UrbanWorth

By **AttaUllah Zahid Bhalli**

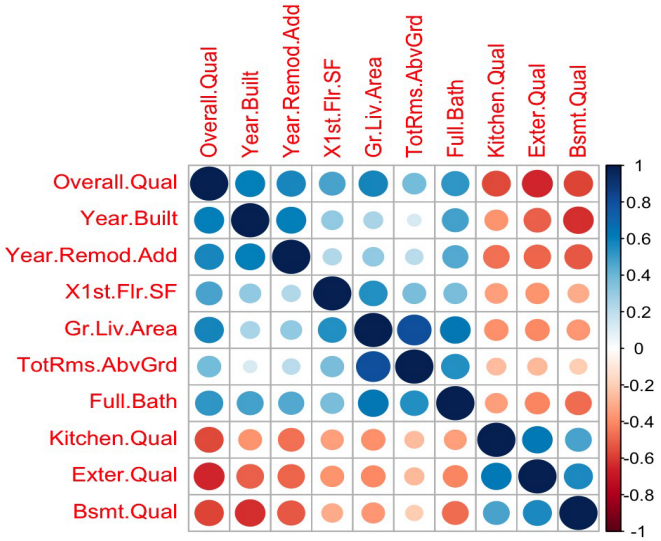
The study

We have the dataset for real estate sales in Ames, Iowa. The data set contains 79 variables that we can use to predict the Sales Price of a house. Each variable gives a description of the features of each house and at the end of the data set the actual price of the house is given. The goal of the study is to use the variables to predict the value of each house. We can answer the goals by first building an appropriate model then constructing a confidence interval from the testing data. The confidence interval should give the range in which our true predicted value is supposed to lie in with a certain amount of 95% confidence.

The data

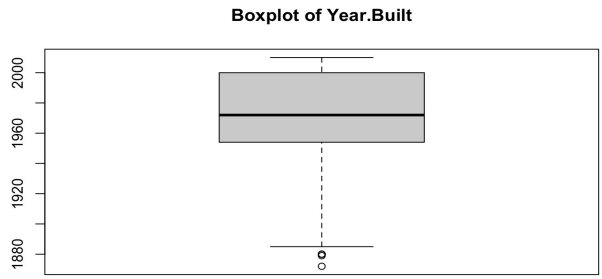
The data set consists of both categorical (ordinal and nominal) variables and quantitative (continuous and discrete) variables. In order to apply functions and carry out tasks on the data we need to convert the categorical variables in the factors using "as.factor()" function. We then shuffle the data to make sure that the training and testing data sets will be random samples. We then calculate the Pearson correlation coefficient of all the variables with the "SalePrice" field. I considered the variables with a correlation coefficient greater than equal to 0.5 to be a predictor of "SalePrice". Then to check whether the predictors were correlated with each other I plotted a correlation matrix between all of them (see below)

Correlation Matrix of Predictor Variables to Sales Price



Here you can observe that the pairs "TotRms.Abv.Grd & Gr.Liv.Area", "Overall.Qual & Exter.Qual" , "Year.Built & Bsmt.Qual" are highly correlated hence, we will have to check their VIF (**V**ariance **I**nflation **F**actor) .

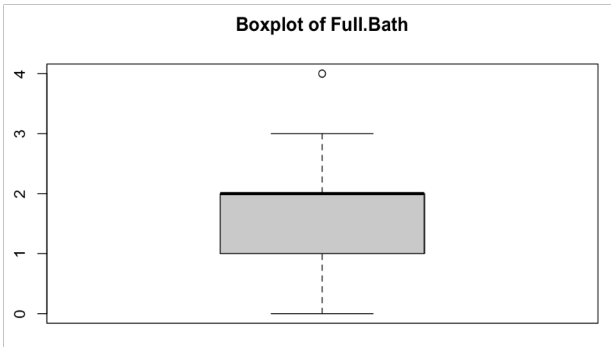
Then I checked the spread of the **quantitative variables** that were highly correlated to the sales price.



