WINE ANALYSIS

September 2019

ABSTRACT

Is it possible to predict a wine varietal by a description? Is the Sommelier profession in jeopardy? The following paper explores wine descriptions to determine if varietal predictability is feasible.

Ian Aliman, Erin Lore, Brittney Sherman, Jo Vivian

IST707

Introduction

From fancy dinner dates to religious offerings to being good for cardiovascular health, wine is and has been promoted as a staple in many domains for many years. The origin of wine is often debated, however, most agree that wine was discovered versus invented. It is said that the ancient people mistakenly discovered wine when their grapes spoiled and fermented. “The earliest archaeological evidence of wine produced from grapes has been found at sites in China (c. 7000 BC), Georgia (c. 6000 BC), Lebanon (c. 50000 BC), Iran (c. 5000 BC), Greece (c. 4500 BC) and Sicily (c. 4000 BC).”[1] There are many tales told about which of the ancient people first discovered wine. One example is of the Persian Princess. Having heard that she was disliked by the King of Persia, the Princess attempted suicide by consuming a jar of spoiled grapes. Instead of succeeding in her suicide attempt, she found herself feeling better and being happier. Once the King saw her in this state, he liked her new attitude and she was again liked by him.[2]

Regardless of when and where wine originated, it is one of the most consumed alcoholic beverages around the World. In the US alone, 966 gallons are consumed each year.[3] This is up over 200% since the 1970’s. Wine popularity has grown in large parts due to people dining out more, wine availability and the growth of wine tourism. The explosion of wineries in the US has been instrumental in the increase of wine consumption. “In 2019, there were a total of 10,043 wineries in the United States producing wine. Since 2009, the number of wineries in the U.S. has grown by over 50 percent.”[4]

With the increase in the number of wineries, there is certainly no lack of wine varieties. Worldwide, there are over 10,000 types of grapes used in the production of wine.[5] While some welcome the choices, others find deciding on which year, which vineyard and ultimately which wine to select to be an overwhelming process. And why drink wine unless it is going to be an enjoyable experience? Wine can range from tasting like the back of a school bus to quaffable to transcendent. Sadly, not everyone knows how to find that transcendent bottle and how to avoid drinking from what tastes like the back of a school bus. True, if one is visiting an upscale restaurant, a Sommelier, or wine steward, can help. But how does the average person, looking to buy a bottle of wine or two, know which would suit their taste preferences and which may go best with a certain cuisine?

The following analysis explores various wine reviews to determine if predicting wine varietals from a description is feasible. This ability could open the door for wine

[1] https://en.wikipedia.org/wiki/History_of_wine

[2] https://wine.lovetoknow.com/wiki/Who_Invented_Wine

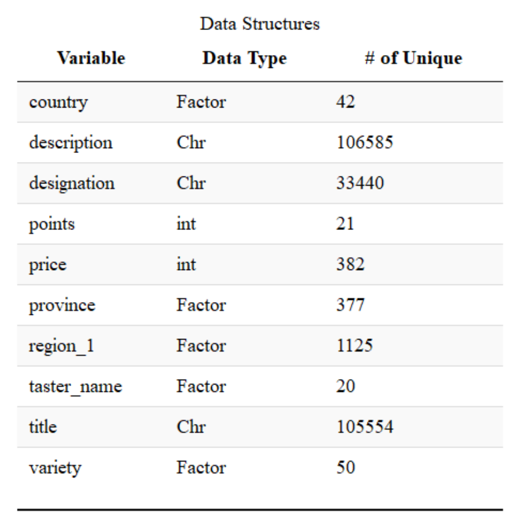
[3] https://www.wineinstitute.org/resources/statistics/article86

[4] https://www.statista.com/statistics/259353/number-of-wineries-in-the-us/

[5] https://www.winestyr.com/wine-guide/how-many-different-types-of-wine-grapes-are-there

distributors to reach additional consumers, perhaps those that are not connoisseurs of their beverages.

About the Data



Data Structures

Variable	Data Type	# of Unique
country	Factor	42
description	Chr	106585
designation	Chr	33440
points	int	21
price	int	382
province	Factor	377
region_1	Factor	1125
taster_name	Factor	20
title	Chr	105554
variety	Factor	50

Figure 1 variables found in winemag-data-130k-v2.csv.

The data set, obtained through Kaggle, includes reviews by several tasters on various wines. The original dataset contains ~130,000 wine reviews with fourteen variables and is titled “winemag-data-130k-v2.” The entire data dictionary can be found in the Appendix Table 1. Figure 1 to the left highlights the variables of interest of analysis.

Locales:

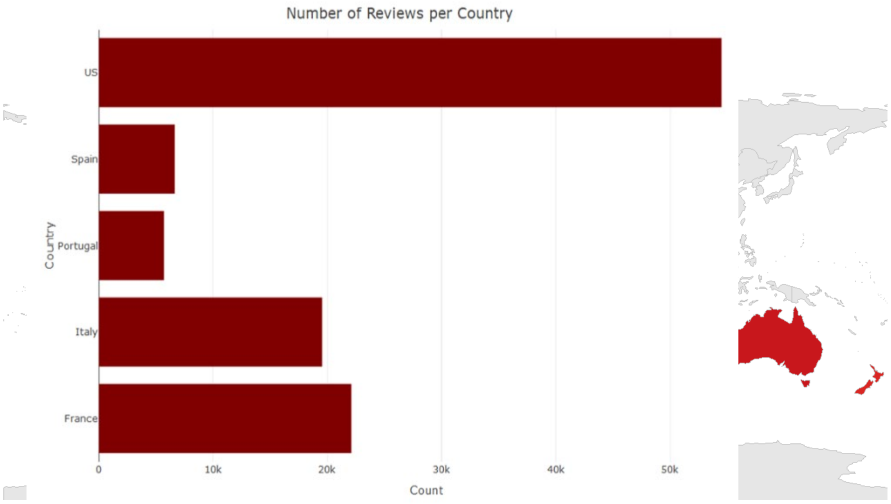


Figure 2 Count of wine reviews by country of wine origin

The wine reviews included wines and varieties from 42 different countries. In Figure 2 to the left, the different countries are visualized. The darker red color indicates a higher frequency of wine reviews, the lighter color indicates fewer reviews, and the gray color indicates that no reviews were provided from that country. While Europe is typically known for its wines, the United States commands the vast majority of the wine reviews. As shown in an alternate graphic provided the United States is responsible for 54,504 of the total 130,000 reviews (41.18% of the total reviews). Not surprisingly, on the list of top 5 are countries more commonly associated with wine consumption and production: France at 22,093 reviews, Italy at 19,540 reviews, Spain at 6,645 reviews, and Portugal at 5,691 reviews. While location provides more salient information regarding varieties and how wine is grown, this analysis focuses on the reviews themselves.

The reviews for the major wine producing countries the United States, France, and non-United States countries were isolated and frequent words were extracted (see Fig. 4). This provides an overview of the qualities and characteristics for wines within different locations. American wines seemed to be more fruit forward and cherry, wines from France seemed to be more acidic, whereas other countries seemed to be producing more aromatic wines.

Figure 3 The top 5 countries by number of wine reviews.

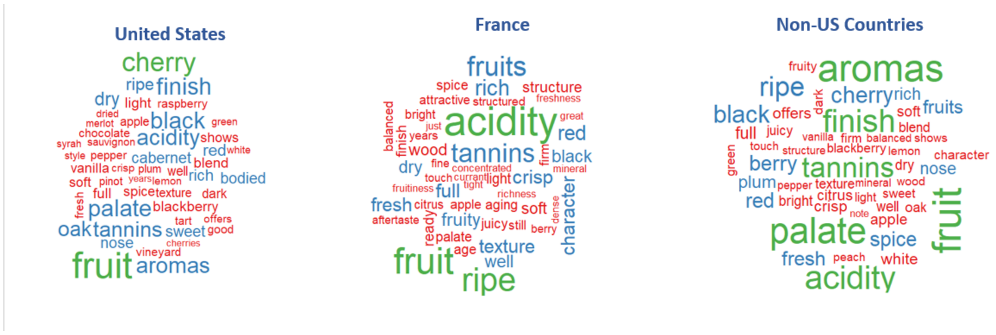


Figure 4 Most frequent words used when reviewing wines produced in US, France, and rest of world non-US countries.

The resulting summary of the dataset is shown in the following tables:

Country	
US	54,504
France	22,093
Italy	19,540
Spain	6,645
Portugal	5,691
Chile	4,472
Other	17,026

Province	
California	36,247
Washington	8,639
Bordeaux	5,941
Tuscany	5,897
Oregon	5,373
Burgundy	3,980
Other	63,894

Winery	
Wines/Winemaker	222
Testarossa	219
DFJ Vinhos	215
Williams Selyem	211
Louis Latour	199
George Duboeuf	196
Other	128,710

Figure 5 Summary of cleansed data

Points and Prices:

The points variable follows a standard rating scale found in Wine Spectator[6]. While Wine Spectators review scale falls as low as 50 points, the data set examined featured more favorable wines.

Rating	Description
95 – 100	Classic – a Great Wine
90 – 94	Outstanding: a wine of superior character and style
85 – 89	Very good: a wine with special qualities
80 – 84	Good: a solid, well-made wine
75 – 79	Mediocre: a drinkable wine that may have minor flaws
50 – 74	Not Recommended

Figure 6 Wine Spectator Review Ranges

The mean of the points is 88 out of a range of 80-100. There are a few wines of note that received scores of 100, they were mostly Bordeaux style red blends (35%) and 58% of those 100 scoring wines were reviewed by one reviewer – Roger Voss who will be introduced below. As shown below, most of the wines fell between 86 and 91 points which is "very good" by Wine Spectator standards. Table 3 in the appendix shows the breakdown of wine varietals and their average point breakdown. Interestingly, there does not seem to be a strong crowd favorite. Sangiovese Grosso scored the highest at 91 points on average. For lowest points at 86 there was a 5-way tie including Moscato and Prosecco.

