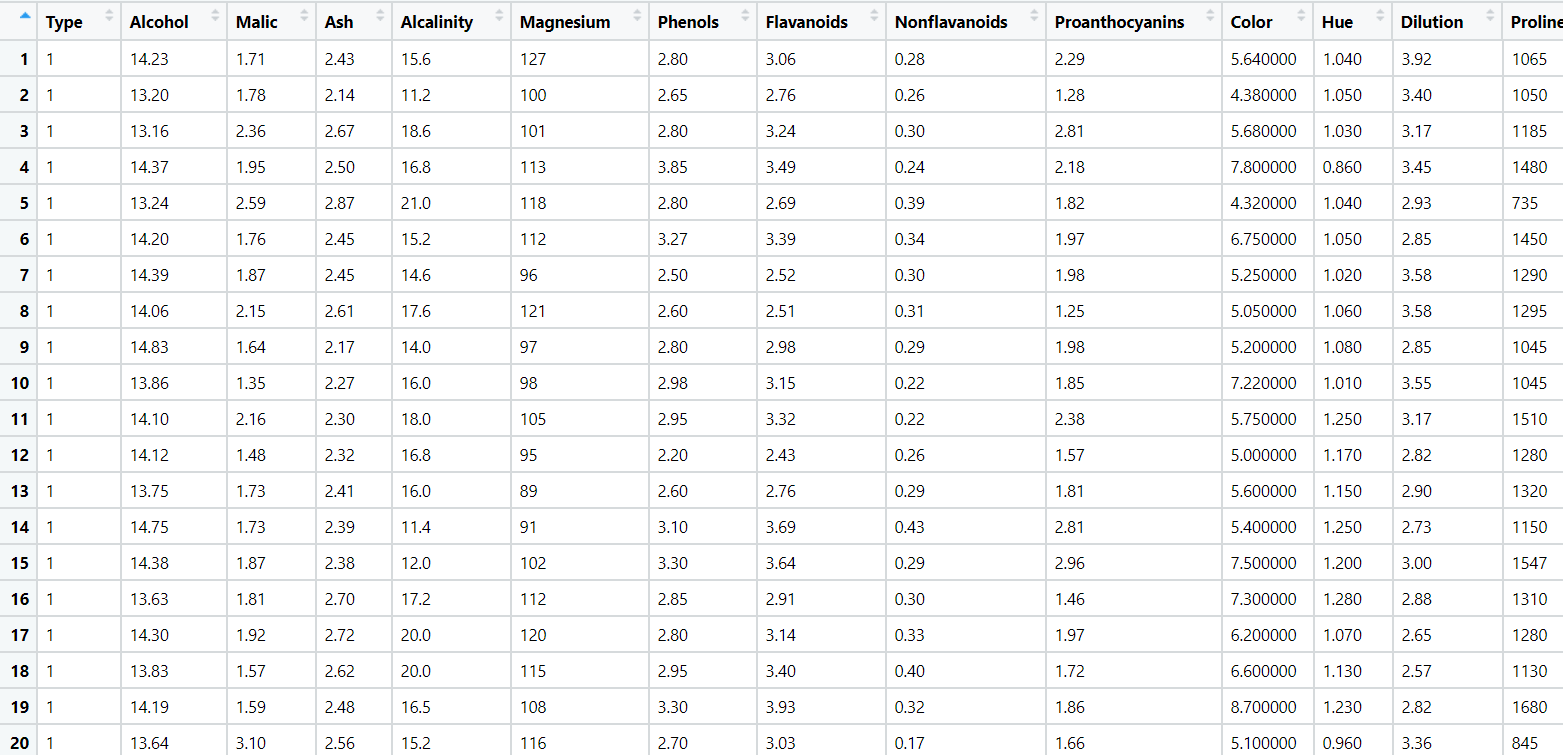
# Topic: Dimension Reduction (PCA)

Perform Principal component analysis and perform clustering using first 3 principal component scores (both Hierarchical & K-Mean clustering). Use Scree plot or elbow curve and obtain optimum number of clusters and check whether we have obtained same number of clusters with the original data



