**ASSIGNMENT 2**

**COMP 4254 FALL 2018**

Submit one file per question in a single zip archive. Each file should be ready to be executed as a script.

The deadline for this assignment is Feb 19 6pm. Late assignments will not be accepted for marking.

You will need to write Python code and use scikit-learn to answer the questions below. You do not need to submit a separate report - when there is a discussion question, please put your answer as a comment in your code.

**Question 1 (6 points)**

Download Carseats.csv file from D2L. This is a simulated dataset containing sales of child car seats at 400 different stores. There are 11 variables in this dataset, with various factors affecting sales.

1. Choose a proper visualization type and plot *Sales* against *CompPrice*, *Price*, *US*, and *Urban* variables. *CompPrice* and *Price* show the price charged by the competitor and the company at each location, respectively. *US* denotes whether the store is in US or not. Similarly, *Urban* denotes whether the store is in an urban or rural location. Notice that not all variables are numeric. Label your axes.
2. Fit a multiple regression model to predict *Sales* using *CompPrice*, *Price*, *Urban*, and *US*. Be careful about the qualitative variables (read the linear regression notebook again if you do not know what this means). Provide an interpretation of each coefficient in the model as a comment in your code.
3. Using **adjusted** R-squared value, find out if a smaller model (i.e. fewer features) can fit your data better. Assuming your features are in dataframe X (where features are the columns), and the values you are trying to predict are in y, and the name of the linear model you fit is lm, here is a one-liner to compute adjusted R-squared:

1 - (1-lm.score(X, y))\*(len(y)-1)/(len(y)-X.shape[1]-1)

If you decide to use fewer variables based on adjusted R-squared, clearly comment in your code the variables you picked.

1. Assuming the company is in US, and the price is 120, predict the sales by using *CompPrice* values that vary between 50 and 200. range function or linspace function from numpy should come in handy. Plot *CompPrice* (x-axis) vs Predicted Sales (y-axis) to show your answers. Label your axes.

**Question 2 (4 points)**

Download the Adults.csv file from D2L. This data was extracted from the census bureau database and it contains various demographic characteristics and salaries of a sample population of the US.

1. Using train\_test\_split function of scikit-learn, divide your dataset into two parts where 80% of the observations are in the training data.
2. Train a k-nearest neighbor classifier **on the training data** where the features are education-num, hours-per-week, capital-gain, capital-loss and the label is salary. Print the accuracy of your model **on the test data**. Do this for the following K values: 3, 5, and 7.
3. Can you add occupation as a feature to be used in a k-nearest neighbor classifier? Why/why not?