SCHOOL OF COMPUTING (SOC)

CA1 Specification

DIPLOMA IN APPLIED AI & ANALYTICS

ST1510
Programming for Data Analytics

2023/2024 Semester 2

Assignment rubrics

1. Demonstrate basic competency in writing Python programs.
2. Demonstrate basic competency in using the Pandas 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.

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, **14 Dec (Thus) 2023 on 2359Hr**.

4. Submissions must 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.
   "YourClass-YourStudentID-YourName.zip"
   e.g. "DAAA1B04-2388888-StevenLee.zip"

6. The Zip file should include the following items:
   - One or more Jupyter Notebook (.ipynb) that accomplishes the given tasks using the Python
   - A set of PowerPoint slides that summarizes the data insights that you have gained through the Python code you have written
   - All datasets (.csv files) used
   - One Declaration of Academic Integrity

7. A compulsory presentation/interview will be conducted. During the session, you must present your work using the submitted PowerPoint slides and the Jupyter Notebook. Your module tutor will ask questions related to the submission and ask you to reproduce certain parts of your code during the session.

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

9. 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 module tutor as soon as reasonably possible. Students are not to assume on their own that their deadline has been extended.

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

    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.

Section 2
Assignment Requirements

1. You must use **at least three** datasets, including **at least two** datasets from data.gov.sg published by either the Ministry of Health (MOH) or the National Environment Agency (NEA).

   Example of datasets from data.gov.sg :
   - Weekly Infectious Disease Bulletin, https://beta.data.gov.sg/collections/508/view
   - Top 10 Conditions of Hospitalisation, https://beta.data.gov.sg/collections/505/view

   Do NOT use dataset from training website such as Kaggle.

2. You are encouraged to select interrelated datasets that align with a central theme of investigation. Clearly define an analysis question or problem statement and provide the answer through your data analysis. **Document the URLs of all datasets** in both the Jupyter Notebook and PowerPoint slides.

3. For each dataset, your task is to conduct exploratory data analysis using Python in a Jupyter Notebook. The Jupyter Notebook should use Pandas for data wrangling and analysis, and Matplotlib for creating informative data visualizations. You should not use additional visualization packages, such as Seaborn, for plotting graphs. The objective is to train you to know Pandas and Matplotlib packages well.

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

4. Your code should generate at least five charts, each representing a different chart type chosen from the following six options. The expectation is to produce a total of 5 charts, each of a distinct chart type.

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

5. While you can submit more than five charts, keep it within a maximum of nine, focusing on the most informative and insightful ones.

6. Your Python codes should extract meaningful insights from the chosen datasets, crafting an engaging and interesting data analysis.

7. Compile your findings into a deck of PowerPoint slides. The PowerPoint slides should include the following sections and to be presented in 6 minutes.

   - 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 peculiarities 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 Pandas 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 of the analysis.

Section 3
Marking Scheme

Marks will be awarded based on the following rubrics:

Component	Weightage
Basic Requirements • Use of at least three different datasets that met the requirement outlined in Section 2 • At least five informative charts, each representing a different chart type	35%
Quality of application • Code quality ▪ Demonstration of mastery of the topics covered ▪ Reusability ▪ Efficiency • Code clarity and documentation **NO mark** can be awarded if Pandas and Matplotlib are not used extensively.	30%
Data analysis and Presentation • Quality of the analysis • Organization and Quality of the presentation and the slides • Question and Answer • Ability to re-produce the submitted code during the presentation/interview **No mark** can be awarded if the PowerPoint slide is not submitted.	35%

