Data Mining Project – Business Report

Table of Contents

1 Introduction

This report contains two different cases and datasets, for **Part 1: Clustering:** and **Part 2: PCA**. These two datasets are analyzed. The **Part 1: Clustering** presents the **State_wise_Health_income.csv dataset**, which contains a country's health and economic conditions in various states. The Group States depending on their situation's similarity for submitting these groups to the government to carry out effective measures by escalating their Health and Economic conditions. The following are the variables of this dataset:

a) **States:** It refers to the state names.
b) **Health_indeces1:** A composite index rolls various associated measures (indicators) into a single score, which gives a summary of the health system's performance in a state.
c) **Health_indeces2:** A composite index rolls various associated measures (indicators) into a single score, which gives a summary of the health system's performance in a particular area of a state.
d) **Per_capita_income:** It denotes the Per capita income (PCI), which is used for measuring the average income a person earned in a provided area (region, city, country, and so on.) in a particular year. It is estimated by dividing the area's total income with the total population.
e) **GDP:** It depicts a state or country's economic condition, and it is utilized for calculating growth rate and size of the economy.

Whereas, the **Part 2: PCA** presents the **Hair Salon.csv**, which contains information about a hair salon chain's hair product service and their market segmentation.

The following are the variables of this dataset:

- ProdQual: It refers to the quality of the product.
- Ecom: It refers to the e-commerce.
- techSup: It refers to the technical support.
- CompRes: It refers to the complaint resolution.
- Advertising: It refers to the advertising.
- ProdLine: It refers to the product line.

- SalesFImage: It refers to the sales force image.
- ComPricing: It refers to the competitive pricing.
- WartClaim: It refers to the warranty and claim.
- OrderBilling: It refers to the order and billing.
- DelSpeed: It refers to the speed of the delivery.
- Satisfaction: It refers to the customer satisfaction.

Thus, the objective of this report is to conduct clustering and PCA on different case studies and their datasets.

2 Data Exploration

The data for this analysis is explored to see a detailed information about the two given datasets. In this part of the report, the dataset information, missing values, bivariate analysis, univariate analysis, outlier detection and so on are discussed.

2.1 State Wise Health Income Dataset

To begin with, the state wise health income dataset was explored as demonstrated below.

```
data.head(5)
```

	Unnamed: 0	States	Health_indices1	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

The dataset information is represented below.

```
# checking datatypes and number of non-null values for each column
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 296 entries, 0 to 295
Data columns (total 6 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Unnamed: 0       296 non-null    int64
 1   States           296 non-null    object
 2   Health_indices1  296 non-null    int64
 3   Health_indices2  296 non-null    int64
 4   Per_capita_income  296 non-null    int64
 5   GDP              296 non-null    int64
dtypes: int64(5), object(1)
memory usage: 14.0+ KB
```

This dataset has 296 rows and 6 columns. All the columns in the data are non-null. The columns 'Health_indices1', 'Health_indices2', 'Per_capita_income' and 'GDP' are integer datatypes variables, while 'States' is an object (string) column representing state names.

