Data Mining

DATA MINING GRADED PROJECT

PREPARED FOR
DATA MINING MENTOR

PREPARED BY
NITIN KUMAR SINGH

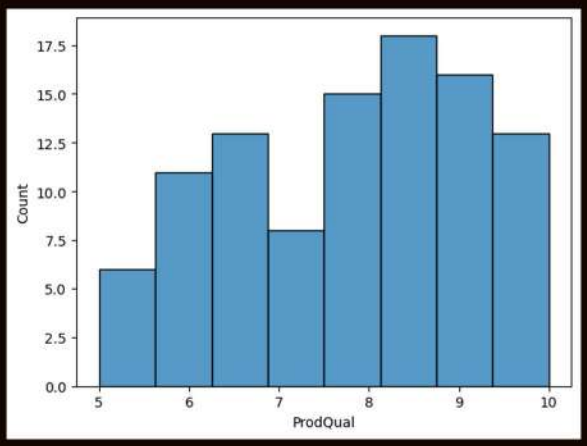
INDEX

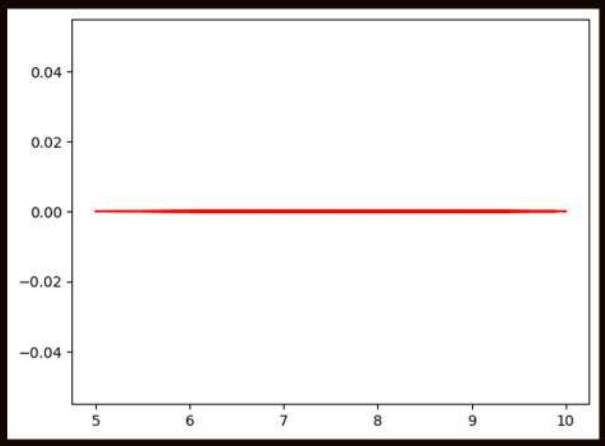
Perform Exploratory Data Analysis

There are 100 Rows * 13 Column are present in our dataset, 1 Columns are int64 (ID) and 12 Column are object64 (ProdQual, Ecom ,TechSup , CompRes , Advertising , ProdLine, SalesFImage, ComPricing, WartyClaim, OrdBilling, DelSpeed, Satisfaction)

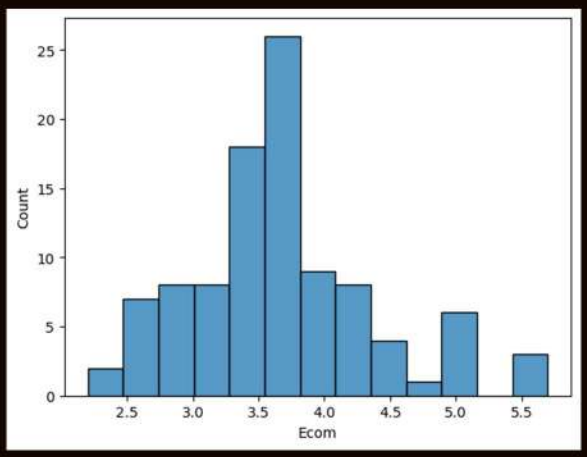
No null value and no duplicated value present in dataset

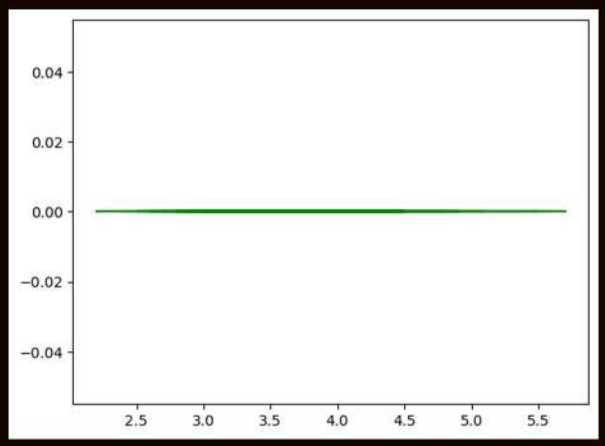
UNIVARIATE ANALYSIS



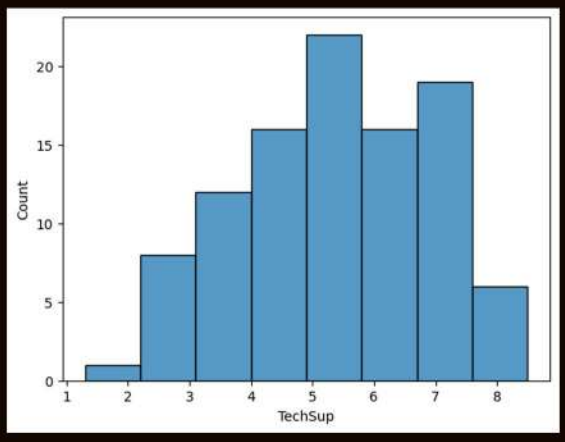


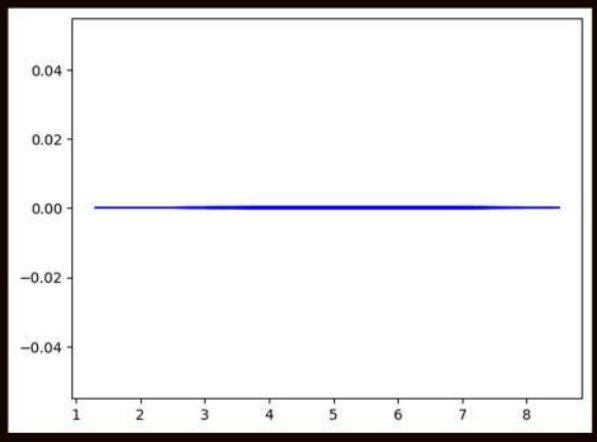
IN ProdQual column data lie between 5 to 10 range, most of the data between lie betwwen 7.5-10 and total sum of 780.9



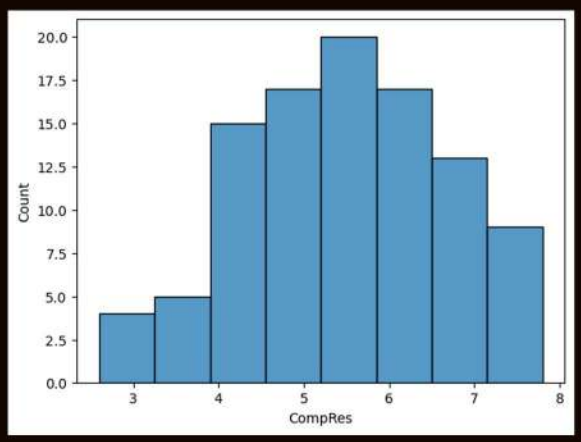


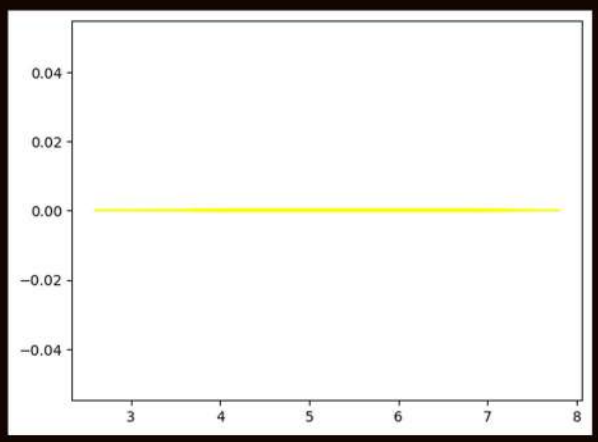
IN Ecom column data lie between 2.2-5.7 range, most of the data between lie betwwen 2.5-4.2 and total sum of 102.2



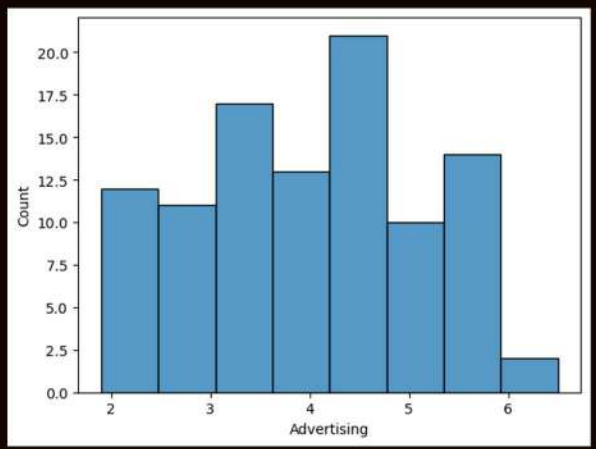


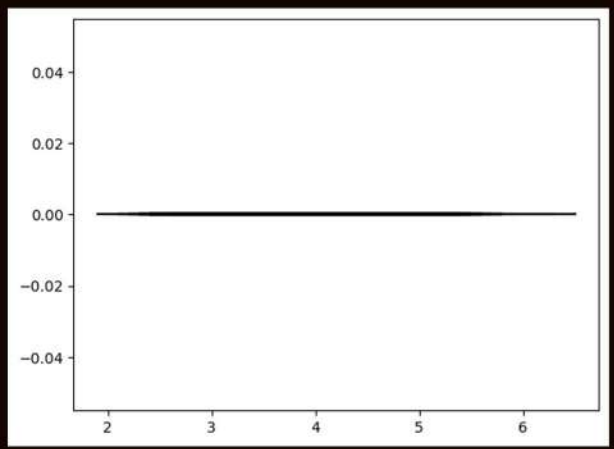
IN TechSup column data lie between 1.3 to 8.5 range, most of the data between lie between approx 2.1-7.8 and total sum of 536.4



