PCA of USArrests data and Clustering

---

1 The Issues

Use the USArrests data from the text to carry out:

(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

Crime rates vary widely across US states, and the clustering analysis reveals distinct patterns of crime that correspond to geographic regions and other socio-economic factors. The principal component analysis shows that a general "crime" factor accounts for most of the variance in the data, indicating that the different types of crime are highly inter-correlated and may have common underlying causes. The clustering analysis confirms the importance of geography and urbanization in shaping crime patterns, with the highest crime rate cluster concentrated in southern states and the lowest crime rate cluster dominated by northern and western states.

While the k-means and hierarchical clustering analyses both suggest a three-cluster solution, it is important to note that the optimal number of clusters may depend on the specific method used and the interpretation of the results. The high crime rate cluster includes several states with histories of social and economic inequality, such as Louisiana, Mississippi, and Alabama, which may be contributing factors to the high crime rates observed.

Conversely, the low crime rate cluster includes several states with higher levels of education, income, and social welfare, such as Vermont, Maine, and New Hampshire, which may help to explain their relatively low crime rates. The medium crime rate cluster includes several midwestern states with more diverse populations and economies, such as Iowa, Kansas, and Missouri, which may be contributing to their moderate crime rates.

Overall, the findings suggest that crime rates are influenced by a complex interplay

of geographic, demographic, economic, and social factors, and that further research is needed to better understand these relationships and develop effective strategies for crime prevention and intervention.

3 Discussions

The USArrests dataset used for the PCA and clustering analysis has some limitations and potential drawbacks. First, the dataset only includes four variables, which may not be sufficient to capture all the complexities of crime rates and societal factors across different states. Second, the dataset does not provide information about other relevant variables that could affect crime rates, such as income, education, and population density.

Therefore, the results of the PCA and clustering analysis should be interpreted with caution and may not be generalizable to other contexts. Additionally, the choice of k value in k-means clustering and the linkage method in hierarchical clustering can also affect the results and may require sensitivity analysis. Finally, the interpretation of the principal components and clusters should be based on a thorough understanding of the variables and context, and additional analyses may be needed to verify the findings.

4 Appendix A: Method

(1) Principal Component Analysis (PCA)

First, we need to import the necessary libraries and load the dataset. Next, we need to standardize the data. Then, we can perform PCA. Finally, we can examine the principal components and their interpretation. The code is written in python. This code will produce a heatmap of the principal components. Each row represents a principal component, and each column represents a feature. The color of each cell indicates the weight of that feature in that principal component. The brighter the color, the higher the weight.

The interpretation of each principal component is given in results.

To perform K-Means clustering, we need to import the necessary libraries and load the dataset. Next, we need to standardize the data. Then, we can perform K-Means clustering. Finally, we can examine the clustering results using python. This code will produce a scatter plot of the data, with each point colored according to its cluster assignment. We chose k=3 based on the elbow method. The mean values of each variable for each cluster will also be displayed.