Discussion on Mission Values

Next, if any missing values are present in the datasets is checked. Based on the output, there were no missing values in the state wise health income dataset. After that, the dataset is summarized as presented below:

```
# Let's look at the statistical summary of the data
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Health_indices1	296.0	2629.195946	2041.890970	-10.0	640.0	2446.5	4102.75	10219.0
Health_indices2	296.0	693.594595	469.738035	0.0	173.5	810.5	1076.00	1508.0
Per_capita_income	296.0	2159.597973	1493.663013	500.0	746.0	1869.0	3138.75	7049.0
GDP	296.0	174717.050676	167439.128587	22.0	8679.5	135748.5	314751.25	728575.0

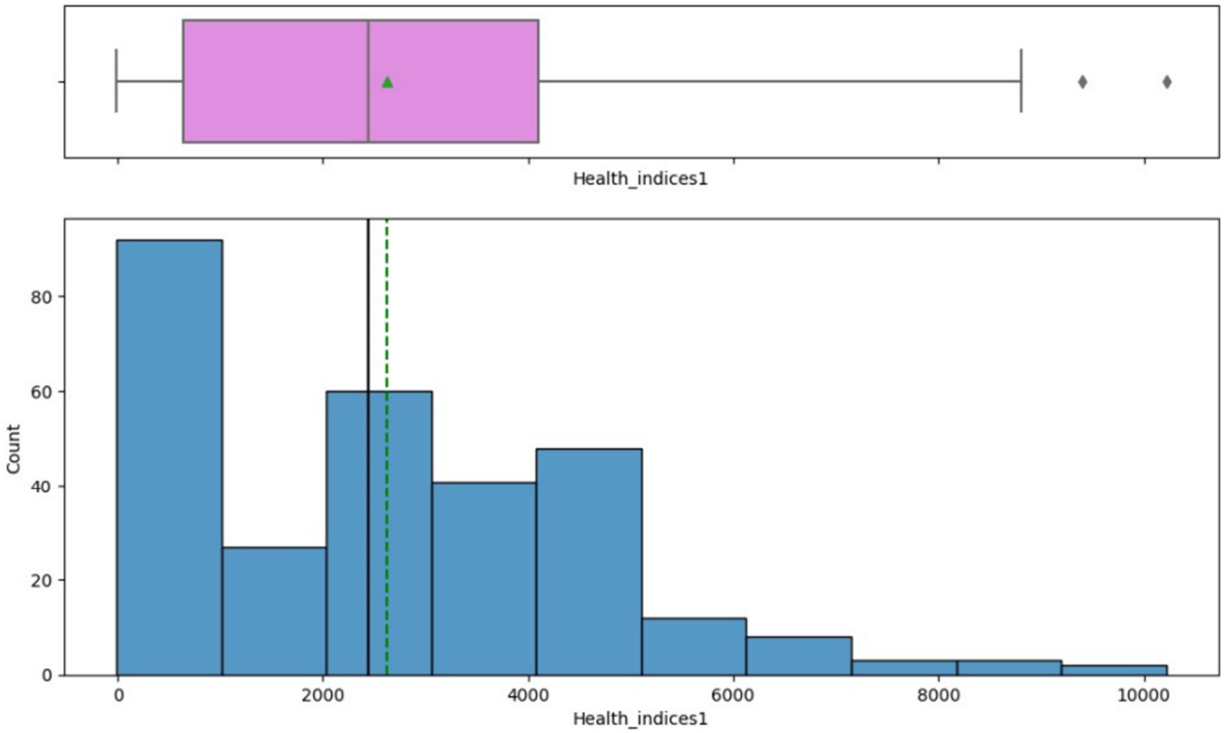
According to the above result,

- The average Health Index (Health_indices1) across different states of the country was around 2630, with the highest was 10,219, and the lowest was -10.
- The average GDP across the states is 167,439, with the highest for a state was 728,575 and lowest was 22.

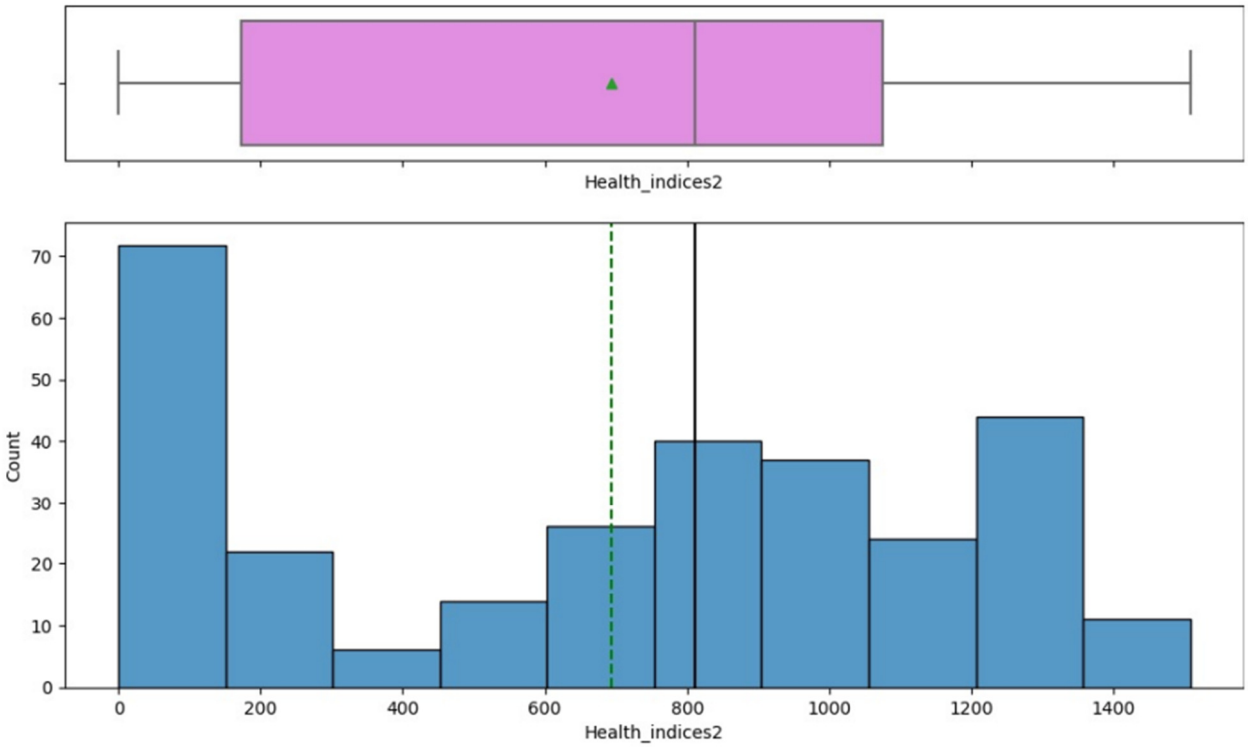
- These summary statistics show variations - with per capita income ranging from 500 to 7049, and GDP from 22 to 728575.

Univariate Analysis

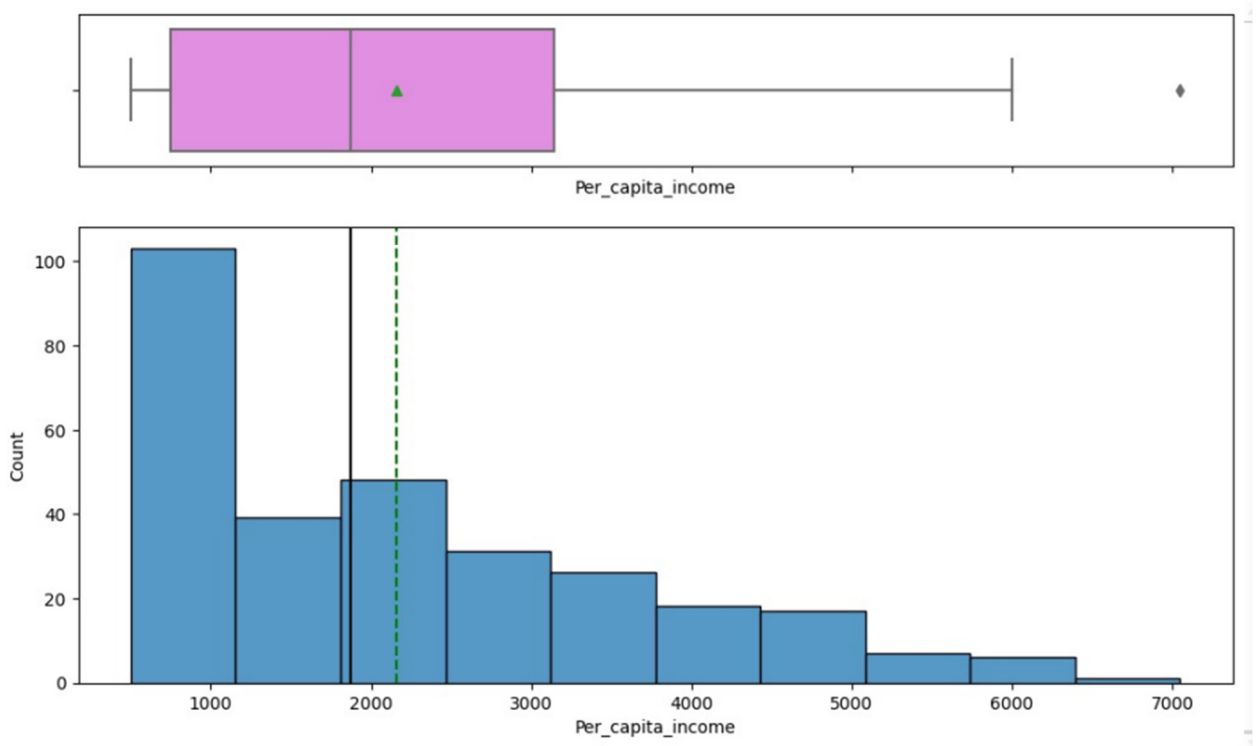
- The health indices 1 attribute has the right-skewed distribution. This means that for majority of the states, the overall Health_indices1 is on the lower side for most of the states. There are few states with Health_indices1 greater than 5000, with the average across all the states being between 2000 to 3000 (mean and median, both). The boxplot and the histogram of this variable indicate the presence of outliers.



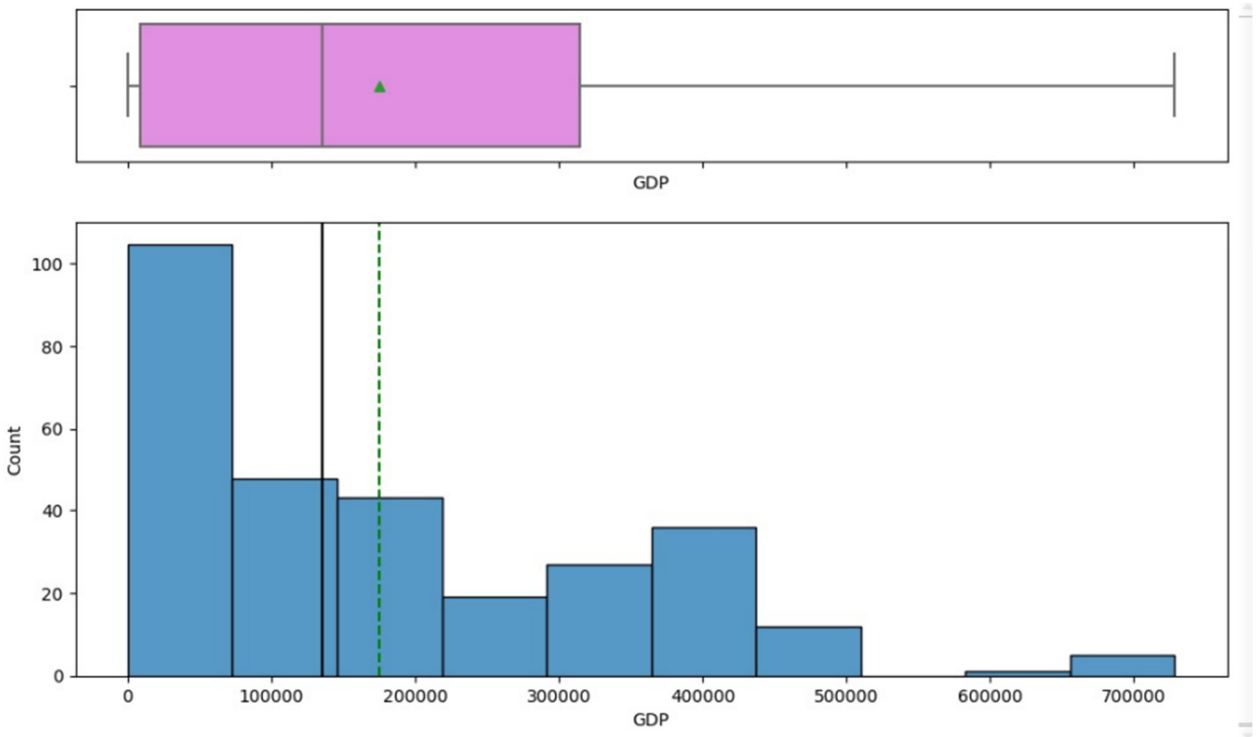
