CLUSTERING

Issue:

The "USArrests" dataset is a collection of data about different characteristics of each state in the United States, including murder rate, assault rate, urban population percentage, rape rate, and the names of the states.To analyze this dataset, we can use various statistical techniques, including principal component analysis (PCA), k-means clustering, and hierarchical clustering.

PCA is a technique used to reduce the dimensionality of a dataset by identifying the most important features that capture the majority of the variance in the data. In the case of "USArrests" dataset, we can use PCA to identify the principal components that explain most of the variation in the dataset. We can then interpret these principal components to gain insights into the relationships between the different variables in the dataset.

K-means clustering is a technique used to partition a dataset into k clusters, where k is a user-defined parameter. In the case of "USArrests" dataset, we can use k-means clustering to group the states based on their similarities in terms of the different characteristics provided in the dataset. We can then interpret these clusters to identify any underlying patterns or relationships between the different variables in the dataset.

Hierarchical clustering is another clustering technique that creates a hierarchy of clusters based on the similarity between data points. In the case of "USArrests" dataset, we can use hierarchical clustering to identify the different levels of clusters and their relationships with each other. We can then interpret these clusters to identify any underlying patterns or relationships between the different variables in the dataset.

Overall, by applying these statistical techniques to the "USArrests" dataset, we can gain insights into the relationships between the different characteristics of each state and identify any underlying patterns or clusters in the data.

Findings:

The results of the Principal Component Analysis (PCA) showed that there is a strong relationship between serious crimes, such as murder, assault, and rape, and the degree of urbanization in the first loading vector. However, this correlation is weaker in the second loading vector. Therefore, it can be concluded that these crimes tend to occur together in states, but there is little connection between these crimes and the urban population.

Using k-means clustering, it was found that increasing the value of "k" resulted in a decrease in "tot.withinss," indicating that the algorithm was successful in creating clusters that were more homogeneous and well-separated from each other.

Additionally, applying hierarchical clustering to the data resulted in dendrograms, which are tree-like structures. When using complete and average linkage methods, the dendrograms were more evenly distributed. This suggests that the hierarchical clustering algorithm was able to identify distinct patterns or clusters within the data that could be interpreted in a meaningful way.

Discussions:

The results of the principal component analysis showed that the first principal component explains 62.0% of the variance in the data, while the second principal component explains 24.7%, and so on. This indicates that the first principal component is the most important factor in explaining the variation in the data.

On the other hand, when applying hierarchical clustering to the data, it was found that the dendrograms produced using complete and average linkage methods were more balanced than the dendrogram produced using single linkage. This suggests that complete and average linkage methods were able to identify distinct patterns or clusters within the data that could be interpreted in a meaningful way, while the single linkage method failed to do so.

It is important to note that the choice of linkage method can have a significant impact on the results of hierarchical clustering, and researchers should carefully consider which method to use based on the nature of the data and their research questions.

Appendix A: Method

Principal Component Analysis (PCA) is a technique that can be used to reduce the dimensionality of a dataset by identifying patterns in the data. To perform PCA in R, we first compute the mean and variance of the dataset's variables. If the variables' mean and variance values differ, we use the prcomp() function to perform PCA. This function provides the center, scale, rotation, sdev, and x values. The scale's center and length show the variables' mean and standard deviation. We can then use the biplot() function to plot the first two principal

components. We can also determine the variance and the percentage of variance explained by each principal component and visualize it using the plot() method.

To perform k-means clustering in R, we start with a matrix-formatted dataset that has been divided into two equal halves. We cluster the data using R's kmeans() function and then plot the data, giving each observation a color based on which cluster it belongs to. We repeat this procedure with k=3 and note the tot.withinss value, which is a metric for the sum of all within-cluster squares. Based on our observations, we can make inferences about the effectiveness of the clustering algorithm.

