Hyperiondev

TASK

Capstone VII - PCA-Unsupervised Machine Learning

Visit our website

Introduction

The aim of this task was to conduct a comprehensive Principle Component Analysis (PCA) on the UsArrests.csv data set. A PCA is used in order to reduce the amount of variables that share the same differences in a dataset, by encoding the data into fewer dimensions. This analysis also used clustering techniques such as K-Nearest Neighbours and Hierarchical Clustering in order to extrapolate key information.

The US Arrests data set consisted of columns denoting the 50 states and the occurrences of crimes such as Rape, Murder and Assault, as well as the percentage of such crimes in Urban populations denoted by UrbanPop.

There were a series of pre-processing steps that were in order before conducting the analysis, such as finding and clearing the data of any missing values and encoding certain columns into the right format for processing.

DATA Cleaning

The data was first visualised using the head() method and seeing that all the columns were given appropriate names, no further action was taken in cleaning. However, for future uses and ease of processing the ‘City’ column was made the index column.

MISSING DATA

After using the isnull() method to identify any columns with missing values, the table generated that the data frame (df) had none. Seeing as there were no missing values in the df there was no need to do any computational methods to replace missing slots.
In the case there were any missing values present, they would be handled using the dropna() function if there were only a few. If there were too many missing values the K-Nearest Neighbour Imputer function could be used to fill them.

Exploring the Data

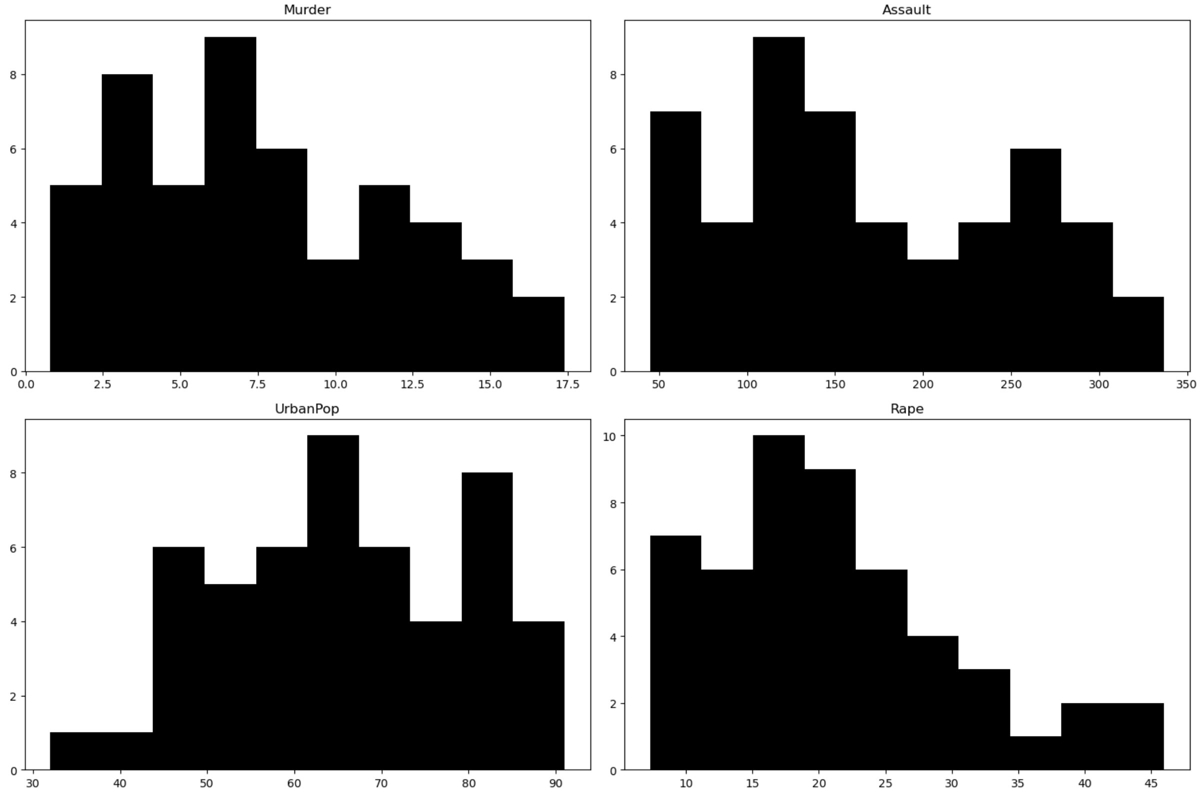
It was first necessary to check the data to make primary observations of the data in regard to which group stood out as an outlier or had any points of interest. The first table was made in such efforts to identify such cases from the of Mean, Standard Deviation, Minimum and Maximum of the groups shown below. As stated before, none of the groups had any missing values accentuating the integrity of the dataset.

