Hyperiondev

TASK

Principal Component Analysis And Clustering Of US Arrests Dataset

Visit our website

Introduction

In this project we will apply principal component analysis and clustering techniques on the UsArrests dataset from the US Arrests Kaggle challenge. The dataset contains information on the state and the percentage of the population that lives in the urban areas in the state (UrbanPop). The data set also contains information on the statistics on Murder, Assault and Rape arrests per state.

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

The data set has 50 rows and 5 columns and the first step is to see if it contains any missing values on any of the rows and columns. Missing data points can be handled using different types of imputation such as the average, mode, median etc, and some machine learning techniques.

MISSING DATA

Missing data can be handled by removing the rows with the missing values, but that would result in the exclusion of information depending on the extent of the missing data. Imputation is the best way to handle the missing data, since it helps with retaining the collected information.

The UsArrests dataset does not have any missing data points as can be seen in the figure below. This is fortunate as it means no data will be removed and no imputation is required. All the information that was collected is in the data set.

	Missing Values
City	0
Murder	0
Assault	0
UrbanPop	0
Rape	0

The next step is to study the relationships that are contained in the data collected and what information can be gained.

DATA EXPLORATION

To remove confusion the City column has been renamed with the proper name of State, as this column contains the names of US States and not cities. The statistics describing the collected data can be seen in the figure below.

	Murder	Assault	UrbanPop	Rape
count	50.00000	50.000000	50.000000	50.000000
mean	7.78800	170.760000	65.540000	21.232000
std	4.35551	83.337661	14.474763	9.366385
min	0.80000	45.000000	32.000000	7.300000
25%	4.07500	109.000000	54.500000	15.075000
50%	7.25000	159.000000	66.000000	20.100000
75%	11.25000	249.000000	77.750000	26.175000
max	17.40000	337.000000	91.000000	46.000000

From the statistical data above we can see that Assault has significantly high values for the min, max , std and mean. This indicates that the data will need scaling as this feature would predominantly affect the analysis of the machine learning.