Appendix B: Results

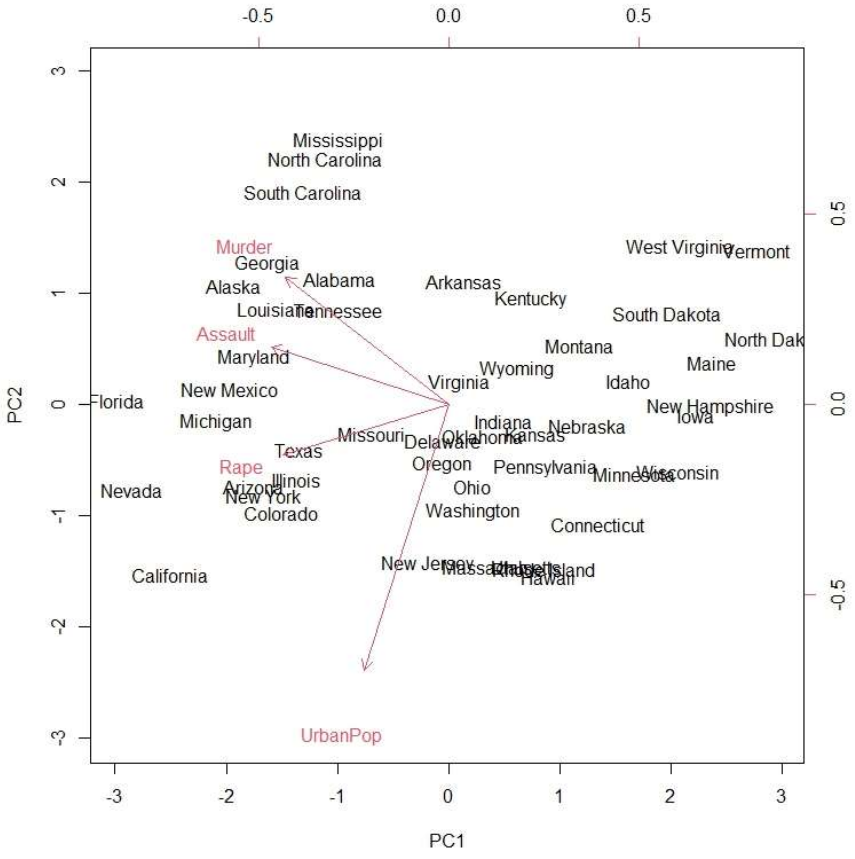

Fig1 : Biplot for first two principal components

```
names(USArrests)
[1] "Murder"   "Assault"  "UrbanPop" "Rape"
> apply(USArrests,2,mean)
  Murder  Assault UrbanPop     Rape
```

```
   7.788  170.760   65.540   21.232
> apply(USArrests,2,var)
    Murder    Assault   UrbanPop       Rape
  18.97047 6945.16571  209.51878   87.72916
> pr.out<- prcomp (USArrests , scale = TRUE)
> names(pr.out)
[1] "sdev"     "rotation" "center"   "scale"    "x"
> pr.out$center
  Murder  Assault UrbanPop     Rape
   7.788  170.760   65.540   21.232
> pr.out$scale
   Murder   Assault  UrbanPop      Rape
 4.355510 83.337661 14.474763  9.366385
> pr.out$rotation
                PC1        PC2        PC3         PC4
Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

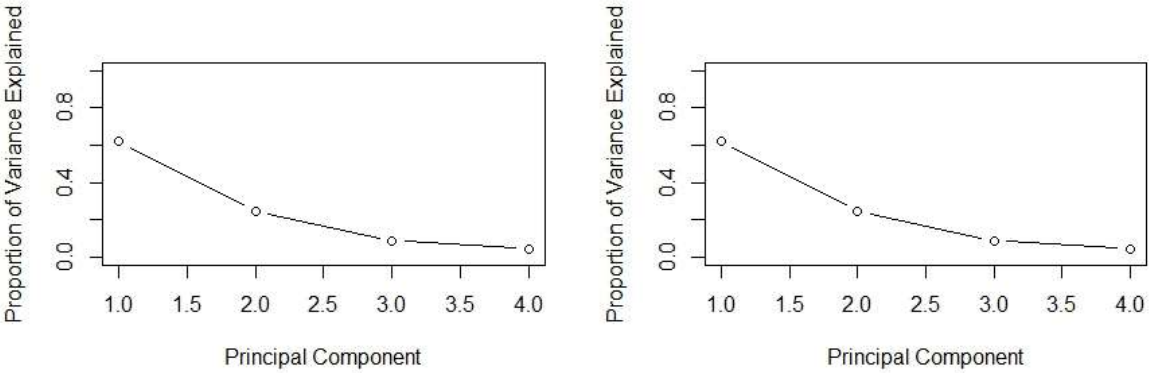

Fig2: Variance Plots

