# **Wine Review NLP Analysis**

Interpretation of Sommelier’s critical reviews of wine from around the world have been collected into one datatable. The analysis of the reviews will focus on its ability to identify the variety of wine from the comments from the Sommelier’s critiques.

## **1. Data**

The data set is found in a downloadable csv or json file on the Kaggle website at the below link:

https://www.kaggle.com/zynicide/wine-reviews

The data consists of 14 columns and 129,971 reviews from wine critics, which includes the wine critique and supportive information such as the winery, variety, price, points(rating given by the sommelier), country, province, regions, critic name, critic’s twitter handle, and the label name of the wine.

There is an abundance of analysis regarding the distribution of the wines globally, provincial pricing and ratings, country pricing and ratings, varieties distribution globally, and much more. This analysis is focused mainly on the possibility of utilizing Natural Language Processing techniques to decipher what ‘variety’ of wine is described in the critic’s descriptions.

## **2. Data Cleaning**

There are approximately 130,000 records available for analysis. There are NaN’s in multiple columns such as the designation, price, region\_1, region\_2, taster\_twitter\_handle, taster\_name (the critic’s name), and one variety is not labeled.

Handling the NaNs is interesting. The missing variety is a critical bit of information that probably can’t be researched accurately to ascertain the proper entry, so this entry will be deleted.

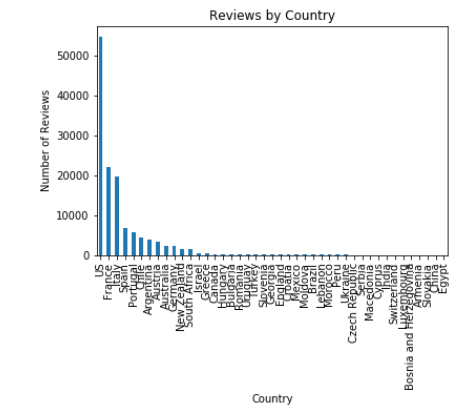
Checking for duplicates in the dataset was found that approximately 10,000 completely duplicate entries existed. We will de-duplicate and that will reduce the number of reviews to under 120,000.

The remaining NaNs are in complicated locations, such as price, rating, description which would be difficult to interpret or substitute a non-invasive value, so we will remove those items as well. This reduces the number of reviews to under 100,000.

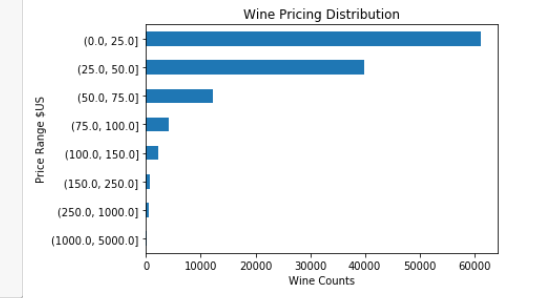
## **3. EDA**

Exploring the dataset provides an interesting array of general statistics such as 16,757 wineries from 43 different countries, in 425 different provinces, and 1,229 different regions from around the world are identified in this dataset. There are 707 different varieties of wines were reviewed in these 130,000 reviews.

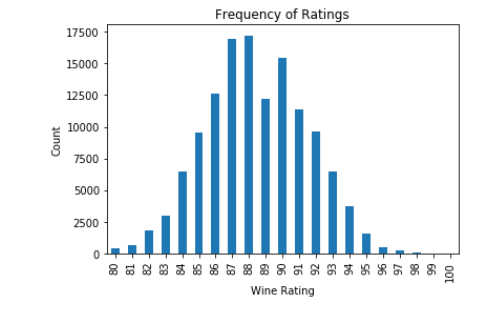
It is evident from the chart below that the US, France and Italy wines dominate the dataset.



There are 390 different prices for the wines reviewed. Approximately 99% of wines are prices less than $113 US. The price distribution is quite obvious in the bar chart below.



