DM Project Report

Contents

List of Figures

List of Tables

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 drop this variable before doing Principal Component Analysis.

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

Data Info:

```
<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
```

Data set contains continuous data (numerical) only.

Duplicate:

There is no duplicate in the given data set.

Null Check:

```
df.isnull().sum()
```

```
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
dtype: int64
```

There is no null data available in the given data set.

Summary of the data sheet:

	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

Removing unnecessary Column:

Index value is not used anymore to calculate PCA.
Since Response is the output variable, we can remove this column.
Ecommerce is also not required to calculate PCA.

Univariate Analysis:

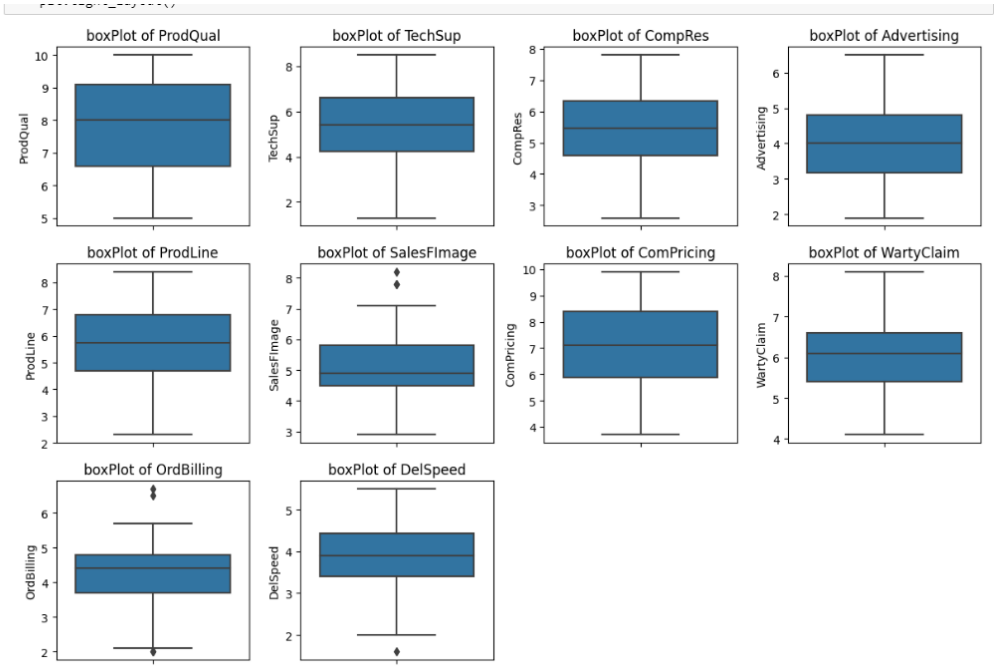

Figure 1: Boxplots of the numerical variables

There are some outliers present in SalesFImage, OrdBilling and DelSpeed.

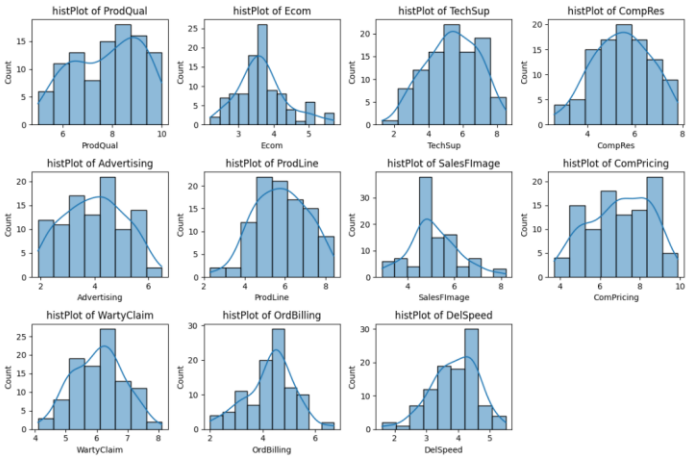

Figure 2: Histogram of the numerical variables

Most of the variants having normal distribution apart from Compricing, Prodqual.

Multivariate Analysis

Pair Plot

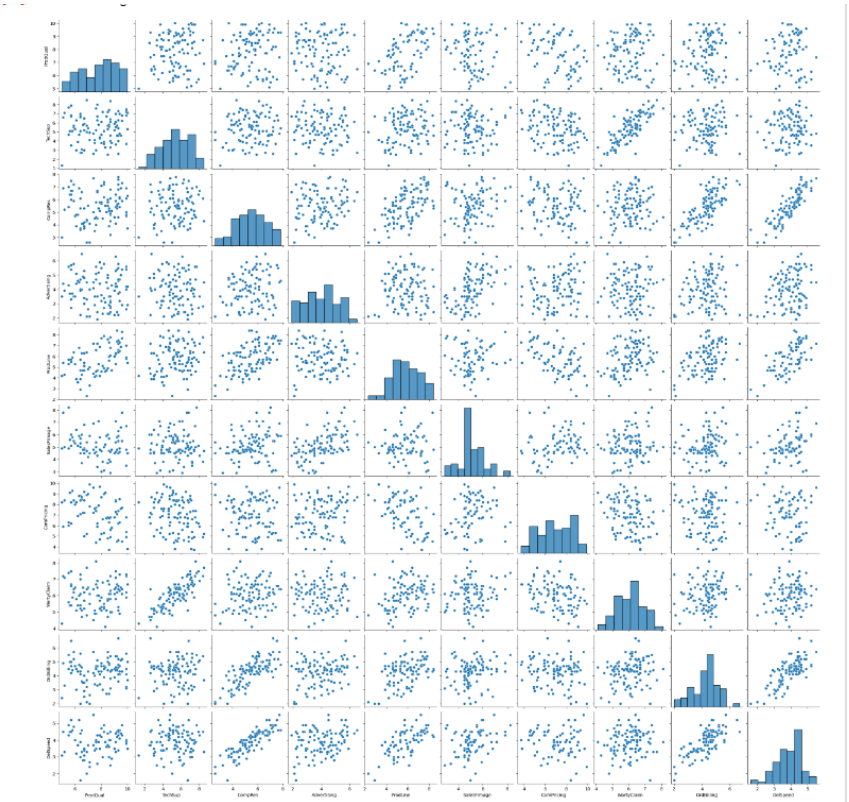

Figure 3: Pair plot of the numerical variables

Co relation analysis:

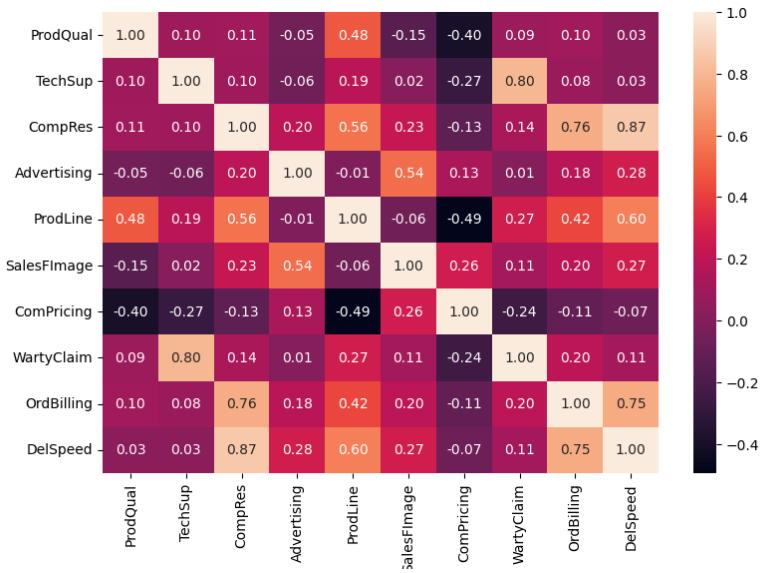

Figure 4: Heat Map of the numerical variables

- Highest positive corelation exist between Delivery Speed and Company Response (87%)
- Highest positive corelation exist between warranty claim and technical support (80%)
- Positive corelation exist between Order Billing and Company Response (76%).
- Highest positive corelation exist between Delivery Speed and order billing (75%)

Part 1: PCA: Scale the variables and write the inference for using the type of scaling function for this case study

Applying Scaling function

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

H0: Correlations are not significant.

