Research Skills: Programming with R

Assignment 1

This graded set of homework assignments must be handed in on Blackboard before Wednesday November 21st, 5:00 PM. It tests your mastery of Worksheets 1 to 3. You will be asked to manipulate, summarise and plot data.

It will be graded as follows:

- 0.5 point each for Questions 1 through 8
- 1.0 point each for Questions 9 through 12
- 1.0 point in total for overall code organisation & style
- 1.0 point in total for complying with the instructions below

The guidelines for overall code organisation & style can be found in the Mini-Worksheet and the slides for Class 3. All questions are independent; copy the data set before modifying it, and start afresh with the original each time.

Questions 1 through 8 will be graded semi-automatically. Answer them exactly as asked, no deviations or elaborations; any exactly correct answer, irrespective of efficiency, will be worth 0.5 point, and any other, 0 point. Note that this *includes* matching the requested spellings & capitalisations exactly.

For Questions 9 through 12, partial credit will be awarded for partially correct answers; however, for full credit, your solution should not require more code than necessary, given the skills taught in the worksheets.

Other instructions:

- solve all the questions in a single R script
- use `Assigment_1_DemoScript.R`, from Blackboard, as the basis of this script
- load the data exactly as shown in this demo; do not adapt the relative path
- use any function from 'base R', `dplyr` and `ggplot2`, and no other packages
- name your script `lastname_u-number_assignment1.R`
- include your name and u-number at the top of your script
- store your final solutions in the objects `answerX` objects as described

This is an individual assignment: I accept that you will discuss it with your fellow students in general terms but directly sharing code is strictly prohibited. Suspected plagiarism will be referred to the Exam Board. Good luck!

Data Set Information

This assignment concerns a data set called `olympics`, a subset of data available from Kaggle.com. It contains information on the participants of all the sports events held at the Summer and Winter Olympics from 2010 - 2016; each row is a separate record. Most columns are self-explanatory; the `Height` column is in centimeters and the `Weight` column is in kilograms. The `NOC` column indicates the 'National Olympic Committee' that sent each athlete.

Question 1.

Create an object that's a copy of `olympics`, but which omits the `Games` column. It should be sorted by `Year`, oldest first, and then alphabetically by `Event` and `Name`, in that order. Create this object with a meaningful name initially, then copy it into an object called `answer1`.

Question 2.

Create an object that's a copy of `olympics`, but which includes only records with `"Relay"` in the name of the `Event`. Create this object with a meaningful name initially, then copy it into an object called `answer2`.

Question 3.

Create an object that's a copy of `olympics`, but with a new final column, `Title`. For each record, it should contain the `Year` of the record, followed by a space, and then the `City`. Create this object with a meaningful name initially, then copy it into an object called `answer3`.

Question 4.

Create an object which contains, for each combination of `Sex`, `Season` and `Year`, the number of different sports in the `olympics` data set. The columns of this object should be called `Competitor_Sex`, `Season`, `Year`, and `Num_Sports`, respectively. Create this object with a meaningful name initially, then copy it into an object called `answer4`.

Question 5.

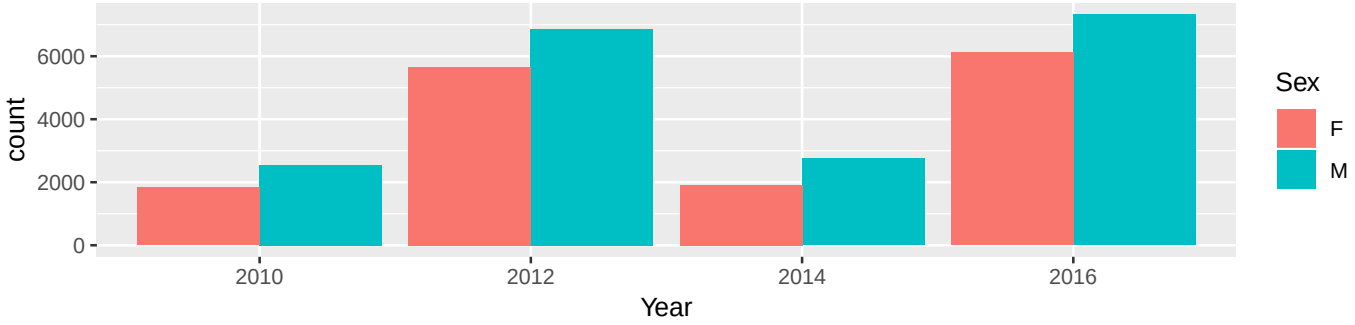
Using `ggplot2`, create a boxplot of `Height` per `Sex`. Hide outliers, and make the inside of the boxes for the women's boxplot `"pink"` and for the men's `"blue"`. Leave all other settings at their defaults. Store this plot in an object called `answer5`.

