**Introduction**

This assignment was a great challenge to test many of the techniques that we have learned in the class thus far. The goal for the second part of this project was to create more accurate models for the Ames, IA housing market. I did this by using multiple linear regression techniques on various combinations of variables for the models. The building blocks for each model were based upon shell code from class and the models from the first part of the assignment. The goal of adding additional variables to the model is to minimize the mean square root errors while also increasing the precision of the model. This will be measured by R squared scores and AIC and BIC values.

**Section 1. Modeling & More**

Model 1

OLS Regression Results

==============================================================================

Dep. Variable: Y R-squared: 0.808

Model: OLS Adj. R-squared: 0.803

Method: Least Squares F-statistic: 167.4

Date: Thu, 14 Feb 2019 Prob (F-statistic): 0.00

Time: 00:25:51 Log-Likelihood: -24204.

No. Observations: 2039 AIC: 4.851e+04

Df Residuals: 1988 BIC: 4.880e+04

Df Model: 50

Covariance Type: nonrobust

==============================================================================================

coef std err t P>|t| [0.025 0.975]

----------------------------------------------------------------------------------------------

Intercept -1.093e+06 1.23e+05 -8.903 0.000 -1.33e+06 -8.52e+05

C(lotconfig)[T.CulDSac] 9959.0566 3853.710 2.584 0.010 2401.323 1.75e+04

C(lotconfig)[T.FR2] -9229.1512 5358.716 -1.722 0.085 -1.97e+04 1280.137

C(lotconfig)[T.FR3] -4.164e+04 1.14e+04 -3.654 0.000 -6.4e+04 -1.93e+04

C(lotconfig)[T.Inside] 1090.6206 2122.872 0.514 0.607 -3072.667 5253.909

C(housestyle)[T.1.5Unf] 3571.6500 9857.072 0.362 0.717 -1.58e+04 2.29e+04

C(housestyle)[T.1Story] -3608.3476 3057.079 -1.180 0.238 -9603.762 2387.067

C(housestyle)[T.2.5Fin] 1.207e+04 1.5e+04 0.805 0.421 -1.73e+04 4.15e+04

C(housestyle)[T.2.5Unf] -3655.5504 9215.417 -0.397 0.692 -2.17e+04 1.44e+04

C(housestyle)[T.2Story] -3197.4247 3239.242 -0.987 0.324 -9550.090 3155.241

C(housestyle)[T.SFoyer] -1.251e+04 5437.443 -2.302 0.021 -2.32e+04 -1851.083

C(housestyle)[T.SLvl] -1.08e+04 4842.278 -2.231 0.026 -2.03e+04 -1306.732

C(roofstyle)[T.Gable] -5849.0281 1.08e+04 -0.542 0.588 -2.7e+04 1.53e+04

C(roofstyle)[T.Gambr] -4787.4578 1.34e+04 -0.357 0.721 -3.11e+04 2.15e+04

C(roofstyle)[T.Hip] 9712.8899 1.09e+04 0.890 0.374 -1.17e+04 3.11e+04

C(roofstyle)[T.Mansa] 5865.3970 1.65e+04 0.354 0.723 -2.66e+04 3.83e+04

C(roofstyle)[T.Shed] -8503.5956 2.05e+04 -0.415 0.678 -4.87e+04 3.17e+04

C(heating)[T.GasA] 9710.3288 3.55e+04 0.274 0.784 -5.98e+04 7.92e+04

C(heating)[T.GasW] 2.029e+04 3.62e+04 0.560 0.576 -5.08e+04 9.14e+04

C(heating)[T.Grav] 6528.5477 3.78e+04 0.173 0.863 -6.76e+04 8.06e+04

C(heating)[T.OthW] -1.5e+04 5.02e+04 -0.299 0.765 -1.13e+05 8.34e+04

C(heating)[T.Wall] 2.924e+04 4.33e+04 0.675 0.500 -5.57e+04 1.14e+05

C(neighborhood)[T.Blueste] -3.636e+04 1.57e+04 -2.311 0.021 -6.72e+04 -5508.840

C(neighborhood)[T.BrDale] -4.718e+04 1.17e+04 -4.020 0.000 -7.02e+04 -2.42e+04

C(neighborhood)[T.BrkSide] -1.747e+04 1.01e+04 -1.732 0.083 -3.73e+04 2308.281

