Introduction:

The purpose of this analysis is to improve upon the first analysis using additional variables and a few additional techniques, such as grouping variables, creating new variables, transformation, and an automated selection algorithm. As with the first analysis, the predictor variables will be used to generate several linear regression models on the train data set, in order to predict the sale prices of homes in the test data set.

The analysis will begin with the cleansed and prepared data from the last analysis.

Data Recap:

The original data was updated in the following ways in order to attempt to make the data as accurate as possible and removing outliers to build the most accurate model:

- Home index number 543 had a garage built year of 2207, which seems to be incorrect. Given that this home was built in 2006 and remodeled in 2007, the garage built data was adjusted to 2007.
- Home index number 2884 had a missing value for garage built year with a detached garage. Given that this home was built in 1910 and remodeled in 1983, the garage year was adjusted to 1983, under the assumption that the home did not have a garage when it was originally built in 1910. Additionally, this garage was assigned an Unfinished type, as this is the most frequent finishing type for detached garages.
- Home index number 1851 had missing values for number of basement bathrooms. This home is a slab home, implying it does not have a basement, therefore these values have been updated to 0.
- Home index number 378 was also missing values for number of basement bathrooms, in addition to all other basement statistics such as basement square footage. Given all of the data is absent, the assumption was made that this home does not have a basement and all basement variables have been set accordingly.
- All missing values for Lot Frontage were replaced with the mean of the existing data due to the fact that no information exists in order to make an inference regarding these variables.
- Any home with Masonry Veneer Type and Masonry Veneer Area blank values were assigned a type of None and an Area of 0, given that these are the most frequent types in their respective variable categories.
- Any home with Masonry Veneer Type None and Masonry Veneer Area greater than 0 was adjusted to Masonry Veneer Area equal to 0.
- Any home with missing Basement Exposure Type was assigned No, given this is the most frequent Basement Exposure Type
- Any home with missing Electrical Type was assigned Standard Circuit Breakers, given this is the most frequent Electrical Type.
- All assumptions can be verified using methods such as using Google Street View (if home addresses were known), real estate sales records, public records such as building permits, or in-person home inspections.
- Appendix A contains a summary of the data with the adjustments described above.

In order to predict home values for a “typical” home, the sample data should contain “typical” homes, therefore, outliers and other oddities will be excluded from the data set. Exclusions are outlined below (and are not mutually exclusive):

- Homes with a sale price greater than or equal to $500k
- Homes with above grade living area greater than or equal to 3,000 square feet
- Homes with Agriculture, Commercial, or Industrial zoning
- Homes with Major or Severe Damage and Salvage Homes
- Homes with lot area greater than or equal to 50,000 square feet
- Only single-family homes were included

The final population of homes included in the data set is 1,640 observations. Summary statistics for the final population are included in **Appendix A**

Additional Model Building:

As with the first analysis, additional variables were chosen to “layer” over the last model chosen in the first analysis.

With the first analysis, the four variables chosen were neighborhood, above grade living area, finished basement square feet type 1, and overall quality. This combination of variables produced the largest R squared value and smallest AIC. This model produces an RMSE score of 35544.86 on Kaggle.

Model 1:

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.878
Model:                            OLS   Adj. R-squared:                  0.876
Method:                 Least Squares   F-statistic:                     504.6
Date:                Wed, 11 Oct 2017   Prob (F-statistic):               0.00
Time:                        21:45:05   Log-Likelihood:                -18956.
No. Observations:                1640   AIC:                         3.796e+04
Df Residuals:                    1616   BIC:                         3.809e+04
Df Model:                          23                                         
Covariance Type:            nonrobust                                         
=============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
Intercept                  -1.38e+04   1.55e+04     -0.889      0.374   -4.43e+04    1.66e+04
neighborhood[T.BrkSide] -2.711e+04   1.51e+04     -1.798      0.072   -5.67e+04    2463.571
neighborhood[T.ClearCr] -1237.5878   1.57e+04     -0.079      0.937   -3.19e+04    2.95e+04
neighborhood[T.CollgCr]  -584.2684   1.49e+04     -0.039      0.969   -2.98e+04    2.86e+04
neighborhood[T.Crawfor]    52.2418   1.51e+04      0.003      0.997   -2.96e+04    2.97e+04
neighborhood[T.Edwards]  -2.22e+04    1.5e+04     -1.479      0.139   -5.16e+04    7235.856
neighborhood[T.Gilbert] -6718.6658    1.5e+04     -0.449      0.653   -3.61e+04    2.26e+04
neighborhood[T.IDOTRR]  -3.416e+04   1.54e+04     -2.224      0.026   -6.43e+04   -4032.332
neighborhood[T.Mitchel] -1.207e+04   1.52e+04     -0.796      0.426   -4.18e+04    1.77e+04
neighborhood[T.NAmes]   -2.102e+04   1.49e+04     -1.412      0.158   -5.02e+04    8189.923
neighborhood[T.NWAmes]  -2.259e+04   1.51e+04     -1.499      0.134   -5.22e+04    6968.601
neighborhood[T.NoRidge]  2.725e+04   1.54e+04      1.775      0.076   -2864.541    5.74e+04
neighborhood[T.NridgHt]  6.597e+04   1.51e+04      4.359      0.000    3.63e+04    9.57e+04
neighborhood[T.OldTown] -3.421e+04   1.49e+04     -2.289      0.022   -6.35e+04   -4891.985
neighborhood[T.SWISU]   -3.655e+04   1.59e+04     -2.303      0.021   -6.77e+04   -5415.391
neighborhood[T.Sawyer]  -1.519e+04   1.51e+04     -1.008      0.314   -4.48e+04    1.44e+04
neighborhood[T.SawyerW] -1.583e+04   1.51e+04     -1.050      0.294   -4.54e+04    1.37e+04
neighborhood[T.Somerst]  2.614e+04    1.5e+04      1.738      0.082   -3364.924    5.57e+04
neighborhood[T.StoneBr]  7.025e+04   1.65e+04      4.266      0.000     3.8e+04    1.03e+05
neighborhood[T.Timber]   1.631e+04   1.52e+04      1.074      0.283   -1.35e+04    4.61e+04
neighborhood[T.Veenker] -6736.3733   1.66e+04     -0.407      0.684   -3.92e+04    2.57e+04
grlivarea                  54.0394      1.907     28.336      0.000      50.299      57.780
overallqual              1.763e+04    781.737     22.557      0.000    1.61e+04    1.92e+04
bsmtfinsf1                 36.0406      1.631     22.099      0.000      32.842      39.239
==============================================================================
Omnibus:                      155.158   Durbin-Watson:                   2.021
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              813.729
Skew:                           0.269   Prob(JB):                    2.00e-177
Kurtosis:                       6.409   Cond. No.                     1.74e+05
==============================================================================
```

In an additional model, lot area is layered with the first four variables. This produces a higher R squared value and a lower AIC. However, this produces an RMSE score of 36055.49 on Kaggle, which is not an improvement over the first analysis.

Model 2:

```
OLS Regression Results
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.881
Model:                            OLS   Adj. R-squared:                  0.880
Method:                 Least Squares   F-statistic:                     499.8
Date:                Sat, 14 Oct 2017   Prob (F-statistic):               0.00
Time:                        13:56:49   Log-Likelihood:                -18932.
No. Observations:                1640   AIC:                         3.791e+04
Df Residuals:                    1615   BIC:                         3.805e+04
Df Model:                          24
Covariance Type:            nonrobust
============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept               -1.651e+04   1.53e+04     -1.079      0.281   -4.65e+04    1.35e+04
neighborhood[T.BrkSide] -3.074e+04   1.49e+04     -2.067      0.039   -5.99e+04   -1572.806
neighborhood[T.ClearCr] -1.512e+04   1.56e+04     -0.972      0.331   -4.56e+04    1.54e+04
neighborhood[T.CollgCr] -7707.6756   1.47e+04     -0.524      0.600   -3.66e+04    2.11e+04
neighborhood[T.Crawfor] -9142.9507    1.5e+04     -0.611      0.541   -3.85e+04    2.02e+04
neighborhood[T.Edwards] -2.919e+04   1.48e+04     -1.969      0.049   -5.83e+04    -111.569
neighborhood[T.Gilbert] -1.428e+04   1.48e+04     -0.966      0.334   -4.33e+04    1.47e+04
neighborhood[T.IDOTRR]  -3.851e+04   1.52e+04     -2.541      0.011   -6.82e+04   -8786.521
neighborhood[T.Mitchel] -2.248e+04    1.5e+04     -1.496      0.135   -5.19e+04    6985.867
neighborhood[T.NAmes]   -2.809e+04   1.47e+04     -1.908      0.057    -5.7e+04     779.882
neighborhood[T.NWAmes]  -3.037e+04   1.49e+04     -2.039      0.042   -5.96e+04   -1149.822
neighborhood[T.NoRidge]  1.967e+04   1.52e+04      1.296      0.195   -1.01e+04    4.94e+04
neighborhood[T.NridgHt]  5.721e+04    1.5e+04      3.821      0.000    2.78e+04    8.66e+04
neighborhood[T.OldTown] -3.856e+04   1.47e+04     -2.614      0.009   -6.75e+04   -9628.880
neighborhood[T.SWISU]   -3.954e+04   1.57e+04     -2.526      0.012   -7.02e+04   -8839.997
neighborhood[T.Sawyer]  -2.298e+04   1.49e+04     -1.543      0.123   -5.22e+04    6240.900
neighborhood[T.SawyerW] -2.274e+04   1.49e+04     -1.527      0.127    -5.2e+04    6478.185
neighborhood[T.Somerst]  1.977e+04   1.49e+04      1.331      0.184   -9372.189    4.89e+04
neighborhood[T.StoneBr]  6.232e+04   1.63e+04      3.830      0.000    3.04e+04    9.42e+04
neighborhood[T.Timber]   7508.7218    1.5e+04      0.500      0.617    -2.2e+04     3.7e+04
neighborhood[T.Veenker] -1.853e+04   1.64e+04     -1.129      0.259   -5.07e+04    1.37e+04
grlivarea                  50.5360      1.946     25.967      0.000      46.719      54.353
overallqual              1.801e+04    772.452     23.317      0.000    1.65e+04    1.95e+04
bsmtfinsf1                 35.0938      1.613     21.753      0.000      31.929      38.258
lotarea                     1.2675      0.182      6.950      0.000       0.910       1.625
==============================================================================
Omnibus:                      157.483   Durbin-Watson:                   2.010
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1003.088
Skew:                           0.162   Prob(JB):                    1.52e-218
Kurtosis:                       6.818   Cond. No.                     1.21e+06
==============================================================================
```

To further explore the data and attempt to improve the model, overall condition was added as a variable. This increases the R squared value and decreases the AIC, but this produces an RMSE score of 35770.32, which is not an improvement from the first analysis.

Model 3:

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.885
Model:                            OLS   Adj. R-squared:                  0.883
Method:                 Least Squares   F-statistic:                     495.5
Date:                Sat, 14 Oct 2017   Prob (F-statistic):               0.00
Time:                        14:23:31   Log-Likelihood:                -18908.
No. Observations:                1640   AIC:                         3.787e+04
Df Residuals:                    1614   BIC:                         3.801e+04
Df Model:                          25                                         
Covariance Type:            nonrobust                                         
=============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
Intercept                 -3.452e+04   1.53e+04     -2.254      0.024   -6.46e+04   -4479.894
neighborhood[T.BrkSide]   -3.772e+04   1.47e+04     -2.567      0.010   -6.65e+04   -8899.622
neighborhood[T.ClearCr]    -1.83e+04   1.53e+04     -1.192      0.233   -4.84e+04    1.18e+04
neighborhood[T.CollgCr]   -9197.0366   1.45e+04     -0.634      0.526   -3.76e+04    1.93e+04
neighborhood[T.Crawfor]   -1.664e+04   1.48e+04     -1.124      0.261   -4.57e+04    1.24e+04
neighborhood[T.Edwards]   -3.358e+04   1.46e+04     -2.295      0.022   -6.23e+04   -4884.172
neighborhood[T.Gilbert]   -1.562e+04   1.46e+04     -1.072      0.284   -4.42e+04     1.3e+04
neighborhood[T.IDOTRR]     -4.48e+04    1.5e+04     -2.993      0.003   -7.42e+04   -1.54e+04
neighborhood[T.Mitchel]   -2.636e+04   1.48e+04     -1.779      0.075   -5.54e+04    2706.604
neighborhood[T.NAmes]     -3.353e+04   1.45e+04     -2.308      0.021    -6.2e+04   -5030.953
neighborhood[T.NWAmes]    -3.527e+04   1.47e+04     -2.399      0.017   -6.41e+04   -6426.917
neighborhood[T.NoRidge]    1.818e+04    1.5e+04      1.215      0.224   -1.12e+04    4.75e+04
neighborhood[T.NridgHt]    5.743e+04   1.48e+04      3.891      0.000    2.85e+04    8.64e+04
neighborhood[T.OldTown]   -4.654e+04   1.46e+04     -3.191      0.001   -7.51e+04   -1.79e+04
neighborhood[T.SWISU]     -4.461e+04   1.54e+04     -2.888      0.004   -7.49e+04   -1.43e+04
neighborhood[T.Sawyer]     -2.82e+04   1.47e+04     -1.917      0.055   -5.71e+04     647.346
neighborhood[T.SawyerW]   -2.427e+04   1.47e+04     -1.652      0.099   -5.31e+04    4542.835
neighborhood[T.Somerst]    1.957e+04   1.46e+04      1.336      0.182   -9155.677    4.83e+04
neighborhood[T.StoneBr]    6.221e+04    1.6e+04      3.878      0.000    3.07e+04    9.37e+04
neighborhood[T.Timber]     6186.1617   1.48e+04      0.417      0.676   -2.29e+04    3.53e+04
neighborhood[T.Veenker]   -2.299e+04   1.62e+04     -1.420      0.156   -5.47e+04    8772.898
grlivarea                    51.6629      1.926     26.829      0.000      47.886      55.440
overallqual                1.728e+04    768.931     22.470      0.000    1.58e+04    1.88e+04
bsmtfinsf1                   34.9186      1.591     21.952      0.000      31.799      38.039
lotarea                       1.2767      0.180      7.101      0.000       0.924       1.629
overallcond                4338.3994    628.732      6.900      0.000    3105.183    5571.616
==============================================================================
Omnibus:                      164.150   Durbin-Watson:                   2.017
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              931.338
Skew:                           0.271   Prob(JB):                    5.79e-203
Kurtosis:                       6.652   Cond. No.                     1.21e+06
==============================================================================
```

The next model combines all of the previous variables and includes year built. This also increases the R squared value and decreases the AIC, but this produces an RMSE score of 36171.45 on Kaggle which is not better than the previous analysis and is not an improvement over model 3.

