STAT4214 Final Project: Predicting House Prices

Chandler Crescentini

May 13, 2019

Introduction

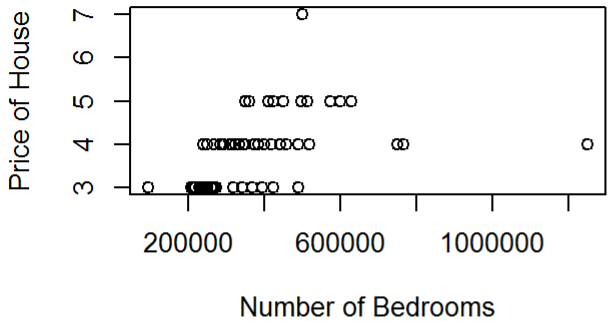
The ability to predict house prices can be advantageous for both buyers and sellers. With an accurate prediction model, buyers can get a sense of whether or not a particular house is a good deal, and sellers can use it as a guideline for setting an initial price when putting a house on the market. The goal of this project is to use publicly available data collected from Realtor.com to create a predictive model for house prices in Hanover County, Virginia.

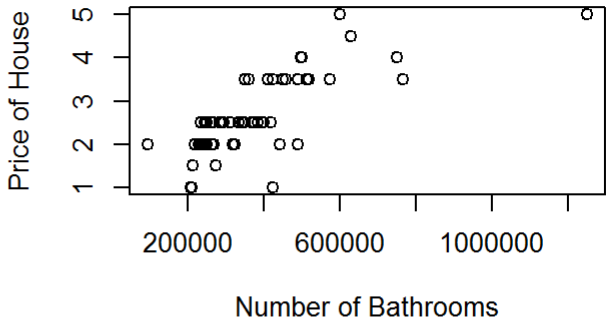
Data Collection

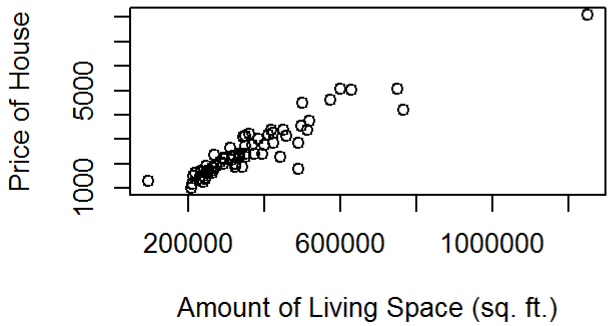
The first step in this, and any other data analysis project, is the collection of data. Here, we collected data on the first 80 single family homes currently listed for sale in Hanover County, Virginia from Realtor.com. We manually collected quantitative data in the form of current price of the home, number of bedrooms, number of bathrooms, amount of living space in square feet, the lot size in acres, the year the house was built, and the total number of rooms. Two categorical varibables, containing four levels each, were also collected: the building exterior type (wood, brick, vinyl, brick/vinyl), and the school district the house is situated in (Atlee, Hanover, Lee Davis, Patrick Henry).

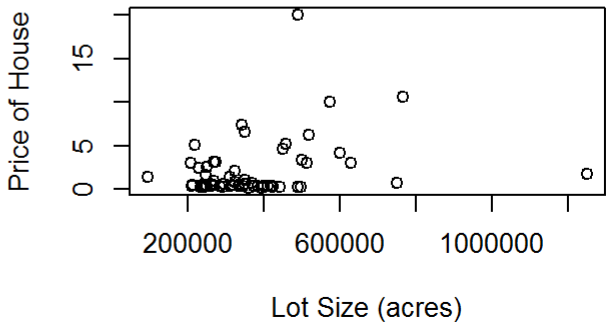
Analysis

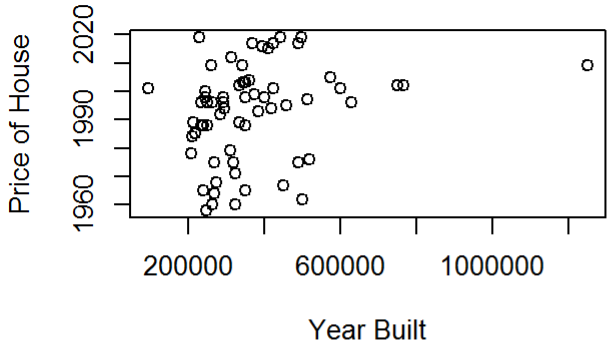
To begin, we take the last 16 observations (20%) of the data, set it aside to serve as our holdout for prediction, and perform some exploratory data analysis on the remaining data. We do this by examining the plots of each individual predictor versus the response in an attempt to identify any relationships that may be present.

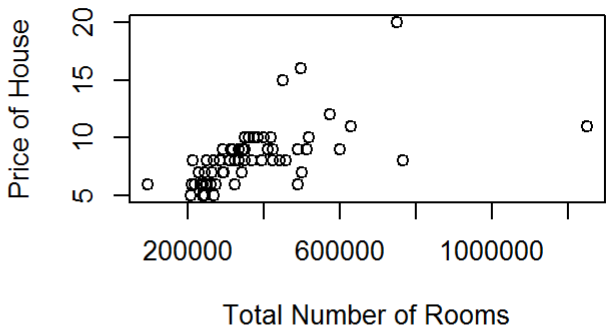

Price of House
3
4
5
6
7
200000
600000
1000000
Number of Bedrooms

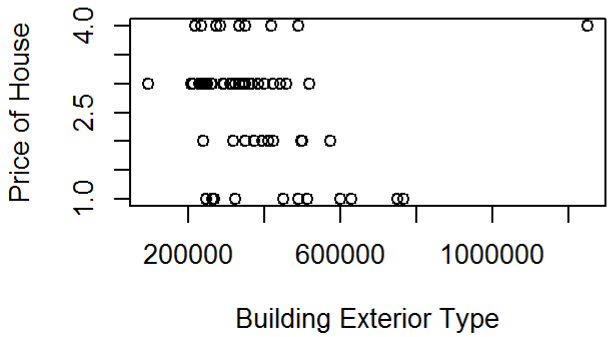

Price of House
1
2
3
4
5
200000
600000
1000000
Number of Bathrooms

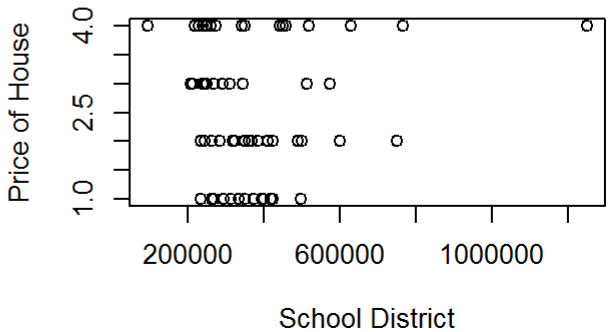

Price of House
1000
5000
200000
600000
1000000
Amount of Living Space (sq. ft.)


