# Homework 2

**Due**: Tuesday 10/15, 11:55pm on Titanium. Prepare your answers as a single PDF file.
**Group work**: You may work in groups of 1-3. Include all group member names in the PDF file. Only one person in the group needs to submit to Titanium.

1. Consider the "Auto MPG" which "concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes."[1] The goal is to model mpg given engine displacement and number of cylinders. Answer the following questions.

   a. Load the `autompg.csv` file on Titanium and convert cylinders variable to a factor. (code, output of str())
   b. Which is the dependent variable? Which are the independent variables?
   c. Plot `mpg` vs. `displacement` (code, plot)
   d. Create a linear model of mpg vs. displacement (only one independent variable). What is the $R^2$? (code, output of summary(model), R2 value)
   e. Create a new transformed variable that is sqrt(displacement). Create a linear model of mpg vs. log(displacement).
      i. Give R code, output of summary(model)
      ii. Is this a better fit than in part (d)?
      iii. Plot `mpg` vs. `sqrt(displacement)` and overlay the best fit model as a straight line. (code, plot)
      iv. Plot `mpg` vs. `displacement` and overlay the best fit model as a curve. (code, plot) [Hint: plot the predictions; use `add_predictions()` and `geom_line()`. You don't have to use `data_grid()`]

   f. Create a linear model of `mpg` vs. `sqrt(displacement)` and `cylinders`.
      i. Give R code, output of summary(model)
      ii. How many dummy (i.e., 0-1) variables were created in the model?
      iii. Is this a better fit than in part (e)?
      iv. Plot `mpg` vs. `sqrt(displacement)` and overlay the **multiple** linear fit lines: one for each value of the discrete variable. (code, plot)
      v. Plot `mpg` vs. `displacement` and overlay the best fit model as a curve. (code, plot) [Hint: plot the predictions; use `add_predictions()` and `geom_line()` and use the color aesthetic for cylinders]

[1] Dataset modified from UC Irvine ML repository https://archive.ics.uci.edu/ml/datasets/Auto+MPG. This is *not* the mpg data that is part of the tidyverse datasets.

2. Consider the toy dataset below which shows if 4 subjects have diabetes or not, along with two diagnostic measurements. [This question is meant to be completed with a calculator; no need to write any R code.]

| **Preg** | **BP** | **HasDiabetes** | **Preg.Norm** | **BP.Norm** |
|---|---|---|---|---|
| 2 | 74 | No | | |
| 3 | 58 | Yes | | |
| 2 | 58 | Yes | | |
| 1 | 54 | No | | |
| 2 | 70 | ? | | |

a. Which variable is the “Class” variable?
b. Normalize the Preg and BP values by scaling the minimum-maximum range of each column to 0-1. Fill in the empty columns in the table.
c. Predict whether a subject with Preg=2, BP=70 will have diabetes using the 1-NN algorithm and
   i. Using Euclidean distance on the original variables:
   ii. Using Manhattan distance on the original variables:
   iii. Using Euclidean distance on the normalized variables:
   iv. Using Manhattan distance on the normalized variables:

3. The `data_banknote_authentication.csv` file attached on Titanium contains instances of genuine and forged banknotes. The first four columns are features calculated from an industrial camera[2] ; the fifth column indicates if the banknote is forged or not. The goal is to see if it is possible to detect a forgery from only the features.

a. Load and pre-process the data. Show code to:
   i. Load the data file on Titanium.
   ii. How many rows and columns are there?
b. Split the dataset into train and test datasets with the *rows 1, 3, 5, ...* for training, and the remaining rows for test (i.e, test using rows 2, 4, 6, …). Do **NOT** randomly sample the data (though resampling is usually done, this hw problem does not use this step for ease of grading). (code)
c. Train and test a k-nearest neighbor classifier with the above datasets. *Consider only variance and skewness columns*. Set k=1. What is the error rate (number of misclassifications)? (code)

[2] For further information, refer to: https://archive.ics.uci.edu/ml/datasets/banknote+authentication

d. Repeat part (c) but *consider only variance , skewness, and curtosis columns*. Set k=1. (show code.) What is the error rate? Will the error rate always decrease with larger number of parameters? Why or why not: answer in 2-3 sentences?

e. Repeat part (d) but set k=5. What is the error rate?

f. Repeat part (e) but set k=11. What is the error rate? Considering your observations from (d)-(f), which is the best value for k?

g. Consider only the ranges of the features - is normalization required?

h. Normalize each column by scaling the minimum-maximum range of each column to 0-1. (Hint: the built-in R function `scale()` can be used for this) (code)

i. Train and test a k-nearest neighbor classifier with the normalized dataset. *Consider only variance, skewness, and curtosis columns*. Set k=1. What is the error rate?