Factor Analysis Study:

Stage 1:

We are doing an exploratory analysis where data reduction is the primary goal. The unit of analysis for factor analysis is variables for this study. We want to reduce the number of variables and divide them into factors that can capture the maximum variance.

Stage 2:

Although our data consists of metric and non-metric variables, this study is limited to metric variables where the sample size consists of 1460 observations of 27 variables. All the variables have been cleaned and here is their assignment to X values. Please note that this assignment is valid for this analysis, multiple regression, and cluster analysis.

X1 = 'X1 - MSSubClass'	X11 = 'X11 - 1stFlrSF'	X21 = 'X21 - GarageCars'
X2 = 'X2 - LotFrontage'	X12 = 'X12 - 2ndFlrSF'	X22 = 'X22 - GarageArea'
X3 = 'X3 - LotArea'	X13 = 'X13 - GrLivArea'	X23 = 'X23 - WoodDeckSF'
X4 = 'X4 - OverallQual'	X14 = 'X14 - BsmtFullBath'	X24 = 'X24 - OpenPorchSF'
X5 = 'X5 - OverallCond'	X15 = 'X15 - FullBath'	X25 = 'X25 - 3SsnPorch'
X6 = 'X6 - YearBuilt'	X16 = 'X16 - BedroomAbvGr'	X26 = 'X26 - MoSold'
X7 = 'X7 - YearRemodAdd'	X17 = 'X17 - KitchenAbvGr'	X27 = 'X27 - YrSold'
X8 = 'X8 - BsmtFinSF1'	X18 = 'X18 - TotRmsAbvGrd'	X28 = 'X28 - SalePrice';
X9 = 'X9 - BsmtUnfSF'	X19 = 'X19 - Fireplaces'	
X10 = 'X10 - TotalBsmtSF'	X20 = 'X20 - GarageYrBlt'	

Stage 3:

Iteration 1: The overall value of MSA for this analysis in the first iteration is 0.755 which is greater than 0.5 hence we can proceed further with the factor analysis. Communality estimates were 18.86 but there were some variables identified, namely X18, X10, and X13, in the 1st iteration that had individual MSA less than 0.5.

Iteration 2: After deleting X10, the overall value of MSA for this analysis increased in the 2nd iteration and became 0.785 and communality estimates decreased to 18.10 but there were still some variables identified, namely X18, X27, and X13, in the 2nd iteration that had individual MSA less than 0.5.

Iteration 3: After deleting X13, the overall value of MSA for this analysis increased in the 3rd iteration and became 0.824 and communality estimates decreased to 17.37 but

there were still some variables identified, namely X18, and X27, in the 3rd iteration that had individual MSA less than 0.5.

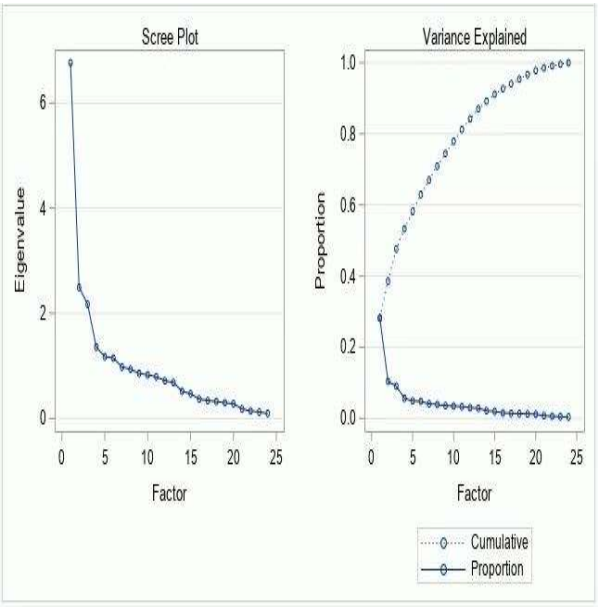
Iteration 4: After deleting X18, the overall value of MSA for this analysis increased in the 4th iteration and became 0.834 and communality estimates decreased to 15.10. There were no variables with individual MSA less than 0.5, so we are good to proceed further.

