HAW HAMBURG

BACHELORTHESIS
Viktor Olkov

Datenexploration – Herausforderungen und Vergleich von Visualisierungsmöglichkeiten

FAKULTÄT TECHNIK UND INFORMATIK
Department Informatik

Faculty of Computer Science and Engineering
Department Computer Science

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Hamburg University of Applied Sciences

Viktor Olkov

Datenexploration - Herausforderungen und Vergleich von Visualisierungsmöglichkeiten

Bachelorarbeit eingereicht im Rahmen der Bachelorprüfung
im Studiengang *Bachelor of Science Informatik Technischer Systeme*
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Marina Tropmann-Frick
Zweitgutachter: Prof. Dr. Olaf Zukunft

Eingereicht am: 20. Juli 2021

Viktor Olkov

Thema der Arbeit

Datenexploration - Herausforderungen und Vergleich von Visualisierungsmöglichkeiten

Stichworte

Datenvisualisierung, Datenanalyse, Automatisierung

Kurzzusammenfassung

Obwohl es heute sehr simpel ist für jeden Datenvisualisierungen zu erstellen, gibt es nur wenige Regeln und Konzepte, welche einheitlich angewendet werden. Im Rahmen dieser Bachelorarbeit werden Daten und Visualisierungen hinsichtlich einer automatischen Datenvisualisierung untersucht. Nach der Identifikation verschiedener Eigenschaften und Möglichkeiten wird ein Tool konzipiert, welches durch das automatische Generieren von Visualisierungen hier einheitliche Konzepte einbringen kann, um die visuelle Datenexploration zu optimieren.

Viktor Olkov

Title of Thesis

Data exploration - Challenges and Comparison of Visualization Options

Keywords

Data visualization, data analysis, automation

Abstract

Although it is very simple today to create data visualizations for everyone, there are only a few rules and concepts that are applied uniformly. In the context of this bachelor thesis, data and visualizations are examined with respect to an automatic data visualization. After the identification of different features and possibilities, a tool will be designed, which can introduce uniform concepts by automatically generating visualizations in order to optimize the visual data exploration.

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

Nahezu jede Aktion, die wir heute tätigen, erzeugt eine für den einzelnen Menschen unüberschaubare Menge an Daten, ob wissentlich oder unwissentlich. Sei es das Fitnessarmband, welches unsere Bewegung im Alltag misst, oder die Kasse, an der wir unseren Einkauf bezahlen. Alles wird automatisch aufgezeichnet. Ohne jedoch diese Daten weiter zu verarbeiten, können wir keine Erkenntnisse hierdurch gewinnen. Selbst bei kleinen Datenmengen fällt es uns ohne passende Darstellung schwer, Muster oder Abhängigkeiten darin festzustellen. Eine Lösung hierfür ist es, Visualisierungen zu erstellen. Musste man dies früher von Hand machen, so ist es heute mit ein wenig Erfahrung im Umgang mit Computern für jeden möglich. Dies bringt jedoch auch Herausforderungen mit sich: Wollen wir diese Visualisierung mit unseren Mitmenschen teilen, erwarten wir, dass unser Gegenüber diese versteht und auf dieselbe Weise interpretiert, wie wir es im Vorfeld getan haben. Durch heutige Tools ist das Erstellen von Visualisierungen so intuitiv und einfach, wie nie zuvor. Ein Thema wird dabei jedoch vernachlässigt: Wie visualisiert man Daten optimal? Lassen wir mehrere Nutzer dieselben Daten visualisieren, so werden unterschiedliche Ergebnisse vorliegen, da jeder Nutzer nach seiner Erfahrung, seinen Interessen und seinem Geschmack die Darstellung gestaltet. Widersprechen sich dann die Vorstellungen von den Nutzern, so kann es zu Verwirrung oder gar Fehlinterpretationen kommen.

1.1 Zielsetzung

Ziel dieser Ausarbeitung ist es, ein Analysetool zu entwickeln, welches aus einem Datensatz automatisch Visualisierungen generiert. Dabei wird geprüft, welche Mindestmenge an Eingaben eines Nutzers notwendig sind. Der Fokus liegt darauf, den Nutzer von der eigentlichen Gestaltung der Visualisierungen vollständig abzulösen. Dadurch können verschiedene Visua-

lisierungsregeln gezielt eingebunden und eine einheitliche Datenvisualisierung erzeugt werden. Fehlinterpretationen sollen dadurch verhindert, die Hürde neue Visualisierungstypen zu verwenden genommen und die Datenanalyse erleichtert werden. Somit können unterschiedliche Nutzer, welche denselben Datensatz analysieren, die Gewissheit haben, dass jeder andere dank der gleichen Darstellung zu denselben Schlussfolgerungen kommt. Die Visualisierungen sollen dabei anhand der Datenstruktur, den Datentypen und durch gewonnene Informationen einer Datenanalyse generiert werden.

Hierfür soll eine prototypische Anwendung entwickelt werden, um die Umsetzbarkeit zu prüfen. Dabei sollen zusätzliche Visualisierungsregeln aus den IBCS berücksichtigt werden.

1.2 Gliederung

Die Ausarbeitung ist in 7 Kapitel gegliedert, wobei das erste Kapitel die Einleitung darstellt. Hier soll ein erster kurzer Einblick in das Thema gegeben und die Zielsetzung definiert werden.

Das zweite Kapitel bildet die Grundlagen. Hier sind Erläuterungen zu Begriffen und Konzepten zu finden, die für das Verständnis der Arbeit gegebenenfalls benötigt werden.

Im dritten Kapitel wird die aktuell angewendete visuelle Datenexploration analysiert und beschrieben. Es werden Konzepte, Anwendungsgebiete und Tools genannt und sowohl Vorteile als auch Herausforderungen aufgezeigt, welche im Rahmen der visuellen Datenexploration auftreten.

Im darauffolgenden vierten Kapitel werden Eigenschaften von Daten aufgezeigt. Es werden unter anderem verschiedene Datentypen, Skalenniveaus und Datenstrukturen verglichen und deren Einfluss auf die visuelle Datenexploration analysiert.

Anschließend werden im Kapitel fünf verschiedene Visualisierungen gezeigt, deren Anwendungsgebiete beschrieben und die Vor- und Nachteile der jeweiligen Visualisierungstypen dargestellt. Zudem wird erläutert welche Arten von Visualisierungsdimensionen es gibt und

welche Gestaltungskonzepte eine Visualisierung aus Sicht der Informationsübertragung verbessern können.

Im sechsten Kapitel der Ausarbeitung wird die Umsetzung eines Tools durchgeführt, welches die Analyse eines Datensatzes mittels automatisiert generierter Visualisierungen ermöglichen soll. Dabei fließen unter anderem die Gestaltungskonzepte aus Kapitel fünf mit ein.

Der siebte Teil der Ausarbeitung enthält eine Zusammenfassung und den Rückblick im Hinblick auf das zu Beginn der Arbeit definierte Ziel. Außerdem wird ein Ausblick auf mögliche zukünftige Erweiterungen gegeben.

2 Grundlagen

In diesem Kapitel werden Bezeichnungen, Begrifflichkeiten und Konzepte erklärt, welche zum Verständnis dieser Ausarbeitung benötigt werden.

2.1 Begriffe

- **Visualisierung**
  Hiermit sind in dieser Arbeit speziell Datenvisualisierungen gemeint. Eine Visualisierung ist eine aufgrund von qualitativen und/oder quantitativen Daten erzeugte visuelle Darstellung, welche den Zweck hat, Einblick in eben diese für den Betrachter zu erleichtern oder zu ermöglichen.

- **Visualisierungstyp**
  Der Visualisierungstyp beschreibt die darstellbaren Visualisierungsdimensionen und definiert eine Ordnung, nach der die Visualisierungsdimensionen mit Eigenschaften der Daten befüllt werden. Zudem beschreibt es den Aufbau der Visualisierungselemente.

- **Visualisierungsdimensionen**
  Dies sind die visuellen Ausprägungen einer Visualisierung, mittels welcher die Eigenschaften der Daten abgebildet werden. Sie können eine räumliche Größe sein wie die x- und y-Achse, oder auch die Farbe der Visualisierungselemente.

- **Visualisierungselement**
  Hiermit sind visuelle Elemente wie Punkte, Linien, Formen und Symbole gemeint, welche letztendlich die Daten repräsentieren. Ihre visuellen Eigenschaften wie Größe, Farbe oder Struktur sollen dabei Teilmengen eines Datensatzes oder Datenpunkte visuell darstellen. Gemeinsam mit einer dazugehörigen visuellen Referenzgröße wie beispielsweise einer Achse lassen sich so Werte für den Betrachter ablesen und vergleichen.

- **Datensatz**
  Als Datensatz wird die Datenmenge bezeichnet, welche in dieser Ausarbeitung analysiert und visualisiert wird.
- **Datenpunkt**
  Ein Datenpunkt ist ein eindeutig bestimmbarer Wert in einem Datensatz. Dieser kann einen beliebigen Datentyp haben.
- **Lageparameter**
  Lageparameter sind in der Statistik zu finden und dienen der Ermittlung der zentralen Lage einer Verteilung von Daten. Beispiele hierfür sind der Median oder der Mittelwert.

2.2 Visualisierungs-Pipeline

Die Visualisierungs-Pipeline ist eine Prozesskette, welche den Weg von Rohdaten bis hin zur bildlichen Darstellung beschreibt. Sie besteht aus einer Reihe von aufeinanderfolgenden Funktionen.

Zuerst werden Schritte angewandt, welche die Rohdaten bereinigen, mit Informationen erweitern und filtern, so dass ein wohlgeformter Datensatz entsteht.

Danach werden die Daten auf Visualisierungselemente abgebildet für einen passenden Visualisierungstyp. Dies wird dann als Visualisierung dargestellt. (Moreland, 2013)

2.3 IBCS

Die International Business Communication Standards, kurz IBCS, ist eine Sammlung von praktischen Vorschlägen für die Gestaltung der Geschäftskommunikation. Dies umfasst sowohl den Aufbau von Berichten als auch die Gestaltung von Diagrammen und Tabellen in Bezug auf ihre inhaltliche Konzeption, ihre visuelle Wahrnehmbarkeit und ihre semantische Notation.

Das Ziel der IBCS kann wie folgt erläutert werden: Anders als beispielsweise in der Musik das Notensystem oder in der Elektrotechnik Schaltpläne, gibt es in der Geschäftskommunikation wenige feste Standards. Auch wenn von vielen Nutzern bereits einheitliche Regeln angewendet werden, so scheitert es an den Details bzw. an der Konstanz. Möchte ein Controller

mittels eines Berichts seine Mitarbeiter auf ein Problem oder einen Fortschritt aufmerksam machen, so kann hierbei die Botschaft anders formuliert sein als gewollt und/oder anders bei den Empfängern ankommen als erwartet. Dies kann gerade bei unerfahrenen oder neuen Mitarbeitern vermehrt passieren. Beim Überwinden dieser Herausforderung soll die sogenannte SUCCESS-Formel helfen, welche ein Akronym der 7 darin enthaltenen Regeln darstellt. Diese befassen sich mit der Terminologie, Beschreibungen, Dimensionen, allen visuellen Elementen und vielem mehr.

Auffallend bei der Gestaltung der Visualisierungen ist der minimalistische optische Aufbau. Oft werden keine sichtbaren X- oder Y-Achsen eingefügt, Gitter werden weggelassen und klassische Legenden sind seltener zu finden. Farben werden nur verwendet, wenn wirklich notwendig und auch dann sehr diskret. Stattdessen werden Werte und Beschreibungen direkt an den Visualisierungselementen positioniert, welche meist in schwarz auftreten. Dies soll den Spielraum für Fehlinterpretationen reduzieren, nichtssagende oder gar ablenkende Elemente reduzieren und damit die eigentliche Botschaft klarer in den Vordergrund stellen. (Hichert & Faisst, 2022)

2.4 Visual information seeking mantra

Das Visual information seeking Mantra ist ein Konzept von Ben Shneiderman, wie Daten möglichst optimal für einen Nutzer dargestellt werden sollten. Es lässt sich wie folgt zusammenfassen: Erst einen Überblick gewinnen, dann zoomen und filtern, Details gibt es dann auf Anfrage. Hierdurch macht sich der Analysierende erst großflächig mit der Struktur und dem Umfang seiner Daten vertraut. Daraufhin können Teilmengen der Daten analysiert werden und mittels der dadurch gewonnenen Erkenntnisse Details erforscht werden. (Craft & Cairns, 2005)

3 Visuelle Datenexploration

Im folgenden Kapitel beschreibe ich, auf welche Weise wir Daten visuell explorieren und welche Herausforderungen es dabei gibt.

