Machine Learning Unsupervised Learning 1

Assignment 2

Step 1:

Uploading Wine Dataset provided in the assignment folder in the python notebook. To do so imported few pandas and NumPy libraries. Column Names that are being used were provided in the Wine Names excel. After loading the data into a data frame got 178 rows and 14 columns wherein the first column used in WineType.

```
WineType  Alcohol  MalicAcid   Ash  Aclacinity of Ash  Magnesium  \
0            1    14.23       1.71  2.43               15.6        127
1            1    13.20       1.78  2.14               11.2        100
2            1    13.16       2.36  2.67               18.6        101
3            1    14.37       1.95  2.50               16.8        113
4            1    13.24       2.59  2.87               21.0        118
..         ...      ...        ...   ...                ...        ...
173          3    13.71       5.65  2.45               20.5         95
174          3    13.40       3.91  2.48               23.0        102
175          3    13.27       4.28  2.26               20.0        120
176          3    13.17       2.59  2.37               20.0        120
177          3    14.13       4.10  2.74               24.5         96

     Total Phenols  Flavanoids  NonFlavanoids phenols  Proanthocyanins
\
0              2.80        3.06                   0.28             2.29
1              2.65        2.76                   0.26             1.28
2              2.80        3.24                   0.30             2.81
3              3.85        3.49                   0.24             2.18
4              2.80        2.69                   0.39             1.82
..              ...         ...                    ...              ...
173            1.68        0.61                   0.52             1.06
174            1.80        0.75                   0.43             1.41
175            1.59        0.69                   0.43             1.35
176            1.65        0.68                   0.53             1.46
177            2.05        0.76                   0.56             1.35

     Color intensity   Hue  OD280/OD315 of diluted wines  Proline
0               5.64  1.04                          3.92     1065
1               4.38  1.05                          3.40     1050
2               5.68  1.03                          3.17     1185
3               7.80  0.86                          3.45     1480
4               4.32  1.04                          2.93      735
..               ...   ...                           ...      ...
173             7.70  0.64                          1.74      740
174             7.30  0.70                          1.56      750
175            10.20  0.59                          1.56      835
176             9.30  0.60                          1.62      840
177             9.20  0.61                          1.60      560

[178 rows x 14 columns]
```

Also performed analysis on the above dataset to understand initial clustering provided in the dataset as per WineType. Interpretation of each WineType is done by taking centroid of each WineType and dividing it into 3 clusters. Below is the analysis result:-

```
Cluster 0 :

 WineType                        1.000000
Alcohol                         13.744746
MalicAcid                        2.010678
Ash                              2.455593
Aclacinity of Ash               17.037288
Magnesium                      106.338983
Total Phenols                    2.840169
Flavanoids                       2.982373
NonFlavanoids phenols            0.290000
Proanthocyanins                  1.899322
Color intensity                  5.528305
Hue                              1.062034
OD280/OD315 of diluted wines     3.157797
Proline                       1115.711864
dtype: float64

 Cluster 1 :

 WineType                        2.000000
Alcohol                         12.278732
MalicAcid                        1.932676
Ash                              2.244789
Aclacinity of Ash               20.238028
Magnesium                       94.549296
Total Phenols                    2.258873
Flavanoids                       2.080845
NonFlavanoids phenols            0.363662
Proanthocyanins                  1.630282
Color intensity                  3.086620
Hue                              1.056282
OD280/OD315 of diluted wines     2.785352
Proline                        519.507042
dtype: float64

 Cluster 2 :

 WineType                        3.000000
Alcohol                         13.153750
MalicAcid                        3.333750
Ash                              2.437083
Aclacinity of Ash               21.416667
Magnesium                       99.312500
Total Phenols                    1.678750
Flavanoids                       0.781458
NonFlavanoids phenols            0.447500
Proanthocyanins                  1.153542
Color intensity                  7.396250
Hue                              0.682708
OD280/OD315 of diluted wines     1.683542
Proline                        629.895833
dtype: float64
```

Cluster 0: Rich and Flavorful:-

WineType: 1.0:
This cluster may represent wines with rich and intense flavors, characterized by higher alcohol conte nt, higher phenolic compounds (such as total phenols and flavonoids), and a higher concentration of proline. The higher OD280/OD315 ratio suggests a more intense and full-bodied flavor profile.

Cluster 1: Balanced and Moderate Flavor:-

WineType: 2.0
This cluster may represent wines with a balanced and moderate flavor profile. They have moderate l evels of key chemical components, indicating a well-rounded composition that strikes a balance bet ween different flavor elements.

Cluster 2: Bold and Robust Flavours:-

WineType: 3.0
This cluster may represent wines with bold and robust flavors. They have lower levels of key chemica l components compared to the other clusters, suggesting a more intense acidity (indicated by higher MalicAcid) and potentially deeper color intensity, which are often associated with robust and pronou nced flavors.

Step 2:

To do the PCA on the data have imported modules from sklearn and seaborn libraries. To perform PCA have excluded WineType column as it is a categorical data.

I have performed PCA for 2 components on normalized wine data , the output and its representation is presented below:

```
Principal Component 1: 0.9773
Principal Component 2: 0.0134
```

The output of pca_wine = PCA(n_components=2).fit_transform(winedata_scaled_df) is a transformed normalized wine dataset that has been reduced to two principal components using Principal Component Analysis (PCA). The resulting pca_wine variable represents the original dataset projected onto a new two-dimensional space, where each sample is represented by two principal component values.

A. Insights on PCA: -

PC1: PC1 explains most of the variance (97.73%) in the dataset. It represents the primary source of variation and can be considered as a measure of overall wine quality or characteristics. The weights (loadings) of the original features in PC1 indicate their contributions to this component. Features with higher absolute weights in PC1 have a stronger influence on differentiating wine samples along this component.

PC2: PC2 explains a relatively small amount of the variance (1.34%) in the dataset. It represents a secondary pattern or variation in the wine samples that is not captured by PC1. Like PC1, the weights of the original features in PC2 can provide insights into the specific characteristics or attributes that contribute to this component. Features with higher absolute weights in PC2 have a stronger influence on differentiating wine samples along this component.

To understand which original features have the most significant impact on PC1 and PC2, we need to analyse the weights of different variables in the dataset.