Price of House
0
5
15
200000
600000
1000000
Lot Size (acres)


Price of House
1960
1990
2020
200000
600000
1000000
Year Built


Price of House
5
10
15
20
200000
600000
1000000
Total Number of Rooms


Price of House
1.0
2.5
4.0
200000
600000
1000000
Building Exterior Type


Price of House
1.0
2.5
4.0
200000
600000
1000000
School District

Examining the above plots, we see pretty clear linear relationships in the plots of number of bedrooms, bathrooms, total number of rooms, and amount of living space. Lot size and year built also appear to have positive relationships, but lot size has a large number of small observations which may dampen any effects in our regression analysis. The plots of our categorical variables are difficult to interpret so further examination will be required. It is also clear from the above plots that we may have a problematic data point with a price that is nearly 60% larger than the next highest value.

```
## # A tibble: 12 x 3
##    Variables               Tolerance   VIF
##    <chr>                       <dbl> <dbl>
##  1 Beds                        0.347  2.88
##  2 Baths                       0.169  5.91
##  3 LivingSpace                 0.230  4.35
##  4 LotSize                     0.735  1.36
##  5 YearBuilt                   0.601  1.66
##  6 TotRooms                    0.489  2.05
##  7 BuildExtTypeBrick/Vinyl     0.483  2.07
##  8 BuildExtTypeVinyl           0.326  3.06
##  9 BuildExtTypeWood            0.524  1.91
## 10 SchoolDistHanover           0.608  1.65
## 11 SchoolDistLee Davis         0.608  1.64
## 12 SchoolDistPatrick Henry     0.503  1.99
```

After examining the individual predictors, we check for any multicollinearity issues by reporting the variance inflation factors for each predictor, then perform forward, backward, and forward/backward stepwise regression in order to build our model. The VIF values are shown above and we see that all reported values are less than ten, so we conclude that there are no serious multicollinearity issues in the full model. We do notice that there are two variables (Baths and LivingSpace) that have VIF values greater than four. This is not a serious issue, but we will keep this in mind and reexamine once we have arrived at a reduced model.

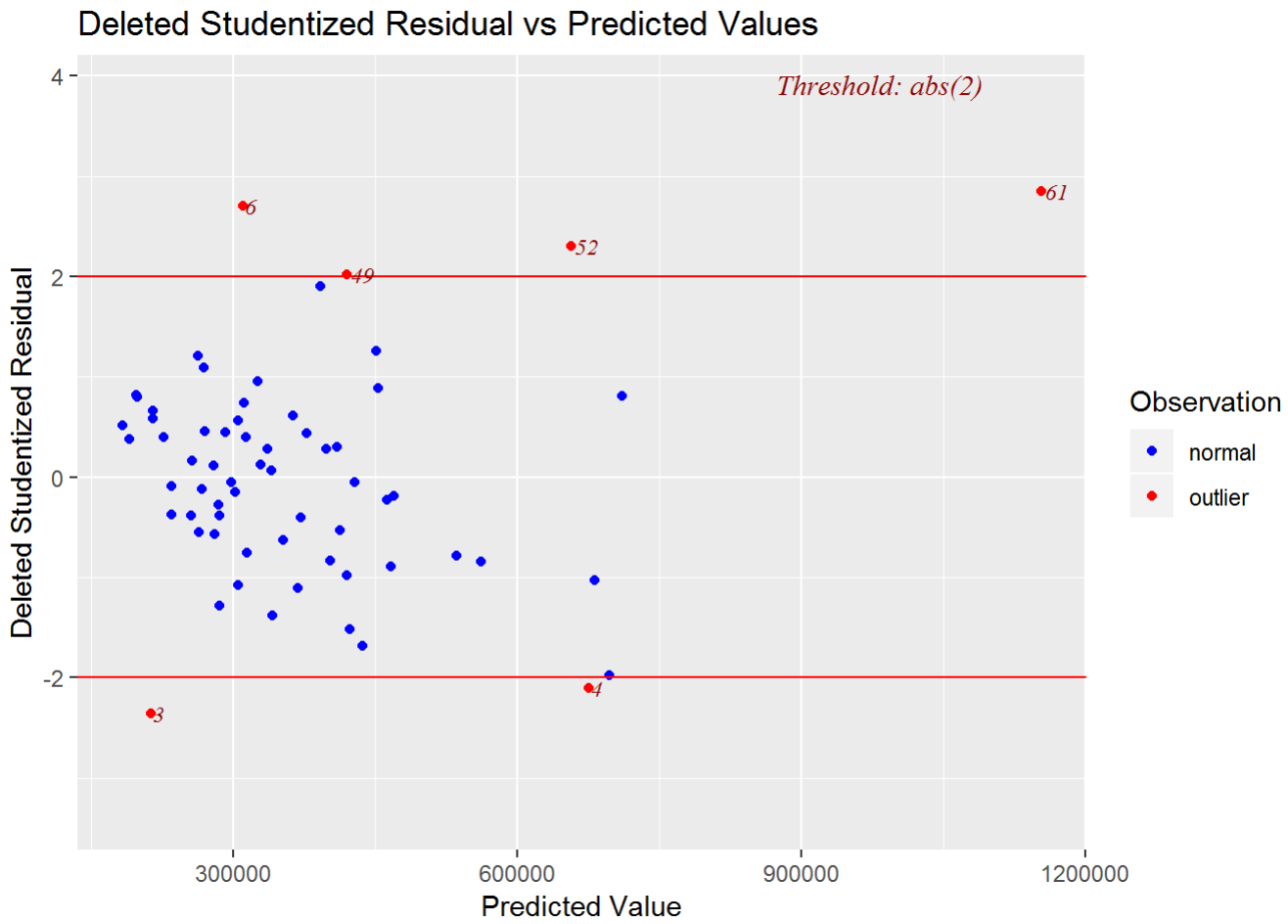
```
## Stepwise Selection Method
## --------------------------
##
## Candidate Terms:
##
## 1. Beds
## 2. Baths
## 3. LivingSpace
## 4. LotSize
## 5. YearBuilt
## 6. TotRooms
## 7. BuildExtType
## 8. SchoolDist
##
## We are selecting variables based on p value...
##
## Variables Entered/Removed:
##
## - LivingSpace added
## - Beds added
## - LotSize added
##
## No more variables to be added/removed.
##
##
## Final Model Output
## ------------------
##
##                             Model Summary
## ------------------------------------------------------------------------
## R                       0.953       RMSE                     52717.489
## R-Squared               0.908       Coef. Var                   14.220
## Adj. R-Squared          0.904       MSE                 2779133644.883
## Pred R-Squared          0.869       MAE                      40618.630
## ------------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                   ANOVA
## --------------------------------------------------------------------------------
##                     Sum of
##                    Squares        DF         Mean Square       F         Sig.
## --------------------------------------------------------------------------------
## Regression    1.652511e+12         3    550837165461.128    198.205    0.0000
## Residual       1.66748e+11        60      2779133644.883
## Total          1.81926e+12        63
## --------------------------------------------------------------------------------
##
##                                       Parameter Estimates
## ----------------------------------------------------------------------------------------
-------
##       model          Beta    Std. Error    Std. Beta      t        Sig          lower
```

```
upper
## ------------------------------------------------------------------------------------------
-------
## (Intercept)    130206.192     32999.099                3.946    0.000     64198.166    1962
14.218
## LivingSpace       143.764         7.065        1.018    20.350    0.000       129.632       1
57.895
##        Beds    -38067.860     10395.747       -0.181    -3.662    0.001    -58862.451    -172
73.270
##     LotSize      7176.852      2063.466        0.138     3.478    0.001      3049.306     113
04.398
## ------------------------------------------------------------------------------------------
-------
```

