DATA MINING PROJECT REPORT

SHUBHAM NAGARGOJE

DSBA July,22

Table Of Content:

Content:

Part 1: PCA:

Part 1: PCA: Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented. (5marks)	5.0 pts
Part 1: PCA: Scale the variables and write the inference for using the type of scaling function for this case study. (3 marks)	3.0 pts
Part 1: PCA: Comment on the comparison between covariance and the correlation matrix after scaling. (2 marks)	2.0 pts
Part 1: PCA: Check the dataset for outliers before and after scaling. Draw your inferences from this exercise. (3 marks)	3.0 pts
Part 1: PCA: Build the covariance matrix, eigenvalues and eigenvector. (4 marks)	4.0 pts
Part 1: PCA: Write the explicit form of the first PC (in terms of Eigen Vectors). (5 marks)	5.0 pts
Part 1: PCA: Discuss the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate? Perform PCA and export the data of the Principal Component scores into a data frame. (8 marks)	8.0 pts
Part 1: PCA: Mention the business implication of using the Principal Component Analysis for this case study. (5 marks)	5.0 pts

Part 2: Clustering:

2.1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, etc, etc)

2.2. Do you think scaling is necessary for clustering in this case? Justify

2.3. Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.

2.4. Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and find the silhouette score.

2.5. Describe cluster profiles for the clusters defined. Recommend different priority based actions that need to be taken for different clusters on the bases of their vulnerability situations according to their Economic and Health Conditions.

Part 1: PCA:

Problem Statement: The 'Hair Salon.csv' dataset contains various variables used for the context of Market Segmentation. This particular case study is based on various parameters of a salon chain of hair products. You are expected to do Principal Component Analysis for this case study according to the instructions given in the rubric. **Kindly refer to the PCA_Data_Dictionary.jpg file for the Data Dictionary of the Dataset.**
Note: This particular dataset contains the target variable satisfaction as well. Please do drop this variable before doing Principal Component Analysis.

Variable	Expansion
ProdQual	Product Quality
Ecom	E-Commerce
TechSup	Technical Support
CompRes	Complaint Resoluti
Advertising	Advertising
ProdLine	Product Line
SalesFImage	Salesforce Image
ComPricing	Competitive Pricing
WartyClaim	Warranty & Claims
OrdBilling	Order & Billing
DelSpeed	Delivery Speed
Satisfaction	Customer Satisfact

Figure Number 1

Part 1: PCA: Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented. (5marks)

	ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
0	1	8.5	3.9	2.5	5.9	4.8	4.9	6.0	6.8	4.7	5.0	3.7	8.2
1	2	8.2	2.7	5.1	7.2	3.4	7.9	3.1	5.3	5.5	3.9	4.9	5.7
2	3	9.2	3.4	5.6	5.6	5.4	7.4	5.8	4.5	6.2	5.4	4.5	8.9
3	4	6.4	3.3	7.0	3.7	4.7	4.7	4.5	8.8	7.0	4.3	3.0	4.8
4	5	9.0	3.4	5.2	4.6	2.2	6.0	4.5	6.8	6.1	4.5	3.5	7.1

Figure Number 2

	ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
95	96	8.6	4.8	5.6	5.3	2.3	6.0	5.7	6.7	5.8	4.9	3.6	7.3
96	97	7.4	3.4	2.6	5.0	4.1	4.4	4.8	7.2	4.5	4.2	3.7	6.3
97	98	8.7	3.2	3.3	3.2	3.1	6.1	2.9	5.6	5.0	3.1	2.5	5.4
98	99	7.8	4.9	5.8	5.3	5.2	5.3	7.1	7.9	6.0	4.3	3.9	6.4
99	100	7.9	3.0	4.4	5.1	5.9	4.2	4.8	9.7	5.7	3.4	3.5	6.4