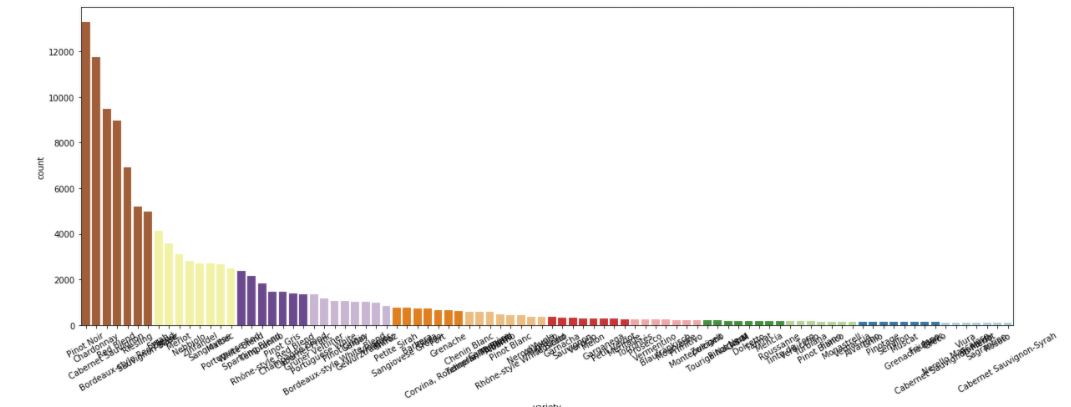
Ratings given by the Sommoliers ranged from 80 to 100. The distribution was normal, see chart below.



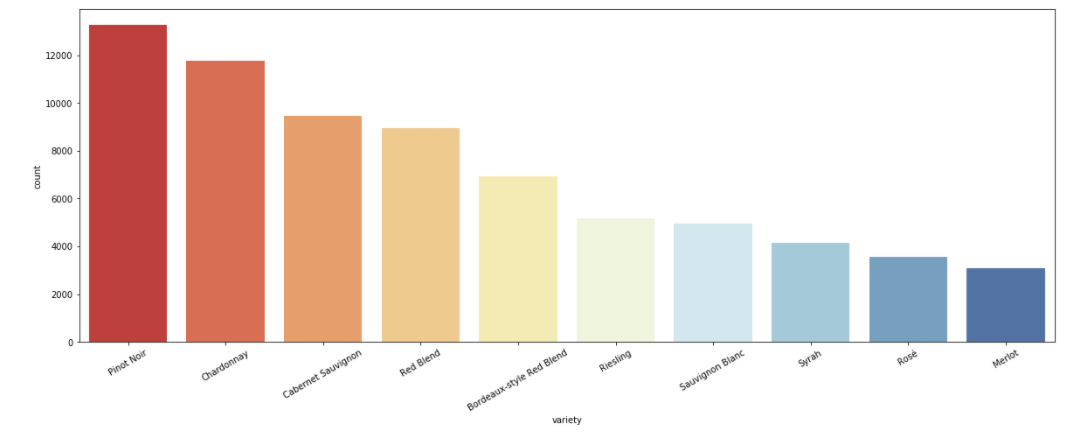
There is a slightly longer tail to 100 and no ratings below 80.

There is an extraordinarily high number of wine varieties evaluated. The 707 different varieties were considerably higher than expected. The distribution of the reviews over the 707 varieties is not very uniform, the top 17 varieties captured 72% of the reviews. These 17 varieties had 2100 or more reviews each, while the top 10 received more than 3000 reviews. There were 617 varieties with fewer than 100 reviews. It will probably make sense to evaluate how well the NLP processes can predict variety from the most populous reviews vs the less reviewed varieties.

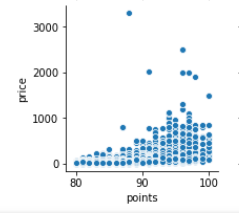
Below is a chart with all the varieties listed. There are a lot of entries with little information.



However, just reviewing the top 10 varieties shows strong number of reviews from around 3000 reviews for Merlot to nearly 13000 for Pinot Noir.



