Hoe betrouwbaar zijn de wijn beoordeling van Robert Parker?

Capstone project - Data mining & exploration

Lisanne Sijtsma

2023-04-08

Web scraping van de website van Robert Parker en een onderzoek naar de betrouwbaarheid van deze wijn recensies

Content

Business understanding

Determine business objectives

Background

Dit project komt voort uit een hobby en is een vervolg op een eerder project genaamd: *Exploration in the quality of wine, the weather and the relationship between both.* Dit voorgaande project was onderdeel van het vak *Programming for Data Science.* Over het algemeen wordt de kwaliteit van wijn onder andere bepaald door:

- De druifsoort
- Het klimaat
- De bodemsoort
- De houtsoort (van het vat)
- Het rijpingsproces in de fles

Echter is de combinatie van allen op de daadwerkelijke kwaliteit nog niet onderzocht. Daarnaast is het ook lastig om kwaliteit ergens in uit te drukken, want wie bepaalt uiteindelijk kwaliteit? Is dat de persoon die de wijn drinkt? Of iemand die verstand heeft van wijn? Het blijft een subjectief oordeel. En dat is de aanleiding voor dit onderzoek.

Om een goed beeld te geven wat er is onderzocht in het voorgaande project volgt hier een toelichting daarvan.

Voorgaande project

Het doel van het voorgaande project was om zowel de kwaliteit van wijn, het weer door de jaren heen en de relatie tussen beide te onderzoeken. Als databron voor de kwaliteit van wijn is de website van Robert Parker gebruikt. Robert Parker is een van 's werelds meest bekende wijn critici. Hij recenseert samen met een team wijnen over de hele wereld. Het voorgaande project richtte zich op een specifiek gebied: Barolo, Italië. Binnen dit gebied zijn van verschillende jaren de recensies opgehaald van alle wijnen. Vervolgens is per jaar hier een gemiddelde van genomen. Als databron voor het weer is gebruik gemaakt van NASA EARTH Data. Dagelijkse weer data van het Barolo gebied is hier uitgehaald en omgezet naar jaarlijkse gemiddelden. Het betrof 18 variabelen die het weer meten. Als laatste stap is er gekeken naar de relatie tussen beide datasets. Geen van de 18 variabelen kon direct in verband worden gebracht met een hogere of lagere gemiddelde wijn-score van Robert Parker.

Zoals omschreven is in het voorgaande project gebruik gemaakt van recensies van Robert Parker voor de kwaliteit van wijn. Maar hoe betrouwbaar zijn deze recensies eigenlijk om daadwerkelijk iets over de kwaliteit te zeggen?

Business objectives

Het doel van dit onderzoek is inzicht krijgen in de recensies die worden gegeven door Robert Parker. Daarmee wil men de volgende vraag beantwoorden:

Hoe betrouwbaar zijn de recensies van Robert Parker?

Daaruit vloeien een aantal andere vragen:

- Zijn er factoren die invloed hebben op de scores die worden gegeven aan een wijn?
- Is de data representatief genoeg voor verder onderzoek? *In verder onderzoek zullen eerdergenoemde factoren (o.a. bodemsoort) tegen deze data worden aangelegd.*

Business success criteria

Dit project is succesvol wanneer er een antwoord kan worden gegeven op bovenstaande vragen. Op basis van die uitkomsten kan er met de data van Robert Parker worden doorgegaan met een breder onderzoek. Mocht de data niet representatief bevonden worden dan zal er opnieuw onderzoek moeten worden gedaan hoe wijnkwaliteit het best te meten is.

Assess situation

Inventory of resources

Voor de data wordt binnen dit project gebruik gemaakt van de website van Robert Parker. De wijnrecensies zullen worden *gescraped*. Dit zal onderdeel zijn van het project. Er wordt gekeken naar de recensies betreft wijnen uit verschillende landen en verschillende jaren.

Requirements, assumptions & constraints

Er is onderzocht of de data van de website van Robert Parker *gescraped* mag worden. Het `robots.txt` bestand op de website laat zien dat dit is toegestaan en dat er geen rate limits zijn.

Risks and contigencies

Ondanks dat er geen *rate limits* zijn gespecificeerd in de `robots.txt`, zal het aantal requests zo veel mogelijk worden beperkt. Zowel in de testfase (bouwen van de *scraper*, als bij het daadwerkelijk ophalen van de data).

Terminology

Er volgt een beschrijving van een aantal begrippen:

Variabele	Uitleg
Rating	De score van de wijn tussen 0 en 100
Drink date	Wanneer de wijn idealiter gedronken zou moeten worden
Issue date	Datum waarop de review is gedaan
Source	Locatie van het bijbehorende artikel
Content	Begeleidende (uitleg) tekst / review tekst
Type	Type van de wijn (bijvoorbeeld Table (tafelwijn), of Sparkling (bruisend))

Variabele	Uitleg
Sweetness	Zoetheid van de wijn
Variety	Druifsoort
Vintage	Het jaar dat de druif is geplukt
Sub-region	Een specifiek gebied binnen de regio
Appellation	Een gedefinieerd gebied waar druiven moeten worden verbouwd om een specifieke naam te mogen dragen
Sub-appellation	Een specifiek gebeid binnen een appellatie

Costs and benefits

Er zijn (voor nu) geen kosten gemoeid bij dit project, behalve tijd van de onderzoeker. Als dit project succesvol is kan er antwoord worden gegeven op de *business objectives* en kan er vervolg onderzoek worden uitgevoerd.

Determine data mining goals

Data mining goals

Om inzicht te krijgen in de wijnrecensies die worden gegeven door Robert Parker zijn de volgende technische doelen opgesteld:

- In wat voor range worden de scores gegeven?
- Wat is het gedrag van de scores over de jaren heen?
- In hoeverre is de wijn / sweetness / color / producer / type / reviewer / variety / locatie / appellation /vintage / begeleidende tekst van invloed op de score van de wijn?

Uiteindelijk moet daarmee antwoord gegeven kunnen worden op de vragen die zijn gesteld in *business objectives*.

