CS 3753 & 5163 Data Science

Homework 3 (110 points)

Submission:

1. Submit a single python script (abc123_hw#.ipynb or abc123_hw#.py) through Blackboard learn. All the results are outputted from your Python code.
2. You should have the instruction of running your code at the beginning of your code (e.g. Download the file abc123_hw#.py; Open the file in Spyder; Run the code by clicking the "Run" button, …). It should run successfully either in the basic command prompt with Python3, Jupyter Notebook, or Spyder.
3. If your code cannot run, we assume your code can run, then we will check whether your code is correct logically. If so, half points will be deducted. Otherwise, more points will be deducted if your code is wrong or there is no code.
4. Do not compress your source code and data files. The compressed files will receive a warning at the first time and will lose 10% points in future assignments. Make sure all your files are in the same folder when you run the code. So, after the graders download your homework, they do not need to set the path for the data file. They can run your code successfully.
5. If there is any plagiarism, you will lose all points on the questions at first time. In next, you will lose all points in the whole homework.
6. You can submit your homework 3 times before the deadline. The late submission will lose 15% of the total points in the assignment. The submission is unacceptable if it is more than 24 hours late. The compressed files will receive a warning at the first time and will lose 10% points later.

Questions

1. Matrix multiplication (10 pts)

   $A=\begin{pmatrix}2 & 8 & 4\\ 5 & 4 & 2\end{pmatrix}$; $B=\begin{pmatrix}4 & 1\\ 6 & 4\\ 5 & 3\end{pmatrix}$; $C=\begin{pmatrix}4 & 1 & 2\\ 6 & 4 & 3\\ 5 & 3 & 4\end{pmatrix}$; $D=\begin{pmatrix}4 & 1 & 2\\ 6 & 4 & 3\end{pmatrix}$

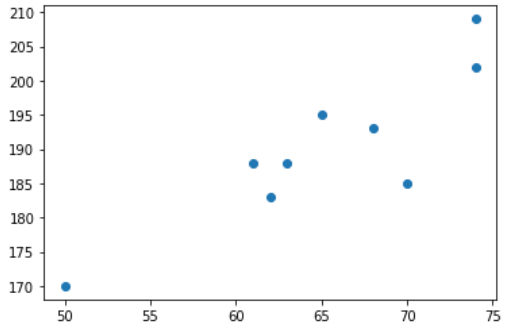
   Write a python code to calculate A·B, A·C, and A·D using dot product function. If any of them cannot be calculated, please explain why.

2. Pearson correlation (10 pts)

   x = np.array([50, 68, 74, 70, 65, 61, 63, 74, 62])

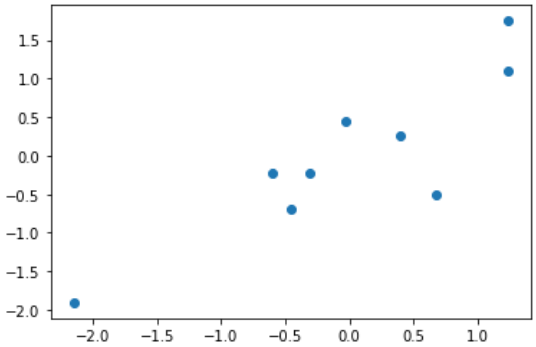
   y = np.array([170, 193, 209, 185, 195, 188, 188, 202, 183]) (10 pts)

   a. Scatter plot of x and y. (2 pts)



b. Define the function of zscore() to compute the zscore of x and y. (3 pts)

c. Scatter plot of the zscore of x and y. (2 pts)



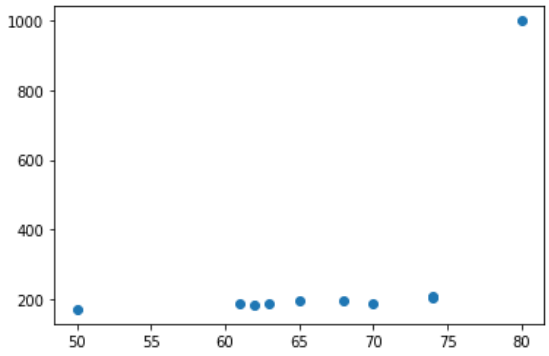
d. Calculate the Pearson correlation coefficient using your own methods and compare your result with the result from the corrcoef function in numpy module. Are they the same? (use the print function to display your answer) (3 pts)