Table 1 – Measurements of Mean, Standard Deviation, Minimum and Maximum

	Missing	Mean	Std	Min	Max
Murder	0	7.79	4.36	0.80	17.40
Assault	0	170.76	83.34	45.00	337.00
UrbanPop	0	65.54	14.47	32.00	91.00
Rape	0	21.23	9.37	7.30	46.00

The value that stands out in the dataset from the summary in Table 1 would be the Assault group. The Assault group shows the highest values across the different measurements of mean, standard deviation, minimum value and maximum values. The mean for Assault is almost three times the size of UrbanPop which has the second highest measurements. However, urban pop describes the percentage of the crimes where the population lived in urban areas, so this mean is out of 100, hence, a higher value would be expected. This means that the second highest measurements of relevance would be Rape crimes followed by Murder. Which makes sense in a realistic frame. The Assault groups higher values compared to the rest means that scaling is necessary to prevent the data from being too disproportionate.

Figure 1 – Shape of each Variable

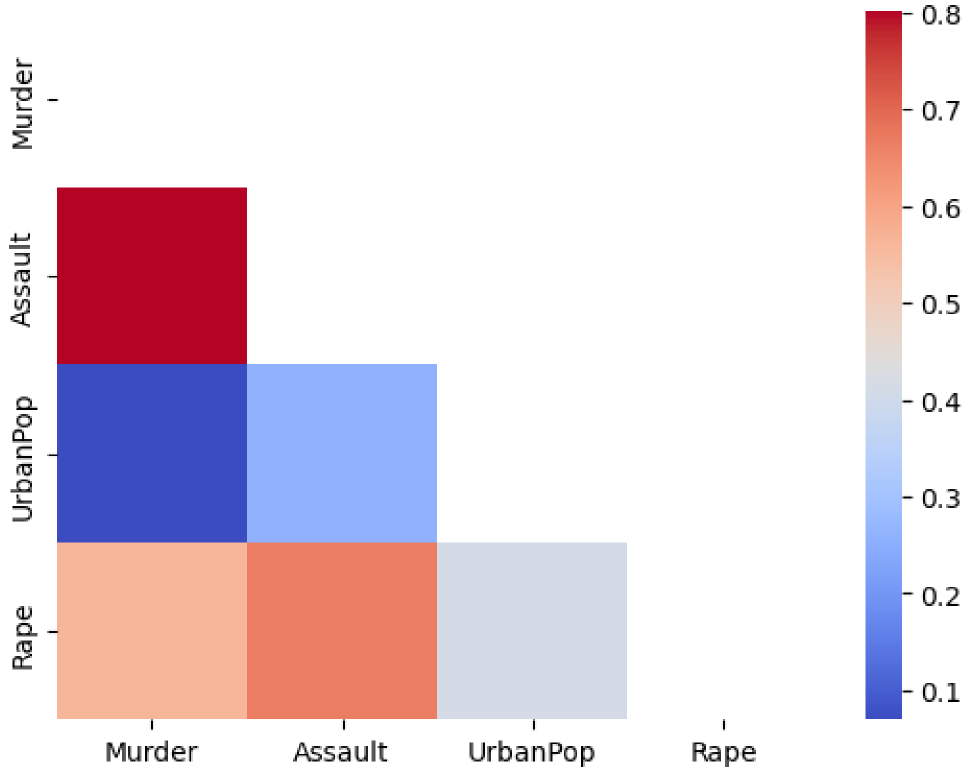


Looking at shape of each variable a similar distribution is evident although not clear due to each of the different maximums. Within each group the distribution of the data covers a wide area within their respective ranges. Amongst these, the distribution for ‘Rape’ shows insightful information as the shape forms descending, meaning ‘Rape’ occurrences decline in certain states the more they increase.

Correlation Analysis

In order to see how different variables relate to one another, a correlation analysis is necessary. This showcases whether one variable has a positive correlation or a negative correlation to another. Using Pandas or Seaborn, it is possible to generate a correlation heatmap using the highly functional corr() method.

Figure 2 – Correlation Heatmap



From the plot, we see a linear correlation between various variables, they are denoted by the many different shades in the heatmap. At first there are not many variables that show a very strong positive or negative correlation indicating that this df does not have many redundancies to declare. The variables that are of a shade of red possess positive correlations as they are near to the value 1. This means as one variable increases the other variable increases proportionally. Whereas there are no variables with a negative correlation i.e. close to -1, meaning as one increases the other decreases and vice versa.

Moreover, more variables appear to have a correlation close to neutral meaning no distinct positive or negative correlation can be identified, indicated by values close to 0. Starting from the distinct red box, this indicates a positive correlation between assault and murder with a value near 0.8. This means that the likelihood of murder rises with assault, which is a statement that proves true in a logical sense.

The dark blue shaded correlation between Urban Populations and Murder is a near neutral one, this indicates that there is no distinct relationship between the two. The light pink shaded correlation between Murder and Rape shows a near positive correlation with a value near 0.6/0.7. This means that the likelihood of Murder increases after Rape. The darker pink shaded correlation between Assault and Rape indicates a similar yet more positive relationship as with Murder and Rape, its value is shown around 0.7/0.8 indicating that many assault cases end with rape.

