Data Mining (Extended Project)

AMAN TANDON

Contents

Principal component analysis (PCA)

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

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

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

- Build the covariance matrix, eigenvalues and eigenvector.

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

- 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.

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

Contents

Clustering

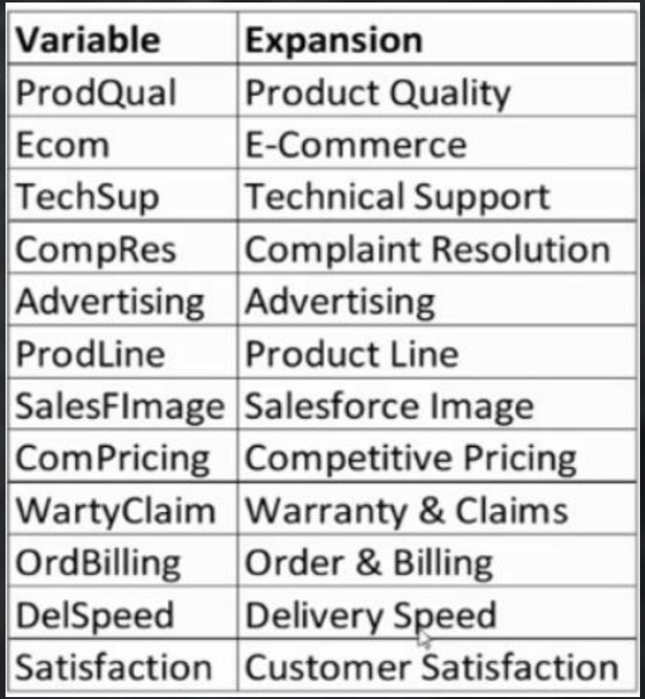
- Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, etc)
- Do you think scaling is necessary for clustering in this case? Justify.
- Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using a Dendrogram and briefly describe them.
- Apply K-Means clustering on scaled data and determine optimum clusters. Apply the elbow curve and find the silhouette score.
- 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.

Principal component analysis (PCA)

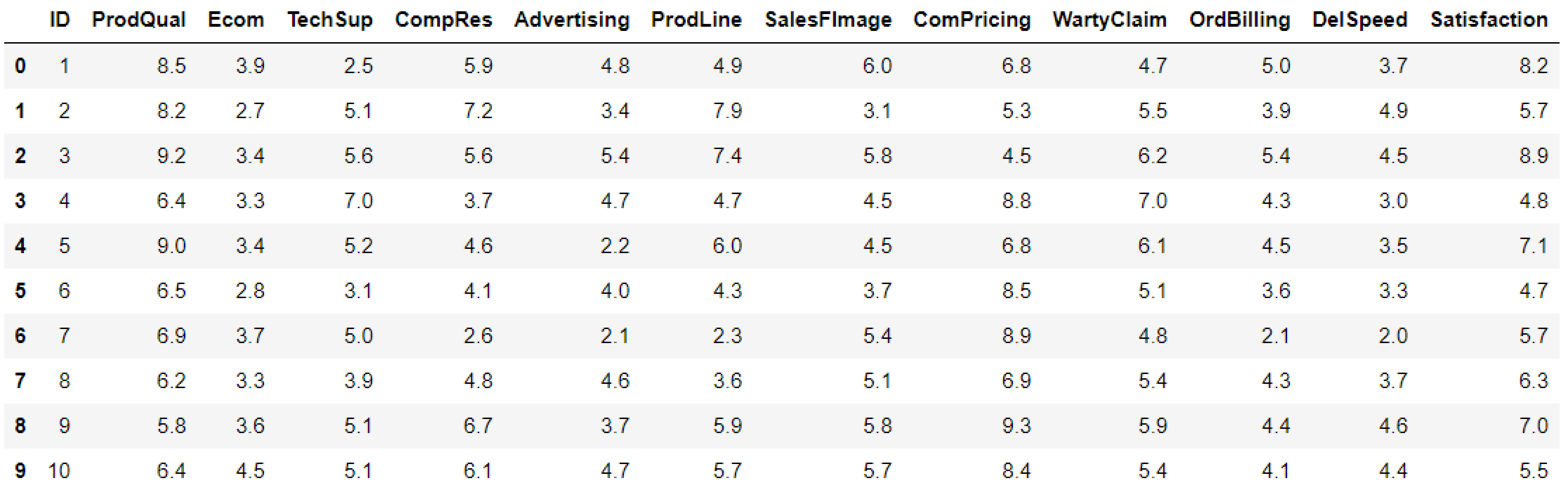
Problem Statement

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 a 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.

Data Dictionary



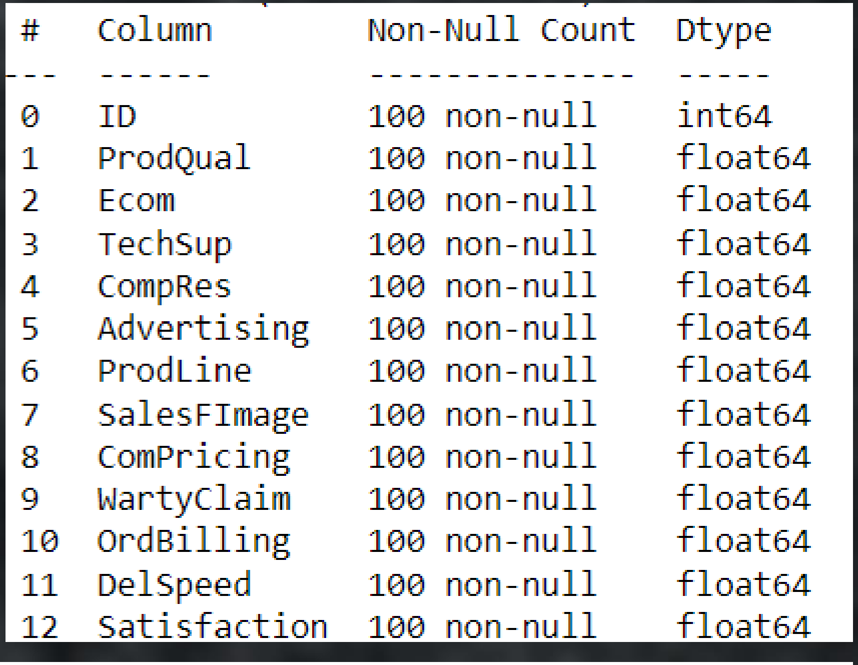
Variable	Expansion
ProdQual	Product Quality
Ecom	E-Commerce
TechSup	Technical Support
CompRes	Complaint Resolution
Advertising	Advertising
ProdLine	Product Line
SalesFImage	Salesforce Image
ComPricing	Competitive Pricing
WartyClaim	Warranty & Claims
OrdBilling	Order & Billing
DelSpeed	Delivery Speed
Satisfaction	Customer Satisfaction



	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
5	6	6.5	2.8	3.1	4.1	4.0	4.3	3.7	8.5	5.1	3.6	3.3	4.7
6	7	6.9	3.7	5.0	2.6	2.1	2.3	5.4	8.9	4.8	2.1	2.0	5.7
7	8	6.2	3.3	3.9	4.8	4.6	3.6	5.1	6.9	5.4	4.3	3.7	6.3
8	9	5.8	3.6	5.1	6.7	3.7	5.9	5.8	9.3	5.9	4.4	4.6	7.0
9	10	6.4	4.5	5.1	6.1	4.7	5.7	5.7	8.4	5.4	4.1	4.4	5.5

Data set with the first 10 rows

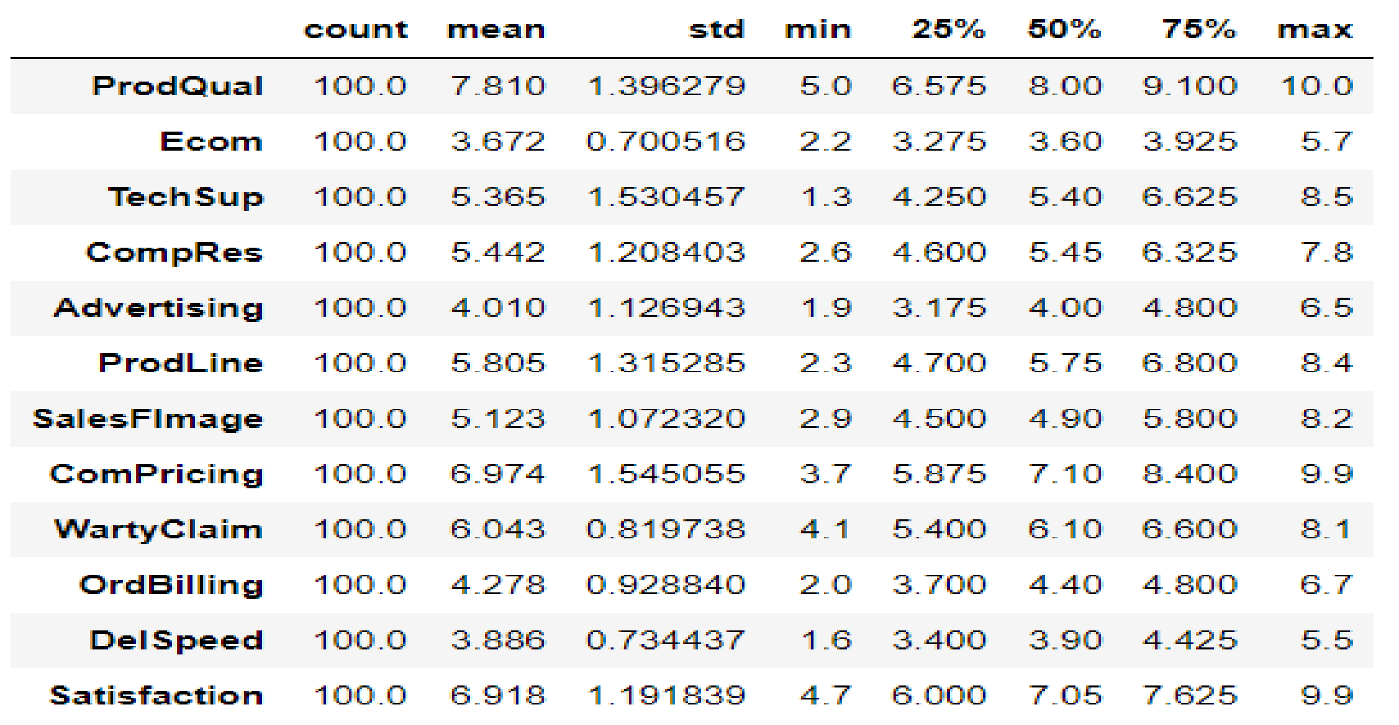
There are 100 rows and 13 columns in the dataset where the 12 columns have float64 data type and 1 columns have int64 data type.



```
#   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
```

There are no Null values and no duplicated values in the data set

Dataset summary



	count	mean	std	min	25%	50%	75%	max
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

- ❑ Product Quality (ProdQual) has a high mean rating of 7.810, indicating that customers generally perceive the product quality to be good. The ratings range from 5.0 to 10.0, with a standard deviation of 1.396279, suggesting some variability in customer opinions.

- ❑ E-commerce (Ecom) has a lower mean rating of 3.672, indicating that customers are relatively less satisfied with the e-commerce experience. The ratings range from 2.2 to 5.7, with a standard deviation of 0.700516, indicating moderate variability in customer opinions.

- ❑ Technical Support (TechSup) has a mean rating of 5.365, suggesting an average satisfaction level. The ratings range from 1.3 to 8.5, with a relatively high standard deviation of 1.530457, indicating significant variability in customer opinions.

- ❑ Complaint Resolution (CompRes): The attribute "CompRes" represents the resolution of customer complaints. The mean rating for CompRes is 5.442, indicating an average level of satisfaction with complaint resolution. The ratings range from 2.6 to 7.8, suggesting that customers have varied experiences when it comes to resolving their complaints. The standard deviation of 1.208403 reflects moderate variability in customer opinions regarding complaint resolution.

- ❑ Advertising (Advertising) has a mean rating of 4.010, suggesting a moderate level of satisfaction. The ratings range from 1.9 to 6.5, with a standard deviation of 1.126943, indicating moderate variability in customer opinions.

- ❑ Product Line (ProdLine) has a mean rating of 5.805, indicating a relatively higher satisfaction level. The ratings range from 2.3 to 8.4, with a standard deviation of 1.315285, suggesting some variability in customer opinions.

- ❑ Sales Force Image (SalesFImage) has a mean rating of 5.123, indicating an average satisfaction level. The ratings range from 2.9 to 8.2, with a standard deviation of 1.072320, indicating moderate variability in customer opinions.

- ❑ Competitive Pricing (ComPricing): The attribute "ComPricing" pertains to the competitive pricing of products or services. It has a mean rating of 6.974, suggesting a relatively higher level of satisfaction with competitive pricing. The ratings range from 3.7 to 9.9, indicating a wide range of opinions on pricing. The standard deviation of 1.545055 indicates some variability in customer perceptions of pricing competitiveness.

- ❑ Warranty Claims (WartyClaim) has a mean rating of 6.043, indicating an average satisfaction level. The ratings range from 4.1 to 8.1, with a relatively low standard deviation of 0.819738, suggesting less variability in customer opinions.

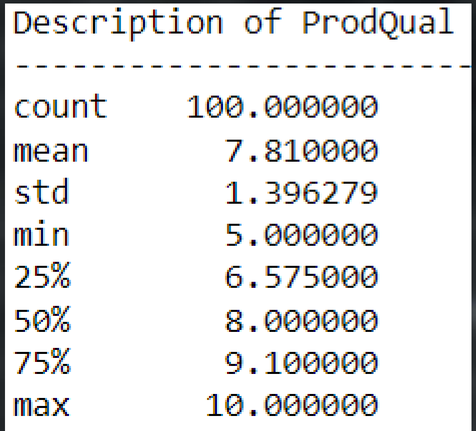
- ❑ Order Billing (OrdBilling) has a mean rating of 4.278, indicating a moderate satisfaction level. The ratings range from 2.0 to 6.7, with a standard deviation of 0.928840, suggesting moderate variability in customer opinions.

- ❑ Delivery Speed (DelSpeed) has a mean rating of 3.886, indicating a moderate satisfaction level. The ratings range from 1.6 to 5.5, with a standard deviation of 0.734437, suggesting moderate variability in customer opinions.

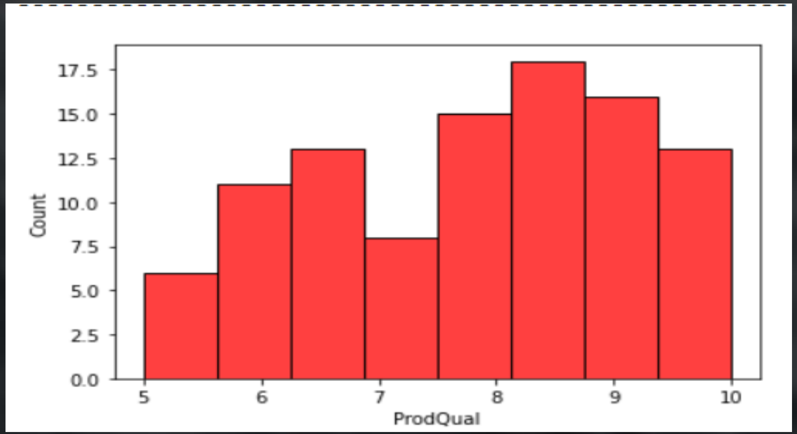
- ❑ Customer Satisfaction (Satisfaction): The attribute "Satisfaction" represents overall customer satisfaction. It has a mean rating of 6.918, indicating a relatively higher level of customer satisfaction. The ratings range from 4.7 to 9.9, suggesting a wide range of satisfaction levels among customers. The standard deviation of 1.191839 indicates some variability in customer satisfaction levels.

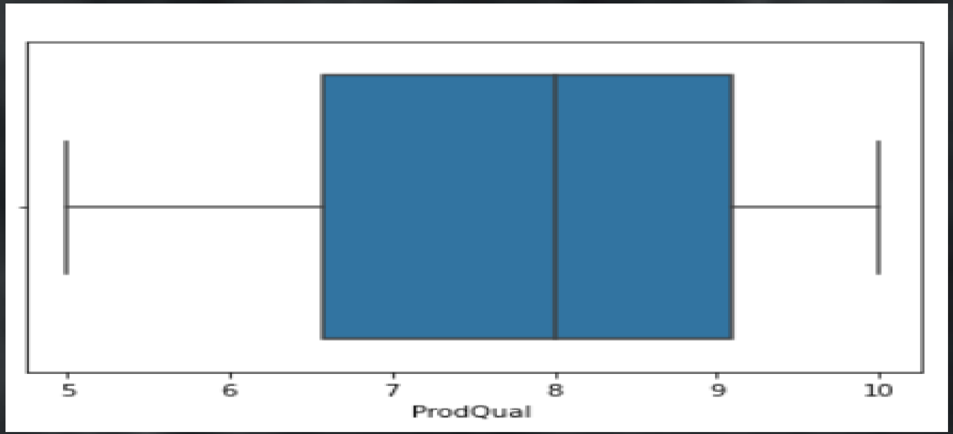
Univariate Analysis

Product Quality (ProdQual)

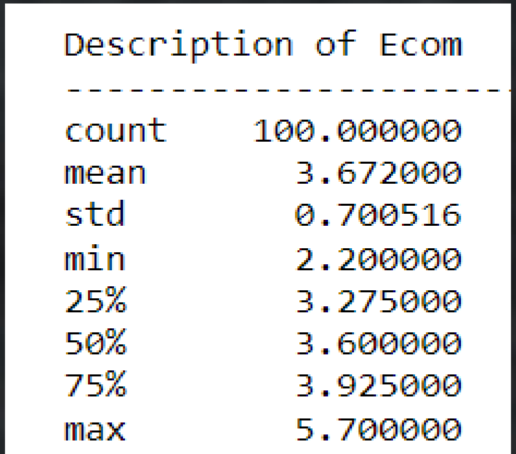


```
Description of ProdQual
------------------------
count     100.000000
mean        7.810000
std         1.396279
min         5.000000
25%         6.575000
50%         8.000000
75%         9.100000
max        10.000000
```

