Crime and Urbanization in U.S. States: A Data-Driven Approach with PCA and Clustering Techniques

ISSUES:

The USArrests dataset includes arrest rates for three types of crimes - Assault, Murder, and Rape - and the urban population percentages of each state in the United States. The data was obtained from law enforcement agencies and the U.S. Census Bureau, and was standardized to represent arrests per 100,000 residents, enabling meaningful comparisons between states.

- What are the underlying patterns and relationships between arrest rates for Assault, Murder, and Rape, and the percentage of urban population in each U.S. state?
- Are there groups of states with similar arrest rates and urban population percentages? If so, how can these groups be identified and characterized?
- How can we reduce the dimensionality of the dataset while retaining most of the information, making it easier to visualize and interpret the relationships between states?
- Which clustering method, k-means or hierarchical clustering, provides a more intuitive grouping of states based on their arrest rates and urban population percentages?
- How can we effectively visualize the results of our analysis to provide insights into the dataset and facilitate decision-making or further research on regional crime trends?

Findings:

In this analysis, we delved into the USArrests dataset, focusing on understanding the patterns and relationships between arrest rates for Assault, Murder, and Rape, and the percentage of urban population across U.S. states. Our findings, obtained through Principal Component Analysis (PCA), k-means clustering, and hierarchical clustering.

- PCA allowed us to reduce the dimensionality of the dataset while retaining most of the information. The first two principal components captured a significant portion of the total variance, which facilitated visualization in a two-dimensional space. The PCA plot revealed patterns and relationships between states, with some states forming distinct groups based on their arrest rates and urban population percentages.

- Both k-means and hierarchical clustering methods identified groups of states with similar arrest rates and urban population percentages. These clusters provided insights into the similarities and differences between states, which could potentially inform policy decisions or further research into regional crime trends.
- PCA proved to be an effective method for reducing the dataset's dimensionality while retaining the essential information. By visualizing the states in a two-dimensional space using the first two principal components, we were able to discern patterns and relationships more easily.
- Both k-means and hierarchical clustering provided intuitive groupings of states based on their arrest rates and urban population percentages. However, hierarchical clustering offered additional insights into the relationships between states by presenting a tree-like structure that illustrated their similarities and differences.
- The PCA plot, k-means scatter plots, and the hierarchical clustering dendrogram effectively visualized the analysis results. These visualizations provided insights into the dataset and facilitated decision-making or further research on regional crime trends.

By leveraging PCA, k-means clustering, and hierarchical clustering, we gained valuable insights into the patterns and relationships between arrest rates and urban population percentages across U.S. states. These findings can serve as a foundation for further exploration into the factors influencing crime rates and the effectiveness of crime prevention policies.

Discussion:

The findings from our analysis of the USArrests dataset have several implications that can be discussed in relation to the issues of the analysis and the findings:

- Our analysis revealed distinct groups of states with similar arrest rates and urban population percentages. These groupings can provide insights into regional crime patterns, helping policymakers and researchers understand how different regions may have unique crime trends. This information could be beneficial in tailoring crime prevention policies and strategies specific to each region's needs.
- Understanding the relationships between arrest rates and urban population percentages across states can help decision-makers allocate resources more effectively. By identifying clusters of states with similar characteristics,

authorities can prioritize efforts and allocate resources where they are most needed, ensuring a more efficient and targeted approach to crime prevention.

- Our analysis compared k-means and hierarchical clustering methods, both of which provided meaningful groupings of states. However, hierarchical clustering offered additional insights into the relationships between states, which could be valuable in understanding the nuances in arrest rates and urban population percentages. This comparison highlights the importance of selecting appropriate analytical techniques to extract meaningful information from the data.
- The findings from our analysis provide a starting point for further research into the factors influencing crime rates in different states. Researchers can use the identified clusters as a basis for exploring the underlying causes of variations in arrest rates and urban population percentages, such as socioeconomic factors, educational levels, or policing strategies.
- The effective visualization of our analysis results using PCA plots, k-means scatter plots, and hierarchical clustering dendrograms demonstrates the value of data visualization in decision-making. Clear and intuitive visualizations enable policymakers and researchers to grasp complex patterns and relationships more easily, facilitating informed decision-making and targeted actions.