Model 4:

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.894
Model:                            OLS   Adj. R-squared:                  0.893
Method:                 Least Squares   F-statistic:                     525.4
Date:                Sat, 14 Oct 2017   Prob (F-statistic):               0.00
Time:                        15:02:00   Log-Likelihood:                -18836.
No. Observations:                1640   AIC:                         3.773e+04
Df Residuals:                    1613   BIC:                         3.787e+04
Df Model:                          26                                         
Covariance Type:            nonrobust                                         
=============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
Intercept                 -1.148e+06   9.29e+04    -12.364      0.000   -1.33e+06   -9.66e+05
neighborhood[T.BrkSide] -1602.1994   1.44e+04     -0.111      0.911   -2.98e+04    2.66e+04
neighborhood[T.ClearCr]  -584.7951   1.48e+04     -0.040      0.968   -2.96e+04    2.84e+04
neighborhood[T.CollgCr] -5532.9942   1.39e+04     -0.398      0.690   -3.28e+04    2.17e+04
neighborhood[T.Crawfor]  1.401e+04   1.44e+04      0.974      0.330   -1.42e+04    4.22e+04
neighborhood[T.Edwards] -9378.5027   1.42e+04     -0.663      0.508   -3.71e+04    1.84e+04
neighborhood[T.Gilbert] -1.395e+04    1.4e+04     -0.999      0.318   -4.13e+04    1.34e+04
neighborhood[T.IDOTRR]  -7038.8908   1.47e+04     -0.480      0.631   -3.58e+04    2.17e+04
neighborhood[T.Mitchel]  -1.57e+04   1.42e+04     -1.104      0.270   -4.36e+04    1.22e+04
neighborhood[T.NAmes]   -1.183e+04    1.4e+04     -0.843      0.399   -3.93e+04    1.57e+04
neighborhood[T.NWAmes]  -2.125e+04   1.41e+04     -1.504      0.133    -4.9e+04    6461.150
neighborhood[T.NoRidge]     2.4e+04   1.43e+04      1.674      0.094   -4113.407    5.21e+04
neighborhood[T.NridgHt]    5.99e+04   1.41e+04      4.238      0.000    3.22e+04    8.76e+04
neighborhood[T.OldTown] -6797.4441   1.43e+04     -0.474      0.636   -3.49e+04    2.13e+04
neighborhood[T.SWISU]   -7482.0078   1.51e+04     -0.495      0.620   -3.71e+04    2.21e+04
neighborhood[T.Sawyer]  -9751.8091   1.42e+04     -0.688      0.491   -3.75e+04     1.8e+04
neighborhood[T.SawyerW] -1.879e+04   1.41e+04     -1.335      0.182   -4.64e+04    8810.890
neighborhood[T.Somerst]    2.06e+04    1.4e+04      1.469      0.142   -6904.630    4.81e+04
neighborhood[T.StoneBr]  6.681e+04   1.54e+04      4.348      0.000    3.67e+04    9.69e+04
neighborhood[T.Timber]   1.049e+04   1.42e+04      0.739      0.460   -1.73e+04    3.83e+04
neighborhood[T.Veenker] -1.001e+04   1.55e+04     -0.644      0.520   -4.05e+04    2.05e+04
grlivarea                   53.3232      1.849     28.842      0.000      49.697      56.950
overallqual               1.513e+04    757.093     19.991      0.000    1.36e+04    1.66e+04
bsmtfinsf1                  32.6117      1.535     21.247      0.000      29.601      35.622
lotarea                      1.4814      0.173      8.564      0.000       1.142       1.821
overallcond               5927.5843    616.062      9.622      0.000    4719.219    7135.950
yearbuilt                  556.9301     45.858     12.145      0.000     466.983     646.877
==============================================================================
Omnibus:                      199.472   Durbin-Watson:                   2.048
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1187.660
Skew:                           0.394   Prob(JB):                    1.27e-258
Kurtosis:                       7.094   Cond. No.                     1.84e+06
==============================================================================
```

The final model built for this exploration combines all previous variables plus total basement square footage. This improves the R squared value and AIC, but this produces an RMSE score of 37184.18 on Kaggle which is not better than the previous analysis and is not an improvement over the previous model.

Model 5:

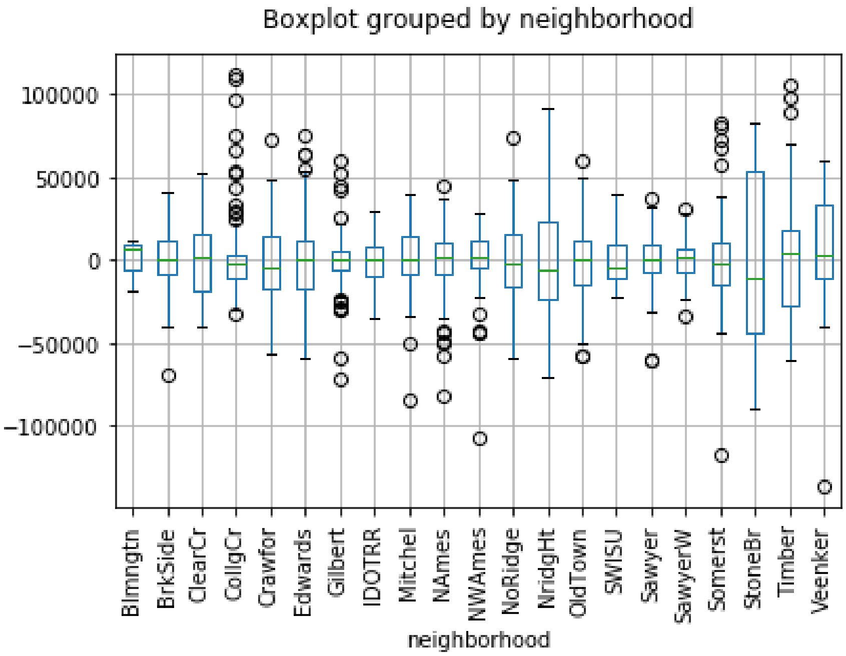
```
                                      OLS Regression Results
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.902
Model:                            OLS   Adj. R-squared:                  0.901
Method:                 Least Squares   F-statistic:                     550.5
Date:                Sat, 14 Oct 2017   Prob (F-statistic):               0.00
Time:                        15:13:43   Log-Likelihood:                -18773.
No. Observations:                1640   AIC:                         3.760e+04
Df Residuals:                    1612   BIC:                         3.775e+04
Df Model:                          27
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Intercept               -1.101e+06   8.95e+04    -12.302      0.000   -1.28e+06   -9.25e+05
neighborhood[T.BrkSide]  4371.6316   1.39e+04      0.316      0.752   -2.28e+04    3.15e+04
neighborhood[T.ClearCr]  5109.6917   1.42e+04      0.359      0.720   -2.28e+04     3.3e+04
neighborhood[T.CollgCr]   125.4971   1.34e+04      0.009      0.993   -2.61e+04    2.64e+04
neighborhood[T.Crawfor]  1.981e+04   1.39e+04      1.429      0.153   -7384.608     4.7e+04
neighborhood[T.Edwards] -2344.4757   1.36e+04     -0.172      0.864   -2.91e+04    2.44e+04
neighborhood[T.Gilbert] -3401.9296   1.35e+04     -0.252      0.801   -2.98e+04     2.3e+04
neighborhood[T.IDOTRR]  -2079.6959   1.41e+04     -0.147      0.883   -2.98e+04    2.56e+04
neighborhood[T.Mitchel] -8652.6101   1.37e+04     -0.631      0.528   -3.55e+04    1.82e+04
neighborhood[T.NAmes]   -7444.5653   1.35e+04     -0.551      0.582   -3.39e+04    1.91e+04
neighborhood[T.NWAmes]  -1.595e+04   1.36e+04     -1.172      0.241   -4.26e+04    1.07e+04
neighborhood[T.NoRidge]  2.939e+04   1.38e+04      2.129      0.033    2309.357    5.65e+04
neighborhood[T.NridgHt]  6.041e+04   1.36e+04      4.439      0.000    3.37e+04    8.71e+04
neighborhood[T.OldTown] -1783.6624   1.38e+04     -0.129      0.897   -2.89e+04    2.53e+04
neighborhood[T.SWISU]   -1706.0068   1.46e+04     -0.117      0.907   -3.02e+04    2.68e+04
neighborhood[T.Sawyer]  -3833.7749   1.36e+04     -0.281      0.779   -3.06e+04    2.29e+04
neighborhood[T.SawyerW] -1.086e+04   1.36e+04     -0.801      0.423   -3.75e+04    1.57e+04
neighborhood[T.Somerst]  2.388e+04   1.35e+04      1.768      0.077   -2609.944    5.04e+04
neighborhood[T.StoneBr]  6.889e+04   1.48e+04      4.657      0.000    3.99e+04    9.79e+04
neighborhood[T.Timber]   1.436e+04   1.37e+04      1.051      0.294   -1.24e+04    4.12e+04
neighborhood[T.Veenker] -7992.0425    1.5e+04     -0.534      0.593   -3.73e+04    2.14e+04
grlivarea                  52.4370      1.782     29.429      0.000      48.942      55.932
overallqual              1.328e+04    747.238     17.766      0.000    1.18e+04    1.47e+04
bsmtfinsf1                 25.4455      1.608     15.826      0.000      22.292      28.599
lotarea                     1.2649      0.168      7.546      0.000       0.936       1.594
overallcond              6825.7220    598.440     11.406      0.000    5651.919    7999.525
yearbuilt                 524.4540     44.246     11.853      0.000     437.669     611.239
totalbsmtsf                23.2964      2.059     11.314      0.000      19.258      27.335
==============================================================================
Omnibus:                      177.122   Durbin-Watson:                   2.039
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1151.967
Skew:                           0.262   Prob(JB):                    7.14e-251
Kurtosis:                       7.072   Cond. No.                     1.85e+06
==============================================================================
```

Although each, additional model has improvements of the R squared and AIC, the RMSE is not improving. This implies that the addition of variables to the model is overfitting the model and not actually improving it.

Neighborhood Accuracy:

In order to attempt to further improve the model, an exploration of the accuracy of sale price by neighborhood was completed.

Figure 1 – Boxplot of residuals by neighborhood:



The boxplot shows that while some neighborhoods are predicted accurately, other neighborhoods are consistently over-predicted or under predicted.

Figure 2 – Heat mapped chart of median residual by neighborhood

Neighborhood	Number of homes	Median Residual
StoneBr	13	-11162.14
NridgHt	74	-5358.71
SWISU	20	-5098.21
Crawfor	66	-4140.75
Somerst	78	-2716.29
CollgCr	176	-2698.60
NoRidge	46	-2660.72
BrkSide	78	-18.46
OldTown	152	183.54
Edwards	122	303.10
Sawyer	85	321.81
Mitchel	62	821.83
Gilbert	111	858.40
IDOTRR	39	976.39
SawyerW	70	1239.23
NAmes	274	1390.64
ClearCr	26	1483.46
NWAmes	80	2254.94
Veenker	12	2793.03
Timber	53	4346.01
Blmngtn	3	6291.93

The chart in Figure 2 illustrates the median residuals by neighborhood. The larger negative median residuals (color coded red) are consistently over predicted, while the larger positive median residuals (color coded blue) are consistently under predicted. The residuals close to zero are lightly colored and fall in the center of the chart and are better predicted by the model.

To explore increasing prediction accuracy in individual neighborhoods, the average actual cost per square foot by neighborhood and the average estimated cost by square foot was calculated.

Figure 3 – Actual cost per square foot by neighborhood (sorted by lowest actual average cost per square foot to highest)

neighborhood	count	mean	std	min	25%	50%	75%	max
SWISU	20	79.58	17.80	53.14	66.87	75.54	92.00	117.98
NAmes	274	82.53	16.45	47.68	71.49	81.37	89.91	134.32
Veenker	12	82.78	20.79	44.72	74.48	83.79	86.69	129.82
NWAmes	80	83.00	14.82	36.95	74.88	81.41	88.22	124.75
ClearCr	26	83.40	13.80	53.53	74.39	84.88	93.35	104.85
Edwards	122	83.73	23.13	47.02	68.06	80.95	94.73	169.05
Sawyer	85	84.16	17.47	45.10	73.26	83.34	89.55	160.71
IDOTRR	39	85.98	19.14	41.08	75.28	86.17	93.47	132.61
OldTown	152	86.07	21.94	39.15	73.15	86.22	99.84	147.08
Mitchel	62	86.36	19.45	33.07	75.14	84.46	93.75	140.35
BrkSide	78	89.17	19.76	49.70	76.68	85.36	100.26	154.57
SawyerW	70	94.99	17.80	68.69	81.77	90.49	106.36	133.43
Crawfor	66	96.47	19.55	63.70	80.44	94.74	111.02	151.62
NoRidge	46	97.71	15.66	71.57	87.71	93.10	104.87	143.43
CollgCr	176	103.23	23.53	68.25	86.64	94.64	117.06	187.08
Gilbert	111	105.13	18.58	69.63	88.75	106.69	118.80	143.71
Timber	53	108.24	31.04	56.49	85.35	106.05	120.16	214.91
StoneBr	13	117.32	21.32	92.72	99.03	116.54	129.79	165.82
Somerst	78	123.76	25.55	81.89	100.93	118.33	148.23	182.64
NridgHt	74	125.54	27.03	88.00	105.70	119.27	136.57	206.44
Blmngtn	3	141.15	13.02	126.30	136.41	146.51	148.57	150.63

Figure 4 – Estimated cost per square foot by neighborhood from the model (sorted by lowest estimated average cost per square foot to highest)

neighborhood	count	mean	std	min	25%	50%	75%	max
SWISU	20	79.80	23.03	46.92	62.63	79.50	92.28	124.62
Veenker	12	82.56	34.84	3.92	73.95	82.85	94.83	146.08
NWAmes	80	83.20	21.26	-10.83	72.68	83.36	94.37	137.38
ClearCr	26	83.67	21.36	44.07	66.02	91.14	100.34	121.76
NAmes	274	83.85	23.92	26.05	69.46	81.20	95.72	165.38
Edwards	122	85.05	36.15	0.69	59.36	75.73	102.84	212.65
Sawyer	85	85.05	24.11	22.74	73.78	83.61	94.52	176.53
Mitchel	62	86.77	27.13	-2.02	73.49	87.62	96.63	156.99
IDOTRR	39	88.07	28.37	28.27	72.37	87.51	101.76	174.77
OldTown	152	89.04	33.90	20.63	67.67	86.18	110.47	182.94
BrkSide	78	89.93	28.02	23.60	74.88	86.76	109.56	157.51
SawyerW	70	94.54	18.16	60.93	78.74	93.47	107.75	133.51
Crawfor	66	96.71	29.42	29.54	75.14	90.84	112.76	173.42
NoRidge	46	97.18	20.63	50.64	84.27	95.09	107.04	163.09
CollgCr	176	102.78	29.53	57.60	83.11	94.65	118.10	237.05
Gilbert	111	105.39	24.58	41.30	87.34	105.16	121.88	186.26
Timber	53	107.26	44.25	38.08	68.95	105.77	126.82	272.80
StoneBr	13	115.94	37.51	67.18	89.22	121.92	127.28	200.78
Somerst	78	121.74	29.81	16.81	101.57	117.95	139.76	206.87
NridgHt	74	124.09	34.71	67.57	102.00	117.53	135.13	239.07
Blmngtn	3	141.20	26.04	111.58	131.53	151.48	156.01	160.53

To attempt to improve the accuracy of sale price prediction by neighborhood, the neighborhood groups were created. Neighborhoods were grouped using a combination of two criteria: similar actual cost per square foot and similar median residual by neighborhood. This created 6 neighborhood groups to be utilized in the model.

Figure 5 – Median residuals, average actual cost per square foot, and neighborhood group

Neighborhood	Number of homes	Median Residual	Mean Actual Cost per Square Foot	Neighborhood Group
StoneBr	13	-11162.14	117.32	1
NridgHt	74	-5358.71	125.54	1
SWISU	20	-5098.21	79.58	2
Crawfor	66	-4140.75	96.47	3
Somerst	78	-2716.29	123.76	3
CollgCr	176	-2698.60	103.23	3
NoRidge	46	-2660.72	97.71	3
BrkSide	78	-18.46	89.17	3
OldTown	152	183.54	86.07	4
Edwards	122	303.10	83.73	4
Sawyer	85	321.81	84.16	4
Mitchel	62	821.83	86.36	4
Gilbert	111	858.40	105.13	6
IDOTRR	39	976.39	85.98	5
SawyerW	70	1239.23	94.99	5
NAmes	274	1390.64	82.53	5
ClearCr	26	1483.46	83.40	5
NWAmes	80	2254.94	83.00	5
Veenker	12	2793.03	82.78	5
Timber	53	4346.01	108.24	6
Blmngtn	3	6291.93	141.15	6

The neighborhood groups were then run in the regression model to view the accuracy of the sale price predictions. Neighborhood group 1 was used as the base group in the regression.

Model 6:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.454
Model:                            OLS   Adj. R-squared:                  0.453
Method:                 Least Squares   F-statistic:                     272.0
Date:                Sun, 15 Oct 2017   Prob (F-statistic):          7.09e-212
Time:                        22:53:31   Log-Likelihood:                -20183.
No. Observations:                1640   AIC:                         4.038e+04
Df Residuals:                    1634   BIC:                         4.041e+04
Df Model:                           5
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept     3.418e+05   5747.623     59.475      0.000    3.31e+05    3.53e+05
nbgroup[T.2] -2.001e+05   1.33e+04    -15.053      0.000   -2.26e+05   -1.74e+05
nbgroup[T.3]  -1.34e+05   6285.561    -21.322      0.000   -1.46e+05   -1.22e+05
nbgroup[T.4] -2.084e+05   6313.630    -33.003      0.000   -2.21e+05   -1.96e+05
nbgroup[T.5] -1.822e+05   6226.702    -29.261      0.000   -1.94e+05    -1.7e+05
nbgroup[T.6]  -1.33e+05   7088.379    -18.769      0.000   -1.47e+05   -1.19e+05
==============================================================================
Omnibus:                      225.242   Durbin-Watson:                   1.977
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              423.277
Skew:                           0.857   Prob(JB):                     1.22e-92
Kurtosis:                       4.805   Cond. No.                         13.3
==============================================================================
```

This model was compared to using ungrouped neighborhoods as a variable alone.

