ASSIGNMENT SUBMISSION FORM

Course Name: Machine Learning (Unsupervised Learning 1)
Assignment Title: Individual Assignment 2
Submitted by:

Student Name	**PG ID**
Lakshmi Harshitha Yechuri	12110035

ISB Honour Code

- I will represent myself in a truthful manner.
- I will not fabricate or plagiarise any information with regard to the curriculum.
- I will not seek, receive or obtain an unfair advantage over other students.
- I will not be a party to any violation of the ISB Honour Code.
- I will personally uphold and abide, in theory and practice, the values, purpose and rules of the ISB Honour Code.
- I will report all violations of the ISB Honour Code by members of the ISB community.
- I will respect the rights and property of all in the ISB community.
- I will abide by all the rules and regulations that are prescribed by ISB.

Note: Lack of awareness of the ISB Honour Code is never an excuse for a violation. Please go through the Honour Code in the student handbook, understand it completely. Please also pay attention to the following points:

- Please do not share your assignment with your fellow students under any circumstances if the Honour Code scheme prohibits it. The HCC considers both parties to be guilty of an Honour Code violation in such circumstances.
- If the assignment allows you to refer to external sources, please make sure that you cite all your sources. Any material that is taken verbatim from an external source (website, news article etc.) must be in quotations. A much better practice is to paraphrase the source material (it still must be cited).

Step 1: Download the Wine data from the UCI machine learning repository (Wine dataset- UCI Repository)

Downloaded and saved as winedata.csv

Step 2: Do a Principal Components Analysis (PCA) on the data. Please include (copy-paste) the relevant software outputs in your submission while answering the following questions.

There are 14 variables in total in the dataset. Out of these, the first variable is categorical, the rest are numeric.

Six point summary

Below is the six point summary of all the variables in the dataset –

Type	Alcohol	Malic acid	Ash
1:59	Min. :11.03	Min. :0.740	Min. :1.360
2:71	1st Qu.:12.36	1st Qu.:1.603	1st Qu.:2.210
3:48	Median :13.05	Median :1.865	Median :2.360
	Mean :13.00	Mean :2.336	Mean :2.367
	3rd Qu.:13.68	3rd Qu.:3.083	3rd Qu.:2.558
	Max. :14.83	Max. :5.800	Max. :3.230

Alcalinity of ash	Magnesium	Total phenols	Flavanoids
Min. :10.60	Min. : 70.00	Min. :0.980	Min. :0.340
1st Qu.:17.20	1st Qu.: 88.00	1st Qu.:1.742	1st Qu.:1.205
Median :19.50	Median : 98.00	Median :2.355	Median :2.135
Mean :19.49	Mean : 99.74	Mean :2.295	Mean :2.029
3rd Qu.:21.50	3rd Qu.:107.00	3rd Qu.:2.800	3rd Qu.:2.875
Max. :30.00	Max. :162.00	Max. :3.880	Max. :5.080