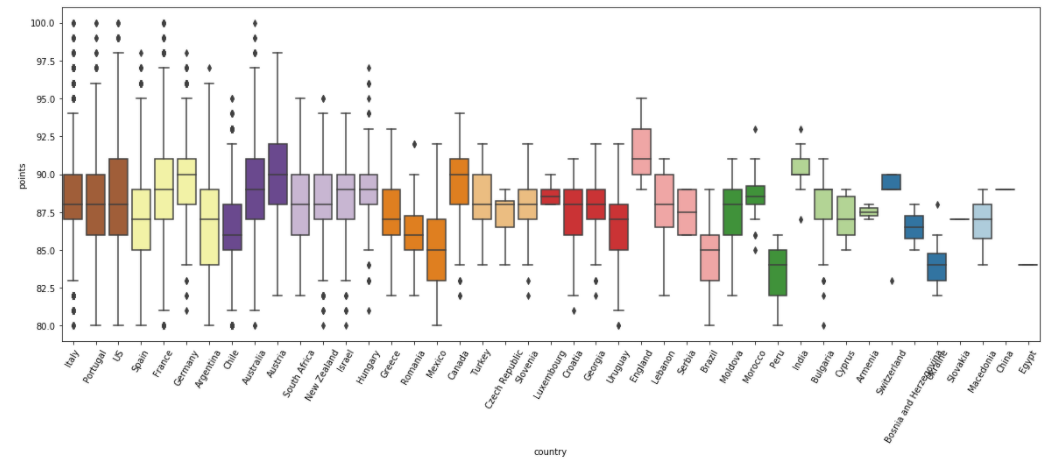
As expected, there is an increasing trend between the ratings and price.



Correlation between ratings and pricing is very strong. There is an evident trend to higher prices with higher ratings.. Obviously, the higher rated wines would naturally have more value, and the general trend is reflected in the above chart.

However, there are several outliers, which upon research, had found that most of the outlier pricing for middle tiered rated wines were probably typos. Unfortunately, I couldn’t confidently find the original pricing of the wines for the review time period, so, I’ll leave those outliers alone.

Ratings evaluated across countries was consistent with stronger outliers and broader range in the countries with more wineries.



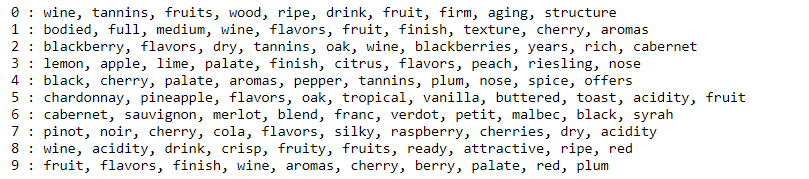
I’ll break the data into two groups for the model process. One with the top 10 varieties with the largest number of reviews, which is more than 3000 reviews and one with the top 90 varieties, ones with over 100 reviews each.

