BUSINESS REPORT-DATA MINING

[PCA& CLUSTERING]

AISHA KHAN

Contents

Table of Figures

List of Tables

PCA

Problem Statement:

The ‘Hair Salon.csv’ dataset contains various variables used for the context of Market Segmentation. This case study is based on various parameters of a salon chain of hair products.

1. Data Summary_PCA:

We will start analysing the data set by performing the basic steps which are as:

1. Checking the shape.
2. Checking head & tail.
3. Checking summary & info.
4. Checking null values.
5. Checking duplicate values.

Checking the shape:

The dataset has **100 rows and 13 features**.

Checking the head & tail:

Head Values are:

ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
1	8.5	3.9	2.5	5.9	4.8	4.9	6	6.8	4.7	5	3.7	8.2
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
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
4	6.4	3.3	7	3.7	4.7	4.7	4.5	8.8	7	4.3	3	4.8
5	9	3.4	5.2	4.6	2.2	6	4.5	6.8	6.1	4.5	3.5	7.1

Table 1. PCA_Head

Tail Values are:

ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
96	8.6	4.8	5.6	5.3	2.3	6	5.7	6.7	5.8	4.9	3.6	7.3
97	7.4	3.4	2.6	5	4.1	4.4	4.8	7.2	4.5	4.2	3.7	6.3
98	8.7	3.2	3.3	3.2	3.1	6.1	2.9	5.6	5	3.1	2.5	5.4
99	7.8	4.9	5.8	5.3	5.2	5.3	7.1	7.9	6	4.3	3.9	6.4
100	7.9	3	4.4	5.1	5.9	4.2	4.8	9.7	5.7	3.4	3.5	6.4

Table 2. PCA_Tail

Checking summary & info:

The dataset includes 1 integer feature and 12 float (decimal feature).

It includes 100 unique ID's and the statistical description of each parameter 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

Table 3. Summary of PCA Data

Type of Data	No of features
Integer Value	1
Float Values	12

Table 4. Data types count

The description of the abbreviated features are as follows:

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 5. Data Dictionary_PCA

Checking Null values:

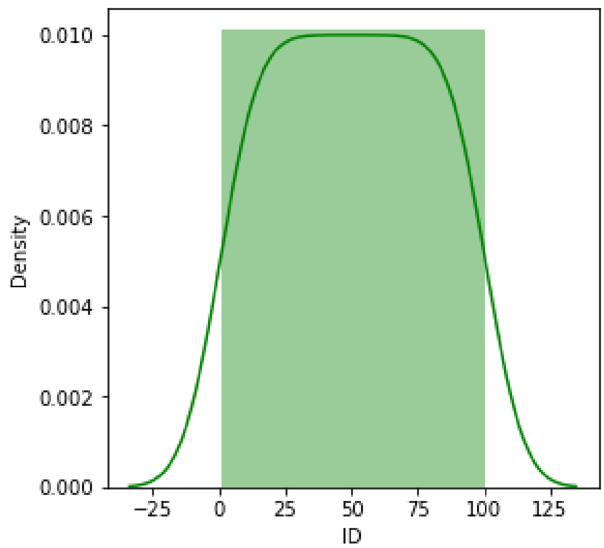
There are no null(missing) values.

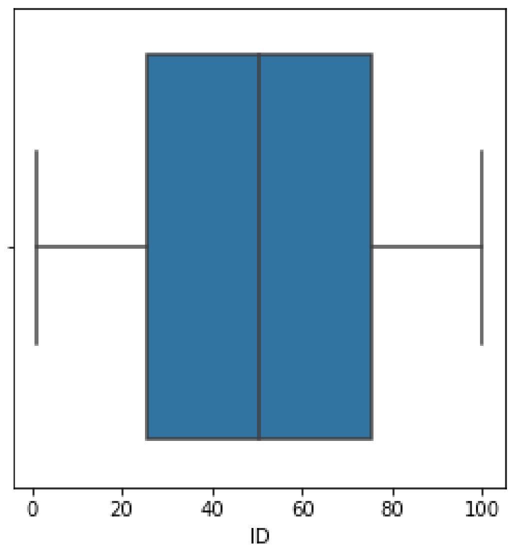
Checking Duplicates:

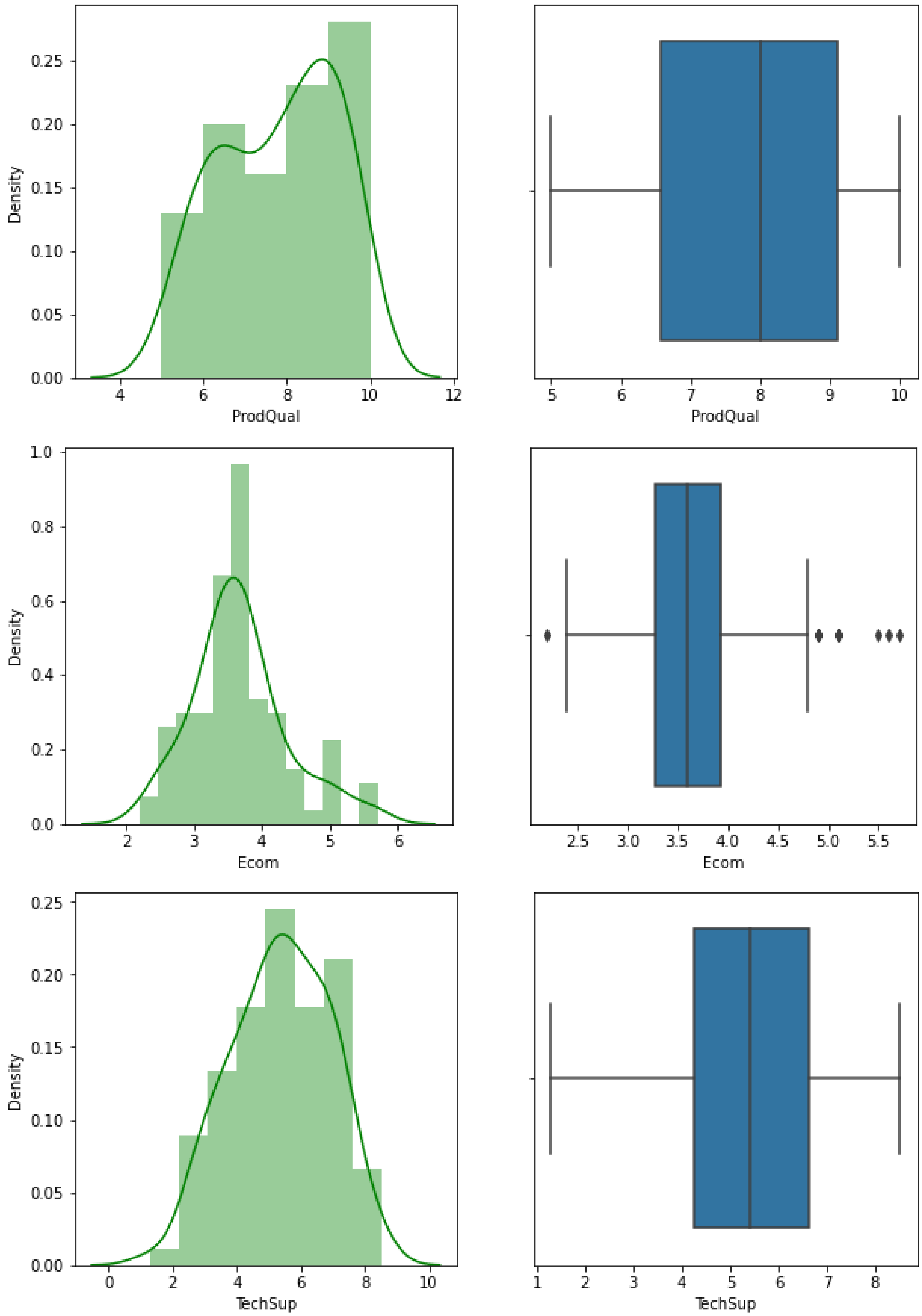
There are no duplicate values in the dataset.

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

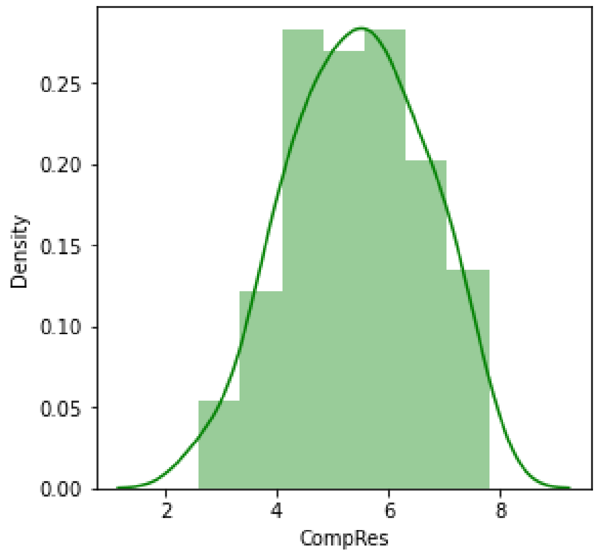
UNIVARIATE ANALYSIS:

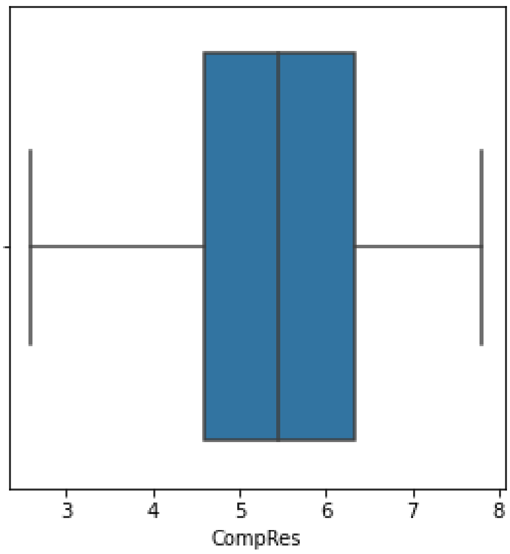


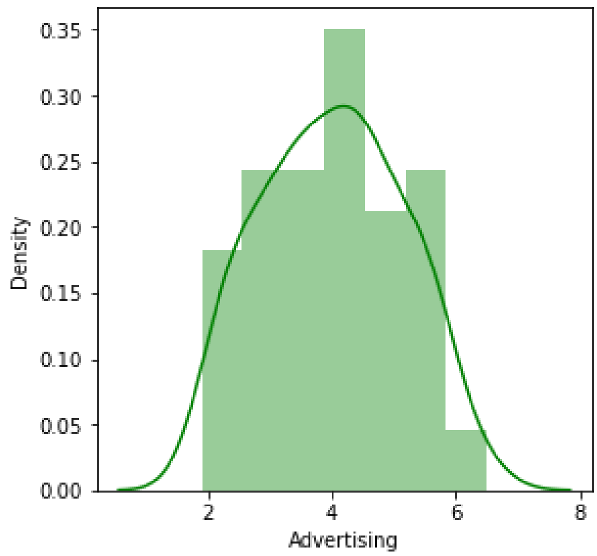


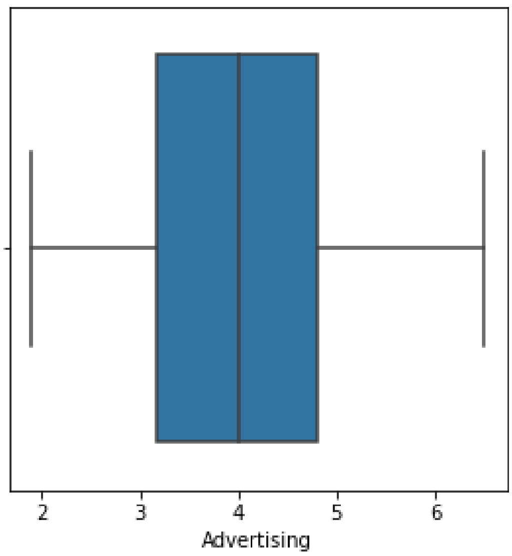


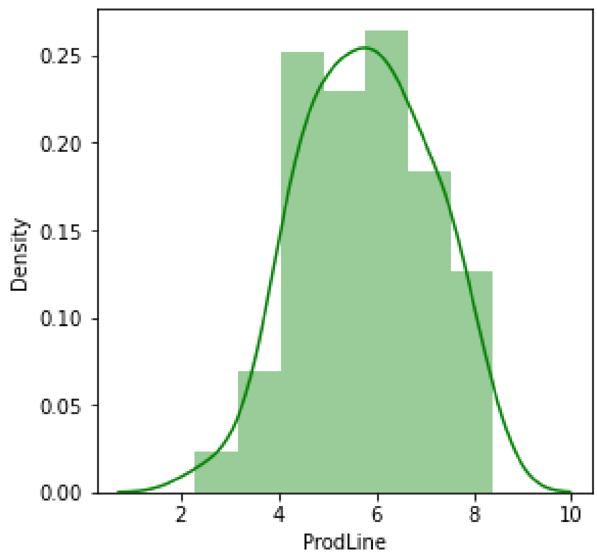
Density
0.00 0.05 0.10 0.15 0.20 0.25
4 6 8 10 12
ProdQual
5 6 7 8 9 10
ProdQual
Density
0.0 0.2 0.4 0.6 0.8 1.0
2 3 4 5 6
Ecom
2.5 3.0 3.5 4.0 4.5 5.0 5.5
Ecom
Density
0.00 0.05 0.10 0.15 0.20 0.25
0 2 4 6 8 10
TechSup
1 2 3 4 5 6 7 8
TechSup

