SecondTerm

Ashmita Bhatta

2024-05-31

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
#----------------------------Q6---------------------------------
#library(ggplot2)
#set.seed(4)
#dataset <- c(c(10:99), c(factor(male,female)), c(factor(noeducation, primary, secondary, Beyound secon
 #            c(factor(low, middle, high)), c(14:38), rnorm = 200)
#colnames <- c("Age", "Sex", "education levels", "socio-economic status", "body mass index")
```

```
#---------------------------------Q7-----------------
library(lawstat)
```

```
## Warning: package 'lawstat' was built under R version 4.3.3
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.3.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.3.3

##
## Attaching package: 'car'

## The following object is masked from 'package:lawstat':
##
##     levene.test
```

```
data("airquality")
str(airquality)
```

```
## 'data.frame':    153 obs. of  6 variables:
##  $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
##  $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
##  $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
##  $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
##  $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

```
table(airquality$Temp)
```

```
##
## 56 57 58 59 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82
##  1  3  2  2  3  2  1  2  2  3  4  4  3  1  3  3  5  4  4  9  7  6  6  5 11  9
## 83 84 85 86 87 88 89 90 91 92 93 94 96 97
##  4  5  5  7  5  3  2  3  2  5  3  2  1  1
```

```
d<- factor(airquality$Temp)
leveneTest(Temp~d, data = airquality)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group  39     NaN    NaN
##       113
```

```
summary(aov(Temp~d, data = airquality))
```

```
##              Df Sum Sq Mean Sq   F value Pr(>F)
## d            39  13618   349.2 1.711e+29 <2e-16 ***
## Residuals   113      0     0.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
shapiro_results <- by(airquality$Temp, airquality$Month, shapiro.test)
shapiro_results
```

```
## airquality$Month: 5
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.94771, p-value = 0.1349
##
## ------------------------------------------------------------
## airquality$Month: 6
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.97158, p-value = 0.5832
##
## ------------------------------------------------------------
```

```
## airquality$Month: 7
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.94579, p-value = 0.1194
##
## ------------------------------------------------------------
## airquality$Month: 8
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.96391, p-value = 0.3688
##
## ------------------------------------------------------------
## airquality$Month: 9
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.9513, p-value = 0.1831
```

```
#Interpretation: The Shapiro-Wilk test assesses whether the data within each month's group
#follows a normal distribution. A p-value greater than the chosen significance level (commonly 0.05)
#indicates that the data follows normal distribution, leading to rejection of the null hypothesis of no

library(lawstat)
library(car)

# Perform the one-way ANOVA
anova_model <- aov(Temp ~ Month, data = airquality)

# Print ANOVA table
summary(anova_model)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Month         1   2413  2413.0   32.52 6.03e-08 ***
## Residuals   151  11205    74.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Interpretation: The ANOVA table provides information on the significance of
#the Month variable in explaining the variation in the Temp variable.
#The p-value associated with the Month variable indicates whether there are
#significant differences Temp and month data.
```

```
#-----------------Q8-----------------------
library(car)
library(lawstat)
data("Arrests")
str(Arrests)
```

```
## 'data.frame':    5226 obs. of  8 variables:
##  $ released: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 2 2 2 ...
##  $ colour  : Factor w/ 2 levels "Black","White": 2 1 2 1 1 1 2 2 1 2 ...
##  $ year    : int  2002 1999 2000 2000 1999 1998 1999 1998 2000 2001 ...
##  $ age     : int  21 17 24 46 27 16 40 34 23 30 ...
##  $ sex     : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 2 1 2 2 ...
##  $ employed: Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 1 2 2 2 ...
##  $ citizen : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ checks  : int  3 3 3 1 1 0 0 1 4 3 ...
```

```
head(Arrests)
```

```
##   released colour year age    sex employed citizen checks
## 1      Yes  White 2002  21   Male      Yes     Yes      3
## 2       No  Black 1999  17   Male      Yes     Yes      3
## 3      Yes  White 2000  24   Male      Yes     Yes      3
## 4       No  Black 2000  46   Male      Yes     Yes      1
## 5      Yes  Black 1999  27 Female      Yes     Yes      1
## 6      Yes  Black 1998  16 Female      Yes     Yes      0
```

```
ind <- sample(2, nrow(Arrests), replace = T, prob = c(0.8,0.2))

train.data <- Arrests[ind ==1, ]
test.data <- Arrests[ind == 2, ]

#library(lawstat)
# Fit Naive Bayes model
#naive_bayes_model <- naiveBayes(released ~ .,
 #                               data = train.data)

# Fit Support Vector Machine (SVM) model
#svm_model <- svm(released ~ .,
 #                data = train.data)
```

```
#-----------------Q9--------------------------
city_distances <- matrix(c(
  0, 587, 1212, 701, 1936, 604, 748, 2139, 2182, 543,
  587, 0, 920, 940, 1745, 1188, 713, 1858, 1737, 597,
  1212, 920, 0, 879, 831, 1726, 1611, 1949, 2204, 1494,
  701, 940, 879, 0, 1374, 968, 1420, 1645, 1891, 1220,
  1936, 1745, 831, 1374, 0, 2339, 2451, 347, 2734, 2300,
  604, 1188, 1726, 968, 2339, 0, 1092, 2594, 2408, 923,
  748, 713, 1611, 1420, 2451, 1092, 0, 2571, 678, 205,
  2139, 1858, 1949, 1645, 347, 2594, 2571, 0, 678, 2442,
  2182, 1737, 2204, 1891, 2734, 2408, 678, 678, 0, 2329,
  543, 597, 1494, 1220, 2300, 923, 205, 2442, 2329, 0
), nrow = 10, byrow = TRUE)

# Assign row and column names
city_names <- c("Atlanta", "Chicago", "Denver", "Houston", "Los Angeles", "Miami",
                "New York", "San Francisco", "Seattle", "Washington D.C.")
rownames(city_distances) <- city_names
colnames(city_distances) <- city_names
```

```r
city.dissimilarity <- as.dist(city_distances)

# Fit the classical MDS model
mds_model <- cmdscale(city.dissimilarity, eig = TRUE, k = 2)

mds_coords <- mds_model$points
print(mds_coords)
```

