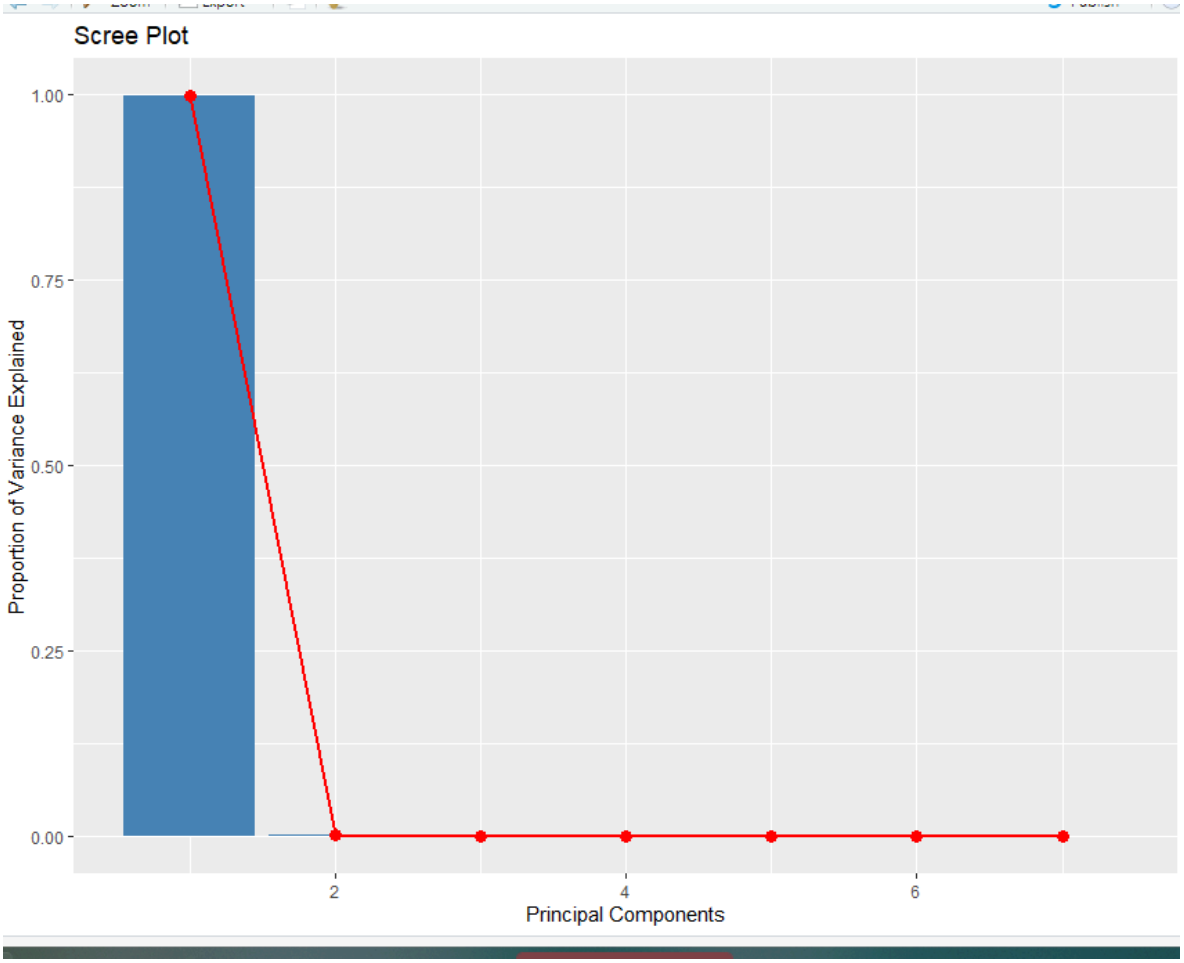
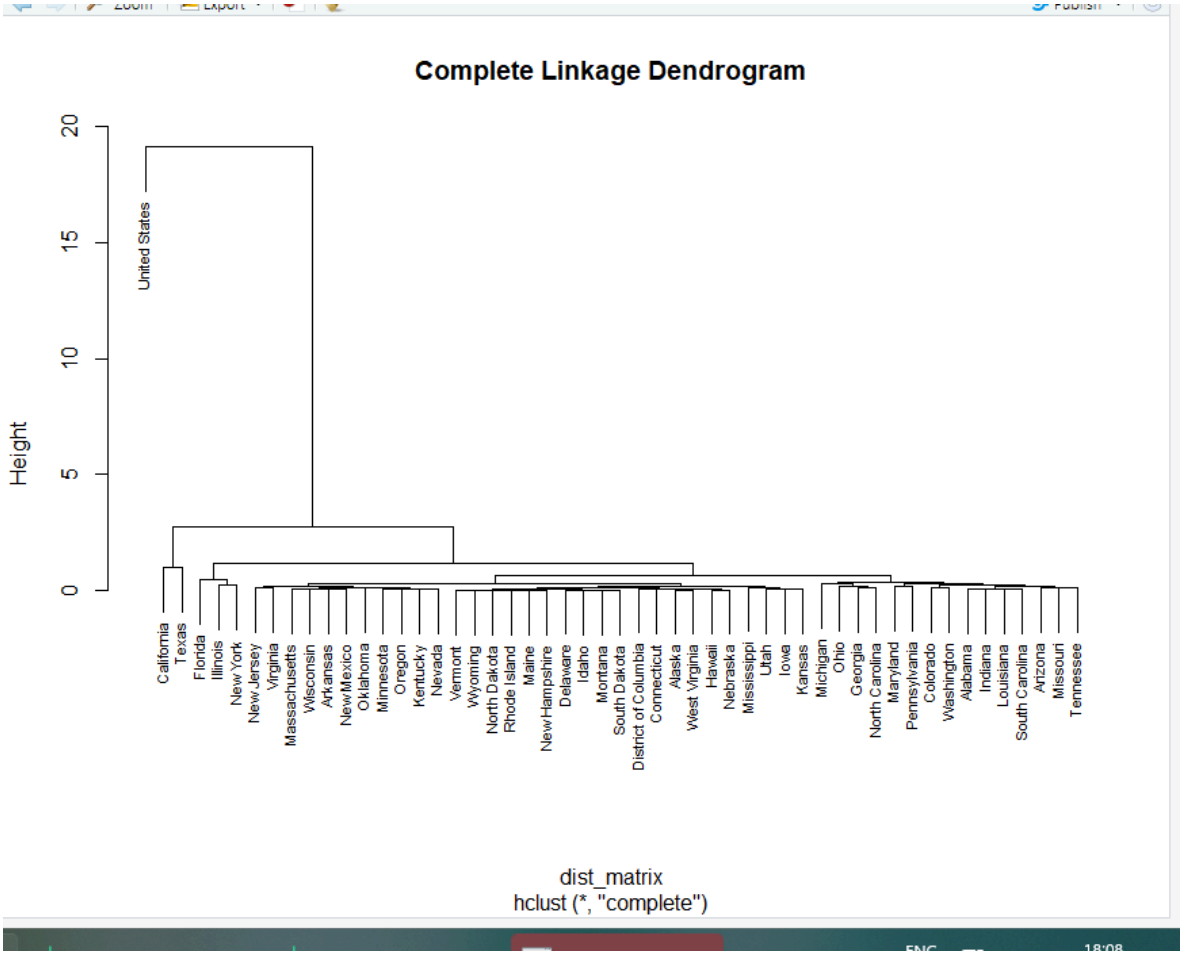
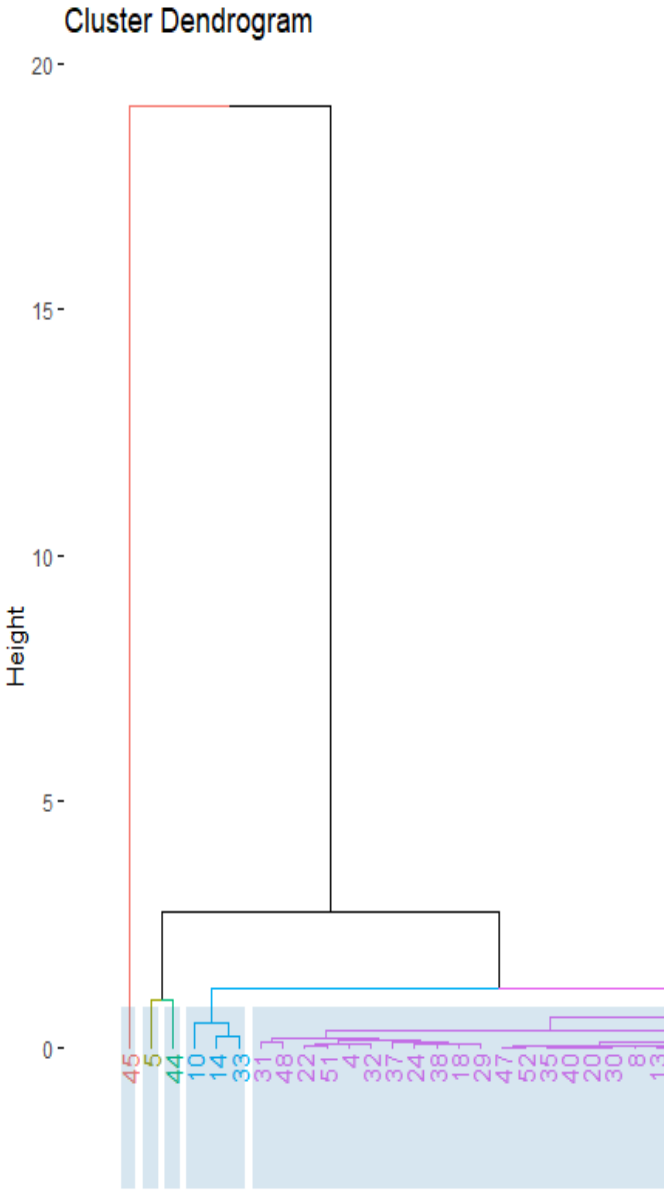
```
> # Load necessary libraries
> library(ggplot2)
> library(cluster)
> library(factoextra)
> library(dendextend)
> library(readr)
> library(tidyverse)
> library(rgl)
> library(MASS)
> library(knitr)
>
> # Load Data
> data <- read.csv("us_crime_data.csv")
>
> # Remove non-numeric columns (assuming first column is State names)
> data_numeric <- data[,-1]
>
> # Standardizing the data
> data_scaled <- scale(data_numeric)
>
> # Perform PCA
> pca_result <- prcomp(data_scaled, center = TRUE, scale. = TRUE)
>
> # Extract first 3 principal components
> pca_data <- as.data.frame(pca_result$x[, 1:3])
> print(head(pca_data))
         PC1           PC2          PC3
1 -0.3722171  0.0242629316 -0.020892408
2 -0.6533620  0.0122761165  0.002631423
3 -0.2533656  0.0000760964 -0.006440092
4 -0.4838226  0.0548924852 -0.028537768
5  1.9620256 -0.6787580644  0.032553634
6 -0.3491026 -0.0278186429 -0.072912955
>
> # 3D PCA Projection
> plot3d(pca_data$PC1, pca_data$PC2, pca_data$PC3, col = "blue", size = 5)
>
> # Outlier Detection using Mahalanobis Distance
> distances <- mahalanobis(pca_data, colMeans(pca_data), cov(pca_data))
> outlier_threshold <- quantile(distances, 0.975)
