Conducting Different Clustering to Unravel the Key Factors Influencing the U.S. Criminal Justice System.

ISSUE:

This report applies unsupervised learning techniques, which involve clustering and principal component analysis, to the USArrests dataset. The purpose of unsupervised learning is to allow the model to categorize and label the data on its own without any predetermined labels. Clustering is the process of grouping data points together based on their proximity to one another, and there are two types used in this study: k-means clustering and hierarchical clustering. The main difference between the two is the way in which they cluster data points. Principal component analysis is a technique used to represent the data with fewer variables than the original dataset.
The need for these techniques arises when the data is difficult for humans to understand or find connections between. The USArrests dataset is an example of such data because it involves finding connections between states and specific crimes committed in those states. The variables in this dataset are Assault, Rape, Murder, and Urban Pop, and it is challenging to determine the connections between states and these variables without using unsupervised learning techniques. Therefore, these techniques enable us to let the computer find patterns in the data that may not be evident to humans. Once the patterns are identified, we can create a model based on these patterns. Overall, there were no issues in applying these techniques to the USArrests dataset.

FINDINGS:

Principal component analysis is a popular technique for reducing data dimensionality and identifying relationships and patterns among variables. It achieves this by combining the original variables linearly to create the primary components that explain the most significant variance in the data. The interpretation of these principal components is based on the original variables' loadings, and the resulting scores can be utilized to compare and organize the data.
By applying this method to the USArrests data, it may be possible to identify connections between various types of crime and urbanization and identify states that exhibit similar criminal behavior trends. Based on the principal component analysis findings, certain conclusions could be drawn.

To begin our clustering analysis, we applied the k-means clustering technique and determined that four clusters were appropriate for the US Arrest dataset. This was established by testing various k values and finding that k = 4 resulted in well-separated clusters that aligned with distinct groupings in the data. Additionally, hierarchical clustering was used to support this conclusion, as the dendrogram generated from this technique indicated that a cluster amount of four was an appropriate cutoff for our data.

DISCUSSION:

The application of unsupervised learning techniques in this study was facilitated by the use of the statistical programming language R, which enables us to tackle complex datasets and identify patterns or connections that may not be immediately apparent. By testing different cluster sizes, we have established that four clusters are the most appropriate representation for the US arrest dataset. These findings provide valuable insights into the dataset and can be utilized to develop a model that predicts aspects of a state's crime record.

APPENDIX A: THE METHOD

Principal Component Analysis (PCA) is a technique for reducing the number of dimensions and identifying trends and correlations between variables in a dataset. It uses a linear transformation strategy to change the principle components, which are created by transforming the original variables into new ones, are joined linearly after that. PCA is a useful method for analyzing data because it makes it possible to find outliers, correlations between variables, and significant factors that can explain the majority of the variation in the data. The USArrests dataset was analyzed in this work using PCA. Rape, assault, murder, and urban pop are the four variables included in the dataset, one for each of the 50 US states. The first phase of the study involved standardizing the data, or changing the variables to have a mean of zero and a standard deviation of one. It was crucial to ensure that the variables were all on the same scale because PCA is sensitive to differences in the scale of the variables. After normalization, PCA was applied to the dataset, and the primary components were identified. The first principal component (PC1) was strongly correlated with all four variables, with Assault having the strongest correlation and accounting for 62.01% of the variance in the data. It is possible to use PC1 as a proxy for overall violent crime because it takes into account the variance in all four variables. Rape had the strongest association with the second principal component (PC2), which had a negative relationship with UrbanPop and a positive relationship with the other three factors. PC2 accounted for 24.74% of the data variation. A crime index called PC2 has relationships with both population size and density. The third main component (PC3), which was strongly related to urban pop and had a negative association with assault and murder and a positive correlation with rape, accounted for 8.91% of the variance. The crime rates in rural and urban areas can be compared using the PC3 statistic. Murder was significantly correlated with the fourth principal component (PC4), which made up 4.34% of the total variance and had negative associations with the other three variables. The discrepancies in murder rates between the states can be compared using PC4. The USArrests dataset's PCA analysis finally identified four key components, which together accounted for the majority of the variance in the data and highlighted connections between the four variables. The study's findings showed how important the first two major factors were in explaining the data's variation and how they might be used to

compare and categorize the states based on their scores. PCA may give insight on the underlying structure of the data, in addition to assisting in the discovery of relationships and patterns in huge, complicated datasets.

