STAT8010 Assignment 1 - Due Sunday 15th November at 23:59

1. Using the xlsx or readxl package or otherwise, read each sheet in the "assignment1.xlsx" file into R.
2. Generate a data frame for each sheet in the file.
3. The dataset in the first sheet is a random selection from a larger dataset. You will never get access to the full dataset so you should regenerate a new identification number for each subject in the dataset. This should be the row number of each entry in Sheet 1. You do not need to do this for Sheet 2.
4. It is also required to have an additional identifier which is the number you have generated in (3) followed by the first letter of each subjects first name and then followed by the first letter of each subject's surname. You do not need to do this for Sheet 2.
5. Although the data is not available for most subjects, some data highlighting subjects state of health is available in Sheet 2. You should use the subjects ID number to match it and merge it with the data in Sheet 1.
6. Not every subject has its ID number included in Sheet 2. You should attempt to match the remaining subjects using their first and surnames. This must be done using tidyverse in a robust manner. Your code for doing this should work again in the case of a new sample of data being provided.
7. You should add a column for age range. This should be

Age Range	Category Name
0-17	1
18-35	2
35-54	3
54-74	4
74-	5

8. You should filter the data by each age category. Generate a bar plot using ggplot2 for the criminal record variable.
9. You should generate an appropriate visualisation examining the relationships between height, weight, age and criminal records. Comment on this.
10. Using filters, you should analyse if there are any interesting results in the dataset regarding the relationships between height, weight and criminal record. Use appropriate visualisations.
11. Generate a smaller data frame for the subjects where health related data is available. Examine if there is a relationship between the different states of health and height, weight or age. Use appropriate visualisations. Note this should include a modelling type analysis such as regression. (S. Weisberg. Applied Linear Regression. Wiley Series in Probability and Statistics, 2005. may be useful)

You should write a short report outlining your results. This report and the code used to generate the report should be submitted to Canvas before **23:59 Sunday 15th November**. You must also record a short 2-3 minute presentation titled "Three interesting things I learned about R during this assignment"; this presentation should be uploaded to a discussions forum that will be made available after the report submission date.

Your code must be clearly annotated in your own words using comments. Failure to do so will result in a mark of 0 being recorded for your assessment. This means that every single line of code must be clearly explained in your script file. This is an **individual assignment**. Any collaboration amongst students is forbidden. Plagiarism is strictly prohibited and will be dealt with by the harshest punishments available.