R. Farhanian, M. Lee, J.Park MSDS 6372 8 June 2017

Multiple Linear Regression of Ames, Iowa Home Prices

In the real estate industry, a properly priced home can make the difference of a quick sale, a prolonged experience, or no sale at all. Although it is common practice for the home owner to defer to the real estate agent to suggest pricing, there are often many factors, such as square footage, year built, and whether or not the property has a garage to name a few. With so many factors involved pricing can be confusing and difficult. A properly priced home should, deliver a fair value to both parties as well as sell quickly. This study will explore many possible categorical and numerical factors (79 appendiin total) that may affect sales price and attempt to determine the most dominant factors leading to an improved and simple sales price model. The study will first explore possible contributing factors to sales price on a high level then a reduced data set will be studied using various auto-selection techniques to determine an accurate prediction model based on the test data set. Finally, the output of that model will be submitted to the Kaggle.com website and scored against a Mean Squared Error criteria.

Problem Statement

Develop a multiple linear regression model based on an observed set of explanatory variables (the Ames Housing data set) that can easily be used to predict future sales prices without overly complicated interpretation.

Constraints and Limitations

The analysis was completed based on data provided on the Kaggle website based on home sales in Ames, Iowa. Some of the data, as provided, was incomplete and was corrected by the data analysts based on assumptions, contextual fit, and in order to keep the model relatively simple. The data was divided into two data sets. The first data set called ''Train" was utilized to establish the theoretical model based on multiple linear regression statistics. The second data set called "Test" contained separate values without known sales price in order to test the model derived from the "Train" data set. Some categorical factors involved in this data set may contain highly unbalanced levels, meaning it is more challenging to determine the usefulness of the data. These data will be considered for incorporation into the model with caution and also with consideration that the data set may grow in the future. Given the provided data sets this can only be considered an observational study and causal inferences may not be drawn. However, with such large and comprehensive data sets the associations between factors and to sales price is still useful information in understanding pricing strategies and also in designing any follow-on studies. Finally, this study did not include the use of techniques, such as "bootstrapping", which have not yet been covered in this course.

Data Set Description

There were 79 potential factors identified that may be useful in building a sales price regression model. Due to the fact that this data set is common to all of the students in the MSDS 6372 class the factors and descriptions are provided in Appendix A. of this report. Each of these variables could possibly impact the eventual sales price of the property.

Exploratory Data Analysis

Given the large number of numerical and categorical factors, many of which, could provide very similar information, it was decided to first group similar variables in order to discover and get a feel for correlations between both Sales Price and other factors. This information was then used to either eliminate factors or inform the data analysts of possible trends when looking at the full model, once constructed. The groupings and assigned variables were:

1. Area Related: MSSubClass, MSZoning, Neighborhood
2. Lot Parameters: LotFrontage, LotArea, LotShape, LandContour, LotConfig, LandSlope, PavedDrive
3. Condition/Quality: Condition1, Condition2, OverallQual, OverallCond, ExterQual, ExterCond, BsmtQual, BsmtCond, LowQualFinSF, KitchenQual
4. Style: BldgType, HouseStyle, RoofStyle, RoofMatl, Exterior1st, Exterior2nd, MasVnrType, MasVnrArea, Foundation
5. Basement: BsmtExposure, BsmtFinType, BsmtFinSF1, BsmtFinType2, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF
6. Mechanical: Heating, HeatingQC, CentralAir, Electrical
7. Fireplace: Fireplaces, FireplaceQu
8. Interior Size: 1stFlrSF, 2ndFlrSF, GrLivArea
9. Rooms: BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, BedroomAbvGr, Kitche-1bvGr, TotRmsAbvGrd
10. Garage: GarageType, GarageYrBlt, GarageFinish, GarageCars, GaragerArea, GarageQual, GarageCond
11. Outdoors: WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea

12. Misc: Street Utilities YearBuilt YearRemoAdd Functional MiscVal MoSold YrSold SaleType SaleCondition

Initial Exploratory Method of Groups and Factors

First each data set was corrected and cleaned in order to produce more consistent comparisons and predictions. Corrections included properly coding the type of data from a character field to a numerical field, substituting values where 'NA', and substituting values where in some cases there were no values. Each adjustment was completed with the other factors and variables in consideration (ran the model several times to verify corrections). In part 2 the dependent variable, Sale Price, for all of the group studies was log transformed based on early indications of funneling of residuals. This was carried forward in the majority of exploration. Each grouping was evaluated based on the following criterion as appropriate: Correlation between factors within the group, correlation between the factors and sales price/log sale price, scatterplots, first pass intra-group regression model to investigate p-value trends with like factors, and Variable Inflation Factor (VIF) to shed light on possible multi-collinearity as a prelude to the final model. Inspection of residuals will be evaluated once the preliminary full model is established with the exception of the dependent variable "sales price", which was determined very early in the exploratory process to need a log transformation. In the interest of maintaining an efficient read of this information only one group example will be included in the main body of this paper. The remaining are included in Appendix B.

Initial Exploratory Method Example Using the Garage Group

Reference Group Garage: GarageType, GarageYrBlt, GarageFinish, GarageCars, GaragerArea, GarageQual, GarageCond

First a correlation test was run to investigate any possible correlation to log sale price and the other Garage related continuous variables. GarageYrBlt seems to have low correlation to log Sale Price and also low significance (p=0.4581) to log Sale Price. Recommend removing GarageYrBlt in the final model. Finally, GarageArea and GarageCars were scatter plotted against one another since their titles would suggest similar data and possible multi-collinearity. In the event there is further evidence of multi-collinearity, the scatter plots may suggest only keeping one variable. We will keep both for now and re-evaluate in full model.

```
Proc CORR data=train;
VAR LogSalePrice
GarageYrBlt GarageCars
GarageArea;
run;
```

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	logSalePrice	GarageYrBlt	GarageCars	GarageArea
logSalePrice	1.00000 1460	0.54107 <.0001 1379	0.68062 <.0001 1460	0.65089 <.0001 1460
GarageYrBlt	0.54107 <.0001 1379	1.00000 1379	0.58892 <.0001 1379	0.56457 <.0001 1379
GarageCars	0.68062 <.0001 1460	0.58892 <.0001 1379	1.00000 1460	0.88248 <.0001 1460
GarageArea	0.65089 <.0001 1460	0.56457 <.0001 1379	0.88248 <.0001 1460	1.00000 1460

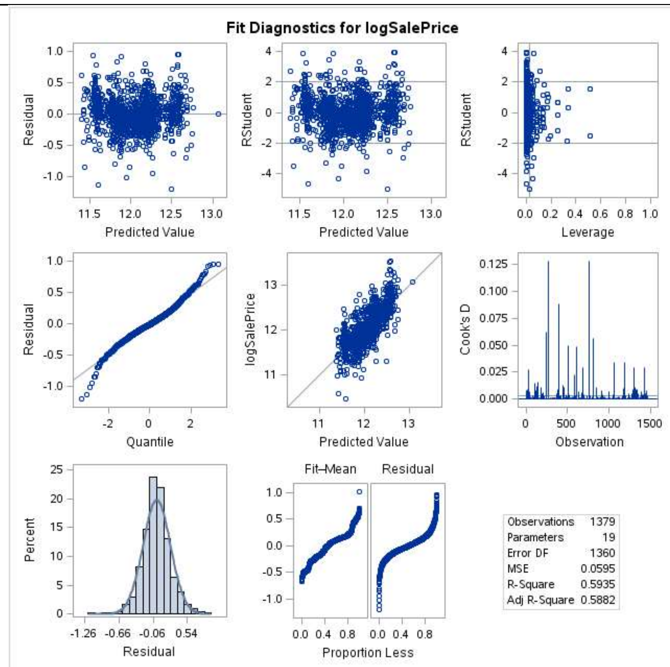
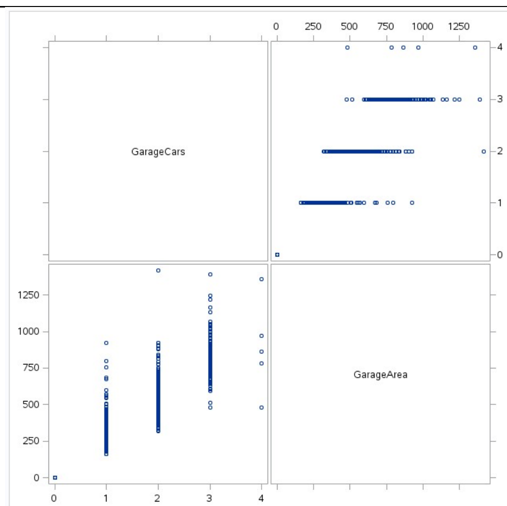
```
proc glm data = train
plots=diagnostics;
class GarageType
GarageFinish GarageQual
GarageCond;
model logsaleprice =
GarageType GarageFinish
GarageQual GarageCond
GarageYrBlt GarageCars
GarageArea;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	118.2059594	6.5669977	110.33	<.0001
Error	1360	80.9494913	0.0595217		
Corrected Total	1378	199.1554508			

R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.593536	2.023761	0.243971	12.05531

Source	DF	Type I SS	Mean Square	F Value	Pr > F
GarageType	5	53.61664125	10.72332825	180.16	<.0001
GarageFinish	2	20.50890840	10.25445420	172.28	<.0001
GarageQual	4	1.32001374	0.33000344	5.54	0.0002
GarageCond	4	2.24433229	0.56108307	9.43	<.0001
GarageYrBlt	1	10.48606013	10.48606013	176.17	<.0001
GarageCars	1	27.08888714	27.08888714	455.11	<.0001
GarageArea	1	2.94111646	2.94111646	49.41	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
GarageType	5	8.27257024	1.65451405	27.80	<.0001
GarageFinish	2	4.23673219	2.11836610	35.59	<.0001
GarageQual	4	1.16129535	0.29032384	4.88	0.0007
GarageCond	4	1.03478563	0.25869641	4.35	0.0017
GarageYrBlt	1	0.03278170	0.03278170	0.55	0.4581
GarageCars	1	4.77137507	4.77137507	80.16	<.0001
GarageArea	1	2.94111646	2.94111646	49.41	<.0001

<table>
<tr><td></td><td>
</td></tr>
<tr><td>proc sgscatter data = train;
matrix GarageCars
GarageArea;
run;</td><td>
</td></tr>
<tr><td>Final Garage group variables selected for incorporation into full model:</td><td>GarageType, GarageFinish, GarageCars, GaragerArea, GarageQual, GarageCond</td></tr>
</table>

Data Cleansing

Cleansing process

The steps were performed to clean the data. Additional, specific, examples may be viewed in Appendix G.

1. We used “proc content”, “proc means”, and “proc print” to have a better understanding of the data.

2. We checked and fixed null, “NA”, “.”, and “-1” and their impact on the full model.
3. Some numerical columns were interpreted as String in some cases. We used input function extensively to fix this issue. After converting the type, we renamed the column, and dropped the original one.
4. We observed that the data of some of the columns were truncated. After doing some research, we figured that the “proc import” has a “guessingrows” argument. It specifies the number of rows of the file to scan to determine the appropriate data type and length for the columns. We used a large number (32676) for “guessingrows”.
5. We went through a trial and error process, and had to go through step 1 to 4 several times.
6. Finally, Kaggle website provided us with which prediction were missing. Subsequently, a code was written to isolate those rows of factors in order to quickly determine what the suspect rows had in common. This was then quickly rectified with the data cleansing steps identified in 1-5 of this section.
7. Other missing or low prediction values were filtered and forcefully assigned to $35,000, which is a round up value of the minimum home sale price from the train data set.

Model Selection

Part 1: Develop a sparse and concise model

Once the data was cleaned and explored to more thoroughly to understand the factors involved, the team proceeded to build a preliminary model based on the reduced number of factors as indicated in the exploration. The full data set with notes to keep (no note indicates keep) or eliminate prior to building this model may be viewed in Appendix C. The preliminary model was run with several auto-selection techniques including forward, stepwise and LASSO. This generated the table viewed in Table 2 below.

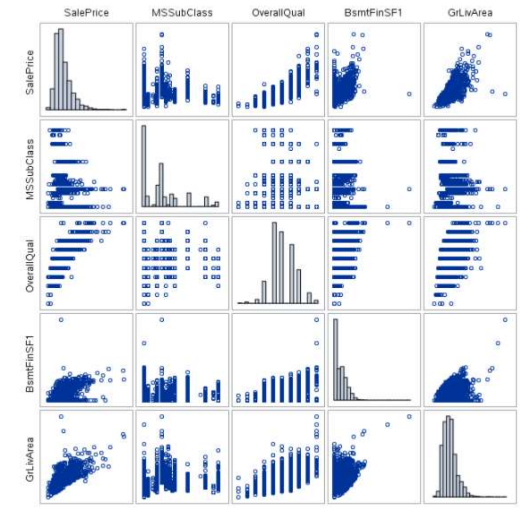
Trials	Effects	R-square	Adj R-Sq
#1. selection=LASSO(choose=SBC stop=SBC)	Intercept OverallQual ExterQual_TA BsmtQual_Ex KitchenQual_Ex _1stFlrSF GrLivArea GarageCars GarageArea	0.7191	0.7176
#2. selection=STEPWISE(choose=SBC stop= SBC)	Intercept Neighborhood MSSubClass LotArea Condition2 OverallQual OverallCond ExterQual BsmtQual KitchenQual RoofMatl MasVnrArea BsmtExposure BsmtFinSF1 BsmtFinSF2	0.9118	0.9072

	BsmtUnfSF _2ndFlrSF GrLivArea BedroomAbvGr KitchenAbvGr GarageArea PoolArea SaleCondition YearBuilt		
#3. selection=FORWARD(choose=SBC stop= SBC)	Intercept Neighborhood MSSubClass LotArea Condition2 OverallQual OverallCond ExterQual BsmtQual KitchenQual RoofMatl MasVnrArea BsmtExposure BsmtFinSF1 BsmtFinSF2 BsmtUnfSF _2ndFlrSF GrLivArea BedroomAbvGr KitchenAbvGr GarageArea PoolArea SaleCondition YearBuilt	0.9118	0.9072
#7. selection = STEPWISE(select = SL SLE = 0.01 SLS = 0.0000000000000000000000001 choose = SBC)	Intercept Neighborhood MSSubClass OverallQual BsmtQual RoofMatl BsmtFinSF1 GrLivArea	0.8609	0.8570

Table 2, Preliminary Models using Auto techniques in SAS

Shwarz Baysian Criteria (SBC) options are highly leveraged to choose the variables in this model since it penalizes for complexity and tends to provide a simple model. The simplest model with the most significant factors appeared to be (#7 in Table 2). This model used the SLE and SLS options which found the most significant variables. As the SLS, the significant level of variables that stay in the model, decreases the lower number of variables are left. The SLS value keeps decreasing until there are only seven factors left.

In further reviewing the data there is a high correlation between ‘OverallQual’ and ‘GrLivArea’ as seen in Table 3. This correlation will be kept in mind as we progress and mature the model.



Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	SalePrice	MSSubClass	OverallQual	BsmtFinSF1	GrLivArea
SalePrice	1.00000 1460	-0.08428 0.0013 1460	0.79098 <.0001 1460	0.38642 <.0001 1460	0.70862 <.0001 1460
MSSubClass	-0.08428 0.0013 1460	1.00000 2919	0.03364 0.0692 2919	-0.06400 0.0005 2919	0.07168 0.0001 2919
OverallQual	0.79098 <.0001 1460	0.03364 0.0692 2919	1.00000 2919	0.28215 <.0001 2919	0.57513 <.0001 2919
BsmtFinSF1	0.38642 <.0001 1460	-0.06400 0.0005 2919	0.28215 <.0001 2919	1.00000 2919	0.21198 <.0001 2919
GrLivArea	0.70862 <.0001 1460	0.07168 0.0001 2919	0.57513 <.0001 2919	0.21198 <.0001 2919	1.00000 2919

Table 3, Correlation and Scatter plots of auto-selected model factors

In addition a class variable association/Chi-square test shows there is a significant association between 'Neighborhood' and 'BsmtQual'. As seen in Table 4, since there are multiple levels F-values are utilized to remove one of the class variable, and the 'Neighborhood' is removed with lower F-value than 'BsmtQual'.

