Data Analysis

Gargi Rajadnya - 23200711

USArrests Dataset

For the purpose of this assignment, the dataset selected is USArrests.

Introduction

The USArrests dataset is a collection of information on arrest rates in the US for various crimes. It contains data for each of the 50 states, including the number of arrests per 100,000 residents for murder, assault, rape, and the percentage of the population living in urban areas.

Description

Let's delve into the data using descriptive statistics, which provides a summary including maximum, minimum, median, and other key metrics, along with correlation analysis. Exploring the correlation among variables is essential during data analysis, particularly when progressing to model development. Understanding the relationships between variables is crucial as it allows us to streamline models, identify the most relevant features, and validate associations. Additionally, correlation analysis guides data exploration, ensures quality monitoring, and assists in financial decision-making by identifying assets with low correlation for effective risk management. In essence, correlation analysis serves as a potent tool for gaining valuable insights and enhancing decision-making processes.

Let's take a look at the data we have

	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
Colorado	7.9	204	78	38.7

Summary of the data

```
     Murder          Assault         UrbanPop          Rape
 Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
 1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
 Median : 7.250   Median :159.0   Median :66.00   Median :20.10
 Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
 3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
 Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
```

***Murder*:**

Rate of murders per 100,000 residents. It ranges from a minimum of 0.8 to a maximum of 17.4. The median is 7.25, indicating that half the states have murder rates below 7.25 and the other half above.

***Assault*:**

Rate of assaults per 100,000 residents. It ranges from a minimum of 45 to a maximum of 337. The median is 159, indicating that half the states have assault rates below 159 and the other half above.

***Rape*:**

Rate of rapes per 100,000 residents. It ranges from a minimum of 7.3 to a maximum of 46. The median is 20.1, indicating that half the states have rape rates below 20.1 and the other half above.

***UrbanPop*:**

Percentage of urban population in each state. Ranges from a minimum of 32% to a maximum of 91%. The median is 66%, indicating that half the states have a lower urban population and the other half a higher urban population.

Correlation Matrix

```
          Murder Assault UrbanPop Rape
Murder      1.00    0.80     0.07 0.56
Assault     0.80    1.00     0.26 0.67
UrbanPop    0.07    0.26     1.00 0.41
Rape        0.56    0.67     0.41 1.00
```

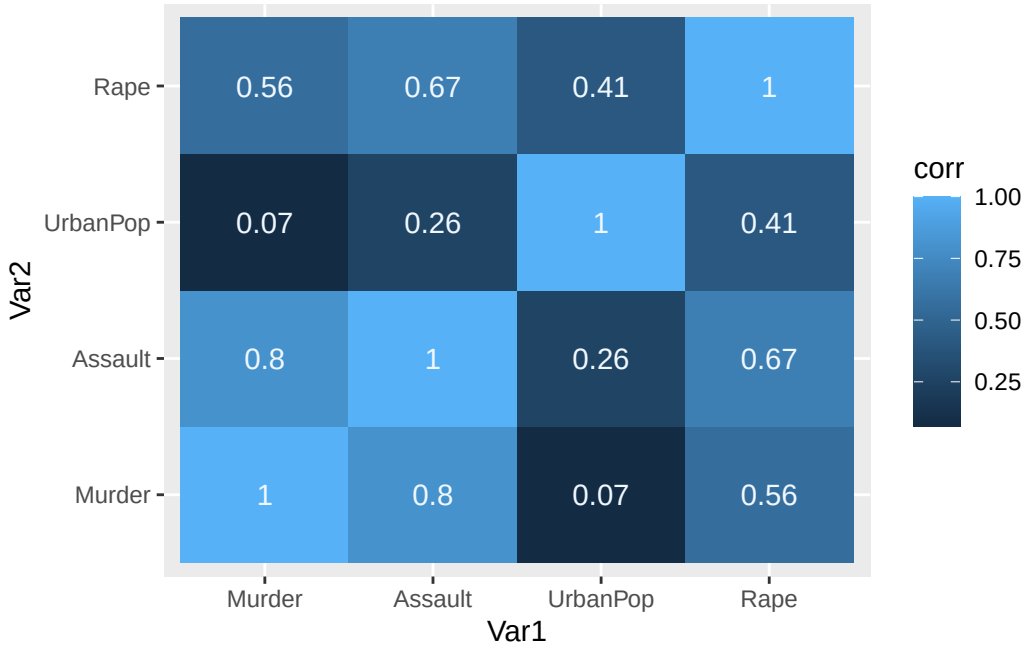
Reshaping

To create a visually appealing heatmap, we need to reshape our correlation matrix into a long format. We'll use the reshape2 library for this purpose and dplyr library to rename our columns.

```
      Var1    Var2 corr
1   Murder  Murder 1.00
2  Assault  Murder 0.80
3 UrbanPop  Murder 0.07
4     Rape  Murder 0.56
5   Murder Assault 0.80
6  Assault Assault 1.00
```

