EXTENDED PROJECT

DATA MINING

SUBMITTED BY,

SHARA GEORGE VAIDIAN

QUESTION 1:

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 following rubric.

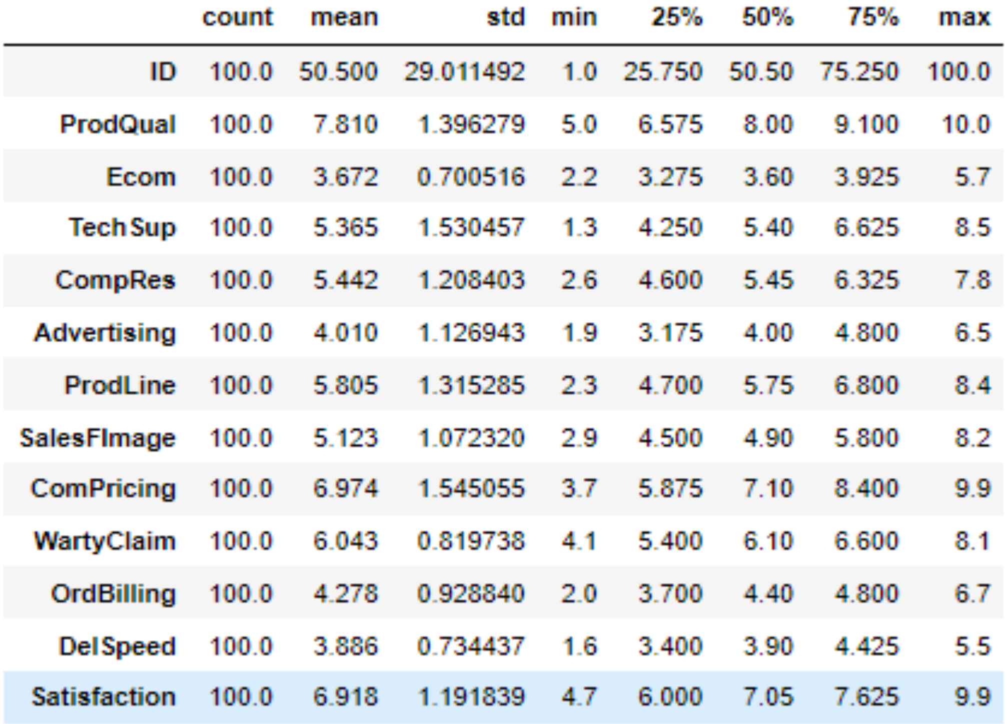
Note: This particular dataset contains the target variable satisfaction as well. Please do drop this variable before doing Principal Component Analysis.

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

Summary of dataset:

- There are 100 rows and 13 columns
- All data types are float data type
- There are no duplicates present in the dataset
- There are no null values present in the data set.

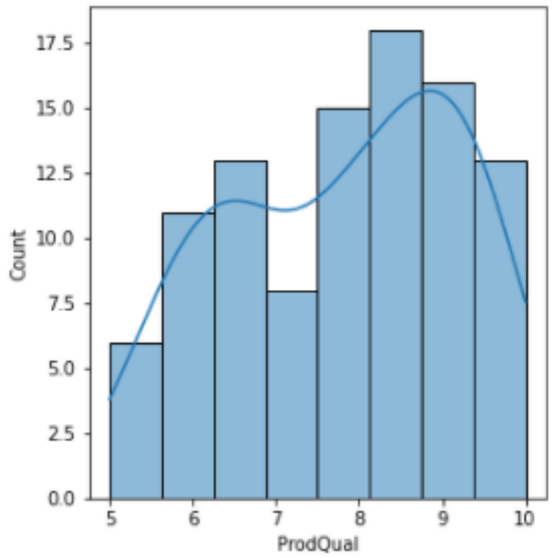
Statistical summary of the data is given below:

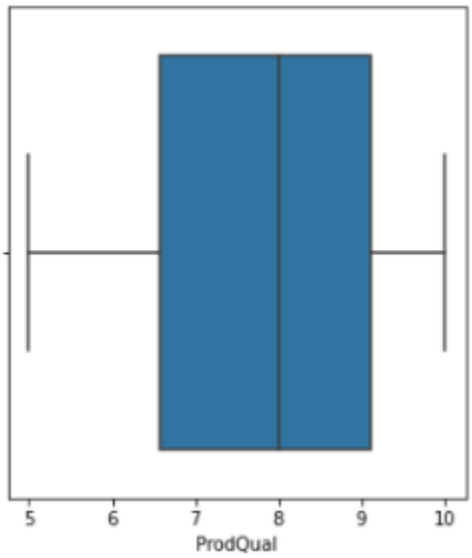


	count	mean	std	min	25%	50%	75%	max
ID	100.0	50.500	29.011492	1.0	25.750	50.50	75.250	100.0
ProdQual	100.0	7.810	1.396279	5.0	6.575	8.00	9.100	10.0
Ecom	100.0	3.672	0.700516	2.2	3.275	3.60	3.925	5.7
TechSup	100.0	5.365	1.530457	1.3	4.250	5.40	6.625	8.5
CompRes	100.0	5.442	1.208403	2.6	4.600	5.45	6.325	7.8
Advertising	100.0	4.010	1.126943	1.9	3.175	4.00	4.800	6.5
ProdLine	100.0	5.805	1.315285	2.3	4.700	5.75	6.800	8.4
SalesFImage	100.0	5.123	1.072320	2.9	4.500	4.90	5.800	8.2
ComPricing	100.0	6.974	1.545055	3.7	5.875	7.10	8.400	9.9
WartyClaim	100.0	6.043	0.819738	4.1	5.400	6.10	6.600	8.1
OrdBilling	100.0	4.278	0.928840	2.0	3.700	4.40	4.800	6.7
DelSpeed	100.0	3.886	0.734437	1.6	3.400	3.90	4.425	5.5
Satisfaction	100.0	6.918	1.191839	4.7	6.000	7.05	7.625	9.9

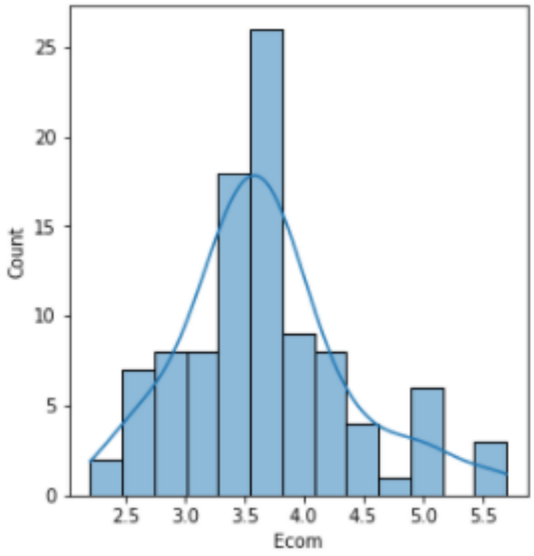
Univariate Analysis:

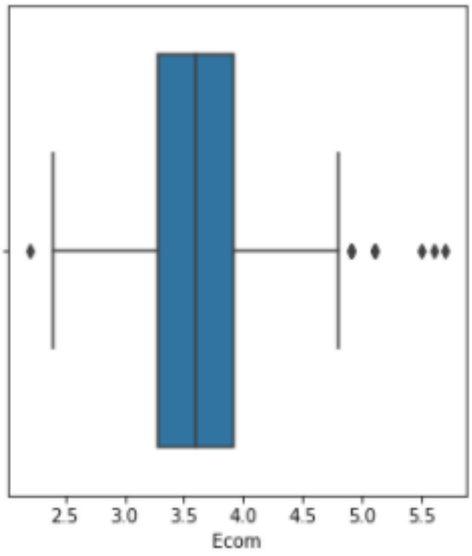
With the help of univariate analysis we can understand the distribution of data



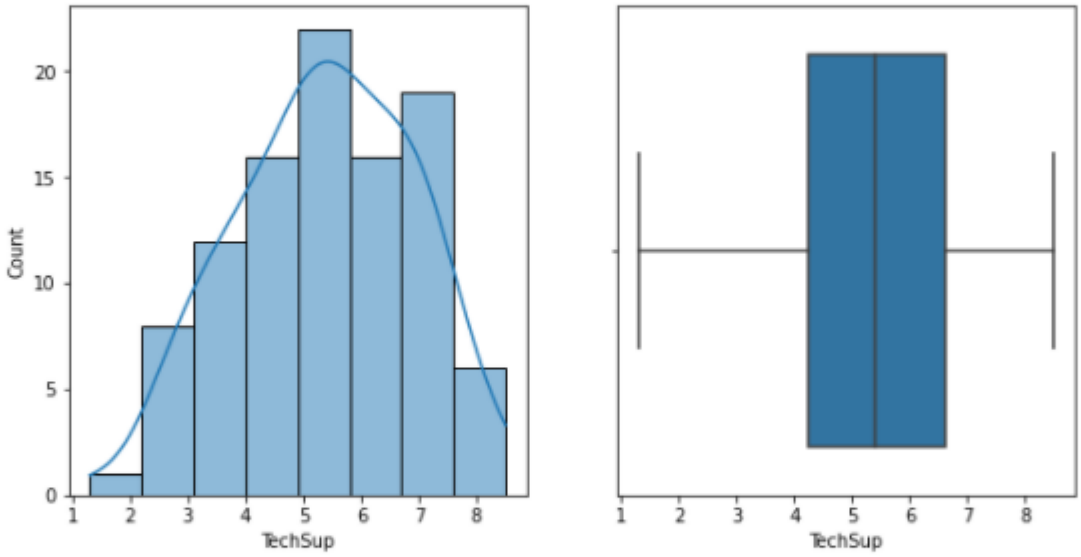


The product quality ranges between 6.5 and 9.1. There are no outliers present.

