BAS 220 Group #1
Carlos Vargas Bayas
Clayton Dombrowski
Elizabeth Etchells
Bradley Jordan

Housing Business Report

Part 1

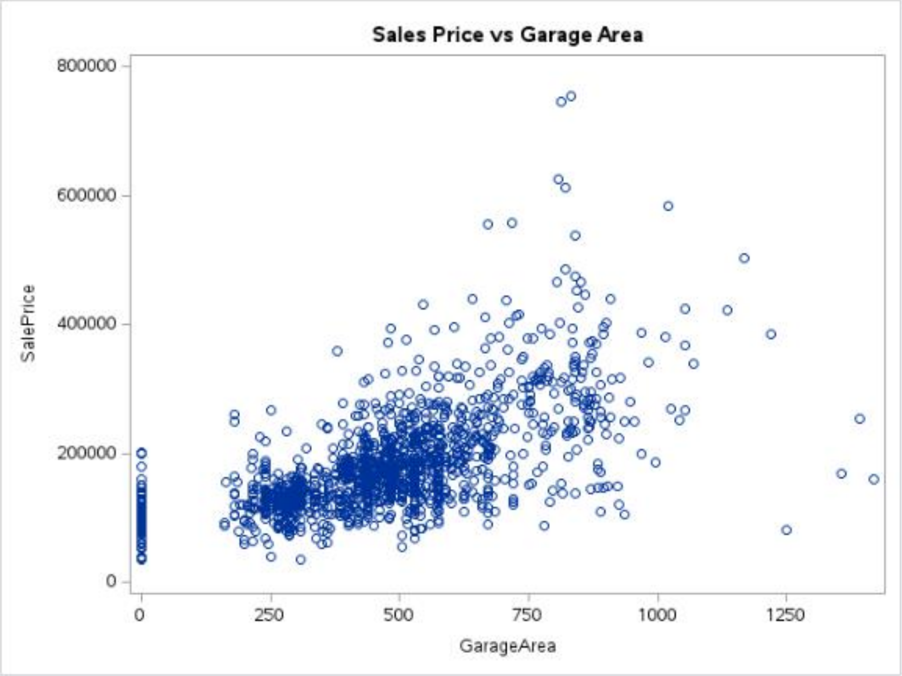
The housing data set contains information on 1,460 housing units in Ames, Iowa. The 23 variables include information on the type of housing, the year built, the lot area, the number of rooms and bedrooms, and the overall conditions of the house. The goal of this project is to use the available data to develop a model and predict the sales prices of the houses.

After much consideration between 3 categorical variables. The team chose KitchenQual as our categorical variable. This is because we see how the value of the house can hold due to the finishes in the kitchen and it can also be a good predictor of the condition of the rest of the house. Therefore, if the condition of the kitchen is excellent, then it will have an effect in the home's sales price.

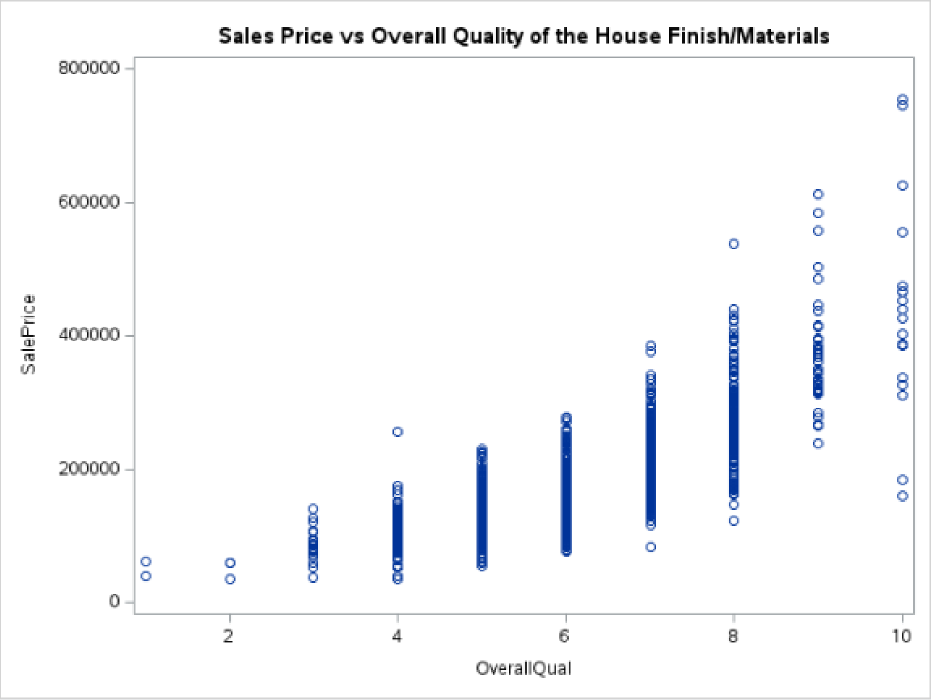
Two variables that have outliers are LotArea and BedroomAbvGr. The LotArea variable has 4 very large outliers. It is likely that these properties are outside the main areas of Ames where there is more space. GarageArea also has 4 outliers. One reason may be that largest houses also have the largest garages.

