Please note that we have made summated scales of only 3 variables because we could not find a proper benchmark to set for the remaining two factors. As we can see, there is no correlations between the scales. High correlations between factor scores and summated scales imply that these summated scales can be used in the place of these 3 factors for any analysis that needs to be done on these 7 variables.

Multiple Regression Analysis

Stage 1:

We are doing multiple regression analysis with the objective of accurately predicting the total House Sales price of the House sales dataset. We are also interested in identifying the variables that lead to help in predicting the sales price.

Stage 2:

This study is based on metric variables in the dataset, with 27 potential independent variables and 1 dependent variable of sale price. The assignment of the variables is the same as it was for factor analysis.

Stage 3:

The three assumptions to be addressed for the individual variables are linearity, constant variance (homoscedasticity), and normality. We have already explained the conditions of linearity and homoscedasticity in the descriptive statistics part.

Stage 4:

In this stage, we will approach model building starting from multiple regression analysis with a single best variable, stepwise regression analysis at alpha = 0.05. Please note that, at professor's instructions, we also tried keeping alpha = 0.01 to more restrict the criteria to enter the variate but only one variable was reduced.

We performed a multiple regression analysis on the House Sales data. The analysis was performed Stepwise wherein we performed 19 steps. In every step, a new variable was added to the analysis.

In the first step, the model F-value is 2436.77 which is significant, and the sales price could explain 62.57% of the variation in the data. In the second step, a new variable X14 was considered, and it led to an improvement in the R-squared value. In the third step, X9 was considered, and so on. The R-square increased gradually. The F-value for the variables was found significant.

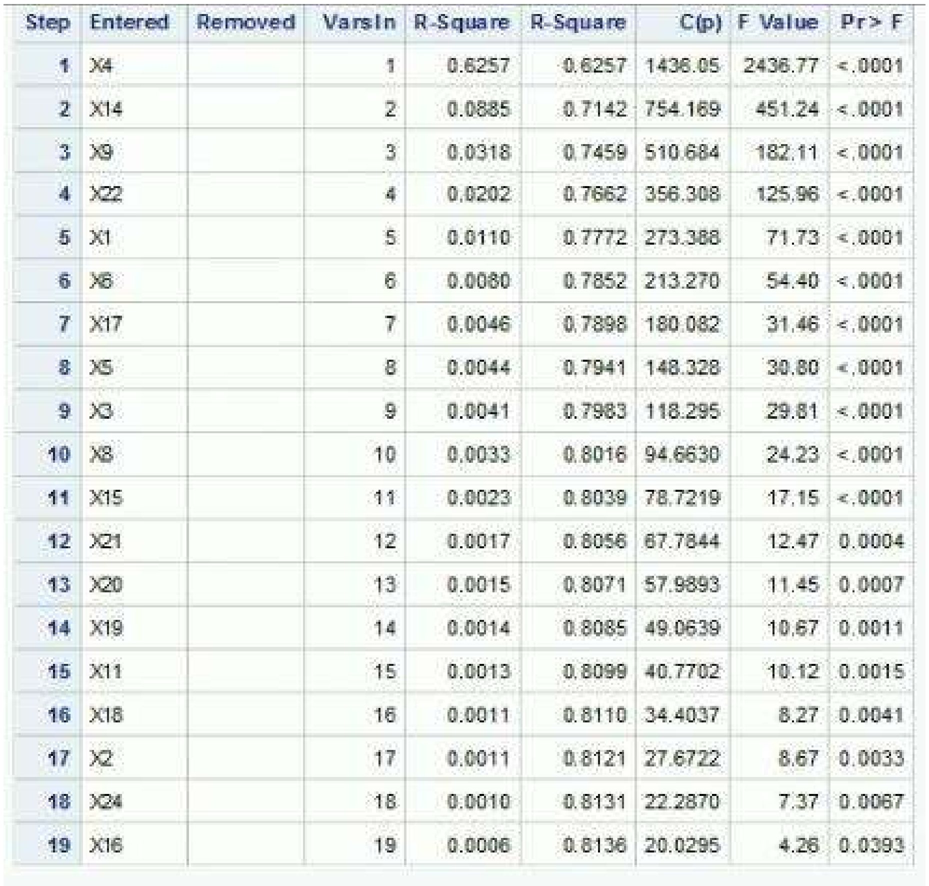
After Step 19, all the variables left are significant for estimating the House sales. The F-value for the model is 330.81 and the individual f-values are also significant. The R-square value is 0.8136 which means 81.36% of the variation can be explained by the regression equation.

