DATA MINING

Name: Vamsi Krishna Bommu

Batch: PGPDSBA.APRIL23.B

CONTENTS

Part 1:

Exploratory Data Analysis:

- This dataset includes various attributes like ID, Product Quality, E-Commerce, Technical Support, Complaint Resolution, Advertising, Product Line, Salesforce Image, Competitive Pricing, Warranty & Claims, Order & Billing, Delivery Speed and 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

- Data is used for descriptive analysis.description() returns a statistical overview of numerical columns, including the mean, median, standard deviation, and distribution range of each numeri property.

	ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Sati
count	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.00000	100.000000	100.000000	100.00000	100.000000	10(
mean	50.500000	7.810000	3.672000	5.365000	5.442000	4.010000	5.805000	5.12300	6.974000	6.043000	4.27800	3.886000	6
std	29.011492	1.396279	0.700516	1.530457	1.208403	1.126943	1.315285	1.07232	1.545055	0.819738	0.92884	0.734437	1
min	1.000000	5.000000	2.200000	1.300000	2.600000	1.900000	2.300000	2.90000	3.700000	4.100000	2.00000	1.600000	4
25%	25.750000	6.575000	3.275000	4.250000	4.600000	3.175000	4.700000	4.50000	5.875000	5.400000	3.70000	3.400000	6
50%	50.500000	8.000000	3.600000	5.400000	5.450000	4.000000	5.750000	4.90000	7.100000	6.100000	4.40000	3.900000	7
75%	75.250000	9.100000	3.925000	6.625000	6.325000	4.800000	6.800000	5.80000	8.400000	6.600000	4.80000	4.425000	7
max	100.000000	10.000000	5.700000	8.500000	7.800000	6.500000	8.400000	8.20000	9.900000	8.100000	6.70000	5.500000	9

- To inspect the dataset shape, use the shape function to determine the number of columns and rows.

