SMDM GRADED PROJECT

A PROJECT REPORT

Submitted by

BALAJI S (PGP-DSBA-JUNE 2023 TO JUNE2024)

Part 1: PCA:

Problem Statement: The [‘Hair Salon.csv’] dataset contains various variables used for the context of Market Segmentation. This case study is based on various parameters of a salon chain of hair products. 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 drop this variable before doing a Principal Component Analysis.**

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

Reading the data and performing basic checks like checking head, info, summary, nulls, duplicates, etc.

	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

HEAD

	ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
95	96	8.6	4.8	5.6	5.3	2.3	6.0	5.7	6.7	5.8	4.9	3.6	7.3
96	97	7.4	3.4	2.6	5.0	4.1	4.4	4.8	7.2	4.5	4.2	3.7	6.3
97	98	8.7	3.2	3.3	3.2	3.1	6.1	2.9	5.6	5.0	3.1	2.5	5.4
98	99	7.8	4.9	5.8	5.3	5.2	5.3	7.1	7.9	6.0	4.3	3.9	6.4
99	100	7.9	3.0	4.4	5.1	5.9	4.2	4.8	9.7	5.7	3.4	3.5	6.4

TAIL

```
<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
```

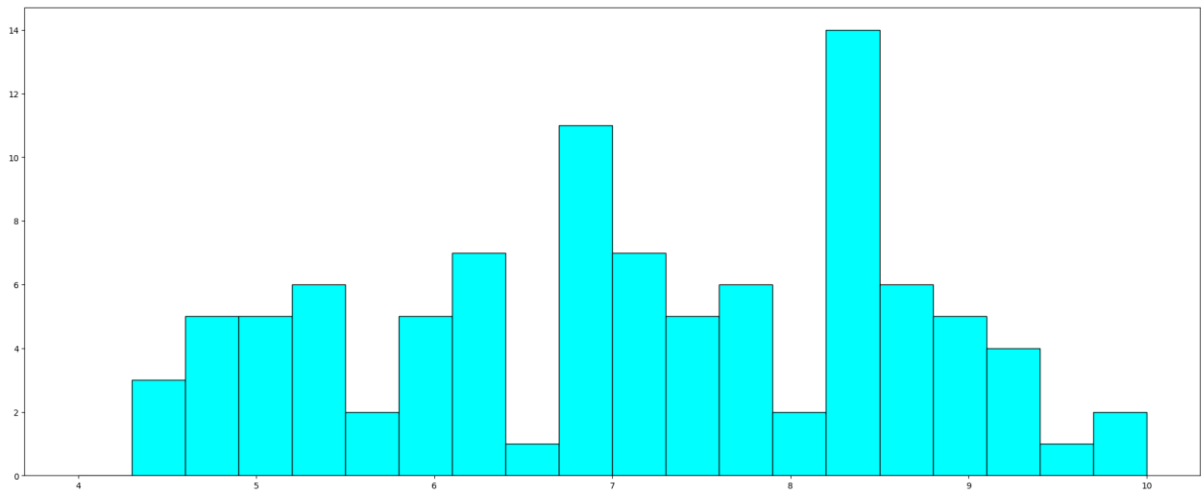
INFORMATION ABOUT THE DATASET

CHECKING THE DUPLICATE AND NULL VALUES

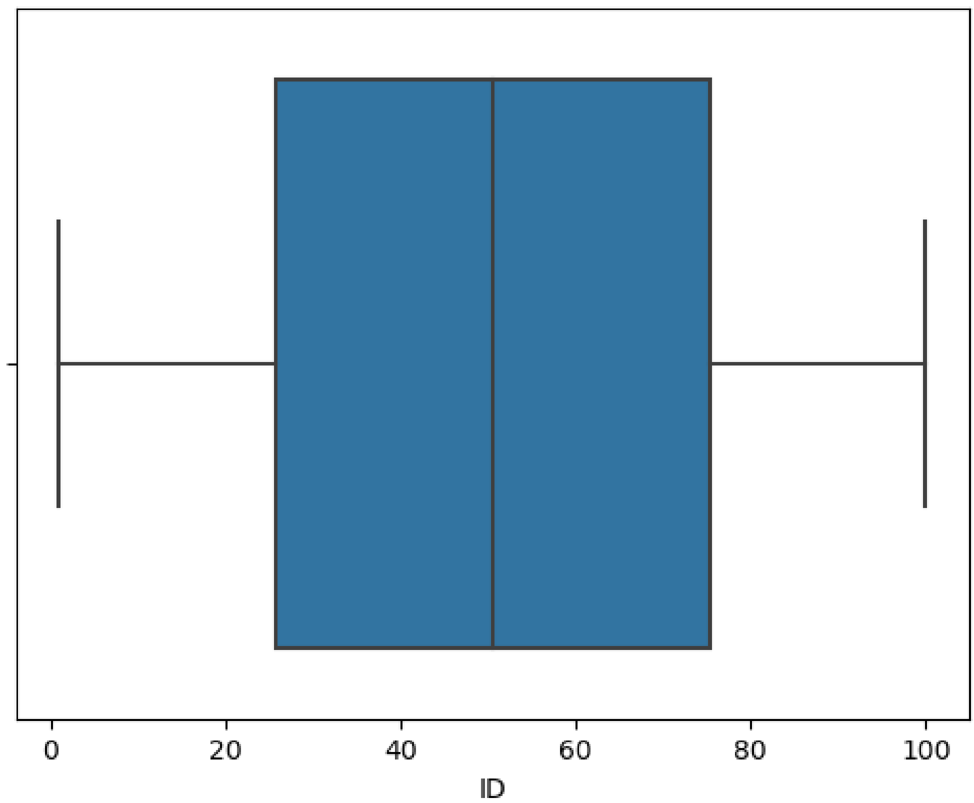
It shows that there is no null and duplicate values in the dataset.

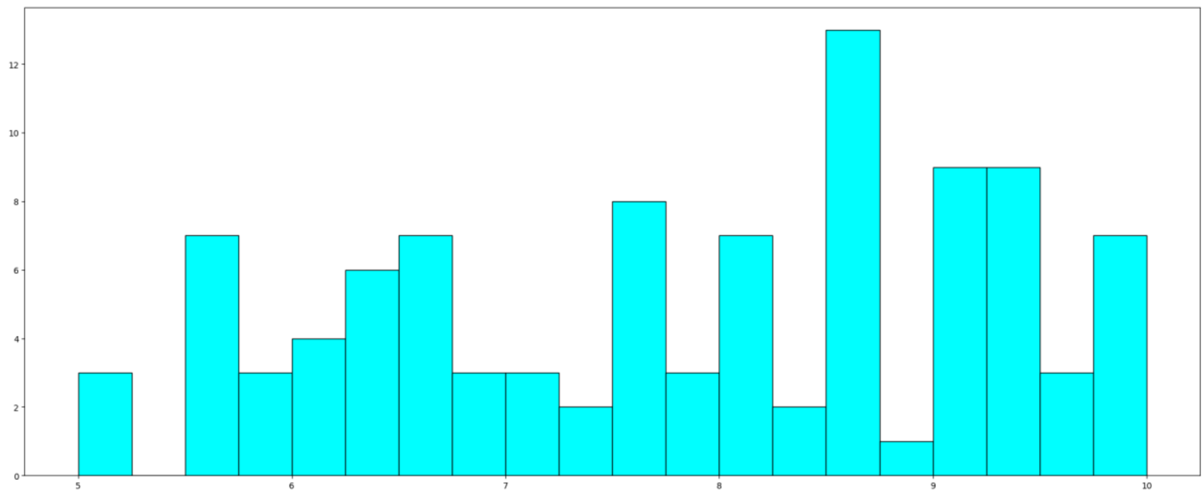
PERFORMING EDA FOR THE GIVEN DATASET:

UNIVAIRENT ANALYSIS:

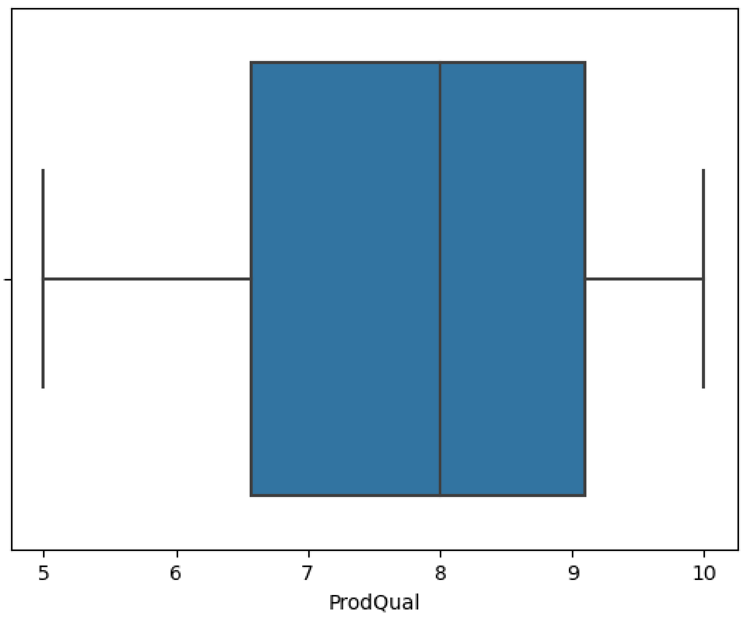


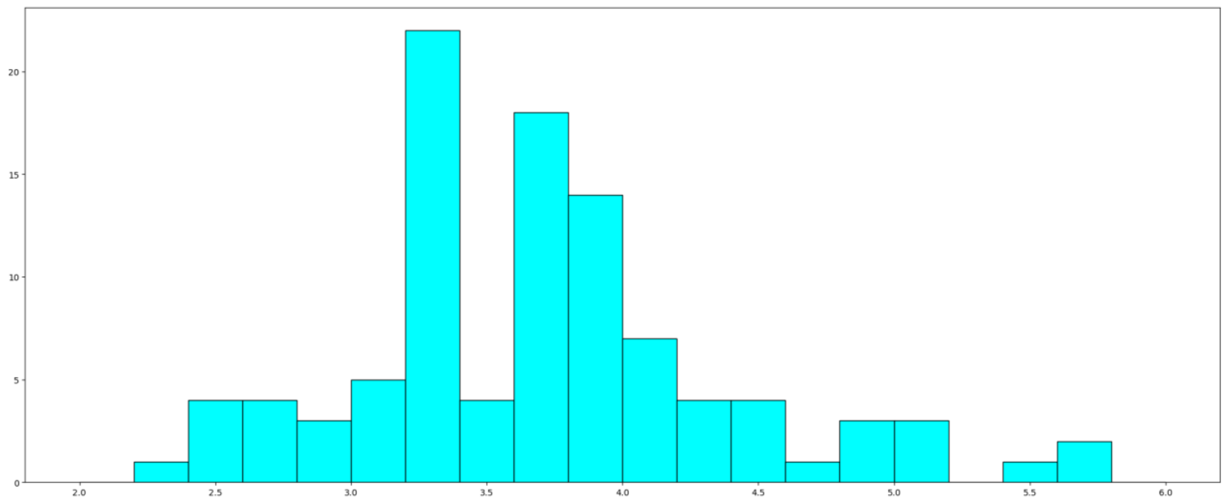

```
Description of ID
-----------------------------------------------
count    100.000000
mean      50.500000
std       29.011492
min        1.000000
25%       25.750000
50%       50.500000
75%       75.250000
max      100.000000
Name: ID, dtype: float64 Distribution of ID
```



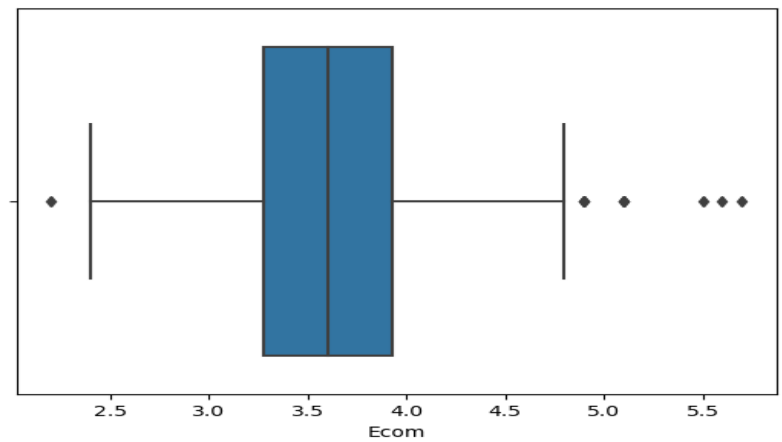


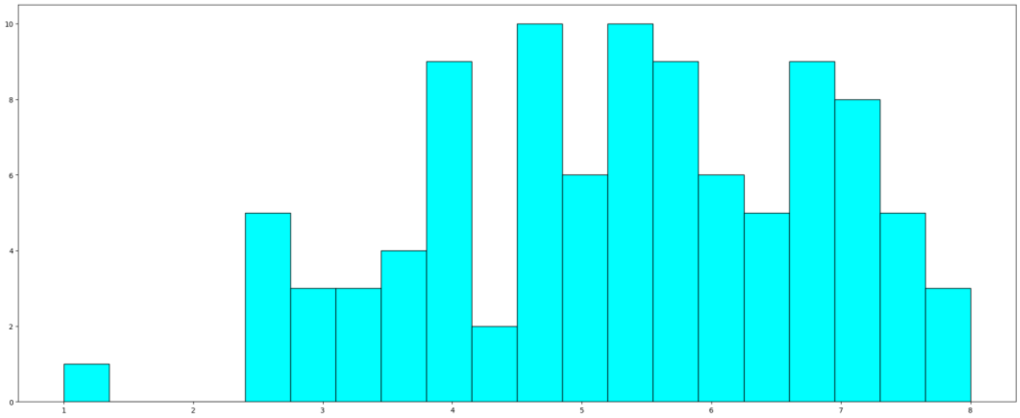

```
Description of ProdQual
----------------------------------------------------------
count    100.000000
mean       7.810000
std        1.396279
min        5.000000
25%        6.575000
50%        8.000000
75%        9.100000
max       10.000000
Name: ProdQual, dtype: float64 Distribution of ProdQual
```



