# Assignment 2 - 50 marks

John O'Sullivan
ESILV-J - Data Statistics
Dorset College Dublin

Due by midnight on Sunday the 10th of May 2020

## Instructions

- This assignment involves three tasks - Question 1 (14 marks) involves clustering analysis, Question 2 (18 marks) involves a logisitic regression, and Question 3 (18 marks) involves PCA.
- You should submit your assignment to the 'Assignment 2' object on Moodle.
- You should submit two files:
  (i) a single .Rmd script file containing all of the commented code you used to obtain your answers
  (ii) the HTML (or pdf) file which you produced from the .Rmd script
- You may need to find some new functions in order to do some of these tasks. Remember to use R's search engine, as well as checking online.
- Make sure that your file is readable and has a neat presentation and clear flow. The HTML (or pdf) output file **must be a stand-alone document containing the answers to all questions and showing all necessary code and all necessary output**. Marks will only be given for clear and detailed answers.
- I advise you to first create an R script with all of your answers. When you are happy with this, convert it piece-by-piece into an .Rmd file.
- There are marks in this assignment for document presentation - your final document should be neat with a clear layout, showing a good use of RMarkdown to mix free-flowing text, code, and output.

# Question 1 - Clustering

The **pottery.csv** dataset on Moodle contains the chemical composition of Romano-British pottery. The data record the chemical composition of 48 pots, determined by atomic absorption spectrophotometry, for nine oxides, as well as the location (the kiln) at which the pottery was found. The nine oxides are Al2O3 (aluminium trioxide), Fe2O3 (iron trioxide), MgO (magnesium oxide), CaO (calcium oxide), Na2O (natrium oxide), K2O (calium oxide), TiO2 (titanium oxide), MnO (mangan oxide) and BaO (barium oxide).

## Question 1.a

Explore the dataset. Briefly summarise the variables, and include any plots which you feel are necessary.

## Question 1.b

Remove the column 'kiln', so that for the remainder of these questions you are working with the numeric columns 2 to 10. Should this data be standardised prior to analysis? Explain your reasoning. (If so, use the standardised dataset for the following questions.)

## Question 1.c

Cluster the 48 pots using hierarchical clustering and average linkage. From the dendrogram, how many clusters would you suggest are present in the dataset? Cut the dendrogram at the desired number of clusters.

## Question 1.d

Cluster the 48 pots using $k$-means clustering. How many clusters would you suggest are present in the dataset? Detail any decisions you make when running this procedure. Provide details of your reasoning.

## Question 1.e

Compare the cluster solutions obtained in 1.c and 1.d using an appropriate measure(s) and comment on the agreement between the two solutions.

## Question 1.f

Comment on the relationship between both clustering solutions and the known 'kiln' variable. Have you any concerns about the reproducibility of your clustering solutions?

# Question 2 - Logistic Regression

The file called **heart-disease.csv** on Moodle contains information from a study on 303 individuals. Details of the variables can be found in **heart-disease.txt**.

A hospital wants to use this information to predict the presence of heart disease in patients. The response variable of interest is the *num* variable. They ask you to write a report where you present a model they can use to do this.

In your report, consider structuring it in the following way: Introduction, Data Exploration, Methodology, Results, Conclusion. You should apply some/all of the steps we have seen in class for logistic regression, and clearly outline the decisions you make as you conduct your analysis.

Your report is intended for the hospital directors to read, and should be neat and clear and concise.

Note that some of these variables are continuous and some are not. Also consider that you may want to include interactions in your model choice. As with the previous question, you may need to do some extra research online to help you.

# Question 3 - Principal Components Analysis

The **ratings.txt** file on Moodle contains the ratings of 329 communities according to 9 criteria: Climate and Terrain; Housing; Health Care and Environment; Crime; Transportation; Education; The Arts; Recreation; and Economics. (The $10^{th}$ column is simply a row index and should not be used in the analysis.)

Within the dataset, except for housing and crime, the higher the score the better. For housing and crime, the lower the score the better. Where some communities might rate better in the arts, other communities might rate better in other areas such as having a lower crime rate and good educational opportunities.

With such a large number of variables, it is quite difficult to assess or visualise the data in a meaningful form. It is neccessary to reduce the number of variables to a few interpretable linear combinations of the data.

Your task is to write a report on the application of PCA to this dataset. In your report, consider structuring it in the following way: Introduction, Data Exploration, Methodology, Results, Conclusion. You should apply some/all of the steps we have seen in class for PCA methods, and clearly outline the decisions you make as you conduct your analysis.