US arrests (1973) Data set analysis

This report contains an analysis of a dataset detailing various crime figures gathered for each US state from the year 1973. It includes the numbers per 100000 for assaults, murders and rapes.

The report here is generated from analysis performed on an IPython Notebook (IPYNB) written in Google COLAB. It includes PCA analysis and a couple of clustering techniques (Hierarchical and K means clustering)

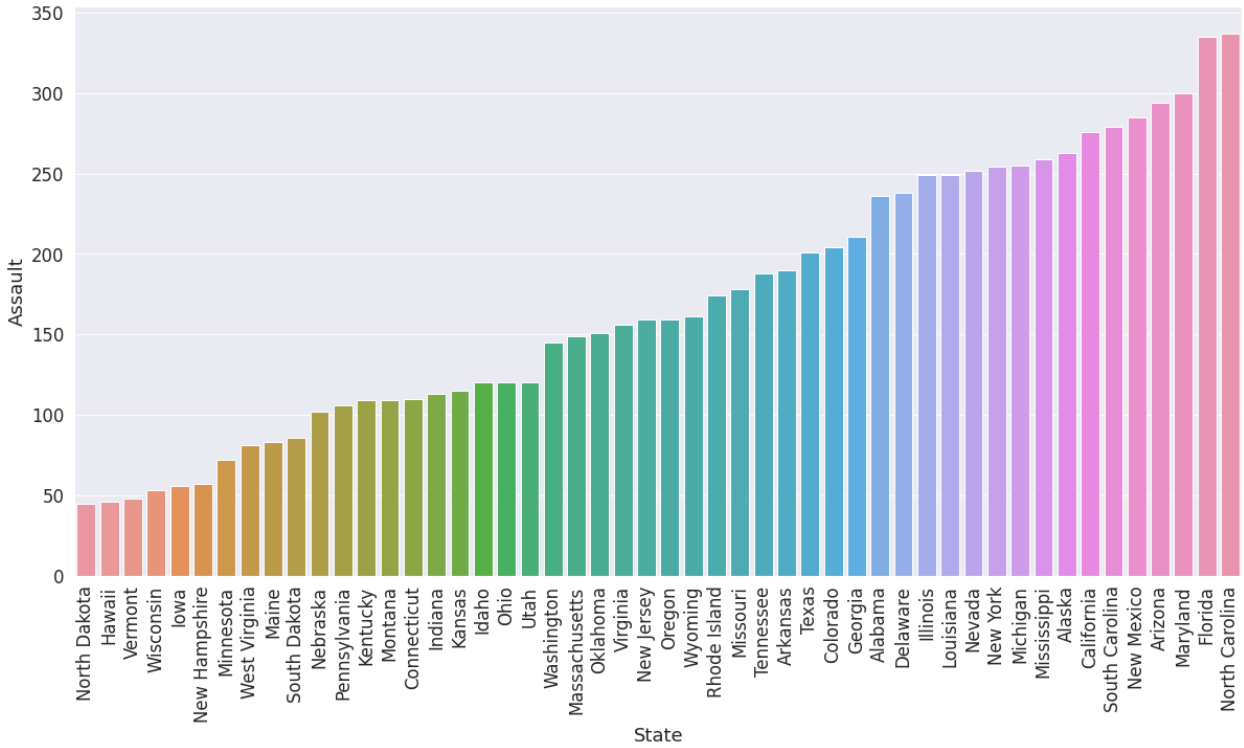
There are 50 rows of data (one for each state) with four columns:

1973 US crime figures by state

- Murders, Assaults, Rapes per 100 000
- UrbanPop: Percentage population living in an urban area (towns, cities and suburbs)

