Instructions:

We will discuss data clustering in this lab.

We will use `pandas` and `numpy` libraries to process the data. We shall use `seaborn` and `matplotlib` libraries for plotting purposes.

We shall discuss K-Means algorithm to cluster data. We will use `scikit-learn` to perform K-Means clustering.

Please follow the instructions given below:

- Please use different notebooks for solving different problems.
- The notebook name for Exercise 1 should be `YOURROLLNUMBER_IE507_Lab8_Ex1.ipynb`.
- Similarly, the notebook name for Exercise 2 should be `YOURROLLNUMBER_IE507_Lab8_Ex2.ipynb`, etc.
- Please post your doubts in MS Teams or Moodle so that TAs can clarify.

For more details on `pandas`, please consult https://pandas.pydata.org/docs/getting_started/intro_tutorials/index.html.

For more details on `matplotlib`, please consult https://matplotlib.org/stable/tutorials/index.html.

For more details on `seaborn`, please consult https://seaborn.pydata.org/.

For more details about `scikit-learn` clustering library, please consult https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

There are only 3 exercises in this lab. Try to solve all problems on your own. If you have difficulties, ask the Instructors or TAs.

Only the questions marked [**R**] need to be answered in the notebook. You can either print the answers using `print` command in your code or you can write the text in a separate text tab. To add text in your notebook, click `+Text`. Some questions require you to provide proper explanations; for such questions, write proper explanations in a text tab. Some questions require you to prepare plots, for such questions write codes to produce the required plots.

After completing this lab's exercises, click File → Download .ipynb and save your files to your local laptop/desktop. Create a folder with name `YOURROLLNUMBER_IE507_Lab7` and copy your `.ipynb` files to the folder. Also copy the `.csv` files to the folder. Some questions require the appropriate files to be included in folder. Please include all related files required to execute your code in the folder. Then zip the folder to create `YOURROLLNUMBER_IE507_Lab7.zip`. Then upload only the `.zip` file to Moodle.

The deadline for today's lab submission is **tomorrow, 11 59 PM Indian Standard Time (IST)**.

Exercise 1: Clustering using data from S1.txt file [25 marks]
Consider the practice code posted in Moodle.

1. Try to understand the **K-Means++** algorithm implemented in `scikit-learn` package.
2. [**R**] For the data in `S1.txt`, vary the number of clusters K by choosing from the set $\{6, 7, 8, 9, 10, 11, 12, 13\}$ and use the data in the KMeans function of scikit learn package.
3. [**R**]For each value of K, prepare the scatter plots depicting the clusters using different colors along with the cluster centers depicted in the same plot with a color different from those used for clusters.
4. [**R**]Explain your observations about the clustering results you obtained when the value of K is increased from 6 till 13.
5. [**R**]Consider the `test.txt` file given in moodle and find the predictions for the points in `test.txt` for clustering obtained for each value of K. Plot the points in the scatter plot and indicate the predicted cluster labels.
6. [**R**]Explain your observations about the predictions obtained for different values of K.
7. [**R**]Can you suggest your procedure which can be used to find the best choice for the number of clusters?
8. [**R**]Implement your procedure for the data from `S1.txt` and report the best choice for the number of clusters.
9. Explain how you can modify the data in `S1.txt` so that the mean of each column is 0 and variance is 1. This procedure is called column normalization.
10. [**R**]Write the appropriate code to do the column normalization.
11. [**R**]On the new data set thus obtained where the columns have mean 0 and variance 1, repeat the clustering for K in $\{5, 6, 7, 8, 9, 10, 11, 12, 13\}$.
12. [**R**]For each value of K, prepare the scatter plots depicting the clusters using different colors along with the cluster centers depicted in the same plot with a color different from those used for clusters.
13. [**R**]Explain your observations about the clustering results you obtained when the value of K is increased from 5 till 13.
14. [**R**]Explain how you will modify the test data given in `test.txt` file so that it can be used for prediction? Using your idea, convert the data in `test.txt` so that it can be used for prediction and report the predicted labels. Prepare scatter plots where you plot the transformed data from `test.txt`.
15. [**R**]Explain your observations about the predictions obtained for different values of K.
16. [**R**]Using your procedure to find the best choice of number of clusters, report the best choice for the number of clusters for the column normalized data.
17. [**R**]Did you observe any differences when the data from `S1.txt` was used without normalization and with normalization? Explain.
18. [**R**]Did you observe any differences during prediction when the data from `S1.txt` was used for clustering without normalization and with normalization? Explain.
19. [**R**]Explain a situation where normalizing the data might help.

Exercise 2: Clustering using data from other files [**15 marks**]
Answer the following questions for the other data sets from `b4.txt`, `e3.txt`, `u1.txt` files posted in moodle.

1. Vary the number of clusters K by choosing from the set $\{3, 5, 7, 9, 11, 13, 15, 17, 19\}$ and use the data in the KMeans function of `scikit-learn` package.
2. [**R**]For each value of K, prepare the scatter plots depicting the clusters using different colors along with the cluster centers depicted in the same plot with a color different from those used for clusters.
3. [**R**]Explain your observations about the clustering results you obtained when the value of K is increased from 3 till 19.
4. [**R**]Consider the `test.txt` given in moodle and find the predictions for the points in `test.txt` for clustering obtained for each value of K. Plot the points in the scatter plot and indicate the predicted cluster labels.
5. [**R**]Explain your observations about the predictions obtained for different values of K.
6. [**R**]Using your procedure to find the best choice of number of clusters derived in Exercise 1, report the best choice for the number of clusters for the column normalized data.
7. Normalize the columns of the data to be of mean 0 and variance 1.
8. [**R**]On the new data set thus obtained where the columns have mean 0 and variance 1, repeat the clustering for K in $\{3, 5, 7, 9, 11, 13, 15, 17, 19\}$.
9. [**R**]For each value of K, prepare the scatter plots depicting the clusters using different colors along with the cluster centers depicted in the same plot with a color different from those used for clusters.
10. [**R**]Explain your observations about the clustering results you obtained when the value of K is increased from 3 till 19.
11. [**R**]Modify the test data given in `test.txt` file so that it can be used for prediction and report the predicted labels. Prepare scatter plots where you plot the transformed data from `test.txt`.
12. [**R**]Explain your observations about the predictions obtained for different values of K.
13. [**R**]Using your procedure to find the best choice of number of clusters, report the best choice for the number of clusters for the column normalized data.
14. [**R**]Did you observe any differences when the data was used without normalization and with normalization? Explain.
15. [**R**]Did you observe any differences during prediction when the data was used for clustering without normalization and with normalization? Explain.

Exercise 3: Clustering with more than 2 dimensions [**15 marks**]
Answer the following questions for the data set from `f8.txt` file posted in moodle.

1. [**R**] Perform K-Means clustering on the data from `f8.txt` using `scikit-learn package`.
2. [**R**] Explain how you will visualize the clusters for the data from `f8.txt`?
3. [**R**] Design and illustrate a suitable idea for visualizing the clusters obtained for data from `f8.txt`. Implement your idea and prepare the required plots to visualize the clusters.
4. [**R**] Plot the cluster centers in the plots thus prepared.