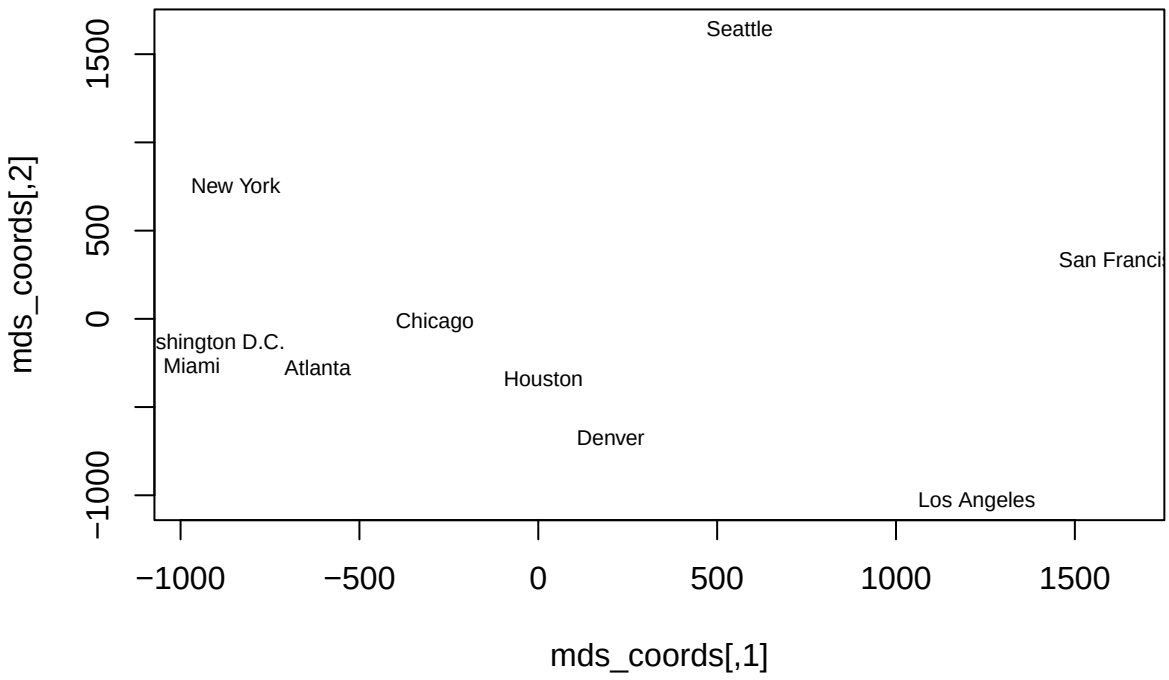
```
##                       [,1]        [,2]
## Atlanta         -616.46326  -277.03319
## Chicago         -288.61063   -22.16151
## Denver           202.61148  -672.61019
## Houston           14.25242  -335.54496
## Los Angeles     1225.78174 -1033.78934
## Miami           -968.45797  -264.31832
## New York        -845.50822   757.66327
## San Francisco   1645.58380   339.92746
## Seattle          563.12009  1646.43854
## Washington D.C. -932.30945  -138.57175
```

```r
summary(mds_model)
```

```
##        Length Class  Mode
## points 20     -none- numeric
## eig    10     -none- numeric
## x       0     -none- NULL
## ac      1     -none- numeric
## GOF     2     -none- numeric
```

```r
#Interpretation : We can interpret that the mds_model dataset has the dissimilarity of different cities

plot(mds_coords, type = "n")
text(mds_coords, labels = city_names, cex = 0.7)
title("Classical MDS of US Cities")
```




```
#Interpretation : It shows the distance of the cities with dissimilarities.
```

```
#----------------Q10---------------------------
library(cluster)
```

```
## Warning: package 'cluster' was built under R version 4.3.3
```

```
library(ClusterR)
```

```
## Warning: package 'ClusterR' was built under R version 4.3.3
```

```
iris <- read.csv("iris.csv")
str(iris)
```

```
## 'data.frame':    150 obs. of  6 variables:
##  $ X           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : chr  "Iris-setosa" "Iris-setosa" "Iris-setosa" "Iris-setosa" ...
```

```r
iris1 <- iris[, -1]
str(iris1)
```