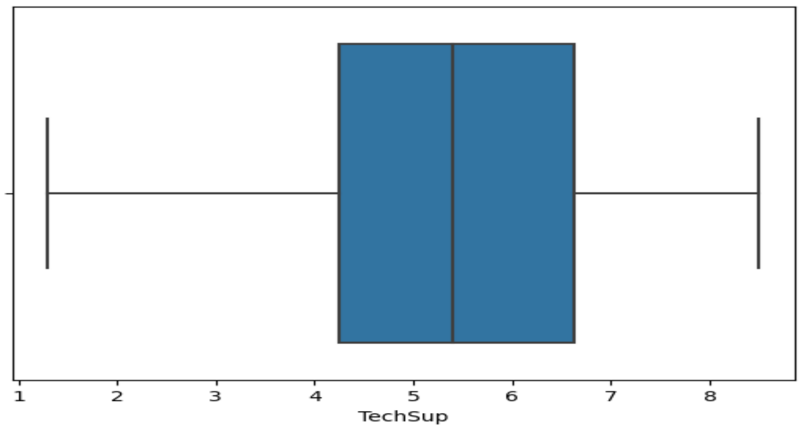


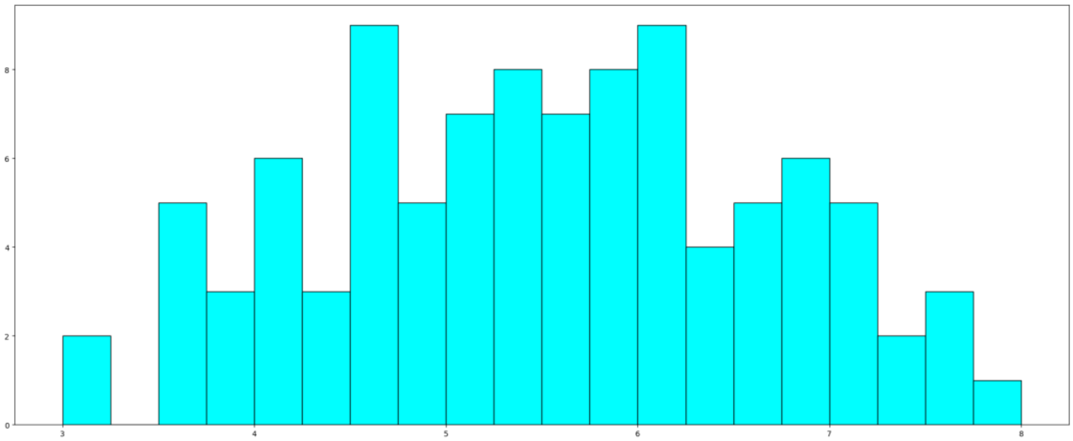

```
Description of Ecom
---------------------------------------------------
count    100.000000
mean       3.672000
std        0.700516
min        2.200000
25%        3.275000
50%        3.600000
75%        3.925000
max        5.700000
Name: Ecom, dtype: float64 Distribution of Ecom
```



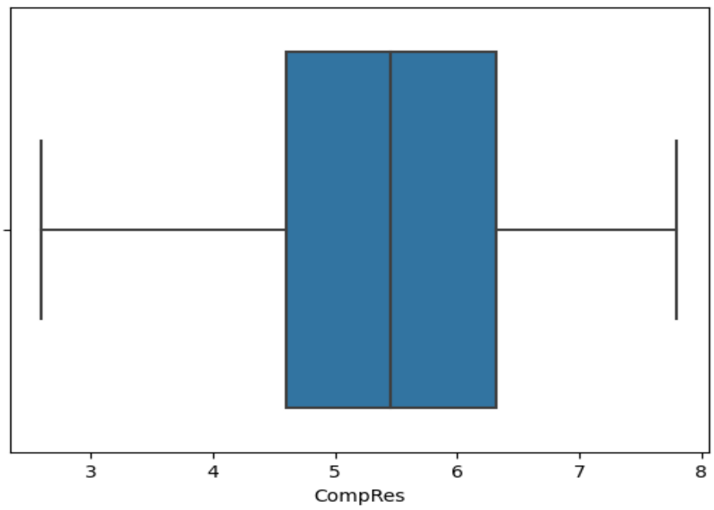


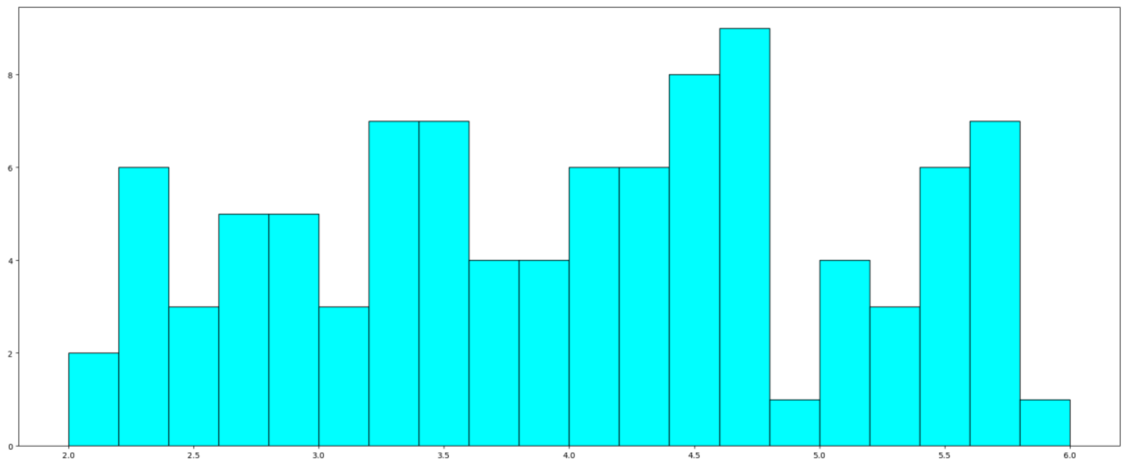

```
Description of TechSup
--------------------------------------------------------
count    100.000000
mean       5.365000
std        1.530457
min        1.300000
25%        4.250000
50%        5.400000
75%        6.625000
max        8.500000
Name: TechSup, dtype: float64 Distribution of TechSup
--------------------------------------------------------
```



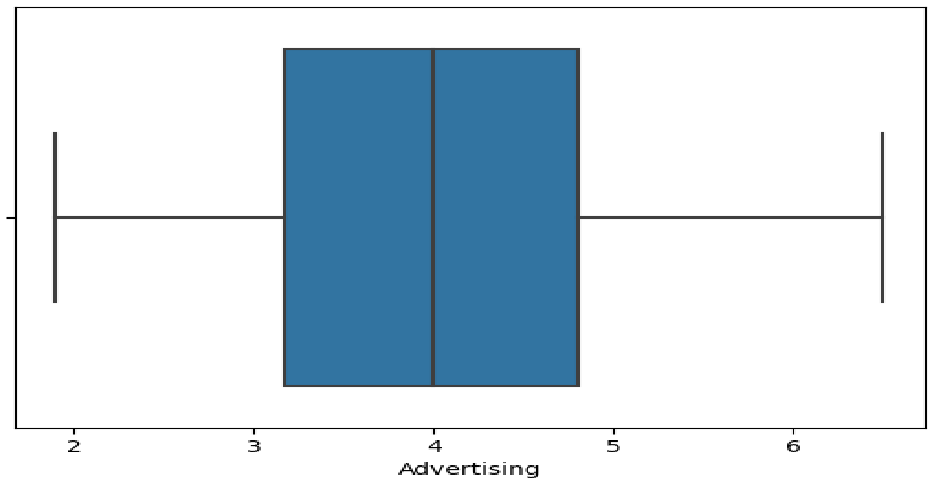


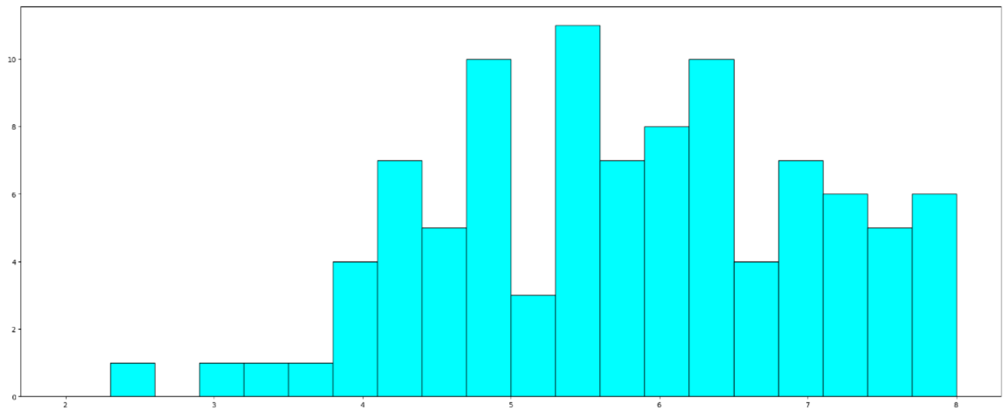

```
Description of CompRes
----------------------------------------------------------
count    100.000000
mean       5.442000
std        1.208403
min        2.600000
25%        4.600000
50%        5.450000
75%        6.325000
max        7.800000
Name: CompRes, dtype: float64 Distribution of CompRes
```



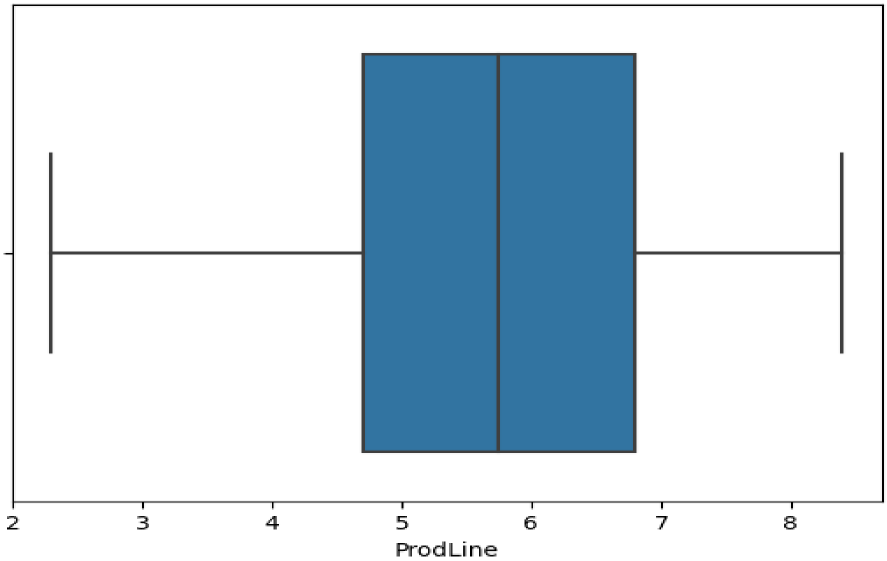


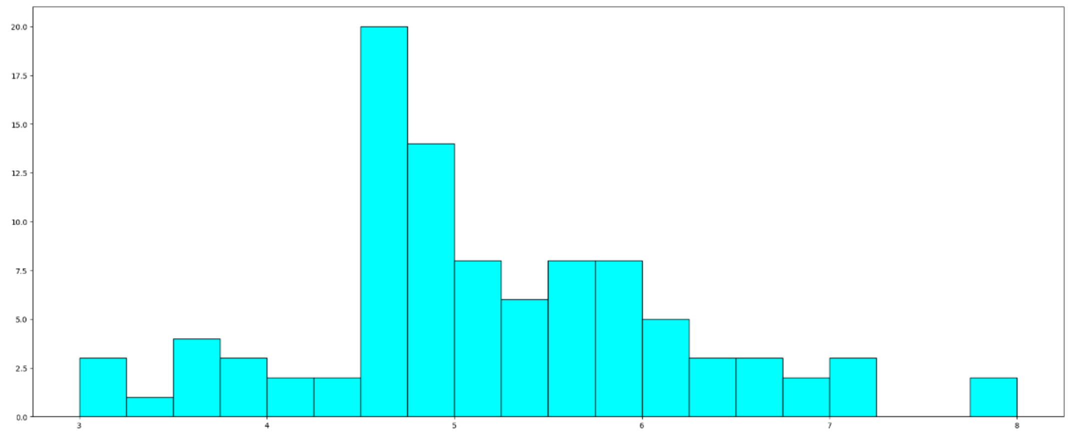

```
Description of Advertising
-----------------------------------------------------------------
count    100.000000
mean       4.010000
std        1.126943
min        1.900000
25%        3.175000
50%        4.000000
75%        4.800000
max        6.500000
Name: Advertising, dtype: float64 Distribution of Advertising
```



