Ames Housing OLS With Neighborhood Accuracy

Introduction

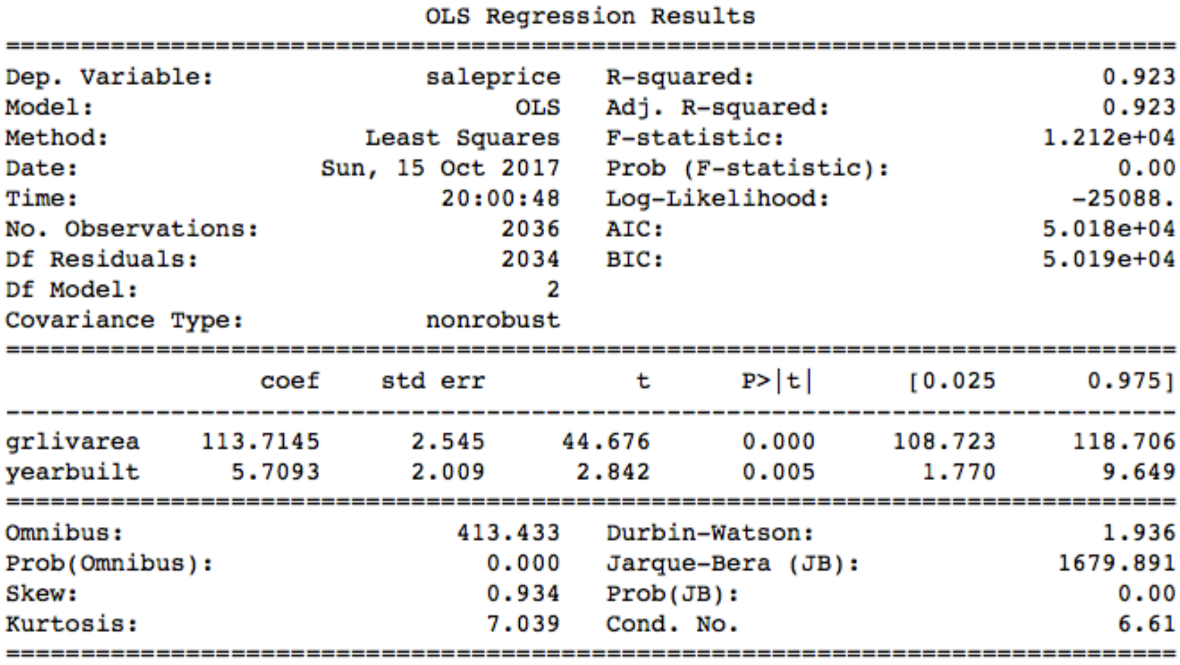
In this document we explore building a predictive model for the sale price of single family homes in Ames, IOWA using Ames Housing Dataset. Both simple and multiple regressions models are built to measure the linear relationship between the parameters. As part of the assignment log transformation of the responsible variable also performed. This analysis mainly focused on variables that consumers commonly ask about when looking for a home, such as the square footage. In previous analyses, data exploration was performed on such variables, and an exploratory data analysis showed that the total living area above grade may be a good predictor of sale price. EDA also produced that the total living area above grade, year built, and mean price per square foot by neighborhood have a positive correlation with sale price. A multiple linear regression with a log transformation of SalePrice produced the best model for predicting SalePrice. Variability in SalePrice by neighborhood needs further exploration, and the elimination of outliers and influential variables in each neighborhood could improve future models.

Section 1: Modeling & More

As described below several models were built to predict the SalePrice. Data transformation was performed on some of the parameters during the exploratory data analysis.

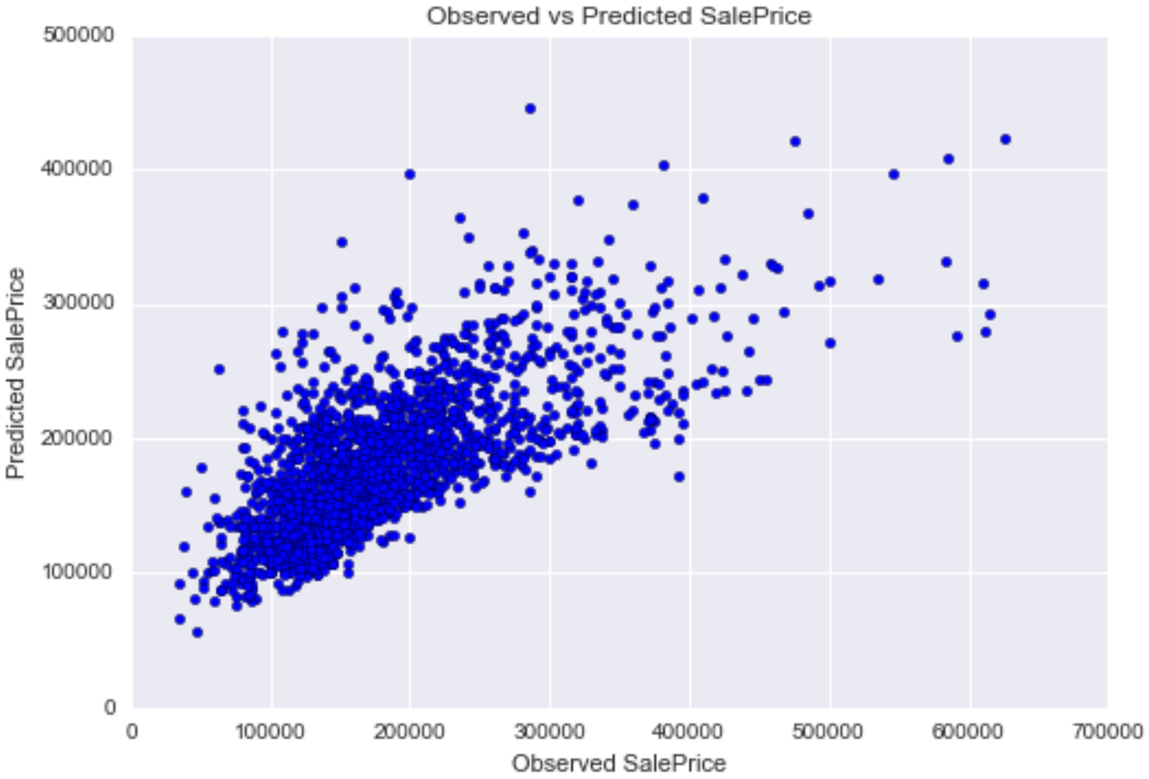
Section 1.1: Model 1 – GrLivArea & YearBuilt

In this model both ‘GrLivArea’ and ‘YearBuilt’ are considered for multiple linear regression. The multiple linear regression model has a r-squared of 0.923, which suggests there is a strong positive correlation between GrLivArea and YearBuilt to SalePrice. This correlation coefficient is higher than that of the simple linear models of GrLivArea to SalePrice or YearBuilt to SalePrice, but this is expected because the correlation coefficient usually gets larger when more variables are included. For this reason, we look at the adjusted r-squared value which adjusts for the number of predictors in the model, however in this case both has the same value. Both coefficients have a p-value less than 0.05, which means we can reject the null hypothesis that there is no relationship between GrLivArea and YearBuilt and SalePrice.



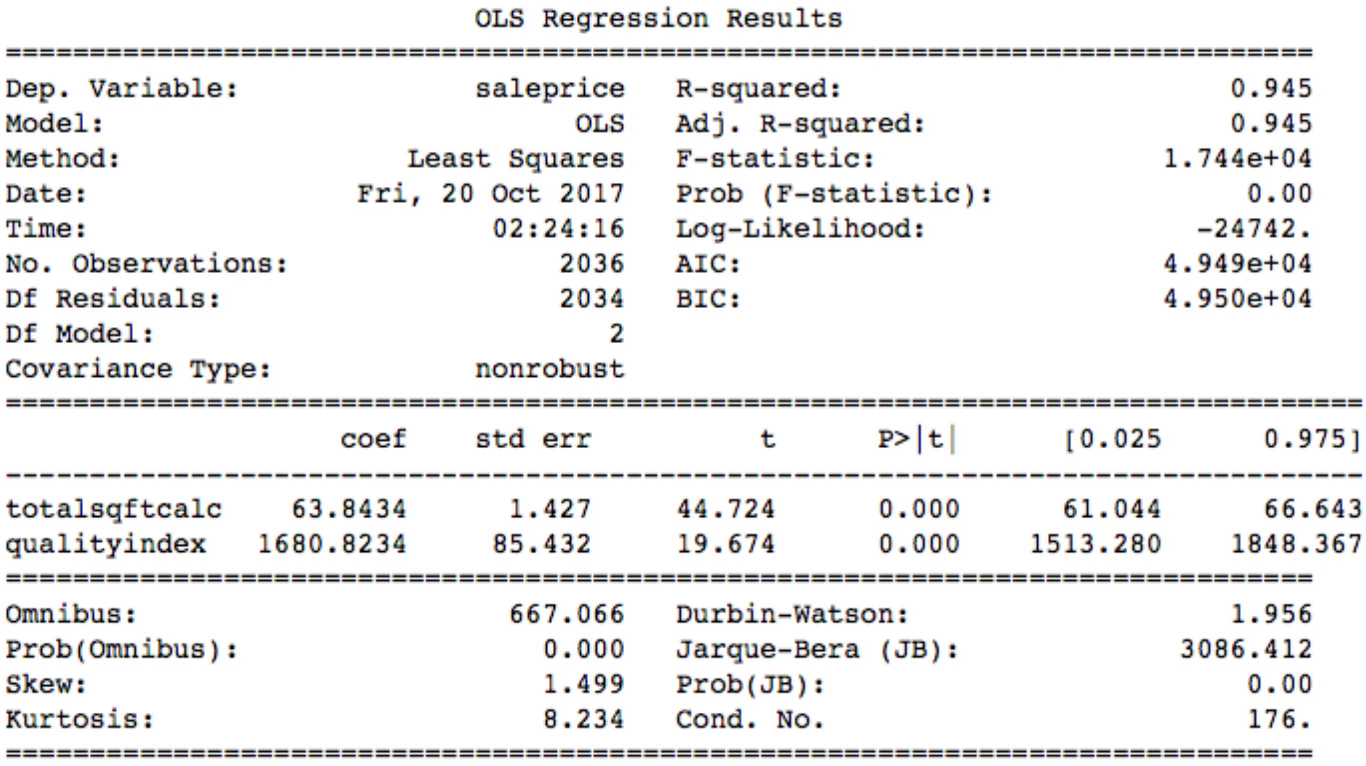
```
                            OLS Regression Results
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.923
Model:                            OLS   Adj. R-squared:                  0.923
Method:                 Least Squares   F-statistic:                 1.212e+04
Date:                Sun, 15 Oct 2017   Prob (F-statistic):               0.00
Time:                        20:00:48   Log-Likelihood:                -25088.
No. Observations:                2036   AIC:                         5.018e+04
Df Residuals:                    2034   BIC:                         5.019e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
grlivarea    113.7145      2.545     44.676      0.000     108.723     118.706
yearbuilt      5.7093      2.009      2.842      0.005       1.770       9.649
==============================================================================
Omnibus:                      413.433   Durbin-Watson:                   1.936
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1679.891
Skew:                           0.934   Prob(JB):                         0.00
Kurtosis:                       7.039   Cond. No.                         6.61
==============================================================================
```

Lastly, a plot of the observed vs predicted SalePrice values shows a stronger linear relationship than our simple linear models. The model is less accurate after a SalePrice of 400,000, which may be because we don't have enough observations that have a SalePrice above 400,000 to make an accurate prediction at that range.



