**Experiment 3.1**

1. **Aim:** To perform the cluster analysis by k-means method using R.

# Objective:

* + To identify natural groupings or clusters within a dataset using the k-means clustering algorithm in R
  + To apply the k-means clustering algorithm in R to a dataset with a known number of clusters, and to evaluate the effectiveness of the clustering method

# Script:

[K Means Clustering](https://www.geeksforgeeks.org/k-means-clustering-introduction/) in [R Programming](https://www.geeksforgeeks.org/introduction-to-r-programming-language/) is an Unsupervised Non-linear algorithm that cluster data based on similarity or similar groups. It seeks to partition the observations into a pre- specified number of clusters. Segmentation of data takes place to assign each training example to a segment called a cluster. In the unsupervised algorithm, high reliance on raw data is given with large expenditure on manual review for review of relevance is given. It is used in a variety of fields like Banking, healthcare, retail, Media, etc.

K-Means clustering groups the data on similar groups. The algorithm is as follows:

* Choose the number **K** clusters.
* Select at random K points, the centroids (Not necessarily from the given data).
* Assign each data point to closest centroid that forms K clusters.
* Compute and place the new centroid of each centroid.
* After final reassignment, name the cluster as Final cluster.

# Code:

# Installing required packages

# ClusterR is an R package for cluster analysis and provides functions for k-means clustering, hierarchical clustering, and more.

## install.packages("ClusterR")

#The cluster package is an R package for cluster analysis, including functions for k-means clustering, hierarchical clustering, and other algorithms.

## install.packages("cluster")

# Loading packages

**library(ClusterR)** # Load ClusterR library

**library(cluster)** # Load cluster library

# Loading Seatbelts dataset

## data(Seatbelts)

# Removing rows with missing values

**Seatbelts\_1 <- na.omit(Seatbelts[, -1])** # Remove rows with missing values in Seatbelts dataset

# Fitting K-Means clustering Model to training dataset

**set.seed(240)** # Setting seed for reproducibility

**kmeans.re <- kmeans(Seatbelts, centers = 3, nstart = 20)** # Fit k-means clustering model to Seatbelts dataset with 3 clusters and 20 starts

# Cluster identification for each observation

**kmeans.re$cluster** # Display the cluster identification for each observation

# Creating a confusion matrix

**cm <- table(Seatbelts$front, kmeans.re$cluster)** # Create a confusion matrix of Seatbelts dataset and k-means clustering result

**cm** # Display the confusion matrix # Model Evaluation and visualization

# Plot drivers vs front for Seatbelts dataset

## plot(Seatbelts[, c("drivers", "front")], ylim = c(0, max(Seatbelts\_1$front)))

# Plot drivers vs front for Seatbelts dataset with cluster colors

## plot(Seatbelts[, c("drivers", "front")], col = kmeans.re$cluster, ylim = c(0, max(Seatbelts$front)))

# Plot drivers vs front for Seatbelts dataset with cluster colors and main title

## plot(Seatbelts[, c("drivers", "front")], col = kmeans.re$cluster, main = "K-means with 3 clusters", ylim = c(0, max(Seatbelts\_1$front)))

