Assignment 2 – Alan Kessler, PREDICT 410, Section 57

Introduction

The purpose of the assignment is to analyze property characteristics to predict the sale price of homes in Ames, Iowa. It is a continuation of the first assignment with the opportunity to fit more complex models to generate stronger predictions. A variety of multivariate linear regression models are evaluated using variable selection techniques. Several models are considered, and the final model is chosen based on model fit statistics and diagnostic plots as well as performance on the Kaggle.com leaderboard. The final model considers 18 engineered features prior to categorical variable encoding.

Bonus

For bonus points, I will use a train-test split to validate that one of the models generalizes to new data well.

To try something new, I will use five-fold cross-validation as the basis for a forward selection algorithm to build my final model.

Modeling & More

For the purpose of comparison, the final model from the previous assignment is used as a baseline. It considers **OverallQual** as a grouped categorical variable and **GrLivArea** as a continuous variable. **Neighborhood** is not included in this model, but as Figure 1 shows, the baseline model fits some neighborhoods well (e.g. "Mitchel" and "Sawyer") but under-predicts for others such as "NridgHt" and over-predicts for "OldTown".

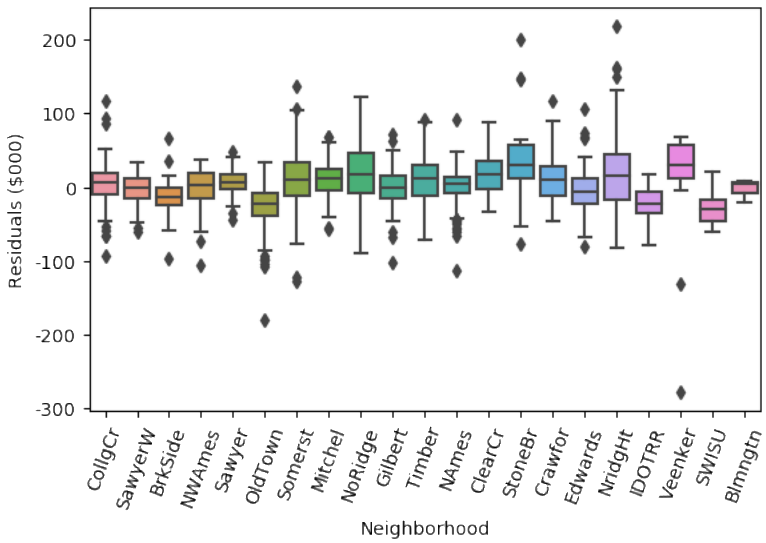

Figure 1 – Baseline Model Residuals by Neighborhood

Figure 2 shows the average and predicted sales price per square foot by **Neighborhood** for the baseline model. The differences in the bars indicate that there is some variance that could be explained by using **Neighborhood** as a predictor.

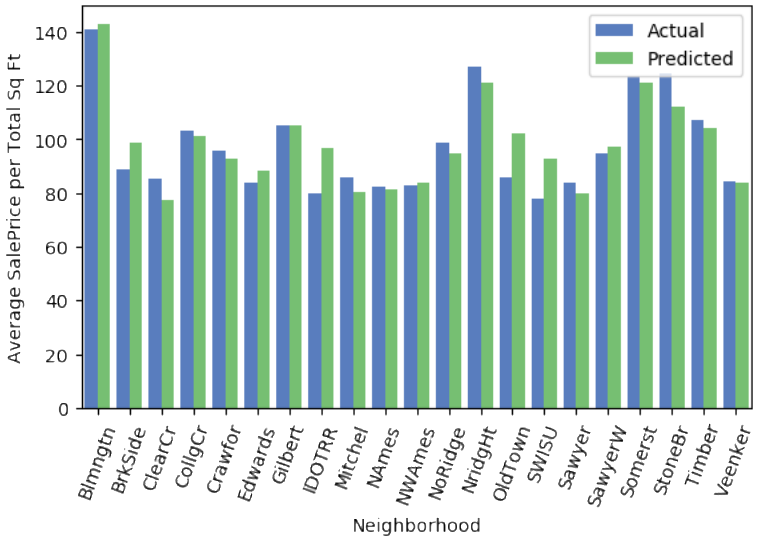

Figure 2 – Actual and Predicted Sale Price per Square Foot by Neighborhood

Model 1 – Baseline with Neighborhood Cluster

The assignment requires grouping **Neighborhood** into three to six groups. This is accomplished by using k-means clustering to group neighborhoods by average actual sale price per square foot. The algorithm's goal is to reduce the distance with in clusters and maximize the distance between clusters. A total of three clusters are chosen to ensure each is sufficiently populated to use in a model reliably.

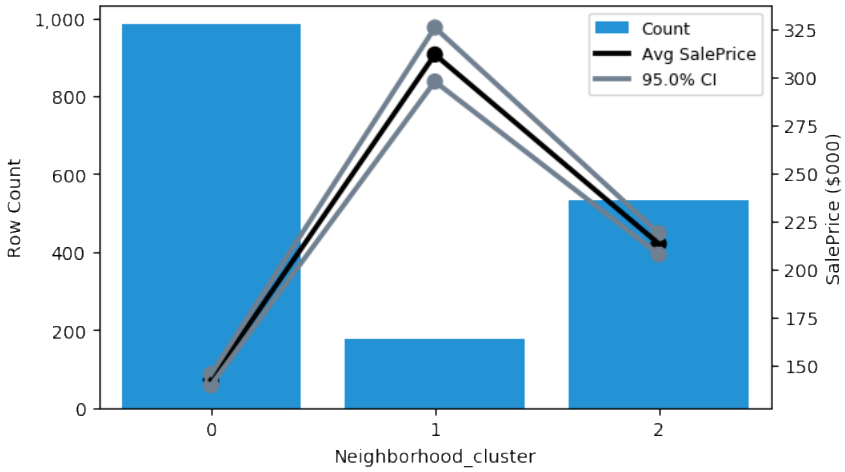

Figure 3 – Average Sale Price by Neighborhood Cluster

Figure 3 illustrates the difference in average sale price by cluster. The group labels are unrelated to the cost of the homes. The results show that there are large differences in sale price by cluster. This makes the newly created variable potentially useful in a predictive model.

The baseline model is compared to a new model where **Neighborhood_cluster** is added. This new model is labeled **Model 1**. The reference levels for the model's categorical variables are selected as the level with the most observations. This allows for significance of the other levels

to be easily interpreted as a comparison to the most common level. Comparing against a rare level would make the comparison difficult because the base may not be significant itself.

Term	Coefficient	Std Error	t	P>|t|
Intercept	58,416.88	2,951.52	19.792	0.000
GrLivArea	61.7511	2.12	29.133	0.000
OverallQual_grp_grp_02	-51,636.45	10,927.94	-4.725	0.000
OverallQual_grp_grp_03	-41,549.82	6,786.06	-6.123	0.000
OverallQual_grp_grp_04	-21,861.24	3,115.56	-7.017	0.000
OverallQual_grp_grp_06	8,370.15	2,262.71	3.699	0.000
OverallQual_grp_grp_07	25,559.94	2,787.19	9.171	0.000
OverallQual_grp_grp_08	69,049.96	3,715.02	18.587	0.000
OverallQual_grp_grp_09	150,684.17	5,638.80	26.723	0.000
OverallQual_grp_grp_10	214,134.46	8,970.02	23.872	0.000
Neighborhood_cluster_1	47,127.36	3,595.05	13.109	0.000
Neighborhood_cluster_2	20,969.01	2,097.72	9.996	0.000

Figure 4 – Model 1 Terms

Figure 4 shows that all terms in **Model 1** are both statistically significant and match the intuitive relationships where larger higher-quality homes are predicted to cost more.