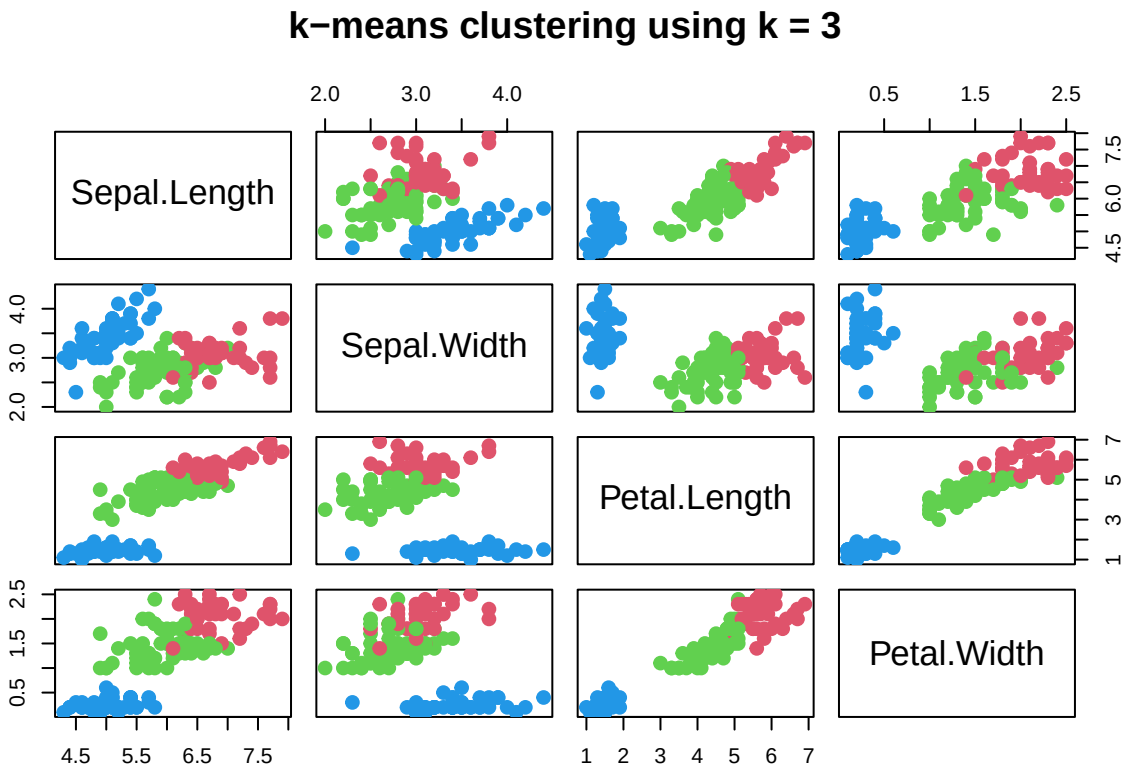
```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : chr  "Iris-setosa" "Iris-setosa" "Iris-setosa" "Iris-setosa" ...
```

```r
iris2 <- iris1[, -5]

set.seed(4)
km2 <- kmeans(iris2, 2, nstart = 20)
km3 <- kmeans(iris2, 3, nstart = 20)
km3$cluster
```

