Assignment #1

Introduction:

The purpose of this assignment is to gain some understanding of the relationships between the predictor variables X1, X2, …, X9 and the response variable Y in the building_prices data set. In order to understand the relationships between these variables we will perform some exploratory data analysis (EDA) steps including calculating Pearson correlation coefficients, creating a scatterplot matrix, and generating scatterplots with a LOESS smoother. These EDA steps will help us gain insight into which predictor variables have a meaningful relationship with the response variable.

Results:

The table below shows the Pearson correlation coefficient for each combination of the variables in the building_prices data set. In each cell of the table the top number is the Pearson correlation coefficient and the bottom number is a p-value representing the probability of observing an R value at least this extreme if no correlation existed between these two variables.

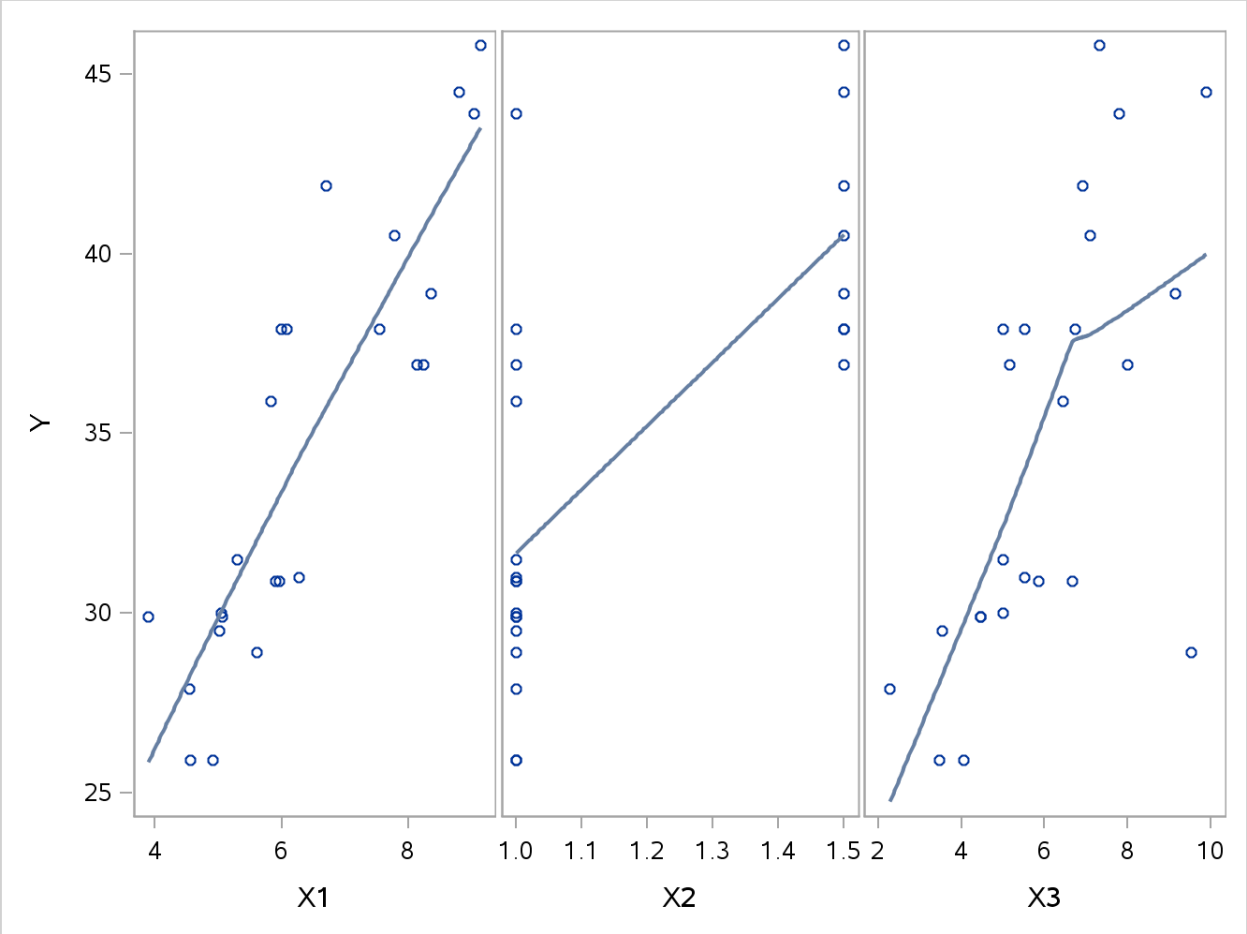
Pearson Correlation Coefficients, N = 24 Prob > |r| under H0: Rho=0										
	X1	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**	**X9**	**Y**
X1	1.00000	0.65127 0.0006	0.68921 0.0002	0.73427 <.0001	0.45856 0.0242	0.64062 0.0007	0.36711 0.0776	-0.43710 0.0327	0.14668 0.4940	0.87391 <.0001
