MATHS 7107 Data Taming
Assignment 1

Trimester 3 2024

Cricket wicket at Takeley Cricket Club, Essex. Source: Acabashi, via Wikipedia

1 Background

A local cricket club has been told that left-handed batters are better than right-handed batters. So for next season, the club is thinking of sacking all of their right-handed players. But they'd like some data analysis before they commit to this plan. (90% of the club's revenue comes from right-handers!)

They've located batting data for players from the international cricket teams of four countries: Australia, England, India and Pakistan. Each of these data sets gives the batting statistics for each player for each year between 2000 and 2005. The data collates the results from all the matches that each player played in that year. A player may not have played any matches in that year, and no player has ever played more than 200 matches in one year.

The cricket club would like you to analyse the data and tell them if left-handers are indeed better batters than right-handers. Conveniently for you, the club president has just started using `R` and `R Markdown`, so she wants your report as a PDF generated using `R Markdown`. She studied Data Taming last trimester, so she wants you to only use commands from the course, so that she can easily see what analysis you've done. In your `R Markdown` code chunks: make sure that you **do not** set `echo = FALSE` so that she can see what `R` code you used to generate your output. But of course, she doesn't want to see irrelevant warnings or messages.

Make sure you write some text (not just comments in the code) explaining what you are doing at step.

1.1 A brief description of batting cricket

In case you are not familiar with the game of cricket, here is a simplified description of batting:

- A match consists of one or two **innings** for each team.
- Each player has the chance to bat (at most) once per innings.
- During the innings the **batter** stands in front of the **wicket** and the **bowler** throws[1] the ball and tries to hit the wicket.
- The batter then attempts to hit the ball away from the wicket.
- If the batter hits the ball, while the other team is running to fetch the ball, the batter can run back and forth to score **runs**.

[1] For those that know cricket, you'll know the word "throw" is controversial, but we're going to stick with it.

  - The batter can score **up to 6 runs** for each ball they hit.

- Once the bowler has retrieved the ball, they then repeat the process by attempting to throw it at the wicket again.
- The batter can go **out** is several ways, including:
  - The bowler succeeds in hitting the wicket with the ball.
  - The opposing side can catch the ball after the batter hits it.

  Once they are out, they cannot score any more runs for that innings.
- Once everybody in the team has gone out, the innings is over and the other team have a turn to bat.

Here's a slightly more detailed description in this video from the International Cricket Council.

1.2 Number of digits

When writing your own text, or **USING** the output from `R`:

- For integer results, report the whole integer.
- For non-integers with absolute value > 1: use 2 decimal places
- For non-integers with absolute value < 1: use 3 significant figures.

  For example:
  - $135.5681 \approx 135.57$
  - $-0.0004586 \approx -0.000459$

If you're just **PRINTING** the output from `R`, then just keep the output as it is.

- Note that if you have `R` do the rounding for you then you need to conform to these two conventions listed above.

2 The data

The company has four datasets labelled `England.csv`, `Australia.csv`, `India.csv` and `Pakistan.csv`. Each dataset contains 8 columns:

- `Player`: The player's name
- `RightHand`: `yes` if the player is right-handed and `no` if they are not.
- `200*`: A description of the performance in the year "200*" (here the asterisk is a **wildcard** character, meaning it stands for any set of characters). There is data from 6 years in total. This column contains information about the player's batting performance that year:
  - The number of innings they played in that year
  - The number of times they went out over all innings that year.
  - The total number of runs they scored that year.
  - The total number of balls they faced.

3 Data cleaning

As you work through the Tasks below, you will need to clean the data.

> **IMPORTANT!**
>
> Make sure you only remove data that you must remove. Do not just delete data because it is inconvenient. You must have specific instructions from the client before you remove any data from your analysis.

Instructions:

- There may be some duplicated rows, in which case, remove one of them.
- Some test data may have been left in. Remove it.
- Any negative numbers should be converted to positive numbers.
- If there are any values that are impossible (in absolute value) then remove the entire row.
- There may be some other typos, so fix them if possible. If they're not possible to fix, then delete the entire row.

4 Your job

> **Note**
>
> Make sure you write text to explain what you are doing at each point and why you are doing it. Also describe the results.

