Wine Data Clustering Report

Arturo M.
Paulette R.

May 6th, 2024

Summary

For this project, we are working with a wine data set adapted from the UCI Machine Learning Repository, where the results are based on the chemical analysis of wines grown from the same region in Italy, and derived from three different cultivars. There are 13 variables and a total of n = 178 observations, with each variable being constituents found in each of the wines analyzed. To summarize the project, we have found a total of three decided clustering groups that best represent the wines collected from the data, arranging from mellow and intense wines to the average consumer. Therefore, we've decided to name each cluster as follows: sweet and warm wine with high alcohol concentration, tart and soft wine with a dark indigo hue, and bitter, flat wine with low alcohol concentration.

Introduction

The goal is to apply clustering algorithms to see if we can group the wines based on having similar chemical properties and give names to the wines that fit those attributes.

The names of the 13 variables, all being numerical values, are alcohol by volume (**Alcohol**); which has wines that taste bolder, warmer, and oilier, and wines with less alcohol are lighter and more delicate. The amount of malic acid (**Malic_Acid**); is mainly attributed to a sour taste but softens the wine's profile and is done to most red wines, while white wines generally retain a crisper edge. The amount of ash, a type of inorganic salt (**Ash**); affects the overall flavor of the wine and can give the wine a fresh feeling. The amount of alkalinity of ash (**Ash_Alcanity**); wine grown in low-nutrient soils has more alkalinity, which results in a more rounded, less sour taste. The amount of magnesium (**Magnesium**); increases sugar and alcohol levels in wines. The total molecules containing polyphenolic substances (**Total_Phenols**); the total phenol content of wine varies by type of wine, with red wine having the highest concentration (216 mg/100 ml, or 1000–3000 mg/L). The amount of Flavonoids or phytochemical compounds (**Flavanoids**); compounds are responsible for the wine's bitterness, astringency, and color, as well as for providing potent antioxidant effects. The amount of non-flavonoids (**Nonflavanoid_Phenols);** enhances and stabilizes the color of red wines, and contributes to their flavor, found in both red and white wine but are the main phenolic compounds in white wines. The amount of Proanthocyanidins or condensed tannins

(**Proanthocyanins**); are flavonoid polymers that are responsible for wine's astringent taste and bitterness. The color intensity or the degree of color shade (**Color_Intensity**); the darker the color, the more intense the wine should taste. The vividness of the color (**Hue**); the hue of red wine can range from pale purple to garnet to almost black, and is determined by pigments in the grape skin. The OD280 method of protein concentration in wine (**OD280**); a higher OD280/OD315 absorbance ratio indicates higher protein purity, making the wine taste bitter. Finally, the amount of proline/sensory attributes in red wine (**Proline**); can contribute to desirable sensory attributes in red wine, such as increased sweetness, viscosity, and red fruit flavor. Proline can decrease bitterness and astringency.

Before applying unsupervised methods on our data, we will begin by scaling our numerical variables to ensure each variable contributes equally to the k-means clustering algorithm as the variables do not have the same unit measure.

Statistical Analysis

When doing statistical analysis for the wine data set, we need to check to make sure there are no null values, observe the distributions of the variables and note variables that show signs of high correlations. Observing if there’s skewness to the variable’s density plots will help provide more information about the data set. That being said, K-means does not work well with data that is heavily skewed and has many outliers so we may transform some of the variables.

```
    Alcohol        Malic_Acid          Ash         Ash_Alcanity      Magnesium       Total_Phenols     Flavanoids
 Min.   :11.03   Min.   :0.740   Min.   :1.360   Min.   :10.60   Min.   : 70.00   Min.   :0.980   Min.   :0.340
 1st Qu.:12.36   1st Qu.:1.603   1st Qu.:2.210   1st Qu.:17.20   1st Qu.: 88.00   1st Qu.:1.742   1st Qu.:1.205
 Median :13.05   Median :1.865   Median :2.360   Median :19.50   Median : 98.00   Median :2.355   Median :2.135
 Mean   :13.00   Mean   :2.336   Mean   :2.367   Mean   :19.49   Mean   : 99.74   Mean   :2.295   Mean   :2.029
 3rd Qu.:13.68   3rd Qu.:3.083   3rd Qu.:2.558   3rd Qu.:21.50   3rd Qu.:107.00   3rd Qu.:2.800   3rd Qu.:2.875
 Max.   :14.83   Max.   :5.800   Max.   :3.230   Max.   :30.00   Max.   :162.00   Max.   :3.880   Max.   :5.080
 Nonflavanoid_Phenols Proanthocyanins Color_Intensity       Hue              OD280          Proline
 Min.   :0.1300       Min.   :0.410   Min.   : 1.280   Min.   :0.4800   Min.   :1.270   Min.   : 278.0
 1st Qu.:0.2700       1st Qu.:1.250   1st Qu.: 3.220   1st Qu.:0.7825   1st Qu.:1.938   1st Qu.: 500.5
 Median :0.3400       Median :1.555   Median : 4.690   Median :0.9650   Median :2.780   Median : 673.5
 Mean   :0.3619       Mean   :1.591   Mean   : 5.058   Mean   :0.9574   Mean   :2.612   Mean   : 746.9
 3rd Qu.:0.4375       3rd Qu.:1.950   3rd Qu.: 6.200   3rd Qu.:1.1200   3rd Qu.:3.170   3rd Qu.: 985.0
 Max.   :0.6600       Max.   :3.580   Max.   :13.000   Max.   :1.7100   Max.   :4.000   Max.   :1680.0
```

