

EXP 13: K-Means Clustering on CO2 Dataset

AIM

The aim of this experiment is to apply the **K-Means clustering** algorithm to the **CO2 dataset** in R. We aim to explore patterns in the data by grouping similar observations into clusters.

EXPERIMENTAL SETUP

- Dataset: CO2 (built-in dataset in R)
- Algorithm: K-Means Clustering
- Key Variables: **conc**, **uptake**
- Preprocessing: Normalization using scaling
- Clustering Evaluation: Elbow Method
- Visualization: Scatter plot with colored clusters

LIBRARIES REQUIRED

```
library(dplyr)
```

```
##
```

```
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
```

```
##
```

```
##      filter, lag
```

```
## The following objects are masked from 'package:base':
```

```
##
```

```
##      intersect, setdiff, setequal, union
```

```
# Load built-in CO2 dataset
```

```
data(CO2)
```

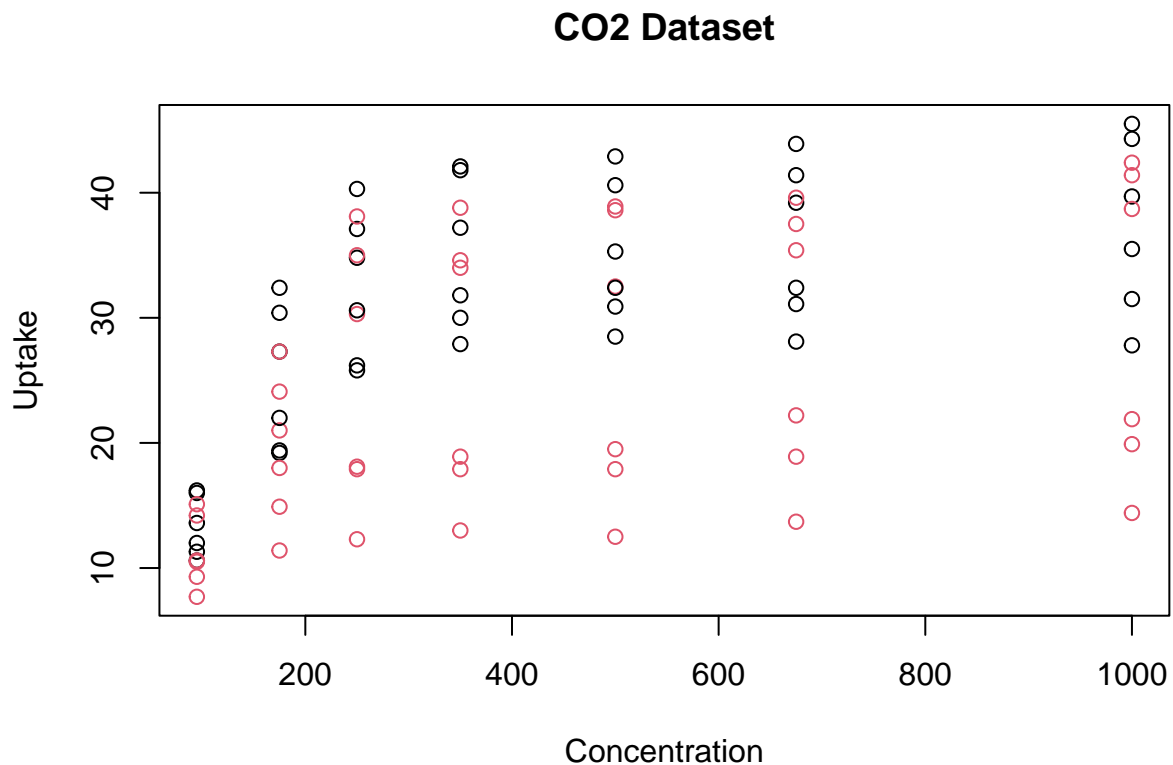
```
# Display first few rows
```

```
head(CO2)
```

```
##   Plant   Type Treatment conc uptake
## 1  Qn1 Quebec nonchilled   95   16.0
## 2  Qn1 Quebec nonchilled  175   30.4
## 3  Qn1 Quebec nonchilled  250   34.8
```

```
## 4   Qn1 Quebec nonchilled 350 37.2
## 5   Qn1 Quebec nonchilled 500 35.3
## 6   Qn1 Quebec nonchilled 675 39.2
```

```
# Scatter plot colored by Treatment
plot(CO2$conc, CO2$uptake,
     col = CO2$Treatment,
     main = "CO2 Dataset",
     xlab = "Concentration",
     ylab = "Uptake")
```



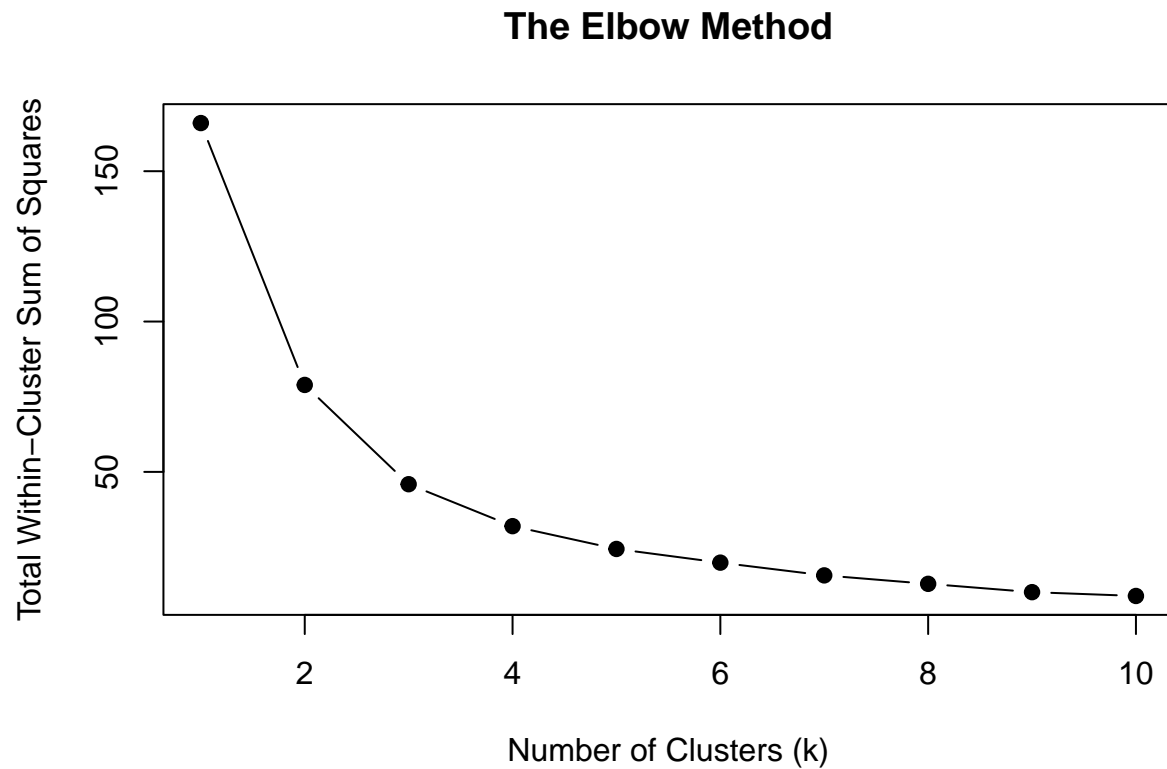
```
# Normalize the selected columns
CO2_normalized <- as.matrix(scale(CO2[, c("conc", "uptake")]))
```

```
# Initialize empty vector to store within-cluster sum of squares
wss <- numeric(10)
k_range <- 1:10
```

```
# Loop to calculate WSS for different values of k
for (k in k_range) {
  km_fit <- kmeans(CO2_normalized, centers = k, nstart = 20)
  wss[k] <- km_fit$tot.withinss
}
```

```
# Plot the Elbow Curve
```

