Which regions produces the best quality wine and what are the notes that describes them?

1. Introduction

Living in the San Francisco Bay Area, we have great proximity to the Napa Valley and Sonoma County, two of the best wine producing regions in the world. With that in mind, I got interested in which wine regions produces the best quality wines in the world.

In addition, wine is a very complex product with many different notes (flavors and aromas). For example, wines can be described as fruity, earthy, chocality, having a subtle or strong finish, or plain, among many other notes. With that in mind, I set out to see how well a computer could correctly classify a wine based on the description of the wine.

This analysis should be interested for an everyday consumer who is interested in trying and purchasing the best wines, as well as becoming more educated about what goes into each varietal.

2. Data Wrangling

After reading the .csv file into python and calling .head() to get a sense for the columns and values of the dataset. My initial observation from this step include:

- Most of the columns have string objects, with the exception of the “points” and “price” columns. These have numerical values.
- The column names are well written. Everything is lowercase and columns with multiple letters are joined by an underscore, and not a space. Which is great since it’ll avoid issues in the future.
- The dataset is organized in tidy format, having one subject per column.

My next step was to look at “.info()” command to verify each columns dtypes, as well if the data has any missing values.

- Turns out the dtypes are correct. Including the “points” and “price” columns, which have int64 and float64 respectively.
- There are a lot of missing values from almost every column in this dataframe. The only columns that do not contain missing values are “description”, “points”, “title”, and “winery”.

From there, since I’m interested in comparing wine quality between countries, I tackled the missing values in country column. Given that the “winery” field has all of its values, created a

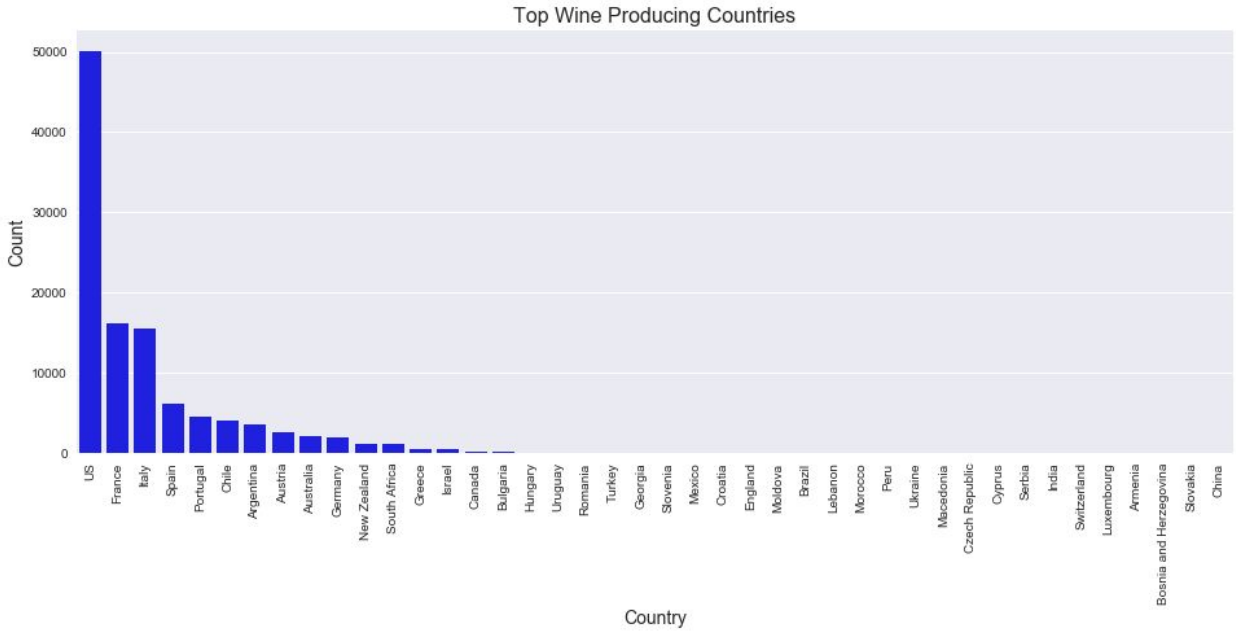
map of {winery:country} value pairs and mapped the missing country values back into the dataframe. This approach worked for 32 missing values. There are still 31 missing country values, out of over 100k. Since 31 is such a small proportion of the dataset, I dropped those rows.

