# Topic: Dimension Reduction (PCA)

**Instructions**

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name: Shreyas A B Batch Id: 05012021-10AM** **Topic: Principal Component Analysis**

1. **Business Problem**
   1. **Objective**
   2. **Constraints (if any)**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**2.1 Make a table as shown above and provide information about the features such as its Data type and its relevance to the model building, if not relevant provide reasons and provide description of the feature.**

**Using R and Python codes perform:**

1. **Data Pre-processing**

**3.1 Data Cleaning, Feature Engineering, etc.**

1. **Exploratory Data Analysis (EDA):**
   1. **Summary**
   2. **Univariate analysis**
   3. **Bivariate analysis**
2. **Model Building**
   1. **Build the model on the scaled data (try multiple options)**
   2. **Perform PCA analysis and get the maximum variance between components**
   3. **Perform clustering before and after applying PCA to cross the number of clusters formed.**
   4. **Briefly explain the model output in the documentation.**

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

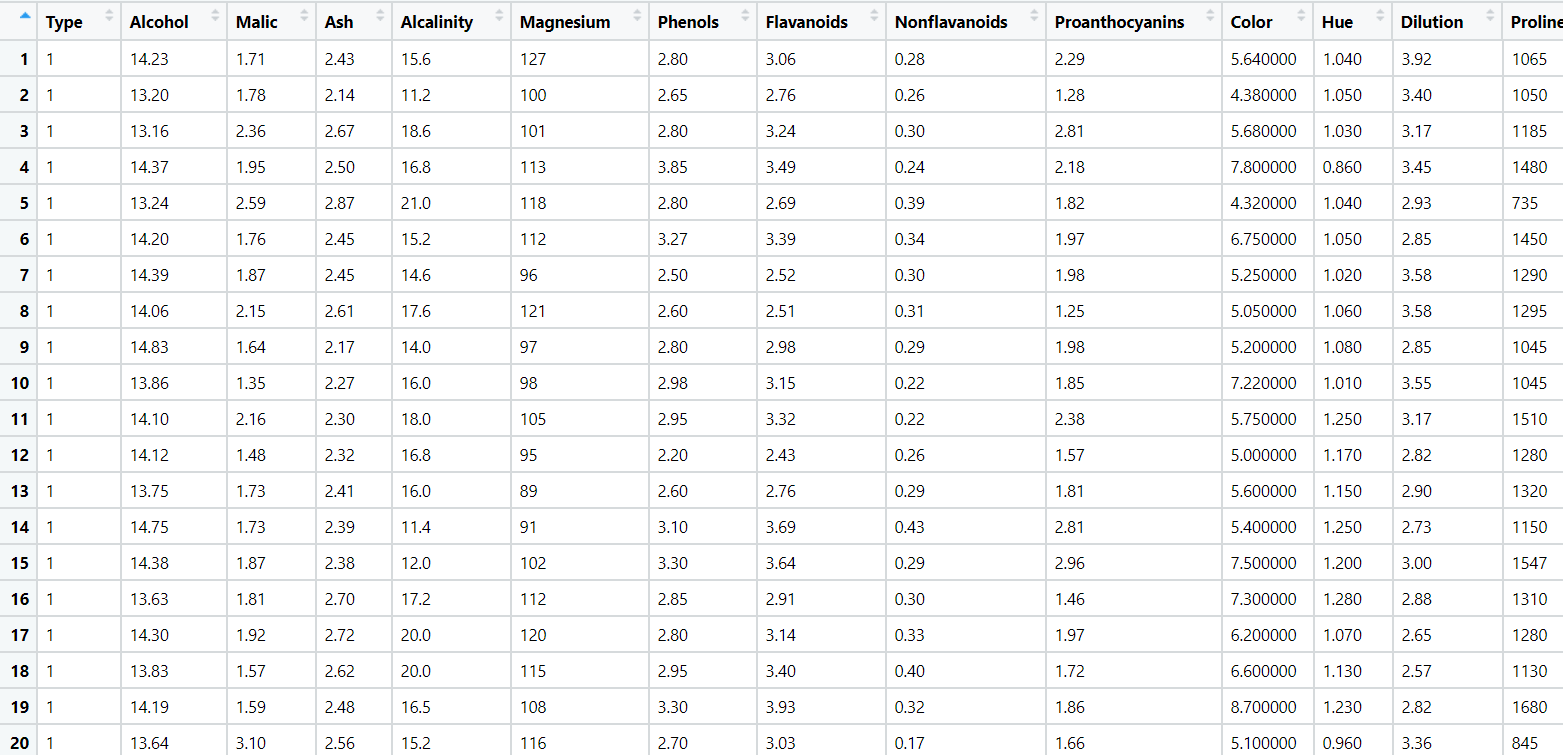
# Note:

The assignment should be submitted in the following format:

* R code
* Python code
* Code Modularization should be maintained
* Documentation of the model building (elaborating on steps mentioned above)

**Problem Statement: -**

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



Python code steps :\_

Ans = Data Frame name is wine. It has 178 rows and 14 columns.

(**Work on each feature of the dataset to create a data dictionary as displayed in the below image:**

Then we create a data frame that’s contain details of each columns ,like- description ,data types ,and save the details named as data\_details .all of them are important

**.Data Pre-processing**

Dataframe has no missing values so we create a new data frame without “Type#” columns named as wine\_1 . We also check for unique values for each columns named as “col\_uni” .

