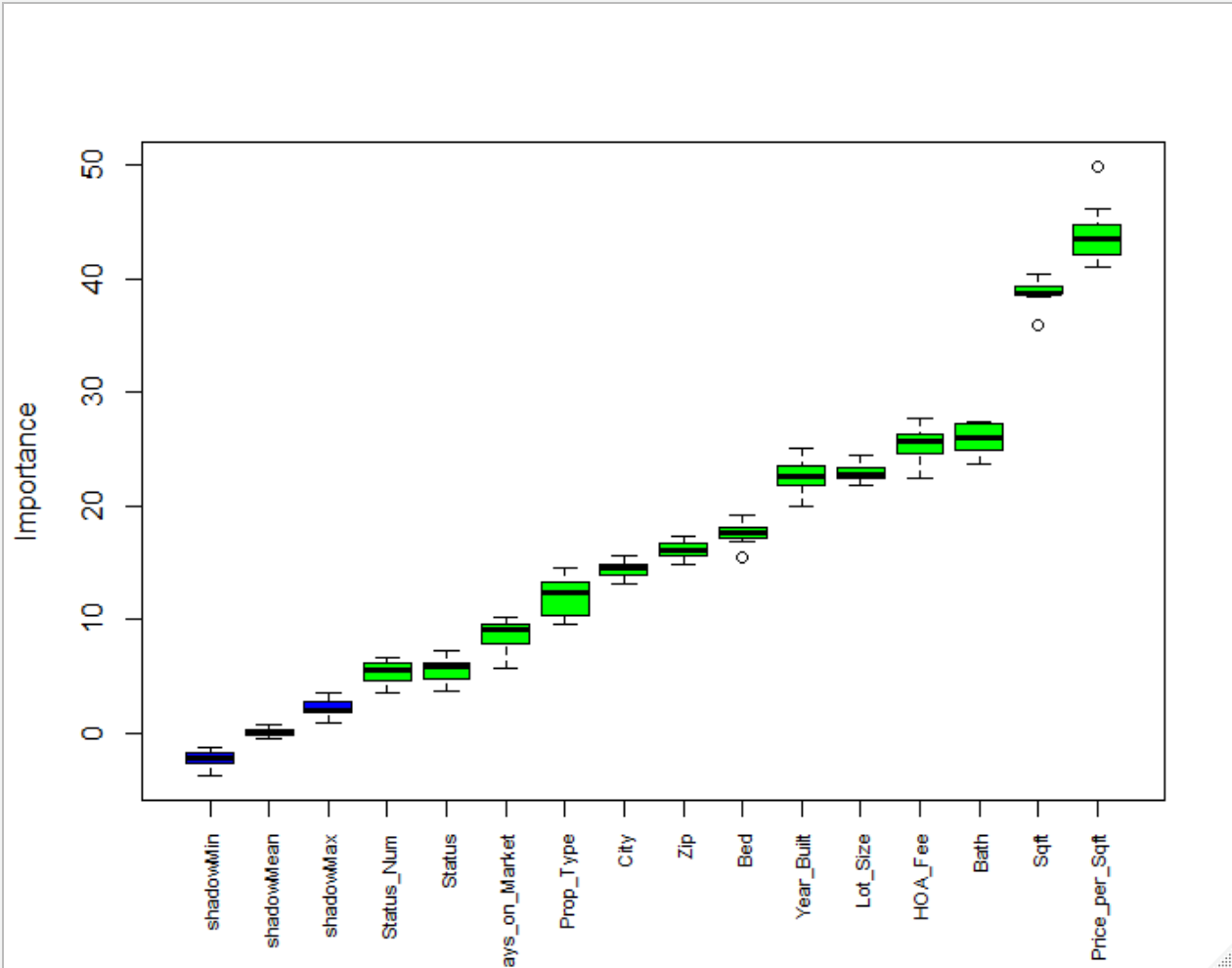
**Introduction**

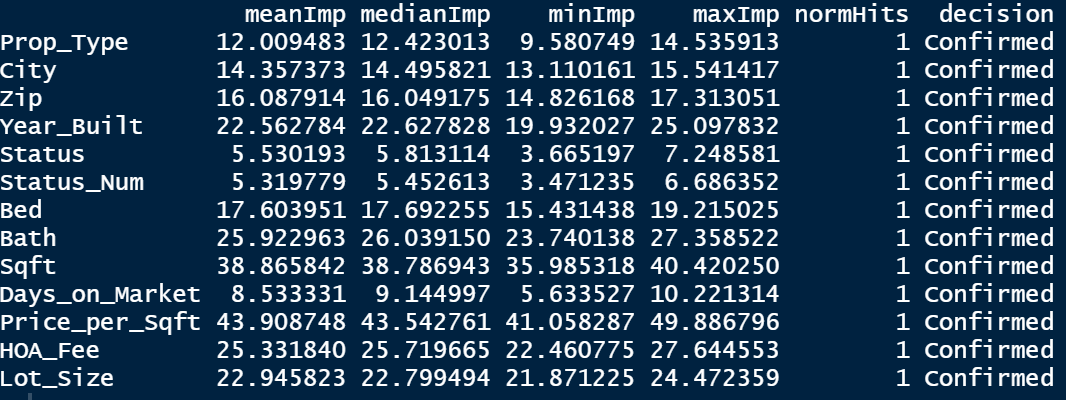
I looked at the housing data for Hamilton County to see what the predictors of the price were. I obtained the data from Redfin.com on 4 different days so that the data was sufficiently different. The data was messy with plenty of missing data so some cleaning had to be done. The accompanying csv and r file will have the whole cleaning of the data, but I will briefly describe it here.

* The status column was duplicated and eventually transformed into integers, 1 being for sold and 0 for active
* Renamed the columns
* Removed vacant land and multi family property types
* Changed the names of the remaining property types
* Checked for NA’s in dataframe
* Calculated the single and multi family lot size median to be imputed into the lot size column
* Combined various cities with not very many values into the neighboring city with more values to make only 5 different values
* Calculated and impute median values for HOA fee by city
* Remove the 2 remain rows with NA because we cant calculate the median for year built
* Restructure and reorder data
* Put latitude and longitude into separate data frame and remove from housing data

After all the cleaning was done then we can explore the data.

