Assignment 4

Data Preparation Techniques (COMP 3400)

Fall 2024

Important notes:

1. You are required to submit your assignment in *one* file in *IPython Notebook* format **(due date: Nov 24).** Please note that:
   - You may use *Markdown* in your IPython Notebook.
   - How you develop your IPython Notebook is your choice. You may use *jupyter.org* or *Google Colab*, or a locally installed Jupyter platform on your machine.
2. For some of the problems you may have to refer to the `matplotlib` API.
3. You are not allowed to use loops, in any of the problems unless otherwise stated (or you'll get a mark of 0 for that problem).
4. Slides covered in this assignment: **4-01** to **4-08**.

Assignment 4 involves three datasets, all available in the Data folder in the Course Shell:

1. `athlete_events` (from Assignment 3)
2. `london_weather`: Weather data recorded by a weather station near Heathrow airport in London, UK (you may find the description of columns here).
3. `user_behavior`: This dataset provides a comprehensive analysis of mobile device usage patterns and user behavior classification (you may find the description of columns here).

Problem 1 (20pts). In the `athlete_events` dataset, visualize the trend of the number of female athletes versus the number of male athletes in the sport of Athletics through the ages. Use *line plots* in your answer; where the x axis indicates the year of the event, and y axis indicates the number of athletes. Your plot should include two lines of different colors, one belonging to the number of female athletes, and one belonging to the number of male athletes.

Problem 2 (30pts). This problem aims at revealing the relationship between *global radiation* and *pressure* in the `london_weather` dataset in terms of distribution. Produce a 2×2 plot consisting of four subplots. The four subplots visualize the *2-d distribution* of *global radiation* vs. *pressure* in four seasons of 2020.

Problem 4 (20pts). In the `london_weather` dataset, use a *line plot* with *error bars* to visualize temperature readings in 2021. Where the x axis indicates the days of the year, and the y axis indicates the temperature readings. The upper/lower bounds of the error bars indicate the maximum/minimum temperature readings of the corresponding day.

Problem 4 (30pts). This problem aims at uncovering the relations in the `user_behavior` dataset between *Operating System*, *App Usage Time*, *Number of Apps Installed*, *Age*, and *Gender*. Draw a *scatter plot*, where x axis indicates *App Usage Time*, y axis indicates *Number of Apps Installed*, *Shape* of the points indicate *OS*, *size* of the points indicate *age*, and *color* of the points indicate *Gender*. Your plot should include legends for *OS*, *Age*, and *Gender*.