CS 3753&5163 HW3

Due: Sunday Nov 18, 11:59pm.

Please read: Submit your source code and writeups via blackboard. Please include your answers to all questions in **one single document**. **Label your figures/tables clearly** with xlabels, ylabels, and legend if necessary, and preferably with fig number and caption. e.g. "Fig1. Boxplot for question Q2a."; "Fig2. Boxplot for question Q2b. Y-axis represents log2 transformed data.". (Fig number and caption should be placed underneath the figure and not part of the image.) Source code for Q1 is not required. **Source code for Q2 and Q3 are required**. Name/document your functions appropriately. To make sure that your program can run by the grader, please explicitly import all needed packages instead of depending on the anaconda environment.)

1. **Pandas basics (40 pts)**

   Let df be a pandas DataFrame constructed with the following code:

   ```
   In [62]: data = np.array([0, 7, 3, 6, 2, 8, 5, 9, 4]).reshape(3, -1)
   In [63]: 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)

	a	b	c
One	0	7	3
Two	6	2	8
Three	5	9	4

   b. df['a']

One	0
Two	6
Three	5

   c. df['One']

   KeyError

   d. df.loc['Two']

a	6
b	2
c	8

   e. df[:2]

	a	b	c
One	0	7	3
Two	6	2	8

f. df.iloc[:,:2]

	a	b
One	0	7
Two	6	2
Three	5	9

g. list(df.columns)

['a', 'b', 'c']

h. list(df.index)

['One', 'Two', 'Three']

i. df['b']['Two']

2

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

[5, 9, 4]

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

	b	c
One	7	3
Two	2	8
Three	9	4

l. df[df.a !=5]

	a	b	c
One	0	7	3
Two	6	2	8

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

[11, 18, 15]

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

	a	c
One	0	3
Two	6	8
Three	5	4

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

	a	b	c
One	0	7	3
Three	5	9	4
Two	6	2	8

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

	b	a	c
One	7	0	3
Two	2	6	8
Three	9	5	4

q. df.T

	One	Two	Three
a	0	5	6
b	7	9	2
c	3	8	4

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

a	True
b	True
c	False

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

	a	b	c
One	-1	13	5
Two	11	5	15
Three	9	17	7

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

One	7
Two	8
Three	9

2. Pandas plots, probability models, and simple linear regression. (30 pts)

Use pandas to load hw3q2.csv file into a dataframe called df2, and then do the following.

a. (3 pts) Show a boxplot of the data
b. (3pts) Apply log2 transformation (with applymap and np.log2) to the data and show the boxplot.
c. (3pts) Use pandas function describe() to print out the summary statistics of the data
d. (6pts) Use pandas function hist to show the histogram of each column of the data frame. (Use option normed = True so it plots probability instead of counts.) Decide an appropriate number of bins and whether to apply log transformation on the data for each column.
e. (5 pts) Based on the information and plots you obtained above, what type of probability distribution do you guess they might belong to? (Hint: data in the four columns come from four different distributions we discussed in class: normal, lognormal, exponential, and pareto. See slides lec4.pptx page 28-44.).

A. EXPONENTIAL
B. LOGNORMAL
C. NORMAL
D. PARETO