Assault in Urban Populations is depicted by the light shade of blue, indicating a weak positive correlation, meaning that there is little relationship between the two with no apparent trend. The correlation between Rape cases in Urban populations shows a slightly more positive correlation compared to its counterpart, indicated by the even lighter shade of blue, and a value close to 0.5. This shows that a good proportion of Rape cases have happened in Urban areas.

Indicatively, there are cases that show a highly positive correlation, this makes this dataset suitable for PCA which will be discussed next.

PCA

Simply put, using Principal Component Analysis (PCA) it is possible to reduce variables that share similarities or correlations with other variables, by encoding the data into fewer dimensions. More specifically, it’s a method that finds the principal variables that denote an observation by looking at the direction where the data is spread. This method focalises on identifying the direction with the highest amounts of variance, hence such variables that possess high amounts of variance tend to dominate. PCA effectively reduces this effect.

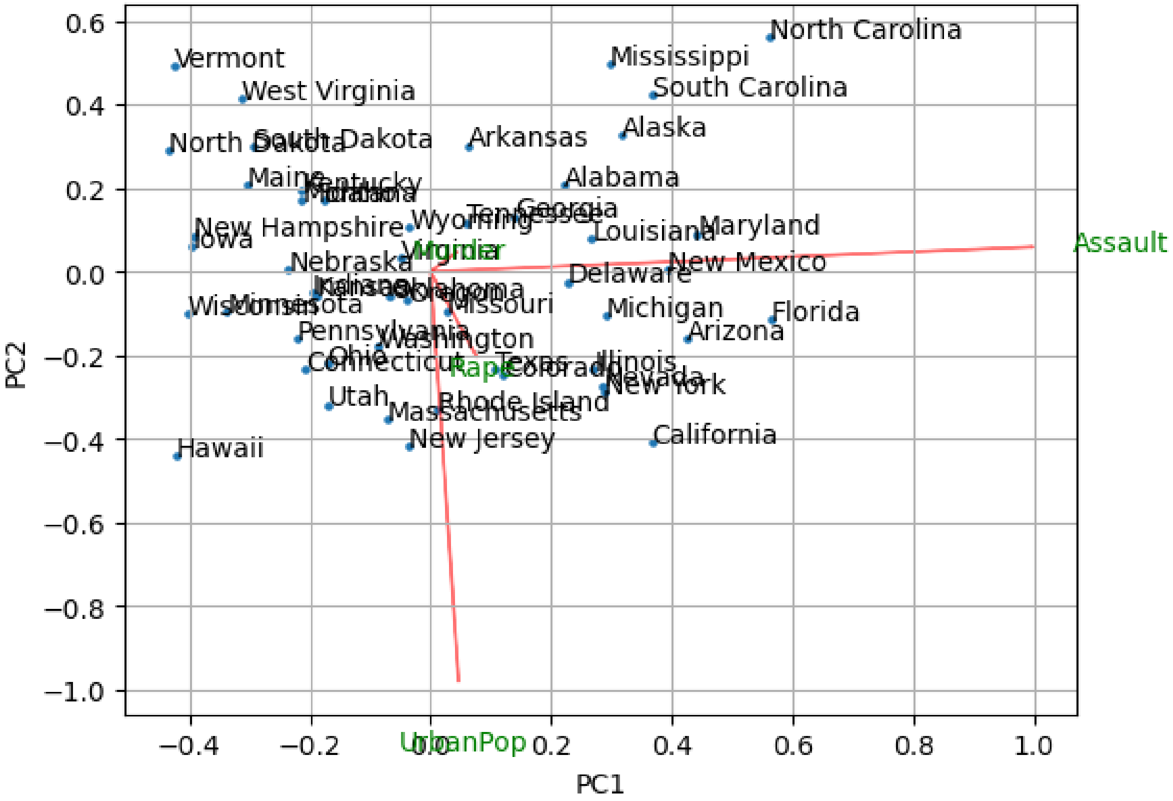
The table below depicts the four principal components in order of their importance, based on the standard deviation and variance as previously stated. It also shows the proportion of variance for each component compared to the total variance, figure 3 below further accentuates this data into more insightful information.

Table 2 – Importance of Components

	PC1	**PC2**	**PC3**	**PC4**
Standard deviation	83.73	14.21	6.49	2.48
Variance Proportion Explained	9.66e-01	2.78e-02	5.80e-03	8.49e-04
Cumulative Proportion	7011.11	7213.11	7255.22	7261.38

Figure 3 below provides evidence that the first principal component is led by the Assault group, as the Assault group lies in the direction that measures the highest amount of variance. Following closely to the assault group is the UrbanPop (Urban Population) group with the second highest reported variance. The findings of this biplot have been consistent with the previous observations mainly accentuating the dominance of the Assault group in this dataset. However, an area of concern is the relative scale of the two leading principal components. The UrbanPop component is a percentage whereas the Assault component is a count, this makes it challenging to differentiate various states from another in regard to where they stand. Indicating, scaling is a necessary requirement for more clarity.