```
#Kmeans clustering
> set.seed (2)
> x <- matrix ( rnorm (50 * 2), ncol = 2)
> x[1:25, 1] <- x[1:25, 1] + 3
> x[1:25, 2] <- x[1:25, 2] - 4
> km.out <- kmeans (x, 2, nstart = 20)
> km.out$cluster
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2
[29] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
> par (mfrow = c(1, 2))
> plot (x, col = (km.out$cluster + 1)
+         ,main = "K- Means Clustering Results with K = 2"
+         ,xlab = "", ylab = "", pch = 20, cex = 2)
> set.seed (4)
> km.out <- kmeans (x, 3, nstart = 20)
> km.out
K-means clustering with 3 clusters of sizes 17, 23, 10

Cluster means:
        [,1]        [,2]
1  3.7789567 -4.56200798
2 -0.3820397 -0.08740753
3  2.3001545 -2.69622023

> plot (x, col = (km.out$cluster + 1),
+       main = "K- Means Clustering Results with K = 3",
+       xlab = "", ylab = "", pch = 20, cex = 2)
> set.seed (4)
> km.out <- kmeans (x, 3, nstart = 1)
> km.out$tot.withinss
[1] 104.3319
```

```
> km.out <- kmeans (x, 3, nstart = 20)
> km.out$tot.withinss
[1] 97.97927
```

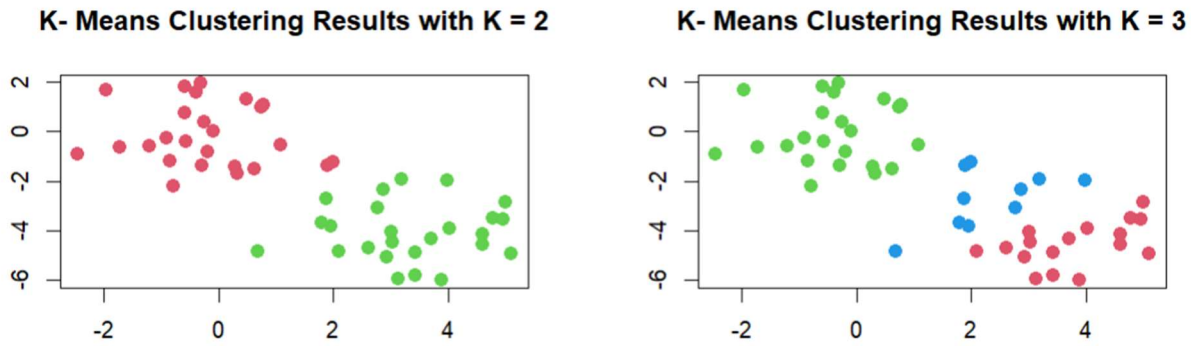

Fig3: Plots for K-Means clustering for different k values