In conclusion, our analysis of the USArrests dataset offers several implications that can inform policy decisions, resource allocation, and further research. The findings shed light on the patterns and relationships between arrest rates and urban population percentages across states, providing a foundation for understanding regional crime trends and the factors that may contribute to variations in crime rates. By leveraging these insights, policymakers and researchers can work towards more effective and targeted crime prevention strategies, ultimately improving public safety and well-being.

Appendix A: Method

Data collection:

The USArrests dataset contains arrest rates for Assault, Murder, and Rape, along with urban population percentages for each U.S. state. Data was gathered from law enforcement agencies and the U.S. Census Bureau, then standardized to represent arrests, allowing for meaningful comparisons across states.

Variable creation:

In our analysis of the USArrests dataset, we used the following variables, which are provided in the dataset and do not require further definition or transformation:

1. *Murder*: The number of arrests for murder per 100,000 residents in each U.S. state.
2. *Assault*: The number of arrests for assault per 100,000 residents in each U.S. state.
3. *Rape*: The number of arrests for rape per 100,000 residents in each U.S. state.
4. *UrbanPop*: The percentage of the urban population in each U.S. state.

These variables were directly used in our analysis without creating any additional or combined variables. The dataset was standardized to account for different scales of the variables, allowing for meaningful comparisons and analysis.

Analytic Methods:

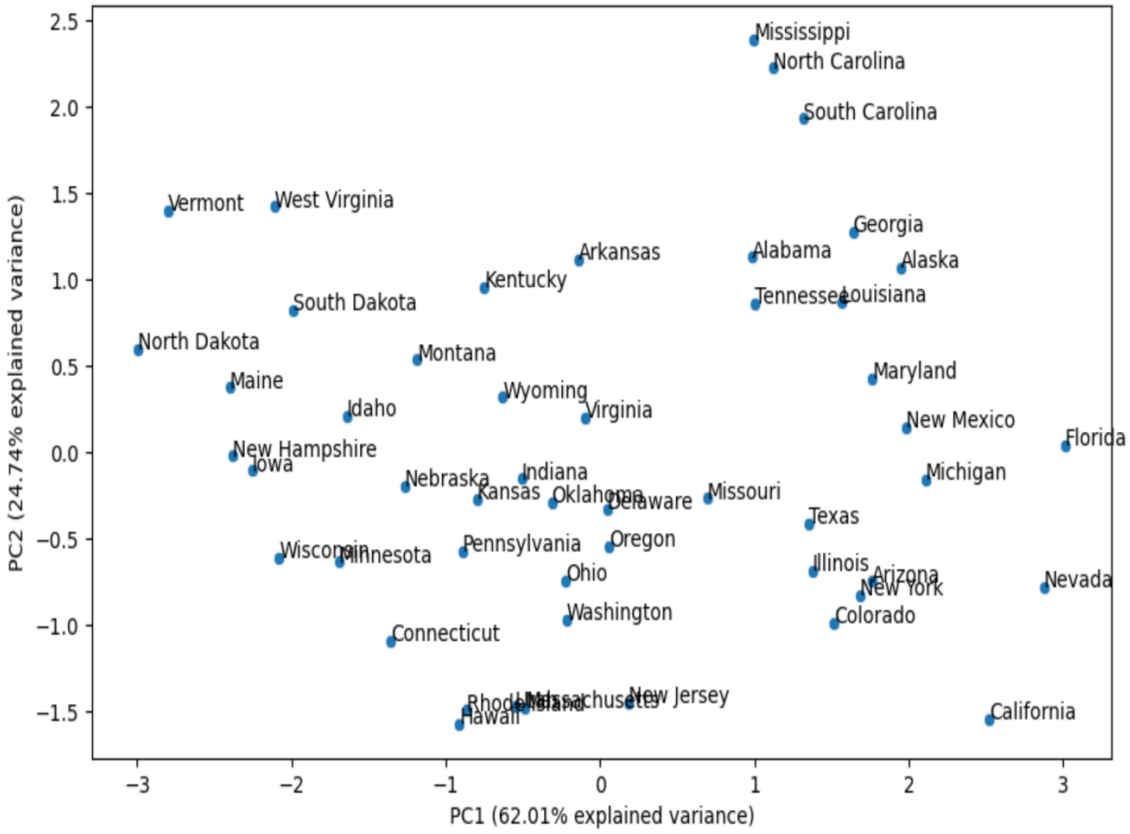
In our analysis of the USArrests dataset, we employed the following statistical procedures to explore patterns and relationships between the variables:

1. Principal Component Analysis (PCA): PCA was used to reduce the dimensionality of the dataset while retaining most of the information. The first two principal components were visualized in a scatter plot to observe patterns and relationships between states.
2. K-means Clustering: We applied the k-means clustering algorithm to group states based on their arrest rates and urban population percentages. We used the elbow method to find the optimal number of clusters (k). The resulting clusters were then visualized using a scatter plot.
3. Hierarchical Clustering: We performed hierarchical clustering to organize states into a tree-like structure based on their similarities in arrest rates and urban population percentages. We used a dendrogram to visualize the hierarchy and chose an appropriate level to cut the tree and obtain a specific number of clusters.
4. Data Visualization: Throughout the analysis, we used various plots and charts to visualize the results, such as PCA scatter plots, k-means scatter plots, and hierarchical clustering dendrograms. These visualizations facilitated the interpretation of our findings and helped convey complex patterns and relationships in a clear and concise manner.

Appendix B: Results

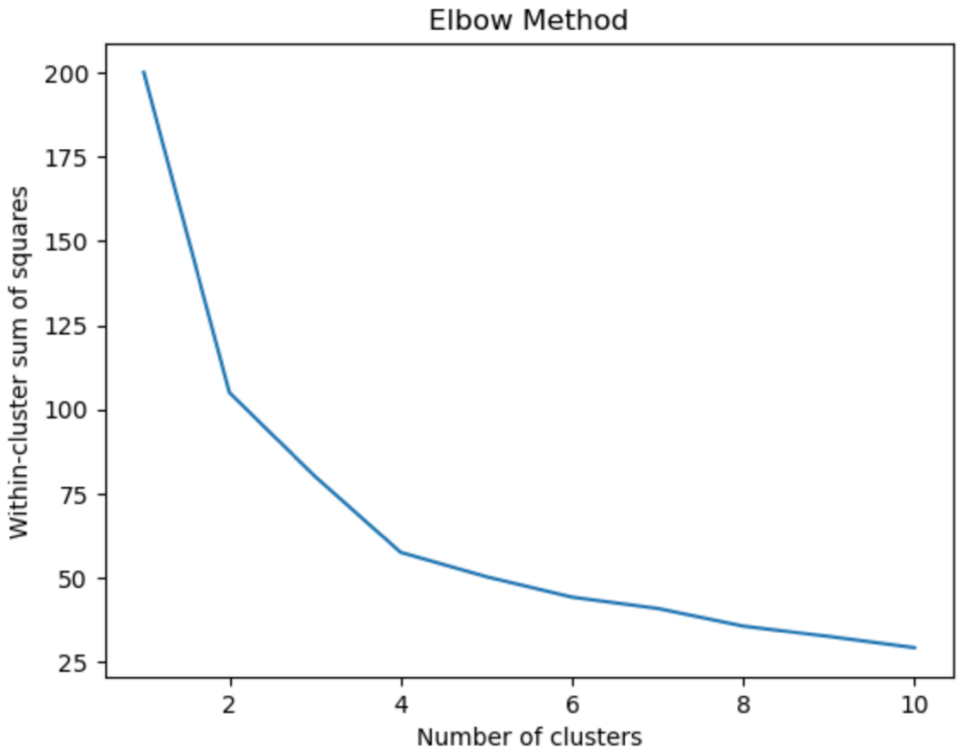
In our analysis of the USArrests dataset, we present the following results, highlighting the statistical information without discussing their impact:

Principal Component Analysis (PCA) Results: PCA was applied to reduce the dimensionality of the dataset. Since PCA is affected by the scale of the variables, it's important to standardize the features to have a mean of 0 and a standard deviation of 1. The first two principal components captured a significant portion of the total variance. The PCA scatter plot showed states in a two-dimensional space, revealing patterns and relationships between them.