Figure 1. Summary Statistics on Unscaled Wine Data Set.

Running the summary function for the data set, it is clear that there are no null values for us to try and fill in. Furthermore, we also want to analyze the correlations among the variables.

	Alcohol	Malic_Acid	Ash	Ash_Alcanity	Magnesium	Total_Phenols	Flavanoids
Alcohol	1.00000000	0.09439694	0.211544596	-0.31023514	0.27079823	0.28910112	0.2368149
Malic_Acid	0.09439694	1.00000000	0.164045470	0.28850040	-0.05457510	-0.33516700	-0.4110066
Ash	0.21154460	0.16404547	1.000000000	0.44336719	0.28658669	0.12897954	0.1150773
Ash_Alcanity	-0.31023514	0.28850040	0.443367187	1.00000000	-0.08333309	-0.32111332	-0.3513699
Magnesium	0.27079823	-0.05457510	0.286586691	-0.08333309	1.00000000	0.21440123	0.1957838
Total_Phenols	0.28910112	-0.33516700	0.128979538	-0.32111332	0.21440123	1.00000000	0.8645635
Flavanoids	0.23681493	-0.41100659	0.115077279	-0.35136986	0.19578377	0.86456350	1.0000000
Nonflavanoid_Phenols	-0.15592947	0.29297713	0.186230446	0.36192172	-0.25629405	-0.44993530	-0.5378996
Proanthocyanins	0.13669791	-0.22074619	0.009651935	-0.19732684	0.23644061	0.61241308	0.6526918
Color_Intensity	0.54636420	0.24898534	0.258887259	0.01873198	0.19995001	-0.05513642	-0.1723794
Hue	-0.07174720	-0.56129569	-0.074666889	-0.27395522	0.05539820	0.43368134	0.5434786
OD280	0.07234319	-0.36871043	0.003911231	-0.27676855	0.06600394	0.69994936	0.7871939
Proline	0.64372004	-0.19201056	0.223626264	-0.44059693	0.39335085	0.49811488	0.4941931

	Nonflavanoid_Phenols	Proanthocyanins	Color_Intensity	Hue	OD280	Proline
Alcohol	-0.1559295	0.136697912	0.54636420	-0.07174720	0.072343187	0.6437200
Malic_Acid	0.2929771	-0.220746187	0.24898534	-0.56129569	-0.368710428	-0.1920106
Ash	0.1862304	0.009651935	0.25888726	-0.07466689	0.003911231	0.2236263
Ash_Alcanity	0.3619217	-0.197326836	0.01873198	-0.27395522	-0.276768549	-0.4405969
Magnesium	-0.2562940	0.236440610	0.19995001	0.05539820	0.066003936	0.3933508
Total_Phenols	-0.4499353	0.612413084	-0.05513642	0.43368134	0.699949365	0.4981149
Flavanoids	-0.5378996	0.652691769	-0.17237940	0.54347857	0.787193902	0.4941931
Nonflavanoid_Phenols	1.0000000	-0.365845099	0.13905701	-0.26263963	-0.503269596	-0.3113852
Proanthocyanins	-0.3658451	1.000000000	-0.02524993	0.29554425	0.519067096	0.3304167
Color_Intensity	0.1390570	-0.025249931	1.00000000	-0.52181319	-0.428814942	0.3161001
Hue	-0.2626396	0.295544253	-0.52181319	1.00000000	0.565468293	0.2361834
OD280	-0.5032696	0.519067096	-0.42881494	0.56546829	1.000000000	0.3127611
Proline	-0.3113852	0.330416700	0.31610011	0.23618345	0.312761075	1.0000000

Figure 2. Correlations for 13 Variables.

