Topic: Dimension Reduction With PCA

# Instructions:

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

# Name: Nukala Ayyappa Bharthwaj Batch ID : DSWDMCON 21012022 Topic: Principal Component Analysis

**Grading Guidelines:**

1. **An assignment submission is considered complete only when correct and executable code(s) are submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered for evaluation.**
2. **Assignments submitted after the deadline will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* + **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline
  + **Grade B: (>= 80 and < 90):**
    - When assignments are submitted on time but less than 80% of problems are completed.

(OR)

* + - All assignments are submitted after the deadline.
  + **Grade C: (>= 70 and < 80):**
    - When assignments are submitted on time but less than 50% of the problems are completed.

(OR)

* + - Less than 80% of problems in the assignments are submitted after the deadline
  + **Grade D: (>= 60 and < 70):**
    - Assignments submitted after the deadline and with 50% or less problems.
  + **Grade E: (>= 50 and < 60):**
    - Less than 30% of problems in the assignments are submitted after the deadline

(OR)

* + - Less than 30% of problems in the assignments are submitted before deadline
  + **Grade F: (< 50):** No submission (or) malpractice.

# Hints:



1. **Business Problem**
   1. **What is the business objective?**

**1.1. Are there any constraints?**

1. **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. And if not relevant, provide reasons and a description of the feature.**

