**Machine Learning Unsupervised Learning 1**

## Individual Assignment – 2

**Paritosh Sinha**

**As mentioned in the problem statement -**

#### **Objective**: To identify distinct clusters of wines based on their chemical composition By clustering the wines based on these attributes, we can uncover hidden similarities and differences amongst the wines based on its Chemical composition.

**Entities being clustered**: Here we are clustering Wines samples

**Attributes on which clustering is being done on**: clustering is done on the basis wine’s chemical composition

**In order to achieve**: It can help achieve better production, quality checks and target customers and decide the price of the Wine as per its composition.

# Answer a.

Enumerate the insights you gathered during your PCA exercise. Please do not clutter your report with too many insignificant insights as it will dilute the value of your other significant findings.

1. Principal Component Analysis (PCA) is performed majorly for Dimensionality Reduction: As we are having 13 dimensions. After doing the PC it has come to light that for covering 13 variables if we go with two components, we will be losing significant amount to information as we are getting coverage of only 55.4% of our dimensions with 2 PCs which states that there is huge variation in our data.
2. Explained Variance: PCA scores of components also suggests that PC1 and PC2 variables do not contribute much to the variance in the data. All the individual PC scores indicates that PC1 captures disproportionally more variance than others, which indicates variables is conveying same underlying factor or do not add additional variables, they are conveying same information from marginally varying angle.

**PC1** – factors like Flavanoids, Total Phenols, Proanthocyanins, OD280/OD315 of diluted

wines, Proline have a strong positive contribution in the derivation of it. Higher levels of the mentioned dimensions are associated with higher values of PC1.

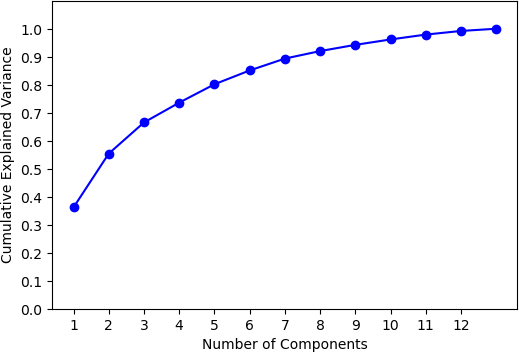
On the other hand Malic\_Acid, Alcalinity\_of\_ash, Nonflavanoid\_phenols have a negative coefficients which suggests negative contribution to it. In the same manner all the PCs can be studied for the relation and impact of different dimensions on the individual PCs

1. Variance explained by each of the all\_components: As the number of Principal Component increases more variance in the data is captured. Below data explains the values of PC as the number of Principal Components increases.

[0.36198848 0.1920749 0.11123631 0.0706903 0.06563294 0.04935823 0.04238679 0.02680749

0.02222153 0.01930019 0.01736836 0.01298233 0.00795215]

### Number of PC and Cumulative Variance Explained



1. Feature Importance: By look at the biplot of our PCA data it is observed ***–***
2. Vector length: Longer the vector indicates higher importance. Here it is evident that Color intensity, Alcohol plays an important role.
3. Dimension Correlation: Smaller the angle between two dimensions higher the correlation between them. Like Total phenols and Flavanoids same is inferred from correlation matrix as well same is the case for Proanthocyanins and Total Phenols.

Similar observation between Hue and Malic Acid showing negative correlation.

Same can be observed in the biplot below.

5. Data Visualization: For data visualization in 2 dimensions, we select PC1 and PC2 whose variance explains 55.4% of information. On plotting this data against class variable, it can be observed that

Class 1 – Is showing maximum variability as its spread is very high and it is even overlapping Class 2 and Class 3 where Class 2 and Class 3 are overlapping each other when explained

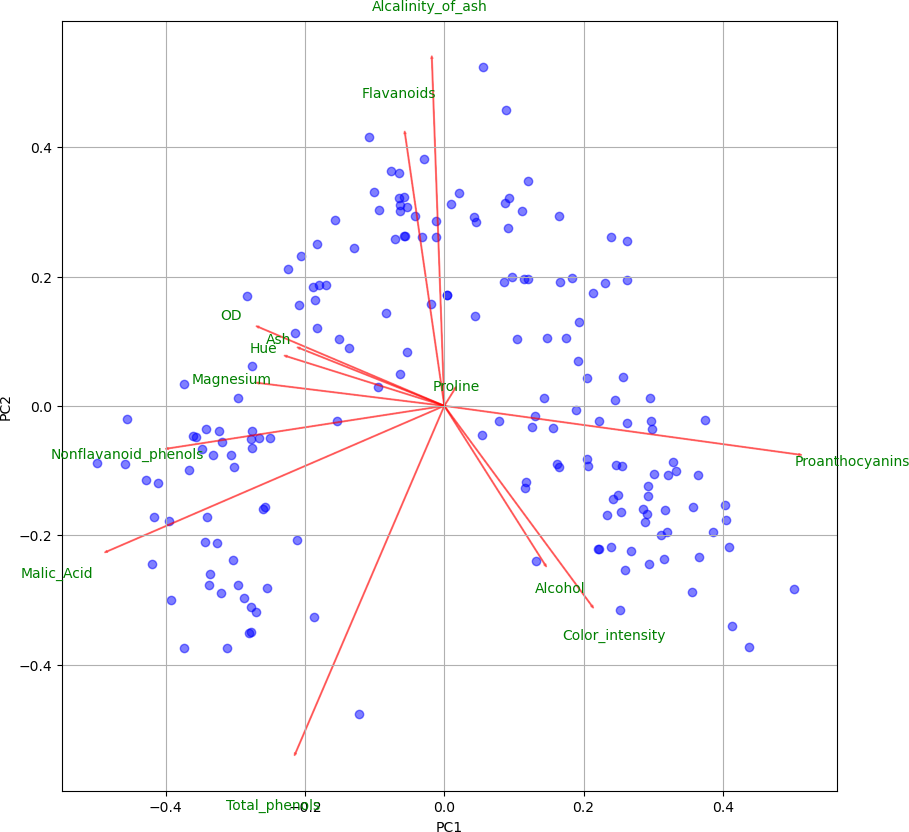
with the help of PCs

Variable influenced by Dimensions as observed in Biplot: The plotting of the observations on the

dimension vectors represents the contribution of each dimension for that observation. Observations with higher coordinates on a specific variable have a stronger influence from that dimension. From biplot it is observed observations in our dataset for one of the clusters are influenced by Malic Acid, Non Flavanoids phenols and another cluster is majorly influenced by Alcalinity\_of\_ash, Hue.