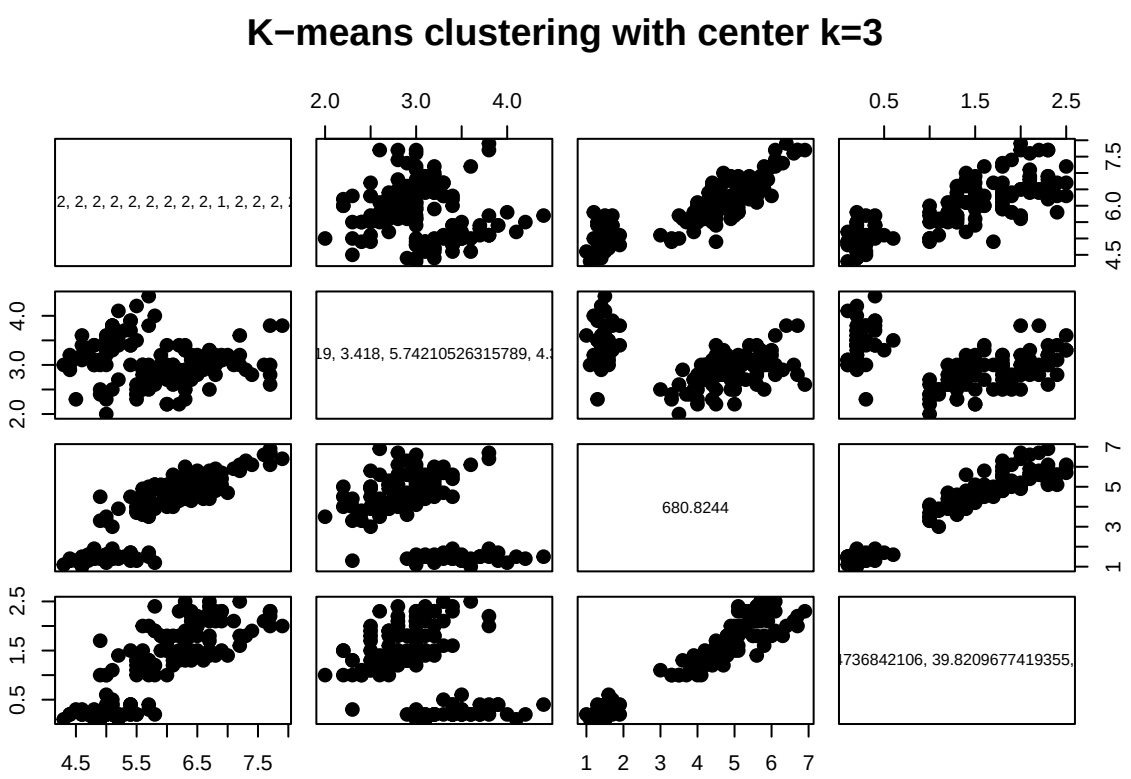
```
##   [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##  [38] 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [75] 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 1 1 1 1 2 1 1 1 1
## [112] 1 1 2 2 1 1 1 1 2 1 2 1 2 1 1 2 2 1 1 1 1 1 2 1 1 1 1 2 1 1 1 2 1 1 1 2 1
## [149] 1 2
```

```r
plot(iris2, col = (km3$cluster + 1), main = "k-means clustering using k = 3",
     pch = 20, cex = 2)
```



```
kmeans3 <- kmeans(iris2, centers = 3, nstart = 20)

plot(iris2, kmeans3, main = "K-means clustering with center k=3",
     pch = 20, cex = 2)
```




```
iris1$Species
```

```
##   [1] "Iris-setosa"     "Iris-setosa"     "Iris-setosa"     "Iris-setosa"
##   [5] "Iris-setosa"     "Iris-setosa"     "Iris-setosa"     "Iris-setosa"
##   [9] "Iris-setosa"     "Iris-setosa"     "Iris-setosa"     "Iris-setosa"
##  [13] "Iris-setosa"     "Iris-setosa"     "Iris-setosa"     "Iris-setosa"
##  [17] "Iris-setosa"     "Iris-setosa"     "Iris-setosa"     "Iris-setosa"
##  [21] "Iris-setosa"     "Iris-setosa"     "Iris-setosa"     "Iris-setosa"
##  [25] "Iris-setosa"     "Iris-setosa"     "Iris-setosa"     "Iris-setosa"
##  [29] "Iris-setosa"     "Iris-setosa"     "Iris-setosa"     "Iris-setosa"
##  [33] "Iris-setosa"     "Iris-setosa"     "Iris-setosa"     "Iris-setosa"
##  [37] "Iris-setosa"     "Iris-setosa"     "Iris-setosa"     "Iris-setosa"
##  [41] "Iris-setosa"     "Iris-setosa"     "Iris-setosa"     "Iris-setosa"
##  [45] "Iris-setosa"     "Iris-setosa"     "Iris-setosa"     "Iris-setosa"
##  [49] "Iris-setosa"     "Iris-setosa"     "Iris-versicolor" "Iris-versicolor"
##  [53] "Iris-versicolor" "Iris-versicolor" "Iris-versicolor" "Iris-versicolor"
##  [57] "Iris-versicolor" "Iris-versicolor" "Iris-versicolor" "Iris-versicolor"
##  [61] "Iris-versicolor" "Iris-versicolor" "Iris-versicolor" "Iris-versicolor"
##  [65] "Iris-versicolor" "Iris-versicolor" "Iris-versicolor" "Iris-versicolor"
##  [69] "Iris-versicolor" "Iris-versicolor" "Iris-versicolor" "Iris-versicolor"
```

