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

1.1. What is the business objective?

1.1. Are there any constraints?

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

Name of Feature	Description	Type	Relevance
ID	Customer ID	Quantitative, Nominal	Irrelevant, ID does not provide useful information

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.

3. Data Pre-processing

3.1 Data Cleaning, Feature Engineering, etc.

4. Exploratory Data Analysis (EDA):

4.1. Summary.

4.2. Univariate analysis.

4.3. Bivariate analysis.

5. Model Building

5.1 Build the model on the scaled data (try multiple options).

5.2 Perform PCA analysis and get the maximum variance between components.

5.3 Perform clustering before and after applying PCA to cross the number of clusters formed.

5.4 Briefly explain the model output in the documentation.

6. 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).

	Type	Alcohol	Malic	Ash	Alcalinity	Magnesium	Phenols	Flavanoids	Nonflavanoids	Proanthocyanins	Color	Hue	Dilution	Proline
1	1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.640000	1.040	3.92	1065
2	1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.380000	1.050	3.40	1050
3	1	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.680000	1.030	3.17	1185
4	1	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.800000	0.860	3.45	1480
5	1	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82	4.320000	1.040	2.93	735
6	1	14.20	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.750000	1.050	2.85	1450
7	1	14.39	1.87	2.45	14.6	96	2.50	2.52	0.30	1.98	5.250000	1.020	3.58	1290
8	1	14.06	2.15	2.61	17.6	121	2.60	2.51	0.31	1.25	5.050000	1.060	3.58	1295
9	1	14.83	1.64	2.17	14.0	97	2.80	2.98	0.29	1.98	5.200000	1.080	2.85	1045
10	1	13.86	1.35	2.27	16.0	98	2.98	3.15	0.22	1.85	7.220000	1.010	3.55	1045
11	1	14.10	2.16	2.30	18.0	105	2.95	3.32	0.22	2.38	5.750000	1.250	3.17	1510
12	1	14.12	1.48	2.32	16.8	95	2.20	2.43	0.26	1.57	5.000000	1.170	2.82	1280
13	1	13.75	1.73	2.41	16.0	89	2.60	2.76	0.29	1.81	5.600000	1.150	2.90	1320
14	1	14.75	1.73	2.39	11.4	91	3.10	3.69	0.43	2.81	5.400000	1.250	2.73	1150
15	1	14.38	1.87	2.38	12.0	102	3.30	3.64	0.29	2.96	7.500000	1.200	3.00	1547
16	1	13.63	1.81	2.70	17.2	112	2.85	2.91	0.30	1.46	7.300000	1.280	2.88	1310
17	1	14.30	1.92	2.72	20.0	120	2.80	3.14	0.33	1.97	6.200000	1.070	2.65	1280
18	1	13.83	1.57	2.62	20.0	115	2.95	3.40	0.40	1.72	6.600000	1.130	2.57	1130
19	1	14.19	1.59	2.48	16.5	108	3.30	3.93	0.32	1.86	8.700000	1.230	2.82	1680
20	1	13.64	3.10	2.56	15.2	116	2.70	3.03	0.17	1.66	5.100000	0.960	3.36	845

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

2.

