Projektbericht zum Modul Information Retrival und

Visualisierung Sommersemester 2021

Visualisierung von verschiedenen Weindaten

Richard Brennecke

Matrikelnummer:

Inhaltsverzeichnis

1. 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 besondere Datensätze gibt, welche in einigen Kategorien besonders hervorstechen. Oder auch ob generelle Trends oder zusammenhänge zwischen den Kategorien vorhanden sind. Diese Fragen werden anschließend in diesem Bericht mithilfe von den drei Visualisierungstechniken des Scatterplots, Baumhierarchie und Parallele Koordinaten beantwortet.

1.1 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]

Diese Visualisierungstechniken visualisieren verschiedene Daten und somit verschiedene Eigenschaften von Wein. Auf diese Eigenschaften wird nun kurz eingegangen, um so ein grundlegendes Verständnis für diese Eigenschaften zu schaffen. Die Eigenschaften, welche in den Visualisierungstechniken verwendet werden, sind Alkohol, Trinktemperatur, Süße, Säuregehalt, Körper, Gerbstoff und Jahr.

Der Alkoholgehalt des Weines entsteht bei der Gehrung des Weins. Dabei handelt es sich um einen natürlichen Prozess, welcher abläuft, nachdem die Trauben zerquetscht worden sind. Je nachdem wie süß und reif die Trauben während der Ernte gewesen sind was einen höheren Zuckergehalt zur Folge hat steigt der Alkoholgehalt des Weines. Der Alkoholgehalt eines Typischen Weines liegt dabei zwischen 9 und 14 Volumenprozent. Falls ein Wein eines höheren Alkoholgehaltes haben sollte, wurde bei der Produktion künstlich Zucker oder Alkohol hinzugefügt. Die Zugabe von Zucker geht dabei auch in der Regel mit einer Minderung der Güteklasse des Weines einher, da der Qualitätsanspruch bei diesen niedriger ist. [5]

Auch die Trinktemperatur hat einen Einfluss darauf wie Wein bei Verkostung schmeckt. Dabei besitzt jeder Wein in der Regel seine eigene ideale Trinktemperatur (welcher meist auf dem Etikett des Weines vermerkt worden ist). Eine Faustregel ist dabei für junge Rotweine eine Trinktemperatur von 14-15 Grad, Kräftige Rotweine hingegeben bei 15-17 Grad und schwere Rotweine bei 17-19 Grad. Falls Rotwein zu warm getrunken werden, kann der Geschmack nicht mehr richtig zur Geltung kommen und der Alkohol überwiegt in dem Wein. Darüber hinaus könnte es sein das dieser Wein brennt im Hals. Bei Weißweinen hingegen liegt die Temperatur niedriger als bei den Rotweinen. So besitzt ein junger Weißwein zwischen neun und elf Grad, ein Reifer Weißwein sollte bei elf bis 13 Grad getrunken werden. Falls Weißwein zu kalt getrunken werden, sollte überwiegt dabei die Säure und die restlichen Aromen des Weines verloren gehen. Da jeder Geschmack individuell ist, ist es auch möglich mit der verschiedenen Trinktemperaturen des Weines zu experimentieren. Als Faustregel gilt dabei auch sobald ein Wein kühler getrunken wird, verstärken sich die Säuren und Tannine wobei sich die Aromatik abschwächt. Wenn jedoch die Trinktemperatur erhöht wird kommt die Aromatik, Körper, Süße und der Alkohol des Weines mehr zum Vorschein. Wenn der Wein serviert wird, ist darüber hinaus in tendenziell etwas kühler als die optimale Trinktemperatur zu servieren, da dieser sich nach dem Ausschenken schnell um zwei bis drei grad erhöht. [6]

Bei der Süße des Weines handelt es sich um eine der wichtigsten Bestandteile des Weines und verleiht mit anderen Komponenten dem Wein erst seinen Geschmack. Die Süße des Weines beschreibt dabei den Restzucker innerhalb des Weines. Dieser kann je nach Anbaugebiet und Rebsorte auch unterschiedlich ausfallen. Dabei gilt aber vor allem je reifer die Trauben sind desto mehr süßer ist der Wein. Innerhalb des EU – Weingesetzes ist festgelegt das ein süßer (oder auch lieblicher) Wein als mindestens eines Restzuckergehalt von 45 Gramm pro Liter aufweisen muss. Darüber hinaus ist es möglich dieses zu erhöhen in dem man im Nachhinein noch Zucker hinzufügt, um die natürliche Süße zu erhöhen. Dieser Vorgang ist als „Aufspriten“ bekannt. [7]