3.1 Vorteile von Datenvisualisierung

Die visuelle Datenexploration bietet viele Vorteile gegenüber einer Datenexploration mittels simpler tabellarischer Darstellung. Sie konzentriert sich auf den wichtigsten und ausgeprägtesten Sinn von uns, dem Sehsinn. Wir erkennen Muster, Verläufe und Unterschiede in visuellen Elementen, ohne dass wir darüber aktiv nachdenken müssen. Aus einer einfachen Reihe von als Zahlen dargestellten Werten können wir zwar Muster ermitteln und weitere Informationen ableiten. Dieser Prozess ist jedoch langsam, aufwendig und kontraintuitiv. Zudem kann es hierbei gerade bei repetitiver Analyse schnell zu Fehlern kommen. Dies lässt sich simpel an den beiden Zahlen 3656 und 3565 zeigen. Wir erkennen zwar den Unterschied zwischen ihnen, jedoch müssen wir diesen zuerst mit einigen Schwierigkeiten interpretieren. Erschwerende Faktoren sind die gleiche Länge, der Aufbau aus den gleichen Zahlen, das ähnliche Aussehen der Zahlen 5 und 6, das gleiche strukturelle Muster der Zahlen und dass der Unterschied zwischen ihnen erst in der Mitte beginnt. All diese Probleme entstehen ausschließlich durch die Darstellung. Erhöhen wir die Menge an Werten, so wird die Analyse immer schwerer.

Ein sehr bekanntes Beispiel für den Nutzen von Datenvisualisierungen ist die Karte der Choleraausbrüche in der Broad Street, London, im Jahr 1854 (Tuthill & Van Wyk, 2003). Der Arzt John Snow hatte dabei auf einer Karte mehrere Ausbruchsorte der Krankheit eingezeichnet und konnte dadurch eine konzentrierte Verteilung im Umfeld einer Wasserpumpe feststellen und somit das Rätsel um den Ursprung lösen.

3.2 Bekannte und verbreitete Visualisierungstools

Auf dem heutigen Markt gibt es eine sehr große Menge an Tools, welche die Erkundung und Analyse von Daten vereinfachen. Sie erleichtern das Visualisieren von Daten deutlich, indem

sie auf eine intuitive Weise interaktive und hoch anpassbare Visualisierungen erzeugen. Oft muss der Nutzer dabei nur den Visualisierungstyp auswählen und die Visualisierungsdimensionen den jeweils gewünschten Daten zuordnen, um ein Ergebnis zu erhalten. In vielen Tools lässt sich dabei auch die komplette Visualisierungspipeline abbilden. Zu den bekanntesten und verbreitetsten Tools nach aktuellem Stand gehören unter anderem Power BI und Tableau.

3.3 Herausforderung bei der visuellen Datenexploration

Auch wenn der Zugang zu hochdynamischen Visualisierungen dank moderner Tools leichter denn je ist, so muss ein Nutzer weiterhin Entscheidungen treffen, welche Erfahrung auf dem Gebiet der Datenanalyse und visuellen Datenexploration voraussetzen. Meist erlauben moderne Tools dem Nutzer, so lange technisch möglich, jeden gewünschten Visualisierungstyp zu generieren und ebenfalls die Visualisierung beliebig anzupassen. Die Verantwortung zu bewerten, ob die Visualisierung sinnvoll ist oder überhaupt richtig angewendet wurde, liegt dabei vollständig beim Anwender.

Zwar wird von den Tools in der Regel eine ansprechende Menge an Visualisierungstypen bereitgestellt, jedoch können dem Nutzer möglicherweise passendere Visualisierungen unbekannt bleiben, da auf diese entweder nicht hingewiesen wird oder diese in dem jeweils angewendeten Tool nicht zur Verfügung stehen.

Zudem kann der Nutzer an einen Punkt kommen, der für ihn „gut genug“ erscheint. Die meisten Nutzer kennen bereits viele Visualisierungstypen und suchen gezielt nach diesen. Selbst wenn eine bessere Alternative zur Verfügung steht, wird diese gegebenenfalls vom Nutzer nicht verwendet. Mögliche Gründe sind eine mangelnde Motivation, sich mit der Funktionsweise eines unbekannten Visualisierungstyps auseinanderzusetzen oder fehlende Kenntnis über geeignete Visualisierungstypen für den konkreten Anwendungsfall.

Die Herausforderungen bestehen also sowohl auf der Seite der Tools als auch auf der des Nutzers und sind dadurch nicht trivial.

3.4 Nachteile von Visualisierungen

Visualisierungen bringen oft eine reduzierte Genauigkeit bei der Darstellung der Daten mit sich und Variationen in der Darstellung können je nach Interpretation auch vollkommen andere Ergebnisse liefern. Daher ist Transparenz besonders wichtig, beispielsweise durch einen mitgegebenen Kontext. Erstellt man beispielsweise ein Liniendiagramm, welches Umsatzzahlen je Monat darstellt, jedoch als kumulierten Wert der zurückliegenden Monate des Jahres, so kann dies als ein stark positiv verlaufendes Jahr interpretiert werden, wohingegen in der Realität stagnierende Verkäufe vorliegen könnten. Dies kann die Folge sein, wenn ein individuelles Visualisierungskonzept, beispielsweise im Unternehmen, als allgemeine Norm fehlinterpretiert wird. Auch kann eine böswillige Absicht dahinterstecken, um Daten besser oder schlechter darzustellen und damit Investoren oder Kunden zu täuschen.

3.5 Regelwerke zur Datenvisualisierung

Um lesbare und einheitliche Visualisierungen zu erstellen, sind Regeln zur Gestaltung und zum Aufbau sehr hilfreich. Heutzutage werden viele davon bereits intuitiv umgesetzt. Wir verwenden Balkendiagramme, um Kategorien zu vergleichen, nutzen Farbpaletten, die die Unterscheidung von Visualisierungselementen erleichtern usw. Oft sind Nutzer mit solchen Regeln vertraut, weil sie deren Auswirkungen durch verschiedene Diagramme aus ihrem Alltag kennen und diese reproduzieren. Es gibt jedoch auch Regeln, welche aus den Visualisierungen nicht direkt ersichtlich sind oder im kleinen Rahmen scheinbar keinen großen Mehrwert geben und dem Nutzer somit erläutert werden müssen.

Abhilfe schaffen viele verschiedene Regelwerke, welche Nutzer beim Erstellen von Visualisierungen unterstützen sollen. Diese können branchenabhängig oder -unabhängig sein.

Branchenunabhängige Regelwerke unterscheiden sich meist in ihrem Umfang und Fokus. Meist unterscheiden sich diese jedoch nur bedingt in den empfohlenen Visualisierungstypen. Auch werden Regeln oft allgemein gehalten und sollen ein Gefühl vermitteln, welche Darstellungen vorteilhaft sind. Der Nutzer kann weiterhin seine Kreativität einfließen lassen und entscheiden, ob die Anforderungen erfüllt sind.

Ebenfalls existieren Regelwerke für bestimme Branchen. Ein Beispiel dafür sind die IBCS, welche sich mit der Gestaltung von finanzbezogenen Visualisierungen befassen. Diese definieren detaillierte Vorgaben, wie Visualisierungen aufgebaut sein sollen. Durch den Branchenfokus können häufig angewandte Kennzahlen einheitlich dargestellt werden, um eine Kommunikation und Analyse zu optimieren. Dadurch bleiben den Nutzern jedoch deutlich weniger Freiheiten bei der Gestaltung.

3.6 Tools mit ähnlichem Ziel

Im Folgenden werden zwei Tools vorgestellt, welche ein ähnliches Ziel wie diese Ausarbeitung verfolgen. Die Idee der automatisierten Datenvisualisierung ist nicht neu. Viele Tools erleichtern uns heute die Erstellung von Visualisierungen. Meist sind dafür nur wenige Klicks notwendig und wir bekommen Vorschläge für unterschiedliche Visualisierungstypen. Oft werden den Nutzern jedoch viele Freiheiten gelassen. Sie können einen beliebigen Visualisierungstyp auswählen, die Visualisierungsdimensionen beliebig zuordnen, die optische Gestaltung ändern oder Elemente wie eine Legende oder Datenbeschriftungen ausblenden. Anders machen es jedoch folgende Tools. Diese erstellen Visualisierungen ohne den Einfluss eines Nutzers auf deren Gestaltung.

3.6.1 DeepEye: Towards Automatic Data Visualization

Durch DeepEye sollen für einen relationalen Datensatz mit wenig Aufwand Visualisierungen zur Datenanalyse automatisiert erzeugt werden. Dieses Ziel wird durch 3 Techniken zu erreichen versucht:

1. Ein Algorithmus zur Bewertung, ob eine Visualisierung „gut“ oder „schlecht“ ist.
2. Die Einstufung, welche von 2 gegebenen Visualisierungen besser ist.
3. Das Generieren von Top-k Visualisierungen.

Anders als in meiner Arbeit wird hier, neben einer direkten Beurteilung der Effizienz einer Visualisierung, unter anderem stark auf Machine Learning gesetzt, wohingegen ich mich auf feste Regeln zur Wahl und Gestaltung von Visualisierungen konzentriert habe. Auch werden

hier nur wenige Visualisierungstypen angeboten (Balken-, Linien-, Punktwolken- und Kuchendiagramm). (Luo, et al., 2018)

3.6.2 Autoviz

Das Tool Autoviz analysiert anhand verschiedener Techniken Merkmale in den Daten, versucht dadurch interessante Muster zu erkennen und will im Anschluss diese über automatisch generierte Visualisierungen dem Nutzer darstellen. Dadurch soll sowohl der erste Einblick als auch die tiefere Analyse erleichtert werden. Durch die zusätzlich hohe Anzahl an verschiedenen integrierten Visualisierungstypen entstehen sehr ansprechende Ergebnisse. Ebenfalls werden interaktive Visualisierungen angeboten, was für eine praktische Anwendbarkeit essenziell ist. (Roth, 2019)

4 Daten

In diesem Kapitel wird erläutert, welche Eigenschaften Daten haben, welche Informationen bei der Analyse von diesen gewonnen werden können und welchen Einfluss diese auf die Erstellung von Visualisierungen haben.

4.1 Datenstruktur

In der Regel erfassen wir mit Daten beliebige Informationen zu Dingen, Ereignissen, Abläufen etc., folgend als Entitäten bezeichnet. Hierfür gibt es verschiedene Ansätze die jeweils Vor- und Nachteile mitbringen. Diese lassen sich einteilen in strukturierte, semi-strukturierte und unstrukturierte Daten. Die Struktur hat großen Einfluss auf die Datenanalyse und die Erstellung von Visualisierungen. Je strukturierter die Daten aufgebaut sind desto generischer können Analyseverfahren und Transformationen auf diese angewendet werden.

Strukturierte Daten sind meistens Daten in tabellarischer Form. Eine Entität wird dabei durch eine Zeile repräsentiert. Durch die Aufteilung in Spalten können wir beliebig viele unterschiedliche Eigenschaften dieser Entität erfassen. Alle Daten einer Spalte sollten dabei ein einheitliches Format haben. Dies erleichtert die Analyse stark. Diese Struktur eignet sich sehr gut für die Erstellung von Visualisierungen, da oft jeweils eine Spalte einer Visualisierungsdimension zugeordnet werden kann. Durch die Aufteilung der Informationen auf Spalten ist auch eine Analyse verglichen mit anderen Strukturen weniger aufwendig, da viele generische Techniken angewendet werden können. Die häufigsten Transformationen finden auf Spaltenebene statt, indem diese beispielsweise aufgeteilt, zusammengefasst oder inhaltlich umgewandelt werden.

Semi-strukturierte Daten finden wir oft im XML oder Json Format vor. Zwar ist eine Struktur vorhanden, diese wird jedoch von den Daten selbst getragen. So kann in einer XML-Datei eine Hierarchie dargestellt werden, indem eine Baumstruktur erzeugt wird, bei der jede Ebene des Baumes eine Ebene der Hierarchie abbildet. Zwar ist dies auch tabellarisch umsetzbar, die Information über den Aufbau der Hierarchie würde dabei jedoch verloren gehen. Werden in einem tabellarischen Datensatz beispielsweise die Information über das Jahr, den Monat

und den Tag separat in einzelnen Spalten gehalten, so können wir zwar interpretieren, wie diese Informationen zusammengehören. Dies gelingt jedoch nur mithilfe der allgemeinen Kenntnis darüber, wie ein Datum aufgebaut ist. Hier bieten semi-strukturierte Daten die Möglichkeit mehr Information durch ihren Aufbau zu tragen. Auch können Informationen über verschiedenartig aufgebaute Entitäten im selben Datensatz gehalten werden. Dies bringt jedoch einen Mehraufwand für eine Analyse bzw. Datenvisualisierung mit sich. So muss geklärt werden, welche Werte vergleichbare Eigenschaften darstellen bzw. gemeinsam darstellbar sind. Eine generische Behandlung ist somit nur stark eingeschränkt möglich, wodurch individuell angepasste Schritte notwendig sind.

