How reliable are Robert Parker's wine reviews?

Capstone project - Data mining & exploration

Lisanne Sijtsma

2023-04-17

Web scraping of Robert Parker's website and research into the reliability of his wine reviews

Content

Business understanding

Determine business objectives

Background

This project is based upon a hobby and a continuation on an earlier project named: *Exploration in the quality of wine, the weather and the relationship between both*. The previous project was part of the course *Programming for Data Science*. In general wine quality is measured by the following attributes:

- Grape
- Climate
- Soil type of vinyard
- Woodtype of the barrel
- The ageing process

There has been no research into the impact of the combined aforementioned attributes on the quality of wine. It is also challenging to quantify the quality of wine, cause it will always remain a matter of taste and is hard to be measured subjectively. That is why this research is started.

To give a clear picture of what has been researched in the previous project. The following will be a summary of the previous project.

Previous project

The goal of the previous project was to research the quality of wine, the weather throughout the years and the relation between both. As a datasource for the quality of wine the reviews from the website of Robert Parker are used. Robert Parker is one of the most well-known wine critic. Together with his team he reviews wines from all over the world. The previous project focused mainly on a single region: Barolo, Italy. From this region multiple reviews of different years have been collected. The score of the reviews has been averaged per year. For the weather a datasource from NASA EARTH Data was used. Daily weather data of the Barolo region was used and converted to yearly averages. The data contained 18 variables measuring the different aspects of the weather. Lastly the relation between the weather and quality of wine was examined. There was no direct relation between the 18 variables and the wine rating of Robert Parker.

As mentioned, the previous project used wine ratings from Robert Parker as a measurement for wine quality. But how reliable are these wine ratings to actually measure quality?

Business objectives

The goal of this research is to get insight in the reviews given by Robert Parker. This research will try to answer the following question:

How reliable are the reviews of Robert Parker?

Forthflowing from this comes the following questions:

- Are there specific factors that influence the scores given to a wine?
- Is the data representative enough for further research? *Factors mentioned earlier will be used in further research.*

This projects research is successful if the aforementioned questions can be answered. Based on the outcomes of this research the data from Robert Parker can be used for further study. If the data is not found representative, further study into quantifying the quality of wine has to be done.

Assess situation

Resources and costs

The wine reviews that are used in this research will be scraped from Robert Parker's website. The scraping process is a part of this project. The wine reviews will be composed of wines from multiple countries and vintages. The only costs made for this research is the time (and thus costs) of the researcher.

Requirements, assumptions & constraints

Research has been done to check if scraping is allowed on Robert Parkers'website. The `robots.txt` file shows that scraping is allowed and there are no rate limits Even though there are no rate limits specified, the number of requests will be limited as much as possible.

Terminology

Variable	Explanation
Rating	Wine score ranging from 0 to 100
Drink date	Ideal drink date of the wine
Issue date	Date of publication
Source	Source of accompanying publication
Content	Review text
Type	Wine type (for example Table or Sparkling)
Sweetness	Taste of the wine (for example Dry or Sweet)
Variety	Grape variety
Vintage	The year the grape is harvested
Region	Region of the vineyard
Sub-region	A specific area within the region
Appellation	A defined area where grapes must be harvested to carry a specific name or label
Sub-appellation	A specfic are within an appellation

Determine data mining goals

The following data mining goals are defined:

- What is the range of the winescores?
- What are the patterns of the winescores over time?
- In what extent do the variables (sweetness, color, producer, type, reviewer, variety, location, appellation, vintage) measured by Robert Parker influece the score of the wine?

The success criteria include:

- Prove which variables influence the rating given to a wine, with a confidence interval of 95 percent.
- Insights on the range of the wine scores and the patterns of rating throughout the years to determine representativity.

Project plan

Phase	Time	Risks
Business understanding	4 week	To correctly formulate the business objectives.
Data understanding	3 days	To find inexplicable patterns in the data
Data preparatie	1 week	Unforeseen data problems & low data quality
Modeling	3 days	Technical problems & no model fit
Evaluation	2 days	
Deployment	2 days	

Tools used:

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

Data understanding

Collect initial data

Robert Parker wijn recensies. The website contains reviews from different countries. With the scraper reviews per country can be obtained. Initially collection will only be done on the following countries: Germany, Italy, Portugal and Austria. Reviews will contain the wine rating and additional information such as:

- Title
- Drink date
- Reviewed by
- Issue date
- Source
- Content
- Producer
- Location
- Color
- Type
- Sweetness
- Variety

Describe data

For Portugal 13.496 wine reviews have been scraped. For Germany 24.889, for Italy 50.358 and for Austria 9488 wine reviews. This results 98.231 wine reviews. All reviews contain the same fields (see *collect initial data*):

- **Numeric fields**: Rating
- **Categorical fields**: Drink date, Reviewed by, Issue date, Source, Producer, Location, Color, Type, Sweetness, Variety
- **Other**:
  - **Title** Contains the year, the producer and the name of the wine.
  - **Content** Free text

The number of fields will differ after data preparation. New fields will be extracted based upon existing fields, for example year.