The Summary statistics are shown below:

All 27 variables were considered as we can see no variable was removed in the process.

Parameter Estimates:

We can find out the regression variate equation based on the parameter estimates
Y= -9.43 + (-114.75)*X1 + (181.76)*X2 + ……… + (21.20)*X19
The P-value for all the variables is satisfactory to the condition of 0.05



Step	Entered	Removed	VarsIn	R-Square	R-Square	C(p)	F Value	Pr> F
1	X4		1	0.6257	0.6257	1436.05	2436.77	<.0001
2	X14		2	0.0885	0.7142	754.169	451.24	<.0001
3	X9		3	0.0318	0.7459	510.684	182.11	<.0001
4	X22		4	0.0202	0.7662	356.308	125.96	<.0001
5	X1		5	0.0110	0.7772	273.388	71.73	<.0001
6	X6		6	0.0080	0.7852	213.270	54.40	<.0001
7	X17		7	0.0046	0.7898	180.082	31.46	<.0001
8	X5		8	0.0044	0.7941	148.328	30.80	<.0001
9	X3		9	0.0041	0.7983	118.295	29.81	<.0001
10	X8		10	0.0033	0.8016	94.6630	24.23	<.0001
11	X15		11	0.0023	0.8039	78.7219	17.15	<.0001
12	X21		12	0.0017	0.8056	67.7844	12.47	0.0004
13	X20		13	0.0015	0.8071	57.9893	11.45	0.0007
14	X19		14	0.0014	0.8085	49.0639	10.67	0.0011
15	X11		15	0.0013	0.8099	40.7702	10.12	0.0015
16	X18		16	0.0011	0.8110	34.4037	8.27	0.0041
17	X2		17	0.0011	0.8121	27.6722	8.67	0.0033
18	X24		18	0.0010	0.8131	22.2870	7.37	0.0067
19	X16		19	0.0006	0.8136	20.0295	4.26	0.0393