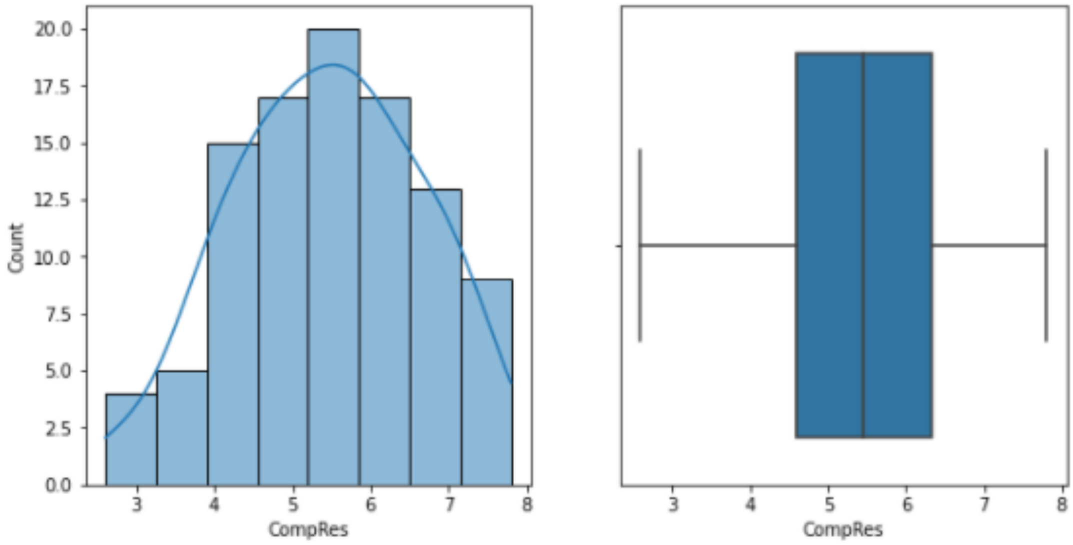




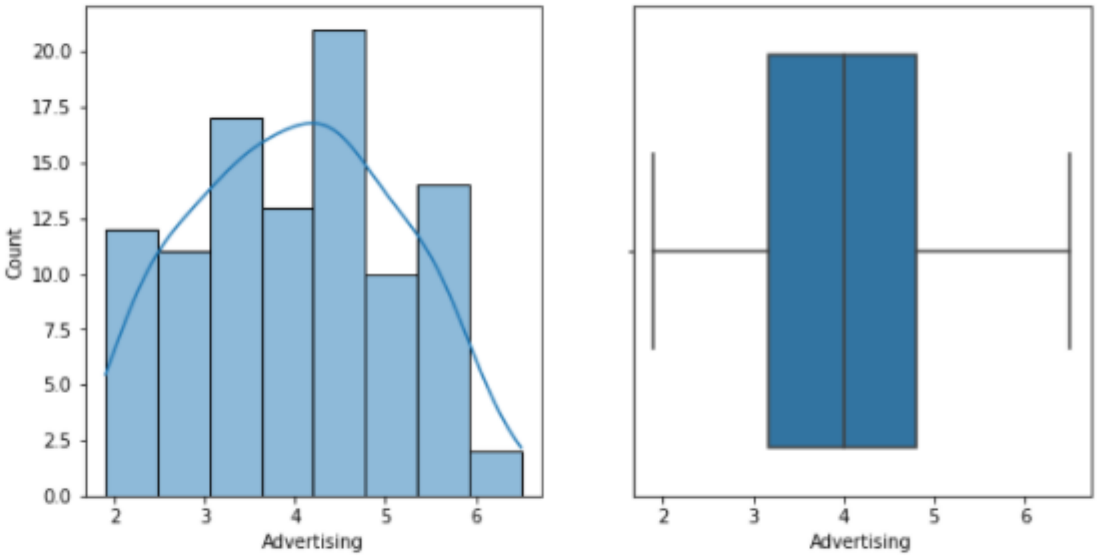
Here E-Commerce is normally distributed. It ranges between 3.25 and 4. There are a few outliers present.



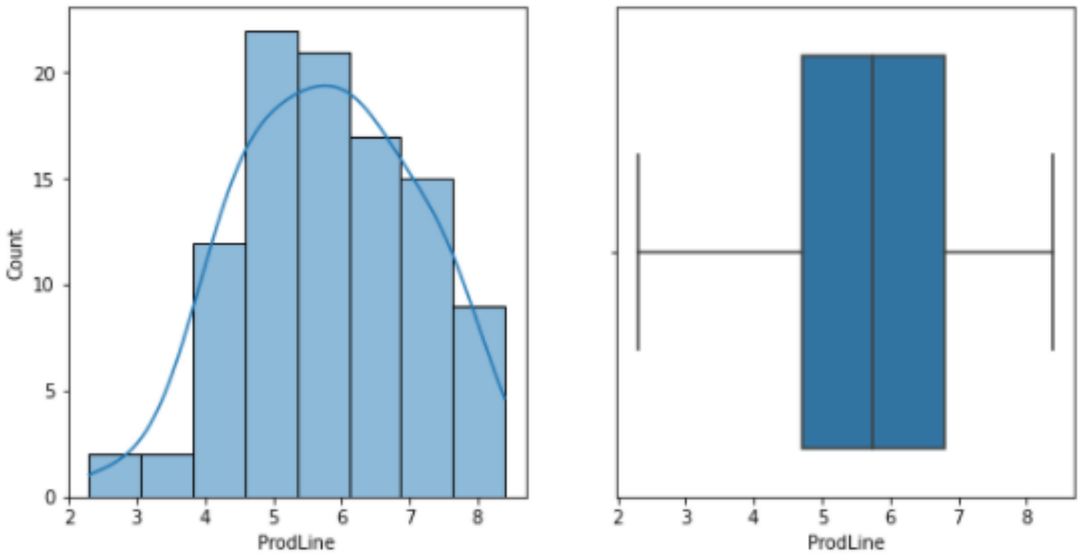
The data for Technical Support is normally distributed. It ranges between 4.1 and 6.9. There are no outliers present



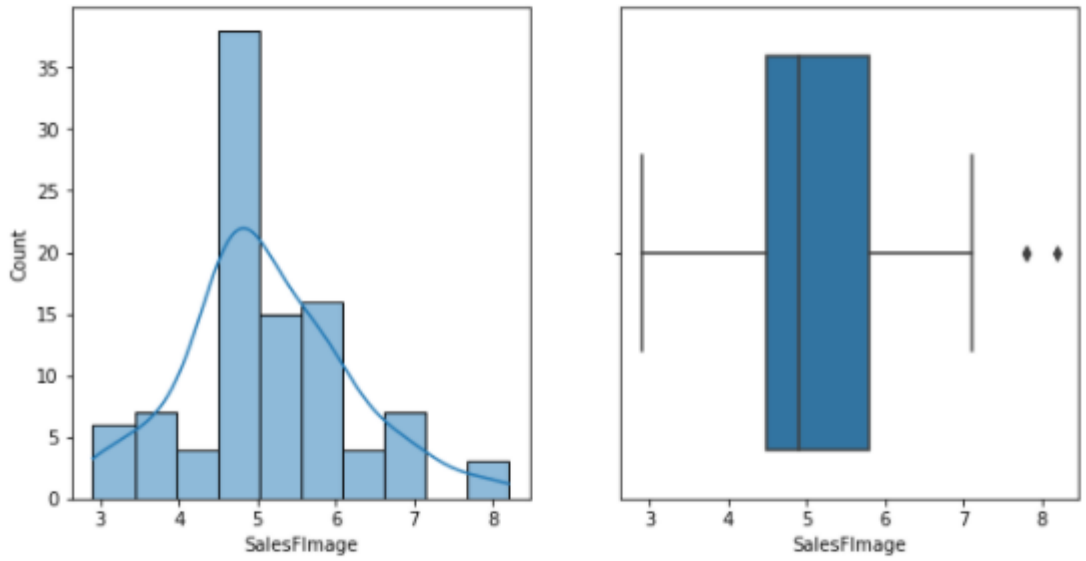
The data for Complaint Resolution is normally distributed. It ranges between 4.5 and 6.5. There are no outliers present



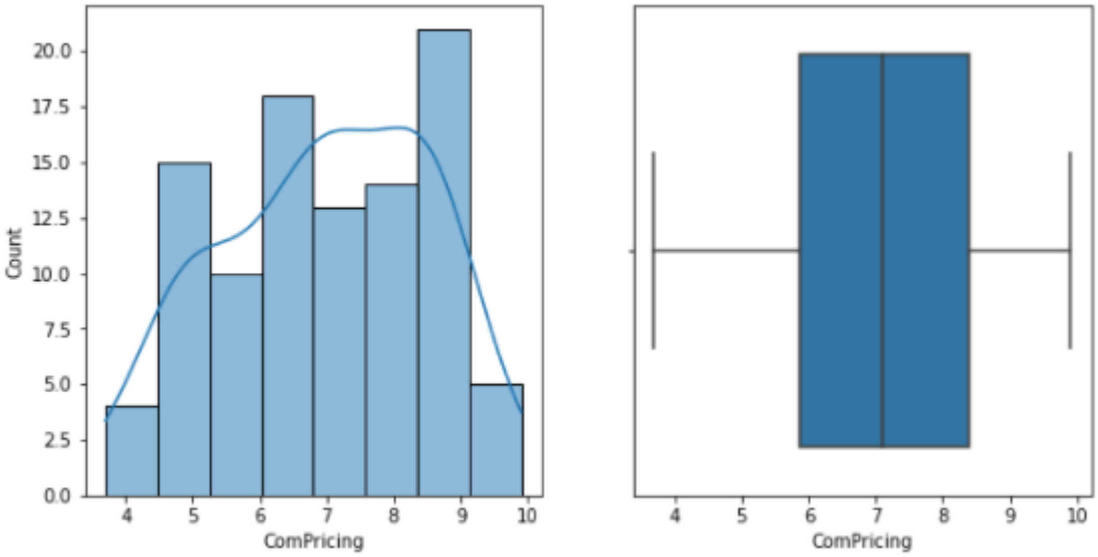
The data for advertising is normally distributed with data ranging from 3 and 5.



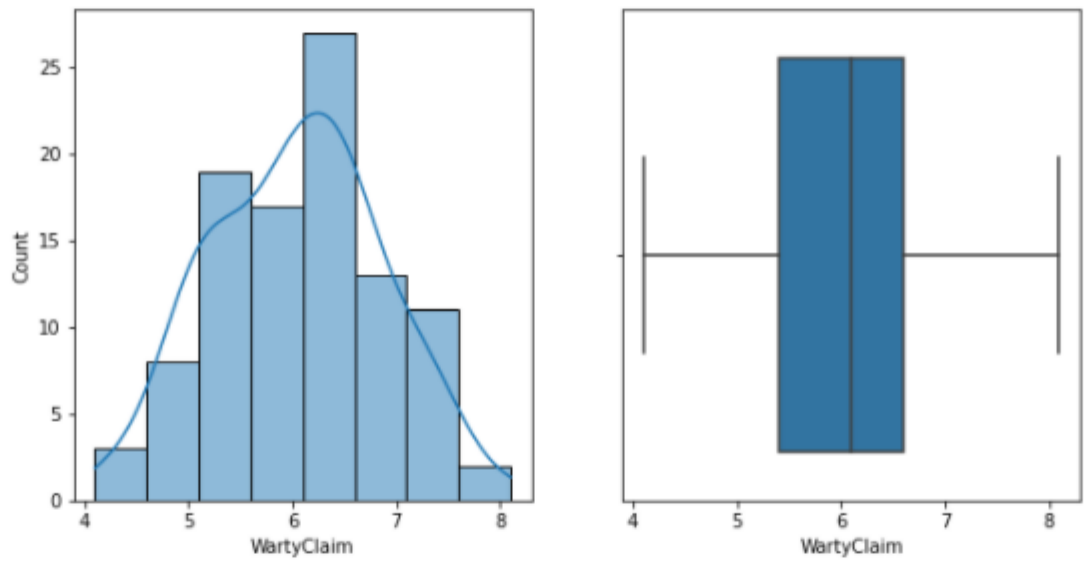
The data for Product Line is normally distributed with the data ranging from 4.5 and 7



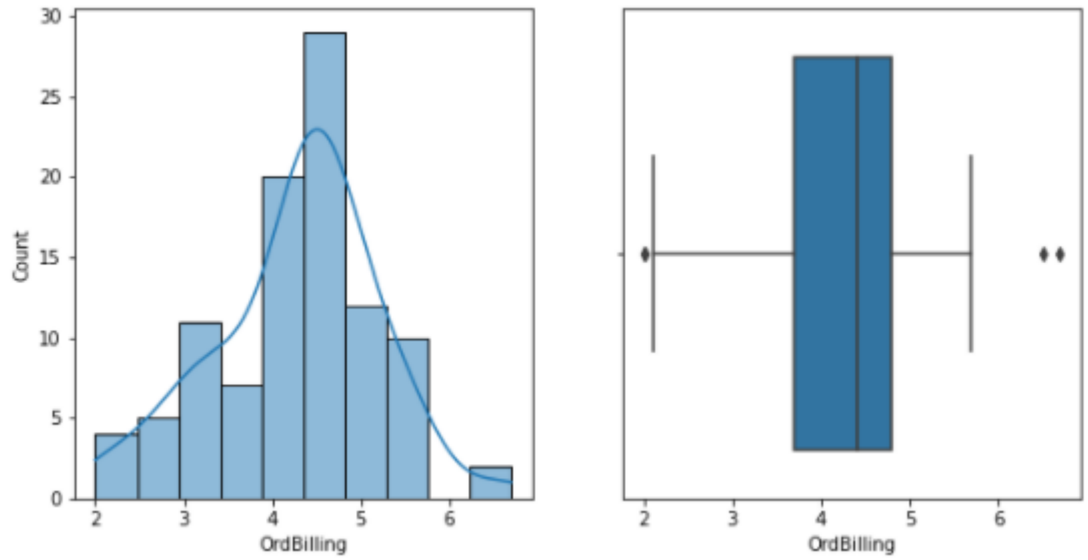
The data for Salesforce Image is normally distributed with data ranging from 4.5 and 6.



