Assignment #1

Daniel S Prusinski

Introduction:

Purpose of Assignment:

The purpose of this assignment is to perform an Exploratory Data Analysis (EDA) on the building_prices data set. In order to effectively carry out an EDA, I will need to follow the following steps that combine EDA and steps in Regression Analysis (RA). The most important aspect is using the data to determine the model. Below is a breakdown on how the EDA utilized the steps in RA to solve the assignment:

1. Statement of the Problem – Perform an EDA on the building_prices data set, such that one can better understand the relationships between the data.
2. Selection of Potentially Relevant Variables – For this problem, 9 predictor variables are already given, thus after conducting analysis, one can better discern which variables are good indicators.
3. Data Collection – For this problem, the data is already collected and the metadata consists of 9 variables and 1 response variable. There are 24 observations per variable.
4. Model Specification: This problem dictates to conduct a Pearson Correlation Coefficients (PCC) and a Scatter Plot Matrix of the predictor variables with Y. The model is assumed to be linear and if conducting a regression equation it would be a multiple regression equation.
    a. The Results portion will reveal which variables have the strongest linear relationship with the response variable. This information is important in determining which predictor variables can accurately predict the response variable. The Pearson correlation coefficient will express as a percent of how much the response variable can be explained by the predictor variable.
5. The next steps are not relevant because for this EDA it is not necessary to actually conduct a regression model; rather it is focusing on the relational dynamics between the predictor variables and the response variable.

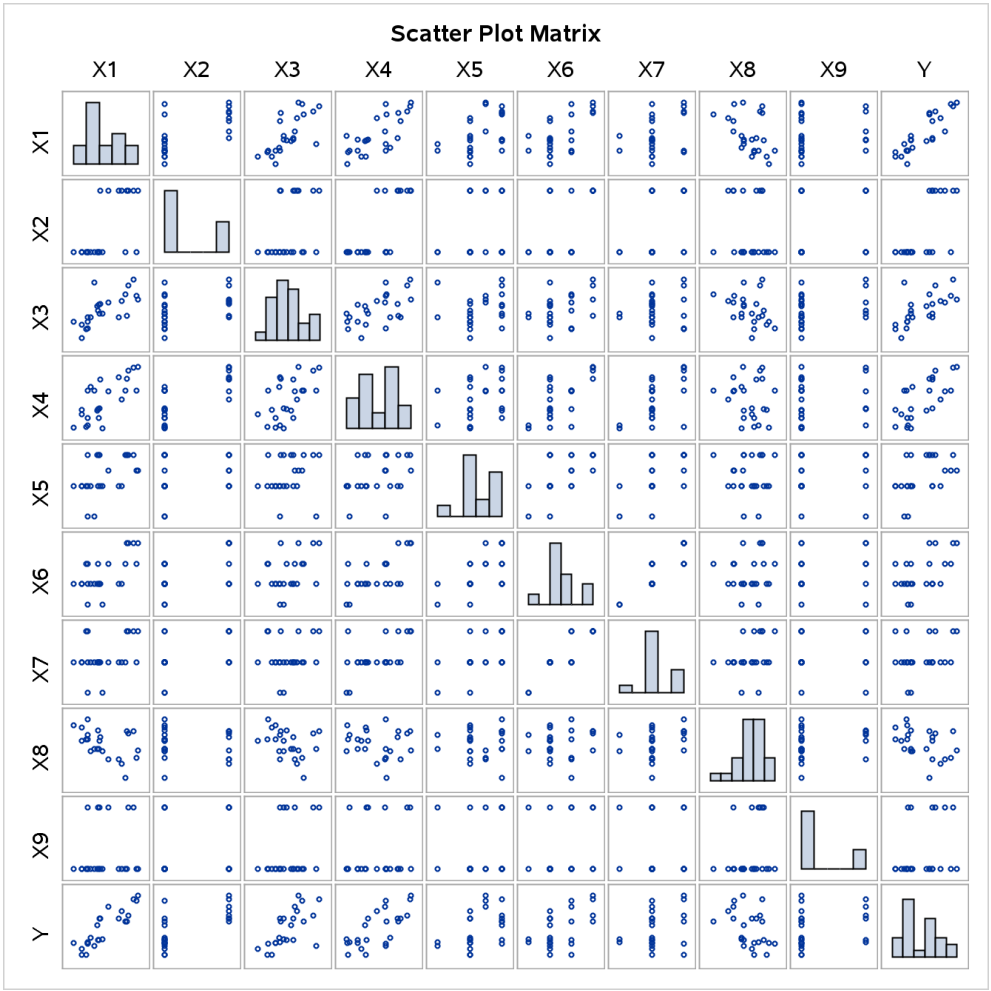
Results:

(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.