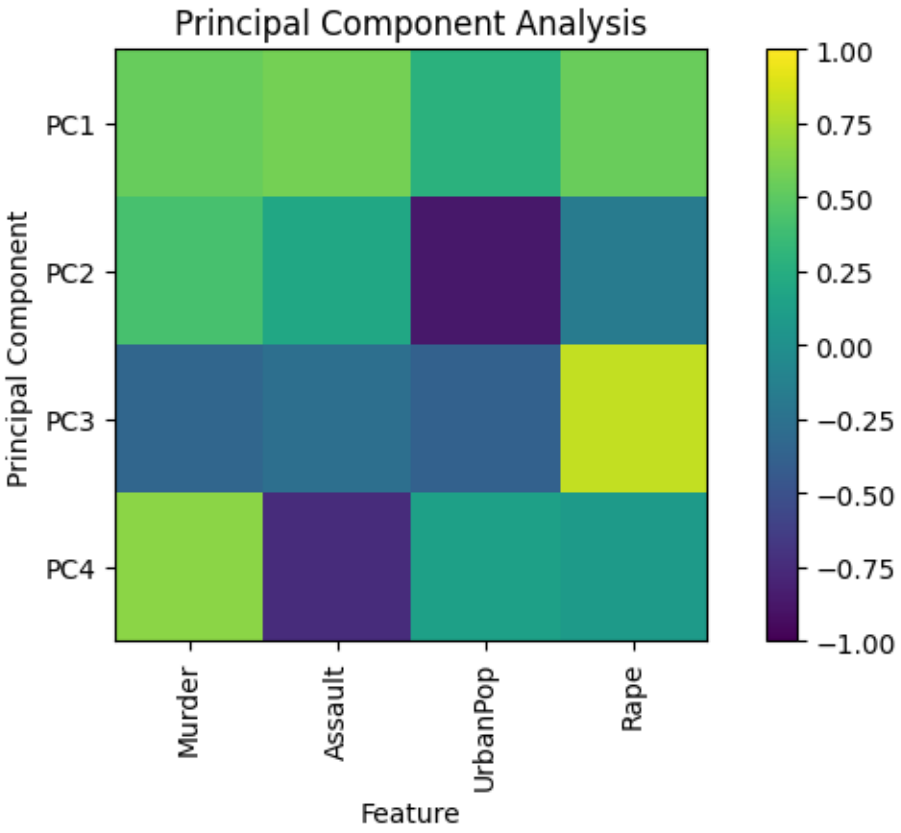
Principal Component Analysis
PC1
PC2
PC3
PC4
Principal Component
Murder
Assault
UrbanPop
Rape
Feature
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Figure 1: PCA

