#import dataset iris.txt as space as Seperator

> data(iris)

 # Print value of iris

> str(iris)

'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 : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...

> print(summary(iris))

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

>

>

> # Identify the value of K

> # make this example reproducible , use Seed

> set.seed(1234)

> **# K max value to 10 to draw the plot between k and within class sum of square(wcss)**

> k.max <- 10

> data <- as.matrix(iris[,1:4])

> wcss <- sapply(1:k.max,function(k){kmeans(data,k,nstart=6,iter.max=18)$tot.withinss})

> wcss

[1] 681.37060 152.34795 78.85144 57.26562 46.44618 41.70442 34.29823 29.98894 27.78609 26.41374

>

> **# Plot wcss for k = 2 ..4..6..8**

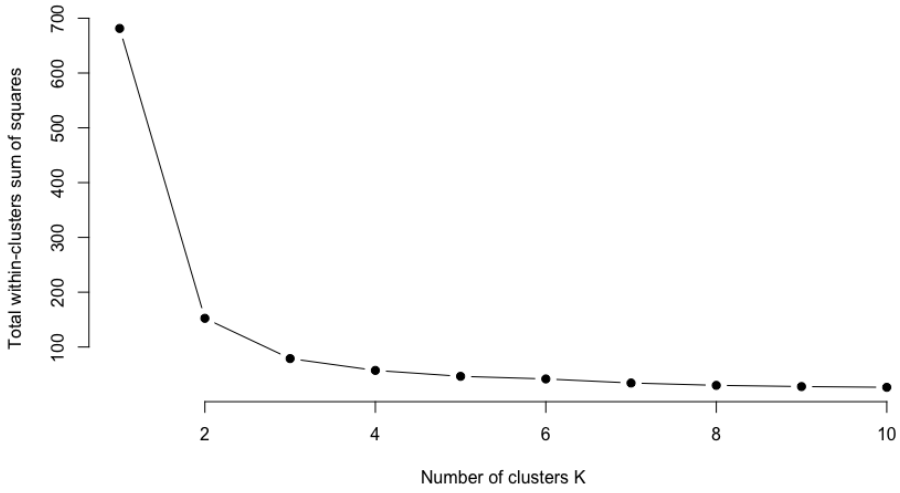
> plot(1:k.max, wcss,

+ type="b", pch = 19, frame = FALSE,

+ xlab="Number of clusters K",

+ ylab="Total within-clusters sum of squares")

>



> **# List the within-class sum of squares for different sets of predictors**

> # From the plot , ideal value would be 2 but from data, have noticed three classess of flowers so select K =3

> find_predictors = function(d){kmeans(d,3,nstart=6,iter.max=18)$tot.withinss}

>

> find_predictors(iris[,1]) #15.8

[1] 15.81662

> find_predictors(iris[,2]) # 5.34

[1] 5.535155

> find_predictors(iris[,3]) #24.66

```
[1] 24.51643
> find_predictors(iris[,4]) #4.91
[1] 4.913174
>
> find_predictors(iris[,1:2]) #37.05
[1] 37.0507
> find_predictors(iris[,2:3]) #40.73
[1] 40.73707
> find_predictors(iris[,3:4]) #31.37
[1] 31.37136
>
> find_predictors(iris[,c(1,3)]) #53.80
[1] 53.80998
> find_predictors(iris[,c(1,4)]) #32.73
[1] 32.73348
> find_predictors(iris[,c(2,3)]) #40.73
[1] 40.73707
> find_predictors(iris[,c(2,4)]) #20.60
[1] 20.6024
>
> find_predictors(iris[,1:3]) #69.42
[1] 69.42974
> find_predictors(iris[,2:4]) #47.86
[1] 47.86643
> find_predictors(iris[,1:4]) #78.85
[1] 78.85144
>
> find_predictors(iris[,c(1,2,4)])# 48.66
[1] 48.66078
```

```
>
> # Based on above predictions, best clustering value 78.85 found for 1:4, using this value for Best Model
> # Suggested K value is 3
>
> best_model <-kmeans(iris[,1:4],3,nstart=6,iter.max=18)
> table(iris[,5], best_model$cluster)

            1  2  3
  setosa     50  0  0
  versicolor  0  2 48
  virginica   0 36 14
>
>
> # using the best model , 1 , 2, 3
> #Setosa 50, 0 , 0
> #VersiColor 0 , 2 , 48
> #Virginica 0 , 36, 14
>
```

R-SCript Only

```
#import dataset iris.txt as space as seperator
data(iris)

str(iris)
print(summary(iris))
```

```
# Identify the value of K

# make this example reproducible

set.seed(1234)


k.max <- 10

data <- as.matrix(iris[,1:4])

wcss <- sapply(1:k.max,function(k){kmeans(data,k,nstart=6,iter.max=18)$tot.withinss})

wcss


# Plot wcss for k = 2 ..4..6..8

plot(1:k.max, wcss,

    type="b", pch = 19, frame = FALSE,

    xlab="Number of clusters K",

    ylab="Total within-clusters sum of squares")


# List the  within-class sum of squares for different sets of predictors

# From the plot , ideal value would be 2 but from data, have noticed three classess of flowers so select K
=3

find_predictors = function(d){kmeans(d,3,nstart=6,iter.max=18)$tot.withinss}


find_predictors(iris[,1]) #15.8

find_predictors(iris[,2]) # 5.34

find_predictors(iris[,3]) #24.66

find_predictors(iris[,4]) #4.91


find_predictors(iris[,1:2]) #37.05

find_predictors(iris[,2:3]) #40.73

find_predictors(iris[,3:4]) #31.37
```

```
find_predictors(iris[,c(1,3)]) #53.80
find_predictors(iris[,c(1,4)]) #32.73
find_predictors(iris[,c(2,3)]) #40.73
find_predictors(iris[,c(2,4)]) #20.60

find_predictors(iris[,1:3]) #69.42
find_predictors(iris[,2:4]) #47.86
find_predictors(iris[,1:4]) #78.85

find_predictors(iris[,c(1,2,4)])# 48.66

# Best clustering value 78.85 found for 1:4, using this value for Best Model
# Suggested K value is 3

best_model <-kmeans(iris[,1:4],3,nstart=6,iter.max=18)
table(iris[,5], best_model$cluster)
```