Data Analysis of Wine

Hallie Hohbein, Niloofar Mansoor, Miracle Modey, Saeideh Samani, Bruno Silva

1. Introduction

Generally, people are using sensational evaluation to differentiate between wines. Different studies have been done in chemical and spectroscopic analysis of wines using statistical techniques [1].

Different methods of data analysis have been used to explore the chemical properties of wine in order to differentiate between a variety of wines [2]. The volatile composition of wines is mostly associated with the sensory analysis which are the normal ways of differentiating wines. The functional way of distinguishing different types of wines are: exploring the amount of alcohol, volatile fatty acids, and esters [3]. The number of wine enthusiast is increasing as years pass by. Spain, Italy, and France were the top three countries as wine exporters in 2017 [4]. For example, Spain exported a total of 22.8 million hectoliters of wine in 2017, a significant sale increase compared to previous years [5].

The wine industry is trying to support this growth by applying new technologies and using data analysis and data modeling techniques for production and consumption purposes.

To evaluate the quality of wine for generating the certificate, both sensory tests and physicochemical are being assessed though since the sensory tests rely primarily on human sense, the wine categorization is an exhausting task. Also, the relationships between sensory tests and physicochemical analysis are not completely studied [6].

Data analysis techniques aided scientists to inspect, transforming, and modeling data which made it possible to discover useful information and eventually in decision making aspects [7]. Data mining as a particular data analysis technique tries to obtain high-level knowledge from raw data [8]. In this project, we aim to analyze sensory wine data to discover correlations between prices, countries, and ratings. We used the *Wine Reviews [9]* dataset specifically because most of the related work about wine data analysis was based on the data about wine's physiochemical properties.

Section 2 describes the challenges we faced while studying the dataset and performing our analysis. Section 3 describes some of the related work about wine data analysis and studies. Section 4 describes our approach to data analysis and visualization of our results. We describe our evaluation in detail in Section 5.

2. Challenges

We chose the *Wine Reviews* dataset [3], which includes structured review data from Twitter users, containing 150926 wines with 11 attributes. We downloaded this dataset from *Kaggle* [10], an online community of data scientists and machine learners which allows users to find and publish data sets, explore and build models to solve challenges. The dataset is publicly available. The data in this dataset has attributes such as the name of the wine, its country and regions of origin, variety, winery, the taster's Twitter handle, price, and the score they have given the wine. The score lies between 80 to 100. The dataset has more than 150,000 rows. In our preliminary studies of the dataset, we realized that there are duplicated rows in the dataset. After cleaning up the duplicates, we had more than 97,000 rows in our dataset.

We also had to deal with some missing values in our dataset. There were a few rows that did not have the *country* of the corresponding wine, but we ignored that because the number was very insignificant (less than 10) and all of those wines were not highly-rated. About 60% of the rows do not have the *region_2* value listed, which is a field we did not consider in our analysis. 9% of the rows did not have a price value, and we did not consider those rows for the part of the analysis that factored in the price of the wine.

3. Related Work

Data analytical techniques play an important role in the interpretation of chemical and instrumental properties of wine [11]. Two methods which can be used to investigate the intrinsic characteristics of wine are principal component analysis and linear discriminant analysis. The principal component analysis is an unsupervised classification technique that is used to observe the underlying structure of a data matrix and its results illustrates the degree of similarity between samples and the impact of variables [12, 13]. Some Madeira wines and their chemical properties were studied and PCA was applied in order to determine the main sources of variability present in the data sets and differentiate and characterize the wines [14]. Cortez et al. [15] propose a data mining approach to predict human wine taste preferences, leveraging three regression techniques and studying the physiochemical properties of the wines. There's also recent work on mining large text data to extract meaningful domain-specific information about the sensory properties of wine [16].

Following related works, there is a similar project that covers the principal component analysis of wine [17]. In this project, the wine was analyzed according to 13 attributes. 7/13 of these attributes were used for data analysis and a cluster plot was generated with K-Means clustering.

4. Proposed Approaches

We needed to explore and analyze our data. We used Microsoft Excel and the Python programming language to do data exploration and analysis. For data visualization, we used D3, JavaScript, and JQuery. Excel was used as our tool for initial data exploration and visualization, and we identified the missing data and cleaned up our data. However, during the initial investigation, we discovered that this tool would not handle the dataset well while answering the majority of our proposed research questions. Even a basic operation such as sorting the data took several seconds which was not ideal for such a big dataset. Therefore, for scenarios that Excel was not able to handle, we used the Python programming language with the Pandas data analysis library [18], along with matplotlib [19], which is a plotting library to help for visualization. These tools have been shown to be more robust and faster than Excel for large datasets.