Figure Number 3

Describing the data, we know that

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   ID            100 non-null    int64
 1   ProdQual      100 non-null    float64
 2   Ecom          100 non-null    float64
 3   TechSup       100 non-null    float64
 4   CompRes       100 non-null    float64
 5   Advertising   100 non-null    float64
 6   ProdLine      100 non-null    float64
 7   SalesFImage   100 non-null    float64
 8   ComPricing    100 non-null    float64
 9   WartyClaim    100 non-null    float64
 10  OrdBilling    100 non-null    float64
 11  DelSpeed      100 non-null    float64
 12  Satisfaction  100 non-null    float64
dtypes: float64(12), int64(1)
memory usage: 10.3 KB
```

Figure Number 4

- **There are 0 non-null values.**
- **All the variables are in float data type.**
- **There are total 13 features and 100 rows in the given dataset.**
- **There are no duplicates found.**

Let's have a look at the univariate, bivariate and multivariate Analysis.

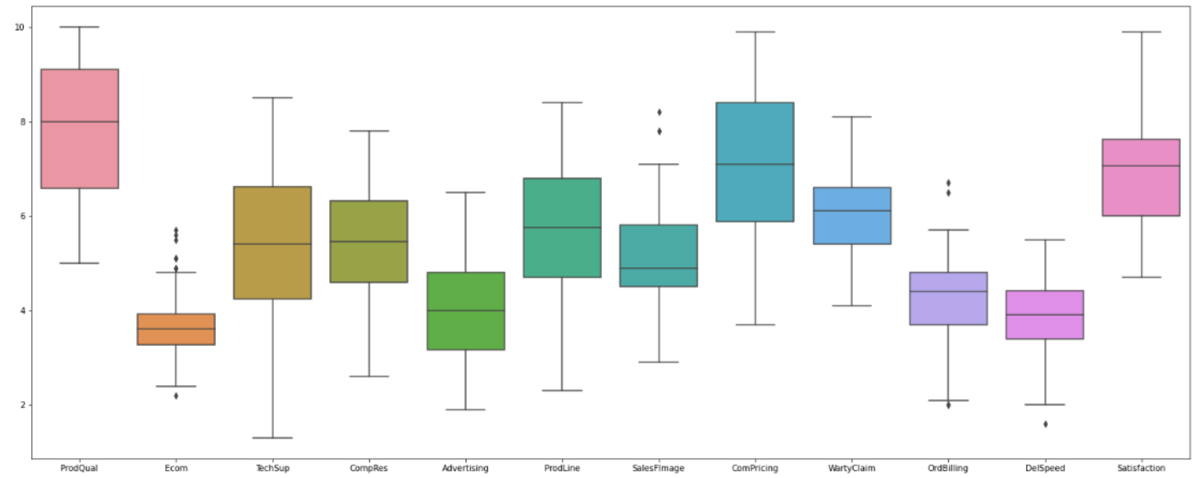


Figure Number 5

There are some outliers present in the Ecom, SalesFimange, OrdBilling and DelSpeed columns.

Now, let us do Bivariate as well as Multivariate analysis.

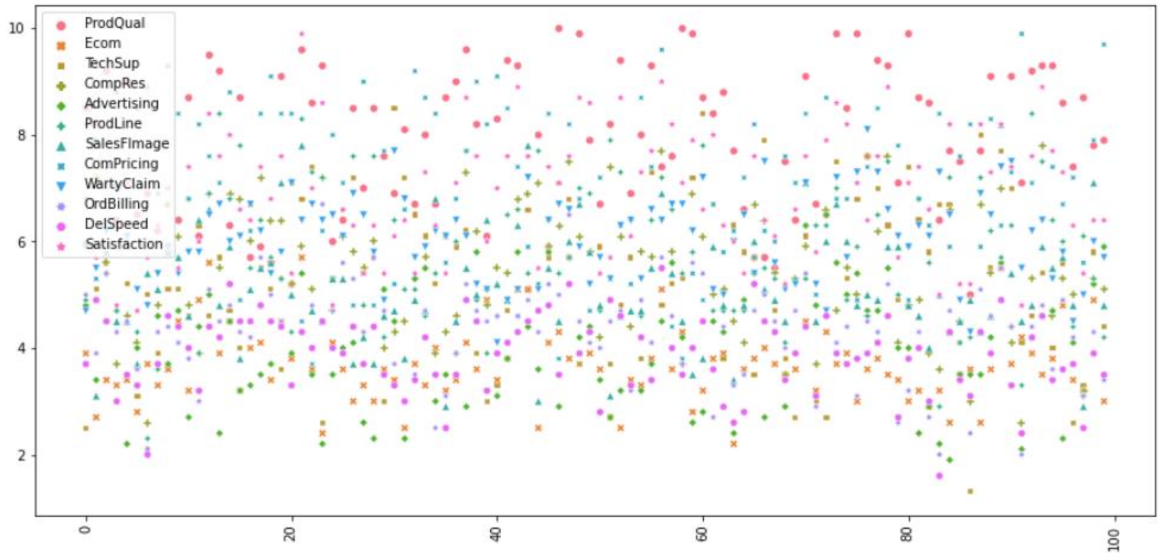


Figure Number6

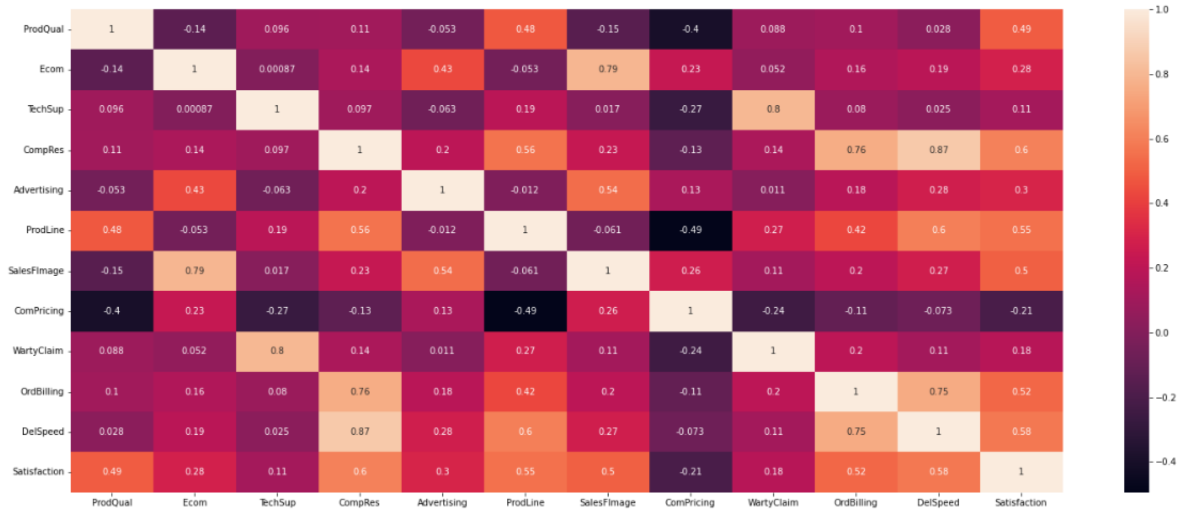


Figure Number 7

*Ecom & salesFimage , TechSup & Wartclaim , CompRes & Odbilling , CompRes & DelSpeed are most corelated .

* All have positive corelations with each other.

-This is bivariate analysis of variables 'Ecom' and 'SalesFimage'.

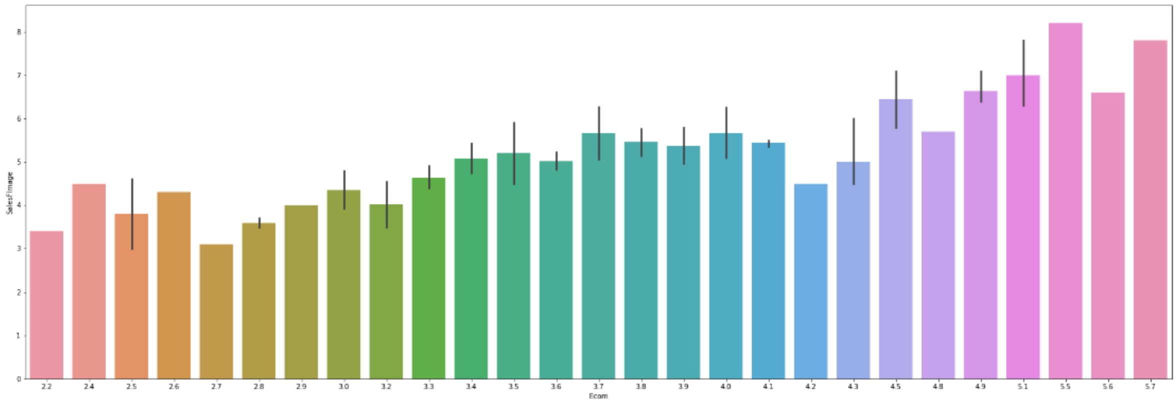
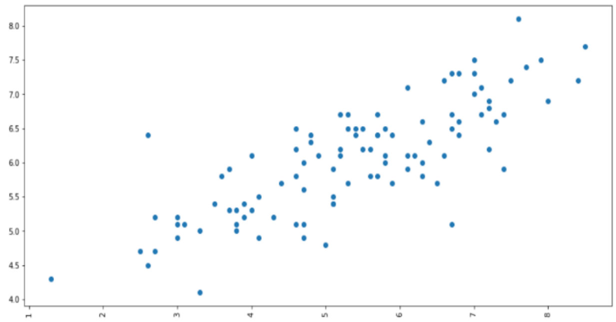
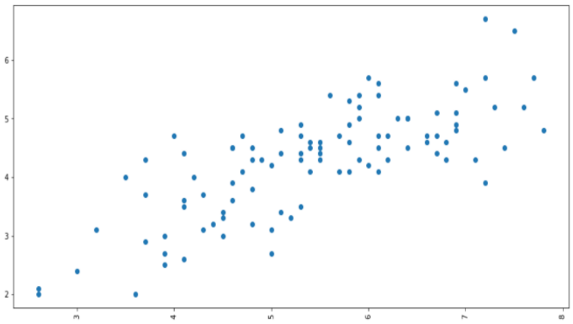


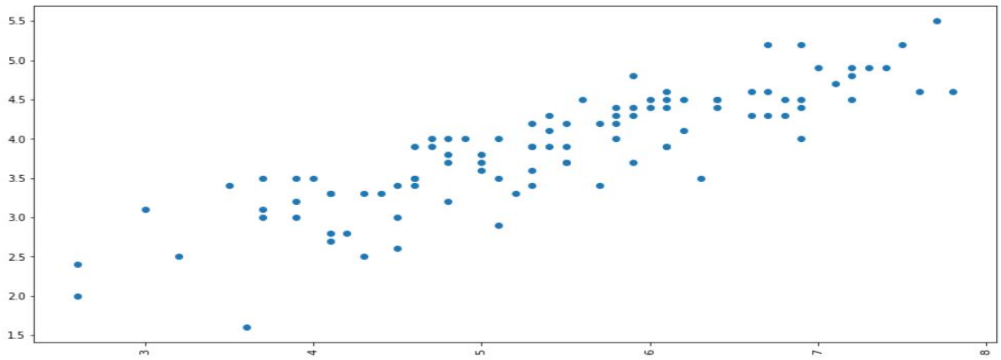
Figure No. 8



(between 'TechSup' and 'WartyClaim')



(between 'CompRes' and 'OrdBilling')



(between 'CompRes' and 'DelSpeed')

Figure Number9

Part 1: PCA: Scale the variables and write the inference for using the type of scaling function for this case study. (3 marks)

After applying Z score, the data gets scaled.

* Z-score is a variation of scaling that represents the number of standard deviations away from the mean.

* I would use z-score to ensure your feature distributions have mean = 0 and std = 1. It's useful when there are a few outliers, but not so extreme that you need clipping.

* As This dataset has minimal outliers ,we used z-score.