Model 7:

```
                                      OLS Regression Results
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.648
Model:                            OLS   Adj. R-squared:                  0.643
Method:                 Least Squares   F-statistic:                     148.8
Date:                Sun, 15 Oct 2017   Prob (F-statistic):               0.00
Time:                        22:47:15   Log-Likelihood:                -19824.
No. Observations:                1640   AIC:                         3.969e+04
Df Residuals:                    1619   BIC:                         3.980e+04
Df Model:                          20
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Intercept                1.777e+05    2.5e+04      7.112      0.000    1.29e+05    2.27e+05
neighborhood[T.BrkSide] -5.165e+04   2.55e+04     -2.029      0.043   -1.02e+05   -1711.633
neighborhood[T.ClearCr]  2.942e+04   2.64e+04      1.115      0.265   -2.23e+04    8.12e+04
neighborhood[T.CollgCr]   2.56e+04   2.52e+04      1.016      0.310   -2.38e+04     7.5e+04
neighborhood[T.Crawfor]  1.831e+04   2.55e+04      0.717      0.474   -3.18e+04    6.84e+04
neighborhood[T.Edwards] -4.896e+04   2.53e+04     -1.936      0.053   -9.86e+04     642.413
neighborhood[T.Gilbert]  1.411e+04   2.53e+04      0.557      0.577   -3.56e+04    6.38e+04
neighborhood[T.IDOTRR]  -6.654e+04   2.59e+04     -2.566      0.010   -1.17e+05   -1.57e+04
neighborhood[T.Mitchel] -1.362e+04   2.56e+04     -0.532      0.595   -6.38e+04    3.66e+04
neighborhood[T.NAmes]    -3.27e+04   2.51e+04     -1.302      0.193    -8.2e+04    1.66e+04
neighborhood[T.NWAmes]   7696.6250   2.54e+04      0.302      0.762   -4.22e+04    5.76e+04
neighborhood[T.NoRidge]  1.239e+05   2.58e+04      4.803      0.000    7.33e+04    1.74e+05
neighborhood[T.NridgHt]  1.627e+05   2.55e+04      6.385      0.000    1.13e+05    2.13e+05
neighborhood[T.OldTown] -5.513e+04   2.52e+04     -2.185      0.029   -1.05e+05   -5638.789
neighborhood[T.SWISU]   -3.597e+04   2.68e+04     -1.342      0.180   -8.85e+04    1.66e+04
neighborhood[T.Sawyer]  -4.023e+04   2.54e+04     -1.583      0.114   -9.01e+04    9631.681
neighborhood[T.SawyerW]  7350.7143   2.55e+04      0.288      0.773   -4.27e+04    5.74e+04
neighborhood[T.Somerst]  7.686e+04   2.55e+04      3.019      0.003    2.69e+04    1.27e+05
neighborhood[T.StoneBr]  1.722e+05   2.77e+04      6.214      0.000    1.18e+05    2.27e+05
neighborhood[T.Timber]   6.846e+04   2.57e+04      2.665      0.008    1.81e+04    1.19e+05
neighborhood[T.Veenker]  5.142e+04   2.79e+04      1.841      0.066   -3366.890    1.06e+05
==============================================================================
Omnibus:                      306.368   Durbin-Watson:                   2.049
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              797.697
Skew:                           0.993   Prob(JB):                    6.06e-174
Kurtosis:                       5.780   Cond. No.                         112.
==============================================================================
```

The model using ungrouped neighborhoods has a higher R squared value and lower AIC value than the model using grouped neighborhoods, therefore, grouped neighborhoods will not be used in the model for this analysis.

Additional variables:

A few variables were created to explore the combination of variables and their impact on their ability to predict sale price.

A "total square foot" variable was created by adding the above grade living area, finished basement type 1, and finished basement type 2. This variable was used in conjunction with the base model developed in the first analysis. This model produces a slightly lower R squared value, a higher AIC value, and a higher RMSE on Kaggle of 36151.03201, so this model will not be used.

Model 8:

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.875
Model:                            OLS   Adj. R-squared:                  0.873
Method:                 Least Squares   F-statistic:                     512.2
Date:                Tue, 17 Oct 2017   Prob (F-statistic):               0.00
Time:                        10:26:34   Log-Likelihood:                -18977.
No. Observations:                1640   AIC:                         3.800e+04
Df Residuals:                    1617   BIC:                         3.813e+04
Df Model:                          22                                         
Covariance Type:            nonrobust                                         
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept               -1.365e+04   1.57e+04     -0.868      0.385   -4.45e+04    1.72e+04
neighborhood[T.BrkSide] -2.593e+04   1.53e+04     -1.698      0.090   -5.59e+04    4024.475
neighborhood[T.ClearCr] -3920.6126   1.59e+04     -0.247      0.805   -3.51e+04    2.72e+04
neighborhood[T.CollgCr] -1958.8355   1.51e+04     -0.130      0.897   -3.15e+04    2.76e+04
neighborhood[T.Crawfor]   797.2269   1.53e+04      0.052      0.959   -2.93e+04    3.09e+04
neighborhood[T.Edwards] -2.216e+04   1.52e+04     -1.458      0.145    -5.2e+04    7656.589
neighborhood[T.Gilbert] -3170.9336   1.51e+04     -0.209      0.834   -3.29e+04    2.65e+04
neighborhood[T.IDOTRR]    -3.21e+04   1.56e+04     -2.063      0.039   -6.26e+04   -1586.312
neighborhood[T.Mitchel] -1.482e+04   1.54e+04     -0.965      0.335    -4.5e+04    1.53e+04
neighborhood[T.NAmes]   -2.416e+04   1.51e+04     -1.601      0.110   -5.38e+04    5437.293
neighborhood[T.NWAmes]  -2.331e+04   1.53e+04     -1.526      0.127   -5.33e+04    6644.518
neighborhood[T.NoRidge]  2.896e+04   1.55e+04      1.863      0.063   -1533.952    5.95e+04
neighborhood[T.NridgHt]  6.689e+04   1.53e+04      4.363      0.000    3.68e+04     9.7e+04
neighborhood[T.OldTown] -3.161e+04   1.51e+04     -2.088      0.037   -6.13e+04   -1915.017
neighborhood[T.SWISU]     -3.26e+04   1.61e+04     -2.029      0.043   -6.41e+04   -1084.427
neighborhood[T.Sawyer]    -2.19e+04   1.53e+04     -1.434      0.152   -5.19e+04    8054.057
neighborhood[T.SawyerW] -1.628e+04   1.53e+04     -1.065      0.287   -4.62e+04    1.37e+04
neighborhood[T.Somerst]  2.653e+04   1.52e+04      1.741      0.082   -3361.051    5.64e+04
neighborhood[T.StoneBr]  7.001e+04   1.67e+04      4.199      0.000    3.73e+04    1.03e+05
neighborhood[T.Timber]   1.509e+04   1.54e+04      0.980      0.327   -1.51e+04    4.53e+04
neighborhood[T.Veenker] -1.397e+04   1.68e+04     -0.831      0.406   -4.69e+04     1.9e+04
totalsqftcalc              42.6387      1.213     35.165      0.000      40.260      45.017
overallqual              1.966e+04    755.004     26.044      0.000    1.82e+04    2.11e+04
==============================================================================
Omnibus:                      156.512   Durbin-Watson:                   2.028
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              602.743
Skew:                           0.402   Prob(JB):                    1.31e-131
Kurtosis:                       5.859   Cond. No.                     2.24e+05
==============================================================================
```

However, using the individual variables that make up the total square food variable does improve the R squared value, AIC, and RMSE score on Kaggle to 35321.76 and can be used in the final model.

Model 9:

```
                                   OLS Regression Results
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.880
Model:                            OLS   Adj. R-squared:                  0.878
Method:                 Least Squares   F-statistic:                     493.5
Date:                Tue, 17 Oct 2017   Prob (F-statistic):               0.00
Time:                        10:31:19   Log-Likelihood:                -18941.
No. Observations:                1640   AIC:                         3.793e+04
Df Residuals:                    1615   BIC:                         3.807e+04
Df Model:                          24
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Intercept               -1.441e+04   1.54e+04     -0.936      0.349   -4.46e+04    1.58e+04
neighborhood[T.BrkSide] -2.752e+04   1.49e+04     -1.842      0.066   -5.68e+04    1791.405
neighborhood[T.ClearCr] -5099.2104   1.55e+04     -0.328      0.743   -3.56e+04    2.54e+04
neighborhood[T.CollgCr] -1703.9494   1.48e+04     -0.115      0.908   -3.07e+04    2.72e+04
neighborhood[T.Crawfor] -1986.6710    1.5e+04     -0.132      0.895   -3.14e+04    2.74e+04
neighborhood[T.Edwards] -2.327e+04   1.49e+04     -1.564      0.118   -5.24e+04    5912.274
neighborhood[T.Gilbert] -6887.2517   1.48e+04     -0.464      0.642    -3.6e+04    2.22e+04
neighborhood[T.IDOTRR]  -3.424e+04   1.52e+04     -2.248      0.025   -6.41e+04   -4370.888
neighborhood[T.Mitchel] -1.348e+04    1.5e+04     -0.897      0.370    -4.3e+04     1.6e+04
neighborhood[T.NAmes]   -2.284e+04   1.48e+04     -1.547      0.122   -5.18e+04    6119.408
neighborhood[T.NWAmes]  -2.432e+04   1.49e+04     -1.628      0.104   -5.36e+04    4985.733
neighborhood[T.NoRidge]  2.438e+04   1.52e+04      1.601      0.110   -5487.526    5.42e+04
neighborhood[T.NridgHt]  6.511e+04    1.5e+04      4.340      0.000    3.57e+04    9.45e+04
neighborhood[T.OldTown] -3.479e+04   1.48e+04     -2.348      0.019   -6.39e+04   -5731.599
neighborhood[T.SWISU]   -3.686e+04   1.57e+04     -2.343      0.019   -6.77e+04   -6005.946
neighborhood[T.Sawyer]  -1.829e+04    1.5e+04     -1.223      0.221   -4.76e+04     1.1e+04
neighborhood[T.SawyerW] -1.705e+04   1.49e+04     -1.141      0.254   -4.64e+04    1.23e+04
neighborhood[T.Somerst]  2.556e+04   1.49e+04      1.714      0.087   -3689.588    5.48e+04
neighborhood[T.StoneBr]  6.937e+04   1.63e+04      4.250      0.000    3.74e+04    1.01e+05
neighborhood[T.Timber]   1.505e+04   1.51e+04      0.999      0.318   -1.45e+04    4.46e+04
neighborhood[T.Veenker] -1.245e+04   1.64e+04     -0.757      0.449   -4.47e+04    1.98e+04
grlivarea                  54.0147      1.890     28.576      0.000      50.307      57.722
bsmtfinsf1                 36.8381      1.623     22.697      0.000      33.655      40.021
bsmtfinsf2                 21.5340      3.935      5.472      0.000      13.815      29.253
overallqual              1.772e+04    775.007     22.871      0.000    1.62e+04    1.92e+04
==============================================================================
Omnibus:                      159.554   Durbin-Watson:                   2.024
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              827.795
Skew:                           0.294   Prob(JB):                    1.76e-180
Kurtosis:                       6.431   Cond. No.                     1.74e+05
==============================================================================
```

Another variable, quality index, was created by multiplying the overall quality rating and overall condition rating. The new variable was substituted in the original model from the first analysis in the place of the overall quality rating.

This produces a lower R squared, higher AIC value, and higher RMSE (35463.68) on Kaggle and will not be used in the final model.

Model 10:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.869
Model:                            OLS   Adj. R-squared:                  0.867
Method:                 Least Squares   F-statistic:                     466.5
Date:                Tue, 17 Oct 2017   Prob (F-statistic):               0.00
Time:                        10:37:15   Log-Likelihood:                -19012.
No. Observations:                1640   AIC:                         3.807e+04
Df Residuals:                    1616   BIC:                         3.820e+04
Df Model:                          23
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------

Intercept                4.009e+04   1.56e+04      2.572      0.010    9516.878    7.07e+04
neighborhood[T.BrkSide] -5.551e+04   1.55e+04     -3.573      0.000    -8.6e+04    -2.5e+04
neighborhood[T.ClearCr] -2.129e+04   1.62e+04     -1.318      0.188    -5.3e+04    1.04e+04
neighborhood[T.CollgCr] -8075.8193   1.54e+04     -0.524      0.600   -3.83e+04    2.21e+04
neighborhood[T.Crawfor] -2.864e+04   1.56e+04     -1.835      0.067   -5.93e+04    1980.478
neighborhood[T.Edwards] -4.767e+04   1.54e+04     -3.086      0.002    -7.8e+04   -1.74e+04
neighborhood[T.Gilbert] -1.646e+04   1.55e+04     -1.064      0.288   -4.68e+04    1.39e+04
neighborhood[T.IDOTRR]  -6.232e+04   1.58e+04     -3.939      0.000   -9.34e+04   -3.13e+04
neighborhood[T.Mitchel] -3.244e+04   1.56e+04     -2.073      0.038   -6.31e+04   -1748.102
neighborhood[T.NAmes]   -4.599e+04   1.54e+04     -2.995      0.003   -7.61e+04   -1.59e+04
neighborhood[T.NWAmes]  -4.292e+04   1.56e+04     -2.757      0.006   -7.35e+04   -1.24e+04
neighborhood[T.NoRidge]  2.052e+04   1.59e+04      1.292      0.197   -1.06e+04    5.17e+04
neighborhood[T.NridgHt]  7.123e+04   1.57e+04      4.548      0.000    4.05e+04    1.02e+05
neighborhood[T.OldTown] -6.595e+04   1.54e+04     -4.284      0.000   -9.61e+04   -3.58e+04
neighborhood[T.SWISU]   -6.201e+04   1.64e+04     -3.789      0.000   -9.41e+04   -2.99e+04
neighborhood[T.Sawyer]  -3.996e+04   1.55e+04     -2.572      0.010   -7.04e+04   -9482.116
neighborhood[T.SawyerW] -2.579e+04   1.56e+04     -1.654      0.098   -5.64e+04    4793.936
neighborhood[T.Somerst]  2.665e+04   1.56e+04      1.712      0.087   -3890.661    5.72e+04
neighborhood[T.StoneBr]  7.361e+04    1.7e+04      4.320      0.000    4.02e+04    1.07e+05
neighborhood[T.Timber]   1.141e+04   1.57e+04      0.725      0.468   -1.94e+04    4.22e+04
neighborhood[T.Veenker] -2.335e+04   1.71e+04     -1.364      0.173   -5.69e+04    1.02e+04
grlivarea                  63.8016      1.835     34.773      0.000      60.203      67.400
bsmtfinsf1                 37.1141      1.685     22.024      0.000      33.809      40.419
qualityindex             1635.7496     85.279     19.181      0.000    1468.480    1803.019
==============================================================================
Omnibus:                      195.017   Durbin-Watson:                   2.018
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              691.449
Skew:                           0.558   Prob(JB):                    7.14e-151
Kurtosis:                       5.979   Cond. No.                     1.74e+05
==============================================================================
```

However, as with the previous calculated variable (totalsqft), using the individual variables (overall quality rating and overall condition rating) improved the R-squared value, AIC, and RMSE on Kaggle to 35172.54.

