# STATS 7022 - Data Science PG
# Assignment 1

Trimester 2, 2024

## Question 1: Data Analysis

A data scientist is routinely required to analyse data. In this question, you will import a dataset, derive some new variables, and produce a simple analysis.

### Question

The `board_game`[1] dataset contains detailed information about board games, collected from a popular board gaming website. Read the `board_game` dataset in to `R` and complete the following steps:

(a) Select the following variables:

- `primary` - the name of each game
- `year` - the year each game was released
- `boardgamemechanic` - a list of mechanics for each game
- `minplaytime` - the minumum playing time (in minutes) for each game
- `maxplaytime` - the maximum playing time (in minutes) for each game
- `average` - the average rating (out of 10) for each game

Drop all other variables from the dataset.

(b) Rename `average` to `rating`.

(c) Remove any games released before 2016 or after 2020.

(d) Create a new variable called `duration` that calculates the average playing time (in minutes) as the geometric mean of `minplaytime` and `maxplaytime`. Remove any games with zero `duration`.

(e) Create a new variable called `dice` that has value `Dice game` if the game's mechanic contains "Dice", and has value `Not a dice game` if the game's category does not contain "Dice" or is `NA`.

(f) Create a new variable called `rating_7.5` that has value `TRUE` if the game has an average `rating` of at least 7.5, and has value `FALSE` if it does not.

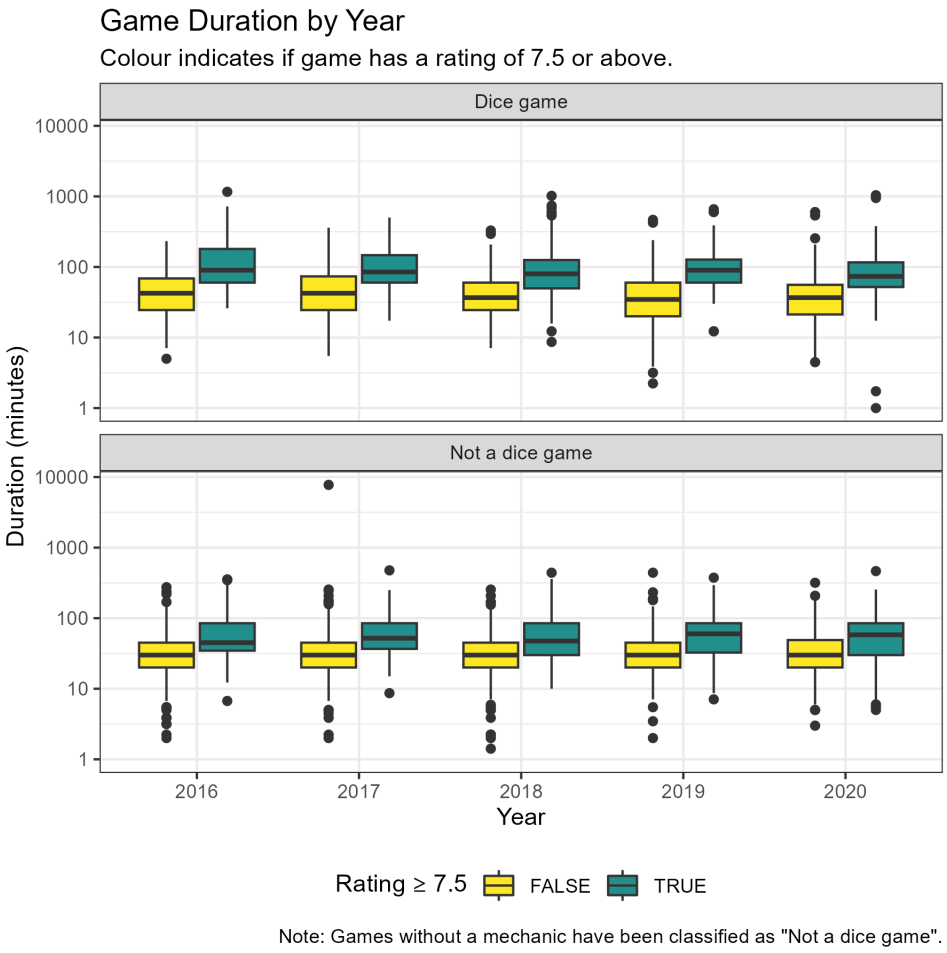


Figure 1: Side-by-side boxplot of game duration (in minutes) by year, separated by whether or not each game is a dice game, and whether or not each game has a rating of at least 7.5.

(g) Produce the plot given in Figure 1.

A copy of the `board_game` dataset is available on MyUni.

For further information, see the assessment rubric on MyUni.

[Question total: 15 marks]

### Submission

A single pdf file documenting your analysis. Your submission should include both your analysis and the `R` code you wrote to generate that analysis.

[1]Data sourced via Tidy Tuesday on 25 January 2022. Note: You do **NOT** need to download the data from Tidy Tuesday. Please download the .rds file from MyUni.

**Checklist**

Please ensure that:

- Your submission includes all `R` output and plots to support your answers where necessary.
- Your submission includes all `R` code.
- Your submission includes a caption for every plot and table.
- Your submission is a single pdf file - correctly oriented, clear, and legible.
- You have submitted your solution by online submission to the correct portal on MyUni. Submissions will not be marked and will receive a grade of 0 if they are not made through MyUni or if they are not made to the correct portal.

The course's regular late policy applies to this assignment. If you need to request an extension, or request a variation to the assessment on medical or compassionate grounds, please contact the course coordinator.

## Question 2: Variance-Bias Trade-Off

A good data scientist understands the mathematics that underpins the methods they use. In this question, we will derive the variance-bias trade-off.