Part 1: PCA: Comment on the comparison between covariance and the correlation matrix after scaling. (2 marks)

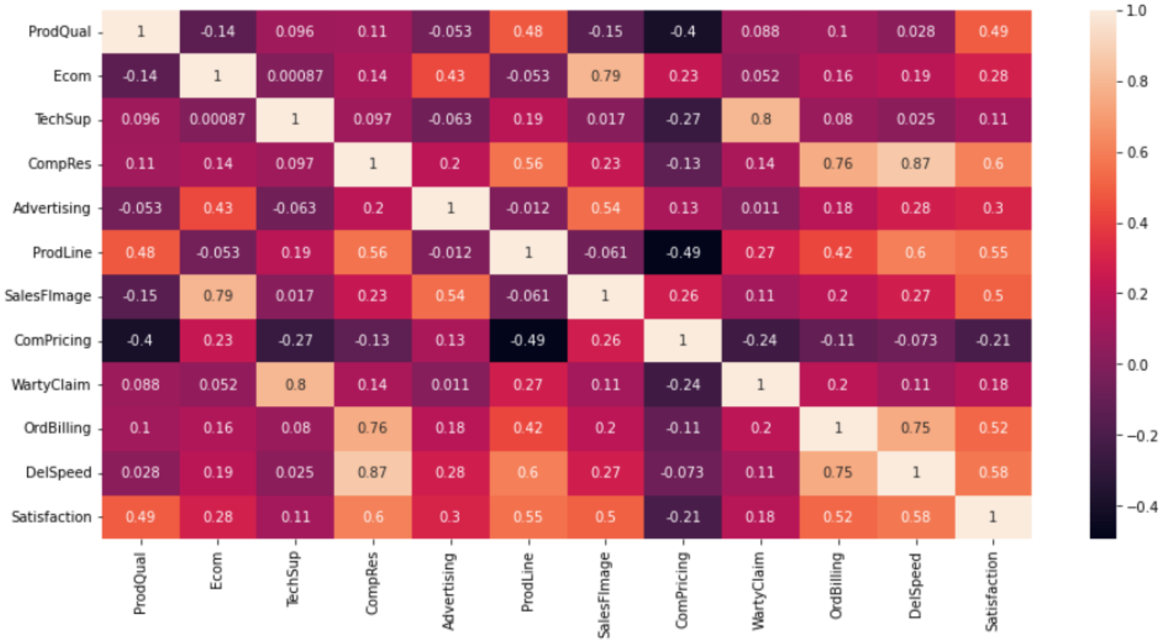


Figure Number 10

As we can see, There is no major difference in correlations before and after scaling

After scaling, lets check the correlation for covariance matrix.

- Covarience Matrix Heat Maps

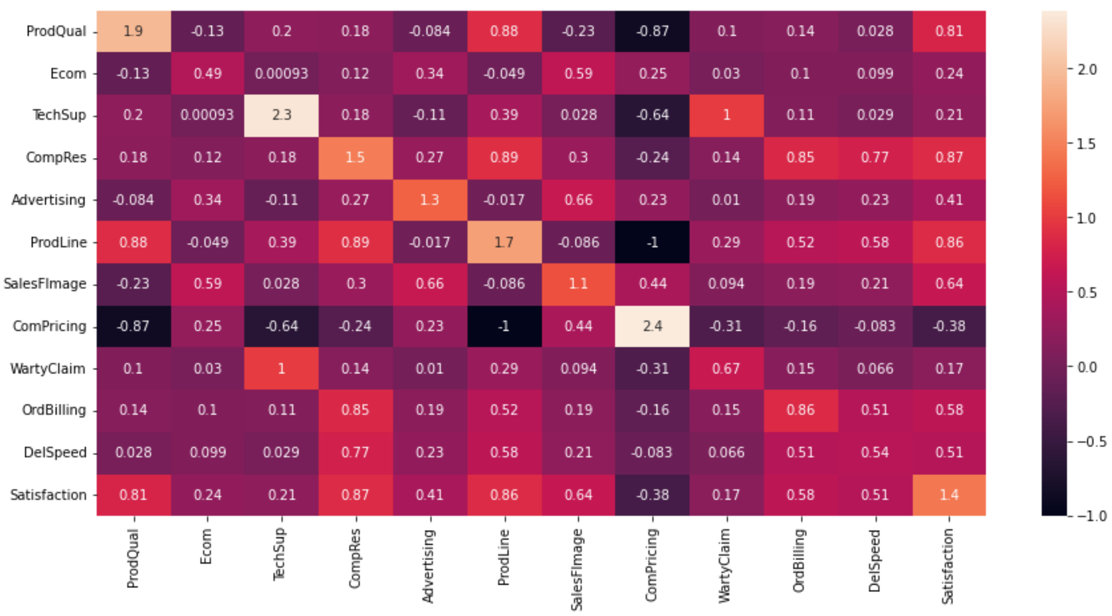


Figure Number 11

Now, let check for scaled covariance matrix.

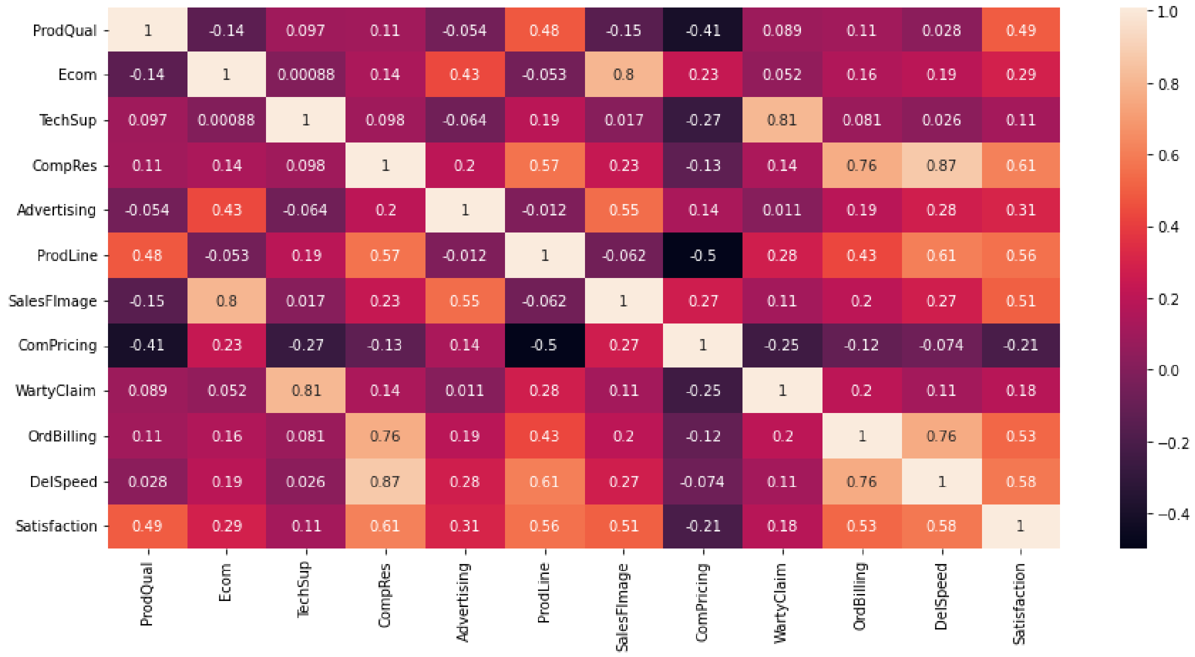


Figure Number 12

- Orignal Data covariance matrix

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
ProdQual	1.949596	-0.134162	0.204293	0.179475	-0.084141	0.876919	-0.227303	-0.865697	0.101081	0.135273	0.028424	0.809313
Ecom	-0.134162	0.490723	0.000929	0.118663	0.339374	-0.048545	0.594590	0.248356	0.029802	0.101600	0.098594	0.236065
TechSup	0.204293	0.000929	2.342298	0.178758	-0.108434	0.387753	0.027884	-0.640313	1.000106	0.113869	0.028596	0.205384
CompRes	0.179475	0.118663	0.178758	1.460238	0.268162	0.892313	0.297711	-0.238897	0.139085	0.849519	0.767766	0.868832
Advertising	-0.084141	0.339374	-0.108434	0.268162	1.270000	-0.017121	0.655222	0.233697	0.009970	0.192848	0.228323	0.409212
ProdLine	0.876919	-0.048545	0.387753	0.892313	-0.017121	1.729975	-0.086480	-1.005828	0.294429	0.518495	0.581384	0.863040
SalesFImage	-0.227303	0.594590	0.027884	0.297711	0.655222	-0.086480	1.149870	0.438382	0.094456	0.194349	0.213861	0.639279
ComPricing	-0.865697	0.248356	-0.640313	-0.238897	0.233697	-1.005828	0.438382	2.387196	-0.310285	-0.164416	-0.082691	-0.383568
WartyClaim	0.101081	0.029802	1.000106	0.139085	0.009970	0.294429	0.094456	-0.310285	0.671971	0.150046	0.065861	0.173461
OrdBilling	0.135273	0.101600	0.113869	0.849519	0.192848	0.518495	0.194349	-0.164416	0.150046	0.862743	0.512315	0.577572
DelSpeed	0.028424	0.098594	0.028596	0.767766	0.228323	0.581384	0.213861	-0.082691	0.065861	0.512315	0.539398	0.505103
Satisfaction	0.809313	0.236065	0.205384	0.868832	0.409212	0.863040	0.639279	-0.383568	0.173461	0.577572	0.505103	1.420481

Table 01

- Scaled data covariance matrix

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
ProdQual	1.949596	-0.134162	0.204293	0.179475	-0.084141	0.876919	-0.227303	-0.865697	0.101081	0.135273	0.028424	0.809313
Ecom	-0.134162	0.490723	0.000929	0.118663	0.339374	-0.048545	0.594590	0.248356	0.029802	0.101600	0.098594	0.236065