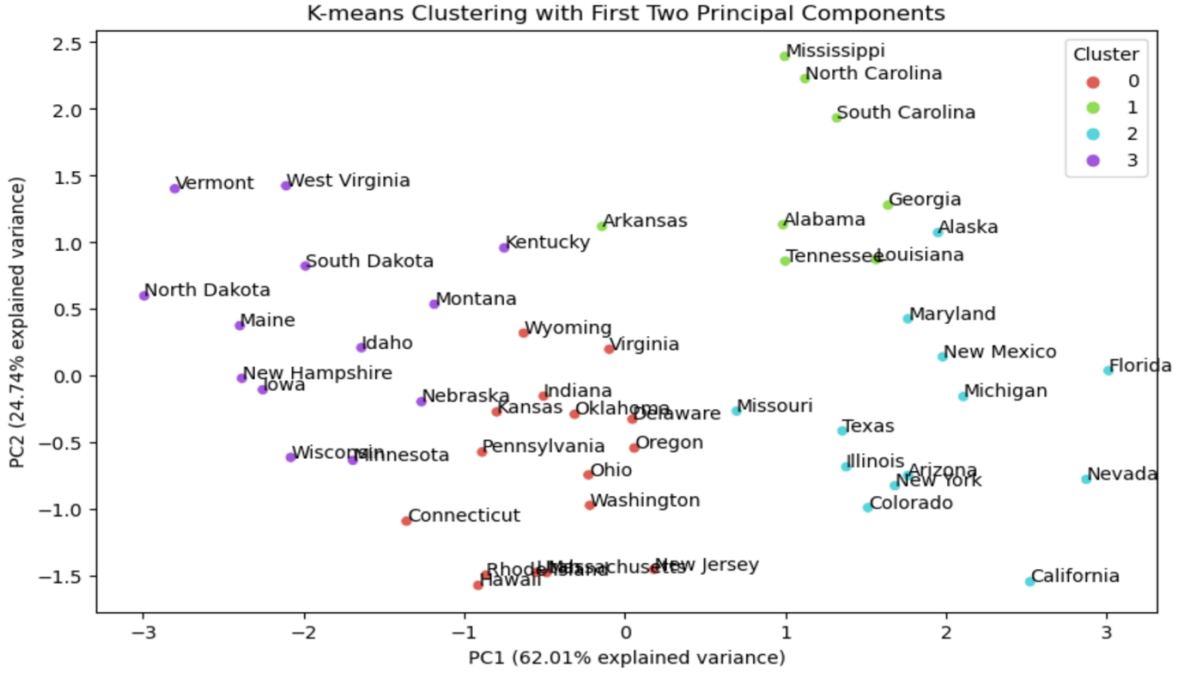


In the scatter plot, each point represents a state, and the axes correspond to the first two principal components. The percentage of explained variance for each component is displayed on the corresponding axis label. The loadings can help you interpret the meaning of the principal components. For example, if the first component has high positive loadings for Murder, Assault, and Rape, it could be interpreted as a measure of overall crime rate.

K-means Clustering Results: The elbow method suggested an optimal number of clusters (k) for k-means clustering.

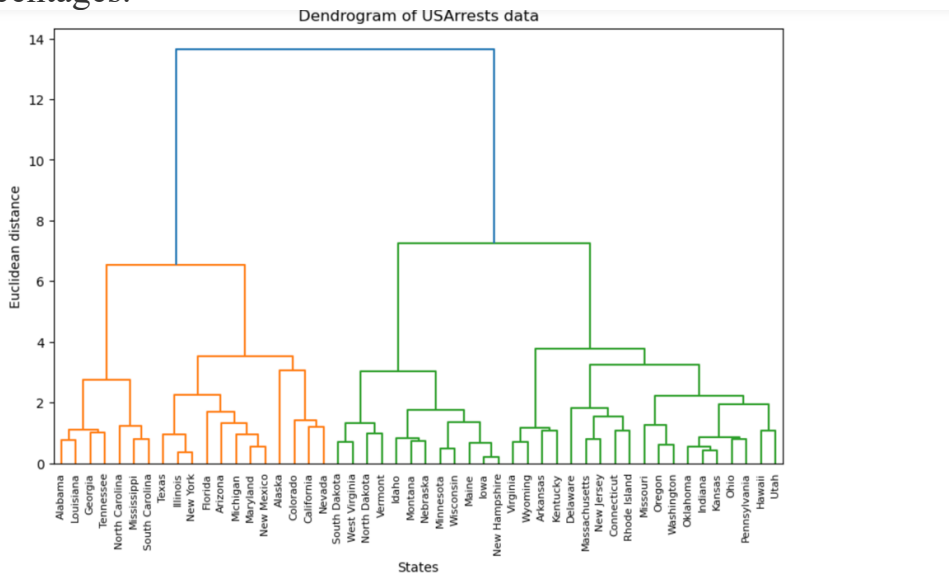


The algorithm grouped states into distinct clusters based on their arrest rates and urban population percentages, which were visualized in a scatter plot using the first two principal components.