- The health indices 2 attribute does not seem to follow conventional distributions. Majority of the states have Health_indices_2 score between 0 to less than 200. From the boxplot and the histogram, it appears that the variable does not have any outliers.



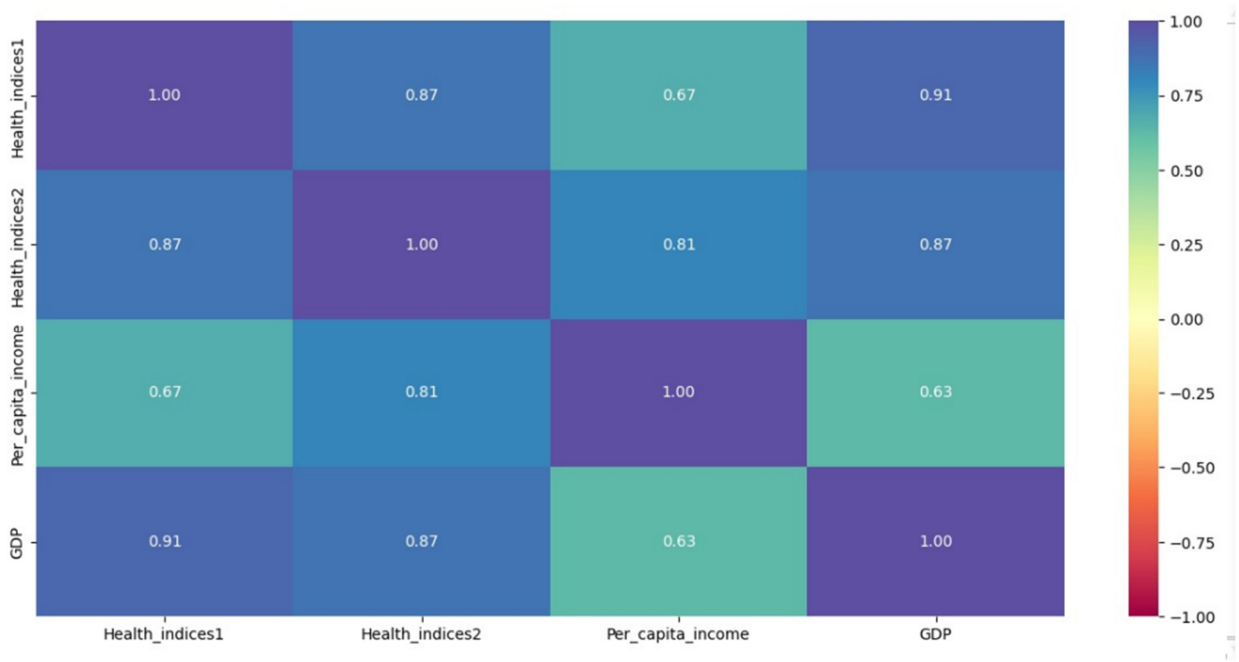
- PCI is financial data and is expected to have right-skewed distribution - meaning, that that majority of the data points, or states, have lower PCI. Few states have high PCI. Average PCI (median) is slightly less than 2000, with 75% of the states having around 3000 PCI. As observed from the plots, the variable has outliers.



- GDP being financial data is also right-skewed. Fewer states have GDP over 300,000, with average GDP across states being less than 200,000 (mean and median, both). There are outliers within this variable.



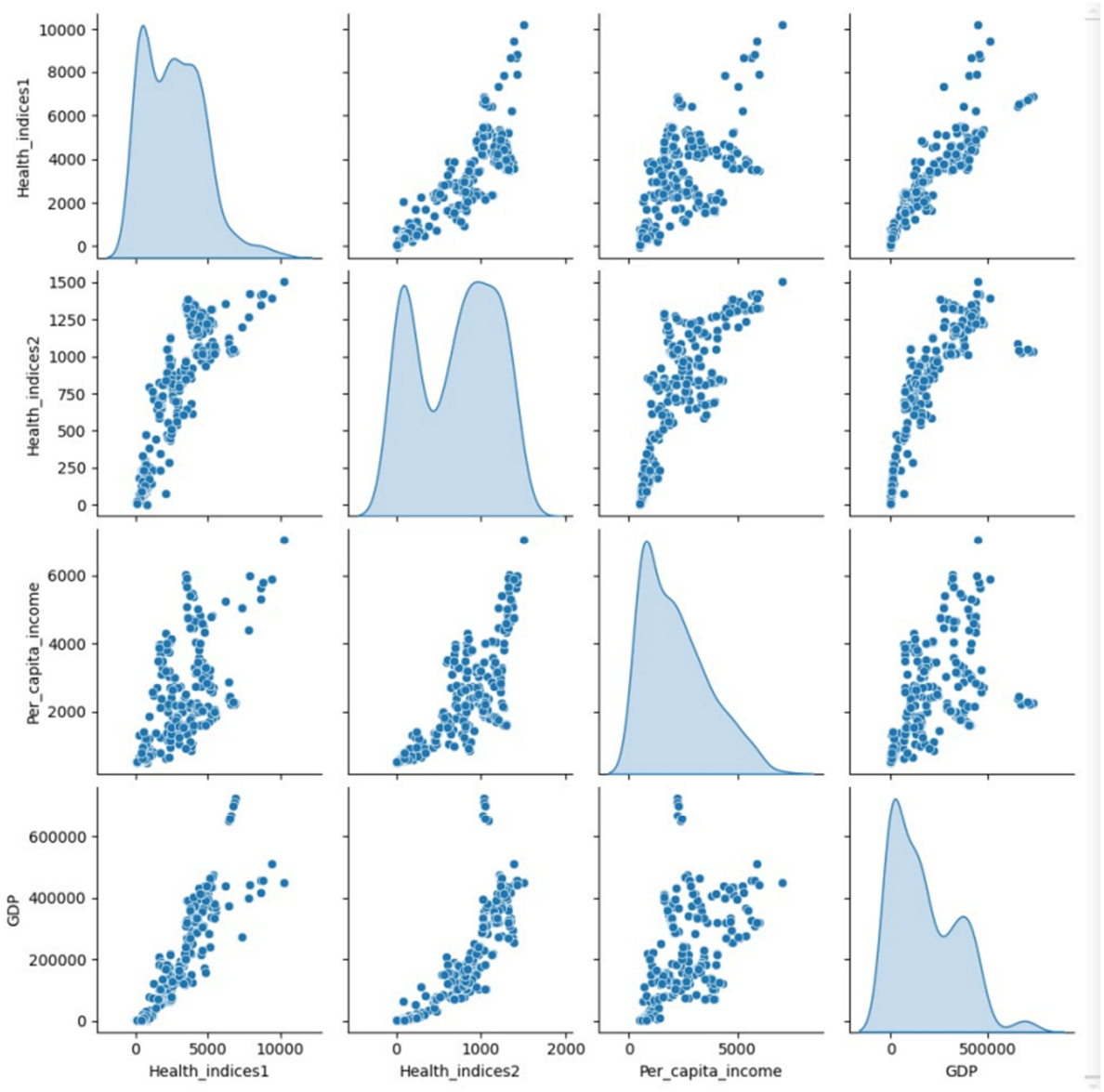
Bivariate Analysis - Correlation



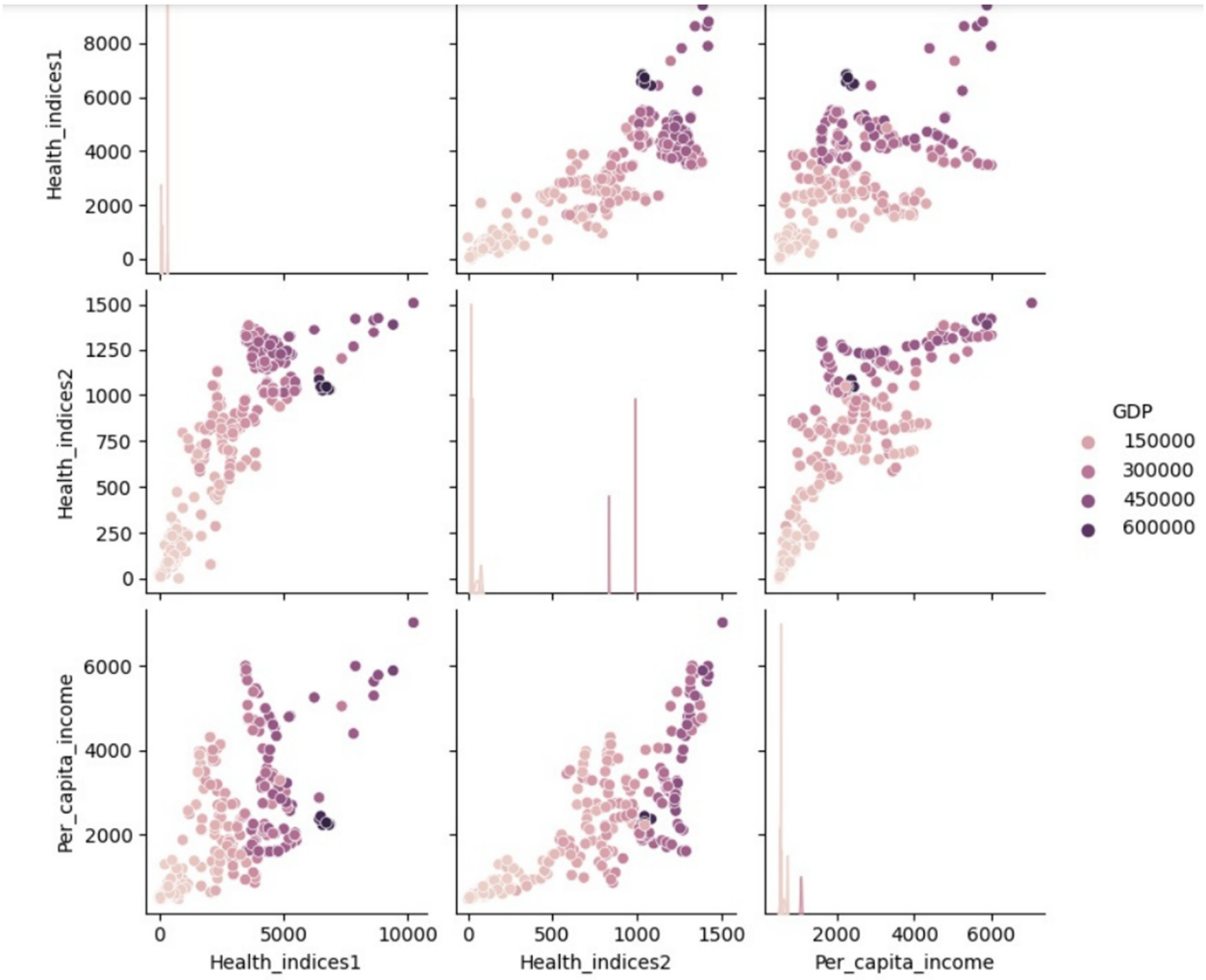
According to the correlation result, the following is obtained:

- A correlation value greater than 0.8 indicates a strong positive linear relationship between two variables.
- The attributes Health_indices1 and Health_indices2 have more than 80% positive correlation.
- GDP and Health_indices1, and GDP and Health_indices2, have more than 80% correlation.
- PCI and Health_indices_2 has more than 80% correlation

Bivariate Analysis – Pair plot



As observed from the correlation matrix, and the above scatter plots, there are positive correlations observed amongst the numerical variables in the dataset. Upon observing closely, some non-linear relationships are also observed, e.g., Health_indices1 and Health_indices2; Health_indices2 and PCI; Health_indices2 and GDP. Next, we can add hue and see if we can see some clustered distributions.



