

**TASK**

**Exploratory Data Analysis and Unsupervised Learning on the US Arrests Set**

[Graphical user interface, text

Description automatically generated](https://www.hyperiondev.com/)

# Introduction

This report presents an analysis of the "UsArrests.csv" dataset. The dataset contains statistics on the number of arrests per 100,000 residents for three different types of crimes (murder, assault, and rape) in each of the 50 US states in 1973. Additionally, it includes the percentage of the population living in urban areas.

In this task, we explore the differences between US States using unsupervised learning methods, Principal Component Analysis (PCA) and various clustering techniques. The dataset has 50 observations for each US State and 4 variables (three for the types of crimes and one for the urban population percentage). The data has been imported and organised into a structured format, as a dataframe.

# Data Exploration

The column containing each state was edited to change state names to their initials (i.e., California was renamed to CA). This was performed to improve the presentation of the plots in this report.

|  |  |  |
| --- | --- | --- |
| **Murder** | **Missing** | 0 |
|  | **Count** | 50 |
| **Mean** | 7.79 |
| **Std** | 4.36 |
| **Min** | 0.8 |
| **25%** | 4.08 |
| **50%** | 7.25 |
| **75%** | 11.25 |
| **Max** | 17.4 |

**Figure 1.** Histogram and statistical information of the ‘Murder’ variable

|  |  |  |
| --- | --- | --- |
| **Assault** | **Missing** | 0 |
|  | **Count** | 50 |
| **Mean** | 170.76 |
| **Std** | 83.34 |
| **Min** | 45 |
| **25%** | 109 |
| **50%** | 159 |
| **75%** | 249 |
| **Max** | 337 |

**Figure 2.** Histogram and statistical information of the ‘Assault’ variable

|  |  |  |
| --- | --- | --- |
| **UrbanPop** | **Missing** | 0 |
|  | **Count** | 50 |
| **Mean** | 65.54 |
| **Std** | 14.47 |
| **Min** | 32 |
| **25%** | 54.5 |
| **50%** | 66 |
| **75%** | 77.75 |
| **Max** | 91 |

**Figure 3.** Histogram and statistical information of the ‘UrbanPop’ variable

|  |  |  |
| --- | --- | --- |
| **Rape** | **Missing** | 0 |
|  | **Count** | 50 |
| **Mean** | 21.23 |
| **Std** | 9.37 |
| **Min** | 7.3 |
| **25%** | 15.07 |
| **50%** | 20.1 |
| **75%** | 26.18 |
| **Max** | 46 |

**Figure 4.** Histogram and statistical information of the ‘Rape’ variable

# Figures 1-4 present the statistical characteristics and histograms for each variable in the dataset. Notably, the variables ‘Assault’ and ‘UrbanPop’ exhibit significantly higher mean and standard deviation values compared to other variables, thereby indicating their potential dominance during PCA analysis. To reduce the bias of these variables, data standardisation may be required. No missing data was identified, thus obviating the need for data imputation.

# Correlation Analysis

Chart, bar chart

Description automatically generated

**Figure 5.** Correlation Analysis of the US Arrest database

In Figure 5, a correlation heatmap of the 4 variables is presented, which depicts the degree of linear correlation between the variables. The heatmap displays strong positive correlation in red and weak positive correlation in blue. It can be observed that the 'Assault' variable has a significant positive correlation with the 'Murder' and 'Rape' variables. In contrast, 'UrbanPop' exhibits the weakest correlation with all other variables.

# Principal Component Analysis

Principal Component Analysis (PCA) is a powerful dimensionality reduction technique used to find the underlying structure in high-dimensional data. By identifying the directions along which the data points contain the most variation, PCA can project the data onto a lower-dimensional space while retaining the most important patterns in the original data. Each direction, known as a principal component, is a linear combination of the original variables.

## Unstandardised Data

|  |  |  |  |
| --- | --- | --- | --- |
| **Principal component** | **Standard Deviation** | **Proportion of Variance Explained** | **Cumulative Proportion** |
| 1 | 83.73 | 0.97 | 7011.11 |
| 2 | 14.21 | 0.03 | 7213.11 |
| 3 | 6.49 | 0.01 | 7255.22 |
| 4 | 2.48 | 0 | 7261.38 |

**Table 1.**  Standard Deviation variance associated with each component

The standard deviation of each component is displayed in Table 1, along with the proportion of total variance that the principal component contributes. Notably, the first two components capture the majority of the variance in the data.

