

Projektbericht zum Modul Information Retrieval und Visualisierung Sommersemester 2021

Leistung von Studenten in Prüfungen

Edward Sabinus

5. September 2021

1 Einleitung

Die Leistung von Studenten in Prüfungen ist je nach Student unterschiedlich. Die Einen legen Prüfungen eher mit hohen Bewertungen ab, die Anderen eher mit niedrigeren. Interessant dabei ist die Frage von welchen Faktoren die Leistungen abhängen. Sind beispielsweise Studenten eher dazu geneigt Prüfungen mit guten Bewertungen abzulegen, wenn ihre Eltern einen hohen Bildungsgrad haben? Oder sind Studenten, die am Vorbereitungskurs für die Prüfung teilgenommen haben erfolgreicher in der Prüfung als diejenigen, die an dem Vorbereitungskurs nicht teilgenommen haben? Dieser Bericht versucht solche Fragen mittels Visualisierungstechniken und einem Datensatz zu klären.

1.1 Zielgruppen und Anwendung

Die Frage wovon die Leistung von Studenten in Prüfungen abhängt kann sowohl für die Lehrenden von Universitäten als auch für die Studenten selbst interessant sein. Beispielsweise die Frage ob die Leistung davon abhängt ob die Studenten am Vorbereitungskurs für die Prüfung teilgenommen haben: Wenn es sich herausstellt, dass die Verteilung der Leistungen sich vollkommen zufällig zur Variable verhält, ob die Studenten am Vorbereitungskurs teilgenommen haben, dann wissen die Lehrenden, dass sie am Vorbereitungskurs etwas ändern müssen: entweder ist ein Vorbereitungskurs für diese Prüfung unnötig oder er muss verbessert werden. Oder wenn es sich herausstellt, dass Studenten, die den Vorbereitungskurs besucht haben bessere Noten bekommen, als diejenigen, die daran nicht teilgenommen haben, dann ist das für alle Studenten interessant, die eine möglichst gute Note bekommen wollen. Solche Information, angewendet auf spezifische Prüfungen und Vorbereitungskurse wird den Studenten helfen sich zu entscheiden ob sie beim Vorbereitungskurs teilnehmen sollten oder lieber die Zeit anderweitig nutzen wollen.

Da der Bericht sich somit auch an Zielgruppen wendet, die sich nicht gut mit Visualisierungstechniken auskennen, versucht der Bericht die einzelnen Visualisierungen so zu erklären, sodass sie auch ohne spezielles Vorwissen verstanden werden können. Allerdings wird davon ausgegangen, dass Lesende wissen wie die Lehre an Universitäten in Verbindung mit Studenten funktioniert.

1.2 Überblick und Beiträge

Die Daten sind zufällig generiert und enthalten Informationen zu Studenten: Geschlecht, Rasse / ethnische Gruppe, Bildungsgrad der Eltern, Ausgaben beim Mittagessen, ob der Prüfungsvorbereitungskurs abgeschlossen wurde und eine Bewertung zwischen 0 bis 100 für je Mathematik, Lesen und Schreiben. Es sind insgesamt 1000 Zeilen an Daten, die publiziert wurden, man kann sich allerdings jederzeit mehr Daten generieren lassen. In dieser Arbeit werden nur die publizierten 1000 Zeilen verwendet.

Für die Lösung der Frage ob Studenten, die den Vorbereitungskurs abgeschlossen haben bessere Leistungen erzielen, werden die Leistungen der Studenten mit abgeschlossenem Vorbereitungskurs den Leistungen der Studenten ohne abgeschlossenem Vorbereitungskurs in einem QQ-Plot gegenübergestellt, aus dem ablesbar sein wird, dass die Studenten mit abgeschlossenem Vorbereitungskurs bessere Leistungen erzielen.