Next I dropped missing values from the dataframe including duplicates, missing values from the price column and variety columns. From here, the dataset is ready for exploration.

3. Exploratory Data Analysis (EDA)

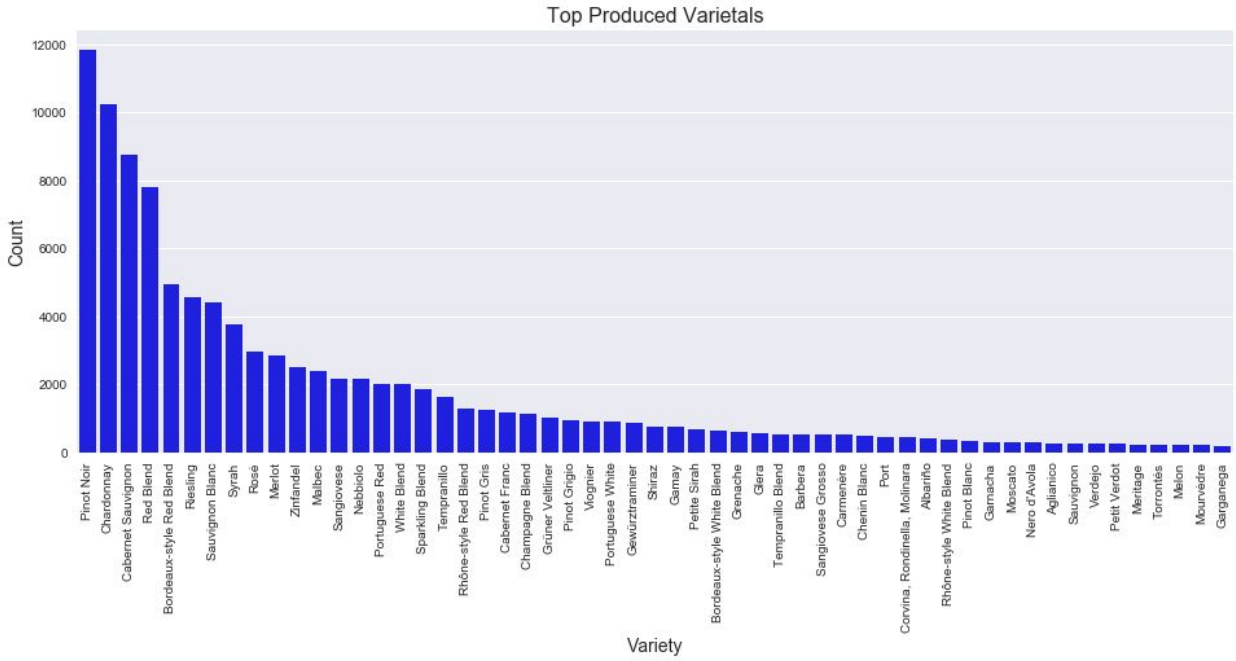
After cleaning the dataset, I wanted to know more about the dataset. The questions and results from my findings and be seen below:

- **Which nations produce the most wine in the world?**

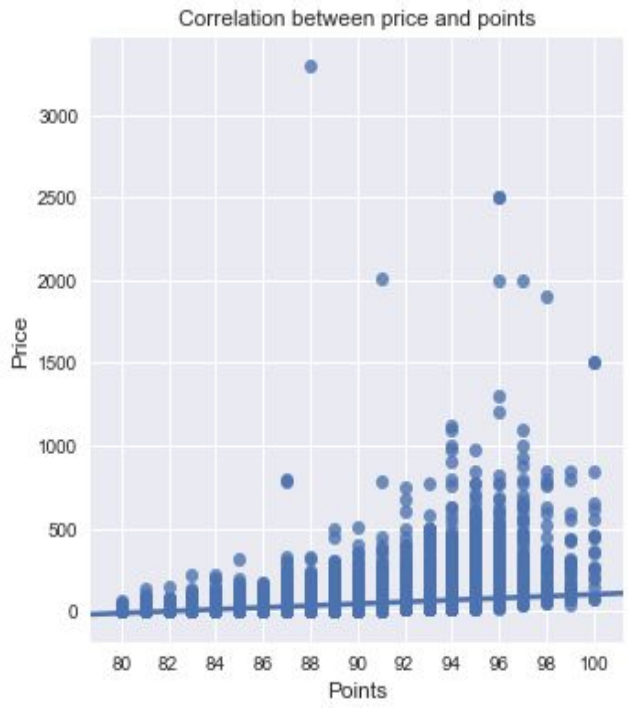


- **How many different varieties are of wine are included in the dataset?**
    - There are 694 different wine varietals in this dataset, including blends and champagne. In order to have a comprehensible graph, I limited the graph to varietals with at least 200 observations in the dataset.

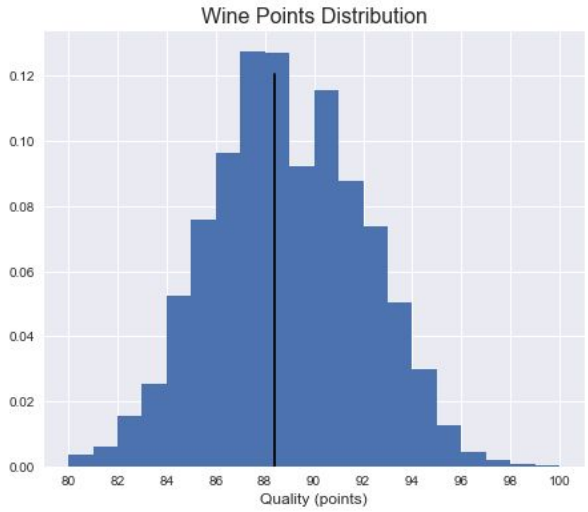
- **What are the top varietals?**

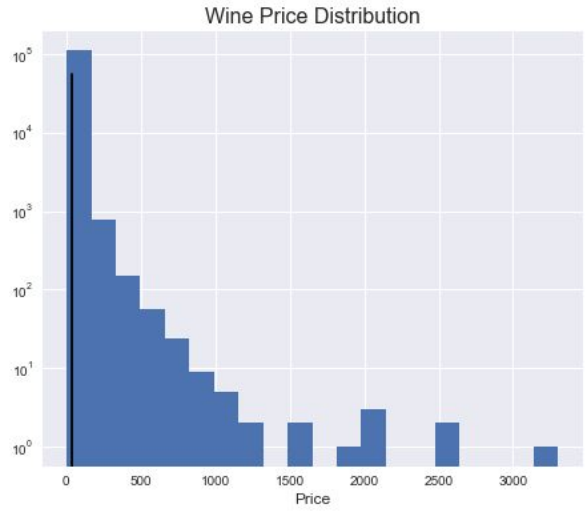


- **What is the relationship between wine price and points (quality)?**



- **What type of distribution do points and price have?**
    - The black vertical bars in the graph below are the averages.

