PROJECT DATA MINING

PGPDSBA

Angela Jose

Figure of contents

List of Figure

List of Equation

Equation

List of Table

Problem Statement 1 : 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.

The data file Hair Salon .csv contains 12 variables used for Market Segmentation in the context of Product Service Management.

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

Table 1: Data dictionary

1.Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented.

	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 2: Data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   ProdQual      100 non-null    float64
 1   Ecom          100 non-null    float64
 2   TechSup       100 non-null    float64
 3   CompRes       100 non-null    float64
 4   Advertising   100 non-null    float64
 5   ProdLine      100 non-null    float64
 6   SalesFImage   100 non-null    float64
 7   ComPricing    100 non-null    float64
 8   WartyClaim    100 non-null    float64
 9   OrdBilling    100 non-null    float64
 10  DelSpeed      100 non-null    float64
 11  Satisfaction  100 non-null    float64
dtypes: float64(12)
memory usage: 9.5 KB
```

Figure 3: Data info

	count	mean	std	min	25%	50%	75%	max
ProdQual	100.0	7.81	1.40	5.0	6.57	8.00	9.10	10.0
Ecom	100.0	3.67	0.70	2.2	3.28	3.60	3.92	5.7
TechSup	100.0	5.36	1.53	1.3	4.25	5.40	6.62	8.5
CompRes	100.0	5.44	1.21	2.6	4.60	5.45	6.32	7.8
Advertising	100.0	4.01	1.13	1.9	3.18	4.00	4.80	6.5
ProdLine	100.0	5.80	1.32	2.3	4.70	5.75	6.80	8.4
SalesFImage	100.0	5.12	1.07	2.9	4.50	4.90	5.80	8.2
ComPricing	100.0	6.97	1.55	3.7	5.88	7.10	8.40	9.9
WartyClaim	100.0	6.04	0.82	4.1	5.40	6.10	6.60	8.1
OrdBilling	100.0	4.28	0.93	2.0	3.70	4.40	4.80	6.7
DelSpeed	100.0	3.89	0.73	1.6	3.40	3.90	4.43	5.5
Satisfaction	100.0	6.92	1.19	4.7	6.00	7.05	7.62	9.9

```
ProdQual        0
Ecom            0
TechSup         0
CompRes         0
Advertising     0
ProdLine        0
SalesFImage     0
ComPricing      0
WartyClaim      0
OrdBilling      0
DelSpeed        0
Satisfaction    0
dtype: int64
```

	Attributes	skewness
0	DelSpeed	-0.46
1	OrdBilling	-0.33
2	ProdQual	-0.24
3	ComPricing	-0.24
4	TechSup	-0.20
5	CompRes	-0.14
6	ProdLine	-0.09
7	WartyClaim	0.01
8	Advertising	0.04
9	Satisfaction	0.08
10	SalesFImage	0.38
11	Ecom	0.66

Feature	kurtosis
ProdQual	-1.13
Ecom	0.74
TechSup	-0.55
CompRes	-0.59
Advertising	-0.89
ProdLine	-0.52
SalesFImage	0.41
ComPricing	-0.90
WartyClaim	-0.44
OrdBilling	0.24
DelSpeed	0.22
Satisfaction	-0.79