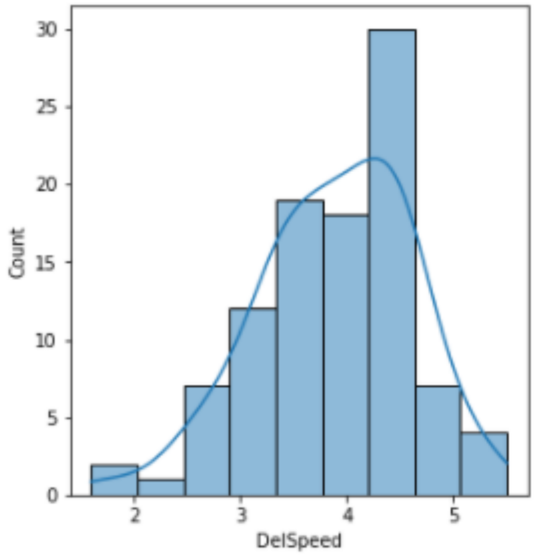
The data for Competitive Pricing ranges between 5.8 and 8.5

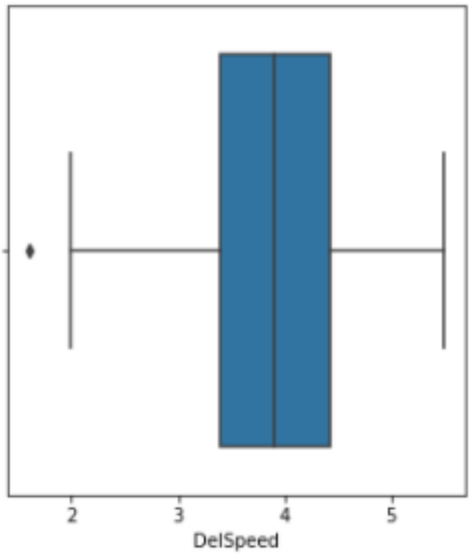


Data for Warranty and Claims ranges between 5.4 and 6.5



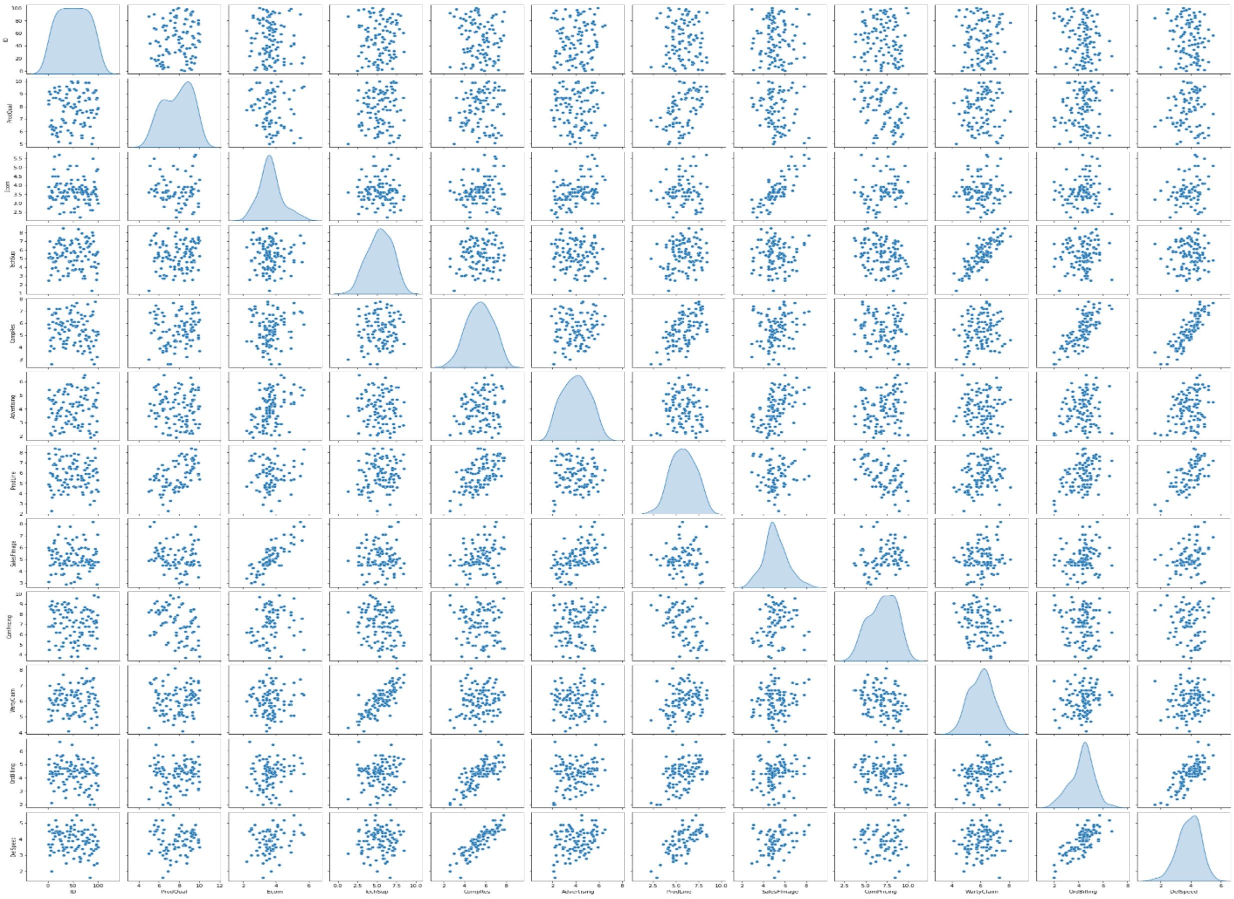
Data for order and Billing ranges between 3.8 and 4.9



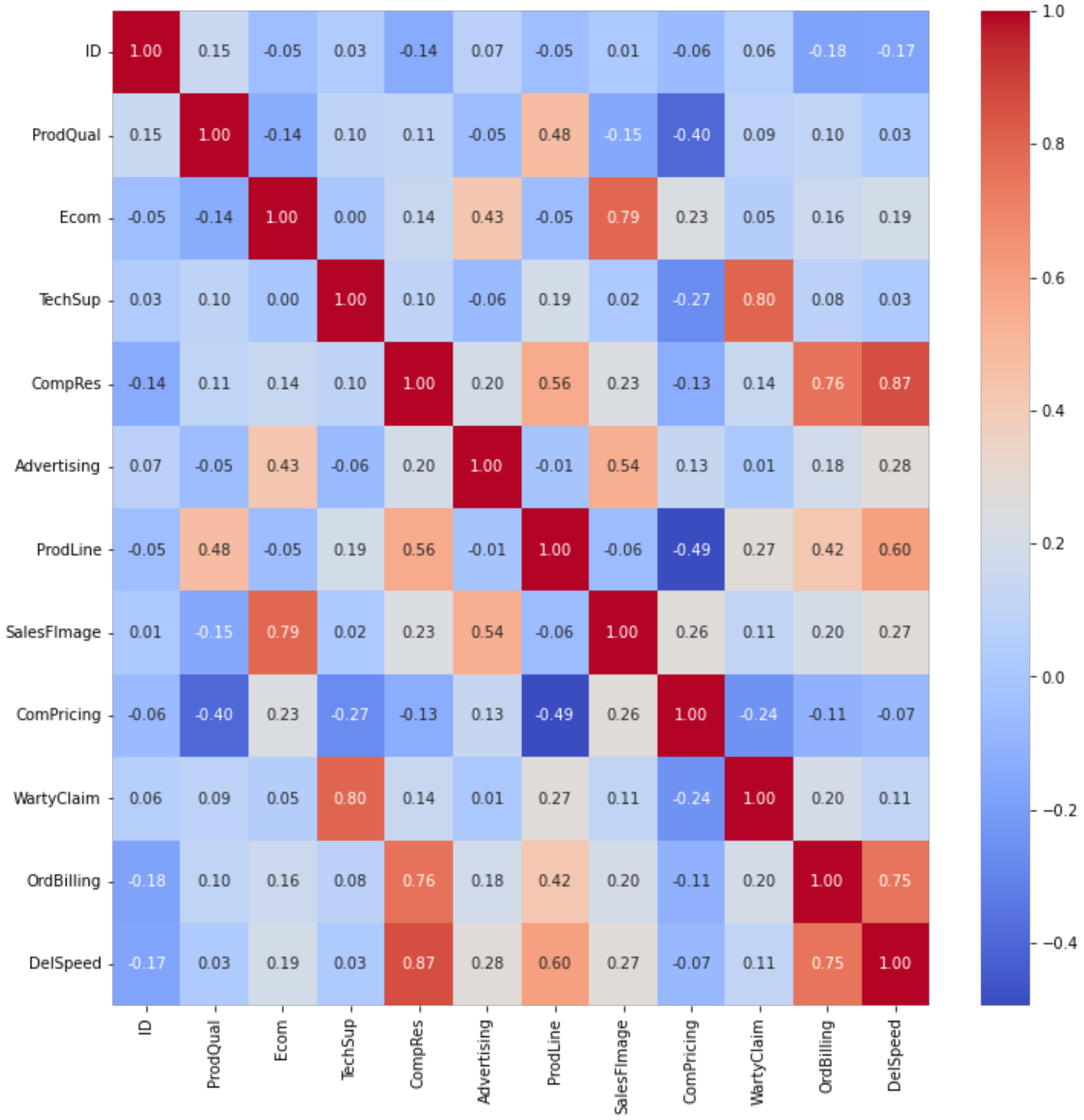


The data for Delivery speed ranges between 3.5 and 4.5.