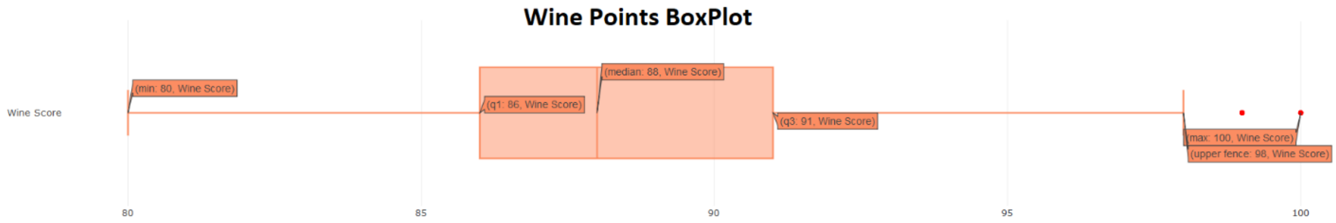


Figure 7 Boxplot of Wine Points.

It was found that prices fell across a range $4-$3,300. The mean price is $35 while the median price is $26. As seen below the mean is heavily shifted by the extreme values of the more premium wines that are in the $1,000 - $3,300 range. Amazingly, the most expensive wine on the list is not a well-aged one. It is a Tempranillo from a 2011 vintage from Covila. In fact, the top 14 most expensive wines are all from 2004 or more recent.

[6] https://www.winespectator.com/articles/scoring-scale

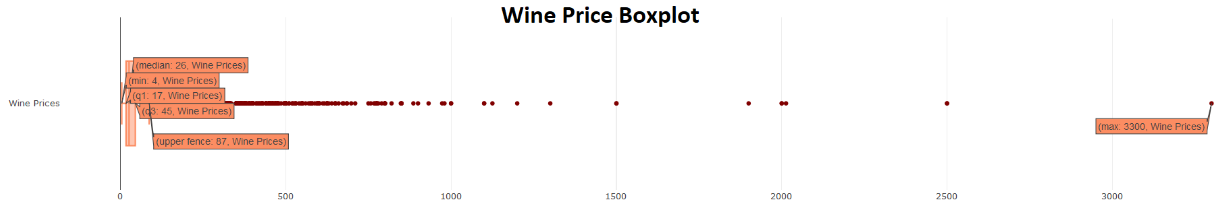


Figure 8 Wine Price Box Plot.
Min: $4 | Max: $3,300 | Median: $26 | Mean $35

Reviewers:

As previously mentioned, for this analysis reviews are paramount. There were 19 unique wine tasters whose reviews were included in this analysis. Unfortunately, there were 26,244 reviews that were not associated with a reviewer, that's about 20% of the available reviews.

Above in the point box plot, it is highlighted that there are 10 wines rated at 100 that were reviewed by Roger Voss. Although he is disproportionately represented in the 17 wines that scored 100, his overall score is in line with other reviewers. There does not seem to be an explicit bias for reviewers that needs correcting. The sidebar to the right provides another table that shows the average points assigned per each wine taster seen in the distribution table below.

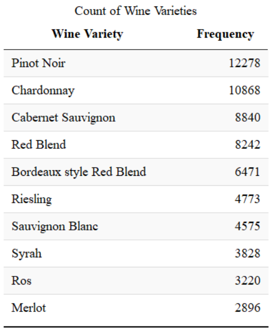
Varietals & Wineries

There are 708 varieties of wine provided in the dataset from a total of 1657 wineries. The most popular wines were Pinot Noir, Chardonnay, and Cabernet Sauvignon which dominate just over 20% of the total wine reviews. Of the wineries, 200 of the wines were from Williams Selyem.

Average Points per Reviewer

Taster	Avg Points
No Name	88
Anna Lee C Iijima	88
Carrie Dykes	87
Fiona Adams	87
Jim Gordon	89
Kerin OKeefe	89
Matt Kettmann	90
Mike DeSimone	89
Roger Voss	89
Susan Kostrzewa	86
Alexander PearTree	86
Anne KrebiehlMW	91
Christina Pickard	88
Jeff Jenssen	88
Joe Czerwinski	89
Lauren Buzzeo	88
Michael Schachner	87
Paul Gregutt	89
Sean P Sullivan	89
Virginie Boone	89

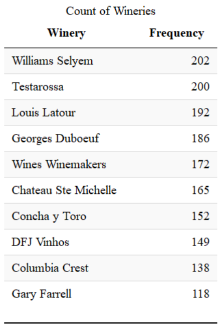
There are 708 Varietals:



Count of Wine Varieties

Wine Variety	Frequency
Pinot Noir	12278
Chardonnay	10868
Cabernet Sauvignon	8840
Red Blend	8242
Bordeaux style Red Blend	6471
Riesling	4773
Sauvignon Blanc	4575
Syrah	3828
Ros	3220
Merlot	2896

There are 1,657 Wineries:



Count of Wineries

Winery	Frequency
Williams Selyem	202
Testarossa	200
Louis Latour	192
Georges Duboeuf	186
Wines Winemakers	172
Chateau Ste Michelle	165
Concha y Toro	152
DFJ Vinhos	149
Columbia Crest	138
Gary Farrell	118

Figure 9 Top 10 Varieties and Wineries

Processing/Cleaning the data

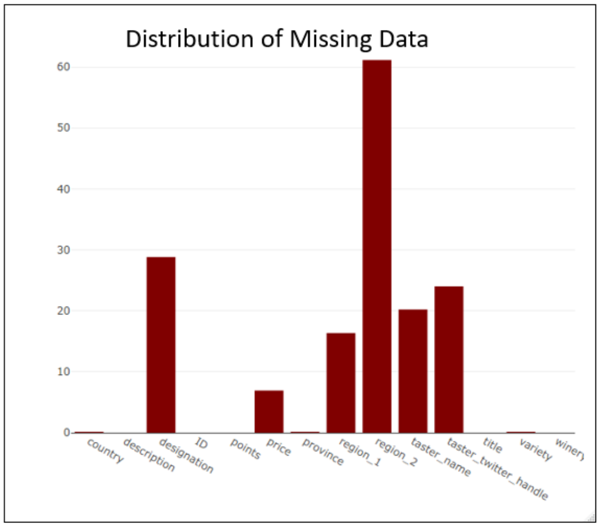


Figure 10 Missing data points

The next step in data exploration is to determine if there are missing values and to remove any unnecessary columns. The below table shows there are several columns with missing data, specifically region 2, taster twitter handle, and designation had high rates of missing values. These categories were eliminated from analysis as there was no accurate methods of imputing the data nor were the variables of interest for analysis.

The columns with high values of missing data were removed. This included the ID column, which was an artifact from the csv just indicating the row number. Retaining the row number was irrelevant to processing as the row index can be called in R; therefore, this was dropped. Additionally, the designation, which is where in the winery the grapes were grown was also removed, as it had nearly 30% of the data missing with no meaningful way to impute the data values nor was it relevant in the prediction of the varietals, therefore this column was also removed. Region 2 was missing nearly 60% of the data and like the designation, impossible to interpret and was removed. Lastly, while the twitter handle of the reviewer could provide more text analysis, for the scope of this analysis these were also dropped from consideration. Now that the meaningful rows were isolated cleansing of the data could begin.

Data Dictionary – Columns Removed		
Variable Name	**Description**	**Reason Removed**
X	ID of review	Unnecessary for analysis
Designation	The vineyard within the winery where the grapes that made the wine are from	Unnecessary for analysis
Region 2	Sometimes there are more specific regions specified within a wine growing area, but this value can sometimes be blank	Unnecessary for analysis
Taster Twitter Handle	Twitter handle of tester	Unnecessary for analysis

Figure 11 Removed columns

Text Processing:

The most troubling aspect of this data set was the special characters included in the dataset. Many of the wines are from overseas and included many special characters. For example, varietal Gewürstraminer being re-written as Gewrstraminer when the umlaut was removed along with its corresponding vowel. This allowed text processing to be run without confusion of the special characters, however, the shortened term "Gewrstraminer" was used as a factor and retained its uniqueness for analytical purposes. This processing of text was run on all columns that contained text which included the reviews, the taster names, region, province, title, variety, and winery.

All numbers and punctuation were removed from the description section using the tm library in R. The description was also transformed into lower case before stop words were removed. Numbers were extracted from the title variable in order to generate the vintage of each wine.

"pinot", "chardonnay","cabernet", "sauvignon", "red", "blend", "champagne", "cab", "pinots", "pinotage", "pinotesque", "campinoti", "pinotphiles", "chardonnays", "reds", "blends", "blended", "blending", "zin", "zinfandels", "zinfully", "zinfandal", "sauv", "blanc", "sauvs", "sauve", "sauvignons", "sauvage","sauvigon", "sauvies", "sauvignonesque", "sauvy", "sauvignonasse", "sauvig non", "sauvion", "sauvgnon", "sauvingon", "sauvingnon", "ssauvignon","cabernert", "cabernets", "cabs", "moscatos", "merlots", "shirazes", "tempranillos", "malbecs", "gris", "gew rztraminer", "sangioveses", "ros", "nebbiolos","mlbech ", "franc", "caberent", "francs", "grigios", "grigio", "white", "whites", "blancs", "barberas'

Figure 12 Wine Terms Removed

General English stop words were removed from the dataset as well as words that were specifically related to wine, which may allow the model to cheat. Amusingly, Sauvignon was the most misspelled varietal which included nearly 15 terms to be removed. These additional wine terms were found and removed using R's View function. Word parts were entered the search box which allowed simplistic searching of the slang terms wine reviewers have created. More sophisticated measures can be used moving forward and there are very likely terms that were missed.

Imputation:

The data contained two numeric variables, price and points. The price as noted above did have around 7% missing variables. As this number was numeric the missing values were imputed with the median. The median was selected over the mean due to the outliers in the dataset. The median of the set was $25 a bottle whereas the mean was $36, and the most expensive bottle was $3,300. Points on the other hand all managed to come over.