Die Säure innerhalb eines Weines soll ihm die Forsche und Eleganz eines Weines geben. Diese entsteht gemeinsam an der Rebe beim Reifungsprozess dieser. Dabei steigt der Zuckergehalt der Traube an, während der Säuregehalt sinkt. Das Verhältnis von dieser Zu und Abnahme ist dabei nicht immer gleich. So können kühle Nächte den Säurerückgang verzögern. Im Wein sind besonders zwei verschiedene Säurearten anzutreffen, welche die Weinsäure und die Apfelsäure sind. Dabei ist die Weinsäure eine meist erwünschte Säure da diese den Wein weich und angenehm schmeckt macht. Die Apfelsäure hingehen macht den Wein kantig und hart. Dies kann bei Weißweinen zum teil erwünscht sei, bei Rotweinen jedoch wird diese meistens in Milchsäure umgewandelt. Darüber hinaus ist der Apfelsäuregehalt auch ein guter Indikator für die gute des Jahrganges, da dieser von der Witterung abhängt. So wird weniger Apfelsäure bei sonnigen Jahrgängen produziert und in kühlen Jahrgängen mehr. [8]

Bei dem Körper des Weines handelt es sich um eine Möglichkeit den Eindruck des Weines zu beschreiben welcher dieser im Mund hinterlässt. Der Körper beschreibt somit die Gerbstoffe, Restsüße und die Säure des Weins. Dabei wird dieser meistens in einen leichten, mittleren und voluminösen Körper unterschieden was zu der Bezeichnung leichtgewichtigem, mittelgewichtigen und schwergewichtigen Wein führt. So sind vor allem die Gerbstoffe für den Körper verantwortlich. So führt meistens ein höherer Gerbstoffgehalt zu einem voluminöseren Körper. Aber auch wenn der Alkohol kein Bestandteil des Körpers ist, spielt dieser beim Körper eine Rolle, denn je höher der Alkoholgehalt des Weines ist desto voluminöser erscheint dieser auch im Mund. [9]

Die Gerbstoffe (auch Tannine genannt) kommen primär im Wein vor bei dem die gaze Traube verarbeitet wird. Dies liegt daran, dass die meisten Gerbstoffe in dem Kern der Traube und in der Schale vorzufinden sind. Dementsprechend besitzt ein Rotwein mehr als ein Weißwein, da hierbei die gesamte Traube verarbeitet wird. Jedoch können auch bei der Gröhrung in Holzfässern weitere Gerbstoffe mit hinzukommen. Beim Trinken des Weines rufen die Gerbstoffe den herben bis bitteren Geschmack und das Gefühl sich die Mundschleimhäute zusammenziehen oder der Mund trocken ist, hervor. Je nach Stärke des Weines sind die Gerbstoffe im hinteren Bereich des Mundes zu finden, bei besonders Kräftigen Gerbstoffen auch im gesamten Mund. Zu viele Gerbstoffe verleihen den Wein einen eher starken herben oder bitteren Geschmack, zu wenig Gerbstoffe führt jedoch zu einem eher flachen Geschmack. Bei der Lagerung kann es darüber hinaus sein das sich die Gerbstoffe Moleküle zu größeren zusammenfügen und somit dem Wein seinem besonderen Geschmack verleihen. [10]

Ein großer Unterschied von Wein zu vielen anderen Getränken ist das dieser weiterhin in der Flasche reifen kann. Dies bedeutet das Weine in der Lage sind mit der Zeit an Qualität zu gewinnen. Jedoch haben die meisten Weine ihren Trinkhöhepunkt schon nach einigen Jahren ihrer Abfüllung erreicht und sollten deswegen nicht übermäßig lange gelagert werden. Da ansonsten mit der Zeit Sauerstoff an dem Verschluss deines Weines vorbeikommt und im schlimmsten Falle Essig bildet. Dabei hängt es vom Gerbstoffgehalt des Weines ab, wie viel Sauerstoff er verträgt. Dieser macht die Gerbstoffe nämlich weicher und somit kann sich die eine gute Frucht innerhalb des Weines Entwickeln. Somit eigenen sich tendenziell Rotweine mehr zum Lagern als Weißweine, da diese weniger Gerbstoffe beinhalten. Beispielsweise ist einer der ältesten Weine ein 1870er Chateau Lafite ein Rotwein. [11]

1.2 Zielgruppen

Für die verschiedene Visualisierungstechniken gibt es drei verschiedene Zielgruppen, welche potenziell an diesen einen Mehrwert finden könnte. Diese Zielgruppen sind: Weininteressierte, Weineinkäufer (-verkäufer) und der Weinexperte. Auf diese wird nachfolgend nun eingegangen.

Die Weininteressierten sind eine Gruppe welche gerne Wein trinken, welche jedoch aber kaum bis gar kein Vorwissen zu dem Thema Wein besitzen. Deswegen könnten diese mithilfe von

verschiedenen Visualisierungen neue Erkenntnisse gewinnen rund um das Thema der Weine. So könnten Sie die Zusammenhänge zwischen der Säure und der Süße des Weines erkennen. Darüber hinaus könnten sie mithilfe dieser Visualisierung die Weine, die diese bisher getrunken haben, viel besser einordnen und somit vielleicht auch neuen Weine entdecken, welche ihnen potenziell schmecken könnte.

