TASK

Principle Component Analysis on the UsArrests Dataset

Visit our website

Introduction

The dataset was obtained from a Kaggle challenge having statistics describing the arrests per 100 000 residents for the crimes of assault, rape, and murder in 1973 across the 50 US states, with a column describing the number of residents from urban areas in the respective state.

Table 1. The first 5 rows of the data in the UsArrests dataset

	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

The data containing 50 rows and 5 columns

Table 2. Summary statistics of the dataset

	count	mean	std	min	25%	50%	75%	max
Murder	50.0	7.79	4.36	0.8	4.08	7.25	11.25	17.4
Assault	50.0	170.76	83.34	45.0	109.00	159.00	249.00	337.0
UrbanPop	50.0	65.54	14.47	32.0	54.50	66.00	77.75	91.0
Rape	50.0	21.23	9.37	7.3	15.08	20.10	26.17	46.0

The table above is a summary statistics table which give information on the average number of arrests per crime of, Murder, Rape and Assault in respective severity order. Immediately from this table assault has seriously high counts, especially above 75% (3rd quartile).

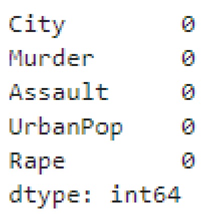
As you can see the Assault column has quite a large mean and standard deviation. This will greatly impact the machine learning algorithm and scaling the data will be required to provide a good fit.

DATA CLEANING

The data had no duplicate cities, indicating that all the data were unique to each city.

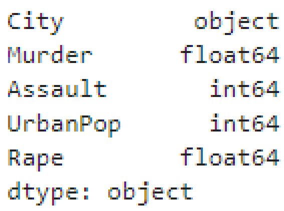
```
array(['Alabama', 'Alaska', 'Arizona', 'Arkansas', 'California',
       'Colorado', 'Connecticut', 'Delaware', 'Florida', 'Georgia',
       'Hawaii', 'Idaho', 'Illinois', 'Indiana', 'Iowa', 'Kansas',
       'Kentucky', 'Louisiana', 'Maine', 'Maryland', 'Massachusetts',
       'Michigan', 'Minnesota', 'Mississippi', 'Missouri', 'Montana',
       'Nebraska', 'Nevada', 'New Hampshire', 'New Jersey', 'New Mexico',
       'New York', 'North Carolina', 'North Dakota', 'Ohio', 'Oklahoma',
       'Oregon', 'Pennsylvania', 'Rhode Island', 'South Carolina',
       'South Dakota', 'Tennessee', 'Texas', 'Utah', 'Vermont',
       'Virginia', 'Washington', 'West Virginia', 'Wisconsin', 'Wyoming'],
      dtype=object)
```

There were no null values in any of the columns, and the data was in the correct datatypes:



```
City        0
Murder      0
Assault     0
UrbanPop    0
Rape        0
dtype: int64
```

Figure 1. Null Data



```
City         object
Murder      float64
Assault       int64
UrbanPop      int64
Rape        float64
dtype: object
```

Figure 2. Data Types

MISSING DATA

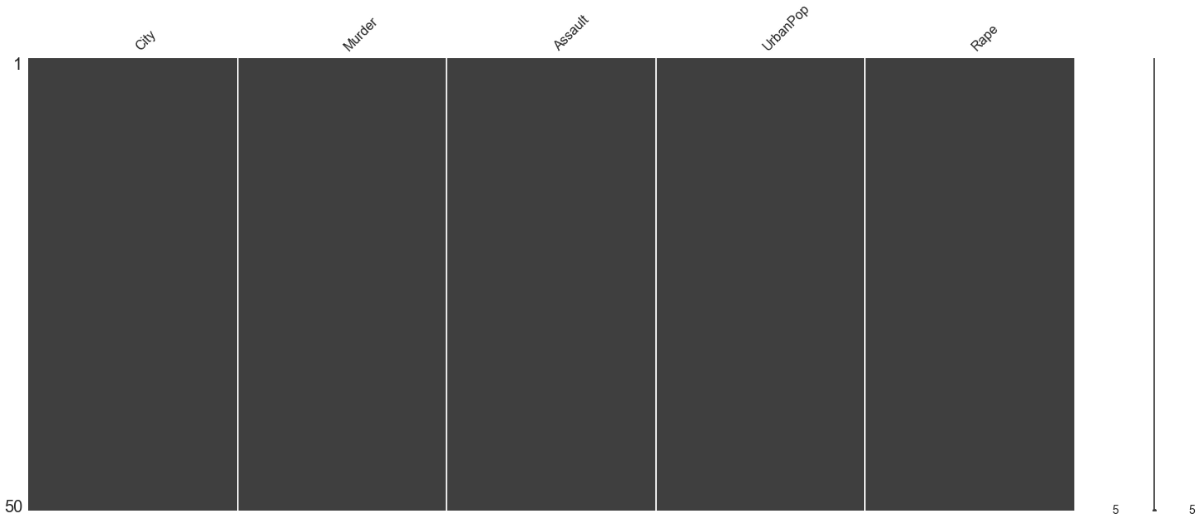


Figure 3: Missing data matrix graph

Used missingno library to produce the above graph to visualize any missing data. There are no white spaces which indicate that all the rows had data in them.