H1: There are significant correlations.

p_value is lesser than 0.5. So, we can reject the null hypothesis here. To conclude this, there is a correlation between variables.

Part 1: PCA: Comment on the comparison between covariance and the correlation matrix after scaling.

Covariance (After Scaling)

	ProdQual	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
ProdQual	1.010101	0.096566	0.107444	-0.054013	0.482317	-0.153346	-0.405335	0.089204	0.105357	0.027998
TechSup	0.096566	1.010101	0.097633	-0.063505	0.194571	0.017162	-0.273522	0.805220	0.080911	0.025698
CompRes	0.107444	0.097633	1.010101	0.198906	0.567088	0.232072	-0.129247	0.141827	0.764514	0.873830
Advertising	-0.054013	-0.063505	0.198906	1.010101	-0.011667	0.547680	0.135573	0.010901	0.186097	0.278650
ProdLine	0.482317	0.194571	0.567088	-0.011667	1.010101	-0.061935	-0.499948	0.275836	0.428695	0.607930
SalesFImage	-0.153346	0.017162	0.232072	0.547680	-0.061935	1.010101	0.267269	0.108541	0.197098	0.274294
ComPricing	-0.405335	-0.273522	-0.129247	0.135573	-0.499948	0.267269	1.010101	-0.247461	-0.115724	-0.073608
WartyClaim	0.089204	0.805220	0.141827	0.010901	0.275836	0.108541	-0.247461	1.010101	0.199056	0.110500
OrdBilling	0.105357	0.080911	0.764514	0.186097	0.428695	0.197098	-0.115724	0.199056	1.010101	0.758589
DelSpeed	0.027998	0.025698	0.873830	0.278650	0.607930	0.274294	-0.073608	0.110500	0.758589	1.010101

Correlation (After Scaling)

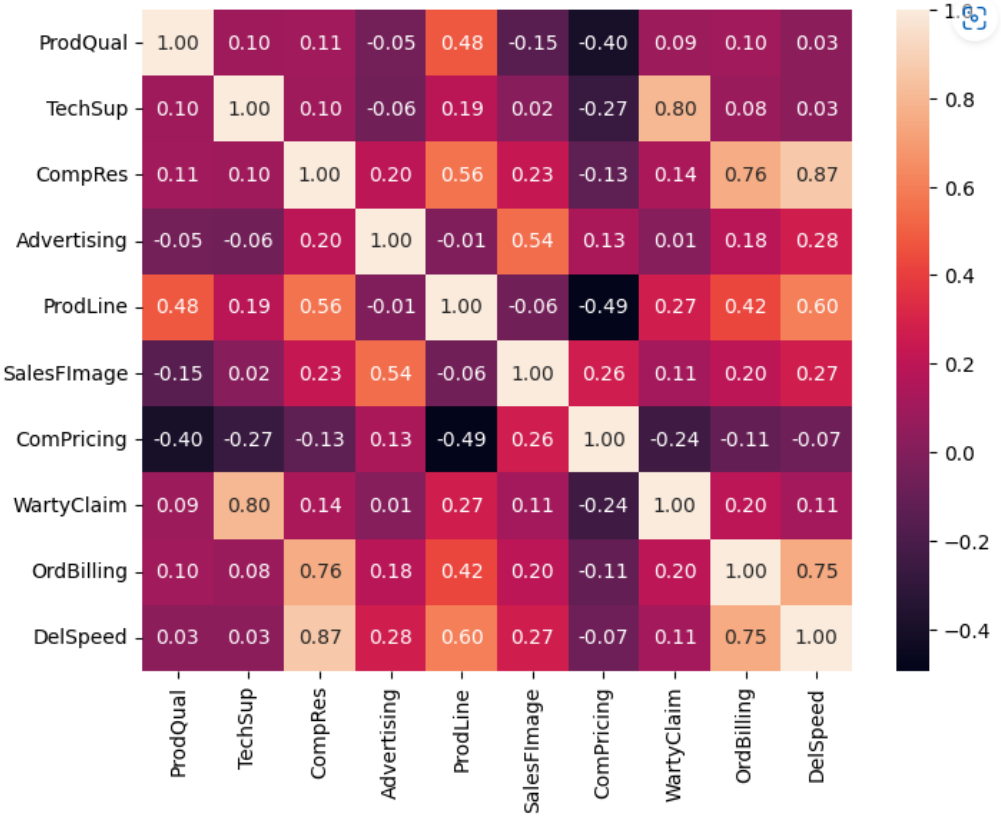

Figure 5: Heat Map of the numerical variables after scaling

- Highest positive corelation exist between Delivery Speed and Company Response (87%)
- Highest positive corelation exist between warranty claim and technical support (80%)
- Positive corelation exist between Order Billing and Company Response (76%).
- Highest positive corelation exist between Delivery Speed and order billing (75%)

Part 1: PCA: Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.

Before Scaling:

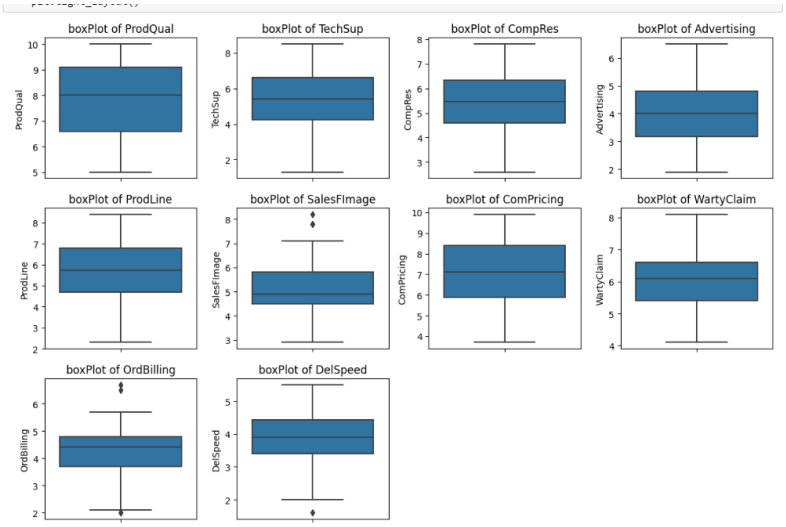

Figure 6: Box Plot of the numerical variables Before scaling

	ProdQual	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
ProdQual	1.949596	0.204293	0.179475	-0.084141	0.876919	-0.227303	-0.865697	0.101081	0.135273	0.028424
TechSup	0.204293	2.342298	0.178758	-0.108434	0.387753	0.027884	-0.640313	1.000106	0.113869	0.028596
CompRes	0.179475	0.178758	1.460238	0.268162	0.892313	0.297711	-0.238897	0.139085	0.849519	0.767766
Advertising	-0.084141	-0.108434	0.268162	1.270000	-0.017121	0.655222	0.233697	0.009970	0.192848	0.228323
ProdLine	0.876919	0.387753	0.892313	-0.017121	1.729975	-0.086480	-1.005828	0.294429	0.518495	0.581384
SalesFImage	-0.227303	0.027884	0.297711	0.655222	-0.086480	1.149870	0.438382	0.094456	0.194349	0.213861
ComPricing	-0.865697	-0.640313	-0.238897	0.233697	-1.005828	0.438382	2.387196	-0.310285	-0.164416	-0.082691
WartyClaim	0.101081	1.000106	0.139085	0.009970	0.294429	0.094456	-0.310285	0.671971	0.150046	0.065861
OrdBilling	0.135273	0.113869	0.849519	0.192848	0.518495	0.194349	-0.164416	0.150046	0.862743	0.512315
DelSpeed	0.028424	0.028596	0.767766	0.228323	0.581384	0.213861	-0.082691	0.065861	0.512315	0.539398