Data mining success criteria

Met een zekerheid van 95% kunnen zeggen welke meegenomen variabelen van invloed zijn op de score die wordt gegeven. Daarnaast op basis van de range van de 'scores' en het gedrag daarvan over de jaren heen inzicht kunnen geven over de representativiteit van de dataset.

Produce project plan

Project plan

Fase	Tijd	Risico's
Business understanding	4 week	Het goed kunnen formuleren van hoofd/deelvragen - data binnen kunnen halen
Data understanding	3 days	Onverklaarbare dingen zien
Data preparatie	1 week	Onvoorziene data problemen - data kwaliteit is anders dan verwacht
Modeling	3 days	Technische problemen - niet kunnen vinden van goed model
Evaluation	2 days	
Deployment	2 days	

Initial assesment of tools and techniques

De volgende tooling wordt gebruikt voor dit project:

- Visual Studio Code (editor)
- Python (code)
- Streamlit (dashboard)
- Quarto (document)

Data understanding

Collect initial data

Robert Parker wijn recensies. Op de website zijn recensies van verschillende landen beschikbaar. Met de gebouwde *scraper* is het mogelijk deze recensies per land op te halen. Initieel zullen recensies van de landen Duitsland, Italië en Portugal worden opgehaald. De recensies bevatten de wijnscore. Ook bevatten ze aanvullende informatie, dit zijn de volgende 12 velden:

- Title
- Drink date
- Reviewed by
- Issue date
- Source
- Content

- Producer
- Location
- Color
- Type
- Sweetness
- Variety

Describe data

Voor Portugal zijn 13.496 wijn recensies opgehaald. Voor Duitsland zijn er 24.889 wijn recensies opgehaald en voor Italië 50.358. Bij elkaar resulteert dit in 88.743 recensies. Voor alle recensies zullen overeenkomende velden beschikbaar zijn (zie *collect initial data*).

- **Numerieke velden**: Rating
- **Categorische velden**: Drink date, Reviewed by, Issue date, Source, Producer, Location, Color, Type, Sweetness, Variety
- **Overig**:
  - **Title** Bevat het jaar, de producer en de naam van de wijn
  - **Content** Vrije tekst

Het aantal velden zal na de data preparatie anders zijn, daar worden nieuwe velden gecreëerd op basis van de bestaande velden, zoals bijvoorbeeld het jaar. Ook zullen voor dit project onnodige velden worden verwijderd.

Explore data

Het is lastig al een exploratie te doen vóór de data preparatie. Er zijn wel een aantal interessante resultaten aan de dataset: 88743 rijen en 88477 unieke titels - dat betekent dat er 266 overeenkomende combinaties van vintage, producent en wijn zijn. 266 overeenkomende wijnen, met een rating dus.

Deze plots laten zien hoe vaak de verschillende categorische waarden van sweetness, type, color en reviewer voorkomen.

Verify data quality

Missing data. In de kolommen ‘rating’, ‘reviewed_by’, ‘source’, ‘content’, ‘color’, ‘type’ en ‘sweetness’ komen missende velden voor, maar in theorie kan dit bij alle velden voorkomen. Dit zal tijdens de data preparatie verder naar voren komen, omdat nu de verschillende velden niet zijn opgesplitst (zoals de locatie). Niet iedere wijn zal een sub-regio of een appellation hebben bijvoorbeeld.

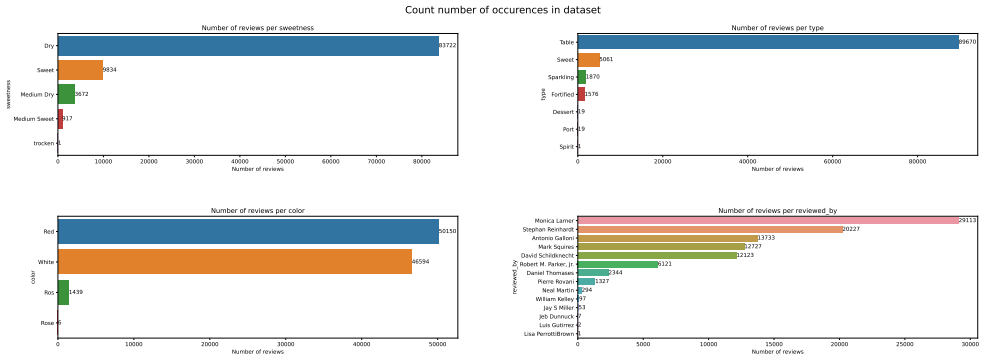


Figure 1: Categorical occurences of sweetness, type, color and reviewer

Data errors. Typfouten kunnen voorkomen bij het invullen van de score. Dit is alleen te achterhalen als er een rating buiten de range 0-100 valt, dat is nu niet het geval. Eventueel kunnen er typfouten gemaakt zijn in een ander veld. Dit zou zijn opgevallen bij het kijken naar de unieke waarden van ieder veld. Het is mogelijk dat dit bij data preparation nog verder naar voren komt, omdat daar nog velden worden opgesplitst.

Measurement errors. Bij de rating komen een aantal verschillende formats voor. Soms is het een range van scores (bijvoorbeeld 84 -86), maar ook een +, - of ? komt voor (bijvoorbeeld 84+). coding inconsistencies bad metadata

Data preparation

Select data

Selecting items. De velden 'Issue date' en 'source' zijn niet nodig voor project en zullen niet worden meegenomen.

Selecting attributes. Dit project zal intieel worden beperkt tot de wijn recensies uit de landen Portugal, Duitsland en Italië. Deze data is initieel opgehaald. Dit betekent dat dus alle data zal worden meegenomen. Het blijft mogelijk deze set uit te breiden.