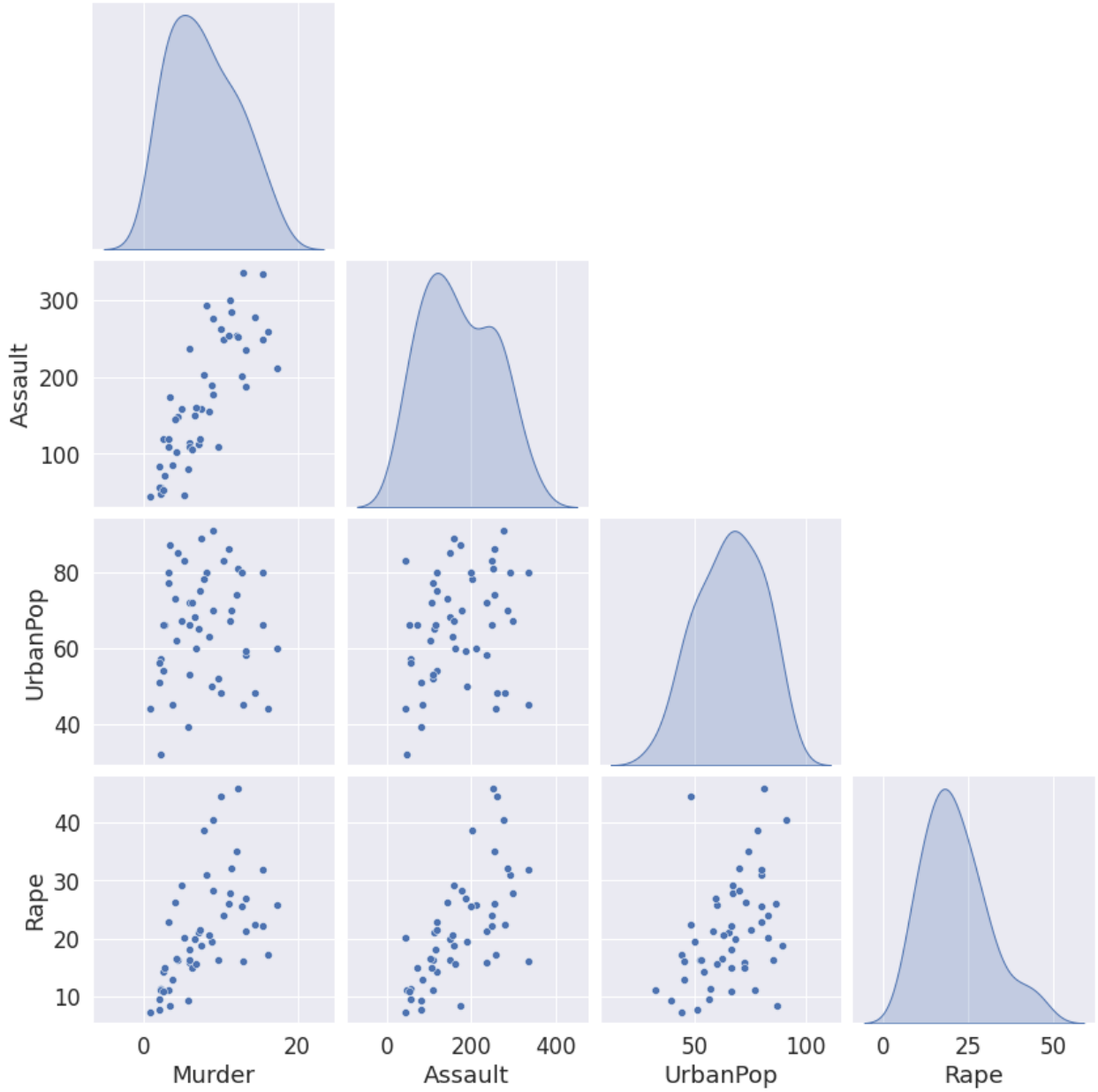
	Murder	Assault	UrbanPop	Rape
mean	7.79	170.76	65.54	21.23
std	4.36	83.34	14.47	9.37
min	0.80	45.00	32.00	7.30
max	17.40	337.00	91.00	46.00

The table shown to the left shows a summary of the data. Looking at the standard deviation, SD, min and max for all the variables we can see that there are significant differences between individual states. This can be better appreciated when we plot the number of assaults in ascending order for each state.