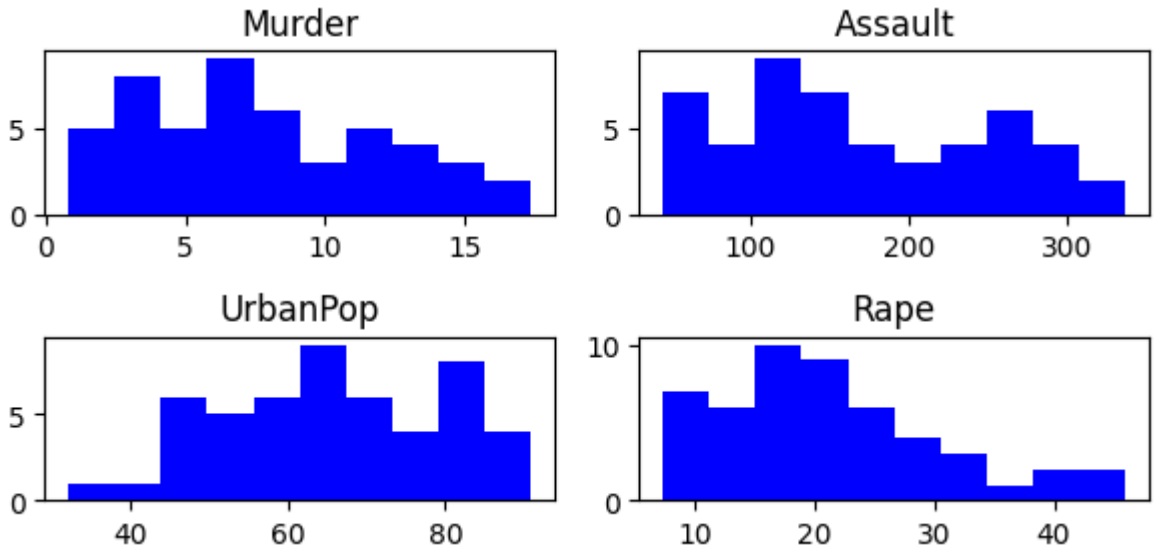
From the figure below we can see that we are working predominantly with continuous variables of type integer and float.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   State     50 non-null     object
 1   Murder    50 non-null     float64
 2   Assault   50 non-null     int64
 3   UrbanPop  50 non-null     int64
 4   Rape      50 non-null     float64
dtypes: float64(2), int64(2), object(1)
memory usage: 2.1+ KB
```

The data can best be summarised by the information in the table below. This table tells us about the mean, minimum, maximum and standard deviation of each feature.

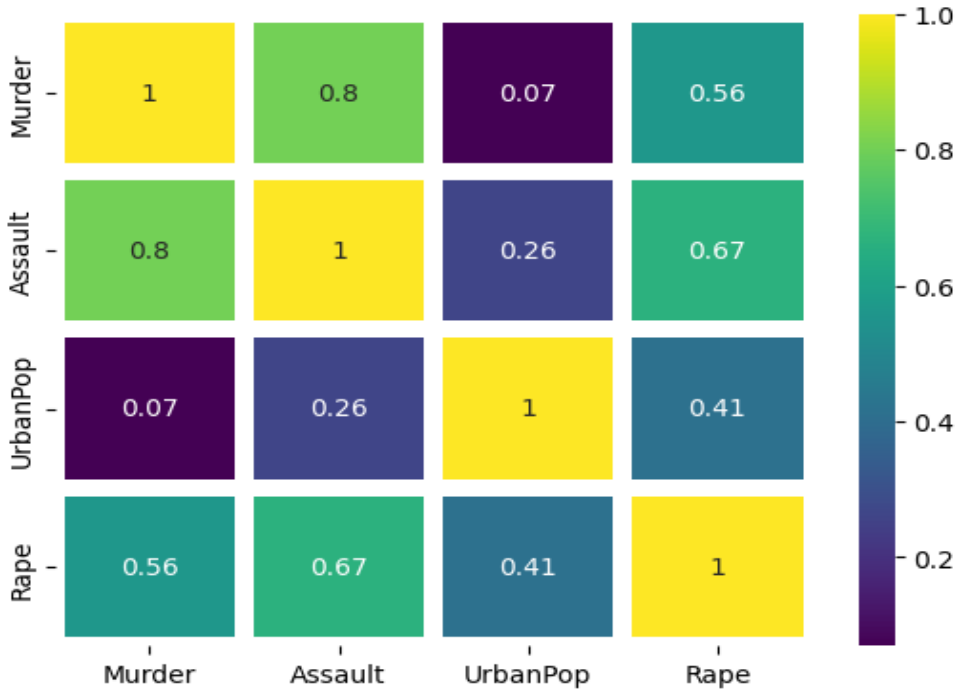
	missing	mean	std	min	max
Murder	0	7.788	4.355510	0.8	17.4
Assault	0	170.760	83.337661	45.0	337.0
UrbanPop	0	65.540	14.474763	32.0	91.0
Rape	0	21.232	9.366385	7.3	46.0

Looking at the above table, we can notice once more that the assault row contains large numbers in comparison to the other features and therefore scaling will be necessary for this data set. The distribution of each of the features of the datasets can be seen in the figure below.



CORRELATION ANALYSIS

From the plot below all the variables are positively correlated with some variables with weak correlation and some with strong correlation. The correlation coefficient can give insight of the linear relationship