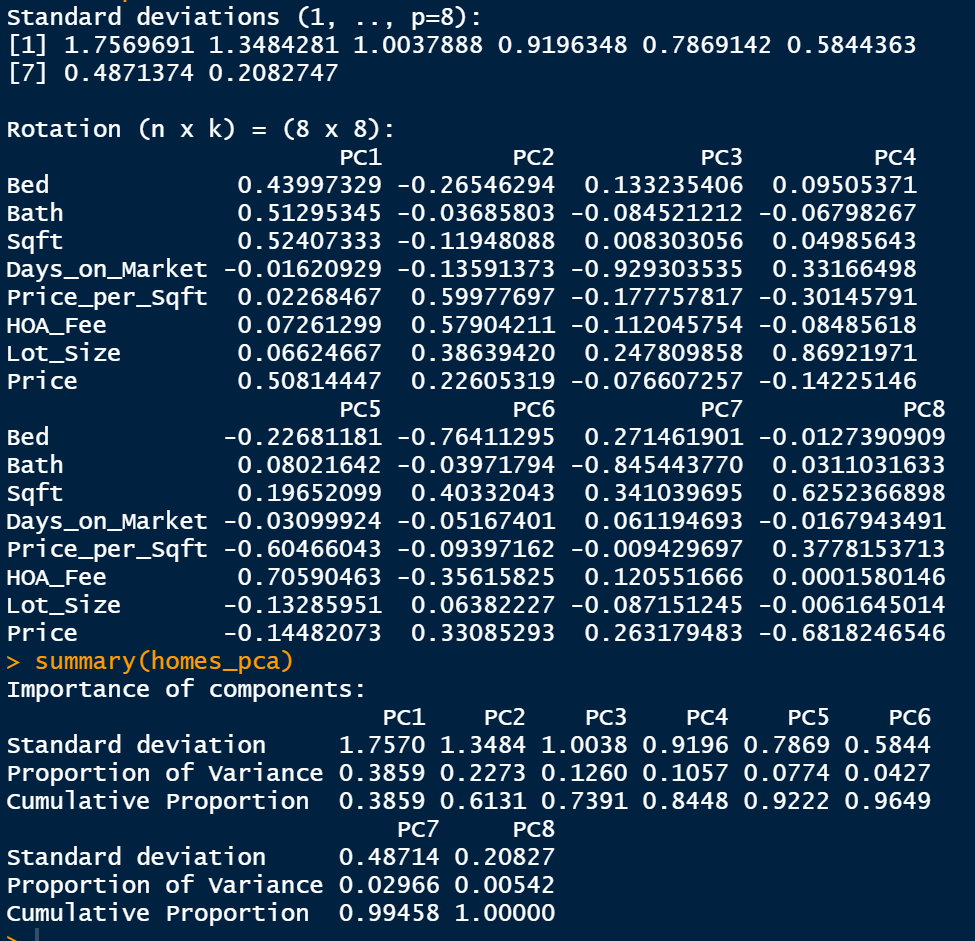
**Primary Component Analysis**

I started by preforming a primary component analysis. The first method was automated using the Boruta R package. As you can see, all the columns are quite important to determining the price of a house especially square feet and price per square feet. The package also output each importance rating for each of the values.

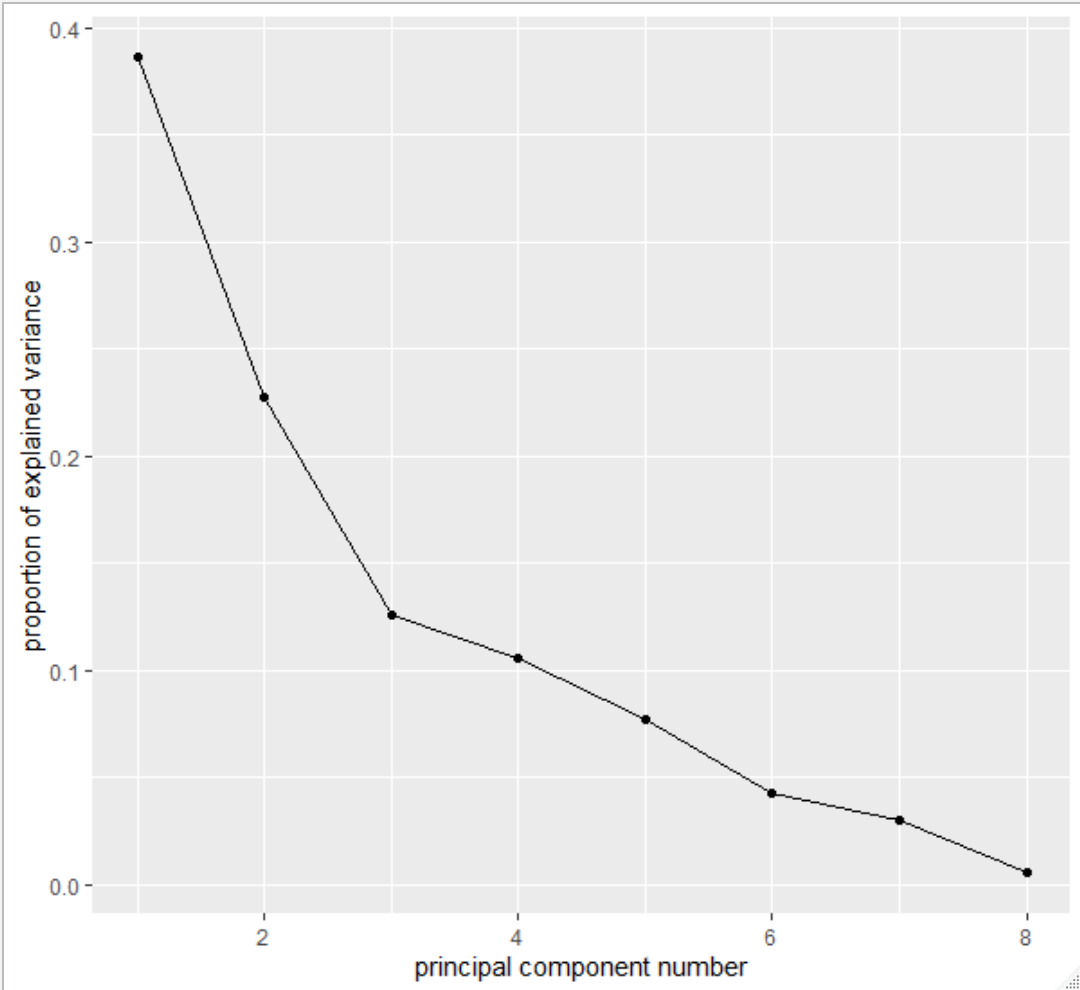


This backs up what the graph showed, and that is that square foot and price per square foot are the highest rated components.

Next, the primary component analysis without the Boruta algorithm using prcomp(). The results were similar.



With only 5 pc I was able to explain 92% of the data and by adding only 3 more primary components 100%. Luckily there was only 8. All of this is explained visually in the following scree plot.



