An Educational look at wine

By: Douglas Malfacini

Overview

- Introduction
- Data Wrangling
- Exploratory Analysis
- Statistical Analysis
- Varietal classification
- Results
- Potential Improvements

Introduction

Wine is an incredibly complex product. From which grapes are used, and where it is grown, to the producing process, there are many steps which can influence the taste of the wine.

In this presentation, we are going to explore the following questions:

- Which countries produce the highest quality wine?
- How does one describe a good quality Pinot Noir? What about Chardonnay?
    - What notes are typically found in different wine varietals?

Data Wrangling

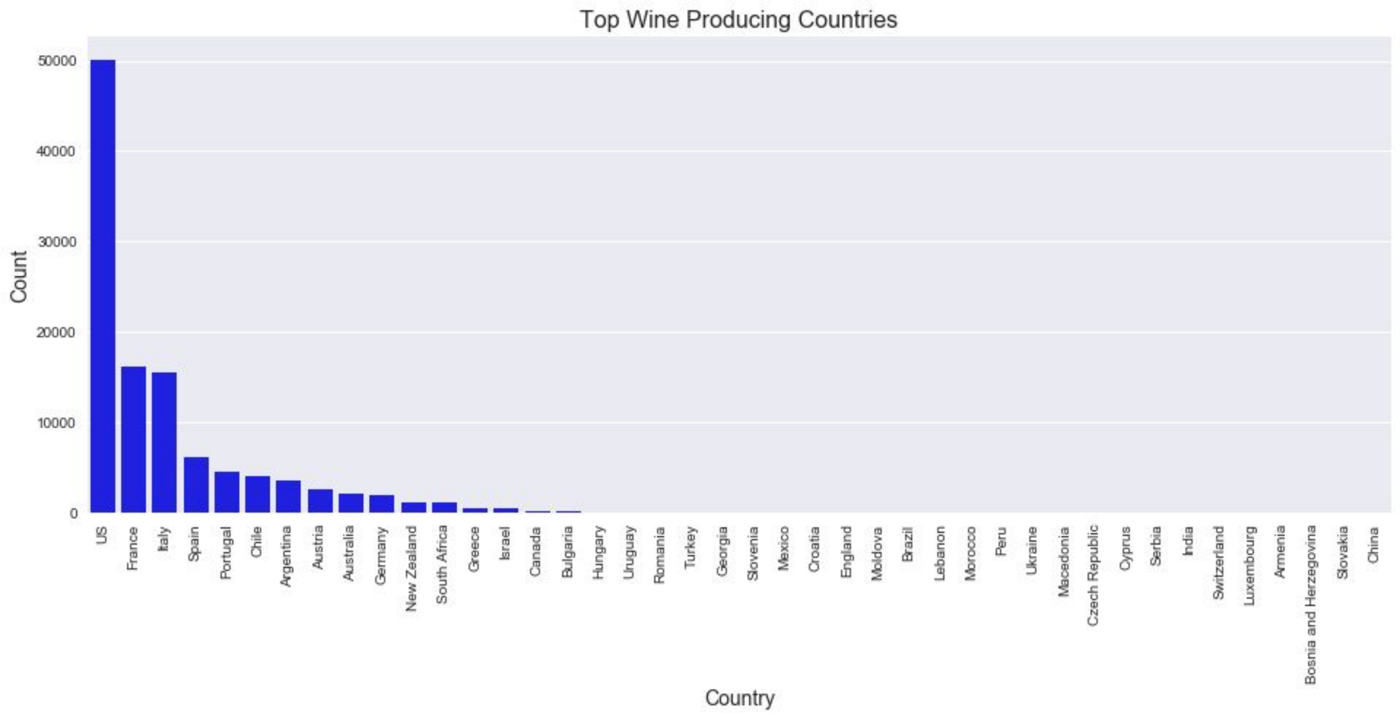
Data Wrangling

But first, we need to get our dataset ready for analysis:

