SCHOOL OF COMPUTING

# CA1 Specification

## EP0302 Programming for Data Science

### 2023/2024 Semester 1

**Assignment rubrics**

1. Demonstrate basic competency in writing Python programs
2. Demonstrate basic competency in using the Python Numpy and Matplotlib packages for data analysis and data visualization
3. Demonstrate basic competency in applying the insights gained from the outputs of your Python programs to deliver a useful data analysis presentation for your stakeholders

# Table of Contents

# Section 1
# Instructions and Guidelines

1. This is an **INDIVIDUAL** assignment which requires the student to write Python code that retrieves data from CSV text files and perform basic data manipulation operations such as, transformation and visualization on the data.

2. The requirements of this assignment are outlined in Section 2 of this document.

3. The deadline of this assignment is on **9 June 2023 (23:59).**

4. Submissions should be made via the **Brightspace CA1 Assignment Submission link** by the stated deadline

5. Deliverable should be a zip file with the following file-naming convention:
   **CA1-[ElectiveClass]-[AdminID]-[Name].zip**

6. Zip file should include the following items:
   - **One** Jupyter notebook that accomplishes the given tasks using the Python programming language. The notebook will also document the data insights that you have gained through the Python code you have written
   - **One** HTML exported version of the Jupyter notebook
   - **All** datasets (.csv files) used (including the recommended datasets)
   - **One** Declaration of Academic Integrity (SOC)

7. As part of the assignment requirements, you will be having an interview using the Jupyter notebook you have prepared. Your module tutor may ask you to reproduce certain parts of your code during this interview session. Codes that are reproduced need not be exactly the same but the code should be able to perform the task in question. Usage of Google will be allowed during this questioning process. AI tools will not be allowed during the interview.

8. This assignment will account for **40%** of the **module grade**.

9. No marks will be awarded, if:

   a. The work is copied or you have allowed others to copy your work.
   b. If you are unable to reproduce most major parts of your code.
   c. If you use other packages besides Numpy and Matplotlib.
   d. Your zip file is corrupted. (please double check by downloading)

10. 50% of the marks will be deducted for assignments that are received within ONE (1) calendar day after the submission deadline. No marks will be given thereafter.

# Section 2
# Scope of the assignment

In this individual assignment, you are required to produce a data analysis presentation for at least 3 datasets belonging to the **Infrastructure Sector** based on the requirements as stated below.

## Basic Requirements

1. We have chosen 2 datasets for you to work on:
   a. https://data.gov.sg/dataset/resale-flat-prices
   b. https://data.gov.sg/dataset/median-rent-by-town-and-flat-type

2. On top of the 2 datasets we have chosen, you are required to choose **at least one more** datasets regarding **Singapore's infrastructure**. One resource you can get this data from is: https://data.gov.sg/search?groups=infrastructure
   You are encouraged to choose datasets which are interrelated and support a central theme of investigation. You must come up with your own questions and answer it based on the analysis of your data.

3. Your Jupyter notebook **MUST** include the following:
   a) Your name and the title of your data analysis
   b) The questions you want to answer to gain deeper insights into the chosen datasets such that you are able to produce an interesting data analysis on it
   c) A list of URLs of all the datasets you have used (including the ones we gave)
   d) For each dataset, write Python code that uses the **Numpy** package to extract useful statistical or summary information about the data and **Matplotlib** package to produce useful data visualizations that explain the data. **Note: You cannot use other packages**.
   e) For each dataset, explain the **nature of that dataset** (i.e. what is in that dataset) or any pecularities about it you wish to highlight and explain the **process** you went through to analyse that dataset.
   f) For each dataset**, describe the insights** you have gained from analysing the data and any conclusions or **recommendations** you want to make as a result of the analysis

4. Your code should produce the following chart types:
   - At least one line chart
   - At least one histogram
   - At least one scatterplot
   - At least one bar chart
   - At least one pie chart

5. A sample output of the text-based analysis and data visualisation requirement are given in Section 4 of this document.