```
proc freq data = train3;
table BsmtQual*RoofMatl / chisq ;
run;quit;
```

Statistics for Table of BsmtQual by RoofMatl

Statistic	DF	Value	Prob
Chi-Square	28	34.0095	0.2006
Likelihood Ratio Chi-Square	28	31.5555	0.2930
Mantel-Haenszel Chi-Square	1	0.8486	0.3570
Phi Coefficient		0.1079	
Contingency Coefficient		0.1073	
Cramer's V		0.0540	
WARNING: 83% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Table 4, Chi-Square test for factor association

The final base model (no transformations, etc) was rerun with the aforementioned variables removed. The model that explains a decent amount of the variation (near 80%) in the home sale price with only the most significant variables is shown below.

SalePrice = -50550 + 70277*BsmtQual_Ex -4776*BsmtQual_Fa + 19341*BsmtQual_Gd +2893*BsmtQual_Na -228*MSSubClass +22363*OverallQual +23.66*BsmtFinSF1 +55.15*GrLivArea;

This model is highly significant (p-value < 0.001), and this model explains 79% of variation in the sale price in the train data set (R-Square =0.79).

Interpretation of the model is same as followings. With everything remaining constant when 'BsmtFinSF1' increases by 1 unit (SqFt) then the mean home sale price will rise by $23.66. The same rule applies to other continuous variables with different units and coefficients. For the class variable, 'BsmtQual', the mean home price will go up by the coefficients of the corresponding quality level. For example, if the basement quality is excellent (EX) then the mean home sale price will go up by $70,277. Appendix D provides additional code and output information relating to our development and decisions in the Part 1 Base Model Development.

The referenced Kaggle score in Figure 1 verifies that the model is working with a score of 0.17250.

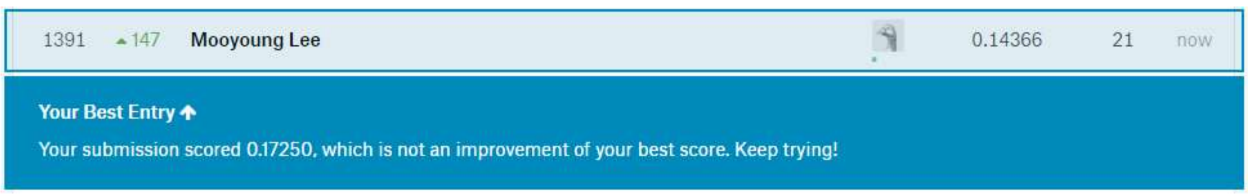


Figure 1, Base model (no transformations, etc.) Kaggle score

Part 2: Strategy and final model build that would provide the best predictions

With the base model determined we now know a lot more about the factors involved, their correlation to each other and to the Sales Price. In this section we will attempt to further improve the model by investigating a log transformation of variables to remove variation, re-investigation of correlation, multi-collinearity, and a higher ordered polynomial term. The objective of this part2 to design a model that describes the given home price values in the train data set, then to test that model against the test data set. The complexity of the model can be increased as compared to the base model developed in Part 1 in order to improve fit. The final model was verified via the Kaggle website. Fitting to the Kaggle data set is not our priority but designing a best fit model using a logical statistical reasoning is our goal here. The detailed procedure to develop the final model is outlined in the 12 steps seen in Appendix F. This approach further improved the Kaggle score to 0.14582 which shows what even a minor amount of factor manipulation may cause the model to be more predictive.

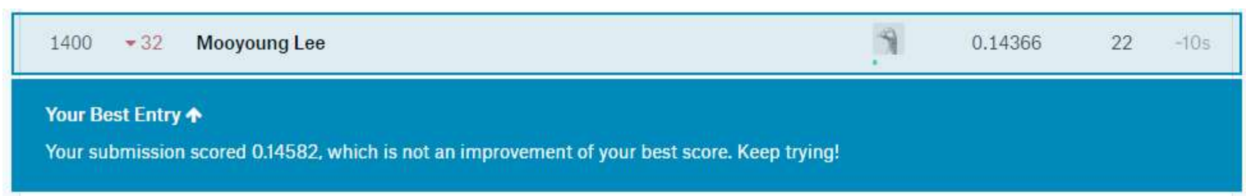


Figure 2. Final Model Kaggle Score

Conclusions

In this study the team conducted extensive data cleaning and explored many options while investigating which factors were most important in predicting Sales Price. There were several challenges along the way. Particularly with incomplete, mis-labeled, or truncated variables in the original data sets. These were all corrected along the way using various techniques learned in the course and with other methods. In the end the initial base model, in Part 1, did fairly well with a Kaggle score of 0.17250. This experience taught the team that a simple method may provide a reasonably predictive model, even without much data manipulation. In our second attempt, in Part 2, the team investigated several transformations, and other interactions and the Kaggle score was improved to 0.14582. The goal of this was to determine if the model could be significantly improved through some broad steps and to learn efficient methods of model building. In addition the team learned that cleaning the data should first be fully verified prior to progressing ahead and building the model. Once the preliminary model was established the team also found there was a significant amount of trial and error and other data manipulation test necessary in order to further improve the fit. Finally, the team discovered that highly complex auto-selection techniques do a great job of getting us very close to a solution. The final improvements, after auto-correction, are much more time consuming to obtain without learning more advanced methods.

Appendix A.

The following are the factor names and descriptions:

- SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of alley access
- LotShape: General shape of property
- LandContour: Flatness of the property
- Utilities: Type of utilities available
- LotConfig: Lot configuration
- LandSlope: Slope of property
- Neighborhood: Physical locations within Ames city limits
- Condition1: Proximity to main road or railroad
- Condition2: Proximity to main road or railroad (if a second is present)
- BldgType: Type of dwelling
- HouseStyle: Style of dwelling
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- RoofStyle: Type of roof
- RoofMatl: Roof material
- Exterior1st: Exterior covering on house
- Exterior2nd: Exterior covering on house (if more than one material)
- MasVnrType: Masonry veneer type
- MasVnrArea: Masonry veneer area in square feet
- ExterQual: Exterior material quality
- ExterCond: Present condition of the material on the exterior
- Foundation: Type of foundation
- BsmtQual: Height of the basement
- BsmtCond: General condition of the basement
- BsmtExposure: Walkout or garden level basement walls
- BsmtFinType1: Quality of basement finished area
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinType2: Quality of second finished area (if present)
- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area
- Heating: Type of heating
- HeatingQC: Heating quality and condition
- CentralAir: Central air conditioning
- Electrical: Electrical system
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- GrLivArea: Above grade (ground) living area square feet
- BsmtFullBath: Basement full bathrooms

- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- Bedroom: Number of bedrooms above basement level
- Kitchen: Number of kitchens
- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Functional: Home functionality rating
- Fireplaces: Number of fireplaces
- FireplaceQu: Fireplace quality
- GarageType: Garage location
- GarageYrBlt: Year garage was built
- GarageFinish: Interior finish of the garage
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- GarageQual: Garage quality
- GarageCond: Garage condition
- PavedDrive: Paved driveway
- WoodDeckSF: Wood deck area in square feet
- OpenPorchSF: Open porch area in square feet
- EnclosedPorch: Enclosed porch area in square feet
- 3SsnPorch: Three season porch area in square feet
- ScreenPorch: Screen porch area in square feet
- PoolArea: Pool area in square feet
- PoolQC: Pool quality
- Fence: Fence quality
- MiscFeature: Miscellaneous feature not covered in other categories
- MiscVal: $Value of miscellaneous feature
- MoSold: Month Sold
- YrSold: Year Sold
- SaleType: Type of sale
- SaleCondition: Condition of sale

Appendix B

1. Mechanical: Heating, HeatingQC, CentralAir, Electrical

First a correlation test was run to investigate any possible correlation to log sale price and the other Mechanical related variables. Heating, HeatingQC, CentralAir, and Electrical are all categorical. They are all recommended to incorporate in the final model. But we will keep in mind that Heating, Electrical and Central Air have 40% multicollinearity in their relevant group.

```
proc corr data =
mechanicalData;
title1
"Mechanical";
run;
```

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	logSalePrice	HeatingType	HeatingQCGroup	ElectricalGroup	CentralAirGroup
logSalePrice	1.00000 1460	-0.10228 <.0001 1457	0.45034 <.0001 1460	-0.30083 <.0001 1459	-0.35160 <.0001 1460
HeatingType	-0.10228 <.0001 1457	1.00000 1457	-0.08228 0.0017 1457	0.19303 <.0001 1456	0.40066 <.0001 1457
HeatingQCGroup	0.45034 <.0001 1460	-0.08228 0.0017 1457	1.00000 1460	-0.14504 <.0001 1459	-0.18213 <.0001 1460
ElectricalGroup	-0.30083 <.0001 1459	0.19303 <.0001 1456	-0.14504 <.0001 1459	1.00000 1459	0.39783 <.0001 1459
CentralAirGroup	-0.35160 <.0001 1460	0.40066 <.0001 1457	-0.18213 <.0001 1460	0.39783 <.0001 1459	1.00000 1460

2. Interior Size: 1stFlrSF, 2ndFlrSF, GrLivArea

First a correlation test was run to determine if the continuous factors were correlated with one another and to investigate potential correlation to the log of the sale price variable. The log of Sale Price was used in all investigations as this produced a more normal residual condition as seen the output below. The variables 1stFlrSF, 2ndFlrSF, GrLivArea will be incorporated in the full model. But we will consider that 1stFlrSF and 2ndFlrSF have both more than 50% multicollinearity with GrLivArea. Regression does not prove the significance of the variable; however, I ran it to investigate significance of each factor to the log of the sale price. Keep for now.

```
proc corr data =
livingAreaData;
run;
```

Pearson Correlation Coefficients, N = 1460
Prob > |r| under H0: Rho=0

	_1stFlrSF	_2ndFlrSF	GrLivArea	logSalePrice
_1stFlrSF	1.00000	-0.20265 <.0001	0.56602 <.0001	0.59698 <.0001
_2ndFlrSF	-0.20265 <.0001	1.00000	0.68750 <.0001	0.31930 <.0001
GrLivArea	0.56602 <.0001	0.68750 <.0001	1.00000	0.70093 <.0001
logSalePrice	0.59698 <.0001	0.31930 <.0001	0.70093 <.0001	1.00000

```
proc glm data =
livingAreaData
plots=diagnostics;
model logSaleprice =
_1stFlrSF _2ndFlrSF
GrLivArea /clparm
solution;
run; quit;
```

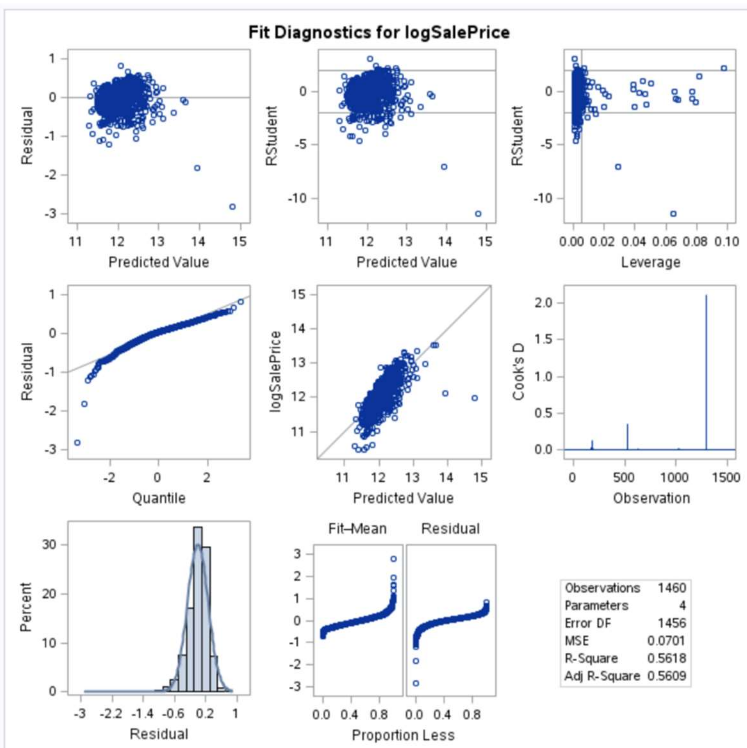
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	130.7916334	43.5972111	622.27	<.0001
Error	1456	102.0090256	0.0700611		
Corrected Total	1459	232.8006590			

R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.561818	2.201343	0.264691	12.02405

Source	DF	Type I SS	Mean Square	F Value	Pr > F
_1stFlrSF	1	82.96698136	82.96698136	1184.21	<.0001
_2ndFlrSF	1	47.05926105	47.05926105	671.69	<.0001
GrLivArea	1	0.76539097	0.76539097	10.92	0.0010

Source	DF	Type III SS	Mean Square	F Value	Pr > F
_1stFlrSF	1	4.74717980	4.74717980	67.76	<.0001
_2ndFlrSF	1	2.68234743	2.68234743	38.29	<.0001
GrLivArea	1	0.76539097	0.76539097	10.92	0.0010

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	11.05104337	0.02411076	458.34	<.0001	11.00374783	11.09833891
_1stFlrSF	0.00118490	0.00014395	8.23	<.0001	0.00090253	0.00146727
_2ndFlrSF	0.00089545	0.00014472	6.19	<.0001	0.00061157	0.00117933
GrLivArea	-0.00047200	0.00014280	-3.31	0.0010	-0.00075213	-0.00019188



3. Rooms: BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, BedroomAbvGr, TotRmsAbvGrd

First a correlation test was run to determine if the non-categorical (continuous) factors were correlated with one another and to investigate potential correlation to the log of the sale price variable. The variable Fireplaces and FireplaceQu were not removed from the model. Finally, a regression of only the Fireplace variables was run to investigate significance of each factor to the log of the sale price. Keep for now.