C(neighborhood)[T.ClearCr] 7050.8946 1.11e+04 0.637 0.524 -1.46e+04 2.88e+04

C(neighborhood)[T.CollgCr] -1057.3237 8718.998 -0.121 0.903 -1.82e+04 1.6e+04

C(neighborhood)[T.Crawfor] 1.409e+04 9833.075 1.433 0.152 -5195.611 3.34e+04

C(neighborhood)[T.Edwards] -2.617e+04 9219.795 -2.839 0.005 -4.43e+04 -8092.035

C(neighborhood)[T.Gilbert] -795.6399 9079.647 -0.088 0.930 -1.86e+04 1.7e+04

C(neighborhood)[T.Greens] -2.381e+04 1.67e+04 -1.425 0.154 -5.66e+04 8957.083

C(neighborhood)[T.GrnHill] 1.029e+05 3.6e+04 2.857 0.004 3.23e+04 1.74e+05

C(neighborhood)[T.IDOTRR] -1.597e+04 1.05e+04 -1.528 0.127 -3.65e+04 4529.173

C(neighborhood)[T.MeadowV] -4.371e+04 1.13e+04 -3.868 0.000 -6.59e+04 -2.15e+04

C(neighborhood)[T.Mitchel] -1.809e+04 9393.775 -1.925 0.054 -3.65e+04 336.029

C(neighborhood)[T.NAmes] -2.392e+04 8949.379 -2.673 0.008 -4.15e+04 -6373.808

C(neighborhood)[T.NPkVill] -3.32e+04 1.22e+04 -2.728 0.006 -5.71e+04 -9335.589

C(neighborhood)[T.NWAmes] -2.214e+04 9352.744 -2.367 0.018 -4.05e+04 -3798.909

C(neighborhood)[T.NoRidge] 5.368e+04 9976.588 5.381 0.000 3.41e+04 7.32e+04

C(neighborhood)[T.NridgHt] 7.071e+04 9023.361 7.836 0.000 5.3e+04 8.84e+04

C(neighborhood)[T.OldTown] -1.739e+04 9824.114 -1.770 0.077 -3.67e+04 1878.279

C(neighborhood)[T.SWISU] -1.96e+04 1.16e+04 -1.686 0.092 -4.24e+04 3202.449

C(neighborhood)[T.Sawyer] -2.548e+04 9361.979 -2.721 0.007 -4.38e+04 -7118.135

C(neighborhood)[T.SawyerW] -1.445e+04 9185.627 -1.573 0.116 -3.25e+04 3569.249

C(neighborhood)[T.Somerst] 1.697e+04 8900.968 1.907 0.057 -482.153 3.44e+04

C(neighborhood)[T.StoneBr] 5.574e+04 1.05e+04 5.323 0.000 3.52e+04 7.63e+04

C(neighborhood)[T.Timber] 2.038e+04 9532.920 2.138 0.033 1684.484 3.91e+04

C(neighborhood)[T.Veenker] 8445.5088 1.28e+04 0.659 0.510 -1.67e+04 3.36e+04

qualityindex 2135.8233 99.771 21.407 0.000 1940.156 2331.490

totalsqftcalc 39.7080 1.274 31.162 0.000 37.209 42.207

yearbuilt 569.4023 58.348 9.759 0.000 454.973 683.832

==============================================================================

Omnibus: 605.595 Durbin-Watson: 2.040

Prob(Omnibus): 0.000 Jarque-Bera (JB): 47418.402

Skew: -0.429 Prob(JB): 0.00

Kurtosis: 26.609 Cond. No. 4.70e+05

==============================================================================

My first model above used: qualityindex, totalsqftcalc, yearbuilt, neighborhood, heating, roofstyle, housestyle and lotconfig. I choose these variables based off of correlation matrix python produced. It has a good R-Squared score and low AIC and BIC scored. Overall most of the nominal variables like heating, seemed to have less significant p-values.

Model 2

OLS Regression Results

==============================================================================

Dep. Variable: Y R-squared: 0.802

Model: OLS Adj. R-squared: 0.798

Method: Least Squares F-statistic: 245.6

Date: Thu, 14 Feb 2019 Prob (F-statistic): 0.00

Time: 00:53:05 Log-Likelihood: -24237.

No. Observations: 2039 AIC: 4.854e+04

Df Residuals: 2005 BIC: 4.873e+04

Df Model: 33

Covariance Type: nonrobust

==============================================================================================

coef std err t P>|t| [0.025 0.975]

----------------------------------------------------------------------------------------------

