Instructions:

Try to solve all problems on your own. If you have difficulties, ask the instructor or TAs.

Please follow the instructions given below to prepare your solution notebooks:

- Please use different notebooks for solving different Exercise problems.
- The notebook name for Exercise 1 should be `ROLLNUMBER-labLL-ex1.ipynb`. `ROLLNUMBER-labLL-ex2.ipynb`, etc for others. 'LL' is the two digit lab number (lab-3 is 03, etc).
- Please ask your doubts to TAs or instructors or post in Moodle Discussion Forum channel.
- You should upload on the .ipynb files on Moodle (one per exercise).

Only the questions marked [**R**] need to be answered on paper. Write legible and to-the-point explanations. The work-sheet on which you write needs to be submitted before leaving the session.

Some other questions require plotting graphs (histograms, trajectories, level-sets etc) or tables. Please make sure that the plots are present in the submitted ipython notebooks.

Submission Time: Please check the submission deadline as show on the assignment web-page in Moodle. Late submissions will be accepted upto 24 hours from the deadline. All late submissions will have a penalty of 3 marks. Submissions later than 24 hours after the deadline will not be accepted.

The fifth laboratory exercise aims to helps us do some data processing and forecasting.

Exercise 1[10 marks] Suppose we would like to make some estimates about emails received by Prof Clue Less at IIT Bombay based on historical information. Two files are available online with past data. Each row contains information about one email.

- File 'unfiltered-2023-24.csv' has data from all emails received by the professor over a period of about two years. We will call it Unfiltered Data.
- File 'filtered-2012-24.csv' has data from some (not all) emails received by the professor over a period of about 13 years. These files were considered 'interesting or useful' and hence saved for future. We will call it the Filtered Data.

The files have columns on 'Date and Time (day, month, year, time, timezone) of sending', domain of sender (iitb, ee.iitb, gmail etc), whether the email was written by the professor to himself, size (bytes), whether the email had the professor in the 'To:' field (otherwise he is in CC or a mailing list or an alias), whether the email is from seminar@iitb, discuss-faculty or faculty-notices mailing lists, whether it is spam, spam-score, whether it has a zip file attachment, html or plain-text, pdf attachment, doc, ppt, calendar invite, User-Agent, and whether image file attached.

1. Suppose we want to understand the rate of arrivals of emails. The data has date of sending. Assuming it is same as the date of receiving, interpret the data for average number of emails received in last two years on each of the seven days of the week. Also compute variance or range of these numbers. Use an appropriate graph or chart to represent this data. Only one graph please, using the Unfiltered Data.
2. **[R]** One can collect this data by looking at the 'original content' of the email. To see the actual email that you recieve (not just the content), you can login into webmail, open any one email, then click on 'more' button on top and then 'show source'. For each of the following data, write the header that contains the details (for example, the 'From:' header has the name and email of the sender.
   (a) Time at which the email is sent
   (b) Spam score
3. **[R]** Which header is appears more than once? Can you guess what it shows?
4. Suppose we want to estimate how many emails are deleted or removed by the professor. Use the two files and estimate this fraction.
5. The professor has some biases. Is an email with 'html' content more likely to be deleted as compared to one without html?
6. Similarly, is an email with any attachments (like pdf, doc, ppt or image) more likely to be deleted than an email without them?
7. The professor is worried that the recieved number of emails is increasing over time. Explain with suitable information and/or a figure whether it is true.
8. Which weeks or months of an year see the highest email traffic for the professor? Explain using a suitable chart.
9. Forecast the number of emails the professor can expect
   (a) on the coming Monday
   (b) in the coming week (Mon-Sun)