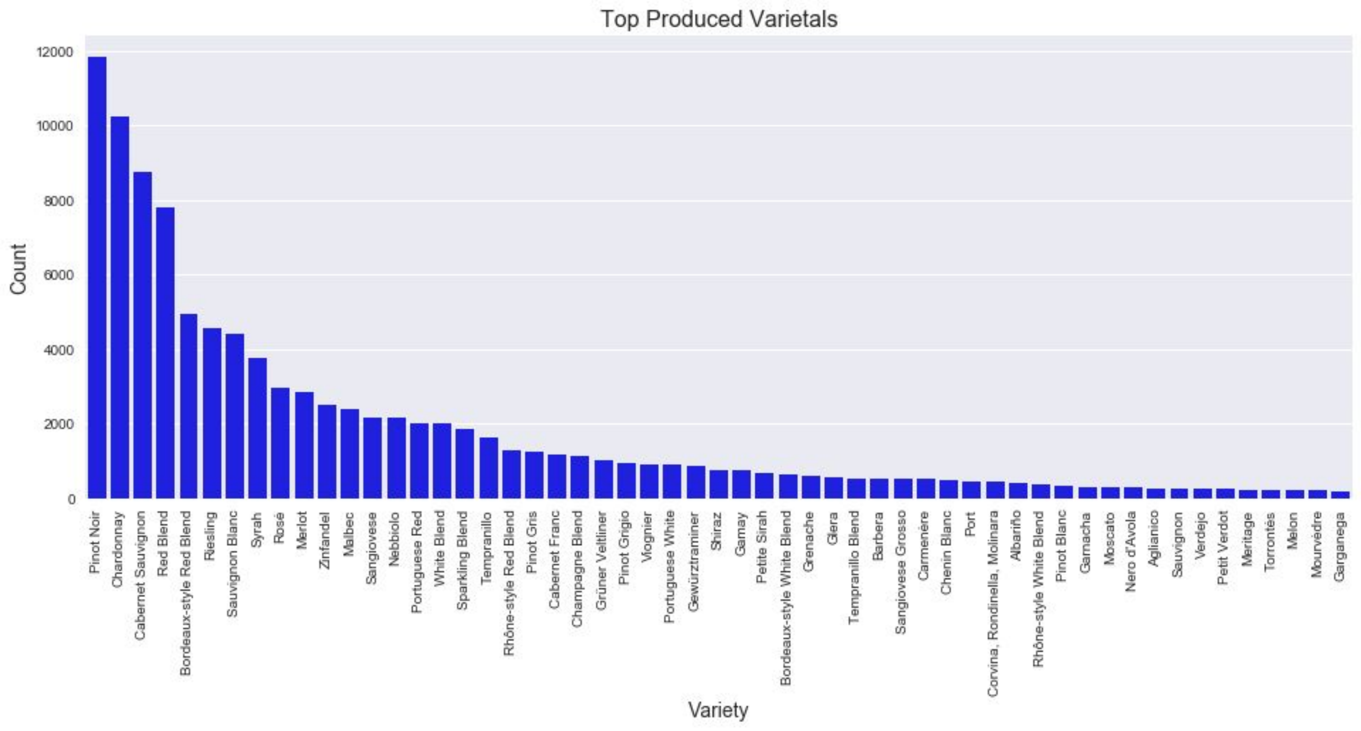
- We’re only going to look at wines varietals that have over 200 observations
- We’ve have filtered out blends, as that will complicate the model
- In our statistical analysis, we’re only comparing the top 5 wine producing countries. They encompass 82.88% of the dataset.

Exploratory Analysis

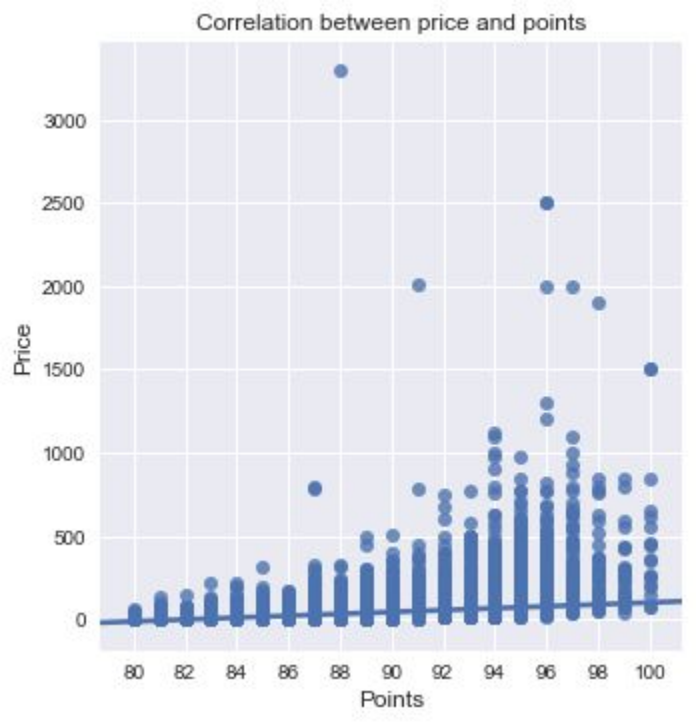
Exploratory Analysis



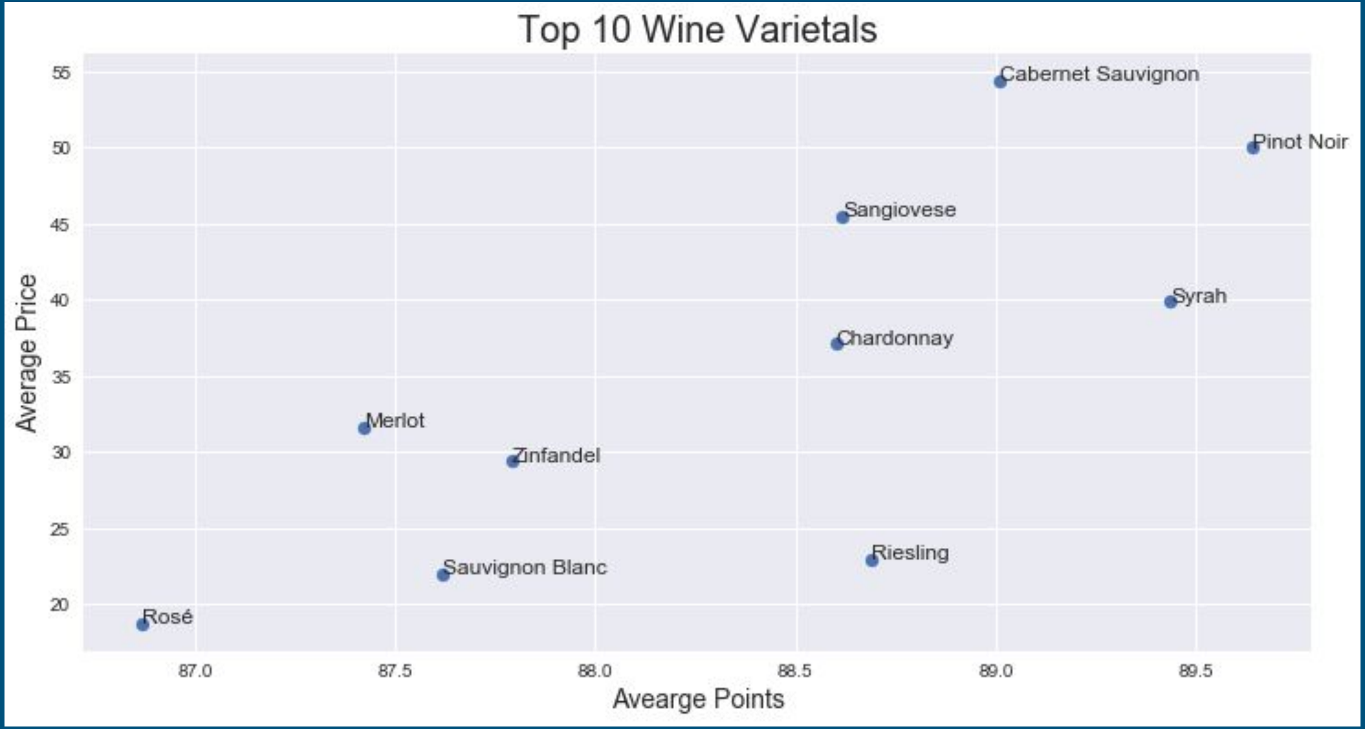
Exploratory Analysis



Exploratory Analysis



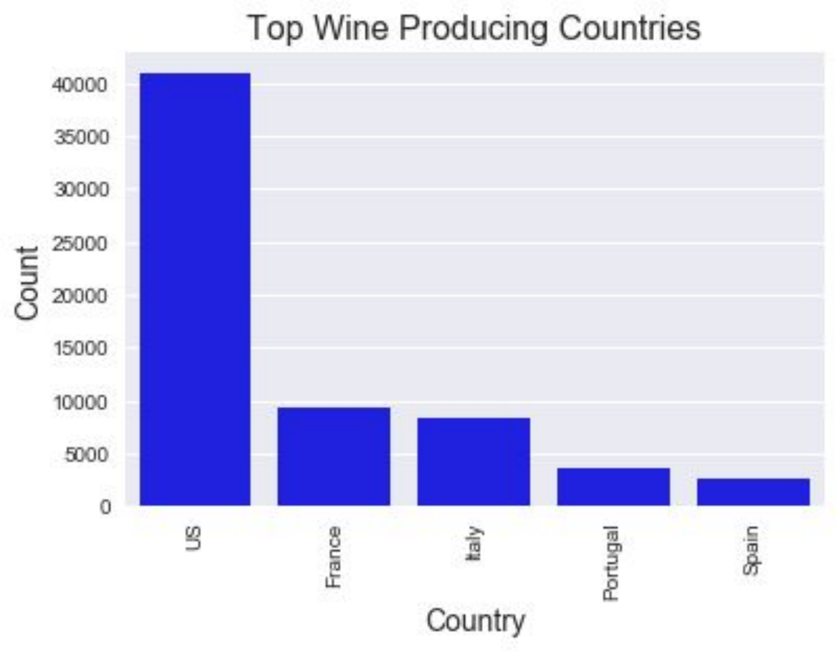
Exploratory Analysis



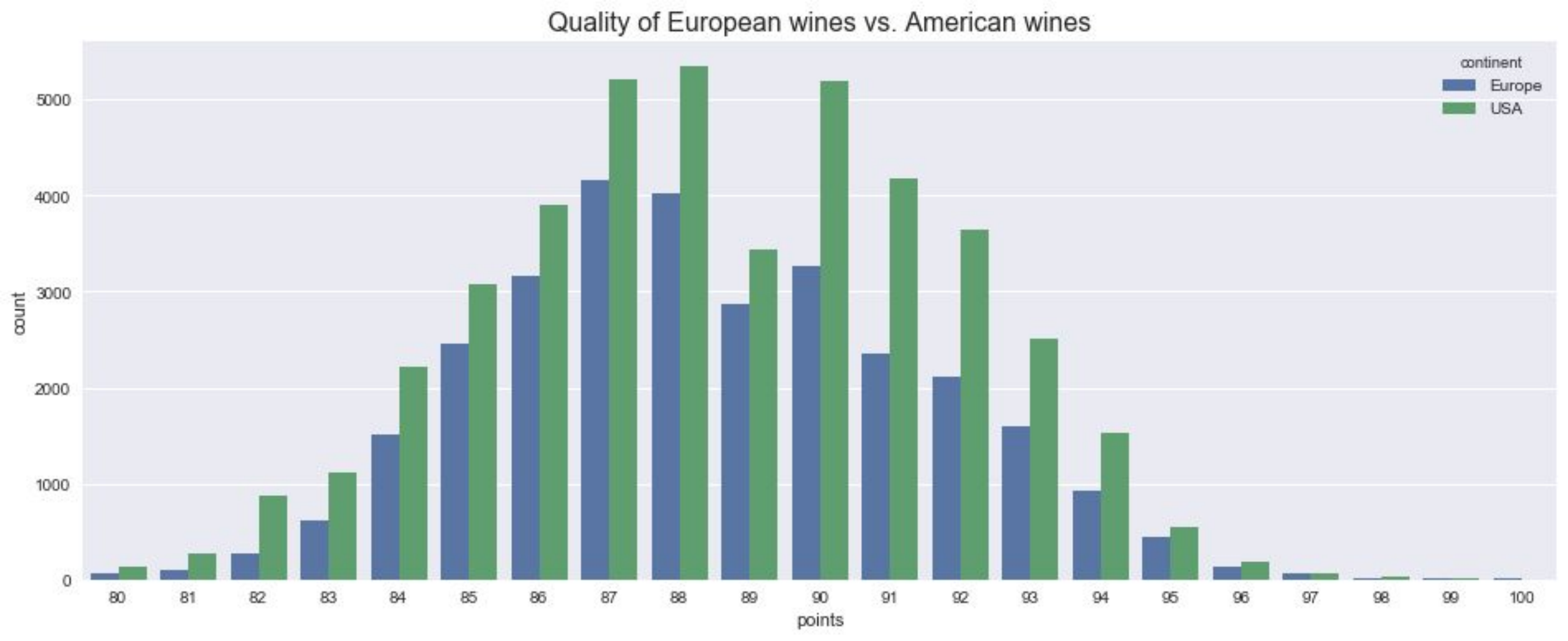
Statistical Analysis

Statistical Analysis

- The top 5 wine producing nations in the world is the United States along with a group of of European nations.
- By aggregating the European nations, we can compare the United States vs. Europe to see who has the highest quality of wines.