Bei der Gruppe der Weineinkäufer bzw. Weinverkäufer handelt es sich um die Gruppe, welche das meiste Wissen über Weine besitzen sollte. Da diese ihre Kundenbetraten sollten und entsprechende gute Weine zu guten Preisen finden und einkaufen müssen. Somit ist ihr Vorwissen seht gut bis exzellent. Mithilfe der Visualisierungen wäre es möglich verschiedene neue Sorten zu entdecken welche in das Sortiment des einzelnen Händlers passen könnten. Darüber hinaus könnten neue oder außergewöhnliche Sorten entdeckt werden, welche vielleicht nur spezielle Kunden infrage kommen würde. Weiterhin könnte diese zur Kundenberatung verwendet werden, um die speziellen Geschmäcker aller Kunden effektiv abdecken zu können.

Die Weinexperten sind eine Gruppe von Leuten welche gerne Wein trinken und sich bereits mit diesem Thema auseinandergesetzt haben. Dementsprechend sollte ihr Wissenstand durchschnittlich bis gut sein. Diese Gruppe könnte mithilfe der Visualisierungen neue Sorten für sich entdecken, welche ihren Geschmack entspricht und so vielleicht noch tiefer in die verschiedenen Weinrichtungen einzutauchen. Weiterhin wäre es auch möglich mithilfe der verschiedenen Visualisierungen ihren Weingeschmack weiter zu erforschen, da diese mithilfe der Darstellungen diesen besser erforschen könnten.

1.3 Überblick und Beiträge

In diesem Projekt wurden die Daten von der Webseite Kaggle verwendet. Dabei handelt es sich Daten rund um das Thema Wein. Dabei sind in den Daten die verschiedenen Namen der Weine zu finden, deren Produzenten und woher diese sind (inkl. der einzelnen Standorte). Weiterhin sind zu dem Wein an sich weitere Informationen vorhanden. So ist der Typ und die Verwendung des Weines zu finden. Weiterhin sind in diesen Daten Eigenschaften wie Alkoholgehalt, Trinktemperatur, Süße, Säure, Körper, Gerbstoffe, Preis, Jahr und größer einer Flasche vermerkt. Genauere Informationen was diese Eigenschaften bedeuten ist im Kapitel Anwendungshintergrund zu finden. Diese Daten werden anschließend anhand ihrer verschiedenen Eigenschaften an einem Scatterplot, Parallelen Koordinaten und einer Baumhierachie dargestellt. Wie diese Diagramme aufgebaut sind, ist im Kapitel Anwendungshintergrund zu finden. Dabei ist es möglich mithilfe des Scatterplots zwei verschiedene Eigenschaften des Weines zu vergleichen, um entsprechende Muster zwischen diesen beiden Eigenschaften zu erkennen und identifizieren. Dies könnte zu neuen Erkenntnissen rund um diese Eigenschaften führen. Mithilfe der Parallelen Koordinaten können die verschiedenen paare an Eigenschaften verglichen werden, um so für einen einzelne Datenpaar herauszufinden, wie es sich gegenüber den anderen Datenpaaren verhält. Somit könnten schnell besondere Datenpaare herausgefiltert werden und entsprechende neue Informationen gewonnen werden. Mithilfe der Baumhierachie wird es möglich sein die verschiedenen Weine ihren Regionen zuzuordnen und zu herauszufinden wie viele Weine pro Region vorhanden sind. Diese Informationen kann dann wiederum mit anderen Daten abgeglichen werden, um einen weiteren Erkenntnisgewinn zu ermöglichen.

2. Daten

- https://www.kaggle.com/dev7halo/wine-information

Die Originaldaten, welche auf der Plattform Kaggle zu finden sind, wurden von einem Nutzer von einer Koreanischen Webseite (welche nicht genauer angegeben worden ist) gesammelt und bereitgestellt. Diese Originaldatei enthält 32 Spalten mit insgesamt 21605 Datensätzen. Da diese Datensätze teilweise Koreanische Symbole enthielten wurde eine zweite Datei angelegt, um die entsprechenden Zeichen herauszufiltern und den Datensatz somit zu bereinigen. Diese Datei heißt „cleasingWine.csv“ und besitzt 31 Spalten und 21600 Datensätze. Auf dieser Grundlage wurden die nachfolgend in diesem Kapitel beschrieben Datenbearbeitungsschritte vorgenommen.[]