5. Evaluation

In completing this project, we aimed to answer these research questions:

- **RQ1:** Is there any correlation between the price and the quality of a wine?
- **RQ2:** Which country has the best wine?
- **RQ3:** Does a well-ranked winery also produce bad wine?
- **RQ4:** Do well-ranked wines have any sort of correlation?

In these following subsections, we answer each research question separately.

5.1. RQ1

For this research question, we wanted to know if there's any correlation between the price and the quality of a wine. We considered the score value and the price value of the wine in the dataset, visualizing this information using JavaScript and D3. Figure 1 is a scatterplot showing the score given to each wine and the wine's price. The *x* axis shows the range of scores (points) given to wine, and the *y* axis shows the range of wine prices. The different colors for the points represent different wineries. Looking at this visualization, we realize that price does play a role in wine quality. Specifically, we can see that in the price range is smaller in the smaller points, meaning that usually the poorly ranked wines are not as expensive as highly ranked wines are. We can also see that the higher prices correlate to higher scores in the chart. We can conclude that highly ranked wines have a higher price than poorly ranked wines and that price plays a role in determining a wine's quality. The relationship is not completely linear and the data does not conclude a specific pattern of a wine's quality when the price is between $0-$200.

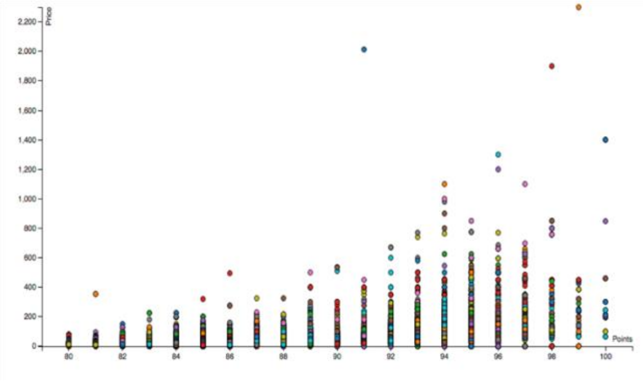

Figure 1: Scatterplot showing price-points relationship

5.2. RQ2

For research question two we wanted to discover which country was producing the best wine in the world.

Our dataset includes more entries from some countries compared to others. Therefore, comparing the number of wines with specific scores from each country was not a good solution to realize which country has the best wine. Figure 2 shows the percentage of entries from top 8 countries with the most entries in regards to the whole dataset.

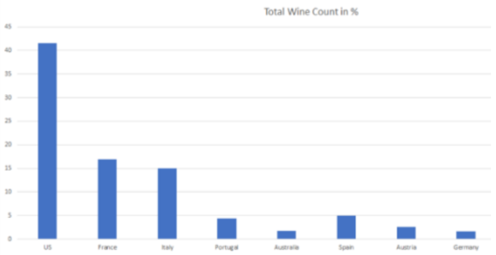

Figure 2: Countries' share of entries in the dataset

To answer the second question, we divided the points given to wines into five categories. We considered wines with points 80-84 in the "Poor" category, points 84-88 in the "Fair" category, points 88-92 in the "Good" category, points 92-96 in the "Very good" category, and finally, points 96-100 in the "Excellent" category. We separated the wines from each country and calculated what percentage of the wines from each country belong in each category. We visualized our result using D3 and styled our webpage with Bootstrap v4.

After reading in the csv format wine data using python, we add this data to the countries properties in a GeoJSON file [20]. Then, we load the GeoJSON output to the application. Next, we make the map using the geometry feature of the GeoJSON file. Using the property feature of the GeoJSON file, we then extract data to build donut charts and bar charts. Each bar chart is a graphical object that will be added to the html file after creating the map and donut charts. To make the web page interactive, JavaScript and jQuery have been used to detect the actions of the cursor.

Our world map visualization also shows the level of wine production for each country appearing in our dataset (Figure 3). Hovering over each country shows the percentage of each category of wine from that country using a donut chart. This can be seen in the webpage that we have created to show our visualization.