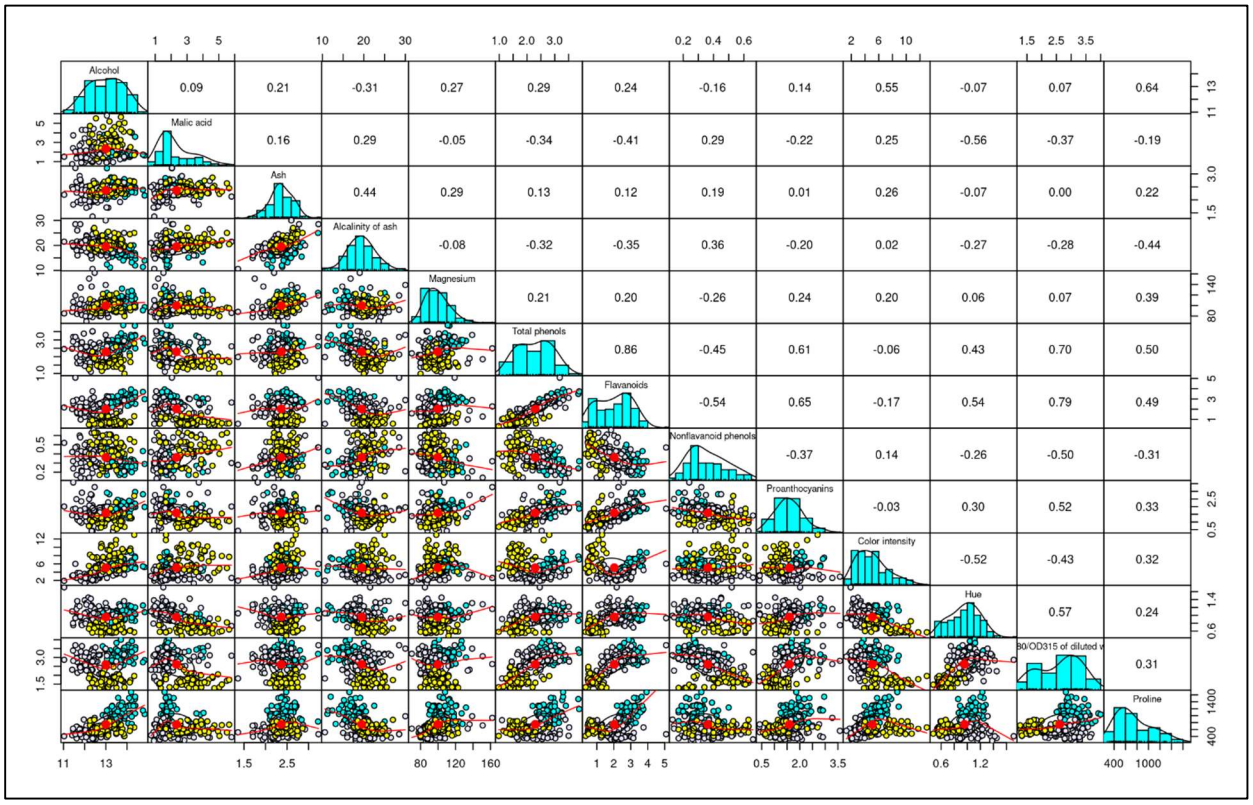
Nonflavanoid phenols	Proanthocyanins	Color intensity	Hue
Min. :0.1300	Min. :0.410	Min. : 1.280	Min. :0.4800
1st Qu.:0.2700	1st Qu.:1.250	1st Qu.: 3.220	1st Qu.:0.7825
Median :0.3400	Median :1.555	Median : 4.690	Median :0.9650
Mean :0.3619	Mean :1.591	Mean : 5.058	Mean :0.9574
3rd Qu.:0.4375	3rd Qu.:1.950	3rd Qu.: 6.200	3rd Qu.:1.1200
Max. :0.6600	Max. :3.580	Max. :13.000	Max. :1.7100

OD280/OD315 of diluted wines	Proline
Min. :1.270	Min. : 278.0
1st Qu.:1.938	1st Qu.: 500.5
Median :2.780	Median : 673.5
Mean :2.612	Mean : 746.9
3rd Qu.:3.170	3rd Qu.: 985.0
Max. :4.000	Max. :1680.0

We can see different scales for different variables. In order to avoid one variable overshadowing the other, it would be a good idea to scale the data.

Correlation

Using the pair panels function, we can create a scatterplot of all variables in the dataset. Below is the output –



From the above chart, we can see highest correlation between Flavonoids and OD280/OD315 (0.74), followed by Total Phenols and OD280/OD315 (0.70). We can see a fair number of variables having correlations, giving rise to multicollinearity problem. Hence, PCA can be used to address this problem.

Principal Component Analysis

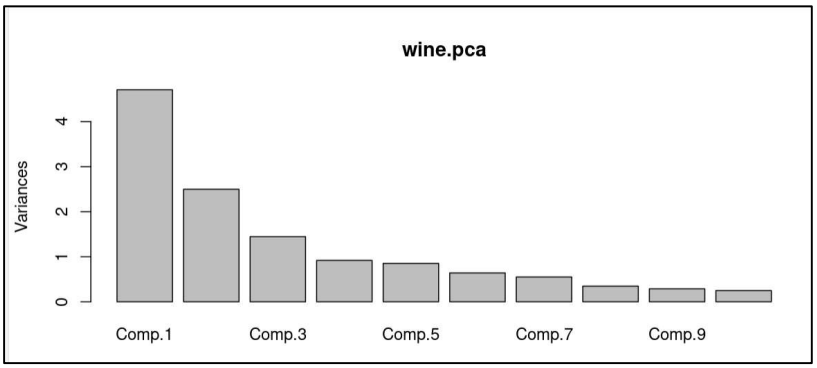
The prcomp function can be used to identify the principal components of the dataset. We have used the arguments center=TRUE and scale.=TRUE to ensure that the data is normalized before performing the PCA. Our PCA is based on the correlation matrix (scale. = TRUE) has been used for the biplot. As summary of prcomp shows us the importance of each PC. The table shows the first two Principal components can explain 55% of the variation in the data. And first 5 PCs explain 80%.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	2.169	1.5802	1.2025	0.95863	0.9237	0.80103	0.74231
Proportion of Variance	0.362	0.1921	0.1112	0.07069	0.06563	0.04936	0.04239
Cumulative Proportion	0.362	0.5541	0.6653	0.73599	0.80162	0.85098	0.89337

	PC8	PC9	PC10	PC11	PC12	PC13
Standard deviation	0.59034	0.53748	0.5009	0.47517	0.41082	0.32152
Proportion of Variance	0.02681	0.02222	0.0193	0.01737	0.01298	0.00795
Cumulative Proportion	0.92018	0.9424	0.9617	0.97907	0.99205	1

Scree plot

This shows the explained variance of each of the principal components and that illustrates the criteria used for selecting the number of principal components to be studied. The scree plot is as below –

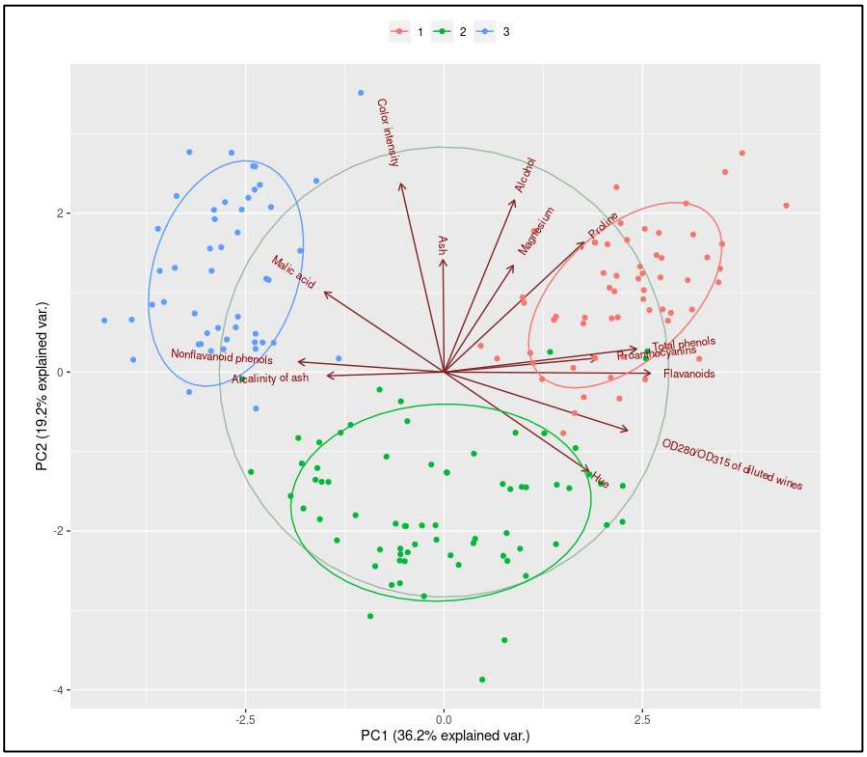


From the principal component 5, we can see that there is not much change in the explanation of the variability of the data. Hence, we can reduce the number of principal components to 5 PCs.

Hence, basis the cumulative proportion of variance and the scree plot, going ahead to interpret basis first five principal components.

Biplot of the data

Here, we have a plot of PC1, that explains 36% of the variability in the dataset, and PC2, which explains 19% of the variability in the dataset. We have the data color coded based on type of wine.