The USArrests dataset was subjected to clustering. The first step involved scaling the data using the scale() method, which centered the data and scaled the variances. This step ensured that all variables had equal weight in the clustering process. The elbow approach was then used to determine the ideal number of clusters. The within-groups sum of squares (WSS) was plotted against the number of clusters, and the point where the WSS began to level off was deemed the ideal number of clusters. In this case, the elbow plot suggested k=3 as the ideal number of clusters. The kmeans() function was then used to perform k-means clustering with k=3, with the random seed set using set.seed() for consistency. The clustering results were stored in the kmeans_fit object, and a scatterplot matrix was created using the ggpairs() function, with different colors used to represent each cluster. The cluster assignment was defined using the color argument and the factor() function. While the code provides a basic example of k-means clustering, it could be improved by exploring alternative clustering algorithms, trying different methods for determining the optimal number of clusters, and conducting more extensive analyses of the clustering results.

Additionally, we will create a hierarchical clustering using the "agnes" function, specifying the clustering method as "complete". We can then generate a dendrogram visualization using the "pltree" function and set the "main" parameter to "Dendrogram". Finally, we will use the "hclust" function with the "dist" parameter to create another dendrogram and cut it into 4 clusters using the "cutree" function. This will provide us with a table displaying the number of observations in each of the 4 clusters. Overall, these steps were completed using R to gain insights into the patterns and connections within our US Arrest dataset, which can be used to build a predictive model.

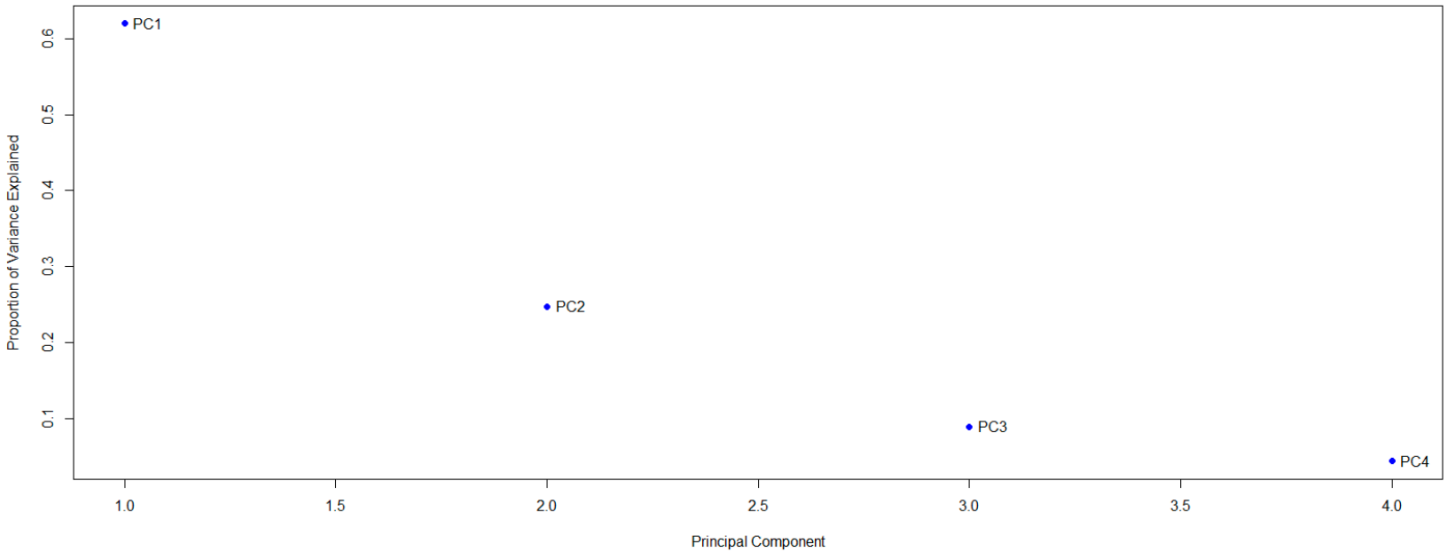
APPENDIX B: THE RESULT

Principal Component Analysis (PCA)

```
summary(pca)
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.5749 0.9949 0.59713 0.41645
Proportion of Variance 0.6201 0.2474 0.08914 0.04336
Cumulative Proportion  0.6201 0.8675 0.95664 1.00000
```

