Introduction to Data Science

**Final Project Guidelines**

As detailed in the syllabus, you are required to complete a final project based on the material learned in class (lectures and tutorials) throughout the semester. This project aims to expose you to the processes and tools that are being used in the industry.

General Information

**Credit** – 30% of the total grade (10% for each part)

**Deadline** – June 20, 2022

**Team Size** – 2

The final project consists of **three parts**, as follows:

1. **Prediction – Kaggle Competition**

In this part, you are required to classify a comment posted on reddit as sarcastic or not given the provided features or by generating new features.

The project's dataset is available for download in the [attached link](https://www.kaggle.com/t/a578c4e527724e9098a4c2bba879fbb8), which will also serve for predictions submission. To submit your predictions, you will need to join the competition and establish a team.

As this is a competition, you can submit two predictions a day. The leaderboard will show your position according to the public score, defined as the classification accuracy of 30% of the test data.

The competition's deadline is **June 19**, make sure to pick your final prediction before the due date.

1. **Exploratory Data Analysis & Inferential Statistics**
   1. Inspect the data by computing the summaries of features which in your opinion are useful features.
   2. Visualize the data in *at least three* distinct graphs. Then, in part 3, report the top 3 most valuable and informative graphs.
      1. You are allowed to use either {ggplot2} or any other special package dedicated to visualizing a specific type of graph. That is, you are not allowed to use {base} or {graphics} packages.
      2. Explain each visualization (e.g., variables choice, graph type).
      3. In *at least one of the visualizations*, add visual representations of the summaries you computed in subsection A.
   3. Conduct two statistical tests of your choice, based on any feature in the data (whether provided or custom-generated).
      1. Write down the hypotheses.
      2. Check the underlying assumptions (if necessary).
      3. Conduct the test appropriately.
      4. Report and interpret the results. State your conclusions clearly.

1. **Report**

Finally, compile your project into a research report, reflecting on your approach to handling the data, creating the visualizations, fitting different models, etc. Walk us through your progress and way of thinking tackling the project.

Your report should include code, results, visualizations, and interpretations of parts A and B. You can use either R Markdown or Word to create a report. If you are using Word, submit your R code separately.

To sum up, you need to submit:

1. Your final prediction to Kaggle until 23:59 on June 19.

2. A written report & R code to Moodle until 23:59 on June 20.