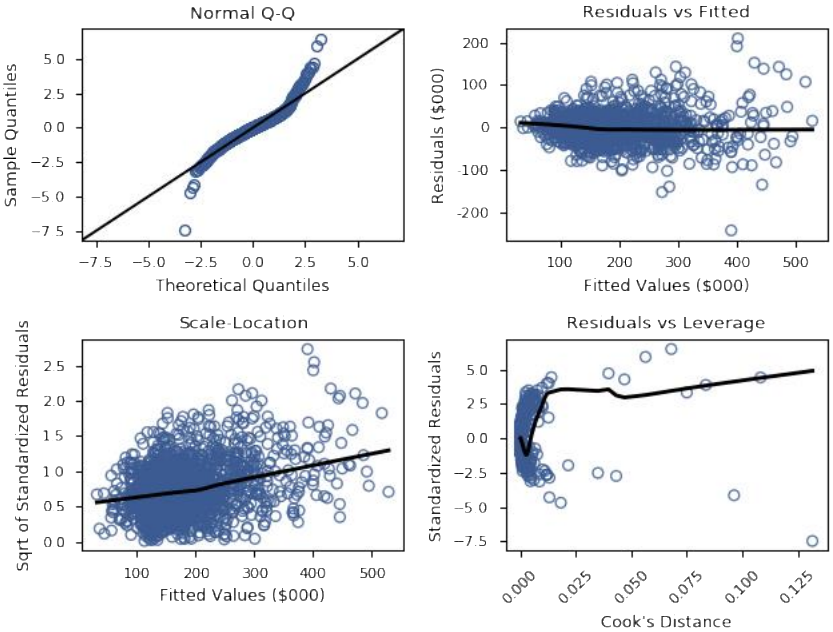

Figure 5 – Model 1 Diagnostic Plots

The diagnostic plots shown in Figure 5 indicate that assumptions underlying linear regression are not adequately met. The normal Q-Q plot shows that residuals are heavy-tailed. Residuals appear unbiased but increase in variance as the fitted value increases. These assumptions should continue to be observed as more terms are added to the model.

Model	R-Squared	Adj R-Squared	AIC
Assignment 1	0.8205	0.8196	40,209.21
Model 1	0.8391	0.8380	40,028.53

Figure 6 – Model 1 Performance

Figure 6 shows that this model is fits the data better than the baseline by R-Squared, Adjusted R-Squared, and AIC.

Model 2 – Neighborhood Cluster with Provided Features

The assignment provides two features: the product of overall quality and overall condition and the sum of livable square footage in the home. This model, including the neighborhood clusters is labeled **Model 2**.

Term	Coefficient	Std Error	t	P>|t|
Intercept	-24,632.91	3,396.09	-7.253	0.000
qualityindex	1,870.38	96.727	19.337	0.000
totalsqftcalc	60.7466	1.263	48.096	0.000
Neighborhood_cluster_1	103,908.14	2,953.43	35.182	0.000
Neighborhood_cluster_2	38,527.28	1,881.82	20.473	0.000

Figure 7 – Model 2 Terms

Figure 7 shows the terms in **Model 2**. All of the terms in the model are statistically significant.

Figure 8 shows the diagnostic plots for **Model 2**. The same issues related to the regression assumptions from **Model 1** persist. The new variables did reduce the heaviness of the left tail of residuals, however, the right tail still does not match the normal distribution.

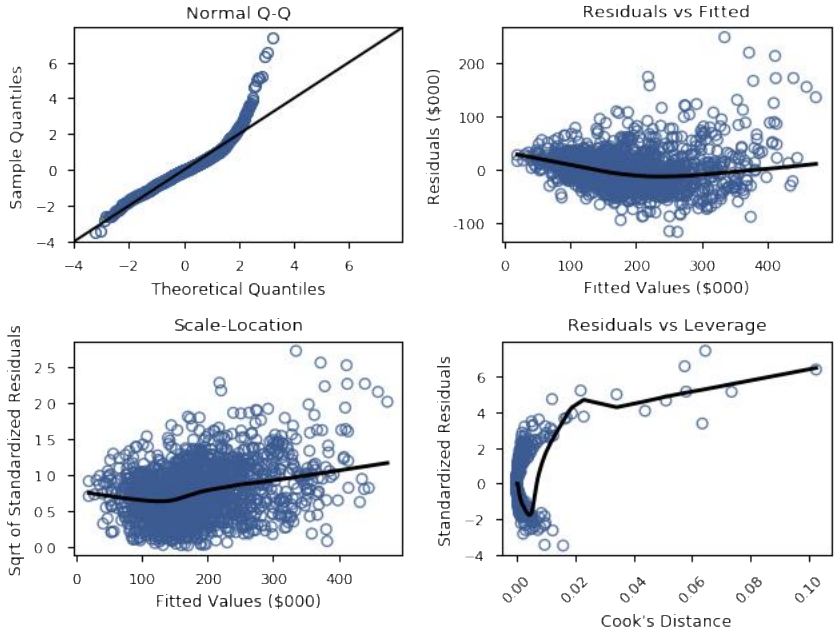

Figure 8 – Model 2 Diagnostic Plots

Figure 8 shows the model performance, indicating that the model is a worse fit relative to **Model 1**.

Model	R-Squared	Adj R-Squared	AIC
Assignment 1	0.8205	0.8196	40,209.21
Model 1	0.8391	0.838	40,028.53
Model 2	0.8271	0.8266	40,136.49

Figure 9 – Model 2 Performance

Continuous Feature Creation

Looking to additional variables to consider, elements of the first assignment analysis are repeated. First, variables are split into two groups: continuous and categorical. Next, missing values are imputed. For continuous variables, missing values are replaced with the mean for **LotFrontage** and zero otherwise. For categorical variables, missing values are imputed with "None", zero, or the mode level depending on the nature of the variable. Outliers are removed but the lot area outlier is capped instead of being removed. This makes a comparison to models 1 and 2 technically different but the impact should be minimal especially as more terms are added to the model.

Three additional features are created:

- **HomeAge_flr** = **YrSold** - **YearBuilt** (minimum 1950)
- **LotArea_cap** – Capped at 20,000 square feet
- **OutdoorSF** – The sum of all porch and deck square footage

Model 3 – Additional Continuous Variables

These newly created features along with the variables in **Model 1** and other continuous predictors are included as **Model 3**. This more complicated model still shows significant terms as shown in Figure 10. Figure 11 shows that the same patterns seen in the diagnostic plots for **Model 1** persist for **Model 3**. Continuous features not found statistically significant were not included in the model.

