Implement K-means and Hierarchical Clustering

AIM:

Implement k-means and Hierarchicalclustering

THEORY:

K-Means Clustering:

- K-means is a partitioning method that divides the dataset into (k) clusters.
- The algorithm works by initializing (k) centroids, assigning each data point to the nearest centroid, and then updating the centroids based on the mean of the assigned points. This process is repeated until convergence.

Hierarchical Clustering:

- Hierarchical clustering creates a tree-like structure (dendrogram) that represents the nested grouping of data points.

PRACTICAL 5 :

CODE :-

```
# Load required libraries
library(ggplot2)
library(cluster)
library(dendextend)

# Load dataset from URL
url <- "https://people.sc.fsu.edu/~jburkardt/data/csv/airtravel.csv"
data <- read.csv(url, header=TRUE)

# Check structure of the dataset
str(data)
```

```
# Convert column names to valid format
colnames(data) <- make.names(colnames(data))

# Convert the first column (Month) to row names for better processing
rownames(data) <- data[,1]
data <- data[,-1]

# Convert data to numeric format
data <- data.frame(lapply(data, as.numeric))

# Handle missing values if any
data[is.na(data)] <- 0

# Normalize the data for clustering
data_scaled <- scale(data)

### K-MEANS CLUSTERING ###
set.seed(123)  # For reproducibility
k <- 3  # Assume 3 clusters
kmeans_result <- kmeans(data_scaled, centers = k, nstart = 25)

# Create a new dataset for K-Means clusters
kmeans_data <- data
kmeans_data$Cluster <- as.factor(kmeans_result$cluster)

# Visualize K-Means Clustering
ggplot(kmeans_data, aes(x = X1958, y = X1959, color = Cluster)) +
  geom_point(size=4) +
```

```
  theme_minimal() +
  labs(title="K-Means Clustering", x="Passengers in 1958", y="Passengers in 1959")

# Print K-Means Cluster Centers
print("K-Means Cluster Centers:")
print(kmeans_result$centers)

# Save K-Means clustered data to CSV
write.csv(kmeans_data, "KMeans_Clusters.csv", row.names=TRUE)

### HIERARCHICAL CLUSTERING ###
dist_matrix <- dist(data_scaled)  # Compute distance matrix
hclust_result <- hclust(dist_matrix, method = "ward.D2")

# Plot the dendrogram
plot(hclust_result, main="Hierarchical Clustering Dendrogram", xlab="", sub="")

# Cut tree into 3 clusters
hclust_clusters <- cutree(hclust_result, k = 3)

# Create a new dataset for Hierarchical clusters
hclust_data <- data
hclust_data$HCluster <- as.factor(hclust_clusters)

# Visualize Hierarchical clustering
ggplot(hclust_data, aes(x = X1958, y = X1959, color = HCluster)) +
  geom_point(size=4) +
  theme_minimal() +
  labs(title="Hierarchical Clustering", x="Passengers in 1958", y="Passengers in 1959")
```

Print Hierarchical Cluster Assignments

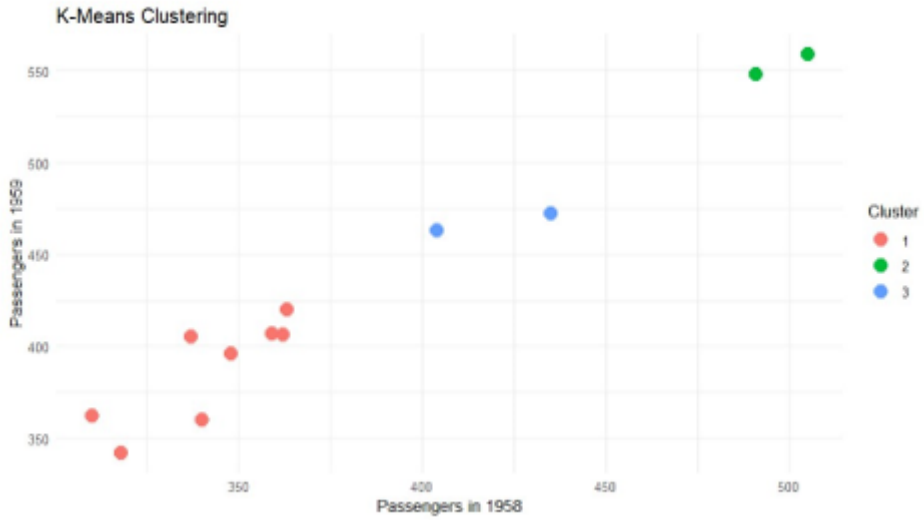
print("Hierarchical Clustering Cluster Assignments:")

print(hclust_clusters)