```
proc corr data =
basementData;
run;
```

Pearson Correlation Coefficients, N = 1460
Prob > |r| under H0: Rho=0

	BsmtFullBath	BsmtHalfBath	FullBath	HalfBath	BedroomAbvGr	TotRmsAbvGrd	logSalePrice
BsmtFullBath	1.00000	-0.14787 <.0001	-0.06451 0.0137	-0.03090 0.2379	-0.15067 <.0001	-0.05328 0.0418	0.23622 <.0001
BsmtHalfBath	-0.14787 <.0001	1.00000	-0.05454 0.0372	-0.01234 0.6376	0.04652 0.0756	-0.02384 0.3628	-0.00515 0.8442
FullBath	-0.06451 0.0137	-0.05454 0.0372	1.00000	0.13638 <.0001	0.36325 <.0001	0.55478 <.0001	0.59477 <.0001
HalfBath	-0.03090 0.2379	-0.01234 0.6376	0.13638 <.0001	1.00000	0.22665 <.0001	0.34341 <.0001	0.31398 <.0001
BedroomAbvGr	-0.15067 <.0001	0.04652 0.0756	0.36325 <.0001	0.22665 <.0001	1.00000	0.67662 <.0001	0.20904 <.0001
TotRmsAbvGrd	-0.05328 0.0418	-0.02384 0.3628	0.55478 <.0001	0.34341 <.0001	0.67662 <.0001	1.00000	0.53442 <.0001
logSalePrice	0.23622 <.0001	-0.00515 0.8442	0.59477 <.0001	0.31398 <.0001	0.20904 <.0001	0.53442 <.0001	1.00000

4. Fireplace: Fireplaces FireplaceQu

First a correlation test was run to determine if the non-categorical (continuous) factors were correlated with one another and to investigate potential correlation to the log of the sale price variable. The log of Sale Price was used in all investigations as this produced a more normal residual condition as seen the output below. The variable Fireplaces and FireplaceQu were not removed from the model. Finally, a regression of only the Fireplace variables was run to investigate significance of each factor to the log of the sale price. Keep for now.

```
proc corr data =
firePlaceData;
run;
```

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Fireplaces	logSalePrice	FireplaceQuGroup
Fireplaces	1.00000 1460	0.48945 <.0001 1460	-0.01414 0.6951 770
logSalePrice	0.48945 <.0001 1460	1.00000 1460	0.06561 0.0688 770
FireplaceQuGroup	-0.01414 0.6951 770	0.06561 0.0688 770	1.00000 770

```
proc glm data =
firePlaceData
plots=diagnostics;
class FireplaceQu;
model logsaleprice =
Fireplaces /clparm
solution;
run; quit;
```

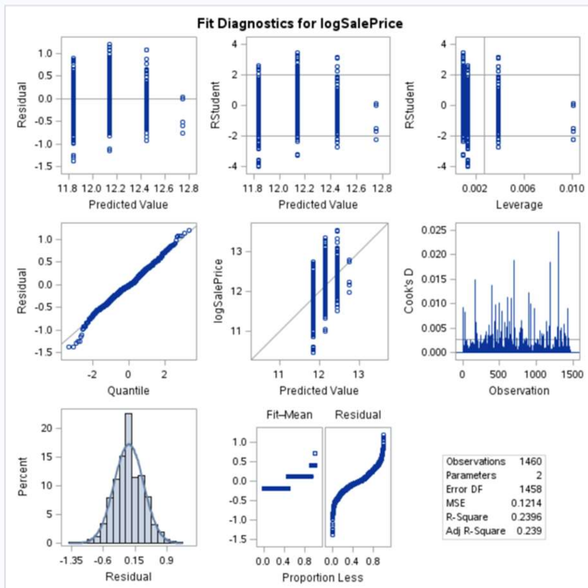
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	55.7698986	55.7698986	459.31	<.0001
Error	1458	177.0307603	0.1214203		
Corrected Total	1459	232.8006590			

R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.239561	2.897976	0.348454	12.02405

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Fireplaces	1	55.76989865	55.76989865	459.31	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Fireplaces	1	55.76989865	55.76989865	459.31	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	11.83813890	0.01258628	940.56	<.0001	11.81344974	11.86282805
Fireplaces	0.30327545	0.01415086	21.43	<.0001	0.27551722	0.33103368



Variance Inflation Factor (VIF) was checked for the Fireplaces variable.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	11.83814	0.01259	940.56	<.0001	0
Fireplaces	1	0.30328	0.01415	21.43	<.0001	1.00000

Final Fireplace variables selected for incorporation into full model:	Fireplaces and FireplaceQu (categorical)

5. Area Related: MSSubClass, MSZoning, Neighborhood

Checked the correlation coefficients among numerical variables, and F-test is performed to check the significance of the variables in a model. MSSubClass variable seems the least significant variable from this group. All three variables (MSSubClass, MSZoning, Neighborhood) will be selected since all individual p-values are lower than 0.05.

```
proc corr data = train3
pearson plots = all;
var SalePrice logSalePrice
MSSubClass
logMSSubClass;
run;
```

Pearson Correlation Coefficients, N = 1460
Prob > |r| under H0: Rho=0

	SalePrice	logSalePrice	MSSubClass	logMSSubClass
SalePrice	1.00000	0.94837 <.0001	-0.08428 0.0013	-0.03361 0.1993
logSalePrice	0.94837 <.0001	1.00000	-0.07396 0.0047	-0.01976 0.4505
MSSubClass	-0.08428 0.0013	-0.07396 0.0047	1.00000	0.93947 <.0001
logMSSubClass	-0.03361 0.1993	-0.01976 0.4505	0.93947 <.0001	1.00000

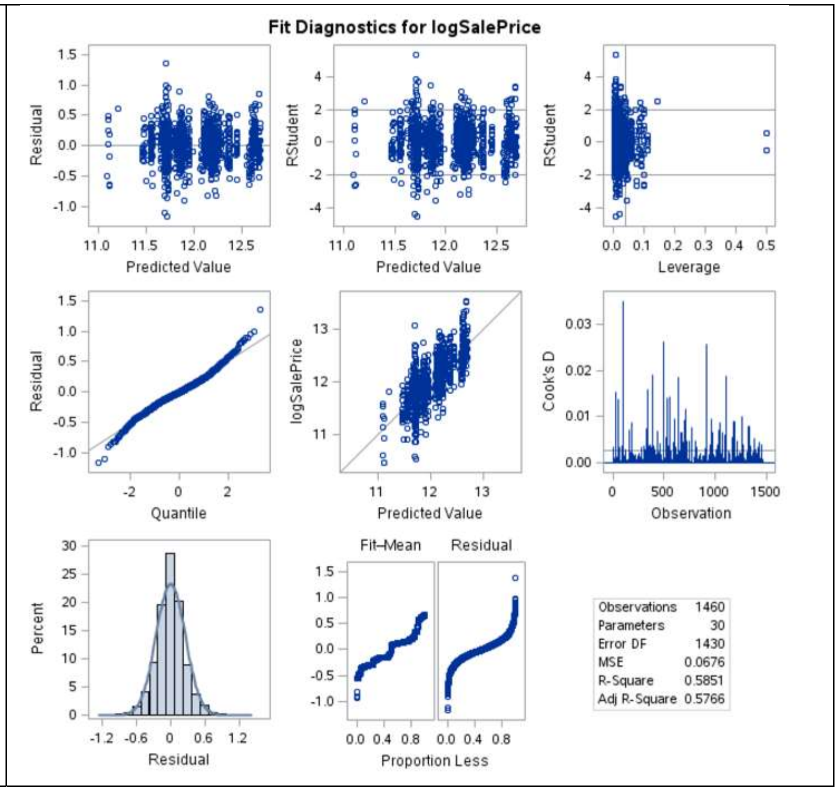
```
proc glm data = train3 plots =
diagnostics;
class MSZoning
Neighborhood;
model logSalePrice =
MSZoning Neighborhood
MSSubClass/ clparm
solution;
run;quit;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	29	136.2009179	4.6965834	69.53	<.0001
Error	1430	96.5997411	0.0675523		
Corrected Total	1459	232.8006590			

R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.585054	2.161569	0.259908	12.02405

Source	DF	Type I SS	Mean Square	F Value	Pr > F
MSZoning	4	40.93539333	10.23384833	151.50	<.0001
Neighborhood	24	94.95023487	3.95625979	58.57	<.0001
MSSubClass	1	0.31528966	0.31528966	4.67	0.0309

Source	DF	Type III SS	Mean Square	F Value	Pr > F
MSZoning	4	2.71994712	0.67998678	10.07	<.0001
Neighborhood	24	95.21360099	3.96723337	58.73	<.0001
MSSubClass	1	0.31528966	0.31528966	4.67	0.0309



6. Lot Parameters: LotFrontage, LotArea, LotShape, LandContour, LotConfig, LandSlope, PavedDrive

LotConfig variable will be eliminated for further analysis since both type I and III SS shows insignificant p-values.

```
proc corr data = train3 pearson
plots = all;
var SalePrice logSalePrice
LotFrontage logLotFrontage
LotArea logLotArea;
run;
```

Pearson Correlation Coefficients, N = 1460
Prob > |r| under H0: Rho=0

	SalePrice	logSalePrice	LotFrontage	logLotFrontage	LotArea	logLotArea
SalePrice	1.00000	0.94837 <.0001	0.02583 0.3240	-0.02630 0.3154	0.26384 <.0001	0.38852 <.0001
logSalePrice	0.94837 <.0001	1.00000	0.03157 0.2280	-0.04701 0.0725	0.25732 <.0001	0.39992 <.0001
LotFrontage	0.02583 0.3240	0.03157 0.2280	1.00000	0.91309 <.0001	-0.09804 0.0002	0.04176 0.1107
logLotFrontage	-0.02630 0.3154	-0.04701 0.0725	0.91309 <.0001	1.00000	-0.13563 <.0001	-0.09357 0.0003
LotArea	0.26384 <.0001	0.25732 <.0001	-0.09804 0.0002	-0.13563 <.0001	1.00000	0.69795 <.0001
logLotArea	0.38852 <.0001	0.39992 <.0001	0.04176 0.1107	-0.09357 0.0003	0.69795 <.0001	1.00000

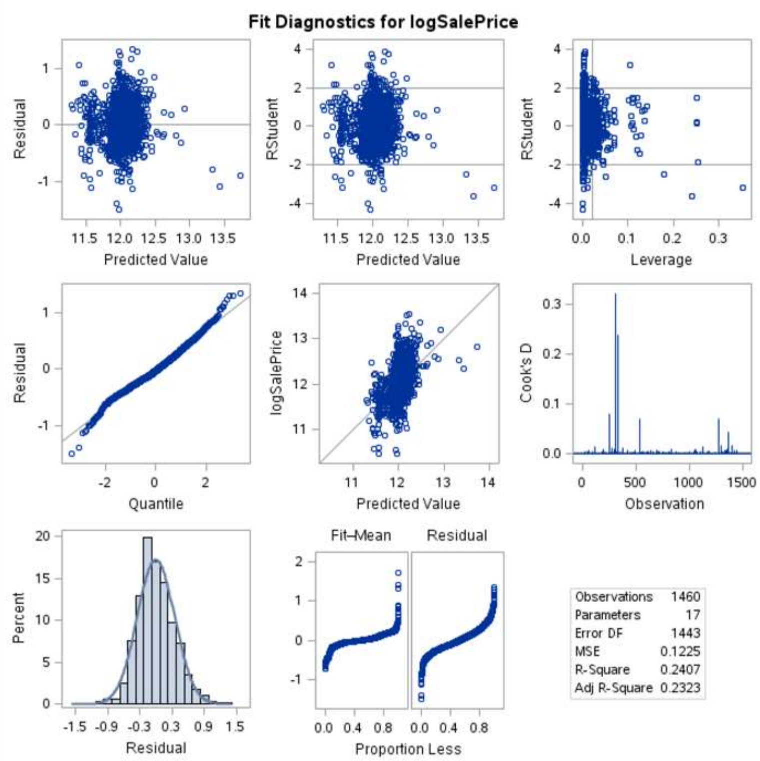
```
proc glm data = train3 plots = diagnostics;
class LotShape LandContour LotConfig LandSlope PavedDrive;
model logSalePrice = LotFrontage LotArea LotShape LandContour LotConfig LandSlope PavedDrive/ clparm solution;
run;quit;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	56.0301950	3.5018872	28.59	<.0001
Error	1443	176.7704640	0.1225021		
Corrected Total	1459	232.8006590			

R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.240679	2.910857	0.350003	12.02405

Source	DF	Type I SS	Mean Square	F Value	Pr > F
LotFrontage	1	0.21970778	0.21970778	1.79	0.1807
LotArea	1	15.93476342	15.93476342	130.08	<.0001
LotShape	3	16.63736580	5.54578860	45.27	<.0001
LandContour	3	4.44818006	1.48272669	12.10	<.0001
LotConfig	4	0.54503364	0.13625841	1.11	0.3491
LandSlope	2	3.16297388	1.58148694	12.91	<.0001
PavedDrive	2	15.08217041	7.54108521	61.56	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
LotFrontage	1	2.40285874	2.40285874	19.61	<.0001
LotArea	1	12.48191058	12.48191058	101.89	<.0001
LotShape	3	9.47086611	3.15695537	25.77	<.0001
LandContour	3	1.95277536	0.65092512	5.31	0.0012
LotConfig	4	0.56233115	0.14058279	1.15	0.3324
LandSlope	2	3.18589456	1.59294728	13.00	<.0001
PavedDrive	2	15.08217041	7.54108521	61.56	<.0001



7. Condition/Quality: Condition1, Condition2, OverallQual, OverallCond, ExterQual, ExterCond, BsmtQual, BsmtCond, LowQualFinSF, KitchenQual

BsmtCond and LowQualFinSF variables will be eliminated since the p-values from both Type I and III SS are insignificant.