> outliers <- which(distances > outlier_threshold)
> print(data[outliers, 1])
[1] "California"    "United States"
>
> # Scree Plot
> fviz_eig(pca_result)
>
> # Proportion of Total Sample Variation Captured
> variance_explained <- summary(pca_result)$importance[2, 1:3]
> prop_variation <- sum(variance_explained)
> print(variance_explained)
    PC1     PC2     PC3
```

```
0.99744 0.00152 0.00048
> print(paste("Total Variation Captured:", round(prop_variation * 100, 2),
"%"))
[1] "Total Variation Captured: 99.94 %"
>
> # Correlation Between First PC and Assault
> if("assault" %in% colnames(data_numeric)) {
+  correlation <- cor(pca_result$x[,1], data_numeric[, "assault"])
+  print(paste("Correlation between PC1 and Assault:", correlation))
+ } else {
+  print("Variable 'assault' not found in dataset.")
+ }
[1] "Variable 'assault' not found in dataset."
>
> # Hierarchical Clustering
> # Compute distance matrix
> dist_matrix <- dist(data_scaled, method = "euclidean")
>
> # Perform hierarchical clustering
> hc <- hclust(dist_matrix, method = "complete")
>
> # Plot dendrogram
> plot(hc, labels = data[,1], main = "Complete Linkage Dendrogram", cex =
0.7)
>
> # Partitioning States into Clusters
> clusters <- cutree(hc, k = 5)
>
> # Assign states to clusters