Henceforth for other clusters can be studied.

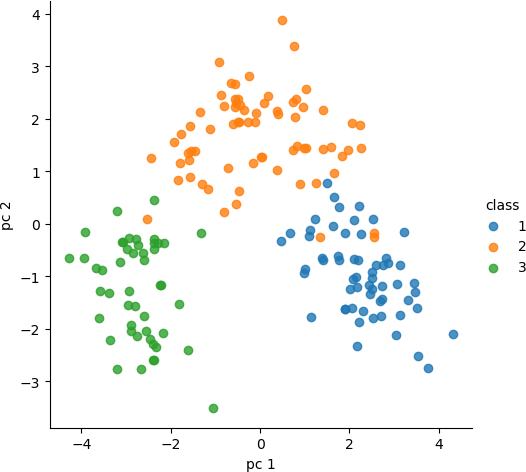
### Biplot of PC1 and PC2



Clustering and Grouping: Clear picturisation of Clusters appear on plotting PC1 and PC2 against Class variable only a few observations are overlapping amongst the clusters. Principal Component Analysis have helped here in putting the observations into clear groups which can help identify

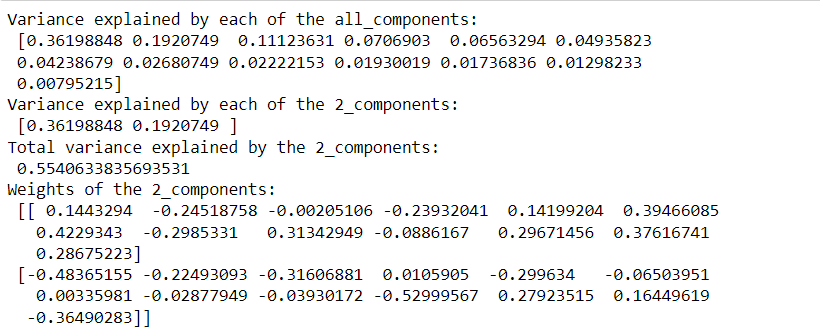
Wines with the help of its constituents.

### Plotting of PC1 PC2 with respect to Class Variable



Variance explained by each of the 2\_components:

[0.36198848 0.1920749]



***\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\****

# Answer b.

Insights harnessed from studying the composition of wine can benefit multiple stakeholders from Wine Cultivars, Wine Researchers, Business selling and buying wines, Stockists and lastly the consumers.

1. Wine cultivars can gain insight from this data, based on the study they can focus and grow a particular kind of grape which produces the best kind of wine or wine which is more in

demand by the market. They can alter the process of cultivation, change the breed of the grape, focus of grafting, budding, propagation of seeds, select proper fertilizer and crop care etc.

1. Wine Researchers insights on composition of all influencing factors can be used to better understand the science of winemaking. Like researchers can use this information on the

chemical constituents of wines to study the impacts of different grape varieties grown by Wine Cultivars, fermentation techniques and impact aging conditions on the flavour, colour, hue and quality of wine.

#### Business buying and selling wines like Hotels, Supermarkets, Stockists based on the

demands of a particular kind(cluster) business can map that demand to a particular clusters chemical composition and suggests wine producers on producing more of what is being in demand and suggest alterations if any required.

1. Wineries - Wine makers can gather insights from researchers, businesses, customer

feedbacks on various wines

Based on the study of Chemical composition many decisions can be make like Grape selection, Vinification (crushing, fermenting, aging, storing in the barrels), Blending of other chemicals during production.

Devise improvement plans

Decide on costs and categorisation of the wines.

1. Customers - Can use insights on the chemical composition of wines to find wines that match their personal preferences.

Insights can be used on the flavour profiles of different wines to compare different wines before making a purchase.

Insights on the Chemical composition can help find health benefits of wine to make informed decisions about their wine consumption.

1. Wine connoisseurs – Understanding of chemical compounds which affects wine’s taste and fragrance to better identify and distinguish variety of wines. This can enhance one’s experience. This analysis or clustering can also help them to better pair variety of foods with different types of wines as per their texture and taste.

Wine connoisseurs can gain a better insight of the wine industry and the factors that drive the quality and price of wine.

***\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\****

# Answer c.

During clustering exercise, we came to know the following:

I tried different types of clustering techniques like K-Mean with PCA, K-means without PCA and

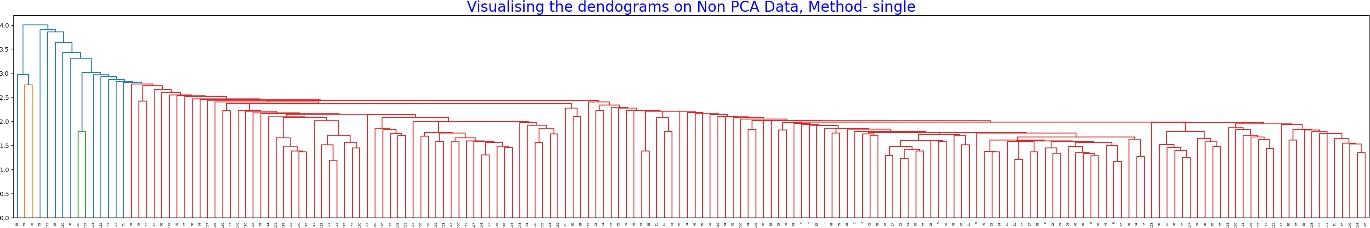
Hierarchal clustering with various options of linkages with Euclidean distance with PCA and without PCA.

