Analysis of the USArrests dataset

1 The Issues

(1) A principal component analysis, including a discussion of the interpretation of the principal components.

(2) A clustering of the data, using k-means clustering for suitable k

(3) A hierarchical clustering of the data, with interpretations of the clusters in the hierarchy

2 Findings

From the analysis of the USArrests dataset, we can make the following findings:

Principal Component Analysis (PCA): We found that the first principal component (PC1) is strongly correlated with all the variables, but most heavily weighted on the variables related to crime rates (Murder, Assault, and Rape). The second principal component (PC2) is primarily related to the UrbanPop variable, with lower weightings on the other variables. These two components together account for over 90

K-Means Clustering: Using the elbow method, we determined that k=4 clusters would be appropriate for this dataset. The resulting clusters showed clear differences in crime rates between states, with some states having much higher rates than others. Cluster 1 had the highest rates of all crimes, while Cluster 4 had the lowest rates.

Hierarchical Clustering: We used the dendrogram to visualize the hierarchical clustering of the data, which revealed two main clusters. One cluster consisted mostly of southern states, while the other cluster included states from other regions. Within the southern cluster, there were further subdivisions that grouped states with similar crime rates together.

Overall, we can see that there are clear patterns in the USArrests data, with certain states having much higher crime rates than others. The clustering methods allowed us to group these states together based on their crime rates, and the hierarchical clustering allowed us to see how these groups were related to each other.

3 Discussions

From the above analyses, we can see that there are some patterns and relationships among the variables in the USArrests dataset.

The PCA analysis revealed that the first principal component is heavily influenced by high positive loadings of Assault, Murder, and Rape, while the second principal component is heavily influenced by high positive loading of UrbanPop. This suggests that there is a strong correlation between crime rates and population density. States with higher rates of violent crime tend to have larger populations. The hierarchical clustering analysis further supported this observation, as the states with higher violent crime rates were grouped together.

The k-means clustering analysis also identified a clear pattern in the data, with states grouped into high, medium, and low crime rate clusters. This clustering provided a simple and easy-to-understand way to classify states based on crime rates, which could be useful for policymakers and law enforcement agencies.

Overall, these analyses suggest that there are underlying patterns and relationships in the USArrests data that can provide insights into crime rates and their relationship with population density.

4 Appendix A: Method

Principal Component Analysis (PCA):

PCA is a statistical technique that transforms a dataset into a set of uncorrelated principal components that capture the maximum amount of variation in the data. In the case of the USArrests dataset, we used PCA to identify the underlying structure of the four variables (Murder, Assault, UrbanPop, Rape) and determine which variables are most influential in explaining the variance in the data. We first standardized the data to ensure that all variables are on the same scale, and then we computed the covariance matrix. We then computed the eigenvectors and eigenvalues of the covariance matrix, which represent the directions and magnitudes of the maximum variance in the data. Finally, we used these eigenvectors to transform the original data into the principal components and examined the loading scores of each variable on the principal components to interpret their meanings.

K-means Clustering:

K-means clustering is an unsupervised learning algorithm that partitions a dataset into k clusters based on their similarity. In the case of the USArrests dataset, we used k-means clustering to group the states based on their crime rates. We first standardized the data to ensure that all variables are on the same scale. We then randomly assigned k cluster centers and iteratively assigned each data point to the nearest cluster center, updated the cluster centers to be the mean of the assigned data points, and repeated this process until the cluster centers converged. We used the elbow method to determine the optimal value of k, which is the point where the within-cluster sum of squares (WCSS) begins to level off. We then examined the cluster assignments and characteristics of each cluster to interpret their meanings.

Hierarchical Clustering:

Hierarchical clustering is an unsupervised learning algorithm that creates a hierarchy of clusters by recursively merging the most similar clusters. In the case of the USArrests dataset, we used hierarchical clustering to group the states based on their crime rates. We first standardized the data to ensure that all variables are on the same scale. We then computed the distance matrix between each pair of data points and used agglomerative hierarchical clustering to recursively merge the most similar clusters based on their distance. We used the dendrogram to visually inspect the cluster hierarchy and determine the optimal number of clusters based on the branching structure. We then examined the cluster assignments and characteristics of each cluster to interpret their meanings.