X2	0.65127 0.0006	1.00000	0.41296 0.0449	0.72859 <.0001	0.22402 0.2926	0.51031 0.0108	0.42640 0.0377	-0.10075 0.6395	0.20412 0.3387	0.70978 0.0001
X3	0.68921 0.0002	0.41296 0.0449	1.00000	0.57155 0.0035	0.20466 0.3374	0.39212 0.0581	0.15161 0.4795	-0.35275 0.0909	0.30599 0.1459	0.64764 0.0006
X4	0.73427 <.0001	0.72859 <.0001	0.57155 0.0035	1.00000	0.35888 0.0850	0.67886 0.0003	0.57434 0.0033	-0.13909 0.5169	0.10656 0.6202	0.70777 0.0001
X5	0.45856 0.0242	0.22402 0.2926	0.20466 0.3374	0.35888 0.0850	1.00000	0.58939 0.0024	0.54130 0.0063	-0.02017 0.9255	0.10162 0.6366	0.46147 0.0232
X6	0.64062 0.0007	0.51031 0.0108	0.39212 0.0581	0.67886 0.0003	0.58939 0.0024	1.00000	0.87039 <.0001	0.12427 0.5629	0.22222 0.2966	0.52844 0.0079
X7	0.36711 0.0776	0.42640 0.0377	0.15161 0.4795	0.57434 0.0033	0.54130 0.0063	0.87039 <.0001	1.00000	0.31351 0.1358	0.00000 1.0000	0.28152 0.1826
X8	-0.43710 0.0327	-0.10075 0.6395	-0.35275 0.0909	-0.13909 0.5169	-0.02017 0.9255	0.12427 0.5629	0.31351 0.1358	1.00000	0.22578 0.2888	-0.39740 0.0545
X9	0.14668 0.4940	0.20412 0.3387	0.30599 0.1459	0.10656 0.6202	0.10162 0.6366	0.22222 0.2966	0.00000 1.0000	0.22578 0.2888	1.00000	0.26688 0.2074
Y	0.87391 <.0001	0.70978 0.0001	0.64764 0.0006	0.70777 0.0001	0.46147 0.0232	0.52844 0.0079	0.28152 0.1826	-0.39740 0.0545	0.26688 0.2074	1.00000

