Homework 3

Bailey Ho | Daria Barbour-Brown | Warren Kennedy

Spring 2023

Note: All group members contributed equally.

Conceptual Problems

Question 1

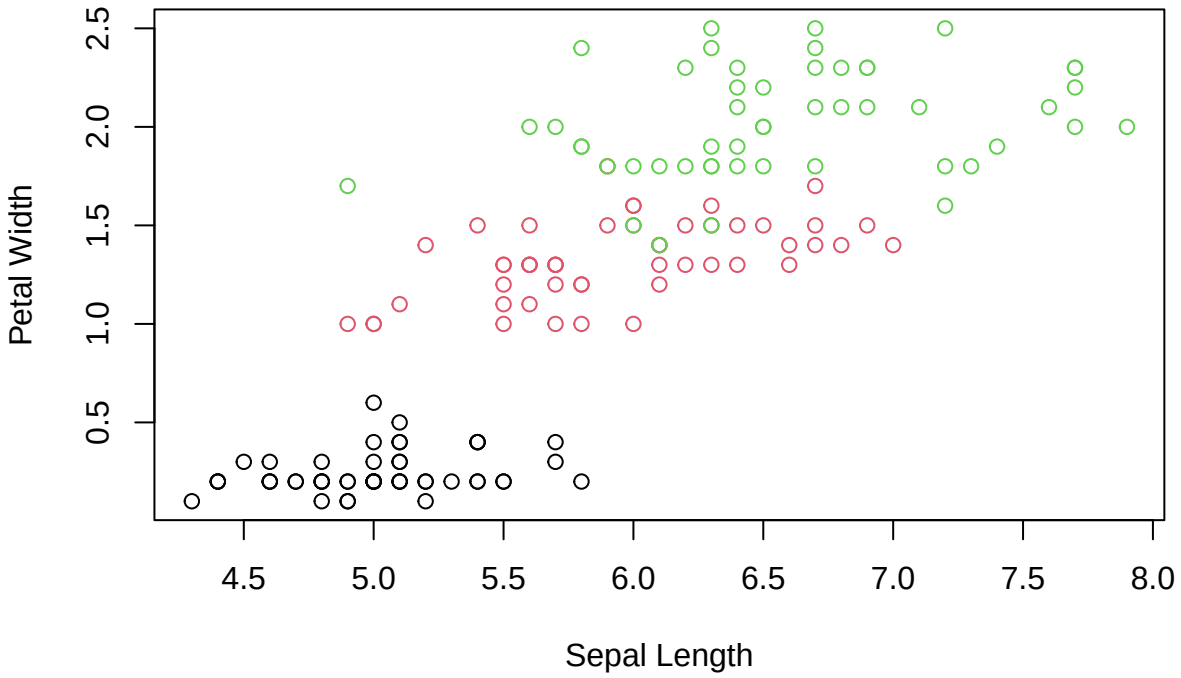
If we don't know how many clusters to organize our data in, we can take a "brute-force" approach by trying different values of K. After trying different values of K, we can use various methods to assess which value of K minimizes the within cluster variation, in Layman's terms, which describes how closely knit each cluster is. One such method is known as the "Elbow" method, where we graph the total within variation vs. K. We then observe the Kth value for which the curve begins to "flatten" telling us that after a certain K, additional clusters no longer provide us with important information.

```
library(ISLR2)
library(fpc)
library(RSSL)
```

Application Problems

Question 2

```
data <- iris
plot(data$Sepal.Length,data$Petal.Width,
  xlab = "Sepal Length", ylab = "Petal Width", col = data$Species)
```



Question 3

Part A For this calculation, we want only the columns that contain numerical data so remove the column with non-numerical data.

```
km3 <- kmeans(data[,1:4], 3, nstart = 20)
km3$withinss
```