Significant Features Contributing to PC1 & PC2:

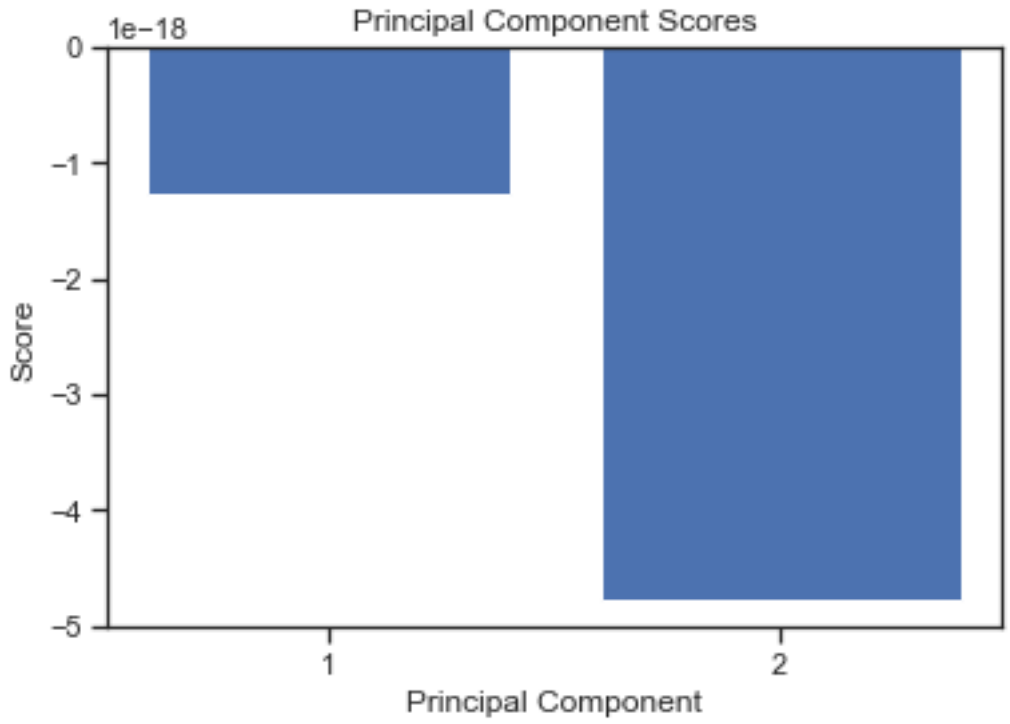
	PC1 Loadings	PC2 Loadings
Magnesium	0.947519	0.260495
Aclacinity of Ash	0.233458	0.912659
Proline	0.173131	0.037152
Alcohol	0.119674	0.258199
OD280/OD315 of diluted wines	0.025794	0.050489
MalicAcid	0.024934	0.112805
Color intensity	0.024025	0.098696
Ash	0.022872	0.064166
Total Phenols	0.020069	0.029461
Flavanoids	0.017301	0.015707
Proanthocyanins	0.014383	0.021773
Hue	0.009841	0.008528
NonFlavanoids phenols	0.003976	0.018834

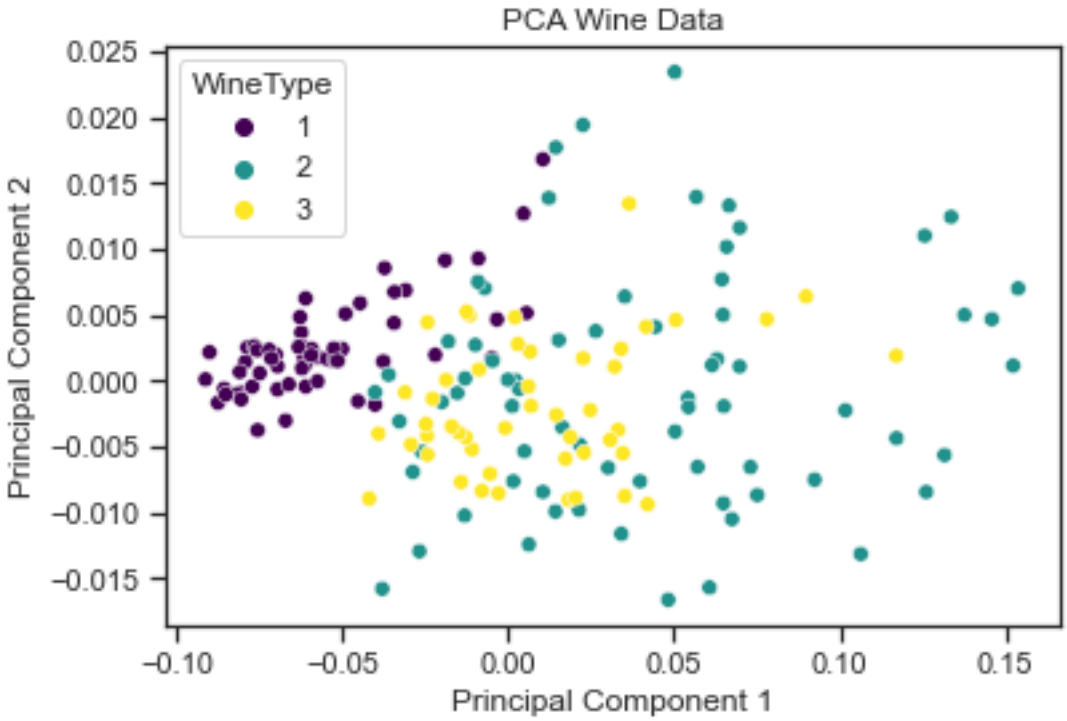
PC1 Significant Features (highest absolute loadings):
Magnesium (0.947519), Aclacinity of Ash (0.233458), Proline (0.173131), Alcohol (0.119674) and OD280/OD315 of diluted wines (0.025794)

PC2 Significant Features (highest absolute loadings):
Aclacinity of Ash (0.912659), Magnesium (0.260495),Alcohol (0.258199), MalicAcid (0.112805) and Color intensity (0.098696) .
These significant features indicate the variables that have the highest influence on each principal component.