Year Built – Original construction date

As we can see the median of the "Year.Built'' variable is around 1970. The range of the variable is from about 1890 - 2010. However, there are a few outliers which indicate that few houses were made before 1890.

The interquartile range is about 50.



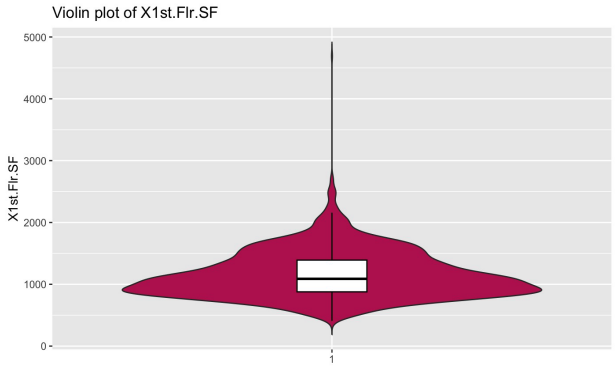
Full Bath - Full bathrooms above grade

The median of "Full.Bath" is 2 .

The range is from 0-3.

There is one outlier at 4 which means that there is one house with 4 full bathrooms.

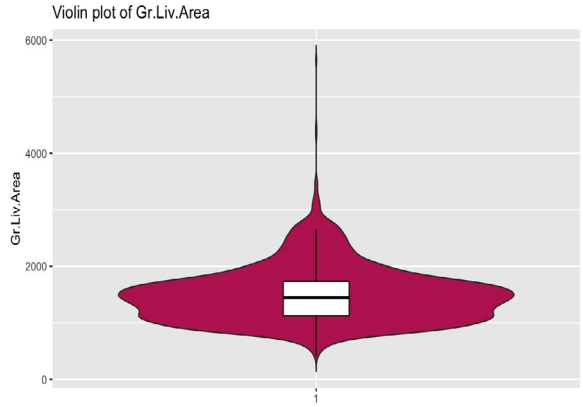
The interquartile range is 2



First Floor square feet

The distribution is right-skewed, and the data is unimodal with a hump around 800-900 sq ft.

The median surface area is a little over a 1000 (about 1250). The range of the surface area is from 500-2250.the inter quartile range is about 550.



Above grade (ground) living area square feet

The data here is somewhat multimodal (2 humps).

The median is about 1500.

The range is approximately from 500-2500 with the major hump at about 1750 sqft.

The interquartile range is about 600

Comparing this to the above plot indicates that a lot of houses also have a second floor

The models

The predictor variables that i will use to predict the sale price are :

1. **Overall.Qual** (Rates the overall material & finish of the house.)
2. **Year.Built** (Original construction date)
3. **Year.Remod.Add** (Remodel date)
4. **X1st.Flr.SF** (1st floor surface area)
5. **Gr.Liv.Area** (Above ground surface area in square feet)
6. **TotRms.AbvGrd** (Total rooms above grade)
7. **Full.Bath** (Number of full baths above grade)
8. **Kitchen.Qual** (Kitchen quality)
9. **Exter.Qual** (Condition of the exterior)
10. **Bsmt.Qual** (Basement height)

Before making the model to predict the sale data we must check for any interactions between the predictor variables. To do this, I used an interaction plot:

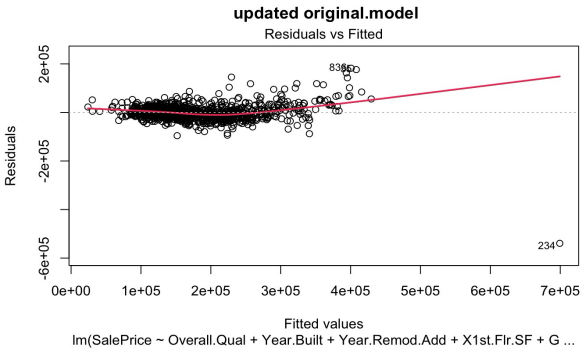
In this plot we can observe there is an interaction between Year.Built & Year.Remod.Add, which makes sense as **remodelling helps reset the age of house**.