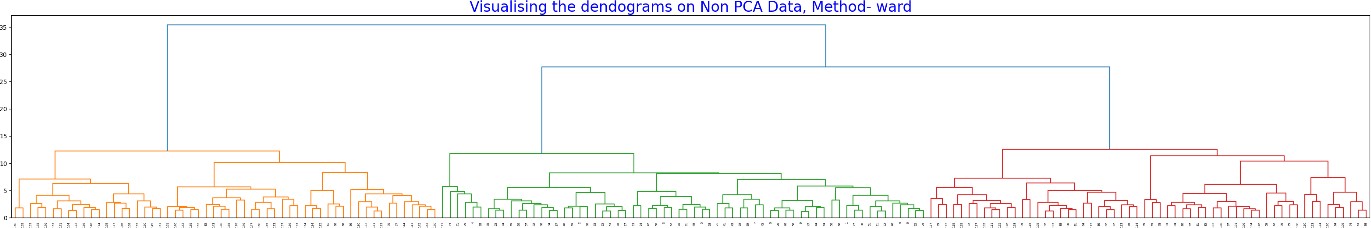
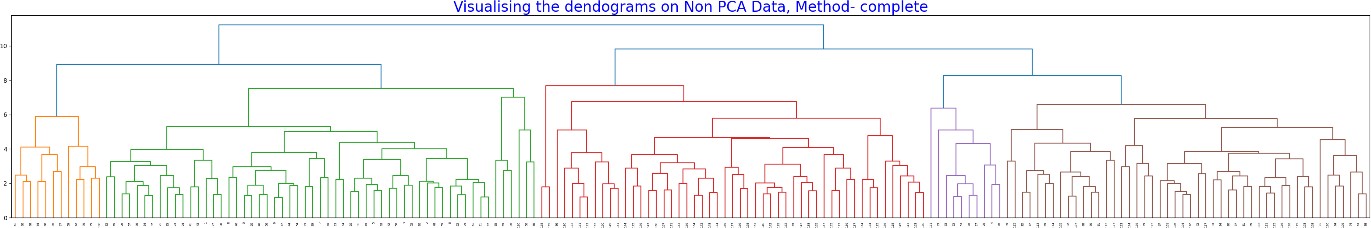
Following are my insights-

1. While doing Hierarchal clustering on Non PCA data I deduced that Ward’s method very clear clusters emerged in the dendogram, As observed three clusters appeared most prominently. There we choose that for our further analysis. Average silhouette\_score also supported the s ame.
2. Where as on hierarchal clustering on PCA data I deduced that Ward as well as Complete both produced very clear clusters which appeared on the dendogram. As observed **four** clusters a ppeared most prominently using Wards.

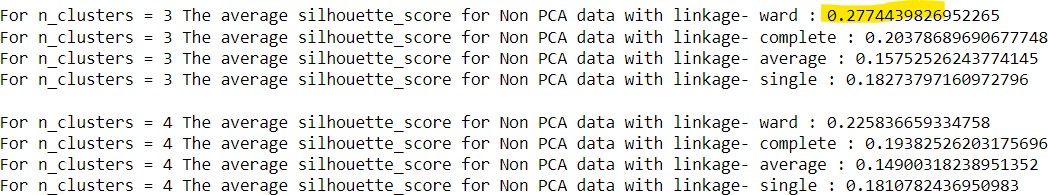
For keeping the comparison between PCA and Non PCA data I went ahead and made 3 clusters for Hierarchal Clustering using Ward’s method.

