Exploring Wine through Machine Learning

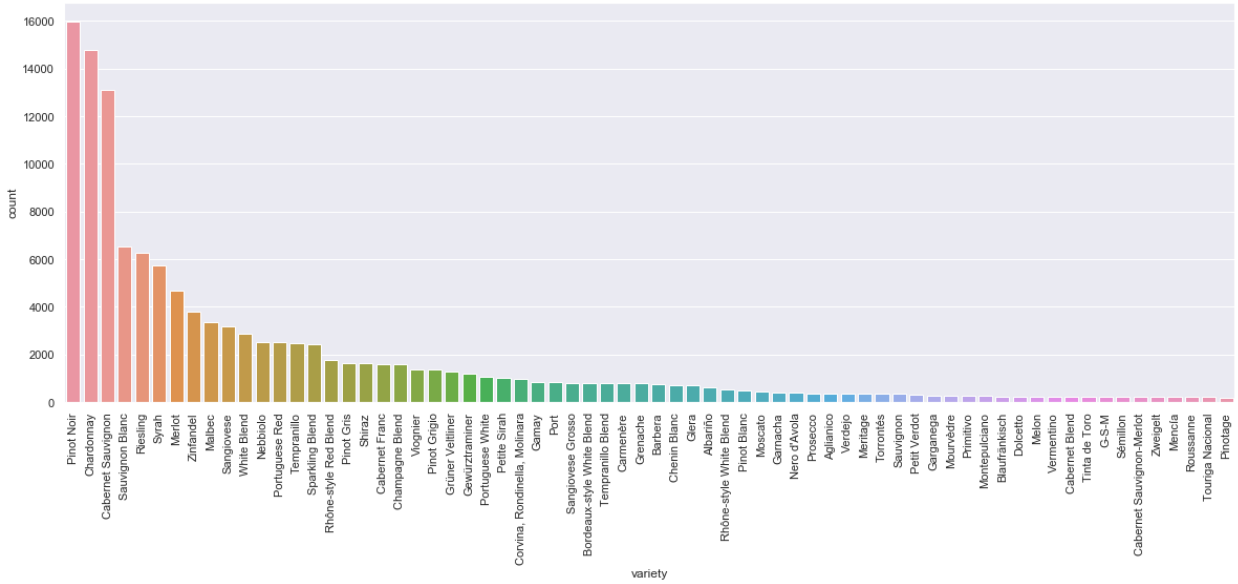
By: Nathan Duffy

Introduction:

For this project I wanted to explore an extensive dataset on Kaggle that had scrapped reviews of wine from the internet. It included features such as: country, description, designation, points, price, province, region_1, region_2, variety, winery, taster_name, taster_twitter_handle, and title. I was particularly interested in the features of variety and description. Variety is the type of grape used in the wine, essentially the type. The description refers to the short spiel that the sommelier gave the wine. The plan for this data was to use several machine learning methods for prediction and exploration. First a logistic regression to predict variety based on description. Second, a random forest just to see if it would have a better accuracy score that the logistic regression when predicting variety. Third, K-Means clustering to group the wines into clusters that attempt to generally describe each unique variety we are exploring.

Data Cleaning:

So there were actually two different CSV files for this dataset because it was too large for just one. My first step in the cleaning process was to merge these two files; after the merge our new dataset was 280,900 lines long. A substantive amount of information that will be too much for some machine learning models to handle, so we will reduce the size by removing unneeded information. First, we dropped any instances of duplicate descriptions. Second, we remove the instances where the price data is incomplete. Third, we removed the generic blends that are prominent in the dataset. After doing these three things, our data downsized to 136,033 lines. Next, I did some exploratory analysis on the dataset. To start, I looked at the distribution of varieties, and through this we find that there are 743 different values in the dataset.



The chart above accounts for the varieties that have more than 200 instances. It gives us a good visual as to which types are showing up most in the data. It will be extremely difficult to train a classifier to predict all of the varieties. Because of this, I will have to simplify the amount of varieties. Going forward I plan on only using the varieties that have more than 3000 instances (or the top ten) to give the model a more realistic chance of predicting the correct label. After filtering out the less prominent varieties our data downsized further to 77,509 lines; a realistic amount of data to use for our machine learning models.