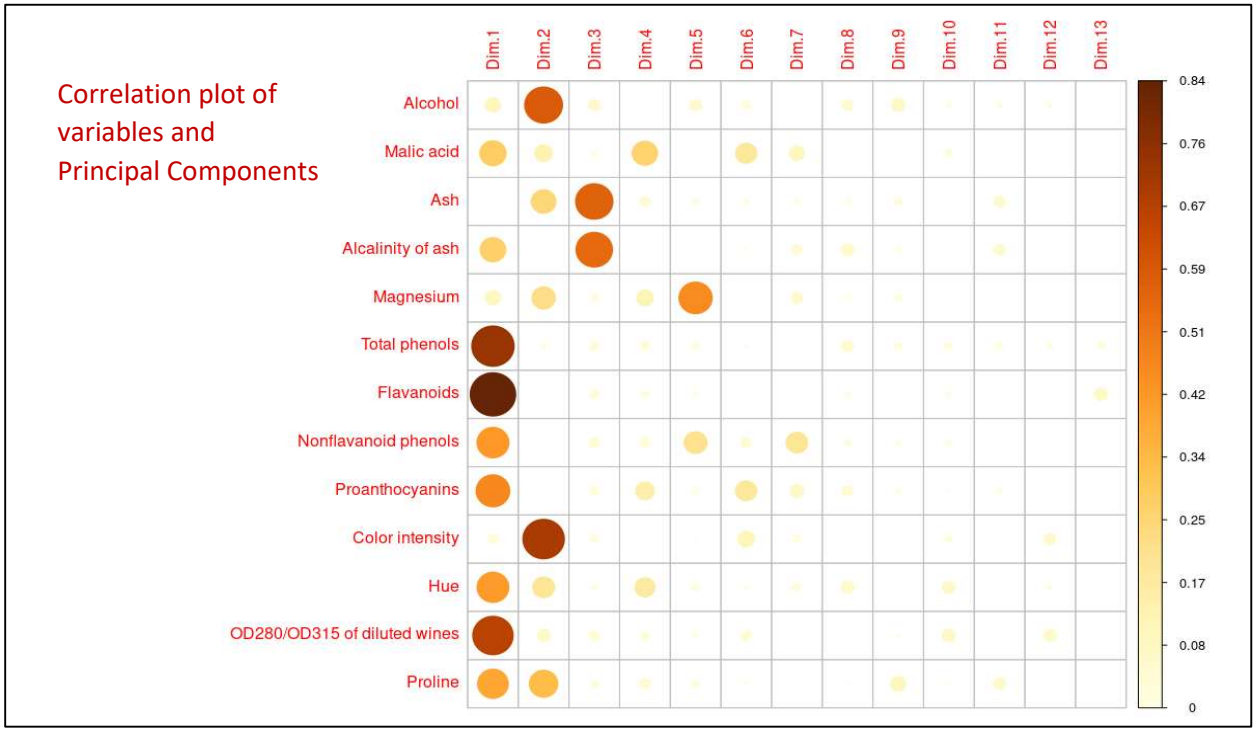


Loadings

Since we have 13 variables, we can see 13 PCs are generated. The rotations shows us the contribution of each variable into the PC. I have considered only the first 5 PCs as most of our variability is explained by that.

Chemical Component	PC1	PC2	PC3	PC4	PC5
Alcohol	-0.144	**0.484**	-0.207	0.018	-0.266
Malic acid	0.245	0.225	0.089	**-0.537**	0.035
Ash	0.002	**0.316**	**0.626**	0.214	-0.143
Alcalinity of ash	0.239	-0.011	**0.612**	-0.061	0.066
Magnesium	-0.142	**0.300**	0.131	**0.352**	**0.727**
Total phenols	**-0.395**	0.065	0.146	-0.198	-0.149
Flavanoids	**-0.423**	-0.003	0.151	-0.152	-0.109
Nonflavanoid phenols	0.299	0.029	0.17	0.203	**-0.501**
Proanthocyanins	**-0.313**	0.039	0.149	**-0.399**	0.137
Color intensity	0.089	**0.530**	-0.137	-0.066	-0.076
Hue	-0.297	-0.279	0.085	**0.428**	-0.174
OD280/OD315 of diluted wines	**-0.376**	-0.164	0.166	-0.184	-0.101
Proline	-0.287	**0.365**	-0.127	0.232	-0.158

A correlation plot will make it easier to understand the impact of each variable on its respective principal component.



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.

- A six-point summary shows us that variables are of different ranges, hence scaling of data is needed for further analysis.

- The correlation plot shows us that fair number of variables show correlation, leading to multicollinearity. Hence, using PCA and removing the multicollinearity will make our dataset more useful for analysis.

- The principal component summary shows us that 55% of the variability in the data can be explained by first two principal components and 80% of the variability can be explained using first five principal components.

- A biplot of the data explains the structure of the first two principal components. We can see the first PC has heavy load on total phenols, flavonoids, proanthocyanins and OD280/OD315 of diluted wines. The second PC has heavy load on alcohol, color intensity and proline. The positive/negative impact can be understood from the direction of the arrow on our plot.

- For the remaining three PCs, we can understand the load based on the loadings table and the correlation plot. The sign of the loading gives us the direction and the magnitude gives us the strength. On this basis, PC3 is loaded on Ash and Alkalinity of ash. PC4 is loaded on proanthocyanins, magnesium, malic acid and hue. PC5 is loaded on magnesium and non-flavonoid phenols.

b. What are the social and/or business values of those insights, and how the value of those insights can be harnessed—enumerate actionable recommendations for the identified stakeholder in this analysis?

