STAT 515 – Applied Statistics & Visualization for Analytics

FINAL PROJECT REPORT ON

Rating on Wines throughout the World

BY

PATHURI NIKHIL REDDY

G01103052

Email: npathuri@gmu.edu

GEORGE MASON UNIVERSITY

FAIRFAX, VA

Abstract

This project aims to visualize a dataset about wines produced in different parts of the world using various techniques and methods which makes us easily understand the data and patterns present in the data by looking at the plots. The challenging part of this project is cleansing and filtering the raw data to help better visualize the data. I have used the ratings and prices of wines to visualize which regions throughout the world produce good and bad wines and many other basic statistics using bar graphs. Determined what are the best varieties of wine and which countries produce them. According to my analysis there was a slight positive relationship between ratings of wine and price of wine (with increase in price there is a slight increase in rating). Extracted the words that are best used to describe a good wine using the description of wines. As the data is mostly comprised of wines produced in US, I have made a deeper analysis on US data and used micromap and plotly to show further statistics on this data. I have also taken advantage of wine description and used it to build a prediction model using Naïve Bayes Classifier. This model would help us to determine whether the wine is good or bad just by looking at the description of wine.

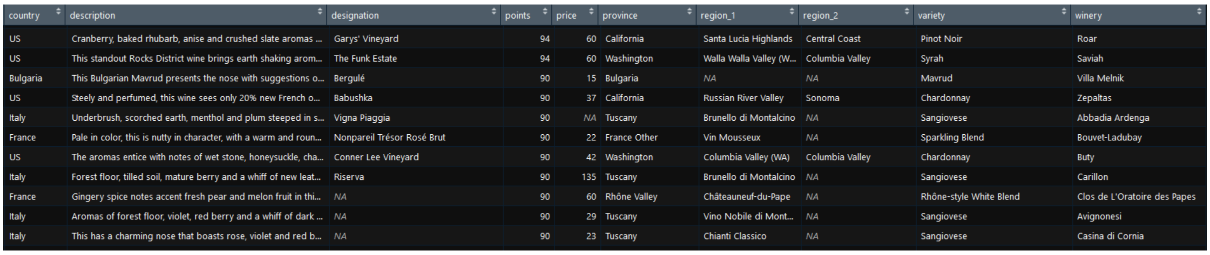
Introduction

Wines are produced throughout the world and there are thousands of different varieties out there. To determine which wine is good, one would check the internet and know the rating of the wine. But, not all the wines have ratings and people would not want to buy a $1000 wine just to know that it tastes bad.

The objective of this analysis is to determine which region in world produces good wines, also which variety has highest rating and help people predict whether a wine is good or bad just by looking at the description.

Dataset Implemented

The data is taken from Kaggle website which is a very popular site to search for various types of datasets. The data has various attributes like country from which the wine is produced, description of wine, designation, points ranging from 80-100, price in x10 USD, province, region, variety of wine and winery from which the wine is produced. The dataset covers data of about more than 150,000 wines.The points are given by various wine testers and the wine which is capable of getting a score more than 80 only are being tested. So, the score in our dataset varies from 80-100.



country	description	designation	points	price	province	region_1	region_2	variety	winery
US	Cranberry, baked rhubarb, anise and crushed slate aromas ...	Garys' Vineyard	94	60	California	Santa Lucia Highlands	Central Coast	Pinot Noir	Roar
US	This standout Rocks District wine brings earth shaking arom...	The Funk Estate	94	60	Washington	Walla Walla Valley (W...	Columbia Valley	Syrah	Saviah
Bulgaria	This Bulgarian Mavrud presents the nose with suggestions o...	Bergulé	90	15	Bulgaria	NA	NA	Mavrud	Villa Melnik
US	Steely and perfumed, this wine sees only 20% new French o...	Babushka	90	37	California	Russian River Valley	Sonoma	Chardonnay	Zepaltas
Italy	Underbrush, scorched earth, menthol and plum steeped in s...	Vigna Piaggia	90	NA	Tuscany	Brunello di Montalcino	NA	Sangiovese	Abbadia Ardenga
France	Pale in color, this is nutty in character, with a warm and roun...	Nonpareil Trésor Rosé Brut	90	22	France Other	Vin Mousseux	NA	Sparkling Blend	Bouvet-Ladubay
US	The aromas entice with notes of wet stone, honeysuckle, cha...	Conner Lee Vineyard	90	42	Washington	Columbia Valley (WA)	Columbia Valley	Chardonnay	Buty
Italy	Forest floor, tilled soil, mature berry and a whiff of new leat...	Riserva	90	135	Tuscany	Brunello di Montalcino	NA	Sangiovese	Carillon
France	Gingery spice notes accent fresh pear and melon fruit in thi...	NA	90	60	Rhône Valley	Châteauneuf-du-Pape	NA	Rhône-style White Blend	Clos de L'Oratoire des Papes
Italy	Aromas of forest floor, violet, red berry and a whiff of dark ...	NA	90	29	Tuscany	Vino Nobile di Mont...	NA	Sangiovese	Avignonesi
Italy	This has a charming nose that boasts rose, violet and red b...	NA	90	23	Tuscany	Chianti Classico	NA	Sangiovese	Casina di Cornia

