Assignment 1 – 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. Both simple and multivariate models are evaluated using the Best Subsets variable selection technique. Four models are considered and the final model is chosen based on model fit statistics and diagnostic plots. The final model considers the overall quality of the property and the square footage of the home above ground level.

Bonus

For bonus points, I will explore using target-encoding to improve model fit for one of the models. I will also explore fitting a gamma log-link GLM.

Data Survey

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

The training data includes the target variable and contains 2,039 records while the test data leaves the target blank and contains 726 single-family homes. In addition to the target variable and an index, there is a total of 79 variables containing property characteristics.

The small size of the data set and the specific nature of its collection suggest that a predictive model generated from the data would be limited in use. The target, sale price, could only be predicted with confidence for other properties within Ames during the time period between 2006 and 2010. Depending on the type of building or unique features of the property, the small sample size would result in uncertain estimates of a sale price.

As the training data contains other building types in addition to single-family homes, it should be considered whether those records contribute to the model fit or worsen it. If these types were likely to worsen the fit, they should be excluded from the analysis.

Define the Sample Population

To check the similarity to other property types, the average sale price for each building type is compared in Figure 1.

BldgType	Count	Percent	Avg SalePrice
1Fam	1,699	83.33%	182,715
TwnhsE	156	7.65%	193,906
Duplex	76	3.73%	139,583
Twnhs	74	3.63%	137,550
2fmCon	34	1.67%	125,372

Figure 1 – Average Sale Price by Building Type

Figure 1 demonstrates that single-family homes make up the majority of the training data. It also shows that the average sale price varies by building type. These observations indicate that building types other than single-family homes will add little predictive power and may actually make it more difficult to build regression models that fit the test data well. For example, the coefficient for another predictor may actually vary significantly by building type, requiring a more complicated model with interaction terms.

For this analysis, building types other than single-family homes are dropped from the dataset. The training data with this drop condition is used in the rest of the analysis.

Data Quality Checks

The first data quality checks are performed on the target variable, sale price. No negative values or unintuitive results are found. One particularly high-valued home is present as the maximum. That value is examined more closely in Figure 2.

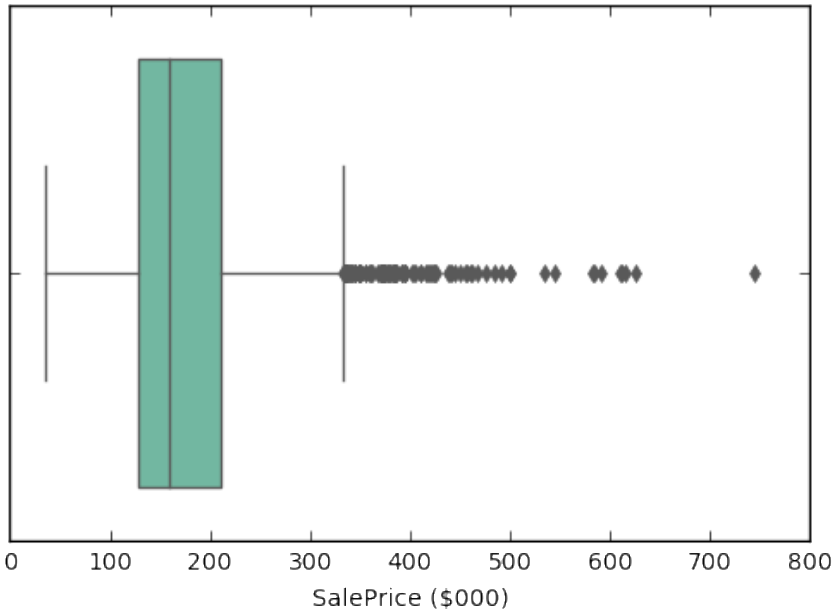

Figure 2 – Box Plot of Sale Price

The value on the far right of the plot represents the high-valued home. Its price is far greater than the typical home's price discussed as the objective of the analysis. As a result, this row is excluded from the training data and subsequent analysis does not use that observation. The box plot also shows that the response is heavily skewed. Special attention should be given to model diagnostics to observe the impact this has on model residuals.

For the predictors, each variable is categorized into one of three groups: variables that are related to distances or areas, variables related to years, and categorical variables including discrete, ordinal, or numeric variables with a small number of unique values.

The first category's variables are examined for missing values and unintuitive results such as negative values. No negative values are found. Missing values are imputed with zero in both the training and the testing data. The exception to this is for **LotFrontage** in which missing values are imputed with the median of the training data for training and testing data. The exception is

made because a value of zero is unintuitive. Not having frontage is not the equivalent of having a small amount of frontage.

For a simple model with only one or two predictors, each variable needs to successfully segment the data well by sale price. If a large majority of observations have the same value for a variable, it is likely not as useful. This results in the following continuous variables warranting further exploratory data analysis (EDA): **LotArea**, **GrLivArea**, **TotalBsmtSF**, and **LotFrontage**. These variables also have an intuitive relationship to home price as larger homes and lots likely cost more.

For variables related to time, **YearBuilt**, **YearRemodel**, and **GarageYrBlt**, no negative values are observed. However, a possible error in **GarageYrBlt** shows a year of 2207 recorded. Additionally, **GarageYrBlt** is missing if the property does not contain a garage. The missing values for **GarageYrBlt** are imputed with zero and the error is imputed with the more likely intended value of 2007. The summary statistics show that the **YearRemodel** data does not have any information prior to 1950, and the garage year built won't have results if the property does not have a garage. For a simple model, **YearBuilt** should be explored further in EDA. The year-related variables are also correlated so choosing one to investigate further will reduce the risk of multicollinearity.