TechSup	0.204293	0.000929	2.342298	0.178758	-0.108434	0.387753	0.027884	-0.640313	1.000106	0.113869	0.028596	0.205384
CompRes	0.179475	0.118663	0.178758	1.460238	0.268162	0.892313	0.297711	-0.238897	0.139085	0.849519	0.767766	0.868832
Advertising	-0.084141	0.339374	-0.108434	0.268162	1.270000	-0.017121	0.655222	0.233697	0.009970	0.192848	0.228323	0.409212
ProdLine	0.876919	-0.048545	0.387753	0.892313	-0.017121	1.729975	-0.086480	-1.005828	0.294429	0.518495	0.581384	0.863040
SalesFImage	-0.227303	0.594590	0.027884	0.297711	0.655222	-0.086480	1.149870	0.438382	0.094456	0.194349	0.213861	0.639279
ComPricing	-0.865697	0.248356	-0.640313	-0.238897	0.233697	-1.005828	0.438382	2.387196	-0.310285	-0.164416	-0.082691	-0.383568
WartyClaim	0.101081	0.029802	1.000106	0.139085	0.009970	0.294429	0.094456	-0.310285	0.671971	0.150046	0.065861	0.173461
OrdBilling	0.135273	0.101600	0.113869	0.849519	0.192848	0.518495	0.194349	-0.164416	0.150046	0.862743	0.512315	0.577572
DelSpeed	0.028424	0.098594	0.028596	0.767766	0.228323	0.581384	0.213861	-0.082691	0.065861	0.512315	0.539398	0.505103
Satisfaction	0.809313	0.236065	0.205384	0.868832	0.409212	0.863040	0.639279	-0.383568	0.173461	0.577572	0.505103	1.420481

Table 02

Part 1: PCA: Check the dataset for outliers before and after scaling. Draw your inferences from this exercise. (3 marks)

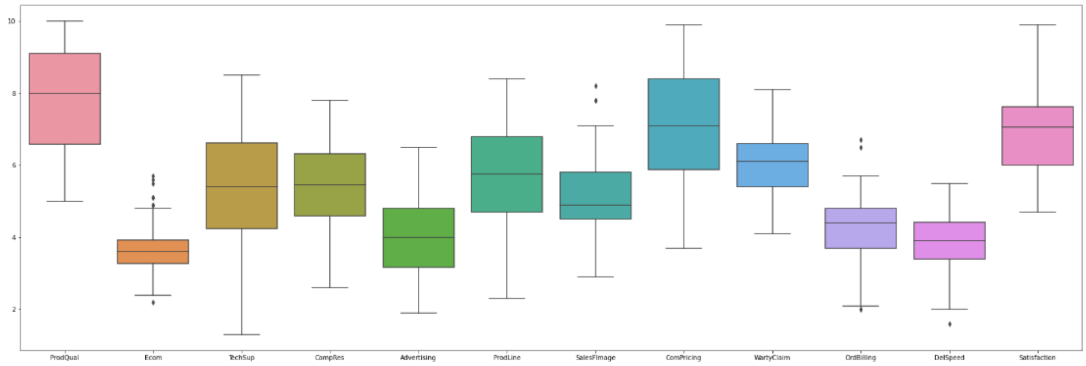


Figure Number 13

(Before Scaling)

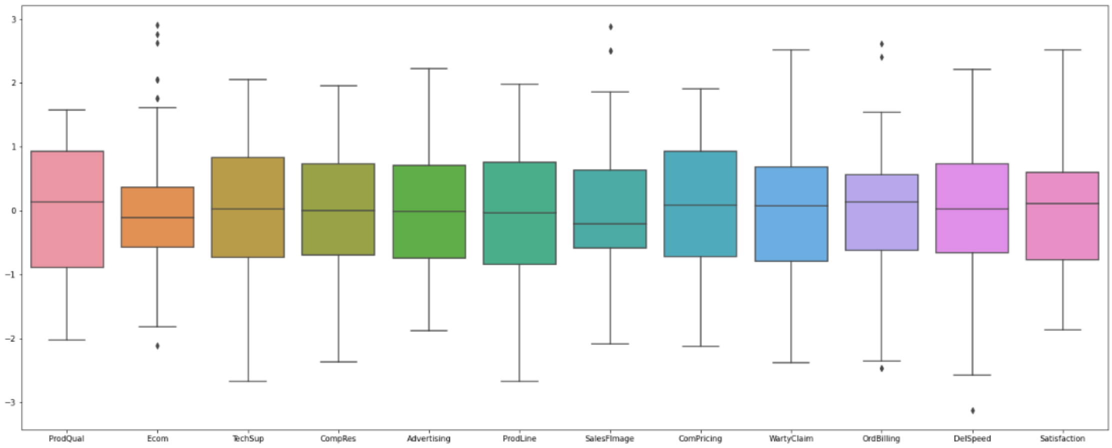


Figure Number 14(After Scaling)

Even after scaling, the outliers has not been treated.

Part 1: PCA: Build the covariance matrix, eigenvalues and eigenvector. (4 marks)

By applying the codes, we performed,

1.PCA taking all features.

2.created covariance matrix.

3. Extracted eigen Values and eigen vectors.

Now, lets have a look at the scree plot to identify the number of components to be built.

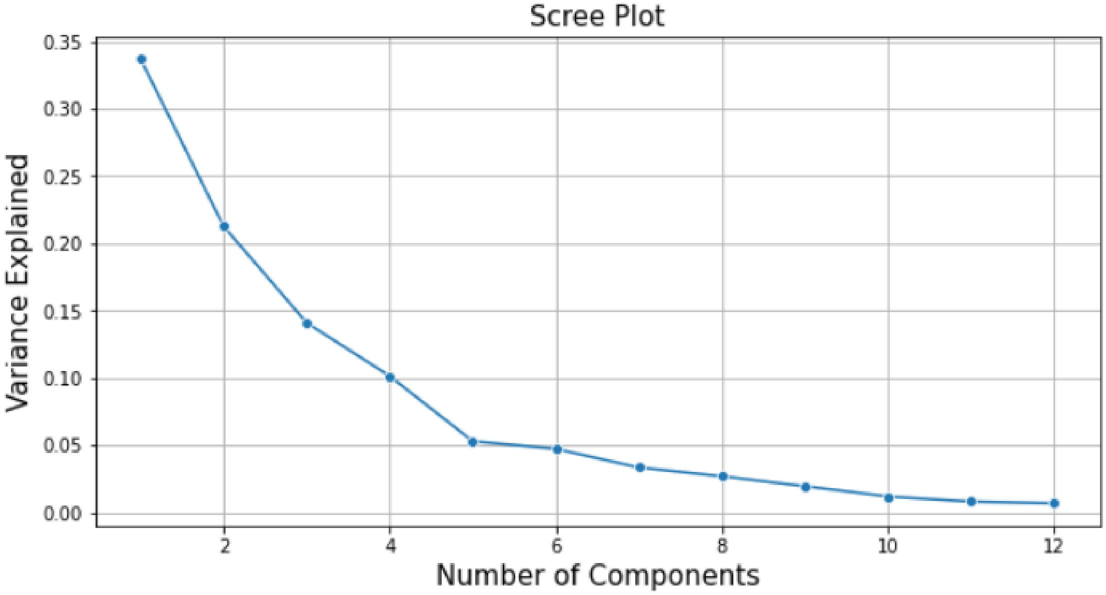


Figure Number 15

Part 1: PCA: Write the explicit form of the first PC (in terms of Eigen Vectors). (5 marks)

The explicit form of first PC is as below:

PC1= 0.15*ProdQual - 0.31*Ecom - 0.07*TechSup - 0.61*CompRes - 0.24*Advertising + 0.36*F.ProdLine - 0.12*P.SalesFImage - 0.32*ComPricing + 0.18*WartyClaim - 0.2*OrdBilling - 0.21*DelSpeed + 0.22*Satisfaction