Fig1: Sample of Dataset (imported to R)

As we can see in the image some of the columns have missing values and not all the columns are necessary to determine a good wine.

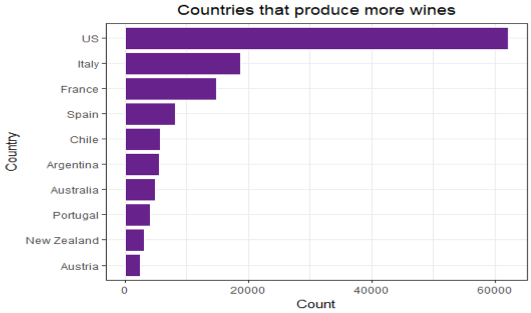
Data Cleaning

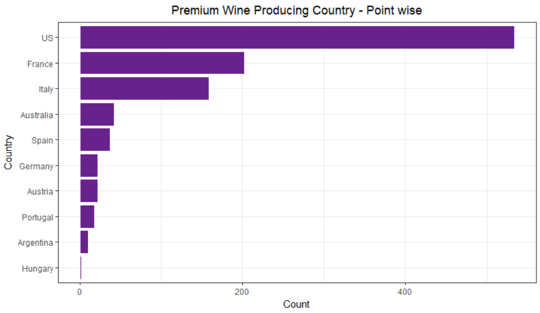
Cleaning is a very important process in order to better visualize the data and avoid creating misleading graphs. The data has different levels of cleaning for usage in different graphs. The first step is to remove the less significant columns like designation, regions and winery. These columns data contribute very less to determine a good wine. As we can see that some rows have missing prices. So, as a second step we need to remove the rows with missing values. Further I have used different filters and aggregations on data in order to make the data suitable for various graphs. As most of the data is from US, I have filtered this data and made a depper analysis. I have also tokenized the data in description and made the data in lower-case, remove numbers, remove stopwords and strip out white spaces to make the data ready for the application of Naïve Bayes Classifier model.

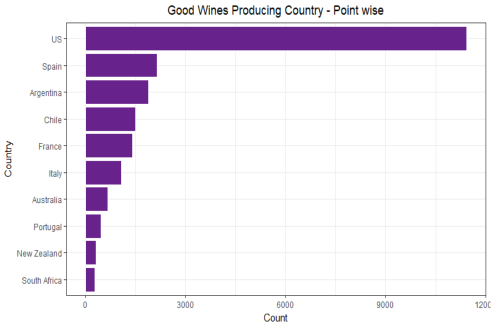
Implementation of Graphs

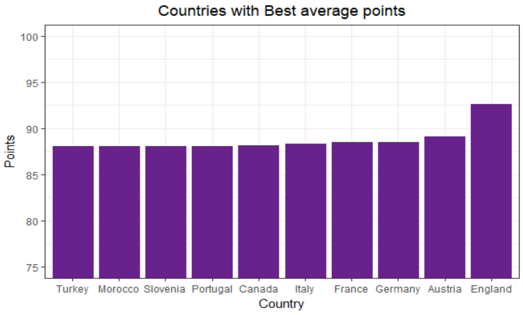
a) Visualization of basic Statistics of the data

First, I have used simple bar graphs to visualize which countries produce the most wines, which countries produce premium and good wines and which countries have good averages.



