CS 299–Introduction to Data Science, HW2

Submit your source code and write-ups via blackboard. Please include your solutions to all questions in **one single document**. Source code can be separated into multiple files if you want and source code for Q1 is optional. Source code for Q2 and Q3 are required.

1. **Pandas basics (40 pts)**
   Let df be a pandas DataFrame constructed with the following code:

   ```
   data = np.array([0, 7, 3, 6, 2, 8, 5, 9, 4]).reshape(3, 3)
   df = pd.DataFrame(data, index=['One', 'Two', 'Three'], columns=['a', 'b', 'c'])
   ```

   What is the output of the following code? (Try to write the output without using python.)

   a. print(df)

   b. df[‘a’]

   c. df.loc[‘Two’]

   d. df[:2]

   e. df.iloc[:,:]

   f. df.iloc[:,:2]

   g. list(df.columns)

   h. list(df.index)

   i. df[‘b’][‘Two’]

   j. list(df.iloc[2, :])

   k. df.drop('a', axis=1)

   l. df[df.a !=5]

   m. list(df.sum(axis=0))

   n. df.iloc[:, list(df.sum(axis=0) < 17)]

   o. df.sort_values(by='c')

   p. df.sort_values(by='Two', axis=1)

   q. df.T

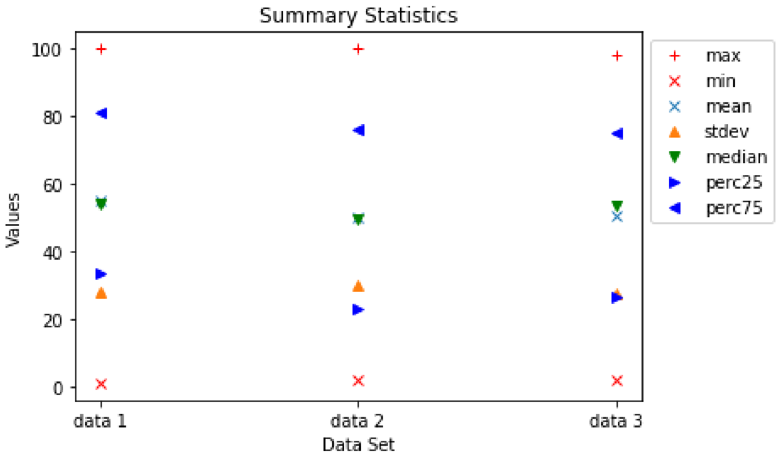
   r. (df<=2).any(axis=0)

   s. df.applymap(lambda x: x*2-1)

   t. df.apply(lambda x: max(x), axis=1)

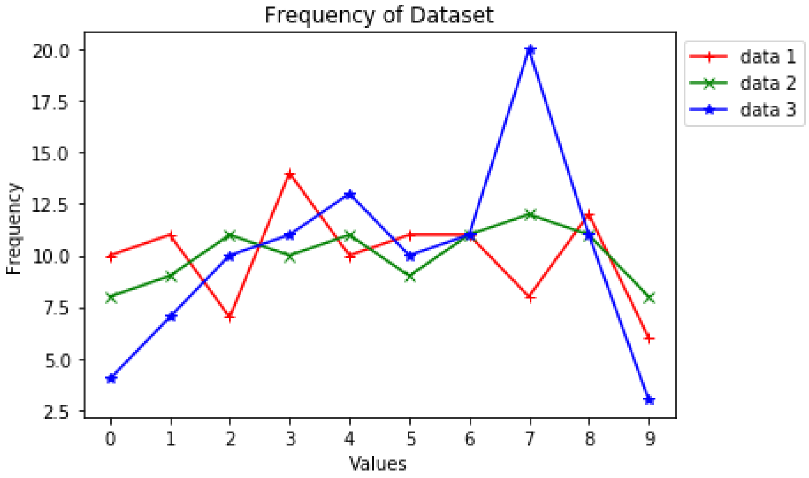
2. (30 pts) Generate three sets of data (at least 100 values in each dataset) using the HW1 part (b), plot the summary statistics in one figure from all three datasets. Try to reproduce all the details of the example figure (see below). You will need to experiment with parameters of plot such as xticks, legend, style as well as set_xlabel() and set_ylabel(). Use the following to move the legend to outside of the plot. (Note: import matplotlib.pyplot as plt)

   ```
   plt.legend(bbox_to_anchor=(1, 1), loc=2)
   ```



3. (30 pts) Use the python function written in HW1 part (c) that accepts a list of numbers in any range, then scales the numbers to [0,1]. Now modify the code to scale the numbers from [0,1] to integers in [0, 9], and then count the number of occurrences of each integer in the dataset. Note: Use collections.Counter() by passing the list. You need to import collections. Test your function on the three datasets used in Question 2 above.

    a. Using the counts you collected, plot the distribution of the three data sets in one figure. Again, try to reproduce all the details of the example below.



What to turn in:

Each file must exactly follow the naming convention: **Lastname-299-A2.zip**
Hw2.py
HW2.pdf (Question 1)
In addition, each .py file must contain the following information at the top:

CS299
HW2
@author: