**SCHOOL OF COMPUTING**

**CA1 Specification**

**EP0302 Programming for Data Science**

**2021/2022 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 2](#_Toc52448502)

[Section 2 Scope of the assignment 3](#_Toc52448503)

[Basic Requirements 3](#_Toc52448504)

[Section 3 Marking Scheme 4](#_Toc52448505)

[Section 4 Sample outputs expected 5](#_Toc52448506)

[Example 1 Simple Text-based Analysis using Numpy 5](#_Toc52448507)

[Example 2 Simple Data Visualization using Matplotlib 6](#_Toc52448508)

# 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 **Friday 11 June 2021 (23:59).**
4. Submissions should be made via the **Blackboard CA1 Assignment Submission link** by the stated deadline
5. Deliverable should be a zip file with the following file-naming convention

**“YourCourse-YourModuleClass-YourStudentID-YourName.zip”**

1. Zip file should include the following items:

* One or more Jupyter notebooks that accomplishes the given tasks using the Python programming language. The notebooks will also document the data insights that you have gained through the Python code you have written
* A self-reflection report that briefly states the challenges you have faced and the take-aways you have gained from doing this assignment

1. As part of the assignment requirements, you will need to give a short presentation / interview using the Jupyter notebooks you have prepared. Your module tutor may ask you questions related to the Python code during this interview / presentation session.
2. This assignment will account for **40%** of the **module grade**.
3. No marks will be awarded, if the work is copied or you have allowed others to copy your work.
4. 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.

Exceptions to this policy will be given to students with valid LOA on medical or compassionate grounds. Students in such cases will need to inform the lecturer as soon as reasonably possible. Students are not to assume on their own that their deadline has been extended.

# 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 Society based on the requirements as stated below.

Basic Requirements

1. You are required to choose at least three datasets from the Health sector at the link, <https://data.gov.sg/search?groups=health>. You are encouraged to choose datasets which are interrelated and support a central theme of investigation.
2. Your Jupyter notebook(s) should include the following:
3. Your name and the title of your data analysis
4. 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
5. A list of URLs of all the datasets you have used
6. 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.
7. 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.
8. 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
9. Your code should produce the following chart types:

* At least one line chart - see trend over time or over continuous data
* At least one histogram - see distribution of univariate data
* At least one scatterplot – see distribution and correlation of the data
* At least one bar chart - compare between categorical data
* At least one box plot - distribution and outliers

For example, you can use one line chart and one histogram for dataset 1, then use one scatterplot for dataset 2, and use one bar chart and one box plot for dataset 3.

1. Analysing real-world data is not an easy task. Reflect on your **challenges** and your **achievements** in completing this assignment and document it using the given “Reflection for CA1” template.
2. 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   * Use of at least 3 different datasets from Society at data.gov.sg * Python codes that extract useful insights from the datasets using only the Numpy library (ie. Not to use other scientfic computing package) * Python codes that produces useful data visualizations from the datasets using the Matplotlib library * Explain the datasets, what was done to process these datasets and summarizes the insights gained from the analysis of the data | 40% |
| Quality of application   * Technical complexity * Code quality * User-friendliness * Aesthetics | 30% |
| Data analysis   * Completeness in the analysis of data * Quality of analysis and presentation | 20% |
| Reflection Report   * Explanation of challenges faced * Explanation of achievements made | 10% |

# 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.

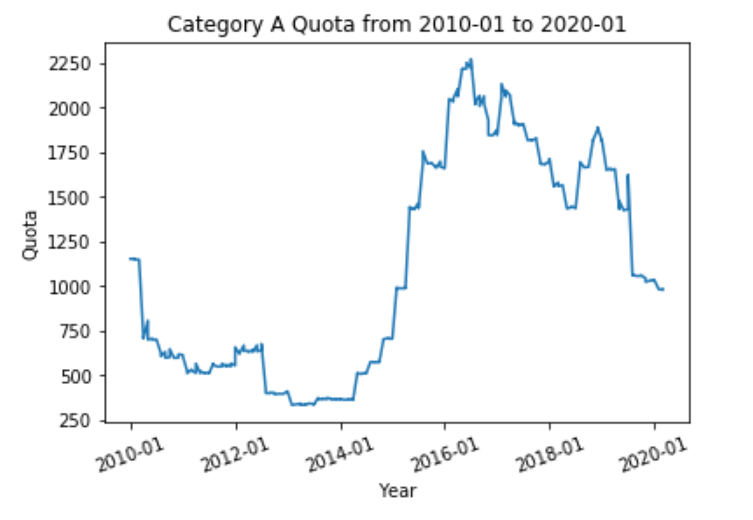
|  |
| --- |
|  |

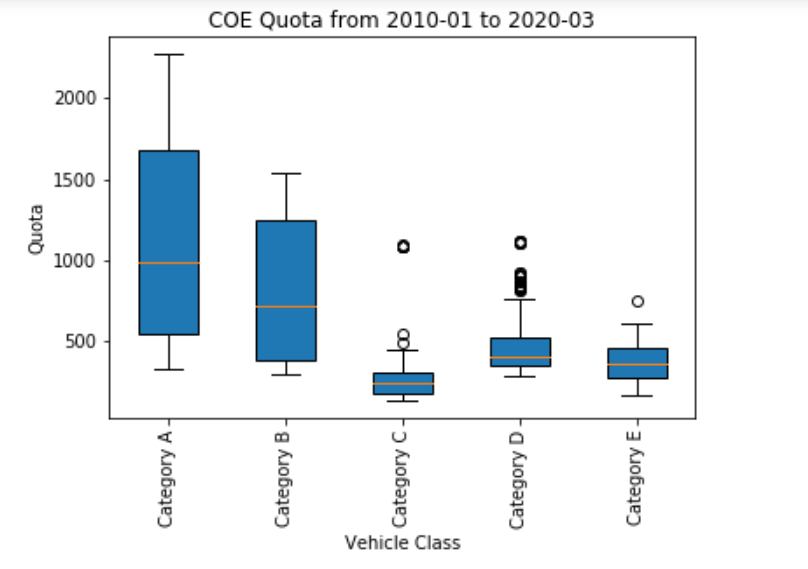
## Example 2 Simple Data Visualization using Matplotlib

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 --**