Statistical Analysis

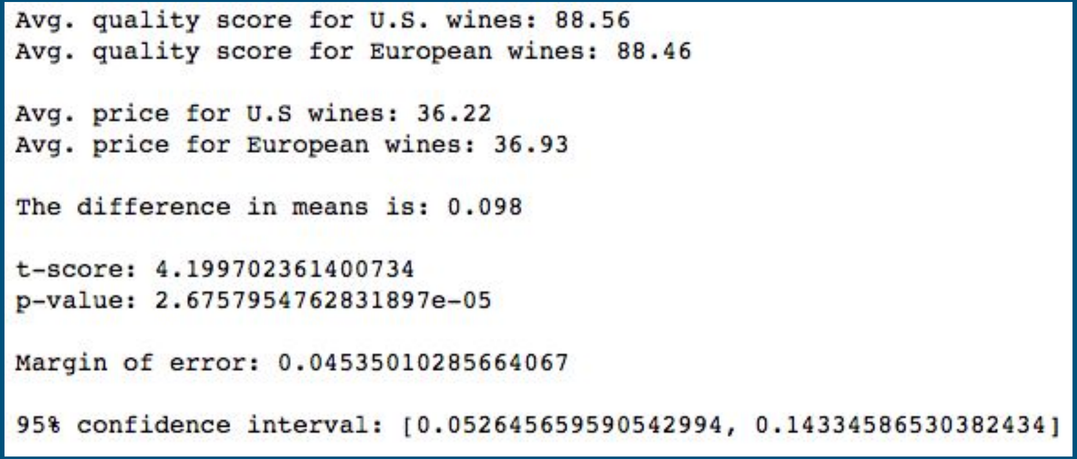


Statistical Analysis

- From an absolute sense, the United States produces a greater quantity of high quality wine.
- This is very likely due to the greater quantity or wine produced, good or bad.
- To be sure who makes higher quality wine, let’s test the null hypothesis that there is no difference between the quality of wines between the United States and Europe (in this study consisting of France, Italy, Spain, and Portugal).

Statistical Analysis

- Given the results below, and the low P-value, we can safely reject the null hypothesis and conclude that on average, the United States makes higher quality wine when compared to Europe.
- From a practical sense however, the wines are so close it may take a Sommelier to differentiate.



```
Avg. quality score for U.S. wines: 88.56
Avg. quality score for European wines: 88.46

Avg. price for U.S wines: 36.22
Avg. price for European wines: 36.93

The difference in means is: 0.098

t-score: 4.199702361400734
p-value: 2.6757954762831897e-05

Margin of error: 0.04535010285664067

95% confidence interval: [0.052645659590542994, 0.14334586530382434]
```

