Iris Dataset K-Means Clustering Analysis

The *iris* data set iris.txt contains 150 data points, each with four predictor variables and one categorical response. The predictors are the width and length of the sepal and petal of flowers and the response is the type of flower. The data is available from the R library datasets and can be accessed with iris once the library is loaded. It is also available at the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Iris). *The response values are only given to see how well a specific method performed and should not be used to build the model.*

Use the R function kmeans to cluster the points as well as possible. Report the best combination of predictors, your suggested value of k, and how well your best clustering predicts flower type.

Solution:

Step 1: Basic data exploration:

```
# Solution for Question 4.2
#load library
#for plotting
library(ggplot2)
#to plot cluster
install.packages("cluster")
library(cluster)
#we do not have to load data from file. iris is a default dataset in R
data <- iris
#prints dataset dimension
cat("Dimensions:", dim(data), "\n")
#prints the head data (top few rows)
print(head(data))
#prints summary of dataset
print(summary(data))
```

OUTPUT

```
> cat("Dimensions:", dim(data), "\n")
Dimensions: 150 5
> #prints the head data (top few rows)
> print(head(data))
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
> #prints summary of dataset
> print(summary(data))
  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
```

Step 2: Plot the data in scatterplot before clustering

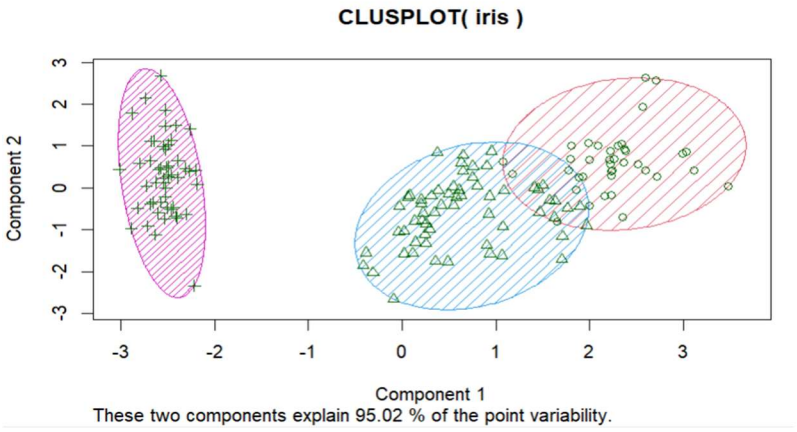
```
#plot data before clustering to find relationship between petal length & width
ggplot(data, aes(Petal.Length, Petal.Width)) + geom_point(aes(col=Species), size=4)
```

Step 3(a): For all length and width of petal and sepal - Try K=3

```
#k-means clustering (k=3) - using petal & sepal length and width
set.seed(101)
iris_clus <- kmeans(data[,1:4], center=3, nstart=20)
iris_clus
#create table to compare with original data & plot
table(iris_clus$cluster, data$Species)
#calculate accuracy for k=3
accuracy1 <- sum(diag(table(iris_clus$cluster, data$Species))) / nrow(data)
cat("Accuracy for k=3:", round(accuracy1 * 100, 2), "%\n")
clusplot(iris, iris_clus$cluster, color=T, shade=T, labels=0, lines=0)
```

OUTPUT:

```
K-means clustering with 3 clusters of sizes 38, 62, 50

Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     6.850000    3.073684     5.742105    2.071053
2     5.901613    2.748387     4.393548    1.433871
3     5.006000    3.428000     1.462000    0.246000

Clustering vector:
  [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 3 3 3 3
 [46] 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 2 2 1 2 2 2 2 2 2 2 2 2
 [91] 2 2 2 2 2 2 2 2 2 2 2 1 2 1 1 1 1 2 1 1 1 1 1 2 2 1 1 1 1 2 1 2 1 2 1 1 2 2 1 1
[136] 1 1 1 2 1 1 1 2 1 1 1 2 1 1 2

Within cluster sum of squares by cluster:
[1] 23.87947 39.82097 15.15100
 (between_SS / total_SS =  88.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
> #create table to compare with original data & plot
> table(iris_clus$cluster, data$Species)

    setosa versicolor virginica
  1      0          2        36
  2      0         48        14
  3     50          0         0
> #calculate accuracy for k=3
> accuracy1 <- sum(diag(table(iris_clus$cluster, data$Species))) / nrow(data)
> cat("Accuracy for k=3:", round(accuracy1 * 100, 2), "%\n")
Accuracy for k=3: 32 %
```

