Homework 3

Chance Kang / A13605546 / csk025@ucsd.edu
Jayden Kim / A16271107 / s0k003@ucsd.edu
Sia Sheth / A16357789 / snsheth@ucsd.edu
Math 189
Spring 2023

Conceptual Question

1. If we were to be not expecting a certain number of clusters (K), then a trial and error approach would be appropriate until an adequate analysis can be done. K-Means Clustering forms K counts of clusters that share similarities. Choosing K and analysis followed by performing K-Means Clustering would be up to what kind of data and project the users are working with. One can attempt to repeatedly use the number of clustering resulting from hierarchical clustering for K as initial trial and error method.

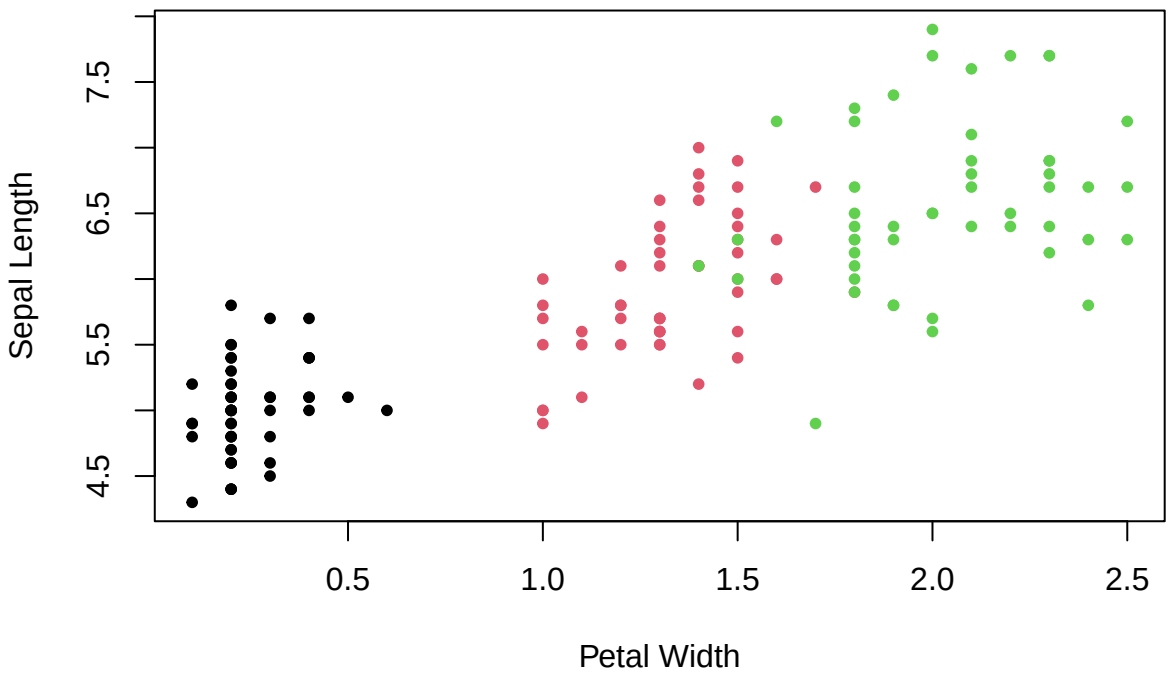
Application Question

Loading packages:

```
library(fpc)
```

2.

```
data(iris)
plot(iris$Petal.Width,iris$Sepal.Length,col=as.numeric(iris$Species),xlab ="Petal Width", ylab="Sepal L
```



3.

```
iris_k <- iris[,1:4]
kmi <- kmeans(iris_k, 3, nstart =20)
```

a.

```
kmi$withinss
```

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

```
kmi$tot.withinss
```

```
## [1] 78.85144
```