Multivariate Analysis:



The heatmap shows the correlation between the data



	ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
ID	1.00	0.15	-0.05	0.03	-0.14	0.07	-0.05	0.01	-0.06	0.06	-0.18	-0.17
ProdQual	0.15	1.00	-0.14	0.10	0.11	-0.05	0.48	-0.15	-0.40	0.09	0.10	0.03
Ecom	-0.05	-0.14	1.00	0.00	0.14	0.43	-0.05	0.79	0.23	0.05	0.16	0.19
TechSup	0.03	0.10	0.00	1.00	0.10	-0.06	0.19	0.02	-0.27	0.80	0.08	0.03
CompRes	-0.14	0.11	0.14	0.10	1.00	0.20	0.56	0.23	-0.13	0.14	0.76	0.87
Advertising	0.07	-0.05	0.43	-0.06	0.20	1.00	-0.01	0.54	0.13	0.01	0.18	0.28
ProdLine	-0.05	0.48	-0.05	0.19	0.56	-0.01	1.00	-0.06	-0.49	0.27	0.42	0.60
SalesFImage	0.01	-0.15	0.79	0.02	0.23	0.54	-0.06	1.00	0.26	0.11	0.20	0.27
ComPricing	-0.06	-0.40	0.23	-0.27	-0.13	0.13	-0.49	0.26	1.00	-0.24	-0.11	-0.07
WartyClaim	0.06	0.09	0.05	0.80	0.14	0.01	0.27	0.11	-0.24	1.00	0.20	0.11
OrdBilling	-0.18	0.10	0.16	0.08	0.76	0.18	0.42	0.20	-0.11	0.20	1.00	0.75
DelSpeed	-0.17	0.03	0.19	0.03	0.87	0.28	0.60	0.27	-0.07	0.11	0.75	1.00

The variables with strong correlation are:

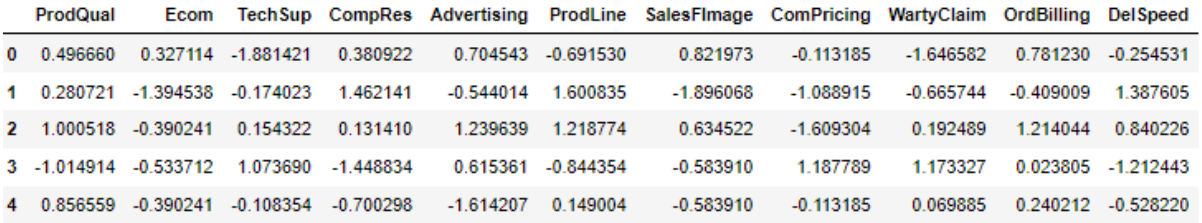
- SalesFImage and EComm (0.79)
- WartyClaim and TechSup (0.7971)
- CompRes and DelSpeed (0.87)
- DelSpeed and ProdLine (0.60)
- OrdBilling and CompRes (0.76)
- OrdBilling and DelSpeed (0.75)

Negative correlation exists between:

- ProdLine and CompPricing (-0.49)
- ProdQual and ComPricing (-0.40)

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

Here, the ID and satisfaction columns are dropped as ID is unique and satisfaction is a target variable. The Z-Score scaling is used to scale the data. Scaling helps in the standardization of data.



	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
0	0.496660	0.327114	-1.881421	0.380922	0.704543	-0.691530	0.821973	-0.113185	-1.646582	0.781230	-0.254531
1	0.280721	-1.394538	-0.174023	1.462141	-0.544014	1.600835	-1.896068	-1.088915	-0.665744	-0.409009	1.387605
2	1.000518	-0.390241	0.154322	0.131410	1.239639	1.218774	0.634522	-1.609304	0.192489	1.214044	0.840226
3	-1.014914	-0.533712	1.073690	-1.448834	0.615361	-0.844354	-0.583910	1.187789	1.173327	0.023805	-1.212443
4	0.856559	-0.390241	-0.108354	-0.700298	-1.614207	0.149004	-0.583910	-0.113185	0.069885	0.240212	-0.528220

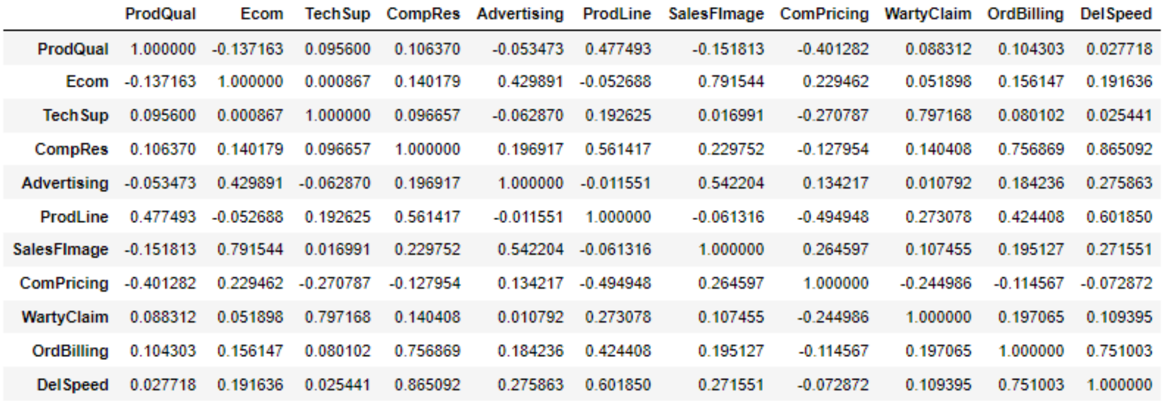
3) Comment on the comparison between covariance and the correlation matrix after scaling.

Covariance and correlation helps in measuring the relationship and dependency between two variables in the dataset. Scaling will not impact the covariance or correlation as scaling is used only for the standardization of data. Covariance helps to find whether two variables are directly or inversely proportional in the data set i.e covariance helps to find linear relationship between the variable. Covariance matrix of the data is shown as below, and it clearly shows the positive and negative linear relationship between the data.

```
[[ 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]
 [-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]
 [ 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]
 [ 1.07444445e-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]
 [-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]
 [ 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]
 [-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]
 [-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]
[ 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]
[ 1.05356640e-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]
[ 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]]
```

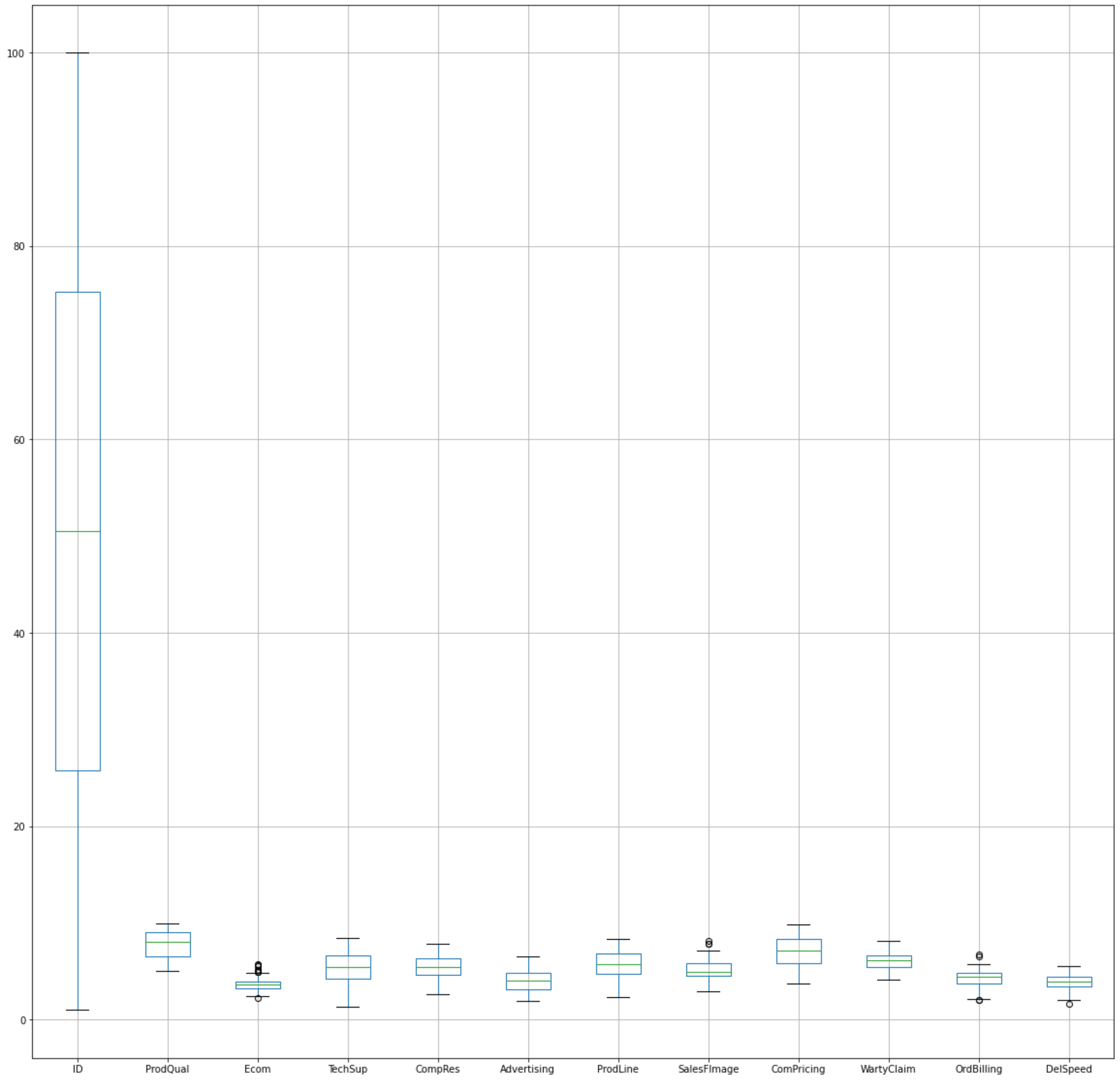
Correlation shows the how much is both the variables are correlated. The correlation matrix between and after the scaling is same. The below matrix shows the correlation between the variables and how strong the relationship between the variable.