Explore data

It is difficult to do the data exploration before preparing the data. The plotted data shows how many times each categorical values of sweetness, type, color and reviewer are seen in the data.

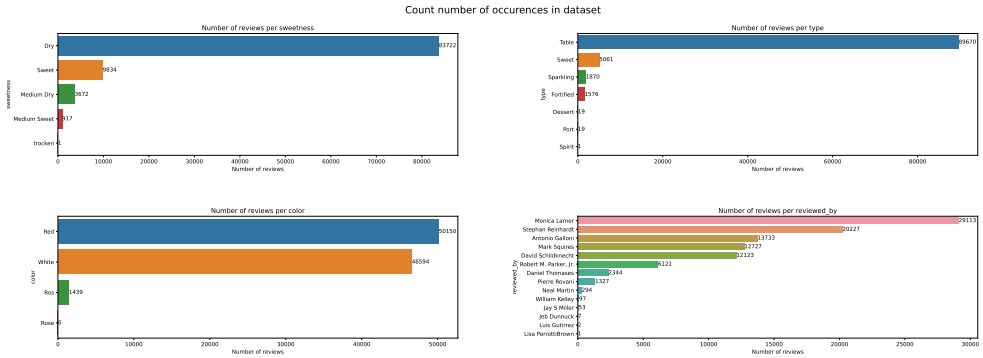


Figure 1: Categorical occurences of sweetness, type, color and reviewer

Verify data quality

Missing data. The columns rating, reviewed_by, source, content, color, type and sweetness contain missing fields, But in theory this can happen in all the fields. In data preparation this

will be elaborated since the fields are not split up. Not every wine will contain a sub-region or appellation for example.

Data Errors. Since human errors can occur during data entry for scores, It can only be assumed that all the scores are entered correctly unless numbers do not fall withing the 0-100 range.

There can also be human errors in the entry for other fields but these can only be found if there is an outlier in the value that is only present in for example one review. Since data preparation is not fully done by this time the aforementioned errors can still show up when the fields are split up.

Measurement errors. The ratings do not have a an overall consistent format, Sometimes it contains a range for example (84 – 86) but it can also have a non-concise rating containing plusses or minuses such as 84+

Data preparation

Select data

The variables 'Issue date' en 'source' are not necessary for further analysis and will be removed from the dataset. Initially the data will be limited with reviews from the countries: Italy, Portugal, Germany and Austria. This data is already scraped.

Clean data

Because scraping is used to collect the data several steps must be taken to clean the data.

- Take fieldnames out of the value for example the field "rating" should only contain numbers and not "rating: 87".
- Remove any special characters, control character and non-alphanumeric characters with a regular expression (Regex)

These steps are done within the script containing the scraper, The data is then stored as a CSV in the folder 'data/raw'.

Blanks: The fields 'Rating', 'Reviewed_by', 'Source', 'Color', 'Type' and 'Sweetness' contains blanks.

The fields will be removed from the set since filling them with blank values or existing values from other reviews will compromise the results.

The field 'Content' will be left as is since this does not contribute directly to the description or quality of the wine.

Outliers: Outliers in the fields however will not be removed but these will be kept for analysis since most ratings are at the top range 80-90 and especially the rating on the lower ranges could be interesting for this research.

Formatting issues: Formatting within de "Rating" field is also an issue since this contains a number of formats such as:"

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

Ratings that contain a "?" have been removed and "+" and "–" are stripped from the value since there is no explanation what a plus and minus mean numerically. Ratings that have a range have been averaged so (86-88) will be transformed to 87.

Data errors: Due to removal of special characters several values within the field "color" have accidentally been renamed for example Rosé turned into Ros. This has been corrected and now all the values are Rose.

Construct data

Various new variables are extracted:

- From the title: **Vintage** and the **Name of the wine**
- From the content: **Length of content**
- From the location: **Country**, **Region**, **Sub-region**, **Appellation** and **Sub-appellation**.

The next cleaning steps are executed for these new variables:

1. Removed winereviews without vintage
2. Changed datatype of vintage to integer
3. Check whether the winename is always filled
4. Fill missing values of the 'content length' with 0, no content means a 'content length' of 0.

