**UNIVERSITY OF CAPE TOWN**
**DEPARTMENT OF STATISTICAL SCIENCES**
**MSc - Data Science Speacialisation 2024**

**STA5092Z - Exploratory Data Analysis Assignment 1**

---

Due date: Tuesday, 28 February 2024 at 6:00pm
Late submissions will be penalised at 10% per day (pro rata)

General Instructions:

- Prepare a report that answers the questions. All analyses to be conducted via R and if you wish to use Python, be aware of the fact that our language is R. All code must be submitted on Vula, and should include comments that clearly explain the purpose of the code.
- **NB:** The length of your report (excluding appendices) is limited to a maximum of **10 pages**. Any information beyond the $10^{\text{th}}$ page will be ignored.
- You need to prepare your coding in R Markdown format and submit both your final pdf as well as the .Rmd file. It is expected that you create the pdf in scientific writing format (ie. article).
- The submissions will be done on VULA under the assignments tab.
- Please attach a plagiarism declaration to your report. Plagiarism of any form will be reported to the university court.

---

## Dilemma of Changing Majors or Persevering

This problem dataset comes from a renowned university's undergraduate students in their first and second year of studies.

Data were collected for 2015-2022.

---

# PART I - Data Wrangling

For this part, follow the instructions below and explain every step of your actions in detail.

1. Download the zipped file of the data from Vula.

2. After unzipping the downloaded file, you will notice that there are 24 .csv files. The following variables are common in all datasets:

   - **Term**: indicates the year of study.

   - **Catalog Nbr**: indicates the courses in consideration. These courses are mathematical courses. A1 and A2 are 1st year courses that cannot be taken together. A1 and A2 can be taken if certain mathematical foundations are obtained. B1 is a 2nd year first semester course with a pass prerequisite of either A1 or A2. B2 is a 2nd year second semester course with a pass prerequisite of B1. For example, if a student is a 1st year student in 2016, then the student can take either A1 or A2. If the student has obtained $\geq 50\%$ then they can take B1 in the 1st semester of 2017. If they fail B1, then they will have to take B2 only in the 2nd semester of 2018 or they can opt not to do B2 all together not to extend their degree.

   - **Acad Prog**: indicates the degree the student is studying. Hence a student might switch between degrees from one year to another.

   - **Grade**: indicates the mark obtained. All in percentage terms. DPR: did not duly perform to write the exam, INC: incomplete, UP: passed the course in a supplementary exam, and AB: absent, or did not write the exam. Basically, DPR, INC, and AB are fail; whereas UP means that a student passed the course with a supplementary exam, no numeric mark is provided. A pass at this university means a student obtained $\geq 50\%$. In year 2020, Covid hit and there were no numeric marks published. If a student passed the course, it was marked as PA.

   - **StudentID**: indicates the ID of the student. This is anonymised.

   Extract the data .csv files into R, convert them into tibble format.

3. Merge the files into one single file called undergraduate.
   Hints:

   (i) You need to take care of the number, text in the grade column.

   (iii) Check the types of the variables in each file. Record any changes that are necessary to make sure that all variables are coded correctly and why?

# PART II - Data Summaries and Visualizations

Use the merged dataset and attempt the following questions with both tables and appropriate figures. Make sure you explain your findings from each figure and table in detail:

1. What is the total number and percentage of students for each year who pass both the 1st and the 2nd year courses in the same year?
2. Are there any outliers in the Grade variable? What is your definition of an outlier?
3. Are there any obvious clusters in the Grade variable? i.e. more marks around 50%.
4. What is the pass rate for the 1st and 2nd year courses for each year? Are there any differences between A1 and A2?
5. How many students fail B1 even if they pass A1 or A2?
6. What would you suggest the minimum mark obtained should be from 1st year courses so that the student is able to pass B1?
7. What would you suggest the minimum mark obtained should be for B1 so that the student is able to pass B2?
8. Is there any hope for students with a UP (supplementary exam) for A1 or A2 to make it to the 3rd year of their studies without failing B1 and/or B2?
9. What is the correlation between the Grades of 1st and 2nd year courses?
10. Determine if answers for any of the above questions have been affected by Covid years? UCT has experienced other disasters, protests. Do you see any impact of these on your answers from one year to another?
11. What kind of an informed suggestion/advice one can make to a student who is registering for A1 or A2 for the first time? What would you look at to decide from the very beginning that the student will be successful in these courses?
12. Suggest two possible additional questions that may be answered with the data and perform the relevant analyses to answer the questions you have suggested.

Good luck!