Model 11:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.881
Model:                            OLS   Adj. R-squared:                  0.879
Method:                 Least Squares   F-statistic:                     498.8
Date:                Tue, 17 Oct 2017   Prob (F-statistic):               0.00
Time:                        11:27:50   Log-Likelihood:                -18933.
No. Observations:                1640   AIC:                         3.792e+04
Df Residuals:                    1615   BIC:                         3.805e+04
Df Model:                          24
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Intercept               -3.165e+04   1.55e+04     -2.037      0.042   -6.21e+04   -1167.401
neighborhood[T.BrkSide] -3.401e+04   1.49e+04     -2.281      0.023   -6.33e+04   -4770.316
neighborhood[T.ClearCr] -4292.7587   1.55e+04     -0.278      0.781   -3.46e+04     2.6e+04
neighborhood[T.CollgCr] -2011.0599   1.47e+04     -0.137      0.891   -3.08e+04    2.68e+04
neighborhood[T.Crawfor] -7318.3195    1.5e+04     -0.489      0.625   -3.67e+04     2.2e+04
neighborhood[T.Edwards] -2.651e+04   1.48e+04     -1.789      0.074   -5.56e+04    2560.513
neighborhood[T.Gilbert] -7997.3459   1.48e+04     -0.542      0.588    -3.7e+04     2.1e+04
neighborhood[T.IDOTRR]  -4.037e+04   1.52e+04     -2.659      0.008   -7.02e+04   -1.06e+04
neighborhood[T.Mitchel] -1.585e+04    1.5e+04     -1.059      0.290   -4.52e+04    1.35e+04
neighborhood[T.NAmes]   -2.638e+04   1.47e+04     -1.793      0.073   -5.52e+04    2484.907
neighborhood[T.NWAmes]   -2.74e+04   1.49e+04     -1.841      0.066   -5.66e+04    1798.541
neighborhood[T.NoRidge]  2.583e+04   1.51e+04      1.705      0.088   -3882.494    5.55e+04
neighborhood[T.NridgHt]  6.625e+04   1.49e+04      4.437      0.000     3.7e+04    9.55e+04
neighborhood[T.OldTown] -4.211e+04   1.48e+04     -2.846      0.004   -7.11e+04   -1.31e+04
neighborhood[T.SWISU]   -4.155e+04   1.57e+04     -2.651      0.008   -7.23e+04   -1.08e+04
neighborhood[T.Sawyer]  -2.032e+04   1.49e+04     -1.365      0.173   -4.95e+04    8887.812
neighborhood[T.SawyerW]  -1.73e+04   1.49e+04     -1.162      0.245   -4.65e+04    1.19e+04
neighborhood[T.Somerst]    2.6e+04   1.48e+04      1.752      0.080   -3114.634    5.51e+04
neighborhood[T.StoneBr]  7.019e+04   1.62e+04      4.321      0.000    3.83e+04    1.02e+05
neighborhood[T.Timber]   1.506e+04    1.5e+04      1.005      0.315   -1.43e+04    4.45e+04
neighborhood[T.Veenker] -1.107e+04   1.63e+04     -0.677      0.498   -4.31e+04     2.1e+04
grlivarea                  55.1829      1.889     29.213      0.000      51.478      58.888
bsmtfinsf1                 35.8735      1.609     22.295      0.000      32.717      39.030
overallqual               1.69e+04    778.768     21.705      0.000    1.54e+04    1.84e+04
overallcond              4305.3477    638.262      6.745      0.000    3053.439    5557.257
==============================================================================
Omnibus:                      165.798   Durbin-Watson:                   2.025
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              762.235
Skew:                           0.370   Prob(JB):                    3.04e-166
Kurtosis:                       6.257   Cond. No.                     1.74e+05
==============================================================================
```

A model was also run to determine if the individual variables of finished basement type 2 and overall condition together could be layered onto the model from the first analysis to improve it. This does not have a major impact on the R squared value or AIC, but does improve the RMSE (to 34927.22) on Kaggle and can be used in the final model.

Model 12:

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.883
Model:                            OLS   Adj. R-squared:                  0.881
Method:                 Least Squares   F-statistic:                     488.2
Date:                Tue, 17 Oct 2017   Prob (F-statistic):               0.00
Time:                        11:36:03   Log-Likelihood:                -18919.
No. Observations:                1640   AIC:                         3.789e+04
Df Residuals:                    1614   BIC:                         3.803e+04
Df Model:                          25                                         
Covariance Type:            nonrobust                                         
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept                -3.186e+04   1.54e+04     -2.067      0.039   -6.21e+04   -1630.720
neighborhood[T.BrkSide] -3.426e+04   1.48e+04     -2.318      0.021   -6.33e+04   -5269.238
neighborhood[T.ClearCr] -7964.1623   1.53e+04     -0.519      0.604    -3.8e+04    2.21e+04
neighborhood[T.CollgCr] -3064.0704   1.46e+04     -0.210      0.833   -3.16e+04    2.55e+04
neighborhood[T.Crawfor] -9134.2250   1.48e+04     -0.615      0.538   -3.82e+04       2e+04
neighborhood[T.Edwards] -2.745e+04   1.47e+04     -1.868      0.062   -5.63e+04    1376.139
neighborhood[T.Gilbert] -8133.2447   1.46e+04     -0.556      0.579   -3.68e+04    2.06e+04
neighborhood[T.IDOTRR]  -4.031e+04   1.51e+04     -2.678      0.007   -6.98e+04   -1.08e+04
neighborhood[T.Mitchel] -1.714e+04   1.48e+04     -1.154      0.249   -4.63e+04     1.2e+04
neighborhood[T.NAmes]   -2.802e+04   1.46e+04     -1.920      0.055   -5.66e+04     599.551
neighborhood[T.NWAmes]  -2.897e+04   1.48e+04     -1.962      0.050   -5.79e+04     -13.417
neighborhood[T.NoRidge]  2.308e+04    1.5e+04      1.536      0.125   -6395.512    5.26e+04
neighborhood[T.NridgHt]  6.542e+04   1.48e+04      4.418      0.000    3.64e+04    9.45e+04
neighborhood[T.OldTown]   -4.25e+04   1.47e+04     -2.897      0.004   -7.13e+04   -1.37e+04
neighborhood[T.SWISU]   -4.175e+04   1.55e+04     -2.686      0.007   -7.22e+04   -1.13e+04
neighborhood[T.Sawyer]   -2.32e+04   1.48e+04     -1.570      0.117   -5.22e+04    5775.713
neighborhood[T.SawyerW] -1.844e+04   1.48e+04     -1.250      0.211   -4.74e+04    1.05e+04
neighborhood[T.Somerst]  2.543e+04   1.47e+04      1.728      0.084   -3430.897    5.43e+04
neighborhood[T.StoneBr]  6.934e+04   1.61e+04      4.305      0.000    3.78e+04    1.01e+05
neighborhood[T.Timber]   1.386e+04   1.49e+04      0.933      0.351   -1.53e+04     4.3e+04
neighborhood[T.Veenker] -1.651e+04   1.62e+04     -1.017      0.310   -4.84e+04    1.53e+04
grlivarea                  55.1346      1.873     29.437      0.000      51.461      58.808
bsmtfinsf1                 36.6487      1.602     22.878      0.000      33.507      39.791
bsmtfinsf2                 20.8360      3.885      5.363      0.000      13.216      28.456
overallqual              1.701e+04    772.399     22.018      0.000    1.55e+04    1.85e+04
overallcond              4213.6893    633.076      6.656      0.000    2971.952    5455.427
==============================================================================
Omnibus:                      172.139   Durbin-Watson:                   2.029
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              796.391
Skew:                           0.391   Prob(JB):                    1.16e-173
Kurtosis:                       6.323   Cond. No.                     1.74e+05
```

An additional model was built using year built with the model from the first analysis along with the variables that have improved the model (bsmtfinsf2 and overallcond.) While this improves the R squared value and AIC, the RMSE on Kaggle is not improved (35088.89) and year built will not be used in the final model.

Model 13:

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.891
Model:                            OLS   Adj. R-squared:                  0.890
Method:                 Least Squares   F-statistic:                     509.5
Date:                Tue, 17 Oct 2017   Prob (F-statistic):               0.00
Time:                        19:43:30   Log-Likelihood:                -18859.
No. Observations:                1640   AIC:                         3.777e+04
Df Residuals:                    1613   BIC:                         3.792e+04
Df Model:                          26                                         
Covariance Type:            nonrobust                                         
=============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
Intercept               -1.056e+06   9.37e+04    -11.272      0.000   -1.24e+06   -8.72e+05
neighborhood[T.BrkSide]  -467.0942   1.46e+04     -0.032      0.974   -2.91e+04    2.81e+04
neighborhood[T.ClearCr]  1.059e+04   1.49e+04      0.711      0.477   -1.86e+04    3.98e+04
neighborhood[T.CollgCr]  1420.8417   1.41e+04      0.101      0.919   -2.61e+04     2.9e+04
neighborhood[T.Crawfor]  2.053e+04   1.46e+04      1.410      0.159   -8025.496    4.91e+04
neighborhood[T.Edwards] -4082.8619   1.43e+04     -0.285      0.776   -3.22e+04     2.4e+04
neighborhood[T.Gilbert] -5459.3562   1.41e+04     -0.387      0.699   -3.31e+04    2.22e+04
neighborhood[T.IDOTRR]  -4914.8450   1.49e+04     -0.331      0.741   -3.41e+04    2.42e+04
neighborhood[T.Mitchel] -5710.2668   1.44e+04     -0.398      0.691   -3.39e+04    2.24e+04
neighborhood[T.NAmes]   -6916.3795   1.42e+04     -0.487      0.626   -3.48e+04    2.09e+04
neighborhood[T.NWAmes]  -1.483e+04   1.43e+04     -1.037      0.300   -4.29e+04    1.32e+04
neighborhood[T.NoRidge]   2.97e+04   1.45e+04      2.048      0.041    1248.594    5.81e+04
neighborhood[T.NridgHt]  6.903e+04   1.43e+04      4.834      0.000     4.1e+04     9.7e+04
neighborhood[T.OldTown] -5256.2871   1.45e+04     -0.362      0.718   -3.38e+04    2.33e+04
neighborhood[T.SWISU]   -7134.7676   1.53e+04     -0.466      0.641   -3.72e+04    2.29e+04
neighborhood[T.Sawyer]  -4913.4786   1.43e+04     -0.343      0.732    -3.3e+04    2.32e+04
neighborhood[T.SawyerW] -1.232e+04   1.42e+04     -0.865      0.387   -4.02e+04    1.56e+04
neighborhood[T.Somerst]  2.736e+04   1.42e+04      1.928      0.054    -479.615    5.52e+04
neighborhood[T.StoneBr]   7.48e+04   1.55e+04      4.813      0.000    4.43e+04    1.05e+05
neighborhood[T.Timber]    1.92e+04   1.43e+04      1.338      0.181   -8935.653    4.73e+04
neighborhood[T.Veenker] -2535.8178   1.57e+04     -0.161      0.872   -3.34e+04    2.83e+04
grlivarea                  57.1842      1.816     31.496      0.000      53.623      60.745
bsmtfinsf1                 34.6277      1.556     22.261      0.000      31.577      37.679
bsmtfinsf2                 19.7830      3.748      5.279      0.000      12.432      27.134
overallqual              1.497e+04    767.133     19.520      0.000    1.35e+04    1.65e+04
overallcond              5675.7275    624.613      9.087      0.000    4450.589    6900.866
yearbuilt                 512.4571     46.283     11.072      0.000     421.677     603.237
==============================================================================
Omnibus:                      213.595   Durbin-Watson:                   2.057
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              989.592
Skew:                           0.531   Prob(JB):                    1.30e-215
Kurtosis:                       6.654   Cond. No.                     4.07e+05
==============================================================================
```

A model was also built using year remodel to determine if this would improve the model. While there is no discernable change in the R squared value or AIC, the RMSE score on Kaggle was not improved (35013.98) and this variable will not be used in the final model.

Model 14:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.886
Model:                            OLS   Adj. R-squared:                  0.885
Method:                 Least Squares   F-statistic:                     484.1
Date:                Tue, 17 Oct 2017   Prob (F-statistic):               0.00
Time:                        19:51:28   Log-Likelihood:                -18896.
No. Observations:                1640   AIC:                         3.785e+04
Df Residuals:                    1613   BIC:                         3.799e+04
Df Model:                          26
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Intercept               -5.676e+05   8.09e+04     -7.014      0.000   -7.26e+05   -4.09e+05
neighborhood[T.BrkSide] -2.272e+04   1.47e+04     -1.547      0.122   -5.15e+04    6079.938
neighborhood[T.ClearCr]   226.1574   1.52e+04      0.015      0.988   -2.95e+04       3e+04
neighborhood[T.CollgCr]  -407.8623   1.44e+04     -0.028      0.977   -2.86e+04    2.78e+04
neighborhood[T.Crawfor]  2414.1440   1.47e+04      0.164      0.870   -2.65e+04    3.13e+04
neighborhood[T.Edwards] -1.831e+04   1.46e+04     -1.258      0.209   -4.69e+04    1.02e+04
neighborhood[T.Gilbert] -5860.8250   1.44e+04     -0.406      0.685   -3.42e+04    2.25e+04
neighborhood[T.IDOTRR]  -2.941e+04   1.49e+04     -1.969      0.049   -5.87e+04    -109.759
neighborhood[T.Mitchel] -1.099e+04   1.47e+04     -0.749      0.454   -3.98e+04    1.78e+04
neighborhood[T.NAmes]    -1.78e+04   1.45e+04     -1.230      0.219   -4.62e+04    1.06e+04
neighborhood[T.NWAmes]  -1.984e+04   1.46e+04     -1.357      0.175   -4.85e+04    8846.780
neighborhood[T.NoRidge]  2.791e+04   1.48e+04      1.880      0.060   -1207.619     5.7e+04
neighborhood[T.NridgHt]   6.77e+04   1.46e+04      4.634      0.000     3.9e+04    9.64e+04
neighborhood[T.OldTown] -3.234e+04   1.45e+04     -2.223      0.026   -6.09e+04   -3802.698
neighborhood[T.SWISU]   -2.901e+04   1.54e+04     -1.878      0.061   -5.93e+04    1294.032
neighborhood[T.Sawyer]  -1.487e+04   1.46e+04     -1.017      0.309   -4.36e+04    1.38e+04
neighborhood[T.SawyerW] -1.513e+04   1.46e+04     -1.039      0.299   -4.37e+04    1.34e+04
neighborhood[T.Somerst]  2.677e+04   1.45e+04      1.844      0.065   -1703.300    5.53e+04
neighborhood[T.StoneBr]  7.257e+04   1.59e+04      4.565      0.000    4.14e+04    1.04e+05
neighborhood[T.Timber]   1.705e+04   1.47e+04      1.162      0.245   -1.17e+04    4.58e+04
neighborhood[T.Veenker] -8511.6945   1.61e+04     -0.530      0.596      -4e+04     2.3e+04
grlivarea                  54.2002      1.853     29.252      0.000      50.566      57.835
bsmtfinsf1                 36.6965      1.580     23.221      0.000      33.597      39.796
bsmtfinsf2                 21.3487      3.833      5.569      0.000      13.830      28.867
overallqual              1.618e+04    771.852     20.959      0.000    1.47e+04    1.77e+04
overallcond              2437.7959    677.834      3.596      0.000    1108.268    3767.323
yearremodel               274.8057     40.769      6.741      0.000     194.841     354.771
==============================================================================
Omnibus:                      180.928   Durbin-Watson:                   2.029
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              878.206
Skew:                           0.405   Prob(JB):                    2.00e-191
Kurtosis:                       6.492   Cond. No.                     3.51e+05
==============================================================================
```

A variable was created using the year built and year remodeled variables to generate the number of years since the home has been remodeled. (Years since remodel will be 0 for homes that have not been remodeled.) The model was run using the variables for the best scoring model thus far plus the years since remodel variable. This improved the RMSE on Kaggle to 34927.23, and can be used in the final model.