The remaining variables are related to the quality, count, or presence of a specific characteristic. Many of these variables have missing values. In some cases, such as **PoolQual**, the variable is missing when the property does not have a pool. For those variables, missing values are given their own category of “Blank”. In the other cases, the mode level of the training data is given the missing data for both training and testing. A one-way analysis showing the observation count and average sale price for each level of each categorical variable is used to both select variables for further EDA and check for that only allowed values are present. Variables to be investigated further in EDA need to demonstrate strong segmentation in the target and have observations spread across levels rather than a single category. The resulting categorical variables warranting further analysis are: **BedroomAbvGr**, **OverallQual**, **Neighborhood**, **KitchenQual**, and **GarageCars**. Homes with larger garages and more rooms in nicer neighborhoods likely cost more, so there is an intuitive relationship between the variables and sale price as well.

Exploratory Data Analysis – Categorical Variables

For each of the categorical variables, the relationship with the target by means of one-way analysis is considered. Levels with small sample size and large variance in sale price are grouped with nearby levels. This type of analysis and plot is common in actuarial ratemaking models. In many of the figures below, an additional axis is added to the plot to show both the one-way relationship with the target as well as the sample size at each level.

BedroomAbvGr shows a strong relationship with the target for the three most populated levels: 2, 3, and 4 bedrooms. The small sample size of other levels results in unintuitive relationships between the number of bedrooms and sale price. For example, zero-bedroom homes have the highest sale price on average. As a result, the levels are grouped into more statistically credible segments as shown in Figure 3.

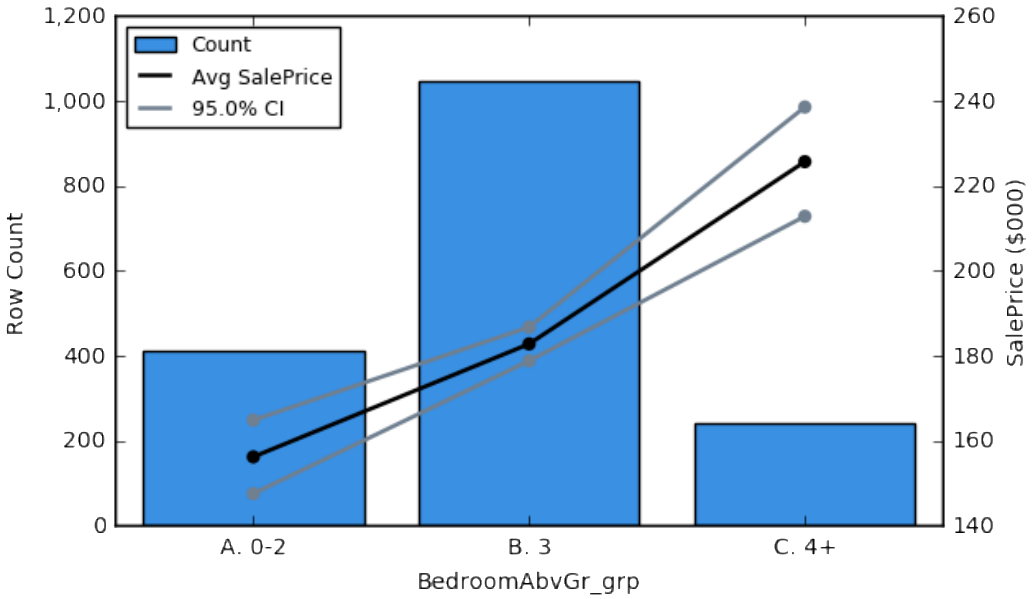

Figure 3 – Grouped BedroomAbvGr Count and Average Sale Price

This variable shows an ability to differentiate by sale price but the variance that required grouping may limit its value. The grouped version of the variable is named **BedroomAbvGr_grp.**

OverallQual has more levels than **BedroomAbvGr** and has a stronger relationship with sale price as shown in Figure 4. Only the lowest quality level, 1, has a sample size worth grouping with the next lowest, 2.

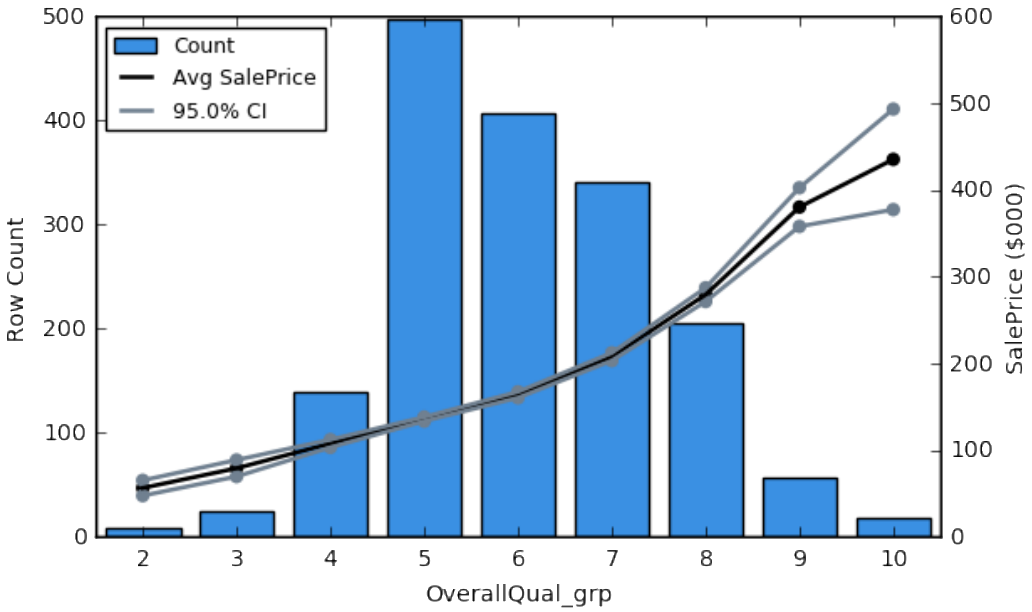

Figure 4 – Grouped OverallQual Count and Average Sale Price

The figure demonstrates a significant relationship between sale price and quality, making the variable as strong candidate for use in a predictive model. Since it is numeric, it is worth using this variable both as numeric and categorical when building a predictive model.