```
proc corr data = train3 pearson
plots = all;
var SalePrice logSalePrice
OverallQual OverallCond
LowQualFinSF
logLowQualFinSF;
run;
```

Pearson Correlation Coefficients, N = 1460
Prob > |r| under H0: Rho=0

	SalePrice	logSalePrice	OverallQual	OverallCond	LowQualFinSF	logLowQualFinSF
SalePrice	1.00000	0.94837 <.0001	0.79098 <.0001	-0.07786 0.0029	-0.02554 0.3295	-0.04434 0.0903
logSalePrice	0.94837 <.0001	1.00000	0.81718 <.0001	-0.03687 0.1591	-0.03790 0.1478	-0.05450 0.0373
OverallQual	0.79098 <.0001	0.81718 <.0001	1.00000	-0.09193 0.0004	-0.03043 0.2453	-0.02993 0.2530
OverallCond	-0.07786 0.0029	-0.03687 0.1591	-0.09193 0.0004	1.00000	0.02548 0.3306	0.02859 0.2750
LowQualFinSF	-0.02554 0.3295	-0.03790 0.1478	-0.03043 0.2453	0.02548 0.3306	1.00000	0.94010 <.0001
logLowQualFinSF	-0.04434 0.0903	-0.05450 0.0373	-0.02993 0.2530	0.02859 0.2750	0.94010 <.0001	1.00000

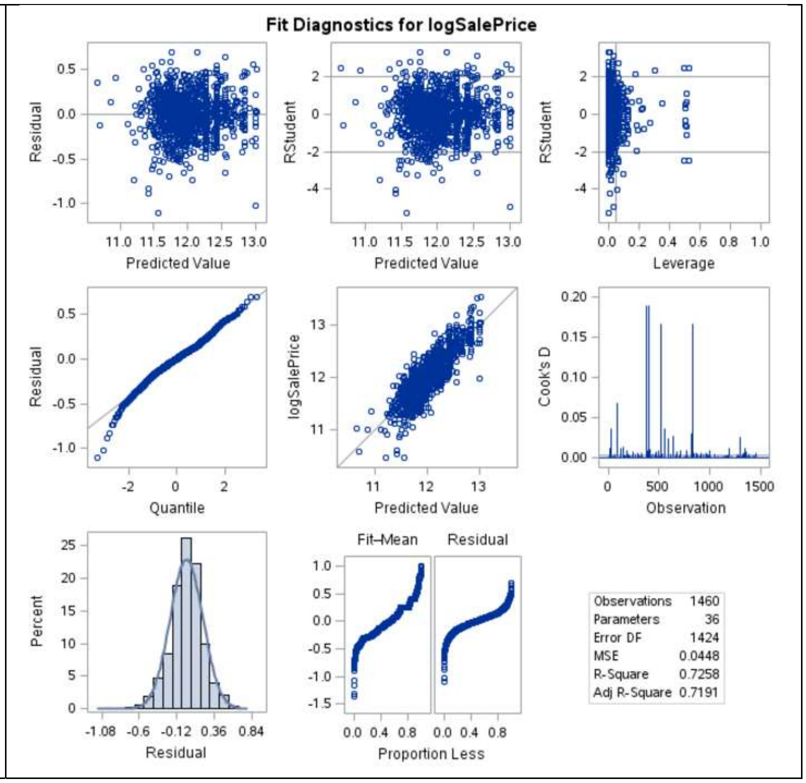
```
proc glm data = train3 plots =
diagnostics;
class Condition1 Condition2
ExterQual ExterCond BsmtQual
BsmtCond KitchenQual;
model logSalePrice =
Condition1 Condition2
OverallQual OverallCond
ExterQual ExterCond BsmtQual
BsmtCond LowQualFinSF
KitchenQual
/ clparm solution;
run;quit;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	35	168.9742219	4.8278349	107.71	<.0001
Error	1424	63.8264370	0.0448219		
Corrected Total	1459	232.8006590			

R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.725832	1.760737	0.211712	12.02405

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Condition1	8	9.8786519	1.2348315	27.55	<.0001
Condition2	7	2.5246741	0.3606677	8.05	<.0001
OverallQual	1	145.9542109	145.9542109	3256.31	<.0001
OverallCond	1	0.4724541	0.4724541	10.54	0.0012
ExterQual	3	3.7655182	1.2551727	28.00	<.0001
ExterCond	4	0.8198059	0.2049515	4.57	0.0011
BsmtQual	4	3.6276362	0.9069091	20.23	<.0001
BsmtCond	3	0.1404881	0.0468294	1.04	0.3717
LowQualFinSF	1	0.0465047	0.0465047	1.04	0.3086
KitchenQual	3	1.7442777	0.5814259	12.97	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Condition1	8	1.19176008	0.14897001	3.32	0.0009
Condition2	7	1.10618831	0.15802690	3.53	0.0009
OverallQual	1	25.62253346	25.62253346	571.65	<.0001
OverallCond	1	0.63470474	0.63470474	14.16	0.0002
ExterQual	3	0.34964110	0.11654703	2.60	0.0507
ExterCond	4	0.43931006	0.10982752	2.45	0.0444
BsmtQual	3	2.61058995	0.87019665	19.41	<.0001
BsmtCond	3	0.13629375	0.04543125	1.01	0.3857
LowQualFinSF	1	0.04512670	0.04512670	1.01	0.3158
KitchenQual	3	1.74427772	0.58142591	12.97	<.0001



8. Style: BldgType, HouseStyle, RoofStyle, RoofMatl, Exterior1st, Exterior2nd, MasVnrType, MasVnrArea, Foundation

Will keep all variables except Exterior2nd since the type III p-value is insignificant.

```
proc corr data = train3 pearson
plots = all;
var SalePrice logSalePrice
MasVnrArea logMasVnrArea;
run;
```

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	SalePrice	logSalePrice	MasVnrArea	logMasVnrArea
SalePrice	1.00000 1460	0.94837 <.0001 1460	0.47739 <.0001 1452	0.41117 <.0001 1452
logSalePrice	0.94837 <.0001 1460	1.00000 1460	0.43056 <.0001 1452	0.41875 <.0001 1452
MasVnrArea	0.47739 <.0001 1452	0.43056 <.0001 1452	1.00000 1452	0.80164 <.0001 1452
logMasVnrArea	0.41117 <.0001 1452	0.41875 <.0001 1452	0.80164 <.0001 1452	1.00000 1452

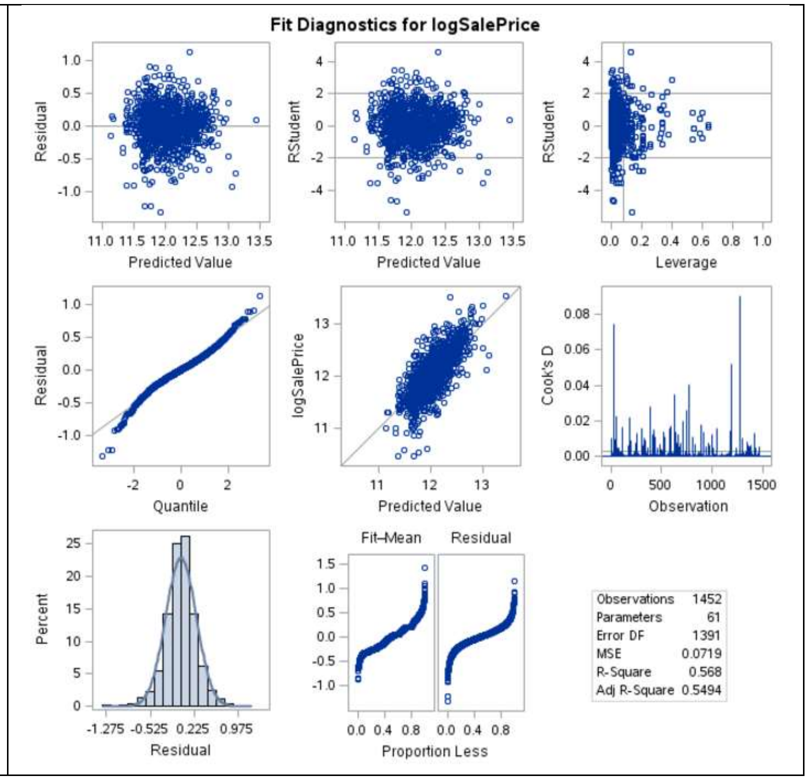
```
proc glm data = train3 plots =
diagnostics;
class BldgType HouseStyle
RoofStyle RoofMatl Exterior1st
Exterior2nd MasVnrType
Foundation;
model logSalePrice = BldgType
HouseStyle RoofStyle RoofMatl
Exterior1st Exterior2nd
MasVnrType MasVnrArea
Foundation
/ clparm solution;
run;quit;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	60	131.4456466	2.1907608	30.48	<.0001
Error	1391	99.9765604	0.0718739		
Corrected Total	1451	231.4222070			

R-Square	Coeff Var	Root MSE	log SalePrice Mean
0.567991	2.229947	0.268093	12.02239

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BldgType	4	9.23218981	2.30804745	32.11	<.0001
HouseStyle	7	25.13538774	3.59076968	49.96	<.0001
RoofStyle	5	10.34637928	2.06927586	28.79	<.0001
RoofMatl	7	3.58275071	0.51182153	7.12	<.0001
Exterior1st	14	30.53550398	2.18110743	30.35	<.0001
Exterior2nd	14	2.08515872	0.14893991	2.07	0.0110
MasVnrType	3	24.33149268	8.11049756	112.84	<.0001
MasVnrArea	1	5.05784545	5.05784545	70.37	<.0001
Foundation	5	21.13893826	4.22778765	58.82	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BldgType	4	8.91857416	2.22964354	31.02	<.0001
HouseStyle	7	6.75495651	0.96499379	13.43	<.0001
RoofStyle	5	2.46213833	0.49242767	6.85	<.0001
RoofMatl	7	3.28969005	0.46995572	6.54	<.0001
Exterior1st	13	5.68413159	0.43724089	6.08	<.0001
Exterior2nd	14	0.96559910	0.06897136	0.96	0.4932
MasVnrType	3	6.04386448	2.01462149	28.03	<.0001
MasVnrArea	1	4.44834195	4.44834195	61.89	<.0001
Foundation	5	21.13893826	4.22778765	58.82	<.0001



9. Basement: BsmtExposure, BsmtFinType1, BsmtFinType2, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, and TotalBsmtSF

First a correlation test was run to determine if the non-categorical (continuous) factors were correlated with one another and to investigate potential correlation to the log of the sale price variable. The log of Sale Price was used in all investigations as this produced a more normal residual condition as seen the output below. Next, the variable TotalBsmtSF was removed for subsequent runs as it was found to be merely a sum of three other variables (BsmtFinSF1, BsmtFinSF2, and BsmtUnfSF). Finally, a regression of only the basement variables was run to investigate significance of each factor to the sale price. BsmtFinType2 was found not be significant (p=0.1090). Keep for now.

```
Proc CORR data=trainJP;
VAR LogSalePrice BsmtFinSF1
BsmtFinSF2 BsmtUnfSF
TotalBsmtSF;
run;
```

Pearson Correlation Coefficients, N = 1460
Prob > |r| under H0: Rho=0

	logSalePrice	BsmtFinSF1	BsmtFinSF2	BsmtUnfSF	TotalBsmtSF
logSalePrice	1.00000	0.37202 <.0001	0.00483 0.8536	0.22199 <.0001	0.61213 <.0001
BsmtFinSF1	0.37202 <.0001	1.00000	-0.05012 0.0556	-0.49525 <.0001	0.52240 <.0001
BsmtFinSF2	0.00483 0.8536	-0.05012 0.0556	1.00000	-0.20929 <.0001	0.10481 <.0001
BsmtUnfSF	0.22199 <.0001	-0.49525 <.0001	-0.20929 <.0001	1.00000	0.41536 <.0001
TotalBsmtSF	0.61213 <.0001	0.52240 <.0001	0.10481 <.0001	0.41536 <.0001	1.00000

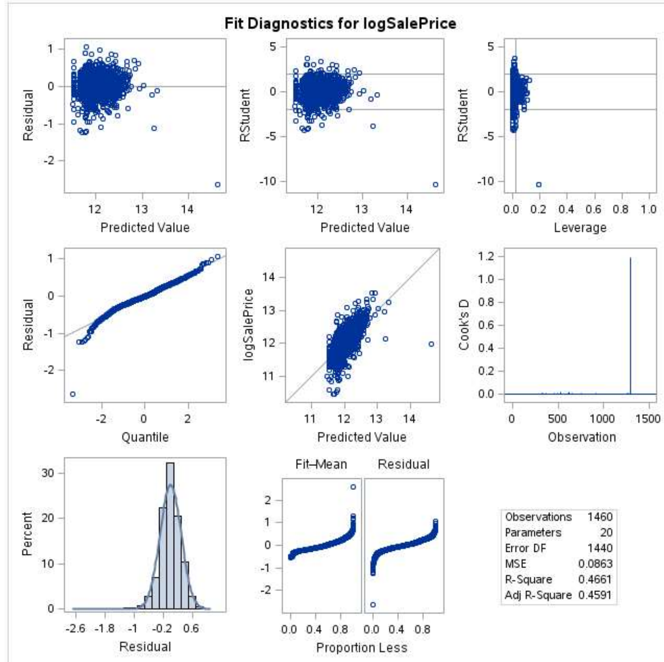
```
proc glm data = trainJP
plots=diagnostics;
class BsmtFinType1
BsmtFinType2 BsmtExposure;
model logsaleprice = BsmtUnfSF
BsmtFinType1 BsmtFinType2
BsmtExposure BsmtFinSF1
BsmtFinSF2 /clparm solution;
run; quit;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	19	108.5181510	5.7114816	66.18	<.0001
Error	1440	124.2825080	0.0863073		
Corrected Total	1459	232.8006590			

R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.466142	2.443278	0.293781	12.02405

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BsmtUnfSF	1	11.47180281	11.47180281	132.92	<.0001
BsmtFinType1	6	60.45381378	10.07563563	116.74	<.0001
BsmtFinType2	6	2.43353525	0.40558921	4.70	<.0001
BsmtExposure	4	8.43459919	2.10864980	24.43	<.0001
BsmtFinSF1	1	23.46211247	23.46211247	271.84	<.0001
BsmtFinSF2	1	2.26228754	2.26228754	26.21	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BsmtUnfSF	1	35.31505333	35.31505333	409.18	<.0001
BsmtFinType1	6	14.23605924	2.37267654	27.49	<.0001
BsmtFinType2	6	0.89917666	0.14986278	1.74	0.1090
BsmtExposure	4	1.40014915	0.35003729	4.06	0.0028
BsmtFinSF1	1	24.91089994	24.91089994	288.63	<.0001
BsmtFinSF2	1	2.26228754	2.26228754	26.21	<.0001

<table>
<tr>
<td></td>
<td>


</td>
</tr>
<tr>
<td>Variance Inflation Factor (VIF) was checked for quantifiable variables. There is further evidence that the TotalBsmtSF is redundant.</td>
<td>
<table>
<tr><th colspan="7">Parameter Estimates</th></tr>
<tr><th>Variable</th><th>DF</th><th>Parameter Estimate</th><th>Standard Error</th><th>t Value</th><th>Pr > |t|</th><th>Variance Inflation</th></tr>
<tr><td>Intercept</td><td>1</td><td>11.44279</td><td>0.02167</td><td>528.16</td><td><.0001</td><td>0</td></tr>
<tr><td>BsmtUnfSF</td><td>B</td><td>0.00053191</td><td>0.00002228</td><td>23.88</td><td><.0001</td><td>1.42902</td></tr>
<tr><td>BsmtFinSF2</td><td>B</td><td>0.00040023</td><td>0.00005308</td><td>7.54</td><td><.0001</td><td>1.08124</td></tr>
<tr><td>BsmtFinSF1</td><td>B</td><td>0.00058812</td><td>0.00002113</td><td>27.83</td><td><.0001</td><td>1.36987</td></tr>
<tr><td>TotalBsmtSF</td><td>0</td><td>0</td><td>.</td><td>.</td><td>.</td><td>.</td></tr>
</table>
</td>
</tr>
<tr>
<td>Final basement group variables selected for incorporation into full model:</td>
<td>BsmtExposure, BsmtFinType1, BsmtFinSF1, BsmtFinSF2, and BsmtUnfSF, BsmtFinType2</td>
</tr>
</table>

First a correlation test was run to investigate any possible correlation to log sale price and the other Garage related continuous variables. GarageYrBlt seems to have low correlation to log Sale Price and also low significance (p=0.4581) to log Sale Price. Recommend removing GarageYrBlt in the final model. Finally, GarageArea and GarageCars were scatter plotted against one another since their titles would suggest similar data. In the event there is further evidence of multi-collinearity, the scatter plots may suggest only keeping one variable. We will keep both for now.

```
Proc CORR data=train;
VAR LogSalePrice
GarageYrBlt GarageCars
GarageArea;
run;
```

