HW#1 (Due February 3rd)

(PLEASE USE BLUE/BLACK PENS OR/AND PRINT FROM COMPUTER ON A BLANK SHEET OF PRINTER PAPER FOR ALL FINAL WORK.)

YOU MUST SHOW ALL WORK WHERE APPLICABLE

Note: LATE HW is not accepted!

1) (a) Using the 'titanic_data' dataset, state code that can be used to create a vector called 'FullName' which contains the first-names and last-names of all the passengers on the titanic. Note that first name should include all nicknames, etc. within the dataset (basically everything after the title).
For example, your code should turn "Braund, Mr. Owen Harris" into "Owen Harris Braund".

Hint: In order to search for the location of a period "." in a string you must use two backslashes first as in the following: regexpr("\\.", "string with a . in it"). This code for example would output the number 15 as the first value. Furthermore the function nchar("string with a . in it") would output the number of characters in a string. This code for example would output 21.

(b) State code which could be used to output the names of all the survivors in the dataset.

2) We will use the NFLTEAMVALUES2018 dataset (containing information for the 2018-2019 NFL season) to study the relationship between the Value (y) of an NFL team and its 'Debt to Value' (x_1), 'Revenue (x_2), and Operating Income (x_3).

(a) State the form of the OLR model for the relationship above.
(b) Estimate the OLR model. Copy and Paste the output generated by Rstudio. Indicate what the slope and intercept estimates are.
(c) State the estimated OLR model.
(d) State code that could be used to output what the model in (c) would predict for each observation in the dataset.
(e) State the "error" made by the model in (c) for the New York Jets.
(f) Include a scatterplot of the value of an NFL team versus its yearly revenue.

3) Generate a dataset consisting of 30 observations on two predictor variables and a response variable. The values for the predictor variables should be generated randomly. The values for the response variable should be generated such that the three variables follow an Ordinary Linear Regression Model (but not the one discussed in RegressionIntro.R. Then include the following:

   (a) The code you used to generate the predictor and response variable values.

   (b) The (TRUE) Ordinary Linear Regression Model which describes the relationship between the three variables.

   (c) Pretending that we do not know the coefficient for the model in (b), use the dataset to estimate them using Rstudio. Provide the estimated values.

   (d) Generate another 30 observations where the TRUE OLR model is the same as in (b) except the variance of the (simulated) residuals here is 10 times larger. Use the dataset to estimate the coefficients for this model. Provide the estimated values. Which of the estimated models does a better job estimating the coefficients in the model in (b)?