Final Project, STAT/Q Sci 403

Ilse Schmitz

June 04 2025

Introduction

Model Selection

In this case, I decided to use AIC to help select which parameters to use, however this is the wrong way to select features. I chose AIC since the added flexibility is preferable to a smaller dataset. Whether the training data set is small may be up to interpretation, but given it's context in real estate, I've deemed the dataset small given the vast amount of real estate data that exists in the world.

```
##   (Intercept)         price      bedrooms     bathrooms   sqft_living
## -1.012988e+01  3.254803e-07  1.462580e-02  3.442956e-03 -1.786466e-03
##      sqft_lot        floors    waterfront          view     condition
##  2.082069e-06  5.017800e-02  1.511741e-01 -1.160520e-02  2.692756e-02
##         grade    sqft_above sqft_basement      yr_built  yr_renovated
##  6.211073e-02  1.730939e-03  1.797815e-03 -1.014626e-03  3.562950e-05
##       zipcode sqft_living15    sqft_lot15
##  1.720314e-04  7.498896e-05 -2.298239e-06
```

Based on the information here, and using AIC as our criteria, from forward selection we end up with 11 predictors, (grade, yr_built, sqft_living15, sqft_lot15, floors, bedrooms, sqft_lot, condition, yr_renovated, zipcode, sqft_above)

and with backward selection we end up with 13 predictors, (grade, yr_built, sqft_living15, sqft_lot15, floors, bedrooms, sqft_lot, condition, yr_renovated, zipcode, sqft_living, view, sqft_basement)

However, as stated previously, this is the wrong way to select the variables in this case. As such, we use prior knowledge to estimate which features to use in our regression and prediction tasks.

By constructing a series of graphs comparing the log10price and price to all other features, I selected features to utilize based on appearence and whether there appeared to be a relationship between the two. the features I chose to keep were:

sqft_living15

sqft_above

grade

floors

sqft_living

bathrooms

bedrooms

The features I chose to discard were:

sqft_lot15

zipcode

yr_renovated

yr_built

condition

waterfront

sqft_lot

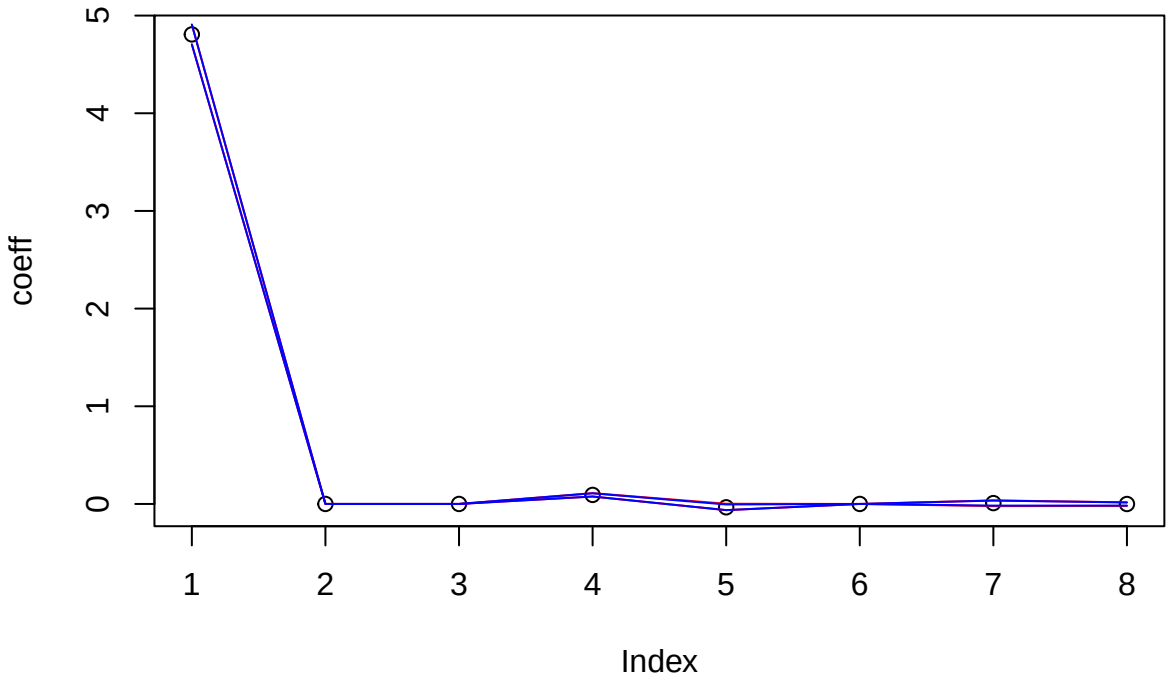
sqft_basement

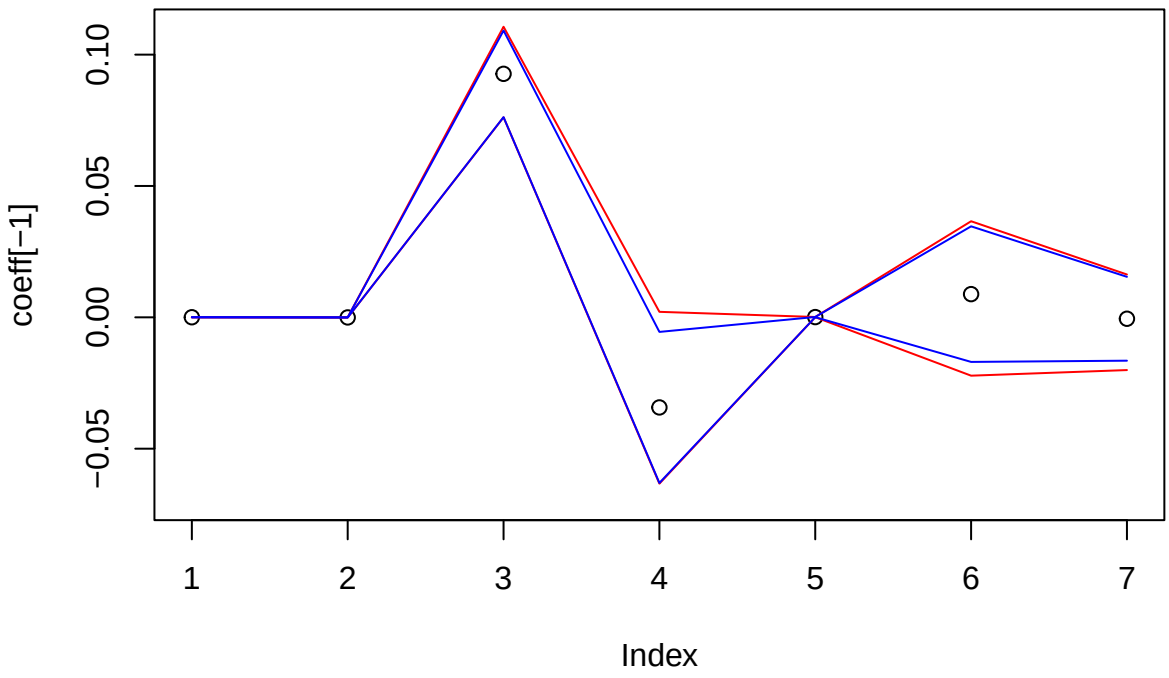
view

These graphs can be found at the bottom of this report, in order to not impede reading, as there are 32 in total.

Regression Task

```
##       5%      95%
## 4.704351 4.906639
##            5%           95%
## -1.604774e-06  6.162290e-05
##            5%           95%
## -7.730599e-05  5.372580e-07
##        5%       95%
## 0.0761488 0.1106130
##           5%          95%
## -0.063306602  0.002099521
##           5%          95%
## 5.545086e-05 1.382196e-04
##          5%         95%
## -0.02220956  0.03657248
##          5%         95%
## -0.02007013  0.01630754
```