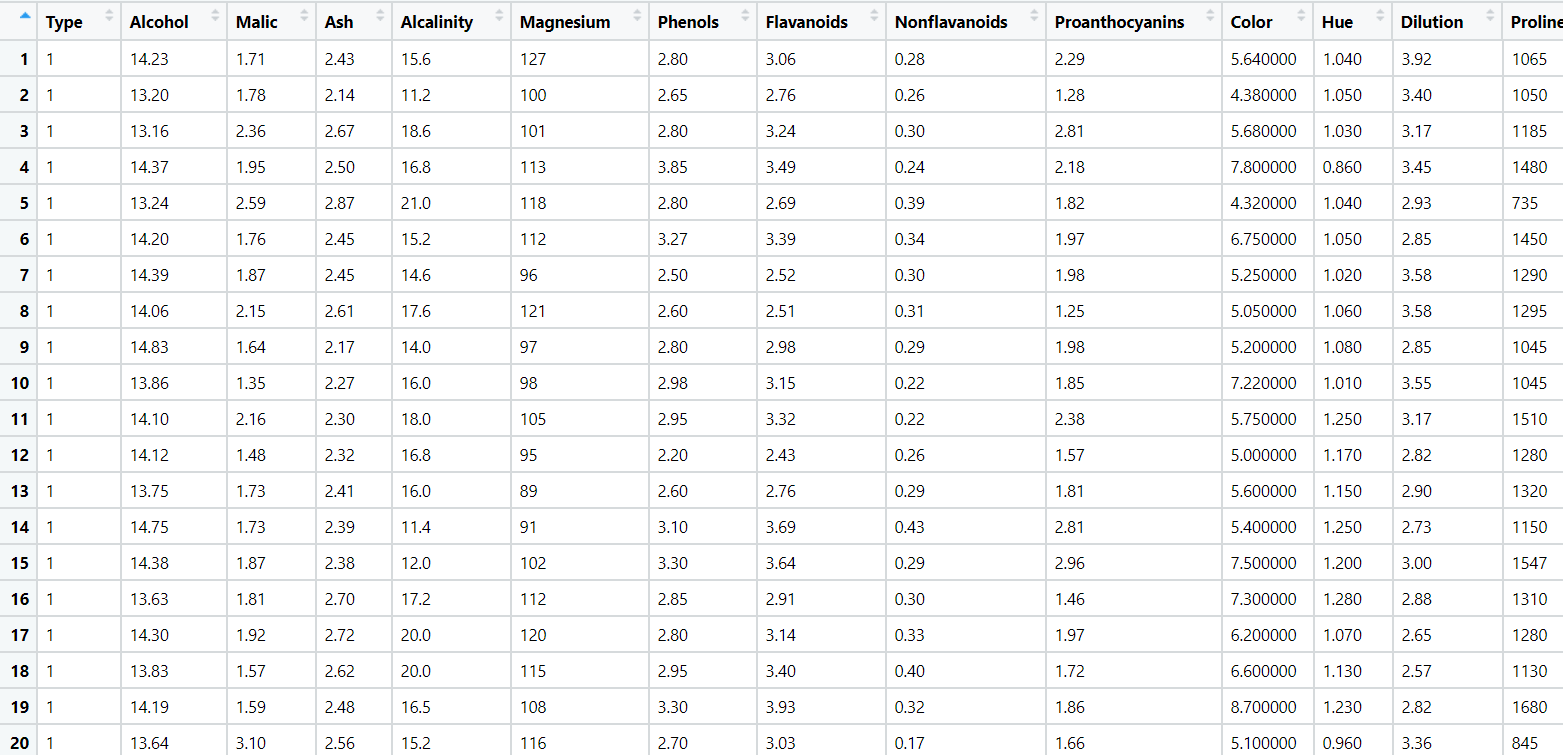
1. **Data Pre-processing**

**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.**
3. **Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provide**

# Problem Statement: -

Perform hierarchical and K-means clustering on the dataset. After that, perform PCA on the dataset and extract the first 3 principal components and make a new dataset with these 3 principal components as the columns. Now, on this new dataset, perform hierarchical and K-means clustering. Compare the results of clustering on the original dataset and clustering on the principal components dataset (use the scree plot technique to obtain the optimum number of clusters in K-means clustering and check if you’re getting similar results with and without PCA).



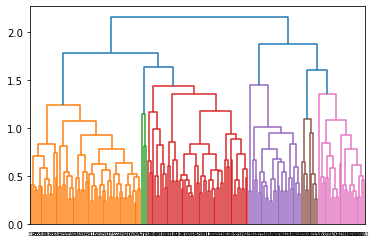
1. objective: To find the percentages of chemical concentration in the wine.

2.

|  |  |  |  |
| --- | --- | --- | --- |
| Name of the  feature | Description | Type of data | relevance |
| Type | Class category of the wine | Quantitative | relevant |
| Alcohol | Amount of Alcohol in that particular wine type | quantitative | relevant |
| Malic | Amount of Malic Acid in that particular wine type | quantitative | relevant |
| Ash | Amount of Ash in that particular wine type | quantitative | relevant |
| Alcalinity | Amount of Alcalinity in that particular wine type | quantitative | relevant |
| Magnesium | Amount of magnesium in that particular wine type | quantitative | relevant |
| Phenols | Amount of phenols in that particular wine type | quantitative | relevant |
| Flavanoids | Amount of flavonoids in that particular wine type | quantitative | relevant |
| Nonflavanoids | Amount of nonflavanoids in that particular wine type | quantitative | relevant |
| Proanthocyanins | Amount of proanthocyanins in that particular wine  type | quantitative | relevant |
| Color | Amount of colour intensity in the particular wine  type | quantitative | relevant |
| Hue | Amount of hue in the particular wine type | quantitative | relevant |

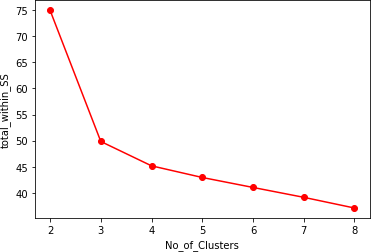
|  |  |  |  |
| --- | --- | --- | --- |
| Dilution | Amount of dilution in the particular wine type | quantitative | relevant |
| Proline | Amount of proline in the particular wine type | quantitative | relevant |

1. DATAPREPROCESSING : checked the type of data .All the data is of numeric type so no need to do the typecasting. There are no duplicates or null values as well.
2. EDA:-From the EDA we found that the mean and median are same which implies that the data is normally distributed. for Malic data ,there is positive skew or right skew.
3. MODEL BUILDING:

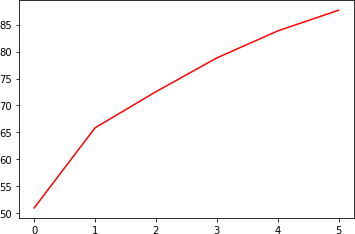


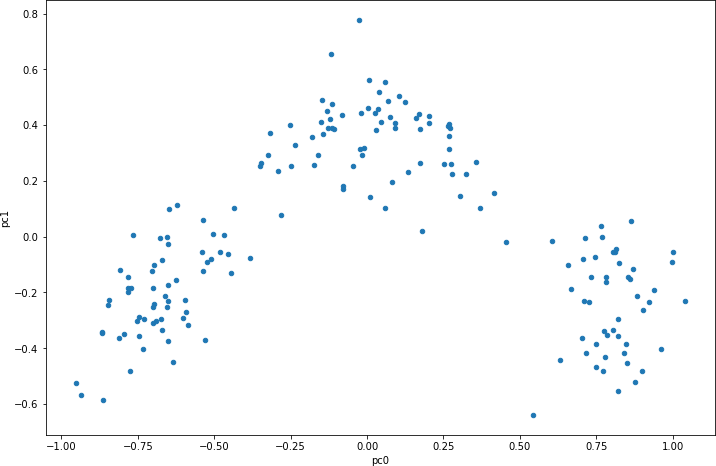
Hierarchical clustering :-

K-means clustering:

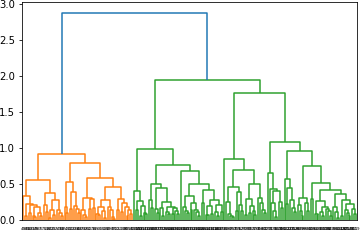


PCA:

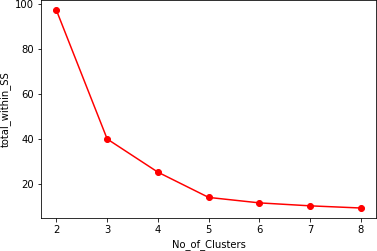


Scatter plot of PCA data:

Dendrogram of Hierarchical clustering on PCA data:



k-means elbow curve on PCA data:

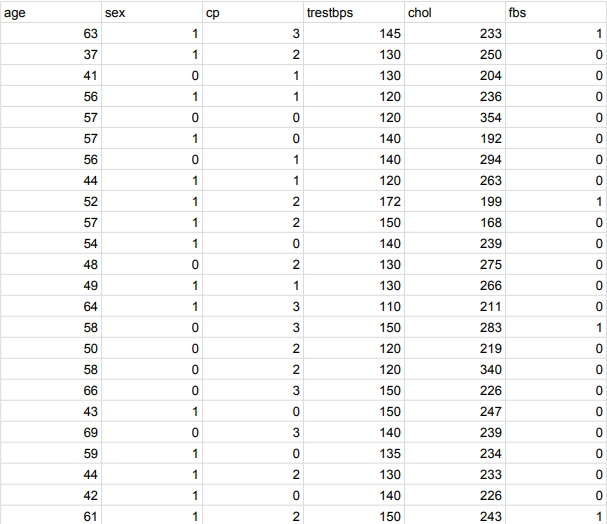


1. BENEFITS: From the model we have found out that before the pca there are 6 clusters each. but after pca ,the number of clusters decreased to 4 and 3.That shows that with the less number of columns we have derived insights .

**problem Statement: -**

A pharmaceuticals manufacturing company is conducting a study on a new medicine to treat heart diseases. The company has gathered data from its secondary sources and 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 given in the data. Perform PCA and clustering algorithms on the dataset and check if the clusters formed before and after PCA are the same and provide a brief report on your model. You can also explore more ways to improve your model.

Note: This is just a snapshot of the data. The datasets can be downloaded from AiSpry LMS in the Hands-On Material section.



Objective: is that whether that particular person has a heart disease or not and other is the experimental task to diagnose and find out various insights from this dataset which could help in understanding the problem more.

|  |  |  |  |
| --- | --- | --- | --- |
| Name of the  feature | description | Type of data | relevance |
| Age | person's age in years | Quantitative | relevant |

