

Projektbericht zum Modul Information Retrieval und Visualisierung Sommersemester 2021

Leistung von Studenten in Prüfungen

Edward Sabinus

10. 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 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.

Außerdem werden gewisse Quantile für die Daten berechnet, die für die Visualisierungen notwendig sind. Quantile sind notwendig um verschiedene Verteilungen von Daten gegenüberzustellen. Es werden immer Quantile zu den durchschnittlichen Leistungen innerhalb gewählter Gruppen berechnet. So ist es möglich zu vergleichen welche der Gruppen leistungsfähiger ist. Im Abschnitt 3 wird geklärt welche Quantile genau berechnet werden.

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. Auch Unterscheidungen hinsichtlich der anderen Attribute kann bei dieser Fragestellung hilfreich sein.

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.

Optional könnten die Visualisierungen zusätzlich zeigen welchen Studentengruppen der Vorbereitungskurs besser hilft und welchen weniger.

3.3 Präsentation der Visualisierungen

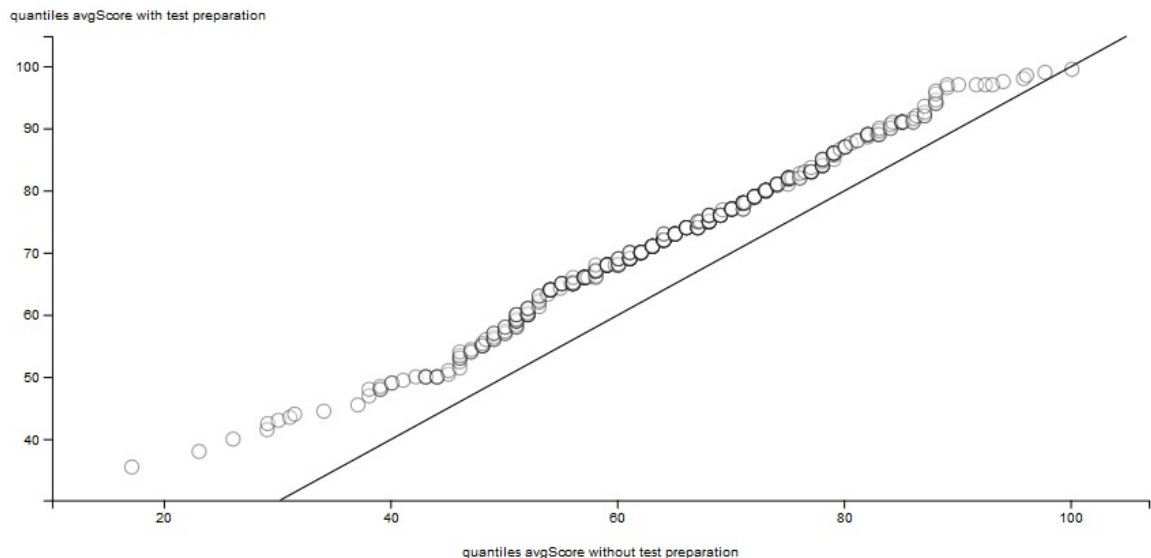
3.3.1 Visualisierung Eins

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 bessere 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

