Data Science Laboratory 1

Assignment 3

The final assignment is all about modelling and clustering. You are expected to bring your own initiative to the task to find a data set(s) of interest to you (you can take inspiration from the links provided in the previous assignments, copied below for your convenience). You have a choice between *one* of the following *two* choices:

- Creating a simple model for a system, and performing optimization (e.g. least squares, MLE) to robustly fit the parameters, including considerations of uncertainty and confidence. For example, the rate of increase of Covid-19 infections can be described by an exponential function, but what are the parameters of that function? Or the rate of production of CO_2 in a country might be a function of the population; what is that function and what are its characteristic parameters. Can the model be used for prediction?
- Identify clusters within a data set of your choosing, using an algorithm or algorithms of your choice and interpret their meaning. For example, in the AirBnB listings of London, are there geographic clusters when one takes into account price, amenities, etc.? What key factors drive the clusters? Remember, clusters can be found in many dimensions!

As before, you are expected to produce a 1 page report conforming to the same guidelines as in Assessment 1 and 2: present your findings with appropriate visualisation, clear narrative, attention to detail and explanation of methodology used. Be sure to make some commentary on the interpretation of your findings. **In addition**, you are also expected to give a **5-10 minute** presentation on your findings. You are expected to introduce your data set and outline what you have done. You are then expected to explain your key results and make any relevant conclusions. There will then be **5 minutes** in which you will be asked questions.

Some ideas for data sources:

- Government: `https://data.gov.uk`
- MET Office: `https://www.metoffice.gov.uk/research/climate/maps-and-data/data/index`
- World Bank Open Data: `https://data.worldbank.org`
- WHO Global Health Observatory: `https://www.who.int/data/gho`
- Google Public Data Explorer: `https://www.google.com/publicdata/directory`
- FiveThirtyEight: `https://data.fivethirtyeight.com`
- UCI Machine Learning Repository: `https://archive.ics.uci.edu/ml/index.php`