In der zweiten Visualisierung wird die Frage beantwortet, ob die Leistung der Studenten vom Bildungsgrad der Eltern abhängt. Dazu werden analog der ersten Visualisierung die Studenten nach Bildungsgrad der Eltern aufgeteilt und es werden jeweils die Quantile für die Durchschnittliche Leistung für jeden Bildungsgrad berechnet. Es gibt insgesamt 6 Bildungsgrade: Associateabschluss, Bachelorabschluss, Masterabschluss, Oberschulabschluss, irgendein Collageabschluss, irgendein Oberschulabschluss. Daher kann der QQ-Plot nicht 2-Dimensional erfolgen, stattdessen wird dieser QQ-Plot mehrdimensional gezeichnet. Zusätzlich gibt es die Interaktionsmöglichkeit bei der Mehrdimensionalen Darstellung 2 Dimensionen auszuwählen und dann einen 2D-Plot anzeigen zu lassen wie in der 1. Visualisierung.

Die dritte Visualisierung gibt einen Überblick über die gesamten Daten. Die Daten werden nach der Gesamtpunktzahl gruppiert und sortiert. Jede Datenzeile wird dann in ein farbiges Icon codiert und dargestellt, sodass die 4 nominale Werte Geschlecht, Rasse / ethnische Gruppe, Bildungsgrad der Eltern, Ausgaben beim Mittagessen und ob der Prüfungsvorbereitungskurs abgeschlossen wurde abgelesen werden können. Die Gesamtpunktzahl kann dann an der Position des Icons abgelesen werden. Mit einer Interaktion ist man dann in der Lage eines der Werte auszuwählen und in einem QQ-Plot der 1. oder 2. Visualisierung, je nach dem ob der ausgewählte Wert binär ist oder nicht, darzustellen.

2 Daten

Die Daten [**Daten**] sind zufällig generiert und enthalten Informationen zu Studenten: Geschlecht, Rasse / ethische Gruppe, Bildungsgrad der Eltern, Ausgaben beim Mittagessen, ob der Prüfungsvorbereitungskurs abgeschlossen wurde und eine Bewertung zwischen 0 bis 100 für je Mathematik, Lesen und Schreiben. Es sind insgesamt 1000 Zeilen an Daten, die publiziert wurden, man kann sich allerdings jederzeit mehr Daten generieren lassen. In dieser Arbeit werden nur die publizierten 1000 Zeilen verwendet, da die Menge für die Visualisierungszwecke hier ausreicht.

Da es keine Daten aus der realen Welt sind, sondern zufällig generierte Daten sind, sind die Beobachtungen auf diesen Daten nur so aussagekräftig für die reale Welt, wie realitätsnah die Generierung dieser Daten ist. Auf http://roycekimmons.com/tools/generated_data/exams, wo man sich die Daten generieren lassen kann, wird empfohlen die Daten nur für Trainingszwecke zu verwenden.

Daher kann es sein, dass die hier gemachten Beobachtungen von der realen Welt abweichen. Wenn es ein guter Generator ist, dann lässt sich allerdings vermuten, dass auch wenn die einzelnen Daten fiktiv sind, deren Zusammenhänge und statistische Eigenschaften realitätsnah sind.

2.1 Technische Bereitstellung der Daten

Die in dieser Arbeit verwendete Stichprobe ist über Kaggle zugänglich. Der Generator dieser Daten kann auf http://roycekimmons.com/tools/generated_data/exams verwendet werden.

Alle Daten sind als Csv-Dateien verfügbar.

2.2 Datenvorverarbeitung

Da es sich um maschinell generierte Daten handelt, sind alle Daten vollständig und gültig. Somit müssen die Daten nicht gefiltert werden.

Sinnvoll ist aber eine Durchschnittliche Leistung für jeden Studenten zu berechnen. In den Daten sind die Leistungen für Mathematik, Lesen und Schreiben gegeben. Ein Durchschnitt über diese drei Werte erlaubt es einfacher die Gesamtleistung der Studenten miteinander zu vergleichen, was für die Fragestellungen äußerst hilfreich ist.