Machine Learning Models:

- Logistic Regression
- Random Forest
- K-Means Clustering

Logistic Regression:

First step in this process was to drop the unneeded features (columns), and reduce the dimensionality of our data to perform a logistic regression. We will only be using description to predict the variety of wine. The accuracy score was surprisingly high after the initial test. The model predicted the correct variety of wine based on description with 80% accuracy. After some thinking and a deep dive into the data, I realized that the description actually has the variety of wine in it sometimes. This takes away from the purpose of the model, so before training I must remove these indicator words and re-run the model updating the TfidfVectorizer() function with our stop_words.

Early in the process, I experimented with having two features predicting variety: price and description. However, with this new feature added we only increased the accuracy of the model by 1%. Therefore it was determined to have no significance in determining variety, and was removed from the model. Another important step in this process was to refine the description column. Right now our model is using the full feature space with all stop words, indicators, and punctuation. Ideally these three types of tokens should be removed, which will reduce dimensionality further and allow the model to give extra weight to more important tokens.

After refining the features, and some trial and error, the model was finally ready for training. When I ran the model I got 77% training accuracy and 72% testing accuracy. A great initial score! Next step was to do error analysis by looking at the confusion matrix. We have the most data on Chardonnay, 14795 instances in the original data frame. This makes it easier for our classifier to learn and make more accurate descriptions with the extra information. Our logistic regression has a 91% chance of

predicting Chardonnay correctly. However, this model is atrociously bad at predicting Merlot and Malbec. Merlot is being classified as Cabernet Sauvignon 46% of the time, while Malbec does the same 23% of the time. That is significant. This makes sense however if we think about how similar these three wines are in taste and look.

Random Forest:

After performing the logistic regression, I wanted to compare the model accuracy to that of a random forest. After tuning the hyperparameters for this model to achieve optimal accuracy we ended with the following values:

- n_estimators = 50
- max_depth = 50
- min_samples_leaf = 5

With these hyperparameters, I trained the model and it outputted an accuracy score of 61%. Which is 10% percent lower than the logistic regression.

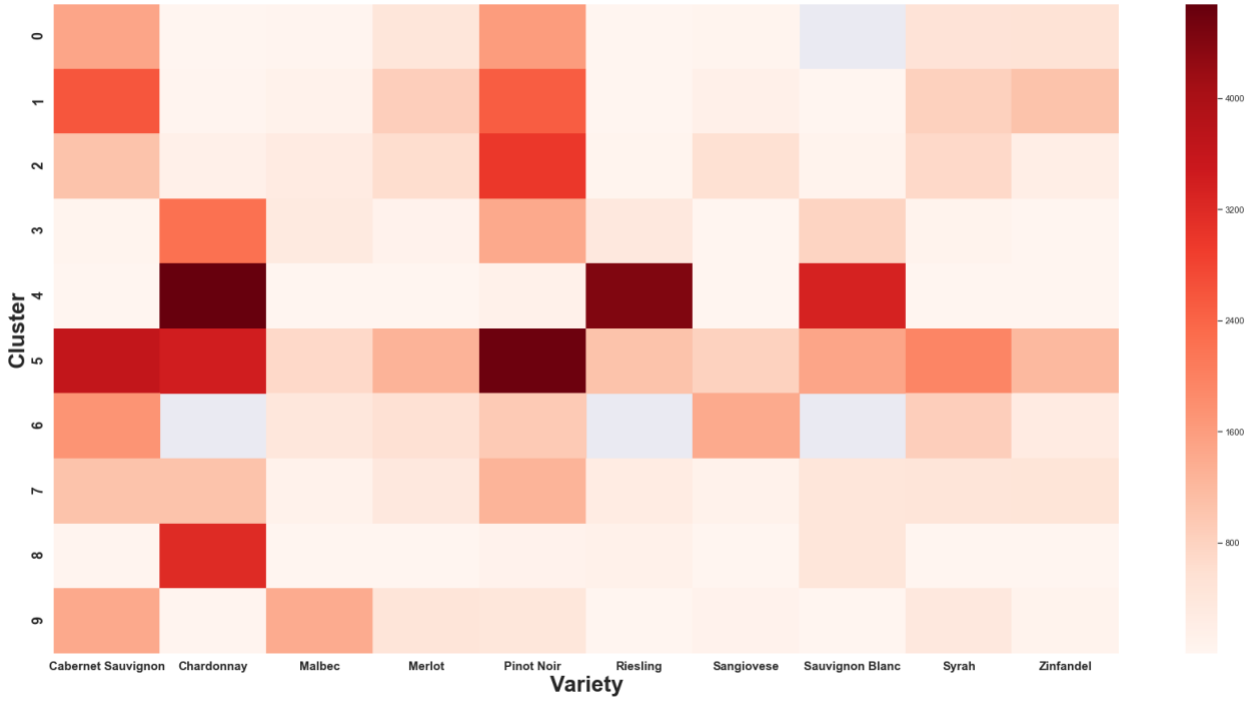
K-Means Clustering:

The idea for this model was to see if we can understand how these 10 wines are being generally described. After training the model and determining the clusters. I wanted to take the top clusters for each variety and make sense of their respective buzz words. I eliminated common words such as: wine, nose, palate, ect. and used the model to create our own artificial sommelier!

Understanding what the 10 most common words are in each cluster our model created:

```
0 : cherries, blackberries, currants, raspberries, dry, flavors, tannins, cola, rich, wine
1 : cherry, flavors, blackberry, dry, tannins, currant, sweet, drink, soft, wine
2 : red, light, cherry, fruit, aromas, palate, cranberry, flavors, nose, raspberry
3 : wine, fruits, drink, acidity, ripe, character, wood, texture, fruit, ready
4 : apple, lemon, palate, lime, citrus, finish, flavors, peach, acidity, fresh
5 : wine, fruit, flavors, oak, finish, tannins, vineyard, cherry, well, new
6 : black, cherry, tannins, palate, aromas, dark, fruit, plum, pepper, wine
7 : bodied, full, medium, wine, flavors, fruit, finish, texture, cherry, aromas
8 : pineapple, buttered, vanilla, flavors, toast, acidity, oak, sweet, rich, crisp
9 : berry, aromas, plum, herbal, finish, flavors, palate, feels, notes, oak
```

Visualizing our varieties and their respective clusters with a heat-map:



With this information, I made a profile for each variety expressing their top clusters and comparing our generalized buzz words to what is on winefolly.com.

Cabernet Sauvignon

Top Clusters: 1, 5

Buzz words:
1 : **cherry**, flavors, **blackberry**, **dry**, tannins, **currant**, **sweet**, drink, **soft**, wine
5 : wine, fruit, flavors, oak, finish, tannins, vineyard, **cherry**, well, new

Artificial Sommelier:

- **Primary Flavors:** Cherry, Blackberry, Currant
- **Taste Profile:** Sweet, Soft, Dry

Wine Folly:

The world's most popular red wine is a natural cross between Cabernet Franc and Sauvignon Blanc that originated around Bordeaux, France. Wines are concentrated and age worthy.

https://winefolly.com/cabernet-sauvignon/

- **Primary Flavors:** Black Cherry, Black Currant, Cedar, Baking Spices, Graphite
- **Taste Profile:** Full Body, Dry, Medium-High Tannin, Medium Acidity, 13.5–15% ABV

Chardonnay

Top Clusters: 4, 8

Buzz words:
4 : **apple**, **lemon**, palate, **lime**, **citrus**, finish, flavors, **peach**, **acidity**, **fresh**
8 : **pineapple**, **buttered**, **vanilla**, flavors, **toast**, **acidity**, oak, **sweet**, **rich**, **crisp**

Artificial Sommelier:

- **Primary Flavors:** Apple, Pear, Citrus (Lemon, Grapefruit, Lime), Buttered, Pineapple, Peach, Vanilla, Toast, Green
- **Taste Profile:** Acidity, Fresh, Rich, Dry, Sweet, Crisp

Wine Folly:

One of the world's most popular grapes, Chardonnay is made in a wide range of styles from lean, sparkling Blanc de Blancs to rich, creamy white wines aged in oak.

https://winefolly.com/chardonnay/

- **Primary Flavors:** Yellow Apple, Starfruit, Pineapple , Vanilla, Butter
- **Taste Profile:** Medium Body, Dry, Low/None Tannin, Medium Acidity, 13.5–15% ABV

