# Assignment 3

## STAT7123/8123: Statistical Graphics

### Submission details

Weight: 40%

Total Marks: 70 (PDF file) + 2 (`.rmd` file) = 72

Due Date/Time: 11:55pm Friday 3rd November 2023

Submission via: the link in iLearn to Turnitin

Formatting details:

- Your assignment must be written in `rmarkdown`, and both the PDF and `.rmd` file must be submitted. The PDF will be submitted via Turnitin on iLearn and the `.rmd` will be submitted via file upload on iLearn.

Late Penalties: Standard Late Penalty applies (see the unit guides/Assessments block of iLearn for details).

### Purpose

The purpose of this assessment is to provide you with the opportunity to demonstrate your ability to communicate your understanding of statistical graphics. This is a key capability in all professions whose role may entail explaining such graphics to non-experts.

### Outcomes addressed

This assessment addresses the following unit outcome/s:

**ULO1:** be familiar with important and contemporary examples of graphics, and be able to use them.

**ULO2:** be aware of the elements of graphical design, and use them to critically appraise presented graphics in articles and web pages and suggest appropriate ways to improve them.

**ULO3:** use the computer to generate appropriate graphics using particular packages or languages and be able to develop the ability to do so in others.

**ULO4:** be familiar with a range of modern multivariate graphical techniques and know when it is appropriate to use them.

**ULO5:** use statistical graphics to investigate and analyse data, check statistical model assumptions and effectively present the results of statistical investigations graphically to a range of audiences.

### Skills assessed

Using the methods and techniques described in Lectures and SGTAs, this task allows you to demonstrate:

1. Your written communication skills
2. Your ability to create different statistical graphics
3. Your ability to create, select, and visualise statistical models.
4. Your proficiency in using rmarkdown to present results.

### Task overview

For this assessment, you will answer a series of research questions using `R` statistical software with the tidyverse suite of commands. You will present the results via a PDF generated by compiling an rmarkdown template file. You will also submit the original file with descriptions of code intent (annotations) within each code chunk.

### Detailed Instructions

1. Access the data from iLearn and load the data into `R`.

2. Use tidyverse to manipulate the data where necessary to answer questions 1 and 2.

3. Use `ggplot` to create the required graphics.

4. Your submission will be made up of the PDF file [70 Marks] and the `.rmd` file [2 Marks].

The questions begin on the page 4.

### Quality Criteria

A high-quality submission will:

1. Use the tidyverse suite of commands to manipulate data and statistical models
2. Present professional-quality graphics using ggplot
3. Provide accurate interpretations of each graph where required
4. Provide detailed and clear written responses where required
5. Provide the assignment with code and code intent (annotations) arising from using `echo=TRUE` in each chunk presented neatly within the PDF.
6. Deliver an rmarkdown file that can be compiled to produce the submitted PDF.

## Question 01 [40 Marks]

The NSW government provides yearly data about student attendance rates at individual government schools from 2011 to 2022. The dataset `school.csv` includes the attendance rate for every year. Some initial cleaning has been done to the data. The variables of this dataset are as follows:

- **school_name:** Official school name.
- **attend_YYYY:** Attendence rate for the year `YYYY` (Takes all values between 2011 and 2022)
- **indigenous_pct:** The percentage of full-time equivalent (FTE) students enrolled, as reported under the NSSC, who identify as Aboriginal or Torres Strait Islanders and are accepted in the community with which they associate.
- **selective_school:** Secondary school selective flag. Options include: fully selective, partially selective, not selective.
- **school_gender:** School population gender status. Options include: co-ed schools, girls schools, boys schools.
- **lga:** Local government area.
- **asgs_remoteness:** Remoteness category based on the ABS' Australian Statistical Geography Standard (ASGS) 2016 remoteness structure. Options include: Major Cities, Inner Regional, Outer Regional, Remote and Very Remote.
- **enrolment_2022:** Full-time equivalent (FTE) student enrolments in 2022 as reported under the Australian Bureau of Statistics (ABS) National Schools Collection (NSSC).

