Principal Component Analysis Report

Data Set: US Arrests

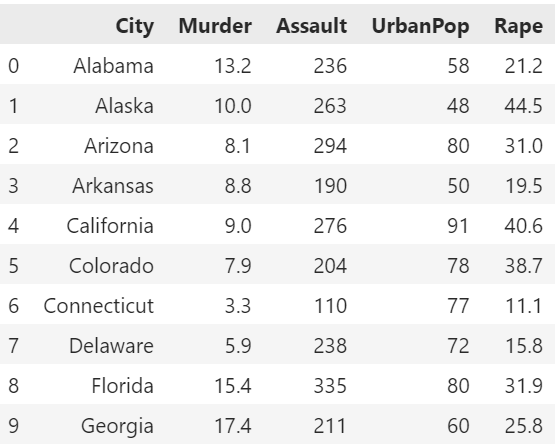
Compiled By: Sadiyya Sheik

# Overview

This report details a PCA analysis that was conducted on the US Arrests data set. This dataset includes information relating to US states. There are 5 variables in the dataset. 1 variable identifies the state using its common name and another indicates the % of the population living in urban areas. 3 other variables indicate the arrests per 100 000 residents for murder, assault and rape.

A sample of the dataset is provided below.

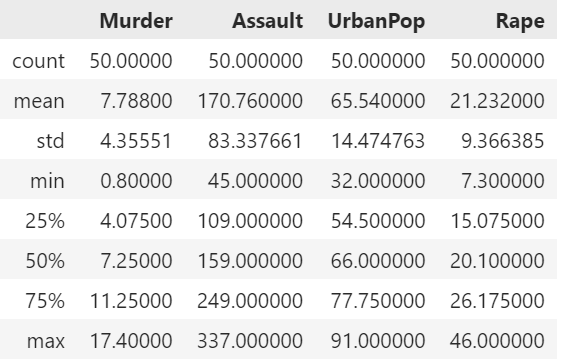
Note that the variable relating to US state was named “City” rather than “State.” This can be confusing and was therefore changed accordingly.



# Exploring the dataset

## Summary Statistics

Summary statistics for the data set are shown in the table below. From this summary, we can see that there are 50 records in each column, this corresponds with the 50 states. We also note that the mean values for the variables vary considerably. This indicates that scaling would be useful.

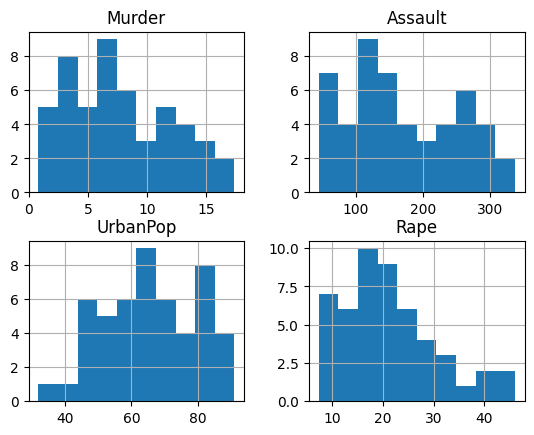


## Missing Data

There are no missing values noted in this data set.

## Visualising Data Spread

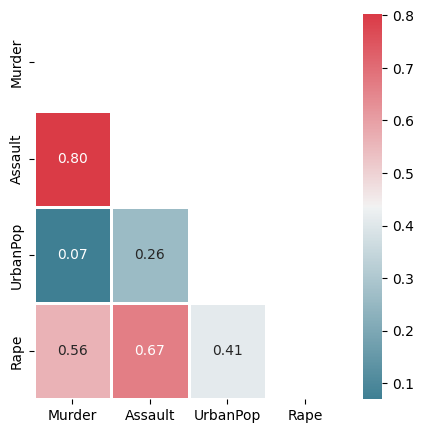
We can tell that most of the urban population plot data sits beyond 50%, this is in keeping with the mean for this variable.



