**SCHOOL OF COMPUTING**

**CA1 Specification**

**Programming for Data Science**

**2019/2020 Semester 2**

**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](#_Toc526873522)

[Section 2 Scope of the assignment 3](#_Toc526873523)

[Basic Requirements 3](#_Toc526873524)

[Section 3 Marking Scheme 5](#_Toc526873525)

[Section 4 Sample outputs expected 6](#_Toc526873526)

[Example 1 Simple Text-based Analysis using Numpy 6](#_Toc526873527)

[Example 2 Simple Data Visualization using Matplotlib 7](#_Toc526873528)

# 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 cleansing, 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 **Sunday 24 November 2019 (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

**“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
* A set of Powerpoint slides that summarizes 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 Powerpoint slides 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 various datasets belonging to the **Land Transport Authority** based on the requirements as stated below.

Basic Requirements

1. You must use **at least three** datasets from Land Transport Authority (LTA) at the link below.

<https://data.gov.sg/dataset?organization=land-transport-authority>.

1. For each dataset you use, you must write Python code that uses the **Numpy** package to extract useful statistical or summary information about the data.

A sample of the expected output of this requirement is given in Section 4 of this document.

1. For each dataset you use, you must write Python code that uses the **Matplotlib** package to produce useful data visualizations that explain the data.

Your code should produce the following chart types:

* At least one bar chart
* At least one line chart
* At least one histogram
* At least one scatterplot
* At least one boxplot

A sample of the expected output of this requirement is given in Section 4 of this document.

1. Your Python codes should help you to gain deeper insights into the chosen datasets such that you are able to produce an interesting data analysis on it.

Compile your findings into a deck of **Powerpoint slides**

Your Powerpoint slides should include the following sections:

* A cover page that lists your name and the title of your data analysis
* A slide that lists the URLs of all the datasets you have used
* For each dataset, one slide or more to briefly explain the **nature of that dataset** (i.e. what is in that dataset) or any pecularities about it you wish to highlight
* For each dataset, one slide or more to explain the **process** you went through to analyse that dataset. Where possible, you should specifically mention how you used the Numpy or Matplotlib functions to achieve a certain outcome e.g. to transform the data or to produce a certain visualization
* For each dataset, the **insights** you have gained from analysing the data and any conclusions or recommendations you want to make as a result of the analysis

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.

# 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 Land Transport Authority 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 only the Matplotlib library * A deck of Powerpoint slides that 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

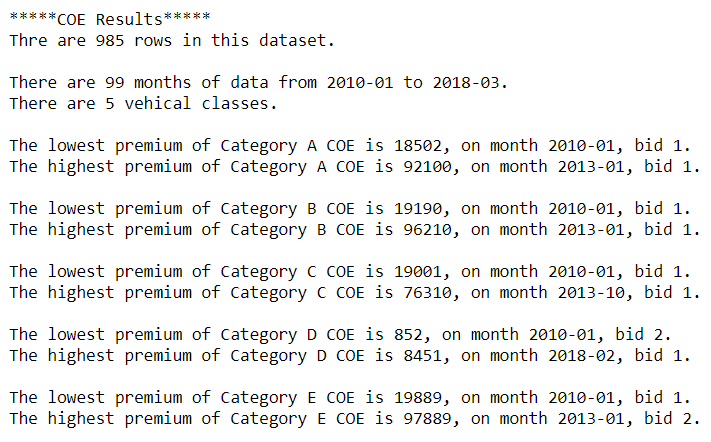
This section contains sample screenshots of how your Python programs may look like.

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 LTA CSV dataset 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.

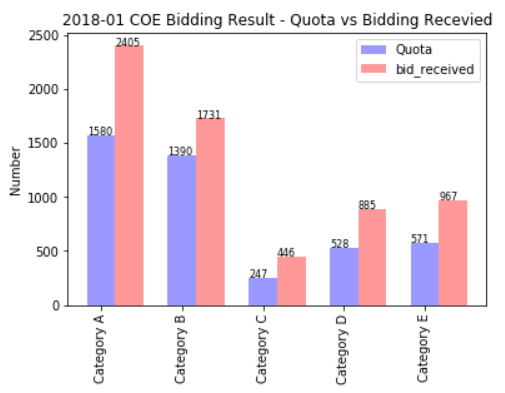


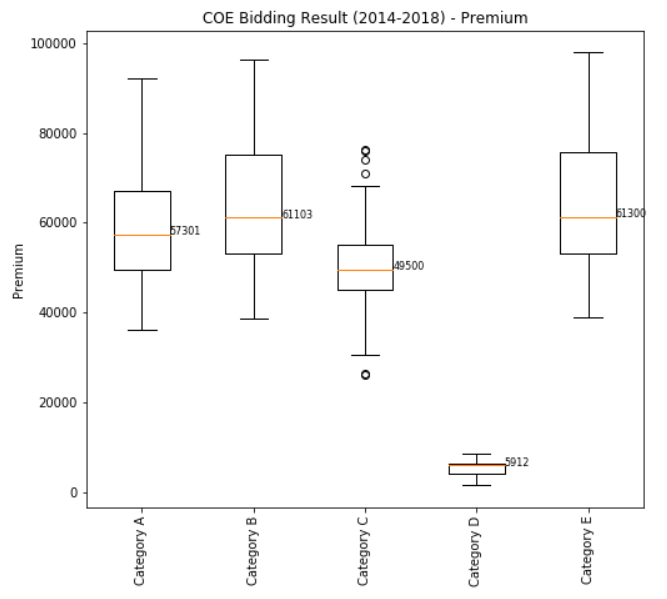
## Example 2 Simple Data Visualization using Matplotlib

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

From the bar chart, it shows that Category A receives the most number of bids compared to other categories. For all categories, more bids were received than the allocated quota. It is a case of demand is greater than than supply.

From the boxplot, it shows that the Category B and the Cartegroy E COE have similar values. It is interested to note that Category C (GOODS VEHICLE & BUS) has several outliers.





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