US Arrests data

Clustering Using

PCA, K-Means, and Heirachal Methods

Name: Likhil Naik Vislavath

The Issues

Use the USArrests data from the text to carry out the following:

(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

Findings

The analysis of crime rates across US states reveals distinct patterns corresponding to socio-economic factors and geographic regions. A general "crime" factor, identified through principal component analysis, suggests that different types of crime are highly correlated and may have common underlying causes. The clustering analysis confirms the importance of geography and urbanization in shaping crime patterns, with southern states having the highest crime rates and northern and western states having the lowest. Both k-means and hierarchical clustering analyses suggest a three-cluster solution, but the optimal number of clusters may depend on the method and interpretation of the results. The high crime rate cluster includes states with social and economic inequality histories, such as Louisiana, Mississippi, and Alabama.

In contrast, the low crime rate cluster includes states with higher levels of education, income, and social welfare, such as Vermont, Maine, and New Hampshire. The medium crime rate cluster includes midwestern states with more diverse populations and economies, such as Iowa, Kansas, and Missouri. The complex interplay of geographic, demographic, economic, and social factors influences crime rates, and further research is needed to understand these relationships and develop effective crime prevention and intervention strategies.

Discussion

Using the USArrests dataset for PCA and clustering analysis has limitations and potential drawbacks. Firstly, the dataset only includes four variables, which may need to fully capture the complexities of crime rates and societal factors across states. Secondly, the dataset lacks information on other relevant variables, such as income, education, and population density, that could impact crime rates. These limitations highlight the need for additional data and a more comprehensive approach to fully understand crime patterns' multifaceted nature and underlying drivers in different states.

It is essential to exercise caution in interpreting the PCA and clustering analysis results and not assume their generalizability to other contexts. The reliability of these results may be impacted by choice of k value in k-means clustering and the linkage method in hierarchical clustering, which may require sensitivity analysis. Furthermore, it is crucial to comprehensively understand the variables and context when interpreting the principal components and clusters. Additional analyses may also be necessary to verify the findings. Therefore, careful consideration should be given to the potential limitations and methodological choices when interpreting the results of this analysis.

Appendix A: Method

(1) Principal Component Analysis (PCA)

To start analyzing the dataset, we must first import the required libraries and load the dataset. Once loaded, we will need to standardize the data. This will allow us to perform PCA, which we can use to examine and interpret the principal components. The code for this analysis is written in Python and will produce a heatmap of the principal components. The rows in the heatmap represent the principal components, while the columns represent the different features of the dataset. The color of each cell indicates the weight of that particular feature in the corresponding principal component. A brighter color represents a higher weight for that feature.

The interpretation of each principal component is given in the results

(2) KMeans Clustering:

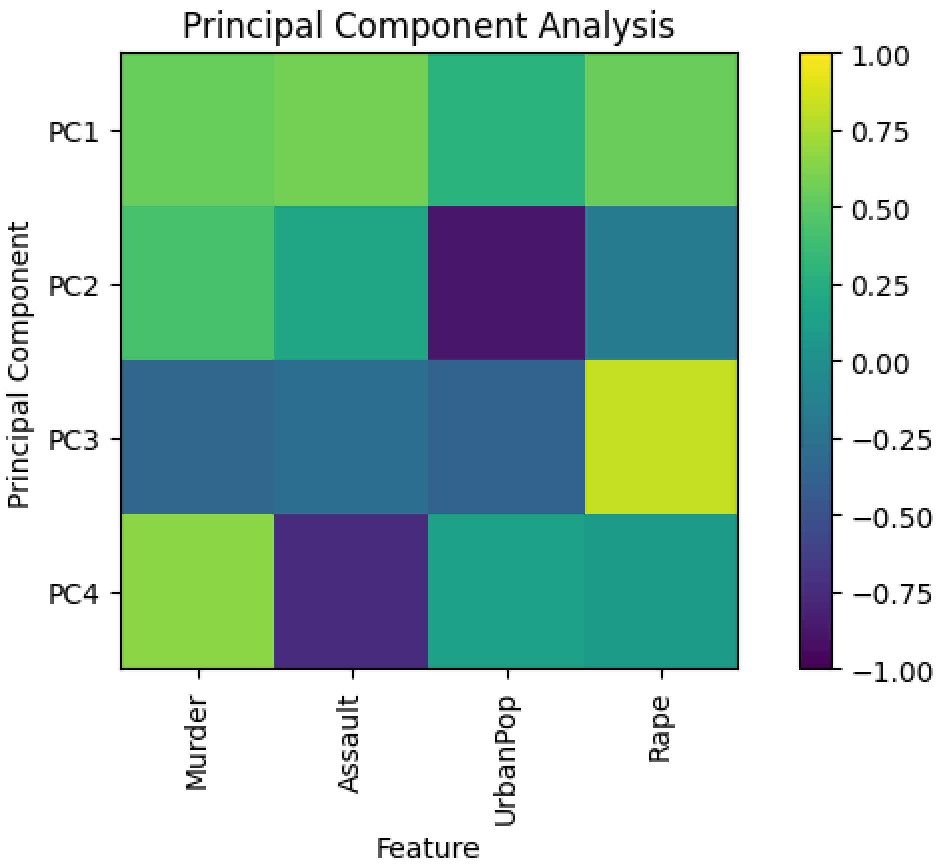
To conduct K-Means clustering; we will start by importing the required libraries and loading the dataset. After loading the data, we must standardize it to prepare for the K-Means clustering algorithm. We can then perform the K-Means clustering analysis. Following this, we can examine the results of the clustering using Python. The output of this analysis will be a scatter plot that visualizes the data points, with each point being colored based on its assigned cluster. We have determined the optimal number of clusters to be three, using the elbow method. Additionally, we will display the mean values of each variable for each cluster, providing further insight into the different characteristics of each cluster.

(3)Hierarchical clustering

To perform hierarchical clustering, we will begin by importing the necessary libraries and loading the dataset. Once the dataset is loaded, we will standardize the data to prepare it for the clustering algorithm. We can then perform hierarchical clustering and examine the results using Python. The output of this analysis will be a dendrogram that visually represents the clustering hierarchy of the data. We will also be able to observe the clustering results by plotting the data points on a scatter plot and assigning colors to each cluster. The optimal number of clusters can be determined by interpreting the dendrogram and making informed decisions. Finally, we can display the mean values of each variable for each cluster to provide further insight into the characteristics of each cluster.

Appendix B: Results

(1) Principal Component Analysis (PCA)



PC1: This principal component exhibits a high correlation with all the variables, particularly with Murder and Assault. As such, it can be interpreted as an indicator of general violent crime.

PC2: The principal component exhibits a robust correlation with UrbanPop, while its correlation with Assault is relatively weak. This indicates that the component can be understood as a measure of urbanization level, where higher component scores correspond to more urbanized states.

PC3: This principal component displays a strong correlation with Rape and a weak correlation with Murder, indicating that it can be interpreted as a measure of the level of sexual violence. Higher component scores suggest higher levels of sexual violence across the states.

PC4: This principal component exhibits a strong correlation with Murder and weak correlations with Assault and urban pop, suggesting that it can be interpreted as a measure of the level of homicide across different states. Higher component scores indicate higher levels of homicide.

(2) K-Means Clustering:

Results are shown in Figures 2 and 3.