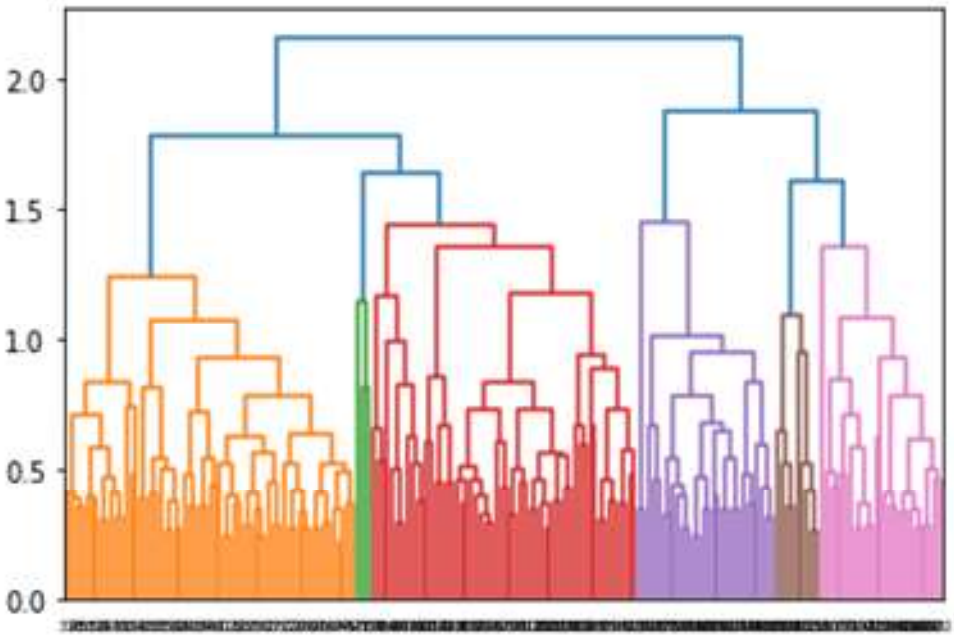
Name of the feature	Description	Type of data	relevance
Type	Class category of the wine	Quantitative	relevant
Alcohol	Amount of Alcohol in that particular wine type	quantitative	relevant
Malic	Amount of Malic Acid in that particular wine type	quantitative	relevant
Ash	Amount of Ash in that particular 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

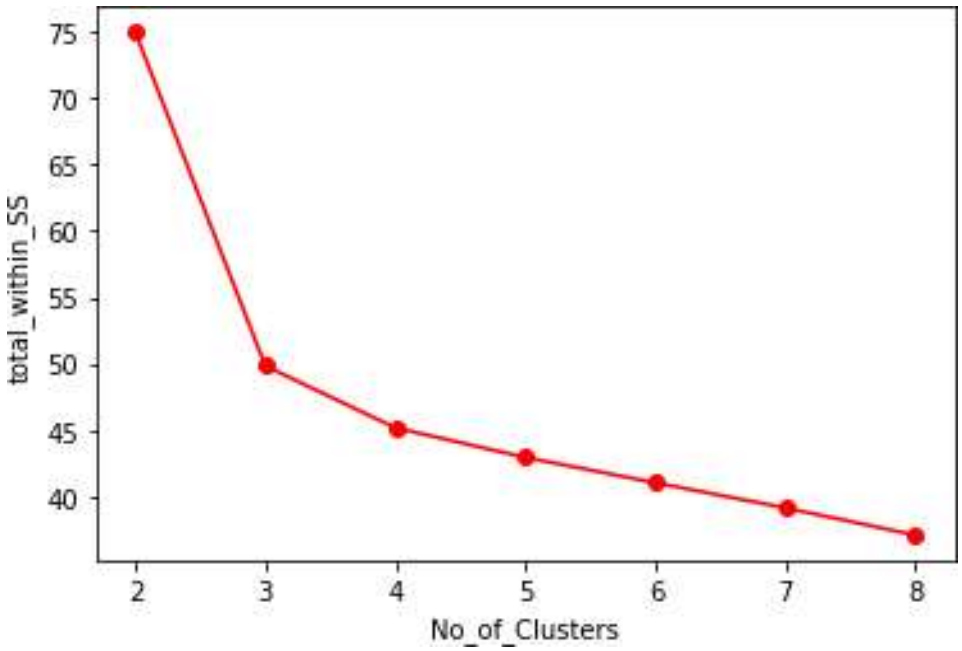
3. DATAPREPROCESSING : checked the type of data .All the data is of numeric type so no need to do the typecasting. There are no duplicates or null values as well.

4. 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.
5. MODEL BUILDING:

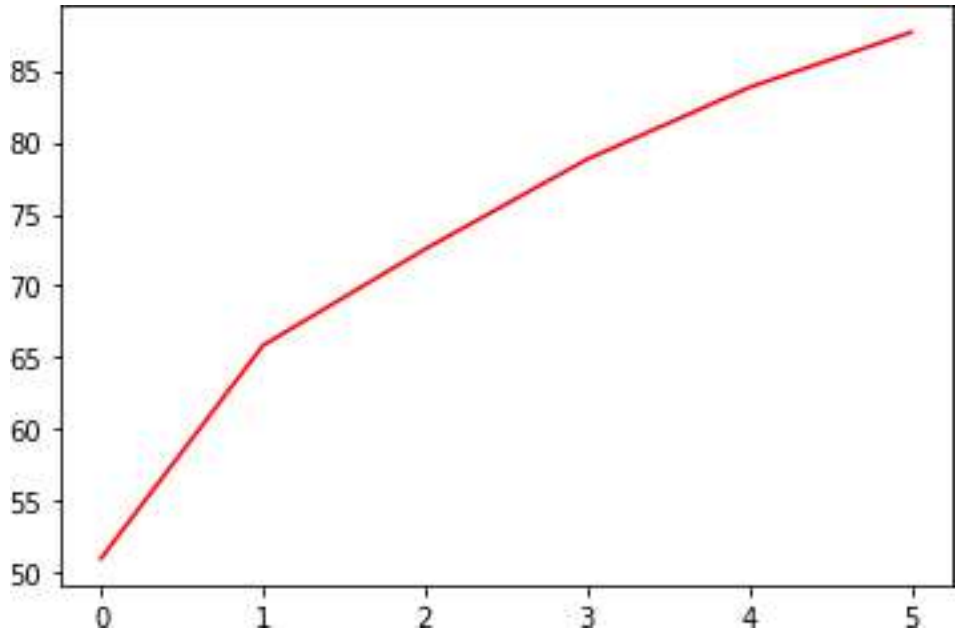
Hierarchical clustering :-



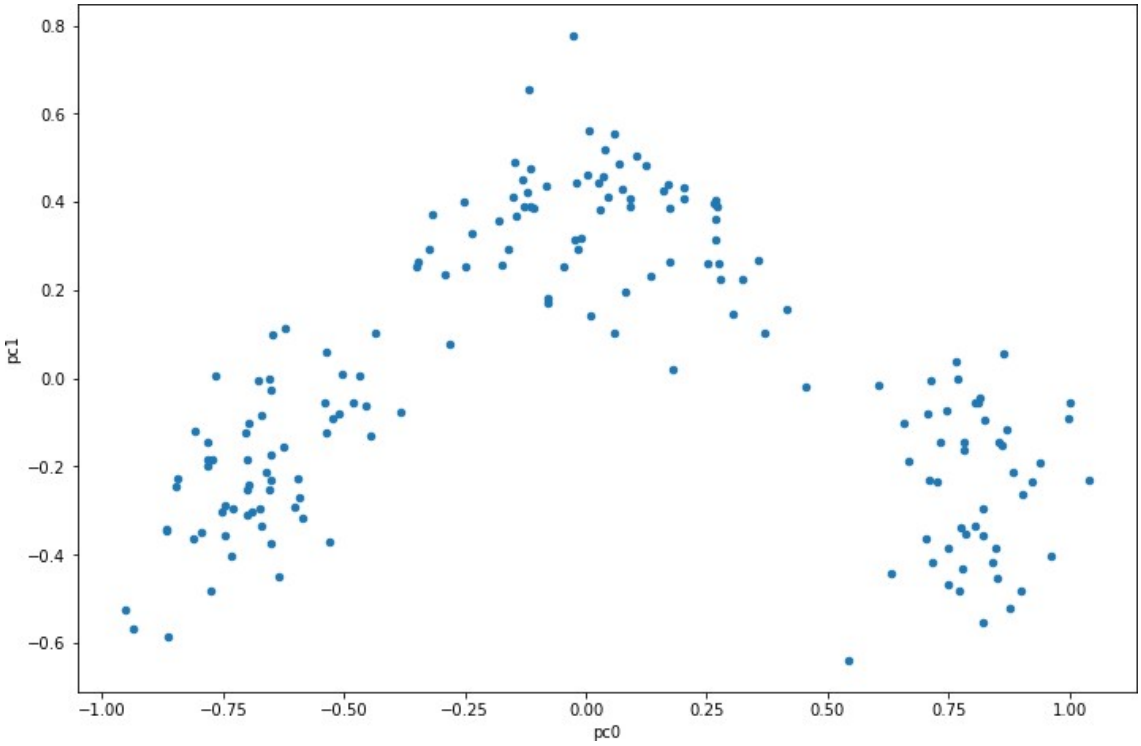
K-means clustering:



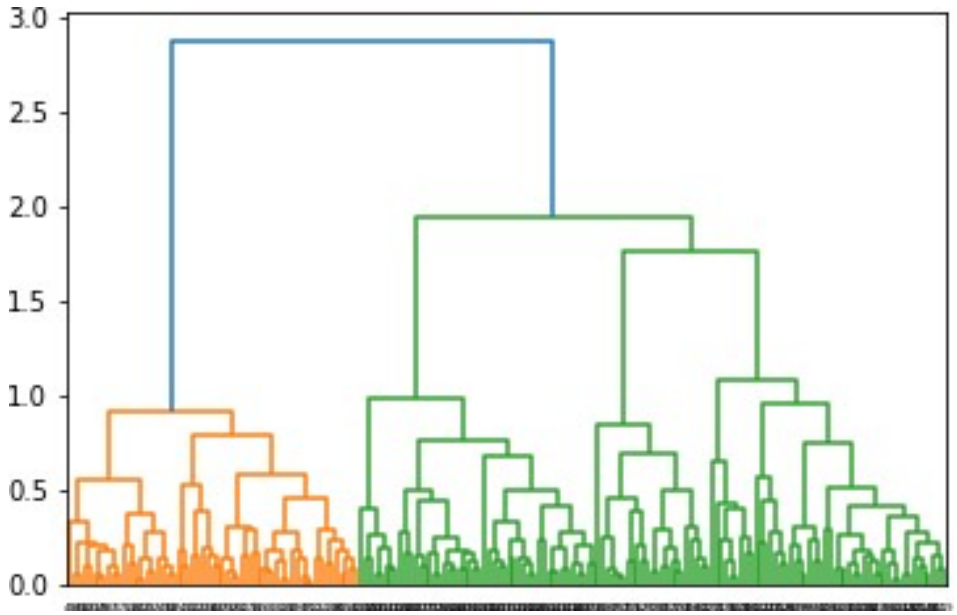
PCA:



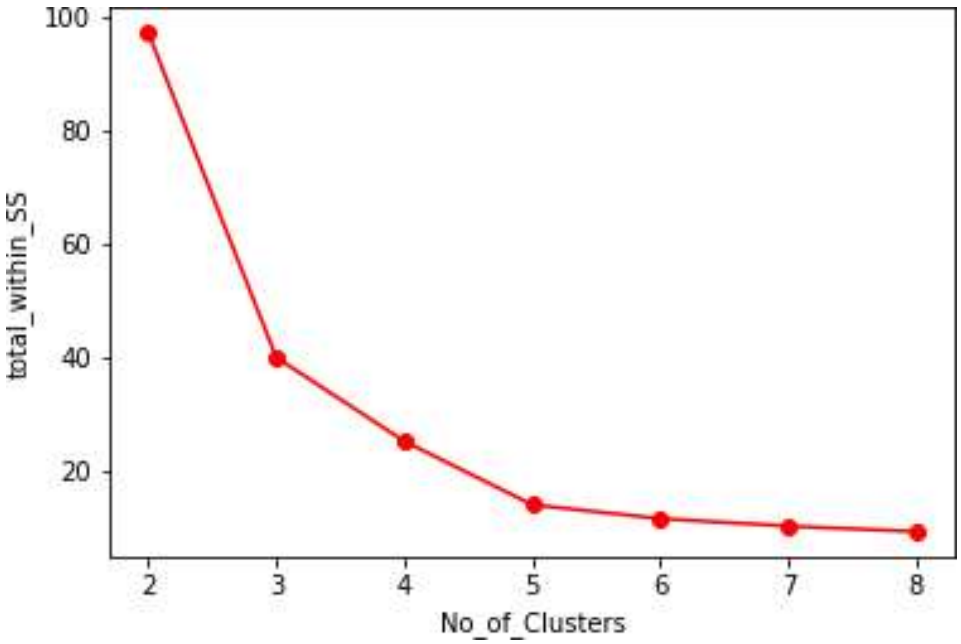
Scatter plot of PCA data:



Dendrogram of Hierarchical clustering on PCA data:



k-means elbow curve on PCA data:



6. 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.

