Projektbericht zum Modul Information Retrieval und Visualisierung Sommersemester 2023

Visualisierungen zur Analyse von Einflussfaktoren auf die Schlaff Effizienz

Mick Stewart Wörner

27. Dezember 2023

1 Einleitung

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 besschreiben.

1.1 Anwendungshintergrund

Sie müssen genug Hintergrund bereitstellen, so dass die Lesenden sich ein Urteil bilden können, ob ihre Lösung funktioniert. Sie sollen die Lesenden jedoch nicht mit Anwendungsdetails so überschütten, dass der Fokus auf die Fragen zur Informationsvisualisierung untergehen. Eine Visualisierung muss **expressiv: Expresivität bedeutet, dass die Daten unverfälscht wiedergegeben werden. Grundsätzlich sollen nur die Informationen dargestellt werden, die auch im Datenamterial vorhanden sind.**
I effektiv (Effektivität hängt nicht nur von den Daten ab, sondern auch von:
1. dem Bearbeitungsziel und
2. den Fähigkeiten des Betrachters
3. Eine effektive Visualisierung versucht die Inhalte auf intuitivenWegen zu präsentierten) I angemessen sein (Angemessenheit beschreibt den Verbrauch an Ressourcen zur Erzeugung der Visualisierung)

1.2 Zielgruppen

Die Schlafqualität hat einen signifikanten positiven Einfluss auf die Lebensqualität von Menschen. [2] Das Menschliche schlafsystem ist allerdings ein hochkompliziertes System das von Multiplen externen Faktoren beeinflusst werden kann. Die in diesem Projekt dargestellten Visualisierungen sollen dabei helfen Einflüsse und Zusammenhänge bei der Schlafqualität zu erkennen und Forschern sowie Privatpersonen dabei zu helfen den Einfluss von Lebensstilen auf verschiedene Maße von Schlafqualität zu erkennen. Die untersuchten Attribute sind der Koffeein, Alkohol, Tabak Konsum und Sport sowie das Alter und Geschlecht der Personen. Dies sollte Personen hefen die Verhaltensweisen zu identifizieren, mit denen Sie den größten Einfluss auf Ihre Schlafqualität haben könnten. Dabei kann ausgegengen werden, dass die Benutzer wissen, dass die Datenlage kritisch in hinblick auf die Aussagekräftigkeit und praktische implikationen reflektiert werden muss. Dies ist wichtig, da der benutze Datensatz n < XXXX Datenpunkte besetitz, welche durch den eingebauten Filter weiter reduziert werden können. ab n<30 ist die Statistische Aussagekraft

ncihtmehr gegeben. klein ist. 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?: **Überarbeiten:**
The data was then analyzed to understand the relationship between lifestyle factors and sleep patterns and to identify any potential areas for intervention to improve sleep

1.3 Überblick und Beiträge

Im diesem Abschnitt wird eine Überblick auf die Daten und verwendeten Visualisierungstechniken. 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

Beschreiben Sie vorhandenen Daten. Der verwendete Datensatz besteht aus: 452 Personen, welche durch Ihre ID identifiziert werden. Ob die ID nur den Datenpunkt oder die Person Identifiziert ist unklar. Daher lässt sich nicht sagen ob Schlafdaten einer Person mehrfach erfasst weurden sind. Auf der Kaggle Seite wurde nach angaben des Authors erwähnt, dass der Datensatz im Kontext einer Studie von der ENSIAS, Marroco gesammelt wurde. Innerhalb einer eingeschränkten Recherche konnten weder auf der Webseite der ENSIAS noch in weitergehender Literaturrecherche eine Quelle identifiziert werden. Daher sollten die Daten und daraus entwickelten Ergebnisse, nicht unreflektiert übernommen werden. Der Datensatz hat 15 Attribute, Nominale (Id, Gender und Raucher) und Quantitative (Rest) .