Removal of the no longer necessary variables:

- title
- rating
- issue_date
- source
- from_location

After these steps it turned out that besides the field 'content', also Region', 'Sub-region', 'Appellation' and 'Sub-appellation' have missing values. This makes sense for 'Appellation' and 'Sub-appellation' since there is a limited number of appellations. It is meaningless to fill missing values for the field 'Region' with data of other reviews. The missing values in 'Region' will be kept.

Integrate data

There are four seperate datasets for the different countries. These contain the same attributes, but different records. The datasets will be combined into one dataset.

Modelling

Modeling technique

There are various data mining goals within the research. In this chapter, the statistical techniques to answer these questions will be discussed in more detail.

- Descriptive statistics. By means of descriptive statistics, an answer can be given to the following questions of the data mining goals:
  - What is the range of the winescores?
  - What are the patterns of the winescores over time?

In addition, the descriptive statistics are also used to look at the other variables. In this way, a total picture of the dataset can be obtained.

- Explanatory statistics. By means of explanatory statistics, the following question can be answered:
  - In what extent do the variables (sweetness, color, producer, type, reviewer, variety, location, appellation, vintage) measured by Robert Parker influece the score of the wine?

The influence of various variables on the winescore is examined. Statistical tests are performed to examine whether the relation between each variable and the score of wine is coincidental or significantly different. To be able to perform a statistical test, it is first examined whether the score of wine is normally distributed. All independent variables (the different variables) have more than 2 groups. In addition, the data consists of independent groups, which means that these are unpaired samples. This means that the test will end up as a 'One-way ANOVA' or 'Kruskal Wallis test', depending on whether the score is normally distributed.

There is one modeling assumptions:

- It is assumed that the samples are independent of each other.

The dependent variable is 'Rating'. This is a ratio variable. The independent variables are the different variables that may affect the dependent variables. These are all nominal / categorical variables.

Generate test design

For all tests an alpha of 0.05 will be used (95 percent confidence interval).

The Lilliefors test was used to test whether the 'rating' is normally distributed. This test is based on the Kolmogorov-Smirnov test, with the difference that no sketched normal distribution (based on the data) needs to be provided in advance. The result comes in a test score and a p-value. This p-value determines whether the 0-hypothesis can be maintained or rejected.

The Lilliefors test showed that the 'rating' is not normally distributed. Therefore, a Kruskal Wallis test is used for the following tests. 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 significantly different than the others. It does not identify which group it is. This test will be executed for each earlier mentioned variable on all available data.

Assess model

- What is the range of the winescores?

