K-Means Clustering

Mr. Gulati

July 24, 2020

```
library("NbClust")
library("ggplot2")
library("factoextra")
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
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
```

Reading the data

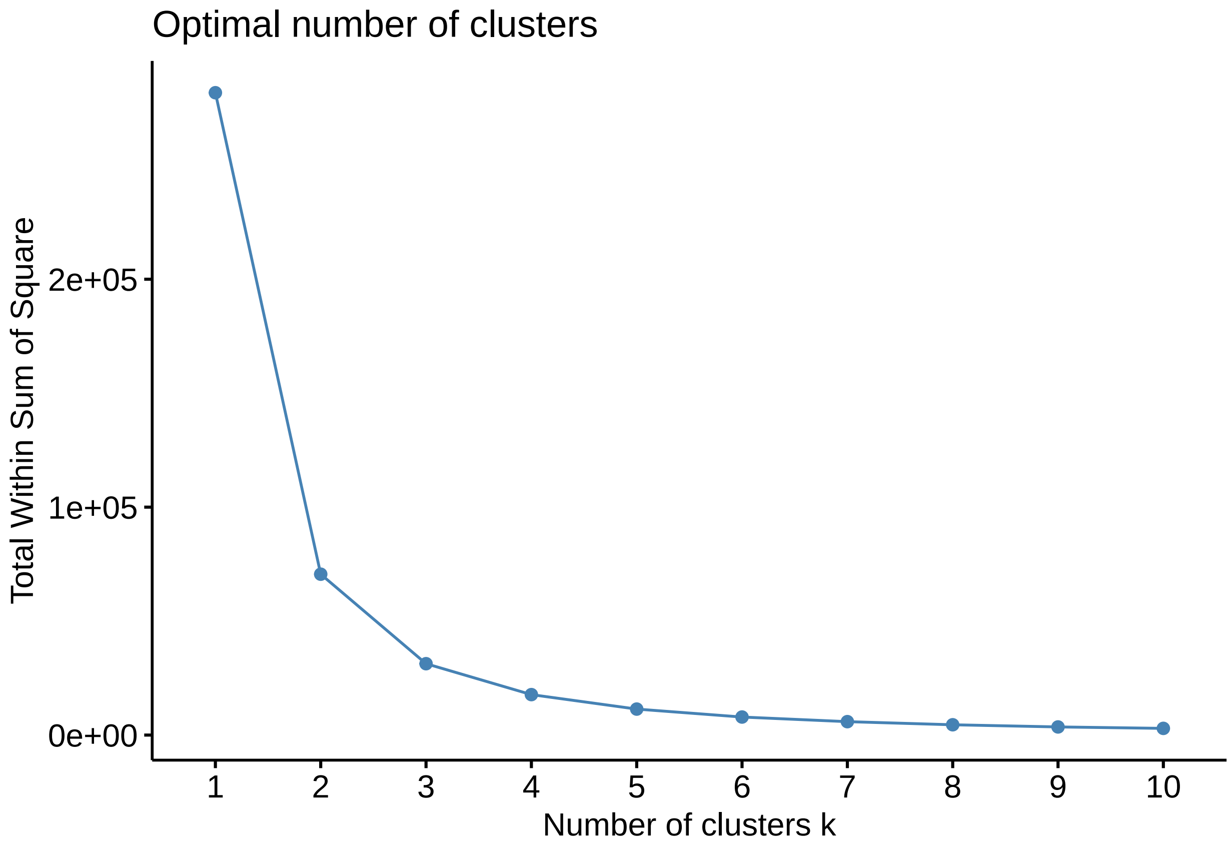
```
dataset <- read.csv("D:/Internship/Task 2/Iris.csv",sep = ",", header = TRUE)
dataset_df <- data.frame(dataset)
dataset_df <- na.omit(dataset_df)
head(dataset_df)
```

```
##   Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm     Species
## 1  1           5.1          3.5           1.4          0.2 Iris-setosa
## 2  2           4.9          3.0           1.4          0.2 Iris-setosa
## 3  3           4.7          3.2           1.3          0.2 Iris-setosa
## 4  4           4.6          3.1           1.5          0.2 Iris-setosa
## 5  5           5.0          3.6           1.4          0.2 Iris-setosa
## 6  6           5.4          3.9           1.7          0.4 Iris-setosa
```

Finding optimum number of clusters

```
set.seed(100)

fviz_nbclust(dataset_df[,c(1,2,3,4)], kmeans, method = "wss")
```



The results suggests the bend appears at k=3

Applying kmeans

```
model <- kmeans(dataset_df[,c(1,2,3,4)], 3, nstart = 25)
model
```