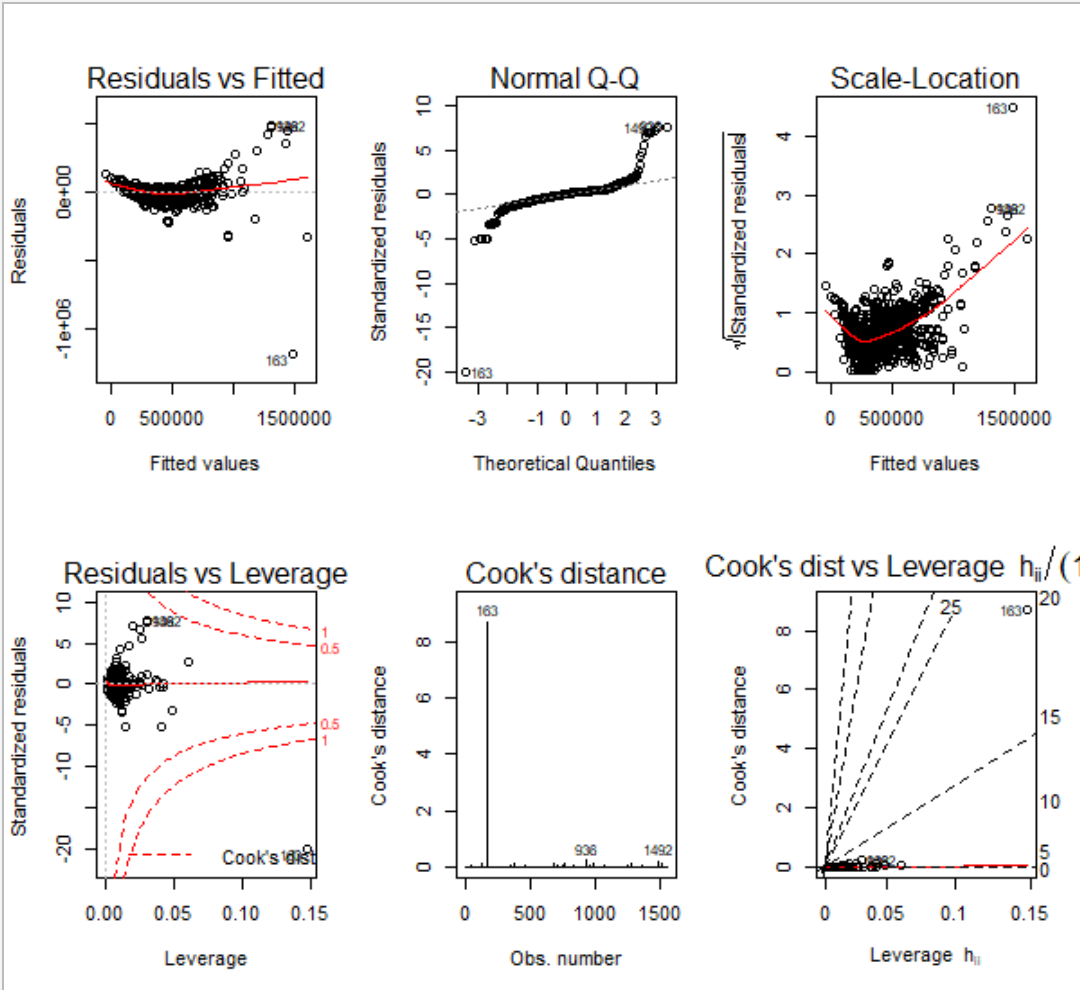
After determining the primary components, now we look at predictions of the data. Starting out with linear regression, the model was fitted with the following formula:



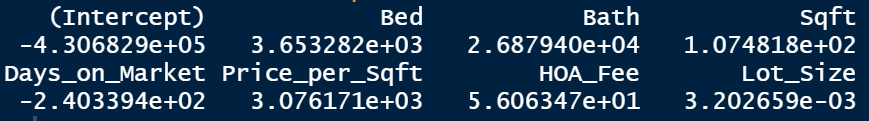
Based on the model bath, sqft, HOA fee, days on market, and price per sqft all had statistically significant p-values further confirming the primary component analysis. The model also produced an r-squared value and adjusted r-squared value of 0.913 and 0.912, respectfully.

**Linear Model**

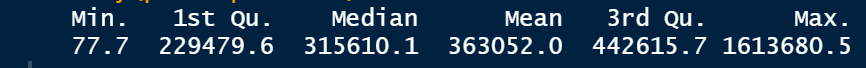
Plotting the model yield some interesting insights. The first graph is the residuals vs fitted plot. While the graph is pretty centralized around 0 it does have a distinctive curve to it where it is supposed to be random and it also contains some outliers. The next plot is the Q-Q normality plot. For the most part the graph is fairly normal, but again there are some outlier values that skew the normality of the plot. The next plot is the scale-location plot. This one, as with the first, should look random and here we have an odd v-shaped curve with a more clearly defined outlier variable. The bottom 3 plots are just variations of cooks distance plot. Cooks distance plot just shows us which variables have the most leverage on the data. In all of the graphs we see that the outlier is number 163. A further look into number 163 doesn’t reveal much out of the ordinary, but with a little more digging we might uncover that it was one of the houses that was in a different city that was combined into a different group. I would consider removing the value and reassess the model.



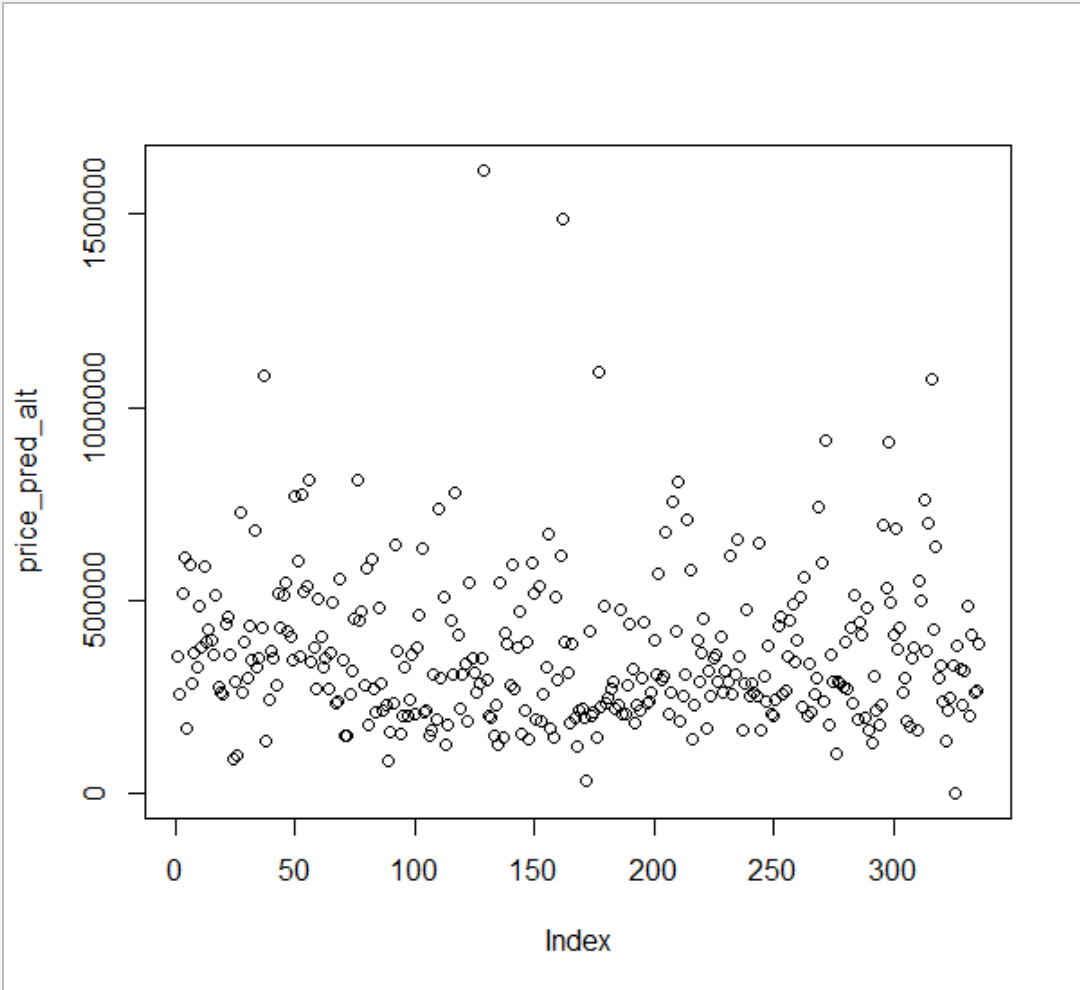
Coefficients for the linear model. Each of the values multiply the predictor. So as the bed value goes up so does the price and as days on market goes up, price goes down.