Intercept -1.031e+06 1.12e+05 -9.217 0.000 -1.25e+06 -8.12e+05

C(lotconfig)[T.CulDSac] 9938.0849 3867.921 2.569 0.010 2352.520 1.75e+04

C(lotconfig)[T.FR2] -8986.4899 5393.883 -1.666 0.096 -1.96e+04 1591.712

C(lotconfig)[T.FR3] -4.599e+04 1.15e+04 -4.003 0.000 -6.85e+04 -2.35e+04

C(lotconfig)[T.Inside] 1375.5535 2134.445 0.644 0.519 -2810.408 5561.515

C(neighborhood)[T.Blueste] -3.917e+04 1.59e+04 -2.470 0.014 -7.03e+04 -8070.036

C(neighborhood)[T.BrDale] -4.974e+04 1.17e+04 -4.260 0.000 -7.26e+04 -2.68e+04

C(neighborhood)[T.BrkSide] -1.868e+04 1.01e+04 -1.849 0.065 -3.85e+04 1136.236

C(neighborhood)[T.ClearCr] 5848.2137 1.09e+04 0.534 0.593 -1.56e+04 2.73e+04

C(neighborhood)[T.CollgCr] -2813.0286 8794.937 -0.320 0.749 -2.01e+04 1.44e+04

C(neighborhood)[T.Crawfor] 1.307e+04 9913.581 1.318 0.188 -6371.445 3.25e+04

C(neighborhood)[T.Edwards] -2.68e+04 9270.142 -2.891 0.004 -4.5e+04 -8616.638

C(neighborhood)[T.Gilbert] -3368.6064 9037.078 -0.373 0.709 -2.11e+04 1.44e+04

C(neighborhood)[T.Greens] -2.797e+04 1.69e+04 -1.656 0.098 -6.11e+04 5148.418

C(neighborhood)[T.GrnHill] 9.827e+04 3.65e+04 2.696 0.007 2.68e+04 1.7e+05

C(neighborhood)[T.IDOTRR] -1.708e+04 1.05e+04 -1.625 0.104 -3.77e+04 3532.006

C(neighborhood)[T.MeadowV] -4.851e+04 1.13e+04 -4.311 0.000 -7.06e+04 -2.64e+04

C(neighborhood)[T.Mitchel] -2.115e+04 9428.589 -2.244 0.025 -3.96e+04 -2664.022

C(neighborhood)[T.NAmes] -2.345e+04 9004.978 -2.604 0.009 -4.11e+04 -5785.905

C(neighborhood)[T.NPkVill] -3.653e+04 1.22e+04 -2.982 0.003 -6.06e+04 -1.25e+04

C(neighborhood)[T.NWAmes] -2.34e+04 9415.411 -2.486 0.013 -4.19e+04 -4938.379

C(neighborhood)[T.NoRidge] 5.517e+04 9990.785 5.522 0.000 3.56e+04 7.48e+04

C(neighborhood)[T.NridgHt] 7.439e+04 9100.565 8.174 0.000 5.65e+04 9.22e+04

C(neighborhood)[T.OldTown] -1.889e+04 9895.376 -1.909 0.056 -3.83e+04 512.709

C(neighborhood)[T.SWISU] -1.994e+04 1.16e+04 -1.717 0.086 -4.27e+04 2835.356

C(neighborhood)[T.Sawyer] -2.689e+04 9392.435 -2.863 0.004 -4.53e+04 -8472.736

C(neighborhood)[T.SawyerW] -1.735e+04 9240.592 -1.877 0.061 -3.55e+04 776.689

C(neighborhood)[T.Somerst] 1.558e+04 8954.068 1.740 0.082 -1982.433 3.31e+04

C(neighborhood)[T.StoneBr] 5.678e+04 1.06e+04 5.361 0.000 3.6e+04 7.75e+04

C(neighborhood)[T.Timber] 2.107e+04 9627.659 2.189 0.029 2188.975 4e+04

C(neighborhood)[T.Veenker] 3203.2929 1.29e+04 0.248 0.804 -2.21e+04 2.85e+04

qualityindex 2154.8081 98.905 21.787 0.000 1960.840 2348.776

totalsqftcalc 41.5578 1.230 33.788 0.000 39.146 43.970

yearbuilt 537.9639 55.761 9.648 0.000 428.609 647.319

==============================================================================

Omnibus: 590.736 Durbin-Watson: 2.039

Prob(Omnibus): 0.000 Jarque-Bera (JB): 48577.684

Skew: -0.354 Prob(JB): 0.00

Kurtosis: 26.902 Cond. No. 4.10e+05

==============================================================================

My second model above used: qualityindex, totalsqftcalc, yearbuilt, neighborhood, and lotconfig. I dropped heating, roof style and housestyle from this model due to high p-values. Unfortunately, it did not improve my model and the AIC and BIC scores went up, along with a decrease in the R-Squared score.

Model 3

OLS Regression Results

==============================================================================

Dep. Variable: Y R-squared: 0.802

Model: OLS Adj. R-squared: 0.800

Method: Least Squares F-statistic: 471.9

Date: Thu, 14 Feb 2019 Prob (F-statistic): 0.00

Time: 01:05:22 Log-Likelihood: -19255.

No. Observations: 1645 AIC: 3.854e+04

Df Residuals: 1630 BIC: 3.862e+04

Df Model: 14

Covariance Type: nonrobust

===========================================================================================

coef std err t P>|t| [0.025 0.975]

-------------------------------------------------------------------------------------------

Intercept -1.514e+06 5.77e+04 -26.219 0.000 -1.63e+06 -1.4e+06

C(lotconfig)[T.CulDSac] -2794.1488 3603.313 -0.775 0.438 -9861.760 4273.463

C(lotconfig)[T.FR2] -1.473e+04 4685.372 -3.144 0.002 -2.39e+04 -5541.025

C(lotconfig)[T.FR3] -1.149e+04 1.06e+04 -1.083 0.279 -3.23e+04 9323.235

C(lotconfig)[T.Inside] -1227.8088 1973.864 -0.622 0.534 -5099.386 2643.768

C(housestyle)[T.1.5Unf] 1.03e+04 8513.429 1.210 0.227 -6398.895 2.7e+04

C(housestyle)[T.1Story] -2862.0665 2608.078 -1.097 0.273 -7977.605 2253.472

C(housestyle)[T.2.5Fin] 2.658e+04 1.5e+04 1.772 0.077 -2840.752 5.6e+04

C(housestyle)[T.2.5Unf] 7862.9373 8506.186 0.924 0.355 -8821.270 2.45e+04

C(housestyle)[T.2Story] -5676.1508 2798.566 -2.028 0.043 -1.12e+04 -186.987

C(housestyle)[T.SFoyer] -2.695e+04 4844.866 -5.563 0.000 -3.65e+04 -1.74e+04

C(housestyle)[T.SLvl] -1.747e+04 4168.740 -4.191 0.000 -2.56e+04 -9294.063

qualityindex 2301.0798 88.532 25.991 0.000 2127.431 2474.729

totalsqftcalc 56.0250 1.267 44.229 0.000 53.540 58.509

yearbuilt 764.7805 30.206 25.319 0.000 705.534 824.027

==============================================================================

Omnibus: 251.950 Durbin-Watson: 1.974

Prob(Omnibus): 0.000 Jarque-Bera (JB): 583.201

Skew: 0.864 Prob(JB): 2.29e-127

Kurtosis: 5.350 Cond. No. 2.23e+05

==============================================================================

The above model produced a better model for me. The p-values from house style are the main concerns I have. I also used the lotconfig, qualityindex, totalsqftcalc and yearbuilt as my other variables. The R-Squared scores are consistent from previous models, but the AIC and BIC scores are also much more improved and help to show evidence of a potentially better model. I did a VIF for this model as well for comparison purposes. Due to high values of correlation for housestyle and lotconfig, I dropped those and did another Model that I call four.