Der Datensatz von CleasingWine beginnt mit der „wineID“ welche eine einfache Index Nummer darstellt für die einzelnen Daten. Nach dieser Spalte folg die Spalte mit dem Namen „name“, in dieser Spalte sind entsprechend der ganzen Namen der einzelnen Weine mit eingetragen. Die Produzenten dieser Weine sind in der darauffolgenden Spalte der „producer“ zu finden. Aus welchem Land dieser Wein stammt wird anschließend in der Spalte der „nation“ beantwortet. Anschließend folgen fünf Spalten mit dem Namen „local“ und der entsprechenden Nummer, welche die entsprechende Region wo der Wein herkommt, darstellen. Anschließend sind finden sich die Spalten „varieties“ und die entsprechende Nummer, welche bis zwölf geht, um die Weine ihre entsprechende Sorte zuordnen zu können. Nachdem diesen Spalten folgt die Spalte „typ“ welche den entsprechenden Typ des Weines enthält. Anschließend wird in der Spalte „use“ die Verwendung des Weines genauer festgelegt. Auf diese Spalte folgt, die der „abv“ welche den Alkoholgehalt pro Volumen darstellt. Danach wird in der Spalte der „degree“ die optimale Trinktemperatur des Weines festgehalten. Darauffolgend wird mit den Spalten „sweet“, „acidity“, „body“, „tannin“, „price“, „year“, „ml“ die jeweilige Süße, Säure, Körper, Gerbstoff, Preis, Herstellungsjahr und Größe der Flasche des Weines definiert. Dabei sind ab der Spalte der „nation“ leer Werte in der Datei enthalten. Zahlenwerte sind aber der Spalte „abv“ in diesem Datensatz zu finden.[]

Dabei eigen sich diese Daten für die Zielgruppe der Weininteressierten gut. Da diese ausreichend tiefe bieten um neue Weine kennen zu lerne und auch Zusammenhänge zwischen den Weinen Eigenschaften zu erkennen. Für die gruppen der Weineinkäufer oder Weinexperten jedoch ist dieser Datensatz ausreichend. Da hierbei vor allem durch die vielen leeren Werte keine vollständige Datenbasis vorhanden ist. Somit ist nicht garantiert das für jeden Wein der ggf. entdeckt wird die entsprechenden Daten vorhanden sind, die ggf. für eine Entscheidung benötigt werden ausreichen. Weiterhin ist die tiefe der Daten für diese Gruppen nicht ausreichend, da die Eigenschaften der Weine (Süße, Säure, Körper und Gerbstoffe) nur auf einer Skala von ein bis fünf angegeben werden. Dies führt dazu, dass diese Daten ungenau sind, wobei diese hätten, genauer sein können und somit weniger geeignet für Gruppen welche genauen Daten zu den Weinen wünschen würden.

Dabei werden reichen diese Daten jedoch aus um neue Erkenntnisse rund um das Thema des Weins und ihren unterschiedlichen Eigenschaften zu generieren. Dementsprechend reichen diese Daten auch aus, um die Fragestellungen der Zielgruppen zu beantworten, wenn auch nicht so ausführlich wie diese es sich wünschen würden. Dabei könnten die Weineinkäufer und Weinexperten diese Daten mehr als Richtwerte nehmen um sich dann entsprechend weiter Recherchieren.

Darüber hinaus wurde noch für das Baumdiagramm eine weiter Datei erstellt, um die Länder welche nur in der Datei vorhanden, warum um die Kontinente und Einteilungen dieser zu ergänzen. Um so eine bessere Hierarchische Darstellung der Weine und ihrer Länder zu erhalten.

2.1 Technische Breitstellung der Daten

Die Technische Bereitstellung der Datenerfolgt mithilfe des GitHubs Repositorie, in welchem das Visualisierungsprojekt umgesetzt worden ist. Dabei sind die Daten innerhalb des Ordners „Daten“ zu finden. In diesem Ordner sind unter dem Unterordner „Quelldaten“ sind die Ursprünglichen Daten

vorhanden welche von der Plattform „Kaggle“ heruntergeladen werden konnten. Eine genauere Beschreibung dieser Daten ist im Kapitel Daten zu finden. Im anderen Unterordner namens „Aufbereitete Daten“ sind alle Daten zu finden welche weiterverarbeitet wurden und entsprechend selektiert wurden. Wie die Weiterverarbeitung erfolgte, ist im Nachfolgendem Kapitel Datenvorverarbeitung zu lesen.

Dabei wurden zwei wesentliche Dateiformate für die Bereitstellung der Datei verwendet. Diese sind CSV- und JSON-Dateien. Die CSV-Datei, auf welche der Scatterplot und die Parallelen Koordinaten zugreifen heißt dabei „WineInformationExcelAufbereitetKlein“. Der Name der JSON Datei auf, welche das Baumdiagramm zugreift, lautet „WineInformationGeoKleinKlein“. Dabei sind alle Daten innerhalb der CSV Datei mithilfe eines Kommas getrennt. Darüber hinaus werden alle Zahlen mit einem Dezimaltrennzeichen als einen Punkt angegeben. Falls eine Null in den Feldern stehen sollte, ist dies gleichbedeutend mit einem Feld welches Leer ist. In den JSON Datei hingegen wurden nur die Beziehungen zwischen den Daten abgebildet. Somit enthalten die „data“-Felder nur eine „id“ welches nur den Namen enthält. Die Beziehung wurde mithilfe der „children“-Felder realisiert. Darüber hinaus wurden in der finalen JSON Datei alle Länder entfernt, welche nach der CSV-Datei keine Weine produzieren.