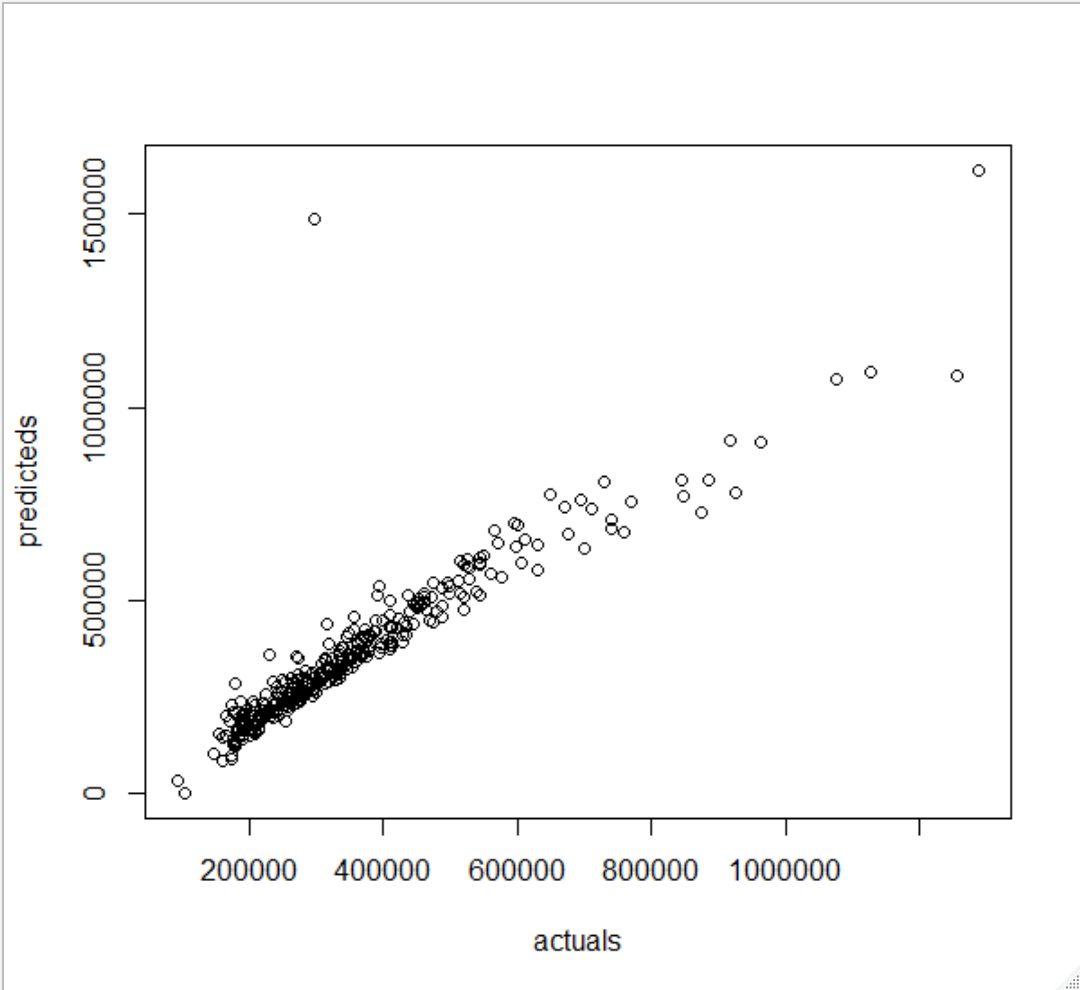
Next we use the model to predict the price. First we split the data, then we run the make the predictions. The following are the range of values of the prediction in dollars.

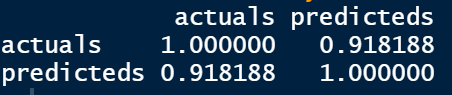


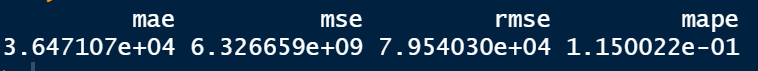
The plot of predicted values is quite good because it looks quite random with most of the values hanging around the 300,000 to 400,000 dollar range.



