**Projektbericht zum Modul Information Retrival und**

**Visualisierung Sommersemester 2021**

**Titel des Dokuments**

Richard Brennecke

Matrikelnummer:

# Inhaltsverzeichnis

[1. Inhaltsverzeichnis 2](#_Toc79327777)

[2. Einleitung 3](#_Toc79327778)

[2.1 Anwendungshintergrund 3](#_Toc79327779)

[2.2 Zielgruppen 3](#_Toc79327780)

[2.3 Überblick und Beiträge 3](#_Toc79327781)

[3. Daten 4](#_Toc79327782)

[3.1 Technische Breitstellung der Daten 4](#_Toc79327783)

[3.2 Datenvorverarbeitung 4](#_Toc79327784)

[4. Visualisierung 5](#_Toc79327785)

[4.1 Analyse der Anwendungsfälle 5](#_Toc79327786)

[4.2 Anforderungen an die Visualisierungen 5](#_Toc79327787)

[4.3 Präsentation der Visualisierung 5](#_Toc79327788)

[4.3.1 Visualisierung Eins 5](#_Toc79327789)

[4.3.2 Visualisierung Zwei 5](#_Toc79327790)

[4.3.3 Visualisierung Drei 5](#_Toc79327791)

[4.4 Interaktion 5](#_Toc79327792)

[5. Implementierung 5](#_Toc79327793)

[6. Anwendungsfälle 6](#_Toc79327794)

[6.1 Anwendung Visualisierung Eins 6](#_Toc79327795)

[6.2 Anwendung Visualisierung Zwei 6](#_Toc79327796)

[6.3 Anwendung Visualisierung Drei 6](#_Toc79327797)

[7. Verwandte Arbeiten 6](#_Toc79327798)

[8. Zusammenfassung und Ausblick 6](#_Toc79327799)

[9. Anhang 7](#_Toc79327800)

Seitenabstand Ränder alle 2

Schriftart: Times new roman

Zeilenabstand: 1,5

Zitierung: mit eckige Klammer mit Nummer dahinter

# Einleitung

Weltweit wurden 2020 über mehr als 260 Millionen Hektoliter (mhl) Wein produziert [1]. Umgerechnet in Liter sind dies 26.000.000.000 Liter (26 Milliarden Liter). Davon wurden von Deutschland allein 8,4 mhl produziert [1]. Was jedoch von der Führenden Weinproduktionsnation Italien mit 49,1 mhl rund um das neunfache übertroffen wird [1]. Aus diesen Zahlen lässt sich bereits erahnen das es Weine in den verschiedensten Arten und Geschmäckern gibt. Aus dieser Vielfalt an Weinen wurde ein Datensatz zusammengestellt mit rund 21 Tausend Datensätzen rund um verschiedene Weine. Diese wurden dabei in den verschiedensten Kategorien bewertet und werden nun mithilfe von verschiedenen Visualisierungstechniken in diesem Projektbericht genauer analysiert.

Dabei ist das Hauptziel aus diesen verschiedenen Daten möglichst viel Erkenntnisse zu generieren umso vielleicht neue Einsichten in das Thema rund um die Weine zu erhalten. Insbesondere ob es verschiedene Zusammenhänge zwischen den verschiedenen Eigenschaften, welche in dem Datensatz analysiert worden sind, gibt. Darüber hinaus wäre es wichtig zu wissen, ob es auch einen Zusammenhang zwischen den Produktionsmengen von Weinen gibt und den verschieden Weinarten innerhalb eines Landes. Zusätzlich sollte überprüft werden, ob es zwischen den einzelnen Kategorien, welche im Datensatz behandelt worden sind, Eigenschaftsübergreifende Eigenschaften gibt. Diese Fragen werden anschließend in diesem Bericht mithilfe von den drei Visualisierungstechniken des Scatterplots, Baumhierarchie und Parallele Koordinaten beantwortet.

## Anwendungshintergrund

Die Visualisierungstechniken, die in diesem Projektbericht verwendet werden, sind die des Scatterplots, der Parallelen Koordinaten und der Baumhierachie. Diese werden nachfolgend kurz vorgestellt, um ein besseres Verständnis für diese zu schaffen.

