**Assignment 2**

**Due Friday, February 16, 2018**

*Submission: Put the data and Jupyter notebook files in a folder. Make sure all links to data are relative to the folder so the TAs can run the notebooks.* Points will be deducted if the TAs can't run the notebooks.

Use the public dataset that you used for cleaning and EDA in Assignment 1. You *MUST* get approval if you wish to use a *different* dataset.

In this assingment you will cluster your data and create predictive linear and logistic models.

* Cluster your data:
  + Use at least two methods to cluster your data. (25 points)

Answer the following questions for the clustering:

\* Do the clustering methods generate the same clusters?

\* Does scaling effect the clustering?

\* Does the clustering produce interesting groupings?

* Generate a linear model for your data:
  + Find a significant linear relation of your choosing in your data. Create a multivariate linear model. (50 points)

Answer the following questions for the multivariate linear model:

\* Is the relationship significant?

\* Are any model assumptions violated?

\* Is there any multi-colinearity in the model?

\* In the multiple regression models are predictor variables independent of all the other predictor variables?

\* In in multiple regression models rank the most significant predictor variables and exclude insignificant ones from the model.

\* Does the model make sense?

\* Cross-validate the model. How well did it do?

\* Does regularization help with creating models that validate better on out of sample data?

* Generate a logistic model for your data:
  + Find a significant logistic linear model of your choosing in your data. Create a logistic linear model. (25 points)

Answer the following questions for the logistic linear model:

\* Is the relationship significant?

\* Are any model assumptions violated?

\* Cross-validate the model. How well did it do?