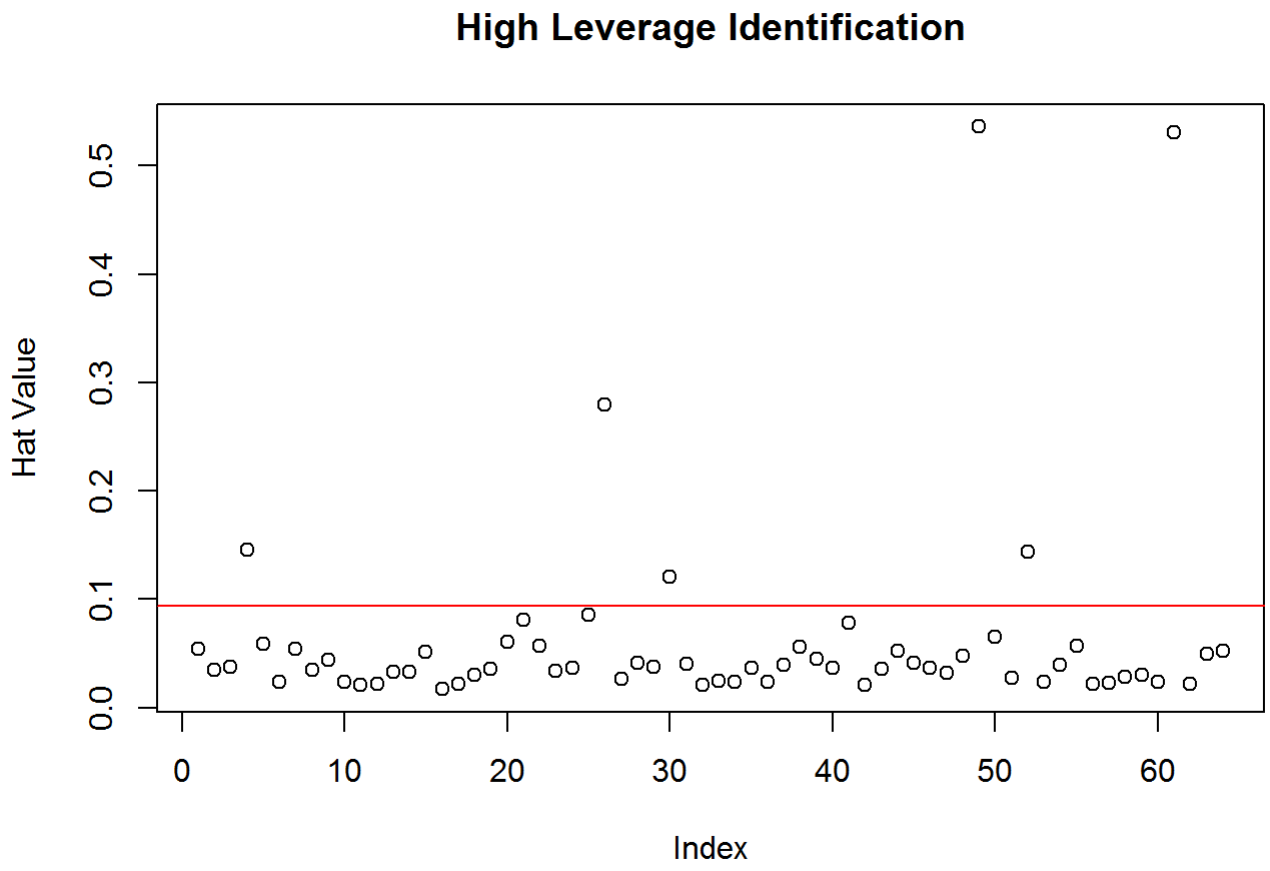
```
##
##                                 Stepwise Selection Summary
## ------------------------------------------------------------------------------------------
##                          Added/                   Adj.
## Step     Variable       Removed     R-Square    R-Square     C(p)         AIC          RMSE
## ------------------------------------------------------------------------------------------
##    1    LivingSpace    addition       0.866       0.864    18.5930    1599.5260    62715.0860
##    2       Beds        addition       0.890       0.886     6.5770    1588.9542    57312.3270
##    3      LotSize      addition       0.908       0.904    -2.2580    1579.1996    52717.4890
## ------------------------------------------------------------------------------------------
```

Each of the three stepwise regression routines return the same three variables (LivingSpace, Beds, and LotSize) with the exact same values returned for each statistic. For succinctness, we only include the output for the forward/backward stepwise regression. We see that the overall F-test, as well as the partial t-tests for each regressor are significant on the $\alpha = 0.05$ level and we report an R^2 and RMSE values of 0.908 and 52717.489 respectively. The R^2 value tells us that 90.8% of the variation in the reponse can be explained by this model, and the RMSE tells us how far from the regression line each of the data points are.