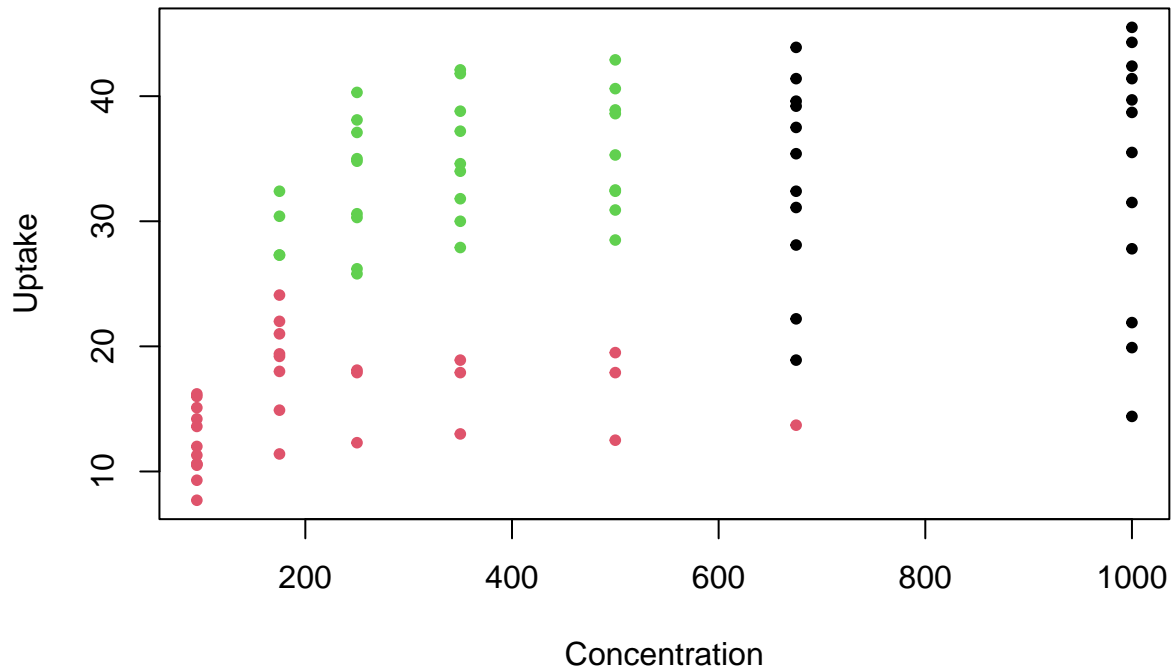
```
plot(k_range, wss, type = "b", pch = 19,
     xlab = "Number of Clusters (k)",
     ylab = "Total Within-Cluster Sum of Squares",
     main = "The Elbow Method")
```



```
# Choose optimal number of clusters
k <- 3
set.seed(123)
km_fit <- kmeans(CO2_normalized, centers = k, nstart = 20)
```

```
# Cluster visualization using scatter plot
plot(CO2$conc, CO2$uptake,
     col = km_fit$cluster,
     main = "CO2 Dataset Clusters",
     xlab = "Concentration",
     ylab = "Uptake",
     pch = 20)
```

CO2 Dataset Clusters



```
# Print cluster assignments for each observation
km_fit$cluster
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
##  2  3  3  3  3  1  1  2  3  3  3  3  1  1  2  3  3  3  3  1  1  2  2  3  3  3
## 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52
##  1  1  2  3  3  3  3  1  1  2  2  3  3  3  1  1  2  2  3  3  3  1  1  2  2  3
## 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78
##  3  3  1  1  2  2  3  3  3  1  1  2  2  2  2  2  1  1  2  2  2  2  2  2  1  2
## 79 80 81 82 83 84
##  2  2  2  2  1  1
```

CONCLUSION

The K-Means algorithm successfully grouped the CO2 dataset into three clusters. The Elbow Method helped identify the optimal number of clusters, and the final visualization showed clear separation between them. This experiment shows how clustering can reveal hidden patterns in data.