```
(100, 12)
```

Figure 4: Data description Figure 5: Data null value check Figure 6: Data skewness Figure 7: Data kurtosis

Interference of dataset:

- The data has 12 columns and 100 rows
- All data features or columns are float
- The data detail description is explained in Figure 3 , the units of various features greatly differ.
- The are no duplicate values in the data
- There are no null value in the data
- Skewness: Except WartyClaim and Satisfaction all other features are skewed.
- Kurtosis: SalesFFigure , ordbilling and delspeed has positive kurtosis and rest are negative.

Univariate Analysis

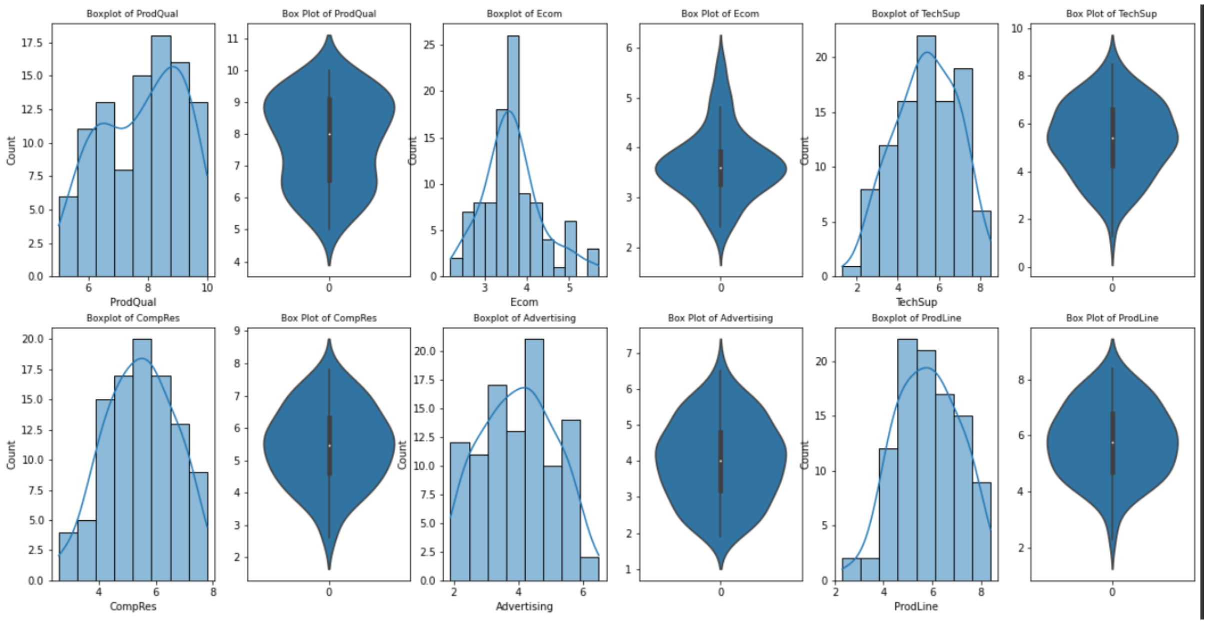


Figure 8: Univariate analysis

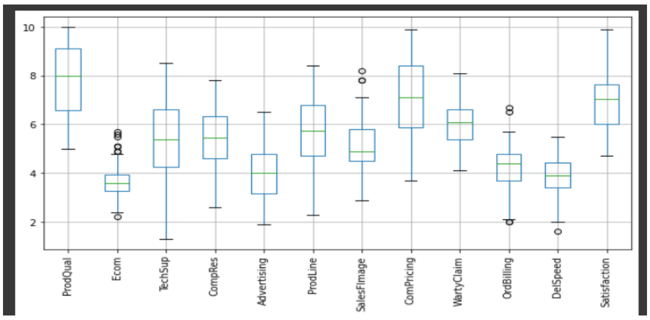


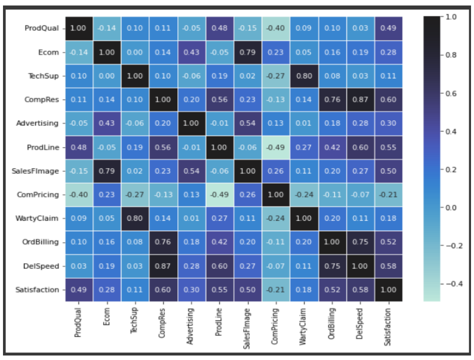
Figure 9: Boxplot analysis

Insight

- Outliers are present for the following features, Ecom, SalesFFigure , Ordbilling and Delspeed.
- All features have almost normal distribution with greatest variation observed for Ecom.
- ProductLine has more features towards the right where prodqual has values in all bins.

Bivariate Analysis

- Features are correlated to each other. CompRes and Delspeed has the maximum correlation.
- Techsup and satisfaction has the least correlation.
- Various features are plotted against satisfaction. Comp pricing has negative linear regression and compRes has most positive linear regression.
- Pairplot of the various features are mapped.



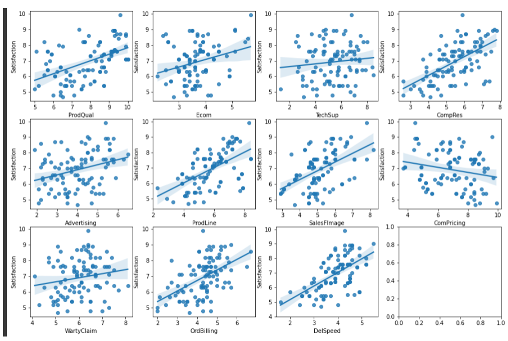


Figure 10: Heatmap Figure 11: Independent vs dependent variable analysis

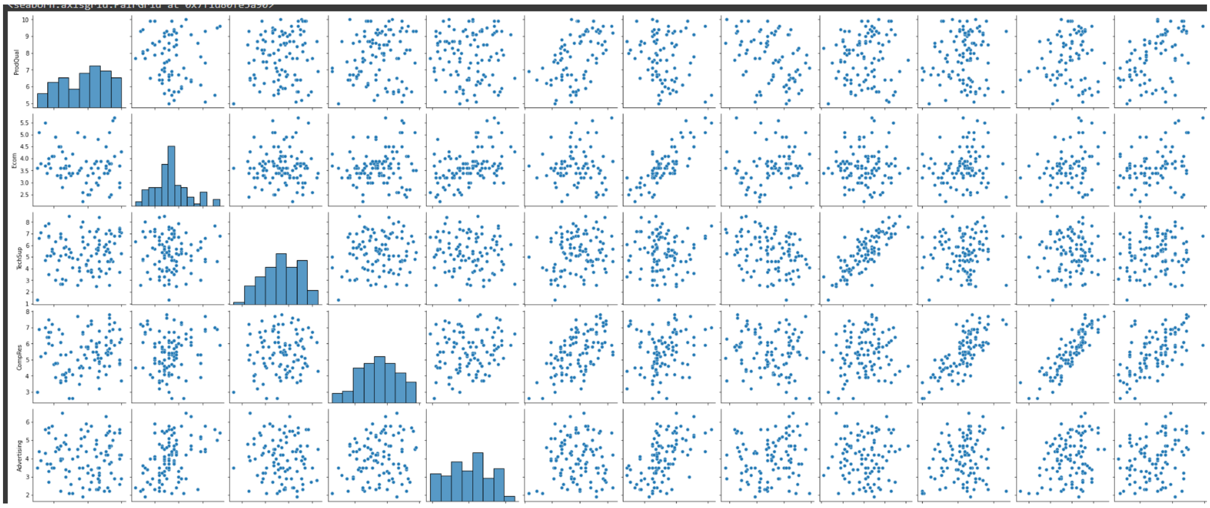
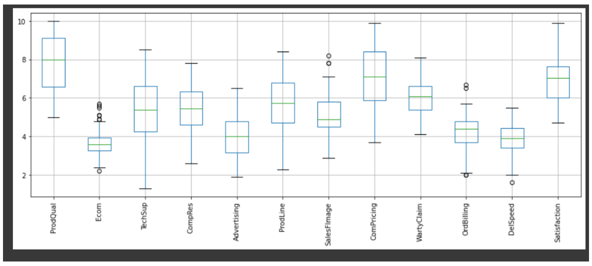


Figure 12: Pairplot

2.Scale the variables and write the inference for using the type of scaling function for this case study.

Outlier -Before outlier treatment data **Outlier -After outlier treatment data**



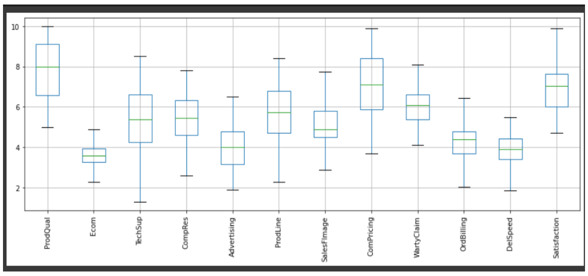


Figure 13: Outlier -Before outlier treatment data Figure 14: Outlier -After outlier treatment data

- Since all the variables are numeric there was no need to remove any columns
- Since we have a dataset with 12 numeric columns of different scales.
- In this case we use both z scaling and min max scaling method.

Min-Max scaling output

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
count	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000
mean	0.562000	0.517692	0.564583	0.546538	0.458696	0.574590	0.457216	0.528065	0.485750	0.505909	0.557010	0.426538
std	0.279256	0.244442	0.212563	0.232385	0.244988	0.215620	0.218065	0.249202	0.204935	0.208840	0.199756	0.229200
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.315000	0.375000	0.409722	0.384615	0.277174	0.393443	0.329897	0.350806	0.325000	0.375000	0.422680	0.250000
50%	0.600000	0.500000	0.569444	0.548077	0.456522	0.565574	0.412371	0.548387	0.500000	0.534091	0.560137	0.451923
75%	0.820000	0.625000	0.739583	0.716346	0.630435	0.737705	0.597938	0.758065	0.625000	0.625000	0.704467	0.562500
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Table 1: Min max scale Equation1: Min max scaling

- The data is between 0 and 1. And the respective features are scaled.

Z Scaling output

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
0	0.496660	0.401668	-1.881421	0.380922	0.704543	-0.691530	0.838627	-0.113185	-1.646582	0.791872	-0.260903	1.081067
1	0.280721	-1.495974	-0.174023	1.462141	-0.544014	1.600835	-1.917200	-1.088915	-0.665744	-0.411249	1.398918	-1.027098
2	1.000518	-0.389017	0.154322	0.131410	1.239639	1.218774	0.648570	-1.609304	0.192489	1.229371	0.845644	1.671354
3	-1.014914	-0.547153	1.073690	-1.448834	0.615361	-0.844354	-0.586801	1.187789	1.173327	0.026250	-1.229132	-1.786038
4	0.856559	-0.389017	-0.108354	-0.700298	-1.614207	0.149004	-0.586801	-0.113185	0.069885	0.244999	-0.537540	0.153474

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   ProdQual      100 non-null    float64
 1   Ecom          100 non-null    float64
 2   TechSup       100 non-null    float64
 3   CompRes       100 non-null    float64
 4   Advertising   100 non-null    float64
 5   ProdLine      100 non-null    float64
 6   SalesFImage   100 non-null    float64
 7   ComPricing    100 non-null    float64
 8   WartyClaim    100 non-null    float64
 9   OrdBilling    100 non-null    float64
 10  DelSpeed      100 non-null    float64
 11  Satisfaction  100 non-null    float64
dtypes: float64(12)
memory usage: 9.5 KB
```

$$Z = \frac{x - \mu}{\sigma}$$

Table 2: z scale Figure 15: Z scale data Equation2: Z scale

- Z score tells us how many standard deviations is the point away from the mean and also the direction. Now, the value is scaled between-1 and 1.