Figure 3 – Plot of principal components

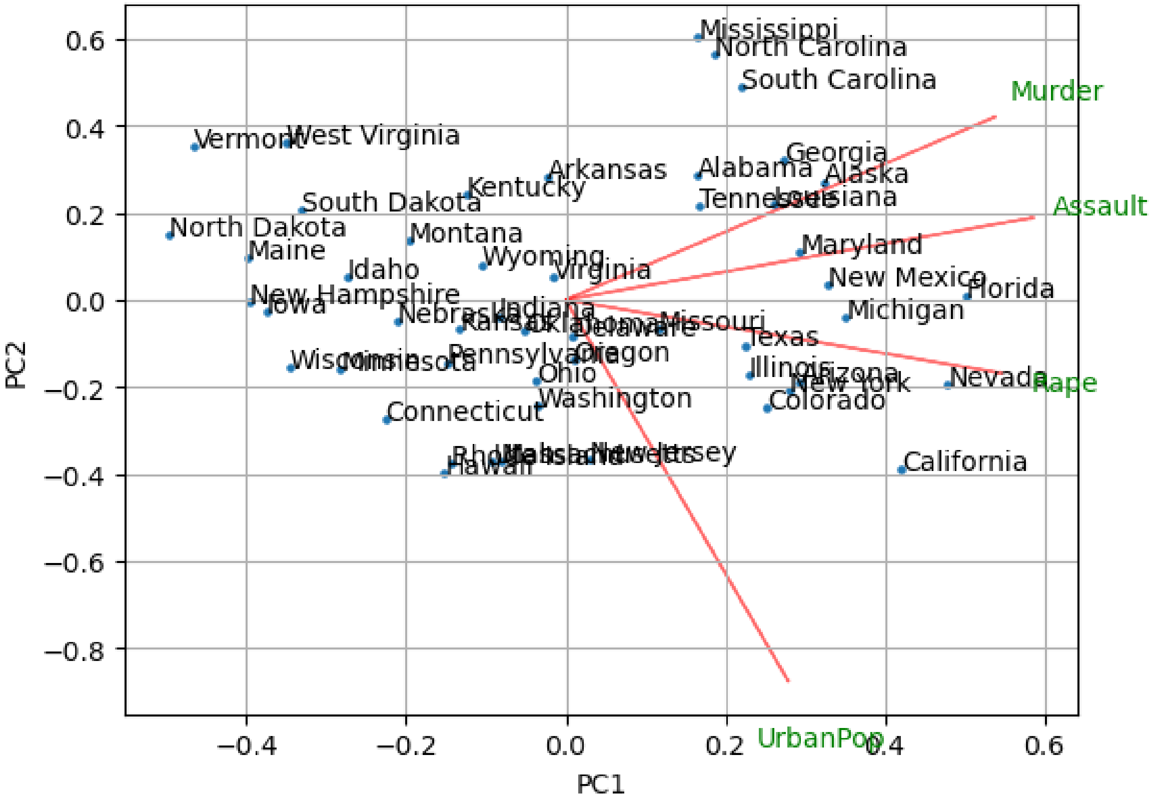


Nonetheless, a few states stand out in this non-scaled data set such as North Carolina, Florida and Maryland amongst some, with Assault cases numbering amongst the highest. Due to the congestion of the data, the exact direction that many states lie in is not clear, more specifically, the states that have a negative or near neutral correlation to PC1 cannot be clearly assorted. The second principal component shows a correlation with Rape from their directions on the plot, although, little variation is shown between the states. In the same manner, some states that show a positive or negative correlation to the PC2 can be identified but its challenging to establish concrete information. However, more scaling is in tow in order to extrapolate additional key information.

PCA standardised data

Upon standardising the data, a clearer image of the plot can be seen. The four components are clearly divided into an axis of their own, moreover, the states that each component has strong correlation with are lined up on their axis, making it easier to distinguish them. Firstly, looking at PC1, we see that the four components are all stretched out in different directions on the right of the plot, this indicates positive correlation between each of them. Along that direction the variables with the highest positive value would be Assault, followed closely by Murder, then Rape and lastly UrbanPop. The states that are situated on the right hand side such as Florida, Nevada, Michigan, California and New Mexico are more prone to Assault, Murder, and Rape crimes compared to the States situated on the left. The states on the left have negative correlations to the first principal component suggesting that these crimes are not as prevalent there. Such states include North Dakota, Maine and Vermont.