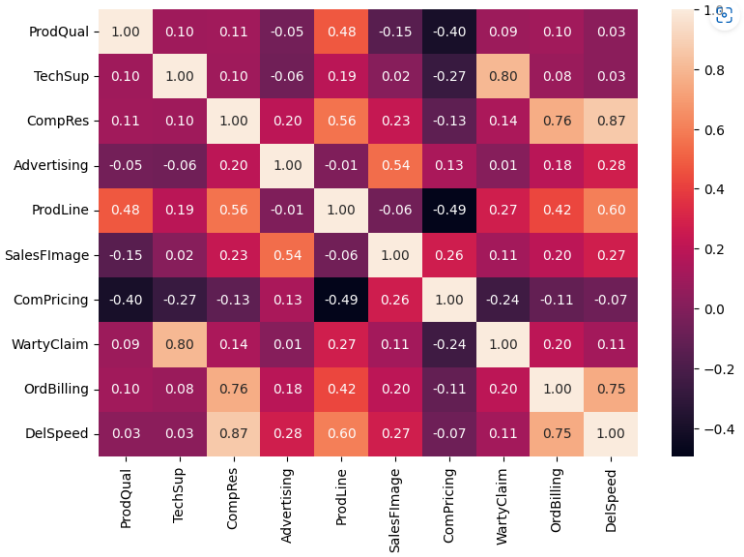

Figure 7: Heat Map of the numerical variables before scaling

- Highest positive corelation exist between Delivery Speed and Company Response (87%)
- Highest positive corelation exist between warranty claim and technical support (80%)
- Positive corelation exist between Order Billing and Company Response (76%).
- Highest positive corelation exist between Delivery Speed and order billing (75%)

After Scaling:

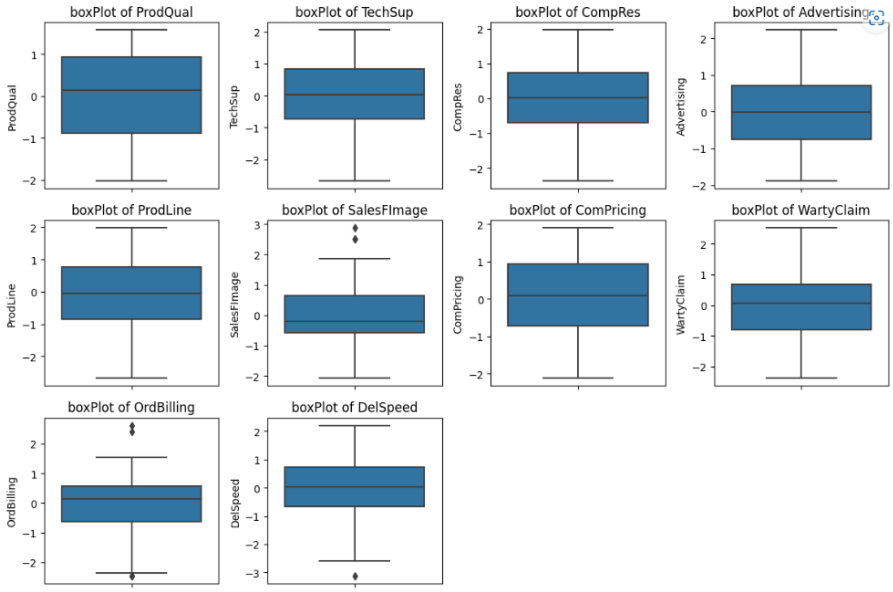

Figure 8: Box Plot of the numerical variables after scaling

	ProdQual	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
ProdQual	1.010101	0.096566	0.107444	-0.054013	0.482317	-0.153346	-0.405335	0.089204	0.105357	0.027998
TechSup	0.096566	1.010101	0.097633	-0.063505	0.194571	0.017162	-0.273522	0.805220	0.080911	0.025698
CompRes	0.107444	0.097633	1.010101	0.198906	0.567088	0.232072	-0.129247	0.141827	0.764514	0.873830
Advertising	-0.054013	-0.063505	0.198906	1.010101	-0.011667	0.547680	0.135573	0.010901	0.186097	0.278650
ProdLine	0.482317	0.194571	0.567088	-0.011667	1.010101	-0.061935	-0.499948	0.275836	0.428695	0.607930
SalesFImage	-0.153346	0.017162	0.232072	0.547680	-0.061935	1.010101	0.267269	0.108541	0.197098	0.274294
ComPricing	-0.405335	-0.273522	-0.129247	0.135573	-0.499948	0.267269	1.010101	-0.247461	-0.115724	-0.073608
WartyClaim	0.089204	0.805220	0.141827	0.010901	0.275836	0.108541	-0.247461	1.010101	0.199056	0.110500
OrdBilling	0.105357	0.080911	0.764514	0.186097	0.428695	0.197098	-0.115724	0.199056	1.010101	0.758589
DelSpeed	0.027998	0.025698	0.873830	0.278650	0.607930	0.274294	-0.073608	0.110500	0.758589	1.010101

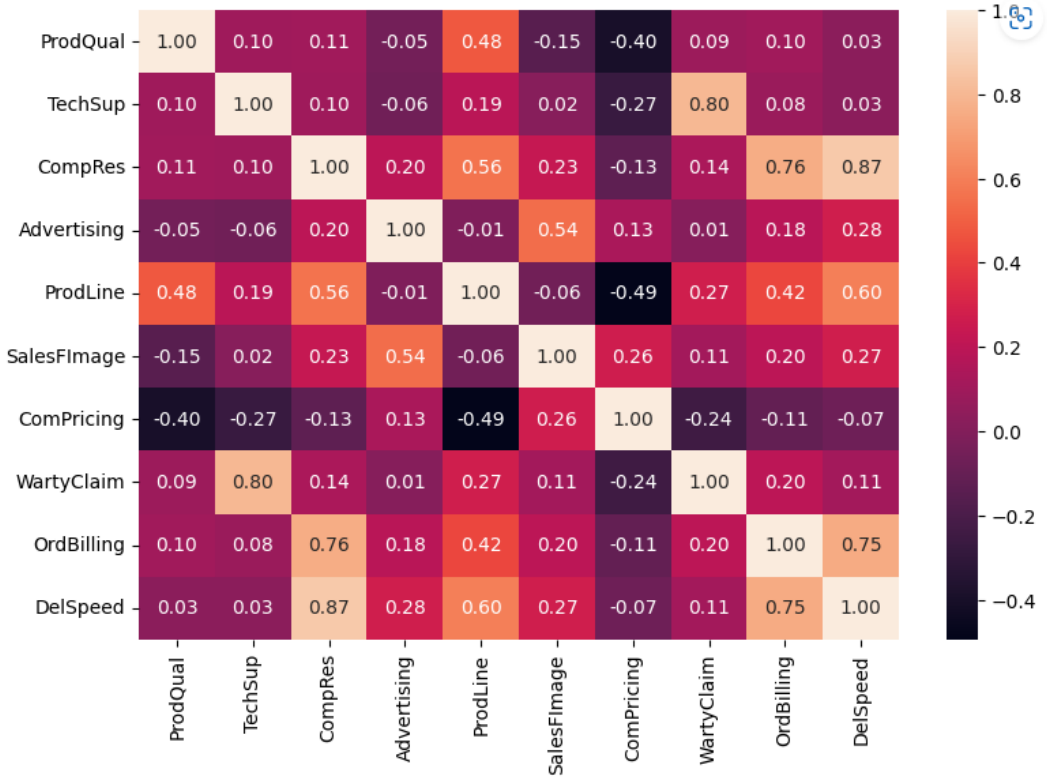

Figure 9: Heat Map of the numerical variables after scaling

- Highest positive corelation exist between Delivery Speed and Company Response (87%)
- Highest positive corelation exist between warranty claim and technical support (80%)
- Positive corelation exist between Order Billing and Company Response (76%).
- Highest positive corelation exist between Delivery Speed and order billing (75%)

Part 1: PCA: Build the covariance matrix, eigenvalues and eigenvector.

Covariance Matrix

	ProdQual	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
ProdQual	1.010101	0.096566	0.107444	-0.054013	0.482317	-0.153346	-0.405335	0.089204	0.105357	0.027998
TechSup	0.096566	1.010101	0.097633	-0.063505	0.194571	0.017162	-0.273522	0.805220	0.080911	0.025698
CompRes	0.107444	0.097633	1.010101	0.198906	0.567088	0.232072	-0.129247	0.141827	0.764514	0.873830
Advertising	-0.054013	-0.063505	0.198906	1.010101	-0.011667	0.547680	0.135573	0.010901	0.186097	0.278650
ProdLine	0.482317	0.194571	0.567088	-0.011667	1.010101	-0.061935	-0.499948	0.275836	0.428695	0.607930
SalesFImage	-0.153346	0.017162	0.232072	0.547680	-0.061935	1.010101	0.267269	0.108541	0.197098	0.274294
ComPricing	-0.405335	-0.273522	-0.129247	0.135573	-0.499948	0.267269	1.010101	-0.247461	-0.115724	-0.073608
WartyClaim	0.089204	0.805220	0.141827	0.010901	0.275836	0.108541	-0.247461	1.010101	0.199056	0.110500
OrdBilling	0.105357	0.080911	0.764514	0.186097	0.428695	0.197098	-0.115724	0.199056	1.010101	0.758589
DelSpeed	0.027998	0.025698	0.873830	0.278650	0.607930	0.274294	-0.073608	0.110500	0.758589	1.010101

Eigen Vector:

```
array([[-0.13862521, -0.13255163, -0.16169816, -0.47353676, -0.1761158 ,
        -0.39268125, -0.18960766,  0.15780895, -0.21620195, -0.44059722,
        -0.47527939],
       [-0.30616255,  0.46140667, -0.22578297,  0.03265571,  0.3640688 ,
        -0.27255203,  0.47203851,  0.40962005, -0.18553872,  0.04338501,
         0.08704893],
       [ 0.06709281, -0.22881574, -0.61626115,  0.20638388, -0.0906638 ,
         0.11796883, -0.23760663,  0.04673072, -0.60432141,  0.15770584,
         0.22342129],
       [ 0.64972486,  0.25658155, -0.18140816, -0.20421824,  0.33265598,
         0.20316157,  0.2349916 , -0.32936184, -0.17150521, -0.22900266,
        -0.19979577],
       [ 0.29035482,  0.40547352, -0.00914631,  0.02569798, -0.78168324,
         0.11128718,  0.19275175,  0.29415447, -0.01839439,  0.04402211,
        -0.03422493],
       [ 0.52977297, -0.30406454,  0.10842621,  0.03136166,  0.25830325,
        -0.10856804, -0.12869836,  0.69840604,  0.13820704,  0.10792548,
        -0.02355011],
       [ 0.19462811,  0.07519199, -0.00340068, -0.00843865, -0.04839602,
        -0.60704084, -0.03532911, -0.2994856 , -0.03912519,  0.66311709,
        -0.23097124],
       [ 0.13317007, -0.18938057,  0.41970237,  0.51401584, -0.07750325,
        -0.34374346,  0.2570409 , -0.10993275, -0.40542731, -0.36861759,
         0.05887779],
       [ 0.01183241, -0.52168045, -0.41237425,  0.01690441, -0.1574079 ,
        -0.0745044 ,  0.62078485, -0.08201505,  0.36496521, -0.02092733,
        -0.03612782],
       [ 0.08936578,  0.28195279, -0.38976411,  0.5056586 ,  0.02731564,
        -0.2654653 , -0.34636059, -0.06113765,  0.44586354, -0.33035426,
        -0.04538755],
       [ 0.17803864,  0.03622843, -0.01479579, -0.4104794 , -0.07704307,
        -0.35982333, -0.05636008, -0.10052761,  0.08293481, -0.16708542,
         0.78408613]])
```

Eigen Value

Eigen value will retrieved from explained_variance. Therefore,

```
array([3.40819146, 2.1897756 , 1.58979714, 1.01213449, 0.57593913,
       0.43260415, 0.39761745, 0.24726987, 0.14760309, 0.10007772])
```

```
array([0.33741095, 0.21678778, 0.15738992, 0.10020131, 0.05701797,
       0.04282781, 0.03936413, 0.02447972, 0.01461271, 0.00990769])
```

Part 1: PCA: Write the explicit form of the first PC (in terms of Eigen Vectors).

```
array([-0.16769139, -0.17596117, -0.4733895 , -0.13800081, -0.42041589,
       -0.13494202,  0.1976868 , -0.22680363, -0.43805204, -0.46947525])
```

```
array([-0.16769139, -0.17596117, -0.4733895 , -0.13800081, -0.42041589,
       -0.13494202,  0.1976868 , -0.22680363, -0.43805204, -0.46947525])
```


(-0.167691 * **ProdQual) + (**-0.17596117***TechSup) + (**-0.4733895 ***CompRes) + (**-0.13800081***Advertising)+(** -0.42041589***ProdLine)+(** -0.13494202* **SalesFImage)+(** 0.1976868 ***ComPricing)+(** -0.22680363***WartyClaim)+(** -0.43805204***OrdBilling)+(** -0.46947525**** OrdBilling)**

Part 1: 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.

Eigen Values (ratio):

```
array([0.33741095, 0.21678778, 0.15738992, 0.10020131, 0.05701797,
0.04282781, 0.03936413, 0.02447972, 0.01461271, 0.00990769])
```

Cumulative:

```
array([0.33741095, 0.55419874, 0.71158866, 0.81178997, 0.86880794,
0.91163576, 0.95099988, 0.9754796 , 0.99009231, 1.        ])
```

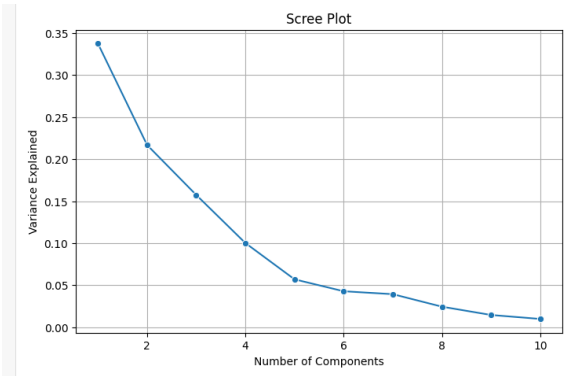

Figure 10: Scree Plot for PCA

- We can see from the plot that there is a consistent dip from 1 to 8 and there doesn't seem to be a clear 'elbow' here. We may chose any from 1 to 8 as our # of clusters.

	PC1	PC2	PC3	PC4	PC5
ProdQual	-0.167691	-0.330597	-0.257431	0.572797	-0.600417
TechSup	-0.175961	-0.357497	0.562634	-0.117578	-0.066298
CompRes	-0.473390	0.169622	-0.118662	-0.200457	-0.034925
Advertising	-0.138001	0.384934	0.231289	0.572578	0.317047
ProdLine	-0.420416	-0.220108	-0.218347	0.119984	-0.002778
SalesFImage	-0.134942	0.413982	0.365982	0.351534	-0.123538
ComPricing	0.197687	0.442764	0.138871	-0.242581	-0.706788
WartyClaim	-0.226804	-0.302891	0.572445	-0.089517	-0.092527
OrdBilling	-0.438052	0.159256	-0.078284	-0.230813	-0.087575
DelSpeed	-0.469475	0.230175	-0.121954	-0.178732	0.048127

