HW#2 (Due February 10th)

You may work in groups of two. Both names must appear on the turned-in assigment. Note that you will not be able to work with this person again for the remainder of the HWs.

For Question (1)-(3) Refer to the HOMES Dataset.
We can note that the first column (**Value**) contains the current price of a home. The other columns are self-explanatory where we note that in any column containing 1s and 0s, 1 represents that the given amenity does appear in the home and 0 that it does not. Note that CAC stands for Central Air Conditioning. For example, the house in row 1 does have an Eat-in-kitchen and CAC, does not have a Fireplace, etc.

1) Suppose we wanted to predict the value of a home using an Ordinary Linear Regression Model.

   (a) Decide on a group of (5) Predictor Variables that could be used from the data we have. You should not be spooked by the columns containing 1s and 0s since these are still numeric values and such columns can be used in Regression Models (we will discuss these kinds of columns soon).

   (b) Formally state what the Linear Regression Model would look like using the variables in (a). The easiest way to do this is to define $x_1 = Lotsize,$ if the lotsize column appears in (a), etc. and then just writing out the model using the notation learned in class.

   (c) Do we know the values of the "beta" parameters in the model? Can we obtain them using the HOMES dataset? If we did know them, **describe** how we could obtain a good prediction of the value of a home.

   (d) Estimate the regression model in (b) using Rstudio. Write the full estimated model out.

   (e) Assume that the predicted values of a home that you would obtain using the estimated model in (d) are what the houses should be priced at. Use this to identify the biggest bargain in the dataset.

2) Generate a correlation table and heatmap for all the predictor variable columns in the dataset. You only need to include the heatmap in your assignment. Do you see any columns that are highly correlated? What issue does this bring up and what should we do to avoid this issue?

3) Split the HOMES dataset into 2 new sets called “HOMES_Train” and “HOMES_Test”, where “HOMES_Traint” consists of the first 60 observations and “HOMES_Test” consists of the remaining observations in HOMES.

   (a) Using the same 5 predictor variables as in 1(a), estimate a regression model using the observations in “HOMES_Train” and use this model to get predictions for the observations in “HOMES_Test”. You must provide the code used, the estimated regression model, and the value of MSE (mean-squared error) for the observations in HOMES_Test.

   (b) Try all subsets of 4 predictor variables from your original 5 and do the same thing as in (a). That is, for each possible subset, estimate a model using “HOMES_Train” and compute MSE for HOMES_Test. You must provide the code used and the value of MSE for each of the models using different subsets of 4 predictor variables.

   (c) Out of the models in (a) and (b), which one contains the optimal set of predictor variables?