1. Load the correct dataset as a tibble. Output the first 10 rows of the dataset.
2. What are the dimensions of the data set?
3. Set the correct seed, then randomly permute all rows in your data set. *(Hint: a random permutation is like doing a random sample of all rows, without replacement.)* Output the first 10 rows of the dataset.
4. We want to clean up our data, but first we'll put in an extra column of row numbers, so we can track some changes we've made to the data.
   - Add a column at the far left of the dataset called `Rows` that contains the row numbers.

   Output the first 10 rows of the dataset.
5. Now clean the data. Make sure you justify every step of cleaning that you do. Then display the first 10 rows of the the dataset, and the dimensions of the dataset.
   - If you discover any problems with the data in the following questions then you should come back and redo this question before you submit. Your data should be clean and shiny from this point.
6. Next, let's tidy the data.
   (a) Convert the data to a long form by converting the `200*` columns to two new columns called `year` and `performance`.
   (b) Using the new `performance` column, create four new columns
   - `INNINGS`: the total number of innings for that year
   - `OUTS`: the total number of outs for that year
   - `RUNS`: the total number of runs for that year
   - `BALLS`: the total number of balls for that year

   (Note that you just want the numbers in these new columns, not the text.)
   (c) Then delete the `performance` column.
   (d) Since we now have a larger number of rows, let's add new numbers to keep track. Add a new column, second from the left, called `T Rows` with the row numbers of the tidy data set.

Output the first 10 rows, and the dimensions, of the data set.

7. Using dot points, identify what types of variables we now have in our data set, i.e., "Quantitative Discrete", "Quantitative Continuous", "Categorical Nominal", "Categorical Ordinal". (Don't just describe what data type they are in the tibble — you need to think about the type of variable in the context of the meaning of the data.) Make sure you provide some justification for your choice of variable types.

   - Don't just provide vague statements, but be very concrete about describing this particular set of data.

8. Now it's time to tame our data.

   - Make your data set correspond to the Tame Data conventions on page 3 of Module 2. You'll need to use your answers to Q7.
   - You'll need to make sure the `R` data types in your tibble match the variable types that you identified in Q7.
   - *(Reminder: Your data should already be clean by this point. You may want to check here if there is any more cleaning required. If so, go back to Q5 and try again.)*

   Output the first 10 rows, and the dimensions, of your clean, tidy and tame data set.

9. We are going to calculate some statistics for each player, and so we only want to keep data that is non-zero. Remove all rows where the player did not face a single ball. Output the first 10 rows, and the dimensions, of this data set.

10. We will just look at a random subset of your data.

    (a) Setting the correct seed again, take a random sample of 70 rows from the dataset in Q9.

    (b) Sort the dataset by the tidy row numbers.

    Output the first 10 rows, and the dimensions, of your sample.

    > **Note**
    >
    > Use this random subset from Q10 for the remainder of the assignment.

11. (a) Insert two new columns, just to the right of the `year` column:

    - `pct_out`: the percentage of innings where the player was out.
    - `run_rate`: the number of runs scored per ball.

    Describe what type of variables these new columns represent ("Quantitative Discrete", "Quantitative Continuous", "Categorical Nominal", "Categorical Ordinal"). Are the data types correct? (Explain your answer.) If they are not correct, make sure you change them.

    Output the first 10 rows, and the dimensions, of the data set.

    (b) Find:

    i. which player/s had the lowest percentage of outs, and in which year/s this occurred?

    ii. which player/s had the highest run rate, and in which year/s this occurred?

    Make sure you report the values as well.

12. We want to produce a boxplot using the `year` variable, however, we need it to be an `<ord>` data type. (Otherwise, `R` won't know how to deal with it.) So first convert `year` to `<ord>` type.

    Then produce a side-by-side boxplot of the run rate, with `year` on the horizontal axis, and use the `fill` option to split the data according to handedness.

13. Produce two more side-by-side boxplots, this time with the handedness on the horizontal axis and `pct_out` and `run_rate` on the vertical axes. (You can fill with the handedness again to create nicer graphs.)

14. Based on your output in Q11–Q13, write a couple of paragraphs about the cricket club's question. (Everybody's output will be different, so there are no strictly correct answers. But make sure you specifically refer to your output in your discussion here.)

5 Submission

You must submit your assignment via MyUni. Do not email it to the teaching staff. Detailed instructions are on the assignment submission page in MyUni. Make sure that all your output is relevant to the questions being asked.

6 Deliverable Specifications (DS)

Before you submit your assignment, make sure you have met all the criteria in the **Deliverable Specifications (DS)**. The client will not be happy if you do not deliver your results in the format that they've asked for.