Unsupervised Machine Learning Report: "*USArrest*" Dataset

Introduction

Summary of the data set

The 'ArrestUS' dataset 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.

This dataset can be located on Kaggle:
https://www.kaggle.com/datasets/kurohana/usarrets

DATA CLEANING

SUMMARY OF THE METHODS AND VISUALISATIONS DONE DURING DATA CLEANING

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

- It is not necessary at this point to remove any of the columns from the database.
- Checked data types: Murder and Rape are float64 and Assault and UrbanPop are int64.
- Float64 data is already rounded to 1 decimal place.
- Data types do not need to be changed.

MISSING DATA

ANY MISSING DATA? HOW DID YOU HANDLE IT

- Visualised the number of missing values.
- There is 0 missing data.

DATA SUMMARY

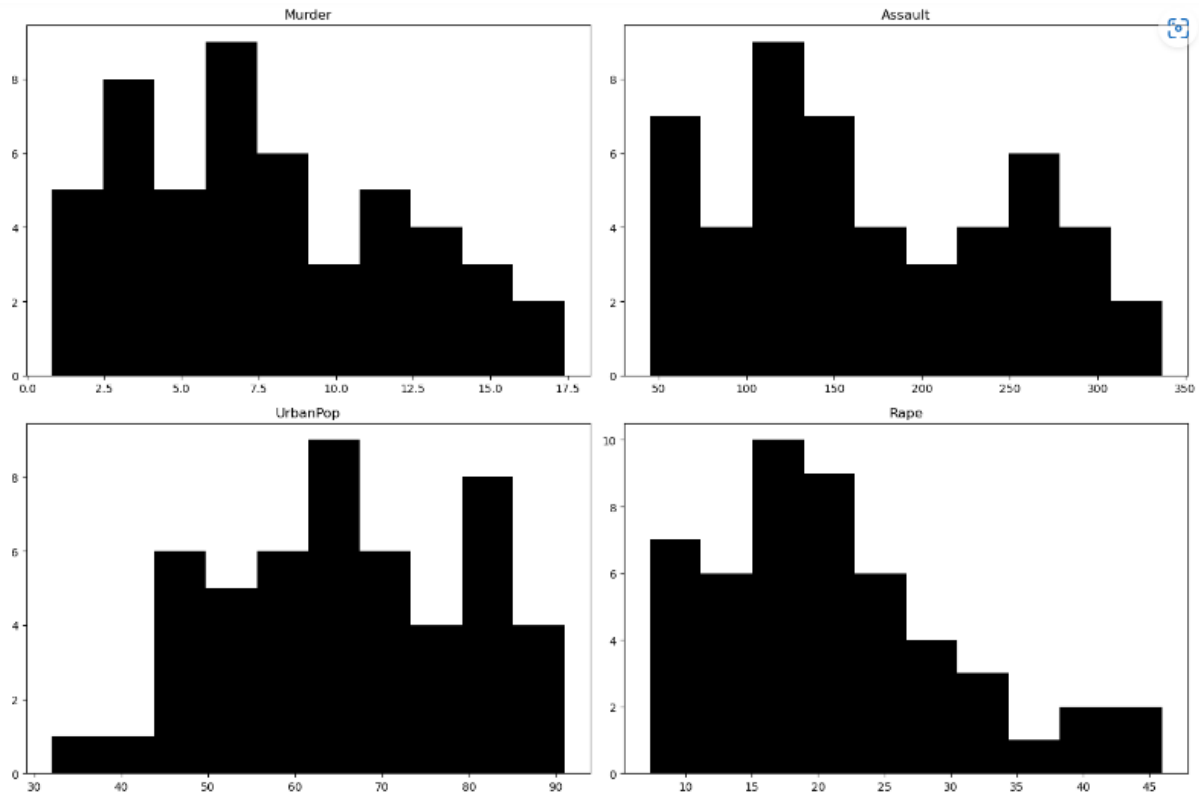
SUMMARY TABLE

Visualised a table showing the mean, standard deviation, minimum, maximum of the four variables.

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

At first glance, there are more assaults per 100,000 residents than any other crime. The variable's mean and standard deviation is dramatically higher than the other variables. This suggests that assault is a more common crime and indicates that scaling the data will be useful to keep the assault variable from impacting the analysis disproportionately.

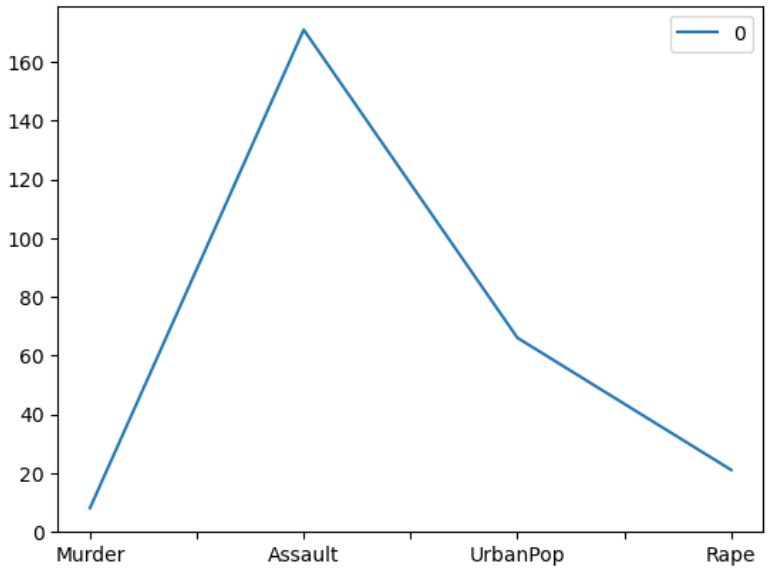
Histograms of the four variables emphasises how the assault variable is disproportionate to the other three variables:



DATA STORIES AND VISUALISATIONS

EXTRACT STORIES AND ASSUMPTIONS BASED ON VISUALISATIONS

1. Visualisation of average of each variable



The line graph shows the averages are:

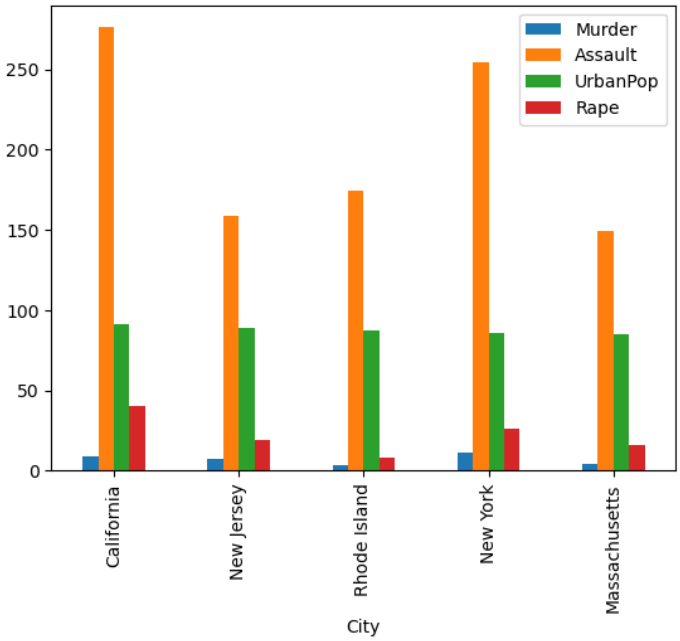
- Murder: 8
- Assault: 171
- UrbanPop: 66
- Rape: 21

2. UrbanPop

Highlighted which 5 cities had the largest percent of the population living in urban area:

City	Murder	Assault	UrbanPop	Rape
California	9.000	276	91	40.600
New Jersey	7.400	159	89	18.800
Rhode Island	3.400	174	87	8.300
New York	11.100	254	86	26.100
Massachusetts	4.400	149	85	16.300

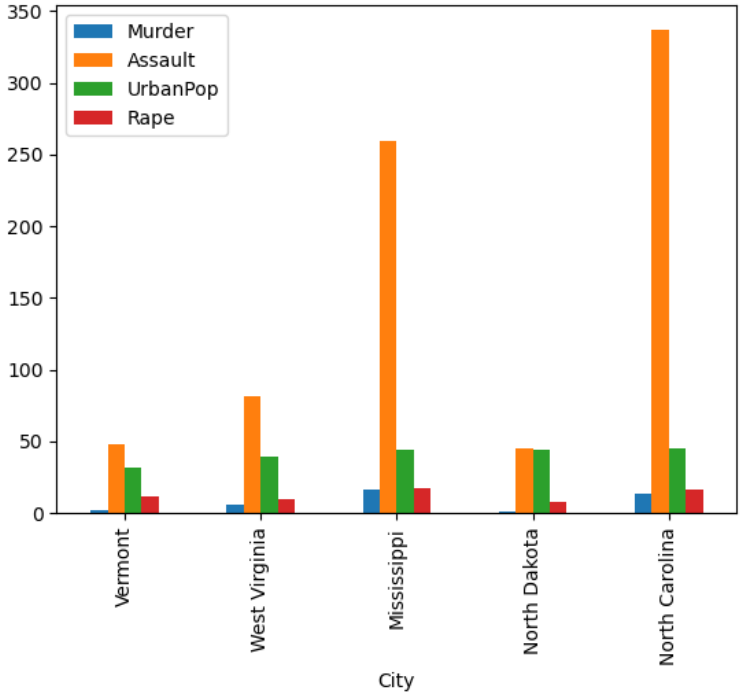
Visualised this further using a bar chart:



Highlighted which 5 cities had the smallest percent of the population living in urban area:

City	Murder	Assault	UrbanPop	Rape
Vermont	2.200	48	32	11.200
West Virginia	5.700	81	39	9.300
Mississippi	16.100	259	44	17.100
North Dakota	0.800	45	44	7.300
North Carolina	13.000	337	45	16.100

Visualised this further using a bar chart:



By comparing the two charts, it can be shown that UrbanPop does not influence the number of arrests per 100,000 residents, for assault, murder, and rape.

For example, North Carolina, with the lowest percent (45) of the population living in urban areas, has an arrest rate of over 300 per 100,000 residents for assault. On the other hand, California, with the highest percent (91) of the population living in urban areas, has an arrest rate of over 250 per 100,000 residents for assault.

Removed UrbanPop from dataset.