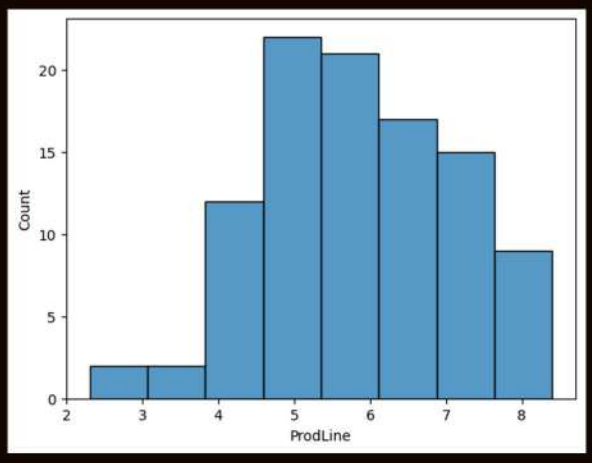


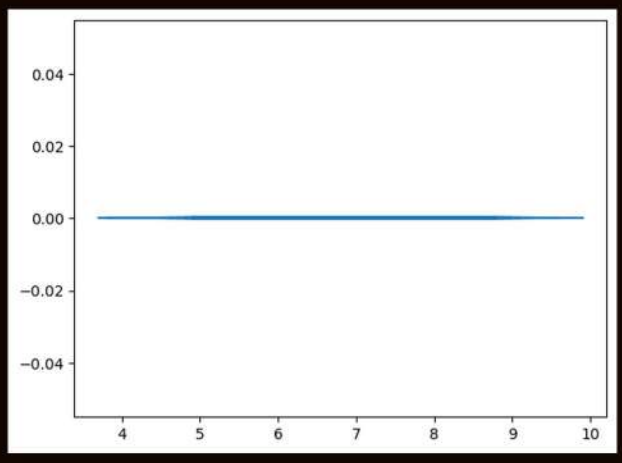
IN CompRes column data lie between 2.6 to 7.8 range, most of the data between lie between approx 3.9-7.8 and total sum of 544.2



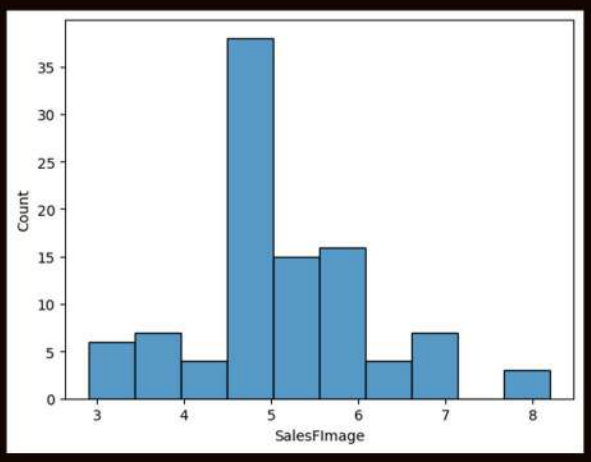


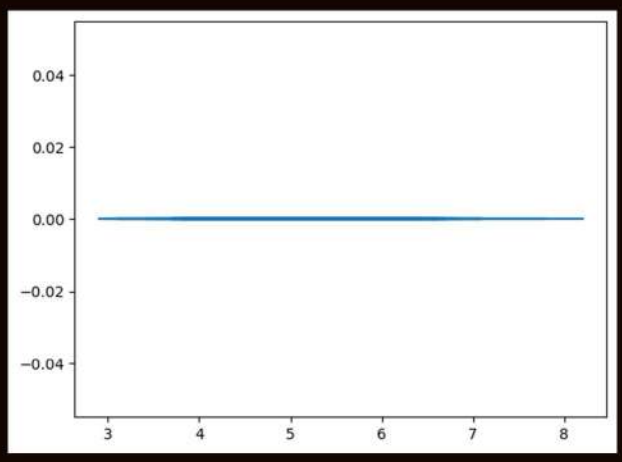
IN Advertising column data lie between 1.9 to 6.5 range, most of the data between lie between approx 1.9-5.9 and total sum of 401.0



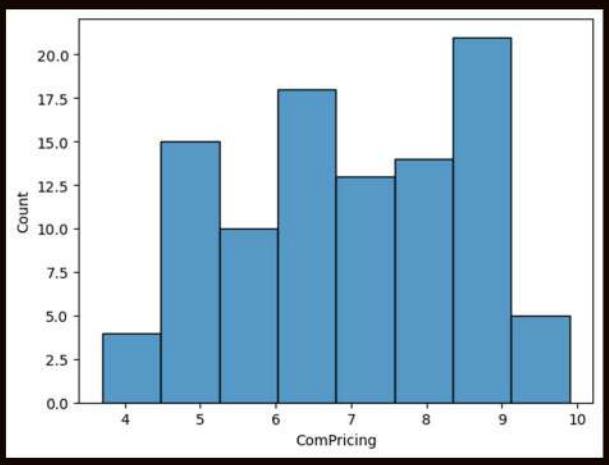


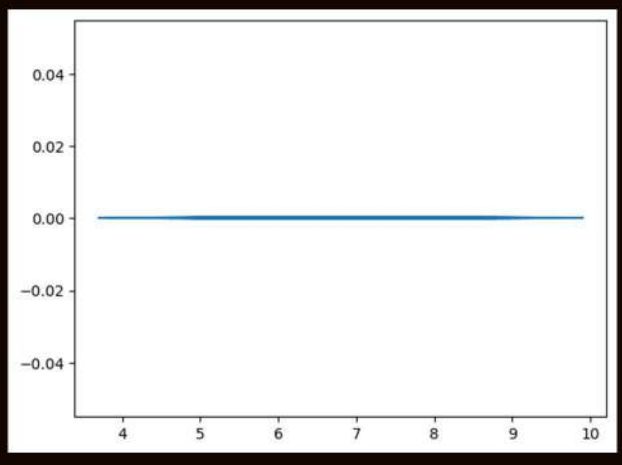
IN ComPricing column data lie between 2.3 to 8.4 range, most of the data between lie between approx 3.8-8.4 and total sum of 580.4



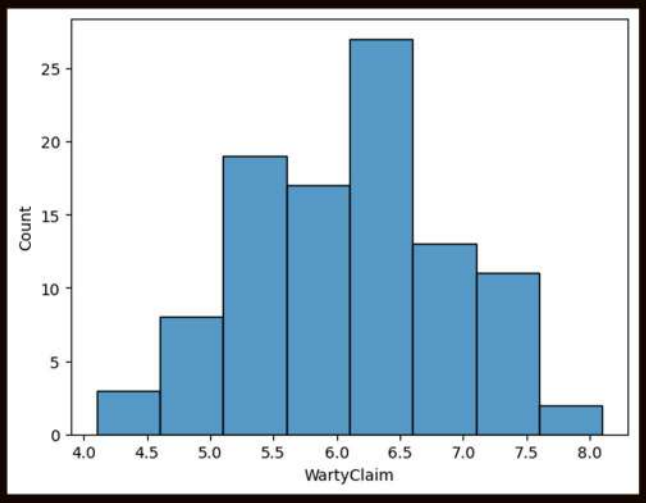


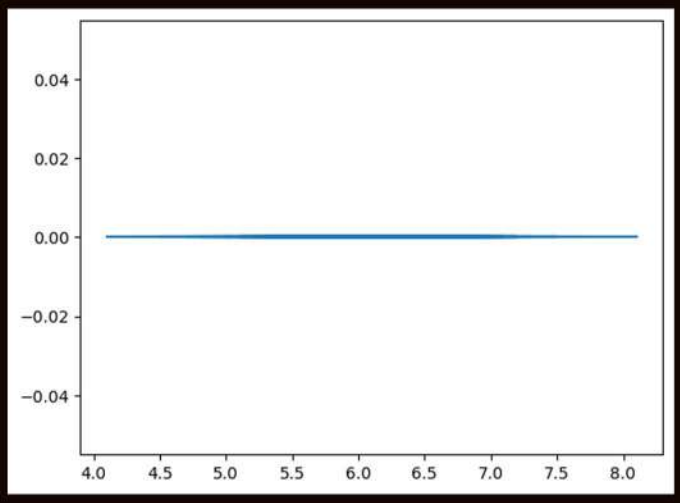
IN SalesFImage column data lie between 2.9 to 8.2 range, most of the data between lie between approx 4.5-6.1 and total sum of 512.3



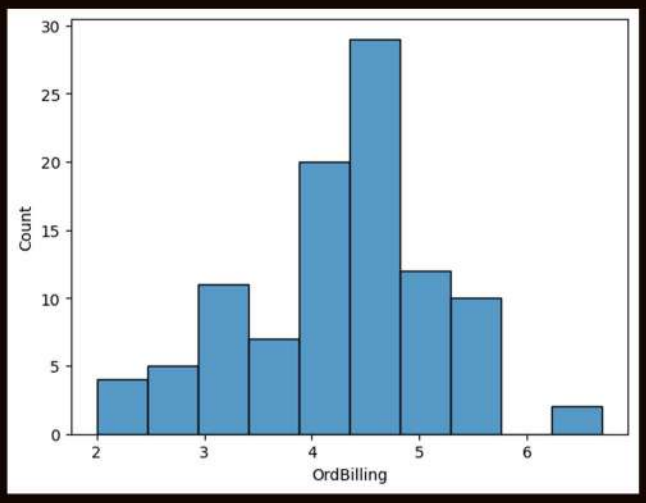


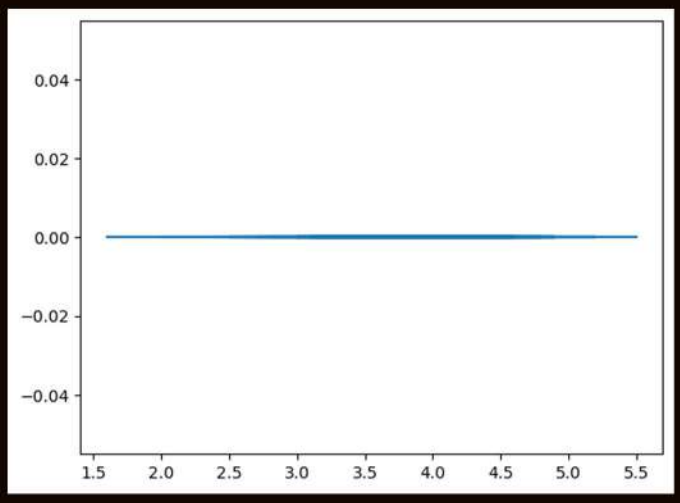
IN ComPricing column data lie between 3.7 to 9.9 range, most of the data between lie between approx 4.5-9.1 and total sum of 697.4