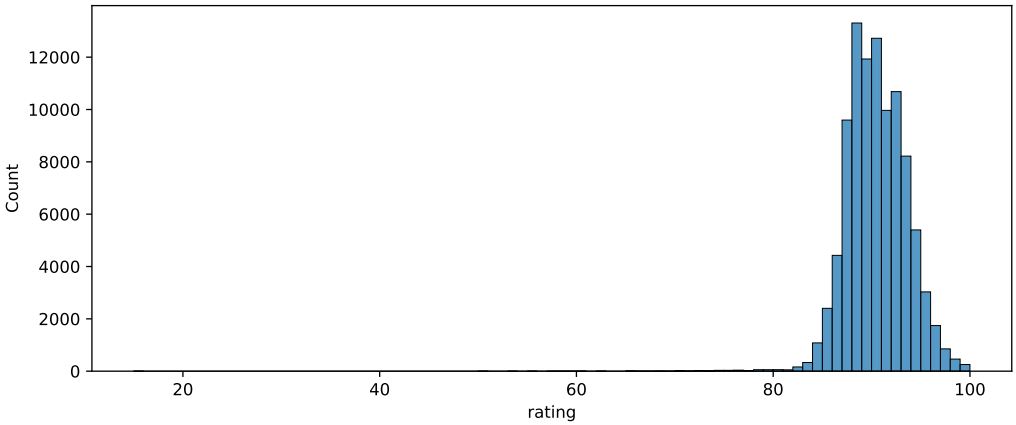


Figure 2: Histogram of ratings

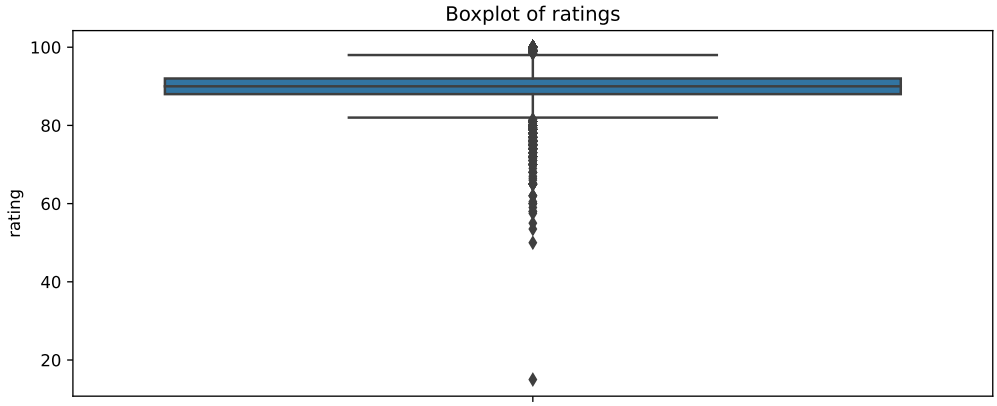


Figure 3: Boxplot of ratings

Measurement	Value
count	96990.000000
mean	90.027585
std	3.033268
min	15.000000
25%	88.000000
50%	90.000000
75%	92.000000
max	100.000000

- What are the patterns of the winescores over time?

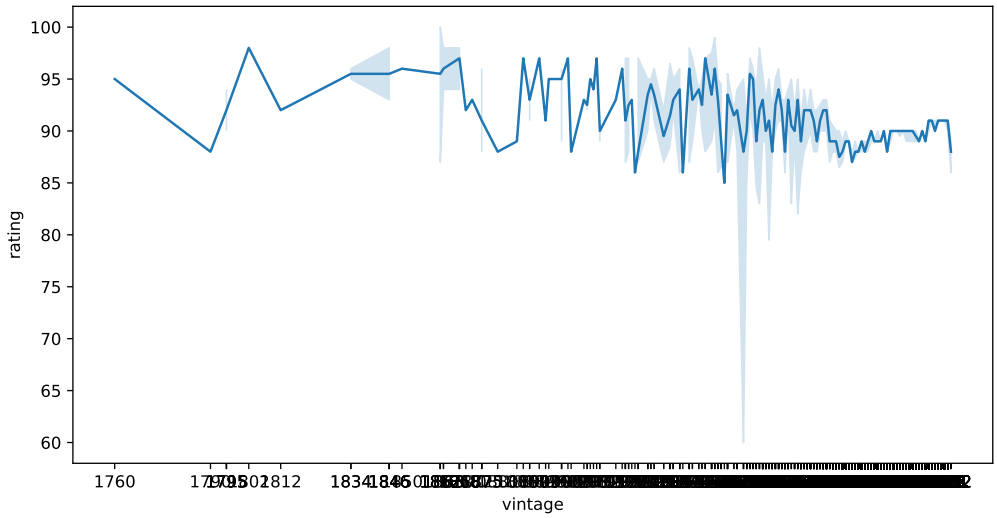


Figure 4: Median ratings over the years

The figure shows the median of the scores over the years. There is no clear rising or falling trend visible.

- In what extent do the variables (sweetness, color, producer, type, reviewer, variety, location, appellation, vintage) measured by Robert Parker influece the score of the wine?

Boxplots have been made for the mentioned variables to provide more insight into their distributions. These can be found in the attachments: appendix I.

As can be seen in the ratings-histogram, the rating appears to be normally distributed. However, this can only be determined with certainty after a test.

alpha = 0.05 H_0 = The ratings followw a normal distribution

H_1 = The ratings do not follow a normal distribution

Test statistic	P-value
0.08359504079899555	0.0009999999999998899

Figure 5: Test results of the Lilliefors normal distribution test

P-value < 0.05 which means that the null hypothesis is rejected. The Lilliefors test shows that the 'rating' is not normally distributed.

The following results have come from the kruskal wallis tests.

The following general hypotheses have been formulated for each test:

H_0 = The groups in the variable are equal in terms of rating

H_1 = At least one group is different than another group in terms of rating

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

Evaluation

Evaluate results

Looking at the total ranges of the scores this varies between 15 and 100, However if we focus on where the majority of the scores are this shows an average that is 90, 'Lower Scores' seem to be outliers. Based upon the modelling (Kruskal Wallis test) it can be stated that all variables have at least one group that significantly differs from another group in terms of rating. With all the tests the 'test statistic' > 'chi critical value' which means that the null hypothesis is rejected and therefore there is a difference.

Data mining results assessment

Based on this research, various insights and new questions have emerged. The modeling chapter showed that the ratings are mainly in a high range (88-92). This is an important finding that cannot be gleaned directly from Robert Parker's website. This finding affects the representativity of the dataset. In the context of further research, it will be difficult to determine the impact of, for example, soil type on the basis of a Robert Parker rating. There is little spread in the ratings which causes the mutual differences to be small. Based on the tests performed, it can be said that within all the variables studied, groups differ (at least one from another) among themselves in the context of rating.