Term	Coefficient	Std Error	t	P>|t|
Intercept	43,233.23	3,856.297	11.211	0.000
GarageArea	23.7443	4.101	5.790	0.000
LotArea_cap	1.6494	0.220	7.489	0.000
HomeAge_flr	-366.9154	51.557	-7.117	0.000
BsmtFSF	31.9126	1.550	20.585	0.000
MasVnrArea	20.4777	4.429	4.623	0.000
GrLivArea	53.7927	1.896	28.379	0.000
OutdoorSF	11.8978	4.420	2.692	0.007
Neighborhood_cluster_1	34,963.42	3,321.711	10.526	0.000
Neighborhood_cluster_2	12,318.61	2,073.323	5.941	0.000
OverallQual_grp_grp_02	-33,295.57	8,878.310	-3.750	0.000

OverallQual_grp_grp_03	-21,191.51	5,519.577	-3.839	0.000
OverallQual_grp_grp_04	-10,959.76	2,539.883	-4.315	0.000
OverallQual_grp_grp_06	6,115.01	1,856.126	3.295	0.001
OverallQual_grp_grp_07	19,892.78	2,421.356	8.216	0.000
OverallQual_grp_grp_08	49,780.47	3,224.554	15.438	0.000
OverallQual_grp_grp_09	110,486.05	4,882.375	22.630	0.000
OverallQual_grp_grp_10	175,105.50	7,404.855	23.647	0.000

Figure 10 – Model 3 Terms

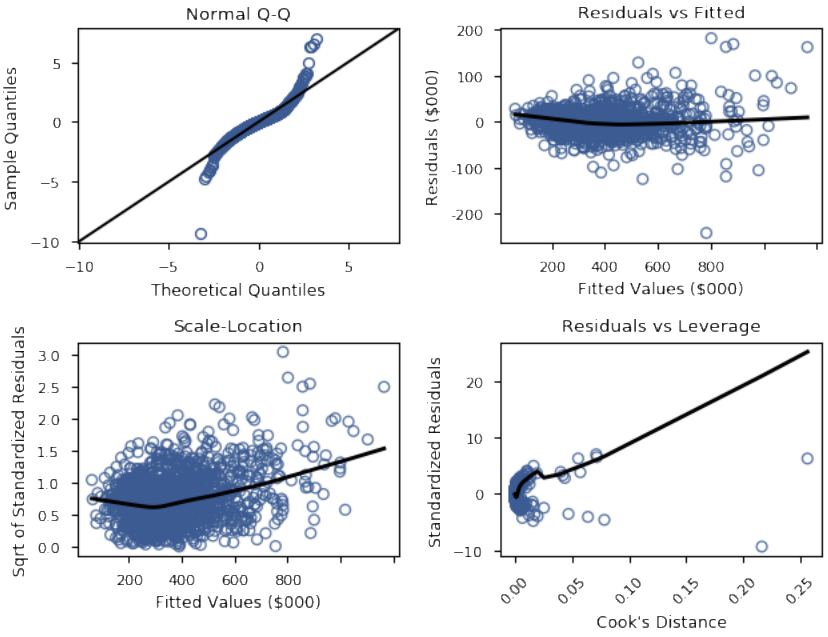

Figure 11 – Model 3 Diagnostic Plots

Model 4 – Log-Transformed Sale Price

The same set of predictors from **Model 3** are used in **Model 4**, however, the target is the natural logarithm of sale price. This model is interpreted differently than **Model 3**. For example, in **Model 3**, a one square foot increase in garage area results in an approximate $23 increase in the estimated sale price. For **Model 4**, a one square foot increase in garage area results in an approximate 0.02% increase in sale price. The coefficients demonstrating this different relationship are shown in Figure 12. Using these terms will generate predictions of the log of the sale price. The model performance for both models is shown in Figure 13. These metrics are not directly comparable because in the case of AIC the likelihood function is changed and in the case of r-squared, the residuals are defined differently because the target has changed. However, there are other ways to compare the two models.

Term	Coefficient	Std Error	t	P>|t|
Intercept	11.3564	0.020	565.527	0.000
GarageArea	0.0002	0.000	8.585	0.000
LotArea_cap	0.0000	0.000	9.548	0.000
HomeAge_flr	-0.0030	0.000	-11.177	0.000
BsmtFSF	0.0001	0.000	18.012	0.000

MasVnrArea	0.0000	0.000	1.076	0.282
GrLivArea	0.0003	0.000	26.613	0.000
OutdoorSF	0.0001	0.000	3.172	0.002
Neighborhood_cluster_1	0.1385	0.017	8.005	0.000
Neighborhood_cluster_2	0.0576	0.011	5.333	0.000
OverallQual_grp_grp_02	-0.6467	0.046	-13.988	0.000
OverallQual_grp_grp_03	-0.3731	0.029	-12.980	0.000
OverallQual_grp_grp_04	-0.1556	0.013	-11.761	0.000
OverallQual_grp_grp_06	0.0639	0.010	6.615	0.000
OverallQual_grp_grp_07	0.1291	0.013	10.240	0.000
OverallQual_grp_grp_08	0.2161	0.017	12.870	0.000
OverallQual_grp_grp_09	0.3328	0.025	13.088	0.000
OverallQual_grp_grp_10	0.3937	0.039	10.209	0.000

Figure 12 – Model 4 Terms

Model	R-Squared	Adj R-Squared	AIC
Assignment 1	0.8205	0.8196	40,209.21
Model 1	0.8391	0.8380	40,028.53
Model 2	0.8271	0.8266	40,136.49
Model 3	0.8988	0.8978	39,352.12
Model 4	0.8898	0.8887	-1,937.39

Figure 13 – Model 3 and 4 Performance

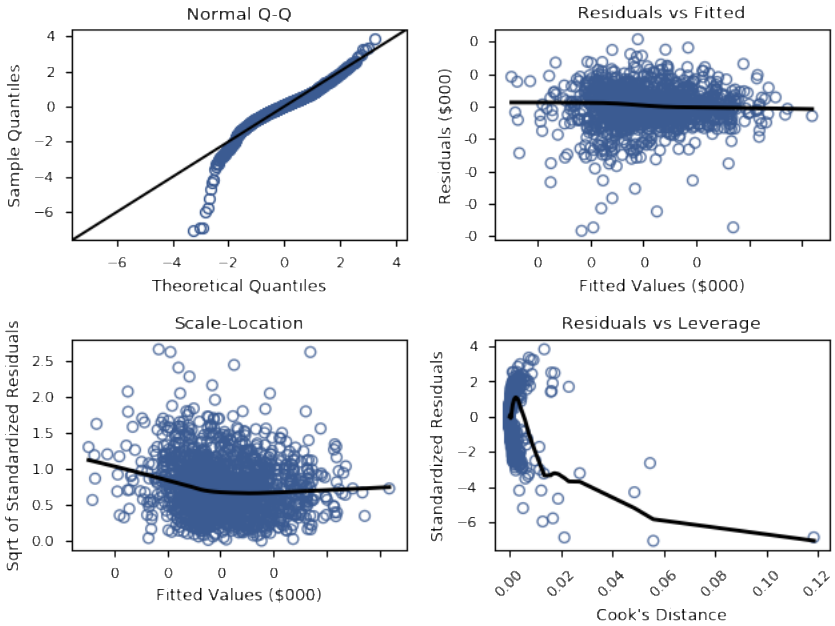

Figure 14 – Model 4 Diagnostic Plots

First, by looking at the diagnostic plots in Figures 11 and 14, the log-transformed model shows that the residuals are closer to normally distributed and do not increase in variance as the fitted value increases, meeting least squares regression assumptions. Another point of comparison is measuring the root mean squared error after undoing the transformation to the predicted values for **Model 4. Model 3** has a RMSE of approximately 26,000 while **Model 4** has a RMSE of approximately 24,000. Overall, both performance and diagnostics suggest that **Model 4** fits the

training data best. In general, a log transformation to the response can improve fit when the response is strictly positive and residuals are skewed.

