SCHOOL OF COMPUTING (SOC)

CA1 Specification

DIPLOMA IN APPLIED AI & ANALYTICS

ST1510
Programming for Data Analytics

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

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 Week 9 **Monday 12 Dec 2022 (8am)**.

4. Submissions should be made via the **BrightSpace CA1 Assignment Submission link** by thestated deadline

5. Deliverable should be a zip file with the following file-naming convention
   "YourClass-YourStudentID-YourName.zip"
   e.g. "DAAA1B04-1928883-StevenLee.zip"

6. 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

7. As part of the assignment requirements, you will need to give a short (not more than 10 minutes) presentation / interview 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 interview / presentation session.

8. 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.

9. This assignment will account for **30%** of the **module grade**.

10. No marks will be awarded, if the work is copied or you have allowed others to copy yourwork.

Warning: Plagiarism means passing off as one's own the ideas, works, writings, etc., which belong to another person. In accordance with this definition, you are committing plagiarism if you copy the work of another person and turning it in as your own, even if you would have the permission of that person.
Plagiarism is a serious offence, and if you are found to have committed, aided, and/or abetted the offence of plagiarism, disciplinary action will be taken against you. If you are guilty of plagiarism, you may fail all modules in the semester, or even be liable for expulsion.

Section 2
Scope of the assignment

In this individual assignment, you are required to produce a data analysis presentation for datasets relating to **Education** based on the requirements as stated below.

Basic Requirements

1. You must use **at least three** datasets, including at least one from Data.gov.sg
   https://data.gov.sg/search?groups=education

2. 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.

3. For each dataset you use, you must perform exploratory data analysis to understand the data, and 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 seaborne to plot your graphs. This is to train you to know the Matplotlib library well.

   Your code should produce at least 5 charts from the following 6 chart types:

   - bar chart
   - line chart
   - pie chart
   - histogram
   - scatterplot
   - boxplot

   A sample of the expected output of this requirement is given in Section 4 of this document. Note that different datasets (HDB) are used in the sample.

   To clarify a point that many students misunderstand, note that we expect a total of at least 5 graphs in all, **not** at least 15 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 and a boxplot

In the above example, you would have satisfied the requirement to use 3 datasets as well as to include at least 5 compulsory graphs

4. 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

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, with at least one from data.gov.sg relating to Education • Python codes that extract useful insights from the datasets using the Numpy library • Presence of the 5 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	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	20%

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.

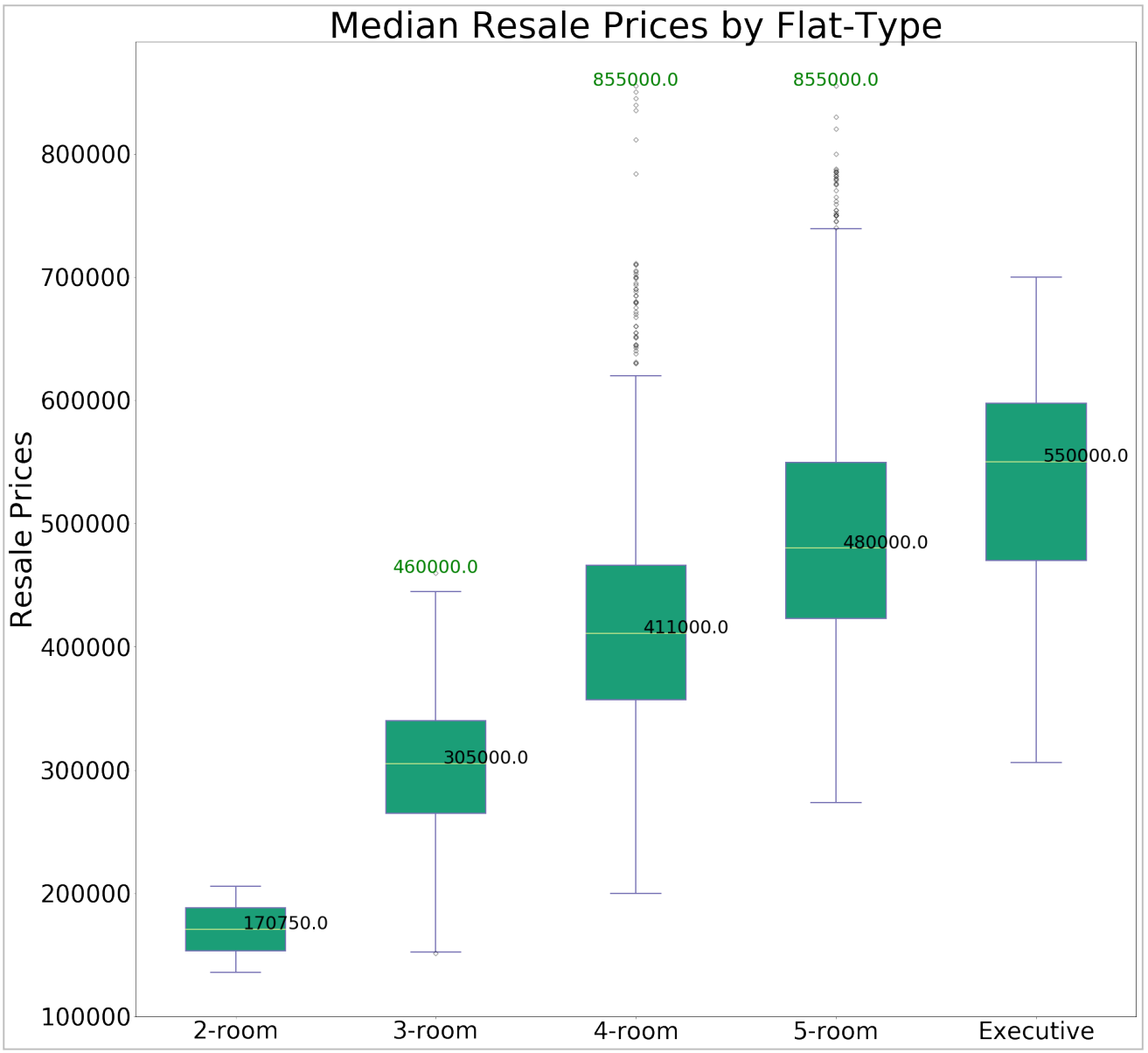
```
***Median Resale Prices for Registered Applications by Town and Flat Type***

There are 6396 rows and 4 columns in this dataset

The names of the columns are:
- quarter <class 'str'> isnumeric: False
- town <class 'str'> isnumeric: False
- flat_type <class 'str'> isnumeric: False
- price <class 'str'> isnumeric: False

41 unique values in quarter column
6 unique values in flat_type column
26 unique values in town column
939 unique values in price column
```

