
TASK

# Capstone Project VII

# Introduction

In this task, we explore the number of arrests for Assault, Murder and Rape and the percentage of the population living in an urban area for all 50 US states during the year 1974. Using unsupervised learning methods such as Principal Component Analysis (PCA) and various clustering techniques we will attempt to generate some analysis on this dataset.

| | City | Murder | Assault | UrbanPop | Rape |
|---|---|---|---|---|---|
| **0** | Alabama | 13.200 | 236 | 58 | 21.200 |
| **1** | Alaska | 10.000 | 263 | 48 | 44.500 |
| **2** | Arizona | 8.100 | 294 | 80 | 31.000 |
| **3** | Arkansas | 8.800 | 190 | 50 | 19.500 |
| **4** | California | 9.000 | 276 | 91 | 40.600 |

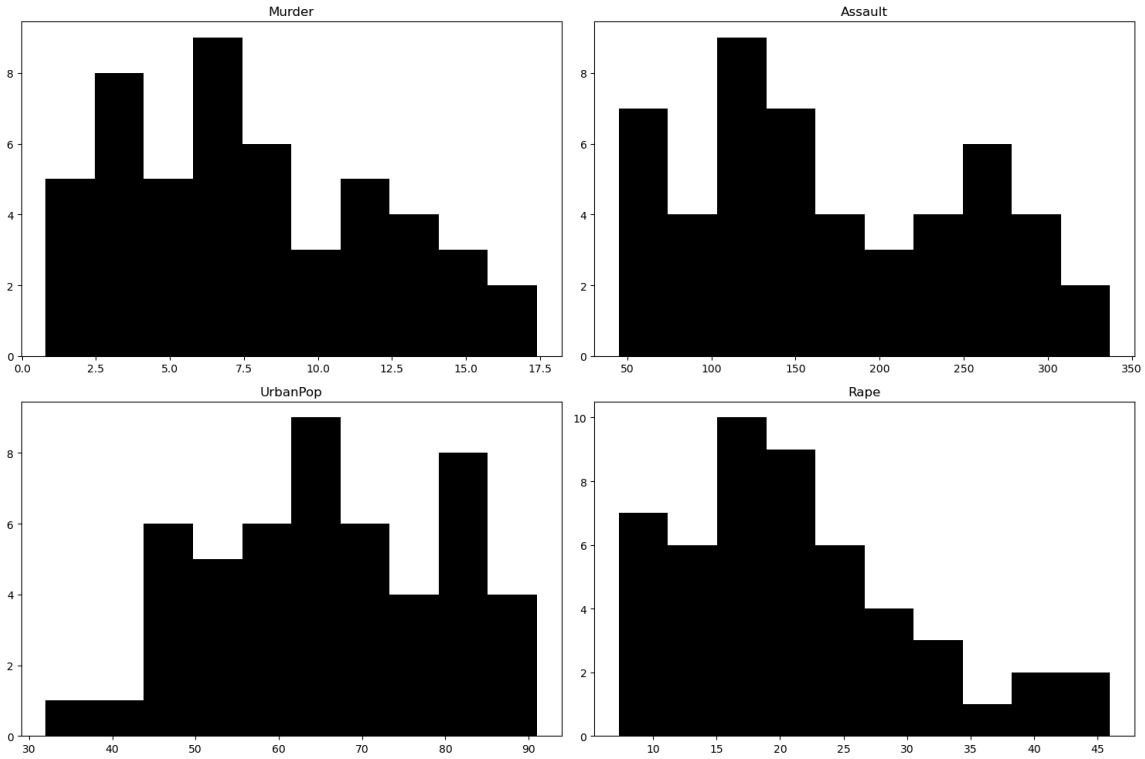
For this data, the number of arrests is per 100,000 residents.

