**SCHOOL OF COMPUTING (SOC)**

**CA1 Specification**

**SPECIALIST DIPLOMA IN DATA SCIENCE**

**IT8701  
Introduction to Programming for Data Science**

**2022/2023 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](#_Toc526247177)

[Section 2 Scope of the assignment 3](#_Toc526247178)

[Basic Requirements 3](#_Toc526247179)

[Section 3 Marking Scheme 5](#_Toc526247180)

[Section 4 Sample outputs expected 6](#_Toc526247181)

[Example 1 Simple Text-based Analysis using Numpy 6](#_Toc526247182)

[Example 2 Simple Data Visualization using Matplotlib 7](#_Toc526247183)

# 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 **Monday 02 Jan 2023 (11:59 pm)**.
4. Submissions should be made via the **POLITEMall CA1 Assignment Submission link** by the stated deadline
5. Deliverable should be a zip file with the following file-naming convention

**“YourModuleLecturerName-YourStudentID-YourName.zip”**

**e.g. “Chanon-2128883-MoffGideon.zip”**

1. Zip file should include the following items:

* One or more Jupyter notebooks (.ipynb) or Python source code files (.py) 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 document 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 (not more than 10 minutes) presentation / demonstration to your module tutor using the Powerpoint slides you have prepared. Your module tutor may ask you questions related to the Python code during this session.
2. Subsequent to the submission of your codes and slides, your Module Lecturer will arrange assignment interviews with you separately. Please take note that the dates of the interviews you arrange with your lecturer do not affect our records of the date that you submitted your assignment.
3. This assignment will account for **40%** of the **module grade**.
4. No marks will be awarded, if the work is copied or you have allowed others to copy your work.
5. There will be a penalty of 5 marks for each day that you submitted the assignment later than the specified deadline. You may submit late assignments to your lecturer via email.
6. There will be no exceptions to this policy unless you are granted a 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.

Note that we will not accept appeals to waive off late submissions due to your work commitments or travel. It does not reflect well on you if you are not able to plan your time well and manage your workload to finish the assignments on-time.

If you have further questions about this, please address them to your Module Tutor.

# Section 2 Scope of the assignment

In this individual assignment, you are required to produce a data analysis presentation for datasets belonging to the **Housing Development Board (HDB)** based on the requirements as stated below.

Basic Requirements

1. You must use **at least three** datasets from the Housing Development Board at the link below

<https://data.gov.sg/dataset?organization=housing-and-development-board>

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. You are not allowed to use Python packages like pandas to extract and transform the data. This is to train you to know the Numpy package well.

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. You are not allowed to use other visualization packages like seaborn to plot your graphs. This is to train you to know the Matplotlib library well.

Your code should produce at least 4 of the following chart types:

* At least one bar chart
* At least one line chart
* At least one pie chart
* At least one histogram
* At least one scatterplot
* At least one boxplot
* Any other type of chart that can be produced by matplotlib

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

To clarify a point that many students misunderstand, note that we expect a total of at least 4 graphs in all, **not** at least 12 graphs in total.

For example, you could

1. use Dataset 1 to plot a barchart and the line chart
2. use Dataset 2 to plot a histogram
3. use Dataset 3 to plot a scatterplot or a boxplot

In the above example, you would have satisfied the requirement to use 3 datasets as well as to include at least 4 compulsory graphs

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 HDB at data.gov.sg * Python codes that extract useful insights from the datasets using the Numpy library * Presence of the 4 compulsory chart types * Python codes that produces useful data visualizations from the datasets using 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 * A self-reflection document outlining your challenges and achievements doing this assignment | 50% |
| Quality of application   * Technical complexity * Code quality * User-friendliness (of the graphs) * Aesthetics * Creativity | 30% |
| Data analysis   * Completeness in the analysis of data * Quality of analysis and presentation | 15% |
| Self-reflection   * Explaination of challenges faced * Explaination of achievements made * Reflection of how you have met the assignment requirements | 5% |

# Section 4 Sample outputs expected

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

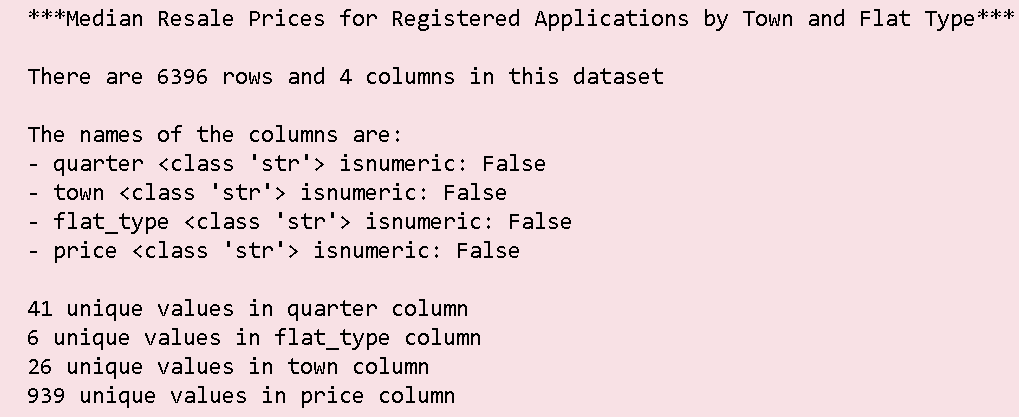
Do note that they are simple examples only, and you are highly encouraged to enhance your own version with more complex features or functionalities than what is shown here.