Check if data available is ok for further PCA check

Bartletts Test of Sphericity

- Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population.
- HO : All variables in the data are uncorrelated
- HA : At least one pair of variables in the data are correlated
- If the null hypothesis cannot be rejected, then PCA is not advisable

P value = 1.521

KMO Test

- The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.
- Generally, if MSA is less than 0.5, PCA is not recommended, since no reduction is expected. On the other hand, MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

MSA = 0.661

3.Comment on the comparison between covariance and the correlation matrix after scaling.

- Covariance measures how two variables are related to each other and whether they increase or decrease together. However, the magnitude of covariance is influenced by the scale of the variables.
- This means that variables with larger scales will have a greater influence on the covariance value than variables with smaller scales.
- Therefore, it can be difficult to compare covariances across variables that have different scales.
- On the other hand, correlation measures the linear relationship between two variables and is scaled to fall between -1 and 1.
- By scaling the values, correlation coefficients can be compared directly, making it easier to understand the strength and direction of relationships between variables.
- This is particularly important when comparing variables with different scales, as correlation is not influenced by the scale of the variables.
- While covariance is a measure of the extent to which two variables change together, correlation measures the strength and direction of the linear relationship between two variables.

```
Covariance Matrix
 [[ 1.01010101e+00 -1.38548704e-01  9.65661154e-02  1.07444445e-01
  -5.40132667e-02  4.82316579e-01 -1.53346338e-01 -4.05335236e-01
   8.92043497e-02  1.05356640e-01  2.79979825e-02  4.91237372e-01]
 [-1.38548704e-01  1.01010101e+00  8.75544162e-04  1.41595213e-01
   4.34233041e-01 -5.32200387e-02  7.99539102e-01  2.31780203e-01
   5.24224157e-02  1.57724577e-01  1.93571786e-01  2.85601025e-01]
 [ 9.65661154e-02  8.75544162e-04  1.01010101e+00  9.76329270e-02
  -6.35051180e-02  1.94571168e-01  1.71621612e-02 -2.73521901e-01
   8.05220127e-01  8.09109340e-02  2.56976702e-02  1.13734524e-01]
 [ 1.07444445e-01  1.41595213e-01  9.76329270e-02  1.01010101e+00
   1.98905906e-01  5.67087831e-01  2.32072486e-01 -1.29246720e-01
   1.41826562e-01  7.64513729e-01  8.73829997e-01  6.09356166e-01]
 [-5.40132667e-02  4.34233041e-01 -6.35051180e-02  1.98905906e-01
   1.01010101e+00 -1.16674936e-02  5.47680463e-01  1.35572620e-01
   1.09010852e-02  1.86096560e-01  2.78649579e-01  3.07746944e-01]
 [ 4.82316579e-01 -5.32200387e-02  1.94571168e-01  5.67087831e-01
  -1.16674936e-02  1.01010101e+00 -6.19348764e-02 -4.99947880e-01
   2.75835887e-01  4.28695202e-01  6.07929503e-01  5.56107006e-01]
 [-1.53346338e-01  7.99539102e-01  1.71621612e-02  2.32072486e-01
   5.47680463e-01 -6.19348764e-02  1.01010101e+00  2.67269246e-01
   1.08540752e-01  1.97098390e-01  2.74294201e-01  5.05257885e-01]
 [-4.05335236e-01  2.31780203e-01 -2.73521901e-01 -1.29246720e-01
   1.35572620e-01 -4.99947880e-01  2.67269246e-01  1.01010101e+00
  -2.47460661e-01 -1.15724268e-01 -7.36078070e-02 -2.10399686e-01]
 [ 8.92043497e-02  5.24224157e-02  8.05220127e-01  1.41826562e-01
   1.09010852e-02  2.75835887e-01  1.08540752e-01 -2.47460661e-01
   1.01010101e+00  1.99055678e-01  1.10499598e-01  1.79338201e-01]
 [ 1.05356640e-01  1.57724577e-01  8.09109340e-02  7.64513729e-01
   1.86096560e-01  4.28695202e-01  1.97098390e-01 -1.15724268e-01
   1.99055678e-01  1.01010101e+00  7.58588957e-01  5.27001932e-01]
 [ 2.79979825e-02  1.93571786e-01  2.56976702e-02  8.73829997e-01
   2.78649579e-01  6.07929503e-01  2.74294201e-01 -7.36078070e-02
   1.10499598e-01  7.58588957e-01  1.01010101e+00  5.82870984e-01]
 [ 4.91237372e-01  2.85601025e-01  1.13734524e-01  6.09356166e-01
   3.07746944e-01  5.56107006e-01  5.05257885e-01 -2.10399686e-01
   1.79338201e-01  5.27001932e-01  5.82870984e-01  1.01010101e+00]]
```

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
ProdQual	1.000000	-0.137163	0.095600	0.106370	-0.053473	0.477493	-0.151813	-0.401282	0.088312	0.104303	0.027718	0.486325
Ecom	-0.137163	1.000000	0.000867	0.140179	0.429891	-0.052688	0.791544	0.229462	0.051898	0.156147	0.191636	0.282745
TechSup	0.095600	0.000867	1.000000	0.096657	-0.062870	0.192625	0.016991	-0.270787	0.797168	0.080102	0.025441	0.112597
CompRes	0.106370	0.140179	0.096657	1.000000	0.196917	0.561417	0.229752	-0.127954	0.140408	0.756869	0.865092	0.603263
Advertising	-0.053473	0.429891	-0.062870	0.196917	1.000000	-0.011551	0.542204	0.134217	0.010792	0.184236	0.275863	0.304669
ProdLine	0.477493	-0.052688	0.192625	0.561417	-0.011551	1.000000	-0.061316	-0.494948	0.273078	0.424408	0.601850	0.550546
SalesFImage	-0.151813	0.791544	0.016991	0.229752	0.542204	-0.061316	1.000000	0.264597	0.107455	0.195127	0.271551	0.500205
ComPricing	-0.401282	0.229462	-0.270787	-0.127954	0.134217	-0.494948	0.264597	1.000000	-0.244986	-0.114567	-0.072872	-0.208296
WartyClaim	0.088312	0.051898	0.797168	0.140408	0.010792	0.273078	0.107455	-0.244986	1.000000	0.197065	0.109395	0.177545
OrdBilling	0.104303	0.156147	0.080102	0.756869	0.184236	0.424408	0.195127	-0.114567	0.197065	1.000000	0.751003	0.521732
DelSpeed	0.027718	0.191636	0.025441	0.865092	0.275863	0.601850	0.271551	-0.072872	0.109395	0.751003	1.000000	0.577042
Satisfaction	0.486325	0.282745	0.112597	0.603263	0.304669	0.550546	0.500205	-0.208296	0.177545	0.521732	0.577042	1.000000

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{N} \qquad Correlation = \frac{Cov\ (x,y)}{\sigma x * \sigma y}$$

Figure16: Covariance matrix Table 3: Correlation matrix Equation3: Covariance Equation4:Correlation

4.Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.