In the case of PC1, features such as Magnesium, Aclacinity of Ash, Proline, Alcohol, and OD280/OD315 of diluted wines contribute the most to the variability captured by PC1. Similarly, PC2 is primarily influenced by features like Aclacinity of Ash, Magnesium, Alcohol, MalicAcid, and Color intensity.

Principal Component Scores:





Below DataSet is used to represent above graph. The data points in the graph represents their coordinates in the PCA component space. All the variables in the original dataset are reduced to PC1 & PC2 and then segregated WineType wise.

	PC1	PC2	WineType
0	-0.037425	0.008566	1
1	-0.061292	0.006263	1
2	-0.069781	-0.000652	1
3	-0.079461	0.001433	1
4	0.005421	0.005151	1
..	...	...	...
173	-0.024926	-0.003294	3
174	-0.017195	-0.003465	3
175	-0.011702	0.004963	3
176	-0.012636	0.005245	3
177	0.020230	-0.008875	3

B. Social and/or business values of those insights and actionable recommendations: -

These significant features indicate the variables that have the highest influence on each principal component. In the case of PC1, features such as Magnesium, Aclacinity of Ash, Proline, Alcohol, and OD280/OD315 of diluted wines contribute the most to the variability captured by PC1. Similarly, PC2 is primarily influenced by features like Aclacinity of Ash, Magnesium, Alcohol, MalicAcid, and Colour intensity.

These insights can help understand the underlying factors driving the patterns in the data and can be used to make informed decisions. Stakeholders can utilize this information for various purposes, such as:

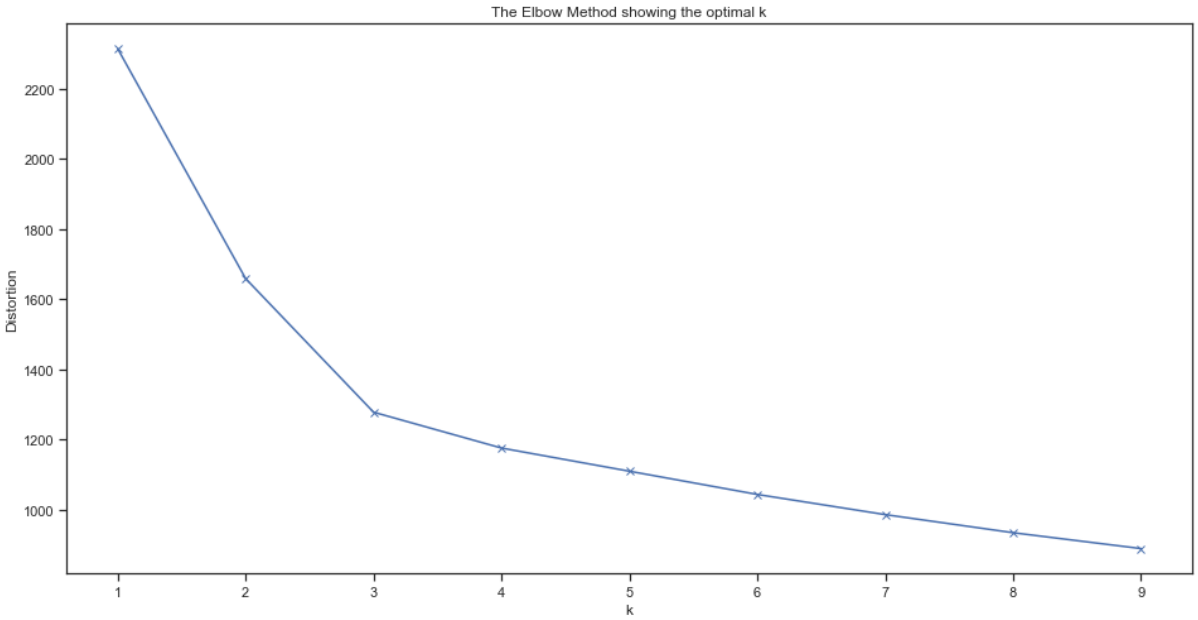
Quality control:
The alcohol content in wine is an important factor that contributes to its overall flavour, body, and balance. Higher alcohol levels can provide richness and warmth, but excessive alcohol can lead to a harsh or burning sensation. Phenolic compounds, such as flavonoids, contribute to the colour, flavour, and mouthfeel of the wine. Hue represents the colour shade of the wine, ranging from reddish to yellowish tones. It provides information about the wine's maturity and can be influenced by various pigments present in the wine. Understanding such significant features can provide insights into the key factors affecting wine quality. Stakeholders can focus on monitoring and controlling these features to maintain consistent quality standards.

Product development:
By identifying the influential features such as Alcohol, Magnesium, Proline, Total Phenols and other features, stakeholders can target specific attributes in the production process to develop wines with desired characteristics. They can focus on optimizing the levels of significant features to create wines with specific flavour profiles. This will also help them to segment their target market and tailor their marketing strategies accordingly.

Decision-making:
By considering the significant features, stakeholders can make data-driven decisions to optimize wine production such as fermentation techniques, grape selection, etc., and meet consumer preferences.

Step 3:

A. Cluster Analysis using KMeans on all chemical measurements:



As per the Elbow Curve analysis, KMean clustering on wine dataset is done on 3 clusters , even initially data was divided into 3 WineTypes.

Before Clustering, data count as per WineType was:

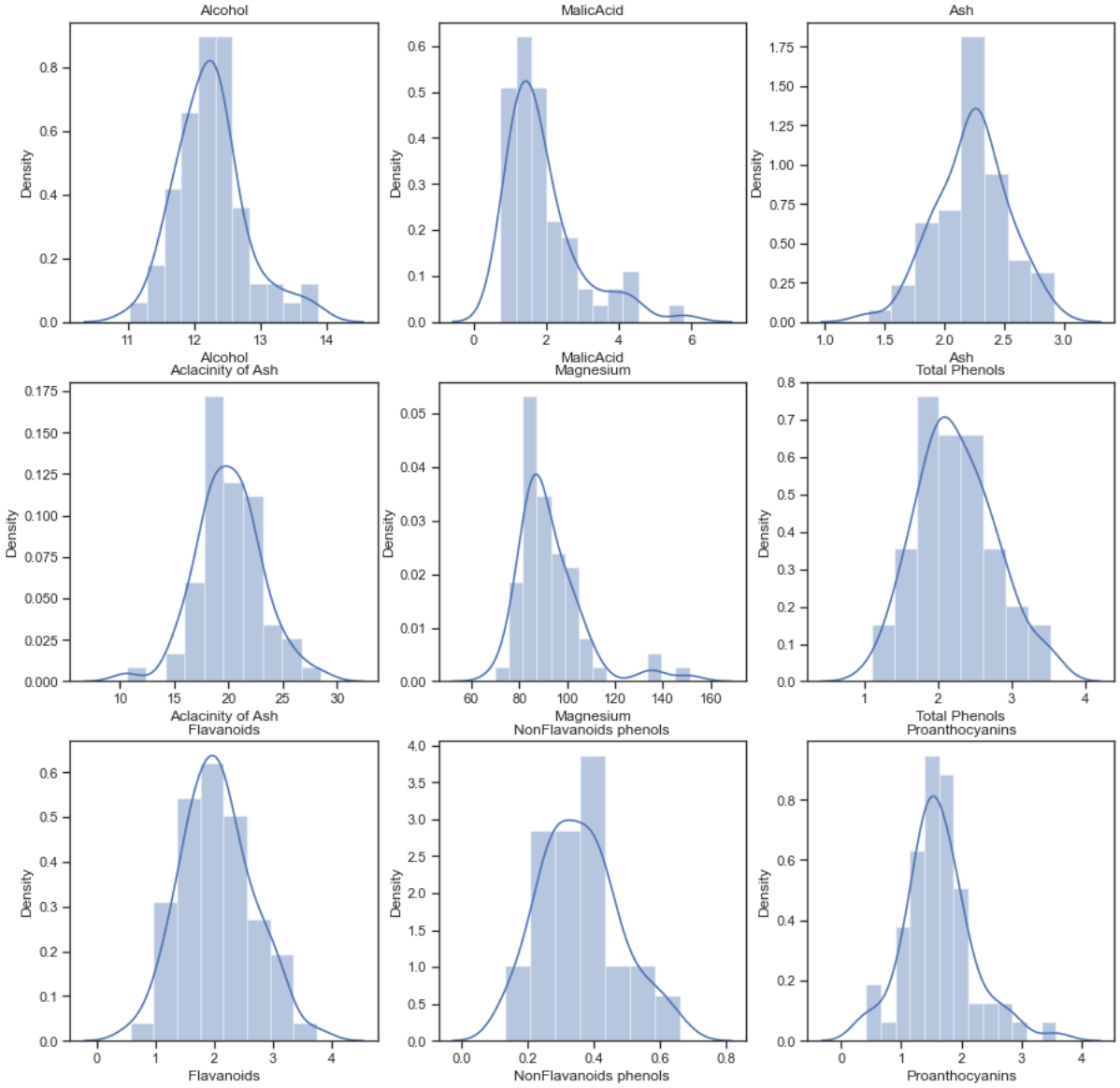
```
2    71
1    59
3    48
```

After Clustering, data count is changed into following:

```
0    65
2    62
1    51
Name: clust, dtype: int64
```

So, it seems clustering has some effect on the data. To analyse further data is plotted and centroid of each cluster is taken to under the data clustering using KMeans.