Der Scatterplot stellt dabei zwei verschiedene numerische Variablen gegenüber, diese werden mithilfe von Punkten dargestellt. Die Position dieser Punkte gibt dabei den Wert auf der Horizontalen und vertikalen Achse an. Diese Darstellung ist besonders wichtig, um verschiedene Zusammenhänge zwischen zwei Variablen herauszufinden. So können diese Punkte als ganzes betrachtet werden verschiedene Muster anzeigen. So ist es möglich einfach verschiedene Korrelationsbeziehungen festzustellen. Bei solchen Fällen ist es dann einfach vorherzusagen wo ein gewisser Horizontaler Wert liegen werden würde wenn wir einen vertikalen Wert haben. Darüber hinaus ist es mithilfe eines Scatterplots auch möglich verschiedene Daten in verschiedene Gruppen zu unterteilen, wenn diese nahe beieinander liegen. So lassen sich Ausreißer oder Lücken in den Daten erkennen, was nützlich ist, wenn die Daten eingeteilt werden sollen. [2]

Bei den Parallelen Koordinaten handelt es sich um einen Ansatz Mehrdimensionale Daten zu analysieren. Dabei werden die Daten auf verschiedenen Achsen eingezeichnet, wobei jede Eigenschaft als eine Dimension dargestellt wird. Diese Daten werden über die verschiedenen Achsen miteinander verbunden. Die Stärke der parallelen Koordinaten liegt vor allem darin mehrdimensionalen Muster und Vergleiche zu tätigen. Um dies zu erreichen ist es wichtig nicht zu denken das Linien eine Kodierung von Zeitreihen darstellen (und somit eine Veränderung des Wertes von Zeitpunkt a nach Zeitpunkt b darstellen). Stattdessen stellen eine Linie im Parallelen Koordinatensystem eine Verbindung von einer Reihe von Werten dar. So lässt sich beispielsweise einfacher erkennen ob sich Werte innerhalb oder außerhalb des Durchschnittes stehen oder aber besondere Ausreißer darstellen. Weiterhin ist es auch so möglich generelle Aussagen über die verschiedenen Werte zu treffen, ob beispielsweise einige Werte insgesamt größer sind als andere die restlichen Werte. [3]

Die Baumhierachie ist eine Darstellung von Daten welche hierarchisch aufgebaut sind. Dabei beginnt die Baumhierachie bei einem Element und verzweigt sich dabei mindestens zweimal. Diese, durch die Verzweigung entstanden Elemente, können sich wiederrum auch verzweigen, um so immer weiter eine Hierarchische Beziehung darzustellen. Das fertige Diagramm ähnelt dabei einen Baum mit seinem Stamm und den Ästen. Dieses Diagramm kann dabei helfen von einer sehr generellen Kategorie mit feinen Detailstufen zu unterteilen. [4]

## Zielgruppen

* Weininteressierte
  + Vorwissen
    - Kaum bis gar nicht
  + Erkenntnisse:
    - Zusammenhang von verschiedenen Kriterien beim Wein
      * Zusammenhänge kurz erklären
    - Informationsgewinnung
      * Kennenlernen von Weinen
    - Entdeckung neuer Weine welche sie trinken möchten
* Weineinkäufer
  + Vorwissen
    - Vorhanden bis Exzellent
  + Erkenntnisse:
    - Entdecken von neuen Sorten die sein Sortiment ergänzen
    - Entdeckung von neuen Sorten die ggf. Außergewöhnlich sind
    - Beratung der Kunden die gewisse Vorlieben haben
* Weinexperte
  + Vorwissen
    - Gut bis ausgeprägt
  + Erkenntnisse:
    - Entdecken von neuen Sorten die seinem Geschmack entsprechen
    - Bessere Einschätzung seiner bisherigen Weine

## Überblick und Beiträge

* Erklären welche Daten verwendet wurden
  + Oberkategorien kurz erklären
* Visualisierungstechniken erklären
* Beiträge
  + Mehrwert der Techniken für die Darstellung der Daten
    - Sacatterplott -> Gegenüberstellung von Dimensionen

# Daten

* Beschreibung der gegebenen Daten
* Eignung der Daten für die Zielgruppen
  + Weintressierte
    - Gut
  + Weinexperte/ Weinverkäufer
    - Teilweise
    - Daten können unvollständig sein -> und haben zu wenig Aussagekraft mit Body, Süße usw.
* Fragestellungen
  + Gut da Dimensionen erkannt werden können
  + Herkunft der Daten erkennbar -> Teilweise aber unvollständig
* Daten Ergänzung
  + Mussten mit Geo Daten ergänzt werden für Baumhierarchie da sonst kein Ursprungsknoten
  + Und wo kommen die her?