## Exploring the data

Before we start our analysis on this data set we will first check for any inconsistencies in the data set and if so cleanse the data so we can then move onto the analysis. Our first check is for any missing values and that all the variables are stored as the correct data type.]

```
<class 'pandas.core.frame.DataFrame'>
Index: 50 entries, Alabama to Wyoming
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Murder    50 non-null     float64
 1   Assault   50 non-null     int64  
 2   UrbanPop  50 non-null     int64  
 3   Rape      50 non-null     float64
dtypes: float64(2), int64(2)
```

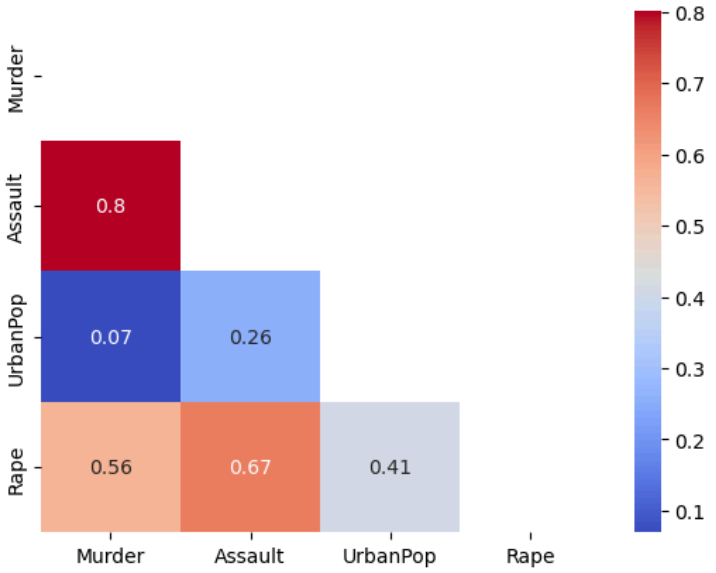
Here we can see that there are no missing entries in the data set and all 4 variables are of the correct data type so we can perform our analysis. We will now plot the histograms of each of our 4 variables to see their distribution and if they will require any scaling.




Here we can see that the Assault variable is going to dominate our analysis as it is considerably bigger than the other variables. This means that scaling will be useful to allow us to see the impact of the other variables better, and allow us to perform a better analysis.

## Correlation Analysis

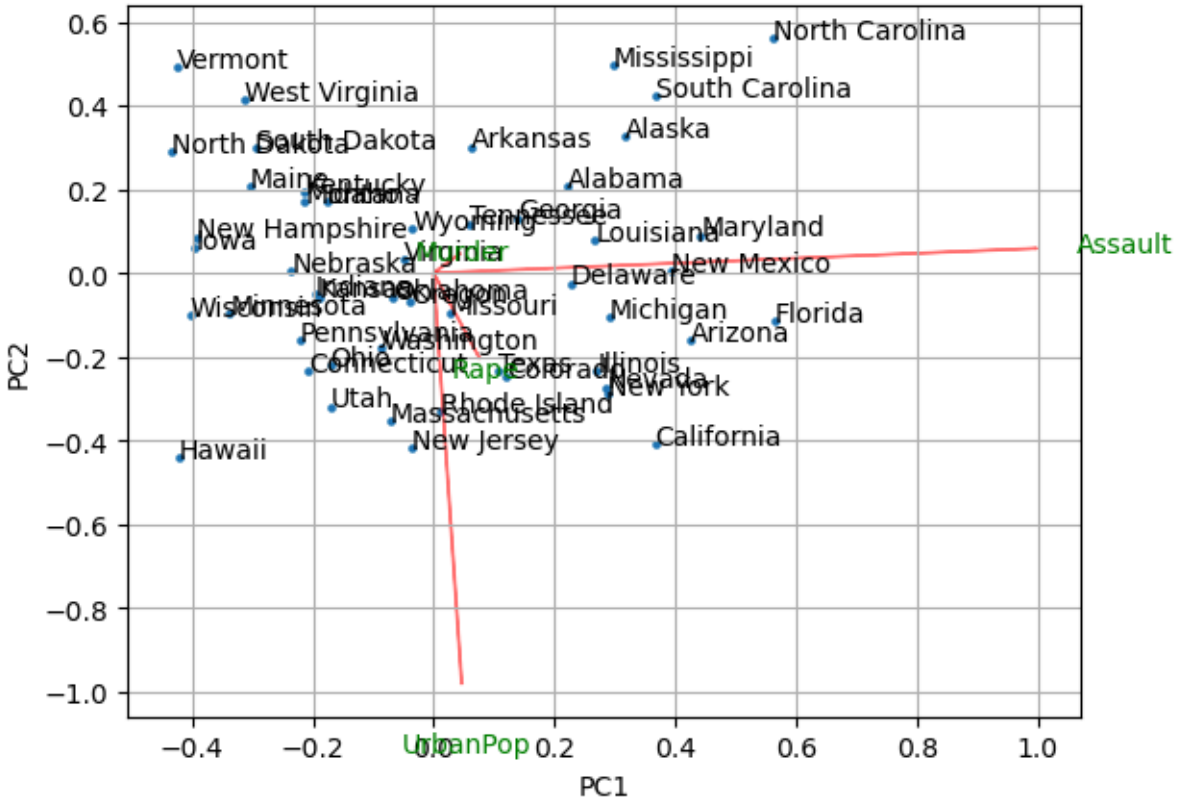
We now want to see how much correlation there is between our 4 variables and if there is a possibility to reduce the number of variables needed for our analysis.



