# Analysis Report of the US Arrests Dataset 1973 Using Unsupervised Machine Learning

Report written by William Bigwood

## Introduction

In this report I will explore the differences between the 50 US States in the 'US Arrests' dataset using unsupervised learning methods such as Principal Component Analysis (PCA) and various clustering techniques. This data set contains statistics, in arrests per 100,000 residents, for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas. The aim of this report is to identify clusters that exist in the data and interpret their meaning.

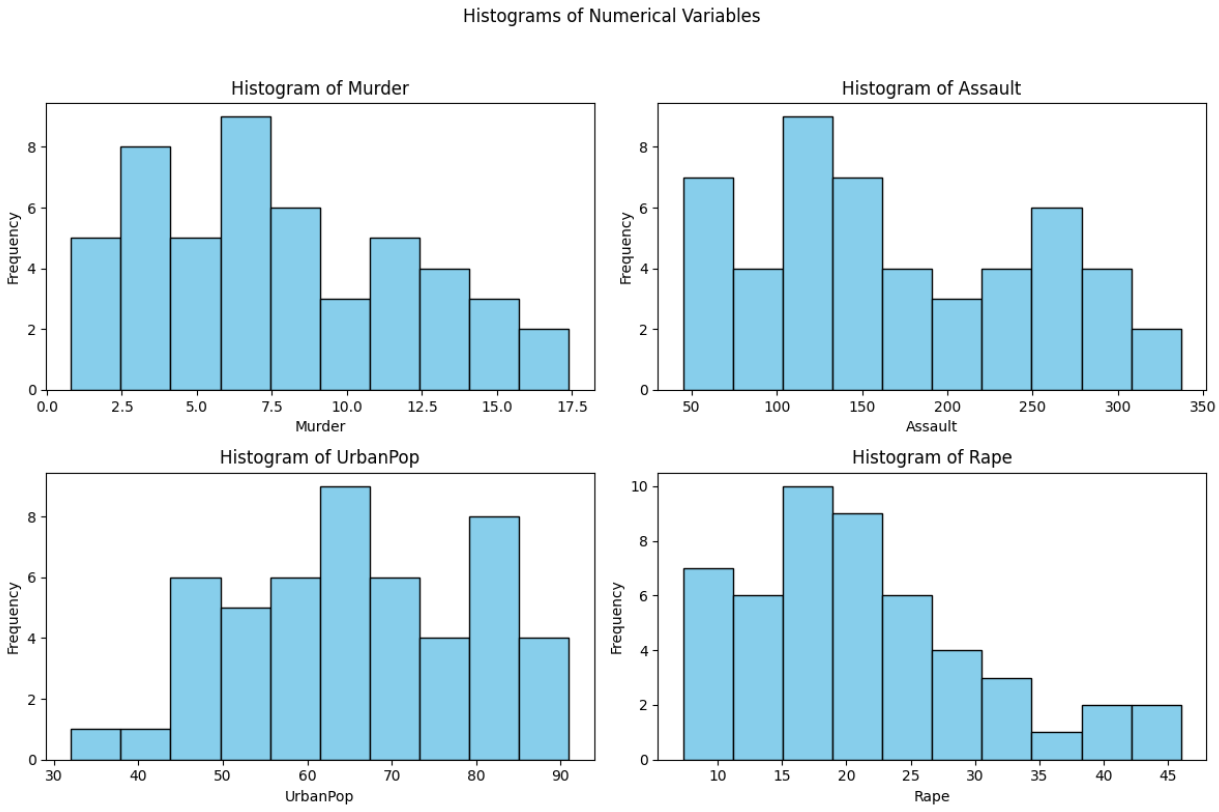
## Data Preprocessing and Exploration

The dataset provided has the name of each US state in the first column alphabetically, but this column is titled 'City'. To improve the understanding of the data in this analysis this column will be renamed as 'State'.

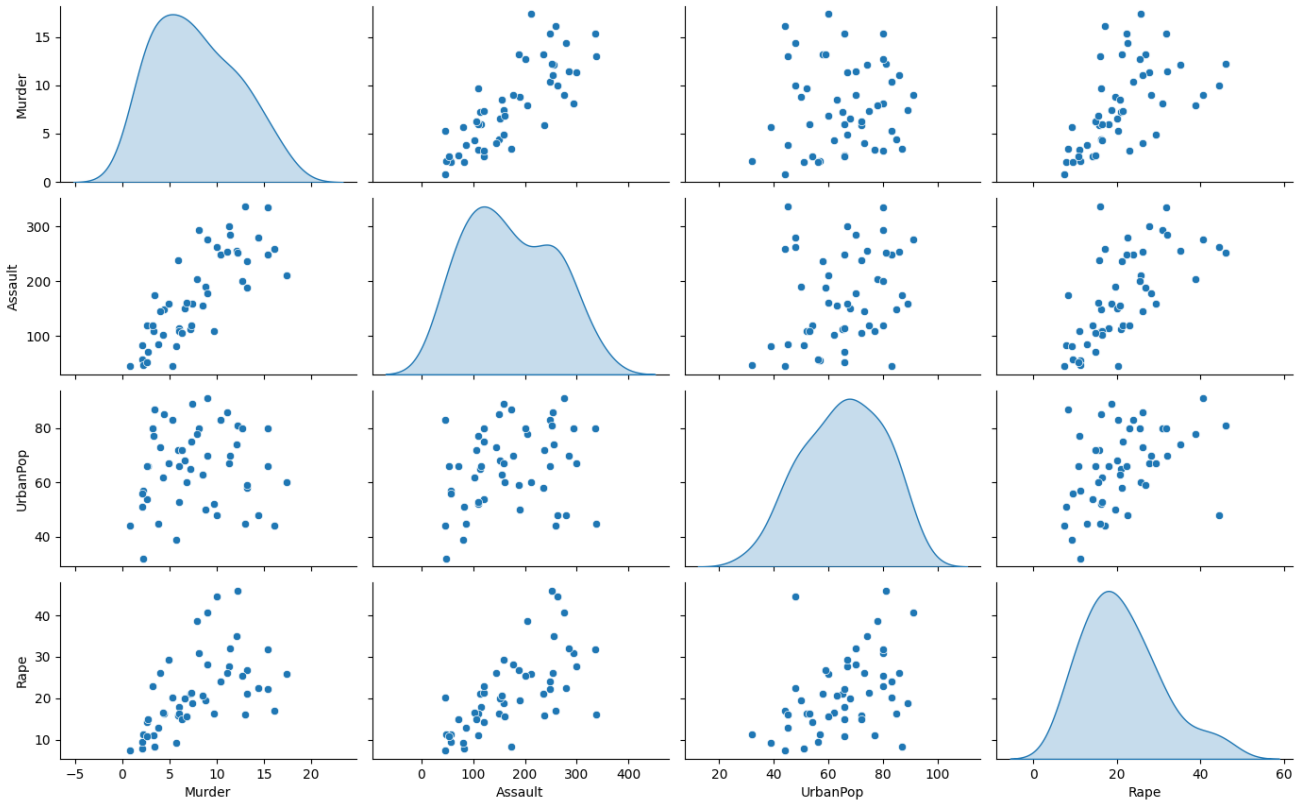
The mean, standard deviation, range and distribution of each numerical variable in the dataset was observed, as well as the number of missing values. This can be seen in the table below. We can see that there are no missing data points at all, which means we can use every row of data for the analysis, without having to remove or impute any data.