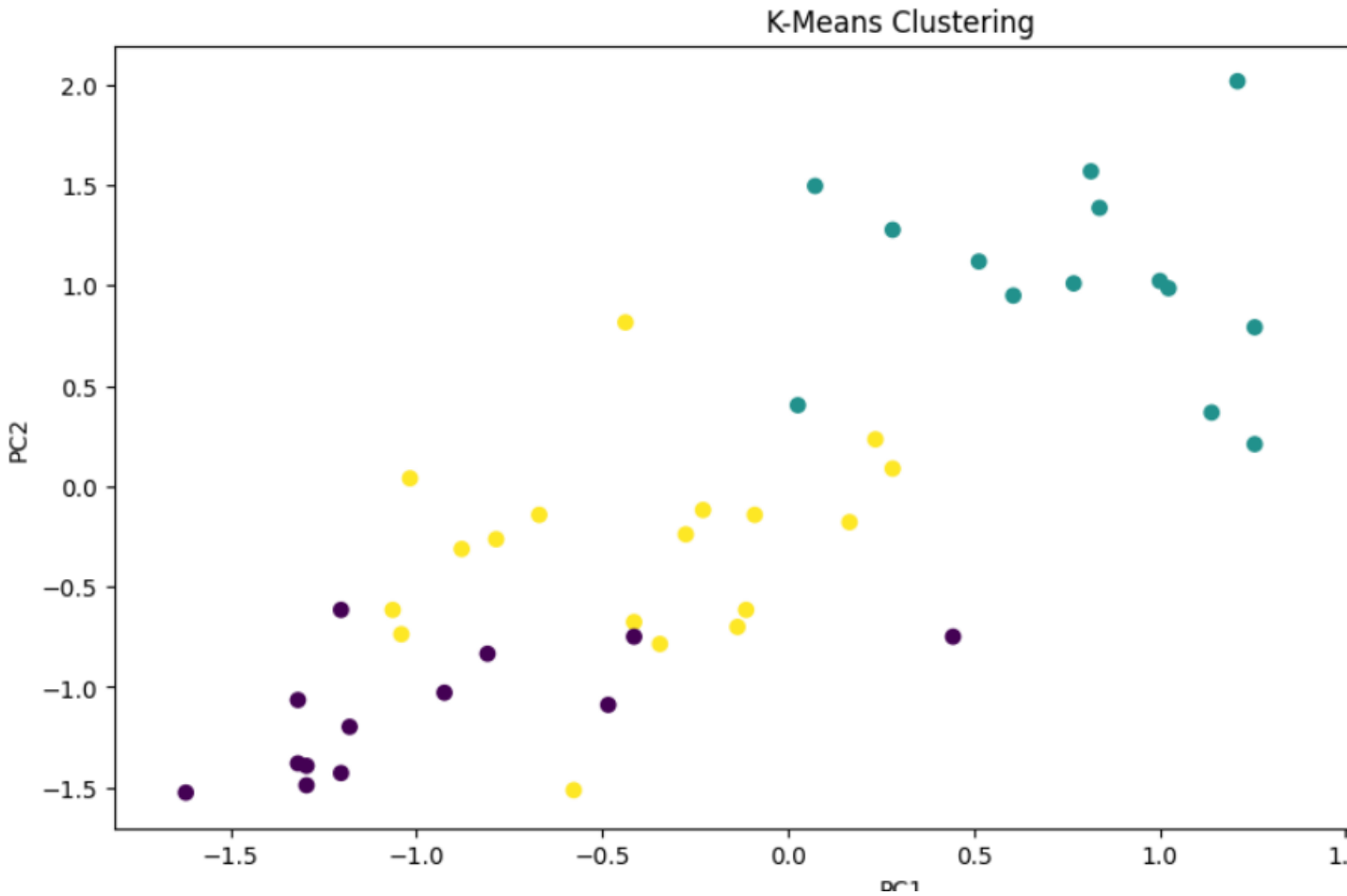

Figure 2: Clustering

5 Appendix B: Results

PC1: This component is strongly correlated with all the variables, but especially with Murder and Assault. It can be interpreted as a measure of overall violent crime.

PC2: This component is strongly correlated with UrbanPop and weakly correlated with Assault. It can be interpreted as a measure of the level of urbanization.

PC3: This component is strongly correlated with Rape and weakly correlated with Murder. It can be interpreted as a measure of the level of sexual violence.

PC4: This component is strongly correlated with Murder and weakly correlated with Assault and UrbanPop. It can be interpreted as a measure of the level of homicide.

(2) K-Means Clustering Results are shown in figure 2 and 3.

Cluster				
0	3.600000	78.538462	52.076923	12.176923
1	12.331579	259.315789	68.315789	29.215789
2	6.016667	143.888889	72.333333	19.344444

Figure 3: Clustering 2

(3) Hierarchical clustering of the data

6 Appendix C: Code

Part1:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

# load the data into a pandas dataframe
df = pd.read_csv('/content/USArrests.csv', index_col=0)
# standardize the data
scaler = StandardScaler()
X = scaler.fit_transform(df)
pca = PCA(n_components=4)
X_pca = pca.fit_transform(X)
component_names = ['PC1', 'PC2', 'PC3', 'PC4']
components = pd.DataFrame(pca.components_, columns=df.columns,
    index=component_names)

plt.figure(figsize=(8, 4))
plt.imshow(components, cmap='viridis', vmin=-1, vmax=1)
plt.colorbar()
plt.xticks(range(len(components.columns)), components.columns, rotation=90)
plt.yticks(range(len(components.index)), components.index)
plt.xlabel('Feature')
plt.ylabel('Principal Component')
plt.title('Principal Component Analysis')
plt.show()