Cluster 0:-

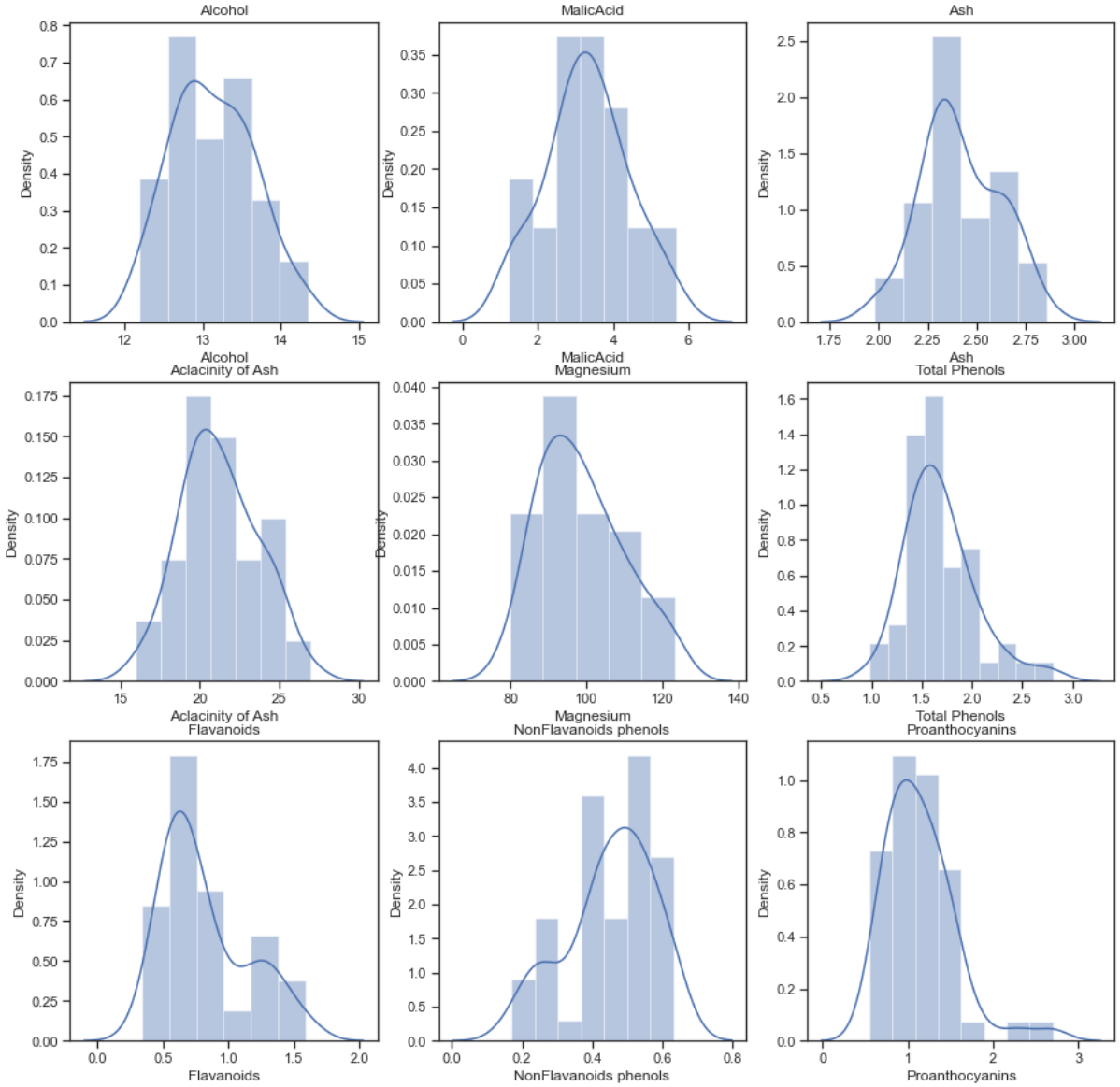


Alcohol	12.250923
MalicAcid	1.897385
Ash	2.231231
Aclacinity of Ash	20.063077
Magnesium	92.738462
Total Phenols	2.247692
Flavanoids	2.050000
NonFlavanoids phenols	0.357692
Proanthocyanins	1.624154
Color intensity	2.973077
Hue	1.062708
OD280/OD315 of diluted wines	2.803385
Proline	510.169231
clust	0.000000

dtype: float64

Cluster 0 (Balanced and Moderate): This cluster has lower values for Alcohol (12.25) compared to the other clusters, indicating wines with relatively lower alcohol content.Aclacinity of Ash is relatively high (20.06) in this cluster, suggesting wines with higher ash content. Magnesium is lower (92.74) compared to the other clusters.Total Phenols (2.25) and Flavanoids (2.05) have moderate values. Proline is also relatively low (510.17) in this cluster. Other features such as MalicAcid, Ash, NonFlavanoids phenols, Proanthocyanins, Color intensity, Hue, and OD280/OD315 of diluted wines have lower to moderate values,indicating a slightly milder and less intense flavor.The Aclacinity of Ash and Total Phenols values suggest a balanced acidity and a moderate.

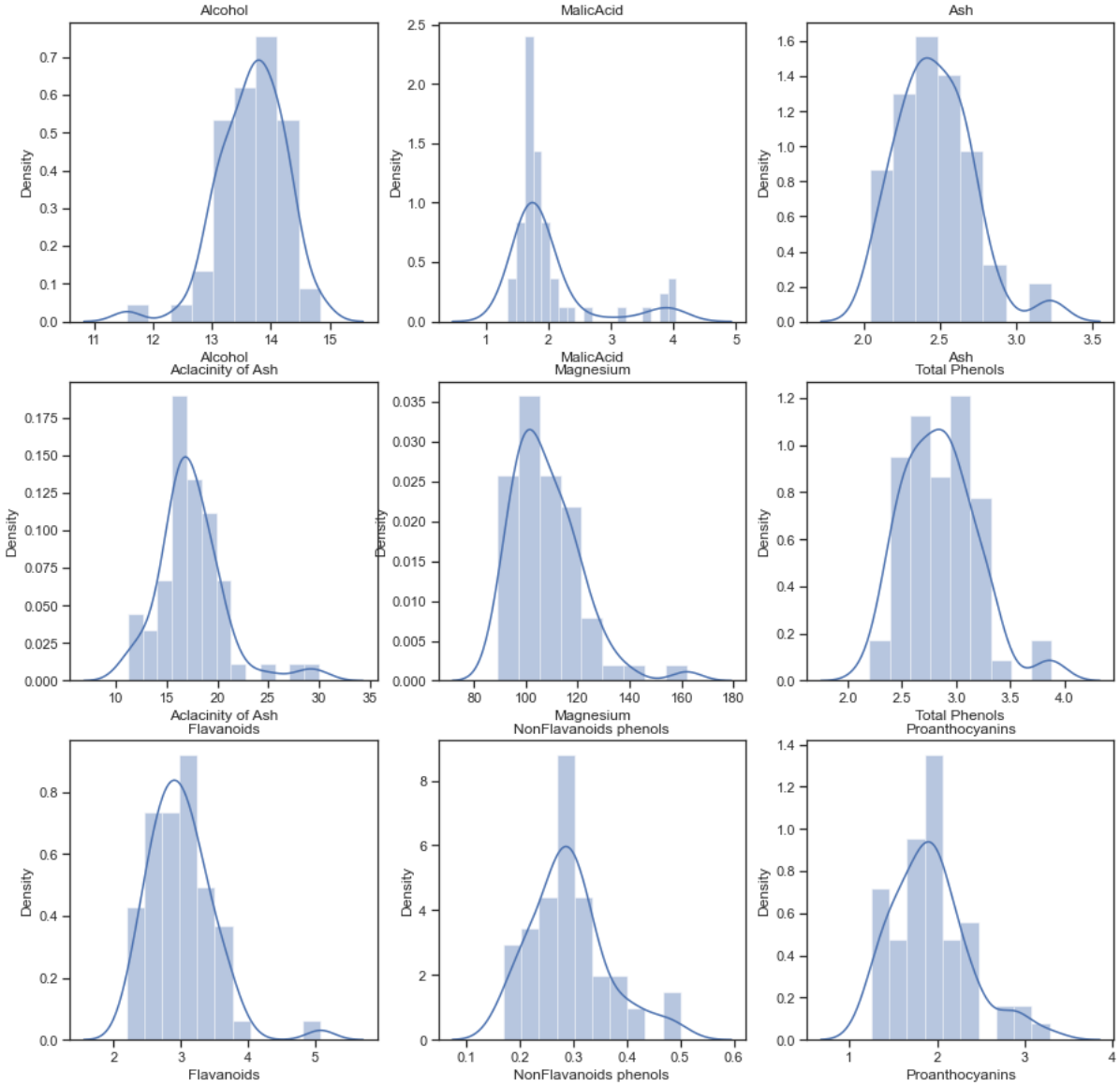
Cluster 1:



```
Alcohol                          13.134118
MalicAcid                         3.307255
Ash                               2.417647
Aclacinity of Ash                21.241176
Magnesium                        98.666667
Total Phenols                     1.683922
Flavanoids                        0.818824
NonFlavanoids phenols             0.451961
Proanthocyanins                   1.145882
Color intensity                   7.234706
Hue                               0.691961
OD280/OD315 of diluted wines      1.696667
Proline                         619.058824
clust                             1.000000
dtype: float64
```