Clean data

Door het *scrapen* van de data, moet er veel worden opgeschoond. Bij de meeste waarden staat de veldnaam nog in de waarde (bijvoorbeeld bij het veld 'Rating' is de waarde: Rating: 87). Dit is voor alle velden opgeschoond door de veldnamen er uit te halen. Daarnaast staan er veel enters, spaties en overige tekens rondom de waarde. Deze zijn ook verwijderd door gebruik te

maken van regex. Dit onderdeel van het opschonen van de data is direct uitgevoerd bij het *scrapen* van de data. Deze data is vervolgens in csv-formaat opgeslagen in de map *raw_data*

In de velden 'Rating', 'Reviewed by', 'Source', 'Content', 'Color', 'Type' en 'Sweetness' komen missende waarden voor. Het is nietszeggend om een missende waarde op te vullen met data van een andere wijnrecensie. Daarom zullen wijnrecensies met missende waarden in de velden 'Rating', 'Reviewed by', 'Source', 'Color', 'Type' en 'Sweetness' worden verwijderd. Een wijnrecensie met missende waarden in het veld content wordt niet verwijderd, omdat deze niet direct de wijn omschrijft zoals de andere velden dat doen. Daarnaast komen er ook 'uitschieters' voor. Binnen dit project is er voor gekozen uitschieters niet te verwijderen maar mee te nemen in de verdere analyse. Er is hiervoor gekozen omdat de ratings aan de bovenkant van de score range zitten, en juist de lager scorende reviews interesant voor verder onderzoek kunnen zijn.

Binnen het veld 'Rating' komen verschillende formats voor:

- 86
- (86-88)
- 88+
- 88-
- ?

De wijnrecensies met een vraagteken voor 'Rating' zijn verwijderd. Daarnaast is de "+" en "-" verwijderd van de ratings, omdat er niet is gedocumenteerd is hoeveel meer of minder dit is op de score. Van een score range is een gemiddelde genomen van de range (Bij een range van (86-88), wordt dit dus 87).

Controle of alle ratings binnen de 0 en 100 zijn.

Voor het veld 'color' kwamen verschillende waarden voor. 'White', 'Ros', 'Red', 'Rose', nan. (Plaatje met hoe vaak iedere waarde voorkomt). Ros is een fout door het binnenhalken van dedata. Daar zijn alle niet a-z characters verwijderd. Rosé is dus ros geworden. Dit aangepast dat Ros Rose is geworden.

Construct data

Er worden verschillende nieuwe velden (kolommen) aangemaakt op basis van de opgehaalde wijnrecensies:

- Vintage - Uit de titel
- Lengte Vintage - Aantal karakters van 'Vintage'
- Name (naam van de wijn) - Uit de titel
- Lengte Content - Aantal karakters van 'Content'
- Country - Uit de locatie
- Region - Uit de locatie

- Sub-region - Uit de locatie
- Appellation - Uit de locatie
- Sub-appellation - Uit de locatie

Vervolgens zijn er ook *cleaning-steps* over deze data gegaan.

1. Wijnrecensies zonder vintage worden verwijderd.
2. Het datatype van vintage wordt omgezet naar integer.
3. Controle of de name altijd gevuld is.
4. Missing values in de length_content opvullen met 0, want geen content is een lengte van 0.

Na de controles worden de velden die niet nodig zijn voor verder onderzoek verwijderd: * 'title' * 'rating' * 'issue_date' * 'source' * 'from_location'

Na deze stappen blijken naast 'Content', ook 'Region', 'Sub-region', 'Appellation' en 'Sub-appellation' missende waarden te hebben. Voor 'Appellation' en 'Sub-appellation' is het logisch dat er veel missende waarden zijn, omdat er maar een beperkt aantal appellations zijn. Voor het veld 'Region' is het nietszeggend deze op te vullen met data uit een andere recensie. Voor nu laten we deze staan als een missende waarde.

Checken op uitschieters. Controle op hoeveelheid unieke velden.

Integrate data

Er zijn drie losse datasets. Deze worden *append*. Ze bevatten dezelfde attributen (velden), maar verschillende records (rijen)

Format data

-

Modelling

Select modeling technique

Modeling technique

Binnen het onderzoek zijn er verschillende data mining goals. Binnen dit hoofdstuk zal verder in worden gegaan op de statistische technieken om deze vragen te beantwoorden.

- Beschrijvende statistiek. Door middel van beschrijvende statistiek kan een antwoord worden gegeven op volgende vragen van de data mining goals:
  - In wat voor range worden de scores gegeven
  - Wat is het gedrag van de scores over de jaren heen

Daarnaast wordt de beschrijvende statistiek ook gebruikt om naar de andere variabelen te kijken. Zo kan een totaal beeld worden verkregen van de dataset.

- Verklarende statistiek. Door middel van verklarende statistiek kan antwoord gegeven worden op de volgende algemene vraag:
- In hoeverre is een variabele van invloed op een andere variabele?

Binnen dit onderzoek wordt gekeken naar de invloed van verschillende variabelen op de score van wijn. Er worden statistische toetsen uitgevoerd om te onderzoeken of de relatie tussen iedere variabele en de score van wijn toeval is, of significant anders. Om een statistische toets uit te kunnen voeren wordt eerst onderzocht of de score van wijn normaal verdeeld is. Alle onafhankelijke variabelen (de veschillende variabelen) hebben meer dan 2 groepen. Daarnaast bestaat de data uit onafhankelijke groepen, wat betekent dat dit ongepaarde samples zijn. Alles bij elkaar betekent dit dat de toets zal uitkomen op een 'One-way ANOVA' of 'Kruskal Wallis toets', afhankelijk van of de score normaal verdeeld is.

Modeling assumptions

Er word aangenomen dat de samples onafhankelijk van elkaar zijn.

De afhankelijke variabele is 'Rating'. Dit is een ratio variabele. De onafhankelijke variabelen zijn de verschillende variabelen die mogelijk invloed hebben op de afhankelijke variabelen. Hier vallen onder: 'Wijn', 'Sweetness', 'Color', 'Producer', 'Type', 'Reviewer', 'Variety', 'Locatie', 'Appellation', 'Vintage', en 'Content'. Dit zijn allemaal nomnimale / categoriale variabelen.

Generate test design

Test design

Voor de toetsen zal een alpha van 0.05 worden aaneghouden (betrouwbaarheid van 95%).

Build model

Om te toetsen of de 'rating' normaal verdeeld is, is gebruik gemaakt van de Lilliefors toets. Deze toets is gebaseerd op de Kolmogorov-Smirnov toets met het verschil dat er geen geschetste normaal verdeling (op basis van de data) vooraf hoeft te worden meegegeven. Het resultaat komt in een testscore en een p waarde. Deze p waarde bepaald of de 0-hypothese kan worden behouden of verworpen.

