WINE CLUSTERING

Moringa School - DSPT03 - Group 3

Colleta Nandutu

Sylvia Ng'ang'a

Yves Havugimana

# Business Understanding

## 1.1. Background

Wine is an alcoholic beverage made with the fermented juice of grapes.

Technically, any fruit is capable of being used for wine (i.e., apples, cranberries, plums, etc.), but if it just says “wine” on the label, then it’s made with grapes. (By the way, wine grapes are different than table grapes).

The difference between two popular drinks, wine and beer, is that brewing beer involves fermented grains. Simply, wine is made from fruit, and beer is made from grains.

## 1.2. Defining the Question

The goal of this project is to use unsupervised learning techniques to identify wine categories. Our dataset is composed of 13 numerical physical-chemical measurements which will be used by a gaussian mixture model to identify these distinct categories. We will not assume any particular number of clusters beforehand but will rather use the silhouette score as an indicator of the best number of clusters to segment our data by.

## 1.3. Metric of Success

The project will be considered a success when we are able to cluster wines based on numerical values which are their chemical compositions and properties.

## 1.4. Experimental Design

Below are the steps taken in this analysis:

* loading the required libraries
* Loading and previewing data
* Cleaning the data
* Univariate analysis
* Bivariate analysis
* Multivariate analysis
* Modelling (K-Means)
* Challenging the solution and giving insights on how improvements can be made.

# Data Understanding

## 2.1. Data Relevance

This dataset is adapted from the Wine Data Set from https://archive.ics.uci.edu/ml/datasets/wine by removing the information about the types of wine for unsupervised learning.

The following descriptions are adapted from the UCI webpage:

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars.

The analysis determined the quantities of 13 constituents found in each of the three types of wines.

The attributes are:

* Alcohol - the alcohol content
* Malic acid - malic acid concentration
* Ash- ash content in wine
* Alkalinity of ash - water balance in wine
* Magnesium - magnesium content
* Total phenols - phenols quantity in wine
* Flavonoids - flavonoids content in wine
* Non Flavonoid phenols - non flavonoids measure
* Pro Anthocyanins - proanthocyanidins composition
* Color intensity - color pigment concentration
* Hue - coloring measure
* OD280/OD315 of diluted wines - protein content
* Proline - amino acid content composition

The dataset has a total of 178 records and 13 columns.

## 2.2. Data Quality Verification

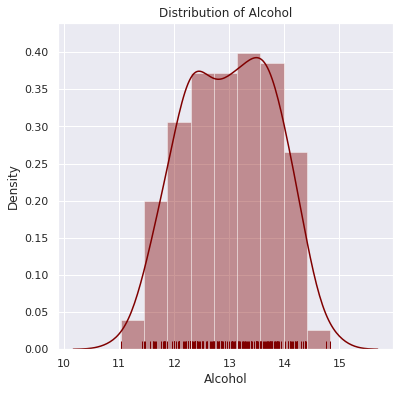
The dataset quality is very good:

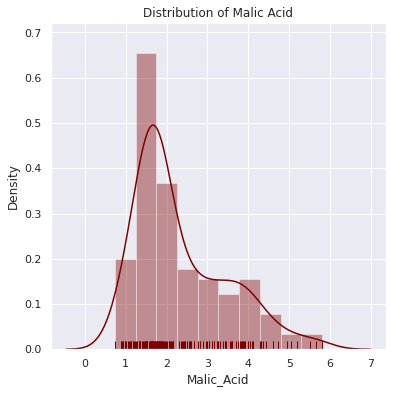
* The naming of the variables is clear and concise
* The dataset contains no missing values
* The dataset has no duplicate records
* The dataset shows very few outliers in the following variables:
  + Malic\_Acid, Proanthocyanidins and Hue : 1 outlier each,
  + Ash\_Alcanity and Color\_Intensity: 2 outliers each and
  + Magnesium: 4 outliers.