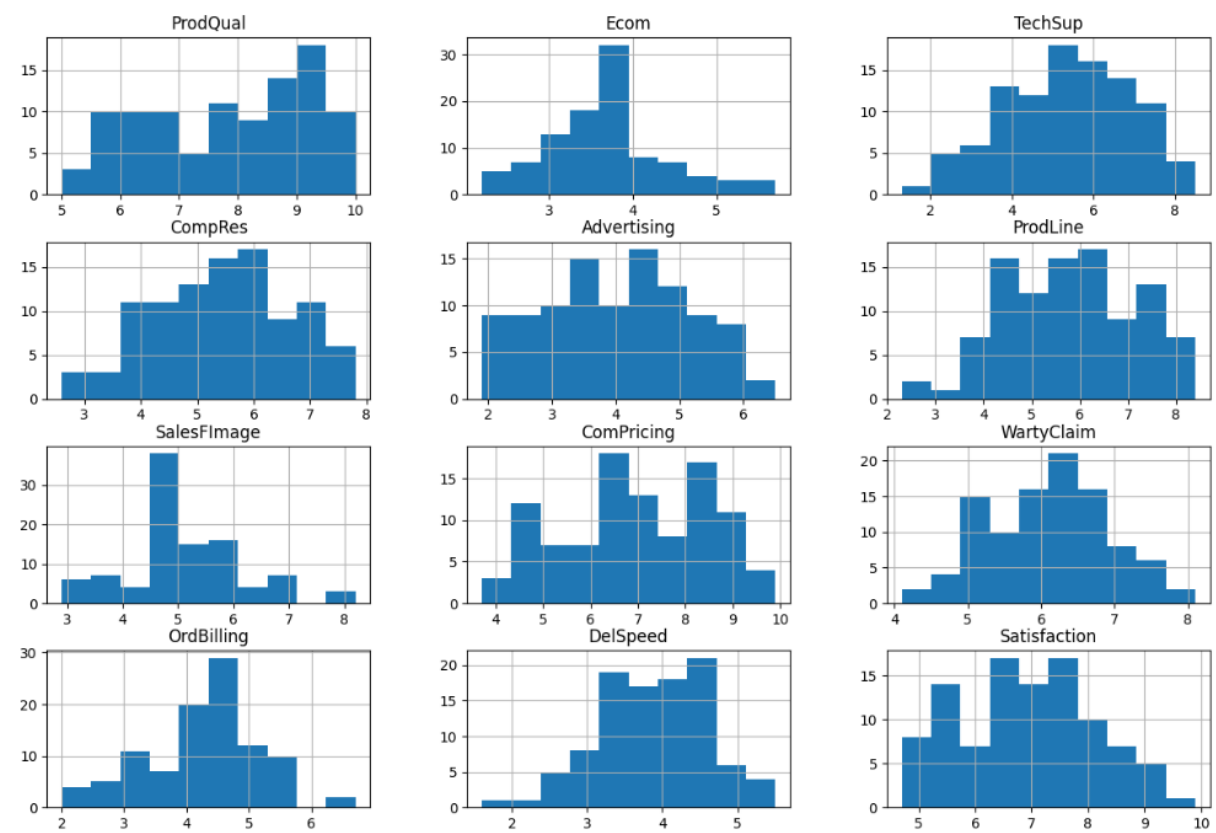
```
df.shape
```

```
(100, 13)
```

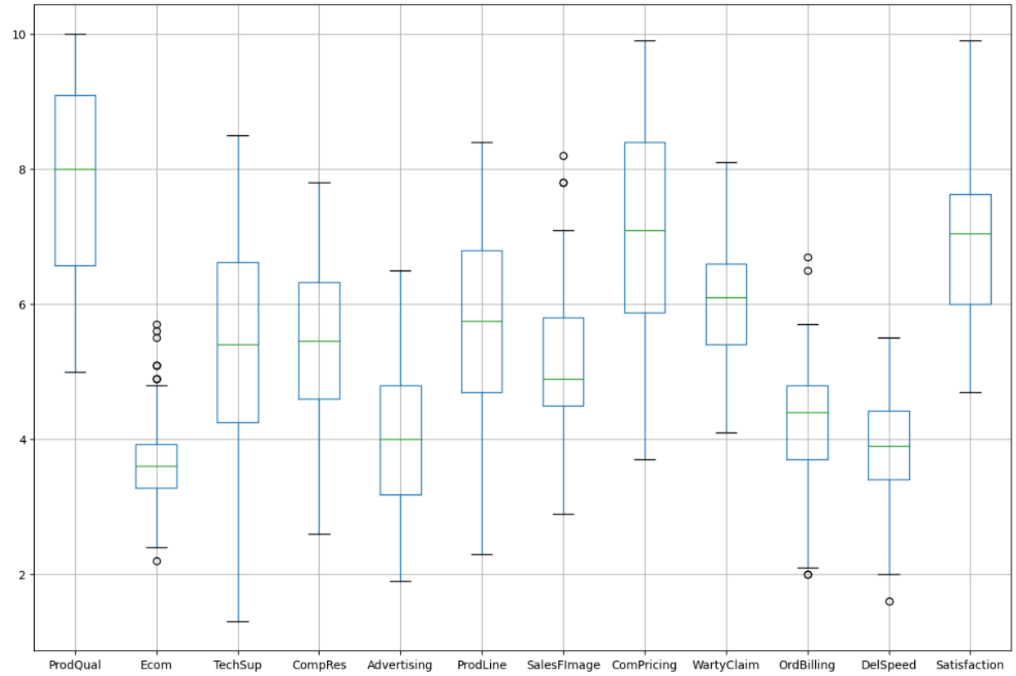
- If null values are presented in the dataset, then we use isnull() function to find the null values in each column.

```
ID              0
ProdQual        0
Ecom            0
TechSup         0
CompRes         0
Advertising     0
ProdLine        0
SalesFImage     0
ComPricing      0
WartyClaim      0
OrdBilling      0
DelSpeed        0
Satisfaction    0
dtype: int64
```

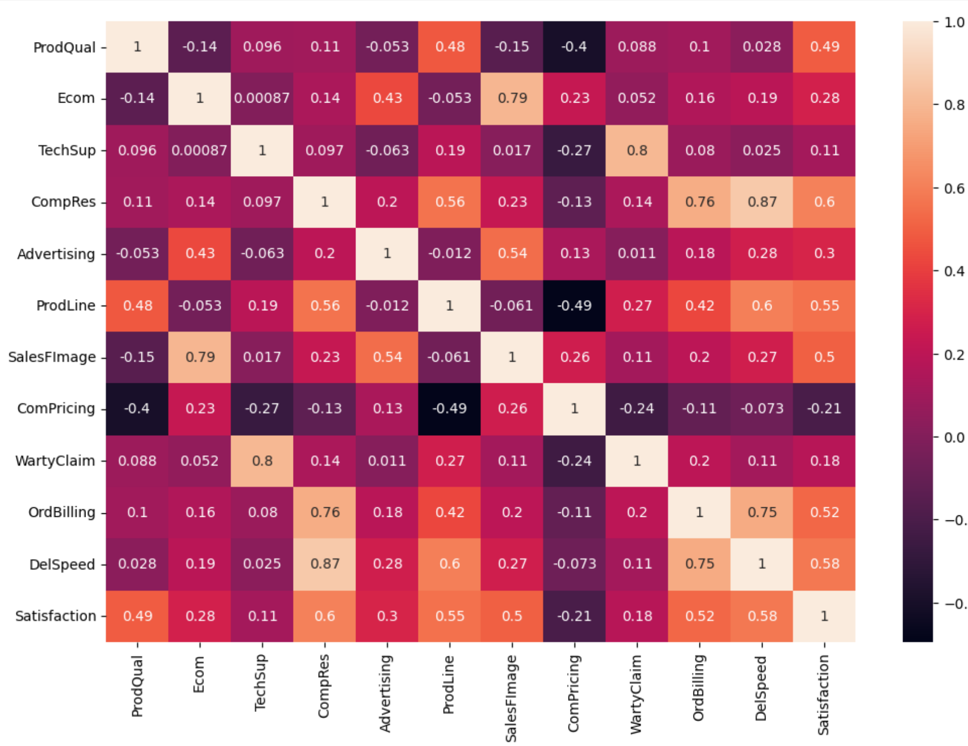
- Univariate Analysis:
    - Histograms represent the distribution of each variable. They suggest that some variables may have skewed distributions.