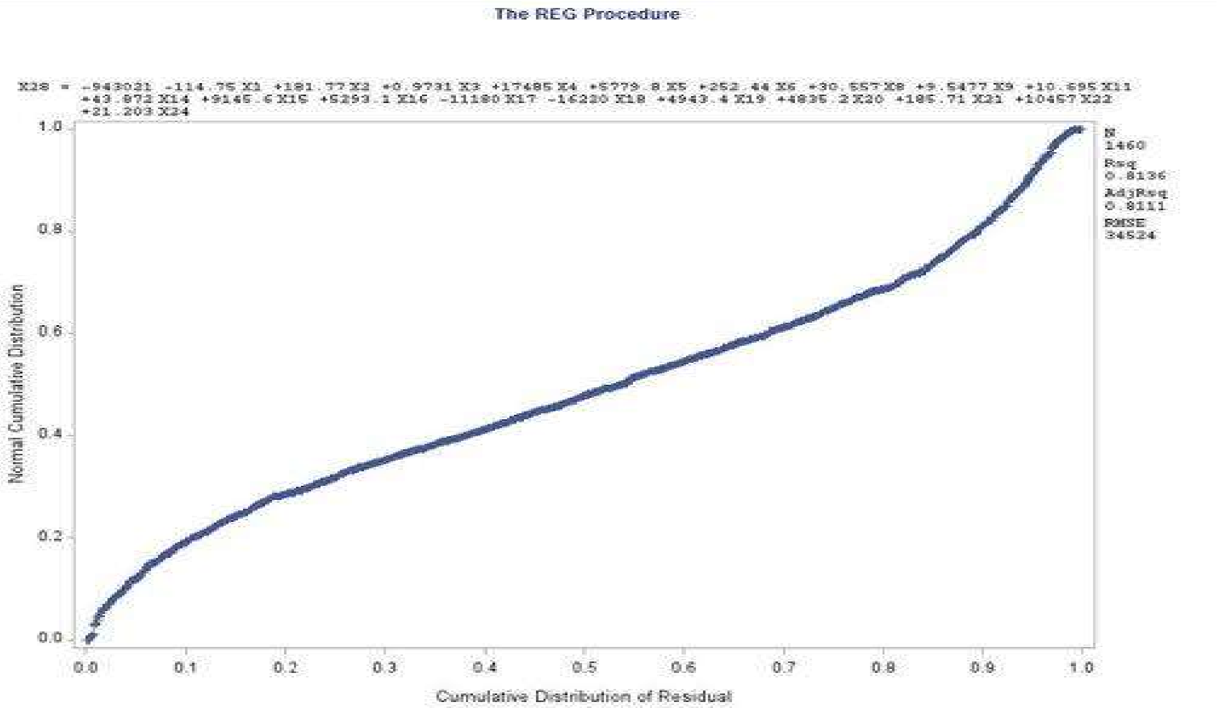
Based on professor's suggestion that we should redo the stepwise regression with 0.05 level of significance. This analysis that we just presented, is done with alpha = 0.05. We tried the similar analysis with level of significance of 0.01 to see if the number of variables entering the variate reduce. Only one variable was reduced at the significance level of 0.01.

Multiple Regression Assumptions:

Linearity: We have checked the scatter plots of every independent variable with the dependent variables. While we conclude the assumption of linearity holds for all the variables, we certainly can see that there are some variables that are more linear than others.

Homoscedasticity: The residual plots of all the independent variables do not seem to show any homoscedasticity. (Please refer to SAS output).

Normality: It seems that the condition of Normality is satisfied in the sense that in both ways the residual normal plot is almost on the diagonal. So, it would be fair to say that it barely satisfies normality.



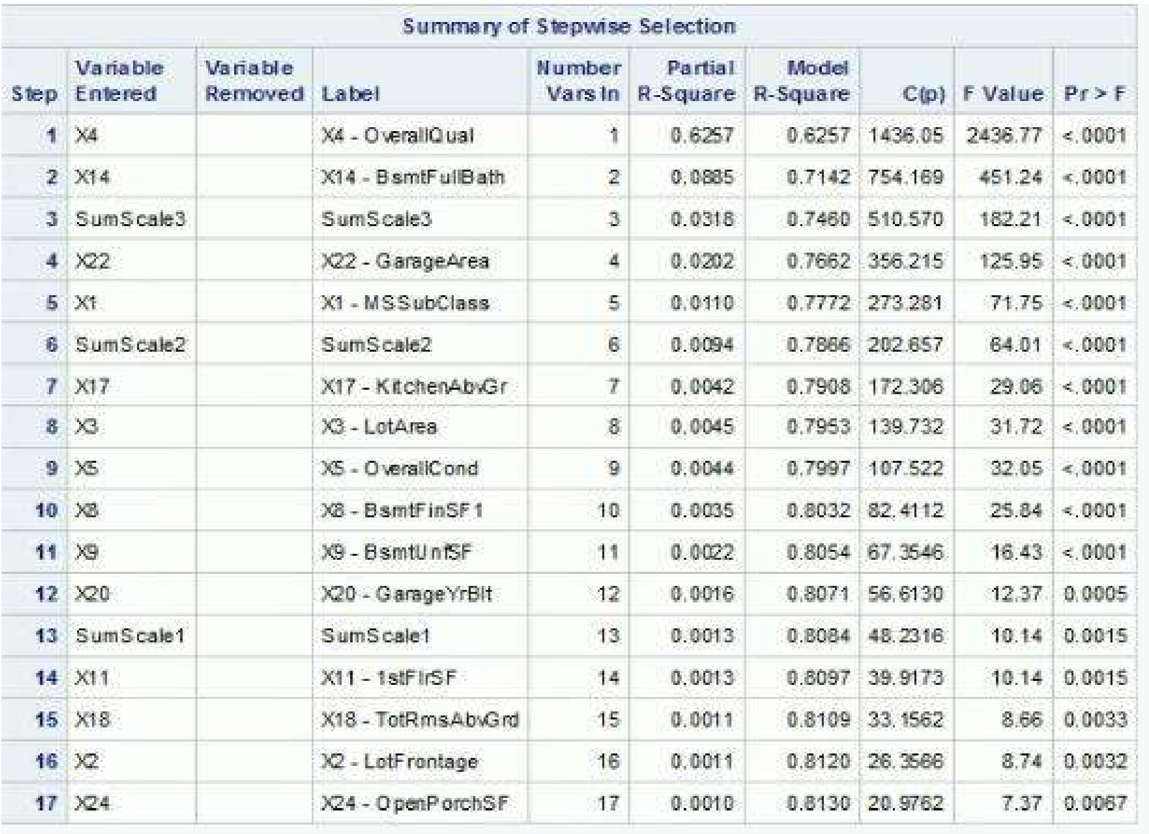
Independence: Since all the plots of the residuals against independent variables are without any pattern, it is safe to say that this condition is satisfied in all the models.