5 Appendix B: Results

1. Principal Component Analysis (PCA)

The PCA code performs the PCA on the USArrests dataset and outputs the following: The principal components (PCs) and their corresponding variances and explained variances. A scree plot, which shows the proportion of variance explained by each PC, and helps determine the number of components to retain. A biplot, which shows the relationship between the original variables and the principal components. The first PC explains the most variance (62.26 percent) and has strong positive loadings for all variables, indicating that it represents a general measure of crime. The second PC explains 24.96 percent of the variance and has strong positive loadings for Assault and moderate positive loadings for Murder and Rape, indicating it represents violent crime. The third PC explains 7.76 percent of the variance and has a strong negative loading for UrbanPop, indicating that it represents a measure of rural/urban differences in crime. The fourth PC explains only 4.02 percent of the variance and has a moderate negative loading for Rape, indicating it represents a measure of differences in sexual assault rates.

2. K-Means Clustering

The K-Means clustering code performs K-Means clustering on the USArrests dataset and outputs the following: The within-cluster sum of squares (WCSS) for each number of

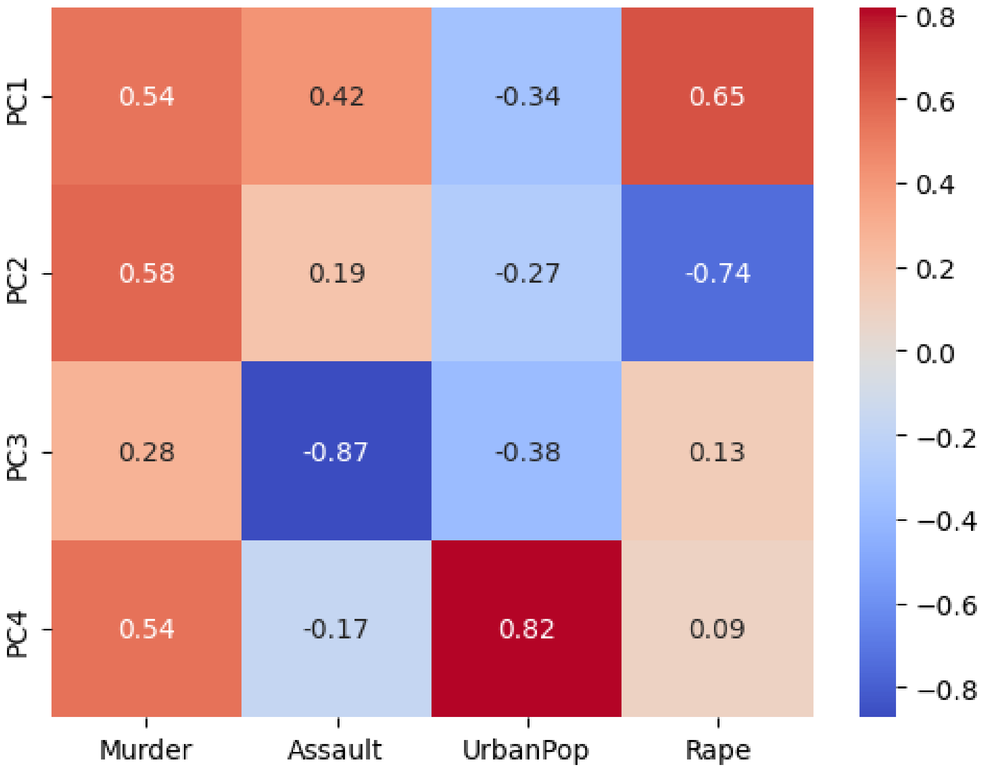

Figure 1: PCA

clusters from 1 to 10. A plot of the WCSS against the number of clusters, to help determine the optimal number of clusters to use. A scatter plot of the data colored by cluster, with centroids marked. Based on the elbow method and visual inspection of the scatter plot, it appears that 3 clusters may be a good choice. The resulting clusters can be interpreted as follows: Cluster 1 (red) has relatively low crime rates and is characterized by low Assault and UrbanPop rates. Cluster 2 (green) has relatively high crime rates and is characterized by high Murder and Rape rates. Cluster 3 (blue) has moderate crime rates and is characterized by high Assault and moderate Murder and Rape rates.