Moreover, there is a three-way interaction between Year.Built ,Year.Remod.Add & Gr.Liv.Area. I believe this might be because renovated or newer homes that have a higher living area check the box on **important factors for prospective home buyers – age of the house and living space**. I will take note of these when building my model. On the left side of the plot, we can observe that vertical lines of several categorical variables are plotted – we will not consider that interaction given these are categorical.

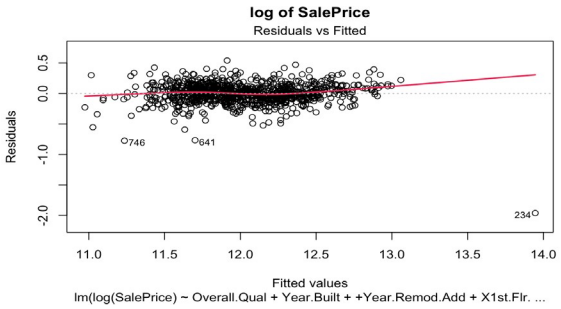
For building the model, I will use two different models: linear regression & random forest. I started off by building a linear model using our original predictor variables and including the appropriate interaction terms. When I calculated the VIF values of the model I observed that the VIF of Gr.Liv.Area is greater than 5 proving there is high

collinearity between Gr.Liv.Area & TotRms.AbvGrd. To fix this I dropped **TotRms.AbvGrd.**

After dropping TotRms.AbGrd, I observed the Residual vs Fitted plot and realized that this is not a good fit, so I tried transformations like the natural log of sale price; the square root of sale price; and squaring sale price. The natural log fit was the best.



Original Model (no transformation)
Here we can see that the points on the plot are not randomly spread , they have a slight curvature pattern which suggests that the relationship is not linear.



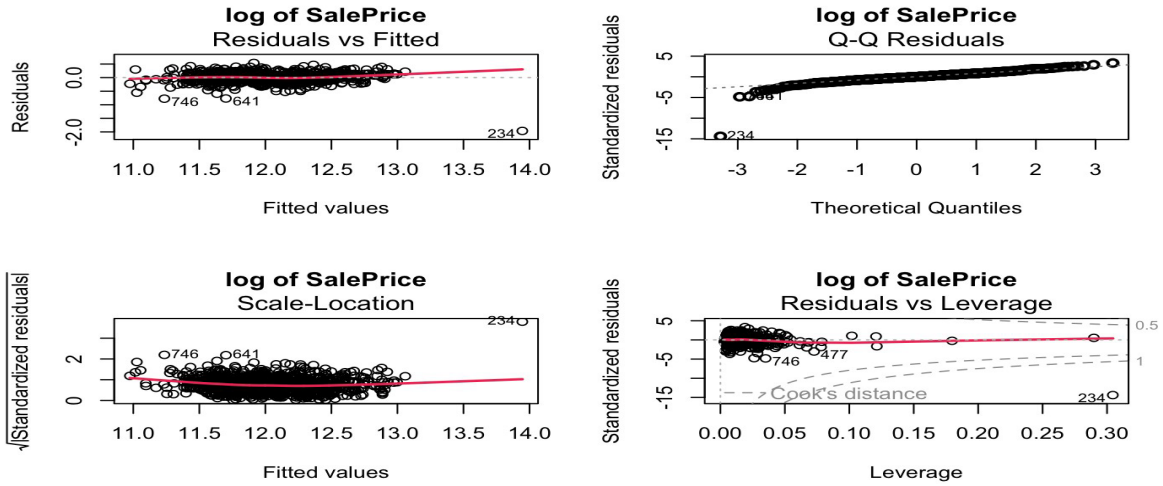
Natural Log of SalesPrice Model Here we can observe that the points form a "horizontal band " around the zero- line . Suggesting that the relationship is more linear.

Checking RandomForest Model: I tried making various random forest models by varying the mtry and ntree inputs. I found the best model was when **mtry was set to square root of 79** (total number of **85.64** variables excluding SalePrice) and **ntree was set to 500**. This model gave me the least mean squared error out of all the other random forest models.

