**SCHOOL OF COMPUTING**

**CA2 Specification**

**EP0302 Programming for Data Science**

**2021/2022 Semester 1**

**Assignment rubrics**

1. Demonstrate competency in using the Python Pandas package for data cleaning and analysis and Python visualization packages for data visualization
2. Demonstrate 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](#_Toc43918027)

[Section 2 Scope of the assignment 3](#_Toc43918028)

[Basic Requirements 3](#_Toc43918029)

[Section 3 Marking Scheme 5](#_Toc43918030)

[Section 4 Sample outputs expected 6](#_Toc43918031)

[Example 1 Simple Text-based Analysis using Pandas 7](#_Toc43918032)

[Example 2 Data Visualization using Matplotlib 8](#_Toc43918033)

[Example 3 Violin Plot Data Visualization using Seaborn 9](#_Toc43918034)

# Section 1 Instructions and Guidelines

1. This is an **INDIVIDUAL** assignment which requires the student to code a Python application that retrieves and combines data from multiple data sources to perform data cleaning, transformation, visualization and analysis on it.
2. The requirements of this assignment are outlined in Section 2 of this document.
3. The deadline of this assignment is on **6 August 2021, 23:59**.
4. Submissions should be made via the Blackboard CA2 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. The notebooks will also document 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 presentation using the Jupyter notebooks you have prepared. Your module tutor may ask you questions related to the Python code during this 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 write Python programs and produce a data analysis presentation for various datasets based on the requirements as stated below.

Basic Requirements

1. You must use at least **three** datasets from data.gov.sg. You can mix the three datasets from any sectors, **except from health sector**. You are also allowed to use additional datasets from other websites, e.g. World Bank Data (<http://databank.worldbank.org/data/home.aspx>)
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 craft a ‘storyline’ or produce an interesting data analysis on it
5. A list of URLs of all the datasets you have used
6. For the chosen datasets, explain the **nature of the dataset** (i.e. what is in the dataset) or any pecularities about it you wish to highlight.
7. Write Python code that uses the pandas package to extract useful statistical or summary information about the data and Python visualisationpackage to produce useful data visualizations that explain the data.
8. Highlight 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. For each dataset you use, you must write a Python codes that uses data visualization package(s) such as Matplotlib, Seaborn, etc to produce useful graphs / charts that explain the data.

Your submission should contain the following graphs / chart types:

* At least one bar chart
* At least one line chart
* At least one pie-chart
* At least one scatterplot
* At least one boxplot

A sample of a possible output of this requirement is given in Section 4 of this document.

You are highly encouraged to utilise other graph types that may aid in the understanding and analysis of your chosen datasets.

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 CA2” template.

# Section 3 Marking Scheme

Marks will be awarded to each student based on the following rubrics.

To score higher marks, you are encouraged to explore and experiment beyond the syllabus and demonstrate your independently-acquired skills via your deliverables / presentation.

|  |  |
| --- | --- |
| Component | Weightage |
| Assignment requirements are met   * Use of at least 3 different datasets at data.gov.sg * Python codes that extract useful insights from the datasets using the Pandas library on the datasets * Python codes that produce useful data visualizations from the datasets using an appropriate data visualization library with the chart types as specified earlier in this document * Explain the 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% |
| Self-reflection   * Explanation of challenges faced * Explanation of achievements made | 10% |

