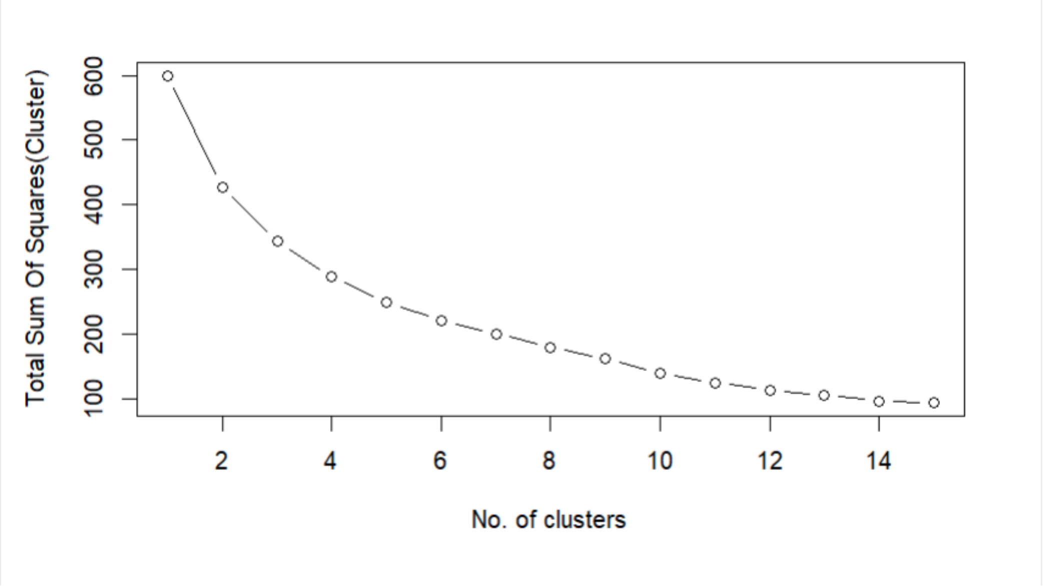
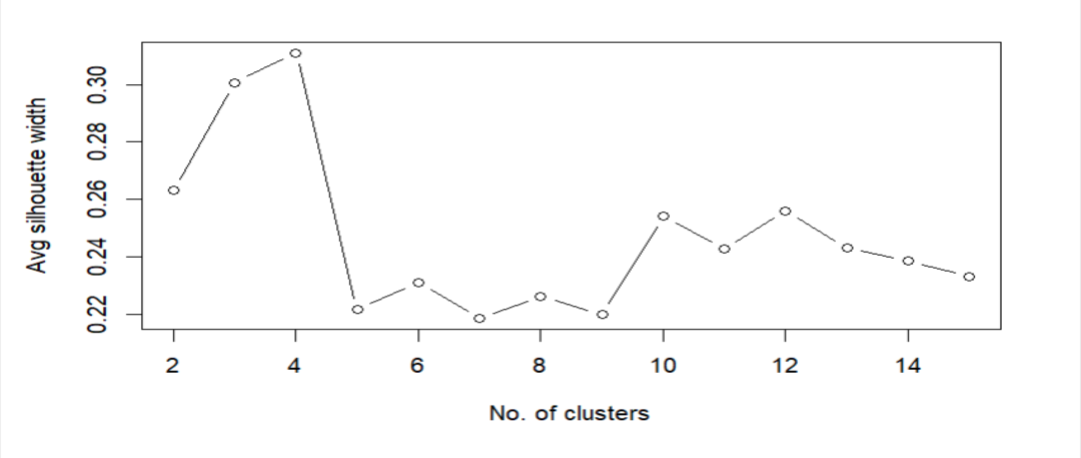
Homework-3

1) **The "chorSub" data from the "cluster" package contains measurements of 10 chemicals in 61 geological samples from the Kola Peninsula. Cluster the data using k-means and hierarchical clustering. What is a good choice of "k" for each of these methods? Justify your selection**

For K-means clustering k can be choosen by analysing the data using elbow method, silhouette method and gap statistic method. here we will use Elbow method and Silhouette analysis.