3. Spearman rank correlation (10 pts)

x = np.array([50, 68, 74, 70, 65, 61, 63, 74, 62, **80**])

y = np.array([170, 193, 209, 185, 195, 188, 188, 202, 183, **1000**])

a. Scatter plot of x and y. (2 pts)



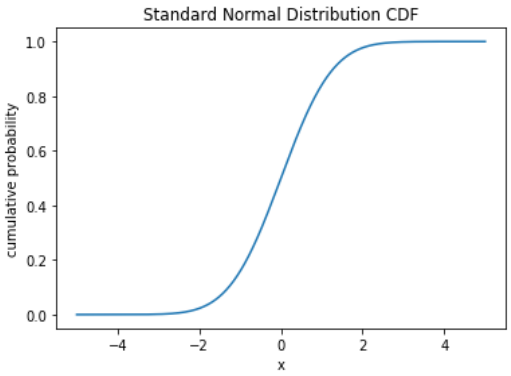
b. Calculate the Pearson correlation coefficient using the corrcoef function. (2 pts)

c. Calculate the Spearman rank correlation coefficient using the corrcoef function. (2 pts)

d. Are the two coefficients the same? Which one is better? Why? (2 pts)
e. What do the values (-1, 0, and 1) of the coefficient indicate correlation? (2 pts)

4. Calculate the following probability in the standard normal distribution and output the questions and results in Python code. (8 pts)

$$P(-0.5 \leq X < 0.5) = \text{CDF}_X(0.5) - \text{CDF}_X(-0.5)$$



a. norm.cdf(0.5) - _____

b. 2*(_____)

5. Properties of normal distribution (12 pts).
(Note: for printing out special characters, e.g. μ, σ, and ∩, you can refer to https://pythonforundergradengineers.com/unicode-characters-in-python.html. You also can copy the special characters from a Word document and paste it into your Python code)
If the distribution of a random variable X is a normal distribution, $X \sim N(\mu, \sigma^2)$.
a. If we have X' = $aX + b$, then what is the distribution if X'? (3 pts)
b. If the distribution of a random variable X is a normal distribution, $X \sim N(\mu, \sigma^2)$. If we convert it into a standard normal distribution, $Z \sim N(0, 1)$. What is the relationship between X and Z? In other words, how do you represent Z using X? (3 pts)
c. Calculate the probability P(2 <= X <= 7) where $X \sim N(5, 9)$ is in the normal distribution and output the major steps and results in Python code. (3 pts)
d. Calculate the probability P(-1.5 <= X <= 1.5) where $X \sim N(0, 1)$ is in the standard normal distribution. (3 pts)

6. Complete the following Probabilistic calculation and output your solution by print function in Python. (8 pts)
a. $P(A \cup B) = P(A) +$_____

b. $P(A|B) =$ _____P(B)

c. $P(A \cap B) =$ P(B)_____

d. If A and B are independent, $P(A \cap B) =$_____

7. Assume d is the tossed number of an 8-face die. Will the probabilities P(d = even) and P(d < 5) be independent? Write down your derivation and intermediate steps and output them in Python code. (8 pts)

8. Use the theorem of total probability and Bayes theorem to solve the following problem and output the major steps in Python code. (10 pts)

   A box of dices: 95% fair, 5% loaded (50% at six). If we get 4 six in a row, what's the chance that the die is loaded?

