Which Criteria Define a Good Wine From a sensory analysis point of view

TIMOTHÉE DARMAILLACQ, TOM PEYPOUDAT

ACM Reference Format:
Timothée Darmaillacq, Tom Peypoudat. 2024. Which Criteria Define a Good Wine From a sensory analysis point of view. 1, 1 (January 2024), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Abstract

This document is the achievement of universitarian project done for the "Data science and IA with Python languages" at uvic. It is not mean to be published or used in different context than the actual one.

1 INTRODUCTION

When we learn that we have this assignment the both quickly thought about the same idea : doing it on wine. We are both from the region of Bordeaux which is well know for it's wine, we were use to talk about and taste different wine with our family. Wine has a cultural importance, made by man since eight thousand years and still enjoyed today, but it's a complex product. It's often challenging to choose between the different wine. There are plenty of region, grape variety, price. For normal people, like us, who are not wine expert it's complicated. It's unfortunate that this choice is made randomly, it would be more interesting having information helping us to choose. In this project this what we will be trying to do using data science.

During our research we came across multiple studies, based on data science, related to wine consumption : Using wine data base to analyse the result of different statistical method other looking for the link between the wine chemical composition and his quality, these papers focuses on the chemical structure of the wine which is interesting because you use accurate and quantitative measure. Us we are more interested in the human way of experiencing wine, and some studies followed the same thinking as we did. The first one find a link between the text review of wines and their prices. The second one is a really interesting paper measuring the accuracy of wine reviews, their conclusion his that wine reviews done by professionals are statistical quite accurate. Wine spectator, a famous wine review publisher, had more than 87 percent of accuracy on their reviews and the other important publisher had little lower accuracy result. This publication comfort us in the idea of doing data science on wine and that using wine reviews as a source of information was relevant and could give us interesting results.

The more human sensory approach that wine reviews can give is for us more interesting. Because it is closer to the experience that any consumer could have and so do you. We really wanted to have general result that could be easily useful, insight that we could take into account the next time we will want to find a good wine. We won't be able to take into account preferences of different people but it's not really our point. Like every food product you have difference in quality, induced by plenty of different reason, that lead to product being factually better than other. We are looking to

Author's address: Timothée Darmaillacq, Tom Peypoudat, timothee.darmaillacq,tom.peypoudat@uvic.cat.

Manuscript submitted to ACM

find this kind of differences that for some fruit for example are easily identifiable but for wine it's not the case. The data from the reviews of wine expert is the most pertinent for the goal of our work. The reviewers are use to drink wine therefore they have a more complete and subtle overview of the wine quality that will be shared, at least partly, by other people. And we make sure to reviews were the variables taken into account was only the quality of the wine, not the packaging neither the price or the prestige, to have more reliable information. There are protocols followed by the reviewers to ensure this. The main one are : the bottle is hided, they are in specific room that ensure a calm atmosphere, and they only taste limited sample each time. Knowing that we collect data from two sources : our first data set is composed of 2'000 wine reviews that we scrapped from "Wine Spectator" as specialised wine media. And the second one is a data set composed of 130'000 wine reviews scrapped from 'Wine Enthusiast" already used is some data science project. With this pertinent data we begin our work to answer "Which are the relevant information to look upon to when trying to buy good wine"

2 MATERIALS AND METHODS

2.1 Data Preparation

We work upon two data bases of wine reviews made by expert, we choose this one because the validity of the reviews have a capital importance in the analysis we are willing to do. The two organism that we got the data from are two well know, experience and open and clear on the method used to do the review.

The first data base is composed of only the best 100 wines of the years for the last 20 years. The ranking is really interesting because it is made in curated conditions respecting a process that assures non biased data. They taste more than 15000 different wines per year, they focus on widely available wine from the US and other markets. The tastings are made by experts from different regions in optimum condition : private room, blindly therefore they can't be influenced by prestige nor price. High scoring wines are frequently re-tasted and they are financially independent and follow an ethic code promoting fairness. The given criteria are : the name of the bottle with the name of the domain and region and sometimes type of grapes, the ranking for a specific years, a grade from 50 to 100 based on the quality of the wine (in the ranking we only have wine with a grade higher than 90, considerate as 'Outstanding' or 'Excellent'), the price and the years of production of the wine.

The second data base compile 130 000 references of wine bottles with numerous details about each bottle. In particular, the province in which the grape was cultivated, its country, the winery in which the wine was made, and the main grape variety in the wine.They might be little less rigorous than the first organism but we though it would be interesting to work on both to be able to compare the result we obtain with this two different samples. Before applying method we sorted the sorted the data, we looked for any incoherence entry but did not find any. The only issues we encounter were missing data in some attribute of few reviews.

2.2 Methods

We begin by analysing the our data following a descriptive methodology to have an overlook and understanding of what we were working on. We did all our work using python and it's libraries like pandas, matplotlib, numpy. We begin by calculating the correlation between the price and the grade of the wine because it's the thing that was the most expected we find a significant correlation for the first data set and a lower for the second.

This result indicates a moderate positive correlation between the two variables. In other words, there is a tendency for wines with higher ratings to have higher prices, and vice versa. However, it is important to note that correlation

```
...  Price caracteristic for the WineSpectator data
     count    2100.000000
     mean       48.052857
     std        40.951110
     min         8.000000
     25%        22.000000
     50%        35.000000
     75%        60.000000
     max       535.000000
     Name: Price, dtype: float64
     The WineSpectator correlation between 'Price' and 'Grade': 0.7126281060632821
     The WineMagasin correlation between 'Price' and 'Age': 0.30033709399960695
     Price caracteristic for the WineMagasin data
     count    120913.000000
     mean         35.368687
     std          41.031345
     min           4.000000
     25%          17.000000
     50%          25.000000
     75%          42.000000
     max        3300.000000
     Name: Price, dtype: float64
     The WineMagasin correlation between 'Price' and 'Grade': 0.4161919001430962
```

Fig. 1. Information about the price and correlation

```
Top 5 wineries, province, countries and varieties with the highest frequency and their average points:

                Winery  Frequency  Average Points
0  Wines & Winemakers        222       87.599099
1         Testarossa        218       90.738532
2          DFJ Vinhos        215       86.669767
3     Williams Selyem        211       92.744076
4       Louis Latour        199       90.537688

     Province  Frequency  Average Points
0  California      36247       88.627876
1  Washington       8639       88.947216
2    Bordeaux       5941       88.658980
3     Tuscany       5897       89.074614
4      Oregon       5373       89.051926

    Country  Frequency  Average Points
0        US      54504       88.563785
1    France      22093       88.845109
2     Italy      19540       88.562055
3     Spain       6645       87.288337
4  Portugal       5691       88.250220

                  Variety  Frequency  Average Points
0              Pinot Noir      13272       89.411468
1              Chardonnay      11753       88.340083
2       Cabernet Sauvignon       9472       88.607580
3               Red Blend       8946       88.380282
4  Bordeaux-style Red Blend       6914       89.106306
```

Fig. 2. frequency analysis

does not imply direct causation. Other factors may influence both the price and the rating of a wine, and correlation provides only a measure of the strength and direction of the linear relationship between these two specific variables. A correlation of 0.713 still suggests a significant association between the price and the rating of wines within the context of your data set.We then calculte mean() to obtain the average price based on a certain rating. In our case, the average price for wines rated 98/100 or higher. With this, we obtained everything we wanted related to the price of wines.

For non numerical data we choose to use a frequency analysis.With the substantial amount of information of the second data base enable us to deduce the provinces, countries, grape varieties, or wineries that result in the most appreciated wines.

And from this analysis we can conclude that the most appreciated variety is Pinot Noir, for the teritory it is Tuscany in Italy, the country from wich the wine is most apreciated is France. The Best winery is Williams Selyem in the US. The wine around 5 years old are the most appreciated and the price around which the wine is most profitable is 50 euros. To conclude we can confirm that the quality of the wine on a taste point of view is linked to his price.

3 DISCUSSION

Contribution : During this project, we quickly agreed on the theme of the project. Subsequently, both of us actively searched for as many resources as possible to proceed with our analyses. Once this was done, Timothée took charge of organizing all the data into an Excel file for easier processing. Tom primarily handled various analyses, such as correlation, analyses related to pricing, and the different top 5 rankings obtained from the database of 130,000 references. Timothée took care of formatting the final report and contributed to some of the interpretations.
We could say that we learn different ability doing this project. Of course about data science but also on the management of this kind of task, the redaction of publication and the adaptation needed to be able to go further when something doesn't work as well as we expected.

The result we find did not shocked us, they seems logical from an exterior point of view. Our project lead to backing some already heard a priori about wine, the more expensive bottle are better and the region well know for their wine produce overall better wine. But maybe in the future this result could change, so we all need to stay open and curious.

References Wineinformatics: A Quantitative Analysis of Wine Reviewers :
https://www.mdpi.com/2311-5637/4/4/82
Wine Quality and Taste Classification Using Machine Learning Model :
https://ijirase.com/assets/paper/issue$_1$/$volume_4$/$V4 - Issue - 4 - 715 - 721.pdf$

Thank you for reading this project.