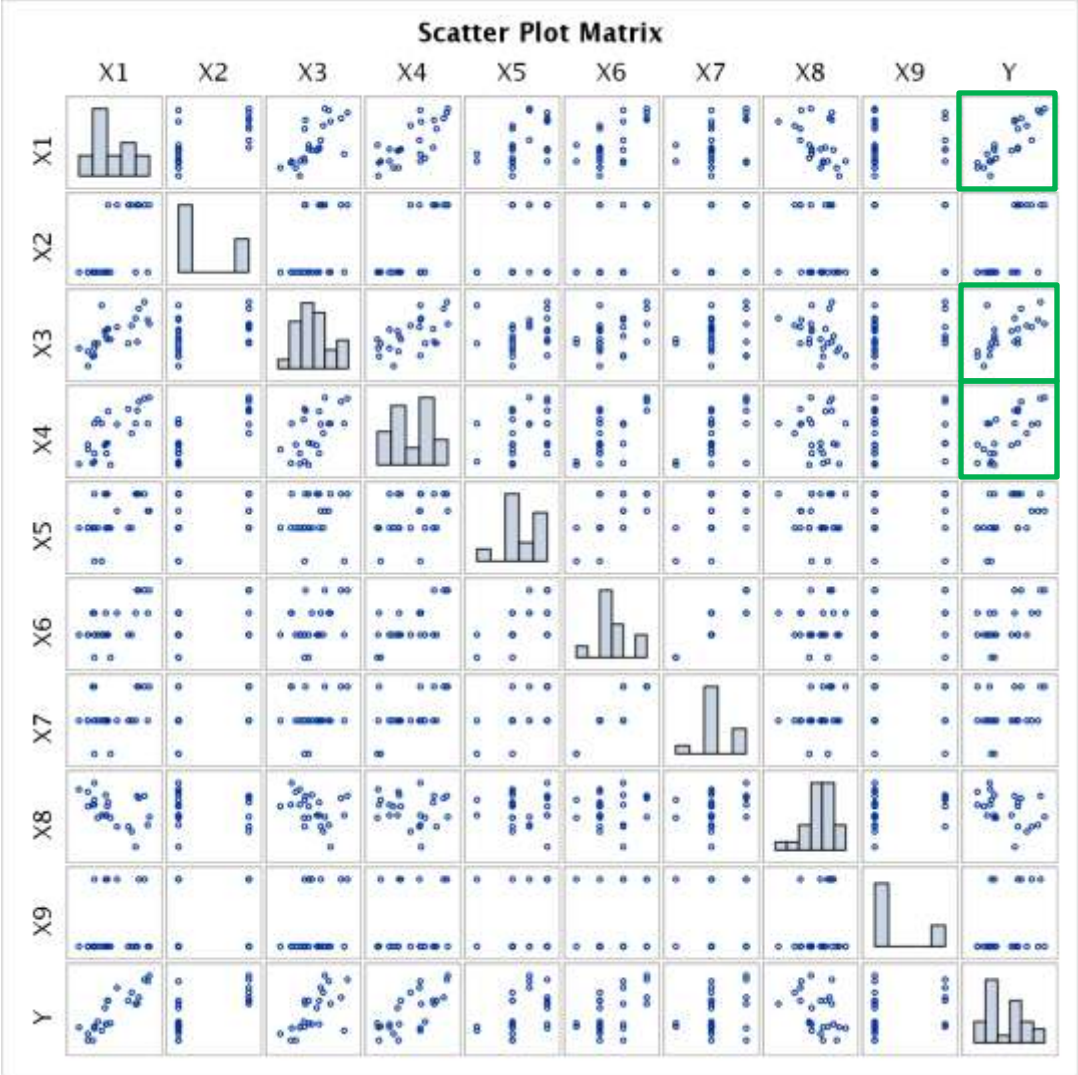
Assignment #1

Introduction:

The purpose of this exercise is to perform an exploratory data analysis (EDA) of the data set, **building_prices**. EDA identifies relationships between the predictor and response variables using a variety of methods, including simple summary of the data as well as plotting the data to visualize patterns. The SAS procedure, PROC CORR, will generate Pearson's Correlation Coefficient and scatterplot matrix, which will provide indication of the strength and direction of the relationship between the predictor and response variables.