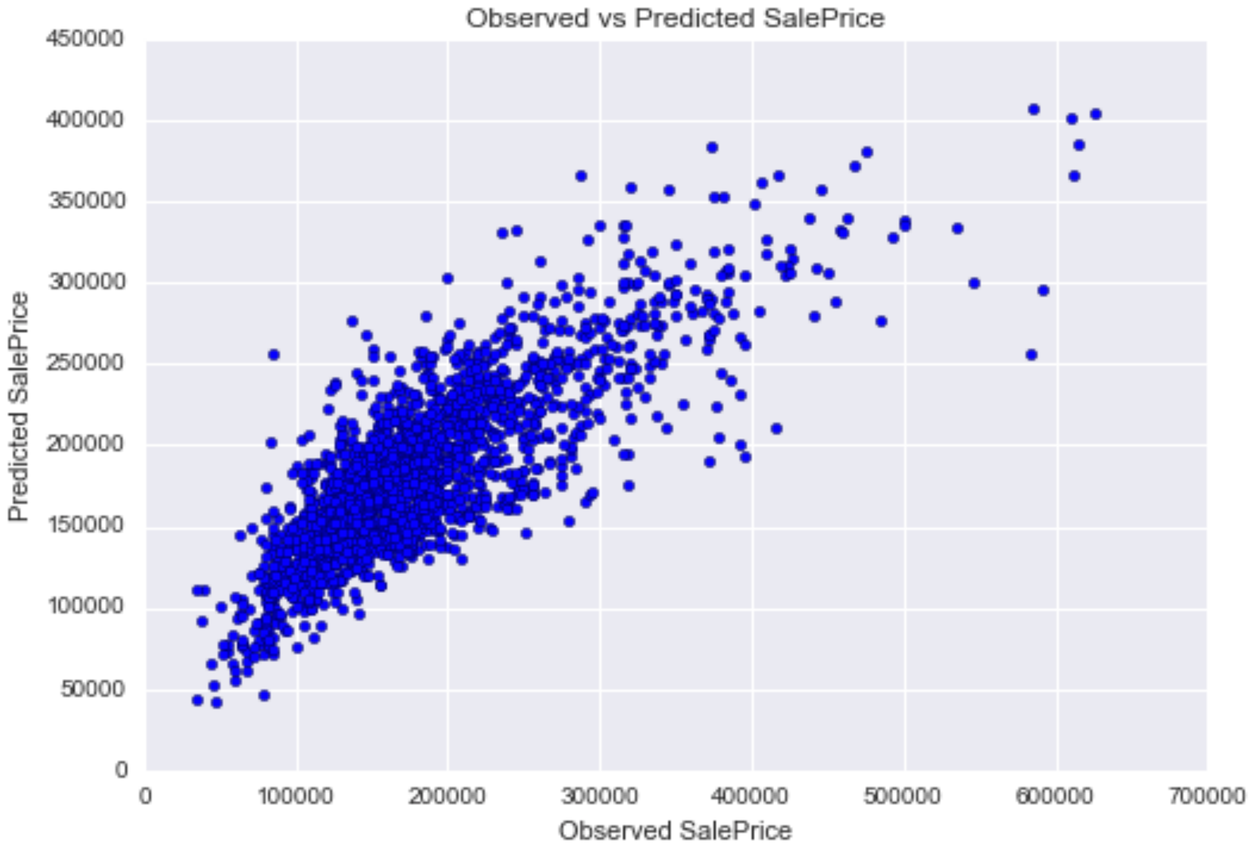
Section 1.2: Model 2 – 'totalsqftcalc' & 'qualityindex'

This model explores the two new variables, totalsqftcalc and qualityindex, that are calculated based on the training data set parameters. The multiple linear regression has a r-squared of 0.945, which suggests there is a strong positive correlation between the ‘totalsqftcalc’ and ‘qualityindex’. The parameter ‘totalsqftcalc’ represents the combined size of the two basement size variable and the total living area size and the second parameter ‘qualityindex’ represents the combined value of overall quality of the property and the overall condition of the property. One thing to note that there is no multicollinearity between these variables, so aggregating these variables will not produce any unwarranted results. When we look at the adjusted r-squared value which adjusts for the number of predictors in the model, however in this case both has the same value. Both coefficients have a p-value less than 0.05, which means we can reject the null hypothesis that there is no relationship between totalsqftcalc and qualityindex and SalePrice.



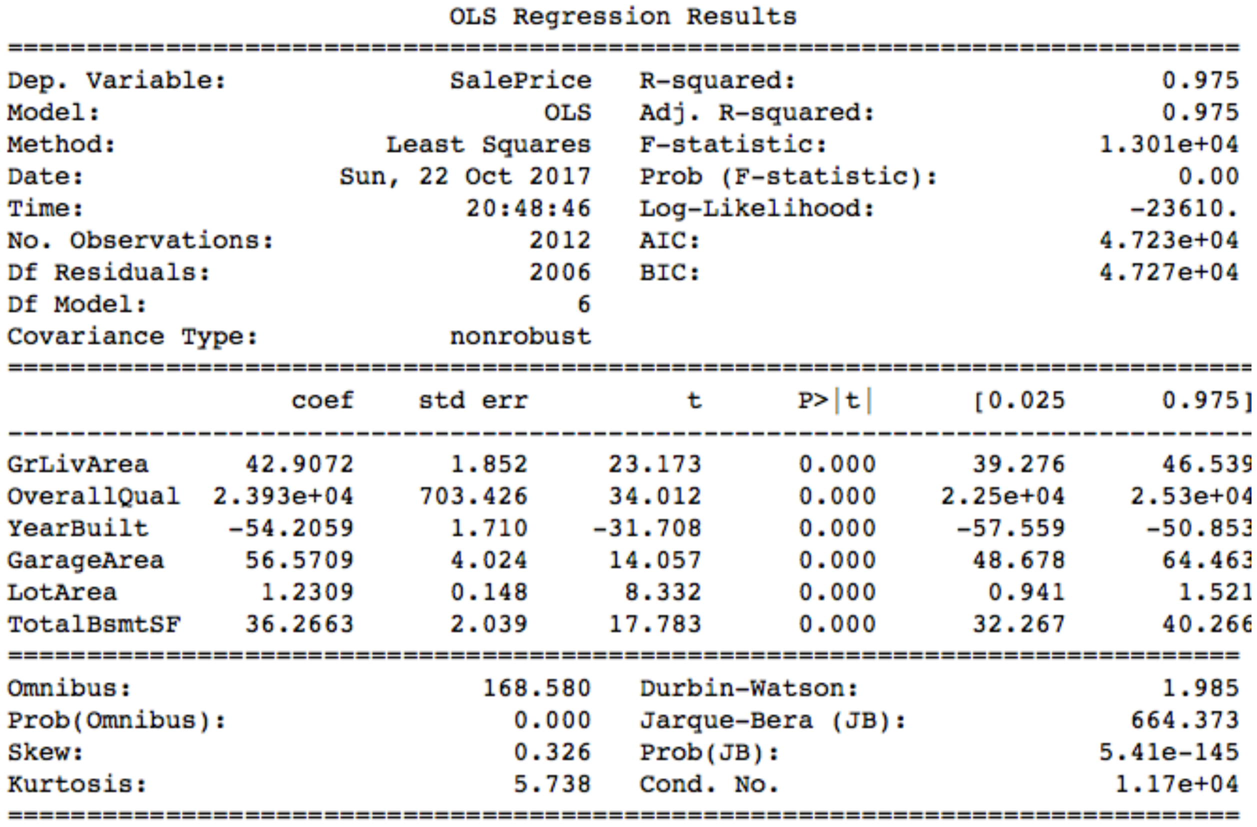
```
                            OLS Regression Results
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.945
Model:                            OLS   Adj. R-squared:                  0.945
Method:                 Least Squares   F-statistic:                 1.744e+04
Date:                Fri, 20 Oct 2017   Prob (F-statistic):               0.00
Time:                        02:24:16   Log-Likelihood:                -24742.
No. Observations:                2036   AIC:                         4.949e+04
Df Residuals:                    2034   BIC:                         4.950e+04
Df Model:                           2
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
totalsqftcalc    63.8434      1.427     44.724      0.000      61.044      66.643
qualityindex   1680.8234     85.432     19.674      0.000    1513.280    1848.367
==============================================================================
Omnibus:                      667.066   Durbin-Watson:                   1.956
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3086.412
Skew:                           1.499   Prob(JB):                         0.00
Kurtosis:                       8.234   Cond. No.                         176.
==============================================================================
```

The plot of the observed vs predicted SalePrice values shows a stronger relationship than the previous models. The model is less accurate after the SalePrice of 500,000 which may be because we don’t have enough observations that have a SalePrices above 500,000 to make an accurate prediction.