**Data Preprocessing:**

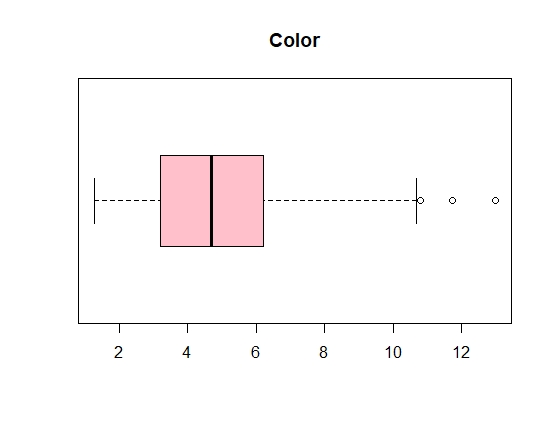
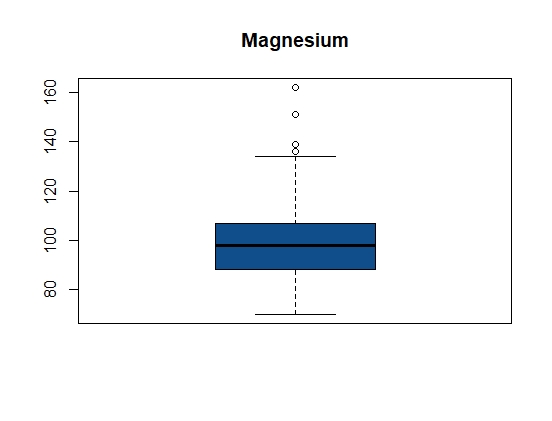
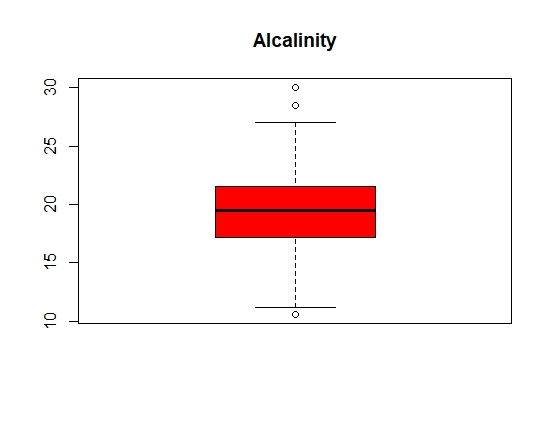
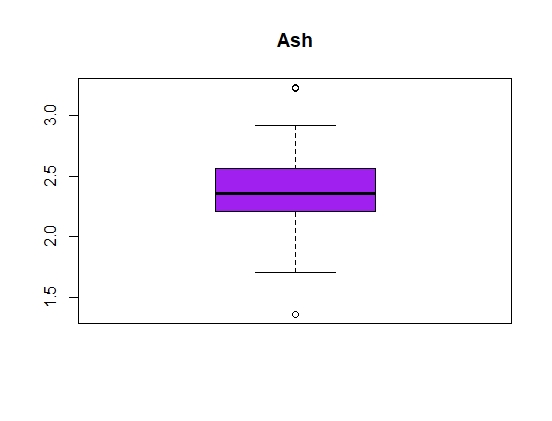
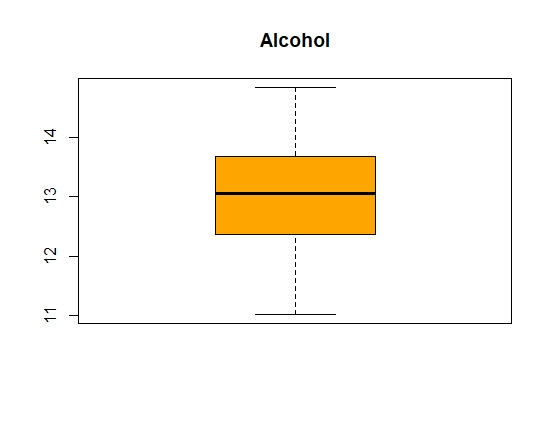
* Importing the dataset using read.csv.
* Removing unnecessary columns.