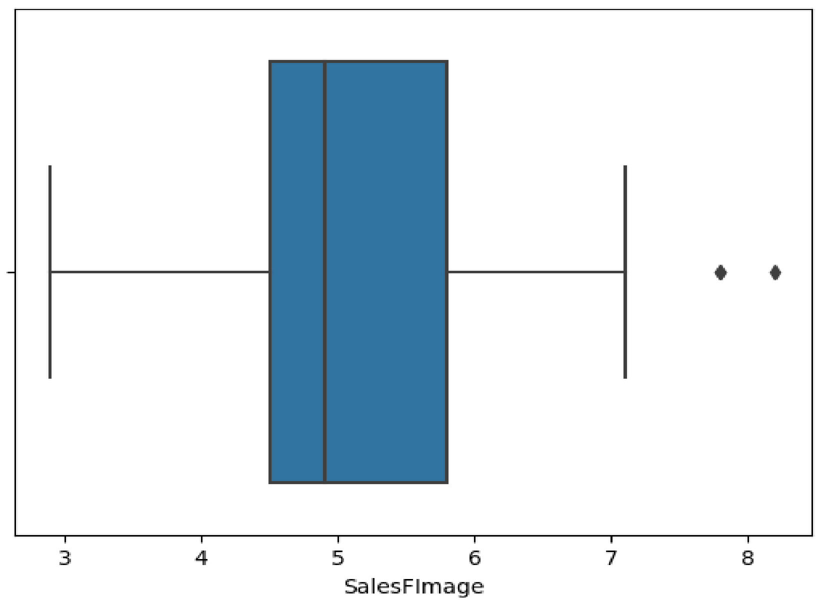


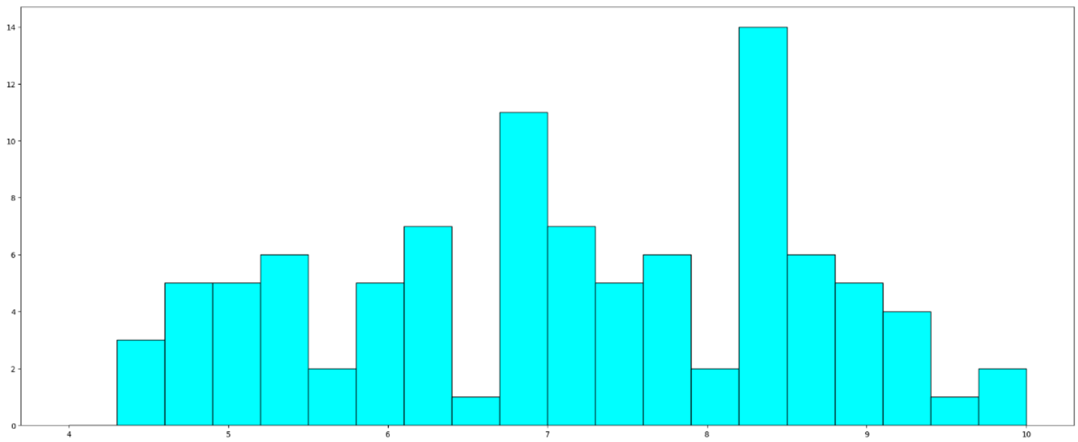
```
Description of ProdLine
------------------------------------------------------------
count    100.000000
mean       5.805000
std        1.315285
min        2.300000
25%        4.700000
50%        5.750000
75%        6.800000
max        8.400000
Name: ProdLine, dtype: float64 Distribution of ProdLine
------------------------------------------------------------
```



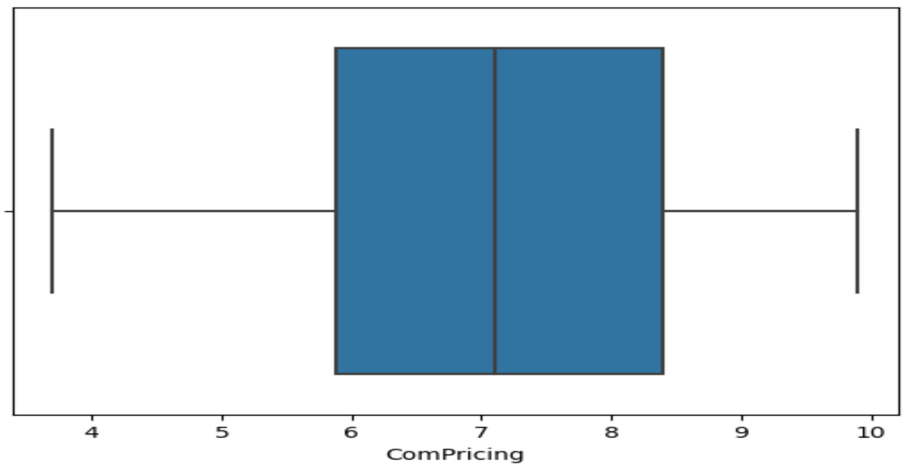


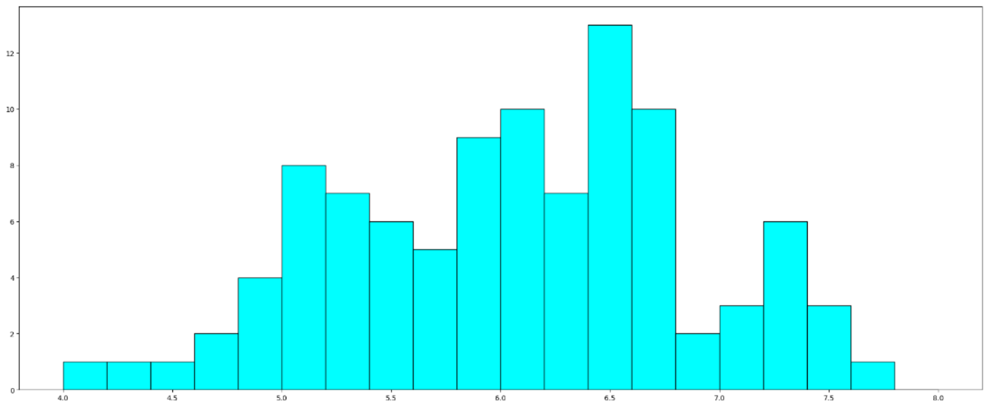

```
Description of SalesFImage
-------------------------------------------------------------------
count    100.00000
mean       5.12300
std        1.07232
min        2.90000
25%        4.50000
50%        4.90000
75%        5.80000
max        8.20000
Name: SalesFImage, dtype: float64 Distribution of SalesFImage
-------------------------------------------------------------------
```



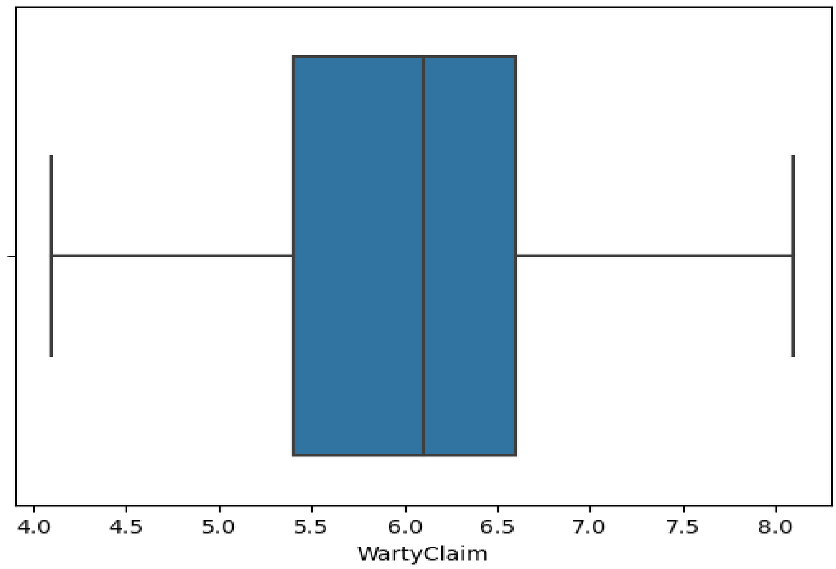


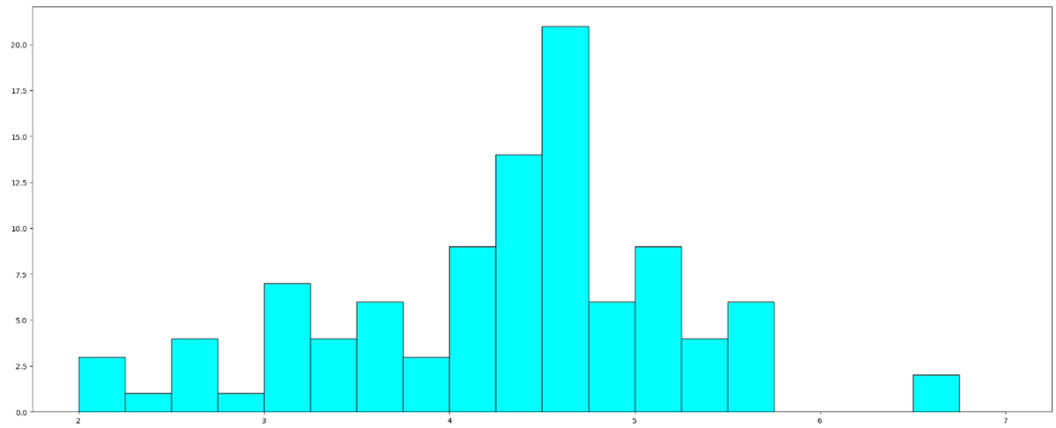

```
Description of ComPricing
----------------------------------------------------------------
count    100.000000
mean       6.974000
std        1.545055
min        3.700000
25%        5.875000
50%        7.100000
75%        8.400000
max        9.900000
Name: ComPricing, dtype: float64 Distribution of ComPricing
```



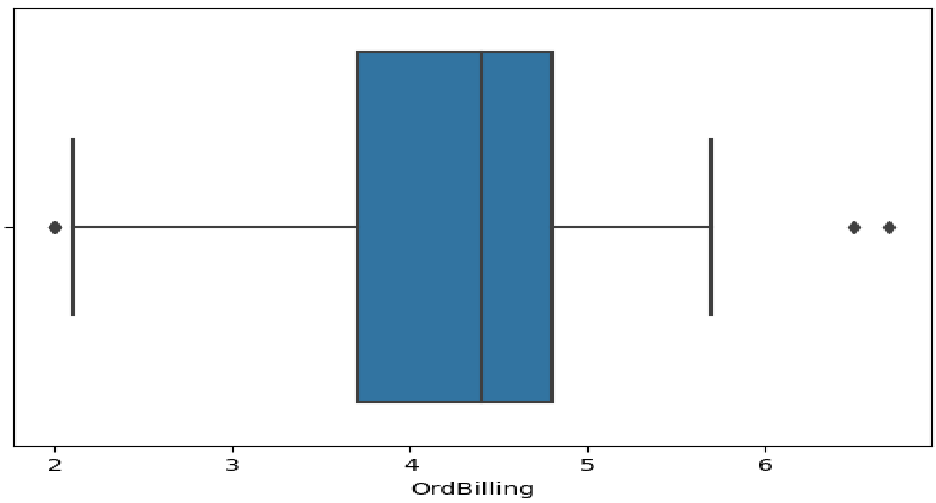


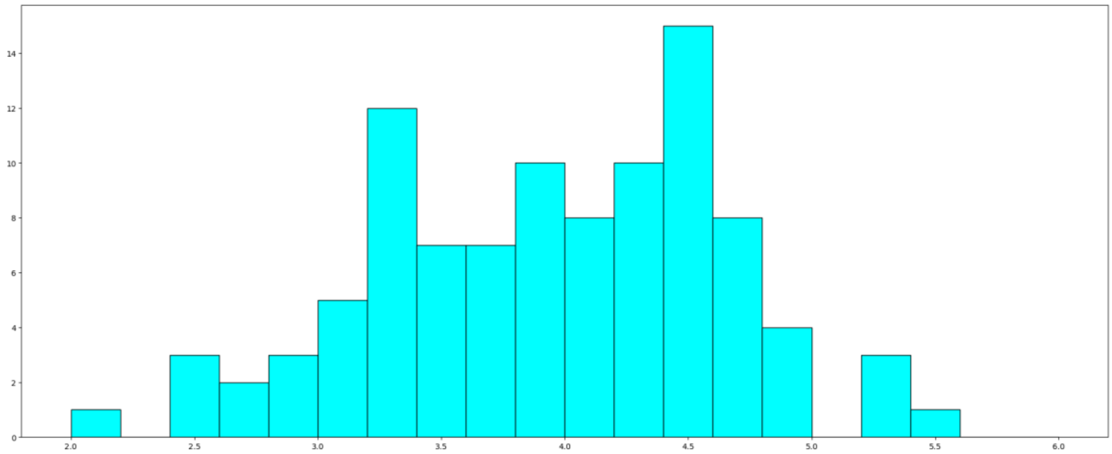

```
Description of WartyClaim
----------------------------------------------------------------------
count    100.000000
mean       6.043000
std        0.819738
min        4.100000
25%        5.400000
50%        6.100000
75%        6.600000
max        8.100000
Name: WartyClaim, dtype: float64 Distribution of WartyClaim
```



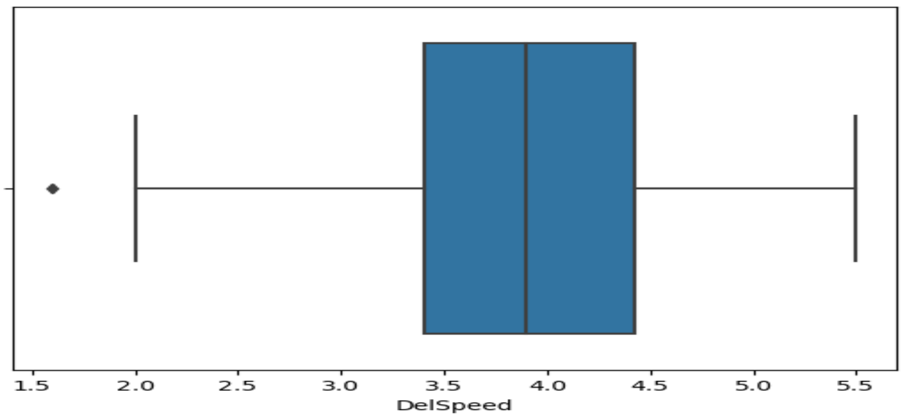


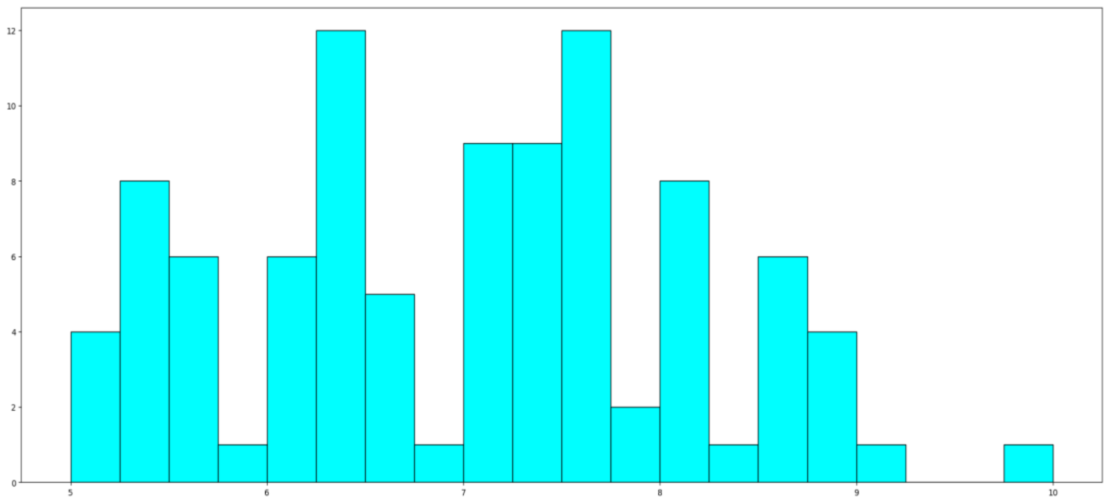

```
Description of OrdBilling
----------------------------------------------------------------------
count    100.00000
mean       4.27800
std        0.92884
min        2.00000
25%        3.70000
50%        4.40000
75%        4.80000
max        6.70000
Name: OrdBilling, dtype: float64 Distribution of OrdBilling
----------------------------------------------------------------------
```



