DS L3T12 Capstone Project

Analysis of US Arrests Data

Archie Macdonald

Examining the Data

This data set consists of crime statistics from the 50 states of the USA. The statistics refer to Assault, Murder and Rape rates per 100,000 people, as well as a fourth feature concerning Urban Population percentage per state.

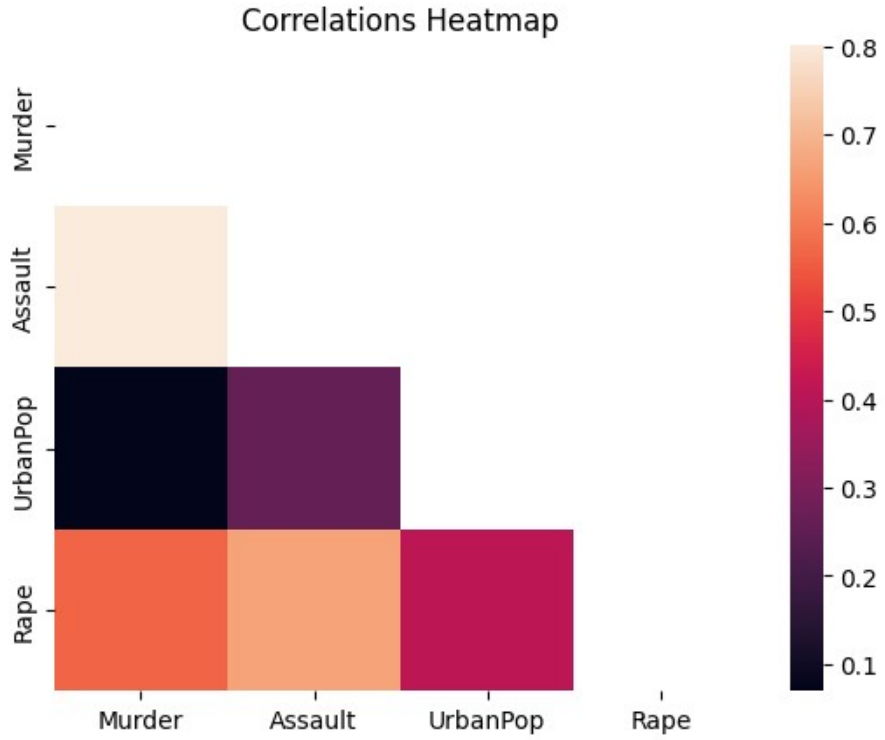
	City	Murder	Assault	UrbanPop	Rape
0	Alabama	13.2	236	58	21.2
1	Alaska	10.0	263	48	44.5
2	Arizona	8.1	294	80	31.0
3	Arkansas	8.8	190	50	19.5
4	California	9.0	276	91	40.6

As can be seen in the sample taken from the data, the values for “Assault” are much higher than the other two crime types, this means the data will need to be scaled at some point to properly balance its effect on the clustering.

A null check on the data revealed no missing values, so no imputation was required prior to the data being manipulated. A check of the fields also confirmed that all dtypes were correctly assigned, so no changes were required here either.

Correlations and PCA Scores

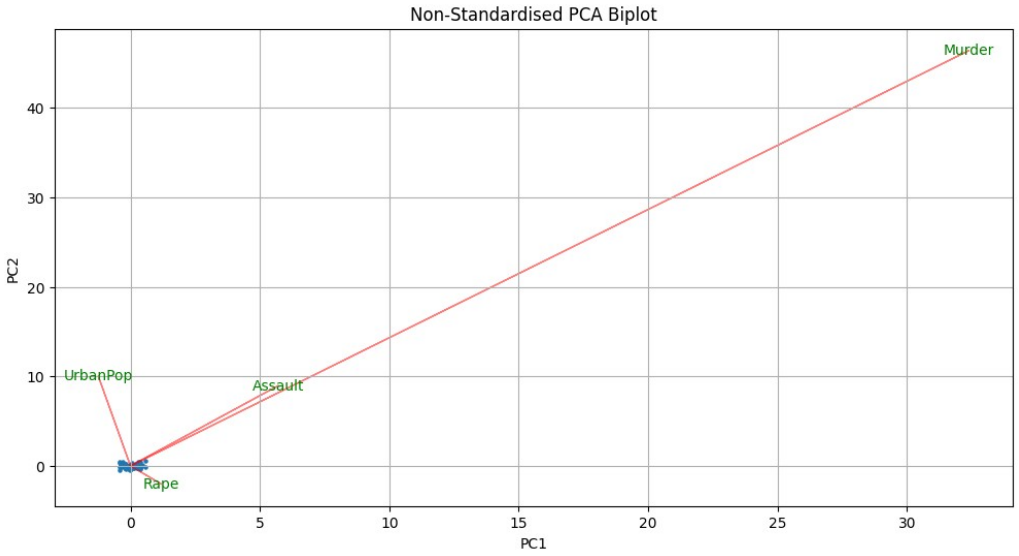
Assessment of the data coefficient values revealed three clear correlation patterns of note.