Part 2

The below plot is included in this report because of the positive correlation between the sale price and the garage area.



This next plot is included in the report because of the positive correlation between sale price and overall quality.



The variables with the strongest correlation to SalePrice are OverallQual, GrLivArea, GarageArea, FullBath, and TotRmsAbvGrd.

Variables that have no or very little correlation include MoSold, YrSold, OverallCond, and BedroomAbvGr.

Part 3

For the regression analysis, we initially included eleven variables: sales price (***SalePrice***) as our Response Variable and total rooms above ground (***TotRmsAbvGrd***), living area above ground (***GrLivArea***), overall quality of the house (***OverallQual***), the garage area (***GarageArea***), number of full baths (***FullBath***), and all four levels of the categorical variable for the kitchen quality (***Fair_vs_Poor, Typical_Average_vs_Poor, Good_vs_Poor, Excellent_vs_Poor***) as our Predictor Variables.

In running the regression analysis to refine the model, we first looked at the Variance Inflation Factor (VIF) for each of the variables included in the model. One by one, we dropped the variables that had a VIF > 5. The VIF test eliminated the ***Typical_Average_vs_Poor*** variable from our regression analysis. Next, we examined the t-tests for the variables left in the model and dropped the variables where Pr > |t| > 0.05, one by one. The T-test eliminated the variables ***TotRmsAbvGrd***, ***FullBath***, and **Fair_vs_Poor**.

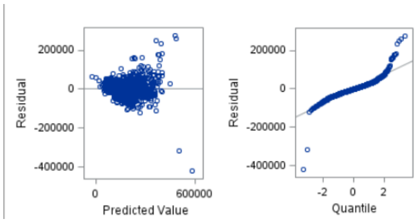
Once these tests were complete, we had five variables remaining in the model to predict ***SalePrice***: ***GrLivArea, OverallQual, GarageArea, Good_vs_Poor***, and ***Excellent_vs_Poor***. The model equation is:

$$\textbf{\textit{SalePrice}} = -64237 + 48.04341*\textbf{\textit{GrLivArea}} + 21532*\textbf{\textit{OverallQual}} + 63.85506*\textbf{\textit{GarageArea}} + 15307*\textbf{\textit{Good\_vs\_Poor}} + 68272*\textbf{\textit{Excellent\_vs\_Poor}}$$

Our Model provides us with the following predictions, assuming we hold all other variables constant:
for every square foot of living area, we can expect a $48.04 increase in sales price.
For every unit increase in the overall quality, we can expect a $21,532 increase in sales price.
For every square foot of garage area, we can expect a $63.86 increase in sales price.
If the kitchen quality is rated good, we can expect a $15,307 increase in sales price.
If the kitchen quality is rated excellent, we can expect a $68,272 increase in sales price.

R^2 is 0.7687 and **Adjusted R^2** is 0.7679. Adjusted R^2 explains the variance of the model after imposing a penalty for adding additional explanatory variables, so we want the value to be relatively close to the value of R^2 otherwise we potentially have created an overfitted model. R^2 and Adjusted R^2 are pretty close together for our model with a difference of 0.0008, so we can be sure that we have NOT overfitted the data set and have NOT used too many predictors in our model. R squared shows the variance of Sales Price that can be explained by our independent variables. So approximately 76.9% of the variance in Sales Price can be explained by the variables in our model.

Looking at the residual plots for our model, the predicted value vs. residual plot appears to have a mostly random pattern and therefore passes the tests of **Linearity**, **Constant Variance**, and **Independence** (Left Plot below). However, the quantile vs. residual ploy appears to have a S shape, which fails the test of **Normality** (Right Plot below).



The MAPE prediction accuracy metric is 15.08 which suggests that our model is accurate approximately 84.92% of the time in predicting a home's Sale Price.