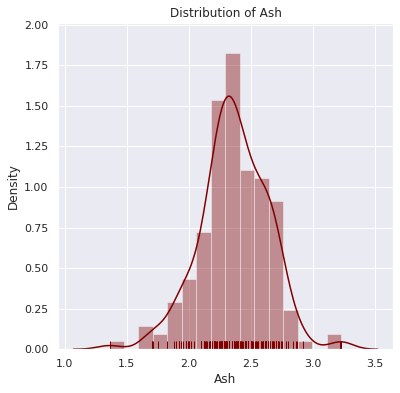
# EDA

Within the EDA, we perform Univariate, Bivariate and Multivariate Analysis.

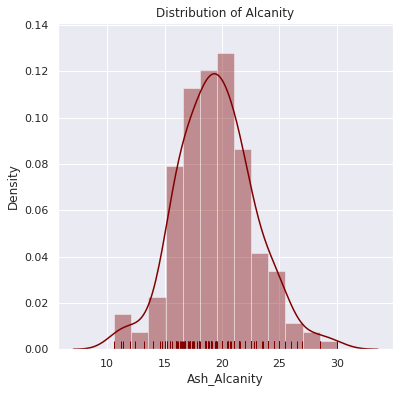
## 3.1. Univariate Analysis

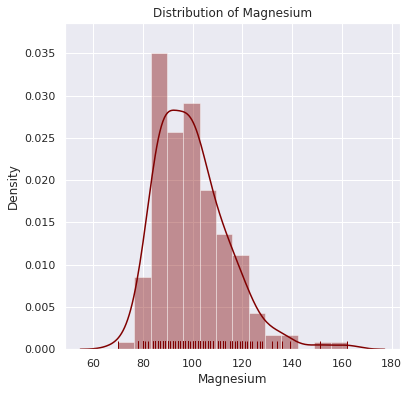


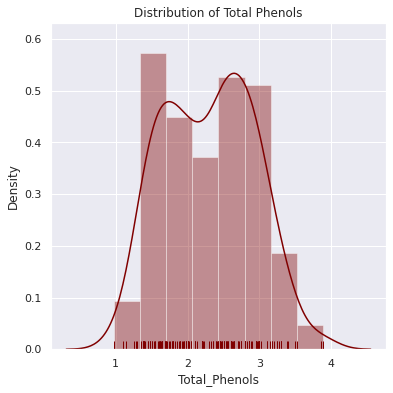


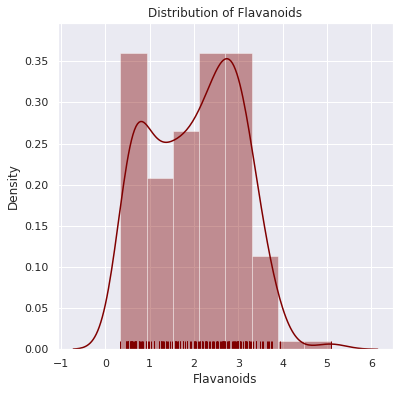


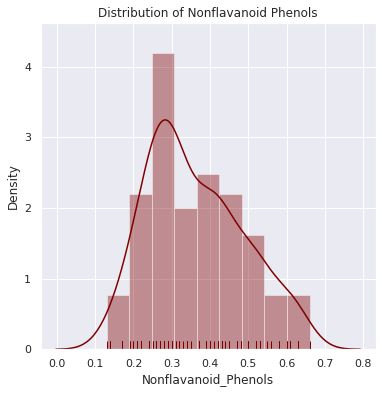
3.1.