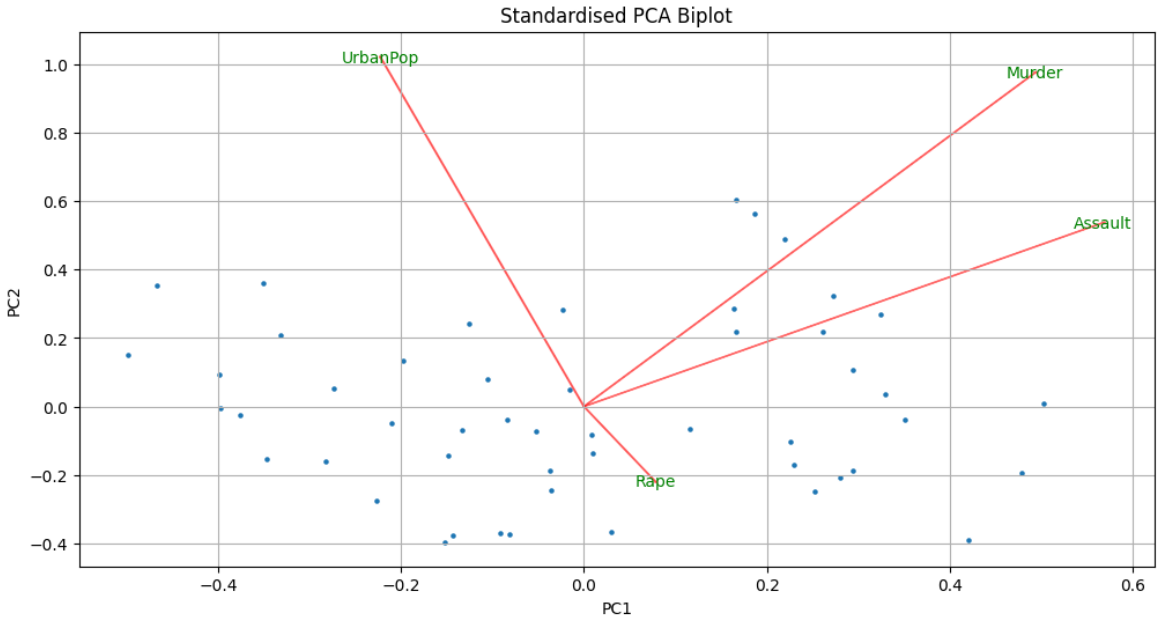
As can be seen in the heat map above, three observations can be made:

- Assault has almost no correlation with murder. This makes sense as all murder is assault with the victim dying, so crimes reported as murder are unlikely to also be reported as an assault.
- There are strong correlations between Urban Population percentage and all crime types.
- There is a slightly stronger correlation between Rape and Murder than between Rape and Assault.

The lack of a strong correlation between Assault and Murder is of particular importance, as this will likely be used by the models in order to draw the strongest patterns from the set. The relatively low values between Rape and Assault/Murder will also likely be quite important as well.



This PCA biplot of the data shows that, as expected, murder – and to a far lesser extent assault - have the largest effect on the distribution of data points. However, the scale of this makes the data hard to understand in this form; the below graph shows the same information with a standardised data set:

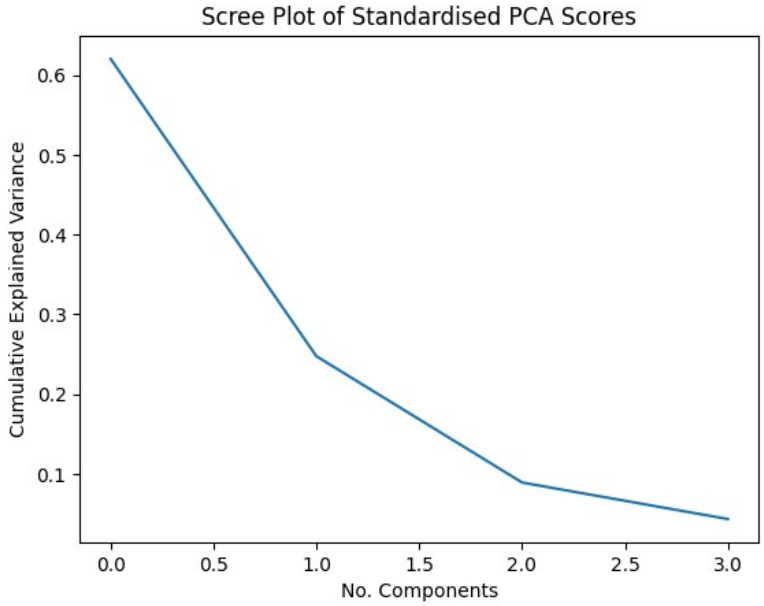


This graph allows us to differentiate between the individual data points much more easily. There do not seem to be any obvious grouping of the data in this plot, besides a somewhat noticeable gap splitting the data down the middle (this gap however is not particularly pronounced due to a single data point occupying the area. As can be seen, assault and

murder still comprise the strongest determining features in this PCA. The gulf between the significance of rape and the other three features is particularly pronounced here.

	Features	PC1 Importance	PC2 Importance	PC3 Importance	PC4 Importance
0	Murder	0.54	0.42	0.34	0.65
1	Assault	0.58	0.19	0.27	0.74
2	UrbanPop	0.28	0.87	0.38	0.13
3	Rape	0.54	0.17	0.82	0.09

This information in the graph is confirmed by the information in this table concerning importance scores for Principal Components (PC). For PC1, the scores for murder, assault and rape are very comparable, with UrbanPop having only roughly 60% of their scores. This is to be expected, as the correlation scores for UrbanPop were high across the board. PC2 however has a very different pattern, with UrbanPop having an extremely high score at 0.87, with only murder holding a somewhat high score at 0.42.



