2023

Data Mining Project

PCA & CLUSTERING

GUDLA SAI SRINIVAS

Table of Contents

Table of Figures

List of Tables

Problem Statement 1:

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.

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

Introduction:

The given dataset contains data of hair products used in Hair Salon. Need to do dimension reductions for this dataset using PCA.

Before working on data, let us look at the below basic details of data.

1. Head

	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

Table 1. Head_PCA data

2. The data contains 13 variables of different market segmentation and 100 IDs for each variable.
3. Information about the datatype and null values given below

Data columns	Rows	Null info	Data Type
ID	100	non-null	int64
ProdQual	100	non-null	float64
Ecom	100	non-null	float64
TechSup	100	non-null	float64
CompRes	100	non-null	float64
Advertising	100	non-null	float64
ProdLine	100	non-null	float64
SalesFImage	100	non-null	float64
ComPricing	100	non-null	float64
WartyClaim	100	non-null	float64
OrdBilling	100	non-null	float64
DelSpeed	100	non-null	float64
Satisfaction	100	non-null	float64

Table 2. Data_information

Data contains no 1 integer data type and 12 float data types.

The variable names give the Expansion as mentioned below.

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 Claim
OrdBilling	Order & Billing
DelSpeed	Delivery Speed
Satisfaction	Customer Satisfaction

Table 3. Data Dictionary

4. There are no missing values in the data.

Variable	Null values
ID	0
ProdQual	0
Ecom	0
TechSup	0
CompRes	0
Advertising	0
ProdLine	0
SalesFImage	0
ComPricing	0
WartyClaim	0
OrdBilling	0
DelSpeed	0
Satisfaction	0

Table 4. Missing Values information

5. Checking summary of the data

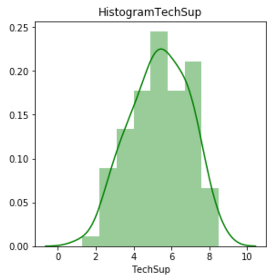
	ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
count	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.00000	100.000000	100.000000	100.00000	100.000000	100.000000
mean	50.500000	7.810000	3.672000	5.365000	5.442000	4.010000	5.805000	5.12300	6.974000	6.043000	4.27800	3.886000	6.918000
std	29.011492	1.396279	0.700516	1.530457	1.208403	1.126943	1.315285	1.07232	1.545055	0.819738	0.92884	0.734437	1.191839
min	1.000000	5.000000	2.200000	1.300000	2.600000	1.900000	2.300000	2.90000	3.700000	4.100000	2.00000	1.600000	4.700000
25%	25.750000	6.575000	3.275000	4.250000	4.600000	3.175000	4.700000	4.50000	5.875000	5.400000	3.70000	3.400000	6.000000
50%	50.500000	8.000000	3.600000	5.400000	5.450000	4.000000	5.750000	4.90000	7.100000	6.100000	4.40000	3.900000	7.050000
75%	75.250000	9.100000	3.925000	6.625000	6.325000	4.800000	6.800000	5.80000	8.400000	6.600000	4.80000	4.425000	7.625000
max	100.000000	10.000000	5.700000	8.500000	7.800000	6.500000	8.400000	8.20000	9.900000	8.100000	6.70000	5.500000	9.900000

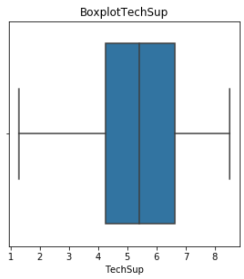
Table 5. Summary of the Data

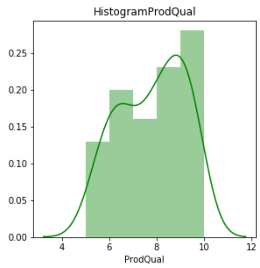
The average satisfaction of the hair products from different market segmentation is **6.91.** We can understand more about the hair products of different segments while doing EDA.

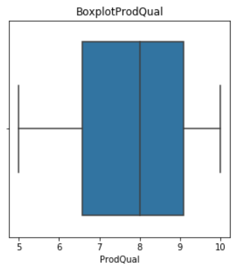
1. PCA: Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented.

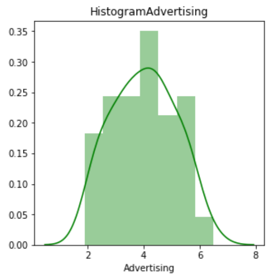
Univariate Analysis:

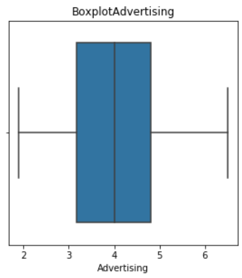


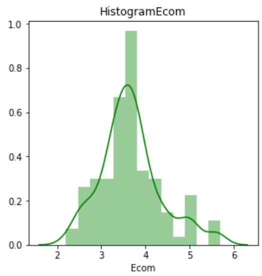


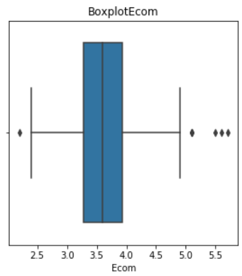












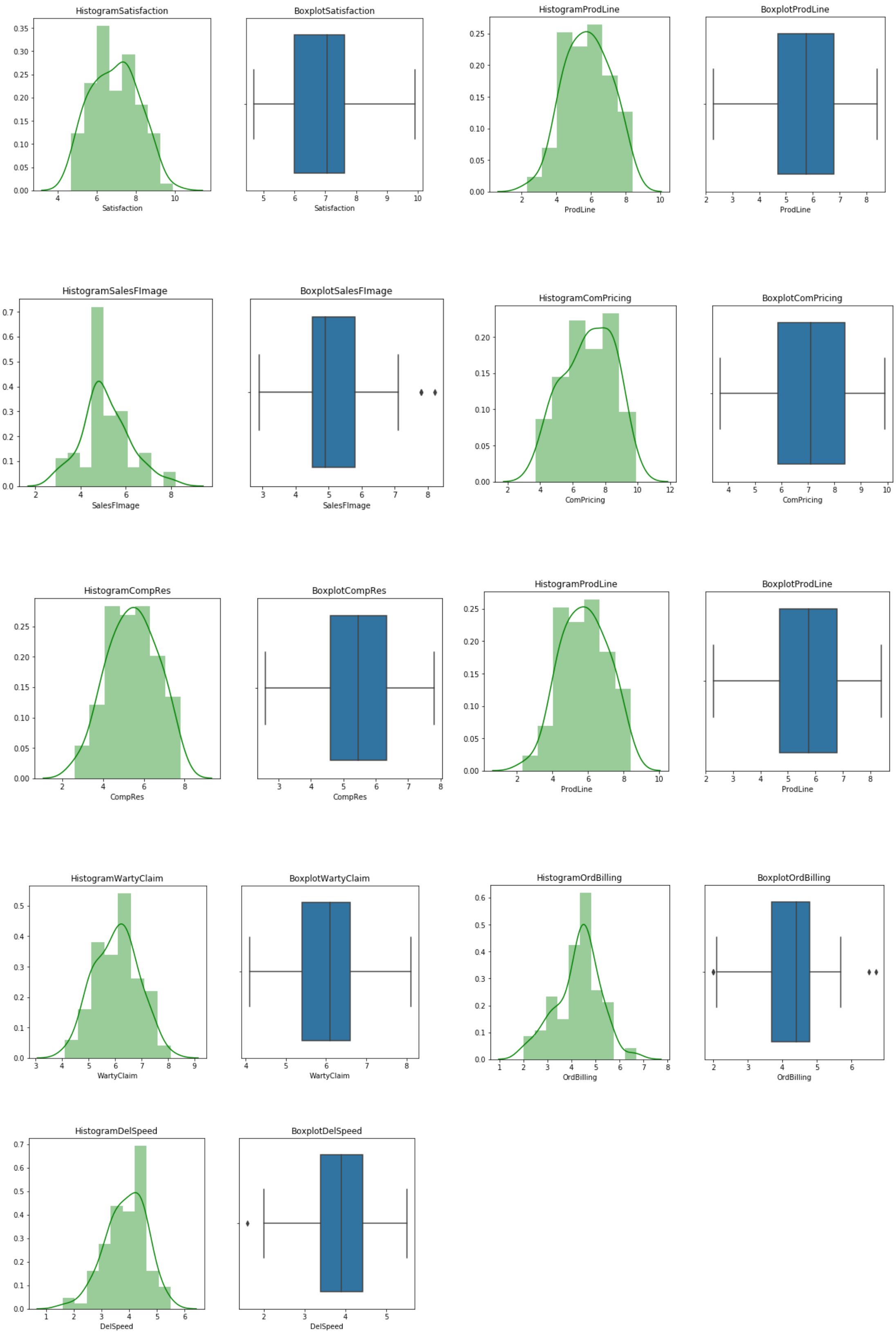


Fig. 1.Univariate analysis graphs (histogram and Boxplots)

From the above graphs Product line, Complaint Resolution, Advertising, Technical support, warranty claim, Sales force Image and Competitive pricing are normally distributed. Product quality is normal with a very little skew. E-Commerce, Order and Billing has some outliers and with normal distribution.

Bivariate Analysis:

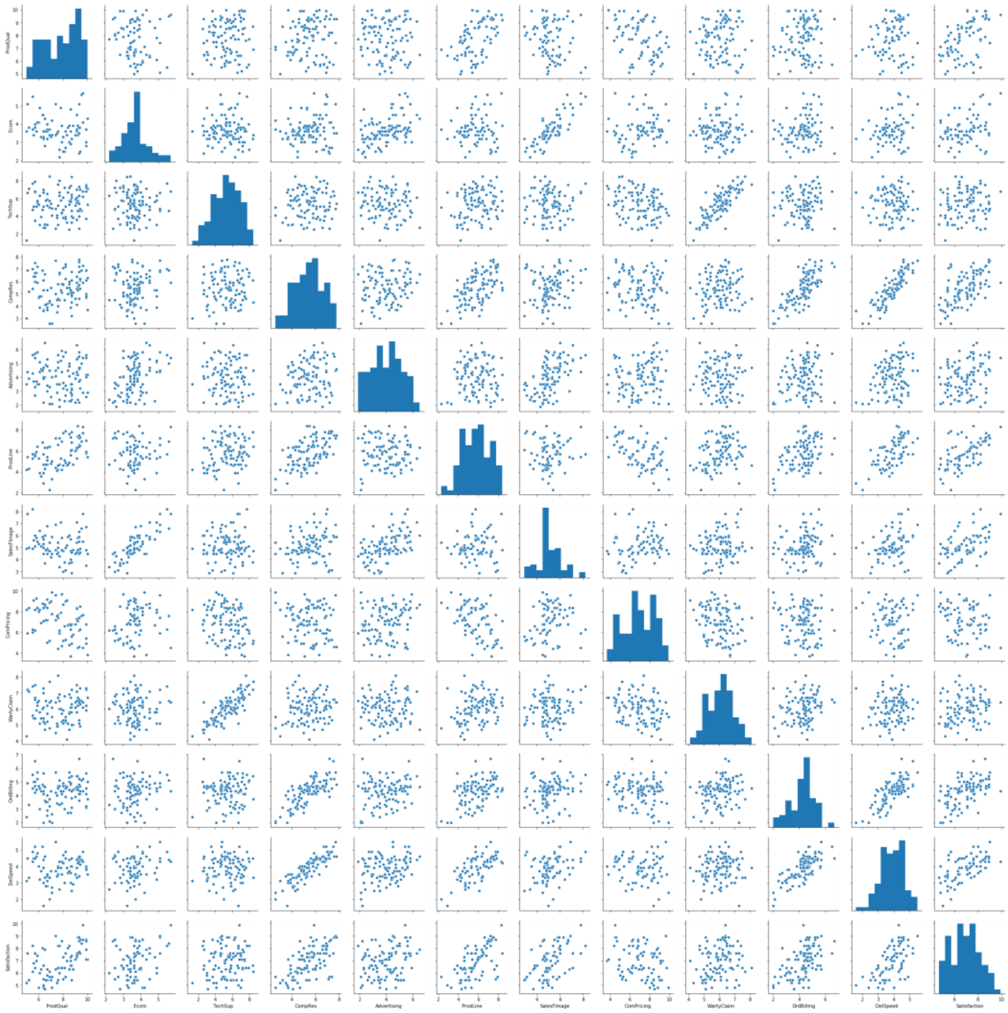


Fig. 2. Pair Plot_PCA

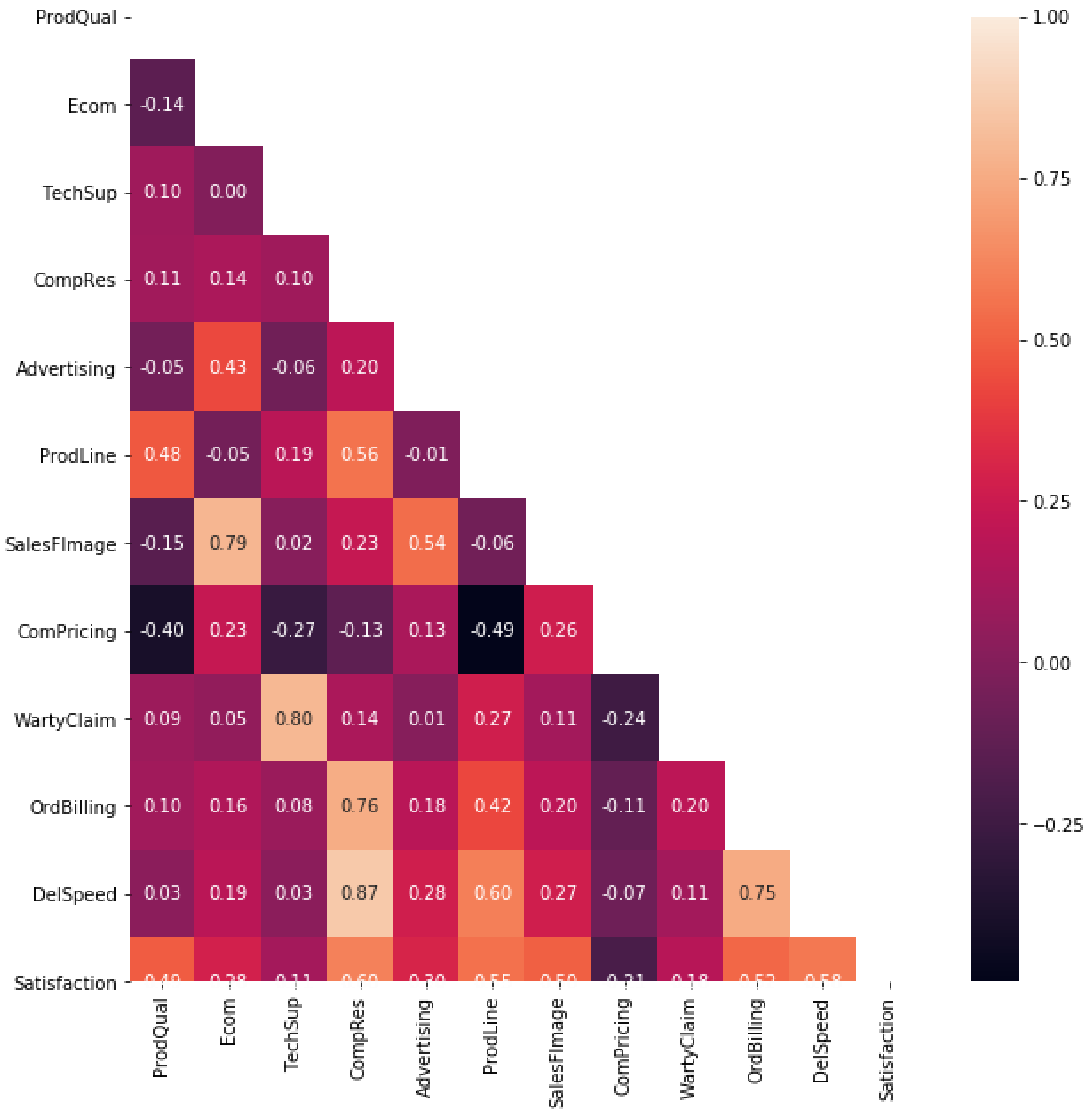


Fig. 3. Heatmap_PCA

From the Pair Plot and Heatmap which gives correlation details having high positive correlation.

- Ecom and Sales Force image
- Techsup and WartyClaim
- CompRes and Delspeed
- OrdBilling and Delspeed
- CompRes and OrdBilling

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

As PCA is affected by scale, so scale the data before applying the PCA. We can use **scipy.stats** to scale the feature of the data by using Z-Score method.

In Z-score method,

$$z = (x-\mu)/s$$

μ = mean of the training samples

s = standard deviation of the training sample

After scaling:

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
0	0.496660	0.327114	-1.881421	0.380922	0.704543	-0.691530	0.821973	-0.113185	-1.646582	0.781230	-0.254531	1.081067
1	0.280721	-1.394538	-0.174023	1.462141	-0.544014	1.600835	-1.896068	-1.088915	-0.665744	-0.409009	1.387605	-1.027098
2	1.000518	-0.390241	0.154322	0.131410	1.239639	1.218774	0.634522	-1.609304	0.192489	1.214044	0.840226	1.671354
3	-1.014914	-0.533712	1.073690	-1.448834	0.615361	-0.844354	-0.583910	1.187789	1.173327	0.023805	-1.212443	-1.786038
4	0.856559	-0.390241	-0.108354	-0.700298	-1.614207	0.149004	-0.583910	-0.113185	0.069885	0.240212	-0.528220	0.153474

Table 6. Scaled Data_PCA

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

	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

Table 7. Scaled Correlation Values

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
ProdQual	1.010101	-0.138549	0.096566	0.107444	-0.054013	0.482317	-0.153346	-0.405335	0.089204	0.105357	0.027998	0.491237
Ecom	-0.138549	1.010101	0.000876	0.141595	0.434233	-0.053220	0.799539	0.231780	0.052422	0.157725	0.193572	0.285601
TechSup	0.096566	0.000876	1.010101	0.097633	-0.063505	0.194571	0.017162	-0.273522	0.805220	0.080911	0.025698	0.113735
CompRes	0.107444	0.141595	0.097633	1.010101	0.198906	0.567088	0.232072	-0.129247	0.141827	0.764514	0.873830	0.609356
Advertising	-0.054013	0.434233	-0.063505	0.198906	1.010101	-0.011667	0.547680	0.135573	0.010901	0.186097	0.278650	0.307747
ProdLine	0.482317	-0.053220	0.194571	0.567088	-0.011667	1.010101	-0.061935	-0.499948	0.275836	0.428695	0.607930	0.556107
SalesFImage	-0.153346	0.799539	0.017162	0.232072	0.547680	-0.061935	1.010101	0.267269	0.108541	0.197098	0.274294	0.505258
ComPricing	-0.405335	0.231780	-0.273522	-0.129247	0.135573	-0.499948	0.267269	1.010101	-0.247461	-0.115724	-0.073608	-0.210400
WartyClaim	0.089204	0.052422	0.805220	0.141827	0.010901	0.275836	0.108541	-0.247461	1.010101	0.199056	0.110500	0.179338
OrdBilling	0.105357	0.157725	0.080911	0.764514	0.186097	0.428695	0.197098	-0.115724	0.199056	1.010101	0.758589	0.527002
DelSpeed	0.027998	0.193572	0.025698	0.873830	0.278650	0.607930	0.274294	-0.073608	0.110500	0.758589	1.010101	0.582871
Satisfaction	0.491237	0.285601	0.113735	0.609356	0.307747	0.556107	0.505258	-0.210400	0.179338	0.527002	0.582871	1.010101

Table 8. Scaled Covariance values

	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

Table 9.Correlation values Unscaled

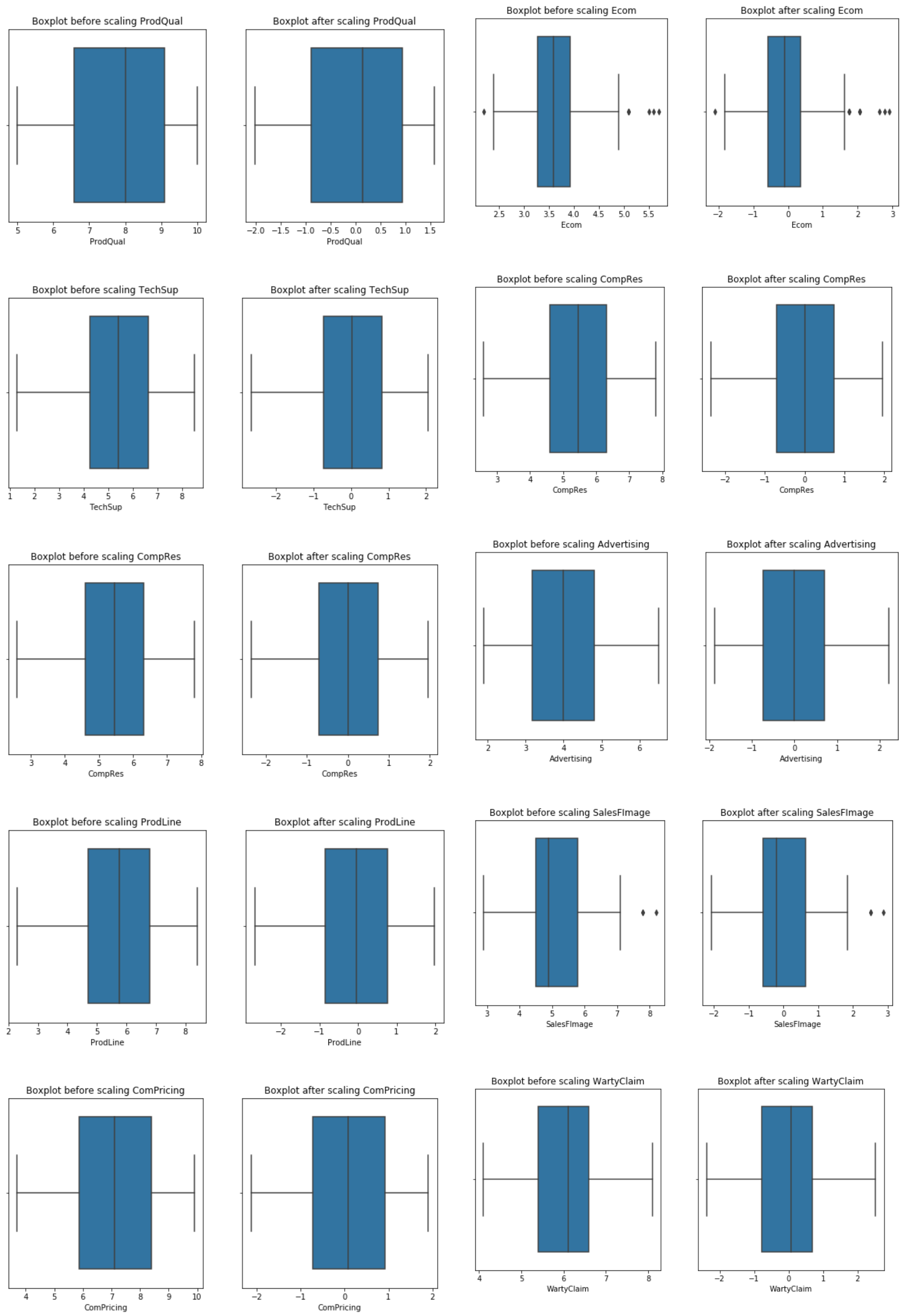
	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
ProdQual	1.949596	-0.134162	0.204293	0.179475	-0.084141	0.876919	-0.227303	-0.865697	0.101081	0.135273	0.028424	0.809313
Ecom	-0.134162	0.490723	0.000929	0.118663	0.339374	-0.048545	0.594590	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 10. Covariance values Unscaled

Covariance indicated the direction of linear relationship between variables and Correlation is the function of covariance.

While observing the values of both scaled and unscaled data of Correlation, there is no difference in between them.

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



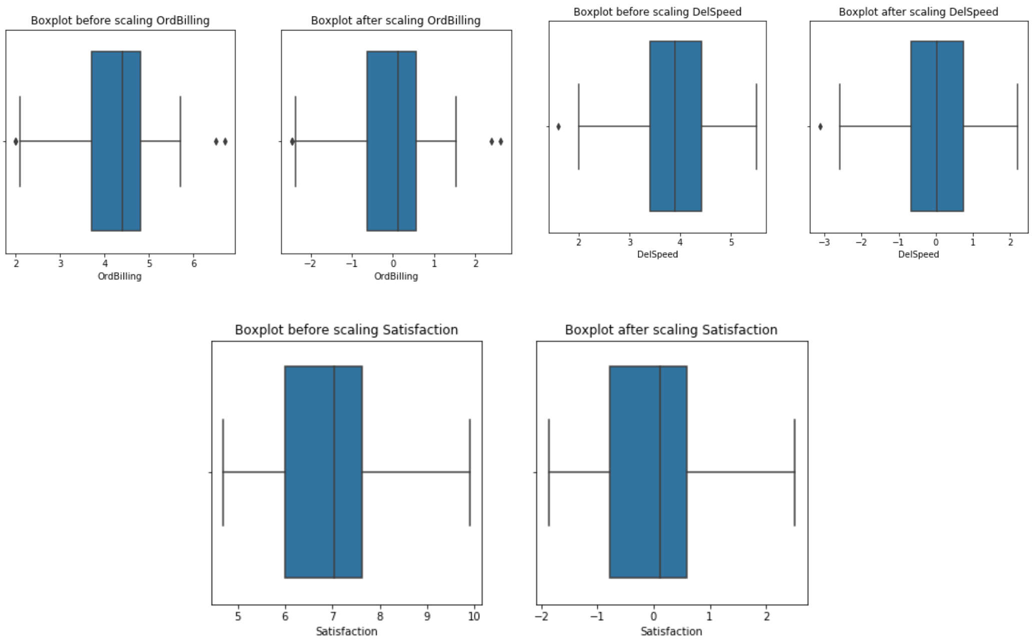


Fig. 4. Boxplots before and after scaling

From the above boxplots shown, there is not much effect of outliers before and after scaling of the variables.

5. PCA: Build the covariance matrix, eigenvalues and eigenvector.

Covariance Matrix

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
ProdQual	1.010101	-0.138549	0.096566	0.107444	-0.054013	0.482317	-0.153346	-0.405335	0.089204	0.105357	0.027998	0.491237
Ecom	-0.138549	1.010101	0.000876	0.141595	0.434233	-0.053220	0.799539	0.231780	0.052422	0.157725	0.193572	0.285601
TechSup	0.096566	0.000876	1.010101	0.097633	-0.063505	0.194571	0.017162	-0.273522	0.805220	0.080911	0.025698	0.113735
CompRes	0.107444	0.141595	0.097633	1.010101	0.198906	0.567088	0.232072	-0.129247	0.141827	0.764514	0.873830	0.609356
Advertising	-0.054013	0.434233	-0.063505	0.198906	1.010101	-0.011667	0.547680	0.135573	0.010901	0.186097	0.278650	0.307747
ProdLine	0.482317	-0.053220	0.194571	0.567088	-0.011667	1.010101	-0.061935	-0.499948	0.275836	0.428695	0.607930	0.556107
SalesFImage	-0.153346	0.799539	0.017162	0.232072	0.547680	-0.061935	1.010101	0.267269	0.108541	0.197098	0.274294	0.505258
ComPricing	-0.405335	0.231780	-0.273522	-0.129247	0.135573	-0.499948	0.267269	1.010101	-0.247461	-0.115724	-0.073608	-0.210400
WartyClaim	0.089204	0.052422	0.805220	0.141827	0.010901	0.275836	0.108541	-0.247461	1.010101	0.199056	0.110500	0.179338
OrdBilling	0.105357	0.157725	0.080911	0.764514	0.186097	0.428695	0.197098	-0.115724	0.199056	1.010101	0.758589	0.527002
DelSpeed	0.027998	0.193572	0.025698	0.873830	0.278650	0.607930	0.274294	-0.073608	0.110500	0.758589	1.010101	0.582871
Satisfaction	0.491237	0.285601	0.113735	0.609356	0.307747	0.556107	0.505258	-0.210400	0.179338	0.527002	0.582871	1.010101

Table 11.Covariance Matrix_PCA

Eigen Values:

array ([3.12504686, 2.23977366, 1.55039912, 1.04281689, 0.6183749,

0.43703311, 0.39005721, 0.24491075, 0.20132541, 0.12424549,

0.0975319])

Eigen Vectors:

```
array([[-0.20874524,  0.00785665, -0.23647319, -0.46958593, -0.0966874 ,
        -0.45803627, -0.05264898,  0.24911169, -0.28377432, -0.35663144,
        -0.43387222],
       [-0.27901965,  0.27209646, -0.27923835,  0.23264359,  0.41174654,
        -0.14942319,  0.43663998,  0.41811331, -0.21457071,  0.19110584,
         0.27589257],
       [ 0.23625851, -0.18375239, -0.59583178,  0.16616098, -0.17133247,
         0.22860203, -0.23694206, -0.12834804, -0.59534637,  0.04810997,
         0.12965979],
       [ 0.60612246,  0.19464119, -0.07431341, -0.21921792,  0.52176118,
         0.12622518,  0.31867405, -0.23236174, -0.05147397, -0.20404537,
        -0.23055294],
       [-0.52943906, -0.21917653, -0.04406656,  0.00663353,  0.54381836,
        -0.01455584, -0.22138005, -0.54033705, -0.05631267, -0.16442907,
         0.05038347],
       [ 0.25419554, -0.54148365,  0.09911307,  0.07774564,  0.41622836,
        -0.02604969, -0.3561701 ,  0.55373193,  0.08586758, -0.0959344 ,
         0.04832013],
       [-0.22648312,  0.0029601 , -0.03950657, -0.05457839, -0.11476415,
         0.62451177,  0.18774638,  0.22094888,  0.06496696, -0.65459062,
         0.15554616],
       [ 0.11488336, -0.17679751,  0.42238718,  0.490482  , -0.12257153,
```

```
-0.33978316,  0.27606619, -0.12991027, -0.38614927, -0.40264991,
  0.04191821],
[ 0.04498111, -0.5250979 , -0.45381987,  0.01808496, -0.14461155,
 -0.22839929,  0.45580819, -0.14810285,  0.45479736, -0.04273674,
  0.06729037],
[ 0.1219168 ,  0.44539783, -0.32074532,  0.4347328 ,  0.0136532 ,
 -0.25436455, -0.38026915, -0.01614618,  0.37966679, -0.37513695,
  0.01610075],
[ 0.17049942,  0.08738203,  0.07072047, -0.45906491, -0.0353389 ,
 -0.27651083, -0.09584195, -0.07865125, -0.03848165, -0.15143703,
  0.79376291]])
```

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

(-0.21) * ProdQual + (0.01) * Ecom + (-0.24) * TechSup + (-0.47) * CompRes + (-0.1) * Advertising + (-0.46) * ProdLine + (-0.05) * SalesFImage + (0.25) * ComPricing + (-0.28) * WartyClaim + (-0.36) * OrdBilling + (-0.43) * DelSpeed

7. 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.

Cumulative values give the % of variance for the n components. For the given cumulative variance, we consider approx. 80% of the variance within the dataset.

With help of cumulative values and scree plot we can find the optimum number of principal components.

Cumulative values of variance:

array ([0.31028567, 0.53267263, 0.68661164, 0.79015285, 0.85155125, 0.89494423, 0.93367298, 0.95799015, 0.97797974, 0.99031606,]