Results:

In the matrix below, the scatterplots highlighted in green depict a relatively clear visual of the positive linear relationship between X (the predictor variable) and Y (response variable).

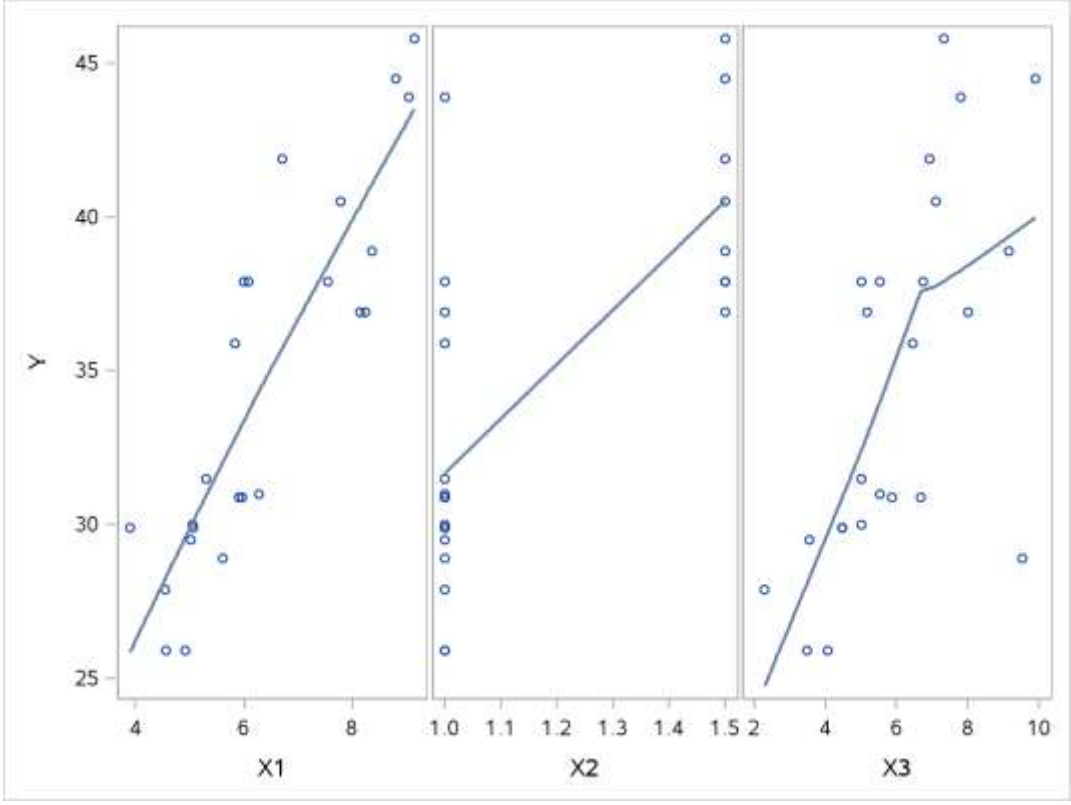


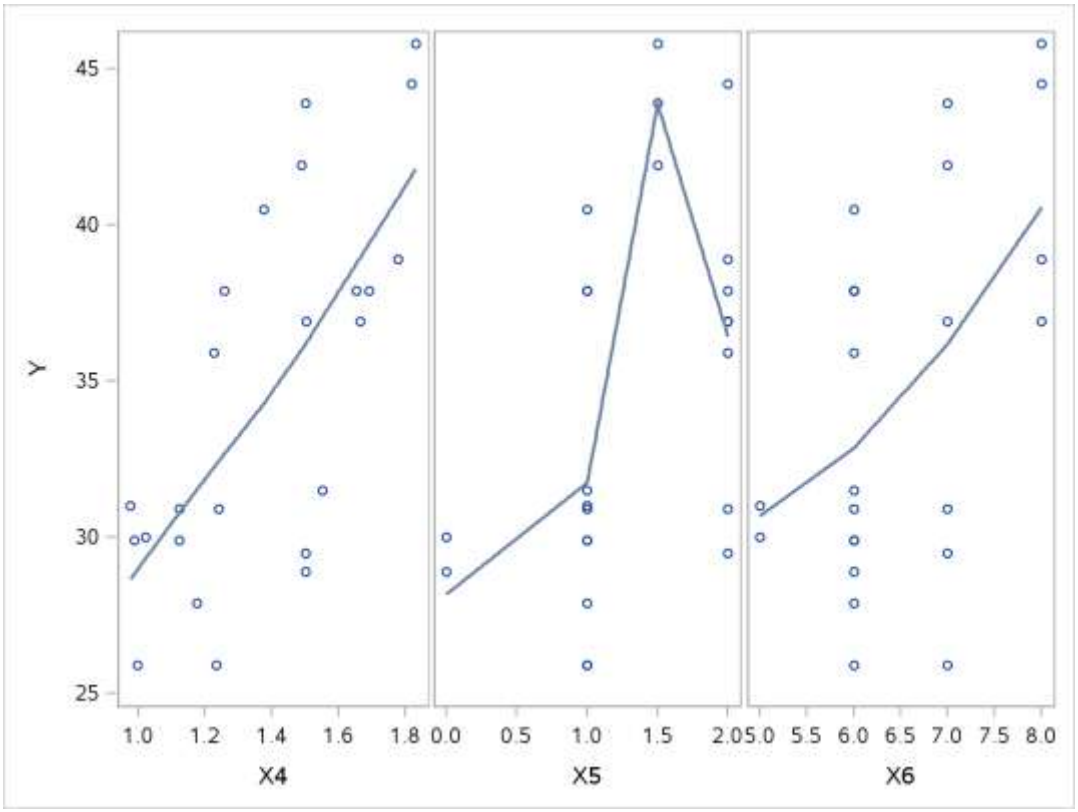
The table below contains the Pearson Correlation Coefficients ranked from highest to lowest (absolute value). The variables, X1, X2 and X4 have a strong positive linear relationship to Y, while