R Practical 1: Reading in and manipulating data

# Instructions

In this R practical, you will have the opportunity to try out some of the commands that were covered in the previous lecture.

# Remember:

## A few notes about R

* To run R commands, highlight the line(s) of code you want to run in the R-file and click the execute button  or press Ctrl+R. If you do not highlight specific lines of code, only the current line will be executed.
* Comments appear after the # symbol. They are ignored by R when code is run.
* R is case-sensitive so if there is a variable called ‘Age’, typing ‘age’ or ‘AGE’ will not work for this variable.

## Data

From SDEC at YGC admissions data from the 3-month period of   
September to November 2020.

Please note, SDEC is the provision of same day care for emergency patients who would otherwise be admitted to hospital.

* + Identifier
  + Age
  + Gender
  + Admittance method
  + Number of visits
  + Length of Stay (LOS)
  + Speciality
  + DISCHARGE WARD
  + Discharged

This data was made freely available by Hannah Dowell (4th year medical student)and Dr. Richard Morgan

**1. Setup**

Firstly, we want to set up our working environment to make it easier to read in data and save our work.

a) Before you open R, create a new folder (you could call it R course) and save the datasets you received via email into this folder.

I strongly recommend creating this folder inside My Documents on your **One Drive**, so that you’ll have easy access to it.

b) Open R, and open a new script. From now on, type all your commands in the script, and execute them from there. If there are errors in your script, correct them and re-run them to make sure they’re working as they should.

c) Use getwd() to find which directory you’re working in. If it’s useful, you can add comments to your script to remind you how a function works, or what output you get from it, by starting the line with a #.

d) Change your working directory to the folder you created in step a) by using setwd(…). Instead of … , you’ll need to provide the path to the folder you created in part a).

If you created a folder called R course in your One Drive, then this will be: setwd(‘C:/Users/yourusername/OneDrive - Bangor University/2028/R course’).

e) Now save your script (using file → save) and it will automatically be saved in the working directory. You can check this by using the command dir() which will list everything in your directory.

**2. Read in csv file**

Now we’re ready to read in some data, we’ll start with the SDEC data set from the lecture.

a) Read in the gusto data set contained in the file *SDECdata\_tidy.csv*, and call it data.

*Hint:* Use the command read.csv check the lecture notes for the exact command.

If it’s in your working directory, you won’t need to specify the full file path, just give the name of the file.

Make sure you don’t get any errors at this point.

**3. Explore dataset**

Now we can find out some general properties of the data set and look at specific values which are of interest to us.

a) How many rows and columns does this data set have?

*Hint:* dim().

b) Which variables are continuous?

*Hint:* str()

c) What is the height and weight of participants 101 and 1001?

*Hint:* data$ Age[c(101,1001)]

d) How many people smoke?

*Hint:* Look at the SMK variable, you might need to use the table on the first page of this worksheet to interpret its values. Then use sum(Gusto$SMK==…)

e) How many over 65s?

*Hint*: You can sum the A65 variable

f) How many over 70s?

*Hint:* Look at the AGE variable; Gusto$AGE>=70 will return a logical vector containing TRUE if a particitpant is 70 or over, and FALSE otherwise.

g) How tall are the tallest and shortest people in the study.

*Hint:* Look at the page of useful functions in the lecture notes for min() and max().

**4. Use help manual**

R has lots of functions and we only have time to cover a very small number in this course. The help functions are vital in learning how to use a new function. Here we’ll use a function you won’t have seen before and use the help pages.

a) Find the .25 and .75 quantiles of participants’ age in the complete and filtered data sets.

Hint: Use ?quantile.

b) Find the .1 and the .9 quantiles of participants’ age in the complete and filtered data sets.

*Hint:* Use ?quantile; look at the probs argument and the Examples right at the bottom.

**7. Use new packages**

R has loads of packages available for it, often another user has written a package to do exactly the same thing you want to do. The xlsx package contains functions to read and write .xlsx files. We’ll use this package to read in Excel files more easily.

a) Install the package xlsx using install.packages(‘xlsx’).

*Hint:* You’ll need to use a personal library, so click yes to those options; then you’ll need to select a CRAN mirror – it doesn’t really matter which you choose, but it’s better to choose one close to your location. I often use one from London, scroll down the list and select it from under UK. You could also do it through the tool bar, clicking packages → install packages… and then you’ll get a list from which you can select xlsx.

You only need to install the package once, so use # to comment this line out of your script once you’ve installed it.

b) Load the library.

Hint: Use library(xlsx)

c) Read in the file called Gusto\_Edited.xlsx and call it GustoX.

Hint: Use read.xlsx() and ?read.xlsx to find out about the inputs. You might have to look at the excel file to find out which sheet number you need.