# DS 3010 Homework 1

**Instructions:** Homework is to be submitted on Canvas by the deadline stated on Canvas. Please clearly print **your name** on your HW.

To receive full credit, **show all necessary work**, including calculations. Please make your submission as **clear and readable** as possible. Do **not** include raw R output or R code in your written submission **unless explicitly requested or necessary for clarity**.

**Option 1: R Markdown (single file)**

- You may submit **one R Markdown file** (.Rmd) and its rendered output (.html or .pdf).
- If you use R Markdown, **you do NOT need to submit a separate code file**.

**Option 2: Separate files**

- Submit a **written file** (Word or PDF or picture of hand-written answer) containing your answers and calculations.
- Submit a **separate file** containing your R code (.R, .txt, or .Rmd).

**Clearly label sections of the code using comments**, for example:

##### Problem 1 #####

**Grading Note**

**Extra credit:** Questions marked with **(practice)** are included to support your learning and provide more flexibility. These questions are graded for **completion** rather than correctness. A substantive, good-faith attempt on all practice questions will earn **10% completion credit**, added to the homework total.

**Problem 1: R review**

Download the **Auto.data** from Module 'Introduction to DS 3010'.

(a) How many rows and columns are in this dataset? Note that in supervised learning, we refer to the rows in the dataset as 'observations' and columns as 'variables' (or predictors/response).

(b) Remove any observations with missing values in the dataset. Now how many rows and columns are in this dataset?

(c) Which columns are stored as factors? Define what factors are in R.

(d) Should the variable name be encoded as a factor? Justify why or why not.

(e) Print a snapshot of the first 10 observations in the dataset. Which R command did you use?

(f) Extract observations 10, 14, and 29 using one line of code. Print your code/output here.

(g) Extract the displacement and horsepower values for observations 10, 14, and 29 using one line of code. Print your code/output here.

(h) Find the average mpg for all observations in the dataset that have a horsepower less than 200. Can you compute that using one line of code?

(i) What kind of plot would be appropriate to examine the relationship between mpg and horsepower? Show that plot here.

(j) What kind of plot should be appropriate to examine the relationship between year and acceleration? Show that plot here. Note you should not be using the same type of plot as the previous question.

(k) Many functions in R are vectorized. Explain in plain language (to someone with no coding background) what that means.

**Problem 2: Multiple linear regression**

For this problem, we will use the Boston data set which is part of the ISLR2 package. To access the data set, install the ISLR2 package and load it into your R session:

```
install.packages("ISLR2") #you only need to do this one time.
library(ISLR2) #you will need to do this every time you open a new R session.
```

To get a snapshot of the data, run **`head(Boston)`**.

To find out more about the data set, we can type **`?Boston`**.

We will now try to predict per capita crime rate using the other variables in this data set. In other words, ***`per capita crime rate`*** is the response (Y ), and the other variables are the predictors (X).

(a) How many rows (n) are in the data set? How many columns are in the data set?

(b) What does the variable ***`lstat`*** represent? (Hint: check ?Boston)

(c) Obtain the average per capita crime rate across all suburbs in the data set. Report that here.

(d) Obtain the average crime rate only for those suburbs who are not near the Charles river (chas *==0) and those suburbs who are near the Charles river (chas ==1).* Report both values here. Is it safer to be near or away from the Charles river?

(e) Do any of the suburbs of Boston appear to have particularly high crime rates? Define what a 'high' crime rate is and provide some summary statistics on the crime rate.

(f) Are any of the other predictors in the data set associated with per capita crime rate? Use your exploratory data analysis skills from DS 2020 to uncover insights. Describe your findings.

(g) Fit a simple linear regression model with **`crim`** as the response and **`lstat`** as the predictor. Describe your results. What are the estimated coefficients from this model? Report them here.
Note: a simple linear regression is just a regression model with a single predictor.

(h) **(practice)** Repeat part (g) for each predictor in the dataset. That means for each predictor, fit a simple linear regression model to predict the response. Describe your results and organize them in a table. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

(i) **(practice)** Fit a multiple regression model to predict the response using all of the predictors. Summarize your results neatly in a table.

(j) **(practice)** How do your results from (h) compare to your results from (i) Create a plot comparing the simple linear regression coefficients from (h) to the multiple regression coefficients from (i). Describe what you observe.

(k) **(practice)** Explain why your results from (j) provide evidence that using many simple linear regression models is not sufficient compared to a multiple linear regression model. What information does a multiple linear regression model capture that many simple linear regression models cannot capture?

### Problem 3: Interpreting Multiple Linear Regression Models

*Suppose we have a data set with five predictors, $X_1$ = GPA, $X_2$ = IQ, and $X_3$ = Level (1 for College and 0 for High School). The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit a multiple linear regression model on our data set and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$ and $\hat{\beta}_3 = 35$.*

(a) Which answer is correct, and why?

i. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.

ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.

iii. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.

iv. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.

(b) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

(c) True or false: Since the coefficient of IQ is very small, the effect of IQ effect on salary is not very important. Justify your answer.