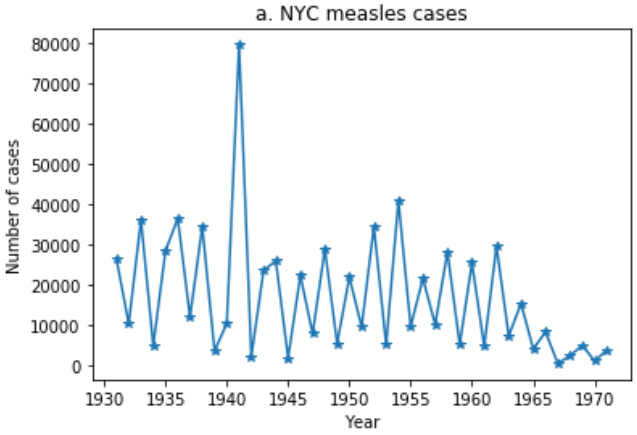
9. Suppose that one person in 10,000 people has a rare genetic disease. There is an excellent test for the disease; 99.9% of people with the disease test positive and only 0.02% who do not have the disease test positive. Output the major steps in Python code (10 points)

   a. What is the probability that someone who tests positive has the genetic disease? (5 pts)

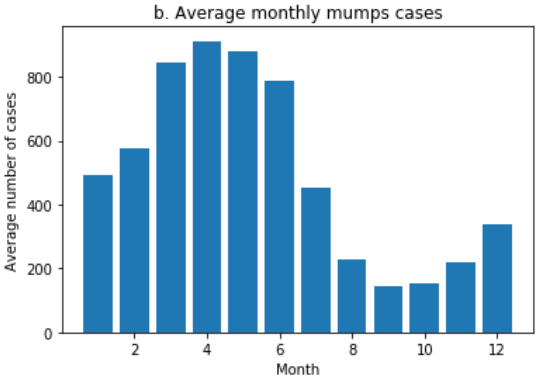
   b. What is the probability that someone who tests negative does not have the disease? (5 pts)

10. To complete this question, you need to download the three csv files, which contain the monthly totals of the number of new cases of measles, mumps, and chicken pox, respectively, for New York City during the years 1931-1971. Each data file contains 41 rows and 13 columns: the first column represents the year (1931 to 1971), and the remaining 12 columns are the number of new cases for each month from January to December. Note that in the chicken pox file, the year is not ordered (i.e., the rows are not chronically ordered), unlike in the other two files. (24 pts)

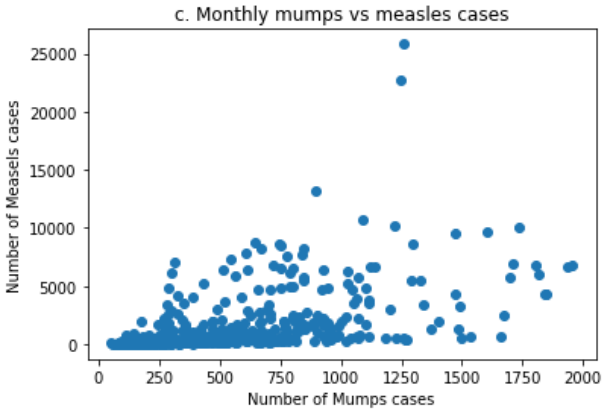
a. Plot the total number of measles cases in each year. (5 pts)



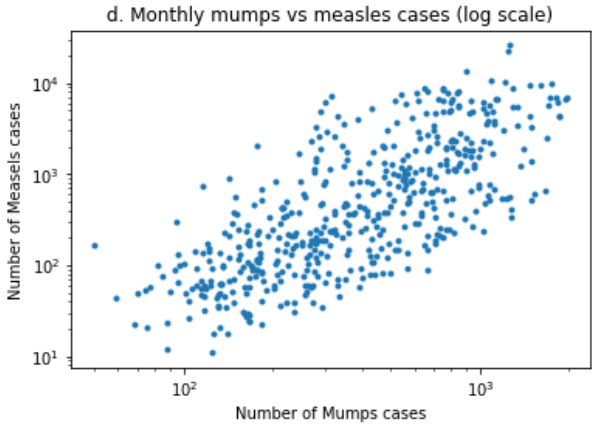
b. Bar plot the average number of mumps cases for each month of the year. (5 pts)



c. Scatter plot the monthly mumps cases against the measles cases. Each dot in the plot represents one month and there is a total of 41 x 12 months. (5pts)



d. Similar to the previous question, but plot both x and y axis in logarithm scale (using loglog function) (5 pts)



e. In some cases, why would we want to plot the y axis in logarithm scale instead of linear scale? (4 pts)