Midterm Exam. (100 Points)

Answer all questions. Submit your answers in a single Jupyter Notebook.

1) **Pandas basics (20 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?

a. print(df)
b. df['a']
c. df['One']
d. df.loc['Two']
e. df[:2]
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. **Numpy and vectorized computing (15 points)**

Let x be a numpy array with 4 rows and 4 columns:

```
x = numpy.array(    [[ 1,  2,  3,  4],
                     [ 5,  6,  7,  8],
                     [ 9, 10, 11, 12],
                     [13, 14, 15, 16]])
```

What is the result of the following operations? (Please try to solve them without using a computer and then use python to validate your results.)

a. y = x[:, 2]; print (y)
b. y = x[-1,:2]; print (y)
c. y = x[:, [True, False, False, True]]; print(y)
d. y = x[0:2, 0:2]; print(y)
e. y = x[[0, 1, 2], [0, 1, 2]]; print(y)
f. y = x[0]**2; print(y)
g. y = x.max(axis=1); print(y)
h. y = x[:2, :2]+x[:2,2:]; print(y)
i. y = x[:2, :3].T; print(y)
j. y = x[:2, :3].reshape((3, 2)); print(y)
k. y=x[:, :2].dot([1, 1]); print(y)
l. y = x[:, :2].dot([[3, 0], [0, 2]]); print(y)

3. Data analysis (40 points)

In this problem we will be analyzing the BRFSS weight vs height data. Download data from Zip file. The five columns in the numpy array represent: age, current_weight (kg), weight_a_year_ago (kg), height (cm), and gender, where gender == 1 represents male and 2 represents female. The sixth column wtkg2 is the same as weight2 but rounded to two decimal places.

a. Read and impute missing data, you may also standardize data if necessary. Make two copies of this data frame, you can use the second copy in Question 4.
b. Produce a summary statistics of this data
c. Is your data normally distributed? Explain your answer with distribution plots.
d. Produce a summary statistics graph on current_weight, weight_a_year_ago, and height.
e. Define weight_change = (current_weight – weight_a_year_ago). Calculate correlation between weight_change and the following variables, and determine which one is most correlated (regardless of sign of correlation) with weight_change. Use scatter plot to support your conclusion.
    i. current_weight
    ii. weight_a_year_ago
    iii. age
f. Calculate and compare the mean and SEM (standard error of the mean) for the weight_change of male and female. Use t-test to test whether there is a significant difference between the weight_change of male and female.
g. Randomly split the subjects into two groups of roughly equal sizes, and use t-test to test whether there is a significant difference between the weight_change of the two groups. What can you say about the difference between male and female in terms of their weight_change?

(Consider both the p-value and the absolute differences between the two means.)

h. Define weight_height _ratio as current_weight/height. Use t-test to test whether there is a significant difference between the weight_height_ratio of male and female.
i. Propose and perform your own analysis that utilizes different skills you learned in class (or reveal additional interesting insight from this data set).

4. Regression Analysis (25 points)

Use the BRFSS weight vs. height data set but create a different data frame. Do not use the data frame you used in Question 3.

a. Read and impute missing data, you may also standardize data if necessary (as you did in 3a, copy of the data frame would be a good idea).
b. Randomly Split the data set 70%-30% ratio for training and testing data.
c. Implement a predictive model for this data using Multiple Linear Regression. Your model should be able to predict if the individual is male or female based on any of the given independent variables. You may also use weight_height_ratio as an independent variable but then do not use weight and height separately.
d. Predict the results for your test data set.
e. Output model intercepts, coefficients, Root Mean Squared Error (RMSE) and R^2 values.
f. Write an explanation of your model based on your results if this model is an acceptable predictive model or not.