**Evaluation & Marking**

Students will be provided with a Jupyter Notebook which contains a set of questions that they will need to work out and complete.

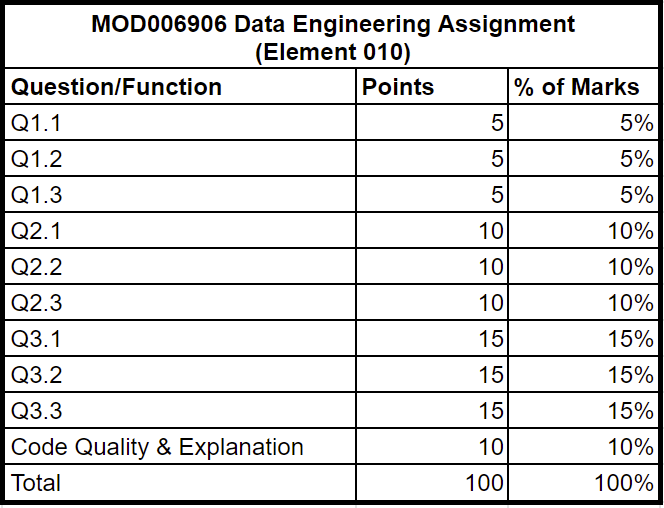
Each question point is weighted equally. However, more complex questions will have more points. Additional marks (10%) will be awarded on the basis of code quality**. You will need to achieve a score of at least 50% to be awarded a Pass in this assignment.**

In this assignment, there are 90 points allocated to 9 different coding questions, as indicated in the marking scheme below. There are an additional 10 marks available for code and explanation quality, making a total of 100 possible marks.

Code and explanation quality will be assessed on the basis of:

* Functions/methods should do one well-defined task
* Variable names should reflect what the variable holds and may contain nouns but should not contain verbs
* Code should be properly indented
* Localise variable and function/method scope to the greatest degree possible
* Code should be as efficient as possible (time, complexity)
* Code should be clearly and appropriately commented

**Mark Scheme**



**Assignment (Element 010) specification**

You are required to implement 9 functions in Python using the template file (Data Engineering 1.ipynb) and the data file (data/sfscores.sqlite) provided. The specification of each function is given below, as well as in the comment part of the function definition in each of the template files.

In this assignment, you are working with a relational dataset that captures information about food health investigations carried out in San Francisco and their outcomes.

Imagine you want to start a new restaurant in San Francisco.

You want to stay away from big restaurant owners who own multiple restaurants.

You also want to pick an area that is trending (hence, a lot of restaurants).

In this exercise, you will write SQL queries to find out the best area and answer business intelligence questions.

The database is stored under data/sfscores.sqlite for you to test your queries with, but for this assignment, your answers should only be the SQL queries you construct.

The database consists of *3 tables*. The schemas are shown in the notebook.

* businesses: information relating to restaurant businesses
* inspections: information about individual inspection events
* violations: information about violation events

The queries you need to implement will increase in difficulty. By the end of this assignment, the focus will be on answering multipart business questions using multistep queries or multiple tables.

# Part 1 - Essentials

## Q1.1 [5 marks]

Write a string which contains a specific SQL query.

The query should find and return the total number of business\_id in the businesses table.

## Q1.2 [5 marks]

Write a string which contains a specific SQL query.

The query should find out how many unique owner names are registered with the San Francisco Food Health Investigation organisation.

The column of unique owner names should be renamed as "unique owner name count".

## Q1.3 [5 marks]

Write a string which contains a specific SQL query.

The query should find out the earliest and latest dates on which a health investigation is recorded in the database.

Give the columns suitable names.

# Part 2 - Groupby

## Q2.1 [10 marks]

Write a string which contains a specific SQL query.

The query should report the distribution of the risk exposure of all violations reported in the database.

The first column of the result should be "risk category".

The second column should be the count of that risk category.

## Q2.2 [10 marks]

Write a string which contains a specific SQL query.

The query should report the distribution of the risk exposure of all the violations reported in the database that are **related to temperature**. That is, they contain the word "temperature".

Sort the results in order of highest to lowest frequency.

## Q2.3 [10 marks]

Write a string which contains a specific SQL query.

The query should find the restaurant owners (by owner\_name) who own one or more restaurants in the city, and the number of restaurants they own.

Report only the top 10 owners.

Order this top 10 in descending order of the number of restaurants.

# Part 3 - Subqueries and joins

## Q3.1 [15 marks]

Write a string which contains a specific SQL query.

From the businesses table, select the top 5 most popular postal\_code.

Filter these to only count the restaurants owned by people/entities that own 10 or more restaurants.

The result should:

* return a row (postal\_code, frequency) for each 5 selections
* sort by descending order to get the most relevant postal codes

## Q3.2 [15 marks]

Write a string which contains a specific SQL query.

Let's get an idea about the inspection score the competition has.

Based on multiple inspections, find out:

* the minimum Score (as min\_score)
* the average Score (as avg\_score)
* the maximum Score (as max\_score)

for all restaurant in postal code 94158.

The average score should be rounded to one decimal.

## Q3.3 [15 marks]

Write a string which contains a specific SQL query.

Look at how many times restaurants with postal code 94158 have committed health violations and group them based on their risk category.

The output should be (risk\_category, count as frequency) and sorted in descending order by frequency.