A large absolute value for the correlation coefficient, and correspondingly a low p-value, indicates a strong linear relationship between the two variables. Evidence of linear relationships can be strengthened, or weakened, through a visual analysis of the scatterplot matrix below.

Using a guideline stating that correlation coefficient values of greater than 0.7 or less than -0.7 indicate a strong linear relationship our table of correlation coefficients reflects strong linear relationships between Y and variables X1, X2, and X4. A visual inspection of the scatterplot for Y and X1 confirms the strong linear relationship between these two variables; the same is true for the scatterplot for Y and X4. Since variable X1 has the highest correlation coefficient and the scatterplot between Y and X1 confirms a strong linear relationship between these variables, there is strong evidence that X1 will be the best single predictor variable. The scatterplot for Y and X2 indicates that X2 is likely a categorical variable.

Detecting a linear relationship on a scatterplot involving at least one categorical can be difficult, but it is clear that the concentration of the Y values is strongly related to the value of X2.

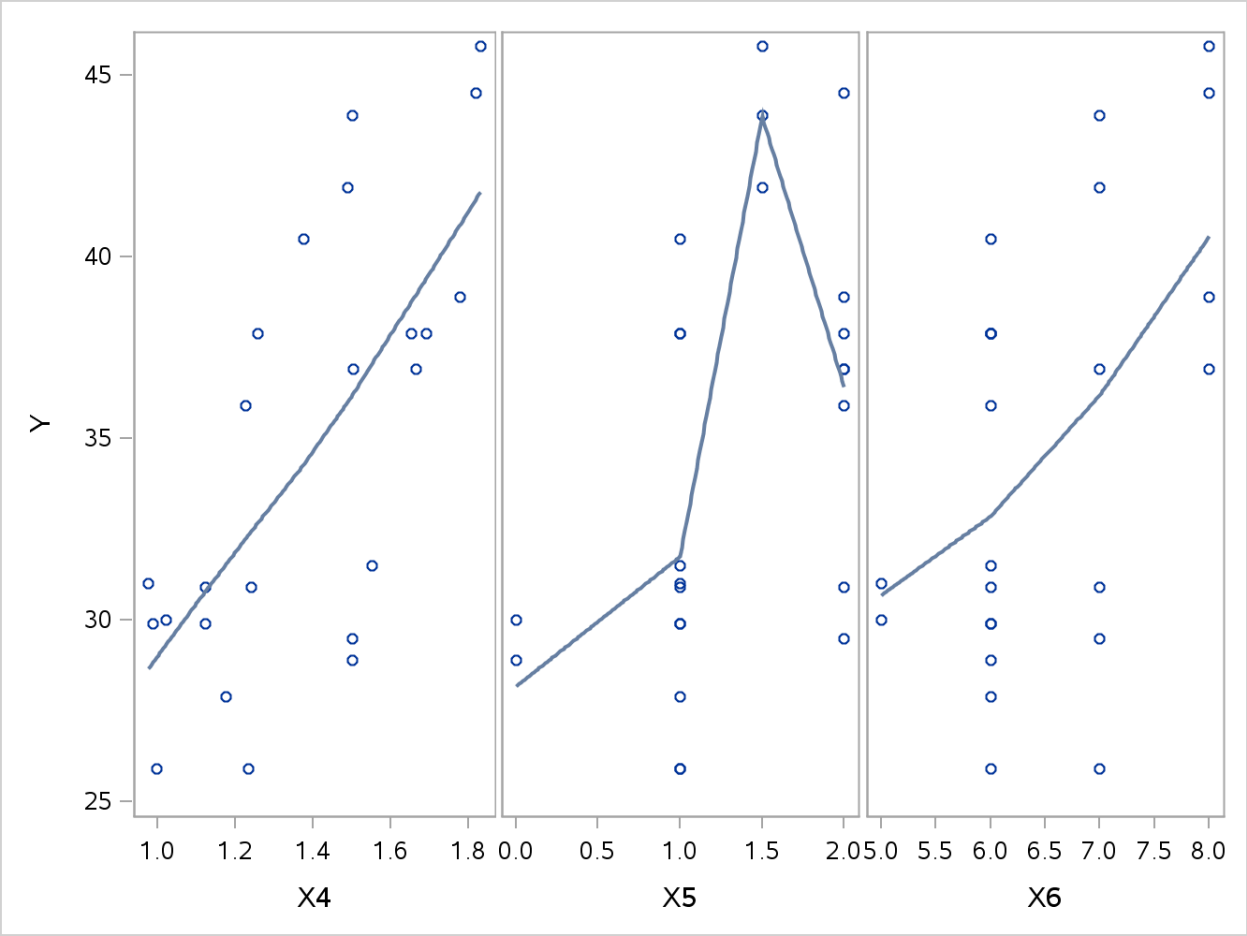
The scatterplots with a loess smoother, which are illustrated below, can provide further evidence for relationships between variables.



The straightness of the loess curve for the plot between variables Y and X1 confirms that the relationship between these two variables is strongly linear. We can now be even more confident in our conclusion that X1 will likely be the best single predictor of Y.