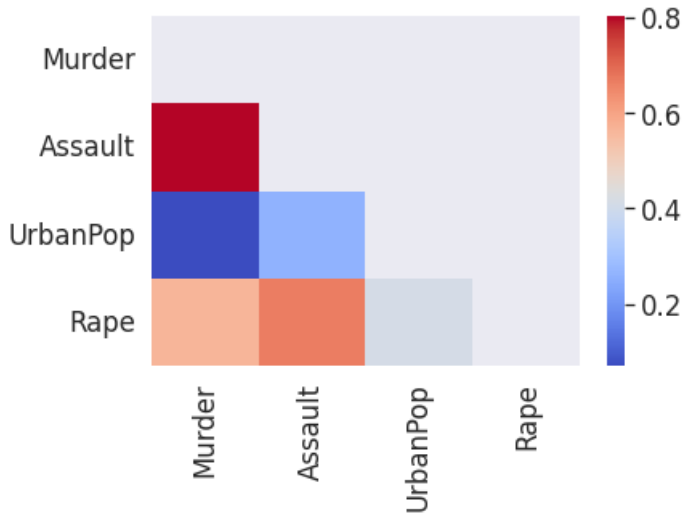
US arrests (1973) Data set analysis

The powerful Seaborn ‘Pairplot’ immediately gives us a feel for the data. We can see positive relationships between the number of assaults and murders; and to a lesser extent we see a positive relationship between assaults and rapes. The role of urban population is less clear cut from the scatter diagrams so we can look at the variable correlations to learn more. It’s worth highlighting the histograms at this point: the data appears to be normally distributed suggesting we can use standardistion to scale the data and ensure there is no bias when examining the variables.



US arrests (1973) Data set analysis

Correlation Analysis

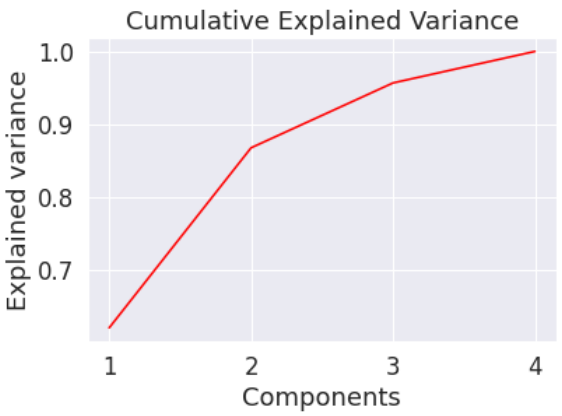


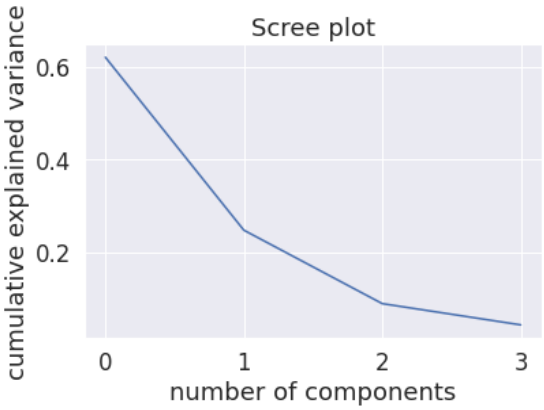
Seaborn allows us to quickly look at the correlations between the variables. As discussed above we can see a strong association between murders and assaults and to a lesser extent the numbers of assaults and rapes.

Principal Component Analysis (PCA)

The correlations above indicate that we may be able to perform PCA analysis on the data. In other words we may be able to reduce the number of variables by describing using one or two key components.

