**CASE STUDY 5**

1. **a.** Data frame dimensions, i.e. Number of rows and columns in data set: (50,8)

**The first 10 records of the dataframe are shown below:**

|  | **States** | **murder** | **rape** | **robbery** | **assault** | **burglary** | **larceny** | **auto** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | ALABAMA | 14.2 | 25.2 | 96.8 | 278.3 | 1135.5 | 1881.9 | 280.7 |
| **1** | ALASKA | 10.8 | 51.6 | 96.8 | 284.0 | 1331.7 | 3369.8 | 753.3 |
| **2** | ARIZONA | 9.5 | 34.2 | 138.2 | 312.3 | 2346.1 | 4467.4 | 439.5 |
| **3** | ARKANSAS | 8.8 | 27.6 | 83.2 | 203.4 | 972.6 | 1862.1 | 183.4 |
| **4** | CALIFORNIA | 11.5 | 49.4 | 287.0 | 358.0 | 2139.4 | 3499.8 | 663.5 |
| **5** | COLORADO | 6.3 | 42.0 | 170.7 | 292.9 | 1935.2 | 3903.2 | 477.1 |
| **6** | CONNECTICUT | 4.2 | 16.8 | 129.5 | 131.8 | 1346.0 | 2620.7 | 593.2 |
| **7** | DELAWARE | 6.0 | 24.9 | 157.0 | 194.2 | 1682.6 | 3678.4 | 467.0 |
| **8** | FLORIDA | 10.2 | 39.6 | 187.9 | 449.1 | 1859.9 | 3840.5 | 351.4 |
| **9** | GEORGIA | 11.7 | 31.1 | 140.5 | 256.5 | 1351.1 | 2170.2 | 297.9 |

**b. Normalized Input Variables for seven crimes**

|  | **murder** | **rape** | **robbery** | **assault** | **burglary** | **larceny** | **auto** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **States** |  |  |  |  |  |  |  |
| **ALABAMA** | 1.75 | -0.05 | -0.31 | 0.67 | -0.36 | -1.09 | -0.50 |
| **ALASKA** | 0.87 | 2.40 | -0.31 | 0.73 | 0.09 | 0.96 | 1.94 |
| **ARIZONA** | 0.53 | 0.79 | 0.16 | 1.01 | 2.44 | 2.47 | 0.32 |
| **ARKANSAS** | 0.35 | 0.17 | -0.46 | -0.08 | -0.74 | -1.11 | -1.00 |
| **CALIFORNIA** | 1.05 | 2.20 | 1.84 | 1.46 | 1.96 | 1.14 | 1.48 |
| **COLORADO** | -0.30 | 1.51 | 0.53 | 0.81 | 1.49 | 1.70 | 0.51 |
| **CONNECTICUT** | -0.84 | -0.83 | 0.06 | -0.79 | 0.13 | -0.07 | 1.12 |
| **DELAWARE** | -0.37 | -0.08 | 0.37 | -0.17 | 0.90 | 1.39 | 0.46 |
| **FLORIDA** | 0.71 | 1.29 | 0.72 | 2.37 | 1.31 | 1.61 | -0.14 |
| **GEORGIA** | 1.10 | 0.50 | 0.19 | 0.45 | 0.14 | -0.69 | -0.41 |

The normalized data was calculated using Pandas library in python because Pandas library use sample standard deviation formula, which is more appropriate in clustering. The standardization is done using the following formula:

Z= (x - )/s

where, = sample mean, s = sample standard deviation

The measure computed above is highly influenced by the scale of each variable, so that variables with larger scales have a much greater influence over the total distance. It is therefore customary to normalize continuous measurements before computing the Euclidean distance. This converts all measurements to the same scale. Normalizing a measurement means subtracting the average and dividing by the standard deviation (normalized values are also called z-scores).