2.2 Datenvorverarbeitung

Bei der Datenverarbeitung wurden im Wesentlichen drei Schritten durchgeführt, um die Daten weiter zu bearbeiten. Diese Schritte sind das Sichten-, Bearbeiten- und anschließende Überführung der Daten. Was in diesen Schritten genauer geschehen ist, wird nachfolgend erklärt.

Bei der Sichtung der Daten war es das Ziel die Daten in ein Lesbares und leicht zu verarbeitendes Format zu bringen. Aufgrund dessen wurde die Datei „cleasingWine“ in eine Exceldatei konvertiert, da eine Exceldatei genau die Zielstellung dieses Schrittes erfüllt. Diese Exceldatei ist unter dem Unterordner „Aufbereitete Daten“ mit dem Namen „WineInformtionExcel“ zu finden. Durch diesen Schritt konnten diese Dateien nun einfacher eingelesen und bearbeitet werden. Bei der JSON Datei wurde hierbei eine entsprechende zu ergänzende Datei gefunden, welche im nächsten Schritt dann ergänzt werden konnte. Diese Datei wurde durch den GitHub Nutzer „Curran Kelleher“ zur Verfügung gestellt [https://gist.github.com/curran/1dd7ab046a4ed32380b21e81a38447aa/]. Die Datei ist dabei unter „Aufbereitete Daten“ mit dem Namen „WineInformationGeo“ zu finden.

In der Bearbeitung der Daten sollten diese so bereitgestellt werden, dass sie von Code, welcher eingesetzt wurde, einfach zu verarbeiten sind. Dabei wurden sechs wesentliche Schritte getätigt um diese Daten entsprechend bereitzustellen. Zuerst wurden die Namen aus den Spalten „sweet“, „adcidity“, „body“ und „tannin“ vor den zahlen entfernt, da diese ansonsten verhindert hätten die Zahlen in ein entsprechendes Datenformat zu bringen. Anschließend wurden die Zahlenwerte überarbeitet. So standen in der Spalte „abv“ und „degree“ eine Tilde, welche die minimale und maximalen Werte miteinander verbunden hat. Da jedoch kein Zahlenformat eine Tilde zulässt wurden diese beiden Werte getrennt und anschließend aus den beiden Werten der Durchschnitt gebildet. Falls nur ein Wert bereits in der Spalte stand, wurde dieser einfach übernommen. Darüber hinaus wurde war der Preis für die Weine in der Spalte „price“ in südkoreanischen Won angegeben. Aufgrund dieser dessen wurde dieser Preis mithilfe eines Währungskurses von 1 Euro zu 1355.382 Won umgerechnet. [] Nachdem diese Zahlen erfolgreich überarbeitet worden sind, wurden die Namen in der Splate „name“ angepasst. So ist es bei der Konvertierung vorgekommen, dass Apostrophe und Umlaute nicht richtig übersetzt worden sind. Diese Fehler wurden korrigiert. Um die Daten anschließend besser in die JSON Datei einfügen zu können wurde außerdem eine neue Spalte erstellt, in welcher die Namen der Weine mit der entsprechenden JSON Schreibweise kombiniert wurden, um so eine zusammenfügen der Weinnamen und der Länder einfacher zu gestallten. Diese

Änderungen wurden alle and er Exceldatei „WineInformationExcel“ durchgeführt. Anschließend wurden in einer neuen auf „WineInformationExcel“ basierenden Datei, die Namen der Spalten von Englischen ins Deutsche übersetzt. Zusätlich dazu wurden die Spalte in der Excel „Column1“ und die Spalte mit den Schreibweisen für die JSON Datei herausgelöst. Dieser Stand präsentiert findet sich in der Exceldatei mit dem Namen „WineInformationExcelAufbereitet“ wieder. Zum Abschluss dieser Bearbeitung der Daten wurden alle Datensätze herausgelöscht welche in den Spalten „alc“ bis „ml“ ein leeres Feld oder eine Null als Werte enthielten entfernt. Diese Änderung ist in der Datei mit dem Namen „WineInformationExcelAufbereitetKlein“ wiederzufinden. Innerhalb der JSON Datei gab es nur zwei Schritte welche zur Bearbeitung der JSON Dateien. Innerhalb desersten Schritts wurden alle Weinnamen dort eingefügt, wo diese bereits auch in der CSV eine Zuordnung zu einem Land erhalten haben. Dieses Ergebnis dieses Schritts ist unter der Datei „WineInformationGeoKlein“ zu finden. Anschließend wurden im letzten Schritt alle Länder entfernt, welche keine Weine nach der CSV-Datei hergestellt haben. Dies ist in der Datei „WineInformationGeoKleinKlein“ zu erkennen.