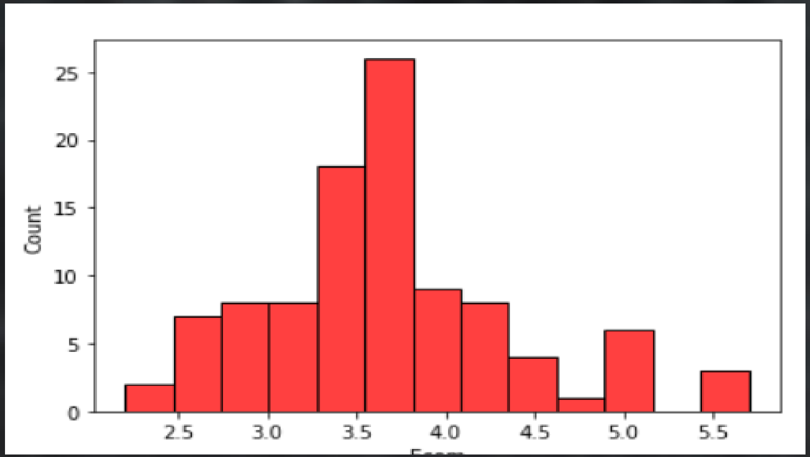


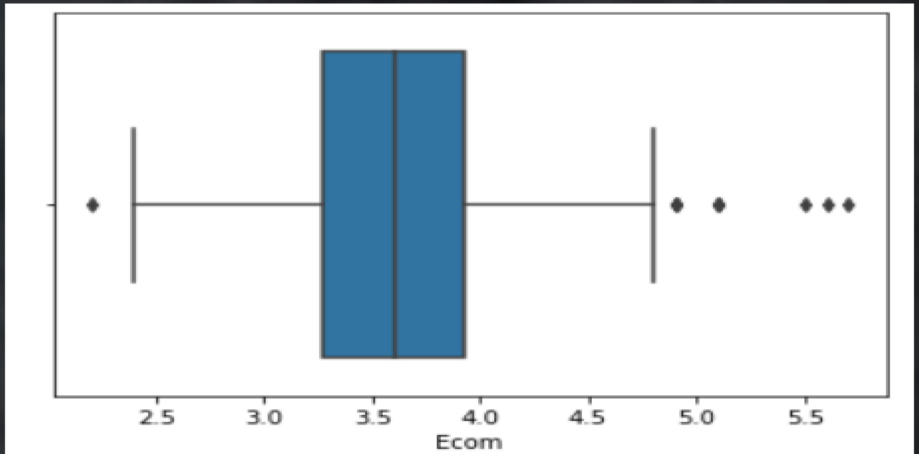


E-commerce (Ecom)

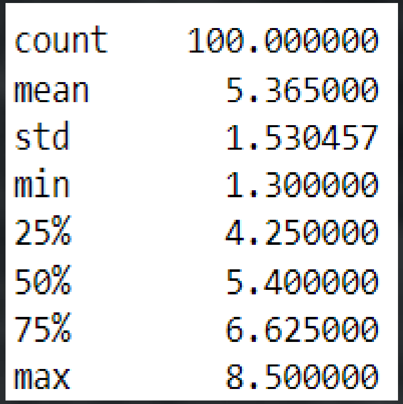


```
Description of Ecom
------------------------
count     100.000000
mean        3.672000
std         0.700516
min         2.200000
25%         3.275000
50%         3.600000
75%         3.925000
max         5.700000
```

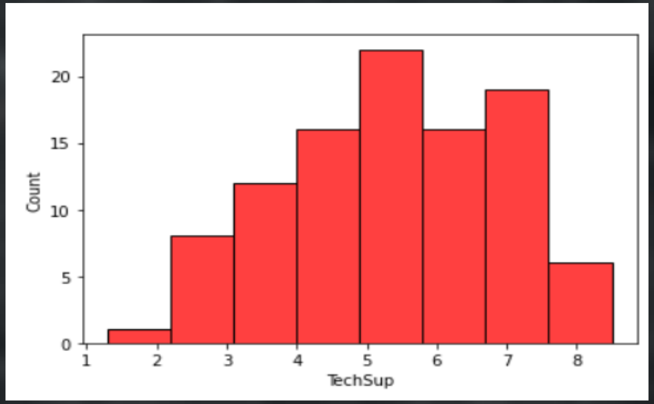


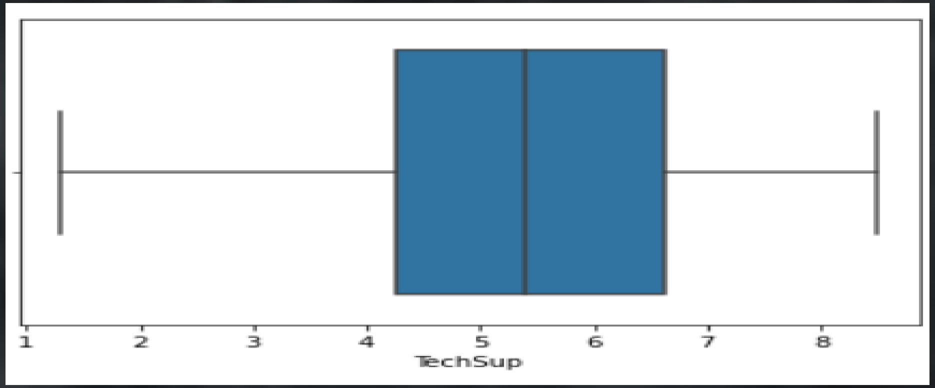


Technical Support (TechSup)

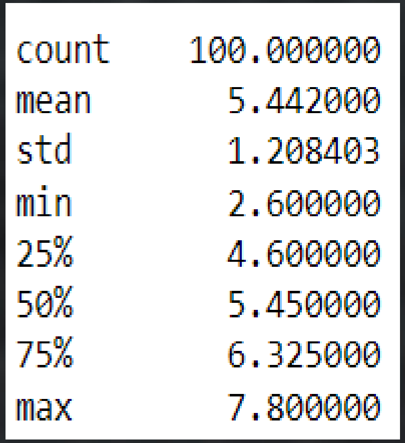


```
count    100.000000
mean       5.365000
std        1.530457
min        1.300000
25%        4.250000
50%        5.400000
75%        6.625000
max        8.500000
```

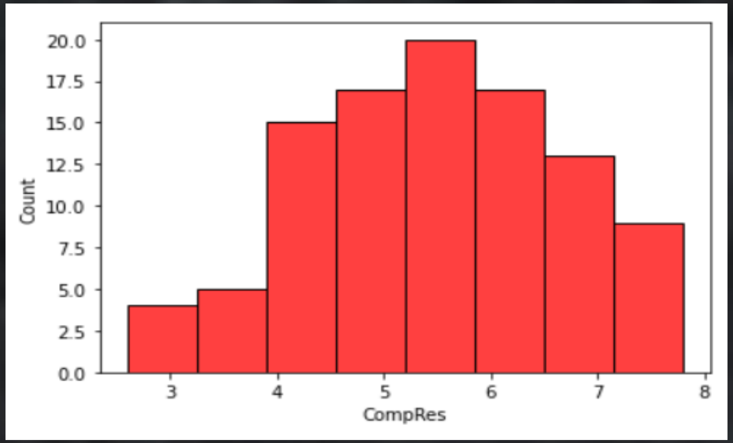


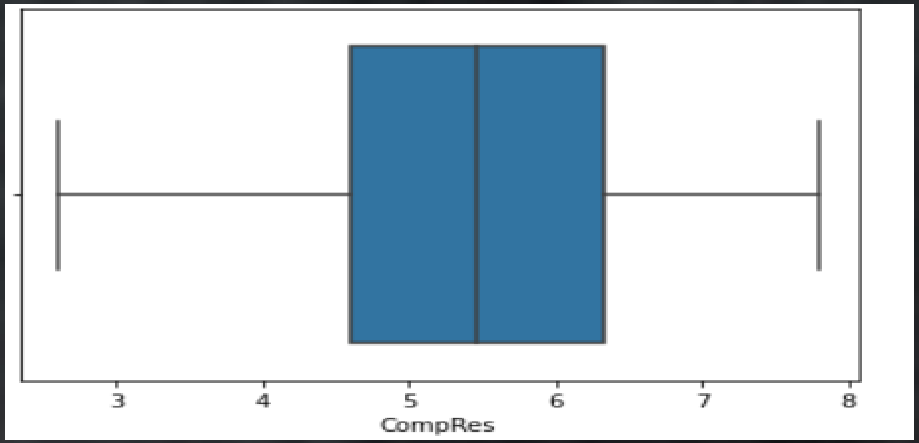


Complaint Resolution (CompRes):

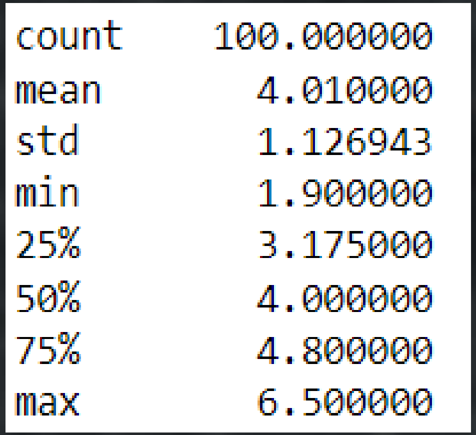


```
count    100.000000
mean       5.442000
std        1.208403
min        2.600000
25%        4.600000
50%        5.450000
75%        6.325000
max        7.800000
```

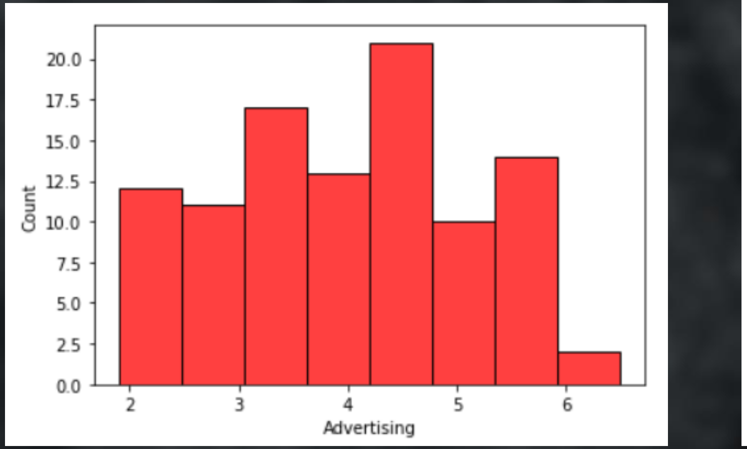


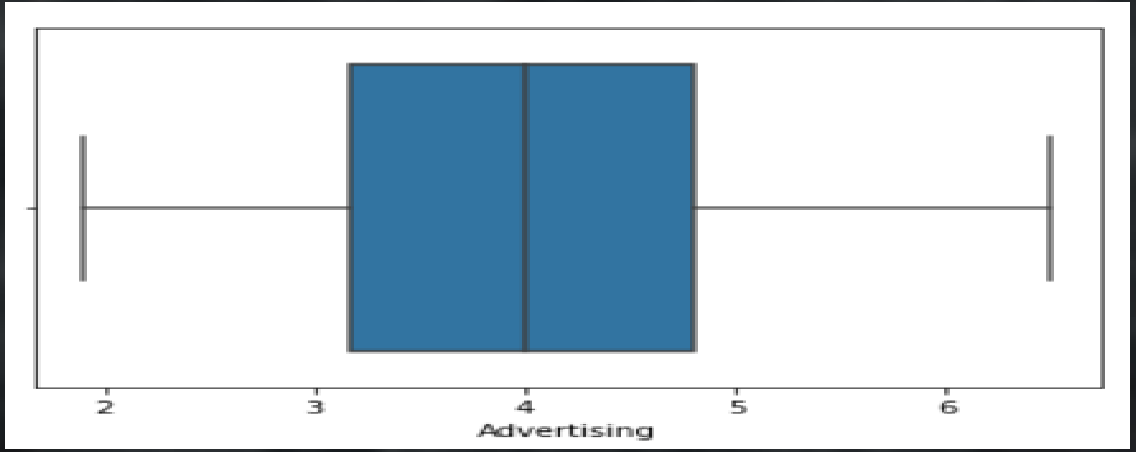


Advertising (Advertising)

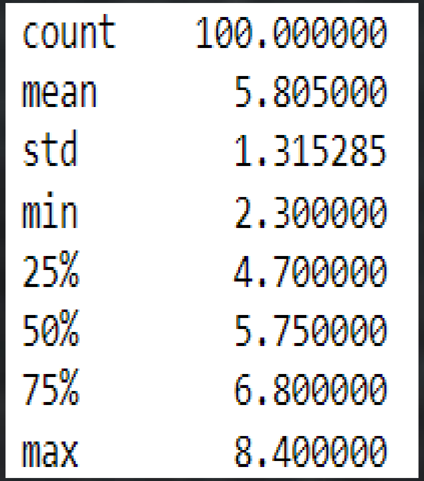


```
count    100.000000
mean       4.010000
std        1.126943
min        1.900000
25%        3.175000
50%        4.000000
75%        4.800000
max        6.500000
```

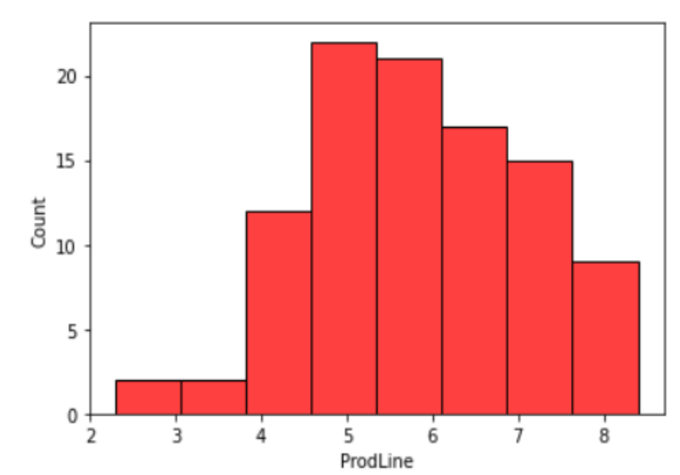


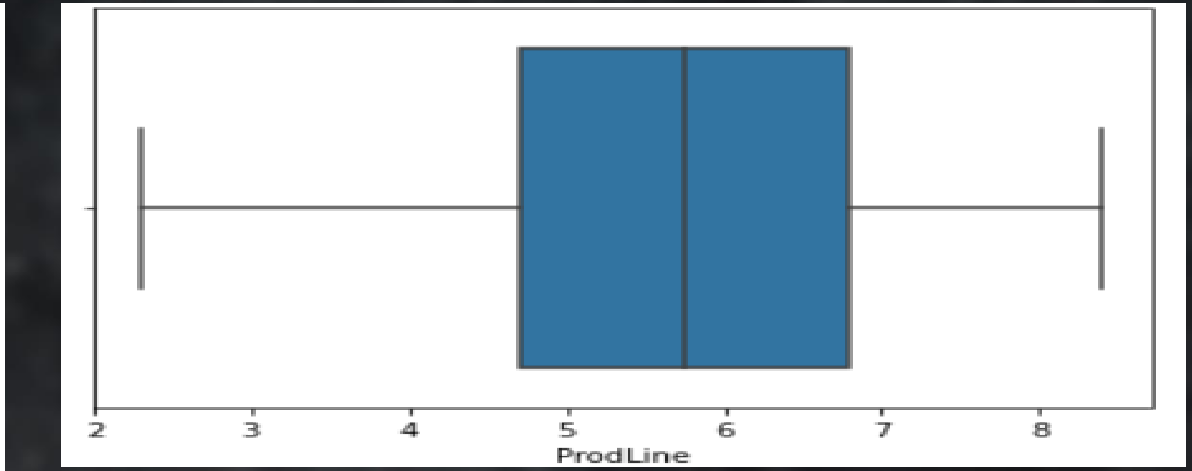


Product Line (ProdLine)

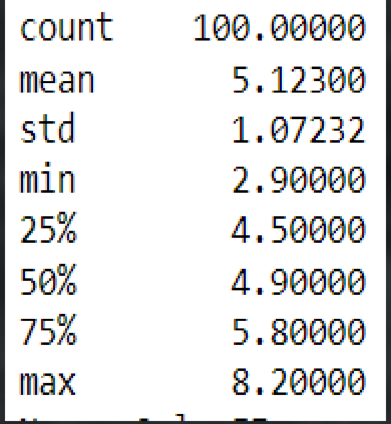


```
count    100.000000
mean       5.805000
std        1.315285
min        2.300000
25%        4.700000
50%        5.750000
75%        6.800000
max        8.400000
```

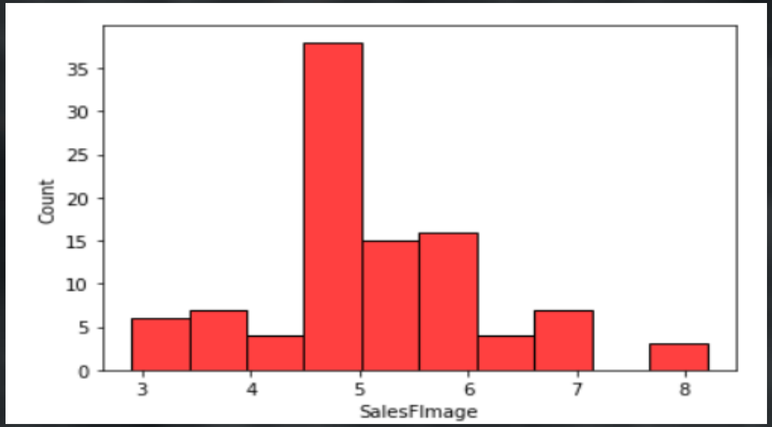


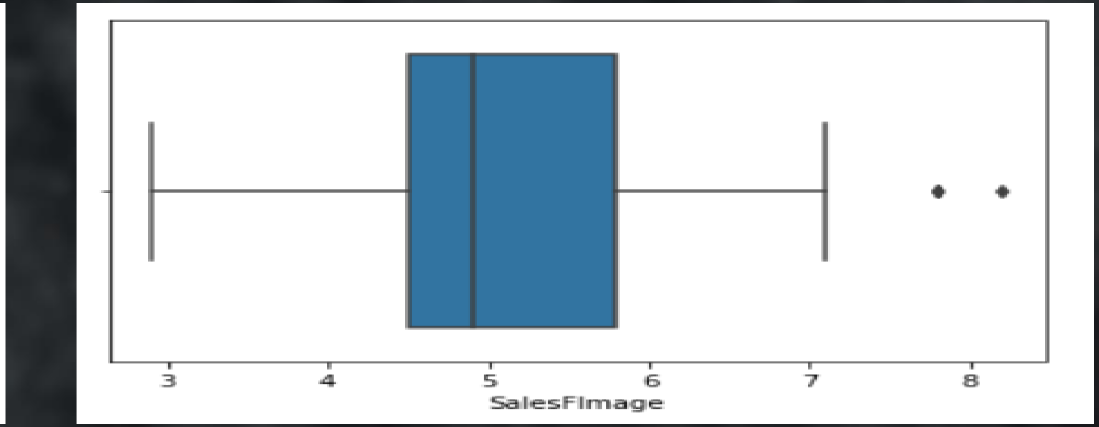


Sales Force Image (SalesFImage)