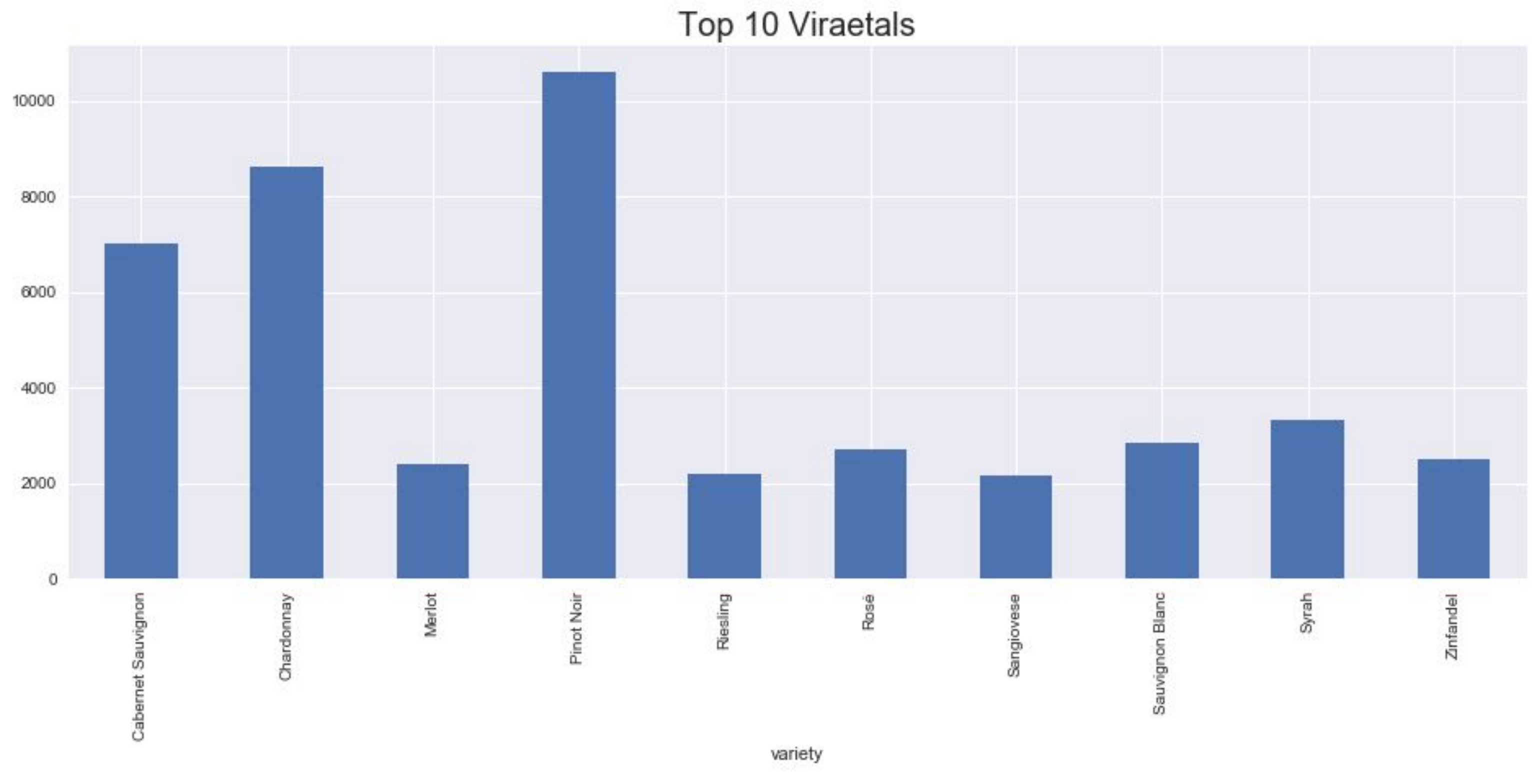
Varietal Classification

Varietal Classification

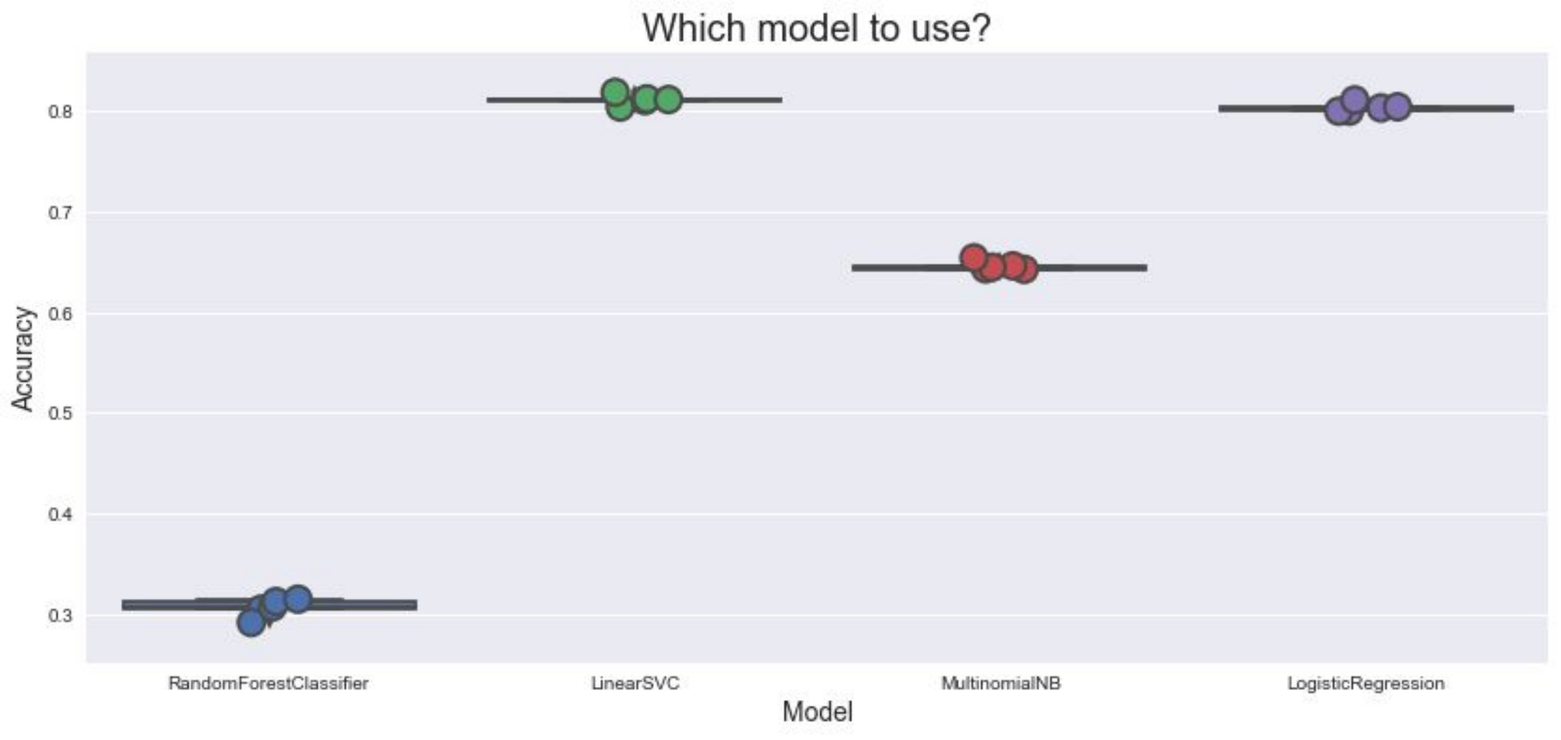
Facts about the model:

- We started with 44,393 observations and 10 classifiers
- The classifiers are skewed, with Pinot Noir being the largest varietal with 23.90%. We'll use this as the baseline for the model.
- After running bag of words model on the text, we ended up with 39,153 features.
- Various steps were taking to clean it up and we ended up with 37,269 features for the final model