Unstrukturierte Daten treten in verschiedenen Formen auf, wie etwa Texte, Bild-, Video- oder Audiodateien, und werden oft im Zusammenhang mit Big Data genannt. Namensgebend kann keine Struktur in ihnen ermittelt werden. So sind die Entitäten hier schwer zu ermitteln bzw. haben diese einen vollkommen anderen Maßstab als bei anderen Datenstrukturen. Eine automatisierte Analyse ist nur sehr bedingt möglich. Sie stellen dadurch die größte Herausforderung dar. Um deren Analyse zu erleichtern bzw. erst zu ermöglichen, müssen diese entweder in eine strukturierte Form umgewandelt werden, oder es werden Verfahren benötigt, welche die Daten anderweitig erweitern bzw. vergleichbar machen. So kann durch Vektorisierung von Texten mittels Machine Learning ermittelt werden, wie verschiedene Wörter zueinanderstehen und somit eine gewisse Vergleichbarkeit geschaffen werden.

4.2 Datentyp

Der Datentyp legt bereits viele Eigenschaften fest. Er limitiert das mögliche Skalenniveau und kann bereits festlegen, ob es sich um qualitative oder quantitative Daten handelt. Wir schauen uns nun einige Datentypen an und halten fest, welche weiteren Informationen wir aus diesen ermitteln können.

- **Ganzzahlen und Gleitkommazahlen (integer und float)**

  Diese beiden Datentypen werden in der Regel als stufenlose Eigenschaften wie Größe, Farbe/Helligkeit oder Position von Visualisierungselementen dargestellt. Auch ist eine indirekte Darstellung möglich, beispielsweise als Sortierung der Visualisierungselemente. Analysiert werden meist die Summe, der Durchschnitt, der Median, der Minimal- und

Maximalwert oder die Anzahl oder die Verteilung der Werte. Dafür ist das Skalierungsniveau entscheidend.

- **Text (string)**
  Texte sind aufwendiger zu analysieren als Zahlen. Es kann nur die Anzahl und Häufigkeit der Werte ermitteln werden. Die meisten Eigenschaften kommen hauptsächlich aus dem Skalenniveau. Es kann abseits von der alphabetischen Reihenfolge keine natürliche Ordnung in Texte oder Wörter interpretiert werden, wodurch diese ohne weitere Informationen immer nominal skaliert sind. Auch kann der Wertebereich begrenzt sein, was aus den Daten jedoch nicht ermittelbar ist. Texte sind qualitative Daten.
- **Ja/Nein (boolean)**
  Grundsätzlich lässt sich dieser Datentyp wie ein Text behandeln, jedoch ist es hier vorteilhaft, die Bedeutung der Werte zu kennen. Der feste Wertebereich mit 2 möglichen Werten erlaubt Visualisierungen zu nutzen, die hierfür optimiert sind.
- **Datum**
  Ein Datum ist ein kardinalskalierter Datentyp, der jedoch verschiedene Größen besitzt, anhand derer die Daten gruppiert werden können. So können Tage zu Monaten und Monate zu Jahren zusammengefasst werden.

4.3 Qualitative und Quantitative Daten

Qualitative und quantitative Daten lassen sich hauptsächlich als numerische und nicht numerische Daten unterteilen. In Visualisierungen treffen wir selten qualitative Daten allein an, da diese nicht mehr als eine Liste von möglichen Werten darstellen. Wir stellen quantitative Werte dar und geben ihnen eine visuelle Zuordnung zu qualitativen Daten. Dies ist besonders relevant für die Wahl des Visualisierungstyps, da dies die Wahl der Visualisierungsdimensionen einschränkt.

4.4 Skalenniveau

Jeder Menge von strukturierten Daten wird einem bestimmten Skalenniveau bzw. auch Messniveau zugeordnet. Hierdurch können wir feststellen, welche Analysemethoden sich für unsere Daten eignen. Welche Skalierung die richtige ist, hängt von verschiedenen Eigen-

schaften ab, welche jedoch je nach Betrachtung variieren können. Hierbei lassen sich folgende aufzählen:

- **Nominalskaliert**

  Diese Daten beinhalten am wenigsten Informationen. Hierunter fallen Datensätze, deren Werte keine natürliche Rangfolge aufweisen. Irrelevant ist hierbei der Datentyp. So können auch Zahlen nominalskaliert auftreten, wie die Postleitzahl oder eine Produktnummer. Weitere Beispiele hierfür sind Farben, Formen oder das Geschlecht. Mit ihnen können am wenigsten Analysen/Berechnungen vorgenommen werden. So lässt sich für das Lageparameter nur der Modus (das am häufigsten auftretende Element) berechnen.
- **Ordinalskaliert**

  Diese Daten besitzen zwar eine natürliche Rangfolge und können somit sortiert werden, es kann jedoch keine Aussage über die absoluten Abstände zwischen den Werten getätigt werden. Auch lassen sich dadurch keine Rechenoperationen zwischen den Werten durchführen. Hierunter fallen Schulnoten. Aus einer 1 und einer 3 wird keine 4. Auch ist der Abstand zwischen einer 1 und einer 2 anders als zwischen einer 3 und einer 4. Weitere Beispiele hierfür sind Dienstränge im Militär oder Tabellenplätze bei Sportturnieren. Hier lässt sich zusätzlich der Median berechnen als Lageparameter. Zwar wird bei Schulnoten im Alltag oft der Durschnitt berechnet, hierfür nehmen wir aber jedoch auch an, dass der absolute Abstand zwischen den Noten gleichmäßig ist, obwohl diese in der Regel für unterschiedlich große Punktebereiche stehen.
- **Intervallskaliert**

  Bei diesen Daten lässt sich sowohl eine Rangordnung als auch ein absoluter Abstand zwischen den Werten feststellen, jedoch kein natürlicher Nullpunkt. Sie ist eine der Kardinalskalierungen. Ein Beispiel wäre die Temperatur in Grad Celsius. Zwischen 10°C und 11°C besteht derselbe Abstand wie zwischen 20°C und 21°C. Durch den fehlenden Nullpunkt kann jedoch nichts über das Verhältnis zwischen den Werten gesagt werden. So sind 20°C nicht doppelt so warm wie 10°C (auf physikalischer Ebene). Hierdurch lässt sich zusätzlich das arithmetische Mittel berechnen (Durchschnitt).
- **Rationalskaliert**

  Die Rationalskala ist die zweite Kardinalskala und das höchste Skalenniveau in der Statistik. Zusätzlich zu einer Rangordnung der Werte und einem absoluten Abstand zwi-

schen diesen existiert ein natürlicher Nullpunkt. Beispiele hierfür sind Entfernung, Masse und Geschwindigkeit. Hier lassen sich Aussagen tätigen zu den Verhältnissen zwischen Werten. Ein Auto mit 50 km/h ist halb so schnell wie der Zug daneben mit 100 km/h. Dadurch lässt sich zusätzlich das geometrische Mittel berechnen.

5 Visualisierungen

5.1 Einleitung

In diesem Kapitel werden einige Visualisierungstypen vorgestellt. Grundsätzlich ist jede Art Daten visuell darzustellen, eine Form der Visualisierung. Als sinnvoll erachten wir jedoch erst diese, die entweder für uns eine große Datenmenge übersichtlich und greifbar machen, oder es uns erleichtern Muster in den Daten zu erkennen. Wie effektiv eine Visualisierung ist, hängt dabei von vielen Faktoren ab. So nutzt die Wissenschaft der Visualisierung von Daten unter anderem die Kenntnisse über die Farbenlehre, über den Aufbau des menschlichen Auges, die Psychophysik und die Kognitionspsychologie.

5.2 Visualisierungsdimensionen

Eine Visualisierung besteht in der Regel aus mehreren verschiedenen Visualisierungsdimensionen. Welche es sind hängt vom Visualisierungstyp ab. Auch wenn verschiedene Visualisierungstypen individuelle Visualisierungsdimensionen darstellen, so haben diese meist ebenfalls Dimensionen, welche in ähnlicher Form oft vorzufinden sind. Da diese generischen Visualisierungsdimensionen oft anzutreffen sind, ist deren Interpretation meist simpel. Es stellt sich die Frage, „was“ diese darstellen, denn mit dem „wie“ sind wir in der Regel vertraut. Hierunter fallen unter anderem die Farbe bzw. der Kontrast von Visualisierungselementen oder auch deren Größe und Position.

Individuelle Visualisierungsdimensionen stellen Werte entweder als eine Kombination von generischen Dimensionen dar oder auf eine von der Norm abweichende Art. Diese erfüllen in der Regel einen speziellen Zweck, um eine Interpretation zu vereinfachen oder Zusammenhänge oder Strukturen in den Daten (deutlicher) darzustellen.

Grundsätzlich kann jede Visualisierungsdimension jeden Wert darstellen, solange dieser den passenden Datentyp hat. Eine willkürliche Zuordnung erschwert jedoch die Interpretation. Neben intuitiven Regeln zur Dimensionszuordnung gibt es Regelwerke wie die IBCS, welche

bei Einhaltung die Qualität der Visualisierungen und somit der visuellen Datenexploration stark verbessern.

5.3 Visualisierungstypen

Auch wenn es eine sehr große Anzahl an verschiedenen Visualisierungstypen gibt, so sind diese oft Erweiterungen oder Kombinationen von simplen Visualisierungstypen. Die grundlegenden Visualisierungstypen werden zusammen mit ihren Stärken und Schwächen in diesem Kapitel beschrieben.

5.3.1 Säulen und Balkendiagramme

Ein Säulendiagramm stellt Zahlenwerte durch die Höhe von Rechtecken dar. Dabei steht jedes Rechteck für ein Element einer Kategorie. Ein Balkendiagramm ist ein um 90° gedrehtes Säulendiagramm und unterscheidet sich technisch ansonsten nicht. In der einfachen Form hat die Breite der Rechtecke hierbei keine Bedeutung. Es eignet sich besonders gut, um eine kleine bis mittlere Anzahl Elemente einer Kategorie darzustellen und zu vergleichen. Durch die simple und intuitiv verständliche Darstellung ist diese Visualisierung sehr verbreitet. Dank ihrer überschaubaren Menge an Eigenschaften sind die Informationen leicht zu

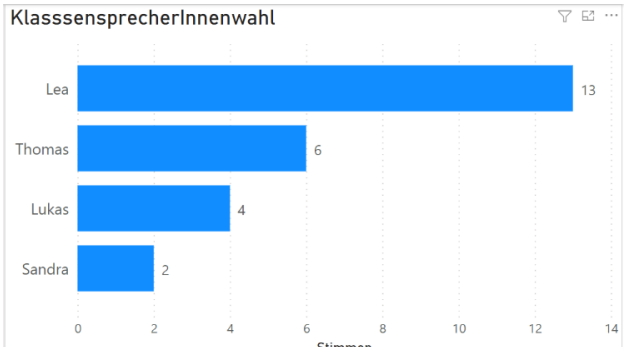

Abbildung 1: Balkendiagramm über die Abstimmung einer fiktiven Schulklasse. (Erstellt mittels Microsoft Power BI)

interpretieren. Die Darstellung der Höhe eines Zahlenwertes als Höhe einer Säule ist für uns sehr einfach zu verarbeiten und lässt einen Betrachter im Bruchteil einer Sekunde erkennen, ob und wie stark sich zwei oder mehrere Werte unterscheiden.

Da die Säulen durch kein grafisches Element direkt miteinander verbunden sind, stehen diese unabhängig voneinander für ihren jeweiligen Datenpunkt. Dadurch können unterschiedliche Sortierungen gewählt werden. So kann im Fall einer ordinalen Kategorie die natürliche Ordnung verwendet werden, was die Lesbarkeit steigert, oder eine Ordnung nach der Höhe der Säulen, je nach Verwendungszweck. Eine zufällige Sortierung sollte vermieden werden, da dies die Lesbarkeit verschlechtert und Potential schlicht ungenutzt lässt.

Die Wahl zwischen Säulen- und Balkendiagramm sollte immer abhängig davon getroffen werden, ob die Achse mit der Kategorie eine Zeit beinhaltet. Wenn ja, ist immer ein Säulendiagramm zu empfehlen.