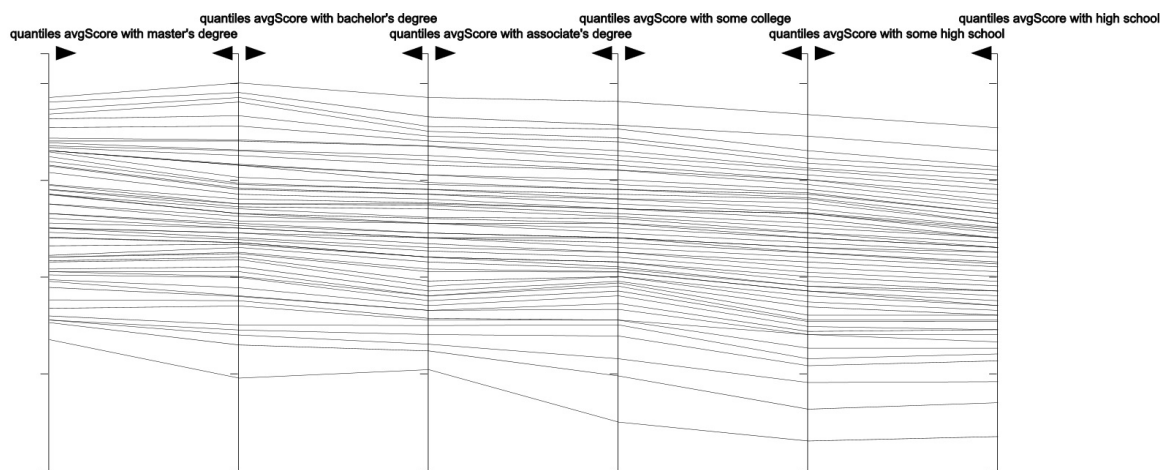
am Vorbereitungskurs nicht teilgenommen haben. Jede Box stellt dann die Verteilung der Leistungen der entsprechenden Studenten dar. Vorteil des Boxplots wäre, dass man die Mediane sowie die Quantile, die den Rand der Box darstellen, der beiden Verteilungen leichter vergleichen kann, da diese durch die 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-

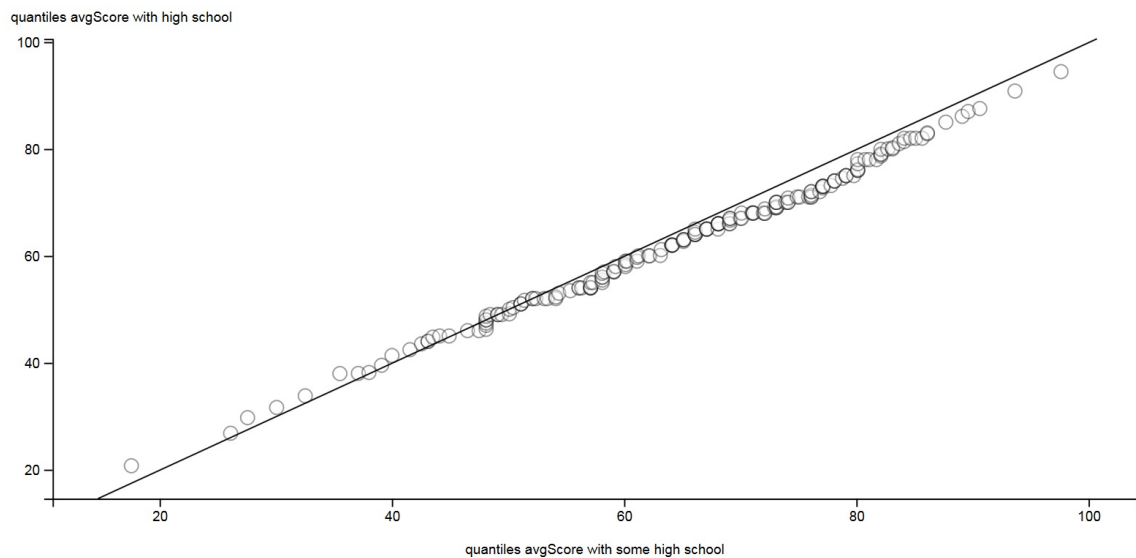
Parallel QQ-Plot for parental Education showing the avgScore



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 Interaktionsmög-

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



lichkeit 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 Studentengruppen, 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