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	logSalePrice	GarageYrBlt	GarageCars	GarageArea
logSalePrice	1.00000 1460	0.54107 <.0001 1379	0.68062 <.0001 1460	0.65089 <.0001 1460
GarageYrBlt	0.54107 <.0001 1379	1.00000 1379	0.58892 <.0001 1379	0.56457 <.0001 1379
GarageCars	0.68062 <.0001 1460	0.58892 <.0001 1379	1.00000 1460	0.88248 <.0001 1460
GarageArea	0.65089 <.0001 1460	0.56457 <.0001 1379	0.88248 <.0001 1460	1.00000 1460

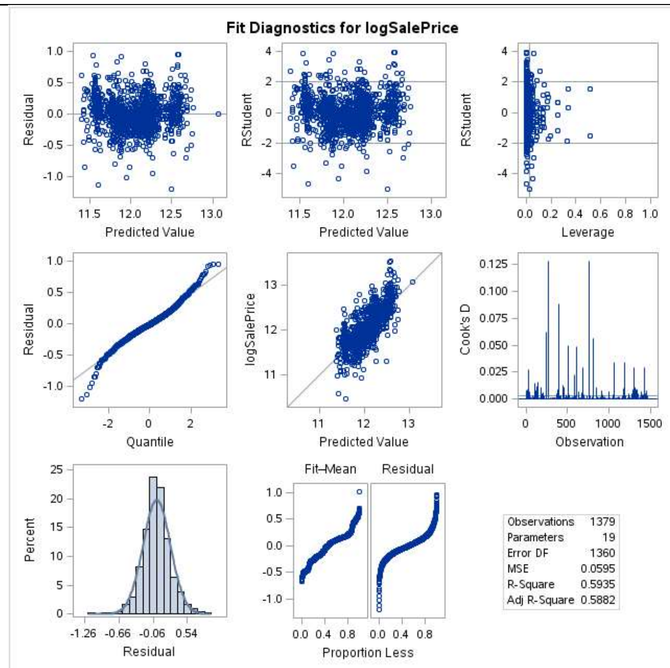
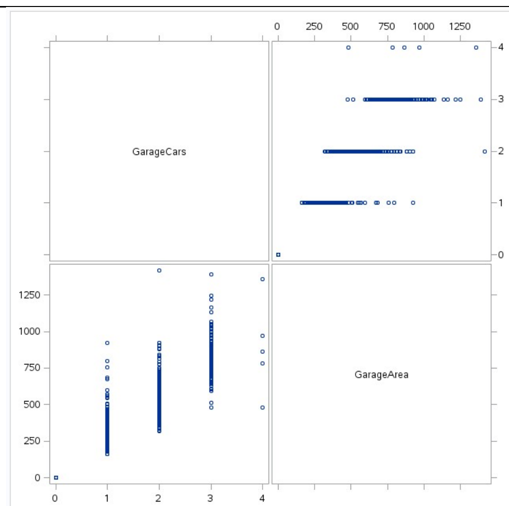
```
proc glm data = train
plots=diagnostics;
class GarageType
GarageFinish GarageQual
GarageCond;
model logsaleprice =
GarageType GarageFinish
GarageQual GarageCond
GarageYrBlt GarageCars
GarageArea;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	118.2059594	6.5669977	110.33	<.0001
Error	1360	80.9494913	0.0595217		
Corrected Total	1378	199.1554508			

R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.593536	2.023761	0.243971	12.05531

Source	DF	Type I SS	Mean Square	F Value	Pr > F
GarageType	5	53.61664125	10.72332825	180.16	<.0001
GarageFinish	2	20.50890840	10.25445420	172.28	<.0001
GarageQual	4	1.32001374	0.33000344	5.54	0.0002
GarageCond	4	2.24433229	0.56108307	9.43	<.0001
GarageYrBlt	1	10.48606013	10.48606013	176.17	<.0001
GarageCars	1	27.08888714	27.08888714	455.11	<.0001
GarageArea	1	2.94111646	2.94111646	49.41	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
GarageType	5	8.27257024	1.65451405	27.80	<.0001
GarageFinish	2	4.23673219	2.11836610	35.59	<.0001
GarageQual	4	1.16129535	0.29032384	4.88	0.0007
GarageCond	4	1.03478563	0.25869641	4.35	0.0017
GarageYrBlt	1	0.03278170	0.03278170	0.55	0.4581
GarageCars	1	4.77137507	4.77137507	80.16	<.0001
GarageArea	1	2.94111646	2.94111646	49.41	<.0001

<table>
<tr><td></td><td>
</td></tr>
<tr><td>proc sgscatter data = train;
matrix GarageCars
GarageArea;
run;</td><td>
</td></tr>
<tr><td>Variance Inflation Factor</td><td></td></tr>
<tr><td>Final basement group variables selected for incorporation into full model:</td><td></td></tr>
</table>

10. Outdoors: WoodDeckSF, OpenPorchSF, EnclosedPorch, _3SsnPorch, ScreenPorch, PoolArea

Correlation and VIF tests were run for exploratory purposes among the Outdoor group of variables. The only log transformation required was on the sale price variable. Of note in this initial exploration were two variables (PoolArea and _3SsnPorch). The PoolArea variable was the least significant of the grouping (p=0.2367). We will still leave this in the full model for now and see how it interacts with the other variables and also if the model improves with another degree of freedom (adjusted R-Square) before considering elimination. The -3SsnPorch variable had a negative effect to the log sale price. This may be interesting to the realtor, or home owner, as they may choose to remove the enclosure to boost the sale price. Again, we will leave this in the full model to determine how it interacts with all of the other variables.

```
Proc CORR data=train;
VAR LogSalePrice
WoodDeckSF
OpenPorchSF
EnclosedPorch
_3SsnPorch ScreenPorch
PoolArea;
run;
```

Pearson Correlation Coefficients, N = 1460
Prob > |r| under H0: Rho=0

	logSalePrice	WoodDeckSF	OpenPorchSF	EnclosedPorch	_3SsnPorch	ScreenPorch	PoolArea
logSalePrice	1.00000	0.33414 <.0001	0.32105 <.0001	-0.14905 <.0001	0.05490 0.0359	0.12121 <.0001	0.06980 0.0076
WoodDeckSF	0.33414 <.0001	1.00000	0.05866 0.0250	-0.12599 <.0001	-0.03277 0.2108	-0.07418 0.0046	0.07338 0.0050
OpenPorchSF	0.32105 <.0001	0.05866 0.0250	1.00000	-0.09308 0.0004	-0.00584 0.8235	0.07430 0.0045	0.06076 0.0202
EnclosedPorch	-0.14905 <.0001	-0.12599 <.0001	-0.09308 0.0004	1.00000	-0.03731 0.1542	-0.08286 0.0015	0.05420 0.0384
_3SsnPorch	0.05490 0.0359	-0.03277 0.2108	-0.00584 0.8235	-0.03731 0.1542	1.00000	-0.03144 0.2300	-0.00799 0.7603
ScreenPorch	0.12121 <.0001	-0.07418 0.0046	0.07430 0.0045	-0.08286 0.0015	-0.03144 0.2300	1.00000	0.05131 0.0500
PoolArea	0.06980 0.0076	0.07338 0.0050	0.06076 0.0202	0.05420 0.0384	-0.00799 0.7603	0.05131 0.0500	1.00000

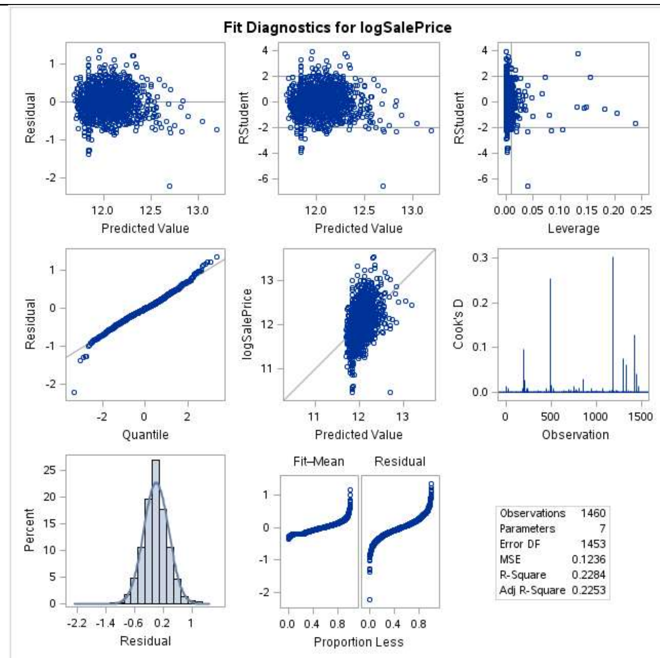
```
proc glm data = train
plots=diagnostics;
model logsaleprice =
WoodDeckSF OpenPorchSF
EnclosedPorch _3SsnPorch
ScreenPorch PoolArea;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	53.1822909	8.8637152	71.70	<.0001
Error	1453	179.6183680	0.1236190		
Corrected Total	1459	232.8006590			

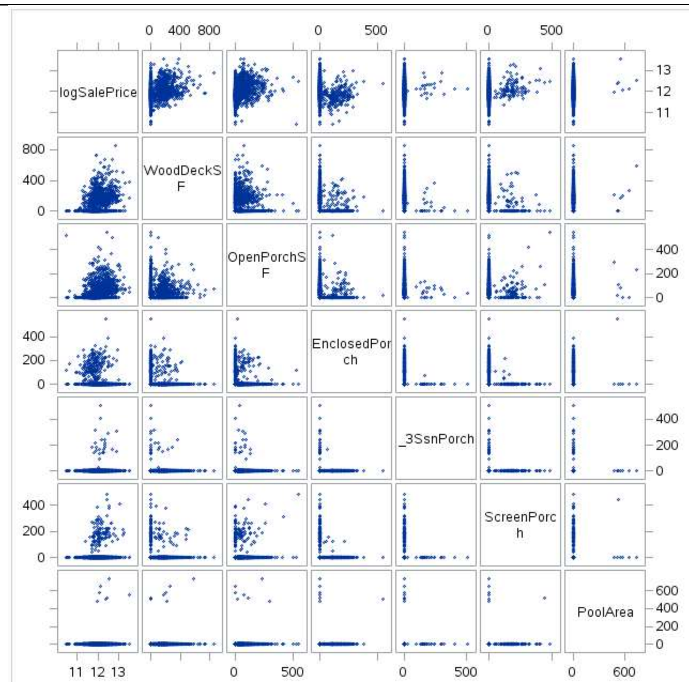
R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.228446	2.924097	0.351595	12.02405

Source	DF	Type I SS	Mean Square	F Value	Pr > F
WoodDeckSF	1	25.99132003	25.99132003	210.25	<.0001
OpenPorchSF	1	21.22847153	21.22847153	171.73	<.0001
EnclosedPorch	1	1.56501148	1.56501148	12.66	0.0004
_3SsnPorch	1	0.94350926	0.94350926	7.63	0.0058
ScreenPorch	1	3.28078764	3.28078764	26.54	<.0001
PoolArea	1	0.17319100	0.17319100	1.40	0.2367

Source	DF	Type III SS	Mean Square	F Value	Pr > F
WoodDeckSF	1	22.64323095	22.64323095	183.17	<.0001
OpenPorchSF	1	18.63572884	18.63572884	150.75	<.0001
EnclosedPorch	1	1.14865208	1.14865208	9.29	0.0023
_3SsnPorch	1	1.07933271	1.07933271	8.73	0.0032
ScreenPorch	1	3.18083547	3.18083547	25.73	<.0001
PoolArea	1	0.17319100	0.17319100	1.40	0.2367




```
proc sgscatter data = train;
matrix logsaleprice
WoodDeckSF OpenPorchSF
EnclosedPorch _3SsnPorch
ScreenPorch PoolArea;
run;
```



Variance Inflation Factor

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	53.18229	8.86372	71.70	<.0001
Error	1453	179.61837	0.12362		
Corrected Total	1459	232.80066			

Root MSE	0.35159	R-Square	0.2284
Dependent Mean	12.02405	Adj R-Sq	0.2253
Coeff Var	2.92410		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	11.84195	0.01429	828.65	<.0001	0
WoodDeckSF	1	0.00101	0.00007473	13.53	<.0001	1.03542
OpenPorchSF	1	0.00172	0.00014029	12.28	<.0001	1.01972
EnclosedPorch	1	-0.00046805	0.00015355	-3.05	0.0023	1.03947
_3SsnPorch	1	0.00092973	0.00031465	2.95	0.0032	1.00431
ScreenPorch	1	0.00084765	0.00016710	5.07	<.0001	1.02460
PoolArea	1	0.00027349	0.00023106	1.18	0.2367	1.01711

Final Outdoor group variables selected for incorporation into full model:	WoodDeckSF OpenPorchSF EnclosedPorch _3SsnPorch ScreenPorch PoolArea;

11. Misc: Street Utilities YearBuilt YearRemoAdd Functional MiscVal MoSold YrSold SaleType SaleCondition

Correlation and VIF tests were run for exploratory purposes among the Misc group of continuous variables. The only log transformation required was on the sale price variable. Three variables (MiscVal Street and Utilities) were not significant to the log of sale price. Their p-values were respectively, 0.8771, 0.1758, and 0.6070. These three variables were not very well populated with unique responses, hence they had insignificant effects on the screening model and will be removed from the combined final model. YearBuilt and YearRemodAdd both had significant p values (<0.0001) and were correlated 59% and 57% to the log sale price. These will stay in to be evaluated further in the final model. The remaining factors exhibited less than 0.12 p-values in this limited screening model. They will all advance for futher analysis in a full model.

```
Proc CORR data=train;
VAR LogSalePrice YearBuilt
YearRemodAdd MiscVal MoSold
YrSold;
run;
```