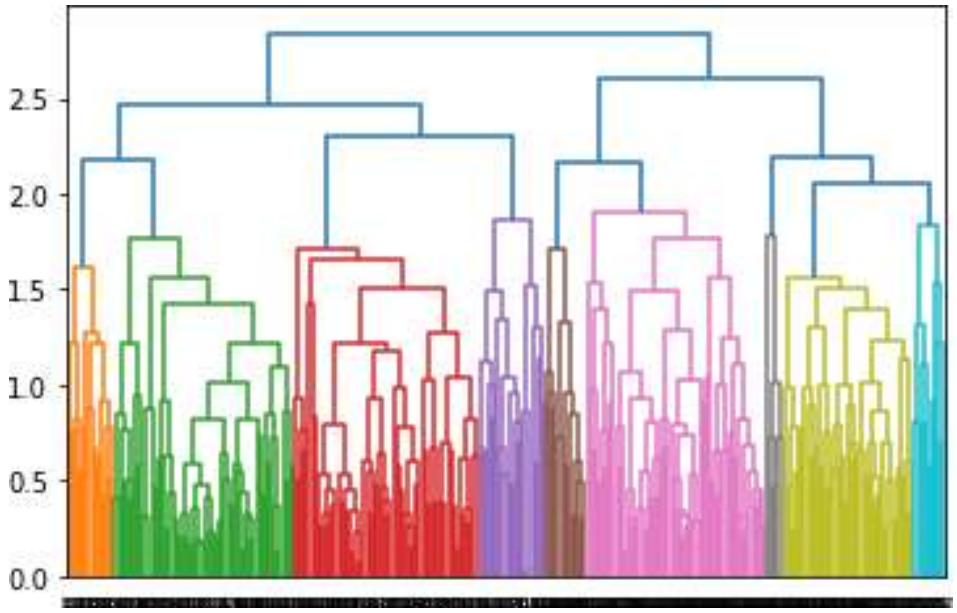
age	sex	cp	trestbps	chol	fbs
63	1	3	145	233	1
37	1	2	130	250	0
41	0	1	130	204	0
56	1	1	120	236	0
57	0	0	120	354	0
57	1	0	140	192	0
56	0	1	140	294	0
44	1	1	120	263	0
52	1	2	172	199	1
57	1	2	150	168	0
54	1	0	140	239	0
48	0	2	130	275	0
49	1	1	130	266	0
64	1	3	110	211	0
58	0	3	150	283	1
50	0	2	120	219	0
58	0	2	120	340	0
66	0	3	150	226	0
43	1	0	150	247	0
69	0	3	140	239	0
59	1	0	135	234	0
44	1	2	130	233	0
42	1	0	140	226	0
61	1	2	150	243	1

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.

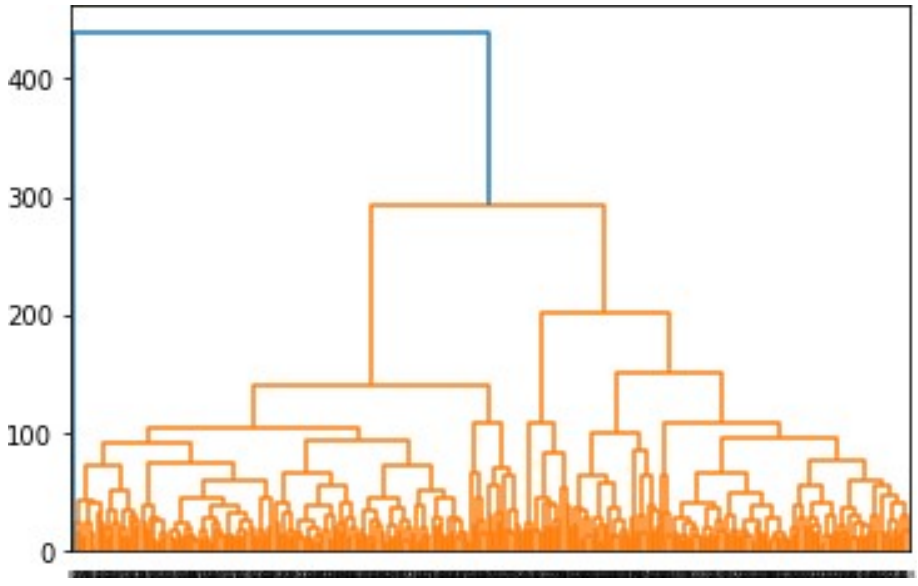
Name of the feature	description	Type of data	relevance
Age	person's age in years	Quantitative	relevant

