Data Mining

PROJECT REPORT

HARI HARAN

16th May, 2023

CONTENTS	PAGE

`

FIGURES AND TABLES	PAGE

Part 1: PCA:

Problem Statement: The ‘Hair Salon.csv’ dataset contains various variables used for the context of Market Segmentation. This particular case study is based on various parameters of a salon chain of hair products. You are expected to do Principal Component Analysis for this case study according to the instructions given in the rubric. Kindly refer to the PCA_Data_Dictionary.jpg file for the Data Dictionary of the Dataset.

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

Q.1.1. Perform Exploratory Data Analysis [both Univariate and Multivariate analysis to be performed]. The inferences drawn from this should be properly documented.

Performing EDA for the given data:

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

```
ID                int64
ProdQual        float64
Ecom            float64
TechSup         float64
CompRes         float64
Advertising     float64
ProdLine        float64
SalesFImage     float64
ComPricing      float64
WartyClaim      float64
OrdBilling      float64
DelSpeed        float64
Satisfaction    float64
dtype: object
```

Table - 1.1. Data Info and Data Types.

There are no missing values and duplicate values in the given data. The data has **100 rows** and **13 columns**. There 12 float data types and 1 integer data type.

	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 - 1.2. Data Description

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 - 1.3. Data- Variable Description

- There are 100 different product IDs.
- The ratings of each variable is between 0-10.
- The minimum satisfaction rating is 4.7 and the maximum is 9.9.
- The 75% of products have 7.6 satisfaction ration.
- E-commerce and Delivery Speed has lowest rating compared to other variables.

Since, **Satisfaction** is a target variable which we will drop for PCA.

Univariate and Multivariate Analysis :

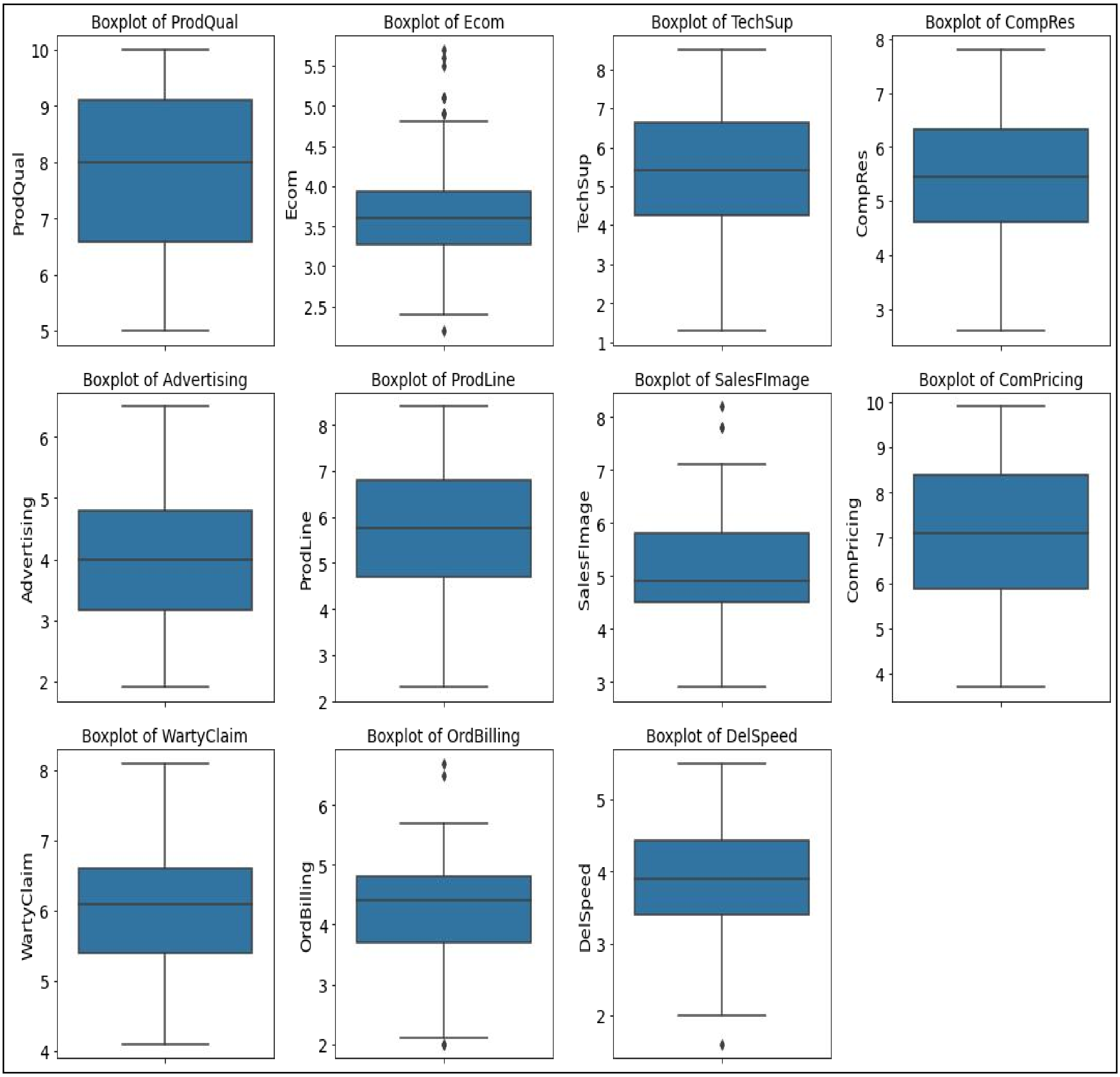

Fig.1.1. Boxplot

Based on th above boxplot, there are few outliers in the given data, which maybe ignored for further analysis.

