CSC380 **Homework 2** **Fall 2024**

This homework is due on Monday, Feb 11^{th} at 3:30pm. Late submissions are not accepted.

Submission Guidelines

- You must make two submissions:
  - Your complete homework as a SINGLE PDF file by the stated deadline to the gradescope. **Include your code and output of the code as texts in the PDF.**
  - Your codes to a separate submission: a single notebook file including codes for questions 1,3,4,5.
- For your PDF submissions:
  - Select the page number for the answer to each question in the Gradescope.
  - You may submit typed or handwritten/scanned answers. If you decide to submit handwritten answers then please ensure that it is easily readable.
  - You can easily scan and upload your answers as a PDF using the Gradescope mobile app.

1. **[15pts]** Download the 2012 London Summer Olympics Data from here. Using the matplotlib library, create a grouped bar chart for the olympics data where each country is shown at the x-axis with the names of the countries as the labels and each country has a group of three bars depicting its number of gold, silver, and bronze medals. Check here for an example of grouped bar charts. What are the pros and cons of this visualization compared with the stacked bar plot we showed in class for olympics data?

2. **[10pts]** Assume that we roll two fair six-sided dice. Let E be the event that the two dice's outcomes sum to 8. What is the probability of E?

3. **[15pts]** Continuing with question 2: Initialize the random seed to 2025 using `numpy.random.seed`. Using `numpy.random.randint`, simulate 1,000 throws of two fair six-sided dice. Paste your code here. From these simulations, what is the empirical frequency of E (i.e., the percentage of times this event occurred in simulation)?

4. **[10pts]** Continuing with question 3: Reset the random seed to 2025 and repeat the above simulation a total of 10 times and report the empirical frequency of E for each of the 10 runs. Paste your code here. The empirical frequency of E from each simulation will differ. Why do these numbers differ? Yet, the probability of E is fixed and was calculated in part (a) above. Why does the probability disagree with the empirical frequencies?

5. **[10pts]** Continuing with question 3, 4: In the above we have estimated the probability of an event by performing $1,000$ rolls of two dice each. We generated 10 different estimates by repeating this procedure. How do our results change if we instead performed $10,000$ rolls, and repeated 10 times? Try it, report the difference, and discuss why.

6. **[15pts]** Assume that we roll two fair six-sided dice. Let A be the event that the two dice sum to 6; let B be the event that the second die is even. Use the inclusion-exclusion principle to calculate $P(A \cup B)$.

7. **[25pts]** Below are the counts of undergraduate students in Bugsville University in different class years and majors:

	Freshman	Sophomore	Junior	Senior
CS Major	200	180	160	140
Non-CS Major	800	820	840	860

Table 1: Student Counts at Bugsville University

Suppose we select a student uniformly at random from the university roster.

(a) **[8pts]** Give the probability table of this random process. The table will have the same row and column names as the count table, and each cell should represent the probability of a randomly selected student belonging to that category and class year.

(b) **[10pts]** Similar to the blood type example discussed in class, create a 'marginal' row and a 'marginal' column in your probability table in (a) and fill in the respective entries using the law of total probability. We will have 7 new entries; explain in plain words the meaning of each entry.

(c) **[7pts]** Suppose we have seen that the student we select is a senior; how likely are we to see that this student is a CS major?

Does knowing that the student is a senior increase or decrease our belief that the student is a CS major?