By applying principal component analysis, we have arrived at 5 principal components that can be used as explanatory variables in any further analysis and clustering. Also, these can be used to describe the factors to explain the nature each type of wine.

First Principal Component – Health and Flavor Factor

This loads on the variables- phenols, flavonoids, proanthocyanins and OD280/OD315. The phenols, flavonoids, and proanthocyanins impact aroma, mouthfeel and bitterness. Also, these impact the nutrients in the wine and its antioxidant properties. They are good for the body. OD280/OD315 indicates high protein component.

Second Principal Component – Alcohol and color factor

This loads on the variables- alcohol, color intensity and proline. Proline impacts the nutrition and flavor of the wine. The color intensity refers darkness/lightness of the color. It reflects the nature of the grapes that make the wine. Therefore, the second common factor can be named as the visual evaluation factor of wine.

Third Principal Component – the Alkalinity factor

This is loaded on two variables- ash and alkalinity of ash. It is known that ash in wine is an effective substance for neutralizing acidity and is essentially an inorganic salt. Alkalinity of ash is a measure of weak alkalinity that is dissolved in the water.

Fourth principal Component – Sweetness and Hue factor

This has a large load on the four variables- of malic acid, hue, proanthocyanins and magnesium. Malic acid balances the sweetness of the wine. The hue refers to the vividness and warmth of the color of the wine.

Fifth principal Component – Mineral and Aroma factor

This is loaded on the variable- magnesium and non-flavonoid phenols. The magnesium can represent the mineral element and non-flavonoid phenols impact the aroma of the wine.

The score against each component with each type of wine can be summarized as below –

Type of Wine	Health & Flavour Factor	Alcohol & colour factor	Alkalinity factor	Sweetness & Hue factor	Mineral & Aroma factor
1	-2.2763	0.9652	-0.1591	0.1120	-0.2419
2	0.0389	-1.6389	0.2609	-0.0508	0.1217
3	2.7405	1.2378	-0.1903	-0.0625	0.1173

- We can see the first type of wine is high in low in health & flavor and mineral & aroma factor. It is high on sweetness & hue factor.
- The second type of wine is low in alcohol & color factor. It is high on Alkalinity and Mineral & aroma factors.
- The third type of wine is high on health & flavor factor, and Alcohol & color factor. It is low on the alkalinity and Sweetness & Hue factor.

Actionable recommendations for various stakeholders –

Vineyard owners - The vineyard owners or the wine sellers can promote their wines by customizing to customer's needs. The first type of wine can be promoted as a Sweet wine with medium alcohol levels. It can be a pairing wine. The second type of wine can be promoted as a low alcohol wine. The third type of wine can be promoted as a healthy and strongly flavored wine.

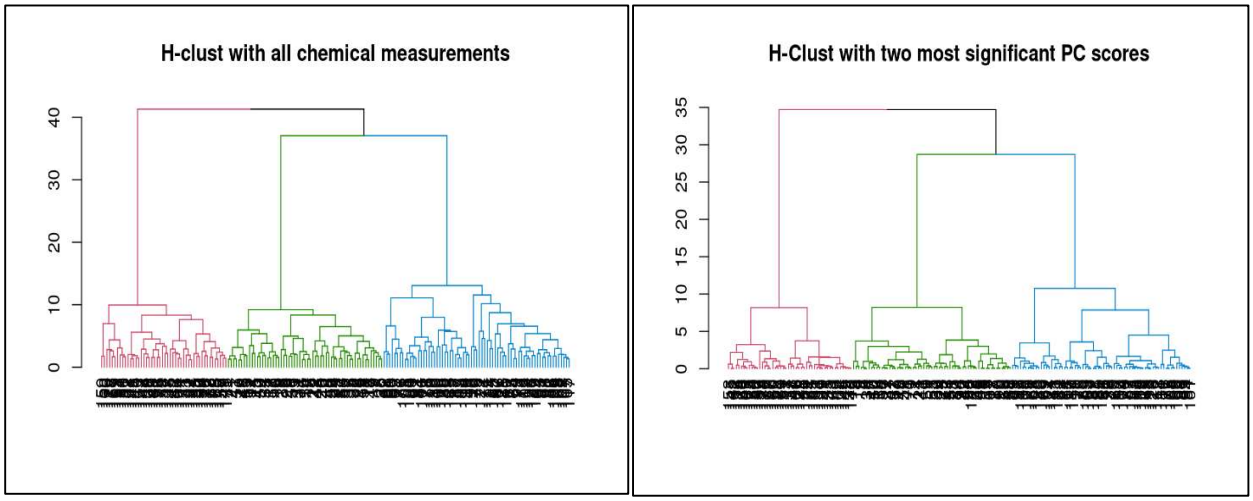
Customers – using the forementioned categories, customers can easily pick the wine of their preference based on their tastes and requirements.

