Hyperiondev

TASK

# PCA Report on UsArrests Dataset

Visit our website

# Introduction

This report summarises the finding and detail analysis of UsArrests.csv dataset which is available at www.kaggle.com open-sourced using unsupervised learning methods such as Principal Component Analysis (PCA) and various other clustering techniques.

The dataset contains the details of different crimes such as Murder, Assault, Urban Pop and Rape in different cities in USA.

The dataset contains data from 50 different cities and 4 features columns for analysis. The first five rows from the dataset are shown below.

| 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 |

## DATA CLEANING

To summarise the findings from the dataset, following tools are used.

**Programming Language: Python 3**
**Environment: Jupyter Notebook with Google Colab**
**Libraries: Pandas, NumPy**
**Data Visualisation Tools: Matplotlib, Seaborn**

There are no missing values found in the dataset. The summary of missing dataset for each feature column is as follows.

```
Murder      0
Assault     0
UrbanPop    0
Rape        0
```

## EXPLORING THE DATA

To understand the data better, mean, standard deviation, range, and distribution of each variable are calculated and presented in the table below.

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

It is useful to have a visual representation of these numbers so that we can compare these statistics in a glance. The bar plot showing above statistics is shown below.

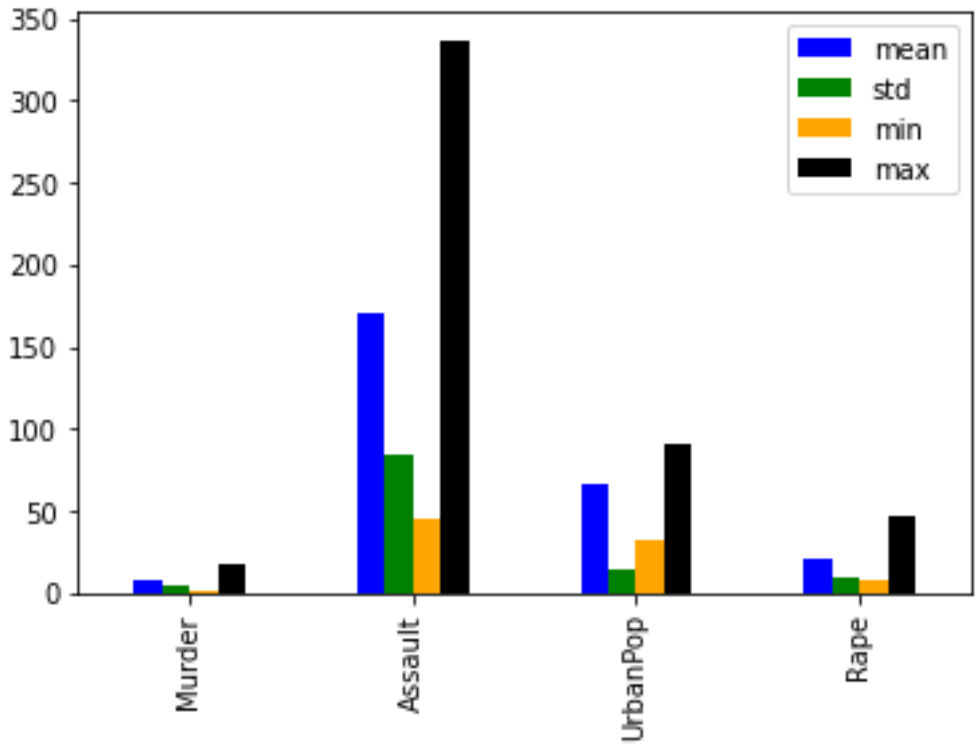



**Fig 1: Bar plot of data statistics**

From the table and graph above, it is clear that Assault is the most committed crime with a mean of 170.76 and std deviation of 88.34 followed by urban Pop with mean of 65.54 and std deviation of 14.47.

Generally, it can be expected that crimes with higher violence tends to decrease as compared to crimes with lower violence hence murder has the least mean value of 7.79 with std deviation of 4.36.

To get better insight of the dataset, histograms for each column are provided below.