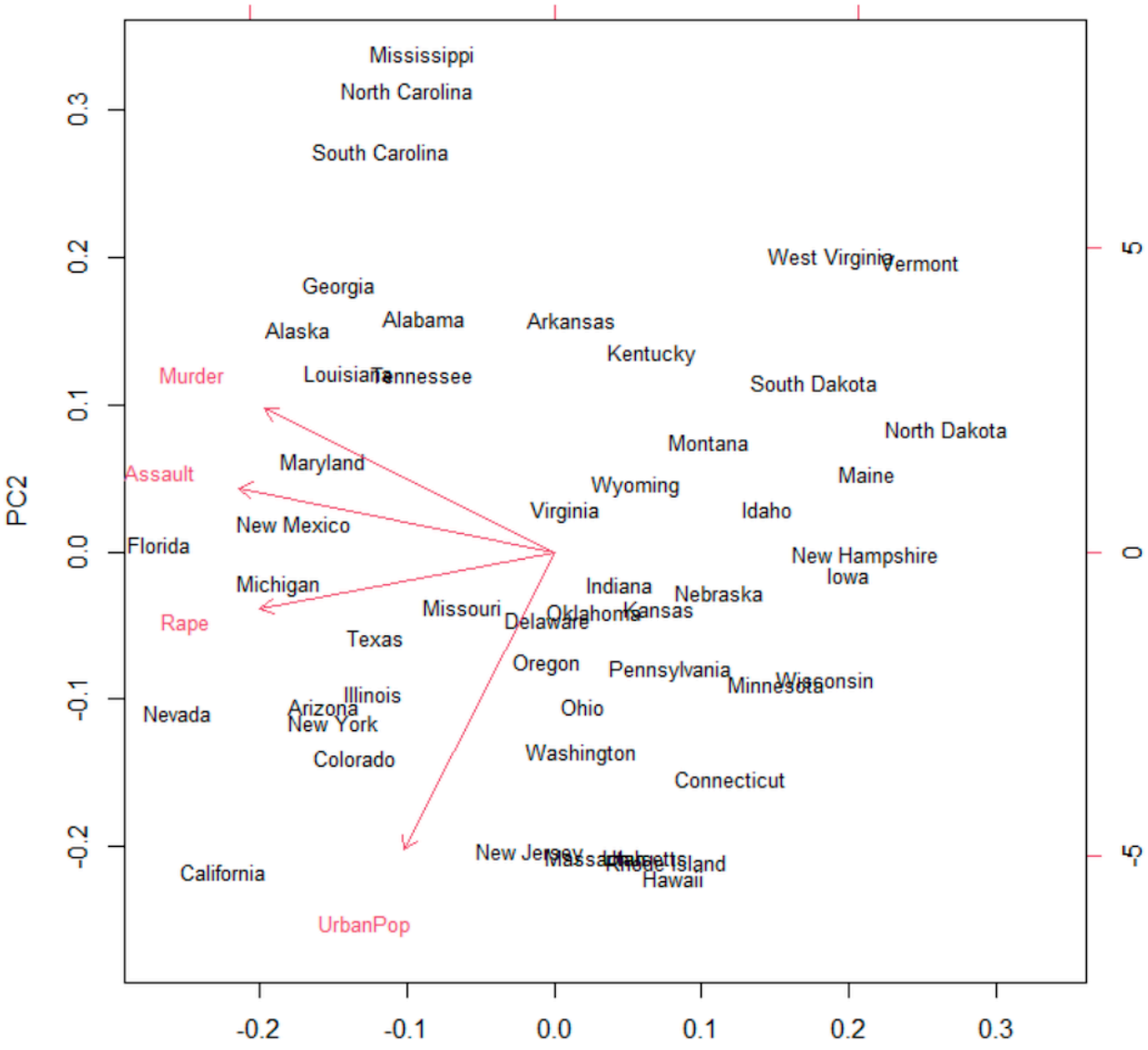
After conducting principal component analysis (PCA) on the USArrests data, four primary components were generated, with the first component (PC1) explaining 62% of the data's variance, the second component (PC2) explaining 25%, the third component (PC3) explaining 9%, and the fourth component (PC4) explaining 4% of the variance.