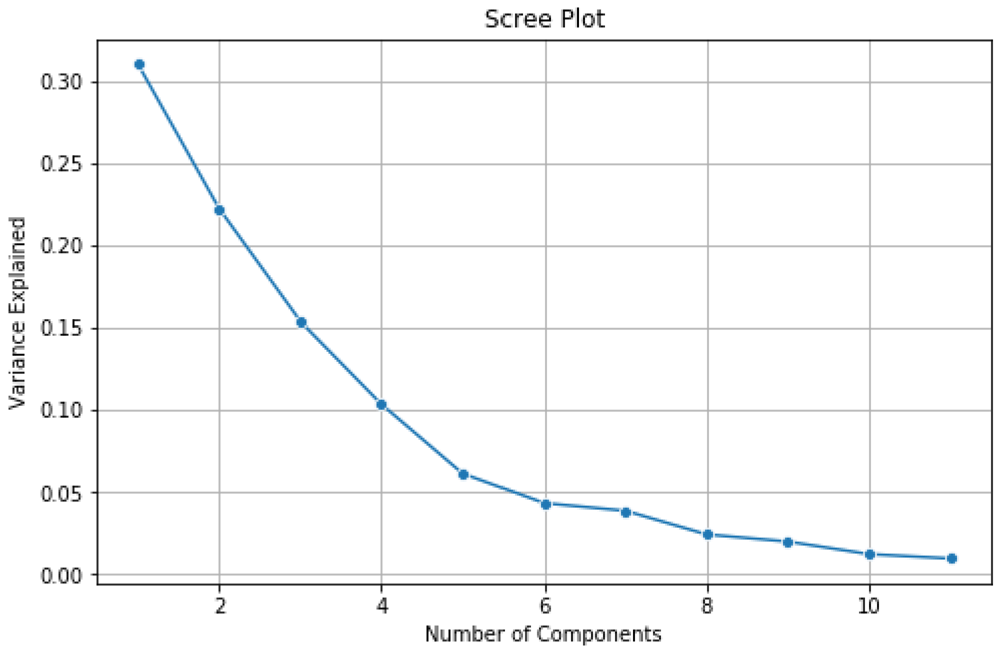


Fig. 5. Screeplot_PCA

On considering % of variance and scree plot, optimum number of Principal components considered where the % reaches 80 approx. i.e.., 4 PCs. In scree plot too there is a steep drop after the 4^{th} Principal Component.

Eigen vectors indicate the amount of weight consider for each variable value within each PC to make those orthogonal values equal to the original co-ordinate system.

Data frame for all principal components is given below.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
ProdQual	-0.208745	-0.279020	0.236259	0.606122	-0.529439	0.254196	-0.226483	0.114883	0.044981	0.121917	0.170499
Ecom	0.007857	0.272096	-0.183752	0.194641	-0.219177	-0.541484	0.002960	-0.176798	-0.525098	0.445398	0.087382
TechSup	-0.236473	-0.279238	-0.595832	-0.074313	-0.044067	0.099113	-0.039507	0.422387	-0.453820	-0.320745	0.070720
CompRes	-0.469586	0.232644	0.166161	-0.219218	0.006634	0.077746	-0.054578	0.490482	0.018085	0.434733	-0.459065
Advertising	-0.096687	0.411747	-0.171332	0.521761	0.543818	0.416228	-0.114764	-0.122572	-0.144612	0.013653	-0.035339
ProdLine	-0.458036	-0.149423	0.228602	0.126225	-0.014556	-0.026050	0.624512	-0.339783	-0.228399	-0.254365	-0.276511
SalesFImage	-0.052649	0.436640	-0.236942	0.318674	-0.221380	-0.356170	0.187746	0.276066	0.455808	-0.380269	-0.095842
ComPricing	0.249112	0.418113	-0.128348	-0.232362	-0.540337	0.553732	0.220949	-0.129910	-0.148103	-0.016146	-0.078651
WartyClaim	-0.283774	-0.214571	-0.595346	-0.051474	-0.056313	0.085868	0.064967	-0.386149	0.454797	0.379667	-0.038482
OrdBilling	-0.356631	0.191106	0.048110	-0.204045	-0.164429	-0.095934	-0.654591	-0.402650	-0.042737	-0.375137	-0.151437
DelSpeed	-0.433872	0.275893	0.129660	-0.230553	0.050383	0.048320	0.155546	0.041918	0.067290	0.016101	0.793763

Table 12. Principal Component Scores Data Frame

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

4P components gives us the variance of 80% that we found with the scree plot and explained variance.

Correlation between the 4 components and features:

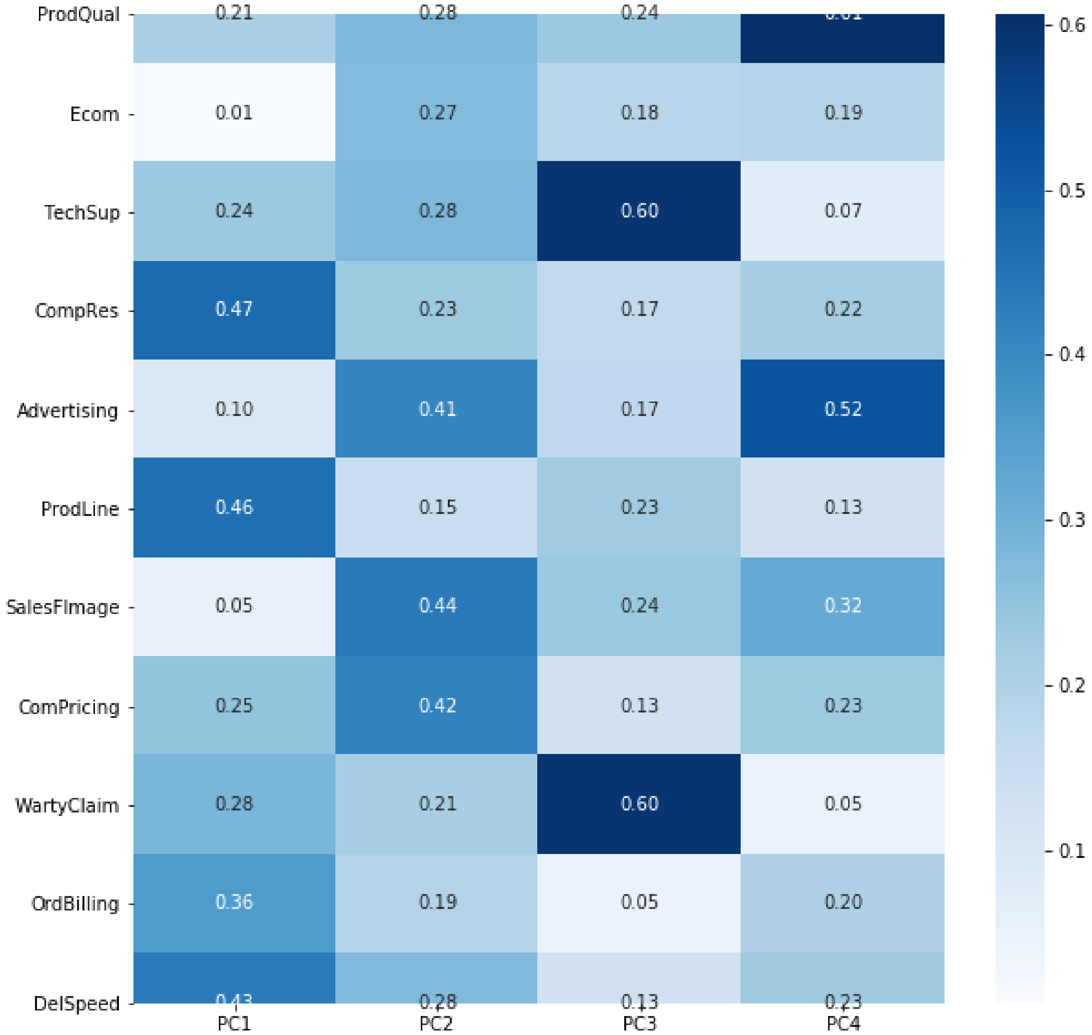


Fig. 6. Heatmap_PCs vs Variables

Heatmap represents the correlation between the optimal principal components and various features of the case study.