Duplication and NA removal:

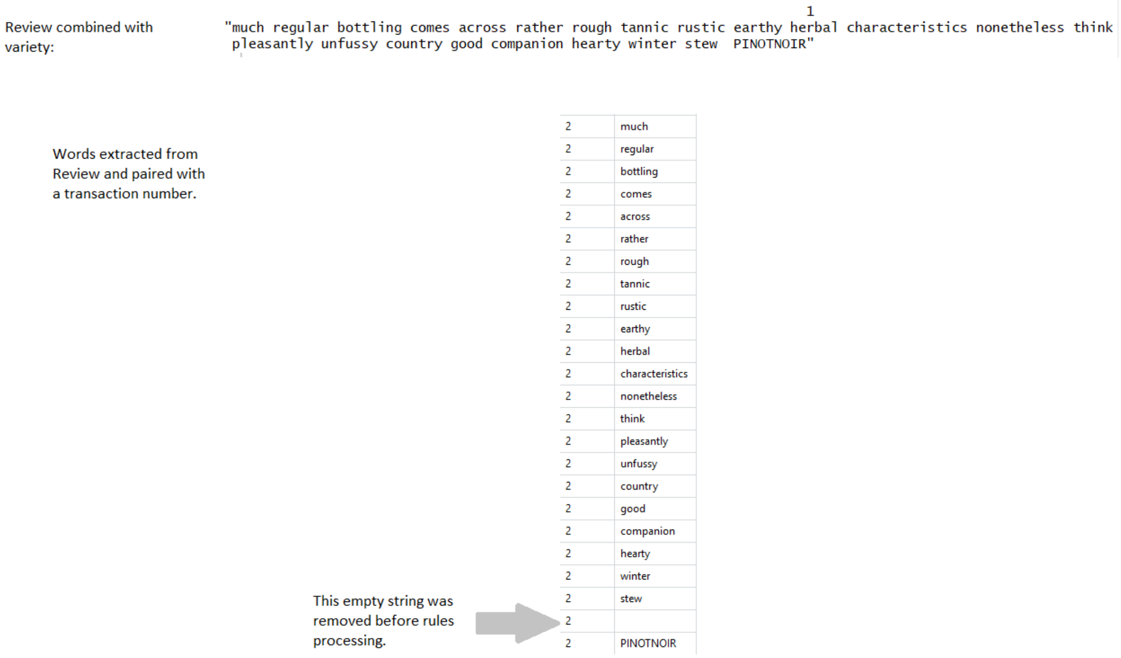
Despite efforts, there were still NA values that were unable to be imputed such as missing reviewer names, missing varietals, and duplicate reviews. This left a data frame of 99,151 reviews.

Model specific cleansing:

Association Rule Mining

For the rule mining the descriptions and varieties were isolated into a data frame. The challenge for this data set was that the reviews were all different lengths and there was no logical way to truncate all the reviews into the same length. In order to obtain meaningful transactions from the data set two transformations took place. Spaces were removed from the varieties and turned uppercase, for example: pinot noir to PINOTNIOR. This was completed to ensure that the terms that were being isolated were in fact the varietal and not an artifact from the wine review in which the variety was referenced. Then the two rows were combined into a single string using a loop and a paste function to contain both the review and the variety within one string.

The second transformation extracted each word and assigned it a transaction number and copied it into a new data frame. The new data frame contained two columns, the transaction ID which was the number of the original review, and the word. This new data frame resulted in 2,485,499 individual words that were in all the reviews. The figure below demonstrates this transformation visually. Once the transaction set was created empty strings were removed. These empty strings were an artifact of punctuation removal. Once this was accomplished the transaction data set was complete.



Review combined with variety:

```
                                                                      1
"much regular bottling comes across rather rough tannic rustic earthy herbal characteristics nonetheless think
 pleasantly unfussy country good companion hearty winter stew  PINOTNOIR"
```

Words extracted from Review and paired with a transaction number.

2	much
2	regular
2	bottling
2	comes
2	across
2	rather
2	rough
2	tannic
2	rustic
2	earthy
2	herbal
2	characteristics
2	nonetheless
2	think
2	pleasantly
2	unfussy
2	country
2	good
2	companion
2	hearty
2	winter
2	stew
2	
2	PINOTNOIR

This empty string was removed before rules processing.

Figure 13 Transformation of Transaction Data

Modeling

Model #1: Association Rule Mining to determine if wine descriptions can predict varietal

Association rule mining was also used to discover the most commonly used words describing each type of wine. Association rule mining is a powerful data exploration tool that allows commonly occurring words be turned into “rules”. These rules can then be used for a suggestion engine or as a starting point for further analysis like a classification algorithm or a clustering algorithm.

The top 50 most frequent wines were analyzed. By identifying the main words used by wine tasters to describe specific varietals, wine distributors could better market and describe wines to various consumer groups.

Once the transaction dataset was prepared association rules were run first on Pinot Noir and Chardonnay as they were the most common wines. The variety was isolated to the right-hand side of the equation. The lift was set at 0.001 as there were many reviews and it was of interest to obtain fewer common rules that have high confidence. The confidence was set to .01 for similar reasons. It was of interest to look at the strong rules, but weak rules can also be of interest to analysis. It can show what terms not to use when describing a wine. Lastly the minimum length was set to 5. It would be overwhelming to look at 2-word pairs for a list of 2,485,499 words and their varieties. Setting the minimum length allows for 4 descriptors and 1 wine variety. This makes trends in words easier to identify.

For the large rule set, the same exploratory function was used as above, however for the righthand side of the equation all 50 varieties of wines were constrained. This means that any rule associated with a variety that has at least 4 words will show.

Model #2: K-Means Clustering

K-means is a clustering technique that identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the distance between the cluster centroids and cluster members as small as possible.

To create the K-Means model a column was added to the data frame to differentiate between red, white, rose, and sparkling wines. Then, due to the small numbers of sparkling and roses, two subsets of only reds and whites were created. The distribution of wine types can be seen in Figure 16 below. Next, a very large matrix was created from a list of tokens created from the description feature. Then, each row was divided by the number of terms to give more weight to words where few were used. For the red and white wines, take the cross products of the matrices to get a matrix with an indication of the correlation between words. Then, divide each row by the sum of the row and apply the natural logarithm to the cross product of both the reds and the whites to get measurements. The k-means model created word clusters based on the rows which consisted of the words in the descriptions. It used a random seed generator of 100 and a k of 50. It narrowed the clusters to only those with clusters greater than 5 words and then divided the cluster variation with log size to get a measurement of how close the cluster in question was. The output described in the Results section consists of the 10 best clusters for red, and the 10 best for white.

Model #3: Decision Trees to Predict White vs Red Wines

To predict the variety of wine based on the description, Decision Tree analysis was used. A corpus and subsequently a document term matrix were created from the description column. The variety column was used as the labels for the training and testing sets.

The first model included both red and white wines from the data frame limited to the top 60 wines as discussed in the Data Exploration section. This model resulted in a very large decision tree with only 20% accuracy. The next two models, seen in the top two squares, separated out the white wine and red wine, but still collectively included all wines from the top 60 selection. These models also still had a very low (22%) accuracy, and the confusion matrix revealed that the descriptions for all wines other than chardonnay, Riesling, sauvignon blanc, pinot noir, red blend, and sauvignon did not provide enough data to make accurate predictions. The models seen in bottom two squares explored only the top 3 wines within each of the red and white wine document term matrix. The classification method was used for every instance of these decision trees.

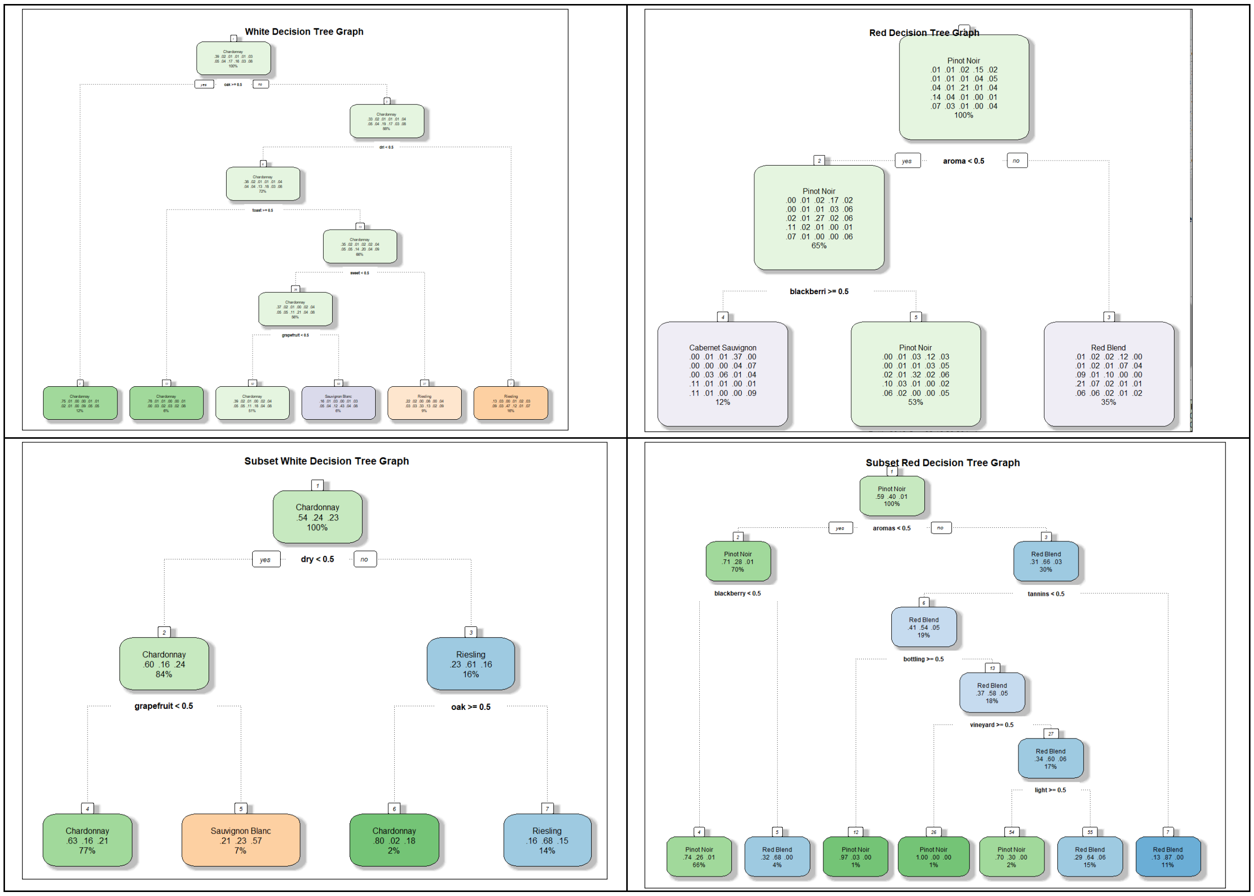


Figure 14 Decision Tree attempts

Model #4: Decision Trees using only Adverbs and Adjectives

Understanding that Decision Tree and Naïve Bayes classifiers were achieving better results with smaller word vector-spaces, revised Decision Tree models were trained using the data from a set of adverbs and adjectives extracted from the description column. Going back to the original scope, it makes sense that a prospective diner, shopper, or sommelier would want to find a varietal using a few choice words and still get accurate results from the wine description predictive model.