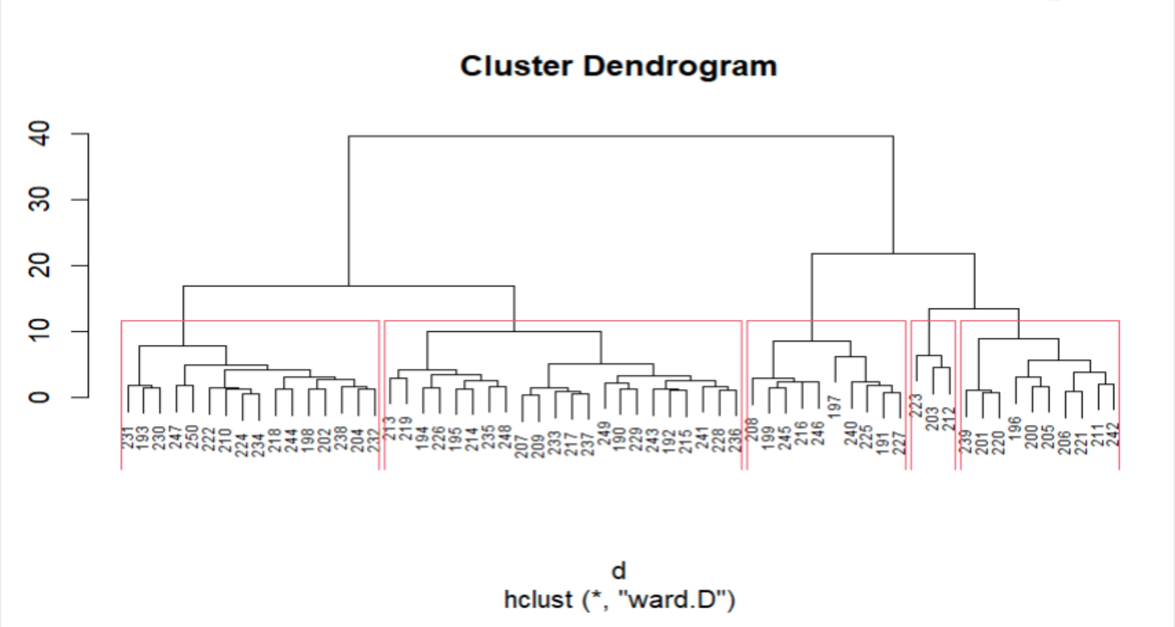


This graph gradually smooths out as the number of clusters increases, but there's a slight bend at k=4. This suggests that adding more clusters after four will provide only marginal improvements to the model's fit.

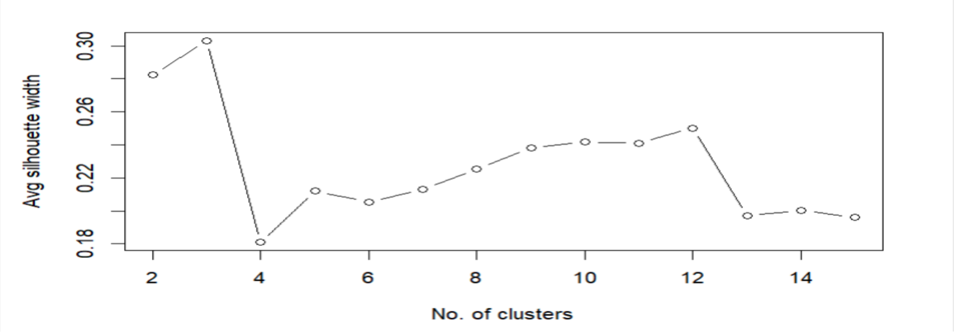


In Silhoutte analysis, the first peak is at k=4, indicating a strong level with two clusters that well-separate the data.

The elbow approach and the silhouette analysis both recommend that k=4 be used for k-means clustering. This choice is accurate. A massive peak can be seen in the silhouette plot at k=4, suggesting a robust structure with distinct clusters that are apart from one another. This is supported by the elbow plot, which exhibits a small bend at the same location and suggests that adding clusters beyond 4 will not significantly reduce variation. This methodological consistency confirms the correctness for k=4, which means a balanced and understandable clustering solution for the chorSub dataset that is both statistically and practically significant.



I searched for notable difference in the linkage distance, which frequently specify a natural cluster separation, when analyzing the dendrogram above. I analysed the dissimilarity between the data points that are clustered together based on the height of the merges. There are multiple levels at which clusters merge in this dendrogram, but when I pay close attention to the areas where there is a significant rise in merge height because these may be possible cuts for significant clusters.Five is the potential K value.