Part 1: PCA: Discuss the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate? Perform PCA and export the data of the Principal Component scores into a data frame. (8 marks)

We first check the cumulative variance described by each component.

Then we checked the cumulative explained variance ratio to find a cut off for selecting the number of Principal Components.

Now, lets take a look at the scree plot to identify the number of components to be built.

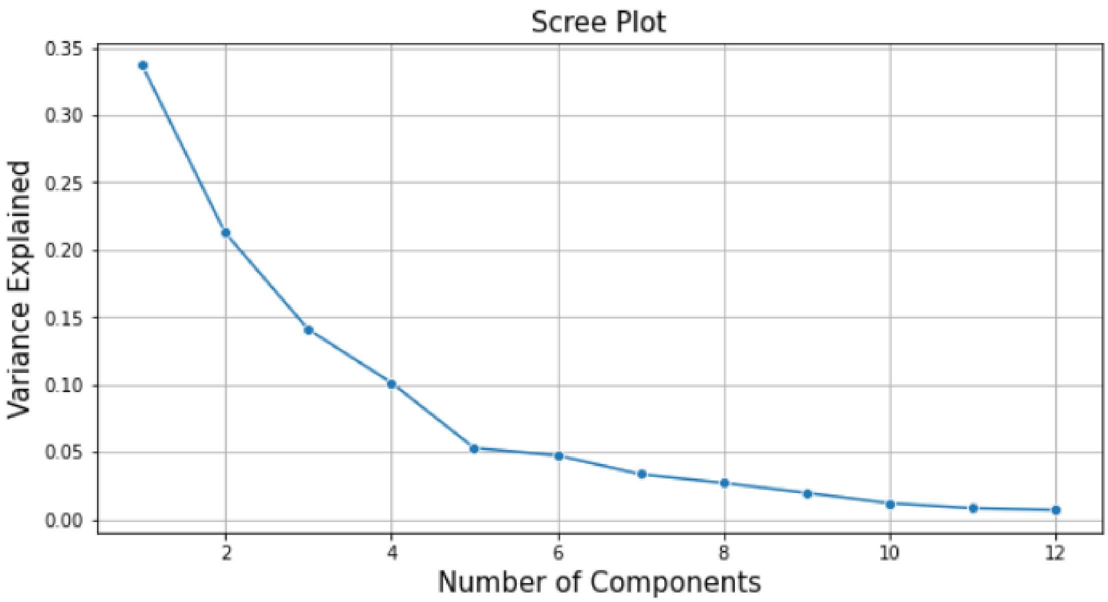


Figure Number 16

- Total no. of optimum variables is 8 as it explains the 95% variance.
- Eigen vectors indicates direction.

We generated only 8 PCA dimensions (dimensionality reduction from 12 to 8)

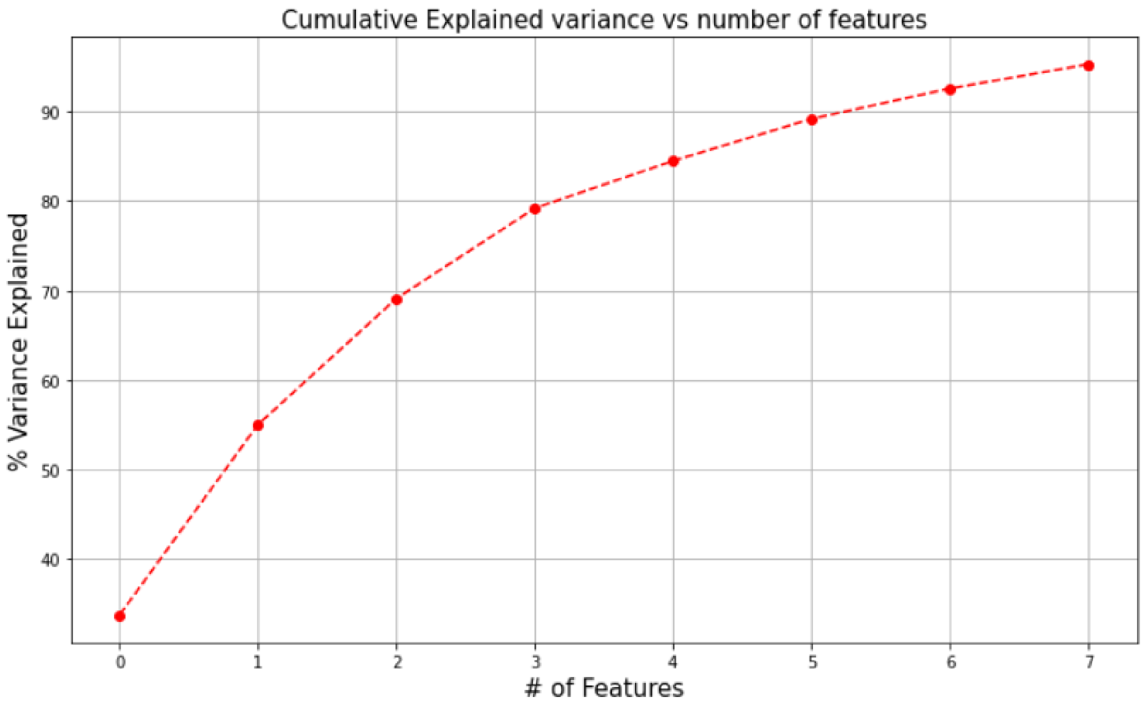


Figure Number 17

Lets see the corelation after data reduction,

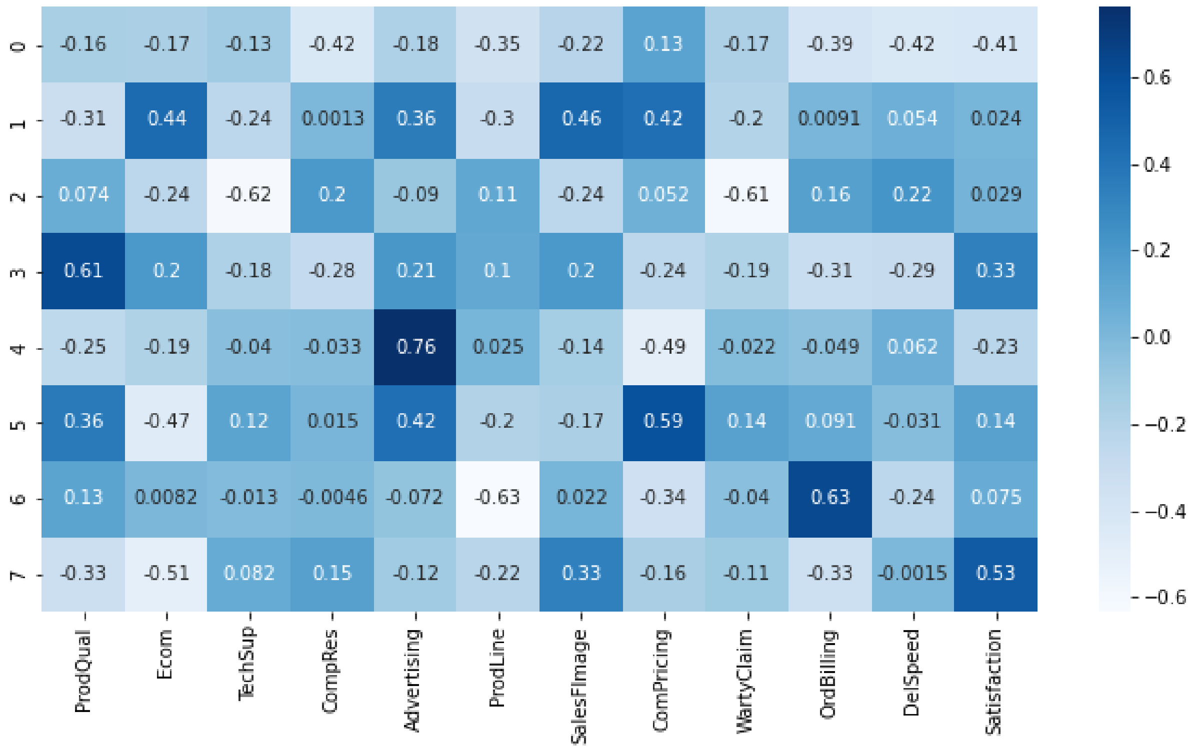


Figure Number 18

Part 1: PCA: Mention the business implication of using the Principal Component Analysis for this case study. (5 marks)

Principal component analysis, or PCA, is a statistical procedure that allows you to summarize the information content in large data tables by means of a smaller set of “summary indices” that can be more easily visualized and analysed.

Following are the advantages of using PCA.

PCA can help us improve performance at a meagre cost of model accuracy.

Other benefits of PCA include reduction of noise in the data,

feature selection (to a certain extent),

and the ability to produce independent,

uncorrelated features of the data.

PCA is used to visualize multidimensional data.

It is used to reduce the number of dimensions in healthcare data.

-**For this case study :-**

We have successfully sorted the most important features that affect the business sales and revenue.

By which we can strategize the sales and factors which will increase the sales.

We also get to know what are thwe affects and the amount of variance explained by the features ,

With the

Part 2: Clustering:

[The State_wise_Health_income.csv] dataset given is about the Health and economic conditions in different States of a country. The Group States based on how similar their situation is, so as to provide these groups to the government so that appropriate measures can be taken to escalate their Health and Economic conditions.
2.1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, etc, etc)
