In this assignment, you will have an opportunity to apply the "linear" methods presented in class and in chapters 14, 15, 16, as well as related items from chapters 8, 10 and 11. You may use our author's code from those chapters, code that I presented in class (in various files in the Files Section on Canvas), or Python libraries/packages of your choice to answer the following unless otherwise specified. As usual, you should include your code in your submission.

In the files section you will find a .zip file (FilesForAssignment7.zip) that contains some data and a Python script that are referred to in the following questions. Please download and use the contents of that zip file. Other code that I presented in class is contained in other files in the Files Section.

Please archive your responses and associated source code into a single .zip archive, and upload that file as your submission for this exercise. Your archive should include an MS-Word or PDF file containing your responses to individual questions.

1.

In the assignment files, you will find automotive data from 1993, courtesy of the American Statistical Association. I have included a link to the original source, the original data file, and the original metadata file. For your convenience, I have also converted this file to CSV format. Use this data to answer the following:

a.

Use Simple Linear Regression to show an "interesting" relationship between two variables presented in this data. Present this relationship using an equation, a plot, the sum of squared errors, the r-squared value and Pearson's Correlation Coefficient. By "interesting", I mean that you should look at the relationship between two items that are not obviously-related. For instance, min-, mid- and max-price data are all "obviously" related. We expect these things to be strongly related. Similarly, we should expect CityMPG and HighwayMPG to be strongly related. Examine a not-so-obvious relationship.

b.

Using the same dependent variable from Exercise 1a, select eight or more independent variables including the one independent variable used in Exercise 1a.

Use Multiple Linear Regression to predict the value of the dependent variable from the independent variables. Present this relationship using an equation, the sum of squared errors, and the r-squared value. This model should perform better than the model produced in Exercise 1a. Explain if it did or did not and why.

c.

Use Lasso Regression to eliminate some of the variables from the model found in Exercise 1b. Present this relationship using an equation, the sum of squared errors, and the r-squared value. Explain how this result compares to the one found in Exercise 1b.

d.

Use Logistic Regression and Horsepower (independent variable) to create a model to predict if the type of car is "Compact" (dependent variable) or not. Present this relationship using an equation. Also present measures of performance including TP, TN, FP, FN, accuracy, precision, recall and F1 measures for your model. How well does this model perform?

e.

Enhance your Logistic Regression model from Exercise 1d by including three more numeric fields in addition to Horsepower as the independent variables and again predict if the car is "Compact" or not. Present this relationship using an equation. Also present measures of performance including TP, TN, FP, FN, accuracy, precision, recall and F1 measures. How well does this model perform as compared to the model from Exercise 1d?

f.

Select 8 or more numeric variables from the cars data set. Use Principal Components Analysis to determine the first three principal components relative to the 8 (or more) numeric variables you selected. (You will want to normalize/rescale your data before computing the Principal Components.) Present these three principal components idividually as linear combinations of the 8 (or more) variables that you selected. For each principal component, indicate which types of vehicles will receive large positive values versus the types of vehicles that will receive large negative values. For instance, if the linear combination for one of the principal components assigns large, positive coefficients to measures of vehicle size and large, negative coefficients to measures of fuel economy (MPG), then that principal component likely distinguishes large vehicles from compact/economy vehicles.

2.

In the assignment files, you will find a Python script named "QuadraticRegressionExample.py". This relatively self-contained script contains functions that compute the optimal quadratic mapping between two variables. One method uses matrices and the other uses gradient descent. The output from both of these methods for the exact same data should be the same. But, as you will see, the output is quite different. Run the script, QuadraticRegressionExample.py, to see this for yourself. Examine the contents of the script. Note that what is contained is roughly equivalent to what I presented in class, although cleaned up a bit to make the file self-contained. Please explain why the results are different for the two methods.

3.

For **Graduate Students Only. Undergraduates to not have to do this problem!** Coordinate Descent is a "derivative-free" alternative to Gradient Descent. It does not require that derivatives be supplied or even approximated. It iteratively uses line searches to optimize relative to individual coordinates. Find further information about Coordinate Descent and implement a simple, derivative free version of Coordinate Descent in Python (this must be your original code) in a manner similar to what was done in our book and by me for batch gradient descent. In turn, demonstrate your implementation by computing the optimal quadratic fit between the two variables used in Exercise 2. How does your result compare to the results of the two methods from Exercise 2. How does the computational (time) performance compare? In general, how useful do you think this method is compared to gradient descent?