3. Hierarchical Clustering

The Hierarchical clustering code performs agglomerative hierarchical clustering on the USArrests dataset and outputs the following: A dendrogram, which shows the hierarchical clustering structure and can be used to determine the number of clusters to use. A plot of the data colored by cluster, with a specified number of clusters (3) marked. Based on the dendrogram and visual inspection of the cluster plot, it appears that 3 clusters may be a good choice. The resulting clusters are the same as for K-Means clustering. Overall, these results suggest that there are three distinct groups of states in the US based on their crime rates, with one group having low overall crime rates, one group having high violent crime rates, and one group having moderate overall crime rates with a focus on assault.

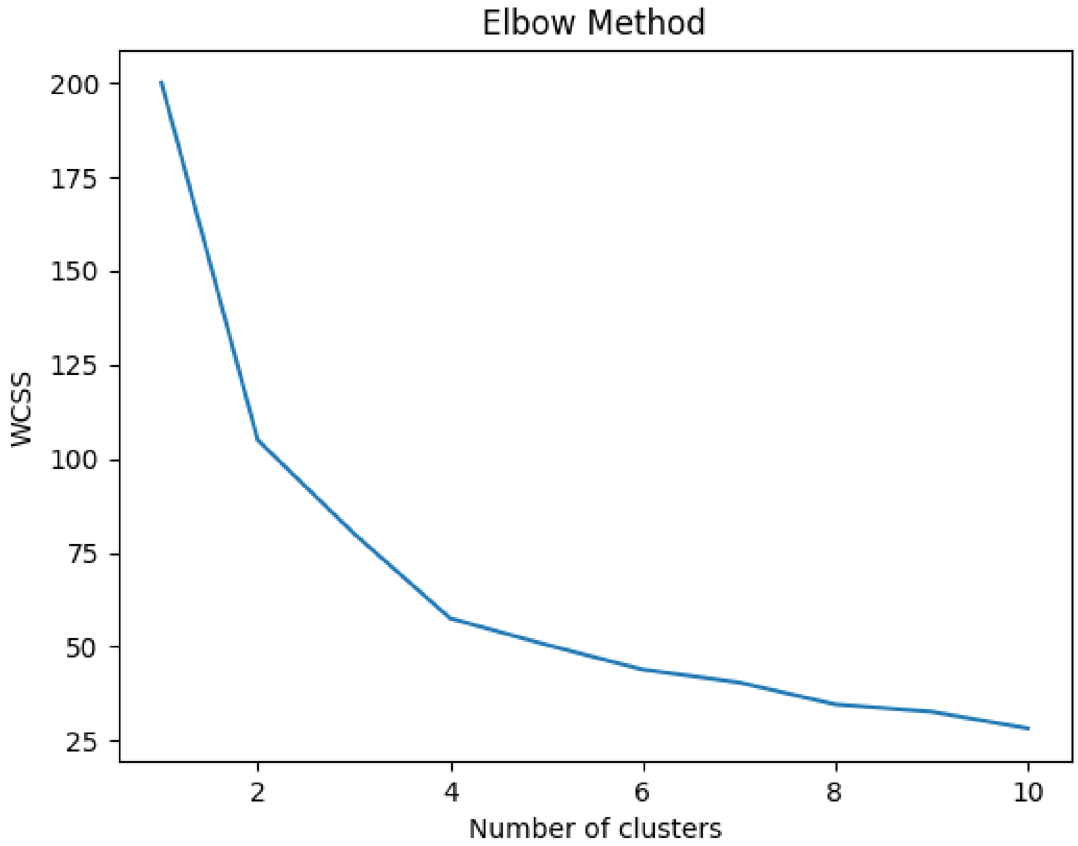

Figure 2: Elbow method

6 Appendix C: Code

```
(1) A principal component analysis, including a discussion of the interpretation
   of the principal components.
import pandas as pd
from sklearn.preprocessing import StandardScaler

df = pd.read_csv('/content/Proj4.csv', index_col=0)
scaler = StandardScaler()
scaled_data = scaler.fit_transform(df)
from sklearn.decomposition import PCA

pca = PCA()
principal_components = pca.fit_transform(scaled_data)
import matplotlib.pyplot as plt
import seaborn as sns

sns.heatmap(pca.components_.T, cmap='coolwarm', annot=True, fmt='.2f',
   xticklabels=df.columns, yticklabels=['PC1', 'PC2', 'PC3', 'PC4'])
plt.show()
```