Malbec

Top Cluster: 5, 9

Buzz words:
5 : wine, fruit, flavors, oak, finish, tannins, vineyard, **cherry**, well, new
9 : **berry**, aromas, **plum**, **herbal**, finish, flavors, palate, feels, notes, oak

Artificial Sommelier:

- **Primary Flavors:** Berry, Plum, Herbal
- **Taste Profile:** None

Wine Folly:

Argentina's most important variety came by way of France, where it's commonly called Côt (sounds like "coat"). Wines are loved for their lusty fruit flavors and smooth chocolatey finish.

https://winefolly.com/malbec/

- **Primary Flavors:** Red Plum, Blackberry, Vanilla, Tobacco, Cocoa
- **Taste Profile:** Full Body, Dry, Medium Tannin, Medium-Low Acidity, 13.5–15% ABV

Merlot

Top Clusters: 1, 5

Buzz words:
1 : **cherry**, flavors, **blackberry**, **dry**, tannins, **currant**, **sweet**, drink, **soft**, wine
5 : wine, fruit, flavors, oak, finish, tannins, vineyard, **cherry**, well, new

Artificial Sommelier:

- **Primary Flavors:** Cherry, Blackberry, Currant
- **Taste Profile:** Sweet, Soft, Dry

Wine Folly:

Merlot is loved for it's boisterous black cherry flavors, supple tannins, and chocolatey finish. On the high end, it's often mistaken with Cabernet Sauvignon and commonly blended with it.

https://winefolly.com/merlot/

- **Primary Flavors:** Cherry, Plum, Chocolate, Bay Leaf, Vanilla
- **Taste Profile:** Medium-Full Body, Bone-Dry, Medium-High Tannin, Medium Acidity, 13.5–15% ABV

Pinot Noir

Top Clusters: 1, 2, 5

Buzz words:
1 : **cherry**, flavors, **blackberry**, **dry**, tannins, **currant**, **sweet**, drink, **soft**, wine
2 : red, **light**, **cherry**, fruit, aromas, palate, **cranberry**, flavors, nose, **raspberry**
5 : wine, fruit, flavors, oak, finish, tannins, vineyard, **cherry**, well, new

Artificial Sommelier:

- **Primary Flavors:** Cherry, Cranberry, Rasberry
- **Taste Profile:** Light, Dry, Sweet, Soft

Wine Folly:

Pinot Noir is the world's most popular light-bodied red wine. It's loved for its red fruit, flower, and spice aromas that are accentuated by a long, smooth finish.

https://winefolly.com/pinot-noir/

- **Primary Flavors:** Cherry, Rasberry, Mushroom, Vanilla, Hibiscus
- **Taste Profile:** Medium Body, Dry, Medium Low Tannin, Medium-High Acidity, 11.5–13.5% ABV

Riesling

Top Cluster: 4

Buzz words:
4 : **apple**, **lemon**, palate, **lime**, **citrus**, finish, flavors, **peach**, **acidity**, **fresh**

Artificial Sommelier:

- **Primary Flavors:** Apple, Citrus (Lemon, Lime), Peach
- **Taste Profile:** Acidity

Wine Folly:

An aromatic white variety that can produce white wines ranging in style from bone-dry to very sweet. Germany is the world's most important producer of Riesling.

https://winefolly.com/riesling/

- **Primary Flavors:** Lime, Green Apple, Beeswax, Jasmine, Petrol
- **Taste Profile:** Light Body, Off-Dry, Low/None Tannin, High Acidity, Under 10% ABV

Sangiovese

Top Clusters: 5, 6

Buzz words:
5 : wine, fruit, flavors, oak, finish, tannins, vineyard, **cherry**, well, new
6 : black, **cherry**, tannins, palate, aromas, **dark**, fruit, **plum**, **pepper**, wine

Artificial Sommelier:

- **Primary Flavors:** Cherry, Plum, Pepper
- **Taste Profile:** Dark

Wine Folly:

Italy's most planted wine variety and the pride of the Tuscan regional wine, Chianti. Sangiovese is a sensitive grape that takes on different stylistic expressions based on where it grows.

https://winefolly.com/sangiovese/

- **Primary Flavors:** Cherry, Roasted Tomato, Oregano, Espresso, Sweet Balsamic
- **Taste Profile:** Medium-Full Body, Bone-Dry, Medium-High Tannin, Medium-High Acidity, 13.5–15% ABV

Sauvignon Blanc

Top Cluster: 4, 5

Buzz words:
4 : **apple**, **lemon**, palate, **lime**, **citrus**, finish, flavors, **peach**, **acidity**, **fresh**
5 : wine, fruit, flavors, oak, finish, tannins, vineyard, **cherry**, well, new

Artificial Sommelier:

- **Primary Flavors:** Apple, Citrus (Lemon, Lime), Peach
- **Taste Profile:** Acidity, Fresh

  Confused as to why cluster 9 was not prominent with Sauvignon Blac as it had "herbal" as one of the buzz words.

