TASK

Unsupervised Machine Learning on the US Arrest Data Set

Visit our website

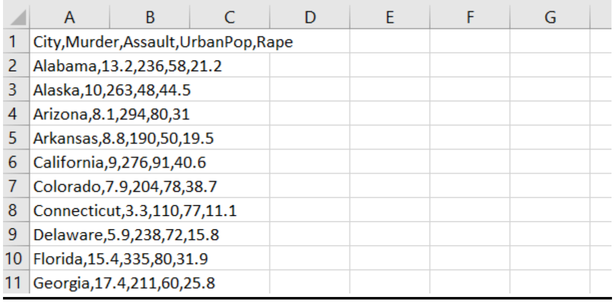
INTORDUCTION

We explore the differences between the 50 US states using unsupervised machine learning methods such as Principal Component Analysis (PCA) and various clustering techniques.

There are 5 variables in total, with 4 variables describing each city/state.

This dataset is from the US Arrests Kaggle challenge.

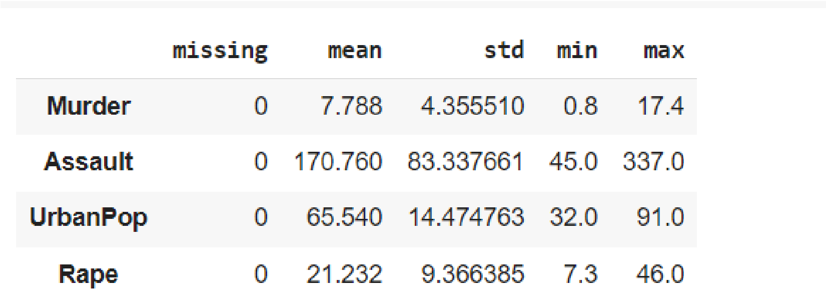
A description of the data is given as: “This data set contains statistics, in arrests per 100,000 residents, for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.” (Kaggle.com)



	A	B	C	D	E	F	G
1	City,Murder,Assault,UrbanPop,Rape						
2	Alabama,13.2,236,58,21.2						
3	Alaska,10,263,48,44.5						
4	Arizona,8.1,294,80,31						
5	Arkansas,8.8,190,50,19.5						
6	California,9,276,91,40.6						
7	Colorado,7.9,204,78,38.7						
8	Connecticut,3.3,110,77,11.1						
9	Delaware,5.9,238,72,15.8						
10	Florida,15.4,335,80,31.9						
11	Georgia,17.4,211,60,25.8						

EXPLORING THE DATA

To get a better understanding of the data a table was created for all variables. The number of missing values, mean, standard deviation (std), min and max values were observed. See the table below

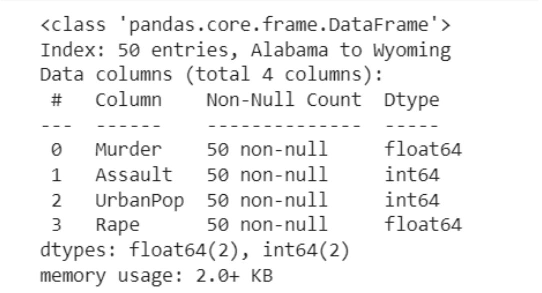


	missing	mean	std	min	max
Murder	0	7.788	4.355510	0.8	17.4
Assault	0	170.760	83.337661	45.0	337.0
UrbanPop	0	65.540	14.474763	32.0	91.0
Rape	0	21.232	9.366385	7.3	46.0

Looking at the data it is clear that Assault has the highest mean and standard deviation compared to the other variables. This is understandable because assault cases are not as severe as rape and murder. However, assault is also more common that rape and murder.

The statistics for the urban population are also high as this variable is a percentage. This difference in scales indicates that it would be necessary to scale the data to prevent the result from being skewed in the direction of the dominants variables.

Using the *info()* function, the data types for all columns was assessed to ensure they were all in a form that could be easily manipulated.