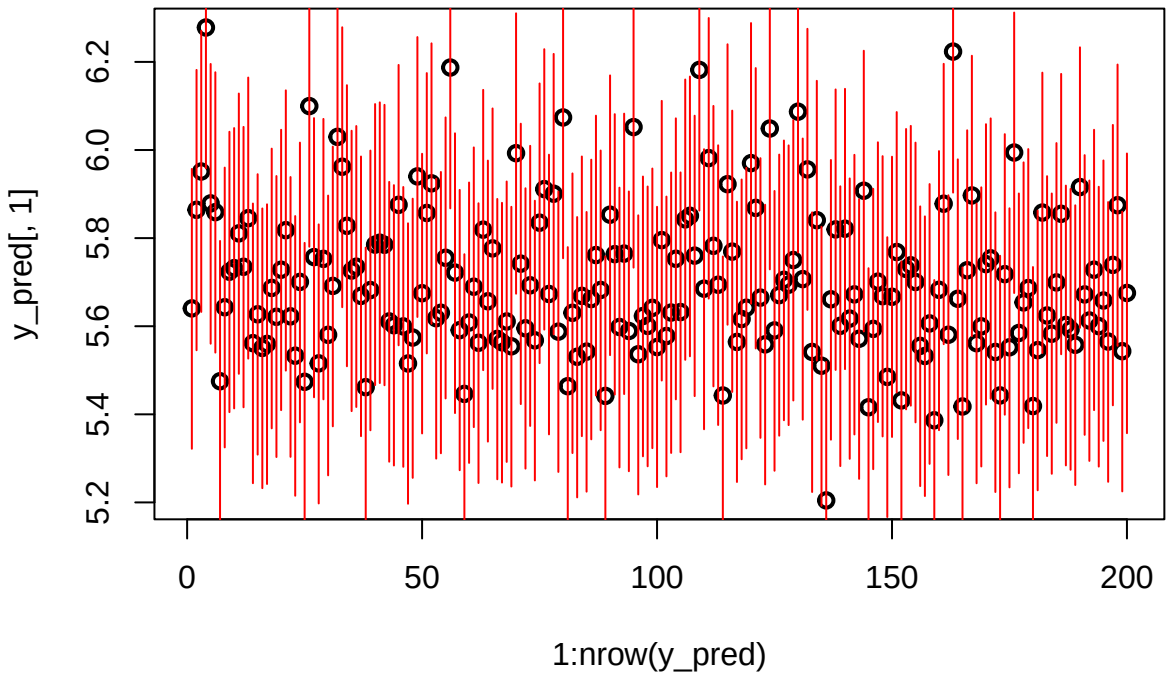




For computing the Gaussian interval, two plots have been provided, one including the intercept, and one without, in order to better interpret the differences between the confidence intervals and the regression coefficients. Blue is the Gaussian confidence intervals, and red is the bootstrap. Based on these graphs, we can see the bedrooms and floors features seem to have the most variation, with grade and floors second. All three features dealing with square footage have shockingly small confidence intervals, implying very high correlation between price and square footage of at least the three different types included within this model.

For the Bootstrap model, our confidence intervals follow very similar patterns, with the only notable exception being the interval of the bedroom and bathroom coefficients being distinctly larger. Additionally, much like the Gaussian CI's, there is a strong confidence in the square footage features and their ability to predict price, which lends credence to those variables being the best indicators.

Conformal Prediction (Jackknife+) Task



For the Guess.dat, the MSE for test_1 was based on the MSE of the training data when sent through the regression model, using both the price and the log10price. Based on prior experience, if the model has been correctly set up test sets usually score a higher MSE than ther training counterparts, but it does tend to be close. As such, my guess for the MSE of the log10price is 0.9, rounding up the existing MSE from the training data.