Zum besseren Verständnis der Visualisierung wird oberhalb der eigentlichen Visualisierung die Legende angezeigt. Dort sieht man welcher Wert von welcher Dimension durch welche Farbe dargestellt wird. Die Farben wurden dabei so gewählt, dass sie möglichst aussagekräftig sind. So wurde beim Geschlecht für männlich blau und für weiblich rot gewählt und bei der Vorbereitung für die Prüfung grün für bestanden und rot für kein Vorbereitungskurs abgeschlossen. Zur Legende gehört auch das weiß gefärbte Icon. Durch das Drübergehen mit der Maus über die einzelnen Flächen dieses Icons wird rechts dargestellt welche Dimension in diesem Teil des Icons dargestellt wird. Bei jedem Icon wird während die Maus auf einen bestimmten Sektor zeigt dieser Sektor schwarz gefärbt, sodass man besser sieht auf welchen Sektor man gerade zeigt. So sieht man im folgenden Bild, dass im rechten oberen Icon der rechte obere Sektor von der im Bild unsichtbaren Maus betreten wird. Deshalb ist die Fläche schwarz. Wie bei dem Legende-Icon werden nun rechts die Daten zu diesem Sektor angezeigt. Es wird immer die durchschnittliche Leistung angezeigt und der in diesem Sektor kodierte Wert. Die Visualisierung ist so aufgebaut, dass in jeder Zeile alle Icons die selbe durchschnittliche Leistung haben. Ganz oben mit der höchsten Leistung von 100 und weiter nach unten mit einer immer niedrigeren Leistung. Auf diese Weise sieht man auch die Verteilung der durchschnittlichen Leistung, wobei Leistungswerte zu denen es keine Studenten gibt weggelassen werden. Klickt man auf einen der Sektoren eines Icons, so werden die Daten an dieser Dimension aufgespaltet und in einem QQ-Plot der 1. oder 2. Visualisierung gegenübergestellt. Ob die 1. oder 2. Visualisierung geöffnet wird hängt davon ab ob die Dimension binär ist, also nur 2 mögliche verschiedene Werte hat oder mehr. Bei einer binären Dimension wird die 1. Visualisierung geöffnet, ansonsten wird die 2. Visualisierung geöffnet.