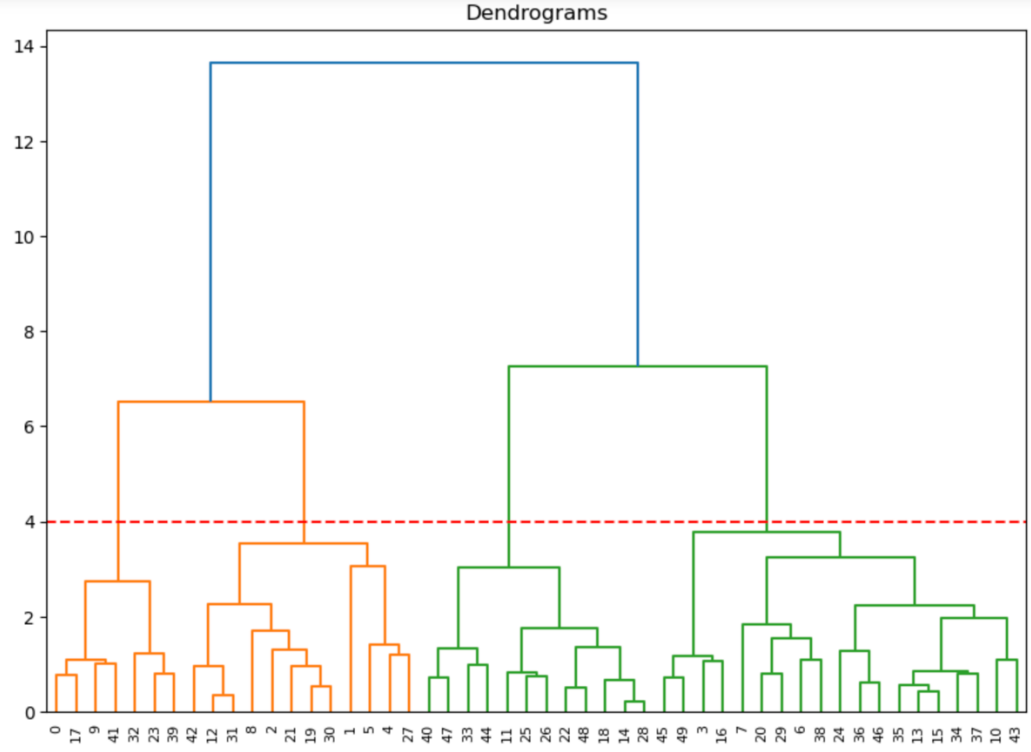
The resulting scatter plot displays the states projected onto the first two principal components, with each point colored by its k-means cluster assignment. This visualization helps in understanding the grouping of states based on the USArrests dataset.

Hierarchical Clustering Results: Hierarchical clustering organized states into a tree-like structure based on their similarities in arrest rates and urban population percentages.