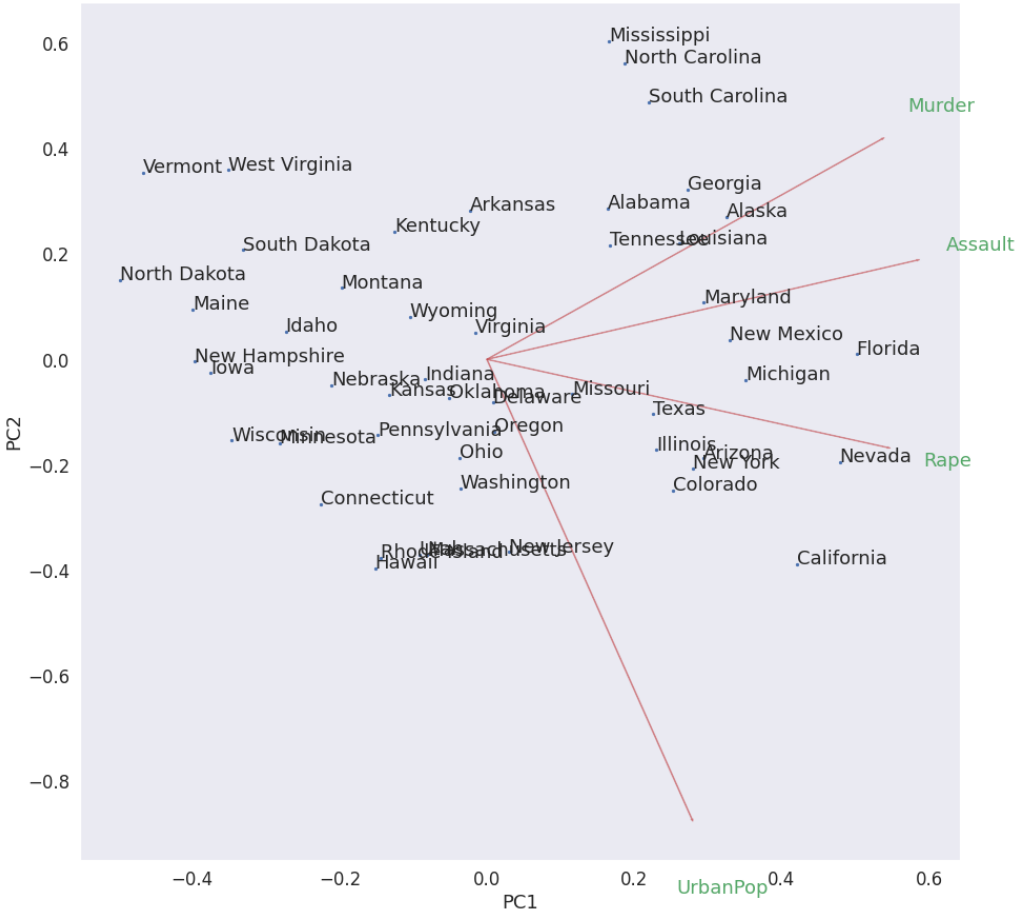
Ideally we want to simplify our analysis by minimising the number of principal components that we need to work with. The scree and cumulative variance plots show us that over 80% of the variance is explained by the first two principal components. From this point on we are able to focus on PCA1 and PCA2.





US arrests (1973) Data set analysis

Biplot of the first two principal components



The plot shows that PCA1 is more dominated by the crime variables. The angles between the lines for murder, assault and rape are relatively small and we can deduce that they are closely related. Percentage urban population has a greater influence over PCA 2.