**This is the VIF for Model 3**

Intercept 6321.423581

C(lotconfig)[T.CulDSac] 1.286576

C(lotconfig)[T.FR2] 1.155319

C(lotconfig)[T.FR3] 1.033131

C(lotconfig)[T.Inside] 1.415540

C(housestyle)[T.1.5Unf] 1.077486

C(housestyle)[T.1Story] 3.223679

C(housestyle)[T.2.5Fin] 1.034751

C(housestyle)[T.2.5Unf] 1.075653

C(housestyle)[T.2Story] 3.105904

C(housestyle)[T.SFoyer] 1.337091

C(housestyle)[T.SLvl] 1.470230

qualityindex 1.185466

totalsqftcalc 1.351872

yearbuilt 1.498649

Model 4

OLS Regression Results

==============================================================================

Dep. Variable: Y R-squared: 0.794

Model: OLS Adj. R-squared: 0.793

Method: Least Squares F-statistic: 2103.

Date: Thu, 14 Feb 2019 Prob (F-statistic): 0.00

Time: 01:14:28 Log-Likelihood: -19289.

No. Observations: 1645 AIC: 3.859e+04

Df Residuals: 1641 BIC: 3.861e+04

Df Model: 3

Covariance Type: nonrobust

=================================================================================

coef std err t P>|t| [0.025 0.975]

---------------------------------------------------------------------------------

Intercept -1.399e+06 5.24e+04 -26.730 0.000 -1.5e+06 -1.3e+06

qualityindex 2312.9261 88.300 26.194 0.000 2139.733 2486.120

totalsqftcalc 56.4671 1.262 44.753 0.000 53.992 58.942

yearbuilt 702.9255 26.917 26.114 0.000 650.130 755.721

==============================================================================

Omnibus: 269.946 Durbin-Watson: 1.972

Prob(Omnibus): 0.000 Jarque-Bera (JB): 635.477

Skew: 0.914 Prob(JB): 1.02e-138

Kurtosis: 5.435 Cond. No. 1.99e+05

==============================================================================

The above model performed better in regards to having better p-values and better BIC score. This model scored in the lower part of the 40,000 RMSE.

Model 5

OLS Regression Results

==============================================================================

Dep. Variable: saleprice R-squared: 0.896

Model: OLS Adj. R-squared: 0.894

Method: Least Squares F-statistic: 435.3

Date: Thu, 14 Feb 2019 Prob (F-statistic): 0.00

Time: 01:19:46 Log-Likelihood: -18723.

No. Observations: 1645 AIC: 3.751e+04

Df Residuals: 1612 BIC: 3.769e+04

Df Model: 32

Covariance Type: nonrobust

===========================================================================================

coef std err t P>|t| [0.025 0.975]

-------------------------------------------------------------------------------------------