Model 15:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.884
Model:                            OLS   Adj. R-squared:                  0.882
Method:                 Least Squares   F-statistic:                     470.9
Date:                Tue, 17 Oct 2017   Prob (F-statistic):               0.00
Time:                        20:05:25   Log-Likelihood:                -18916.
No. Observations:                1640   AIC:                         3.789e+04
Df Residuals:                    1613   BIC:                         3.803e+04
Df Model:                          26
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Intercept               -3.598e+04   1.55e+04     -2.322      0.020   -6.64e+04   -5583.564
neighborhood[T.BrkSide] -3.225e+04   1.48e+04     -2.180      0.029   -6.13e+04   -3238.639
neighborhood[T.ClearCr] -7425.1402   1.53e+04     -0.485      0.628   -3.75e+04    2.26e+04
neighborhood[T.CollgCr] -3141.0521   1.45e+04     -0.216      0.829   -3.17e+04    2.54e+04
neighborhood[T.Crawfor] -7797.7031   1.48e+04     -0.526      0.599   -3.69e+04    2.13e+04
neighborhood[T.Edwards]  -2.64e+04   1.47e+04     -1.798      0.072   -5.52e+04    2394.057
neighborhood[T.Gilbert] -8390.2532   1.46e+04     -0.574      0.566   -3.71e+04    2.03e+04
neighborhood[T.IDOTRR]  -3.783e+04   1.51e+04     -2.510      0.012   -6.74e+04   -8266.280
neighborhood[T.Mitchel] -1.714e+04   1.48e+04     -1.156      0.248   -4.62e+04    1.19e+04
neighborhood[T.NAmes]   -2.769e+04   1.46e+04     -1.900      0.058   -5.63e+04     899.298
neighborhood[T.NWAmes]  -2.944e+04   1.47e+04     -1.997      0.046   -5.84e+04    -521.245
neighborhood[T.NoRidge]  2.269e+04    1.5e+04      1.512      0.131   -6750.218    5.21e+04
neighborhood[T.NridgHt]  6.531e+04   1.48e+04      4.417      0.000    3.63e+04    9.43e+04
neighborhood[T.OldTown] -3.949e+04   1.47e+04     -2.685      0.007   -6.83e+04   -1.06e+04
neighborhood[T.SWISU]   -3.997e+04   1.55e+04     -2.572      0.010   -7.05e+04   -9484.270
neighborhood[T.Sawyer]  -2.275e+04   1.48e+04     -1.542      0.123   -5.17e+04    6192.859
neighborhood[T.SawyerW] -1.845e+04   1.47e+04     -1.252      0.211   -4.74e+04    1.04e+04
neighborhood[T.Somerst]  2.534e+04   1.47e+04      1.724      0.085   -3488.343    5.42e+04
neighborhood[T.StoneBr]  6.925e+04   1.61e+04      4.305      0.000    3.77e+04    1.01e+05
neighborhood[T.Timber]   1.376e+04   1.48e+04      0.927      0.354   -1.54e+04    4.29e+04
neighborhood[T.Veenker] -1.666e+04   1.62e+04     -1.027      0.304   -4.85e+04    1.52e+04
grlivarea                  55.7579      1.890     29.495      0.000      52.050      59.466
bsmtfinsf1                 36.3019      1.607     22.589      0.000      33.150      39.454
bsmtfinsf2                 20.5057      3.883      5.281      0.000      12.890      28.121
overallqual              1.693e+04    772.190     21.920      0.000    1.54e+04    1.84e+04
overallcond              4998.2864    719.829      6.944      0.000    3586.387    6410.186
remodelage                -84.2262     36.940     -2.280      0.023    -156.681     -11.771
==============================================================================
Omnibus:                      174.553   Durbin-Watson:                   2.034
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              791.461
Skew:                           0.407   Prob(JB):                    1.37e-172
Kurtosis:                       6.304   Cond. No.                     1.74e+05
==============================================================================
```

Another variable was created using the external quality and external condition variables, in a simple concatenation of the variables, to determine the impact on the model. (Please note: Any home with poor exterior condition in the test file was replaced with fair exterior condition, as no homes with poor exterior condition exist in the cleansed training file. This resulted in 2 replacements.) This increased the R-squared value, decreased the AIC and improved the RMSE value on Kaggle to 33621.98 and can be used in the final model.

Model 16:

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.898
Model:                            OLS   Adj. R-squared:                  0.896
Method:                 Least Squares   F-statistic:                     361.2
Date:                Tue, 17 Oct 2017   Prob (F-statistic):               0.00
Time:                        20:57:43   Log-Likelihood:                -18807.
No. Observations:                1640   AIC:                         3.769e+04
Df Residuals:                    1600   BIC:                         3.791e+04
Df Model:                          39                                         
Covariance Type:            nonrobust                                         
=============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
Intercept                  4.562e+04   2.91e+04      1.567      0.117   -1.15e+04    1.03e+05
neighborhood[T.BrkSide]    -2.22e+04    1.4e+04     -1.587      0.113   -4.96e+04    5233.147
neighborhood[T.ClearCr]    1779.2209   1.45e+04      0.123      0.902   -2.66e+04    3.01e+04
neighborhood[T.CollgCr]     804.6144   1.37e+04      0.059      0.953    -2.6e+04    2.76e+04
neighborhood[T.Crawfor]    4308.5569    1.4e+04      0.307      0.759   -2.32e+04    3.19e+04
neighborhood[T.Edwards]   -1.843e+04   1.39e+04     -1.328      0.184   -4.56e+04    8783.397
neighborhood[T.Gilbert]   -2514.5188   1.38e+04     -0.183      0.855   -2.95e+04    2.45e+04
neighborhood[T.IDOTRR]    -2.895e+04   1.43e+04     -2.029      0.043   -5.69e+04    -965.177
neighborhood[T.Mitchel]   -8033.6056    1.4e+04     -0.574      0.566   -3.55e+04    1.94e+04
neighborhood[T.NAmes]     -1.805e+04   1.38e+04     -1.310      0.191   -4.51e+04    8981.929
neighborhood[T.NWAmes]    -1.584e+04    1.4e+04     -1.135      0.257   -4.32e+04    1.15e+04
neighborhood[T.NoRidge]    2.786e+04   1.41e+04      1.974      0.049     179.527    5.55e+04
neighborhood[T.NridgHt]    5.493e+04    1.4e+04      3.937      0.000    2.76e+04    8.23e+04
neighborhood[T.OldTown]   -3.085e+04   1.39e+04     -2.219      0.027   -5.81e+04   -3582.162
neighborhood[T.SWISU]     -2.961e+04   1.47e+04     -2.014      0.044   -5.84e+04    -771.201
neighborhood[T.Sawyer]    -1.346e+04   1.39e+04     -0.965      0.335   -4.08e+04    1.39e+04
neighborhood[T.SawyerW]   -1.377e+04   1.39e+04     -0.993      0.321    -4.1e+04    1.34e+04
neighborhood[T.Somerst]    2.572e+04   1.38e+04      1.862      0.063   -1375.971    5.28e+04
neighborhood[T.StoneBr]     6.94e+04   1.52e+04      4.570      0.000    3.96e+04    9.92e+04
neighborhood[T.Timber]     1.525e+04    1.4e+04      1.092      0.275   -1.21e+04    4.27e+04
neighborhood[T.Veenker]   -1.244e+04   1.56e+04     -0.799      0.425    -4.3e+04    1.81e+04
qualcond[T.ExGd]          -5.631e+04   2.82e+04     -2.000      0.046   -1.12e+05   -1080.412
qualcond[T.ExTA]          -3331.5023    2.5e+04     -0.133      0.894   -5.24e+04    4.57e+04
qualcond[T.FaFa]          -7.662e+04    2.7e+04     -2.837      0.005    -1.3e+05   -2.36e+04
qualcond[T.FaGd]          -6.976e+04   3.43e+04     -2.036      0.042   -1.37e+05   -2562.378
qualcond[T.FaTA]          -6.642e+04    2.7e+04     -2.463      0.014   -1.19e+05   -1.35e+04
qualcond[T.GdEx]          -3.682e+04      3e+04     -1.228      0.220   -9.56e+04     2.2e+04
qualcond[T.GdFa]           -7.55e+04   3.33e+04     -2.264      0.024   -1.41e+05   -1.01e+04
qualcond[T.GdGd]          -6.027e+04   2.48e+04     -2.434      0.015   -1.09e+05   -1.17e+04
qualcond[T.GdTA]          -5.302e+04   2.47e+04     -2.143      0.032   -1.02e+05   -4501.708
qualcond[T.TAEx]          -6.361e+04   2.73e+04     -2.329      0.020   -1.17e+05       -1e+04
qualcond[T.TAFa]          -7.623e+04   2.54e+04     -3.002      0.003   -1.26e+05   -2.64e+04
qualcond[T.TAGd]          -7.149e+04   2.48e+04     -2.882      0.004    -1.2e+05   -2.28e+04
qualcond[T.TATA]           -6.97e+04   2.48e+04     -2.815      0.005   -1.18e+05   -2.11e+04
grlivarea                    55.3506      1.805     30.669      0.000      51.811      58.891
bsmtfinsf1                   34.5088      1.527     22.605      0.000      31.514      37.503
bsmtfinsf2                   20.7138      3.659      5.661      0.000      13.537      27.891
overallqual                1.281e+04    791.019     16.190      0.000    1.13e+04    1.44e+04
overallcond                5156.5636    721.863      7.143      0.000    3740.667    6572.460
remodelage                  -99.0440     35.118     -2.820      0.005    -167.927     -30.161
==============================================================================
Omnibus:                      141.194   Durbin-Watson:                   2.042
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              711.804
Skew:                           0.220   Prob(JB):                    2.71e-155
Kurtosis:                       6.197   Cond. No.                     2.60e+05
==============================================================================
```

The final variable added to the model is kitchen quality. This was done using the iterative method developed in the first analysis to layer additional variables in the model and check the R squared and AIC values and RMSE score on Kaggle. This model improved the RMSE to 33340.86.