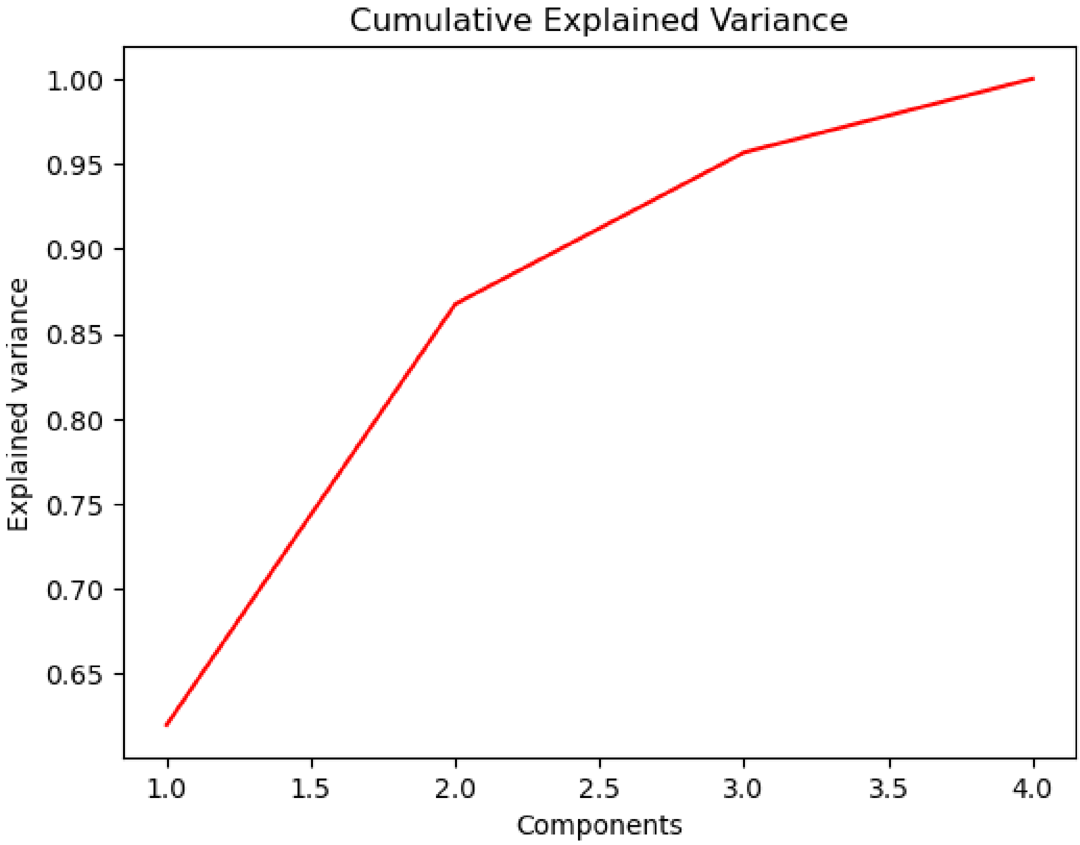
Figure 4 – Scaled Plot of Principal Components



In comparison, the second principal component is dominated by UrbanPop, followed by Murder, Assault and lastly, Rape. Both UrbanPop and Rape variables have negative correlations, the states in these regions consist of California, Rhode Island, New Jersey, and Connecticut amongst some. This indicates that crime in urban areas in these states is significantly lower compared to other states. Murder holds the second highest importance amongst the second principal components, the main states populating this region are Mississippi, North Carolina and South Carolina . Many states lie amongst the centre of the plot between Assault and Rape, meaning no distinct positive or negative correlation could be identified in regard to the variables. The states in this region consist of Iowa, New Hampshire, Nebraska, and Indiana amongst some. Standardising the data has proved to be useful in clearly differentiating the variance of certain variables compared to others, allowing for useful insights to be gained.