Cluster 1 (Bold and Intense): This cluster has lower values for Alcohol (13.13) compared to the other clusters, indicating wines with lower alcohol content.It shows higher levels of Aclacinity of Ash (21.24), suggesting wines with higher ash content.Magnesium is relatively lower (98.67) in this cluster compared to others.Total Phenols (1.68) and Flavanoids (0.82) have lower values, suggesting a slightly less pronounced aromatic profile.Proline is also relatively low (619.06) in this cluster,indicating a moderate level of richness in the wine.Other features such as MalicAcid, Ash, NonFlavanoids phenols, Proanthocyanins, Color intensity, Hue, and OD280/OD315 of diluted wines have moderate values.

Cluster 2:



Alcohol	13.676774
MalicAcid	1.997903
Ash	2.466290
Aclacinity of Ash	17.462903
Magnesium	107.967742
Total Phenols	2.847581
Flavanoids	3.003226
NonFlavanoids phenols	0.292097
Proanthocyanins	1.922097
Color intensity	5.453548
Hue	1.065484
OD280/OD315 of diluted wines	3.163387
Proline	1100.225806
clust	2.000000

dtype: float64

Cluster 2 (Rich Flavors) : This cluster has a relatively higher value for Alcohol (13.68) compared to the other clusters, indicating wines with higher alcohol content.It also has a moderate level of Aclacinity of Ash (17.46) and Magnesium (107.97).The cluster shows higher values for Total Phenols (2.85) and Flavanoids (3.00), suggesting wines with higher phenolic compounds.Proline, a measure of wine quality and intensity, is relatively high in this cluster (1100.23).Other features such as MalicAcid, Ash, NonFlavanoids phenols, Color intensity, Hue, and OD280/OD315 of diluted wines have intermediate values.Overall,these wines exhibit rich flavors and intense aromas and represents a robust and full-bodied profile.

Even after KMean Clustering we have arrived at the same clusters, where data is divided into Rich and Falavourful wine , Bold and Robust Flavour, and Balance and Moderate Flavour.

The output of this clustering is saved in this file **Snigdha_Bhattacharjee_12220067__Kmean.csv**

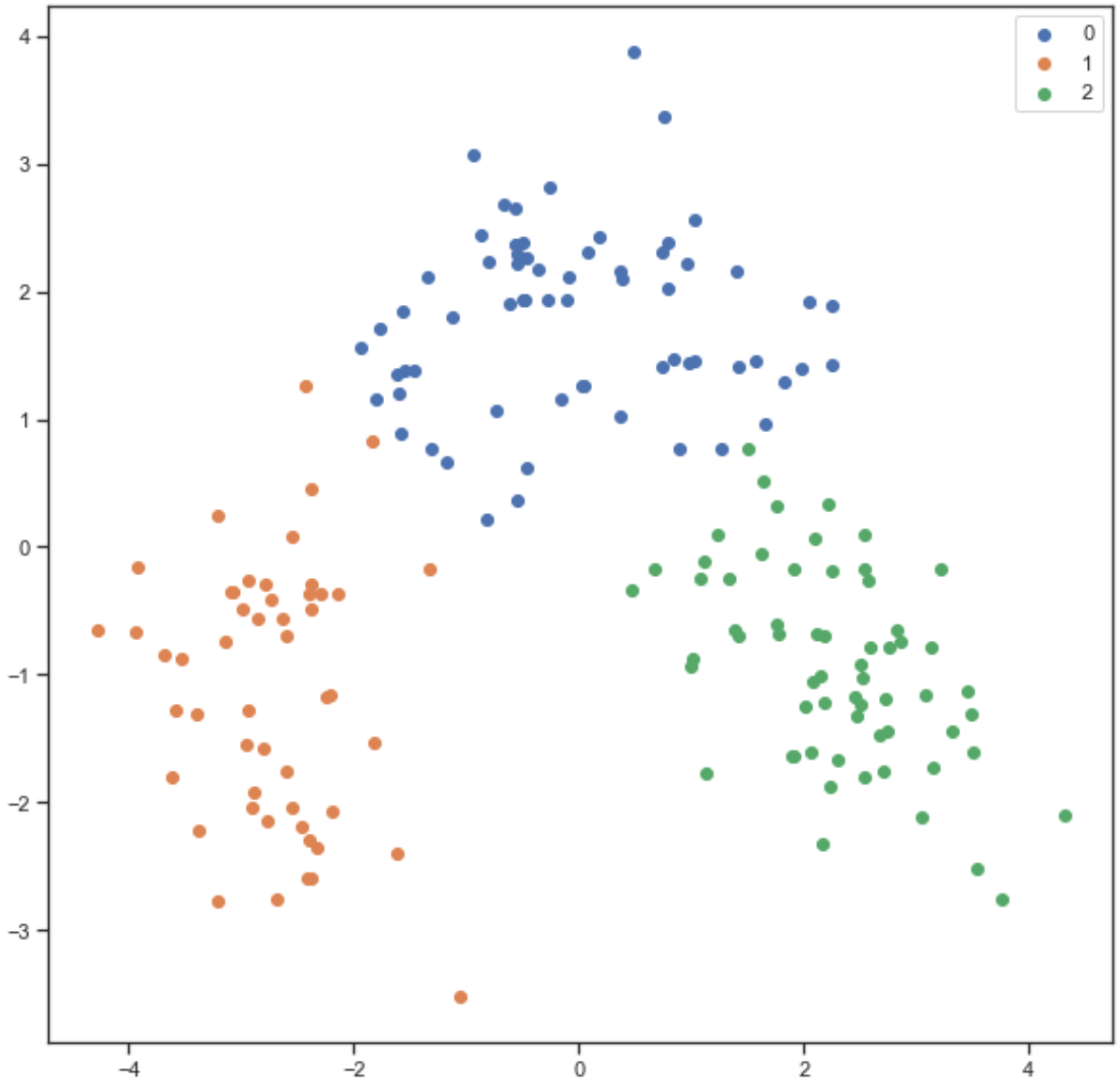
B. Cluster Analysis using KMeans on PCA.