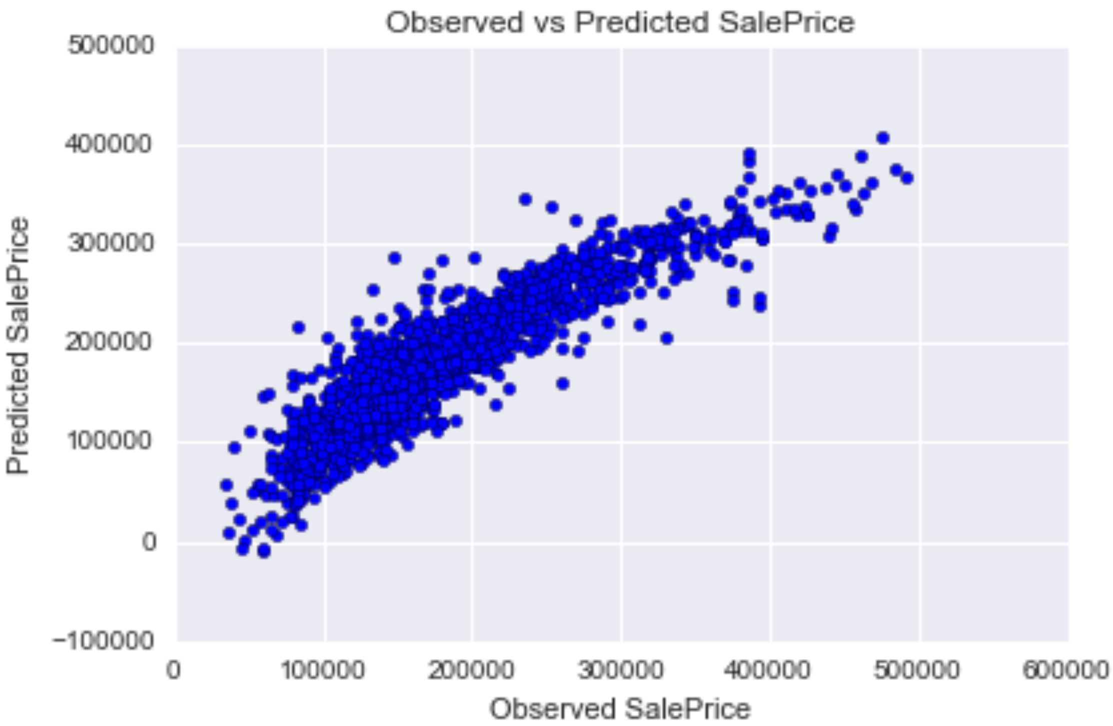


Section 1.3: Model 3 – Highly correlated

This model was built using all the variables that were highlighted as part of the EDA, including the two new variables that were created in model 2. The multiple linear regression has a r-squared of 0.975, which suggests there is a strong positive correlation between the ‘GrLivArea’, ‘OverallQual’, ‘yearbuilt’, ‘LotArea’, ‘TotalBsmtSF’ and the ‘garagearea’. One thing to note that there is no multicollinearity between these variables, so aggregating these variables will not produce any unwarranted results. When we look at the adjusted r-squared value which adjusts for the number of predictors in the model, however in this case both has the same value. Both coefficients have a p-value less than 0.05, which means we can reject the null hypothesis that there is no relationship between totalsqftcalc and qualityindex and SalePrice. The plot of the observed vs predicted SalePrice values shows a stronger relationship than the previous models. The model is less accurate after the SalePrice of 400,000 which may be because we don’t have enough observations that have a SalePrices above 400,000 to make an accurate prediction.

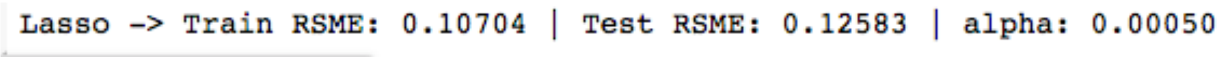


```
                            OLS Regression Results
==============================================================================
Dep. Variable:              SalePrice   R-squared:                       0.975
Model:                            OLS   Adj. R-squared:                  0.975
Method:                 Least Squares   F-statistic:                 1.301e+04
Date:                Sun, 22 Oct 2017   Prob (F-statistic):               0.00
Time:                        20:48:46   Log-Likelihood:                -23610.
No. Observations:                2012   AIC:                         4.723e+04
Df Residuals:                    2006   BIC:                         4.727e+04
Df Model:                           6
Covariance Type:            nonrobust
===============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
GrLivArea     42.9072      1.852     23.173      0.000      39.276      46.539
OverallQual 2.393e+04    703.426     34.012      0.000    2.25e+04    2.53e+04
YearBuilt    -54.2059      1.710    -31.708      0.000     -57.559     -50.853
GarageArea    56.5709      4.024     14.057      0.000      48.678      64.463
LotArea        1.2309      0.148      8.332      0.000       0.941       1.521
TotalBsmtSF   36.2663      2.039     17.783      0.000      32.267      40.266
==============================================================================
Omnibus:                      168.580   Durbin-Watson:                   1.985
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              664.373
Skew:                           0.326   Prob(JB):                    5.41e-145
Kurtosis:                       5.738   Cond. No.                     1.17e+04
==============================================================================
```