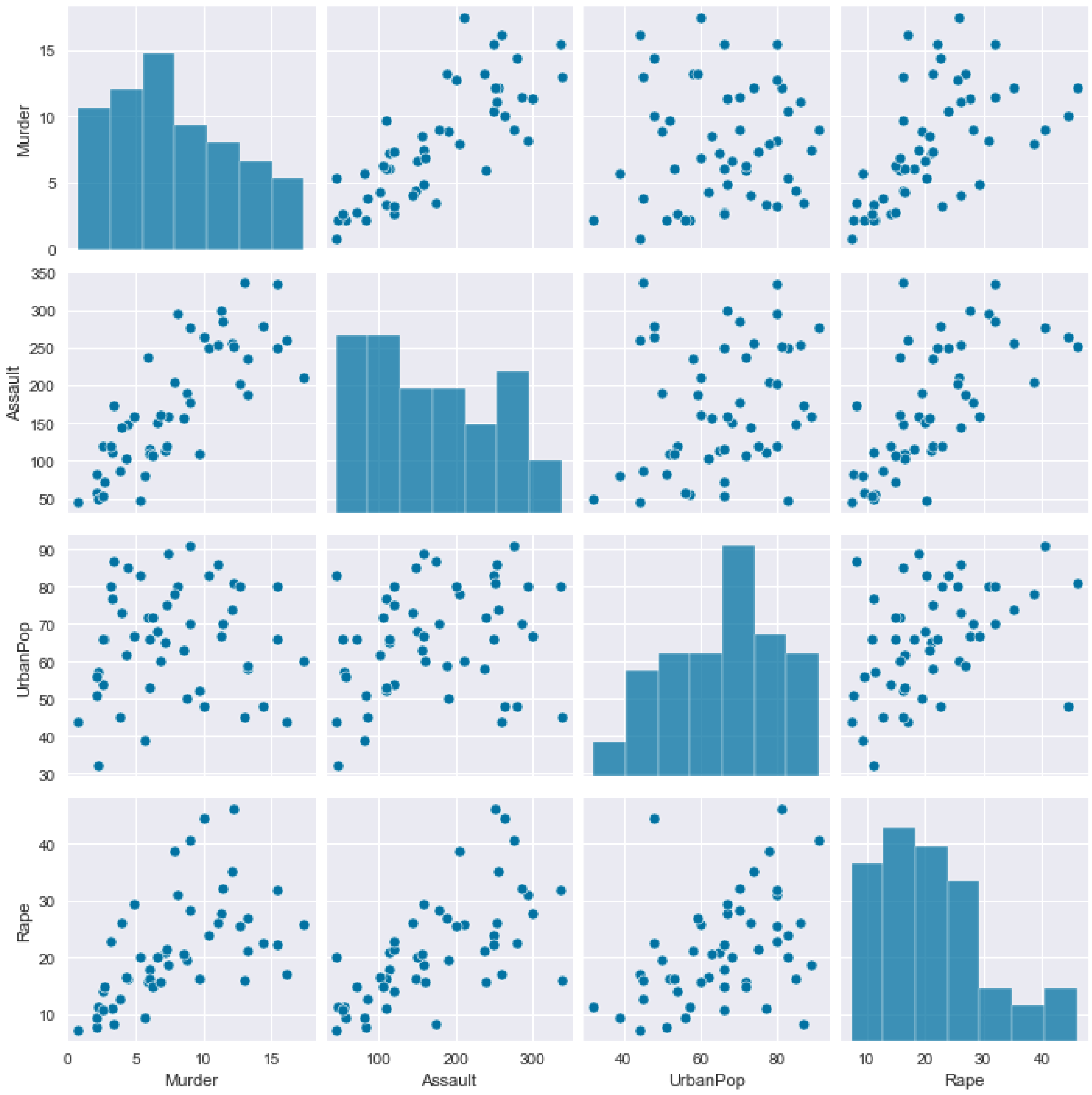


Figure 4. Pair plot of all the data columns

The pair plot gives an in-depth glance of the distribution of the data, and it provides a good way of quickly scanning the data. We can see that Murder and UrbanPop distribute normally. We can also see that the general trend here is that as the UrbanPop increase so do the crime, especially in areas of Rape and Assault, with assault indicating a high trend of Rape associated.

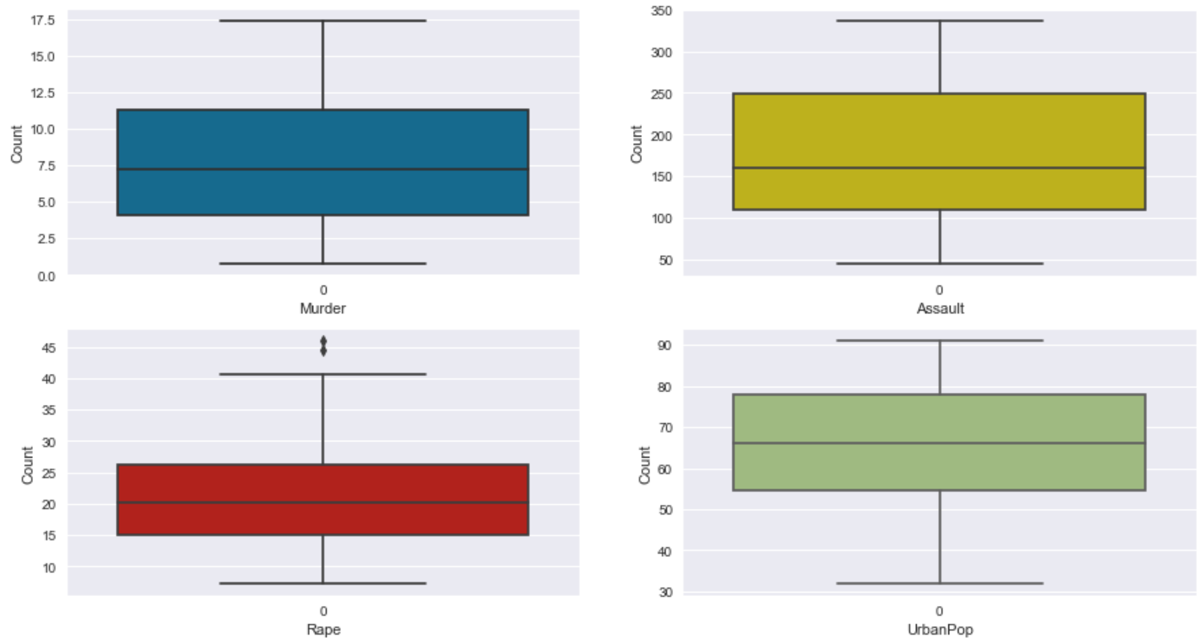


Figure 5. Boxplot of the Dataset

The box plots display assault with a high upper quartile range, indicating that most states have a very high count of assault. We also notice that rape has a few outliers indicating that even though most states have a low count of rape, there are a few states which have counts way above the normal range. To have a better understanding of how each of these crimes measure in respect to the counts of each other a box plot of the same y-axis is used.

