# Introduction to Data Science

Instructor: Daniel D. Gutierrez

# HOMEWORK 3

**Question 1**

Using the `Auto` data set found in the `ISLR` package, perform the tasks below using the supervised machine learning algorithm `lm()` for simple linear regression:

1. Use the `lm()` function to perform a simple linear regression with `mpg` as the response variable and `horsepower` as the predictor.
    i. Use the `summary()` function to print the results.
    ii. Comment on the output of `summary()`, for example: is there a relationship between the predictor and the response variable? If so, how strong is the relationship? Is the relationship positive or negative?
    iii. What is the predicted `mpg` associated with a `horsepower` of 98?
2. Plot the response variable and predictor. In addition, use the `abline()` function to display the least square regression line.
3. Use the `plot()` function to produce diagnostic plots of the fit. Comment on any problems you see with the fit.

**Question 2**

Using the `Auto` data set, perform the tasks below using the supervised machine learning algorithm `glm()` for logistic regression. Develop a model to predict whether a given car gets high or low gas mileage:

1. Create a binary categorical variable `mpg01` that contains a 1 if `mpg` contains a value above its median, and a 0 if `mpg` contains a value below its median. You can use the `median()` function to calculate the median.

Create a new data frame containing all the variables from `Auto` plus the new `mpg01` variable.

2. Explore the data using EDA techniques in order to investigate the association between `mpg01` and the other features. Which of the other features seem most likely to be useful in predicting `mpg01`? Scatterplots and boxplots may be useful tools to answer this question.
3. Split the data into a training set and test set.
4. Perform logistic regression on the training set in order to predict `mpg01` using the variables that seemed most associated with `mpg01` above. What is the test error of the model obtained?

**Question 3**

Use the K-means clustering algorithm on the `iris` data set for the `Sepal.Length` and `Sepal.Width` variables. Perform the following steps:

1. Set the number of centroids to 3
2. Call the `kmeans()` algorithm and store the resulting `kmeans` class object to a variable named `kc`. You need to set seed to get reproducible results because `kmeans()` uses a random number generator to come up with the centers if you use the `centers` argument.
3. Review and print the `cluster` component of the `kmeans` object.
4. Review and print the `centers` component of the `kmeans` object.
5. Produce a data visualization to plot each of the resulting clusters of data points and their centers. Use different colors for the data points residing in each cluster. Also, plot a special character showing the centroid of each cluster.