## **4. Algorithms & Modeling**

**Clusters:**

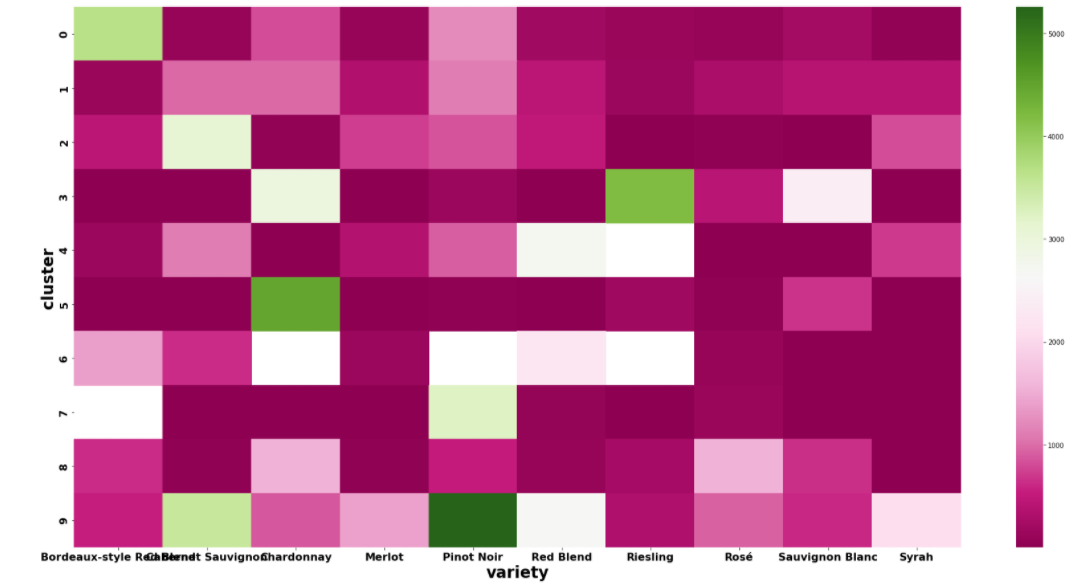
Kmeans modeling was based on two word counting algorithms. To be consistent with the split dataset, a KMeans test was conducted for the Top10 dataset with 10 clusters and the Top100 dataset with 90 clusters.

The resulting clustering of words for the 10 clusters is as follows:

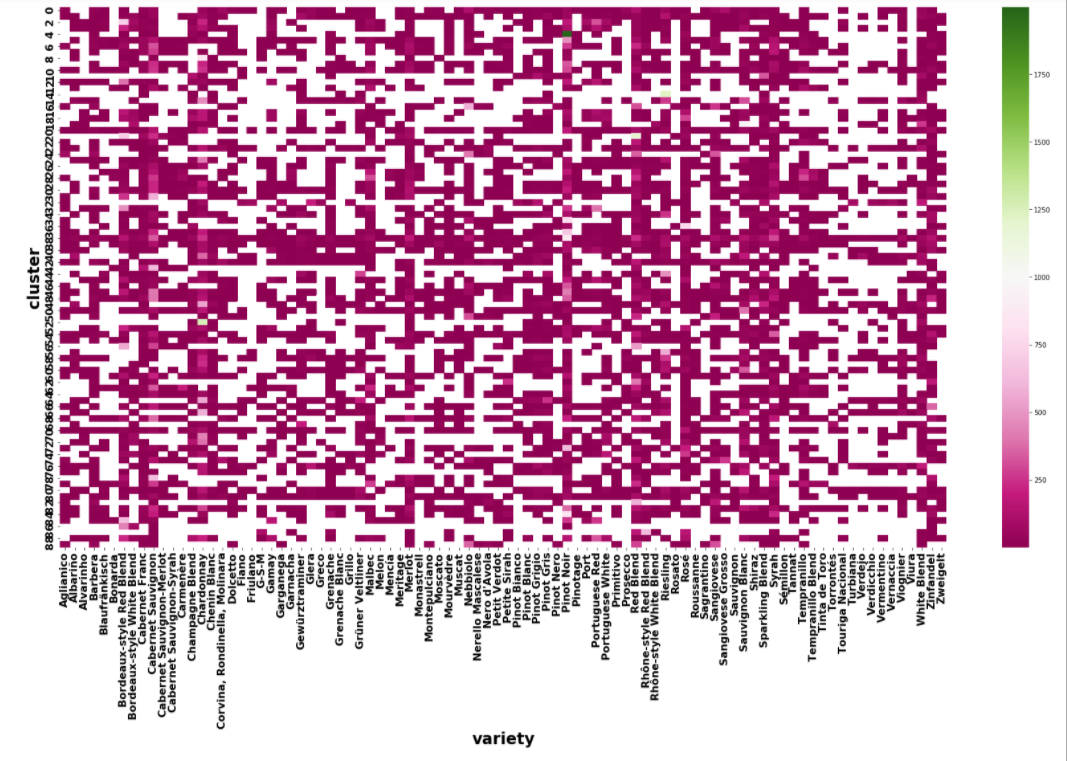


Utilizing the Pearson Correlation Coefficient, evaluating the top 10 correlation strength vs the top 90 correlations found minimal differences between the two groups. The R-Squared values for the two groups were 0.42 and 0.44.