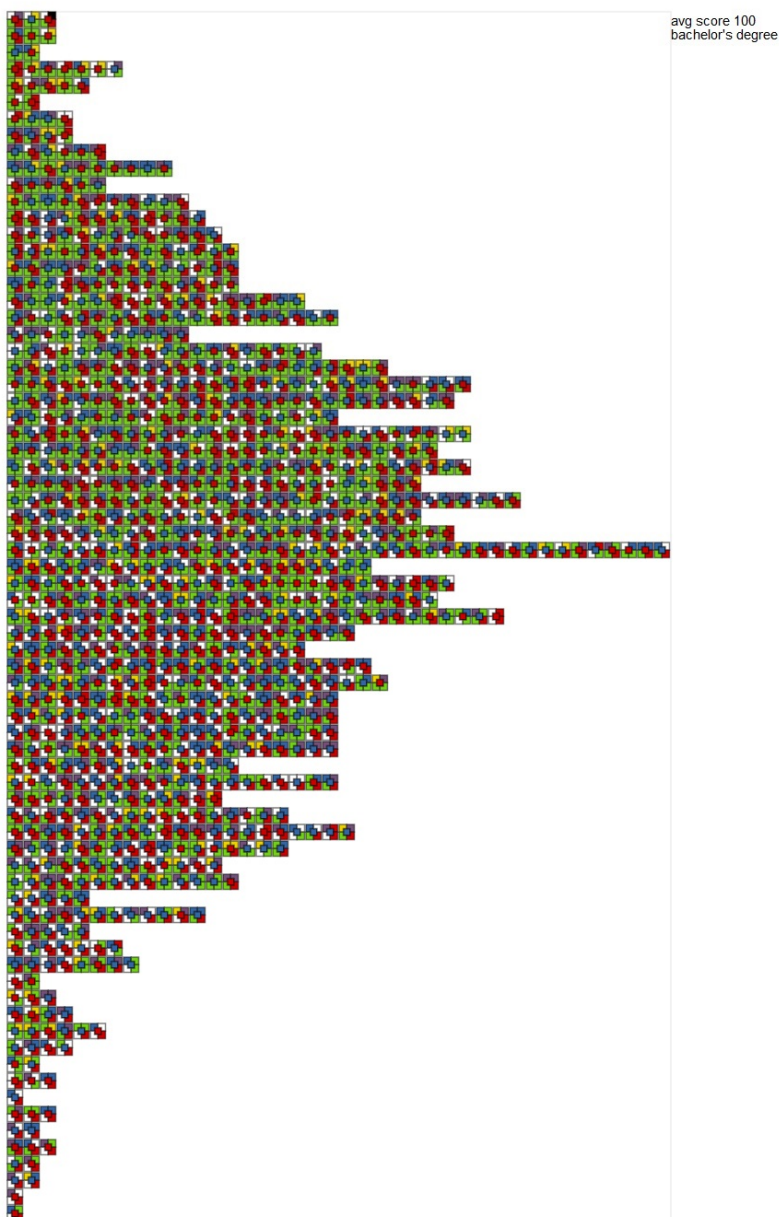
Diese Visualisierung gibt einen Überblick darüber welche Studenten eher dazu geneigt sind bessere Leistungen zu erzielen und welche Studenten eher dazu geneigt sind schlechtere Leistungen zu erzielen. So kann man in dieser Visualisierung gleichzeitig mehrere Einflussfaktoren miteinander vergleichen und schauen welche Bedingungen erfüllen die Studenten, die gute Leistungen erzielen und welche Bedingungen erfüllen die Studenten, die schlechte Leistungen erzielen. So ist beispielsweise auffallend, dass die meisten Studenten, die gute Leistungen erzielen mindestens eine von 2 Bedingungen erfüllen: sie haben den Vorbereitungskurs abgeschlossen oder ihre Eltern haben einen hohen Bildungsgrad. Studenten, die einen Bildungsgrad geringer als blau haben und den Vorbereitungskurs nicht abgeschlossen haben sind eher selten unter den besten Leistungen zu finden. Ebenso kann man erkennen, dass unter den schlechtesten Leistungen die meisten den Vorbereitungskurs nicht abgeschlossen haben und / oder deren Eltern einen niedrigen Bildungsgrad haben. Durch die vielen Daten die in dieser Visualisierung dargestellt sind, ist es aber auch nicht leicht in dieser Menge von bunten Icons Informationen zu finden. Da muss

man etwas genauer hingucken und sich erstmal mit der Codierung der Daten, die oben in der Legende gut abgelesen werden kann, vertraut machen um Aussagen treffen zu können. Genau deswegen dient diese Visualisierung auch nur dem Überblick und gibt die Interaktionsmöglichkeit zu anderen Visualisierungen zu wechseln, in denen man die Daten etwas genauer betrachten kann.

Überblick: Eigenschaften der Studenten nach Durchschnittsleistung sortiert

Legende:

gender: female: red, male: blue,
 race/ethnicity: group E: white, group D: green, group C: blue, group B: purple, group A: yellow,
 parental education: high school: white, some high school: green, some college: blue, associate's degree: purple, bachelor's degree: yellow, master's degree: red,
 lunch: free/reduced: white, standard: green,
 test preparation: none: red, completed: green,