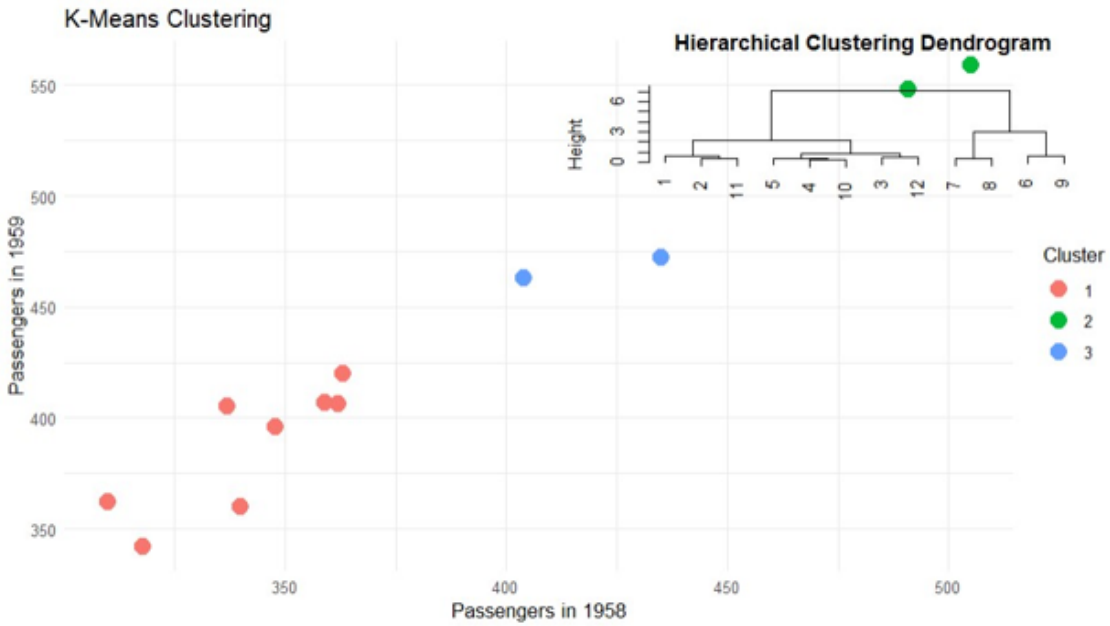
Save Hierarchical clustered data to CSV

write.csv(hclust_data, "Hierarchical_Clusters.csv", row.names=TRUE)

