**SCHOOL OF COMPUTING (SOC)**

**CA2 Specification**

**SPECIALIST DIPLOMA IN DATA SCIENCE**

**IT8701  
Introduction to Programming for Data Science**

**2022/2023 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 2](#_Toc503286506)

[Section 2 Scope of the assignment 3](#_Toc503286507)

[Basic Requirements 3](#_Toc503286508)

[Section 3 Marking Scheme 5](#_Toc503286509)

[Section 4 Sample outputs expected 6](#_Toc503286510)

[Example 1 Simple Text-based Analysis using Pandas 7](#_Toc503286511)

[Example 2 Viewing data inserted in mySQL 8](#_Toc503286512)

[Example 3 Data Visualization using Matplotlib 9](#_Toc503286513)

[Example 4 Violin Plot Data Visualization using Seaborn 10](#_Toc503286514)

[Example 5 Interactive Data Visualization using Bokeh 11](#_Toc503286515)

[Example 6 Interactive Data Visualization using Pygal 13](#_Toc503286516)

# 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, including relational databases and NoSQL datastores and 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 **13 Feb 2023 (Mon) 11:59 PM**.
4. Submissions should be made via the Polymall CA2 Assignment Submission link by the stated deadline
5. Deliverable should be a zip file with the following file-naming convention

**“YourModuleTutorsName-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
* 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

1. As part of the assignment requirements, you will need to give a short 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.
2. 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.
3. This assignment will account for **40%** of the **module grade**.
4. No marks will be awarded, if the work is copied or you have allowed others to copy your work.
5. There will be a penalty of 5 marks for each day that you submitted the assignment later than the specified deadline.
6. There will be no exceptions to this policy unless you are granted a 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.

Note that we will not accept appeals to waive off late submissions due to your work commitments or travel. It does not reflect well on you if you are not able to plan your time well and manage your workload to finish the assignments on-time.

If you have further questions about this, please address them to your Module Tutor.

# 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 **all** datasets

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

1. For each dataset you use, you must write a Python program that uses a data visualization package such as Matplotlib, Seaborn, Bokeh, Pygal etc to produce useful graphs / charts that explain the data.

Your submission should contain at least **four** the following graphs / chart types:

* A bar chart
* A pie chart
* A line chart
* A histogram
* A scatterplot
* A 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.

Though it is not compulsory, bonus marks would be given to students who are able to:

* Store and retrieve data in a relational database such as mySQL or a NoSQL database like MongoDB on at least one of the datasets.
* Produce *interactive* visualizations that allow the users to view the data in different ways.

1. 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 the **process** you went through to analyse that dataset. Where possible, you should specifically mention how you used the relational database or NoSQL 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

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 / interview.

|  |  |
| --- | --- |
| Component | Weightage |
| Assignment requirements are met   * Use of at least 3 different datasets * Python codes that extract useful insights from the datasets using the Pandas library on all 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 * A self-reflection document outlining your challenges and achievements doing this assignment * Python codes that store / retrieve data from a relational database or NoSQL database on at least one of the datasets, and/or that produce Interactive Charts*[Optional – bonus marks will be given to those who manage to incorportate these functions]* | 50% |
| 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 | 15% |
| Self-reflection   * Explaination of challenges faced * Explaination of achievements made | 5% |