Use the `school.csv` data set to answer the following questions. For each question, provide one or more plots created using `ggplot`, followed by an accurate interpretation of each plot (if instructed to do so). You are also required to annotate your `.rmd` file.

a. [6 Marks] Make a dumbbell plot that shows that change in average attendance between 2011 and 2022 for the 5 different school remoteness categories. In 2-4 sentences, comment on what the dumbbell plot shows in regards to the changes in student attendance rates.

b. [8 Marks] Create an alluvial plot with 2022 data, using three categorical variables on the axes: `selective_school`, `school_gender`, and `asgs_remoteness`. In 2-4 sentences, comment on what the alluvial plot shows in regards to school types.

c. [8 Marks] For this question, select 5 different schools (You are required to show which schools you picked). Create two plots; (i) a line plot, and (ii) a stacked area plot for the 5 schools showing the changes in percentage attendance over time. Which school had the highest attendance percentage in 2018? Provide 2-4 sentences which comment on the benefits and drawbacks of each plot.

d. [5 Marks per graphic = 10 Marks available] Create **two** different plots of your choice to explore the school attendance data further.

   **Note:** Higher marks will be awarded for; using more complex graphics to represent relationships (e.g., plots where you have used more data manipulation, added more layers, judiciously used faceting), using a variety of graphics (e.g., two different graphics, rather than all in the same format), and using multiple variables in one plot to provide greater insight into the data (e.g., comparing trends). Plots should present different information.

e. [8 Marks] Understanding, analysing, and communicating data is essential for any analytics role. Job descriptions for such positions often emphasise the importance of statistics skills to interpret and communicate data effectively with a broad audience. Within 150 words, provide an informative, coherent and precise summary of your finding from the graphical analysis of the given data set. On top of the 150 words, provide 2-4 insights from the data (these can be done as bullet points).

   **Note:** Tools for generating text cannot be used.

## Question 02 [30 Marks]

The data set, `soccer.csv`, includes information on soccer match data from $2021 - 2022$. The dataset has been altered to contain information from games where one of the teams won. The dataset contains information about the number of shots, number of shots on target, number of fouls, number of corners, number of yellow cards, and number of red cards by both the home and away teams. The names of the teams have been removed from the dataset.

- **hs:** Number of shots taken by the home team
- **as:** Number of shots taken by the away team
- **hst:** Number of shots on target by the home team
- **ast:** Number of shots on target by the away team
- **hf:** Number of fouls by the home team
- **af:** Number of fouls by the away team
- **hc:** Number of corners taken by the home team
- **ac:** Number of corners taken by the away team
- **hy:** Number of yellow cards received by the home team
- **ay:** Number of yellow cards received by the away team
- **hr:** Number of red cards received by the home team
- **ar:** Number of red cards received by the away team
- **home_win:** Indicator of whether the home team won (0 = "No", 1 = "Yes")

Use the `soccer.csv` data set to answer the following questions. For each question, provide one or more plots created using `ggplot`, followed by an accurate interpretation of each plot (if instructed to do so). You are also required to annotate your `.rmd` file.

a. [5 Marks] Plot the correlation matrix for all the appropriate variabes from the `soccer` dataset. In 2-4 sentences, comment on the different correlations between variables.

b. [5 Marks] Graphically explore the relationship between `hs`, `as`, and `home_win`.

   **Note:** Marks are awarded for describing what the plot(s) are showing.

c. [15 Marks] Fit a model for home side wins, `home_wins`, to understand what factors are possibly influencing home side wins. This model should be a fit a `glm` model with a binomial family (`family = "binomial"`). Then, use the `broom` package to tidy up your model output and provide an interpretation of the model. Provide 2-2 sentences on if and how any of the model diagnostic plots look different to those seen in the lecture content and SGTA material.

   **Note:** Higher marks are associated with fitting a better model. You should also breifly comment why you ended up with the model you provide.

d. [5 Marks] Plot the confidence interval(s) for the regression coefficient(s) from the model in part (b). Write 2-3 sentences that discuss what is shown in the plot.