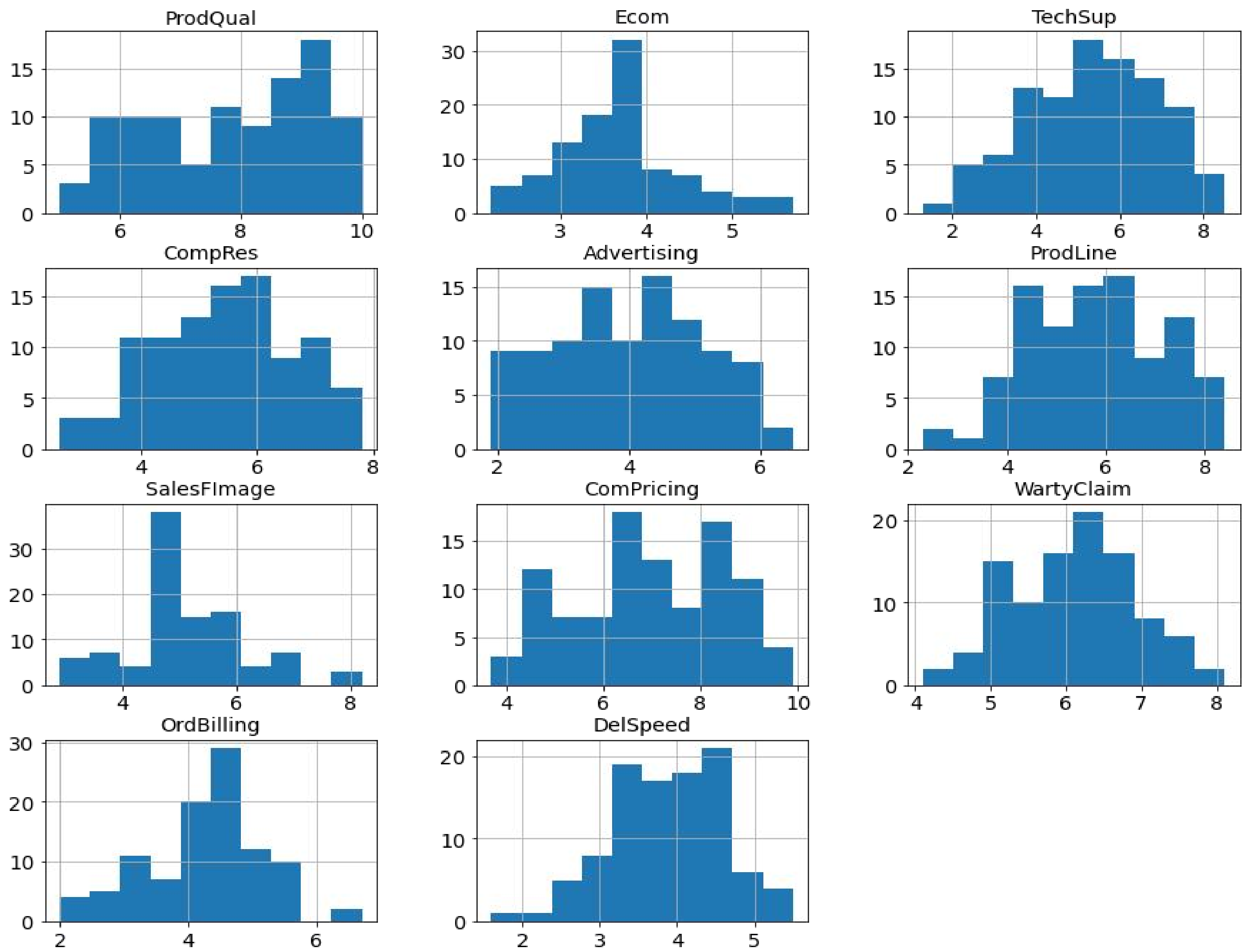

Fig.1.2. Histogram

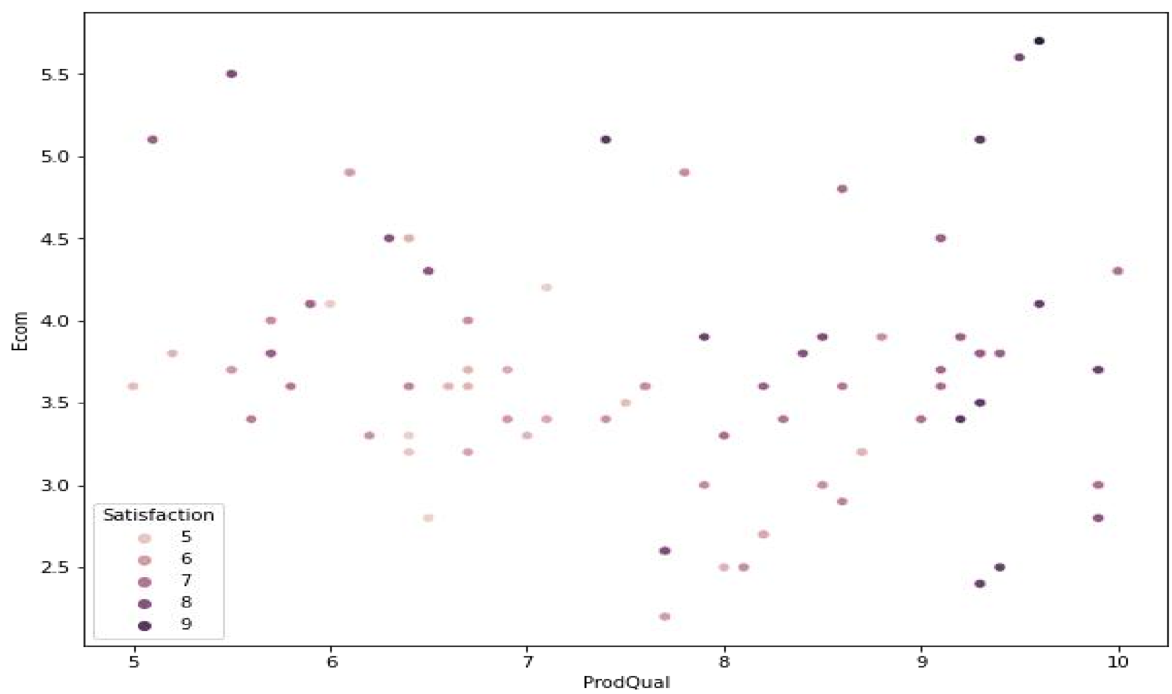

Fig.1.3. Scatter Plot - Product Quality Vs ECom (hue = Satisfaction)

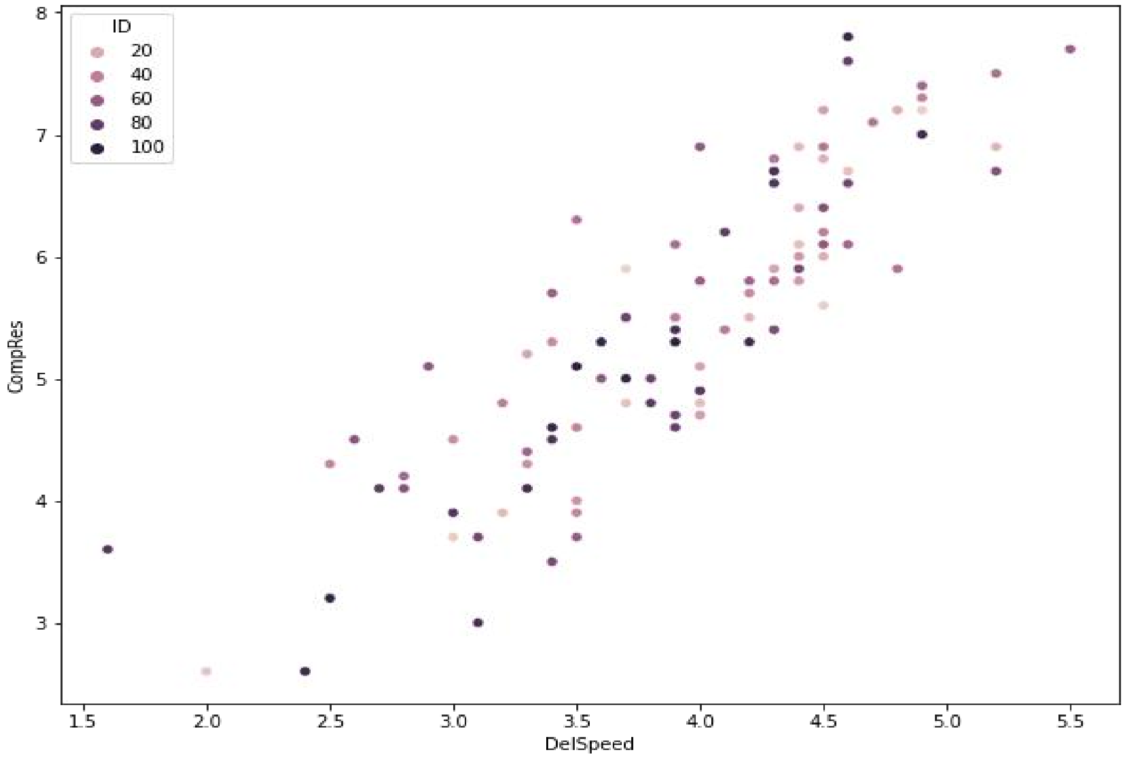

Fig.1.4. Scatter Plot - Delivery Speed Vs Complaint Resolution (hue = ID)

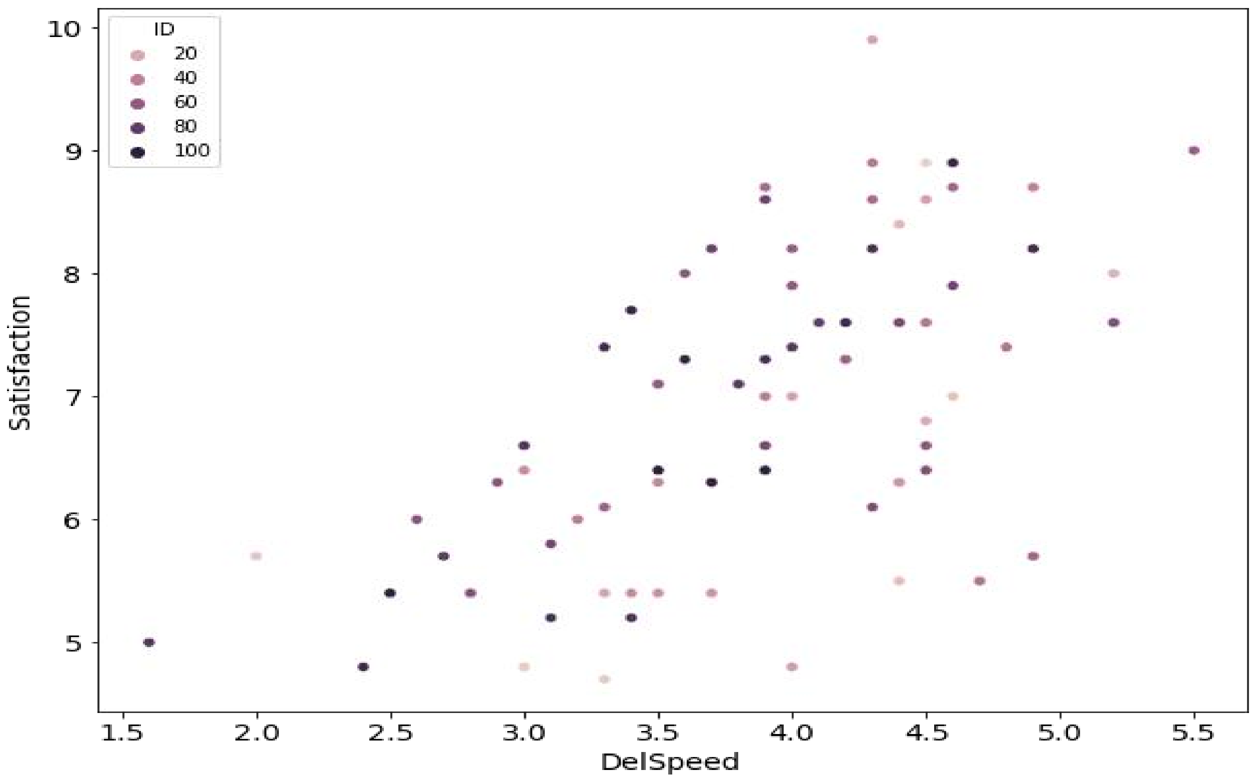

Fig.1.5. Scatter Plot - Delivery Speed Vs Satisfaction. (hue = ID)