| | Murder | Assault | UrbanPopulation | Rape |
|---|---|---|---|---|
| **Mean** | 7.79 | 170.76 | 65.54 | 21.23 |
| **StdDev** | 4.36 | 83.34 | 14.47 | 9.37 |
| **Min** | 0.8 | 45 | 32 | 7.3 |
| **Max** | 17.4 | 337 | 91 | 46 |
| **Missing Values** | 0 | 0 | 0 | 0 |

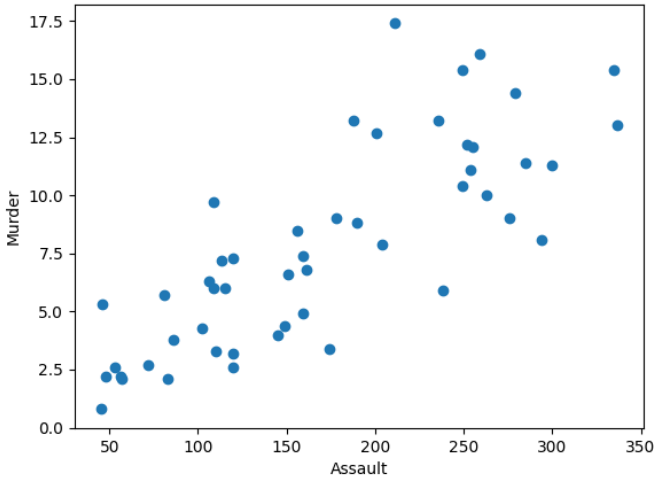
At first glance the 'Assault' variable stands out as having a mean and standard deviation which are significantly higher than the other variables. This indicates that scaling the data may be useful for further analysis. We can also gain insight into the spread of the data through histogram plots for each of the variables.

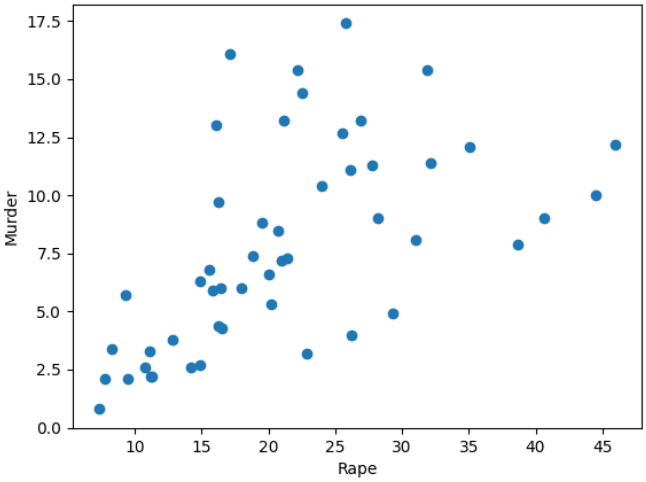


We can see from the above histogram plots of numerical values, that per 100,000 residents, across all of the 50 states, the most common number of murders is around 6-7.5. The most common number of assaults is around 105-130. The most common number of rapes is around 15-19. And the percentage of the population living in urban areas that is most common is around 61-68%. The variables are also not normally distributed. Knowledge of this may be useful when scaling the data later in the analysis.



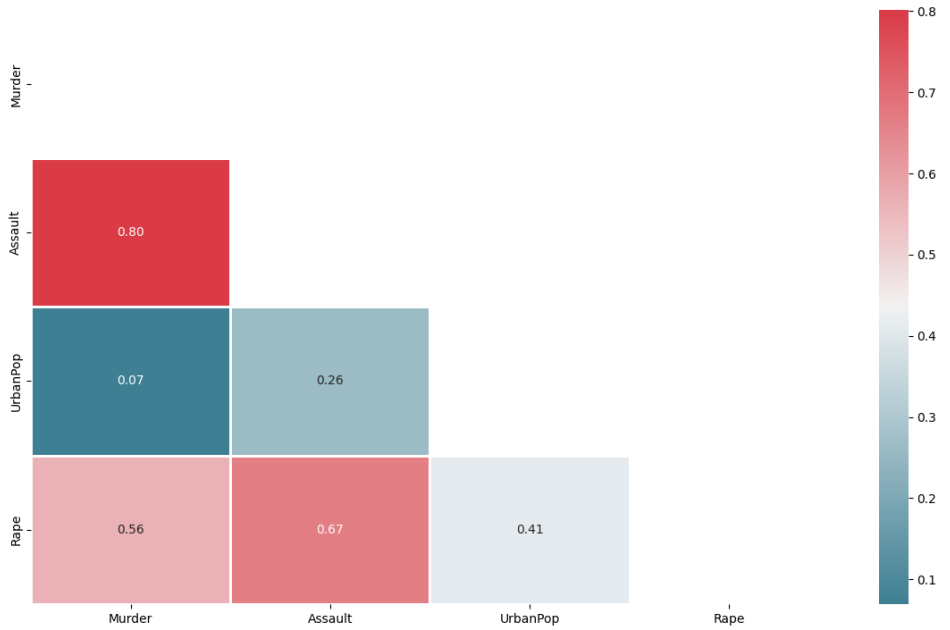

In order to explore how the variables interact with each other, the above pair-plot was created showing the relationships between all the numerical variables. This allows us to visually look for grouping or clusters in the data. At first glance there appears to be some obvious grouping in the Assault vs Murder scatter graph and the Assault vs Rape scatter graph. Looking more closely at the two graphs below we can see two or three possible groups in the graph of Murder vs Assault, and similarly two or three possible groups in the graph of Murder vs Rape. Knowledge of this may be useful for clustering techniques utilised later in the analysis.



## Correlation analysis

Looking at the plot below, we can see that most of the variables in the data set have some correlation, especially Assault, Rape and Murder which all seem to be highly correlated with each other. These correlations are intuitive as all three variables measure different types of violent crimes, and places with high rates of one type of violent crime are more likely to have high rates of another type of violent crime.

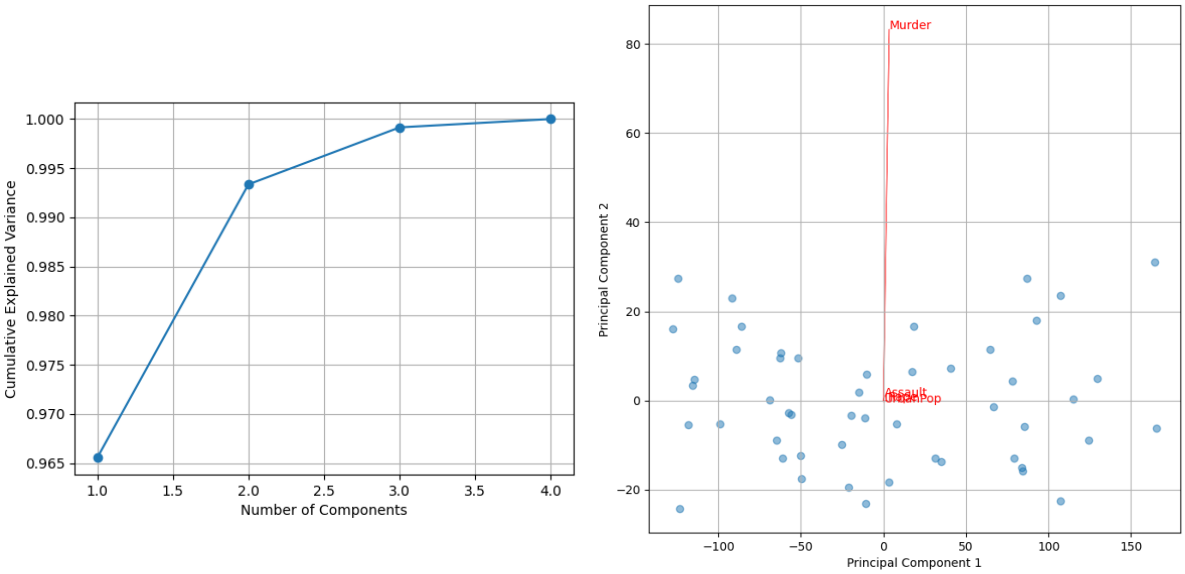



The correlation between assault and murder is the highest with a correlation coefficient of 0.8, followed by the correlation between assault and rape with a correlation coefficient of 0.67. Then closely after follows the correlation between murder and rape with a correlation coefficient of 0.56. These are all relatively strong correlations. The UrbanPop variable (measuring the percentage of the population living in urban areas) is not very strongly correlated to the other variables, however the correlation coefficient between UrbanPop and rape is the strongest at 0.41. Given that most of the variables are highly correlated, this dataset is a good candidate for Principal Components Analysis (PCA), which can allow for dimensionality reduction.