Figure 16: **Outlier before scaling**

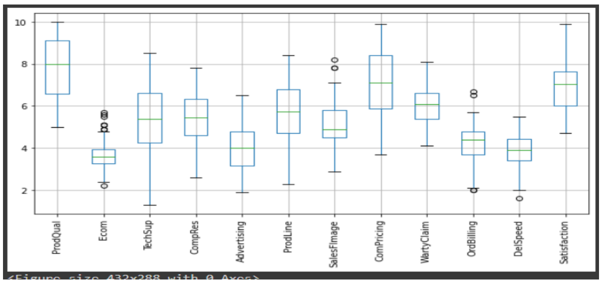


Figure 17: **Outlier after scaling**

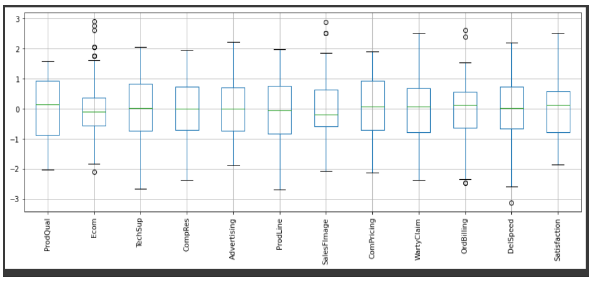
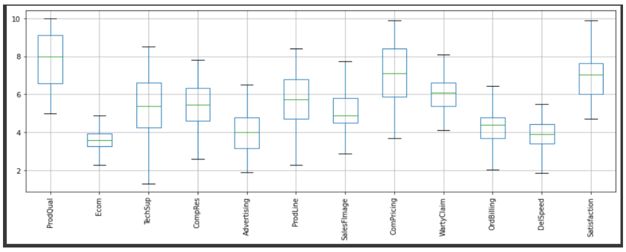


Figure 18:**Outlier - After outlier treatment**



Inference

- Outlier are present in both scaled and unscaled data
- Scaling does not remove outliers, scaling here is done with z scaling.
- Used capping to remove outliers . I.e any value above 3 IQR is imputed with IQR value.

5.Build the covariance matrix, eigenvalues and eigen vector.

The covariance matrix is denoted in the Figure and eigen value, eigen vector.

A **covariance matrix** is a square matrix with diagonal elements which represent the variance and the non-diagonal components that express covariance. The covariance of a variable can take any real value- positive, negative, or zero. A positive covariance suggests that the two variables have a positive relationship, whereas a negative covariance indicates that they do not. If two elements do not vary together, they have a zero covariance.

Eigen Values

The factor by which the magnitude of an eigenvector is changed by a given transformation.

The change in magnitude of a vector that does not change in direction under a given linear transformation; a scalar factor by which an eigenvector is multiplied under such a transformation.

(mathematics) any number such that a given square matrix minus that number times the identity matrix has a zero determinant

Eigenvector of a square matrix is defined as a non-vector in which when a given matrix is multiplied, it is equal to a scalar multiple of that vector. Let us suppose that A is an n x n square matrix, and if v be a non-zero vector, then the product of matrix A, and vector v is defined as the product of a scalar quantity λ and the given vector, such that:

$Av = \lambda v$

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
ProdQual	1.01	-0.16	0.10	0.11	-0.05	0.48	-0.15	-0.41	0.09	0.10	0.02	0.49
Ecom	-0.16	1.01	-0.02	0.11	0.43	-0.10	0.79	0.27	0.03	0.15	0.17	0.24
TechSup	0.10	-0.02	1.01	0.10	-0.06	0.19	0.01	-0.27	0.81	0.09	0.03	0.11
CompRes	0.11	0.11	0.10	1.01	0.20	0.57	0.23	-0.13	0.14	0.77	0.88	0.61
Advertising	-0.05	0.43	-0.06	0.20	1.01	-0.01	0.55	0.14	0.01	0.19	0.28	0.31
ProdLine	0.48	-0.10	0.19	0.57	-0.01	1.01	-0.06	-0.50	0.28	0.43	0.61	0.56
SalesFImage	-0.15	0.79	0.01	0.23	0.55	-0.06	1.01	0.27	0.10	0.20	0.27	0.51
ComPricing	-0.41	0.27	-0.27	-0.13	0.14	-0.50	0.27	1.01	-0.25	-0.11	-0.07	-0.21
WartyClaim	0.09	0.03	0.81	0.14	0.01	0.28	0.10	-0.25	1.01	0.20	0.12	0.18
OrdBilling	0.10	0.15	0.09	0.77	0.19	0.43	0.20	-0.11	0.20	1.01	0.76	0.53
DelSpeed	0.02	0.17	0.03	0.88	0.28	0.61	0.27	-0.07	0.12	0.76	1.01	0.58
Satisfaction	0.49	0.24	0.11	0.61	0.31	0.56	0.51	-0.21	0.18	0.53	0.58	1.01

```
Eigen Values:
 [4.0604 2.6156 1.6962 1.2299 0.6447 0.5631 0.4064 0.3394 0.2373 0.146
 0.0982 0.0842]


Eigen Vectors:
 [[-1.613e-01 -1.390e-01 -1.271e-01 -4.255e-01 -1.773e-01 -3.565e-01
  -2.104e-01  1.376e-01 -1.767e-01 -3.912e-01 -4.250e-01 -4.133e-01]
 [-3.063e-01  4.549e-01 -2.353e-01  8.900e-03  3.559e-01 -2.899e-01
   4.649e-01  4.155e-01 -1.978e-01  2.060e-02  6.260e-02  2.960e-02]
 [ 7.950e-02 -2.299e-01 -6.217e-01  1.918e-01 -9.220e-02  1.128e-01
  -2.366e-01  4.500e-02 -6.114e-01  1.428e-01  2.077e-01  3.040e-02]
 [ 6.165e-01  1.838e-01 -1.665e-01 -2.799e-01  2.147e-01  9.850e-02
   2.130e-01 -2.369e-01 -1.755e-01 -3.034e-01 -2.939e-01  3.370e-01]
 [-2.567e-01 -1.960e-01 -4.320e-02 -3.100e-02  7.633e-01  1.960e-02
  -1.387e-01 -4.843e-01 -2.290e-02 -4.970e-02  5.540e-02 -2.237e-01]
 [ 3.497e-01 -4.721e-01  1.190e-01  2.270e-02  4.105e-01 -1.943e-01
  -1.703e-01  6.007e-01  1.370e-01  7.620e-02 -2.670e-02  1.372e-01]
 [ 1.596e-01  4.580e-02 -1.900e-03 -5.700e-03 -5.500e-02 -6.243e-01
  -2.160e-02 -3.186e-01 -4.430e-02  6.478e-01 -2.333e-01  4.140e-02]
 [-3.288e-01 -5.096e-01  5.570e-02  1.366e-01 -1.422e-01 -2.709e-01
   3.525e-01 -1.804e-01 -9.000e-02 -2.793e-01 -2.220e-02  5.229e-01]
 [-1.685e-01 -1.981e-01 -5.563e-01 -4.360e-01 -4.160e-02  2.173e-01
   1.581e-01  3.150e-02  5.126e-01  2.764e-01 -7.820e-02  1.123e-01]
 [ 2.266e-01  4.240e-02 -4.160e-01  5.641e-01 -3.510e-02 -2.764e-01
   4.980e-02 -9.660e-02  4.511e-01 -3.269e-01 -6.700e-03 -2.361e-01]
 [ 1.979e-01 -2.000e-03  6.000e-04 -4.185e-01 -8.360e-02 -3.444e-01
   1.080e-02 -1.017e-01  6.250e-02 -1.497e-01  7.881e-01 -4.750e-02]
 [-2.305e-01  3.507e-01 -1.121e-01 -1.210e-02  5.510e-02 -1.515e-01
  -6.616e-01  1.570e-02  1.598e-01 -1.500e-01 -5.000e-03  5.470e-01]]
```