Plotting: Correlation Heatmap

Below is a correlation heatmap plotted with the help of ggplot library. It is a graphical representation of the correlation matrix, which shows how variables in the dataset are related to each other. The below also demonstrates correlation coefficients, which are a measure that represents how strong the relationship is between two variables. The higher the absolute value of the coefficient, the higher is the correlation.



The above heatmap shows the correlation between the variables: Murder, Assault, Rape, UrbanPop

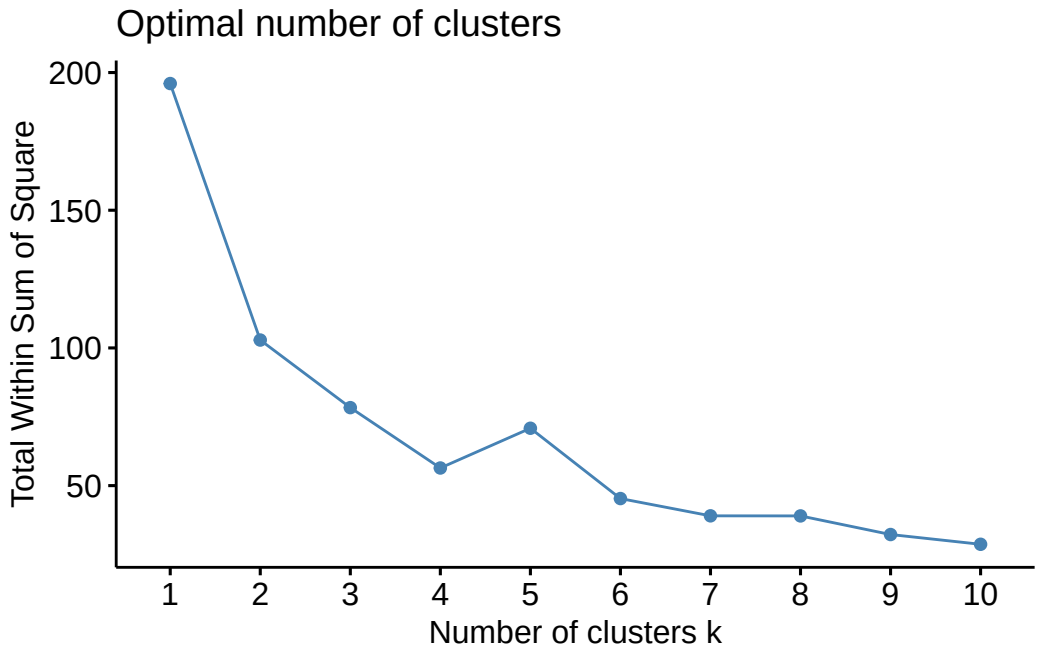
It can be interpreted that Assault and Murder have the highest correlation among the variables. This means that places with higher murder rates also tend to have higher assault rates.