Diese Icons ermöglichen kompakt mehrdimensionale Daten darzustellen, sodass man gleichzeitig alle Daten auf einen Bildschirm bekommen kann und trotzdem die verschiedenen Werte der verschiedenen Daten ablesen kann. Die Verwendung von Farbe ist dabei notwendig um auch Dimensionen darzustellen, die mehr als 2 verschiedene Werte einnehmen können. Ebenso bauen Menschen gerne zu Farben bestimmte Assoziationen auf. So würde beispielsweise jeder dem männlichen Geschlecht einen Blauton vergeben und dem weiblichen Geschlecht einen Rotton. Solche Assoziationen bei der Wahl der Farben zu verwenden hilft auch den Betrachtern die Daten aus den Farben wieder zu decodieren. Solche Assoziationen wurden bei der Visualisierung auch versucht bei den nominalen Daten, die übrigens auch binär sind, zu verwenden. Die Daten Rasse / ethnische Gruppe, die nur von A bis E durchnummeriert sind und die Daten Bildungsgrad der Eltern können als Ordinale Daten betrachtet werden. Für Ordinale Daten passt gut eine Farbabstufung. Um die verschiedenen Werte aber trotzdem gut voneinander unterscheiden zu können wurde eine Farbabstufung gewählt, die möglichst gut voneinander verschiedene Farben beinhaltet und dabei auch eine recht bekannte Farbabstufung sein sollte: Weiß, Grün, Blau, Purpur, Gelb, Rot. Die Sortierung der Daten nach durchschnittlicher Punktzahl ermöglicht es leicht zu erkennen wo Icons zu Studenten mit hoher Leistung und wo Icons zu Studenten mit geringerer Leistung sind, sodass Vergleiche in Bezug auf die Leistung vereinfacht werden. Alternativ statt solcher Bunter Icons könnte man beispielsweise Chernoff-Gesichter verwenden. Allerdings sollte es schwierig sein beispielsweise anhand der Breite der Nase den Bildungsgrad abzulesen, sodass Chernoff-Gesichter für nicht-binäre Daten nicht so gut geeignet sind, während bei farbigen Icons verschiedene Farben genutzt werden können. Ebenso passen Farben gut zu nominalen und ordinalen Daten.

3.4 Interaktion

Bei der 1. Visualisierung (2D-Plot) kann man mit der Maus über die einzelnen Punkte gehen. Wenn die Maus auf einen der abgebildeten Punkte zeigt, dann wird dieser Punkt farblich hervorgehoben, damit die Nutzer einfacher erkennen können auf welchen Punkt sie gerade zeigen. Ebenso werden die genauen Koordinaten des Punktes eingeblendet, sodass die beiden Werte des selben Quantils besser miteinander verglichen werden können.

Bei der 2. Visualisierung (multidimensionaler Plot) kann man mit der Maus über die Dimensionsachsen gehen. Ist die Maus über einer solchen Achse, dann wird die Beschriftung der Achse eingeblendet. So kann man während die Maus nicht über der Achse ist, alle Linien genau erkennen und während die Maus über einer Achse ist, stört die Beschriftung ein wenig bei der genauen Erkennung der Linien, dafür kann man grob die Werte der Linien ablesen. Geht man mit der Maus über eines der Dreiecke oben, so signalisiert das Dreieck durch einen Helligkeitsunterschied, dass die Maus gerade auf das Dreieck zeigt. Und einen klick auf das Dreieck tauschen 2 benachbarte Dimensionen ihre Plätze. So können Betrachter alle möglichen Paare von Dimensionen nebeneinander legen und diese so einfacher miteinander vergleichen. Ebenso hilft das wenn man die Dimensionen nach eigenen bestimmten Eigenschaften sortieren will, so kann

man das gleich in der Visualisierung machen. Beispielsweise könnte man absteigend sortieren nach der niedrigsten Leistung oder nach der höchsten Leistung oder auch nach der geschätzten mittleren Leistung.

Bei der 3. Visualisierung (Überblick) gibt es Interaktionsmöglichkeiten mit den Icons. Beim Drübergehen mit der Maus wird der Sektor des Icons, auf welches gerade die Maus zeigt schwarz, damit man erkennt auf welchen Sektor man gerade zeigt. Ebenso wird rechts die durchschnittliche Leistung des Studenten, welcher durch dieses Icon dargestellt wird angezeigt sowie der decodierte Wert des Sektors. Bei dem weißen Icon der Legende wird stattdessen rechts dargestellt, dass es sich um die Legende handelt und der Name der Dimension des Sektors wird auch dargestellt. Das Legenden-Icon ist notwendig, damit Nutzer sich einfacher mit der Codierung der Daten vertraut machen können. Bei den anderen Icons wird rechts der decodierte Wert dargestellt, für den Fall, dass Nutzer gerade vergessen haben welchen Wert diese Farbe repräsentiert. Auch das hilft diese Visualisierung zu verstehen. Es wäre anstrengender für die Nutzer jedes Mal zurück nach oben zur Legende zu schauen. Die durchschnittliche Leistung wird rechts angezeigt, damit man sieht welche Leistung das Icon hat. Da anderweitig die genaue Leistung nicht abgelesen werden kann, ist das notwendig.

