INFO411
Laboratory exercises on
Data handling, outliers and noise

Important notes:
This lab is a non-assessed group exercise. All students must work as part of a group. Student must be a member of one group (and one group only) and group sizes must be 3 or 4 (no less than 3, no more than 4 students in each group). You can chose your own group, or you can ask the tutor during the lab to be allocated to a group. Each group member makes a submission as a proof of participation.

Objectives: Learning outcomes:
- grasp the importance of understanding the data/domain
- ability to work with fixed field data sets
- ability to understand and work with data that contain errors, outliers, noise
- ability to work with dates and time stamps

Overview:
The job of a data miner starts with a dataset. Datasets can come in a myriad of formats. There are literally an uncountable number of different data formats. The first job of a data miner is hence to understand the format of the data file and to find ways of loading the data into a computer for processing. A very good introduction to loading various data formats in R can be found in https://www.datacamp.com/community/tutorials/importing-data-r-part-two.
In this tutorial task we are given a small subset of patient transaction records, and a description of the data format. The data in this dataset is an extract from a real world dataset which is more than 100 times in size. Since we have neither the time nor the resources to process the full set and hence we will use a small subset. It is also useful to know that the data was entered manual (i.e. a human being entered the data into the database).
The second step of a Data Miner is to develop an understanding of the data and domain. A statistical analysis can be very useful. It is during this phase that a Data Miner gets an insight in the reliability of the data (i.e. whether there is noise or outliers in the dataset). Note that noise and outliers are not normally obvious. It is commonly necessary to ask the right questions in order to identify noise and/or outliers.
There can be some pitfalls when working with dates. If a dataset contains dates and/or timestamps then this requires extra caution.
Today's lab task is to guide you through the first steps of a Data Miner. To expose you to common challenges and how these need to be addressed.

What you need:
1. R software package (already installed on the lab computers)
2. The files "DataFormat.pdf", "Step1.R", "Step2.R", "Step3.R", "Step4.R", and "valid_postcodes.csv" on Moodle.
3. **Internet access and some free disk space**

Preparation:
1. Work in groups. Minimum group size is 3, maximum size of a group is 4.
2. Boot computer into Windows mode.
3. Download "DataFormat.pdf", "Step1.R", "Step2.R", "Step3.R", "Step4.R" then save to an arbitrary folder, say "C:\Users\yourname\Desktop"
4. Read the file DataFormat.pdf

5. Start "R"
6. Change the working directory by entering: setwd("C:/Users/yourname/Desktop")
   (Note that R expects forward slashes rather than backward slashes as used by Windows.)

Your task:
Each of the R-files contain a set of tasks and questions. You are to submit a PDF document which contains your answers to the questions. One document is to be submitted by each student. Clearly indicate which question you have answered. You answers should be thoroughly explained and justified.

Work through the following steps and answer given questions:
Step1: Open file Step1.R by using a text editor (i.e. Notepad)
Execute one line of code at a time. Try to understand as many of the copied commands as possible (do not just blindly copy). **Answer the question in Step1.R.**

Step 2: Open file Step2.r by using a text editor.
Execute each line (one at a time) in this file in R. Try to understand all of the copied commands.
Answer the questions in Step2.R. Do not move to Step3 until you have thought very carefully about the question in step2.

Step 3: Open the file Step3.R
You will need the datafile "valid_postcodes.csv" for this task.
Execute each line in Step3.R (one line at a time) in R. Try to understand as many of the copied commands as possible. **Answer the questions in Step3.R.**

Step 4: Open file Step4.R by using a text editor.
Execute each line (one at a time) in this file in R. Try to understand all of the copied commands.
Answer the questions in Step4.R.

Step 5:
Submit your answers (one PDF file) via the provided submission link for this lab on the subjects' Moodle site.