This is a first step, to really say something about the influence of specific factors on the rating, it will have to be investigated which groups these are and whether a combination of these factors leads to a higher or lower rating.

All in all, this means that an answer cannot be given directly to the main question: 'How reliable are Robert Parker's reviews'. Further research will first have to be done on the aforementioned questions.

Approved models

This project produced a git repository which contains the following:

- Scraper to extract data from Robert Parker's website
- The actual data from Robert parker's website
- Dashboard to visualize the data

Review process

Within the project, the main challenge was to clarify the business understanding and the process leading up to it. This also meant changing the subject of the Capstone project multiple times. This took a lot of time, so there was a lot of work to do in a short time.

Within the process a number of improvements are to be considered:

- Set a deadline for obtaining the business objectives. This is a bit more difficult with this project because there is no external client. Setting a deadline on obtaining the business objectives will help to leave enough time for the other topics.
- An assumption was made of a normal distribution for the score, however when tested, it turned out that there was no normal distribution. Next time do not start on an assumption.
- Following the previous project, a scraper had already been built to extract ratings from Robert Parker's website. However, this scraper was built completely different, so this took more time than expected. In addition, it was also not efficient enough to retrieve larger amounts of data.
- Data has currently been collected from four countries. Improvements can be made by collecting data from different countries aswell.
- Quality remains an issue to measure and quantify. A point for improvement could be to look at alternative data to do comparisons, for example data from Vivino. However, this did not fit into the timeframe of this project.

Discussion

- In what range are the scores given?

50% of the scores are between 88 and 92. The minimum score is 15 and the maximum score is 100. In general, the scores are therefore in a small cluster in the top end of the range. The 'skewed' distribution of the wine scores raises some question marks. Are the wines that Robert Parker reviews really all good? Or does it still play a role that a specific colour, type or grape variety is generally found to be tastier?

- What is the pattern in the scores over the years?

The scores have remained stable over the years. There is no clear trend in the median scores over the years.

- Are there factors that influence the scores given to the wine?

A test was done with each variable to determine whether the difference between the variables has an influence on the rating of the wine. All these tests have shown that there is a significant difference between at least one group and another group in terms of rating. This does not say anything about which groups these are. Further research will have to be done on this.

Conclusions

The business objective: How reliable are Robert Parker's reviews? Is the data representative enough for further research? In further research, the aforementioned factors (for example soil type) will be compared to these data.

With the currently available results it is not possible to answer the main question. To this end, further research will have to be done. However the small spread in the ratings currently collected does indicate that representativeness of the reviews might not be a good fit to use in research to measure quality.

Determine next steps

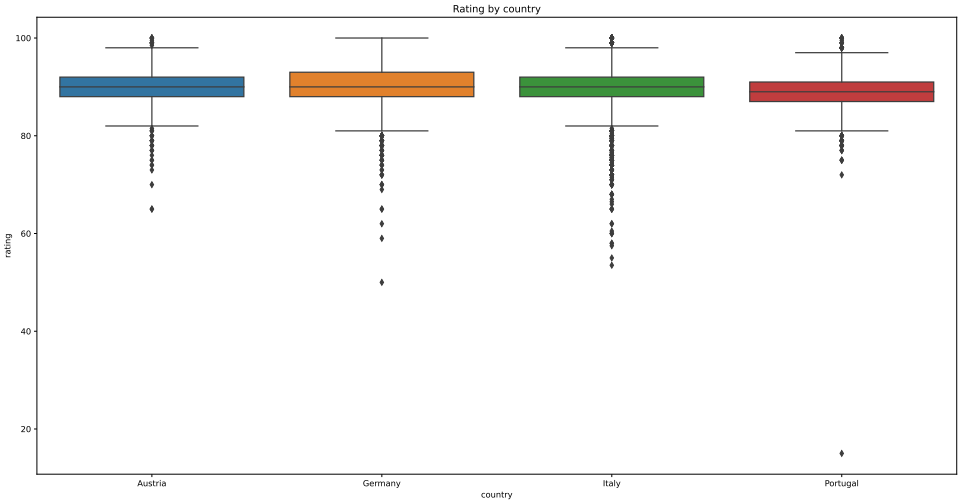
- Adding more data to the research to be able to draw conclusions with more substantiation.
- Comparing data from other 'review' sites such as *Vivino* which is crowd based to rule out any possibility of a reviewers specific taste.
- Review the processes Robert Parker has to determine a wine rating and if it is not based in a pay-for-rating system.
- Study more in depth articles such as A Pragmatic Approach to Using Wine Ratings.