Chart, scatter chart

Description automatically generated

**Figure 5.** Principal component biplot of unstandardised data

| **Features** | **PC1 Importance** | **PC2 Importance** |
| --- | --- | --- |
| Murder | 0.04 | 0.04 |
| Assault | 1.00 | 0.06 |
| UrbanPop | 0.05 | 0.98 |
| Rape | 0.08 | 0.20 |

**Table 2.** Feature importance of unstandardised data on PC 1 and PC2

Figure 5 presents a biplot of the first two principal components (PC1 and PC2), which illustrates the projection of each point from the original data onto a two-dimensional space. The importance of each feature is represented by the length of the lines on the biplot, which corresponds to the magnitude of the values in the eigenvectors. The biplot reveals that 'Assault' and 'UrbanPop' are the dominant features for PC1 and PC2, respectively. These findings are supported by Table 2, which reports the feature importance associated with each of the components. To account for this, standardisation of the data is necessary to ensure that all features have equal weight in the analysis.

## Standardised Data

# Chart, scatter chart Description automatically generated

**Figure 6.** Principal component biplot of standardised data

| **Features** | **PC1 Importance** | **PC2 Importance** |
| --- | --- | --- |
| Murder | 0.54 | 0.42 |
| Assault | 0.58 | 0.19 |
| UrbanPop | 0.28 | 0.87 |
| Rape | 0.54 | 0.17 |

**Table 3.** Feature importance of standardised data on PC 1 and PC2

The biplot of the principal components of the standardised data is shown in Figure 6, along with the feature importance quantified in Table 3. The standardisation has ensured that all the features have equal weight, and no larger features are dominating the plot. The biplot reveals that the data points are clustered into two major groups separated by PC1, with some overlap between the groups. The crime-related features, namely Murder, Assault, and Rape, contribute most to the variation of PC1, while PC2 is most influenced by UrbanPop. Interestingly, the direction of the UrbanPop feature does not correlate with the other features, indicating that the percentage of the urban population is not necessarily a reliable indicator of crime rates.

Graphical user interface, chart

Description automatically generated

**Figure 7.** Scree plot and Cumulative Explained Variance for the number of components.

Figure 7 presents the Scree plot and Cumulative Explained Variance for the PCA. The plots illustrate the amount of variance explained by each principal component for a given number of components. These plots can aid in selecting an optimal number of components for dimensionality reduction. In this study, the plots indicate that 2 principal components can account for over 85% of the total variance in the data. This suggests that reducing the dimensionality of the data to 2 components could provide a satisfactory representation of the original data.

# Cluster Analysis

## Hierarchical Clustering

Hierarchical clustering is a method of clustering data objects based on their similarity or distance. It builds a hierarchy of clusters by recursively partitioning the data set into smaller subsets, and then merging them into larger ones. This process results in a tree-like structure called a dendrogram, which displays the hierarchy of the clusters.

## Chart Description automatically generated

**Figure 8.** Hierarchical Clustering dendrograms using various linkage methods

Figure 8 depicts the dendrograms resulting from different linkage methods used to calculate the distances between clusters. The linkage methods used are single, complete, and average. Euclidean distance was chosen as the distance metric to measure the dissimilarity between the data points. The dendrograms show how the clusters are merged as the linkage threshold increases. By observing the dendrograms, the number of clusters can be determined based on the linkage method and distance metric.

Chart, histogram

Description automatically generated

**Figure 9.** Hierarchical cluster plot with complete linkage

Chart, scatter chart

Description automatically generated

**Figure 10.** Hierarchical cluster plot with complete linkage

It appears that the hierarchical clustering using the complete linkage method yielded the most balanced dispersion of clusters with a resulting k value of 4. As shown in Figure 9, the bi-plot of the resulting clusters appears to be strongly dependent on the crime rate, with the green cluster having the highest crime rate and highest urban population, while the yellow group have the lowest crime rates and urban population. The blue group contains states with the highest murder rate. This information can provide useful insights for further analysis and understanding of the factors influencing crime rates in different states.

## K-means clustering

K-means clustering is a popular unsupervised machine learning algorithm used for grouping data points into a predefined number of clusters. The algorithm selects random points as cluster centroids and assigns each data point to the nearest cluster based on the Euclidean distance. The centroid of each cluster is then updated by taking the mean of all the data points assigned to that cluster, and the process of assigning points to the closest cluster and updating centroids is repeated until convergence. The algorithm converges when the cluster assignments no longer change. The technique requires that k be specified in advance.

Chart, scatter chart

Description automatically generated

**Figure 11.** K-means cluster plot with k=4

The k-means bi-plot (figure 11) was generated with a k-value of 4, and the clustering pattern was found to be similar to that of the hierarchical cluster plot, with clusters grouped by crime rates. The bi-plot shows the principal components and the centroids of each cluster. The yellow cluster has a lower crime rate and a lower urban population, while the purple cluster has a higher crime rate and a higher urban population. The centroids of the green and purple clusters have higher values in the Assault and Murder variables, indicating that these two variables play a significant role in differentiating these two clusters. The bi-plot shows that the k-means algorithm is effective in clustering the data, and the results are consistent with the hierarchical clustering plot.

THIS REPORT WAS WRITTEN BY: Joshua Houlden