## PCA: Unstandardised Data

Principal Components Analysis is a technique used to identify the fundamental factors, known as principal components, that effectively distinguish observations. It does this by identifying the directions in which data points exhibit the greatest dispersion. Choosing the right number of principal components to retain is an important step and can significantly impact the performance and interpretability of the model. The choice depends on the specific characteristics of the data.

| | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| **StdDev** | 83.73 | 14.21 | 6.49 | 2.48 |
| **Proportion of Variance Explained** | 0.9655 | 0.0278 | 0.0058 | 0.0008 |
| **Cumulative Proportion Explained** | 0.9655 | 0.9934 | 0.9992 | 1 |




After performing PCA on the unstandardised data we can see that almost all of the data (96.55%) is explained by the first principle component. This can be visualised in the graph above (on the left) showing the number of principal components vs the cumulative proportion of data explained.

During the identification of principal components, particular attention is paid to the direction that maximises variance, which can lead to certain variables with significantly higher variances dominating the analysis primarily because of their scale. You can see from the bi-plot above (on the right) that the second principal component is dominated by the 'Murder' variable. This makes it difficult to see how States vary with respect to the other variables, or to read the plot as the other labels are overlapping. Due to the scaling issue there seems to be little variation based on the first principal component.

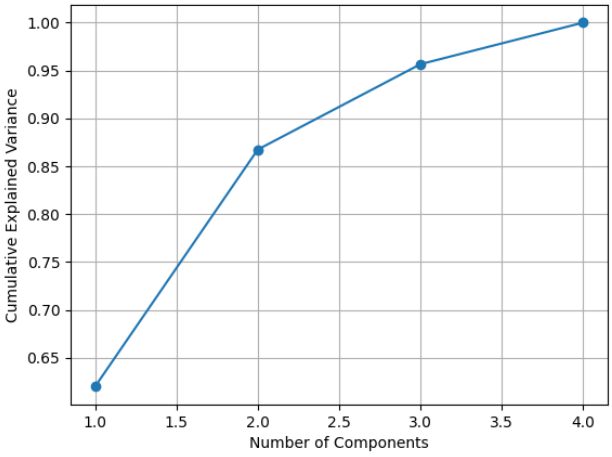
## PCA: Standardised Data

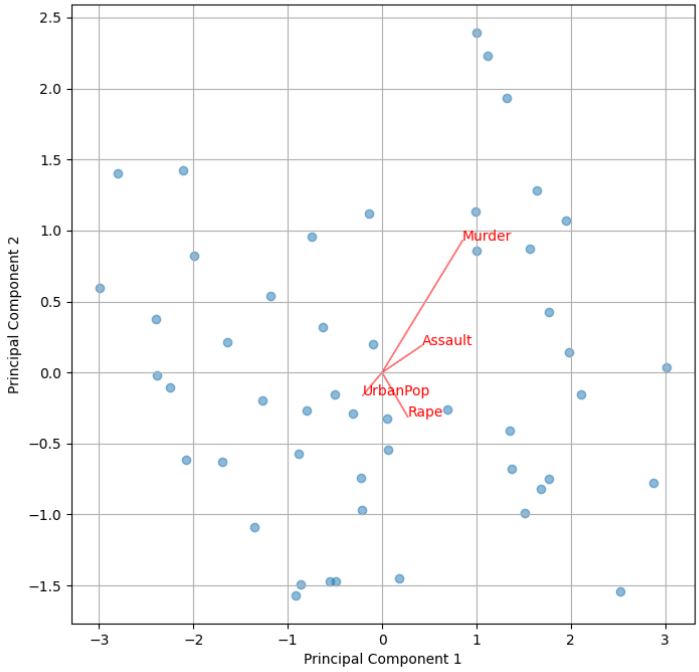
There are a few different options when it comes to scaling the data for PCA. StandardScalar is particularly useful when you want to interpret the loadings of the original variables on the principal components. It ensures that each variable has equal influence on the analysis, regardless of its units or scale. This also aligns well with bi-plot visualisations.

The data was scaled using a StandardScalar and PCA was performed. This time we can see that the variance is explained over more of the primary components, with 62% explained by the first component, 24.7% explained by the second component and 8.9% explained by the third component.

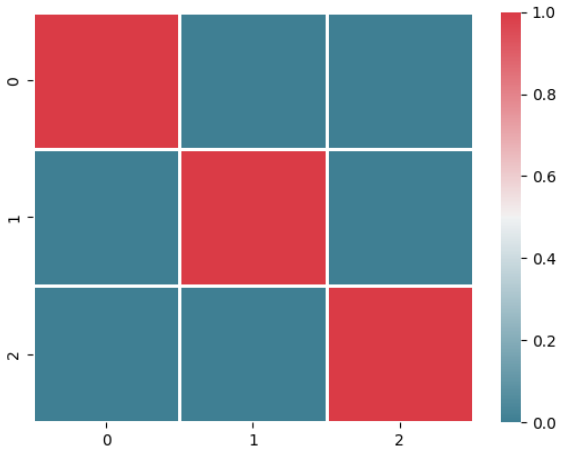
| | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| **StdDev** | 1.59 | 1.00 | 0.60 | 0.42 |
| **Proportion of Variance Explained** | 0.6201 | 0.2474 | 0.0891 | 0.0434 |
| **Cumulative Proportion Explained** | 0.6201 | 0.8675 | 0.9566 | 1 |

The first three principal components together explain about 95.66% of the variance. This can be seen in the Scree Plot below. The fourth component only explains 0.08% of the variance, and so we can therefore remove it and use only the first three to perform cluster analysis. We began with a dataset of 4 numerical variables and we have reduced the dimensionality, so we now have 3 variables that explain most of the variability.




We can see in the above bi-plot that scaling the data has increased the readability of the other 3 variables and the data points are more spread out. We now perform PCA one final time on the scaled data, this time specifying that we want 3 principal components in total. The correlation heat map below shows that there is little to no correlation between each of the three principal components that are produced, meaning that PCA has effectively summarised all related information in a smaller number of variables. This is known as dimensionality reduction.