# Section 3
# Marking Scheme

Marks will be awarded to each student based on the following rubrics:

| Component | Weightage |
|---|---|
| **Assignment requirements are met**<br>• Find at least one more dataset from the link provided in Section 2<br>• Python codes that extract useful insights from the datasets using only the Numpy library (ie. Not to use other scientfic computing package)<br>• Python codes that produces useful data visualizations from the datasets using only the Matplotlib library<br>• Explain the datasets, what was done to process these datasets and summarizes the insights gained from the analysis of the data | 30% |
| **Quality of code**<br>• Technical complexity<br>• Code quality<br>• User-friendliness<br>• Aesthetics<br>• Usage of markdown cells for text | 15% |
| **Data analysis**<br>• Completeness in the analysis of data<br>• Depth of questions asked<br>• Quality of answers you provide | 30% |
| **Interview**<br>• Ability to re-produce codes<br>• Explanation of insights<br>• Preparation, confidence and flow of content | 25% |

# Section 4
# Sample outputs expected

This section contains sample screenshots of how your Python programs may look like.

Do note that they are simple examples only, and you are highly encouraged to enhance your own version with more complex features or functionalities than what is shown here.

## Example 1
## Simple Text-based Analysis using Numpy

This output uses the Numpy library to load a Transport CSV dataset (from data.gov.sg) with 'Certificate of Entitlement (COE) Bidding Results' and quickly breaks down the data with some simple and useful information.

It helps us to think about how we may want to extract subsets of this dataset and the choice of chart type for data visualization later.

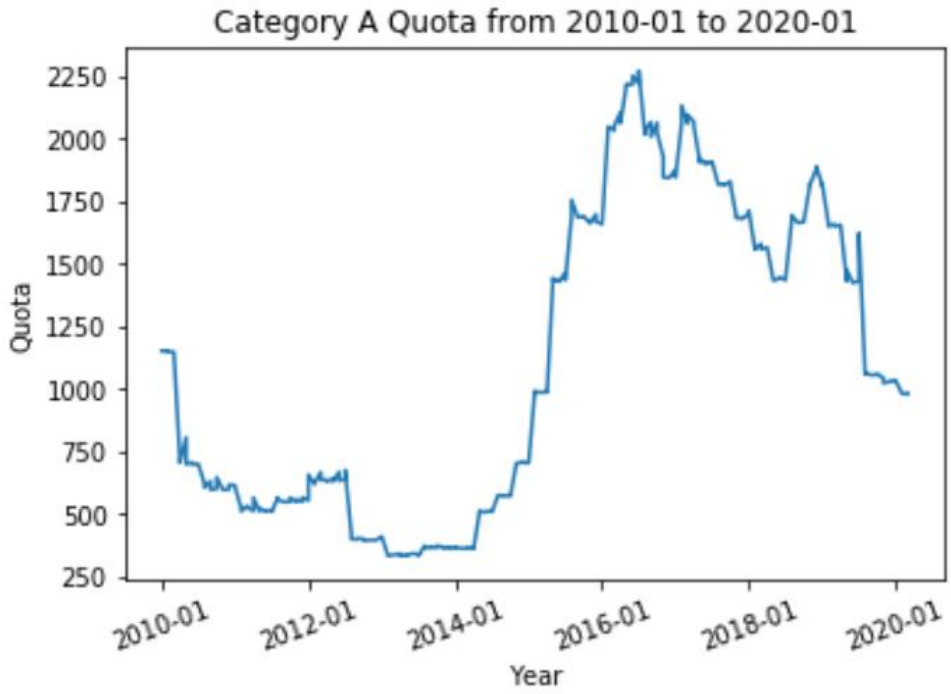
```
***** COE Results *****
There are 1230 rows in this dataset
There are 123 months of data from 2010-01 to 2020-03.
There are 5 vehicle classes

The lowest premium of Category A COE is 18502, on month 2010-01, bid 1.
The highest premium of Category A COE is 92100, on month 2013-01, bid 1.

The lowest premium of Category B COE is 19190, on month 2010-01, bid 1.
The highest premium of Category B COE is 96210, on month 2013-01, bid 1.

The lowest premium of Category C COE is 19001, on month 2010-01, bid 1.
The highest premium of Category C COE is 76310, on month 2013-10, bid 1.

The lowest premium of Category D COE is 852, on month 2010-01, bid 2.
The highest premium of Category D COE is 8451, on month 2018-02, bid 1.

The lowest premium of Category E COE is 19889, on month 2010-01, bid 1.
The highest premium of Category E COE is 97889, on month 2013-01, bid 2.
```

