US Arrest data PCA Report

In this report we are going to deal with the arrest data for all the 50 states in America. The data is described as the different arrests per 100 000 within the different states separated by: rape, assault, murder, and urban population. Apparently urban crime is described as all the crimes that are done in urban areas. For starters we will be exploring how the data looks like and how it corresponds to itself. The goal of this analysis is to understand how crime relates to area within the United States. As it stands crime is very high and understanding how then crimes correspond per are and per other crimes will help us understand how to possibly solve the crimes and prevent them.

Exploring the data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Murder | Assault | Urban Crime | Rape |
| Count | 50.00000 | 50.000000 | 50.000000 | 50.000000 |
| Mean | 7.78800 | 170.760000 | 65.540000 | 21.232000 |
| Std | 4.35551 | 83.337661 | 14.474763 | 9.366385 |
| Min | 0.80000 | 45.000000 | 32.000000 | 7.300000 |
| 25% | 4.07500 | 109.000000 | 54.500000 | 15.075000 |
| 50% | 7.25000 | 159.000000 | 66.000000 | 20.100000 |
| 75% | 11.25000 | 249.000000 | 77.750000 | 26.175000 |
| Max | 17.40000 | 337.000000 | 91.000000 | 46.000000 |

For starters we will be replacing the column UrbanPop with urban crime for a better readability and understanding because clearly that’s what the column is supposed to be.

It isn’t difficult to return the statistical summary for the data and so here we have it. It shows a slightly greater variance for the Assault crime compared to other crimes but as they are all on the same scale it is very clear that it just shows that more assaults happen per 10000 arrests than other crimes. This is not a good sign but we should forge on.

The first thing we want to explore is the relationships between all of this information in such a way as to understand what kind of data we could possibly yield from our analysis. As we can see below the pair plot describes the different relationships between the data and the histograms for the columns that we have on the data. It seems as though the distribution of values within the various crimes is fairly normal and usually spikes at the point closest to the mean. The data then describes scatter plots between all the different crimes and generally shows really cool distributions that will be explored later with clustering techniques. The most interesting relationships we can see here are the ones between the Assault x Murder and Assault x Rape crimes. They show a rather linear correlation and will probably have a high correlation score. This will look quite similar to the corresponding heatmap.

Chart, histogram

Description automatically generated

Correlation Analysis

We will now try to find out how the data looks when it is related to itself. This is fairly simple and can be represented using a heatmap .

Chart, treemap chart

Description automatically generated

As we can see the heatmap shows interesting relationship scores between all the data. None of the data present shows negative correlation. Like we had guessed the data shows great favouring to Assault x Murder and Assault x Rape Crimes, with the former being more favourable. These correlations are intuitive though. It is as though they have some close some close connection of sorts. This will be the principal focus of our analysis. The PCA algorithm will probably use the first 2 components to describe as well as it can, the relationship between murder and Assault. Due to assault being of a higher value we will have to standardize the data so that it can be made to show a more accurate analysis.

PCA Analysis

The goal of PCA is to explain most of the variability in the data with a smaller number of variables than the original data set. Even though we have a relatively small data set in this instance it still has some use for the PCA analysis. It finds a low-dimensional representation of a data set that contains as much of the variation as possible. The idea is that each of the *n* observations lives in *p*-dimensional space, but not all of these dimensions are equally interesting. PCA seeks a small number of dimensions that are as interesting as possible, where the concept of *interesting* is measured by the amount that the observations vary along each dimension.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 |
| Count | 5.000000e+01 | 5.000000e+01 | 5.000000e+01 | 5.000000e+01 |
| Mean | -1.998401e-17 | -3.330669e-17 | -3.330669e-17 | -1.054712e-17 |
| Std | 1.590867e+00 | 1.004970e+00 | 6.031915e-01 | 4.206774e-01 |
| Min | -2.992226e+00 | -1.570460e+00 | -1.369946e+00 | -9.543817e-01 |
| 25% | -1.117253e+00 | -7.271390e-01 | -4.333588e-01 | -2.113386e-01 |
| 50% | -1.791618e-01 | -1.534797e-01 | 3.120435e-02 | -6.009741e-03 |
| 75% | 1.372968e+00 | 7.672871e-01 | 2.565733e-01 | 2.144737e-01 |
| Max | 3.013042e+00 | 2.393796e+00 | 2.040003e+00 | 1.076797e+00 |

Table

Description automatically generatedIt’s usually beneficial for each variable to be centered around zero for PCA analysis, due to the fact that it makes comparing each principal component to the mean a lot easier. This also eliminates potential problems with the scale of each variable. For example, the variance of *Assault* is is greater than the variance of *Murder, which* is only 18.97. The *Assault* data isn’t necessarily more variable, it’s just on a different scale relative to *Murder*.

By examining the principal components above, we can infer the the first principal component (PC1) roughly corresponds to an overall rate of serious crimes since *Murder, Assault,* and *Rape* have the largest values. The second component (PC2) is affected by *UrbanPop* more than the other three variables, so it roughly corresponds to the level of urbanization of the state, with some opposite, smaller influence by murder rate.

Chart

Description automatically generatedChart, line chart

Description automatically generatedThe first principal component explains 62% of the variability, and the second principal component explains 25%. Together, the first two principal components explain 87% of the variability. As seen in the graphs below we show that by describing the proportion of variance explained and the cumulative PVE. This is another method that describes how the variance is mapped by the pca.

Deciding on how many components to use is a difficult job. The easiest answer is that there is no ‘correct’ method for determining how many components to use. As the number of observations, the number of variables, and the application vary, a different level of accuracy and variable reduction are desirable. The most common way for determining how many principal components to keep is checking the *scree plot*, which is the plot shown above.

The biplot below consolidates what we have been discussing to show the clustering of data based on the variance and this plots the PC1 and PC2.

Graphical user interface

Description automatically generated with low confidence

Cluster analysis

Agglomerative Clustering

Chart, histogram

Description automatically generatedAgglomerative Clustering( also called hierarchal clustering) is a clustering technique that uses the similarity of objects to create clusters. The clusters can be represented visually using dendrograms. The distance used was Euclidean with complete linkage. I use both scatter plot and dendrograms to show how the data is clustered.

Based on the plot above the crimes are grouped into 2 groups that vary widely on one side but are relatively the same and one that isn’t quite like the others. This shows the same thing we say when we were looking at our PCA data. It really did separate into the murder, rape, assault data and the urbanpop data. This shows rather significant indications on the crimes.

Chart, scatter chart

Description automatically generatedDue to the fact that the data describes 2 clusters instead of 3 or 4 we showed how they map onto scatter plots below. The plot describes exactly the same thing that the dendrogram describes but with the addition of variance around 0.

Kmeans clustering

Chart, scatter chart

Description automatically generatedKmeans is a great clustering algorithm that uses quite efficient methods to partition data into clusters that show low variance and describes quite well the dispersion of said data. The following plot describes a kmean partitioning that uses 3 clustering due to the fact that 3 clusters has a better uniform dispersion.

With a silhouette score of 0.4 the data shows a really terrible clustering. Anything below 0.5 is recognized as bad. As you can also see most of the data does not even come close to the means that are marked by an x. They are widely dispersed and lack focus. This is an indication of a really big variance within the data. As if the murdering to the assault and rape don’t quite fit the same bill.