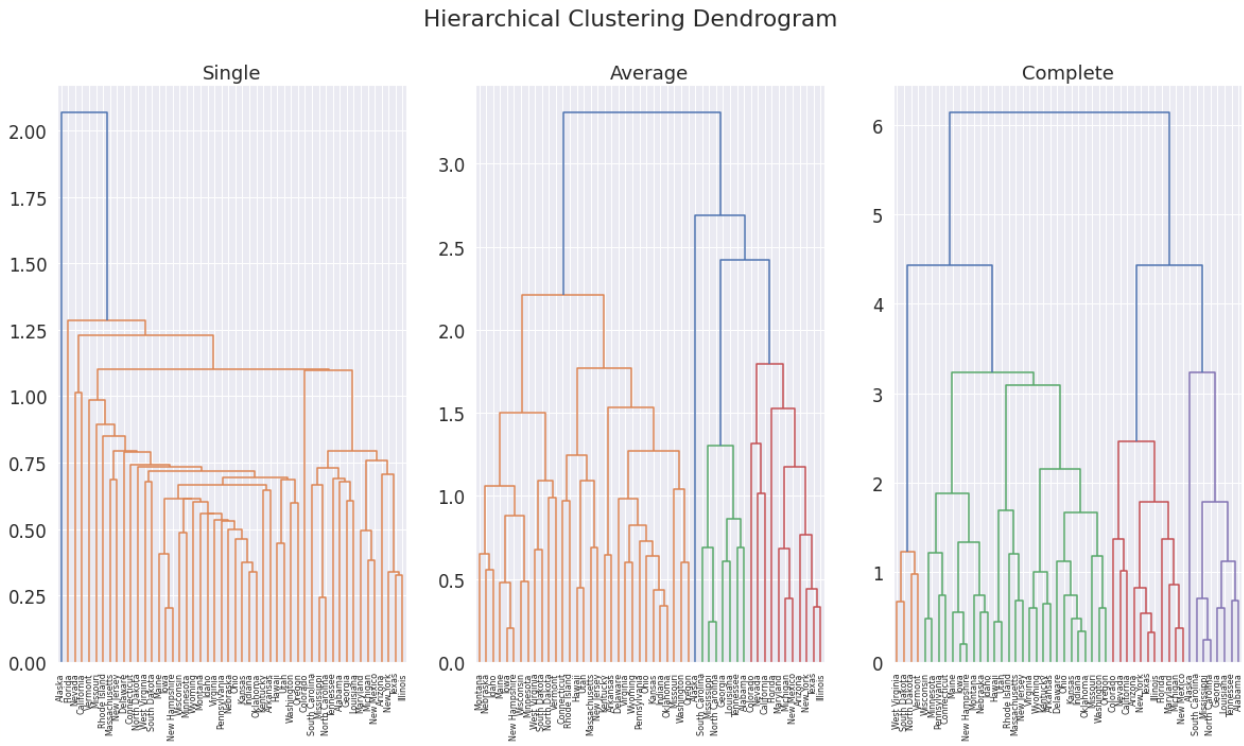
States such as Florida, Nevada and New Mexico seem to lie in a similar cluster with high crime rates. On the opposite end of the scale we can see a group with much lower crime rates including North Dakota, Maine and New Hampshire.

Clustering

We can now look at two powerful clustering techniques to help us create groups of states that are behaving in a similar way in terms of crime and urban populations.