3. Crime Comparison

Comparing the 5 states with the smallest and highest murder rate per 100,000 residents:

Smallest

City	Murder	Assault	Rape
North Dakota	0.800	45	7.300
Maine	2.100	83	7.800
New Hampshire	2.100	57	9.500
Iowa	2.200	56	11.300
Vermont	2.200	48	11.200

Highest

City	Murder	Assault	Rape
Georgia	17.400	211	25.800
Mississippi	16.100	259	17.100
Florida	15.400	335	31.900
Louisiana	15.400	249	22.200
South Carolina	14.400	279	22.500

Comparing the 5 states with the smallest and highest assault rate per 100,000 residents:

Smallest

City	Murder	Assault	Rape
North Dakota	0.800	45	7.300
Hawaii	5.300	46	20.200
Vermont	2.200	48	11.200
Wisconsin	2.600	53	10.800
Iowa	2.200	56	11.300

Highest

City	Murder	Assault	Rape
North Carolina	13.000	337	16.100
Florida	15.400	335	31.900
Maryland	11.300	300	27.800
Arizona	8.100	294	31.000
New Mexico	11.400	285	32.100

Comparing the 5 states with the smallest and highest rape rate per 100,000 residents:

Smallest

City	Murder	Assault	Rape
North Dakota	0.800	45	7.300
Maine	2.100	83	7.800
Rhode Island	3.400	174	8.300
West Virginia	5.700	81	9.300
New Hampshire	2.100	57	9.500

Highest

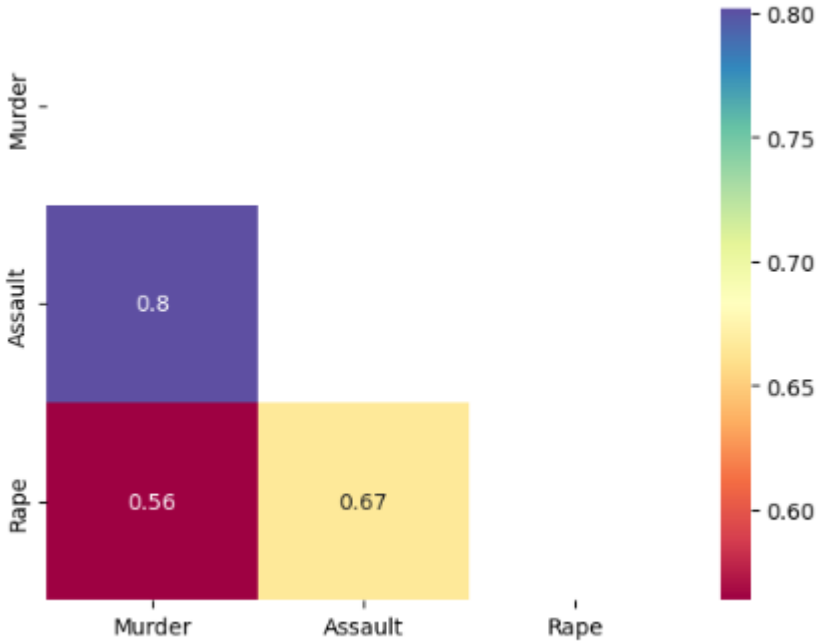
City	Murder	Assault	Rape
Nevada	12.200	252	46.000
Alaska	10.000	263	44.500
California	9.000	276	40.600
Colorado	7.900	204	38.700
Michigan	12.100	255	35.100

If the murder arrest rate is low in a state, you can assume that the other crimes arrest rate is low too. If the murder arrest rate is high in a state, you can assume that the other crimes arrest rate is high too. This trend is similar for all arrests.

CORRELATION ANALYSIS

COMPUTE CORRELATIONS BETWEEN THE DIFFERENT COLUMNS

Pandas offers a highly useful function, corr, which allows us to compute correlations between the different columns. The standard correlation coefficient is the Pearson coefficient. It returns a matrix of values. It is often useful to visualise these as a plot. Both Pandas and Seaborn have functions for plotting correlation heatmaps, but Seaborn's offers more adjustability.



The variables only have positive correlations which mean that an increase in one will also leads to an increase in another.

From the correlation plot, it is evident that murder has a relatively strong positive correlation to assault and rape. These correlations are intuitive as states with a high murder rate are probably areas where other violent crimes occur.

The correlation plot also indicates that assault has a strong positive correlation with rape.

Overall, the variables that have strong positive correlations with each other. This makes the data a good candidate for Principal Components Analysis.

PRINCIPAL COMPONENTS ANALYSIS (PCA)