According to the above plot, A general trend that is observed for Health_indices1 and Health_indices2 scores with GDP is that GDP is higher for data points with high health indices scores. Similarly, GDP is higher for data points with high PCI and health index scores.

2.2 Hair Salon Dataset

The hair salon dataset was imported and the irrelevant variables for salon chain's market segmentation prediction were removed as represented below.

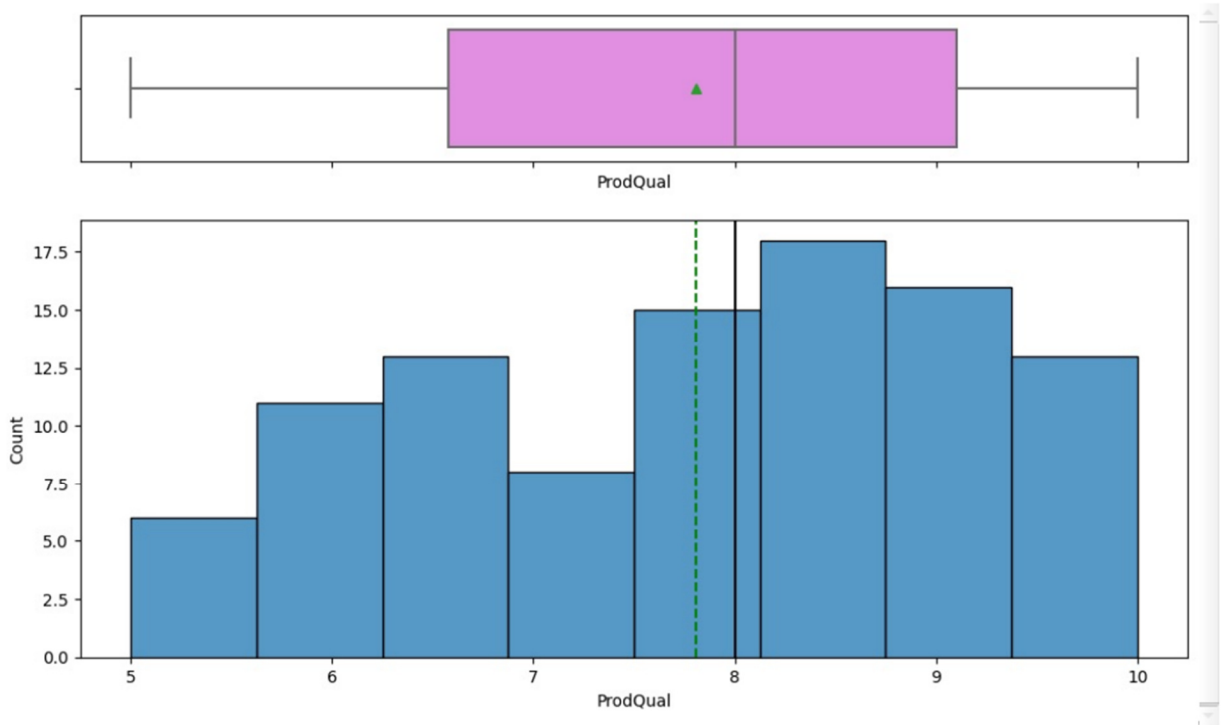
```
data.head(5)        ## Complete the code to check the head
```

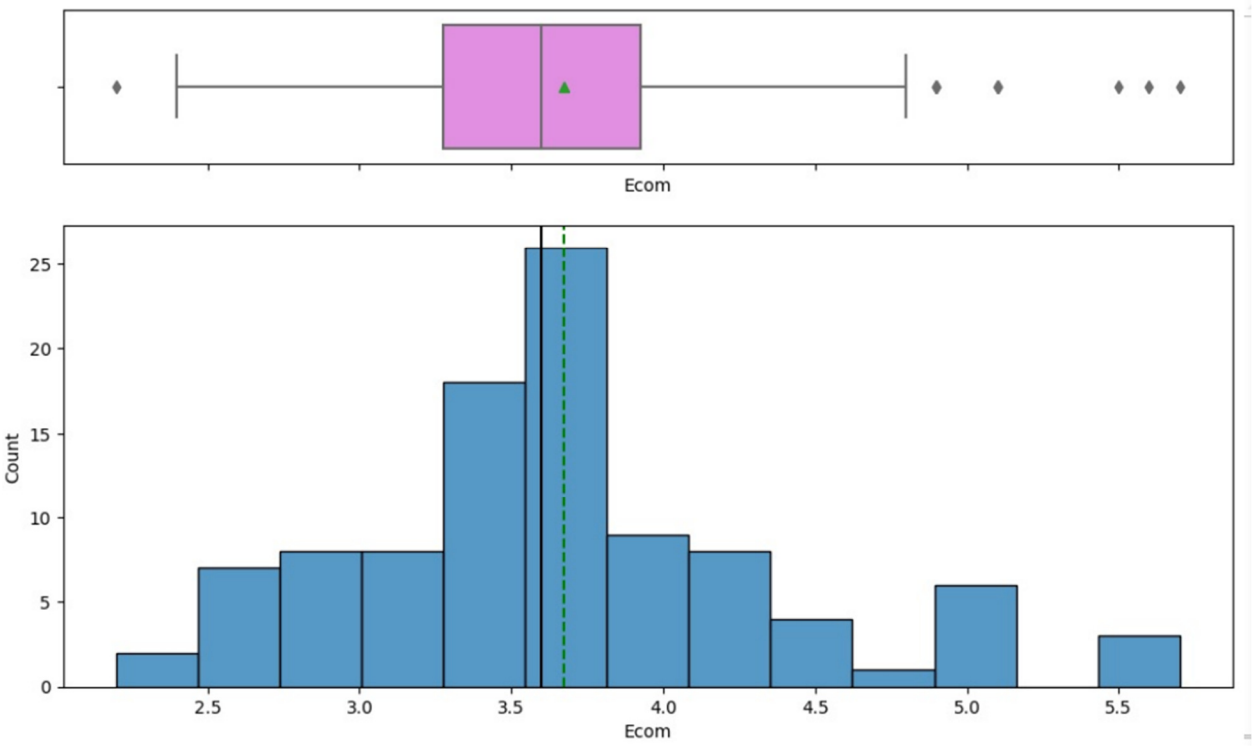
	ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfactio
0	1	8.5	3.9	2.5	5.9	4.8	4.9	6.0	6.8	4.7	5.0	3.7	8
1	2	8.2	2.7	5.1	7.2	3.4	7.9	3.1	5.3	5.5	3.9	4.9	5
2	3	9.2	3.4	5.6	5.6	5.4	7.4	5.8	4.5	6.2	5.4	4.5	8
3	4	6.4	3.3	7.0	3.7	4.7	4.7	4.5	8.8	7.0	4.3	3.0	4
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

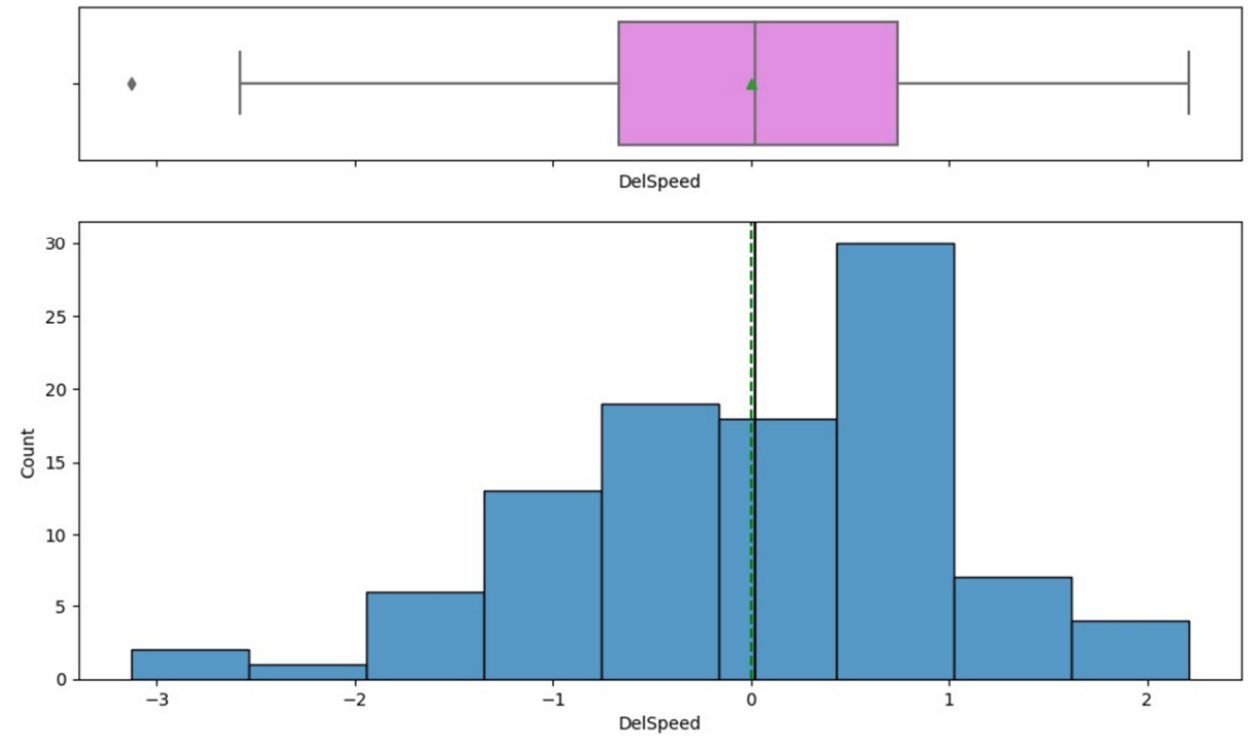
```
data.drop(['ID','Satisfaction'],axis=1, inplace = True)     ## Complete the code to drop the ID and Satisfaction columns
```

Univariate and Bi variable Analysis

According to the univariate and bivariate analysis result, the data fields appear the same as before. Variables 'ProdQual', 'TechSup', 'CompRes', 'Advertising', 'ProdLine', 'CompPricing', 'WartyClaim' are quite normally distributed. Other variables such as 'DelSpeed', 'OrdBilling', 'SalesFImage', and 'Ecom' still show some outliers. Scaling has not affected the outliers as demonstrated below.







3 Data Preprocessing

The datasets are prepared for conducting the analysis.

3.1 State wise Health Income Dataset

Outlier Detection

First, the outliers in the data will be identified by using the z-score with the help of the threshold obtained in the above plot. The result is shown as follows.