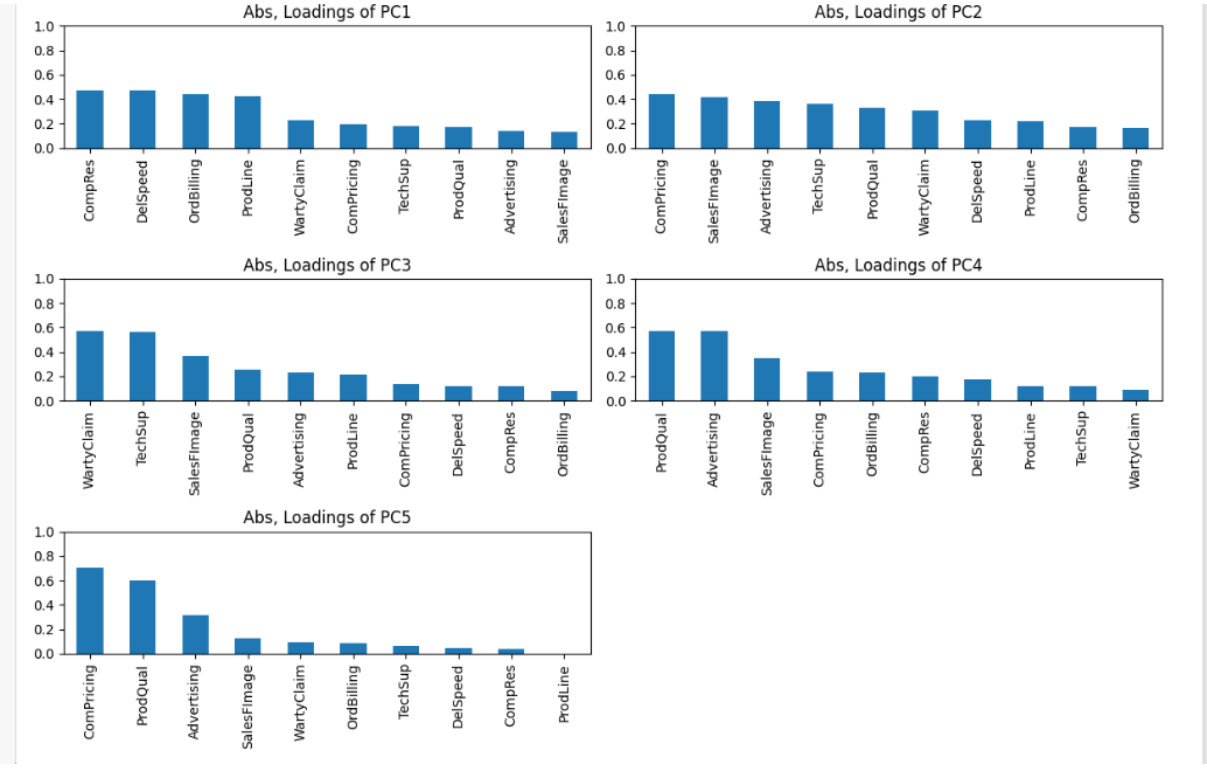

Figure 11: Detailed plot of PC1, PC2, PC3, PC4 and PC5

Part 1: PCA: Mention the business implication of using the Principal Component Analysis for this case study.

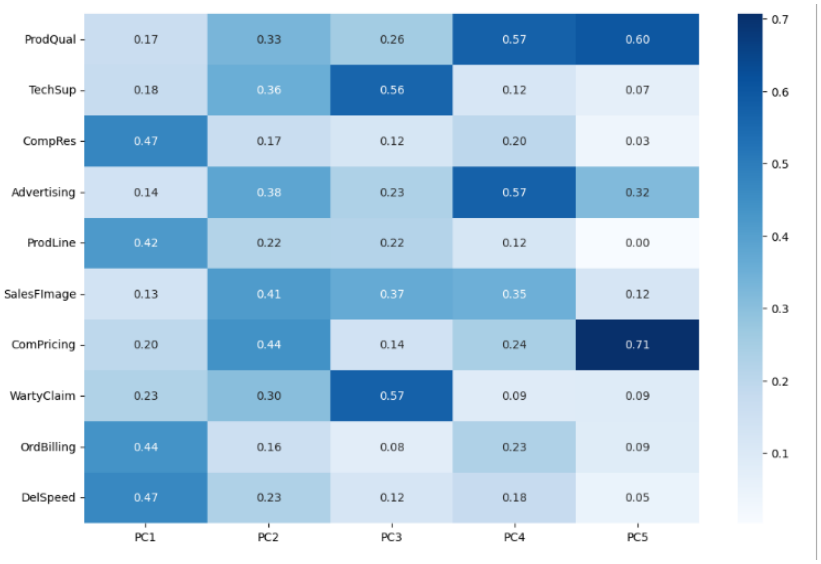

Figure 12: Heat map of Principle Components

- PC1 having highest values of Delivery Speed , Order Billing, Product Line and Company response.
- Where as PC2 having highest values Company Pricing, Sales F Image, and Advertising.
- PC3 has it own characteristics like higher values of warranty claim, Technical Support and Sales F Image.
- PC4 having highest values of Advertising and Product Quality.
- And finally PC5, having highest values of company pricing and Product Quality.
- So we use 5 columns instead of 10 columns from the calculated data.
  For e.g. below mentioned values are first index of scaled data (df_pca_scaled.iloc[0])

```
ProdQual       0.496660
TechSup       -1.881421
CompRes        0.380922
Advertising    0.704543
ProdLine      -0.691530
SalesFImage    0.821973
ComPricing    -0.113185
WartyClaim    -1.646582
OrdBilling     0.781230
DelSpeed      -0.254531
Name: 0, dtype: float64
```

After performing dot product, (replacement value of above 10 columns)

Score of 1st row :

```
0.2783831, .851167, -1.6052461, .0787530, .088661
```

Part 2: Clustering:

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.

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.

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

Null Check

```
Unnamed: 0          0
States              0
Health_indeces1     0
Health_indices2     0
Per_capita_income   0
GDP                 0
dtype: int64
```

There are no null values in the given data sets.

Duplicate Check:

There are no duplicate values in the given data.

Shape of the Data Set: (297, 6)

This shows the total number of rows = 297 and total number of columns = 6.

Data Type:

sno	Data Type	Count
1	Int64	5
2	Object	1

Table 1: Data Type and Count

Dropping Unnecessary columns:

There are two variables “Unamed: 0” and States that signifies only the id in the dataset and are not required in clustering process.

Summary of the Data set:

	count	mean	std	min	25%	50%	75%	max
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