    - Boxplots show the distribution and central tendency of each variable. They reveal the presence of outliers in certain variables.



- Multivariate Analysis:
  - Correlation The heatmap shows the correlation between variables. It aids in the identification of potential correlations among variables. Strong correlations imply possible multicollinearity, but weak correlations may indicate independence.



PCA Scaling Variables

- For this case study, the variables were scaled with StandardScaler. StandardScaler was chosen because it normalizes features by removing the mean and scaling to unit variance. This scaling is useful for variables with multiple units or scales, ensuring that all variables contribute equally to the study.

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	
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	1
mean	9.188483e-16	1.011413e-15	1.029177e-15	-1.432188e-16	-6.061818e-16	2.531308e-16	6.178391e-16	-7.105427e-16	-1.247891e-15	4.751755e-16	4
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	1
min	-2.022630e+00	-2.111893e+00	-2.669451e+00	-2.363712e+00	-1.881755e+00	-2.678246e+00	-2.083519e+00	-2.129693e+00	-2.382210e+00	-2.464877e+00	-3
25%	-8.889494e-01	-5.695798e-01	-7.322109e-01	-7.002976e-01	-7.446754e-01	-8.443545e-01	-5.839103e-01	-7.148848e-01	-7.883484e-01	-6.254166e-01	-6
50%	1.367614e-01	-1.032991e-01	2.298420e-02	6.653659e-03	-8.918268e-03	-4.202669e-02	-2.090080e-01	8.196131e-02	6.988470e-02	1.320083e-01	1
75%	9.285383e-01	3.629816e-01	8.274312e-01	7.343976e-01	7.045432e-01	7.603011e-01	6.345221e-01	9.275939e-01	6.829084e-01	5.648226e-01	7
max	1.576356e+00	2.909592e+00	2.058728e+00	1.961166e+00	2.220649e+00	1.982896e+00	2.883936e+00	1.903324e+00	2.521979e+00	2.620690e+00	2

PCA Comparison between Covariance and Correlation Matrix after Scaling

- The Covariance Matrix reflects the relationship between variables in absolute values, regardless of unit of measurement.

```
Covariance Matrix:
 [[ 1.01010101e+00 -1.38548704e-01  9.65661154e-02  1.07444445e-01
  -5.40132667e-02  4.82316579e-01 -1.53346338e-01 -4.05335236e-01
   8.92043497e-02  1.05356640e-01  2.79979825e-02  4.91237372e-01]
 [-1.38548704e-01  1.01010101e+00  8.75544162e-04  1.41595213e-01
   4.34233041e-01 -5.32200387e-02  7.99539102e-01  2.31780203e-01
   5.24224157e-02  1.57724577e-01  1.93571786e-01  2.85601025e-01]
 [ 9.65661154e-02  8.75544162e-04  1.01010101e+00  9.76329270e-02
  -6.35051180e-02  1.94571168e-01  1.71621612e-02 -2.73521901e-01
   8.05220127e-01  8.09109340e-02  2.56976702e-02  1.13734524e-01]
 [ 1.07444445e-01  1.41595213e-01  9.76329270e-02  1.01010101e+00
   1.98905906e-01  5.67087831e-01  2.32072486e-01 -1.29246720e-01
   1.41826562e-01  7.64513729e-01  8.73829997e-01  6.09356166e-01]
 [-5.40132667e-02  4.34233041e-01 -6.35051180e-02  1.98905906e-01
   1.01010101e+00 -1.16674936e-02  5.47680463e-01  1.35572620e-01
   1.09010852e-02  1.86096560e-01  2.78649579e-01  3.07746944e-01]
 [ 4.82316579e-01 -5.32200387e-02  1.94571168e-01  5.67087831e-01
  -1.16674936e-02  1.01010101e+00 -6.19348764e-02 -4.99947880e-01
   2.75835887e-01  4.28695202e-01  6.07929503e-01  5.56107006e-01]
 [-1.53346338e-01  7.99539102e-01  1.71621612e-02  2.32072486e-01
   5.47680463e-01 -6.19348764e-02  1.01010101e+00  2.67269246e-01
   1.08540752e-01  1.97098390e-01  2.74294201e-01  5.05257885e-01]
 [-4.05335236e-01  2.31780203e-01 -2.73521901e-01 -1.29246720e-01
   1.35572620e-01 -4.99947880e-01  2.67269246e-01  1.01010101e+00
  -2.47460661e-01 -1.15724268e-01 -7.36078070e-02 -2.10399686e-01]
 [ 8.92043497e-02  5.24224157e-02  8.05220127e-01  1.41826562e-01
   1.09010852e-02  2.75835887e-01  1.08540752e-01 -2.47460661e-01
   1.01010101e+00  1.99055678e-01  1.10499598e-01  1.79338201e-01]
 [ 1.05356640e-01  1.57724577e-01  8.09109340e-02  7.64513729e-01
   1.86096560e-01  4.28695202e-01  1.97098390e-01 -1.15724268e-01
   1.99055678e-01  1.01010101e+00  7.58588957e-01  5.27001932e-01]
 [ 2.79979825e-02  1.93571786e-01  2.56976702e-02  8.73829997e-01
   2.78649579e-01  6.07929503e-01  2.74294201e-01 -7.36078070e-02
   1.10499598e-01  7.58588957e-01  1.01010101e+00  5.82870984e-01]
 [ 4.91237372e-01  2.85601025e-01  1.13734524e-01  6.09356166e-01
   3.07746944e-01  5.56107006e-01  5.05257885e-01 -2.10399686e-01
   1.79338201e-01  5.27001932e-01  5.82870984e-01  1.01010101e+00]]
```

