Coursework: Programming for Data Science (20COP504)

Dr Bangli Liu(b.liu2@lboro.ac.uk)

Coursework 2 (60%)

In this coursework, you will analyse a publicly available dataset named **IMDB-Movie-Data**. This is a data set of 1,000 most popular movies on IMDB in the last 10 years. The data points included are Title, Genre, Description, Director, Actors, Year, Runtime, Rating, Votes, Revenue, Metascore.

Go to https://www.kaggle.com/PromptCloudHQ/imdb-data and download the **IMDB-Movie-Data.csv**.

1.1

(1) Read the data into a DataFrame called **movies** and get the number of columns and rows.

[2 marks]

(2) Get the minimum, maximum, and mean value of each numerical column.

[4 marks]

(3) Rename **Runtime (Minutes)** to **Runtime_Minutes** and **Revenue (Millions)** to **Revenue_Millions**, respectively.

[4 marks]

1.2

Find all null values in the DataFrame **movies**. Fill the null values with the mean value of the corresponding column and produce a cleaned DataFrame called **clean_movies**. Please note that all the following tasks will be using **clean_movies**.

[10 marks]

1.3

Create a histogram to illustrate the distribution of the **Runtime**. Please set the bin size to 10.

[10 marks]

1.4

Rank the **Directors** according to their average **Rating**. Select the top ten **Directors** and draw a proper chart to compare their average **Rating**.

[10 marks]

1.5

(1) Use **NumPy** to generate three random integers as indices.

[2 marks]

(2) Use the generated three indices to select the corresponding three **Directors**.

[4 marks]

(3) Calculate the annual **Metascore** of the selected three **Directors**.

[8 marks]

(4) Use three subplots to illustrate their **Metascore** by year. The range of the x-axis should be from 2006 to 2016.

[6 marks]

1.6

Use **Revenue** and **Metascore** to select samples that satisfy some conditions (the selection should include three conditional operators and two logical operators). Based on the selected samples, draw a scatter chart to show the relationship between **Revenue** and **Metascore**.

[15 marks]

1.7

Use **NumPy** to generate 30 random integers from 0 to 1000. Use the generated 30 integers as indices and extract relative rows from the DataFrame **clean_movies**.

Draw a chart to present the relationship between **Year** and **Votes**. Please use **Year** as the x-axis and **Votes** as the y-axis. The range of the x-axis should be from 2006 to 2016.

[15 marks]

Format

- Please ensure that detailed comments are embedded inside the code.

[4 marks]

- Please ensure that your charts are labelled with a suitable colour scheme, title, and legend.

[6 marks]

Files to submit

- You will submit your coursework in the form of a single Jupyter notebook (i.e. the **.ipynb** file).
    - Please ensure that all your results are presented in the Jupyter notebook.
    - Please ensure that your codes are executable.
    - Wherever necessary, add meaningful comments to your code.
    - Use of additional Python packages is allowed.