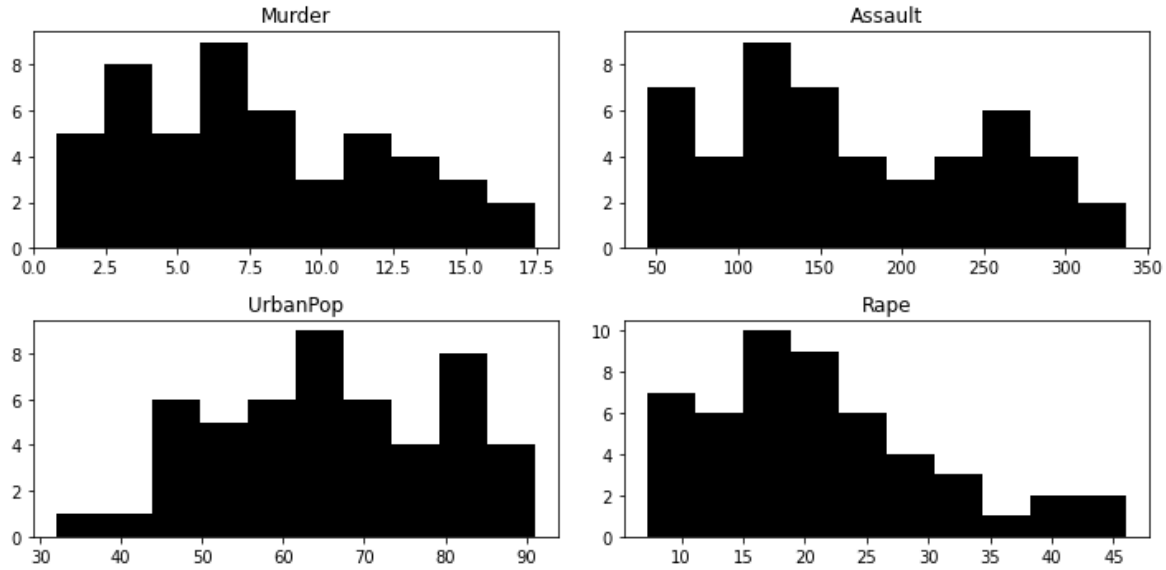



**Fig 2: Histogram of crimes on different cities in USA**

## CORRELATION ANALYSIS

Fig 3 below shows the heatmap of features correlation.

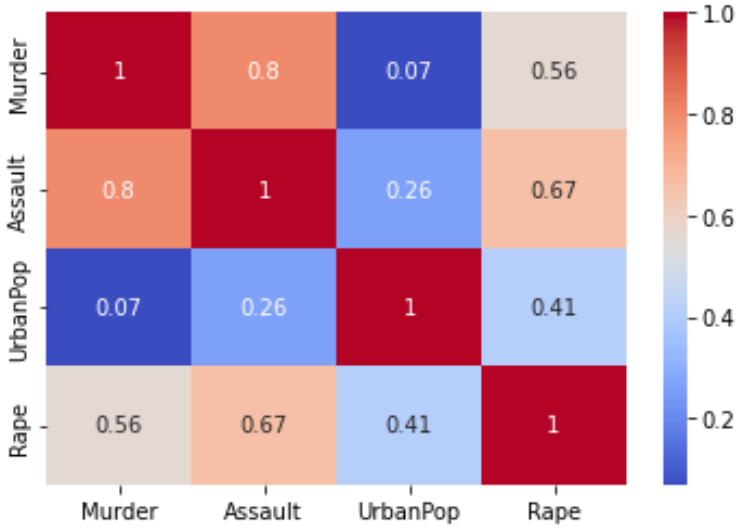



**Fig 3: Heatmap of feature correlation**

From the plot above, most features do not have a high positive or negative correlation with each other except for Assault and Murder and Assault and Rape. There is small positive correlation between Rape and Murder and Rape and Urban Pop as well.

These correlations can be considered intuitive as extreme Assault might result in a murder and Rape is indeed a type of assault. On the other hand, Murder might be a consequence of Rape hence as lower correlation.

The below scatter matrix can provide a broader view of the correlation between each feature.