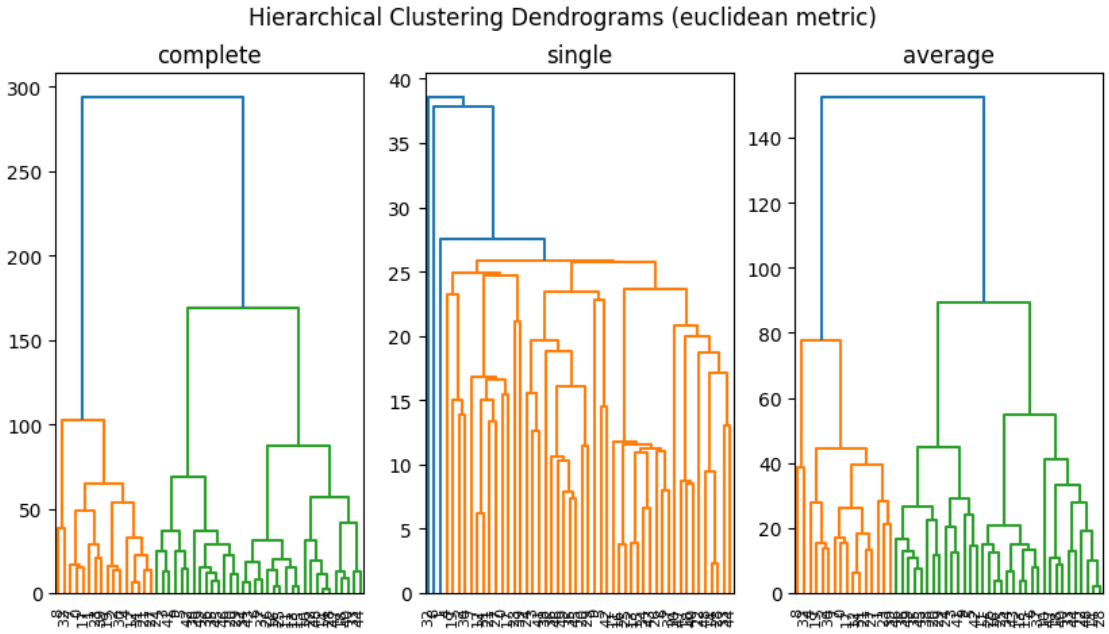
As can be seen in this scree graph, 90% of the cumulative variance is made up in the first three components, the the final component making up only around 5% of the total. If this was a larger data set, I would exclude all components beyond the third for efficiency's sake. However, due to the small number of features in the set, I will leave all features in as removing one will have minimal effect on the efficiency of the modelling.

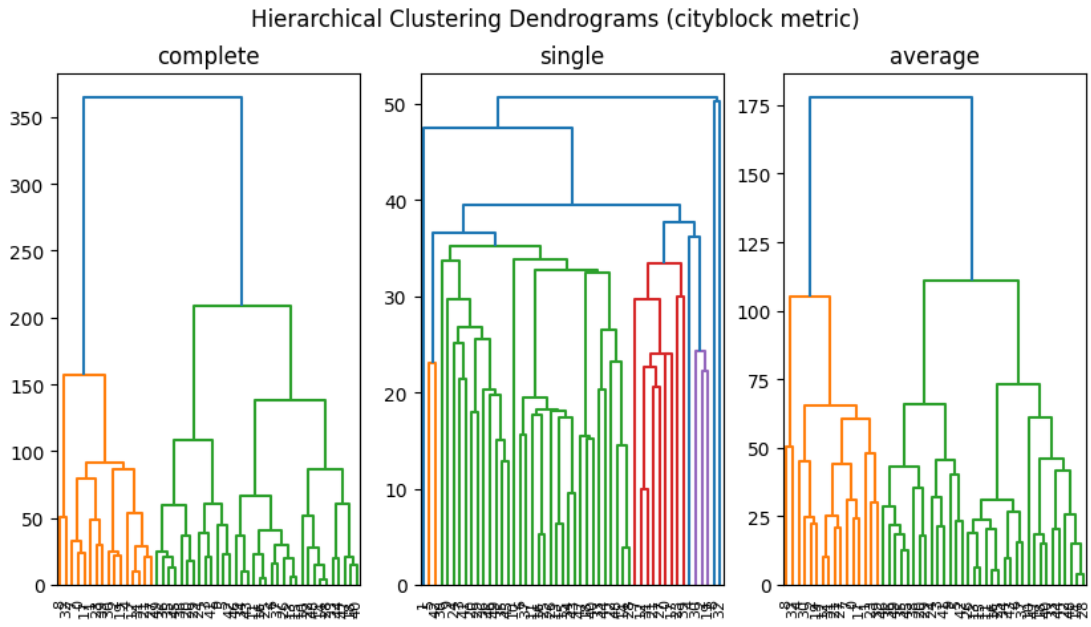
Clustering Models

I have clustered the data using three models: Agglomerative, K-Mean and Mean Shift.

Agglomerative

Prior to fitting the agglomerative model, I created dendrograms for six metric/method combinations, as can be seen below:

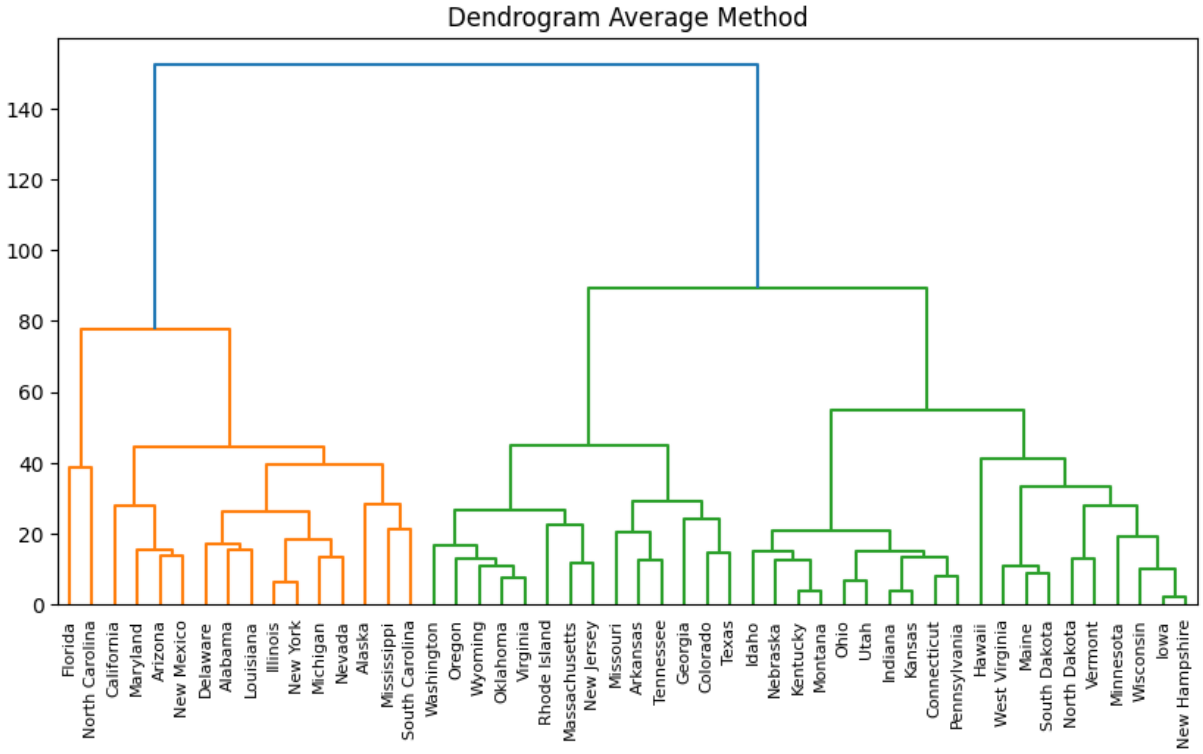




After comparing the above results, I determined that the most effective parameters for the agglomerative clustering would be as follows:

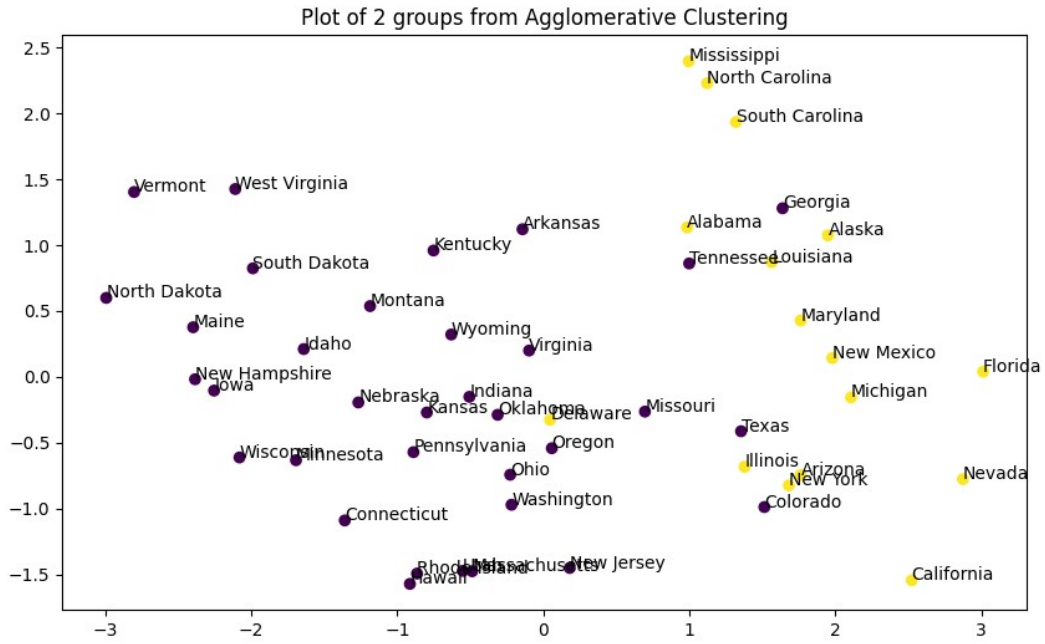
- Method: Average
- Metric: Euclidean
- Number of Clusters: 2

The method was chosen due to it producing the clearest grouping of the three, both for the Euclidean and the City Block metrics. Complete also created satisfactory groupings, with only the two "single" method graphs producing what I would consider unsatisfactory results. As can be seen, the differing metrics only seemed to make large changes in the "single" plots, with the "average" and "complete" outputs remaining largely the same. The dendrogram for the chosen metrics is as follows:



As can be seen, the two groups are somewhat dissimilar in size, although the clusters are both more than large enough to prove useful in analysis. Without further information on the states in the groupings, it's difficult to identify patterns, although from the looks of things both groups contain a reasonably even distribution of urban and rural states, which would confirm the low UrbanPop PCA score we saw earlier. The green cluster above has two noticeably even sub-clusters within the group, which may allow for larger numbers of clusters if so desired.

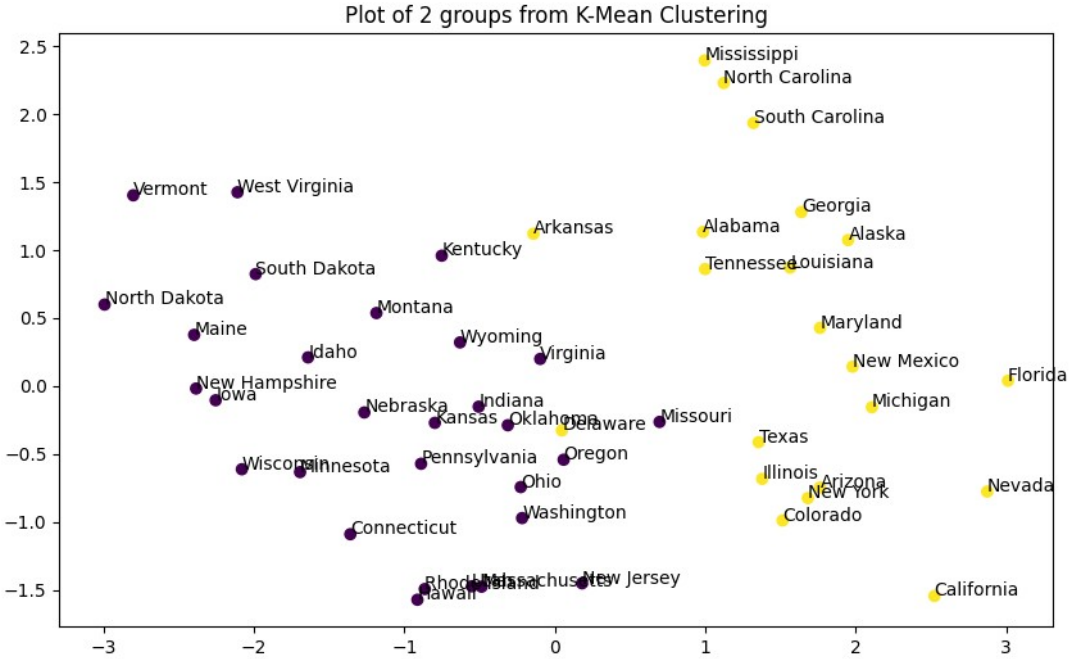
The scatter plot for the agglomerative clustering is as such:



Unfortunately, there does seem to be some noticeable overlap between the groups, with Georgia and Delaware being the two most standout points in this regard. This being said, the two groups do seem to be largely coherent and split in a largely logical fashion.

K-Mean Clustering

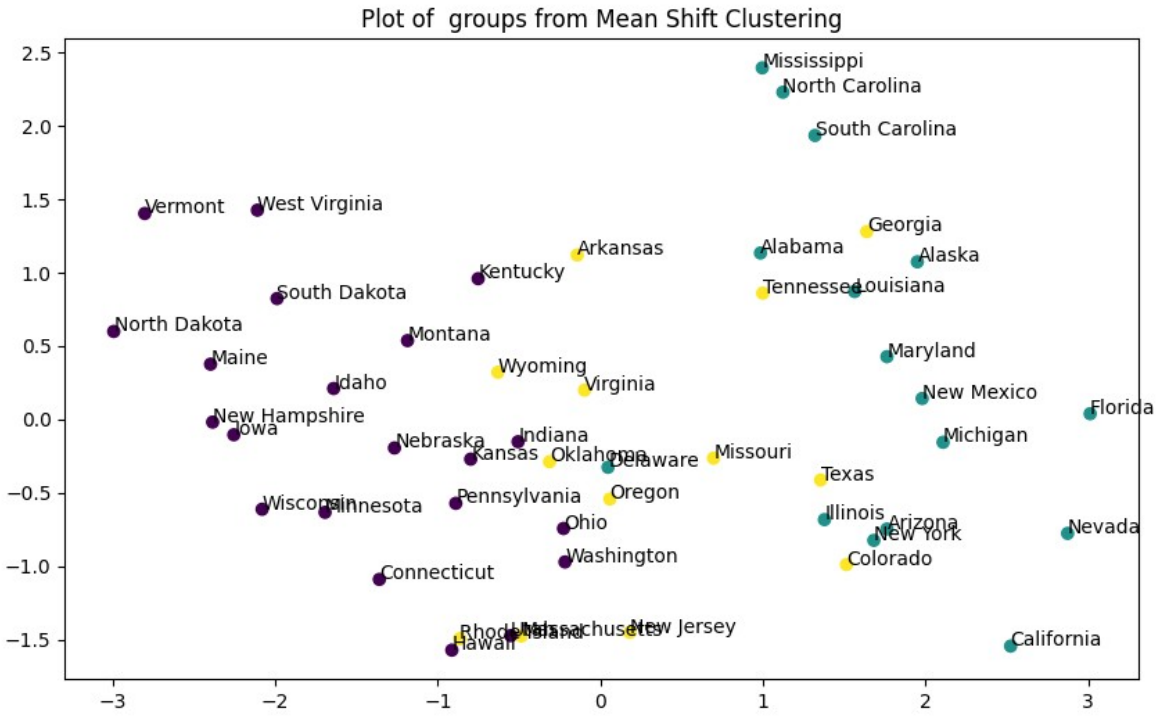
For the K-Mean clustering, I have used the same number of clusters as used in the Agglomerative model (which is 2). This produced the following scatter graph:



The clusters produced are largely similar to Agglomerative model, with two clear groups separated left from right. There is however a noticeable improvement in overlap, with Texas and Colorado being grouped with the right cluster. Delaware remains a noticeable exception, though the overall effect seems much more logical to an observer.

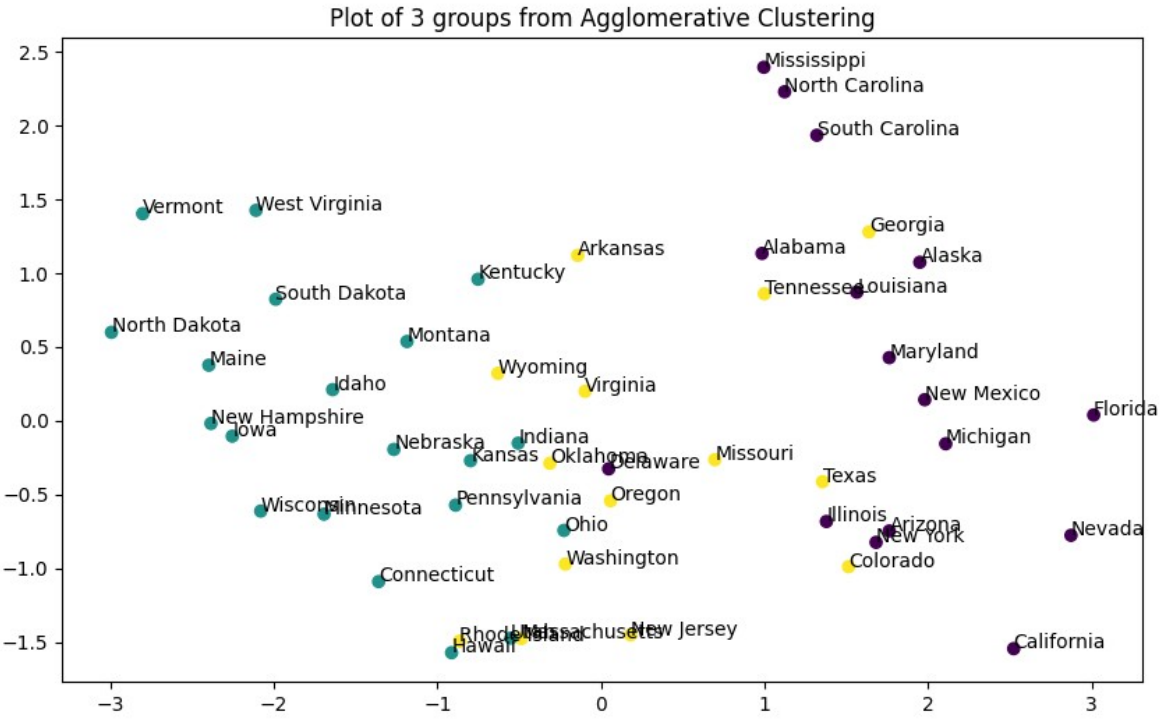
Mean Shift

To provide a little more variety in, I also fit a Mean Shift model as well to see how it would handle the data. Mean Shift doesn't have any common parameters with the other two models, so I passed the data to the model with default settings:

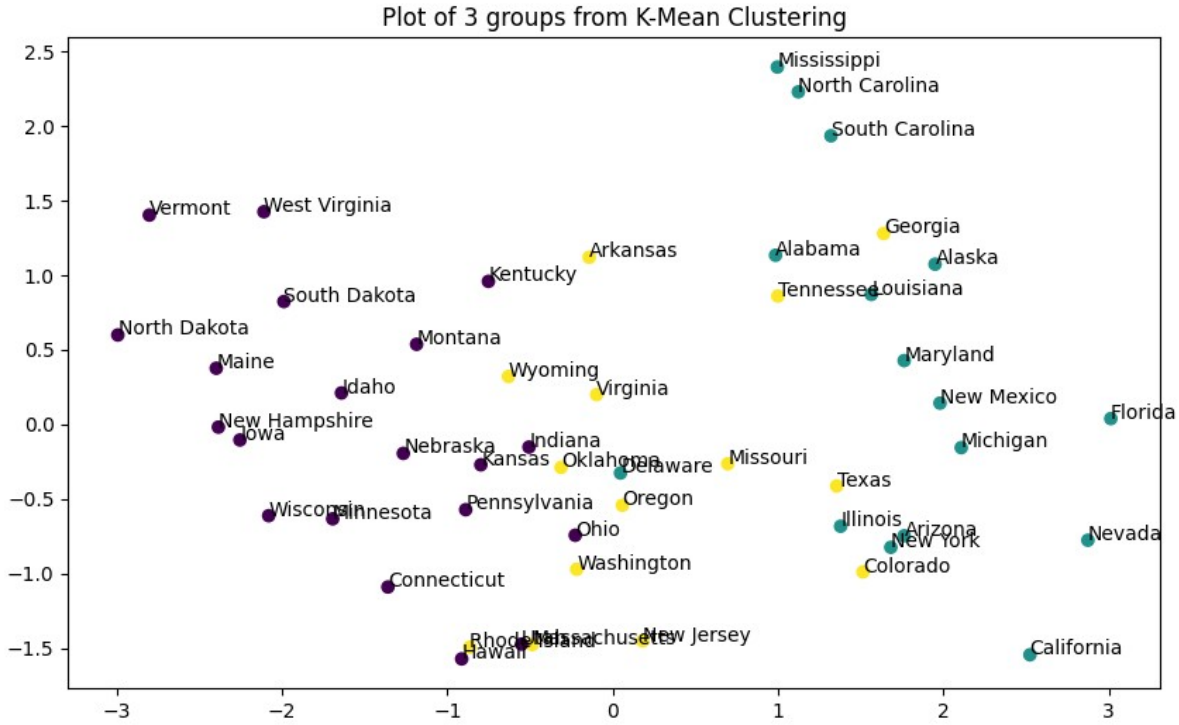


The most notable difference here is that the model decided upon three clusters as being the most effective option. There is now a third group occupying the middle ground between the left and right clusters, with minor overlap between all three. Interestingly, Delaware is still a notable outlier for the right cluster, with it remaining part of that group rather than joining the centre group which it lays deep inside. This would suggest that the initial components used to group this data are ineffective at grouping Delaware in a cohesive manner.

In order to compare the models with Mean Shift, I created two further models of Agglomerative and K-Mean clustering with identical parameters to the first two, except with three clusters.



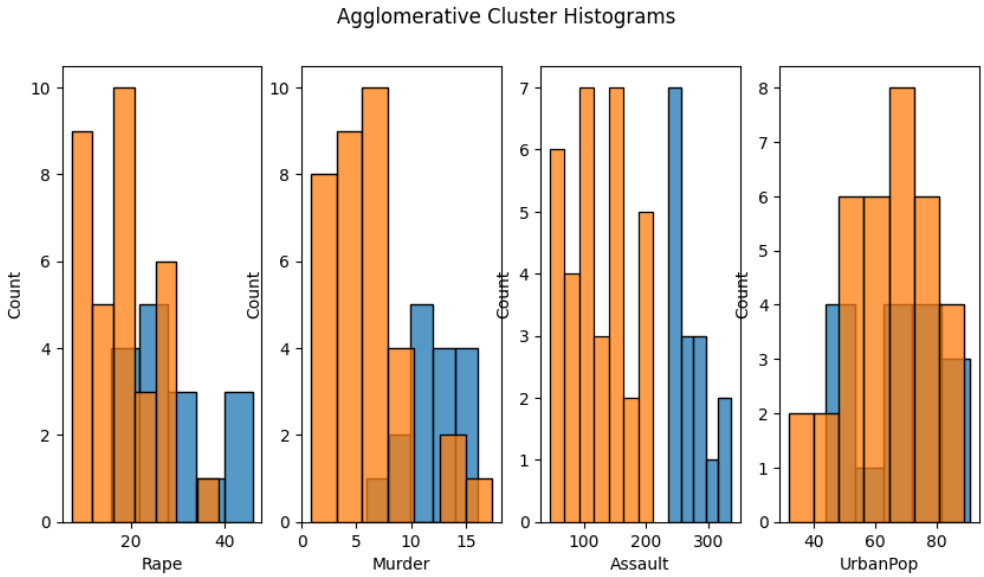
Perhaps unsurprisingly, the groupings are very similar to the Mean Shift model for the agglomerative model, with the only difference being the inclusion of Washington in the central cluster.



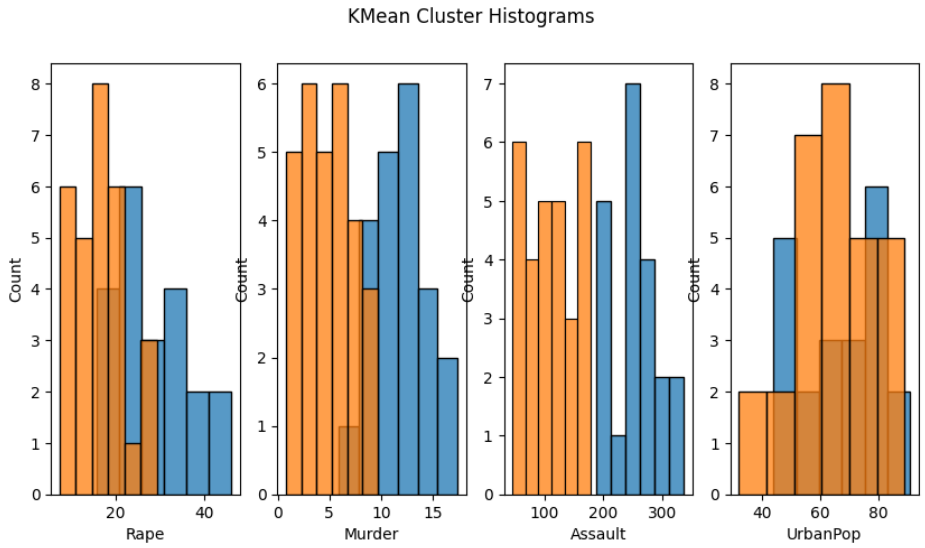
Interestingly, the groupings for 3 clusters for the Agglomerative and K-Mean models are identical. This may suggest that 3 groups, rather than two, would be more useful overall. As always, Delaware remains a staunch contrarian and stands out as an outlier in both of these models.

Histograms

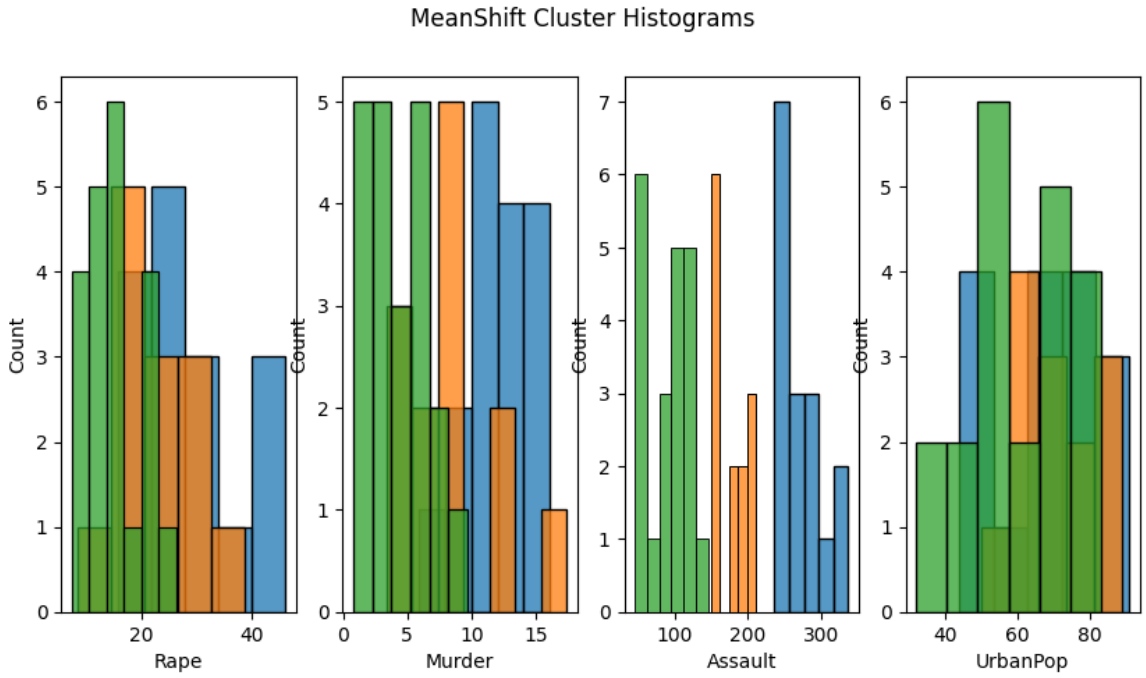
As a final point of analysis, I have created histograms for each model detailing the distribution of clusters for each feature in the data set.



As can be seen for the agglomerative model, the two most important features are (as predicted earlier) Murder and Assault. Of the two, Assault seems to be the most significant determiner, as the two clusters for this model are clearly separated within the "Assault" field. Rape sees some separation of the groups, but they are largely indistinguishable for "UrbanPop" with almost total overlap.

