Predicting Course Grades Using Midterm Scores
Part I: Predicting Final Exam Scores

In this exercise, we will analyze real, historical, student performance data from the class. The goal is to model the relationship between midterm and final exam scores so that we can later predict final exam scores based on midterm scores. The dataset we will use is in the *grades.csv* file. Table 1 shows the names and descriptions of the variables in this dataset, where the unit of observation is students.

variable	description
midterm	students' scores in the midterm (from 0 to 100 points)
final	students' scores in the final exam (from 0 to 100 points)
overall	students' scores in the class overall (from 0 to 100 points)
gradeA	identifies students who earned an A or an A minus in the class

Table 1: Variables in "grades.csv"

In this problem set, we practice fitting a line to make predictions when Y is non-binary, including computing correlations, creating scatter plots, adding the fitted line to the plot, and computing R^2.

As always, we start by loading and looking at the data:

```
## load and look at the data
grades <- read.csv("grades.csv") # reads and stores data
head(grades) # shows first observations
##   midterm final overall gradeA
## 1   79.25 47.00    69.2      0
## 2   96.25 87.75    94.3      1
## 3   58.25 37.75    62.0      0
## 4   54.50 62.00    72.4      0
## 5   83.00 39.75    72.4      0
## 6   41.75 49.50    59.5      0
```

1. First, let's figure out what each observation represents, identify our X and Y variables, and explore whether they are moderately or strongly linearly associated with each other.

   a. In this dataset, what does each observation represent? (2.5 points)

   b. What should be our X variable? In other words, which variable are we going to use as the predictor? Please provide the name of the variable and identify whether it is binary or non-binary. (2.5 points)

   c. What should be our Y variable? In other words, which variable are we going to use as the outcome variable? Please provide the name of the variable and identify whether it is binary or non-binary. (2.5 points)

d. Compute the correlation coefficient between X and Y. Is the relationship between X and Y moderately or strongly linear? A yes/no answer will suffice. (2.5 points)

2. Second, let's fit the linear model that we will use to make predictions.

   a. Use the function lm() to fit a linear model to summarize the relationship between X and Y and store the output in an object called *fit*. Then, ask R to provide the contents of *fit* by running its name. (R code only.) (5 points)

   b. What is the fitted line? In other words, provide the formula $\widehat{Y} = \widehat{\alpha} + \widehat{\beta}X$ where you specify each term (i.e., substitute Y for the name of the outcome variable, substitute $\widehat{\alpha}$ for the estimated value of the intercept coefficient, substitute $\widehat{\beta}$ for the estimated value of the slope coefficient, and substitute X for the name of the predictor.) (5 points)

   c. Create a visualization of the relationship between X and Y and add the fitted line to the graph using the function abline(). (R code only.). (5 points)

3. Now, let's use the fitted line to make some predictions.

   a. Computing $\widehat{Y}$ based on X: Suppose that you earn 80 points in the midterm. What would be your best guess of your predicted final exam score based on your performance in the midterm? Please show your calculations and then answer the question with a full sentence (including units of measurement). (5 points)

   b. Computing $\widehat{Y}$ based on X: Now, suppose that you earn 90 points in the midterm. What would be your best guess of your predicted final exam score based on your performance in the midterm? Please show your calculations and then answer the question with a full sentence (including units of measurement). (5 points)

   c. Computing $\triangle\widehat{Y}$ based on $\triangle X$: Suppose that you study a few extra hours and you earn 10 *extra* points in the midterm as a result. What would be your best guess of by how much your predicted final exam score will *change* as a result of these 10 *extra* points in the midterm? Please show your calculations and then answer the question with a full sentence (including units of measurement). (10 points)

4. What is the R^2 of the fitted model? And, how would you interpret it? (Hint: the function cor() might be helpful here.) (5 points)