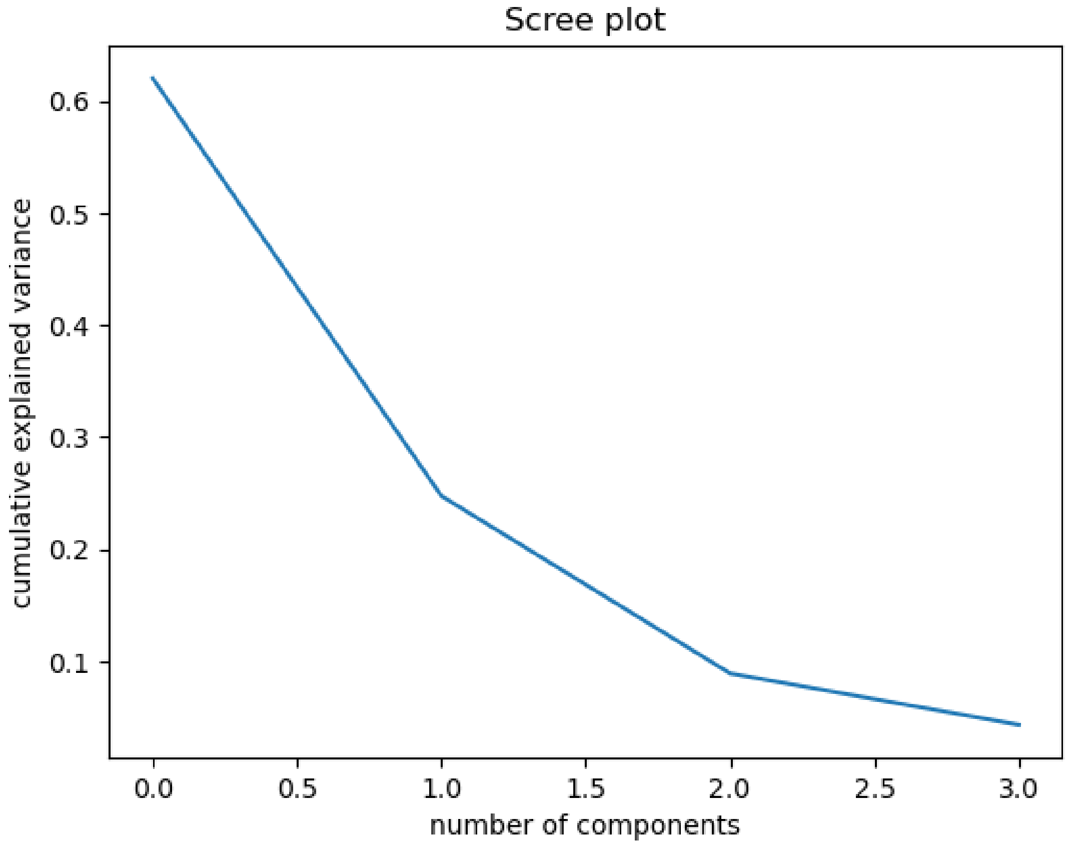
Additionally, when there are many variables in a df, the first few principal components are most useful for explaining variance in the data. Hence why, during PCA it is essential to choose a reasonable quantity of principal components that will explain majority of the variation in the data. Plots such as Scree plots and Cumulative plots are very useful for such endeavours, as shown in figure 5 and 6 below.

Figure 5 - Cumulative Explained Variance Plot



These plots provide clear visualisation of which components attribute for most of the variance in the data. As shown above, the first two variables attribute for 60-80% of the variance whereas the latter two account for 80-100% of the variance. In this manner, dimensionality reduction can be achieved, essentially reducing many variables to few essential components. The same effect can be seen in Figure 6 below in the Scree plot.

Figure 6 – Scree Plot



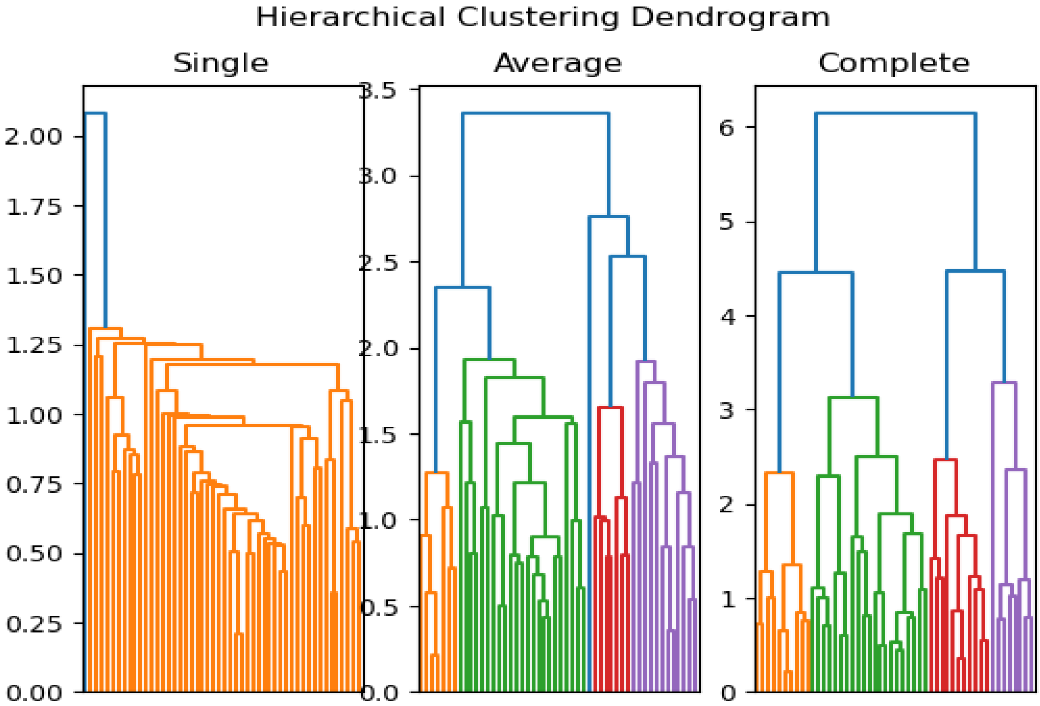
Cluster Analysis

The two clustering techniques that will be used in this dataset are Hierarchical Clustering and K-means.

