CSE 351 - Introduction to Data Science (Spring 2024)
Assignment 1 (Homework 1):
Exploring Data Science and Essential Math Review

Due Date: Assignment 1 is due by 11:59 PM New York Time on Wednesday, February 21, 2024

This homework will reinforce the concepts taught in the class in lectures 1-2, familiarizing you with the real-world examples of these concepts.

Textbook Exercises
A solution manual of the textbook for selected exercises and problems are publicly available to the SBU community on web. You can use this resource for exams, quizzes and homework preparation, however, the direct use (copying) of suggested solutions for homework assignments are **prohibited** unless specified otherwise in the problem description.

Use of Online Resources
You are allowed to use online resources, including **Generative AI (GAI)** tools such as Chat **GPT** on the conditions that

1. you will list the used sources and the exact prompt in the case of GAI.
2. you will **not** directly copy the GAI answer.
3. you will **double check** GAI answers and list a **reference** (other than the GAI) for that.

Report the used resources at the last page of your assignment, under a section called **Used Resources** and use in-text citation for referencing.

Submission Guidelines
1. This assignment must be done individually and independently by every student. Your answers and codes will be checked thoroughly to detect copying/plagiarism. Do your own work!
2. If you do not have much experience with Python and the associated tools, this homework will be a substantial amount of work. Get started on it as early as possible! Below are some resources for practicing Python on your own.: Python.org, Python REU Tutorial by Prof. Michael Zingale
3. Please use **Piazza** to ask any questions.
4. Submit everything through **BrightSpace**. You will need to upload:
    1. The Jupyter notebook all your work is in (.ipynb file) – if your submission includes coding
    2. Python file (export the notebook as .py) – if your submission includes coding
    3. PDF (export the notebook as a pdf file)
    4. Please do **not** include the questions.

    These files should be named with the following format, where the italicized parts should be replaced with the corresponding values:

    1. cse351_hw1_lastname_firstname_sbuid.ipynb – if your submission includes coding
    2. cse351_hw1_lastname_firstname_sbuid.py – if your submission includes coding
    3. cse351_hw1_lastname_firstname_sbuid.pdf

Please keep in mind that:
(a) Academic dishonesty will result in an 'F' grade and referral to the Academic Judiciary.
(b) Late submission, beyond the 'due' date/time, will result in a score of 0 on this assignment.

Tasks [Points]

Exploring Data Science

1. Textbook [1-2]. Propose relevant data sources for the following *The Quant Shop* prediction challenges. Provide two sources and distinguish between sources of data that you are sure somebody must have (but may not be readily accessible to the public), and those where the data is clearly available to you. (**5 Points**)
   (a) Art auction price.
   (b) White Christmas.
   (c) Football champions.

2. Textbook [1-3]. Search the web for publicly available datasets. Identify three datasets that sound interesting to you. For each, write a brief description, and propose three interesting things you might do with them. You can use US Government's Open Data World Bank Open Data, Open Data on AWS, or NASA's Open Data Portal (**5 Points**)

3. Textbook [1-6]. You would like to conduct an experiment to establish whether your friends prefer the taste of regular Coke or Diet Coke. Briefly outline a design for such a study. (**5 Points**)

Math Review

Logarithm

4. Logarithms are very useful mathematical tools that play a crucial role in data analysis, offering powerful methods for transforming, analyzing, and interpreting data.. Briefly discuss three roles/use cases. (**5 Points**)

   Extra Credits - discuss two more roles/use cases that are not discussed in lectures or the textbook. (+**2 Points**)

Correlation

5. What is the difference between Correlation and Causation? Explain completely in a paragraph or two. (**5 Points**)

6. Consider the following paired data for variables X (hours of study per week) and Y (exam score out of 100). (**5 Points**)

Student	Hours of Study per Week (X)	Exam Score (Y)
1	5	62
2	8	74
3	7	69
4	9	76
5	11	85
6	4.5	60
7	10	80
8	3	55
9	12	90
10	6	65
11	7	70
12	10	82
13	2	50
14	13	92

15	8	78
16	14	95
17	5.5	63
18	3.5	58
19	12.5	88
20	11.5	86

a. What is the mean of X and Y?
b. What are the standard deviations of X and Y?
c. What is the covariance between X and Y?
d. Calculate the Pearson correlation coefficient, r, and interpret its value. Does it indicate a strong or weak linear relationship? Is the relationship positive or negative?

7. Consider the data in question 6. (**5 Points**)
   a. What are the ranks for X and Y?
   b. What are the differences (d) between the ranks of X and Y, and what are their squared values (d^2)?
   c. What is the sum of all squared differences ($\sum d^2$)?
   d. Calculate the Spearman Rank Correlation Coefficient, r_s , and interpret its value. Does it indicate a strong or weak monotonic relationship? Is the relationship positive or negative?

8. Compare your results from question 6 and 7. Explain the similarities and difference of the two correlation measures (Pearson's and Spearman's) based on the data. (**5 Points**)

   Extra Credits - Use plots to better illustrate your explanations. (+**5 Points**)

Statistics

9. Textbook [2-6]. Compare each pair of distributions to decide which one has the greater mean and the greater standard deviation. You do not need to calculate the actual values of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, just how they compare with each other. (**10 points**)
   (a) i. 3; 5; 5; 5; 8; 11; 11; 11; 13.
   ii. 3; 5; 5; 5; 8; 11; 11; 11; 20.

   (b) i. -20; 0; 0; 0; 15; 25; 30; 30.
   ii. -40; 0; 0; 0; 15; 25; 30; 30.

   (c) i. 0; 2; 4; 6; 8; 10.
   ii. 20; 22; 24; 26; 28; 30.

   (d) i. 100; 200; 300; 400; 500.
   ii. 0; 50; 300; 550; 600.

   Extra Credits - Write a Python program that computes the $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ for each pair of distributions above and validate your answer based on the results. (+**5 Points**)

Probability

10. Textbook [2-2]. Suppose that P(A) = 0.3 and P(B) = 0.7. (**5 Points**)
    (a) Can you compute P(A and B) if you only know P(A) and P(B)?
    (b) Assuming that events A and B arise from independent random processes:
    I. What is P(A and B)?

II. What is P(A or B)?
III. What is P(A|B)?

11. What is the difference between Probability and Statistics? Explain completely in a paragraph or two. (**5 Points**)

12. A particular disease affects a small percentage of the population. A screening test has been developed to detect this disease. However, like all tests, it is not perfectly accurate. Some people who have the disease will test negative (false negatives), and some people who do not have the disease will test positive (false positives). Consider the following data, (**5 Points**)

Outcome	Disease Present (D+)	Disease Absent (D-)	Total
Test Positive (T+)	95	105	200
Test Negative (T-)	5	1795	1800
Total	100	1900	2000

Calculate the following probabilities:
a. P (D +) the prior probability of having the disease
b. P (T +) the total probability of testing positive
c. P(D + | T +) the probability of having the disease given a positive test result
d. P(T + | D +) the probability of testing positive given the presence of the disease (sensitivity of the test)
e. P (T - | D +) the probability of testing negative given the absence of the disease (specificity of the test)
g. Interpret the implications of your findings for the effectiveness of the screening test.