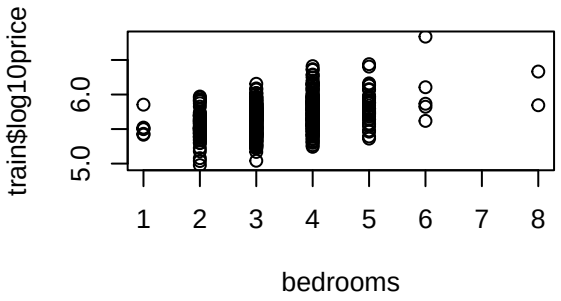
```
## [1] 57726344874
```

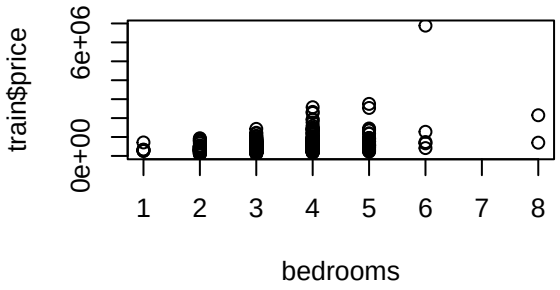
```
## [1] 0.08743939
```

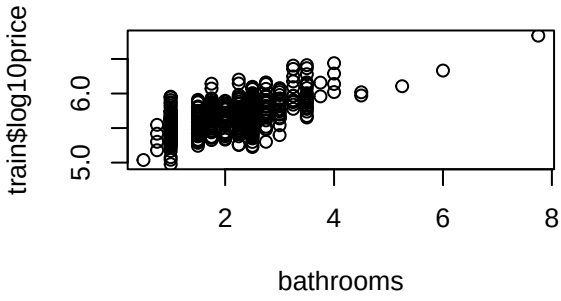
We know the MSE of a test set is very likely to be larger than the MSE of the training set, and based on this, I would assume the MSE is around 60,000,000,000, close to the MSE of the training set, but slightly larger. For the confidence interval, We can use the coverage of the test_2 set as a basis for the true coverage. I would suspect there will be some variability, but for the sake of a clean guess, I would be the true coverage would be around 0.935, to match the existing prediction data.