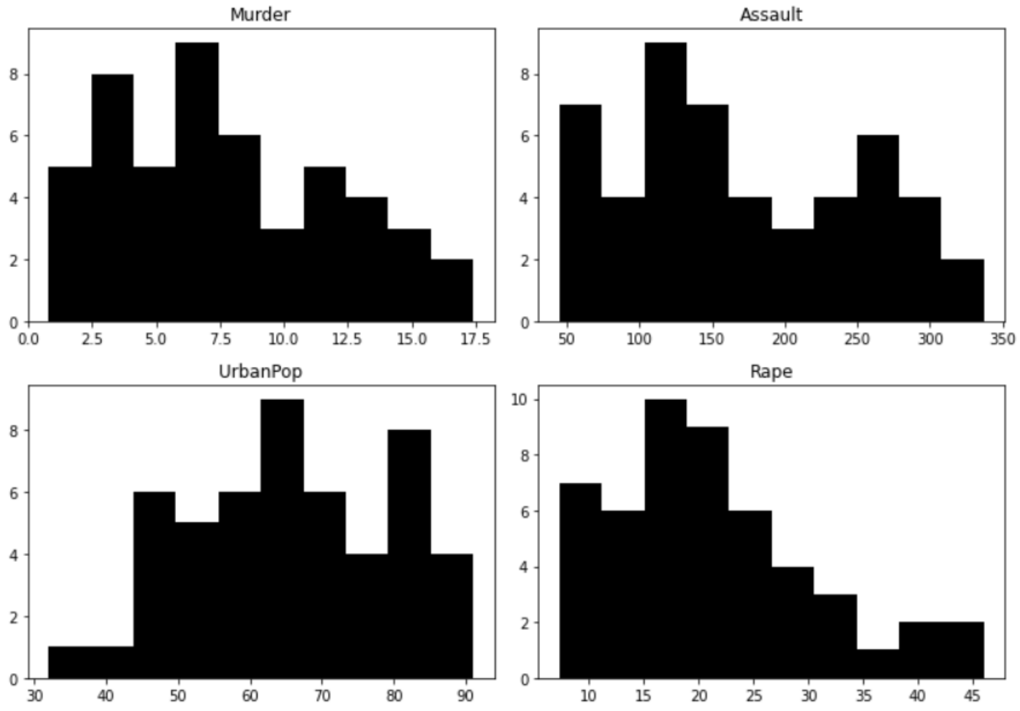


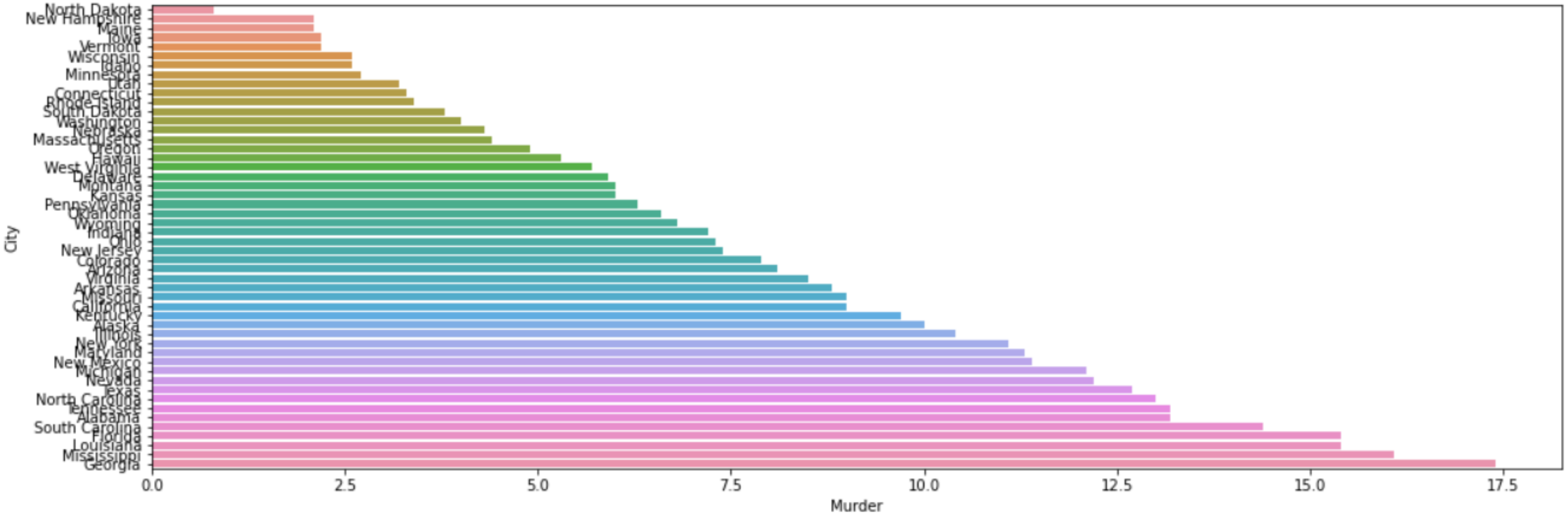
```
<class 'pandas.core.frame.DataFrame'>
Index: 50 entries, Alabama to Wyoming
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Murder    50 non-null     float64
 1   Assault   50 non-null     int64  
 2   UrbanPop  50 non-null     int64  
 3   Rape      50 non-null     float64
dtypes: float64(2), int64(2)
memory usage: 2.0+ KB
```