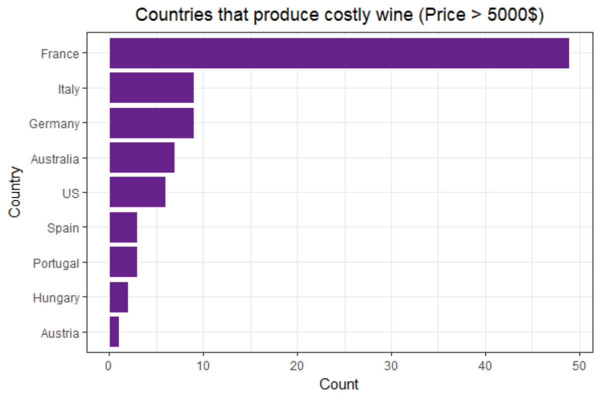


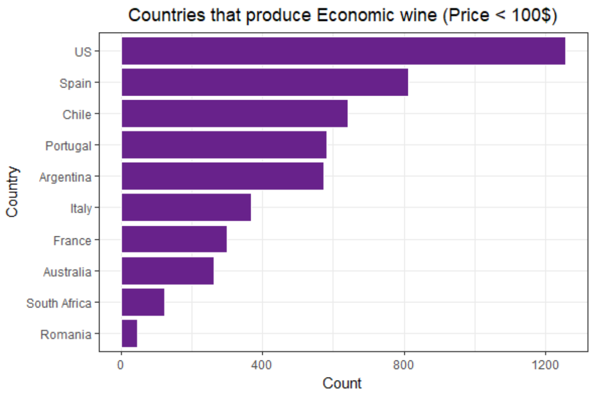


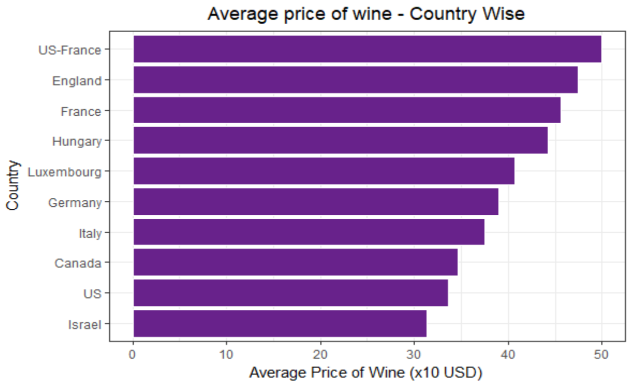


From the first bar graph we can see that US produces the most amount of wines according to the data. From second, we can see that US, France and Italy top for producing premium wines (points > 90). From, third we can see that US, Spain and Argentina produce good wines (points < 90). Just by looking at 2nd and 3rd graph we can not say that a particular country is good and bad. The best way to know a country produces good wines is by looking at averages. From the 4th graph we can see that average points are high for England and Austria. From this we can say that wines from these countries are generally better than other countries.

Now let's take a look on prices also. The below bar graphs give us an idea on which countries produce costly and economic wines and which country has high average values.

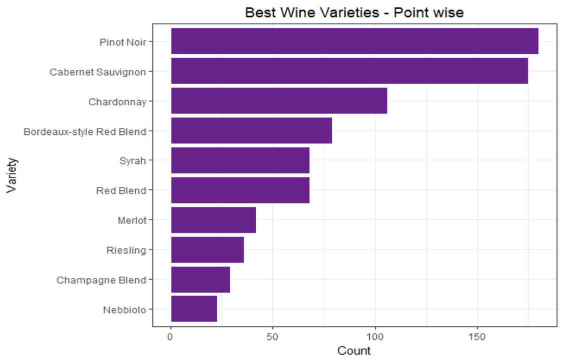


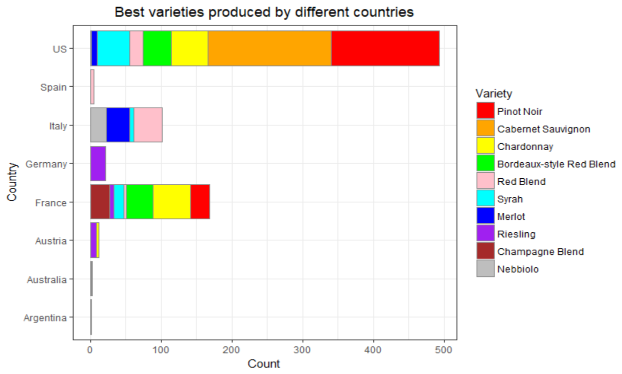




We can see that France produces the costliest wines and whereas economy wines are produced by US, Spain and Chile. The average price is highest for US and France with $500.