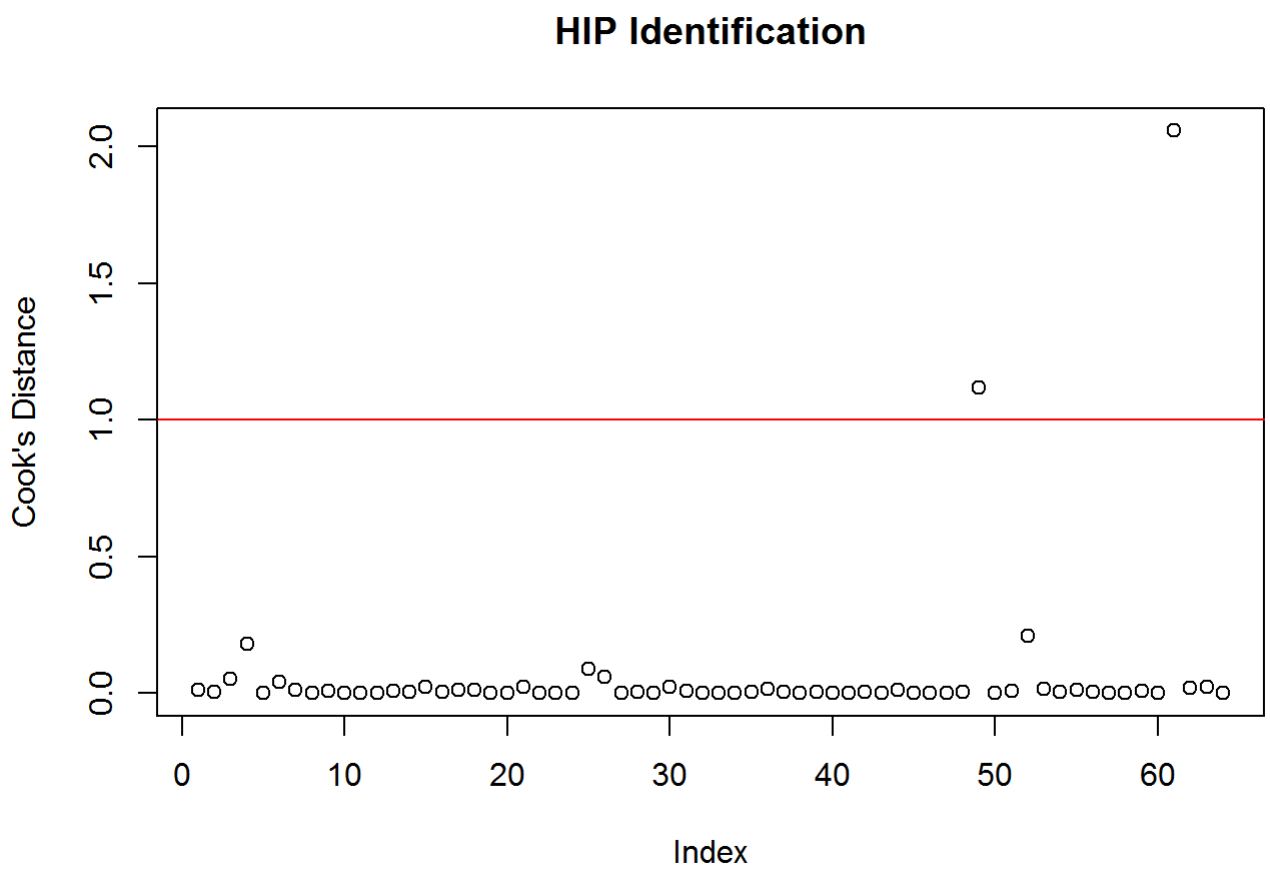
In the context of this project, it does not make sense to interpret the intercept, but in the interest of completeness, the intercept of 130206.192 represents the average price of a home with zero square feet, zero bedrooms, and zero acres of land. The beta value for LivingSpace, 143.764 means that for a one square foot increase in living space and everything else held constant, we expect the price of a home to increase by \$143.76, on average. For the LotSize variable, for a one acre increase in property size with everything else held constant, we expect the price of a home to increase by \$7176.85, on average. Finally, for the Beds variable, for each additional bedroom with everything else held constant, we expect the price of a home to decrease by \$38067.86, on average. This seems to be a very strange result, especially based on the graph of number of bedrooms versus price of home in our exploratory data analysis above. Looking back at this plot, the high leverage point on the far right of the graph might also be highly influential, and unduly affecting the regression result. Next, we perform residual analysis to determine if this is the case.



In order to identify outliers, we examine the deleted studentized residuals versus the predicted values. Looking at the plot, we see the majority of data points far left quarter, and all of the points, except for observation 61, in the left half. Although we have several outliers based on our rule of thumb $|R_i| > 2$, the only point of real concern is observation 61. As a note, if we had used a more conservative rule of thumb $|R_i| > 3$, none of the points would be flagged as outliers.



Next, we examine a plot of the hat values, which measure the remoteness of the predictor value of the i^{th} observation relative to the others, in order to identify any high leverage points in the data. Looking at the plot, we have six data points that are flagged based on our rule of thumb $h_{ii} > \frac{2p}{n} = 0.09375$. Of these, two (49 and 61) are significantly greater than the rest. Next, we will examine whether or not these two points are also identified as highly influential points using Cook's Distance.