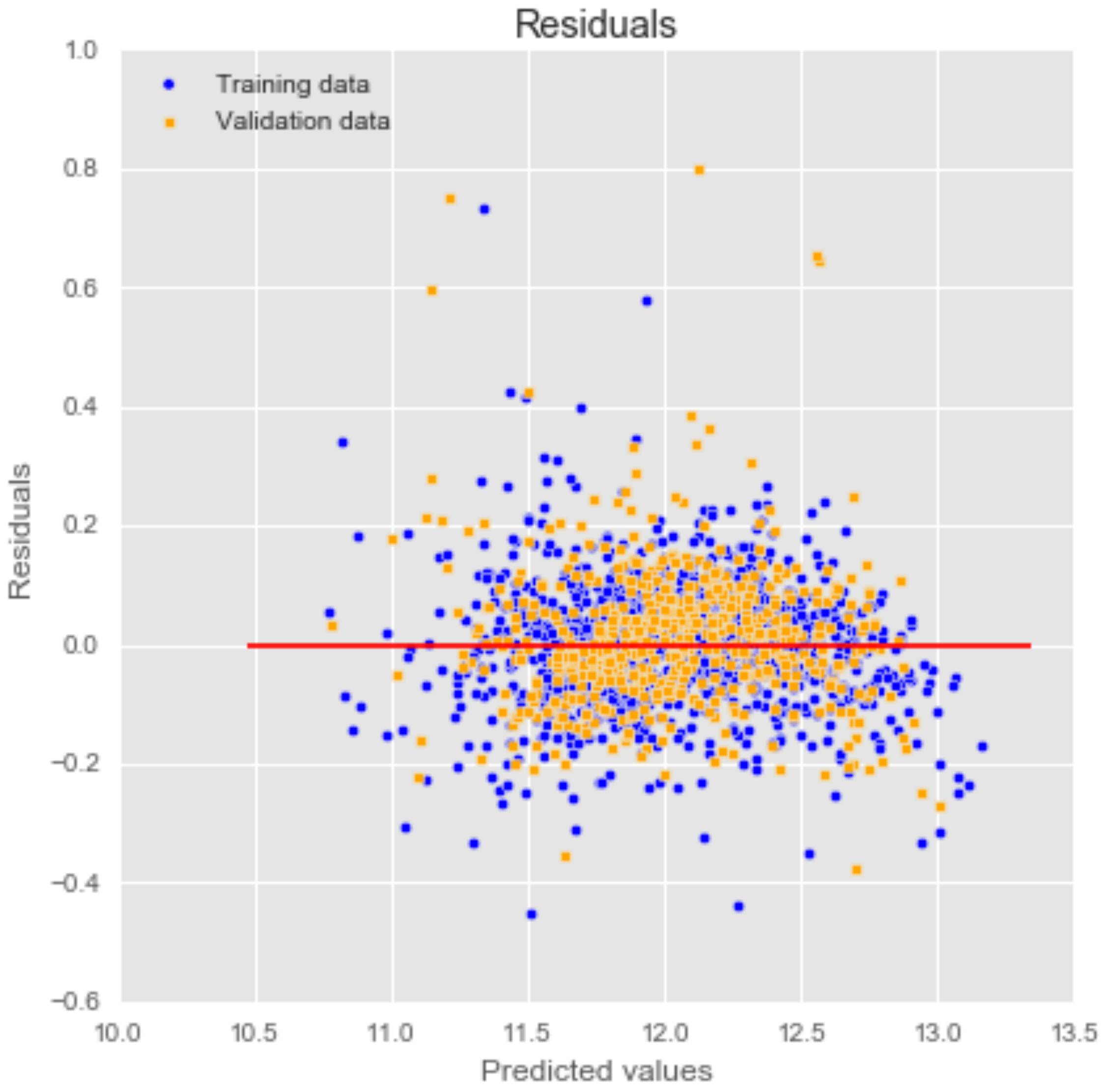


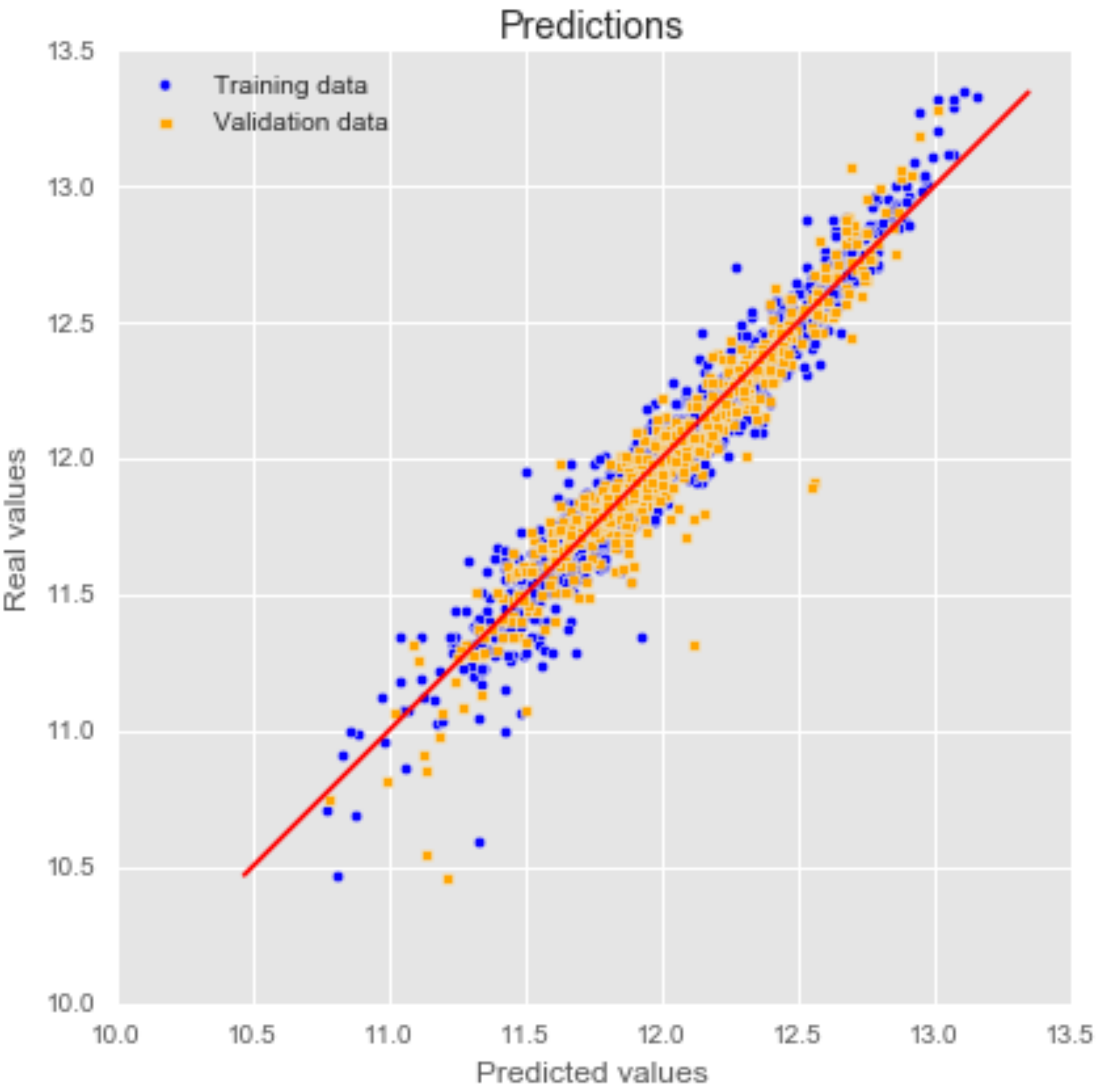
Section 1.3: Model 4 – Lasso Regression

Lasso Regression, least absolute shrinkage and selection operator, is a L1 regularization model in which sum of weights is added to the cost function. This model uses the grid search to find the optimal alpha values. Interestingly Lasso model dropped about two thirds of the features and provided the better RSME values for train and test data sets.



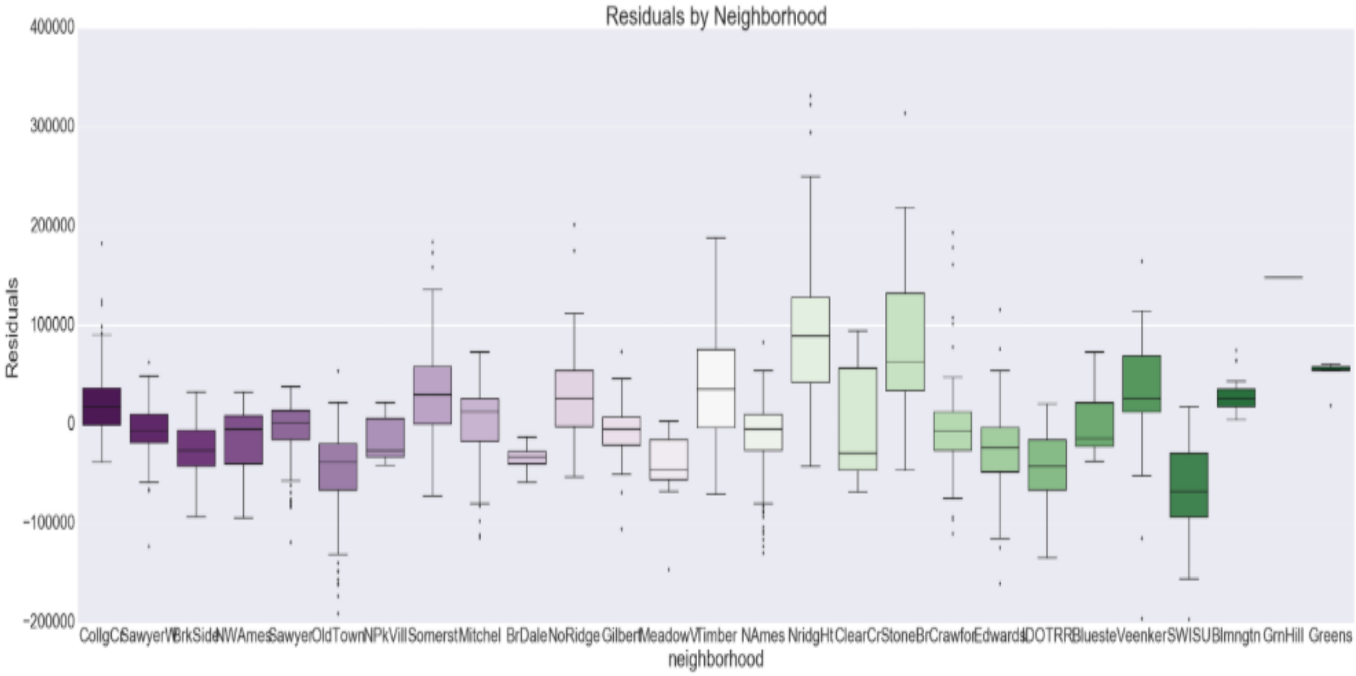
```
Lasso -> Train RSME: 0.10704 | Test RSME: 0.12583 | alpha: 0.00050
```




Predictions
Training data
Validation data
Real values
Predicted values
10.0
10.5
11.0
11.5
12.0
12.5
13.0
13.5
10.0
10.5
11.0
11.5
12.0
12.5
13.0
13.5

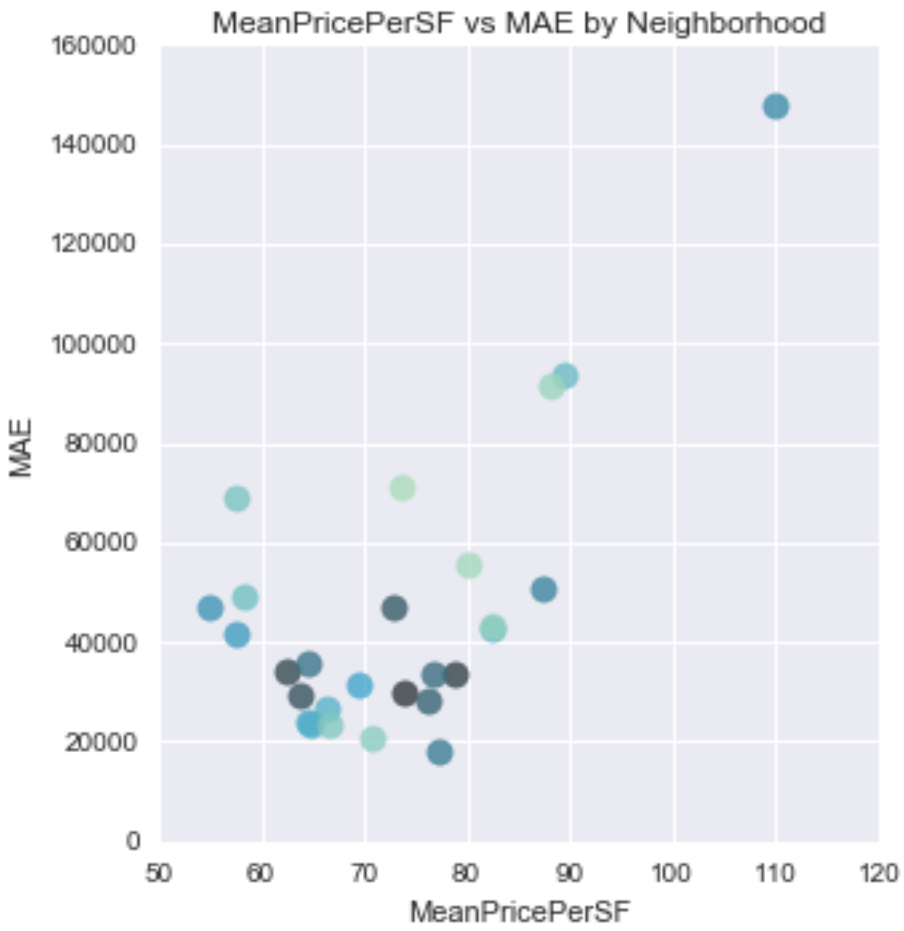
Section 1.4: Neighborhood Accuracy

The following box plot represents the residuals by neighborhood was created in order to evaluate whether some neighborhoods are better fit by the model than others. The neighborhoods of NridgeHt, ClearCr and StoneBr are consistently over predicted with larger residuals than the other neighborhoods. Neighborhoods NoRidge, Somerst, Crawfor, Timber and Veenker neighborhoods are also consistently over predicted by the model, but with smaller residuals. Under predicted neighborhoods include Gilbert, GmHill, BrDale, SWIsu, CollgCr, and Blmngtn, although there appears to be only one property in the GmHill neighborhood and very few observations in the Greens neighborhood, so we may not have enough data to predict SalePrices in these neighborhoods. Both over and under prediction could result from either bias in our sample or from a missing variable in our model. We may have left out crucial environmental or demographic factors that vary across neighborhoods and influence SalePrice, such as demographics information.



There also seem to be some extreme outliers for most of the neighborhoods. These outliers could be influencing the model to over predict or under predict the SalePrices for a certain neighborhood. For example, both NridgHt and NoRidge neighborhoods appear to have extreme outliers that may cause over prediction in our model. The StoneBr neighborhood in particular has a very skewed distribution to the right, suggesting there are outliers or influential values that could be removed to improve our model. Long boxplot whiskers in both directions, such as the boxplot for ClearCr neighborhood, could suggest that the neighborhood is being developed. Additional variables that cause or represent the variation among Neighborhoods could also be used to improve our model.