Type of random forest: regression
Number of trees: 500
No. of variables tried at each split: 9
Mean of squared residuals: 890763556
% Var explained:

Results

I chose my final model to be the linear model with the log transformation. Because the mean squared error was lower compared to the random forest model. To check whether our model fits well or not we will use the standard diagnostic plots.



From these diagnostic plots we can observe that our model fits well. In the residual vs fitted plot the points form a "Horizontal Band " around the zero-line. Moreover, the red line is approximately horizontal at zero . Both ideas suggest that the fit is adequate. However, the point with index value of 234 is an outlier in this model . In the residual vs leverage plot, we can observe that the coordinate with index 234 is actually an influence point and a leverage point. Additionally, the last 2 points to the far right of the plot are also leverage points.

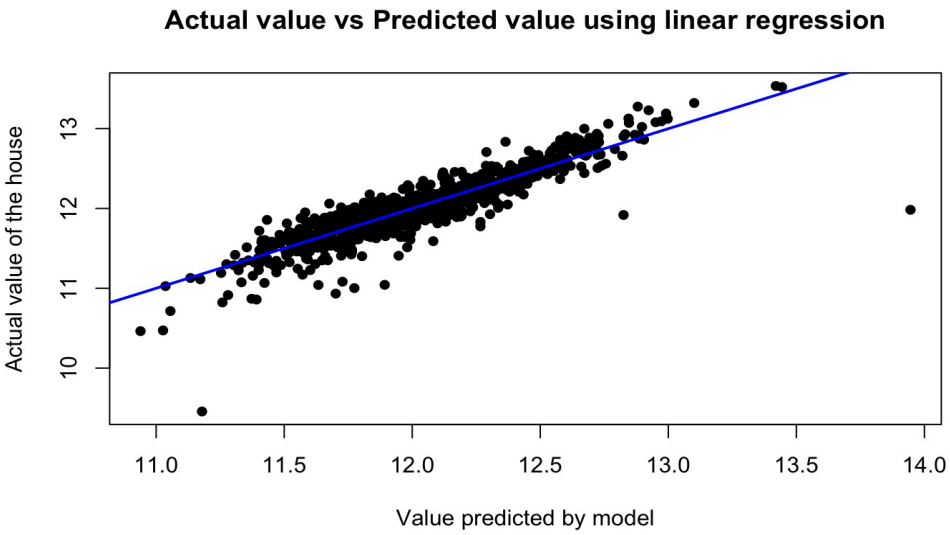
The final confidence interval calculated are below:

	2.50%	97.50%
(Intercept)	-3.96199E+02	9.82799E+01
Overall.Qual	9.69920E-02	1.22733E-01
Year.Built	-4.95768E-02	2.03777E-01
Year.Remod.Add	-4.67459E-02	2.01346E-01
X1st.Flr.SF	1.04951E-04	1.73239E-04
Gr.Liv.Area	-1.27807E-01	1.92275E-01
Full.Bath	-3.25816E-02	2.22015E-02
Kitchen.Qual	-4.88353E-02	-1.51617E-02
Exter.Qual	-2.39857E-02	1.99389E-02
Bsmt.Qual	-1.34552E-02	4.16882E-03
Year.Built:Year.Remod.Add	-1.00726E-04	2.63723E-05
Year.Built:Gr.Liv.Area	-9.62626E-05	6.77736E-05
Year.Remod.Add:Gr.Liv.Area	-9.58717E-05	6.45583E-05
Year.Built:Year.Remod.Add:Gr.Liv.Area	-3.41600E-08	4.80476E-08

The confidence interval is sensitive to the type of model we choose . Moreover, the confidence interval for the training set and testing set isn't that different because the data is spit randomly. This model is quite accurate and has a Mean Squared Error of only 0.031.

Checking the accuracy of our model:

Here is a plot showing the actual value plotted against the predicted value.



The values lie approximately on the line y=x(blue line) which means the model is quite accurate.