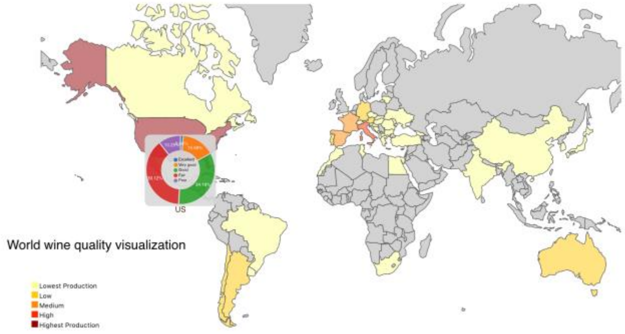

Figure 3: World map showing production level of wine

After gathering this data, we compared the percentages of excellent, very good, good, fair and poor wine from each county with others and created bar charts that indicate which country has the best wine. Figure 4 shows that France has the best wine, since 1.8% of the produced wine in France belongs in the excellent category. Australia and Germany, and US follow France in this ranking. It should be noted that US's wine entries in our dataset are far more than France's, but the quality of the French wine considering the number of entries is better. The rest of the bar charts (for all the different categories) are available on our webpage. One interesting point is that the number of countries having entries in those lower categories kept increasing as the point range decreased. We can conclude that despite the fact that a large number of countries do produce wine, but only a few of them produce excellent or very good wine.

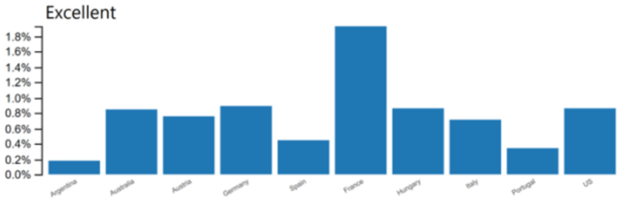

Figure 4: Comparison of excellent wines in different countries

5.3. RQ3

Research question 3 looked to discover if a well-ranked winery also produced bad wine.

To find the answer to this question, first we found the highest and poorest ranked wineries seen in *Table 1* and *Table 2* respectively. This data was found by first using Javascript to generate and order the highest and lowest ranked wineries and print it into the console. There is a column for the highest and a column for the lowest. Then, in Excel we used the advanced filter feature to look at both columns and highlight the boxes that are the same in both columns. From this, we found that Tobreck and Williams Selyem were found on both lists.

Table 1: Best wineries

Country	Winery			
France	*Louis Jadot*	*Château Ausone*	*Domaine Leflaive*	*Château Lafite Rothschild*
Germany	*Geh. Rat Dr. von Bassermann-Jordan*	*Schloss Johannisberg*		
Australia	*Campbells*	*Penfolds*	*Chambers Rosewood Vineyards*	*Torbreck*
US	*Williams Selyem*	*Cayuse*	*Charles Smith*	*Goldeneye*
Italy	*Tenuta dell'Ornellaia*	*Le Macchiole*	*Marchesi Antinori*	*Avignonesi*

Table 2: Poor wineries

Country	Winery			
France	*Domaines Barons de Rothschild*	*Millésimé*	*Vignobles de France*	
Germany	*Meilen*	*Rudolf Müller*		
Australia	*Torbreck*			
US	*Williams Selyem*	*10 Knots*	*Adobe Road*	*Agate Ridge*
Chile	*Alba*	*De Martino*	*Miguel Torres*	*MontGras*

To create the visualization, which is the word cloud seen in *Figure 5*, we used a D3 [21] visualization to analyze and display wine data. It starts by reading the csv file and checking each wine to see if it is excellent wine, poor wine, or neither. If it is neither, then the line is discarded. If it is excellent, then the winery where it was produced is added to the excellent winery set. If the line is poor, then the winery where it was produced is added to the poor winery set. This is so that each winery only shows up once in a visualization. It should be noted that when the winery is added, the spaces are replaced with underscores so that wineries with more than one word do not end up being multiple, separate words in the word cloud. Then, the JavaScript code iterates through both the excellent and poor sets. If a winery is found in both sets, then it is added as a key-pair to a list with a larger font size. Otherwise the element is added as a key-pair element with a smaller font size. This final list of key-pair elements is then given to the D3 implementation and the program displays the wineries, with the wineries found in both the poor and excellent lists being large and red. In this case, there are only 2 wineries that are in both data sets.