```
The following are the outliers in the data:

Health_indices1 : [8802, 9403, 10219]

Health_indices2 : []

Per_capita_income : [7049]

GDP : [703190, 713295, 728575]
```

According to the outlier result, the following points are determined:

- Identification of outliers was done based on a threshold of 3 standard deviations from the mean.
- Health_indices1 outliers were [8802, 9403, 10219]. Three states had high values for "Health_indices1" that are considered outliers.
- These states may exhibit significantly different characteristics in terms of health indices compared to the majority. There were no identified outliers for "Health_indices2".
- Per_capita_income outlier is [7049]. One state had a high per capita income value (7049), considered an outlier compared to the rest.
- GDP Outliers are [703190, 713295, 728575]. Three states had high GDP values that are considered outliers. These states may significantly differ in terms of economic output compared to others.

Data Scaling

Next, scaling of dataset was done with the help of standard scaler function, as represented below.

```
# scaling the data before clustering
scaler = StandardScaler()
subset = df.copy()
subset_scaled = scaler.fit_transform(subset)
```

```
# creating a dataframe of the scaled data
subset_scaled_df = pd.DataFrame(subset_scaled, columns=subset.columns)
```

Now, clustering analysis was possible to conduct, because the dataset was cleaned and preprocessed for this analysis.

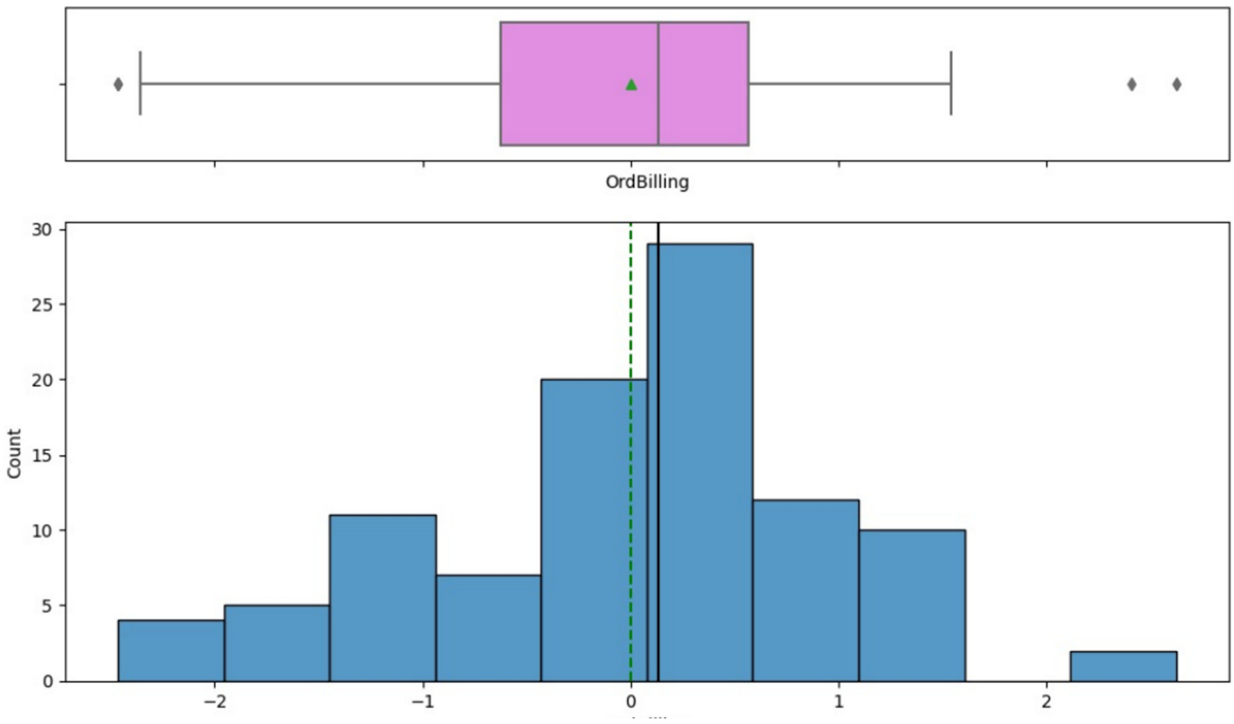
3.2 Hair Salon Dataset

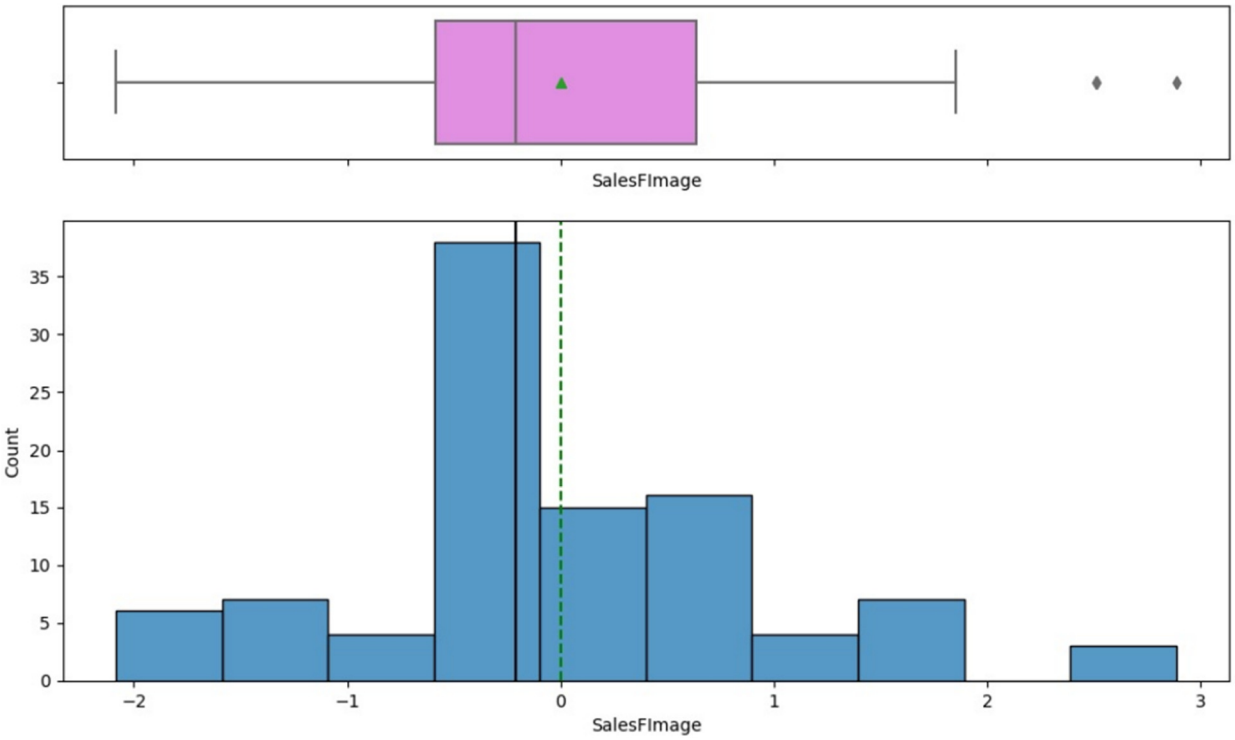
The dataset for the PCA analysis was prepared by conducting scaling and outlier detection. The scaling of data is represented below.

```
# scaling the data before clustering
scaler = StandardScaler()
subset = data.copy()
subset_scaled = scaler.fit_transform(subset)
```

```
# creating a dataframe of the scaled data
subset_scaled_df = pd.DataFrame(subset_scaled, columns=subset.columns)
```

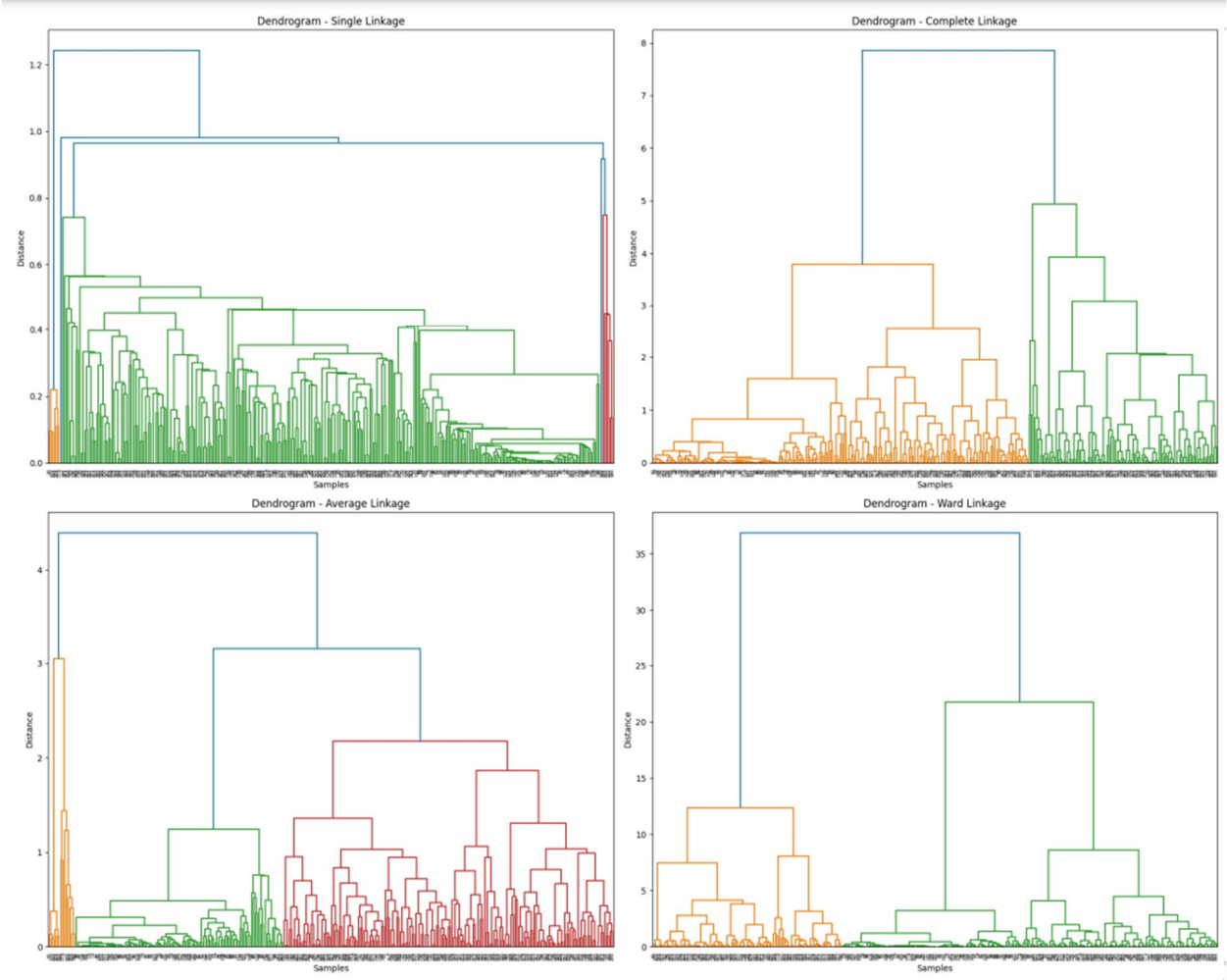
Next, the outlier detection was conducted. The result showed, Post Scaling of data, the data fields appeared the same as before. Variables 'ProdQual', 'TechSup', 'CompRes', 'Advertising', 'ProdLine', 'CompPricing', 'WartyClaim' were quite normally distributed. Other variables such as 'DelSpeed', 'OrdBilling', 'SalesFImage', and 'Ecom' still showed some outliers. Scaling had not affected the outliers as represented below.





4 Clustering Analysis

Clustering analysis was conducted on state wise health income dataset. To begin with, the hierarchical clustering analysis was performed as demonstrated below.



According to the above result, upon examining the dendrograms, it was observed that when Ward Linkage was applied, well separated clusters were obtained. In this case, the dendrogram followed a clear, hierarchical structure. The clusters are seen to merge at a higher vertical distance/height, which indicate the distinctness of the clusters.

Next, the number of clustered was determined based on the ward linkage technique as demonstrated below.

```
linkage_matrix_ward = linkage(subset_scaled_df, method='ward')
max_d = 7  # Maximum distance threshold for forming clusters
clusters = fcluster(linkage_matrix_ward, t=max_d, criterion='distance')

num_clusters = len(np.unique(clusters))
print(f"Number of clusters determined using Ward linkage: {num_clusters}")
```

```
Number of clusters determined using Ward linkage: 7
```

Based on the result, the ward linkage technique told the number of clusters as 7 for this prediction. Then, silhouette score technique can be checked to confirm the number of clusters is 7, which is best for this prediction. The result is shown below.

```
from sklearn.cluster import AgglomerativeClustering
from sklearn.metrics import silhouette_score
hc_df = subset_scaled_df.copy()
sil_score_hc = []
cluster_list = list(range(2, 10))
for n_clusters in cluster_list:
    clusterer = AgglomerativeClustering(n_clusters=n_clusters)
    preds = clusterer.fit_predict((subset_scaled_df))
    score = silhouette_score(hc_df, preds)
    sil_score_hc.append(score)
    print("For n_clusters = {}, silhouette score is {}".format(n_clusters, score))
```

