**CS 3753/5163 HW5 (100 points + 30 bonus)**

1. **Pandas plots, probability models. (24 pts)**

Use pandas to load hw5q1.xlsx file into a dataframe. The first line in the file is column header, and there is no row index. (Your data should have 1000 rows and 4 columns.) Do the following.

* 1. (4pts) Show a boxplot of the data.
  2. (4pts) Apply log2 transformation (with applymap and np.log2) to the data and show the boxplot.
  3. (4pts) Use pandas function describe() to print out the summary statistics of the data
  4. (4pts) 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.)
  5. (4pts) Use pandas function hist to show the histogram of the log transformed data frame. (Use option normed = True so it plots probability instead of counts.)
  6. (4 pts) Based on the information and plots you obtained above, which column do you think is from normal distribution and which column is from log normal distribution? (FYI: 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.).

1. **Pandas DataFrame operations (36 points)** 
   1. (6 pts) Load data stored in brfss.csv format into a python DataFrame. The first line in the file is column header, and the first column in the file is row index. Drop the rows that have any NaN values, and drop the 'wtyrago' and 'wtkg2' column. Change the sex column so that 1 (True) means male and 0 (False) means female. (Originally, sex == 2 means female and sex == 1 means male.) Rename the columns to age, weight, height, and male. Name the new DataFrame as brfss.
   2. (30 pts – 3 pts each) Based on the DataFrame brfss, print out answers to the following questions using pandas DataFrame functionality.
      1. What is the max age for people in the dataset?
      2. What is the mean weight for people in the dataset?
      3. What is the mean weight for male in the dataset?
      4. What is the median height for female in the dataset?
      5. What is the mean weight for female younger than 20 years old?
      6. How many males are in the dataset?
      7. How many individuals in the dataset has height > 190cm and weight < 50kg?
      8. What is the average height of females whose weight is between 59 and 61 kg?
      9. Print out row 2001 to row 2010 (inclusive, a total of ten rows) from the dataframe.
      10. Print out rows with rowID from 2001 to 2010 (including 2010, but maybe less than 10 rows due to NaNs being dropped out) from the dataframe.
2. **Simple and multiple linear regression** **(40 points. Useful examples on slides #24, 35, etc.)**
   1. (10 pts) From the brfss DataFrame created in Q2a, use the height column as Y, and weight column as X to perform a simple linear regression (See slide #24, lr.fit). Print out the equation that is obtained by the linear regression in the form of “height = a \* weight + b” (replace a and b with the values obtained from the linear regression, lr.coef\_ and lr.intercept\_.) (Note: you need to either make X a DataFrame or reshape X to be a n x 1 numpy array using reshape(-1, 1)).
   2. (4 pts) Use the linear regression object that you obtained in 3a to predict the height of an individual whose weight is 60kg (with lr.predict). Print out the predicted height.
   3. (6 pts) Compute the MSE and R-square of the linear regression. (Slide #24, Imported as r2 and mse.)
   4. (10 pts) From the brfss DataFrame above, use the height column as Y, and two columns, weight and male, as X, to perform a multiple linear regression (slides #35). Print out the equation that is obtained by the linear regression in the form of “height = a \* weight + b \* male + c” (replace a, b and c with the values obtained from the linear regression.)
   5. (3 pts) Use the linear regression object that you obtained in 3d to predict the height of a male whose weight is 60kg. Print out the predicted height.
   6. (3 pts) Use the linear regression object that you obtained in 3d to predict the height of a female whose weight is 60kg. Print out the predicted height. (Just FYI, how does it compare with the value in Q2b.viii?)
   7. (4 pts) Compute the MSE and R-square of the linear regression (Imported as r2 and mse.)
3. **Multiple linear regression (30 points, optional)** 
   1. (5 pts) Load data stored in HDF5 format into python using the following statement: hdfstore = pd.HDFStore('hw5q4.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’].