Here we can see that there is a very strong correlation between Murder and Assault which is what we would expect since they are both violent crimes and you would expect that a city that has a large number of assaults would also have a similarly high number of murders. The Urban Population has the weakest correlation among the other 3 variables. Since there is a high level of correlation between the variables this data set is a good candidate for Principal Component Analysis.

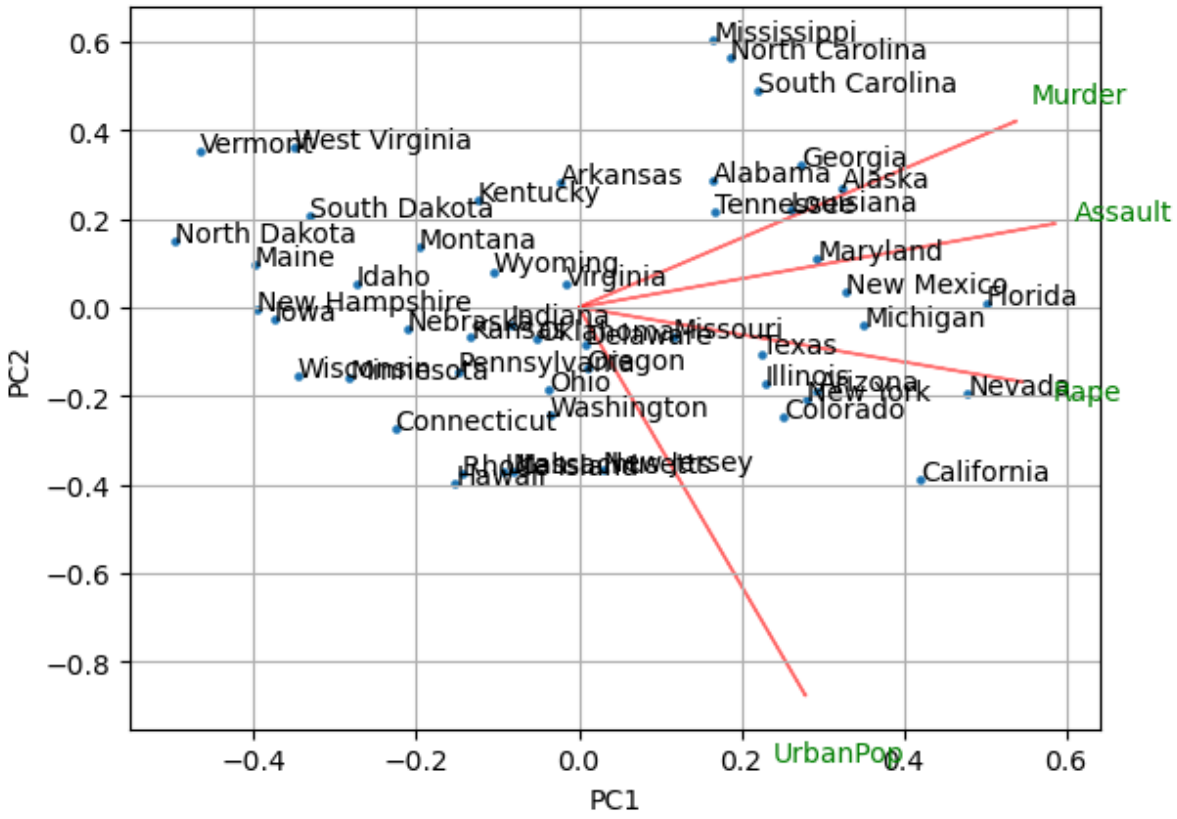
## PCA: UNSTANDARDISED DATA

Principal Components Analysis (PCA) is a method for finding the underlying variables (i.e. principal components) that best differentiate the observations by determining the directions along which your data points are most spread out. We shall first perform PCA on our data set without scaling the data to see the impact scaling has on the analysis.