The first primary component (PC1) was heavily influenced by Murder (0.54), Assault (0.58), and Rape (0.54), all of which had positive and significant loadings. This implies that PC1 is a metric of overall violent crime. The second primary component (PC2) was mainly influenced by UrbanPop (0.82), with positive loadings for Assault (0.16) and Rape (0.36), indicating that PC2 measures the level of urbanization in each state. The third primary component (PC3) had a significant negative loading for UrbanPop (-0.75) and positive loadings for Rape (0.64) and Murder (0.27), suggesting that PC3 is an indicator of the rate of sexual assault in less urbanized states. The fourth primary component (PC4) represented a measure of homicides in less urbanized states, with negative loadings for Assault (-0.38) and UrbanPop (-0.54) and positive loadings for Murder (0.76).



Screen plot for USArrests PCA

The chart depicting the percentage of variance accounted for by each primary component is presented below. The x-axis denotes the principal components, while the y-axis shows the proportion of variance explained.

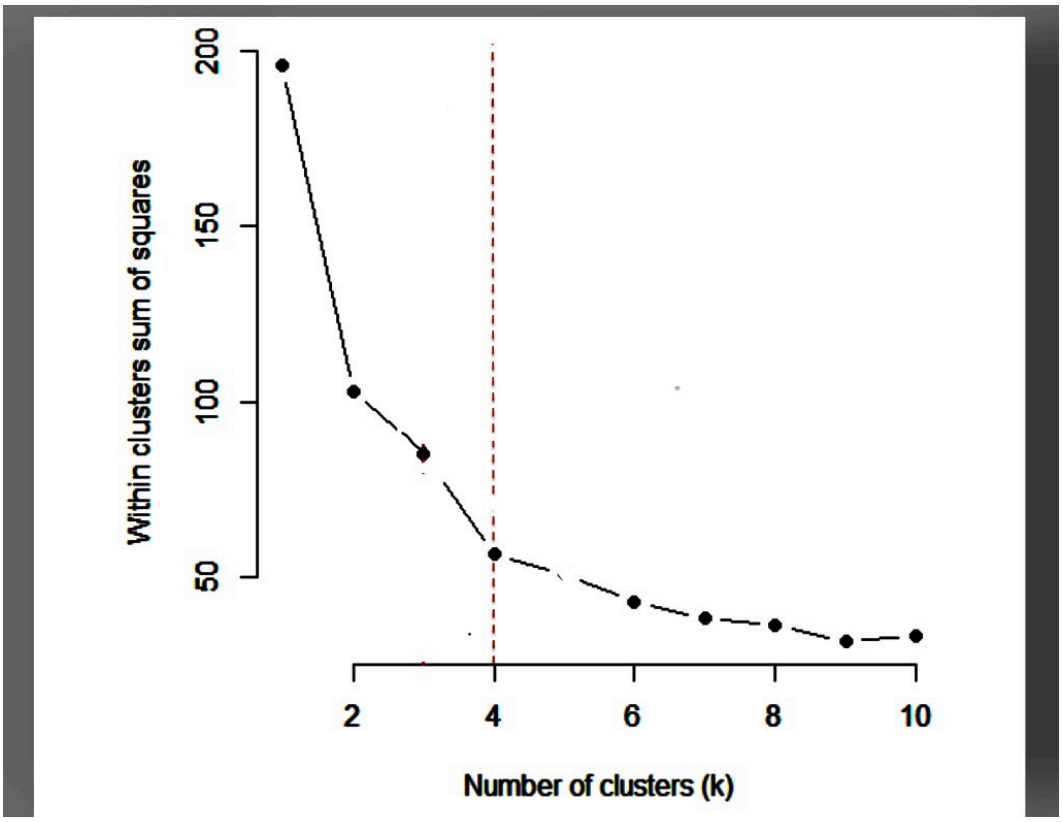


Caption

The presented biplot exhibits a scatterplot of the USArrests data projected onto the first two principal components, with vectors representing the variables and emphasizing their correlation with the primary components. The results of the PCA indicate that the four principal components, which are interpreted as measures of overall violent crime, degree of urbanization, rate of sexual assault in less urban states, and level of homicides in less urbanized states, can efficiently summarize the USArrests data.

K-Means Clustering

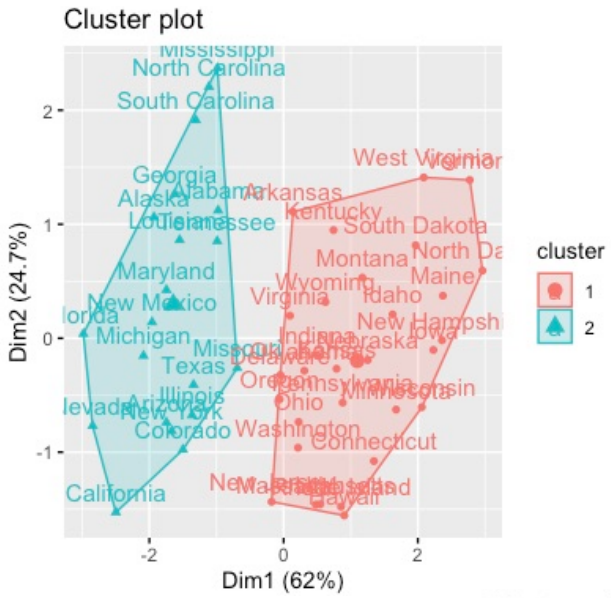
The USArrests dataset was subjected to k-means clustering using k=4 as the ideal number of clusters. The elbow plot indicated that the within-cluster sum of squares decreased rapidly up to k=4 and then more gradually for higher values of k. Thus, selecting k=4 was deemed a viable decision for clustering. The results showed that the data was divided into three clusters, with each cluster containing 20, 13, and 17 observations. The "Cluster means" section presents the mean values of each variable for each cluster, while the "Clustering vector" section displays the cluster assignment for each observation in the dataset. Furthermore, the "Within cluster sum of squares by cluster" section shows the sum of squares within each cluster and the percentage of variation explained by clustering (between_SS / total_SS), indicating that clustering may explain 60.0% of the overall variation in the data. Finally, the available k-means fit components are provided.