Some of the predictors in **Model 4** should be log-transformed to linearize the relationship or reduce skewness in partial residuals. Additionally, variables are "floored" to recognize situations where value does not get lower for distances less than a specific value. The following variables are created and used in **Model 5**:

- **LotArea_cap_log** – Log of capped lot area
- **OutdoorSF_flr_log** – Outdoor square footage floored at 200 sq. ft. and log-transformed
- **GarageArea_flr_log** – Garage area floored at 250 sq. ft. and log-transformed
- **BsmtFSF_flr_log** – Basement (finished) area floored at 600 sq. ft. and log-transformed
- **GarageArea_0** – Indicator of no garage area created due to floor
- **BsmtFSF_0** – Indicator of no finished basement space created due to floor

Model 5 improves on the RSME of **Model 4** by approximately 300, a small improvement but with additional categorical variables incorporated, this could lead to greater improvement. All terms are statistically significant as shown in Figure 15.

Term	Coefficient	Std Error	t	P>|t|
Intercept	8.6401	0.147	58.851	0.000
HomeAge_flr	-0.0029	0.000	-10.777	0.000
BsmtUnfSF	0.0001	0.000	5.189	0.000
LotArea_cap_log	0.1164	0.012	9.628	0.000
GrLivArea	0.0003	0.000	27.624	0.000
OutdoorSF_flr_log	0.0429	0.012	3.680	0.000
GarageArea_flr_log	0.0624	0.012	5.016	0.000
BsmtFSF_flr_log	0.1999	0.017	11.876	0.000
GarageArea_0	-0.1000	0.018	-5.606	0.000
BsmtFSF_0	-0.0925	0.009	-10.376	0.000
Neighborhood_cluster_1	0.1328	0.017	7.737	0.000
Neighborhood_cluster_2	0.0573	0.011	5.362	0.000
OverallQual_grp_grp_02	-0.6067	0.046	-13.139	0.000
OverallQual_grp_grp_03	-0.3353	0.029	-11.481	0.000
OverallQual_grp_grp_04	-0.1478	0.013	-11.175	0.000
OverallQual_grp_grp_06	0.0586	0.010	6.113	0.000
OverallQual_grp_grp_07	0.1230	0.013	9.789	0.000
OverallQual_grp_grp_08	0.2102	0.017	12.516	0.000
OverallQual_grp_grp_09	0.3309	0.025	13.152	0.000
OverallQual_grp_grp_10	0.4096	0.038	10.804	0.000

Figure 15 – Model 5 Terms

Diagnostic plots for **Model 5** are shown in Figure 16 and results are similar to that of **Model 4**.

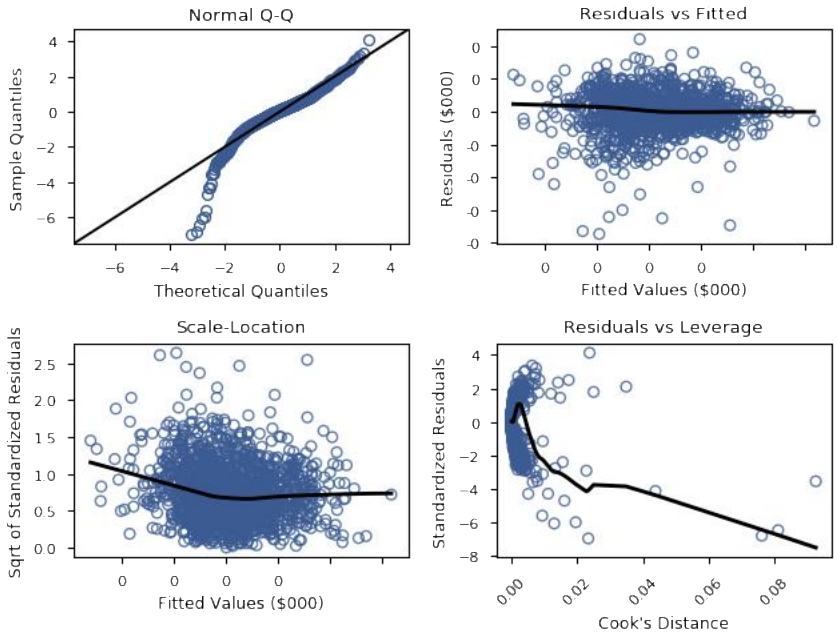

Figure 16 – Model 5 Diagnostic Plots

The variance inflation factors (VIF) are calculated for **Model 5** terms and are shown in Figure 17. The results show that multicollinearity is not a large concern as factors are all under five for non-intercept terms.

Feature	VIF
Intercept	2016.52
HomeAge_flr	3.49
BsmtUnfSF	2.05
LotArea_cap_log	1.42
GrLivArea	2.14
OutdoorSF_flr_log	1.21
GarageArea_flr_log	2.19
BsmtFSF_flr_log	2.00
GarageArea_0	1.21
BsmtFSF_0	1.64

Feature	VIF
Neighborhood_cluster_1	2.58
Neighborhood_cluster_2	2.31
OverallQual_grp_grp_02	1.05
OverallQual_grp_grp_03	1.11
OverallQual_grp_grp_04	1.23
OverallQual_grp_grp_06	1.57
OverallQual_grp_grp_07	2.37
OverallQual_grp_grp_08	2.80
OverallQual_grp_grp_09	1.92
OverallQual_grp_grp_10	1.41

Figure 17 – Model 5 VIFs

Additional Categorical Variables

Prior to additional models being created, a thorough analysis of categorical variables is completed. The result is that levels of the categorical variables are grouped to reduce cardinality. Groupings are done based on the ordinal relationship of levels, their intuitive relationships, and one-way analyses between the variable and sale price. The final output of the analysis is a list of grouped variables to consider for future models.

Model 6 – Automatic Feature Selection

Scikit-Learn's F_regression feature selection algorithm is used to determine the top ten features to use in **Model 6**. This algorithm requires dummy variables be created prior to selection. This has the downside of making results less intuitive when not all levels are selected for a given categorical variable. The terms selected in **Model 6** are shown in Figure 18. In spite of the feature selection algorithm, one of the terms is found to be insignificant. The diagnostic plots are contained in Figure 19 and demonstrate the same characteristics as models 4 and 5. The resulting RMSE for the model is approximately 29,000. This is a clear increase from **Model 5**. The same holds true for other fit statistics such as r-squared. As a result, **Model 6** is not considered for the final model.