FIND THE UNDERLYING VARIABLES THAT BEST DIFFERIENTIATE THE OBSERVATIONS

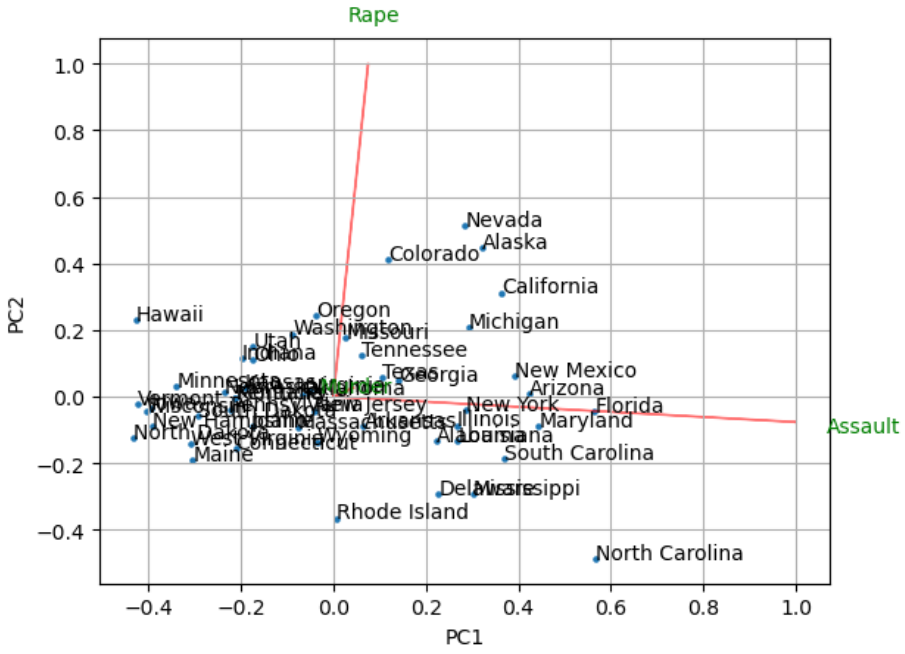
Principal Components Analysis (PCA) is a method for finding the underlying variables (i.e. principal components) that best differentiate the observations by determining the directions along which your data points are most spread out. Since the determination of the principal components is based on finding the direction that maximises the variance, variables with variance that are much higher than the other variables tend to dominate the analysis purely due to their scale. PCA is a dimensionality reduction technique which is useful when we have a lot of variables and need to reduce these.

PCA data:

	0	1	2
0	65.223	-4.860	2.813
1	93.737	16.227	-2.124
2	123.530	0.362	-4.868
3	19.081	-3.165	0.295
4	106.355	11.325	-3.528

PCA standard deviation = 83.64, 6.98, 2.59
PCA proportion of variance explained = 9.92e-01, 6.90e-03, 9.54e-04
PCA cumulative proportion = 6996.48, 7045.14, 7051.87

Unscaled Biplot:



Each point on a biplot is the projected observation, transformed from the original data. The importance of each feature is indicated by the length of the arrows on the biplot. This corresponds to the magnitude of the values in the eigenvectors. From this biplot you can see that assault and rape are the most important features as the arrows to each of those dominate the biplot.

Most of the cities are clustered together which indicates that they have similar principal component scores.

From the unscaled biplot, it is difficult to define relationships. It appears the features overpower each other.

The unscaled biplot data can be summarised as:

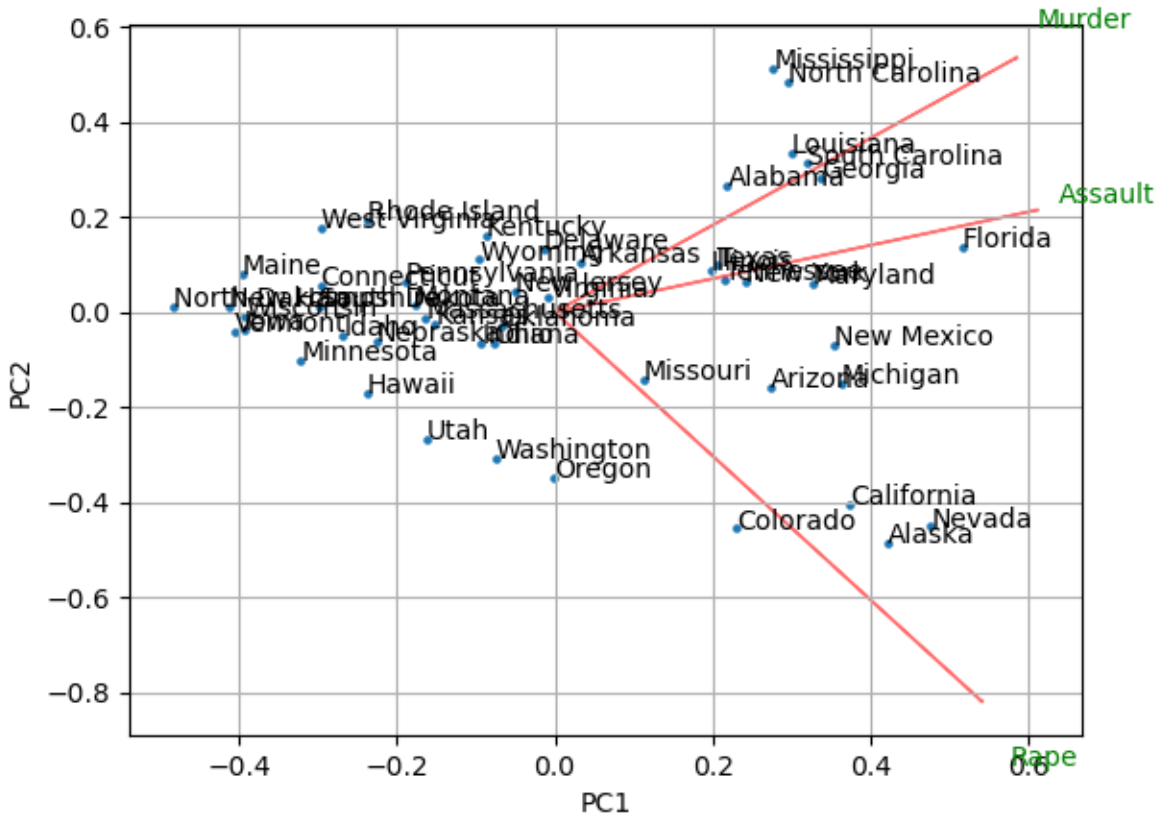
	Features	PC1 Importance	PC2 Importance
0	Murder	0.042	0.026
1	Assault	0.996	0.076
2	Rape	0.075	0.997

Scaled Biplot:

We standardise the data so that some features do not swamp the others.