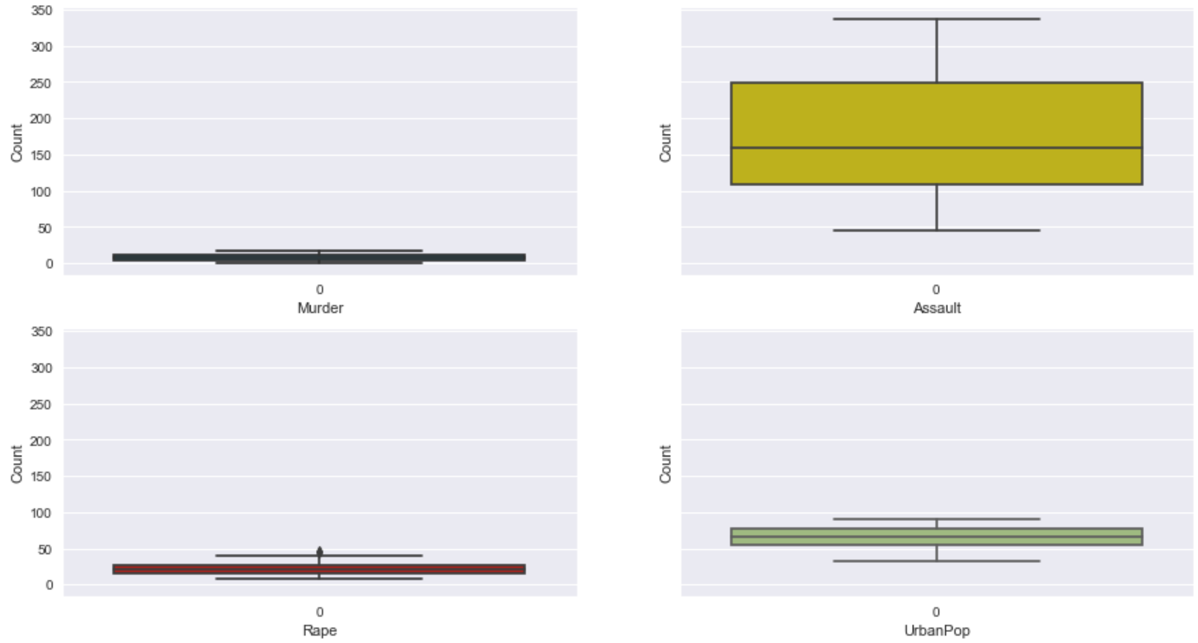


Figure 6. Box plot with the same y-axis used

The boxplot at figure 6, indicates that the lowest crime at that time was murder, as you can barely see a defined plot well below the 50 range. Assault was very common in the states which is indicative by such high counts. People living in urban areas is also very low, and the possibility that poverty was quite high may be a contributing factor to the overwhelming counts of assault. This doesn't indicate a corrupt government as this would have indicated very high counts of rape and murder as well.

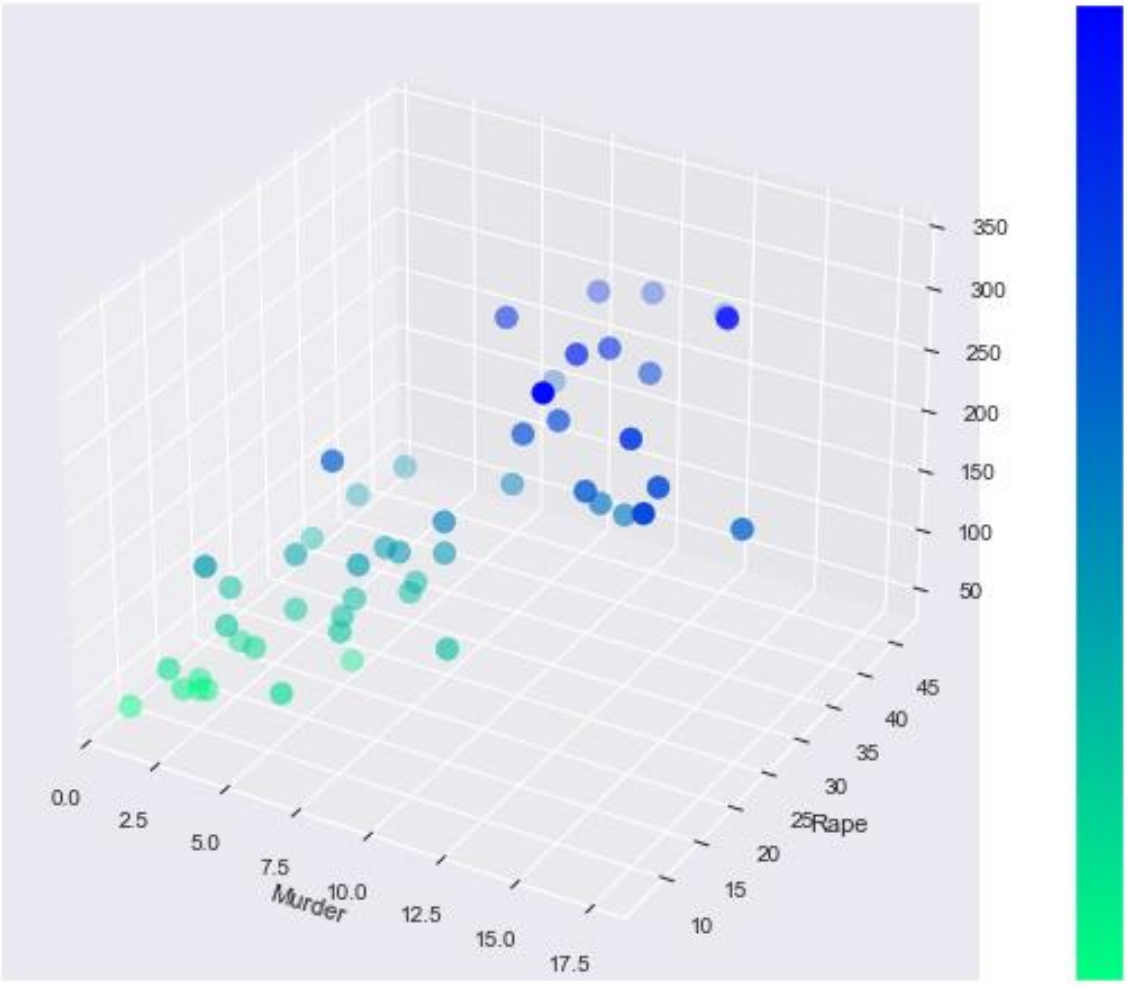


Figure 7. 3D scatter plot of Murder vs Rape vs Assault

The 3D scatter plot is a very nice way of visualizing the arrests. You can see a trend where assault (z-axis) has high counts in relation to murder, and this would indicate that assault is highly likely in a murder arrest than rape, however very few points indicate a positive relationship between rape and murder.

PRINCIPAL COMPONENT ANALYSIS

To find out more about the relationships between the arrests, we do a correlation heatmap.