US arrests (1973) Data set analysis

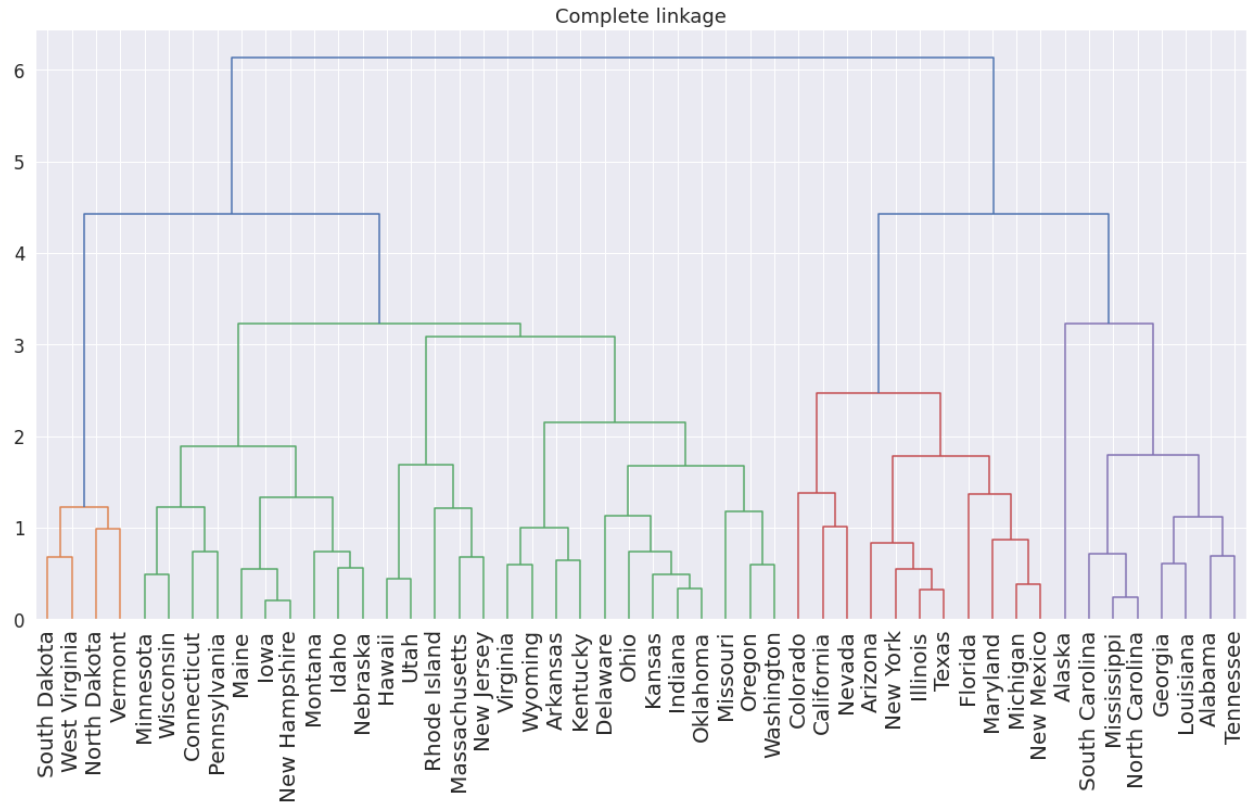
Hierarchical clustering



Single linkage: computes the minimum distance between clusters before merging them.
Complete linkage: computes the maximum distance between clusters before merging them.
Average linkage: computes the average distance between clusters before merging them.

The advantage of Hierarchical Clustering is not necessary to define the number of clusters. The dendograms make it easier to understand and compare the methods for calculating the clusters. Hierarchical clustering does not work very well on vast amounts of data or huge datasets - not an issue in this case.
Here the use of complete linkage seems to give the most balanced group of cluster so let's look at this in more detail.