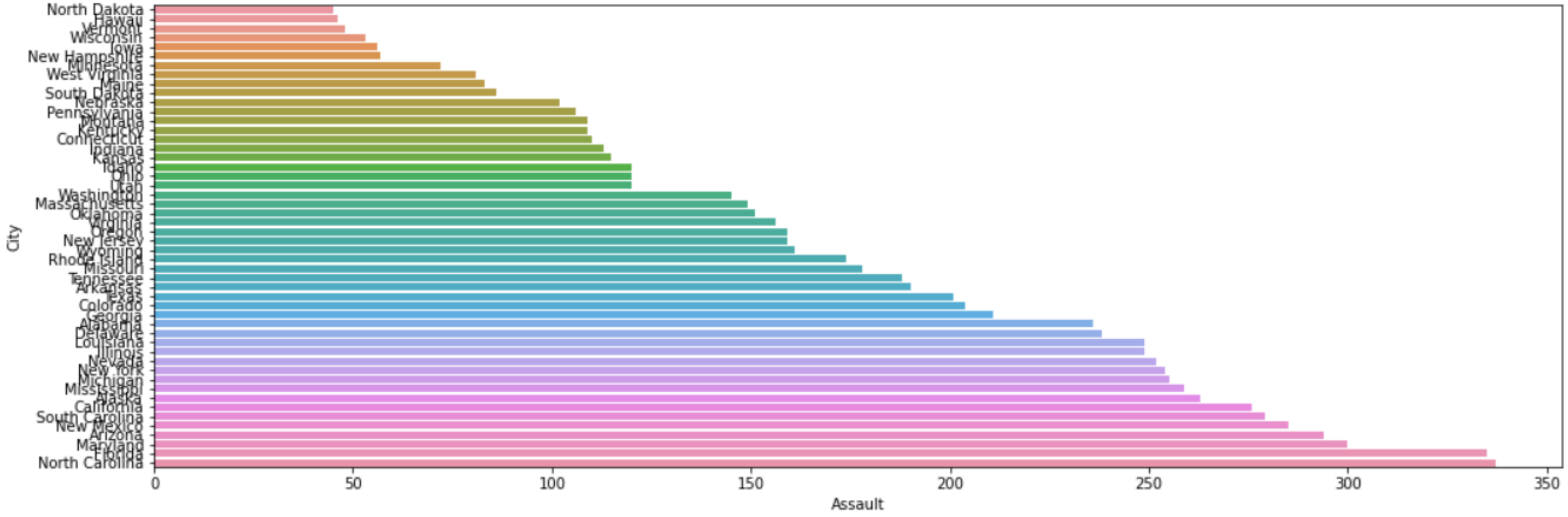
Checks were done for duplicate rows using Pandas' *drop_duplicates()* function. No duplicate rows were found.