```
> hc.complete <- hclust ( dist (x), method = "complete")
> hc.average <- hclust ( dist (x), method = "average")
> hc.single <- hclust ( dist (x), method = "single")
> par (mfrow = c(1, 3))
> plot (hc.complete, main = "Complete Linkage",
+         xlab = "", sub = "", cex = .9)
> plot (hc.average , main = "Average Linkage",
+         xlab = "", sub = "", cex = .9)
> plot (hc.single, main = "Single Linkage",
+         xlab = "", sub = "", cex = .9)
> cutree (hc.complete,2)
> cutree (hc.complete,2)
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2
[29] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
> cutree (hc.average,2)
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2
[29] 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 1 2 1 2 2 2 2
> cutree (hc.single,2)
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1
[29] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
> cutree (hc.single,4)
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 3 3 3
[29] 3 3 3 3 3 3 3 3 3 3 3 3 3 4 3 3 3 3 3 3 3 3
> xsc <- scale (x)
> plot ( hclust ( dist (xsc), method = "complete")
+          ,main = " Hierarchical Clustering with Scaled Features")
> x <- matrix ( rnorm (30 * 3), ncol = 3)
> dd <- as.dist (1 - cor (t(x)))
> plot ( hclust (dd, method = "complete")
+          ,main = "Complete Linkage with Correlation - Based Distance "
+          ,xlab = "", sub = "")
```

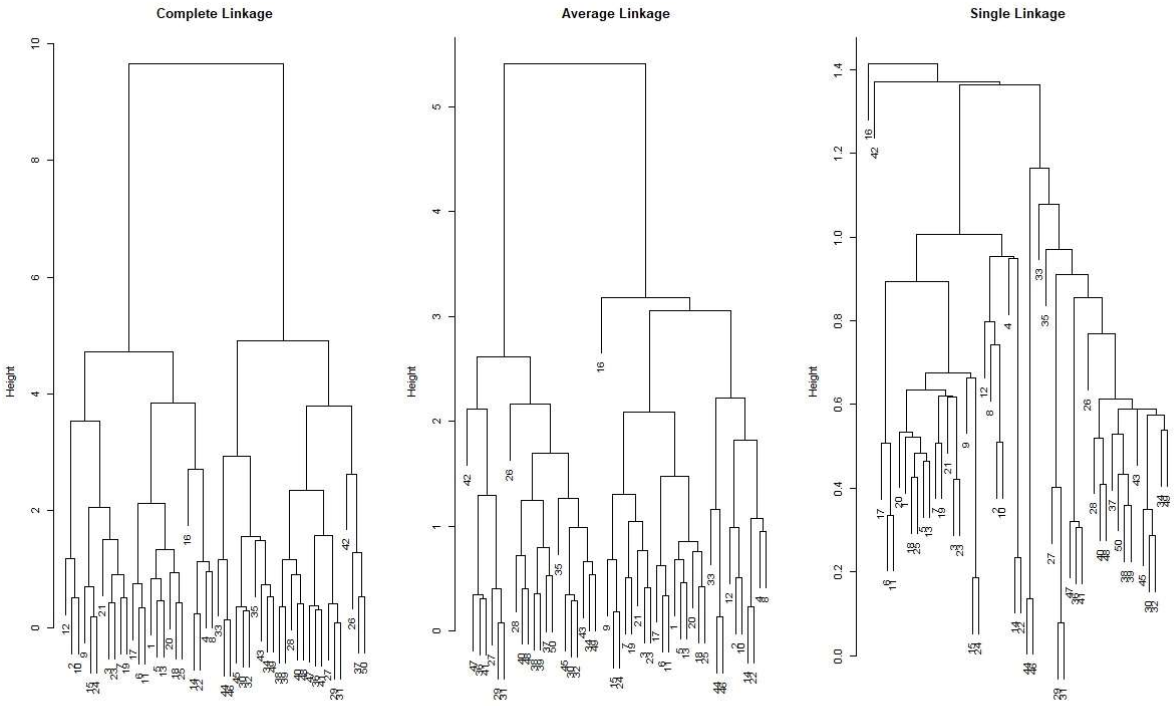

Fig4: Hierarchical clustering Performing complete, average and single linkage

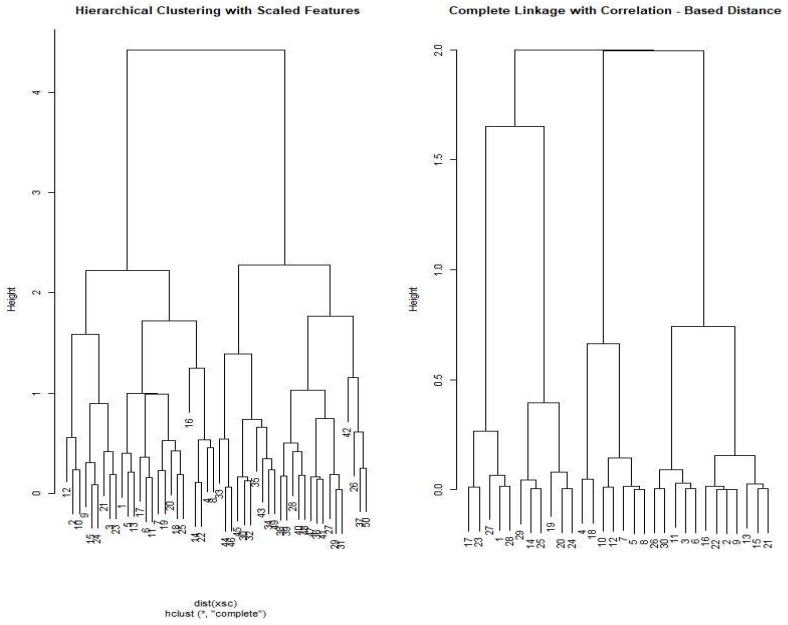

Fig5: Hierarchical clustering with Scaled features

Appendix C: Code

Code for Principal Component Analysis