Erste Gruppe: Identifikatoren, sind zur eindeutigen bestimmung eines Datentupels oder Person. Dieser Gruppe gehöhrt nur die "ID" an. Die ID identifiziert eine Person einmalig. Da keine ID wurde mehrfach aufgeführt wird ist anzunehmen, das jede Person nur einmalig an der Studie teilgenommen hat. Die ID wird als Integer bereitgestellt. Der Datenbereich geht von [1-452]
Age: Gibt das Alter an, welches die Person zum Zeitpunkt der Erfassung hatte. Das Alter wird als Integer angegben uns ist dahingehend Diskret z.B. 43 Jahre. Die Verteilung des Datenbereiches, werden im folgendem in diesem Formati angegeben. (Quantile [Min, 25, 50, 75, Max]), Quantile [9, 29, 40, 52, 69]
Gender: Das Geschlecht wird als String abgespeichert nimmt aber nur zwei Werte an: "Male" oder "Female". Dabei gibt es einen Anteil von 50 Prozent Männern und 50 Prozent Frauen.
Bedtime: Gibt die Uhrzeit an zu der die Person ins Bett gegangen ist hierbei ist nicht klar ob damit der Zeitpunkt gemeinst ist, zu dem die Person eingeschlafen ist oder zu dem die Person

sich ins Bett gelegt hat. Die Information wird als DateTime angegeben. Die Daten steigen in 30 Minuten schritten und ist trotz DateTime somit Diskret.
WakeUp Time = Gibt das Datum und die Uhrzeit an zu dem die Person erwacht steigt analog zu der Bedtime in halben Stunden Schritten an. Das Datum ist bei beiden DateTime formaten nicht von weiterem interesse, da es keine zeitliche Entwicklung der erfassten Personen gibt.
Sleep Duration: Die Schlafdauer ist wie die Bed Time unklar in Ihrer interpretation, da sich der Wert immer aus der Differenz zwischen bedtime und WakeupTime berechnet. Daher ist unklar ob es sich um die geschlafene oder um die im Bett verbrachte Zeit handelt. Die Spalte wird als Float angegeben und steigt aufgrund der halbstündlichen Sprünge der BedTime und WakeupTime auch in 0.5 Schritten. Die Daten haben Quantile von [5.0, 7.0, 7.5, 8.0, 10.0] Stunden. Die Interprettation der Sleep Duration wird weiter dadurch erschwert, dass im weiterem Datensatz die Anzahl angegeben wird wie oft eine Person in der Nacht wach wird. Aber ohne Anganbe wie lange diese SSchlafpausenßpezifisch sind.
Schlaf Effizienz = Gibt den prozentualen Anteil an, die eine Person Schlafend im Bett verbracht hat. Die Daten werden als Float mit zwei Nachkommastellen angeben. Die Daten haben die Quantile: [0.5, 0.7, 0.82, 0.9, 0.99]. Eine Person die 5 Stunden im Bett verbracht hat und davon eine Stunde wach war. Hat also eine Schlafeffizienz von 80 Prozent. Da hier wieder die Interpetations problematik besteht. Wird im weiteren davon ausgegangen, dass die Schlafeffizienz angibt welchen Anteil die person nach dem Einschlafen schlafend, also in einem der drei Schlafzyklen verbracht hat.
REM Sleep percentage = Die REM steht für Rapid Eye Movement Schlaf, dies ist einer der drei Schlafzyklen die ein Mensch im Schlaf durchführt. Die CDC empfiehlt einen Anteil von 25 Prozent Healthline sollte. Der REM Percentage gibt den Prozentualen Anteil an den die schlafende Person im REM Verbracht hat. Also Anteil REM an Schlaff Effizienz. Die Daten werden als Integer abgespeichert und haben Quantile von [15, 20, 22, 25 30]
Deep sleep percentage? = Der Tiefschlaf Prozentsatzt gibt den Anteil am Schlaf an, der im Tiefschlaf verbracht wurde. Die Daten werden als Integer angegeben und haben Quantile:[18, 51, 58, 63, 75]
Light sleep percentage = Gibt den Prozentualen Anteil am Schlaf an, der im Leichtschlaf verbracht wurde. Die Daten werden als Integer angegeben und haben Quantile von [7, 15, 18, 40, 63]
Awakenings = Gibt die absolute Anzahl an, wie oft eine Person aufgewacht ist. Die Daten werden im Datensatz als Float abgespeichert. 0.0 bedeutet eine Person hat durchgeschlafen und ist nur einmal Final am morgen aufgewacht. Die Daten reichen von [0.0, 1.0, 1.0, 3.0, 4.0]
Caffeine Intake = Gibt an wie viel Koffeein die Person in den letzten 24 Stunden zu sich genommen hat. Die Maßeinheit hierbei beträgt mg. Die Daten werden als Float abgespeichert und haben Quantile von [0.0, 0.0, 25.0, 50.0, 200.0].
Alcohol Intake = Gibt an wie viel Alkohol die Personen ind en letzten 24 Stunden zu sich genommen hat in Oz. Die Daten haben Quantile von [0.0, 0.0, 0.0, 2.0, 5.0]