2.2. Do you think scaling is necessary for clustering in this case? Justify
2.3. Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.
2.4. Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and find the silhouette score.
2.5. Describe cluster profiles for the clusters defined. Recommend different priority based actions that need to be taken for different clusters on the bases of their vulnerability situations according to their Economic and Health Conditions.
Data Dictionary for State_wise_Health_income Dataset:
1. States: names of States
2. Health_indeces1: A composite index rolls several related measures (indicators) into a single score that provides a summary of how the health system is performing in the State.
3. Health_indeces2: A composite index rolls several related measures (indicators) into a single score that provides a summary of how the health system is performing in certain areas of the States.
4. Per_capita_income-Per capita income (PCI) measures the average income earned per person in a given area (city, region, country, etc.) in a specified year. It is calculated by dividing the area's total income by its total population.
5. GDP: GDP provides an economic snapshot of a country/state, used to estimate the size of an economy and growth rate.

Dataset for Part 1: PCA: Hair Salon.csv

Dataset for Part 2: Clustering: State_wise_Health_income.csv

Part 2: Clustering: Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, etc)

	Unnamed: 0	States	Health_indeces1	Health_indices2	Per_capita_income	GDP
0	0	Bachevo	417	66	564	1823
1	1	Balgarchevo	1485	646	2710	73662
2	2	Belasitsa	654	299	1104	27318
3	3	Belo_Pole	192	25	573	250
4	4	Beslen	43	8	528	22
...	...	...	...	...	...	...
292	292	Greencastle	3443	970	2499	238636
293	293	Greenisland	2963	793	1257	162831
294	294	Greyabbey	3276	609	1522	120184
295	295	Greysteel	3463	847	934	199403
296	296	Groggan	2070	838	3179	166767

297 rows × 6 columns

Figure Number 19

Describing the data we know that,

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 297 entries, 0 to 296
Data columns (total 4 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Health_indeces1    297 non-null    int64
 1   Health_indices2    297 non-null    int64
 2   Per_capita_income  297 non-null    int64
 3   GDP                297 non-null    int64
dtypes: int64(4)
memory usage: 9.4 KB
None
```