```
For n_clusters = 2, silhouette score is 0.5005875144522923
For n_clusters = 3, silhouette score is 0.5260929381868132
For n_clusters = 4, silhouette score is 0.5374753668213341
For n_clusters = 5, silhouette score is 0.5136233841673663
For n_clusters = 6, silhouette score is 0.5251014259304009
For n_clusters = 7, silhouette score is 0.5475499898084455
For n_clusters = 8, silhouette score is 0.5214864942797744
For n_clusters = 9, silhouette score is 0.494016002953426
```

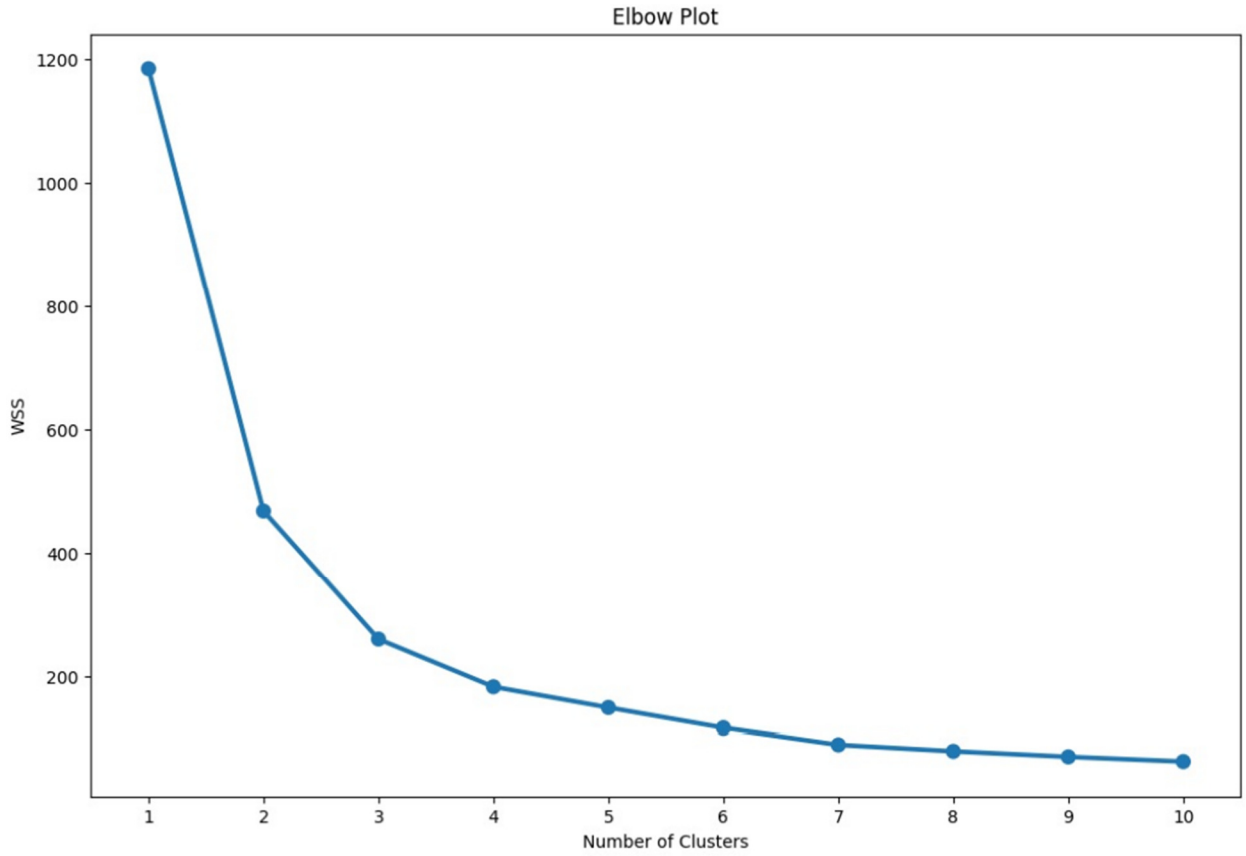
Based on the silhouette score result, the cluster 7 had the highest silhouette score when compared with the other number of clusters. Therefore, the number of clusters was 7, which is best for this prediction. The final model of agglomerative clustering is demonstrated below.

```
HCmodel = AgglomerativeClustering(n_clusters=7, affinity="euclidean", linkage="average")
HCmodel.fit(hc_df)
```

```
CPU times: user 8.65 ms, sys: 1.98 ms, total: 10.6 ms
Wall time: 6.71 ms
```

```
AgglomerativeClustering(affinity='euclidean', linkage='average', n_clusters=7)
```

Next, the K means clustering is conducted for the State wise health income dataset. Before doing this, the number of clusters is determined which are best for this prediction, by using the elbow method and silhouette score technique. The elbow method plot is demonstrated below.

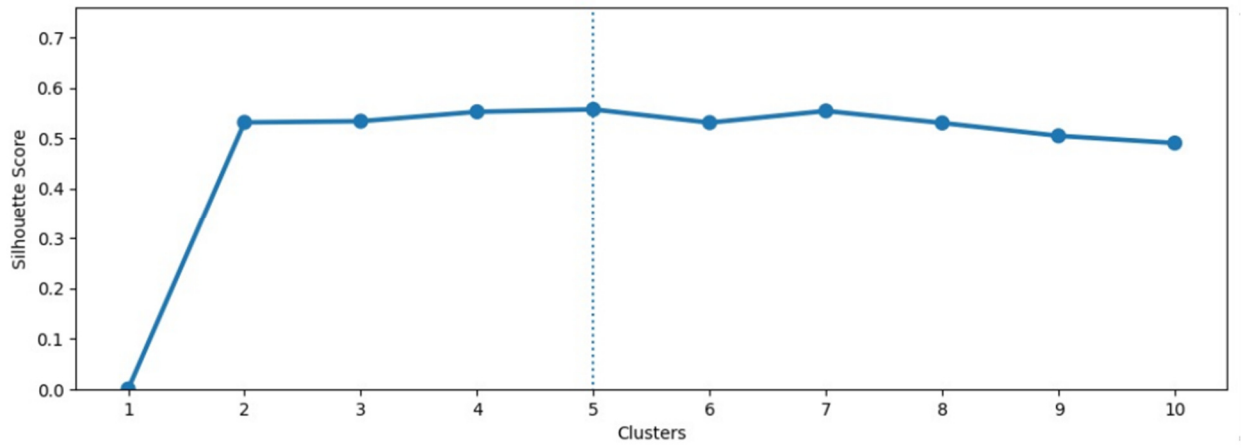


According to the plot, the optimal number of clusters can be either 3 or 4, based on the Elbow Plot method. Next, the Silhouette Scores are checked. The result is shows as follows:

```
ss={1:0}
for i in range(2, 11):
 clusterer = KMeans(n_clusters = i, init = 'k-means++', random_state = 1)
 y=clusterer.fit_predict(k_means_df)                                    ## Co
 s =silhouette_score(k_means_df, y )                                     ## C
 ss[i]=round(s,5)
 print("The Average Silhouette Score for {} clusters is {}".format(i,round(s,5)))
```

```
The Average Silhouette Score for 2 clusters is 0.53121
The Average Silhouette Score for 3 clusters is 0.53353
The Average Silhouette Score for 4 clusters is 0.55225
The Average Silhouette Score for 5 clusters is 0.55701
The Average Silhouette Score for 6 clusters is 0.53076
The Average Silhouette Score for 7 clusters is 0.55395
The Average Silhouette Score for 8 clusters is 0.53008
The Average Silhouette Score for 9 clusters is 0.50449
The Average Silhouette Score for 10 clusters is 0.49018
```

From the silhouette scores, which k value had the highest value and elbow in elbow curve was determined as demonstrated below.



According to the result, Silhouette score was the highest for k=5. However, the elbow plot suggested 3 or 4 clusters as optimal k. These k values also had similar silhouette scores as k=5. Based on the elbow plot and silhouette scores, k = 4 should be an optimal number of clusters.

Next, a k means clustering model was built as demonstrated below.

```
kmeans = KMeans(n_clusters = 4 , init = 'k-means++', random_state = 1)