Uit deze toets is gekomen dat de 'rating' niet normaal verdeeld is. Daarom wordt er voor de volgende toetsen een Kruskal Wallis toets gebruikt.

The Kruskal Wallis test is a non-parametric method for testing. No normal distribution is assumed.

H_0 = Median of all groups are equal

H_1 = At least one population median of one group is different

- N = Total number of observations across all groups
- g = Number of groups
- n_i = Number of observations in group i
- r_{ij} = The rank (among all observations) of observation j from group i
- $\overline{r}_i = \dfrac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$is the average rank of all observations in group i
- $\overline{r} = \dfrac{1}{2}(N+1)$is the average of all the r_{ij}

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{g} n_i \overline{r}_i^2 - 3(N+1)$$

If significant this means there is at least one group different than the others. It does not identify which group it is.

This test will be executed for each earlier mentioned variables on all available data.

Assess model

- In wat voor range worden de scores gegeven

Er zijn 96990 ratings. De gemiddelde rating is 90.03. Er is een standaarddeviatie van 3.03. De laagste rating is 15 en de hoogste 100. (terug te vinden in de logger).

Meting	Waarde
count	96990.000000
mean	90.027585

Meting	Waarde
std	3.033268
min	15.000000
25%	88.000000
50%	90.000000
75%	92.000000
max	100.000000

- Wat is het gedrag van de ratings over de jaren heen

Figuur 4 laat de mediaan van de ratings over de jaren zien. Er is geen duidelijke stijgende of dalende trend te zien.

Voor de variabelen (…) zijn boxplots gemaakt om meer inzicht te geven in de verdelingen daarvan. Deze zijn te vinden in bijlage x.

Zoals te zien in fig x histogram of ratings lijkt de rating normaal verdeeld te zijn. Echter kan dit pas met zekerheid worden vastgesteld na een toets.

alpha = 0.05 H_0 = De ratings volgen een normal distribution

H_1 = De ratings volgen geen normal distribution

P-value < 0.05 wat betekent dat de nulhypothese wordt verworpen. Uit de Lilliefors toets komt dat de 'rating' niet normaal verdeeld is.

Uit de kruskall wallis toetsen zijn de volgende resultaten gekomen: Voor iedere toets is de volgende algemene hypotheses opgesteld:

H_0 = The x variabelen are equal in terms of rating

H_1 = At least one variabele is different from the other x variabelen in terms of rating

Voor alle gevallen geldt: test_statistic >= critical_value. Wat betekent dat de nulhypothese wordt verworpen en er dus significante verschillen zitten tussen op zijn minst één groep en een andere groep binnen iedere variabele.

Evaluation

Van data mining goals naar business objectives.

Kijken naar de range van de scores is dit tussen 15 en 100. Echter, als men daadwerkelijk kijkt naar de verdeling zit het gemiddelde op 90. 'Lagere scores' komen zeer weinig voor.

Nav modelling (kruskal wallis toets) kan nu worden gezegd dat alle variabelen op zijn minst één groep hebben die significant anders is dan een andere group in termen van rating.