Innerhalb der Überführung der Daten sollten die Daten wieder für den Programmcode lesbar gemacht werden. Somit wurden alle Daten, welche in einer Exceldatei gespeichert waren, wieder in eine CSV-Datei zurücküberführt. Dabei wurde darauf geachtet, dass die Daten mithilfe eines Kommas getrennt worden sind und das Dezimaltrennzeichen ein Punkt war. Da die JSON Datei nicht umgewandelt wurde, war hierbei auch keine Überführung der Daten in eine anderes Datenformat nicht nötig.

Im Schritt „Bearbeitung der Daten“ wurden die verschiedenen Durchschnitte gebildet, um die Daten besser lesbar sowohl für die Menschen als auch die Maschine zu machen. So hätten ansonsten noch mehr Kategorien zur Auswahl gestanden, welche durch die Differenz der Zahlen keinen großen Mehrwert für die Visualisierungen gebracht hätte. Weiterhin wurden verschieden Datensätze in diesem Schritt entfernt. Dies wurde getan, um eine gemeinsame Konsistente Datenbasis für alle Visualisierungen zu schaffen, so jeder Wein in allen Visualisierungen zu finden. Darüber hinaus führte die Reduktion der Datensatze dazu, dass die Visualisierungen nicht überfüllt wirken. Zusätzlich wurden alle Datensätze entfernt, welche nicht ins Deutsche übersetzt werden konnten, da diese noch in Koreanischer Sprache geschrieben waren und somit die Weine nicht mehr eindeutig identifiziert werden konnten.

3. Visualisierung

In dem Nachfolgendem Kapiteln wird genauer auf die einzelnen Visualisierungen und ihren Zweck eingegangen.

3.1 Analyse der Anwendungsaufgaben

Wie bereits im Kapitel Einleitung erwähnt worden ist, gibt es ein großes Hauptziel, welches mit diesen Visualisierungen erreicht werden soll. Dieses ist möglichst viele Erkenntnisse zu gewinnen, rund um das Thema der Weine. Dabei kann das Thema vor allem aus den Eigenschaften des Weines beleuchtet werden, welche innerhalb der Ursprünglichen CSV Datei vorhanden waren. Also an den Eigenschaften des Alkoholgehaltes, Trinktemperatur, Süße, Säure, Körpers, Gerbstoffe, Preis, Jahres und Größe der Flasche. Diese Eigenschaften können nun mithilfe von den verschiedenen Visualisierungen gegenübergestellt werden und so sollten neue Erkenntnisse rund um das Thema wein gewonnen werden können. Als Unterstützung für dieses Hauptziel gibt es noch drei etwas konkretere Fragen welche bereits in der Einleitung genannt worden sind. Auf diese wird nachfolgend eingegangen.

Um möglicherweise verschieden Zusammenhänge zwischen verschiedenen Eigenschaften herauszufinden, wird eine Darstellung benötigt, in welcher die verschiedenen Eigenschaften des Weines gegenübergestellt werden kann. Darüber hinaus sollte ein gewisses Vorwissen über verschiedene Weine bereits vorhanden sein, um die anzeigten Daten interpretieren zu können. Mit dieser Kombination ist es nun möglich verschiedene Erkenntnisse, welche bereits bekannt sind zu bestätigen oder auch neue Erkenntnisse aus diesen Daten zu ziehen.

Bei der Frage, ob es einen Zusammenhang zwischen den Produktionsmengen und der Anzahl der von Weinen in einem Land, ist es nötig diese zwei Informationen bereitzustellen und anschließend miteinander zu vergleichen. Dafür ist die Information der einzelne Produktionsmangen der Länder nötigt. Diese Information ist beispielsweise im „State of the Vitivinicultural World in 2020“ [leer] nachzulesen. So finden sich unter dem Top 3 der am meisten Produzierenden Ländern Italien mit 49,1 mhl, Frankreich mit 46,6 mhl und Spanien mit 40,7 mhl. Diese Angaben sollten nun mit einer Darstellung vergleichen werden aus der die Weine entsprechend ihren Ländern zugeordnet werden können, um so einen Abgleich mit diesen Zahlen durchführen zu können.

In der letzten unterstützenden Frage geht um besondere Datensätze und verschiedene Trends, die sich in diesem gesamten Datensatz erkennen lassen. Um diese Frage beantworten wird eine Darstellung benötig, in welcher die verschiedenen Eigenschaften dargestellt werden können, aber gleichzeitig die verschiedenen Datensatze noch unterscheidbar sind. Weiterhin sollte die Darstellung so aufgebaut das gewisse Trends in den Daten erkennbar sind. Gemeinsam mit einem gewissen Vorwissen, was besonders für die einzelnen Eigenschaften ist, könnte diese Frage beantwortet werden.