# Section 4 Sample outputs expected

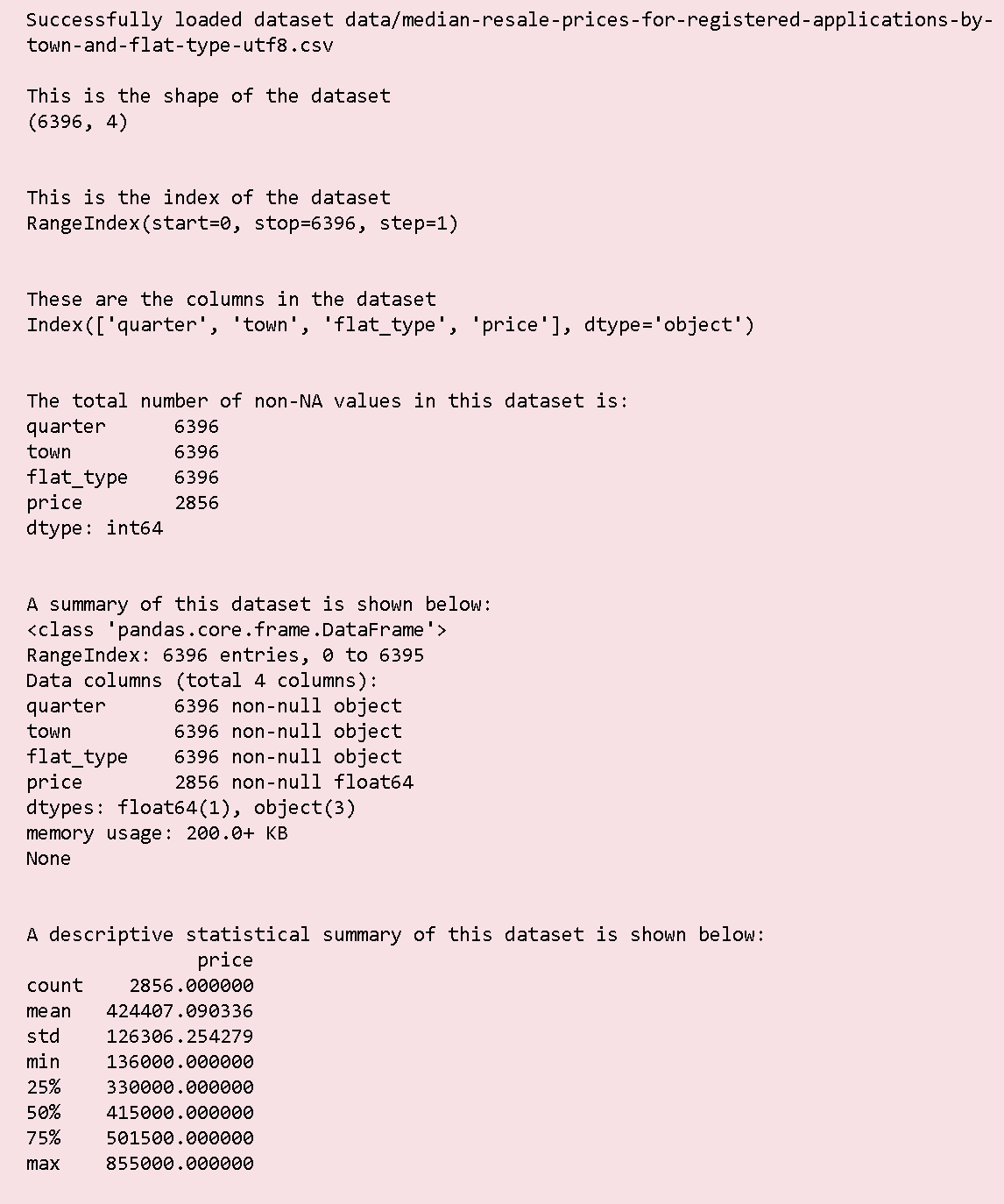
This section contains sample screenshots of how your Python programs may look like.

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
* **Bokeh** – This library allows you to create a web-based data visualization program using Python with the ability to let the user interact with your graph. It is not so easy to use for a beginner but if you are able to master it, you can produce much more interesting and useful visualizations in your data analysis tasks
* **Pygal** – This library is really quite cool and allows you to create visualizations REALLY easily! You can use it stand-alone or with a web-based framework for Python like Flask. However, if you are ambitious and want to create a web-based application, do note that the latter is much more difficult to achieve because it would require you to have knowledge of not just Python, but also HTML, CSS, Javascript and jQuery.

