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 two 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: PGA and LPGA Tour 2008 Driving Accuracy

In golf, the opening stroke on each golf hole (called "the drive") can determine how well a player performs, hence "driving" is an important component of becoming a successful professional golfer. If a golfer's drive lands in one of the various hazards on the course, such as the rough grass or the sand bunkers, then it can impact on their chances of finishing the hole with as few strokes as possible. Thus, it is important for a golfer's drive to accurately land on the fairway (i.e. the cut grass) in order to become competitive on the professional golf tours, e.g. the PGA Tour for men and the LPGA Tour for women.

One measure of drive accuracy is the percentage of a golfer's drives that land on the fairway. Data is available for the drive accuracy (%) of 354 professional golfers who competed on the PGA and LPGA Tours in 2008. The data is contained within the `PGALPGA2008.csv` file. Use what you have learned to produce a report on the following question of interest:

Using a linear model, describe the drive accuracies (%) of PGA (male) and LPGA (female) professional golfers. What does the model say about the difference in drive accuracies, on average, between male and female golfers?

25 MARKS

Further Question 1

A farmer is interested in comparing the effect of different fertilizers on crop yield, and decides to undertake an experiment. He wants to compare three different fertilizers, labelled `A`, `B` and `C`, respectively, against a control group with no fertilizer, labelled `D` . He partitions his field of potatoes into 40 plots, and applies each of the four treatments A, B, C and D to 10 plots at random. At the end of the experiment he measures the total weight (in kilograms) of potatoes grown in each of the 40 plots. The results of this experiment are stored in `test1.csv`.

(a) Use the `gather()` function to convert the data into the `tidy` format. Ensure the `Fertilizer` categorical variable is a factor.

3 MARKS

(b) Produce an appropriately labelled plot of the data using `ggplot()` that compares the yield distributions of the four different fertilizers. Comment on what you see from your plot.

4 MARKS

Further Question 2

(a) Simulate two continuous random variables X and Y, each consisting of 100 observations, where $\mathbb{E}(X) = 10$ and $\mathbb{E}(Y) = 18$ and $\text{Var}(X) = \text{Var}(Y) = 1$. X and Y should have a correlation coefficient between 0.5 and 0.7.

Hint: You may want to use the `mvrnorm()` function from the `MASS` library.

4 MARKS

(b) Produce an appropriately labelled scatterplot of your simulated data using `ggplot()` and comment on the relationship between X and Y. Using the `cor()` function, ensure that the correlation coefficient of your simulated X and Y lies between 0.5 and 0.7.

2 MARKS

Total: 38 MARKS (+ 2 for pdf upload)