STAD29 / STA 1007 assignment 1

Due Tuesday Jan 14 at 11:59pm on Blackboard

Hand in the indicated questions. In preparation for the questions you hand in, it is worth your while to work through (or at least read through) the other questions as well.

What you hand in needs to include (i) your code, (ii) the output that your code produced and (iii) your comments on the output as asked for in the questions. The easiest way to get this is to use an R Notebook and preview the results (to HTML or Word or PDF) when you are done.

Hand in your work on Quercus. If you did STAC32 last fall, it's the same procedure. A reminder is here: `https://www.utsc.utoronto.ca/~butler/c32/quercus1.nb.html`

You are reminded that work handed in with your name on it must be *entirely your own work*. It is as if you have signed your name under it. If it was done wholly or partly by someone else, *you have committed an academic offence*, and you can expect to be asked to explain yourself. The same applies if you allow someone else to copy your work. The grader will be watching out for assignments that look suspiciously similar to each other (or to my solutions). Besides which, if you do not do your own assignments, you *will* do badly on the exams, because the struggle to figure things out for yourself is an important part of the learning process.

Before you start, you'll need this:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------- tidyverse 1.2.1 --
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
## The following object is masked from 'package:base':
##
##     date
```

1. Work through, or at least read, Chapter 15 of PASIAS: `http://ritsokiguess.site/pasias/`
2. Manchester United is one of the most famous soccer clubs in England and indeed the world. Information about the players in the current squad (as at December 10, 2019) is here: `http://ritsokiguess.site/STAD29/manu.csv` (it's a CSV file). We are going to learn something about the ages of the players.
   (a) (2 marks) Read in the file and display some of the resulting data frame.
   (b) (3 marks) What kind of thing is the column `date_of_birth`? Create a new column that contains the players' dates of birth as actual R dates, and display the old dates of birth alongside your new column (or at least the first few rows of them). Save your updated data frame.

(c) (2 marks) Treating the new column of dates as quantitative, make a suitable plot of these with the players' `position` on the field. What do you see on the quantitative axis?

(d) (2 marks) Is there a position where the players tend to be older? Explain briefly.

(e) (3 marks) (This is to prepare you for the next thing.) Work out how many years old *you* are by using something like `as.Date("1966-04-13")` to turn your birth date into an R date, create a `period` from the interval from it to now (use `Sys.Date()` to get today's date), and pull out the number of (completed) years. (If you don't want to share your birth date, use any other date. I'm not checking.) Make sure to have the right number of brackets in the right places.

(f) (3 marks) Go back to the Manchester United players. Calculate a new column containing the age, in completed years, of each player as measured today (thus, for example, a player who is currently 29 years and some number of days old should be counted as 29 years old, even if the number of days is something like 364). Display your new column side by side with the one called `age`. (This uses the same technique that you used to calculate your own age in the previous part, except that you don't need `as.Date` because you converted the birth dates into R `Date`s in an earlier part.)

(g) (3 marks) Display the names and (original) birth dates of the players whose age (in the original data frame) and whose age (as you calculated it) are different. What do these players have in common?

3. Work through, or at least read, problems 13.12, 13.13, 13.14, and 13.18 in PASIAS: `http://ritsokiguess.site/pasias/`

4. In the sport of baseball, the pitcher has a very important role: to stop the batters of the other team from scoring runs, and thereby helping their team to win the game. One way that a pitcher can get a batter out is called a "strikeout"; this, roughly speaking, is done by throwing three accurate pitches that the batter cannot hit. We would suspect that a pitcher who has more strikeouts would also help their team to win more games, but is that actually true?

   The data in `http://ritsokiguess.site/STAD29/pitchers.txt` is from 40 pitchers in Major League Baseball in the 2011 season. Pitchers earn a win for themselves each time they help their team win a game. Technical definition here.

   (a) (2 marks) Read in the (space-delimited) data and confirm that you have the right number of rows and the right columns (the column `sos` contains the number of strikeouts).

   (b) (2 marks) Fit a linear regression (predicting wins from strikeouts), and display the output.

   (c) (3 marks) Obtain confidence intervals for the mean number of wins for pitchers that have 125 and 175 strikeouts, based on your regression of the previous part.

   (d) (2 marks) Which one of your intervals in the previous part is longer? Explain briefly how that makes sense.

   (e) (2 marks) Find the prediction intervals for the numbers of wins for new pitchers (ones not in this data set) who have 125 and 175 strikeouts.

   (f) (2 marks) For a pitcher or pitchers with 175 strikeouts, compare the lengths of the confidence and prediction intervals. Which one is longer? Explain briefly why that happened.

Hand in your answers to the questions with marks attached to them (questions 2 and 4, here). I want to see your code, output and comments; a good way to get those is to use an R Notebook and Preview it (producing an HTML or PDF or Word file that you hand in).

Notes

[1]If you're colour-blind, you might find it difficult to distinguish everything. If that's you, let me know and we'll work on something you can actually see.

[2]So is $\lambda = 0$ or log, which you could also try. The data are not very revealing about what transformation would be good.

[3]This is a slightly artificial use of `select`, but I wanted to show you what happened.

[4]This would work with any transformation, for example `exp` to undo `log`.