## Example 1 Simple Text-based Analysis using Pandas

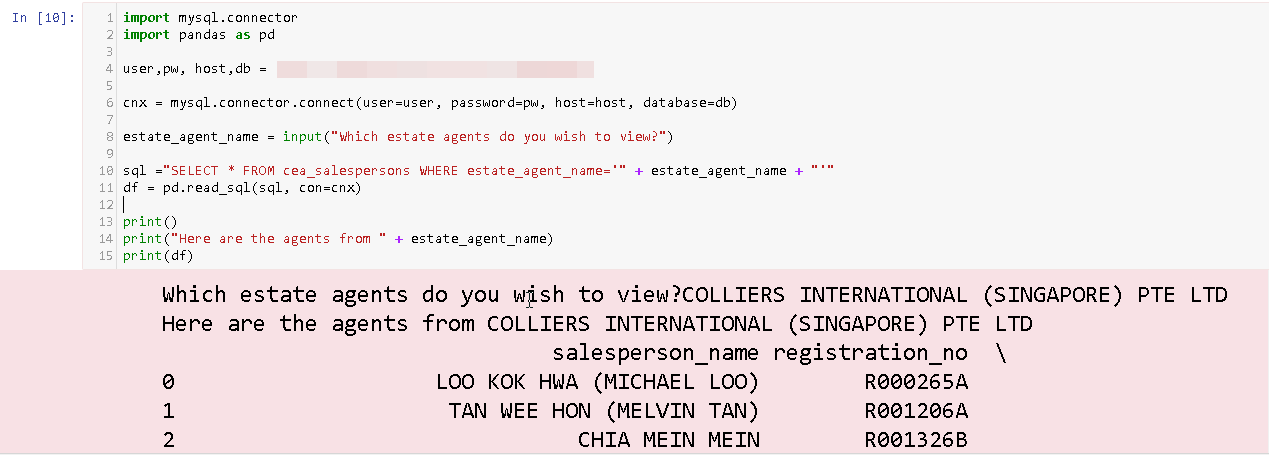


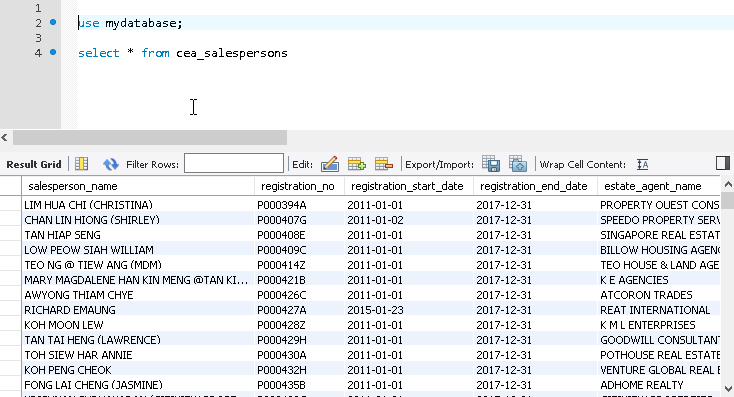
## Example 2 Viewing data inserted in mySQL

Your assignment submission must include a component where you demonstrate your competency in loading and storing data to an external relational database or NoSQL database.

A simple way where you can practise your skills in this area and fulfill the basic requirement of this assignment is to choose any one dataset, read it using Pandas, then store it in your preferred database. Subsequently, allow the user to query the database as shown in this simple example below.

For this example, the data was loaded from https://data.gov.sg/dataset/cea-salesperson-info





## Example 3 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)

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.