MeanPricePerSF versus MAE by Neighborhood

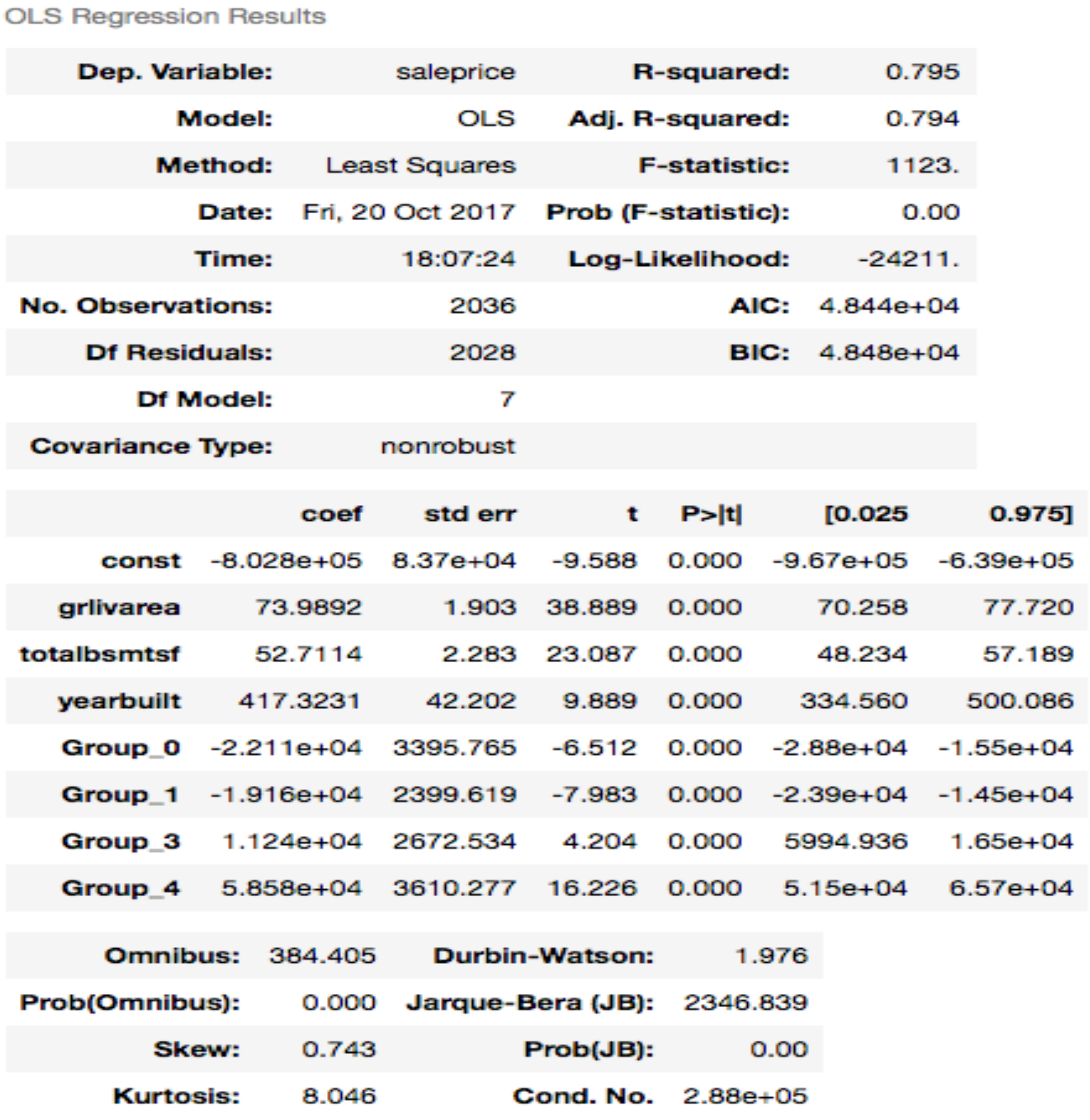


We found a positive relationship between the mean price per square foot (MeanPricePerSF) and the mean absolute error (MAE) in predicting SalePrice. The data for each neighborhood tends to be clustered below a mean absolute error of 60000. We see that the lowest MeanPricePerSF has a MAE of 20000 to 45000, and the MAE slightly decreases as the MeanPricePerSF increases from 70 to 80, and then the MAE and MeanPricePerSF increase together. There is a large difference between the MAE for a MeanPricePerSF under 85 and the MAE for a MeanPricePerSF around 90. This suggests to us that, while the MAE in predicting SalePrice generally increases with MeanPricePerSF, there may be more variability in SalePrice as the MeanPricePerSF increases. Neighborhoods are grouped based on their MeanPricePerSF to further explore the relationship between the neighborhoods and the SalePrice. After removing empty neighborhoods from the dataset, they have been categorized into total of five groups based on MeanPricePerSF.

- Neighborhoods that have less than 64 MeanPricePerSF are categorized into Group_0. IDOTRR, SWISU, MeadowV, Old Town, BrDale and BrkSide fall under this bucket
- Neighborhoods that have greater than 64 and less than 70 MeanPricePerSF are categorized into Group_1. Edwards, NAmes, NPkVill, NWAmes, Sawyer, Mitchel fall under this bucket.

- Neighborhoods that have greater than 70 and less than 76 MeanPricePerSF are categorized into Group_2. SawyerW, ClearCr, Veenker, Blmngtn, CollgCr, Crawfor are part of this group.
- Neighborhoods that have greater than 76 and less than 85 MeanPricePerSF are categorized into Group_3. Gilbert, Blueste, Timber, Somerst, NoRidge are part of this group.
- Neighborhoods that have greater than 75 MeanPricePerSF are categorized into Group_4. Greens, StoneBr, NridgHt, GrnHill are in the final group

As part of the next step, dummy variables are created for each group and selected Group_2 as the base category, which was chosen based on the group with largest number of neighborhoods. Below table shows the output from multiple linear regression model of SalePrice that includes ‘GrLivArea’, ‘YearBuilt’, ‘TotalBsmtSF’ and the neighborhood groups as the predictors.



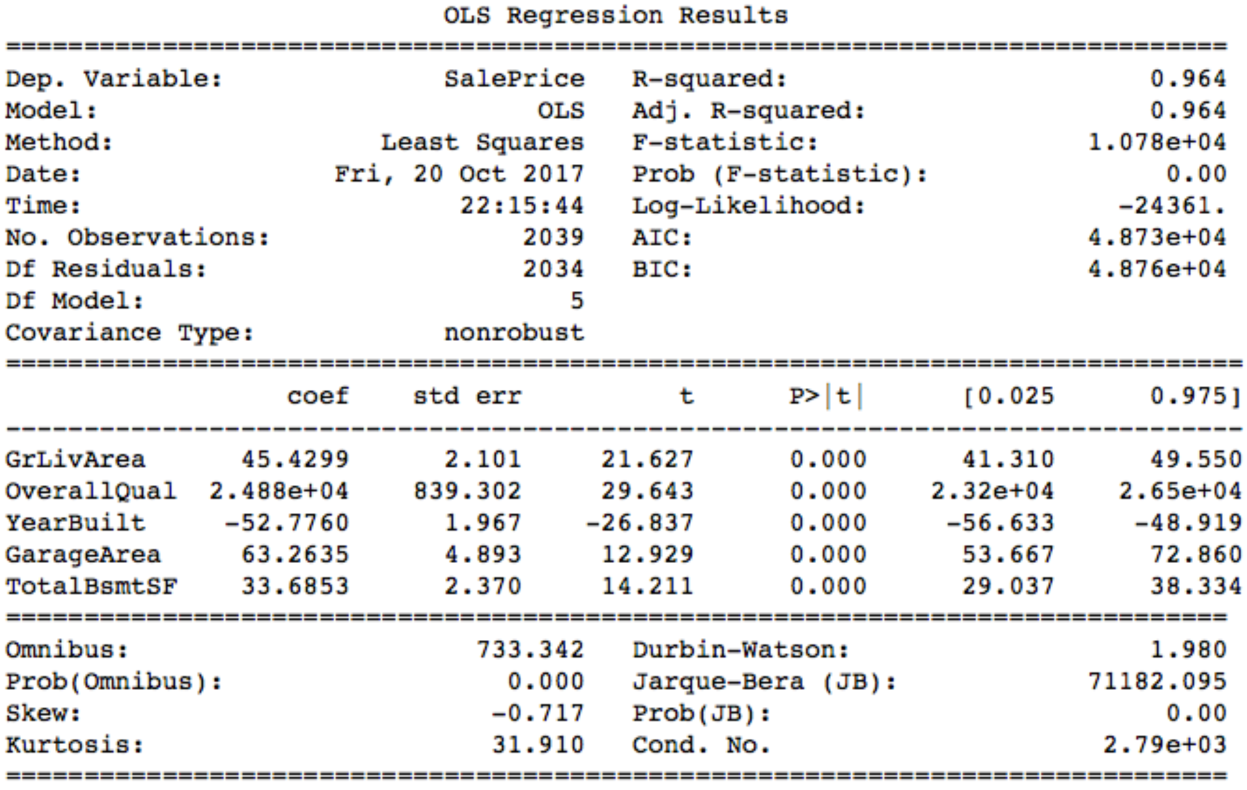
```
OLS Regression Results
      Dep. Variable:        saleprice          R-squared:      0.795
             Model:              OLS      Adj. R-squared:      0.794
            Method:    Least Squares         F-statistic:      1123.
              Date: Fri, 20 Oct 2017  Prob (F-statistic):       0.00
              Time:         18:07:24      Log-Likelihood:    -24211.
 No. Observations:             2036                 AIC:  4.844e+04
      Df Residuals:             2028                 BIC:  4.848e+04
          Df Model:                7
   Covariance Type:        nonrobust

                 coef    std err        t   P>|t|      [0.025      0.975]
      const -8.028e+05   8.37e+04   -9.588   0.000   -9.67e+05   -6.39e+05
  grlivarea    73.9892      1.903   38.889   0.000      70.258      77.720
totalbsmtsf    52.7114      2.283   23.087   0.000      48.234      57.189
  yearbuilt   417.3231     42.202    9.889   0.000     334.560     500.086
    Group_0 -2.211e+04   3395.765   -6.512   0.000   -2.88e+04   -1.55e+04
    Group_1 -1.916e+04   2399.619   -7.983   0.000   -2.39e+04   -1.45e+04
    Group_3  1.124e+04   2672.534    4.204   0.000    5994.936    1.65e+04
    Group_4  5.858e+04   3610.277   16.226   0.000    5.15e+04    6.57e+04

      Omnibus:  384.405      Durbin-Watson:     1.976
Prob(Omnibus):    0.000   Jarque-Bera (JB):  2346.839
         Skew:    0.743           Prob(JB):      0.00
     Kurtosis:    8.046          Cond. No.  2.88e+05
```

Section 2: Model Comparison of Y vs. log(Y)

This section provides the comparison between linear regression model of SalePrice vs. log(SalePrice) to explore whether log transformation would improve our model. Same five continuous parameters are chosen as the predictors for both the models ('GrLivArea', 'OverallQual', 'YearBuilt', 'GarageArea','TotalBsmtSF').