Eine zu große Anzahl an Säulen kann es schwer machen Informationen zu interpretieren. Dadurch ist die effektive Menge an Visualisierungselementen limitiert. Folgende nennenswerte Variationen gibt es von diesem Visualisierungstyp:

- **Gestapeltes Säulen-/Balkendiagramm**
  Dies zeichnet sich durch eine weitere Achse aus, die die Säulen/Balken farblich unterteilt. Somit erlangt das Diagramm eine weitere Dimension. Die farblich dargestellte Achse lässt sich dadurch jedoch nicht so genau interpretieren wie die des einfachen Säulen-/Balkendiagramms und eignet sich dadurch nur für Kategorien, die nicht so detailliert betrachtet werden müssen.
- **Gruppiertes Säulen-/Balkendiagramm**
  Ähnlich dem gestapelten Säulendiagramm kann hier eine weitere Achse in das Diagramm eingefügt werden. Jedoch teilen sich die Säulen in Gruppen von mehreren aneinander liegenden Säulen auf. Dadurch kann die farbliche Achse besser analysiert werden. Anders als beim gestapelten Säulendiagramm verlieren wir jedoch die Darstellung des Gesamtwertes über die Säulengruppen.
- **Wasserfalldiagramm**
  Ähnlich dem Säulen-/Balkendiagramm, jedoch wird hier eine klare Ordnung impliziert, da ein Visualisierungselement in das nächste übergeht. Ebenfalls wird hier der Unterschied zwischen den benachbarten Visualisierungselementen stark in den Vordergrund gestellt.

5.3.2 Tortendiagramm

Dieser sehr simple Visualisierungstyp stellt numerische Werte abhängig von ihrer relativen Größe zum Gesamtwert als Abteile eines Kreises dar. Dieser Visualisierungstyp impliziert, dass die Kombination aller Werte einen Gesamtwert darstellt und bietet keine sinnvoll darstellbare Ordnung der Kategorien. Variationen dieses Visualisierungstyps sind:

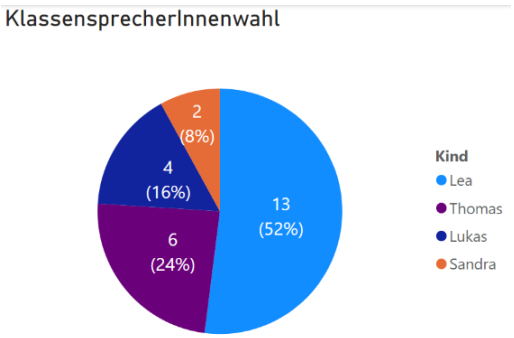

Abbildung 2: Kreisdiagramm über die Abstimmung einer fiktiven Schulklasse. (Erstellt mittels Microsoft Power BI)

- **Ringdiagramm**

  Das Ringdiagramm ähnelt stark dem Tortendiagramm, jedoch mit einer leeren Fläche in der Mitte. Meist wird in diesem Leerraum eine weitere Information oder ein Titel für die Visualisierung dargestellt, wodurch der Informationsgehalt bei gleicher Visualisierungsgröße erhöht wird.

- **Sunburstdiagramm**

  Diese Art der Visualisierung zeigt Hierarchien durch eine Reihe von Ringen, die für jede Kategorie in Segmente unterteilt sind. Jeder Ring entspricht einer Ebene in der Hierarchie, wobei der zentrale Kreis die Wurzel darstellt und sich die Hierarchie von ihr aus nach außen bewegt.

5.3.3 Streudiagramm

Diese auch als Punktwolken bekannten Visualisierungen stellen die Daten meist in einer nicht zusammengefassten Form dar und eignen sich dadurch besonders gut, um das Verhalten zwi-

schen zwei Variablen zu analysieren. Durch sich abzeichnende Muster lassen sich intuitiv Gruppen, Abweichungen oder Verläufe in den Daten erkennen.

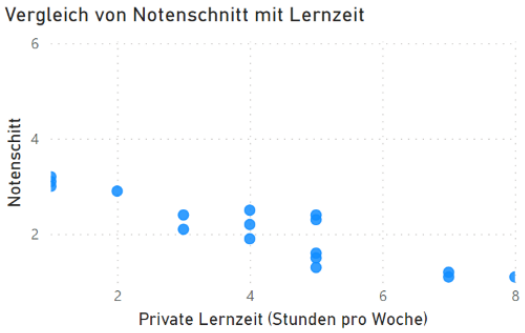

Abbildung 3: Streudiagramm über den Zusammenhang zwischen Lernzeit und Notendurschnitt einer fiktiven Schulklasse. (Erstellt mittels Microsoft Power BI)

Eine mögliche Variation des Streudiagramms ist:

- **3D-Streudiagramm**
  Hierbei wird das Diagramm um eine Z-Achse in die Tiefe erweitert, wodurch Muster als dreidimensionale Formen erkennbar werden. Dadurch sind sowohl Darstellung und Analyse dieses Diagrammes herausfordernd und nur mittels interaktiver Tools zu empfehlen.

5.3.4 Liniendiagramm

In diesem Visualisierungstyp werden die Daten durch eine Linie repräsentiert, welche entweder Punkte miteinander verbindet oder eine Kurve darstellt. Dadurch wird ein fließender Übergang von einem Datenpunkt in den nächsten impliziert. Hiermit werden also meist zeitliche Verläufe dargestellt. Ein Vorteil ist die Möglichkeit, mehrere Kategorien gleichzeitig für eine Periode zu vergleichen, indem mehrere Linien dargestellt werden.

Dieser Visualisierungstyp sollte nur verwendet werden, wenn ein klarer Verlauf von einem Datenpunkt in den nächsten dargestellt werden soll, da es sonst zu Fehlinterpretationen kommen kann.

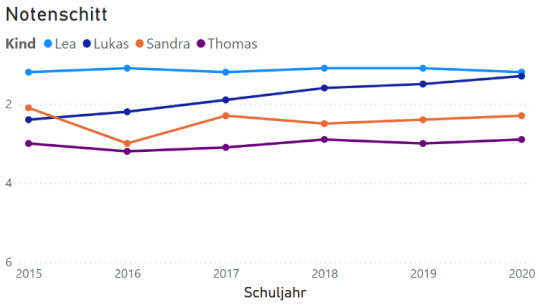

Abbildung 4: Liniendiagramm über den Verlauf des Notenschnitts einer fiktiven Schulklasse. (Erstellt mittels Microsoft Power BI)

5.3.5 Netzdiagramm

Ähnlich dem Liniendiagramm werden mehrere Datenpunkte durch eine Linie verbunden. Der Unterschied liegt jedoch bei der Anzahl an Achsen und der fehlenden Ordnung. Das Netzdiagramm benötigt mindesten 3 Achsen, da sonst nur eine gerade Linie dargestellt werden kann, aus welcher keine Informationen gezogen werden können. Technisch ist die Anzahl nach oben nicht limitiert, sollte der Leserlichkeit wegen jedoch nicht zu hoch ausfallen.

5.3.6 Paralleles Koordinatendiagramm

Dieser Visualisierungstyp dient dazu, hochdimensionale Daten darzustellen. Anders als beispielsweise beim Streudiagramm sind hier die Achsen namensgebend parallel zueinander angeordnet. Dadurch ist die Anzahl an Achsen erweiterbar. Eine Datenpunkt wird dabei durch eine Linie dargestellt, die durch jede Achse läuft und anhand der Höhe des Schnittpunktes ihre Werte darstellt.

5.4 Visualisierungsparadigmen

Wir sind sehr gut darin Muster in Visualisierungen zu erkennen. Um dies zu erleichtern, gibt es visuelle Konzepte, die wir intuitiv interpretieren. Ein größeres Element bedeutet einen

größeren Wert. Wenn es rote und grüne Elemente gibt, dann sind die grünen Werte die besseren. Dadurch haben wir bereits einen ersten Eindruck, was eine Visualisierung aussagen könnte, ohne uns mit dem Inhalt genauer auseinanderzusetzen. Detaillierte Informationen erhalten wir dann bei der genaueren Betrachtung. Leitet uns ein Konzept jedoch zu einem falschen Ersteindruck, so sind wir zu Anfang entweder verwirrt und müssen uns neu orientieren oder wir bemerken es nicht und ziehen die falschen Schlüsse. Diese Regeln bei der Visualisierungsgestaltung bezeichne ich als Visualisierungsparadigmen bzw. kurz als Paradigmen.

Dies sind einige Paradigmen für das Erstellen von Visualisierungen. Sie sind inspiriert von den Vorschlägen der Visualisierungswahl und Gestaltung der IBCS. Durch deren Einhaltung soll die Kommunikation von Informationen erleichtert, die Interpretation beschleunigt und die Wahrscheinlichkeit für Fehlannahmen reduziert werden.

1. Eine Zeitdimension wird primär auf der X-Achse, eine Nicht-Zeitdimension primär auf der Y-Achse dargestellt

   Dadurch kann eine Visualisierung auf den ersten Blick thematisch eingeordnet werden, ohne dass man sich inhaltlich weiter mit dieser auseinandersetzten muss. Wird dieses Paradigma einheitlich verwendet, kann der Nutzer bereits beim Überfliegen von Berichten erkennen, ob der Inhalt für ihn interessant ist und somit schneller die gesuchten Informationen finden.

   Zwar wird dieses Paradigma bereits häufig angewandt, die Anordnung der Visualisierungen in einem Bericht hat jedoch oft eine höhere Priorität. So kommt es dennoch vor, dass durch Platzmangel notwendige Visualisierungen mit vertauschten Achsen verwendet werden.

2. Achsen werden nicht gekürzt, sie zeigen immer Werte bis 0 oder bis zum Minimum und Maximum an

   Damit der relative Abstand zwischen zwei Visualisierungselementen korrekt interpretiert werden kann, muss dem Nutzer immer ein Kontext über die Größe der Visualisierungselemente gegeben sein. Sonst kann es zu Fehlannahmen wie der Annahme einer zu großen Abweichung führen.

Oft wird bei hohen Werten mit geringen Abweichungen in Verbindung mit zu wenig Platz das Kürzen der Achsen als Lösung gesehen, wodurch jedoch Abstände als viel zu groß interpretiert werden können.

3. Visualisierungselemente werden immer ausreichend beschriftet

Eine Visualisierung hat das Ziel, einem Betrachter die Erkennung von Mustern in den Daten oder den Vergleich von Werten zu erleichtern. Werden die Visualisierungselemente nicht ausreichend mit ihren Werten oder ihrer Bedeutung beschriftet, so können Fehlinterpretationen häufiger vorkommen.

Hier werden jedoch oft Ausnahmen gemacht, wenn zu viele Informationen in einem kleinen Bereich dargestellt werden. So werden entweder zu wenige Beschriftungen eingefügt, oder diese werden ganz weggelassen.

4. Farben werden seltener bzw. mit einer geringeren Priorität als Dimension verwendet

Erwähnt sei, dass damit nicht die grundsätzliche Verwendung von Farben vermieden werden soll, sondern Farben dem Zweck entsprechend sinnvoll eingesetzt werden sollten. Farben bieten oft eine weniger präzise Information als beispielsweise eine Beschriftung oder ein anderer Visualisierungstyp, welcher die Dimension anders darstellt.

Ein Bericht mit mehrfarbigen Visualisierungen wird oft als optisch ansprechender bewertet und ihm wird subjektiv mehr Qualität zugesprochen, obwohl hierdurch kein Mehrwert gegeben ist und es bessere Alternativen in Bezug auf die Lesbarkeit gibt.

6 Umsetzung

In diesem Kapitel beschreibe ich die praktische Umsetzung dieser Arbeit und welche Erkenntnisse ich dabei gesammelt habe. Entwickelt habe ich diese vollständig in Python 3. Das wichtigste Auswahlkriterium waren dabei die vielen Bibliotheken, die sowohl die Datenanalyse als auch die Erstellung von Visualisierungen vereinfachen.

6.1 Bewertung des Ergebnisses und Umfang

Um die Qualität der im Rahmen dieser Ausarbeitung erstellten Umsetzung sicherzustellen, braucht es qualitative Merkmale und Ziele abseits der Zielsetzung aus der Einleitung.

Die Effizienz bzw. Qualität der Umsetzung bewerte ich anhand dessen, wie viel Aufwand der Nutzer bis zur ersten Visualisierung betreiben muss, wie viele unbrauchbare Visualisierungen entstehen und ob eine nützliche Visualisierung durch das Tool erstellt wurde. Zudem ist ein hohes Maß an Transparenz notwendig, damit der Nutzer keine falschen Schlüsse aus den Ergebnissen zieht. Ebenfalls ist relevant, wie präzise die Visualisierungsparadigmen aus Kapitel 5.4 eingebunden wurden.