Die 2. Visualisierung ist mit der 1. verknüpft: Wenn man 2 Achsen durch klicken auf diese ausgewählt hat, dann werden in der ersten Visualisierung die 2 Datenreihen gegenüber gestellt, sodass man diese 2 Dimensionen einfacher miteinander vergleichen kann. Da der eindimensionale Abstand im 2D-Plot einfacher wahrgenommen werden kann als unterschiedliche Neigungen von Geraden im multidimensionalen Plot, vereinfacht diese Interaktionsmöglichkeit den Vergleich.

Die 3. Visualisierung ist mit der 2. und der 1. verknüpft: Wenn man auf einen Sektor eines Icons klickt, dann werden die gesamten Daten an der Dimension dieses Sektors aufgespalten und die Leistungen der verschiedenen Werte dieser Dimension werden dann entweder in der 1. oder in der 2. Visualisierung gegenübergestellt. Welche der beiden Visualisierungen gewählt wird, hängt davon ab wie viele mögliche Werte diese Daten haben können. Gibt es nur 2 mögliche Werte, so werden nur 2 Dimensionen benötigt und es wird in der 1. Visualisierung angezeigt. Gibt es mehr als 2 mögliche Werte, so kann das nicht mehr in einem 2D-Plot dargestellt werden und dann wird es in der 2. Visualisierung dargestellt. Diese Interaktionsmöglichkeit ermöglicht es zunächst den Benutzern sich einen Überblick in der 3. Visualisierung zu verschaffen und dann an einem beliebigen Sektor die Daten aufzuspalten und miteinander genauer vergleichen zu können.

Damit unterstützen die Visualisierungen so ziemlich alle sinnvolle Interaktionsmöglichkeiten. Eine Interaktionsmöglichkeit, die auf der ersten blick sinnvoll zu sein scheint wäre beispielsweise bei der 2. Visualisierung noch hinzufügen, dass beim Drübergehen mit der Maus über eine Linie die Koordinaten dieser Linie angezeigt werden. Das würde den Vergleich sicherlich vereinfachen, wenn man die genauen Zahlen sieht, allerdings gibt es sehr viele Linien im Plot der parallelen Koordinaten, sodass es schwierig sein sollte genau eine Linie zu treffen. Außerdem würden die Zahlen der Koordinaten über der ohnehin schon vollen Visualisierung schweben und sicherlich nicht besonders einfach zu lesen sein, wenn andere Linien darüber liegen. Ebenso würde die-

se Interaktionsmöglichkeit es verhindern, dass Nutzer die Linien mit der Maus verfolgen, was für viele Nutzer sicherlich nicht so gut wäre. Denn beim verfolgen einer Linie mit der Maus würden dann ständig verschiedene Zahlen erscheinen, da man immer wieder auf andere Linien zeigt, sodass der Bildschirm quasi anfängt zu flackern, was dem Wahrnehmen der Visualisierung ordentlich stören würde.