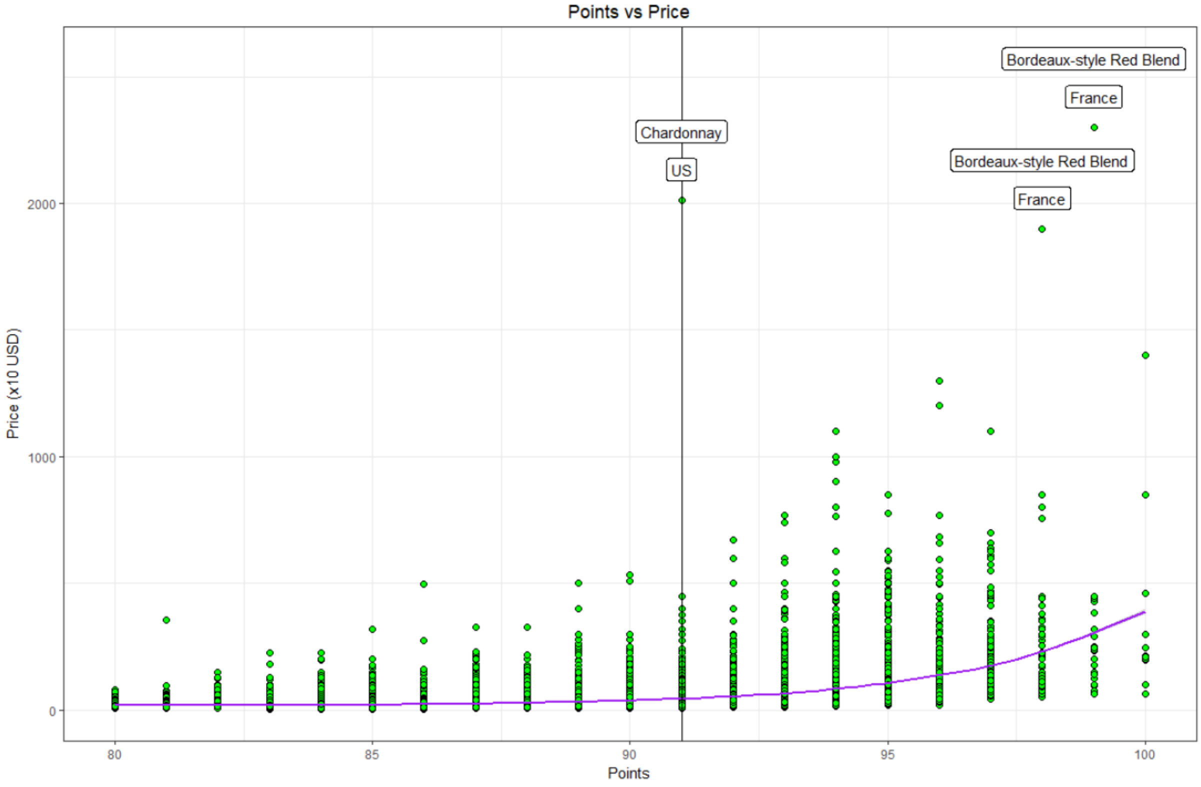
b) Visualizing different varieties of wine





The first bar graph represents the top 10 varieties with high points (>95). Pinot Noir and Cabernet Sauvignon top the list of best varieties. The second graph represents which countries are producing these top varieties. The legend has the same order as ranking from 1st bar graph. I have also used basic colours to easily determine the variety. The interesting point we can infer from this graph is that some varieties are only produced in a particular country. For example, Cabernet Sauvignon is only produced in US.