Model 17:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.902
Model:                            OLS   Adj. R-squared:                  0.900
Method:                 Least Squares   F-statistic:                     343.4
Date:                Tue, 17 Oct 2017   Prob (F-statistic):               0.00
Time:                        21:16:08   Log-Likelihood:                -18771.
No. Observations:                1640   AIC:                         3.763e+04
Df Residuals:                    1596   BIC:                         3.787e+04
Df Model:                          43
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Intercept                5.704e+04   2.86e+04      1.997      0.046    1019.487    1.13e+05
neighborhood[T.BrkSide] -2.032e+04   1.37e+04     -1.481      0.139   -4.72e+04    6592.437
neighborhood[T.ClearCr]  5770.3560   1.42e+04      0.406      0.685   -2.21e+04    3.36e+04
neighborhood[T.CollgCr]  1499.6751   1.34e+04      0.112      0.911   -2.48e+04    2.78e+04
neighborhood[T.Crawfor]  6031.9044   1.38e+04      0.438      0.661    -2.1e+04     3.3e+04
neighborhood[T.Edwards] -1.723e+04   1.36e+04     -1.267      0.205   -4.39e+04    9450.830
neighborhood[T.Gilbert]  -982.9886   1.35e+04     -0.073      0.942   -2.74e+04    2.55e+04
neighborhood[T.IDOTRR]  -2.671e+04    1.4e+04     -1.909      0.056   -5.42e+04     738.787
neighborhood[T.Mitchel] -5928.9617   1.37e+04     -0.432      0.666   -3.28e+04     2.1e+04
neighborhood[T.NAmes]   -1.608e+04   1.35e+04     -1.190      0.234   -4.26e+04    1.04e+04
neighborhood[T.NWAmes]  -1.367e+04   1.37e+04     -0.999      0.318   -4.05e+04    1.32e+04
neighborhood[T.NoRidge]  3.065e+04   1.38e+04      2.217      0.027    3529.757    5.78e+04
neighborhood[T.NridgHt]  4.977e+04   1.37e+04      3.634      0.000    2.29e+04    7.66e+04
neighborhood[T.OldTown]  -2.95e+04   1.36e+04     -2.164      0.031   -5.62e+04   -2763.512
neighborhood[T.SWISU]   -2.708e+04   1.44e+04     -1.878      0.061   -5.54e+04    1203.833
neighborhood[T.Sawyer]  -1.128e+04   1.37e+04     -0.825      0.409   -3.81e+04    1.55e+04
neighborhood[T.SawyerW] -1.314e+04   1.36e+04     -0.968      0.333   -3.98e+04    1.35e+04
neighborhood[T.Somerst]  2.484e+04   1.35e+04      1.836      0.067   -1698.584    5.14e+04
neighborhood[T.StoneBr]  6.723e+04   1.49e+04      4.516      0.000     3.8e+04    9.64e+04
neighborhood[T.Timber]   1.521e+04   1.37e+04      1.112      0.266   -1.16e+04    4.21e+04
neighborhood[T.Veenker] -9060.5843   1.53e+04     -0.594      0.553    -3.9e+04    2.09e+04
qualcond[T.ExGd]         -3.27e+04   2.78e+04     -1.178      0.239   -8.71e+04    2.17e+04
qualcond[T.ExTA]         4843.0328   2.45e+04      0.197      0.843   -4.33e+04     5.3e+04
qualcond[T.FaFa]        -6.192e+04   2.65e+04     -2.333      0.020   -1.14e+05   -9855.901
qualcond[T.FaGd]        -4.382e+04   3.37e+04     -1.300      0.194    -1.1e+05    2.23e+04
qualcond[T.FaTA]        -4.629e+04   2.66e+04     -1.743      0.082   -9.84e+04    5816.796
qualcond[T.GdEx]        -2.153e+04   2.94e+04     -0.732      0.465   -7.92e+04    3.62e+04
qualcond[T.GdFa]        -5.624e+04   3.28e+04     -1.714      0.087   -1.21e+05    8112.686
qualcond[T.GdGd]        -3.958e+04   2.44e+04     -1.621      0.105   -8.75e+04    8326.490
qualcond[T.GdTA]        -3.351e+04   2.44e+04     -1.373      0.170   -8.14e+04    1.44e+04
qualcond[T.TAEx]        -4.614e+04   2.68e+04     -1.719      0.086   -9.88e+04    6500.472
qualcond[T.TAFa]        -5.253e+04    2.5e+04     -2.097      0.036   -1.02e+05   -3396.086
qualcond[T.TAGd]        -4.849e+04   2.45e+04     -1.981      0.048   -9.65e+04    -485.298
qualcond[T.TATA]        -4.671e+04   2.44e+04     -1.912      0.056   -9.46e+04    1203.176
kitchenqual[T.Fa]       -2.295e+04   5350.957     -4.290      0.000   -3.34e+04   -1.25e+04
kitchenqual[T.Gd]       -2.148e+04   3176.690     -6.762      0.000   -2.77e+04   -1.53e+04
kitchenqual[T.Po]       -4.485e+04   2.38e+04     -1.883      0.060   -9.16e+04    1859.622
kitchenqual[T.TA]        -2.88e+04   3451.953     -8.343      0.000   -3.56e+04    -2.2e+04
grlivarea                  54.1586      1.776     30.500      0.000      50.676      57.642
bsmtfinsf1                 32.9770      1.507     21.884      0.000      30.021      35.933
bsmtfinsf2                 20.2707      3.585      5.655      0.000      13.240      27.302
overallqual              1.189e+04    785.423     15.136      0.000    1.03e+04    1.34e+04
overallcond              4851.8248    713.260      6.802      0.000    3452.800    6250.850
remodelage               -122.4095     34.676     -3.530      0.000    -190.425     -54.395
==============================================================================
Omnibus:                      152.446   Durbin-Watson:                   2.046
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              792.868
Skew:                           0.260   Prob(JB):                    6.77e-173
Kurtosis:                       6.366   Cond. No.                     2.61e+05
==============================================================================
```

Log Transformation:

An example model was fitted with several variables and then fitted with the same variables and a log transformation of the sales price to determine if a log transformation could improve the model

Model 18 (non-log model):

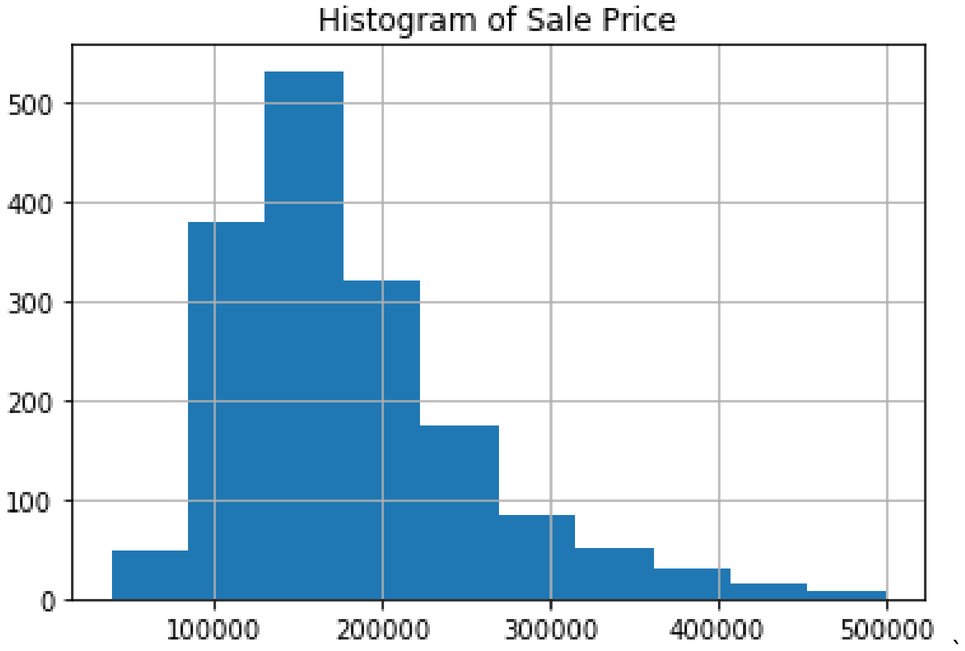
```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.909
Model:                            OLS   Adj. R-squared:                  0.906
Method:                 Least Squares   F-statistic:                     360.9
Date:                Wed, 18 Oct 2017   Prob (F-statistic):               0.00
Time:                        19:52:35   Log-Likelihood:                -18716.
No. Observations:                1640   AIC:                         3.752e+04
Df Residuals:                    1595   BIC:                         3.777e+04
Df Model:                          44                                         
Covariance Type:            nonrobust                                         
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Intercept                3.437e+04   2.77e+04      1.240      0.215      -2e+04    8.87e+04
neighborhood[T.BrkSide] -1.473e+04   1.33e+04     -1.109      0.268   -4.08e+04    1.13e+04
neighborhood[T.ClearCr]  9698.8261   1.37e+04      0.706      0.481   -1.73e+04    3.67e+04
neighborhood[T.CollgCr]  5717.3546    1.3e+04      0.441      0.659   -1.97e+04    3.12e+04
neighborhood[T.Crawfor]  1.082e+04   1.33e+04      0.811      0.417   -1.53e+04     3.7e+04
neighborhood[T.Edwards] -1.126e+04   1.32e+04     -0.855      0.393   -3.71e+04    1.46e+04
neighborhood[T.Gilbert]  6999.3906   1.31e+04      0.536      0.592   -1.86e+04    3.26e+04
neighborhood[T.IDOTRR]  -2.241e+04   1.35e+04     -1.654      0.098    -4.9e+04    4159.335
neighborhood[T.Mitchel] -1242.0115   1.33e+04     -0.093      0.926   -2.73e+04    2.48e+04
neighborhood[T.NAmes]    -1.27e+04   1.31e+04     -0.972      0.331   -3.83e+04    1.29e+04
neighborhood[T.NWAmes]  -1.021e+04   1.32e+04     -0.771      0.441   -3.62e+04    1.58e+04
neighborhood[T.NoRidge]  3.529e+04   1.34e+04      2.636      0.008    9031.468    6.15e+04
neighborhood[T.NridgHt]  4.993e+04   1.33e+04      3.767      0.000    2.39e+04    7.59e+04
neighborhood[T.OldTown] -2.473e+04   1.32e+04     -1.873      0.061   -5.06e+04    1163.737
neighborhood[T.SWISU]   -2.153e+04    1.4e+04     -1.542      0.123   -4.89e+04    5853.257
neighborhood[T.Sawyer]  -6362.8612   1.32e+04     -0.481      0.631   -3.23e+04    1.96e+04
neighborhood[T.SawyerW] -6701.4630   1.32e+04     -0.510      0.610   -3.25e+04    1.91e+04
neighborhood[T.Somerst]  2.707e+04   1.31e+04      2.067      0.039    1378.809    5.28e+04
neighborhood[T.StoneBr]  6.768e+04   1.44e+04      4.699      0.000    3.94e+04    9.59e+04
neighborhood[T.Timber]   1.772e+04   1.32e+04      1.338      0.181   -8251.593    4.37e+04
neighborhood[T.Veenker] -6943.9373   1.48e+04     -0.470      0.638   -3.59e+04     2.2e+04
qualcond[T.ExGd]        -2.931e+04   2.69e+04     -1.091      0.275    -8.2e+04    2.34e+04
qualcond[T.ExTA]         5697.5137   2.37e+04      0.240      0.810   -4.09e+04    5.23e+04
qualcond[T.FaFa]         -5.94e+04   2.57e+04     -2.312      0.021    -1.1e+05   -9016.163
qualcond[T.FaGd]        -3.365e+04   3.26e+04     -1.032      0.302   -9.76e+04    3.03e+04
qualcond[T.FaTA]        -3.499e+04   2.57e+04     -1.360      0.174   -8.55e+04    1.55e+04
qualcond[T.GdEx]        -1.058e+04   2.85e+04     -0.371      0.710   -6.65e+04    4.53e+04
qualcond[T.GdFa]        -5.334e+04   3.17e+04     -1.680      0.093   -1.16e+05    8935.676
qualcond[T.GdGd]        -3.505e+04   2.36e+04     -1.483      0.138   -8.14e+04    1.13e+04
qualcond[T.GdTA]        -3.057e+04   2.36e+04     -1.294      0.196   -7.69e+04    1.58e+04
qualcond[T.TAEx]        -4.157e+04    2.6e+04     -1.600      0.110   -9.25e+04    9380.988
qualcond[T.TAFa]        -4.957e+04   2.42e+04     -2.045      0.041   -9.71e+04   -2027.183
qualcond[T.TAGd]        -4.292e+04   2.37e+04     -1.812      0.070   -8.94e+04    3547.106
qualcond[T.TATA]          -4.2e+04   2.36e+04     -1.777      0.076   -8.84e+04    4364.806
kitchenqual[T.Fa]       -2.345e+04   5178.134     -4.529      0.000   -3.36e+04   -1.33e+04
kitchenqual[T.Gd]       -2.063e+04   3075.044     -6.708      0.000   -2.67e+04   -1.46e+04
kitchenqual[T.Po]       -3.977e+04   2.31e+04     -1.725      0.085    -8.5e+04    5445.327
kitchenqual[T.TA]       -2.752e+04   3342.550     -8.235      0.000   -3.41e+04    -2.1e+04
grlivarea                  52.9393      1.722     30.739      0.000      49.561      56.317
bsmtfinsf1                 47.3941      2.006     23.623      0.000      43.459      51.329
bsmtfinsf2                 34.4042      3.722      9.242      0.000      27.103      41.706
overallqual               1.05e+04    771.451     13.616      0.000    8991.273     1.2e+04
overallcond              5615.1901    694.040      8.091      0.000    4253.863    6976.517
remodelage               -113.0227     33.566     -3.367      0.001    -178.862     -47.184
bsmtunfsf                  21.4307      2.048     10.462      0.000      17.413      25.449
==============================================================================
Omnibus:                      147.559   Durbin-Watson:                   2.034
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              841.686
Skew:                           0.179   Prob(JB):                    1.70e-183
Kurtosis:                       6.491   Cond. No.                     2.76e+05
==============================================================================
```

Model 19 (log transformed model):

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      np.log(saleprice)   R-squared:                       0.898
Model:                            OLS   Adj. R-squared:                  0.896
Method:                 Least Squares   F-statistic:                     320.4
Date:                Wed, 18 Oct 2017   Prob (F-statistic):               0.00
Time:                        21:34:34   Log-Likelihood:                 1148.7
No. Observations:                1640   AIC:                            -2207.
Df Residuals:                    1595   BIC:                            -1964.
Df Model:                          44
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Intercept                  10.9299      0.152     71.836      0.000      10.631      11.228
neighborhood[T.BrkSide]    -0.1691      0.073     -2.319      0.021      -0.312      -0.026
neighborhood[T.ClearCr]     0.0283      0.075      0.375      0.708      -0.120       0.176
neighborhood[T.CollgCr]     0.0087      0.071      0.122      0.903      -0.131       0.148
neighborhood[T.Crawfor]     0.0176      0.073      0.240      0.810      -0.126       0.161
neighborhood[T.Edwards]    -0.1437      0.072     -1.988      0.047      -0.286      -0.002
neighborhood[T.Gilbert]     0.0326      0.072      0.455      0.649      -0.108       0.173
neighborhood[T.IDOTRR]     -0.2489      0.074     -3.348      0.001      -0.395      -0.103
neighborhood[T.Mitchel]    -0.0242      0.073     -0.332      0.740      -0.167       0.119
neighborhood[T.NAmes]      -0.1016      0.072     -1.415      0.157      -0.242       0.039
neighborhood[T.NWAmes]     -0.0789      0.073     -1.085      0.278      -0.221       0.064
neighborhood[T.NoRidge]     0.0335      0.073      0.456      0.649      -0.111       0.178
neighborhood[T.NridgHt]     0.1110      0.073      1.526      0.127      -0.032       0.254
neighborhood[T.OldTown]    -0.2440      0.072     -3.368      0.001      -0.386      -0.102
neighborhood[T.SWISU]      -0.1582      0.077     -2.064      0.039      -0.308      -0.008
neighborhood[T.Sawyer]     -0.0681      0.073     -0.938      0.349      -0.211       0.074
neighborhood[T.SawyerW]    -0.0511      0.072     -0.708      0.479      -0.193       0.090
neighborhood[T.Somerst]     0.0918      0.072      1.278      0.202      -0.049       0.233
neighborhood[T.StoneBr]     0.1474      0.079      1.865      0.062      -0.008       0.302
neighborhood[T.Timber]      0.0392      0.073      0.540      0.589      -0.103       0.182
neighborhood[T.Veenker]    -0.0922      0.081     -1.137      0.256      -0.251       0.067
qualcond[T.ExGd]           -0.1649      0.147     -1.119      0.263      -0.454       0.124
qualcond[T.ExTA]           -0.0179      0.130     -0.137      0.891      -0.273       0.238
qualcond[T.FaFa]           -0.2694      0.141     -1.911      0.056      -0.546       0.007
qualcond[T.FaGd]           -0.0480      0.179     -0.268      0.789      -0.399       0.303
qualcond[T.FaTA]           -0.1507      0.141     -1.068      0.286      -0.428       0.126
qualcond[T.GdEx]           -0.0875      0.156     -0.560      0.576      -0.394       0.219
qualcond[T.GdFa]           -0.1777      0.174     -1.020      0.308      -0.519       0.164
qualcond[T.GdGd]           -0.0881      0.130     -0.679      0.497      -0.343       0.166
qualcond[T.GdTA]           -0.0687      0.130     -0.530      0.596      -0.323       0.186
qualcond[T.TAEx]           -0.0863      0.143     -0.605      0.545      -0.366       0.193
qualcond[T.TAFa]           -0.1811      0.133     -1.361      0.174      -0.442       0.080
qualcond[T.TAGd]           -0.1190      0.130     -0.915      0.360      -0.374       0.136
qualcond[T.TATA]           -0.1059      0.130     -0.817      0.414      -0.360       0.149
kitchenqual[T.Fa]          -0.0959      0.028     -3.375      0.001      -0.152      -0.040
kitchenqual[T.Gd]          -0.0361      0.017     -2.139      0.033      -0.069      -0.003
kitchenqual[T.Po]          -0.1907      0.127     -1.508      0.132      -0.439       0.057
kitchenqual[T.TA]          -0.0806      0.018     -4.392      0.000      -0.117      -0.045
grlivarea                   0.0003   9.45e-06     31.997      0.000       0.000       0.000
bsmtfinsf1                  0.0002    1.1e-05     21.268      0.000       0.000       0.000
bsmtfinsf2                  0.0002   2.04e-05      9.652      0.000       0.000       0.000
overallqual                 0.0696      0.004     16.440      0.000       0.061       0.078
overallcond                 0.0477      0.004     12.527      0.000       0.040       0.055
remodelage                 -0.0007      0.000     -3.950      0.000      -0.001      -0.000
bsmtunfsf                   0.0001   1.12e-05     11.151      0.000       0.000       0.000
==============================================================================
Omnibus:                      284.584   Durbin-Watson:                   2.028
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1364.113
Skew:                          -0.736   Prob(JB):                    6.12e-297
Kurtosis:                       7.218   Cond. No.                     2.76e+05
==============================================================================
```

Although the log transformed model had an R squared value just slightly less than the non-transformed model, the RMSE on Kaggle is significantly worse for the log transformed model (88925.35 vs 34726.32). The transformed and untransformed models are interpreted differently because the resulting sale price predictions need to be reverse transformed before they can be interpreted. While this did not improve the model, in general, log transformation can be used to make highly skewed distributions less skewed. The sale price, in this case, does not appear to be highly skewed as shown in the histogram of the sales price data in the train file in figure 6:

Figure 6 – histogram of sale price:



There are no additional variables that are highly skewed, per the exploratory data analysis done in the first analysis, that would lend themselves to a log transformation or any other transformation.

Computation of VIF values for the model:

In order to determine if the model has highly correlated pairs, the VIF values were calculated for the model that produces the best RMSE score, highest R squared value and lowest AIC. This model produced some VIF scores that are very large (over 50), so the calculated exterior quality and exterior condition variable were removed from the model and the VIF was recalculated.

VIF Table:

VIF	Variable
2543.8	Intercept
26.6	neighborhood[T.BrkSide]
9.8	neighborhood[T.ClearCr]
53.6	neighborhood[T.CollgCr]
22.8	neighborhood[T.Crawfor]
39.7	neighborhood[T.Edwards]
35.8	neighborhood[T.Gilbert]
14.2	neighborhood[T.IDOTRR]
21.4	neighborhood[T.Mitchel]
79.2	neighborhood[T.NAmes]
27.1	neighborhood[T.NWAmes]
16.3	neighborhood[T.NoRidge]
25.2	neighborhood[T.NridgHt]
48.7	neighborhood[T.OldTown]
7.8	neighborhood[T.SWISU]
28.6	neighborhood[T.Sawyer]
23.5	neighborhood[T.SawyerW]
25.9	neighborhood[T.Somerst]
5.4	neighborhood[T.StoneBr]
18.3	neighborhood[T.Timber]
5.3	neighborhood[T.Veenker]
4.4	qualcond[T.ExGd]
52.2	qualcond[T.ExTA]
6.7	qualcond[T.FaFa]
2.2	qualcond[T.FaGd]
6.7	qualcond[T.FaTA]
3.3	qualcond[T.GdEx]
2.0	qualcond[T.GdFa]
58.2	qualcond[T.GdGd]
388.7	qualcond[T.GdTA]
5.5	qualcond[T.TAEx]
23.6	qualcond[T.TAFa]
136.3	qualcond[T.TAGd]
462.7	qualcond[T.TATA]
1.8	kitchenqual[T.Fa]
7.6	kitchenqual[T.Gd]
1.1	kitchenqual[T.Po]
9.3	kitchenqual[T.TA]
2.1	grlivarea
1.3	bsmtfinsf1
1.1	bsmtfinsf2
3.5	overallqual
1.9	overallcond
2.4	remodelage

Recalculated VIF Table:

VIF	Variable
688.3	Intercept
26.4	neighborhood[T.BrkSide]
9.8	neighborhood[T.ClearCr]
53.6	neighborhood[T.CollgCr]
22.6	neighborhood[T.Crawfor]
39.4	neighborhood[T.Edwards]
35.7	neighborhood[T.Gilbert]
14.0	neighborhood[T.IDOTRR]
21.2	neighborhood[T.Mitchel]
78.6	neighborhood[T.NAmes]
26.8	neighborhood[T.NWAmes]
16.2	neighborhood[T.NoRidge]
25.1	neighborhood[T.NridgHt]
48.4	neighborhood[T.OldTown]
7.7	neighborhood[T.SWISU]
28.4	neighborhood[T.Sawyer]
23.4	neighborhood[T.SawyerW]
25.9	neighborhood[T.Somerst]
5.4	neighborhood[T.StoneBr]
18.2	neighborhood[T.Timber]
5.1	neighborhood[T.Veenker]
1.6	kitchenqual[T.Fa]
6.2	kitchenqual[T.Gd]
1.1	kitchenqual[T.Po]
7.8	kitchenqual[T.TA]
2.0	grlivarea
1.3	bsmtfinsf1
1.1	bsmtfinsf2
3.1	overallqual
1.7	overallcond
2.3	remodelage

This produces, overall, more favorable VIF values, although some neighborhoods have high VIF values.