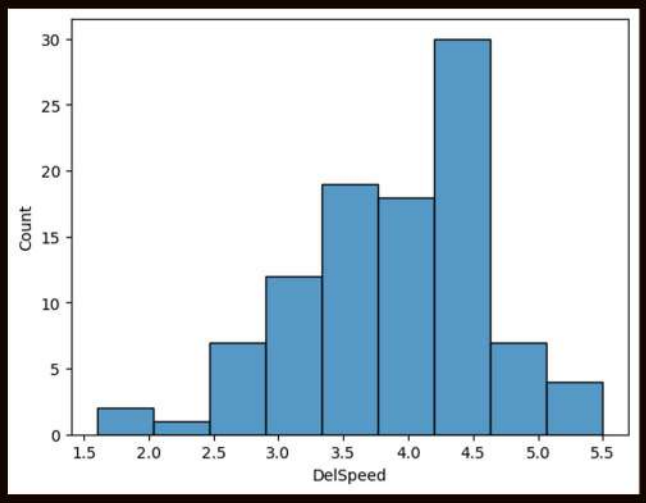


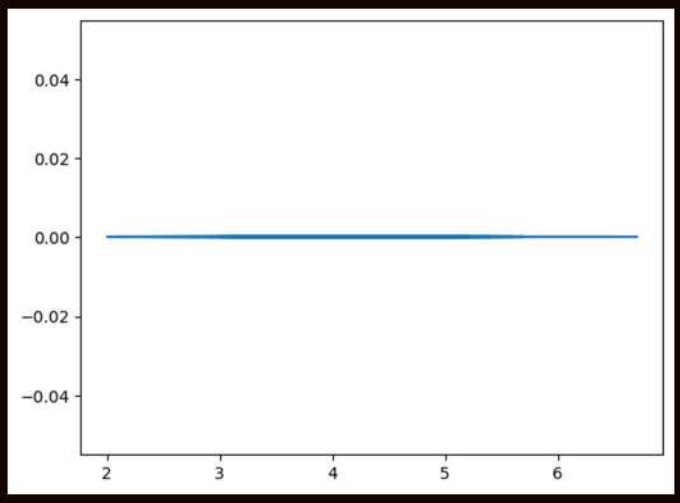
IN WartyClaim column data lie between 4.1to 8.1range, most of the data between lie between approx 5.1-7.6 and total sum of 604.3



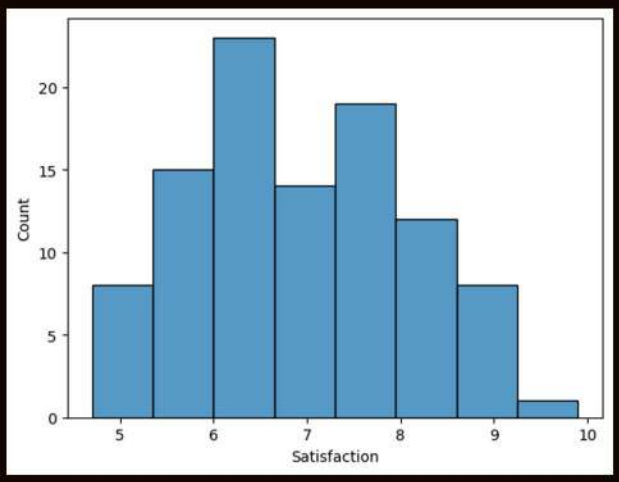


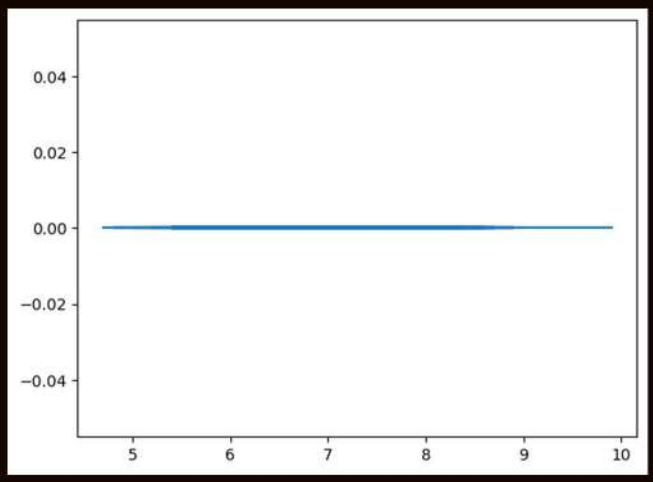
IN WartyClaim column data lie between 2.0 to 6.7 range, most of the data between lie between approx 3.9 -5.8 and total sum of 427.8



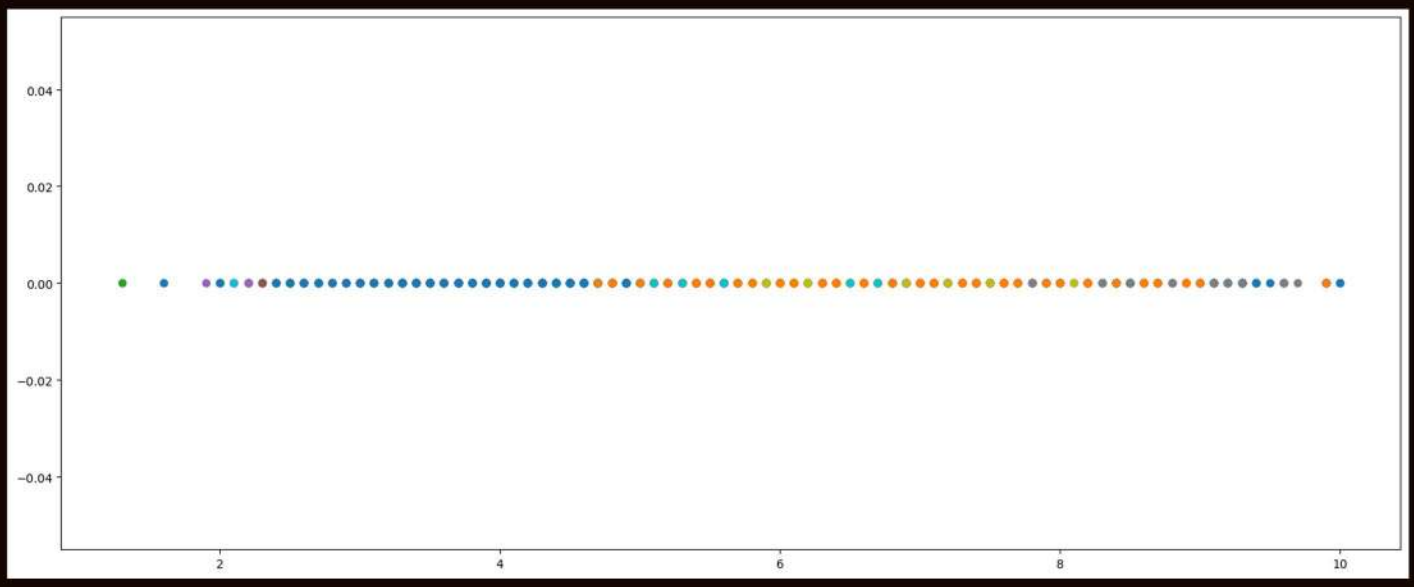


IN DelSpeed column data lie between 1.6 to 5.5 range, most of the data between lie between approx 2.9 -4.2 and total sum of 388.5





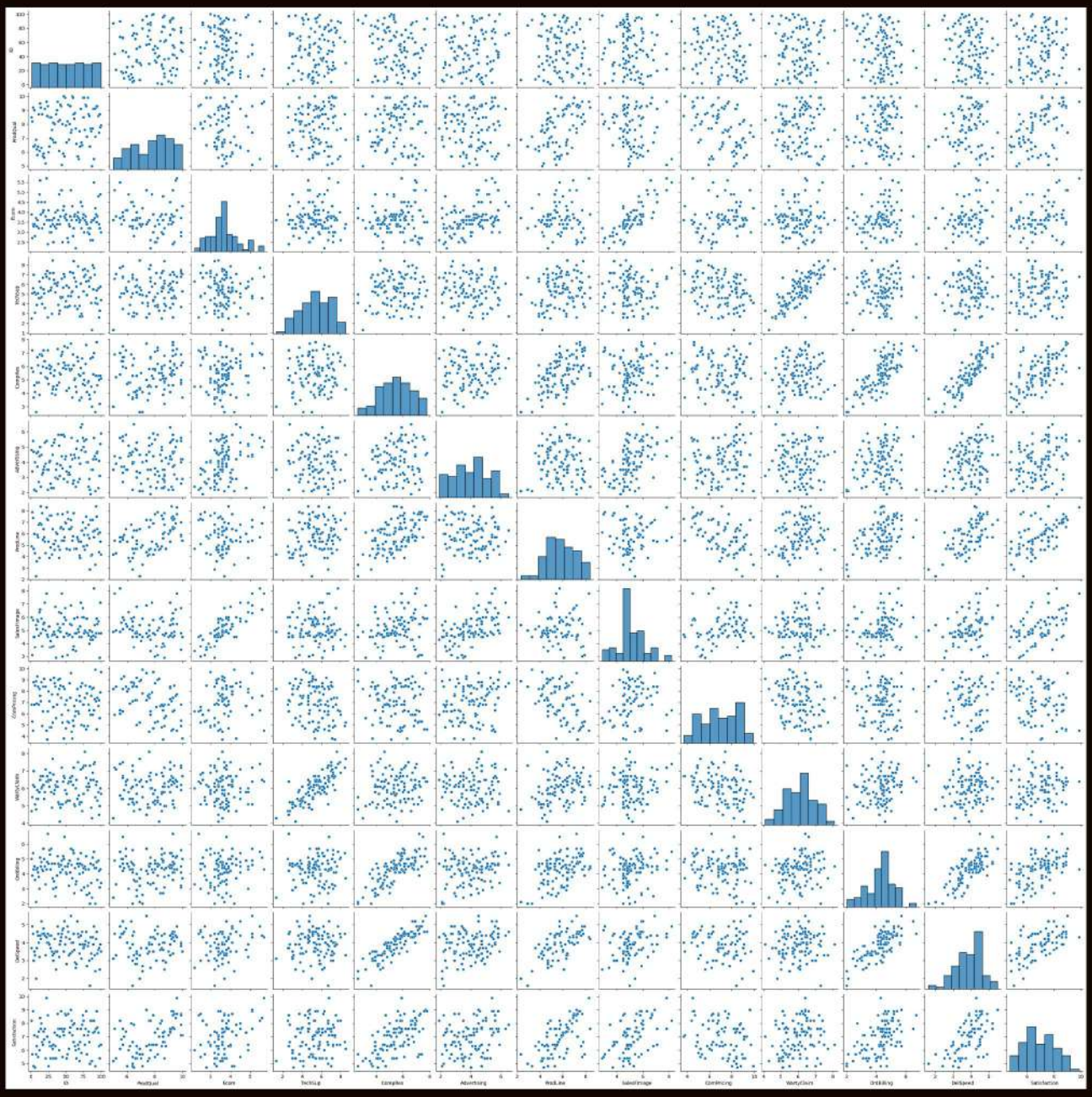
IN Satisfaction column data lie between 4.7 to 9.9range, most of the data between lie between approx 5.4-9.2 and total sum of 691.8



