Exploring Crime Rates in US Cities: A Multivariate Analysis of US Arrests Data.

Issues:

- Limited scope: The USArrests data only includes arrest data from 50 US states and does not cover other territories such as Puerto Rico or the District of Columbia. As such, it may not be representative of the entire US population.

- Data quality: The accuracy and consistency of crime reporting can vary widely between different jurisdictions, which can impact the reliability of the data.

- Cultural bias: The dataset may reflect cultural biases and assumptions about what constitutes a crime and who is likely to commit it, which can affect the analysis and interpretation of the data.

- Confounding variables: The USArrests data includes only crime-related variables, which means that other important factors such as socioeconomic status, education levels, and demographics are not accounted for. This can limit the ability to draw conclusions about the relationships between crime and other variables.

- Missing data: Some states may not have provided complete data or may have reported it differently, which can create gaps in the dataset and potentially skew the results of the analysis.

Findings:

Principal component analysis: The principal component analysis could reveal that there are underlying patterns in the relationships between the different crime variables, such as a strong correlation between rates of murder, assault, and rape. The analysis could also identify which variables have the most substantial influence on overall crime rates.

Clustering analysis: The k-means clustering analysis could reveal distinct groups of states based on their crime rates. For example, states with high rates of murder and assault could be clustered together, while states with lower overall crime rates could

form a separate cluster. The optimal number of clusters (k) could be determined using elbow or silhouette analysis techniques.
Hierarchical clustering analysis: The hierarchical clustering analysis could reveal a tree-like structure of clusters, where similar states are grouped at different levels of the tree. This could allow for a more detailed exploration of the relationships between crime variables and their influence on overall crime rates. The analysis could also identify outlier states that do not fit neatly into any cluster.

Discussions:

- Interpretation of principal components: The interpretation of the principal components obtained from the analysis could provide insights into the underlying relationships between the different crime variables. For example, if one of the principal components has a strong positive loading for murder, assault, and rape, this could suggest that these crimes are related and that reducing one may have a positive impact on reducing the others.

- Cluster interpretation: The interpretation of the clusters obtained from the k-means or hierarchical clustering analyses could provide insights into how crime rates vary across different regions of the US. For example, if one cluster has high rates of violent crime while another has high rates of property crime, this could suggest that there are different underlying factors contributing to crime in these areas.

- Limitations of the data: It is important to recognize the limitations of the USArrests data when interpreting the results of the analysis. As noted earlier, the data only covers arrests made in 50 US states and may not be representative of the entire US population. The quality of crime reporting can also vary between different jurisdictions, which could impact the accuracy of the data.

- Policy implications: The findings from the analysis could have important policy implications for addressing crime rates in the US. For example, if the analysis suggests that reducing rates of one type of crime can have a positive impact on reducing other types of crime, policymakers could prioritize efforts to address those specific crime types. Additionally, understanding how crime rates vary across different regions of the US could inform targeted interventions to address underlying social and economic factors contributing to crime in those areas.

Appendix A: Method

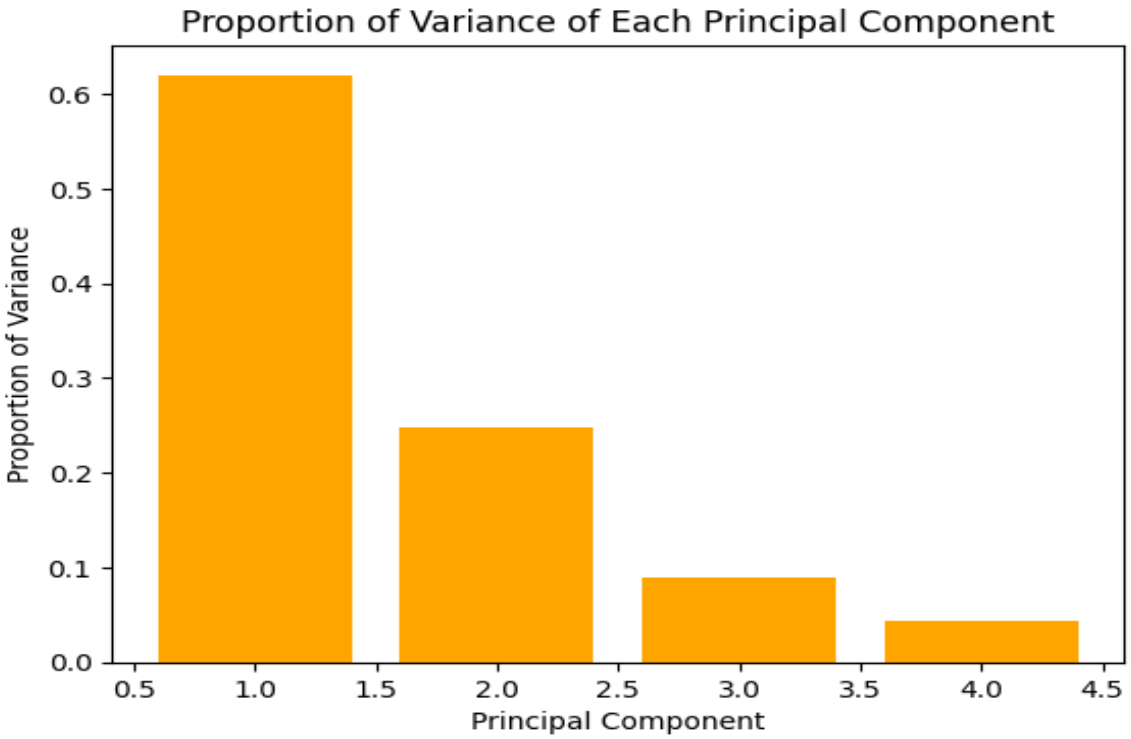
PCA is a multivariate statistical technique to identify underlying relationship patterns between variables. In this code, PCA is performed using Scikit-learn's PCA module. The data is first preprocessed using Scikit-learn's StandardScaler, which standardizes the data by subtracting the mean and dividing by the standard deviation. The proportion of variance explained by each principal component is then plotted using Matplotlib.