This plot presents a peak at k=3, which is another viable option but not as distinct as k=5

Although the silhouette analysis for hierarchical clustering indicates that k = 3 would be a good option, the dendrogram offers an alternative viewpoint, suggesting that k = 5 might offer more precise clustering. Due to significant jumps in the connection distance at this level, which indicate a more meaningful division of the data into natural groups , we selected k = 5 after analyzing the dendrogram.

2) Consider the "diamonds" data from ggplot2. Use principal components on the variables{caret, x, y, z, depth, table}, and answer the following quesHons.

a) How much of the total variance does the first principal component account for? How

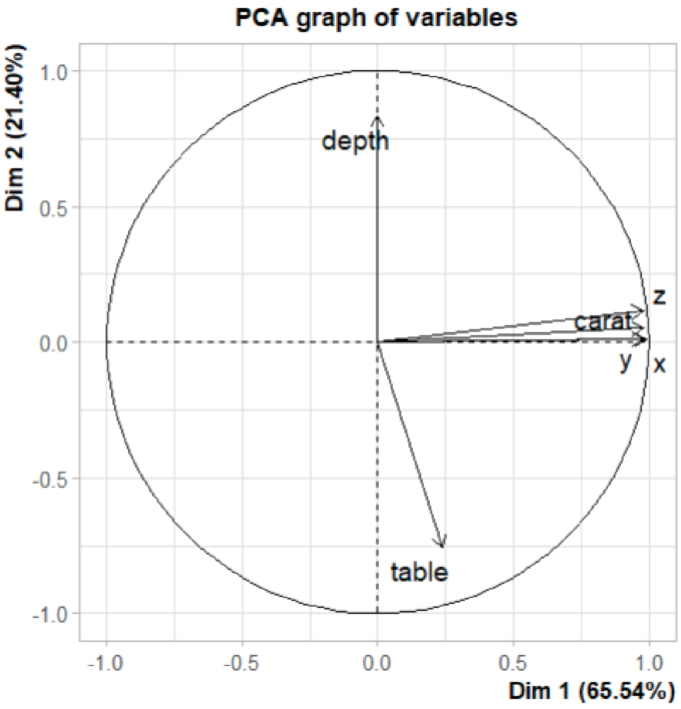
many components are needed to account for at least 90% of the total variance?

b) Judging by the loadings, what do the first two principal components measure?

c) What is the correlaHon between the first principal component and price?

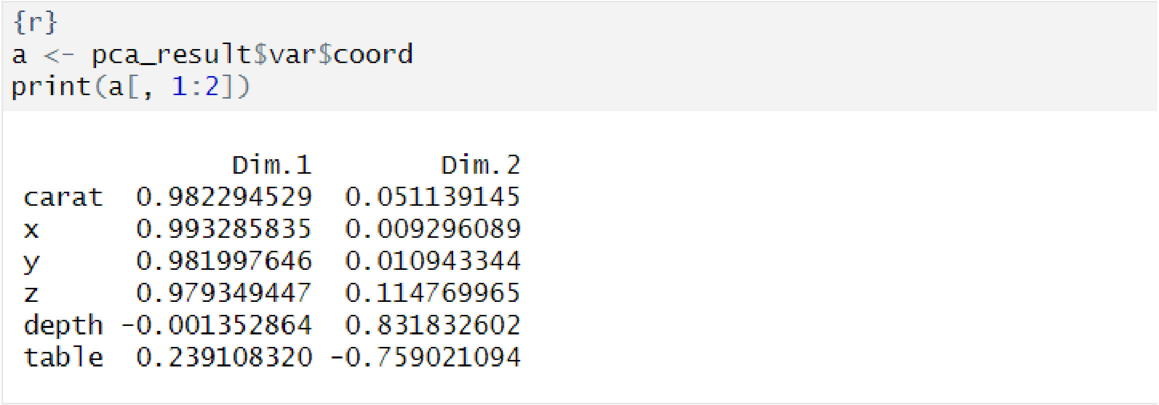
d) Can the first two principal components be used to disHnguish between diamonds

with different cuts



The first principal component accounts for about 65.54% of the total variance. To achieve at least 90% of the total variance explained, the first three components are necessary, cumulatively which is approximately 98.33%. This highlights the importance of these components in representing the dataset's variability.

b. Judging by the loadings, what do the first two principal components measure?

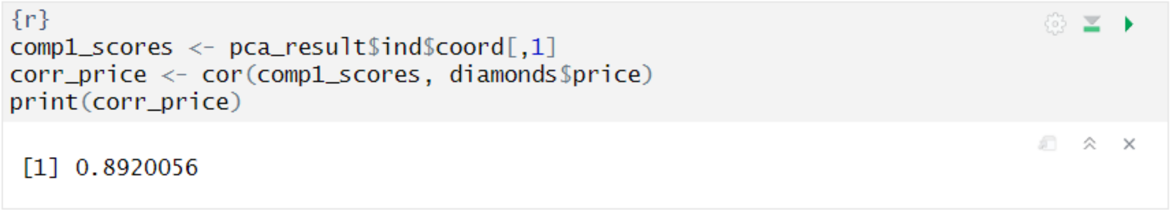


```
{r}
a <- pca_result$var$coord
print(a[, 1:2])
```

```
             Dim.1        Dim.2
carat  0.982294529  0.051139145
x      0.993285835  0.009296089
y      0.981997646  0.010943344
z      0.979349447  0.114769965
depth -0.001352864  0.831832602
table  0.239108320 -0.759021094
```

The first primary component, which has large loadings on carat, x, y, and z, indicates the overall size of the diamonds. The second principal component evaluates features of cut quality and proportions orthogonal to size. It highlights variations in the form and cut of the diamonds by differing loadings on depth and table.

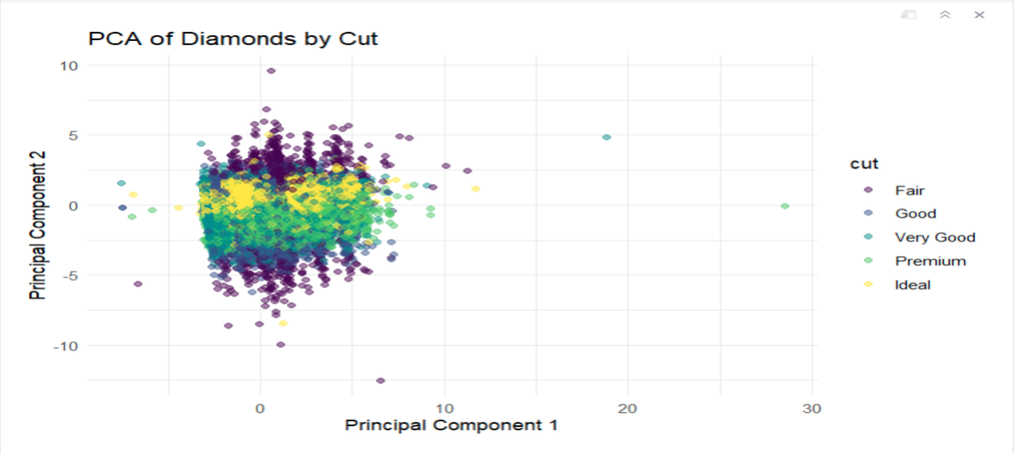
c. What is the correlation between the first principal component and price?



```
{r}
comp1_scores <- pca_result$ind$coord[,1]
corr_price <- cor(comp1_scores, diamonds$price)
print(corr_price)
```

```
[1] 0.8920056
```

The correlation between the first principal component and the price of the diamonds is 0.8920056, indicating a very strong positive relationship. This indicate that the principal component, which primarily measures the overall size of the diamonds, is a significant predictor of their price. The closer this value is to 1, the stronger the linear relationship. The size-related measures increase, so does the price, in a strong linear way.



The first two primary component’s scatter plot, colored by diamond cut, shows how the various cut qualities overlap. There is some clustering, with 'Ideal' cuts appearing to be more central and denser, but no group is totally isolated from the others. This suggests that although there can be patterns linking specific cuts to areas inside the PCA space, the initial two principal components do not effectively distinguish diamonds based just on their cut quality.