# Principal Component Analysis

## Correlation Analysis

A heatmap of the correlations between the variables was created to view the positive and negative associations between the variables. From the heatmap below, we can see that murder and assault had a high positive correlation. This intuitively makes sense- states with higher murder arrests would be expected to have higher assault arrests - the two crimes are related. To a lesser extent, there was also a positive correlation between assault and rape. Again, it makes sense that states with higher assault arrests would have higher rape arrests, because rape is technically a form of assault. There were no negative correlations identified. The presence of correlation between the variables indicates that principal component analysis could be useful here.



## Performing Principal Component Analysis on scaled data

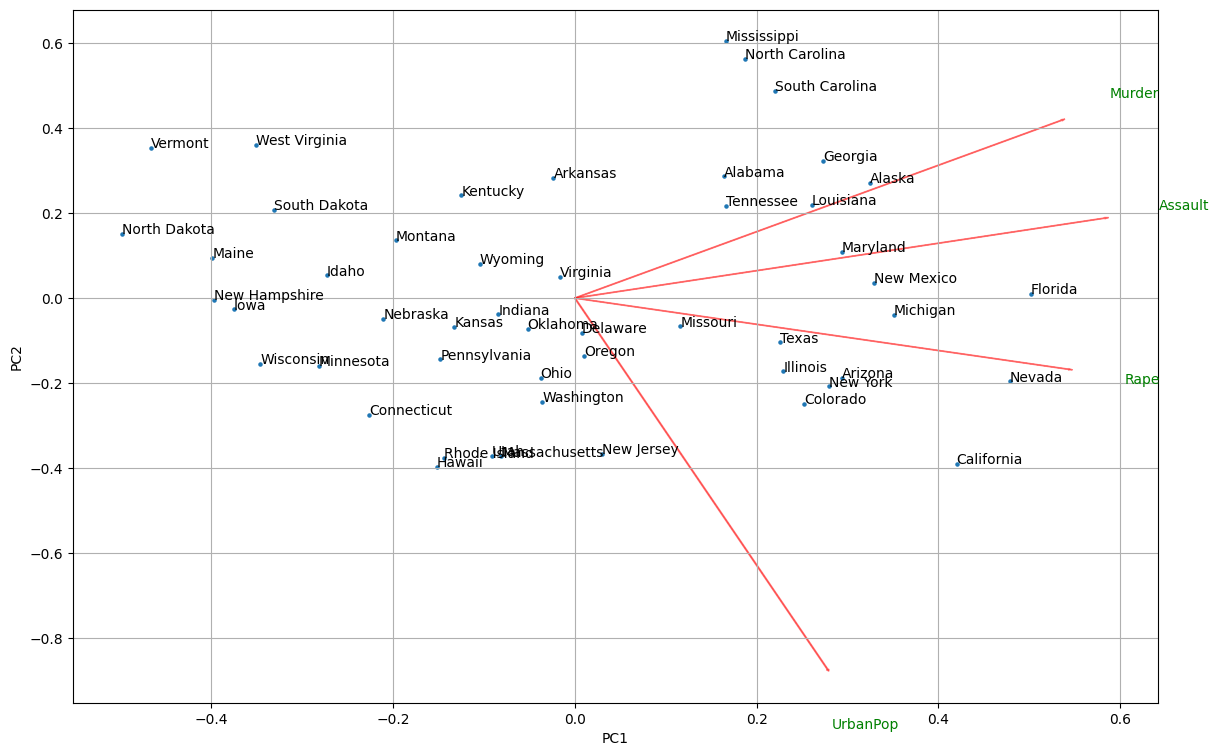
The data was scaled prior to performing PCA. This was done due to the differences we saw in the data as part of the initial exploration.

### Biplot

The biplot below shoes the first principal component on the x axis against the 2nd principal component on the y axis. The original variables are shown by the arrows.

From the biplot we can see that all 4 variables are on the right of the plot. This indicates they have positive loading values for PC1. The murder, assault and rape have higher values than urban population. It is likely that PC 1 relates to crime in general.

Murder and assault have higher PC2 values than the other two variables. Urban population has a negative loading value and therefore PC2 may relate to the “urbanness” of the population. Rape also has a (smaller) negative loading value for PC2 and this may reflect more to the type of crime i.e. sexual assault vs non-sexual assault.