Neighborhood contains 21 unique levels, too many to include in a plot similar to Figure 4. Observations are spread across these neighborhoods and the average sale price varies significantly as well. This make the variable a strong candidate for a predictive model. However, some neighborhoods have very few observations. In cases where the number of observations for a neighborhood is under 2%, the neighborhood is grouped with the neighborhood with the closest average sale price. The resulting table showing the groupings and their relationship to sale price is shown in Figure 5.

Neighborhood	Count	Percent	Avg SalePrice	Std Dev SalePrice
NAmes	276	16.25%	145,189	31,205
CollgCr	176	10.37%	203,288	53,964
OldTown	155	9.13%	124,158	44,805
Edwards	125	7.36%	130,672	45,033
Gilbert	111	6.54%	191,801	28,742
Sawyer and SWISU	108	6.36%	137,793	22,629
Crawfor and ClearCr	99	5.83%	205,220	56,827
NridgHt and StoneBr	96	5.65%	362,671	91,459
NWAmes and Blmngtn	83	4.89%	185,107	34,404
BrkSide	81	4.77%	124,530	34,961
Somerst	78	4.59%	254,550	57,111
SawyerW	71	4.18%	186,940	43,606
Timber and Veenker	69	4.06%	245,880	75,166
Mitchel	63	3.71%	162,761	44,105
IDOTRR	58	3.42%	97,677	32,642
NoRidge	49	2.89%	316,135	75,554

Figure 5 – Grouped Neighborhood Count and Average Sale Price

From the figure, different neighborhoods have varying average sale prices. That makes sense as homes in the same neighborhood are used as “comps” during the appraisal process.

KitchenQual makes sense as a predictor due because it is intuitive that better kitchens would result in a higher sale price. “Fair” and “Poor” quality kitchens have similar sale prices and small sample sizes and as a result are grouped. The final groupings are shown in Figure 6.

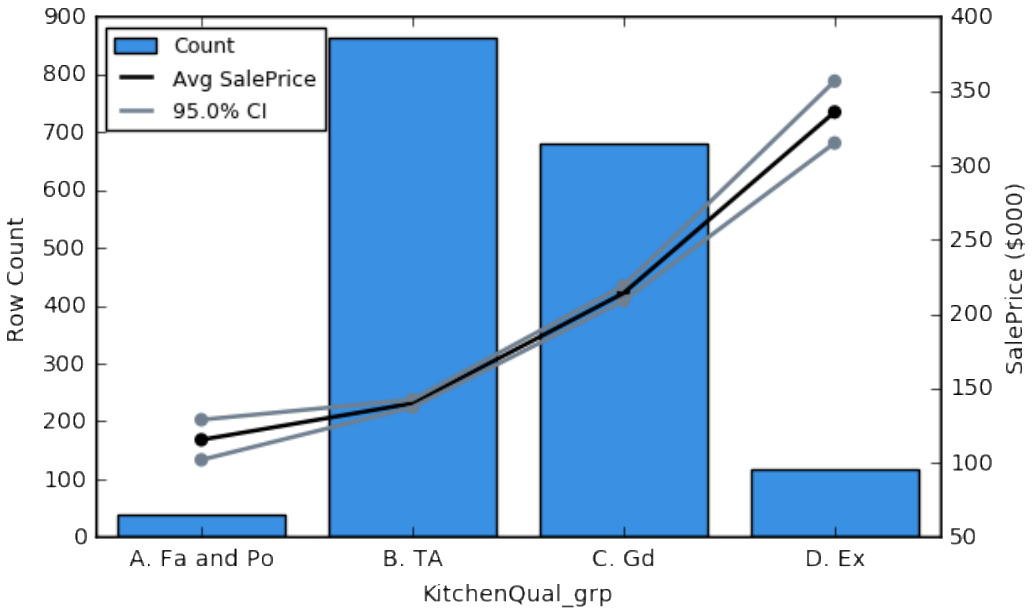

Figure 6 - Grouped KitchenQual Count and Average Sale Price

The figure demonstrates the clear relationship between the variable and sale price although the fact that the majority of the observations are in the two most common levels limit its use in simple models.

GarageCars is technically a discrete variable but is has few possible values allowing it to be treated as categorical. Properties with more than three garage spots are rare and are grouped with properties with three garage spots. The grouped data is shown in Figure 7.