Es gibt viele verschiedene Visualisierungstypen und diese können in einer Vielzahl von Szenerien verwendet werden. Das vollständige Abdecken von alldem ist in diesem Rahmen nicht möglich. Deswegen reduziere ich den Umfang sowohl bei der Anzahl der angebotenen Visualisierungstypen als auch bei den visualisierbaren Kombinationen von Daten. Ich halte dies nicht für notwendig, um die Umsetzbarkeit zu prüfen.

6.2 Zielgruppe

Hier wird die Zielgruppe für diese Umsetzung beschrieben bzw. erwartete Fähigkeiten und Kenntnisse des Nutzers definiert. Der Nutzer soll einfache IT-Kenntnisse und erste Erfahrungen mit Visualisierungen haben. Er soll in der Lage sein, selbstständig eine Visualisierung zu verstehen, wenn ihm genügend Zusatzinformationen gegeben werden, und er ist mit seinen zu visualisierenden Daten insoweit vertraut, dass er deren Aufbau und Bedeutung kennt.

6.3 Konzept der visuellen Datenexploration

Inspiriert durch das Visual information seeking mantra habe ich versucht, die visuelle Datenanalyse so zu gestalten, dass der Nutzer sich zuerst einen groben Überblick verschaffen kann. Dann kann dieser die Daten filtern und die betrachteten Bereiche eingrenzen, um somit auf Details zu stoßen, die er genauer analysieren kann. Hierdurch braucht der Nutzer mit seinen Daten nicht inhaltlich tiefer vertraut zu sein, da er so Stück für Stück seine Fragen beantworten und neue Fragen stellen kann.

6.4 Datenstruktur

Wie effizient ein Datensatz analysiert werden kann, hängt neben der Datenqualität stark von der Datenstruktur ab.

Bei meiner Umsetzung einer automatisierten Datenvisualisierung erwiesen sich strukturierte Daten als ideal. Semi-strukturierte Daten bringen durch ihre Freiheiten im Aufbau Herausforderungen mit. Unstrukturierte Daten bieten hierbei am wenigsten Vorteile. Je strukturierter die Daten sind, desto effektiver bzw. simpler lassen sich diese automatisiert visualisieren und somit visuell explorieren.

Erklären lässt sich dies durch die Form der Daten, welche Visualisierungen zu ihrer Erstellung benötigen. Grundsätzlich stellt ein Visualisierungselement Werte aus dem Datensatz dar. Je simpler diese Werte zu ermitteln bzw. berechnen sind, desto einfacher kann eine Visualisierung erstellt werden.

Um dies zu erläutern, schauen wir uns einen einfachen Vergleich zwischen strukturierten und semi-strukturierten Daten an: Ein Unternehmen speichert Informationen über seine Produkte, wobei es 2 unterschiedliche Produktgruppen gibt, Bücher und Kaffeebohnen. Als strukturierte Daten wäre dies eine Tabelle mit Spalten für jede Eigenschaft. Für semi-strukturierte Daten bietet sich eine XML-Datei an.

	Kategorie	Sub-Kategorie	Größeneinheit	Menge	Bezeichnung	Autor	Lieferant
1	Kaffee	Entkoffeiniert	Gramm	500	Tchibo - Entkoffeiniert		Tchibo
2	Kaffee	Entkoffeiniert	Gramm	500	Jacobs Krönung - Entkoffeiniert		Jacobs Krönung
3	Kaffee	Koffeinhaltig	Gramm	1000	Tchibo - Crema		Tchibo
4	Kaffee	Koffeinhaltig	Gramm	1000	Jacobs Krönung - Crema		Jacobs Krönung
5	Kaffee	Koffeinhaltig	Gramm	1000	Tchibo - Espresso		Tchibo
6	Buch	Thriller	Seitenzahl	368	Das Paket	Sebastian Fitzek	Knaur Taschenbuch
7	Buch	Thriller	Seitenzahl	432	Passagier 23	Sebastian Fitzek	Knaur Taschenbuch
8	Buch	Fantasy	Seitenzahl	335	Harry Potter und der Stein der Weisen	J.K. Rowling	Carlsen
9	Buch	Fantasy	Seitenzahl	352	Harry Potter und die Kammer des Schreckens	J.K. Rowling	Carlsen
+							

Abbildung 5: Tabelle mit Produktinformationen

```xml
<?xml version="1.0"?>
<Produktinformationen>
    <Kaffee>
        <Bohnen>
            <Bezeichnung>Tchibo - Entkoffeiniert</Bezeichnung>
            <Art>Entkoffeiniert</Art>
            <Gewicht>500</Gewicht>
            <Marke>Tchibo</Marke>
        </Bohnen>
        <Bohnen>
            <Bezeichnung>Jacobs Krönung - Entkoffeiniert</Bezeichnung>
            <Art>Entkoffeiniert</Art>
            <Gewicht>500</Gewicht>
            <Marke>Jacobs Krönung</Marke>
        </Bohnen>
        <Bohnen>
        <Bohnen>
        <Bohnen>
    </Kaffee>
    <Buecher>
        <Autor>
            <Name>Sebastian Fitzek</Name>
            <Genre>Thriller</Genre>
            <Verlag>Knaur Taschenbuch</Verlag>
            <Buch>
                <Bezeichnung>Das Paket</Bezeichnung>
                <Seitenzahl>368</Seitenzahl>
            </Buch>
            <Buch>
        </Autor>
        <Autor>
            <Name>J.K. Rowling</Name>
            <Genre>Fantasy</Genre>
            <Verlag>Carlsen</Verlag>
            <Buch>
                <Bezeichnung>Harry Potter und der Stein der Weisen</Bezeichnung>
                <Seitenzahl>335</Seitenzahl>
            </Buch>
            <Buch>
        </Autor>
    </Buecher>
</Produktinformationen>
```

Abbildung 6: XML-Datei mit Produktinformationen
(Elemente aus Platzgründen eingeklappt)

Da die beiden Produktgruppen unterschiedliche Eigenschaften aufweisen, muss hiermit entsprechend der gewählten Struktur umgegangen werden. Während in einer Tabelle beispielsweise das Genre der Bücher und die Kaffeeart als „Sub-Kategorie“ zusammengefasst werden, können diese Informationen im XML präzise benannt werden. Der wohl größte Vorteil bietet jedoch der Aufbau der XML-Datei. Zwar können wir in der Tabelle erkennen, dass jeder Autor nur in je einer Sub-Kategorie auftaucht. Ob dies jedoch nur Zufall ist, können wir nicht sagen. In der XML-Datei ist dies jedoch so hinterlegt, dass das Genre vom Autor abhängt, da dies eine direkte Eigenschaft von diesem ist. Es scheint also, als wäre die XML-Datei eine bessere Struktur für diesen Datensatz.

Wollen wir jedoch nun automatisch eine Visualisierung erstellen, so müssen wir zwei essenzielle Fragen klären. Welche Visualisierungsart erstellen wir? Und was stellen die Visualisierungsachsen dar? Gehen wir von einem einfachen Balkendiagramm aus:

- Eine Tabelle bieten eine generische Lösung, indem die zu analysierenden Spalten je einer Visualisierungsachse zugeordnet werden. Die Zuordnung wird dabei anhand der Spalteneigenschaften in deren Inhalt sinnvoll gewählt.
- Bei der XML-Datei kann hier keine generische Regel angewendet werden. Weder die Ebenen, die Namen der Eigenschaften noch die Anordnung sind hier einheitlich gewählt. Auch unterscheiden sich die tiefsten Elemente in der Anzahl der Eigenschaften. Um hier also ein sinnvolles Ergebnis zu erstellen, braucht es zusätzliche Informationen vom Nutzer. Ohne diese kann in dem Fall nicht einmal ermittelt werden, worum es sich bei den tatsächlichen Produkten handelt.

Wir sehen, die beiden Datenstrukturen unterscheiden sich stark im Aufwand der Vorbereitung. Natürlich ist der Aufbau der XML-Datei nicht optimal für diese Umsetzung und ein anderer Aufbau könnte ähnlich gut, wie eine Tabelle verwendet werden. Dies kann jedoch nicht von einem Nutzer erwartet werden, da dies bereits beim Erstellen der XML-Datei berücksichtigt werden muss.

Dies ist ein komplexes Problem, bei dem durch simple Ansätze keine zufriedenstellenden und generischen Ergebnisse entstanden. Durch eine Umwandlung in eine flache Tabellenform, bei der Werte mit gleicher Bezeichnung und gleicher Tiefe zu einer Spalte zusammengefasst werden, konnte ich simple Datensätze durchaus analysieren und Visualisierungen erstellen. Diese Methode setzte jedoch voraus, dass in den Daten einheitliche Entitäten abgebildet werden. Abweichungen müssen händisch vom Nutzer angepasst werden. Dafür müssen die Gründe für das nicht optimale Ergebnis ermittelt werden und ein intuitiver Weg existieren diese anzupassen. Es wird also ein Konzept benötigt, wodurch unter anderem vergleichbare Informationen vom Nutzer gruppiert werden können.

Da sich dies während der Umsetzung als eine aufwendige Herausforderung herausstellte, habe ich mich schließlich auf strukturierte Daten im tabellarischen Format konzentriert und alle anderen Ansätze verworfen. Diese sind stark verbreitet und bieten für eine visuelle Exploration die meisten Vorteile. Semi-strukturelle Daten bieten sich durch ihre Eigenschaften für eine spätere Iteration als eine Erweiterung an. Dasselbe gilt ebenfalls für unstrukturierte Daten.

6.5 Nutzerschnittstelle und Eingabe

Um dem Nutzer eine einfache und angenehme Anwendung zu bieten, braucht mein Tool eine intuitive und komfortable Nutzerschnittstelle.

Eine Herausforderung dabei war es, eine Lösung zu finden, wie für den Nutzer mehrere Visualisierungen darzustellen sind, ohne dass dies für ihn überladen wirkt. Es sollen sowohl alle notwenigen Eingaben und Datenanpassungen direkt im Tool möglich sein als auch das Navigieren und Interagieren mit den verschiedenen Visualisierungen.

Hierfür habe ich mittels der Bibliothek „Dash“ eine Oberfläche entwickelt, welche über mehrere Seiten die notwendigen Eingaben abfragt und die Visualisierungen darstellt.

Ziel war es, dass die Lösung erweiterbar ist und dass der Nutzer mit so wenig Eingaben wie möglich zu einem Ergebnis kommt. Grundsätzlich sollten keine Informationen mehrfach eingegeben werden müssen. Im finalen Ergebnis stehen dem Nutzer die folgenden Seiten zur Verfügung, durch welche dieser Schritt für Schritt navigiert:

1. Auswahl der Art der Datenquelle
2. Anbindung der Datenquelle
3. Anpassung und Erweiterung der Daten
4. Auswahl und Erforschung der Visualisierung

Als erstes muss der Nutzer eine Datenquelle auswählen und die dafür erforderlichen Verbindungsinformationen angeben. Da in einer Datenquelle oft nicht alle Daten relevant sind, können diese hier ausgewählt und reduziert werden.

Da die Daten nicht immer in optimaler Form vorliegen, kann der Nutzer als nächstes seine Daten anpassen und/oder erweitern. Dies geschieht auf Spaltenebene. Hierbei hat er folgende Optionen:

- Ersetzen eines Wertes in einer Spalte
- Zeilen mit einem bestimmen Wert in einer Spalte entfernen
- Datentyp einer Spalte ändern
- Eine Spalte duplizieren

- Das Skalenniveau einer Spalte anpassen

Hiernach kann er eine Teilmenge der geladenen und erzeugten Spalten und die Art der Datenzusammenfassung auswählen. Anhand dieser Auswahl werden nun Visualisierungen generiert.

Der Nutzer kann nun zwischen den Visualisierungen wählen und diese erforschen. Es wird jedoch nur eine Visualisierung zur gleichen Zeit dargestellt. Dies fördert die Übersichtlichkeit. Er kann in den Visualisierungen auf Teilbereiche zoomen, Bereiche markieren oder Visualisierungselemente mit Hilfe einer Legende ausblenden, wenn diese vorhanden ist.

Der Nutzer kann immer wieder zu den einzelnen Seiten zurücknavigieren und dort gewonnene Erkenntnisse miteinfließen lassen.

6.6 Datenquelle und Datenhaltung

Um die Komplexität zu reduzieren habe ich mich auf eine Datenquelle konzentriert. Meine Wahl fiel auf eine MSSQL Datenbank. Hierfür nutze ich die Python Bibliothek „Pandas“. Durch diese kann man die Daten laden - im Fall einer Datenbank mittels einer SQL-Abfrage - und in sogenannten „Data Frames“ speichern. Den Code hierfür erzeuge ich automatisch anhand der Nutzereingaben. Ein Data Frame ist eine zweidimensionale, größenveränderliche, potenziell heterogene tabellarische Datenstruktur.