Term	Coefficient	Std Error	t	P>|t|
Intercept	8.4438	0.163	51.661	0.000
HomeAge_flr	-0.0057	0.000	-18.007	0.000
LotArea_cap_log	0.1202	0.015	8.052	0.000
GrLivArea	0.0004	0.000	30.959	0.000
GarageArea_flr_log	0.1068	0.015	7.225	0.000
BsmtFSF_flr_log	0.2154	0.019	11.629	0.000
Bath_grp_D. 2.5+	-0.0474	0.014	-3.448	0.001
BsmtFinType1_GLQ	0.0440	0.012	3.697	0.000
BsmtQual_grp_C. Ex	0.0749	0.019	3.960	0.000
ExterQual_Gd	0.0094	0.012	0.805	0.421
Neighborhood_cluster_1	0.1013	0.017	5.915	0.000

Figure 18 – Model 6 Terms

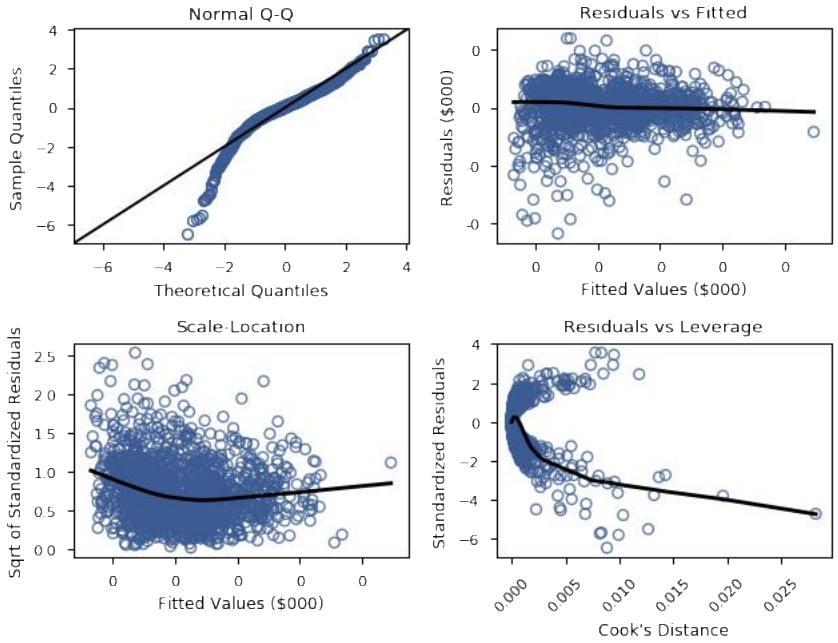

Figure 19 – Model 6 Diagnostic Plots

Bonus – Cross-Validation Train-Test Split

To further test the specification of **Model 5**, the data is split into training and testing subsets made up of 70% and 30% of the training data respectively. The model is fit to the training subset and scored on the testing to determine how well the model generalizes to new data. The RMSE of the model fit to the entire training data set is approximately 23,800 while the RMSE of the out-of-sample validation subset built on the split training data is approximately 24,200. This shows that while as expected, new data has worse performance, the performance is fairly similar.

Model 7 – Bonus Cross-Validation with Forwards Selection

Due to the automatic feature selection shortfalls, a different process is used to select terms. This process uses five-fold cross-validation to minimize the out-of-fold RMSE. The variable that improves the out-of-fold RMSE most is added to the model at each step until it is no longer improved. This in effect mimics the behavior of the SAS GLMSELECT procedure. The RMSE at each step is analyzed to determine where improvement becomes marginal to generate a more parsimonious model. Rather than consider each level of categorical variables separately, the variable as a whole is considered. The grouping discussed previously allows for this interpretation as sample sizes are greater for fewer groups. The chart for this process is shown in Figure 20.

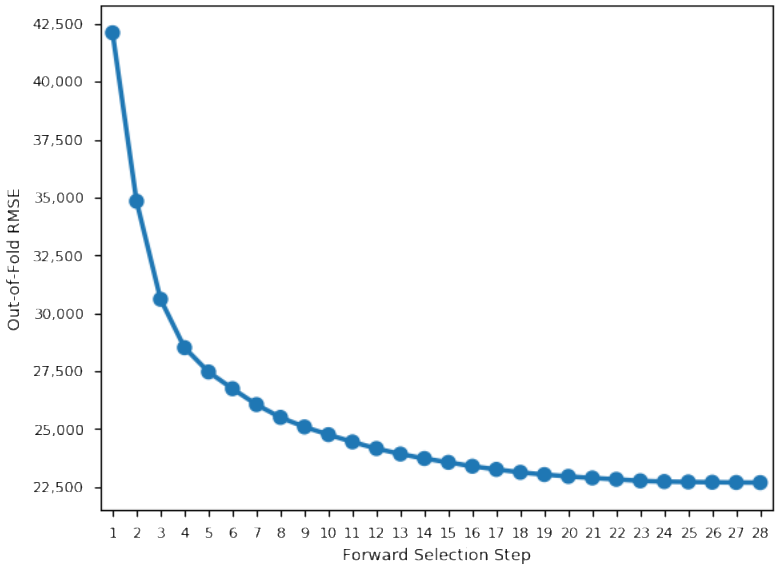

Figure 20 – Cross-Validation Forward Selection Performance

Based on analysis of Figure 20, RMSE starts to improve more slowly after the 18th step. The model with the first 18 terms are included as **Model 7**. The terms selected are included in Figure 21. Individual levels are not all significant but with the exception of basement exposure where small coefficients appear counterintuitive and the indicator for second floors suggesting homes with two floors should be worth less, all coefficients match their intuitive relationships with sale price. Despite adding a multitude of terms, the fit diagnostic plots in Figure 23 show heavy tails in the model residuals. However, figure 22 shows that from a RMSE perspective this new model outperforms the alternatives. Additionally, **Model 7** has an adjusted r-squared of over 0.92.