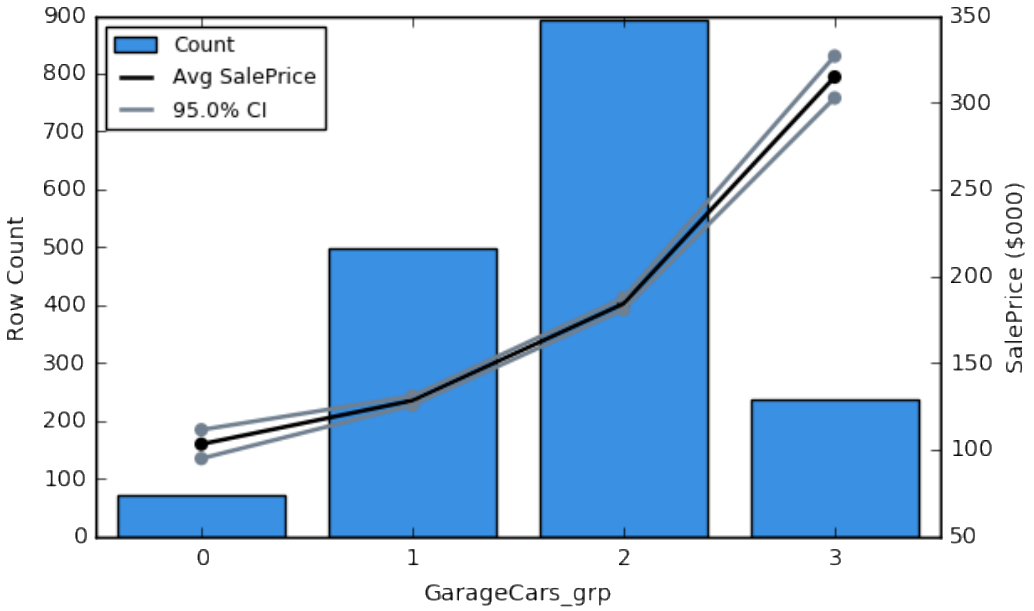

Figure 7 - Grouped GarageCars_grp Count and Average Sale Price

The figure shows an intuitive relationship as homes with more garage spots cost more.

Exploratory Data Analysis – Numeric Variables

YearBuilt shows a relationship with sale price. Figure 8 shows this relationship a scatter-plot with a smooth curve fit. Note that years prior to 1950 do not show much differentiation in sale price.

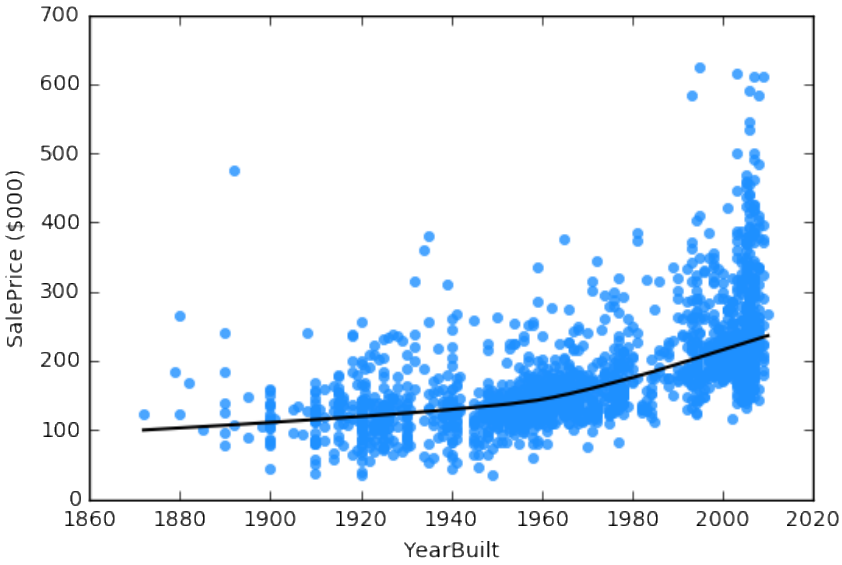

Figure 8 – Sale Price by YearBuilt

Instead, **YearRemodel**, that has a minimum value of 1950 can be used instead. The figure also shows that for more recent years, the variance in sale price is much greater. This limits the value of this variable in a predictive model. That variable is shown in Figure 9.

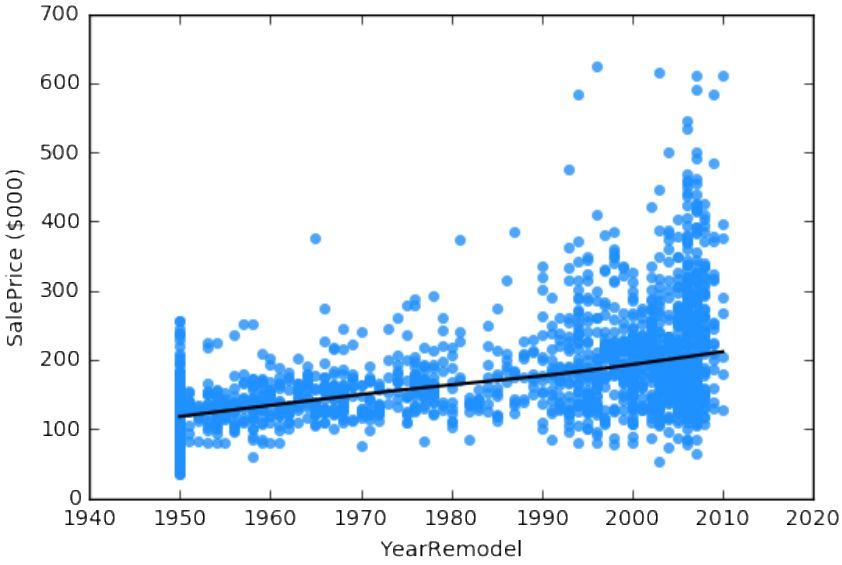

Figure 9 – Sale Price by YearRemodel

This variable has a clear linear relationship with sale price and should be considered in a predictive model.

Initial EDA of **LotArea** and **TotalBsmtSF** show significant outliers. These are shown visually in Figure 10. The observation with the maximum **LotArea** and the two observations with the highest **TotalBsmtSF** values are removed from the data. Additional analysis shows that these outliers are present for the other variables looked at for EDA.