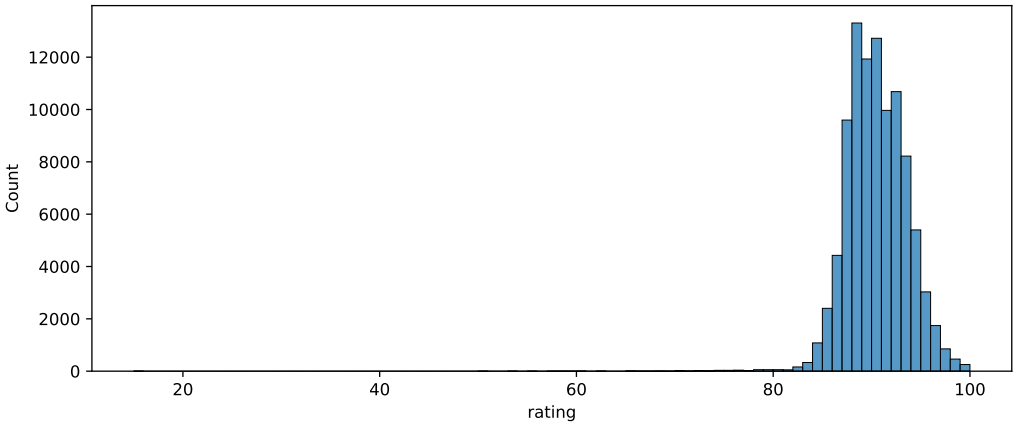


Figure 2: Histogram of ratings

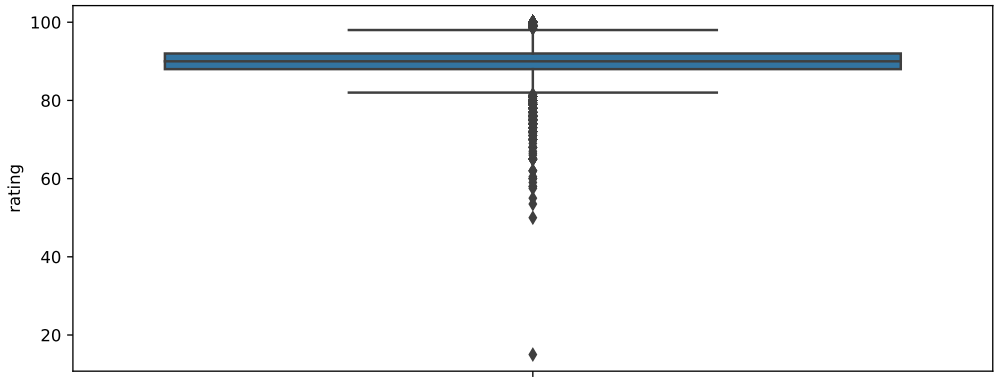


Figure 3: Boxplot of ratings

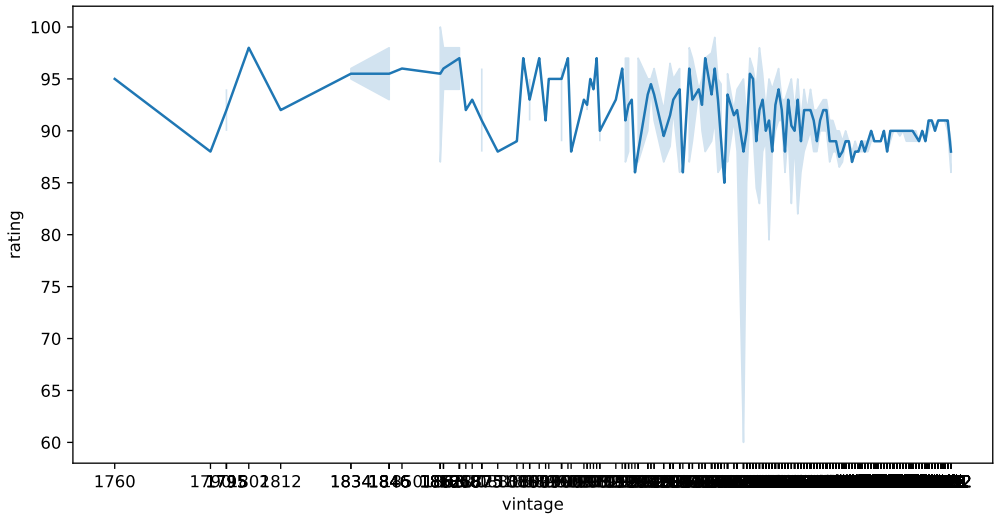


Figure 4: Median ratings over the years

bij alle toetsen is de test_statstic >= chi_critical value. Wat betekent dat de nulhypothese wordt verworpen er er dus verschil is. Er is ook gekeken naar de range van de scores en …

Evaluate results

Data mining results assessment

Aan de hand van dit onderzoek zijn er diverse inzichten en nieuwe vragen naar voren gekomen. Uit het hoofstuk modeling bleek dat de ratings zich voornamelijk in een hoge range bevinden (88-92). Dit is een belangrijke bevinding die niet direct uit de website van Robert Parker is te halen. Deze bevinding heeft invloed op de presentativiteit van de dataset. In het kader van verder onderzoek, zal het lastig zijn de impact van bijvoorbeeld bodemsoort te bepalen aan de hand van een rating van Robert Parker. Er is weinig spreiding in de ratings, en de onderlinge verschillen zijn klein omdat het in een kleine range zit.

Aan de hand van de uitgevoerde toetsen kan worden gezegd dat binnen alle onderzochte variabelen groepen onderling verschillen in het kader van rating. Dit is een eerste stap, om echt iets te kunnen zeggen over de beinvloeding van specifieke facoren op de rating zal moeten ondenrzocht welke groepen dit zijn en of een combinatie van deze factoren leidt tot een hogere rating.

Alles samen betekent dit dat er niet direct een antwoord kan worden gegeven op de hoofvraag: ‘Hoe betrouwbaar zijn de recensies van Robert Parker’. Er zal eerst vervolg onderzoek moeten

Test statistic	P-value
0.08359504079899555	0.0009999999999998899

Figure 5: Test results of the Lilliefors normal distribution test

Variable	Test Statistic	Chi critical value	P-Value	Alpha
color	720.1	5.99	41688313994874e	0.05
type	1784.29	12.59	0	0.05
reviewer	7874.42	21.03	0	0.05
sweetness	4184.36	7.81	0	0.05
producer	30725.65	4980.59	0	0.05
variety	7603.73	526.81	0	0.05
vintage	4231.59	149.88	0	0.05
appellation	5914185.24	125.46	0	0.05
country	2132.26	7.81	0	0.05

Figure 6: Test results of all variables related to the rating

worden gedaan naar eerder genoemde vragen.

Approved models

Dit project heeft opgeleverd een git repository (hier te vinden) met dit erin: - scraper om data van website van robert parker te halen - de daadwerkelijke data van robert parkers' website. - dashboard om deze data inzichtelijk te maken

Review process

Binnen het project was het voornamelijk een uitdaging om de business understanding helder te krijgen en het proces daaraan voorafgaand. In het kader van het capstone project diverse keren gewisseld van onderwerp. Dit kostte veel tijd, waardoor er veel werk te doen was in een korte tijd.

Binnen het proces zijn er de volgende verbeterpunten: - Deadline stellen op het verkrijgen van de business objectives. Bij dit project is dit iets lastiger omdat er geen externe opdrachtgever is. Het stellen van een deadline op dit eerste stuk zal helpen om genoeg tijd over te houden voor de andere onderwerpen. - Te vroeg verdiept in specfieke toetsen, want aanname normaal verdeling. Bij daadwerkelijke toetsing bleek er geen sprake te zijn van een normaal verdeling. - Na aanleiding van het vorige project was er al een *scraper* gebouwd om recensies van de website van Robert Parker te halen. Echter, was deze *scraper* totaal anders opgebouwd, waardoor dit meer tijd kostte dan verwacht. Daarnaast was deze ook niet efficient genoeg om grotere hoeveelheden data op te halen. - Er is op dit moment data binnengehaald van vier landen. Er kan worden verbeterd door data van meer verschillende landen op te halen. - Kwaliteit blijft een lastig punt om te meten. Een verbeterpunt zou kunnen zijn om te kijken naar alternatieve data om ernaast te leggen, bijvoorbeeld data van Vivino (link). Echter paste dit qua tijd niet binnen het project.

Discussie

Systematisch deelvragen(=data mining goals) langslopen.

- In wat voor range worden de scores gegeven? 75% van de scores wordt tussen de 88 en 92. Minimale score is 15 en de maximale score 100. Over het algemeen worden de scores dus in een kleine range gegeven, die aan de bovenkant ligt van de mogelijke scores.

De 'scheve' verdeling van de scores van de wijn rijst wat vraagtekens. Zijn de wijnen die Robert Parker beoordeeld daadwerkelijk allemaal goed? Of speelt het toch mee dat een specifieke kleur, type of druifsoort over het algemeen lekkerder wordt bevonden?