3 Visualisierungen

3.1 Analyse der Anwendungsaufgaben

Bei den Lehrveranstaltungen und Vorbereitungskursen für Prüfungen haben Lehrende die Aufgabe diese Veranstaltungen so zu gestalten, dass diese den Studenten beim lernen helfen. Bei dieser Gestaltung ist unter anderem auch die Frage wichtig wie sehr und ob die Veranstaltung für die Studenten nützlich ist. Um das herauszufinden hilft eine Gegenüberstellung der Verteilungen der Leistung der Studenten, die den Vorbereitungskurs abgeschlossen haben und der Leistung

der Studenten, die den Vorbereitungskurs nicht abgeschlossen haben. Dadurch bekommen die Lehrenden ein Bild davon ob und wie ihr Vorbereitungskurs hilfreich ist. Nützlich dabei zu wissen ist auch für welche Studenten es eher hilfreich ist und für welche weniger. Das hilft nicht nur den Lehrenden sich auf gewisse Studentengruppen etwas mehr zu konzentrieren, sondern auch den Studenten zu entscheiden, ob sie an dem Vorbereitungskurs teilnehmen sollten oder nicht. Denn dann sehen sie anhand der Studentengruppen ob sie eher in der Gruppe sind, denen der Vorbereitungskurs viel bringt oder eher in der Gruppe sind, denen der Vorbereitungskurs weniger hilft. Bei der Frage an welche Studentengruppen sich Lehrende eher konzentrieren sollten, die mehr Hilfe beim lernen benötigen als andere Studenten, ist eine der Unterscheidungen welchen Bildungsgrad die Eltern haben. Daher ist es wichtig zu analysieren ob und welchen Einfluss der Bildungsgrad der Eltern auf die Leistungen der Studenten hat.

3.2 Anforderungen an die Visualisierungen

Die Visualisierungen sollen ein Bild darüber vermitteln ob der Vorbereitungskurs hilfreich ist. Dazu bietet sich eine Gegenüberstellung der Verteilungen der entsprechenden Leistungen der Studenten an.

Die Visualisierungen sollen auch vermitteln welche Studentengruppen eher dazu geneigt sind schlechtere Leistungen zu erzielen und somit mehr Hilfe beim lernen von den Lehrenden benötigen als andere Studentengruppen; mit dem Fokus auf die Unterscheidung des Bildungsgrades der Eltern.

Weiterhin sollen die Visualisierungen zeigen welchen Studentengruppen der Vorbereitungskurs besser hilft und welchen weniger.