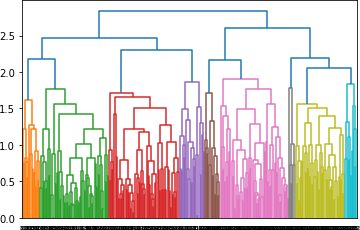
|  |  |  |  |
| --- | --- | --- | --- |
| Sex | sex of the person(male or female) | Quantitative | relevant |
| Cp | chestpain type | Quantitative | relevant |
| Trestbps | person's resting blood pressure on admission to the hospital | Quantitative | relevant |
| Chol | person's cholesterol measurement | Quantitative | relevant |
| Fbs | person's fasting blood sugar | Quantitative | relevant |
| Restecg | resting electrocardiographic  results | Quantitative | relevant |
| Thalach | persons maximum heartrate  achieved | Quantitative | relevant |
| Exang | exercise induced angina | Quantitative | relevant |
| Oldpeak | ST depression induced by exercise relative to rest | Quantitative | relevant |
| Slope | the slope of the peak exercise ST segment — 0: downsloping; 1: flat;  2: upsloping | Quantitative | relevant |
| Ca | no. of major vessels | Quantitative | relevant |
| Thal | A blood disorder called thalassemia Value | Quantitative | relevant |
| Target | heart disease yes=1.no=0 | Quantitative | relevant |

Data preprocessing:-checked for the data types and found that one column has float value which is changed to int data type. checked for the duplicate value ,found 1 and it is removed. also checked for the null values. There is no presence of null values.

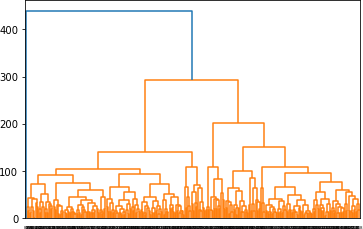
EDA:-exploratory data analysis is done on the data and also the data visualization.it is found that the data is right skewed or positive skew.

Model building:-

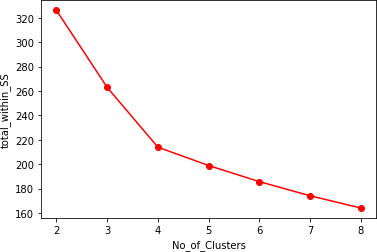
Dendrogram of hierarchical clustering



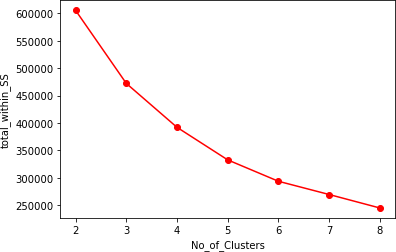
Dendrogram of hierarchical clustering after the PCA



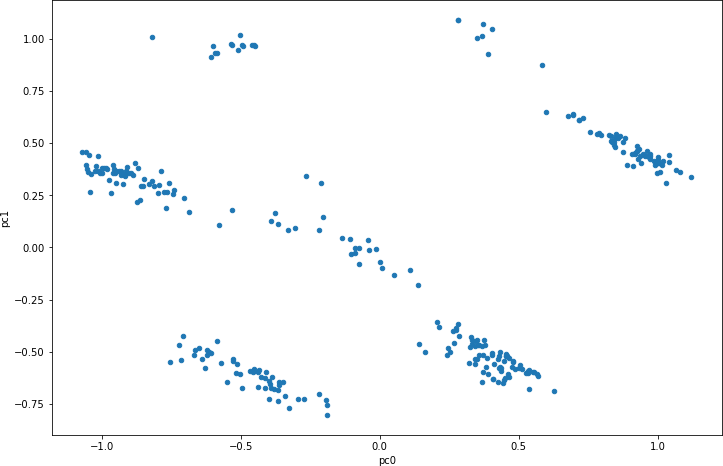
Elbow curve of Kmeans clustering:

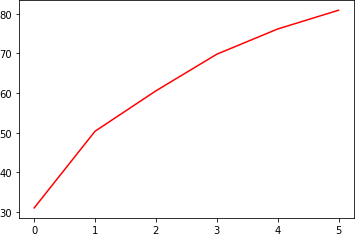


Elbow curve of Kmeans after PCA :



Scatter plot :



PCA:

Inferences: from the given data set after the clustering techniques I found out that number of clusters dropped after the PCA. we get insights from less data.in the

hierarchical clustering it is found that cluster 0 and 1 has more heart diseases where as other clusters has no heart disease.whereas after PCA there is presence of heart disease in the cluster 2.