Cook's Distance is a measure that represents how all of the predicted values would change if that particular observation were left out of the regression. Based on our rule of thumb $D_i > 1$, we see two data points (49 and 61) are classified as highly influential points. Point 61 is the most egregious, and stands out substantially compared to the rest of the observations. Since we have classified this point as an outlier, a high leverage point, and a highly influential point, we will rerun the regression analysis excluding this point. Although point 49 has also been flagged by all three tests, we will wait and see if removing point 61 fixes its problems before deleting that one as well.

```
## Stepwise Selection Method
## --------------------------
##
## Candidate Terms:
##
## 1. Beds
## 2. Baths
## 3. LivingSpace
## 4. LotSize
## 5. YearBuilt
## 6. TotRooms
## 7. BuildExtType
## 8. SchoolDist
##
## We are selecting variables based on p value...
##
## Variables Entered/Removed:
##
## - LivingSpace added
## - LotSize added
## - YearBuilt added
## - BuildExtType added
## - TotRooms added
## - TotRooms added
##
## No more variables to be added/removed.
##
##
## Final Model Output
## ------------------
##
##                             Model Summary
## ----------------------------------------------------------------------
## R                       0.936       RMSE                    47742.222
## R-Squared               0.877       Coef. Var                  13.381
## Adj. R-Squared          0.863       MSE                2279319789.322
## Pred R-Squared          0.832       MAE                     32636.239
## ----------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                    ANOVA
## --------------------------------------------------------------------------------
##                         Sum of
##                        Squares        DF         Mean Square      F         Sig.
## --------------------------------------------------------------------------------
## Regression    906244124365.893         6    151040687394.315    66.266    0.0000
## Residual      127641908202.044        56      2279319789.322
## Total             1.033886e+12        62
## --------------------------------------------------------------------------------
##
##                                              Parameter Estimates
```

```
## ------------------------------------------------------------------------------------------------
------------------------
##                   model            Beta    Std. Error    Std. Beta      t        Sig
lower          upper
## ------------------------------------------------------------------------------------------------
------------------------
##             (Intercept)    -2445070.913    828327.803                -2.952    0.005    -410
4410.897    -785730.928
##              LivingSpace          98.034         7.385        0.751    13.274    0.000
83.239        112.829
##                  LotSize        8493.160      1992.033        0.217     4.264    0.000
4502.638      12483.682
##                YearBuilt        1294.774       422.242        0.168     3.066    0.003
448.922       2140.626
## BuildExtTypeBrick/Vinyl      -31610.198     21918.648       -0.090    -1.442    0.155      -7
5518.527      12298.130
##       BuildExtTypeVinyl      -55200.442     20372.852       -0.215    -2.710    0.009      -9
6012.168     -14388.715
##        BuildExtTypeWood      -48402.500     23928.219       -0.126    -2.023    0.048      -9
6336.484       -468.517
## ------------------------------------------------------------------------------------------------
------------------------
```

```
##
##                                 Stepwise Selection Summary
## ------------------------------------------------------------------------------------------------
-
##                          Added/                   Adj.
## Step      Variable      Removed     R-Square    R-Square     C(p)         AIC          RMSE
## ------------------------------------------------------------------------------------------------
-
##    1    LivingSpace     addition       0.801       0.797    29.0720    1564.9980    58115.745
1
##    2      LotSize       addition       0.850       0.845     9.1800    1548.9943    50795.839
7
##    3     YearBuilt      addition       0.860       0.853     6.9960    1546.8800    49578.871
0
##    4    BuildExtType    addition       0.877       0.863     1.5650    1544.8359    47742.222
3
##    5      TotRooms      addition       0.880       0.865     2.0000    1545.0025    47478.381
9
##    6      TotRooms      removal        0.877       0.863     1.5650    1544.8359    47742.222
3
## ------------------------------------------------------------------------------------------------
-
```

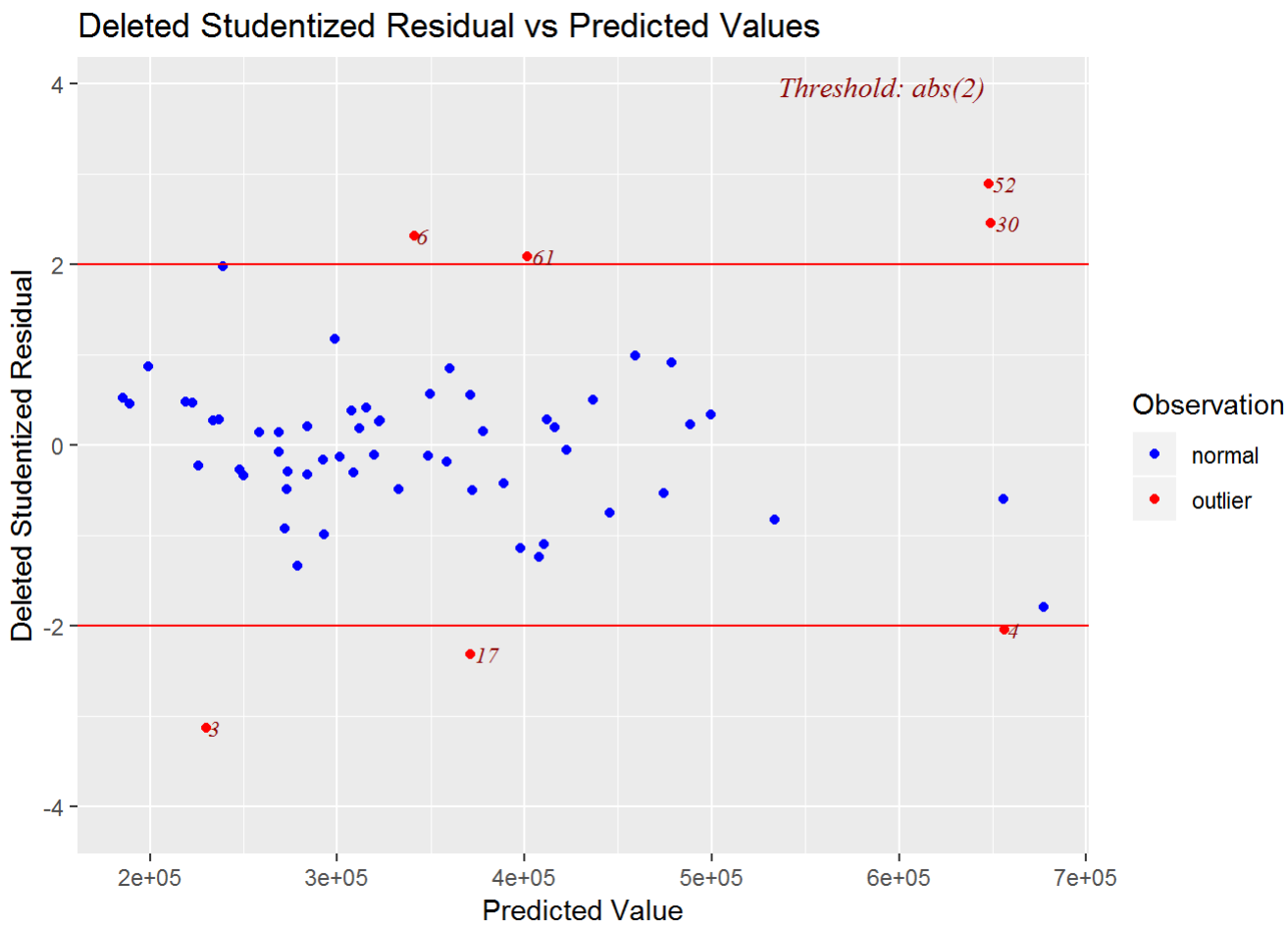
We repeat same process as above for the dataset without observation 61. Once again, for succinctness, we only include the forward/backward stepwise procedure since it is the exact same as the backward procedure, and illustrates the fact that the forward procedure includes the TotRooms variable even though it does not significantly contribute to the model containing all other variables and should be removed. We notice that this model contains