Term	Coefficient	Std Error	t	P>|t|
Intercept	9.3026	0.130	71.473	0.000
GrLivArea	0.0003	0.000	26.217	0.000
BsmtFSF_flr_log	0.1180	0.014	8.465	0.000
HomeAge_flr	-0.0035	0.000	-11.561	0.000
SecondFlr_ind	-0.0526	0.009	-5.692	0.000
LotArea_cap_log	0.1190	0.010	11.359	0.000
LowQualFin_ind	-0.1056	0.024	-4.471	0.000
GarageArea_flr_log	0.0712	0.010	6.998	0.000
BsmtFSF_0	-0.0537	0.007	-7.590	0.000
OverallQual_grp_grp_02	-0.4015	0.041	-9.908	0.000
OverallQual_grp_grp_03	-0.2265	0.025	-9.039	0.000
OverallQual_grp_grp_04	-0.1123	0.011	-9.924	0.000
OverallQual_grp_grp_06	0.0376	0.008	4.523	0.000
OverallQual_grp_grp_07	0.0861	0.011	7.823	0.000
OverallQual_grp_grp_08	0.1549	0.015	10.641	0.000
OverallQual_grp_grp_09	0.2539	0.023	11.185	0.000
OverallQual_grp_grp_10	0.3089	0.034	9.166	0.000
OverallCond_grp_grp_03	-0.2392	0.021	-11.295	0.000
OverallCond_grp_grp_04	-0.1207	0.017	-7.002	0.000
OverallCond_grp_grp_06	0.0281	0.008	3.342	0.001
OverallCond_grp_grp_07	0.0904	0.009	9.621	0.000
OverallCond_grp_grp_08	0.1066	0.014	7.759	0.000
OverallCond_grp_grp_09	0.1606	0.022	7.245	0.000
Neighborhood_cluster_1	0.1198	0.015	8.056	0.000
Neighborhood_cluster_2	0.0474	0.009	5.163	0.000
BsmtExposure_grp_A. No Bsmt	-0.0212	0.023	-0.918	0.359
BsmtExposure_grp_C. Mn and Av	-0.0002	0.007	-0.022	0.983
BsmtExposure_grp_D. Gd	0.0448	0.012	3.823	0.000
Fireplaces_grp_B. 1	0.0400	0.007	5.960	0.000
Fireplaces_grp_C. 2+	0.0861	0.012	7.143	0.000
SaleCondition_grp_Ab Adj	-0.0849	0.011	-7.386	0.000
SaleCondition_grp_Alloca	0.0410	0.047	0.871	0.384
SaleCondition_grp_Family	-0.1192	0.022	-5.318	0.000
SaleCondition_grp_Partial	0.0349	0.012	2.981	0.003
BsmtQual_grp_A. Po Fa no Bsmt	-0.0394	0.015	-2.589	0.010
BsmtQual_grp_C. Ex	0.0331	0.014	2.380	0.017
Condition1_grp_Arter	-0.0920	0.015	-6.031	0.000
Condition1_grp_Feedr	-0.0395	0.013	-3.108	0.002
Condition1_grp_Pos	0.0051	0.019	0.274	0.784
Condition1_grp_RRe	-0.0888	0.025	-3.531	0.000
Condition1_grp_RRn	-0.0540	0.019	-2.834	0.005
LandContour_grp_HLS	0.0517	0.014	3.573	0.000
Bath_grp_B. 1.5	0.0255	0.010	2.503	0.012
Bath_grp_C. 2	0.0242	0.010	2.406	0.016
Bath_grp_D. 2.5+	0.0226	0.015	1.487	0.137

Figure 21 – Model 7 Terms

Model	RMSE
3	25,979.61
4	24,134.84
5	23,810.17
6	28,996.25
7	19,729.31

Figure 22 – RMSE Comparison Across Models

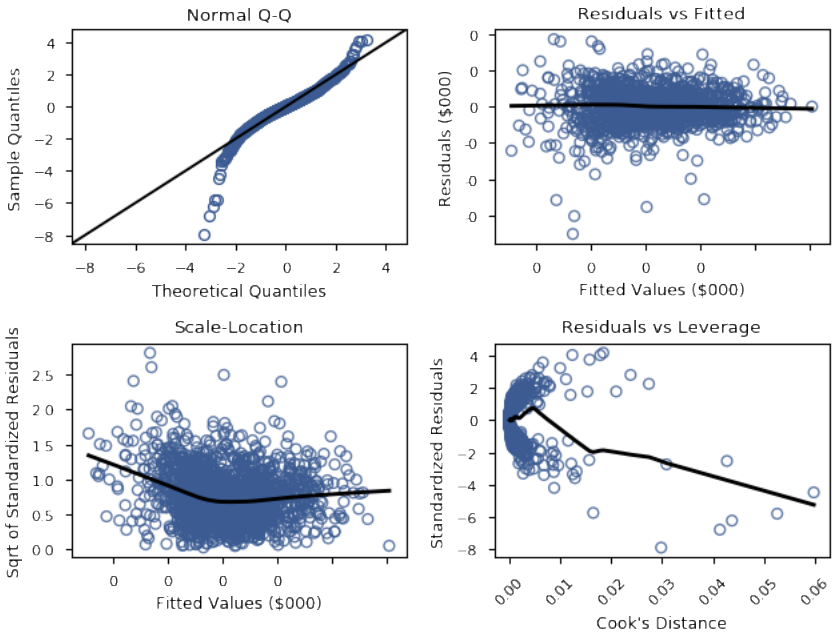

Figure 23 – Model 7 Diagnostic Plots

Model 7 is submitted to Kaggle to assess the fit on out-of-sample data. The result is surprising in that the RMSE is almost 50% worse relative to the score from the first assignment. Given that **Model 7** is specified on cross-validation out-of-fold predictions, the test scores should not have a RMSE much different than the RMSE from step 17 in the forwards selection. Given that RMSE will be heavily influenced by outliers and the residuals of **Model 7** still have heavy tails, it is worth analyzing the residuals more closely.

The model residuals are attached to the training data and explored in more depth. After sorting by the absolute value of the residuals, it is clear that the top six largest errors are partial sales which indicates that the home was not finished when last assessed. Figure 24 shows that this phenomenon generates errors in both directions. It also demonstrates a violation of the least squares regression assumptions. Residual variance depends on a variable included in the model rather than a constant variance.

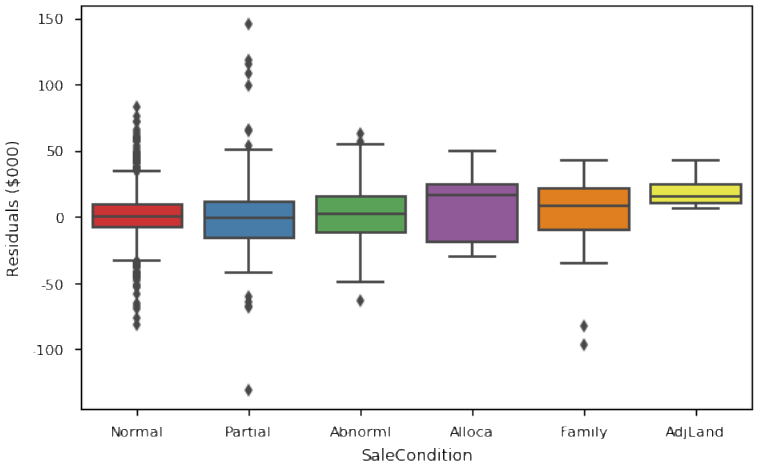

Figure 24 – Model 7 Residuals by Sale Condition

Looking at the outliers more closely, there appears to be a trend that the largest outliers are partial sales where the home was built close to when it was sold but that the home was remodeled after being sold. This could indicate that the home was sold without being finished but with outliers in both directions, it is not possible to know which is the case with the information given. Using these criteria, a low confidence indicator shown in Figure 25 demonstrates that the largest outliers are a part of this 35-observation count group.