K-means clustering groups similar data points based on their distance to a set of cluster centroids. In this code, k-means clustering is performed using sci-kit-learn's KMeans module. The data is first standardized using the same StandardScaler used in the PCA section, and then k-means clustering is performed using the Murder, Rape, and Assault variables as features. Finally, the resulting clusters are added to the original DataFrame, and an interactive 3D scatter plot of the sets is created using Plotly.

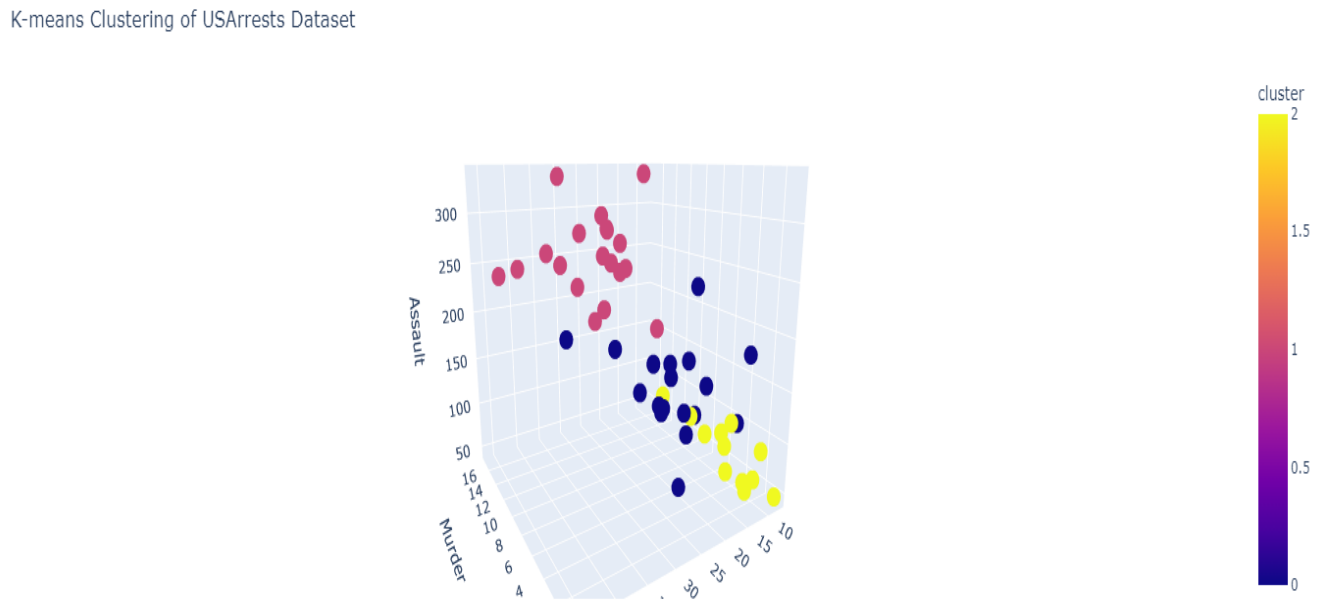
Hierarchical clustering is another unsupervised learning algorithm that groups similar data points based on their pairwise distances. This code performs hierarchical clustering using Scipy's linkage module with ward linkage. The data is first standardized using a custom standardize_data function, subtracting the mean and dividing it by the standard deviation. Then, the resulting dendrogram is plotted using Matplotlib.