Die eingesetzten Darstellungen des Scatterplots, Parallelen Koordinaten und Baumdiagramms können dabei helfen dem Betrachtetem das Verstehen der Daten vereinfachen. So ist es für diesen möglich sich mithilfe der Parallelen Koordinationen einen überblick über die gesamten Datensatz zu erhalten und bereits gewisse Trends zu erkennen. Falls dieser anschließend tiefere Analyse und Darstellung dieser Daten haben möchte kann, dieser zu dem Scatterplot greifen. Hierbei ist es möglich zwei Eigenschaften der Daten gegenüberzustellen und diese Gegenüberstellung weitere Analysen oder Trends zu erkennen. Weiterhin ist es möglich mithilfe des Buamdiagramms herauszufinden was die Herkunft dieser Weine ist, wenn es dort gewisse Präferenzen vom Betrachteter geben sollte.

Diese Struktur der Darstellungen ist dabei nicht zwingen nötig um die verschiedenen oben genauer beschrieben Fragen beantworten zu können. Da diese verschiedenen Fragen sich mit einer Darstellung beantworten lassen. Somit muss es keine Vernetzung dieser Darstellungen geben. Es könnte jedoch sein, dass eine solche Struktur gebraucht werden kann, um die Hauptfrage zu beantworten, da Erkenntnisse gegeben falls auch zwischen den einzelnen Darstellungen rekombiniert werden können. Dann wäre eine solche Strukturierung sinnvoll und würde bei der Erkenntnisgewinn hilfreich sein.

3.2 Anforderungen an die Visualisierungen

Durch die Analyse der verschiedenen Ziele ergeben sich verschiedene Anforderungen an die Darstellungen. Für das Hauptziel ist es wichtig das die Visualisierungen so dargestellt sind, dass die Darstellungen entsprechend einfach verständlich und es sich trotzdem Erkenntnisse zu dem Thema Weine ziehen lassen. Darüber hinaus sollten die verschiedenen Eigenschaften innerhalb der Anwendungen integriert sein.

Die Anforderungen aus der Frage nach den Zusammenhängen zwischen Eigenschaften sind das sich verschiedene Eigenschaften miteinander vergleichen lassen. Dementsprechend sollten es eine Auswahl geben, in welcher der Anwender nach seinen individuellen Wünschen Eigenschaften auswählen kann, welche Anschließend automatisch angezeigt werden. Darüber hinaus sollten sich in der Darstellung verschieden Trends und Besonderheiten erkennen lassen.

Bei der Frage von Produktionsmengen und der Anzahl von Weinen in einem Land sollten in der Darstellung erkennbar sein, wo die Weine herkommen. Darüber hinaus sollte die Anzahl der Weine mit in der Darstellung vermerkt sein. Weiterhin sollten die aktuellen Produktionsmengen der verschiedenen Länder vermerkt sein, um diese mit der Anzahl der Weine im Land vergleichen zu können. Zusätzlich wäre eine Darstellung, welche diese beiden Größen miteinander vergleicht bei der Beantwortung dieser Frage sehr hilfreich.

In der letzten Frage, welche sich um die Trends und besondere Datensätze herausfinden möchte, sollte die Darstellung es ermöglichen verschiedene Eigenschaften mit dem gesamten Datensatz darzustellen. Dabei sollte jeder Datensatz trotzdem noch nachverfolgbar sein, um diesen gegeben falls als besonderen Datensatz zu identifizieren. Darüber hinaus sollte die Darstellung des Datensatz so gut erfolgen, dass sich noch Trends aus der Darstellung ablesen lassen, und dies nicht untergehen aufgrund einer beispielsweise gedrängten Darstellungsweise.

3.3 Präsentation der Visualisierung

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

Nachfolgend werden die verschiedenen Visualisierungen vorgestellt, welche bei der Realisierung des Projektes verwenden worden sind. Dies sind der Scatterplot, Parallele Koordinaten und die Buamdiagramm.

3.3.1 Visualisierung Eins

- Wird ein Scatterplot
- Präsentation -> Abbildung, Kodierung der Daten, Interaktionsmöglichkeiten
- Erfüllung und wie gut die Anforderungen erfüllt werden
- Warum ist die Visuelle Darstellung passend für das Problem? (Diskussion der Auswahl von Darstellungen)

Die erste Visualisierung innerhalb dieses Projektes ist ein Scatterplot, in welchem immer zwei verschiedene Eigenschaften des Weines gegenübergestellt werden. Dabei nimmt immer eine Eigenschaft eine Achse des Scatterplots ein und anschließend werden die Werte wie XY Koordinaten in das entstandene Koordinatensystem eingetragen. Die dabei entstanden Punkte werden in dieser Visualisierung als Kreise dargestellt. Wenn mit der Maus über diese Kreise gefahren wird, werden farblich und der Text, welcher über diesen Kreis auftaucht, zeigt den Namen des Weins, die X und Y Eigenschaft dieses Punktes an. Darüber hinaus können die Eigenschaften angepasst werden. Dafür sind über dem Diagramm verschiedene Buttons zu finden welche die Eigenschaften beinhalten, welche es insgesamt in diesem Scatterplot dargestellt werden können. Wenn diese Buttons angeklickt werden, ändert sich je nachdem in welcher Reihe der Button angeklickt wurde die jeweilige Achse des Scatterplots. Der Scatterplot ist in Abbildung 1 zu erkennen.