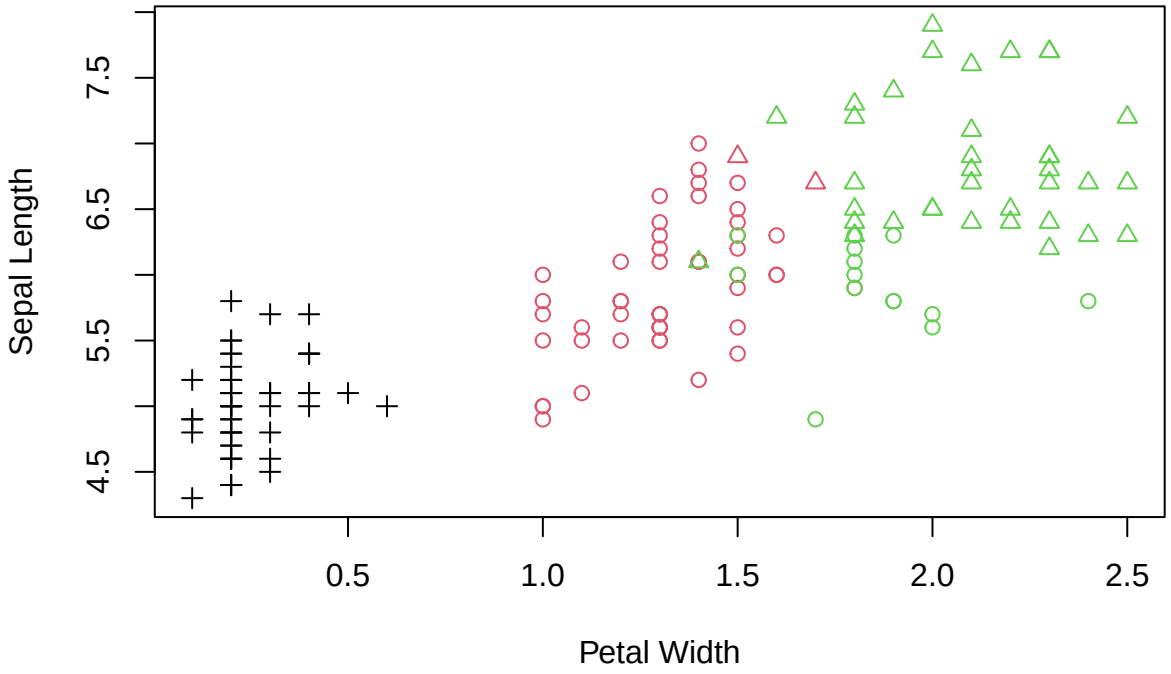
b.

```
kmi$betweenss
```

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

c.

```
plot(iris$Petal.Width, iris$Sepal.Length, col = as.numeric(iris$Species), xlab ="Petal Width", ylab="Se
```

