**Case Study**

**1 a)**

**Data frame**



**Dimensions of data frame (Without Adding Index)**

A close-up of a computer screen

Description automatically generated

**Adding of Index**

****

**Dimensions after adding index:**

A screenshot of a computer

Description automatically generated

**1 b)**

**Displaying the First 10 Records of Normalized Data**

A screenshot of a computer

Description automatically generated

Calculation of Normalized Data, Normalized Data scores are also called Z Score Values.

Formula

|  |
| --- |
| Zj = (xj – Uj) / Sj |

Uj = Mean values in Column j.

Sj = Standard Deviation of the numbers in column j.

They are calculated by subtracting the Value of the Variable in a column with a mean value of that column, divided by the standard deviation of that column.

A screenshot of a computer

Description automatically generated

**Example:**

Mean in Murder Column = 7.44

Mean in Murder for Alabama = 14.2

Standard Deviation for Murder Column = 3.87

(14.2-7.44)/3.87 = 1.746(1.75 rounded)

Normalized Value of Alabama for Murder = 1.75.

Clustering with normalized data is better than the original data due to the reason that it makes variables equal contribution levels to the clustering process. All variables are brought to a single scale through normalization, allowing for the cancellation of magnitude differences. Thus, discrepancies in the scale of variables are avoided, as the larger ones do not dominate clustering by default. This provides the opportunity for a more balanced and correct representation of the data, enabling more meaningful and true clustering results.

**2 a)**

Creating of Dendrogram

A screenshot of a computer code

Description automatically generated

In below picture-**Dendrogram** showing Hierarchical Clustering using Complete Linkage Method

A diagram of a clustering method

Description automatically generated

The chart above was generated using a threshold value of 5, resulting in a total of five clusters. Increasing the threshold leads to fewer clusters, while decreasing it results in more clusters being

formed.

**2 b)**

A screenshot of a computer program

Description automatically generated

The above image shows the Clusters formed in our Complete Link Method and the states present in each of the five clusters.

**Mean Values for Each Cluster along with their Input Variable.**

A screenshot of a computer

Description automatically generated

**Normalized Mean Values with their Input Variable with Complete Linkage Method**

A screenshot of a computer program

Description automatically generated

**2 c)**

**Code for the Profile Plots for Normalized Means of Cluster**

A screenshot of a computer code

Description automatically generated

**Plot For Profile Plots**

A graph of different colored lines

Description automatically generated

When comparing to other clusters, states in Cluster 1 exhibit high rates of both violent and non-violent crimes. However, they have lower rankings in robbery and auto crime compared to other clusters, while all other crime rates are the highest within Cluster 1.

In Cluster 2, robbery and auto crimes are prevalent with the remaining crimes at moderate levels.

States in Cluster 3 demonstrate moderate levels of crime across all categories, with intermediate rates of both violent and non-violent crimes.

Cluster 4 states have moderate rates of non-violent crimes, except for larceny and auto theft, where rates are high. Moreover, murder and assault have high crime rates within Cluster 4.

For Cluster 5, murder has the lowest crime rate, while auto theft is the most prevalent crime.

**For labeling purposes, we classify crimes into Violent and Non-Violent**

**Violent Crimes = Rape, Murder, Assault**

**Non-Violent Crimes = Robbery, Burglary, Larceny, auto.**

**Cluster Labeling Using Common Features.**

Cluster One: This cluster exhibits high levels of both violent and non-violent behavior.

Cluster Two: Like Cluster One, this cluster also shows high levels of both violent and non-violent behavior.

Cluster Three: This cluster is characterized by low levels of both violent and non-violent behavior.

Cluster Four: In this cluster, there are low levels of violent behavior but moderate levels of non-violent behavior.

Cluster Five: This cluster displays moderate levels of both violent and non-violent behavior.