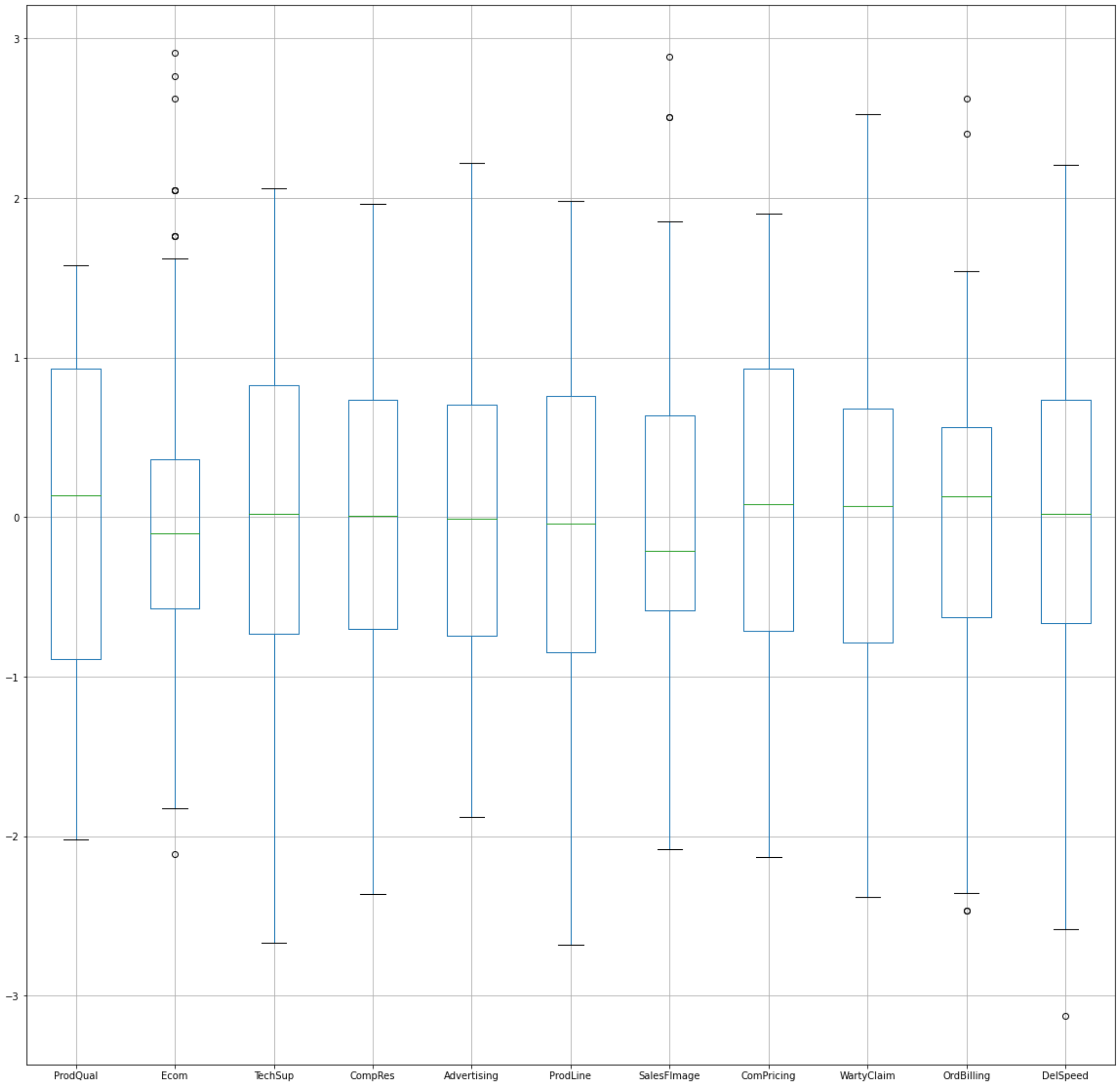
	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
ProdQual	1.000000	-0.137163	0.095600	0.106370	-0.053473	0.477493	-0.151813	-0.401282	0.088312	0.104303	0.027718
Ecom	-0.137163	1.000000	0.000867	0.140179	0.429891	-0.052688	0.791544	0.229462	0.051898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

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

Checking for outliers before scaling:



Checking for outliers after scaling:



There are no changes to outliers before and after scaling which means that scaling will not impact outliers. Ecommerce, salesforce image, order billing and delivery speed have few outliers.

5) Build the covariance matrix, eigenvalues and eigenvector.

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]
 [-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]
 [ 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]
 [ 1.07444445e-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]
 [-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]
 [ 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]
 [-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]
 [-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]
 [ 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]
 [ 1.05356640e-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]
 [ 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]]
Eigen Vectors:
[[ 0.13378962 -0.31349802  0.06227164  0.6431362  -0.2316662  -0.56456996
   0.19164132  0.18279209  0.06659717 -0.13547311  0.0313281 ]
 [ 0.16595278  0.44650918 -0.23524791  0.27238033 -0.42228844  0.26325703
   0.05962621  0.06233863  0.28155772  0.12202642 -0.54251104]
 [ 0.15769263 -0.23096734 -0.61095105 -0.19339314  0.02395667 -0.10876896
 -0.01719992 -0.05192956 -0.3881709  -0.46470964 -0.35929961]
 [ 0.47068359  0.01944394  0.21035078 -0.20632037 -0.02865743 -0.02815231
  -0.0084996  -0.36253352  0.53467243 -0.51339754  0.09324751]
 [ 0.18373495  0.36366471 -0.08809705  0.31789448  0.80387024 -0.20056937
  -0.06306962 -0.08118684  0.03715799  0.05347713 -0.15468169]
 [ 0.38676517 -0.28478056  0.11627864  0.20290226 -0.11667416  0.09819533
  -0.60814755 -0.38507778 -0.23479794  0.3332071  -0.08415534]
 [ 0.2036696   0.47069599 -0.2413421   0.22217722 -0.20437283  0.10497225
   0.00143735 -0.08469869 -0.35341191 -0.16910665  0.64489911]
 [-0.15168864  0.4134565   0.05304529 -0.33354348 -0.24892601 -0.70973595
  -0.30824887 -0.10295751 -0.04518224  0.09883227 -0.09414389]
 [ 0.21293363 -0.19167191 -0.59856398 -0.18530205  0.03292706 -0.13983966
  -0.03064024  0.12893245  0.43534752  0.4435404   0.31756604]
 [ 0.43721774  0.02639905  0.16892981 -0.23685365 -0.02675377 -0.11947974
   0.65931989 -0.19415064 -0.30386545  0.36601754 -0.09907265]
 [ 0.47308914  0.07305172  0.23262477 -0.1973299   0.03543294  0.02979992
  -0.23423927  0.77563222 -0.12010386 -0.06539059 -0.02188514]]
```

Eigen values:

```
[3.4615872  2.57666335 1.70805705 1.09753137 0.61557989 0.55745836
 0.40557389 0.09942123 0.13418341 0.249446   0.20560936]
```

6) Write the explicit form of the first PC (in terms of Eigen Vectors)

The explicit form of the first PC:

-0.13 * ProdQual + -0.17 * Ecom + -0.16 * TechSup + -0.47 * CompRes + -0.18 * Advertising + -0.39 * ProdLine + -0.2 * SalesFImage + 0.15 * ComPricing + -0.21 * WartyClaim + -0.44 * OrdBilling + -0.47 * DelSpeed +

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.

```
Array ([ 31.1542848 ,  54.34425491,  69.71676832,  79.59455066,
        85.1347697 ,  90.15189496,  93.80205993,  96.04707397,
       97.89755822,  99.10520892, 100.  ])
```

From the above array, we can clearly find that total sums to 100.

- Check for cumulative variance upto 90%. Check the corresponding associated
- The incremental value between the components should not be less than 5%
- We select 5 as the principal components as after 6 the incremental value between thecomponents is more than 5%.
- So we select 5 principal components for this analysis.

PCA Components:

```
([[-0.13378962,  -0.16595278,  -0.15769263,  -0.47068359,  -0.18373495,-0.38676517,  -0.2036696  ,
0.15168864, -0.21293363, -0.43721774,-0.47308914],[-0.31349802, 0.44650918, -0.23096734, 0.01944394,
0.36366471,-0.28478056, 0.47069599, 0.4134565 , -0.19167191, 0.02639905,0.07305172],[ 0.06227164, -
0.23524791, -0.61095105, 0.21035078, -0.08809705,0.11627864, -0.2413421 , 0.05304529, -0.59856398,
0.16892981,0.23262477],[ 0.6431362 , 0.27238033, -0.19339314, -0.20632037, 0.31789448,0.20290226,
0.22217722, -0.33354348, -0.18530205, -0.23685365,-0.1973299 ],[ 0.2316662 , 0.42228844, -0.02395667,
0.02865743, -0.80387024,0.11667416, 0.20437283, 0.24892601, -0.03292706, 0.02675377,-0.03543294]])
```

- The first component explains 31.15% variance
- The first two components explains about 54.43% variance
- The first three components explains about 69.71% variance
- The first four components explains about 79.59% variance
- The first five components explains about 85.13% variance

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

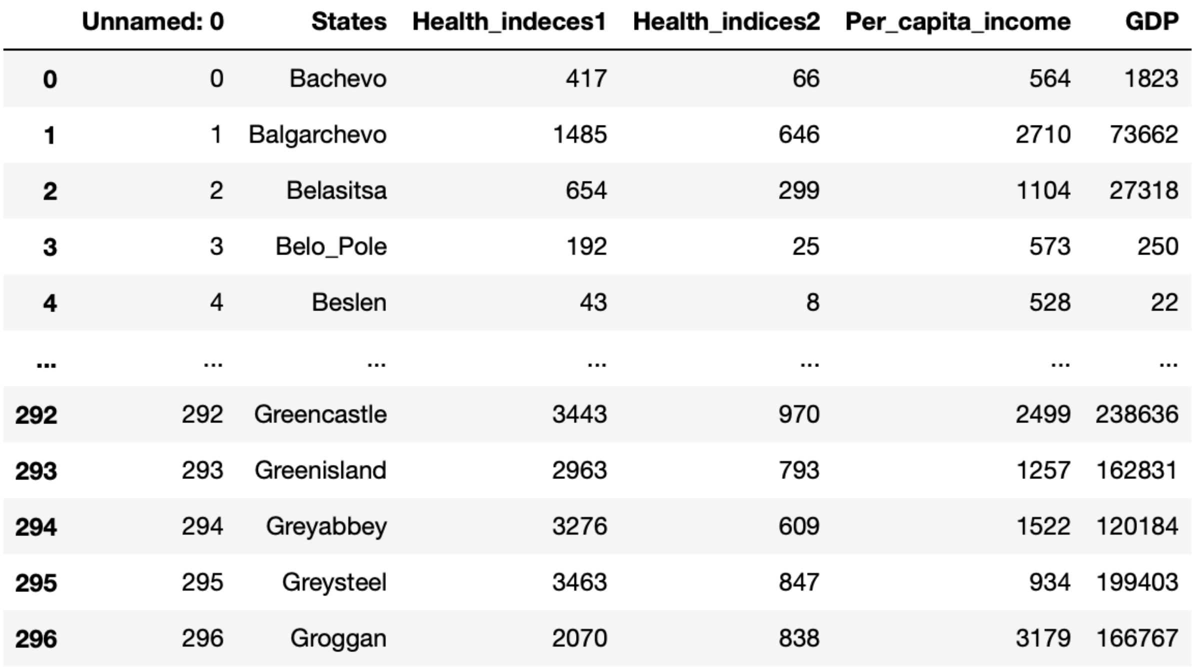
The case study is based on the various hair salon products used for segmentation of market. From the univariate analysis and multivariate analysis we understood the relationship between the variables and distribution of data for each variable. PCA helps us to reduce multicolinearity and helps us to analyze the 5 factors for the next step of analysis rather than analysing all the 11 variables.These five components can be combined together to make a common portfolio to formulate the strategies for the segmentation.

QUESTION 2:

PROBLEM STATEMENT: 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, to provide these groups to the government so that appropriate measures can be taken to escalate their Health and Economic conditions.

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

The dataset:

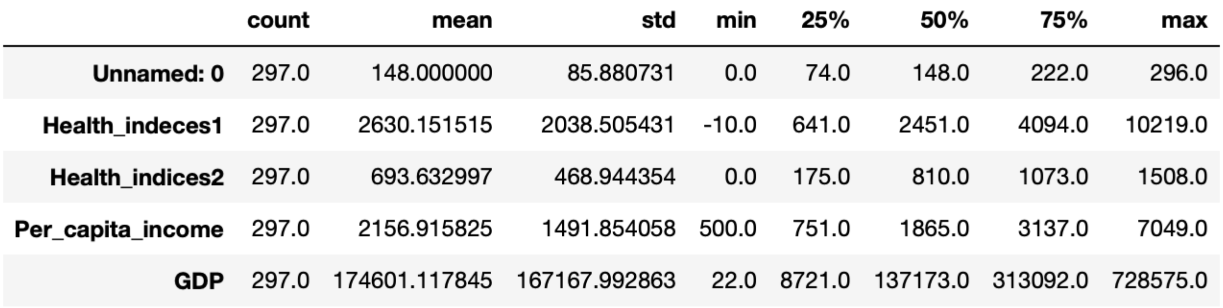


	Unnamed: 0	States	Health_indeces1	Health_indices2	Per_capita_income	GDP
0	0	Bachevo	417	66	564	1823
1	1	Balgarchevo	1485	646	2710	73662
2	2	Belasitsa	654	299	1104	27318
3	3	Belo_Pole	192	25	573	250
4	4	Beslen	43	8	528	22
...	...	...	...	...	...	...
292	292	Greencastle	3443	970	2499	238636
293	293	Greenisland	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

The data set has 297 rows and 6 columns.

- It has 5 integer and 1 object data type.
- There are no null values
- There are no duplicate values

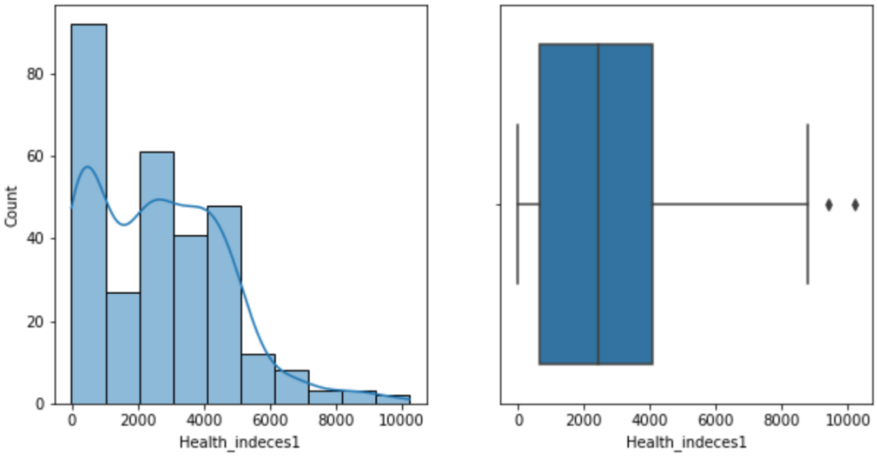
Statistical summary of the data is given below:



	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	297.0	148.000000	85.880731	0.0	74.0	148.0	222.0	296.0
Health_indeces1	297.0	2630.151515	2038.505431	-10.0	641.0	2451.0	4094.0	10219.0
Health_indices2	297.0	693.632997	468.944354	0.0	175.0	810.0	1073.0	1508.0
Per_capita_income	297.0	2156.915825	1491.854058	500.0	751.0	1865.0	3137.0	7049.0
GDP	297.0	174601.117845	167167.992863	22.0	8721.0	137173.0	313092.0	728575.0

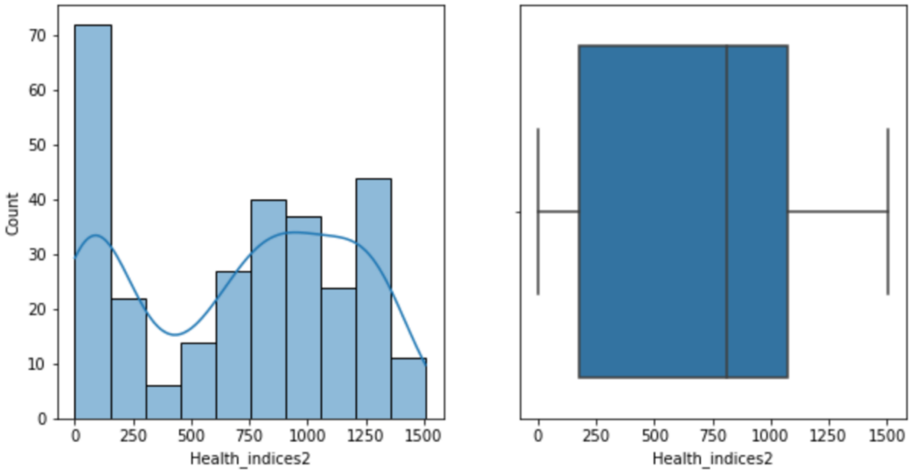
Univariate analysis:

Health_indeces1

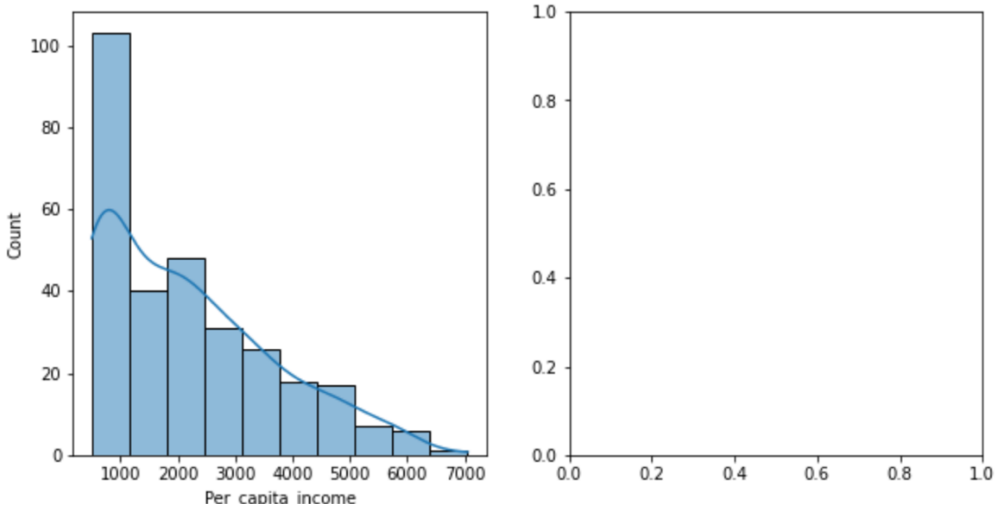


Outliers are present in Health_indeces1

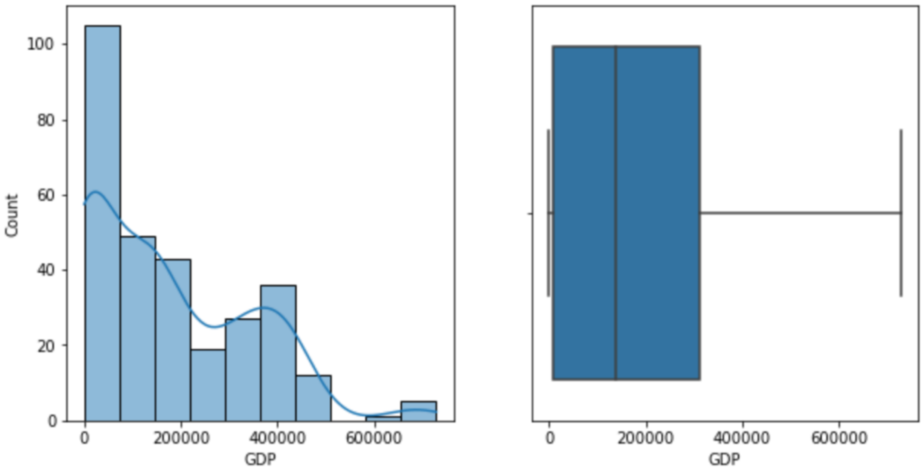
Health_indices2:



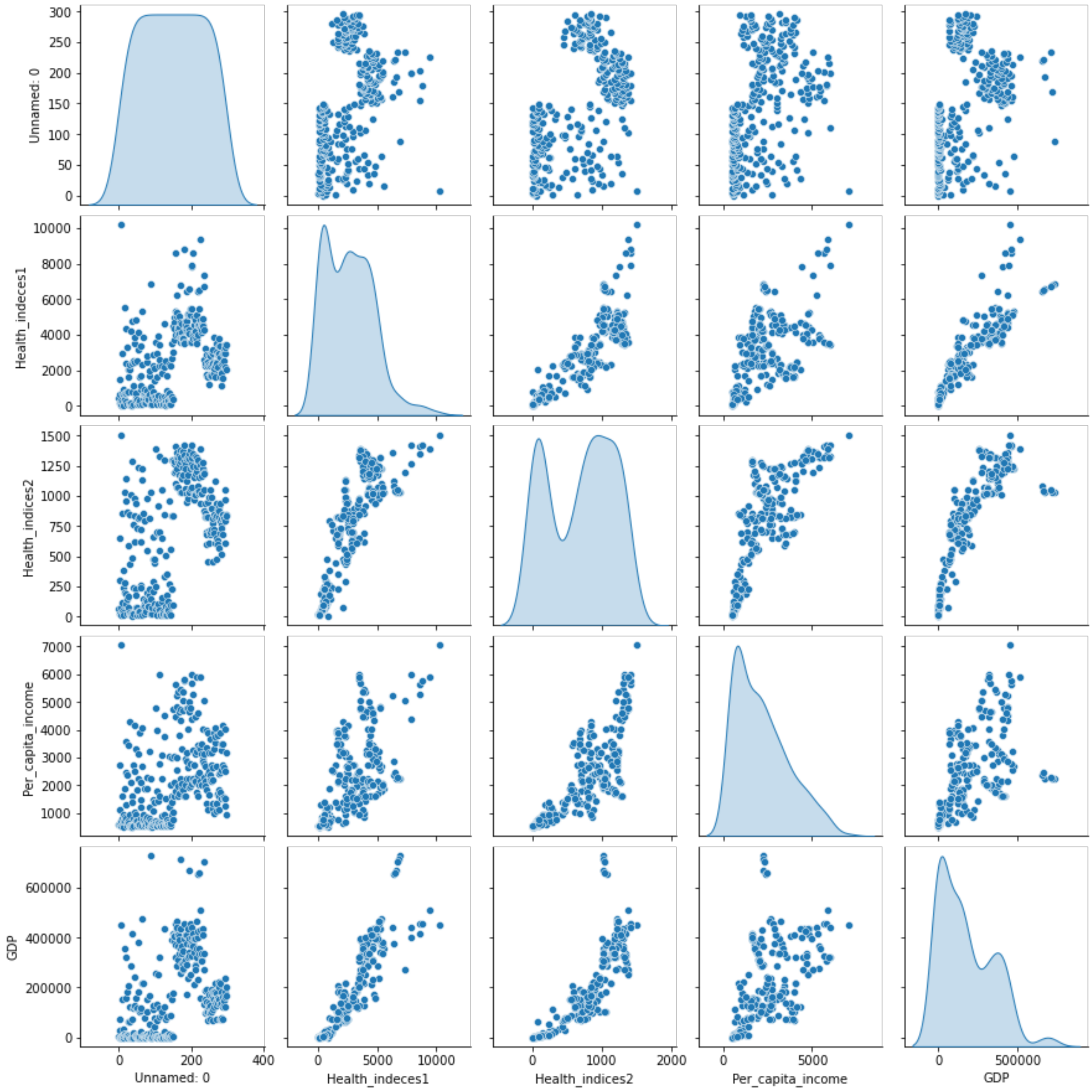
Per_capita Income:



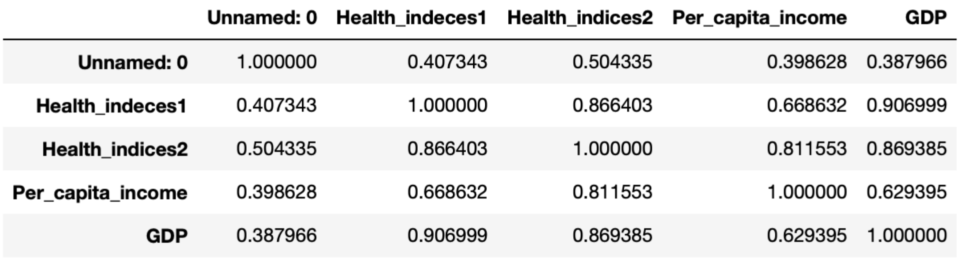
GDP:



Multivariate Analysis:

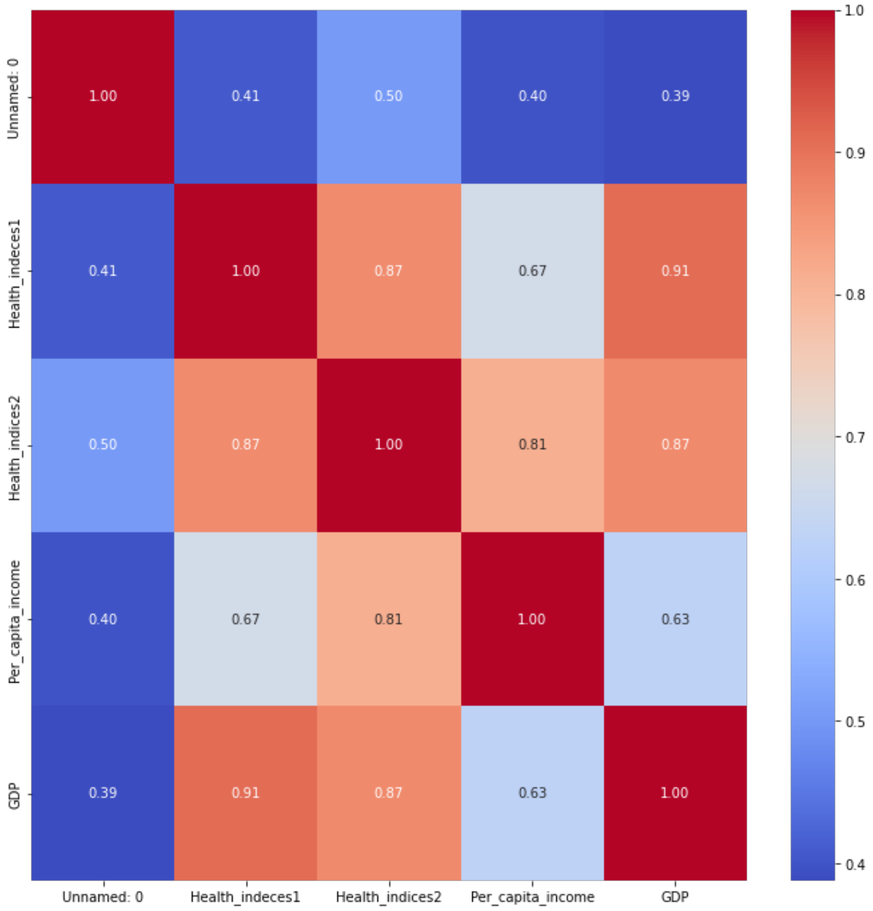


Correlation:



	Unnamed: 0	Health_indeces1	Health_indices2	Per_capita_income	GDP
Unnamed: 0	1.000000	0.407343	0.504335	0.398628	0.387966
Health_indeces1	0.407343	1.000000	0.866403	0.668632	0.906999
Health_indices2	0.504335	0.866403	1.000000	0.811553	0.869385
Per_capita_income	0.398628	0.668632	0.811553	1.000000	0.629395
GDP	0.387966	0.906999	0.869385	0.629395	1.000000

The heat map shows the correlation:



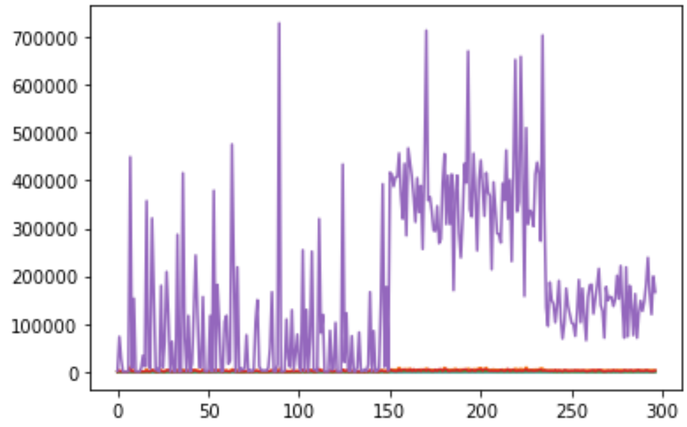
High positive correlation exists between Health index 1 and health index 2 (0.87).

High negative correlation exists between Health index 2 and per capita income (-0.72).

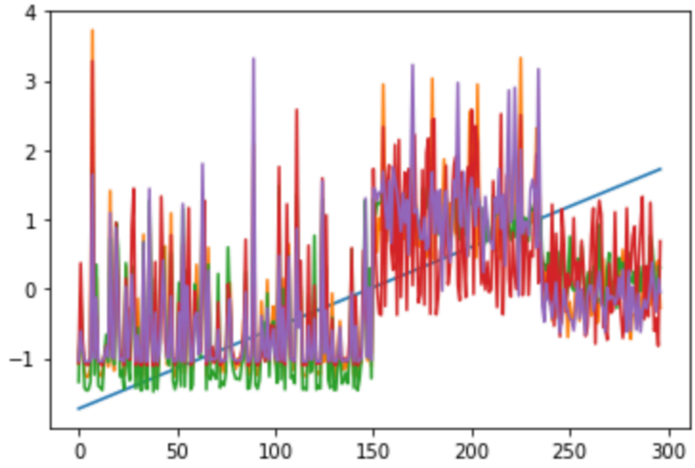
2. Do you think scaling is necessary for clustering in this case? Justify

Scaling is necessary as the values of the variables are different. spending, advance_payments are in different values and this may get higher weightage. With scaling we can have all the values in the relatively same range.

Before scaling:

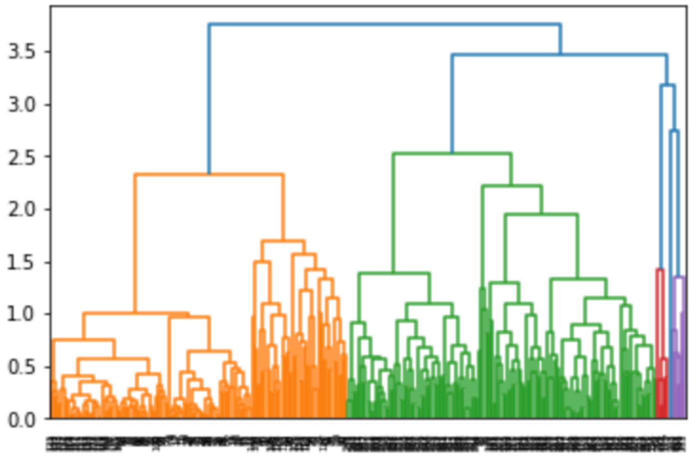


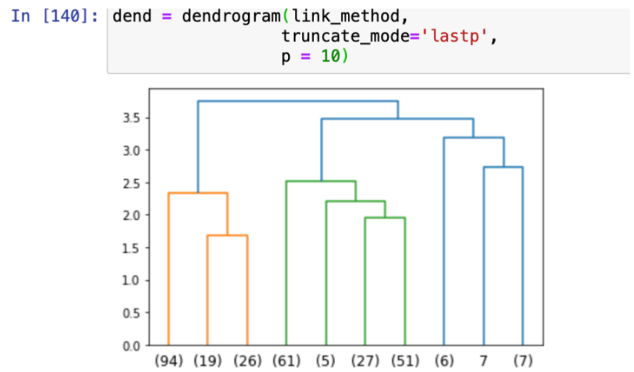
After scaling:



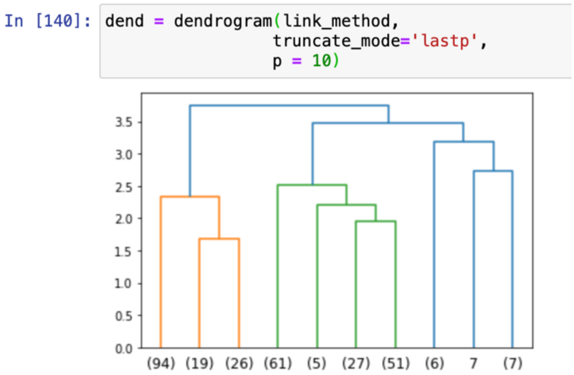
3. Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.

Dendogram (Average Link Method)

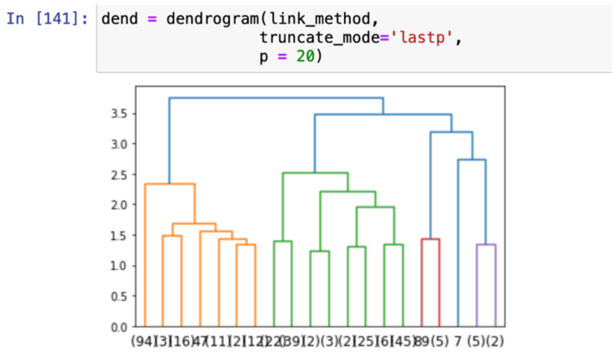