kmeans.fit_predict(k_means_df)                                    ## Com
```

```
array([1, 0, 1, 1, 1, 1, 1, 2, 1, 0, 1, 1, 1, 1, 1, 1, 3, 1, 1, 3, 0, 1,
       1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 3, 1, 1, 3, 0, 1, 0, 1, 1, 0, 0,
       0, 1, 1, 0, 1, 1, 1, 0, 1, 3, 1, 0, 0, 1, 1, 0, 0, 1, 1, 3, 0, 1,
       0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1,
       1, 3, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 2, 1, 0, 1, 0, 0, 1, 1,
       1, 2, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 2, 1, 0, 1, 1, 0, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 3, 1, 0, 1, 2, 3, 3, 3,
       3, 2, 3, 2, 2, 3, 3, 2, 3, 2, 3, 2, 2, 3, 2, 2, 3, 2, 3, 3, 3, 2,
       3, 2, 2, 3, 2, 2, 3, 3, 3, 0, 3, 2, 3, 3, 2, 2, 2, 3, 3, 2, 3, 3,
       2, 2, 2, 3, 2, 2, 3, 3, 0, 3, 3, 3, 3, 2, 3, 3, 3, 2, 3, 3, 2, 3,
       3, 3, 3, 3, 0, 2, 3, 3, 3, 3, 3, 3, 2, 2, 3, 3, 0, 0, 0, 0, 0, 0,
       0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0], dtype=int32)
```

Next, cluster profiling was done that will provide the value count of the each k mean segments as demonstrated below.

K_means_segments	0	1	2	3
Health_indices1	2593.93	499.16	5146.44	4799.36
Health_indices2	783.80	116.36	1327.14	1142.29
Per_capita_income	2475.14	693.77	5047.08	2372.22
GDP	141273.93	9428.10	367196.92	396907.24
freq	100.00	101.00	36.00	59.00

According to the cluster profiling result, there were 4 clusters obtained as a result of k-means clustering.

- Cluster 0: 100 data points.
- Cluster 1: 101 data points.
- Cluster 2: 36 data points.
- Cluster 3: 59 data points.

Afterwards, the cluster profiling was compared with clustering. The below result compared the cluster profiling with the K means clustering.

km_cluster_profile

K_means_segments	Health_indices1	Health_indices2	Per_capita_income	GDP	count_in_each_segment
0	2593.930000	783.800000	2475.140000	141273.930000	100
1	499.158416	116.356436	693.772277	9428.099010	101
2	5146.444444	1327.138889	5047.083333	367196.916667	36
3	4799.355932	1142.288136	2372.220339	396907.237288	59

Based on above result,

- **Cluster 0:** The average GDP of this cluster is 141,274. This somewhat lies in the middle when compared to the rest of the clusters. The health index scores and PCI are also somewhat in the middle, when compared to other clusters.
- **Cluster 1:** The average GDP of this cluster is 9428. This is the lowest when compared to the rest of the clusters. This cluster happens to have the lowest average PCI, and lowest average health index scores.

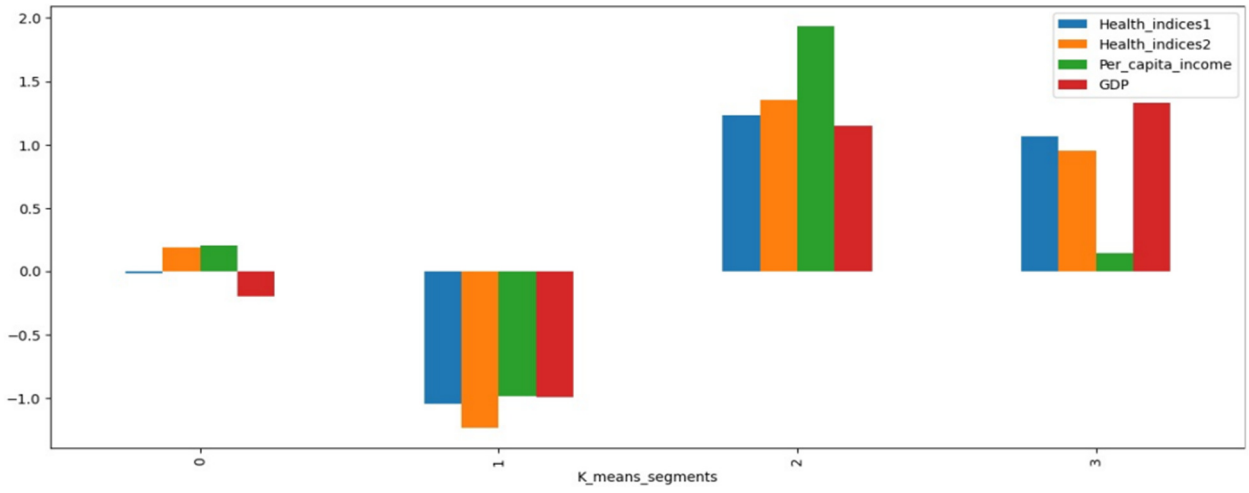
- **Cluster 2:** The average GDP of this cluster was 367,196. This was the second highest when compared to the rest of the clusters. The cluster had the highest average PCI, and highest average health index scores.
- **Cluster 3:** The average GDP of this cluster was 396,907. This was the highest when compared to the rest of the clusters. The cluster had the second highest average PCI, and average health index scores.

The below result compared the cluster profiling with Hierarchical clustering.

HC_segments	Health_indices1	Health_indices2	Per_capita_income	GDP	count_in_each_segment
0	2518.715909	802.022727	2535.636364	145730.011364	88
1	4601.105263	1138.982456	2432.368421	350721.543860	57
2	634.532110	144.614679	751.981651	14274.229358	109
3	6649.333333	1044.000000	2299.833333	687649.666667	6
4	8927.666667	1417.166667	5940.166667	454834.333333	6
5	4162.750000	1314.607143	4879.035714	350573.285714	28
6	7574.500000	1232.500000	4720.500000	337015.500000	2

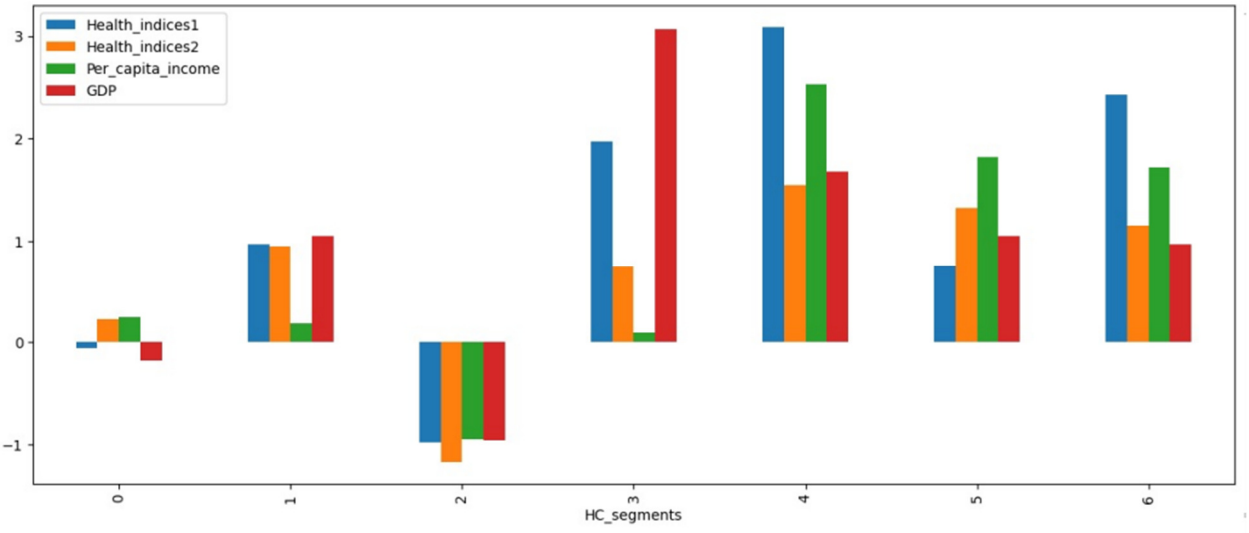
When compared with k-means clustering, hierarchical clustering suggested a greater number of clusters. A few of these clusters such as clusters 3, 4 and 6 had a very few data points. With that in consideration, k-means clustering seemed to perform better.

Further, the K means clustering and Hierarchical Clustering were compared as demonstrated below.



According to the above result, there are 4 clusters suggested by this method of clustering. Though four, 3 distinct groups are observed to appear within the dataset:

- **Excellent Health and Economic Conditions:** Cluster 2 and 3. As deciphered from the graphs, these clusters have highest average PCI, GDP, as well as health index scores.
- **Adequate Health and Economic Conditions:** Cluster 0. As deciphered from the graphs, these clusters have medium level PCI, GDP, as well as health index scores.
- **Poor Health and Economic Conditions:** Clsuter 1: As deciphered from the graphs, these clusters have lowest average PCI, GDP, as well as health index scores.
- Most data points belonged to the Adequate/Poor Health and Economic Conditions.



According to the above plot,

- Cluster 4 had the highest average health index scores, and second highest GDP, but very few data points. Cluster 3 was similar to Cluster 4 in these aspects.
- The cluster with highest number of instances was Cluster 2, which happens to have the lowest average health index scores, PCI, and GDP.
- There was no clear distinction amongst, but the remaining clusters fell within the category of intermediate health scores and economic conditions.

5 PCA Analysis

PCA analysis was conducted for Hair salon dataset to do salon chain's market segmentation. Covariance matrix was created with PCA as demonstrated below.

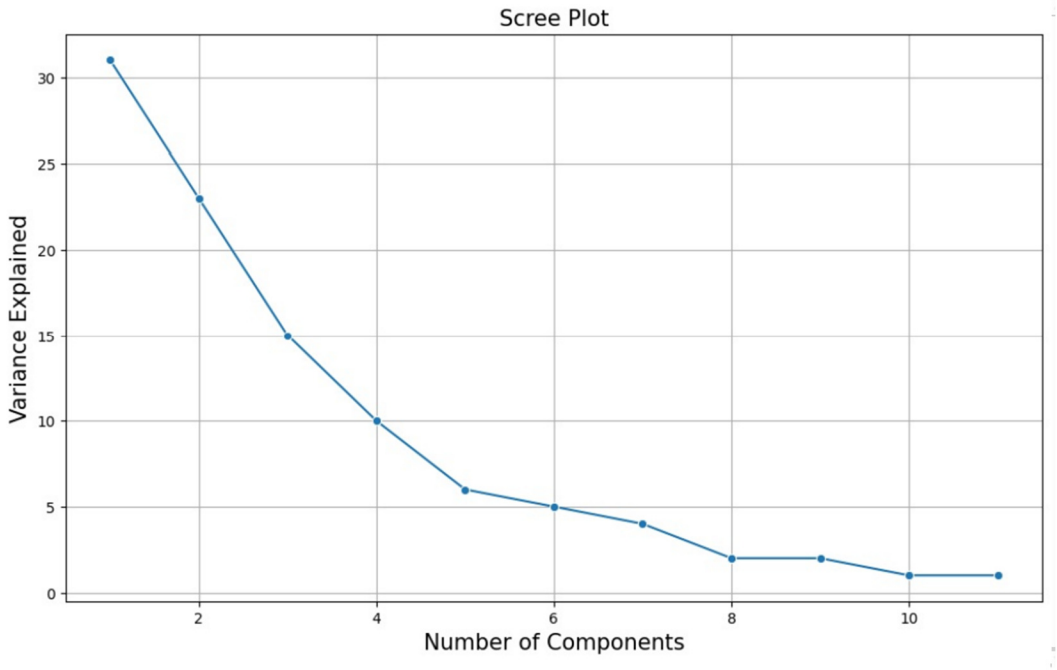
```
pd.DataFrame(np.round(pca.get_covariance(),2),columns=subset_scaled_df.columns,index=subset_scaled_df.columns) #cov matrix
```

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
ProdQual	1.01	-0.14	0.10	0.11	-0.05	0.48	-0.15	-0.41	0.09	0.11	0.03
Ecom	-0.14	1.01	0.00	0.14	0.43	-0.05	0.80	0.23	0.05	0.16	0.19
TechSup	0.10	0.00	1.01	0.10	-0.06	0.19	0.02	-0.27	0.81	0.08	0.03
CompRes	0.11	0.14	0.10	1.01	0.20	0.57	0.23	-0.13	0.14	0.76	0.87
Advertising	-0.05	0.43	-0.06	0.20	1.01	-0.01	0.55	0.14	0.01	0.19	0.28
ProdLine	0.48	-0.05	0.19	0.57	-0.01	1.01	-0.06	-0.50	0.28	0.43	0.61
SalesFImage	-0.15	0.80	0.02	0.23	0.55	-0.06	1.01	0.27	0.11	0.20	0.27
ComPricing	-0.41	0.23	-0.27	-0.13	0.14	-0.50	0.27	1.01	-0.25	-0.12	-0.07
WartyClaim	0.09	0.05	0.81	0.14	0.01	0.28	0.11	-0.25	1.01	0.20	0.11
OrdBilling	0.11	0.16	0.08	0.76	0.19	0.43	0.20	-0.12	0.20	1.01	0.76
DelSpeed	0.03	0.19	0.03	0.87	0.28	0.61	0.27	-0.07	0.11	0.76	1.01

Next, eigen values and eigen vector are obtained as represented below.

