**Forecasting with AR and ARIMA Models**

**1. Upload, explore, clean, and preprocess data for clustering.**

**a. Create a rates\_df data frame by uploading the original data set into Python. Determine and present in this report the data frame dimensions, i.e., number of rows and columns. Display and present in your report the first 10 records of the rates\_df data frame.**

To create a rates\_df data frame and determine the data frame dimensions:

Rectangle

Description automatically generated with medium confidence

Therefore, the data frame dimensions are (50, 8).

To display the first 10 records of the rates\_df data frame:

Table

Description automatically generated

**b. Use Pandas to normalize the crime data, display the first 10 records of the normalized data and present the table in your report. Briefly explain how the normalized data was calculated and what it means. Why is the normalized data used in clustering instead of the original data? Briefly explain.**

To normalize the crime data using Pandas and display the first 10 records:

Graphical user interface, text, application

Description automatically generated

Outcome:

Table

Description automatically generated

The normalized data is calculated using the Z-score method which is also often called ‘standardization’. In simple words, it is calculated by subtracting the mean and dividing by the standard deviation of the feature.

In Python however, to normalize input variables (measurements), we apply preprocessing.scale from scikit-learn or standard deviation .std() from Pandas. To identify normalized distances between records, we use pairwise() from pairwise package in scikit-learn.

The Clustering algorithm utilizes normalized data instead of the original data because it evaluates the similarity between attributes by combining all the features data into a numeric outcome. Consequently, it is necessary for these features to be on the same scale. Normalizing the data ensures that all features are brought to a similar scale. By normalizing the data, we bring all the features to a similar scale. This is important because clustering algorithms are based on distance calculations, and features with larger scales could dominate the clustering process. Normalization ensures that all features contribute equally to the clustering analysis. Normalizing the data can also improves the accuracy and effectiveness of clustering algorithms.

**2. Apply hierarchical clustering to classify the states into clusters based on the normalized crime data.**

**a. Develop the hierarchical clustering (hi\_average) based on the average distance (average linkage) method (method=’average’). Develop and display the hierarchical dendrogram with the cluster threshold of 3.3 (color\_threshold=3.3). Provide the dendrogram in your report and explain how many clusters are shown on the dendrogram. Develop and present in your report the cluster membership based on the number of clusters you received in the dendrogram.**

To develop the hierarchical clustering based on the average distance method and to display the hierarchical dendrogram with the cluster threshold of 3.3:

Graphical user interface, text, application, email

Description automatically generated

The hierarchical dendrogram:

Chart

Description automatically generated

Total Number of clusters shown on the dendrogram: 5

To develop the cluster membership:

Text

Description automatically generated

Outcome:

Text

Description automatically generated

**b. Identify a data frame with the normalized mean values for each cluster and input variable. Display these data frame and provide it in your report. In addition, present in your report the profile plots of the normalized means of the clusters for the input variables. Briefly explain how the clusters can be characterized by their respective means.**

To identify a data frame with the normalized mean values for each cluster and input variable:

Graphical user interface, text, application, email

Description automatically generated

Outcome:

Table

Description automatically generated

To generate the profile plots of the normalized means of the clusters for the input variables:

Text

Description automatically generated

The Profile Plot:

Chart, line chart

Description automatically generated

The profile plot provides a visual representation of how the clusters can be characterized by their respective means for each input variable.

Through the profile plots, we can observe the relative differences in the means of the input variables across the clusters. Clusters with similar profiles and mean values for certain input variables may indicate similar characteristics or patterns in the data. We can identify clusters that have higher or lower mean values for specific input variables, which may represent distinct groups with different crime rate tendencies. Therefore, through profile plots, we can gain insights into how the clusters are characterized in terms of their respective means for the input variables.

**c. Based on the clusters’ profile plots and normalized mean values, provide cluster labeling using some common feature(s) or variable(s) means of clusters.**

Cluster Labeling based on Assaults and Burglaries:

Cluster 1: High Assault and Burglary Rates

Cluster 2: Moderate Assault and Burglary Rates

Cluster 3: Low Assault and Burglary Rates

Cluster 4: Very Low Assault and Burglary Rates

Cluster 5: Extremely Low Assault and Burglary Rates

Cluster Labeling based on Robberies and Larceny (Theft):

Cluster 1: High Robbery and Larceny Rates

Cluster 2: Moderate Robbery and Larceny Rates

Cluster 3: Low Robbery and Larceny Rates

Cluster 4: Very Low Robbery and Larceny Rates

Cluster 5: Extremely Low Robbery and Larceny Rates

Cluster Labeling based on Murder and Auto Theft:

Cluster 1: High Murder and Auto Theft Rates

Cluster 2: Moderate Murder and Auto Theft Rates

Cluster 3: Low Murder and Auto Theft Rates

Cluster 4: Very Low Murder and Auto Theft Rates

Cluster 5: Extremely Low Murder and Auto Theft Rates

Cluster Labeling based on Total Crime Rate:

Cluster 1: High Total Crime Rate

Cluster 2: Moderate Total Crime Rate

Cluster 3: Low Total Crime Rate

Cluster 4: Very Low Total Crime Rate

Cluster 5: Extremely Low Total Crime Rate

**3. Apply k-means clustering to classify the states into clusters based on the crime data.**

**a. Develop in Python k-means clustering with k = 6. Develop cluster membership for k-means clusters and provide it in your report. What are the two main differences between the algorithms used in hierarchical and k-means clustering?**

To develop k-means clustering with k = 6 and cluster membership for k-means clusters:

Graphical user interface, text, application

Description automatically generated

Outcome:

Text

Description automatically generated

The two main differences between the algorithms used in hierarchical and k-means clustering:

Hierarchical clustering provides a visual depiction of various levels of clustering. However, its non-iterative nature makes it prone to instability, as its results can greatly differ depending on the chosen settings. Additionally, hierarchical clustering can be computationally expensive.

