Department of Computer Science & Engineering
Indian Institute of Technology Kharagpur

Part-III
Data Visualization with Python

Instructions:

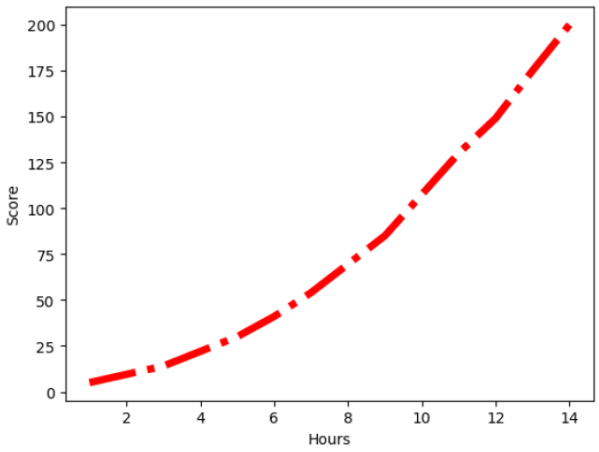
- Solve the problems using "Jupyter Notebook" linked to "Anaconda Navigator".
- For each problem, create a "Markdown" window, mention the problem you have asked to solve followed by "Code" window containing your program and output of the program.
- As comments as much as you can.
- Upload the .ipynb file. Give the name of your file as **A7-<RollNo>.ipynb.** You have to upload only one file in this lab.

Time: 2 hours Full Marks: 100

Problem 1:

```
# Considering the following dataset
x = [1, 3, 5, 6 ,7, 9, 11, 12, 13, 14]
y = [5, 14, 30, 41, 54, 85, 130, 149, 175, 200]
```

Draw a line graph that looks like the one shown below.



[5 points]

Problem 2:

You have given the following data points:

```
# Dataset 1
time = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]
speed = [30, 50, 70, 90, 110, 130, 150, 170, 190, 205]
# Dataset 2
x = [1, 2, 3, 4, 5, 6, 7, 8]
y = [6, 7, 8, 9, 14, 11, 13, 21]
```

Plot the lines that represent the two sets of data points.
Convert the data into a normalized scale (e.g., 0-1 normalization) and then plot the line graph with the normalized values.

Hint: for a data x_i, in a set x, it's 0-1 normalized value z_i will be:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Your graph should include labels and different colours, line styles for differ lines.

[15 points]

Problem 3:

Following are a few set of points:

Set 1: [1, 2, 3, 4, 5]
Set 2: [4, 6, 8, 10, 12]
Set 3: [3, 7, 9, 4, 10]

Plot the data as lines, bars, and histograms as a multiplot graph. Give suitable titles to each polt and their corresponding x-labels and y-labels;

[10 points]

Problem 4:

```
# Considfer the following information
```

Section-1:[89, 65, 81, 90], Section-2:[72, 70, 75, 95],
Section-3:[65, 60, 70, 80], Section-4:[80, 75, 85, 82]}

Data in each section are the average marks obtained in "Physics", "Chemistry", "Maths", "Biology".