KMean Clustering is performed on the above PCA components = 2 and again based on the previous elbow curve ,clustering component is 3.

PCA Cluster Dataset:

	PC1	PC2	clust
0	3.316751	-1.443463	1
1	2.209465	0.333393	1
2	2.516740	-1.031151	1
3	3.757066	-2.756372	1
4	1.008908	-0.869831	1
..	...	...	...
173	-3.370524	-2.216289	0
174	-2.601956	-1.757229	0
175	-2.677839	-2.760899	0
176	-2.387017	-2.297347	0
177	-3.208758	-2.768920	0

Scatter plot of both the PCA components as per newly formed clusters:



Interpretating the cluster formed using its centroid data:

```
Cluster 0 :

 Alcohol                          12.238308
MalicAcid                          1.931385
Ash                                2.219385
Aclacinity of Ash                 19.898462
Magnesium                         92.830769
Total Phenols                      2.204308
Flavanoids                         1.989231
NonFlavanoids phenols              0.365538
Proanthocyanins                    1.587692
Color intensity                    2.992615
Hue                                1.051631
OD280/OD315 of diluted wines       2.769231
Proline                          506.353846
PC1                               -0.162785
PC2                                1.767588
clust                              0.000000
dtype: float64

 Cluster 1 :

 Alcohol                          13.151633
MalicAcid                          3.344490
Ash                                2.434694
Aclacinity of Ash                 21.438776
Magnesium                         99.020408
Total Phenols                      1.678163
Flavanoids                         0.797959
NonFlavanoids phenols              0.450816
Proanthocyanins                    1.163061
Color intensity                    7.343265
Hue                                0.685918
OD280/OD315 of diluted wines       1.690204
Proline                          627.551020
PC1                               -2.743930
PC2                               -1.214191
clust                              1.000000
dtype: float64

 Cluster 2 :

 Alcohol                          13.659219
MalicAcid                          1.975781
Ash                                2.463750
Aclacinity of Ash                 17.596875
```

```
Magnesium                        107.312500
Total Phenols                      2.859688
Flavanoids                         3.012656
NonFlavanoids phenols              0.290000
Proanthocyanins                    1.921719
Color intensity                    5.406250
Hue                                1.069688
OD280/OD315 of diluted wines       3.157188
Proline                         1082.562500
PC1                                2.266150
PC2                               -0.865592
clust                              2.000000
dtype: float64
```

Cluster 0 (Balanced and Moderate): This cluster has lower values for Alcohol (12.24) compared to the other clusters, indicating wines with relatively lower alcohol content.Aclacinity of Ash is relatively high (19.9) in this cluster, suggesting wines with higher ash content. Magnesium is lower (92.83) compared to the other clusters.Total Phenols (2.20) and Flavanoids (2.0) have moderate values. Proline is also relatively low (506.35) in this cluster. Other features such as MalicAcid, Ash, NonFlavanoids phenols, Proanthocyanins, Color intensity, Hue, and OD280/OD315 of diluted wines have lower to moderate values,indicating a slightly milder and less intense flavor.The Aclacinity of Ash and Total Phenols values suggest a balanced acidity and a moderate.

Cluster 1 (Bold and Intense): This cluster has lower values for Alcohol (13.15) compared to the other clusters, indicating wines with lower alcohol content.It shows higher levels of Aclacinity of Ash (21.43), suggesting wines with higher ash content.Magnesium is relatively lower (99.02) in this cluster compared to others.Total Phenols (1.68) and Flavanoids (0.8) have lower values, suggesting a slightly less pronounced aromatic profile.Proline is also relatively low (627.6) in this cluster,indicating a moderate level of richness in the wine.Other features such as MalicAcid, Ash, NonFlavanoids phenols, Proanthocyanins, Color intensity, Hue, and OD280/OD315 of diluted wines have moderate values.

Cluster 2 (Rich Flavors) : This cluster has a relatively higher value for Alcohol (13.65) compared to the other clusters, indicating wines with higher alcohol content.It also has a moderate level of Aclacinity of Ash (17.6) and Magnesium (107.31).The cluster shows higher values for Total Phenols (2.85) and Flavanoids (3.01), suggesting wines with higher phenolic compounds.Proline, a measure of wine quality and intensity, is relatively high in this cluster (1082.23).Other features such as MalicAcid, Ash, NonFlavanoids phenols, Color intensity, Hue, and OD280/OD315 of diluted wines have intermediate values.Overall,these wines exhibit rich flavors and intense aromas and represents a robust and full-bodied profile.

Performing KMean Clustering on 2 PCA components has arrived us to the same clusters, where data is divided into Rich and Falavourful wine , Bold and Robust Flavour, and Balance and Moderate Flavour. Same cluster interpretation happened as the 2 PCA components used summarizes 99% data of the original dataset.

The output of the above clustering result is saved in **Snigdha_Bhattacharjee_12220067_PCA.csv**