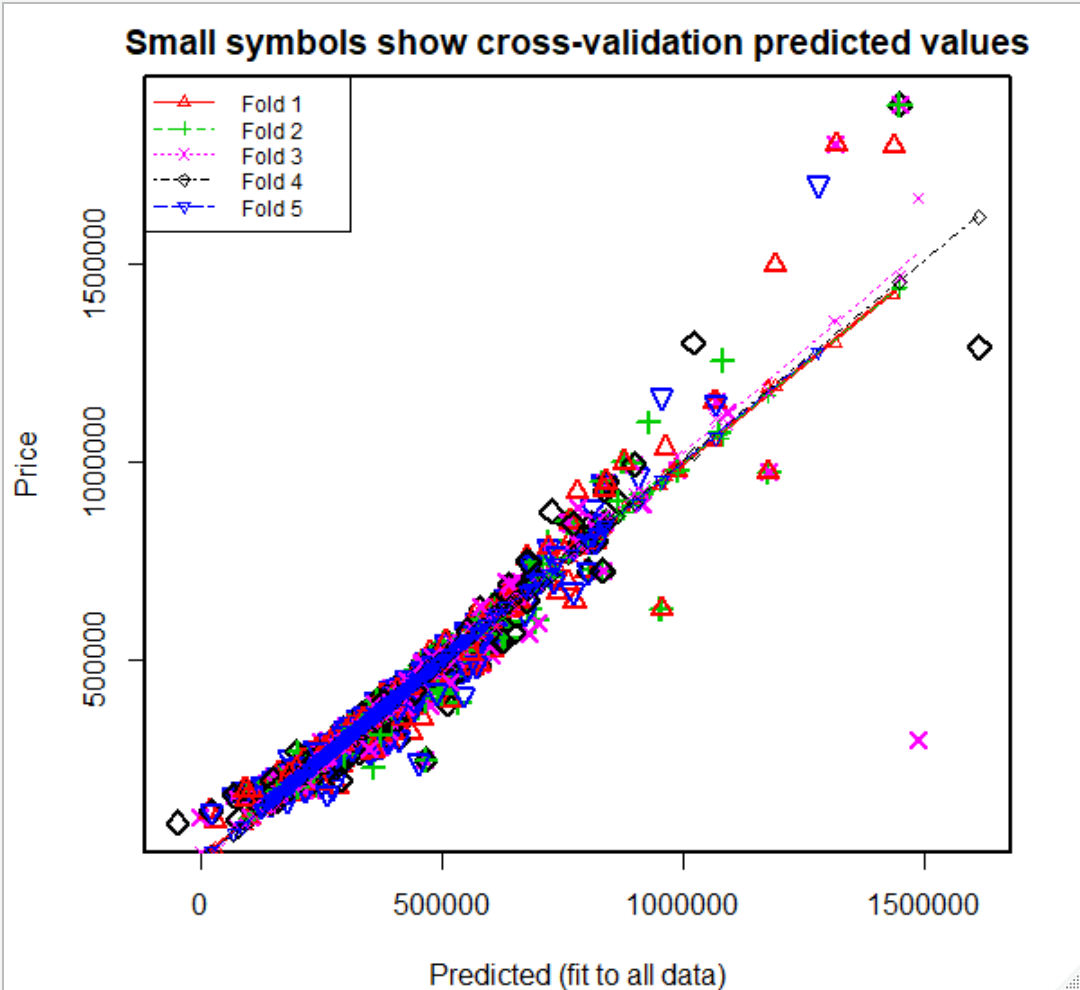
After predictions the values and comparing them to the actuals, we can plot both values. Here we see a pretty positive upward linear trend, which means the model predicted the values quite accurately. And running the correlation accuracy we see that, indeed, the predictions was 92% accurate. Along with the correlation accuracy we also have very small MAE, MSE, RMSE, and MAPE values.





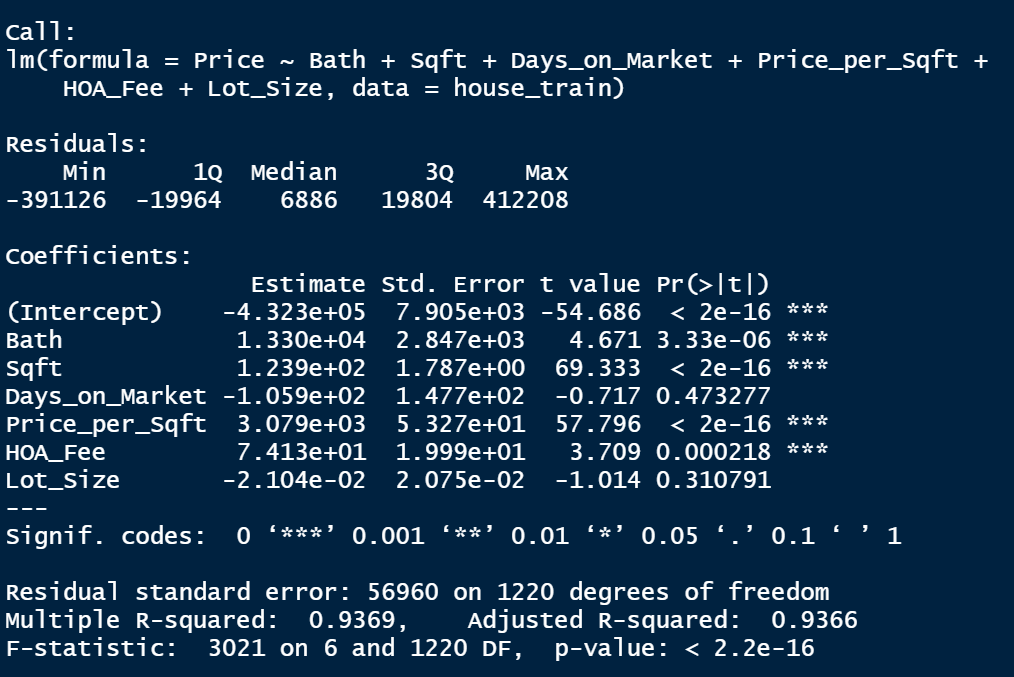


Running a cross validation on the linear model we can again confirm what was previously discussed.

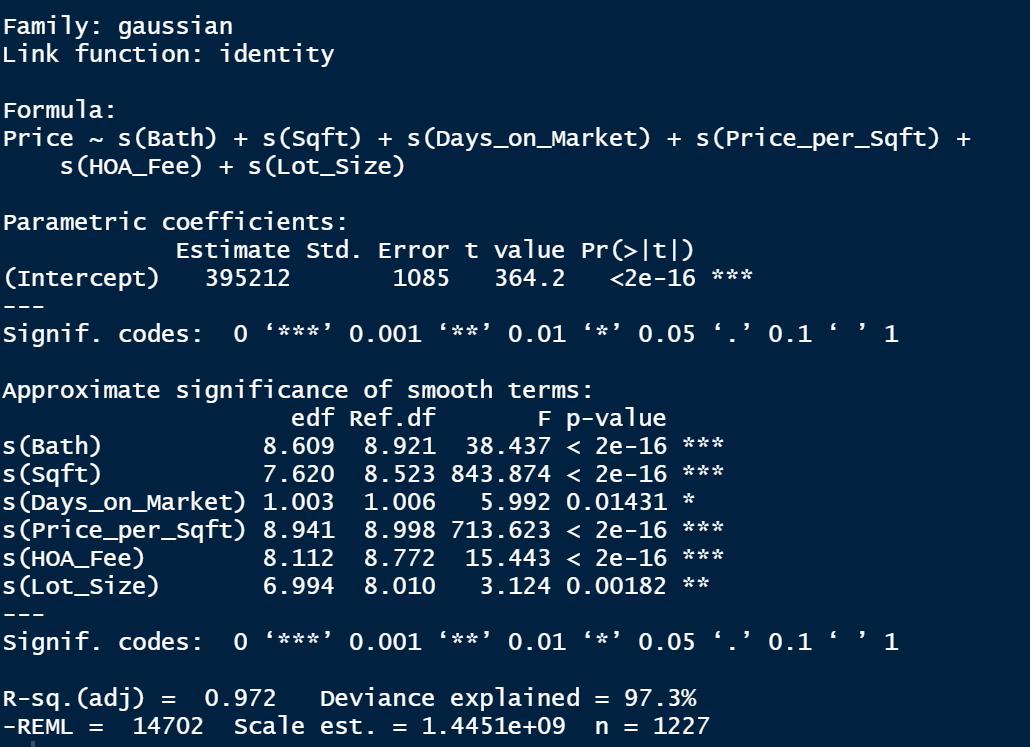


**Generalized Additive Model**

Here we compare the linear regression model to the GAM. Here is the linear model used.