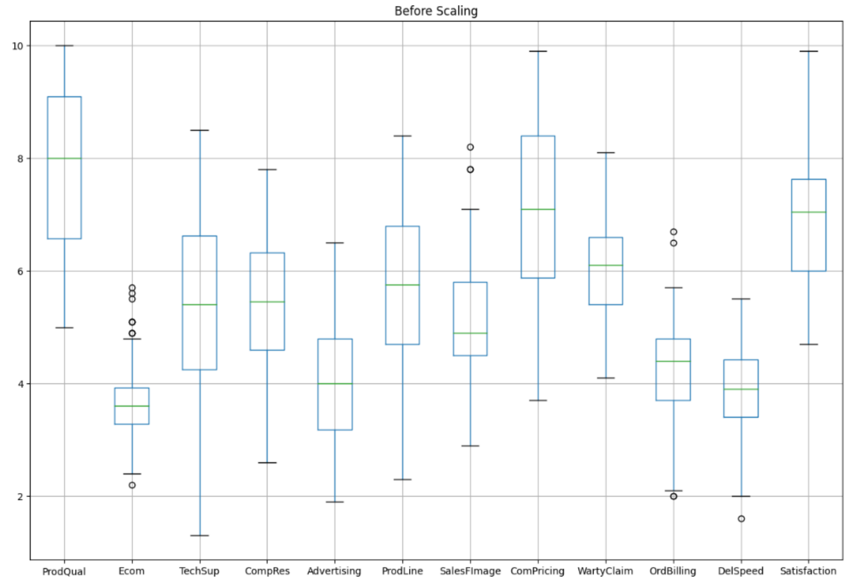
- Correlation Matrix: Shows the strength and direction of a linear relationship between pairs of variables, normalized to [-1, 1].

```
Correlation matrix:
              ProdQual      Ecom   TechSup   CompRes  Advertising  ProdLine  \
ProdQual      1.000000 -0.137163  0.095600  0.106370    -0.053473  0.477493
Ecom         -0.137163  1.000000  0.000867  0.140179     0.429891 -0.052688
TechSup       0.095600  0.000867  1.000000  0.096657    -0.062870  0.192625
CompRes       0.106370  0.140179  0.096657  1.000000     0.196917  0.561417
Advertising  -0.053473  0.429891 -0.062870  0.196917     1.000000 -0.011551
ProdLine      0.477493 -0.052688  0.192625  0.561417    -0.011551  1.000000
SalesFImage  -0.151813  0.791544  0.016991  0.229752     0.542204 -0.061316
ComPricing   -0.401282  0.229462 -0.270787 -0.127954     0.134217 -0.494948
WartyClaim    0.088312  0.051898  0.797168  0.140408     0.010792  0.273078
OrdBilling    0.104303  0.156147  0.080102  0.756869     0.184236  0.424408
DelSpeed      0.027718  0.191636  0.025441  0.865092     0.275863  0.601850
Satisfaction  0.486325  0.282745  0.112597  0.603263     0.304669  0.550546

              SalesFImage  ComPricing  WartyClaim  OrdBilling  DelSpeed  \
ProdQual        -0.151813   -0.401282    0.088312    0.104303  0.027718
Ecom             0.791544    0.229462    0.051898    0.156147  0.191636
TechSup          0.016991   -0.270787    0.797168    0.080102  0.025441
CompRes          0.229752   -0.127954    0.140408    0.756869  0.865092
Advertising      0.542204    0.134217    0.010792    0.184236  0.275863
ProdLine        -0.061316   -0.494948    0.273078    0.424408  0.601850
SalesFImage      1.000000    0.264597    0.107455    0.195127  0.271551
ComPricing       0.264597    1.000000   -0.244986   -0.114567 -0.072872
WartyClaim       0.107455   -0.244986    1.000000    0.197065  0.109395
OrdBilling       0.195127   -0.114567    0.197065    1.000000  0.751003
DelSpeed         0.271551   -0.072872    0.109395    0.751003  1.000000
Satisfaction     0.500205   -0.208296    0.177545    0.521732  0.577042

              Satisfaction
ProdQual          0.486325
Ecom              0.282745
TechSup           0.112597
CompRes           0.603263
Advertising       0.304669
ProdLine          0.550546
SalesFImage       0.500205
ComPricing       -0.208296
WartyClaim        0.177545
OrdBilling        0.521732
DelSpeed          0.577042
Satisfaction      1.000000
```