The heatmap of the top 10 varieties is below. It is noticeable that there are some standout varieties that appear to be potentially identifiable through the clustered dataset.



Doing the same purview with the top 90 varieties the heatmap analysis is a lot more vague and not very many easily noticeable hotspots. There are very few combinations that exceed the white threshold in the middle of the scale. I’m sure expanding from 10 to 15 or 20 varieties may be feasible to identify the number of reviews that would be necessary to quantify any differentiation between the various wines in this set.



This 90 variety heatmap is not a good look. Very few higher rated combinations but there appears to be some trending patterns in the white areas vertically on the right 10 or so varieties, but the significance is probably minimal.

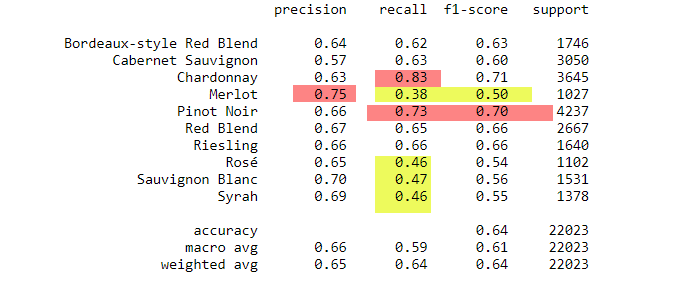
It would seem reasonable to test at a review threshold of 2000 reviews, which is the top 17 varieties, and then possibly a threshold of 1000 reviews which is the top 28 varieties.

Initially, the identification of several wines in the top 10 group could probably benefit from additional experimentation with the hperparameters, but we will save the tweaking for the regression or randomforest models

The heatmap and clustering is based on 1 word groups. Utilizing the feature engineering of CountVectorizer and TfidfVectorizer to groups of 2 and 3 words should help enhance the predictive qualities of other modeling methods.

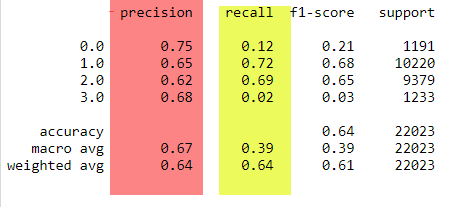
**Logistic Regression:**

Logistic Regression utilizing the CounterVectorize and TFIDFVectorized words provide an interesting display of precision and recall results for wine variety classification test.



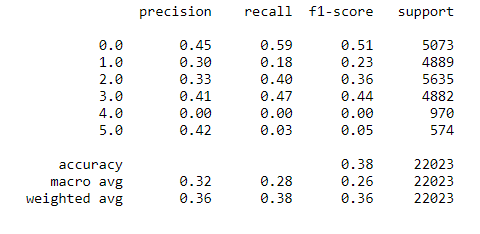
These results show a significant difference in prediction accuracy dependent over the ten varieties in this dataset. It should be noted that with the support level somewhere between 1500 and 1600 there is a noticeable stability between the precision, recall and f-1 scores. Those of the higher support level have more consistent levels versus the lower support levels that vary wildly from upper 30’s to mid 70’s in accuracy levels.

Using logistic regression techniques to predict ratings from the critics description of each wine proved even less reliable. Binning the ratings into 4 groups80,83,88,93,100,low (80 to 83), medium low (84-88), medium high (89-93) and high (94 - 100) seemed to be a decent breaking down of the rating numbers. Utilzing these four groups the resulting classification results were:



The precision was comparable with the wine variety outcomes, however, the recall was wildly off on the edges due to low support levels. It is another issue of not enough information to make adequate estimates for lowest and highest rated wines.