Consider the Iris data

>data(iris)

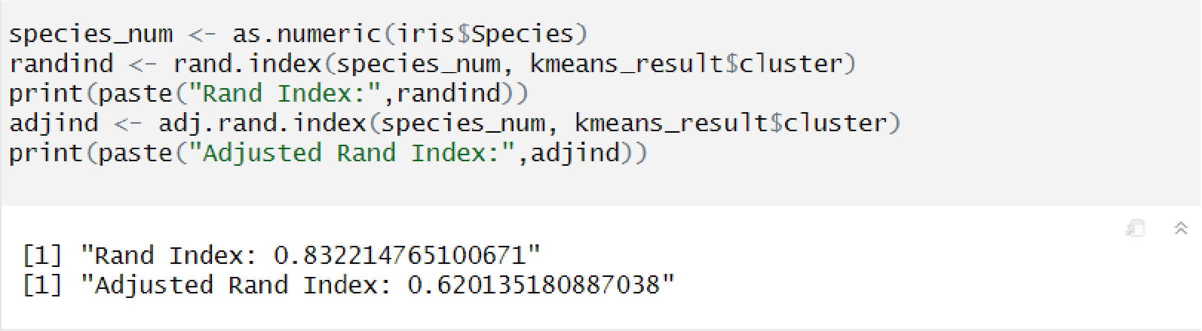
a) Create a plot using the first two principal components, and color the iris species by class.

b) Perform k-means clustering on the first two principal components of the iris data. Plot the clusters different colors, and the specify different symbols to depict the species labels.

c) Use rand index and adjusted rand index to assess how well the cluster assignments capture the species labels.

d) Use the gap staHsHc and silhoue[e plots to determine the number of clusters.

e) Reflect on the results, especially c-d. What does this tell us about the clustering?



```
species_num <- as.numeric(iris$Species)
randind <- rand.index(species_num, kmeans_result$cluster)
print(paste("Rand Index:",randind))
adjind <- adj.rand.index(species_num, kmeans_result$cluster)
print(paste("Adjusted Rand Index:",adjind))
```

```
[1] "Rand Index: 0.832214765100671"
[1] "Adjusted Rand Index: 0.620135180887038"
```

The Rand Index of 0.8322 indicates a high similarity between k-means clusters and the true iris species labels, while the Adjusted Rand Index of 0.6201, suggests a strong agreement. These values demonstrate that the clustering captures the natural groupings of the Iris dataset.

Reflect on the results, especially c-d. What does this tell us about clustering?

The results of the Rand Index and Adjusted Rand Index show that k-means clustering and the genuine Iris species labels match well. the Adjusted Rand Index accounts for chance to give a more accurate assessment. The best number of clusters, as determined by the gap statistics is three,which matches with the actual number of species. This indicates that PCA and k-means were accurate in revealing the underlying structure of the Iris dataset. These results validate the strategy for exploratory analysis and pattern recognition by highlighting the ability of unsupervised learning approaches to identify occurring groupings within data.

Consider the wine quality data (h[ps://archive.ics.uci.edu/dataset/186/wine+quality)

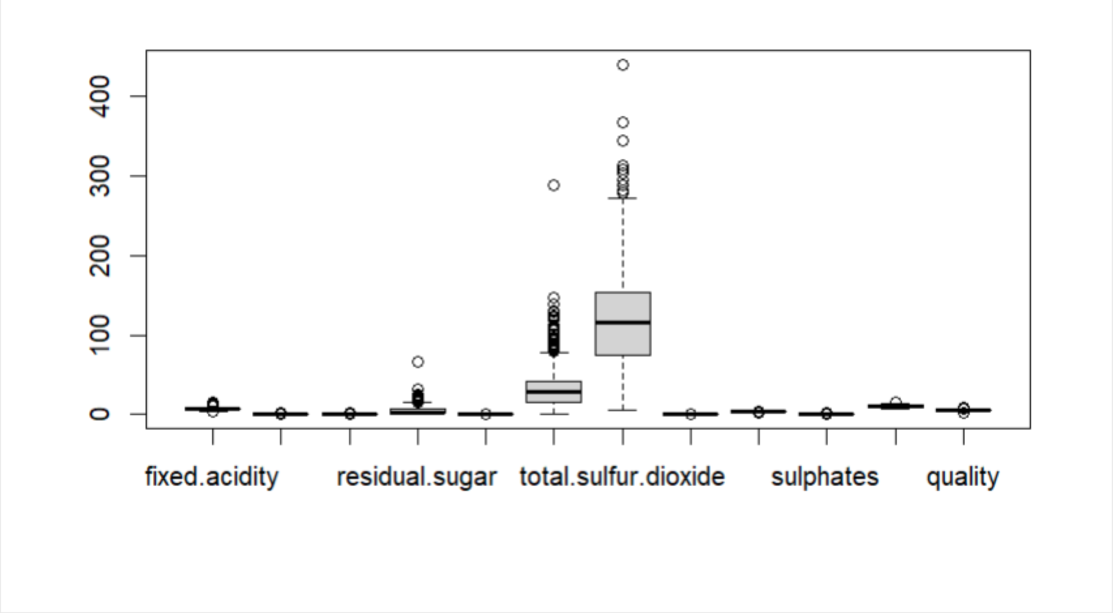
a) Perform exploratory data analysis on the data. Summarize the data quality and characterisHcs. Discuss any apparent outliers and associaHons.