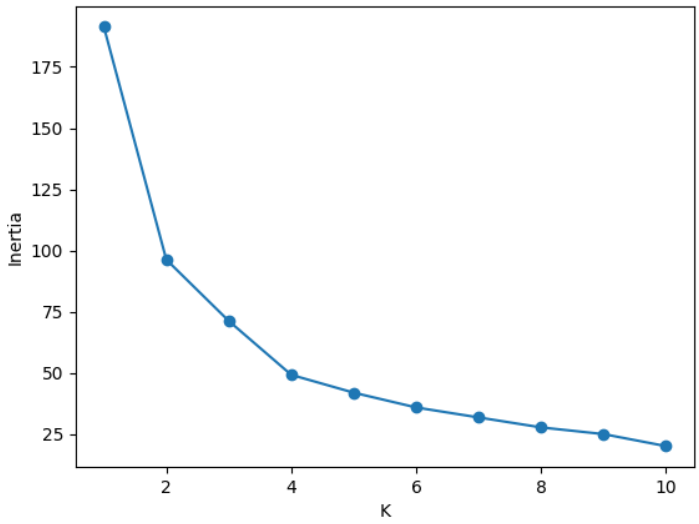



## K - Means Clustering

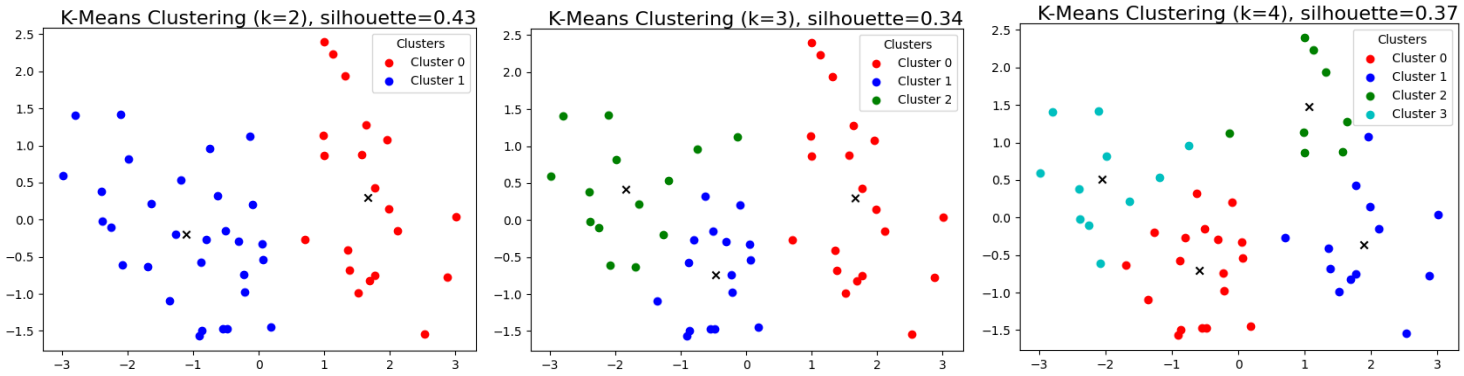
K- means clustering is a well known algorithm for grouping data points into K distinct clusters. First we specify the desired number of clusters, denoted as K, and then assign each data point to one of these clusters. K-means clustering aims to achieve clusters where the within-cluster variation is minimised. This variation gauges how much individual data points within a cluster deviate from one another, signifying that the data points within a cluster are similar, while the variation between different clusters should be maximised, indicating distinct groupings.

Feature selection is a crucial step in preparing data for clustering, as the choice of features can significantly impact the quality and interpretability of the clustering results. Earlier we saw that the scatter graph of Murder vs. Assault had some visually obvious clustering. If we had not conducted PCA then these would potentially be good features to base the model on. However, we will base the K-means clustering model on the primary components from our PCA. Principal components are orthogonal to each other, which means they are uncorrelated (as seen above). They are constructed to maximise the variance, and hence they are good features to select for clustering models.

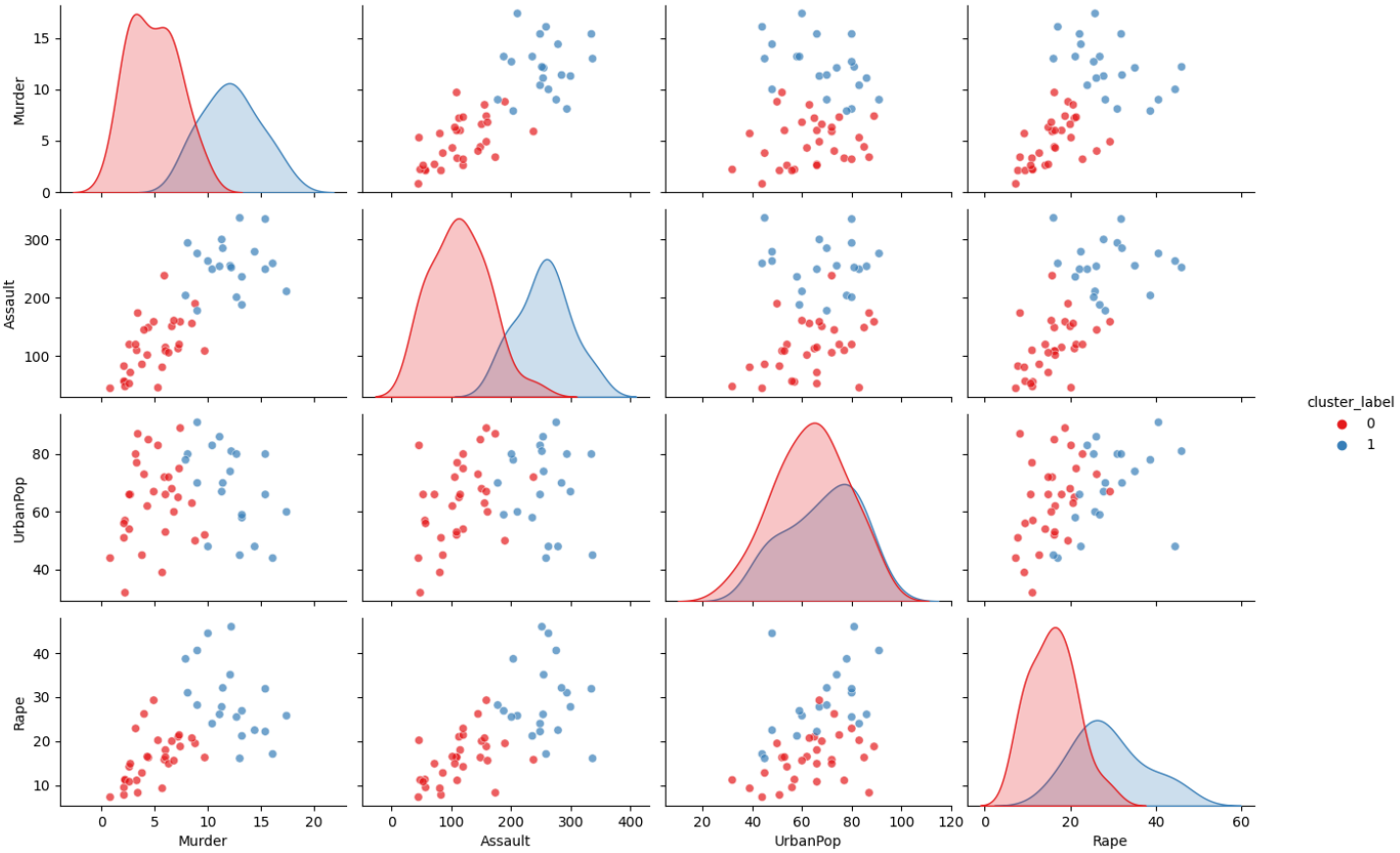
The elbow method helps you chose the appropriate value of K by looking for an 'elbow point' in the plot of within-cluster variance as a function of K. The 'elbow' represents a point of diminishing return, where increasing the number of clusters does not significantly improve the clustering quality. In other words, the elbow point if the value of K at which the rate of decrease in inertia sharply changes. This is typically the optimal number of clusters for the model. We can see from the plot below that the elbow point from this dataset appears to be somewhere between k=2 and k=4.




