MATHS 7107 Data Taming
Assignment 1

Trimester 2 2024

Australian Rules goal posts, RA Cook Reserve, Bedford, WA. Source: Steelkamp, via Wikipedia

1 Background

The Australian Football League (AFL) is the largest Australian Rules football league in the country. Each week during the season, is a new Round, where the teams are paired up to play a match.

There are two types of scores in the game:

- goals: when the ball goes between the two tall posts, worth 6 points
- behinds: when the ball goes between a tall and a short post, worth 1 point

The total score from the game for each team is given by adding up the contributions from goals and behinds. The total number of "shots-on-goal" is the number of goals plus the number of behinds. No team has ever made more than 300 shots-on-goal in a single game.

Every team has a "home" ground, where they are based, and so each game is classed as a "home" game when they are playing at home, or an "away" game when they are not. Currently there are 16 teams in the league (it is the year 2008), and it is expected that in the near future more teams will be added. It is not known what the name of the teams will be, but we know they will be based in Queensland or New South Wales.

Crowds love to see a high-scoring game. The company that sells tickets at the football stadia would like some information about how the total score is related to the team playing home or away. Coincidentally, the company uses `R` and `R Markdown`, so they want your report as a PDF generated using `R Markdown`. In your `R Markdown` code chunks: make sure that you **do not** set `echo = FALSE` so that the client can see what `R` code you used to generate your output.

1.1 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 three datasets labelled `afl_0.csv`, `afl_1.csv` and `afl_2.csv`. Each dataset contains 24 columns:

- `Team`: The team's name
- `State`: The team's home state, one of "VIC, SA, WA, NSW, QLD".
- `Round*`: A description of the team's game in round "*" (here the asterisk is a **wildcard** character, meaning it stands for any set of characters). There are 22 rounds in total. This column contains information about:
  - Home or away status
  - The number of goals
  - The number of behinds

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 convenient. 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 impossibly large (in absolute value) then remove 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 all 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.)* Output all 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 `Row Num` that contains the row numbers.

   Output all rows of the dataset.
5. Now we will do a bit of data cleaning.

   (a) Remove any rows that need removing.
   (b) Correct any team and state names that need correcting.
   (c) Sort the data by the team name.

   Output all rows of the dataset.
6. Next, let's tidy the data.

   (a) Convert the data to a long form by converting the `Round*` columns to two new columns called `round` and `details`.
   (b) Delete the characters "Round" from the `round` column (leaving just the numbers).
   (c) Using the `details` column, create a new column called `home` which contains `TRUE` or `FALSE` (based on the data).
   (d) Still using the `details` column, create two new columns containing the number of "goals" and "behinds".
   (e) Delete the `details` column.
   (f) 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 `Tidy Row Num` 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.
8. Now it's time to tame our data.

   - Make your data set correspond to the Tame Data conventions on page 2 of Module 2. You'll need to use your answers to Q7.
   - Also check if there is any more cleaning that is required. If so, clean the data now. *(Hint: It might be good to check one last time if there is any missing data. The* `is.na()` *command on the Reminder Sheet might come in handy.)*

   Output the first 10 rows, and the dimensions, of your clean, tidy and tame data set.
9. We will just look at a random subset of your data. Setting the correct seed again, take a random sample of 200 rows from the dataset in Q8. Output the first 10 rows, and the dimensions, of your sample.

Note

Use this random subset from Q9 for the remainder of the assignment.

10. (a) Insert two new columns at the right of your dataset

    - `score`: the team's total score in that game
    - `accuracy`: the proportion of shots-on-goal that were goals.

    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. the team/s with the highest average score
    ii. the team/s with the lowest average score
    iii. the team/s with the highest accuracy
    iv. the team/s with the lowest accuracy

    Give both the team name/s and the value.

11. Produce:

    (a) a side-by-side boxplot of the score, with `home` on the horizontal axis. (Use `fill=home` for a nicer looking graph.)

    (b) a side-by-side boxplot of the accuracy, with `home` on the horizontal axis. (Use `fill=home` for a nicer looking graph.)

    From these plots, does it look like the home or away team has an advantage? (Make sure you justify your claim.)

12. Create two new datasets called `afl_home` and `afl_away`, with the data for the home and away rows. Output the first 10 rows, and the dimensions, of each dataset.

13. Use `inspect_num` to find the summary statistics for the home and away data sets. Does this data support the claim you made in Q11? (Make sure you justify your claim.)

14. For each of home and away sets of rows produce a scatterplot of `score` against `accuracy`, with a smoothed trend line. Make sure you put your **independent variable/predictor** on the horizontal axis, and give a brief explanation of why you chose that variable as the predictor.

15. Describe the relationship you see between `score` and `accuracy`. Is the relationship similar for both home and away teams?

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.