Projektbericht zum Modul Information Retrieval und Visualisierung Sommersemester 2023

Visualisierungen zur Analyse von Einflussfaktoren auf die Schlaff Effizienz

Mick Stewart Wörner

26. 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 [1]: 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 we-**

der 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, diese werden in drei Gruppen eingeteilt. 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 be-

steht. 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 Einehiten Sport haben. Die Daten haben Quantile von [0, 0, 2, 3 , 5], es gibt 6 fehlende Werte.
Gehen sie kritisch darauf ein, in wie weit sich die Daten für die Bearbeitung der Fragestellungen und dem Erreichen von Lösungen für die oben beschriebene Zielgruppen eignen. **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 und die**

länge der Unterbrechungen werden nicht differenziert und nicht den Schlafphasen zugeordnet. So wäre es Interessant, zu wissen welche der Schlafphasen durch welche Verhaltensweisen gestört werden. Weiter lässt sich argumentiern, dass die verteilung der Verhaltensweisen sehr linksseitig ist. Dies mag auf kulturelle Untershciede zurückzuführen sein. Daher muss die Aussagekräftigkeit des Datensatzen auf den Marrokanischen Lebensstil relativiert werden. Haben sie die Daten sinnvoll mit weiteren Datenquellen ergänzt? Wenn ja, wie? Erklären sie die technische Bereitstellung der Daten.

Daten werden als amerikanisches CSV auf Kaggle.com bereitgestellt. (MICK: Genauen Typ der CSV herausfinden. (trennung mit , statt dem Europäischem ;))

Die Daten Wie sind die Daten zugänglich? Welche Formate werden genutzt. Gibt es Besonderheiten beim Lesen der Formate? Beschreiben sie die Datenvorverarbeitung. Welche Datenvorverarbeitungsschritte sind notwendig? **Einspielen des CSV als String, -> Werden aus dme Github Repository gezogen.** Beschreiben Sie die einzelnen Schritte und begründen sie sie, z.B. warum werden manche Daten weggelassen, über welche Mengen werden Durchschnitte berechnet, warum sind die so berechneten Werte aussagekräftiger als andere Werte. Wenn möglich sollen sie die Datenvorverarbeitung in Elm programmieren, so dass ihre Anwendung auf eine Änderung der Rohdaten reagieren kan. **Umgang mit den leeren Feldern-> Mit List.map (Bei einem Nothing, wird der Tupel komplett gelöscht.), da wir somit nur mit einem sauberem Datensatz arbeiten. Datensatz hat nur Fehlende werte bei gewisse Datenpunkten. (Koffeein, Alkohol, Tabak, Sport).** Realisiert eine Daten-zu-Daten-Abbildung

Mögliche Operationen:
- Vervollständigung, Interpolation
- Projektion (Reduzierung der Variablen)
- Selektion (Anwendung von Filterkriterien, Glättung,Ausreißereliminierung)
- Berechnung impliziter Eigenschaften (z.B. Maximum,Gradient)
- Konvertierung

Die Daten sollten so konvertiert werden, dass Sie den Visualisierungsanwendungen entsprechen. Veränderung der Merkmalsausprägungs Daten alle Strings außer ID in eine Interpretierbaren Wert. Unsere Strings sind Boolean, male oder female und Raucher Yes und No

3 Visualisierungen

3.1 Analyse der Anwendungsaufgaben

Die Aufgaben die durch die Visualisierung gelöst werden sollen:
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. 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 und die
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** 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

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 zulösenden 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 Anforderungenen 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: QQ-Plot Datenlage Normalverteilt?

3.3.2 Visualisierung Zwei: Boxplott

Ziel der Visualisierung: Erklärung, wie diese aufgebaut sind. Daher

3.3.3 Visualisierung Drei: Force Graph.

Ziel der Visualisierung: Erklärung, wie diese aufgebaut sind.

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

Literatur

[1] Tamara Munzner. “Process and Pitfalls in Writing Information Visualization Research Papers”. In: *Information Visualization: Human-Centered Issues and Perspectives*. Hrsg. von Andreas Kerren u. a. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, S. 134–153. ISBN: 978-3-540-70956-5. DOI: 10.1007/978-3-540-70956-5_6. URL: https://doi.org/10.1007/978-3-540-70956-5_6.