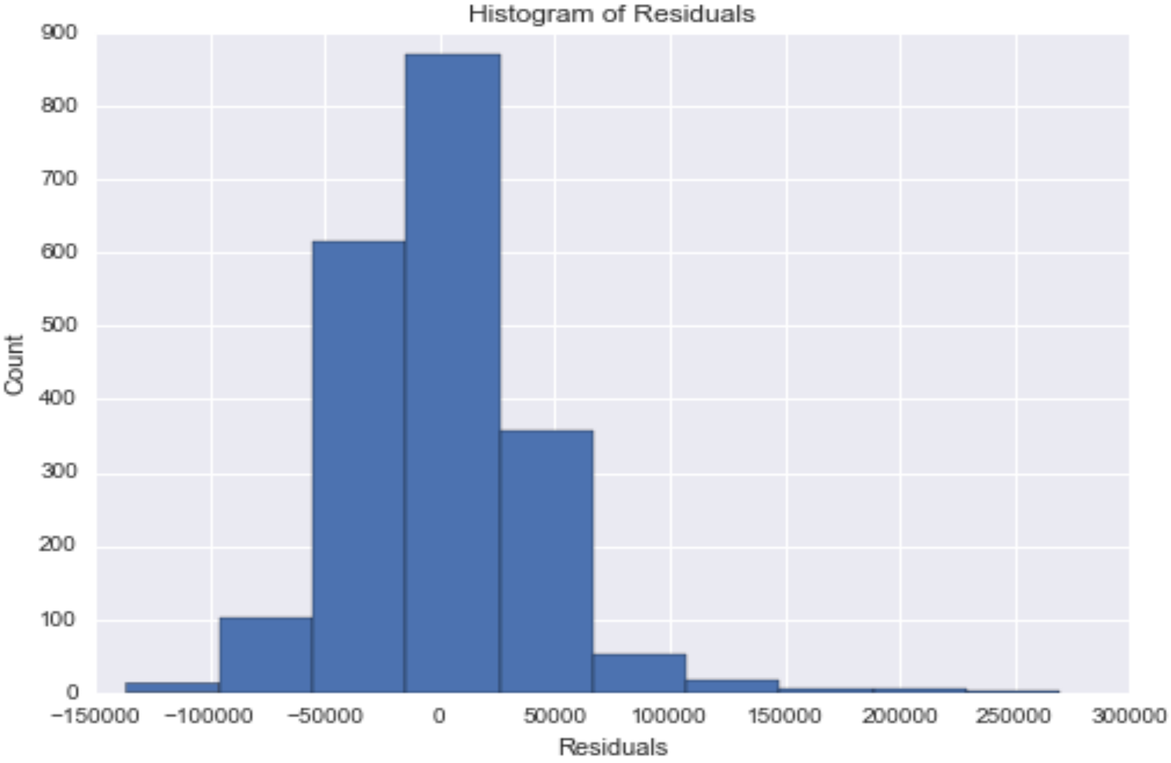
Section 2.1: Model of Y

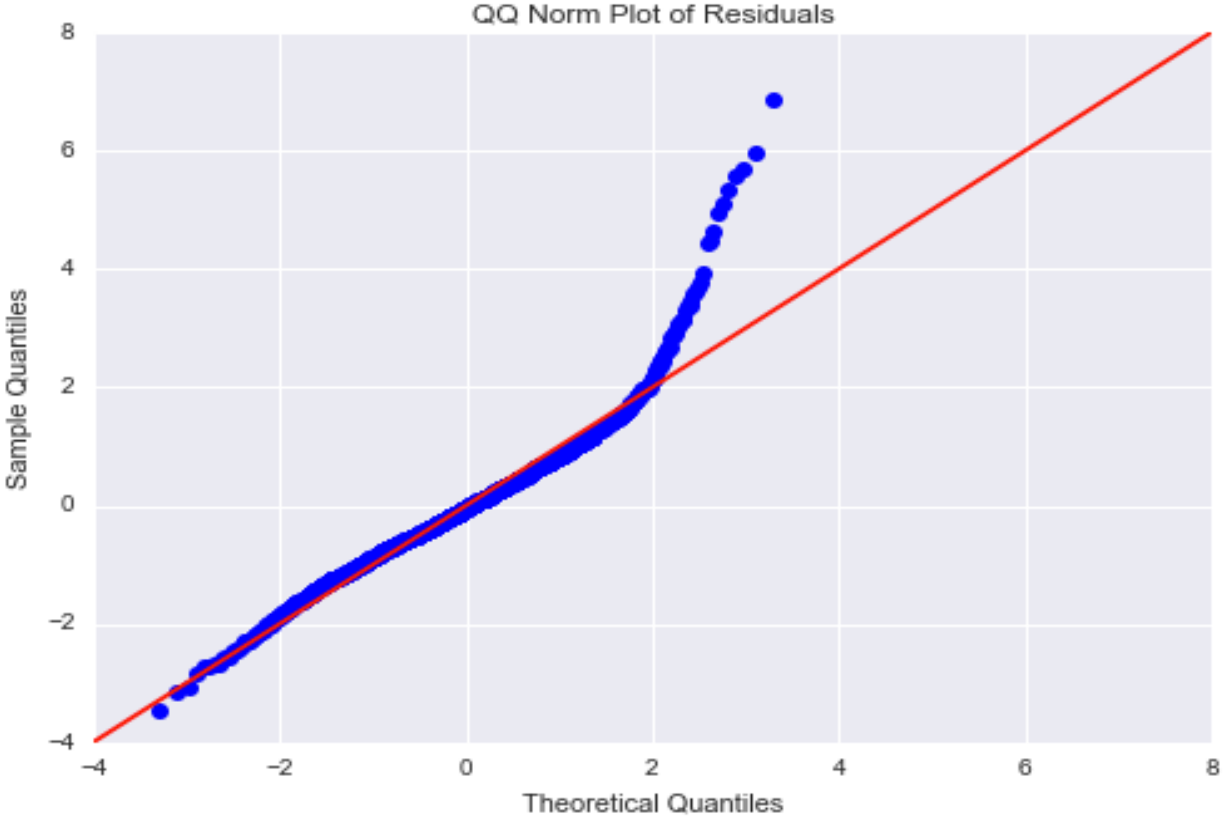
The OLS Regression Results from our SalePrice model has a very high correlation coefficient of 0.964. Although the F-statistic and t-statistic for all the coefficients are significant, we should look at the interaction effects between predictors and explore whether there is multicollinearity impacting our model in future analyses.

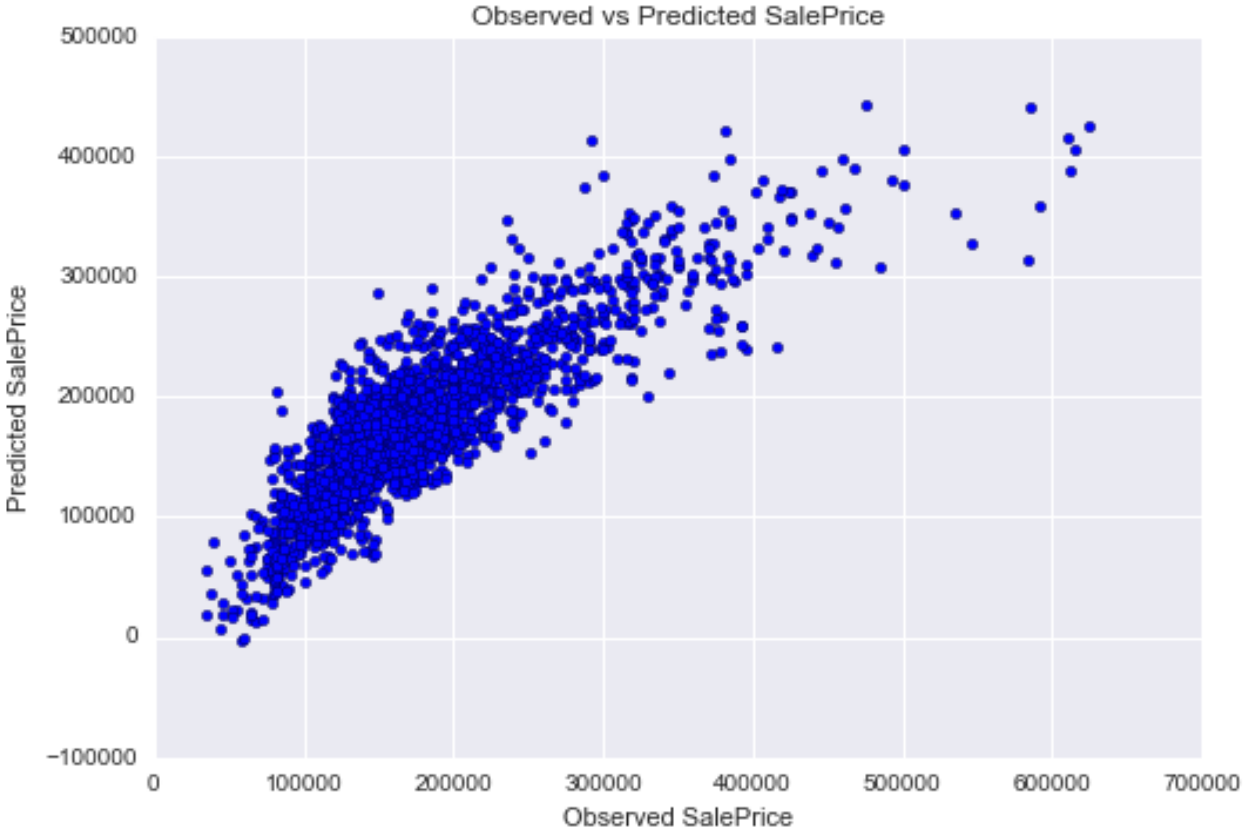


```
                            OLS Regression Results
==============================================================================
Dep. Variable:              SalePrice   R-squared:                       0.964
Model:                            OLS   Adj. R-squared:                  0.964
Method:                 Least Squares   F-statistic:                 1.078e+04
Date:                Fri, 20 Oct 2017   Prob (F-statistic):               0.00
Time:                        22:15:44   Log-Likelihood:                -24361.
No. Observations:                2039   AIC:                         4.873e+04
Df Residuals:                    2034   BIC:                         4.876e+04
Df Model:                           5
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
GrLivArea      45.4299      2.101     21.627      0.000      41.310      49.550
OverallQual  2.488e+04    839.302     29.643      0.000    2.32e+04    2.65e+04
YearBuilt     -52.7760      1.967    -26.837      0.000     -56.633     -48.919
GarageArea     63.2635      4.893     12.929      0.000      53.667      72.860
TotalBsmtSF    33.6853      2.370     14.211      0.000      29.037      38.334
==============================================================================
Omnibus:                      733.342   Durbin-Watson:                   1.980
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            71182.095
Skew:                          -0.717   Prob(JB):                         0.00
Kurtosis:                      31.910   Cond. No.                     2.79e+03
==============================================================================
```