Appendix B: Results

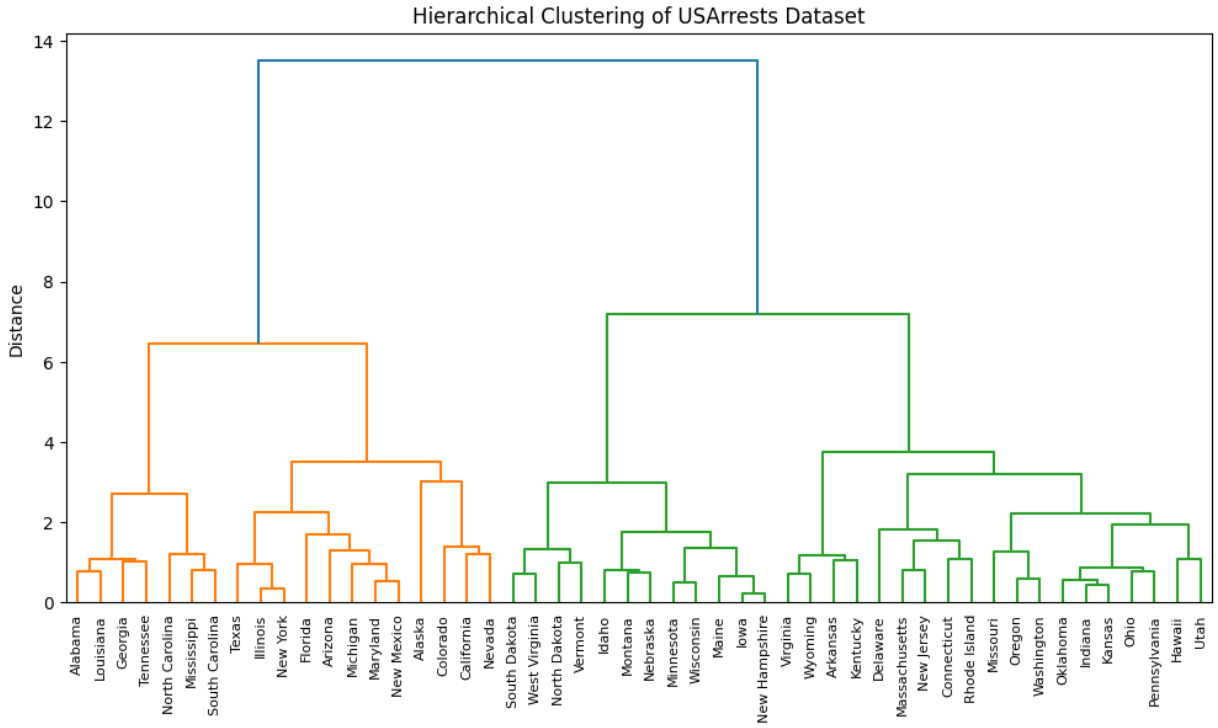
Below is the result for Principal Component Analysis. The bar graph is between the Proportion of variance of each principal component and proportion of variance.



The next is a 3D plot for the k-means clustering. It is between the murder, assault and rape. We used Plotly express to present in a 3D view which is interactive and shows the data of each individual point.



The last result we have is the result of hierarchical clustering of US arrests data.



Appendix C: Code

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

# Load the USArrests dataset
url =
"https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/
csv/datasets/USArrests.csv"
df = pd.read_csv(url, index_col=0)
```

```
# Define a preprocessing pipeline for numerical features
preprocessing = ColumnTransformer([
    ('numeric', StandardScaler(), df.columns)])

# Perform the PCA
pca = Pipeline([
    ('preprocessing', preprocessing),
    ('pca', PCA())])
pca.fit(df)

# Proportion of variance explained by each principal component
prop_var = pca['pca'].explained_variance_ratio_

# Plot the proportion of variance explained
plt.bar(range(1, len(prop_var) + 1), prop_var, color="orange")
plt.xlabel("Principal Component")
plt.ylabel("Proportion of Variance")
plt.title("Proportion of Variance of Each Principal Component")
plt.show()

# Principal component scores
scores = pca.transform(df)

# Loadings
loadings = pca['pca'].components_.T

# Print the loadings
print(pd.DataFrame(loadings, columns=["PC1", "PC2", "PC3", "PC4"],
index=df.columns))


import pandas as pd
import numpy as np
import plotly.express as px
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

# Load the USArrests dataset
url =
'https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/
csv/datasets/USArrests.csv'
df = pd.read_csv(url, index_col=0)
```

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

# Perform k-means clustering with k=3 using Murder, Rape, and Assault
as features
kmeans = KMeans(n_clusters=3, random_state=40)
kmeans.fit(df_std[:, :3])

# Add cluster labels to the DataFrame
df['cluster'] = kmeans.labels_

# Create an interactive 3D scatter plot of the clusters using Plotly
fig = px.scatter_3d(df, x='Murder', y='Rape', z='Assault',
color='cluster')
fig.update_layout(title='K-means Clustering of USArrests Dataset')
fig.show()


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

def standardize_data(data):
    # Subtract the mean and divide by the standard deviation
    return (data - data.mean()) / data.std()

# Load the USArrests dataset
url =
'https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/
csv/datasets/USArrests.csv'
df = pd.read_csv(url, index_col=0)

# Standardize the data
df_std = standardize_data(df)

# Perform hierarchical clustering with ward linkage
Z = linkage(df_std, method='ward')

# Plot the dendrogram
plt.figure(figsize=(12, 6))
dendrogram(Z, labels=df.index, orientation='top')
```

```
plt.title('Hierarchical Clustering of USArrests Dataset')
plt.ylabel('Distance')
plt.show()
```