Feature Importance

From the table below, we see the feature importance. We note that murder, assault and rape has similar importance for PC1. Urban population has the highest importance for PC2.   
A picture containing table

Description automatically generated

### Explained Variance and Scree Plot

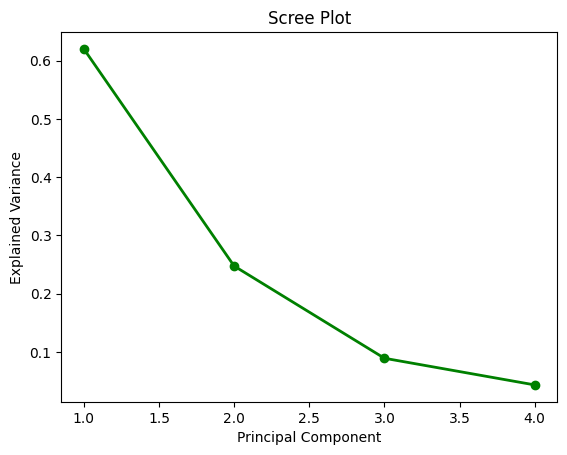
Table

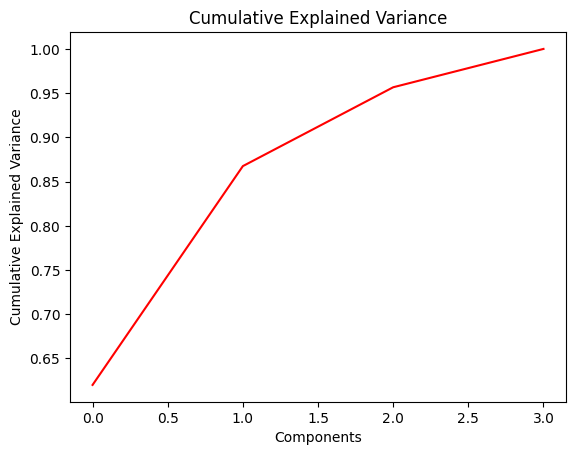
Description automatically generated

From the table above, we can see the variance explained by each of the Principal Components. PC1 explains 62% of the variance, PC2 explains 25%, PC3 explains 9% and PC4 explains 4%. This means that the first three principal components explain 96% of the variance.

If we visualise the Scree plot to determine the n\_components that would be appropriate to use, we can see that the “elbow” of the graph is at 3.

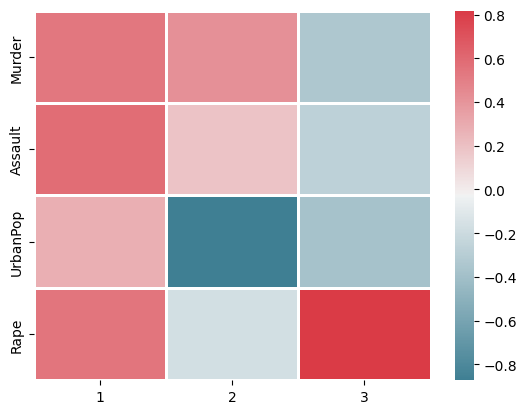
In the Cumulative Explained Variance graph, we can see that the first three components explain close to 100% of the variation.





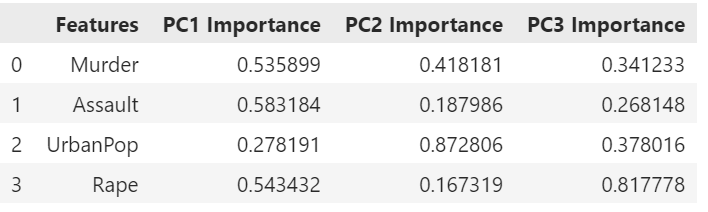
The PCA analysis was then conducted using n\_components = 3.

The heat map for this analysis is shown below. We can see that PC1 has positive correlation with the three types of crimes, PC2 has negative correlation with the urban population and PC3 has a positive correlation with rape.



## Feature Importance- Repeated

After repeating the feature importance and including PC3, we see the above observation confirmed. Rape has the highest importance for PC3.