4 Implementierung

Das Elm-Programm besteht aus 3 Modulen: Das Hauptmodul `StudentsPerformanceInExams`, `Plots` und `Util`. `StudentsPerformanceInExams` ist das Hauptmodul, dessen `main`-funktion verwendet wird. Das Model hat einige Einträge für das Verwalten der Http-Anfrage: In `httpState` wird gespeichert ob es gerade am Laden ist oder schon erfolgreich ist oder fehlgeschlagen ist. In `fullTexts` wird der empfangene Text aus der Http-Anfrage gespeichert. In `error` wird der Fehler gespeichert, falls es einen Fehler gibt. Falls alles gut läuft, dann wird zunächst aus der erhaltenen Html-Datei der csv-String extrahiert und dann an den csv-Decoder weitergegeben, der dann die Daten aufbaut und ebenfalls in das Model einspeichert. Im Model wird außerdem gespeichert welche Visualisierung gerade gezeigt wird. Beim Start des Programms wird zunächst einfach nur eine Tabelle mit allen Daten angezeigt. In einem `select` (Funktioniert nicht auf allen Browsern. Funktionsfähig unter Firefox.) kann man dann die Visualisierung auswählen. Dann wird noch das Model des Plots im Model gespeichert, welches besonders für die 2. Visualisierung notwendig ist, damit dort die Interaktionen funktionieren. Das Hauptmodul enthält im Grunde genommen nur die `main`-Funktion und alle ihre Bestandteile sowie alle Funktionen, die fest mit der Datenstruktur der spezifischen Daten zusammenhängt. In `Plots` sind alle Plots sowie die Datenverarbeitung, welche unabhängig von den spezifischen Daten verläuft (also nur noch Typen wie `XyData` oder Listen verwendet). Dort ist die `Scatterplotfunktion` vorzufinden, die in Visualisierung 1 verwendet wird. Diese musste nicht mehr verändert werden seit der letzten Übungsserie, in der sie verwendet wurde. Ebenso ist die `ParallelCoordsPlot-Funktion` in diesem Modul. Diese Funktion musste nur noch geringfügig erweitert werden im Vergleich zur Übung, um die Auswahl von 2 Dimensionen, die dann in einem Scatterplot dargestellt werden. Das aufwändigste war die neue `coloredShapesPlot-Funktion` zu erstellen, die die 3. Visualisierung aufbaut. Mit Icons hatten wir uns in der Übung nicht beschäftigt, sodass ich diese Funktion vollständig neu aufbauen musste. Für diesen Plot habe ich neue Datenstrukturen erstellt, die die notwendigen Daten Strukturiert an diese Plot-Funktion übergeben. Eines davon ist `MultiDimNominalData`. Dieser Typ ist ähnlich zu der bisher bekannten `MultiDimData` mit dem Unterschied, dass nicht nur die Beschriftung der Dimensionen notwendig ist, sondern auch die Beschriftung jedes einzelnen Wertes jeder Dimension, damit die Zuordnung der restlichen Attribute zu dem Wert möglich ist. Da die Punkte gruppiert nach durchschnittlicher Leistung sind, ist hier auch eine Liste von Listen von entsprechenden Punkten notwendig. Ein solcher Punkt ist analog dem `MultiDimPoint` mit dem Unterschied, dass die Daten selbst Strings sind. Auch notwendig für die

Plotfunktion ist die Information zu den Icons selbst. Mit dem neuen Typ `ColoredShape` wird die Art der Shapes definiert: Für jeden Wert jeder Dimension wird eine Farbe gespeichert in insgesamt einer Liste von Listen von Farben. Außerdem werden die Koordinaten der Sektoren hier gespeichert. Zusätzlich wird die gesamte Breite und Höhe eines Icons abgespeichert. Diese werden mit der Konstruktorfunktion dieses Typs automatisch berechnet. Die Plot-Funktion selbst nimmt dann nur noch diese Daten und platziert die Icons mit den festgelegten Koordinaten, den Farben und den Werten immer hin. Die Umrechnung von den relativen Koordinaten, die in der Definition des Shapes gespeichert sind, in die im Plot tatsächlich verwendeten Koordinaten erfolgt per `Svg.translate`. Das ergibt im Ergebnis eine Architektur für die `ColoredShapes`, in der leicht die Form sowie die Farben des Icons angepasst werden können. Dafür müssen nur die Argumente an die Plot-Funktion verändert werden. Im Modul `Util` sind einige allgemein benutzbare Funktionen. Hauptsächlich für Listen aber auch ein paar andere.