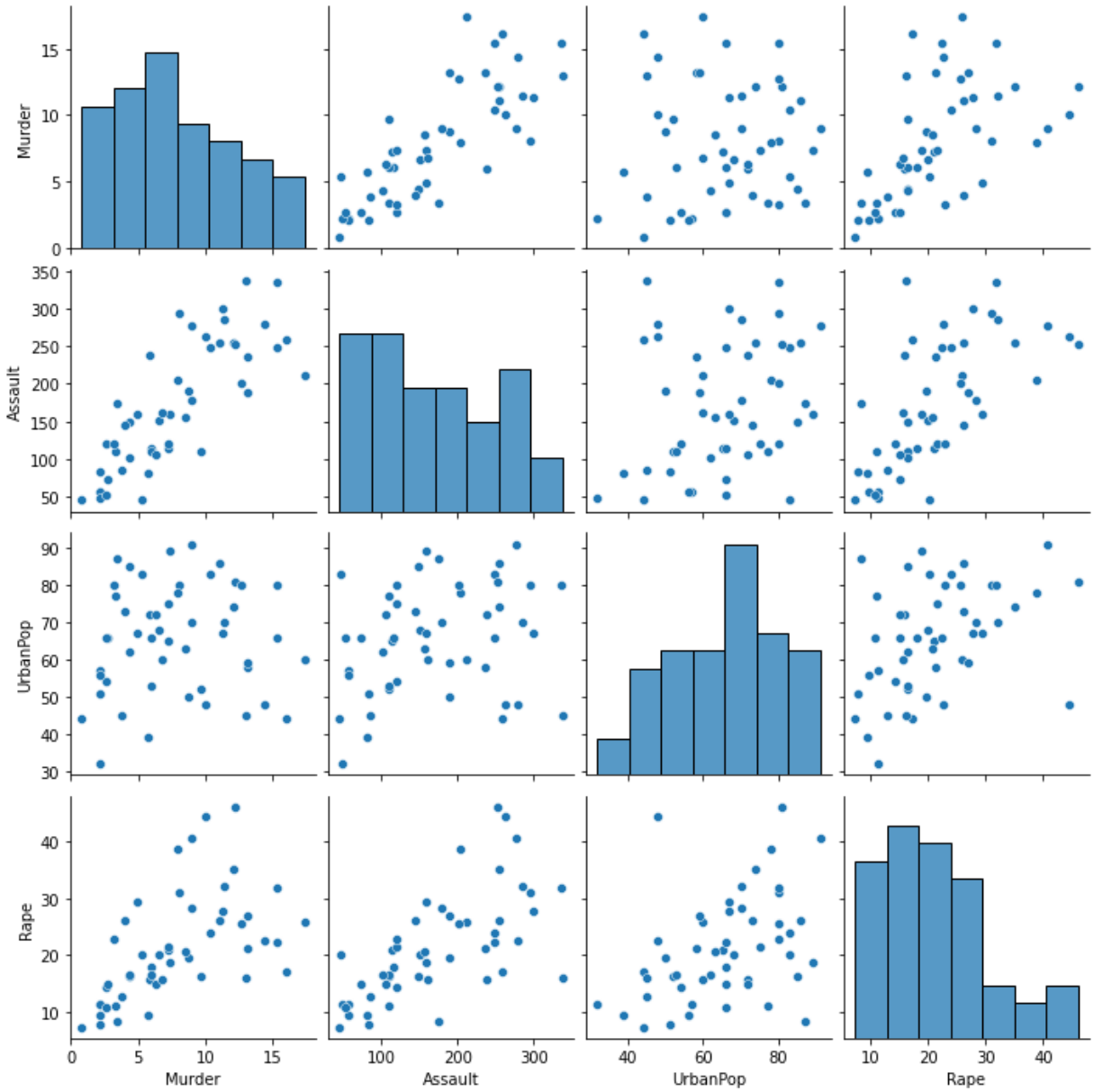



**Fig 4: Scatter Matrix of feature correlation**

## PCA: UNSTANDARDISED DATA

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 our 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.

### Importance of components

The procedure shows the standard deviation associated with each of the 4 components. It also shows the amount of variance that the principal component comprises in comparison to the total variance.

| | PC1 | PC2 | PC2 | PC4 |
|---|---|---|---|---|
| **Standard deviation** | 83.73 | 14.21 | 6.48 | 2.48 |
| **Proportion of Variance Explained** | 9.65e-01 | 2.78e-02 | 5.79e-03 | 8.48e-04 |
| **Cumulative Proportion** | 7011.11 | 7213.10 | 7255.21 | 7261.38 |

The biplot of each component is shown in fig 5 below.

If we consider the biplot for these components, as shown in the data analysis section, the first principal component is dominated by Assault which is on a much larger scale than the other variables (as seen during data exploration). Also, as expected, the second principal component is dominated by Urban Pop followed by Rape and Murder.

Indeed, this makes it difficult to see how cities vary with respect to the other variables, but biplot is reasonability understandable as most cities are spread out all over the biplot.

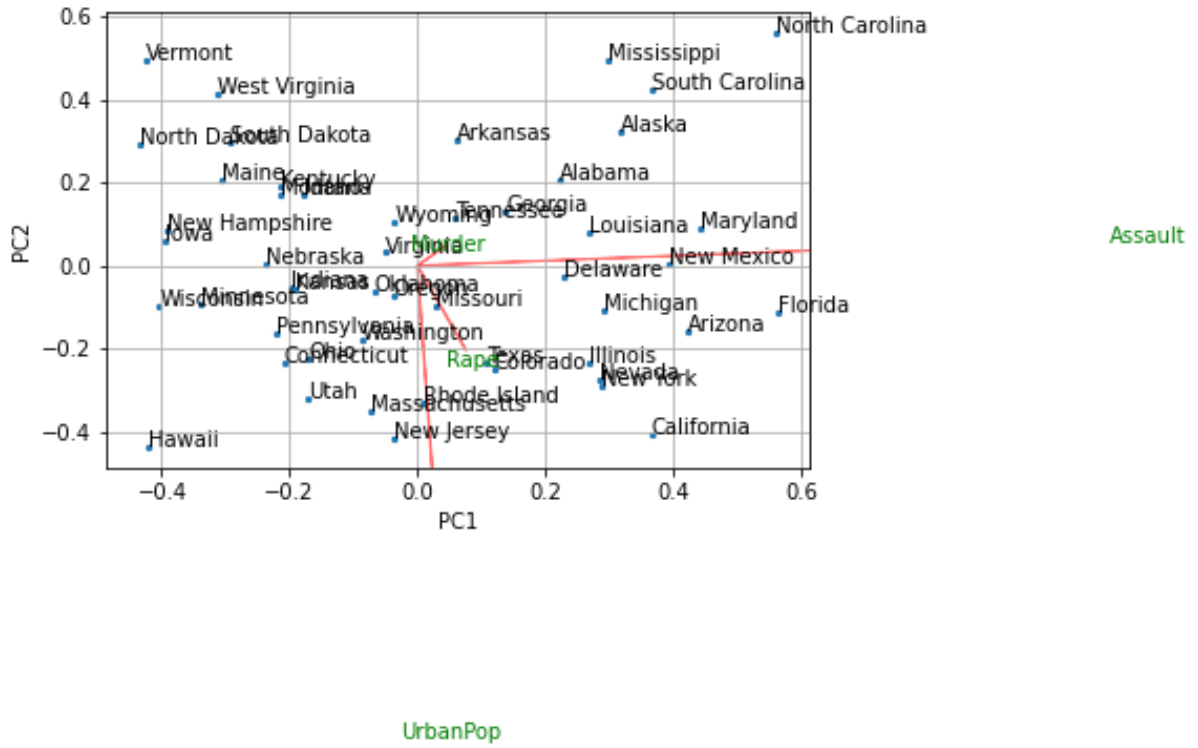
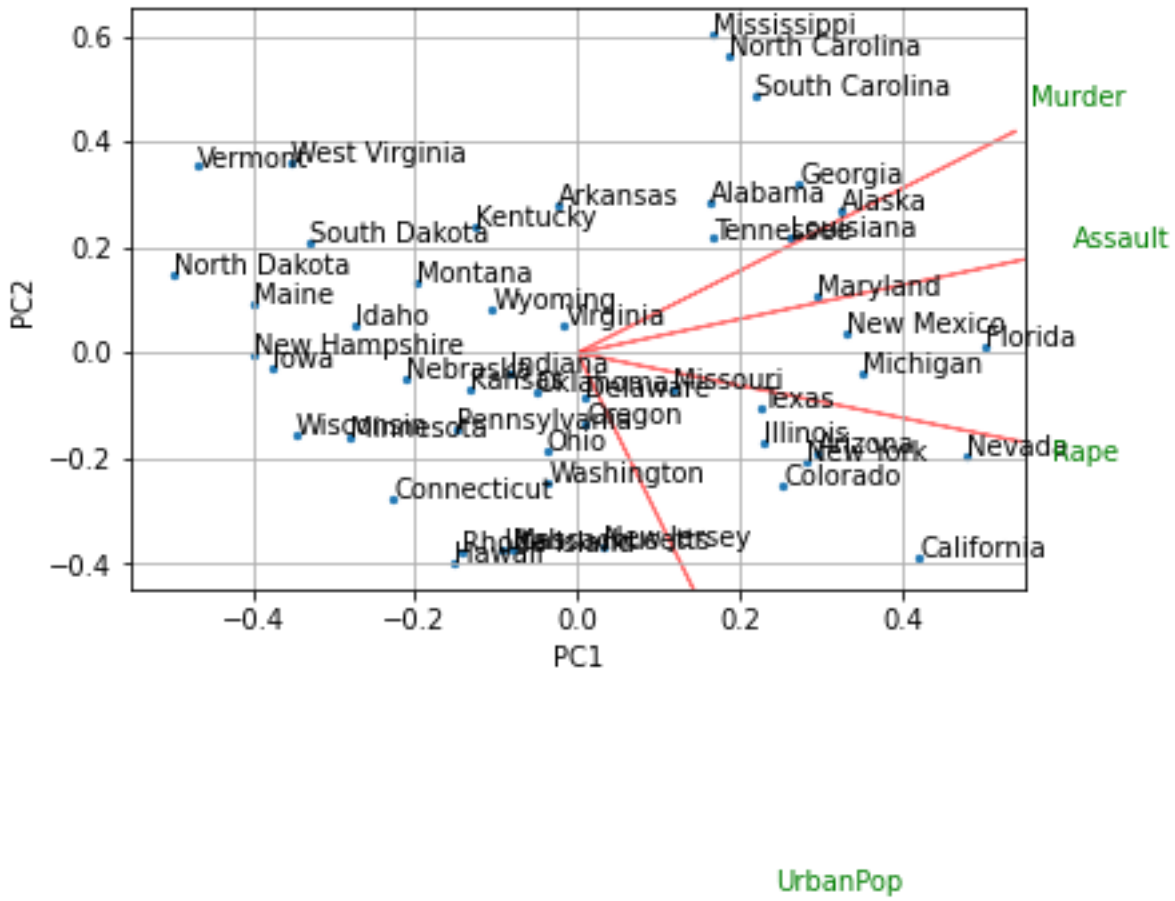



**Fig 5: Biplot for each component**

From the biplot, it can be observed that like the data distribution shown in the descriptive statistics plots, the data assault and Urban Pop are distributed uniformly hence in the biplot as well. Rape, however, is skewed on the right indicating that there are few cities where this crime takes place more often than others which also can be observed from the biplot in fig5.

In order to learn more about the data through PCA, the data was scaled prior to performing PCA. This measure makes it possible to read the biplot (below) more easily, while gaining more insight into possible clusters in the data.

**PCA - Standardised Data**




UrbanPop

**Fig 6: Biplot for each component with standard values**

From fig 6, we can see that almost all variables are being used to explain the variance, on the other hand, there are not many clusters formed in the dataset. By further investigating the feature importance, we can see that for first principal component, all four features contribute equally except Urban Pop and for second principal component, Urban pop still dominate it followed by murder.

The first loading vector places approximately equal weight on Assault, Murder, and Rape, with much less weight on Urban Pop. Hence this component roughly corresponds to a measure of overall rates of serious crimes.

The second loading vector places most of its weight on Urban Pop and much less weight on the other three features. Hence, this component roughly corresponds to the level of urbanization of the state. Overall, we see that the crime-related variables (Murder, Assault, and Rape) are located close to each other, and that the Urban Pop variable is far from the other

three. This indicates that the crime-related variables are correlated with each other—states with high murder rates tend to have high assault and rape rates—and that the Urban Pop variable is less correlated with the other three.

States with large positive scores on the first component, such as California, Nevada and Florida, have high crime rates, while states like North Dakota, with negative scores on the first component, have low crime rates. California also has a high score on the second component, indicating a high level of urbanization, while the opposite is true for states like Mississippi. States close to zero on both components, such as Indiana, have approximately average levels of both crime and urbanization.

In PCA, the first few principal components are the variables that explain most of the variation in the data. As such, 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, below.

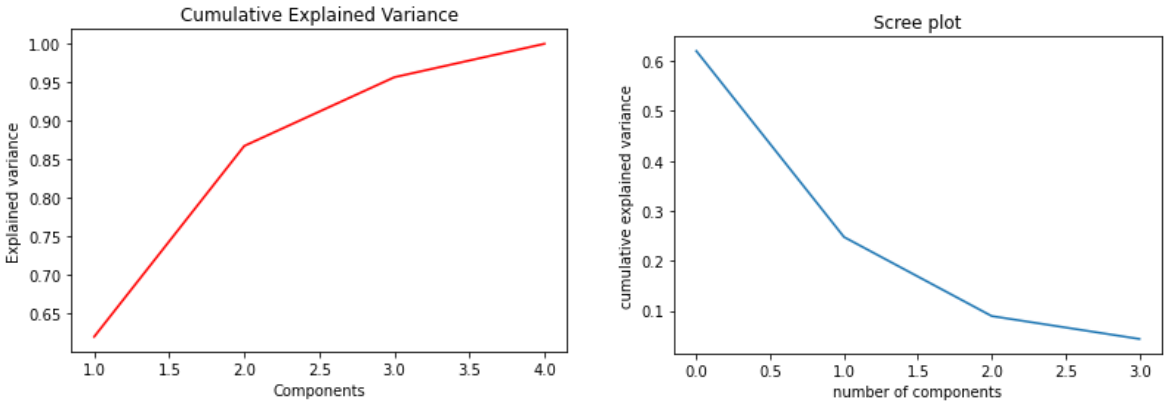



**Fig 8: Cumulative variance and scree plot**

The first 2 principal components together explain around 87% of the variance. We can therefore use them to perform cluster analysis. This is what we refer to as dimensionality reduction. We began with 4 variables and now we have 2 variables explaining most of the variability.

## CLUSTER ANALYSIS

We will perform both Hierarchical Clustering and K-means with these data and compare the results.

### Hierarchical clustering

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. However, we will have to decide the number of clusters after the algorithm runs.

For the distance metric between observations, Euclidean distance was used, which is the most common way to measure distance. In order to determine the method used to measure the distance between clusters, we plotted the various dendrograms for the single, Ward, and average linkage methods.