- Wat is het gedrag van de scores over de jaren heen? De scores gedragen zich over de jaren heen redelijk stabiel. Er is geen duidelijke trend te zien in de mediaan van de scores over de jaren heen.

Zijn er factoren die invloed hebben op de scores die worden gegeven aan de wijn? * In hoeverre is de wijn / sweetness / color / producer / type / reviewer / variety / locatie/ appellation /vintage invloed op de score van de wijn? Met iedere variabele is een test gedaan om te bepalen of het verschil tussen de variabele invloed heeft op de rating van de wijn. Uit al deze toetsen is gekomen dat er significant verschil zit tussen tenminste één groep ten opzichte van een andere groep in termen van rating. Dit zegt nog niets over welke groepen dit dan zijn. Daar zal verder onderzoek naar gedaan moeten worden.

Conclusie

Hoofvraag: Hoe betrouwbaar zijn de recensies van Robert Parker? Is de data representatief genoeg voor verder onderzoek? In verder onderzoek zulleneerdergenoemde factoren (o.a. bodemsoort) tegen deze data worden aangelegd.3 Er kan op dit moment nog geen antwoord worden gegeven op de hoofdvraag. Daarvoor zal eerst verder onderzoek moeten worden gedaan. Wel is er iets te zeggen over de representativiteit voor verder onderzoek. Omdat op basis van de opdit moment binnengehaalde data de spreiding tussen de ratings zo klein is, zal het lastig zijn deze daarvoor te gebruiken.

Determine next steps

Meer data toevoegen aan het onderzoek om met meer onderbouwing conclusies te kunnen trekken. Betekent een hoge score een goede wijn? Of blijft dit een kwestie van smaak bij de specifieke reviewer? Toevoegen van “mensen” data - vivino? Review van het proces wanneer worden wijnen door Robert Parker beoordeeld? Moet het daarvoor eerst aanbepaalde eisen voldoen? Komen er daardoor alleen ‘goede’ wijnen binnen om te beoordelen? Moet er geld worden betaald? etc.

Deployment

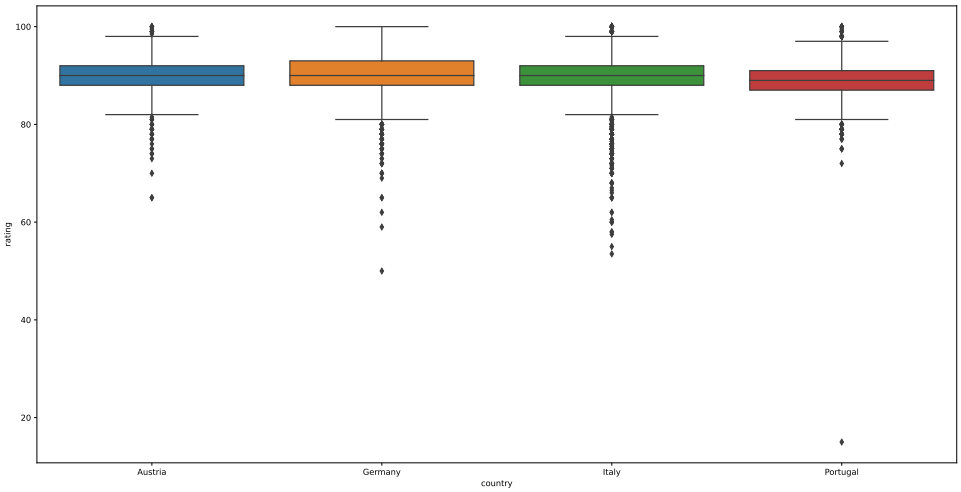
Voor nu is er gebruik gemaakt van een persoonlijke subscription voor Robert Parker. Voor in een daadwerkelijke deployment zou dit een npa (non personal account) moeten zijn.

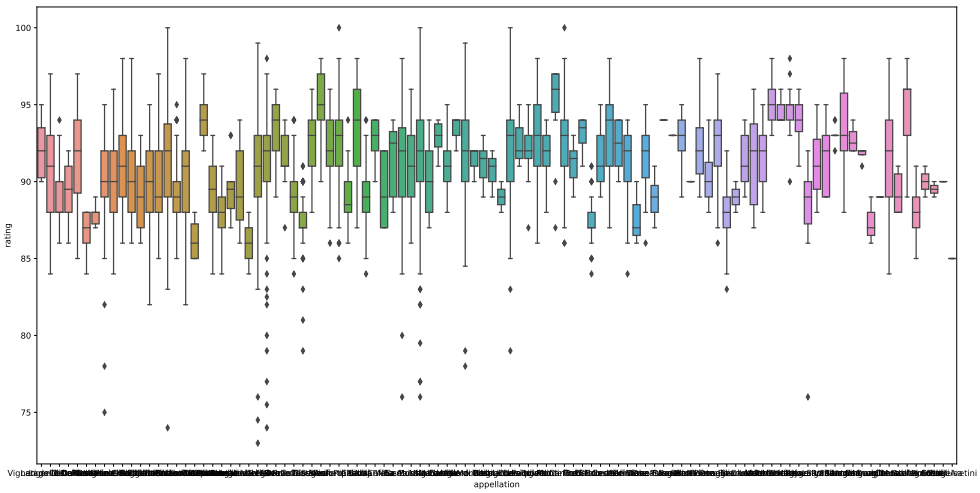
Verder zou het een goede stap zijn om het binnehalen van data te automatiseren. Er zal een intiele bulk moeten worden gedaan om de wijn recensies die er nu zijn op te halen. In het vervolg kan bijvoorbeeld per afgelopen x tijd de ‘nieuwe’ beoordelingen worden opgehaald.