The highest correlation value is 0.864 for Total_Phenols and Flavonoids, which makes sense since flavonoids are "phytochemical compounds present in many plants, fruits, vegetables, and leaves", and the main group for phenolic compounds include flavonoids, phenolic acids, and more (Ayad et. al.). Furthermore, the second highest correlation value is 0.787 for Flavonoids and the OD280 variables. Since OD280 calculates the protein concentration in wine, then the fact that flavonoid bioavailability is significantly increased in the presence of proteins remains true in this data set (Wang et. al.). Therefore, wines that show high OD280 levels will show an increase in the amount of Flavonoids. Since our variables show relatively moderate to high correlations, we will explore principal component analysis (PCA) to make the dataset more interpretable and help explain the high correlation among the variables.

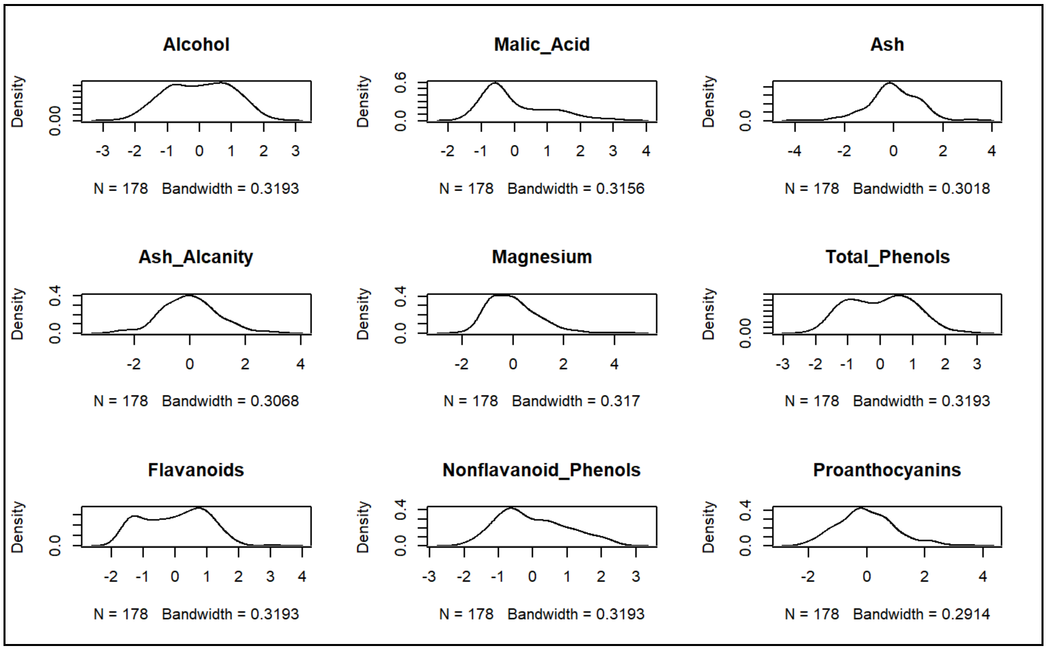

Figure 3.1. Density Plots for Scaled Variables

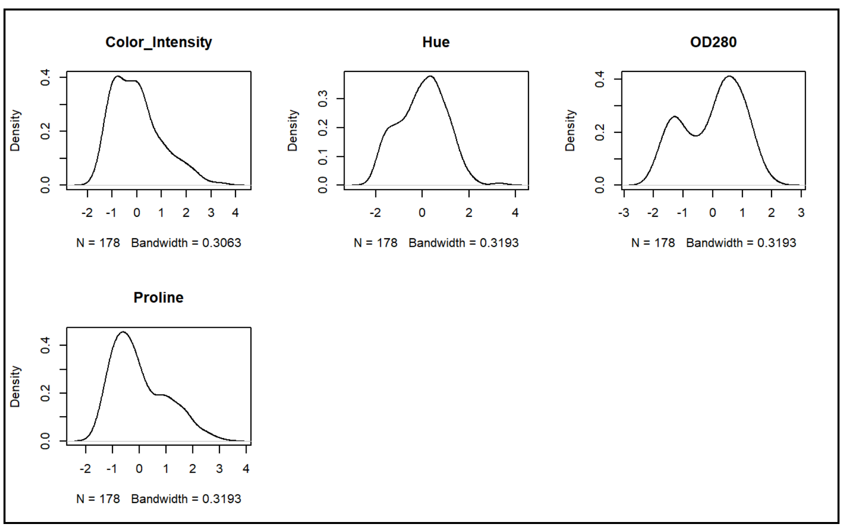

Figure 3.2. Density Plots for Scaled Variables (Cont.)

