ST464/ST674/ST684
Assignment 1
Katarina Domijan
Due on Thursday 28th Feb 1pm

- Do all questions. Submit questions: 3, 4, 5.
- Use R markdown to knit your results to a .pdf file. Print and submit to your tutor's box.
- Place your name and student number under author in the YAML header. E.g.

```
---
title: "Assignment 1"
output: pdf_document
author: Jane Doe 1234567
---
```

- There will be a tutorial on Monday 25th Feb.
- The datasets are on the RStudio share folder (rstudio_files/ST464/data/)

1. For the dataset below, do these calculations by hand.

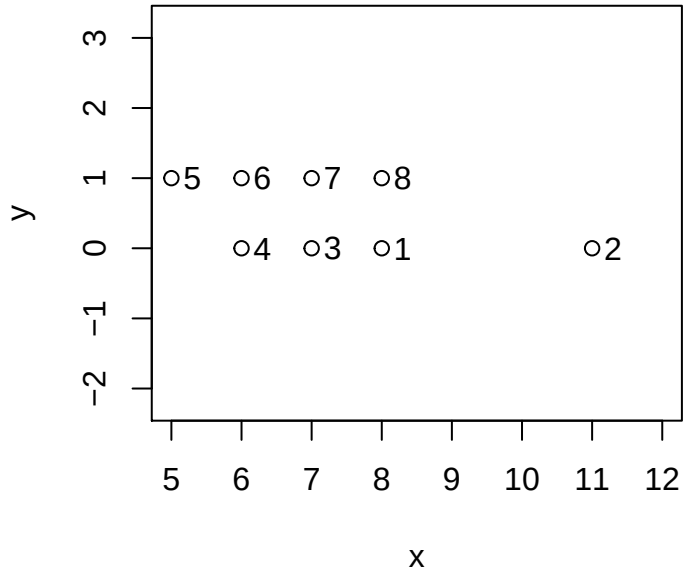
	U	V
a	-2	-5
b	2	-3
c	5	3
d	4	-4
e	1	0

   (a) Calculate a distance matrix using squared euclidean distance.
   (b) Use hierarchical clustering with single linkage to cluster the data. Draw the dendrogram and identify the two-cluster solution.
   (c) Use hierarchical clustering with average linkage to cluster the data. Draw the dendrogram and identify the two-cluster solution.
   (d) Cluster the data using kmeans with $k = 2$. Use starting clusters of (a,b,c) and (d,e).

2. Eight online shoppers buy 8, 11, 7 , 6, 5, 6 , 7, 8 pairs of socks. The same eight shoppers buy 0,0,0,0,1,1,1,1 computers.

   (a) If you run kmeans on this data with $k = 2$, with no scaling, what result would you expect?

```
x <- c(8,11,7,6,5,6,7,8)
y <- c(0,0,0,0,1,1,1,1)
plot(x,y, xlim = c(5, 12), asp=1)
text(x+.3,y, 1:8 )
```




```
d <- data.frame(x=x,y=y)
kmeans(d,2)$cluster
kmeans(d,2, nstart=10)$cluster
```

(b) If both variables are scaled to unit standard deviation, what will kmeans with $k = 2$ give you?

```
d1 <- scale(d, center=F)
```

(c) Suppose socks cost 2 euro and the computer is 2000 euro. What is you clustered the amount spent by each customer using kmeans with $k = 2$, with no scaling?

```
d <- data.frame(x=2*x,y=2000*y)
```

You should be able to do this question without running any R code.

3. The file `eupop.txt` contains the population and percentage distribution by age for EU countries in 1999. The age categories are 0-14 years, 15-44 years, 45-64 years and 65 years and over.

```
eupop <- read.table("yourfolder/eupop.txt", header=T, row.names=1)
eupop <- eupop[,-5]
```

(a) Construct the euclidean distance matrix of the percentage variables. Use it to cluster the countries, using average linkage. Draw the dendrogram and interpret. Are there any outlier countries?

(b) Examine the 3-cluster solution. Which countries belong to each of the three clusters? Summarise the partitions with sumPartition (in h1code.R) Interpret your findings.

(c) Use the kmeans algorithm to find another 3-cluster grouping of countries. Which countries belong to each of the three clusters?

(d) Construct a stars plot which shows the data and clustering obtained from kmeans. Optional: can you think of a better way of showing the clusters? Can you think of a way to present the data and the clustering results of both methods on the same graphical display?

4. Music data from class.

   (a) Run the k-means algorithm over the range $k = 1, \ldots, 15$ clusters and record the total within cluster sum of squares (TWSS). Let nstart = 25. Plot k versus TWSS and choose the best fitting number of clusters. What do you observe? Note: remember to scale the data.

   (b) Make a table of artist vs cluster solution from $k = 5$.

5. Protein data. We want to study the similarities and differences in the protein composition of the diets of different countries. Using any methods that you choose from this course or otherwise, write a brief summary.