It is to be noted that at the start of the analysis, we have total of 27 variables which were giving an overall MSA of greater than 0.5 but individual MSA of some variables to be less than 0.5. Which prompted us to delete the variables X10, X13, and X18 from the dataset and then we achieved satisfactory overall and individual MSA to run factor analysis.

Stage 4:

We have decided to perform Principal component analysis for this study because here data reduction is a primary concern, focusing on the minimum number of factors needed to account for the maximum portion of the total variance represented in the original set of variables. Next, we must decide on the number of factors to extract. Here we will prefer the latent root criterion which is widely used and most satisfactory if the number of variables in the analysis are greater than 20. Using the latest root criterion, at the end of the 4th iteration, the number of factors with eigen value greater than 1 are 6. We have found the scree-test criterion to be always 1 factor behind the latent root criterion. If we were to use the scree-test criterion, we see then the point at which the curve first begins to straighten out is 5, which is considered the maximum number of factors to extract. Preferring the latent root, we will be extracting 6 factors in our 4th iteration.

Eigenvalues of the Correlation Matrix: Total = 24 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	6.76762997	4.27476787	0.2820	0.2820
2	2.49286210	0.32340453	0.1039	0.3859
3	2.16945757	0.81187029	0.0904	0.4762
4	1.35758728	0.18501670	0.0566	0.5328
5	1.17257057	0.02481353	0.0489	0.5817
6	1.14775704	0.17183783	0.0478	0.6295
7	0.97591922	0.04262535	0.0407	0.6702
8	0.93329386	0.07686271	0.0389	0.7090
9	0.85643115	0.03026970	0.0357	0.7447
10	0.82616145	0.03530722	0.0344	0.7792
11	0.79085423	0.07757854	0.0330	0.8121



Stage 5:

The unrotated factor matrix is not conclusive as we can see that every variable is loading on Factor 1 with residual variances loading on the remaining factors. Based on the statistical significance, with more than 250 observations, a factor loading of 0.30 would be considered significant.

Factor Pattern							
		Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
X4	X4 - OverallQual	0.79123	-0.15730	-0.16419	0.04086	0.04448	0.16554
X14	X14 - BsmtFullBath	0.77303	0.41511	-0.21637	0.24625	0.00404	-0.03232
X22	X22 - GarageArea	0.75608	-0.14119	-0.05539	-0.19645	-0.00720	0.03893
X23	X23 - WoodDeckSF	0.75329	-0.12385	0.05284	-0.19701	-0.00148	0.04291
X12	X12 - 2ndFlrSF	0.73609	0.13409	0.35179	0.01287	-0.07363	-0.04892
X11	X11 - 1stFlrSF	0.72572	-0.05142	0.38870	0.00707	-0.08418	-0.04008
X16	X16 - BedroomAbvGr	0.68049	0.04323	-0.46527	0.03173	0.02941	-0.06920
X6	X6 - YearBuilt	0.62099	-0.58597	-0.16187	-0.19543	-0.01683	-0.09121
X19	X19 - Fireplaces	0.61814	0.54352	-0.33271	0.18227	0.03652	-0.10163
X7	X7 - YearRemodAdd	0.54132	-0.42214	-0.25229	-0.05926	0.32033	0.34503
X8	X8 - BsmtFinSF1	0.52319	-0.03381	0.01592	0.12207	-0.07896	-0.18155
X20	X20 - GarageYrBlt	0.49516	0.22337	0.17462	0.25412	-0.07208	0.10375
X25	X25 - 3SsnPorch	0.38761	0.01941	-0.10224	0.04388	-0.08776	0.29715
X24	X24 - OpenPorchSF	0.36557	-0.06858	0.10132	0.19778	0.18734	0.11887
X17	X17 - KitchenAbvGr	0.27844	0.65897	-0.35233	0.08674	0.03252	-0.16426

We rotated this matrix orthogonally by Varimax method and obtained the rotated factor matrix. As we can see, this gives much clearer interpretation of the variable loading on the factors. We can see that there are a lot of violations based on dual/cross loadings. Hence these variables have become a candidate for deletion. We have to delete X7, X4, X23, X16, X11, X12, X20, X8, X3, a total of 9 variables due to cross loading. We can say that cross loading has become one of the major roadblocks for this factor analysis.