Univariant Analysis

Health Indeces1

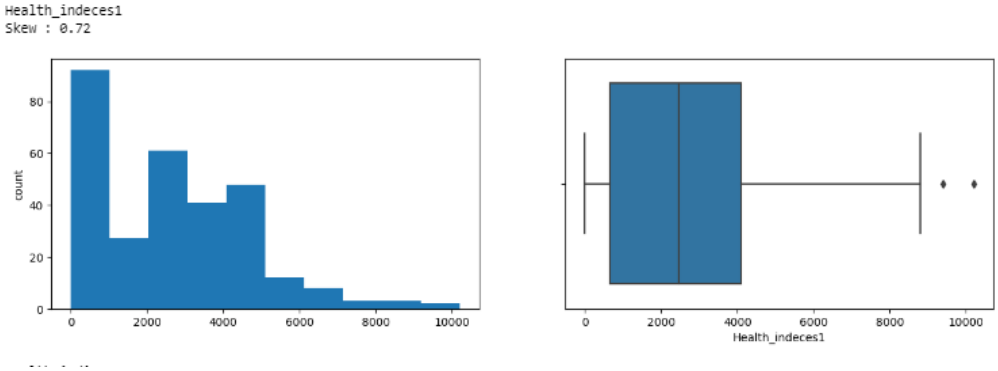

Figure 13: Boxplot and histogram of Health_Indeces1

Health Indeces2

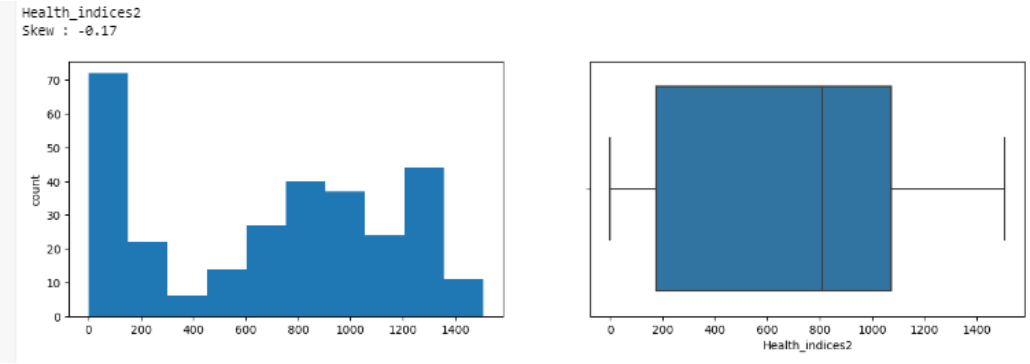

Figure 14: Boxplot and histogram of Health_Indeces2

Per Capita Income

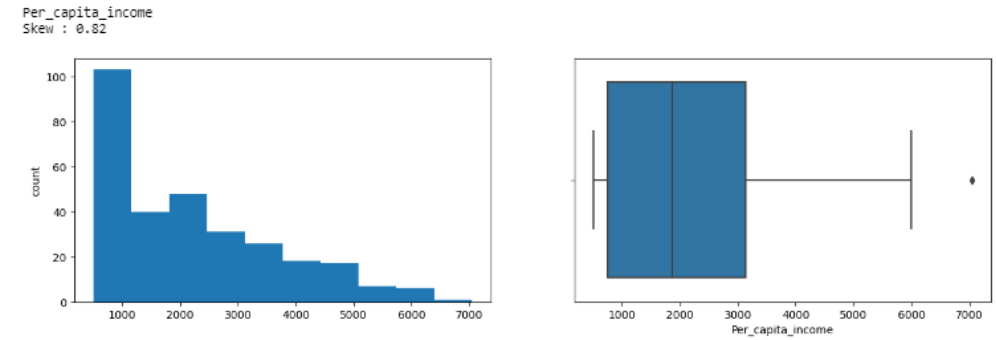

Figure 15: Boxplot and histogram of Per_Capita_Income

GDP

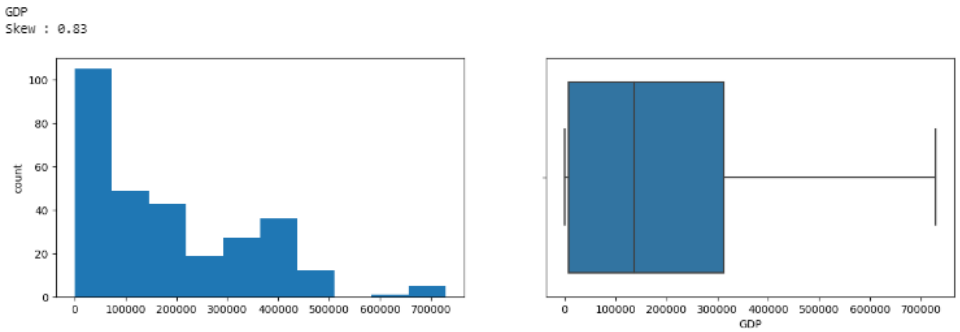

Figure 16: Boxplot and histogram of GDP

- All the variables except for Health_indices2 are right skewed.
- Health_indices2 is negatively skewed.
- There are outliers present “Health_indices1” and “Per_capita_income”

Multivariate Analysis:

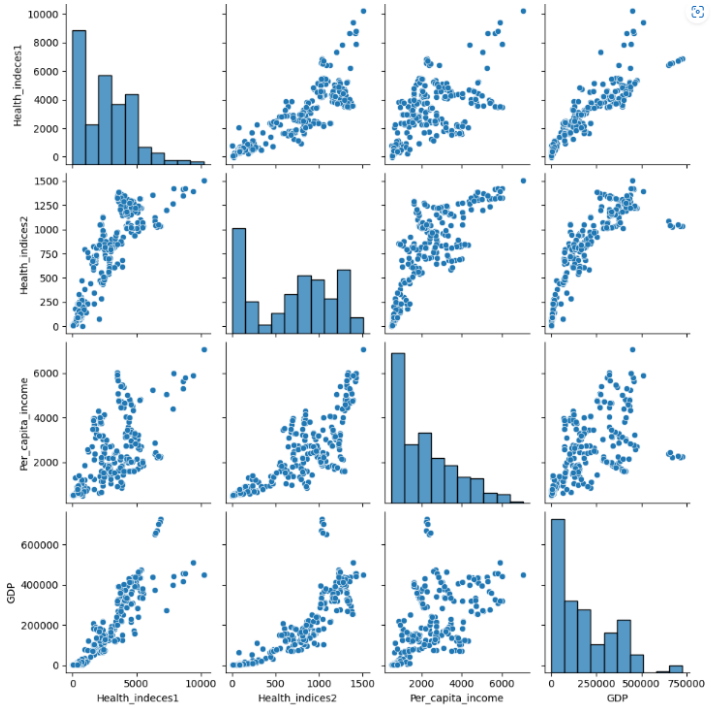

Figure 17: Pair Plot of numerical values

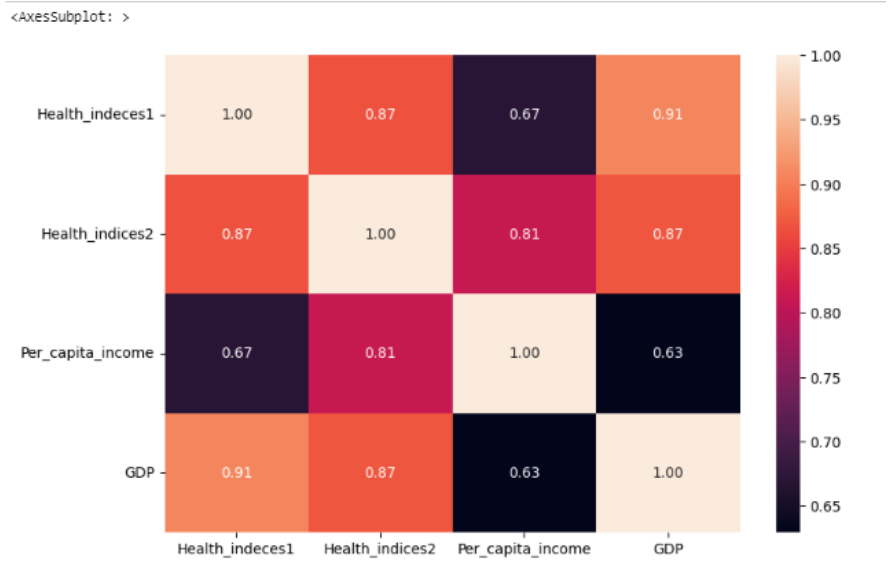

Figure 18: Heat Map of numerical values

- Highest correlation between “Health_indeces1” and “GDP”

Outlier Treatment:

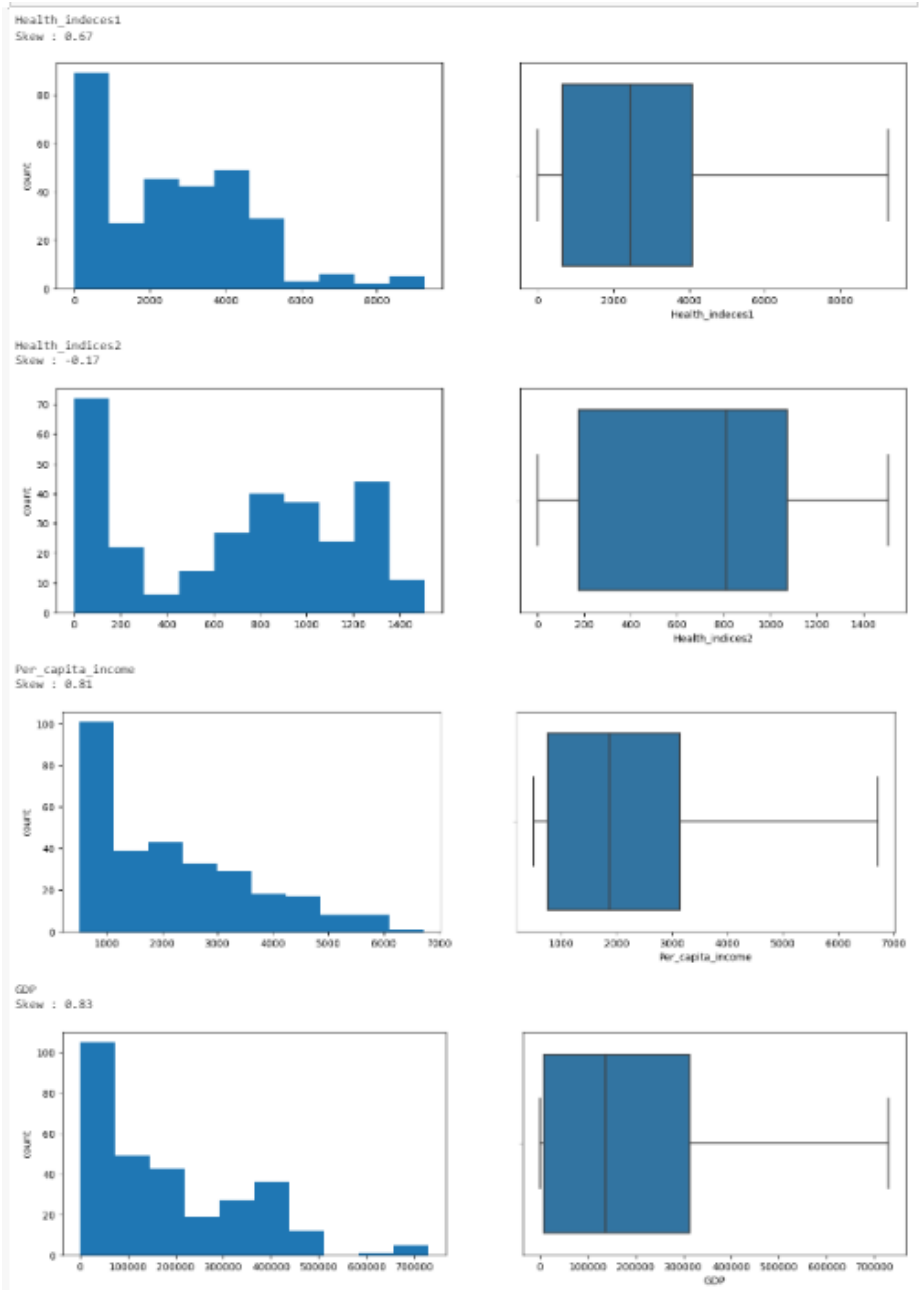

Figure 19: Box plot of numerical values

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

Yes, Scaling is necessary as clustering algorithm such as k-means do need feature scaling before they feed to the algorithm. Since clustering techniques use Euclidean distance, it will be wise to scale the data consisting of attributes with different units of measurements.

Before Scaling:

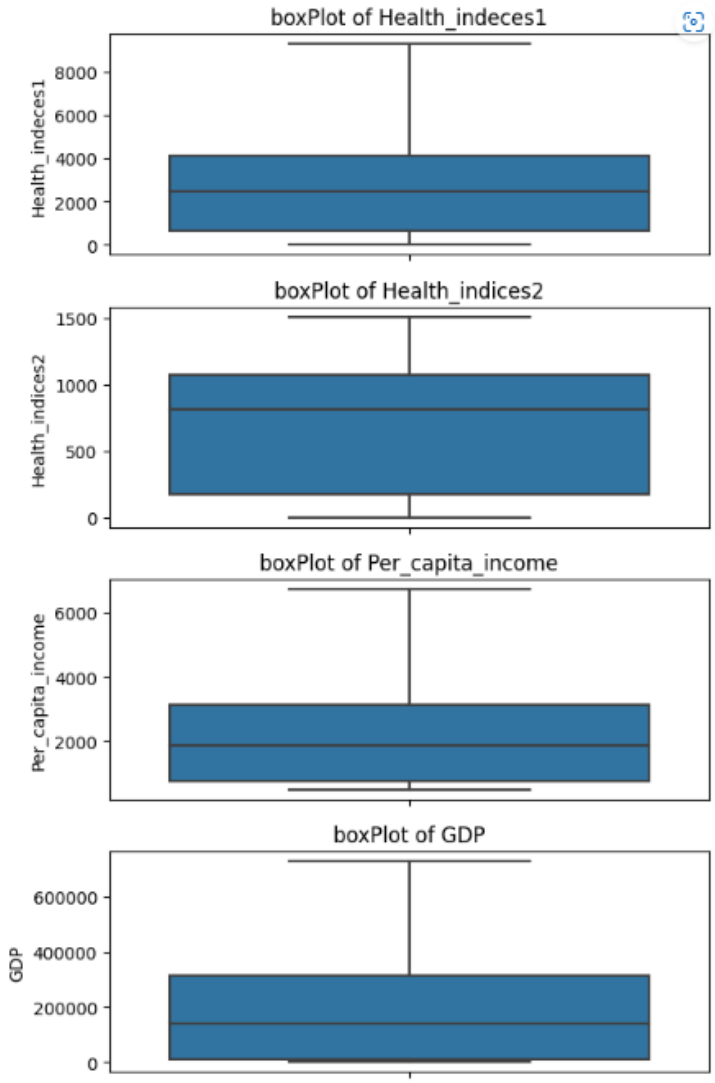

Figure 20: Box plot of numerical values (before scaling)

After Scaling:

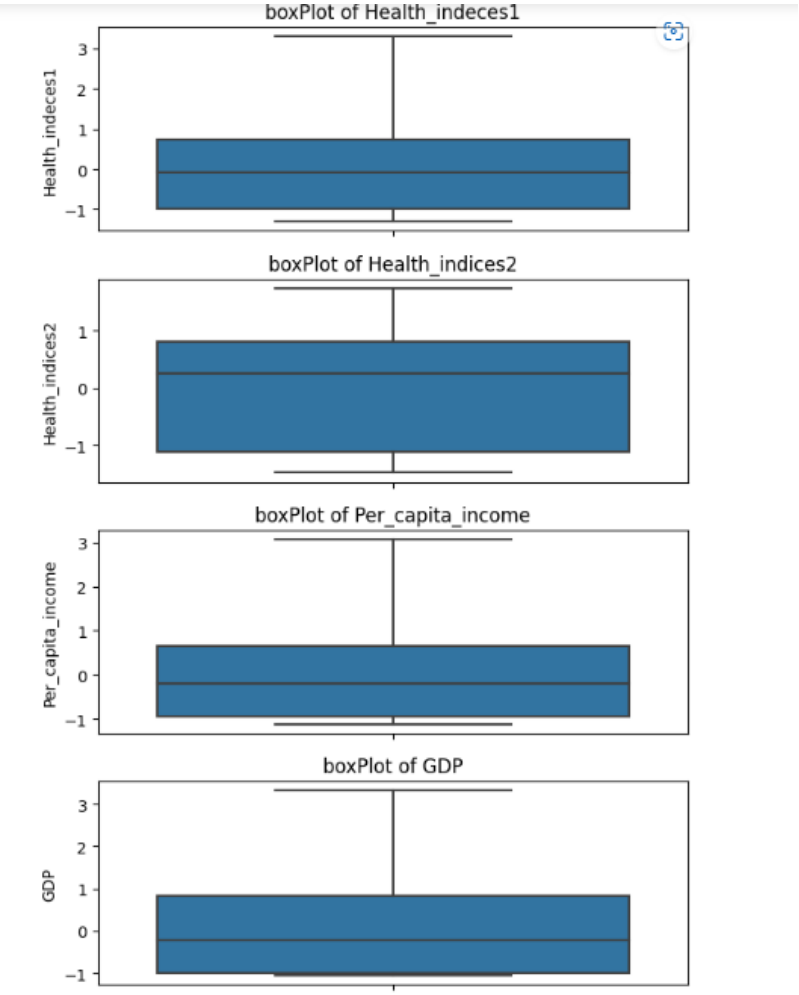

Figure 21: Box plot of numerical values (after scaling)

All numerical data have been scaled between -3 to 3.