The residuals histogram for the SalePrice model below shows a leptokurtic distribution with slight right skew, with most values clustered around 0. There may be outliers causing the skewness. The QQ Norm Plot of residuals also suggests that the residuals may not follow a normal distribution: the residuals have light tails and are skewed on both ends. Because the residuals do not appear to be normal, future analysis should seek to remove outliers or influential values that are skewing the residual distribution. Last, a scatterplot of the SalePrice model reveals a linear pattern that was present in previous models as well.


Histogram of Residuals
Count
900
800
700
600
500
400
300
200
100
0
−150000
−100000
−50000
0
50000
100000
150000
200000
250000
300000
Residuals


QQ Norm Plot of Residuals
Sample Quantiles
8
6
4
2
0
−2
−4
−4
−2
0
2
4
6
8
Theoretical Quantiles



Section 2.2: Model of log(Y)

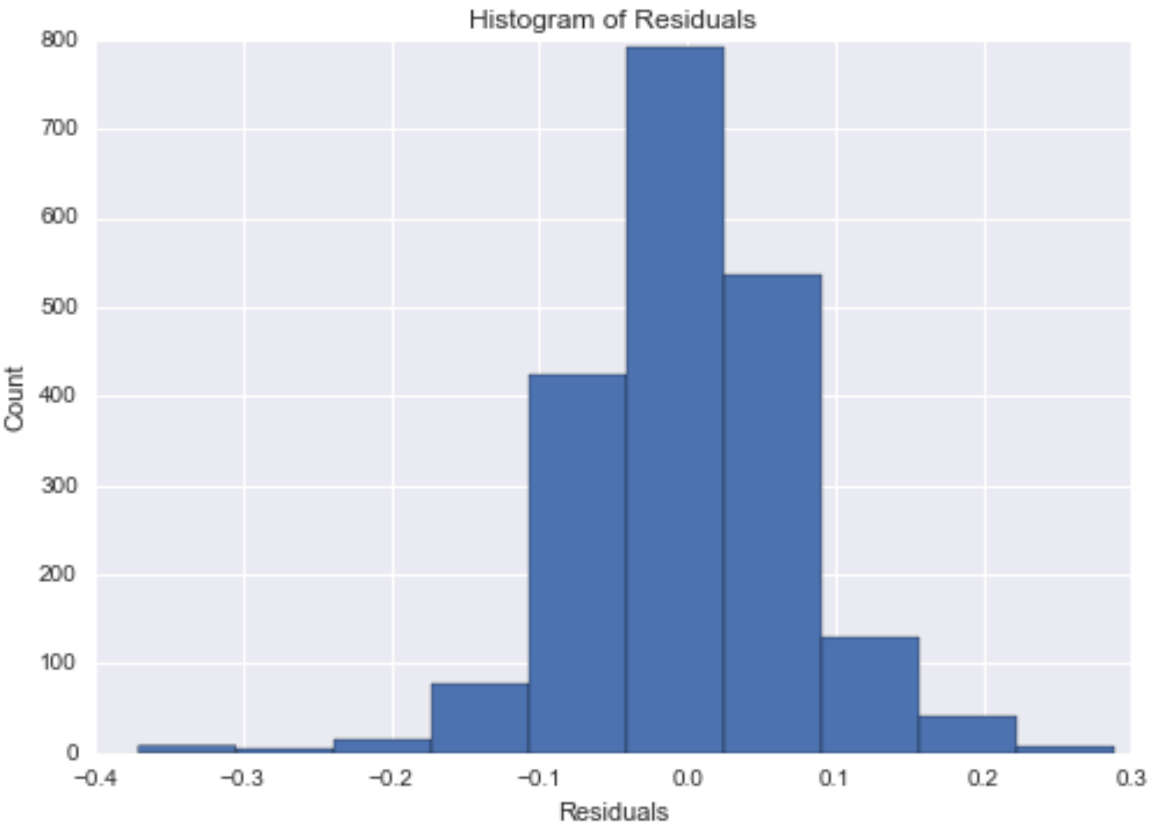
The OLS Regression Results from log(SalePrice) model has a very high correlation coefficient of 1.00, which is higher than the correlation coefficient of the SalePrice model. Although the F-statistic and t-statistic for all the coefficients are significant, similar to the SalePrice model, we should also look at the interaction effects between predictors in future analysis. A correlation coefficient value of 1.00 is very unusual and highly suggestive of multicollinearity. The residuals histogram of log(SalePrice) model reveals a more normal distribution than the SalePrice model. There is less kurtosis and skewness in the residual distribution of the log(SalePrice) model. In the QQ plot, the residuals tails more closely follow the line of normal distribution and are less skewed than those in the SalePrice model. The scatterplot of predicted log(SalePrice) and the residuals reveals a distribution more closely clustered in the center of the plot and more symmetrically distributed around zero than the SalePrice model residuals.