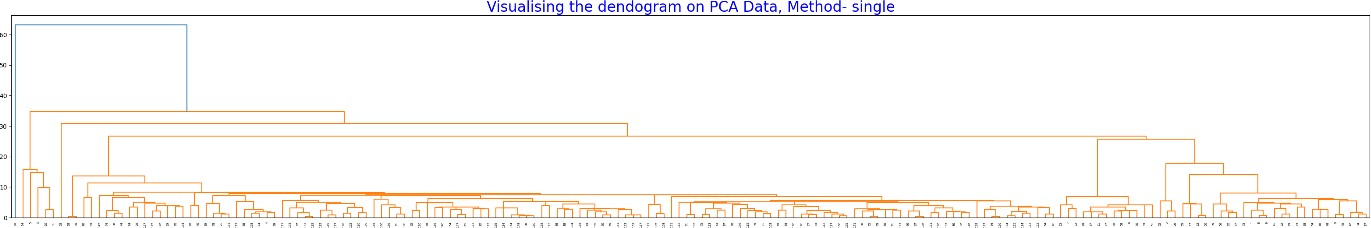
### Dendogram

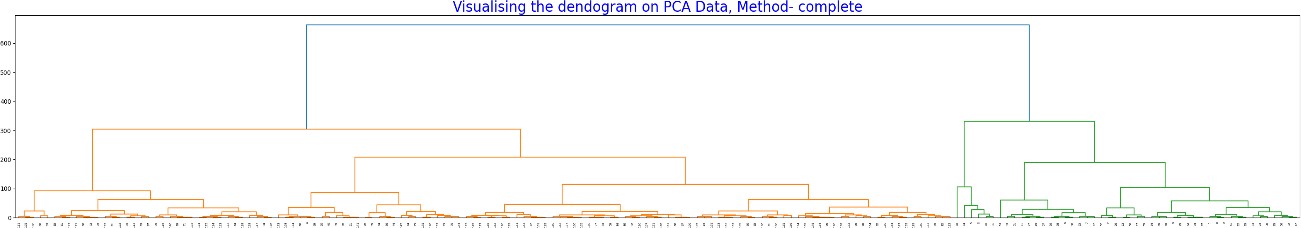


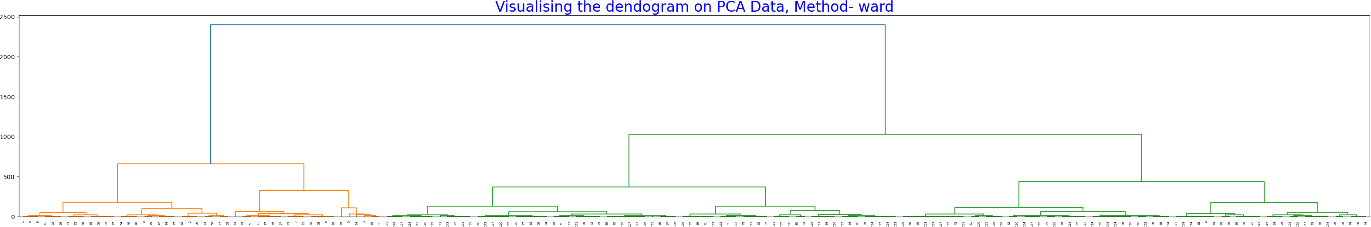


Silhouette Score

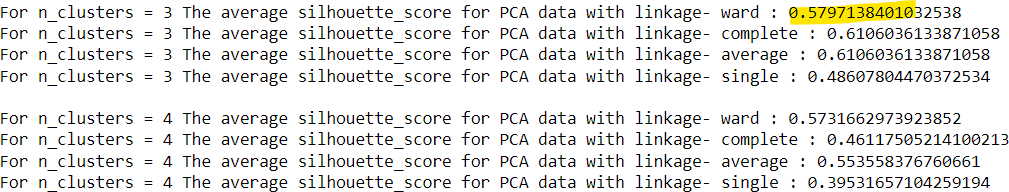




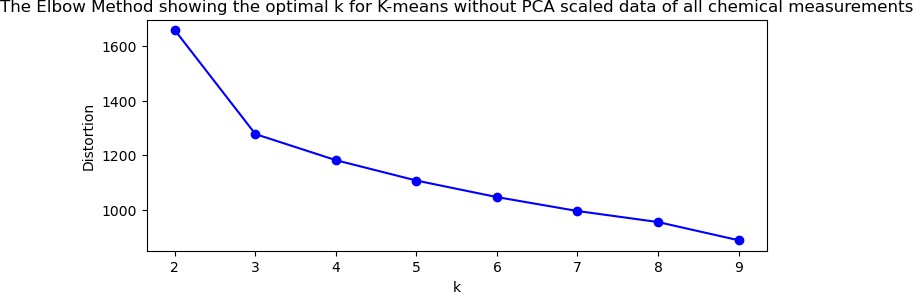




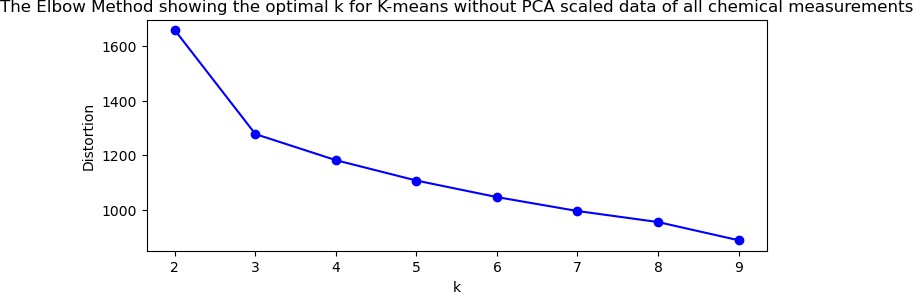
### Silhouette Score



1. While doing K-Means clustering on Non PCA data, I plotted the Elbow graph to deduce the n umber of clusters to create which is K so we choose 4 in our case.



1. Whereas on K-Means clustering on PCA data, I plotted the Elbow graph which produced a ver y clear number for K which is 4 in our case.



1. For keeping the comparison between PCA and Non PCA data I went ahead for **4 clusters** for K-Means.

##### K-Mean Inertia Reading - For PCA Data - 294156.231

**For Non PCA Data - 1175.705**

A lower value of inertia\_ indicates that the data points within each cluster are closer to their cluster centre, suggesting tighter and more compact clusters. A lower inertia value generally indicates better cluster quality.

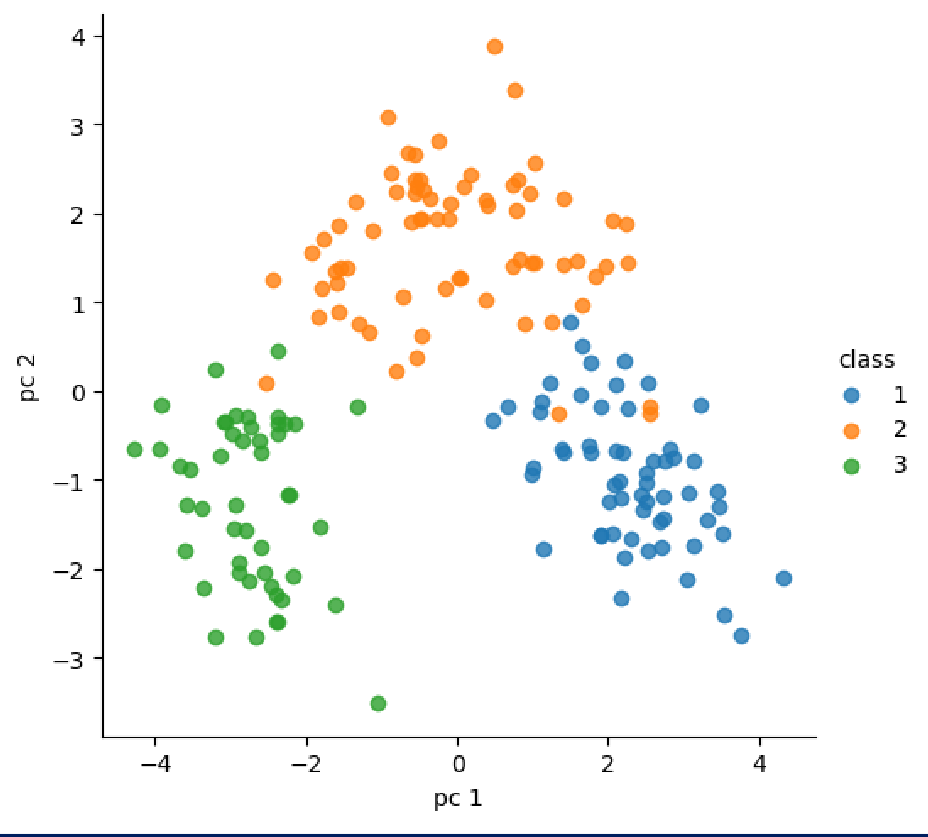
Here it can be interpreted that Cluster formed by Non PCA data are better than PCA Data. As for PCA the coverage of variance in data is only 55.4%