Intercept -8.341e+05 7.85e+04 -10.623 0.000 -9.88e+05 -6.8e+05

neighborhood[T.Blueste] -1.713e+04 1.09e+04 -1.572 0.116 -3.85e+04 4239.347

neighborhood[T.BrDale] -2.471e+04 9000.007 -2.745 0.006 -4.24e+04 -7052.075

neighborhood[T.BrkSide] -1532.4518 8096.964 -0.189 0.850 -1.74e+04 1.43e+04

neighborhood[T.ClearCr] 8014.4751 8662.334 0.925 0.355 -8976.144 2.5e+04

neighborhood[T.CollgCr] 87.4515 7385.889 0.012 0.991 -1.44e+04 1.46e+04

neighborhood[T.Crawfor] 1.561e+04 8040.486 1.942 0.052 -158.060 3.14e+04

neighborhood[T.Edwards] -8651.3826 7681.778 -1.126 0.260 -2.37e+04 6415.938

neighborhood[T.Gilbert] 2292.6064 7583.504 0.302 0.762 -1.26e+04 1.72e+04

neighborhood[T.Greens] -391.6243 1.14e+04 -0.034 0.973 -2.28e+04 2.2e+04

neighborhood[T.GrnHill] 1.171e+05 2.26e+04 5.172 0.000 7.27e+04 1.62e+05

neighborhood[T.IDOTRR] -8670.2039 8490.853 -1.021 0.307 -2.53e+04 7984.067

neighborhood[T.MeadowV] -2.211e+04 8677.187 -2.548 0.011 -3.91e+04 -5088.364

neighborhood[T.Mitchel] -8381.0179 7719.579 -1.086 0.278 -2.35e+04 6760.447

neighborhood[T.NAmes] -1.281e+04 7502.862 -1.708 0.088 -2.75e+04 1903.047

neighborhood[T.NPkVill] -2.497e+04 9031.525 -2.764 0.006 -4.27e+04 -7250.268

neighborhood[T.NWAmes] -1.741e+04 7701.140 -2.261 0.024 -3.25e+04 -2308.925

neighborhood[T.NoRidge] 2.984e+04 8021.701 3.720 0.000 1.41e+04 4.56e+04

neighborhood[T.NridgHt] 4.205e+04 7700.404 5.461 0.000 2.69e+04 5.72e+04

neighborhood[T.OldTown] -1.18e+04 7998.975 -1.475 0.141 -2.75e+04 3894.200

neighborhood[T.SWISU] -1.393e+04 9062.906 -1.537 0.124 -3.17e+04 3845.856

neighborhood[T.Sawyer] -1.182e+04 7707.552 -1.533 0.125 -2.69e+04 3299.222

neighborhood[T.SawyerW] -8603.3201 7619.053 -1.129 0.259 -2.35e+04 6340.971

neighborhood[T.Somerst] 1.368e+04 7582.818 1.804 0.071 -1196.262 2.86e+04

neighborhood[T.StoneBr] 2.836e+04 8440.220 3.360 0.001 1.18e+04 4.49e+04

neighborhood[T.Timber] 1.622e+04 8007.961 2.026 0.043 513.577 3.19e+04

neighborhood[T.Veenker] 8500.2699 9774.036 0.870 0.385 -1.07e+04 2.77e+04

qualityindex 1677.5519 67.639 24.801 0.000 1544.882 1810.222

totalsqftcalc 55.5622 1.530 36.320 0.000 52.562 58.563

totalbsmtsf -8.2319 2.549 -3.230 0.001 -13.231 -3.233

garagearea 39.6388 3.510 11.292 0.000 32.754 46.524

bsmtunfsf 32.0698 1.997 16.059 0.000 28.153 35.987

yearbuilt 415.4386 39.110 10.622 0.000 338.728 492.149

==============================================================================

Omnibus: 204.301 Durbin-Watson: 1.957

Prob(Omnibus): 0.000 Jarque-Bera (JB): 853.462

Skew: 0.533 Prob(JB): 4.71e-186

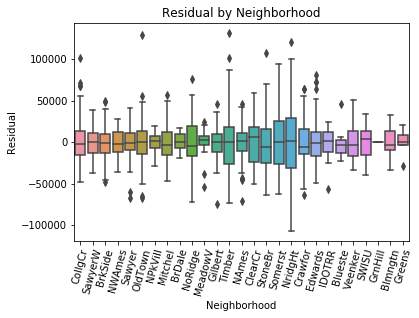
Kurtosis: 6.364 Cond. No. 4.61e+05

==============================================================================

The above model was the best one I created. It scored a 41,973 on Kaggle. I used neighborhood, qualityindex, totalsqftcalc, totalbsmtsf, garagearea, bsmtunfsf and yearbuilt. It was also the same model I did the log transformation on with the response variable. Overall, this model had the best R-Squared, AIC and BIC scores.

**Neighborhood Accuracy**

The neighborhood accuracy was the most challenging and rewarding part of the assignment. The python code I used allowed me to map the neighborhoods to indicator variables. This was done after looking at the boxplot of residuals below. According to the box plot NridgHt,Crawfor and StoneBr were some of the most overpredicted markets. Timber was one of the most underpredicted markets. OldTown and Sawyer appear to have some of the better fits by residual.



Below is the actual and estimated mean price per square foot for each neighborhood that I was able to get through python. I grouped the neighborhoods by three groups and went in descending order based on actual price per sq. foot. Below is the output table. I also did a plot of actual vs. predicted below to help with a visual of the data.

neighborhood actual\_ppsf predicted\_ppsf Neighborhood\_Group

0 GrnHill 123.318386 123.318386 1

1 Blmngtn 118.892612 117.839159 1

2 NridgHt 116.384053 119.373685 1

3 Somerst 110.542232 111.702139 1

4 StoneBr 103.866523 105.737218 1

5 Timber 103.230292 103.698200 1

6 Gilbert 101.491760 100.569597 1

7 CollgCr 99.736694 99.986461 1

8 NoRidge 97.265357 97.242276 1

9 Crawfor 96.726724 96.512449 2

10 Blueste 95.722814 95.435617 2

11 SawyerW 92.704506 92.312512 2

12 Greens 91.696116 91.476015 2

13 BrkSide 89.799890 90.008079 2

14 Veenker 88.186158 88.504960 2

15 Mitchel 86.356609 86.364205 2

16 IDOTRR 85.426961 83.905890 2

17 OldTown 85.169046 83.848181 2

18 ClearCr 85.027752 84.657259 3

19 NWAmes 83.790941 83.103876 3

20 NPkVill 82.931444 82.055189 3

21 NAmes 82.390030 81.178871 3

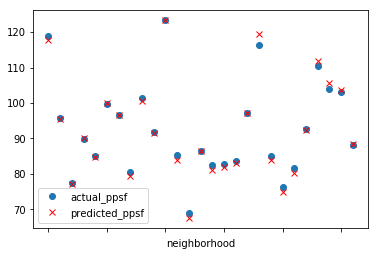
22 Sawyer 81.652367 80.182541 3

23 Edwards 80.688669 79.396191 3

24 BrDale 77.510648 77.060749 3

25 SWISU 76.376106 75.010812 3

26 MeadowV 68.985885 67.620480 3



In order to get a clear idea of how my groupings affected the model, I refit the new variables against the response variable and compared the results below. My group 1 (highest cost per square foot0 ended up being selected as the reference. Overall the AIC and BIC stayed consistent, but the R-squared score decreased and was not good.

OLS Regression Results

==============================================================================

Dep. Variable: Y R-squared: 0.313

Model: OLS Adj. R-squared: 0.313

Method: Least Squares F-statistic: 374.7

Date: Thu, 14 Feb 2019 Prob (F-statistic): 8.74e-135

Time: 01:51:37 Log-Likelihood: -20278.