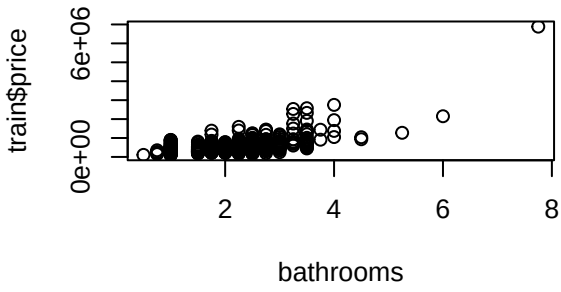
```
## [1] 0.935
```

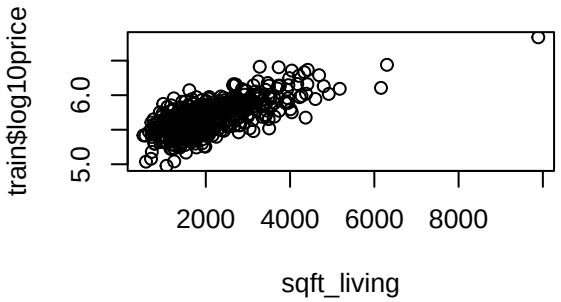
Additional graphs

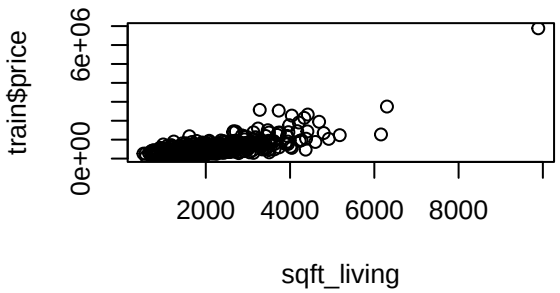


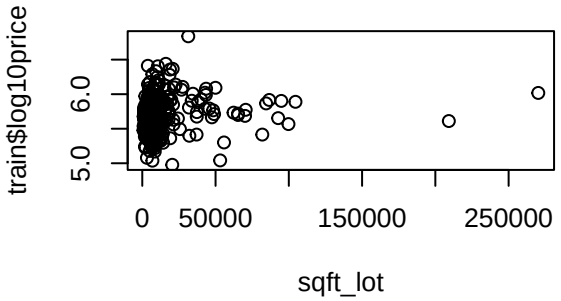


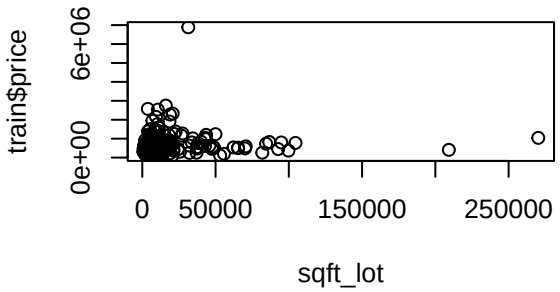


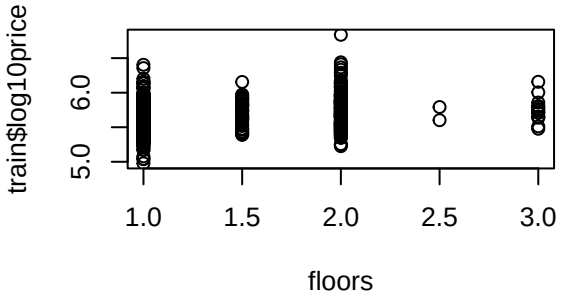


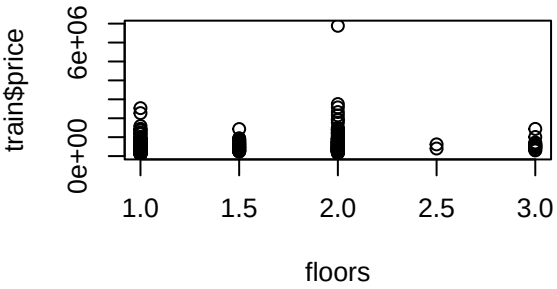


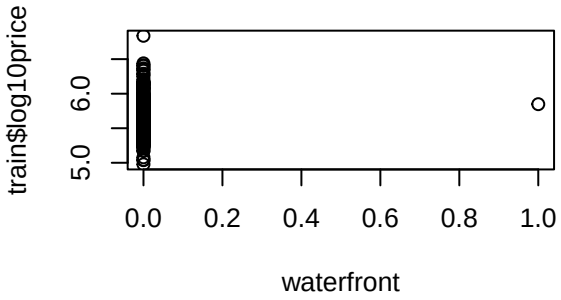


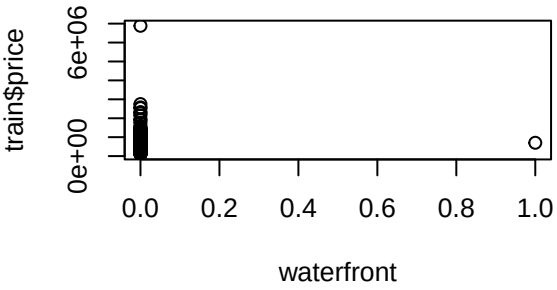


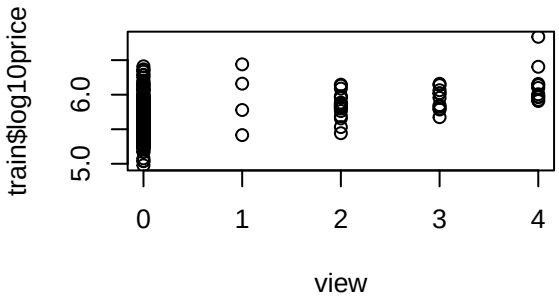


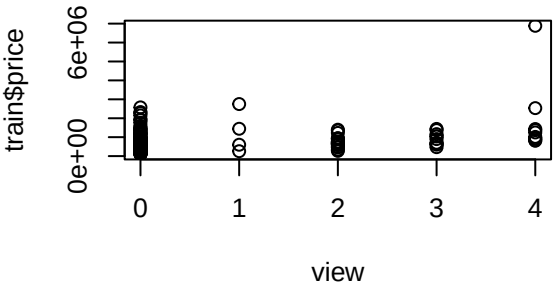

train$log10price
5.0
6.0
1.0
1.5
2.0
2.5
3.0
floors

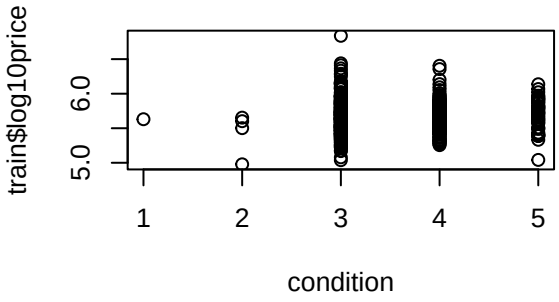