Tobacco Intake = Gibt an ob die Person Raucht. Die Daten sind als String abgespeichert: Yes für Raucher und No für Nichtgeraucht. 154 Personen geben an zu Rauchen und 298 geben an NichtRaucher zu sein.
Exercise Intake = Gibt an wie viele Einheiten Sport die Person in der Woche macht. Dabei ist nicht angeben welche Maßeinheit diese Einheiten Sport haben. Die Daten haben Quantile von [0, 0, 2, 3, 5], es gibt 6 fehlende Werte.

Bei Annahme, dass die Daten legitim sind und die angegebene Interpettation der Attribute korrekt ist, ermöglicht dieser Datensatz geneu Einblicke in die Schlafqualität vorallem die Schlafphasen und Dauer dieser. Zusätzlich werden relevante Verhaltsweisen und Einflüsse erfasst. Die Erfassung der Einnahme von Kaffee und Alkohol ist, suboptimal, da die beiden Substanzen innerhalb des im menschlichen Körpers eine geringe Halbwertszeit aufweisen. Dahingehend wäre der Zeitpunkt der Einnahme relevant. Weitere Faktoren die die Schlafqualität beinflussen werden nicht erfasst weiter werden die länge der individuellen Schlafunterbrechungen nicht differenziert und den drei Schlafphasen nicht zugeordnet. So könnte es für den Anwender von Interesse zu sein, welche der Schlafphasen durch welche Verhaltensweisen gestört werden. Weiter lässt sich argumentiern, dass die Verteilung der Verhaltensweisen sehr linksseitig ist. Dies mag auf kulturelle Unterschiede zurückzuführen sein. Daher muss die Aussagekräftigkeit des Datensatzen auf den Marrokanischen Lebensstil eingeschränkt werden.

Die Daten Der Datensatz wurde auf Kaggle veröffentlicht und unterliegem keinem Copyright Schutz. Die Daten sind als CSV mit einer Größe von 9kB zugänglich. Die Daten sind in dem Github Repository abgepeichert, dies ermöglicht eine zentrale aktualisierung der Daten, falls dies Nötig sein sollte. Das Programm ruft die CVS Datei mittels eines Http.get request auf und speichert diese als einen String ab. Falls der requst klappt wird die Message 'Got Text Ok fulltext' an die Funktion 'update' gesendet. Der String ist hierbei repräsentiert durch 'fullText'. Daraufhin wird im 'Model' das 'datenladen' auf Success gesetzt. Auf den String wird die Funktion 'stringtoUnverarbeitete', deren Ziel es ist den String in eine List(UnverarbeiteteDaten) zu transformieren. Die Funktion benutzt das Paket BrianHicks/elm-csv (Im Code als Decode) Diese zieht die Namen der Felder aus der ersten Reihe in dem String. Die Funktion decode decodiert die Inputdaten, relevant hierbei ist das, auch leere Felder in dem String vorkommen dürfen. Die Funktion Decode.blank gibt ein 'Nothing' zurück, falls das Feld leer sein sollte. Wenn es doch zu einem Fehler kommen sollte gibt die Funktion stringtoUnverarbeitete eine leere Liste zurück an das Modell. Wenn die Daten nun in der Form Unverarbeitete Daten sind wird die Funktion 'sleep2Point' angewandt. Diese entfernt einen Tupel, wenn eines seiner Felder ein 'Nothing' beinhaltet. Weiter werden die Werte für REM, Tiefschlaf und Leitschlaf in wirklich in diesen verbrachten Stunden transformiert indem diese, mit 0.1 in Prozente übertragen wurden und dann mit der Schlafeffizienz und Schlandauer multipliziert. Dies ist besser, da somit die interpretation erleichtert wird, und die Vergleichbarkeit wird hergestellt. Vor der Transformation waren