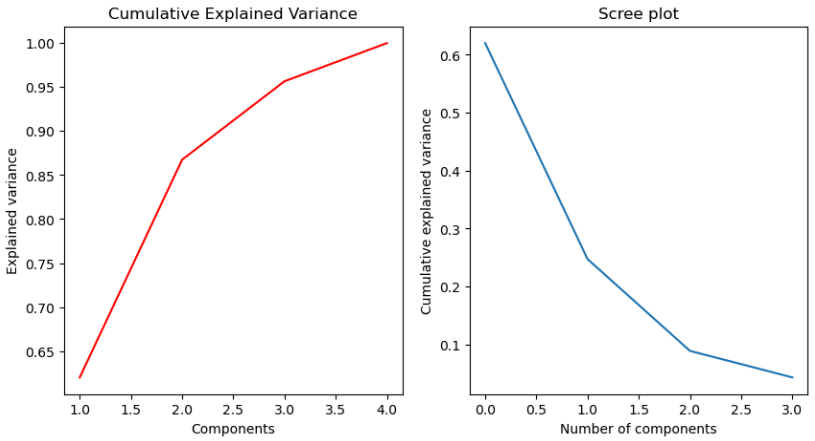

Here we can see a Bi-plot of the data where we have plotted the data using our first 2 principal components. From this graph, we can see that the Assault variable is dominating the 1st principal component and you struggle to see the contributions of the Murder and Rape variables. The 2nd principal component is dominated by the Urban Population variable. We will now apply standard scaling to the data and perform our PCA again and see how this impacts the analysis.

## PCA - Standardised data




After applying the scaling to the data we can now see the impact of all the variables much more clearer. We can now see how all 4 variables contribute more to the 1st principal component compared to before, also although Urban Pop is still the main contributor to the 2nd component the other 3 variables contribute more than before. Also now we can see the cities have started to form into more distinct clusters than before without the scaling.

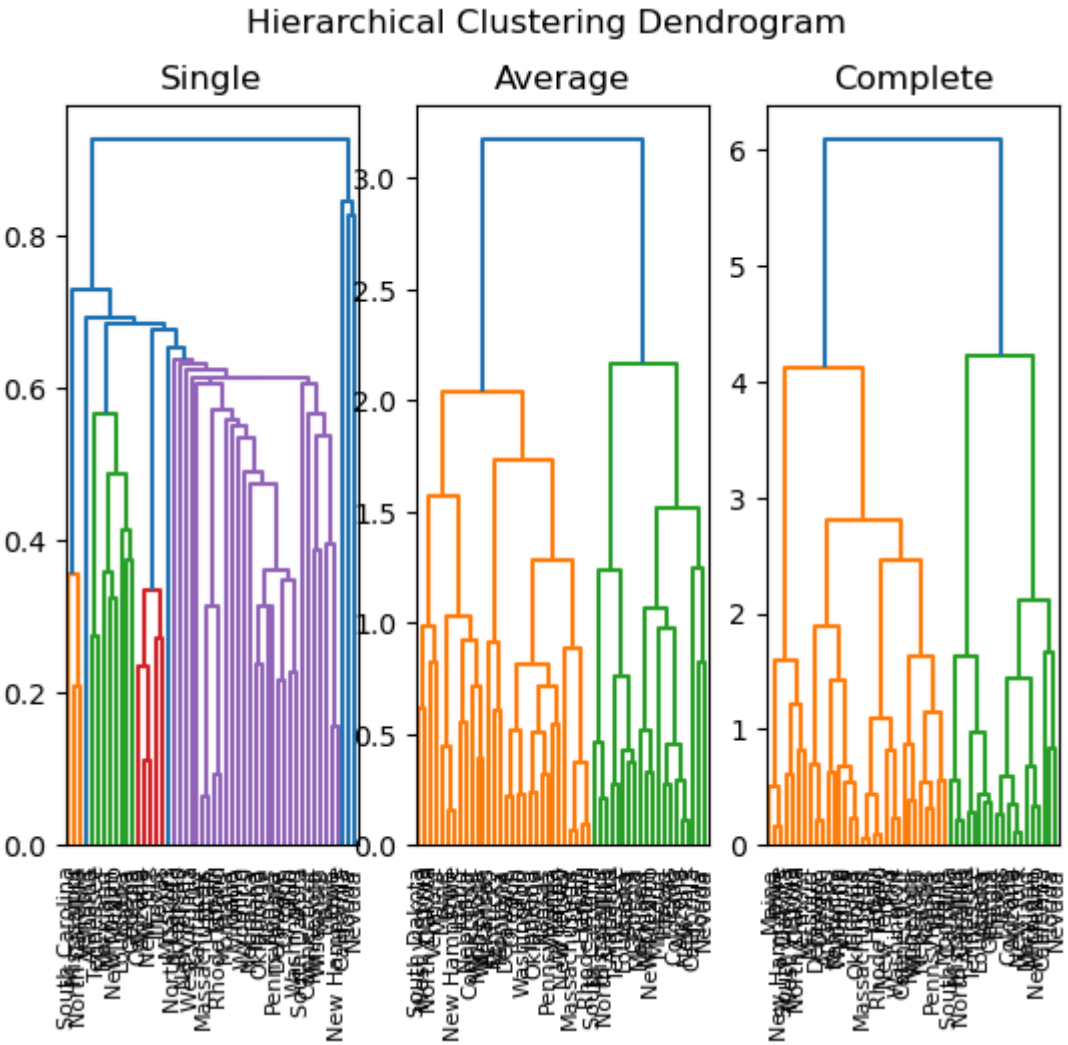
In PCA, the first few principal components are the variables that explain most of the variation in the data. As such, when using PCA for dimensionality reduction, we need to choose an appropriate number of principal components that explain a significant portion of the variation in our data. This decision will be aided by the Scree plot and Cumulative Explained Variance plot, below.