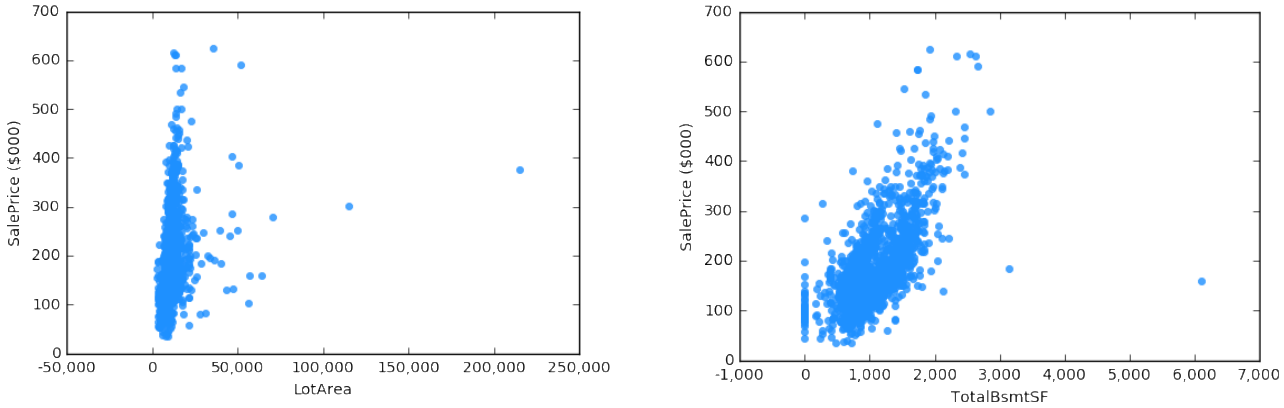

Figure 10 – Display of Outlier in Continuous Variables

LotArea shows a relationship with sale price which makes intuitive sense as larger lots would be more valuable. Even after removing the largest value, there are still large lots that impact the potential fit of a model. As a result, the values are capped at 14,000 square feet.

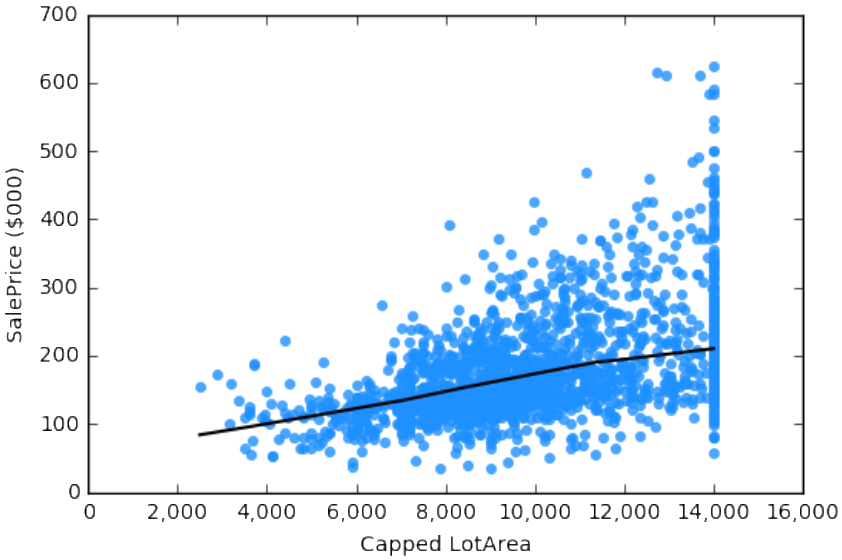

Figure 11 – Sale Price by Capped LotArea with LOESS smooth curve

The capped version of the variable shows a linear relationship with the target but as the area gets larger the variation increases. This could indicate that in rural areas, the lot area can be large without increasing the size of the sale price. The heteroscedasticity observed in this variable and others should be monitored as predictive models are built using these data elements.

GrLivArea also shows a strong relationship with sale price as shown in Figure 12.

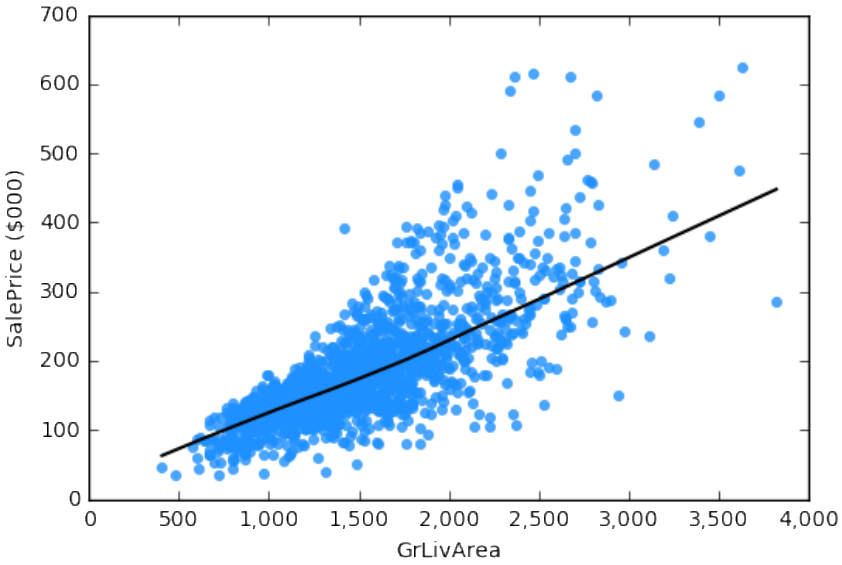

Figure 12 – Sale Price by GrLivArea with LOESS smooth curve

The figure demonstrates the same heteroscedasticity shown with **LotArea** but capping does not appear to be required. The linear relationship appears visually significant, so this variable is a candidate for a predictive model.

TotalBsmtSq is complicated by the fact that just under 5% of the observations do not have a basement. The relationship between this variable and sale price is shown in Figure 13.