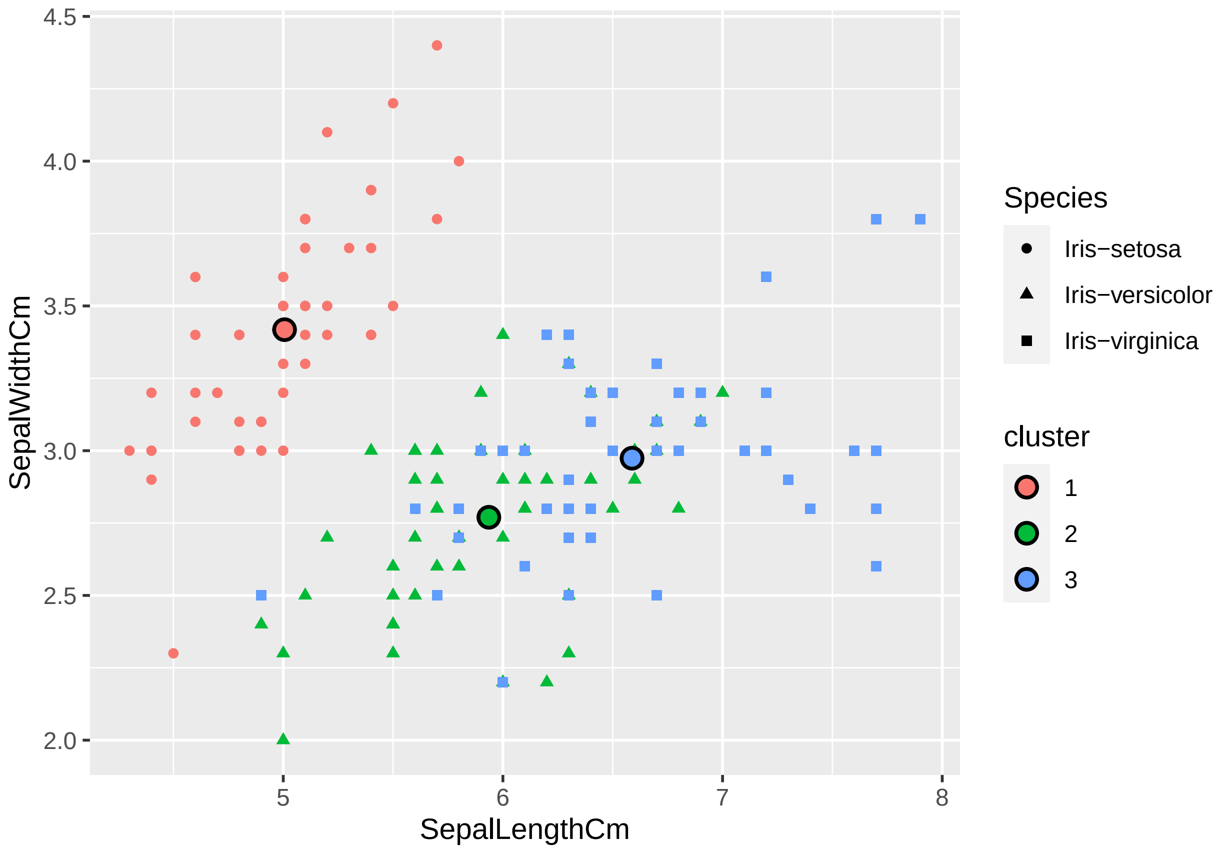
```
## K-means clustering with 3 clusters of sizes 50, 50, 50
##
## Cluster means:
##      Id SepalLengthCm SepalWidthCm PetalLengthCm
## 1  25.5         5.006        3.418         1.464
## 2  75.5         5.936        2.770         4.260
## 3 125.5         6.588        2.974         5.552
##
## Clustering vector:
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
##   1   1   1   1   1   1   1   1   1   1   2   2   2   2   2   2   2   2   2   2
##  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
##  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
##   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3
## 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
##   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3
## 141 142 143 144 145 146 147 148 149 150
##   3   3   3   3   3   3   3   3   3   3
##
## Within cluster sum of squares by cluster:
## [1] 10427.18 10441.20 10452.33
##  (between_SS / total_SS =  88.9 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Visualising Clusters

```
iris_clustered <- data.frame(dataset_df, cluster=factor(model$cluster))

centroids <- data.frame(model$centers, cluster=factor(1:3))

ggplot(iris_clustered, aes(x=SepalLengthCm, y=SepalWidthCm, color=cluster, shape=Species)) + geom_point
  # individual points from the ‘iris_clustered‘ data frame
  geom_point(data=centroids, aes(fill=cluster), shape=21, color="black", size=3, stroke=1) # centroids
```


SepalWidthCm
SepalLengthCm
Species
● Iris−setosa
▲ Iris−versicolor
■ Iris−virginica
cluster
● 1
● 2
● 3