MIE 223 Data Science
Lab and Assignment 03:
Data Visualization and Feature Analysis

In this lab and assignment, you will learn and gain hands-on experience on the use of Matplotlib and Seaborn for data visualization and the use of Pandas for feature analysis.

- Programming language: Python (Google Colab Environment)
- Due Date: Posted in Syllabus

Marking scheme and requirements: You are required to use the Jupyter notebook named `Assignment_03.ipynb` to provide appropriate answers to the questions in this assignment. All changes to the `Assignment_03.ipynb` notebook should be committed and pushed to your assignment repository on Github.

This assignment has *4 points* in total and the point allocation is shown below:

- Solution to Questions (1 point):
  - Q1(a): 0.2 point
  - Q1(b): 0.1 point
  - Q1(c): 0.1 point
  - Q2(a): 0.1 point
  - Q2(b): 0.1 point
  - Q2(c): 0.1 point
  - Q2(d): 0.1 point
  - Q2(e): 0.2 point
- Code review (1 point)
- Quiz (2 points)

What/how to submit your work:

- All your code should be included in the provided notebook named `Assignment_03.ipynb`.
- Commit and push your work to your github repository in order to submit it. Your last commit and push before the assignment deadline will be considered to be your submission. You can check your repository online to make sure that all required files have actually been committed and pushed to your repository.

- A link to create a personal repository for this assignment is posted on QUERCUS.

Notes that you should pay attention to

1. All your code should be written in the provided notebook named `Assignment_03.ipynb`.
2. All functions must *return* the specified return type.
3. Commit and push your work to your GitHub repository in order to submit it. **Your last commit and push will be considered to be your submission.** You can check your repository online to make sure that all required code has actually been committed and pushed to your repository.
4. **Your code should have output for all assignment questions unless otherwise stated**. If any of the assignment questions asks you to explain your outputs, please use markdown cells to clearly explain them.
5. As a backup plan, also upload a PDF copy of your assignment notebook to GitHub. This is useful especially when GitHub cannot render some images or outputs.
6. During the code review, TAs will go over your answer submissions to assess your understanding of the concepts relating to the assignment and the logic behind your codes. Please, ensure your codes are readable and well commented using python comments and markdown cells when necessary. Not have readable code may impact your code review score.
7. Please note the plagiarism policy in the syllabus.

1 Main Assignment

Please answer the questions below in the provided notebook named `Assignment_03.ipynb` using Google Colab. **Numpy** and **Pandas** functions should be used for all computations. Unless otherwise stated, **Matplotlib** should be used to generate plots. **Clearly give titles and labels to all plots**.

Q1. Data Visualization

You are provided with a Pandas Dataframe named **"anscombe_quartet_df"** which has 10 rows and 8 columns. The columns were obtained from four datasets, each with two variables x and y. The column names in the **"anscombe_quartet_df"** dataframe are such that they indicate the dataset and the variable in the dataset from which they were obtained. Specifically, column c_i is the variable c in dataset i $\forall$ $c \in \{x, y\}, i \in \{1, 2, 3, 4\}$. Use the **"anscombe_quartet_df"** dataframe to answer the following questions:

(a) Using Matplotlib, create a 2 by 2 subplot that shows the scatter plot of each of the datasets, such that subplot i (using Matplotlib's subplot indexing) contains the plot for dataset i $\forall$ $i \in \{1, 2, 3, 4\}$. In each subplot, the x variable of its corresponding dataset should be on the horizontal axis, while the y variable should be on the vertical axis. The summary statistics calculated by calling the provided function named **"summary_statistics"** on the **"anscombe_quartet_df"** dataframe are the same for all the datasets, which may indicate that the datasets are similar. In one sentence, what does visualizing each dataset tell you about the similarity of the datasets (are they similar or not)?

(b) Using Seaborn, create multiple boxplots that show the distribution plots of the x variable in each datasets. All the boxplots should be in a single plot axes.

(c) Repeat **Q1b** for variable y.

Q2. Feature Analysis

You are provided with a data about some Android apps on Google Play Store. The data was obtained from Kaggle. You can see the data dictionary here. Use the data to answer the following questions:

(a) Using Numpy, Seaborn and/or Matplotlib; create a correlation heatmap showing the Pearson correlation coefficient between all pairs of the continuous features in the subset of the data containing ONLY paid apps. The list of the continuous features to use is provided in the variable named **"continuous_cols"**.

(b) Using Seaborn, create a pairplot showing the relationship between all pairs of the continuous features (except **"Rating"** in the list named **"continuous_cols"**) in the subset of the data containing ONLY paid apps. Use **"is_good_rating"** variable as the hue.

(c) Repeat **Q2b** for the subset of the data containing ONLY free apps.

(d) Use a horizontal barplot to rank the discrete features (in the list named **"discrete_cols"**) in descending order of their Mutual Information (MI) with **"is_good_rating"** variable. Use the provided function named **"cal_MI"** to calculate MI.

(e) Use a 1 by 2 subplot to do a Frequency-MI analyses for **"Installs"** feature. Subplot 1 (using Matplotlib's subplot indexing) should use a horizontal barplot to rank the categories in **"Installs"** column in descending order of their frequency. Subplot 2 should use a horizontal barplot to rank the categories in **"Installs"** column in descending order of their Mutual Information (MI) with **"is_good_rating"** variable. Use the provided function named **"cal_MI"** to calculate MI.