Looking at the density plots for all our variables, it is shown that the variables Malic_Acid, Nonflavanoid_Phenols, Color_Intensity, and Proline all show normal distributions with heavy right-skewness. Furthermore, the variables Total_Phenols, Flavonoids, and OD280 show bimodal histograms by a slight margin. Although some of our variables show skewness, we can use a log transformation to the variables with right-skewed histograms and see if that will help lessen its severity.

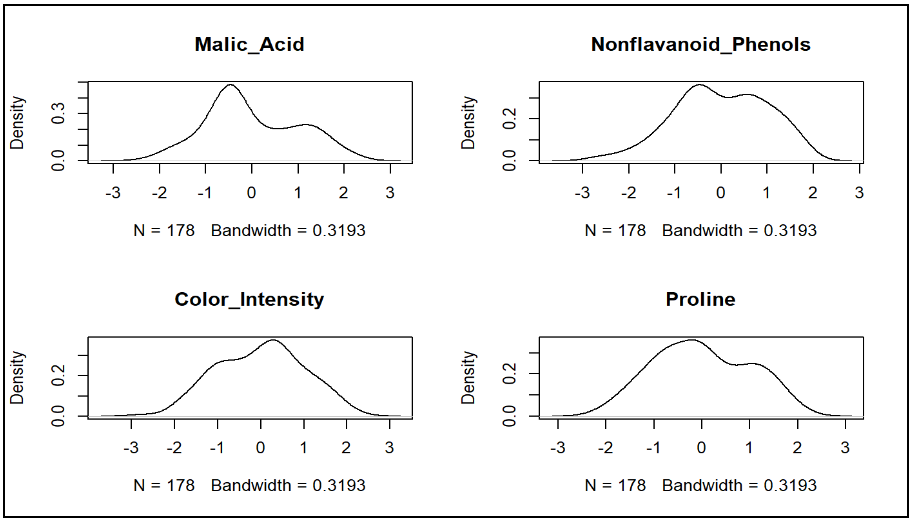

Figure 4. Density Plots for Transformed Variables

As seen in Figure 4, applying log transformations did help lessen the severity of the right-skewness for the scaled variables Malic_Acid, NonFlavonoid_Phenols, Color_Intensity, and Proline. Therefore, we will proceed to move forward with the transformed scaled variables in our data set.

PCA and Clustering Method

Continuing with the number of principal components to use, we can observe from the Scree Plot that by the "elbow" rule, we should utilize the first 4 principal components. Furthermore, the proportion explained by the first 4 principal components has the value of 0.7360, meaning 74% of our observations can be described by the first four PCs. To describe the minimum of 90% of observations, we would need to utilize the first 7 principal components.

var_explained <dbl>	cumsum.var_explained. <dbl>
0.349560518	0.3495605
0.204489486	0.5540500
0.112539175	0.6665892
0.069428502	0.7360177
0.067978296	0.8039960
0.045755574	0.8497515
0.042357477	0.8921090
0.026665784	0.9187748
0.023959996	0.9427348
0.019741640	0.9624764

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Alcohol	0.0000	0.4670	0.0000	0.0000	0.0000	0.0000	0.0000
Ash	0.0000	0.3061	-0.6191	0.0000	0.0000	0.0000	0.0000
Ash_Alcanity	0.0000	0.0000	-0.6081	0.0000	0.0000	0.0000	0.0000
Magnesium	0.0000	0.0000	0.0000	-0.4011	0.6667	0.0000	0.3407
Total_Phenols	-0.4044	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Flavanoids	-0.4283	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Proanthocyanins	-0.3235	0.0000	0.0000	0.3616	0.0000	0.6049	0.3278
Hue	0.0000	0.0000	0.0000	-0.4455	0.0000	0.0000	0.0000
OD280	-0.3794	0.0000	0.0000	0.0000	0.0000	-0.3288	0.0000
logMalic_Acid	0.0000	0.0000	0.0000	0.5286	0.0000	-0.4746	0.4646
logNonflavonoid_Phenols	0.0000	0.0000	0.0000	0.0000	-0.5657	0.3131	0.5362
logColor_Intensity	0.0000	0.5334	0.0000	0.0000	0.0000	0.3156	0.0000
logProline	0.0000	0.3833	0.0000	0.0000	0.0000	0.0000	0.0000

