Mathematical Foundations of Data Science
Assignment 1

Trimester 3, 2024

1. Suppose that Darling West G14 (the room in which seminar classes for MATHS 7027 are held) is completely filled with golf balls. What would be the total mass of the golf balls?
   Note that you are not expected to get an exact answer! This is an exercise in *estimation* - there is no one correct number that will get you full marks. Rather, you must give a reasonable argument with plausible estimates following the principles presented in the course materials. You need to justify your assumptions and make clear how you arrived at your answer.

2. Consider the following statement in set notation:

   $$\forall x \in \mathbb{Z},\ \exists y \in \mathbb{Q} \text{ S.T. } x = 2y + 1$$

   Explain in words what this statement means, and state whether it is a true or false statement, giving justification.

3. Consider the function

   $$f(x) = \frac{3 + 2x^2}{5},\ x \geq 0.$$

   Find $f^{-1}(x)$, clearly stating the domain of f^{-1} and showing all working.

4. Let $A = \{\frac{1}{2}, 1, 2, 3\}$, $B = \{x \in \mathbb{Q} \mid 3x + 1 \in \mathbb{N}\}$, and $C = (0, 1] \subset \mathbb{R}$. Determine the following, showing all working:

   (a) $A \cap C$.

   (b) $C \setminus B$.
   Hint: Express your answer as a union of intervals.

   (c) $A \cup (B \cap C)$.

5. You should complete this question using a Jupyter Notebook. All of the code you will need to complete this question can be taken directly or generalised from the week 1 practical, or will be given to you in the question.

   Download the file `series_data.csv`[1]. This file contains data about television series listed on IMDB. Use `Python` to do the following:

[1] This data was originally sourced from Kaggle.

(a) Using `pandas`, read the data into a dataframe and print out its `head()`.

(b) Create a histogram of the average IMDB rating (`IMDB_Rating`) for all TV series in the dataset.
Hint: Remember to add axis labels. Note that we are asking for the series' rating, not how many votes it received.

(c) Calculate the mean IMDB rating for all TV series. Print out a statement showing the mean rating, rounded to the nearest 3 decimal places.
Hint: In the week 1 practical, we saw how to print out a line containing text and numbers. The `round()` *function might also be useful here.*

(d) Among TV series with an IMDB rating above 8, find the series with the most votes.
Hint: Try creating a new dataframe that only contains series with an IMDB rating of more than 8.

(e) Now find the "Crime" series with the most votes among series with an IMDB rating above 8.
Hint: Many series have multiple genres. We want to include all series with "Crime" in the genre, not just series where "Crime" is the only genre. The string method `.str.contains()` *might be useful here.*

Present your answers as a full Jupyter Notebook. Your notebook must include code to find the results, and text answering the questions based on the output of your code. Download this notebook and convert to a PDF and submit with your assignment.

Please note: You must include the code you used to find results. Each answer submitted without code will receive a mark of 0.

Hint for submitting: You can "Download As PDF" in Jupyter, but that may not work on your computer. If it doesn't, you can download as HTML and convert that to a PDF. Make sure you join it to your assignment to make a single PDF when submitting! You might want to try googling things like "convert html to pdf" and "combine multiple pdfs". There is also a video in the Python Module on MyUni demonstrating how to save a Jupyter Notebook as a PDF.