## Example 1 Simple Text-based Analysis using Numpy

This output uses the Numpy library to load a HDB CSV dataset with the median resale prices by town and flat type and quickly breaks down the data with some simple useful-to-know information.

With this quick breakdown, we quickly realise the price column may have n/a values since the isnumeric is False for this column.

It also 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

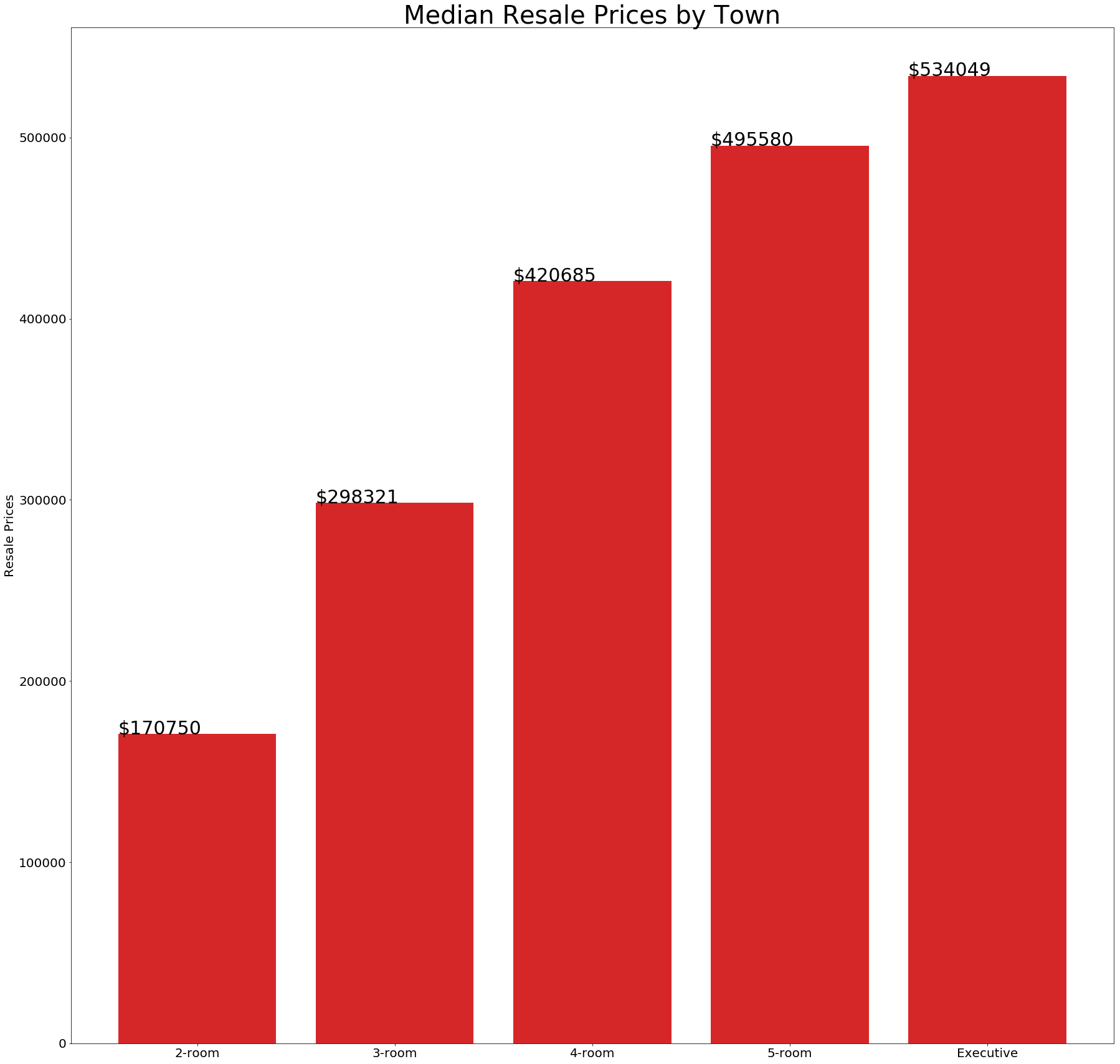
This sample output uses the Matplotlib library to plot a bar chart and a box plot to allow the user to perform a simple data analysis of the prices of resales flats across flat-types.

For example, from the boxplot, you can clearly see the median prices of each flat-type as well as the extreme outliers that were sold at a price level much higher than the median.

The barchart is computed by averaging the prices of flats sold by flat-type and gives you a good comparison of how the average price may differ from the median price of each flat-type.

A close up of text on a white background

Description generated with very high confidence



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