die Schlafphasen einer Person die 5 Stunden schläft und einer Person die 10 Stunden schläft nicht unterscheidbar. Jetzt läsäst sich klar anzeigen wie viel die Personen in den Schlafphasen verbracht haben. Gender und Raucher werden von einem String zu einem Float mittels case handeling konvertiert (genderToFloat und raucherToFloat). Dies ermöglicht es mit den Daten zu rechnen und erleichtert das weiter visualisieren. Die Funktion sleep2Point hat einen Output von Typ Alias 'Aussortierte Daten'. Weiter werden die Attribute Bedtime und Wakeuptime in den aktuellen Visualisierungen nicht benutzt, könnte es bei der Weiterentwicklung interessant werden, daher wurden Sie im Datenkonstrukt belassen aber nicht weiter behandelt. Man könnte von einer sanften Projektion sprechen. Innerhalb der Einstellungen kann der Anwender eine Selektion durchführen und Einschränkungen auf das zu untersuchende Attribut anwenden. So werden nur Datentupel an die Visualisierungen übergeben die dem Kriterium entsprechen.

3 Visualisierungen

3.1 Analyse der Anwendungsaufgaben

Die Aufgaben die durch die Visualisierung gelöst werden sollen:
Die Hauptaufgabe der visualisierungen ist es dem Nutzer dabei behilflich zu sein sich einen Überblick über die gegebenen Daten zu machen, und die Attribute auf Ihren zusammenhang zu anderen zu analysieren und dann die größe der Einflüsse der Verhaltensweisen auf das untersuchte Attrbiut zu differenzieren. Der Anwender muss zunächst die Datenlage kritisch analysieren und die güte der einzelnen Attribute auf Ihre Aussagekräftigkeit zu bewerten. Nach Bewertung der Attribute vergleicht er diese mit anderen Attributen und kann so komplexe Zusammenhänge und Beziehungen erkennen. Daraufhin müssen die Einflüsse des Verhaltens auf die Attribute analysiert werden. Dabei muss der Anwender die Grundlagen der statistischen Interpretation und relevanz von Normalverteilungen für die Interpretation von Datenlagen kennen. Er sollte also wissen, dass die Datenlage normalverteilt sein sollte, um statistische Aussagen treffen zu können. Weiter ist es von Nutzen die Korrelation interpretieren zu können. Die Interpreatation von Rot als negative und Grün als positive ist von Nutzen um die ergebnisse schneller zu erfassen, ist allerdings nicht notwendig.

Die ersten beiden Qualität der Daten zu überprüfen, die Interaktion und zusammenhänge zu identifizieren und einflüssel des verhaltens auf die Attribute zu identifizieren. Das Problem ist folgendes: In hochkomplexen Systemen sind die Einflussfaktoren die die Ausprägung eines Merkmals beeinflussen nicht immer klar. Daher ist es wichtig, dass wir die Daten visualisieren um die Zusammenhänge zu erkennen. Die Visualisierung soll uns also erlauben den Einfluss von multiplen Verhaltensindikatoren auf die Ausprägung eines Merkmales zu schätzen.

Analysieren sie die konkreten Anwendungsaufgaben, die die Lösung des Zielproblems durch die Anwender:innen bearbeitet werden müssen.

Welche sinnvollen mentale Modelle helfen den Personen bei der Bearbeitung. Aufgabenstellung: Analyse der Variablen und den Einfuss der Verhaltensindikatoren Handelt sich um explorative Visualisierung Menatle Modelle: Was ist eine Normalverteilung? Abweichungen von der Linie: Ein Verständnis dafür, wie Abweichungen von der diagonalen Linie im QQ-Plot auf nicht-normale Verteilungen oder systematische Abweichungen hinweisen können. Schwänze und Spitzen: Das mentale Modell von Schwänzen und Spitzen in einer Verteilung hilft dabei zu verstehen, wie sich Ausreißer oder starke Konzentrationen von Werten auf den QQ-Plot auswirken können.

