Data Analysis Skills: Class Test

Introduction

This class test consists of three tasks worth a total of **40 MARKS** broken down as follows:

Task 1. A report on a statistical analysis of a given data set: **25 MARKS**;

Task 2. A Further Task: **13 MARKS**;

Task 3. Successful upload of appropriate `.pdf` and `.Rmd` files: **2 MARKS**.

Tasks 1. and 2. are to be completed within the same R Markdown document (and uploaded in Task 3).

Task 1, the written report on a statistical analysis, should include:

- An appropriate Title, the author's student number and Introduction detailing the data set and question of interest **4 MARKS**;
- An Exploratory Analysis of the data **7 MARKS**;
- A Formal Analysis of the data **10 MARKS** and
- Conclusions **2 MARKS**.

2 additional MARKS will be awarded for an appropriate report layout and good English.

Instructions

1. Go to the **Class Test Files** folder in the **Week 5: Class Test** section of the Data Analysis Skills Moodle page.
2. Download the files in the **Class Test Files** folder to the **same folder**:
    - The two `.csv` files containing the required data sets;
    - `ClassTestTemplate.Rmd` - a R Markdown template for the class test submission files. It includes the R packages necessary to complete the tasks.
3. Open **RStudio**
4. From within RStudio open `ClassTestTemplate.Rmd` then save it as `YourStudentNumber_ClassTest.Rmd` in the **same folder** as the `.csv` files are saved.
5. **Before you start to work**, compile `YourStudentNumber_ClassTest.Rmd` (using `Knit`) and check that the `YourStudentNumber_ClassTest.pdf` file is produced as you expected. **It is strongly recommended** to periodically compile and check the `.pdf` file as you create your document so you can fix any bugs in your code as you go.

6. For the report part of the class test (Task 1) 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 **Task 2: Further Task** you **need to provide your R code** in the `.pdf` file, and hence should include `echo=TRUE` in any R code chunks relating to the Further Task.

7. When you are ready to submit your .pdf document, click on the **Class Test .pdf Upload** link in the **Week 5: Class Test** section and upload and submit the file `YourStudentNumber_ClassTest.pdf`.
8. Also upload and submit the R Markdown file `YourStudentNumber_ClassTest.Rmd` using the **Class Test .Rmd Upload** link. Please note that only the `.pdf` file will be marked. The `.Rmd` file will only be considered if there was a problem when compiling the `.pdf` file.

Examination Conditions

- You have four hours to complete the class test and you can submit your `.pdf` and `.Rmd` files anytime within that time.
- You are required to use `tidyverse` functions (including `ggplot2`) for the analysis and `RMarkdown` to produce your .pdf document.
- You may consult resources (hard copy or online), e.g. `tidyverse` "cheat sheets" and/or the online labs from the course.
- You **must not** communicate or correspond with anyone about the class test during the time that submissions are open. You **must not** submit material you have discussed with or copied from others. If your work is similar to those of any other candidate(s) you will both/all be suspected of collusion and referred to Student Conduct.
- The work you submit must be entirely your own effort and must demonstrate your understanding rather than reproduce text from notes, slides, books, or online sources (which is plagiarism). We may conduct a further oral examination to check your knowledge and establish that the submitted files are your own work.

Task 1. Heights of International Rugby Players

In the sport of rugby there are many different positions a player could play but, in general, players are classified into one of two playing positions, namely "forwards" and "backs". Each position requires different traits of a player such as speed, strength and accuracy. Data were collected on rugby players from the 2015 Men's Rugby World Cup to investigate the heights (in cm) of players in the "forward" and "back" positions. The data are stored in the file `rugby_height.csv`.

Summarise the heights of the two positions, using both numerical and graphical summaries with appropriate comments. Using a linear model, comment on the difference in height, on average, between the "forward" and "back" positions. Based on the fitted linear model, is there a statistically significant difference in the heights between these two positions?

NB: For the purposes of Task 1, you are **NOT** required to include checks of the model assumptions.

Task 2. Further Task

In this task you will work with a larger data set detailing further attributes of rugby players in the 2015 Men's Rugby World Cup. In addition to information on the players' heights data is also available on players' weights in kg (`weight_kg`), age at the tournament (`age`), age at their first international appearance (`age_debut`), number of international appearances (`caps`), number of years playing for their international team (`years_playing`) and the country they played for (`team`). The `position` variable contains either the "Forward" or "Back" playing position.

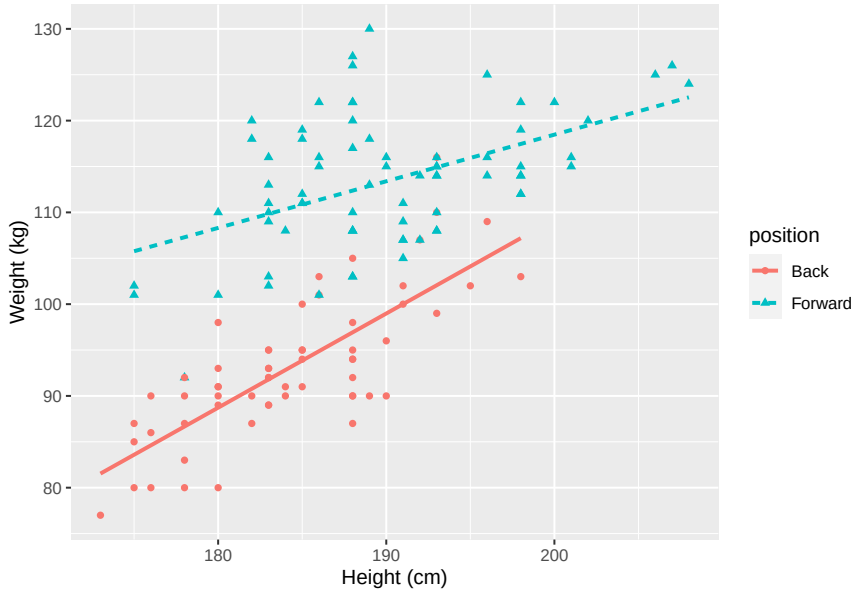
The data are stored in the file `rugby_full.csv`. Import this data into R and answer the following questions.

a. Add a new variable `games_per_year` to the data which contains the average number of games per year of their international career a player has played for their international team. **1 MARK**

b. Is the data set in `tidy` format (as defined in Week 2's Lab)? If not, write R code that converts the data to a `tidy` format, and include the code in your submitted `.pdf` document. **2 MARKS**

c. Produce a plot of the number of international appearances against number of years playing with separate panels for each team. The points should be different shapes and colours for each player position and a legend should be included showing which colour/shape refers to which position. Include meaningful axes labels and a figure caption and number the figure so that it follows after any figures included in the report in Task 1. Include the R code used to produce the plot together with the plot in your submitted `.pdf` document. **5 MARKS**

d. Produce the following scatterplot of height and weight of players in all British Isles teams (i.e. England, Ireland, Scotland and Wales) with "Forward" and "Back" positions shown separately with fitted linear regression lines superimposed (with no confidence bands). Add a figure caption and number the figure so that it follows previous figures. Include the R code used to produce the plot together with the plot in your submitted `.pdf` document. **5 MARKS**



Task 3. Upload appropriate `.pdf` and `.Rmd` files. 2 MARKS