Question 6.

Using `ggplot2`, create a histogram of the `Age` variable. Each bar should represent 3 years. Label the y-axis `"Number of Records"`. Leave all other settings at their defaults. Store this plot in an object called `answer6`.

Question 7.

Using `ggplot2`, re-create this barplot. It shows the total number of records in the data set for each combination of `Sex` and `Year`. Store this plot in an object called `answer7`.



Note: You do not have to replicate the overall dimensions of the plot; these are controlled using RMarkdown.

Question 8.

Using `ggplot2`, create a scatterplot of `Weight` versus `Height`, with `Weight` on the y-axis and `Height` on the x-axis. The points' shape should represent the competitor's `Sex`, and all points should be `"purple"`. Leave all other settings at their defaults. Store this plot in an object called `answer8`.

Question 9.

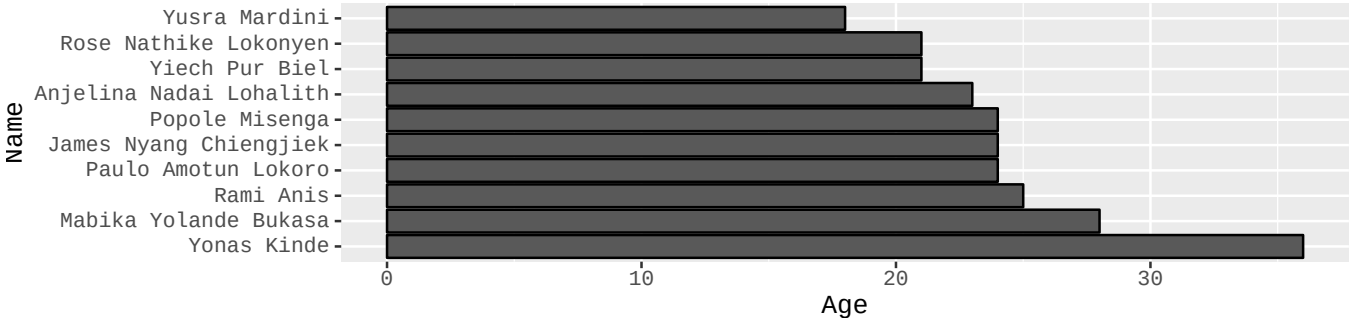
For each `Team` in the data set, calculate the median length of the athlete's `Name` in all corresponding records. Store the top 8 results in terms of median name length in a new object, consisting of the columns `Team` and `Median_Length` only; make sure it sorted alphabetically by `Team`. Create this object with a meaningful name initially, then store it in an object called `answer9`.

Question 10.

Using `ggplot2`, create a barplot showing the number of records relating to the `NED`, `BEL` and `LUX` 'National Olympic Committees'. Each `Season` should be represented by its own panel, with a separate bar for each country in each panel. Ensure that the y-axis is labelled `"Number of Records"`, that it runs from 0 to 600, and that it is placed on the right side of the plot. The title of the x-axis should be `"National Olympic Committee"`. Leave all other settings at their defaults. Store this plot in an object called `answer10`.

Question 11.

Re-create this barplot, concerning the `Team "Refugee Olympic Athletes"`; all text uses the `"mono"` font family. Store this plot in an object called `answer11`.



Note: You do not have to replicate the overall dimensions of the plot; these are controlled using RMarkdown.

Question 12.

Using `ggplot2`, create a scatterplot with one point per `Season`, each showing the mean `Weight` of all records related to that `Season`. Each point should be surrounded by error bars with `width = 0.1`, and these should show the standard error of the corresponding mean, calculated as the standard deviation divided by the square root of the number of records. The plotting area should be surrounded by a `"cadetblue"` border, consisting of a line that is 3 mm thick. Leave all other settings at their defaults. Store this plot in an object called `answer12`.