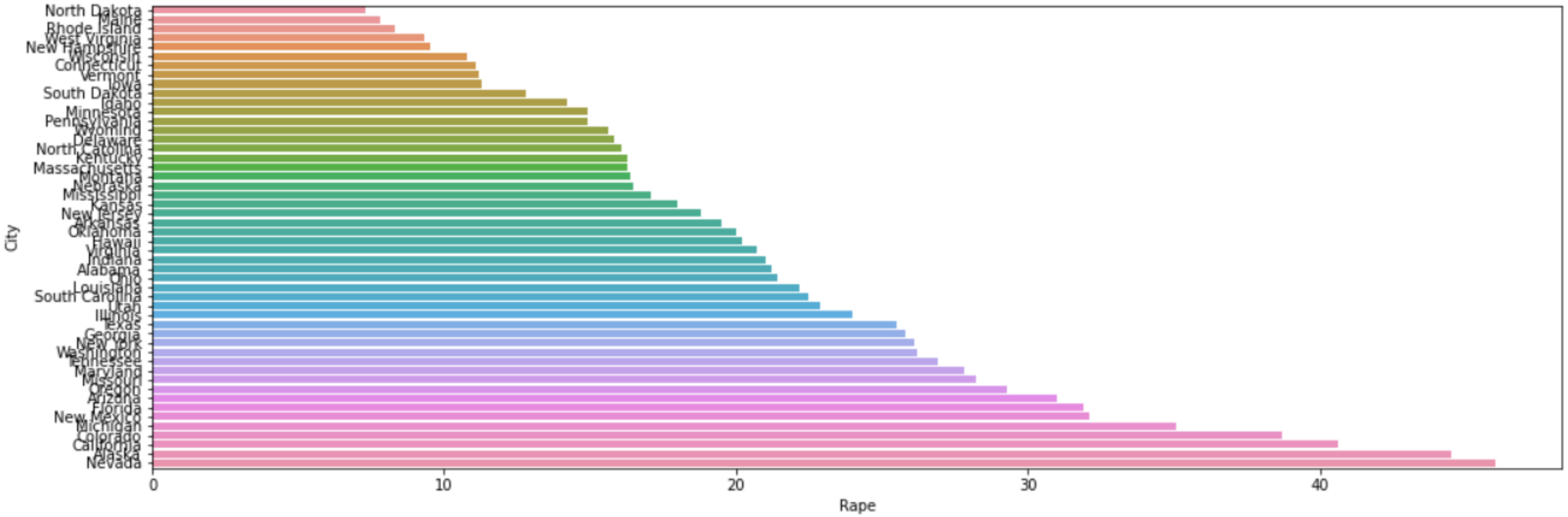
DATA VISUALISATIONS

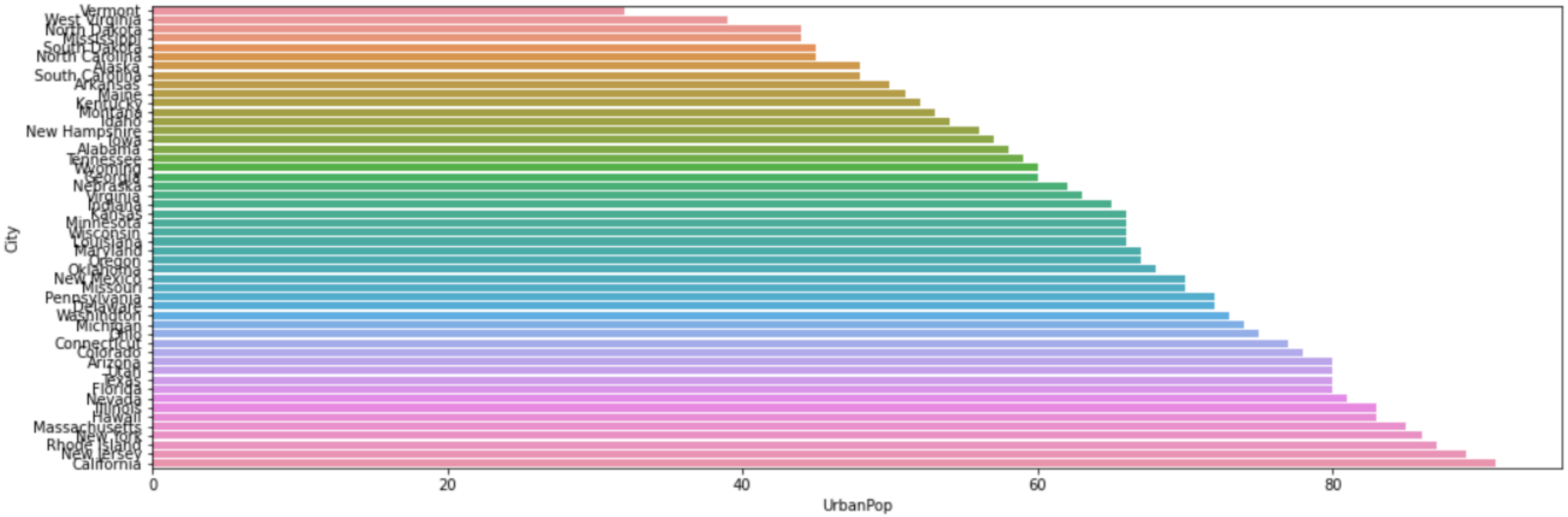
We can also get insight into the spread of the data by plotting histograms for each variable.




City
Murder
North Dakota
New Hampshire
Maine
Iowa
Vermont
Wisconsin
Idaho
Minnesota
Utah
Connecticut
Rhode Island
South Dakota
Washington
Nebraska
Massachusetts
Oregon
Hawaii
West Virginia
Delaware
Montana
Kansas
Pennsylvania
Oklahoma
Wyoming
Indiana
Ohio
New Jersey
Colorado
Arizona
Virginia
Arkansas
Missouri
California
Kentucky
Alaska
Illinois
New York
Maryland
New Mexico
Michigan
Nevada
Texas
North Carolina
Tennessee
Alabama
South Carolina
Florida
Louisiana
Mississippi
Georgia
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


City
Assault
North Dakota
Hawaii
Vermont
Wisconsin
Iowa
New Hampshire
Minnesota
West Virginia
Maine
South Dakota
Nebraska
Pennsylvania
Montana
Kentucky
Connecticut
Indiana
Kansas
Idaho
Ohio
Utah
Washington
Massachusetts
Oklahoma
Virginia
Oregon
New Jersey
Wyoming
Rhode Island
Missouri
Tennessee
Arkansas
Texas
Colorado
Georgia
Alabama
Delaware
Louisiana
Illinois
Nevada
New York
Michigan
Mississippi
Alaska
California
South Carolina
New Mexico
Arizona
Maryland
Florida
North Carolina
0
50
100
150
200
250
300
350