Wine Connoisseurs – This type of factors will help wine connoisseurs to better understand the nature of the wine and helps them include all types of wines in their collections.

Step 3: Do a cluster analysis—you may try different algorithms or approaches and go with the one that you find most appropriate— using (i) all chemical measurements (ii) using two most significant PC scores.

Hierarchical Clustering

Hierarchical clusters have been created using both the approaches. The dendrograms are as below –



The above two plots show us that the clusters formed with both approaches indeed seem quite similar, we cannot find a strong distinction on using the principal components vs the actual data. This shows us that the first two principal components seem to be sufficient for clustering the data. We can indeed go ahead with the principal components for analysis, and we do not need to consider all the variables.

Hierarchical Cluster aggregates with chemical measurements

Cluster	1	2	3
Freq	59	71	48
Type1	1	0	0
Type2	0	1	0
Type3	0	0	1
Alcohol	13.745	12.279	13.154
Malic.acid	2.011	1.933	3.334
Ash	2.456	2.245	2.437
Alcalinity.of.ash	17.037	20.238	21.417
Magnesium	106.339	94.549	99.313
Total.phenols	2.840	2.259	1.679
Flavanoids	2.982	2.081	0.781
Nonflavanoid.phenols	0.290	0.364	0.448
Proanthocyanins	1.899	1.630	1.154
Color.intensity	5.528	3.087	7.396
Hue	1.062	1.056	0.683
OD280.OD315.of.diluted.wines	3.158	2.785	1.684
Proline	1115.712	519.507	629.896

Cluster 1 is clearly type 1 wine, cluster 2 is type 2 wine and cluster 3 is type 3 wine. The components that are high in each type are highlighted with a darker shade of red, and components that are low in each type of wine are left white.

Hierarchical Cluster Aggregates with Principal components –

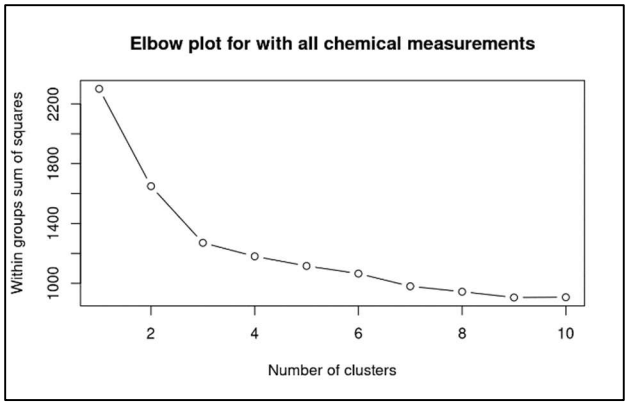
Cluster	**1**	**2**	**3**
Freq	**61**	**69**	**48**
PC1 - Health and Flavour Factor	-2.276	0.038	2.740
PC2 - Alcohol & colour factor	0.965	-1.638	1.237

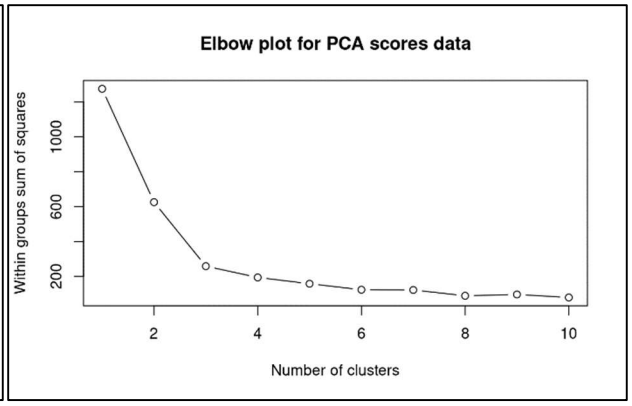
Findings –

- Type 1 wines (cluster 1) are low in health & flavor factor; slightly high in alcohol & color factor.
- Type 2 wines (cluster 2) are moderate in health & flavor factor; low in alcohol & color factor.
- Type 3 wines (cluster 3) are high in health & flavor factor; high in alcohol & color factor.