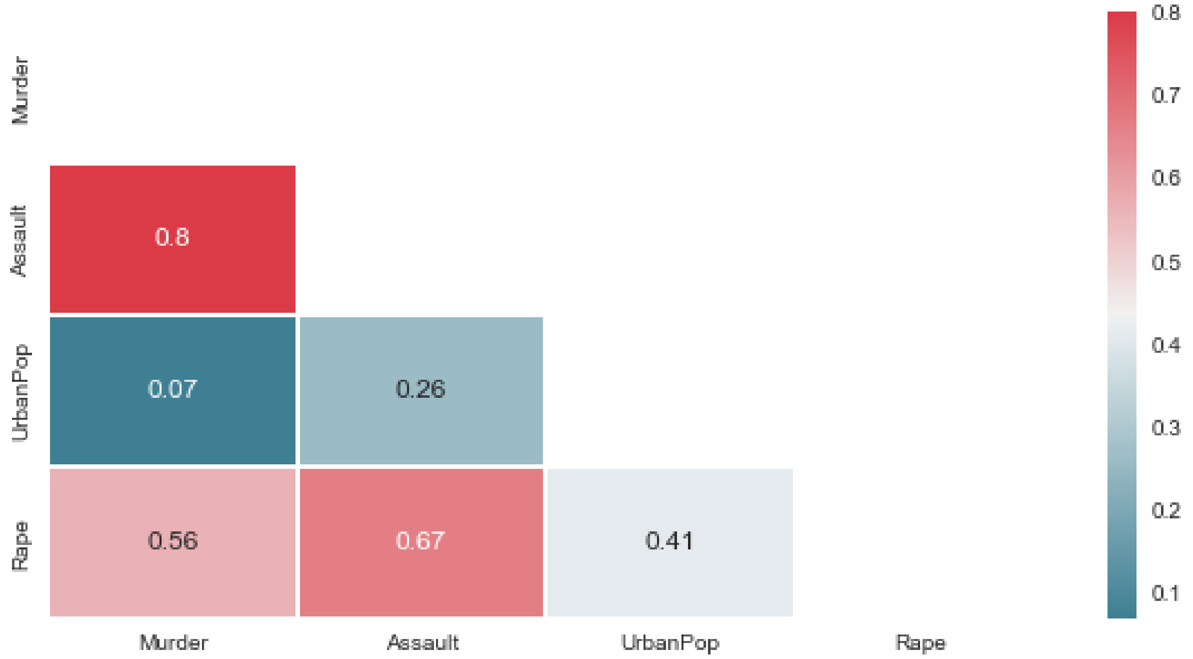


Figure 8. Correlation heatmap

The heatmap provides striking reason to believe that assault is highly associated with murder and rape and strongest positive correlation with murder. And murder being lower in a case of rape associated with murder.

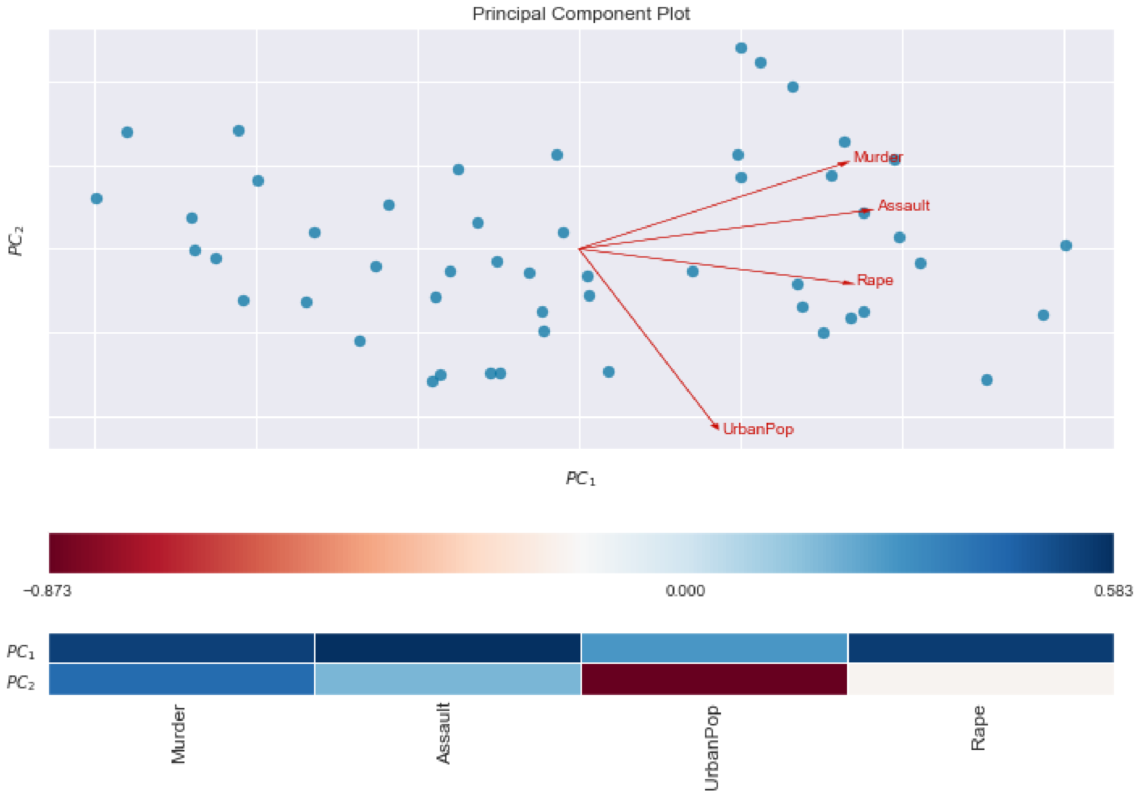


Figure 9. Principal Component Biplot

The first principal component (PC_1) seems to spread the data in a single direction. Indicating high positive correlation between the arrests. The urban population component is positive loading but further away from these arrests.

To investigate how many components to conduct PCA with, a scree plot is generated to find how many components describe at least 90% of the data.