Hierarchical clustering

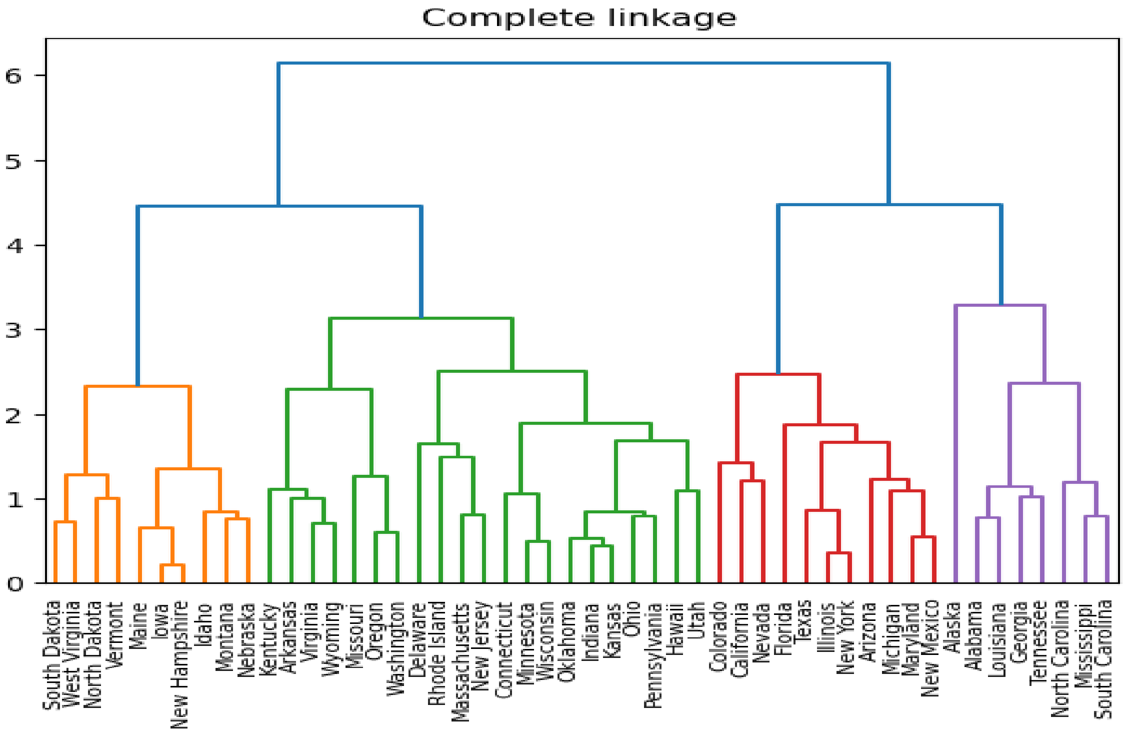
As its name suggests, this clustering method builds a hierarchy of clusters to form a tree like structure known as a ‘dendrogram’. Using this method, the advantage lies in that the clusters are visible from the dendrogram with there being no need to specify how many clusters you want before compiling the algorithm. Although, following the run, its necessary to input how many clusters you want to divide the data into. Distance metrics measure the dissimilarity between each feature in a block and is classified under many types, however, for this analysis Euclidean distance was used. All three linkage criterion, namely, Single, Average and Complete, were used in order to see which measured the distance between each cluster the best. Figure 7 below depicts the plotted dendrograms of the three linkage methods.

Figure 7 – Hierarchical Clustering Dendrogram



From the linkage methods above, it was decided that the complete linkage method provided the most balanced and clearest depiction of the clusters, and was therefore the main method used for the following analysis. Shown in Figure 8 below.

Figure 8 – Complete Linkage Dendrogram



With k=4, the clusters were of size 10, 21, 11, and 8 respectively. Starting from the left-most branch in figure 8 above, the first cluster shaded orange contain the states that possessed a negative correlation to the first principal component and were close to neutral for the second principle component. Most the states that were located in leftmost region of the plot in figure 4 have been grouped together as expected. They are the states that do not experience a large scale of crime compared to the other states. The largest group in green consists of the states that had neither a positive nor negative correlation to the variables, they were mainly neutral, with some possessing slightly higher correlations to the different crimes than others hence, the different clusters within the group. This means that the crime in these regions were all similar in level but not very high, some states were more particular towards some variables than others, for example, Hawaii, Utah, and Ohio were states in which crime did not occur much in their urban populations, whereas, Kentucky, Arkansas and Virginia would have a good amount of crime taking place in urban areas, hence, the division of groups within the cluster.