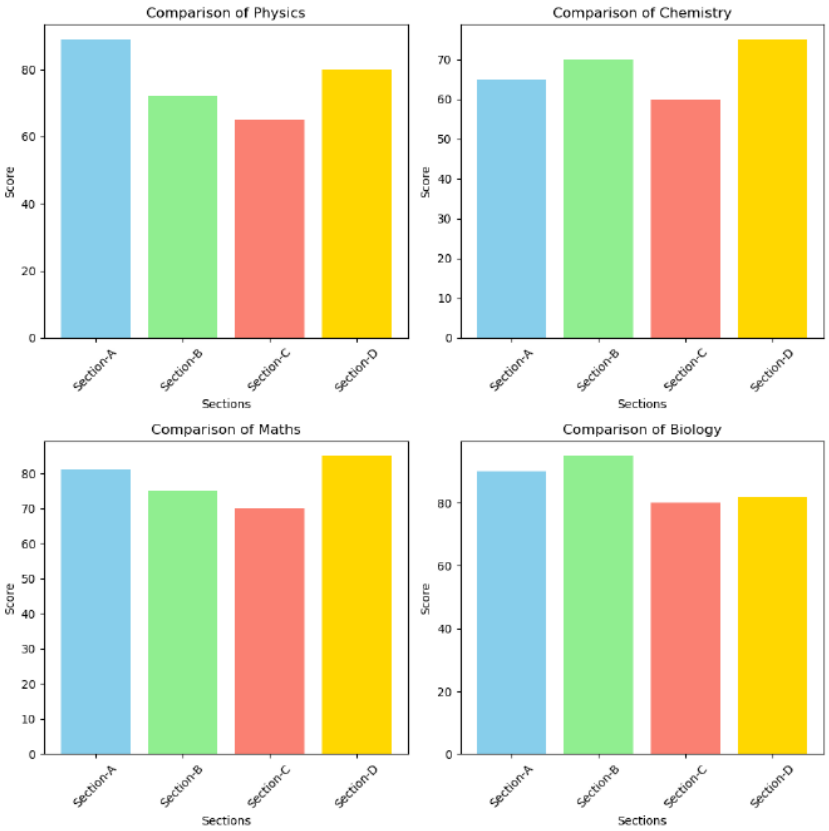
Create a data frame using pandas.
Hint:

```
import pandas as pd
data = {
'Section-1':[89, 65, 81, 90], 'Section-2':[72, 70, 75, 95],
'Section-3':[65, 60, 70, 80], 'Section-4':[80, 75, 85, 82]}
Subjects = ["Physics", "Chemistry", "Maths", "Biology"]
# Creating the dataframe with the row index
df = pd.DataFrame(data, index= subjects)
```

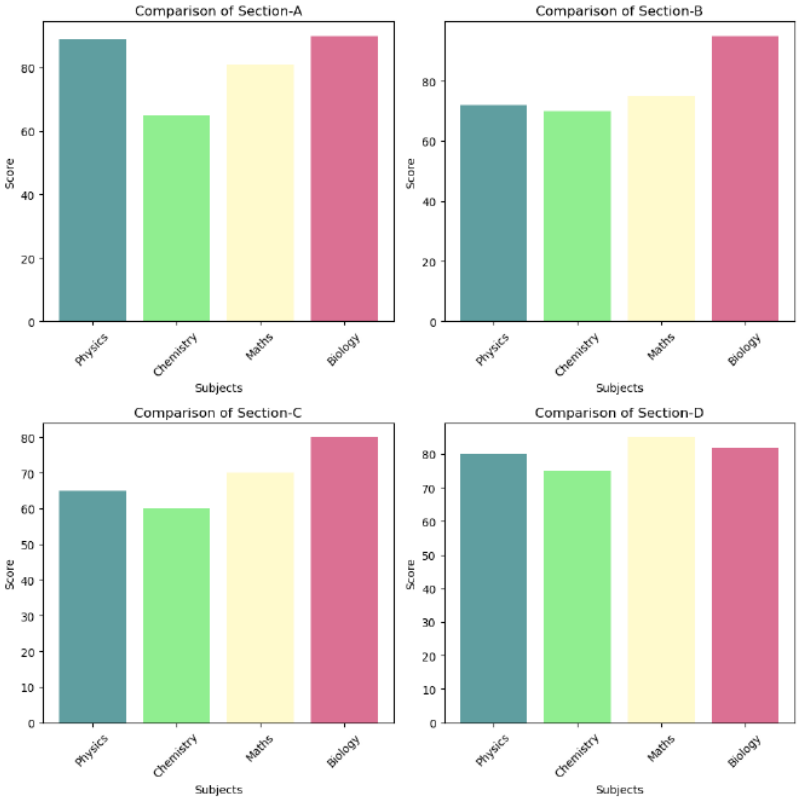
Display the dat which would look as shown below:

	Section-1	Section-2	Section-3	Section-4
Physics	89	72	65	80
Chemistry	65	70	60	75
Maths	81	75	70	85
Biology	90	95	80	82