Heatmap for Utilities Hierarchical Clustering with Complete Linkage Method

A computer code with red and green text

Description automatically generated

A screenshot of a computer screen

Description automatically generated

**3 a)**

**Developing of Cluster using K-Means**

A screenshot of a computer program

Description automatically generated

K-means clustering is a top-down approach where the number of clusters (k) is predetermined, and observations are randomly assigned to clusters. In contrast, hierarchical clustering is a bottom-up method where each observation begins in its cluster, and clusters are merged iteratively until a single cluster is formed.

While k-means clustering lacks a visual representation of inter-cluster relationships, hierarchical clustering generates dendrograms that illustrate the hierarchical structure.

Due to inconsistency, hierarchical clustering can be computationally intensive and less reliable, making it less suitable for larger datasets. Conversely, k-means clustering is computationally more efficient and highly reliable, making it preferable for larger datasets and developing an elbow chart for the normalized Crime Data.

**3 b)**

**Elbow Chart**

A screen shot of a graph

Description automatically generated

The elbow chart presented above illustrates a substantial decrease in the average within-cluster squared distances from cluster one to cluster four, indicating a significant improvement in the homogeneity of clusters. However, the decrease in these distances from cluster six to cluster eight is relatively minor, suggesting minimal variation in cluster homogeneity within this range. Consequently, based on the k-means clustering of the crime data, it is reasonable to consider k=5 as an appropriate number of groups.

Alternatively, k=3 and k=4 can also be considered, as evidenced by the less pronounced decrease in centroid distance on the chart, indicating higher homogeneity and potentially superior clustering performance.

**3 c)**

**Normalized Cluster Centroids for each cluster**

A screenshot of a computer program

Description automatically generated

**Plot for Normalized Each Cluster**

A graph with lines and text

Description automatically generated with medium confidence

In Cluster 0, crime rates are generally low across all categories except for larceny. Compared to other clusters, Cluster 0 consistently exhibits the lowest crime rates.

Cluster 1 stands out with the highest crime rates across all categories except for auto-related crimes. Both violent and non-violent crimes rank notably high within this cluster.

Similarly, Cluster 2 mirrors the trends seen in Cluster 1, with the second-highest crime rates recorded across most categories, excluding murder and auto-related crimes.

Cluster 3 shows moderate crime rates across all categories except for murder. Crime rates for both violent and non-violent crimes fall within a moderate range in this cluster.

In contrast, Cluster 4 displays low crime rates for violent crimes such as murder and rape, while exhibiting moderate rates for assault. However, auto-related crimes stand out in this cluster, recording the highest crime rate among all clusters.

**Cluster labelling based on Profile Plot of Normalized Centroids**

Cluster zero: This cluster represents areas with very low levels of both violent and non-violent crimes.

Cluster one: Areas in this cluster exhibit very high levels of both violent and non-violent crimes.

Cluster two: This cluster encompasses areas with high levels of both non-violent and violent crimes.

Cluster three: Areas in this cluster experience moderate levels of both violent and non-violent crimes.

Cluster four: This cluster consists of areas with low levels of both violent and non-violent crimes.

Cluster five: Represents areas with moderate levels of violent crimes but low levels of non-violent crimes.

**Heatmap for k-Means Clustering with 5 Clusters**

A screenshot of a computer screen

Description automatically generated

**4)**

Understanding state crime rates benefits from both hierarchical and k-means clustering, each highlighting different data aspects. K-means divides states into six distinct clusters based on crime statistics, offering clear categorization. This method simplifies data analysis with each state belonging to a specific cluster.

Hierarchical clustering delves deeper into state relationships regarding crime rates. Dendrogram illustrates various levels of similarity and dissimilarity among states, providing a nuanced perspective.

Both clustering methods offer valuable insights into state-level crime rates, albeit from different angles. K-means provides a simpler and more understandable approach, while hierarchical clustering offers complexity and depth.

In this scenario, k-means is favored for its ability to generate useful insights, particularly through profile plots of the clusters. Therefore, it's chosen as the preferred clustering method for analyzing state crime rates.