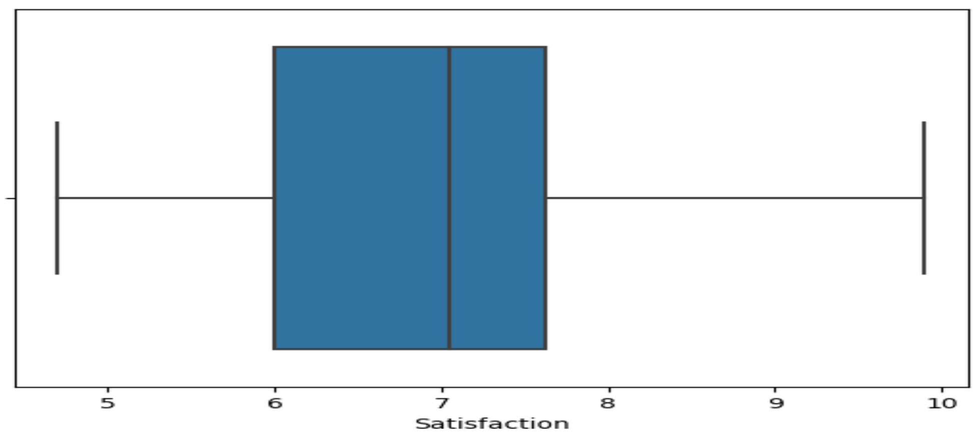



```
Description of DelSpeed
----------------------------------------------------------------
count    100.000000
mean       3.886000
std        0.734437
min        1.600000
25%        3.400000
50%        3.900000
75%        4.425000
max        5.500000
Name: DelSpeed, dtype: float64 Distribution of DelSpeed
```

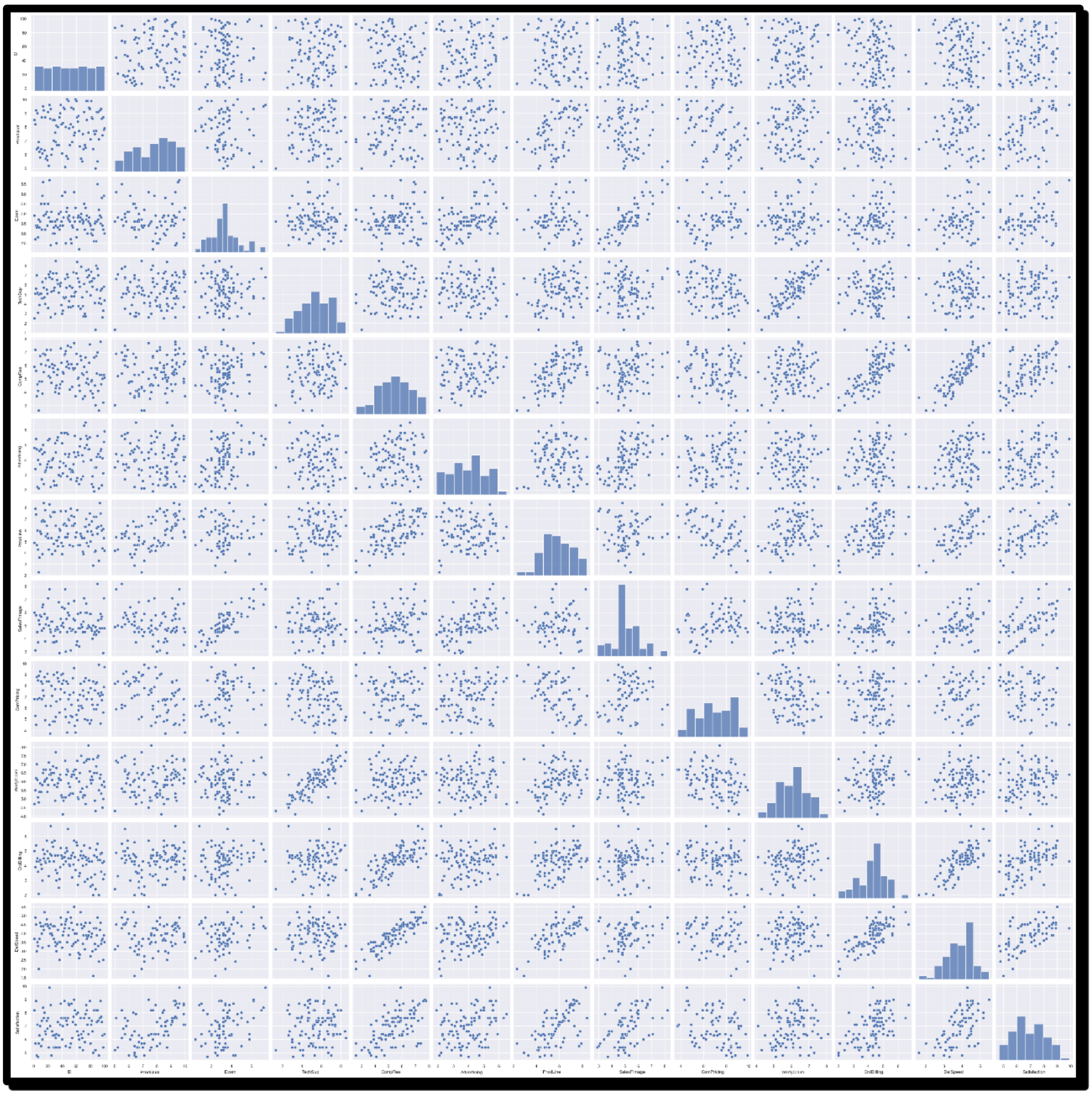


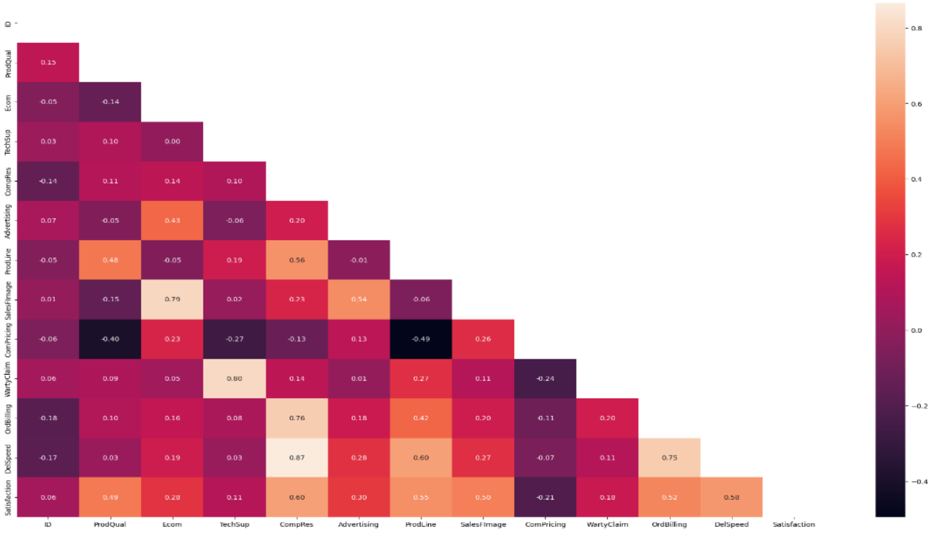



```
Description of Satisfaction
---------------------------------------------------------------------
count    100.000000
mean       6.918000
std        1.191839
min        4.700000
25%        6.000000
50%        7.050000
75%        7.625000
max        9.900000
Name: Satisfaction, dtype: float64 Distribution of Satisfaction
```



MULTI-VARIENT ANALYSIS:



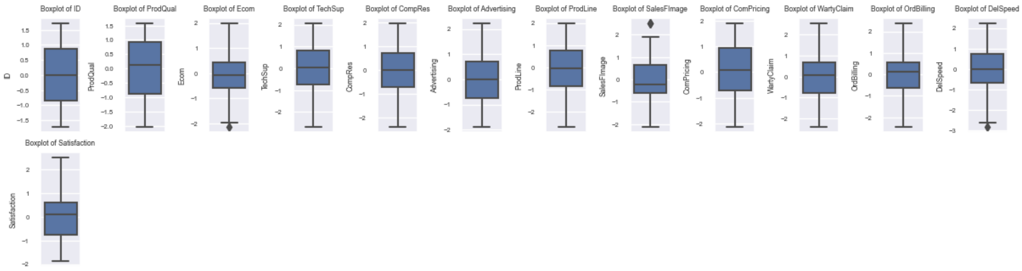


HEATMAP FOR CORRELATION

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

Here we use Standardscalar

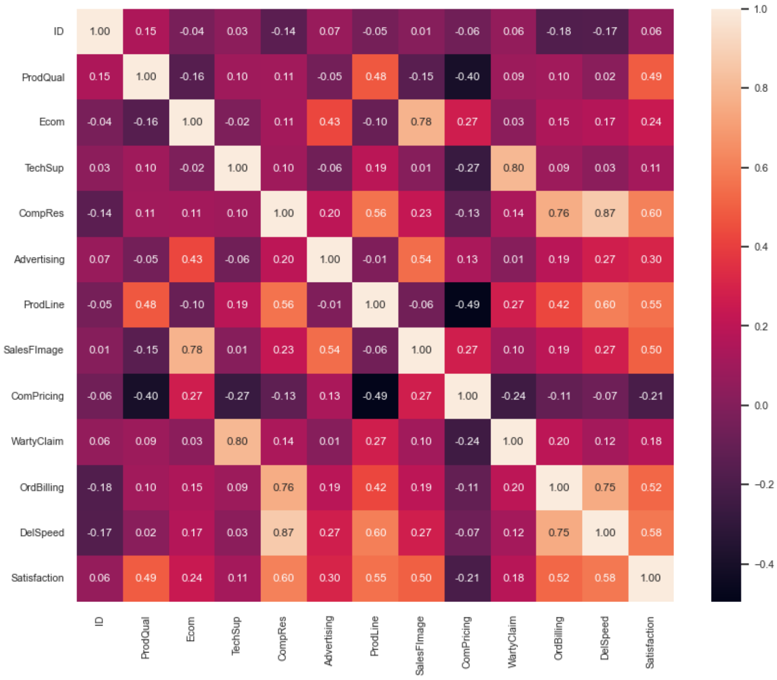
	ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
0	-1.714816	0.496660	0.401668	-1.881421	0.380922	0.704543	-0.691530	0.838627	-0.113185	-1.646582	0.791872	-0.260903	1.081067
1	-1.680173	0.280721	-1.495974	-0.174023	1.462141	-0.544014	1.600835	-1.917200	-1.088915	-0.665744	-0.411249	1.398918	-1.027098
2	-1.645531	1.000518	-0.389017	0.154322	0.131410	1.239639	1.218774	0.648570	-1.609304	0.192489	1.229371	0.845644	1.671354
3	-1.610888	-1.014914	-0.547153	1.073690	-1.448834	0.615361	-0.844354	-0.586801	1.187789	1.173327	0.026250	-1.229132	-1.786038
4	-1.576245	0.856559	-0.389017	-0.108354	-0.700298	-1.614207	0.149004	-0.586801	-0.113185	0.069885	0.244999	-0.537540	0.153474



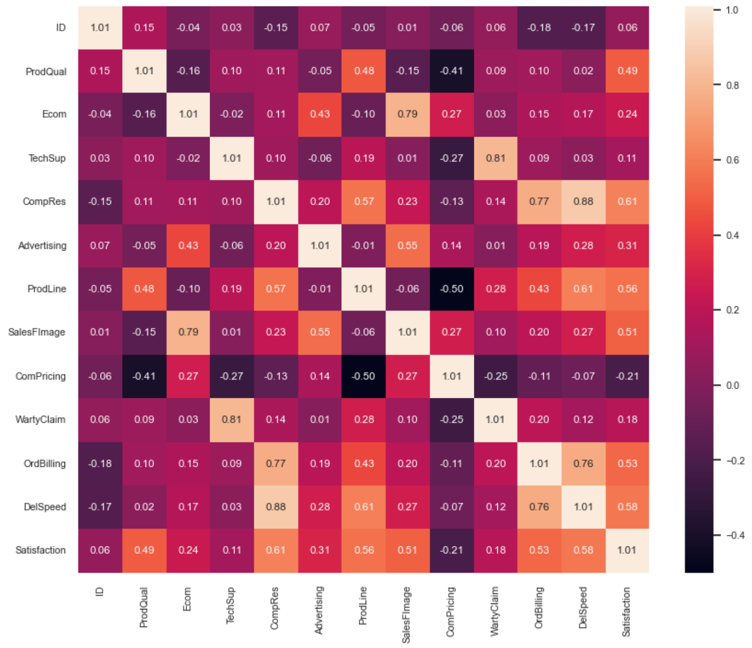
BOXPLOT OF FEATUERD COLUMN

	ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim
count	1.000000e+02	1.000000e+02	1.000000e+02	1.000000e+02	1.000000e+02	1.000000e+02	1.000000e+02	1.000000e+02	1.000000e+02	1.000000e+02
mean	4.440892e-18	9.188483e-16	7.216450e-16	1.029177e-15	-1.432188e-16	-6.061818e-16	2.531308e-16	2.592371e-16	-7.105427e-16	-1.247891e-15
std	1.005038e+00	1.005038e+00	1.005038e+00	1.005038e+00	1.005038e+00	1.005038e+00	1.005038e+00	1.005038e+00	1.005038e+00	1.005038e+00
min	-1.714816e+00	-2.022630e+00	-2.128522e+00	-2.669451e+00	-2.363712e+00	-1.881755e+00	-2.678246e+00	-2.107257e+00	-2.129693e+00	-2.382210e+00
25%	-8.574080e-01	-8.889494e-01	-5.866876e-01	-7.322109e-01	-7.002976e-01	-7.446754e-01	-8.443545e-01	-5.868010e-01	-7.148848e-01	-7.883484e-01
50%	0.000000e+00	1.367614e-01	-7.274293e-02	2.298420e-02	6.653659e-03	-8.918268e-03	-4.202669e-02	-2.066870e-01	8.196131e-02	6.988470e-02
75%	8.574080e-01	9.285383e-01	4.412017e-01	8.274312e-01	7.343976e-01	7.045432e-01	7.603011e-01	6.485695e-01	9.275939e-01	6.829084e-01
max	1.714816e+00	1.576356e+00	1.983036e+00	2.058728e+00	1.961166e+00	2.220649e+00	1.982896e+00	2.501625e+00	1.903324e+00	2.521979e+00