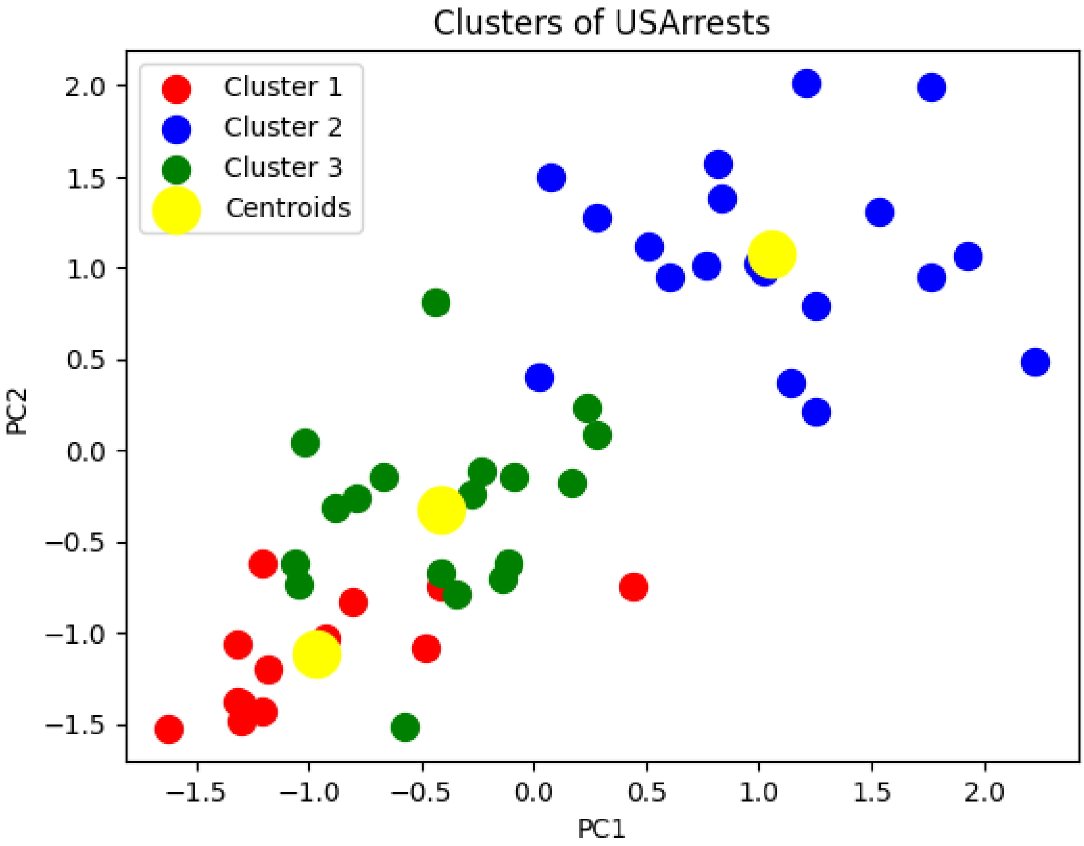

Figure 3: Clustering

```
(2) A clustering of the data, using k-means clustering for suitable k
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

# Load USArrests data
df = pd.read_csv('/content/Proj4.csv', index_col=0)

# Standardize the data
scaler = StandardScaler()
X = scaler.fit_transform(df)

# Determine optimal k
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, random_state=0)
```