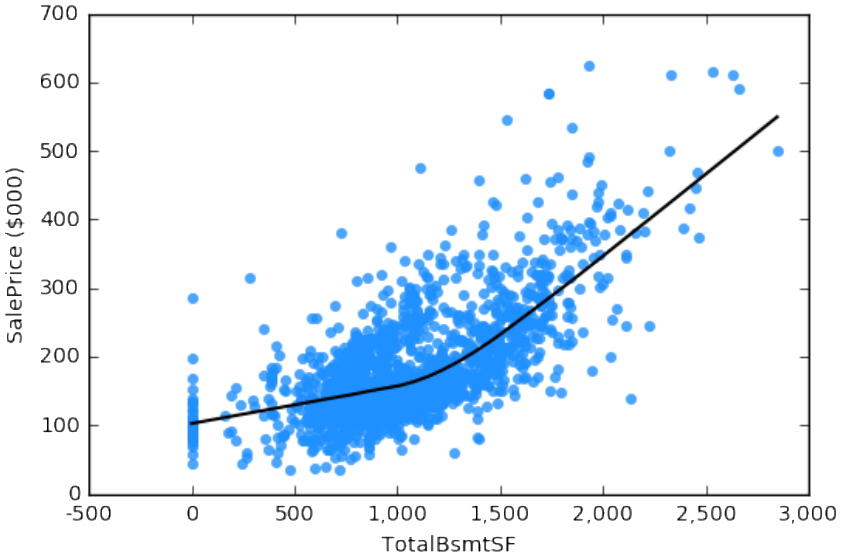

Figure 13 – Sale Price by TotalBsmt with LOESS smooth curve

The figure shows that the absence of a basement results in a nonlinear relationship, making it not a prime candidate for a linear predictive model.

LotFrontage would seem to be an intuitive choice as a large value indicates a wider distance between homes. However, it would be correlated with **LotArea**. Figure 14 shows the relationship with Sale Price.

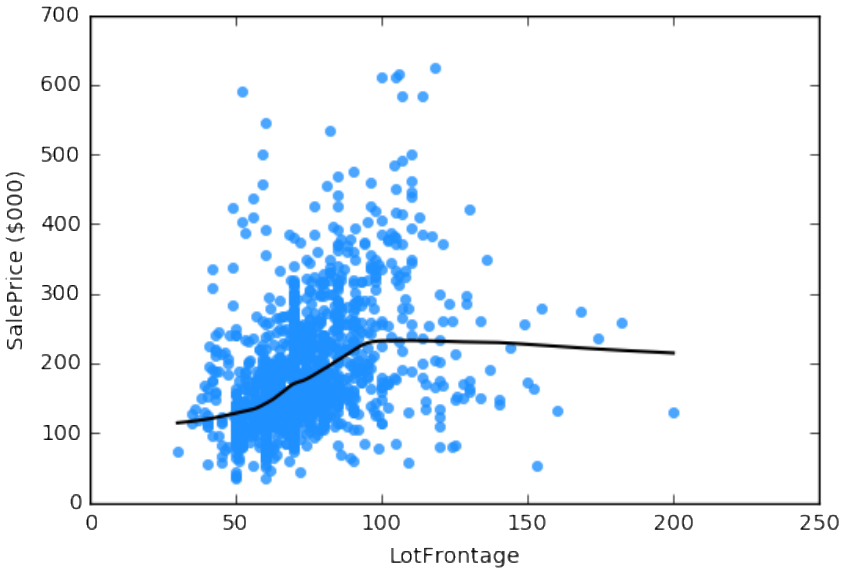

Figure 14 – Sale Price by LotFrontage with LOESS smooth curve

The figure shows a lack of a strong relationship between the variables. The smooth fit does not appear to indicate a strong relationship either. With the large number of missing values that were required to be imputed, the variable is not a strong candidate for a predictive model.

The EDA process leaves the following variables to use in model building: **YearRemodel, GarageCars_grp, KitchenQual_grp, Neighborhood_grp, BedroomAbvGr_grp, LotArea_cap, GrLivArea, OverallQual_grp,** and **OverallQual.**

Build Models – Simple

For each of the final variables listed above, a simple linear regression is fit to the training data For simple models, Best Subsets regression is identical to Forward or Backwards variable selection. In the case of categorical variables with suffix "_grp", dummy variables are created prior to fitting. Figure 15 shows the R-squared, adjusted R-squared, and AIC for each single-variable model on the training data.

Variable	R-Squared	Adjusted R-Squared	AIC
OverallQual_grp	0.7411	0.7399	40,828.30
OverallQual	0.6715	0.6713	41,217.95
Neighborhood_grp	0.6359	0.6326	41,420.38
GrLivArea	0.5700	0.5698	41,674.13
GarageCars_grp	0.5528	0.5520	41,744.75
KitchenQual_grp	0.4755	0.4745	42,015.13
YearRemodel	0.2799	0.2795	42,548.19
LotArea_cap	0.2452	0.2448	42,627.91
BedroomAbvGr_grp	0.0669	0.0658	42,989.51

Figure 15 – Simple Model Fit Statistics Sorted by AIC

The figure shows that the model with **OverallQual_grp** is the best fitting by a large margin by all three statistics. Due to the dummy variables, the R-squared is greater than the adjusted R-squared value by a larger margin than continuous variables. The penalty for a more complicated model, however does not impact the relative performance of this leading model, **Model 1**. Figure 16 shows that almost all of the **OverallQual** levels are statistically significant and the variable as a whole is statistically significant.

Variable	P-Value
Intercept	0.000
OverallQual_grp_03	0.156
OverallQual_grp_04	0.000
OverallQual_grp_05	0.000
OverallQual_grp_06	0.000
OverallQual_grp_07	0.000
OverallQual_grp_08	0.000
OverallQual_grp_09	0.000
OverallQual_grp_10	0.000

Figure 16 – Model 1 Variable Significance

Build Models – Two Variable Prediction

For each of the final variables, a linear regression is fit to the each of the combinations of two variables similar to the approach of the simple regression. However, in this case Best Subsets is not equivalent to other variable selection techniques. The fit statistics for the top ten models by AIC are shown in Figure 17.