1. **a.** Developed the hierarchical clustering (*hi\_complete*) based on the complete (maximum) linkage method (*method=’complete’). T*he hierarchical dendrogram with the cluster threshold of *5.0 is developed in Python and is displayed below.*

Chart, histogram

Description automatically generated

In the dendrogram, it is observed that there are 4 clusters with colours orange, green, red and purple. By choosing a cutoff distance (threshold = 5) on the y-axis, a set of clusters is created. Visually, this means drawing a horizontal line on a dendrogram. Records with connections below the horizontal line (i.e., their distance is smaller than the cutoff distance) belong to the same cluster.

Cluster Membership for 4 Clusters (based on the number of clusters received in the dendrogram) using complete Linkage Method

1 : ARIZONA , CALIFORNIA , FLORIDA , NEVADA , NEW YORK

2 : IDAHO , INDIANA , IOWA , KANSAS , KENTUCKY , MAINE , MINNESOTA , MONTANA , NEBRASKA , NEW HAMPSHIRE , NORTH DAKOTA , PENNSYLVANIA , SOUTH DAKOTA , UTAH , VERMONT , VIRGINIA , WEST VIRGINIA , WISCONSIN , WYOMING

3 : ALABAMA , ARKANSAS , GEORGIA , LOUISIANA , MISSISSIPPI , NEW MEXICO , NORTH CAROLINA , OKLAHOMA , SOUTH CAROLINA , TENNESSEE

4 : ALASKA , COLORADO , CONNECTICUT , DELAWARE , HAWAII , ILLINOIS , MARYLAND , MASSACHUSETTS , MICHIGAN , MISSOURI , NEW JERSEY , OHIO , OREGON , RHODE ISLAND , TEXAS , WASHINGTON

**b.** Normalized Means of Input Variables for Clusters with Complete Linkage Method

|  | **murder** | **rape** | **robbery** | **assault** | **burglary** | **larceny** | **auto** | **Cluster** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | 1.059 | 1.357 | 1.785 | 1.470 | 1.881 | 1.500 | 0.902 | Cluster 1 |
| **2** | -0.753 | -0.801 | -0.734 | -0.852 | -0.828 | -0.479 | -0.623 | Cluster 2 |
| **3** | 1.036 | 0.233 | -0.244 | 0.704 | -0.149 | -0.785 | -0.617 | Cluster 3 |
| **4** | -0.084 | 0.382 | 0.467 | 0.113 | 0.489 | 0.590 | 0.844 | Cluster 4 |

Chart, line chart

Description automatically generated

We can further characterize each of the clusters by examining the summary statistics of their measurements (normalized means) and in the line chart (“profile plot”).

We see, for instance, that cluster 1 is characterized by crimes with a high rate of burglary and robbery compared to other crimes; cluster 2 is characterized by very low crimes; cluster 3 has high murder and low larceny crimes; cluster 4 has low murder and high auto crimes. We can also see which variables do the best job of separating the clusters. For example, the spread of clusters for Burglary and Robbery is quite high, and not so high for the other variables.

**c.** Labelling the clusters based on clusters’ profile plots and normalized mean values.

cluster 1: high burglary and robbery

cluster 2: Relatively very low crimes

cluster 3: high murder and low larceny

cluster 4: low murder and high auto

1. **a.** Applied k-means clustering to classify the states into clusters based on the crime data.

Cluster Membership for 5 Clusters Using k-Means Clustering

0 : ALABAMA, ARKANSAS, GEORGIA, LOUISIANA, MISSISSIPPI, MISSOURI, NEW MEXICO, NORTH CAROLINA, OKLAHOMA, SOUTH CAROLINA, TENNESSEE, TEXAS, VIRGINIA

1 : IDAHO, IOWA, KANSAS, KENTUCKY, MAINE, MINNESOTA, MONTANA, NEBRASKA, NEW HAMPSHIRE, NORTH DAKOTA, PENNSYLVANIA, SOUTH DAKOTA, UTAH, VERMONT, WEST VIRGINIA, WISCONSIN, WYOMING

