SCHOOL OF COMPUTING

CA2 Specification

EP0302 Programming for Data Science

2023/2024 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

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 **07 August 2023, 15:00**.

4. Submissions should be made via the **Brightspace CA2 Assignment Submission link** by the stated deadline.

5. Deliverable should be a zip file with the following file-naming convention:
   CA1-[ElectiveClass]-[AdminID]-[Name].zip

6. Zip file should include the following items:
   a. **One** Jupyter notebook that accomplishes the given tasks using the Python programming language. The notebook will also document the data insights that you have gained through the Python code you have written
   b. **One** HTML exported version of the Jupyter notebook
   c. **All** datasets (.csv files) used (including the recommended datasets)
   d. **One** Declaration of Academic Integrity (SOC)

7. As part of the assignment requirements, you will be having an interview using the Jupyter notebook you have prepared. Your module tutor may ask you to reproduce certain parts of your code during this interview session. Codes that are reproduced need not be exactly the same but the code should be able to perform the task in question. Usage of Google will be allowed during this questioning process. AI tools will not be allowed during the interview.

8. This assignment will account for **40%** of the **module grade**.

9. No marks will be awarded, if:

   a. The work is copied or you have allowed others to copy your work.
   b. If you are unable to reproduce most major parts of your code.
   c. Your zip file is corrupted. (please double check by downloading)

10. 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.

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 with at least 100 rows. The recommended website to use is data.gov.sg. You can mix the three datasets from any sectors, **except from Infrastructure 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 must include the following:

   a) Your name and the title of your data analysis

   b) 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

   c) A list of URLs of all the datasets you have used

   d) 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 and explain the **process** you went through to analyse that dataset.

   e) Write Python code that uses the Pandas package to extract useful statistical or summary information about the data and Python visualisation package to produce useful data visualizations that explain the data.

   f) 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

3. 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. You are allowed to use other packages for plotting graphs.

   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.

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 with more than 5000 rows • Use 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	30%
Quality of code • Technical complexity • Code quality • User-friendliness • Aesthetics • Usage for markdown cells for text	15%
Data analysis • Completeness in the analysis of data • Depth of questions asked • Quality of answers you provide	30%
Interview • Ability to re-produce codes • Explanation of insights • Preparation, confidence and flow of content	25%

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

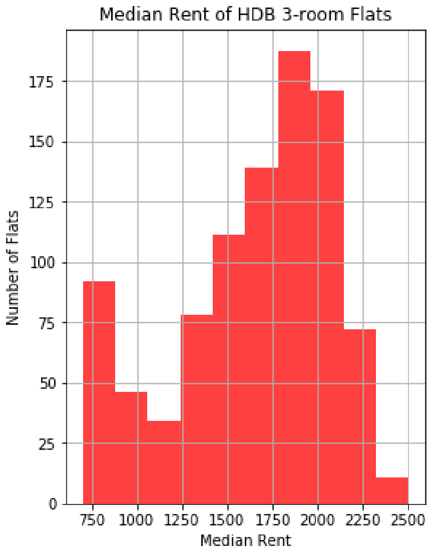
Example 1
Simple Text-based Analysis using Pandas

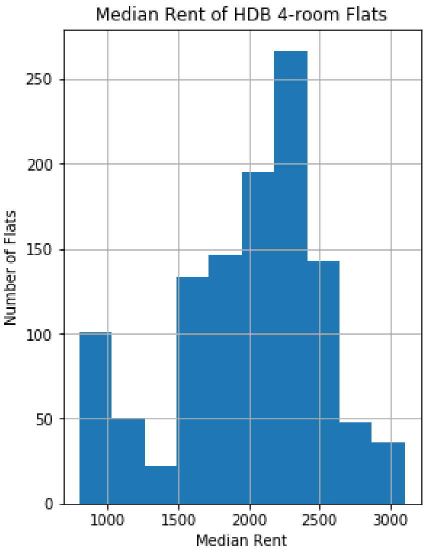
```
Successfully loaded dataset data/median-resale-prices-for-registered-applications-by-
town-and-flat-type-utf8.csv

This is the shape of the dataset
(6396, 4)


This is the index of the dataset
RangeIndex(start=0, stop=6396, step=1)


These are the columns in the dataset
Index(['quarter', 'town', 'flat_type', 'price'], dtype='object')


The total number of non-NA values in this dataset is:
quarter      6396
town         6396
flat_type    6396
price        2856
dtype: int64


A summary of this dataset is shown below:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6396 entries, 0 to 6395
Data columns (total 4 columns):
quarter      6396 non-null object
town         6396 non-null object
flat_type    6396 non-null object
price        2856 non-null float64
dtypes: float64(1), object(3)
memory usage: 200.0+ KB
None


A descriptive statistical summary of this dataset is shown below:
               price
count    2856.000000
mean   424407.090336
std    126306.254279
min    136000.000000
25%    330000.000000
50%    415000.000000
75%    501500.000000
max    855000.000000
```