**Normalizing the data:**

* Using the normalization function, the data is normalized and the data will come under same level.

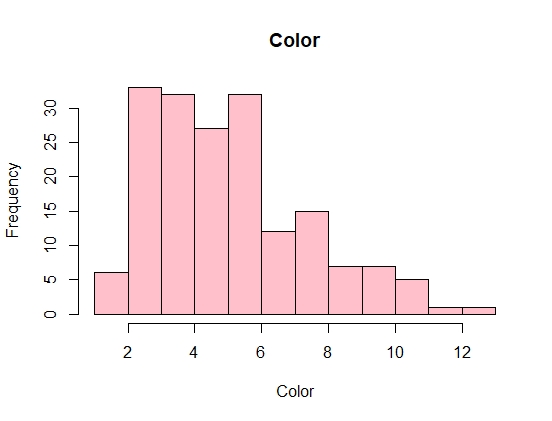
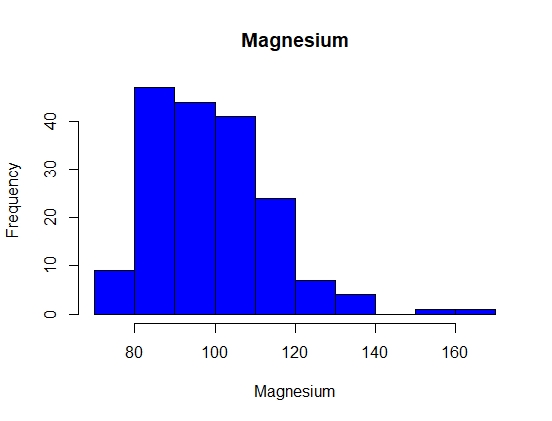
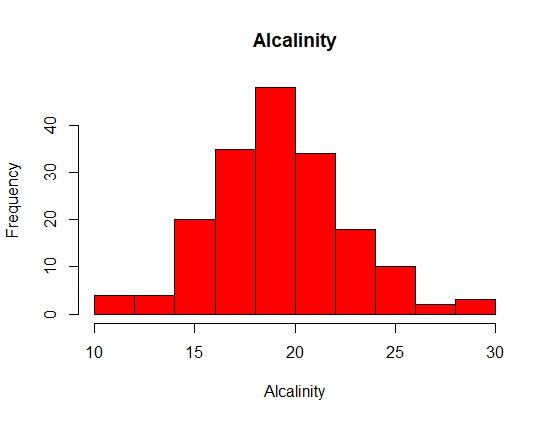
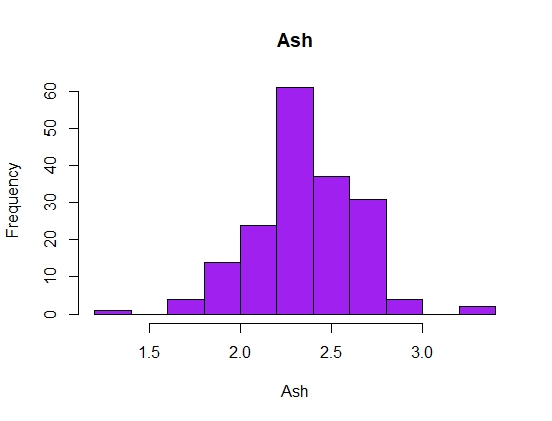
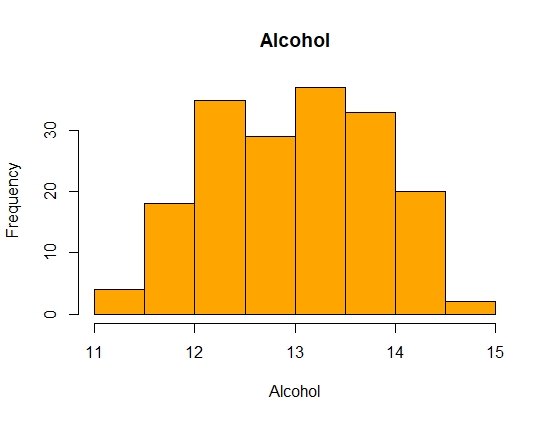
**Exploratory Data Analysis:**

**Box Plot Representation:**



* From the above graphical representation, it shows that outliers exist in all variables except Alcohol.