Have removed the multicollinearity between the variables of different market segments present in the hair products by reducing the columns from 11 to 4.

Problem Statement 2:

The dataset given is about the health and economic conditions in different States of a country. 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.

Data Dictionary for State_wise_Health_income:

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.

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

Data set has 5 important features and 297 rows. States is an object feature and remaining all are int64 data type i.e.., integer.

There are no missing values in the given data

Explore the first 5 rows of the data set.

	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 13. Head_Clustering

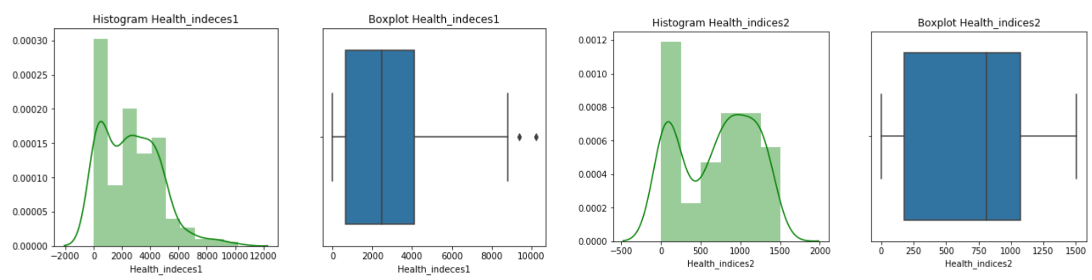
Checking summary:

The mean per capital income of all the states is 2156.92 and GDP is 174601.12.

	Unnamed: 0	Health_indeces1	Health_indices2	Per_capita_income	GDP
count	297.000000	297.000000	297.000000	297.000000	297.000000
mean	148.000000	2630.151515	693.632997	2156.915825	174601.117845
std	85.880731	2038.505431	468.944354	1491.854058	167167.992863
min	0.000000	-10.000000	0.000000	500.000000	22.000000
25%	74.000000	641.000000	175.000000	751.000000	8721.000000
50%	148.000000	2451.000000	810.000000	1865.000000	137173.000000
75%	222.000000	4094.000000	1073.000000	3137.000000	313092.000000
max	296.000000	10219.000000	1508.000000	7049.000000	728575.000000

Table 14. Summary_Clustering

Univariate Analysis:



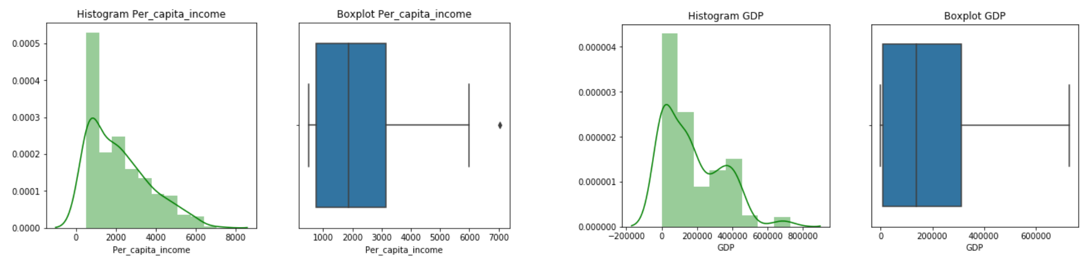


Fig. 7. Univariate Analysis

The above graph shows, Health Indices_1, Per_capita_income and GDP are positively skewed.

Per_capita_income and Health Indices are having outliers.

Bivariate Analysis:

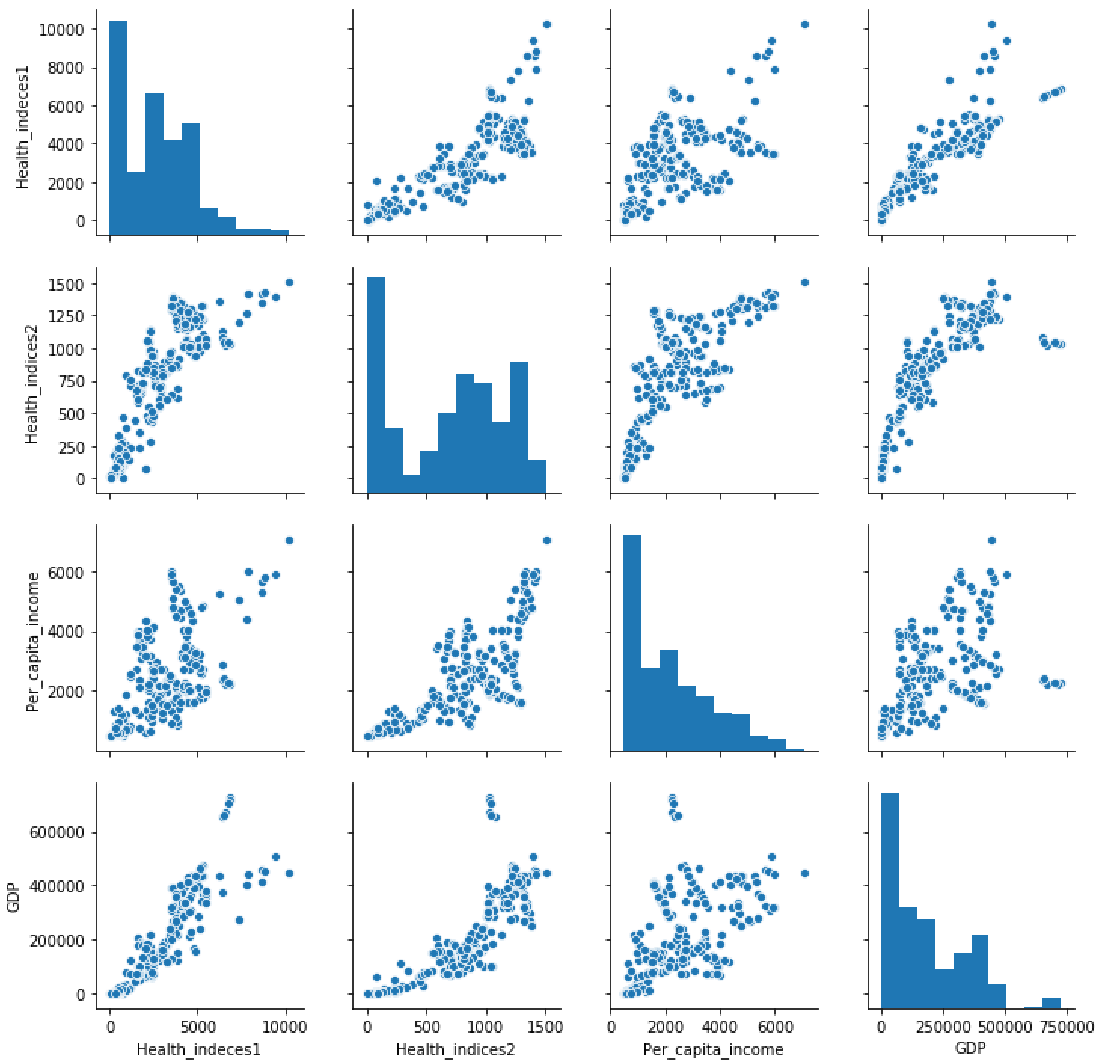


Fig. 8. Pairplot_Clustering

There is a high correlation is observed between Health_indices1 and 2, Health_indices2 and Per_capita_income, Health_indices2 and GDP, Health_indice1 and GDP.

