Assignment 6

Due Tuesday March 17 at 11:59pm on Quercus

The assignment is due on the date shown above. An assignment handed in after the deadline is late, and may or may not be accepted (see course outline). My solutions to the assignment questions will be available when everyone has handed in their assignment.

You are reminded that work handed in with your name on it must be *entirely your own work*.

Assignments are to be handed in on Quercus. See `https://www.utsc.utoronto.ca/~butler/c32/quercus1.nb.html` for instructions on handing in assignments in Quercus. Markers' comments and grades will be available there as well.

Begin with the usual:

```
library(tidyverse)
```

Hand in question 2 below. Question 3 is a bonus question, if you want an extra challenge. If you want the bonus points, hand in question 3 as well.

1. Work through chapter 13 of PASIAS.
2. The file `http://ritsokiguess.site/STAC33/xgrades.csv` contains a data frame with some marks for some students on some tests.
   (a) (2 marks) Read in the data frame and display at least some of it. (It has 12 rows).
   (b) (6 marks) The instructor who awarded these marks wants to rearrange the data frame as shown below:

```
## # A tibble: 18 x 5
##       ID  Year Quarter  Math Writing
##    <dbl> <dbl> <chr>   <dbl>   <dbl>
##  1     1  2008 Fall       15      22
##  2     2  2008 Fall       12      13
##  3     3  2008 Fall       11      17
##  4     1  2008 Spring     16      22
##  5     2  2008 Spring     13      11
##  6     3  2008 Spring     12      12
##  7     1  2008 Winter     19      24
##  8     2  2008 Winter     25      29
##  9     3  2008 Winter     22      23
## 10     1  2009 Fall       12      10
## 11     2  2009 Fall       16      23
## 12     3  2009 Fall       13      14
## 13     1  2009 Spring     13      14
## 14     2  2009 Spring     14      20
## 15     3  2009 Spring     11       9
## 16     1  2009 Winter     27      20
```

```
## 17     2  2009 Winter     21      26
## 18     3  2009 Winter     27      31
```

By making the data frame longer and/or wider or using other tools as appropriate, convert the data frame you read in in the previous part to be laid out this way.

3. This is a bonus question; there are 4 bonus points for a complete answer to this one, which, if earned, allow you to score more than the maximum for this assignment. If you want a shot at the bonus points, hand in your answer to this one as well.

   The Toronto Wolfpack play rugby league, in a league with a lot of English teams (and one French one). They play at Lamport Stadium on King. The file `http://www.utsc.utoronto.ca/~butler/assgt_data/rl.txt` contains some scores from the league that the Wolfpack play in, along with some other leagues. Unfortunately, the data are rather untidy, so we have a fair bit of tidying work to do. Our aim is to create a data frame with the following columns: the date on which each game was played (as text), the name of the home team, the score of the home team, the name of the away team, the score of the away team, and a code for the league in which each game was played (the things like `CH` and `L1` that you see at the end of the line in the data file). Note that the team names have a variable number of words (compare Bradford Bulls and York City Knights, for example). You'll also have to deal with some of the rows being dates and some of them being game results without dates (how do you tell the difference?)

   (a) The data file has one column that has *no* column name. Pretend that the file is a `.csv`, read it in, and display some of the data frame. What name has the one column acquired?

   (b) Construct a (rather long) pipeline that converts the data frame you read in from the file into the desired format. There is some flexibility about how you do this, but you might want to use some of the following tools. If you have not seen them before, you'll need to find out how they work:

   - `mutate` (you'll probably use this a lot)
   - `str_detect`
   - `ifelse`
   - `fill`
   - `filter`
   - `select`
   - `separate`
   - `str_count` for counting words
   - `word` for extracting words from text
   - `extract` if you are clever with regular expressions

   You should hand in your pipeline code and the output it produces.

   (c) Now we can finally do some analysis. How many games were played on each date?

   (d) What were the two highest scores obtained by home teams, and which teams obtained them? Hint: sorting.

   (e) Which were the two lowest away scores, and the teams that scored them? Try the obvious idea first, then find out what goes wrong with it and then fix it (hint: turn text into numbers).

Notes

[1]I don't think there are any rugby league team names like that, but if you were doing this with German soccer teams, there are quite a lot of those with numbers in their names, usually the year the club was formed. The best known of these is Schalke 04, who were formed in 1904 and play in Gelsenkirchen.