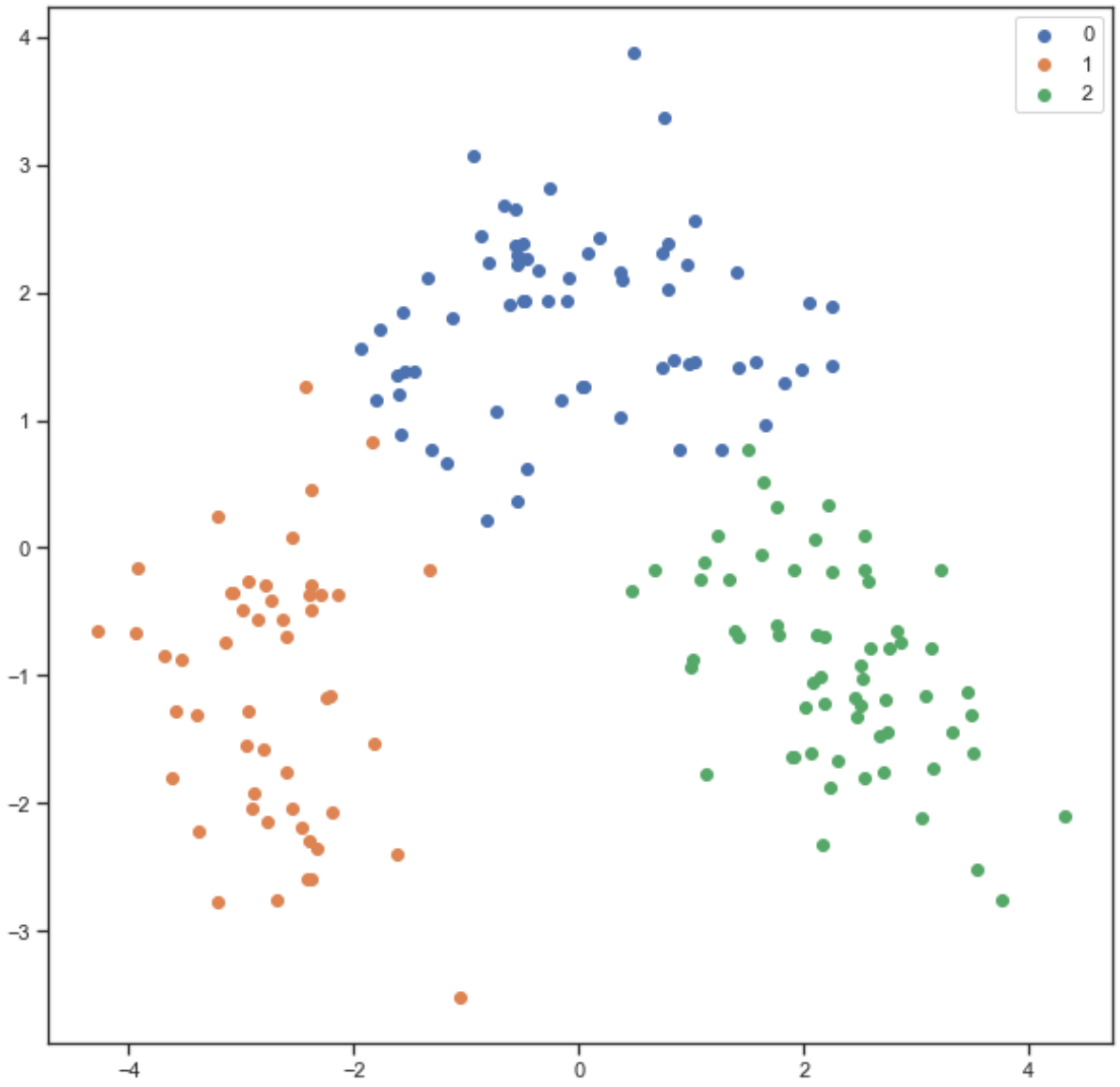
C. Clustering in the original data had below data count

2	71
1	59
3	48

After Clustering, data count is changed into following:

0	65
2	62
1	51

Even though the data count changed but there were no major changes in the centroid data and arriv ed to the same clusters. Upon analysing the scatterplot obtained during PCA clustering, it can been s een that some of the datapoints are in proximity with the data points of other clusters. May be these are the datapoints shifted from one cluster to another based on chemical composition. But this shift had no significance in centroid data as the distance is too small to have any effect.



D. As explained above there are no clearly separable clusters due to a few overlapping data points. Cluster chosen for this analysis is 3 obtained through Elbow curve. The elbow bent was at 3 and the distortions hereafter was smooth so went with 3 clusters.

The detailed analysis is provided in the above section while representing the data. But qualitatively t here is no major difference between the clusters obtained from chemical composition and from PCA scores. One of the reasons for this is the two PCA component selected for the clustering covers 99% of the features of original dataset. So, the clustering obtained through both means were same.3 clus ters that are formed denote the flavours of the three-wine type.

E. To calculate subset of the chemical measurements that can separate wines more distinctly I have used centroid data:

- Alcohol: The Alcohol content varies between Cluster 0 (13.744746) and Cluster 1 (12.278732), indicating a potential distinguishing factor.
- MalicAcid: Cluster 0 (2.010678) and Cluster 2 (3.333750) have distinct MalicAcid levels, sugg esting it as a potential discriminative feature.
- Ash: There is a difference in Ash content between Cluster 0 (2.455593) and Cluster 1 (2.2447 89), indicating its potential for separation.
- Aclacinity of Ash: Cluster 0 (17.037288) and Cluster 2 (21.416667) show significant variation i n Aclacinity of Ash, making it a potential distinguishing measurement.
- Magnesium: The Magnesium levels vary noticeably between Cluster 0 (106.338983) and Cluster 1 (94.549296), suggesting it as a differentiating factor.
- Total Phenols: Cluster 0 (2.840169) and Cluster 2 (1.678750) exhibit a substantial difference in Total Phenols, indicating its potential for separation.
- Flavanoids: There is a significant variation in Flavanoids content between Cluster 0 (2.98237 3) and Cluster 2 (0.781458), making it a potential discriminative feature.
- NonFlavanoids phenols: Cluster 0 (0.290000) and Cluster 2 (0.447500) show distinct NonFlav anoids phenols levels, suggesting its potential for separation.
- Proanthocyanins: The Proanthocyanins levels vary noticeably between Cluster 0 (1.899322) and Cluster 2 (1.153542), indicating its potential as a distinguishing measurement.
- Color intensity: Cluster 0 (5.528305) and Cluster 2 (7.396250) exhibit a significant difference i n Color intensity, making it a potential discriminative feature.
- Proline: Cluster 0 (1115.711864) has a significantly higher Proline value compared to Cluster 1 (519.507042) and Cluster 2 (629.895833).

Below measurements show relatively smaller variations across the clusters. Hence may not have Contribution in the cluster formation in terms of distance.

- Hue: The Hue measurement shows variation across the clusters, with Cluster 0 (1.062034) an d Cluster 1 (1.056282) having similar values.

- OD280/OD315 of diluted wines: Cluster 1 (2.785352) and Cluster 2 (1.683542), have very minimal distinguishing factor.