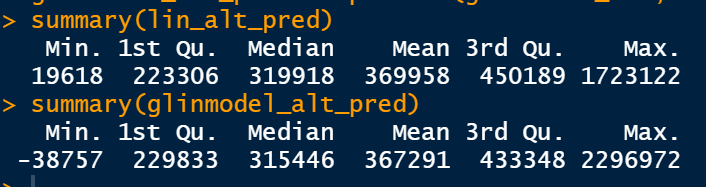


Here is the GAM used.

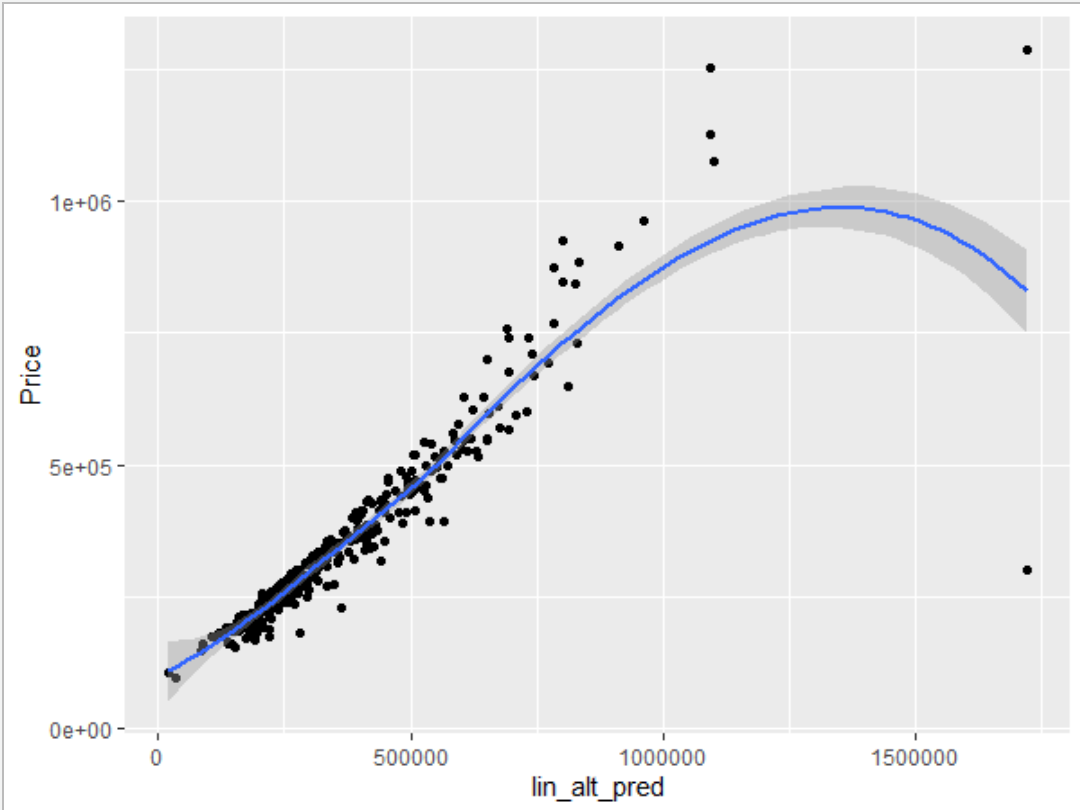


Looking at the linear model we see that if we plot it, they all looks quite similar to the first model even though there are less values. This is shown in the r-squared values is slightly better than the first model at 0.93.

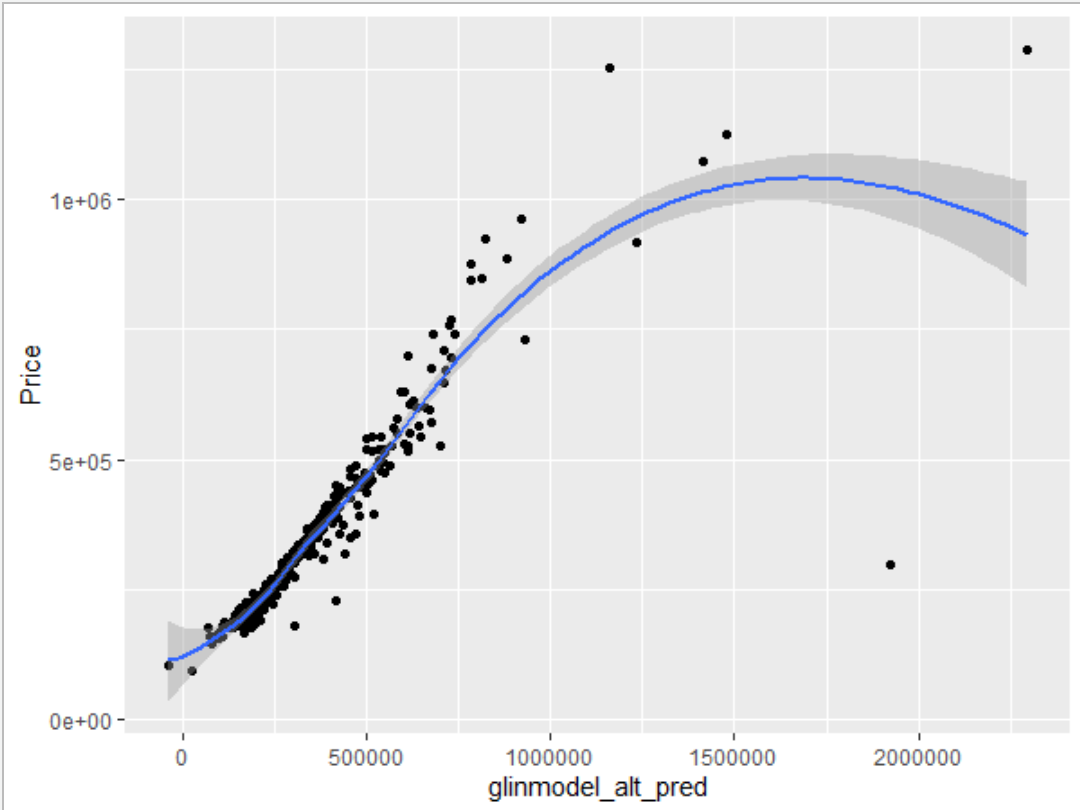
The GAM appears to have done a better job than the linear model with an adjusted r-squared of 0.97 with 97% of the deviance explained. Both summaries of the predicted values are quite close.



If we plot the predictions versus the residuals for the linear model we can see a similar pattern with a strong positive trend with some outliers with high leverage.



