Assignment #1: Exploratory Data Analysis for Regression (30 points)

Data Directory: Data can be accessed on the SAS OnDemand server using this libname statement.

```
libname mydata        '/courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/' access=readonly;
```

Data Set: mydata.building_prices

Data Description: See the data dictionary or pp. 328-329 of *Regression Analysis By Example*.

Assignment Instructions:

For this assignment we will perform an Exploratory Data Analysis (EDA) for the building_prices data set.

This EDA will consist of three parts:

(1) Use PROC CORR to produce the Pearson correlation coefficients and a scatterplot matrix of the predictor variables X1 through X9 with the response variable Y (sale price of the house in thousands of dollars). For help with SAS see Chapter 8 pp. 111-117 in *SAS Statistics By Example*.

```
ods graphics on;
proc corr data=temp plot=matrix(histogram nvar=all);
run;
ods graphics off;
```

(2) Comment on which predictor variables have the strongest relationships with the response variable? What do you notice about the relationship between the numeric correlation measure and the graphical relationship in the scatterplot? Which predictor variable you think will be the best single predictor variable (hint: see Anscombe's example pp. 28-30 in *Regression Analysis By Example*).

(3) Produce a scatterplot with a LOESS smoother for Y with each of the predictor variables X1 through X9. See Section 8.6 in *The Little SAS Book*.

```
ods graphics on;
proc sgscatter data=temp;
compare x=(x1 - x2)
      y=Y / loess;
run; quit;
ods graphics off;
```

For your own learning follow some of the examples in *SAS Statistics By Example* and learn how to make a histogram, a bar chart, and a scatterplot matrix using PROC SGPLOT and PROC SGSCATTER.

Assignment Document:

All assignment reports should conform to the standards and style of the report template provided to you. Results should be presented and discussed in an organized manner with the discussion in close proximity of the results. The report should not contain unnecessary results or information. At a minimum the report for this assignment should contain the correlation matrix and the correlation scatterplot matrix. The document should be submitted in pdf format.