Figure 5. Proportion explained by the first 10 Principal Components; PCs with Variables

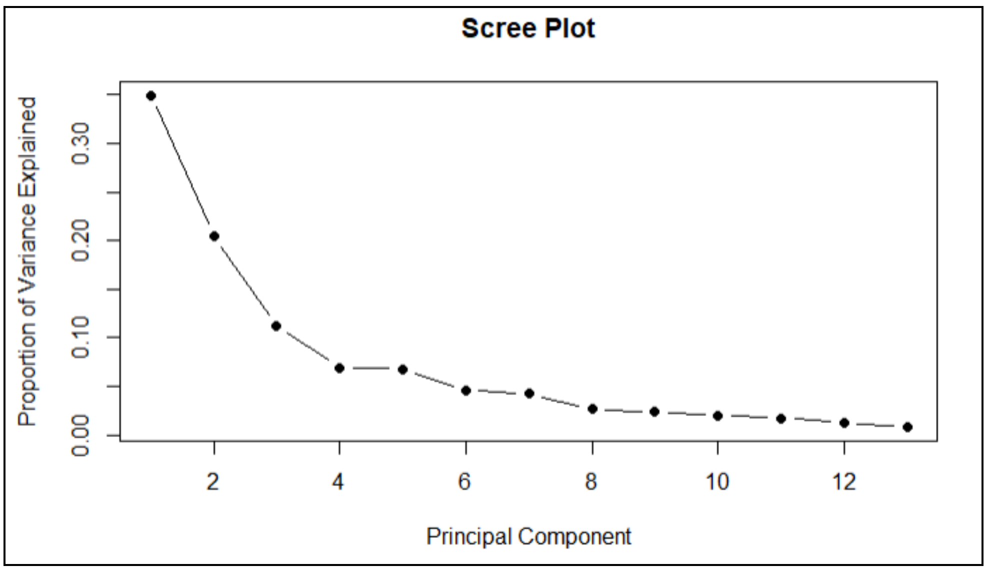

Figure 6. Scree Plot, Finding First j^{th} Principal Components

We can also observe by plotting the PCs with different variables, the first and second principal components make for a great visualization of separation between the variables Alcohol and Total_Phenols.

