# CREDIT TASK

## About this task

**Step-1**

This task is designed to assess the Credit level expectations. There are some coding based questions in this task, please get familiar with the data set first and then answer the questions with your code and necessary explanation.

**Step-2**

Your tutor will then review your submission and will give you feedback. If your submission is incomplete the tutor will ask you to include missing parts. Tutor can also ask follow-up questions, either to clarify something that you have submitted or to assess your understanding of certain topics.

## Feedback and submission deadlines

**Feedback deadline:** Monday 15 April (No submission before this date means no feedback!)
**Submission deadline:** Before creating and submitting portfolio.

**Background**
The Victorian Government is committed to improving the oral health of Victorians. A person's oral health is key to their overall general health and wellbeing. Extending community water fluoridation is important for public health. Melbourne has had fluoridated water since 1977. Other parts of Australia have had fluoridated drinking water for more than 50 years. Community water fluoridation is the most effective population-wide intervention to prevent tooth decay.

**Dataset**
Dataset file name: FluoridationData.csv

Dataset description: The dataset contains different features along with the fluoridation status of different water companies. It contains total 7 features. It contains different types of data including boolean, float and string. Feature names, data type and values are described in the following link. Each observation is a datapoint along the row of the dataset. The data set can be found in :
https://discover.data.vic.gov.au/dataset/victorian-water-fluoridation-status-by-postcode .

# Evidence of Learning – SIT307/SIT720

Answer the following questions in a .jpynb file, execute your code and keep the output, submit the .jpynb file to **Ontrack** (**https://ontrack.deakin.edu.au**)

1. Load the data from supplied data file. Print the data dimension.

2. Continue from question 1. Display the data type of all features. If the data type is float, print the median values of the features.

3. Continue from question 2. Print all the possible values of the feature "fluoride_level" and calculate the ratio of each "fluoride_level" value.

4. Is there any association between "melbourne" and "fluoride_level"? Explain your results from given dataset.

5. Print the number of water companies for different suburbs. Please report the pattern found in the result, if any.

6. Continue from question 5, which suburb has the biggest number of water companies?

7. Continue from question 6, which suburb has the biggest number of fluoridated companies?

8. Create and print a data frame of the number of water companies at different fluoride levels for different suburbs.

9. Continue from question 8. Draw a histogram of the top 10 suburbs against its number of fluoridated companies. Explain the result.

10. Based on the original dataset, use the available features and perform clustering on all the water companies and determine the number of clusters. Is this the same as the number of suburbs in the data set?

11. Continue from question 10, choose the best K and perform K-Means on the data set, report the purity score.

12. Continue from question 11, perform K-Means++ on the data set, report the purity score and explain whether the K-Means++ returns better/worse result than that of K-Means.

13. (This question is for SIT720 students only ) Apart from K-Means and K-Means++, try another clustering method, and compare the results.