No. Observations: 1645 AIC: 4.056e+04

Df Residuals: 1642 BIC: 4.058e+04

Df Model: 2

Covariance Type: nonrobust

===========================================================================================

coef std err t P>|t| [0.025 0.975]

-------------------------------------------------------------------------------------------

Intercept 2.284e+05 2444.625 93.419 0.000 2.24e+05 2.33e+05

Neighborhood\_Group[T.2] -7.531e+04 3473.028 -21.686 0.000 -8.21e+04 -6.85e+04

Neighborhood\_Group[T.3] -8.333e+04 3247.329 -25.661 0.000 -8.97e+04 -7.7e+04

==============================================================================

Omnibus: 379.662 Durbin-Watson: 1.984

Prob(Omnibus): 0.000 Jarque-Bera (JB): 947.769

Skew: 1.234 Prob(JB): 1.57e-206

Kurtosis: 5.781 Cond. No. 3.91

==============================================================================

**Section 2. Model Comparison of Y versus log(y)**

I went back and used my best performing Kaggle model, which was number five for the log comparison section. I will re-copy the table down below here for reference along with the log transformed one. The corresponding VIF scores are also right below the print outs. The VIF scores were the same for both versions of the model and nothing was significant enough to justify dropping any attributes. Overall, the log transformed model had the highest R-squared score but also had high AIC and BIC scores. I was not able to test that model on Kaggle and get a proper score to compare to the one that I got for the non-log version. Overall, I believe that the log transformation does a good job of normalizing the variables as witnessed by the improved p-values. Due to the high AIC and BIC scores of the log transformed model, I would have to keep the original as the better fitting model. After seeing the results below, I can’t justify doing another log transform to a response variable. It would change what this model is trying to predict as well. Overall, I think that log transforms are best done on large continuous variables that have a significant variance. That could improve the model fit, depending on the situation.

OLS Regression Results (non-log)

==============================================================================

Dep. Variable: saleprice R-squared: 0.896

Model: OLS Adj. R-squared: 0.894

Method: Least Squares F-statistic: 435.3

Date: Thu, 14 Feb 2019 Prob (F-statistic): 0.00

Time: 01:19:46 Log-Likelihood: -18723.

No. Observations: 1645 AIC: 3.751e+04

Df Residuals: 1612 BIC: 3.769e+04

Df Model: 32

Covariance Type: nonrobust

===========================================================================================

coef std err t P>|t| [0.025 0.975]

-------------------------------------------------------------------------------------------