All feature lie between in this Range

MULTIVARIATE ANALYSIS

CORRELATION IN FEATURES



	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

COVARIANCE IN FEATURES

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
ProdQual	1.949596	-0.134162	0.204293	0.179475	-0.084141	0.876919	-0.227303	-0.865697	0.101081	0.135273	0.028424	0.809313
Ecom	-0.134162	0.490723	0.000929	0.118663	0.339374	-0.048545	0.594590	0.248356	0.029802	0.101600	0.098594	0.236065
TechSup	0.204293	0.000929	2.342298	0.178758	-0.108434	0.387753	0.027884	-0.640313	1.000106	0.113869	0.028596	0.205384
CompRes	0.179475	0.118663	0.178758	1.460238	0.268162	0.892313	0.297711	-0.238897	0.139085	0.849519	0.767766	0.868832
Advertising	-0.084141	0.339374	-0.108434	0.268162	1.270000	-0.017121	0.655222	0.233697	0.009970	0.192848	0.228323	0.409212
ProdLine	0.876919	-0.048545	0.387753	0.892313	-0.017121	1.729975	-0.086480	-1.005828	0.294429	0.518495	0.581384	0.863040
SalesFImage	-0.227303	0.594590	0.027884	0.297711	0.655222	-0.086480	1.149870	0.438382	0.094456	0.194349	0.213861	0.639279
ComPricing	-0.865697	0.248356	-0.640313	-0.238897	0.233697	-1.005828	0.438382	2.387196	-0.310285	-0.164416	-0.082691	-0.383568
WartyClaim	0.101081	0.029802	1.000106	0.139085	0.009970	0.294429	0.094456	-0.310285	0.671971	0.150046	0.065861	0.173461
OrdBilling	0.135273	0.101600	0.113869	0.849519	0.192848	0.518495	0.194349	-0.164416	0.150046	0.862743	0.512315	0.577572
DelSpeed	0.028424	0.098594	0.028596	0.767766	0.228323	0.581384	0.213861	-0.082691	0.065861	0.512315	0.539398	0.505103
Satisfaction	0.809313	0.236065	0.205384	0.868832	0.409212	0.863040	0.639279	-0.383568	0.173461	0.577572	0.505103	1.420481

SCALE THE VARIABLE

Using a scipy.stats function import z-score after using this funcion is DATA IS SCALED

scaled data is pre-proccessing procces for apply PCA it is use to mean centering of data and SD-1

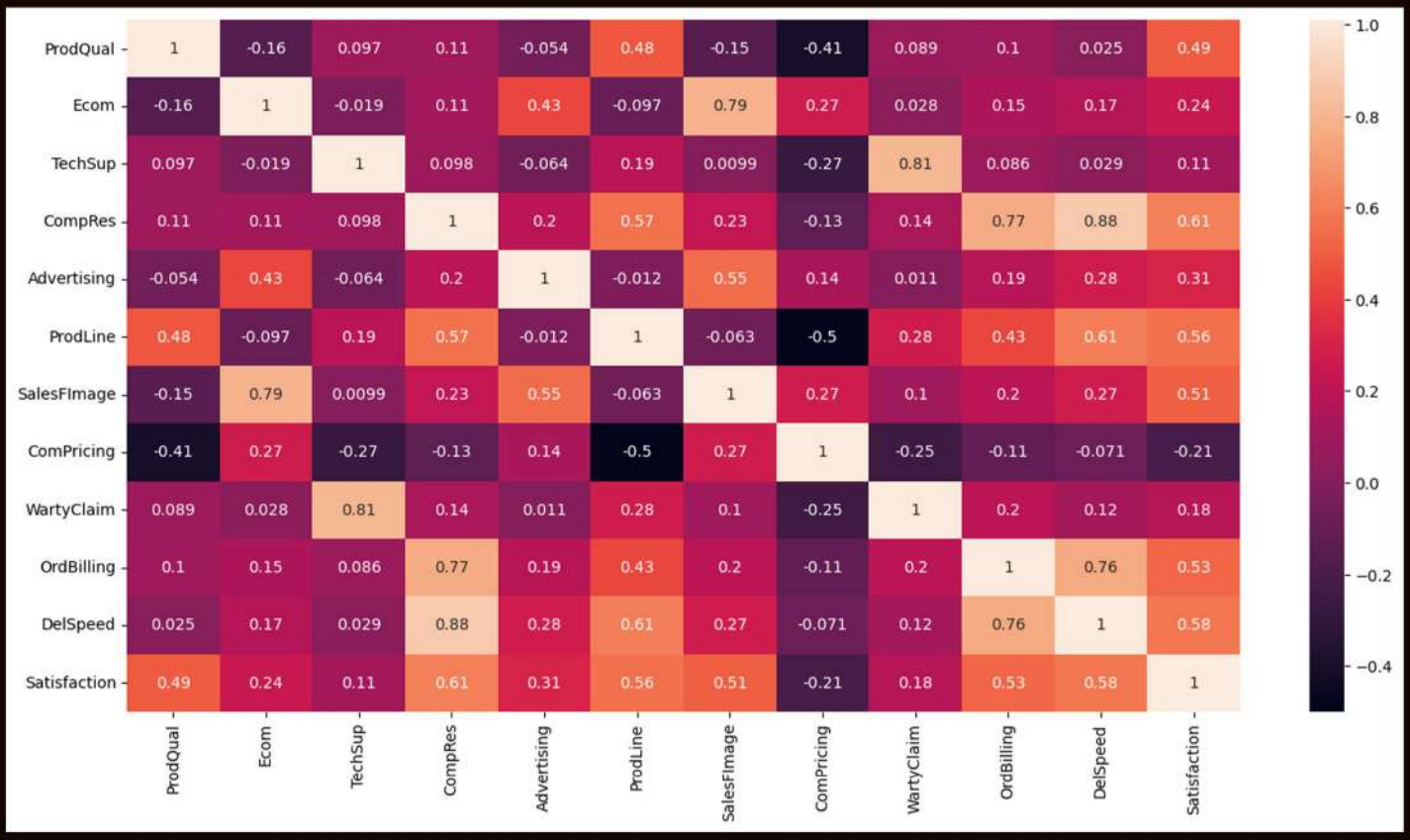
benefit of apply scaling all data point is come in center

most data lie between -3 to +3

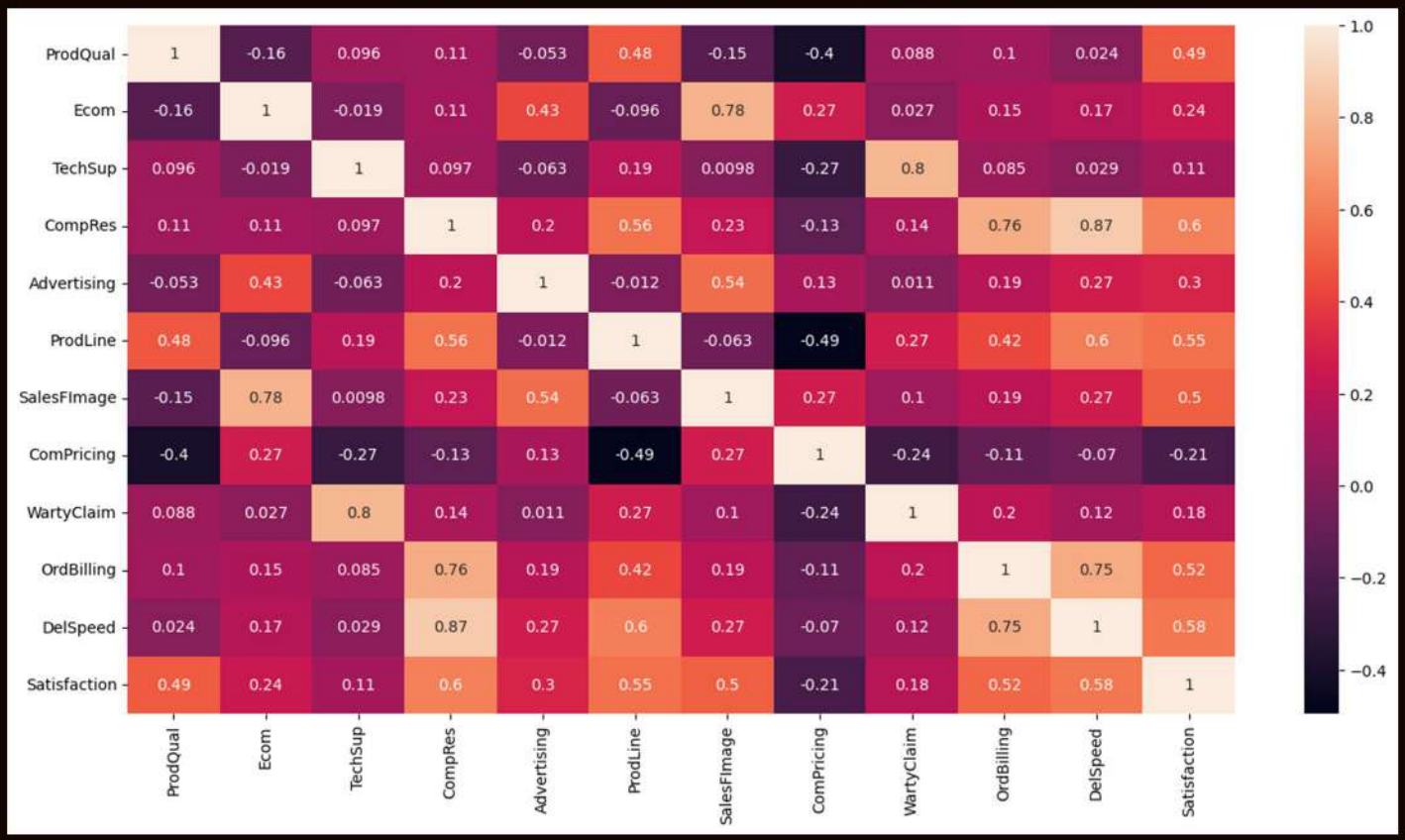
	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
0	0.496660	0.401668	-1.881421	0.380922	0.704543	-0.691530	0.838627	-0.113185	-1.646582	0.791872	-0.260903	1.081067
1	0.280721	-1.495974	-0.174023	1.462141	-0.544014	1.600835	-1.917200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