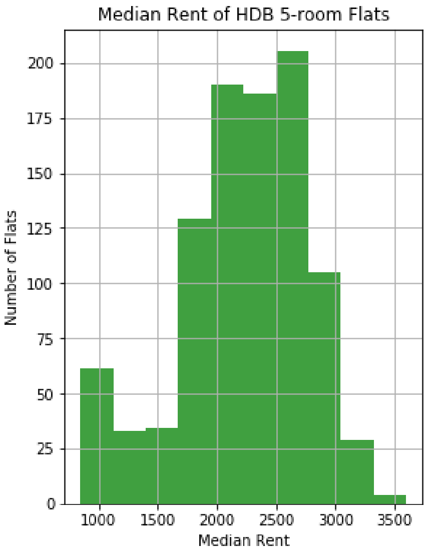
Example 2
Data Visualization using Matplotlib

This sample output uses the Matplotlib 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.



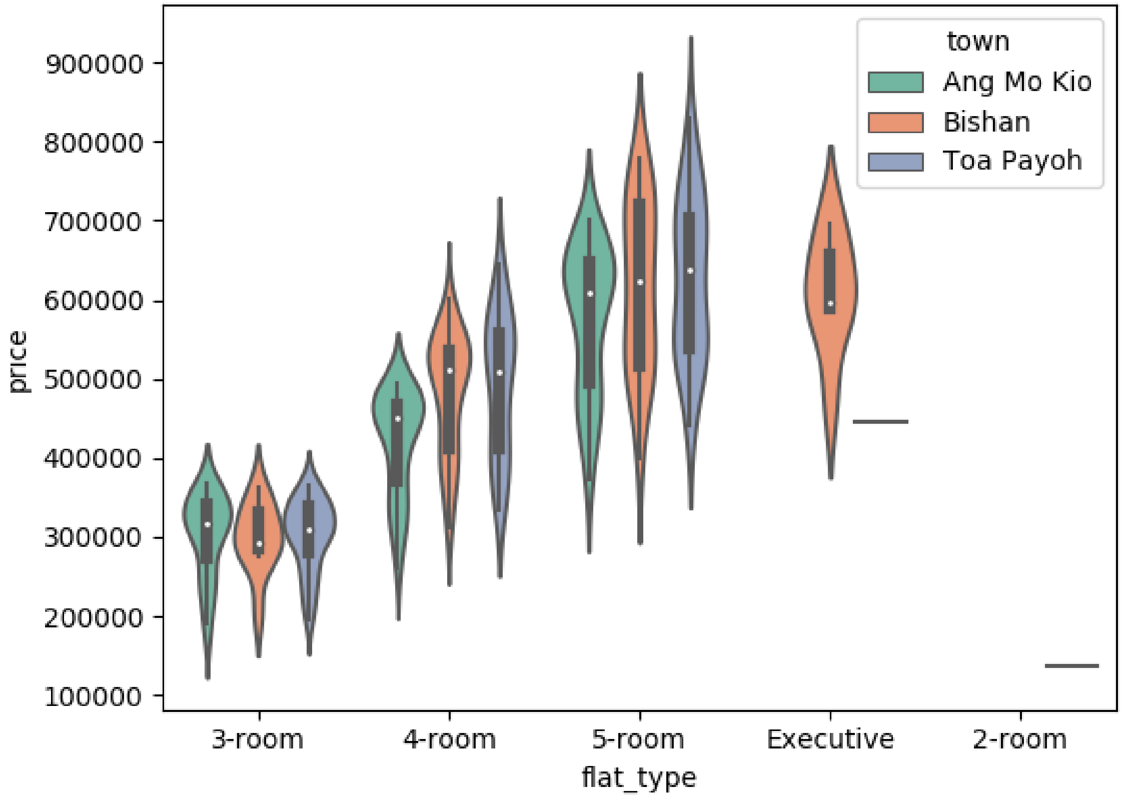




Example 3
Violin Plot Data Visualization using Seaborn

This sample output uses the Seaborn 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!



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