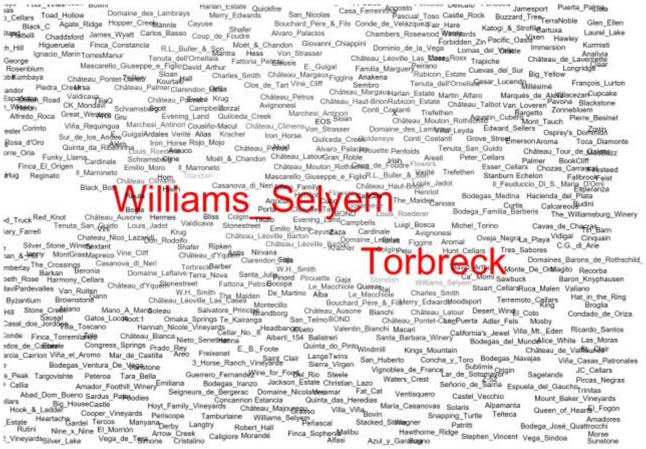

Figure 5: Word Cloud showing the wineries with both excellent and poor wines

5.3. RQ4

For this research question, we want to find what attributes correlate to a wine being well-ranked.

As mentioned above, we selected a subset of values in our dataset containing only well-ranked wines. It is important to mention that the dataset contains only wines with the attribute indicating its quality equal to or greater than 80. Therefore, we decided to create a subset called *well-ranked wines*, where their *points* are greater than or equal to 95. It was observed that this cluster has duplicated values. Duplication removal was possible through the method *drop_duplicates* provided by Pandas, leaving 2368 wines to be analyzed.

The majority of attributes in the dataset are in string format. A conversion to integer is necessary in order to find correlations between columns and the *points* attribute. For this, we mapped all the unique values for each column attribute to integers. For example, instead of displaying that a wine is from Australia, we say it is from country 1. The mapping is possible through the method *replace* from Pandas library.

After removing duplicates, clustering wines, and converting string values to integers, we are ready to draw correlations about the dataset to find what attributes influence the *points* attribute. In other words, what columns contribute for high *points* values. For this, we calculated the Pearson correlation between all the attributes except the *region_2* and *points* columns. The column region_2 was not considered because 60% of the wines in the cluster do not have a value for that attribute. The result of the Pearson correlation [22] is shown in Figure 6. The result shows that there is a moderate correlation between *points* and *price*, and a weak correlation between *points* and *winery*. Due to the moderate correlation, we can conclude that price is significant when determining if a given wine has a high classification rate. It is important to emphasize that price only gives minimum insight if a wine is highly rated. To visualize this correlation, we divided wines in groups based on their *points* column. We conclude that, on average, better wine indeed costs more, as shown in Figure 7.

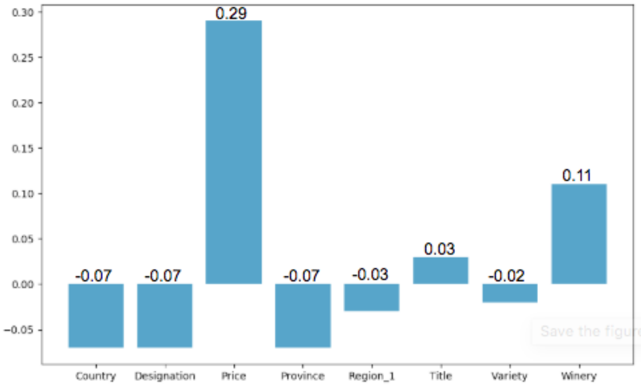

Figure 6: Pearson correlation between Points and other attributes

To interpret the weak correlation, we selected the top 4 wineries that produce most of the wines in the dataset, which are Louis Jadot, Domaine Zind-Humbrecht, Cayuse and Williams Selyem. We then plotted the point of each wine for the 4 wineries, as pictured in Figure 8. We can observe that wineries produce both high and low ranked wines. There is not any strong support to conclude that high rated wines depend only on the winery that they came from.

The steps to reproduce the results for this research question can be found in the Jupyter Notebook file *wine_rq4.ipynb*.