COMPARISION BETWEEN COVARIANCE AND CORRELATION MATRIX

COVARIANCE MATRIX

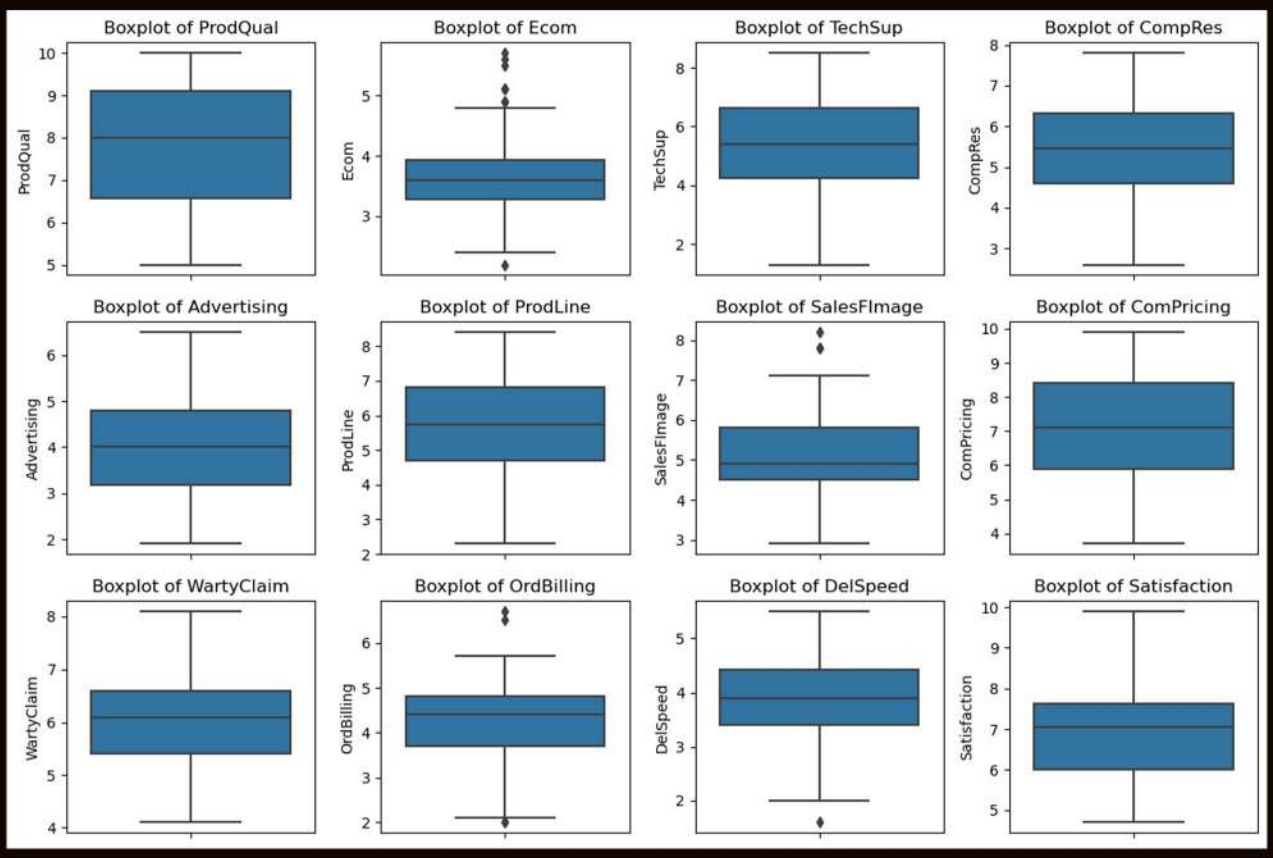


CORRELATION MATRIX

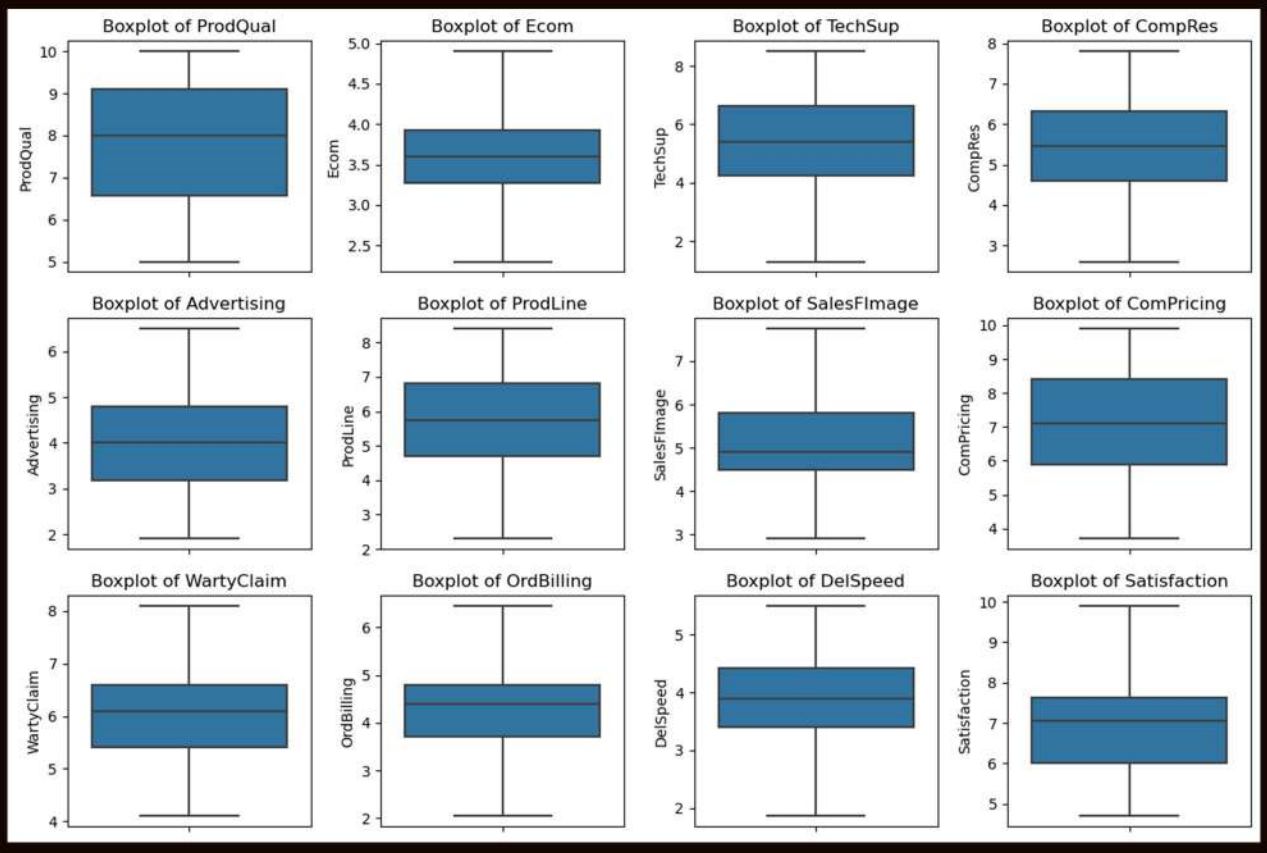


CHECK OUTLIERS BEFORE AND AFTER SCALING

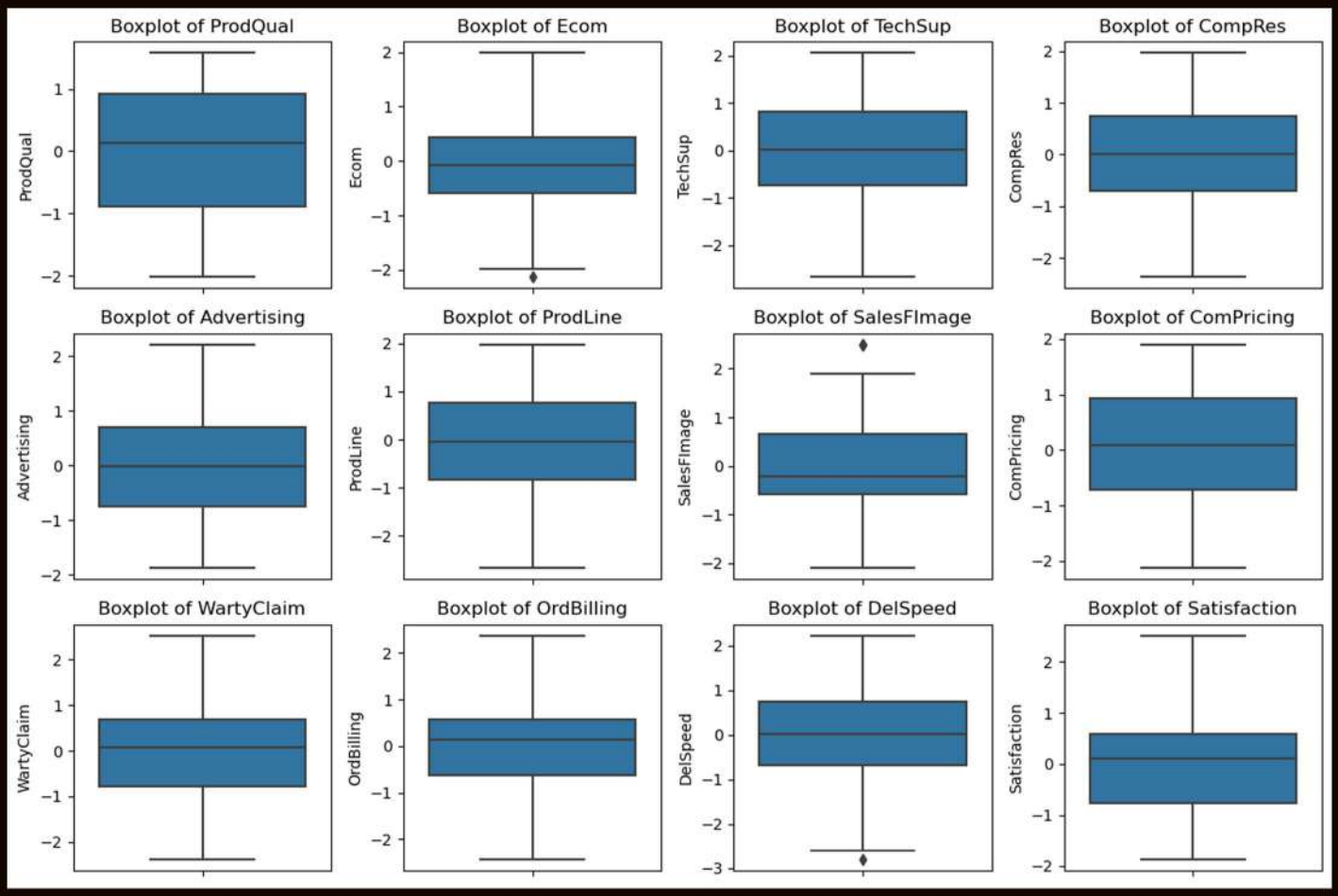
BEFORE SCALING DATA HAD OUTLIERS



BEFORE SCALING DATA HAD REMOVE OUTLIERS



AFTER SCALING DATA



after scaling data mean centering of data data lie between -3 to +3 benefit to do this easy to apply PCA

COVARIANCE MATRIX

EIGEN VECTORS

```
array([[-1.61322427e-01, -1.38992261e-01, -1.27131534e-01,
        -4.25502348e-01, -1.77257763e-01, -3.56524003e-01,
        -2.10387616e-01,  1.37603805e-01, -1.76706227e-01,
        -3.91237634e-01, -4.24958585e-01, -4.13318690e-01],
       [-3.06272359e-01,  4.54921339e-01, -2.35263231e-01,
         8.86065191e-03,  3.55907323e-01, -2.89852601e-01,
         4.64926872e-01,  4.15466718e-01, -1.97843283e-01,
         2.05739475e-02,  6.26377391e-02,  2.95573690e-02],
       [ 7.95045575e-02, -2.29883744e-01, -6.21730460e-01,
         1.91750596e-01, -9.22380787e-02,  1.12809185e-01,
        -2.36626212e-01,  4.49919990e-02, -6.11385841e-01,
         1.42820217e-01,  2.07727869e-01,  3.04020598e-02],
       [ 6.16476615e-01,  1.83792626e-01, -1.66476236e-01,
        -2.79905722e-01,  2.14732458e-01,  9.85304039e-02,
         2.12995164e-01, -2.36864713e-01, -1.75501531e-01,
        -3.03399090e-01, -2.93932094e-01,  3.37012361e-01],
       [-2.56708792e-01, -1.95989018e-01, -4.32018329e-02,
        -3.10014556e-02,  7.63273860e-01,  1.96214033e-02,
        -1.38679963e-01, -4.84289239e-01, -2.28877328e-02,
        -4.96697647e-02,  5.53883726e-02, -2.23746335e-01],
       [ 3.49665681e-01, -4.72109013e-01,  1.18961241e-01,
         2.27476118e-02,  4.10458402e-01, -1.94306175e-01,
        -1.70268215e-01,  6.00687375e-01,  1.37026464e-01,
         7.61903184e-02, -2.66931588e-02,  1.37245917e-01],
       [ 1.59566085e-01,  4.58420598e-02, -1.85228111e-03,
        -5.65413580e-03, -5.50372696e-02, -6.24254550e-01,