of two variables, negative correlation meaning one variable increases as the other decreases, correlation of zero meaning no linear relationship between variables and positive correlation meaning there is an increasing relationship between the variables.



The correlation plot above shows that assault is strongly correlated with murder, this makes sense as violence can usually lead to someone dying.

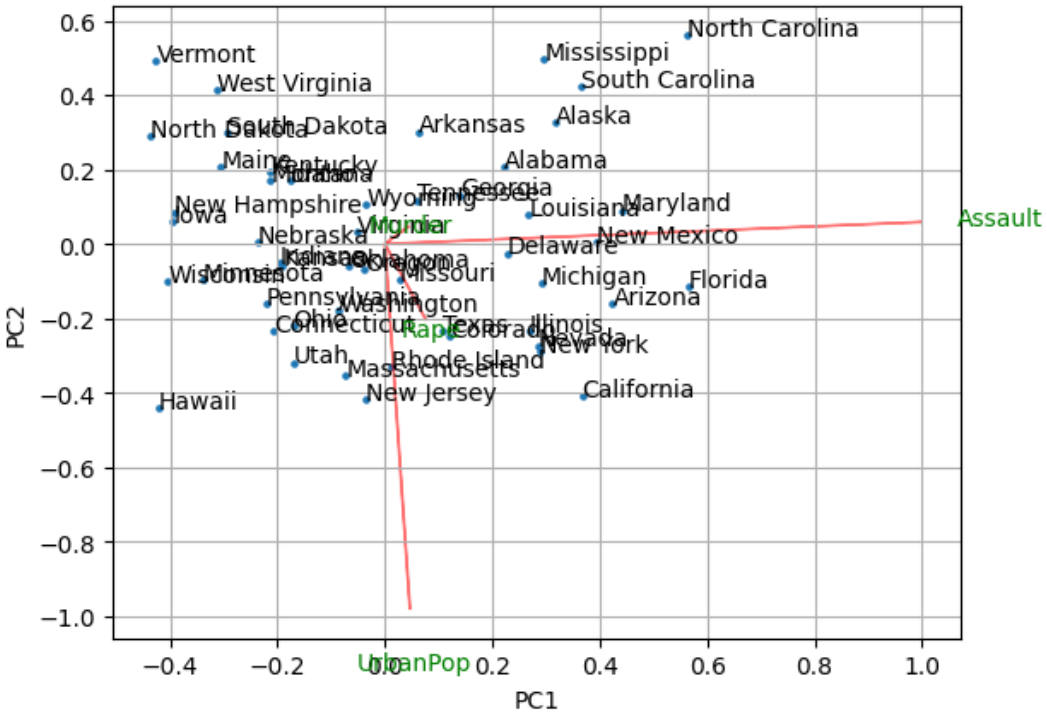
In contrast the correlation between murder and rape is not as strong, this may be due to the fact that rape occurs in many different levels of society with different levels of violence involved, and this may show that most people are not killed during the incident but may just be intimidated into silence and fear is used to keep them silent so that this act may occur multiple times without reporting it.

The correlation plot also suggests that living in a more urban area you are likely to be assaulted and more likely to be a victim of rape, but least likely to be a victim of murder as the correlation coefficient is close to zero. But this coefficient does suggest that it is rare but does still occur.

PCA: UNSTANDARDISED DATA

Principal component analysis is a technique that is used to reduce the dimensionality of a data set. This helps with removing redundancies in the data set by looking for principal components and thus optimise the learning of the machine learning algorithm.