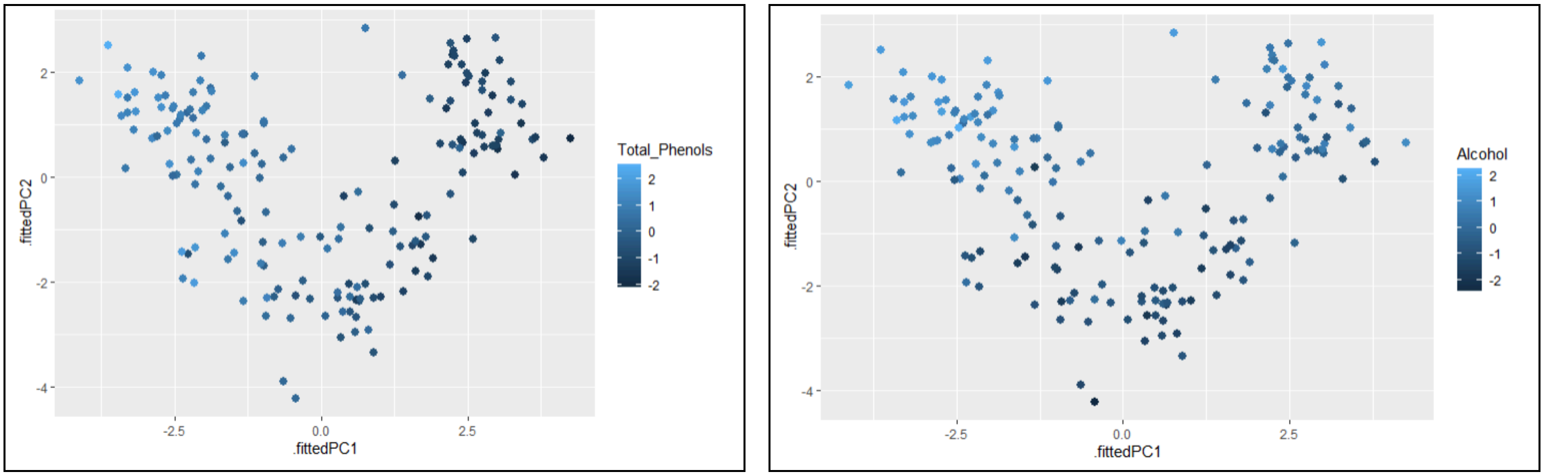

Figure 7. Principal Components 1 and 2 Plotted for Total_Phenols and Alcohol

As shown in the graph for Total_Phenols, there are fewer levels of Total_Phenols for the first principal component indicated to the right of the graph, while on the left side, there are fewer values attributed to the first principal component. For the graph with Alcohol, we can see that a higher alcohol content is closely aligned with a higher second principal component, whereas the lower alcohol content is found more with the lower values of the second principal component. This helps illustrate that PCA can be a good choice to use for our clustering method, especially when trends in the first and second principal components are showing clear trends.

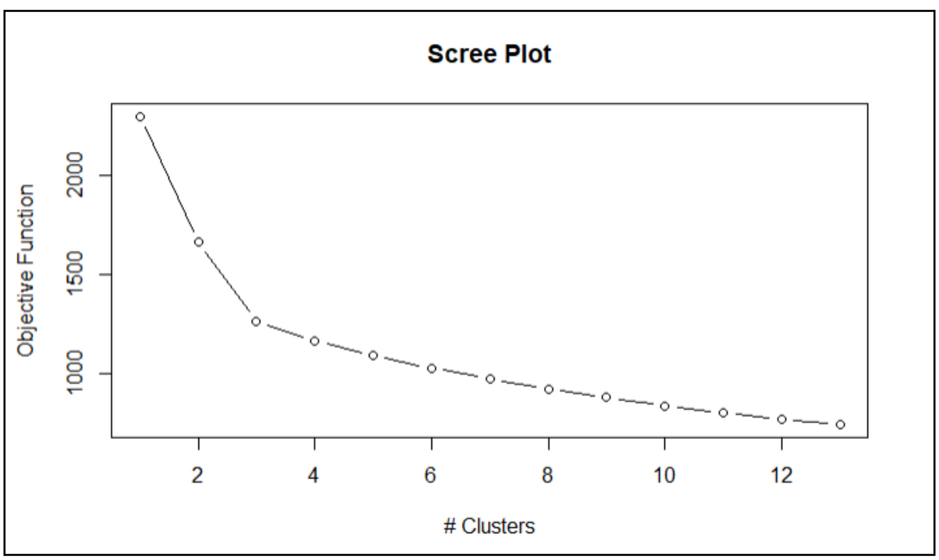

Figure 8. Scree Plot for Number of Clusters

Notice that by the “elbow rule”, a total of 3 clusters will be necessary. We may also consider the number of clusters to be 2 as well, but further analysis tells us that three clusters help create distinct wine categories that are drastically different from one another. Therefore, keeping the number of clusters to 3 will help determine further analysis for categorizing.