Figure Number 20

- There are 297 rows and 4 columns present in the dataset.
- All the variables are in integer datatype.
- There are no null values present in the dataset.
- There is only one duplicate row.

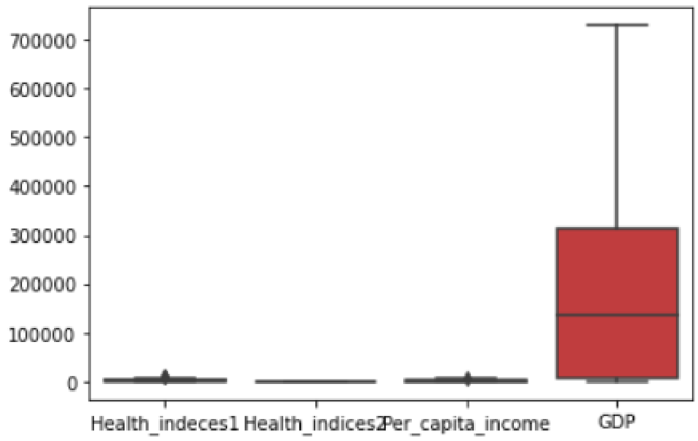


Figure Number 21

Part 2: Clustering: Do you think scaling is necessary for clustering in this case? Justify.

Yes, scaling is required in this data set as all features have different weights and to ensure that none of the feature is identified as important only because of the weight, scaling is mandatory for this data set.

After Scaling,

	count	mean	std	min	25%	50%	75%	max
Health_indeces1	297.0	-6.803387e-17	1.001688	-1.297327	-0.977436	-0.088032	0.719311	3.729034
Health_indices2	297.0	1.252272e-17	1.001688	-1.481634	-1.107825	0.248566	0.810346	1.739527
Per_capita_income	297.0	-1.566274e-16	1.001688	-1.112517	-0.943986	-0.196003	0.658066	3.284732
GDP	297.0	8.032295e-17	1.001688	-1.046096	-0.993971	-0.224273	0.829852	3.319468

Figure Number 21

Now data is scaled and has std = 1 , it seems the presence of outliers but due to scaling the effect is reduced.

