Assignment 1. ST464/ST684
Catherine Hurley
Due Friday February 26, 6pm.

- Do all questions.
- Upload solution to question 4 only to Moodle by Friday February 26, 6pm.
- You should complete the assignment in Rmarkdown, knit the file to html and upload the html file to Moodle. If you are new to Rmarkdown, ask for help in the tutorial.
- Please use the skeleton Rmarkdown file h1student.Rmd available on Moodle.
- Fill in your name and student number on the solution file in the space provided.
- The upload must be completed by time and date given above or the assignment will not be accepted.
- The tutorial on week 4 will provide assistance with this material.

1. For the dataset below, do calculations for (a) to (d) by hand.

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

   (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).
   (e) Verify your answers in R.

2. The dataset alcohol contains physicochemical characteristics of 44 aliphatic alcohols.

```
data(alcohol, package="robustbase")
alcohol <- alcohol[,-7]
```

   (a) Scale the data using the default options. Construct the euclidean distance matrix of the cases. Cluster the cases, using average linkage. Draw the dendrogram.
   (b) Examine the 4-cluster solution. How many cases are in each cluster? Summarise the partitions with sumPartition (in h1code.R) Be sure to use the scaled data. Comment on the chemical composition of the four clusters.
   (c) Verify your findings by drawing a parallel coordinate plot of the data coloured by the clusters. Choose a suitable scaling.
   (d) Use the kmeans algorithm to find another 4-cluster grouping. Use the scaled data and nstart=10. How many cases are in each cluster?

(e) Construct a stars plot which shows the data and clustering obtained from kmeans. Arrange the stars by cluster. (Advanced: arrange the stars by size).

3. 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?

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

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

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

4. The worldhappiness2019.csv data was obtained from Kaggle, see `https://www.kaggle.com/unsdsn/world-happiness` for a description of the data.

```
w <- read.csv("worldhappiness2019.csv")
names(w)<- c( "Rank", "Country", "Score", "GDP", "Family",
              "Life", "Freedom", "Generosity", "Trust")
rownames(w)<- w$Country
wscores <- w[,-(1:3)]
```

   (a) Calculate the correlation matrix of the scores. Which pair of variables have the highest correlation? Make a scatterplot of the scores for these two variables. Find the name of the outlying country on this graph.

   [4 marks]

   (b) Construct the euclidean distance matrix wscores (no standardisation). Use it to cluster the countries, using average linkage. Draw the dendrogram. Are there any outlier countries?

   [3 marks]

   (c) How many countries are in the clusters of the three cluster solution? Which cluster does Ireland belong to? India? Which cluster has the countries with the highest Family score? (use sumPartition). Which cluster has the highest scores?

   [4 marks]

   (d) Make a parallel coordinate plot of the scores coloured by the clusters. Choose a suitable scaling. What is unusual about the lowest-scoring cluster from this graph?

   [3 marks]

   (e) Cluster the countries using kmeans (no standardisation), use nstart=10, centers=3, and a seed of 123. Are there any outlier countries?

   [2 marks]

   (f) Make a parallel coordinate plot of the scores coloured by the kmeans clusters. Choose a suitable scaling. From the graph, which cluster has the countries with the highest Family score? Which cluster has the highest scores overall?

   [4 marks]

   (g) For the kmeans clustering result, make a boxplot of Score (in w) for the three clusters. Comment on how the Scores of the clusters compare.

   [4 marks]