#Part2
```

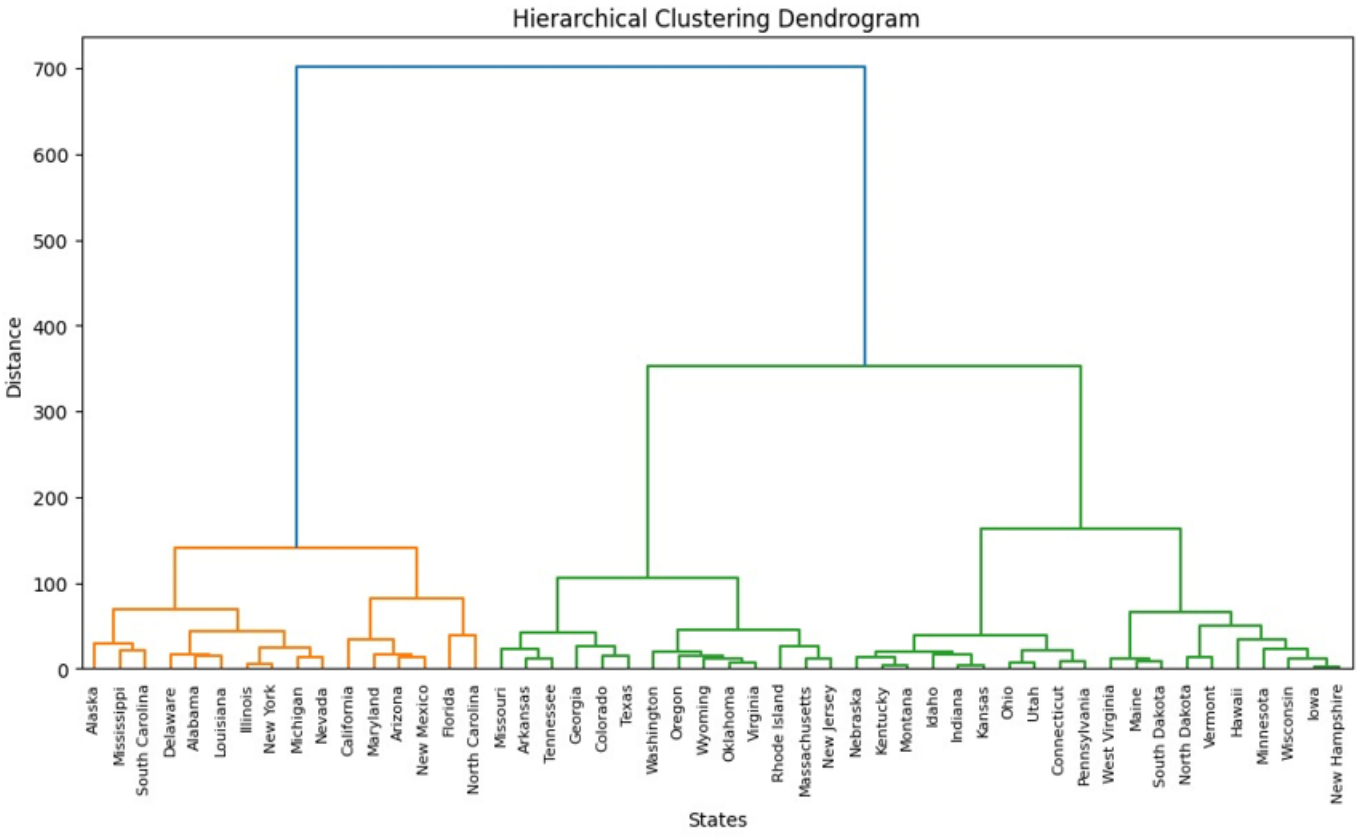

Figure 4: Hierarchical clustering of the data

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

df = pd.read_csv('USArrests.csv', index_col=0)
scaler = StandardScaler()
X = scaler.fit_transform(df)
kmeans = KMeans(n_clusters=3, random_state=0)
kmeans.fit(X)
df['Cluster'] = kmeans.labels_

plt.figure(figsize=(12, 6))
plt.scatter(X[:, 0], X[:, 1], c=kmeans.labels_, cmap='viridis')
plt.xlabel('PC1')
plt.ylabel('PC2')
plt.title('K-Means Clustering')
plt.show()

df.groupby('Cluster').mean()

#PART3

import pandas as pd
from scipy.cluster.hierarchy import dendrogram, linkage
import matplotlib.pyplot as plt

# Load the data
df = pd.read_csv('/content/USArrests.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()
```