Exercise 1

```
##############   The Validation Set Approach   ##########################

library(ISLR)

## The Auto data set is contained in ISLR package

str(Auto) # inspect it
attach(Auto) # when a dataset is attached, its objects can be accessed by simply giving their names.


## Define training and test data

set.seed(1)
(train=sample(nrow(Auto),nrow(Auto)/2)) # generate indices of a training data set by random sampling half of the observation indices of Auto
AutoTrain <- Auto[train,] # training data
AutoTest <- Auto[-train,] # test data

## Fit 3 different regression models to the training data: Which model gives a better fit?

(lm.fit = lm(mpg~horsepower,data=AutoTrain)) # linear regression
(lm.fit2 = lm(mpg~poly(horsepower,2),data=AutoTrain)) # quadratic regression
(lm.fit3=lm(mpg~poly(horsepower,3),data=Auto,subset=train)) # cubic regression

## Plot the fits

plot(mpg~horsepower,data=AutoTrain) # training data

abline(lm.fit$coefficients, col="blue4") # plot linear fit

x <- with(Auto, seq(min(horsepower), max(horsepower), length.out=2000)) # define x values to plot polynomial fits

y2 <- predict(lm.fit2, newdata = data.frame(horsepower = x)) # corresponding predicted values
lines(x, y2, col = "red4") # quadratic fit

y3 <- predict(lm.fit3, newdata = data.frame(horsepower = x)) # corresponding predicted values
lines(x, y3, col = "green4") # cubic fit
```

- For this exercises, we use the Auto data set, which is included in ISLR. It is about cars and their properties.
- Looking at the plot, which model do you think will evaluate best?

Exercise 1

- Now we evaluate the 3 models by calculating their MSEs: Did you have the right hunch?
- Note the results on a piece of paper.

```
## Calculate the test MSEs of the 3 models (mean of squared errors on test set)

(MSE <- mean((mpg-predict(lm.fit,Auto))[-train]^2))
(MSE2 <- mean((mpg-predict(lm.fit2,Auto))[-train]^2))
(MSE3 <- mean((mpg-predict(lm.fit3,Auto))[-train]^2))
```

Exercise 1

- Now we fit and evaluate the 3 models on a different training data set. To do this, we just need to set a different seed for the random sampling with sample when we generate the indices for the training data set.
- Compare the results to the results of your first run: Is the same model winning?
- Repeat the procedure for a few more seeds. Write down the corrsponding MSEs and compare. How similar / different are the results?

```
## Train & evaluate the 3 models on a different training data set (i.e., set a different seed)

set.seed(2)
(train=sample(nrow(Auto),nrow(Auto)/2)) # generate indices of a training data set by random sampling half of the observation indices of Auto
AutoTrain <- Auto[train,] # training data
AutoTest <- Auto[-train,] # test data

(lm.fit = lm(mpg~horsepower,data=AutoTrain)) # linear regression
(lm.fit2 = lm(mpg~poly(horsepower,2),data=AutoTrain)) # quadratic regression
(lm.fit3=lm(mpg~poly(horsepower,3),data=Auto,subset=train)) # cubic regression

(MSE <- mean((mpg-predict(lm.fit,Auto))[-train]^2))
(MSE2 <- mean((mpg-predict(lm.fit2,Auto))[-train]^2))
(MSE3 <- mean((mpg-predict(lm.fit3,Auto))[-train]^2))
```