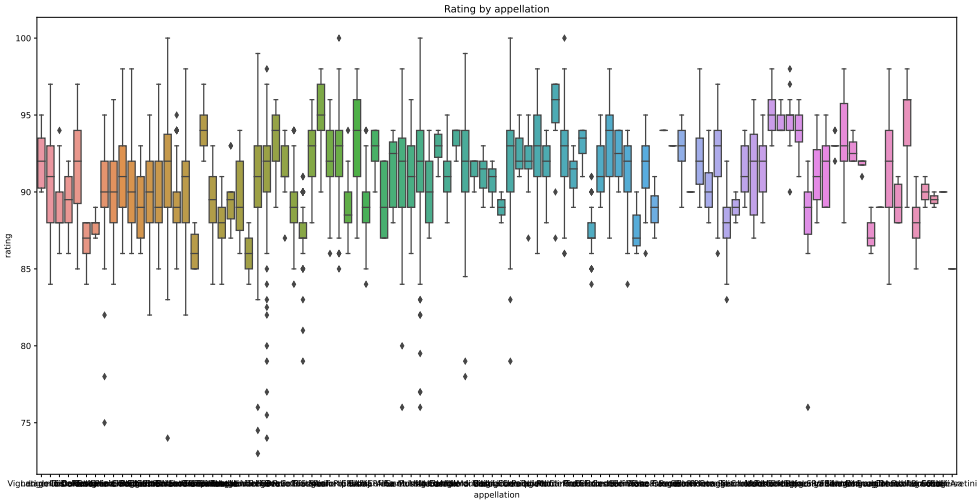
Deployment

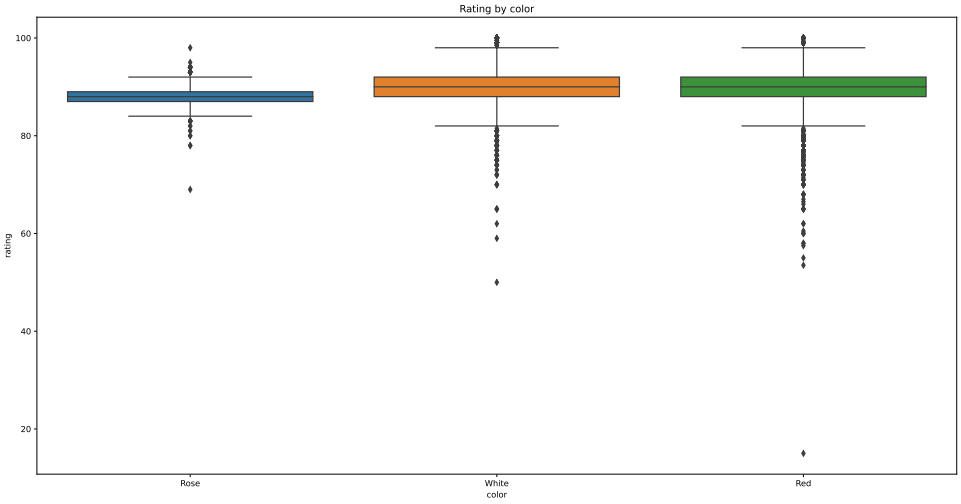
- For now, a personal subscription for Robert Parker's website has been used. For an actual deployment this should be a NPA (non personal account).
- Automate the collection of data. This should be done with a full initial load and then incremental changes.
- Since it cannot be said with certainty whether the data will remain the same, it would be a good addition to add tests. These tests can match on the names of variables that are retrieved.
- A git repository has been created for this project. Changes to the project will be deployed there.