```
In [140]: dend = dendrogram(link_method,
                            truncate_mode='lastp',
                            p = 10)
```

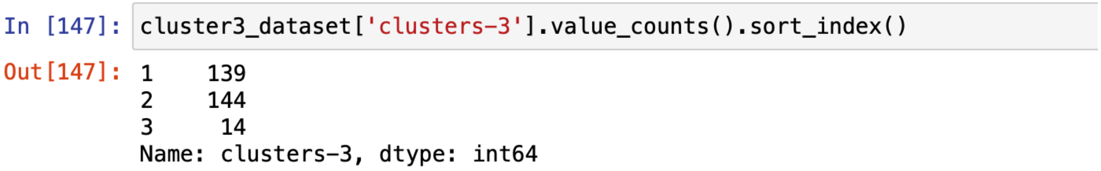


```
In [140]: dend = dendrogram(link_method,
                            truncate_mode='lastp',
                            p = 10)
```

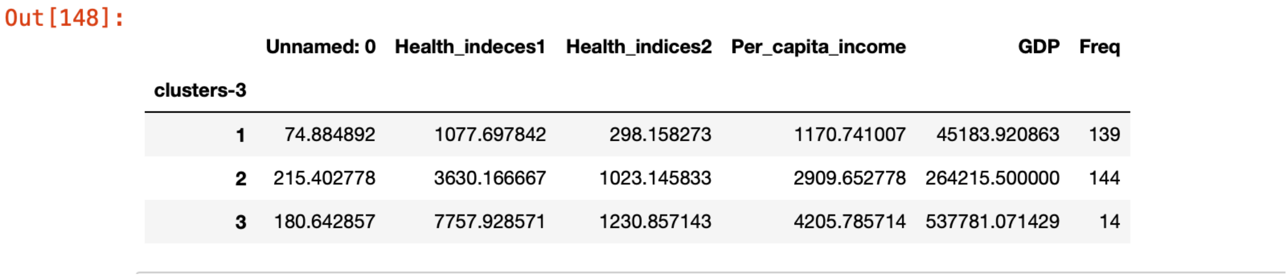