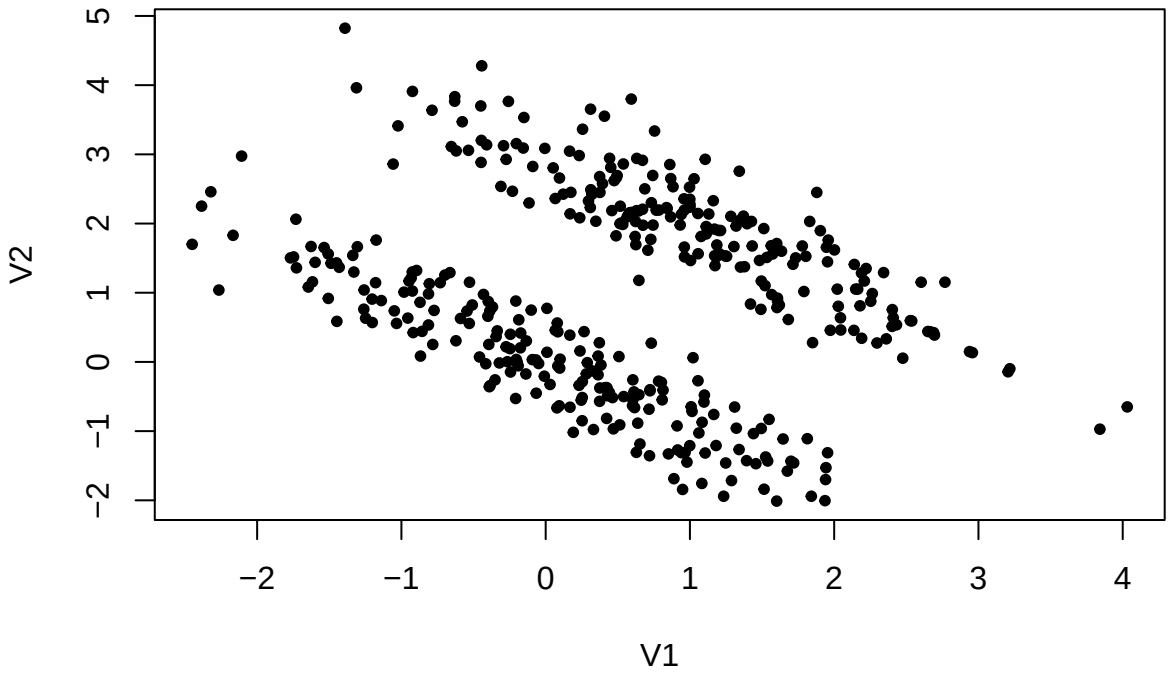


4.

```
hw3<-read.csv(file= "./homework3_clustering.csv")
```

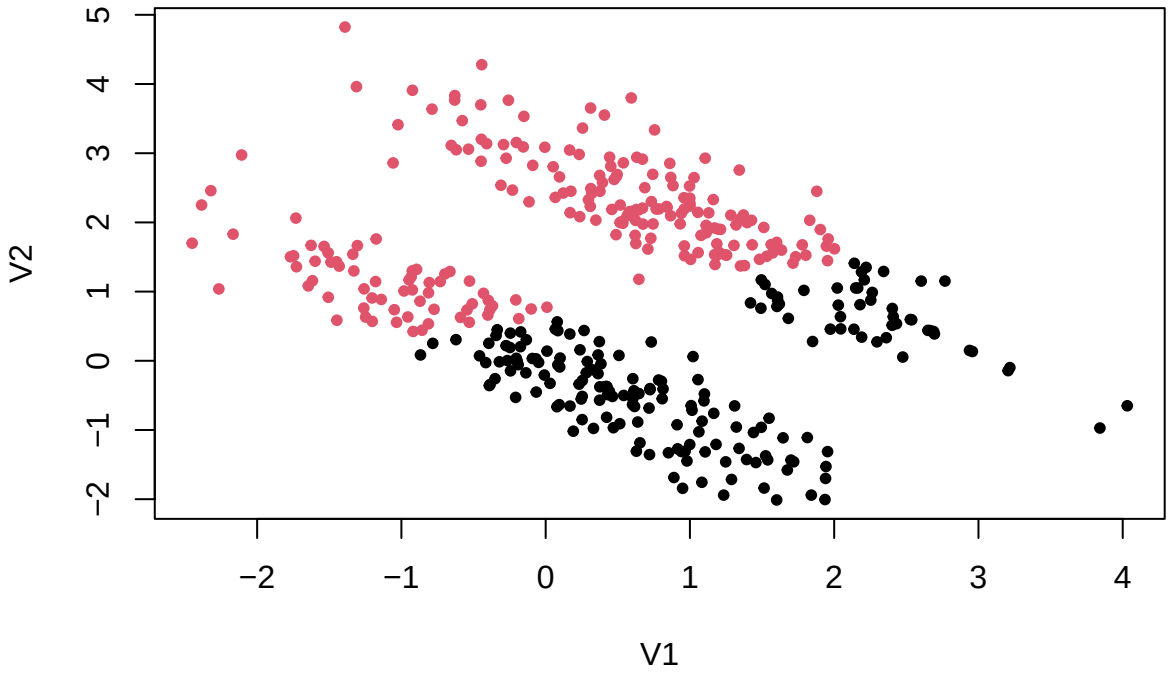
a.

```
plot(hw3$V1, hw3$V2, xlab = "V1", ylab = "V2", pch = 20)
```