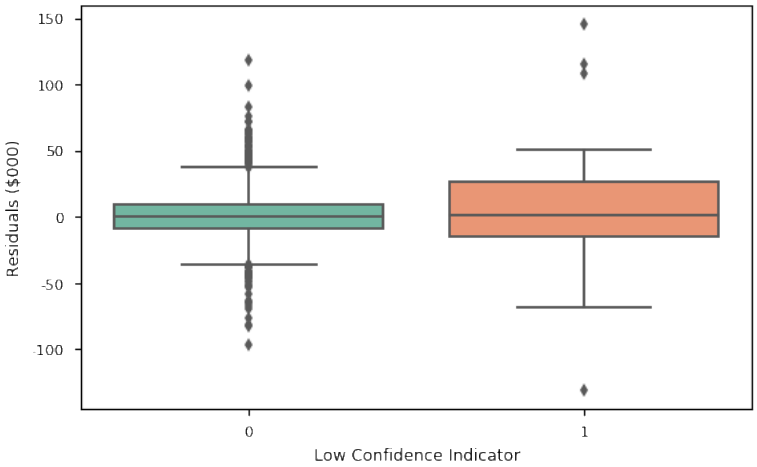

Figure 25 – Model 7 Residuals by Low Confidence Indicator

The data is from the Ames Iowa Assessor's Office and contains characteristics regarding residential properties sold in Ames from 2006 to 2010.

Model 8 - Simple Model for Low Confidence Values

The Kaggle performance dropped when moving from a simple to complex model. This indicates that the assumption of equal variance for least squares regression is violated. There is a defined population with more uncertain results than can be covered by the variables available. Either a Bayesian approach could be considered or a simpler model could be applied to the subpopulation with less certainty to avoid overfitting. This latter approach is similar to what is done in credit modeling when an individual does not have a full credit history.

This model, **Model 8**, is fit to the entire training data but controls for sale condition. In the simple model, the sale price is based entirely on the lot area and neighborhood because the property characteristics cannot be relied upon. **Model 8** would apply to the uncertain low confidence values while **Model 7** would apply otherwise. The terms in **Model 8** are shown in Figure 26, showing the relative simplicity of that model. Overall, the fit as shown in Figure 28 is worse than other models but the diagnostic plots in Figure 27 show that this simple model is resistant to outliers and fit the least squares regression assumptions well.

Term	Coefficient	Std Error	t	P>|t|
Intercept	7.7609	0.189	41.086	0.000
LotArea_cap_log	0.4467	0.021	21.589	0.000
Neighborhood_cluster_1	0.6613	0.025	26.375	0.000
Neighborhood_cluster_2	0.3300	0.015	22.014	0.000

SaleCondition_grp_Ab Adj	-0.1608	0.026	-6.083	0.000
SaleCondition_grp_Alloca	0.0805	0.109	0.739	0.460
SaleCondition_grp_Family	-0.1883	0.053	-3.575	0.000
SaleCondition_grp_Partial	0.0603	0.026	2.279	0.023

Figure 26 – Model 8 Terms

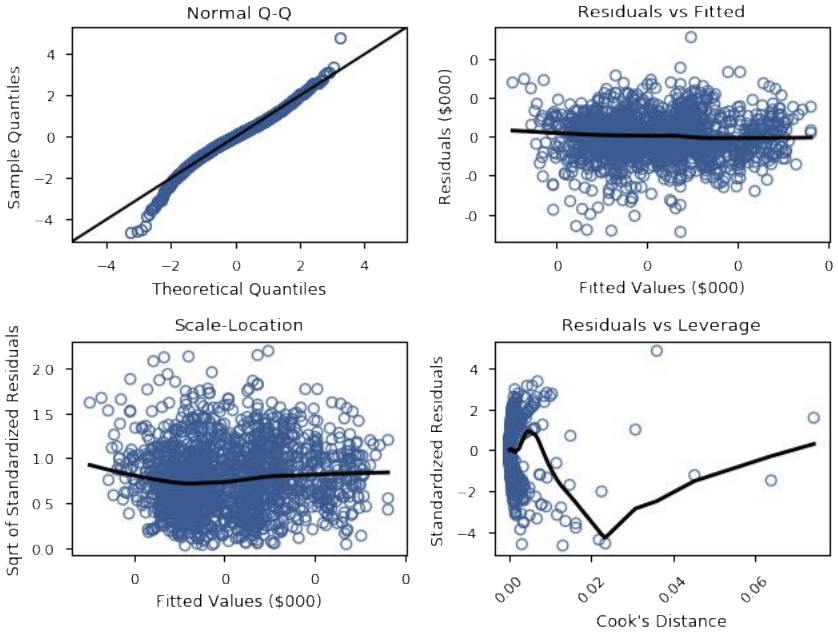

Figure 27 – Model 8 Diagnostic Plots

Model	R-Squared	Adj R-Squared	AIC
Assignment 1	0.8205	0.8196	40,209.21
Model 1	0.8391	0.838	40,028.53
Model 2	0.8271	0.8266	40,136.49
Model 3	0.8988	0.8978	39,352.12
Model 4	0.8898	0.8887	-1,937.39
Model 5	0.8921	0.8909	-1,968.71
Model 6	0.8301	0.8291	-1,215.83
Model 7	0.9266	0.9246	-2,571.71
Model 8	0.5768	0.5751	326.36

Figure 28 – Model Performance Comparison

The combination of **Model 7** and **Model 8** has worse performance overall but is less likely to lend credibility to situations where outliers could occur and impact overall RMSE. The performance on the training data reflecting this point is shown in Figure 29.

R-Squared	Adj R-Squared	RMSE
0.918	0.915	23,396

Figure 29 – Model Combination Training Performance

Select Models

Of the models created, **Model 7** fit the training data best by r-squared, adjusted r-squared, AIC, and RMSE as shown in Figure 30. However, it did not fit the test data on Kaggle well. This was due to overfitting situations where high variance in sale price exists. By combining the model with **Model 8** in those situations, the Kaggle RMSE is reduced to approximately 24,000. Based on that performance and the fit statistics, the combined model is chosen as the final model.

Model	R-Squared	Adj R-Squared	RMSE
Assignment 1	0.821	0.820	40,209
Model 3	0.899	0.898	25,980
Model 4	0.890	0.889	24,135
Model 5	0.892	0.891	23,810
Model 6	0.830	0.829	28,996
Model 7	0.927	0.925	19,729
Model 8	0.577	0.575	53,409
Model 7 & 8	0.918	0.915	23,396

Figure 30 – Grouped Fit Statistics on Training Data

Final Model Formula

The formula representing the final model (combination of **Model 7** and **Model 8**) is shown below. The model uses a log transformation for the response so the scorecard shows that the model result needs to be transformed back to a dollar amount. Additionally, when predictors are log-transformed, a value of one is added prior to the transformation.