b) Perform k-means using Principal Components of the wine data. JusHfy your choice of "k". Visualize the result using a biplot and color the points (samples) according to "wine color".

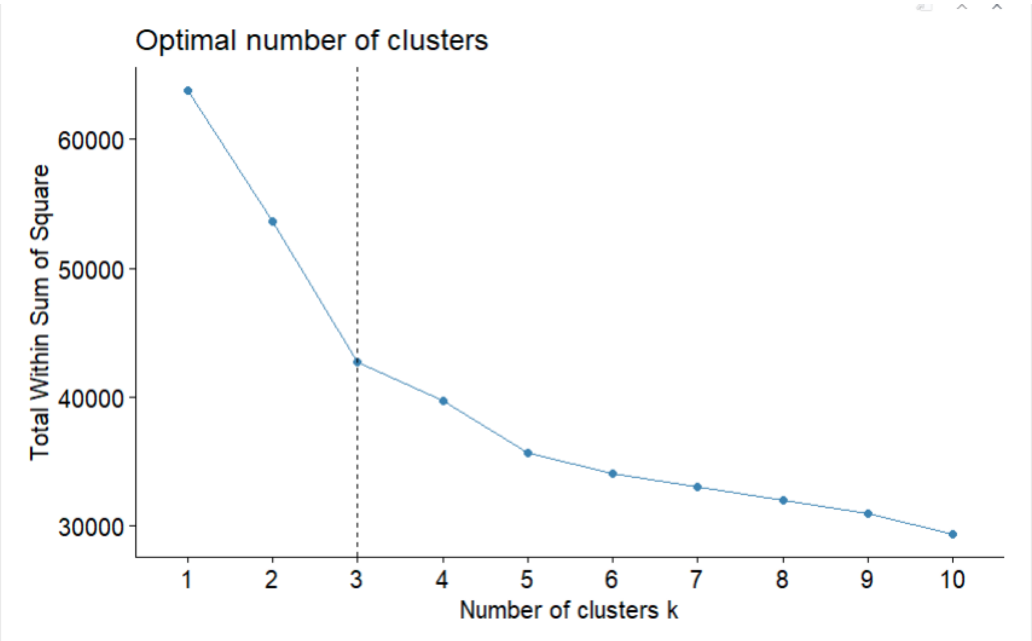
c) Fit an SOM and color the samples according to wine color. Cluster the codebook vectors of the prototypes using hclust.

d) Construct phase-plots (aka component planes) for some of the variables in the dataset.