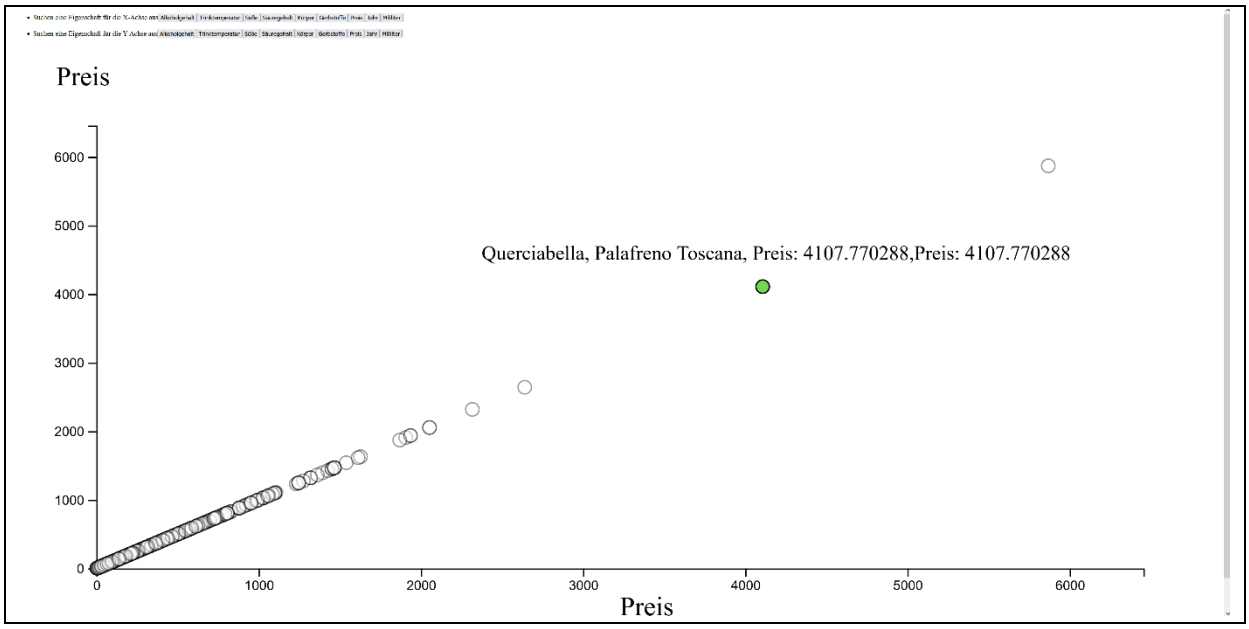

Abbildung 1: Scatterplot (Quelle: eigene Darstellung)

Die Anforderungen an den Scatterplot, welche es gibt konnten erfüllt werden. So kann der Scatterplot zwei Eigenschaften des Weines gegenüberstellen. Drüber hinaus können die Angezeigten Eigenschaften mithilfe der Buttons geändert werden und diese Veränderung passiert auch komplett ohne neu laden der Seite. Weiterhin lassen sich mithilfe der Darstellung der Koordinaten als Kreise entsprechende Trends oder Besonderheiten in den Daten gut erkennen.

3.3.2 Visualisierung Zwei

- Wird Parallele Koordinaten

3.3.3 Visualisierung Drei

- Wird eine Baumhierarchie

3.4 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

4. 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

5. 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

5.1 Anwendung Visualisierung Eins

- Anwendungsfall für Scatterplot

5.2 Anwendung Visualisierung Zwei

- Anwendungsfall für Parallele Koordinaten

5.3 Anwendung Visualisierung Drei

- Anwendungsfall für Baumhierarchie

6. Verwandte Arbeiten

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

7. 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.**

[5] L. Beilmann, *Alkoholgehalt in Wein - Das solltest du unbedingt Wissen!,* https://wein-fuer-laien.de/weinwissen/alkoholgehalt-im-wein/, **o. J.**

[6] M. Teufel, *Die richtige Trinktemperatur für Wein - warum wichtig?,* https://swisscave.com/de/swisscave-blog/post/die-richtige-trinktemperatur-fur-wein-warum-wichtig, **2021**.

[7] Brogsitter Weinversand, *Süße,* https://www.brogsitter.de/weinlexikon/suesse/#, **o. J.**

[8] Weinkenner GmbH, *Die Säure | Weinkenner.de,* https://www.weinkenner.de/die-saeure/, **2011**.

[9] Brogsitter Weinversand, *Körper,* https://www.brogsitter.de/weinlexikon/koerper/, **o. J.**

[10] Brogsitter Weinversand, *Gerbstoffe,* https://www.brogsitter.de/weinlexikon/gerbstoffe/, **o. J.**

[11] Vineyard99, *Wein & Wissen: Weinjahrgang – Geheimnis der Weinalterung,* https://www.vineyard99.de/weinjahrgang-und-weinalterung/, **2020**.

Anhang