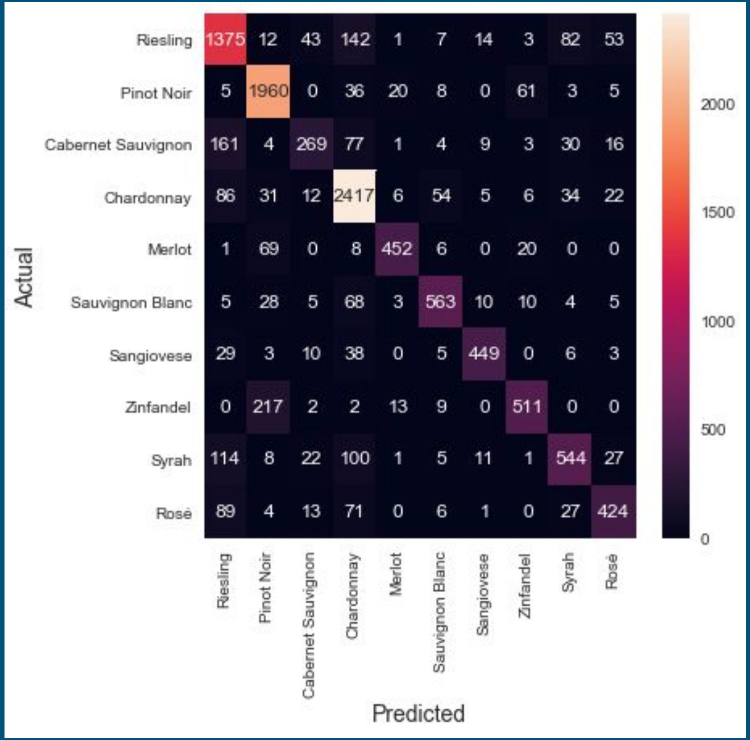
Varietal Classification



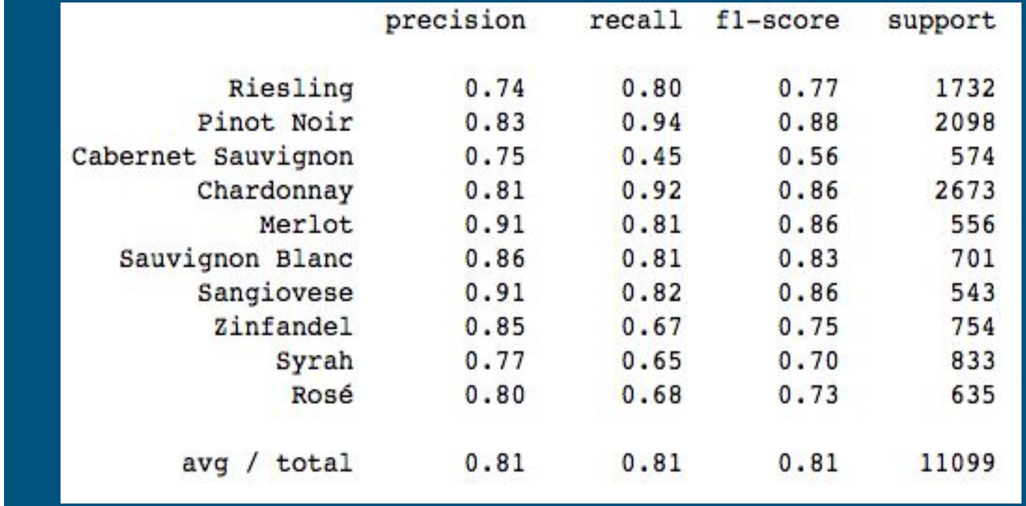
Varietal Classification



Varietal Classification



	Riesling	Pinot Noir	Cabernet Sauvignon	Chardonnay	Merlot	Sauvignon Blanc	Sangiovese	Zinfandel	Syrah	Rosé
Riesling	1375	12	43	142	1	7	14	3	82	53
Pinot Noir	5	1960	0	36	20	8	0	61	3	5
Cabernet Sauvignon	161	4	269	77	1	4	9	3	30	16
Chardonnay	86	31	12	2417	6	54	5	6	34	22
Merlot	1	69	0	8	452	6	0	20	0	0
Sauvignon Blanc	5	28	5	68	3	563	10	10	4	5
Sangiovese	29	3	10	38	0	5	449	0	6	3
Zinfandel	0	217	2	2	13	9	0	511	0	0
Syrah	114	8	22	100	1	5	11	1	544	27
Rosé	89	4	13	71	0	6	1	0	27	424



```
                    precision    recall  f1-score   support

          Riesling       0.74      0.80      0.77      1732
        Pinot Noir       0.83      0.94      0.88      2098
Cabernet Sauvignon       0.75      0.45      0.56       574
        Chardonnay       0.81      0.92      0.86      2673
            Merlot       0.91      0.81      0.86       556
   Sauvignon Blanc       0.86      0.81      0.83       701
        Sangiovese       0.91      0.82      0.86       543
         Zinfandel       0.85      0.67      0.75       754
             Syrah       0.77      0.65      0.70       833
              Rosé       0.80      0.68      0.73       635

       avg / total       0.81      0.81      0.81     11099
```

Varietal Classification

- This is a high accuracy model with all the different steps that were taken to improve it.
- This is also a great improvement over our baseline, as we went from 23.90% to 81% accuracy.

Potential Improvements

- Take a closer look at the classifiers and see if any consolidation is possible
    - For example: “Zinfandel” varietal is called “Primitivo” in Italy.
- We can look into splitting the dataset between red wine, white wine, and champagne and check for improvements in accuracy.
- Re-run the model with blends included and fine tune the results.
- We may add the wines varietal themselves into stopwords as different varietals may be used as comparison.

Final Thoughts

Next time you’re out wine tasting, you can look for these notes in your wines!

- Pinot Noir - Apple, butter, tropical fruits
- Chardonnay - Cherry, Cola, Pomegranate, Cranberry Fruit
- Cabernet Sauvignon - Plum, dark, raspberry, tannin

Thank you and Cheers!