The following three plots show different outcomes of the algorithm depending on the value chosen for K. You can see that where k=2 there are 2 clusters, where k=3 there are 3 clusters, and where k=4 there are 4 clusters. The silhouette score is a metric used to measure the quality of clusters. It measures the degree of separation between different clusters and the degree of similarity within the clusters. The silhouette score ranges from -1 to 1, where a high silhouette score (close to 1) indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters. This is a sign of a good clustering. We can see that the silhouette score is the highest (0.43) in the plot where there are two clusters (k=2).



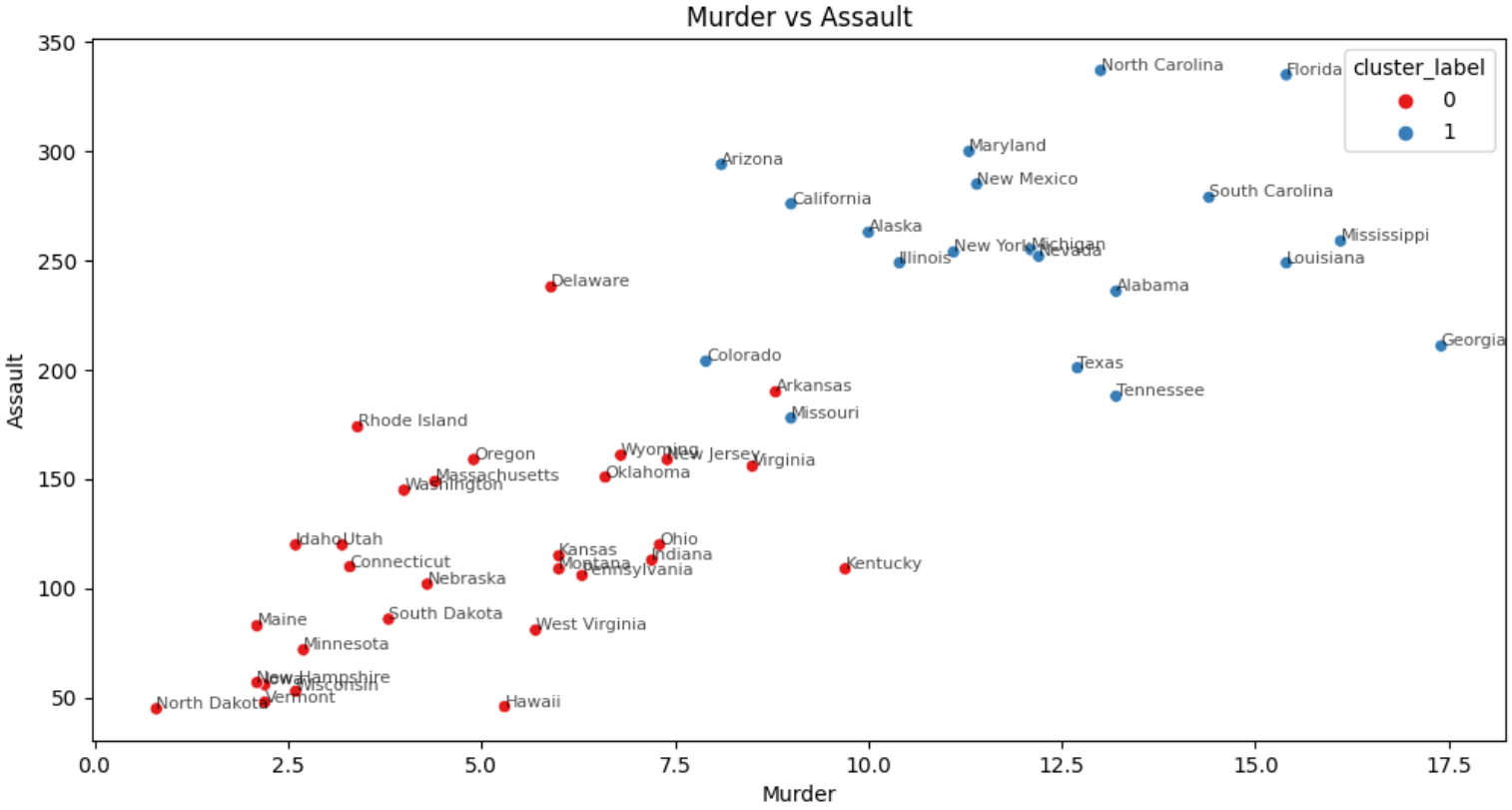

Having established the optimal value for K, we created a K-means clustering model with k=2 and fit the model to the PCA data to find the cluster centres and assign each datapoint to a cluster. With two clusters we find that 30 of the data points are assigned to the first cluster (cluster 0) and 20 are assigned to the second cluster (cluster 1). This is a relatively even distribution. The silhouette score for the final model is measured at 0.43. Now we can use the K-means model to visualise the clusters with respect to different variables, as seen in the following bi-plot.