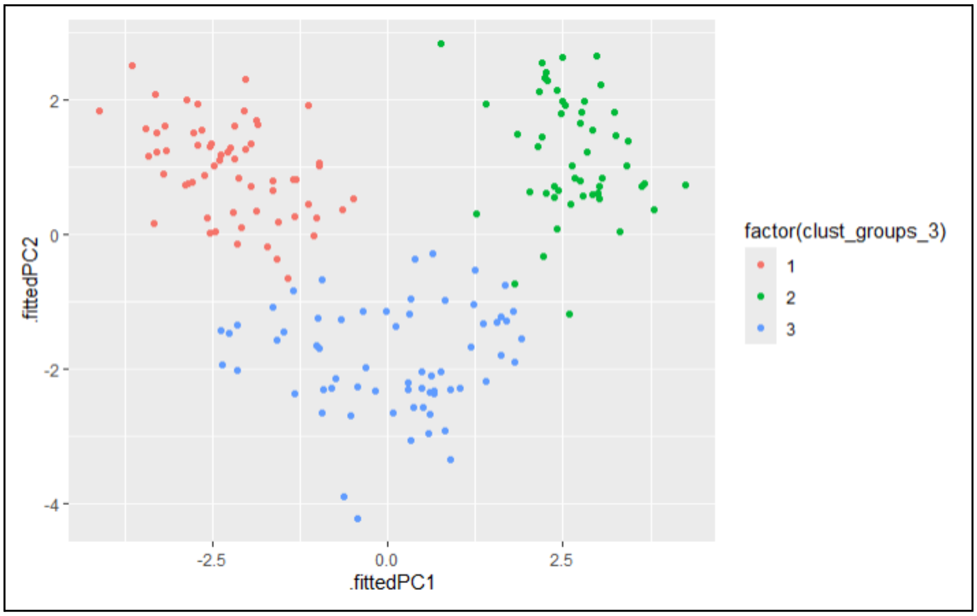

Figure 9. PC1 and PC2 with Three Clusters

To help further visualize the data when doing 3 clusters, we can see from Figure 9 a clear distinction in the boundaries in all three clusters for the first two principal components. PC1 and PC2 show us that the variables Total Phenols, Flavanoids, Proanthocyanins, OD280, Alcohol, Ash, Color Intensity, and Proline can be used to distinguish the clusters. The first PC1 describes the “Mouth Feel” as these variables consider the bitterness and dryness of wine. PC2 describes the “Taste” as these variables consider the main ingredients that affect the taste of a wine. As shown from the graph, Cluster 1 has a small PC1 and a large PC2. This can be interpreted as the cluster having a weak Mouth Feel and strong Taste. Equivalently, this wine does not have a strong bitter or dry taste to it while having a good amount of alcohol in it that is balanced by a sweet and crisp taste. Cluster two can be seen as having a large PC1 and PC2. This cluster seems to have a strong Mouth Feel and strong Taste to it as well. This wine can best be described as being bitter and dry while also being high in alcohol and having a sweet crisp taste to it. Lastly, cluster 3 has a balanced PC1 and a small PC2. The results suggest the cluster can range from a mild to moderate to strong bitter and dry feel to the wine and lower amounts of alcohol and sweetness to it. Given this overview of the data, a further evaluation can be done to better understand how all the variables affect the cluster groupings.