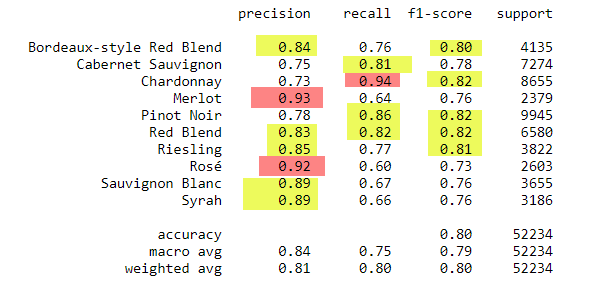
Utilizing the logistic regression to predict pricing from the description. The pricing was place in bins of 4-17,18-25, 26-42, 43-76, 77-117,118-3300. There appears to be little predictability in this process.



These results are less than interesting.

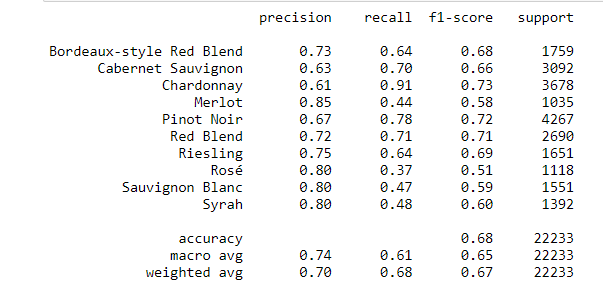
**Random Forest:**

The next step was using the CountVectorizer and TF IDF models in a Random Forest model with hopes of improving the predictability of wine variety, rating, and pricing. Using the Bag-of-Words model produced a nice improvement for wine variety predictions.



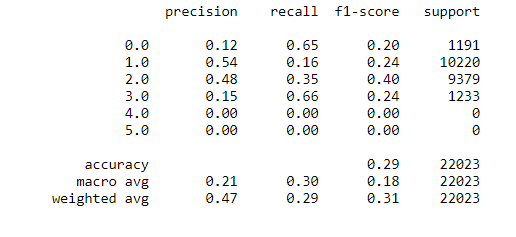
BAG OF WORDS - RF Results again showed that there are high precision and low support for Merlot and Rose’ wines, though their recall is 30 points lower. However, it is obvious with more support 4,000 plus provides a more stable precision and recall percentages generally less than 10 points. Though the 3,000 support level generally has a 20 point differential between recall and precision.

Using the TF\_IDF model within a Random Forest model produced the following table of results.



The overall f1-scores were worse for all the top 10 wine varieties and produced wildly different precision and recall percentages up to 40 points difference with numerous 30+.

Using the TF\_IDF model in RandomForest for predicting the rating was rather dismal as seen in the chart below.



Utilizing a simpler Decision Tree Classification on the Rating value was found to have an accuracy of 24%. Using the Decision Tree model on the TF\_IDF data to estimate prices was slightly worse at 15%.

## **5. Predictions**

There appears to be a few stand out varieties of wine that provide a unique basis for identification. Predicting price and ratings seems to be a bit more problematic, however, I believe including the winery, country, and region information may provide additional features to base these predictions on instead of pure Natural Language Processing methods on the critics description of each wine.

## **6. Future Improvements**

The wine reviews dataset is large with 130,000 reviews, minus the duplicated entries. There are a myriad of additional evaluations that could be possible. The three topics of interest was the predicting the wine variety by the critics wine description. Utilizing that description can the rating and price be estimated as well?

Focusing on the wine variety, there are over 700 varieties of wine in the database. Most varieties only have less than 100 reviews. Improvement of the analytics results would involve using different models, such as XGBoost, Gradient Boosting, or Deep Learning techniques.

Adding back additional features might improve the current results. Specifically, the winery and country features would seem to make an impact on boosting predictability of ratings and pricing. A PCA analysis should help delineate which of the remaining features would be most beneficial.

Working through the word salad created by the bag-of-words and TF\_IDF methods to improve spelling and tenses might provide an increase in performance, though expanding the word groupings from 2 words to 3 provided zero improvement at the Random Forest stats.

## 