```
##  [73] "Iris-versicolor" "Iris-versicolor" "Iris-versicolor" "Iris-versicolor"
##  [77] "Iris-versicolor" "Iris-versicolor" "Iris-versicolor" "Iris-versicolor"
##  [81] "Iris-versicolor" "Iris-versicolor" "Iris-versicolor" "Iris-versicolor"
##  [85] "Iris-versicolor" "Iris-versicolor" "Iris-versicolor" "Iris-versicolor"
##  [89] "Iris-versicolor" "Iris-versicolor" "Iris-versicolor" "Iris-versicolor"
##  [93] "Iris-versicolor" "Iris-versicolor" "Iris-versicolor" "Iris-versicolor"
##  [97] "Iris-versicolor" "Iris-versicolor" "Iris-versicolor" "Iris-versicolor"
## [101] "Iris-virginica"  "Iris-virginica"  "Iris-virginica"  "Iris-virginica"
## [105] "Iris-virginica"  "Iris-virginica"  "Iris-virginica"  "Iris-virginica"
## [109] "Iris-virginica"  "Iris-virginica"  "Iris-virginica"  "Iris-virginica"
## [113] "Iris-virginica"  "Iris-virginica"  "Iris-virginica"  "Iris-virginica"
## [117] "Iris-virginica"  "Iris-virginica"  "Iris-virginica"  "Iris-virginica"
## [121] "Iris-virginica"  "Iris-virginica"  "Iris-virginica"  "Iris-virginica"
## [125] "Iris-virginica"  "Iris-virginica"  "Iris-virginica"  "Iris-virginica"
## [129] "Iris-virginica"  "Iris-virginica"  "Iris-virginica"  "Iris-virginica"
## [133] "Iris-virginica"  "Iris-virginica"  "Iris-virginica"  "Iris-virginica"
## [137] "Iris-virginica"  "Iris-virginica"  "Iris-virginica"  "Iris-virginica"
## [141] "Iris-virginica"  "Iris-virginica"  "Iris-virginica"  "Iris-virginica"
## [145] "Iris-virginica"  "Iris-virginica"  "Iris-virginica"  "Iris-virginica"
## [149] "Iris-virginica"  "Iris-virginica"
```