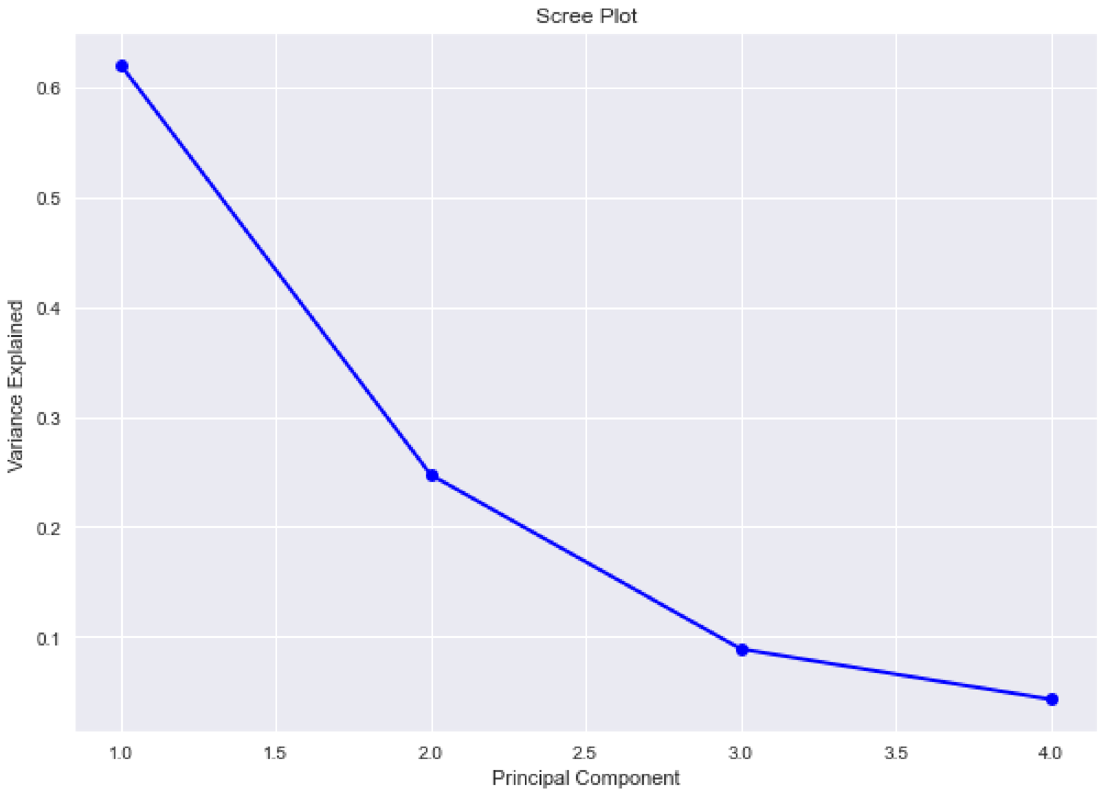


Figure 10. Scree plot relationship pf the data

The above plot indicates that the first 3 components describe 90% of the variance and we can use that in the PCA analysis.

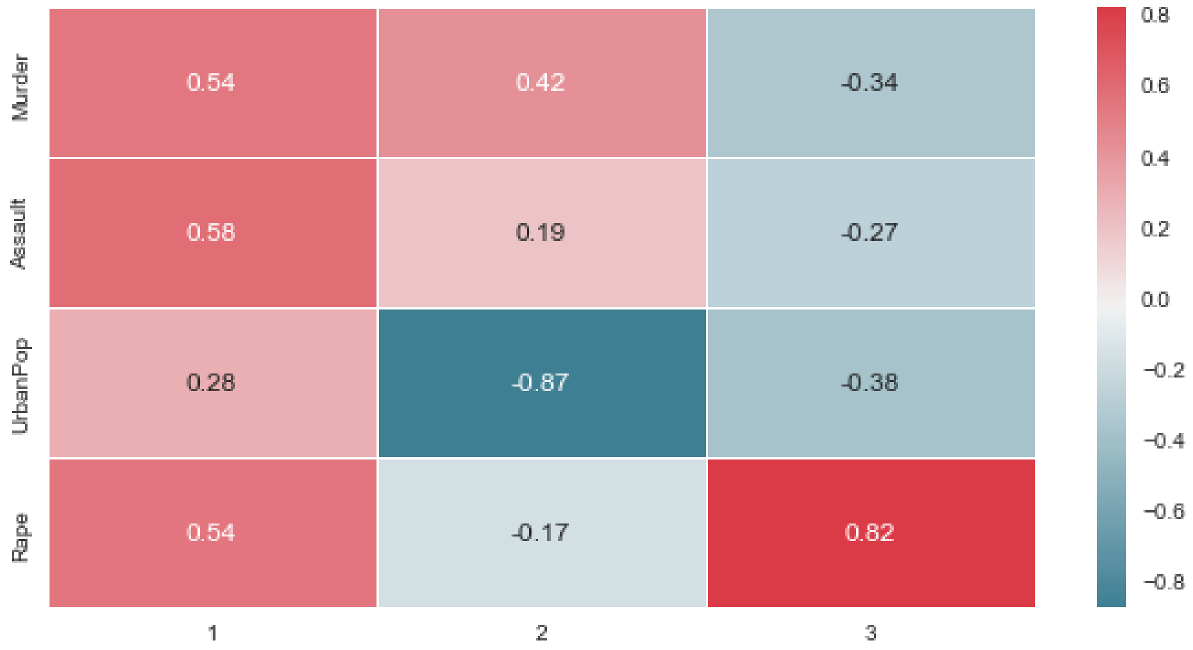


Figure 11. Heatmap of the PCA generated model

The heatmap indicates that there is now variance in the model and can be used for cluster analysis, as the initial heatmap (fig. 8) had little to no variance in the data.