# Section 4 Sample outputs expected

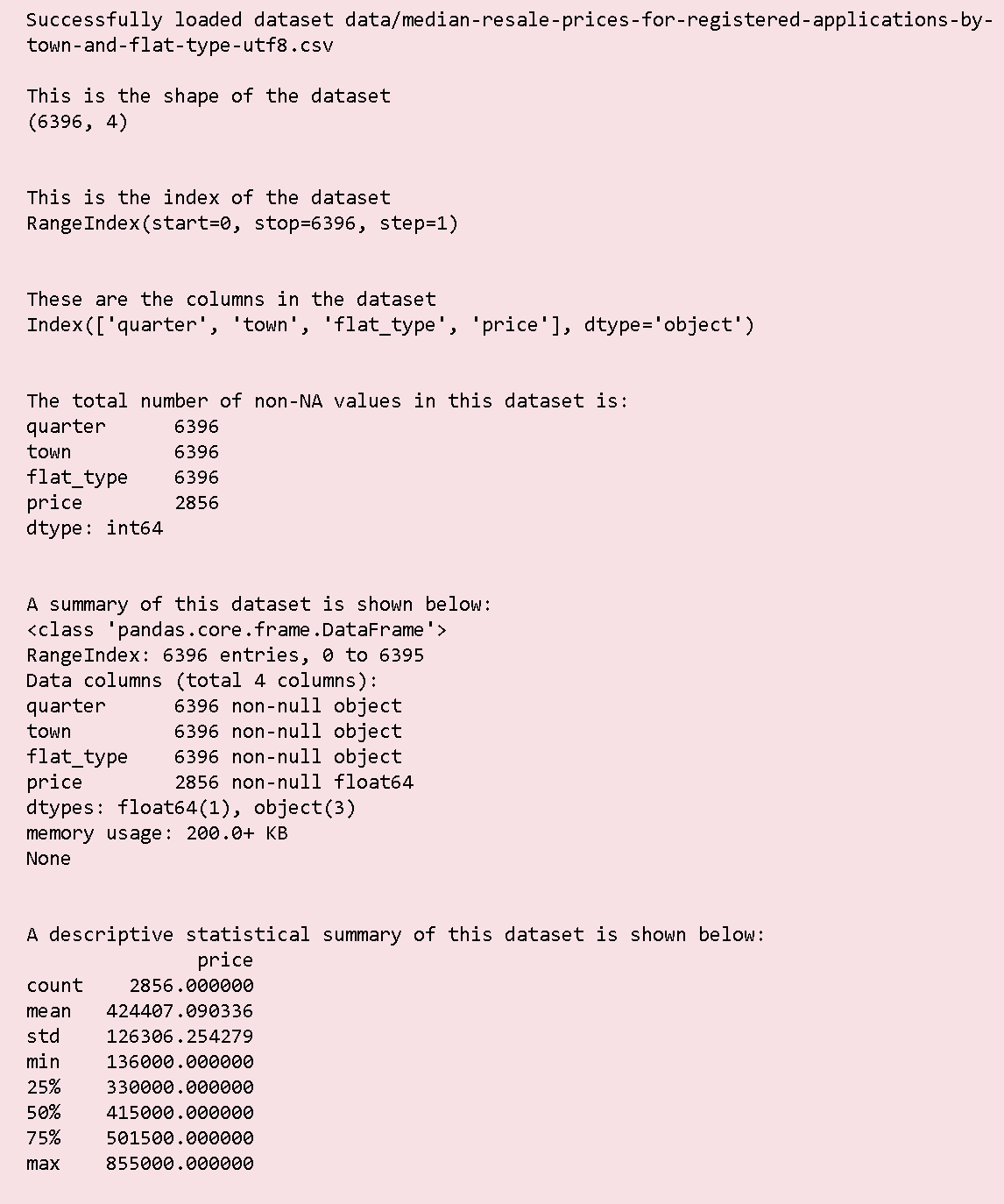
This section contains sample screenshots of how your Python programs may look like.

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.

To encourage you to explore beyond the syllabus, we have included samples of outputs from data visualization libraries not taught during the lessons.

* **Seaborn** – This library is a high-level library built on top of Matplotlib that allows you to create more attractive graphs much more easily

## Example 1 Simple Text-based Analysis using Pandas



## Example 2 Data Visualization using Matplotlib

This sample output uses the [Matplotlib](https://matplotlib.org/) library to plot a histogram of the median rents of different flat-types (data from data.gov.sg)

If you prefer not to dabble in other libraries which are shown in this document, feel free to go ahead and use Matplotlib instead.

A picture containing text, crossword puzzle

Description generated with very high confidence

## Example 3 Violin Plot Data Visualization using Seaborn

This sample output uses the [Seaborn](https://seaborn.pydata.org/) library to plot a static violin chart visualization showing the median resale prices for different flat types in 3 locations (data from data.gov.sg)

Seaborn is quite easy to use and does produce much more aesthetically-pleasing charts than Matplotlib, so go ahead and try it if you are adventurous!

A screenshot of a cell phone

Description generated with very high confidence

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