3.3 Präsentation der Visualisierungen

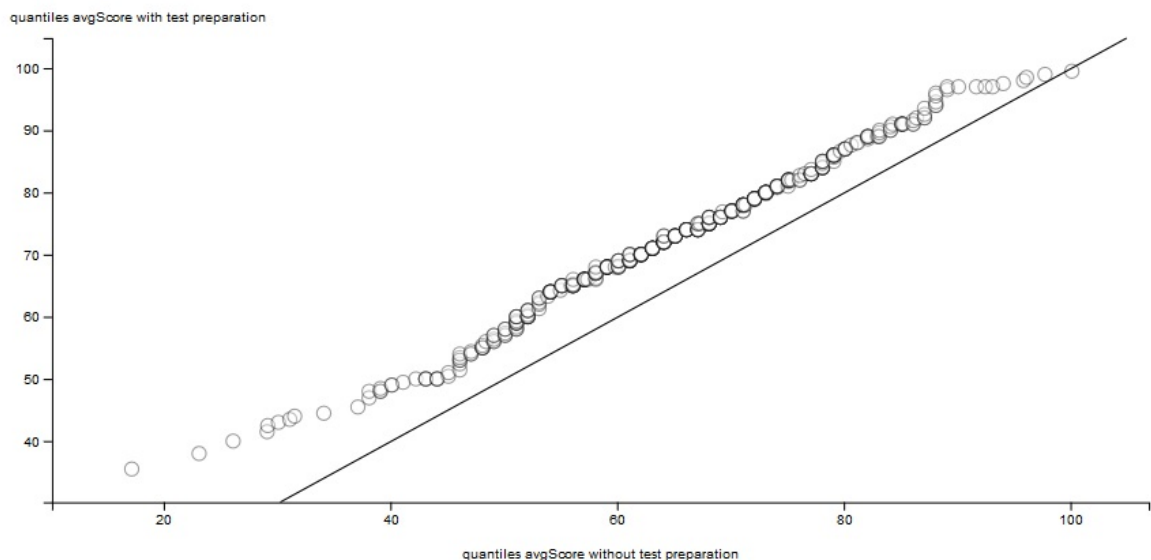
3.3.1 Visualisierung Eins

In der folgenden Visualisierung wird ein Quantil-Quantil-Plot dargestellt. Auf der X-Achse ist die Durchschnittliche Leistung der Studenten dargestellt, die den Vorbereitungskurs nicht abgeschlossen haben. Auf der Y-Achse ist die Durchschnittliche Leistung der Studenten dargestellt, die den Vorbereitungskurs abgeschlossen haben. Ein Punkt (x,y) auf dem Quantil-Quantil-Plot wird genau dann eingezeichnet, wenn das Quantil von x gleich dem Quantil von y ist. Ein Quantil ist dabei die Wahrscheinlichkeit dafür, dass die Werte der Verteilung kleiner dem gegebenen Wert sind. Außerdem ist eine Gerade bei $x=y$ eingezeichnet. Wenn nämlich die Leistung unabhängig davon wäre ob der Vorbereitungskurs abgeschlossen wurde, dann würden bei einer genügend großen Stichprobe alle Punkte diese Geraden bilden. Daher gibt diese Gerade eine Orientierung ob die Studenten von der X-Achse bessere Leistungen erbringen als die Studenten der Y-Achse oder umgekehrt. Man erkennt, dass fast alle Punkte über dieser Orientierungsgeraden liegen, also die Y-Werte größer sind als die X-Werte am entsprechenden Quantil. Das bedeutet, dass die Studenten, die den Vorbereitungskurs abgeschlossen haben bessere Chancen haben bes-

sere Leistungen zu erbringen, als die Studenten, die den Vorbereitungskurs nicht abgeschlossen haben.

Zusätzlich hat man noch die Interaktionsmöglichkeit für jeden angezeigten Punkt die genauen Leistungswerte, die sich gegenüber stehen anzeigen zu lassen.

QQPlot Students avgScore with test preparation vs without test preparation



Durch die Entfernung der Punkte von der Geraden erkennt man gut ob und in welche Richtung der Vorbereitungskurs Einfluss auf die Leistung der Studenten hat. Konkret erkennt man an dieser Visualisierung, dass der Vorbereitungskurs tatsächlich hilfreich ist. Ebenso ist erkennbar, dass die Studenten, die ohnehin schlechtere Leistungen erzielen ohne den Vorbereitungskurs noch viel schlechter sind, da am linken unteren Rand die Punkte viel weiter entfernt von der Gerade sind, als die Punkte am oberen rechten Rand. So sieht man auch, dass diejenigen Studenten, die gute Leistungen erzielen, dass sich dessen Leistung nur leicht bis kaum verbessert, wenn sie den Vorbereitungskurs besuchen. Somit kann man schon durch diese Visualisierung 2 Gruppen von Studenten unterscheiden, auf die sich der Vorbereitungskurs unterschiedlich stark auswirkt. Alternativ zum Quantil-Quantil-Plot könnte man in einem Boxplot 2 Boxen zeichnen: eine für die Studenten, die am Vorbereitungskurs teilgenommen haben und eine für die Studenten, die am Vorbereitungskurs nicht teilgenommen haben. Jede Box stellt dann die Verteilung der Leistungen der entsprechenden Studenten dar. Vorteil des Boxplots wäre, dass man die Mediane sowie die Quantile, die den Rand der Box darstellen, der beiden Verteilungen leichter vergleichen kann, da diese durch die Horizontale Position verglichen werden können und diese explizit durch Linien markiert sind, sodass der Vergleich leichter wahrzunehmen ist als beim Quantil-Quantil-Plot, bei dem man die Quantile selbst nicht direkt ablesen kann. Nachteil des Boxplots ist aber, dass nur wenige Quantile miteinander verglichen werden können, nämlich nur die, die durch die Linien eingezeichnet wurden, während im QQ-Plot zu jedem einzelnen eingetragenen