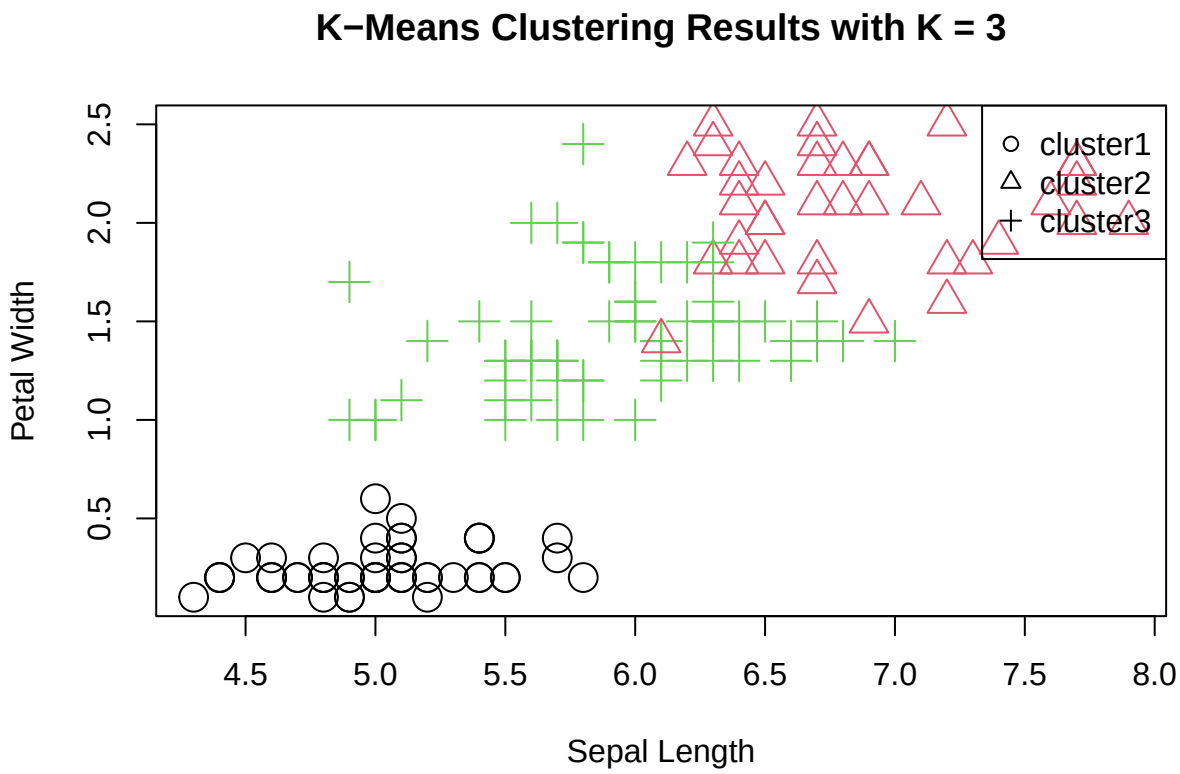
```
## [1] 15.15100 23.87947 39.82097
```

```
km3$betweenss
```

Part B

```
## [1] 602.5192
```

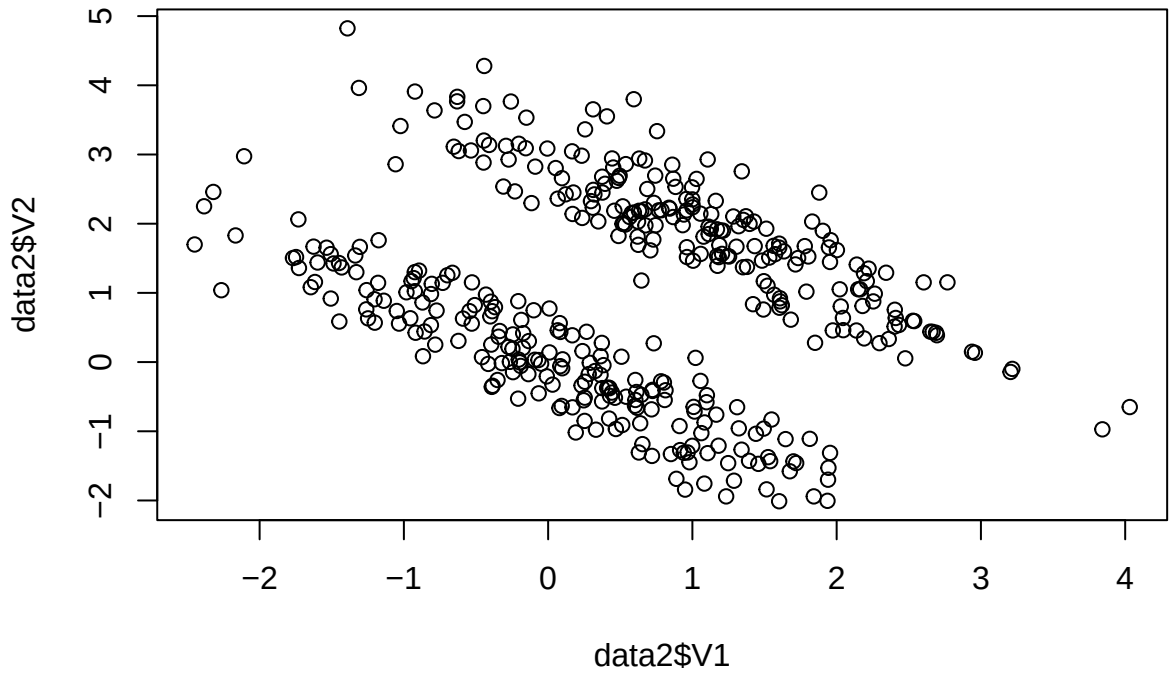
```
plot(data$Sepal.Length, data$Petal.Width, col = (km3$cluster),
     main = "K-Means Clustering Results with K = 3",xlab = "Sepal Length",
     ylab = "Petal Width", cex = 2, pch=as.numeric(km3$cluster))
legend("topright",c("cluster1","cluster2", "cluster3"), pch=c(1,2,3))
```