Decision tree V2 preprocessing

Two data sets were used to create a DTM matrix for the decision trees. The columns used were the Variety label (the prediction) and the simplified descriptions that contained adverbs and adjectives. These words were important in categorical analysis and were therefore a likely space to investigate. The top four wines were predicted using a multi-class decision tree model -- Pinot Noir, Chardonnay, Cabernet Sauvignon, and Riesling. Hybrid blends were specifically left out, as the blend may have qualities of one wine or another, but in general, the flavor profile would be dependent on the grapes of the prior batch. Cabernet Sauvignon, and Riesling. The Variety label was joined to the AR words table using sqldf's inner join to create a "wide," or sparse document term matrix (DTM). *a priori* tuning was performed based on previous model results. Wine varietals were reduced to four wines and parameters were revised for the decision tree model. The data were then randomized and split into train and test groups, 70% and 30%, respectively. The process of growing trees with different parameter settings yielded many of the EDA results that were helpful when exploring the best ways to grow trees before pruning. During pruning, the tree X-error was used to prune the trees. Pruning was conducted by evaluating the local minima of the complexity parameter (cp), which helps understand when data partitioning has been optimized, and adding one standard deviation to that value. *a posteriori* tweaks were made to models and new model variants were built.

Two methods were used to evaluate the decision trees:

- **Confusion matrix.** A multi-class confusion matrix was used to assess accuracy.
- **Receiver Operating Characteristic (ROC):** ROC curves were used to assess True Positive Rate (TPR) and True Failure Rate (TFR).

For reference, a list of parameters used are included,

- **Split index: information gain.** Information gain is similar to GINI.
- **complexity parameter (cp) = 0.03**. Any split that does not decrease the overall lack of fit by a factor of cp
- **bucket minimum (min_bucket) = 300:** The minimum number of observations that must exist in every *leaf* node.
- **xval: 2:5.** Cross validation parameter was tested at 2,3,4, and 5 folds.
- **minsplit = 200.** The minimum number of observations that must exist in any node (leaf or branch)

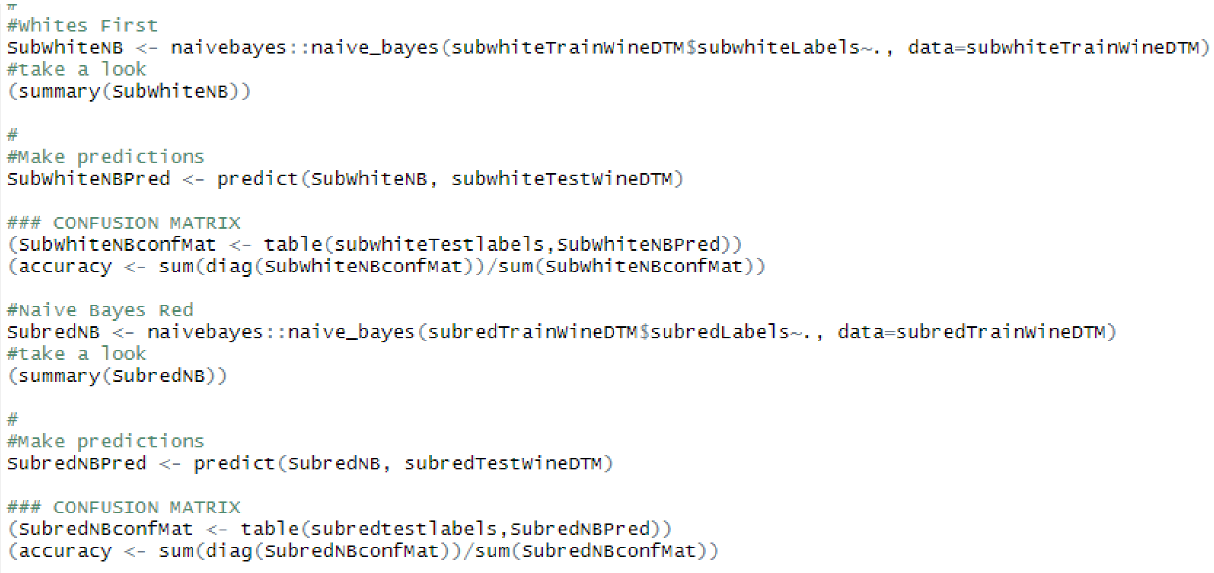
The following parameters were tested and modeled. The final parameter configuration produced the most pleasing results and was selected to highlight.

1. PARAMETERS -- Split index = information gain. complexity parameter (cp) = 0.03. Bucket minimum (min_bucket) = 300.
2. PARAMETERS -- Split index = information gain. complexity parameter (cp) = 0.03. Bucket minimum (min_bucket) = 300, xval(2:5).
   a. *Model 10 was a replica of #9, but with cross-validation added.*

3. PARAMETERS -- Split index = information gain. complexity parameter (cp) = 0.002.
    a. *The goal of above model was to cast a wider net than the first decision tree attempt, with aggressive pruning parameters.*
4. PARAMETERS -- Split index = information gain. complexity parameter (cp) = 0.002. Minsplit = 200.
    a. *The goal of this decision model was to ensure each class was represented equally. The over-represented classes were all down sampled to match the number of Riesling wine reviews found in the training data (4685).*

Model #5 Naïve Bayes:

Once again, the goal is to predict the variety of wine based on the description. And, once again, the first model was run on the entire Document Term Matrix (DTM) with very poor accuracy, so it was re-run on the subset of whites and the subset of reds.

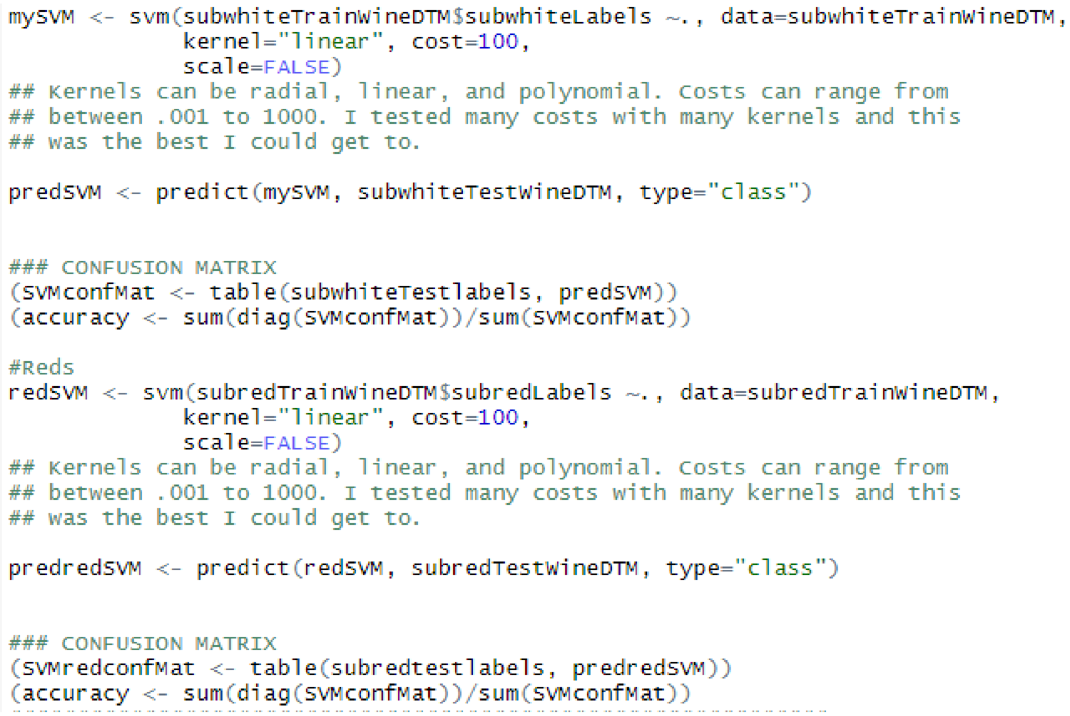


```
#
#Whites First
SubWhiteNB <- naivebayes::naive_bayes(subwhiteTrainWineDTM$subwhiteLabels~., data=subwhiteTrainWineDTM)
#take a look
(summary(SubWhiteNB))

#
#Make predictions
SubWhiteNBPred <- predict(SubWhiteNB, subwhiteTestWineDTM)

### CONFUSION MATRIX
(SubWhiteNBconfMat <- table(subwhiteTestlabels,SubWhiteNBPred))
(accuracy <- sum(diag(SubWhiteNBconfMat))/sum(SubWhiteNBconfMat))

#Naive Bayes Red
SubredNB <- naivebayes::naive_bayes(subredTrainWineDTM$subredLabels~., data=subredTrainWineDTM)
#take a look
(summary(SubredNB))

#
#Make predictions
SubredNBPred <- predict(SubredNB, subredTestWineDTM)

### CONFUSION MATRIX
(SubredNBconfMat <- table(subredtestlabels,SubredNBPred))
(accuracy <- sum(diag(SubredNBconfMat))/sum(SubredNBconfMat))
```

Figure 15 R Studio Output for Naive Bayes

Model #6 Support Vector Machines

After the decision trees and Naïve Bayes, it no longer made sense to keep trying to get predictions for the larger dataset, so only the subset data was used for the remaining models. For the SVM, polynomial, radial, and linear kernels were all attempted, with linear producing the best results. A cost ranging from 0 to 1000 was modeled on the linear kernel and a cost of 100 was used in the final model.