Table4:Covariance matrix Figure 19: Eigen value and eigen vector

6.Write the explicit form of the first PC (in terms of EigenVectors).

```
Index(['ProdQual', 'Ecom', 'TechSup', 'CompRes', 'Advertising', 'ProdLine',
       'SalesFImage', 'ComPricing', 'WartyClaim', 'OrdBilling', 'DelSpeed',
       'Satisfaction'],
      dtype='object')
```

Figure 20: Column name of data Figure 21: Eigen values

```
array([[-1.61322427e-01, -1.38992261e-01, -1.27131534e-01,
        -4.25502348e-01, -1.77257763e-01, -3.56524003e-01,
        -2.10387616e-01,  1.37603805e-01, -1.76706227e-01,
        -3.91237634e-01, -4.24958585e-01, -4.13318690e-01]
       [-3.06272359e-01,  4.54921339e-01, -2.35263231e-01,
         8.86065191e-03,  3.55907323e-01, -2.89852601e-01,
         4.64926872e-01,  4.15466718e-01, -1.97843283e-01,
         2.05739475e-02,  6.26377391e-02,  2.95573690e-02]
       [ 7.95045575e-02, -2.29883744e-01, -6.21730460e-01,
         1.91750596e-01, -9.22380787e-02,  1.12809185e-01,
        -2.36626212e-01,  4.49919990e-02, -6.11385841e-01,
         1.42820217e-01,  2.07727869e-01,  3.04020598e-02]
       [ 6.16476615e-01,  1.83792626e-01, -1.66476236e-01,
        -2.79905722e-01,  2.14732458e-01,  9.85304039e-02,
         2.12995164e-01, -2.36864713e-01, -1.75501531e-01,
        -3.03399090e-01, -2.93932094e-01,  3.37012361e-01]
       [-2.56708792e-01, -1.95989018e-01, -4.32018329e-02,
        -3.10014556e-02,  7.63273860e-01,  1.96214033e-02,
        -1.38679963e-01, -4.84289239e-01, -2.28877328e-02,
        -4.96697647e-02,  5.53883726e-02, -2.23746335e-01]
       [ 3.49665681e-01, -4.72109013e-01,  1.18961241e-01,
         2.27476118e-02,  4.10458402e-01, -1.94306175e-01,
        -1.70268215e-01,  6.00687375e-01,  1.37026464e-01,
         7.61903184e-02, -2.66931588e-02,  1.37245917e-01]
       [ 1.59566085e-01,  4.58420598e-02, -1.85228111e-03,
        -5.65413580e-03, -5.50372696e-02, -6.24254550e-01,
        -2.15976650e-02, -3.18607905e-01, -4.42538118e-02,
         6.47750462e-01, -2.33344595e-01,  4.13848752e-02]
       [-3.28835920e-01, -5.09595822e-01,  5.57062806e-02,
         1.36570710e-01, -1.42161086e-01, -2.70924156e-01,
         3.52510892e-01, -1.80376001e-01, -9.00155002e-02,
        -2.79292694e-01, -2.22438761e-02,  5.22889275e-01]
       [-1.68511089e-01, -1.98053391e-01, -5.56287356e-01,
        -4.35955358e-01, -4.16389528e-02,  2.17278275e-01,
         1.58074558e-01,  3.15183851e-02,  5.12637449e-01,
         2.76449344e-01, -7.81797655e-02,  1.12285686e-01]
       [ 2.26630723e-01,  4.24260261e-02, -4.16015115e-01,
         5.64127007e-01, -3.51473948e-02, -2.76430381e-01,
         4.97639173e-02, -9.66456687e-02,  4.51055742e-01,
        -3.26877107e-01, -6.69243234e-03, -2.36050949e-01]
```

The Linear eq of 1st component:

-0.161 * ProdQual + -0.139 * Ecom + -0.127 * TechSup + -0.426 * CompRes + -0.177 * Advertising + -0.357 * ProdLine + -0.21 * SalesFFigure + 0.138 * ComPricing + -0.177 * WartyClaim + -0.391 * OrdBilling + -0.425 * DelSpeed + -0.413 * Satisfaction

7.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

Table 5:Cumulative value of eigen value

```
array([0.3349843 , 0.55076722, 0.6907002 , 0.79217069, 0.84535601,
       0.89181057, 0.92533555, 0.95333455, 0.97290782, 0.98495303,
       0.99305436, 1.        ])
```

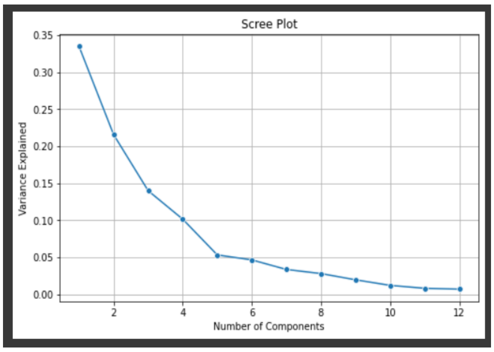


Figure 22: Scree plot

Adding all eigen value equals to 1. After 5 the cumulative sum incremental is not much (<5%).
So based on this the optimum number of the cluster is 5.
The eigen vectors or Pc for the case study is 5. With this eigen vectors we can understand which variables has more weightage and influences the dataset in the principal components. pca helps to reduce collinearity and improves efficiency scores.

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
0	-0.161322	-0.138992	-0.127132	-0.425502	-0.177258	-0.356524	-0.210388	0.137604	-0.176706	-0.391238	-0.424959	-0.413319
1	-0.306272	0.454921	-0.235263	0.008861	0.355907	-0.289853	0.464927	0.415467	-0.197843	0.020574	0.062638	0.029557
2	0.079505	-0.229884	-0.621730	0.191751	-0.092238	0.112809	-0.236626	0.044992	-0.611386	0.142820	0.207728	0.030402
3	0.616477	0.183793	-0.166476	-0.279906	0.214732	0.098530	0.212995	-0.236865	-0.175502	-0.303399	-0.293932	0.337012
4	-0.256709	-0.195989	-0.043202	-0.031001	0.763274	0.019621	-0.138680	-0.484289	-0.022888	-0.049670	0.055388	-0.223746

Table 6: Data of the Principal Component scores into a data frame
PCA is performed and it is exported into a dataframe. After pca the multicollinearity is highly reduced.

8.Mention the business implication of using the Principal Component Analysis for this case study.