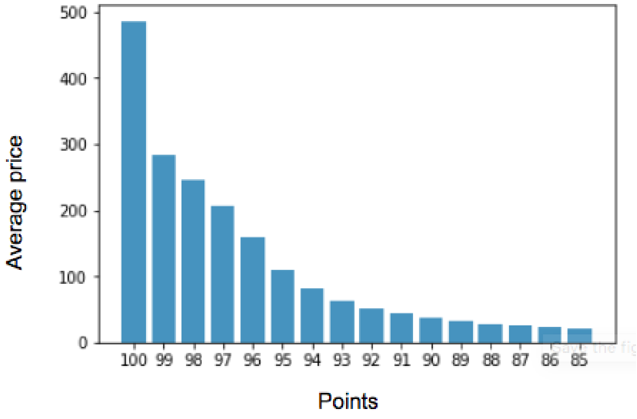

Figure 7: Average price of wine grouped by *points* attribute

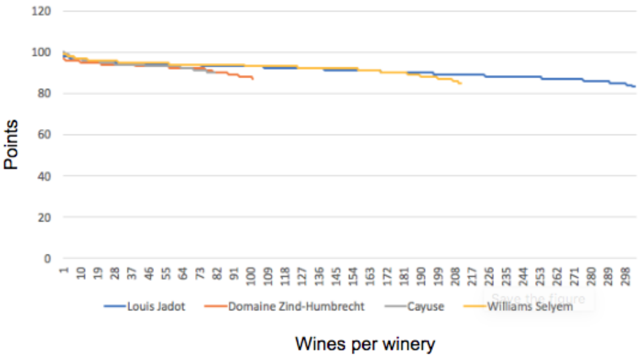

Figure 8: Points for each wine from top 4 wineries

6. Conclusion

This project aims to analyze wine data in order to investigate what attributes collaborate for a wine to be well-ranked. The data was gathered from a from a publicly available dataset, containing several wine attributes such as country, points, price and winery.

Based on the data analysis of the dataset, we concluded that France produces the best wines in the world, price helps to identify good wines but, it is not the main determining factor, and good wineries also produce low quality wines.

References

1. Kwan W, Kowalski BJJoFS. Classification of wines by applying pattern recognition to chemical composition data. 1978;43(4):1320-3.
2. Cozzolino D, Smyth HE, Gishen MJJoA, Chemistry F. Feasibility study on the use of visible and near-infrared spectroscopy together with chemometrics to discriminate between commercial white wines of different varietal origins. 2003;51(26):7703-8.
3. Louw L, Roux K, Tredoux A, Tomic O, Naes T, Nieuwoudt HH, et al. Characterization of selected South African young cultivar wines using FTMIR spectroscopy, gas chromatography, and multivariate data analysis. 2009;57(7):2623-32.
4. Wine Statistics [Available from: https://www.statista.com/statistics/240649/top-wine-exporting-countries-since-2007/.
5. Wine Sales Statistics [Available from: https://www.foodswinesfromspain.com/spanishfoodwine/global/whats-new/news/new-detail/review-spanish-wine-exports.html.
6. Legin A, Rudnitskaya A, Lvova L, Vlasov Y, Di Natale C, D'amico AJACA. Evaluation of Italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception. 2003;484(1):33-44.
7. Data Analysis [Available from: https://en.wikipedia.org/wiki/Data_analysis.
8. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. 2008;14(1):1-37.
9. Wine Reviews Dataset [Available from: https://www.kaggle.com/zynicide/wine-reviews.
10. Kaggle [Available from: https://www.kaggle.com.
11. Kaufmann AJJoAI. Multivariate statistics as a classification tool in the food laboratory. 1997.

12. Bevin CJ, Dambergs RG, Fergusson AJ, Cozzolino DJACA. Varietal discrimination of Australian wines by means of mid-infrared spectroscopy and multivariate analysis. 2008;621(1):19-23.

13. Næs T, Isaksson T, Fearn T, Davies T. A user friendly guide to multivariate calibration and classification: NIR publications; 2002.

14. Câmara JS, Alves MA, Marques JCJT. Multivariate analysis for the classification and differentiation of Madeira wines according to main grape varieties. 2006;68(5):1512-21.

15. Cortez P, Cerdeira A, Almeida F, Matos T, Reis JJDSS. Modeling wine preferences by data mining from physicochemical properties. 2009;47(4):547-53.

16. Valente CC, Bauer FF, Venter F, Watson B, Nieuwoudt HHJSr. Modelling the sensory space of varietal wines: Mining of large, unstructured text data and visualisation of style patterns. 2018;8(1):4987.

17. Wine Data - PCA [Available from: https://rstudio-pubs-static.s3.amazonaws.com/289957_d17f7ff98bd94d2ba5631b16f48fc6c6.html.

18. Pandas [Available from: https://pandas.pydata.org/.

19. Matplotlib.

20. World GeoJSON [Available from: https://github.com/emeeks/d3-carto-map/blob/master/examples/sampledata/world.geojson.

21. D3 Word Cloud [Available from: http://bl.ocks.org/ericcoopey/6382449.

22. Benesty J, Chen J, Huang Y, Cohen I. Pearson correlation coefficient. Noise reduction in speech processing: Springer; 2009. p. 1-4.