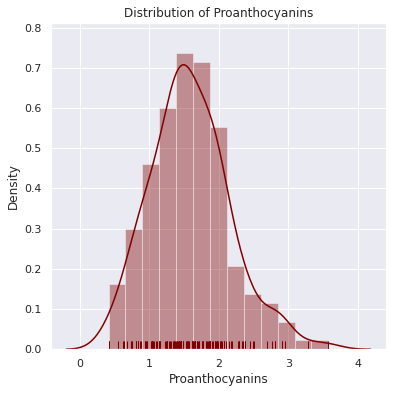


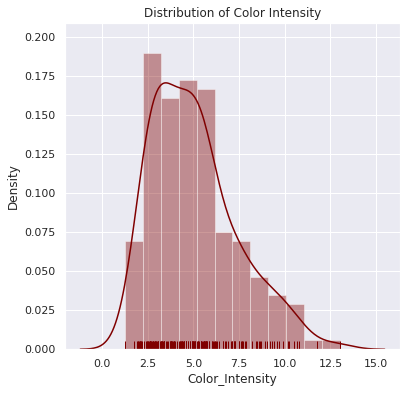


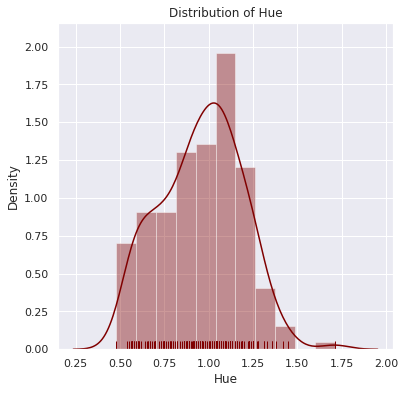


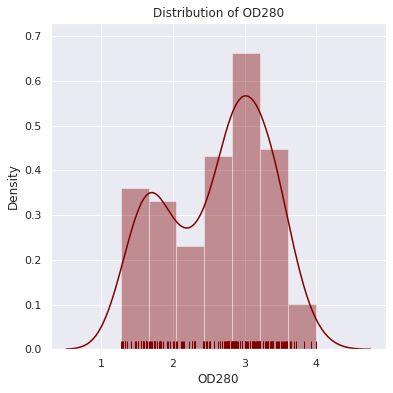


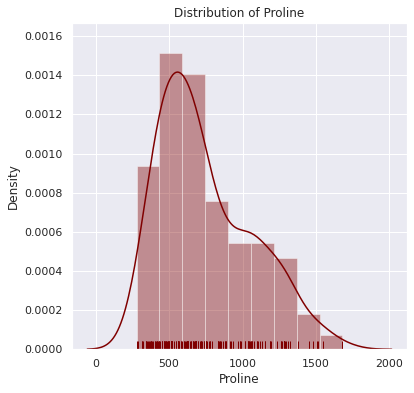




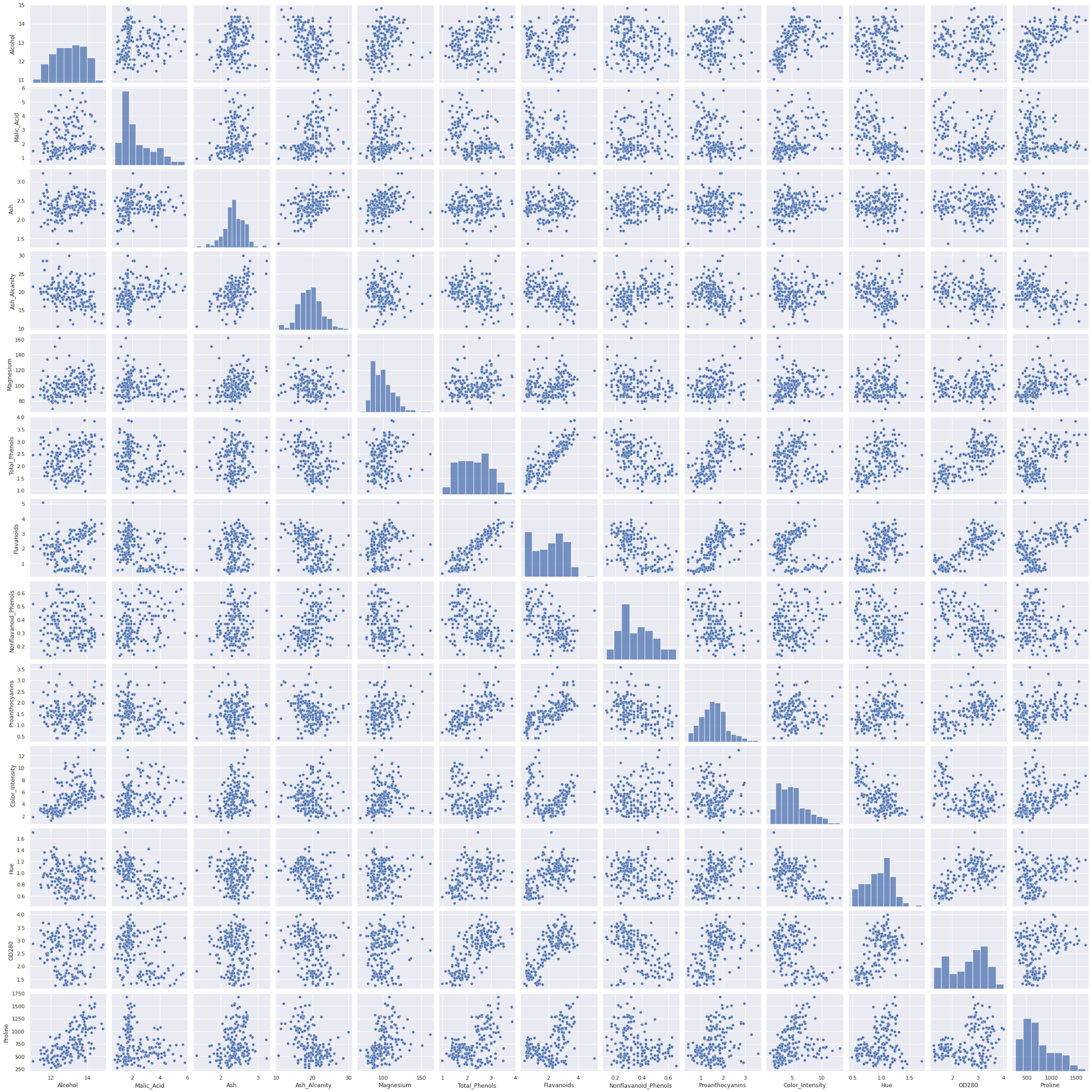






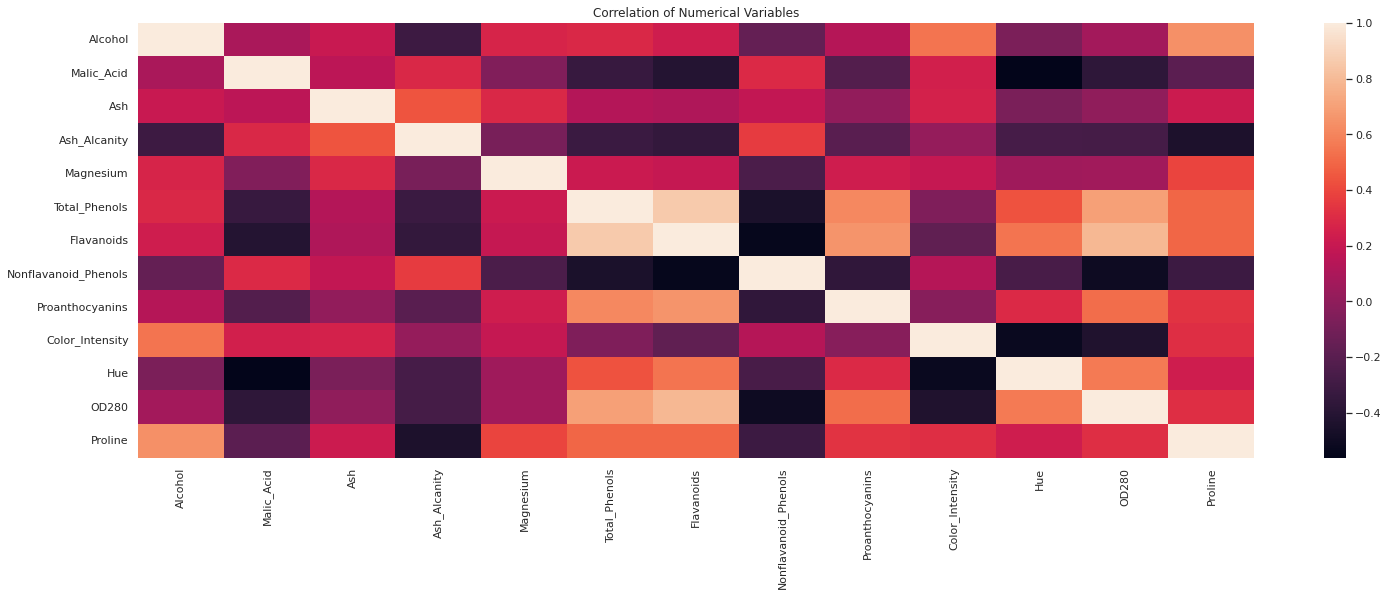


## 3.2. Bivariate Analysis



We can observe some of the scatterplots showing upward trends, where an increase in one variable increases the other.