Part C

Question 4

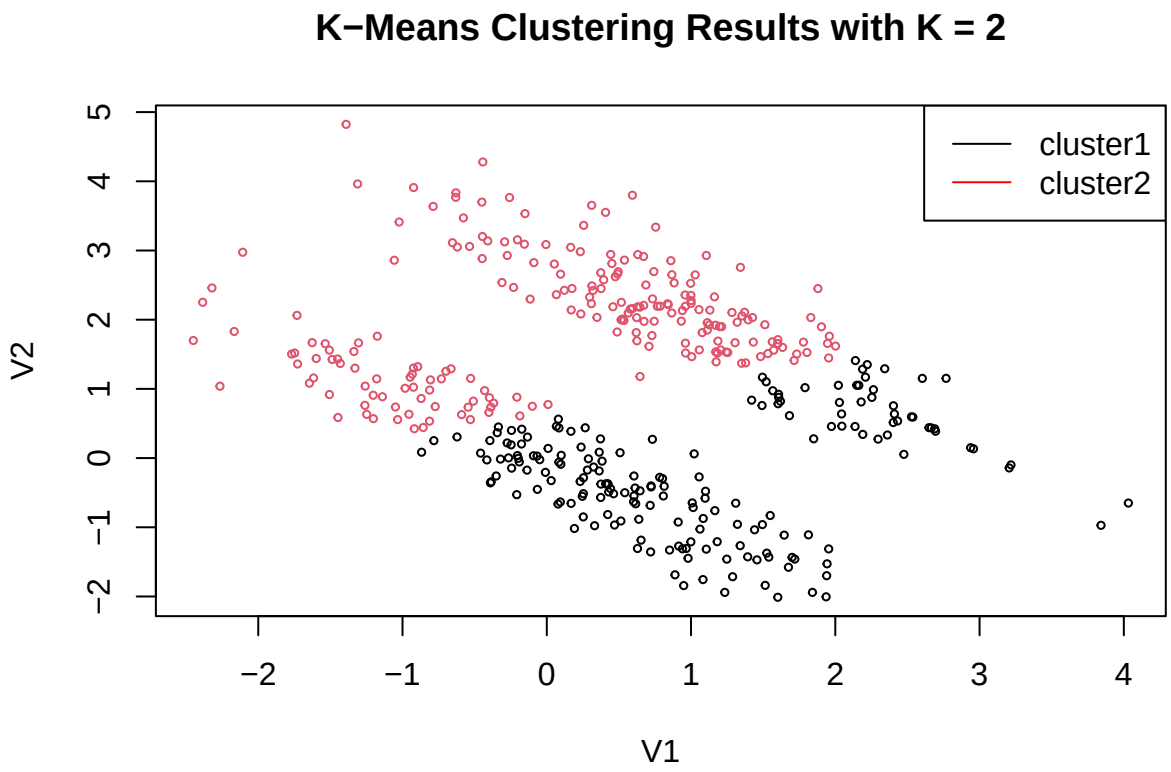
```
data2 <- read.csv("homework3_clustering.csv")
plot(data2$V1, data2$V2)
```



