SCHOOL OF COMPUTING

CA2 Specification

Programming for Data Science

2019/2020 Semester 2

Assignment rubrics

1. Demonstrate competency in using the Python Numpy, Pandas and Matplotlib packages for data analysis and 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 cleansing, 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 **16 Feb 2020, 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"**

6. Zip file should include the following items:
   - One or more Jupyter notebooks that accomplishes the given tasks using the Python programming language.
   - HTML version of each of the Jupyter notebooks used.
   - 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.

7. As part of the assignment requirements, you will need to give a short presentation / interview 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. This assignment will account for **40%** of the **module grade**.

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

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. 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. The topic is unrestricted, i.e. you can mix the three datasets from any topic. You are also permitted to use additional datasets from other websites, e.g. World Bank Data (http://databank.worldbank.org/data/home.aspx)

2. You must use the **Pandas** package on **at least one** of the datasets

   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 write a Python program that uses a data visualization package 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 histogram / 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.

4. Your Python programs should help you 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.

   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 if data manipulation/cleaning has been carried as well as the extend of it being done
   - 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 Python libraries, Pandas or data visualization library 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

5. 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 / interview.

Component	Weightage
Assignment requirements are met • Use of at least 3 different datasets at data.gov.sg (as approved by your tutor) • Python codes that manipulate/clean and extract useful insights from the datasets using the **Pandas** library on **at least one** of the datasets • Python codes that produce useful **data visualizations** from the datasets using an appropriate data visualization library such as Matplotlib, Seaborn, Bokeh, Pygal etc with the chart types as specified earlier in this document • A deck of Powerpoint slides that 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 • Creativity	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

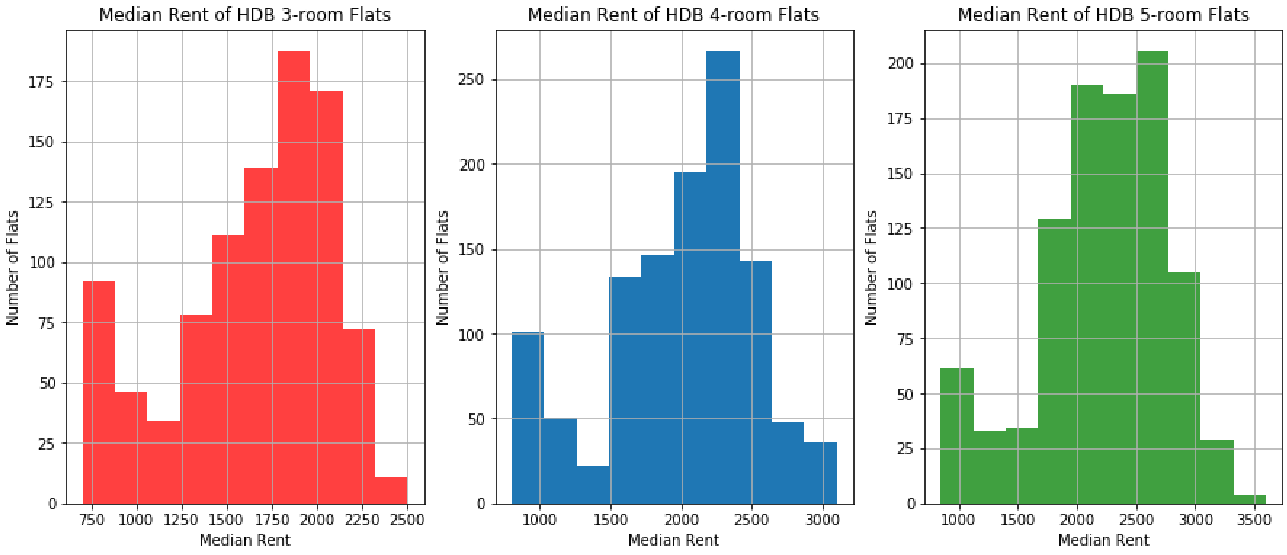
Example 1
Simple Text-based Analysis using Pandas

```
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)

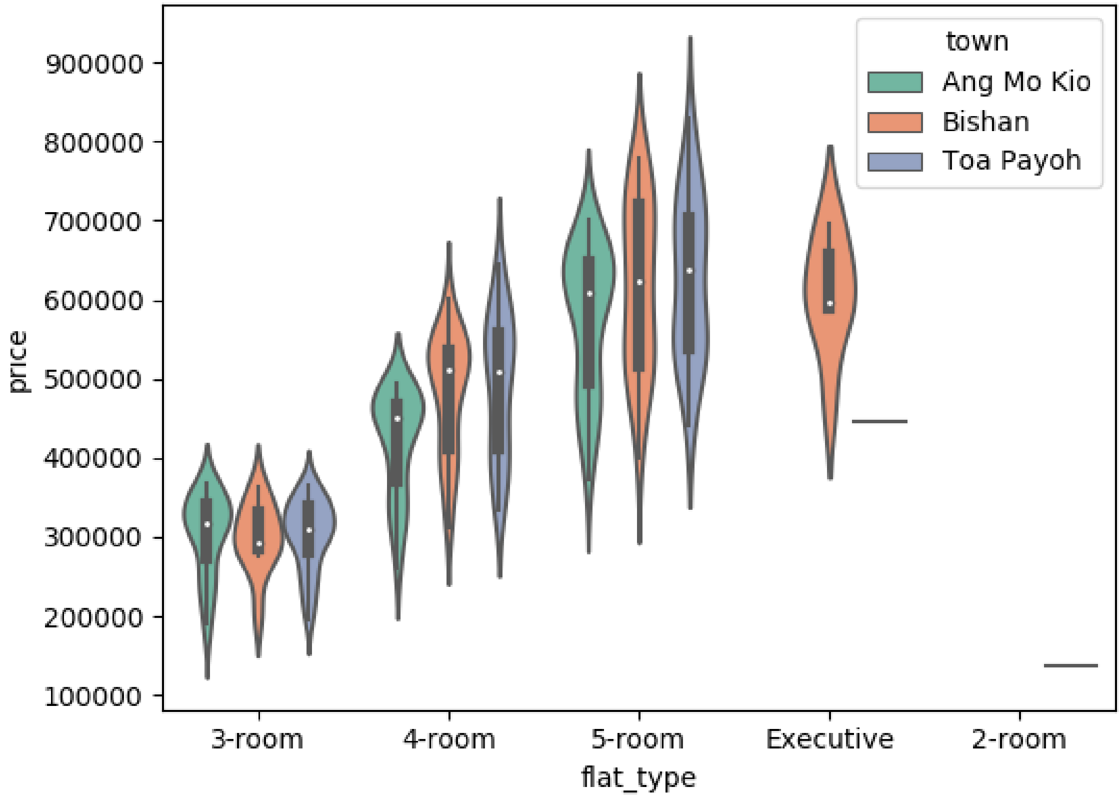
By now, you should be really an expert at Matplotlib! 😊 If you prefer not to dabble in other libraries which are shown in this document, feel free to go ahead and use Matplotlib instead.



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** --