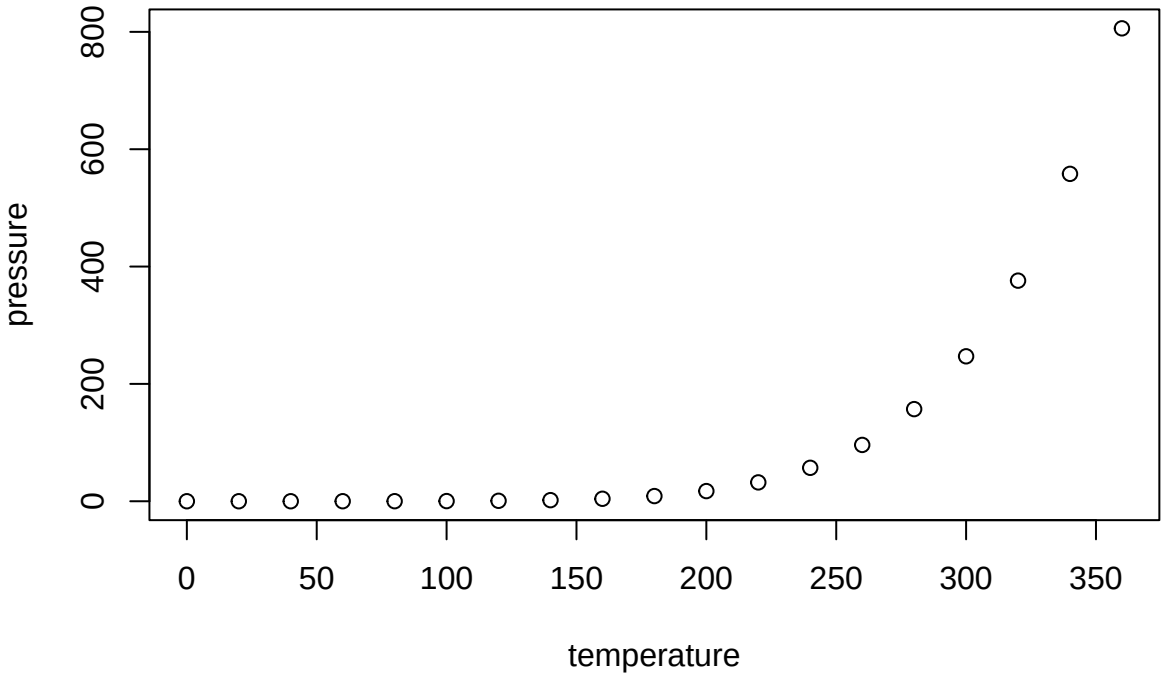
```
cm <- table(iris1$Species, kmeans3$cluster)
cm
```

```
##
##                   1  2  3
##   Iris-setosa     0  0 50
##   Iris-versicolor  2 48  0
##   Iris-virginica  36 14  0
```

```
#Interpretation : The confusion matrix cm gives the matrix of the species of iris fliwer with
#the gropuing of 3 clusters in respect to how many times it was repeated in the column.
#This means that the speecies of iris when divided into 3 clusters will give the respective confusion m
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.