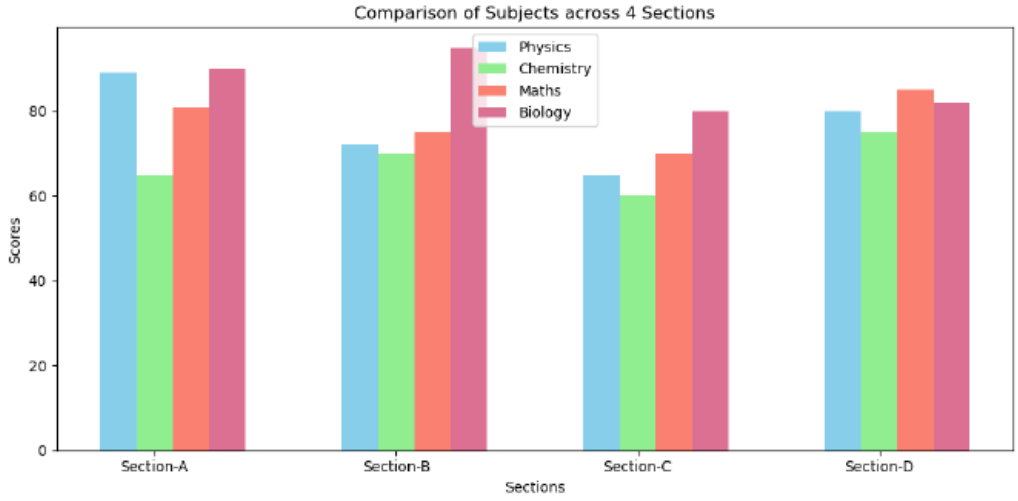
Plot the bar charts, which show the mark distribution as shown in the following graphs.



Draw another graph which shows the comparison of the average marks obtained in different subjects across the different sections. A sample representation is given below.



Draw another grouped bar chart that visualises the data shown below.



[30 points]

Problem 5:

Consider the following information.

Training Data1 = [1500, 800, 950, 1210]

Test Data1 = [500, 600, 550, 400]

#For Set2

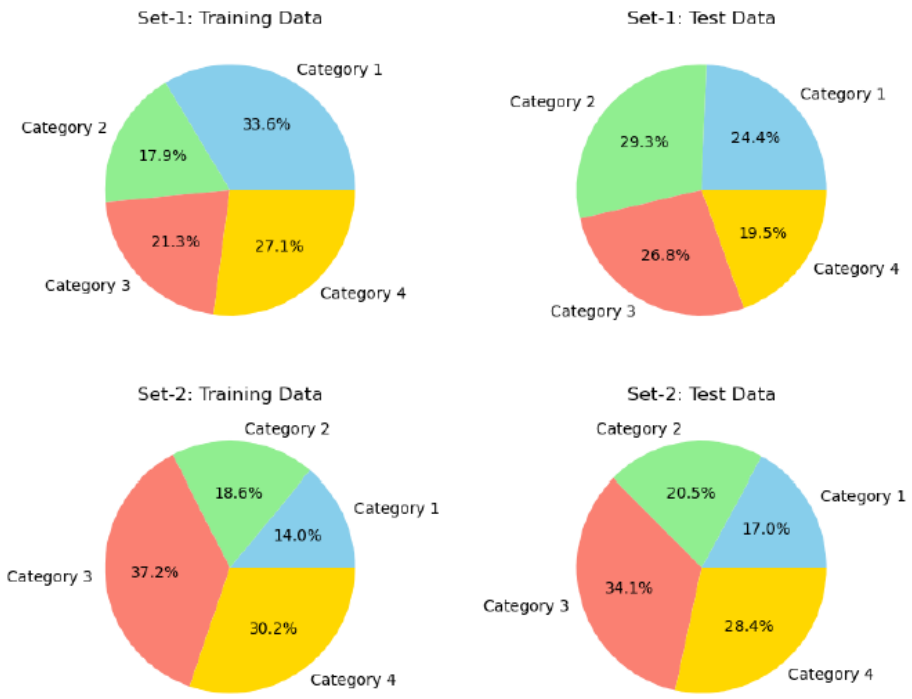
Training Data2 = [2400, 3200, 6400, 5200]

Test Data2 = [1500, 1800, 3000, 2500]

 # Labels for categories

labels = ["Category 1", "Category 2", "Category 3", "Category 4"]

