ST447 Data Analysis and Statistical Methods: Individual Project

1 Your Task

1.1 Meeting Your Friend, XYZ

XYZ is a student from LSE with the profile pseudo-randomly generated using the R function `XYZprofile` (with the argument being the **numerical** value of your 9-digit LSE Student ID).

This function is contained in the file `XYZprofile.r` on Moodle (under the Section "Project"). For example, if your LSE student ID is `202212345`, then you should use the following code:

```
> # Replace the number below by your LSE ID
> ID = 202212345
> # Then copy XYZprofile.r into your R working directory
> source("XYZprofile.r")
> # Now run the function XYZprofile with argument ID
> XYZprofile(ID)
```

You should get the following output:

```
The profile of XYZ:
- Age:  25
- Gender:  Female
- Home address:  Heckmondwike
```

1.2 Problem Description

XYZ has been learning driving for some time, and is thinking of taking the practical car test in UK.[1]. There are two sensible options for XYZ:

1. either take the practical test at the nearest test centre to his/her home;
2. or take it at the nearest test centre to the LSE.

Note that in XYZ's generated profile, the entry "Home address" actually gives the name of the nearest test driving test centre to XYZ's home. In addition, the test centre closest to the LSE is **Wood Green (London)**.[2]

XYZ thinks his/her driving skill is about average. It is widely believed that the driving test routes around some centres are probably more difficult than others (e.g. there are far less bus lanes, roundabouts and cyclists in rural areas than in London).

XYZ knows that you (who is his/her best friend) are taking ST447 this term. XYZ wants to rely on your data analysis skills, and would like you to answer the following questions:

1. What is XYZ's expected passing rate at the nearest test centre to his/her home?
2. What is XYZ's expected passing rate at the nearest test centre to the LSE?
3. Of these two locations, where should XYZ take the test? Is there any evidence to (statistically) support this suggestion?

2 Project Details

2.1 Data Source

The dataset **DVSA1203** is available at

`https://www.gov.uk/government/statistical-data-sets/car-driving-test-data-by-test-centre`

which contains information on car pass rates by age (17 to 25 year olds), gender, year (2007-2022) and test centre.

This dataset is also available on Moodle (under the Section "Project"). Note that this dataset is of "`*.ods`" format, so some data preparation *might* be required.

[1] More information about the test could be found at: `https://en.wikipedia.org/wiki/United_Kingdom_driving_test`

[2] Our school's postcode is WC2A 2AE. Given a postcode, you could find the nearest centre via `http://www.dft.gov.uk/fyn/practical.php`

2.2 Methods

Our intention is to simulate a real-life scenario, so this problem is open-ended. You could choose whatever method you believe that makes most sense. For example, you could either combine many years of data, or just based your analysis on data from a particular year, or investigate the yearly trend, etc; you could use logistic regression or just the Wald test, etc. However, no matter what you choose to do, you will need to briefly justify your choice (of data, method, etc) in the report.

2.3 Final Report

Your final report needs to be understood by a **non-expert** in statistics (as XYZ has limited previous training in statistics).

In addition to the answers to XYZ's three questions, your report should also include:

1. the profile of XYZ you used for this analysis;
2. briefly explanation of the data you used, methodology, as well as the assumptions behind the scene;
3. the relevant R code (with enough comments) so that XYZ could mimic your analysis;
4. the strengths (and potential weaknesses) of your approach to this particular problem.

Your report should be word-processed by, for example, Microsoft word, latex or Rmarkdown). There is a strict page limit for the report (**maximum 8 A4 sides**, including figures, tables and relevant R code). You should use an 11 point standard font (for example, times new roman) and 1.5 spacing.

Note: The limit of 8 pages is the upper bound of the length. A well-structured and clearly-written report can be much shorter than that.

Please ensure that your report has a title and your **5-digit candidate number** (i.e. the number to be used in exams, available on LfY). It should be anonymous, as your name must not appear anywhere in the report.

3 Submission of your report

3.1 Deadline for submission

Save your report as "xxxxx.pdf", where xxxxx stands for your 5-digit candidate number, and submit it to Moodle by **16:00 on Friday, 2 December 2022**.

Late submission entails penalties: 5 marks (out of maximum 100) will be deducted for every half-day (12 hours). This will result in a maximum penalty of 10 marks for the first 24 hours. Then further 5 marks will be deducted per 24 hour period thereafter. Submissions after 9 December 2022 cannot be accepted.

3.2 Assessment Criteria

We will mark your report by its **correctness, concreteness, clarity and conciseness**. In addition, you could show a **critical awareness** of any weaknesses in the analysis you present and discuss possible extensions and improvements. See also a separate file 'Assessment Criteria' for more details.

3.3 Plagiarism

Plagiarism is taking someone else's work or ideas and passing them off as your own (adapted from Concise Oxford Dictionary definition). This arises in course work as sections of text lifted from books, internet sources or someone else's work and submitted as your own work. This is a very serious offense that is quite easy to detect. Plagiarism will result in instant failure (mark 0).