Data Analysis
Week 5: Class Test 1

Introduction

This week is the first of two class tests for Data Analysis and is worth 35% of your final grade. The class test consists of 3 tasks worth a total of **40 MARKS** broken down as follows:

- A report on a statistical analysis of a given data set: **25 MARKS**;
- Further question 1: **7 MARKS**;
- Further question 2: **6 MARKS**;
- Successful upload of `.pdf` document: **2 MARKS**

All tasks will be completed within the same R Markdown document. The written report should include:

- An appropriate **Title** and **Introduction** detailing the data and question of interest; **2 MARKS**
- An **Exploratory Analysis** of the data; **7 MARKS**
- A **Formal Analysis** of the data; **12 MARKS**
- Finish with your **Conclusions**; and **2 MARKS**
- Have an appropriate report layout. **2 MARKS**

Instructions

1. **Do NOT** open RStudio until you have downloaded the required files described in Instructions 2. and 3.
2. Go to the **Class Test 1 Files** folder in the **Week 5: Class Test 1** section of the **Data Analysis Moodle page**.
3. Download the files in the **Class Test 1 Files** folder into the **same folder** on your **M: drive**:
    - `.csv` files contain the required data sets; and
    - `ClassTest1Template.Rmd` - an R Markdown template for this class test. It loads the R packages necessary to complete the set tasks.
4. Open RStudio and open `ClassTest1Template.Rmd` then save it as `ClassTest1YourStudentNumber.Rmd` in the **same folder** as the `.csv` files are saved on your **M: drive**.
5. **Before you start to work**, compile `ClassTest1YourStudentNumber.Rmd` (using `Knit`) and check that the `ClassTest1YourStudentNumber.pdf` file is compiled as expected. It is wise to periodically compile and check the `.pdf` file as you work through the tasks so you can more easily debug your code as you go. You will **NOT** receive any assistance with compiling your document.
6. For the report part of the class test you **are NOT required** to **include** your R code in the `.pdf` file, hence `echo=FALSE` is set as the default in the `.Rmd` template. However, for the further questions you will need to provide your R code in the `.pdf` file, and hence should include `echo=TRUE` in any corresponding R code chunks relating to the further questions.
7. When you are ready to submit your class test document, click on the **Class Test 1 .pdf Upload** link under **Data Analysis > Week 5: Class Test 1** and upload and submit the file `ClassTest1YourStudentNumber.pdf`. **1 MARK** will be deducted if the document is not named as instructed.

8. Also, upload and submit the R Markdown file `ClassTest1YourStudentNumber.Rmd` using the **Class Test 1 .Rmd Upload** link. Again, **1 MARK** will be deducted if the document is not named as instructed. Please note that only the `.pdf` file will be marked. The `.Rmd` file will only be considered if there was a problem compiling the `.pdf` file. **Note**, the `.pdf` file uploaded to Moodle will be considered as your **complete** class test, and as such any partial working files **should not** be uploaded in an attempt to obtain **2 MARKS**.

Examination Conditions

- You have 24 hours to complete the class test and can submit your completed tasks anytime within that time.
- You must work on your own - **NO communication** by any means with anyone is permissible.
- You may consult ANY resources (hardcopy or online), e.g. `tidyverse` "cheat sheets" and/or the online tutorials from the course.

Class Test Tasks

Report: Gambling among teenagers in Britain

A study was conducted into the gambling habits of teenagers living in Britain. Within the study 28 males and 19 females aged 16-19 were asked how much they spend on gambling in pounds per year. Here, we shall examine the expenditure of teenagers in Britain on gambling and determine whether there is a difference in spending habits between males and females. The data is contained within the `gambling.csv` file. Use what you have learned in previous weeks to produce a report on the following question of interest:

Using a linear model, what is the relationship between gambling expenditure and gender?

25 MARKS

Further Question 1

Observations from two random variables, X and Y, are provided in the `FQ1.csv` file.

(a) Produce an appropriately labelled plot of the data using `ggplot()` to look at any relationship between the two variables. Comment on any relationship observed.

1 MARK

(b) Given n observations from (X_i, Y_i) for $i, \ldots, n$, the sample correlation coefficient is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y. Create an R function named `corr.func` that has two arguments which read in the observations of X and Y and returns their sample correlation. You are **NOT** allowed to use any existing functions within R to compute the mean, variance, covariance or correlation. Once you have written the function compare it with the `cor()` function in R to check if you get the same solution.

6 MARKS

Further Question 2

In statistics, compositional data are quantitative parts of some whole, conveying relative information. For example, measurements may involve proportions, percentages or probabilities. The data file `FQ2.csv` contains

10 observations from compositional data based on proportions, such that the three components for each observation should sum to 1. However, there is a missing value for each observation. Taking into account the fact that each row should sum to 1, write a function named `comp.data` that replaces all of the missing values (`NA`'s) with the relative proportion. The function should be general, such that given any number of observations it should be able to appropriately replace all of the missing values.

Hint: the functions `is.na()`, `which()` and `rowSums()` may be helpful.

6 MARKS

Total: 38 MARKS (+ 2 for pdf upload)