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

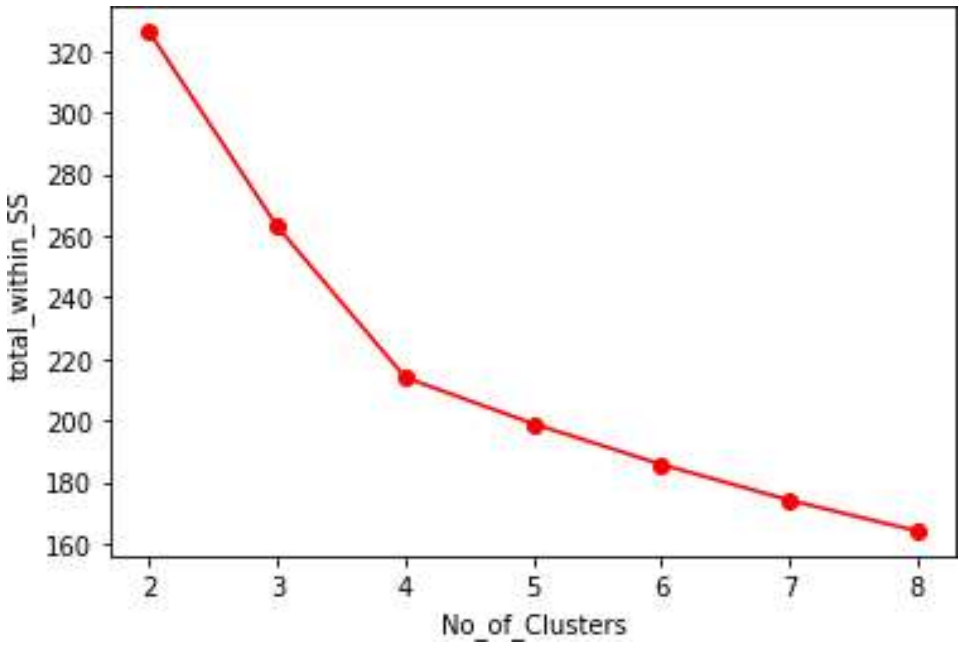
Data preprocessing:-checked for the data types and found that one column has float value which is changed to int data type. checked for the duplicate value ,found 1 and it is removed. also checked for the null values. There is no presence of null values.
EDA:-exploratory data analysis is done on the data and also the data visualization.it is found that the data is right skewed or positive skew.
Model building:-
Dendrogram of hierarchical clustering



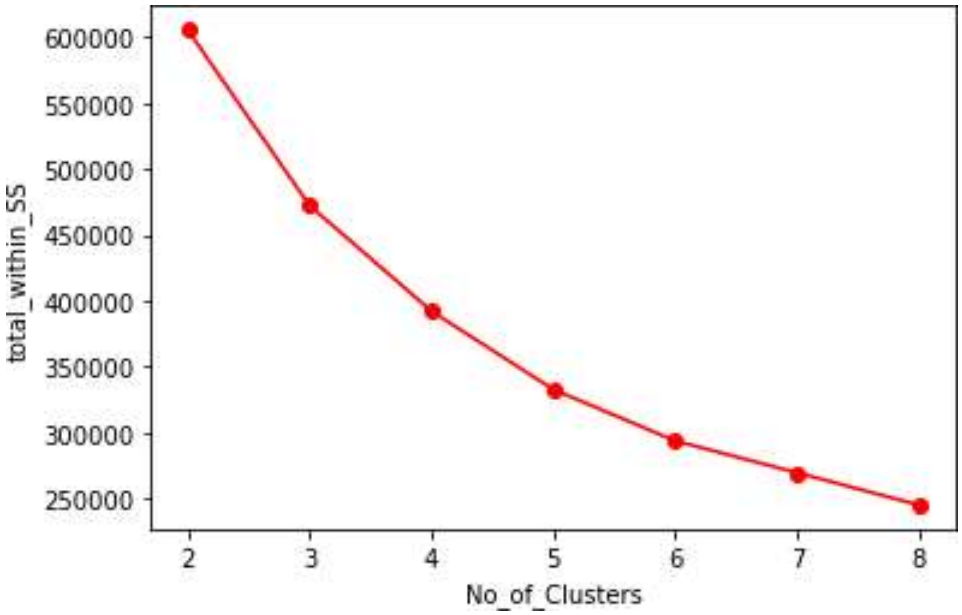
Dendrogram of hierarchical clustering after the PCA