# Plotting cluster centers

# Display the cluster centers

## kmeans.re$centers

# Display the cluster centers for drivers and front features

## kmeans.re$centers[, c("drivers", "front")]

# Plot the cluster centers with different colors, shapes and size

## points(kmeans.re$centers[, c("drivers", "front")], col = 1:3, pch = 8, cex = 3)

# Visualizing clusters

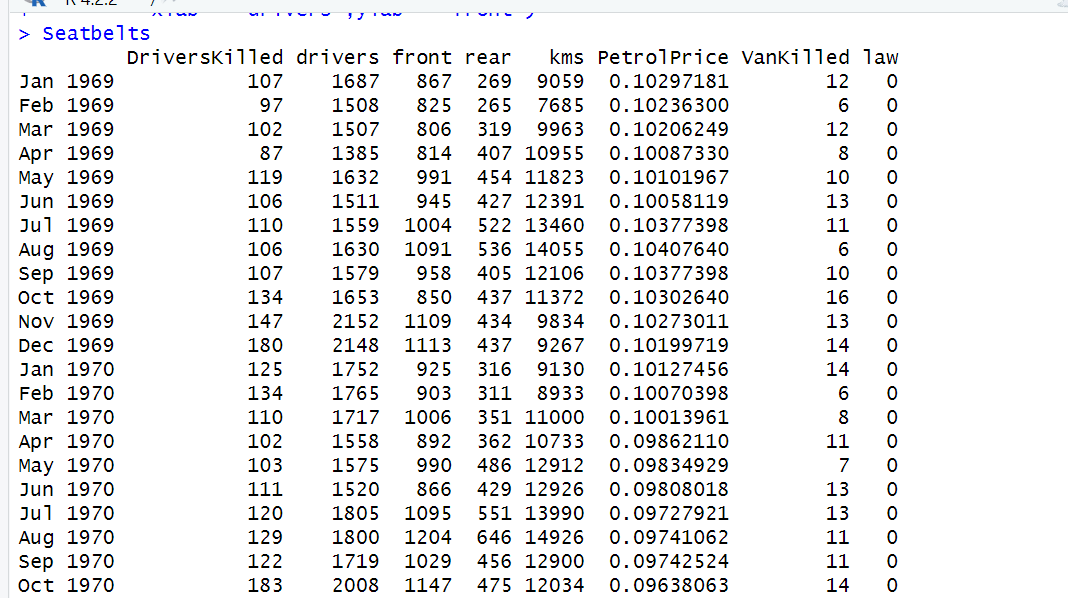
**y\_kmeans <- kmeans.re$cluster** # Assign the cluster identification to y\_kmeans variable **clusplot(Seatbelts[, c("drivers", "front")],** # Plot a cluster plot of drivers vs front for Seatbelts dataset

y\_kmeans, lines = 0, shade = TRUE, color = TRUE, labels = 2,

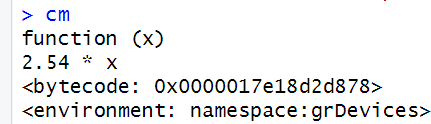
plotchar = FALSE, span = TRUE,

main = "Cluster Seatbelt", xlab = 'drivers',ylab = 'front')

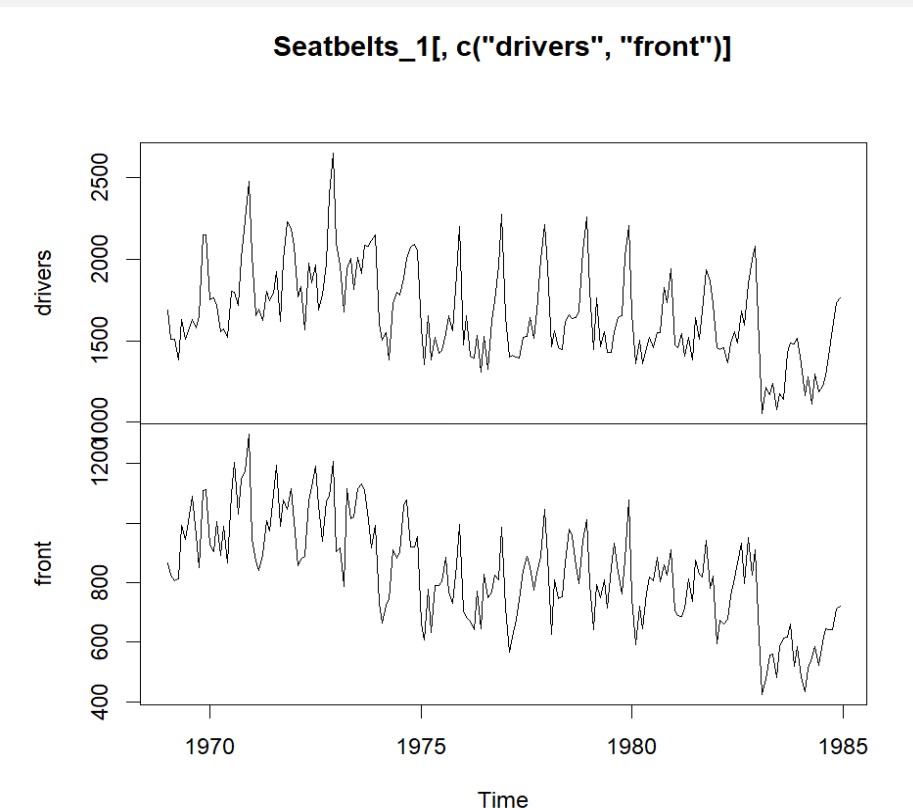
# Output:

1. Seatbelts dataset

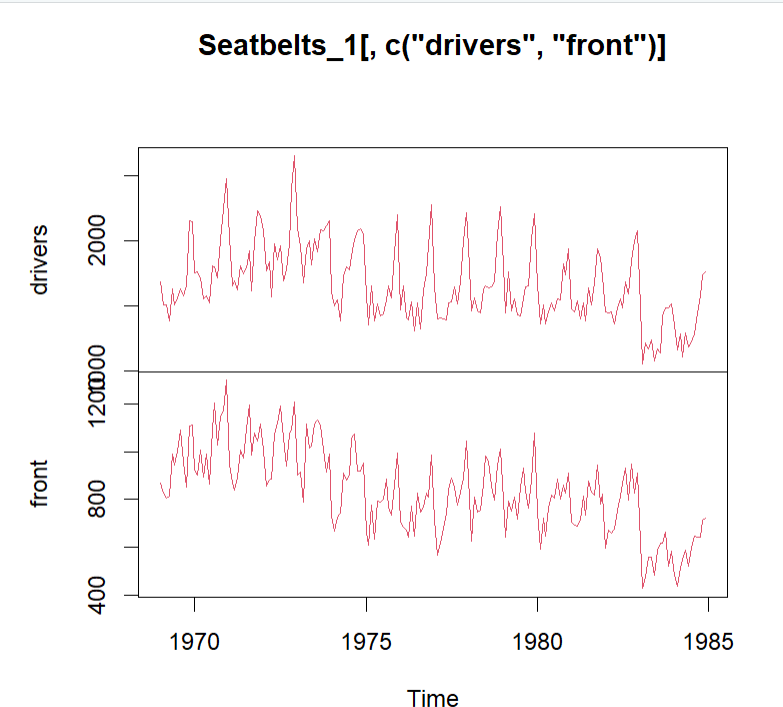
## Clustering Identification:

**Confusion Matrix**

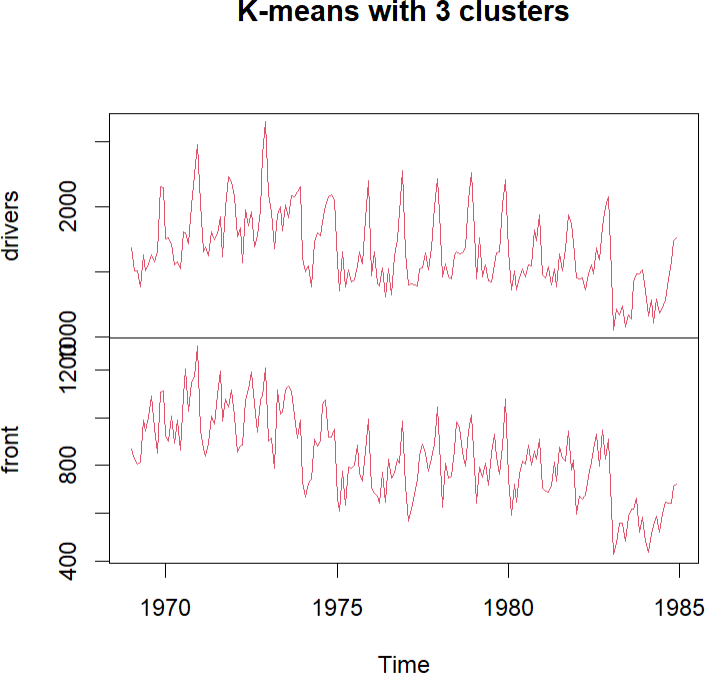
## # Plot drivers vs front for Seatbelts dataset