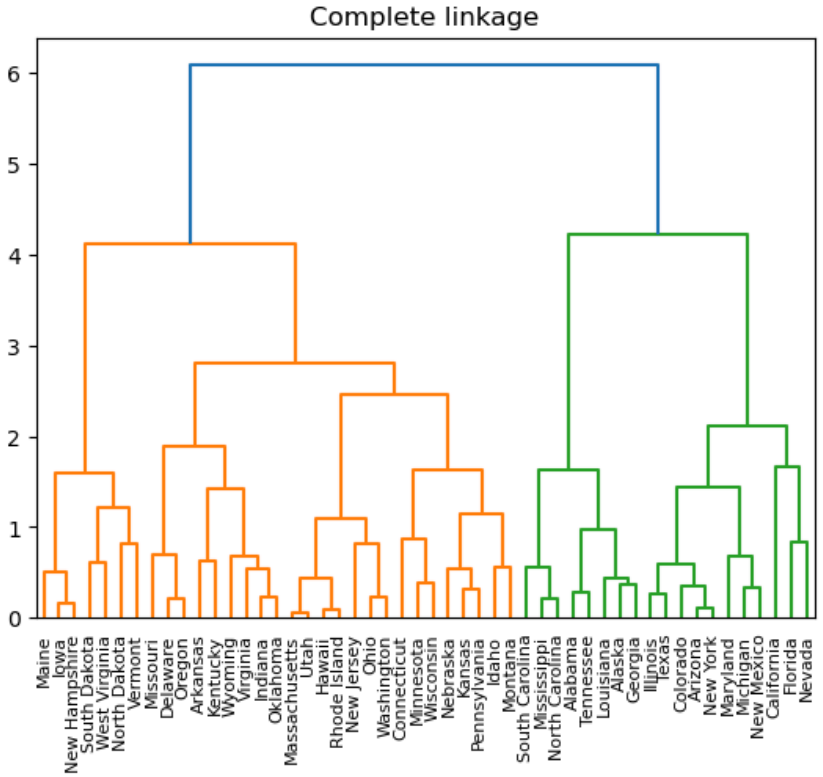
From these graphs, we can see that we can explain almost all of the variance of the data when using only 2 components which means we can reduce the dimensions of our data and just concentrate on using 2 components without missing out on any important features of the data.

## Cluster Analysis - Hierarchical

We will perform two types of cluster analysis the first being Hierarchical clustering. Hierarchical clustering has the advantage that we can see the clusters visually in a dendrogram and doesn't have to specify the number of clusters before running the algorithm. However, we will have to decide the number of clusters after the algorithm runs. For the distance metric between observations, Euclidean distance was used, which is the most common way to measure distance. To determine the method used to measure the distance between clusters, we plotted the various dendrograms for the single, complete, and average linkage methods.