DESCRIBING THE SCALED DATA



HEATMAP OF CORRELATION OF THE SCALED DATA



HEATMAP OF COVARIANCE OF THE SCALED DATA

	ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfac
ID	1.010101	0.147247	-0.035803	0.032160	-0.145780	0.073868	-0.049132	0.008854	-0.063643	0.059184	-0.178085	-0.171489	0.06
ProdQual	0.147247	1.010101	-0.163220	0.096566	0.107444	-0.054013	0.482317	-0.147978	-0.405335	0.089204	0.103531	0.024577	0.49
Ecom	-0.035803	-0.163220	1.010101	-0.018976	0.110490	0.429417	-0.097316	0.787115	0.270772	0.027657	0.147985	0.169845	0.24
TechSup	0.032160	0.096566	-0.018976	1.010101	0.097633	-0.063505	0.194571	0.009936	-0.273522	0.805220	0.086307	0.029190	0.11
CompRes	-0.145780	0.107444	0.110490	0.097633	1.010101	0.198906	0.567088	0.228937	-0.129247	0.141827	0.765652	0.877623	0.60
Advertising	0.073868	-0.054013	0.429417	-0.063505	0.198906	1.010101	-0.011667	0.548407	0.135573	0.010901	0.189904	0.275730	0.30
ProdLine	-0.049132	0.482317	-0.097316	0.194571	0.567088	-0.011667	1.010101	-0.063216	-0.499948	0.275836	0.428152	0.606336	0.55
SalesFImage	0.008854	-0.147978	0.787115	0.009936	0.228937	0.548407	-0.063216	1.010101	0.273986	0.101972	0.196662	0.273952	0.50
ComPricing	-0.063643	-0.405335	0.270772	-0.273522	-0.129247	0.135573	-0.499948	0.273986	1.010101	-0.247461	-0.114463	-0.070999	-0.21
WartyClaim	0.059184	0.089204	0.027657	0.805220	0.141827	0.010901	0.275836	0.101972	-0.247461	1.010101	0.200107	0.117342	0.17
OrdBilling	-0.178085	0.103531	0.147985	0.086307	0.765652	0.189904	0.428152	0.196662	-0.114463	0.200107	1.010101	0.759896	0.52
DelSpeed	-0.171489	0.024577	0.169845	0.029190	0.877623	0.275730	0.606336	0.273952	-0.070999	0.117342	0.759896	1.010101	0.58
Satisfaction	0.061761	0.491237	0.244305	0.113735	0.609356	0.307747	0.556107	0.506129	-0.210400	0.179338	0.526590	0.583217	1.01

COVARIANCE OF THE SCALED DATA

	ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfac
ID	1.000000	0.145774	-0.035445	0.031838	-0.144322	0.073129	-0.048641	0.008765	-0.063007	0.058592	-0.176304	-0.169774	0.06
ProdQual	0.145774	1.000000	-0.161588	0.095600	0.106370	-0.053473	0.477493	-0.146498	-0.401282	0.088312	0.102495	0.024332	0.48
Ecom	-0.035445	-0.161588	1.000000	-0.018786	0.109386	0.425123	-0.096342	0.779244	0.268064	0.027380	0.146505	0.168147	0.24
TechSup	0.031838	0.095600	-0.018786	1.000000	0.096657	-0.062870	0.192625	0.009836	-0.270787	0.797168	0.085443	0.028898	0.11
CompRes	-0.144322	0.106370	0.109386	0.096657	1.000000	0.196917	0.561417	0.226647	-0.127954	0.140408	0.757995	0.868846	0.60
Advertising	0.073129	-0.053473	0.425123	-0.062870	0.196917	1.000000	-0.011551	0.542923	0.134217	0.010792	0.188005	0.272973	0.30
ProdLine	-0.048641	0.477493	-0.096342	0.192625	0.561417	-0.011551	1.000000	-0.062584	-0.494948	0.273078	0.423870	0.600272	0.55
SalesFImage	0.008765	-0.146498	0.779244	0.009836	0.226647	0.542923	-0.062584	1.000000	0.271246	0.100953	0.194695	0.271213	0.50
ComPricing	-0.063007	-0.401282	0.268064	-0.270787	-0.127954	0.134217	-0.494948	0.271246	1.000000	-0.244986	-0.113318	-0.070289	-0.20
WartyClaim	0.058592	0.088312	0.027380	0.797168	0.140408	0.010792	0.273078	0.100953	-0.244986	1.000000	0.198106	0.116168	0.17
OrdBilling	-0.176304	0.102495	0.146505	0.085443	0.757995	0.188005	0.423870	0.194695	-0.113318	0.198106	1.000000	0.752298	0.52
DelSpeed	-0.169774	0.024332	0.168147	0.028898	0.868846	0.272973	0.600272	0.271213	-0.070289	0.116168	0.752298	1.000000	0.57
Satisfaction	0.061143	0.486325	0.241862	0.112597	0.603263	0.304669	0.550546	0.501068	-0.208296	0.177545	0.521324	0.577385	1.00

CORRELATION OF THE SCALED DATA

CHI-SQUARE VALUE METHOD:

Confirm the statistical significance of correlations

- H0: Correlations are not significant, H1: There are significant correlations
- Reject H0 if p-value < 0.05

CHI-SQUARE VALUE: `6.813866448568822e-116`

KMO-MODEL

Confirm the adequacy of sample size.

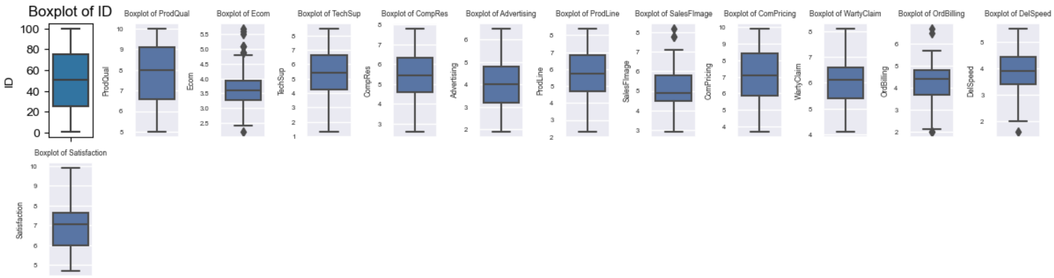
Note: Above 0.7 is good, below 0.5 is not acceptable

KMO-MODEL: `0.6608581001716486`

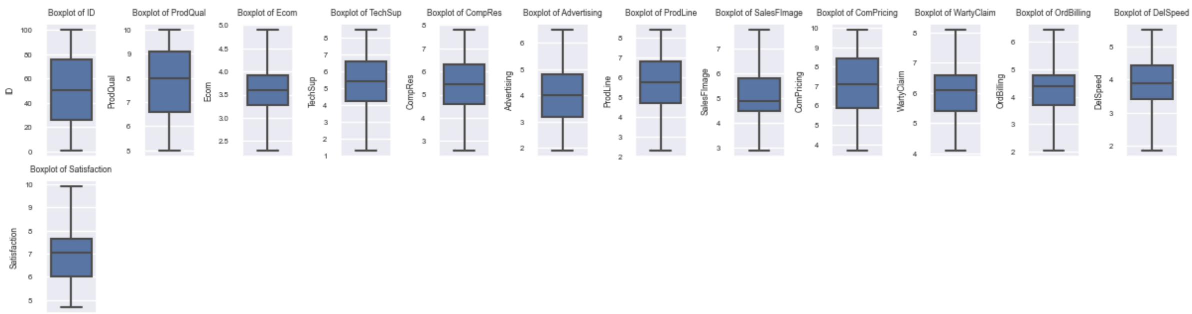
3.Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.

DATASETS BEFORE TREATING OUTLIERS(BEFORE SCALING)

Here we see that there are outliers that need to be treated as they may cause some error data during the calculation



DATASETS AFTER TREATING OUTLIERS: (AFTER SCALING)



4.Build the covariance matrix, eigenvalues and eigenvector.

```
array([[ 1.01010101,  0.14724686, -0.03580259,  0.03215965, -0.14578025,
         0.07386793, -0.04913229,  0.00885361, -0.0636433 ,  0.05918424,
        -0.17808478, -0.17148892,  0.06176054],
       [ 0.14724686,  1.01010101, -0.16322019,  0.09656612,  0.10744445,
        -0.05401327,  0.48231658, -0.14797813, -0.40533524,  0.08920435,
         0.1035307 ,  0.02457729,  0.49123737],
       [-0.03580259, -0.16322019,  1.01010101, -0.01897569,  0.11049041,
         0.42941698, -0.09731551,  0.78711486,  0.27077209,  0.02765676,
         0.14798521,  0.16984515,  0.24430522],
       [ 0.03215965,  0.09656612, -0.01897569,  1.01010101,  0.09763293,
        -0.06350512,  0.19457117,  0.00993585, -0.2735219 ,  0.80522013,
         0.08630654,  0.0291897 ,  0.11373452],
       [-0.14578025,  0.10744445,  0.11049041,  0.09763293,  1.01010101,
         0.19890591,  0.56708783,  0.22893666, -0.12924672,  0.14182656,
         0.765652  ,  0.87762252,  0.60935617],
       [ 0.07386793, -0.05401327,  0.42941698, -0.06350512,  0.19890591,
         1.01010101, -0.01166749,  0.54840714,  0.13557262,  0.01090109,
         0.18990387,  0.2757305 ,  0.30774694],
       [-0.04913229,  0.48231658, -0.09731551,  0.19457117,  0.56708783,
```

COVARIANCE MATRIX

```
Eigen Vectors
 %s [[ 4.62000728e-02 -1.58027218e-01 -1.39347642e-01 -1.25558836e-(
  -4.26608591e-01 -1.76000531e-01 -3.55715200e-01 -2.09925922e-01
   1.35540139e-01 -1.74756616e-01 -3.92818518e-01 -4.26486740e-01
  -4.11130831e-01]
 [-4.92986487e-02 -3.11729134e-01  4.51705349e-01 -2.38476094e-01
   1.05651742e-02  3.50505791e-01 -2.90553749e-01  4.59905998e-01
   4.17706010e-01 -2.01996322e-01  2.30352008e-02  6.48426938e-02
   2.38549586e-02]
 [-2.31074705e-01  6.06171213e-03 -2.42961427e-01 -5.75362555e-01
   2.10670802e-01 -1.36436961e-01  1.03155648e-01 -2.63464211e-01
   6.23167208e-02 -5.71736650e-01  1.72192934e-01  2.29051867e-01
  -2.37142718e-02]
 [ 4.98031848e-01  5.26361720e-01  8.21611606e-02 -2.86955669e-01
  -1.72431608e-01  2.10843618e-01  9.21428385e-02  1.35175834e-01
  -1.67593652e-01 -2.73014963e-01 -2.19362558e-01 -1.85250608e-01
   3.15912306e-01]
 [-7.86280201e-01  3.17591283e-01  2.75713175e-01 -1.07734443e-03
  -2.05117398e-01 -1.00017277e-01  9.63858666e-02  1.66993009e-01
  -1.77708048e-01 -6.01352770e-02 -1.63727955e-01 -2.00866396e-01
   1.15222615e-01]
 [-1.29092812e-01 -2.54237184e-01 -1.09902859e-01 -5.12484690e-02
  -5.95073442e-02  7.09513778e-01  5.31849876e-02 -1.05573224e-01
```

EIGEN VECTORS

```
array([[ 4.62000728e-02, -1.58027218e-01, -1.39347642e-01,
        -1.25558836e-01, -4.26608591e-01, -1.76000531e-01,
        -3.55715200e-01, -2.09925922e-01,  1.35540139e-01,
        -1.74756616e-01, -3.92818518e-01, -4.26486740e-01,
        -4.11130831e-01],
       [-4.92986487e-02, -3.11729134e-01,  4.51705349e-01,
        -2.38476094e-01,  1.05651742e-02,  3.50505791e-01,
        -2.90553749e-01,  4.59905998e-01,  4.17706010e-01,
        -2.01996322e-01,  2.30352008e-02,  6.48426938e-02,
         2.38549586e-02],
       [-2.31074705e-01,  6.06171213e-03, -2.42961427e-01,
        -5.75362555e-01,  2.10670802e-01, -1.36436961e-01,
         1.03155648e-01, -2.63464211e-01,  6.23167208e-02,
        -5.71736650e-01,  1.72192934e-01,  2.29051867e-01,
        -2.37142718e-02],
       [ 4.98031848e-01,  5.26361720e-01,  8.21611606e-02,
        -2.86955669e-01, -1.72431608e-01,  2.10843618e-01,
         9.21428385e-02,  1.35175834e-01, -1.67593652e-01,
        -2.73014963e-01, -2.19362558e-01, -1.85250608e-01,
         3.15912306e-01],
       [-7.86280201e-01,  3.17591283e-01,  2.75713175e-01,
        -1.07734443e-03, -2.05117398e-01, -1.00017277e-01,
         9.63858666e-02,  1.66993009e-01, -1.77708048e-01,
```