Scaled PCA data:

	0	1	2
0	1.210	0.842	0.164
1	2.332	-1.539	-0.039
2	1.519	-0.503	-0.887
3	0.178	0.328	-0.072
4	2.066	-1.285	-0.385



Immediately you can spot the difference between the unscaled and scaled biplot. as the features are much clearer. The scaled biplot shows that murder and assault are closely related. If you group murder and assault together it can be argued that they have little correlation with rape.

It is shown that Colorado, California, Alaska, and Nevada have the highest number of arrests per 100,000 residents for rape.

Florida has the highest number of arrests per 100,000 residents for assault.

Mississippi, North Caroline, Louisiana, South Carolina, Georgia and Alabama have the highest arrests per 100,000 residents for murder.

There are still a lot of cities grouped together which indicates that they have similar principal component scores.

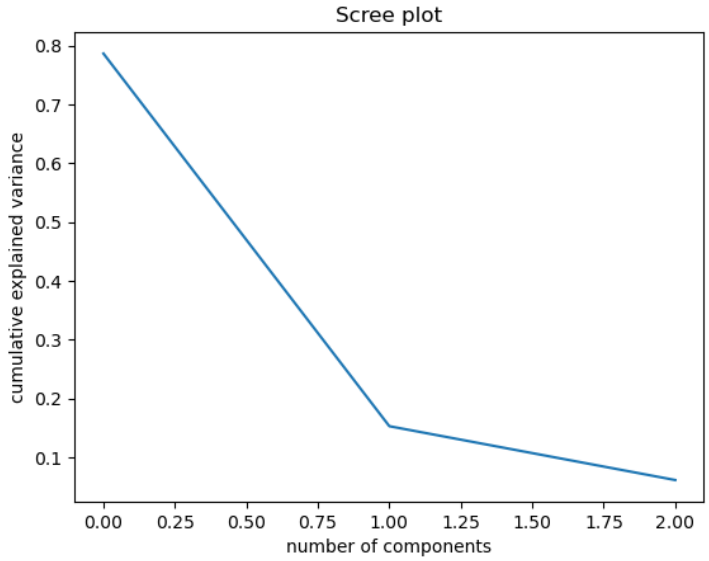
The scaled biplot data can be summarised as:

	Features	PC1 Importance	PC2 Importance
0	Murder	0.583	0.534
1	Assault	0.608	0.214
2	Rape	0.539	0.818

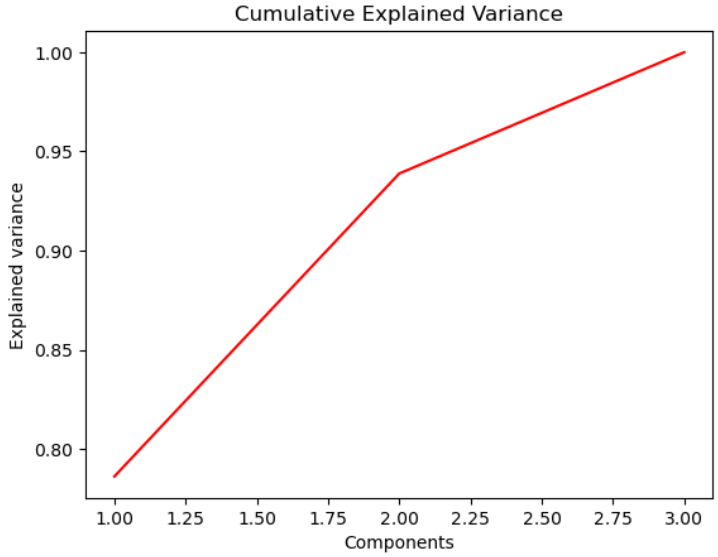
SCREE PLOT AND CUMULATIVE EXPLAINED VARIANCE PLOT

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.

Scree Plot:



Cumulative Explained Variance plot:



It appears the principal components are spread through 75% of the data. Therefore, it is better to use whole dataset for cluster analysis.