## 3.3. Correlation Matrix

“ Flavonoids” show a high correlation (above 0.8) with “ Total Phenols” and a moderate correlation (0.6 to 0.8) with the variable “ OD280“

# Modeling

## 4.1. Normalization

Using MinMaxScaler, we normalize our data and hence prepare it for typical machine learning algorithms.

## 4.2. Principal component analysis (PCA)

PCA is a technique in which a set of normal axes are greedily chosen such that each axis is oriented such that variance is maximized along that axis. We perform PCA in order to determine several things:

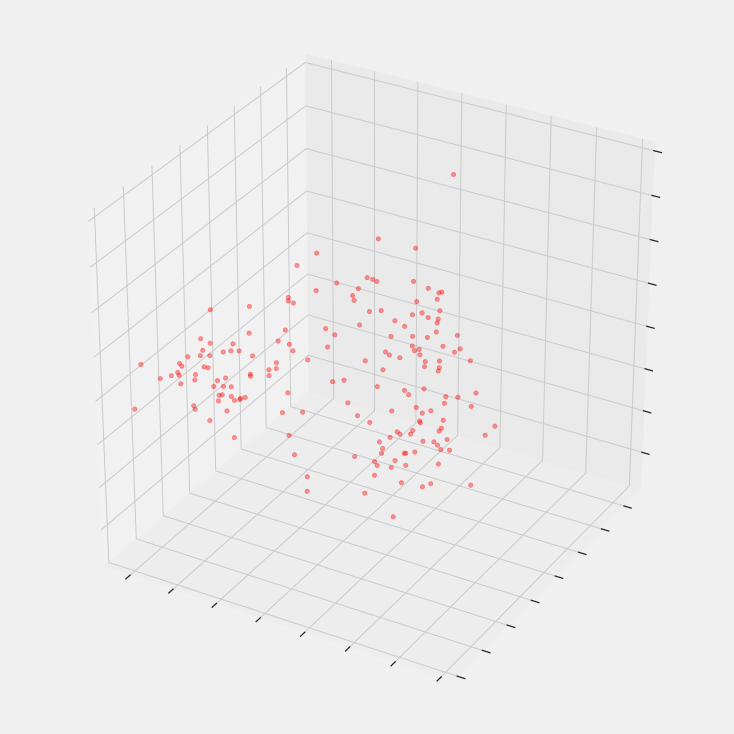
* 1) The minimum number of dimensions necessary to preserve the majority of the information (variance) contained in our data. This is of interest to use if we decide to use fewer dimensions to our ML algorithm in the interest of boosting performance.
* 2) This helps us determine whether a 3D visualization will be sufficient to see some of the inherent structure in our data.