X3, X6 and X5 have a moderately positive relationship. X8 has a moderately negative relationship to Y, and X7 and X9 have a weak linear relationship (using the scale outlined in Ratner, pg. 18).

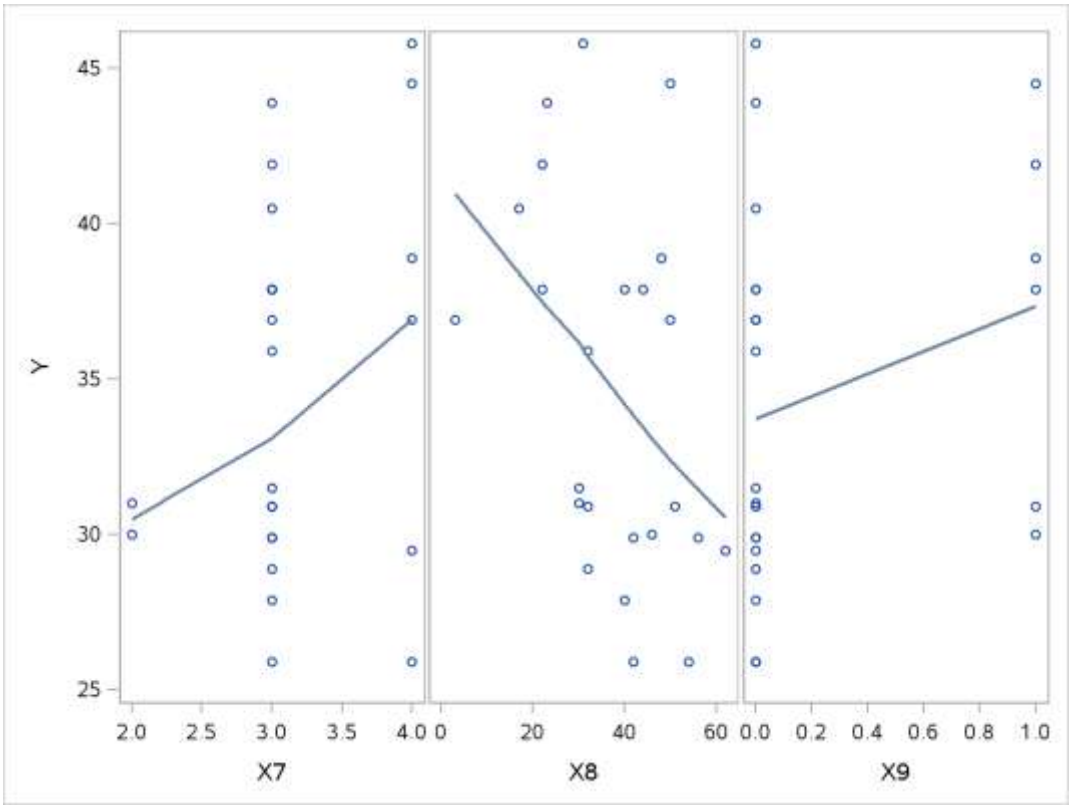
Pearson Correlation Coefficients, N = 24 Prob > |r| under H0: Rho=0									
Y	X1	X2	X4	X3	X6	X5	X8	X7	X9
	0.87391	0.70978	0.70777	0.64764	0.52844	0.46147	-0.39740	0.28152	0.26688
	<.0001	0.0001	0.0001	0.0006	0.0079	0.0232	0.0545	0.1826	0.2074