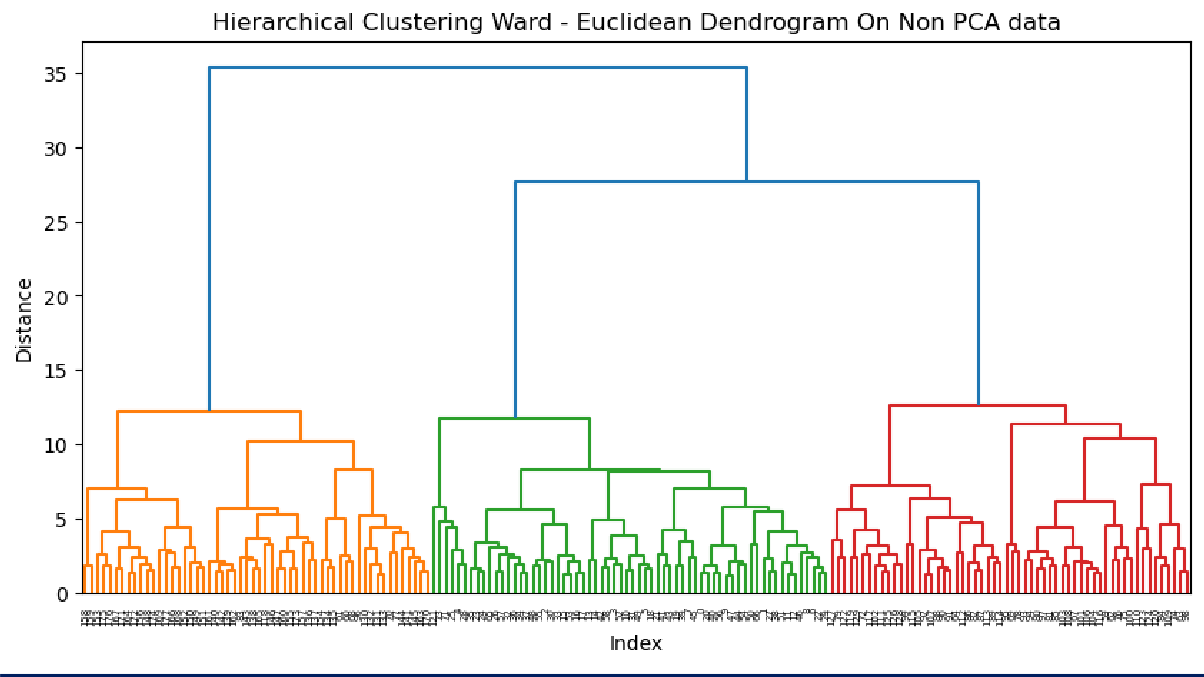
***\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\****

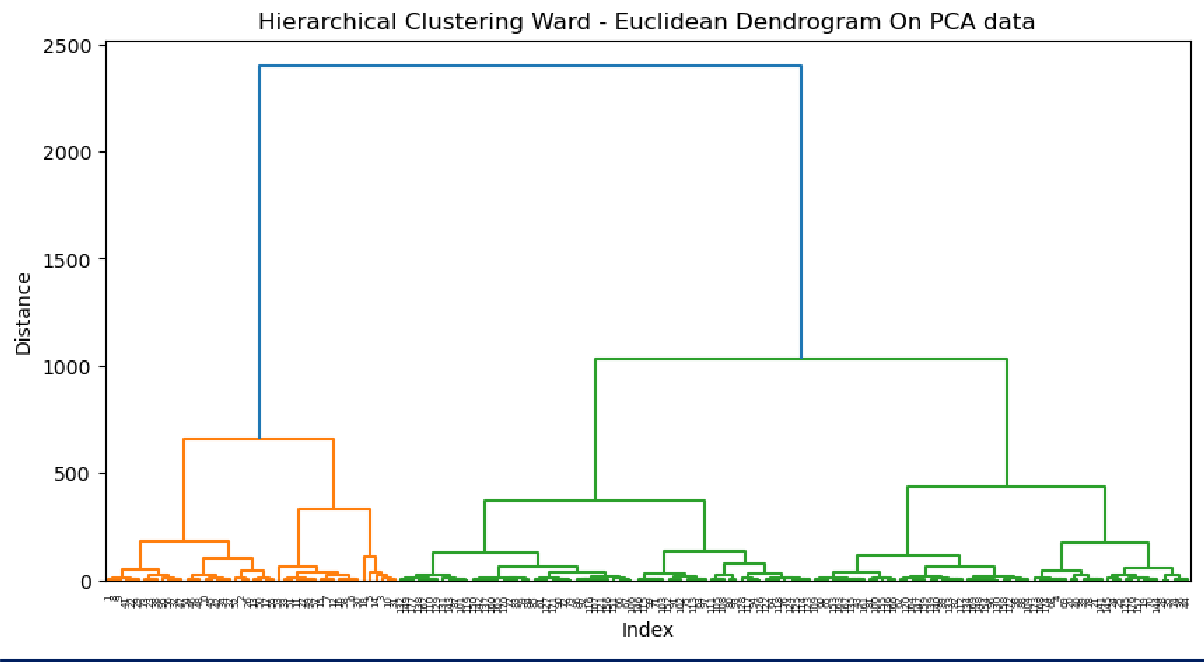
# Answer d.

As described above in part c, yes, we are getting clear separable cluster of wines, and number of clusters are varying on technique used for clustering.

### PCA – PC1 and PC2 with Class Variable

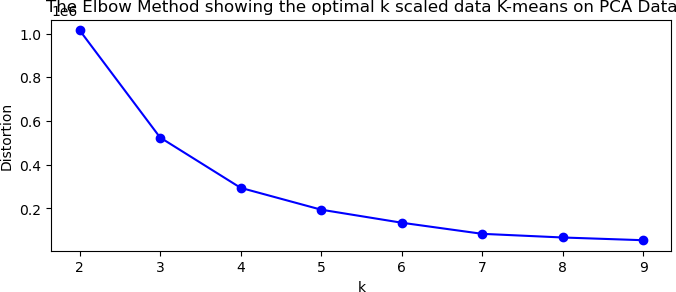


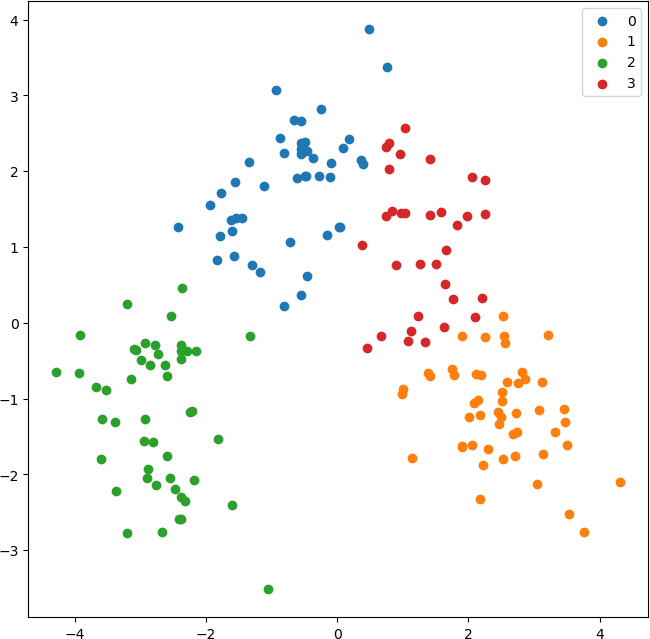




For K-Means and Hierarchal Clustering the clear cluster marking is observed. No of Clusters formed for both the cases is 4