train$price
0e+00
6e+06
1.0
1.5
2.0
2.5
3.0
floors

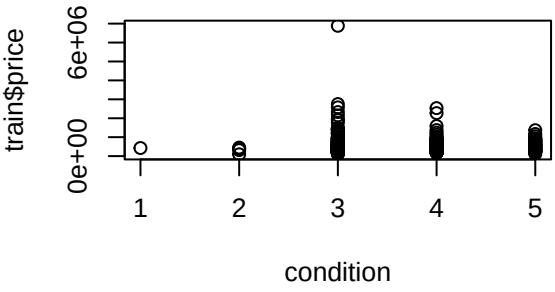

train$log10price
5.0
6.0
0.0
0.2
0.4
0.6
0.8
1.0
waterfront

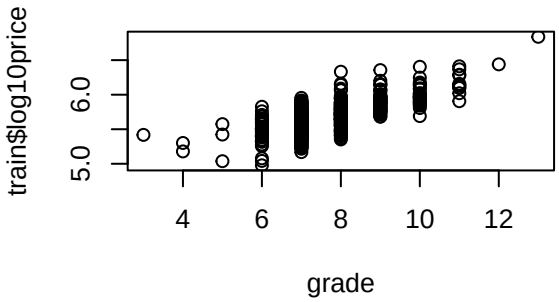

train$price
0e+00
6e+06
0.0
0.2
0.4
0.6
0.8
1.0
waterfront

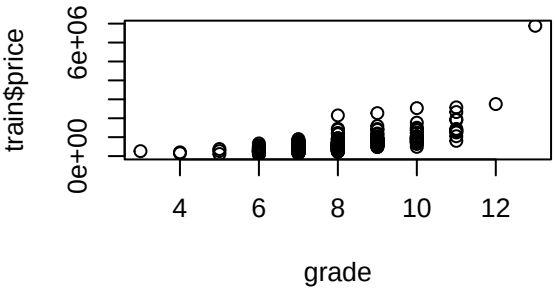

train$log10price
5.0
6.0
0
1
2
3
4
view

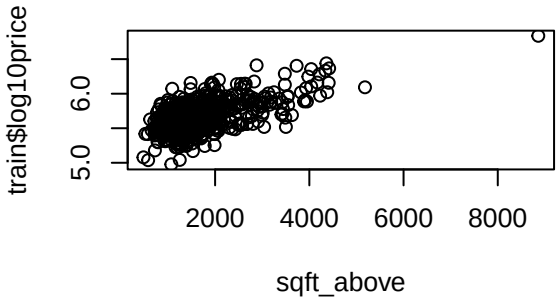

train$price
0e+00
6e+06
0
1
2
3
4
view

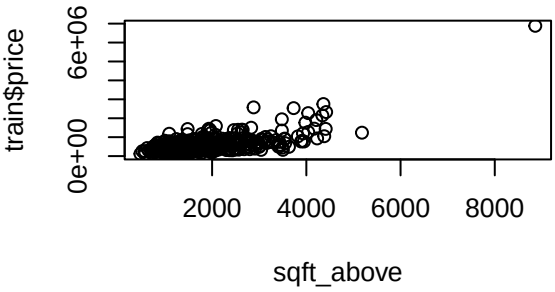

train$log10price
5.0
6.0
1
2
3
4
5
condition

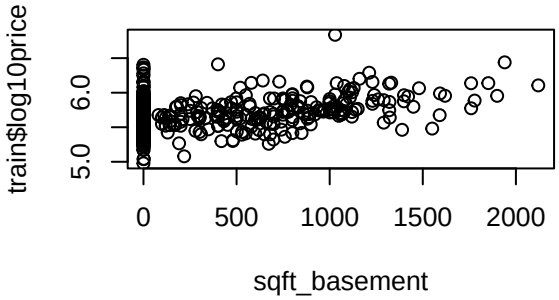

train$price
0e+00
6e+06
1