6.7 Datenanalyse

Um eine bessere visuelle Darstellung zu ermöglichen habe ich neben der Analyse der Metadaten auch eine automatische Analyse der Daten selbst vorgenommen. Dank der „Pandas“ Bibliothek stehen sowohl Möglichkeiten zum Filtern der Daten als auch Optionen zum Sortieren, Zusammenfassen und Auswerten von Grenzwerten zur Verfügung. Ich prüfe in jeder Spalte auf verschiedene Eigenschaften, indem ich über die Zeilen iteriere. Je mehr Zeilen es in den Daten gibt, desto genauer sind die in diesem Schritt gewonnen Annahmen. Zwar lassen sich so viele Informationen gewinnen, die für die Erstellung der Visualisierungen vorteilhaft sind. Eine absolute Gewissheit ist jedoch selten gegeben. Optional kann man hier dem Nutzer die Möglichkeit geben, viele weitere Informationen anzugeben. Mein Fokus war es

jedoch, den Nutzer vor so wenig Aufwand wie möglich zu stellen, weshalb ich mich dagegen entschied. Unabhängig davon, wie nützlich die Eigenschaften letztendlich waren, habe ich versucht so viel wie möglich aus den Daten zu schließen.

Die wohl grundlegendste Unterscheidung bei den Daten ist diejenige zwischen Zahlenwerten und Nicht-Zahlenwerten bzw. quantitativen und qualitativen Datentypen. Diese Unterscheidung nehme ich anhand der Spaltentypen der Datenbank vor.

Nun werden die Daten schrittweise geprüft und folgende Eigenschaften je Spalte bzw. als Kombination aus Spalten gespeichert:

- Alle einzigartig: Wenn kein Wert in der Spalte mehr als einmal auftritt
- Alle gleich: Wenn alle Werte in der Spalte gleich sind (leere Werte zählen als eigener Wert)
- Beinhaltet leere Werte: Wenn die Spalte mindestens einen leeren Wert beinhaltet
- Ganzzahlentext: Wenn die Spalte vom Datentyp Text ist, jedoch alle Werte nur aus den Zeichen 0 bis 9 bestehen
- Gleitkommazahlentext: Wenn die Spalte den Datentyp Text hat, jedoch alle Werte eine Gleitkommazahl darstellen
- Alle Werte positiv: Wenn die Spalte nur positive Zahlenwerte beinhaltet
- Linear periodisch: Bei numerischen Spalten wird geprüft, ob alle Werte im sortierten Zustand den gleichen Abstand zueinander haben.
- Datumsperiodisch: Hierbei soll eine grobe zeitliche Periode erkannt werden. Existiert beispielsweise für jeden Monat ein Wert, jedoch immer nur der letzte Tag des Monats, dann schwankt zwar der Abstand in Tagen, jedoch interpretieren wir dies trotzdem als regelmäßig.
- Beinhaltet leeren Text: Wenn der Spaltentyp Text ist und einen Wert der Länge 0 beinhaltet
- Mehrere Wörter: Wenn der Spaltentyp Text ist und mindestens ein Wert zwischen zwei Nicht-Leerzeichen ein Leerzeichen beinhaltet
- Lange Wörter: Wenn der Spaltentyp Text ist und mindestens einer der Werte ein Wort beinhaltet, welches länger als 29 Zeichen ist (Trennzeichen ist das Leerzeichen)

- Langer Text: Wenn der Spaltentyp Text ist und mindestens ein Wert mehr als 59 Zeichen beinhaltet
- Maximale Dezimallänge: Für jede Spalte mit Zahlenwerten wird hier die maximale Anzahl an Dezimalstellen gespeichert.

Zwischen Spalten:

- In Beziehung: Wenn für jeden Wert **a** aus Spalte **A** immer auf den gleichen Wert **b** aus Spalte **B** geschlossen werden kann, dann wird [**A → B**] markiert.

Hier ist das Ausmaß der fehlenden Präzision in der Interpretation der Eigenschaften besonders deutlich geworden. Zwar kann bei einer entsprechenden Menge an Daten eine gewisse Genauigkeit erzielt werden, diese hängt jedoch stark von der analysierten Eigenschaft ab. Auch sind einige Eigenschaften wie Hierarchien oder die Bedeutung von leeren Werten nicht aus den Daten ablesbar.

6.8 Vorbereitung der Daten

Hier wird erläutert, wie die Daten für den Schritt der Visualisierungserstellung vorbereitet wurden. Aufgrund der Einschränkung auf strukturierte Daten werden keine strukturellen Transformationen benötigt.

Die Besonderheit des Datentyps „Datum“ ist, dass dieser sich sinnvoll gruppieren lässt. So können Daten anhand des Jahres, des Monats etc. zusammengefasst werden.

Für andere Datentypen werden keine Gruppierungen angewendet.

Eine besondere Betrachtung benötigen leere Werte. Diese können als Fehler in den Daten verstanden werden, wodurch diese möglicherweise ersetzt, entfernt oder unabhängig analysiert werden sollten. Sie können jedoch auch als bewusst gesetzter Wert verstanden werden, so dass sie mit in die allgemeine Analyse eingebunden werden müssen. Da beide Optionen im produktiven Alltag anzutreffen sind, wäre es nicht sinnvoll, nur eine der beiden Optionen als repräsentativ zu bestimmen und zu verwenden. Für meine erste Iteration müssen die leeren Werte ersetzt werden und übrige nicht vom Nutzer ersetzte Werte werden für den Schritt der Visualisierung entfernt, um die Komplexität zu reduzieren.

Je nach Auswahl der Zusammenfassungsart werden die Daten nun gegebenenfalls zusammengefasst.

Für einige Visualisierungen werden zusätzliche Informationen benötigt, welche aus verschiedenen Analysemethoden stammen:

Das Punktwolkendiagramm kann um eine Trendlinie erweitert werden, welche über das „ordinary least squares“-Verfahren berechnet wird.

Für das Boxplot werden sowohl Quartile als auch der Median berechnet.

Diese zusätzlichen Berechnungen und Darstellungen werden durch die „plotly“ Bibliothek nativ angeboten.

6.9 Anzahl der generierten Visualisierungen

In meinem ersten Ansatz habe ich das Ziel verfolgt, eine einzelne Visualisierung anhand der Daten und der angegebenen Zusatzinformationen des Nutzers zu erstellen. Dies resultierte jedoch immer in einem von zwei unzufriedenstellenden Ergebnissen: Entweder war der Aufwand für den Nutzer durch die große Menge an anzugebenden Zusatzinformationen so groß, dass meine Lösung kaum einen Mehrwert bot, oder die Visualisierung wurde durch zu viel Inhalt unübersichtlich bzw. zu schwer zu interpretieren. Ich möchte also die Frage klären, ob es grundsätzlich ein sinnvoller bzw. praktikabler Weg sein könnte, eine einzelne Visualisierung für einen definierten Datensatz zu generieren:

Auch simpel aufgebaute Daten können viele verschiedene Botschaften beinhalten. Welche Botschaft hier für den Nutzer am interessantesten ist, kann zwar über viele Fragen und Angaben erahnt, jedoch nicht genau ermittelt werden. Oft kann der Nutzer dies selbst nicht im Vorfeld mit Sicherheit bewerten. Ergibt sich bei der Datenanalyse, dass es viele oder besonders starke Ausreißer in den Werten gibt, so scheint dies ein guter Ansatz zu sein für einen Fokus. Gleichmäßigere Verläufe wären im Umkehrschluss weniger interessant. Betrachtet ein Nutzer beispielsweise die Verkaufszahlen mehrerer Produkte, so könnte es sein Ziel sein, besonders schlecht verkaufte Artikel zu ermitteln und aus dem Sortiment zu nehmen. Solch eine Annahme ist jedoch nicht in jeder Situation richtig. Verkaufszahlen sind an Sonn- und Feiertagen in der Regel besonders niedrig bzw. 0, da hier Geschäfte meist geschlossen sind.

Generieren wir nun eine Darstellung, die auf die stark abweichenden Werte fokussiert ist, so bekommen wir ein wenig interessantes Ergebnis. Nehmen wir nun an, wir fragen den Nutzer nach seinem Interesse, um auf dieser Grundlage Visualisierungen auszuwählen und zu gestalten. Wie oben bereits dargestellt, kann aufgrund einer falschen Annahme auch der Nutzer eine für ihn nicht optimale Auswahl treffen.

Da es also keinen sinnvollen Weg gibt, die interessanteste Botschaft zu ermitteln, müssten also verschiedene Interessen mit einer einzelnen Visualisierung abdecken werden. Auch wenn dies zu Anfang sinnvoll erscheint, darf hier nicht einfach ein Interesse ausgeschlossen werden, weil es sich in den Daten nicht signifikant abzeichnet. Gibt es in den Daten keine Ausreißer, so kann auch dies die gewünschte Botschaft sein. „Normale“ Verkaufszahlen an einem Sonntag, die denen eines Arbeitstages entsprechen, weisen auf einen Fehler in den Daten hin, welcher bei einer automatisierten Analyse mit Fokus auf Ausreißer jedoch nicht beachtet wird. Ein anderes Beispiel ist der zeitliche Verlauf der Daten. Werte, welche weder steigen noch fallen, können gar ein alarmierendes Ergebnis sein im Fall eines Unternehmens, welches das Ziel hat, seine Umsätze zu steigern. Es braucht also eine Autorität, welche die Interessen des Nutzers besser beurteilen kann als dieser im Vorfeld selbst. Somit ist der Ansatz einer einzelnen Visualisierung nicht ratsam. Damit sei - wenn auch in einem kleinen Rahmen - auch die Frage geklärt, ob es überhaupt eine Visualisierung gibt, mit der alle Interessen abgedeckt werden können.

Schauen wir uns dazu folgendes Szenario an: Uns liegen Verkaufsdaten über mehrere Jahre vor. Darin sind neben dem Datum und der Zeit auch die Quantität und Informationen über die verkauften Produkte enthalten. Wir stellen Fragen, welche wir über Visualisierungen beantworten möchten.

- Wie entwickelt sich unser Unternehmen?
- In welchen Monaten verkaufen wir am meisten Produkte?
- Zu welchen Zeiten kaufen die meisten Kunden ein?
- Gibt es Produkte, die häufig zusammengekauft werden?
- Gibt es eine Verbindung zwischen Produktart und Kaufdatum?

Wir fragen also nach allgemeinen Informationen, für die wir das große Ganze betrachten und Dimensionen ignorieren müssen, als auch nach Details, welche dadurch verschwindend klein werden. All das ist in einer Visualisierung nicht darzustellen. Somit können wir davon ausgehen, dass eine Visualisierung allein nicht ausreicht.

Ich erstelle somit mehrere Visualisierungen, die jeweils einen eigenen Vorteil bieten und reduziere so die Wahrscheinlichkeit, eine Frage des Nutzers nicht beantworten zu können.

6.10 Interaktion mit Visualisierungen

Wenn uns im Alltag ein Verhalten in beliebigen Daten dargestellt werden soll, dann geschieht dies in der Regel durch statische bzw. für uns nicht interaktive Visualisierungen. Die Wettervorhersage im Fernsehen, die aktuellen Aktienkurse oder auch aktuelle Pandemiezahlen, im Fall der Covid-19 Pandemie. Uns wird vorgegeben, welchen Teil der Daten wir sehen und wie dieser dargestellt wird. Das ist sinnvoll möglich, da im Vorhinein Einsichten und Erkenntnisse durch eine Analyse gewonnen wurden. In der Regel wird also eine Botschaft übermittelt, die die Visualisierung untermalen oder beweisen soll. Wie jedoch bereits erläutert, ist es im Rahmen einer automatisierten Analyse nicht möglich, die relevanteste Botschaft zu ermitteln und diese dem Nutzer zu präsentieren.

Eine Möglichkeit wäre, viele statische Visualisierungen zu erstellen, welche auf verschiedene Bereiche der Daten zoomen und so mit einer hohen Wahrscheinlichkeit dem Nutzer ein zufriedenstellendes Ergebnis bieten. Versuche so eine Anwendung umzusetzen erwiesen sich jedoch als nicht praktikabel. Selbst bei kleinen Datenmengen, welche mittels eines einzigen Visualisierungstyps ausreichend repräsentierbar waren, musste eine große und unüberschaubare Menge an statischen Visualisierungen erzeugt werden. Dadurch muss der Nutzer kontraintuitiv nach interessanten Mustern in den Visualisierungen suchen. Auch muss er sich jedes Mal neu orientieren, wenn er die nächste Visualisierung betrachtet. Hiervon ist also abzuraten.