Rotated Factor Pattern							
		Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
X6	X6 - YearBuilt	0.83542	-0.05408	0.15790	0.00071	-0.27574	-0.01290
X7	X7 - YearRemodAdd	0.81139	0.02657	0.04773	-0.03704	0.30954	-0.03918
X21	X21 - GarageCars	0.78745	-0.04472	0.00657	-0.12152	-0.20797	-0.05991
X4	X4 - OverallQual	0.70012	0.35350	0.26429	0.07642	0.07195	0.11319
X22	X22 - GarageArea	0.65834	0.24834	0.20739	0.28088	-0.10082	0.08757
X23	X23 - WoodDeckSF	0.61169	0.21387	0.27753	0.34007	-0.08673	0.07842
X25	X25 - 3SsnPorch	0.30626	0.21728	0.11126	0.04979	0.16866	0.27329
X19	X19 - Fireplaces	0.16238	0.88990	0.04323	0.11247	0.00720	0.01481
X14	X14 - BsmtFullBath	0.29096	0.84437	0.24718	0.11685	0.03275	0.07043
X17	X17 - KitchenAbvGr	-0.11278	0.78132	-0.18470	0.11347	-0.02350	-0.02738
X16	X16 - BedroomAbvGr	0.57499	0.58847	-0.01871	-0.02722	-0.09761	0.02328
X9	X9 - BsmtUnfSF	0.08632	-0.04483	0.86298	0.06400	-0.03450	-0.01420
X15	X15 - FullBath	0.03468	-0.17334	0.77510	-0.05474	-0.00269	-0.08942
X11	X11 - 1stFlrSF	0.36289	0.18801	0.60899	0.36359	-0.11981	0.06845
X12	X12 - 2ndFlrSF	0.27515	0.33590	0.55888	0.42069	-0.09233	0.07187
X20	X20 - GarageYrBlt	0.08967	0.39400	0.43660	0.13806	0.11007	0.14503
X8	X8 - BsmtFinSF1	0.29375	0.30612	0.32646	0.04343	-0.20176	-0.00762
X24	X24 - OpenPorchSF	0.25266	0.14501	0.32056	0.01427	0.20590	-0.08204
X2	X2 - LotFrontage	0.10238	0.24324	0.05931	0.73220	-0.06420	-0.02880
X3	X3 - LotArea	0.00162	0.34681	0.19199	0.63244	0.04981	0.00565
X1	X1 - MSSubClass	0.05037	0.20329	0.01593	-0.81690	-0.07061	0.00087
X5	X5 - OverallCond	-0.17969	-0.03226	-0.04418	0.01213	0.88981	-0.01937
X26	X26 - MoSold	0.03475	0.00001	-0.02560	-0.00897	0.09664	0.72213

Iteration 5: After deleting these variables, there have been significant changes in the number of factors criteria, so we must revisit the stage 4. Let's have a look at the latent root criterion graph again for the number of factors to extract given the deletion of these variables because of dual loading. Given the deletion of variables, the overall MSA has also decreased to 0.65 but is still satisfactory.

Eigenvalues of the Correlation Matrix: Total = 13 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.10553498	0.98442641	0.2389	0.2389
2	2.12110857	0.73363606	0.1632	0.4020
3	1.38747251	0.24226097	0.1067	0.5088
4	1.14521155	0.13277771	0.0881	0.5969
5	1.01243383	0.09535372	0.0779	0.6748
6	0.91708011	0.06109248	0.0705	0.7453
7	0.85598763	0.08167754	0.0658	0.8111
8	0.77431009	0.14507599	0.0596	0.8707

Latent root criterion mentions 5 factors to extract as appropriate for this analysis, as compared to 6 in the previous iteration.