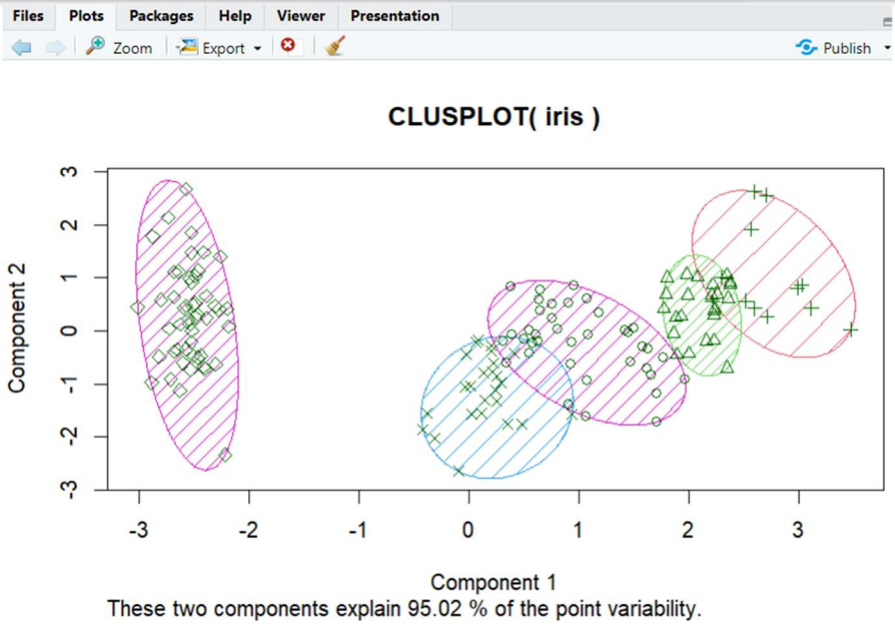


Step 3(b): For all length and width of petal and sepal - Try K=5

```
#k-means clustering (k=5) - using petal & sepal length and width
set.seed(101)
iris_clus <- kmeans(data[,1:4], center=5, nstart=20)
iris_clus
#create table to compare with original data & plot
table(iris_clus$cluster, data$Species)
#calculate accuracy for k=5
accuracy2 <- sum(diag(table(iris_clus$cluster, data$Species))) / nrow(data)
cat("Accuracy for k=5:", round(accuracy2 * 100, 2), "%\n")
clusplot(iris, iris_clus$cluster, color=T, shade=T, labels=0, lines=0)

#k-means clustering (k=5)
set.seed(101)
iris_clus <- kmeans(data[,1:4], center=5, nstart=20)
iris_clus

#create table to compare with original data
table(iris_clus$cluster, data$Species)

clusplot(iris, iris_clus$cluster, color=T, shade=T, labels=0, lines=0)
```

OUTPUT:

```
K-means clustering with 5 clusters of sizes 39, 24, 12, 25, 50

Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     6.207692    2.853846     4.746154    1.564103
2     6.529167    3.058333     5.508333    2.162500
3     7.475000    3.125000     6.300000    2.050000
4     5.508000    2.600000     3.908000    1.204000
5     5.006000    3.428000     1.462000    0.246000

Clustering vector:
  [1] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
 [46] 5 5 5 5 5 1 1 1 4 1 1 1 4 1 4 4 1 4 1 4 1 1 4 1 4 1 4 1 1 1 1 1 1 1 4 4 4 4 1 4 1 1 1 4 4
 [91] 4 1 4 4 4 4 4 1 4 4 2 1 3 2 2 3 4 3 2 3 2 2 2 1 2 2 2 3 3 1 2 1 3 1 2 3 1 1 2 3 3 3 2 1 1
[136] 3 2 2 1 2 2 2 1 2 2 2 1 2 2 1

Within cluster sum of squares by cluster:
[1] 12.81128  5.46250  4.65500  8.36640 15.15100
 (between_SS / total_SS =  93.2 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
> #create table to compare with original data & plot
> table(iris_clus$cluster, data$Species)

    setosa versicolor virginica
  1      0         26        13
  2      0          0        24
  3      0          0        12
  4      0         24         1
  5     50          0         0
> #calculate accuracy for k=5
> accuracy2 <- sum(diag(table(iris_clus$cluster, data$Species))) / nrow(data)
> cat("Accuracy for k=5:", round(accuracy2 * 100, 2), "%\n")
Accuracy for k=5: 8 %
```



