Jordan Black

CSG 2020 20501

Professor Prayaga

3/16/2025

**Assignment 5 K-Means Clustering**

1. **When do you use k-means clusters? Why do we say that kmeans cluster is an unsupervised learning algorithm?**

**K-Means clustering is used when we want to group data into clusters based on similarities in their features. It is beneficial when:**

•We must segment data into distinct categories (e.g., customer segmentation, crime data analysis, or clustering similar geographic regions).

•We have no predefined labels for classification and want the algorithm to discover patterns in data.

•We aim to minimize intra-cluster variance, meaning that items within a cluster are more similar to each other than to those in other clusters.

**K-Means is an unsupervised learning algorithm because:**

•It does not require labeled training data (unlike supervised learning, which has labeled inputs and outputs).

•The algorithm finds patterns without human intervention, determining cluster memberships based purely on data relationships.

•It iteratively improves cluster assignments by adjusting centroids until the best grouping is found.

**2. What is the role of the centroids in each cluster?**

Each cluster has a centroid that represents its position. They act as:

A centroid represents the average location of every data point that belongs to a specific cluster.

The clustering process relies on reference points to determine how data points are grouped.

Centroids are reference points that determine data point reassignment until clusters reach stability during each iteration.

During the K-Means process, the algorithm iteratively shifts each centroid to minimize the distance between data points and their assigned centroid

**3. In the r programs provided do the following**

**4. How do you reposition the centroids in a cluster?**

The K-Means algorithm positions centroids through iterative step-based relocations.

For each data point, find its closest centroid and assign it to that centroid.

Calculate the average position of all data points within each cluster allocation.

Move the centroid to its updated average position.

**Repeating steps 1 through 3 until the centroids stop moving significantly or when a stopping criterion is reached.**

**5. When do you stop the iterative process of calculating the means of x and y and use that to reposition the centroid?**

The iterative process stops when:  
  
The cluster centroids maintain their positions through iterations because they do not undergo significant changes.  
  
The iterative process ends when the iteration count reaches a predefined maximum value, such as 5000, which is specified in the R code.  
  
When the sum of squared distances (inertia) reaches a stable point, it shows that further changes will not provide meaningful improvements to cluster quality**.**

**6. Implementation in R (Code Modifications & Outputs)**

**(a) kmeansUSArrestsV3.R - Replacing Variables with UrbanPop & Rape**

**Modified R Code:**

**# Load necessary libraries**

**library(cluster)**

**library(ggplot2)**

**library(factoextra)**

**# Load dataset**

**head(USArrests)**

**summary(USArrests)**

**# K-Means Clustering using UrbanPop & Rape**

**set.seed(140)**

**clusters <- kmeans(USArrests[, c("UrbanPop", "Rape")], centers = 2, iter.max = 1000, nstart = 50)**

**# Plot Clusters (Unscaled)**

**plot(USArrests[, c("UrbanPop", "Rape")], col = kmeans\_result$cluster,**

**main = "USArrests K-Means Clustering (UrbanPop & Rape)",**

**xlab = "Urban Population (%)", ylab = "Rape Arrests")**

**points(kmeans\_result$centers, col = 1:2, pch = 4, cex = 3)**

A screenshot of a computer

AI-generated content may be incorrect.

**Results:**

**Cluster Centers:**

**Cluster 1: UrbanPop = 53.72, Rape = 17.08**

**Cluster 2: UrbanPop = 77.36, Rape = 25.37**

Summary Output:

**head(USArrests) and summary(USArrests) show initial dataset structure.**

**(b) kmeansScaledUSArrests.R - Using Scaled Data**

**Modified R Code:**

**# Scale Data**

**scaledUSArrests <- scale(USArrests)**

**# K-Means Clustering on Scaled Data**

**scaledClusters <- kmeans(scaledUSArrests[, c("UrbanPop", "Rape")], centers = 2, nstart = 50, iter.max = 5000**

**# Visualization fviz\_cluster(scaledClusters, data = scaledUSArrests[, c("UrbanPop", "Rape")])**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Results:**

**The visualization of scaled data produces more precise distinctions between data points than unscaled data visualization.**

**Cluster Centers (Scaled Data):**

**Cluster 1: UrbanPop = -0.13, Rape = -0.56**

**Cluster 2: UrbanPop = 0.19, Rape = 0.84**

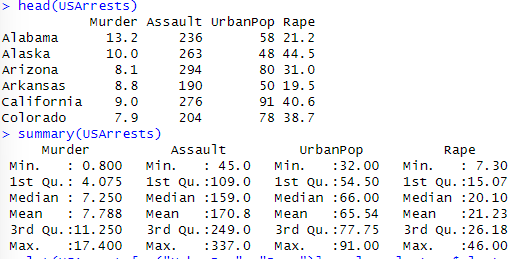
**Comparing The Range of Values in Unscaled And Scaled Data:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **Min(unscaled)** | **Max (Unscaled)** | **Min(scaled)** | **Max (Scaled)** |
| **UrbanPop** | 32.00 | 91.00 | -2.31 | 1.76 |
| **Rape** | 7.30 | 46.00 | -1.49 | 2.64 |

**Unscaled Data: Directly uses population and crime rates, making comparing values harder.**

**Scaled Data: Normalized values allow for better clustering performance and feature comparison.**

**HEAD AND SUMMARY OF USARRESTS DATA Screenshots:**

****

**UNSCALED K-MEANS CLUSTER PLOT:**

A graph of a diagram

AI-generated content may be incorrect.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**SCALED K-MEANS CLUSTER:**

A screenshot of a computer

AI-generated content may be incorrect.