the YearBuilt and BuildExtType variables instead of the Beds variable in the previous model. Although we have a lower R^2_{adj} value for this model (0.863 vs. 0.904), we also have lower AIC (1544.84 vs. 1579.20) and RMSE values (47742.22 vs. 52717.49), indicating this model is an improvement over the previous one.

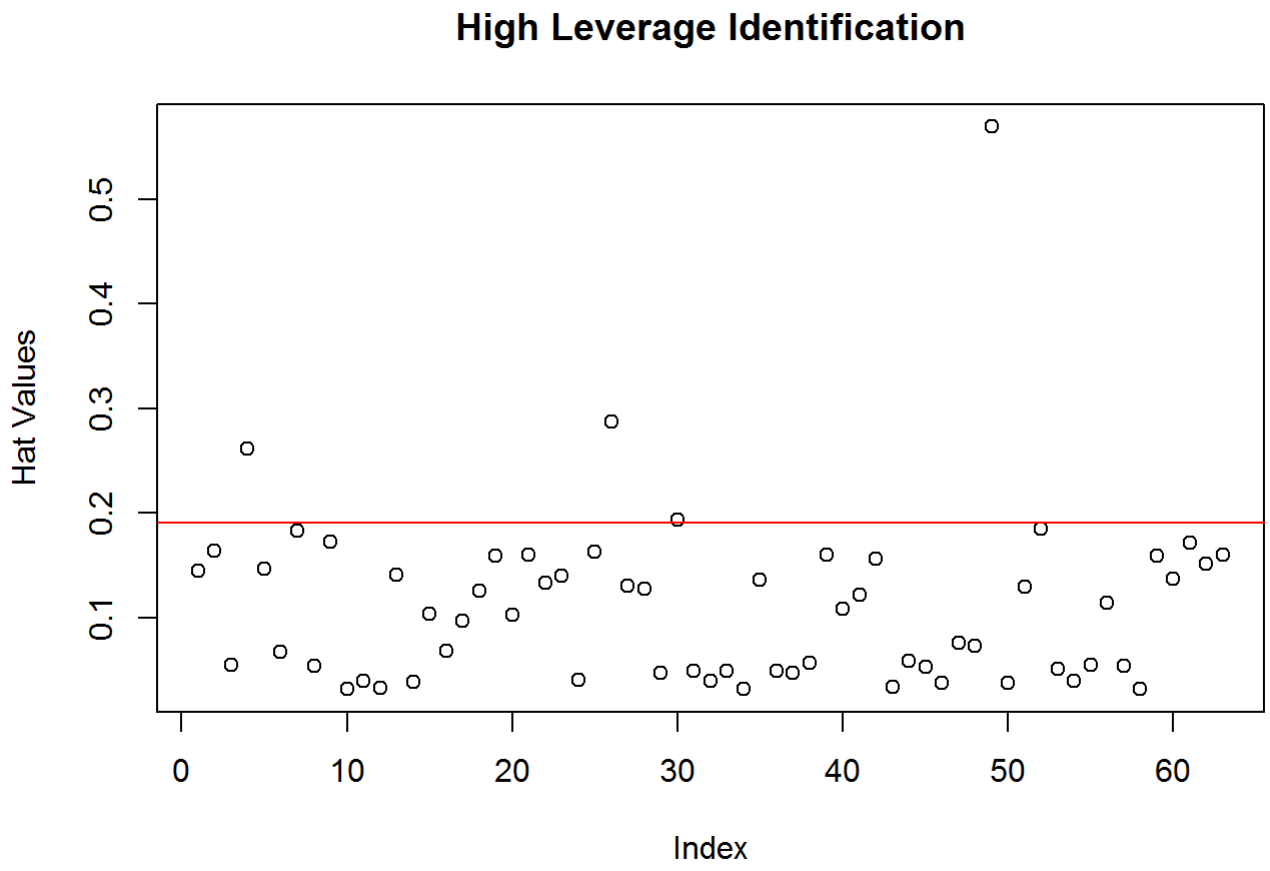
As stated previously, interpreting the intercept does not make sense in the context of this project, but we will do so anyways. Since this model contains one of our categorical variables (BuildExtType), the intercept (β_0) value of -2445070.913 represents the average price of a brick house with zero square feet of living space, zero acres of land, and built in year zero. The beta value for LivingSpace (β_1) means that for each additional square foot of living space, we expect the price of a house will increase by $98.034, on average, provided all other variables are held constant. The beta value for LotSize (β_2) means that for each additional acre of property size, we expect the price of a house will increase by $8493.16, on average, provided all other variables are held constant. The beta value for YearBuilt (β_3) means that for a one year increase in built date, we expect the price of a house to increase by $1294.18, on average, provided all other variables are held constant. For the categorical variable BuildExtType, β_4 represents the mean price for a house with brick/vinyl is $-31610.20 compared to brick, provided all other variables are held constant. β_5 represents the price for a house with a vinyl exterior is $-55200.44 compared to brick, provided all other variables are held constant. Finally, β_5 represents the mean price for a house with wood siding is $-48402.50 compared to a house with brick, provided all other variables are held constant.

```
## # A tibble: 6 x 3
##   Variables               Tolerance   VIF
##   <chr>                       <dbl> <dbl>
## 1 LotSize                     0.852  1.17
## 2 LivingSpace                 0.690  1.45
## 3 YearBuilt                   0.736  1.36
## 4 BuildExtTypeBrick/Vinyl     0.564  1.77
## 5 BuildExtTypeVinyl           0.349  2.86
## 6 BuildExtTypeWood            0.570  1.75
```