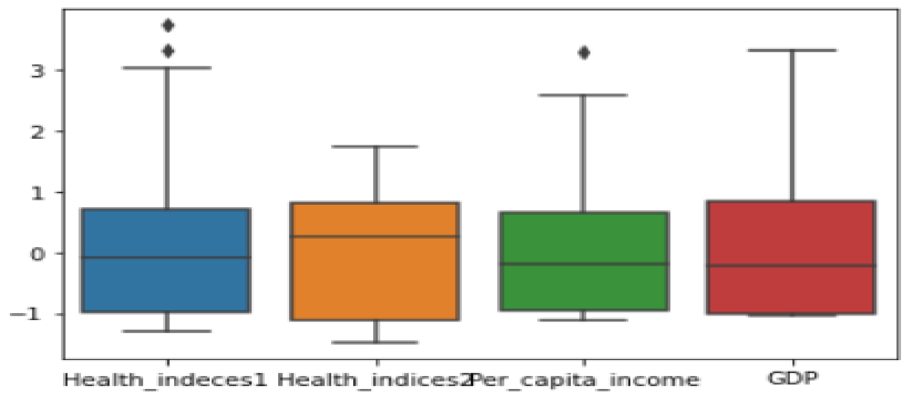


Figure Number 22

Part 2: Clustering: Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.

Below figure is the required dendrogram obtained after applying hierarchical clustering to scaled data,

the last p merged clusters are as follows,

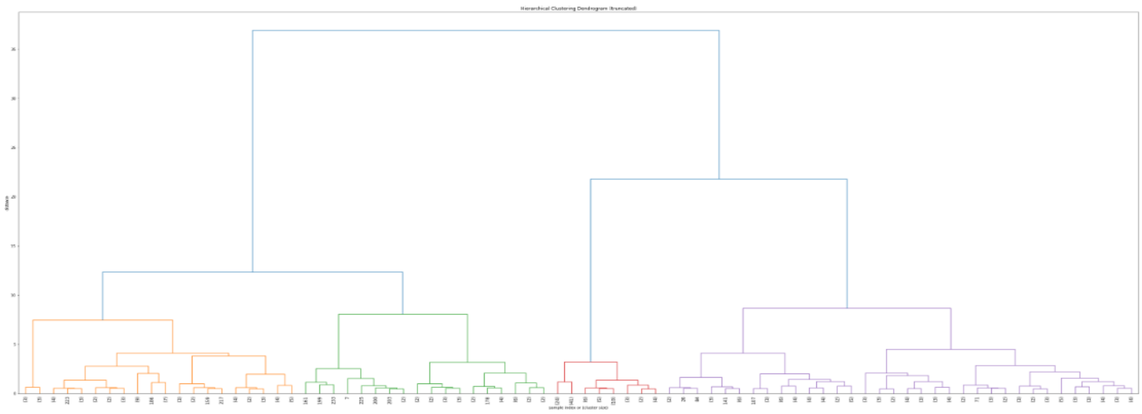


Figure Number 24

Government would like know more than “good” and “not so good” states and hence more insight we are able to generate with more than 2 clusters, better it is for the business. Hence let’s consider 4 clusters and plot the clusters to confirm if the derived clusters are providing the required segmentation details.

Part 2: Clustering: Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and find the silhouette score.

The figure below is an required elbow curve ,

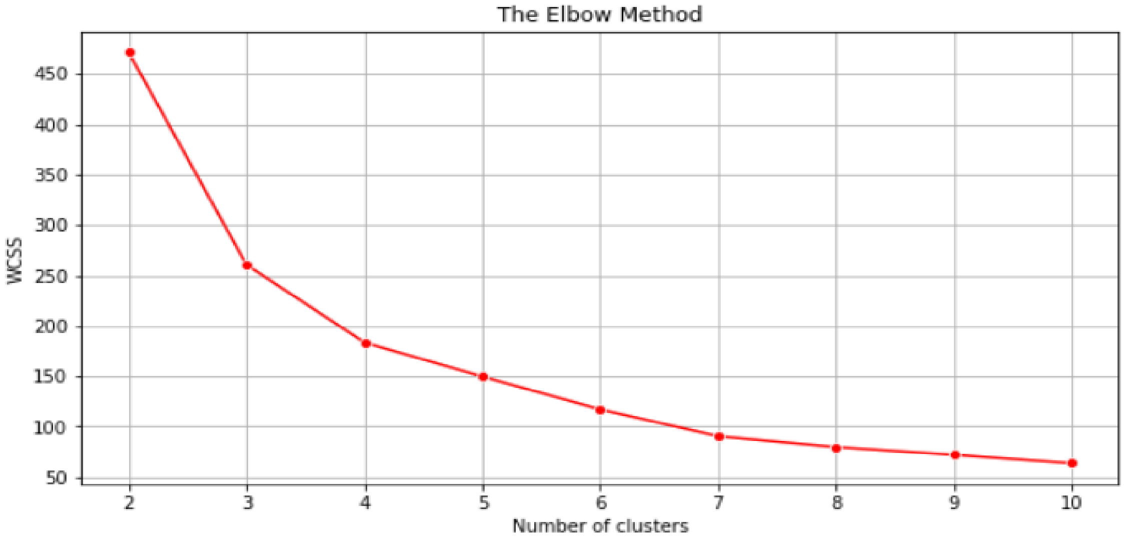


Figure Number 25

K-means clustering technique was used along with elbow curve to define the optimum clusters for this data set. 4 clusters were identified as an optimum number.

Both hierarchical clustering and k-means have provided good segmentation and either one can be used to define strategies.

Part 2: Clustering: Describe cluster profiles for the clusters defined. Recommend different priority based actions that need to be taken for different clusters on the bases of their vulnerability situations according to their Economic and Health Conditions.

	Health_indeces1	Health_indices2	Per_capita_income	GDP	cluster_1	kmeans_cluster_4
0	417	66	564	1823	3	0
1	1485	646	2710	73662	4	2
2	654	299	1104	27318	3	0
3	192	25	573	250	3	0
4	43	8	528	22	3	0

kmeans_cluster_4	Health_indeces1	Health_indices2	Per_capita_income	GDP	cluster_1	cluster count
0	499.158416	116.356436	693.772277	9428.099010	3.059406	101
1	4799.355932	1142.288136	2372.220339	396907.237288	1.000000	59
2	2597.089109	783.019802	2464.128713	141264.138614	3.881188	101
3	5146.444444	1327.138889	5047.083333	367196.916667	2.000000	36