Die meisten Anwendungen erzeugen heute hoch interaktive und dynamische Visualisierungen. So kann beispielsweise in Microsoft Power BI ein Balkendiagramm mithilfe eines sogenannten „Drilldowns“ interaktiv erforscht werden. Durch Klick auf ein Visualisierungsele-

ment wechselt die Visualisierung auf eine andere, vorher definierte Ebene und stellt somit eine Teilmenge der Daten dar. Indem der Nutzer mit dem Mauszeiger auf Visualisierungselemente zeigt, lassen sich bei Bedarf mehr Details einsehen und so trotz hoher Informationsmenge überschaubare Visualisierungen erstellen.

Die interaktive Erforschung von Visualisierungen bietet eine sehr tiefe und immersive Datenexploration, welche ich bei den heute sehr komplexen Daten für unerlässlich halte. Alle von mir erzeugten Visualisierungen haben also die Mindestanforderung, dass der Nutzer Details durch Zoomen vergrößern oder anderweitig hervorheben kann.

6.11 Wahl des Visualisierungstyps

Für die Wahl der zu erstellenden Visualisierungstypen habe ich mich für einen Ansatz mit klaren Regeln entschieden, unter welchen Bedingungen welcher Visualisierungstyp erstellt wird. Der Vorteil hier liegt sowohl in einer hohen Transparenz als auch in einer guten Anpassbarkeit und Erweiterbarkeit. Hierbei fließen die Visualisierungsparadigmen aus dem Kapitel 5.4 ein. Letztendlich entstand eine Baumstruktur. Die Knoten stellen dabei Kriterien dar und die Blätter eine Menge an Visualisierungstypen, welche letztendlich erstellt werden. Die verschiedenen Domains, auf welche ich im Kapitel 6.12 eingehe, bilde ich dabei über jeweils einen eigenen Baum ab. Dadurch können diese unabhängig voneinander angepasst und erweitert werden.

Einige der Kriterien liste ich hier auf:

- Anzahl der zu visualisierenden Spalten

Dies limitiert die Wahl der Visualisierungstypen stark. Effizient lassen sich 4 Spalten darstellen via X-Achse, Y-Achse und der Farbe und Größe des Visualisierungselements. Hierfür müssen diese jedoch auch ein notwendiges Skalenniveau aufweisen, da die X- und Y-Achse mindestens ordinal skaliert sein müssen. Würden wir noch eine weitere Dimension als Z-Achse einfügen, dann würde die Visualisierung schwer lesbar werden und somit ihren Nutzen verlieren. Es gibt jedoch Visualisierungstypen, wie das parallele Koordinatendiagramm oder das Radardiagramm, welche die Möglichkeit bieten noch mehr Dimensionen darzustellen.

- Qualitative und quantitative Werte

Qualitative Werte eignen sich zum Gruppieren der Daten. In der Regel stellt ein Visualisierungselement einen quantitativen Wert aus den Daten dar. Die Eigenschaften wie Lage, Größe etc. stellen quantitative Werte dar. Die Farbe der Visualisierungselemente kann dabei entweder einen quantitativen oder einen qualitativen Wert darstellen. Verschiedene Visualisierungstypen können unterschiedlich viele qualitative oder quantitative Ausprägungen darstellen. Das Punktwolkendiagramm als Beispiel hat für qualitative Ausprägungen nur die Farbe der Elemente zur Verfügung, wohingegen das gestapelte Balkendiagramm sowohl die einzelnen Balken als auch die Farben der Balken für die Darstellung verwenden kann.

- Art der Datenzusammenfassung

Manche Visualisierungstypen eignen sich nicht dazu auf bestimme Weise zusammengefasste Daten darzustellen. So wird in einem Punktwolkendiagramm in der Regel ein Punkt pro Zeile in den Daten dargestellt. Diese sind somit optimal für nicht zusammengefasste Daten. Zusammengefasste Daten haben meistens eine qualitative Ausprägung, anhand welcher diese gruppiert werden.

- Skalenniveau

Das Skalenniveau kann uns bei der Beantwortung einer Frage unterstützen: Inwieweit können die Daten sinnvoll statistisch analysiert werden?

Dies erfahren wir durch die Lageparameter und die messbaren Eigenschaften. Treffen wir auf nominalskalierte Daten, steht uns nur die Information über die Häufigkeit einzelner Elemente und damit der Modalwert zur Verfügung. Bei ordinalskalierten Daten erhalten wir zusätzlich den Median, da wir eine natürliche Ordnung in den Daten finden etc. Durch diese Information erlangen wir die Möglichkeit weitere Visualisierungselemente (bspw. eine Durchschnittskurve) in unsere Visualisierungen einzubauen.

- Vorhandene Datumsausprägung

Der Typ „Datum“ gibt durch seine Eigenschaften eine bestimmte Art der Darstellung vor. Manche Visualisierungstypen wie bspw. das Liniendiagramm benötigen einen kardinalskalierten Datensatz, da ein Datenpunkt mit weiteren Datenpunkten direkt verbunden ist. Aus

dem Vorhandensein eines Datums kann man eben dies schließen und diese Visualisierungen entsprechend umsetzen.

6.12 Domains

Aufgrund der Entscheidung mehrere Visualisierungen auf einmal zu generieren, braucht es ein Konzept, um dem Nutzer einen guten und klaren Überblick zu präsentieren, welche Visualisierungen zur Verfügung stehen und gleichzeitig die Wahl und Gestaltung der Visualisierungen nicht zu stark zu limitieren. Mein erster Ansatz war es, eine Limitation in den Visualisierungstypen einzuhalten, sodass nur eine Visualisierung je Visualisierungstyp möglich ist. Dies erwies sich jedoch als unpraktikabel. So stieß ich auf eine Grenze bei nicht qualitativ unterscheidbarer Zuordnung von Spalten zu Visualisierungsdimensionen. Es müsste hier eine zufällige Wahl getroffen werden. So können beispielsweise die Farbe und Größe der Visualisierungselemente jeweils eine Spalte in den Daten darstellen. Es kann jedoch einen großen Unterschied machen, wie diese zugeordnet sind bzw. können beide Zuordnungen einen Nutzen darstellen. Auch darf somit ein Visualisierungstyp nur auf eine Weise gestaltet und aufgebaut werden. Durch Designvorschläge wie die der IBCS wäre es jedoch ebenfalls möglich, Visualisierungen speziell für gewisse Branchen zu gestalten. Diese können unter anderem eine Farbpalette oder die Positionierung und den Aufbau von Elementen wie einer Legende oder Datenbeschriftungen vorgeben. Aufgrund der zu starken Limitation habe ich diesen Ansatz schließlich verworfen.

Diese Entscheidung erhöhte nun die Anzahl der erstellbaren Visualisierungen. Diese konnten zwar nach ihrem Typ geordnet werden, jedoch wurde es dadurch ebenfalls erschwert, den Überblick über alle Visualisierungen zu behalten.

Als finale Lösung entschied ich mich, die Visualisierungen in sogenannte Domains zu gruppieren. Diese können als Verwendungsgebiet oder Branche verstanden werden. Da aus den Daten nicht geschlossen werden kann, aus welchem Bereich diese kommen und ob der Nutzer nur einen allgemeinen Überblick wünscht oder eine tiefe statistische Analyse, halte ich diese Unterscheidung für notwendig. Umgesetzt ist dies durch eine Einteilung in Gruppen, welche die jeweiligen Namen der Domains tragen. In diesen werden die Visualisierungen nach dem Generieren zur Verfügung gestellt.

Im aktuellen Stand des Projekts gibt es folgenden Domains:

- **IBCS**
  Diese Visualisierungen sind nach den Vorschlägen des IBCS designt und dienen hauptsächlich dem Finanzwesen. Weitere Informationen hierfür sind in Kapitel 2.3 zu finden.
- **Standard**
  Diese Visualisierungen sollen einen einfachen Aufbau haben mit dem Fokus, dass der Nutzer sich mit dem Inhalt der Daten vertraut machen kann. Es kommen selten und nur simple zusätzliche Visualisierungselemente hinzu, welche aus den Daten errechnet wurden.
- **Analytics**
  Hier finden sich Visualisierungen mit zusätzlichen Analyseelementen, welche dem Zweck dienen, dem Nutzer einen tieferen Einblick in seine Daten zu gewähren.

6.13 Erstellen der Visualisierungen

Zur Erstellung der Visualisierungen gehören 3 Schritte: Es muss eine Visualisierungsbibliothek ausgesucht und eine passende Zuordnung von Daten und Visualisierungsdimensionen gewählt werden. Danach wird die Visualisierung optisch angepasst.

Um eine passende Visualisierungsbibliothek auszusuchen, habe ich verschiedene testweise verwendet. Hierbei habe ich folgende Eigenschaften geprüft und anhand dieser die Bibliotheken verglichen:

- Aufwand und Komfort bei der Erstellung von Visualisierungen
- Allgemeine Anpassbarkeit der Darstellung. Dazu zählen die Visualisierungselemente, der Hintergrund, die Legende, Schrift etc.
- Dynamische Anpassung der Darstellung. Passen Datenbeschriftungen sich automatisiert an, wechseln ihre Farbe, Größe oder Position automatisch, um die Lesbarkeit zu verbessern etc.
- Umfang der Visualisierungstypen

- Interaktivität der erstellten Visualisierungen. Da ich mich entschieden habe, ausschließlich dynamische Visualisierungen zu generieren, hatte diese Eigenschaft eine hohe Priorität.

Folgenden Bibliotheken habe ich miteinander verglichen:

- **Matplotlib**

  Hier werden sehr viele, meist einfach gestaltete Visualisierungen angeboten, welche einen großen Umfang an Anpassungsmöglichkeiten haben. Es fehlen jedoch native Funktionen wie automatische Datenbeschriftungen in den Visualisierungen. Diese müssen manuell eingebaut werden und erzeugen einen hohen Entwicklungsaufwand. Die generierten Visualisierungen bieten ebenfalls eine geringe Interaktivität.
- **Seaborn**

  Hier gibt es ebenfalls viele verschiedene Visualisierungstypen. Diese halte ich für grafisch ansprechender als die von Matplotlib. Hier ist die Anpassungsmöglichkeit ebenfalls etwas besser, da einige Funktionen hier nativ eingebaut sind und somit nur eingestellt werden müssen. Hier ist jedoch ebenfalls eine geringe Interaktivität mit den Visualisierungen gegeben.
- **Plotly**

  Zwar bietet diese Bibliothek eine im Vergleich limitierte Anzahl an Visualisierungstypen. Die erstellten Visualisierungen sind jedoch die meiner Meinung nach visuell ansprechendsten. Diese Bibliothek bietet ein hohes Maß an Funktionalitäten zur Anpassung an. Sie ist jedoch teils limitiert bei der manuellen Erweiterung von nicht nativ vorhandenen Funktionen, wodurch manche Darstellungen entweder gar nicht oder nur über einen Umweg umsetzbar waren. Den jedoch größten Vorteil bietet die dynamische und interaktive Verwendbarkeit der Visualisierungen.

Die Visualisierungen werden final ausschließlich durch die Bibliothek „plotly“ erzeugt. Ausschlaggebend waren das sehr hohe Maß an Interaktivität, welches durch andere Bibliotheken nur umständlich oder nicht in vergleichbarer Qualität umsetzbar ist, sowie die umfangreichen Anpassungsmöglichkeiten.

Für die Zuordnung der Visualisierungsdimensionen zu den Daten nutze ich hauptsächlich die Datentypen. Diese ermöglichen via Ausschlussverfahren meist bereits eine eindeutige Zuordnung. Hierbei fließen die Paradigmen aus Kapitel 5.4 und gegebenenfalls Vorgaben aus der jeweiligen Domain mit ein. Ist eine Zuordnung nicht eindeutig möglich und erfüllen alle Varianten einen sinnvollen Zweck, so erstelle ich alle möglichen Varianten der Visualisierung. Zwar könnte man hier ebenfalls dem Nutzer die Option geben, die Dimensionen aus seinen Daten mit den Visualisierungsdimensionen zu verbinden. Mein Ziel war es jedoch, den Nutzer so weit wie möglich von der Erstellung der Visualisierungen zu trennen. Die Anzahl der zusätzlich generierten Visualisierungen hatte dabei einen marginalen Einfluss auf die Übersichtlichkeit.