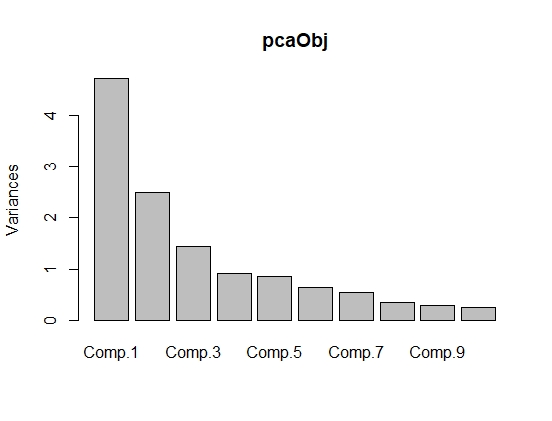
**Histogram Representation:**



* From the above graphical representation, It shows that the data is normally distributed in all the variables except color, the data is right skewed.

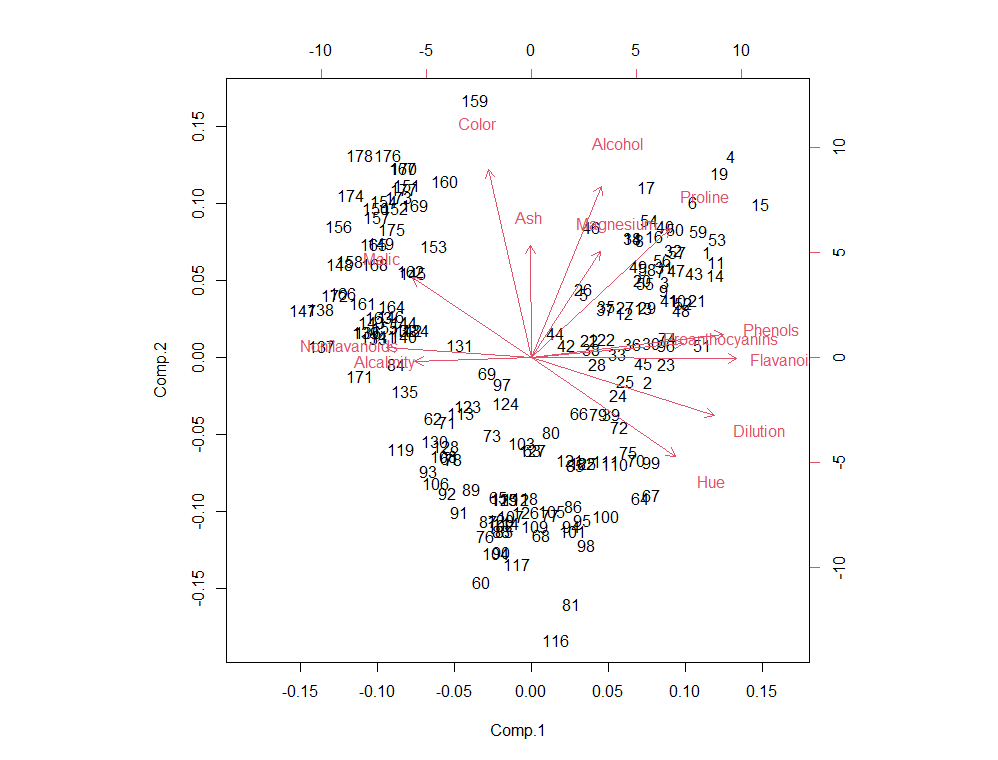
**DIMENSION REDUCTION (PCA) ANALYSIS:**

1. Two main goals of dimensionality reduction
   1. find structure in features
   2. aid in visualization
2. PCA has 3 goals
   1. find a linear combination of variables to create principle components
   2. maintain as much variance in the data as possible
   3. principal components are uncorrelated
3. Intuition
   1. with an x y correlation scatter plot, the best 1 dimension to explain the variance in the data is the linear regression line
   2. this is the first principal component
   3. then the distance of the points from the line is the component score