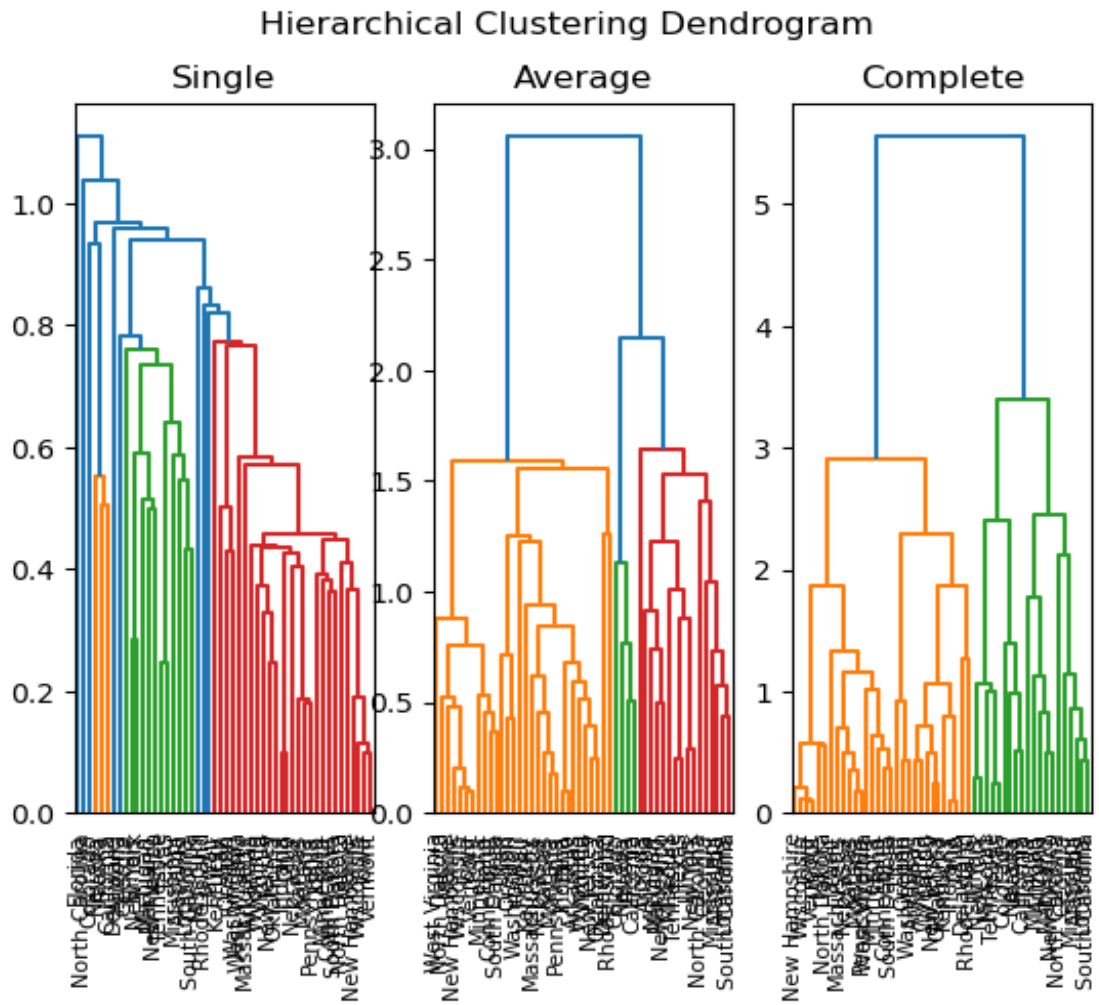
Also, there are only three variables so it is not appropriate to scale down the database.

HIERACHICAL CLUSTERING

VISUALISE THE CLUSTERS IN DENDROGRAMS

Hierarchical clustering has the advantage that we can see the clusters visually in a dendrogram and don't have to specify the number of clusters before running the algorithm.

In order to determine the method used to measure the distance between clusters, we plotted the various dendrograms for the single, complete, and average linkage methods.



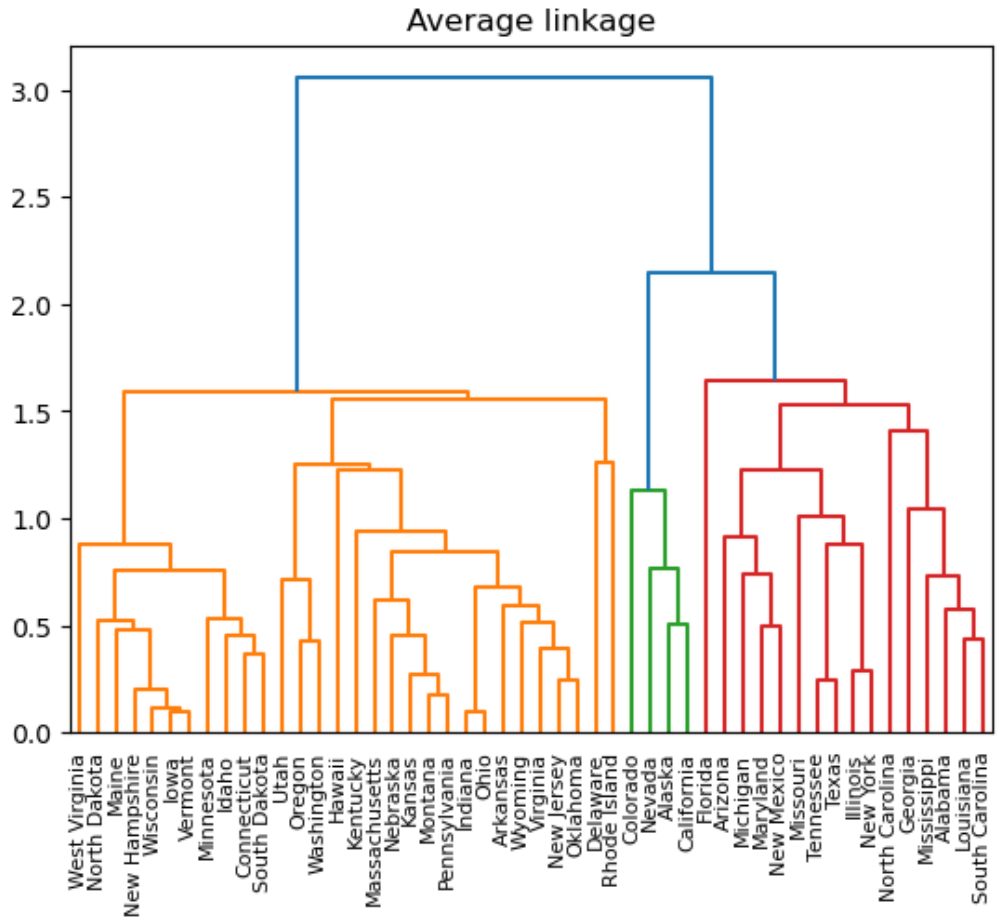
By calculating the silhouette scores, it can be determined which method is best:

- Single = 0.20166299602779528
- Average = 0.5238335497369089
- Complete = 0.5238335497369089

The silhouette scores show that the single dendrogram has the lowest score (0.20), indicating the clusters are not well defined in this method. The silhouette score for the average and complete dendrograms are the same (0.52). The difference between the two is that the average method has four clusters whilst the complete method has three clusters.

Average Linkage Method:

A clearer dendrogram for the average linkage method is shown below:



Clusters:

Orange = West Virginia, North Dakota, Maine, New Hampshire, Wisconsin, Iowa, Vermont, Minnesota, Idaho, Connecticut, South Dakota, Utah, Oregon, Washington, Hawaii, Kentucky, Massachusetts, Nebraska, Kansas, Montana Pennsylvania, Indiana, Ohio, Arkansas, Wyoming, Virginia, New Jersey, Oklahoma, Delaware, Rhode Island

Green = Colorado, Nevada, Alaska, California

Red = Florida, Arizona, Michigan, Maryland, New Mexico, Missouri, Tennessee, Texas,

Illinois, New York, North Carolina, Georgia, Mississippi, Alabama, Louisiana, South Carolina
Blue = Uta, Oregon, Washington, Hawaii, Kentucky, Massachusetts, Nebraska, Kansas, Montana Pennsylvania, Indiana, Ohio, Arkansas, Wyoming, Virginia, New Jersey, Oklahoma, Delaware, Rhode Island, Colorado, Nevada, Alaska, California, Florida, Arizona, Michigan, Maryland

Analysis:

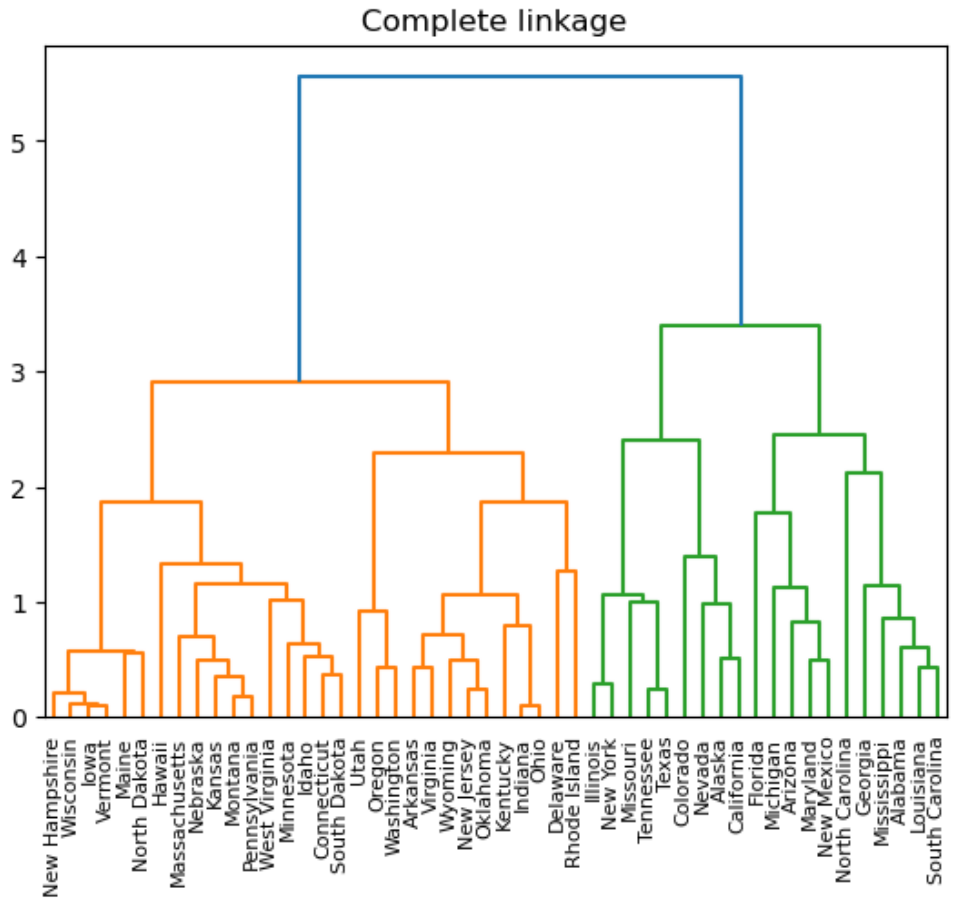
The orange cluster contains the states with the lowest number of arrests per 100,000 residents for assault, murder, and rape.