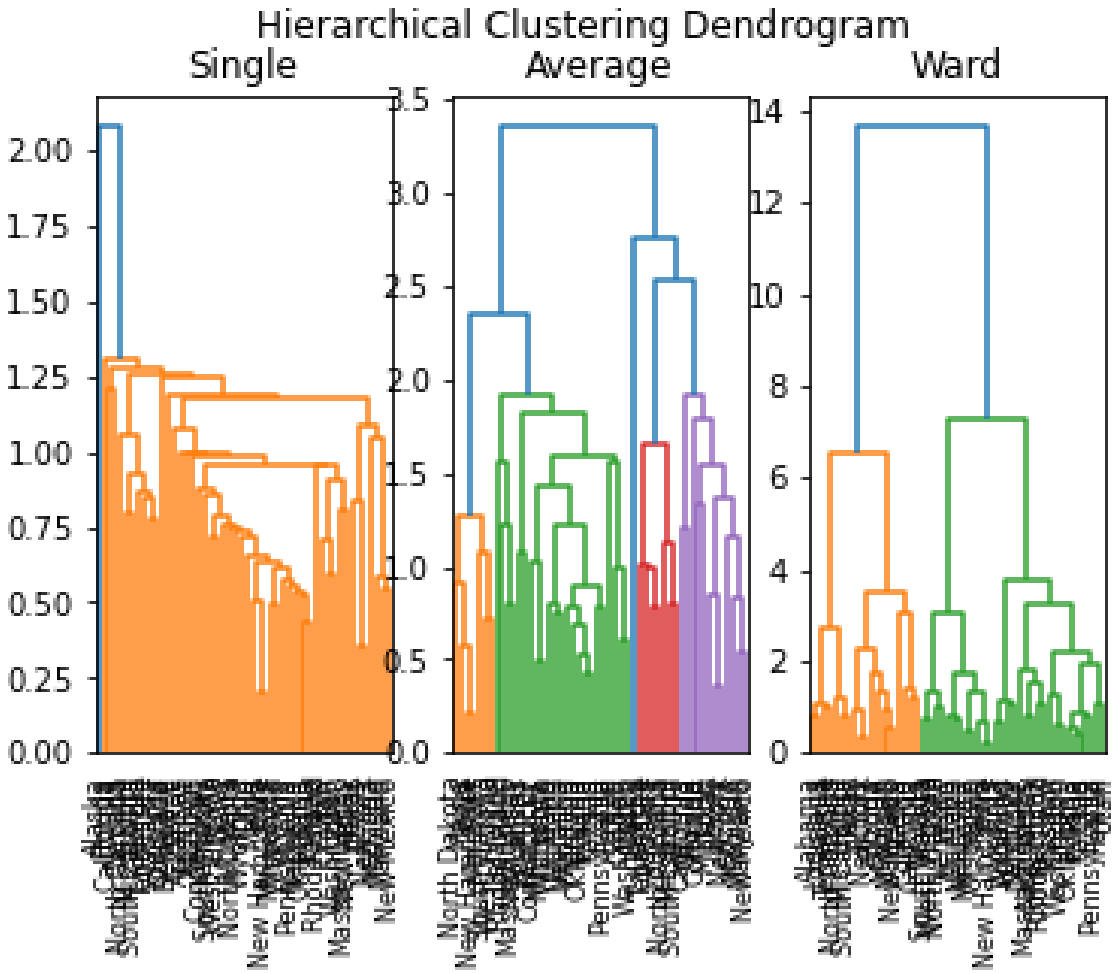



**Fig 9: Clustering Dendrogram**

From the dendrograms above, the ward linkage method creates the most balanced dispersion of clusters and will therefore be the method of choice for the rest of this analysis. A clearer dendrogram for the ward linkage method is shown below.