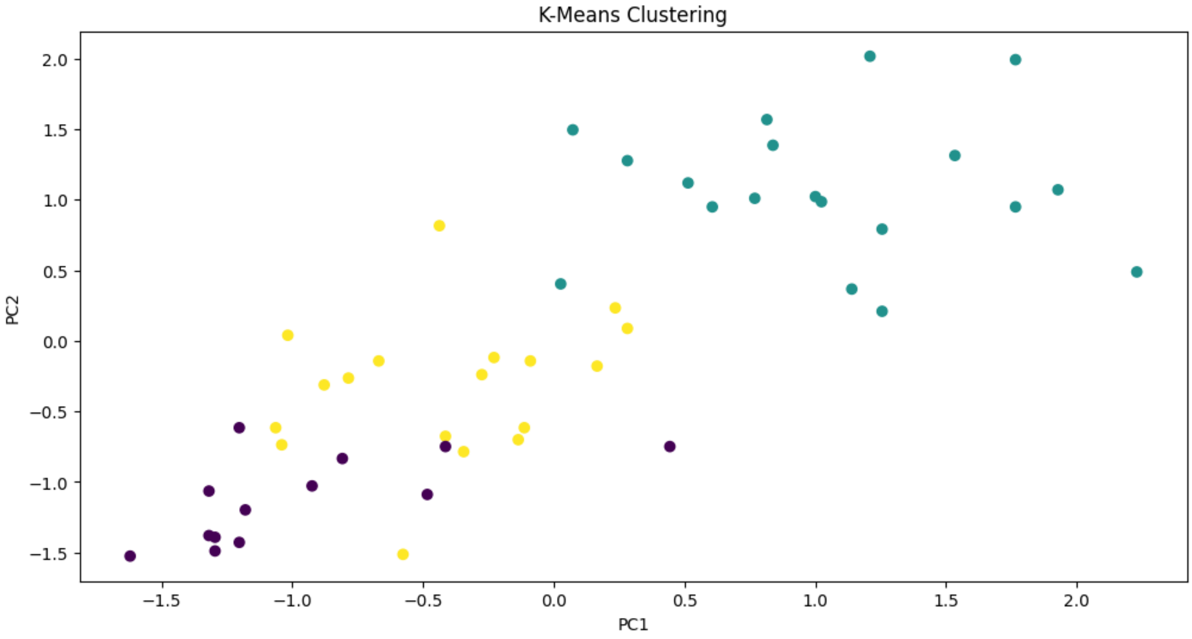
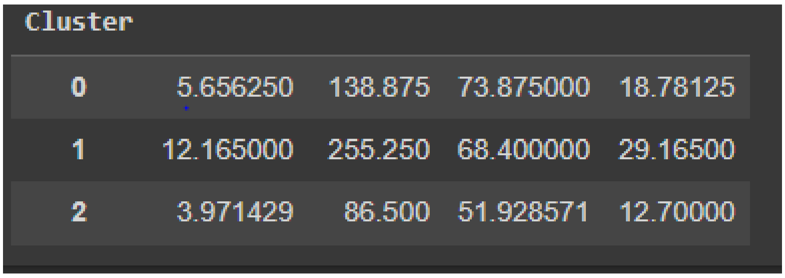