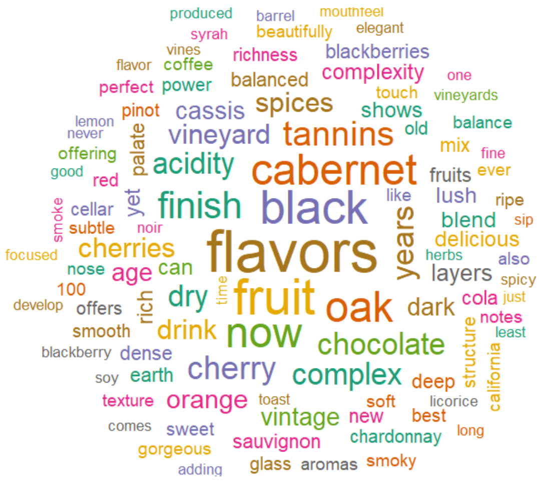
c) Relationship between Points and Price



In this graph we can see the relationship between Points and Price. The curve tells us that there is a slight positive relation between price and points. With increase in points of wine there is a very slight increase in price. Points from 80-91 has barely any relation with price but from points above 91 there is a slight increase. I have used a vertical line to show from where there is high relationship between the variables. I have also labelled the country name and variety of few outliers. From this we can infer that a costly wine is not always a good wine.

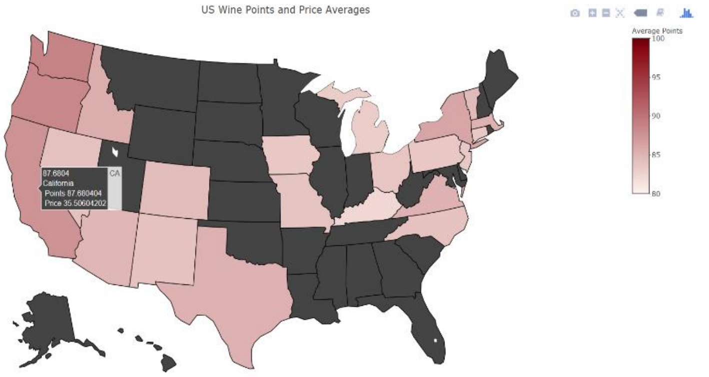
d) Wordcloud

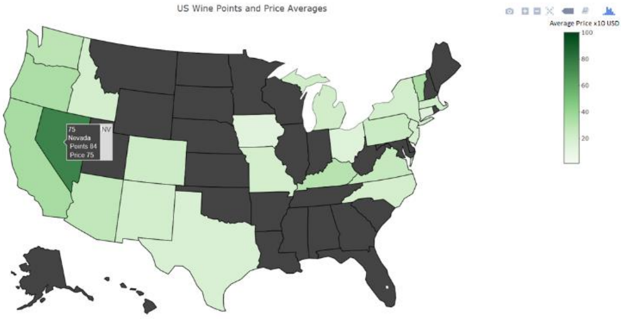
Wordcloud is the best graph for describing an item or event with a group of words. It is very easy to understand and also looks attractive. I have taken advantage of the description of wine. I have first tokenized the data, then filtered few usual words and passed this into wordcloud.



The words present in the above graph are best used to describe a good wine. We can just look for these words in the description of wine and tell whether a wine is good or not.