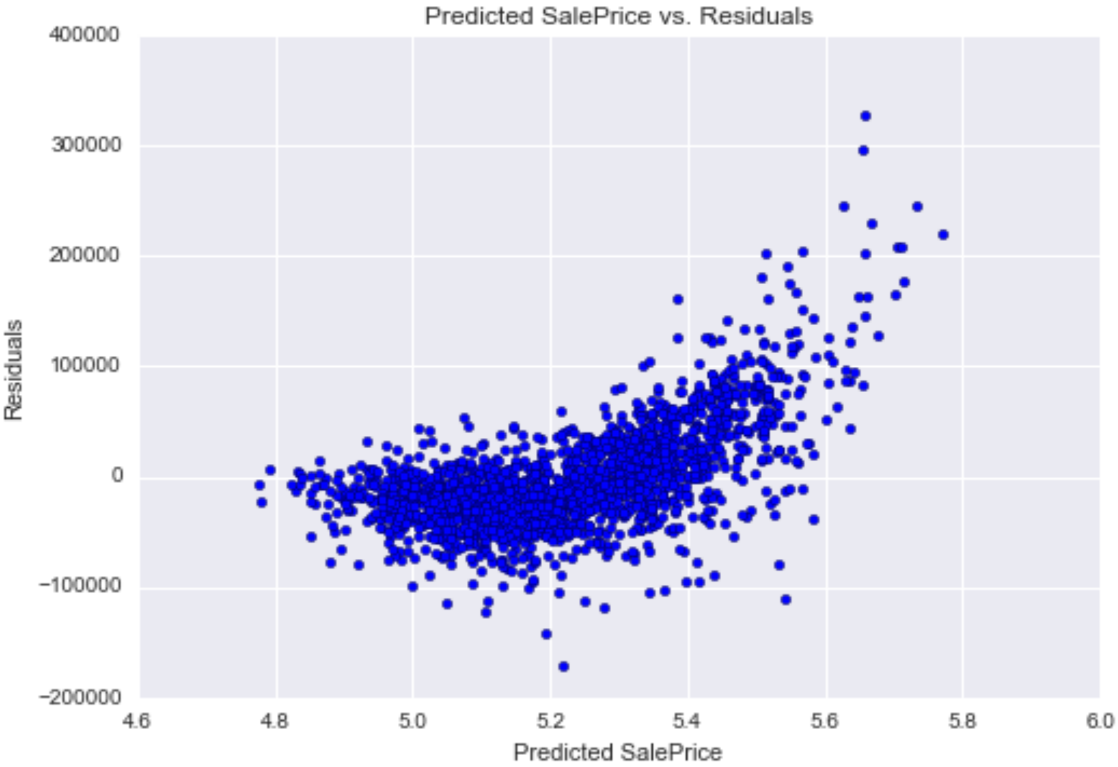
OLS Regression Results

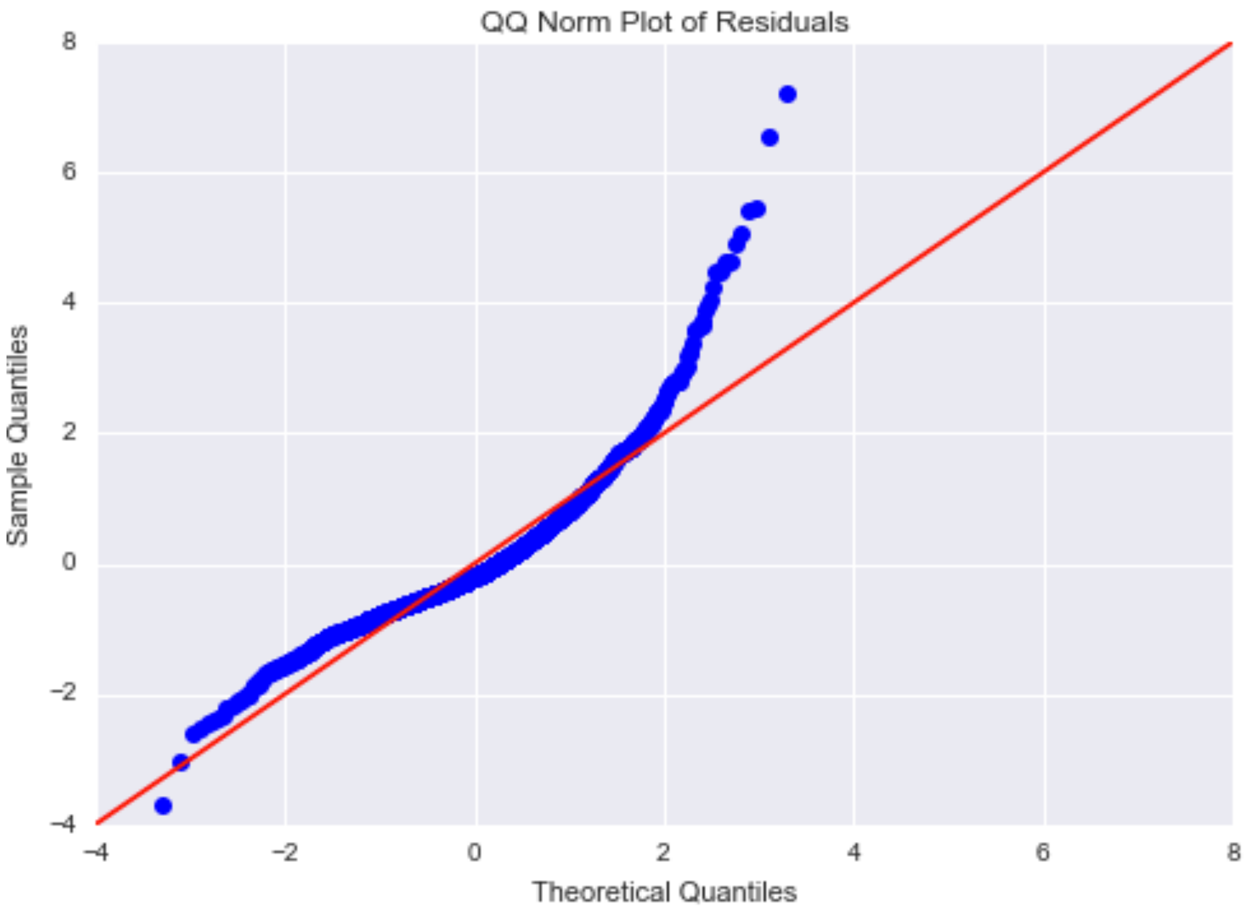
Dep. Variable:	saleprice	**R-squared:**	1.000
Model:	OLS	**Adj. R-squared:**	1.000
Method:	Least Squares	**F-statistic:**	2.069e+06
Date:	Fri, 20 Oct 2017	**Prob (F-statistic):**	0.00
Time:	21:18:54	**Log-Likelihood:**	2435.8
No. Observations:	2036	**AIC:**	-4862.
Df Residuals:	2031	**BIC:**	-4834.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
grlivarea	0.0001	4.22e-06	33.011	0.000	0.000	0.000
yearbuilt	0.0024	3.89e-06	618.138	0.000	0.002	0.002
overallqual	0.0292	0.002	17.630	0.000	0.026	0.032
totalbsmtsf	6.802e-05	4.79e-06	14.189	0.000	5.86e-05	7.74e-05
garagearea	5.841e-05	9.59e-06	6.090	0.000	3.96e-05	7.72e-05

Omnibus:	142.488	**Durbin-Watson:**	2.008
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	533.663
Skew:	-0.250	**Prob(JB):**	1.31e-116
Kurtosis:	5.458	**Cond. No.**	2.80e+03


Histogram of Residuals
Count
Residuals
800
700
600
500
400
300
200
100
0
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3

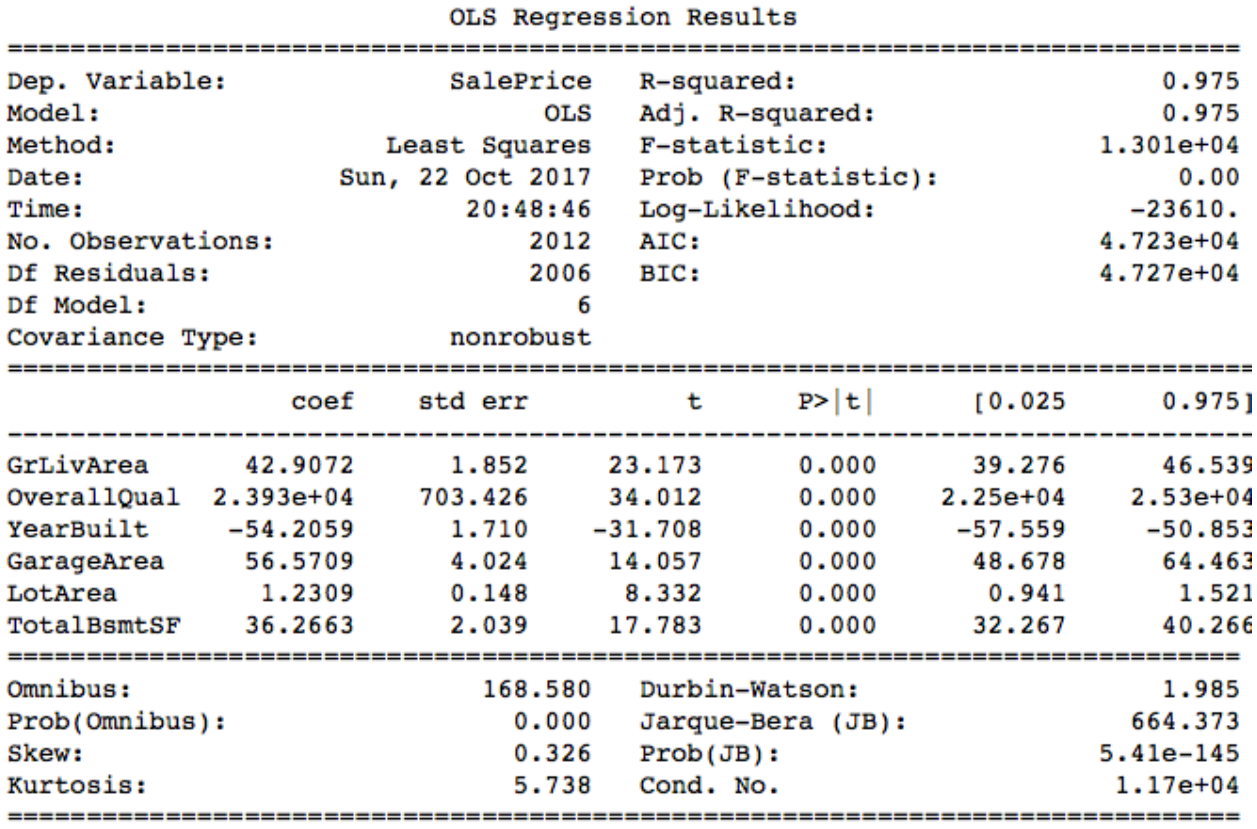

Predicted SalePrice vs. Residuals
Residuals
Predicted SalePrice
400000
300000
200000
100000
0
−100000
−200000
4.6
4.8
5.0
5.2
5.4
5.6
5.8
6.0



To conclude the comparison between these two models, the log transformation of response variable can help to eliminate the non-normal distribution. This transformation complies with the general assumption of the linear regression, that residuals follow a normal distribution. Techiniques like log transformation of predictor variables help to eliminate the outlier and helps to create normal distribution. Log transformation of both predictor and response variables can transform the non-linear model into linear one. However, we should be careful with techniques similar to log transformation because they can lead to an incorrect modeling by transforming non-linear relationships into linear.

Section 3. SELECT MODELS

Model 3 has chosen as the final product. This model producing the better adjusted r-squared value. Also P value and coefficients indicate that there is a strong correlation between all five variables and the SalePrice.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              SalePrice   R-squared:                       0.975
Model:                            OLS   Adj. R-squared:                  0.975
Method:                 Least Squares   F-statistic:                 1.301e+04
Date:                Sun, 22 Oct 2017   Prob (F-statistic):               0.00
Time:                        20:48:46   Log-Likelihood:                -23610.
No. Observations:                2012   AIC:                         4.723e+04
Df Residuals:                    2006   BIC:                         4.727e+04
Df Model:                           6
Covariance Type:            nonrobust
===============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
GrLivArea     42.9072      1.852     23.173      0.000      39.276      46.539
OverallQual 2.393e+04    703.426     34.012      0.000    2.25e+04    2.53e+04
YearBuilt    -54.2059      1.710    -31.708      0.000     -57.559     -50.853
GarageArea    56.5709      4.024     14.057      0.000      48.678      64.463
LotArea        1.2309      0.148      8.332      0.000       0.941       1.521
TotalBsmtSF   36.2663      2.039     17.783      0.000      32.267      40.266
==============================================================================
Omnibus:                      168.580   Durbin-Watson:                   1.985
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              664.373
Skew:                           0.326   Prob(JB):                    5.41e-145
Kurtosis:                       5.738   Cond. No.                     1.17e+04
==============================================================================
```

Section 4. Model Formula

```
p_saleprice = 42.907245     * GrLivArea    +
              23925.067709  * OverallQual  -
              54.205937     * yearbuilt    +
              56.570916     * garagearea   +
              1.230910      * lotarea      +
              36.266345     * totalbsmtsf
```

Conclusion

To summarize the findings TotalSqFtCalc, QualityIndex, YearBuilt and GarageArea appear to have the positive relationship with SalePrice and we reject the null hypothesis that they have no relationship with the SalePrice. A log transformation of the response variable validated that the assumption of normally distributed residuals for linear regression. Our models consistently performed better at SalePrices below 400,000, and an expanded dataset of high SalePrices could lessen the heterodasticity present in many of our models. In general, there are improvements that can be made to fix slight imbalances or abnormal features in our data.