K means clustering

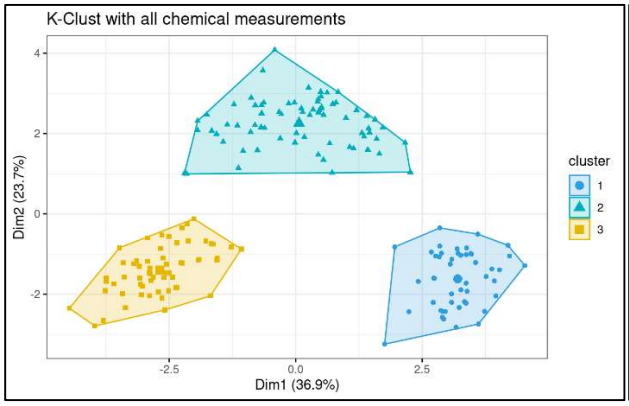
First, we can create elbow plots to clearly define the number of clusters to be formed.

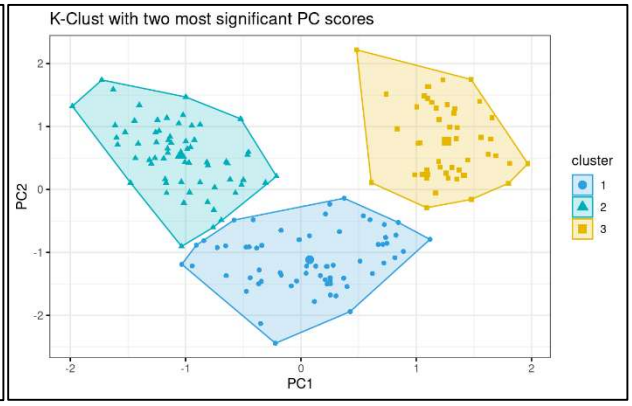




For both approaches, we can see that three clusters would be sufficient to form clearly defined clusters.

Going ahead to chart with three clusters using both the approaches (all chemical measurements approach and PCA approach.





We can see clear cluster formation even with PCA being applied. The clusters are clearly defined and hence, we should be fine with using the principal component analysis instead of using all chemical measurements.

K means Cluster Aggregates with all chemical measurements

Cluster	1	2	3
Freq	**59**	**71**	**48**
Type1	0	0	1
Type2	0	1	0
Type3	1	0	0
Alcohol	13.15375	12.27873	13.74475
Malic.acid	3.33375	1.932676	2.010678
Ash	2.437083	2.244789	2.455593
Alcalinity.of.ash	21.41667	20.23803	17.03729
Magnesium	99.3125	94.5493	106.339
Total.phenols	1.67875	2.258873	2.840169
Flavanoids	0.781458	2.080845	2.982373
Nonflavanoid.phenols	0.4475	0.363662	0.29
Proanthocyanins	1.153542	1.630282	1.899322
Color.intensity	7.39625	3.08662	5.528305
Hue	0.682708	1.056282	1.062034
OD280.OD315.of.diluted.wines	1.683542	2.785352	3.157797
Proline	629.8958	519.507	1,115.71

K means Cluster Aggregates with top two principal components

Cluster	1	2	3
Freq	64	49	65
PC1 - Health and Flavour Factor	-2.259	2.736	0.162
PC2 - Alcohol & colour factor	0.863	1.210	-1.763

Findings –

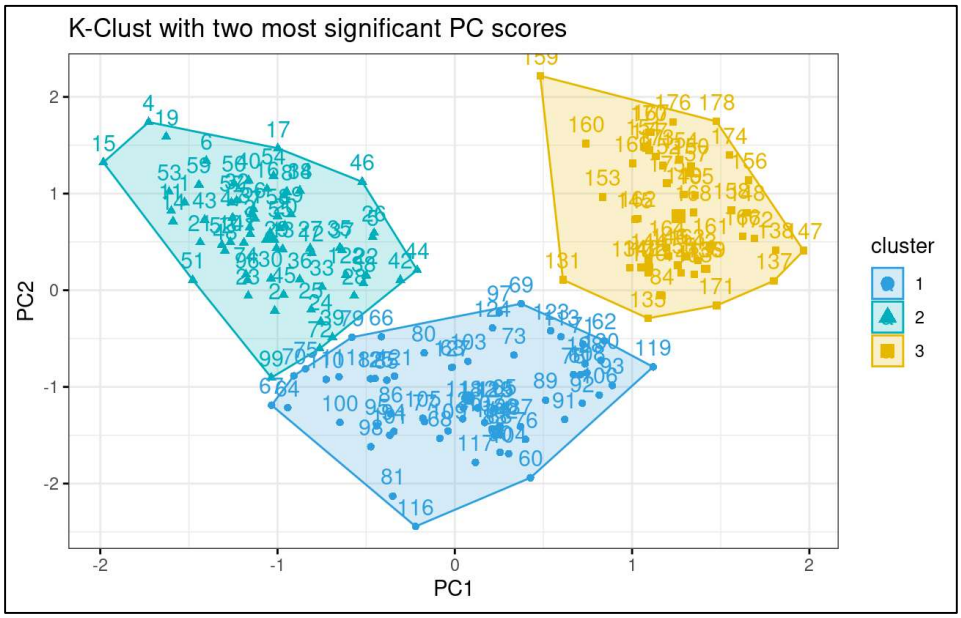
- Type 1 wines (cluster 1) are low in health & flavor factor; slightly high in alcohol & color factor.
- Type 2 wines (cluster 3) are moderate in health & flavor factor; low in alcohol & color factor.
- Type 3 wines (cluster 2) are high in health & flavor factor; high in alcohol & color factor.