        -2.15976650e-02, -3.18607905e-01, -4.42538118e-02,
         6.47750462e-01, -2.33344595e-01,  4.13848752e-02],
       [-3.28835920e-01, -5.09595822e-01,  5.57062806e-02,
         1.36570710e-01, -1.42161086e-01, -2.70924156e-01,
         3.52510892e-01, -1.80376001e-01, -9.00155002e-02,
        -2.79292694e-01, -2.22438761e-02,  5.22889275e-01],
       [-1.68511089e-01, -1.98053391e-01, -5.56287356e-01,
        -4.35955358e-01, -4.16389528e-02,  2.17278275e-01,
         1.58074558e-01,  3.15183851e-02,  5.12637449e-01,
         2.76449344e-01, -7.81797655e-02,  1.12285686e-01],
       [ 2.26630723e-01,  4.24260261e-02, -4.16015115e-01,
         5.64127007e-01, -3.51473948e-02, -2.76430381e-01,
         4.97639173e-02, -9.66456687e-02,  4.51055742e-01,
        -3.26877107e-01, -6.69243234e-03, -2.36050949e-01],
       [ 1.97914841e-01, -2.03297721e-03,  5.88820878e-04,
        -4.18506907e-01, -8.35995268e-02, -3.44414628e-01,
         1.08184495e-02, -1.01667061e-01,  6.24508414e-02,
        -1.49741626e-01,  7.88056217e-01, -4.75225625e-02],
       [-2.30478370e-01,  3.50691037e-01, -1.12144294e-01,
        -1.21358431e-02,  5.50616703e-02, -1.51536732e-01,
        -6.61578652e-01,  1.56701195e-02,  1.59780088e-01,
        -1.50028819e-01, -5.04169815e-03,  5.46974515e-01]])
```

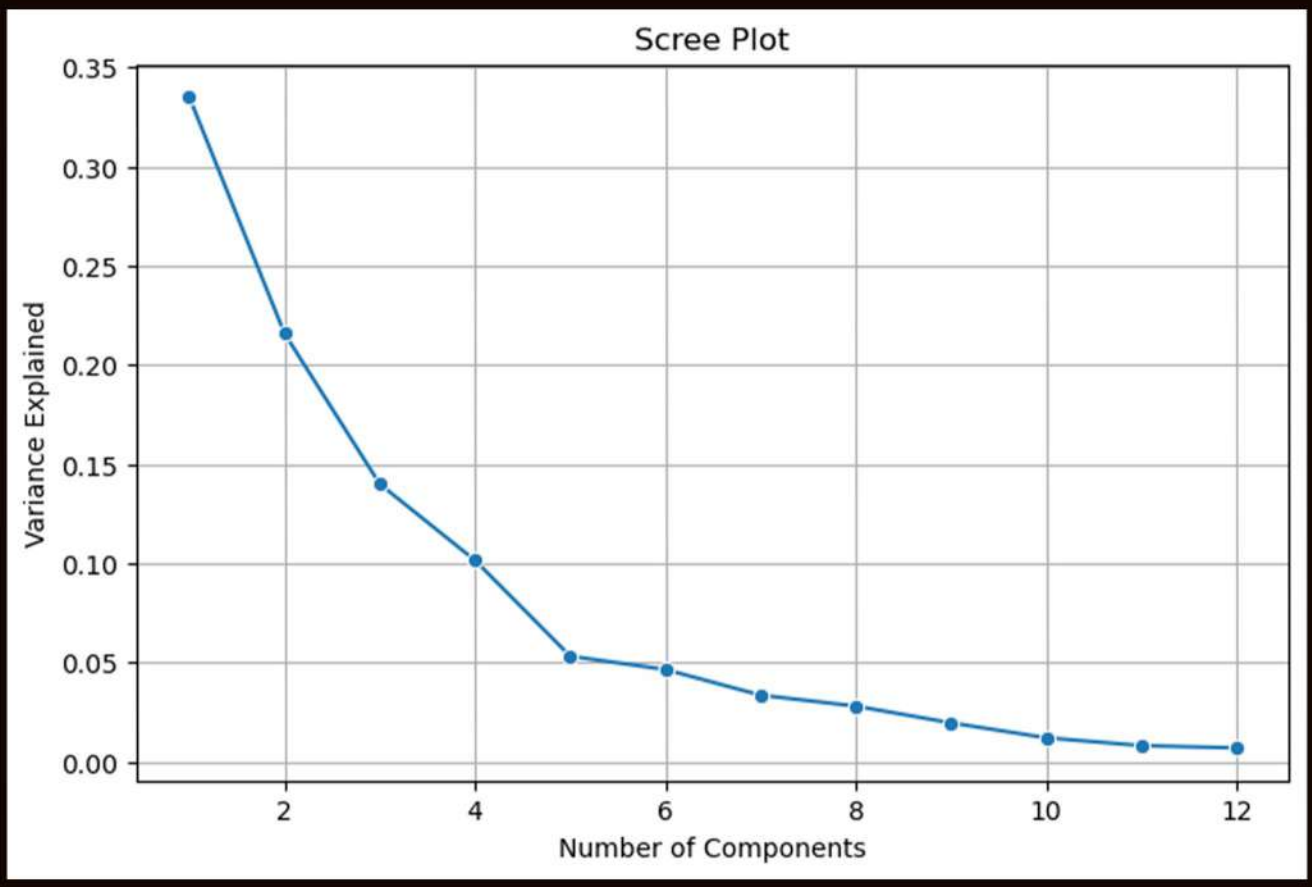
EIGEN VALUES

array([4.06041575, 2.61555054, 1.69615738, 1.22994533, 0.64467056,
0.56308559, 0.40636329, 0.33938181, 0.2372518 , 0.14600259,
0.09819793, 0.08418956])

EXPLICT FIRST PC

array([0.3349843 , 0.21578292, 0.13993298, 0.10147049, 0.05318532,
0.04645456, 0.03352497, 0.027999 , 0.01957327, 0.01204521,
0.00810133, 0.00694564])

FIRST PC COVERED 33% OF DATA



CUMULATIVE VALUES AND EIGEN VALUES

array([0.3349843 , 0.21578292, 0.13993298, 0.10147049, 0.05318532,
0.04645456, 0.03352497, 0.027999 , 0.01957327, 0.01204521,
0.00810133, 0.00694564])

array([0.3349843 , 0.55076722, 0.6907002 , 0.79217069, 0.84535601,
0.89181057, 0.92533555, 0.95333455, 0.97290782, 0.98495303,
0.99305436, 1.])