HIERARCHICAL AGGLOMERATIVE CLUSTERING

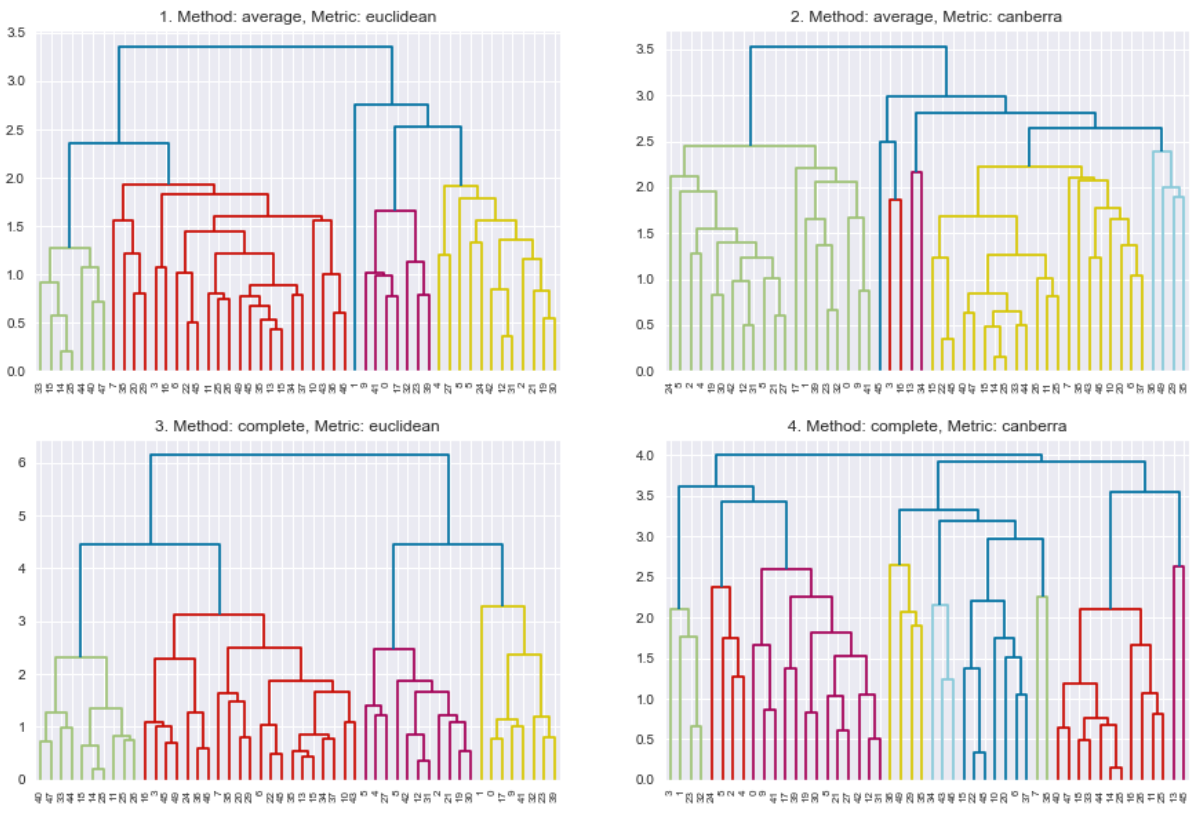


Figure 12. Dendrograms indicating method and metric used

The dendrograms (fig. 12) display clusters generated by the bottom-up approach using various methods and metrics to generate a balanced structure. Model 3, with complete method and Euclidean metric generates the most balanced dispersion of clusters.

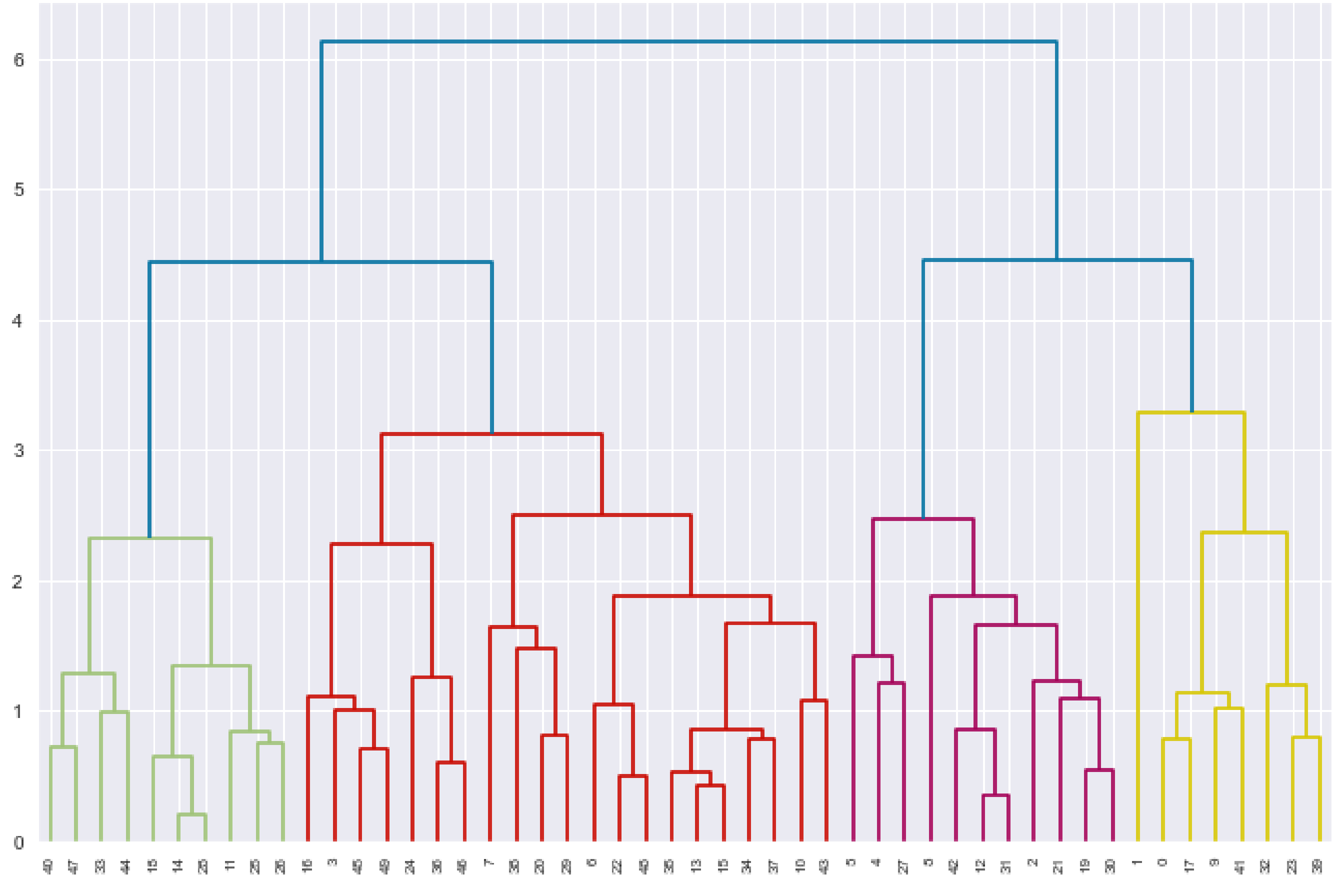
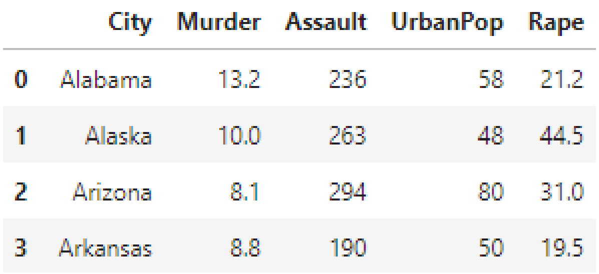


Figure 13. Dendrogram of model 3

In model 3 we can see well defined clusters of size 10, 21, 11 and 8. The model is built on the counts of murder, rape, and assault where the highest is the cluster on the extreme left increasing in severity to the right. Thus, the extreme right (yellow cluster) are cities with the highest arrest counts.