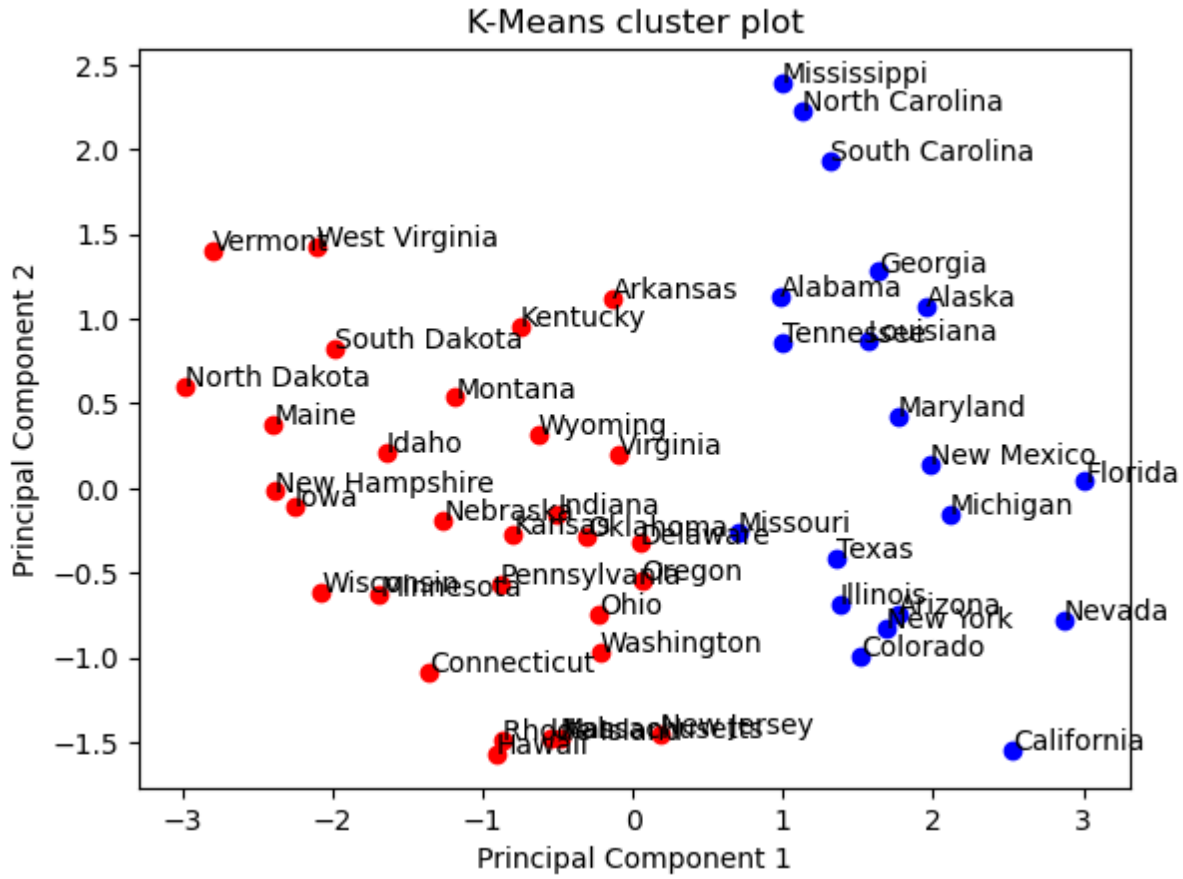
From these dendrograms we can see that Average and Complete perform the most balanced dispersion of clusters so for the rest of our analysis we shall choose to use Complete as there isn't much difference between that and Average. Below is a clearer view of the Complete linkage method.




From the dendrogram we get the biggest distance between the clusters when there are 2 clusters which are represented by the blue lines, this means by separating the data into two clusters the clusters should be more dissimilar than separating them into a greater number of clusters. From this, we can now move onto K-means clustering using a value of k=2 which we have decided from our dendrogram.