Bei der Gestaltung der Visualisierungen flossen ebenfalls die Paradigmen sowie gegebenenfalls vorhandene Vorgaben aus der jeweiligen Domain mit ein. Die „plotly“ Bibliothek erleichterte dies durch die umfangreiche Anpassbarkeit der erzeugten Visualisierungen. Ebenfalls bietet „plotly“ bei manchen Visualisierungen eine direkte Berechnung und Darstellung von Trendlinien an.

7 Auswertung

7.1 Zusammenfassung

Im Rahmen dieser Ausarbeitung konnte ich zeigen, dass es durch das Abbilden von Regeln zur Wahl und Gestaltung von Visualisierungen effektiv möglich ist, automatisiert Visualisierungen zu generieren. Ein Nutzer kann dadurch zum großen Teil vom Entwerfen der Visualisierungen abgekoppelt werden. Auch wenn stark abhängig vom gegebenen Datensatz, der Datenqualität und der Datenstruktur, können so unpassende Visualisierungen ermittelt und verworfen werden (zum Beispiel Liniendiagramme, welche nicht für die Darstellung von nominalen Kategorien verwendet werden sollten) und grundsätzlich als kritisch oder ineffizient angesehene Methoden vermieden werden (bspw. das Verschieben der X- oder Y-Achse vom 0-Wert, um Größenunterschiede zwischen Visualisierungselementen deutlicher darzustellen). Der Nutzer kann sich dadurch besser auf die Auswertung der Daten konzentrieren.

Darüber hinaus können somit Nutzer, welche unabhängig voneinander denselben Datensatz analysieren und Visualisierungen erstellen lassen, davon ausgehen, dass alle dasselbe Ergebnis erhalten. Dies kann den Austausch und die Kommunikation erleichtern.

Die Daten können dabei unterschiedlich viele Dimensionen haben, verschiedene Datentypen aufweisen etc. Es wird jedoch eine hohe Datenqualität und eine sinnvolle Struktur erwartet.

Beim Design der Visualisierungen konnte ich verschiedene Visualisierungsparadigmen beachten, ohne dass Mehraufwand beim Nutzer entstand.

Durch das Konzept der Domains konnte ich der Herausforderung überwinden, dass unterschiedliche Fachgebiete und Verwendungszwecke zum Teil individuelle Visualisierungstypen mit einer speziellen Gestaltung erfordern und somit die Qualität der Visualisierungen deutlich verbessern.

Bei der Umsetzung wurde jedoch auch deutlich, wie wichtig Transparenz und Grenzen für solch ein Konzept sind. In der heutigen Zeit sind bereits viele Dinge automatisiert. Wir sagen heute, *was* wir wollen und nicht *wie* wir es wollen. Das Navi im Auto zeigt uns den schnells-

ten Weg von A nach B, Musik-Apps im Smartphone zeigen uns die schönsten Lieder zu unserem gewünschten Genre und mit wenigen Worten wissen Suchmaschinen genau, was wir erfahren möchten. Die Frage, die wir uns dabei stellen sollten, ist also, wie viel Kontrolle wir hier abgeben wollen. Heute lernen wir mehr denn je aus Daten. Alles um uns herum produziert diese und wir sammeln sie selbst ohne die Notwendigkeit eines genauen Grundes. Da ist es nur verständlich, dass wir die Analyse von Daten so komfortabel wie möglich gestalten wollen. Durch meine Ausarbeitung konnte ich jedoch auch einen Einblick gewinnen, wie einfach es sein kann, durch Automatisierung eine gewisse Abhängigkeit zu schaffen. So wie der Orientierungssinn nicht mehr trainiert wird, wenn regelmäßig das Navigationssystem im Auto genutzt wird, kann auch das automatische Generieren von Visualisierungen unsere Fähigkeit mindern, Daten kritisch zu hinterfragen.

Während der Umsetzung stellten sich außerdem nach und nach auch Grenzen heraus. Einige Aspekte sind nicht oder nur unter großem Aufwand automatisiert umsetzbar. Auch wenn viele Informationen aus den Daten gewonnen werden können, bleiben diese nur Annahmen, mittels derer keine eindeutigen Aussagen über die Daten getroffen werden können. Es werden tiefere Analysen benötigt, um genauere Ergebnisse zu erzielen.

Diese weiteren Limitationen ergaben sich bei der Umsetzung:

- **Erkennung von Hierarchien**

  Enthalten 3 Spalten die jeweiligen Werte Land, Stadt und Postleitzahl, so lässt sich dies nur über den Abgleich mit einer Sammlung von vollständigen Ortsdaten ermitteln. Auch bei einem aufgeteilten Datum können die Werte abgeglichen werden. Dadurch können Daten aber auch falsch interpretiert werden, wodurch nie eine vollständige Genauigkeit gegeben ist. Basiert jedoch eine Hierarchie auf nicht öffentlich zugänglichen Daten bzw. enthält hierbei Fehler, so ist eine automatisierte Erkennung nicht möglich. Hier sind zusätzliche Informationen notwendig, welche nicht in den Daten enthalten sind.
- **Automatisierte Erkennung einer Zusammenfassungsart**

  Manche Daten lassen sich zusammenfassen und ergeben hierdurch Sinn, andere jedoch nicht. Um Visualisierungen zu erstellen, ist es oft notwendig eine Menge von Daten zusammenzufassen. Stückzahlen, die verkauft wurden, lassen sich sinnvoll auf Ebene der Zeit summieren. Auch ein Durchschnittswert kann wertvolle Informationen liefern. Die

Temperatur in Wetterdaten summiert auf der Ebene der Zeit bietet jedoch keine sinnvolle Information. Man würde solch eine Darstellung grundsätzlich als falsch ansehen. Hier wäre der Wert als Durchschnitt zusammengefasst sinnvoll. Es ist also eine inhaltliche Kenntnis der Daten von Nöten und eine Visualisierung somit ohne zusätzliche Informationen nicht automatisiert umsetzbar.

- **Interpretieren von leeren Werten**

  Daten können unsauber gepflegt sein oder Lücken aufweisen und dadurch leere Werte beinhalten. Leere Werte können jedoch auch bewusst gewählt sein. Es gibt viele Möglichkeiten, leere Werte zu verarbeiten. So können diese beim Bestimmen eines Durchschnitts beachtet oder ignoriert werden. Beide Optionen haben gegebenenfalls einen sinnvollen Verwendungszweck. Welche jedoch vom Anwender gewünscht ist, kann nicht ermitteln werden.

- **Interpretation des Datentyps**

  Auch wenn die Information über den Datentyp meist in einer Datenquelle gehalten wird, kann diese auch fehlen. Sind im Fall von strukturierten Daten in einer Spalte ausschließlich Zahlen vorhanden, so ist eine Aussage über den Datentyp ohne zusätzliche Informationen immer noch nicht genau möglich. Wird ein Datensatz erweitert, so kann sich beispielsweise der vorher interpretierte Datentyp einer Spalte unerwartet ändern und somit zu einem Fehler führen.

- **Skalenniveau der Daten automatisiert ermitteln**

  Werden Zahlen betrachtet, so können diese beispielsweise Postleitzahlen darstellen und somit nominalskaliert sein. Oder aber sie stellen ein Gewicht in Gramm dar und sind somit rational skaliert. Die automatisierte Auswahl des Skalenniveaus eines numerischen Datentyps ist hierbei somit kritisch zu betrachten. Bei Texten und binären Werten kann grundsätzlich von einer nominalen Skalierung ausgegangen werden, wenn keine Ordnung vorhanden ist.

7.2 Ausblick

Die Ausarbeitung hat gezeigt, dass eine automatisierte Datenvisualisierung die Datenanalyse und Kommunikation vereinfachen kann und ein hohes Potential besitzt. Für die Akzeptanz und somit die praktische Anwendung ist jedoch der Umfang essenziell. Hierfür können folgende mögliche Erweiterungen in die Umsetzung einfließen.

- **Semi-strukturierte und unstrukturierte Daten**

  Die zusätzlichen Informationen aus semi-strukturierten Daten haben ein hohes Potential, die visuelle Datenexploration zu bereichern. In der Regel ist deren Analyse für Nutzer aufwendiger als bei strukturierten Daten.

  Unstrukturierte Daten haben einen speziellen Aufbau und bieten viele tiefe Eigenschaften, welche untersucht werden können. Ein einfaches Beispiel wäre eine Sammlung von Fotodateien. Es wäre möglich, die Dateigröße darzustellen, oder aber die Verteilung der verwendeten Farben. Auch eine variierende Auflösung, das Seitenverhältnis, das Vorkommen von Mustern oder Strukturen oder gar das Aussehen eines bestimmten Bereiches wie dem Rand könnte analysiert werden. Anders als bei strukturierten Daten muss ein anderer Ansatz der Datenexploration angewandt werden, welche diese große Variabilität beachtet.

- **Weitere Datenquellen**

  Es kann zum einen die Konnektivität zu weiteren ähnlichen Datenquellen, wie anderen Datenbanktypen oder anderen tabellarisch aufgebauten Dateien erweitert werden. Dies sind simple Erweiterungen, die nur den Umfang, jedoch nicht die Komplexität erhöhen.

  Darüber hinaus können anders strukturierte Datenquellen verwendet werden, die weitere Informationen beinhalten, die für die Erstellung von Visualisierungen wertvoll sind.

- **Weitere Dateneigenschaften**

  Hier ist besonders die Einheit der Daten gemeint, speziell die von quantitativen Daten. Zwar ist diese selten in den Datenquellen zu finden, kann jedoch vom Nutzer angegeben werden und erweitert den Prozess der automatischen Visualisierungswahl stark. Auch

kann hierdurch die Nutzereingabe vereinfacht werden, da auf diese Weise viele Informationen abgeleitet werden können.

- **Weitere Domains**

  Wie anhand der IBCS Domain dargestellt, können Visualisierungen mit speziellem Aufbau für besondere Fachgebiete erstellt werden. Dies kann ebenfalls für weitere Verwendungszwecke und Fachgebiete umgesetzt werden.

- **Machine Learning**

  Durch den Einsatz von Machine Learning können die Daten um zusätzliche Informationen erweitert werden oder Muster automatisiert erkannt werden, um die Auswahl an Visualisierungen zu optimieren. Hierbei sollten aus meiner Sich klare Grenzen gesetzt sein. Automatisierte Erweiterungen und Anpassungen an den Daten müssen für den Nutzer verständlich und transparent einsehbar sein.

- **Weitere Visualisierungsparadigmen**

  Je mehr Darstellungsregeln für uns zur Norm werden, desto effektiver können wir Daten visuell explorieren und analysieren. Wichtig ist, dass diese universell angewendet werden können und sich nicht erheblich mit Visualisierungsregeln von wichtigen bzw. nennenswerten Fachgebieten widersprechen.

- **Barrierefreiheit von Visualisierungen**

  Durch das Anbieten einer farbenblindengerechten Farbpalette kann Menschen mit einer Einschränkung im Sehvermögen die Arbeit mit Berichten und Visualisierungen erleichtert oder gar erst ermöglicht werden.

  Es sind etwa 8% der biologisch männlichen Bevölkerung (biologische Frauen sind deutlich seltener betroffen) Europas farbenblind oder farbenfehlsichtig (Wissinger, et al., 2005). Diese Zahl wird von vielen Berichterstellern entweder ignoriert oder sie ist ihnen gar nicht bewusst. Priorität hat oft die Wiedererkennung der Gestaltungsrichtlinien des Unternehmens (beispielsweise das Firmenlogo) oder die persönliche Präferenz des Erstellers.

Literaturverzeichnis

Craft, B. & Cairns, P., 2005. *Beyond Guidelines: What Can We Learn from the Visual Information Seeking Mantra?,* London: IEEE.

Hichert, R. & Faisst, J., 2022. *International Business Communication Standards.* Hilden: IBCS Media.

Luo, Y., Qin, X., Tang, N. & Li, G., 2018. *DeepEye: Towards Automatic Data Visualization.* Paris: IEEE.

Moreland, K., 2013. *A Survey of Visualization Pipelines.* 19 Hrsg. s.l.:IEEE.

Roth, D., 2019. *towardsdatascience.com.* [Online]
Available at: https://towardsdatascience.com/autoviz-a-new-tool-for-automated-visualization-ec9c1744a6ad
[Zugriff am 14 Juli 2022].

Tuthill, K. & Van Wyk, R., 2003. *John Snow and the Broad Street Pump.* [Online]
Available at: https://www.ph.ucla.edu/epi/snow/snowcricketarticle.html
[Zugriff am 10 Juli 2022].

Wissinger, B., Kohl, S. & Labor, M., 2005. *Genetische Ursachen der Farbenblindheit.* s.l.:s.n.

Erklärung zur selbstständigen Bearbeitung einer Abschlussarbeit

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht.

_____	_____	_____
Ort	Datum	Unterschrift im Original