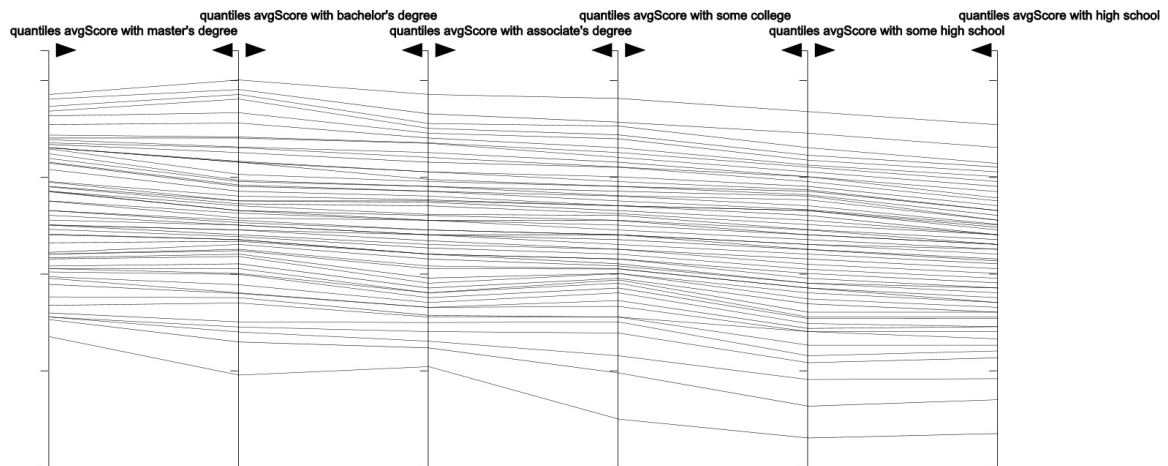
zum Quantil zugehörigen Wert der entsprechende Wert des selben Quantils der anderen Verteilung abgelesen und so verglichen werden kann. Auf die Weise bekommt man ein schärferes Bild des Vergleichs. Durch die zusätzlich eingetragene Linie bei $x=y$ kann man außerdem sehen welche Quantile näher am Gleich-sein sind und welche einen größeren Unterschied aufweisen. Diese Linie erlaubt einen Vergleich, bei dem man lediglich eindimensionale Abstände wahrnehmen muss. Eindimensionale, parallele Abstände lassen sich sehr einfach wahrnehmen. Dadurch überbringt diese Visualisierung ihre Hauptinformation, dessen Wahrnehmung den Betrachtern leicht fällt. Durch dieses schärfere Bild kann man somit nicht nur erkennen welche der Studentengruppen (mit abgeschlossenem Vorbereitungskurs, ohne abgeschlossenem Vorbereitungskurs) bessere Leistungen abscheidet, sondern auch die Studentengruppen aufzeigt, auf die sich der Vorbereitungskurs unterschiedlich stark auswirkt. So sieht man, dass die Wahrscheinlichkeit, dass Studenten, die den Vorbereitungskurs abgeschlossen haben eine sehr schlechte Leistung in der Prüfung aufzeigen deutlich geringer ist als die Wahrscheinlichkeit, dass Studenten, die den Vorbereitungskurs nicht abgeschlossen haben eine sehr schlechte Leistung in der Prüfung aufzeigen. Wohingegen die Wahrscheinlichkeit, dass Studenten, die den Vorbereitungskurs abgeschlossen haben eine sehr gute Leistung vollbringen ähnlich der Wahrscheinlichkeit, dass Studenten, die den Vorbereitungskurs nicht abgeschlossen haben und trotzdem eine sehr gute Leistung vollbringen, ist.