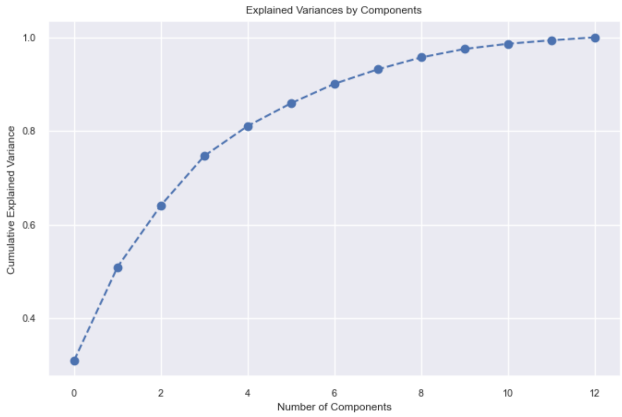
EIGEN VALUES

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

```
array([[ 4.62000728e-02, -1.58027218e-01, -1.39347642e-01,
        -1.25558836e-01, -4.26608591e-01, -1.76000531e-01,
        -3.55715200e-01, -2.09925922e-01,  1.35540139e-01,
        -1.74756616e-01, -3.92818518e-01, -4.26486740e-01,
        -4.11130831e-01],
       [-4.92986487e-02, -3.11729134e-01,  4.51705349e-01,
        -2.38476094e-01,  1.05651742e-02,  3.50505791e-01,
        -2.90553749e-01,  4.59905998e-01,  4.17706010e-01,
        -2.01996322e-01,  2.30352008e-02,  6.48426938e-02,
         2.38549586e-02],
       [-2.31074705e-01,  6.06171213e-03, -2.42961427e-01,
        -5.75362555e-01,  2.10670802e-01, -1.36436961e-01,
         1.03155648e-01, -2.63464211e-01,  6.23167208e-02,
        -5.71736650e-01,  1.72192934e-01,  2.29051867e-01,
        -2.37142718e-02],
       [ 4.98031848e-01,  5.26361720e-01,  8.21611606e-02,
        -2.86955669e-01, -1.72431608e-01,  2.10843618e-01,
         9.21428385e-02,  1.35175834e-01, -1.67593652e-01,
        -2.73014963e-01, -2.19362558e-01, -1.85250608e-01,
         3.15912306e-01],
       [-7.86280201e-01,  3.17591283e-01,  2.75713175e-01,
        -1.07734443e-03, -2.05117398e-01, -1.00017277e-01,
         9.63858666e-02,  1.66993009e-01, -1.77708048e-01,
```

6. Discuss the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate? Perform PCA and export the data of the Principal Component scores into a data frame.

```
Cumulative Variance Explained in Percentage: [ 30.97  50.92  64.09  74.65  81.04  85.91  90.06  93.16  95.71  97.51
  98.62  99.37 100.  ]
```




```
array([4.06686907, 2.61931822, 1.72985238, 1.3867214 , 0.83893283,
       0.63920482, 0.54565893, 0.40628835, 0.33492402, 0.23690301,
       0.14516191, 0.0981979 ])
```

EXPLAINED VARIANCE

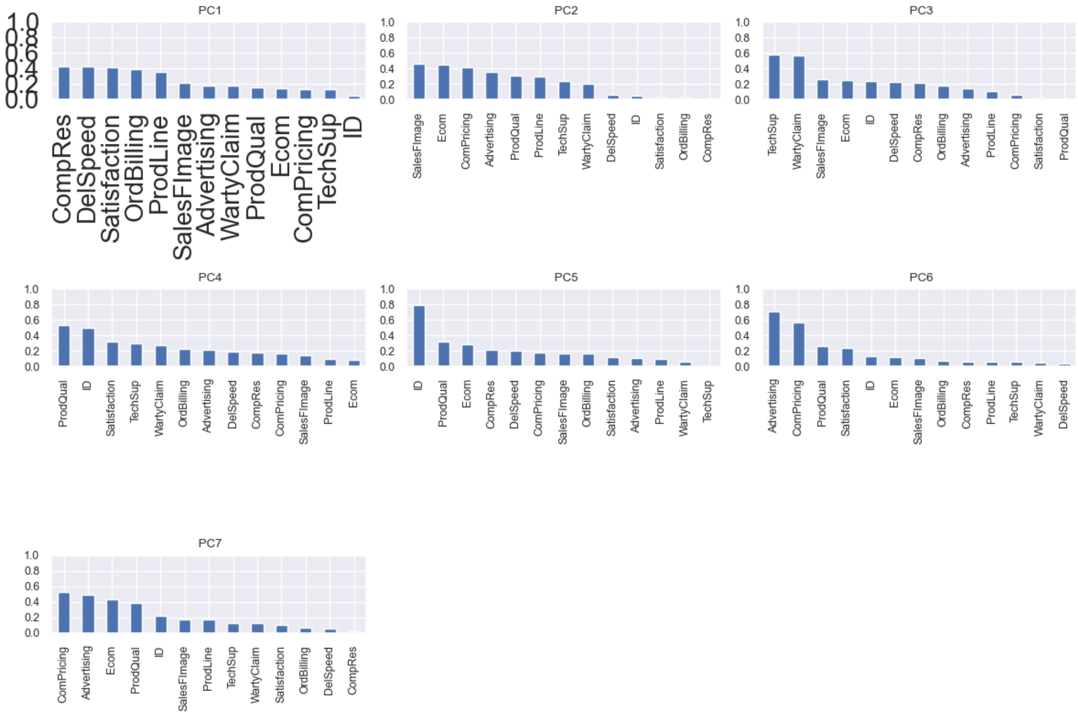
```
array([0.30970772, 0.19947116, 0.13173491, 0.10560417, 0.06388796,
       0.04867791, 0.04155403, 0.03094042, 0.02550575, 0.01804108,
       0.01105464, 0.00747815])
```

EXPLAINED VARIANCE RATIO

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
ID	0.046200	-0.049299	-0.231075	0.498032	-0.786280	-0.129093	-0.219561	0.012301	-0.089345	-0.023189	-0.033219	0.000188
ProdQual	-0.158027	-0.311729	0.006062	0.526362	0.317591	-0.254237	0.385338	0.152471	-0.303293	-0.167786	0.232317	0.197792
Ecom	-0.139348	0.451705	-0.242961	0.082161	0.275713	-0.109903	-0.432703	0.043870	-0.519890	-0.207306	0.029892	-0.001877
TechSup	-0.125559	-0.238476	-0.575363	-0.286956	-0.001077	-0.051248	0.124556	-0.003410	0.073852	-0.551308	-0.415346	0.000592
CompRes	-0.426609	0.010565	0.210671	-0.172432	-0.205117	-0.059507	-0.023714	-0.002732	0.125006	-0.440302	0.559335	-0.418546
Advertising	-0.176001	0.350506	-0.136437	0.210844	-0.100017	0.709514	0.489888	-0.061191	-0.114272	-0.038029	-0.031297	-0.083601
ProdLine	-0.355715	-0.290554	0.103156	0.092143	0.096386	0.053185	-0.169502	-0.625947	-0.267388	0.216321	-0.275454	-0.344435
SalesFImage	-0.209926	0.459906	-0.263464	0.135176	0.166993	-0.105573	-0.171123	-0.018366	0.353388	0.164451	0.063658	0.010554
ComPricing	0.135540	0.417706	0.062317	-0.167594	-0.177708	-0.565659	0.519245	-0.321743	-0.171635	0.030685	-0.097910	-0.101646
WartyClaim	-0.174757	-0.201996	-0.571737	-0.273015	-0.060135	-0.044456	0.122392	-0.044712	-0.098315	0.508872	0.451598	0.062451
OrdBilling	-0.392819	0.023035	0.172193	-0.219363	-0.163728	-0.070717	0.060498	0.646085	-0.290184	0.273157	-0.328822	-0.149740
DelSpeed	-0.426487	0.064843	0.229052	-0.185251	-0.200866	0.034993	-0.055222	-0.231918	-0.030467	-0.080806	-0.010522	0.788076
Satisfaction	-0.411131	0.023855	-0.023714	0.315912	0.115223	-0.234301	0.103396	0.045140	0.524374	0.119191	-0.236632	-0.047335

EXTRACTED LOADINGS OF THE DATASCALED

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



PC1 has **compres ,delspeed satisfaction,ord billing,prod line** equal and high followed by **salesfimage,advertising, warty claim,prod qual,Ecom,comp pricing,tech sup** features almost equal.

PC2 has all the features in decreasing trend starting from **salefimage** as the highest and **compres** as the lowest...

In PC 3 tech sup and **waranty** claim features have higher values up to 0.6 and remaining all features are in a decreasing trend starting from **salesfimage** as 0.25 ending with prod quality as the lowest

Pc 4 prod quality has the highest value upto 0.55 after that all features are seems to be in decreasing trend from satisfaction of 0.3 to Ecom as the lowest of 0.1

Pc 5 prod quality has the highest value of 0.3 to the waranty claim as the lowest of 0.5

Pc 6 advertising has the highest value of 0.7 to the delspeed having the lowest of 0.2 .

Pc 7 also all the features are in the decreasing trend starting from comprising with value of 0.5 to the lowest of compres with value of 0.1

As a conclusion pc1 is expaling the most variance with almost equal values in two sets..

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

Data Dictionary

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.)

	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

HEAD VALUES

	Unnamed: 0	States	Health_indeces1	Health_indices2	Per_capita_income	GDP
287	287	Gortnahey	2458	846	4137	124253
288	288	Goshedan	3109	818	1511	148660
289	289	Gracehill	2499	817	2649	127105
290	290	Grange_Corner	2953	811	1567	147103
291	291	Granville	2155	1052	4009	182653
292	292	Greencastle	3443	970	2499	238636
293	293	Greenisland	2963	793	1257	162831
294	294	Greyabbey	3276	609	1522	120184
295	295	Greysteel	3463	847	934	199403
296	296	Groggan	2070	838	3179	166767

TAIL VALUES

```
<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
```

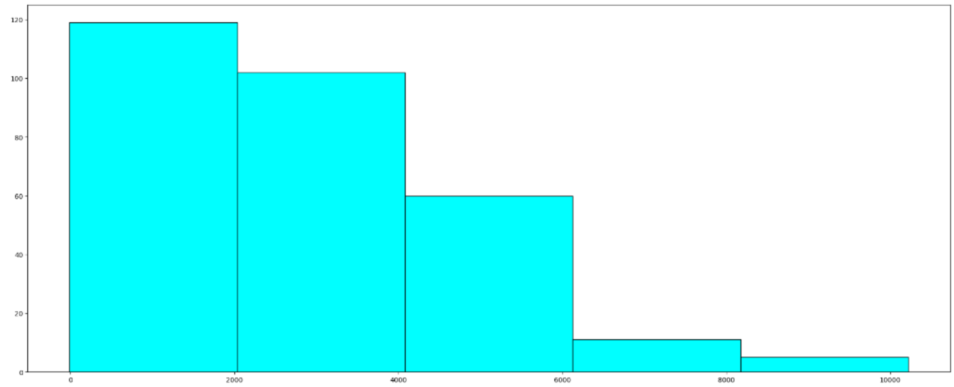
BASIC INFORMATION OF THE DATASET

There are two variables "Unnamed O' and States that signify only the id in the dataset and are not required in the clustering process. Hence, these can be dropped. After dropping these variables

