Lab 10: Clustering

```r
library(tidyverse) ## data manipulation
library(knitr) ## tables

## reproducible
set.seed(445)
```

0.1 Data Preparation

We will make some simulated data to see how clustering works.

Run the following code to create the data.

```r
n <- 50
p <- 2
x <- matrix(rnorm(n * p), ncol = p)

## shift the center of one group
x[1:25, 1] <- x[1:25, 1] + 3
x[1:25, 2] <- x[1:25, 1] - 4
```

1. Make a scatterplot to inspect the data. Describe what you see.

0.2 K-means Clustering

We will use the `kmeans` function to perform K-means clustering. We can specify how many random initializations to use with the `nstart` parameter. For this lab, used `nstart = 20`.

1. Perform K-means clustering with $K = 2$.

2. Create a scatterplot of your data, colored by the resulting clustering. Describe what you see.

3. Repeat 1-2 with $K = 3$.

4. The total within sum of squares is available in the `kmeans` object under the name `tot.withinss`. Compare your two clusterings from 1. and 3. Which should you

choose?

0.3 Hierarchical Clustering

The `hclust` function implements hierarchical clustering in `R`.

1. Use the `dist` function to create a dissimilarity matrix corresponding to euclidean distance for the data you have simulated.

2. Create and plot the dendrograms for complete, single, and average linkage using the `hclust` function.

3. Cut each dendrogram to result in 2 clusters using the `cutree` function.

4. Create 4 scatterplots of your data, colored by the resulting clusterings from 3. Describe what you see.

5. Repeat 1-4. after scaling your data using `scale`. Are there any changes?