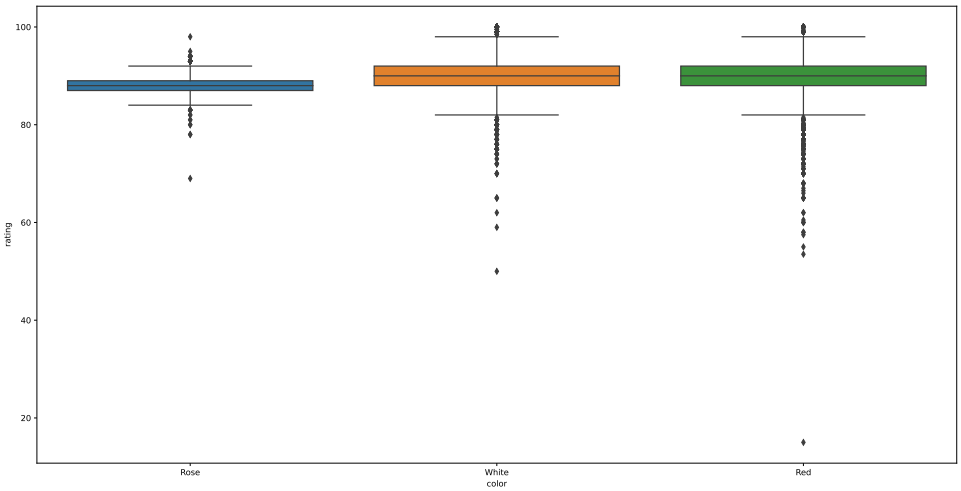
Omdat er niet met zekerheid kan worden gezegd of de data hetzelfde blijft zou het eengoede toevoeging zijn om testen toe ete voegen. Deze testen kunnen matchen op de namen van variabelen die worden opgehaald. Ook kan er iedere x tijd een dummy call worden gedaan om te chekcen of de data die daaruit komt overeenkomend is.

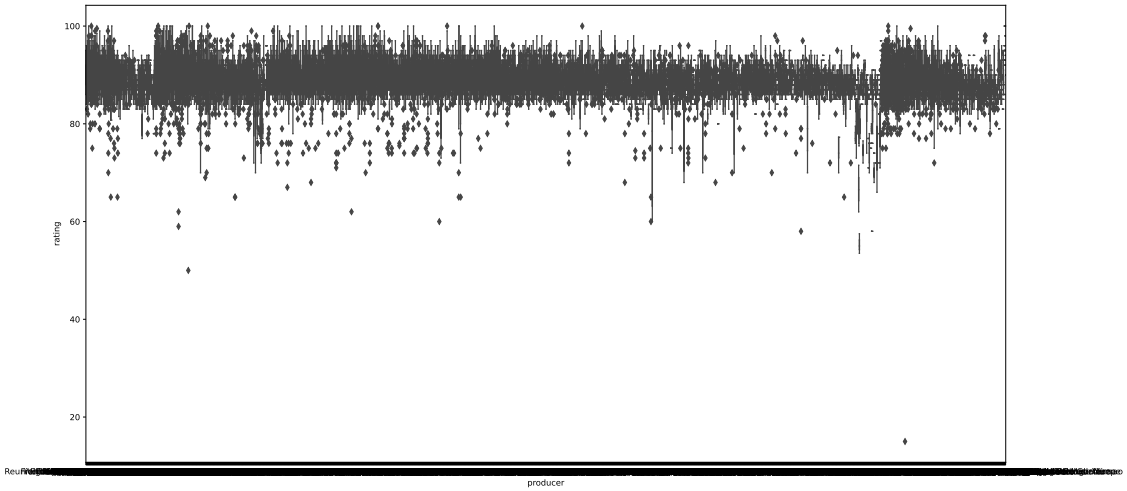
Er is een git repository aangemaakt voor dit project. Changes aan het project zullen daar worden deployed.