A picture containing text, crossword puzzle

Description generated with very high confidence

## Example 4 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

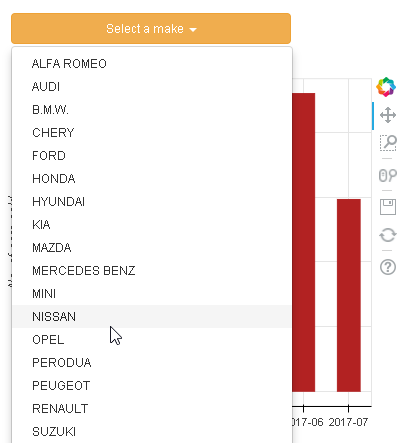
Description generated with very high confidence

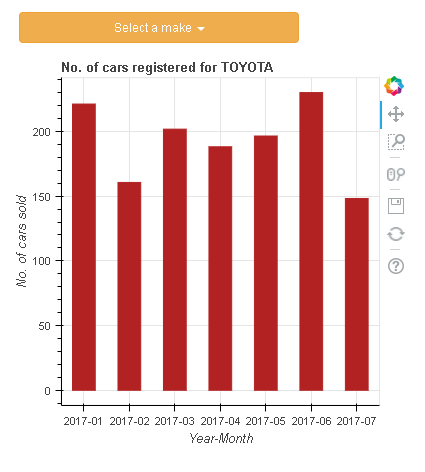
## Example 5 Interactive Data Visualization using Bokeh

This sample output uses the [Bokeh](https://bokeh.pydata.org) library to plot an interactive bar chart visualization showing the number of cars sold over the past year for a certain make of cars (data from data.gov.sg)

The user can select from a list of car-makes to view the information for that specific car make only. The graph changes dynamically based on the user’s choices.

Bokeh is quite a nice library to use and you can do so much more with it than using Seaborn or Matplotlib, but do note that its learning curve is less trivial than Seaborn.