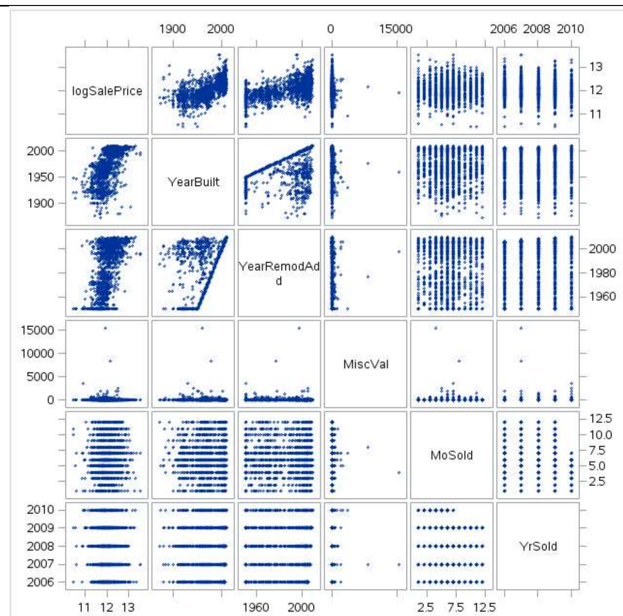
Pearson Correlation Coefficients, N = 1460
Prob > |r| under H0: Rho=0

	logSalePrice	YearBuilt	YearRemodAdd	MiscVal	MoSold	YrSold
logSalePrice	1.00000	0.58657 <.0001	0.56561 <.0001	-0.02002 0.4446	0.05733 0.0285	-0.03726 0.1547
YearBuilt	0.58657 <.0001	1.00000	0.59285 <.0001	-0.03438 0.1892	0.01240 0.6360	-0.01362 0.6031
YearRemodAdd	0.56561 <.0001	0.59285 <.0001	1.00000	-0.01029 0.6945	0.02149 0.4119	0.03574 0.1722
MiscVal	-0.02002 0.4446	-0.03438 0.1892	-0.01029 0.6945	1.00000	-0.00649 0.8042	0.00491 0.8514
MoSold	0.05733 0.0285	0.01240 0.6360	0.02149 0.4119	-0.00649 0.8042	1.00000	-0.14572 <.0001
YrSold	-0.03726 0.1547	-0.01362 0.6031	0.03574 0.1722	0.00491 0.8514	-0.14572 <.0001	1.00000

```
proc glm data = train
plots=diagnostics;
class Street Utilities Functional
SaleType SaleCondition;
model logsaleprice = Street Utilities
Functional SaleType SaleCondition
YearBuilt YearRemodAdd MiscVal
MoSold YrSold; *Y = X is the correct
format;
run;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Street	1	0.16535961	0.16535961	1.83	0.1758
Utilities	1	0.02386215	0.02386215	0.26	0.6070
Functional	6	1.26367548	0.21061258	2.34	0.0300
SaleType	5	0.79124256	0.15824851	1.76	0.1190
SaleCondition	5	1.51970834	0.30394167	3.37	0.0049
YearBuilt	1	17.58015852	17.58015852	195.05	<.0001
YearRemodAdd	1	13.82528744	13.82528744	153.39	<.0001
MiscVal	1	0.00215792	0.00215792	0.02	0.8771
MoSold	1	0.25697029	0.25697029	2.85	0.0915
YrSold	1	0.22348341	0.22348341	2.48	0.1156

```
proc sgscatter data = train;
matrix logsaleprice YearBuilt
YearRemodAdd MiscVal MoSold
YrSold;
run;
```



Variance Inflation Factor

```
proc reg data = train;
*plots(unpack)=residuals;
model logsaleprice = YearBuilt
YearRemodAdd MiscVal MoSold
YrSold / VIF; run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	12.09622	12.20524	0.99	0.3218	0
YearBuilt	1	0.00510	0.00032831	15.52	<.0001	1.54682
YearRemodAdd	1	0.00653	0.00048046	13.60	<.0001	1.54781
MiscVal	1	-0.00000230	0.00001608	-0.14	0.8863	1.00139
MoSold	1	0.00587	0.00298	1.97	0.0494	1.02252
YrSold	1	-0.01152	0.00608	-1.89	0.0584	1.02525

Final MISC group variables selected for incorporation into full model:	YearBuilt YearRemoAdd Functional MoSold YrSold SaleType SaleCondition

Appendix C
Listing of variables kept (kept if no note) and eliminated based on initial exploration.

BsmtExposure
BsmtFinType1
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF
BsmtFinType2
* TotalBsmtSF (Dropped due to buplicate to BsmtFinSF1 BsmtFinSF2 BsmtUnfSF);

* GrLivArea priority 1. May want to delete others;
GrLivArea
1stFlrSF
2ndFlrSF

BsmtFullBath
BsmtHalfBath
FullBath (keep with lower priority than BedroomAbvGr)
HalfBath (keep with lower priority than BedroomAbvGr)
BedroomAbvGr (higher priority)
Kitche-1bvGr
* TotRmsAbvGrd (removed because of multi-collinearity with FullBath and BedroomAbvGr)

GarageType
*GarageYrBlt (Remove due to multi-collinearity and p value).
GarageFinish
GarageCars (Leave in until full model, then remove 1. multi-collinearity with GarageArea and GarageYrBlt)
GarageArea (Leave in until full model, then remove 1. multi-collinearity with GarageCars and GarageYrBlt)
GarageQual
GarageCond

WoodDeckSF
OpenPorchSF
EnclosedPorch
_3SsnPorch (negative value, but no reason to remove at this point)
ScreenPorch
PoolArea (low p value in exploration run. Look at closer in full model)

MSSubClass (keep)
MSZoning (keep)
Neighborhood (keep)

LotFrontage (num. keep for now. high p-value)
LotArea (num. keep for now. More significant than LotFrontage)
LotShape (class keep)
LandContour (class keep for now. weak)

LotConfig (high p-value. Keep for now may need to eliminate)
LandSlope (keep)
PavedDrive (keep. more significant than LandSlope)

Condition1 (cat, keep)
Condition2 (cat, keep)
OverallQual (num, keep)
OverallCond (num, keep for now p .07)
ExterQual (keep)
*ExterCond (Remove due to low p value .43)
BsmtQual (keep)
*BsmtCond (Remove due to low p valaue .73)
*LowQualFinSF (num, Remove p .46)
KitchenQual (keep)

BldgType (keep, low p)
HouseStyle (keep low p)
RoofStyle (keep low p)
RoofMatl (keep low p)
Exterior1st (keep low p)
Exterior2nd (keep for now, See how it fits in model)
MasVnrType (keep low p)
MasVnrArea (num, keep low p)
Foundation (keep low p)

Heating (Keep, 40% multi-colinearity w Central Air)
HeatingQC (Keep it)
CentralAir (Keep it, but consider that it has 40% multi-colinearity with Heating and Electrical. It might be a good candidate to be removed later.)
Electrical (Keep it, but consider that it has 40% multi-colinearity with CentralAir.)

Fireplaces (keep)
FireplaceQu (Keep, .8 collinearity with Electical)

YearBuilt (keep)
YearRemoAdd(keep)
Functional(keep)
MoSold(keep)
YrSold(keep)
SaleType(keep)
SaleCondition(keep)
Misc Val (remove, high p)
Street (remove, high p)
Utilities (remove, high p)

Street (Remove, high p value and not well populated)
Utilities (Remove, high p value and not well populated)
YearBuilt
YearRemoAdd
Functional
MiscVal (Remove, high p value and not well populated)

MoSold
YrSold
SaleType
SaleCondition

Appendix D
Base Model Development and Output

```
proc freq data = train3;
table Neighborhood*BsmtQual / chisq ;
run;quit;
```

Statistics for Table of Neighborhood by BsmtQual

Statistic	DF	Value	Prob
Chi-Square	96	2768.4925	<.0001
Likelihood Ratio Chi-Square	96	2734.8212	<.0001
Mantel-Haenszel Chi-Square	1	90.7401	<.0001
Phi Coefficient		0.9739	
Contingency Coefficient		0.6977	
Cramer's V		0.4869	

WARNING: 40% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

```
proc freq data = train3;
table Neighborhood*RoofMatl / chisq ;
run;quit;
```

Statistics for Table of Neighborhood by RoofMatl

Statistic	DF	Value	Prob
Chi-Square	168	415.0376	<.0001
Likelihood Ratio Chi-Square	168	149.9436	0.8379
Mantel-Haenszel Chi-Square	1	0.6927	0.4053
Phi Coefficient		0.3771	
Contingency Coefficient		0.3528	
Cramer's V		0.1425	

WARNING: 88% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

```
proc freq data = train3;
table BsmtQual*RoofMatl / chisq ;
run;quit;
```

Statistics for Table of BsmtQual by RoofMatl

Statistic	DF	Value	Prob
Chi-Square	28	34.0095	0.2006
Likelihood Ratio Chi-Square	28	31.5555	0.2930
Mantel-Haenszel Chi-Square	1	0.8486	0.3570
Phi Coefficient		0.1079	
Contingency Coefficient		0.1073	
Cramer's V		0.0540	

WARNING: 83% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

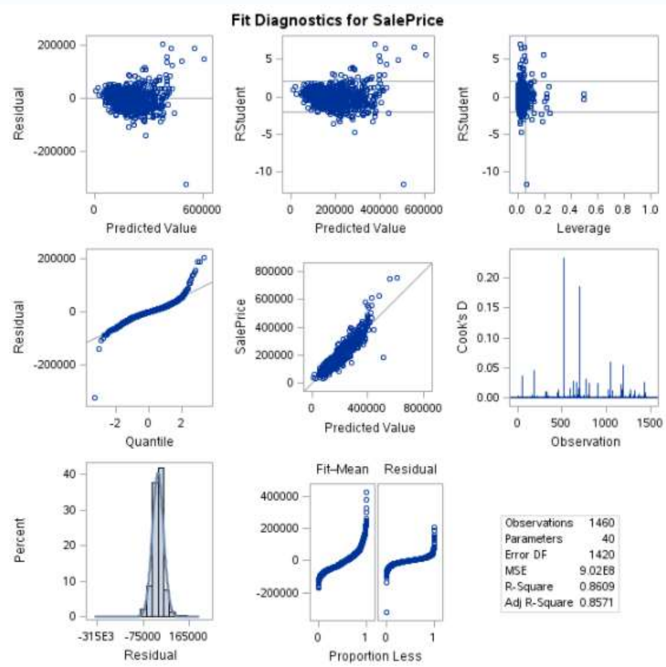
```
proc glm data = train3 plots=all;
class Neighborhood BsmtQual
RoofMatl;
model SalePrice = Neighborhood
BsmtQual RoofMatl MSSubClass
OverallQual BsmtFinSF1 GrLivArea/
clparm clm;
run;quit;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	39	7.9271268E12	203259682176	225.35	<.0001
Error	1420	1.2807845E12	901960922.35		
Corrected Total	1459	9.2079113E12			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.860904	16.59986	30032.66	180921.2

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Neighborhood	24	5.0236061E12	209316922573	232.07	<.0001
BsmtQual	4	922130654789	230532663697	255.59	<.0001
RoofMatl	7	161795413984	23113630569	25.63	<.0001
MSSubClass	1	37360360668	37360360668	41.42	<.0001
OverallQual	1	854259678140	854259678140	947.11	<.0001
BsmtFinSF1	1	186962514821	186962514821	207.28	<.0001
GrLivArea	1	741012060722	741012060722	821.56	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Neighborhood	24	295377102355	12307379265	13.65	<.0001
BsmtQual	4	123890990937	30972747734	34.34	<.0001
RoofMatl	7	261203375034	37314767862	41.37	<.0001
MSSubClass	1	102524531486	102524531486	113.67	<.0001
OverallQual	1	195014906193	195014906193	216.21	<.0001
BsmtFinSF1	1	183951489111	183951489111	203.95	<.0001
GrLivArea	1	741012060722	741012060722	821.56	<.0001



```
proc glm data = train3 plots=all;
class BsmtQual RoofMatl;
model SalePrice = BsmtQual RoofMatl
MSSubClass OverallQual BsmtFinSF1
GrLivArea/ solution;
run;quit;

* Model with the BEST 6 variables.;
```

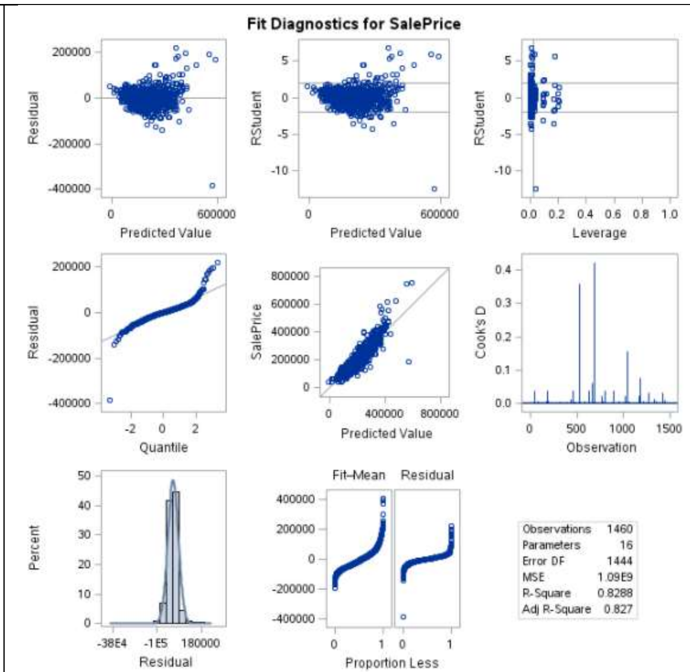
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	7.6317497E12	508783314835	466.12	<.0001
Error	1444	1.5761616E12	1091524662.1		
Corrected Total	1459	9.2079113E12			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.828825	18.26112	33038.23	180921.2

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BsmtQual	4	4.2816215E12	1.0704054E12	980.65	<.0001
RoofMatl	7	208691727751	29813103964	27.31	<.0001
MSSubClass	1	110856269704	110856269704	101.56	<.0001
OverallQual	1	1.8265366E12	1.8265366E12	1673.38	<.0001
BsmtFinSF1	1	304312546231	304312546231	278.80	<.0001
GrLivArea	1	899731148526	899731148526	824.29	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BsmtQual	4	299613532213	74903383053	68.62	<.0001
RoofMatl	7	332820103775	47545729111	43.56	<.0001
MSSubClass	1	122590317015	122590317015	112.31	<.0001
OverallQual	1	489247198693	489247198693	448.22	<.0001
BsmtFinSF1	1	262947182872	262947182872	240.90	<.0001
GrLivArea	1	899731148526	899731148526	824.29	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	10165.8653	B	14988.88188	0.68	0.4977
BsmtQual Ex	69493.2220	B	4224.79951	16.45	<.0001
BsmtQual Fa	-2509.1361	B	5771.27574	-0.43	0.6638
BsmtQual Gd	18738.9954	B	2227.70715	8.41	<.0001
BsmtQual NA	5672.9787	B	5726.84360	0.99	0.3220
BsmtQual TA	0.0000	B	.	.	.
RoofMatl ClyTile	-646948.7679	B	37362.72536	-17.32	<.0001
RoofMatl CompShg	-65810.6078	B	13675.20412	-4.81	<.0001
RoofMatl Membran	-21076.4019	B	35801.43542	-0.59	0.5562
RoofMatl Metal	-30014.9830	B	35804.84583	-0.84	0.4020
RoofMatl Roll	-82894.4499	B	35781.48144	-2.32	0.0207
RoofMatl Tar&Grv	-67528.3748	B	16896.83120	-4.00	<.0001
RoofMatl WdShake	-84750.5737	B	20086.68189	-4.22	<.0001
RoofMatl WdShngl	0.0000	B	.	.	.
MSSubClass	-221.9922		20.94722	-10.60	<.0001
OverallQual	21082.0102		995.78280	21.17	<.0001
BsmtFinSF1	32.9233		2.12122	15.52	<.0001
GrLivArea	60.9660		2.12348	28.71	<.0001



Final Model with the BEST 5 variables.

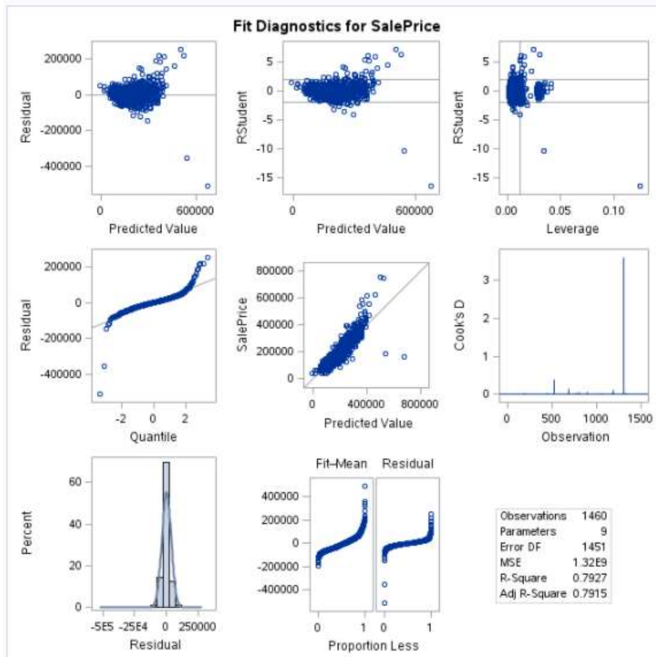
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	7.2989296E12	912366202344	693.48	<.0001
Error	1451	1.9089817E12	1315631782.1		
Corrected Total	1459	9.2079113E12			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.792680	20.04831	36271.64	180921.2

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BsmtQual	4	4.2816215E12	1.0704054E12	813.61	<.0001
MSSubClass	1	118178935078	118178935078	89.83	<.0001
OverallQual	1	1.8998849E12	1.8998849E12	1444.09	<.0001
BsmtFinSF1	1	225852338699	225852338699	171.67	<.0001
GrLivArea	1	773391935153	773391935153	587.85	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BsmtQual	4	306352645141	76588161285	58.21	<.0001
MSSubClass	1	129936088342	129936088342	98.76	<.0001
OverallQual	1	556207479845	556207479845	422.77	<.0001
BsmtFinSF1	1	148164504664	148164504664	112.62	<.0001
GrLivArea	1	773391935153	773391935153	587.85	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-50550.64036	B	5424.318828	-9.32	<.0001
BsmtQual Ex	70277.33892	B	4631.932548	15.17	<.0001
BsmtQual Fa	-4776.42972	B	6333.334588	-0.75	0.4509
BsmtQual Gd	19341.77160	B	2434.863948	7.94	<.0001
BsmtQual NA	2893.50657	B	6279.273830	0.46	0.6450
BsmtQual TA	0.00000	B	.	.	.
MSSubClass	-228.01373		22.943690	-9.94	<.0001
OverallQual	22363.23372		1087.635485	20.56	<.0001
BsmtFinSF1	23.66338		2.229830	10.61	<.0001
GrLivArea	55.14910		2.274604	24.25	<.0001



Appendix E

Final Model SAS Code