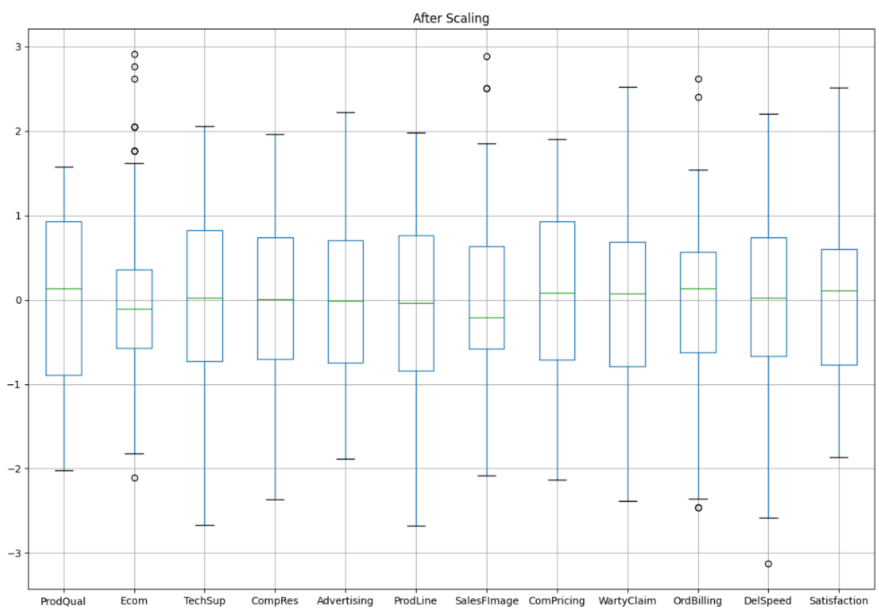
- After scaling, both matrices should have comparable patterns. However, the correlation matrix is recommended when variables are measured in different units because it standardizes the results.

PCA Checking Outliers Before and After Scaling

- Before Scaling: Prior to scaling, outliers in some variables may influence PCA performance.



- After Scaling: Scaling reduces the impact of outliers, resulting in a more uniform spread of data. However, outliers may still exist, especially in variables with extreme values.



