Please complete all the following exercises and submit a pdf file that includes your Python code and relevant discussion in your submission to the Canvas dropbox. This assignment is due February 16th, 2024 at 11:59pm.

Question 1 (10 points)

Data drift is a phenomenon where the process generating input data to a machine learning model changes over time. These changes lead to performance degradation because the model is no longer representative of the process it is emulating. Some causes of data drift are:

- In physical systems, parts of the system are replaced or change condition. For example, a sensor measuring vibrations in a machine may slowly come loose.
- The customer base of a store changes demographics over time.
- Natural disasters such as floods and pandemics significantly alter people's behavior

The sample data provides (question1.csv) provides the number of customers visiting a store and the total sales from that store every day for 400 days. Using this data, perform the following:

a. Load the data and split it into two parts, one for the first 200 days and another for the next 200 days (1 point)
b. Fit a linear regression model linking daily sales to daily customer visits based on data from the first 200 days (1 point)
c. Assume that the rate of customer arrival is constant that arrivals occur independently. Fit a probability distribution for customers arrival in the first 200 days of data. Using this distribution as your null hypothesis, perform a likelihood ratio test using data from the second 200 days. Has the data drifted? (6 points)
d. Evaluate the linear regression model you created in part (b) on the seconds 200 days of data. How does it perform relative to the original data? (2 points)

Question 2

A fabric maker wants to analyze the frequency of yarn breaking in their looms. They have recorded the number of yarn breaks per length of yarn across 9 looms using two different wools, three different tension levels. Because the temperature in the factory changes over time, they also recorded the temperature the test was performed at. They would like to know how these three factors affect their operations. Using the data provided, perform the following:

1. Load the data and format it as needed to create a generalized linear model. (2 point)
2. Create a generalized linear model using a Poisson distribution containing all the independent variables. Examining the results of this model, what can you determine about the cause of yarn breakage? (3 points)
3. Regress breakage against tension and wool separately. Using your favourite regression metric, how do these models compare to the full model? (3 points)

Question 3

A hospital wants to predict the duration of surgeries so that they can create better schedules on the day of surgery. They have recorded the duration of previous surgeries along with some information about the surgery performed. Using the data provided, perform the following:

1. Load the data & calculate surgery duration. For each of the independent variables, group the data by this variable and calculate the average procedure duration. Based on these findings, what predictors do you expect to be meaningful? (5 points)
2. Plot a histogram of surgical duration for all surgeries. Fit at least two distributions of your choice to this data. Which distribution do you think is appropriate? (4 points)
3. Using the features identified in part 1 and the distribution identified in part 2, create a generalized linear model to predict surgical duration. Demonstrate using figures and metrics the additional value of this model over simply using the mean of surgical durations. (5 points)