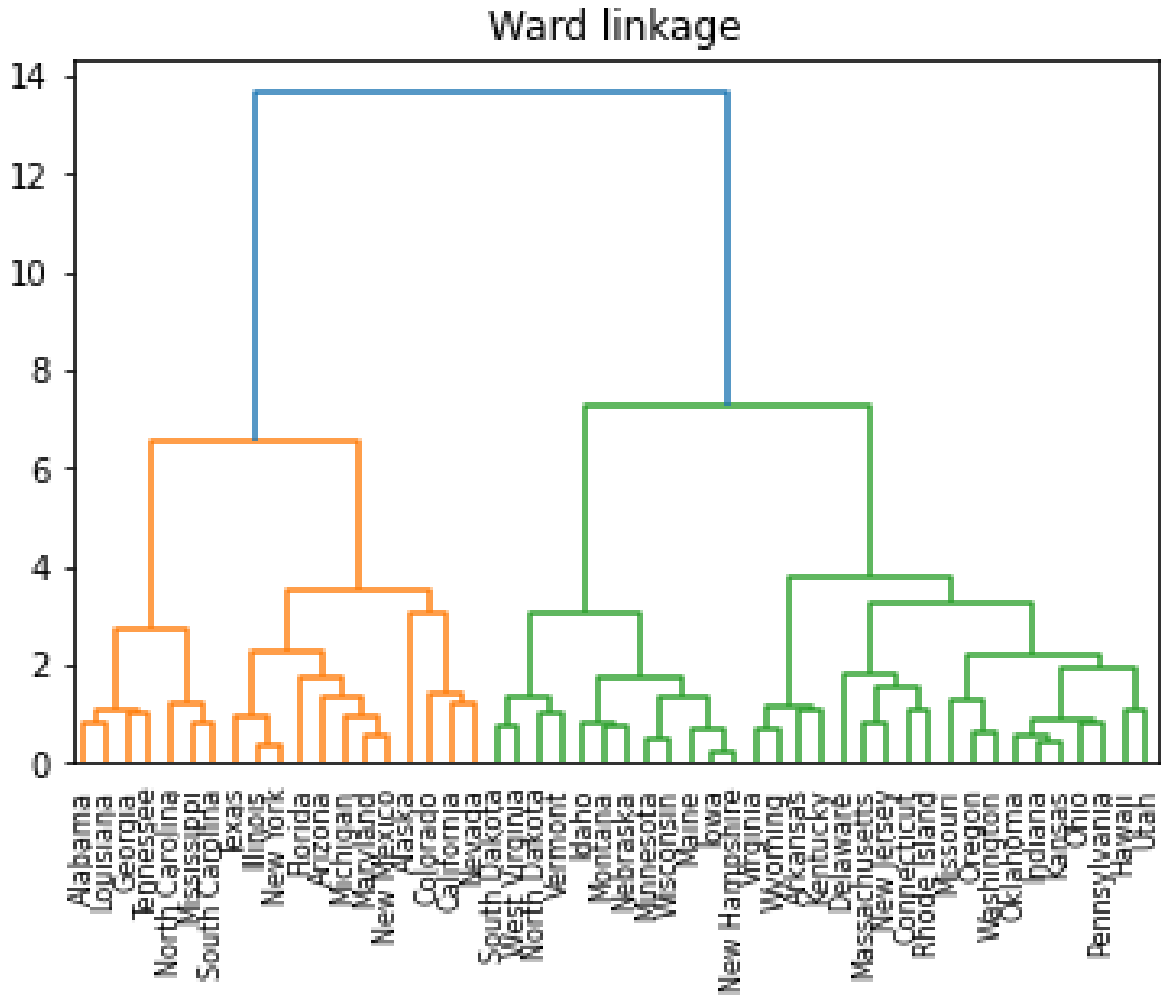



**Fig 10: Ward Linkage Clustering Dendrogram**

With k=2, the clusters are of size 19, and 31 respectively. It would also not be wrong to consider k = 3, but as the data has only 4 features let's continue with k = 2.

### 1. K–means

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 2. A visualisation of the clusters is shown in the figure below.

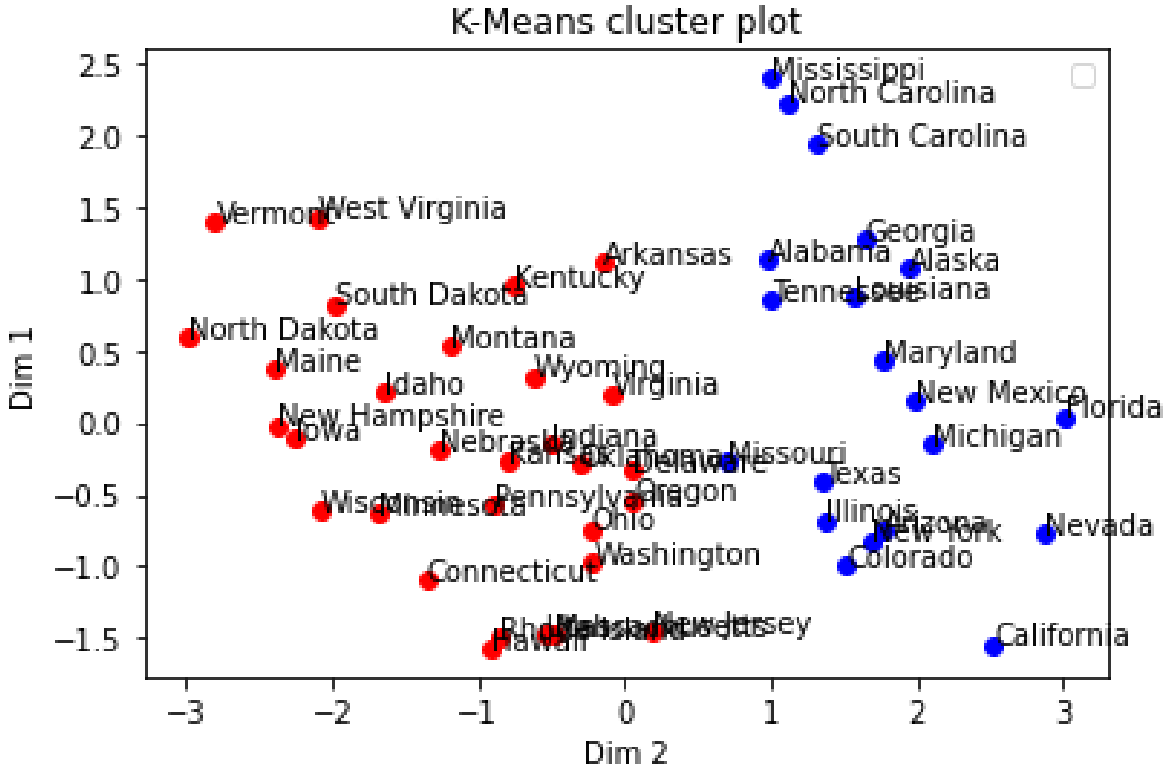



**Fig 11: K-means with K = 2**

Based on the clustering, it seems that K-means has clustered the cities based on number of murders and rape. In the group one, number of murders are quite low as compared to the number of murders in group 2. The number of rapes also shows similar proportion in these two groups. In group one, there are less rape as compared to group 2. In the group two, number of other crimes are also high. We can safely assume that the cities in the group one is safer.