As stated before, this data set requires scaling before the principal component analysis is conducted. The biplot below is to demonstrate the effects of un-scaling the data on our analysis.



The biplot above indicates the importance of each feature by the length of its arrow (red) as seen in the biplot above, this arrow corresponds to the magnitude values of the eigenvectors. From the above plot we can see that Assault and UrbanPop are the two most important features as the arrows dominate the biplot.

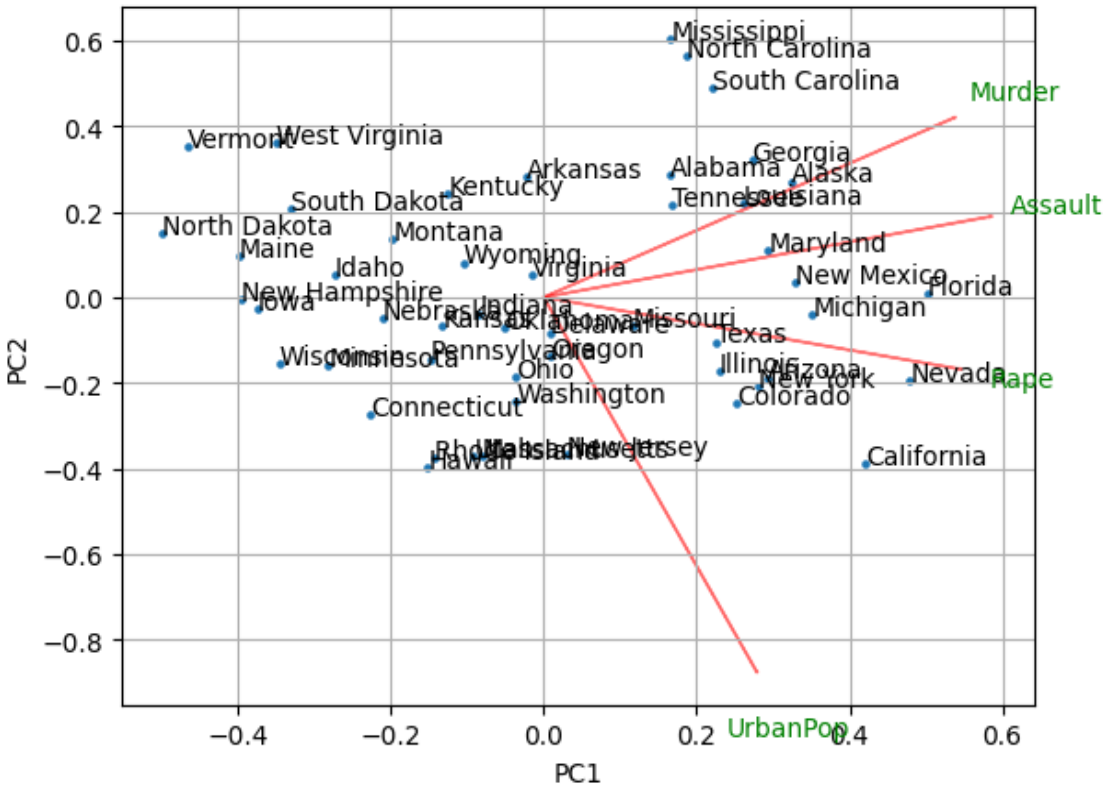
The table below summarises the importance of these features and we can see that Assault has the highest and entirely dominates importance on the first principal component, while the same is true for UrbarnPop on the second principal component.

	Features	PC1 Importance	PC2 Importance
0	Murder	0.041704	0.044822
1	Assault	0.995221	0.058760
2	UrbanPop	0.046336	0.976857
3	Rape	0.075156	0.200718

This makes sense as the magnitudes of these features are significantly higher than those of Murder and Rape. The next step is to standardise the data and plot another biplot after this feature scaling to see if the same results will be communicated.

PCA: STANDARDISED DATA