Variables	R-Squared	Adj R-Squared	AIC
('GrLivArea', 'OverallQual_grp')	0.8205	0.8196	40,209.21
('Neighborhood_grp', 'OverallQual_grp')	0.8034	0.8007	40,391.52
('Neighborhood_grp', 'GrLivArea')	0.7906	0.7886	40,484.41
('LotArea_cap', 'OverallQual_grp')	0.7835	0.7823	40,527.36
('GarageCars_grp', 'OverallQual_grp')	0.7816	0.7801	40,546.29
('Neighborhood_grp', 'OverallQual')	0.7713	0.7691	40,633.96
('GrLivArea', 'OverallQual')	0.7606	0.7603	40,683.69
('KitchenQual_grp', 'OverallQual_grp')	0.7571	0.7555	40,726.38
('BedroomAbvGr_grp', 'OverallQual_grp')	0.7528	0.7513	40,754.03
('YearRemodel', 'OverallQual_grp')	0.7520	0.7506	40,757.74

Figure 17 – Two-Variable Model Fit Statistics Sorted by AIC

The figure shows that the combination of **GrLivArea** and **OverallQual_grp** is the best fitting model buy all fit statistics. **OverallQual_grp** is shown to be the most important variable, showing up in almost every top ten model. Fit statistics show that the terms as a whole in this top model, **Model 2,** are significant (Figure 18).

Variable	P-Value
Intercept	0.516
GrLivArea	0.000
OverallQual_grp_03	0.442
OverallQual_grp_04	0.008
OverallQual_grp_05	0.000
OverallQual_grp_06	0.000
OverallQual_grp_07	0.000
OverallQual_grp_08	0.000
OverallQual_grp_09	0.000
OverallQual_grp_10	0.000

Figure 18 – Model 2 Variable Significance

Figure 19 shows diagnostic plots for the top performing model. The figure shows that standardized model residuals are approximately normally distributed and not correlated with the response variable. It does show however that the variance in residuals is largest for higher groups of both variables.

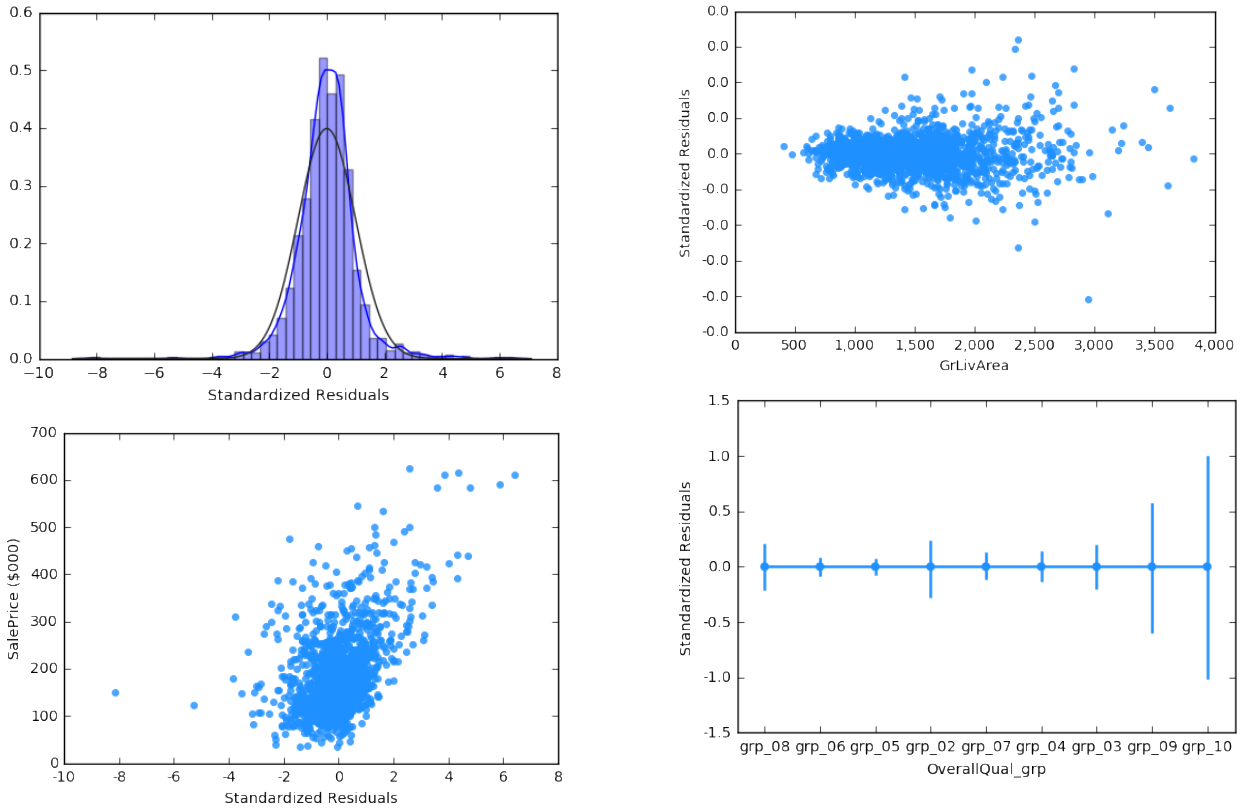

Figure 19 – Regression Diagnostic Plots for Model 2

Model 3 – Two-Variable Model with Transformations

To adjust for this heteroskedasticity, a third model is attempted using the natural log of **GrLivArea** and **OverallQual_grp's** highest value groups (9 and 10) combined. The fit statistics in this new model are show in Figure 20.