Step 3(c): For all length and width of petal and sepal - Try k=7

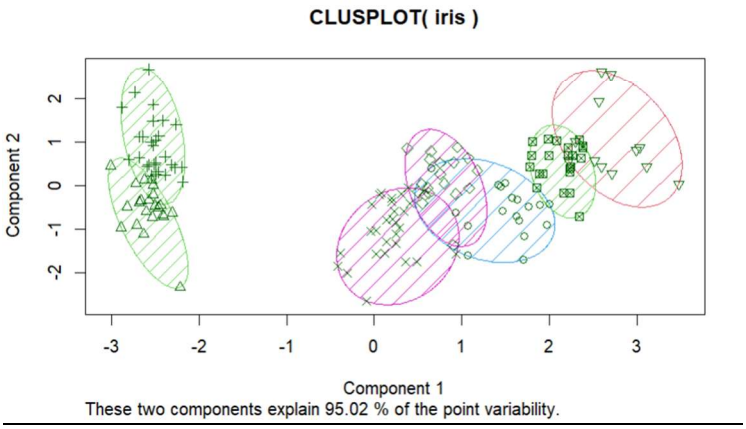
```
#k-means clustering (k=7) - using petal & sepal length and width
set.seed(101)
iris_clus <- kmeans(data[,1:4], center=7, nstart=20)
iris_clus
#create table to compare with original data & plot
table(iris_clus$cluster, data$Species)
#calculate accuracy for k=7
accuracy3 <- sum(diag(table(iris_clus$cluster, data$Species))) / nrow(data)
cat("Accuracy for k=7:", round(accuracy3 * 100, 2), "%\n")
clusplot(iris, iris_clus$cluster, color=T, shade=T, labels=0, lines=0)
```

OUTPUT

```
K-means clustering with 7 clusters of sizes 19, 22, 28, 28, 19, 12, 22

Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     6.036842    2.705263     5.000000   1.7789474
2     4.704545    3.122727     1.413636   0.2000000
3     5.242857    3.667857     1.500000   0.2821429
4     5.532143    2.635714     3.960714   1.2285714
5     6.442105    2.978947     4.594737   1.4315789
6     7.475000    3.125000     6.300000   2.0500000
7     6.568182    3.086364     5.536364   2.1636364

Clustering vector:
  [1] 3 2 2 2 3 3 2 3 2 2 3 2 2 2 3 3 3 3 3 3 3 2 3 2 2 3 3 3 2 2 3 3 3 2 3 3 2 3
 [46] 2 3 2 3 2 5 5 5 4 5 4 5 4 5 4 4 4 4 5 4 5 4 4 1 4 1 4 1 5 5 5 5 5 5 4 4 4 4 1 4
 [91] 4 5 4 4 4 4 4 5 4 4 7 1 6 7 7 6 4 6 7 6 7 1 7 1 1 7 7 6 6 1 7 1 6 1 7 6 1 1 7 6
[136] 6 7 7 1 7 7 7 1 7 7 7 1 7 7 1

Within cluster sum of squares by cluster:
[1] 4.125263 3.114091 4.630714 9.749286 3.708421 4.655000 4.315455
 (between_SS / total_SS =  95.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betwee
[7] "size"         "iter"         "ifault"
> #create table to compare with original data & plot
> table(iris_clus$cluster, data$Species)

    setosa versicolor virginica
  1      0          4        15
  2     22          0         0
  3     28          0         0
  4      0         27         1
  5      0         19         0
  6      0          0        12
  7      0          0        22
> #calculate accuracy for k=7
> accuracy3 <- sum(diag(table(iris_clus$cluster, data$Species))) / nrow(data)
> cat("Accuracy for k=7:", round(accuracy3 * 100, 2), "%\n")
Accuracy for k=7: 0 %
```