e) Plotly Choropleth Graphs

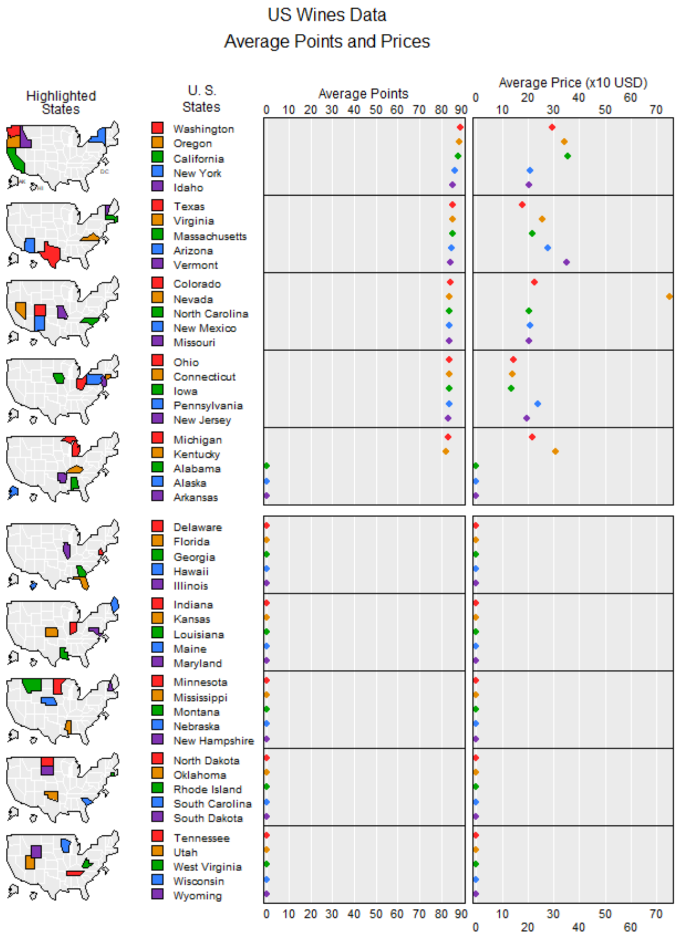




Plotly are interactive graphs mainly used in web pages. These graphs can be used to visualize geographical data. We can view the data of different regions just by hovering the mouse.

In this graph the darker the colour the higher the value. The legend shows us the sorting variable and also the scale. The first graph is sorted based on Points and second is based on Price.

f) Micromap



Micromap is another graph used to show data based on geographical locations. It is a very powerful graph which can help visualize various attributes based on geographical regions in one single graph. The colours and representation of regions is very helpful in easily understanding the graph.

From the Micromap and Plotly graphs we can observe that west coast states have higher average points than rest of the states and Nevada has the highest average price of more than $700.

Naïve Bayes Classifier Model

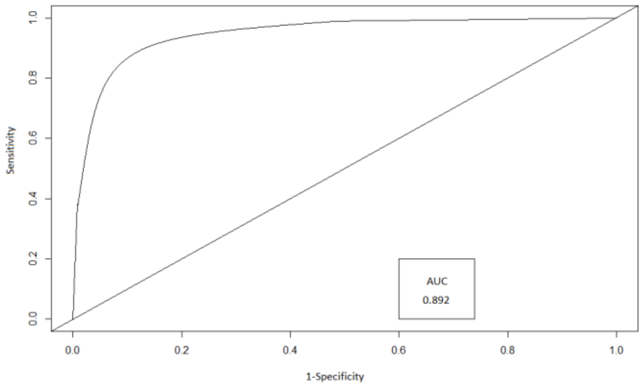
Naïve Bayes Classifier is one of the best and simple model to perform classification using text data. First, I have divided my data into training and testing data. Then classified my training data into premium and good wines if points >90 and points <90 respectively. After this I have tokenized and cleaned the data from description and passed the document term matrix to my model. Then I have used the testing data to predict the values using the model.

Confusion Matrix

	0	1
0	529	59
1	42	870

From the above confusion matrix, we can see that from out of 1500 rows our model predicted 1399 values correctly. Out of 571 good wines our model wrongly predicted only 42 wines as premium wines and out of 929 premium wines our model predicted 59 wines as good wines. This model has an Accuracy of 92.3%, Sensitivity of 0.936 and Specificity of 0.916. These values tell us that the model is very good.

ROC Curve



The ROC Curve with an area under curve value of 0.892 suggests that our model is very good for predicting good and premium wines.

Conclusion

We have seen that just by looking at some max and min values of points for different countries doesn't mean that a country produces good or bad wine, but by looking at averages we can have a clear idea. In most of the cases US stands at top but we can explain this because most of the dataset has wines from US. So, by doing a deeper analysis on US we now know the statistics of each of the states in US. From micromap and plotly graphs we can easily find that states on west coast produce better wines than other states. The Wordcloud and Naïve Bayes Classifier model help us in easily predict whether a wine is good or premium just by looking at description. Finally, the confusion matrix and ROC curve tells us that the model works very effectively.

References

- https://www.kaggle.com/wine-reviews
- http://www.inf.ed.ac.uk/teaching/courses/inf2b/learnnotes/inf2b-learn-note07-2up.pdf
- https://stackoverflow.com/questions/30386890/making-a-simple-wordcloud-in-r
- https://cran.r-project.org/web/packages/micromapST/micromapST.pdf
- https://www.youtube.com/watch?v=ypO1DPEKYFo
- https://plot.ly/r/choropleth-maps/