INSTRUCTIONS:

- Please complete the below exercises using Python in a Jupyter notebook.
- For the SQL exercises (#3), include your query and results. You do not need to query this dataset directly from the notebook.

1. Consider data set 1 (ds1.csv). The data set comprises features (the Five xs) along with three sequences that may or may not be generated from the features (3 ys).
    a. Describe the data set in a few sentences. E.g. What are the distributions of each feature? Summary statistics?
    b. Try to come up with a predictive model, e.g. $y = f(x_1 , \dots , x_n)$ for each y sequence. Describe your models and how you came up with them. What (if any) are the predictive variables? How good would you say each of your models is?

2. Consider data set 2 (ds2.csv). The dataset comprises a set of observations that correspond to multiple groups.
    a. Describe the data in a few sentences
    b. How would you visualize this data set?
    c. Can you identify the number of groups in the data and assign each row to its group?
    d. Can you create a good visualization of your groupings?

3. Stack Overflow provides a tool at https://data.stackexchange.com/stackoverflow/query/new that allows SQL queries to be run against their data. After reviewing the database schema provided on their site, please answer the questions below by providing both your answer and the query used to derive it.
    a. How many posts were created in 2017?
    b. What post/question received the most answers?
    c. For posts created in 2020, what were the top 10 tags?