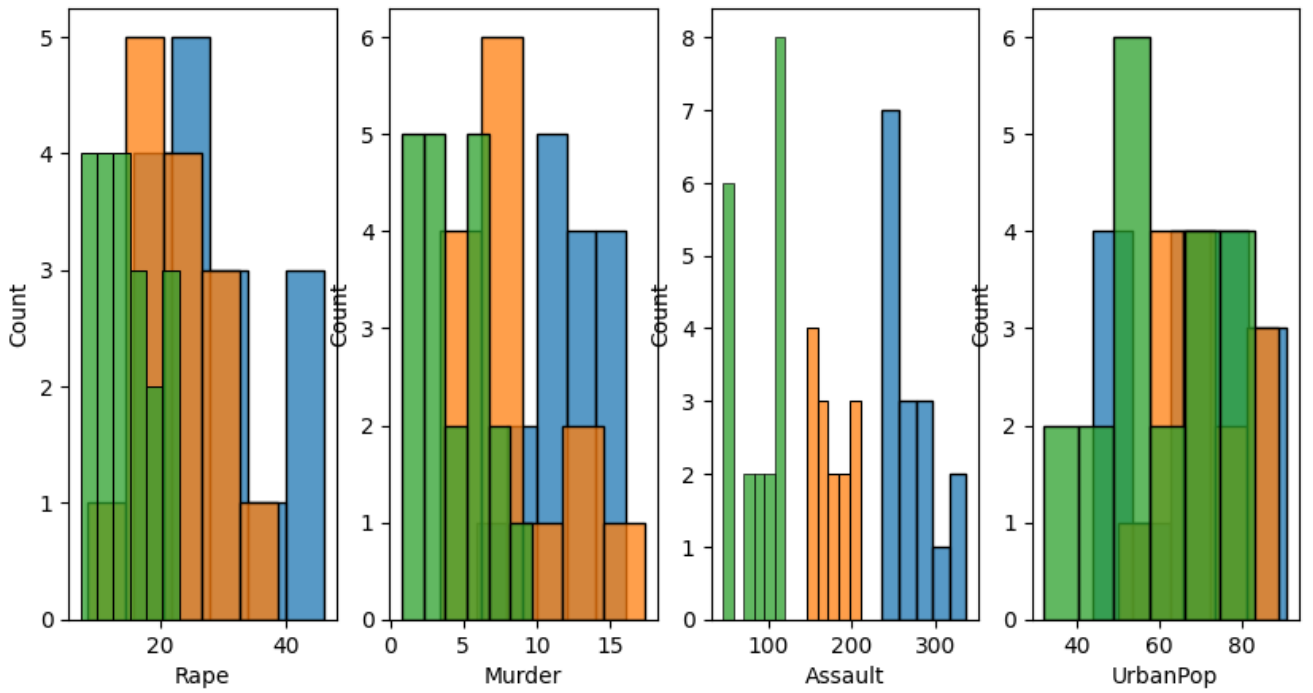


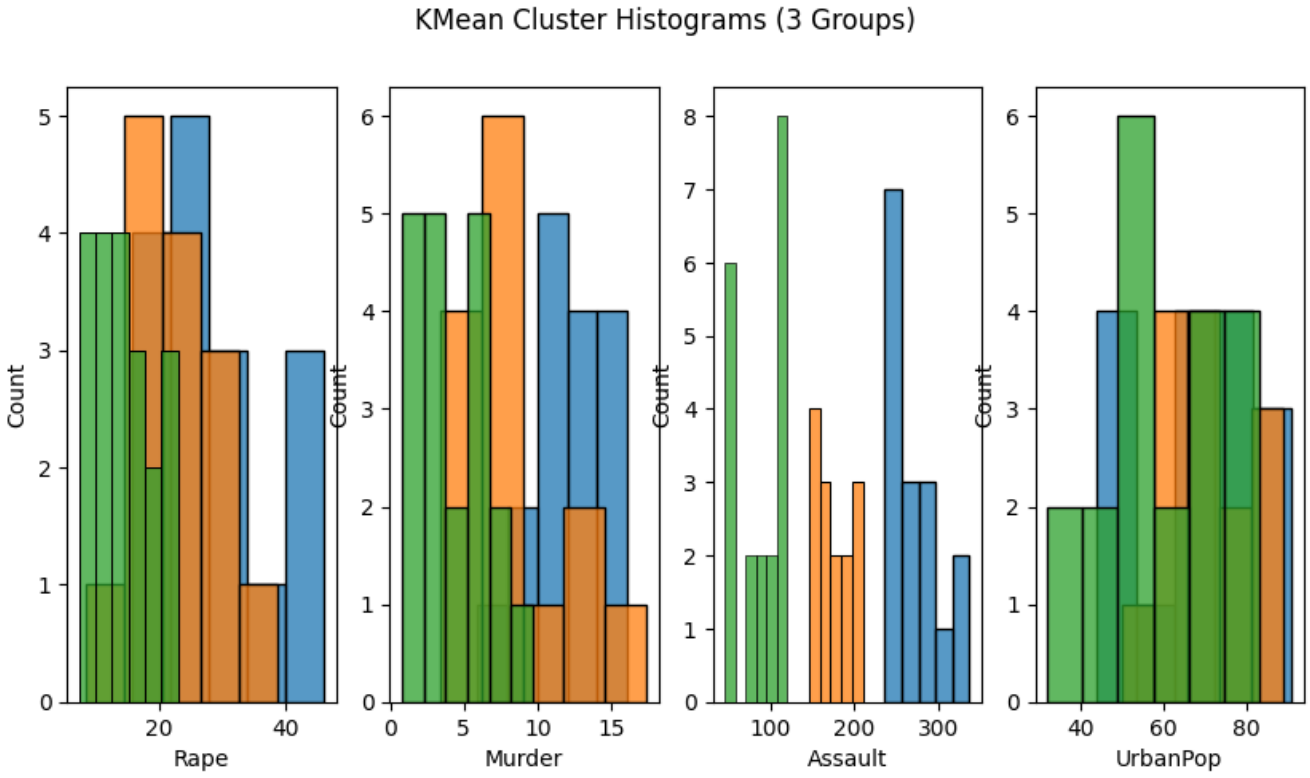
The K-Mean model is very similar to the Agglomerative model, as seen in the scatter plot, with the major change being in the “murder” feature, which seems to be much more evenly distributed between clusters. Interestingly, the “assault” feature also seems to be slightly more evenly distributed between the two, with the orange cluster being slightly smaller than with the Agglomerative model.



Due to the third group in the Mean Shift model, it is difficult to make comparisons with the first two, there are some characteristics of note however. Firstly, the groupings within the “Assault” field are just as well defined, with no overlap at all. “Rape” and “Murder” both have a slightly messier appearance, although there are three noticeable clusters with somewhat even distribution within both. UrbanPop, as with the other two groups, is heavily overlapped, with little to suggest clear groupings within that feature.

As mentioned with the scatter plots, the groupings for K-Mean and Agglomerative clustering are identical when a third cluster is introduced. This is born out when comparing histograms for the two models, which clearly show that the models separate the data in an identical fashion, with a clear difference between these two models and the MeanShift model.


Agglomerative Cluster Histograms (3 Groups)
Count
Rape
Murder
Assault
UrbanPop


KMean Cluster Histograms (3 Groups)
Count
Rape
Murder
Assault
UrbanPop