If we were to set K=3, then we can see in the fig below that group one can be considered the safest cities and group 3 cities would be the ones where crimes are high.

The plot with K = 3 is show below.

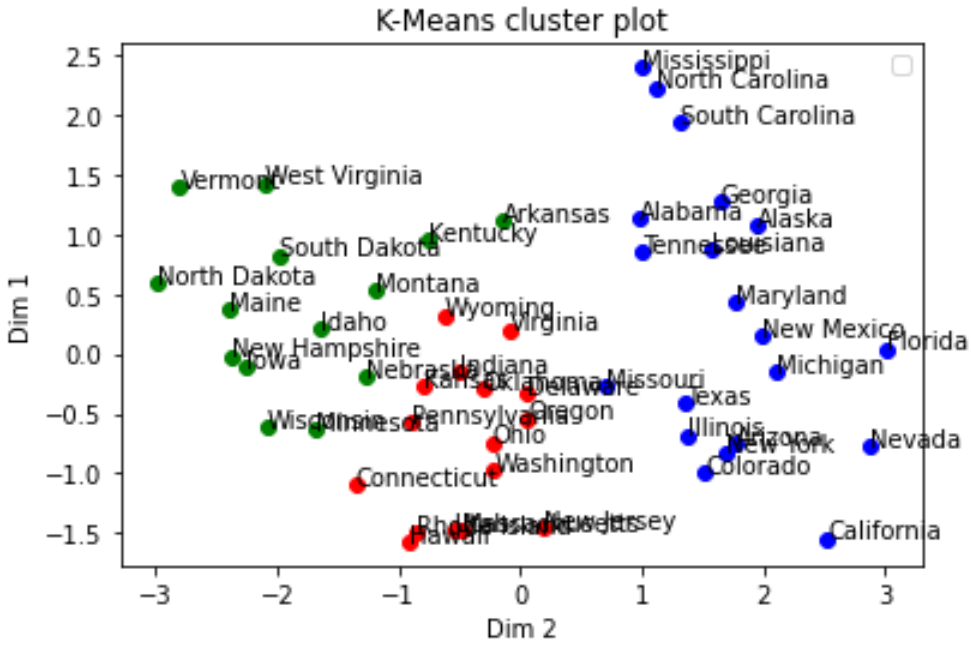



**Fig 11: K-means with K = 3**

## 2. Agglomerative Clustering

Now, we will try to cluster the data with Agglomerative clustering method. Below plot shows the three-clusters formed.

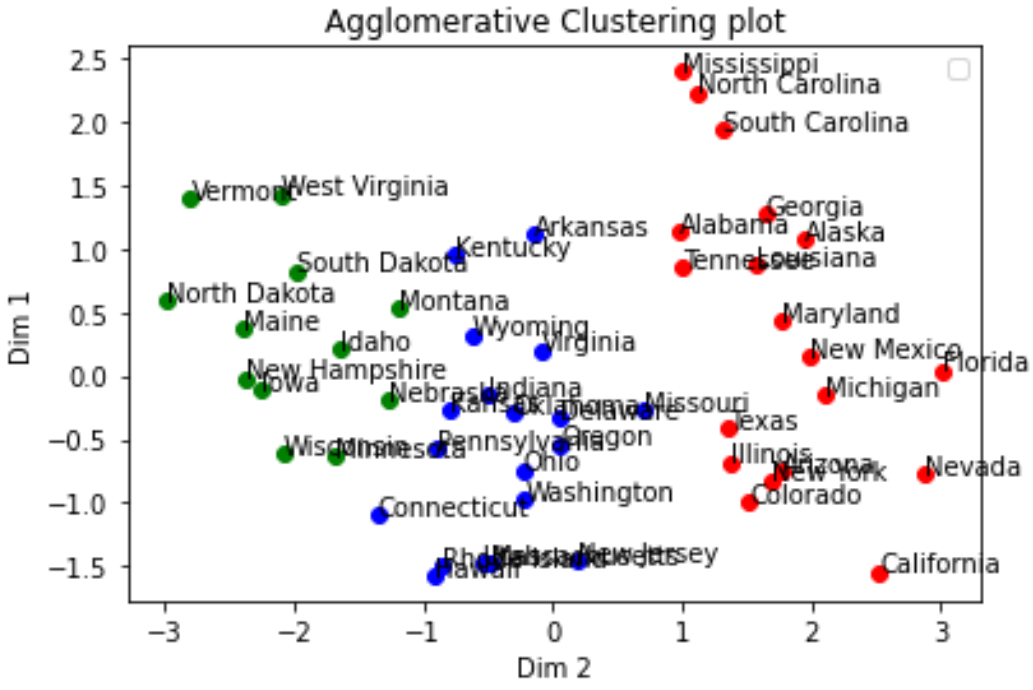



**Fig 12: Agglomerative Clustering plot**

Here also we can see similar results as found by K mean algorithms with same parameters. In fact, the plot is almost identical to the K means with K = 3.

# END OF THE REPORT

**THIS REPORT WAS WRITTEN BY: Shima Maleki**