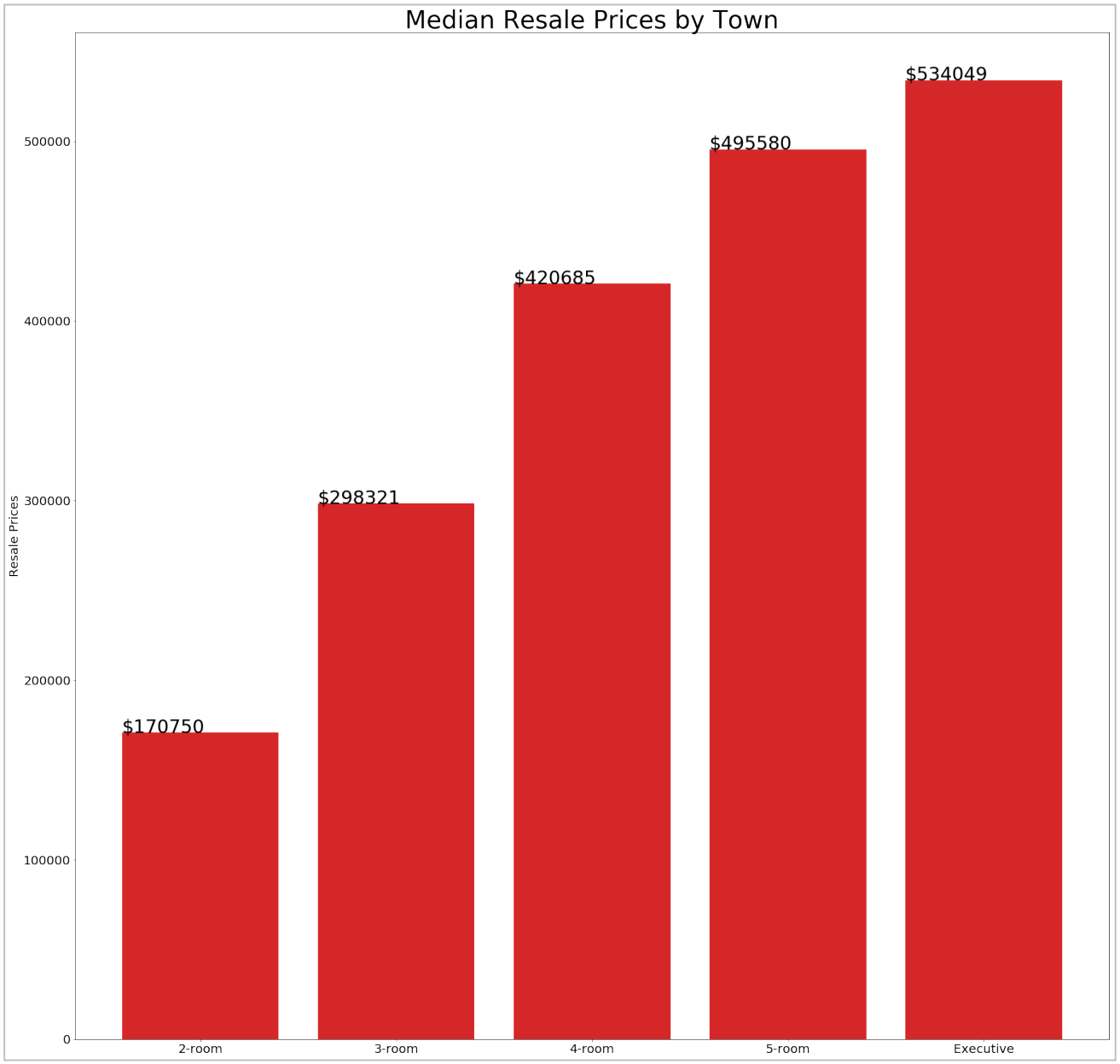
Example 2
Simple Data Visualization using Matplotlib

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.





-- End of Assignment Specifications --