On the other hand, k-means (non-hierarchical) clustering is computationally efficient and offers greater stability, particularly when dealing with larger datasets. However, it requires the user to specify the value of k, which represents the desired number of clusters. K-means clustering involves dividing a dataset into distinct subsets (clusters) without any overlap, ensuring that each data object belongs to only one subset. In contrast, hierarchical clustering entails organizing clusters in a hierarchical structure, forming a tree-like arrangement of nested clusters.

**b. Develop the Elbow chart for k-means clustering (k varies from 1 to 12) of the normalized crime data, present the chart in your report, and explain if k = 6 is an appropriate number of clusters in k-means clustering of the crime data.**

To develop the Elbow chart for k-means clustering of the normalized crime data:

Graphical user interface, text, application, email

Description automatically generated

Outcome:

Shape

Description automatically generated

To assess if k = 6 is an appropriate number of clusters, we can examine the Elbow chart for any distinct bend or significant drop in the Average Within-Cluster Squared Distances at k = 6. Since, there is a distinct bend or significant drop in the Average Within-Cluster Squared Distances at k = 6, it suggests that dividing the data into six clusters captures a substantial amount of variation. Therefore, it indicates that k = 6 is an appropriate number of clusters in k-means clustering of the crime data. Adding more clusters beyond 6 brings less improvement to cluster homogeneity.

**c. Identify a data frame with the normalized cluster centroids for each cluster and input variable. Display these data frame and provide it in your report. In addition, present in your report the profile plots of the normalized clusters’ centroids. Briefly explain how the clusters can be characterized by their respective centroids.**

To identify a data frame with the normalized cluster centroids for each cluster and input variable:

A screenshot of a computer code

Description automatically generated with low confidence

Outcome:

A screenshot of a computer

Description automatically generated with low confidence

To generate the profile plots of the normalized clusters’ centroids:

A screenshot of a computer code

Description automatically generated with low confidence

The profile plot:

A picture containing diagram, line, plot

Description automatically generated

The cluster centroids represent the mean values of the input variables for each cluster. By analyzing the profile plots of the normalized centroids, we can observe the relative differences in the means of the input variables across the clusters. We can identify clusters that have higher or lower centroid values for specific input variables, which may indicate distinct characteristics or patterns in the data. By examining the patterns and differences in the centroids across the clusters, we can gain insights into how the clusters are characterized in terms of their respective centroids for the input variables.

**d. Based on the k-means clusters’ profile plots and normalized centroids, provide cluster labeling using some common feature(s) or variable(s) means of clusters.**

Cluster Labeling based on Assaults and Burglaries:

Cluster 1: High Assault and Burglary Rates

Cluster 2: Moderate Assault and Burglary Rates

Cluster 3: Low Assault and Burglary Rates

Cluster 4: Very Low Assault and Burglary Rates

Cluster 5: Extremely Low Assault and Burglary Rates

Cluster Labeling based on Murder and Auto Theft:

Cluster 1: High Murder and Auto Theft Rates

Cluster 2: Moderate Murder and Auto Theft Rates

Cluster 3: Low Murder and Auto Theft Rates

Cluster 4: Very Low Murder and Auto Theft Rates

Cluster 5: Extremely Low Murder and Auto Theft Rates

Cluster Labeling based on Total Crime Rate:

Cluster 1: High Total Crime Rate

Cluster 2: Moderate Total Crime Rate

Cluster 3: Low Total Crime Rate

Cluster 4: Very Low Total Crime Rate

Cluster 5: Extremely Low Total Crime Rate

**4. Compare the clusters from parts 2 and 3. From your standpoint, which clustering, hierarchical or k-means, provides more useful insights of the states’ crime rates.**

The Profile Plot for hierarchical clustering:

Chart, line chart

Description automatically generated

The Profile Plot for k-means clustering:

A picture containing diagram, line, plot

Description automatically generated

Both hierarchical clustering and k-means clustering provide useful insights into the states' crime rates. However, the two methods differ in their approach and strengths.

Hierarchical clustering has the advantage of providing a visual representation of the clusters through the dendrogram, allowing us to see the hierarchy of the clusters and how they are related to each other. Additionally, hierarchical clustering can handle different types of distances and linkage methods, providing flexibility in how the clusters are formed. However, hierarchical clustering can be computationally expensive and may not be suitable for larger datasets.

On the other hand, k-means clustering is a faster and more computationally efficient method that is better suited for larger datasets. K-means clustering also allows us to specify the number of clusters we want to form, providing more control over the clustering process. However, k-means clustering requires us to predefine the number of clusters, and the clusters formed may be sensitive to the initial starting points.

Based on the analysis and observations, it can be concluded that K-means clustering offers a clearer and more distinct separation of the dataset compared to the hierarchical clustering method. The K-means algorithm demonstrates a stronger ability to partition the data into well-defined clusters, making it more suitable for identifying distinct groups within the states' crime rates. This conclusion is based on the observed graphs and the analysis conducted.