angepasst und können so auch flexibel mit den unterschiedlichen reingeparsten Daten interagieren und sich verändern. In ?? enthält in diesem Beispiel die X-Achse als Werte die Daten von College Mark und die Y-Achse die Werte der Daten von 12th Mark.

3.3.2 Visualisierung Zwei

3.3.3 Visualisierung Drei

3.4 Interaktion

Die präsentierten Visualisierungstechniken müssen interaktiv zu einer Anwendung verknüpft werden. Die Interaktion mit einer Visualisierung soll in den anderen Visualisierungen zu einer

Änderung führen. Erklären sie die möglichen Interaktionen mit den einzelnen Visualisierungen und die möglichen Verknüpfungen zwischen ihnen. Begründen Sie warum die konkreten Interaktionen umgesetzt wurden und welche Zwecke für die Anwenderinnen mit ihnen unterstützt werden. Begründen sie ebenfalls warum sie andere Interaktionsmöglichkeiten nicht umgesetzt haben. Wenn sie keine der geforderten Interaktionen umsetzen, erhalten Sie im gesamten Projekt deutlichen Punktabzug.

4 Implementierung

Beschreiben Sie die Implementierung ihrer Visualisierungsanwendung in Elm. Stellen die Gliederung ihres Quellcodes vor. Haben Sie verschiedene Elm-Module erstellt. Was war aufwändig umzusetzen, was ließ sich mit dem vorhandenen Code aus den Übungen relativ einfach umsetzen?

Wie sieht die Elm-Datenstruktur für das Model aus, in dem die verschiedenen Zustände der Interaktion gespeichert werden können.

5 Anwendungsfälle

Präsentieren sie für jede der drei Visualisierungen einen sinnvollen Anwendungsfall in dem ein bestimmter Fakt, ein Muster oder die Abwesenheit eines Musters visuell festgestellt wird. Begründen sie warum dieser Anwendungsfall wichtig für die Zielgruppe der Anwenderinnen ist. Diskutieren sie weiterhin, ob die oben beschriebene Information auch mit anderen Visualisierungstechniken hätte gefunden werden können. Falls dies möglich wäre, vergleichen sie die den Aufwand und die Schwierigkeiten ihres Ansatzes und der Alternativen.

5.1 Anwendung Visualisierung Eins

5.2 Anwendung Visualisierung Zwei

5.3 Anwendung Visualisierung Drei

6 Verwandte Arbeiten

Führen sie eine kurze Literatursuche in der wissenschaftlichen Literatur zu Informationsvisualisierung und Visual Analytics nach ähnlichen Anwendungen durch. Diskutieren sie mindestens zwei Artikel. Stellen sie Gemeinsamkeiten und Unterschiede dar.

7 Zusammenfassung und Ausblick

Fassen sie die Beiträge ihre Visualisierungsanwendung zusammen. Wo bietet sie für die Personen der Zielgruppe einen echten Mehrwert.

Was wären mögliche sinnvolle Erweiterungen, entweder auf der Ebene der Visualisierungen und/oder auf der Datenebene?

Anhang: Git-Historie