BDA/IDAR Coursework 1

- Please submit TWO files to Dropbox 1 on moodle: a .rmd file and knit it to one of the following (.doc/.pdf/.html) files. Please include any R code, plots or results.
- Your files should be named as follows: CW1_xxxxxxxx_initial_lastname.rmd (.pdf/.html/.doc), where xxxxxxxx is your student id. For instance, CW1_12345678_T_Han.rmd.
- Don't forget to write down your programme (MSc or BSc), name and student id in your files as well.
- Each question below has two weightings. The first weighting is for MSc students and the second weighting is for BSc students. For instance, (10% | 0%) means that the question is worth 10% for MSc students and 0% for BSc students (optional).

1. Statistical learning methods **(10% | 20%)**

For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(a) The sample size n is extremely large, and the number of predictors p is small.

(b) The number of predictors p is extremely large, and the number of observations n is small.

(c) The relationship between the predictors and response is highly non-linear.

(d) The variance of the error terms, i.e. $\sigma^2 = \text{Var}(\epsilon)$, is extremely high.

2. Descriptive analysis **(10% | 20%)**

In a higher educational institution the comprehensive applied mathematics exam is comprised of two parts. On the first day, 20 students took the exam, the results of which are presented below:

Oral exam results:	4, 1, 4, 5, 3, 2, 3, 4, 3, 5, 2, 2, 4, 3, 5, 5, 1, 1, 1, 2.
Written exam results:	2, 3, 1, 4, 2, 5, 3, 1, 2, 1, 2, 2, 1, 1, 2, 3, 1, 2, 3, 4.

(a) Use R to calculate the mean, the mode, the median, the variance and the standard deviation of the oral and written exams separately and together as well.

(b) Find the covariance and correlation between the oral and written exam scores.

(c) Is there a positive or negative or no correlation between the two?

(d) Is there causation between the two? Justify your answers.

3. Descriptive analysis **(10% | 0%)**

This exercise involves the `Auto` data set studied in the class. Make sure that the missing values have been removed from the data.

(a) Which of the predictors are quantitative, and which are qualitative?

(b) What is the range of each quantitative predictor? You can answer this using the `range()` function.

(c) What is the mean and standard deviation of each quantitative predictor?

(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

(f) Suppose that we wish to predict gas mileage (`mpg`) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting `mpg`? Justify your answer.

4. Linear regression (20% | 20%)

This question involves the use of simple linear regression on the `Auto` data set.

(a) Use the `lm()` function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the `summary()` function to print the results. Comment on the output. For example:

i. Is there a relationship between the predictor and the response?

ii. How strong is the relationship between the predictor and the response?

iii. Is the relationship between the predictor and the response positive or negative?

iv. What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?

(b) Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.

(c) Plot the 95% confidence interval and prediction interval in the same plot as (b) using different colours and legends.

5. Logistic regression (10% | 20%)

Using the Boston data set, fit classification models in order to predict whether a given suburb has a crime rate above or below the median. Explore logistic regression models using various subsets of the predictors. Describe your findings.

6. Resampling methods (20% | 0%)

Suppose that we use some statistical learning method to make a prediction for the response Y for a particular value of the predictor X. Carefully describe how we might estimate the standard deviation of our prediction.

7. Resampling methods (20% | 20%)

We will now perform cross-validation on a simulated data set.

(a) Generate a simulated data set as follows:

```
set.seed(500)
y = rnorm(500)
x = 4 - rnorm(500)
y = x - 2*x^2 + 3*x^4 + rnorm(500)
```

In this data set, what is n and what is p? Write out the model used to generate the data in equation form.

(b) Create a scatterplot of X against Y. Comment on what you find.

(c) Set the seed to be 23, and then compute the LOOCV and 10-fold CV errors that result from fitting the following four models using least squares:

i. $Y = \beta_0 + \beta_1 X + \epsilon$

ii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$

iii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$
iv. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$.

Note you may find it helpful to use the `data.frame()` function to create a single data set containing both X and Y.

(d) Repeat (c) using random seed 46, and report your results. Are your results the same as what you got in (c)? Why?

(e) Which of the models in (c) had the smallest LOOCV and 10-fold CV error? Is this what you expected? Explain your answer.

(f) Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in (c) using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?