```
/* STATII HomePrice Project1  */
/* Home Sale Price Model */

/* Final Model */



/* Import train data */

FILENAME REFFILE '/home/mooyoungl0/MSDS 6371 STAT1/train.csv';

PROC IMPORT DATAFILE=REFFILE
	DBMS=CSV
	OUT=WORK.train;
	GETNAMES=YES;
	guessingrows=32767;
RUN;


data train1;
set train;
	MasVnrArea1 = input(MasVnrArea, 8.);
	drop MasVnrArea;
	rename MasVnrArea1=MasVnrArea;
run;

/* Data Manipulation */
data train2;
set train1;
	logSalePrice = log(SalePrice);

	if LotFrontage ="NA" then LotFrontage = 0;
	LotFrontage1 = input(LotFrontage, 8.);
	drop LotFrontage;
	rename LotFrontage1=LotFrontage;
run; quit;


/* Import test data */
FILENAME REFFILE '/home/mooyoungl0/MSDS 6371 STAT1/test.csv';

PROC IMPORT DATAFILE=REFFILE
	DBMS=CSV
	OUT=WORK.test;
	GETNAMES=YES;
	guessingrows=32767;
```

```
RUN;

**Corrects from CHAR to NUM type of variable, from import;
data test1;
set test;
	BsmtFinSF11 = input(BsmtFinSF1, 8.);
	drop BsmtFinSF1;
	rename BsmtFinSF11=BsmtFinSF1;
	BsmtFinSF21 = input(BsmtFinSF2, 8.);
	drop BsmtFinSF2;
	rename BsmtFinSF21=BsmtFinSF2;
	BsmtUnfSF1 = input(BsmtUnfSF, 8.);
	drop BsmtUnfSF;
	rename BsmtUnfSF1=BsmtUnfSF;
	TotalBsmtSF1 = input(TotalBsmtSF, 8.);
	drop TotalBsmtSF;
	rename TotalBsmtSF1=TotalBsmtSF;
	BsmtFullBath1 = input(BsmtFullBath, 8.);
	drop BsmtFullBath;
	rename BsmtFullBath1=BsmtFullBath;
	BsmtHalfBath1 = input(BsmtHalfBath, 8.);
	drop BsmtHalfBath;
	rename BsmtHalfBath1=BsmtHalfBath;
	GarageCars1 = input(GarageCars, 8.);
	drop GarageCars;
	rename GarageCars1=GarageCars;
	GarageArea1 = input(GarageArea, 8.);
	drop GarageArea;
	rename GarageArea1=GarageArea;
	MasVnrArea1 = input(MasVnrArea, 8.);
	drop MasVnrArea;
	rename MasVnrArea1=MasVnrArea;
	if LotFrontage ="NA" then LotFrontage = 0;
	LotFrontage1 = input(LotFrontage, 8.);
	drop LotFrontage;
	rename LotFrontage1=LotFrontage;
run;

/* Test data modification to fix missed predictions */

data test2;
set test1;
	if GarageCars = . then GarageCars = 0;
	if GarageArea = . then GarageArea = 0;
	if BsmtFullBath = . then BsmtFullBath = 0;
	if BsmtHalfBath = . then BsmtHalfBath = 0;
	if BsmtFinSF1 = . then BsmtFinSF1 = 0;
	if BsmtFinSF2 = . then BsmtFinSF2 = 0;
	if BsmtUnfSF = . then BsmtUnfSF = 0;
run;quit;
```

```
/* Merged train data */
data test3;
set test2;
SalePrice = .;
run;quit;

data train3;
set train2 test3;
run; quit;


/* Log Transformation */
data train4;
set train3;
logSalePrice = log(SalePrice);
logLotFrontage = log(LotFrontage+1);
logLotArea = log(LotArea+1);
logBsmtFinSF1 = log(BsmtFinSF1+1);
logBsmtFinSF2 = log(BsmtFinSF2+1);
logBsmtUnfSF = log(BsmtUnfSF+1);
logTotalBsmtSF = log(TotalBsmtSF+1);
log_1stFlrSF = log(_1stFlrSF+1);
log_2ndFlrSF = log(_2ndFlrSF+1);
logGrLivArea = log(GrLivArea+1);
logWoodDeckSF= log(WoodDeckSF+1);
logOpenPorchSF = log(OpenPorchSF+1);
logEnclosedPorch = log(EnclosedPorch+1);
log_3SsnPorch = log(_3SsnPorch+1);
logScreenPorch = log(ScreenPorch+1);
logPoolArea = log(PoolArea+1);

run;quit;


/* Adding Squar Terms */

data Dsqr;
set train4;
sqrOverallQual = OverallQual*OverallQual;
sqrOverallCond =  OverallCond*OverallCond;
sqrBsmtFinSF1 = BsmtFinSF1*BsmtFinSF1 ;
sqrBsmtFinSF2 = BsmtFinSF2*BsmtFinSF2 ;
sqrBsmtUnfSF = BsmtUnfSF*BsmtUnfSF ;
sqr_2ndFlrSF = _2ndFlrSF*_2ndFlrSF;
sqrBsmtFullBath = BsmtFullBath*BsmtFullBath ;
sqrBedroomAbvGr = BedroomAbvGr*BedroomAbvGr;
sqrGarageCars = GarageCars*GarageCars ;
sqrOpenPorchSF = OpenPorchSF*OpenPorchSF ;
sqrScreenPorch = ScreenPorch*ScreenPorch;
sqrPoolArea = PoolArea*PoolArea ;
```

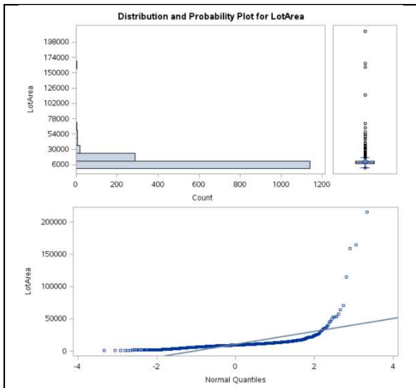
```
sqrYearBuilt = YearBuilt*YearBuilt ;
sqrYearRemodAdd = YearRemodAdd*YearRemodAdd;

sqrlogBsmtFinSF1 = logBsmtFinSF1*logBsmtFinSF1;
sqrlogTotalBsmtSF = logTotalBsmtSF*logTotalBsmtSF;
sqrlog_1stFlrSF = log_1stFlrSF*log_1stFlrSF;
sqrlog_2ndFlrSF = log_2ndFlrSF*log_2ndFlrSF;
sqrlogWoodDeckSF = logWoodDeckSF*logWoodDeckSF;
sqrlogOpenPorchSF = logOpenPorchSF*logOpenPorchSF;
run;quit;


/* Final Model */

proc glm data = Dsqr plots = all;
class
MSZoning Neighborhood Condition2 BsmtQual KitchenQual
RoofMatl CentralAir SaleCondition Functional;

Model logSalePrice =
MSZoning Neighborhood Condition2 BsmtQual KitchenQual
RoofMatl CentralAir SaleCondition Functional

OverallCond FullBath KitchenAbvGr Fireplaces GarageArea EnclosedPorch
logLotArea log_1stFlrSF
sqrOverallQual sqrBsmtFinSF1 sqrBsmtFinSF2 sqrBsmtFullBath sqrOpenPorchSF
sqrScreenPorch sqrYearBuilt sqrYearRemodAdd
sqrlogBsmtFinSF1 sqrlogTotalBsmtSF sqrlog_2ndFlrSF sqrlogWoodDeckSF
sqrlogOpenPorchSF
/ solution;

output out = results p = Predict;
run;quit;