North Dakota
Maine
Rhode Island
West Virginia
New Hampshire
Wisconsin
Connecticut
Vermont
Iowa
South Dakota
Idaho
Minnesota
Pennsylvania
Wyoming
Delaware
North Carolina
Kentucky
Massachusetts
Montana
Nebraska
Mississippi
Kansas
New Jersey
Arkansas
Oklahoma
Hawaii
Virginia
Indiana
Alabama
Ohio
Louisiana
South Carolina
Utah
Illinois
Texas
Georgia
New York
Washington
Tennessee
Maryland
Missouri
Oregon
Arizona
Florida
New Mexico
Michigan
Colorado
California
Alaska
Nevada
City
0
10
20
30
40
Rape

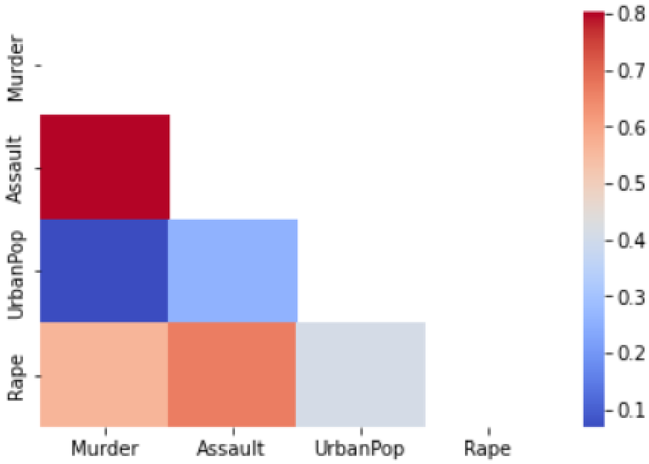

Vermont
West Virginia
North Dakota
Mississippi
South Dakota
North Carolina
Alaska
South Carolina
Arkansas
Maine
Kentucky
Montana
Idaho
New Hampshire
Iowa
Alabama
Tennessee
Wyoming
Georgia
Nebraska
Virginia
Indiana
Kansas
Minnesota
Wisconsin
Louisiana
Maryland
Oregon
Oklahoma
New Mexico
Missouri
Pennsylvania
Delaware
Washington
Michigan
Ohio
Connecticut
Colorado
Arizona
Utah
Texas
Florida
Nevada
Illinois
Hawaii
Massachusetts
New York
Rhode Island
New Jersey
California
City
0
20
40
60
80
UrbanPop

From the bar plots above, the following States are the leaders in murder, rape, assault and urban population:

- Murder - Georgia and Mississippi
- Assault - Florida and North Carolina
- Rape - Nevada and Alaska
- Urban Pop - California and New Jersey

CORRELATION ANALYSIS

In the heat map below, all variables appear to be positively correlated to varying degrees.



From the correlation plot it is clear that there is a strong positive correlation between assault and murder, assault and rape and rape and murder.

These make sense, as assault can be defined as a physical attack on a person (Google dictionary, 2023). So, rape and murder can be considered as types of assault. The more one exists, the greater the chances of the other two existing.

A larger urban population size also increases the likelihood of rape cases.

While there is little to no correlation between the Urban population (UrbanPop) and murder. This is also understandable, as murders can occur anywhere and are not necessarily linked to how urban or rural a city is. There are other factors to consider, such as demographics, education, employment, substance abuse, etc.

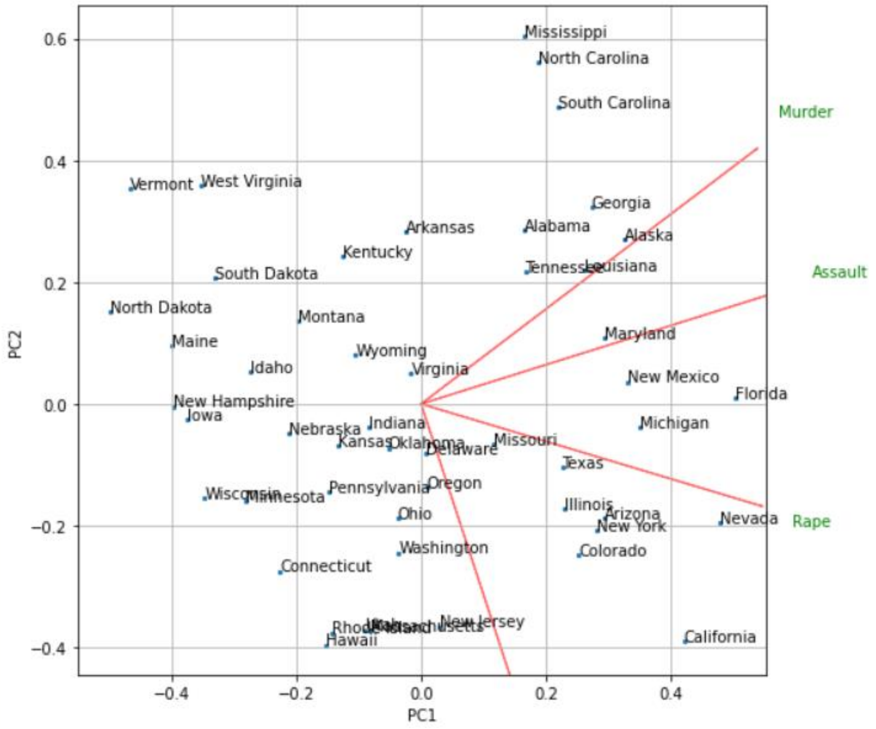
We can use PCA to determine if we can further reduce the dimensions of the data set.

PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is a dimensionality reduction techinque which is useful when we have a lot of variables, and need to reduce them (Hyperiondev$_{b,c}$, 2021). However, we can also use it in this case to check if we can further reduce the dimensions.

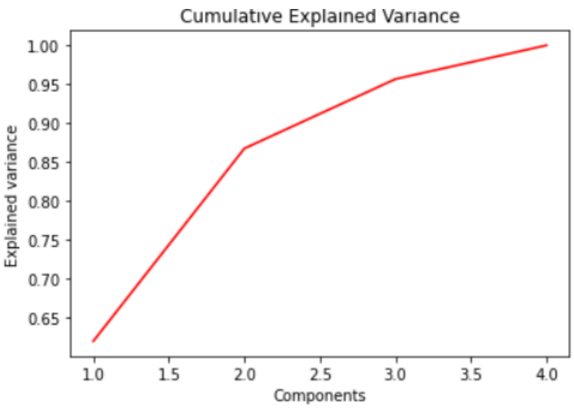
PCA is largely affected by scales and different features might have different scales (Prasad Ostwal, 2019). Recall that Assault's standard deviation and mean were higher than the other variables. This will affect the results if left unchanged. Hence the need to standardize the data before finding PCA components. The data was standardised using Sklearn’s *StandardScalar()* function.

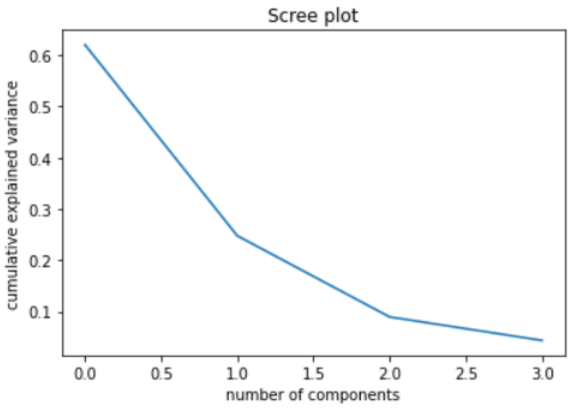
The biplot below shows the directions that the data is distributed as well as possible clusters in the data.



As previously noted in the correlation plot, all the variables are positively correlated. The principal components are all directed towards the right-hand side of the plot. Cities/States with higher crime statistics are grouped on the right-hand side. These also happen to be states with a higher urban population. Cities/ States will lower crime rates are grouped on the left-hand side.

In PCA, the components are listed in order of importance. The data can be reduced to the variables that explain most of the variation in the dataset (Hyperiondev$_{,b,c}$, 2021). To be able to select an appropriate number of principal components we can use a Scree plot and Cumulative Explained Variance plot (Hyperiondev$_{,c}$, 2021). As their names suggest, these are plots of variance for the number of variables (components) in the data set. See the plots below.





From the plots above, it can be seen that the first 3 principal components together explain around 95% of the variance. This allows us to reduce the variables we have from 4 down to 3. We can therefore use them to perform cluster analysis.

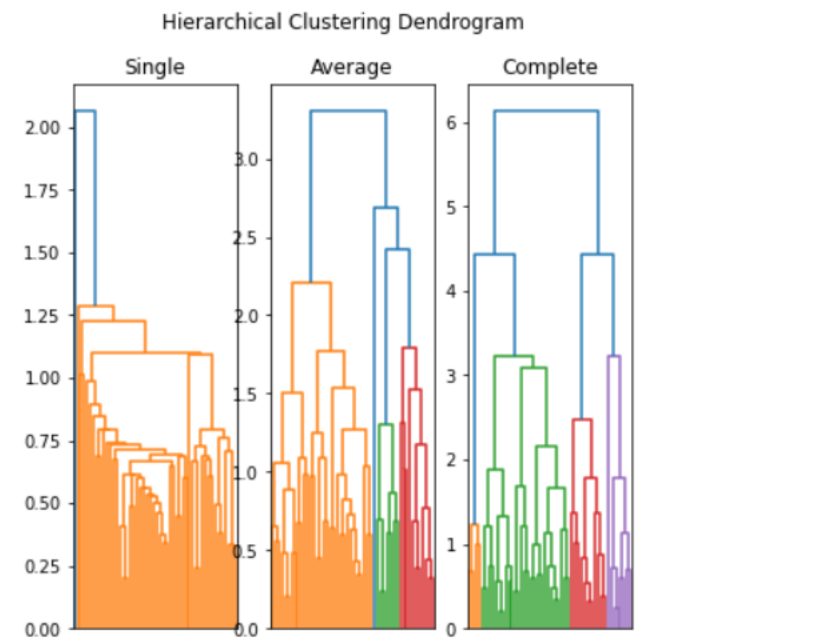
CLUSTER ANALYSIS

On this data set we will look at Hierarchical and K-mean clustering.

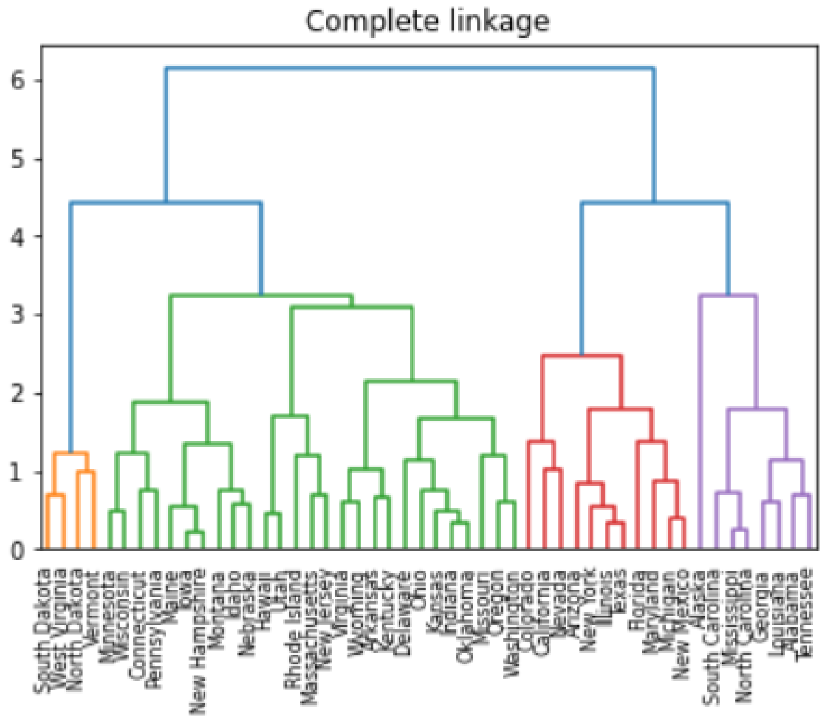
Hierarchical clustering

Hierarchical clustering separates the data into different clusters and gives you an idea of how many clusters you could have without you having to specify that in advance. These clusters are plotted in a dendrogram.

The default setting for the distance metric was used (Euclidean distance), as it is the most common way to measure distance (Hyperiondev,c, 2021). Dendrograms of varying linkages were plotted using the selected distance metric. See plot below.



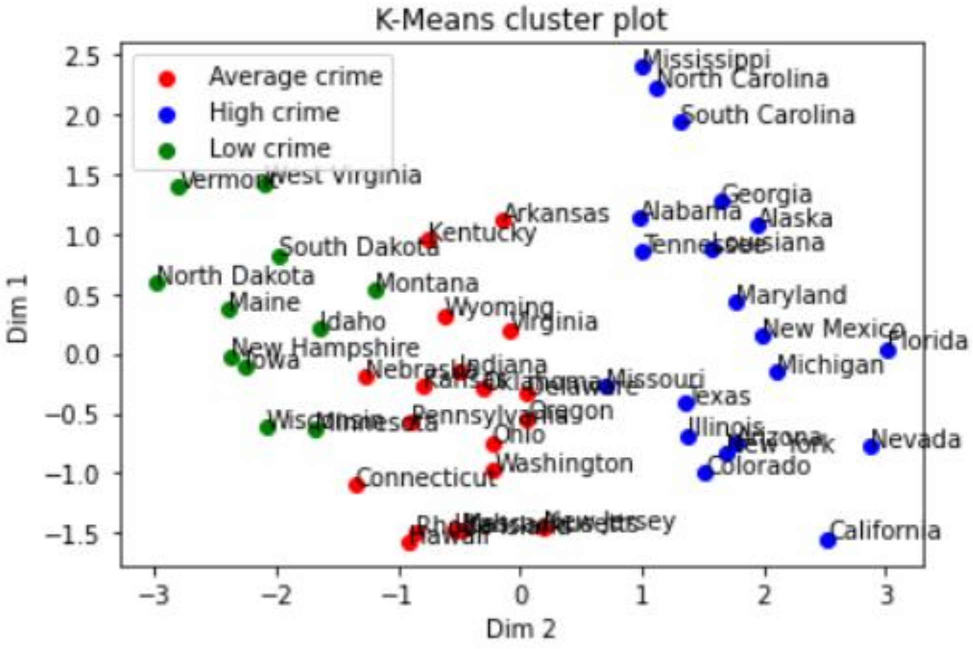
From the plot above, the complete linkage method was selected as it shows the most balanced spread of the clusters. A dendrogram for the complete linkage method is shown below.