We will perform a PCA transformation using the same number of dimensions as our existing dataset. We will then calculate the percentage of the overall variance that can be encoded in 1, 2, ... n dimensions. We will then examine the first component of our PCA transformed data to identify which metrics account for the most variance.

We apply PCA by fitting the data with the same number of dimensions as features. As a result we have the following:

* Percent variance explained by first 1 components: 39.3%
* Percent variance explained by first 2 components: 60.1%
* Percent variance explained by first 3 components: 69.0%
* Percent variance explained by first 4 components: 75.9%
* Percent variance explained by first 5 components: 81.3%
* Percent variance explained by first 6 components: 85.9%
* Percent variance explained by first 7 components: 89.4%
* Percent variance explained by first 8 components: 92.0%
* Percent variance explained by first 9 components: 94.5%
* Percent variance explained by first 10 components: 96.5%
* Percent variance explained by first 11 components: 98.0%
* Percent variance explained by first 12 components: 99.2%
* Percent variance explained by first 13 components: 100.0%

A visualization of our PCA transformed data:



## 4.3. Clustering

We will use an analytical approach before performing our clustering task rather than assuming that 3 clusters is the most reasonable number of groups.

This analytical will be done as follows:

1) We will fit our clustering algorithm using several different values of k where k is the number of clusters. We will have k range from 2 to 20.

2) For each value of k, we will evaluate the clustering results using the average silhouette score. We will plot the silhouette score against each k value and identify which number of clusters leads to the best results.

3) We will then assign cluster labels to our dataset using the fitted clustering model with the optimal number of clusters k.

The silhouette score can be roughly described as a measure of how close a sample is to members of its own cluster as compared to members of other clusters. The silhouette score ranges from -1 to 1.

A score close to one indicates that a record is very close to other members of its cluster and far from members of other clusters.

A score of 0 indicates that a record lies on the decision boundary between two clusters.

A negative score indicates that a sample is closer to members of a cluster other than its own.

By taking the average silhouette score for all records when various number of clusters are used in our clustering algorithm, we can find the optimal number of clusters that promotes cohesion within individual clusters and good separability between the clusters.

### 4.3.1. Gaussian Mixture Clustering

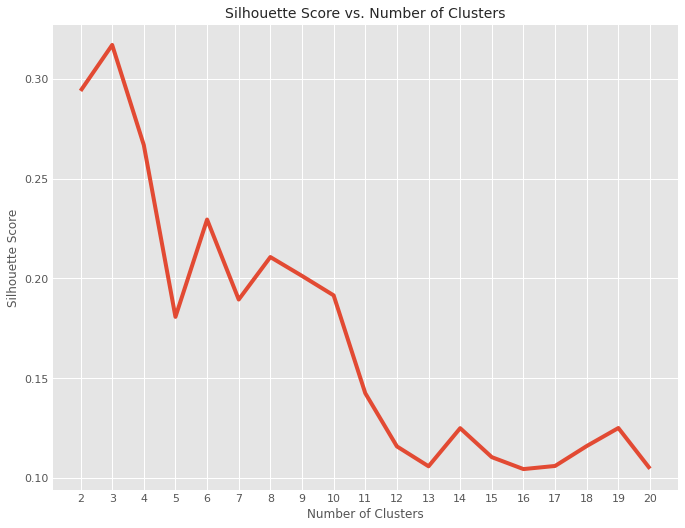
For the clustering algorithm itself, a Gaussian Mixture Model was chosen. This was chosen for several reasons, including the fact that gaussian mixture models allow for mixed membership; GMM models assign probabilities that a given record belongs to a given cluster. This property may be useful for classifying wines which are blends of multiple types.