Vars	R-Squared	Adj R-Squared	AIC
('GrLivArea_log', 'OverallQual_grp2')	0.7574	0.7563	40,716.44

Figure 20 – Fit Statistics for Model 3

The fit statistics are not an improvement in terms of an of the statistics shown. In looking at the summary statistics shown in Figure 21, multiple dummy levels are now not statistically significant

Variable	P-Value
Intercept	0.000
GrLivArea_log	0.000
OverallQual_grp2_03	0.972
OverallQual_grp2_04	0.164
OverallQual_grp2_05	0.001
OverallQual_grp2_06	0.000
OverallQual_grp2_07	0.000
OverallQual_grp_08	0.000
OverallQual_grp_09	0.000

Figure 21 – Model 3 Variable Significance

This model is checked to determine if the transformed variables do a better job of satisfying the linear regression assumptions. In this case, the assumptions are violated as standardized residuals are actually correlated with the predictors. As a result this model is not recommend for further use as a final model.

Model 4 – Two-Variable Model with Encoding for Neighborhood (Bonus)

Target encoding is a technique common in consumer credit modeling. It replaces the label of a categorical variable with the average of the target variable for that level. It is particularly useful when there are a large number of categorical levels or when monotonicity needs to be enforced for business purposes. For example, as costs of a home rises, the rate that the homeowner pays for insurance should increase all else being equal. This encoding can force that relationship. **Neighborhood**, even after grouping has a number of levels. For **Model 4**, the target encoded version of **Neighborhood_grp** is used in place of **GrLivArea**. The fit statistics are shown in Figure 22 and the variable significance is shown in Figure 23.

Vars	R-Squared	Adj R-Squared	AIC
('OverallQual_grp', 'Neighborhood_grp_enc')	0.7957	0.7946	40,428.83

Figure 22 – Fit Statistics for Model 4

Variable	P-Value
Intercept	0.843
Neighborhood_grp_enc	0.000
OverallQual_grp_03	0.124
OverallQual_grp_04	0.000
OverallQual_grp_05	0.000
OverallQual_grp_06	0.000
OverallQual_grp_07	0.000
OverallQual_grp_08	0.000
OverallQual_grp_09	0.000
OverallQual_grp_10	0.000

Figure 23 – Model 4 Variable Significance

The figures show that the model does not fit as well as the unencoded version. This is likely due in part to the fact that very few variables are included in the model to begin with. Perhaps there is greater benefit in this technique when more complicated models are produced with several variables. The disadvantage of the technique is that it makes the model more complicated to implement and can lead to overfitting because it uses the target variable to create the variable. Summary statistics and diagnostic plots, showed that this encoding did not violate the assumptions that **Model 3** does.

Select Models

Of the four final models, **Model 2** fit the training data best by r-squared, adjusted r-squared, and AIC. Additionally, **Model 3** violated multiple linear regression assumptions and using target encoding in **Model 4** is more complicated to implement and runs the risk of overfitting. The difference in fit statistics between **Model 1** and **Model 2** is large enough to merit using a more complicated model. As a result, **Model 2**, with **GrLivArea** and grouped **OverallQual** is selected as the final model. For reference, the fit statistics of the four models on the training data is shown in Figure 24.

Model	R-Squared	Adj R-Squared	AIC
Model 1	0.7411	0.7399	40,828.30
Model 2	0.8205	0.8196	40,209.21
Model 3	0.7574	0.7563	40,716.44
Model 4	0.7957	0.7946	40,428.83

Figure 24 – Grouped Fit Statistics on Training Data

Final Model Formula

The formula representing the final model (**Model 2**) is as follows:

p_saleprice = 7,488.24
*+ 60.87 * GrLivArea*
+ 10,267.28 * (1 if OverallQual_grp = "grp_03", 0 otherwise)
*+ 31,122.14 * (1 if OverallQual_grp = "grp_04", 0 otherwise)*
*+ 54,777.53 * (1 if OverallQual_grp = "grp_05", 0 otherwise)*
*+ 68,020.99 * (1 if OverallQual_grp = "grp_06", 0 otherwise)*
+ 95,673.09 * (1 if OverallQual_grp = "grp_07", 0 otherwise)
*+ 151,438.14 * (1 if OverallQual_grp = "grp_08", 0 otherwise)*
*+ 241,277.31 * (1 if OverallQual_grp = "grp_09", 0 otherwise)*
*+ 308,512.74 * (1 if OverallQual_grp = "grp_10", 0 otherwise)*

The formula makes intuitive sense as when **GrLivArea** increases, the predicted sale price increases. This fits the one-way analysis results and the general idea that as home square footage increases, the home will cost more. For **OverallQual_grp**, the coefficients are monotonically increasing, showing that as quality increases, predicted sale price increases as well. This makes intuitive sense and fits the one-way analysis. These coefficients are used to predict sale price for the test data and save those predictions to a CSV.

Conclusion

Four 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 included **GrLivArea**, the square footage of the home above ground level, and **OverallQual_grp**, grouped categories referring to the quality of the property. While the final model's coefficients made intuitive sense and the model fit the data well, other variables show promise in more complicated models.

Bonus – Gamma GLM

In looking at the response, it is a currency and is strictly positive. If this model were to be implemented, using a gamma log-link generalized linear model could allow for inflation to be considered in future years more easily. In some ways, this problem is similar to building a model to predict claim severity for insurance pricing.

A gamma log-link GLM is fit to the same data as Model 2. The AIC for this model is lower than Model 2 although not directly comparable due to such large changes in model form. This indicates that while out of the scope of the current assignment, a GLM could be a good approach for this data.