The cities can be referenced by index from 0 – 49. Similar trends to PCA where, countries such as Alaska (0) and Alabama (1) have high counts on arrests for associated crimes, these high counts of arrests have been



	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

Figure 14. slice of the data

clustered into one group. You can see that Alaska (1) has a very long link, and this can be attributed to the fact that Alaska and Alabama are similar in respect to Murder and Urban population, however approximately 11% greater when it comes to assault and rape arrests. And this is what attributes to the longer link but still in the same cluster.

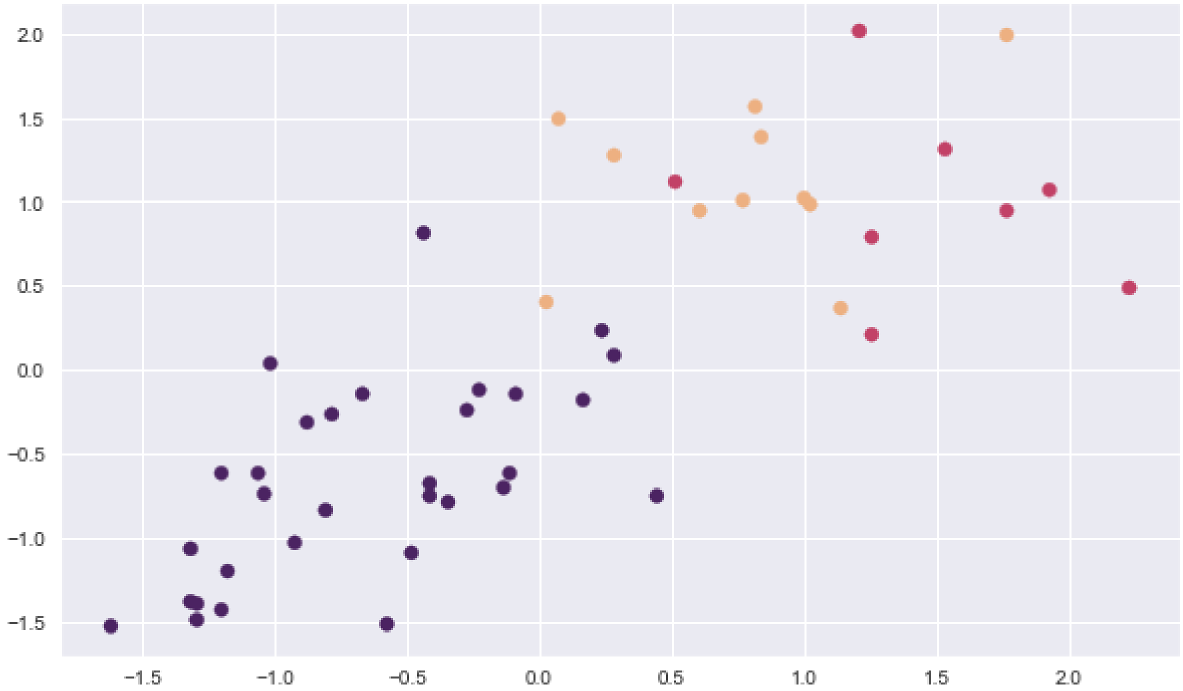


Figure 15. Scatter plot of the predicted clusters

The silhouette score was 0.37, generated from fig 15, which a scatter plot of the predicted clusters. We can see reason why the score is not as high as expected. Many of the data points did not separate well due to low variance in the data and the upper right cluster is mixed in with another cluster. This greatly reduced the silhouette score significantly.

This also indicates that the model found it difficult to separate those clusters entirely due to the high similarity of two or more indicators, as we have seen with the dendrograms where Alaska was similar and different in contrast to Alabama yet still grouped in the same cluster. So this might play a role in poor separation of other data points and thus a low silhouette score.

K-MEANS

The K-means clustering technique is an iterative method which is fast and more widely used. The method segments data into clusters by assigning the observations to the nearest mean which a cluster centroid (fig 18).

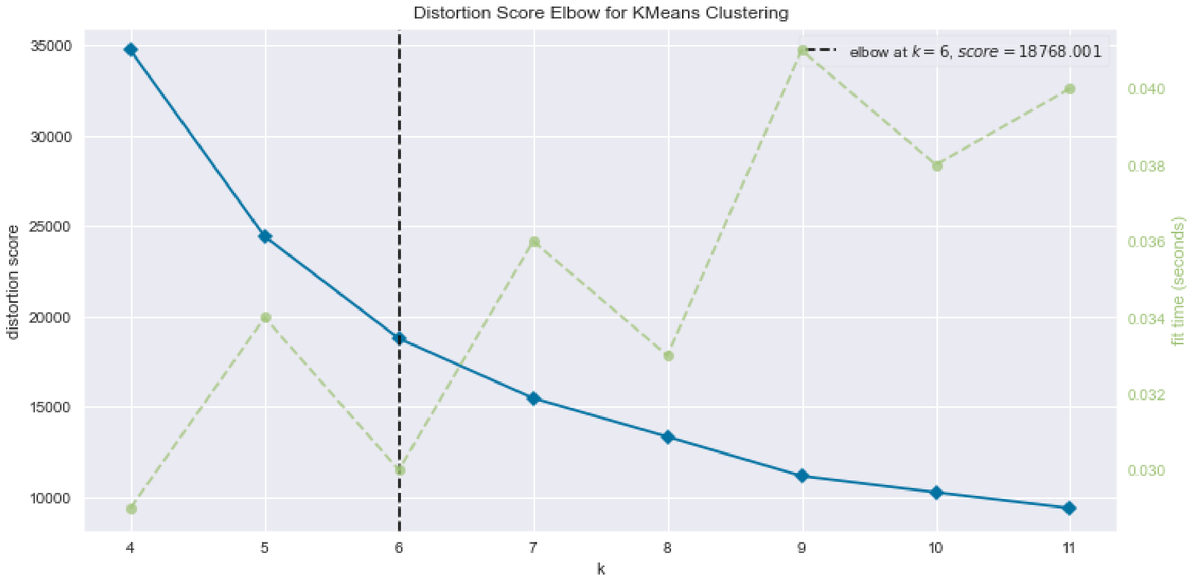


Figure 16. Elbow plot for K-means

In order to cut the repetitious work for each cluster (K), we need to specify K in advance. The above graph (fig 16) displays the optimum distortion score at 6, which was used in this case.

