ECE 3502 / CS 3501: Foundations of Data Analysis **Due Th 3/19**

Homework 3: Linear Regression

Instructions: Submit a single Jupyter notebook (.ipynb) of your work to Collab by 11:59pm on the due date. All code should be written in Python. **Be sure to show all the work involved in deriving your answers! If you just give a final answer without explanation, you may not receive credit for that question.**

You may discuss the concepts with your classmates, but write up the answers entirely on your own. Do not look at another student's answers, do not use answers from the internet, and do not show your answers to anyone.

1. (20%) Answer the following questions about these two vectors (make sure to write out all steps for your solutions, otherwise you may receive partial scores):

$$v = \begin{pmatrix} 15 \\ 34 \\ 18 \\ 22 \end{pmatrix}, \quad w = \begin{pmatrix} 1 \\ 3 \\ 2 \\ 7 \end{pmatrix}$$

   (a) What are the norms $\|v\|$ and $\|w\|$?

   (b) What is the dot product $\langle v, w \rangle$?

   (c) What is the distance between v and w as points?

   (d) What is the projection of v onto w? (looking for a vector here)

2. (80%) In this problem we will be analyzing data from the Old Faithful geyser in Yellowstone National Park. (See `https://en.wikipedia.org/wiki/Old_Faithful` to learn more!) Download the CSV of the data from the class schedule page. The data consists of two variables: the duration of eruptions in minutes (`eruptions` column) and the length of time until the next eruption (`waiting` column). **Hint:** The function `numpy.asarray()` can convert a pandas column into a numpy array.

   (a) Load the `eruptions` column as your `x` variable and the `waiting` column as your `y` variable. Plot the data with a scatter plot. Do you think there is a relationship between eruption time and waiting time?

   (b) What is the mean of the eruption time? What is the mean of the waiting time? Use these values to center your `x` and `y` data.

   (c) Using the dot product formula we discussed in class, compute the correlation of eruption and waiting time. Does the value (and sign) of the correlation match what you would expect from the plot?

   (d) Using the formulas from lecture for $\hat{\alpha}$ and $\hat{\beta}$, compute the intercept and slope for a linear regression. Now plot a scatterplot with your regression line on top of it. How does the value and the sign of the slope compare to the correlation?

   (e) Say you are watching the Old Faithful geyser, and you time an eruption to be 2.2 minutes. Based on your regression analysis, how long should you expect to wait for the next eruption?

(f) Using the formula for the R^2 statistic from class, what is the proportion of variance explained by your regression?