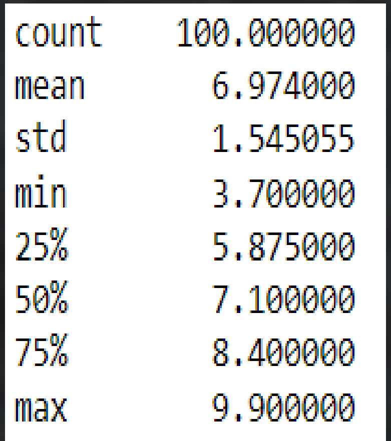


```
count    100.00000
mean       5.12300
std        1.07232
min        2.90000
25%        4.50000
50%        4.90000
75%        5.80000
max        8.20000
```

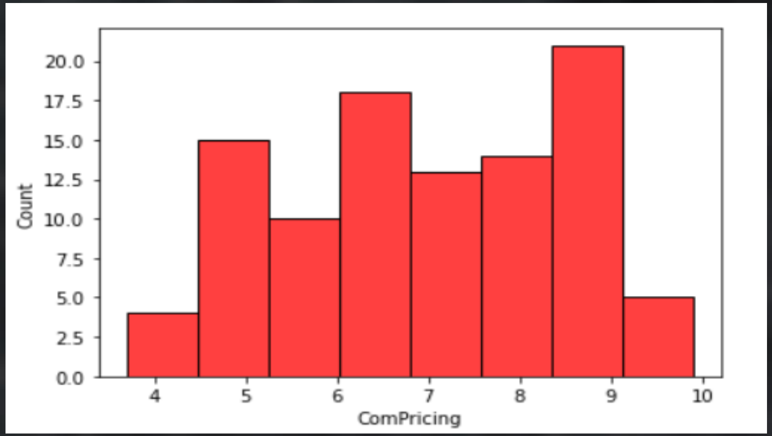


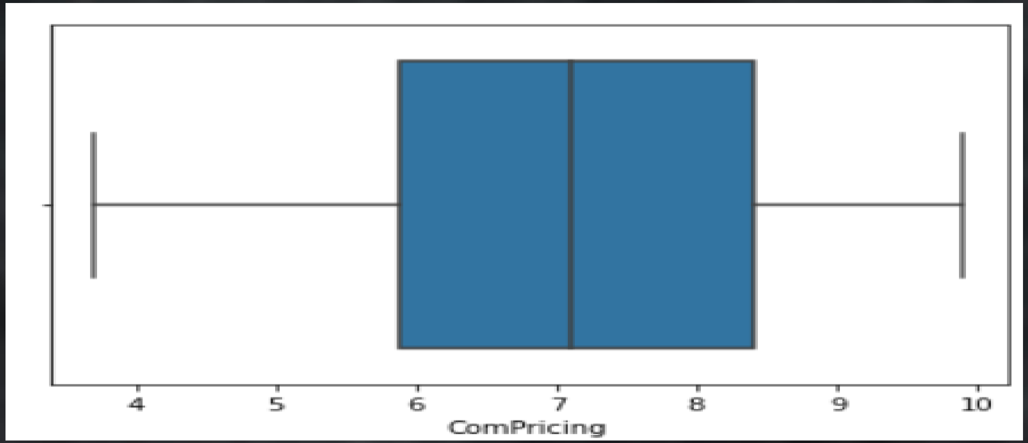


Competitive Pricing (ComPricing):

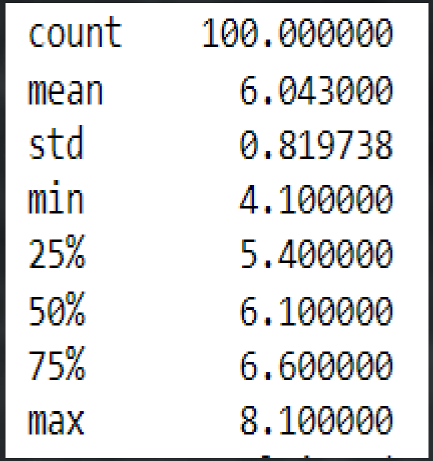


```
count    100.000000
mean       6.974000
std        1.545055
min        3.700000
25%        5.875000
50%        7.100000
75%        8.400000
max        9.900000
```

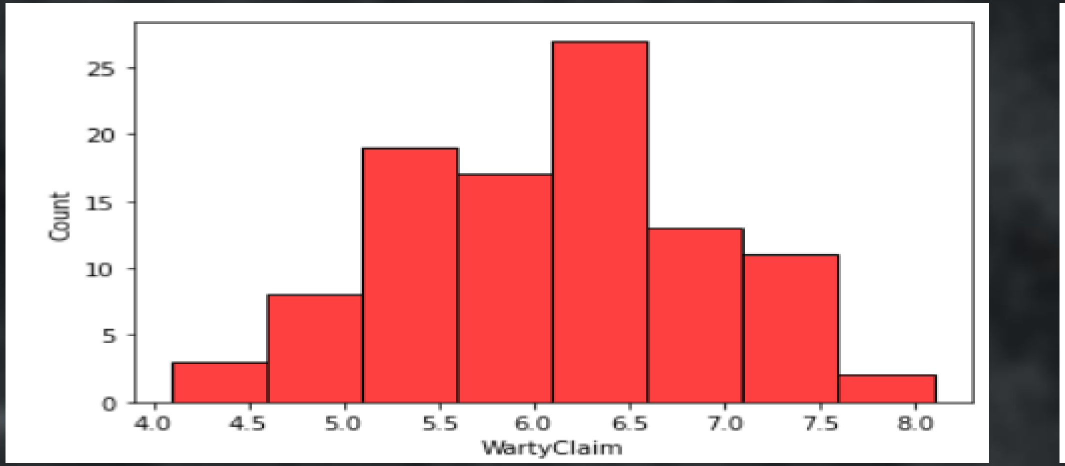


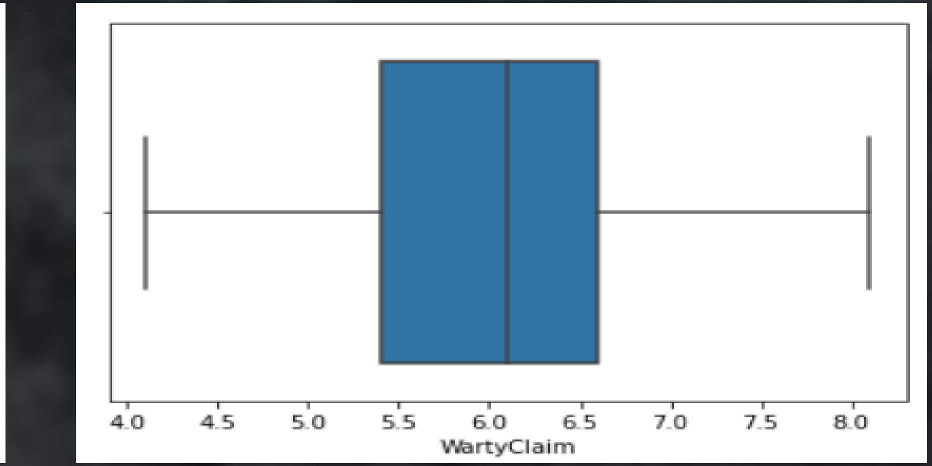


Warranty Claims (WartyClaim)

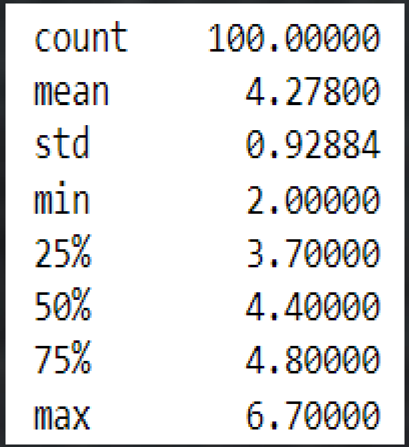


```
count    100.000000
mean       6.043000
std        0.819738
min        4.100000
25%        5.400000
50%        6.100000
75%        6.600000
max        8.100000
```

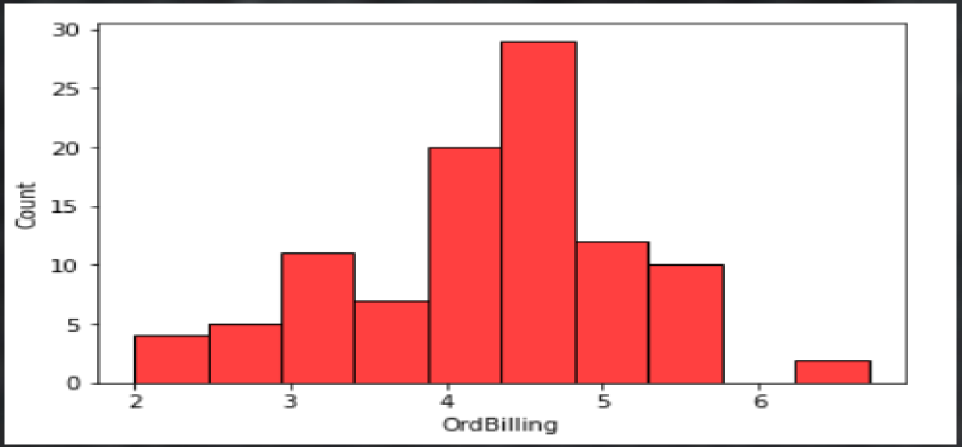


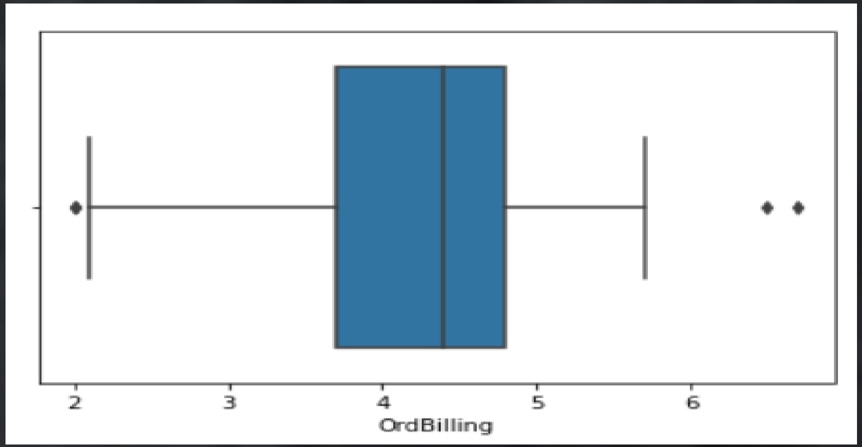


Order Billing (OrdBilling)

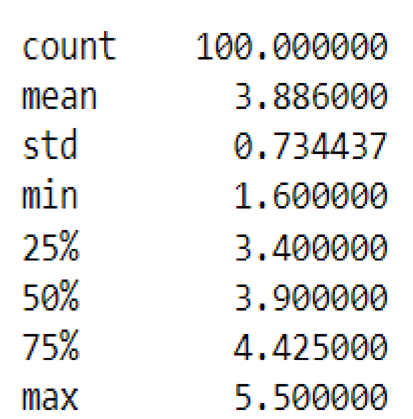


```
count    100.00000
mean       4.27800
std        0.92884
min        2.00000
25%        3.70000
50%        4.40000
75%        4.80000
max        6.70000
```

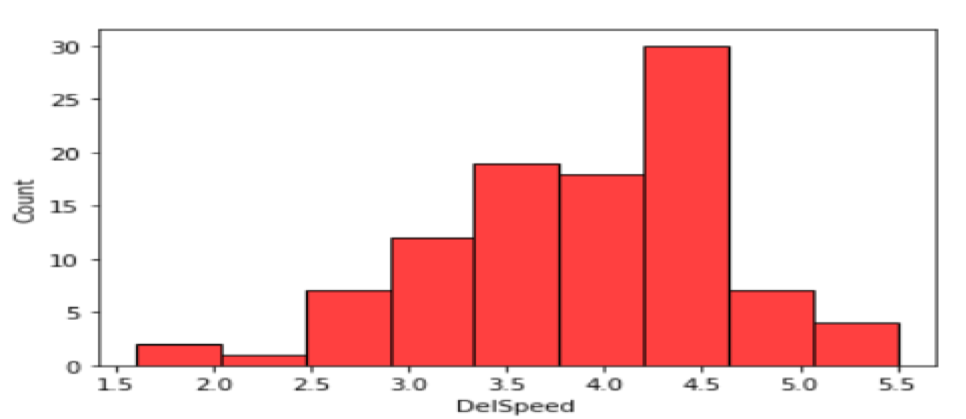


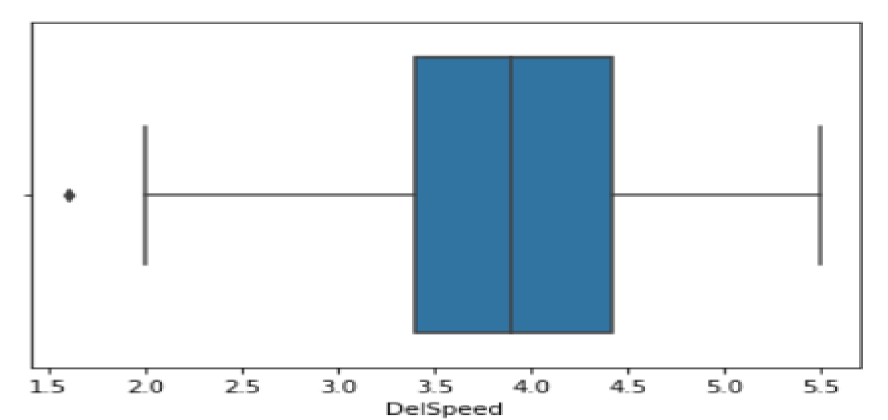


Delivery Speed (DelSpeed)

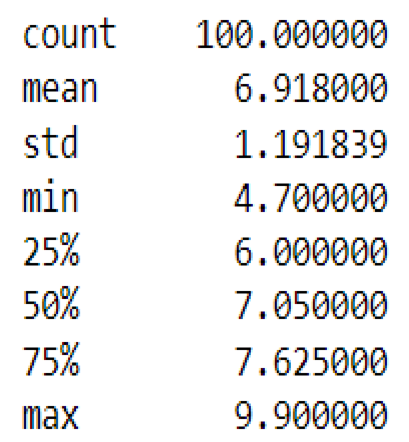


```
count    100.000000
mean       3.886000
std        0.734437
min        1.600000
25%        3.400000
50%        3.900000
75%        4.425000
max        5.500000
```

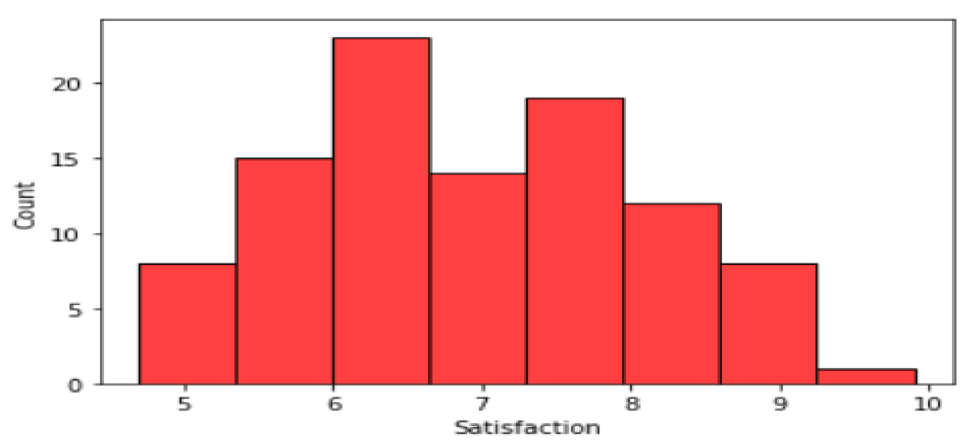


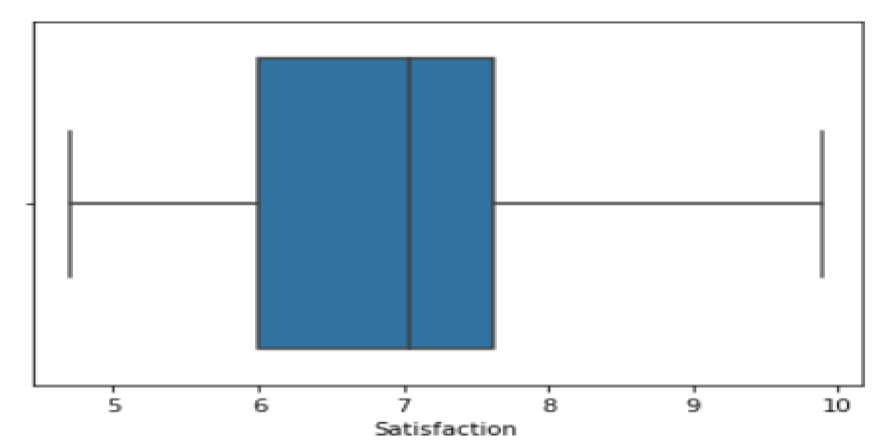


Customer Satisfaction (Satisfaction):