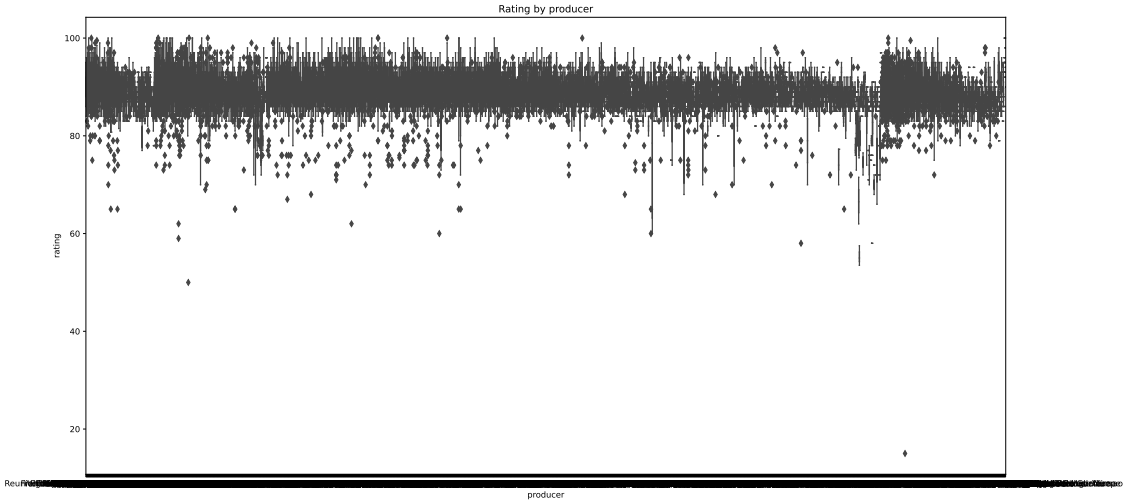
Appendixes

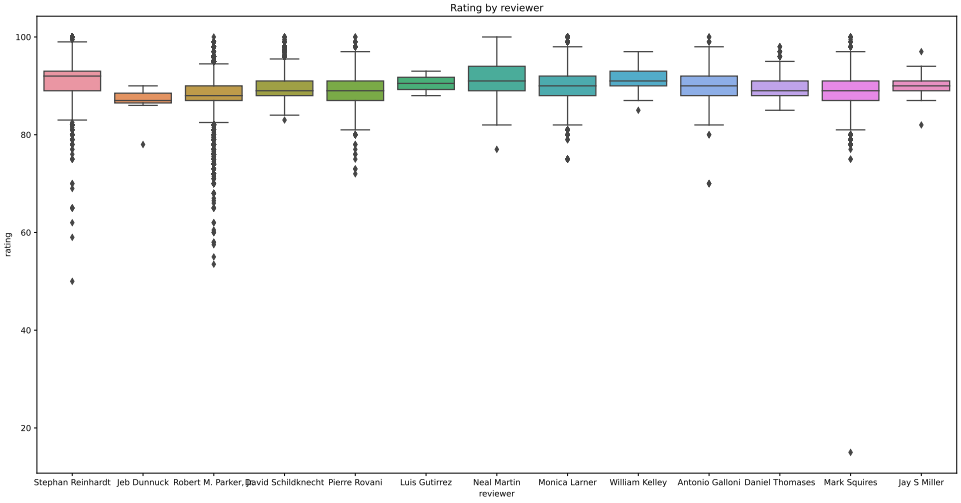
Appendix I

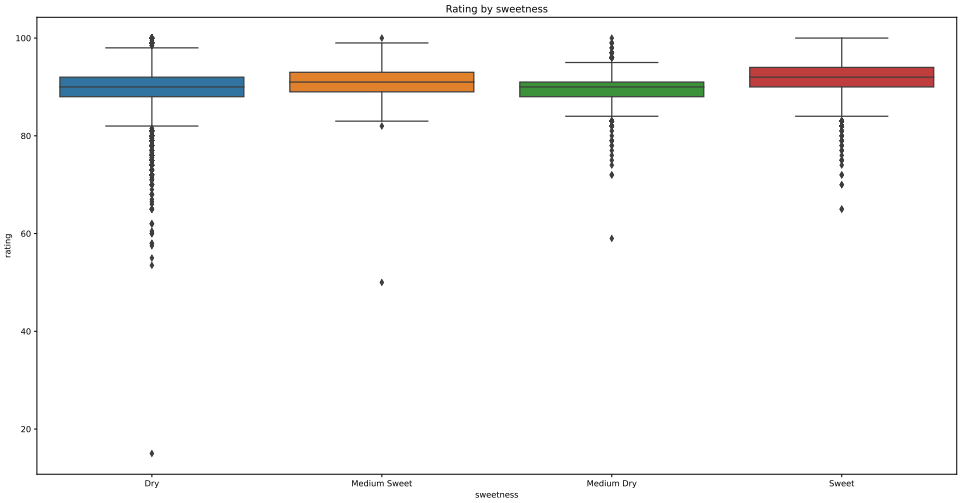


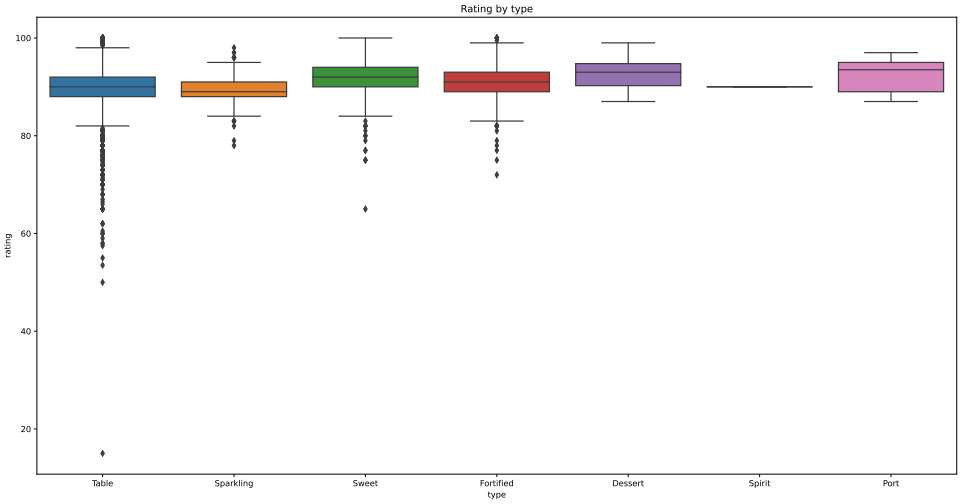


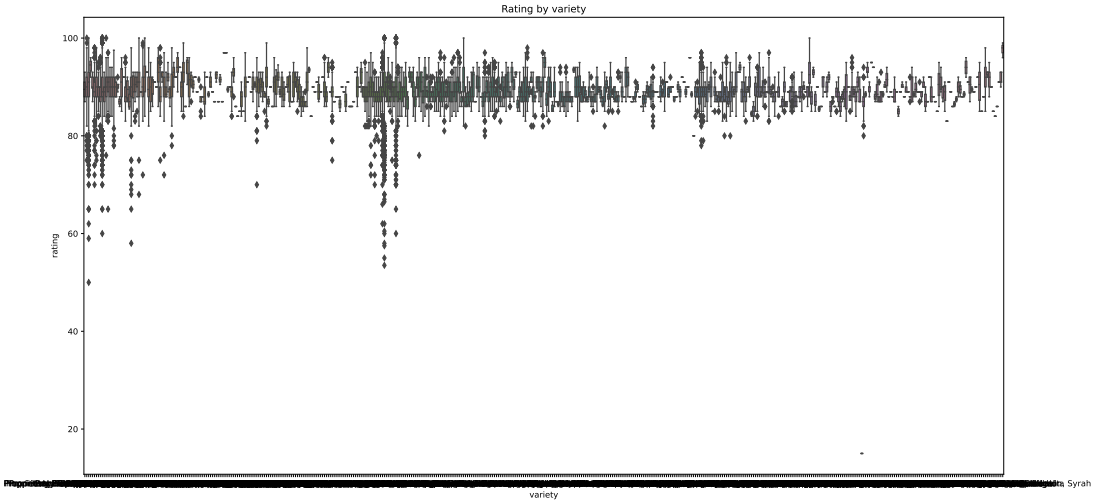

Rating by color
rating
100
80
60
40
20
Rose
White
Red
color

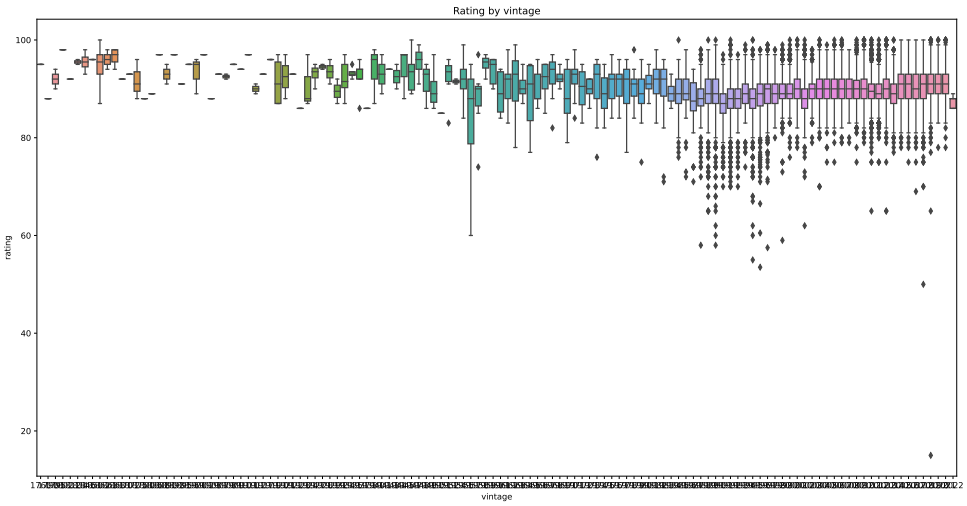

Rating by producer
rating
100
80
60
40
20
producer


Rating by reviewer
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 by sweetness
rating
100
80
60
40
20
Dry
Medium Sweet
Medium Dry
Sweet
sweetness


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


Rating by variety
rating
100
80
60
40
20
variety


Rating by vintage
rating
vintage
100
80
60
40
20