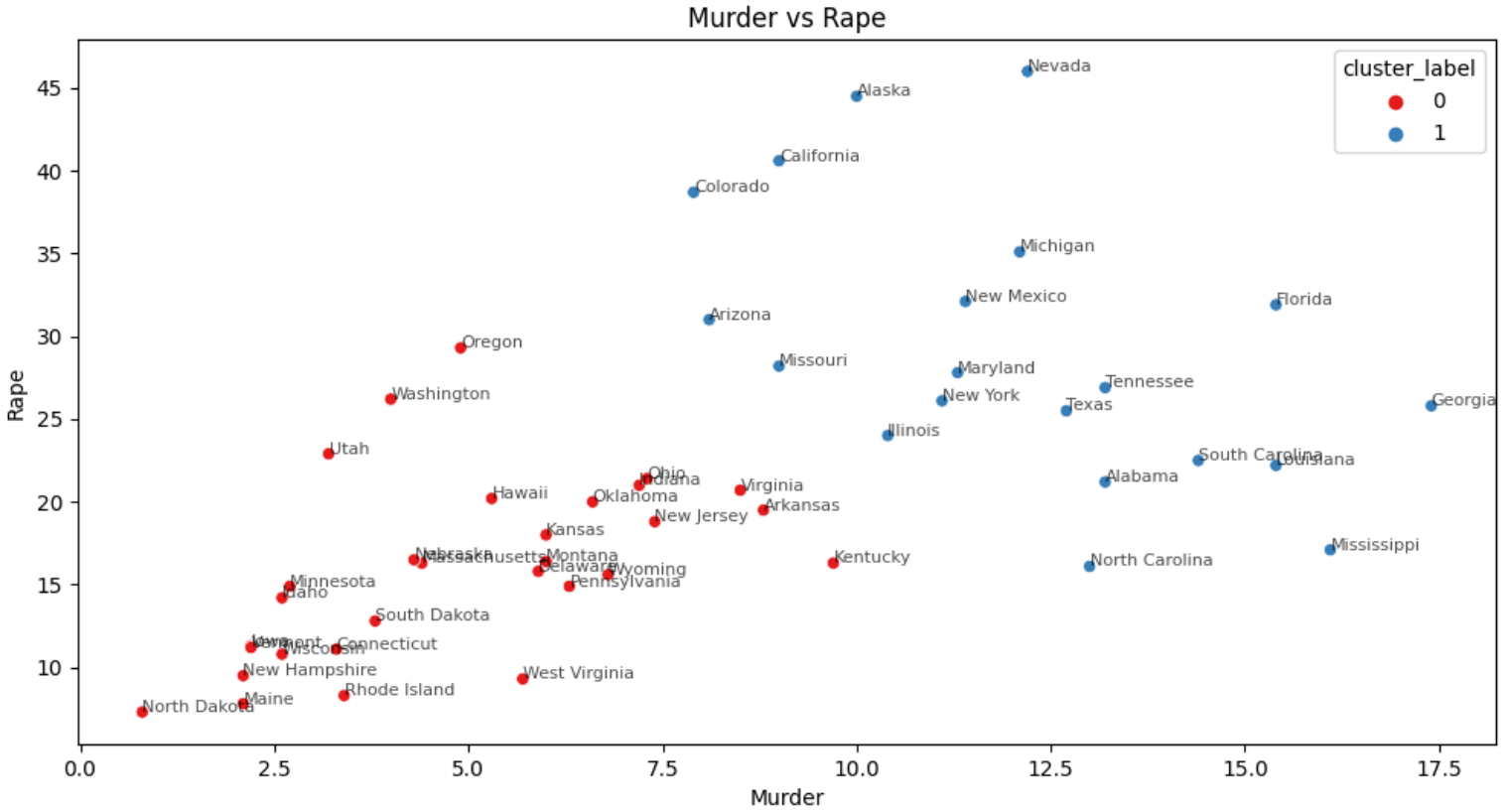



In the bi-plot above we can see that the graphs showing the relationship between the rape, assault and murder variables typically show the red cluster (cluster 0) with low crime rates and the blue cluster (cluster 1) with high crime rates.

As seen before, the percentage of the population living in urban areas (UrbanPop) has little correlation with the crime variables. It could be said that US states in cluster 1 tend to have a slightly higher UrbanPop than states in cluster 2, especially on the UrbanPop vs Rape scatter graph, but the difference is not very significant. This seems counterintuitive as one might expect a higher percentage of UrbanPop to increase the rate of violent crime, however the data seems to suggest that violent crime in more rural US states is just as common as violent crime in more urban states.

Looking more closely at the Murder vs Assault and Murder vs Rape scatter graphs below, we have labeled each state to get a better insight into the difference between each individual state in the dataset. We can see that the K-means clustering model has clustered the data in the way we assumed it would during the data exploration phase.


Murder vs Assault
cluster_label
0
1
Assault
Murder
North Carolina
Florida
Maryland
Arizona
New Mexico
California
South Carolina
Alaska
Mississippi
New York
Michigan
Nevada
Illinois
Louisiana
Delaware
Alabama
Georgia
Colorado
Texas
Arkansas
Tennessee
Missouri
Rhode Island
Oregon
Wyoming
New Jersey
Virginia
Massachusetts
Oklahoma
Washington
Idaho
Utah
Ohio
Kansas
Connecticut
Indiana
Montana
Kentucky
Pennsylvania
Nebraska
Maine
South Dakota
West Virginia
Minnesota
New Hampshire
Iowa
Wisconsin
Vermont
North Dakota
Hawaii


Murder vs Rape
cluster_label
0
1
Rape
Murder
Nevada
Alaska
California
Colorado
Michigan
New Mexico
Florida
Arizona
Oregon
Missouri
Maryland
Tennessee
Washington
New York
Texas
Georgia
Illinois
Utah
South Carolina
Louisiana
Ohio
Indiana
Alabama
Hawaii
Oklahoma
Virginia
Arkansas
New Jersey
Kansas
Nebraska
Massachusetts
Montana
Kentucky
North Carolina
Mississippi
Delaware
Wyoming
Minnesota
Pennsylvania
Idaho
South Dakota
Iowa
Vermont
Connecticut
Wisconsin
New Hampshire
West Virginia
Rhode Island
Maine
North Dakota

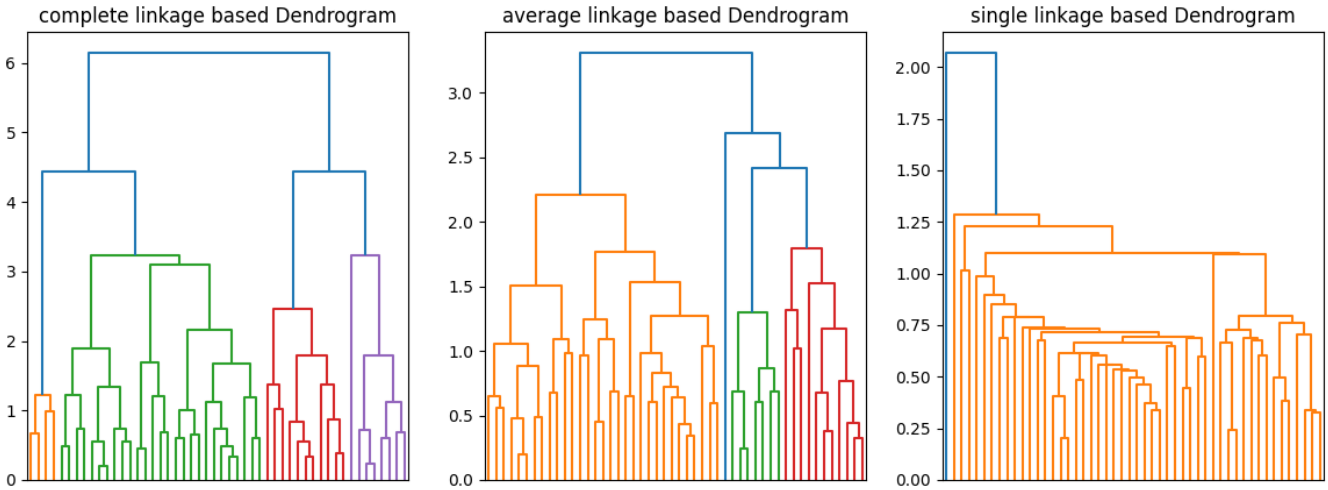
## Hierarchical Clustering

We will now use a different clustering technique on the principal components generated in the PCA section of this report. Using multiple different clustering techniques for an analysis and comparing their results can offer several advantages. By using multiple clustering techniques you can assess the robustness of the clusters. If multiple methods produce similar results, you can be more confident in the clusters found. Additionally different clustering algorithms may be sensitive to different aspects of the data, so using different techniques allows you to assess the clusters from different angles and address the limitations of each individual algorithm. We can also compare the silhouette score of each model to assess which model has better clustering.

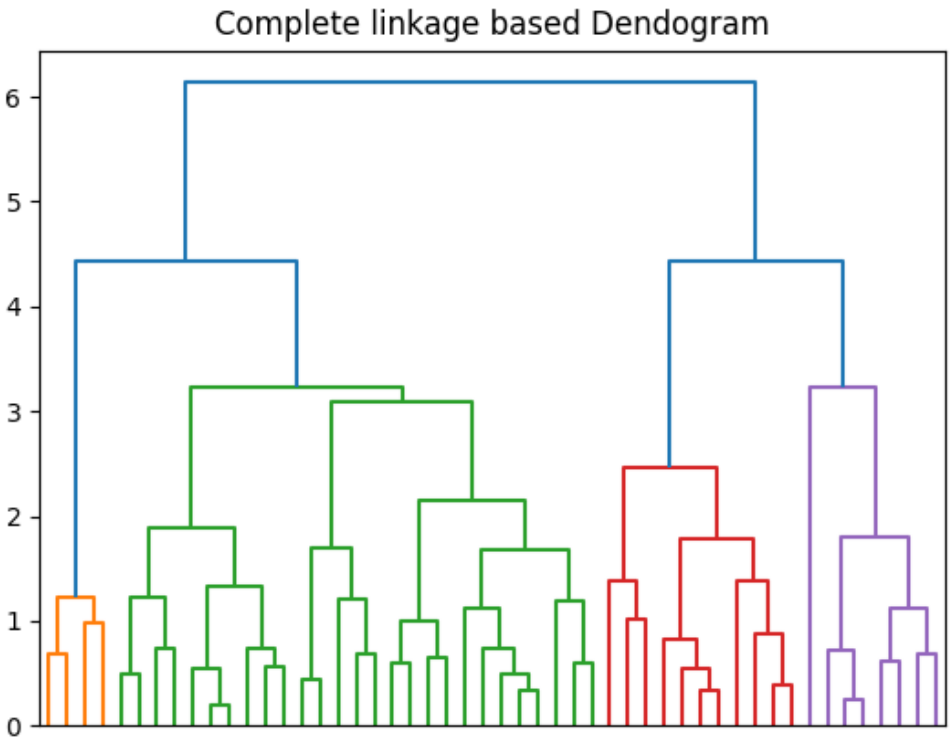
Hierarchical clustering is an algorithm that builds a hierarchy of clusters which form a tree-like structure called a dendrogram. Initially each data point is assigned to a mini-cluster of its own and a measure of dissimilarity is defined. The algorithm then iteratively merges the most similar two clusters into one larger cluster, working upwards, until all data points belong to one single cluster. The dendrogram can then be 'cut' at a particular height to obtain clusters.