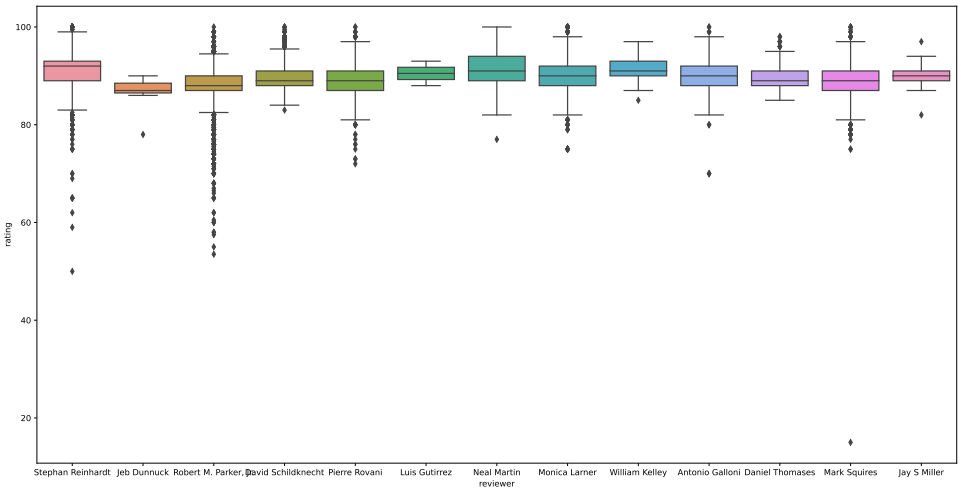
Bijlagen

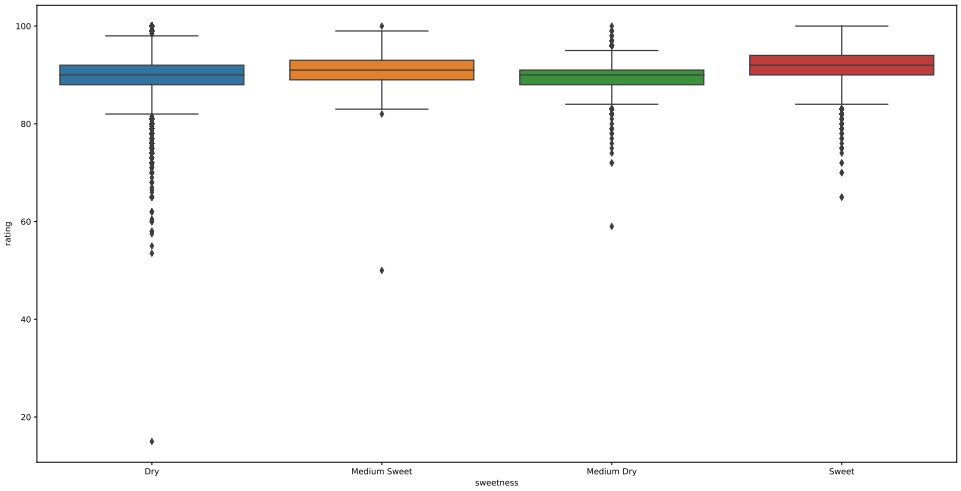

100
80
60
40
20
rating
Austria
Germany
Italy
Portugal
country

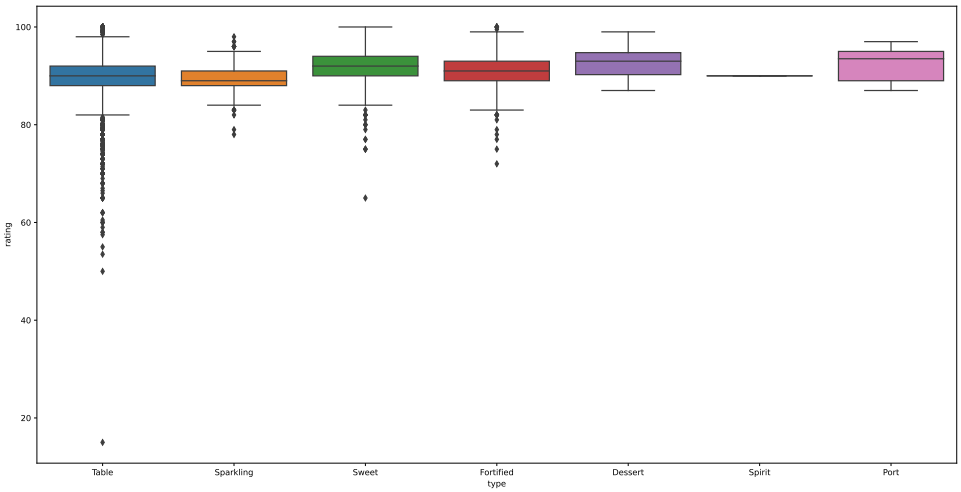

100
95
90
85
80
75
rating
appellation

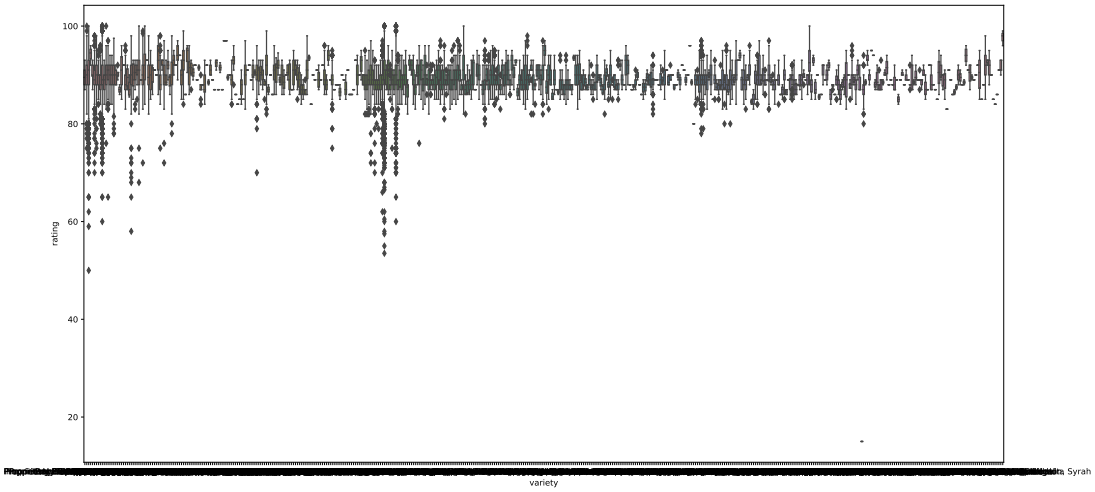

100
80
60
40
20
rating
Rose
White
Red
color

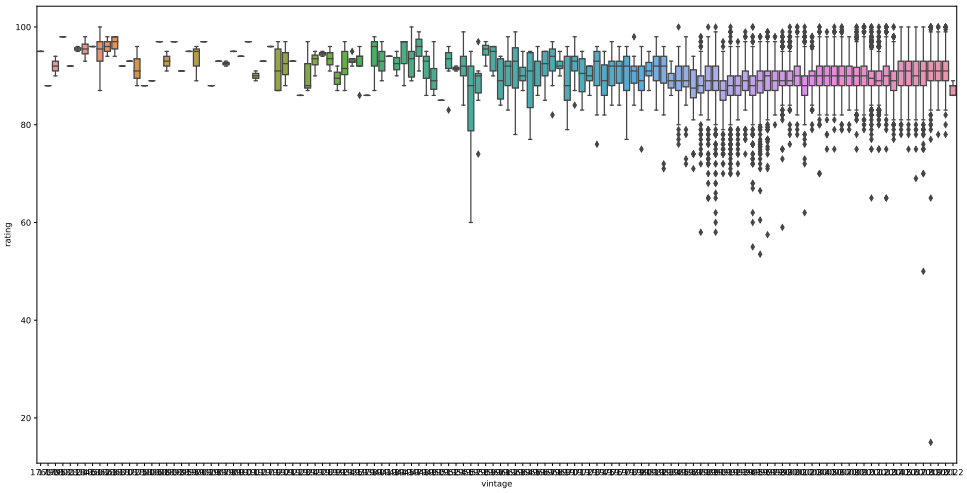

100
80
60
40
20
rating
producer


rating
100
80
60
40
20
Stephan Reinhardt
Jeb Dunnuck
Robert M. Parker, Jr.
David Schildknecht
Pierre Rovani
Luis Gutirrez
Neal Martin
Monica Larner
William Kelley
Antonio Galloni
Daniel Thomases
Mark Squires
Jay S Miller
reviewer


rating
100
80
60
40
20
Dry
Medium Sweet
Medium Dry
Sweet
sweetness


100
80
60
40
20
rating
Table
Sparkling
Sweet
Fortified
Dessert
Spirit
Port
type


100
80
60
40
20
rating
variety


100
80
60
40
20
rating
vintage