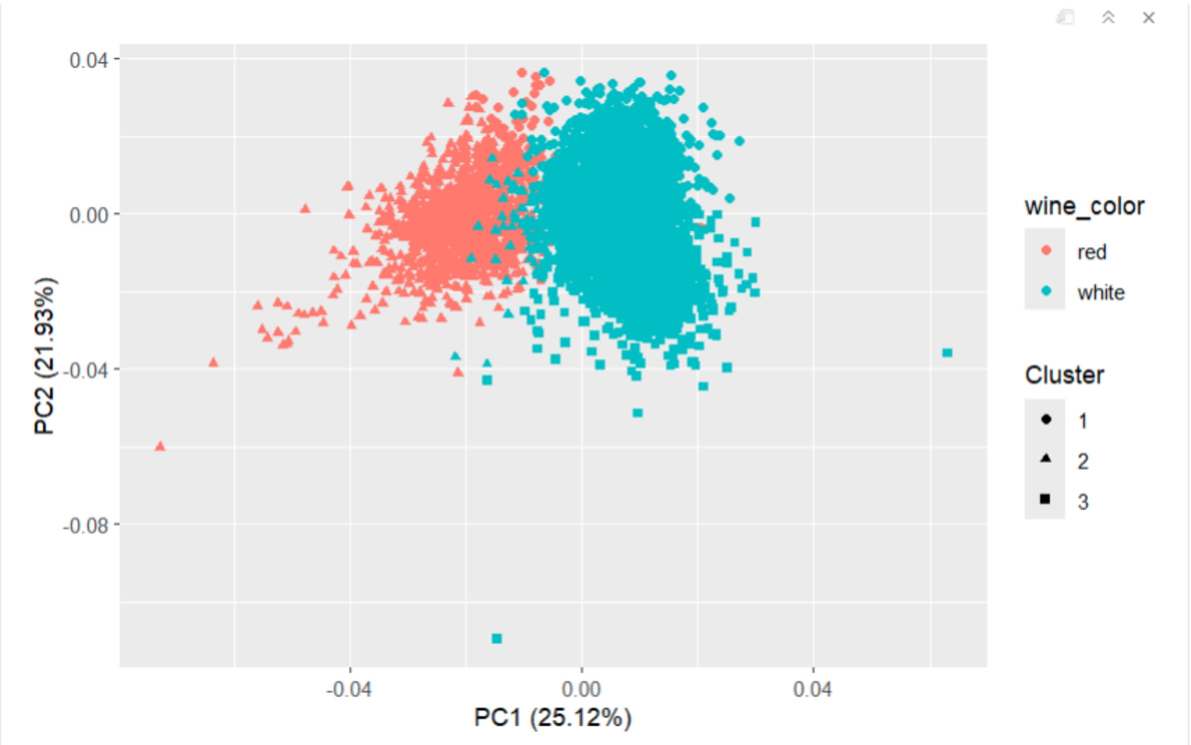
e) Comment on the differences between b and c.



The boxplot for the wine quality dataset reveals numerous outliers, particularly in total.sulfur.dioxide, which could affect the clustering outcomes. Distributions vary across variables, indicating the need for data normalization.The dataset contains duplicates that should be addressed to ensure the integrity of Dataset.



The elbow plot indicates an optimal k-value of 3 for clustering by the point where the within-cluster sum of squares begins to decline more slowly, marking the elbow. Choosing k = 3 will balance between minimizing within-cluster variance and avoiding overfitting with too many clusters. This number represents a meaningful separation in the wine dataset, capturing significant variance between the clusters while keeping the model simple.



About 47% of the variance is explained by the first two principal components, indicating that while they capture significant aspects of wine, they do not be considered for all of it. Wines are not strictly separated by color in k-means-formed clusters, suggesting that there are more complex elements to consider than just hue. This plot helps explain how wines are grouped according to their characteristics and may suggest more slight differences than color classification can convey. Clearer separations might result from additional research using more components or by using different clustering techniques.

Comment on the differences between b and c.

The k-means clustering applied to the principal components of the wine data (b) indicated clusters that, while not strictly separating wines by color, did show some variation based on the most significant variances captured by PCA. As seen in the biplot, the k-means method offered a straightforward, linear segmentation of the data that is very helpful for locating large differences in the dataset.

The mapping of the wine data, however, was more complex and thorough in the SOM analysis (c). The SOM grid, which uses color coding based on wine color, revealed that it is more difficult to distinguish between red and white wines, pointing to a more intricate interaction between the many wine characteristics. Additional levels of structure were produced by the hierarchical clustering that was then applied to the SOM codebook vectors seen in the dendrogram. This highlighted a hierarchical organization within the data that is not revealed by k-means clustering.

Part (d)'s component planes provide additional information about the contributions of the individual variables throughout the grid, enhancing our comprehension of how each chemical measurement affects cluster formation.

The principal difference between the two methods is how they deal with data complexity. By distinct, non-overlapping clusters, the k-means method aims to reduce complexity and is perfect for spotting broad trends. SOMs have complexity, permitting smooth transitions between various wine properties and overlapping clusters, which may better capture the insights of the real world found in the data.

The analytical goals should guide the decision between SOM and k-means. K-means provides an easily interpreted high-level view that might serve as a useful foundation for additional investigation. SOMs, on the other hand, work well in exploratory analysis, where catching the complex patterns in the data is more crucial than achieving instant clarity. If the objective of the wine quality dataset is to establish distinct market segments, then k-means might be a better fit. SOM would offer a deeper and more thorough viewpoint if the objective is to investigate the minute details of wine composition.