Hierarchical clustering is advantageous because we can see the clusters visually and don't have to specify the number of clusters before running the algorithm. But we do have to decide the number of clusters after the algorithm runs.

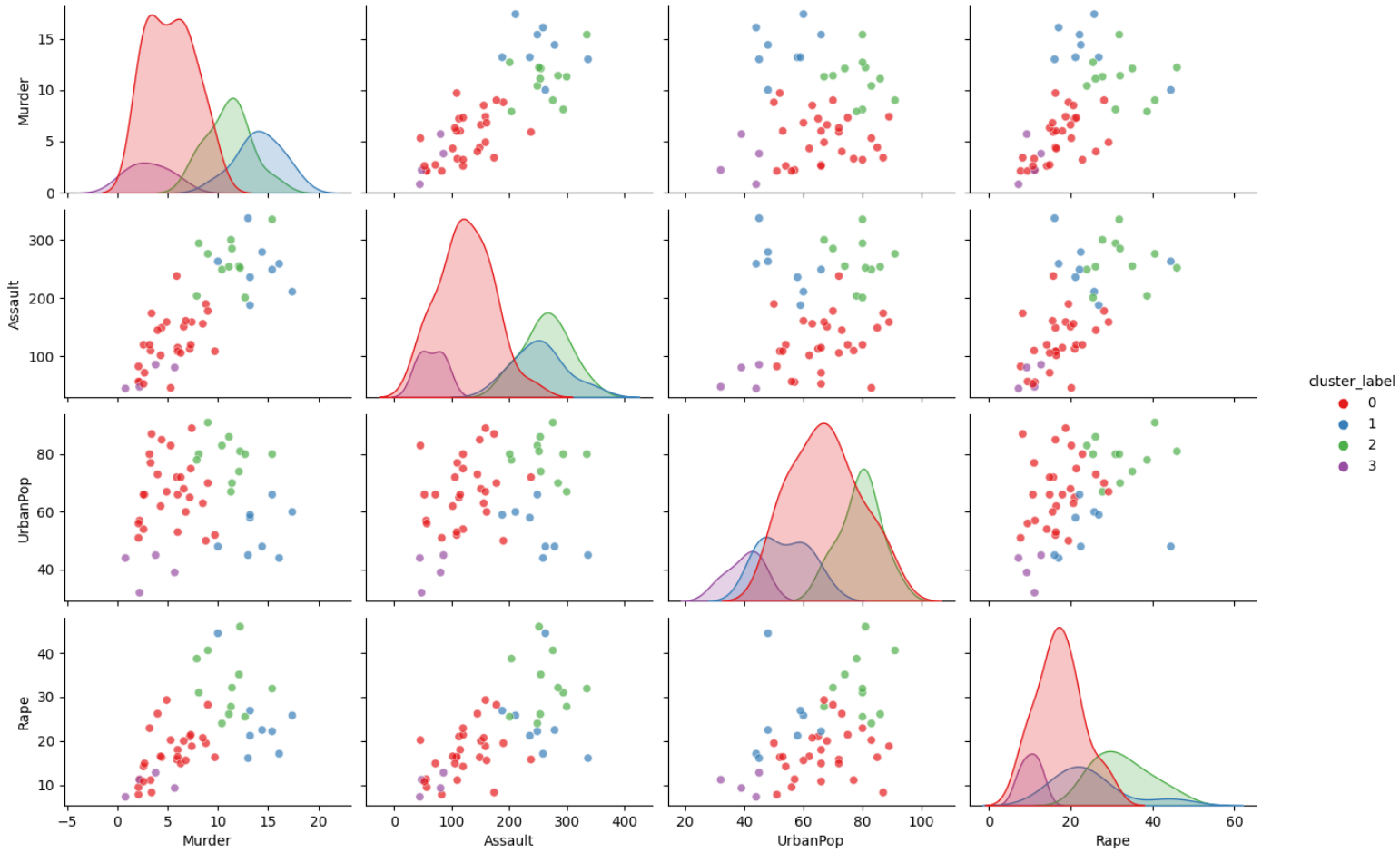
In this analysis, Euclidian distance was used for the distance metric between observations, which is the most common way to measure distance. In order to determine the distance between clusters, the following dendrograms were plotted for the 'single', 'complete', and 'average' linkage methods.




From the dendrograms above, the 'complete' linkage method creates the most balanced dispersion of clusters and will therefore be the method of choice for the rest of this analysis. A clearer dendrogram of the complete linkage method is shown below.



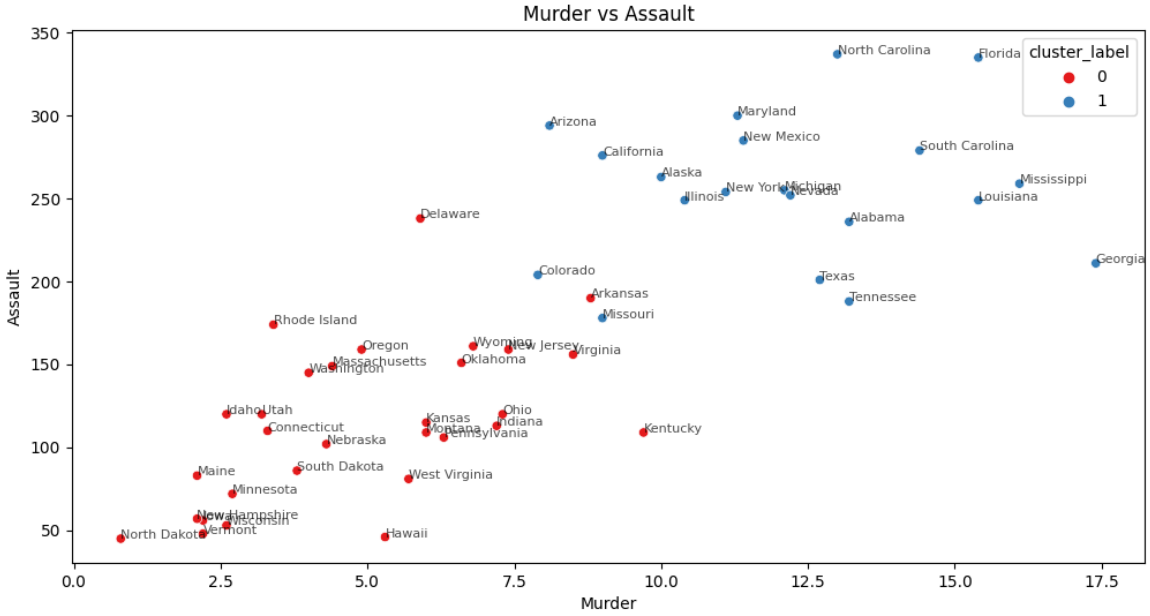

From the complete linkage dendrogram above we can see that there are 4 defined clusters. So we created a hierarchical clustering model that has 4 clusters, a Euclidian distance metric and a complete linkage method, and then fit it to the PCA data and predicted the cluster labels for each data point. We can again look at a pair plot to visualise the four clusters on different variable pairs.