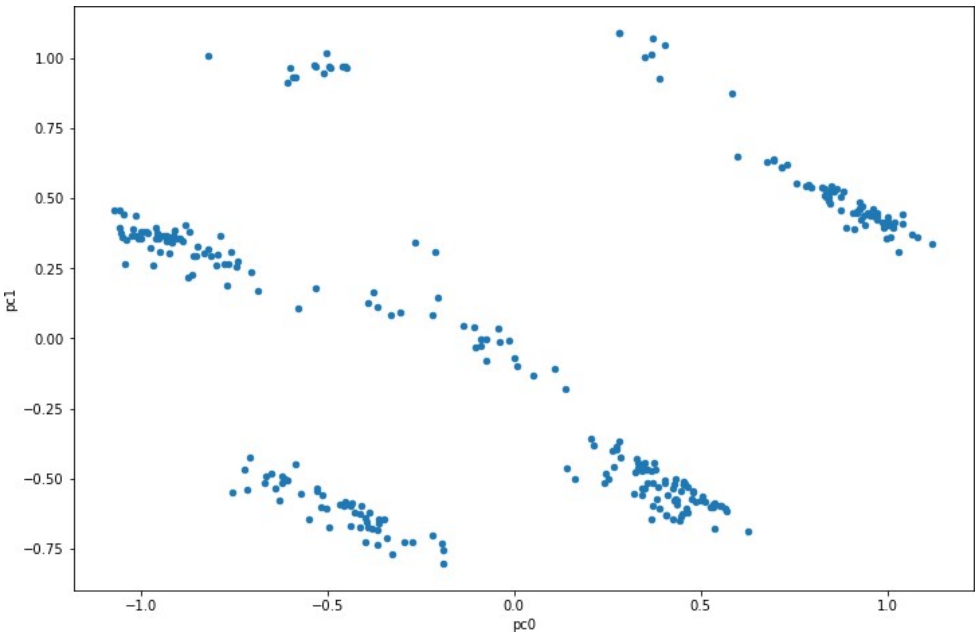
Elbow curve of Kmeans clustering:



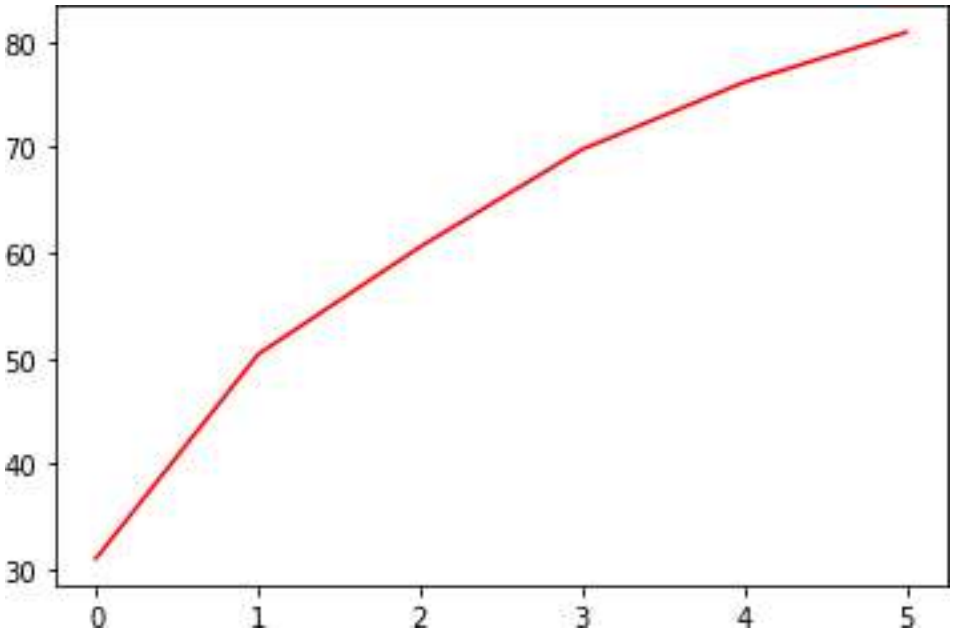
Elbow curve of Kmeans after PCA :



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.