The first step is determining which model should be run based on the time between building and selling (assuming mid-year for month built),

If YrSold + MoSold – (YearBuilt + 0.5) is less than or equal to 1 and SaleCondition is Partial and YearRemodel is greater than YearBuilt then run **Model 8**:

p_saleprice =	exp(7.7609			
+	0.4467	*	LotArea_cap_log	(LotArea capped at 20,000 sq. ft. and log-transformed)
+	0.6613	*	Neighborhood_cluster_1	(1 if Neighborhood is Blmngtn, NridgHt, Somerst, or StoneBr otherwise 0)
+	0.3300	*	Neighborhood_cluster_2	(1 if Neighborhood is CollgCr, Crawfor, Gilbert, NoRidge, SawyerW, or Timber otherwise 0)
+	-0.1608	*	SaleCondition_grp_Ab Adj	(1 if SaleCondition is Ab or Adj otherwise 0)
+	0.0805	*	SaleCondition_grp_Alloca	(1 if SaleCondition is Alloca otherwise 0)
+	-0.1883	*	SaleCondition_grp_Family	(1 if SaleCondition is Family otherwise 0)
+	0.0603	*	SaleCondition_grp_Partial)	(1 if SaleCondition is Partial otherwise 0)

Otherwise, run **Model 7**:

p_saleprice =	exp(9.3026			
+	0.0003	*	GrLivArea	(GrLivArea untransformed)
+	0.1180	*	BsmtFSF_flr_log	(Finished basement sq. ft. floored at 600 and log-transformed)
+	-0.0035	*	HomeAge_flr	(YrSold less the maximum of YearBuilt and 1950)
+	-0.0526	*	SecondFlr_ind	(1 for no second floor otherwise 0)
+	0.1190	*	LotArea_cap_log	(LotArea capped at 20,000 sq. ft. and log-transformed)
+	-0.1056	*	LowQualFin_ind	(1 for positve LowQualFinSF otherwise 0)
+	0.0712	*	GarageArea_flr_log	(GarageArea floored at 250 and log-transformed)
+	-0.0537	*	BsmtFSF_0	(1 no finished basement otherwise 0)
+	-0.4015	*	OverallQual_grp_grp_02	(1 if OverallQual is 1 or 2 otherwise 0)
+	-0.2265	*	OverallQual_grp_grp_03	(1 if OverallQual is 3 otherwise 0)
+	-0.1123	*	OverallQual_grp_grp_04	(1 if OverallQual is 4 otherwise 0)
+	0.0376	*	OverallQual_grp_grp_06	(1 if OverallQual is 6 otherwise 0)
+	0.0861	*	OverallQual_grp_grp_07	(1 if OverallQual is 7 otherwise 0)
+	0.1549	*	OverallQual_grp_grp_08	(1 if OverallQual is 8 otherwise 0)
+	0.2539	*	OverallQual_grp_grp_09	(1 if OverallQual is 9 otherwise 0)
+	0.3089	*	OverallQual_grp_grp_10	(1 if OverallQual is 10 otherwise 0)
+	-0.2392	*	OverallCond_grp_grp_03	(1 if OverallCond is 3 or lower otherwise 0)
+	-0.1207	*	OverallCond_grp_grp_04	(1 if OverallCond is 4 otherwise 0)
+	0.0281	*	OverallCond_grp_grp_06	(1 if OverallCond is 6 otherwise 0)
+	0.0904	*	OverallCond_grp_grp_07	(1 if OverallCond is 7 otherwise 0)
+	0.1066	*	OverallCond_grp_grp_08	(1 if OverallCond is 8 otherwise 0)
+	0.1606	*	OverallCond_grp_grp_09	(1 if OverallCond is 9 otherwise 0)
+	0.1198	*	Neighborhood_cluster_1	(1 if Neighborhood is Blmngtn, NridgHt, Somerst, or StoneBr otherwise 0)
+	0.0474	*	Neighborhood_cluster_2	(1 if Neighborhood is CollgCr, Crawfor, Gilbert, NoRidge, SawyerW, or Timber otherwise 0)
+	-0.0212	*	BsmtExposure_grp_A. No Bsmt	(1 if BsmtExposure is missing otherwise 0)
+	-0.0002	*	BsmtExposure_grp_C. Mn and Av	(1 if BsmtExposure is Mn or Av otherwise 0)
+	0.0448	*	BsmtExposure_grp_D. Gd	(1 if BsmtExposure is Gd otherwise 0)
+	0.0400	*	Fireplaces_grp_B. 1	(1 if one fireplace otherwise 0)
+	0.0861	*	Fireplaces_grp_C. 2+	(1 if two or more fireplaces otherwise 0)
+	-0.0849	*	SaleCondition_grp_Ab Adj	(1 if SaleCondition is Ab or Adj otherwise 0)
+	0.0410	*	SaleCondition_grp_Alloca	(1 if SaleCondition is Alloca otherwise 0)
+	-0.1192	*	SaleCondition_grp_Family	(1 if SaleCondition is Family otherwise 0)

+	0.0349	*	SaleCondition_grp_Partial	(1 if SaleCondition is Partial otherwise 0)
+	-0.0394	*	BsmtQual_grp_A. Po Fa no Bsmt	(1 if BsmtQual is missing, Po, Fa otherwise 0)
+	0.0331	*	BsmtQual_grp_C. Ex	(1 if BsmtQual is Ex otherwise 0)
+	-0.0920	*	Condition1_grp_Arter	(1 if Condition1 is Arter otherwise 0)
+	-0.0395	*	Condition1_grp_Feedr	(1 if Condition1 is Feedr otherwise 0)
+	0.0051	*	Condition1_grp_Pos	(1 if Condition1 is PosA or PosN otherwise 0)
+	-0.0888	*	Condition1_grp_RRe	(1 if Condition1 is RRNe or RRAe otherwise 0)
+	-0.0540	*	Condition1_grp_RRn	(1 if Condition 1 is RRNn or RRAn otherwise 0)
+	0.0517	*	LandContour_grp_HLS	(1 if LandContour is HLS otherwise 0)
+	0.0255	*	Bath_grp_B. 1.5	(1 if 1.5 bathrooms above grade otherwise 0)
+	0.0242	*	Bath_grp_C. 2	(1 if 2 bathrooms above grade otherwise 0)
+	0.0226	*	Bath_grp_D. 2.5+)	(1 if 2.5 or more bathrooms above grade otherwise 0)

The formula is quite complex. However, while parsimony is important, this model is a significant improvement over simpler models. In a corporate setting, a model like this would be fairly simple to implement due to the linear relationship. The usage of the model would also dictate how complex of a model the client would be willing to take on. If a marginal increase in model fit results in a competitive advantage, then complexity would be tolerated.

Each of the terms is intuitive to the one-way relationship with sale price except for the second story indicator and minimum/average basement exposure. These variables are likely impacted by correlation. However, their coefficients are small and have minimal impact on results. Using cross-validation to specify the model provides reassurance in the variable selections.

Conclusion

Several different models were fit to Ames, IA housing data adjusted to generalize to testing data. The goal of the analysis was to predict the sale price of the testing data. Ultimately, the best model is a combination of two models including 18 different features of the data. The resulting model fits the training data as well as performing well on the Kaggle leaderboard for the assignment. The most valuable lesson from the assignment is the understanding that when the linear regression assumptions are violated, the RMSE statistic is impacted greatly. In the future, I would consider incorporating Bayesian techniques to account for the varying confidence that I have in each observation. I would also consider using a generalized linear model with a gamma error distribution to match the skew of the errors associated with dollar values as the response. The final model contains only main effects, so interactions could further increase the quality of the fit or reduce the number of features needed.