FIRST ARRAY DENOTE WHICH PC COVERED HOW MUCH DATA

first pc covered 33% of data, second 21% , third 13%, fourth 10%, fifth 5% ,sixth 4%,,seventh 3%, eight 2% ninth 1.9%, tenth 1.2%, elevnth 0.8%, twelth 0.6% approximately

SECOND ARRAY DENOTE TOTAL DATA COVERD STEP BY STEP

first 5 pc coverd 84.5% of data

ITS HELP TO TAKE DECISION WHICH PC HAVE SELECT AND HOW MUCH PC SELECT THOSE ARE COVERD MAXIMUM DATA

EIGEN VECTOR SHOW:

the direction of our main axes (principal components) of our data The greater the eigenvalue, the greater the variation along this axis. So the eigenvector with the largest eigenvalue corresponds to the axis with the most variance

CUMULATIVE VALUES AND EIGEN VALUES

array([0.3349843 , 0.21578292, 0.13993298, 0.10147049, 0.05318532,
0.04645456, 0.03352497, 0.027999 , 0.01957327, 0.01204521,
0.00810133, 0.00694564])

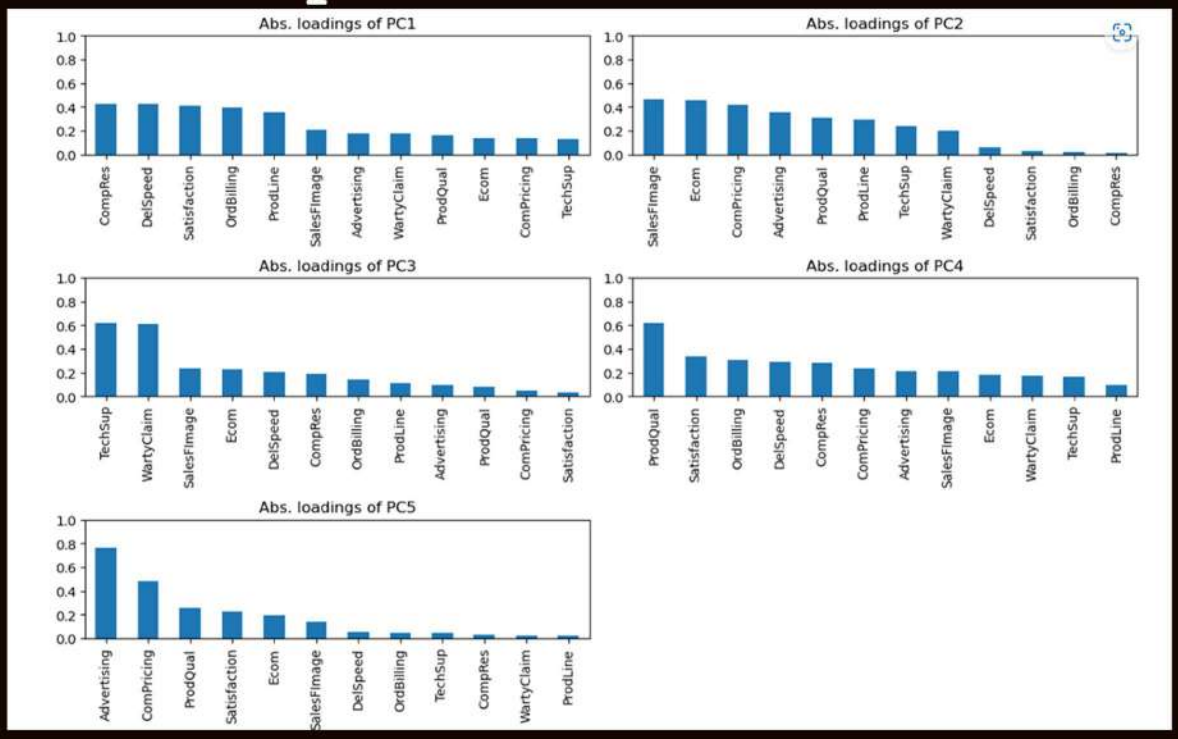
array([0.3349843 , 0.55076722, 0.6907002 , 0.79217069, 0.84535601,
0.89181057, 0.92533555, 0.95333455, 0.97290782, 0.98495303,
0.99305436, 1.])

FIRST ARRAY DENOTE WHICH PC COVERED HOW MUCH DATA

first pc covered 33% of data, second 21% , third 13%, fourth 10%, fifth 5% ,sixth 4%,,seventh 3%, eight 2% ninth 1.9%, tenth 1.2%, elevnth 0.8%, twelth 0.6% approximately

SECOND ARRAY DENOTE TOTAL DATA COVERD STEP BY STEP

first 5 pc covered 84.5% of data



THE SUM OF TOTAL SECOND ARRAY IS 1 BECAUSE WE ASSUME TOTAL VARIANCE OF DATA IS 1

its explained ratio of variance

EIGEN VECTOR SHOW:

the direction of our main axes (principal components) of our data The greater the eigenvalue, the greater the variation along this axis. So the eigenvector with the largest eigenvalue corresponds to the axis with the most variance

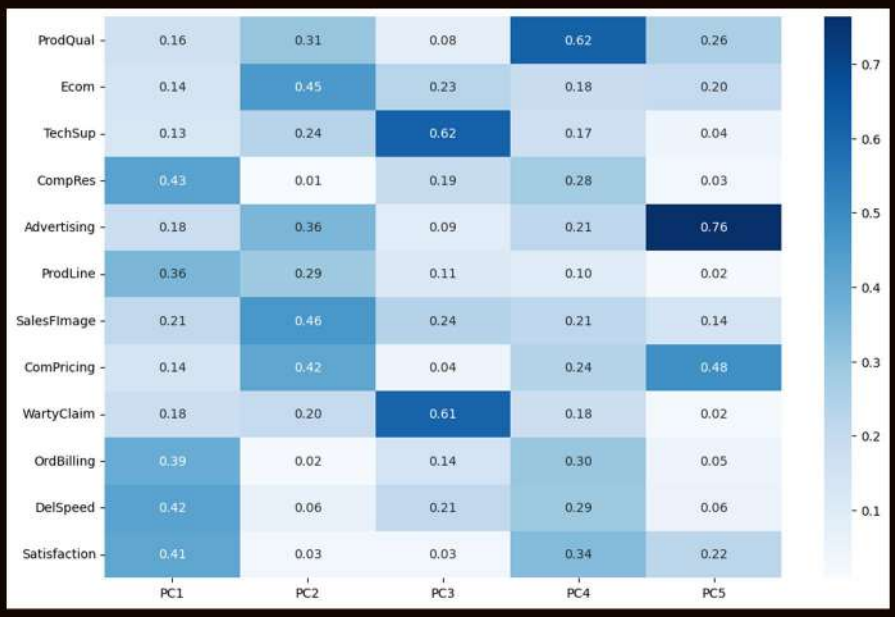
ITS HELP TO TAKE DECISION WHICH PC HAVE SELECT AND HOW MUCH PC SELECT THOSE ARE COVERD MAXIMUM DATA

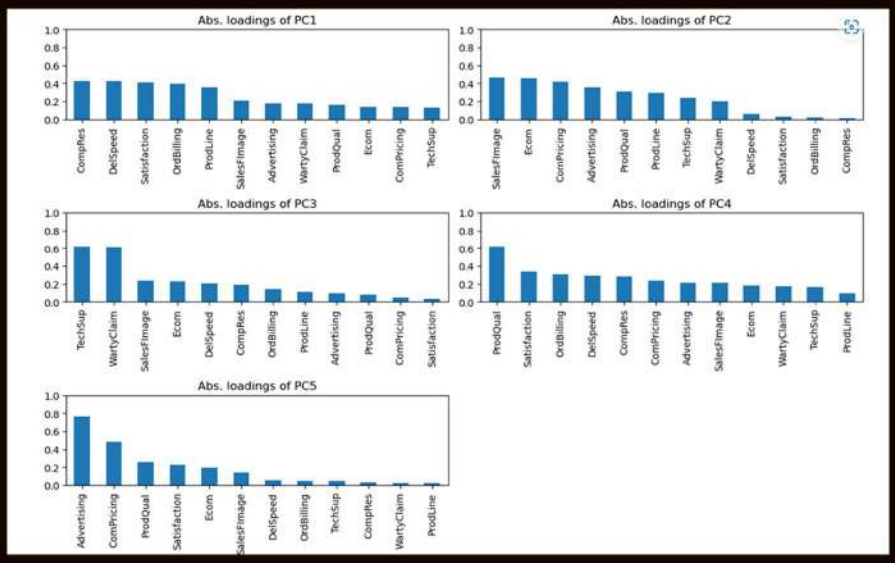
THE SUM OF TOTAL SECOND ARRAY IS 1 BECAUSE WE ASSUME TOTAL VARIANCE OF DATA IS 1

data of the Principal Component scores into a data frame:

	PC1	PC2	PC3	PC4	PC5
ProdQual	-0.161322	-0.306272	0.079505	0.616477	-0.256709
Ecom	-0.138992	0.454921	-0.229884	0.183793	-0.195989
TechSup	-0.127132	-0.235263	-0.621730	-0.166476	-0.043202
CompRes	-0.425502	0.008861	0.191751	-0.279906	-0.031001
Advertising	-0.177258	0.355907	-0.092238	0.214732	0.763274
ProdLine	-0.356524	-0.289853	0.112809	0.098530	0.019621
SalesFImage	-0.210388	0.464927	-0.236626	0.212995	-0.138680
ComPricing	0.137604	0.415467	0.044992	-0.236865	-0.484289
WartyClaim	-0.176706	-0.197843	-0.611386	-0.175502	-0.022888
OrdBilling	-0.391238	0.020574	0.142820	-0.303399	-0.049670
DelSpeed	-0.424959	0.062638	0.207728	-0.293932	0.055388
Satisfaction	-0.413319	0.029557	0.030402	0.337012	-0.223746

BUSINESS IMPLICATION





first pc covered 33% of data, second 21% , third 13%, fourth 10%, fifth 5% of data

first 5 pc covered 84.5% of data

according to first pca(compres,ordbilling,prodline,delspeed, satisfaction) is more implication of bussiness and other approx implict 15% in bussiness and first pc cover 33% of data

according to first 4 pca take (ecom,proqua,salesflmage,wart claim) is more implication of bussiness comapre to other

according to first 4 pca is covered approx 80% of data

CLUSTERING

There are 297 rows and 7 columns present in dataset

(States,Health_indeces1,Health_indices2 ,Per_capita_income,,GDP)
is int64 column and (States) is object column and 1 column is (unnamed)
No null and no duplicated value present in dataset