```
Eigenvectors: [[-0.13 -0.17 -0.16 -0.47 -0.18 -0.39 -0.2   0.15 -0.21 -0.44 -0.47]
 [-0.31  0.45 -0.23  0.02  0.36 -0.28  0.47  0.41 -0.19  0.03  0.07]
 [ 0.06 -0.24 -0.61  0.21 -0.09  0.12 -0.24  0.05 -0.6   0.17  0.23]
 [ 0.64  0.27 -0.19 -0.21  0.32  0.2   0.22 -0.33 -0.19 -0.24 -0.2 ]
 [ 0.23  0.42 -0.02  0.03 -0.8   0.12  0.2   0.25 -0.03  0.03 -0.04]
 [-0.56  0.26 -0.11 -0.03 -0.2   0.1   0.1  -0.71 -0.14 -0.12  0.03]
 [ 0.19  0.06 -0.02 -0.01 -0.06 -0.61  0.   -0.31 -0.03  0.66 -0.23]
 [ 0.14 -0.12  0.46  0.51 -0.05 -0.33  0.17 -0.1  -0.44 -0.37  0.07]
 [ 0.03 -0.54 -0.36  0.09 -0.15 -0.08  0.64 -0.09  0.32 -0.1  -0.02]
 [ 0.07  0.28 -0.39  0.53  0.04 -0.23 -0.35 -0.05  0.44 -0.3  -0.12]
 [ 0.18  0.06 -0.05 -0.36 -0.08 -0.39 -0.08 -0.1   0.13 -0.19  0.78]]
```

Then, a scree plot was created to determine the number of component, which was best for this prediction as demonstrated below.



According to the above result, the first two principal components explained more than 50% variance in the data.

Next, PCA was applied for the number of decided components to get the loadings and component output. The result is shown as follows.

```
df_pca_loading = pd.DataFrame(pca.components_,columns=list(subset_scaled_df),index=['PC0','PC1'])
df_pca_loading.shape
```

```
(2, 11)
```

```
df_pca_loading = np.round(df_pca_loading,2)
```

```
df_pca_loading.style.highlight_max(color = 'lightgreen', axis = 0)
```

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
PC0	-0.130000	-0.170000	-0.160000	-0.470000	-0.180000	-0.390000	-0.200000	0.150000	-0.210000	-0.440000	-0.470000
PC1	-0.310000	0.450000	-0.230000	0.020000	0.360000	-0.280000	0.470000	0.410000	-0.190000	0.030000	0.070000

Then, the explained variance ratio, principal component 1 and 2 loadings are determined as demonstrated below.

```
# Display results
print("Explained Variance Ratio:", explained_var_ratio)
print("Principal Component 1 Loadings:", pc1_loadings)
print("Principal Component 2 Loadings:", pc2_loadings)
```

```
Explained Variance Ratio: [0.31154285 0.2318997 ]
Principal Component 1 Loadings: [-0.13378962 -0.16595278 -0.15769263 -0.47068359 -0.18373495 -0.38676517
 -0.2036696   0.15168864 -0.21293363 -0.43721774 -0.47308914]
Principal Component 2 Loadings: [-0.31349802  0.44650918 -0.23096734  0.01944394  0.36366471 -0.28478056
  0.47069599  0.4134565  -0.19167191  0.02639905  0.07305172]
```

According to the above result, the created PCA model explained variance ratio as 0.31 and 0.23.

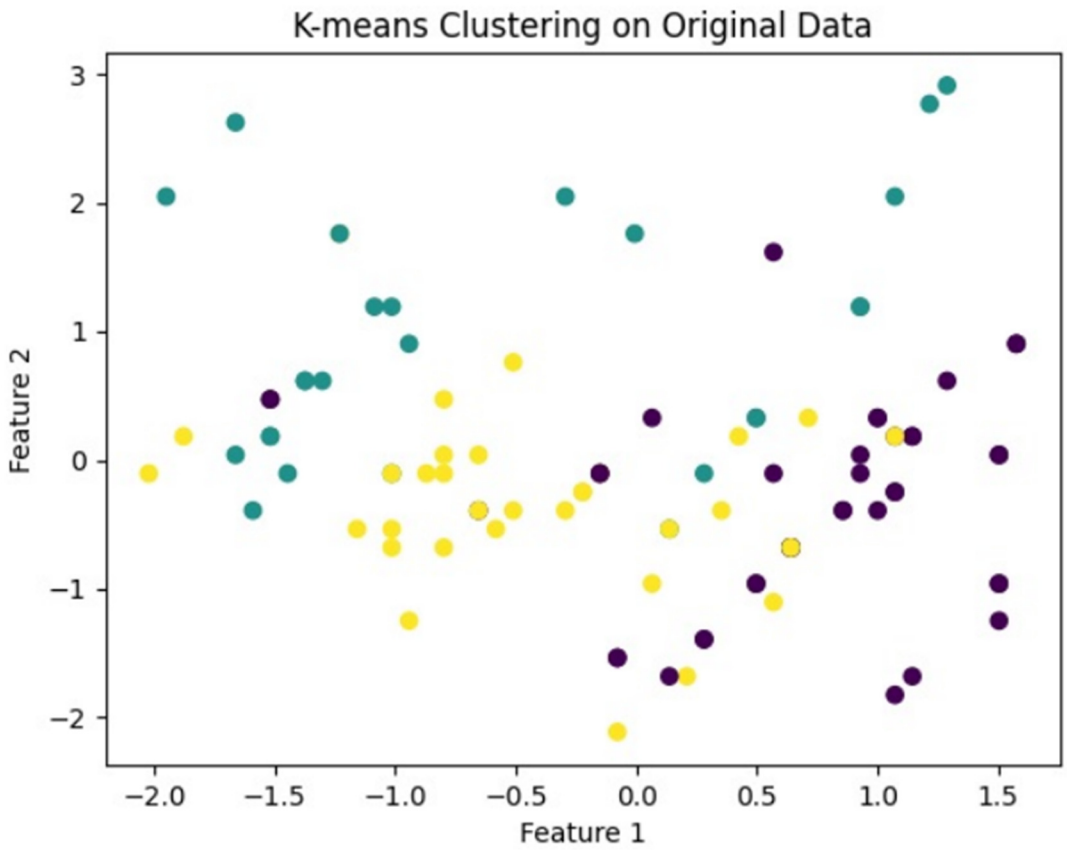
Next, the cumulative explained variance was determines as presented below.

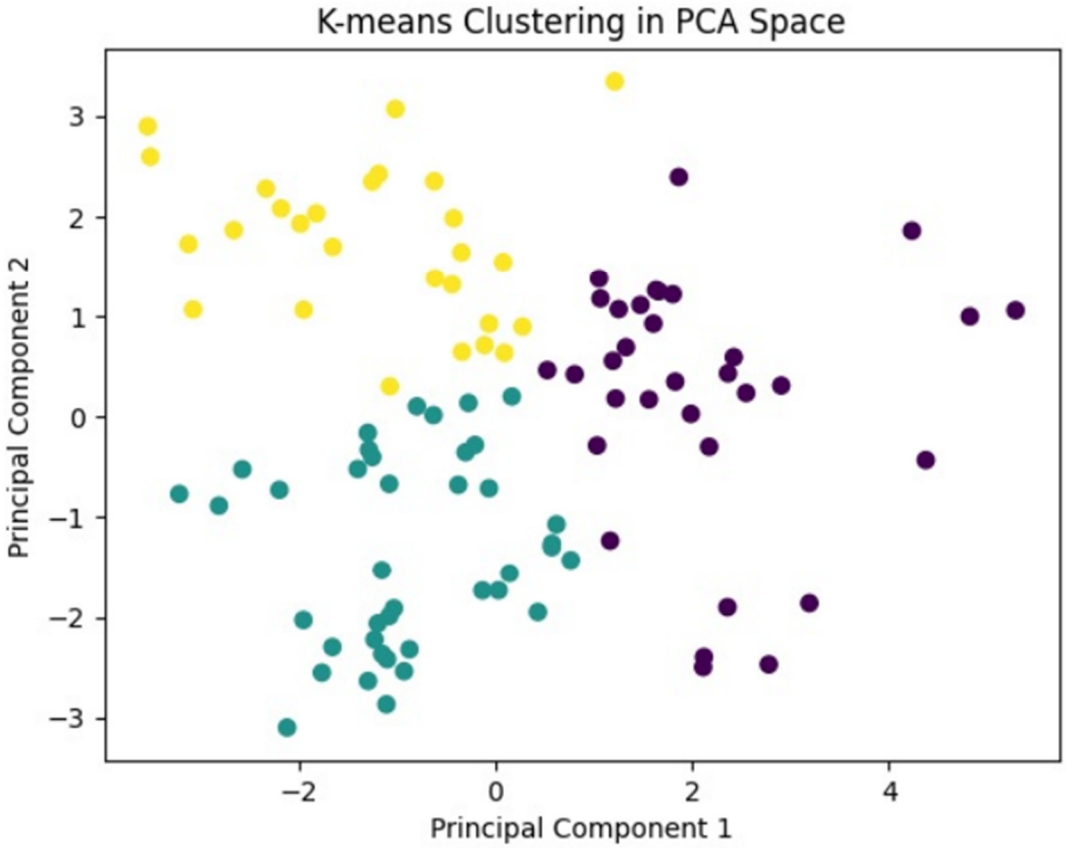
```
# Cumulative explained variance ratio
cumulative_explained_var = np.cumsum(pca.explained_variance_ratio_)

# Display results
print("Cumulative Explained Variance:", cumulative_explained_var)
```

```
Cumulative Explained Variance: [0.31154285 0.54344255]
```

The created PCA model cumulative explained variance ratio as 0.31 and 0.54.

Then, the k means clustering was applied on PCA data and original data as demonstrated below.





Quite evidently, K-Means clusters developed on PCA transformed data are much more distinct when compared with the original data. Therefore, it was stated that PCA analysis is helpful to do salon chain's market segmentation.

6 Results and Discussion

K Means Clustering suggested a clear distinction among the different groups within the dataset, and 3 profiles were obtained:

- Excellent Health and Economic Conditions
- Adequate Health and Economic Conditions
- Poor Health and Economic Conditions

Such a clear distinction is not obtained with Hierarchical Clustering. However, analysis reveals formation of similar cluster profiles. Elite, high-income groups had higher average health index scores, and vice versa. The insights derived from K-Means clustering, and to a certain extent, hierarchical clustering, offer valuable information for governmental bodies and their counterparts.

These findings suggested that marginalized communities, as well as states with lower average GDP and income levels, may encounter challenges in accessing adequate health facilities. This, in turn, contributes to a decline in their health index scores. States affected with low/medium PCI and GDP should be focused on.

Based on PCA analysis conducted on the Hair salon dataset, it determined the number of components that are best for this prediction. The results showed that the first two principal components explained more than 50% variance in the data. Next, PCA was applied for the number of decided components to get the loadings and component output. The result determined the variance ratio, principal component 1 and 2 loadings. Thus, the further obtained result showed that the created PCA model explained variance ratio as 0.31 and 0.23. Next, the cumulative explained variance was determined and explained variance ratio as 0.31 and 0.54. Then, the k means clustering was applied on PCA data and original data and stated that PCA analysis was helpful to do salon chain's market segmentation.

7 Conclusion and Recommendations

Both the datasets were analyzed successfully. Government may involve itself in creating employment opportunities in these areas. It is recommended to ensure there is improvement in economic conditions, as it can help in improving the health conditions.