Removing this variable lowers the R squared value slightly, increases the AIC and produces a slightly higher RMSE score of 34024.86.

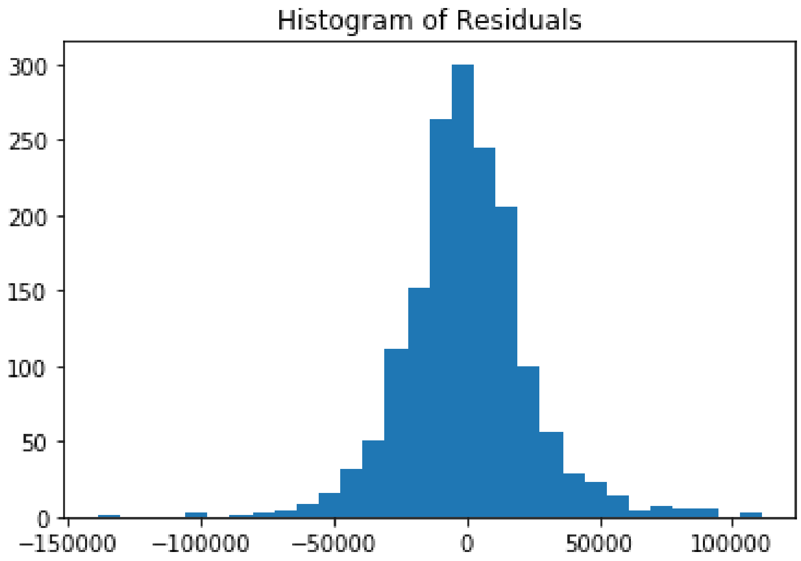
Model 20:

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.895
Model:                            OLS   Adj. R-squared:                  0.893
Method:                 Least Squares   F-statistic:                     456.1
Date:                Thu, 19 Oct 2017   Prob (F-statistic):               0.00
Time:                        21:43:15   Log-Likelihood:                -18833.
No. Observations:                1640   AIC:                         3.773e+04
Df Residuals:                    1609   BIC:                         3.790e+04
Df Model:                          30                                         
Covariance Type:            nonrobust                                         
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Intercept                1.854e+04   1.54e+04      1.206      0.228   -1.16e+04    4.87e+04
neighborhood[T.BrkSide] -2.748e+04   1.41e+04     -1.944      0.052   -5.52e+04     245.151
neighborhood[T.ClearCr]   294.1454   1.47e+04      0.020      0.984   -2.85e+04     2.9e+04
neighborhood[T.CollgCr] -1212.8987   1.39e+04     -0.088      0.930   -2.84e+04     2.6e+04
neighborhood[T.Crawfor] -2334.7790   1.42e+04     -0.165      0.869   -3.01e+04    2.55e+04
neighborhood[T.Edwards] -2.314e+04    1.4e+04     -1.650      0.099   -5.06e+04    4363.077
neighborhood[T.Gilbert] -4646.3273   1.39e+04     -0.334      0.739    -3.2e+04    2.27e+04
neighborhood[T.IDOTRR]  -3.265e+04   1.44e+04     -2.266      0.024   -6.09e+04   -4388.874
neighborhood[T.Mitchel] -1.191e+04   1.41e+04     -0.842      0.400   -3.97e+04    1.58e+04
neighborhood[T.NAmes]   -2.272e+04   1.39e+04     -1.632      0.103      -5e+04    4579.807
neighborhood[T.NWAmes]  -2.267e+04   1.41e+04     -1.611      0.107   -5.03e+04    4936.171
neighborhood[T.NoRidge]  2.822e+04   1.43e+04      1.974      0.049     181.555    5.63e+04
neighborhood[T.NridgHt]  5.325e+04   1.41e+04      3.764      0.000    2.55e+04     8.1e+04
neighborhood[T.OldTown] -3.596e+04   1.41e+04     -2.559      0.011   -6.35e+04   -8395.941
neighborhood[T.SWISU]   -3.375e+04   1.49e+04     -2.272      0.023   -6.29e+04   -4613.723
neighborhood[T.Sawyer]  -1.755e+04   1.41e+04     -1.245      0.213   -4.52e+04    1.01e+04
neighborhood[T.SawyerW] -1.641e+04    1.4e+04     -1.170      0.242   -4.39e+04    1.11e+04
neighborhood[T.Somerst]  2.381e+04    1.4e+04      1.701      0.089   -3644.126    5.13e+04
neighborhood[T.StoneBr]  6.522e+04   1.53e+04      4.252      0.000    3.51e+04    9.53e+04
neighborhood[T.Timber]   1.366e+04   1.41e+04      0.966      0.334   -1.41e+04    4.14e+04
neighborhood[T.Veenker] -1.263e+04   1.55e+04     -0.817      0.414   -4.29e+04    1.77e+04
kitchenqual[T.Fa]       -3.535e+04   5352.415     -6.605      0.000   -4.58e+04   -2.49e+04
kitchenqual[T.Gd]       -3.277e+04   2960.291    -11.069      0.000   -3.86e+04    -2.7e+04
kitchenqual[T.Po]       -5.927e+04   2.46e+04     -2.411      0.016   -1.07e+05   -1.11e+04
kitchenqual[T.TA]       -4.247e+04   3272.587    -12.977      0.000   -4.89e+04    -3.6e+04
grlivarea                  53.9799      1.811     29.808      0.000      50.428      57.532
bsmtfinsf1                 33.3390      1.549     21.524      0.000      30.301      36.377
bsmtfinsf2                 19.9935      3.697      5.408      0.000      12.742      27.246
overallqual              1.439e+04    768.059     18.729      0.000    1.29e+04    1.59e+04
overallcond              4657.6880    690.810      6.742      0.000    3302.706    6012.670
remodelage               -122.1876     35.497     -3.442      0.001    -191.813     -52.562
==============================================================================
Omnibus:                      150.511   Durbin-Watson:                   2.045
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              680.648
Skew:                           0.316   Prob(JB):                    1.58e-148
Kurtosis:                       6.092   Cond. No.                     1.75e+05
==============================================================================
```

Testing goodness of fit:

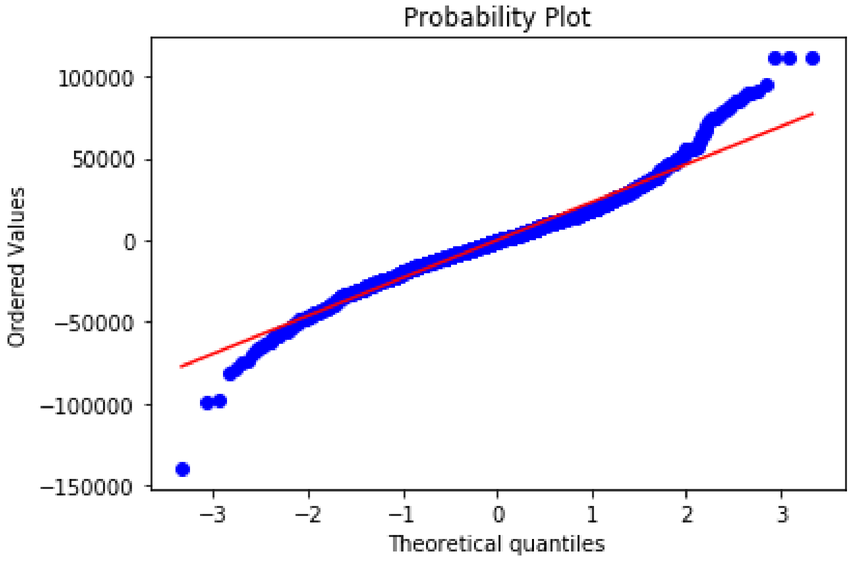
To test the goodness of fit of the model, the residuals were plotted. In figure 7a, the resulting histogram shows a normal distribution, which indicates a well fitted model.

Figure 7a:



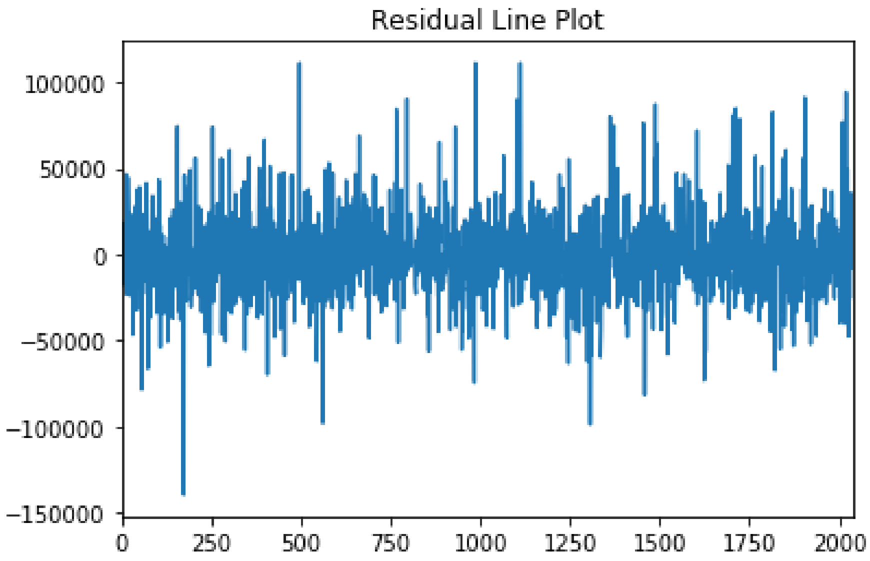
A QQ plot was also generated to confirm the histogram. As the histogram in figure 7a shows, the data is mostly normal with some outliers:

Figure 7b:



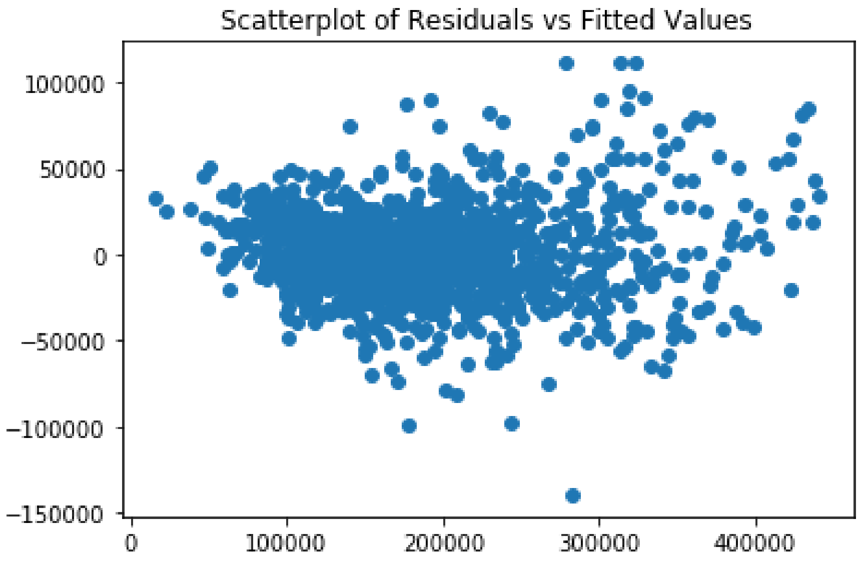
The Residual Line Plot in Figure 7c indicates that the residuals are random and centered around zero:

Figure 7c:

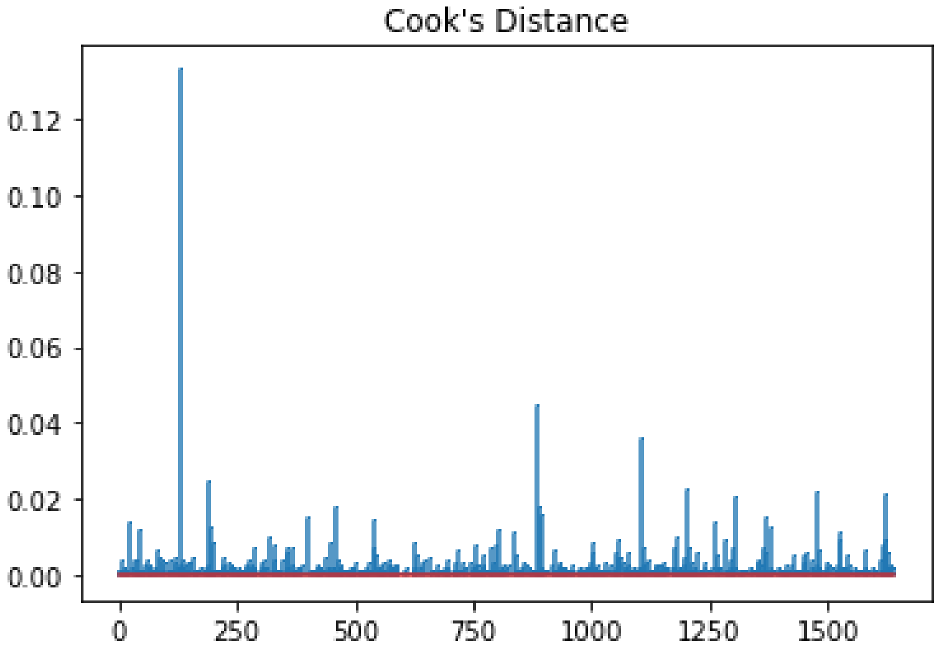


Additionally, the scatter plot of residuals versus fitted values confirms the conclusion of the line plot in figure 7c, the residuals are random and centered around zero:

Figure 7d:



Finally, Cook's Distance was calculated for this data set and plotted in figure 8. The values of Cook's Distance are low, all below 0.15, indicates the observations are equally influential on the least squares results.



Automated Selection Algorithm:

In order to explore the model further and attempt to improve upon it, a forward selection algorithm was utilized in Python. Several numerical variables were run using the f_regression formula, the results are presented in figure 9.

Figure 9 – Forward Regression Algorithm Results:

score	variable
3374.257	overallqual
2045.736	grlivarea
1329.906	garagearea
1177.089	firstflrsf
1031.153	yearbuilt
675.4785	yearremodel
407.5023	bsmtfinsf1
272.9384	lotarea
62.08469	overallcond
49.73479	bsmtunfsf
7.057965	subclass
0.177619	bsmtfinsf2

Based on the results of the forward regression, the first-floor square feet variable may improve the model. To check this result, the best performing model was run including this variable. The R squared value and AIC improve, but the RMSE score increases to 34726.33, so this variable will not be used in the final model.

Model 21:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.908
Model:                            OLS   Adj. R-squared:                  0.905
Method:                 Least Squares   F-statistic:                     356.8
Date:                Wed, 18 Oct 2017   Prob (F-statistic):               0.00
Time:                        21:20:32   Log-Likelihood:                -18725.
No. Observations:                1640   AIC:                         3.754e+04
Df Residuals:                    1595   BIC:                         3.778e+04
Df Model:                          44
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Intercept                 3.52e+04   2.79e+04      1.263      0.207   -1.95e+04    8.99e+04
neighborhood[T.BrkSide] -1.481e+04   1.34e+04     -1.109      0.268    -4.1e+04    1.14e+04
neighborhood[T.ClearCr]  8727.7910   1.38e+04      0.632      0.528   -1.84e+04    3.58e+04
neighborhood[T.CollgCr]  6457.1657    1.3e+04      0.495      0.621   -1.91e+04     3.2e+04
neighborhood[T.Crawfor]  9224.2430   1.34e+04      0.689      0.491    -1.7e+04    3.55e+04
neighborhood[T.Edwards] -1.324e+04   1.32e+04     -1.001      0.317   -3.92e+04    1.27e+04
neighborhood[T.Gilbert]  8207.8452   1.31e+04      0.624      0.533   -1.76e+04     3.4e+04
neighborhood[T.IDOTRR]  -2.156e+04   1.36e+04     -1.583      0.114   -4.83e+04    5153.329
neighborhood[T.Mitchel] -2468.5678   1.34e+04     -0.185      0.853   -2.87e+04    2.37e+04
neighborhood[T.NAmes]   -1.368e+04   1.31e+04     -1.041      0.298   -3.95e+04    1.21e+04
neighborhood[T.NWAmes]  -1.065e+04   1.33e+04     -0.800      0.424   -3.68e+04    1.55e+04
neighborhood[T.NoRidge]   3.72e+04   1.35e+04      2.762      0.006    1.08e+04    6.36e+04
neighborhood[T.NridgHt]  5.106e+04   1.33e+04      3.832      0.000    2.49e+04    7.72e+04
neighborhood[T.OldTown] -2.412e+04   1.33e+04     -1.818      0.069   -5.02e+04    1909.615
neighborhood[T.SWISU]   -2.038e+04    1.4e+04     -1.451      0.147   -4.79e+04    7169.874
neighborhood[T.Sawyer]  -6961.0816   1.33e+04     -0.523      0.601   -3.31e+04    1.91e+04
neighborhood[T.SawyerW] -6408.9908   1.32e+04     -0.485      0.628   -3.23e+04    1.95e+04
neighborhood[T.Somerst]  2.797e+04   1.32e+04      2.124      0.034    2139.162    5.38e+04
neighborhood[T.StoneBr]  6.734e+04   1.45e+04      4.651      0.000    3.89e+04    9.57e+04
neighborhood[T.Timber]   1.645e+04   1.33e+04      1.236      0.217   -9658.732    4.26e+04
neighborhood[T.Veenker] -9055.6252   1.48e+04     -0.610      0.542   -3.82e+04    2.01e+04
qualcond[T.ExGd]        -2.616e+04     2.7e+04     -0.969      0.333   -7.91e+04    2.68e+04
qualcond[T.ExTA]         3195.8433   2.39e+04      0.134      0.893   -4.36e+04       5e+04
qualcond[T.FaFa]        -5.946e+04   2.58e+04     -2.303      0.021    -1.1e+05   -8813.689
qualcond[T.FaGd]        -3.791e+04   3.28e+04     -1.157      0.248   -1.02e+05    2.64e+04
qualcond[T.FaTA]        -4.318e+04   2.58e+04     -1.671      0.095   -9.39e+04    7516.118
qualcond[T.GdEx]        -8520.1634   2.87e+04     -0.297      0.766   -6.47e+04    4.77e+04
qualcond[T.GdFa]         -5.77e+04   3.19e+04     -1.808      0.071    -1.2e+05    4895.353
qualcond[T.GdGd]        -3.727e+04   2.38e+04     -1.569      0.117   -8.39e+04    9328.690
qualcond[T.GdTA]        -3.182e+04   2.37e+04     -1.340      0.180   -7.84e+04    1.47e+04
qualcond[T.TAEx]        -4.214e+04   2.61e+04     -1.614      0.107   -9.34e+04    9078.481
qualcond[T.TAFa]        -5.083e+04   2.44e+04     -2.086      0.037   -9.86e+04   -3034.261
qualcond[T.TAGd]         -4.59e+04   2.38e+04     -1.928      0.054   -9.26e+04     795.545
qualcond[T.TATA]        -4.411e+04   2.38e+04     -1.856      0.064   -9.07e+04    2501.451
kitchenqual[T.Fa]       -2.223e+04   5205.516     -4.271      0.000   -3.24e+04     -1.2e+04
kitchenqual[T.Gd]       -2.089e+04   3090.642     -6.759      0.000    -2.7e+04   -1.48e+04
kitchenqual[T.Po]       -3.696e+04   2.32e+04     -1.594      0.111   -8.24e+04    8507.019
kitchenqual[T.TA]       -2.768e+04   3359.828     -8.238      0.000   -3.43e+04   -2.11e+04
grlivarea                  48.5450      1.824     26.616      0.000      44.968      52.123
bsmtfinsf1                 28.4029      1.542     18.424      0.000      25.379      31.427
bsmtfinsf2                 14.2458      3.543      4.021      0.000       7.296      21.195
overallqual              1.145e+04    765.342     14.965      0.000    9952.390     1.3e+04
overallcond              5377.7356    695.969      7.727      0.000    4012.626    6742.846
remodelage               -113.5883     33.742     -3.366      0.001    -179.772     -47.404
firstflrsf                 21.4647      2.240      9.580      0.000      17.070      25.859
==============================================================================
Omnibus:                      145.059   Durbin-Watson:                   2.033
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              865.800
Skew:                           0.121   Prob(JB):                    9.86e-189
Kurtosis:                       6.551   Cond. No.                     3.23e+05
==============================================================================
```