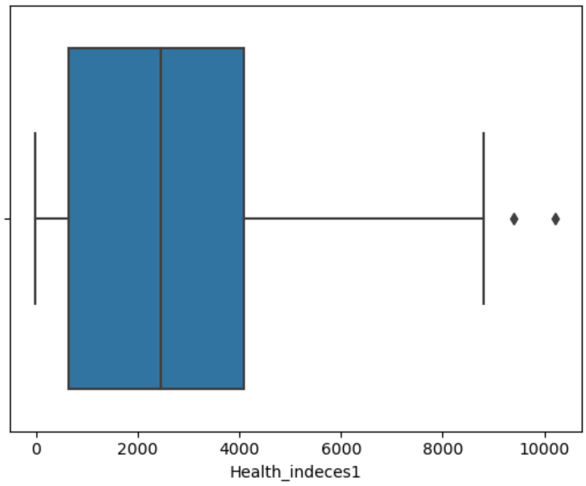
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 297 entries, 0 to 296
Data columns (total 5 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   States             297 non-null    object
 1   Health_indeces1    297 non-null    int64
 2   Health_indices2    297 non-null    int64
 3   Per_capita_income  297 non-null    int64
 4   GDP                297 non-null    int64
dtypes: int64(4), object(1)
memory usage: 11.7+ KB
```

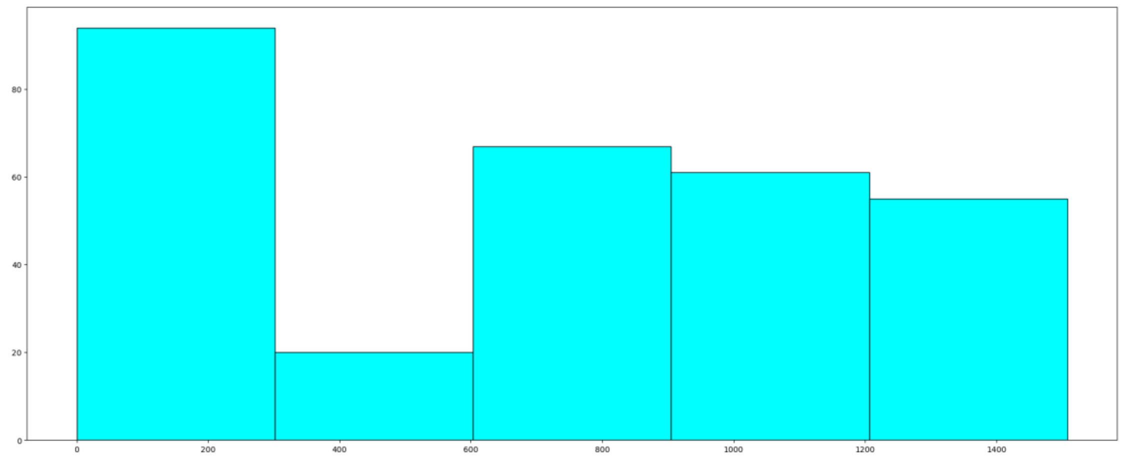
- There are 4 variables and 297 records.
- No missing record based on initial analysis.
- All the variables are integer-type variables.
- Shape of the Dataset: (297, 4)
- This shows the total number of rows = 297 and the total number of columns = 4.

UNIVARIATE -ANALYSIS:

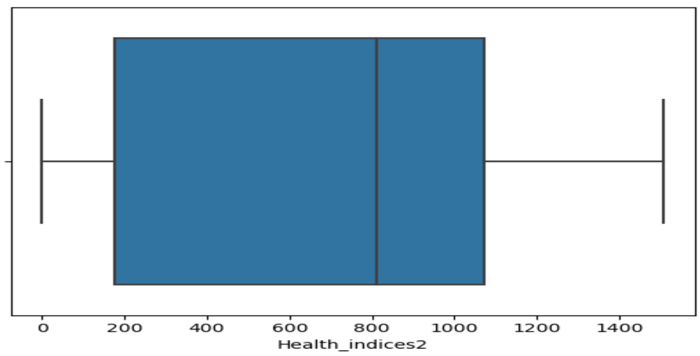


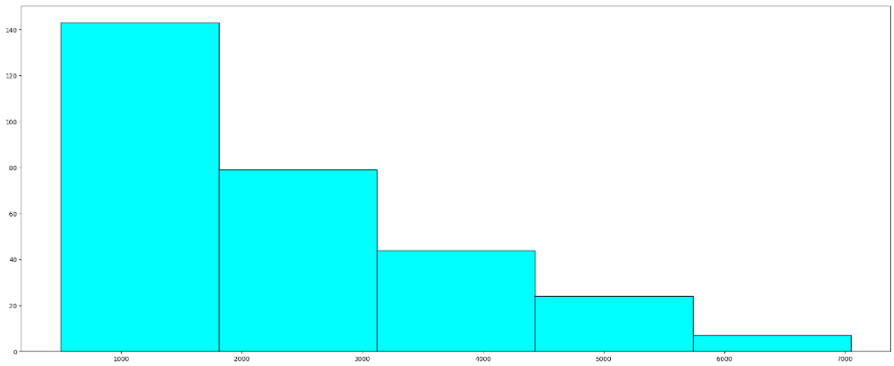

```
Description of Health_indeces1
----------------------------------------------------------------------------
count      297.000000
mean      2630.151515
std       2038.505431
min        -10.000000
25%        641.000000
50%       2451.000000
75%       4094.000000
max      10219.000000
Name: Health_indeces1, dtype: float64 Distribution of Health_indeces1
----------------------------------------------------------------------------
```



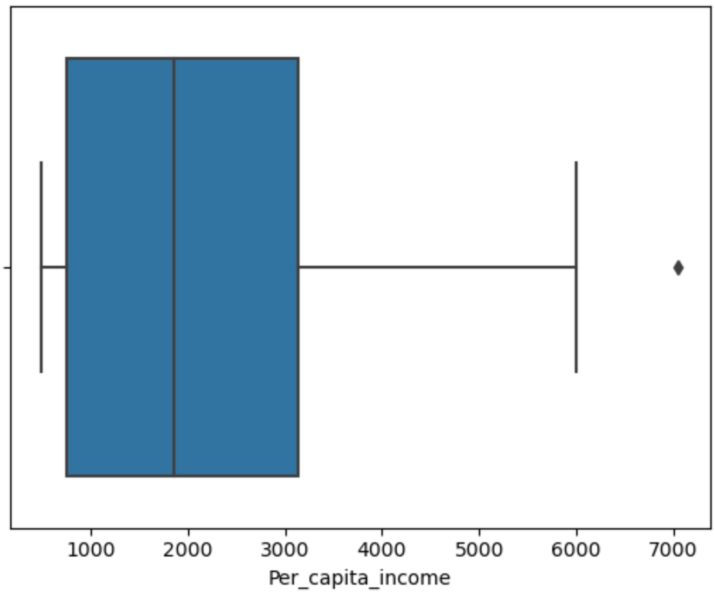


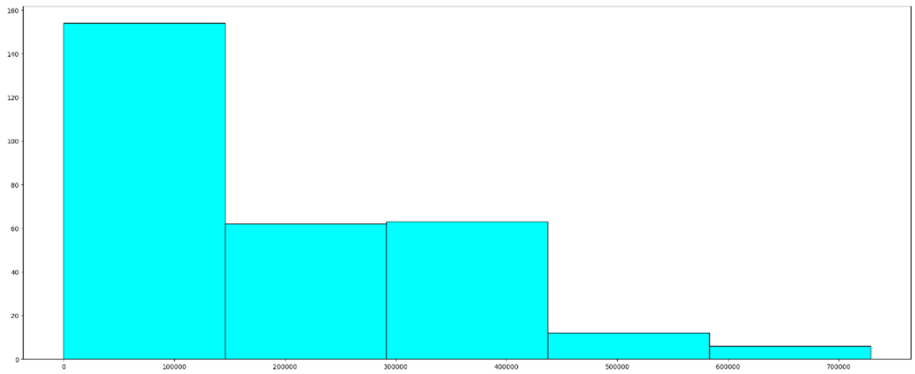

```
Description of Health_indices2
----------------------------------------------------------------------------
count     297.000000
mean      693.632997
std       468.944354
min         0.000000
25%       175.000000
50%       810.000000
75%      1073.000000
max      1508.000000
Name: Health_indices2, dtype: float64 Distribution of Health_indices2
```



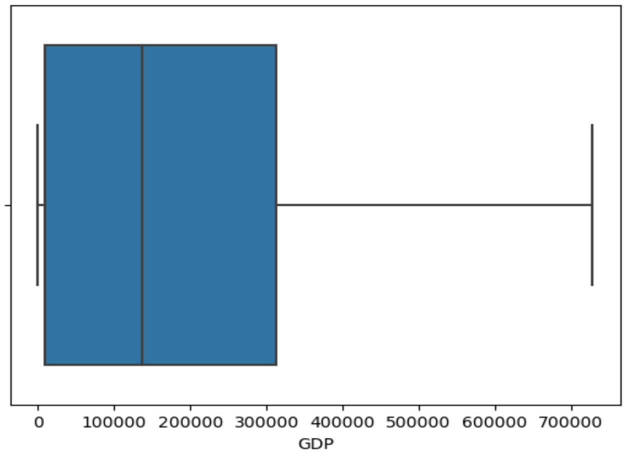


```
Description of Per_capita_income
----------------------------------------------------------------------------
count      297.000000
mean      2156.915825
std       1491.854058
min        500.000000
25%        751.000000
50%       1865.000000
75%       3137.000000
max       7049.000000
Name: Per_capita_income, dtype: float64 Distribution of Per_capita_income
```

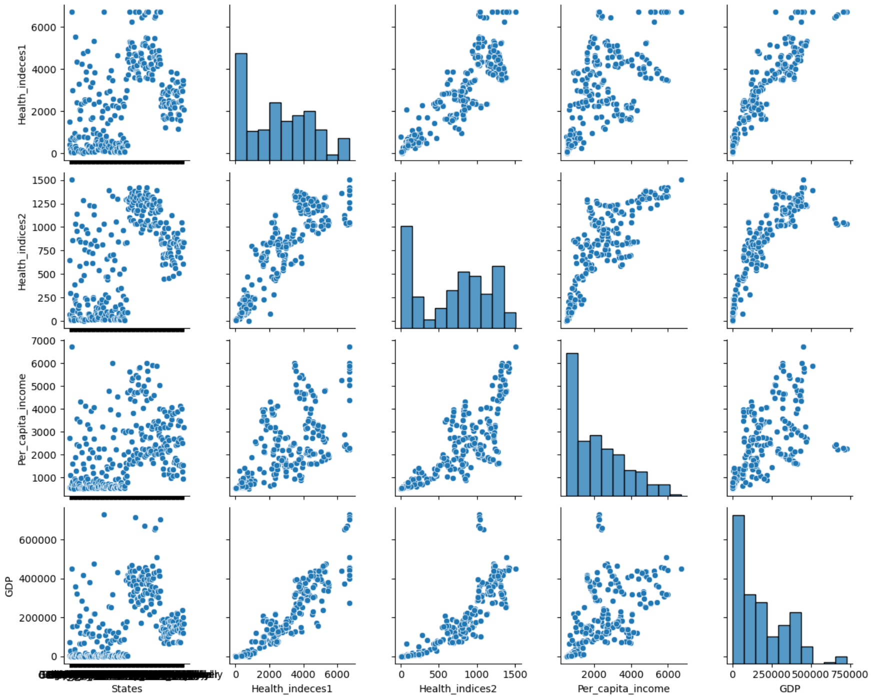