### Implementing K-means on PCA Data



Cluster Visualization

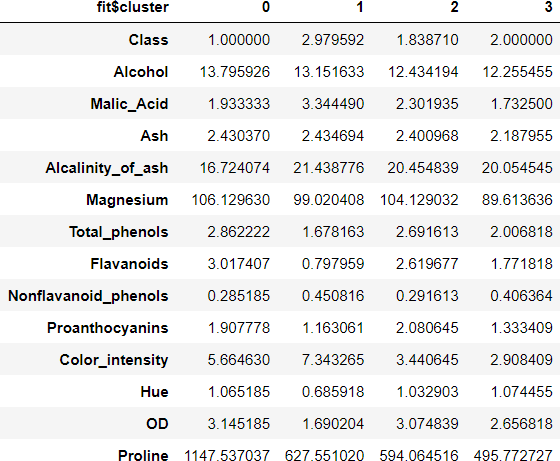
### Implementing Elbow Graph on K-means without PCA on scaled data of all chemical measurements

using (i) all chemical measurements (centroids)

### K-Means Without PCA

**Count in each cluster:**

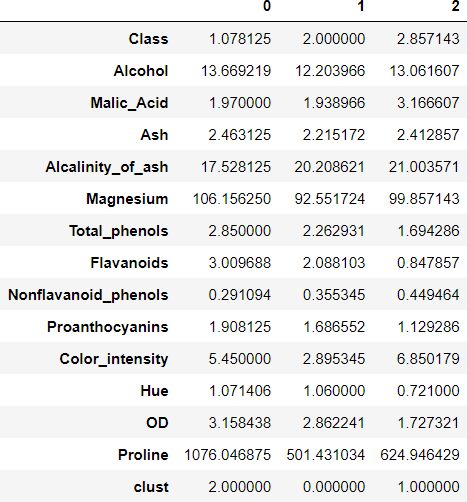
|  |  |
| --- | --- |
| **0** | **54** |
| **1** | **49** |
| **3** | **44** |
| **2** | **31** |



### Hierarchal-Clustering without PCA

**Count in each cluster: Hierarchal**

|  |  |
| --- | --- |
| **2** | **64** |
| **0** | **58** |
| **1** | **56** |

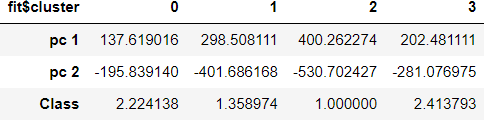


1. using two most significant PC scores (PC Weights)

### K-Means With PCA

**Count in each cluster: Kmeans**

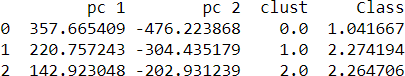
|  |  |
| --- | --- |
| **3** | **58** |
| **0** | **58** |
| **1** | **39** |
| **2** | **23** |



### Hierarchal-Clustering with PCA

**Count in each cluster: Hierarchal**

|  |  |
| --- | --- |
| **2** | **68** |
| **1** | **62** |
| **0** | **48** |



#### Qualitative differences amongst the clusters which we obtained:

* 1. **Number of Clusters formed** - K-Mean PCA – 4 Clusters

K-Mean Without PCA – 4 Clusters Hierarchal PCA – 3 Clusters

Hierarchal Without PCA – 3 Clusters

* 1. **Class Distributions in Clusters** – K-Mean PCA – 4 Clusters
     + Cluster 0 – Have combination of Class 2 & 3 where Class 2 are more
     + Cluster 1 – Mix of 1 ,2 & 3 where 2 & 3 are very less
     + Cluster 2 – Have Class 1
     + Cluster 3 – Mix of 1 , 2 & 3 and all are equal K-Mean Without PCA – 4 Clusters
     + Cluster 0 – Have Class 1
     + Cluster 1 – Mix of Class 2 and 3 where Class 3 is dominating
     + Cluster 2 – Mix of Class 1 and 2 where Class 2 is dominating
     + Cluster 3 – Class 2

Hierarchal PCA – 3 Clusters

* + - Cluster 0 – Mix of Class 1 and 2 where Class 2 are more
    - Cluster 1 – Mix of Class 1 ,2 & 3
    - Cluster 2 – Mix of Class 1 ,2 & 3 and Class 3 is dominating Hierarchal Without PCA – 3 Clusters
    - Cluster 0 – All Class 1 few Class 2
    - Cluster 1 – All Class 2
    - Cluster 2 – Mix of Class 2 and 3

##### Clusters are formed on the basis of –

K-Mean PCA – 4 Clusters

* + - Cluster 0 – Members of this cluster Is dominated by OD280/OD315 of diluted wines, Ash, Hue, Flavanoids and Alcalinity of ash
    - Cluster 1 – Is formed on Proanthocyanins, Alcohol, Color Intensity
    - Cluster 2 – Nonflavanoid phenols, Malic acid, Total phenols
    - Cluster 3 – Alcalinity of ash, Proanthocyanins K-Mean Without PCA – 4 Clusters
    - Are formed on the basis of Proline, Magnesium and Color Intensity Hierarchal PCA – 3 Clusters
    - Cluster 0 – Are dominated by Proanthocyanins, Alcohol, Color Intensity
    - Cluster 1 – OD280/OD315 of diluted wines, Ash, Hue, Flavanoids, Alcalinity of ash, Proanthocyanins, Magnesium
    - Cluster 2 – Nonflavanoid phenols, Malic Acid, Total phenols, Magnesium is dominating

Hierarchal Without PCA – 3 Clusters

* + - Formation of clusters in this show difference amongst them on Proline, Magnesium and Flavanoids.

##### Number of Observations in Clusters formed -

K-Mean PCA – Almost equal in Cluster 0,1 and lesser observation in 2 and 3

K-Mean Without PCA – Almost equal in Cluster 0,1,2 and lesser observation in 3 Hierarchal PCA – Almost equal in 1 and 2 but less in 3

Hierarchal Without PCA – Almost equal in 3 Clusters

***\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\****

# Answer e.

Yes, it is possible to do the same study with the subset of the chemical composition dimension and we can further compare the analysis of the clusters which are created post that.