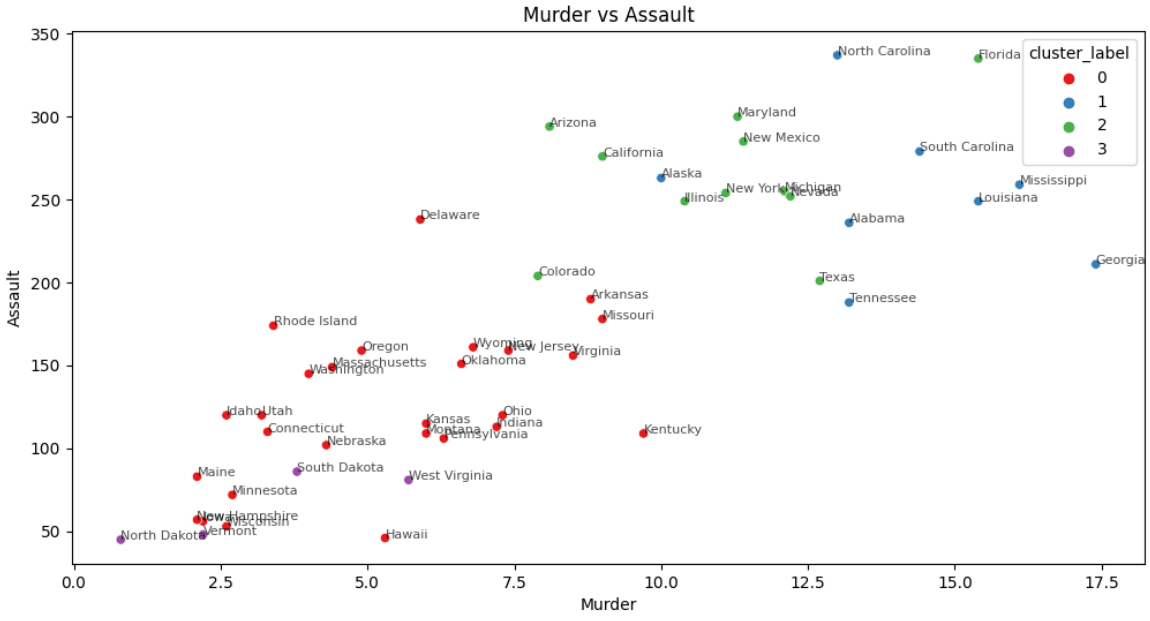


The table below shows the distribution of data points across the four clusters. Most US states fall in clusters 0 and 2, while clusters 1 and 3 are minority 'extreme cases' at each end of the crime spectrum. The silhouette score for this model is 0.31.

| | Number of US States within the cluster |
|---|---|
| **Cluster 0** | 27 |
| **Cluster 1** | 8 |
| **Cluster 2** | 11 |
| **Cluster 3** | 4 |

The two scatter graphs below show the same data as each other, but with different clusters from the different clustering models. The top graph is the K - Means clusters as seen earlier, and the bottom graph is the Hierarchical clusters. It could be said that the clusters 0 and 3 in the bottom graph are largely just two subsets of cluster 0 in the top graph, and that clusters 1 and 2 in the bottom graph are largely just two subsets of cluster 1 in the top graph.



It appears that the cluster with the lowest rate of violent crime is the purple cluster (cluster 3). The US states in this cluster have extremely low rates of violent crime. They also tend to have the smallest UrbanPop percentage than the other clusters. This clustering technique has allowed us to gain more insight into the UrbanPop variable than the K - means clustering technique did. It appears that the cluster with the lowest rates of violent crime (cluster 3) also has the lowest percentage of the population in urban areas, and the cluster with the highest rate of violent crime (cluster 2) tends to have a high percentage of the population in urban areas. The red cluster (cluster 0) has medium to low rates of violent crime, and the blue cluster (cluster 1) has medium to high levels of violent crime.

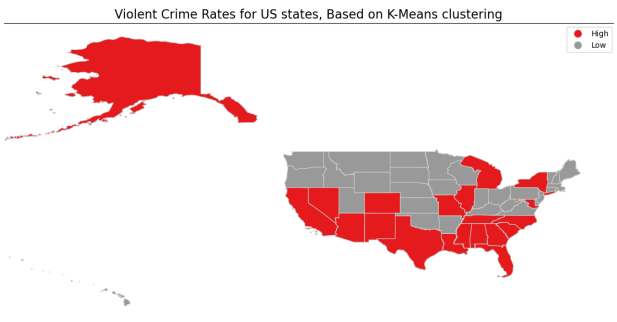
We could rename the clusters for better interpretability, for example renaming clusters 3, 0, 2, and 1 'Low Violent Crime Rate', 'Mid/Low Violent Crime Rate', 'Mid/High Violent Crime Rate', and 'High Violent Crime Rate' respectively, and then visualise the clusters on a map of the US showing which cluster each state belongs to. This would be a useful visualisation for real life uses.

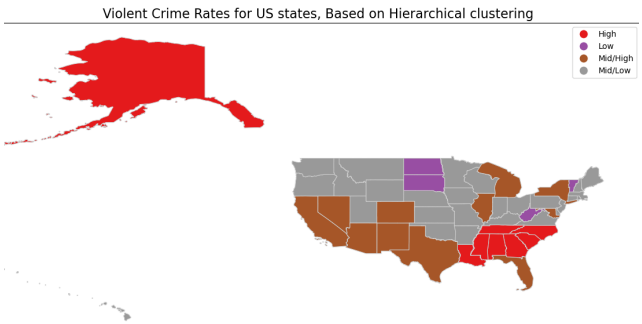
## Conclusion:

In conclusion, we have explored the data, and successfully utilised Principal Components Analysis to implement dimensionality reduction in the dataset. We have performed two clustering techniques, K - Means and Hierarchical clustering to establish clusters in the dataset and interpret their meaning.

Comparing the two clustering models, we see that the K - Means model has a higher silhouette score (0.43) than the Hierarchical model (0.31) which means the US states are better matched to their own cluster in the K - means model and are less closely matched to neighbouring clusters. In other words, the K - means model is a better model than the Hierarchical model. However the Hierarchical model provided insight into the UrbanPop variable that was undetectable in the K - Means model. In this case it has proven useful that multiple models were run as we can understand different things about the dataset from each one, where the K - Means model is more accurate but the Hierarchical model provided more of an in depth insight into the spread of violent crime across the different US states.

We can better visualise these clusters on the two maps below, showing which US states have high rates of violent crime, and which states have low rates of violent crime. From these maps we can gain insight into other factors that aren't in the dataset, like how regional data might impact violent crime rates. It appears that in general, states in the north have higher rates of violent crime than states in the south.




This information has many real life uses such as law enforcement agencies being able to appropriately allocate resources. Potential home-buyers and renters can use this information to make informed decisions about where to live, and similarly business owners looking to establish new ventures may consider crime rates when choosing locations for their operations. Finally, tourists can use this information to make safer decisions about holiday destinations and places to avoid when visiting the US. This data captures a snapshot of the crime rates in 1973. Similar analysis could be done on data from multiple years, and results for different timeframes compared, to get a feel for the violent crime rates over time in different states in the US.