Instructions:

We will discuss data visualization techniques in this lab.

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

We shall discuss about different types of plots which could be used to visualize patterns in the data.

Please follow the instructions given below:

- Please use different notebooks for solving different problems.
- The notebook name for Exercise 1 should be `YOURROLLNUMBER_IE507_Lab7_Ex1.ipynb`.
- Similarly, the notebook name for Exercise 2 should be `YOURROLLNUMBER_IE507_Lab7_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/.

There are only 2 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: Questions about visualization tools [30 marks]
Consider the practice code posted in Moodle.

1. [**R**] After loading the data into the pandas dataframe `df`, write code to identify the number of rows and columns that `df` has, and print them.
2. [**R**] Why are the `num_major_vessels_fluroscopy` and `thal` columns considered object types? Write the reason.
3. [**R**] From the histogram on `age` attribute, identify the number of bins and bin size. Report these quantities.
4. [**R**] Plot the histogram on `age` attribute for 50 bins and report the bin size and your observations.
5. [**R**] What is the KDE option useful for in `histplot()`? Explain the details.
6. [**R**] Plot pandas based histogram and seaborn based histogram for `serum_cholesterol` attribute. Use bin sizes from {default, 20, 50, 100, 200, 500}. For seaborn, use KDE. Report the observations.
7. [**R**] In the plot depicting the histogram of `serum_cholesterol` attribute containing mean and median, add also the vertical lines to represent the 25 percentile and 75 percentile values in the `serum_cholesterol` attribute. Use different colors and appropriate legend.
8. [**R**] Change the order in the bar plots for `gender` vs `serum_cholesterol` from male, female to female, male and replot.
9. [**R**] Explain the difference between the bar plot obtained using the median estimator for `gender` vs `serum_cholesterol` and the bar plot obtained before.
10. [**R**] Explain the observations from the bar plot containing `gender` vs `serum_cholesterol` grouped according to `chest_pain_type`.
11. [**R**] Note that the `chest_pain_type` attribute is numerical and hence is of less value in the bar plot obtained for `gender` vs `serum_cholesterol` grouped according to `chest_pain_type`. To make the plot more meaningful, insert a new column to the dataframe which contains the description according to the corresponding `chest_pain_type` code. Name this column as `chest_pain_type_description`. To fill the values in this `chest_pain_type_description` column, take the description for `chest_pain_type` from description file. Construct the bar plot for `gender` vs `serum_cholesterol` grouped according to `chest_pain_type_description`. Add an appropriate legend and display the legend in a position where the bar graphs are clearly visible.
12. [**R**] Add an appropriate annotation indicating the value of the upper boundary values of the bar plots in the `gender` vs `serum_cholesterol` grouped according to `chest_pain_type`.
13. [**R**] Add an appropriate annotation with pointed arrows and with textual description in bar plot of `gender` vs `serum_cholesterol` grouped according to `chest_pain_type`. Color the arrow with a color other than red.
14. [**R**] Explain your observations from the scatter plot obtained for `age` vs `serum_cholesterol`.
15. [**R**] What do the light-colored bands and the dark central line indicate in the line plot of `age` vs `serum_cholesterol` indicate?

16. [**R**] What do the upper and lower boundaries of the box of `chest_pain_type` and `serum_cholesterol` indicate? What does the line inside the box indicate? What are the points marked beyond the error bars? Explain.

17. [**R**] Discuss the observations made from the box plot for `chest_pain_type` and `serum_cholesterol` grouped according to `gender`.

18. [**R**] Use violin plot to plot the relationship between `chest_pain_type` and `serum_cholesterol` and discuss the observations. Group the violinplots based on `gender` information and discuss the observations.

Exercise 2: Data visualization on a different data set [25 marks]
Consider the `cars.csv` posted in Moodle.

1. Load the data in `cars.csv` to a pandas data frame.
2. [**R**] Plot a histogram of `mpg` attribute using seaborn library. Use bin sizes from {default, 20, 50, 100, 200, 500}. Use KDE to plot the density graphs. Report the observations.
3. [**R**] Prepare a bar plot for `mpg` vs `displacement`. Report your observations. Add arrow-based textual annotations to the **highest bars** in the bar plot.
4. [**R**] Prepare a bar plot for `mpg` vs `displacement` and group according to `model_year` using median estimator. Add a legend at an appropriate location. Report your observations. Add annotations to denote the median values at the top boundary of the **highest bars** in the bar plots.
5. [**R**] Prepare a bar plot for `mpg` vs `displacement` and group according to `origin` using median estimator. Add a legend at an appropriate location. Report your observations. Add annotations to denote the median values at the top boundary of the **lowest bars** in the bar plots.
6. [**R**] Prepare a scatter plot between `mpg` and `horsepower`. Based on the plot, discuss if there is correlation between these attributes?
7. [**R**] Prepare a scatter plot between `mpg` and `acceleration`. Based on the plot, discuss if there is correlation between these attributes?
8. [**R**] Prepare a line plot between `model_year` and `horsepower`. Discuss the observations.
9. [**R**] Prepare a box plot between `model_year` and `weight`. Discuss the observations. Add annotations to the box plots corresponding to the years $70, 74$ at the top boundary of the corresponding boxes.
10. [**R**] Prepare a box plot between `model_year` and `displacement` and group according to `origin`. Discuss the observations.
11. [**R**] Prepare a violin plot between `model_year` and `acceleration` and group according to `origin`. Discuss the observations.