```
states <- row.names(USArrests)
states
names(USArrests)
apply(USArrests,2,mean)
apply(USArrests,2,var)
pr.out<- prcomp (USArrests , scale = TRUE)
names(pr.out)
pr.out$center
pr.out$scale
pr.out$rotation
dim(pr.out$x)
biplot(pr.out,scale= 0)
pr.out$rotation = -pr.out$rotation
pr.out$x = -pr.out$x
biplot(pr.out,scale = 0)
pr.out$sdev
pr.var <- pr.out$sdev^2
pr.var
pve <- pr.var / sum (pr.var)
pve
par (mfrow = c(1, 2))
plot (pve , xlab = " Principal Component ", ylab = " Proportion of Variance Explained ", ylim = c(0, 1) ,type = "b")
plot ( cumsum (pve), xlab = " Principal Component ",ylab = " Cumulative Proportion of Variance Explained ",ylim = c(0, 1), type = "b")
a <- c(1, 2, 8, -3)
```

```
cumsum (a)
```

#Code for K-means clustering

```
set.seed (2)
x <- matrix ( rnorm (50 * 2), ncol = 2)
x[1:25, 1] <- x[1:25, 1] + 3
x[1:25, 2] <- x[1:25, 2] - 4
km.out <- kmeans (x, 2, nstart = 20)
km.out$cluster
par (mfrow = c(1, 2))
plot (x, col = (km.out$cluster + 1)  ,main = "K- Means Clustering Results with K = 2",xlab = "", ylab = "", pch = 20, cex = 2)
set.seed (4)
km.out <- kmeans (x, 3, nstart = 20)
km.out
plot (x, col = (km.out$cluster + 1),main = "K- Means Clustering Results with K = 3",xlab = "", ylab = "", pch = 20, cex = 2)
set.seed (4)
km.out <- kmeans (x, 3, nstart = 1)
km.out$tot.withinss
km.out <- kmeans (x, 3, nstart = 20)
km.out$tot.withinss
```

Code for Hierarchical Clustering

```
hc.complete <- hclust ( dist (x), method = "complete")
hc.average <- hclust ( dist (x), method = "average")
hc.single <- hclust ( dist (x), method = "single")
par (mfrow = c(1, 3))
```

```
plot (hc.complete, main = "Complete Linkage",xlab = "", sub = "", cex = .9)

plot (hc.average , main = "Average Linkage",  xlab = "", sub = "", cex = .9)

plot (hc.single, main = "Single Linkage", xlab = "", sub = "", cex = .9)

cutree (hc.complete,2)

cutree (hc.average,2)

cutree (hc.single,2)

cutree (hc.single,4)

xsc <- scale (x)

plot ( hclust ( dist (xsc), method = "complete")   ,main = " Hierarchical Clustering with Scaled Features")

x <- matrix ( rnorm (30 * 3), ncol = 3)

dd <- as.dist (1 - cor (t(x)))

plot ( hclust (dd, method = "complete") ,main = "Complete Linkage with Correlation - Based Distance ",xlab = "", sub = "")
```