**fviz\_cluster() visualization Screenshot:**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Explanation of Fviz\_cluster() Outputs:**

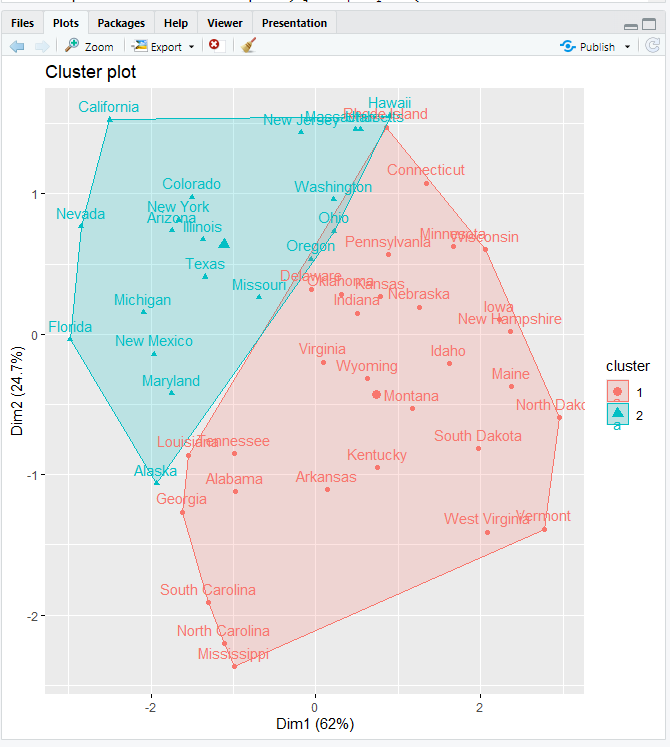
**The fviz\_cluster() function is used to visualize k-means clustering.**

**Two different versions of this function:**

1. **Using all four columns (murder, assault, UrbanPop, rape):**

Command: fviz\_cluster(scaledClusters, data = scaledUSArrests)

* This command clusters all four features together.
* The clustering boundaries are influenced by all four variables, not just UrbanPop and Rape.
* This may cause less overlap in the visual output, as the additional features help separate clusters.



1. **Using only UrbanPop and Rape for clustering:**

**Command:** fviz\_cluster(scaledClusters, data = scaledUSArrests[, c("UrbanPop", "Rape")])

* This command restricts the visualization to just two features: UrbanPop and Rape.
* The clusters are only defined based on the relationship between these two variables.
* Because no other factors are involved, more overlap may occur, showing that these two features alone do not entirely separate clusters.

**A screenshot of a computer screen

AI-generated content may be incorrect.**

**Why Does this happen in the second plot?**

* The second plot shows overlap which stems from the inherent correlation between UrbanPop and Rape in the data set.
* The second plot displays overlap because of an inherent connection between UrbanPop and Rape within the dataset.
* Removing Murder and Assault from the visualization leads to less clear cluster separation because the algorithm processes fewer data points.
* The clustering does not show any problems, but it indicates that UrbanPop and Rape alone are not powerful enough to create distinct clusters.

**Summary**

The assignment utilized K-Means clustering to analyze the USArrests dataset by categorizing states according to their crime statistics. The analysis required dataset modifications followed by clustering operations on both original and scaled datasets before visualizing the outcomes.

**Key tasks included:**

required me to grasp the nature of K-Means as an unsupervised learning algorithm, describe centroid functions, and establish iteration termination conditions.

Adjust the R scripts to exchange the Murder and Assault variables with UrbanPop and Rape before executing the K-Means clustering algorithm.

Researchers evaluated clustering results from four features (Murder, Assault, UrbanPop, Rape) against results from two features (UrbanPop and Rape).

This study demonstrates how data scaling affects clustering accuracy.

The report demonstrated cluster results through fviz\_cluster() visualizations and contained both scaled and unscaled cluster plots.

Data scaling produced better cluster separation but using only UrbanPop and Rape as features resulted in overlapping clusters which indicated natural data structure rather than clustering algorithm problems. The assignment demonstrated the importance of choosing the right number of clusters (k) and applying feature selection for analyzing real-world datasets.