Section 4
Sample outputs expected

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

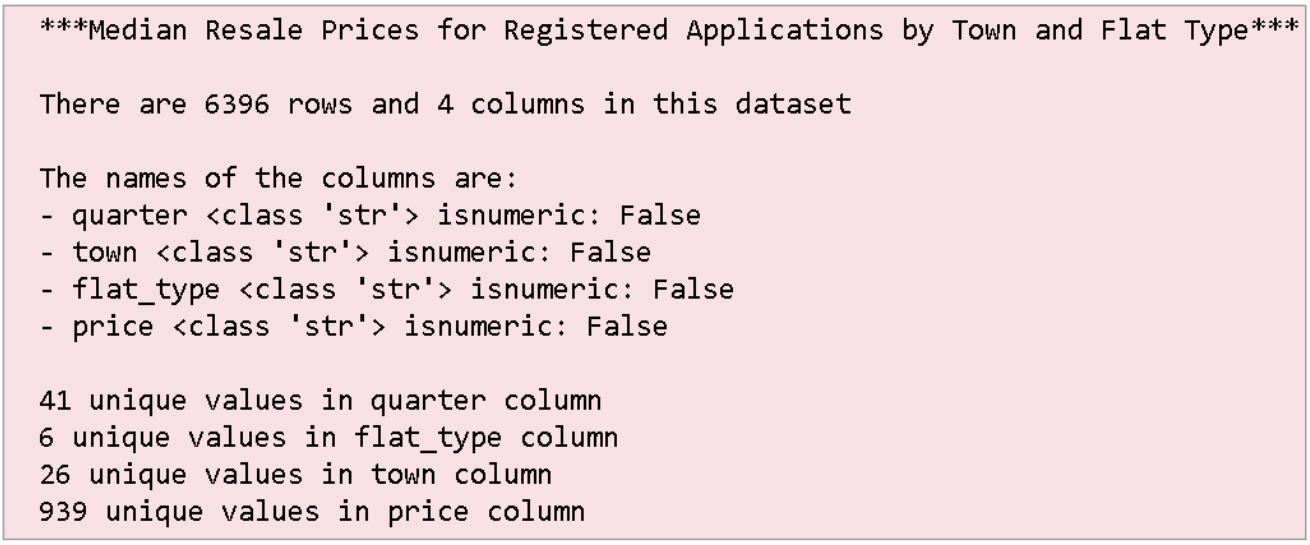
Do note that they are simple examples only, and you are highly encouraged to enhance your own version with more complex features or functionalities than what is shown here.

Example 1
Simple Text-based Analysis

This output uses the Pandas 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.



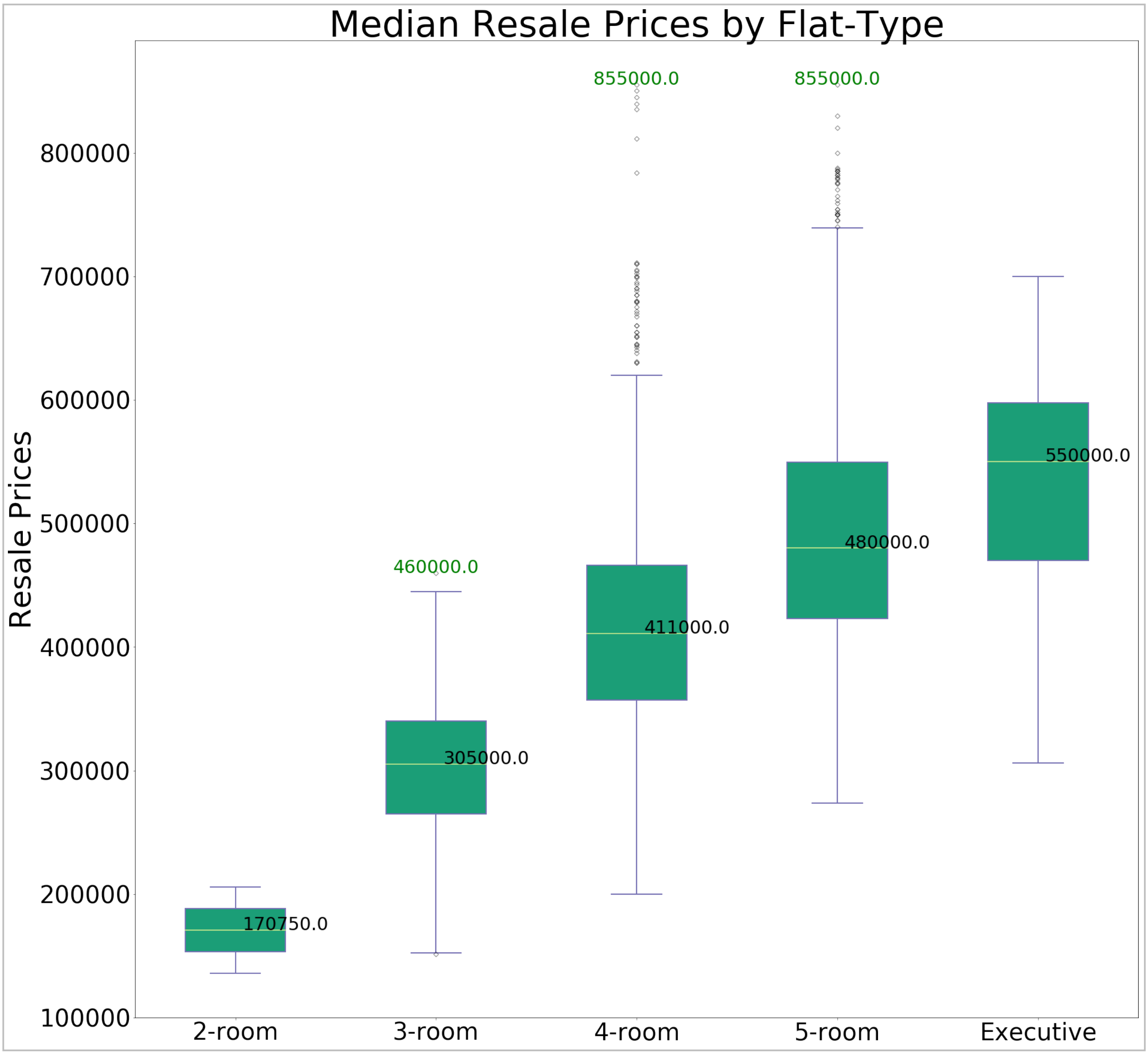
```
***Median Resale Prices for Registered Applications by Town and Flat Type***

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
```