PCA Building Covariance Matrix, Eigenvalues, and Eigenvectors

- Covariance Matrix: Captures the variance between pairs of variables, providing insights into their relationships.
- Eigenvalues: Represent the variance explained by each principal component.
- Eigenvectors: Indicate the direction or pattern of the variance captured by each principal component.

```
Eigenvalues:
 [4.08369694 2.57871152 1.70931735 1.22984483 0.6423868  0.57427406
 0.40689671 0.32775774 0.23852472 0.14568036 0.08398124 0.10013985]
Eigenvectors:
 [[ 0.15855116 -0.31313152 -0.07356137 -0.61407082 -0.24964531  0.36499541
  -0.12640774 -0.32687751 -0.18602426  0.2037033   0.21787575  0.22885317]
 [ 0.1661857   0.44059261  0.23651951 -0.19628244 -0.18886909 -0.46540483
  -0.00824784 -0.50785197 -0.21574952  0.03718659 -0.35323725 -0.02881148]
 [ 0.12514332 -0.23828985  0.61631236  0.17941402 -0.03977108  0.12392836
   0.01346077  0.08182818 -0.54753081 -0.42475155  0.10580091 -0.01766533]
 [ 0.42263337  0.00134121 -0.19665426  0.27970497 -0.03340857  0.01495235
   0.00463818  0.14929932 -0.43697539  0.58601845  0.05627641 -0.37853377]
 [ 0.1807615   0.35724531  0.0898675  -0.20600014  0.76107633  0.4189084
   0.07155058 -0.12282896 -0.04176506 -0.02836138 -0.04824083 -0.0968768 ]
 [ 0.35283874 -0.29778667 -0.11122737 -0.10008828  0.0250607  -0.1958228
   0.63397913 -0.22319134  0.23246141 -0.25391841  0.18600871 -0.34728677]
 [ 0.21794995  0.46488879  0.2409419  -0.19948826 -0.14209236 -0.16711795
  -0.02165026  0.33410983  0.1703657   0.03993494  0.66500583  0.07388433]
 [-0.13483701  0.41776317 -0.0516667   0.24079483 -0.4896484   0.58557549
   0.34280528 -0.16338764  0.02851369 -0.08642644 -0.01139137 -0.10660117]
 [ 0.17499123 -0.2011842   0.60545958  0.18959933 -0.02158615  0.1422959
   0.04011921 -0.10701582  0.50449856  0.45392345 -0.15868264  0.0827785 ]
 [ 0.38797945  0.00906156 -0.15503653  0.30668573 -0.04908379  0.09117472
  -0.62874224 -0.33498375  0.25197455 -0.32105097  0.14716762 -0.15754746]
 [ 0.4223407   0.05445737 -0.21799023  0.28990302  0.06222027 -0.03060577
   0.23692774 -0.00146402 -0.07544805 -0.05793177 -0.06069937  0.78321219]
 [ 0.41302455  0.02390379 -0.02873859 -0.33118987 -0.22967423  0.14296626
  -0.07520655  0.52854183  0.13706617 -0.21557147 -0.53252314 -0.10623326]]
```

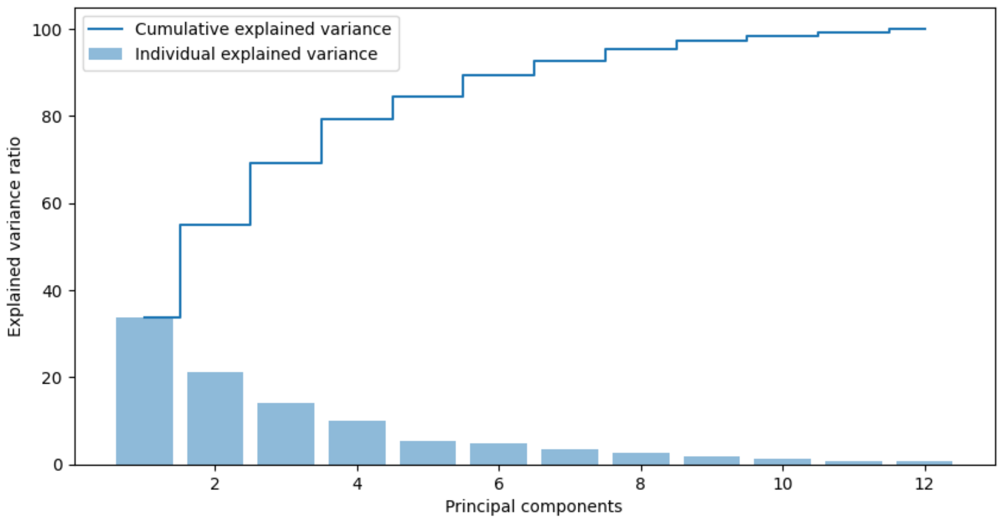
PCA First Principal Component

- The explicit form of the first principal component can be expressed as a linear combination of the original variables using the eigenvectors corresponding to the largest eigenvalue.

```
Eigenvector of the first PC:
 [ 0.15855116  0.1661857   0.12514332  0.42263337  0.1807615   0.35283874
  0.21794995 -0.13483701  0.17499123  0.38797945  0.4223407   0.41302455]
```

PCA Cumulative Values of Eigenvalues

- Help in Optimum Component Selection: Cumulative values indicate the proportion of total variance explained by each principal component.
- Decision-Making: Optimal number of principal components can be determined based on the cumulative explained variance. Typically, a threshold (e.g., 80-90%) is set, and components contributing to that threshold are selected.
- Eigenvectors: They indicate the direction of maximum variance in the data.



PCA Business Implications

- Dimensionality Reduction: PCA reduces the number of variables while retaining most of the information. This simplifies analysis and visualization.
- Insights into Data Structure: PCA identifies patterns and relationships within the data, aiding in decision-making.
- Improved Model Performance: Reduced dimensionality can lead to more efficient modeling, especially when dealing with multicollinearity or high-dimensional data.
- Feature Engineering: PCA can help in feature engineering by creating new variables that capture the most important information from the original features. This can improve the performance of machine learning models.
- By applying PCA, the hair salon can gain insights into customer behavior, preferences, and trends, enabling targeted marketing strategies, resource allocation, and service customization.