5 Anwendungsfälle

5.1 Anwendung Visualisierung Eins

Ein sinnvoller Anwendungsfall für die erste Visualisierung ist wenn Lehrende einen optionalen Vorbereitungskurs für eine Prüfung führen und wissen möchten ob und wie sehr ihr Kurs für die Studenten hilfreich ist. Um sich diese Frage zu beantworten sammeln sich diese Lehrende Daten zu ihrem Kurs: die Leistungen der Studenten in der Prüfung, die ihren Kurs besucht haben und die Leistungen der Studenten in der Prüfung, die ihren Kurs nicht besucht haben. Setzen diese Lehrende dann diese Daten in die Visualisierung ein, so werden ihnen die 2 Studentengruppen gegenüber gestellt. In der Visualisierung wird eine Kurve erkennbar sein, die aus den Punkten gebildet wird. Anhand der Lage und der Form dieser Kurve können die Lehrenden sich unterschiedliche Schlüsse ziehen: Liegen die Punkte alle näher an der Achse, auf der die Studenten sind, die den Kurs nicht abgeschlossen haben, dann wissen die Lehrenden, dass ihr Kurs den Leistungen der Studenten schadet. Liegen die Punkte alle näher an der Achse, auf der die Studenten sind, die den Kurs abgeschlossen haben, dann wissen die Lehrenden, dass ihr Kurs den Leistungen der Studenten hilft. Liegen die Punkte alle so ziemlich auf der $x=y$ Geraden, so wissen die Lehrenden, dass ihr Kurs den Studenten nicht hilft aber auch nicht schadet. Außerdem kann es sein, dass manche Punkte über der $x=y$ Geraden sind und manche Punkte unter dieser Geraden sind oder was viel üblicher ist, dass verschiedene Punkte unterschiedlich weit von der $x=y$ Geraden liegen. Wenn die Entfernung zu dieser Geraden stark variiert und irgendwelche Formen, die diese Punkte bilden erkennbar sind, dann sehen die Lehrenden, dass Studenten, die gewisse Leistungen erbringen, der Vorbereitungskurs mehr oder weniger hilfreich ist. Der Eigentliche Vergleich beruht hier auf der Wahrnehmung der eindimensionalen Entfernung der Punkte von der $x=y$ Geraden. Für die Frage welchen Studenten der Kurs eher hilfreich ist muss man hier den visuellen Vergleich der Entfernungen wahrnehmen. Die Wahrnehmung einer Entfernung, also ob eine Entfernung vorhanden ist bzw von welcher Seite der Geraden die

Punkte liegen ist sehr einfach. Somit ist die Visualisierung sehr effektiv für die Lösung dieser Fragestellung. Eine Alternative zu dieser Visualisierung wäre der Box-Plot, wie schon im Abschnitt 3 erwähnt. Beim Boxplot lässt sich diese Fragestellung, also ob ein Vorbereitungskurs hilfreich ist auch sehr einfach vergleichen. Denn beim Boxplot werden einige wichtige Quantile wie z.B. der Median markiert. Legt man 2 Boxplots untereinander, sodass die gleiche horizontale Stelle die gleiche Leistung bedeutet, so kann man auch hier anhand der Entfernung zweier Markierungen feststellen welche der beiden Studentengruppen leistungsfähiger ist. Allerdings muss man hier alle Quantile einzeln miteinander Vergleichen, da sie nur als einzelne Markierungen auf der Box dargestellt sind, was etwas schwieriger sein kann als bei einem QQ-Plot wie in der ersten Visualisierung: Denn beim QQ-Plot gibt es sehr viele Punkte, die wir Menschen schnell zu einer Kurve oder Geraden umwandeln, was uns die Möglichkeit gibt auf ein mal für alle Punkte festzustellen ob sie über der $x=y$ Geraden liegen oder unterhalb. Für die Fragestellung wie sehr hilfreich der Vorbereitungskurs ist für die unterschiedlichen Studentengruppen, die unterschiedliche Leistungen erbringen, ist der Boxplot eher wenig geeignet, da dort nur wenige Markierungen miteinander verglichen werden können, während beim QQ-Plot sehr viele Punkte sind, sodass man ein schärferes Bild erhält.

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