Sind diese mentalen Modelle für sie notwendig, um die Aufgaben lösen zu können? Gehen sie bei ihrer Argumentation von den Anwendungsaufgaben aus und kommen sie dann zu den

mentalen Modellen, deren Aufbau durch Visualisierungen unterstützt wird.

3.2 Anforderungen an die Visualisierungen

Leiten sie Anforderungen an das Design der Visualisierungen ab, die sich durch ihre Analyse des Zielproblems ergeben.

Anforderungen an das Design.

3.3 Präsentation der Visualisierungen

3.3.1 Visualisierung Eins: QQ-Plot Datenlage Normalverteilt?

Präsentieren der Visualisierung und Interaktionsmöglichkeiten.

Dieses Diagramm zeigt auf der x Achse die Normalverteilung an und auf der Y achse die Daten des gweählten Attributes. Das Ziel der Visualisierung ist es die Daten

Wieso sind diese gut und erfüllen die Anforderungen an das Design?

Der QQ-Plot ermöglicht eine visuelle Überprüfung von Abweichungen zwischen den empirischen und theoretischen Quantilen. Abweichungen können auf nicht-normalverteilte Daten hinweisen. Abweichungen nach oben könnten auf eine rechtssteilere Verteilung hindeuten, während Abweichungen nach unten auf eine linkssteilere Verteilung hinweisen könnten.

Wieso erfüllt die Visualisierung den anfordeurngen der Anwender?

Wahrnehmungsprinzipien und Theorie über Informationsvisualisierung verweisen.

Wieso diese Visualisierung und nicht andere? Wir wollen echte mögliche Alternativen die zu einem ähnlichem mentalem Modell führen.

Expressivität und Effektivität der Visualisierung diskutieren.

Die Zielgruppe der Visualisierungen sind Nutzer mit einem Interesse an den zu grundeliegenden Daten. Es soll also nicht einfach Ergebnisse präsentiert werden sondern die hinterlegten Inputdaten dargestellt werden. Die Visualisierung ist sinnvoll, da sie es dem Anwender ermöglicht sich mit den hinterliegenden Daten vertraut zu machen und die verschiedenen Ausprägungen zu analysieren und auf ihre Güte zu überprüfen. Datenlage zu überprüfen. Falls im Normal-Q-Q-Plot sich eine Gerade ergibt, ist die empirische Verteilung etwa normalverteilt Anstieg der Gerade ist σ, Verschiebung der Gerade ist μ. Die klassische Herangehensweise in den Einfluss zu überprüfen sind hochkomplizierte und benötigen statistisches Hintergrundwissen auf Seiten des Anwenders. Die Visualisierungsanwendung soll es dem Anwender ermöglichen den Datensatz und die Merkmale derer zu untersuchen und Rückschlüsse auf die Beziehungen von Ausprägungen untereinander und mit Verhaltensindikatoren und zu treffen. Dies ermöglicht dem Anweder die Daten auf Validität zu überprüfen.