	PC1	PC2
0	-0.490389	1.580229
1	-0.495644	-2.485075
2	-2.727909	-0.761250
3	2.236864	0.176334
4	0.644061	-1.392029
...	...	...
95	-0.424050	0.138239
96	1.630872	0.975224
97	3.417553	-1.765533
98	-0.483486	2.318799
99	1.628277	1.309759

100 rows × 2 columns

Part 2:

Data Overview:

- Dataset: State-wise data on health indices, per capita income, and GDP.

	ID	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

- Shape: The dataset has 297 rows and 6 columns.

```
data.shape
```

```
(297, 6)
```

- Missing Values: There are no missing values in the dataset.

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

- Describe: To view the data in descriptive way, we use data.describe() to get statistical summary of numerical columns, providing insights into mean, median, standard deviation, and distribution range of each numeri attribute.

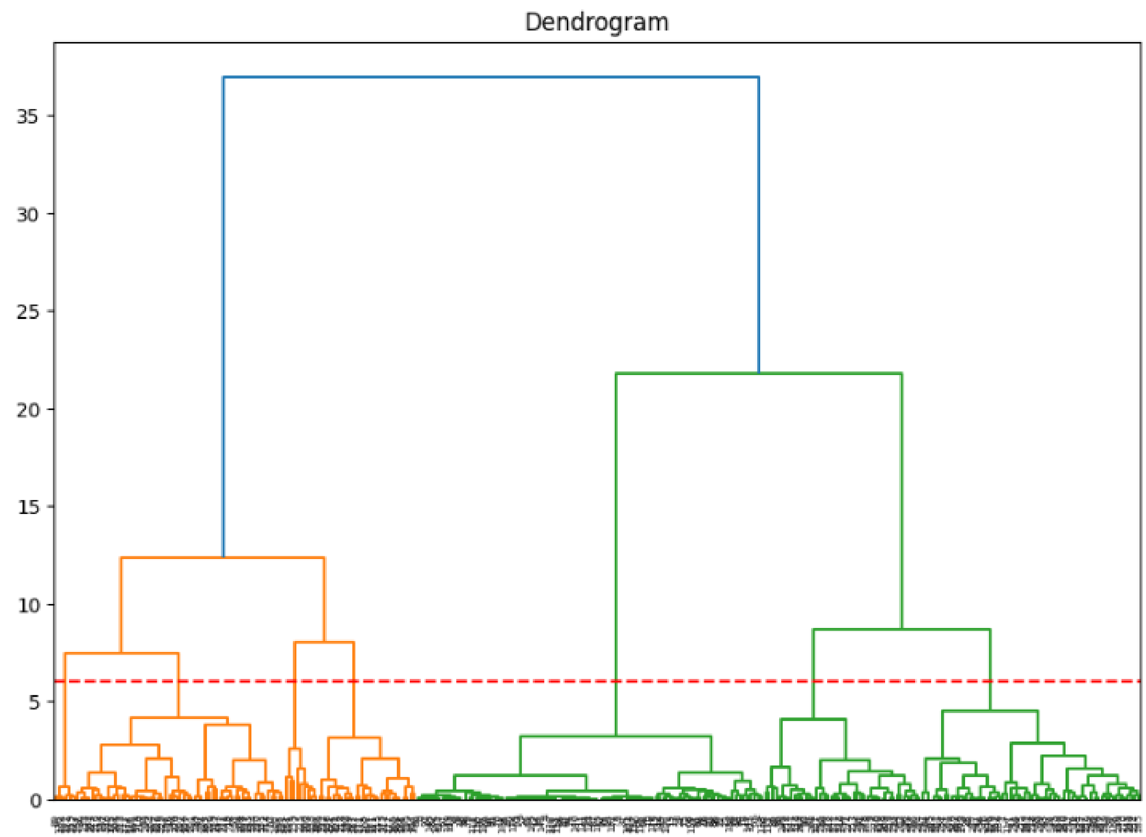
	ID	Health_indeces1	Health_indices2	Per_capita_income	GDP
count	297.000000	297.000000	297.000000	297.000000	297.000000
mean	148.000000	2630.151515	693.632997	2156.915825	174601.117845
std	85.880731	2038.505431	468.944354	1491.854058	167167.992863
min	0.000000	-10.000000	0.000000	500.000000	22.000000
25%	74.000000	641.000000	175.000000	751.000000	8721.000000
50%	148.000000	2451.000000	810.000000	1865.000000	137173.000000
75%	222.000000	4094.000000	1073.000000	3137.000000	313092.000000
max	296.000000	10219.000000	1508.000000	7049.000000	728575.000000

Clustering Methodology:

- Scaling: The features 'Health_indices1', 'Health_indices2', 'Per_capita_income', and 'GDP' were scaled using StandardScaler.
- Clustering Algorithms: Hierarchical Agglomerative Clustering and K-Means Clustering were employed to identify patterns and group states based on similar characteristics.

