**National University of Singapore**
**School of Computing**

# SWS3018 Predictive Analytics
# Lab 3

## Learning Objectives

- Perform linear regression analysis using R

1. In this exercise, we will be using the `Auto` (`Auto.csv`) data set which can be downloaded from IVLE. This data set contains 397 different car observations. The variables are:

   - **mpg** : Miles per gallon
   - **cylinders** : Number of cylinder for this car
   - **displacement** : Engine displacement (in cubic inches)
   - **horsepower** : Horsepower of car
   - **weight** : Weight of car (pound)
   - **acceleration** : Time to accelerate from 0 to 60 (in secs)
   - **year** : Model year (1900s)
   - **origin** : Country of origin (1 = American, 2 = European, 3 = Japanese)
   - **name** : Car name

   a) There are some missing values in the csv file (denoted using the "?" character). Read the help pages for `read.csv` to see how to convert the missing values to be NA and ignore these values in the analysis.

   b) Use the `lm()` function to perform a simple linear regression using `mpg` as the response and `horsepower` as the predictor. Use the `summary()` function to print the results of the model. Comment on the output.

   c) Generate the scatter plot of the response and the predictor.

   d) Use the `abline()` function to display the regression linear on the scatter plot in (c)

   e) By looking at the scatter plot, it appears that the relationship between the predictor and the response is not linear. Try executing `plot(model)` and look at the first diagnostic plot of the least squares regression fit. If the relationship is linear, the **Residuals vs Fitted** plot should show no strong pattern.

   f) You can try a non-linear model by executing this:

```
model2 <- lm(mpg~horsepower + I(horsepower^2), data=auto)
```

   g) The curve can be added to the scatter plot by executing the following:

```
xrange <-
seq(from=min(auto$horsepower),to=max(auto$horsepower))
```

```
lines(xrange,  predict(model2,data.frame(horsepower=xrange)),
col="red")
```

h) Try execute `plot(model2)` and look at the **Residuals vs Fitted** plot. What is the difference you see now? How about the results from `summary()`?

i) Compute the matrix of correlations between the variables using the `cor()` function. You will need to exclude the `name` variable which is not numeric.

j) Use the `lm()` function to perform a multiple linear regression using `mpg` as the response and the other variables (except `name`) as the predictors. Use the `summary()` function to print the results of the model. Comment on the output.