Clustering

Kmeans clustering

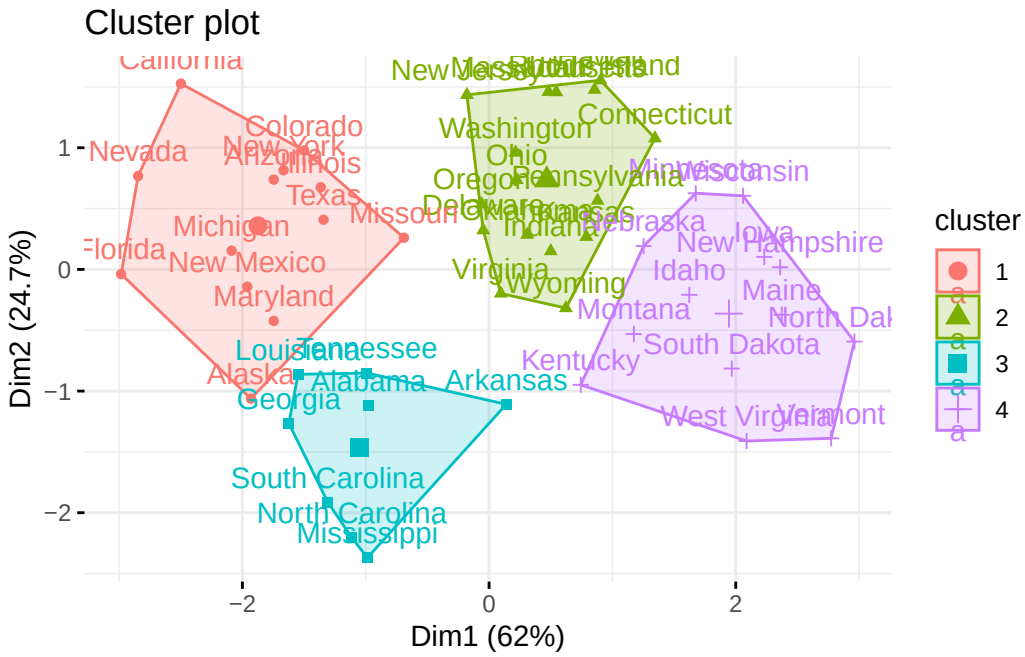
Applying K-means clustering to uncover patterns in crime statistics and group US states into clusters.

```
               Murder   Assault   UrbanPop         Rape
Alabama    1.24256408 0.7828393 -0.5209066 -0.003416473
Alaska     0.50786248 1.1068225 -1.2117642  2.484202941
Arizona    0.07163341 1.4788032  0.9989801  1.042878388
Arkansas   0.23234938 0.2308680 -1.0735927 -0.184916602
California 0.27826823 1.2628144  1.7589234  2.067820292
Colorado   0.02571456 0.3988593  0.8608085  1.864967207
```

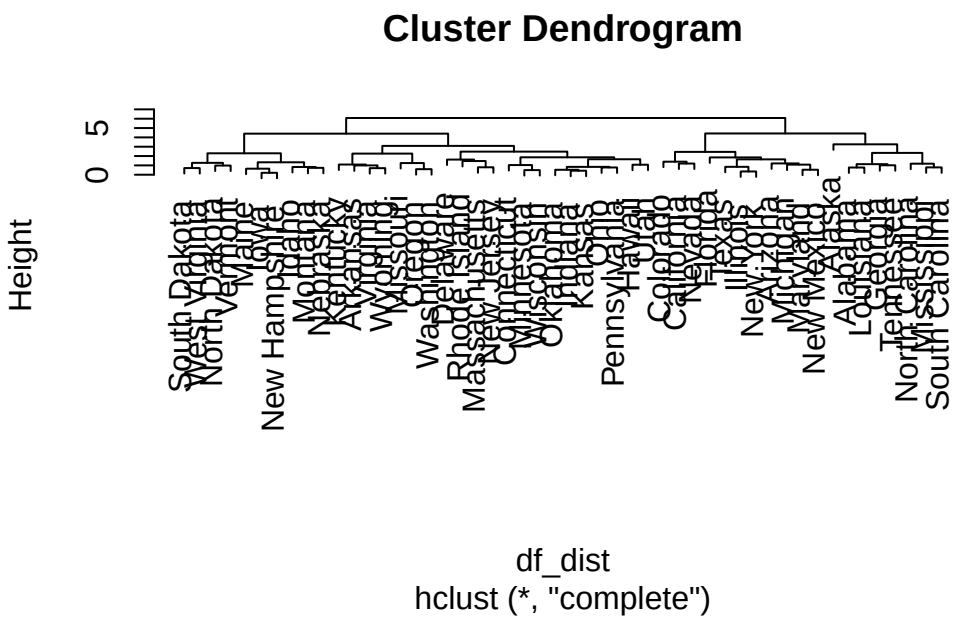


After observation, it can be said, 4 clusters are possibly optimal!

```
  cluster    Murder   Assault UrbanPop     Rape
1       1 10.81538 257.38462 76.00000 33.19231
2       2  5.65625 138.87500 73.87500 18.78125
3       3 13.93750 243.62500 53.75000 21.41250
4       4  3.60000  78.53846 52.07692 12.17692
```



Hierarchical clustering



Again, we can say- 3 or 4 clusters are optimal!