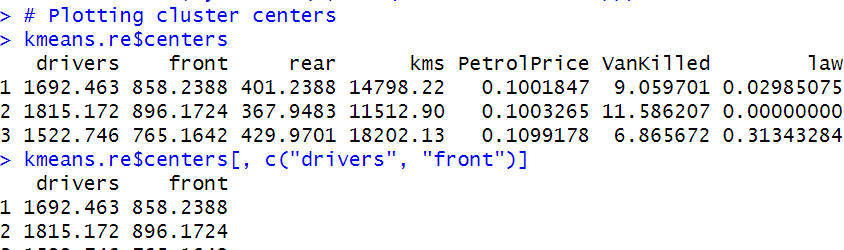


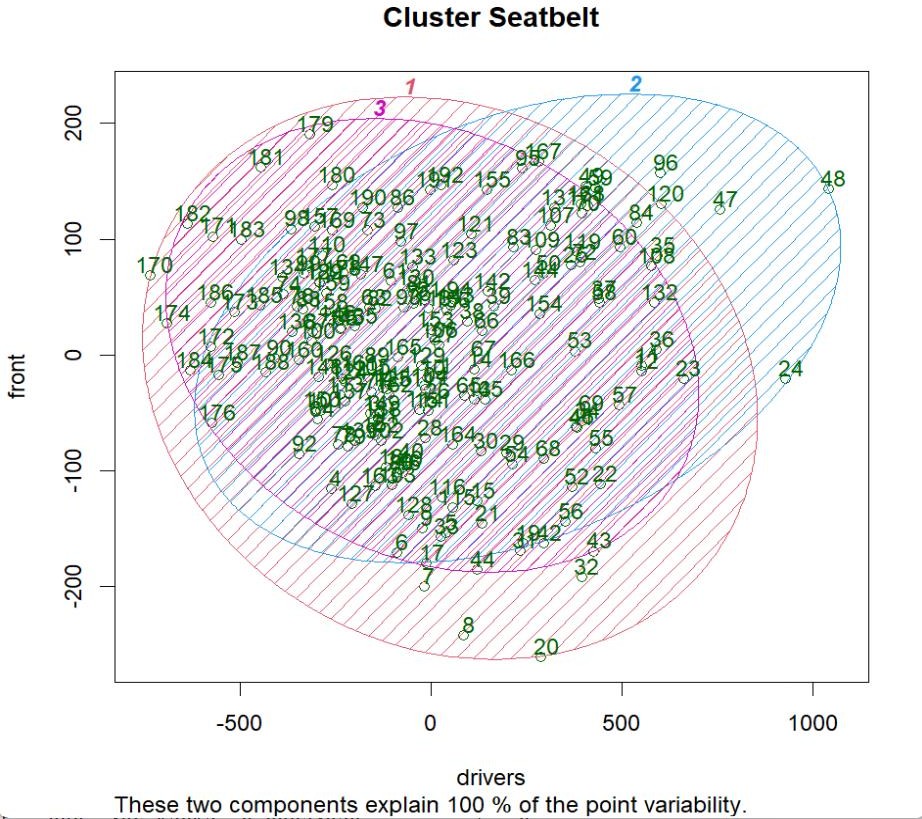
**# Plot drivers vs front for Seatbelts dataset with cluster colors**



## # Plot drivers vs front for Seatbelts dataset with cluster colors and main title





**# Plot a cluster plot of drivers vs front for Seatbelts dataset**