Another reason for choosing a GMM is that they are more flexible with regards to cluster shapes which deviate from a hyper-spherical one. It is impossible for us to directly observe the actual cluster shapes visually since they exist in a 13-dimensional space, so it is helpful to have a clustering algorithm with such flexibility.

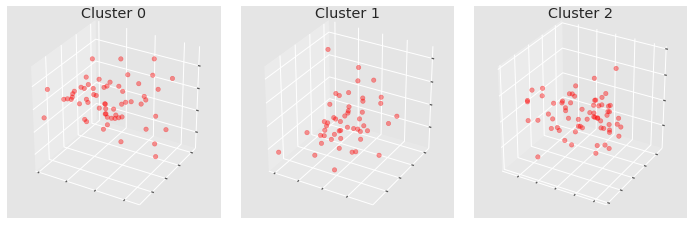
After building and applying the Gaussian Mixture model to our data, we have the following:

* The highest silhouette score is 0.317
* The maximum number of clusters is 20
* The first three silhouette scores are:
  + 2 components: 0.294
  + 3 components: 0.317
  + 4 components: 0.267

In order to have a look at the silhouette score in comparison to the number of clusters, we have a look at the graph below:



We see that the silhouette score is maximized when 3 clusters are used. Therefore, we fit the GMM using 3 clusters. We plot the three clusters below.



On the left, we have cluster 0, on the right we have cluster 2 and in the middle, we have cluster 1.

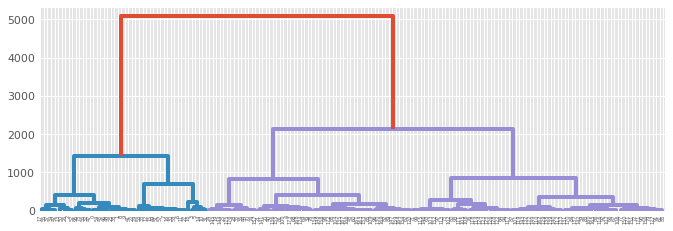
A look at the different clusters and the average values we notice differences:

* The average Alcohol content is highest in cluster 0 and lowest in cluster 2.
* The average Ash, Magnesium and Proline content follows the same trend as Alcohol
* Malic Acid, Ash Alcanity, Nonflavanoid\_Phenols and Color Intensity show higher averages in Cluster 1.

| Cluster | Cluster 0 | Cluster 1 | Cluster 2 |
| --- | --- | --- | --- |
| Count | 61 | 50 | 67 |
| Alcohol | 0.699 | 0.54744 | 0.311316 |
| Malic\_Acid | 0.53 | 0.754012 | 0.457892 |
| Ash | 0.545 | 0.528521 | 0.432834 |
| Ash\_Alcanity | 0.378 | 0.588403 | 0.526674 |
| Magnesium | 0.631 | 0.536055 | 0.461313 |
| Total\_Phenols | 0.682 | 0.279881 | 0.487069 |
| Flavonoids | 0.622 | 0.136898 | 0.43639 |
| Nonflavanoid\_Phenols | 0.423 | 0.70225 | 0.552469 |
| Proanthocyanidins | 0.567 | 0.321092 | 0.476879 |
| Color\_Intensity | 0.595 | 0.709898 | 0.336621 |
| Hue | 0.494 | 0.178061 | 0.486708 |
| OD280 | 0.655 | 0.130482 | 0.51397 |
| Proline | 0.772 | 0.459704 | 0.335533 |

Table showing clusters and their average contents.

### 4.3.2. Hierarchical Clustering



Hierarchical Clustering shows there are three clusters.

# Conclusions

1.Cluster 0 wines have a higher mean alcohol content and thus can be classified as dry wines.

2.Cluster 2 has the lowest alcohol content and thus will be sweeter that cluster 0 and 1 wines.

3.Given that cluster 1 wines have a higher Malic acid content,they will taste abit more sour as compared to cluster 0 and 2 wines despite cluster 0 wines being drier than cluster 1.