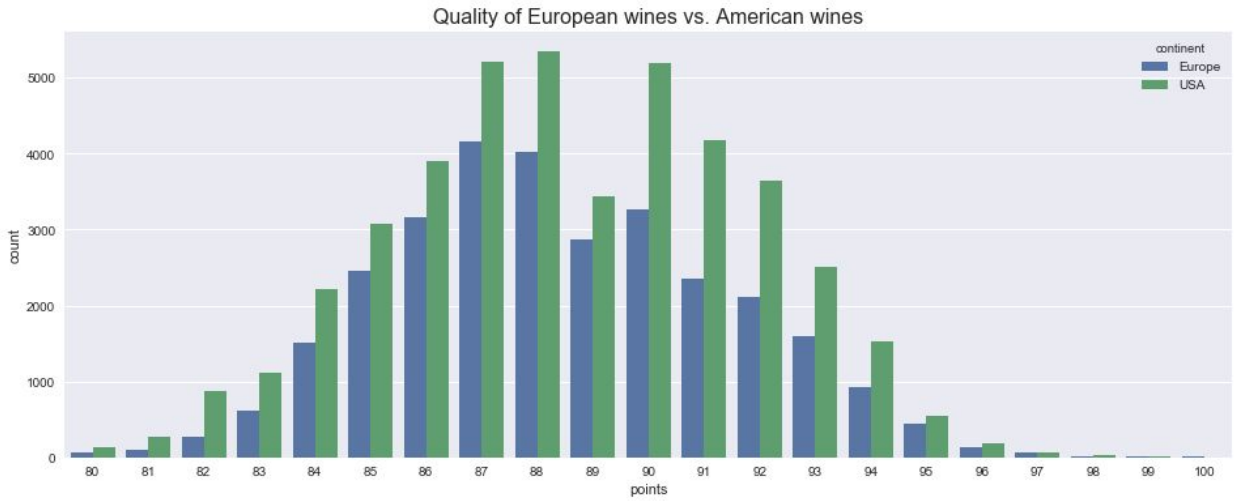




4. Statistics

Given all the findings above, I set out to see if there's a significant difference in wine quality (points) between the United States and France, Italy, Spain, and Portugal, the next four wine producing countries. From here on out our refer to those four European nations simply as Europe.

I put together an initial graph to visualize the data:



You can see that in absolute terms, the United States produces a greater quality of wine when compared to Europe. This, however, is likely because the United States has more wine compared when compared to Europe.

In order to avoid this issue, I compared means with the null hypothesis that there is no difference between the quality of wines from the United States versus the wines from Europe.

As can be seen from the calculations and low P-value shown below, we can reject the null hypothesis and confirm that on average, wines from the United States have greater quality than the wines from Europe.

5. Machine Learning

The goal in the machine learning section of the project is to train the model to identify wine varietals based on the wines description.

To start, I needed to pare down the amount of data I have. I did so in the following ways:

- There are 694 different varietals in the dataset, with many making a single entry. I filtered the data to only include the top 10 varietals, simplifying the dataset
- In addition, I also filtered out blends from the varietals as those could easily confuse the model and undermine accuracy

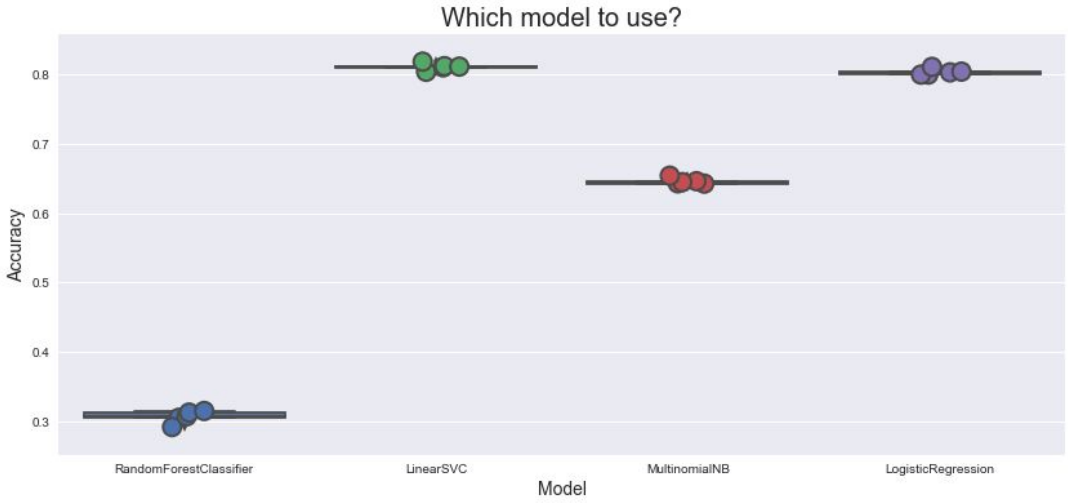
Next, I needed to establish a baseline for the model. Since I have a skewed classification, I took the most frequent varietal (Pinot Noir) and divided it by the whole dataset. This gives me a 23.90% baseline.