<u>Interpretations of Cluster Groups</u>

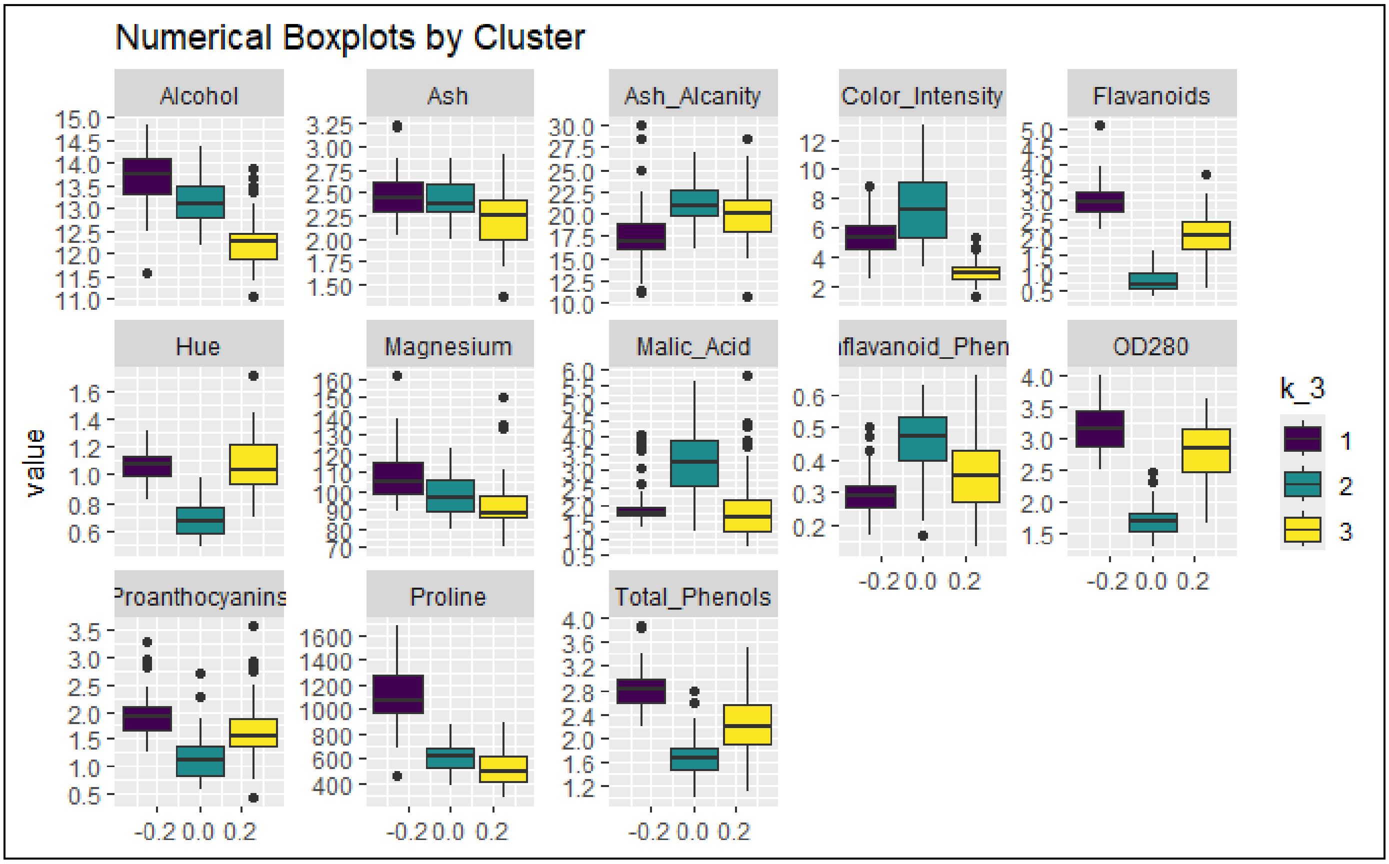

Figure 10. All 13 Variables with Cluster's Boxplots

Besides interpreting the three cluster groups with the 7 principal components, we can also show by plotting the cluster's boxplots to visually inspect which clusters show high or low attributes of all 13 variables from the project. Therefore, we classify the clusters as follows:

Cluster 1: The first cluster of wines shows higher than normal levels of alcohol, high levels of proline, high indication of proteins from the OD280 method, moderately low Ash Alkalinity, higher portion of Magnesium, high amount of Flavonoids, and high levels of Total Phenols. All in all, this cluster of wine has traits that make it taste that's warmer and oilier, an acidic taste that makes it crisp, and refreshing, as well as being an opaque red wine, with a subtle hint of bitterness and fruitiness/sweetness to help combat it. A name that describes this category could be called “Port” wines!

Cluster 2: The second cluster has a high ash alkalinity pH, high concentration of malic acid, low indication of total phenols, a great deal of color intensity, low hue, low OD280 levels, high indication of Nonflavanoid phenols, low flavonoids, and low Proanthocyanidins. This cluster of wines can be described as a softer, rounder tasting wine with a hint of tartness, still within the red wine category, an amount of bitterness that can be neglected, with an intense dark purple/blue color acidic taste. This wine is surely one to have many satisfied if they seek out a high-acidic wine, which is why the name for this category is close to “Grenache” wines.

Cluster 3: The third cluster contains wines of low proline levels, high OD280 indication, brighter hue with low color intensity, low alcohol content, low magnesium, and malic acid. This wine can be described as a brighter, translucent bright purple/garnet wine with a bitter taste that leaves a lighter, and delicate feel with little to no sweetness. The type of wine that can closely mimic this group could be the “Pinot Noir” wines.

Conclusion

Overall, the progression for the wine data set has concluded with three different classifications of wines and labeled them with the attributes of sweet and warm wine with high alcohol concentration, tart and soft wine with a dark indigo hue, and bitter, flat wine with low alcohol concentration. Each of these clusters can help wine enthusiasts branch out to the different types of Italian wines that best suit their tastes and experiences for all sorts of occasions.

During the time of analyzing and visualizing, some factors would have contributed more to the types of wines we are dealing with, such as the ages of grapes, the amount of time to process the wine, the wine’s price per ounce/liter, and more. Many factors can also explain the rarity or quality of the grapes used in the wine to help categorize the clusters into more groups, and perhaps an even wider range of wines yet to be seen and tasted. Besides the addition of new variables, we could also have experimented by not transforming the dataset and see if our variables would have performed similarly. We could also have omitted some variables such as Flavonoids and Total Phenols, since the two variables share similar attributes. Overall, this was a great experience to apply clustering methods and classify a data set under unsupervised learning methods.