The below mentioned rotated factor mix with 5 factors and cross loading variables deleted, tells us the variables to be grouped together under 5 factors.

Rotated Factor Pattern						
		Factor1	Factor2	Factor3	Factor4	Factor5
X19	X19 - Fireplaces	0.93540	0.07211	-0.00641	0.00841	-0.01396
X14	X14 - BsmtFullBath	0.88008	0.18285	0.15876	0.03797	-0.04246
X17	X17 - KitchenAbvGr	0.81437	-0.14849	-0.17588	0.01149	0.03118
X21	X21 - GarageCars	0.03138	0.91006	0.00573	-0.02131	-0.01318
X6	X6 - YearBuilt	0.02592	0.88641	0.14918	0.00950	0.02598
X9	X9 - BsmtUnfSF	0.04946	0.10112	0.89606	-0.00261	0.01310
X15	X15 - FullBath	-0.08143	0.05318	0.88684	-0.05229	0.03459
X26	X26 - MoSold	0.03184	-0.00447	-0.00794	0.78024	0.12046
X27	X27 - YrSold	-0.00606	-0.00048	0.03639	-0.72299	0.15402
X24	X24 - OpenPorchSF	0.23765	0.30482	0.25253	0.06605	0.68278
X25	X25 - 3SsnPorch	0.30136	0.30458	0.17917	0.13015	-0.65751

Here is the variance explained by each factor. The size of the communality is a useful index for assessing how much variance in a particular variable is accounted for by the factor solution. Higher communality values indicate that a large amount of the variance in a variable has been extracted by the factor solution. All of the variables here have high communality

Variance Explained by Each Factor				
Factor1	Factor2	Factor3	Factor4	Factor5
2.7098053	2.0942597	1.3478628	1.1412633	0.9167645

Final Communality Estimates: Total = 8.209956										
X6	X9	X14	X15	X17	X19	X21	X24	X25	X26	X27
0.80941954	0.81578122	0.83642405	0.79987061	0.71728289	0.88047306	0.82985953	0.68371848	0.66494822	0.62437710	0.54780095

We have tried to classify these factors into explanatory names in the broader dimensions based on the variables the captured:

Factor 1 Amenities: X19, Fireplace; X14, Basement Full Bath; X17, Kitchens Above Grade.

Factor 2: X21, Garage Cars; X6, Year Built

Factor 3 Product value: X9, Unfinished square feet of basement area; X15, Full Bath

Factor 4 Sold Time: X26, Month Sold; X27, Year Sold.

Factor 5 Porch Area: X24, Open porch area in square feet; X25, Three season porch area in square feet.

Correlations Between Summated Scales and Factors:

Pearson Correlation Coefficients, N = 1460 Prob > |r| under H0: Rho=0						
	Factor1	Factor2	Factor3	SumScale1	SumScale2	SumScale3
Factor1	1.00000	0.00000 1.0000	0.00000 1.0000	0.88127 <.0001	0.03075 0.2402	0.04933 0.0595
Factor2	0.00000 1.0000	1.00000	0.00000 1.0000	0.18223 <.0001	0.97213 <.0001	0.10110 0.0001
Factor3	0.00000 1.0000	0.00000 1.0000	1.00000	0.15793 <.0001	0.09165 0.0005	0.89641 <.0001
SumScale1 SumScale1	0.88127 <.0001	0.18223 <.0001	0.15793 <.0001	1.00000	0.20505 <.0001	0.20733 <.0001
SumScale2 SumScale2	0.03075 0.2402	0.97213 <.0001	0.09165 0.0005	0.20505 <.0001	1.00000	0.20685 <.0001
SumScale3 SumScale3	0.04933 0.0595	0.10110 0.0001	0.89641 <.0001	0.20733 <.0001	0.20685 <.0001	1.00000

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-