**DATA MINING**

Name: Vamsi Krishna Bommu

Batch: PGPDSBA.APRIL23.B

**CONTENTS**

* Part 1
  + Exploratory Data Analysis
  + Univariate Analysis
  + Multivariate Analysis
  + Scaling the Variables
  + Covariance and Correlation Matrix After Scaling
  + Checking Outliers before and after scaling
  + Creating Covariance Matrix, Eigen Values, and Eigen Vectors
  + First Principal Component
  + Cumulative Values of the Eigen Values
  + Performing PCA and Exporting Scores
* Part 2
  + Exploratory Data Analysis
  + Clustering and Scaling
  + Hierarchical Clustering
  + K-Means Clustering
  + Elbow Curve and Silhouette Score
  + Cluster Profiles for the Cluster Defined

**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.

A screenshot of a graph

Description automatically generated

* 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. A screenshot of a computer screen

  Description automatically generated
* To inspect the dataset shape, use the shape function to determine the number of columns and rows.

A close up of a text

Description automatically generated

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

A screenshot of a computer screen

Description automatically generated

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

A group of blue and white bars

Description automatically generated

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

A graph with blue and white lines

Description automatically generated with medium confidence

* 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.

A colorful squares with white text

Description automatically generated

PCA Scaling Variables

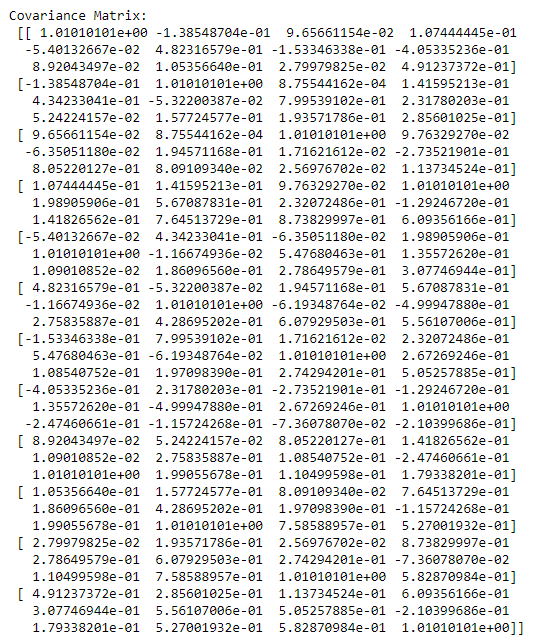
* For this case study, the variables were scaled with StandardScaler. StandardScaler was chosen because it normalizes features by removing the mean and scaling to unit variance. This scaling is useful for variables with multiple units or scales, ensuring that all variables contribute equally to the study.

A screenshot of a computer

Description automatically generated

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.



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

A screenshot of a computer

Description automatically generated

* 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.

A graph with blue and green lines

Description automatically generated

* 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.

A graph with blue lines and black dots

Description automatically generated with medium confidence

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.

A number of numbers on a white background

Description automatically generated

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.

A number with numbers on it

Description automatically generated with medium confidence

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.

A graph of a graph

Description automatically generated

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.

A screenshot of a computer screen

Description automatically generated

**Part 2:**

Data Overview:

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

A screenshot of a graph

Description automatically generated

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

A close-up of a number

Description automatically generated

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

A white background with black text

Description automatically generated

* 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.

A table with numbers and numbers

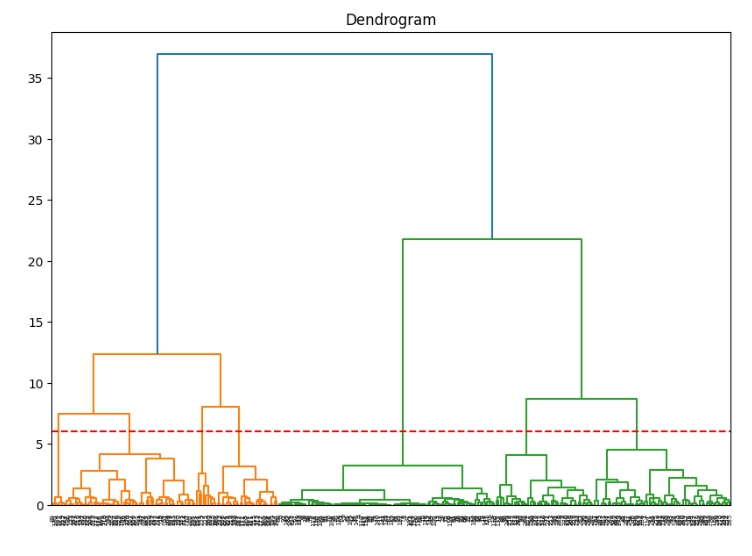
Description automatically generated

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.

A graph of a number of clusters

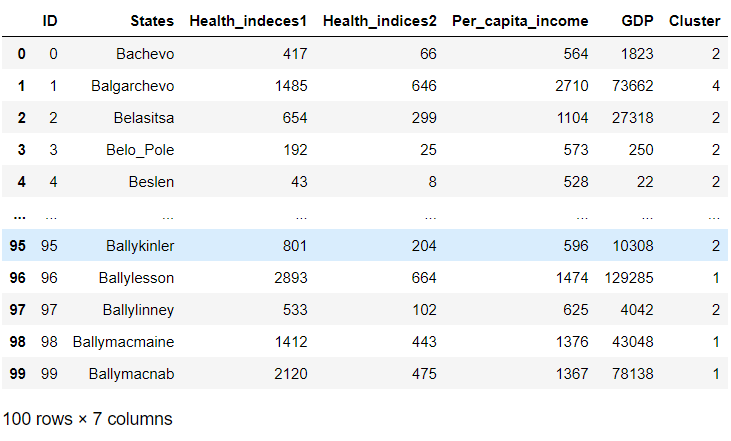
Description automatically generated

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



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.



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.