/* Minimum House Price Filter */

data results6;
set results;
Predict = exp(Predict);
Predict = Predict;
if Predict > 0 then SalePrice = Predict;
if Predict < 35000 then SalePrice = 35000;
keep id SalePrice;
where id > 1460;
run; quit;
/*  */
/* proc univariate data = results2 plots; */
/* var SalePrice; */
/* run;quit; */
```

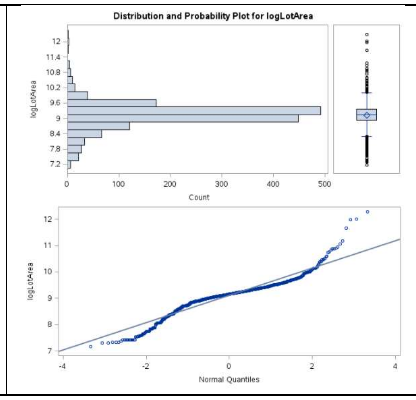
```
/* Export Output */
proc export data=results6
    outfile='/home/mooyoungl0/MSDS6372/HomeModelPart2_Final.csv'
    dbms=csv
    replace;
run;
```

Appendix F,

Final model iterations in 12 steps
Procedure:

1. Variable selection by grouping since there are about 80 variables which is a lot to process at once
2. Log transform some of the variables that are not normally distributed. individual distribution is not a required assumption to design a regression model but it showed increase in correlations so the transformed values are included. The LogSalePrice is shown in Table 6.





Pearson Correlation Coefficients, N = 1460
Prob > |r| under H0: Rho=0

	logSalePrice	LotArea	logLotArea
logSalePrice	1.00000	0.25732 <.0001	0.39992 <.0001

Table 6, LogSalePrice

3. Interaction term and higher order terms can be added to see if increase the fit.
4. Find most influential effects to the model fit using auto variable selection procedures. Use the AIC and CV to choose the variables because SBC penalize the complexity in the model which may lead to a poor fit.
5. Variables need to be examined to eliminate the multi-collinearity.
6. Kaggle data set will be used to check the model is working.

Step 1. Variables found from each groups shown in Table 7.

class variables	numerical variables
MSZoning Neighborhood LotShape LandContour LandSlope PavedDrive Condition1 Condition2 ExterQual ExterCond BsmtQual KitchenQual BldgType HouseStyle RoofStyle RoofMatl Exterior1st MasVnrType Foundation BsmtExposure BsmtFinType1 BsmtFinType2 HeatingQC CentralAir FireplaceQu GarageType GarageFinish GarageQual GarageCond SaleType SaleCondition	MSSubClass LotFrontage LotArea OverallQual OverallCond MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF _1stFlrSF _2ndFlrSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch _3SsnPorch ScreenPorch PoolArea YrSold YearBuilt YearRemodAdd MoSold

Functional;	

Table 7, Variables selected.

Step 2. Variables transformed in Table 8.

logLotFrontage logLotArea logBsmtFinSF1 logBsmtFinSF2 logBsmtUnfSF logTotalBsmtSF log_1stFlrSF log_2ndFlrSF logGrLivArea logWoodDeckSF logOpenPorchSF logEnclosedPorch log_3SsnPorch logScreenPorch logPoolArea

Table 8, Variables transformed

Step 3. Interaction term and higher order terms added, see Table 8.

MSSubClass*MSSubClass LotFrontage*LotFrontage LotArea*LotArea OverallQual*OverallQual OverallCond*OverallCond MasVnrArea*MasVnrArea BsmtFinSF1*BsmtFinSF1 BsmtFinSF2*BsmtFinSF2 BsmtUnfSF*BsmtUnfSF _1stFlrSF*_1stFlrSF _2ndFlrSF*_2ndFlrSF GrLivArea*GrLivArea BsmtFullBath*BsmtFullBath BsmtHalfBath*BsmtHalfBath FullBath*FullBath HalfBath*HalfBath BedroomAbvGr*BedroomAbvGr KitchenAbvGr*KitchenAbvGr Fireplaces*Fireplaces GarageCars*GarageCars GarageArea*GarageArea WoodDeckSF*WoodDeckSF OpenPorchSF*OpenPorchSF EnclosedPorch*EnclosedPorch _3SsnPorch*_3SsnPorch ScreenPorch*ScreenPorch PoolArea*PoolArea YrSold*YrSold YearBuilt*YearBuilt YearRemodAdd*YearRemodAdd MoSold*MoSold logLotFrontage*logLotFrontage logLotArea*logLotArea logBsmtFinSF1*logBsmtFinSF1 logBsmtFinSF2*logBsmtFinSF2 logBsmtUnfSF*logBsmtUnfSF logTotalBsmtSF*logTotalBsmtSF log_1stFlrSF*log_1stFlrSF log_2ndFlrSF*log_2ndFlrSF logGrLivArea*logGrLivArea logWoodDeckSF*logWoodDeckSF logOpenPorchSF*logOpenPorchSF logEnclosedPorch*logEnclosedPorch log_3SsnPorch*log_3SsnPorch

logScreenPorch*logScreenPorch logPoolArea*logPoolArea

Table 8, Interaction Terms investigated

Step4. Effects Found from auto model selection procedures found in Table 9.

Procedure	Effects Found	R-squre
forward(choose=AIC stop=AIC)	Intercept MSZoning Neighborhood Condition2 BsmtQual KitchenQual RoofMatl CentralAir SaleCondition Functional OverallQual OverallCond BsmtUnfSF FullBath KitchenAbvGr Fireplaces GarageCars GarageArea OpenPorchSF EnclosedPorch OverallQu*OverallQua OverallCo*OverallCon BsmtFinSF*BsmtFinSF1 BsmtFinSF*BsmtFinSF2 BsmtUnfSF*BsmtUnfSF _2ndFlrSF*_2ndFlrSF BsmtFullB*BsmtFullBa BedroomAb*BedroomAbv GarageCar*GarageCars OpenPorch*OpenPorchS ScreenPor*ScreenPorc PoolArea*PoolArea YearBuilt*YearBuilt YearRemod*YearRemodA logLotArea logBsmtFinSF2 log_1stFlrSF logGrLivArea logPoolArea logBsmtFi*logBsmtFin logTotalB*logTotalBs log_1stFl*log_1stFlr logWoodDe*logWoodDec logOpenPo*logOpenPor	0.9356
forward(choose=CV stop=CV)	Intercept Neighborhood OverallQual OverallCond GarageCar*GarageCars YearBuilt*YearBuilt logLotArea logGrLivArea logBsmtFi*logBsmtFin	0.8907
stepwise(choose=AIC stop=AIC)	Intercept MSZoning Neighborhood Condition2 BsmtQual RoofMatl SaleCondition Functional OverallCond KitchenAbvGr Fireplaces GarageArea OverallQu*OverallQua BsmtFullB*BsmtFullBa GarageCar*GarageCars ScreenPor*ScreenPorc YearBuilt*YearBuilt YearRemod*YearRemodA logLotArea logGrLivArea logBsmtFi*logBsmtFin logTotalB*logTotalBs	0.9287
stepwise(choose=CV stop=CV)	Intercept Neighborhood OverallQual OverallCond GarageCar*GarageCars YearBuilt*YearBuilt logLotArea logGrLivArea logBsmtFi*logBsmtFin	0.8907
LASSO(choose=AIC stop=AIC)	Intercept MSZoning_C (all) MSZoning_RM BsmtQual_Ex KitchenQual_TA RoofMatl_ClyTile HeatingQC_Ex CentralAir_N FireplaceQu_NA OverallQual OverallCond Fireplaces GarageCars GarageArea OverallQu*OverallQua YearBuilt*YearBuilt YearRemod*YearRemodA logLotArea log_1stFlrSF logGrLivArea logBsmtFi*logBsmtFin logTotalB*logTotalBs logGrLivA*logGrLivAr	0.8725
LASSO AIC effects found above will be reduced based on the all class level information shown below. If effects do not show all levels,	Intercept OverallQual OverallCond Fireplaces GarageCars GarageArea OverallQu*OverallQua YearBuilt*YearBuilt YearRemod*YearRemodA logLotArea log_1stFlrSF logGrLivArea logBsmtFi*logBsmtFin logTotalB*logTotalBs logGrLivA*logGrLivAr	

the whole variable will be removed.		

Table 9, various auto-selections with transformed variable and associated R-Square values

Additional information regarding factor levels may be found in Table 10 below.

Class Level Information		
Class	Levels	Values
MSZoning	6	C (all) FV NA RH RL RM
Neighborhood	25	Blmngtn Blueste BrDale BrkSide ClearCr CollgCr Crawfor Edwards Gilbert IDOTRR MeadowV Mitchel NAmes NPkVill NWAmes NoRidge NridgHt OldTown SWISU Sawyer SawyerW Somerst StoneBr Timber Veenker
LotShape	4	IR1 IR2 IR3 Reg
LandContour	4	Bnk HLS Low Lvl
LandSlope	3	Gtl Mod Sev
PavedDrive	3	N P Y
Condition1	9	Artery Feedr Norm PosA PosN RRAe RRAn RRNe RRNn
Condition2	8	Artery Feedr Norm PosA PosN RRAe RRAn RRNn
ExterQual	4	Ex Fa Gd TA
ExterCond	5	Ex Fa Gd Po TA
BsmtQual	5	Ex Fa Gd NA TA
KitchenQual	5	Ex Fa Gd NA TA
BldgType	5	1Fam 2fmCon Duplex Twnhs TwnhsE
HouseStyle	8	1.5Fin 1.5Unf 1Story 2.5Fin 2.5Unf 2Story SFoyer SLvl
RoofStyle	6	Flat Gable Gambrel Hip Mansard Shed
RoofMatl	8	ClyTile CompShg Membran Metal Roll Tar&Grv WdShake WdShngl
Exterior1st	16	AsbShng AsphShn BrkComm BrkFace CBlock CemntBd HdBoard ImStucc MetalSd NA Plywood Stone Stucco VinylSd Wd Sdng WdShing
MasVnrType	5	BrkCmn BrkFace NA None Stone
Foundation	6	BrkTil CBlock PConc Slab Stone Wood
BsmtExposure	5	Av Gd Mn NA No
BsmtFinType1	7	ALQ BLQ GLQ LwQ NA Rec Unf
BsmtFinType2	7	ALQ BLQ GLQ LwQ NA Rec Unf
HeatingQC	5	Ex Fa Gd Po TA
CentralAir	2	N Y
FireplaceQu	6	Ex Fa Gd NA Po TA
GarageType	7	2Types Attchd Basment BuiltIn CarPort Detchd NA
GarageFinish	4	Fin NA RFn Unf
GarageQual	6	Ex Fa Gd NA Po TA
GarageCond	6	Ex Fa Gd NA Po TA
SaleType	10	COD CWD Con ConLD ConLI ConLw NA New Oth WD
SaleCondition	6	Abnorml AdjLand Alloca Family Normal Partial
Functional	8	Maj1 Maj2 Min1 Min2 Mod NA Sev Typ

Table 10, Description of factors and levels for class variables.

Step 5. Variable examination for multi-collinearity and model selection

The model found by the ‘forward(choose=AIC stop=AIC)’ option will be used since it has the highest R-square value.

Using the below VIF values in Table 11, model VIF values are checked repeatedly by removing one highest VIF variables at a time.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-1.05981	1.48594	-0.71	0.4758	0
OverallQual	1	-0.05494	0.01979	-2.78	0.0056	68.59276
OverallCond	1	0.12639	0.02142	5.90	<.0001	52.02342
BsmtUnfSF	1	-0.00004759	0.00003424	-1.39	0.1648	20.96322
FullBath	1	0.02058	0.00948	2.17	0.0301	2.49692
KitchenAbvGr	1	-0.07618	0.01784	-4.27	<.0001	1.41460
Fireplaces	1	0.03690	0.00645	5.72	<.0001	1.58550
GarageCars	1	0.01894	0.01856	1.02	0.3076	17.62055
GarageArea	1	0.00005424	0.00003583	1.51	0.1303	5.37343
OpenPorchSF	1	0.00143	0.00031491	4.53	<.0001	39.86685
EnclosedPorch	1	0.00014682	0.00006155	2.39	0.0172	1.29616
logLotArea	1	0.08113	0.00792	10.24	<.0001	1.53814
logBsmtFinSF2	1	-0.00719	0.00260	-2.76	0.0058	2.11072
log_1stFlrSF	1	1.21035	0.41402	2.92	0.0035	1581.74328
logGrLivArea	1	0.42277	0.05671	7.45	<.0001	32.72329
logPoolArea	1	-0.13881	0.03449	-4.02	<.0001	20.96603
sqrOverallQual	1	0.01076	0.00156	6.89	<.0001	68.31659
sqrOverallCond	1	-0.00627	0.00181	-3.46	0.0006	54.05774
sqrBsmtFinSF1	1	-3.27664E-8	5.931126E-9	-5.52	<.0001	3.26230
sqrBsmtFinSF2	1	7.028131E-8	3.524826E-8	1.99	0.0464	2.10082
sqrBsmtUnfSF	1	2.717744E-8	1.55007E-8	1.75	0.0798	11.55751
sqr_2ndFlrSF	1	5.291209E-8	1.492583E-8	3.55	0.0004	4.85414
sqrBsmtFullBath	1	0.02687	0.00686	3.92	<.0001	1.81032
sqrBedroomAbvGr	1	-0.00222	0.00093975	-2.36	0.0185	2.02640
sqrGarageCars	1	0.00629	0.00472	1.33	0.1822	13.55986
sqrOpenPorchSF	1	-0.00000364	6.009277E-7	-6.06	<.0001	12.16272
sqrScreenPorch	1	0.00000126	2.306586E-7	5.48	<.0001	1.18188
sqrPoolArea	1	0.00000237	5.880853E-7	4.02	<.0001	18.85191
sqrYearBuilt	1	7.474674E-7	5.770134E-8	12.95	<.0001	4.28309
sqrYearRemodAdd	1	2.928136E-7	6.258006E-8	4.68	<.0001	2.39643
sqrlogBsmtFinSF1	1	0.00204	0.00042076	4.85	<.0001	6.52988
sqrlogTotalBsmtSF	1	0.00287	0.00078626	3.65	0.0003	4.60090
sqrlog_1stFlrSF	1	-0.08277	0.03076	-2.69	0.0072	1724.00198
sqrlog_2ndFlrSF	1	-0.00131	0.00066603	-1.97	0.0495	19.76476
sqrlogWoodDeckSF	1	0.00059455	0.00026601	2.24	0.0256	1.25673
sqrlogOpenPorchSF	1	-0.00431	0.00135	-3.19	0.0014	16.90480

Numerical Variable Removed	Max VIF
sqrlog_1stFlrSF	64.78
OverallQual	52.15
sqrOverallCond	39.85
OpenPorchSF	28.86
logGrLivArea	20.80
BsmtUnfSF	20.65
logPoolArea	16.14
GarageCars	4.76
sqr_2ndFlrSF	4.21
sqrBsmtUnfSF	4.10
sqrGarageCars	3.90
sqrBedroomAbvGr (due to high p-value)	3.56
sqrPoolArea (due to high p-value)	3.46

Final VIF values for the selected numerical variables.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	3.32210	0.29847	11.13	<.0001	0
OverallCond	1	0.04937	0.00386	12.79	<.0001	1.52512
FullBath	1	0.03029	0.00965	3.14	0.0017	2.33884
KitchenAbvGr	1	-0.04974	0.01822	-2.73	0.0064	1.33397
Fireplaces	1	0.04130	0.00660	6.26	<.0001	1.49842
GarageArea	1	0.00019106	0.00002211	8.64	<.0001	1.84832
EnclosedPorch	1	0.00016400	0.00006366	2.58	0.0101	1.25245
logLotArea	1	0.08844	0.00814	10.87	<.0001	1.46596
logBsmtFinSF2	1	-0.00820	0.00261	-3.14	0.0017	1.92241
log_1stFlrSF	1	0.38141	0.02039	18.71	<.0001	3.46513
sqrOverallQual	1	0.00717	0.00033781	21.23	<.0001	2.89109
sqrBsmtFinSF1	1	-5.0112E-8	4.282989E-9	-11.70	<.0001	1.53671
sqrBsmtFinSF2	1	8.4476E-8	3.492683E-8	2.42	0.0157	1.86329
sqrBsmtFullBath	1	0.02953	0.00695	4.25	<.0001	1.67817
sqrOpenPorchSF	1	-0.00000109	2.387304E-7	-4.58	<.0001	1.73401
sqrScreenPorch	1	0.00000114	2.367622E-7	4.81	<.0001	1.12489
sqrYearBuilt	1	8.064146E-7	5.532268E-8	14.58	<.0001	3.55665
sqrYearRemodAdd	1	2.903546E-7	6.329607E-8	4.59	<.0001	2.21460
sqrlogBsmtFinSF1	1	0.00236	0.00023336	10.11	<.0001	1.81440
sqrlogTotalBsmtSF	1	0.00248	0.00048641	5.11	<.0001	1.59061
sqrlog_2ndFlrSF	1	0.00435	0.00023605	18.45	<.0001	2.24264
sqrlogWoodDeckSF	1	0.00074215	0.00027581	2.69	0.0072	1.22047
sqrlogOpenPorchSF	1	0.00155	0.00049966	3.10	0.0020	2.09510

Table 11, Investigation by iterative removal of high VIF values.

The class variable association is not examined since there were 9 class variables left and there are too many pairs to test. The final model was run with all class and numerical variables left, and p-values are examined one more time. 'logBsmtFinSF2' is removed since the Type III SS p-value is insignificant (p-value = 0.1132). The final model effects and fit statistics shown in Table 12 below.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	82	216.2337926	2.6369975	219.18	<.0001
Error	1377	16.5668663	0.0120311		
Corrected Total	1459	232.8006590			

R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.928837	0.912226	0.109687	12.02405

Table 12, Final Model Statistics

Step 6. Verify model with Kaggle data set

The Kaggle score as shown in Figure 2 was 0.14582 which means the model is working as expected.

Appendix G

Data Cleansing Process

"LotFrontage" data type when it was imported from csv file was a character type so the "NA" values are replaced with 0 and the data type is modified to numerical type using below code.

```
data train2;
set train1;
        if LotFrontage ="NA" then LotFrontage = 0;
        LotFrontage1 = input(LotFrontage, 8.);
        drop LotFrontage;
        rename LotFrontage1=LotFrontage;
run; quit;
```

Some data strings were cut off during the import process and resulted inconsistent data level names between the train and test data sets. The "guessingrows" option is utilized in order solve the inconsistent data cut off length.

```
FILENAME REFFILE '/home/mooyoungl0/MSDS 6371 STAT1/train.csv';

PROC IMPORT DATAFILE=REFFILE
        DBMS=CSV
        OUT=WORK.train;
        GETNAMES=YES;
        guessingrows=32767;
RUN;
```

There were some missing values from garage and basement parameters. It occurred from only three samples, houses. A House (ID 1118) inputs were examined closely, and it was decided to assign zero instead of "NA" based on the conditions of house. Other two houses (ID 2121 and 2189) were having no basement inputs at all, and it was determined to assign zero instead of "NA" since the houses are old and the type of house was farm. Thus only sample data was manually changed using below code in order to resolve missing prediction values.

```
data test2;
set test1;
```

```
            if GarageCars = . then GarageCars = 0;
            if GarageArea = . then GarageArea = 0;
            if BsmtFullBath = . then BsmtFullBath = 0;
            if BsmtHalfBath = . then BsmtHalfBath = 0;
            if BsmtFinSF1 = . then BsmtFinSF1 = 0;
            if BsmtFinSF2 = . then BsmtFinSF2 = 0;
            if BsmtUnfSF = . then BsmtUnfSF = 0;
run;quit;
```

Other missing or low prediction values were filtered and forcefully assigned to $35,000, which is a round up value of the minimum home sale price from the train data set.

```
data results6;
set results;
Predict = exp(Predict);
Predict = Predict;
if Predict > 0 then SalePrice = Predict;
if Predict < 35000 then SalePrice = 35000;
keep id SalePrice;
where id > 1460;
run; quit;
```