## 1. OBJECTIVES

In this assignment, you will use a given dataset to:

* prepare data for analysis
* conduct data exploration
* interpret the data using statistical techniques
* document the analysis, using Jupyter notebook with python codes and markdown text.

**2. DATASET: STUDENT-GRADES**

Students score (‘**student-grades.csv**’) have been randomly sampled from 5 different fictitious diploma courses (ACAD\_PROG).

Their results for 4 different modules (Math, Science, Comms, Tech) are recorded.

Each module has a final score (denoted with a \_SemMark).

Each module’s final score is calculated from components score, and each module have different components.

Each row corresponds with the grades of one student.

There are some missing values as some students are exempted from certain modules.

There are a total of 335 students in this random sample.

Detailed information (i.e. column description) is provided below.

* ACAD\_PROG: There are 5 different diplomas, DipID, DipMedia, DipNet, DipCloud and DipFacilities.

* Math.SemMark is the final mathematics marks. It has the following components: Math.CA, Math.TEST1, Math.TEST2 and Math.TEST3.

* Science.SemMark is the final science marks. It has the following components: Science.CA1, Science.CA2, and Science.ASG1.

* Comms.SemMark is the final report writing marks. It has the following components: Comms.ASG1, Comms.ASG2, Comms.CA1 and Comms.CA2.

* Tech.SemMark is the final Technical marks. It has the following components: Tech.CA1, Tech.CA2 and Tech.ASG1.

## 3. TASKS

You have been tasked to answer some questions about the students’ performance. Using the appropriate data from the dataset provided, perform statistical techniques learnt to provide answers to these questions using python in a Jupyter notebook:

Question 1: **Is there a correlation between good performance in one module and another? (20 Marks)**

Hint: Perform suitable technique to deal with missing values. There are multiple ways to deal with missing values, provide reasonable explanation for the technique used. Pick an academic program and determine if there is a correlation between the results of two modules of your choice. How did the correlation compare across students from different academic programs?

Question 2: **Is the distribution of marks similar across academic programs? Is there a particular program that fared better?**

### (15 Marks)

Hint: Pick one module to compare across different academic programs. Comment on the central tendency and dispersion, and distribution across different programs taking the same module.

Question 3: **Students from DipNet consistently perform better. Is that a fair statement? Use hypothesis testing to prove or disprove this statement.**

### (15 Marks)

Hint: Set a suitable significance level (0.05?) and use an appropriate hypothesis test to compare DipNet students results to the rest of the students.

Question 4: **Is the distribution of grades independent of the academic program? (15 Marks)**

Hint: Use a suitable chi-square test to examine if grade distribution is indeed independent of the academic program the student is in.

Question 5: **What is the breakdown of each component to the final SemMark? For example, CA, Test 1, Test 2 and Test 3 contribute 40%, 20%, 20% and 20% respectively to the final semester mark for Math (Math.SemMark). (15 Marks)**

Hint: Pick one module and using linear regression, find the contribution of each component toward the final Semester mark. Explain using the statsmodel summary table to support your findings.

Question 6: **If we know a student’s Math, Comms, Tech SemMarks, can we predict the Science SemMark? Will knowing the academic program increase the accuracy of the prediction?**

### (20 Marks)

Hint: Perform linear regression on the dataset to produce a model. Examine the data for outliers that could have high influence or leverage. Split the data into train and test set. Perform linear regression on the dataset to produce a model. For categorical data, use encoding before including it in the model. Evaluate the model.