Part A

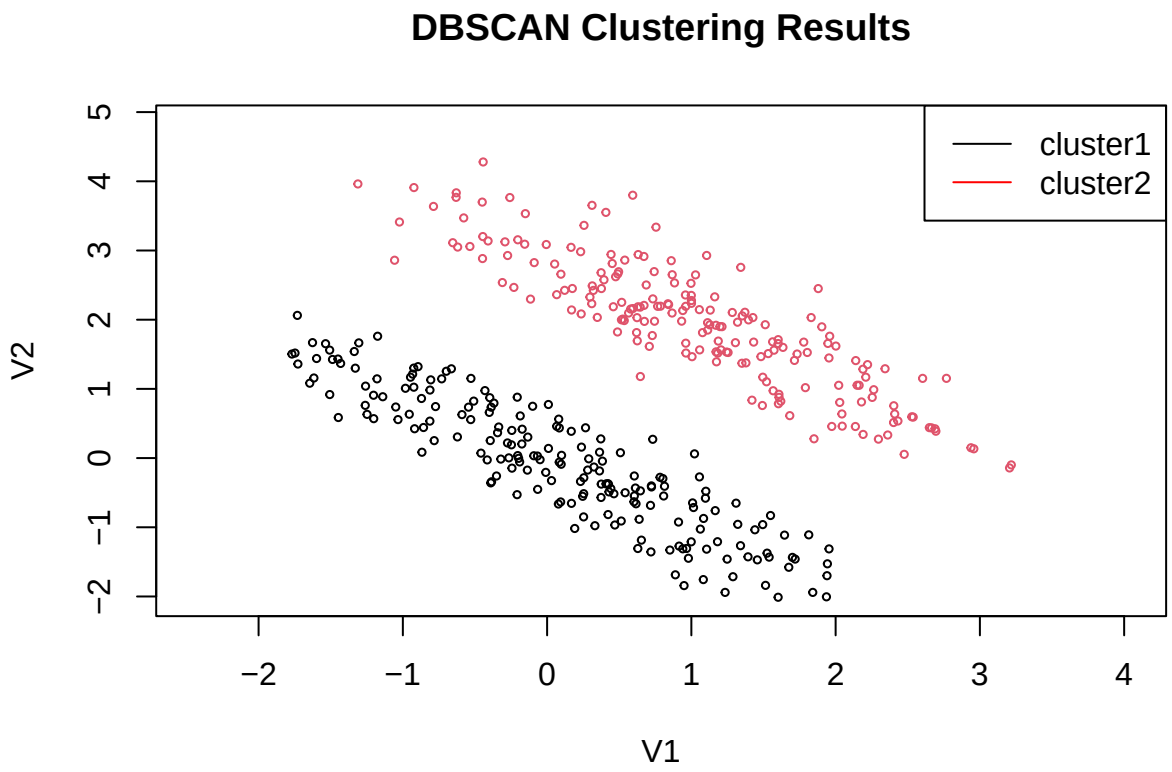
```
km2 <- kmeans(data2, 2)
```

```
plot(data2$V1, data2$V2, col = (km2$cluster),
     main = "K-Means Clustering Results with K = 2",xlab = "V1",
     ylab = "V2", cex = .5, )
legend("topright",c("cluster1","cluster2"), col=c("black", "red"), lty=1:1)
```



Part B

```
dbscan_data2 <- dbscan(data2, eps = .5, MinPts = 5)
plot(data2$V1, data2$V2, col = dbscan_data2$cluster,
     main = "DBSCAN Clustering Results", xlab = "V1", ylab = "V2", cex = .5)
legend("topright",c("cluster1","cluster2"), col=c("black", "red"), lty=1:1)
```



Part C

Part D We expected the DB Scan to give us better results due to the algorithm used to choose clusters. While the K-means algorithm attempts to cluster the data by measuring the distance of each data point to a set of central points, DB Scan clusters the data by grouping lower density areas and higher density areas separately. In this case, the density-based DB Scan method is able to better detect the distinct groupings that we can see in the data. The K-means algorithm was not able to detect this shape, yielding a less informative result.