Files Plots Packages Help Viewer Presentation

Zoom Export Publish

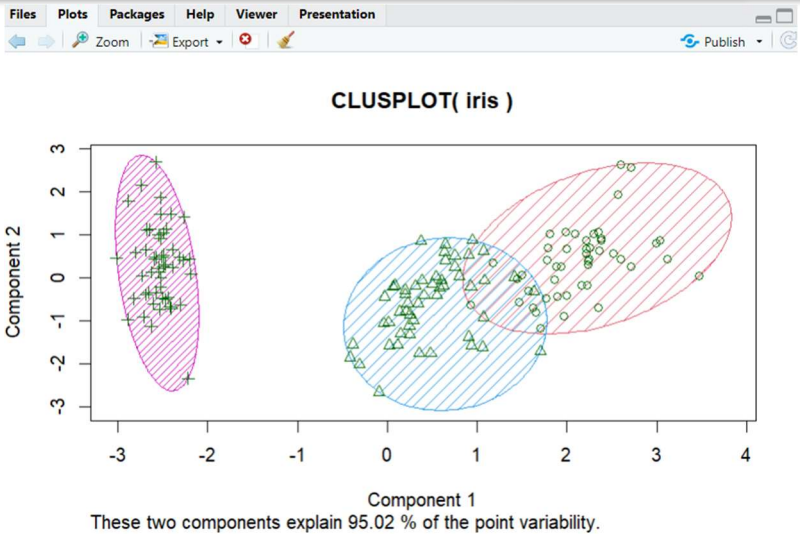


Step 4(a): For just length and width of petal only – k=3

```
#k-means clustering (k=3) - using petal length and width only
set.seed(101)
iris_clus <- kmeans(data[,3:4], center=3, nstart=20)
iris_clus
#create table to compare with original data
table(iris_clus$cluster, data$Species)
#calculate accuracy for k=3 (petal only)
confusion_matrix <- table(iris_clus$cluster, data$Species)
accuracy4 <- sum(apply(confusion_matrix, 1, max)) / nrow(data)
cat("Accuracy for k=3 (petal only):", round(accuracy4 * 100, 2), "%\n")
clusplot(iris, iris_clus$cluster, color=T, shade=T, labels=0, lines=0)
```

OUTPUT:

```
K-means clustering with 3 clusters of sizes 48, 52, 50

Cluster means:
  Petal.Length Petal.Width
1     5.595833    2.037500
2     4.269231    1.342308
3     1.462000    0.246000

Clustering vector:
  [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 3 3 3 3
 [46] 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2 2 2
 [91] 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1
[136] 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1

Within cluster sum of squares by cluster:
[1] 16.29167 13.05769  2.02200
 (between_SS / total_SS =  94.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
> #create table to compare with original data
> table(iris_clus$cluster, data$Species)

    setosa versicolor virginica
  1      0          2        46
  2      0         48         4
  3     50          0         0
> #calculate accuracy for k=3 (petal only)
> confusion_matrix <- table(iris_clus$cluster, data$Species)
> accuracy4 <- sum(apply(confusion_matrix, 1, max)) / nrow(data)
> cat("Accuracy for k=3 (petal only):", round(accuracy4 * 100, 2), "%\n")
Accuracy for k=3 (petal only): 96 %
```



Step 5: Compare the above accuracies to find best k value

```
#find maximum accuracy
all_accuracies <- c(accuracy1, accuracy2, accuracy3, accuracy4)
accuracy_names <- c("k=3 (all features)", "k=5 (all features)", "k=7 (all features)", "k=3 (petal only)")
max_accuracy <- max(accuracy1, accuracy2, accuracy3, accuracy4)
best_model <- accuracy_names[which.max(all_accuracies)]
cat("\nMaximum accuracy:", round(max_accuracy * 100, 2), "% achieved by", best_model, "\n")
```

OUTPUT:

```
> #find maximum accuracy
> all_accuracies <- c(accuracy1, accuracy2, accuracy3, accuracy4)
> accuracy_names <- c("k=3 (all features)", "k=5 (all features)", "k=7 (all features)", "k=3 (petal only)")
> max_accuracy <- max(accuracy1, accuracy2, accuracy3, accuracy4)
> best_model <- accuracy_names[which.max(all_accuracies)]
> cat("\nMaximum accuracy:", round(max_accuracy * 100, 2), "% achieved by", best_model, "\n")

Maximum accuracy: 96 % achieved by k=3 (petal only)
```

CONCLUSION:

After manually comparing different k values,
Best Predictors: Petal Length + Petal Width
Optimal K: 3
Accuracy: 96%