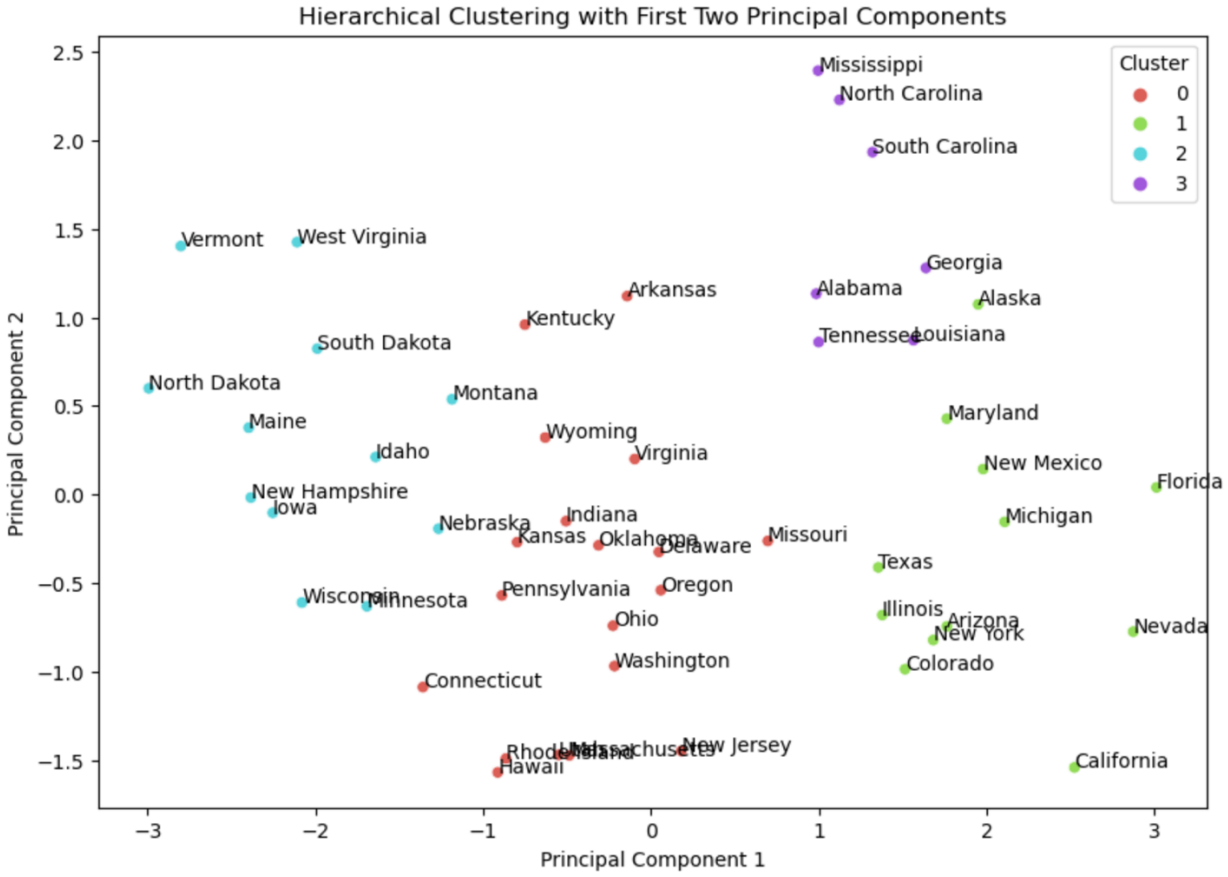
The dendrogram displays the hierarchical clustering of the states based on the USArrests dataset. The vertical axis represents the Euclidean distance between clusters, and the horizontal axis shows the states.

The dendrogram displayed the hierarchy, and we chose an appropriate level to cut the tree (at4 y=4) to obtain a specific number of clusters.



The x-axis contains the samples and y-axis represents the distance between these samples. The vertical line with maximum distance is the line and hence we can decide a threshold of 4 and cut the dendrogram:

This scatter plot shows the hierarchical clustering of the states projected onto the first two principal components, with each point colored according to its cluster assignment. It provides a 2D visualization of the hierarchical clustering, which can help in understanding the grouping of states based on the USArrests dataset.



Appendix C: Data and code:

Import necessary libraries:

```
]: import pandas as pd
   import numpy as np
   import matplotlib.pyplot as plt
   import seaborn as sns
   from sklearn.decomposition import PCA
   from sklearn.preprocessing import StandardScaler
   url = "https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/csv/datasets/USArrests.csv"
   data = pd.read_csv(url, index_col=0)
   data.head()
```

]:

	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6

PCA code:

In [116]:
```
#Standardize the data
scaler = StandardScaler()
scaled_data = scaler.fit_transform(data)
#Perform PCA
pca = PCA()
principal_components = pca.fit_transform(scaled_data)
#Create a DataFrame with the principal components
principal_components_df = pd.DataFrame(principal_components, columns=[f"PC{i+1}" for i in range(len(data.columns))],
                                       index=data.index)
loadings = pd.DataFrame(pca.components_.T, columns=[f"PC{i+1}" for i in range(len(data.columns))], index=data.columns)
explained_variance_ratio = pca.explained_variance_ratio_
plt.figure(figsize=(10, 6))
sns.scatterplot(x="PC1", y="PC2", data=principal_components_df)
plt.xlabel(f'PC1 ({explained_variance_ratio[0]*100:.2f}% explained variance)')
plt.ylabel(f'PC2 ({explained_variance_ratio[1]*100:.2f}% explained variance)')
for i, state in enumerate(principal_components_df.index):
    plt.annotate(state, (principal_components_df.iloc[i, 0], principal_components_df.iloc[i, 1]))
plt.show()
```