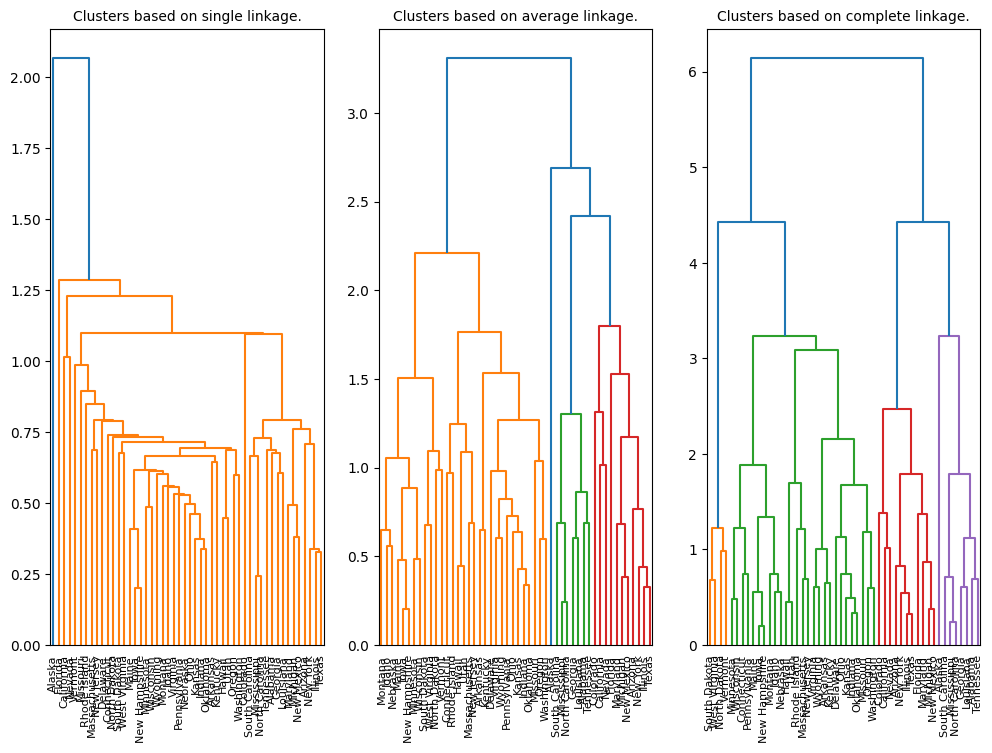


## Cluster Analysis

### Hierarchical Clustering

Hierarchical clustering was performed first because there is no requirement to specify a value for K. This will enable us to find an appropriate K value to use for the K means clustering.

The figure below shows three dendrograms using Euclidean distance and either single, average or complete linkage methods.



In the single linkage method, we can see that there is virtually no clustering that is visualised. In the average linkage method, we see that there are clusters identified but that there are very dissimilar clusters being linked. The complete linkage dendrogram appears to be the most appropriate of the three methods and creates the most balanced dispersion of the clusters. Since there are 4 vertical lines, we see that there are 4 clusters.

The plot below shows the complete linkage clusters more clearly. Here we can see which states are being clustered together: e.g. South Dakota, West Virginia, North Dakota, Vermont are in 1 cluster. If we look back to the biplot we created, we see that these states are grouped closely together in the biplot.