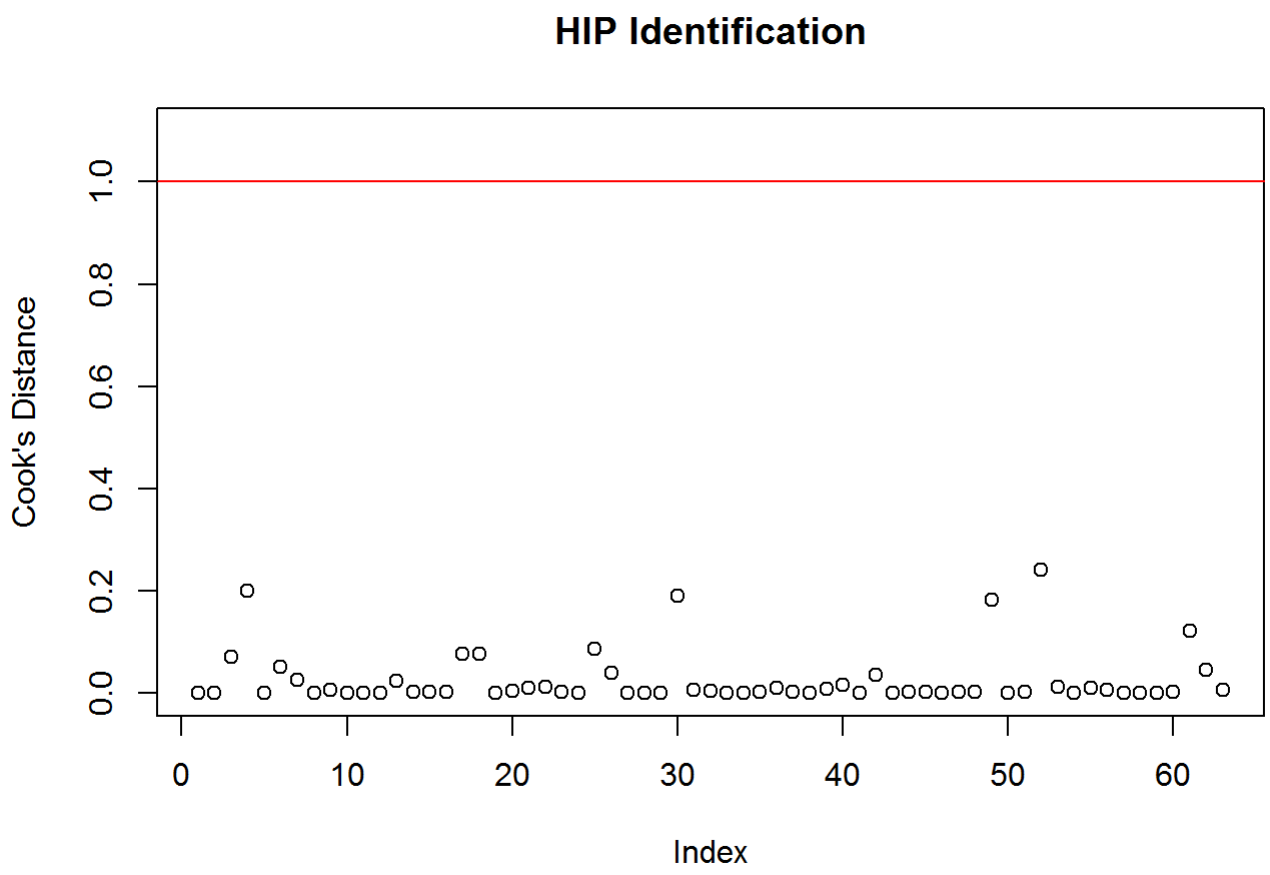
To ensure we do not have any multicollinearity issues we check the VIF values for each of the regressors in our reduced model, and conclude that there are no issues since all values are less than three.



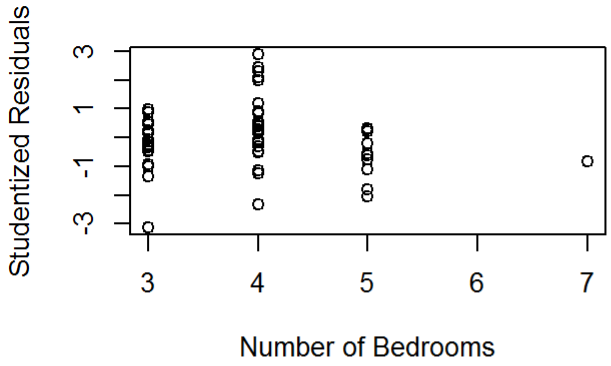
The plot of the deleted studentized residuals versus the predicted values allows us to check the constant variance and linearity assumptions, as well as identify outliers. Looking at the plot, it appears that we have constant variance with all of the data points centered around zero. Although we have some pertubations, we would be reading too much into it to assume that the linearity assumption is violated. There are several outliers based on our rule of thumb stated previously, however, with a slightly more conservative rule of thumb, we eliminate almost all of the outliers.

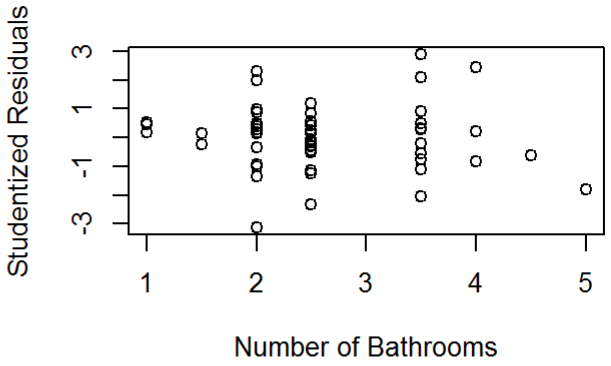


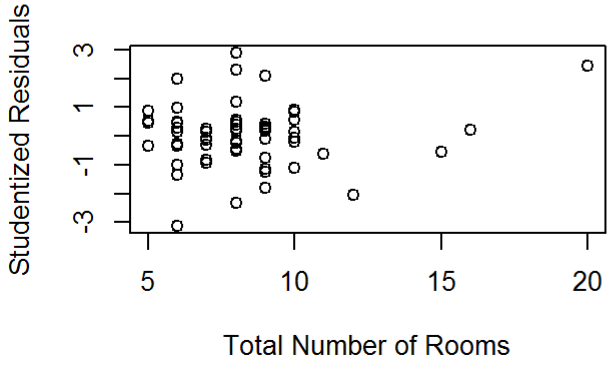
Next, we look at a plot of the hat values in order to identify the high leverage points. Since this model contains more regressors, our rule of thumb cutoff increases in value. Our new model has increased the variance of the hat values, but reduced the amount that fall above the rule of thumb cutoff line. We notice that point 49 still has the potential to be a high influence point, so we examine the Cook's Distance in order to determine if it is or not.