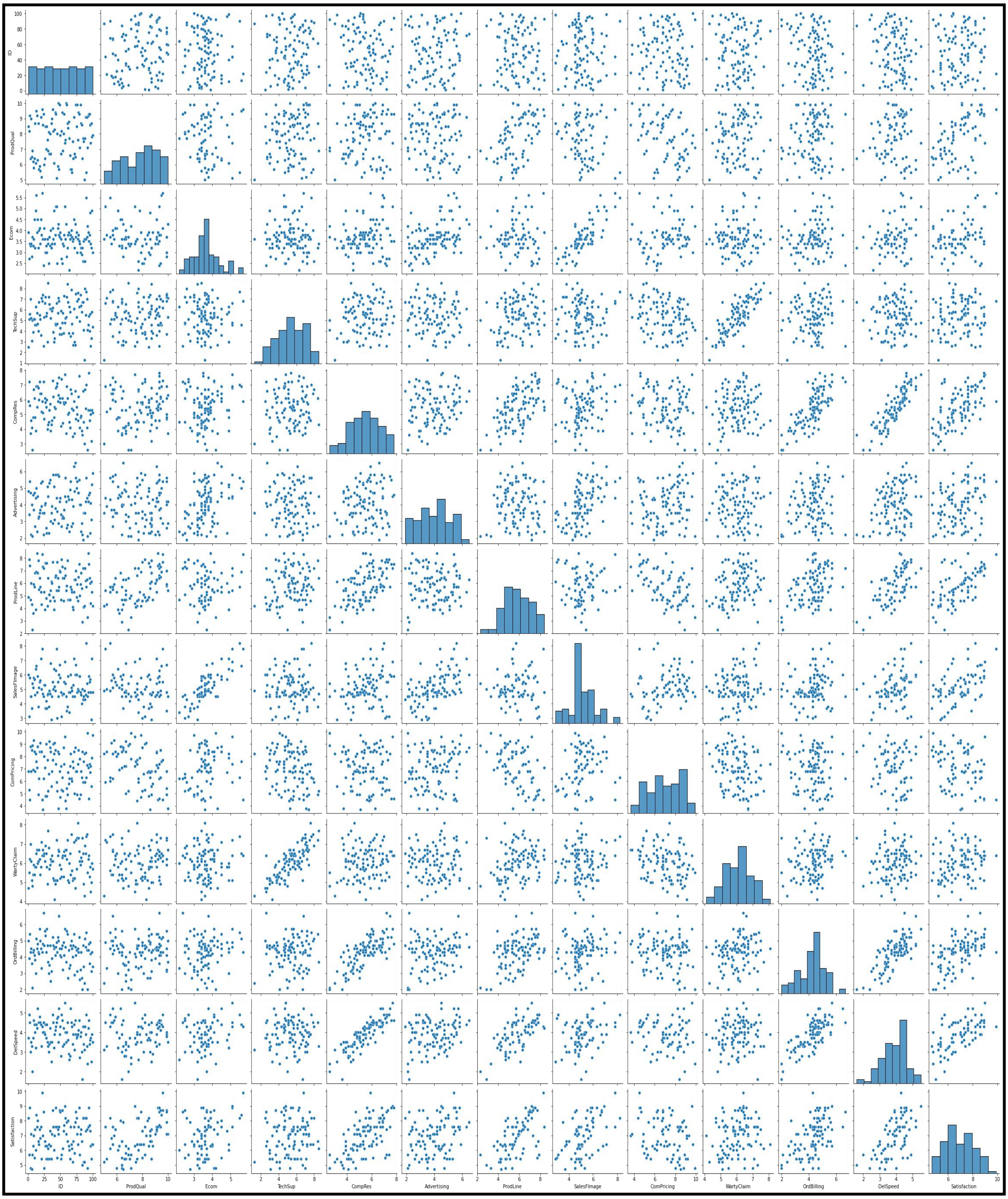

Fig.1.6. Pair plot (Relation between all the Variables)

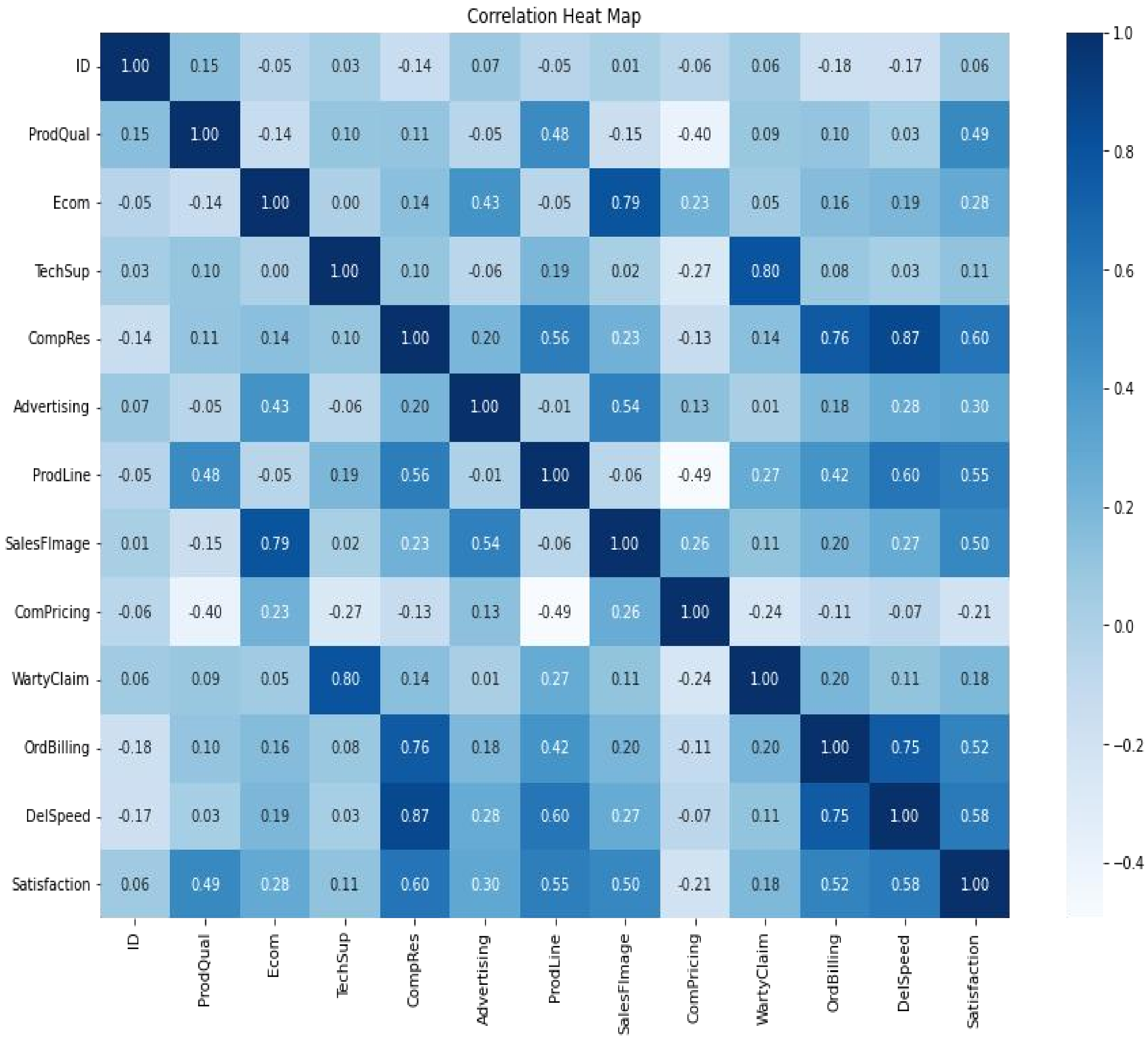

Fig.1.7. Correlation Heat map between all the Variables.

