**Homework 2**

**Upload Excel File and RData file to Isidore**

1. **Histograms**
2. Develop Histograms for all of the UD housing characteristics.
3. Comment based upon the histogram which of the factors should be excluded if developing a model to predict the Gas.Heating.Intensity factor.
4. **Input Factor Correlation**

Ignoring the factors: House.Number and Address, and the factors you’ve chosen to eliminate in Problem 1, develop a Pearson pairwise correlation for all input factors. Note: you could use Rattle to do this more easily.

Identify the factors which have a high degree of correlation (>0.9 and 0.95).

1. **Removing Highly Correlated Factors**

Create dataframe that includes the input factors considered in #2 (e.g., with the noted excluded factors). Create new dataframes with highly correlated factors removed (for correlation > 0.9, and correlation > 0.95). Print the header rows of each of the created dataframes. Copy into Excel, problem 3 tab.

1. **Removing Multi-Colinearity**

Using the remaining input factors from #3, with factors having correlation > 0.95, Identify the factors (if any) that should be retained after removal for co-linearity. Create a new dataframe based on these factors. Note: that these should be the factors used to develop a model from.

1. **Principal Component Regression**

Using the remaining input factors from #4, along with the target Gas.Heat.kBTU.sf, first identify how many principal components you need to account for 95% of the variation. Also, develop a principal component regression. Plot the R2 as a function of number of principal components. Lastly, comment upon how many principal components you should use to develop a decent linear model.