```
Description of GDP
--------------------------------------------------
count       297.000000
mean     174601.117845
std      167167.992863
min          22.000000
25%        8721.000000
50%      137173.000000
75%      313092.000000
max      728575.000000
Name: GDP, dtype: float64 Distribution of GDP
```



MULTI-VAIRENT ANALYSIS:



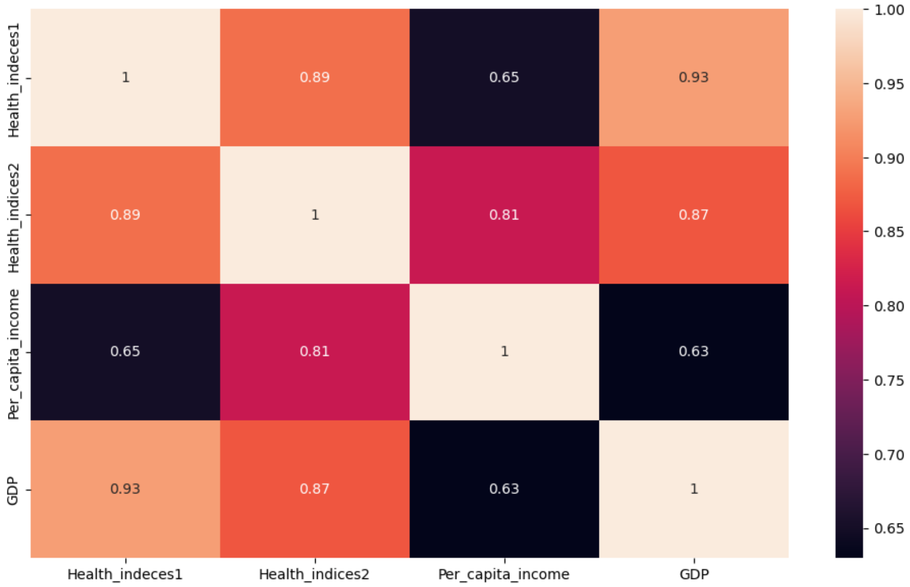
COVARIANCE MATRIX:

	Health_indeces1	Health_indices2	Per_capita_income	GDP
Health_indeces1	1.00000	0.887970	0.649780	0.927470
Health_indices2	0.88797	1.000000	0.812186	0.869385
Per_capita_income	0.64978	0.812186	1.000000	0.629663
GDP	0.92747	0.869385	0.629663	1.000000

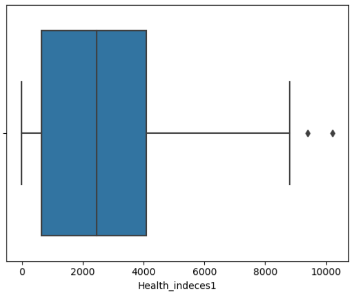
CORRELATED MATRIX:

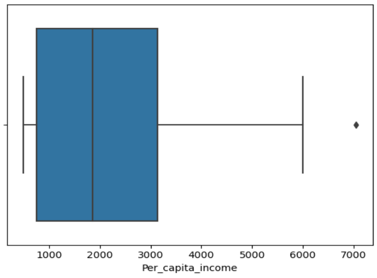
	Health_indeces1	Health_indices2	Per_capita_income	GDP
Health_indeces1	3.616758e+06	7.919162e+05	1.839129e+06	2.948580e+08
Health_indices2	7.919162e+05	2.199088e+05	5.668432e+05	6.815322e+07
Per_capita_income	1.839129e+06	5.668432e+05	2.214995e+06	1.566562e+08
GDP	2.948580e+08	6.815322e+07	1.566562e+08	2.794514e+10

HEATMAP OF CORRELATED MATRIX:



BEFORE TREATING OUTLIERS



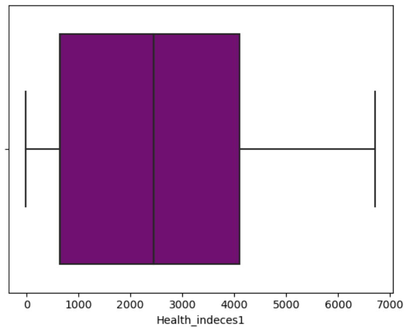


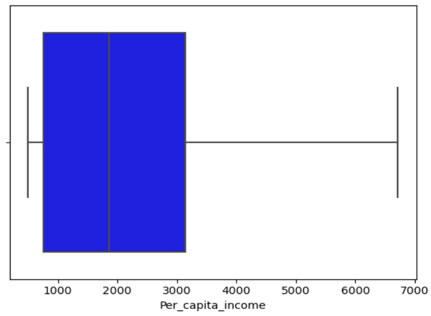
No. of outliers in Health indeces1: 2

No. of outliers in Per capita income: 1

Outlier Treatment- Instead of Imputing which causes data loss we will define a custom function- If for a particular column, the value is greater than the max value, then assign that max value to it. Same logic for the min value as well. This is known as min-max substitution.

AFTER TREATING OUTLIERS:



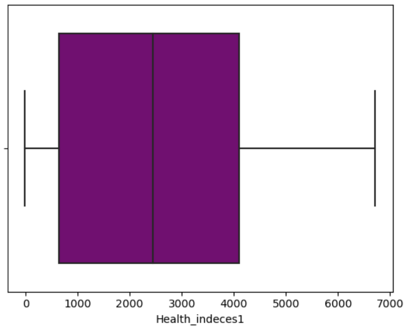


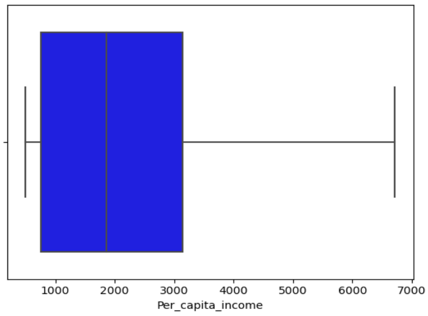
2.2. Do you think scaling is necessary for clustering in this case? Justify

Yes, Scaling is necessary as Clustering algorithms such as K-means do need feature scaling before they are fed 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

The above dataset consists of data with different units of measurement also known as weights, thus scaling them will form a common space and data will be from a relative range

We will use z-score scaling here, in which means and standard deviation1



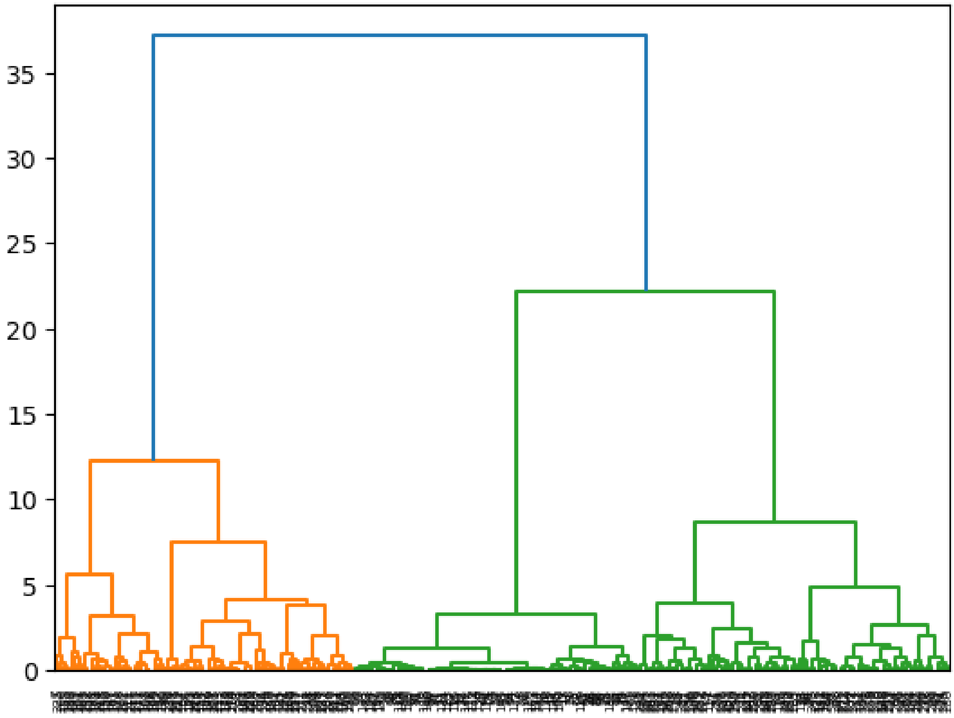


2.3. Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using a Dendrogram and briefly describe them.

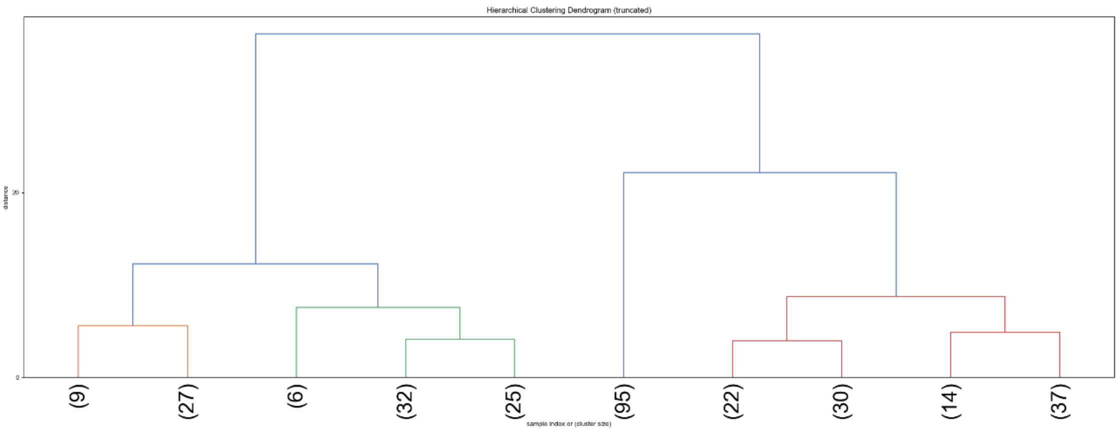
There are different methods of clustering, in this dataset we will use "Average" and "Ward" linkage methods. Average Linkage-

In this method, the distance between each pair of observations in each cluster is added up and divided by the number of pairs to get an average inter-cluster distance.

Average-Linkage and complete linkage are the two most popular distance metrics in hierarchical clustering



To make it clear we use truncate mode



Ward Linkage-

In this method, the linkage function describing the distance between two clusters is computed as the increase in the "error sum of squares" (ESS) after fusing two clusters into a single cluster. Ward's method chooses the successive steps in order to minimize the increase in ESS at each step.

```
[1188.0,
 455.29129911020243,
 244.47930171964813,
 168.20705625177374,
 133.7467703918479,
 106.23235342467379,
 90.67213692538147,
 79.40258262639014,
 70.72722521337673,
 63.082384046707325]
```

-WSS

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

	Health_indeces1	Health_indices2	Per_capita_income	GDP
0	-1.138661	-1.340654	-1.071354	-1.035304
1	-0.576133	-0.101746	0.373007	-0.604838
2	-1.013831	-0.842955	-0.707908	-0.882536
3	-1.257171	-1.428232	-1.065297	-1.044730
4	-1.335651	-1.464545	-1.095584	-1.046096
...	...	...	...	...
292	0.455167	0.590333	0.230994	0.383704
293	0.202346	0.212253	-0.604932	-0.070528
294	0.367206	-0.180780	-0.426574	-0.326073
295	0.465701	0.327599	-0.822326	0.148615
296	-0.268007	0.308375	0.688666	-0.046943

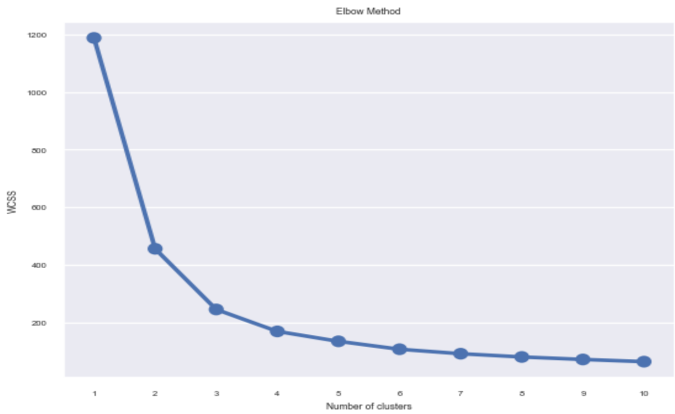
SCALED DATA

K-Mean Clustering. This is an Iterative method of partitioning the data into K-predefined distinct non-overlapping subgroups also known as clusters. In this, Each data point belongs to a single group. In the intra-cluster data points are as similar as possible while the distance between different clusters is as far as possible.

Working steps of an algorithm Specify the number of clusters K

- Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement
- Keep iterating until there is no change to the centroids, The assignment of data points to clusters isn't changing
- Compute the sum of the squared distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid)

K- means from k=1 to k=10

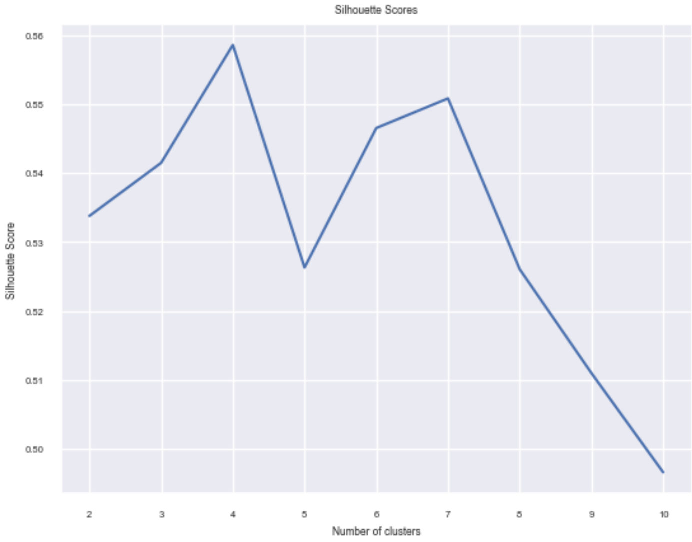


ELBOW METHOD

Silhouette Method- In this we compute the silhouette coefficients for each data point. It is the measure of how close it is to its own cluster rather than other clusters.

Silhouette Score-0.5340151343712788 i.e.,(k=2 to k=10)

```
[0.5337921355008507,
 0.5414933372475436,
 0.5585921135132221,
 0.5263164972079487,
 0.5465444819504095,
 0.5508371876705965,
 0.526060813993696,
 0.5109978994519919,
 0.4966104843213452]
```



	States	Health_indeces1	Health_indices2	Per_capita_income	GDP	Clus_kmeans5	cluster	Clus_kmeans3
0	Bachevo	417.0	66	564.0	1823	2	2	2
1	Balgarchevo	1485.0	646	2710.0	73662	3	0	3
2	Belasitsa	654.0	299	1104.0	27318	2	2	2
3	Belo_Pole	192.0	25	573.0	250	2	2	2
4	Beslen	43.0	8	528.0	22	2	2	2

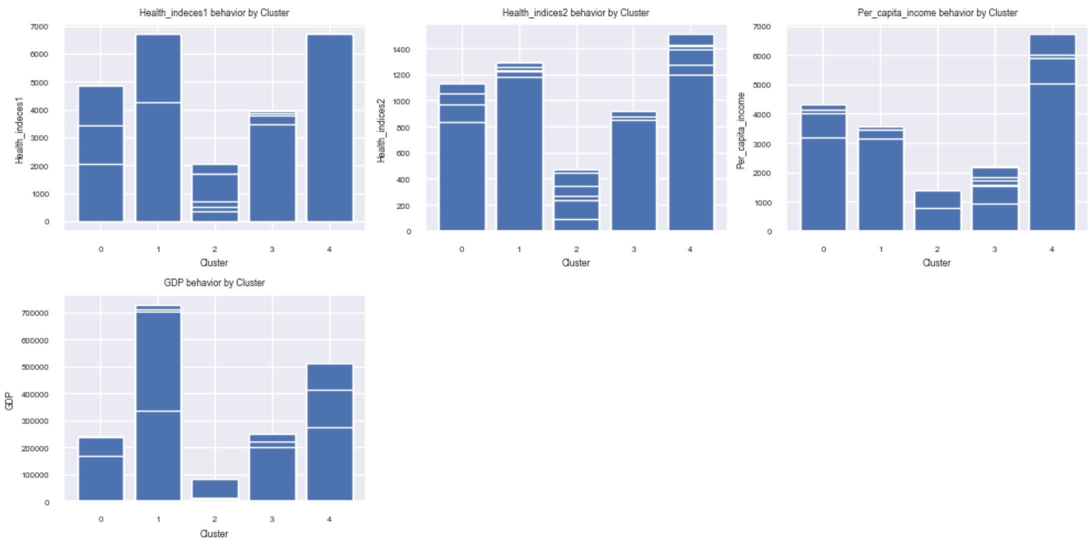
Observations - Based on the above cluster solution, 3 cluster solution seems to be the best fit as it c best fit as it differentiate the 3 clusters as

- High GDP per capita area
- Medium GDP per capita ama
- Low GDP per capita area

1.5. 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

Our main objective was to divide the data into an optimal number of clusters. From both the hierarchal clustering and K-means clustering, we get 3 as the optimal number of clusters

Clus_kmeans5	Health_indeces1	Health_indices2	Per_capita_income	GDP	cluster	Clus_kmeans3	freq
0	1957.736842	505.947368	1451.736842	75629.157895	0.631579	0.0	19
1	4700.434783	1189.130435	3150.130435	381874.304348	1.000000	1.0	23
2	404.680851	96.734043	668.691489	5359.744681	2.000000	2.0	94
3	1872.000000	793.724138	3596.310345	133221.344828	0.000000	3.0	29
4	4059.520000	1316.440000	4941.440000	342414.080000	1.000000	4.0	25
5	4499.718750	1135.000000	1923.000000	353950.781250	1.000000	5.0	32
6	6606.833333	1044.000000	2299.833333	687649.666667	1.000000	6.0	6
7	2977.300000	908.033333	2652.033333	158369.400000	0.000000	7.0	30
8	6662.444444	1369.666667	5555.444444	426759.111111	1.000000	8.0	9
9	3129.233333	729.066667	1488.466667	155488.166667	0.000000	9.0	30



Cluster 1: High GDP per capita Areas:

- These are the areas which have the highest growth rate.
- The health and economic conditions in these areas are excellent Per capita income in these areas is very high.

Cluster 2: Low GDP per capita Areas

- These are the areas which have very low growth rates.
- The health and economic conditions are not good in these areas Per capita income in these areas is very low.

Cluster 3: Medium GDP per capita Areas

- These are the areas which have an average growth rate.
- The health and economic conditions in these areas are adequate-Per capita income in these areas is average.

Recommendations for each cluster profile.

The main features that affect the Health and Economic conditions are workforce and productivity, The Higher these attributes higher the GDP per capita and thus higher the Health and Economic conditions