(unnamed,states) column are drop because in clustering there no use of this column

mean,std,min.max,25%,50%,75% of data shown as:

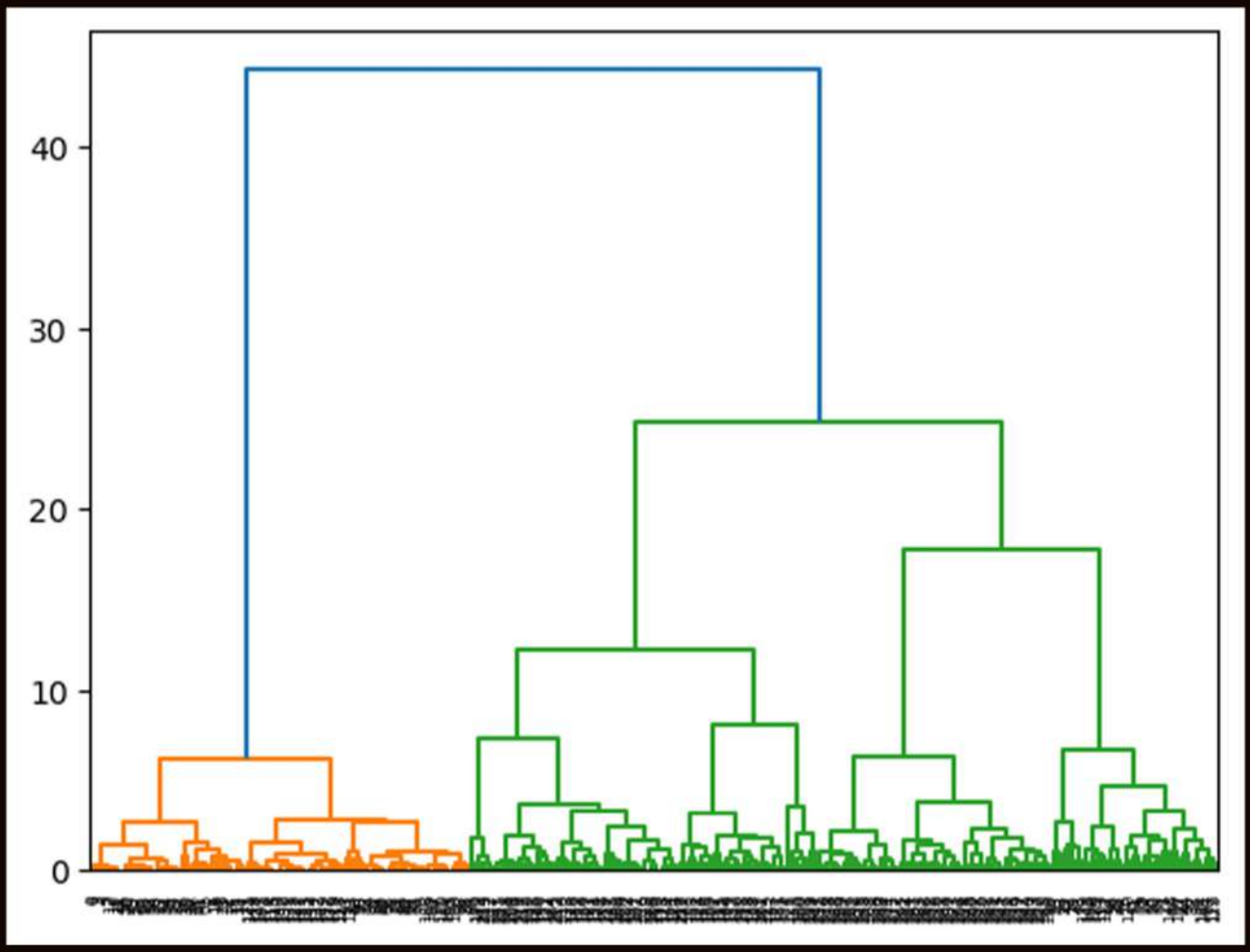
	count	mean	std	min	25%	50%	75%	max
Health_indeces1	297.0	2630.151515	2038.505431	-10.0	641.0	2451.0	4094.0	10219.0
Health_indices2	297.0	693.632997	468.944354	0.0	175.0	810.0	1073.0	1508.0
Per_capita_income	297.0	2156.915825	1491.854058	500.0	751.0	1865.0	3137.0	7049.0
GDP	297.0	174601.117845	167167.992863	22.0	8721.0	137173.0	313092.0	728575.0

SCALING THE DATA

scaling the data is pre processing work it is neccesay to do because it is mean centering of data which means (mean of data is 0 and SD is 1) and data lie between -3 to +3 this proccess is known as (standardiation)

```
array([[-1.72622877, -1.08750595, -1.34065436, -1.06954449, -1.03530421],
       [-1.71456506, -0.56270846, -0.10174638,  0.37136183, -0.60483755],
       [-1.70290136, -0.97104808, -0.84295512, -0.70696788, -0.88253555],
       ...,
       [ 1.70290136,  0.31735924, -0.18078016, -0.42630672, -0.32607295],
       [ 1.71456506,  0.40924793,  0.32759932, -0.82111237,  0.14861542],
       [ 1.72622877, -0.27524917,  0.30837488,  0.68626633, -0.04694284]])
```

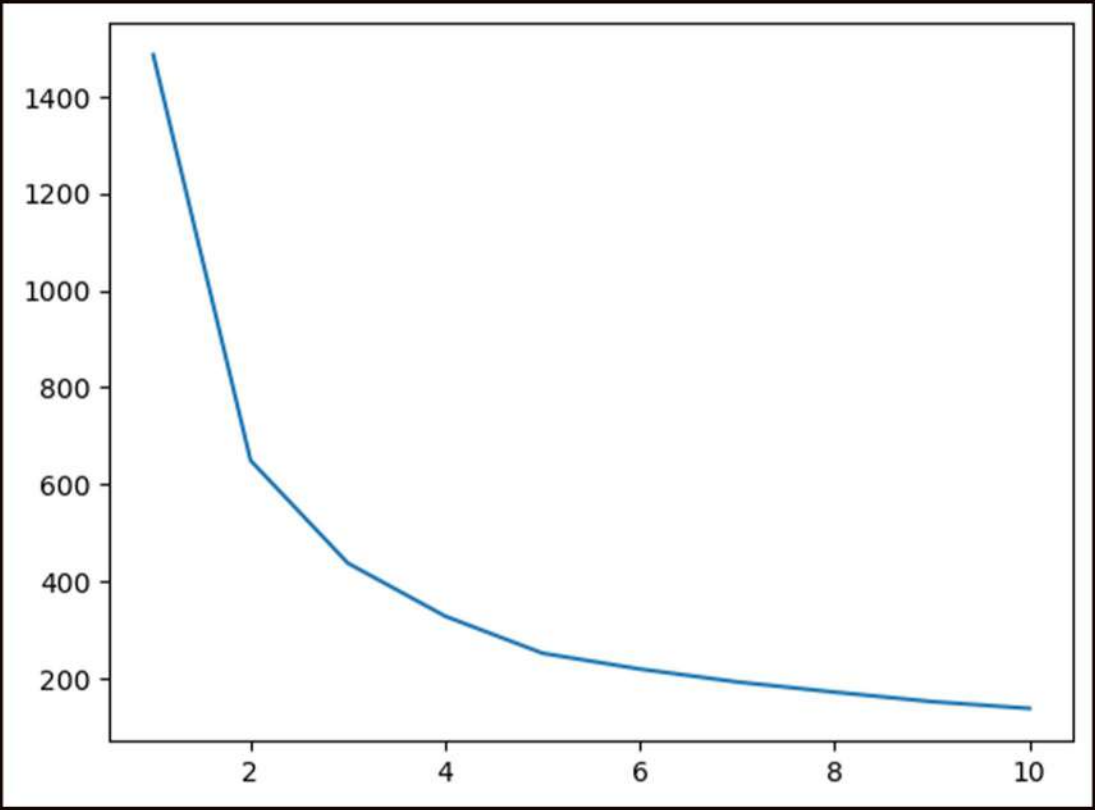
HIERARCHICAL CLUSTERING



2 clusters will be formed in this graph(data) because according hierarchical clustering search who vertical line to all most bari line then cut the verical line

look at this graph blue vertical line is most bari line to all so cut the line it will formed two groups(cluster) green and orange, its mean in this data 2 cluster formed

K-MEAN CLUSTERING



elbow curve method generally used to find no cluster in dataset in this graph shown blue line bend coming in 3 point , its mean 3 should be made 3 cluster in this dataset

silhouette score

this indicate more than value 0.5 well distinguish cluster

```
silhouette_score(scaled_df,labels)
```

```
0.5282460986891512
```

DESCRIBE CLUSTER PROFILE

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

```
In [123]: df_clu[df_clu["Clus"]==1].sum()

Out[123]: Unnamed: 0                                                  8424
          States              BachevoBalgarchevoBelasitsaBelo_PoleBeslenBogo...
          Health_indeces1                                            86601
          Health_indices2                                            21819
          Per_capita_income                                          99439
          GDP                                                      2719290
          Clus                                                         116
          dtype: object

In [124]: df_clu[df_clu["Clus"]==2].sum()

Out[124]: Unnamed: 0                                                 16893
          States              BuchinoDoleneFargovoKolibiteKribulMihnevoObidi...
          Health_indeces1                                           479099
          Health_indices2                                           118218
          Per_capita_income                                         331503
          GDP                                                     37254378
          Clus                                                        196
          dtype: object

In [126]: df_clu[df_clu["Clus"]==0].sum()

Out[126]: Unnamed: 0                                                 18639
          States              Gorna_BreznitsaKalimantsiMendovoPravo_BardoRib...
          Health_indeces1                                           215455
          Health_indices2                                            65972
          Per_capita_income                                         209662
          GDP                                                     11882864
          Clus                                                           0
          dtype: object
```

> THERE ARE 3 CLUSTER IN DATASET

Its is a unsupervised learing, so cluster is nothing it is a groups of rows and column are present in dataset...which mean data look like same - same so this data group in make 1 cluster..

through clusters is easy to analysis the data. it is a very powerful algortithm to analysis data.

FOR EXAMPLE

more column in data you cannot vizulize this which data is lie between them...clustering is help to this analysis and in making a decision