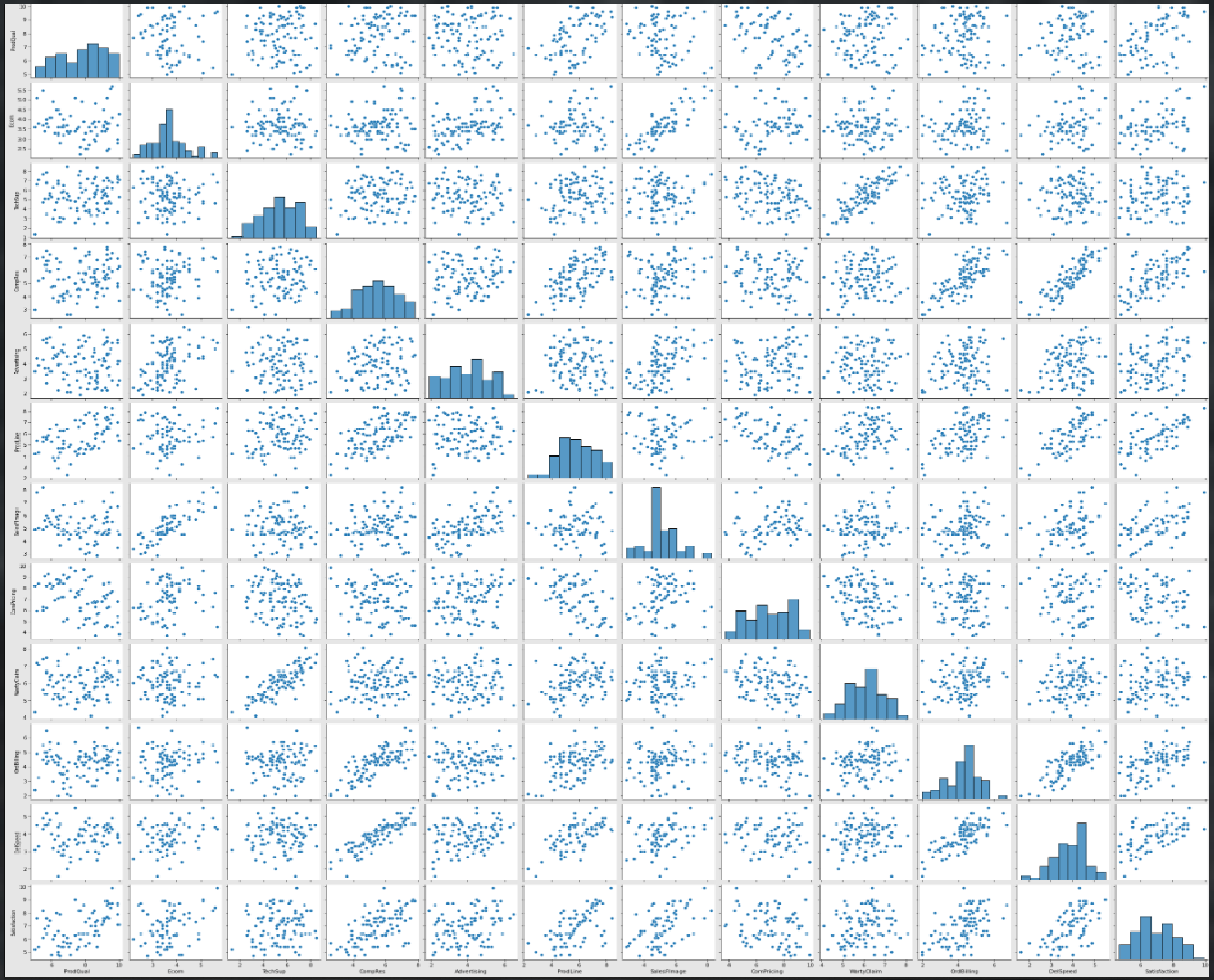
```
count    100.000000
mean       6.918000
std        1.191839
min        4.700000
25%        6.000000
50%        7.050000
75%        7.625000
max        9.900000
```





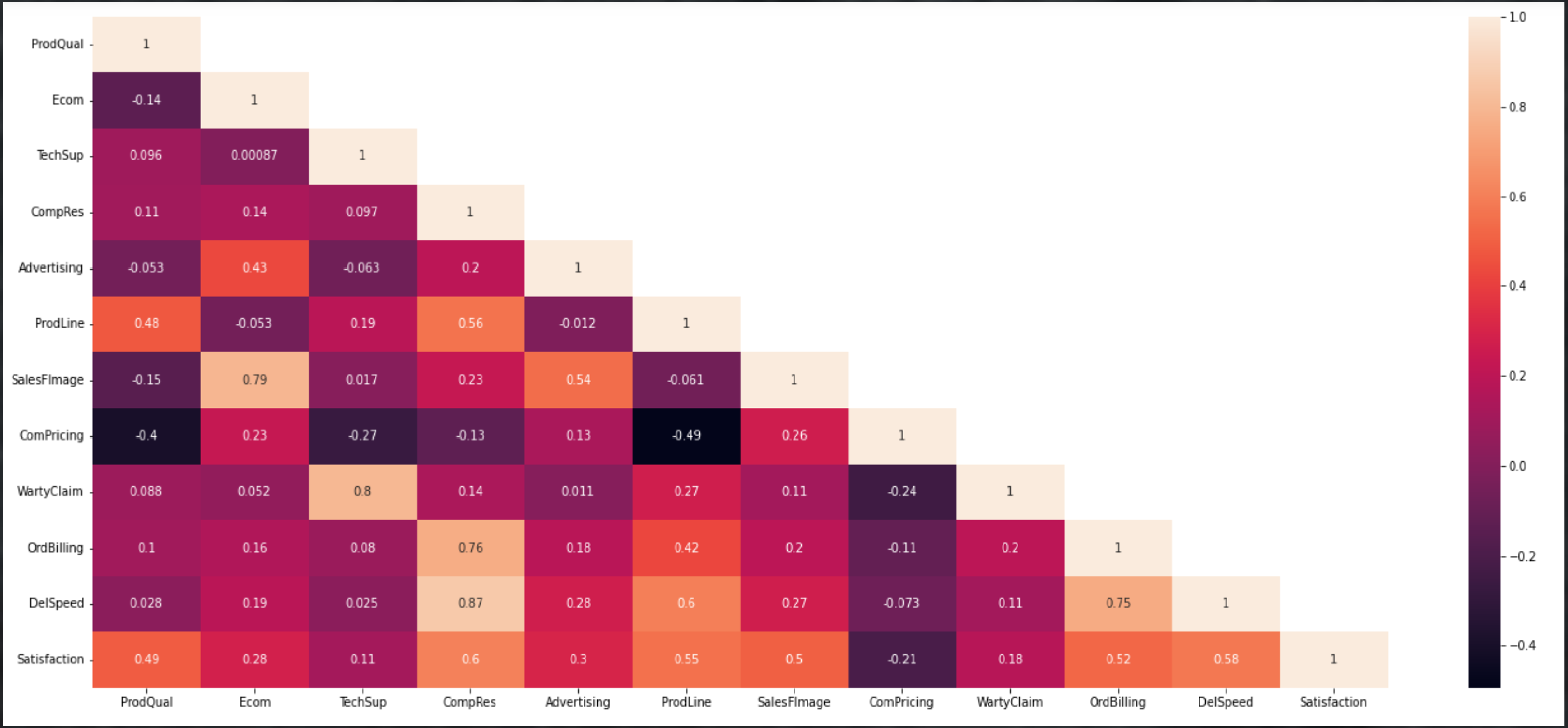
Multivariate Analysis

Pair plot



We can observe that certain variables in the data exhibit correlation.

Heatmap



There is a correlation between the variables in the above graph. So we have visualize this correlation using a heatmap, where lighter colors indicate stronger correlation and darker colors indicate weaker correlation.

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

Since all the variables in this scenario are on the same scale, scaling is not necessary. However, for the sake of standardization, we will perform scaling here. It should be noted that scaling is being done primarily from a covariance perspective.

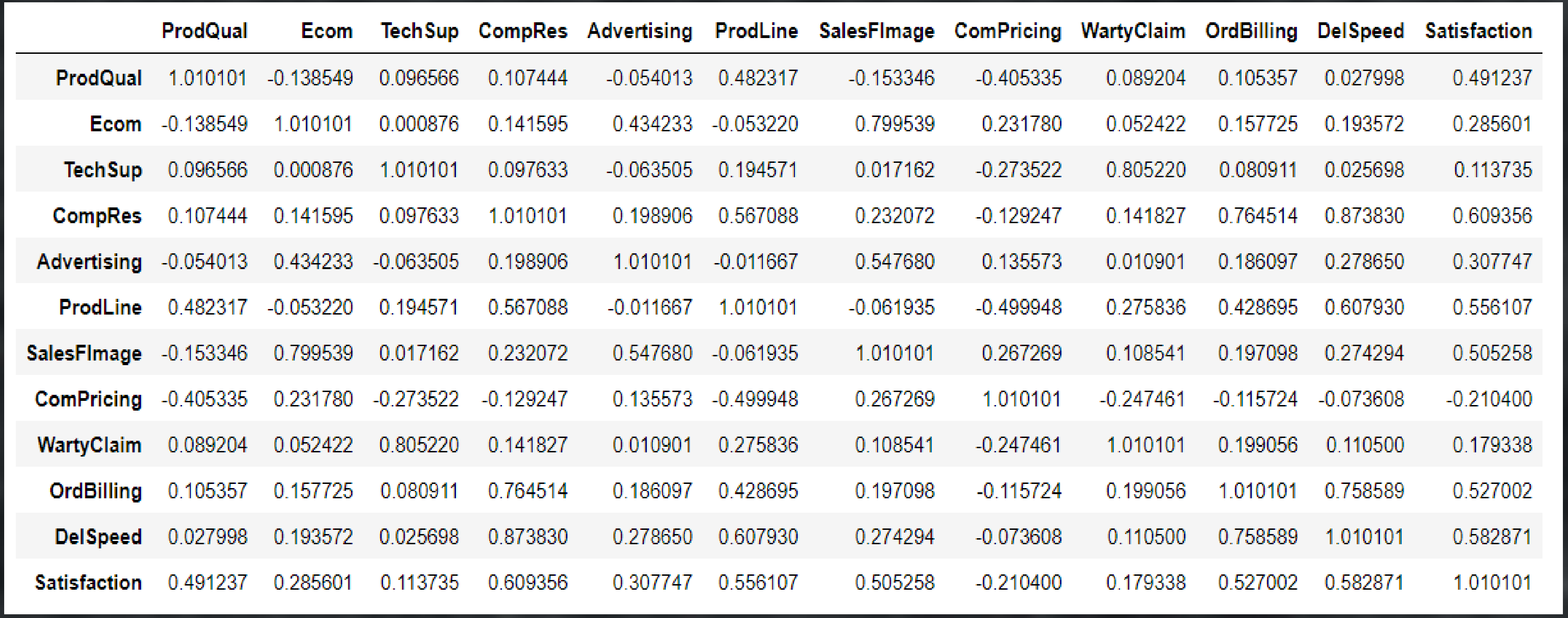
We will perform data scaling using the Z-score method with the help of the “scipy.stats’ library :

- The primary reason for scaling data using the z-score is to ensure that all features (variables) are on a similar scale. When features have different scales, it can lead to certain variables dominating the learning process, potentially resulting in biased outcomes.

- The z-score scaling method transforms the values of each feature so that they have a mean of 0 and a standard deviation of 1. This is achieved by subtracting the mean of the feature from each value and dividing by the standard deviation. As a result, the standardized values will have a mean of 0 and a standard deviation of 1, making them comparable across all features.

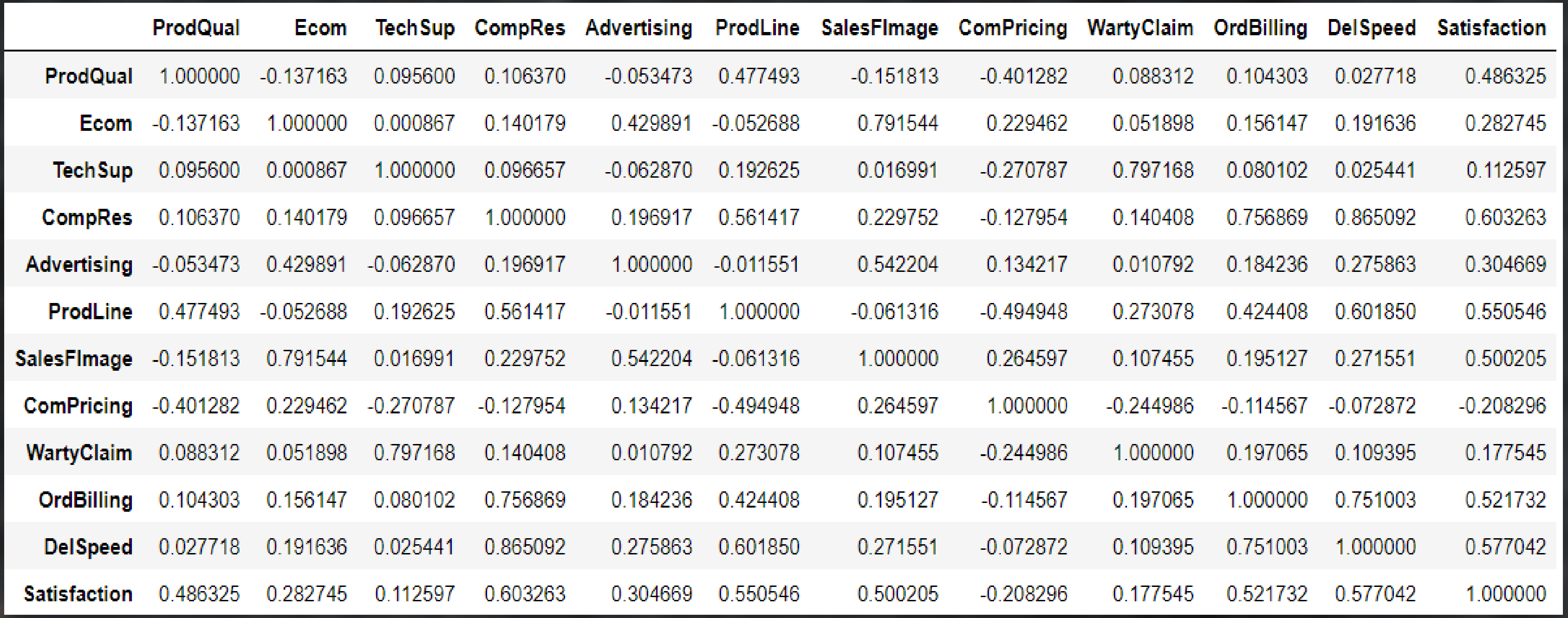
Comment on the comparison between covariance and the correlation matrix after scaling.

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

Correlation Matrix

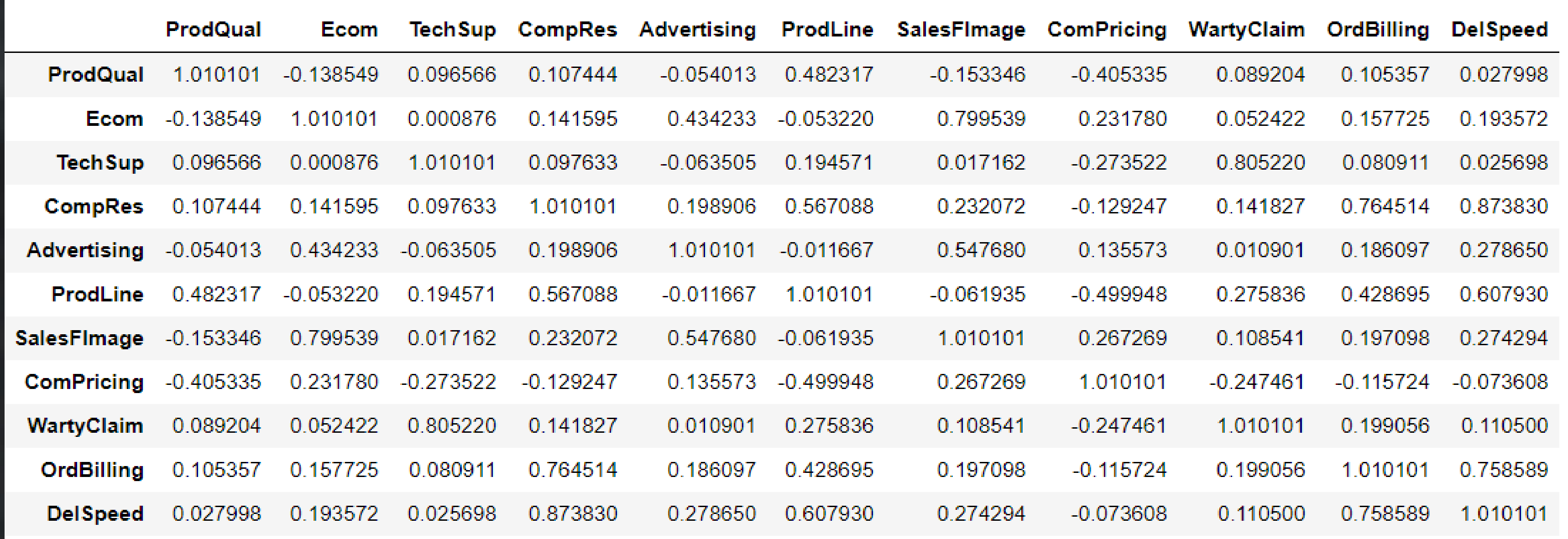


	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
ProdQual	1.000000	-0.137163	0.095600	0.106370	-0.053473	0.477493	-0.151813	-0.401282	0.088312	0.104303	0.027718	0.486325
Ecom	-0.137163	1.000000	0.000867	0.140179	0.429891	-0.052688	0.791544	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

When comparing the two matrices, it becomes evident that the values in the correlation matrix closely resemble those in the covariance matrix. This similarity arises because correlation is derived from covariance by dividing each covariance value by the product of the standard deviations of the corresponding variables. Consequently, the correlation matrix presents a standardized perspective of the relationships, enabling easier interpretation and comparison across variables.