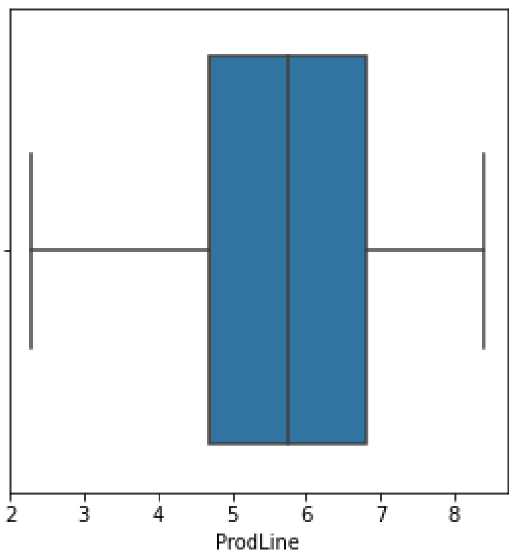

Density
0.00
0.05
0.10
0.15
0.20
0.25
2
4
6
8
CompRes

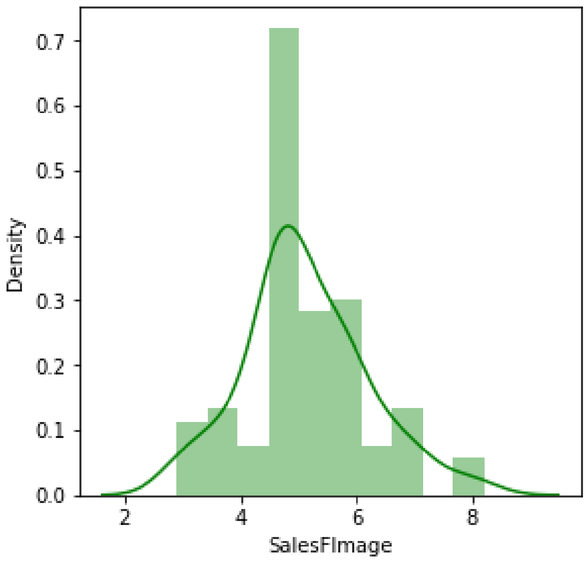

3
4
5
6
7
8
CompRes

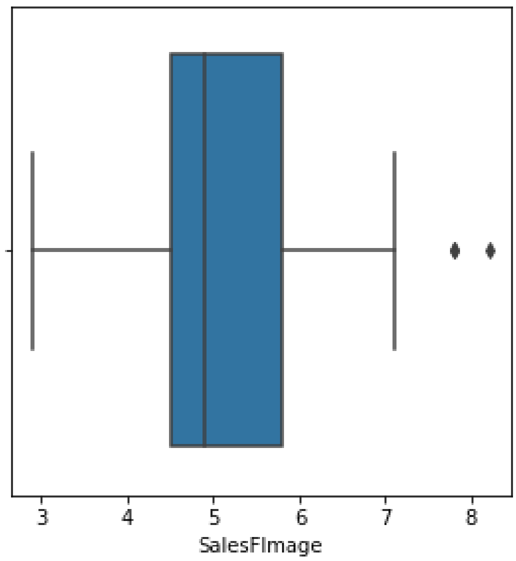

Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
2
4
6
8
Advertising

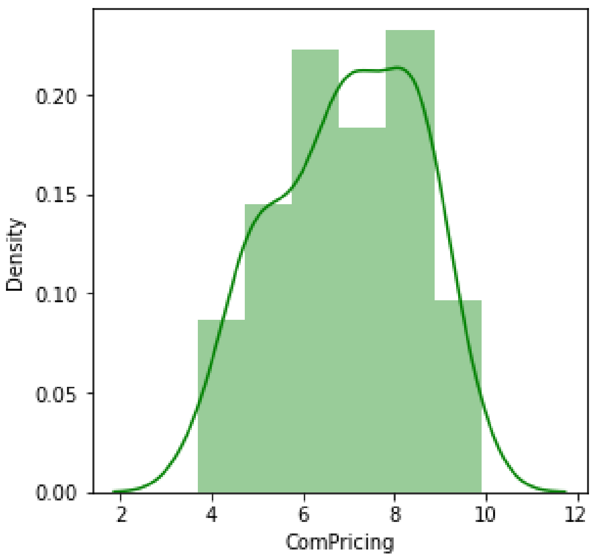

2
3
4
5
6
Advertising

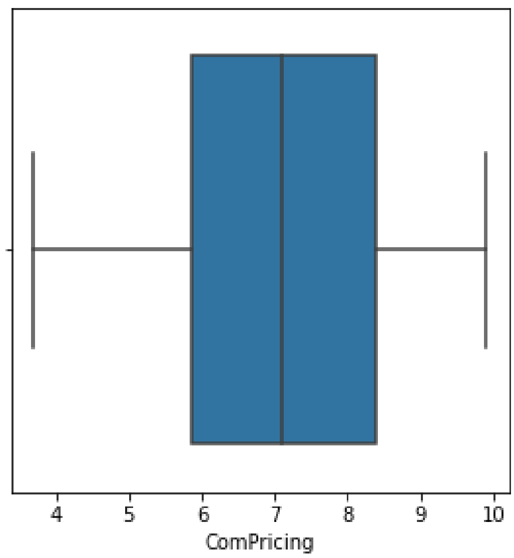

Density
0.00
0.05
0.10
0.15
0.20
0.25
2
4
6
8
10
ProdLine

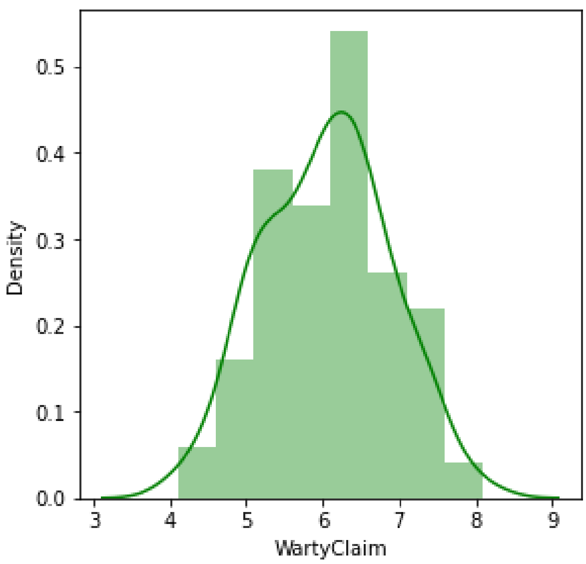

2
3
4
5
6
7
8
ProdLine

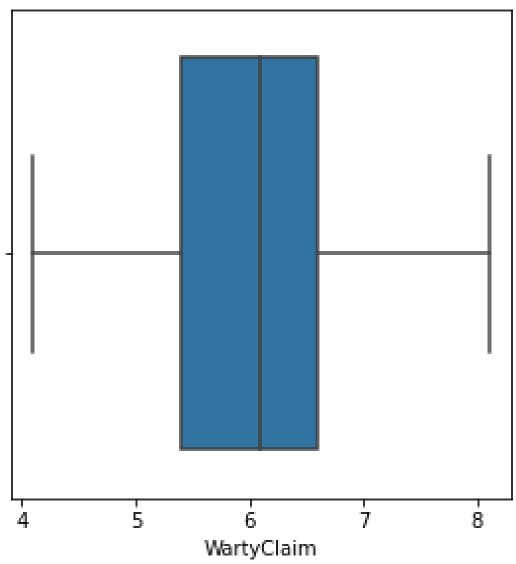

Density
SalesFImage


SalesFImage


Density
ComPricing


ComPricing


Density
WartyClaim


WartyClaim

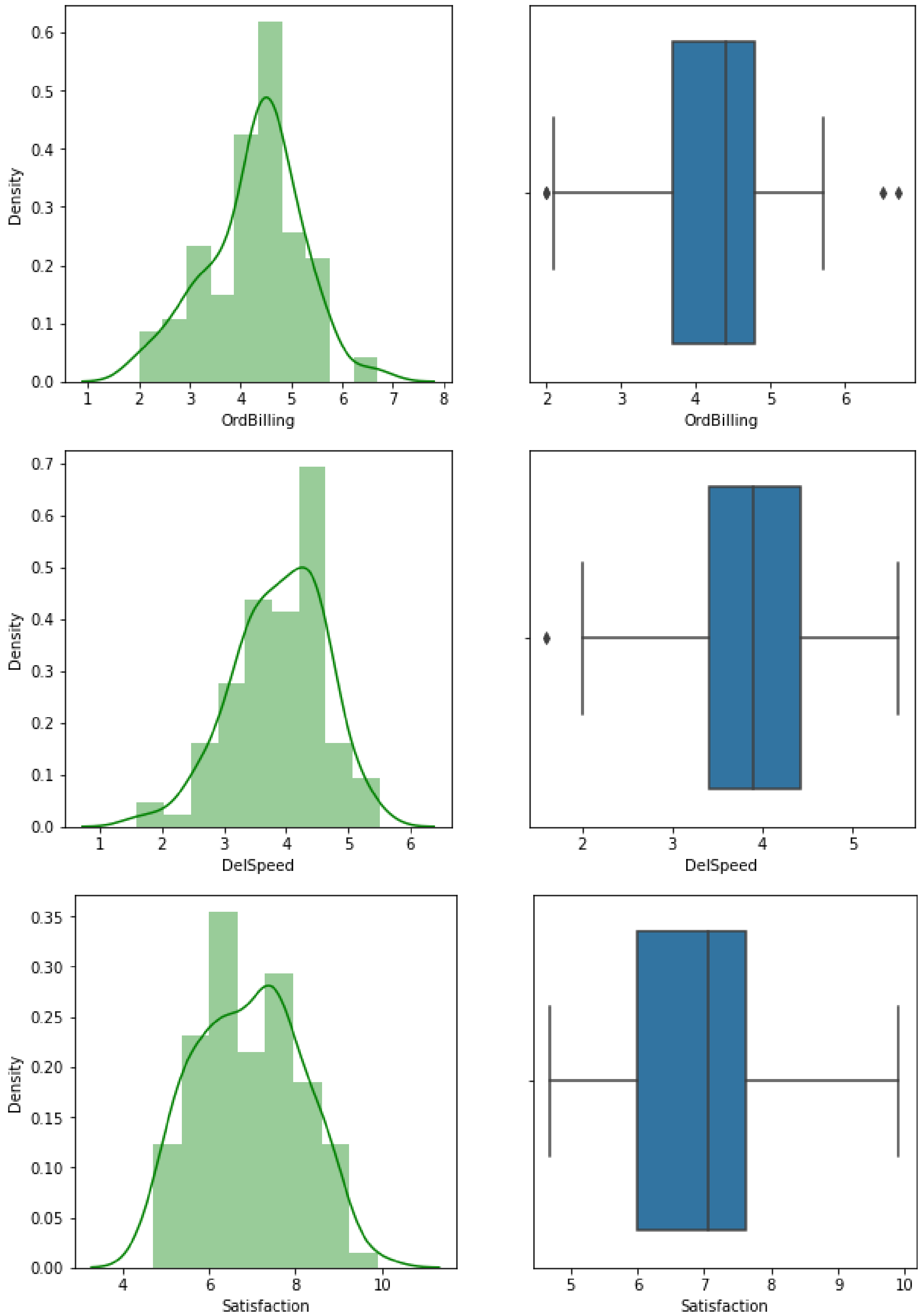


Fig. 1. Univaraite analysis Graphs (Histogram and Box Plot)

From the above graphs we can analyse that:

ProdLine, CompRes,Advertising,TechSup,WartyClaim,SalesFImage, CompPricing & ProdQual are observed to be normally distributed.

ProdQual is observed to be slightly skewed followed by Satisfaction which is slightly left skewed.

Outliers are observed in DelSpeed,OrderedBilling,Ecom & SalesFImage.

MULTIVARIATE ANALYSIS:

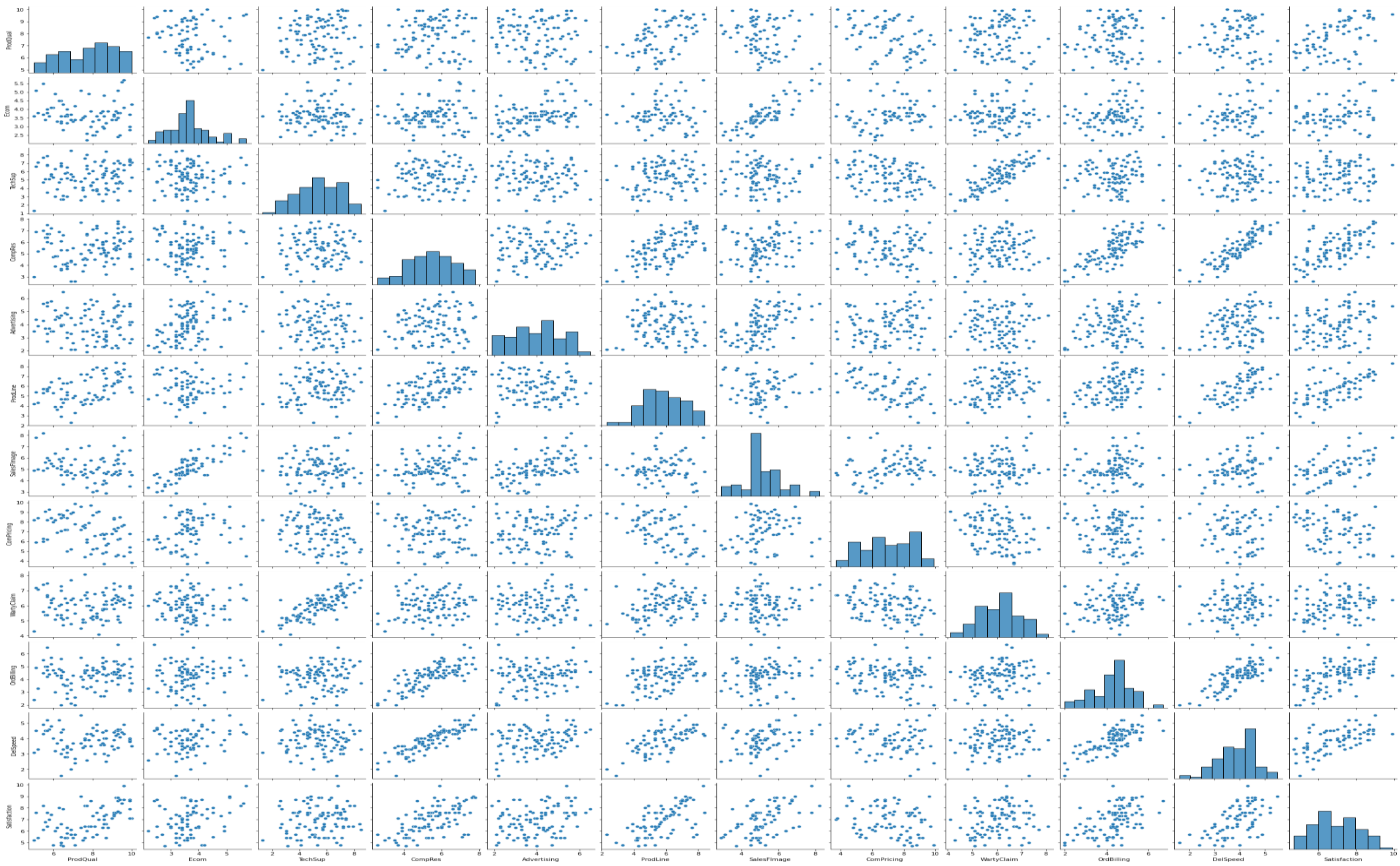


Fig. 2. Pair plot_Bivariate Analysis

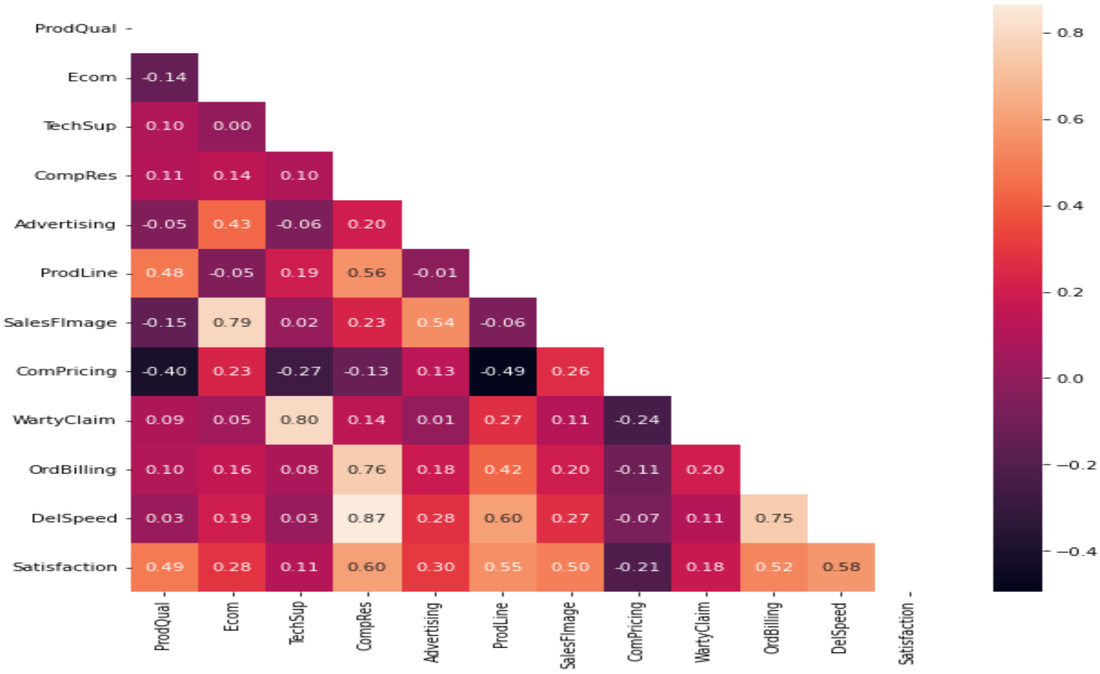


Fig. 3. Heatmap_Bivariate Analysis

The highly correlated features are:

- ECom & SalesFImage.
- TechSup & WartyClaim
- CompRes & DelSpeed
- OrdBilling & DelSpeed
- CompRes & OrdBilling

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

Scaling is required before implying PCA on the dataset as PCA is affected by scaling. Scipy is opted for the dataset to attain optimal values. Zscore method is used which is calculated as

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

Where,

μ=mean of training samples

s=standard deviation

Results after scaling are:

ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
0.50	0.33	-1.88	0.38	0.70	-0.69	0.82	-0.11	-1.65	0.78	-0.25	1.08
0.28	-1.39	-0.17	1.46	-0.54	1.60	-1.90	-1.09	-0.67	-0.41	1.39	-1.03
1.00	-0.39	0.15	0.13	1.24	1.22	0.63	-1.61	0.19	1.21	0.84	1.67
-1.01	-0.53	1.07	-1.45	0.62	-0.84	-0.58	1.19	1.17	0.02	-1.21	-1.79
0.86	-0.39	-0.11	-0.70	-1.61	0.15	-0.58	-0.11	0.07	0.24	-0.53	0.15

Table 6. PCA_Scaled data Head

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

Results before scaling:

Covariance

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
ProdQual	1.95	-0.13	0.20	0.18	-0.08	0.88	-0.23	-0.87	0.10	0.14	0.03	0.81
Ecom	-0.13	0.49	0.00	0.12	0.34	-0.05	0.59	0.25	0.03	0.10	0.10	0.24
TechSup	0.20	0.00	2.34	0.18	-0.11	0.39	0.03	-0.64	1.00	0.11	0.03	0.21
CompRes	0.18	0.12	0.18	1.46	0.27	0.89	0.30	-0.24	0.14	0.85	0.77	0.87
Advertising	-0.08	0.34	-0.11	0.27	1.27	-0.02	0.66	0.23	0.01	0.19	0.23	0.41
ProdLine	0.88	-0.05	0.39	0.89	-0.02	1.73	-0.09	-1.01	0.29	0.52	0.58	0.86
SalesFImage	-0.23	0.59	0.03	0.30	0.66	-0.09	1.15	0.44	0.09	0.19	0.21	0.64
ComPricing	-0.87	0.25	-0.64	-0.24	0.23	-1.01	0.44	2.39	-0.31	-0.16	-0.08	-0.38
WartyClaim	0.10	0.03	1.00	0.14	0.01	0.29	0.09	-0.31	0.67	0.15	0.07	0.17
OrdBilling	0.14	0.10	0.11	0.85	0.19	0.52	0.19	-0.16	0.15	0.86	0.51	0.58
DelSpeed	0.03	0.10	0.03	0.77	0.23	0.58	0.21	-0.08	0.07	0.51	0.54	0.51
Satisfaction	0.81	0.24	0.21	0.87	0.41	0.86	0.64	-0.38	0.17	0.58	0.51	1.42

Table 7. Covariance before scaling

Correlation

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
ProdQual	1.00	-0.14	0.10	0.11	-0.05	0.48	-0.15	-0.40	0.09	0.10	0.03	0.49
Ecom	-0.14	1.00	0.00	0.14	0.43	-0.05	0.79	0.23	0.05	0.16	0.19	0.28
TechSup	0.10	0.00	1.00	0.10	-0.06	0.19	0.02	-0.27	0.80	0.08	0.03	0.11
CompRes	0.11	0.14	0.10	1.00	0.20	0.56	0.23	-0.13	0.14	0.76	0.87	0.60
Advertising	-0.05	0.43	-0.06	0.20	1.00	-0.01	0.54	0.13	0.01	0.18	0.28	0.30
ProdLine	0.48	-0.05	0.19	0.56	-0.01	1.00	-0.06	-0.49	0.27	0.42	0.60	0.55
SalesFImage	-0.15	0.79	0.02	0.23	0.54	-0.06	1.00	0.26	0.11	0.20	0.27	0.50
ComPricing	-0.40	0.23	-0.27	-0.13	0.13	-0.49	0.26	1.00	-0.24	-0.11	-0.07	-0.21
WartyClaim	0.09	0.05	0.80	0.14	0.01	0.27	0.11	-0.24	1.00	0.20	0.11	0.18
OrdBilling	0.10	0.16	0.08	0.76	0.18	0.42	0.20	-0.11	0.20	1.00	0.75	0.52
DelSpeed	0.03	0.19	0.03	0.87	0.28	0.60	0.27	-0.07	0.11	0.75	1.00	0.58
Satisfaction	0.49	0.28	0.11	0.60	0.30	0.55	0.50	-0.21	0.18	0.52	0.58	1.00

Table 8. Correlation after scaling

Results after scaling:

Covariance

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
ProdQual	1.01	-0.14	0.10	0.11	-0.05	0.48	-0.15	-0.41	0.09	0.11	0.03	0.49
Ecom	-0.14	1.01	0.00	0.14	0.43	-0.05	0.80	0.23	0.05	0.16	0.19	0.29
TechSup	0.10	0.00	1.01	0.10	-0.06	0.19	0.02	-0.27	0.81	0.08	0.03	0.11
CompRes	0.11	0.14	0.10	1.01	0.20	0.57	0.23	-0.13	0.14	0.76	0.87	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	-0.05	0.19	0.57	-0.01	1.01	-0.06	-0.50	0.28	0.43	0.61	0.56
SalesFImage	-0.15	0.80	0.02	0.23	0.55	-0.06	1.01	0.27	0.11	0.20	0.27	0.51
ComPricing	-0.41	0.23	-0.27	-0.13	0.14	-0.50	0.27	1.01	-0.25	-0.12	-0.07	-0.21
WartyClaim	0.09	0.05	0.81	0.14	0.01	0.28	0.11	-0.25	1.01	0.20	0.11	0.18
OrdBilling	0.11	0.16	0.08	0.76	0.19	0.43	0.20	-0.12	0.20	1.01	0.76	0.53
DelSpeed	0.03	0.19	0.03	0.87	0.28	0.61	0.27	-0.07	0.11	0.76	1.01	0.58
Satisfaction	0.49	0.29	0.11	0.61	0.31	0.56	0.51	-0.21	0.18	0.53	0.58	1.01

Table 9. Covariance_after scaling

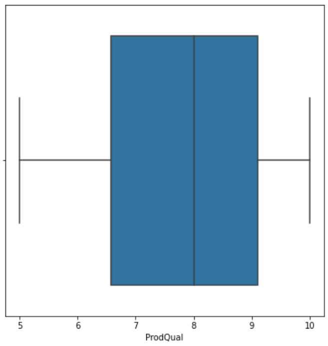
Correlation:

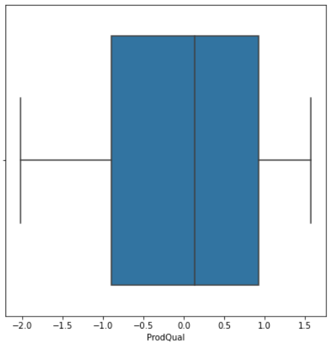
	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
ProdQual	1.00	-0.14	0.10	0.11	-0.05	0.48	-0.15	-0.40	0.09	0.10	0.03	0.49
Ecom	-0.14	1.00	0.00	0.14	0.43	-0.05	0.79	0.23	0.05	0.16	0.19	0.28
TechSup	0.10	0.00	1.00	0.10	-0.06	0.19	0.02	-0.27	0.80	0.08	0.03	0.11
CompRes	0.11	0.14	0.10	1.00	0.20	0.56	0.23	-0.13	0.14	0.76	0.87	0.60
Advertising	-0.05	0.43	-0.06	0.20	1.00	-0.01	0.54	0.13	0.01	0.18	0.28	0.30
ProdLine	0.48	-0.05	0.19	0.56	-0.01	1.00	-0.06	-0.49	0.27	0.42	0.60	0.55
SalesFImage	-0.15	0.79	0.02	0.23	0.54	-0.06	1.00	0.26	0.11	0.20	0.27	0.50
ComPricing	-0.40	0.23	-0.27	-0.13	0.13	-0.49	0.26	1.00	-0.24	-0.11	-0.07	-0.21
WartyClaim	0.09	0.05	0.80	0.14	0.01	0.27	0.11	-0.24	1.00	0.20	0.11	0.18
OrdBilling	0.10	0.16	0.08	0.76	0.18	0.42	0.20	-0.11	0.20	1.00	0.75	0.52
DelSpeed	0.03	0.19	0.03	0.87	0.28	0.60	0.27	-0.07	0.11	0.75	1.00	0.58
Satisfaction	0.49	0.28	0.11	0.60	0.30	0.55	0.50	-0.21	0.18	0.52	0.58	1.00

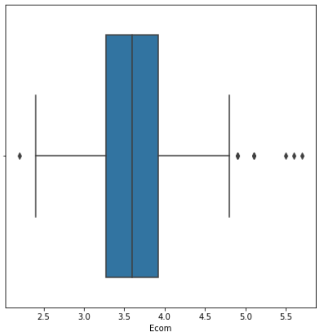
Table 10. Correlation_after scaling

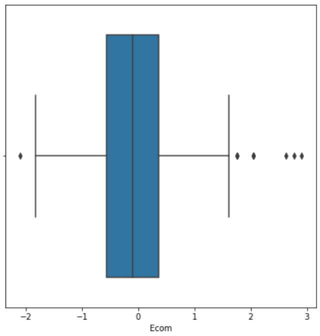
From the above values we observe that the correlation does not get affected by correlation.

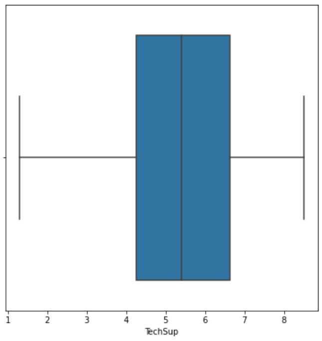
5. PCA: Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.

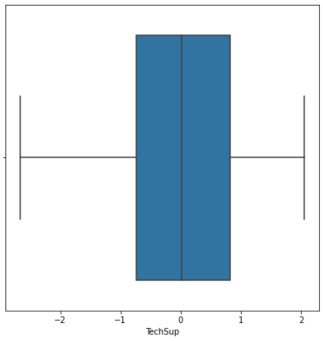


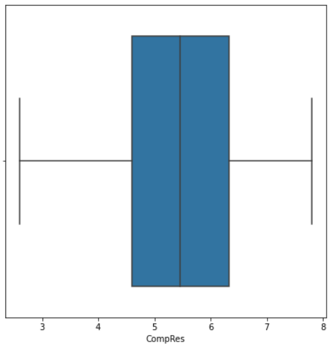


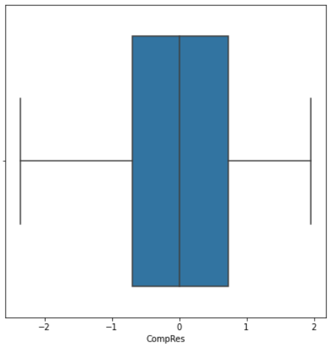


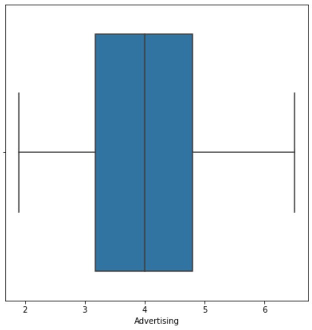


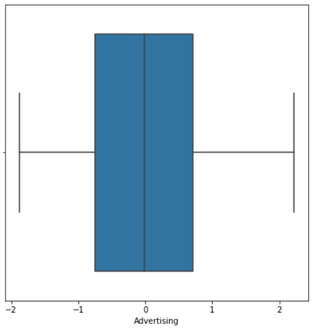


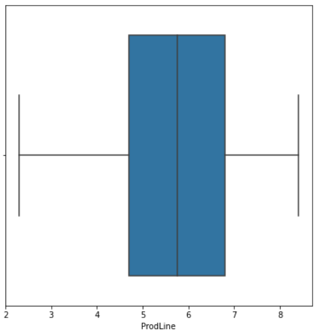


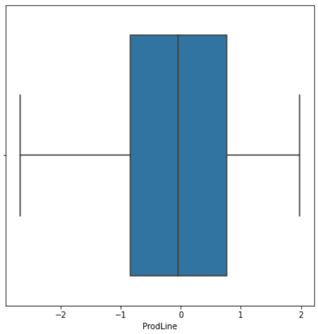


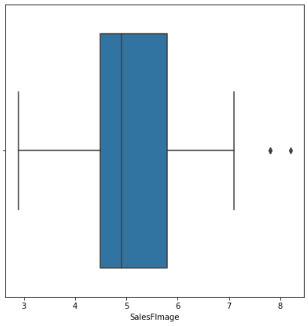


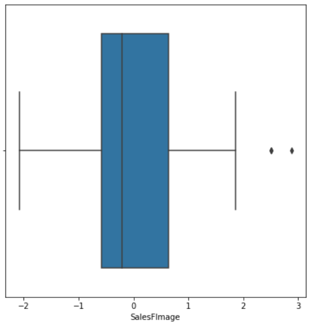


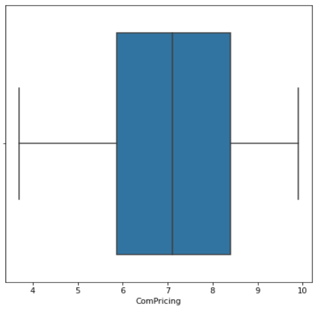


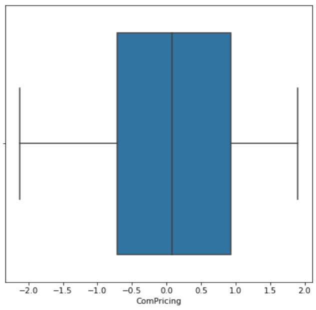












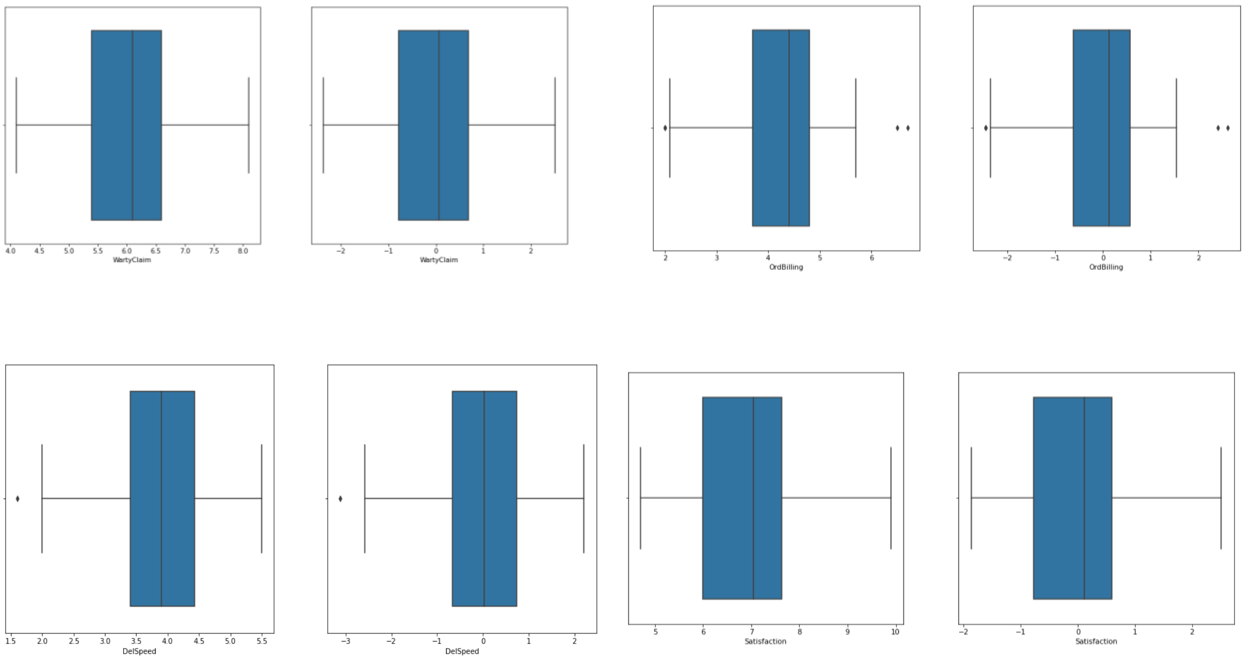


Fig. 4. Boxplots_with or without outliers

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

Covariance Matrix:

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
ProdQual	1.01	-0.14	0.10	0.11	-0.05	0.48	-0.15	-0.41	0.09	0.11	0.03	0.49
Ecom	-0.14	1.01	0.00	0.14	0.43	-0.05	0.80	0.23	0.05	0.16	0.19	0.29
TechSup	0.10	0.00	1.01	0.10	-0.06	0.19	0.02	-0.27	0.81	0.08	0.03	0.11
CompRes	0.11	0.14	0.10	1.01	0.20	0.57	0.23	-0.13	0.14	0.76	0.87	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	-0.05	0.19	0.57	-0.01	1.01	-0.06	-0.50	0.28	0.43	0.61	0.56
SalesFImage	-0.15	0.80	0.02	0.23	0.55	-0.06	1.01	0.27	0.11	0.20	0.27	0.51
ComPricing	-0.41	0.23	-0.27	-0.13	0.14	-0.50	0.27	1.01	-0.25	-0.12	-0.07	-0.21
WartyClaim	0.09	0.05	0.81	0.14	0.01	0.28	0.11	-0.25	1.01	0.20	0.11	0.18
OrdBilling	0.11	0.16	0.08	0.76	0.19	0.43	0.20	-0.12	0.20	1.01	0.76	0.53
DelSpeed	0.03	0.19	0.03	0.87	0.28	0.61	0.27	-0.07	0.11	0.76	1.01	0.58
Satisfaction	0.49	0.29	0.11	0.61	0.31	0.56	0.51	-0.21	0.18	0.53	0.58	1.01