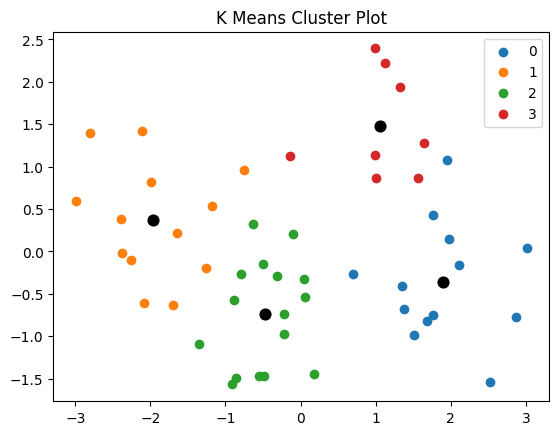
Chart

Description automatically generated

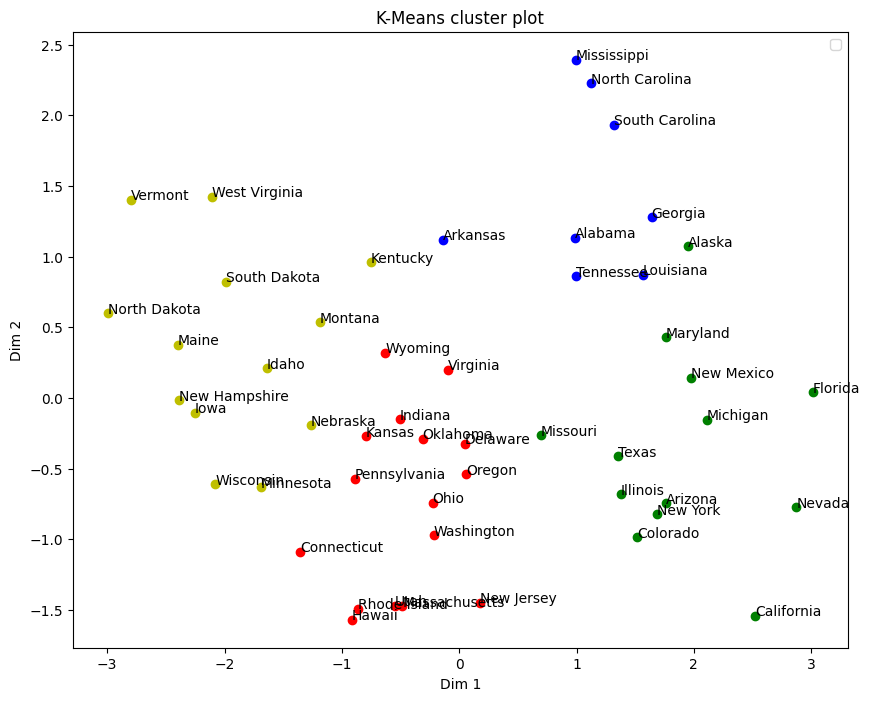
Based on this analysis, a value of K= 4 was used in the next step i.e. K means clustering.

### K means Clustering

K means clustering was performed using K=4. The plot below shows the results of this analysis. Each colour represents a different cluster identified in the data and the black points show the centroids. The K means clustering also determined that there were 4 clusters in the data.



The silhouette score that was determined for this clustering is 0.37. The score ranges between -1 and 1. A score of 1 would mean that the clusters are well defined and a score of -1 would mean that the clusters are very poorly defined. A score close to 0 indicates that there is overlap between the clusters. In this exercise, the score of 0.3 indicates that there is overlap between the clusters – this can be visualised in the plot above which shows for example a point that belongs to the blue cluster but is nearest to a red point. And similarly, a point which belongs to the red cluster but is nearest to an orange point.



The K means plot above shows the clusters together with the labels for each of the data points so that we can visualise which states are clustered together. We can see that as with previously, the states North Dakota, South Dakota, West Virginia and Vermont are clustered together. However, in this clustering we can see that other states are included in the same cluster – e.g. Maine and Idaho. We can also visualise the states that have a neighbouring point that is from another cluster. This indicates overlap between the clusters- as we described when we determined the silhouette score for the K means clustering.

# Conclusion

Principal Component Analysis was useful in the analysis of this dataset as it allowed for dimension reduction. Three principal components explained 96% of the variance in the data. The first component related to crime in general, the second related to “urban-ness” and the third related to rape/sexual assault in particular.

We performed two clustering methods on the data. In agglomerative hierarchical clustering, we determined that the complete linkage method using Euclidean distance derived the most dispersed clusters. We noted four clusters in the data.

We also performed K means clustering and noted the four clusters in the data. We identified commonalities in the clustering via the two methods -e.g. in terms of which states were included in which cluster. The silhouette score for the K means clustering was 0.3- indicating that there was overlap in the clusters. This was visualised on a K means cluster plot.

Interpretation of the four clusters:

The yellow cluster appears to refer to states that have low % urban population and low crime levels. The blue cluster appears to refer to states that have low % urban population and high crime levels. The green cluster appears to refer to states that have high % urban population and high crime levels. The red cluster appears to refer to states that have high % urban population and the crime levels are moderate (i.e. neither high nor low).