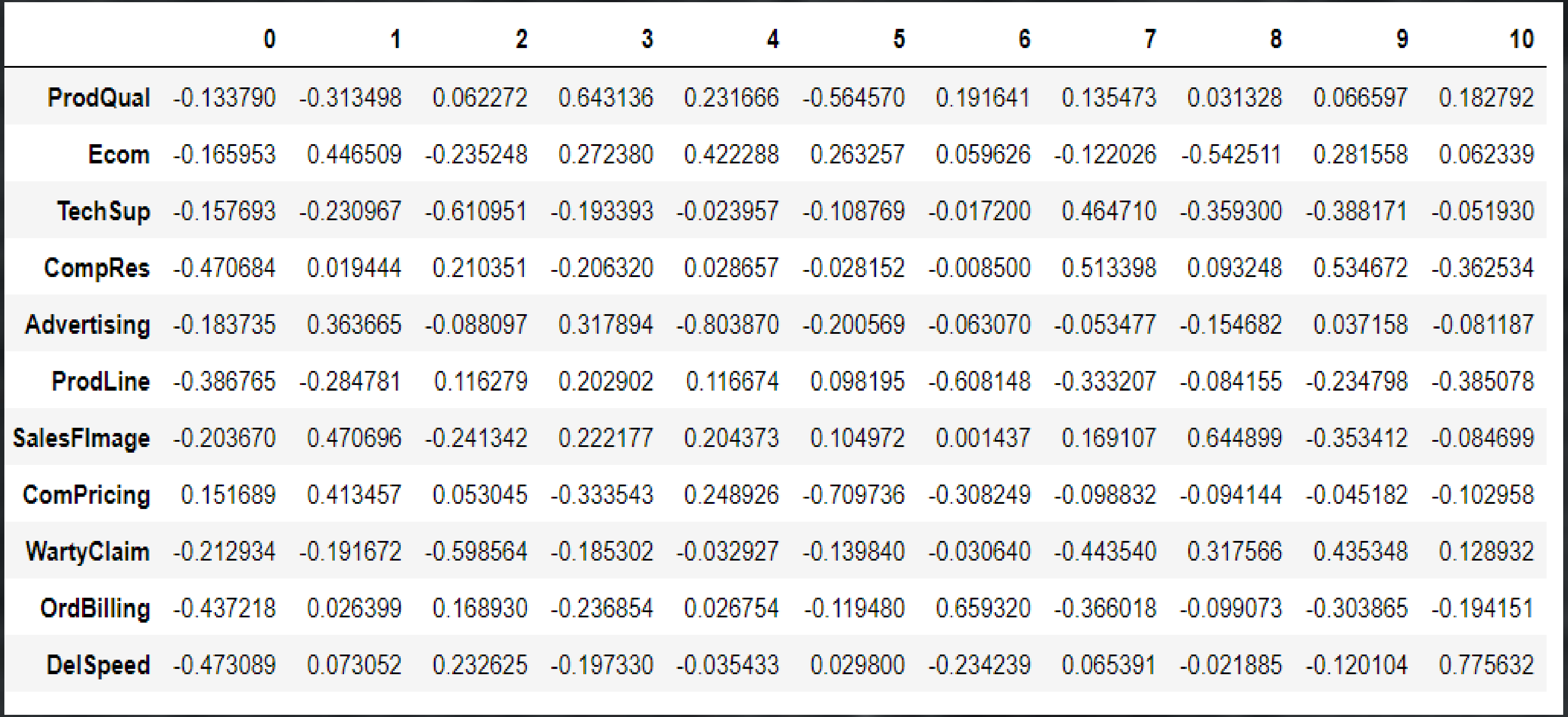
To summarize, the covariance matrix offers insights into the strength and direction of relationships between variables, whereas the correlation matrix provides a standardized measure of the linear relationships. This standardization makes the correlation matrix more convenient for analysis and interpretation purposes.

Build the covariance matrix, eigenvalues and eigenvector



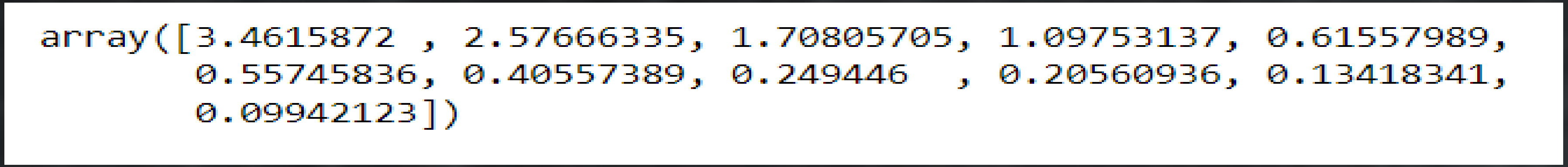
	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
ProdQual	1.010101	-0.138549	0.096566	0.107444	-0.054013	0.482317	-0.153346	-0.405335	0.089204	0.105357	0.027998
Ecom	-0.138549	1.010101	0.000876	0.141595	0.434233	-0.053220	0.799539	0.231780	0.052422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

Covariance Matrix



	0	1	2	3	4	5	6	7	8	9	10
ProdQual	-0.133790	-0.313498	0.062272	0.643136	0.231666	-0.564570	0.191641	0.135473	0.031328	0.066597	0.182792
Ecom	-0.165953	0.446509	-0.235248	0.272380	0.422288	0.263257	0.059626	-0.122026	-0.542511	0.281558	0.062339
TechSup	-0.157693	-0.230967	-0.610951	-0.193393	-0.023957	-0.108769	-0.017200	0.464710	-0.359300	-0.388171	-0.051930
CompRes	-0.470684	0.019444	0.210351	-0.206320	0.028657	-0.028152	-0.008500	0.513398	0.093248	0.534672	-0.362534
Advertising	-0.183735	0.363665	-0.088097	0.317894	-0.803870	-0.200569	-0.063070	-0.053477	-0.154682	0.037158	-0.081187
ProdLine	-0.386765	-0.284781	0.116279	0.202902	0.116674	0.098195	-0.608148	-0.333207	-0.084155	-0.234798	-0.385078
SalesFImage	-0.203670	0.470696	-0.241342	0.222177	0.204373	0.104972	0.001437	0.169107	0.644899	-0.353412	-0.084699
ComPricing	0.151689	0.413457	0.053045	-0.333543	0.248926	-0.709736	-0.308249	-0.098832	-0.094144	-0.045182	-0.102958
WartyClaim	-0.212934	-0.191672	-0.598564	-0.185302	-0.032927	-0.139840	-0.030640	-0.443540	0.317566	0.435348	0.128932
OrdBilling	-0.437218	0.026399	0.168930	-0.236854	0.026754	-0.119480	0.659320	-0.366018	-0.099073	-0.303865	-0.194151
DelSpeed	-0.473089	0.073052	0.232625	-0.197330	-0.035433	0.029800	-0.234239	0.065391	-0.021885	-0.120104	0.775632

Eigen Vectors



```
array([3.4615872 , 2.57666335, 1.70805705, 1.09753137, 0.61557989,
       0.55745836, 0.40557389, 0.249446  , 0.20560936, 0.13418341,
       0.09942123])
```

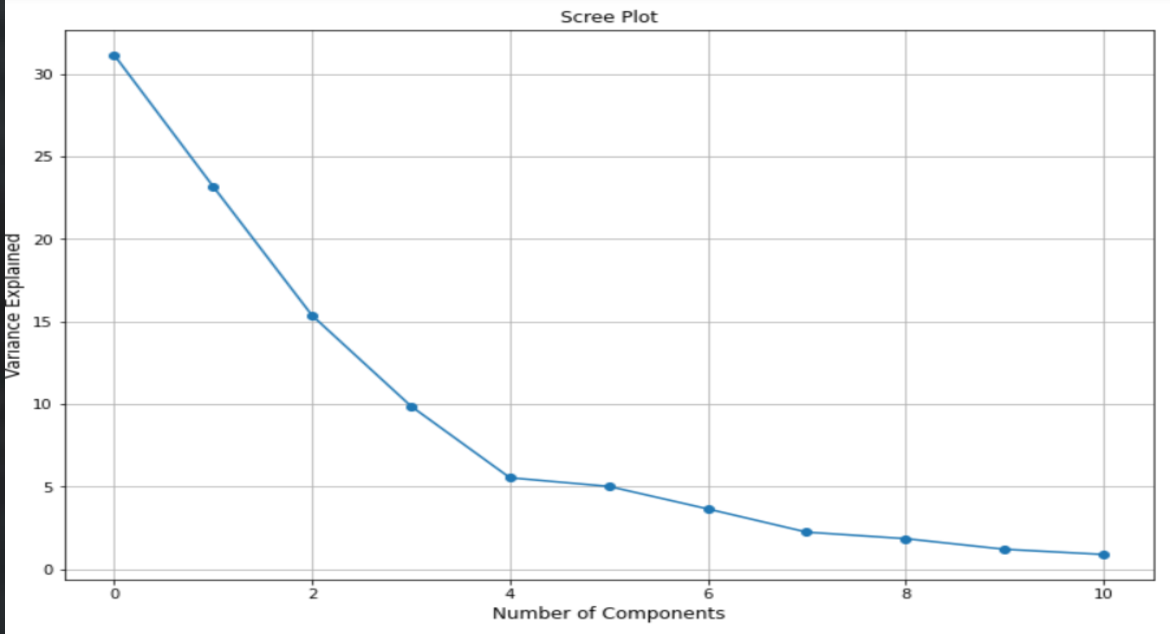
Eigen Values

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

PC1 = -0.134 * ProdQual - 0.166 * Ecom - 0.158 * TechSup - 0.471 * CompRes - 0.184 * Advertising - 0.387 * ProdLine - 0.204 * SalesFImage + 0.152 * ComPricing - 0.213 * WartyClaim - 0.437 * OrdBilling - 0.473 * DelSpeed

In this equation, each variable is multiplied by its respective eigen vector coefficient and then summed together. The first principal component represents a linear combination of the variables, weighted according to their corresponding eigen vector coefficients.

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.



From the plot, it is evident that we can choose 5 components to capture the majority of the variance in the data. Additionally, by selecting these 5 components, we are able to explain approximately 85.135% of the total variance in the dataset.

The cumulative values of the eigenvalues help in determining the optimal number of principal components to select. By examining the cumulative variance explained, we can see the percentage of total variance accounted for by including a certain number of components. In this case, the cumulative variance explained is shown as follows:

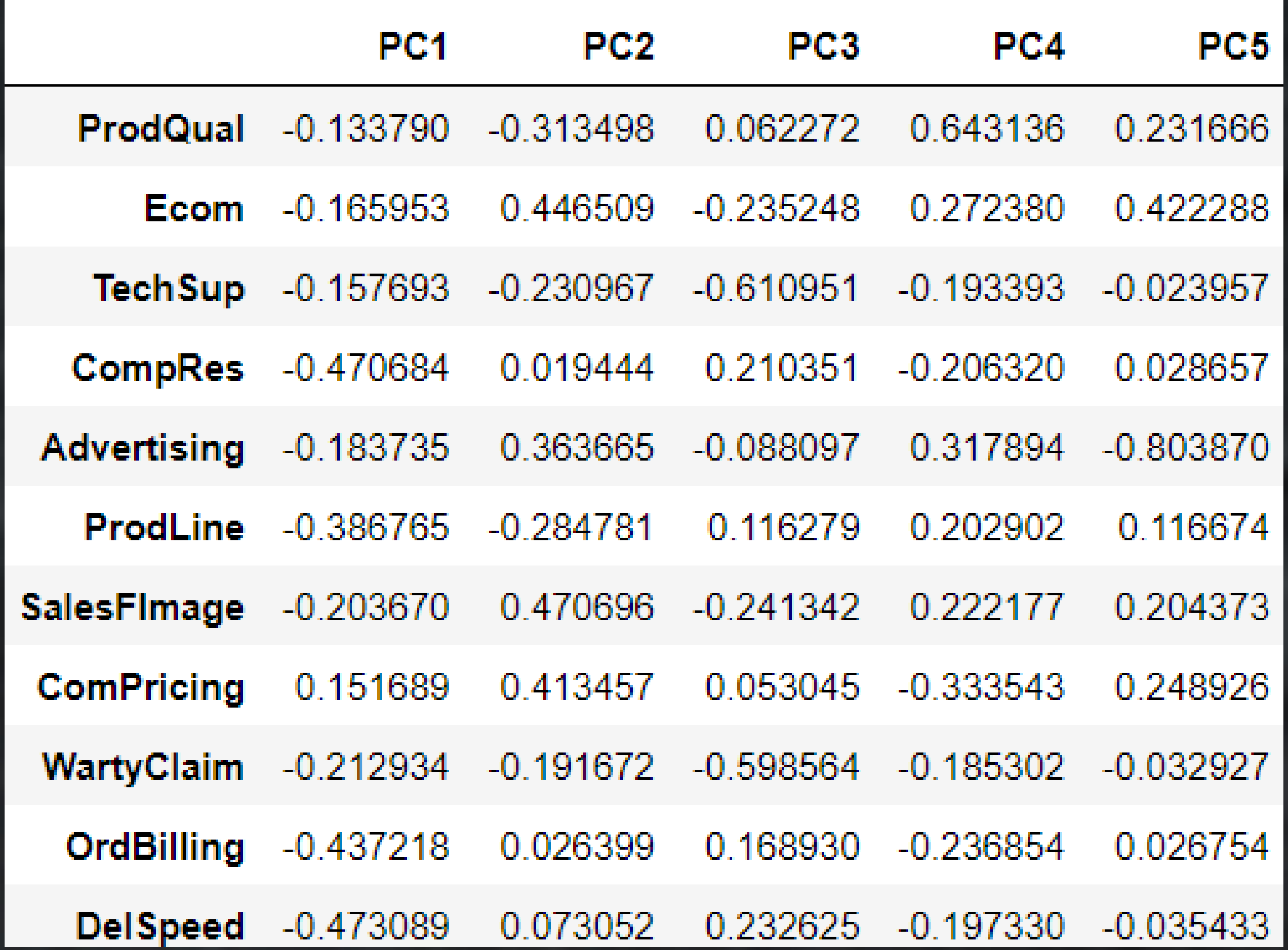
[31.154, 54.344, 69.717, 79.595, 85.135, 90.152, 93.802 ,96.047, 97.898, 99.105, 100.]

These values indicate that as we include more principal components, the cumulative variance explained gradually increases. The cumulative variance explained is a useful metric for deciding the number of components to retain. It helps determine the point at which adding additional components does not significantly contribute to explaining the overall variance in the data.

The eigenvectors in the PCA analysis represent the directions or axes in the original feature space along which the data varies the most. Each eigenvector corresponds to a principal component and indicates the relative contribution of each original variable to that component.

By examining the eigenvectors, we can understand the weights or loadings of each variable on the principal components.

The eigenvectors for the first five principal components (PC1 to PC5) are displayed:

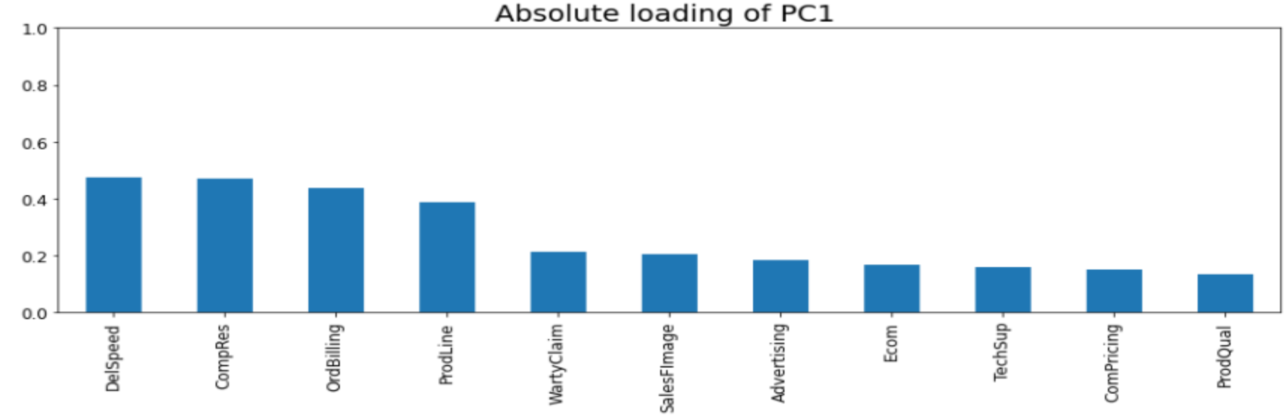


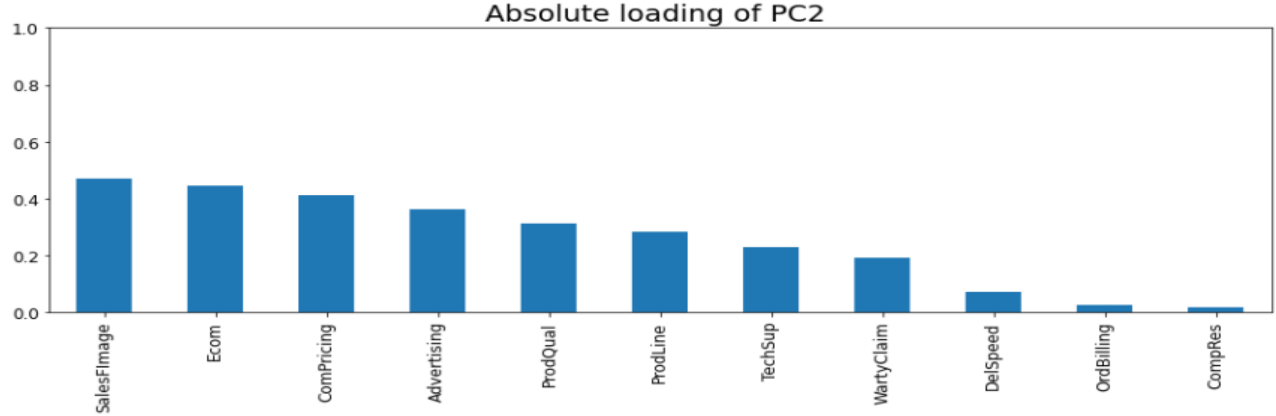
	PC1	PC2	PC3	PC4	PC5
ProdQual	-0.133790	-0.313498	0.062272	0.643136	0.231666
Ecom	-0.165953	0.446509	-0.235248	0.272380	0.422288
TechSup	-0.157693	-0.230967	-0.610951	-0.193393	-0.023957
CompRes	-0.470684	0.019444	0.210351	-0.206320	0.028657
Advertising	-0.183735	0.363665	-0.088097	0.317894	-0.803870
ProdLine	-0.386765	-0.284781	0.116279	0.202902	0.116674
SalesFImage	-0.203670	0.470696	-0.241342	0.222177	0.204373
ComPricing	0.151689	0.413457	0.053045	-0.333543	0.248926
WartyClaim	-0.212934	-0.191672	-0.598564	-0.185302	-0.032927
OrdBilling	-0.437218	0.026399	0.168930	-0.236854	0.026754
DelSpeed	-0.473089	0.073052	0.232625	-0.197330	-0.035433

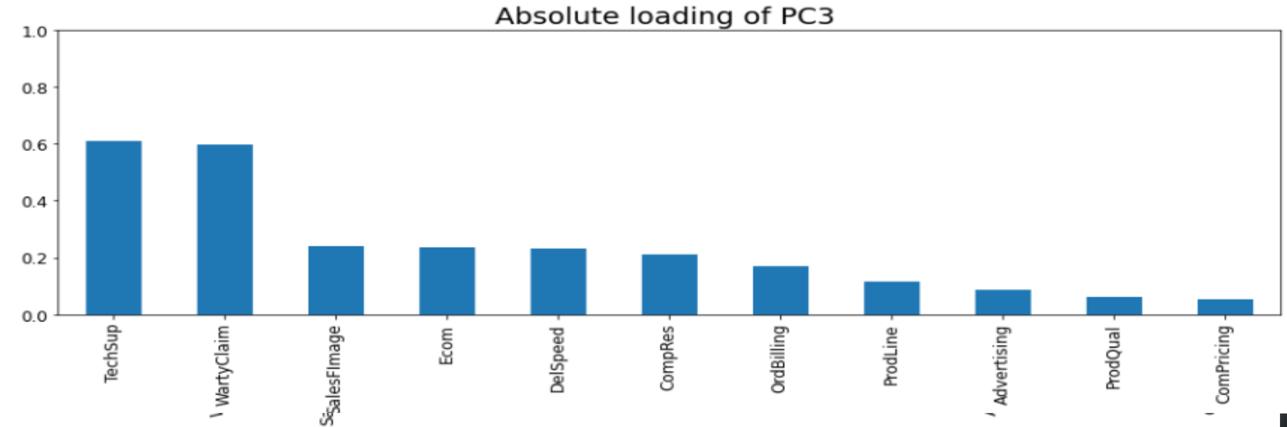
These eigenvectors provide insights into the contribution and direction of each variable in the principal components.