I have based my analysis around biplot which is formed after PCA

**We can form the subset on the basis of three major observations:**

1. Identify redundant variables:

As noticed in the biplot that two or more variables have similar directions and are very close to each other in the biplot even almost overlapping, it suggests that they capture similar

information and also have high correlation. As a result, we may consider dropping one of the redundant variables to reduce dimensionality without significant loss of information.

#### Assess variable importance:

The length of the arrows in the biplot represents the importance or contribution of dimension to the principal components.

Longer arrows represent stronger influence on the principal components and contribute more to the overall variation in the data.

Short arrows indicates that they have less impact on the principal components and may be candidates for dropping.

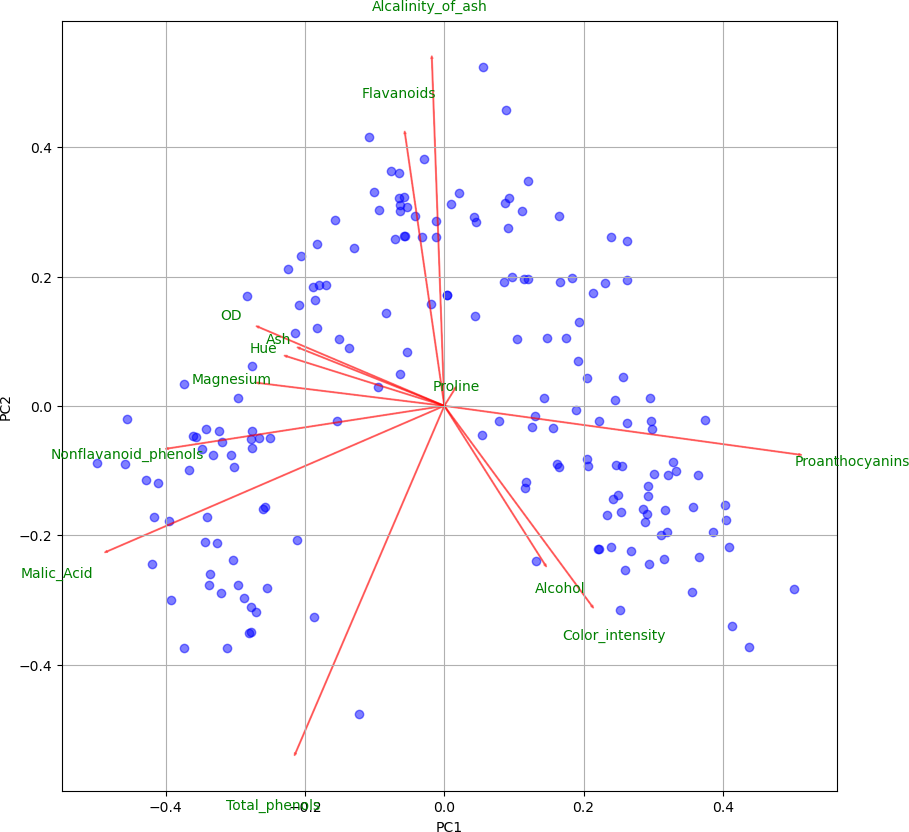
#### Focus on the most informative dimensions:

By examining the biplot or even contribution of dimension in the PCs, we can identify the principal components that capture the most significant variation in the data. If you observe that the variable in the PC have less impact on first PC component then we can drop.

Inference:

**By analysis small angle between dimension and shorter arrows of the dimension on the biplot I am removing the following dimensions – Alcohol, Ash, Hue, Flavanoid, Proline.**

Initial Biplot of PCA



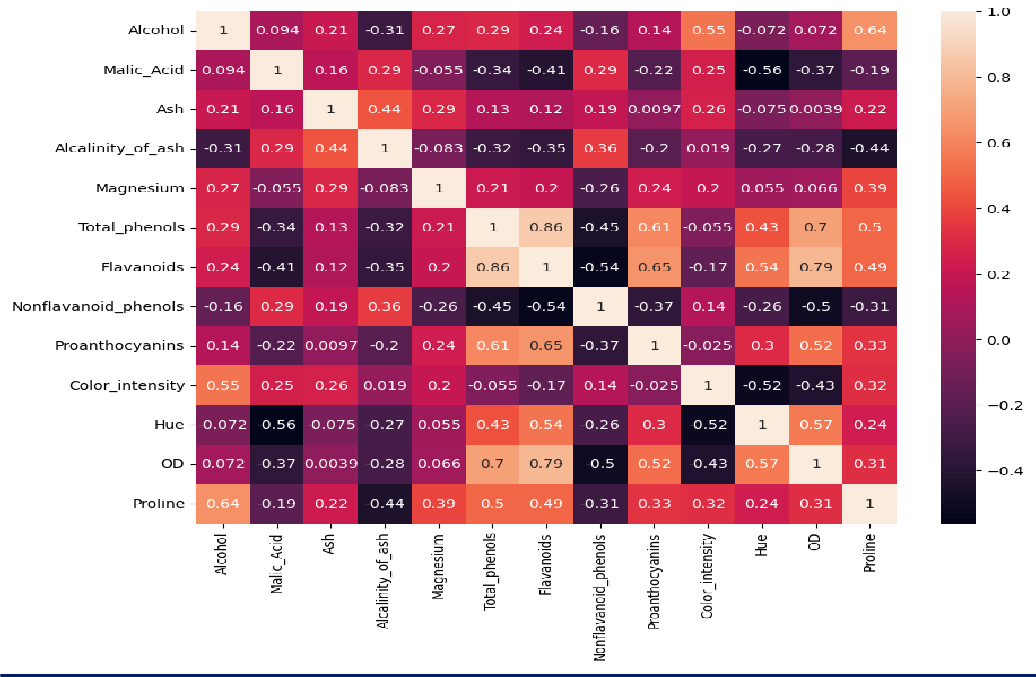
Correlation from the initial EDA is a statistical measure that describes the relationship or association between two dimensions. It measures the strength and direction of the linear relationship between

the variables. The correlation coefficient ranges from -1 to 1.

As per the Statistical Analysis it is stated that any coefficients having value above 0.6 is highly positively correlated whereas -0.6 is highly negatively correlated.