Multicollinearity: Since there is only one independent variable and stepwise regression model chooses variables with no multicollinearity.

Summary:
The Multiple regression model product predictions with good accuracy. The variation explained is 81.36%. These levels, in addition to the results, confirming model validity, give the greatest degrees of confidence regarding the quality and accuracy of the regression models used as the basis for developing business strategies.

Multiple Regression with Summated Scales:

We tried doing stepwise multiple regression analysis again by using the summated scales created in the factor analysis study to see how our model behaves differently. Below is the summary of stepwise regression.



Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	X4		X4 - OverallQual	1	0.6257	0.6257	1436.05	2436.77	<.0001
2	X14		X14 - BsmtFullBath	2	0.0885	0.7142	754.169	451.24	<.0001
3	SumScale3		SumScale3	3	0.0318	0.7460	510.570	182.21	<.0001
4	X22		X22 - GarageArea	4	0.0202	0.7662	356.215	125.95	<.0001
5	X1		X1 - MSSubClass	5	0.0110	0.7772	273.281	71.75	<.0001
6	SumScale2		SumScale2	6	0.0094	0.7866	202.657	64.01	<.0001
7	X17		X17 - KitchenAbvGr	7	0.0042	0.7908	172.306	29.06	<.0001
8	X3		X3 - LotArea	8	0.0045	0.7953	139.732	31.72	<.0001
9	X5		X5 - OverallCond	9	0.0044	0.7997	107.522	32.05	<.0001
10	X8		X8 - BsmtFinSF1	10	0.0035	0.8032	82.4112	25.84	<.0001
11	X9		X9 - BsmtUnfSF	11	0.0022	0.8054	67.3546	16.43	<.0001
12	X20		X20 - GarageYrBlt	12	0.0016	0.8071	56.6130	12.37	0.0005
13	SumScale1		SumScale1	13	0.0013	0.8084	48.2316	10.14	0.0015
14	X11		X11 - 1stFlrSF	14	0.0013	0.8097	39.9173	10.14	0.0015
15	X18		X18 - TotRmsAbvGrd	15	0.0011	0.8109	33.1562	8.66	0.0033
16	X2		X2 - LotFrontage	16	0.0011	0.8120	26.3566	8.74	0.0032
17	X24		X24 - OpenPorchSF	17	0.0010	0.8130	20.9762	7.37	0.0067

We can clearly see that the summated scale variables have been used, with sum scale 3 entering in the model at the 3rd place signifying the importance of the variables created through factor analysis.