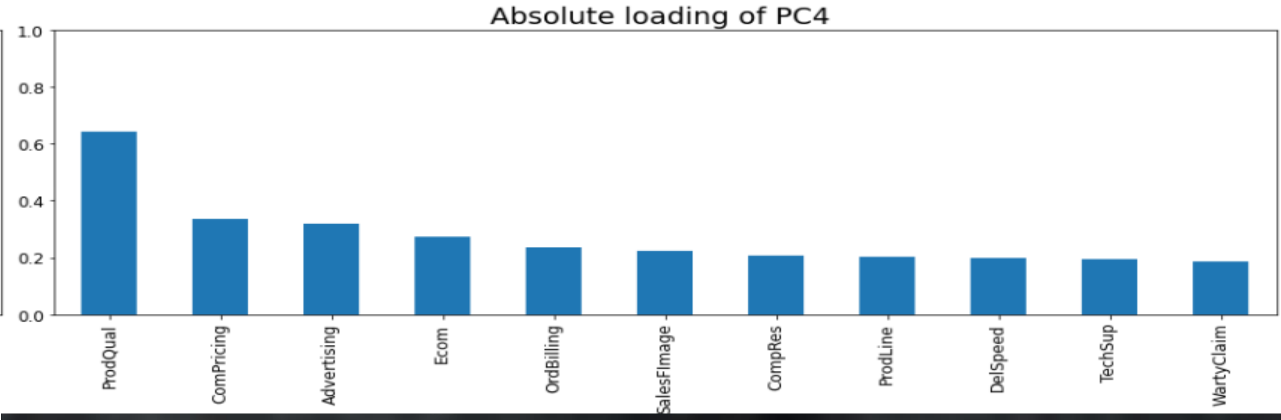
Each principal component represents a linear combination of the original variables, and the eigenvectors determine the weights or coefficients for each variable in that combination.

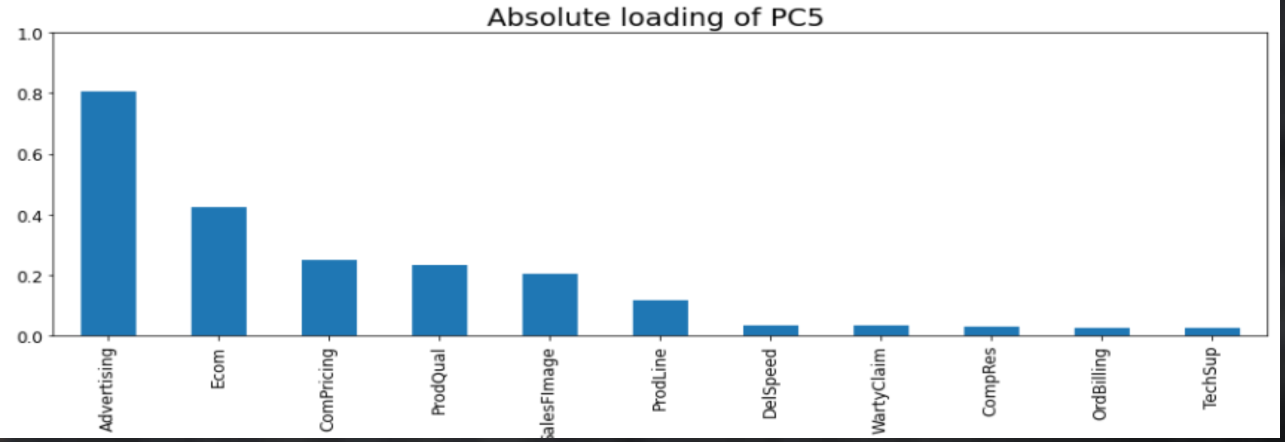
Mention the business implication of using the Principal Component Analysis for this case study.


Absolute loading of PC1


Absolute loading of PC2


Absolute loading of PC3


Absolute loading of PC4


Absolute loading of PC5

Principal components (PCs) are linear combinations of the original variables that are obtained through the process of Principal Component Analysis (PCA). Each PC represents a different pattern or direction of variation in the data.

PC1: PC1 captures the most significant pattern of variation in the data. It is a combination of variables such as DelSpeed, CompRes, and OrdBilling, which contribute positively to PC1. This suggests that higher values of these variables tend to occur together and contribute to the overall variation represented by PC1.

PC2: PC2 represents another important pattern of variation in the data. It is influenced by variables like Ecom, Advertising, SalesFImage, and ComPricing, which have positive correlations with PC2. This indicates that higher values of these variables tend to occur together and contribute to the variation captured by PC2.

PC3: PC3 captures a different pattern of variation characterized by negative correlations with variables like TechSup and WartyClaim. Higher values of TechSup and WartyClaim lead to lower values of PC3. This suggests that these variables are associated with a distinct pattern of variation that is opposite to the one represented by PC3.

PC4: PC4 represents a pattern of variation associated with variables like ProdQual and ProdLine, which have positive correlations with PC4. Higher values of ProdQual and ProdLine contribute to higher values of PC4, indicating a specific pattern of variation captured by PC4.

PC5: PC5 captures a distinct pattern of variation characterized by a negative correlation with Advertising. Higher values of Advertising are associated with lower values of PC5. This indicates that Advertising contributes to a unique pattern of variation represented by PC5.

Understanding the contribution of each variable to the principal components allows us to identify the underlying patterns and factors driving the variation in the data. By analyzing the relationships between the original variables and the PCs, we gain insights into the key factors that shape the data and can use this knowledge for further analysis and decision-making.

Clustering

Problem Statement

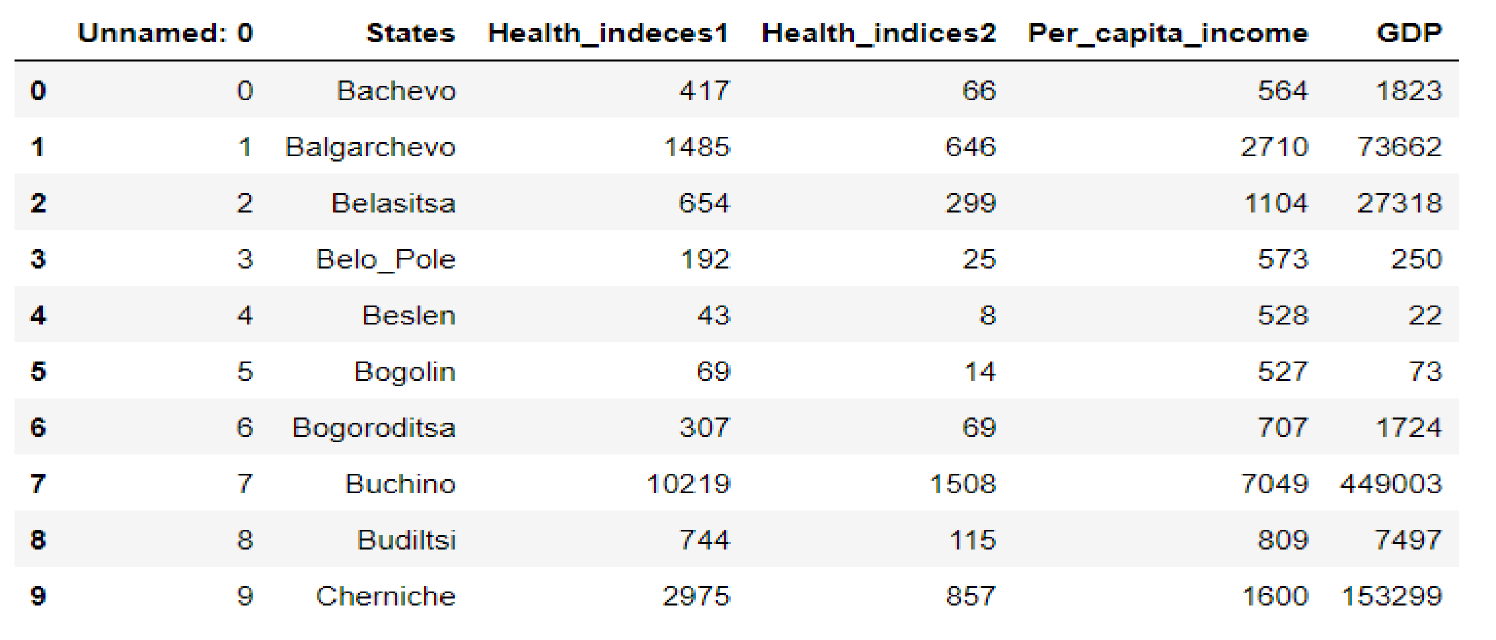
The State_wise_Health_income.csv dataset given is about the Health and economic conditions in different States of a country. The Group States based on how similar their situation is, so as to provide these groups to the government so that appropriate measures can be taken to escalate their Health and Economic conditions.

Data Dictionary

- States: names of States
- 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.
- 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 State.
- 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.
- GDP: GDP provides an economic snapshot of a country/state, used to estimate the size of an economy and its growth rate.

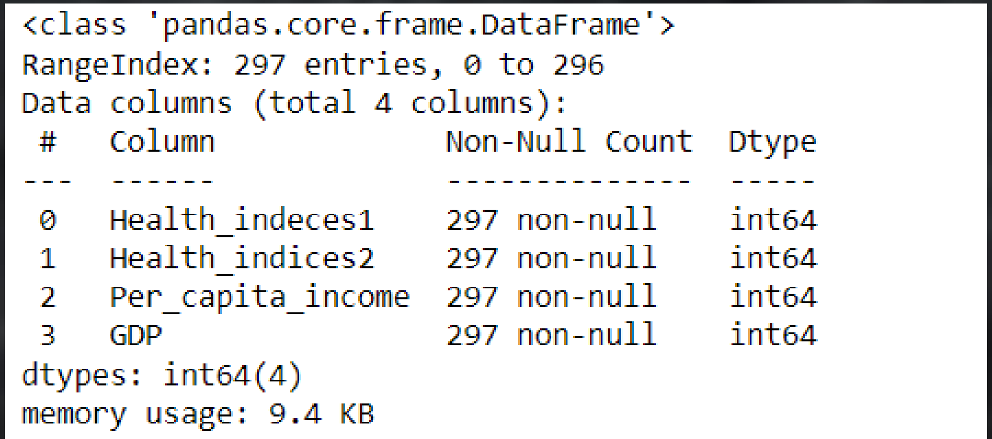
Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, etc)

Data set with the first 10 rows



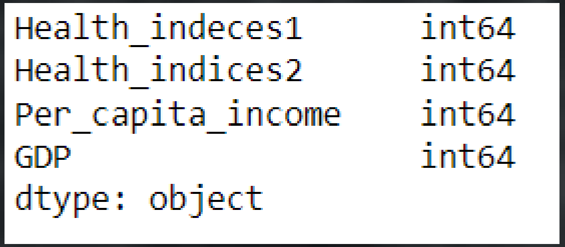
	Unnamed: 0	States	Health_indeces1	Health_indices2	Per_capita_income	GDP
0	0	Bachevo	417	66	564	1823
1	1	Balgarchevo	1485	646	2710	73662
2	2	Belasitsa	654	299	1104	27318
3	3	Belo_Pole	192	25	573	250
4	4	Beslen	43	8	528	22
5	5	Bogolin	69	14	527	73
6	6	Bogoroditsa	307	69	707	1724
7	7	Buchino	10219	1508	7049	449003
8	8	Budiltsi	744	115	809	7497
9	9	Cherniche	2975	857	1600	153299

As we don't need ID column and States column which is the categorical data we can drop these two variables from the data set.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 297 entries, 0 to 296
Data columns (total 4 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Health_indeces1    297 non-null    int64
 1   Health_indices2    297 non-null    int64
 2   Per_capita_income  297 non-null    int64
 3   GDP                297 non-null    int64
dtypes: int64(4)
memory usage: 9.4 KB
```

Data types of the data set

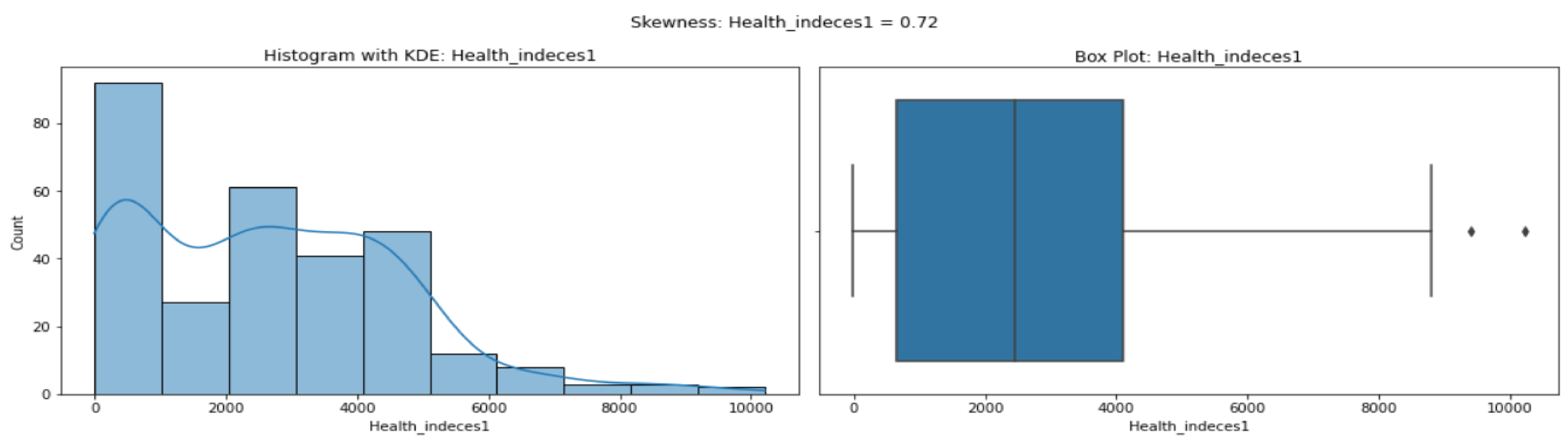