Inferences:

- There is good positive correlation between **Delivery speed** and **Complaint Resolution**. That means consumers are satisfied with after sales services of the product and the issues are resolved in time.
- Ecom and Salesforce are positively co-related.
- Ecom, Advertising and Product line are negatively co-related. The business need to invest more time and resources their advertising of product lines.

There are lot of other plots can be compared between each of the components present in the dataset, but our primary objective is PCA.

Q.1.2. Scale the variables and write the inference for using the type of scaling function for this case study.

Applying **Z-score** method for **scaling**,

Scaling is the process of standardization of data to transform the data in such a way that it will have a **mean 0** and **standard deviation 1**. Here we have used **Z-score method** to standardize the 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

Table - 1.4. Data Description - Scaled Data

Q.1.3.Comment on the comparison between covariance and the correlation matrix after scaling.

Before we apply PCA, scaling is performed on a given dataset so that each feature will have a variance equal to 1 and a mean of 0 and they contribute equally to the analysis. If the features will have **large differences** in variances then the features having the **largest variance will dominate** over other features having a smaller variance which will lead to a biased result. Therefore standardization is performed **before performing PCA**.

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
ProdQual	1.000	-0.137	0.096	0.106	-0.053	0.477	-0.152	-0.401	0.088	0.104	0.028
Ecom	-0.137	1.000	0.001	0.140	0.430	-0.053	0.792	0.229	0.052	0.156	0.192
TechSup	0.096	0.001	1.000	0.097	-0.063	0.193	0.017	-0.271	0.797	0.080	0.025
CompRes	0.106	0.140	0.097	1.000	0.197	0.561	0.230	-0.128	0.140	0.757	0.865
Advertising	-0.053	0.430	-0.063	0.197	1.000	-0.012	0.542	0.134	0.011	0.184	0.276
ProdLine	0.477	-0.053	0.193	0.561	-0.012	1.000	-0.061	-0.495	0.273	0.424	0.602
SalesFImage	-0.152	0.792	0.017	0.230	0.542	-0.061	1.000	0.265	0.107	0.195	0.272
ComPricing	-0.401	0.229	-0.271	-0.128	0.134	-0.495	0.265	1.000	-0.245	-0.115	-0.073
WartyClaim	0.088	0.052	0.797	0.140	0.011	0.273	0.107	-0.245	1.000	0.197	0.109
OrdBilling	0.104	0.156	0.080	0.757	0.184	0.424	0.195	-0.115	0.197	1.000	0.751
DelSpeed	0.028	0.192	0.025	0.865	0.276	0.602	0.272	-0.073	0.109	0.751	1.000

Table - 1.5. Correlation Table without Scaling

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
ProdQual	1.000	-0.137	0.096	0.106	-0.053	0.477	-0.152	-0.401	0.088	0.104	0.028
Ecom	-0.137	1.000	0.001	0.140	0.430	-0.053	0.792	0.229	0.052	0.156	0.192
TechSup	0.096	0.001	1.000	0.097	-0.063	0.193	0.017	-0.271	0.797	0.080	0.025
CompRes	0.106	0.140	0.097	1.000	0.197	0.561	0.230	-0.128	0.140	0.757	0.865
Advertising	-0.053	0.430	-0.063	0.197	1.000	-0.012	0.542	0.134	0.011	0.184	0.276
ProdLine	0.477	-0.053	0.193	0.561	-0.012	1.000	-0.061	-0.495	0.273	0.424	0.602
SalesFImage	-0.152	0.792	0.017	0.230	0.542	-0.061	1.000	0.265	0.107	0.195	0.272
ComPricing	-0.401	0.229	-0.271	-0.128	0.134	-0.495	0.265	1.000	-0.245	-0.115	-0.073
WartyClaim	0.088	0.052	0.797	0.140	0.011	0.273	0.107	-0.245	1.000	0.197	0.109
OrdBilling	0.104	0.156	0.080	0.757	0.184	0.424	0.195	-0.115	0.197	1.000	0.751
DelSpeed	0.028	0.192	0.025	0.865	0.276	0.602	0.272	-0.073	0.109	0.751	1.000

Table - 1.6. Correlation Table after Scaling

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
ProdQual	1.950	-0.134	0.204	0.179	-0.084	0.877	-0.227	-0.866	0.101	0.135	0.028
Ecom	-0.134	0.491	0.001	0.119	0.339	-0.049	0.595	0.248	0.030	0.102	0.099
TechSup	0.204	0.001	2.342	0.179	-0.108	0.388	0.028	-0.640	1.000	0.114	0.029
CompRes	0.179	0.119	0.179	1.460	0.268	0.892	0.298	-0.239	0.139	0.850	0.768
Advertising	-0.084	0.339	-0.108	0.268	1.270	-0.017	0.655	0.234	0.010	0.193	0.228
ProdLine	0.877	-0.049	0.388	0.892	-0.017	1.730	-0.086	-1.006	0.294	0.518	0.581
SalesFImage	-0.227	0.595	0.028	0.298	0.655	-0.086	1.150	0.438	0.094	0.194	0.214
ComPricing	-0.866	0.248	-0.640	-0.239	0.234	-1.006	0.438	2.387	-0.310	-0.164	-0.083
WartyClaim	0.101	0.030	1.000	0.139	0.010	0.294	0.094	-0.310	0.672	0.150	0.066
OrdBilling	0.135	0.102	0.114	0.850	0.193	0.518	0.194	-0.164	0.150	0.863	0.512
DelSpeed	0.028	0.099	0.029	0.768	0.228	0.581	0.214	-0.083	0.066	0.512	0.539

Table - 1.7. Covariance Table without scaling

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
ProdQual	1.010	-0.139	0.097	0.107	-0.054	0.482	-0.153	-0.405	0.089	0.105	0.028
Ecom	-0.139	1.010	0.001	0.142	0.434	-0.053	0.800	0.232	0.052	0.158	0.194
TechSup	0.097	0.001	1.010	0.098	-0.064	0.195	0.017	-0.274	0.805	0.081	0.026
CompRes	0.107	0.142	0.098	1.010	0.199	0.567	0.232	-0.129	0.142	0.765	0.874
Advertising	-0.054	0.434	-0.064	0.199	1.010	-0.012	0.548	0.136	0.011	0.186	0.279
ProdLine	0.482	-0.053	0.195	0.567	-0.012	1.010	-0.062	-0.500	0.276	0.429	0.608
SalesFImage	-0.153	0.800	0.017	0.232	0.548	-0.062	1.010	0.267	0.109	0.197	0.274
ComPricing	-0.405	0.232	-0.274	-0.129	0.136	-0.500	0.267	1.010	-0.247	-0.116	-0.074
WartyClaim	0.089	0.052	0.805	0.142	0.011	0.276	0.109	-0.247	1.010	0.199	0.110
OrdBilling	0.105	0.158	0.081	0.765	0.186	0.429	0.197	-0.116	0.199	1.010	0.759
DelSpeed	0.028	0.194	0.026	0.874	0.279	0.608	0.274	-0.074	0.110	0.759	1.010

Table - 1.8. Covariance Table after scaling

Q.1.4.Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.

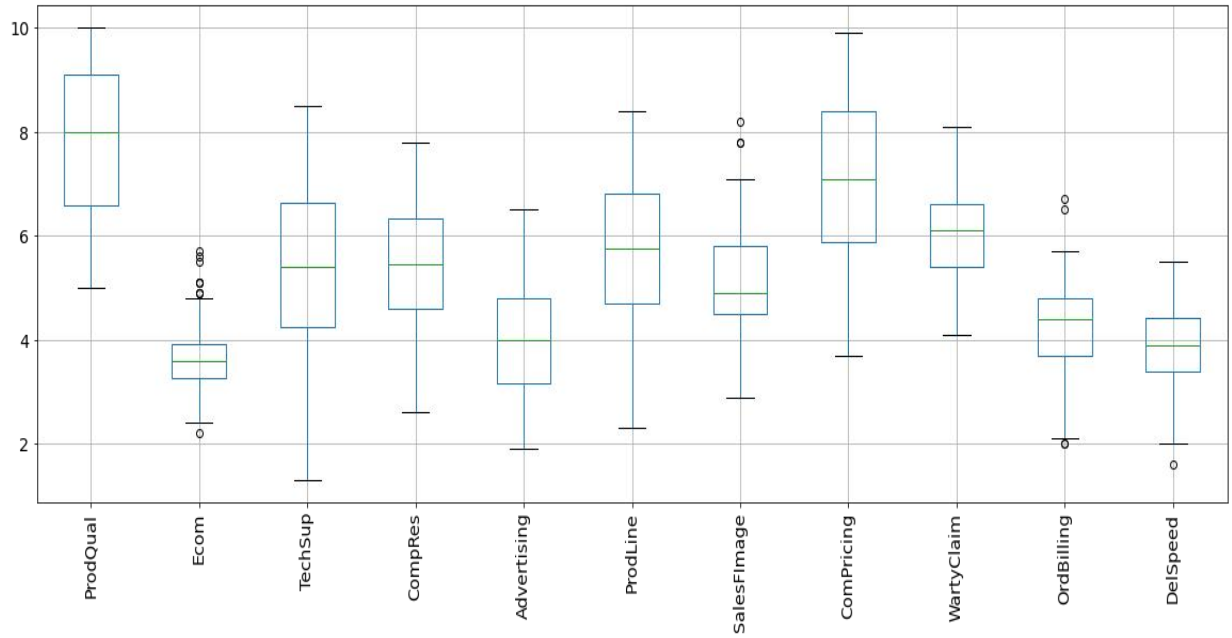

Fig.1.8. Boxplot before Scaling.

