**CS 5163 HW3**

**Due: Sunday Oct 29, 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.)

* 1. print(df)
  2. df[‘a’]
  3. df[‘One’]
  4. df.loc[‘Two’]
  5. df[:2]
  6. df.iloc[:,:2]
  7. list(df.columns)
  8. list(df.index)
  9. df[‘b’][‘Two’]
  10. list(df.iloc[2, :])
  11. df.drop('a', axis=1)
  12. df[df.a !=5]
  13. list(df.sum(axis=0))
  14. df.iloc[:, list(df.sum(axis=0) < 17)]
  15. df.sort\_values(by='c')
  16. df.sort\_values(by='Two', axis=1)
  17. df.T
  18. (df<=2).any(axis=0)
  19. df.applymap(lambda x: x\*2-1)
  20. df.apply(lambda x: max(x), axis=1)

1. **Pandas plots, probability models, and simple linear regression. (30 pts + 10 pts)**

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

* 1. (3 pts) Show a boxplot of the data
  2. (3pts) Apply log2 transformation (with applymap and np.log2) to the data and show the boxplot.
  3. (3pts) Use pandas function describe() to print out the summary statistics of the data
  4. (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.
  5. (5 pts) Based on the information and plots you obtained above, what type of probability distribution do you think they 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.).
  6. (10 pts) Use the characteristic plot of each probability distribution to prove that your answers in 2e is correct. (Hint: for norm and lognormal, use norm probability plot. For exponential and pareto, plot data against CCDF. See example on slide #28, 36, 41, 43.)
  7. (Bonus 10 points): Enhance your plots above in 2f with the least square linear regression line. Try to fit the data in each column of df2 with each of the four distributions. Present the R-squared measures of the linear regressions in a table (with 4x4 entries) or a figure (e.g. imshow). Does your R-squared show that the distribution you choose is the best fit for the data? Also, in the case of exponential and pareto distribution what does the slope of the regression mean?

1. **Multiple linear regression (30 points)** 
   1. (5 pts) Load data stored in HDF5 format into python using the following statement: hdfstore = pd.HDFStore('hw3q3.h5'). Perform a least square multiple linear regression between the objects x and y in hdfstore (hdfstore[‘x’] and hdfstore[‘y’]). Report the R-squared and Mean Square Error (MSE) of the regression. Plot the coefficients in a bar chart.
   2. (10 pts) Perform bootstrap to estimate the standard error of the coefficients obtained in 3a, and calculate the statistical significance (p-value) of each coefficient (the probability that the coefficient is equal to zero). Plot the -log10(p-value) in a bar chart. (See example in slide #38 and #40.)
   3. (10 pts) Perform lasso regression between x and y using alpha = 2\*\*i, for -6 < i <6. For each value of alpha, compute the R-squared as well as the sum of coefficients. Plot the R-squared, MSE, and the sum of absolute value of the coefficients against the alpha values, in three lines in the same graph. Based on the graph, what is the recommended value(s) of alpha that you should use? What is the R2 and MSE of the fit? Plot the coefficients resulted from the lasso regression with the alpha parameter you choose.
   4. (5 pts) Transform the x matrix by dividing each column with the scaling factor stored in the object hdfstore[‘sf’], and then perform a least square multiple linear regression between the transformed x matrix and the y vector. Report the R-squared and the Mean Square Error of the regression. Use a graph to compare the coefficients from the regression with the expected coefficient stored in hdfstore[‘coef’].