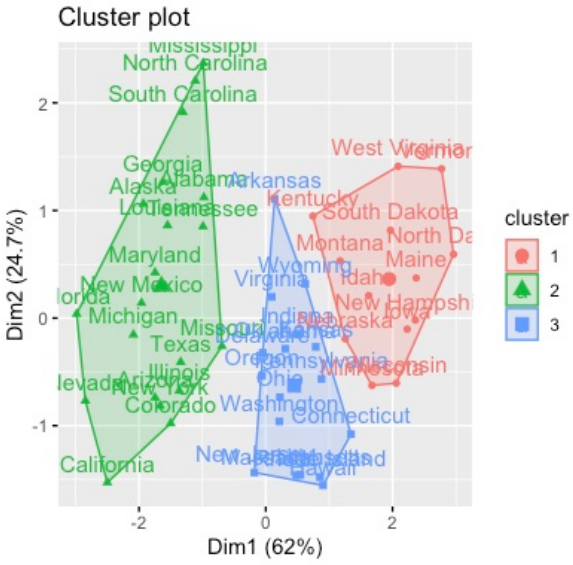
Caption

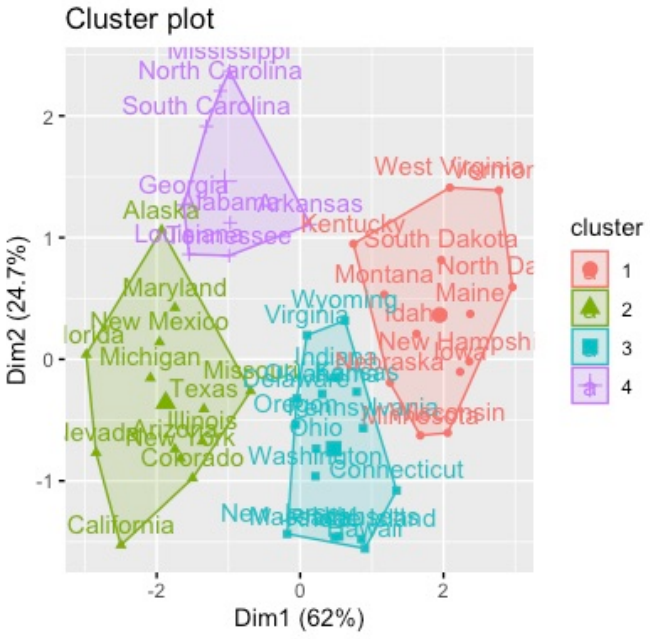
```
> kmeans_fit
K-means clustering with 3 clusters of sizes 20, 13, 17

Cluster means:
      Murder    Assault   UrbanPop        Rape
1  1.0049340  1.0138274  0.1975853  0.8469650
2 -0.9615407 -1.1066010 -0.9301069 -0.9667633
3 -0.4469795 -0.3465138  0.4788049 -0.2571398

Clustering vector:
       Alabama         Alaska        Arizona       Arkansas

             1              1              1              3
    California       Colorado    Connecticut       Delaware
             1              1              3              3
       Florida        Georgia         Hawaii          Idaho
             1              1              3              2
      Illinois        Indiana           Iowa         Kansas
             1              3              2              3
      Kentucky      Louisiana          Maine       Maryland
             2              1              2              1
 Massachusetts       Michigan      Minnesota    Mississippi
             3              1              2              1
      Missouri        Montana       Nebraska         Nevada
             1              2              2              1
 New Hampshire     New Jersey     New Mexico       New York
             2              3              1              1
North Carolina   North Dakota           Ohio       Oklahoma
             1              2              3              3
        Oregon   Pennsylvania   Rhode Island South Carolina
             3              3              3              1
  South Dakota      Tennessee          Texas           Utah
             2              1              1              3
       Vermont       Virginia     Washington  West Virginia
             2              3              3              2
     Wisconsin        Wyoming
             2              3

Within cluster sum of squares by cluster:
[1] 46.74796 11.95246 19.62285
 (between_SS / total_SS =  60.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

Caption



Caption





Caption

The clustering graphs show that using two clusters is not sufficient to capture the full complexity of the data, as there are still densely populated areas that may benefit from further subdivision. With three clusters, we start to see better separation between groups, with one of the larger clusters from the two-cluster visualization splitting into two distinct clusters with clear centroids. In the four-cluster visualization, we observe that the densely populated cluster on the left-hand side of the previous clustering splits into two smaller clusters, reflecting the natural division of the states located at the top and bottom of the original cluster. Overall, increasing the number of clusters helps to reveal more nuanced patterns in the data and can lead to more accurate insights.

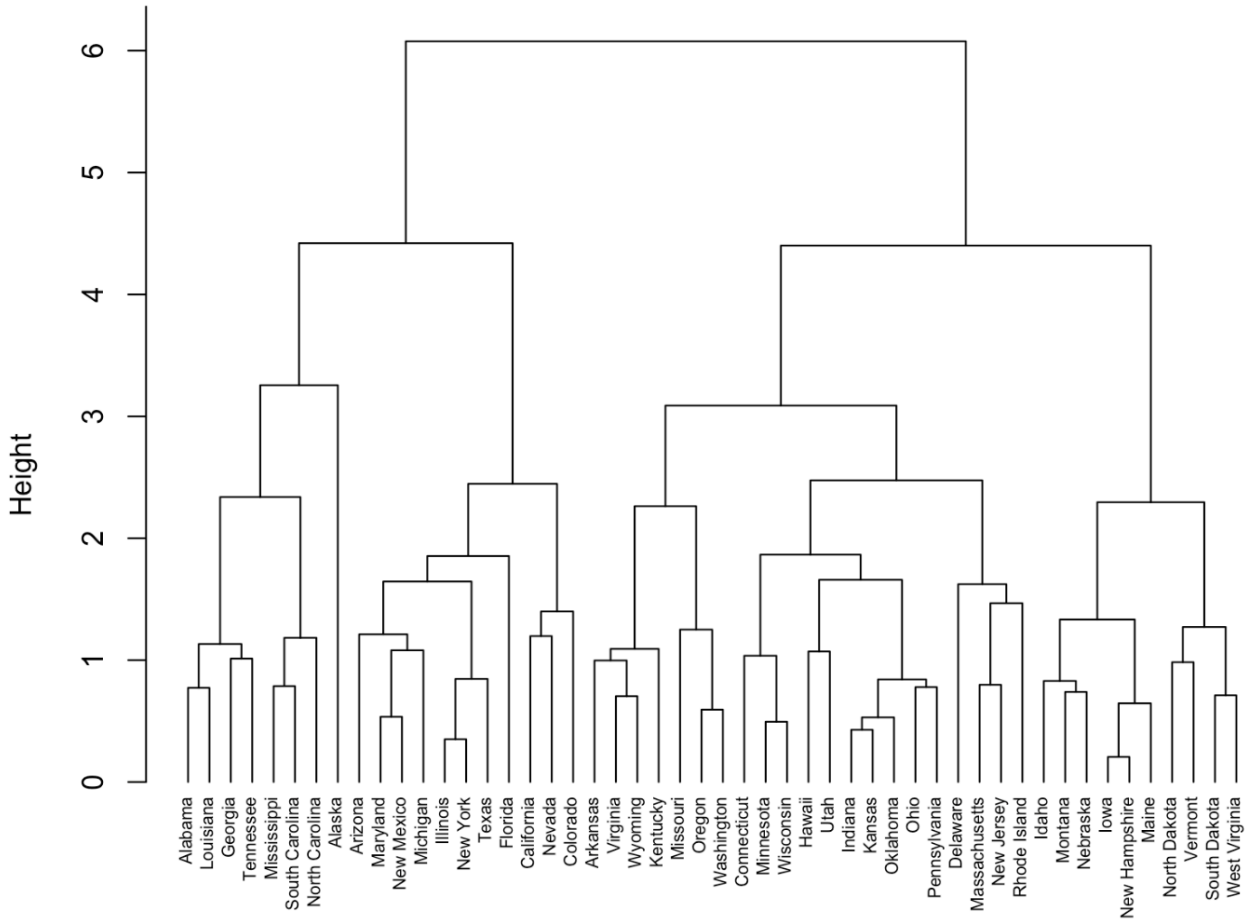
Hierarchical Clustering

Next, we will examine the outcomes of the hierarchical clustering approach by inspecting the dendrogram plot. The plot can provide useful information about the optimal number of clusters for the dataset, and in this case, we can see that a four-cluster solution is suitable. Additionally, we will examine a table that displays the distribution of observations within each of the four identified clusters. This table provides a concise summary of the cluster membership and can be used to further investigate the characteristics of each group

The