Pearson Correlation Coefficients, N = 24
Prob > |r| under H0: Rho=0

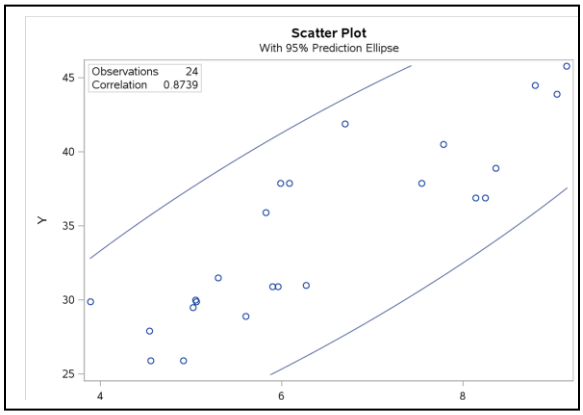
	Y	X1	X2	X4	X3	X6	X5	X8	X7	X9
Y	1.00000	0.87391	0.70978	0.70777	0.64764	0.52844	0.46147	-0.39740	0.28152	0.26688
		<.0001	0.0001	0.0001	0.0006	0.0079	0.0232	0.0545	0.1826	0.2074

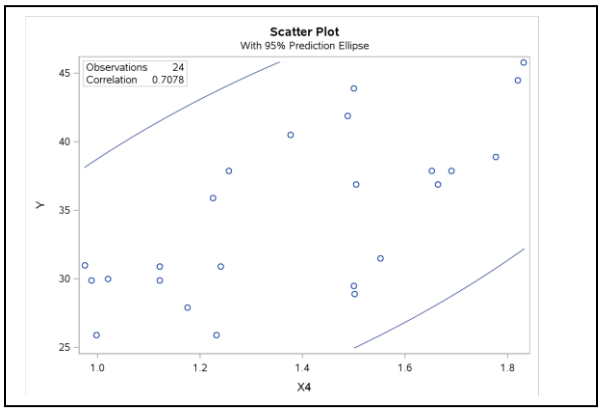
* I ranked the variables from greatest to least to make is easier to see which variables have the greater Pearson correlation coefficient (PCC).



*I find it very interesting that variables X2, X5, X6, X7, and X9 seem to not have a linear relationship when plotted on a scatterplot.

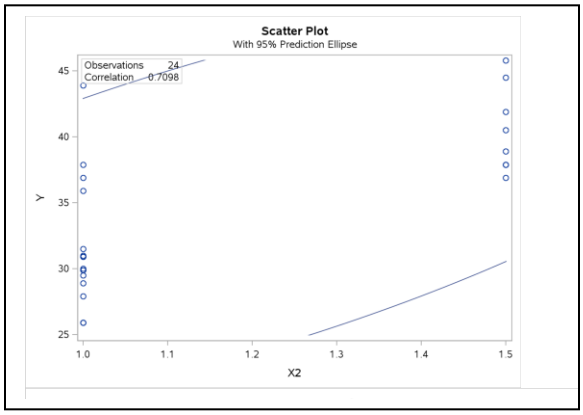
(2) Comment on which predictor variables have the strongest relationships with the response variable?

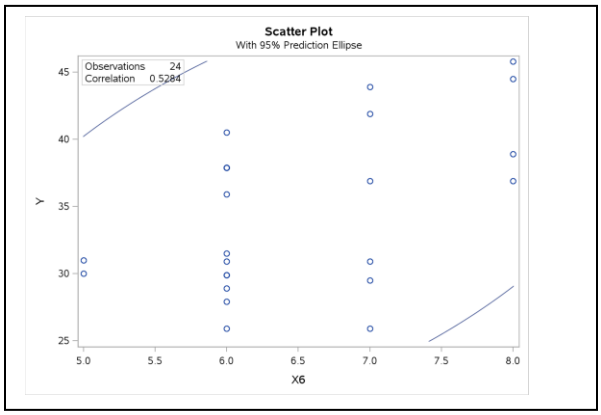




X1 and X4 appear to have the strongest relationships based on the strong PCC and the data visually appears to follow a linear relationship without major outliers. Furthermore, X1 has the strongest PCC and demonstrates a linear relationship, thus I find that this variable is the strongest predictor variable.