The green cluster contains the four states with the highest arrests per 100,000 residents for rape.

The red cluster contains the states with the highest number of arrests per 100,000 residents for assault and murder.

Complete Linkage Method:

A clearer dendrogram for the complete linkage method is shown below:



Clusters:

The complete orange cluster contains the same states as the average method's orange cluster.

The complete green cluster combines all the states in the average method's red and green clusters.

The blue clusters were not an exact match. The complete method's blue cluster contained the following whilst the average method's blue cluster did not: Minnesota, Idaho, Connecticut, South Dakota, Illinois, New York, Missouri, Tennessee, Texas. On the other hand, the average method's blue cluster contained the following whilst the complete method's blue cluster did not Hawaii, Massachusetts, Nebraska, Kansas, Montana Pennsylvania, Florida, Arizona, Michigan, Maryland.

Analysis:

The orange cluster contains the states with the lowest number of arrests per 100,000 residents for assault, murder, and rape.

The green cluster on the complete model contains the states with the highest number of arrests per 100,000 residents for assault, murder, and rape.

Overall:

The average method appears to be more reliable as it breaks the clusters down per crime.

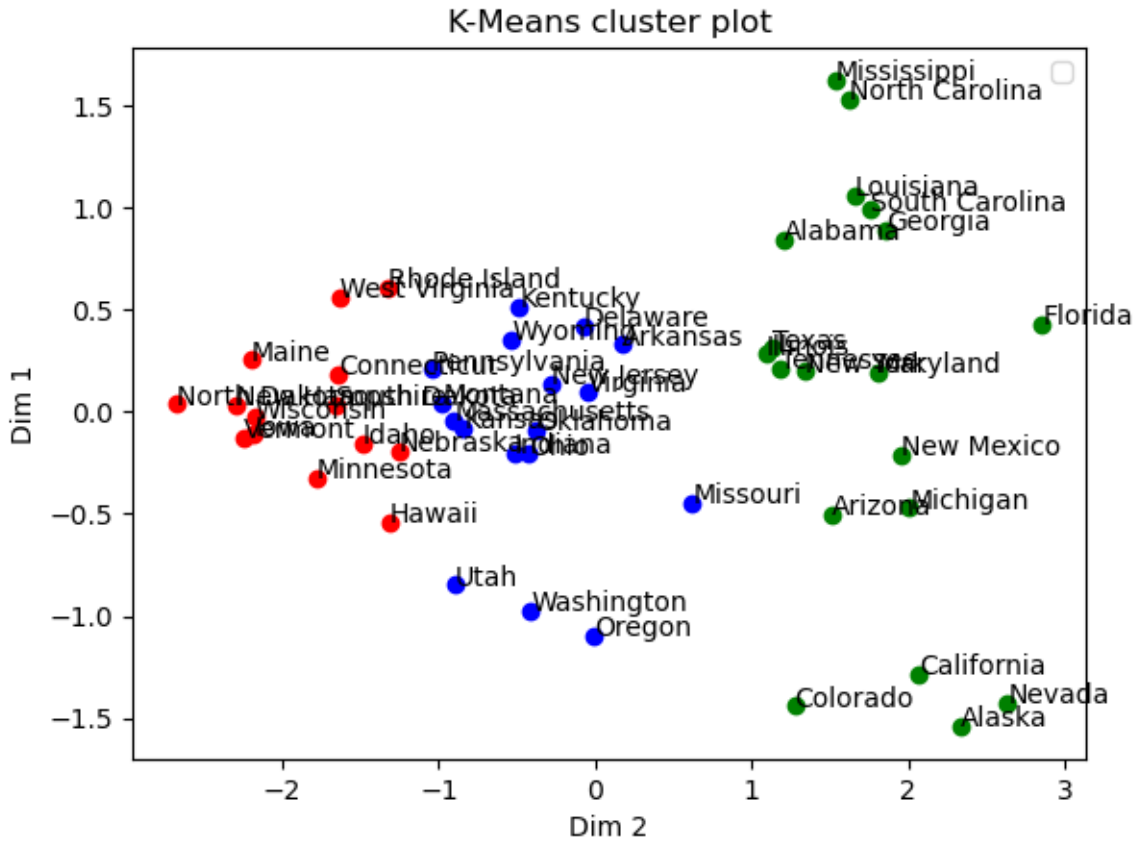
K-MEANS CLUSTERING

K_MEANS CLUSTER PLOT

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.

For our analysis, it seems that the shape of clusters is likely to be regular based on

the PCA biplot. K will be set to 3. A visualisation of the clusters is shown in the figure below.



The K-means cluster plot has clustered the states based on the amount of arrests made as a whole. The green clusters are the states where there are more arrests, the red clusters are where there are the least number of arrests, and the blue clusters represent a middle ground. It groups the states very similarly to the hierarchal clustering.

The silhouette score for the plot is 0.38058945538162325. This means the clusters are not well defined.

When you compare K-means clustering and hierarchal clustering, hierarchal clustering provides a more accurate result.

Github link

https://github.com/RachelBirrell97

THIS REPORT WAS WRITTEN BY: RACHEL BIRRELL