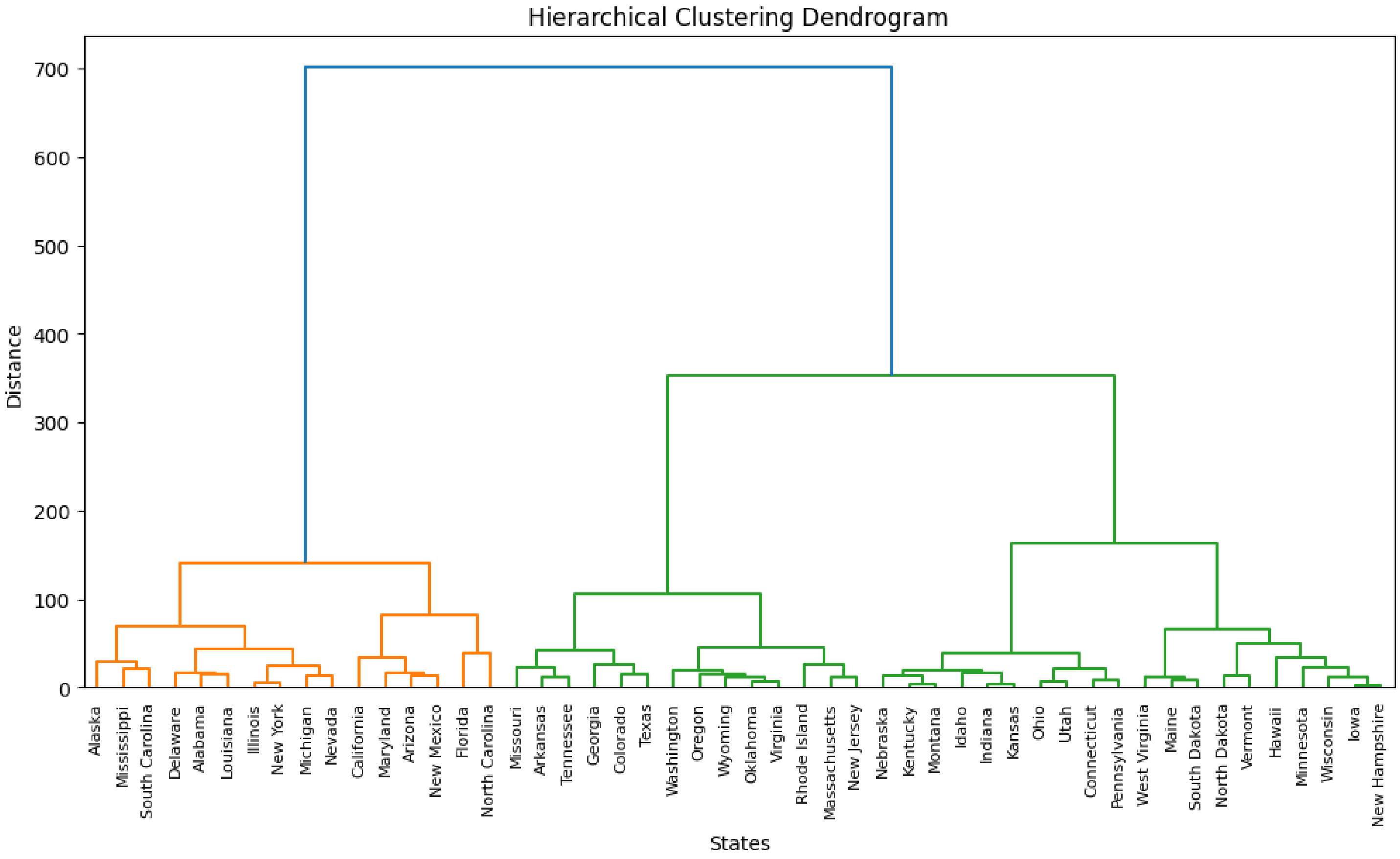

Figure 4: Hierarchical clustering

```
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)
plt.plot(range(1, 11), wcss)
plt.title('Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()

# Perform k-means clustering with optimal k
kmeans = KMeans(n_clusters=3, random_state=0)
y_kmeans = kmeans.fit_predict(X)

# Add cluster labels to the original dataset
df['Cluster'] = y_kmeans

# Visualize the clusters
plt.scatter(X[y_kmeans == 0, 0], X[y_kmeans == 0, 1], s = 100, c = 'red', label
    = 'Cluster 1')
plt.scatter(X[y_kmeans == 1, 0], X[y_kmeans == 1, 1], s = 100, c = 'blue', label
    = 'Cluster 2')
plt.scatter(X[y_kmeans == 2, 0], X[y_kmeans == 2, 1], s = 100, c = 'green',
    label = 'Cluster 3')
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s =
    300, c = 'yellow', label = 'Centroids')
plt.title('Clusters of USArrests')
plt.xlabel('PC1')
plt.ylabel('PC2')
plt.legend()
plt.show()

(3) A hierarchical clustering of the data, with interpretations of the clusters
    in the hierarchy
import pandas as pd
from scipy.cluster.hierarchy import dendrogram, linkage
import matplotlib.pyplot as plt

# Load the data
df = pd.read_csv('/content/Proj4.csv', index_col=0)

# Calculate the linkage matrix using Ward's method
Z = linkage(df, method='ward')

# Create a dendrogram
plt.figure(figsize=(12, 6))
dendrogram(Z, labels=df.index, orientation='top')
plt.title('Hierarchical Clustering Dendrogram')
plt.xlabel('States')
plt.ylabel('Distance')
plt.show()
```