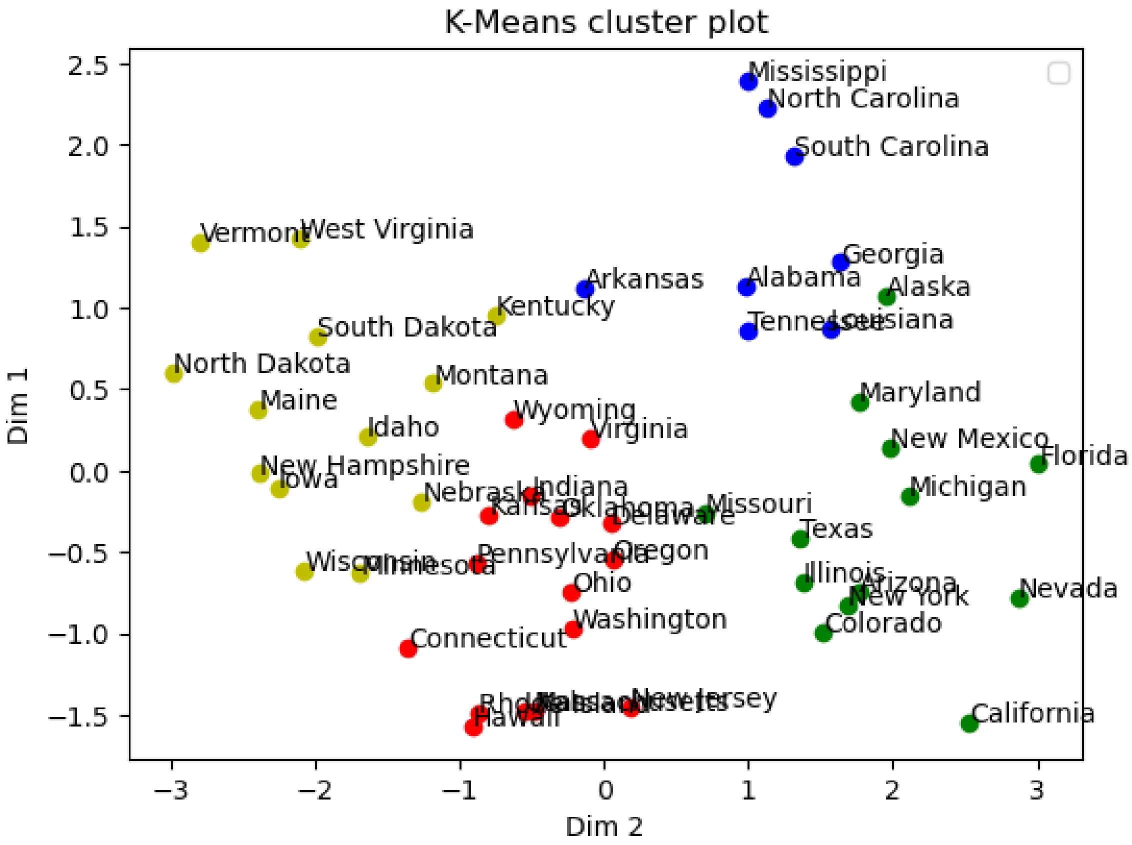
The clusters shaded in red represent the States that were positively correlated to the first principal component and were located on the rightmost region of the plot in Figure 4. Some states within the cluster were more inclined towards a specific variable than others such as California, Colorado, and Nevada, which were strongly correlated to Rape crimes. Whereas Arizona, Michigan, New Mexico and Maryland were more prone towards Assault crimes. The last cluster shaded purple are the states situated on the upper rightmost corner of figure 4. This cluster represents the states that were positively correlated to the second principal component and were more prone to murder crimes. This suggests that states such as Alaska, Mississippi, North Carolina etc are the states in which murder in urban areas would be most common.

K-Means Clustering

K-means is very distinct from hierarchical clustering in that the number of clusters must be specified before running, moreover each observation must be assigned to one K cluster. Nonetheless, this method is still of immense use as it is able to partition the data into specific clusters. The underlying principle for this method is that the smaller the within-cluster variation, the better the clustering, more specifically this is a measure of how much one cluster differs from another. In other words, good clustering means the within-cluster variation is low and bad clustering means that it is high.

With k=4, figure 9 below depicts the K-means cluster plot. Along the x- axis of the plot below, it can be seen that the states are arrayed similar to that of figure 4, in which Assault was the most important principal component. For example we can see the green points at the bottom right hand corner of the plot below consists of the states that were most positively correlated to Assault. Whereas the blue points on the top right hand side are likewise the states that were more prone to cases of murder following assault. The red points represent the states close to neutral i.e., not a lot of Assault, Murder, or Rape related crimes. Whereas the last points in yellow depict the states which had a negative correlation to Assault, meaning there's not much crime in these states. Overall, the shape of the plot followed closely to that of figure 4 and was in tandem with the observations gained from figure 8 above.

Figure 9 – K means cluster plot



Evidently, both Hierarchical and K-Means clustering work in tow to provide clearer confluences on the data, such as how the Assault component was most important in this dataset and most of the variance between states is based on this factor. Essentially, a greater quantity of the states has less Assault, Rape, and Murder related crimes compared to the percentage that do. In each case the plots were divided into two, with the left side consisting of those near neutral states whereas the right side consisted of those states more inclined to those crimes. In other words, this PCA analysis has brought to light where certain states lie in terms of their crime rate regarding Assault, Murder, and Rape.

THIS REPORT WAS WRITTEN BY: Joshua Clegg