Intercept -8.341e+05 7.85e+04 -10.623 0.000 -9.88e+05 -6.8e+05

neighborhood[T.Blueste] -1.713e+04 1.09e+04 -1.572 0.116 -3.85e+04 4239.347

neighborhood[T.BrDale] -2.471e+04 9000.007 -2.745 0.006 -4.24e+04 -7052.075

neighborhood[T.BrkSide] -1532.4518 8096.964 -0.189 0.850 -1.74e+04 1.43e+04

neighborhood[T.ClearCr] 8014.4751 8662.334 0.925 0.355 -8976.144 2.5e+04

neighborhood[T.CollgCr] 87.4515 7385.889 0.012 0.991 -1.44e+04 1.46e+04

neighborhood[T.Crawfor] 1.561e+04 8040.486 1.942 0.052 -158.060 3.14e+04

neighborhood[T.Edwards] -8651.3826 7681.778 -1.126 0.260 -2.37e+04 6415.938

neighborhood[T.Gilbert] 2292.6064 7583.504 0.302 0.762 -1.26e+04 1.72e+04

neighborhood[T.Greens] -391.6243 1.14e+04 -0.034 0.973 -2.28e+04 2.2e+04

neighborhood[T.GrnHill] 1.171e+05 2.26e+04 5.172 0.000 7.27e+04 1.62e+05

neighborhood[T.IDOTRR] -8670.2039 8490.853 -1.021 0.307 -2.53e+04 7984.067

neighborhood[T.MeadowV] -2.211e+04 8677.187 -2.548 0.011 -3.91e+04 -5088.364

neighborhood[T.Mitchel] -8381.0179 7719.579 -1.086 0.278 -2.35e+04 6760.447

neighborhood[T.NAmes] -1.281e+04 7502.862 -1.708 0.088 -2.75e+04 1903.047

neighborhood[T.NPkVill] -2.497e+04 9031.525 -2.764 0.006 -4.27e+04 -7250.268

neighborhood[T.NWAmes] -1.741e+04 7701.140 -2.261 0.024 -3.25e+04 -2308.925

neighborhood[T.NoRidge] 2.984e+04 8021.701 3.720 0.000 1.41e+04 4.56e+04

neighborhood[T.NridgHt] 4.205e+04 7700.404 5.461 0.000 2.69e+04 5.72e+04

neighborhood[T.OldTown] -1.18e+04 7998.975 -1.475 0.141 -2.75e+04 3894.200

neighborhood[T.SWISU] -1.393e+04 9062.906 -1.537 0.124 -3.17e+04 3845.856

neighborhood[T.Sawyer] -1.182e+04 7707.552 -1.533 0.125 -2.69e+04 3299.222

neighborhood[T.SawyerW] -8603.3201 7619.053 -1.129 0.259 -2.35e+04 6340.971

neighborhood[T.Somerst] 1.368e+04 7582.818 1.804 0.071 -1196.262 2.86e+04

neighborhood[T.StoneBr] 2.836e+04 8440.220 3.360 0.001 1.18e+04 4.49e+04

neighborhood[T.Timber] 1.622e+04 8007.961 2.026 0.043 513.577 3.19e+04

neighborhood[T.Veenker] 8500.2699 9774.036 0.870 0.385 -1.07e+04 2.77e+04

qualityindex 1677.5519 67.639 24.801 0.000 1544.882 1810.222

totalsqftcalc 55.5622 1.530 36.320 0.000 52.562 58.563

totalbsmtsf -8.2319 2.549 -3.230 0.001 -13.231 -3.233

garagearea 39.6388 3.510 11.292 0.000 32.754 46.524

bsmtunfsf 32.0698 1.997 16.059 0.000 28.153 35.987

yearbuilt 415.4386 39.110 10.622 0.000 338.728 492.149

==============================================================================

Omnibus: 204.301 Durbin-Watson: 1.957

Prob(Omnibus): 0.000 Jarque-Bera (JB): 853.462

Skew: 0.533 Prob(JB): 4.71e-186

Kurtosis: 6.364 Cond. No. 4.61e+05

**This is the VIF for Model 5 non-log**

Intercept 22053.844215

neighborhood[T.Blueste] 1.800078

neighborhood[T.BrDale] 2.963730

neighborhood[T.BrkSide] 9.946482

neighborhood[T.ClearCr] 3.542099

neighborhood[T.CollgCr] 16.752300

neighborhood[T.Crawfor] 8.777780

neighborhood[T.Edwards] 14.713771

neighborhood[T.Gilbert] 10.193750

neighborhood[T.Greens] 1.694636

neighborhood[T.GrnHill] 1.114889

neighborhood[T.IDOTRR] 5.670744

neighborhood[T.MeadowV] 3.554256

neighborhood[T.Mitchel] 8.210373

neighborhood[T.NAmes] 27.629309

neighborhood[T.NPkVill] 2.810689

neighborhood[T.NWAmes] 9.115227

neighborhood[T.NoRidge] 5.992755

neighborhood[T.NridgHt] 7.932105

neighborhood[T.OldTown] 19.082966

neighborhood[T.SWISU] 3.529132

neighborhood[T.Sawyer] 10.876325

neighborhood[T.SawyerW] 9.036756

neighborhood[T.Somerst] 9.404371

neighborhood[T.StoneBr] 3.663966

neighborhood[T.Timber] 5.044072

neighborhood[T.Veenker] 2.270085

qualityindex 1.305557

totalsqftcalc 3.637908

totalbsmtsf 1.917187

grlivarea 3.053692

garagearea 1.765713

yearbuilt 4.740209

OLS Regression Results Log Version

==============================================================================

Dep. Variable: log\_saleprice R-squared: 0.918

Model: OLS Adj. R-squared: 0.917

Method: Least Squares F-statistic: 565.4

Date: Thu, 14 Feb 2019 Prob (F-statistic): 0.00

Time: 02:31:08 Log-Likelihood: 1424.0

No. Observations: 1645 AIC: -2782.

Df Residuals: 1612 BIC: -2604.

Df Model: 32

Covariance Type: nonrobust

===========================================================================================

coef std err t P>|t| [0.025 0.975]

-------------------------------------------------------------------------------------------