```
Health_indeces1       int64
Health_indices2       int64
Per_capita_income     int64
GDP                   int64
dtype: object
```

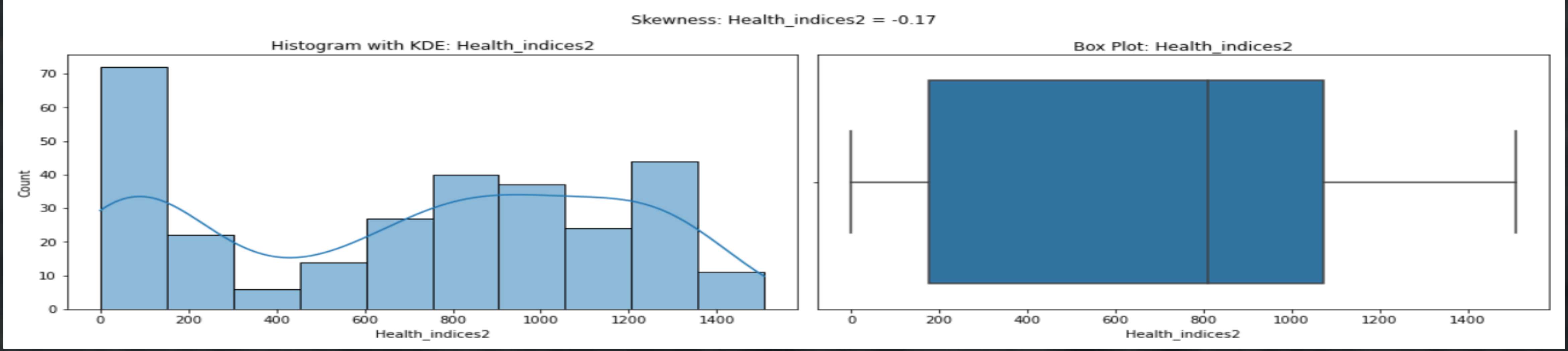
- The data set has 297 rows and 4 columns.
- No Null values were there in the dataset.

Univariate Analysis

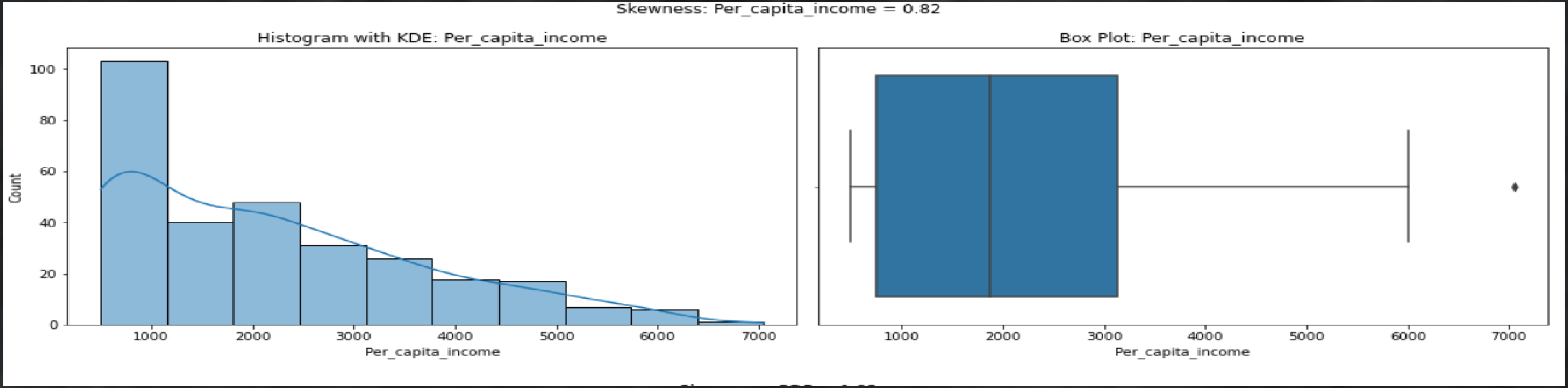
Health_indeces1



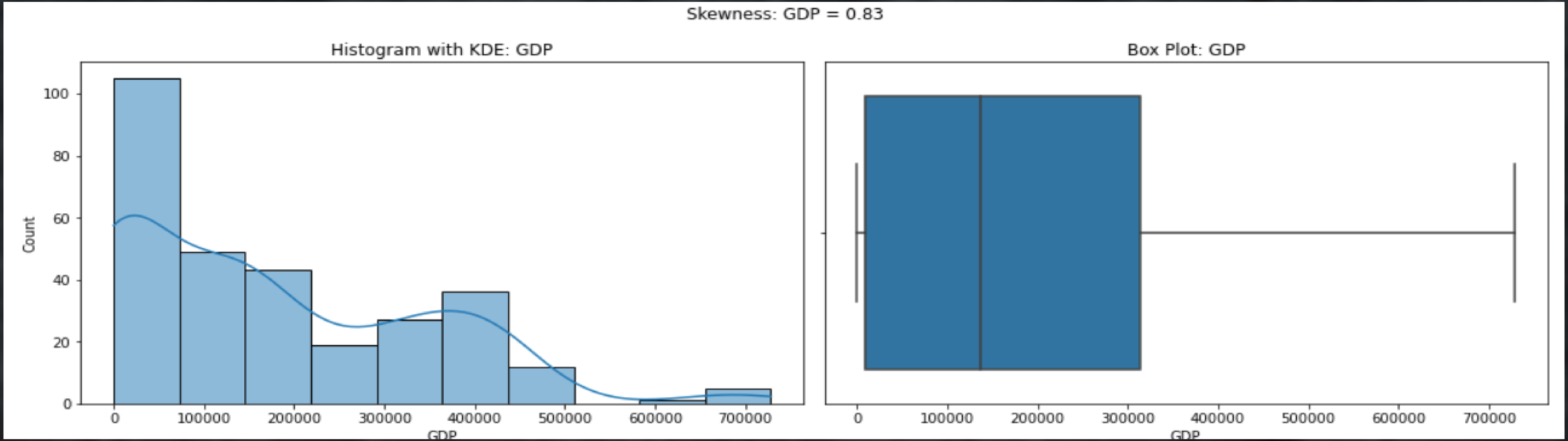
Health_indices2



Per_capita_income

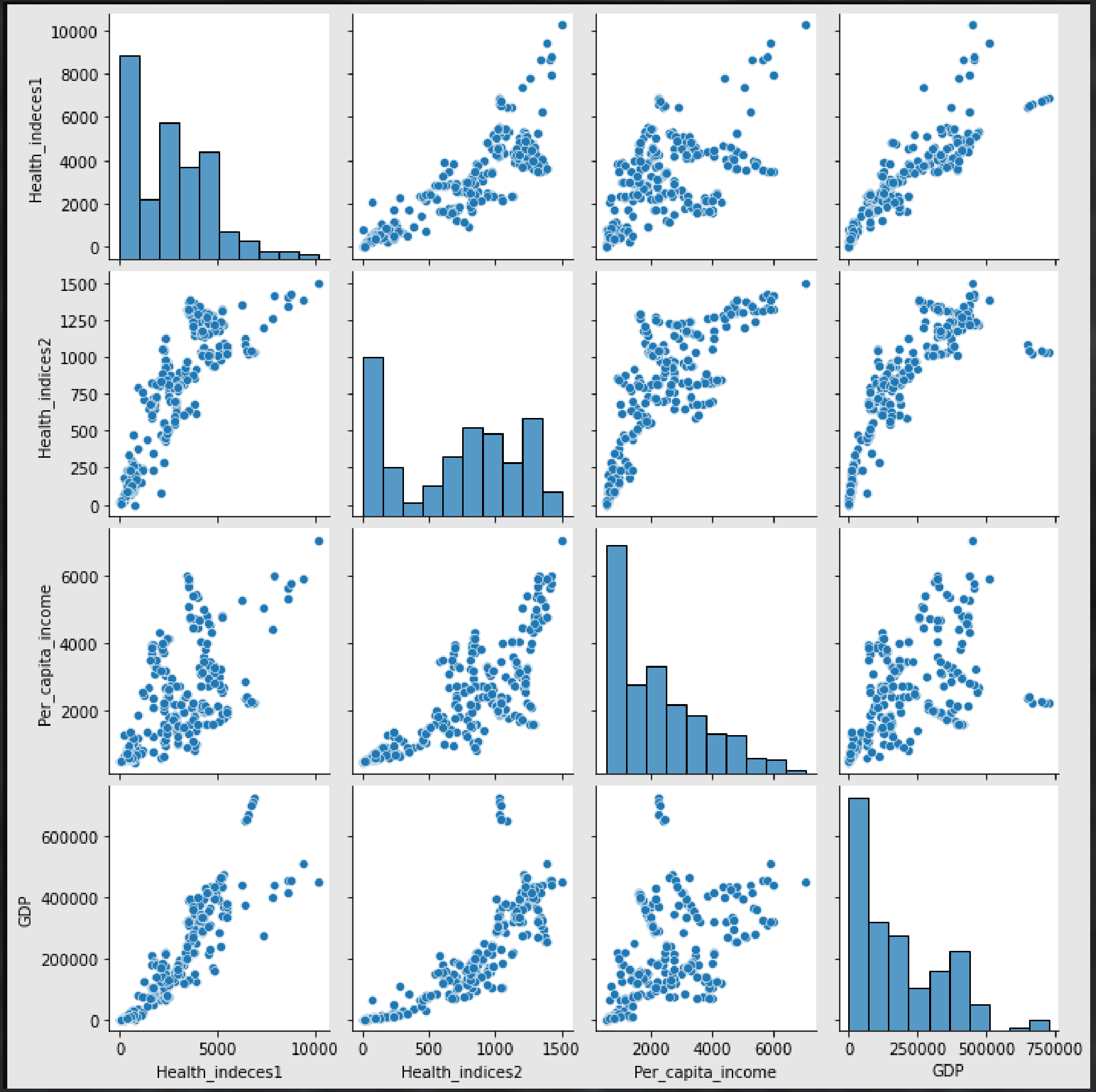


GDP


Skewness: GDP = 0.83
Histogram with KDE: GDP
Count
0
20
40
60
80
100
0
100000
200000
300000
400000
500000
600000
700000
GDP
Box Plot: GDP
0
100000
200000
300000
400000
500000
600000
700000
GDP

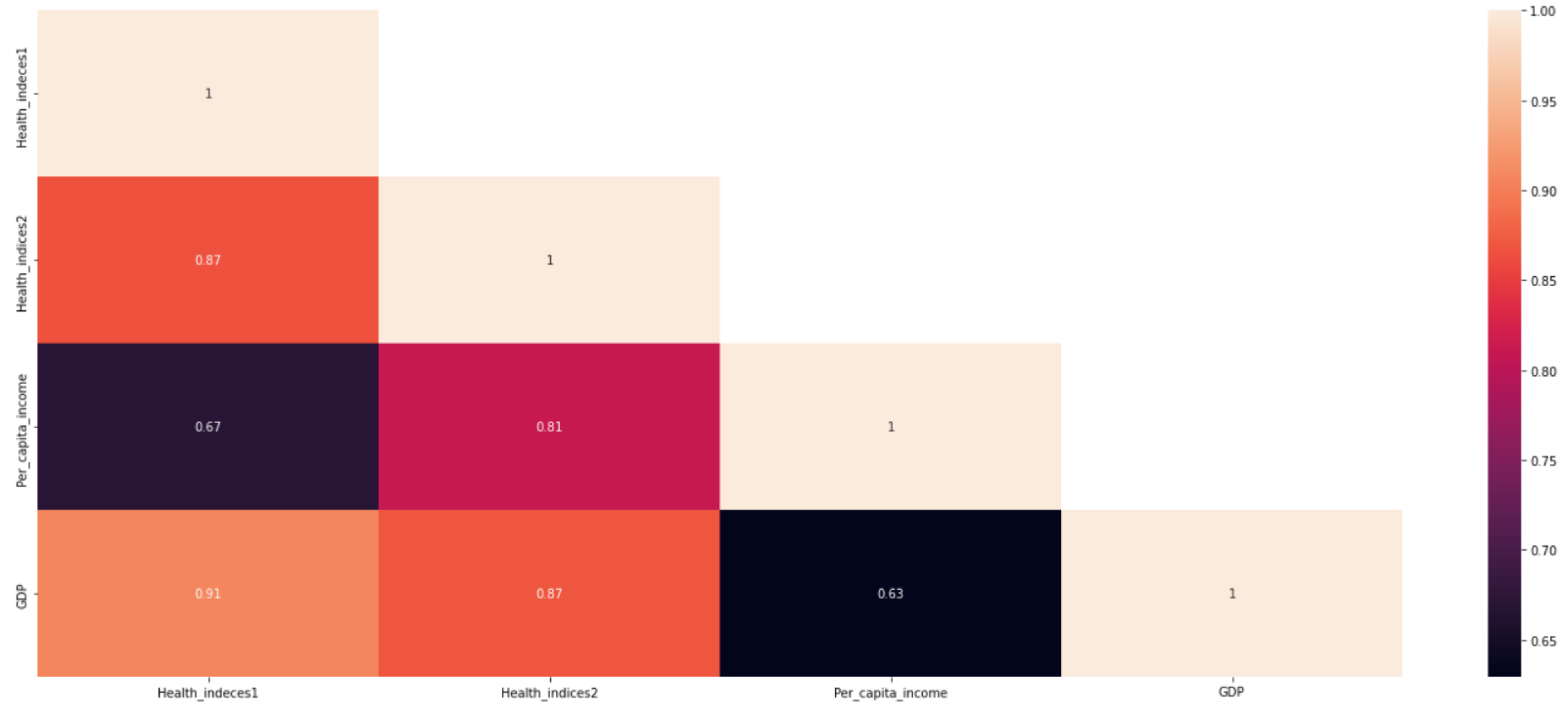
Multivariate Analysis

Pair plot



We can observe that all the variables in the data exhibit correlation.

Heatmap



Above heatmap show that there is a strong co-relation between variables

- Health Indices 1 and GDP have a strong positive correlation of 0.907, indicating that there is a significant relationship between a region's health indices and its GDP. Higher health indices tend to be associated with higher GDP values. This suggests that investing in improving health indicators could have a positive impact on the economic performance of a region.

- Health Indices 1 and Health Indices 2 have a strong positive correlation of 0.866, indicating a high degree of similarity between the two sets of health indices. This suggests that they measure similar aspects of health and can potentially be used interchangeably in certain analyses.

- Health Indices 2 and GDP also show a positive correlation of 0.869, although slightly lower than the correlation between Health Indices 1 and GDP. This suggests that both sets of health indices are moderately related to economic performance.

- Per Capita Income shows a moderate positive correlation of 0.630 with GDP, indicating that regions with higher per capita income tend to have higher GDP values. This suggests that the economic prosperity of a region is influenced by the income levels of its residents.

- There is a moderate positive correlation of 0.669 between Per Capita Income and Health Indices 1. This implies that regions with higher per capita income tend to have better health indices. It suggests that improving income levels can potentially lead to better healthcare outcomes.

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

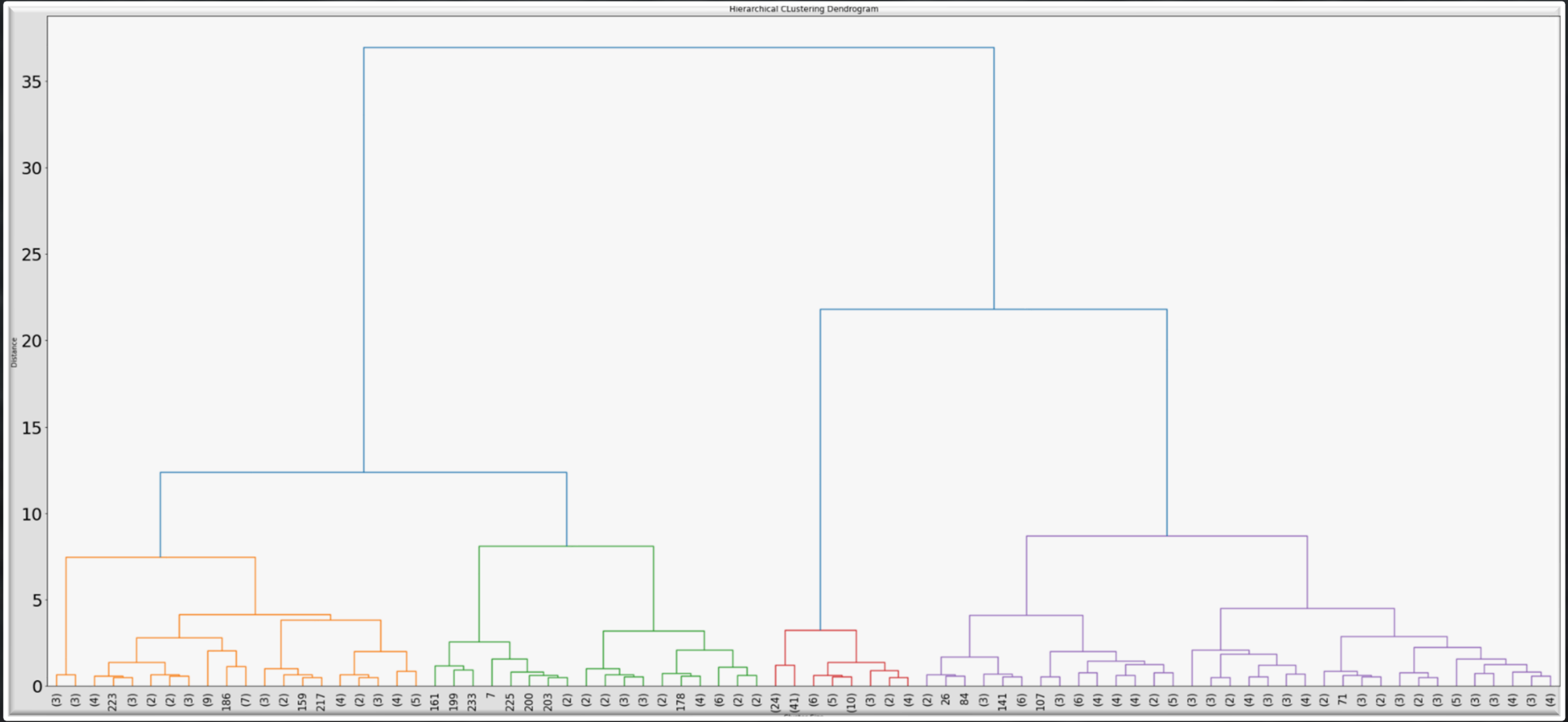
Yes, scaling is necessary for clustering algorithms like K-means. Before applying the K-means algorithm to the data, it is important to perform feature scaling. Clustering techniques, including K-means, rely on calculating distances between data points, usually using Euclidean distance. Scaling the data is beneficial when the attributes have different units of measurement.

In the given dataset, the attributes have different units of measurement, such as health indices, per capita income, and GDP. Scaling the data will create a common scale where the variables have a relative range. One commonly used scaling technique is z-score scaling, which transforms the data to have a mean of 0 and a standard deviation of 1. This scaling method ensures that the data is centered around zero and has a standardized distribution.

Performing feature scaling allows for a fair comparison between the variables and helps the clustering algorithm to effectively group the data based on their similarities.

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

Applying hierarchical clustering to the scaled data using Euclidean distance and 'Ward's' linkage method



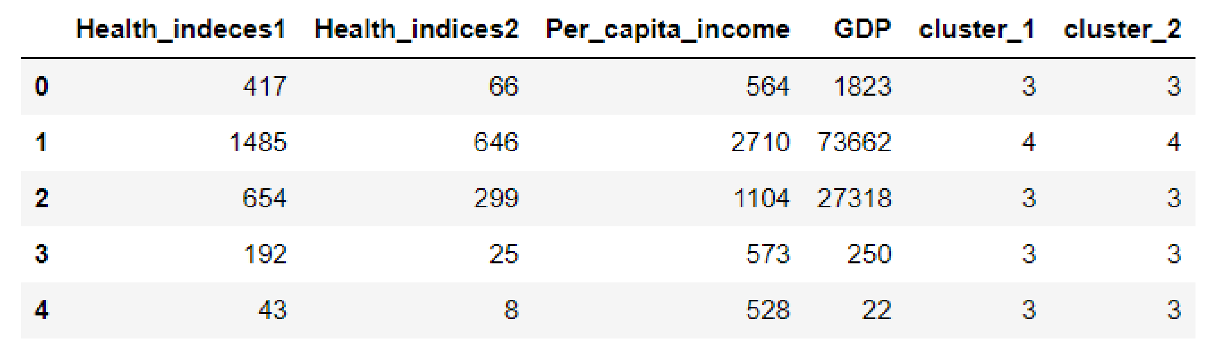
The government aims to gain a deeper understanding of the states beyond simplistic categories of "good" and "not so good." By generating more insights through clustering analysis, we can provide a finer segmentation for businesses. Therefore, let's consider using four clusters and visualize the results to verify if the derived clusters offer the desired level of segmentation detail.

Agglomerative clustering is an approach that follows a "bottom-up" strategy.

In this method, each observation begins as its own cluster, and clusters are progressively merged as we move up the hierarchical structure. Initially, every data point is treated as a separate cluster. During each iteration, clusters that exhibit similarity are merged with other clusters until a specified number, K, of clusters is formed.

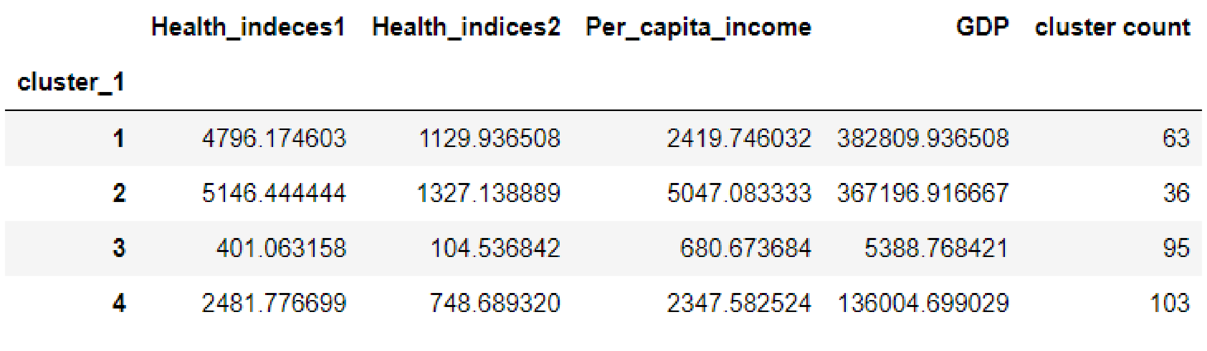
The primary advantage of agglomerative clustering is that it eliminates the need to predefine the number of clusters. Instead, the algorithm automatically determines the optimal number of clusters based on the similarity between data points.

To gain insights into the clusters obtained from hierarchical clustering, we can profile them based on the means of each variable used in the clustering process. By examining the average values of the variables within each cluster, we can understand the characteristic features or trends associated with each cluster. This profiling allows us to identify the distinguishing factors and patterns that contribute to the formation of these clusters.



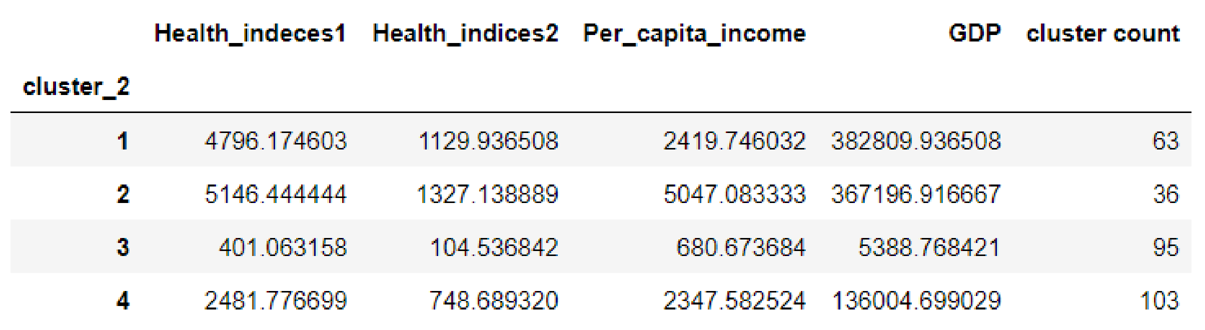
	Health_indeces1	Health_indices2	Per_capita_income	GDP	cluster_1	cluster_2
0	417	66	564	1823	3	3
1	1485	646	2710	73662	4	4
2	654	299	1104	27318	3	3
3	192	25	573	250	3	3
4	43	8	528	22	3	3

This allows you to inspect the updated data frame and verify the assigned cluster values in the 'cluster_1' and 'cluster_2' columns.



	Health_indeces1	Health_indices2	Per_capita_income	GDP	cluster count
cluster_1					
1	4796.174603	1129.936508	2419.746032	382809.936508	63
2	5146.444444	1327.138889	5047.083333	367196.916667	36
3	401.063158	104.536842	680.673684	5388.768421	95
4	2481.776699	748.689320	2347.582524	136004.699029	103

The data frame on the left contains the mean values of each variable within each cluster obtained from 'df1', along with the corresponding cluster count. It provides a summary of the average values and the frequency count for each cluster in the 'cluster_1' column.



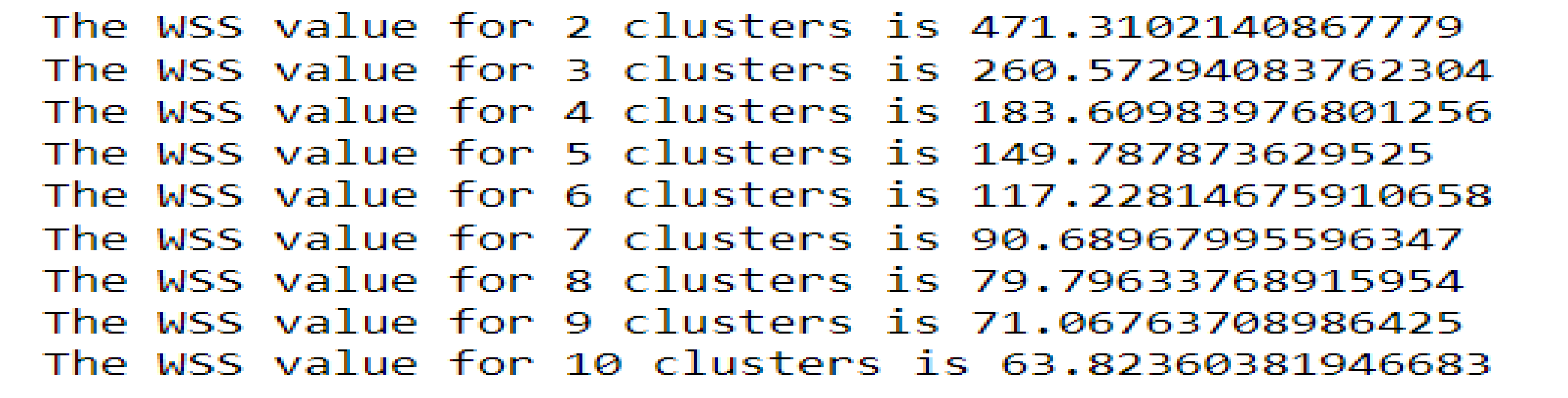
	Health_indeces1	Health_indices2	Per_capita_income	GDP	cluster count
cluster_2					
1	4796.174603	1129.936508	2419.746032	382809.936508	63
2	5146.444444	1327.138889	5047.083333	367196.916667	36
3	401.063158	104.536842	680.673684	5388.768421	95
4	2481.776699	748.689320	2347.582524	136004.699029	103

The data frame on the left contains the mean values of variables for each cluster ('cluster_2') and the corresponding cluster counts. This allows for an analysis of the average characteristics of each cluster and the distribution of data points among the clusters.

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

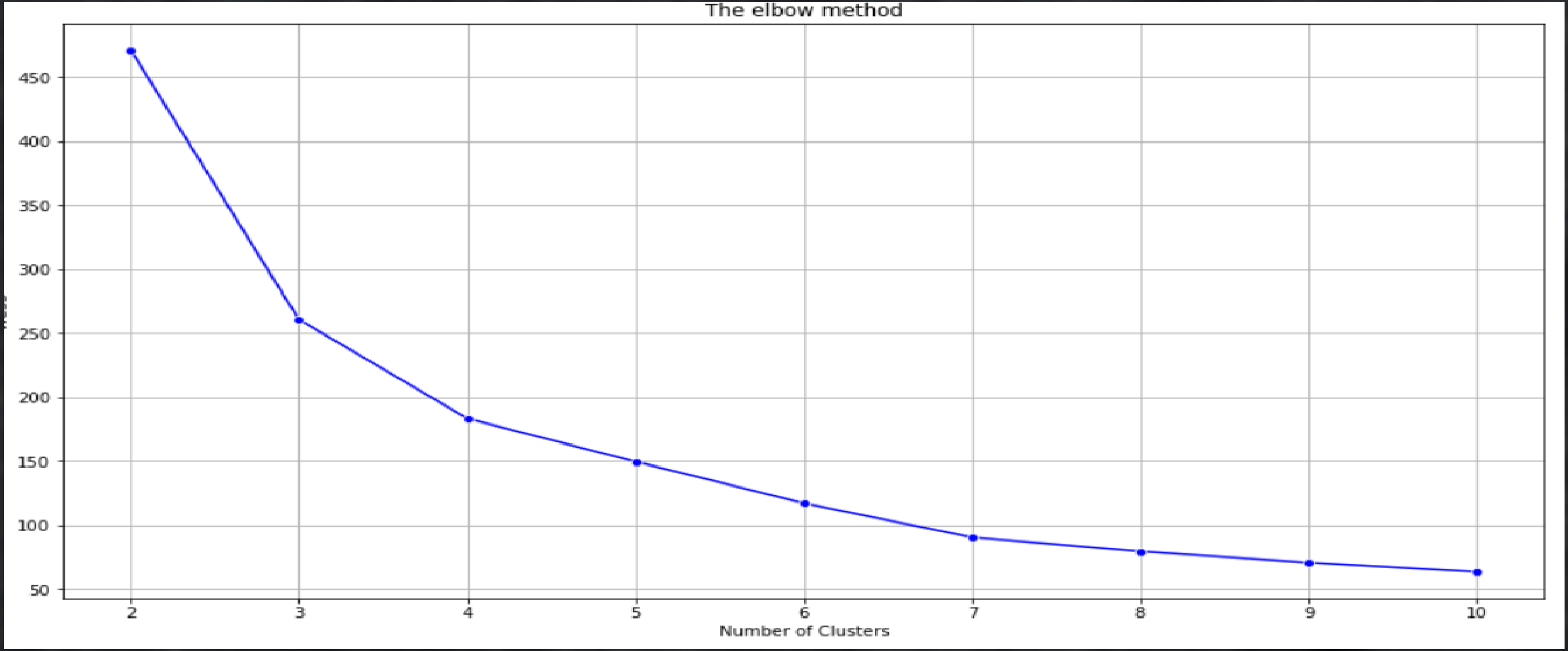
Within-Cluster Sum of Squares (WSS) for different numbers of clusters in a clustering algorithm. The WSS is a measure of the compactness or cohesion of data points within each cluster.

A lower WSS value indicates that the data points within each cluster are closer to each other, suggesting a more compact and well-separated clustering. Conversely, a higher WSS value indicates that the data points within clusters are more scattered and less compact.