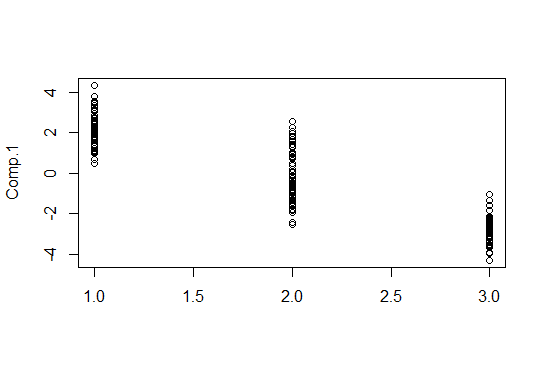
* In the above graph the first three components show higher variance compare to rest all components.

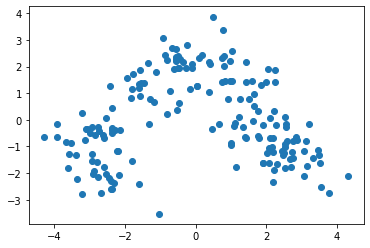
**Biplot**

****

* The graph shows at least two clearly distinguishable clusters. This factoid tells us that the observations in the dataset can be grouped.
* Biplot does not show all the meaningful principal components, K-means clustering will give clearer visuality to the minute level in terms of other reduced dataset. Also, in this scenario or the dataset we do not have the target variable to label the groups. Hence, it is unknown what the labels are in this scenario.

**Scatter Plot**

****

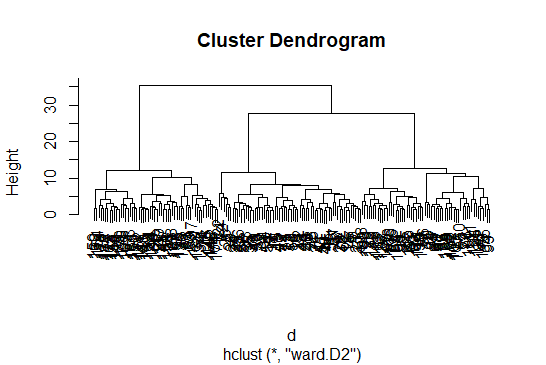
****

* The above graphs represent how the components are scattered and form a cluster.

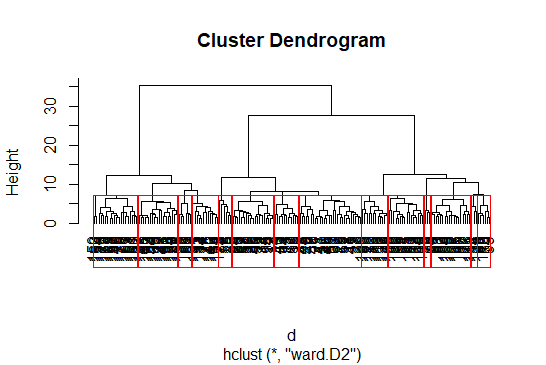
**HIERARCHICAL CLUSTERING**

**Dendrogram:**

1. Standardization/Normalization of data is required otherwise the issues will influence the model as shown in first dendogram.
2. Distance measure will be calculated incorrectly, if all the variables are not brought to equal weight.
3. Largest scale dominating the measure
4. The below dendogram generated with the given data set with standardization is as below.
5. Using **ward.D2** method the dendogram is prepared

****

1. Now the dendogram is divided into clusters and the aggregate of each cluster is compared with the other clusters.



1. Lastly, the output is stored into working directory in a csv format.

**K MEANS CLUSTERING**

**Elbow curve to decide the k value:**

str(fit) --🡪 function to fit the structure / information of K means

List of 9

$ cluster : int [1:178] 3 3 3 3 3 3 3 3 3 3 ...

$ centers : num [1:3, 1:13] -0.923 0.164 0.833 -0.393 0.869 ...