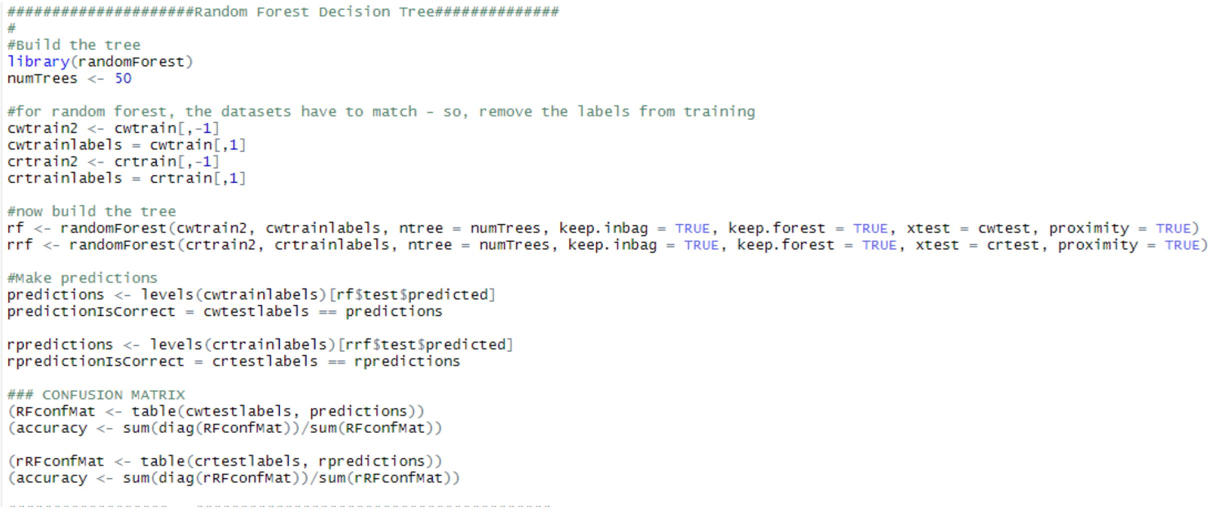


```
mySVM <- svm(subwhiteTrainWineDTM$subwhiteLabels ~., data=subwhiteTrainWineDTM,
             kernel="linear", cost=100,
             scale=FALSE)
## Kernels can be radial, linear, and polynomial. Costs can range from
## between .001 to 1000. I tested many costs with many kernels and this
## was the best I could get to.

predSVM <- predict(mySVM, subwhiteTestWineDTM, type="class")


### CONFUSION MATRIX
(SVMconfMat <- table(subwhiteTestlabels, predSVM))
(accuracy <- sum(diag(SVMconfMat))/sum(SVMconfMat))

#Reds
redSVM <- svm(subredTrainWineDTM$subredLabels ~., data=subredTrainWineDTM,
             kernel="linear", cost=100,
             scale=FALSE)
## Kernels can be radial, linear, and polynomial. Costs can range from
## between .001 to 1000. I tested many costs with many kernels and this
## was the best I could get to.

predredSVM <- predict(redSVM, subredTestWineDTM, type="class")


### CONFUSION MATRIX
(SVMredconfMat <- table(subredtestlabels, predredSVM))
(accuracy <- sum(diag(SVMconfMat))/sum(SVMconfMat))
```

Figure 16 R Studio Output for SVM

Model #7 Random Forest

The random forest model contained 50 decision trees and was run on the subset white and subset red data.

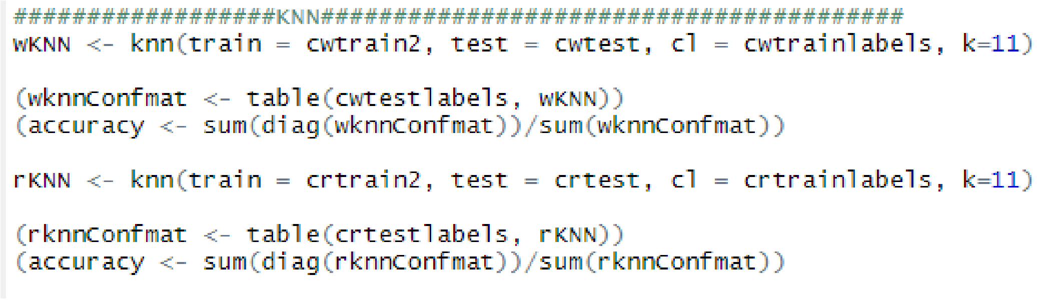


```
####################Random Forest Decision Tree##############
#
#Build the tree
library(randomForest)
numTrees <- 50

#for random forest, the datasets have to match - so, remove the labels from training
cwtrain2 <- cwtrain[,-1]
cwtrainlabels = cwtrain[,1]
crtrain2 <- crtrain[,-1]
crtrainlabels = crtrain[,1]

#now build the tree
rf <- randomForest(cwtrain2, cwtrainlabels, ntree = numTrees, keep.inbag = TRUE, keep.forest = TRUE, xtest = cwtest, proximity = TRUE)
rrf <- randomForest(crtrain2, crtrainlabels, ntree = numTrees, keep.inbag = TRUE, keep.forest = TRUE, xtest = crtest, proximity = TRUE)

#Make predictions
predictions <- levels(cwtrainlabels)[rf$test$predicted]
predictionIsCorrect = cwtestlabels == predictions

rpredictions <- levels(crtrainlabels)[rrf$test$predicted]
rpredictionIsCorrect = crtestlabels == rpredictions

### CONFUSION MATRIX
(RFconfMat <- table(cwtestlabels, predictions))
(accuracy <- sum(diag(RFconfMat))/sum(RFconfMat))

(rRFconfMat <- table(crtestlabels, rpredictions))
(accuracy <- sum(diag(rRFconfMat))/sum(rRFconfMat))
```

Figure 17 R Studio Output for

Model #8 K-Nearest Neighbor

KNN was run on the subset of white and the subset of red and a range from 10 – 100 used for k. Accuracy decreased above 50 and showed little change between 10 – 50, so the square root of the number of variables was used.



```
#################KNN###########################################
wKNN <- knn(train = cwtrain2, test = cwtest, cl = cwtrainlabels, k=11)

(wknnConfmat <- table(cwtestlabels, wKNN))
(accuracy <- sum(diag(wknnConfmat))/sum(wknnConfmat))

rKNN <- knn(train = crtrain2, test = crtest, cl = crtrainlabels, k=11)

(rknnConfmat <- table(crtestlabels, rKNN))
(accuracy <- sum(diag(rknnConfmat))/sum(rknnConfmat))
```

Figure 18 R Studio Output for KNN

Results

Model #1 Results: Association Rule Mining

Association rule mining provided a fundamental insight into common words that were used for each of the wine varietals. The two selected outputs were provided for example as they were the two most prolific wines. Pinot Noir shows raspberry, cranberry, and cherry as common words in the top 5 rules. This would indicate that if a client were to request a wine with cherry notes, based on the descriptions of the wine reviewers, Pinot Noir would be a safe bet. Similarly, for the Chardonnay's buttered, toast, and pineapple are all terms a customer could use to describe a Chardonnay. While rule mining is an unsupervised training technique, it does not necessarily allow for testing for accuracy; however, the information gained from the rules can help influence future models by eliminating wines without key terms. For example, if a customer requested a oak wine, both Pinot Noir and Chardonnay may be viable recommendations, however, if the client offers the terms "oaky" and "pineapple" Pinot Noir can be eliminated as a possibility as none of the rules use that term.

Pinot Noir

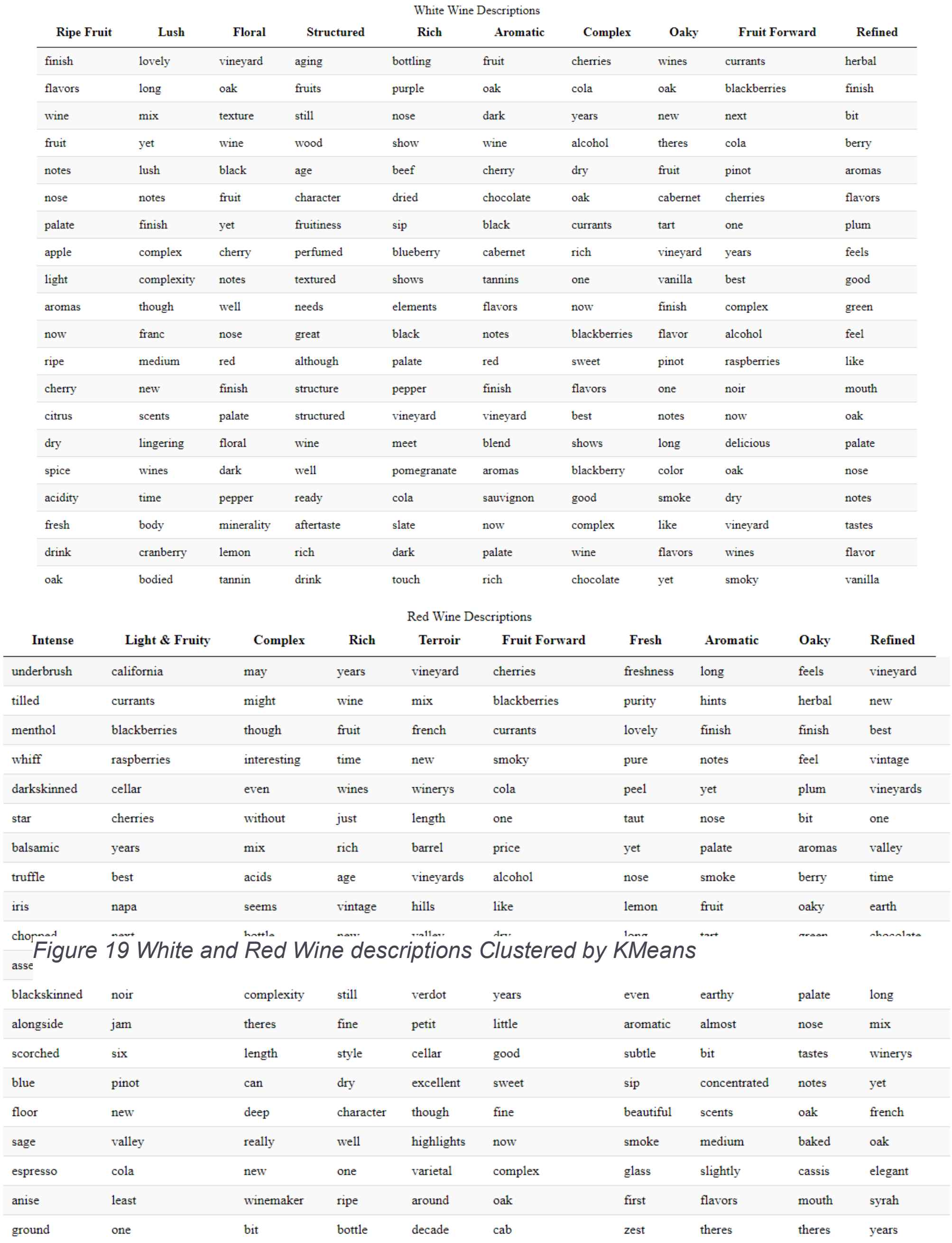
Lhs	rhs	Support	confidence	lift	count
{cherry,cola,dry,silky} =>	{PINOTNOIR}	0.001022274	0.90833333	7.8881774	109
{bottling,cranberry,nose,palate} =>	{PINOTNOIR}	0.001153576	0.85416667	7.4177815	123
{cranberry,fruit,nose,palate} =>	{PINOTNOIR}	0.001041032	0.6686747	5.8069262	111
{cherry,fruit,oak,raspberry} =>	{PINOTNOIR}	0.001031653	0.58823529	5.1083717	110
{acidity,cherry,fruit,raspberry} =>	{PINOTNOIR}	0.001228605	0.52822581	4.5872354	131