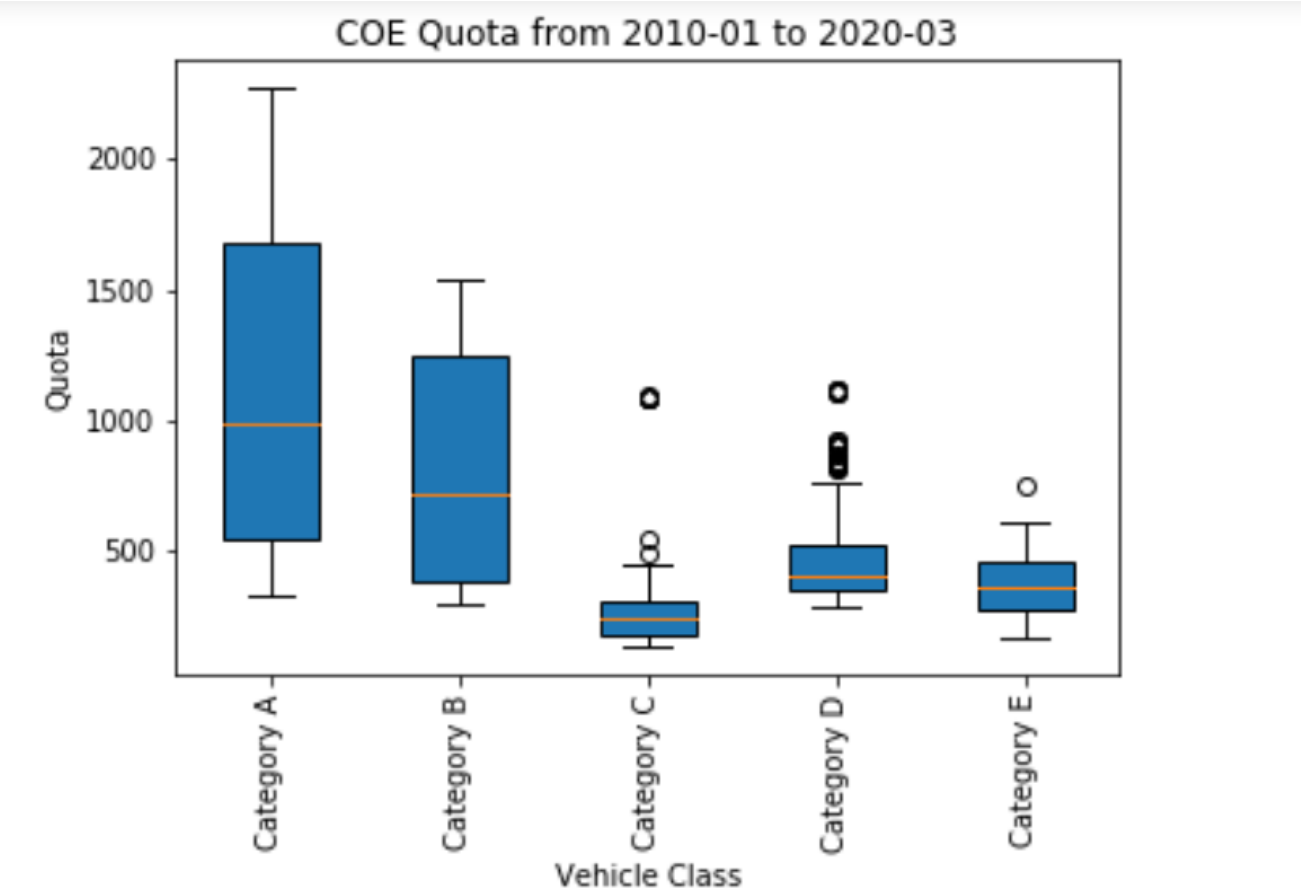
# Example 2
# Simple Data Visualization using Matplotlib

This sample output uses the Matplotlib library to plot a line chart and a boxplot to allow the user to perform a simple data analysis of the COE bidding exercise.

From the line chart, it shows the number of Category A quota over the years. It is quite low before 2014 and it rises quite steeply and reaches a high point around 2016.

From the boxplot, it shows that the range Category A and the Category B quota values are quite wide. It is interested to note that Category C and Category D has several outliers.






**-- End of Assignment Specifications --**