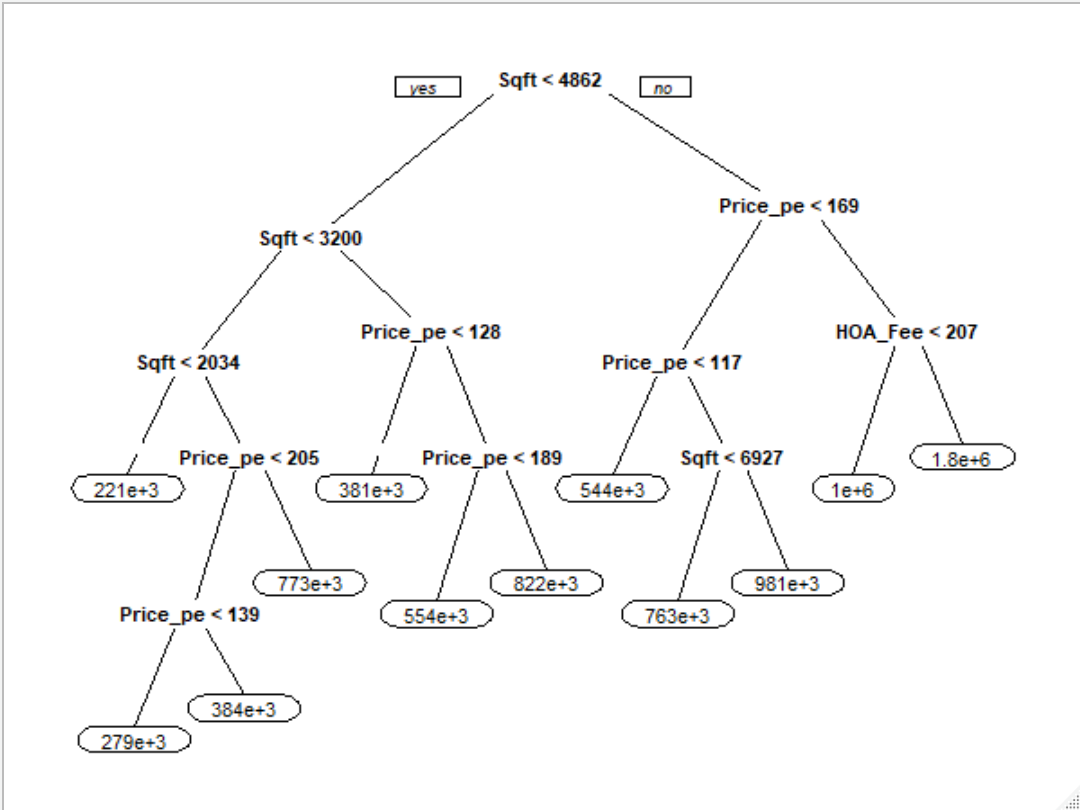
We see something similar when we plot the GAM against the price



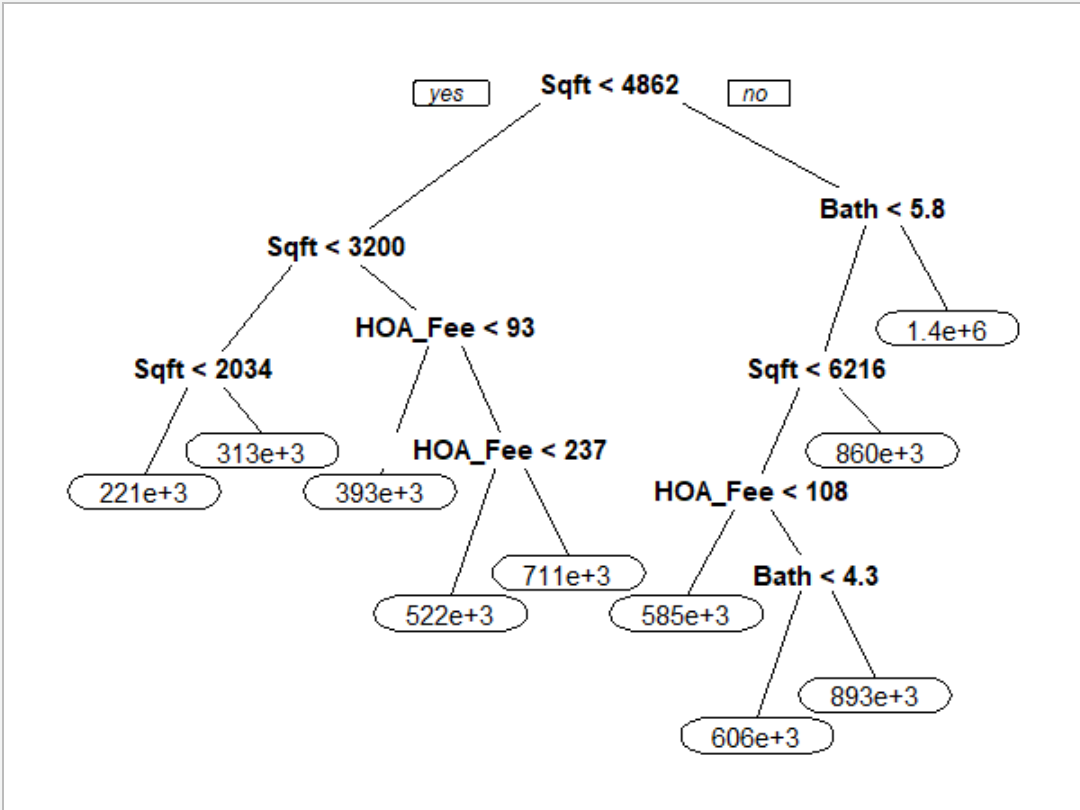
**Advanced Methods**

A couple decision trees were grown on the housing data. After visualizing the data, two trees were grown, one with all the numeric data and one with less than all of the numeric data.