Chardonnay

Lhs		rhs	support	confidence	lift	count
{buttered,oak,pineapple,toast}	=>	{CHARDONNAY}	0.001003517	0.96396396	9.4573664	107
{acidity,buttered,pineapple,toast}	=>	{CHARDONNAY}	0.001162954	0.96124031	9.4306448	124
{buttered,pineapple,toast,vanilla}	=>	{CHARDONNAY}	0.001209848	0.94852941	9.3059393	129
{buttered,fruit,toast,tropical}	=>	{CHARDONNAY}	0.001012896	0.92307692	9.0562272	108
{acidity,buttered,crisp,toast}	=>	{CHARDONNAY}	0.001069168	0.86363636	8.473061	114

Model #2 Results K-Means Clustering for descriptions

Using K-Means clustering provided 10 lists of words to describe the varieties of wine for both the red and white types of wines. These lists can be seen in the tables below. These lists revealed distinct types of wine, because the wines are described in specific ways. For example, descriptions for complex reds such as Cabernet Sauvignon, and Bordeaux style blends clustered together and can be seen in the Complex, Rich, and Refined columns, while descriptions of wines like Pinot Noir, which tend to be lighter and fresh clustered in columns such as Light & Fruity, Fruit Forward, and Fresh.



White Wine Descriptions

Ripe Fruit	Lush	Floral	Structured	Rich	Aromatic	Complex	Oaky	Fruit Forward	Refined
finish	lovely	vineyard	aging	bottling	fruit	cherries	wines	currants	herbal
flavors	long	oak	fruits	purple	oak	cola	oak	blackberries	finish
wine	mix	texture	still	nose	dark	years	new	next	bit
fruit	yet	wine	wood	show	wine	alcohol	theres	cola	berry
notes	lush	black	age	beef	cherry	dry	fruit	pinot	aromas
nose	notes	fruit	character	dried	chocolate	oak	cabernet	cherries	flavors
palate	finish	yet	fruitiness	sip	black	currants	tart	one	plum
apple	complex	cherry	perfumed	blueberry	cabernet	rich	vineyard	years	feels
light	complexity	notes	textured	shows	tannins	one	vanilla	best	good
aromas	though	well	needs	elements	flavors	now	finish	complex	green
now	franc	nose	great	black	notes	blackberries	flavor	alcohol	feel
ripe	medium	red	although	palate	red	sweet	pinot	raspberries	like
cherry	new	finish	structure	pepper	finish	flavors	one	noir	mouth
citrus	scents	palate	structured	vineyard	vineyard	best	notes	now	oak
dry	lingering	floral	wine	meet	blend	shows	long	delicious	palate
spice	wines	dark	well	pomegranate	aromas	blackberry	color	oak	nose
acidity	time	pepper	ready	cola	sauvignon	good	smoke	dry	notes
fresh	body	minerality	aftertaste	slate	now	complex	like	vineyard	tastes
drink	cranberry	lemon	rich	dark	palate	wine	flavors	wines	flavor
oak	bodied	tannin	drink	touch	rich	chocolate	yet	smoky	vanilla

Red Wine Descriptions

Intense	Light & Fruity	Complex	Rich	Terroir	Fruit Forward	Fresh	Aromatic	Oaky	Refined
underbrush	california	may	years	vineyard	cherries	freshness	long	feels	vineyard
tilled	currants	might	wine	mix	blackberries	purity	hints	herbal	new
menthol	blackberries	though	fruit	french	currants	lovely	finish	finish	best
whiff	raspberries	interesting	time	new	smoky	pure	notes	feel	vintage
darkskinned	cellar	even	wines	winerys	cola	peel	yet	plum	vineyards
star	cherries	without	just	length	one	taut	nose	bit	one
balsamic	years	mix	rich	barrel	price	yet	palate	aromas	valley
truffle	best	acids	age	vineyards	alcohol	nose	smoke	berry	time
iris	napa	seems	vintage	hills	like	lemon	fruit	oaky	earth
chopped	next	bottle	new	valley	dry	long	tart	green	chocolate
asse									
blackskinned	noir	complexity	still	verdot	years	even	earthy	palate	long
alongside	jam	theres	fine	petit	little	aromatic	almost	nose	mix
scorched	six	length	style	cellar	good	subtle	bit	tastes	winerys
blue	pinot	can	dry	excellent	sweet	sip	concentrated	notes	yet
floor	new	deep	character	though	fine	beautiful	scents	oak	french
sage	valley	really	well	highlights	now	smoke	medium	baked	oak
espresso	cola	new	one	varietal	complex	glass	slightly	cassis	elegant
anise	least	winemaker	ripe	around	oak	first	flavors	mouth	syrah
ground	one	bit	bottle	decade	cab	zest	theres	theres	years

Figure 19 White and Red Wine descriptions Clustered by KMeans

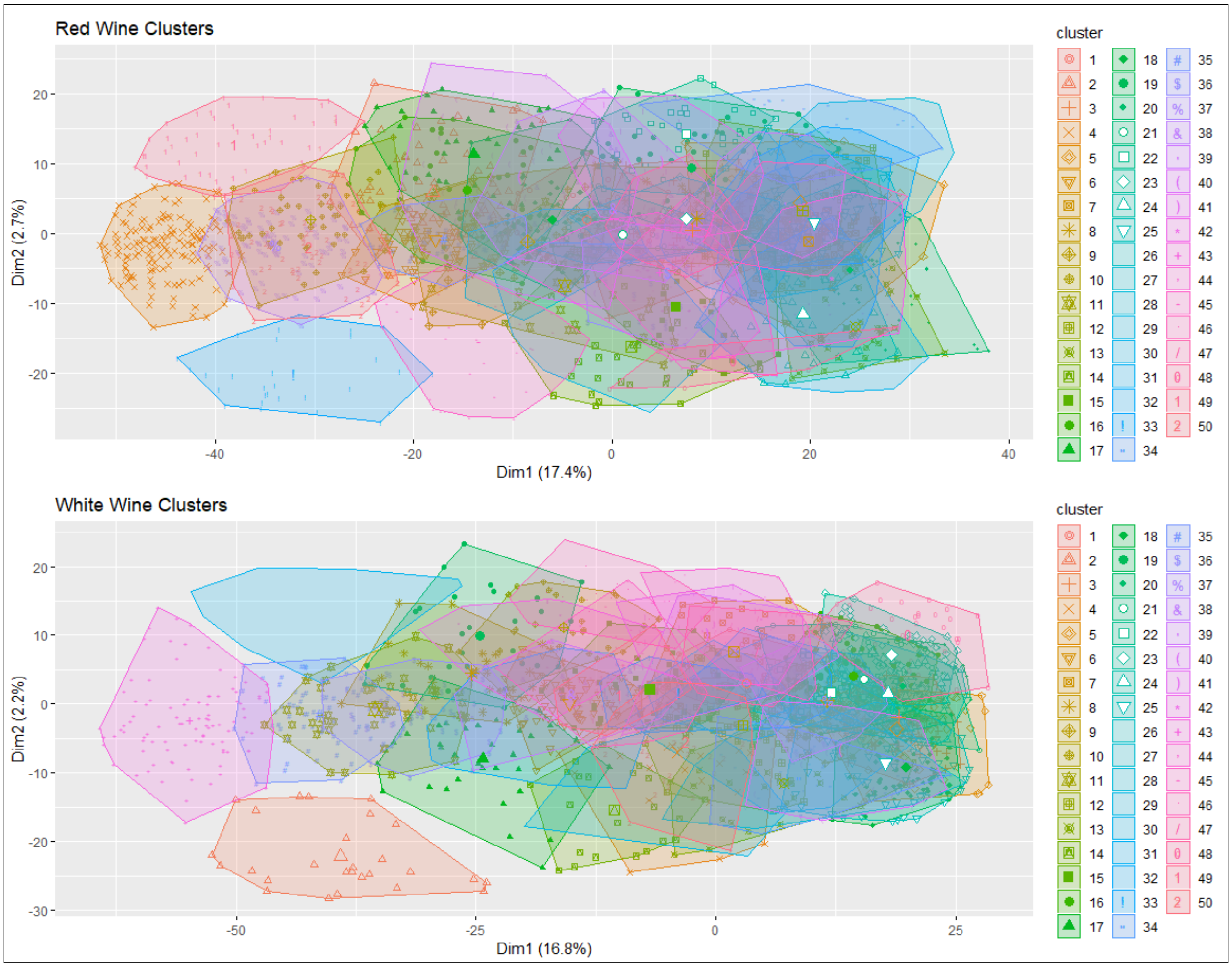


Figure 20 - Red and White Word Clusters

Model # 3 Results Decision Tree to Predict Red Vs White wines

Decision Trees

Subset White Wine: 63% Accuracy

	Chardonnay	**Riesling**	**Sauvignon Blanc**
Chardonnay	8525	477	313
Riesling	2042	1774	275
Sauvignon Blanc	2866	377	679

