**Exploratory Data Analysis Assignment 2**

The attached data (*superbru-10k.Rds*) contains predictions made by 10 000 people for 119 rugby games using the online sports prediction game Superbru. The predictions are for the 2011 Super Rugby tournament, the largest rugby tournament in the Southern hemisphere. In 2011 it consisted of 15 regional teams from New Zealand, Australia and South Africa. These predictions cover the league phase of the tournament (119 games).

If you don’t know anything or have an interest in rugby don’t worry – neither do I! This project is about the judgmental or subjective forecasts that people make, and whether there are any patterns in these forecasts. This could be applicable for just about any kinds of subjective forecasts that people make – financial forecasts, weather forecasts, product sales forecasts, etc.

The superbru.Rds file contains the prediction data and has the following columns

* user\_id: an ID variable to uniquely identify the predictions made by each user
* user\_team\_id: indicates which team the user is a fan of
* game\_id: an ID variable uniquely identifying each of the 119 games
* home\_predicted\_margin: the predicted score difference for that game.

Predictions are made as follows. Each game has a “home team” and an “away team”. The prediction given is the **expected winning margin for the home team** (the home team’s score minus the away team’s score). If this number is positive, that indicates that the home team is predicted to get more points, and hence win. If the number is negative, that indicates that the away team is predicted to win. If the number is zero, a draw/tie is predicted.

The *fixtures.Rds* file has game information and has the following columns:

* game\_id: an ID variable uniquely identifying each of the 119 games (can be used to merge with data in *superbru-10k.Rds*)
* week: an ID variable indicating in which week of the competition the game was held
* home: the name of the home team
* away: the name of the away team
* result: the actual score difference for that game.

The format of the “result” column is the same as for the predictions above – these indicate the points difference between the home team’s score and the away team’s score in the actual game.

**Getting started**

1. Merge together the two datasets so that each row contains both the predicted and actual result. From this, calculate the prediction error.
2. Superbru evaluates users’ predictions by scoring them according to the following rule: the user gets 10 points if the winning team is correctly predicted, and 5 points if the actual margin is within 5 points of the predicted margin. Here are a few examples:
   1. Prediction = +3 (home team wins by 3 points); Result = +2 (home team wins by 2 points)
   2. Prediction = +3 (home team wins by 3 points); Result = +20 (home team wins by 20 points)
   3. Prediction = +3 (home team wins by 3 points); Result = -2 (home team loses by 2 points)
   4. Prediction = +3 (home team wins by 3 points); Result = -20 (home team loses by 20 points)

In case (a) the user would get 15 points (10 points for predicting the winning team and another 5 points for being within 5 points of the actual margin). In case (b) the user gets the team right (10 points) but the predicted margin is more than 5 points away from the true margin (0 points), so the user gets 10 points. In case (c) the winning team is incorrectly identified (0 points) but the margin is within 5 points (5 points), so the user gets 5 points. In case (d) the winning team is incorrectly identified and the margin is way out, so the user gets no points.

1. Add a new column that works out the Superbru points allocated to each prediction.
2. Add your own measure of prediction accuracy (not the Superbru one). So now you have two measures of accuracy: the Superbru one, and your own one.
3. For each of the accuracy measures you have, work out each user’s *cumulative* accuracy after each game (i.e. total up the “points” they have gained in all games up to and including the current one)

The assignment is made up of a few tasks:

**Task 1: Surprising results**

Some games are harder to predict than others. Find some way of assessing which games were most surprising or difficult to predict for the user base as a whole. Show your results using a plot.

**Task 2: Make some rankings**

Construct a leaderboard showing the top 20 users after each week (this is something like what Superbru posts online). Also provide code that will return any user’s rank (their position in the rank order from most accurate to least accurate) at any stage in the tournament (i.e. cumulatively, up to any game).

**Task 3: Check for “the wisdom of the crowd”**

Roughly, the “wisdom of the crowd” effect states that if you take a bunch of people and average their prediction, the result will often be pretty good. One way of interpreting this is that the crowd prediction (the average of everyone’s predictions) will be better than the median prediction (this representing an “average” user). Check if the wisdom of the crowd effect holds for these predictions (there are several ways to reasonably do this). Show your results using a plot.

**Task 4: Fan effects**

You might expect fans of a team to be worse at predicting their team’s performance, because they’re biased. Or you might expect them to be better, because they keep up to date with information and know the team well. See if you can assess which one of these explanations is best supported by the data. You don’t need to use any statistical tests, although you can if you want. Show your results using a plot.

**Task 5: Find the experts**

Are there “expert” forecasters? There are many ways of looking at this question. First, think of what expertise means in forecasting. I’d say there are two kinds of expertise: relative expertise and absolute expertise. Relative expertise just means you’re better than most other people. That doesn’t measure absolute quality though. Absolute expertise means that your predictions are in some objective sense good – but this needs some benchmark for defining good.

The goal of this last section is to use the data to assess whether there is consistent good performance over time. To do this, you could divide your data into two periods: a first period that you use to select your experts, and a second period that you use to test whether your experts are actually any good. Again, show your results graphically.

Write up your work in the form of a short report (max 10 pages) in whatever word processing software you like (e.g. word, latex). The report should contain an introduction to the problem but the majority of the document should be about your results. Your code should be included, either embedded in the document if an Rmd file, or as a separate .R or .Rmd file. The key thing is that I need to be able to run your code and reproduce your results, so there should be clear instructions on how to use your script(s). The code itself should not be displayed in the final typeset document (use “echo = FALSE” for Rmd files) and not pasted as an appendix in your report.

Assignment hand-in is via the "Assignments" tab on Vula, and you should submit a single .zip file containing your report and all code. The submission deadline is on or before 23:59:59 on 16 April 2020.