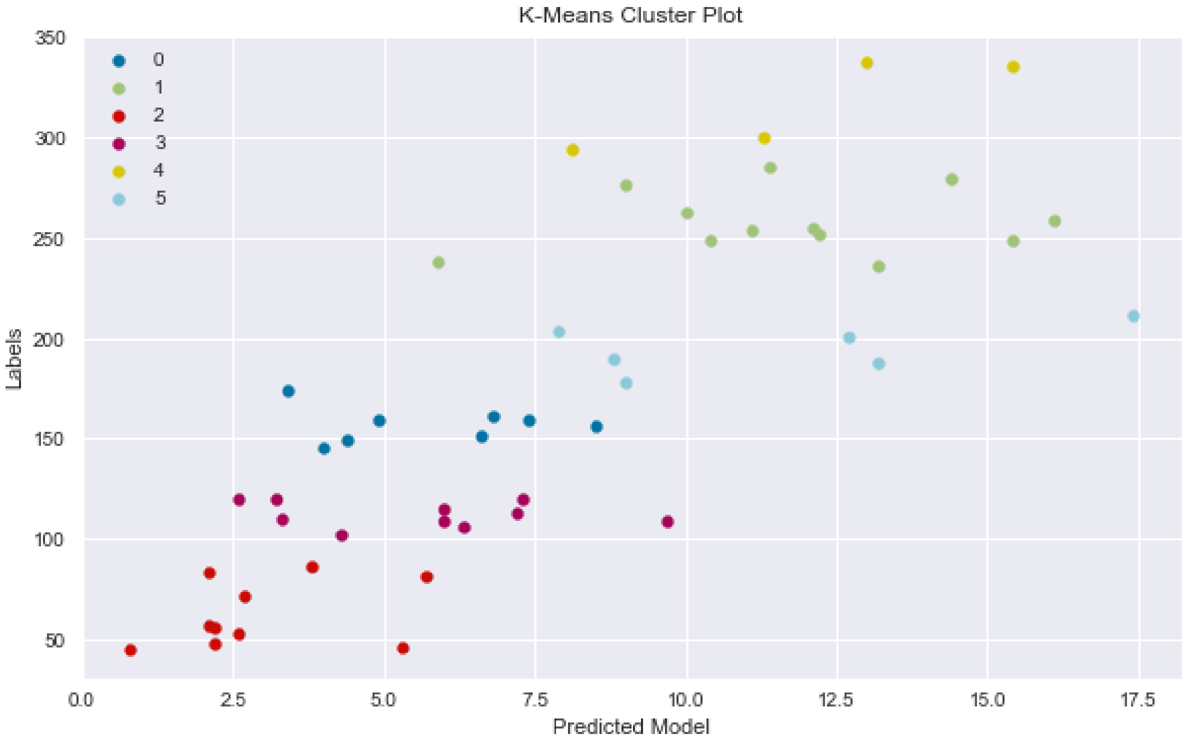


Figure 17. K-means scatter plot of the model

The K-means scatter plot displays 6 clusters where we can see the centroids of each of those clusters. The further away from the cluster centroid the observation is the lower the average is for that observation in respect to the centroid.

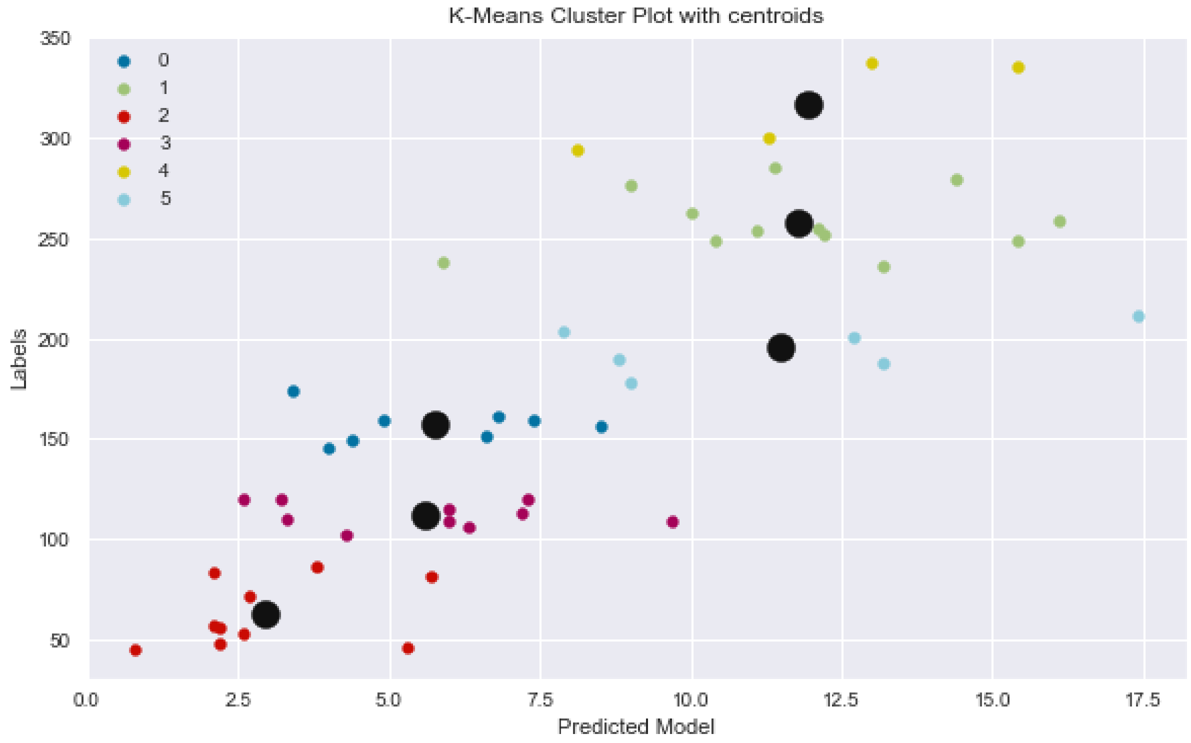


Figure 18. K-means scatter plot of the model with cluster centroids

The silhouette score for this model was 0.45 and we can pretty much see these centroids are quite close, indicating high amounts of similarity in the data.

CONCLUSIONS

The data set was clean with no missing data and allowed for easy machine modelling. The K-means separation was slightly better than Hierarchical agglomerative clustering (HAC) however it is apparent that the data had low variance, and this attributed to low separation scores. The data was separated into 4 clusters of cities from low to high arrests in HAC and 6 in K-means. Even with such low variance the models were able to separate them quite well.

THIS REPORT WAS WRITTEN BY : RAMAN SEWJUGATH