Wine Folly:

A popular and unmistakable white that's loved for its "green" herbal flavors and sky high acidity. Sauvignon Blanc grows nearly everywhere and is produced in a variety of methods resulting in a wide that range from lean to bountiful.

https://winefolly.com/sauvignon-blanc/

- **Primary Flavors:** Gooseberry, Honeydew Melon, Grapefruit, White Peach, Passion Fruit
- **Taste Profile:** Medium-Light Body, Dry, Low/None Tannin, High Acidity, 11.5–13.5% ABV

Syrah¶

Top Clusters: 1, 5, 6

Buzz words:
1 : **cherry**, flavors, **blackberry**, **dry**, tannins, **currant**, **sweet**, drink, **soft**, wine
5 : wine, fruit, flavors, oak, finish, tannins, vineyard, **cherry**, well, new
6 : black, **cherry**, tannins, palate, aromas, **dark**, fruit, **plum**, **pepper**, wine

Artificial Sommelier:

- **Primary Flavors:** Cherry, Plum, Blackberry, Currant, Pepper
- **Taste Profile:** Dark, Sweet, Soft, Dry

Wine Folly:

A rich, powerful, and sometimes meaty red wine that originated in the Rhône Valley of France. Syrah is the most planted grape of Australia where they call it Shiraz.

https://winefolly.com/syrah/

- **Primary Flavors:** Blueberry, Black Plum, Milk Chocolate, Tobacco, Green Peppercorn
- **Taste Profile:** Full Body, Dry, Medium-High Tannin, Medium Acidity, 13.5–15% ABV

Zinfandel

Top Clusters: 1, 5

Buzz words:
1 : **cherry**, flavors, **blackberry**, **dry**, tannins, **currant**, **sweet**, drink, **soft**, wine
5 : wine, fruit, flavors, oak, finish, tannins, vineyard, **cherry**, well, new

Artificial Sommelier:

- **Primary Flavors:** Cherry, Blackberry, Currant
- **Taste Profile:** Sweet, Soft, Dry

Wine Folly:

A fruit-forward-yet-bold red that's loved for its red fruit flavors and smoky exotic spice notes. Originally from Croatia and related to top Croatian grape, Plavic Mali.

https://winefolly.com/search/zinfandel

- **Primary Flavors:** Blackberry, Strawberry, Peach Preserves, Cinnamon, Sweet Tobacco
- **Taste Profile:** Medium-Full Body, Dry, Medium-High Tannin, Medium-Low Acidity, Over 15% ABV

From these profiles, we can see that they are generalizing some wines very well such as chardonnay, cabernet sauvignon, and syrah. Then generalizing some not so well such as merlot, malbec, and sauvignon blanc. It makes sense for the good ones because we have more information on them in our data. Then on the other hand,

varieties such as merlot and malbec are very similar to cabernets, and as a result tend to get mixed up.

Findings:

It would also be interesting to present this data to Wine Folly. Even though our flavors and taste profiles didn't always line up, we could be discovering new flavors and profiles that should be added to the webpage. Given all the data we have, we are looking at the buzz words most used by professional sommeliers to describe certain wines. Wine Folly could be missing some distinct flavors that are widely used, but not necessarily prominently known or understood.

Next Steps:

It would be interesting to further tune the logistic regression to get better accuracy scores by playing around with the hyperparameters and introducing more feature variables. I would also like to experiment with removing varieties that get mixed up often. With the clustering model, I want to see how hard it would be to only give weight to the buzz words in the clustering method. Taking out words like nose, wine, palate ect. to get more accurate generalizations of how the wines taste. Finally, I could also look at the most common 15 or 20 words instead of 10 to see if I'm potentially missing any important buzz words that just barely didn't make the cut.