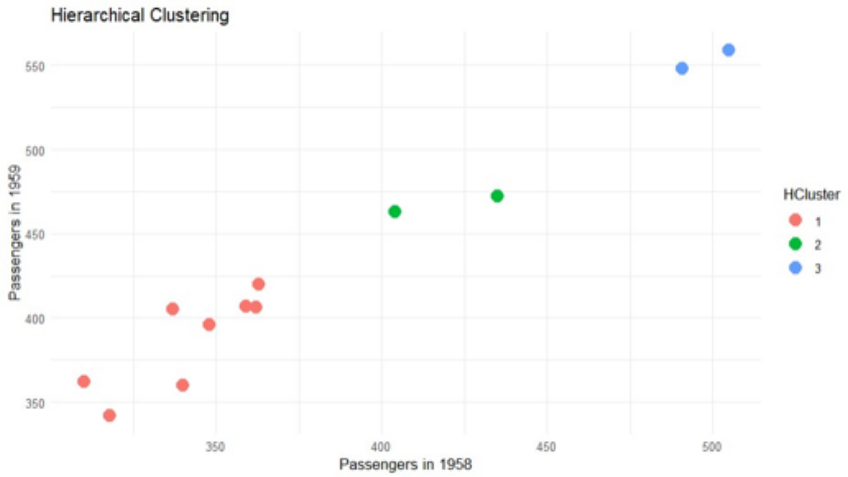
```
# Load required libraries
install.packages("ggplot2")
install.packages("cluster")
install.packages("dendextend")
library(ggplot2)
library(cluster)
library(dendextend)
> # Load dataset from URL
> url <- "https://people.sc.fsu.edu/~jburkardt/data/csv/airtravel.csv"
> data <- read.csv(url, header=TRUE)
>
> # Check structure of the dataset
> str(data)
'data.frame':   12 obs. of  4 variables:
 $ Month: chr  "JAN" "FEB" "MAR" "APR" ...
 $ X1958: int  340 318 362 348 363 435 491 505 404 359 ...
 $ X1959: int  360 342 406 396 420 472 548 559 463 407 ...
 $ X1960: int  417 391 419 461 472 535 622 606 508 461 ...
>
> # Convert column names to valid format
> colnames(data) <- make.names(colnames(data))
> # Convert the first column (Month) to row names for better processing
> rownames(data) <- data[,1]
> data <- data[,-1]
>
> # Convert data to numeric format
> data <- data.frame(lapply(data, as.numeric))
>
> # Handle missing values if any
> data[is.na(data)] <- 0
> # Normalize the data for clustering
> data_scaled <- scale(data)
>
> ### K-MEANS CLUSTERING ###
> set.seed(123) # For reproducibility
> k <- 3 # Assume 3 clusters
> kmeans_result <- kmeans(data_scaled, centers = k, nstart = 25)
>
> # Create a new dataset for K-Means clusters
> kmeans_data <- data
> kmeans_data$Cluster <- as.factor(kmeans_result$cluster)
> # Visualize K-Means Clustering
> ggplot(kmeans_data, aes(x = X1958, y = X1959, color = Cluster)) +
+   geom_point(size=4) +
+   theme_minimal() +
+   labs(title="K-Means Clustering", x="Passengers in 1958", y="Passengers in 1959")
```




```
> # Print K-Means Cluster Centers
> print("K-Means Cluster Centers:")
[1] "K-Means Cluster Centers:"
> print(kmeans_result$centers)
       X1958      X1959      X1960
1 -0.6024286 -0.5883328 -0.5890579
2  1.8130969  1.7924458  1.7730696
3  0.5966174  0.5608852  0.5831619
>
> # Save K-Means clustered data to CSV
> write.csv(kmeans_data, "KMeans_Clusters.csv", row.names=TRUE)
>
> ### HIERARCHICAL CLUSTERING ###
> dist_matrix <- dist(data_scaled) # Compute distance matrix
> hclust_result <- hclust(dist_matrix, method = "ward.D2")
>
> # Plot the dendrogram
> plot(hclust_result, main="Hierarchical Clustering Dendrogram", xlab="", sub="")
```



```
> # Cut tree into 3 clusters
> hclust_clusters <- cutree(hclust_result, k = 3)
>
> # Create a new dataset for Hierarchical clusters
> hclust_data <- data
> hclust_data$HCluster <- as.factor(hclust_clusters)
>
> # Visualize Hierarchical clustering
> ggplot(hclust_data, aes(x = X1958, y = X1959, color = HCluster)) +
+   geom_point(size=4) +
+   theme_minimal() +
+   labs(title="Hierarchical Clustering", x="Passengers in 1958", y="Passengers in 1959")
```




```
> # Print Hierarchical Cluster Assignments
> print("Hierarchical Clustering Cluster Assignments:")
[1] "Hierarchical Clustering Cluster Assignments:"
> print(hclust_clusters)
 [1] 1 1 1 1 1 2 3 3 2 1 1 1
>
> # Save Hierarchical clustered data to CSV
> write.csv(hclust_data, "Hierarchical_Clusters.csv", row.names=TRUE)
```

Conclusion :

- K-Means is generally faster and more efficient for large datasets, while Hierarchical Clustering provides a more detailed view of the data structure through the dendrogram.
- The choice between K-Means and Hierarchical Clustering depends on the specific characteristics of the dataset and the goals of the analysis.

For Faculty Use

Correction Parameters	Formative Assessment [40%]	Timely completion of Practical [40%]	Attendance / Learning Attitude [20%]	
Marks Obtained				

TCET **R-PROGRAMMING** **AIDS-34**

NAME: ujwal sahu **BATCH:02**

SUB: R-PROGRAMMING **PRACTICAL-05** **SYBVOC SEM IV**