Assignment 1: Exploratory data analysis and visualization

(30% of total mark)

Word limit: 1,500 words maximum, excluding the R script

Deadline: **Monday, February 19, 2024, 12pm**

How to submit

This assignment must be submitted online via the Assignment area of our Moodle site. Please do not submit work by e-mail. Each copy must have a completed Coversheet for Coursework attached to it. The coversheet can be downloaded from our Moodle page.

When completing the assignment, please remember that

- you should answer all of the questions and all the parts of each question;
- your assignment submission must include a complete log of your R session, including your R script with the commands and any graphical outputs produced. The log and the plots don't count towards the word limit;
- you do not need to write a lot in response to each question, but you should present your results, etc., in a neat, orderly, and professional manner, as if you were including them in a formal paper (e.g., a report or a dissertation). That is, do not just dump raw R outputs into the main text of your assignment;
- the word-limit for this assignment is a maximum upper limit only. You must be thorough and precise to do well, but you need not "pad out" your assignment to get anywhere close to this limit;
- adhere to APA formatting style for reporting your findings.

Assignment Information

Download the `crimes-sep21.csv` dataset from Moodle and save it in your working directory. The dataset contains data on recorded crime rates from police-force areas in England. The data cover the 12-month period ending September 2021. The data were retrieved online from the Office for National Statistics. All the figures in the file are expressed as rates per 1,000 head of population.

The following abbreviations are used as column headings:

- stalk = stalking and harassment
- pubord = public-order offences
- shoplift = shoplifting
- crimdam = criminal damage and arson
- drugs = drug offences
- vioper = violence against the person
- theft = theft (!)

The column headed "region" sorts the individual police-forces into 8 broader regional areas, as follows. (The data for London, which has only two police-forces, have been omitted for the present task.)

- 1 = North East
- 2 = North West
- 3 = Yorkshire and the Humber
- 4 = East Midlands
- 5 = West Midlands
- 6 = East of England
- 7 = South East
- 8 = South West

Please read the data file into RStudio and then use it to answer the following questions. You will need to have the tidyverse package loaded.

Question 1

Using R, calculate the mean rates and standard deviations for (a) stalking and harassment, and (b) theft. Report your results in one or two complete sentences, as if you were including them in a research report, rounded to two decimal places.

Question 2

Using R, calculate the median rates and interquartile ranges for (a) drugs offences, and (b) violence against the person. Report your results in one or two complete sentences, as if you were including them in a research report, rounded to two decimal places.

Question 3

The following R code will draw a set of side-by-side boxplots for the rates of public-order offences in the eight regions. (Hint: Remember to name your object appropriately)

```
crime_data |>
 ggplot(aes(reorder(region, pubord, median), pubord)) +
 geom_boxplot()
```

Produce the set of plots, then explain briefly what each element of the plot shows. That is, label and explain, in general terms, its component parts. Then go on to outline what you can infer from it about the present data set.Include the plot in your report. (You may like to cross-refer to the original data table to get the most out of this.)

Question 4

a. Modify the R commands from Question 3 to produce a set of side-by-side boxplots for the rates of criminal damage and arson in the eight regions. Outline briefly what your plots tell us about these data. Comment especially on any peculiar-looking features of these plots, and attempt to explain why they have occurred. Include the plot in your report. (Again, you may wish to cross-refer to the data table here)

b. modify the plot for question 4 and give it appropriate label axes (both x and y), and change the region labels from numbers to the names of the regional areas (e.g. 1 should be North East). Hint: you will need to use `mutate()` to update the data like we did in week 3. Include the publication-ready plot in your report with a suitable caption.