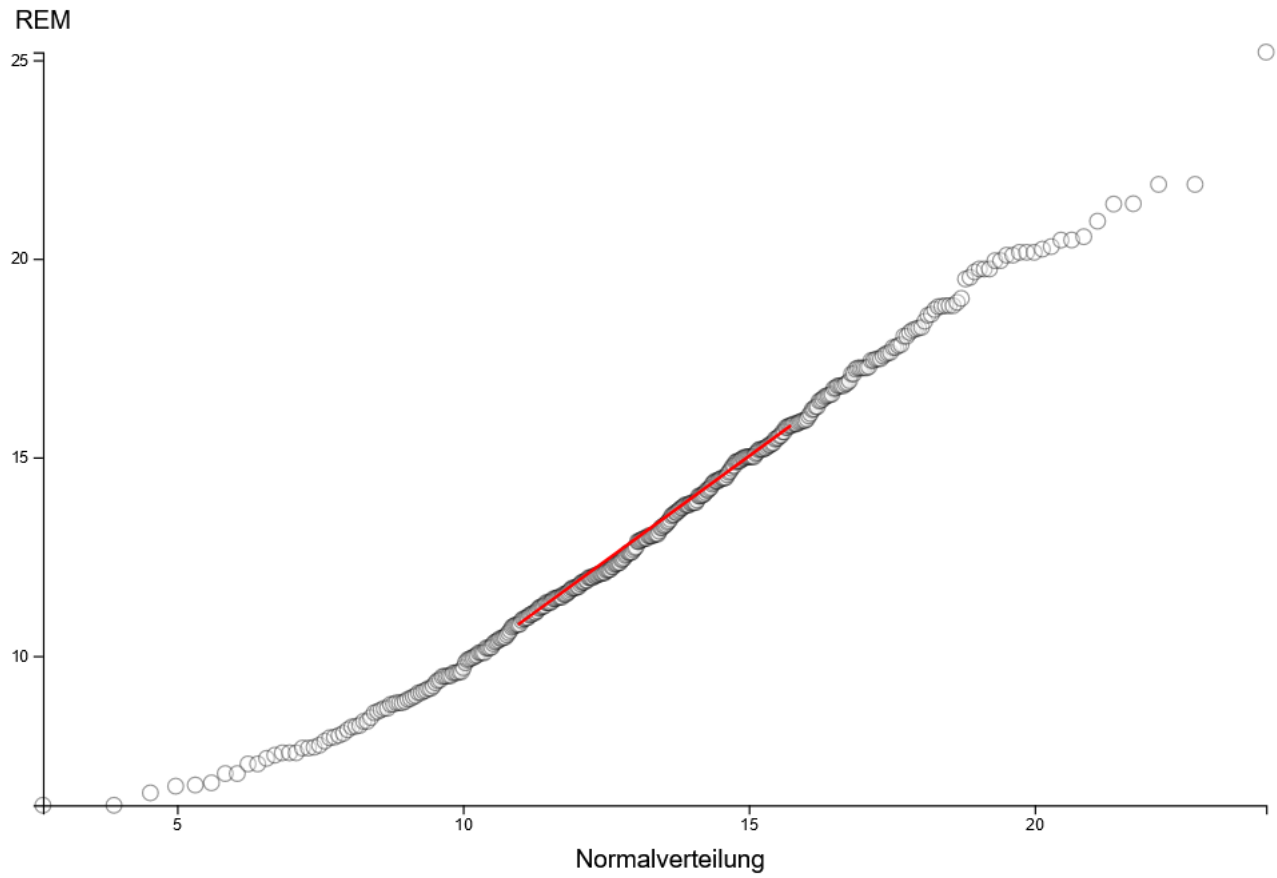

Abbildung 1: Visualisierung Eins - Norm QQ-Plot REM

3.3.2 Visualisierung Zwei: Boxplott

Präsentieren der Visualisierung und Interaktionsmöglichkeiten.

Wieso sind diese gut und erfüllen die Anforderungen an das Design?

Wieso erfüllt die Visualisierung den Anforderungen der Anwender?

Wahrnehmungsprinzipien und Theorie über Informationsvisualisierung verweisen.

Wieso diese Visualisierung und nicht andere? Wir wollen echte mögliche Alternativen die zu einem ähnlichem mentalem Modell führen.

Expressivität und Effektivität der Visualisierung diskutieren.

3.3.3 Visualisierung Drei: Force Graph

Test

Präsentieren der Visualisierung und Interaktionsmöglichkeiten. Wieso sind diese gut und erfüllen die Anforderungen an das Design?

Wieso erfüllt di eVisualisierung den anfordeurngen der Anwender?

Wahrnehmungsprinzipien und Theorie über Informationsvisualisierung verweisen: Wir bauen ein Kausales Mentales Modell auf, Diese verhaltensweisen beinflussen dieses Attribut in dieser Art und weise.

Wieso diese Visualisierung und nicht andere? Wir wollen echte mögliche Alternativen die zu einem ähnlichem mentalem Modell führen.

Expressivität und Effektivität der Visualisierung diskutieren.

Kausales Mentales Modell

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 vorhanden 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