K-means using elbow method to find K:

:
```
wcss = []
max_clusters = 10
for i in range(1, max_clusters + 1):
    kmeans = KMeans(n_clusters=i, random_state=42)
    kmeans.fit(scaled_data)
    wcss.append(kmeans.inertia_)

plt.plot(range(1, max_clusters + 1), wcss)
plt.xlabel("Number of clusters")
plt.ylabel("Within-cluster sum of squares")
plt.title("Elbow Method")
plt.show()
```

In [124]:
```
k = 4 # assumed from elbow method
kmeans = KMeans(n_clusters=k, random_state=42)
clusters = kmeans.fit_predict(scaled_data)
data['Cluster'] = clusters
principal_components_df = pd.DataFrame(principal_components, columns=[f"PC{i+1}" for i in range(len(data.columns)-1)],
                                       index=data.index)
principal_components_df['Cluster'] = clusters
import seaborn as sns
plt.figure(figsize=(10, 6))
palette = sns.color_palette("hls", k)
sns.scatterplot(x="PC1", y="PC2", hue="Cluster", data=principal_components_df, palette=palette, legend="full")
plt.xlabel(f'PC1 ({explained_variance_ratio[0]*100:.2f}% explained variance)')
plt.ylabel(f'PC2 ({explained_variance_ratio[1]*100:.2f}% explained variance)')
for i, state in enumerate(principal_components_df.index):
    plt.annotate(state, (principal_components_df.iloc[i, 0], principal_components_df.iloc[i, 1]))
plt.title("K-means Clustering with First Two Principal Components")
plt.show()
```

Hierarchical clustering:

Perform hierarchical clustering and generate a dendrogram: We'll use the "linkage" function from SciPy's "cluster.hierarchy" module to compute the hierarchical clustering, and the "dendrogram" function to visualize the hierarchy.

```
8]: Z = linkage(scaled_data, method='ward')

plt.figure(figsize=(10, 6))
dendrogram(Z, labels=data.index, leaf_rotation=90)
plt.xlabel("States")
plt.ylabel("Euclidean distance")
plt.title("Dendrogram of USArrests data")
plt.show()
```

Cut the hierarchy and assign cluster labels: Based on the dendrogram, you can decide where to cut the hierarchy to obtain a suitable number of clusters. For example, let's say you want to have 4 clusters.

```
[129]: #Add the cluster labels to the original dataset
data['Cluster'] = cluster_labels
import scipy.cluster.hierarchy as shc4
plt.figure(figsize=(10, 7))
plt.title("Dendrograms")
dend = shc.dendrogram(shc.linkage(scaled_data, method='ward'))
plt.axhline(y=4, color='r', linestyle='--')
```

```
[134]: n_clusters = 4
agglomerative_clustering = AgglomerativeClustering(n_clusters=n_clusters)
cluster_labels = agglomerative_clustering.fit_predict(scaled_data)
from sklearn.decomposition import PCA

pca = PCA(n_components=2)
principal_components = pca.fit_transform(scaled_data)
principal_components_df = pd.DataFrame(principal_components, columns=["PC1", "PC2"], index=data.index)
principal_components_df["Cluster"] = cluster_labels
import seaborn as sns
plt.figure(figsize=(10, 7))
palette = sns.color_palette("hls", n_clusters)
sns.scatterplot(x="PC1", y="PC2", hue="Cluster", data=principal_components_df, palette=palette, legend="full")
plt.xlabel("Principal Component 1")
plt.ylabel("Principal Component 2")
plt.title("Hierarchical Clustering with First Two Principal Components")
for i, state in enumerate(principal_components_df.index):
    plt.annotate(state, (principal_components_df.iloc[i, 0], principal_components_df.iloc[i, 1]))
plt.show()
```

References:

1. Scikit-learn: Machine Learning in Python. (n.d.). Retrieved from https://scikit-learn.org/stable/
2. Jolliffe, I. T. (2002). Principal Component Analysis. Wiley Online Library.
3. Kaufman, L., & Rousseeuw, P. J. (1990). Finding groups in data: an introduction to cluster analysis (Vol. 344).
4. John Wiley & Sons.Everitt, B. S., Landau, S., & Leese, M. (2001). Cluster analysis (Vol. 26). Arnold, London.
5. Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion?. Journal of Classification, 31(3), 274-295.