The feature scaling of the data has been performed using the sklearn StandardScaler. The biplot below demonstrates the effects of scaling the data on our analysis.



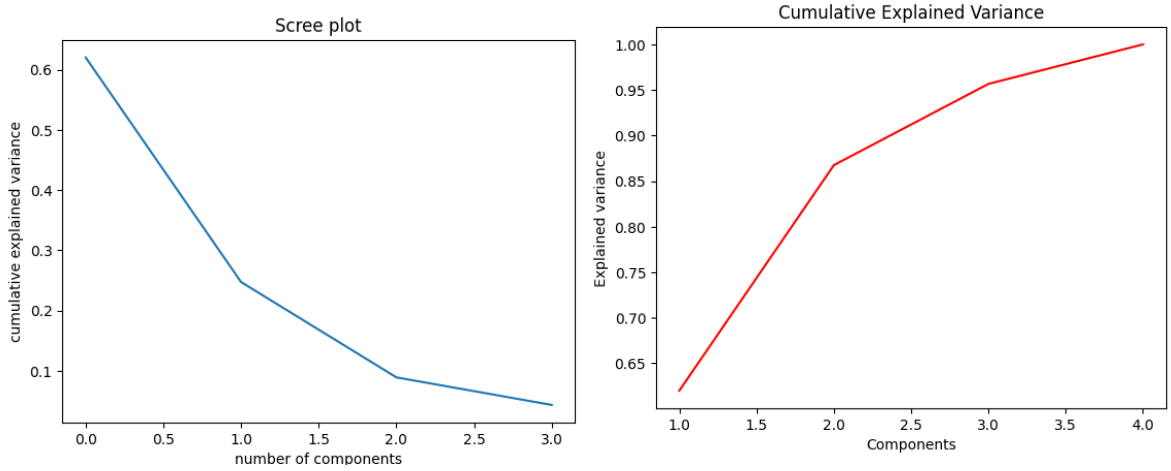
The biplot above shows that scaling reduces the effects of large scale data dominating the principal component analysis. It can be noted that states like Nevada, Arizona and New York are dominated by rape, while states such as Maryland, New Mexico and Florida are dominated by assault and states such as Alaska, Georgia and Louisiana are dominated by murder.

The importance of the features is summarised in the table below. Assault still has higher importance on the first principal component but Murder and Rape are not that far off, while on the second principal component UrbanPop is still dominating. This makes sense as the UrbanPop is least correlated with the other features.

	Features	PC1 Importance	PC2 Importance
0	Murder	0.535899	0.418181
1	Assault	0.583184	0.187986
2	UrbanPop	0.278191	0.872806
3	Rape	0.543432	0.167319

Principal component analysis is useful in reducing the dimensionality of the data set, and thus when using this technique it is important to choose the appropriate number of principal components that explain most of the variance in the data set.

This can be achieved by using the Scree plot and Cumulative Explained Variance plot. These plots can be seen below.