Subset Red Wine: 74% Accuracy

	Pinot Noir	**Red Blend**	**Sauvignon**
Pinot Noir	9200	1324	0

Red Blend	3082	3982	0
Sauvignon	95	147	0

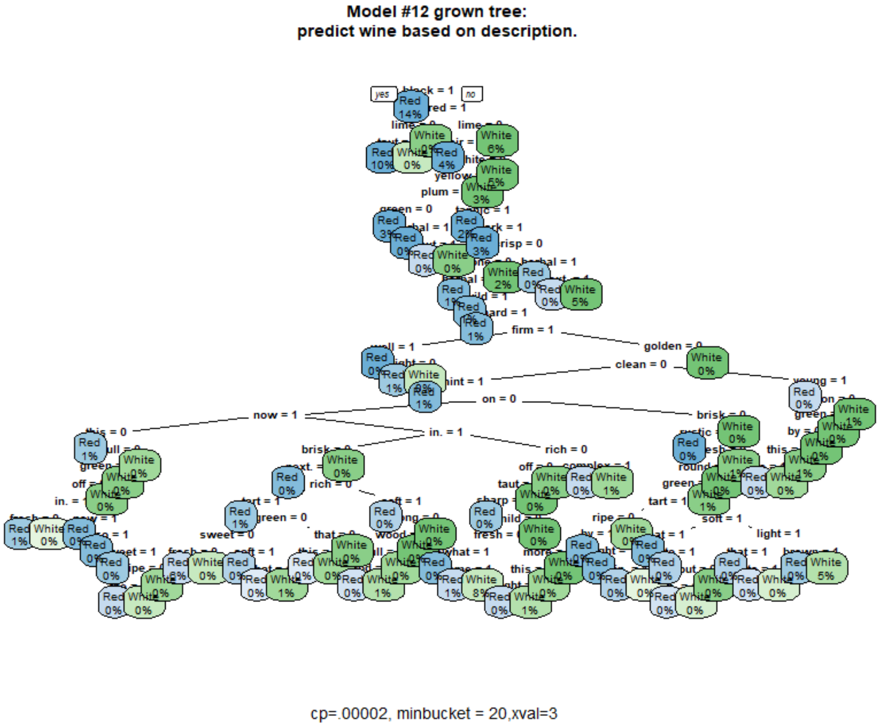
Subset top 4 wines (2 white, 2 red): 79.7%

	Red (Pinot Noir, Sauvignon)	**White** (Riesling, Chardonnay)
Red (Pinot Noir, Sauvignon)	6194	1241
White (Riesling, Chardonnay)	1496	4558

P-Value [Acc > NIR] : < 2.2e-16

Model #4 Results (accuracy: 79.1%)

An overgrown tree was produced to find the right prune point using x-error value.



Model #12 grown tree:
predict wine based on description.

cp=.00002, minbucket = 20,xval=3

Figure 21 Unpruned tree

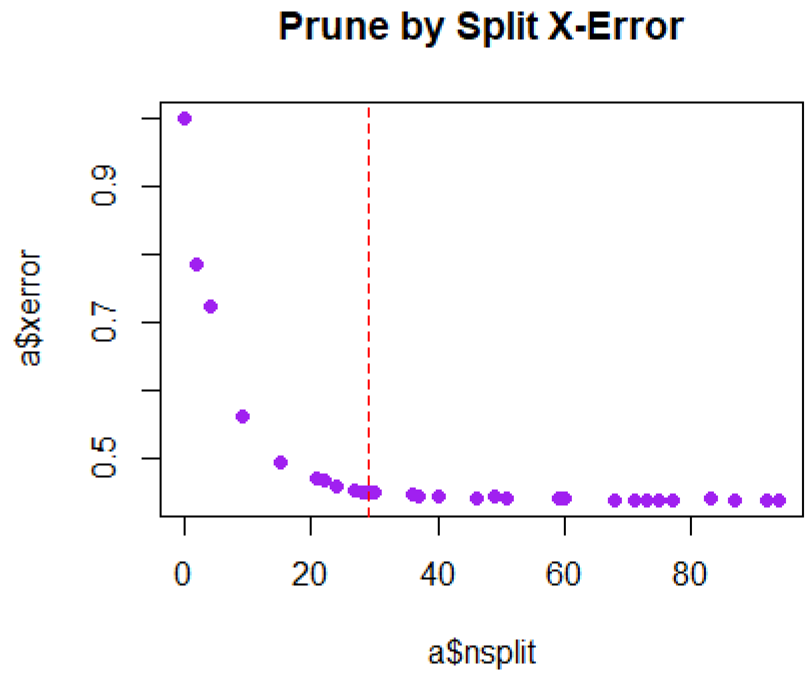


Figure 23 x-error pruning point

The pruned tree shows preference for the most obvious descriptors (e.g., red, yellow).

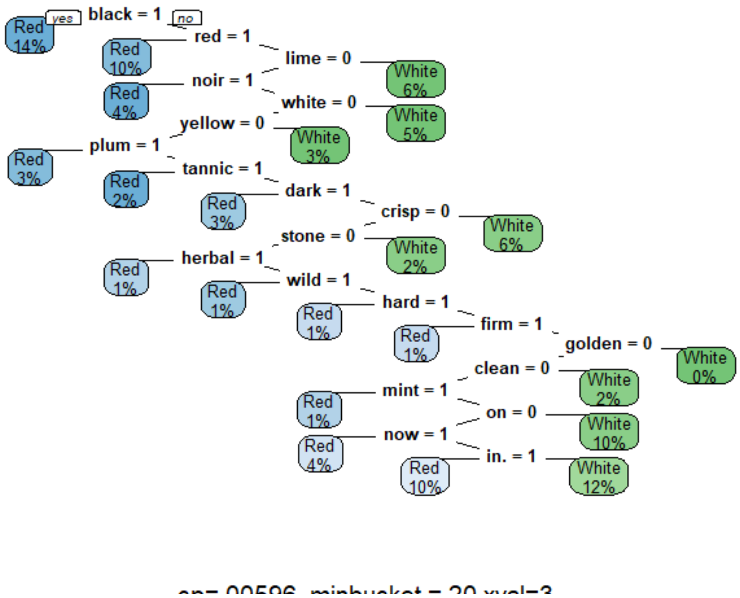


Figure 22 Pruned White Wine Tree

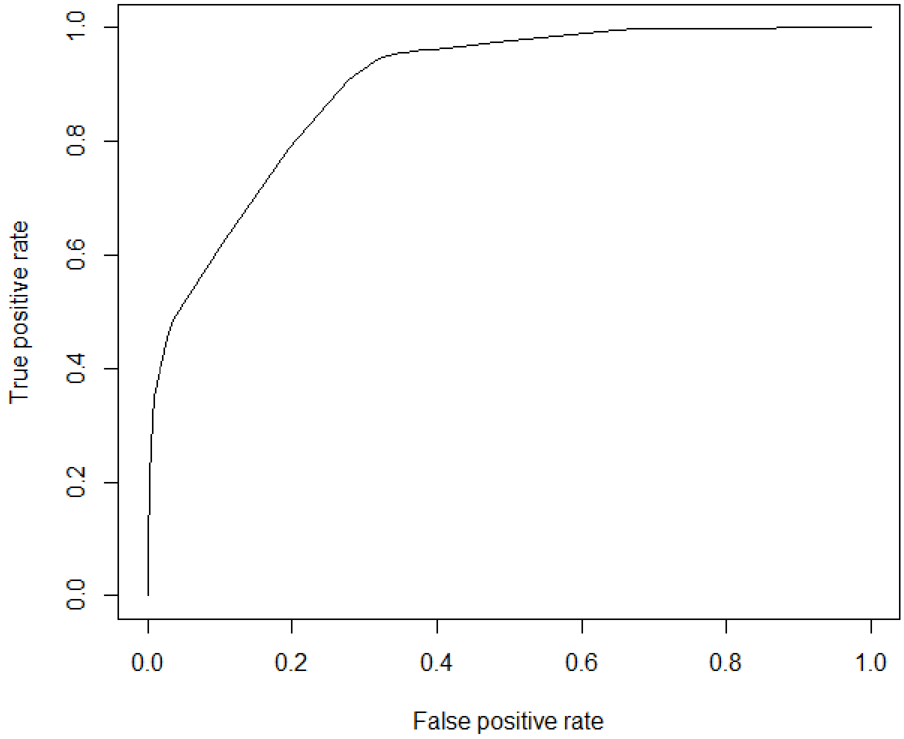


Figure 24 False positive rate vs true positive rate

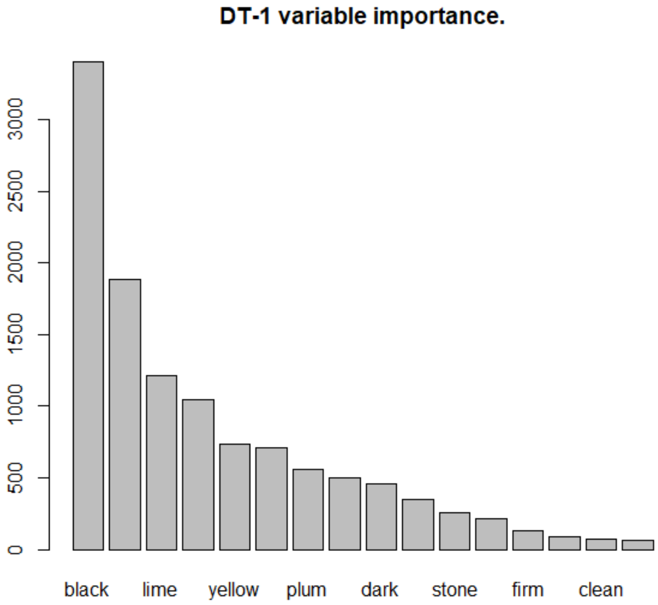


Figure 25 Feature importance by information gain

Model #5 Results: Naïve Bayes

Subset White Wine: 56% Accuracy

	Chardonnay	Riesling	Sauvignon Blanc
Chardonnay	4286	2350	2679
Riesling	181	3394	516
Sauvignon Blanc	454	1443	2025