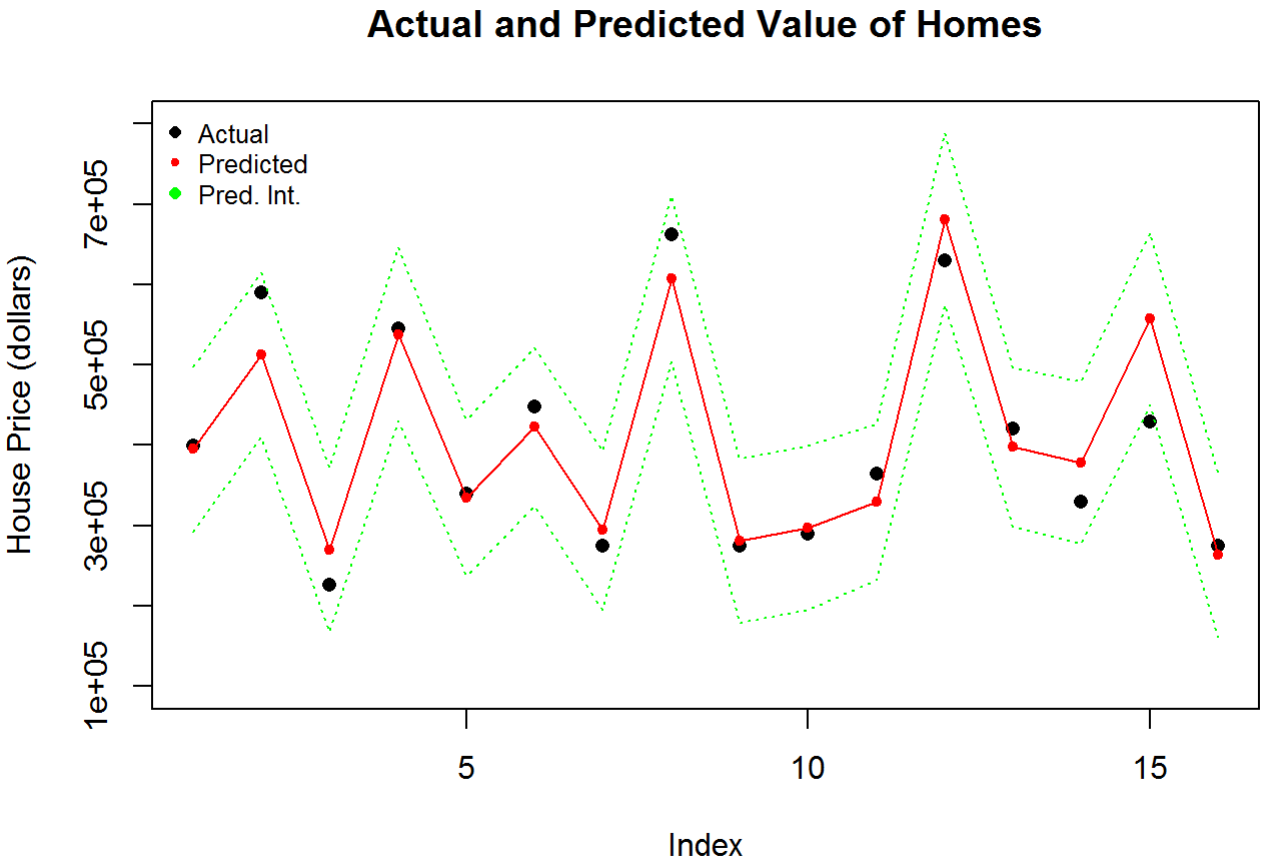
The Cook's Distance plot indicates that we have no highly influential points. Removing data point 61 and fitting a new model has caused our other previously problematic data point to be fixed. Plotting the excluded regressors against the deleted studentized residuals below can help to show whether or not we should include any of them in the model. Since there are no clear trends in any of the plots, we can conclude that none of the excluded regressors should be included in the model. This confirms the stepwise selection process computed above.







Finally, we use our model to predict the 16 properties in the holdout. Below we plot the actual price of the home with a black dot, the predicted price of a home based on our model in red, and provide the prediction interval with a dashed green line. Overall, the model seems to do a pretty decent job of predicting the home price. We predict several of the points almost exactly, are very close on several more, and only one observation falls outside of the prediction interval (point 15).



Conclusion

In conclusion, of the variables collected: number of bedrooms, number of bathrooms, amount of living space in square feet, lot size in acres, year the house was built, total number of rooms, building exterior type, and school district, the most important predictors of price of a single family home in Hanover County, Virginia are amount of living space, lot size, year built, and building exterior type. The RMSE value for the model was found to be 47742.222 with an R^2_{adj} value of 0.863 and an R^2_{pred} value of 0.832. We also saw how a single highly influential point can affect the model to such an extent as to change which variables we include in the model. For our prediction of the 16 holdout datapoints, the RMSE value was found to be 47085.71, which is in line the the RMSE for the rest of the dataset, and only one actual value fell outside of the prediction interval.

In terms of model improvement, there are so many tangible and intangible factors that go into home prices, that it may be impossible to predict them completely accurately based off of publicly available data on the internet. For example, it is impossible to tell from a dataset if the previous owners didn't take care of the house very well or if the house has been on the market for a long time and the price has dropped significantly in that time. Some issues in my model are that the majority of the houses collected for this study were in neightborhoods and thus had very small lot sizes with similar prices compared to others in a more rural area. This caused the overestimation of newer houses with a lot of land. If this project were to be completed again, we might be interested in adding a categorical variable of whether or not a house is in a neighborhood, or a quatitative variable such as commute time to nearest city.