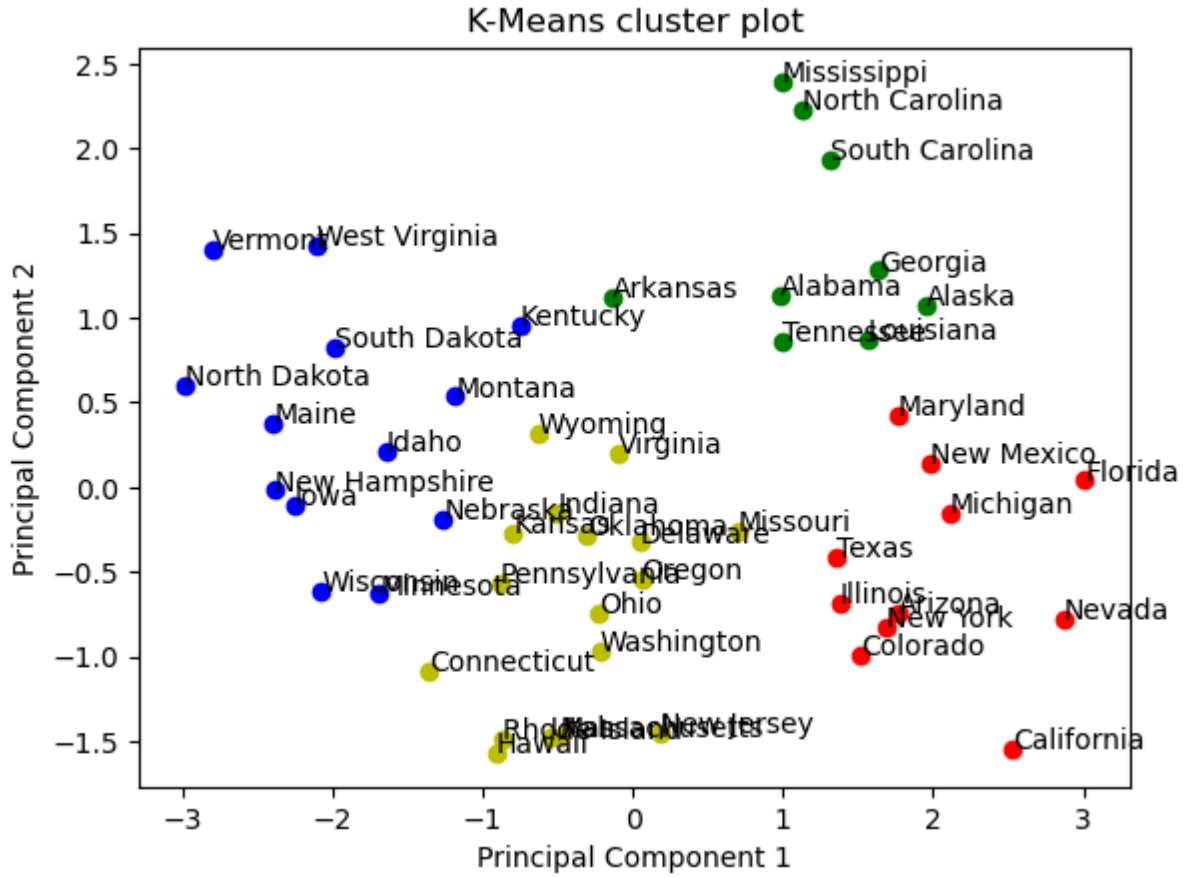
## K means Clustering

K-means is a very popular clustering partitioning algorithm that is fast and efficient and scales well for large datasets. It is an iterative process, so observations can switch between clusters while the algorithm runs until it converges at a local optimum. This method is not robust when it comes to noise data and outliers and is not suitable for clusters with non-convex shapes. Another drawback with K-means is the necessity of specifying K in advance.




Here is the result of our K-means clustering when k has been set to 2. The data has been grouped into 2 distinct clusters but from looking at the plot we can see that the clusters are not particularly tightly grouped. We can assess the quality of our clustering method by calculating the silhouette score which is 0.41 which isn't a particularly great score and suggests that the data isn't well suited to clustering techniques.

From the dendrogram, the other potential value for k is 4 so we will plot the K means cluster again and see if we obtain a better result.




Again we can see there are 4 distinct clusters but within the clusters, the data is spread out and not grouped tightly together. If we calculate the silhouette score again for these clusters we get a score of 0.34 which is worse than before. Therefore the best possible clustering is obtained when using 2 clusters, but even then it is not a particularly strong clustering technique.

To conclude we can conclude that we can obtain that the cities in the US can be separated into 2 clusters based on a combination of their Assault, Murder and Rape arrests but these clusters are not tightly formed so it is hard to draw much further analysis from the data set.