US arrests (1973) Data set analysis

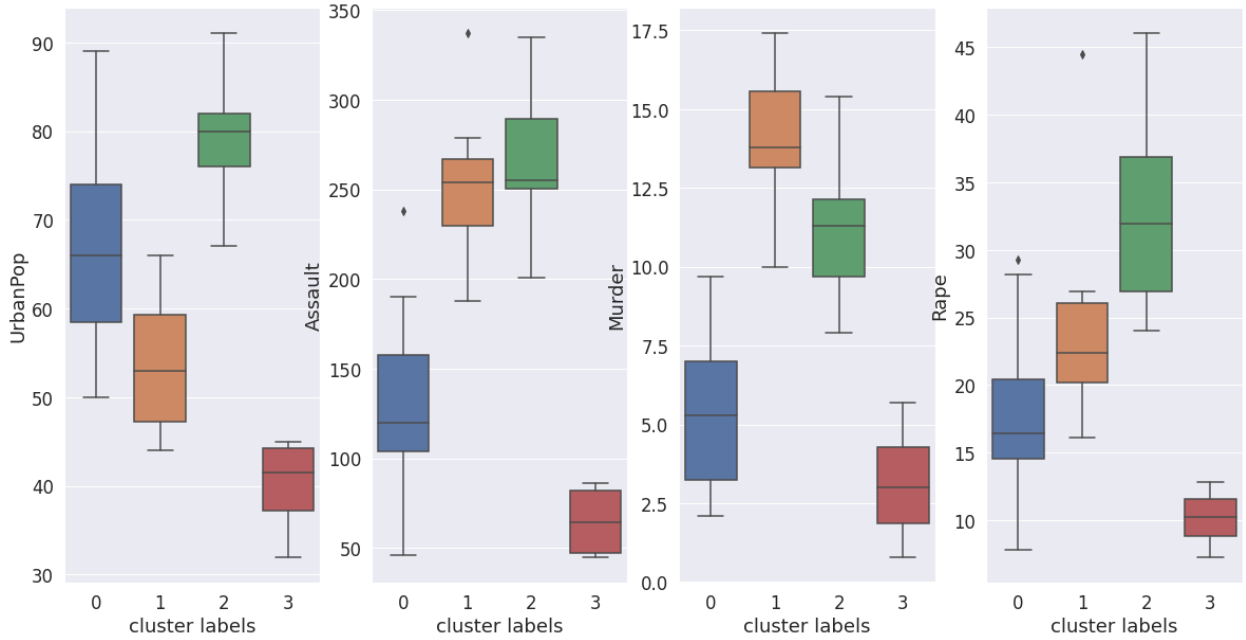


The dendogram above shows us that the hierarchical clustering technique has split the states into 4 distinct groups. We can now examine these groups more closely using the classifications that the states have been divided up into.

The box plots below detail the four clusters. The groups do look distinct with few outliers suggesting that this method of classification has worked reasonably well. An interesting observation is that percentage urban population is not necessarily a predictor of crime rate. For example, Group 1 has a relatively low percentage urban population but relatively high levels of crime. Group 0 has a relatively high percentage urban population but relatively low levels of crime. Clustering methods such as hierarchical clustering allow us to see such anomalies more clearly. In this case they trigger the need to find different explanations to account for the data. For example -percentage urban population could be very misleading if comparing states with low and high populations.

US arrests (1973) Data set analysis

Box plots of the variables based on hierarchical clusters



GROUP 0 states have relatively high percentage urban populations but relatively low crime levels:

['Arkansas' 'Connecticut' 'Delaware' 'Hawaii' 'Idaho' 'Indiana' 'Iowa'
'Kansas' 'Kentucky' 'Maine' 'Massachusetts' 'Minnesota' 'Missouri'
'Montana' 'Nebraska' 'New Hampshire' 'New Jersey' 'Ohio' 'Oklahoma'
'Oregon' 'Pennsylvania' 'Rhode Island' 'Utah' 'Virginia' 'Washington'
'Wisconsin' 'Wyoming']

GROUP 1 states have the highest percentage urban populations and high crime levels:
['Alabama' 'Alaska' 'Georgia' 'Louisiana' 'Mississippi' 'North Carolina'
'South Carolina' 'Tennessee']

GROUP 2 states have relatively low percentage urban populations but high crime levels:
['Arizona' 'California' 'Colorado' 'Florida' 'Illinois' 'Maryland'
'Michigan' 'Nevada' 'New Mexico' 'New York' 'Texas']

GROUP 3 states have the lowest percentage urban populations and the lowest crime levels:
['North Dakota' 'South Dakota' 'Vermont' 'West Virginia']

US arrests (1973) Data set analysis

K means clustering

The K means clustering technique is simple and quick and works well with large data sets. One of the downsides is that we need to start the algorithm with the number of clusters as an input. The hierarchical clustering from above produced four clusters and so this seems like a good place to start and should allow us to compare the two methods.

Below is a direct comparison of the box plot analysis of the state groups created by the two clustering techniques.Interestingly, when the plots of the crime statistics are compared they show a very similar distribution. The plots of urban population though are a lot less similar - the hierarchical clustering technique having produced four distinct groups. The K means clustering is indicating that lower percent populations means lower crime figures which is potentially

misleading. A fuller investigation of such serious crime is likely to require analysis of many variables.