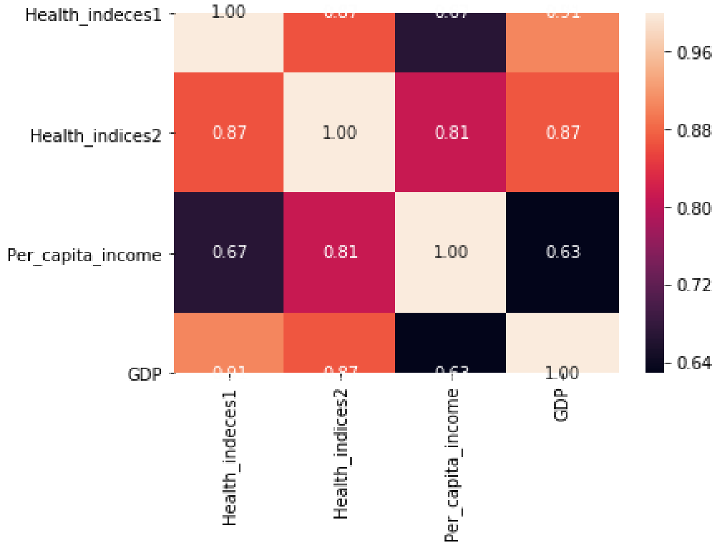


Fig. 9. Heatmap_Clustering

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

To ensure that none of the feature is identified as important only because of weight, as weight of all features are different. So scaling is required.

By using scipy.stats library function Z-score we scaled the data to reduce the weights.

In Z-score method,

z = (x-μ)/s

μ = mean of the training samples

s = standard deviation of the training sample

Scaled data first 5 rows:

	Health_indeces1	Health_indices2	Per_capita_income	GDP
0	-1.087506	-1.340654	-1.069544	-1.035304
1	-0.562708	-0.101746	0.371362	-0.604838
2	-0.971048	-0.842955	-0.706968	-0.882536
3	-1.198067	-1.428232	-1.063502	-1.044730
4	-1.271283	-1.464545	-1.093716	-1.046096

Table 15.Clustering data_after scaling

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

On applying Hierarchical clustering to scaled dataset, clusters are created. Below is the Dendrogram.

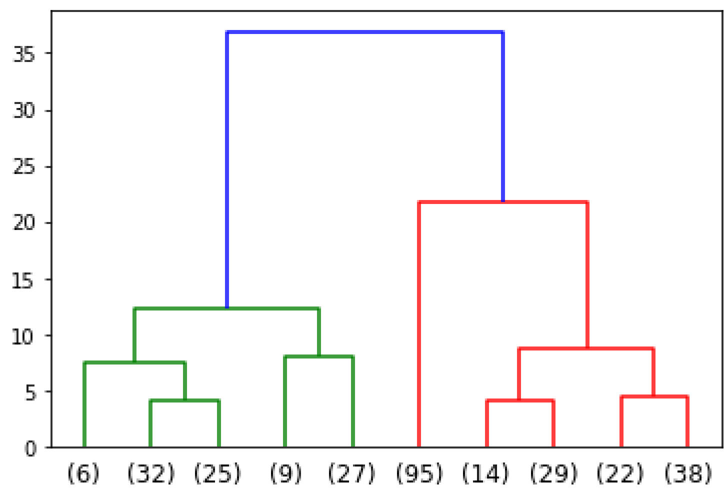


Fig. 10. Dendrogram

2 clusters cannot be considered, we cannot get many insights in the business as business already aware about the 2. Hence, to generate more insights need to take more than 2 clusters.

The optimum number of clusters that can be considered as 4.

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

To determine optimum number of clusters K-means clustering has used along with elbow curve. And the optimum number of clusters with K-means is 4.

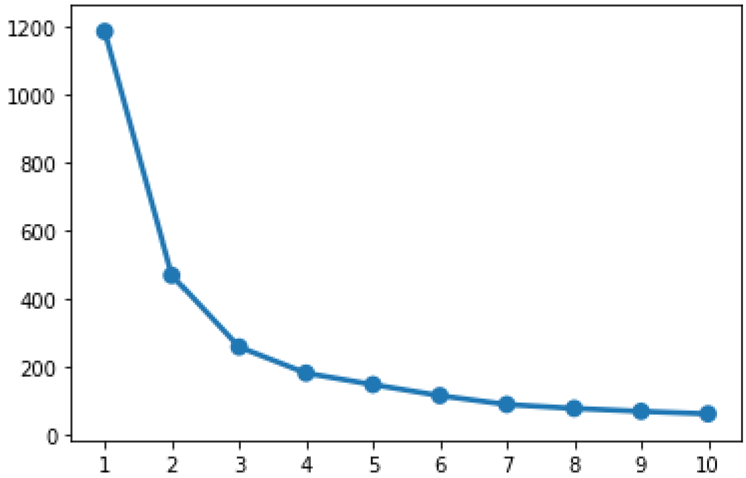


Fig. 11. Elbow plot_Kmeans Clustering

In the above elbow plot after 4 clustering values the graph is having less steep.

As per the silhouette score and inertia values the differences in it reduced after 4 clusters. So optimum number of clusters in 4.

13. 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.

From the 4 optimum clusters we can conclude that:

CLUSTER 1: This group of states or segment is least prioritized in terms of health condition as the Health indices scores are high and have enough money as per capita income to invest for improve health camps.

CLUSTER 2: As the GDP is highest in this group but health indices need little focus in view of increasing the scores as they are having enough money to deploy some health programs.

CLUSTER 3: In view of health indices this segment is very poor. Hence priority should be given for this group of states where GDP and Per capital income low to focus on their health. Government should allocate more funds for programs to improve the health of the people living in these areas.

CLUSTER 4: This is group is second most important group after cluster 3. Where health index and GDP is very low. Government should focus on deploying new programs and allocate funds too