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

Ward Coverage:

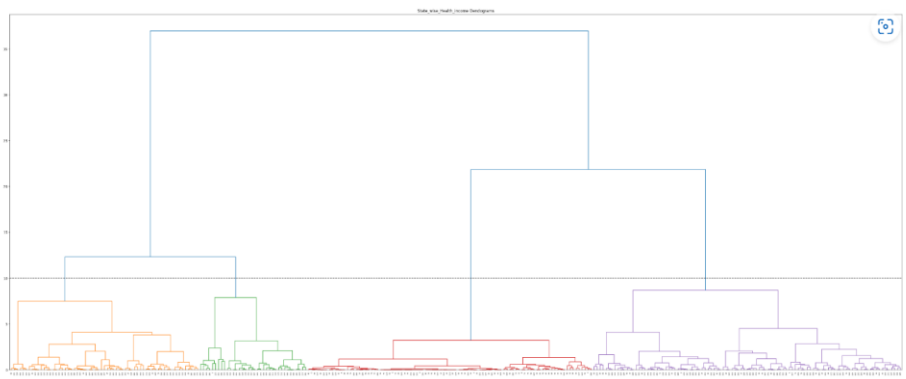

Figure 22: Dendrogram using ward coverage method

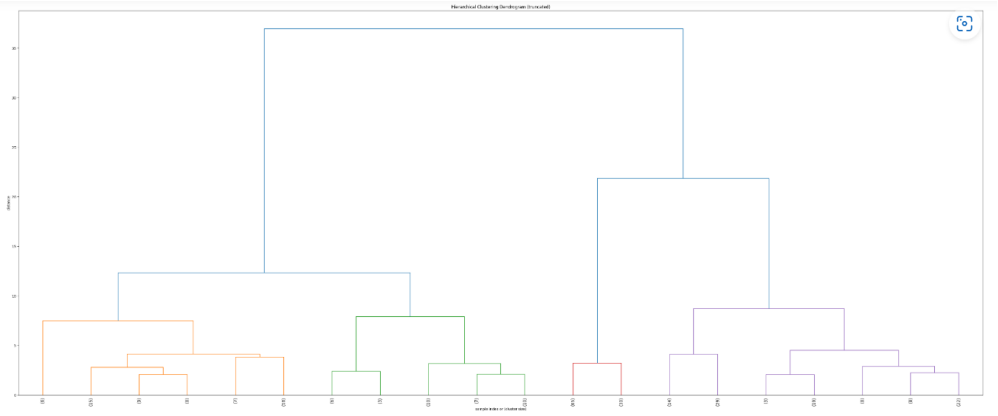

Figure 23: Dendrogram using ward coverage method (p=20)

From the above dendrogram, we have formed 3 clusters.

Cluster Frequency:

cluster_1	Unnamed: 0	Health_indeces1	Health_indices2	Per_capita_income	GDP	cluster count
1	175.767677	4923.545455	1201.646465	3375.141414	377132.474747	99
2	75.021053	401.063158	104.536842	680.673684	5388.768421	95
3	188.621359	2481.776699	748.689320	2347.582524	136004.699029	103

```
<AxesSubplot: xlabel='GDP', ylabel='Health_indeces1'>
```

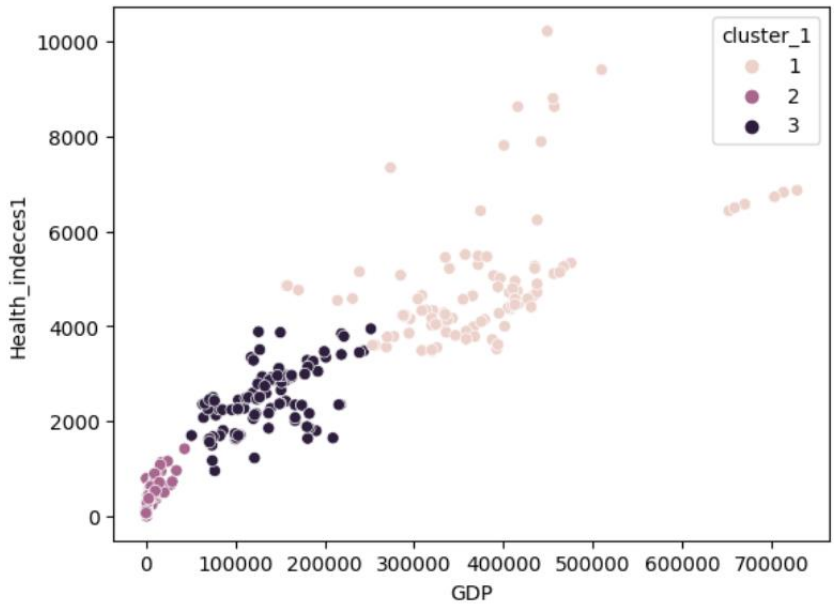

Figure 24: Scatter plot between Health_indeces1 vs GDP

Three group cluster solutions gives a pattern based on high, medium and low GDP.

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

silhouette score

```
2: 0.5282573570427488
3: 0.5340151343712788
4: 0.5524561729411546
5: 0.5204779018816421
6: 0.5307157026086741
7: 0.5550906360809267
8: 0.5342932176693953
9: 0.530194646842701
```

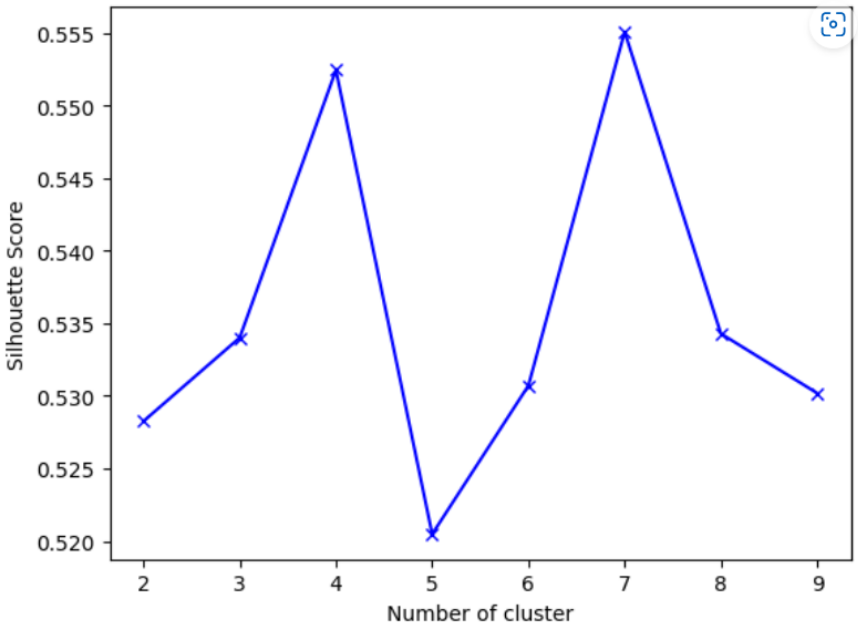

Figure 24: silhouette score

We can use either 3 or 4 clusters from the silhouette score. But there is no significant changes between 3 & 4 from the score.

So, 3 clusters solution seems to be the best fit to differentiate the GDP as,

High/Medium/Low GDP per capita area.

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_1	Unnamed: 0	Health_indeces1	Health_indices2	Per_capita_income	GDP	cluster count
1	175.767677	4923.545455	1201.646465	3375.141414	377132.474747	99
2	75.021053	401.063158	104.536842	680.673684	5388.768421	95
3	188.621359	2481.776699	748.689320	2347.582524	136004.699029	103

Cluster 1= High GDP per capita

Cluster 2=Low GDP per capita

Cluster 3=Medium GDP per capita

Cluster 1: High GDP per capita

- These are having high growth rate.
- The Health and Economic conditions are excellent in these areas.
- Income in this area is very high.

Cluster 2: Low GDP per capita

- These are having low growth rate.
- The Health and Economic conditions aren't good in these areas.
- Income in this area is very low.

Cluster 3: Medium GDP per capita

- These are having average growth rate.
- The Health and Economic conditions are adequate in these areas.
- Income in this area is average.

Recommendations

Cluster 1: High GDP per capita

- Maintaining the growth in productivity and the size of the workforce will keep the Health and Economic conditions high.

Cluster 2: Low GDP per capita

- More employment opportunities should be created to increase the productivity and economic growth.
- More industries/companies should be opened in these areas for the better development.
- Give some tax benefits to start business in these areas.

Cluster 3: Medium GDP per capita

- New business will help in the growth of development.
- Cutting tax rates will also help in the growth of development.