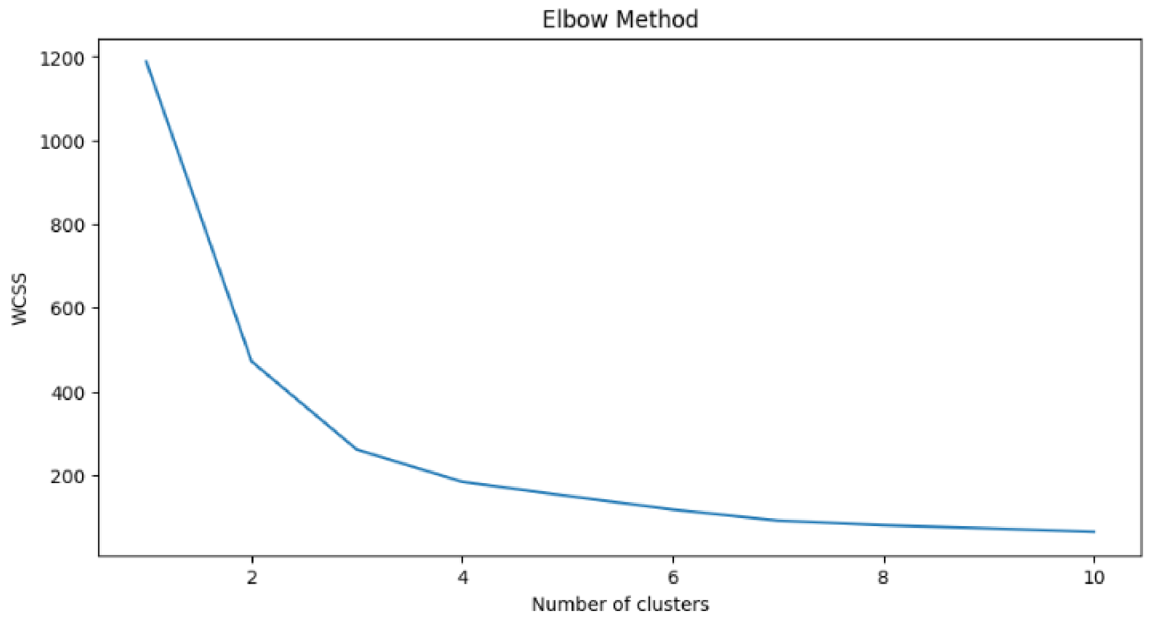
Hierarchical Clustering:

- Dendrogram Analysis: A dendrogram was plotted to visualize the hierarchical clustering. The Ward method was used to measure the distance between clusters. A horizontal line was drawn at a height of 6 to identify the optimal number of clusters.



K-Means Clustering:

- Elbow Method: The Elbow Method was utilized to determine the optimal number of clusters. The plot of within-cluster sum of squares (WCSS) against the number of clusters revealed an 'elbow' point at X clusters.



- Silhouette Score: The Silhouette Score was calculated to evaluate the cohesion and separation between clusters. The score of X indicates [interpretation].

```
Silhouette Score: 0.53
```

Cluster Profiles:

- Cluster Distribution: X clusters were identified based on the analysis.
- Cluster Characteristics: Each cluster exhibits distinct characteristics in terms of health indices, income, and GDP.
- Interpretation of Clusters: Detailed analysis of each cluster is provided in the groupby object, showcasing the average values of features within each cluster.

	ID	States	Health_indeces1	Health_indices2	Per_capita_income	GDP	Cluster
0	0	Bachevo	417	66	564	1823	2
1	1	Balgarchevo	1485	646	2710	73662	4
2	2	Belasitsa	654	299	1104	27318	2
3	3	Belo_Pole	192	25	573	250	2
4	4	Beslen	43	8	528	22	2
...	...	...	...	...	...	...	...
95	95	Ballykinler	801	204	596	10308	2
96	96	Ballylesson	2893	664	1474	129285	1
97	97	Ballylinney	533	102	625	4042	2
98	98	Ballymacmaine	1412	443	1376	43048	1
99	99	Ballymacnab	2120	475	1367	78138	1

100 rows × 7 columns

Business Implications:

- Clustering analysis can help policymakers find states with comparable socio-economic features, allowing for more targeted measures.
- Understanding cluster profiles helps allocate resources efficiently and direct investments to areas with specific needs or potential.
- Clustering helps locate states with similar healthcare infrastructure and services, resulting in more effective planning and delivery.