Table 11. Covariance matrix_scaled

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([[-0.21,  0.01, -0.24, -0.47, -0.1 , -0.46, -0.05,  0.25, -0.28,

        -0.36, -0.43],
       [-0.28,  0.27, -0.28,  0.23,  0.41, -0.15,  0.44,  0.42, -0.21,
         0.19,  0.28],
       [ 0.24, -0.18, -0.6,  0.17, -0.17,  0.23, -0.24, -0.13, -0.6 ,
         0.05,  0.13],
       [ 0.61,  0.19, -0.07, -0.22,  0.52,  0.13,  0.32, -0.23, -0.05,
        -0.2 , -0.23],
       [-0.53, -0.22, -0.04,  0.01,  0.54, -0.01, -0.22, -0.54, -0.06,
        -0.16,  0.05],
       [ 0.25, -0.54,  0.1 ,  0.08,  0.42, -0.03, -0.36,  0.55,  0.09,
        -0.1 ,  0.05],
       [-0.23,  0.  , -0.04, -0.05, -0.11,  0.62,  0.19,  0.22,  0.06,
        -0.65,  0.16],
       [ 0.11, -0.18,  0.42,  0.49, -0.12, -0.34,  0.28, -0.13, -0.39,
        -0.4 ,  0.04],
       [ 0.04, -0.53, -0.45,  0.02, -0.14, -0.23,  0.46, -0.15,  0.45,
        -0.04,  0.07],
       [ 0.12,  0.45, -0.32,  0.43,  0.01, -0.25, -0.38, -0.02,  0.38,
        -0.38,  0.02],
       [ 0.17,  0.09,  0.07, -0.46, -0.04, -0.28, -0.1 , -0.08, -0.04,
        -0.15,  0.79]])
```

7. PCA: 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 + ( -0.3
6 ) * OrdBilling + ( -0.43 ) * DelSpeed +
```

8. 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.

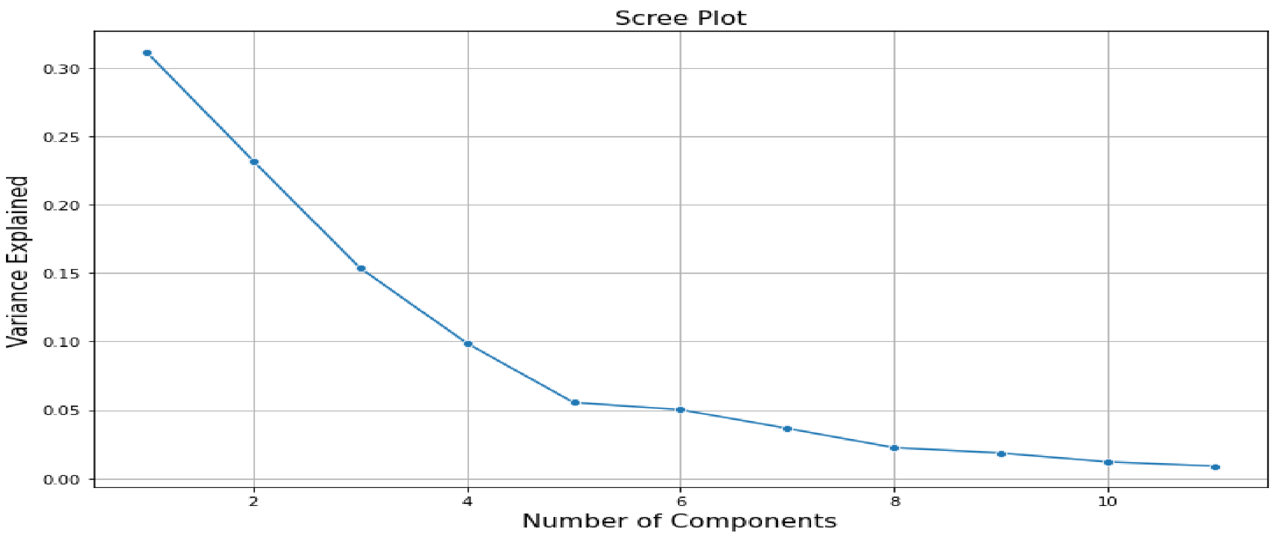


Fig. 5. Scree plot_variance

From the above Scree Plot, we observe that there is a drop from the 4th PC component, we can conclude that 4 PC components give us the maximum variance of the dataset ~80%.

The Eigen Vectors indicate the weight of each principal component towards the variables.

Dataframe for the PC's are as follows:

	pc1	pc2	pc3	pc4	pc5	pc6	pc7	pc8	pc9	pc10	pc11
ProdQual	-0.21	-0.28	0.24	0.61	-0.53	0.25	-0.23	0.11	0.04	0.12	0.17
Ecom	0.01	0.27	-0.18	0.19	-0.22	-0.54	0.00	-0.18	-0.53	0.45	0.09
TechSup	-0.24	-0.28	-0.60	-0.07	-0.04	0.10	-0.04	0.42	-0.45	-0.32	0.07
CompRes	-0.47	0.23	0.17	-0.22	0.01	0.08	-0.05	0.49	0.02	0.43	-0.46
Advertising	-0.10	0.41	-0.17	0.52	0.54	0.42	-0.11	-0.12	-0.14	0.01	-0.04

Table 12. Principal components Dataframe

Cumulative Values of the variance:

```
array([31.03, 53.27, 68.66, 79.01, 85.15, 89.49, 93.36, 95.79, 97.79,
       99.02, 99.99])
```

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

Fig. 6. Heatmap_PCs vs Features

Heat map represents the correlation between the optimal 4 PC components with the other features available in the dataset.

Optimal number of PC components concluded are 4 which is giving a variance of ~80%.

CLUSTERING:

Part 2: Clustering:

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)
2.2. Do you think scaling is necessary for clustering in this case? Justify
2.3. Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.
2.4. Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and find the silhouette score.
2.5. Describe cluster profiles for the clusters defined. Recommend different priority-based actions that need to be taken for different clusters on the bases of their vulnerability situations according to their Economic and Health Conditions.

Data Dictionary for State_wise_Health_income Dataset:

1. States: names of States
2. Health_indeces1: A composite index rolls several related measures (indicators) into a single score that provides a summary of how the health system is performing in the State.
3. Health_indeces2: A composite index rolls several related measures (indicators) into a single score that provides a summary of how the health system is performing in certain areas of the States.
4. Per_capita_income-Per capita income (PCI) measures the average income earned per person in a given area (city, region, country, etc.) in a specified year. It is calculated by dividing the area's total income by its total population.
5. GDP: GDP provides an economic snapshot of a country/state, used to estimate the size of an economy and growth rate.

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

Data Summary:

We will start analysing the data set by performing the basic steps which are as:

1. Checking the shape.
2. Checking head & tail.
3. Checking summary & info.
4. Checking null values.

5. Checking duplicate values.

Checking the shape:

The dataset has 297 rows & 6 features.

Checking the head & tail:

	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

	Unnamed: 0	States	Health_indeces1	Health_indices2	Per_capita_income	GDP
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

Table 14. Tail_Clustering

Checking summary & info:

The dataset includes 5 integer value and 1 object variable:

	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

Table 15. Summary of Clustering data

Type of Data	No of features
Integer Value	5
Object Values	1

Table 16. Datatypes_Clustering

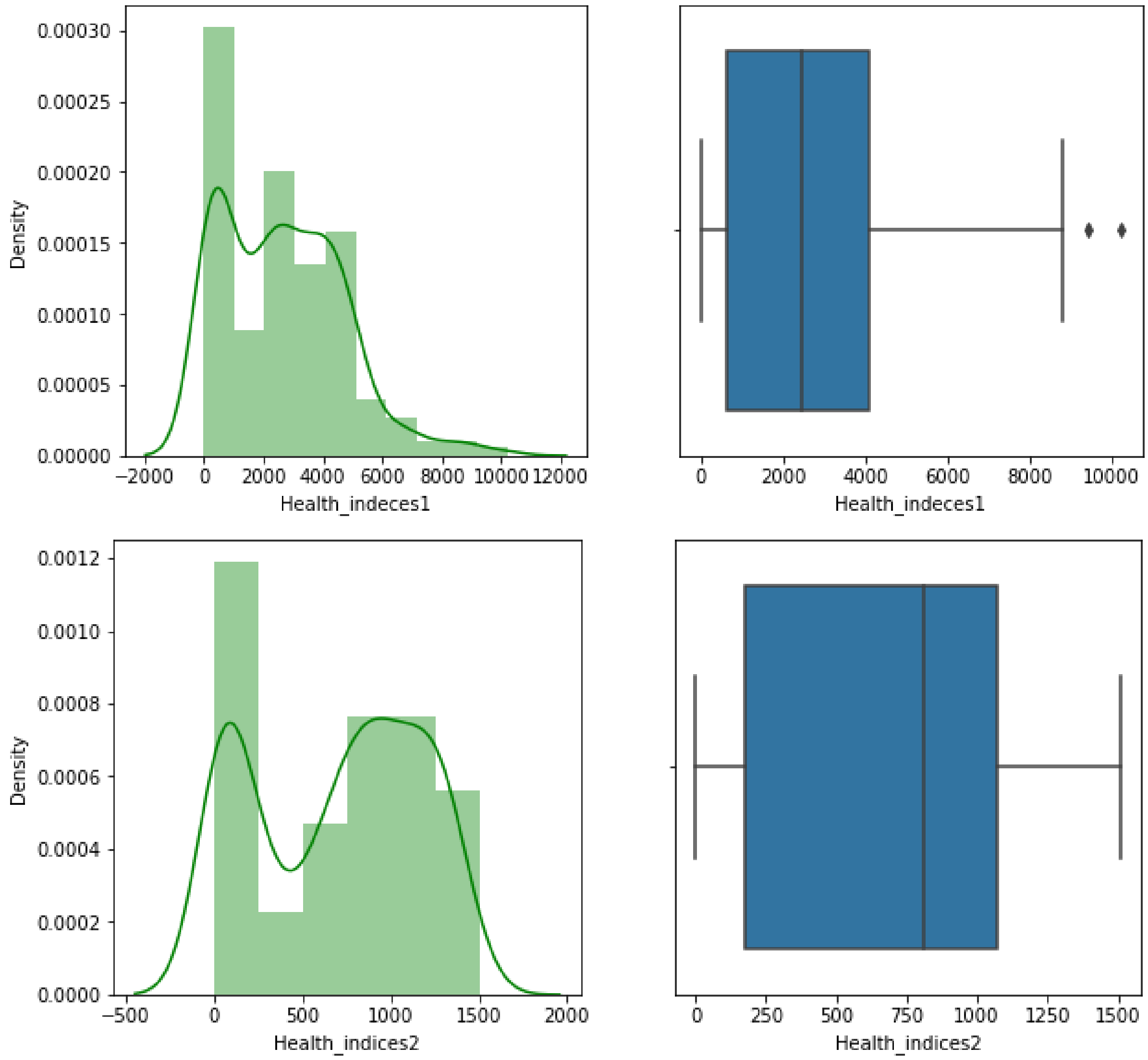
Checking null values:

There are no null (missing) values in the dataset.

Checking duplicate values:

There are no duplicate values in the dataset.

UNIVARIATE ANALYSIS:



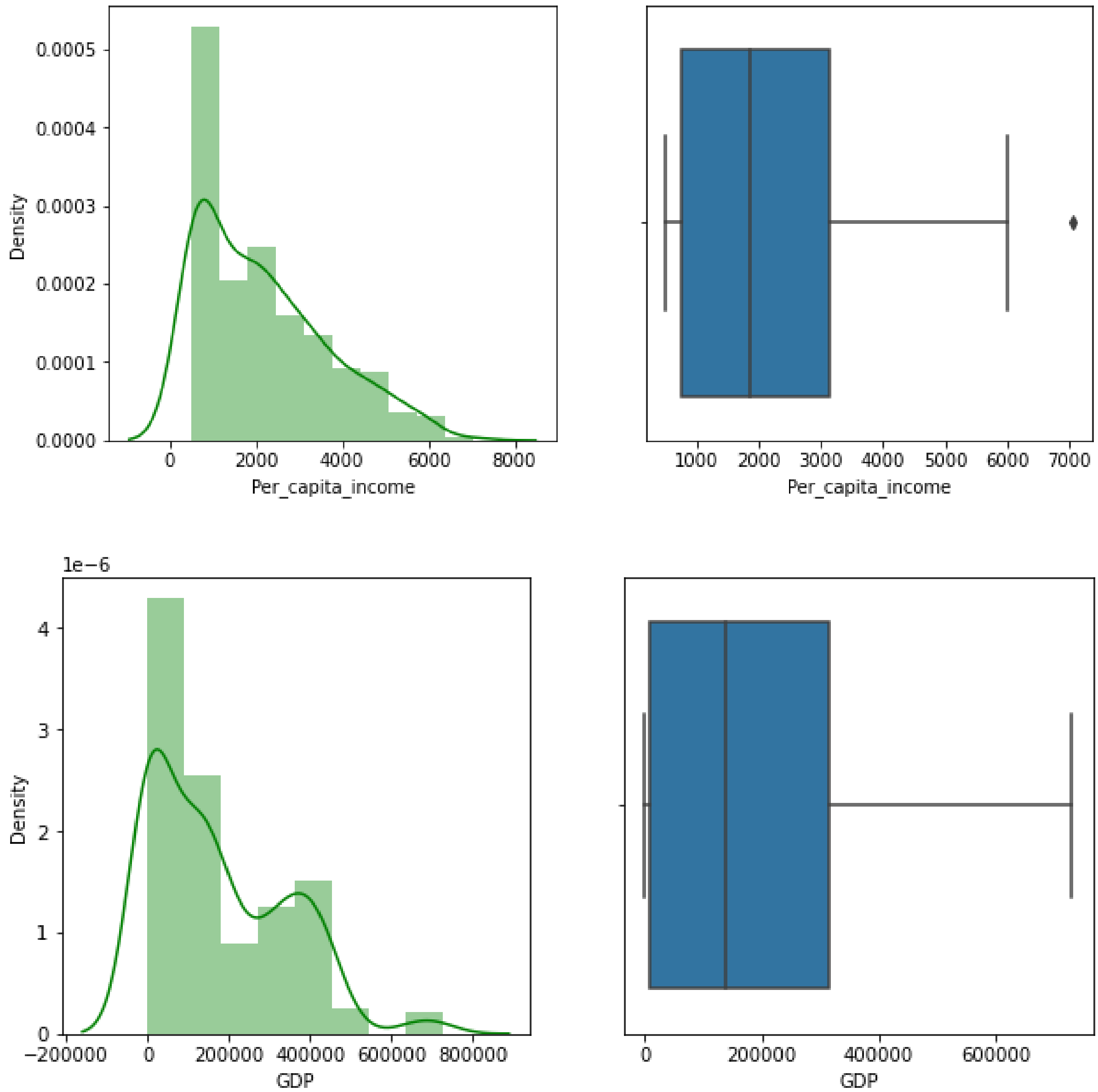


Fig. 7. Univariate analysis

Skewness(Right Skewed) is observed in Health_indices1, Per_capita_income & GDP.

Outliers are observed in Per_capita_income & Health_indices1.

BIVARIATE ANALYSIS:

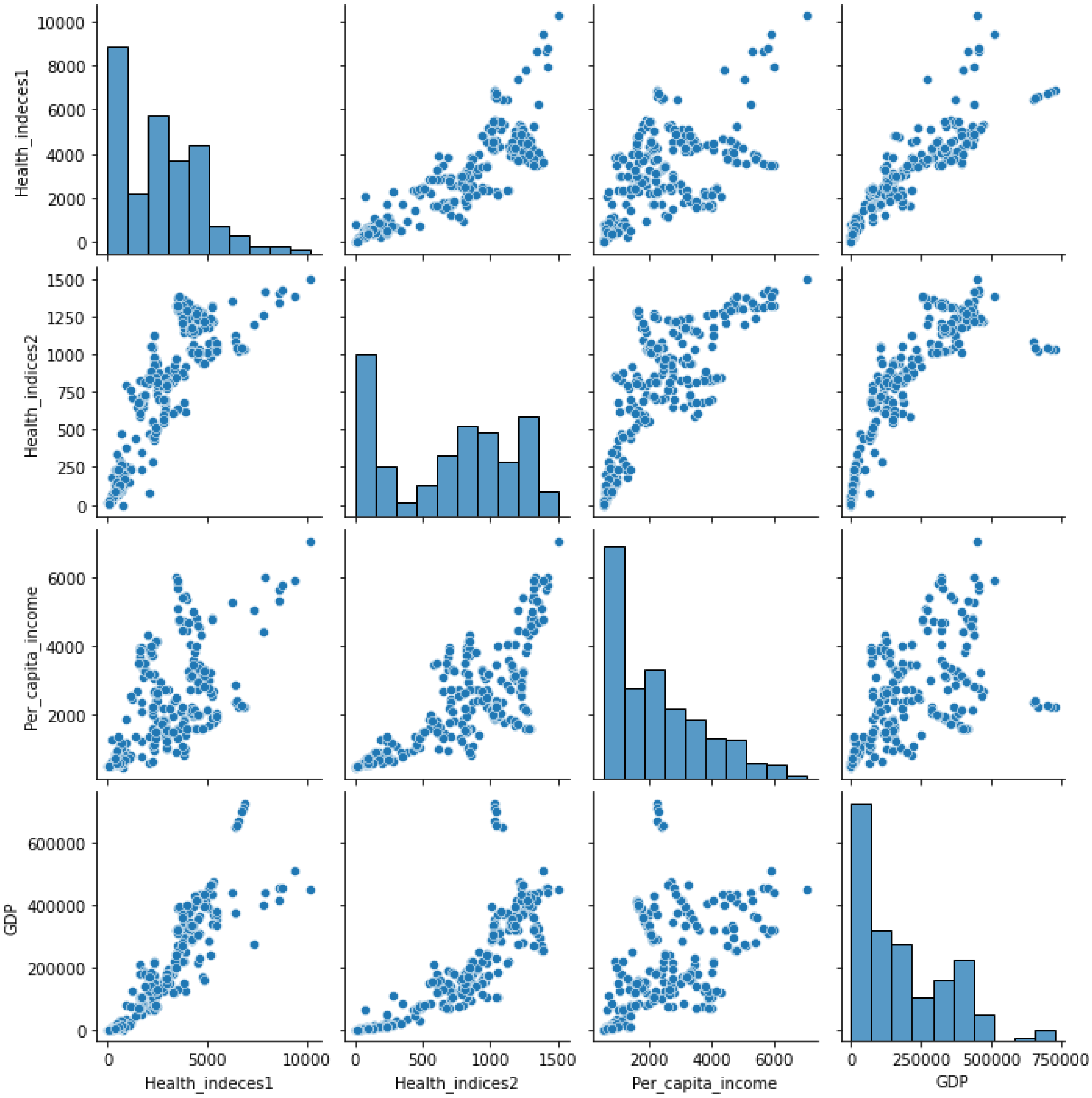


Fig. 8. Pair plot_Bivariate analysis

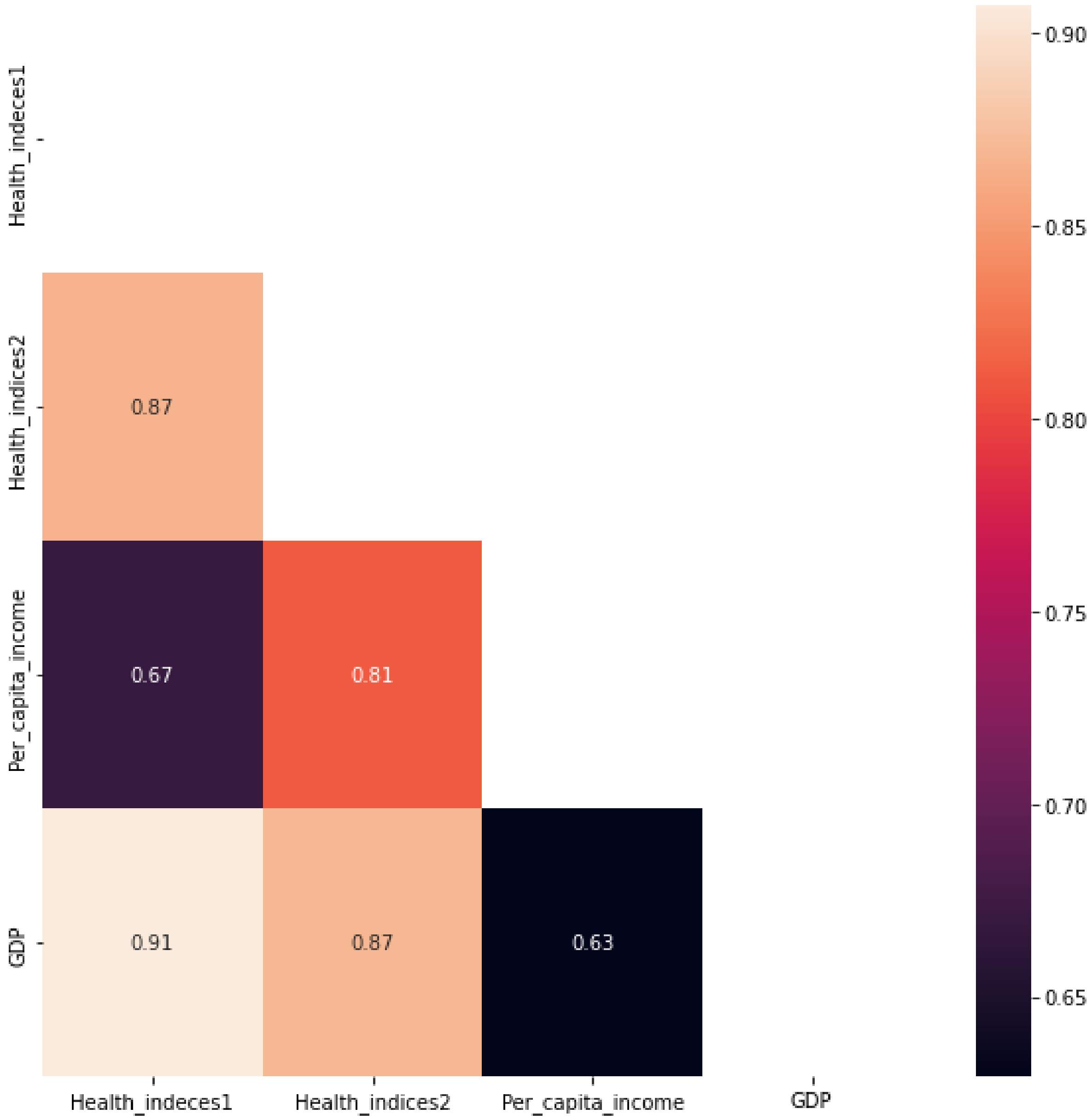


Fig. 9. Heatmap_Bivariate analysis

With reference to the above graphs, we can say that all continuous variables show high correlation among each other.

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

YES, scaling is required in this dataset to avoid prioritising any feature due to heavy weightage.

Scipy is used to scale the dataset. . Zscore method is used which is calculated as

Z=(x-μ)/s

Where,

μ=mean of training samples

s=standard deviation

Result of the dataset after scaling is as follows:

	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	-0.97	-0.84	-0.71	-0.88
3	-1.20	-1.43	-1.06	-1.04
4	-1.27	-1.46	-1.09	-1.05

Table 17. Scaled_clustering Data

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

Clustering is applied on the scaled dataset.

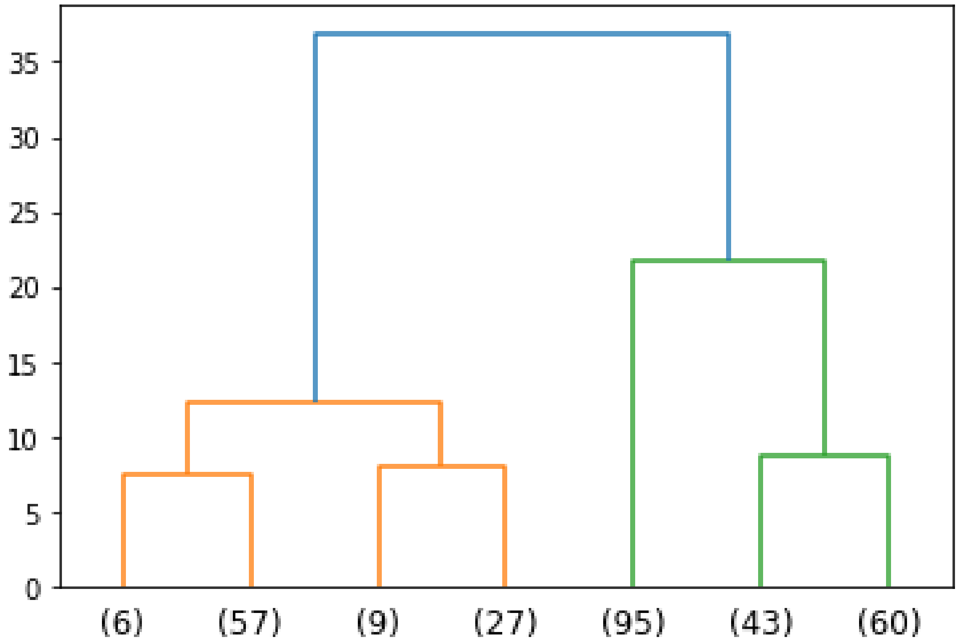


Fig. 10. Dendogram_Hierarchical clustering

2 clusters are not preferred generally to avoid confusion with class distribution among the dataset. To gain more accuracy on insights, segmentation with more than 2 clusters are preferred. The optimum number of clusters which can be considered are 4.

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

K-means clustering was performed on the dataset along with the elbow curve to defined the optimum number of clusters.4 clusters were identified as the optimum number considering the drop in inertia values of each cluster and 0.5520 being the silhouette score.

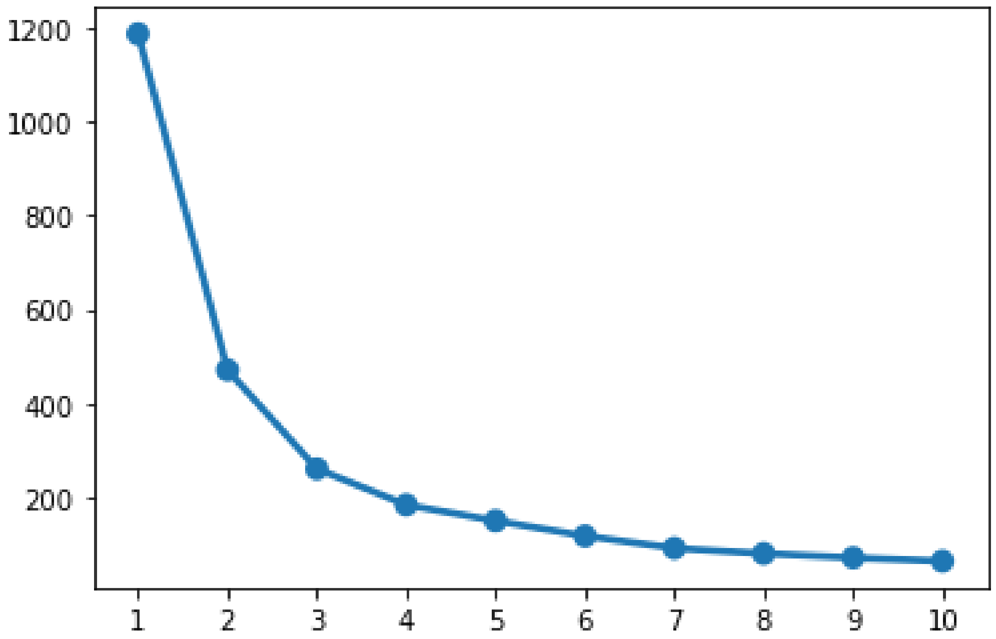


Fig. 11. Elbow Plot

14. 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: Though the GDP is the highest and Health_indeces 2 are low compared to cluster 2 which can be the least prioritized segment in terms of improvisation on health conditions.

CLUSTER 2: The GDP is higher compared to the clusters 3 & 4 with the other features like Health_indeces1 Health_indeces 2 & Per Capita Income also being on the higher bracket making this cluster the 2nd least prioritized segment to be focussed on in terms on health conditions.

CLUSTER 3: Cluster 3 is a red flag with least Health_indeces1, Health_indeces2, Per Capita Income & GDP. Hence Priority should be given to cluster 3 for improving its economy and health condition

CLUSTER 4: With improved Health_indeces1, Health_indeces2, Per Capita Inco GDP compared to cluster 3, this cluster should be the second priority for improving on economy and health conditions.