Based on the dendrogram above, k=4. With the cities/states in purple and red being the States with high crime rates/arrests and the states with lower crime rates/arrests being clustered in the green and orange group.

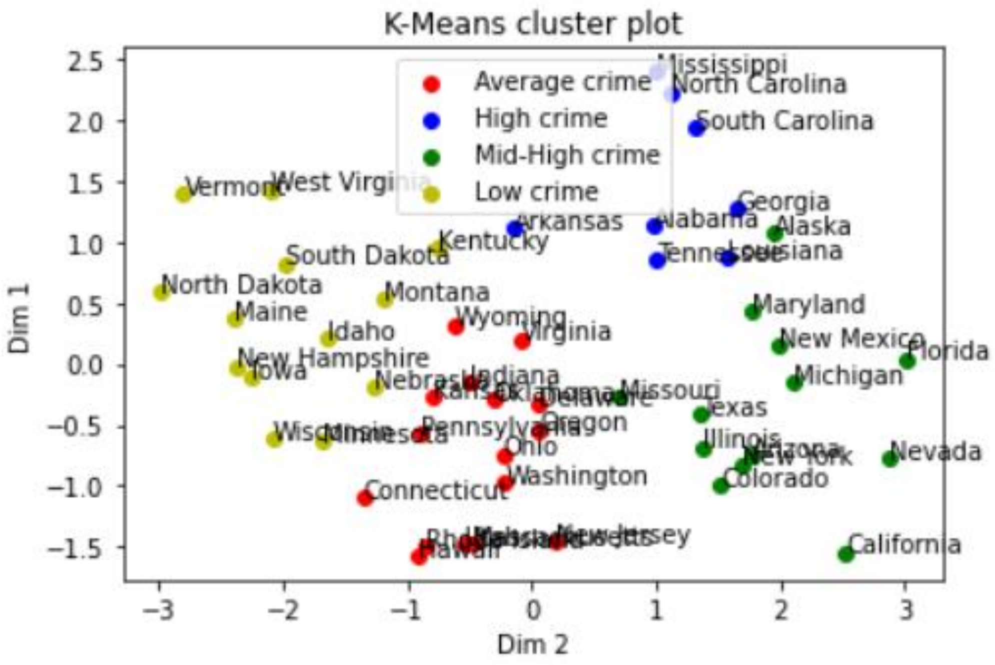
K-means clustering

K-means clustering groups data points with similar centres together based on the number of clusters (K) specified (Hyperiondev$_{,a}$, 2021). This method does not work very well with outliers. For this data set K will be set to 3. See the plot below for a visualisation of the clusters.



K was set to 3 because the data points appeared to be distributed based on crime rate / rate of arrest. The states on the right are the states with the highest arrests in blue, red points have an average/ medium number of arrests and green points are the lowest number of arrests for the three crimes in this data set (rape, murder, assault) .

In an attempt to compare the hierarchical clustering and K-means clustering, changing the K value to 4 for the K-means clustering the states on both plots are similar for the high and mid-high crime rates but are different for the average and low crime rate. See the plot below.



THIS REPORT WAS WRITTEN BY: KP USEH

References

1. Kaggle.com, https://www.kaggle.com/datasets/kurohana/usarrets
2. Google dictionary, https://www.google.com/search?q=assault+meaning&oq=&aqs=chrome.0.69i59i450l8.110063j0j7&sourceid=chrome&ie=UTF-8
3. https://ostwalprasad.github.io/machine-learning/PCA-using-python.html
4. Hyperiondev$_a$, 2021, 'Unsupervised Learning, Clustering'
5. Hyperiondev$_b$, 2021, 'Unsupervised Learning, PCA'
6. Hyperiondev$_c$, 2021, 'Capstone Project, Unsupervised Machine Learning'