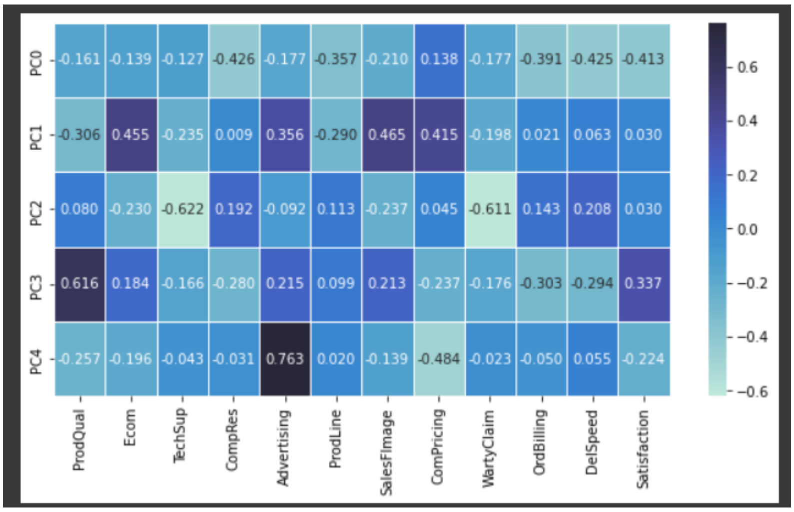


Table 7: 5 PC value wrt to each features

Conclusion

- Based on the above analysis, I observed that maximum data can be captured within 5 PCs.
- Each PC varies from each other based on the information they convey.
- For instance in PC0 compres (Complaint resolution), Prodline(Product line) and ordbilling(order and billing) contributes the most. So taking care of these parameters will lead to maximum customer satisfaction.
- Similar for PC1, Ecom(e-commerce), salesfFigure (Salesforce Figure) and comprising plays an important role in customer satisfaction.
- For PC2 advertisement plays the most important role in customer satisfaction
- For PC3 Product quantity plays the most important role in customer satisfaction
- For PC3 advertisement plays the most important role in customer satisfaction
- So by properly monitoring the following features the customer satisfaction can be increased.
- Another advantage is it helped to reduce multicollinearity and helped to reduce dimensions while maintaining maximum variation as possible.

Problem Statement 2:

The 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)

	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

Table 8: Sample of the dataset 2

Remove unwanted column - Unnamed : 0

Size of dataset:

There are 5 columns and 297 rows

Data type:

All features are integer except States which is object

```
States               object
Health_indeces1       int64
Health_indices2       int64
Per_capita_income     int64
GDP                   int64
dtype: object
```

Figure 23 : Data type of dataset 2

Basic information of data:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 297 entries, 0 to 296
Data columns (total 5 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   States             297 non-null    object
 1   Health_indeces1    297 non-null    int64 
 2   Health_indices2    297 non-null    int64 
 3   Per_capita_income  297 non-null    int64 
 4   GDP                297 non-null    int64 
dtypes: int64(4), object(1)
memory usage: 11.7+ KB
```

Figure 24: Data info of dataset 2

Data Description:

	count	mean	std	min	25%	50%	75%	max
Health_indeces1	297.0	2630.15	2038.51	-10.0	641.0	2451.0	4094.0	10219.0
Health_indices2	297.0	693.63	468.94	0.0	175.0	810.0	1073.0	1508.0
Per_capita_income	297.0	2156.92	1491.85	500.0	751.0	1865.0	3137.0	7049.0
GDP	297.0	174601.12	167167.99	22.0	8721.0	137173.0	313092.0	728575.0

Table 9:Data description

Exploratory Data Analysis

Check for null value:

	0
States	0
Health_indeces1	0
Health_indices2	0
Per_capita_income	0
GDP	0

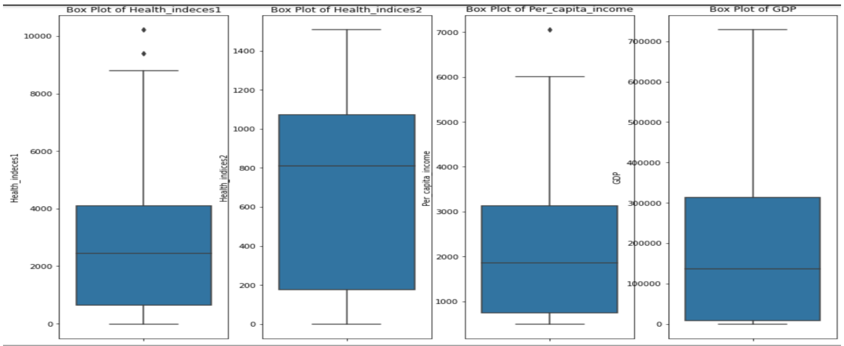
Figure 25: Check for null value

There are no null value in the data

Check for Duplicates:

There are no duplicate values

Figure 26 :Check for Outlier:



Insight:

- There are outliers in health_index1 and per capita income
- Outlier are treated in this case by capping method

Univariate Analysis

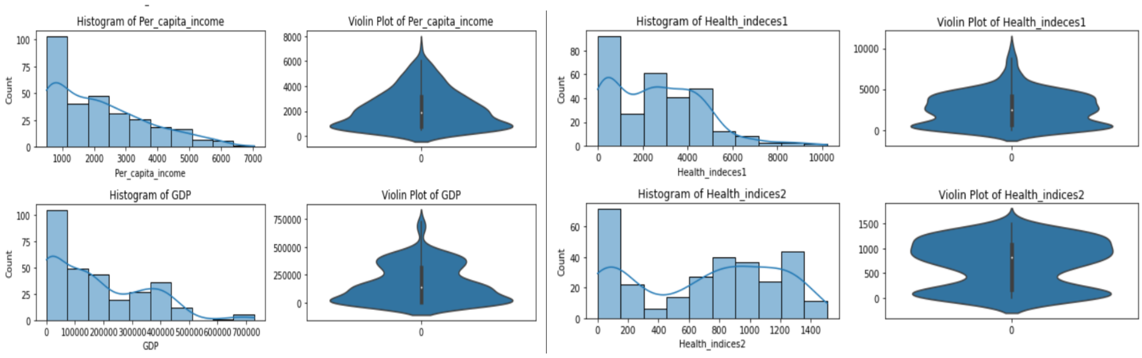


Figure 27 :Univariate Analysis

	Attributes	skewness
0	Health_indices2	-0.17
1	Health_indeces1	0.67
2	Per_capita_income	0.81
3	GDP	0.83

Feature	kurtosis
Health_indeces1	0.44
Health_indices2	-1.40
Per_capita_income	-0.12
GDP	0.06

Figure 28 :Skewness Figure 29 :Kurtosis

Insight:

From above plots and tables, we can conclude below points,

- Health Indices 1 and GDP features have positive kurtosis.
- Health Indices 2 and Per capita income features have negative kurtosis.
- The health_indice2 is left skewed and all other parameters are right skewed.

Bivariate Analysis

Pairplot between numeric continuous variable

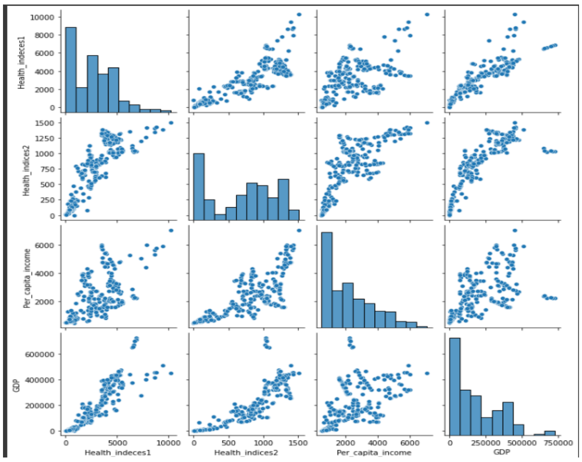


Figure30 :Pairplot

Heatmap

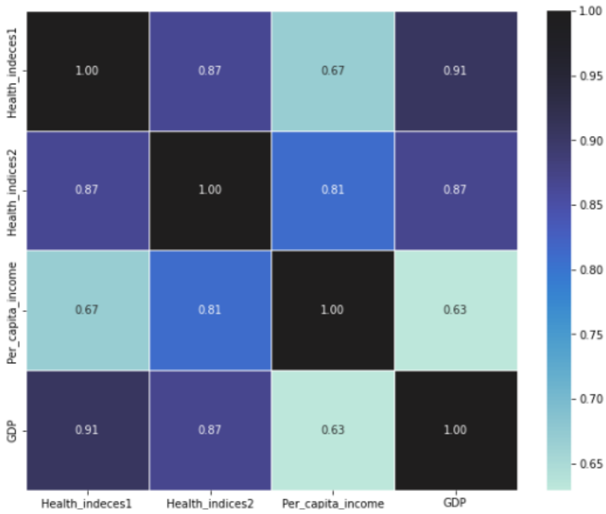


Figure31 :Heatmap

Insight:

From pairplot and heatmap observed the following points:

- Few features have strong correlations between them . Health_index1 and GDP(0.91) and Health_index2 and GDP(0.87)
- Few features have mild correlations per capita income and health index1(0.67)

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

- Scaling is required to bring all the features into a common scale before proceeding to clustering. It is necessary for all distance based models.
- If we don't scale the data, it gives higher weightage to features which have higher magnitude.

Feature	mean	std	min	max	variance
Health_indeces1	2626.53	2025.87	-10.0	9273.5	4.104149e+06
Health_indices2	693.63	468.94	0.0	1508.0	2.199047e+05
Per_capita_income	2155.79	1488.29	500.0	6716.0	2.215007e+06
GDP	174601.12	167167.99	22.0	728575.0	2.794514e+10

Table 10:Data description of each feature

Here the mean, min, max, std and variance are highly varied as the data is not scaled.

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

Scale the data using z score method : The data was scaled using the z score method.

	Health_indeces1	Health_indices2	Per_capita_income	GDP
0	-1.09	-1.34	-1.07	-1.04
1	-0.56	-0.10	0.37	-0.60
2	-0.98	-0.84	-0.71	-0.88
3	-1.20	-1.43	-1.07	-1.04
4	-1.28	-1.46	-1.10	-1.05

Table 11:Data zscore scaled

Dendrogram

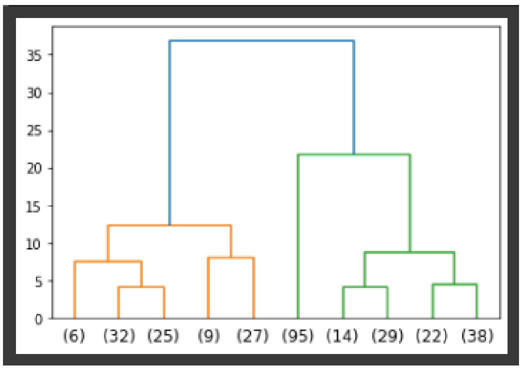


Figure32 :Dendrogram

Selecting the optimum number of cluster

- From the above truncated dendrogram, the distance or increase within sum squares (WSS) is large to merge the last two clusters into a single final cluster.
- We would not get additional information with 2 numbers of clusters.
- Therefore based on the next optimum number of clusters selected based on distance or increase within sum squares (WSS) are three.

Hierarchy cluster label:

```
array([2, 3, 2, 2, 2, 2, 2, 1, 2, 3, 2, 2, 2, 2, 2, 2, 1, 2, 2, 1, 3, 2,
       2, 2, 3, 2, 3, 3, 3, 2, 3, 2, 2, 1, 2, 2, 1, 3, 2, 3, 2, 2, 3, 3,
       3, 2, 2, 1, 2, 2, 2, 3, 2, 1, 2, 3, 3, 2, 2, 3, 3, 2, 2, 1, 3, 2,
       3, 2, 2, 2, 2, 3, 2, 2, 2, 2, 3, 3, 2, 2, 2, 2, 2, 2, 3, 3, 2, 2,
       2, 1, 2, 2, 2, 3, 2, 2, 3, 2, 2, 3, 2, 2, 1, 2, 3, 2, 3, 3, 2, 2,
       2, 1, 3, 3, 2, 2, 2, 3, 2, 2, 3, 2, 2, 2, 1, 2, 3, 2, 2, 3, 2, 2,
       2, 3, 2, 2, 2, 2, 2, 3, 2, 3, 2, 2, 2, 3, 1, 2, 3, 2, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3, 3, 3, 3,
       3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
       3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
       3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3], dtype=int32)
```

Figure33 :Hierarchy cluster

Following is an array of cluster numbers for all observations in the dataset.

Sample of clustered dataset:

	States	Health_indeces1	Health_indices2	Per_capita_income	GDP	Hclusters
0	Bachevo	417	66	564	1823	2
1	Balgarchevo	1485	646	2710	73662	3
2	Belasitsa	654	299	1104	27318	2
3	Belo_Pole	192	25	573	250	2
4	Beslen	43	8	528	22	2

Table 12:Sample of clustered dataset

Customer segmentation:

	Health_indeces1	Health_indices2	Per_capita_income	GDP
Hclusters				
1	4923.5	1201.6	3375.1	377132.5
2	401.1	104.5	680.7	5388.8
3	2481.8	748.7	2347.6	136004.7

Table13:Customer segmentation

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

K-Means clustering

- K means clustering is used to find similar groups or assign the data points to cluster on the basis of their similarity
- First step is to find the number of clusters before applying the model and optimum number of clusters can be found by plotting elbow curve for within sum squares (WSS)

Optimum no. of clusters by elbow plot method

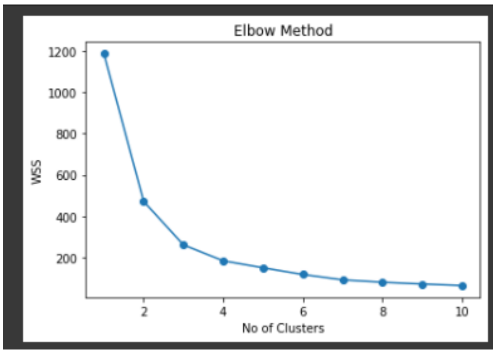
- This method is based on plotting the values of within sum squares against different no. of clusters (k). As the number of clusters increases, WSS decreases.
- The no. of clusters at which decrease in WSS is not significant is known as optimum no. of clusters.

	Number_of_Clusters	WSS
0	1	1188.00
1	2	471.36
2	3	260.57
3	4	183.61
4	5	149.79
5	6	117.32
6	7	91.32
7	8	79.80
8	9	71.07
9	10	63.73

Table14: No.of clusters vs WSS

- WSS is plotted along y axis and number of clusters along x axis
- From the elbow plot, we observed that the optimal cluster number is three.

Figure34 :Elbow plot



Optimum no of clusters by silhouette score method:

- Silhouette scores are calculated for different no of clusters and tabulated .
- Silhouette score plot is drawn by taking no of clusters (k) on x axis and silhouette score values on y axis

	Number_of_Clusters	Silhouette_Score
0	2	0.53
1	3	0.53
2	4	0.55
3	5	0.52
4	6	0.53
5	7	0.56
6	8	0.53
7	9	0.51
8	10	0.51

Table15: No.of clusters vs Silhouette score Figure 35 :Silhouette score plot

- From the above plot, I noticed that maximum silhouette scores exist at four clusters (0.55) and seven clusters (0.56).
- Based on this we conclude the optimum number of clusters is three.

K Mean cluster label:

```
array([0, 1, 0, 0, 0, 0, 0, 2, 0, 1, 0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 1, 0,
       0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 2, 0, 0, 2, 1, 0, 1, 0, 0, 1, 1,
       1, 0, 0, 1, 0, 0, 0, 1, 0, 2, 0, 1, 1, 0, 0, 1, 1, 0, 0, 2, 1, 0,
       1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
       0, 2, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 2, 0, 1, 0, 1, 1, 0, 0,
       0, 2, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 2, 0, 1, 0, 0, 1, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 2, 0, 1, 0, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int32)
```

Figure 36 :K Mean cluster

	States	Health_indeces1	Health_indices2	Per_capita_income	GDP	kmclusters
0	Bachevo	417	66	564	1823	0
1	Balgarchevo	1485	646	2710	73662	1
2	Belasitsa	654	299	1104	27318	0
3	Belo_Pole	192	25	573	250	0
4	Beslen	43	8	528	22	0

Table16: K mean cluster data classification

Distribution of states among K mean cluster

0	101
1	101
2	95

Table17: K mean cluster states distribution

The distribution of states are almost uniform in k mean clusters.

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.

Hierarchical Clustering

Hclusters	Health_indeces1	Health_indices2	Per_capita_income	GDP
1	4912.7	1201.6	3371.8	377132.5
2	401.1	104.5	680.7	5388.8
3	2481.8	748.7	2347.6	136004.7

Table18: Hierarchical Clustering

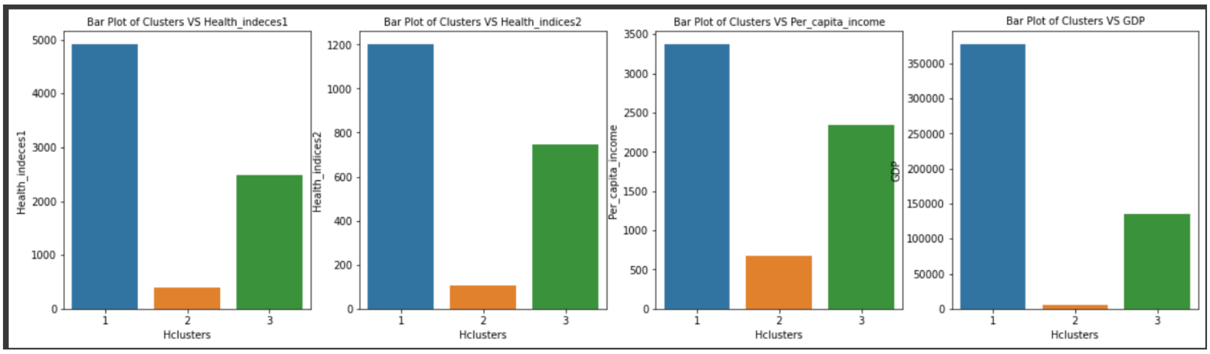


Figure 37: Barplot of Hierarchy cluster

Distribution of states among Hierarchical cluster

3	103
1	99
2	95

Name: Hclusters, dtype: int64

Table19: Hierarchical Clustering

K-Means Clustering

kmclusters	Health_indeces1	Health_indices2	Per_capita_income	GDP
0	499.2	116.4	693.8	9428.1
1	4919.6	1212.3	3382.3	385648.6
2	2597.1	783.0	2464.1	141264.1

Table20:K-Means Clustering

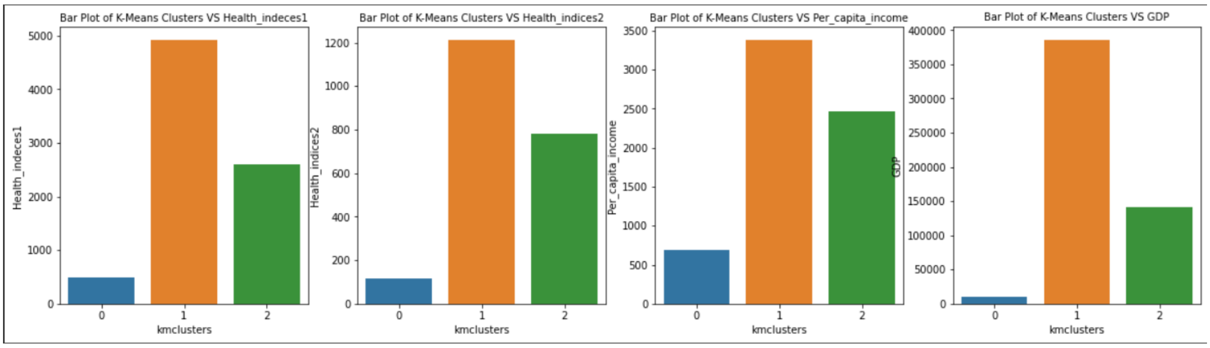


Figure 38: Barplot of k means cluster

Distribution of states among K-Means Clustering

0	101
2	101
1	95

Table21:Distribution of states K-Means Clustering

Conclusion:

By comparing means of different features in Hierarchical Clustering & K-Means Clustering, we can notice below key points.

- Cluster 1 in Hierarchical Clustering (high health indices, high Per capita income and high GDP) is equivalent to Cluster 2 in K-Means Clustering.
- Cluster 2 in Hierarchical Clustering (low health indices, low Per capita income and low GDP) is equivalent to Cluster 0 in K-Means Clustering.
- Cluster 3 in Hierarchical Clustering (moderate health indices, moderate Per capita income and moderate GDP) is equivalent to Cluster 1 in K-Means Clustering.
- States in K Means cluster 2 have high health indices, high Per capita income and high GDP. Hence, we can notice that these states may be considered as developed states. Based on the budget availability, the government should introduce new strategies to improve health indices, per capita income and GDP and also government should strictly keep implementing the strategies which are already being executed in healthcare and financial departments (Equivalent to Cluster 3 in Hierarchical Clustering).
- States in K Means cluster 0 have low health indices, low Per capita income and low GDP. Hence, we can notice that these states may be considered as underdeveloped states.Immediate actions are required by the government to develop the states in the health care and financial sectors.
- Government should introduce new strategies to improve health indices, per capita income and GDP
- Government should review the strategies which are being already executed in healthcare and financial departments and those strategies have to be reformed or discontinued based on in depth analysis