2 : ALASKA, ARIZONA, COLORADO, DELAWARE, FLORIDA, HAWAII, MARYLAND, MICHIGAN, OREGON, WASHINGTON

3 : CONNECTICUT, ILLINOIS, INDIANA, MASSACHUSETTS, NEW JERSEY, OHIO, RHODE ISLAND

4 : CALIFORNIA, NEVADA, NEW YORK

The main differences between hierarchial and k-means clustering are:

1. In hierarchial clustering, clustering is done based on three methods (Minimum, Average and complete/maximum). Here in this study, complete linkage clustering is adopted, the distance between two clusters is the maximum distance (between the farthest pair of records). In k-means clustering, clustering is done based on the Centroid linkage, i.e. centroid distance, where clusters are represented by their mean values for each variable, which forms a vector of means.
2. The number of clusters are prespecified in the k-means clustering whereas in the hierarchial clustering it either starts with n clusters for *n* observations (agglomerative)or with one cluster that includes all observations (divisive).
3. In hierarchial clustering, goal is to arrange the clusters into a natural hierarchy, whereas the k-means clustering is a most popular nonhierarchical method.

**b. Elbow Chart for k-means clustering (k varies from 1 to 12) of the normalized crime data:**

An “elbow chart” is a line chart depicting the decline in cluster heterogeneity as we add more clusters. The below figure shows the overall average within-cluster distance (normalized) for different choices of k. Moving from 1 to 2 tightens clusters considerably (reflected by the large reduction in within-cluster distance), and so does moving from 2 to 3 and even to 4. Adding more clusters beyond 4 brings less improvement to cluster homogeneity. Yes, k = 5 is an appropriate number of clusters in k-means clustering of the crime data. From 6 to 8 the best k can be chosen, as the clusters increases further, the improvement to cluster homogeneity doesn’t show significant change.

**Shape

Description automatically generated**

**c. Normalized cluster Centroids for k-Means Clustering with k = 5**

|  | **murder** | **rape** | **robbery** | **assault** | **burglary** | **larceny** | **auto** | **Cluster** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0.987 | 0.238 | -0.134 | 0.521 | -0.109 | -0.612 | -0.526 | Cluster 0 |
| **1** | -0.865 | -0.887 | -0.799 | -0.892 | -0.856 | -0.509 | -0.650 | Cluster 1 |
| **2** | 0.053 | 1.056 | 0.478 | 0.627 | 0.997 | 1.336 | 0.479 | Cluster 2 |
| **3** | -0.388 | -0.471 | 0.359 | -0.265 | 0.049 | -0.024 | 1.256 | Cluster 3 |
| **4** | 1.351 | 1.571 | 2.680 | 1.324 | 1.884 | 1.139 | 1.441 | Cluster 4 |

**Profile plot:**

**Chart, line chart

Description automatically generated**

We can further characterize each of the clusters by examining the summary statistics of their measurements (normalized means) and in the line chart (profile plot).

We see, for instance, that cluster 0 is characterized by crimes with a high rate of murder and low larceny; cluster 1 is characterized by very low crimes compared to other clusters; cluster 2 has low murder and high larceny crimes; cluster 3 has low rape and high auto crimes; cluster 4 has low murder and larceny crimes and high robbery. We can also see which variables do the best job of separating the clusters. For example, the spread of clusters for Burglary and Robbery is quite high, and not so high for the other variables.

**d.**  Labelling the clusters based on clusters’ profile plots and normalized mean values.

cluster 0: high murder and low larceny

cluster 1: Relatively very low crimes

cluster 2: low murder and high larceny

cluster 3: low rape and high auto

cluster 4: low murder and larceny and high robbery

1. The spread of two clusters (farthest clusters) is more for in both the hierarchical clustering and k-means clustering, whereas the spread of the remaining clustering is more for hierarchical clustering than the k-means clustering. Hence the hierarchical clustering provides more useful insights regarding crimes’ rates. The farthest the clusters are, the better the clustering and better the insights.