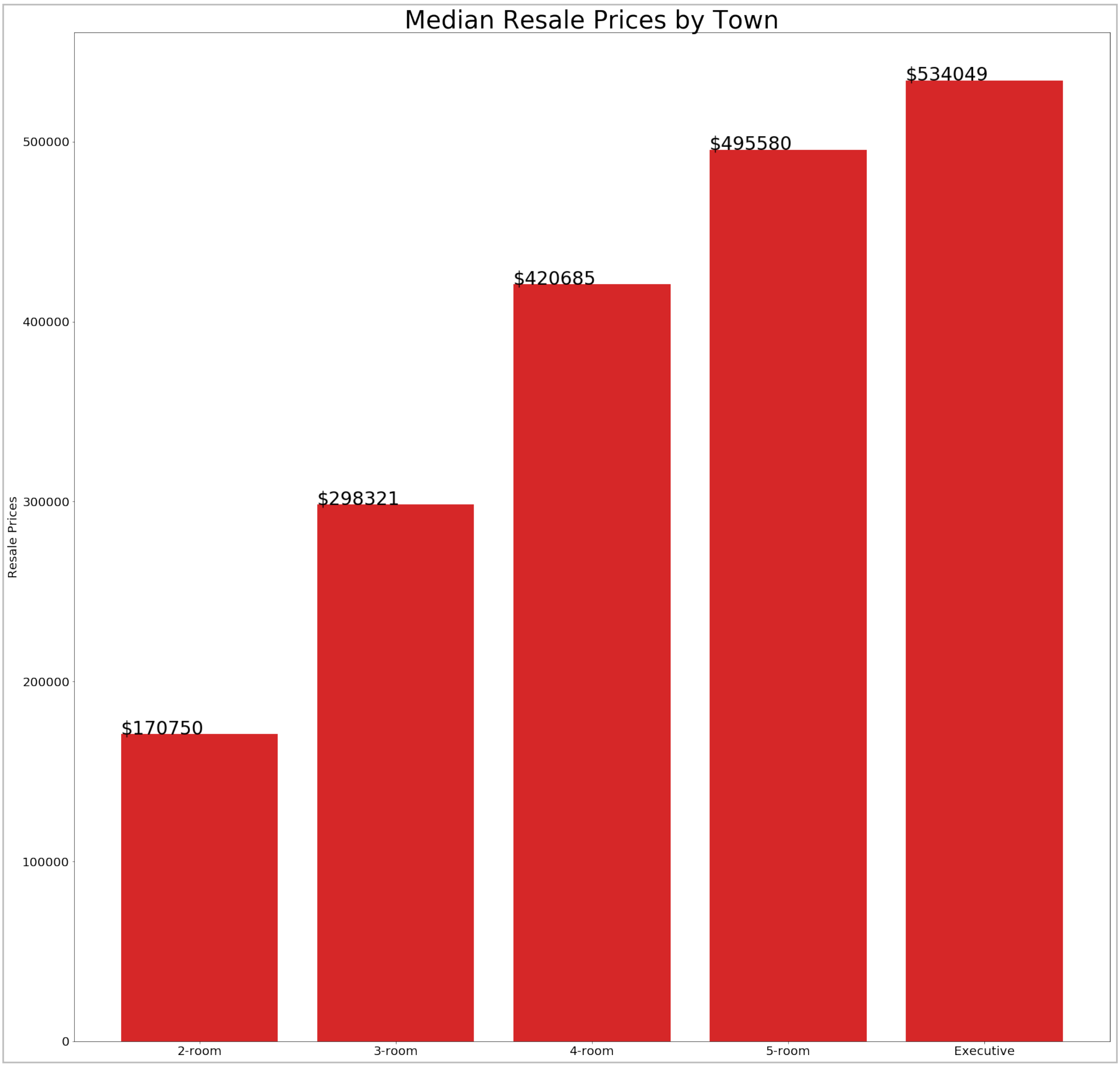
Example 2
Simple Data Visualization using Matplotlib

This sample output uses the Matplotlib library to plot a bar chart and a boxplot 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 bar chart is computed by averaging the prices of flats sold by flat-type and gives you a goodcomparison of how the average price may differ from the median price of each flat-type.





-- End of Assignment Specifications --