3.3.2 Visualisierung Zwei

In der zweiten Visualisierung wird die Frage beantwortet, ob die Leistung der Studenten vom Bildungsgrad der Eltern abhängt. Dazu werden analog der ersten Visualisierung die Studenten nach Bildungsgrad der Eltern aufgeteilt und es werden jeweils die Quantile für die Durchschnittliche Leistung für jeden Bildungsgrad berechnet. Es gibt insgesamt 6 Bildungsgrade: Associateabschluss, Bachelorabschluss, Masterabschluss, Oberschulabschluss, irgendein Collageabschluss, irgendein Oberschulabschluss. Somit gibt es insgesamt 6 Dimensionen. Diese Daten werden in parallelen Koordinaten visualisiert. Jede Dimension wird hier auf einer senkrechten Achse dargestellt. Alle Achsen sind parallel zu einander. Ein Vergleichspunkt auf dieser Visualisierung ist ein Pfad, der über alle Dimensionen geht und somit die Werte aller Dimensionen miteinander verbindet. Über jeder Achse ist deren Beschriftung. Wenn man auf eines der schwarzen Dreiecke klickt, dann werden 2 benachbarte Achsen miteinander vertauscht. So kann man sich die Anordnung der Achsen so zurechtlegen, wie man daraus am besten die Information, nach der man sucht ablesen kann. Üblicherweise hat jede Achse seine eigene Skalierung, da unterschiedliche Dimensionen im Allgemeinen unterschiedliche Wertebereiche haben können. Bei dieser Visualisierung werden aber alle Achsen gleich skaliert. Das wird dadurch ermöglicht, dass jeder Wertebereich ein Teilbereich von 0 bis 100 ist. Die gleiche Skalierung der Achsen ermöglicht es die Dimensionen anhand des Höhenunterschiedes eines Pfads miteinander zu vergleichen. So sieht man leicht, dass der Wert des untersten Quantils beim Masterabschluss wesentlich besser ist als der Wert des selben Quantils beim Bachelorabschluss, da diese Gerade in Richtung der

Bachelor-Achse nach unten geneigt ist. So kann man anhand der Neigungen der Geraden zwischen 2 Dimensionen erkennen welche der beiden Studentengruppen bessere Leistungen erreicht. Allerdings sieht man ebenso, dass zwischen 2 Dimensionen es oft sowohl steigende als auch fal-

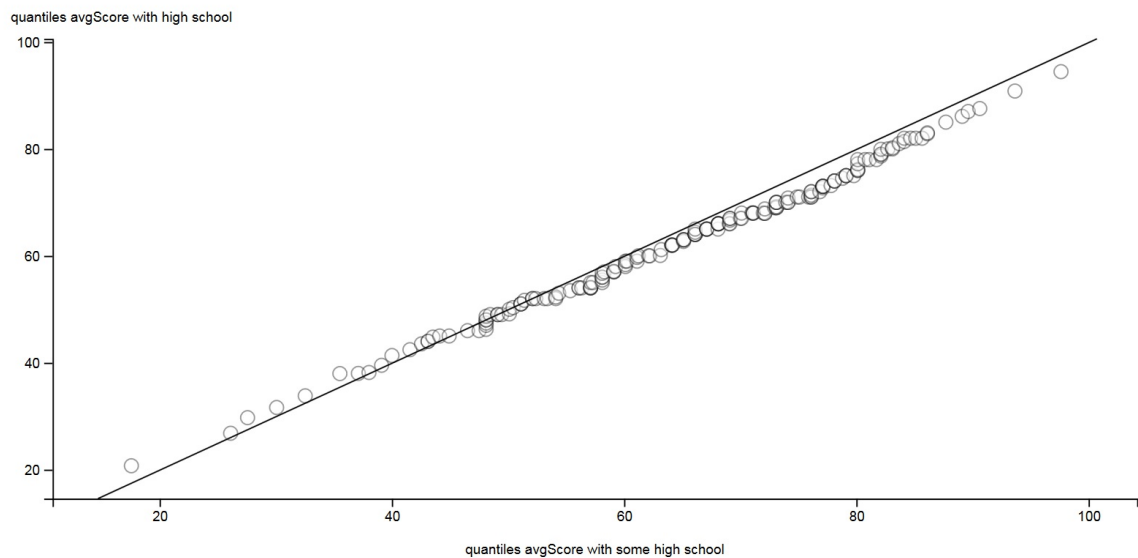
Parallel QQ-Plot for parental Education showing the avgScore