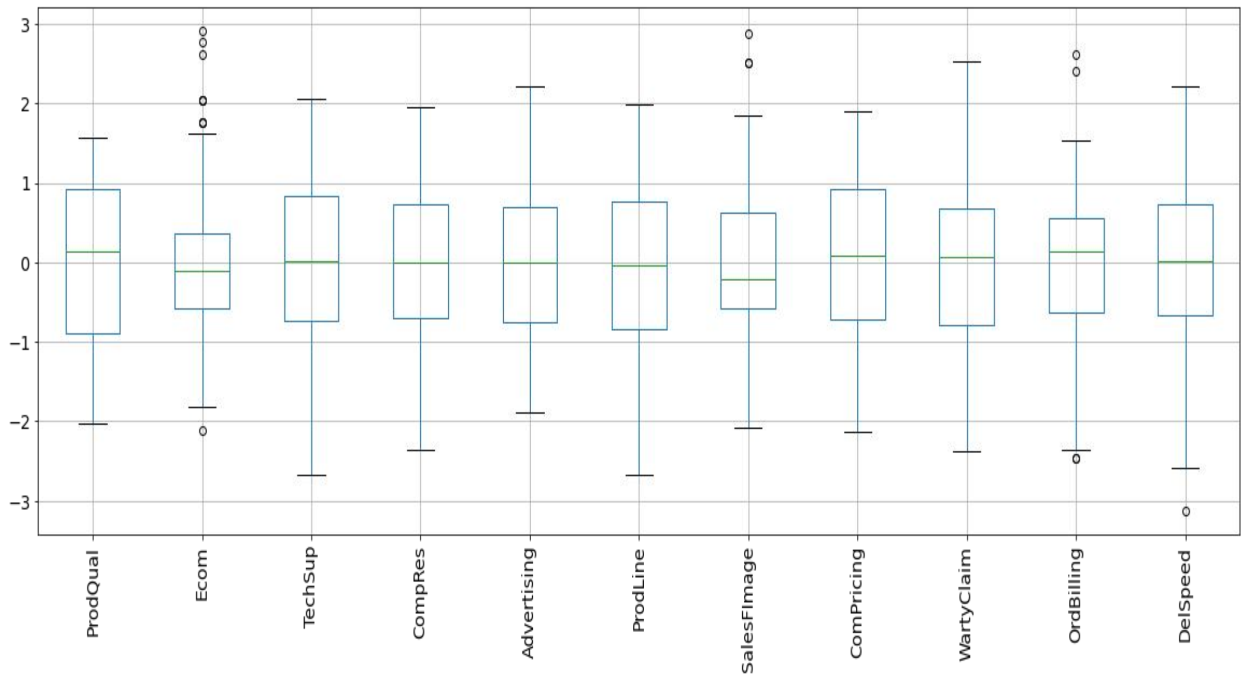

Fig.1.9. Boxplot After Scaling.

Based on the above plots, we can see that the scaling has helped us to standardize the variables with outliers.

Q.1.5. Build the covariance matrix, eigenvalues and eigenvector.

	ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
ID	1.010	0.147	-0.047	0.032	-0.146	0.074	-0.049	0.014	-0.064	0.059	-0.180	-0.174	0.062
ProdQual	0.147	1.010	-0.139	0.097	0.107	-0.054	0.482	-0.153	-0.405	0.089	0.105	0.028	0.491
Ecom	-0.047	-0.139	1.010	0.001	0.142	0.434	-0.053	0.800	0.232	0.052	0.158	0.194	0.286
TechSup	0.032	0.097	0.001	1.010	0.098	-0.064	0.195	0.017	-0.274	0.805	0.081	0.026	0.114
CompRes	-0.146	0.107	0.142	0.098	1.010	0.199	0.567	0.232	-0.129	0.142	0.765	0.874	0.609
Advertising	0.074	-0.054	0.434	-0.064	0.199	1.010	-0.012	0.548	0.136	0.011	0.186	0.279	0.308
ProdLine	-0.049	0.482	-0.053	0.195	0.567	-0.012	1.010	-0.062	-0.500	0.276	0.429	0.608	0.556
SalesFImage	0.014	-0.153	0.800	0.017	0.232	0.548	-0.062	1.010	0.267	0.109	0.197	0.274	0.505
ComPricing	-0.064	-0.405	0.232	-0.274	-0.129	0.136	-0.500	0.267	1.010	-0.247	-0.116	-0.074	-0.210
WartyClaim	0.059	0.089	0.052	0.805	0.142	0.011	0.276	0.109	-0.247	1.010	0.199	0.110	0.179
OrdBilling	-0.180	0.105	0.158	0.081	0.765	0.186	0.429	0.197	-0.116	0.199	1.010	0.759	0.527
DelSpeed	-0.174	0.028	0.194	0.026	0.874	0.279	0.608	0.274	-0.074	0.110	0.759	1.010	0.583
Satisfaction	0.062	0.491	0.286	0.114	0.609	0.308	0.556	0.505	-0.210	0.179	0.527	0.583	1.010

Table - 1.9. Covariance Matrix after scaling

Eigen Values :

```
Eigenvalues: [3.46 2.58 1.71 1.1  0.62 0.56 0.41 0.25 0.21 0.13 0.1 ]
```

Eigen Vectors :

```
Eigenvectors: [[-0.13 -0.17 -0.16 -0.47 -0.18 -0.39 -0.2   0.15 -0.21 -0.44 -0.47]
 [-0.31  0.45 -0.23  0.02  0.36 -0.28  0.47  0.41 -0.19  0.03  0.07]
 [ 0.06 -0.24 -0.61  0.21 -0.09  0.12 -0.24  0.05 -0.6   0.17  0.23]
 [ 0.64  0.27 -0.19 -0.21  0.32  0.2   0.22 -0.33 -0.19 -0.24 -0.2 ]
 [ 0.23  0.42 -0.02  0.03 -0.8   0.12  0.2   0.25 -0.03  0.03 -0.04]
 [-0.56  0.26 -0.11 -0.03 -0.2   0.1   0.1  -0.71 -0.14 -0.12  0.03]
 [ 0.19  0.06 -0.02 -0.01 -0.06 -0.61  0.   -0.31 -0.03  0.66 -0.23]
 [ 0.14 -0.12  0.46  0.51 -0.05 -0.33  0.17 -0.1  -0.44 -0.37  0.07]
 [ 0.03 -0.54 -0.36  0.09 -0.15 -0.08  0.64 -0.09  0.32 -0.1  -0.02]
 [ 0.07  0.28 -0.39  0.53  0.04 -0.23 -0.35 -0.05  0.44 -0.3  -0.12]
 [ 0.18  0.06 -0.05 -0.36 -0.08 -0.39 -0.08 -0.1   0.13 -0.19  0.78]]
```

Q.1.6. Write the explicit form of the first PC (in terms of Eigen Vectors).

Liner equation of the First PC along with scaled variables.

```
In [257]: for i in range(0,11):
              print("(",np.round(pca.components_[0][i],2),")",'*',hsdf_scaled.columns[i], end=' + ')

( -0.13 ) * ProdQual + ( -0.17 ) * Ecom + ( -0.16 ) * TechSup + ( -0.47 ) * CompRes + ( -0.18 ) * Advertising + ( -0.39 ) * Pro
dLine + ( -0.2 ) * SalesFImage + ( 0.15 ) * ComPricing + ( -0.21 ) * WartyClaim + ( -0.44 ) * OrdBilling + ( -0.47 ) * DelSpeed
+
```

Q.1.7.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.

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
PC0	-0.130000	-0.170000	-0.160000	-0.470000	-0.180000	-0.390000	-0.200000	0.150000	-0.210000	-0.440000	-0.470000
PC1	-0.310000	0.450000	-0.230000	0.020000	0.360000	-0.280000	0.470000	0.410000	-0.190000	0.030000	0.070000
PC2	0.060000	-0.240000	-0.610000	0.210000	-0.090000	0.120000	-0.240000	0.050000	-0.600000	0.170000	0.230000
PC3	0.640000	0.270000	-0.190000	-0.210000	0.320000	0.200000	0.220000	-0.330000	-0.190000	-0.240000	-0.200000
PC4	0.230000	0.420000	-0.020000	0.030000	-0.800000	0.120000	0.200000	0.250000	-0.030000	0.030000	-0.040000
PC5	-0.560000	0.260000	-0.110000	-0.030000	-0.200000	0.100000	0.100000	-0.710000	-0.140000	-0.120000	0.030000
PC6	0.190000	0.060000	-0.020000	-0.010000	-0.060000	-0.610000	0.000000	-0.310000	-0.030000	0.660000	-0.230000
PC7	0.140000	-0.120000	0.460000	0.510000	-0.050000	-0.330000	0.170000	-0.100000	-0.440000	-0.370000	0.070000

Table - 1.10. Principal Components Table

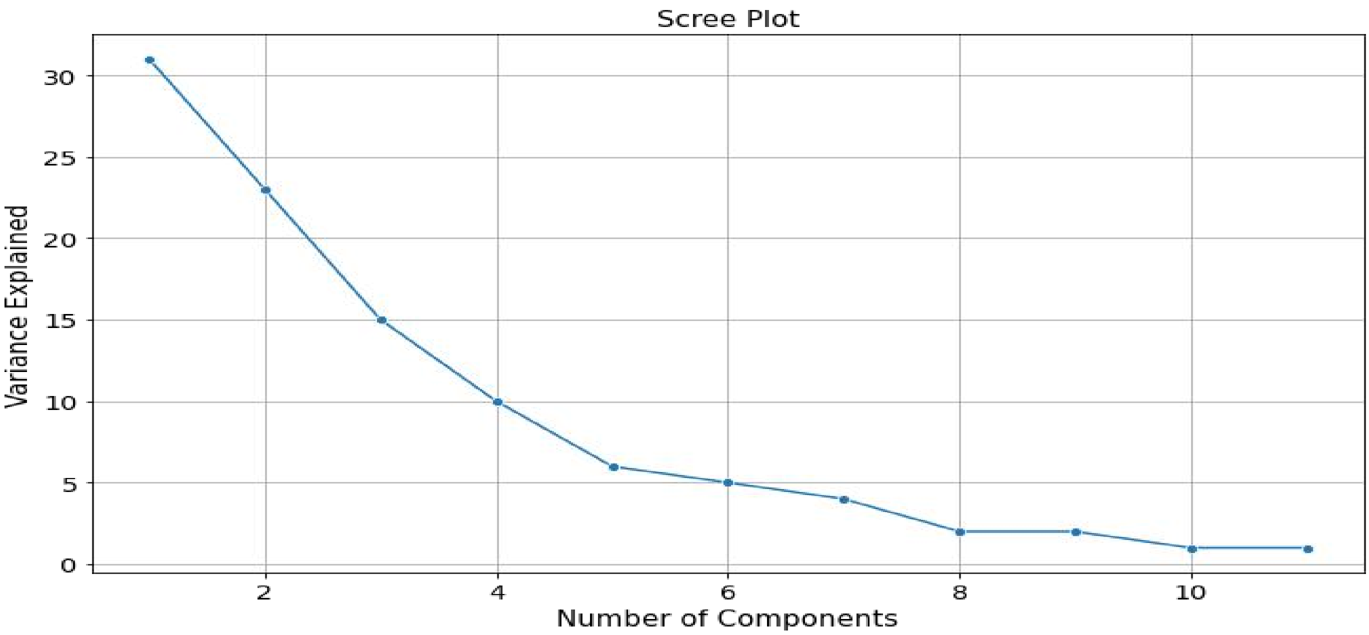

Fig.1.10. Scree Plot

Observations :

- The First Principal component (PC0) is negatively correlated with all the features of the data set, except marginal correlation with ComPricing. These variables explain the most variance in the data i.e. 31%.

- The Second PC (PC1) is strongly and positively correlated with E-commerce, Advertising, Salesforce Image and Competitor Pricing.The Second PC explains about 23% of variation in the data.

- The Third PC (PC2) explains about 15% variation in the data. It is strongly (positively)correlates with Delivery Speed. It is marginally correlated with Product Quality, Complaint Resolution, Product Line, Competitor Pricing and Order Billing.

- The Fourth PC (PC3) correlated positively with Product Quality . It explains about 10% of variation in the data.

- The Fifth PC (PC4) explains about 6% variation in data. It is negatively correlation with Warranty Claim. But it is postively correlated with ProdQual, Ecom, SalesFimage, ComPricing. It is marginally correlated with CompRes, ProdLine and OrdBilling.

- The Sixth PC (PC5) explains about 5% variation in data. It has a good correlation with Ecom compared to other variables in the data.

- The Seventh PC (PC6) explains about 4% variation in data. It is positively correlated with Female Marginal Other workers (0-3,3-6), Main & Marginal Households Female population.

- The Eighth PC(PC7) explains about 2% variation in data. It is positively correlated with Female Marginal Other workers (0-3,3-6), Main & Marginal Households Female population.

- Overall the first 8 PCs contributes to 96% variation in the data. Each of these PCs correlates with the different variables explaining how other features of the data influences the variation in data set.

Part 2: Clustering:

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

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 297 entries, 0 to 296
Data columns (total 6 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Unnamed: 0         297 non-null    int64
 1   States             297 non-null    object
 2   Health_indeces1    297 non-null    int64
 3   Health_indices2    297 non-null    int64
 4   Per_capita_income  297 non-null    int64
 5   GDP                297 non-null    int64
dtypes: int64(5), object(1)
memory usage: 14.0+ KB
```

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

Table - 2.1. Data Info and Data Types.

	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 - 2.2. Data Description

There no missing (null values) or duplicate values in the data set. The data consists of **297 rows** and **6 columns** .

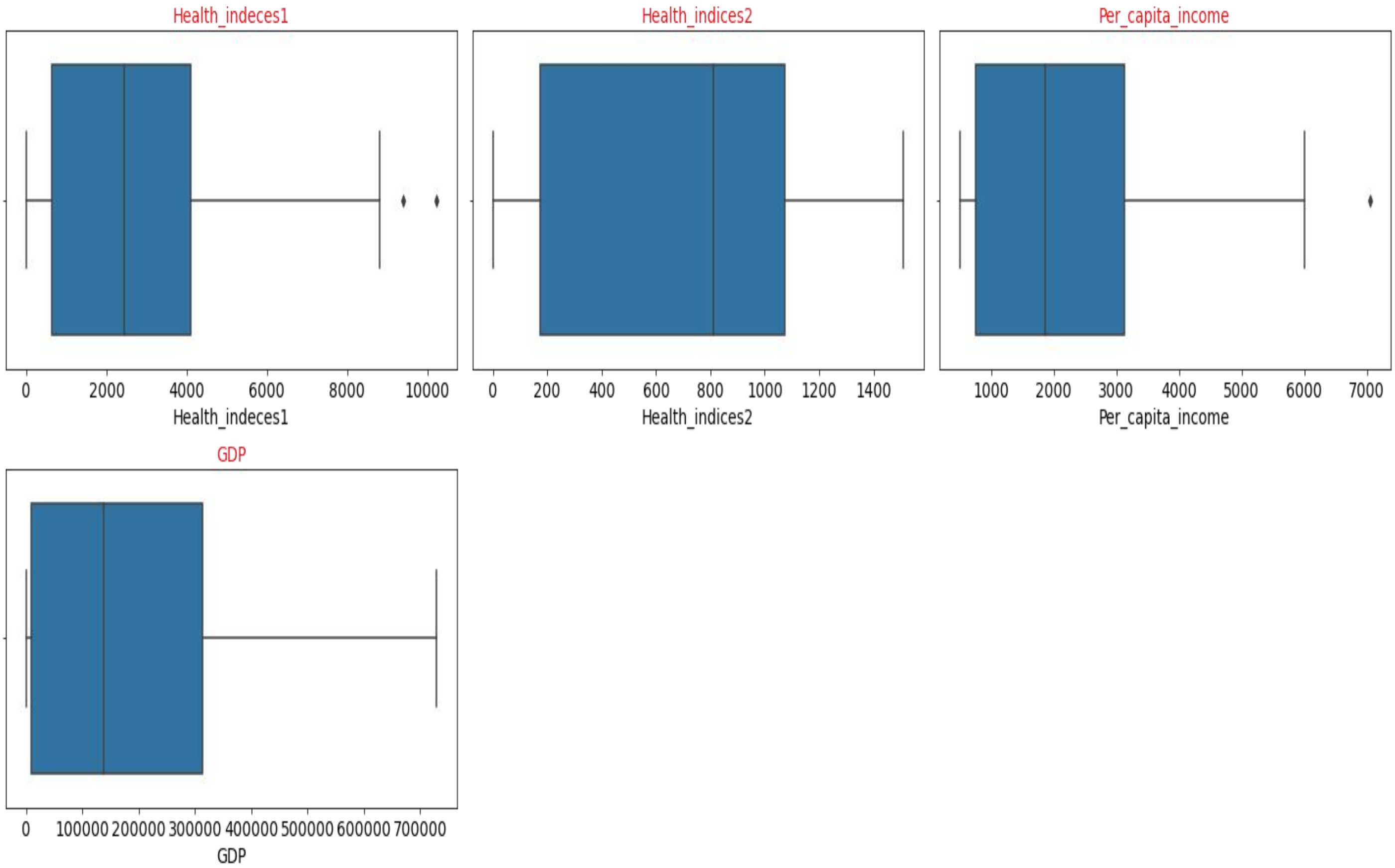

Fig.2.1. Box Plots

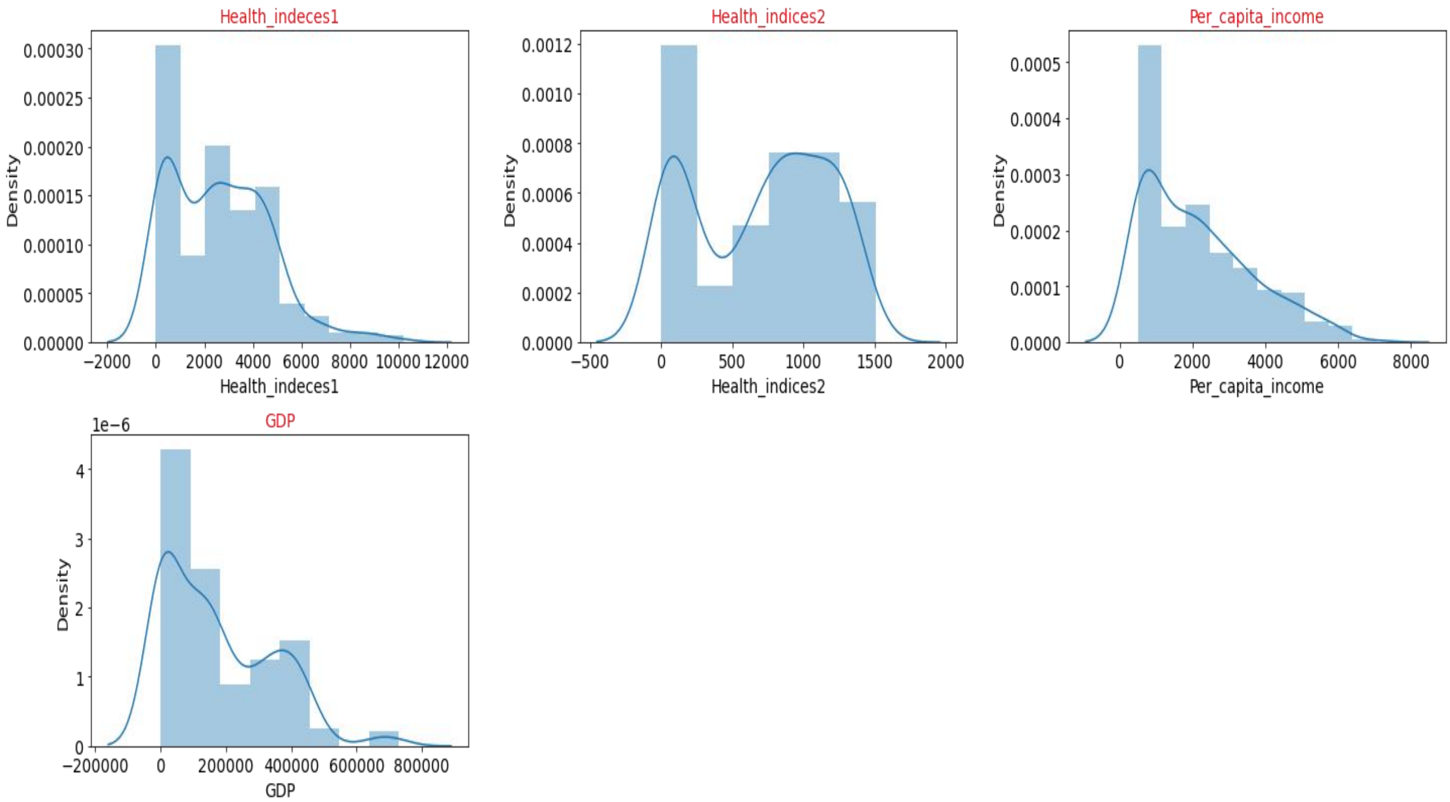

Fig.2.2. Distribution Plots

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

Scaling of variables is important for clustering to stabilize the weights of the different variables. If there is wide discrepancy (or difference) in the range of variables cluster formation may be affected by weight differential. We can observe in the **Table 2.3** before scaling, there are discrepancies in the range of variables.

We are using **Standard Scaler** Method from SKlearn to perform scaling. The below table shows data before and after scaling.

	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 - 2.3. Data Description (Before Scaling)

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

Table - 2.4. Data Description (After Scaling)

The data is standardized after scaling and we can proceed further for clustering.

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

Applying Hierarchical Clustering,

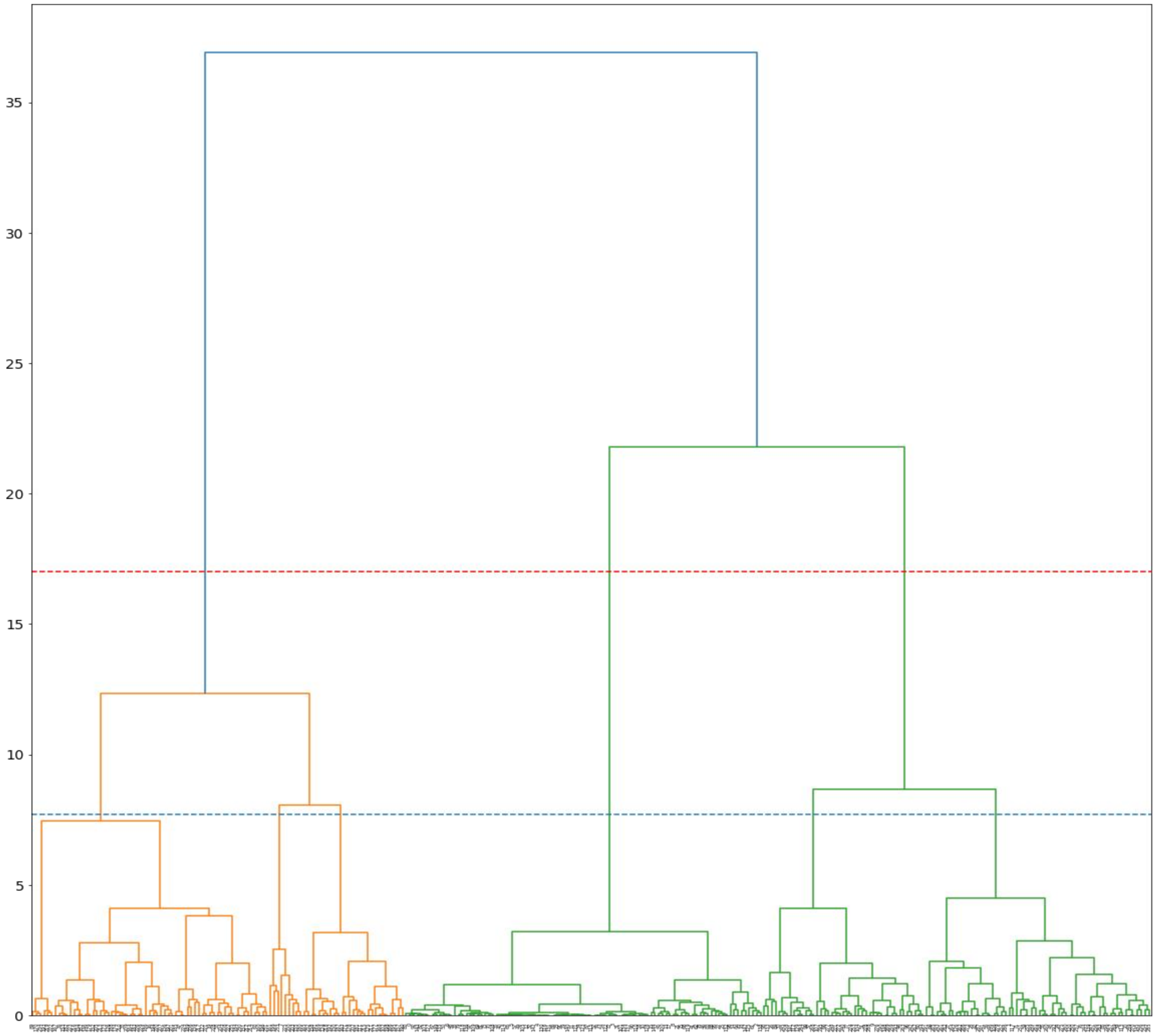

Fig.2.3. Dendogram

The optimum amount of clusters is 6. The height of the branching points indicates how similar or different they are from each other. The greater the height, the greater the difference. Keeping the above reference as base, we can see the longest branch (tallest branch) is in blue. If we see that only blue, it will result in only 2 clusters which is not acceptable in business. However, segmentation at green branches are we get obtain only 3 clusters (market with red line). But for this project we are considering 6 clusters (market with blue line) with an optimum height (4 or 5 clusters can also be considered).

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

Applying K-Means clustering,

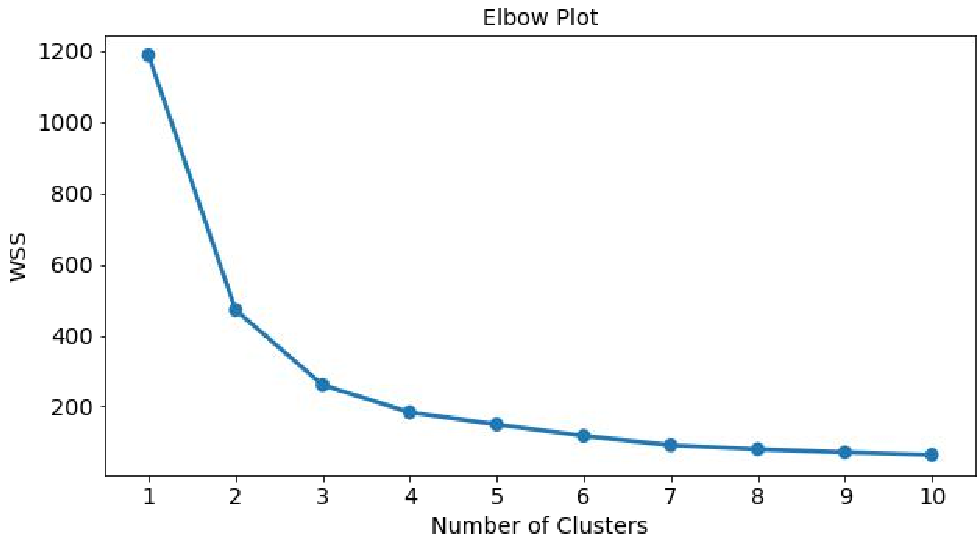

Fig.2.4. Elbow Cruve

The optimum number of clusters is 6.

Silhouette Score,