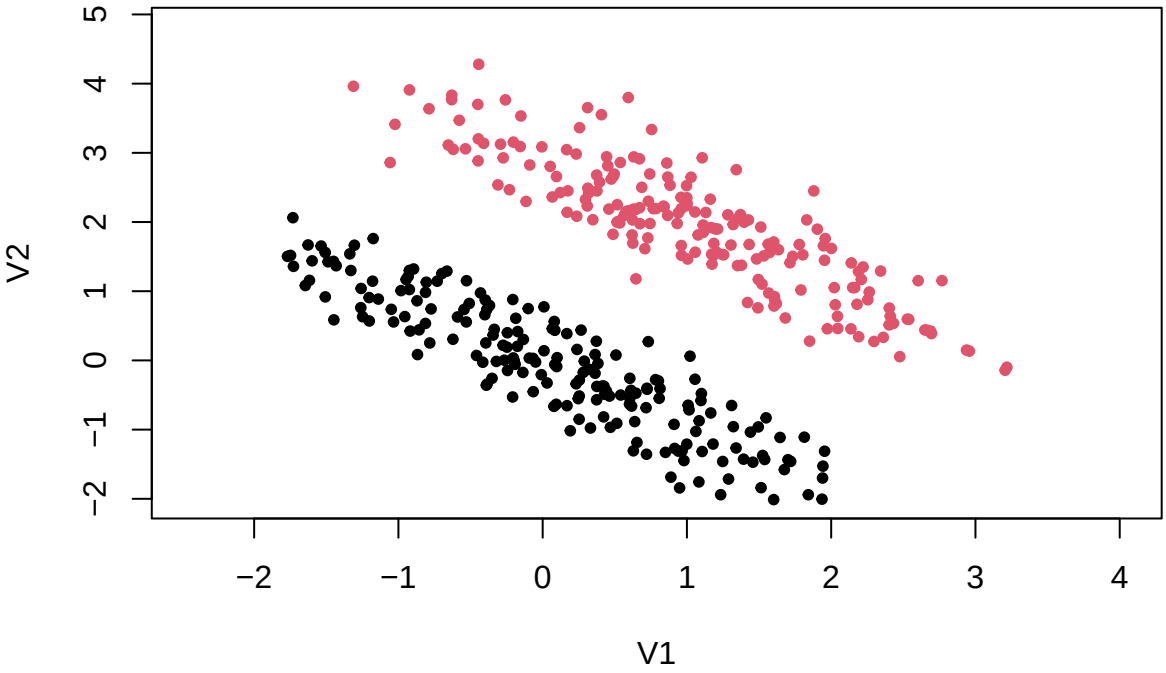
b. We chose to use $K = 2$, since we can visually see two "clusters" on the plot.

```
kmhw3 = kmeans(hw3, 2, nstart =100)
plot(hw3$V1, hw3$V2, xlab = "V1", ylab = "V2", pch = 20, col = kmhw3$cluster)
```



c.

```
dbshw3 = dbscan(hw3,eps=0.5, MinPts = 5)
plot(hw3$V1, hw3$V2, xlab = "V1", ylab = "V2", pch = 20, col = dbshw3$cluster)
```



d. The plot of the original data visually shows two distinct clusters and DBSCAN showed better results in terms of distinguishing the two clusters. The result is because K-Means picks $K = 2$ centroids that minimizes the square distance, hence the above result. However, DBSCAN picks the point to include in a cluster by choosing MinPts count of points that are within the given distance eps, hence relies on density and can distinguish boundaries on its own.

Contribution

Our group homework process goes as the following:
1. Each member attempts to complete the homework
2. Compare and discuss the answers
3. Complete a finalized version to submit

All members have contributed about the equal amount to complete this homework.