Model Selection:

The final model selected to predict housing prices and improve upon the model selected in the first analysis improves the R squared, AIC, and reduces the RMSE as much as possible while avoiding variables that are highly correlated.

To recap, the figure 10 contains a summary of all the models evaluated for selection in this analysis:

Figure 10 – Regression Models:

	R Squared	Adj R Squared	AIC	RMSE
neighborhood, grlivarea, overallqual, bsmtfinsf1	0.878	0.876	3.80E+04	35544.85939
neighborhood, grlivarea, overallqual, bsmtfinsf1, lotarea	0.881	0.880	3.79E+04	36055.49455
neighborhood, grlivarea, overallqual, bsmtfinsf1, lotarea, overallcond	0.885	0.883	3.79E+04	35770.31651
neighborhood, grlivarea, overallqual, bsmtfinsf1, lotarea, overallcond, yearbuilt	0.894	0.893	3.77E+04	36171.45013
neighborhood, grlivarea, overallqual, bsmtfinsf1, lotarea, overallcond, yearbuilt, totalbsmtsf	0.902	0.901	3.76E+04	37184.17778
neighborhood, totalsqftcalc, overallqual	0.875	0.873	3.80E+04	36151.03201
neighborhood, grlivarea,bsmtfinsf1, bsmtfinsf2, overallqual	0.880	0.878	3.79E+04	35321.76
neighborhood, grlivarea,bsmtfinsf1, qualityindex	0.869	0.867	3.81E+04	35463.68
neighborhood, grlivarea,bsmtfinsf1, overallqual, overallcond	0.880	0.879	3.79E+04	35172.54
neighborhood, grlivarea,bsmtfinsf1,bsmtfinsf2, overallqual, overallcond	0.883	0.881	3.79E+04	34927.22
neighborhood, grlivarea,bsmtfinsf1,bsmtfinsf2, overallqual, overallcond,yearbuilt	0.891	0.890	3.78E+04	35088.89
neighborhood, grlivarea,bsmtfinsf1,bsmtfinsf2, overallqual, overallcond,yearremodel	0.886	0.885	3.79E+04	35013.98
neighborhood, grlivarea,bsmtfinsf1,bsmtfinsf2, overallqual, overallcond,remodelage	0.884	0.882	3.79E+04	34927.23
neighborhood, qualcond, grlivarea,bsmtfinsf1,bsmtfinsf2, overallqual, overallcond,remodelage	0.898	0.896	3.77E+04	33621.98
neighborhood, qualcond, kitchenqual, grlivarea,bsmtfinsf1,bsmtfinsf2, overallqual, overallcond,remodelage	0.902	0.900	3.76E+04	33340.86
neighborhood, qualcond, kitchenqual, grlivarea,bsmtfinsf1,bsmtfinsf2, overallqual, overallcond,remodelage,bsmtunfsf	0.909	0.906	3.75E+04	34726.32
log(saleprice)~neighborhood, qualcond, kitchenqual, grlivarea,bsmtfinsf1,bsmtfinsf2, overallqual, overallcond,remodelage,bsmtunfsf	0.898	0.896	-2.21E+03	88925.35
neighborhood, kitchenqual, grlivarea,bsmtfinsf1,bsmtfinsf2, overallqual, overallcond,remodelage	0.895	0.893	3.77E+04	34024.86

The model that was chosen has the lowest RMSE score of 33340.86 and combines the neighborhood, finished basement square feet type 1 and 2, kitchen quality, above grade living area, overall quality, overall condition along with the created variables of qualcond (exterior quality and exterior condition) and remodel age (remodel year subtracted from built year.)

Model Formula:

$$\begin{aligned}\boldsymbol{p\_saleprice} &= (\boldsymbol{lookupneighborhoodvalue}) + (\boldsymbol{lookupqualcondvalue}) \\ &+ (\boldsymbol{lookupkitchenqualvalue}) + \ \mathbf{54.1586} * (\boldsymbol{grlivearea}) + \mathbf{32.9770} \\ &* (\boldsymbol{bsmtfinsf1}) + \mathbf{20.2707} * (\boldsymbol{bsmtfinsf2}) + \mathbf{11890} * (\boldsymbol{overallqual}) \\ &+ \mathbf{4851.8248} * (\boldsymbol{overallcond}) - \mathbf{122.4095} * (\boldsymbol{remodelage}) + \mathbf{57040}\end{aligned}$$

lookupneighborhoodvalue	
BrkSide	-20320
ClearCr	5770.356
CollgCr	1499.6751
Crawfor	6031.9044
Edwards	-17230
Gilbert	-982.9886
IDOTRR	-26710
Mitchel	-5928.9617
NAmes	-16080
NWAmes	-13670
NoRidge	30650
NridgHt	49770
OldTown	-29500
SWISU	-27080
Sawyer	-11280
SawyerW	-13140
Somerst	24840
StoneBr	67230
Timber	15210
Veenker	-9060.5843

lookupqualcondvalue		
externalqual	externalcond	value
Ex	Gd	-32700
Ex	TA	4843
Fa	Fa	-61920
Fa	Gd	-43820
Fa	TA	-46290
Gd	Ex	-21530
Gd	Fa	-56240
Gd	Gd	-39580
Gd	TA	-33510
TA	Ex	-46140
TA	Fa	-52530
TA	Gd	-48490
TA	TA	-46710

lookupkitchenqualvalue	
Fa	-22950
Gd	-21480
Po	-44850
TA	-28800

The first step of the equation requires the user determine the Neighborhood the home resides in and lookup the value of that Neighborhood to substitute into the equation, followed by determining the exterior quality and exterior condition values to look up the value of the combined variable to add to the neighborhood value. Next, the kitchen quality value is looked up in its respective table and added to the equation. The values for above grade living area, finished basement type 1, finished basement type 1, overall quality, overall condition, and remodel age (a function of remodel year minus built year) are multiplied by their coefficients and added together to the intercept. The result is the predicted home value.

Please note:

- Any sale price that was produced by the model that was negative has been replaced with zero, assuming that the homes that produce a negative value are undesirable but do not have negative value.

Conclusion:

The cleansed training file from the previous analysis was used to improve the model created prior in order to predict the sales price of a “typical” home in Ames, Iowa. Several variables were “layered” upon the model chosen in the previous analysis and, though these improved the R squared value, the RMSE did not improve, indicated the models were not a good fit.

Additional fitting techniques were used to improve the R squared value, such as creating variable groups (in this case, grouping neighborhood based on housing cost per square foot and over/under predictions) and creating new variables by combining existing variables in new ways. These calculated variables were used in the model to test to determine if they improved the model. A log transformation was performed on sale price to attempt to improve the model. The log transformation did not improve model.

The goodness of fit of the model was tested using several graphs of the residuals, such as a histogram and Cook’s distance and it was determined the model was a good fit.

Additionally, an automated selection algorithm was utilized to determine if other variables could be added to the model to improve the performance. The variable that the algorithm identified, but was not already utilized in the model, were tested in the regression model, but did not improve the RMSE score.

The final equation uses nine variables. The variables chosen have an adjusted R-squared value of 0.900 and an RMSE score of 33340.86 on Kaggle.

The results were reviewed and implausible results were replaced with null values to make the resulting data file more realistic.

Appendix A - Summary of adjusted data set with outliers and anomalies removed (subclass, month sold, year sold, and index removed from summary):

	lotfrontage	lotarea	overallqual	overallcond	yearbuilt	yearremodel	masvnrarea
count	1640	1640	1640	1640	1640	1640	1640
mean	71.968469	10346.40427	6.071951	5.70122	1969.390854	1983.938415	93.20061
std	17.291501	3994.96215	1.343756	1.101222	30.593962	21.535486	167.997638
min	30	2500	2	3	1872	1950	0
25%	60	8097	5	5	1950	1962	0
50%	68.58168	9644	6	5	1968.5	1994	0
75%	80	11700	7	6	2000	2004	145
max	200	47280	10	9	2010	2010	1290

	bsmtfinsf1	bsmtfinsf2	bsmtunfsf	totalbsmtsf	firstflrsf	secondflrsf	lowqualfinsf	grlivarea
count	1640	1640	1640	1640	1640	1640	1640	1640
mean	429.911585	49.739634	560.95122	1040.602439	1149.926829	327.879878	3.953049	1481.759756
std	428.042406	164.415109	416.707645	394.76659	354.446084	415.217594	40.066749	462.614861
min	0	0	0	0	407	0	0	407
25%	0	0	238.75	797.5	884	0	0	1120
50%	369.5	0	481.5	972	1072	0	0	1444
75%	716	0	796.25	1264.25	1362.25	714	0	1749.25
max	2158	1526	2336	2846	2898	1611	697	2978

	bsmtfullbath	bsmthalfbath	fullbath	halfbath	kitchenabvgr	totrmsabvgrd	fireplaces
count	1640	1640	1640	1640	1640	1640	1640
mean	0.409146	0.065244	1.514634	0.37378	1.003049	6.442683	0.613415
std	0.49799	0.249489	0.530721	0.490217	0.07404	1.40265	0.647379
min	0	0	0	0	0	3	0
25%	0	0	1	0	1	5	0
50%	0	0	2	0	1	6	1
75%	1	0	2	1	1	7	1
max	2	2	3	2	3	12	4

	garageyrblt	garagecars	garagearea	wooddecksf	openporchsf	enclosedporch	threessnporch	screenporch
count	1576	1640	1640	1640	1640	1640	1640	1640
mean	1976.18401	1.765854	475.89878	96.401829	47.283537	23.415854	2.762805	16.105488
std	25.904448	0.737759	210.841976	125.83348	64.363401	62.651103	25.575532	54.965022
min	1900	0	0	0	0	0	0	0
25%	1958	1	315.75	0	0	0	0	0
50%	1977	2	477.5	0	26	0	0	0
75%	2001	2	576	172.25	72	0	0	0
max	2010	5	1488	690	547	584	407	576

	poolarea	miscval	mosold	saleprice	qualityindex	totalsqftcalc	remodelage
count	1640	1640	1640	1640	1640	1640	1640
mean	2.389634	46.085366	6.277439	180418.025	34.343902	1961.410976	14.547561
std	38.497466	403.237408	2.710651	72458.81818	8.718009	693.252426	25.135221
min	0	0	1	37900	8	407	-1
25%	0	0	4	129975	30	1489	0
50%	0	0	6	163000	35	1834	0
75%	0	0	8	214925	40	2352.25	24
max	800	12500	12	500000	72	4958	123