```
The WSS value for 2 clusters is 471.3102140867779
The WSS value for 3 clusters is 260.57294083762304
The WSS value for 4 clusters is 183.60983976801256
The WSS value for 5 clusters is 149.787873629525
The WSS value for 6 clusters is 117.22814675910658
The WSS value for 7 clusters is 90.68967995596347
The WSS value for 8 clusters is 79.79633768915954
The WSS value for 9 clusters is 71.06763708986425
The WSS value for 10 clusters is 63.82360381946683
```

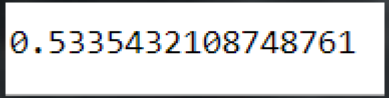
In the given values, we can observe that the WSS value decreases as the number of clusters increases from 2 to 10. This suggests that the data points become more compact and well-separated as more clusters are considered. The decreasing trend indicates that the clustering algorithm is finding better groupings of data points as the number of clusters increases.



The K-means clustering technique was applied to the dataset, and the elbow curve method was used to determine the optimal number of clusters. Based on the analysis, it was determined that either 3 or 4 clusters would be the most suitable number of clusters for this dataset

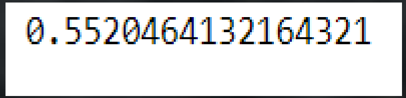
The Silhouette score

Silhouette score for 3 clusters



```
0.5335432108748761
```

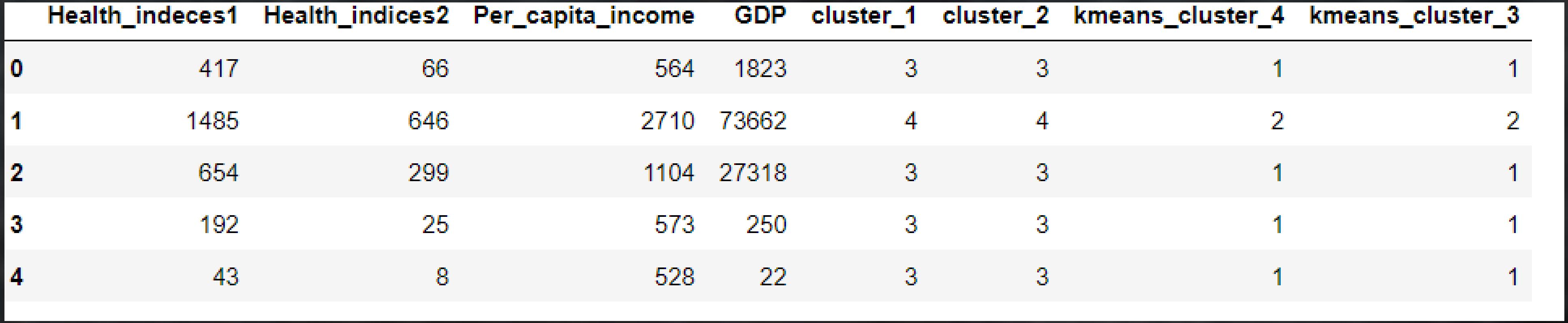
Silhouette score for 4 clusters



```
0.5520464132164321
```

As per the silhouette score Label 4 is slightly better than label 3

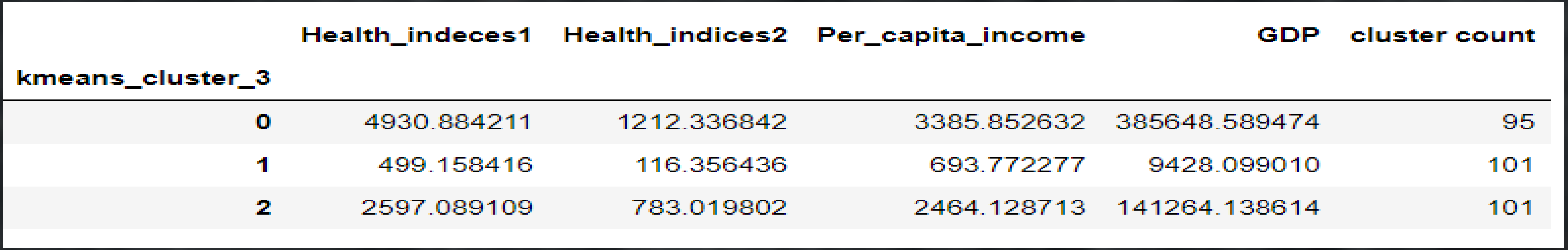
Describe cluster profiles for the clusters defined. Recommend different priority-based actions that need to be taken for different clusters on the bases of their vulnerability situations according to their Economic and Health Conditions



	Health_indeces1	Health_indices2	Per_capita_income	GDP	cluster_1	cluster_2	kmeans_cluster_4	kmeans_cluster_3
0	417	66	564	1823	3	3	1	1
1	1485	646	2710	73662	4	4	2	2
2	654	299	1104	27318	3	3	1	1
3	192	25	573	250	3	3	1	1
4	43	8	528	22	3	3	1	1

kmeans cluster 4 and kmeans cluster 3 were added to the data frame

Cluster Profiles for k-means with 3 Clusters:



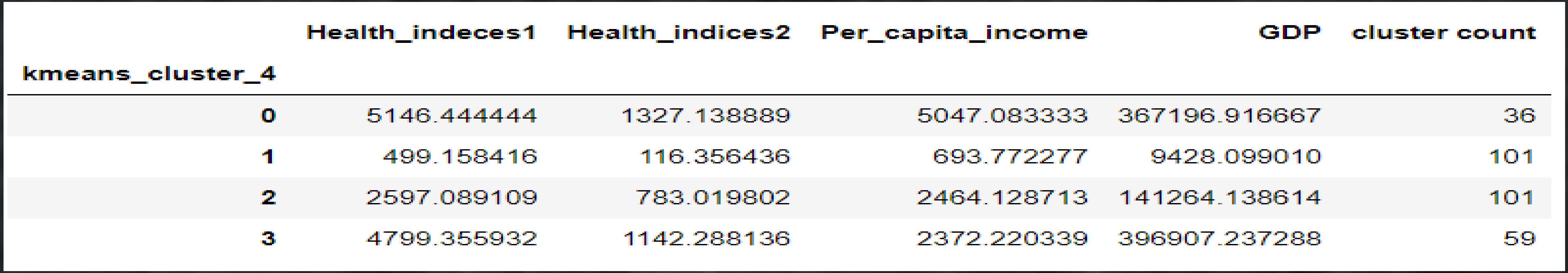
	Health_indeces1	Health_indices2	Per_capita_income	GDP	cluster count
kmeans_cluster_3					
0	4930.884211	1212.336842	3385.852632	385648.589474	95
1	499.158416	116.356436	693.772277	9428.099010	101
2	2597.089109	783.019802	2464.128713	141264.138614	101

- Cluster 0: This cluster exhibits high health indices, per capita income, and GDP. Priority actions for this cluster could focus on sustaining and improving the already favorable economic and health conditions. Investments in infrastructure, education, and healthcare systems can help maintain the high standards in these areas.

- Cluster 1: This cluster represents regions with lower health indices, per capita income, and GDP. Priority actions should be directed towards improving healthcare accessibility, education, and employment opportunities to uplift the economic and health conditions. Initiatives targeting poverty reduction, skill development, and healthcare infrastructure enhancement can be beneficial.

- Cluster 2: This cluster represents regions with moderate health indices, per capita income, and GDP. Priority actions could focus on maintaining and enhancing the current economic and health conditions. Investments in healthcare facilities, skill development programs, and business diversification initiatives can further improve the economic and health outcomes.

Cluster Profiles for k-means with 4 Clusters:



	Health_indeces1	Health_indices2	Per_capita_income	GDP	cluster count
kmeans_cluster_4					
0	5146.444444	1327.138889	5047.083333	367196.916667	36
1	499.158416	116.356436	693.772277	9428.099010	101
2	2597.089109	783.019802	2464.128713	141264.138614	101
3	4799.355932	1142.288136	2372.220339	396907.237288	59

- ➢ Cluster 0: This cluster represents regions with high health indices, per capita income, and GDP. Priority actions should be focused on maintaining and further enhancing the already favorable economic and health conditions. Continued investments in infrastructure, education, healthcare, and innovation can contribute to sustained growth and well-being.

- ➢ Cluster 1: This cluster includes regions with lower health indices, per capita income, and GDP. Priority actions should be directed towards improving healthcare facilities, education, and employment opportunities to uplift the economic and health conditions. Emphasizing poverty reduction, vocational training, and social welfare programs can help improve the overall well-being of this cluster.

- ➢ Cluster 2: This cluster represents regions with moderate health indices, per capita income, and GDP. Priority actions should concentrate on maintaining and enhancing the current economic and health conditions. Investments in healthcare infrastructure, skill development initiatives, and diversification of industries can contribute to sustained growth and improved well-being.

- ➢ Cluster 3: This cluster exhibits high health indices, moderate per capita income, and high GDP. Priority actions can focus on leveraging the existing economic and health advantages to drive further growth and development. Encouraging entrepreneurship, promoting innovation, and investing in research and development can contribute to sustained economic progress and improved health outcomes.

Thank you