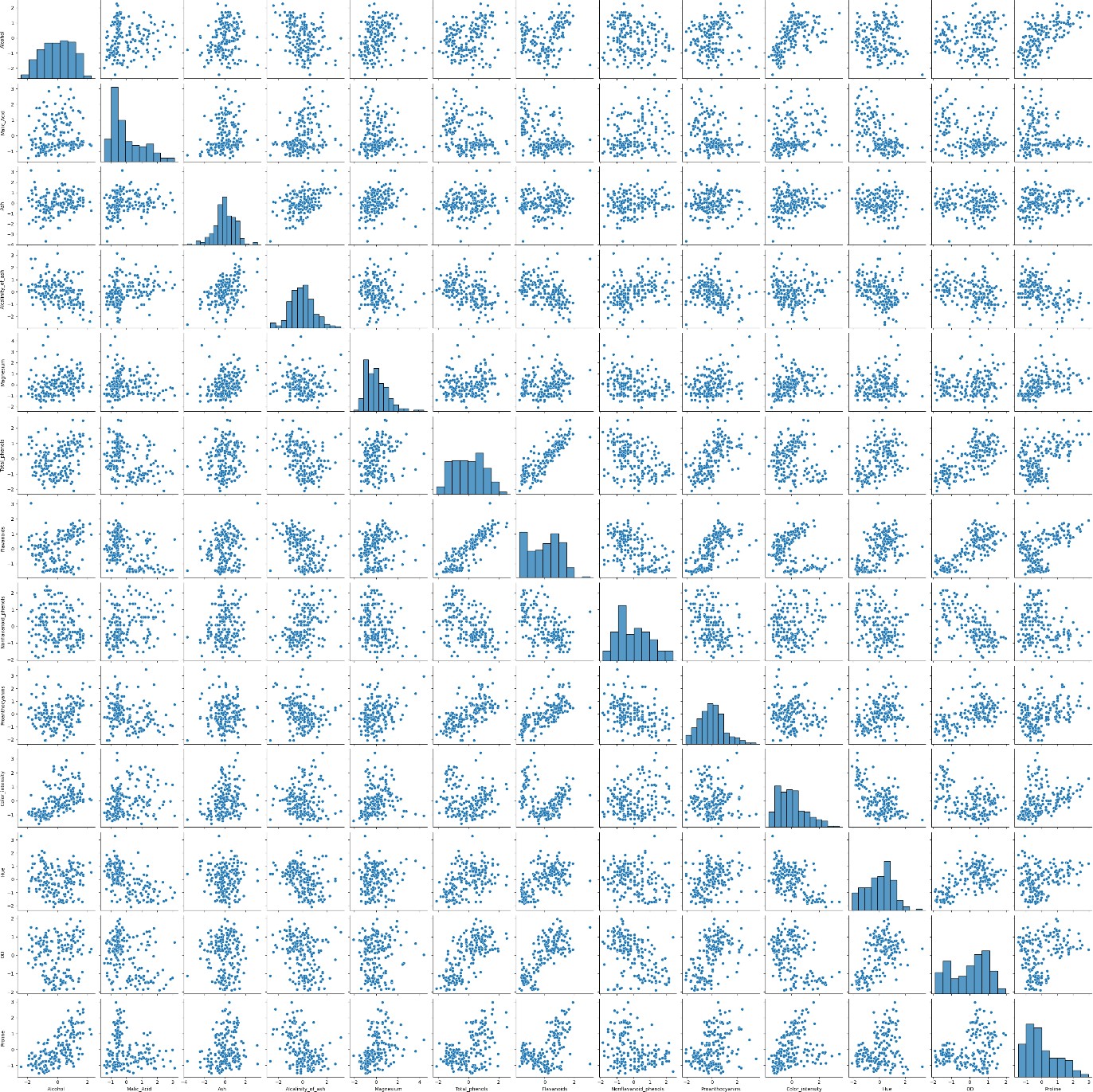
Hence we will be analysing the coefficients and basing our analysis on that.

### Correlation Matrix



Similar study can be sone using graphical method of pair plotting our dimensions to understand the relation amongst them. For supporting the analysis done on Correlation matrix.

Pair Plots Matrix



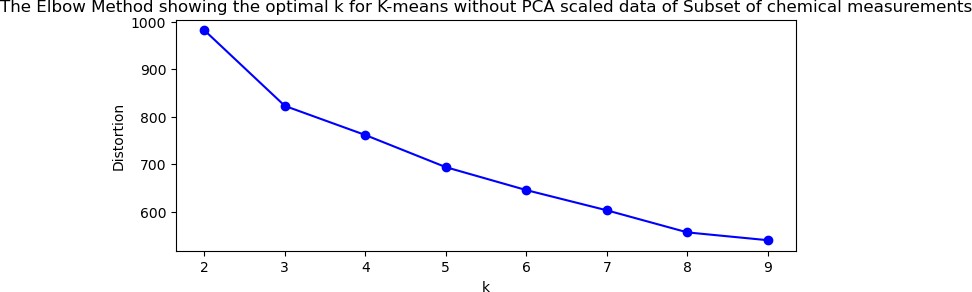
**Performing the Analysis on the Subset after removing the dimensions identified from the biplot**

##### Hierarchal Clustering on Subset Data

As observed on the above dendogram which is constricted using Ward method and Euclidean

distance it can be seen that **same number of clusters** are appearing which were appearing on the entire data.

##### K-Means Elbow Graph

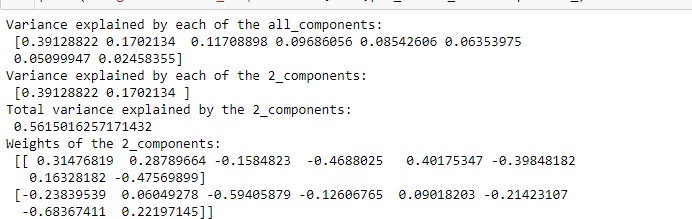


**K-Mean Inertia Reading - For PCA Data - 755.19**

As compared with other inertia values this is having lower value of inertia which indicates that the data points within each cluster are closer to their cluster centre, **suggesting tighter and more compact clusters**. A lower inertia value generally indicates better cluster quality.

##### PCA Biplot-





From the PCA it is showing a slightly higher Variance coverage from PC1 and PC2. Cluster formation is still looking the same.

##### Conclusion: We can go ahead by removing the variable which we found similar after studying initial Biplot graph. As by removing it our number of clusters and their formation is not getting impacted.

***\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\****