Clustering Example 6: Density-based Clustering of Iris Data

Amy Wagaman, Nathan Carter

July 2020

Load necessary libraries.

```
library(mosaic)
library(dbscan)
library(GGally)
```

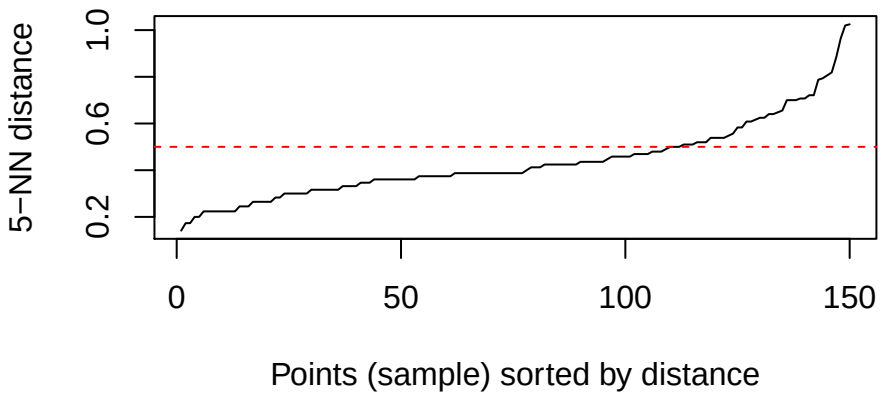
Load the famous dataset of Fisher's irises.

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

Find suitable `eps` parameter.

We use a k-nearest-neighbors plot with k equal to the number of dimensions plus 1, and look for the "knee" in the curve, which we mark with a red line below.

```
kNNdistplot(iris, k = 5)
abline(h=.5, col = "red", lty=2)
```



Apply the DBSCAN algorithm with `eps=0.5`.

The output includes the size of each cluster.

```
res <- dbscan::dbscan(iris, eps = .5, minPts = 5)
res
```

```
## DBSCAN clustering for 150 objects.
## Parameters: eps = 0.5, minPts = 5
## The clustering contains 2 cluster(s) and 17 noise points.
##
```

```
##  0  1  2
## 17 49 84
##
## Available fields: cluster, eps, minPts
```

Visualize clustering in a pair plot.

The pair plot includes the four quantitative variables in the dataset and shows the clustering through the shapes of the data points with in each plot (circle, square, triangle).

```
data(iris)
ggpairs(iris, columns = 1:4,
        mapping = ggplot2::aes(alpha = 0.3, shape = factor(res$cluster)),
        columnLabels = c("Sepal Length", "Sepal Width", "Petal Length", "Petal Width"),
        upper=list(continuous="points"))
```