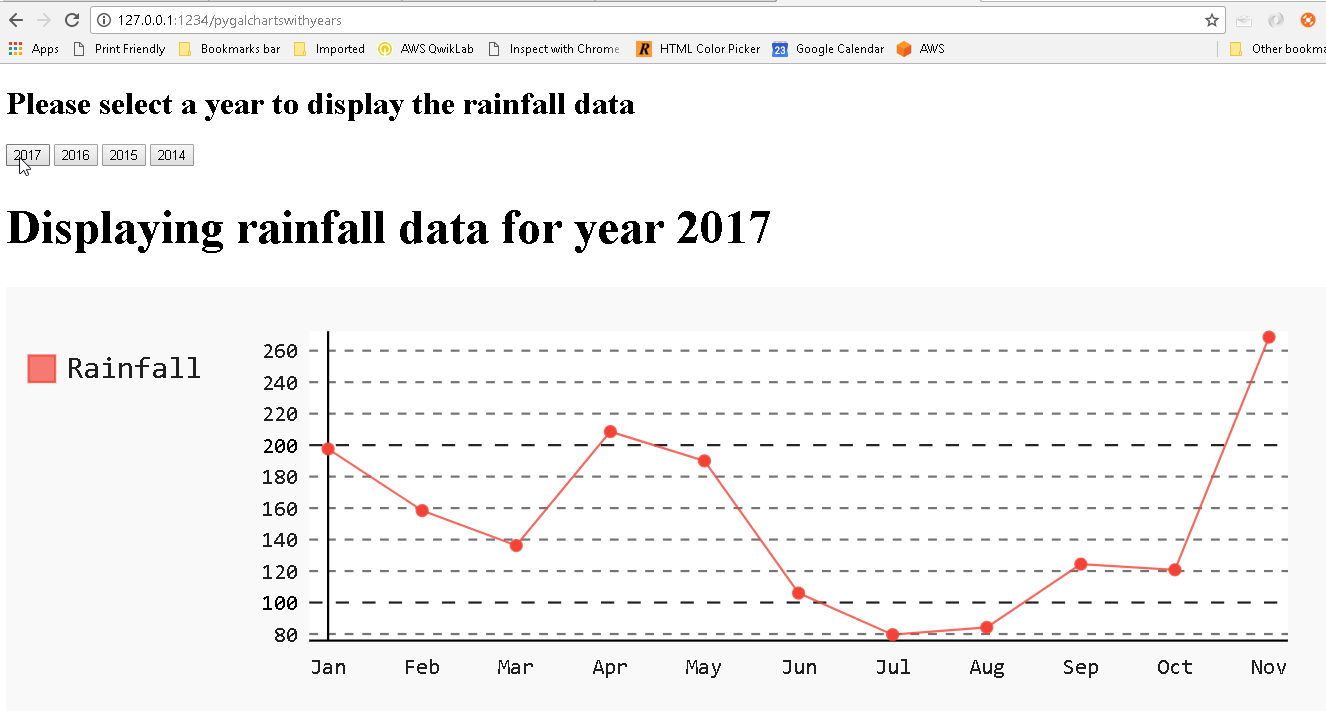


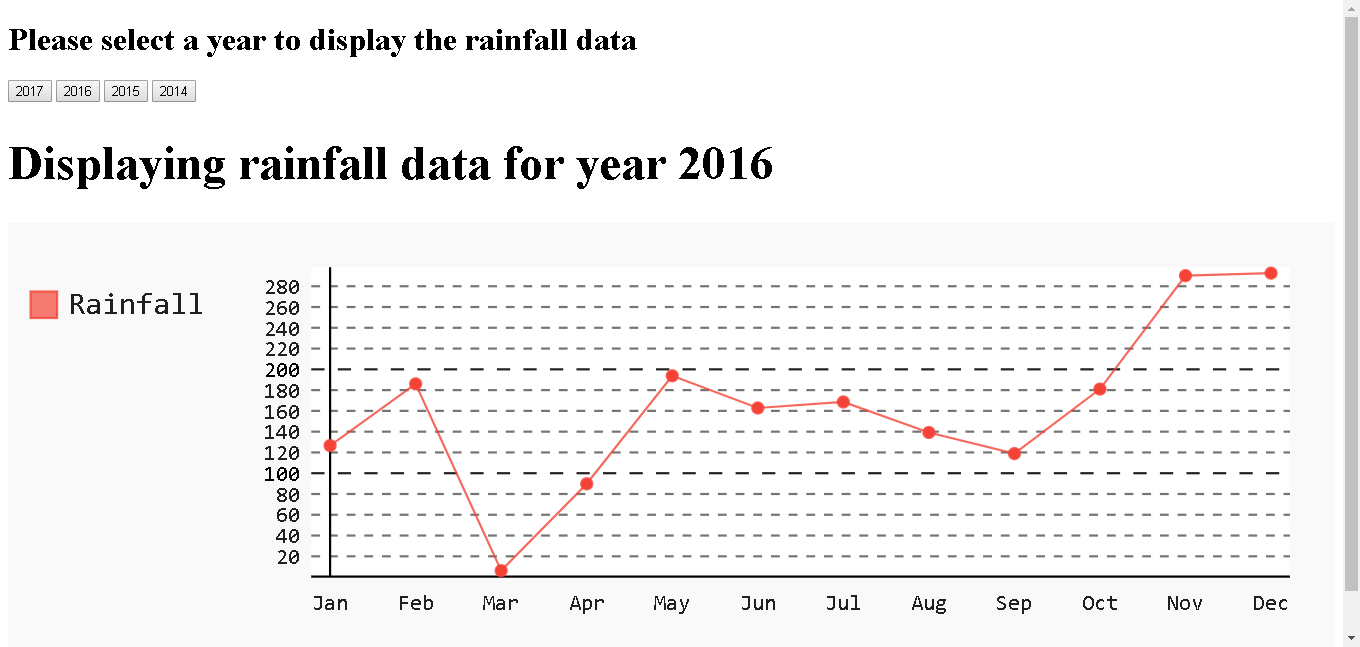
## Example 6 Interactive Data Visualization using Pygal

This sample output uses the [Pygal](http://www.pygal.org/en/stable/index.html) library to plot an interactive line chart visualization on a webpage showing rainfall data from 2014 to 2017 (data from data.gov.sg)

The user can select from the year to display the data. The graph changes dynamically based on the user’s choices.

This library is really awesome, so we encourage to play around with it if you have the time!





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