```
The Average Silhouette Score for 2 clusters is 0.53092
The Average Silhouette Score for 3 clusters is 0.53354
The Average Silhouette Score for 4 clusters is 0.55205
The Average Silhouette Score for 5 clusters is 0.52026
The Average Silhouette Score for 6 clusters is 0.52997
The Average Silhouette Score for 7 clusters is 0.55595
The Average Silhouette Score for 8 clusters is 0.53301
The Average Silhouette Score for 9 clusters is 0.5138
The Average Silhouette Score for 10 clusters is 0.51142
```

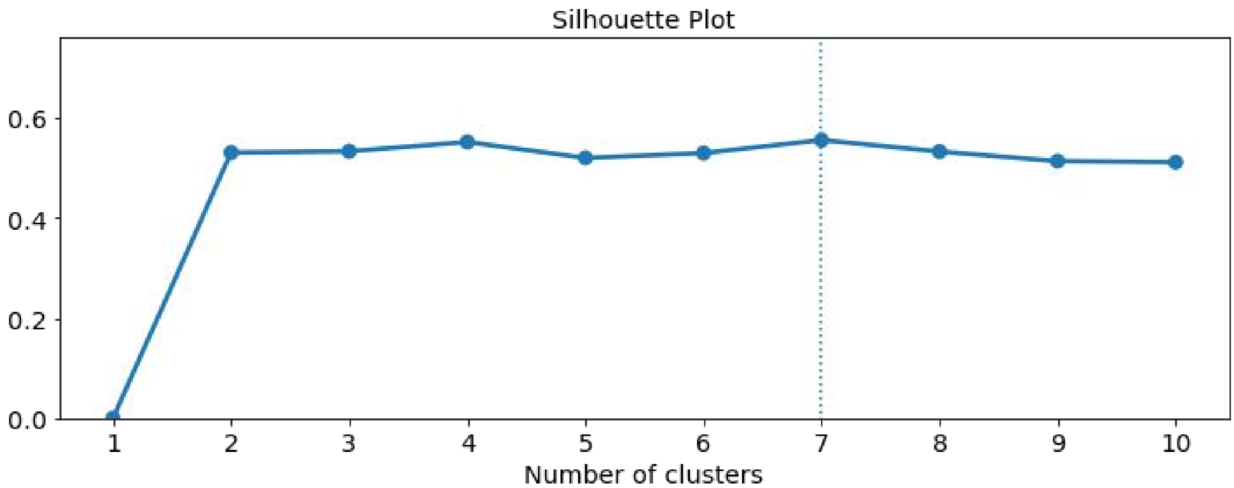

Fig.2.5. Silhouette Score Plot

Q.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.

Cluster profile,

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

Table - 2.5. Data head with K-Means

	Health_indeces1	Health_indices2	Per_capita_income	GDP	KMEANS_LABELS
count	297.000000	297.000000	297.000000	297.000000	297.000000
mean	2630.151515	693.632997	2156.915825	174601.117845	2.010101
std	2038.505431	468.944354	1491.854058	167167.992863	1.685560
min	-10.000000	0.000000	500.000000	22.000000	0.000000
25%	641.000000	175.000000	751.000000	8721.000000	1.000000
50%	2451.000000	810.000000	1865.000000	137173.000000	1.000000
75%	4094.000000	1073.000000	3137.000000	313092.000000	3.000000
max	10219.000000	1508.000000	7049.000000	728575.000000	5.000000

Table - 2.6. Data Description with K-Means

KMEANS_LABELS	0	1	2	3	4	5
Health_indeces1	4816.07	444.40	4116.97	2362.12	8327.67	2815.98
Health_indices2	1140.82	108.02	1293.00	848.38	1369.67	675.96
Per_capita_income	2319.30	686.81	4728.33	3160.00	5592.44	1530.96
GDP	399053.82	7241.66	342126.67	143591.27	426759.11	133085.91
freq	57.00	98.00	30.00	56.00	9.00	47.00

Table - 2.7. Cluster Profile

The minimum **Silhouette Coefficient for each sample is** -0.022892537665382358

The **Silhouette Coefficient** is a measure of how well samples are clustered with samples that are similar to themselves. Clustering models with a high Silhouette Coefficient are said to be dense, where samples in the same cluster are similar to each other, and well separated, where samples in different clusters are not very similar to each other.

Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

- 1: Best value, it means clusters are well apart from each other and clearly distinguished.
- 0: Overlapping clusters, it means clusters are indifferent, or we can say that the distance between clusters is not significant.
- -1: This is not a good value. It means clusters are assigned in the wrong way.

Cluster Profile Observations (From Table 2.7) :

- Cluster **0** and **4** has higher mean **Health_indeces1** compared to other clusters. Cluster **1** h as the lowest mean **Health_indeces1**.

- Cluster **2** and **4** has higher mean **Health_indeces2** compared to other clusters. Cluster **1** h as the lowest mean Health_indeces2.

- Cluster **2** and **4** has higher mean **Per_capita_income** compared to other clusters. Cluster **1** has the lowest mean **Per_capita_income**.

- Cluster **0** and **4** has higher mean **GDP** compared to other clusters. Cluster **1** has the lowest mean **GDP**.

- Cluster **1** has highest mean **freq** compared to other clusters. Cluster **1** has the lowest mea n **Health_indeces1, Health_indeces2, Per_capita_income and GDP**.