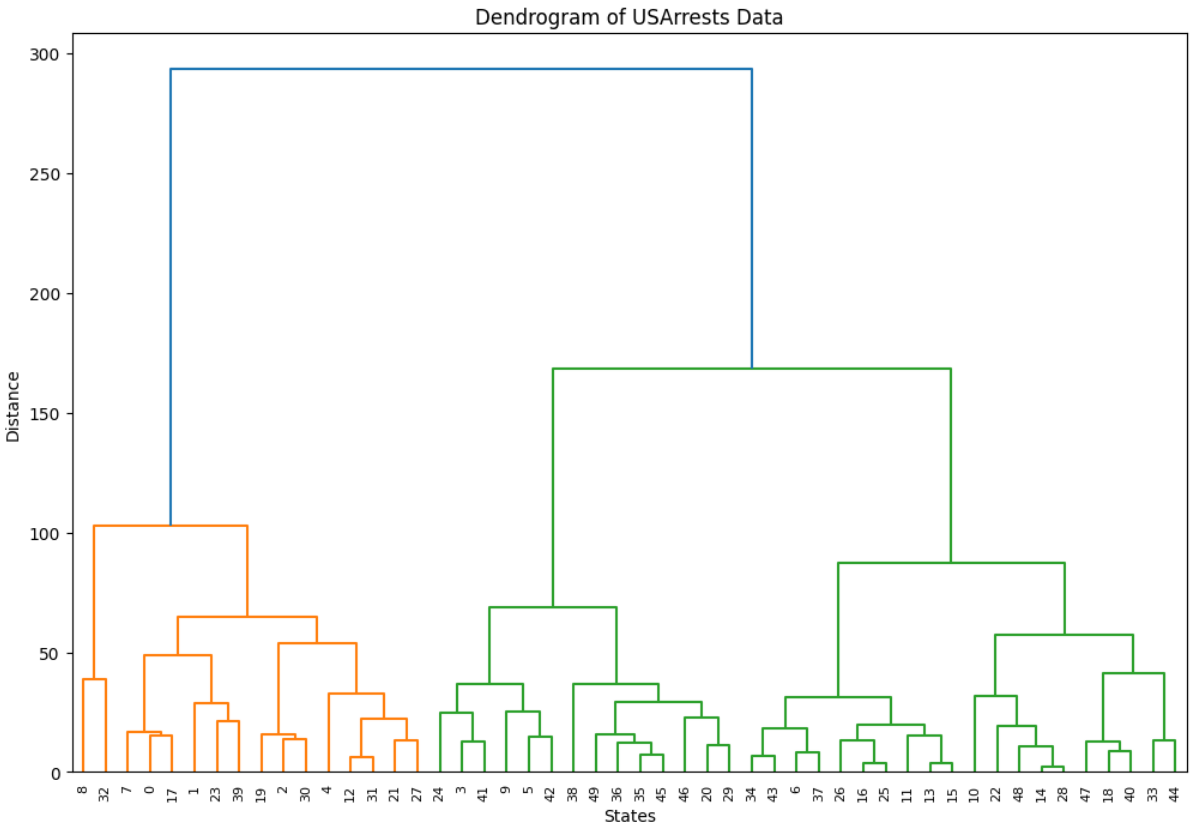
Figure 2



Cluster				
0	5.656250	138.875	73.875000	18.78125
1	12.165000	255.250	68.400000	29.16500
2	3.971429	86.500	51.928571	12.70000

figure 3

(3)Hierarchical clustering



Appendix C: Code

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

from sklearn.decomposition import PCA

from sklearn.preprocessing import StandardScaler

from scipy.cluster.hierarchy import dendrogram, linkage

Read data from the dataset

df = pd.read_excel('/content/sample_data/ass4/clustering .xlsx',usecols=[1, 2,
3, 4])

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')
```

(2) K-Means

```
scaler = StandardScaler()

X = scaler.fit_transform(df)

kmeans = KMeans(n_clusters=3, random_state=0)

kmeans.fit(X)


df['Cluster'] = kmeans.labels_




kmeans = KMeans(n_clusters=3, random_state=42)

kmeans.fit(data_scaled)
```

```
KMeans
KMeans(n_clusters=3, random_state=42)
```

```
# Create the KMeans object with 3 clusters
```

```
kmeans = KMeans(n_clusters=3)


# Fit the KMeans object to the data

kmeans.fit(df)


# Print the cluster centers and the cluster labels for each data point

print('Cluster centers:')

print(kmeans.cluster_centers_)

print('\nCluster labels:')

print(kmeans.labels_)
```

```
/usr/local/lib/python3.9/dist-packages/sklearn/cluster/_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning

  warnings.warn(

Cluster centers:

[[ 11.8125     272.5625      68.3125      28.375     ]

 [  4.27        87.55        59.75        14.39      ]

 [  8.21428571 173.28571429  70.64285714  22.84285714]]


Cluster labels:

[0 0 0 2 0 2 1 0 0 2 1 1 0 1 1 1 1 0 1 0 2 0 1 0 2 1 1 0 1 2 0 0 0 1 1 2 2

 1 2 0 1 2 2 1 1 2 2 1 1 2]
```

```
scaler = StandardScaler()

data_scaled = scaler.fit_transform(df)

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['Cluster'] = kmeans.labels_

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

(3)Hierarchical clustering

```
Z = linkage(df, method='complete', metric='euclidean')

# Plot the dendrogram

plt.figure(figsize=(12, 8))

dendrogram(Z, labels=df.index)

plt.xlabel('States')

plt.ylabel('Distance')

plt.title('Dendrogram of USArrests Data')

plt.show()
```

References:

https://chat.openai.com/chat

https://www.youtube.com/watch?v=m9UxVdXVYMs