lende Geraden gibt, wodurch es schwieriger wird eine Aussage für die 2 gesamten Gruppen zu treffen. Daher wurde noch die Interaktionsmöglichkeit hinzugefügt, die erlaubt bis zu 2 Achsen auszuwählen. Diese beiden Achsen werden dann in einem 2-dimensionalen QQ-Plot gegenübergestellt (analog Visualisierung 1). Da der 1-dimensionale Abstand in der 1. Visualisierung viel leichter wahrgenommen werden kann als geringe Anstiegsunterschiede von Geraden, lässt es sich mit der 2-dimensionalen Darstellung leichter ermitteln welche der beiden Studentengruppen dann doch besser ist und wenn deren Leistung dann durchschnittlich doch auf das gleiche hinauskommt, kann man trotzdem mit der 2-dimensionalen Darstellung konkretere Aussagen treffen. Außerdem gibt es noch die Interaktionsmöglichkeit mit der Maus über eine der Achsen zu fahren und damit die Beschriftung der Werte der Achse ablesen zu können.

Diese Visualisierung zeigt gut wie die Leistung der Studenten vom Bildungsgrad der Eltern abhängt: Man hat sowohl einen Überblick über alle Bildungsgrade durch die Mehrdimensionale Darstellung als auch ist man in der Lage 2 Dimensionen miteinander optisch zu vergleichen anhand der Neigungen der Geraden zwischen den 2 Dimensionen. Durch die Interaktionsmöglichkeit die Achsen beliebig anzuordnen, kann man so 2 beliebige Dimensionen miteinander vergleichen. Für einen detailreicheren Vergleich hilft die Auswahl der zu vergleichenden Achsen und deren Gegenüberstellung in einem QQ-Plot der gleichen Art wie der ersten Visualisierung. Die Interaktionsmöglichkeit mit der Maus über eine Achse zu fahren und dabei dann die Beschriftung der Werte dieser Achse zu sehen ermöglicht es sowohl alle Pfade vollständig sehen zu können, während die Maus nicht auf einer dieser Achsen ist, als auch den Wert dieser Dimension dieses Mehrdimensionalen Punktes näherungsweise ablesen zu können, während die Maus auf einer dieser Achsen ist. Insgesamt kann man die Leistungen der verschiedenen Studenten-

QQ-Plot quantiles avgScore with some high school quantiles avgScore with high school



gruppen, die nach dem Bildungsgrad der Eltern gebildet wurden, sehr einfach und übersichtlich miteinander vergleichen.

Es gibt auch Alternativen wie man Mehrdimensionale Daten darstellen kann: beispielsweise Chernoff-Gesichter, Sternkoordinaten oder Sternförmige Koordinaten. Aber keine dieser Alternativen ermöglicht es so einfach die Dimensionen miteinander zu vergleichen wie die Parallelen Koordinaten: Chernoff-Gesichter bilden Elemente auf einer Fläche. Aufgrund der Kompliziertheit jedes Gesichts ist es schwierig den Vergleich wahrzunehmen. Außerdem ist es schwierig bei ihnen konkrete Werte abzulesen. Eignet sich also nicht so gut für Vergleiche. Sternkoordinaten platzieren Punkte auf einer Fläche, deren Werte der verschiedenen Dimensionen nicht mehr eindeutig ablesbar sind, daher eignet sich das auch nicht für Vergleiche. Sternförmige Koordinaten sind da etwas näher dran: Die Punkte in Sternförmigen Koordinaten werden genauso dargestellt wie in parallelen Koordinaten mit dem einzigen Unterschied, dass alle Achsen sich in einem Punkt treffen und sternförmig voneinander gehen. Bei diesen Koordinaten wäre ein Vergleich machbar, da man 2 Werte von 2 Dimensionen anhand der Entfernung zum Mittelpunkt vergleichen könnte. Allerdings ist es schwieriger die Längen von 2 nichtparallelen Linien, die sich in einem Endpunkt treffen zu vergleichen, als den Anstieg einer geraden Linie wahrzunehmen. Daher sind parallele Koordinaten die beste Wahl.

3.3.3 Visualisierung Drei

3.4 Interaktion

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.

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