However, there are some discrepancies between the visual analysis based on the matrix and the correlation coefficient. X2 has a correlation coefficient of .70978, a value which might normally be interpreted as a strong positive linear relationship. However, having seen the scatterplot for X2, there is <u>not</u> a strong correlation between (X2, Y) since the data does not satisfy the assumption of linearity. X2 is an example where plotting the data points is a critical step prior to interpreting the Pearson Correlation Coefficient.

The following graphs are scatterplots of (X,Y) with a Loess smoother. Again, this graphical representation illustrates the positive linear relationship between (X1,Y), (X3,Y) and (X4,Y), and the non-linear scatterplot of (X2,Y).




Y
45
40
35
30
25
1.0 1.2 1.4 1.6 1.8
X4
0.0 0.5 1.0 1.5 2.0
X5
5.0 5.5 6.0 6.5 7.0 7.5 8.0
X6


Y
45
40
35
30
25
2.0 2.5 3.0 3.5 4.0
X7
0 20 40 60
X8
0.0 0.2 0.4 0.6 0.8 1.0
X9

Of the potential predictor variables (X1, X3, X4), **X1** is likely the best predictor variable. Of the three, X1 has the highest correlation coefficient. In addition, the scatterplot of X1 appears most linear, without any clear outliers. Lastly, X1 has the highest r^2 value of .763, or X1 accounts for 76.3% of the total variability in Y.

Within the context of the data, that is to say that of the 9 potential predictor variables, X1, the amount of taxes paid, is the best predictor of Y, the sale price of the house.

Conclusions:

This exercise in Exploratory Data Analysis demonstrated some aspects of the summary statistics and graphical displays used to analyze the data prior to constructing a model. The Pearson Correlation Coefficient provided a numerical basis to quantify the direction of the linear relationship (positive vs. negative). However, the coefficient cannot be relied upon unless the scatterplot of the data also indicates that the data is linear. As a result of EDA, one can conclude that X1 appears to the be strongest predictor variable out of X1-X9 based on high correlation coefficient paired with a visibly linear scatterplot.

Code:

```
/*****************************************/
/** PRE 410 Section 55                   **/
/** Assignment #1                        **/
/*****************************************/

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

data temp ;
      set mydata.building_prices ;
run ;

/*****************************************/
/** Scatterplot matrix                   **/
/*****************************************/

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

/*****************************************/
/** Pearson Correlation Coefficient      **/
/*****************************************/

proc corr data=temp rank;
      var x1 x2 x3 x4 x5 x6 x7 x8 x9;
      with y;
run;
```

```
/****************************************/
/** Scatterplot with loess smoother     **/
/****************************************/

ods graphics on;
title 'X1 - X3';
proc sgscatter data=temp;
      compare  x=(x1-x3)
      y=Y / loess;
run; quit;
ods graphics off;

ods graphics on;
title 'X4 - X6';
      proc sgscatter data=temp;
      compare  x=(x4-x6)
y=Y / loess;
run; quit;
ods graphics off;

ods graphics on;
title 'X7 - X9';
proc sgscatter data=temp;
      compare  x=(x7-x9)
      y=Y / loess;
run; quit;
ods graphics off;

title '';
```