..- attr(\*, "dimnames")=List of 2

.. ..$ : chr [1:3] "1" "2" "3"

.. ..$ : chr [1:13] "Alcohol" "Malic" "Ash" "Alcalinity" ...

$ totss : num 2301

$ withinss : num [1:3] 559 326 386

$ tot.withinss: num 1271

$ betweenss : num 1030

$ size : int [1:3] 65 51 62

$ iter : int 3

$ ifault : int 0

- attr(\*, "class")= chr "kmeans"

1. The str () function gives the structure of the kmeans which includes various parameters like withinss, betweenss, etc, analyzing which you can find out the performance of kmeans.
2. betweenss: Between sum of squares i.e. Intracluster similarity
3. withinss: Within sum of square i.e. Intercluster similarity
4. totwithinss: Sum of all the withinss of all the clusters i.e. Total intra-cluster similarity
5. A good clustering, will have a lower value of withinss and higher value of betweenss which depends on the number of clusters ‘k’ chosen initially. Let us see how we can find the optimal value of ‘k’.

**Finding the optimal value of ‘k’**

An optimal value of ‘k’ is the value which gives us a converged set of clusters with minimum distortion. Greater the distortion, worse will be the clusters formed

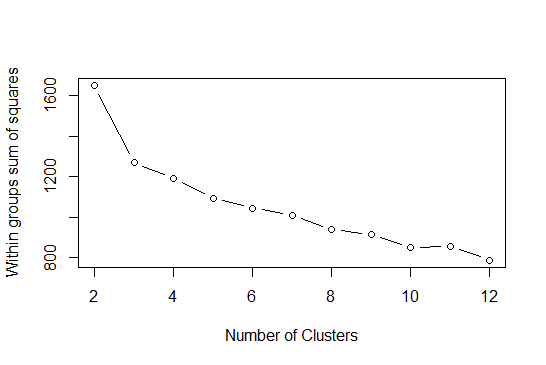
twss <- NULL

for (i in 2:8) {

twss <- c(twss, kmeans(normalized\_data, centers = i)$tot.withinss)

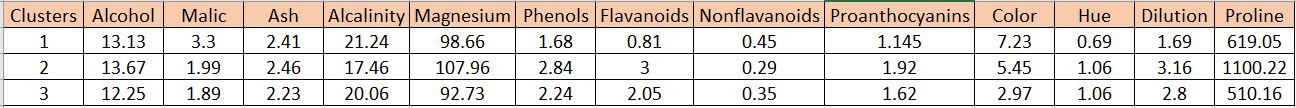
}

Twss



This is the plot between ‘k’, the number of clusters and the ‘totwithinss’ (or distortion) for each value of k. You can see when the number of cluster is less, there is a gradual decrease in distortion but as we keep on increasing the value of k, the rate of reduction of distortion values becomes constant.

This value of k beyond which the distortion rate becomes constant is the optimal value. Here k=3.

****

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**INFERENCES DRAWN FROM ABOVE ANALYSIS**

* **Cluster1 has high in color , alcalinity and Nonflavonoids, the other factors are moderate when compared with cluster2**
* **Cluster2 has high in all the factors except Malic, Nonflavonoids, alcalinity**
* **Cluster 3 all factors are low when compared to cluster 1 and cluster 2**

# Hints:

1. Business Problem
   1. Objective
   2. Constraints (if any)
2. Data Pre-processing

2.1 Data cleaning, Feature Engineering, EDA etc.

1. Model Building
   1. Partition the dataset
   2. Model(s) - Reasons to choose any algorithm
   3. Model(s) Improvement steps
   4. Model Evaluation
   5. Python and R codes
2. Deployment

4.1 Deploy solutions using R shiny and Python Flask.

1. Result Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.

**Note:**

1. For each assignment the solution should be submitted in the format
2. Research and Perform all possible steps for improving the model(s) accuracy Ex: Feature Engineering, Hyper Parameter tuning etc.
3. All the codes (executable programs) are running without errors
4. Documentation of the module should be submitted along with R & Python codes, elaborating on every step mentioned here