> cluster_assignments <- data.frame(State = data[,1], Cluster = clusters)
> print(head(cluster_assignments, 10))
                 State Cluster
1              Alabama       1
2               Alaska       1
3              Arizona       1
4             Arkansas       1
5           California       2
6             Colorado       1
7          Connecticut       1
8             Delaware       1
9  District of Columbia       1
10             Florida       3
```


Scree Plot
1.00
0.75
0.50
0.25
0.00
Proportion of Variance Explained
2
4
6
Principal Components


Complete Linkage Dendrogram
Height
0
5
10
15
20
United States
California
Texas
Florida
Illinois
New York
New Jersey
Virginia
Massachusetts
Wisconsin
Arkansas
New Mexico
Oklahoma
Minnesota
Oregon
Kentucky
Nevada
Vermont
Wyoming
North Dakota
Rhode Island
Maine
New Hampshire
Delaware
Idaho
Montana
South Dakota
District of Columbia
Connecticut
Alaska
West Virginia
Hawaii
Nebraska
Mississippi
Utah
Iowa
Kansas
Michigan
Ohio
Georgia
North Carolina
Maryland
Pennsylvania
Colorado
Washington
Alabama
Indiana
Louisiana
South Carolina
Arizona
Missouri
Tennessee
dist_matrix
hclust (*, "complete")

>


Cluster Dendrogram
Height
20
15
10
5
0
45
5
44
10
14
33
31
48
22
51
4
32
37
24
38
18
29
47
52
35
40
20
30
8
13