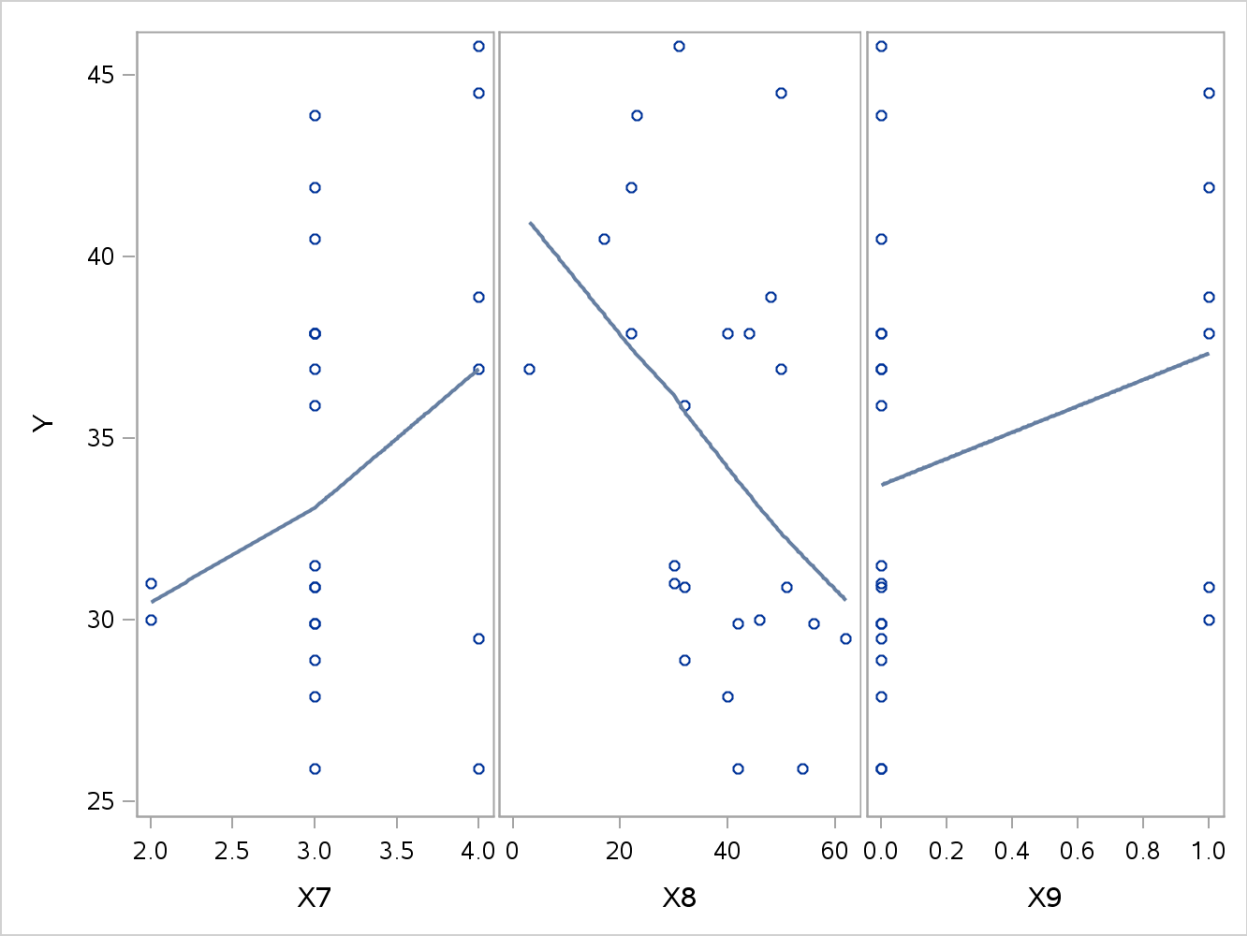
The loess curve involving X2 illustrates the difference in the concentration of data points for the two values of the predictor variable which is evidence of a relationship between X2 and Y.

The two distinct slopes of the loess curve in the third panel alerts us to the impact of an extreme value in the bottom right portion of the graph, around an X3 value of 10 and a Y value of less than 30. This outlier value reduces the value of the correlation coefficient. This data point should be investigated further to see if it was caused by a data error or by some mechanism that should be accounted for in our model. Depending on the cause of this outlier it is possible that the linear relationship between Y and X3 is stronger than indicated by the correlation coefficient shown earlier.



The shape of the loess curve in the graph of Y and X4 is strongly linear, with just a slight upward curve.

The graph involving X5 and the graph involving X6 show that X5 and X6 are categorical variables. It appears that there four distinct values for each of these predictor variables. It appears that the limited number of data points for some of the categories may be having an impact on the shape of the loess curve that would make it difficult to draw meaningful conclusions regarding these variables. Additional data points could improve the accuracy of our analysis.



The graph involving the apparrently categorical variable X7 shows a moderately linear loess curve, but with only two observations with an X7 value of 2.0 it is risky to give much weight to this outcome.

The loess curve for the graph of Y and X8 shows a strong negative linear relationship between these two variables, however the R value of -0.39740 merits a second look. A closer look at the graph shows that there is a relatively large distance between the loess curve and many of the data points. This is an indication of error which lowers the absolute value of the correlation coefficient and weakens the predictive power of the X8 variable.

The graph of Y and X9 shows that the categorical variable X9 has two distinct values which make the loess curve a perfectly straight line. However, the slope of this line is less than graph of Y and X2. This is an indication that there is a weaker relationship between Y and X9 than between Y and X2. This illustrates the value of the loess curve in evaluating the relative importance of different categorical variables.

Conclusions:

Evaluating the same variables through these three EDA methods demonstrates the importance of utilizing a variety of techniques in order to maximize the amount of information that can be gleaned from the data. There are important conclusions about the data that were not apparent after only calculating correlations coefficients. Additional techniques also help to confirm our conclusions and can strengthen our confidence in knowing which variables are important.

Code:

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

data temp;
      set mydata.building_prices;
run;

ods graphics on;
title "Compute Pearson Correlation Coefficients and Create Scatterplot
      Matrix";
proc corr data=temp nosimple plots = matrix(histogram nvar=all);
run;
ods graphics off;

ods graphics on;
title "Scatterplots with a LOESS Smoother";
proc sgscatter data=temp;
compare x=(X1--X3)
        y=Y / loess;
run;
proc sgscatter data=temp;
compare x=(X4--X6)
        y=Y / loess;
run;
proc sgscatter data=temp;
compare x=(X7--X9)
        y=Y / loess;
run; quit;
ods graphics off;
```