```
In [141]: dend = dendrogram(link_method,
                            truncate_mode='lastp',
                            p = 20)
```

Three clusters were created:



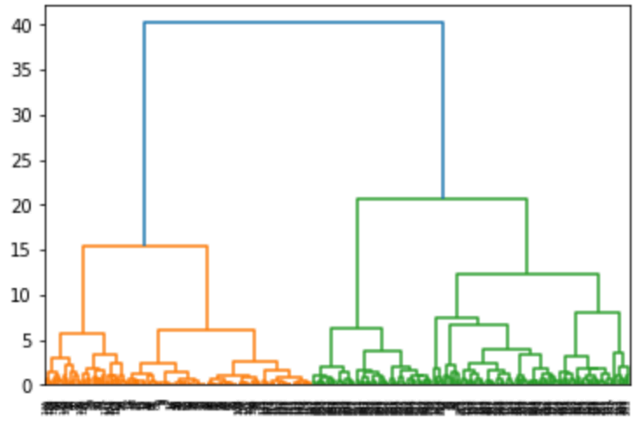
```
In [147]: cluster3_dataset['clusters-3'].value_counts().sort_index()

Out[147]: 1    139
          2    144
          3     14
          Name: clusters-3, dtype: int64
```

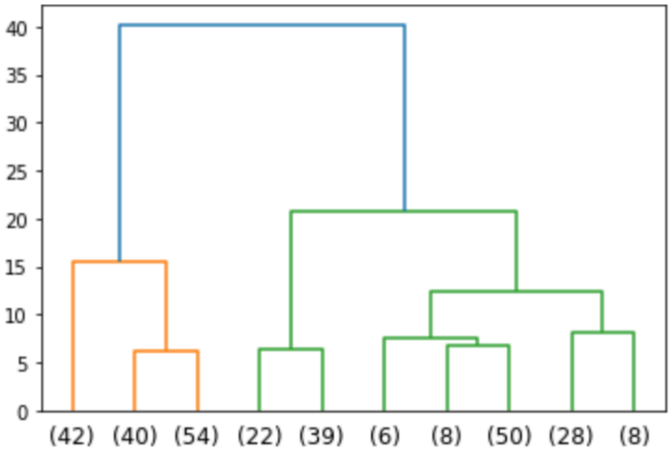


Out[148]:

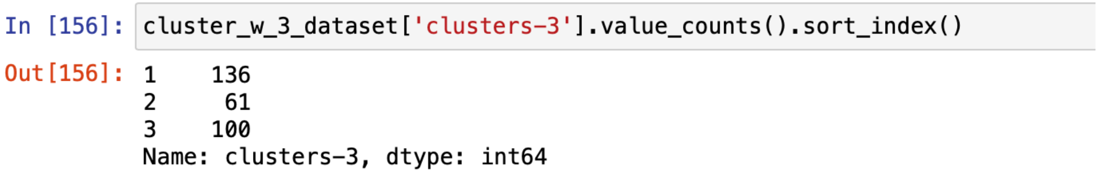
	Unnamed: 0	Health_indeces1	Health_indices2	Per_capita_income	GDP	Freq
clusters-3						
1	74.884892	1077.697842	298.158273	1170.741007	45183.920863	139
2	215.402778	3630.166667	1023.145833	2909.652778	264215.500000	144
3	180.642857	7757.928571	1230.857143	4205.785714	537781.071429	14

Dendogram – Ward link

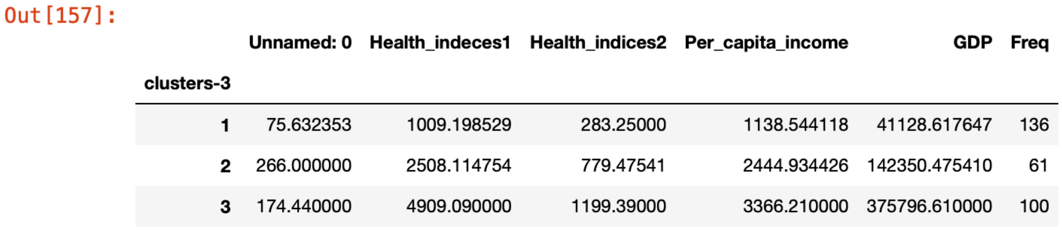




Three clusters were created:



```
In [156]: cluster_w_3_dataset['clusters-3'].value_counts().sort_index()
Out[156]: 1    136
          2     61
          3    100
          Name: clusters-3, dtype: int64
```



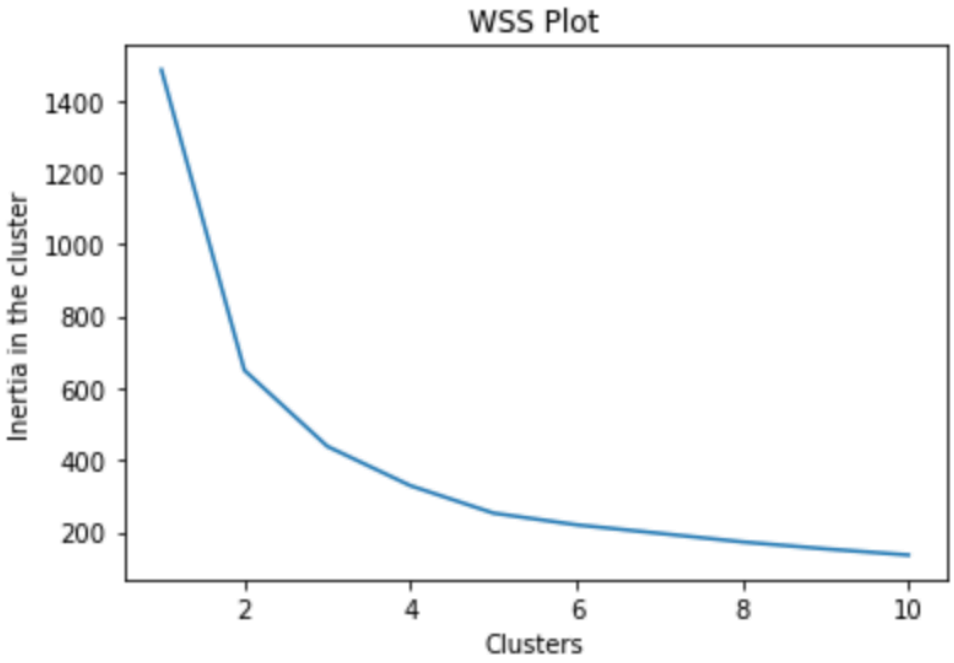
Out[157]:

clusters-3	Unnamed: 0	Health_indeces1	Health_indices2	Per_capita_income	GDP	Freq
1	75.632353	1009.198529	283.25000	1138.544118	41128.617647	136
2	266.000000	2508.114754	779.47541	2444.934426	142350.475410	61
3	174.440000	4909.090000	1199.39000	3366.210000	375796.610000	100

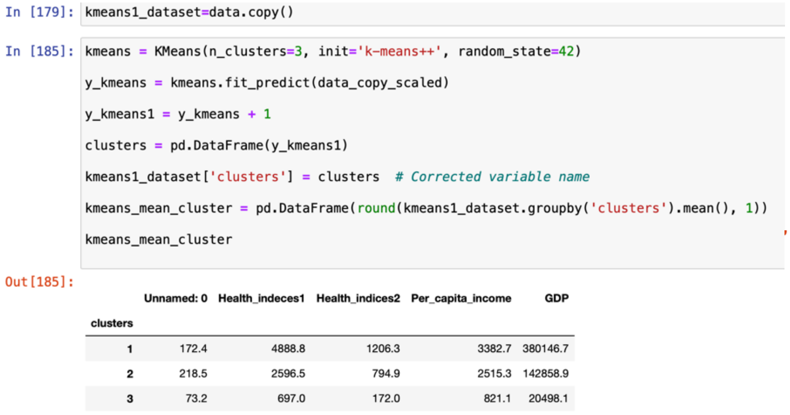
4. Clustering: Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and find the silhouette score.

Insights:

From SC Score, the number of optimal clusters could be 3 or 4. But we are going with 3 clusters as it can be seen in the graph that the curve gets flat after 3.

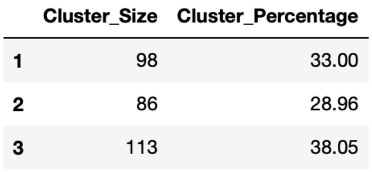


K means clustering and cluster information:

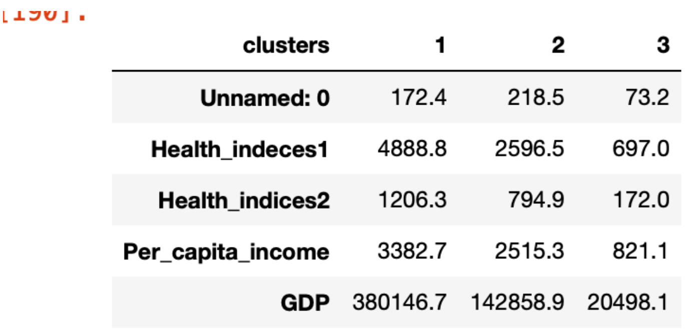


```
In [179]: kmeans1_dataset=data.copy()

In [185]: kmeans = KMeans(n_clusters=3, init='k-means++', random_state=42)

          y_kmeans = kmeans.fit_predict(data_copy_scaled)

          y_kmeans1 = y_kmeans + 1

          clusters = pd.DataFrame(y_kmeans1)

          kmeans1_dataset['clusters'] = clusters  # Corrected variable name

          kmeans_mean_cluster = pd.DataFrame(round(kmeans1_dataset.groupby('clusters').mean(), 1))

          kmeans_mean_cluster

Out[185]:
```

	Unnamed: 0	Health_indeces1	Health_indices2	Per_capita_income	GDP
clusters					
1	172.4	4888.8	1206.3	3382.7	380146.7
2	218.5	2596.5	794.9	2515.3	142858.9
3	73.2	697.0	172.0	821.1	20498.1



	Cluster_Size	Cluster_Percentage
1	98	33.00
2	86	28.96
3	113	38.05



[190]:

clusters	1	2	3
Unnamed: 0	172.4	218.5	73.2
Health_indeces1	4888.8	2596.5	697.0
Health_indices2	1206.3	794.9	172.0
Per_capita_income	3382.7	2515.3	821.1
GDP	380146.7	142858.9	20498.1

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.

We got 3 as the optimum number of clusters.

Cluster Group Profiles-

Cluster 1: High GDP per capita Areas

- These are the areas which have the highest growth rate.
- The health and economic conditions in these ares excellent.
- Per capita income in these areas is very high.

Cluster 2: Low GDP per capita Areas

- These are the areas which have very low growth rate.
- The health and economic conditions are not good in these areas.
- Per capita income in these areas is very low.

Cluster 3: Medium GDP per capita Areas

- These are the areas which have an average growth rate.
- The health and economic conditions in these areas are adequate.
- Per capita income in these areas is average.

Recommendations for each cluster profile.

Main features that affect the Health and Economic conditions are workforce and productivity. Higher these attributes higher is the GDP per capita and thus higher the Health and Economic conditions