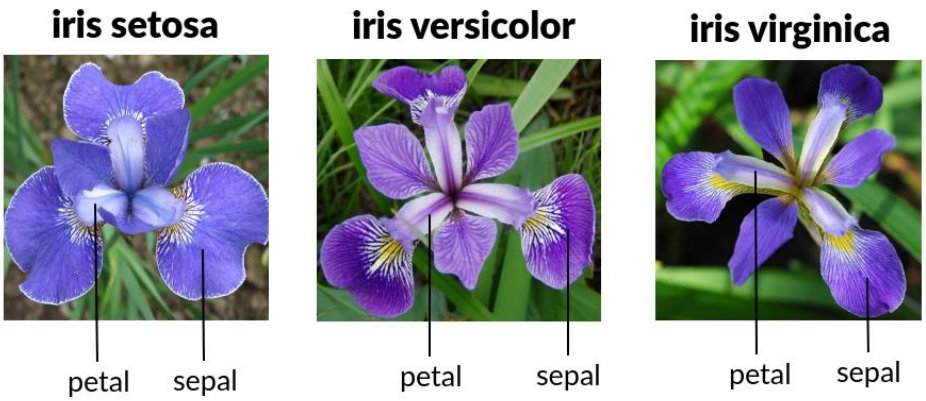
Draw the pier-carts which will visualize the data as shown below:



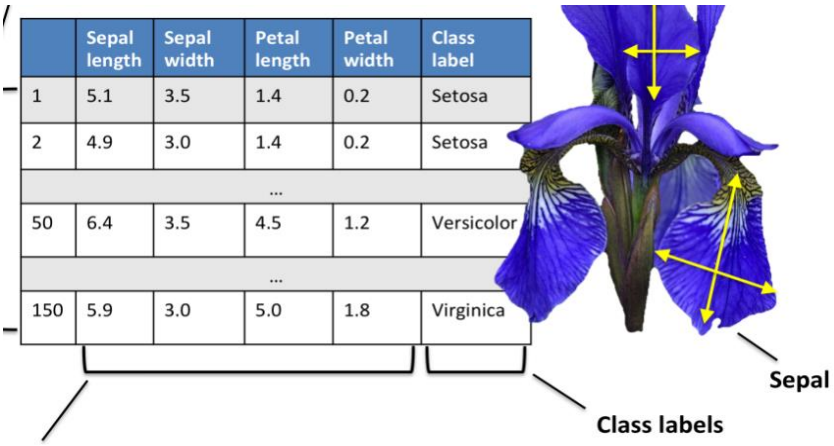
[20 points]

Problem 6:

The Iris flower data set or Fisher's Iris data set is a multivariate data set used and made famous by the British statistician and biologist Ronald Fisher in his 1936 paper. The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. The IRIS dataset can be downloaded from Index of /ml/machine-learning-databases/iris (uci.edu).



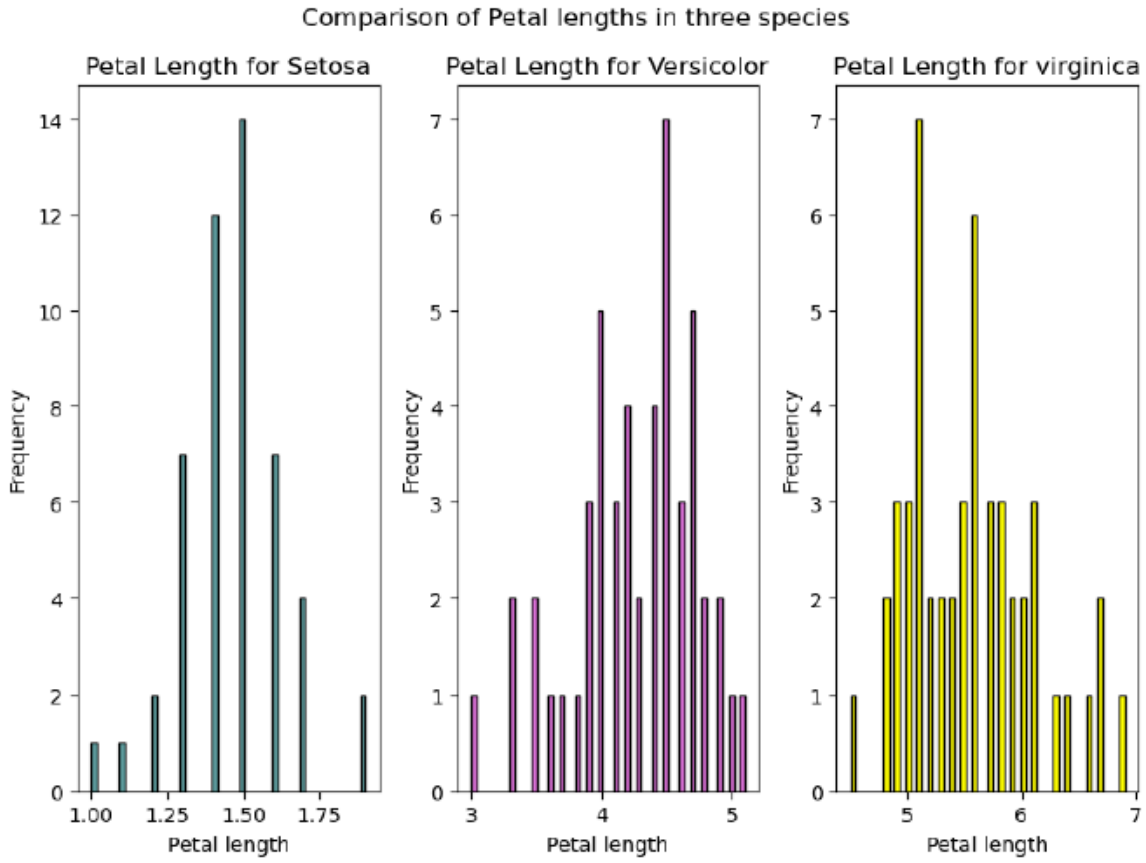
Consider the features of the flowers as shown below:

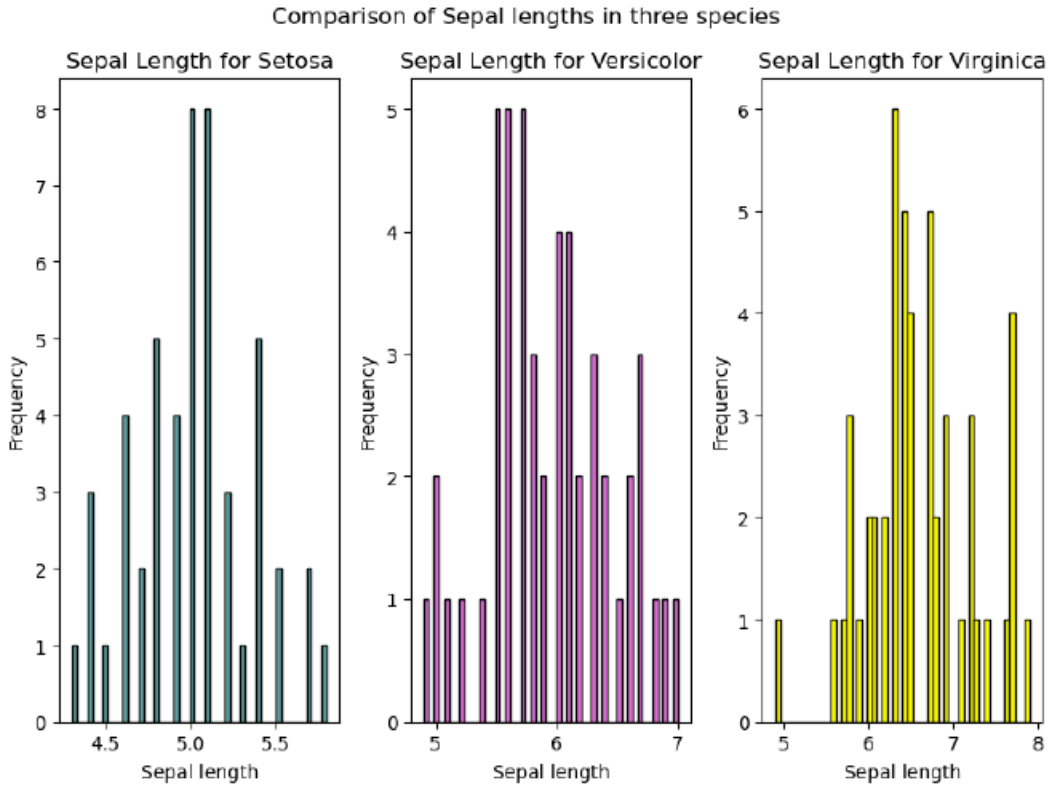


Note: The input is the features of the flowers, and not the image of the flowers!

Download the iris.csv data set from the link given above.

Use the data set and draw six histogram plots showing how sepal and petal lengths vary across the different types of colours. Your histogram is expected to resemble the following graphs.


Comparison of Petal lengths in three species
Petal Length for Setosa
Petal Length for Versicolor
Petal Length for virginica
Frequency
Petal length


Comparison of Sepal lengths in three species
Sepal Length for Setosa
Sepal Length for Versicolor
Sepal Length for Virginica
Frequency
Sepal length

[20 points]