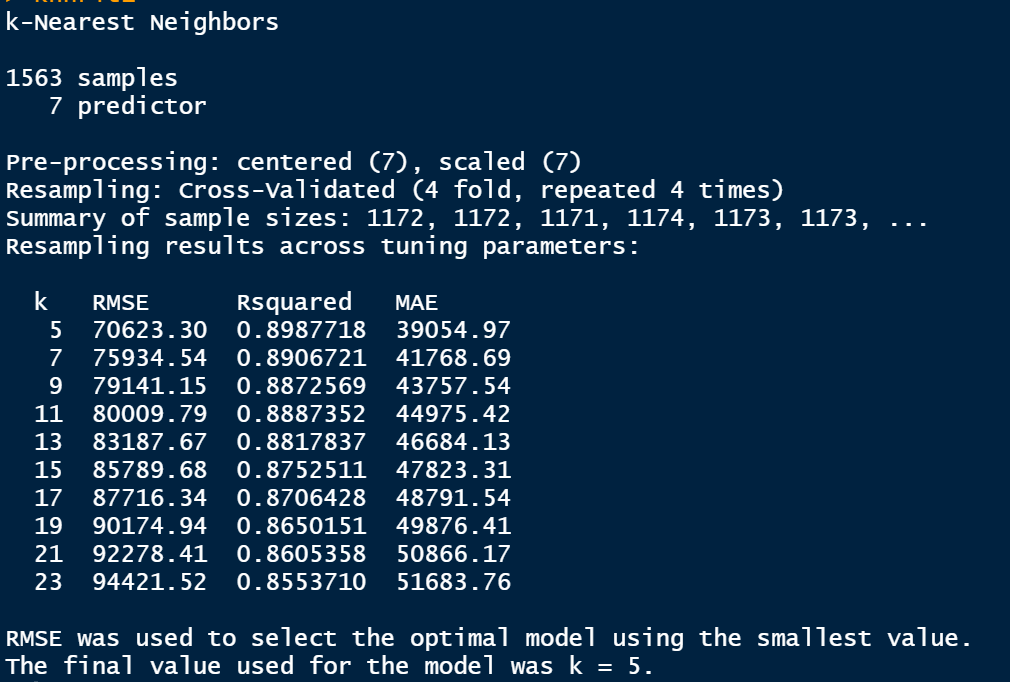
The one grown with all the numeric data determined it was square feet that was the most important and then price. The prediction for this this tree was statistically significant, 1.582e12.

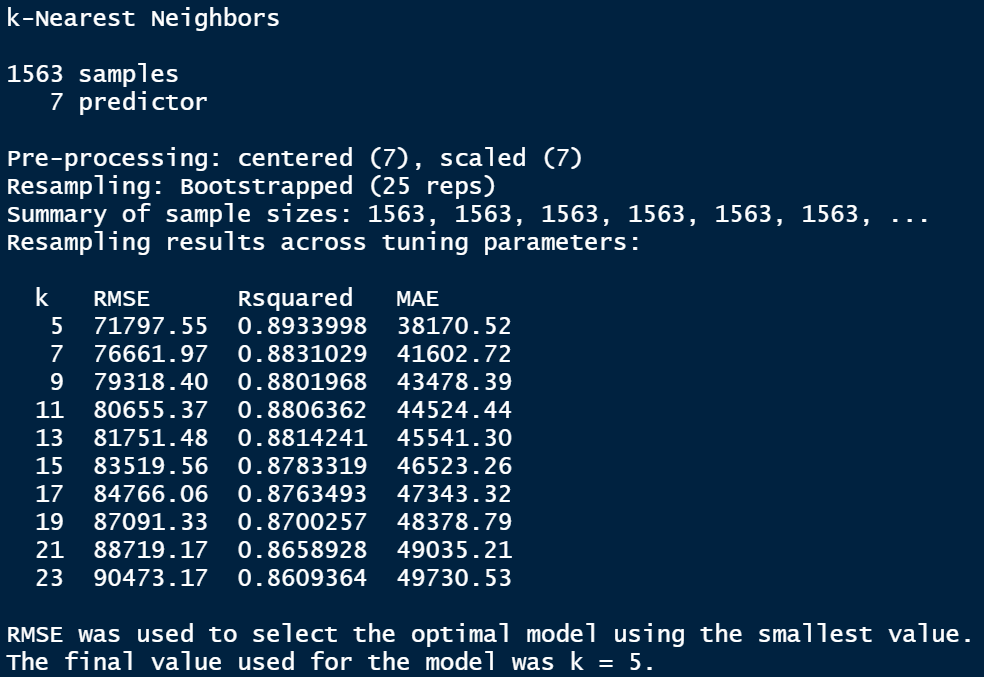


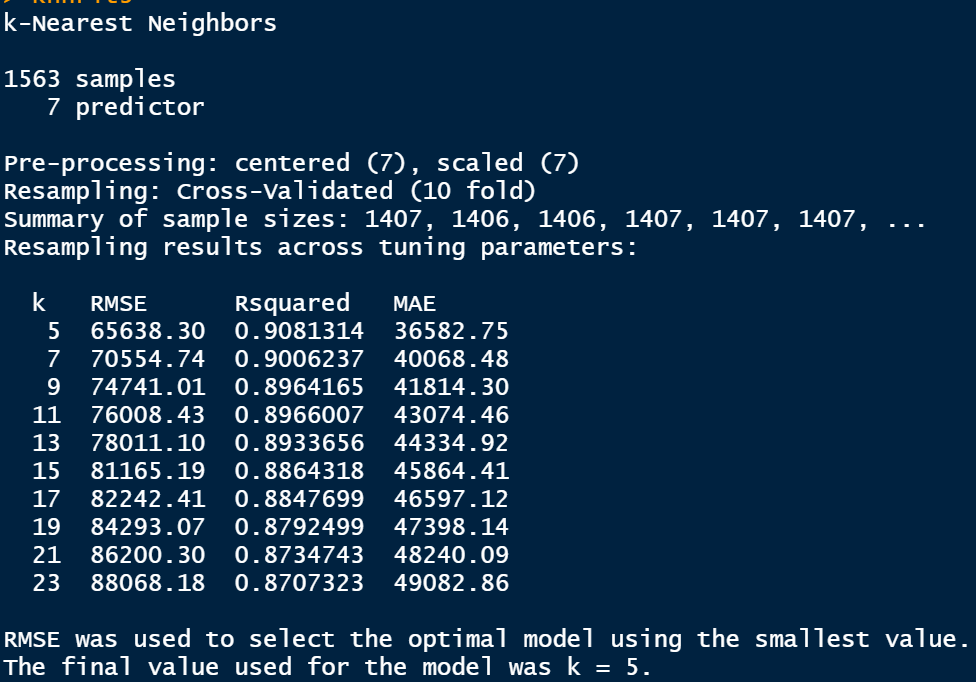
The second tree had less data to use and produced almost the same plot, but a little smaller and this one also used HOA fee and bathroom number. The prediction for the tree was even more significant than the other tree, 4.59e12.



The last advance method uses was KNN. 3 methods were used to determine the best clusters. One of them has a customer control, one control was “boot” (short for bootstrap I assume), and cv (cross-validation).

 KNN custom control

 KNN boot control

 KNN cv control

As you can see the are all quite accurate, but the most accurate was the one that used cross validation as a control. Each of the runs determined that 5 clusters are the best.

**Conclusion**

While I am quite impressed with the analysis performed, I feel that it can be improved. First improvement would be to normalize all the data and transform it obtain a better “randomness” to the plotted data. Second improvement would be to calculated some additional fields such as price derived from square foot and price per square foot, maybe derive acreage from lot size. I would also like to further explore a way to have a model determined what values help sell the house more than others.

I feel that the GAM model was the most accurate of the models used. It was able to explain 97% of the data with an adjusted r-squared of 0.972. Of all the variables used square foot, bath, price per square foot, and HOA fee seem to be the main predictors of a house price. These findings are not something that I would have guessed.

I would have guessed that lot size, bed, bath, square foot, and days on the market would have ultimately determined the price better.