Intercept 5.3755 0.377 14.274 0.000 4.637 6.114

neighborhood[T.Blueste] -0.1215 0.052 -2.324 0.020 -0.224 -0.019

neighborhood[T.BrDale] -0.2343 0.043 -5.427 0.000 -0.319 -0.150

neighborhood[T.BrkSide] -0.0546 0.039 -1.406 0.160 -0.131 0.022

neighborhood[T.ClearCr] 0.0473 0.042 1.138 0.255 -0.034 0.129

neighborhood[T.CollgCr] -0.0113 0.035 -0.320 0.749 -0.081 0.058

neighborhood[T.Crawfor] 0.0833 0.039 2.159 0.031 0.008 0.159

neighborhood[T.Edwards] -0.0958 0.037 -2.600 0.009 -0.168 -0.024

neighborhood[T.Gilbert] 0.0225 0.036 0.620 0.535 -0.049 0.094

neighborhood[T.Greens] 0.0169 0.055 0.309 0.757 -0.090 0.124

neighborhood[T.GrnHill] 0.4632 0.109 4.264 0.000 0.250 0.676

neighborhood[T.IDOTRR] -0.1277 0.041 -3.136 0.002 -0.208 -0.048

neighborhood[T.MeadowV] -0.2588 0.042 -6.218 0.000 -0.340 -0.177

neighborhood[T.Mitchel] -0.0417 0.037 -1.126 0.260 -0.114 0.031

neighborhood[T.NAmes] -0.0674 0.036 -1.873 0.061 -0.138 0.003

neighborhood[T.NPkVill] -0.1414 0.043 -3.264 0.001 -0.226 -0.056

neighborhood[T.NWAmes] -0.0779 0.037 -2.109 0.035 -0.150 -0.005

neighborhood[T.NoRidge] 0.0254 0.038 0.660 0.510 -0.050 0.101

neighborhood[T.NridgHt] 0.0661 0.037 1.789 0.074 -0.006 0.139

neighborhood[T.OldTown] -0.1186 0.038 -3.091 0.002 -0.194 -0.043

neighborhood[T.SWISU] -0.0756 0.043 -1.740 0.082 -0.161 0.010

neighborhood[T.Sawyer] -0.0704 0.037 -1.904 0.057 -0.143 0.002

neighborhood[T.SawyerW] -0.0502 0.037 -1.373 0.170 -0.122 0.022

neighborhood[T.Somerst] 0.0370 0.036 1.018 0.309 -0.034 0.108

neighborhood[T.StoneBr] 0.0585 0.040 1.444 0.149 -0.021 0.138

neighborhood[T.Timber] 0.0368 0.038 0.959 0.338 -0.039 0.112

neighborhood[T.Veenker] 0.0055 0.047 0.118 0.906 -0.086 0.097

qualityindex 0.0106 0.000 32.798 0.000 0.010 0.011

totalsqftcalc 0.0003 7.34e-06 40.605 0.000 0.000 0.000

totalbsmtsf -6.66e-05 1.22e-05 -5.448 0.000 -9.06e-05 -4.26e-05

garagearea 0.0002 1.68e-05 13.028 0.000 0.000 0.000

bsmtunfsf 0.0002 9.58e-06 19.992 0.000 0.000 0.000

yearbuilt 0.0028 0.000 15.136 0.000 0.002 0.003

==============================================================================

Omnibus: 71.332 Durbin-Watson: 1.969

Prob(Omnibus): 0.000 Jarque-Bera (JB): 197.438

Skew: -0.158 Prob(JB): 1.34e-43

Kurtosis: 4.668 Cond. No. 4.61e+05

==============================================================================

**This is the VIF for Model 5 (Log-Transformed Model)**

Intercept 22053.844215

neighborhood[T.Blueste] 1.800078

neighborhood[T.BrDale] 2.963730

neighborhood[T.BrkSide] 9.946482

neighborhood[T.ClearCr] 3.542099

neighborhood[T.CollgCr] 16.752300

neighborhood[T.Crawfor] 8.777780

neighborhood[T.Edwards] 14.713771

neighborhood[T.Gilbert] 10.193750

neighborhood[T.Greens] 1.694636

neighborhood[T.GrnHill] 1.114889

neighborhood[T.IDOTRR] 5.670744

neighborhood[T.MeadowV] 3.554256

neighborhood[T.Mitchel] 8.210373

neighborhood[T.NAmes] 27.629309

neighborhood[T.NPkVill] 2.810689

neighborhood[T.NWAmes] 9.115227

neighborhood[T.NoRidge] 5.992755

neighborhood[T.NridgHt] 7.932105

neighborhood[T.OldTown] 19.082966

neighborhood[T.SWISU] 3.529132

neighborhood[T.Sawyer] 10.876325

neighborhood[T.SawyerW] 9.036756

neighborhood[T.Somerst] 9.404371

neighborhood[T.StoneBr] 3.663966

neighborhood[T.Timber] 5.044072

neighborhood[T.Veenker] 2.270085

qualityindex 1.305557

totalsqftcalc 3.720253

totalbsmtsf 3.625469

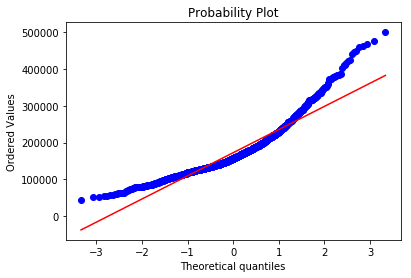
garagearea 1.765713

bsmtunfsf 2.388449

yearbuilt 4.740209

Goodness of Fit

To check the goodness of fit, I built the QQ plot below. The data looks to be normal distributed with a skewness to the upper half, as the residual line starts to elevate from the regression line. I don’t see enough to alarm me that a possible change would need to be made.



**Section 3. Select Models**

I used a multitude of criteria to select the best model. Overall, I relied on the Kaggle testing and scoring tool as my main means of section. The best score was received by Model 5 which was a score of 41973.798. It also had some of the best AIC/BIC scores and also the best R-squared score. I attached a table below with summary stats from the OLS outputs.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 | Model 4 | **Model 5** |
| R sq. | 0.808 | 0.802 | 0.802 | 0.794 | **0.896** |
| AIC | 4.85E+04 | 4.85E+04 | 3.85E+04 | 3.86E+04 | **3.75E+04** |
| BIC | 4.88E+04 | 4.87E+04 | 3.86E+04 | 3.86E+04 | **3.77E+04** |

**Section 4. Model Formula**

For the model I choose some of the variables will have to be referred back to the coefficient table for the output. For example ,which neighborhood a person lived in. I used the OLS output for model 5 to get the formula based on coefficients. First, I had to start with the intercept coefficient and work down from there.

p\_salesprice = -834100 + or - the neighborhood coefficient value multiplied + (1677.5519 \* qualityindex) + (55.5622\*totalsqftcalc) – (8.2319\*totalbsmtsf) + (39.6388\*garagearea) + (32.0698\*bsmtunfsf) + (415.4386\*yearbuilt)

**Section 5. Scored Data File**

This was submitted to canvas with extra files.

**Conclusion**

This exercise was an excellent opportunity to blend together all of the techniques learned so far in the class to build a successful model. There was significant improvement from the first assignment to the most recent one in model predictive capabilities. By using different techniques like a correlation matrix or other tools from sklearn, I was able to build a more accurate predictive model. A seven variable model called Model 5 ended up being the best choice for this project. The continous variables seemed to be the most accurate predictors for the model with the the help of some ordinal values. A good data cleaning up front helped to make sure the model was able to perform as well. In this case, I got rid of the erroneous zoning attributes that didn’t make sense for the model along with outliars in lotarea, salesprice, sale condition and totalsqftcalc. I would be curious to see what my model what do in real life if given the change to apply the same model to recently sold houses today in that market. Overall, this model should be accurate based on the train and test data. I used a QQ plot and other visuals as well to verify the results and eliminate outliers.