## Technische Breitstellung der Daten

* Daten Zugänglich?
  + Sind über GitHub für die verschiedenen Darstellungen erreichbar
* Formate
  + CSV -> Für alle Daten die nicht ergänzt wurden
  + JSON -> Für alle Geo Daten
* Besonderheiten
  + CSV
    - 0 oder nichts bedeutet dort ist nichts vorhanden
    - Trennung durch normales komma
  + JSON
    - Nur Name und Beziehung (Eltern – Kind) in der Datei vorhanden
    - Länder welche keine Weine Produzieren wurden außen vor gelassen

## Datenvorverarbeitung

* Datenverarbeitungsschritte
  + Sichten der Daten (Umwandeln in besser Lesbares Format -> Excel)
  + Bearbeiten der Daten
    - Durchschnitte Bilden
    - Namen überarbeiten
    - Übersetzten
  + Rücküberführung zur CSV
* Daten weggelassen
  + Daten nicht mehr lesbar (eine Japanischer Wein wo nicht mehr Rückschlüsse gezogen werden könnten)
  + Nichts Werte
* Durschnitte
  + Gebildet über die Trinktemperatur, Alkoholgehalt
  + Daten konnten ansonsten nicht eingelesen werden
  + Außerdem lag der unterschied durchschnittlich nicht bei mehr als 2 Gard
* Aussagekräftiger?
  + So hat man noch Toleranz beim der Temperatur und Alkohol ohne das sich die anderen Werte ändern müssten (Body, Süße usw.)

# Visualisierung

* Analyse kann erst gemacht werden, wenn Visualisierungen fertig sind

## Analyse der Anwendungsaufgaben

* Analyse kann erst gemacht werden, wenn Visualisierungen fertig sind
* Anwendungsaufgaben
  + Wie helfen die Darstellungen die genannten Problemstellungen zu beantworten?
* Mentale Modelle
  + Welche Visualisieren eigenen sich um das alles zu kombinieren von Wissen und den Modellen (intuitiv erkennbar)

## Anforderungen an die Visualisierungen

* Analyse kann erst gemacht werden, wenn Visualisierungen fertig sind

## Präsentation der Visualisierung

* Analyse kann erst gemacht werden, wenn Visualisierungen fertig sind
* Vorstellen, Interaktivität, Designentschiedungen begründen Diskutieren wieso nicht anderen Techniken verwenden worden sind

### Visualisierung Eins

* Wird ein Scatterplot

### Visualisierung Zwei

* Wird Parallele Koordinaten

### Visualisierung Drei

* Wird eine Baumhierarchie

## Interaktion

* Scatterplot und Parallele Koordinaten
  + Buttons zum verändern/ verschieben der Dimensionen
* Baumhierarchie
  + Keine nur anschauen
* Zweck der Interaktion
* Warum wurden andere Interaktionen umgesetzt und nicht andere?
* Begründung Interaktion zwischen denen nicht mit dabei

# Implementierung

* Kann erst eingeschätzt werden, nachdem es fertig gestellt worden ist
  + Aktuell hoher Aufwand und nur Baumhierarchie konnte sehr einfach aus Übung übernommen werden
* Gliederung des ELM Codes
* Übungsadaption
* Datenstruktur Modells bei den verschiedenen Interaktionen
* Bei uns in einem Record gespeichert im Main und dann im Update wird auf einen record zugegriffen

# Anwendungsfälle

* Erst nach Fertigstellung der Visualisierungen möglich
* Spezifischer Anwedungsfall -> wo Muster da sind oder nicht was es zu was besonderen macht
* Relevanz für die Zielgruppe
* Möglichkeit Umsetzung mit anderen Personen

## Anwendung Visualisierung Eins

* Anwendungsfall für Scatterplot

## Anwendung Visualisierung Zwei

* Anwendungsfall für Parallele Koordinaten

## Anwendung Visualisierung Drei

* Anwendungsfall für Baumhierarchie

# Verwandte Arbeiten

* Aktuell noch nicht recherchiert
* Zwei Artikel diskutieren
  + Gemeinsamkeiten und Unterschiede dabei herausstellen

# Zusammenfassung und Ausblick

* Ausblick er bei fertigem Projekt möglich
* Zusammenfassung der Beiträge
* Mehrwert für Zielgruppe und Personen
* Erweiterungen für Ebene und Datenebene

# Literaturverzeichnis

[1] P. Roca, *State of the Vitivinicultural World in 2020,* o. O., **20.04.2021**.

[2] M. Yi, *A Complete Guide to Scatter Plots,* https://chartio.com/learn/charts/what-is-a-scatter-plot/, **2019**.

[3] S. Few, *Multivariate Analysis Using Parallel Coordinates,* http://www.perceptualedge.com/articles/b-eye/parallel\_coordinates.pdf, **2006**.

[4] American Society for Quality, *What is a Tree Diagram? Systemic or Hierarchy Analysis | ASQ,* https://asq.org/quality-resources/tree-diagram, **o. J.**

# Anhang