**Exploratory Data Analysis (EDA):**

. now we do exploratory data analysis for whole data frame and saved the details as “EDA”.

For graphical representation we have done scatter and histogram using seaborn libarary .

for continuous data we have done min max scaling and for discrete data we performed onehot encoding and saved as wine\_norm .

**Perform clustering before applying PCA to cross the number of clusters formed.**

**For Hierarchical clustering wine\_norm dataframe**

after that we go for dendrogram on normalized data (wine\_norm).and look for clusters ,so we can check for cluster number .

After we have seen dendrogram we go for 3 cluster number and check the details of each cluster .

we go with 3 cluster and save the details as 'Hclust3\_bfPCA' as a column in wine\_1 df.

**Perform clustering before applying PCA to cross the number of clusters formed.**

**For kmeans clustering wine\_norm dataframe**

we go for scree plot(elbow curve ) where k range from 2 to 6 on normalized data (wine\_norm).and look for clusters ,so we can check for cluster number .

After we have seen elbow curve we go for 3cluster number and check the details of each cluster .

we go with 3 cluster and save the details as Kclust3\_bfPCA.

**Perform PCA analysis and get the maximum variance between components**

for PCA 1st we have set componets number (equal number to columns) and randomly assign weight and The amount of variance that each PCA save as var.

and after all we go with 6 PCA value .pca scores save as pca\_values

**Perform clustering after applying PCA to cross the number of clusters formed.**

**For Hierarchical clustering on wine\_PCA data frame**

after that we go for dendrogram on normalized data (wine\_PCA).and look for clusters ,so we can check for cluster number .

After we have seen dendrogram we go for 4 cluster number and check the details of each cluster .

we go with 3 cluster and save the details as 'Hclust4\_afPCA' as a column in wine\_1 df.

**Perform clustering before applying PCA to cross the number of clusters formed.**

**For kmeans clustering on wine\_PCA data frame**

we go for scree plot(elbow curve ) where k range from 2 to 6 on normalized data (wine\_PCA).and look for clusters ,so we can check for cluster number .

After we have seen elbow curve we go for 3cluster number and check the details of each cluster .

we go with 3 cluster and save the details as Kclust3\_afPCA as a column in wine\_1 df.

**Problem Statement: -**

A Pharmaceutical drug manufacturing company is studying on a new medicine to treat Heart diseases, it has gathered data from its secondary sources, and it would like you to provide high level analytical insights on the data, its aim is to segregate patients depending on their age group and other factors as given in the data, perform PCA and Clustering Machine learning Algorithm on the dataset given, and check if the clusters formed before and after PCA are same and provide a brief report on your model. You can also explore more on ways to improve your model.

Note: - This is just a snap shot of the data, the datasets can be downloaded from Aispry LMS in the Hands on Material section.

A screenshot of a cell phone

Description automatically generated

Python code steps :\_

Ans = Data Frame name is heart. It has 303 rows and 14 columns.

(**Work on each feature of the dataset to create a data dictionary as displayed in the below image:**

Then we create a data frame that’s contain details of each columns ,like- description ,data types ,and save the details named as data\_details .all of them are important

**.Data Pre-processing**

Dataframe has no missing values so we create a new data frame with rearrange of columns named as heart\_1 . We also check for unique values for each columns named as “col\_uni” .

**Exploratory Data Analysis (EDA):**

. now we do exploratory data analysis for whole data frame and saved the details as “EDA”.

For graphical representation we have done scatter and histogram using seaborn libarary .

for continuous data we have done min max scaling and for discrete data we performed onehot encoding and saved as heart\_norm .

**Perform clustering before applying PCA to cross the number of clusters formed.**

**For Hierarchical clustering heart\_norm dataframe**

after that we go for dendrogram on normalized data (heart\_norm).and look for clusters ,so we can check for cluster number .

After we have seen dendrogram we go for 4 cluster number and check the details of each cluster .

we go with 4 cluster and save the details as 'Hclust4\_bfPCA' as a column in heart\_1 df.

**Perform clustering before applying PCA to cross the number of clusters formed.**

**For kmeans clustering heart\_norm dataframe**

we go for scree plot(elbow curve ) where k range from 2 to 6 on normalized data (air\_norm).and look for clusters ,so we can check for cluster number .

After we have seen elbow curve we go for 4cluster number and check the details of each cluster .

we go with 4 cluster and save the details as Kclust4\_bfPCA.

**Perform PCA analysis and get the maximum variance between components**

for PCA 1st we have set componets number (equal number to columns) and randomly assign weight and The amount of variance that each PCA save as var.

and after all we go with 8 PCA value .pca scores save as pca\_values

**Perform clustering after applying PCA to cross the number of clusters formed.**

**For Hierarchical clustering on heart\_PCA data frame**

after that we go for dendrogram on normalized data (heart\_PCA).and look for clusters ,so we can check for cluster number .

After we have seen dendrogram we go for 4 cluster number and check the details of each cluster .

we go with 4 cluster and save the details as 'Hclust4\_afPCA' as a column in heart\_1 df.

**Perform clustering before applying PCA to cross the number of clusters formed.**

**For kmeans clustering on heart\_PCA data frame**

we go for scree plot(elbow curve ) where k range from 2 to 6 on normalized data (air\_norm).and look for clusters ,so we can check for cluster number .

After we have seen elbow curve we go for 4cluster number and check the details of each cluster .

we go with 4 cluster and save the details as Kclust4\_afPCA as a column in heart\_1 df