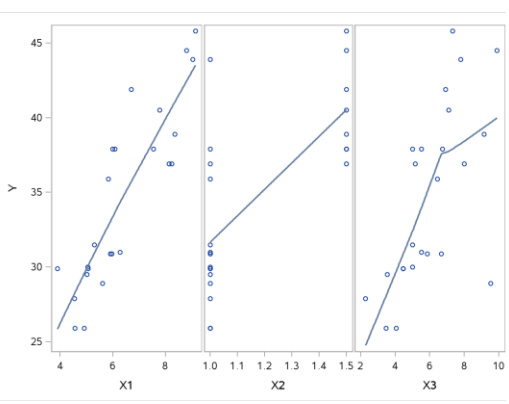
What do you notice about the relationship between the numeric correlation measure and the graphical relationship in the scatterplot?

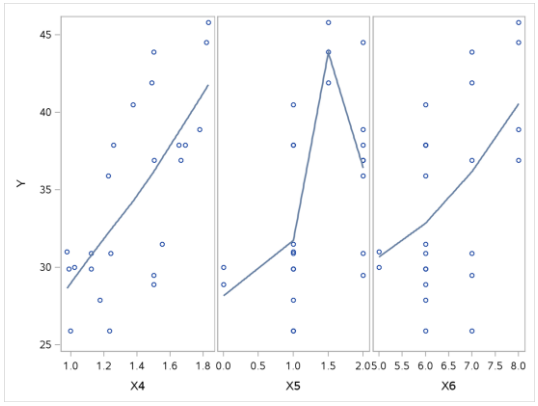


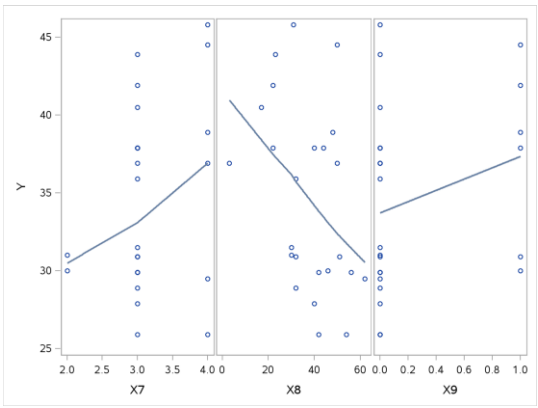


The numeric correlation does not necessarily demonstrate a linear correlative relationship. One needs to also look at the graphical relationship to visually determine that the data follow a linear plot. For example, variables X2 and X6 have relatively strong PCC's but upon visual inspection one can see the relationship is not linear.

(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*.







Conclusions:

Through this assignment, I learned how to produce PCCs, Scatterplots, Scatterplot Matrixs, LOESS smoothers, regression lines, and how to analyze data using SAS. One must combine statistical calculations with graphical analysis to verify a linear relationship.

Code:

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

*****This program will run Perason CC and create a scatter plot matrix of all
the data points*****;
ods graphics on;
title "computing Pearson CC";
proc corr data=mydata.building_prices nosimple rank
      plots = matrix(histogram nvar=all);
run;

*****This program will create a scatterplot for each variable on its own
page*****;
title "computing Pearson CC";
proc corr data=mydata.building_prices nosimple rank plots (only)=scatter
(nvar=all);
      var x1 x2 x3 x4 x5 x6 x7 x8 x9;
      with y;
run;
```

```
*****This program creates the LOESS smoother for Y with each of the predictor
variables*****;
title "Scatter Plot with Loess Smoother X1, X2, X3";
proc sgscatter data=mydata.building_prices;
compare x = (x1 x2 x3)
y=Y / loess;
run;

title "Scatter Plot with Loess Smoother X4, X5, X6";
proc sgscatter data=mydata.building_prices;
compare x = (x4 x5 x6)
y=Y / loess;
run;

title "Scatter Plot with Loess Smoother X7, X8, X9";
proc sgscatter data=mydata.building_prices;
compare x = (x7 x8 x9)
y=Y / loess;
run;
ODS graphics off;
```