Subset Red Wine: 1.4% Accuracy

	Pinot Noir	Red Blend	Sauvignon
Pinot Noir	0	0	10524
Red Blend	0	0	7064
Sauvignon	0	0	242

Model #6 Results: SVM

Subset White = 72%

	Chardonnay	Riesling	Sauvignon Blanc
Chardonnay	7844	623	848
Riesling	881	2776	434
Sauvignon Blanc	1758	387	1777

Subset Red = 72%

	Pinot Noir	Red Blend	Sauvignon
Pinot Noir	9275	1188	61
Red Blend	1915	5059	90
Sauvignon	64	48	130

Model # 7 Results: Random Forest

Subset White = 72%

	Chardonnay	Riesling	Sauvignon Blanc
Chardonnay	8199	471	645
Riesling	1028	2679	384
Sauvignon Blanc	1962	383	1577

Subset Red = 81%

	Pinot Noir	Red Blend	Sauvignon
Pinot Noir	9298	1223	3
Red Blend	1970	5093	1
Sauvignon	77	97	68

Model #8 Results: KNN

Subset White = 64%

	Chardonnay	Riesling	Sauvignon Blanc
Chardonnay	8301	83	931
Riesling	2032	1166	893
Sauvignon Blanc	2229	64	1629

Subset Red = 76%

	Pinot Noir	Red Blend	Sauvignon
Pinot Noir	8830	1691	3
Red Blend	2408	4648	8
Sauvignon	97	63	82

Conclusion

Based on the above analysis it is possible to use data analytic modeling techniques to predict the wine varietal based on a description. For example, “buttery, toasty”, would predict chardonnay. Modeling accuracy ranged from 64-81% and therefore, the techniques with the highest accuracies would be advised.

By predicting wine varietals, a wine distributor could reach the consumer that doesn’t already know the types of wines they would like. According to a recent study conducted for a large wine distributor, there are six types of consumers: Enthusiast, Image Seekers, Savvy Shopper, Traditionalist, Satisfied Sipper and Overwhelmed. At one end of the spectrum, the enthusiasts “are very knowledgeable and particularly passionate about the entire subject of wine. They study and explore wine’s global aspects -often sharing their discoveries with their wine drinking friends.” On the opposite end of the spectrum, the overwhelmed consumer “simply cannot deal with the wine buying experience. There are too many options on the shelves, and this creates a fretful, intimidating experience for them."[7] Varietal predictions could save those consumers who do not like the wine selection process a lot of angst and time.

Imagine a kiosk at the liquor store or at a table at a restaurant. The kiosk could allow for natural language entries of the description of flavors one prefers, or even the meal they would be pairing the wine with. The kiosk would return the various wine varietals that meet the description. As with all industries, technology is discovering new ways to do business and to save corporations every day – perhaps the profession of Master Sommelier, which earns on average of $US150,000 per year[8], will be replaced by a kiosk at a restaurant near you.

[7] R&D Study conducted for a large wine distributor, which needs to remain confidential.

[8] https://www.businessinsider.com/what-it-takes-to-become-a-master-sommelier-2015-6

Appendix

Data Dictionary – Original Dataset		
Variable Name	**Description**	**Data Type**
ID	Line item/row count	Integer
Country	The country that the wine is from	Factor w/ 44 levels
Description	Description of wine	Factor w/ 119,955 levels
Designation	The vineyard within the winery where the grapes that made the wine are from	Factor w/ 37,977 levels
Points	The # of points wineenthusiast.com rated the wine on a scale of 1-100 (they only post reviews that score >=80)	Integer
Price	Price of the wine	Integer
Province	The province or state that the wine is from	Factor w/ 426 levels
Region 1	The wine growing area in a province or state (i.e., Napa)	Factor w/ 1,230 levels
Region 2	Sometimes there are more specific regions specified within a wine growing area, but this value can sometimes be blank	Factor w/ 18 levels
Taster Name	Name of person who taste tested the wine	Factor w/ 20 levels
Taster Twitter Handle	Twitter handle of tester	Factor w/ 16 levels
Title	The title of the wine review, which often contains the vintage	Factor w/ 118,8840 levels
Variety	The type of grapes used to make the wine (i.e. Pinot Noir)	Factor w/ 708 levels
Winery	The winery that made the wine	Factor w/ 16,757 levels

Table 1 Data Dictionary

Province	Variety	Count	Province	Variety	Count
California	Pinot Noir	6,896	California	Cabernet Sauvignon	5,693
California	Chardonnay	5,183	Bordeaux	Bordeaux Style Red Blend	4,617
Oregon	Pinot Noir	2,786	Piedmont	Nebbiolo	2,664
California	Zinfandel	2,639	Burgundy	Chardonnay	2,312
Tuscany	Red Blend	2,174	Tuscany	Sangiovese	2,152
California	Syrah	1,870	California	Sauvignon Blanc	1,807
California	Red Blend	1,804	Burgundy	Pinot Noir	1,550
California	Merlot	1,391	Mendoza Province	Malbec	1,367
Washington	Cabernet Sauvignon	1,365	Champagne	Champagne Blend	1,221
N. Spain	Tempranillo	1,212	Washington	Syrah	1,129
Provence	Rosè	1,062	Mosel	Riesling	1,011
Beaujolais	Gamay	987	Bordeaux	Bordeaux Style White Blend	951
Douro	Portuguese Red	880	California	Bordeaux Style Red Blend	876
Washington	Red Blend	854	Loire Valley	Sauvignon Blanc	752
Tuscany	Sangiovese Grosso	750	Washington	Bordeaux Style Red Blend	743
New York	Riesling	734	Alsace	Riesling	718
Veneto	Glera	704	California	Petite Sirah	694
Washington	Merlot	666	Veneto	Corvina Rondinella Molinara	619
Washington	Chardonnay	612	Port	Port	600
N. Spain	Tempranillo Blend	558	Veneto	Red Blend	552
California	Sparkling Blend	547	Alentejano	Portuguese Red	535
California	Rosè	513	N. Italy	Pinot Grigio	513
California	Rhone Style Red	499	Oregon	Chardonnay	498
Mendoza Province	Cabernet Sauvignon	486	Rhone Valley	Rhone Style Red Blend	484
Piedmont	Barbera	479	California	Sparkling Blend	477
Alsace	Gewürstraminer	476	Oregon	Pinot Gris	474
Alsace	Pinot Gris	458	Marlborough	Sauvignon Blanc	446
S. Australia	Shiraz	443	California	Viognier	436
Washington	Riesling	422	S.W. France	Malbec	419
Sicily Sardinia	Red Blend	401	California	Grenache	373

Table 2 Breakdown of Wine varieties by Province(or State)

Variety	Points	Variety	Points	Variety	Points
Aglianico	89	Albarino	88	Barbera	89
Blaufränkisch	90	Bordeaux Style Red Blend	89	Bordeaux Style White Blend	89
Cabernet Franc	88	Cabernet Sauvignon	89	Carmenere	87
Champagne Blend	90	Chardonnay	88	Chenin Blanc	89
Corvina Rondinella Molinara	88	Gamay	88	Garganega	88
Garnacha	86	Gewürztraminer	89	Glera	87
Grenache	89	Grüner Veltliner	90	Malbec	88
Melon	88	Meritage	88	Merlot	87
Montepulciano	87	Moscato	86	Mourvdre	89
Nebbiolo	90	Nero d'Avola	87	Petit Verdot	88
Petite Sirah	88	Pinot Blanc	88	Pinot Grigio	86
Pinot Gris	88	Pinot Noir	89	Port	90
Portuguese Red	89	Portuguese White	87	Primitivo	87
Prosecco	86	Red Blend	88	Rhonestyle Red Blend	89
Rhone Style White Blend	88	Riesling	89	Rosè	87
Sangiovese	89	Sangiovese Grosso	91	Sauvignon	88
Sauvignon Blanc	87	Shiraz	89	Sparkling Blend	88
Syrah	89	Tempranillo	88	Tempranillo Blend	88
Torrontes	86	Verdejo	88	Vermentino	88
Viognier	88	White Blend	87	Zinfandel	88

Table 3 Average point rating per Wine

The top sixty varietals can be seen below. Data exploration was limited to the top sixty since wine varieties beyond that number became too small for analysis and contained unfamiliar varieties.

Varietal	Count	Varietal	Count	Varietal	Count
Pinot Noir	13,272	Chardonnay	11,753	Cabernet Sauvignon	9,472
Red Blend	8,946	Bordeaux Style Red Blend	6,915	Riesling	5,189
Sauvignon Blanc	4,967	Syrah	4,142	Rose	3,564
Merlot	3,102	Nebbiolo	2,804	Zinfandel	2,714
Sangiovese	2,707	Malbec	2,652	Portuguese Red	2,466
White Blend	2,370	Sparkling Blend	2,153	Tempranillo	1,810
Rhine Style Red Blend	1,471	Pinot Gris	1,455	Champagne Blend	1,396
Cabernet Franc	1,353	Grüner Veltliner	1,345	Portuguese White	1,159
Bordeaux Style White Blend	1,056	Pinot Grigio	1,052	Gamay	1,025
Gewürztraminer	1,012	Viognier	996	Shiraz	836
Petite Sirah	770	Sangiovese Grosso	751	Barbera	721
Glera	709	Port	668	Grenache	651
Corvina Rondinella Molinara	619	Chenin Blanc	591	Tempranillo Blend	588
Carmenàre	575	Albariño	477	Pinot Blanc	442
Rhine Style White Blend	425	Nero d'Avola	365	Aglianico	359
Moscato	358	Garnacha	326	Sauvignon	316
Verdejo	294	Melon	280	Garganega	270
Petit Verdot	269	Meritage	260	Torrontès	260
Prosecco	236	Blaufränkisch	232	Vermentino	232
Mourvedre	229	Primitivo	222	Montepulciano	215

Table 4 Top 60 Varietals