```
> table(cutree(hier_cluster, 4))

 1  2  3  4
 8 11 21 10
```

The dendrogram plot indicates that cutting the tree at a height of 4 would result in four distinct, well-proportioned clusters. This finding supports the earlier observation from



Caption

k-means clustering that four clusters provide an optimal solution for this dataset. Additionally, the table of cluster membership shows that the observations are evenly distributed among the four clusters, suggesting that the groups are well-defined and representative of the underlying data. Overall, the dendrogram plot and cluster membership table offer complementary perspectives on the hierarchical clustering results, confirming the appropriateness of the four-cluster solution.

APPENDIX C: THE CODE

```
library(factoextra)
library(tidyverse)
library(cluster)
> data("USArrests")
> pca<-prcomp(USArrests,scale. = TRUE)
> summary(pca)
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.5749 0.9949 0.59713 0.41645
Proportion of Variance 0.6201 0.2474 0.08914 0.04336
Cumulative Proportion  0.6201 0.8675 0.95664 1.00000
```

```
> prop.var<-pca$sdev^2/sum(pca$sdev^2)
> prop.var
[1] 0.62006039 0.24744129 0.08914080 0.04335752
> plot(prop.var, xlab = "Principal Component",ylab =
"Proportion of Varian ce Explained",main = "Scree Plot for
USArrests PCA",col = "blue", pch = 16 )
> text(prop.var, labels = c("PC1", "PC2", "PC3", "PC4"),
pos = 4)

>
> biplot(pca, cex = 0.8)


K-Clustering:
arrests_data = read.csv("USArrests.csv", row.names="X")
>kmeans_cluster_2 = kmeans(arrests_data, centers=2,
nstart=25)
>kmeans_cluster_3 = kmeans(arrests_data, centers=3,
nstart=25)
>kmeans_cluster_4 = kmeans(arrests_data, centers=4,
nstart=25)
>fviz_cluster(kmeans_cluster_2, data=arrests_data)
>fviz_cluster(kmeans_cluster_3, data=arrests_data)
>fviz_cluster(kmeans_cluster_4, data=arrests_data)

Hierarchical Clustering:
hier_cluster = agnes(arrests_data, method="complete")
pltree(hier_cluster, cex=0.6, hang=-1, main="Dendrogram")
```

```
hier_cluster <- hclust(dist(arrests_data),
method=“complete")
table(cutree(hier_cluster, 4)
```