From there, I used various different steps to clean the data to get it into a format that'll make improve the likelihood of accuracy in the model. These steps include:

- Removing HTML related tags from the corpus.
- Removing special characters
- Lemmatizing the corpus
- Removing extra whitespace
- Removing numbers
- Removing stopwords

These steps reduced my features from 39,153 to 37,269.

In order to decide which model would give me the best accuracy, I run a for loop with the following four models, which gave me the results below:

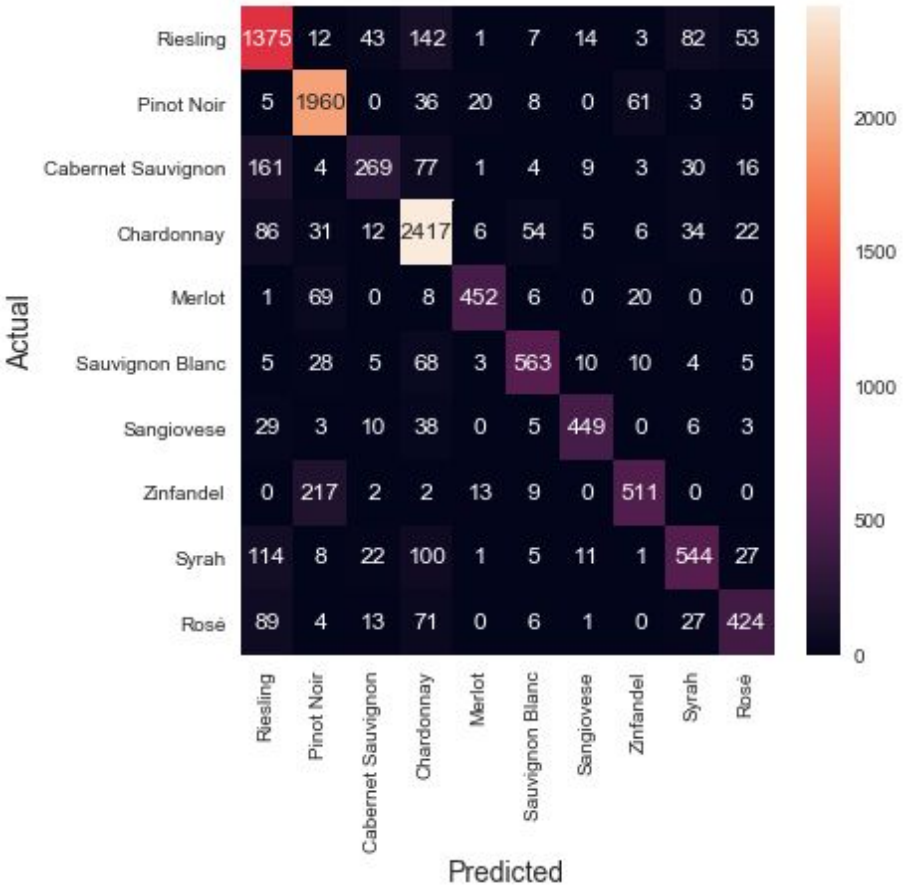



```
model_name
LinearSVC                0.810984
LogisticRegression       0.803326
MultinomialNB            0.645959
RandomForestClassifier   0.306221
Name: accuracy, dtype: float64
```

Give that the LinearSVC and LogisticRegression are so close, I reran the test with both models but not before tuning the hyperparameters for the set. LinearSVC still came out as the best performing model for this dataset, with an accuracy of 0.8151.

```
model_name
LinearSVC              0.815152
LogisticRegression     0.803326
Name: accuracy, dtype: float64
```

From here I ran the model using LinearSVC to get the following results:



	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