This is in line with our findings from all chemicals data; as well as our findings from PCA.

c. Any more insights you come across during the clustering exercise?

- The clusters are very similar for the entire chemical data and for the top two principal components. Hence, we can conclude minimal impact on clustering exercise due to PCA.

- Each cluster in the clustering analysis seems to correspond to each type of wine. We can confirm this by labelling the data. We are aware that the data is sorted by type of wine and the below chart with labels confirms our observation that each cluster represents each type of wine.



- As the clusters are very clearly defined, the findings from hierarchical clustering are same as the findings of K means clustering when all the variables are used.

- The clustering of top two PCA scores has shown a slight variation from Hierarchical to K means clustering. This is understandable because we are explaining only about 55% of the variability of the data using PCA scores.

d. Are there clearly separable clusters of wines? How many clusters did you go with? How the clusters obtained in part (i) are different from or similar to clusters obtained in part (ii), qualitatively?

There are clear clusters being formed with both approaches. The elbow plot suggested that we can go with 3 clusters. Very similar clusters are being formed even if we go with all chemical measurements; or with two most significant PC scores. Hence, we can qualitatively conclude that we can go ahead with clustering using PC scores without much of information loss and provide insights of same quality with much lesser variables.

e. Could you suggest a subset of the chemical measurements that can separate wines more distinctly? How did you go about choosing that subset? How do the rest of the measurements that were not included while clustering, vary across those clusters?

By identifying the chemicals that have the maximum influence on Principal component 1 and principal component 2; we can find the chemical measurements that separate the wines distinctly. In this way, the chemicals identified are –

- Flavonoids
- Total phenols
- OD280/OD315 of diluted wines
- Proanthocyanins
- Color intensity
- Proline
- Alcohol

Below are the ranges for all variables We can see there are clear ranges for the variables that have been listed above. Proline, total phenols and color intensity are good examples.

However, for variables like Ash, Malic acid, Magnesium, etc. which are not included in the above list, we do have a lot of overlap in the ranges. This means there is a lot of variation across clusters for values of these chemical components.

Sl no	Chemical Component	Range for Type 1	Range for Type 2	Range for Type 3
1	**Flavanoids**	(2.19 , 3.93)	(0.57 , 5.08)	(0.34 , 1.57)
2	**Total.phenols**	(2.2 , 3.88)	(1.1 , 3.52)	(0.98 , 2.8)
3	**OD280.OD315.of.diluted.wines**	(2.51 , 4)	(1.59 , 3.69)	(1.27 , 2.47)
4	**Proanthocyanins**	(1.25 , 2.96)	(0.41 , 3.58)	(0.55 , 2.7)
5	**Color.intensity**	(3.52 , 8.9)	(1.28 , 6)	(3.85 , 13)
6	**Proline**	(680 , 1680)	(278 , 985)	(415 , 880)
7	**Alcohol**	(12.85 , 14.83)	(11.03 , 13.86)	(12.2 , 14.34)
8	**Malic.acid**	(1.35 , 4.04)	(0.74 , 5.8)	(1.24 , 5.65)
9	**Ash**	(2.04 , 3.22)	(1.36 , 3.23)	(2.1 , 2.86)
10	**Alcalinity.of.ash**	(11.2 , 25)	(10.6 , 30)	(17.5 , 27)
11	**Magnesium**	(89 , 132)	(70 , 162)	(80 , 123)
12	**Nonflavanoid.phenols**	(0.17 , 0.5)	(0.13 , 0.66)	(0.17 , 0.63)
13	**Hue**	(0.82 , 1.28)	(0.69 , 1.71)	(0.48 , 0.96)