The first 3 principal components together explain around at least 95% of the variance, therefore we can use them to perform cluster analysis. The dimensions of the data set can be reduced from 4 to 3.

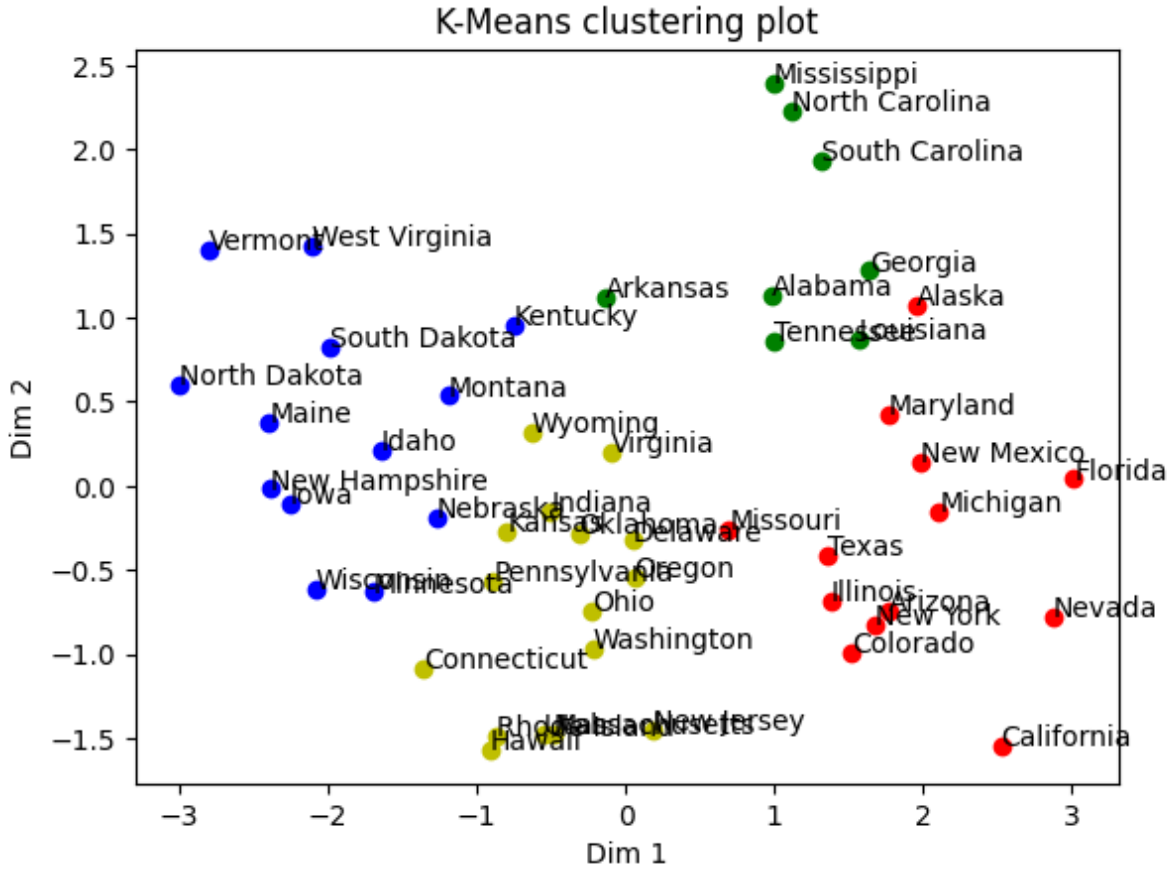
CLUSTER ANALYSIS

For the clustering we will look at two clustering techniques, namely k-means and hierarchical clustering. First we will look at k-means clustering as seen below.

K-MEANS CLUSTERING

K-means clustering is a simple unsupervised machine learning algorithm that is used to group together similar data points of a data set and find underlying patterns using a fixed number of specified clusters popularly denoted as k.

For our analysis the chosen value of k is 4 meaning we are looking to group the data into 4 clusters. A visualisation of the clusters can be seen in the figure below.

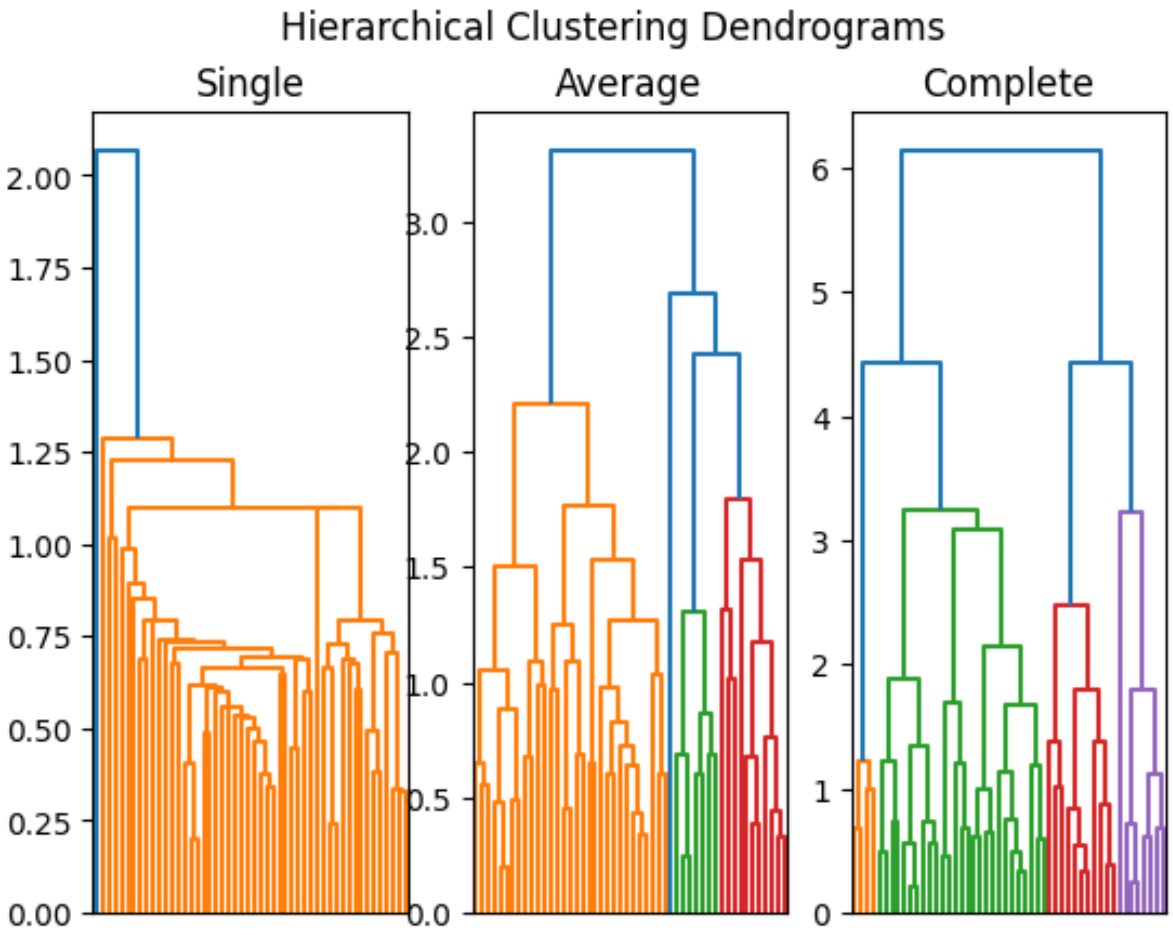


It seems the k-means algorithm has clustered together these states based on the severity of the crimes or magnitudes of the numbers on each of the features. States with large numbers are mostly clustered together and those with the least numbers are clustered together. States with the blue dots have the least numbers in arrest for these crimes while states in red dots have the highest numbers in at least two features (crime categories).

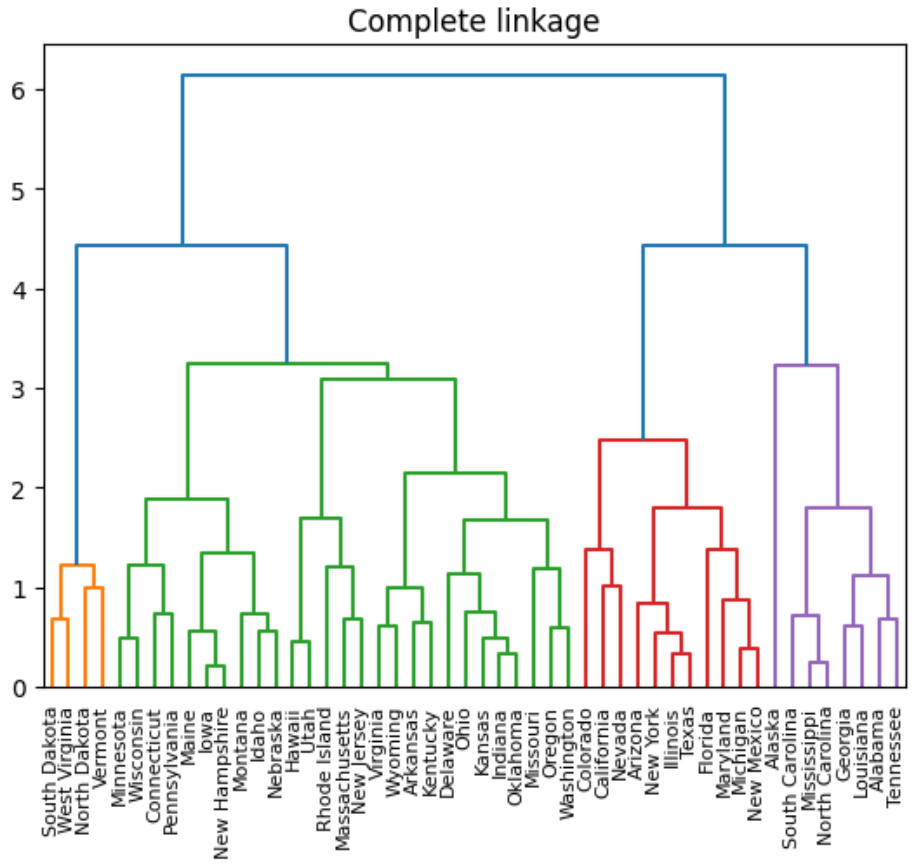
HIERARCHICAL CLUSTERING

The dendrogram plot can be used to visualise the clusters of the hierarchical clustering algorithm before the algorithm is run, this is possible as a dendrogram does not require a specification of the number clusters and can be used to decide the number of clusters after running the hierarchical clustering algorithm.

The linkage method to be used to measure the distance between clusters is determined by plotting and comparing different dendrograms for single, complete and average. The distance metric used is the Euclidean.



The complete linkage method is the method of choice to use as it creates the most balanced spread of clusters. The dendrogram using this method can be seen in the figure below, this is to isolate the dendrogram for clear visualisation.



From the dendrogram above we can see that there are four clusters of sizes 4, 27, 11 and 8. The clustering numbers are not far off from those gotten from the k-means algorithm and some states are clustered similarly in both cases. The clustering is done based on the severity of the crime arrests in the states.

The states with the least numbers in crime arrests are in orange, followed by the ones in green, then the ones in red and the states with the highest numbers in crime arrests clustered together in purple.

THIS REPORT WAS WRITTEN BY : Thabiso Maqhajana