Good data scientists are also good communicators. This includes communicating mathematical theory. Your answer to this question should use mathematical notation that is clear, consistent, and correct, and you should justify all steps beyond simple algebraic manipulation. Marks will be allocated for the clarity of your answer.

### Question

Consider a quantitative response variable $Y$, and $p$ different predictors $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$. Assume that the relationship between $Y$ and $\boldsymbol{X}$ can be written in the general form

$$Y = f(\boldsymbol{X}) + \epsilon,$$

where $f$ is some fixed but unknown function of $\boldsymbol{X}$ and $\epsilon$ is a random error term that is independent of $\boldsymbol{X}$ and has mean zero. In this formulation, $f$ represents the systematic information that $\boldsymbol{X}$ provides about $Y$.

Suppose, using some training data, we estimate $f$ as $\hat{f}$. If the testing data comprises a single new observation $(\boldsymbol{x}_0, Y_0)$, independent of the training data, show that the test MSE

$$E\left[(Y_0 - \hat{f}(\boldsymbol{x}_0))^2\right] = \text{var}(\hat{f}(\boldsymbol{x}_0)) + b_f(\hat{f}(\boldsymbol{x}_0))^2 + \text{var}(\epsilon).$$

[Question total: 11 marks]

### Submission

A single pdf file containing your solutions. You may handwrite your answers and scan them, or typeset your answers.

### Checklist

Please ensure that:

- Your submission is a single pdf file - correctly oriented, clear, and legible.
- You have shown all of your working, including probability notation where necessary.
- You have submitted your solution by online submission to the correct portal on MyUni. Submissions will not be marked and will receive a grade of 0 if they are not made through MyUni or if they are not made to the correct portal.

The course's regular late policy applies to this assignment. If you need to request an extension, or request a variation to the assessment on medical or compassionate grounds, please contact the course coordinator.

## Question 3: ROC Function

One of the best ways to understand an algorithm is to implement it yourself. In this question, we will implement the ROC algorithm.

### Question

In R, write a function called `get_ROC()`. Your function should take two vector inputs:

- `obs`: a vector of the true observed values, ‘‘A’’ and ‘‘B’’.
- `pred`: a vector of predicted probabilities that each observation is ‘‘A’’.

Your function should treat ‘‘A’’ as positive/success and ‘‘B’’ as negative/failure.

Your function should return a tibble with three columns:

- `threshold`
- `specificity`
- `sensitivity`

Your results in these columns should be consistent with the output of the function `yardstick::roc_curve()`.

Here is an example to illustrate its use:

```
df <- tibble(
  obs = rep(factor(c("A","B")), each = 2),
  A = rep(c(0.8,0.2), each = 2)
)
get_ROC(df$obs, df$A)
```

```
## # A tibble 4 x 3
##   threshold specificity sensitivity
##       <dbl>       <dbl>       <dbl>
## 1    -Inf             0           1
## 2       0.2           0           1
## 3       0.8           1           1
## 4     Inf             1           0
```

Here is a template you can use:

```
# Function to return ROC
#
# INPUT
# Two vectors:
```

```
#   obs: a factor with two levels, "A" and "B"
#   A: the predicted probability that each observation
#      is "A"
#
# OUTPUT
# A tibble with 3 columns: threshold, specificity, and
# sensitivity.

get_ROC <- function(obs, A) {
  # YOUR CODE HERE
}
```

Your function should also ensure that the correct type of input is passed to the function. For further information, see the assessment rubric on MyUni.

[Question total: 10 marks]

### Submission

A single R file containing your function. Your code will be automatically unit-tested, so any changes to names etc. may result in a loss of marks. Your function must not use ROC functions from other packages - the code must be your own work.

### Checklist

Please ensure that:

- You have submitted your solution by online submission to the correct portal on MyUni. Submissions will not be marked and will receive a grade of 0 if they are not made through MyUni or if they are not made to the correct portal.

The course's regular late policy applies to this assignment. If you need to request an extension, or request a variation to the assessment on medical or compassionate grounds, please contact the course coordinator.

## Question 4: Technical Communication

There are four key skills in a data scientist's arsenal: analysis, mathematics, coding, and communication. In this question, we'll practice your communication skills.

Whenever you are communicating, it's important to consider your audience. Different audiences will need to know different information, and be able to understand different levels of technical detail.

### Question

Write a brief overview explaining data cleaning and how to perform it in `R` using `tidyverse/tidymodels`. Your overview should discuss:

- why data cleaning is important;
- what data cleaning entails; and
- how data cleaning is performed in `R` using `tidyverse` and `tidymodels`.

Your overview should be suitable for a student who is currently studying STATS 7022 Data Science PG. Marks will be allocated for the quality and correctness of your writing, as well as how suitable your writing is for the intended audience. Your overview must use Harvard-style referencing as detailed here:
https://www.adelaide.edu.au/library/referencing-support

For further information, see the assessment rubric on MyUni.

[Question total: 14 marks]

### Submission

A single doc or pdf file containing your overview. Your overview should be 1-2 A4 pages, with 10-12pt font, single line spacing, and standard page margins.

### Checklist

Please ensure that:

- Your submission is a single doc or pdf file - correctly oriented, clear, and legible.
- Your submission is not longer than 2 pages (including tables and figures).
- Your submission is correctly referenced. For more information, see the link above or visit the Writing Centre.

- You have submitted your solution by online submission to the correct portal on MyUni. Submissions will not be marked and will receive a grade of 0 if they are not made through MyUni or if they are not made to the correct portal.

The course's regular late policy applies to this assignment. If you need to request an extension, or request a variation to the assessment on medical or compassionate grounds, please contact the course coordinator.