2
3
4
5
condition

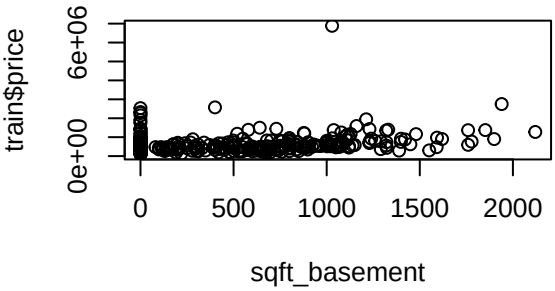

train$log10price
5.0
6.0
4
6
8
10
12
grade

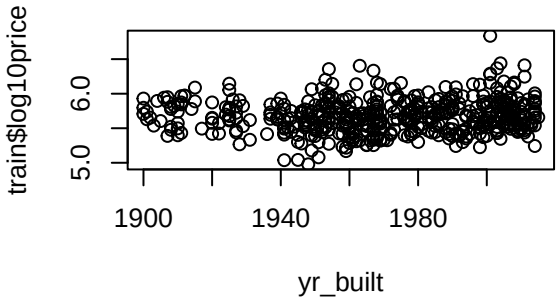

train$price
0e+00
6e+06
4
6
8
10
12
grade

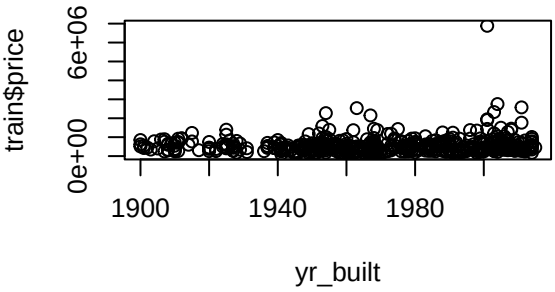

train$log10price
5.0
6.0
2000
4000
6000
8000
sqft_above

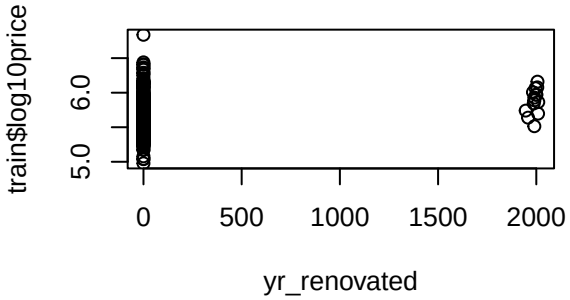

train$price
0e+00
6e+06
2000
4000
6000
8000
sqft_above

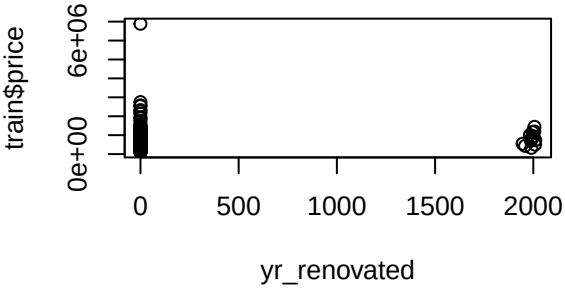

train$log10price
5.0
6.0
0
500
1000
1500
2000
sqft_basement

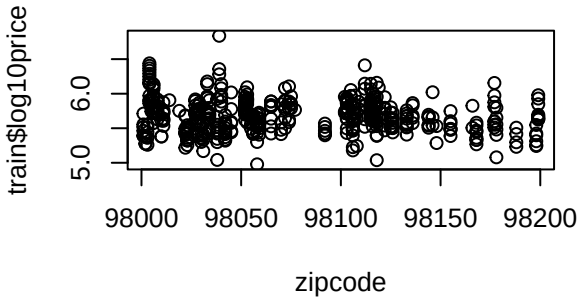

train$price
0e+00
6e+06
0
500
1000
1500
2000
sqft_basement

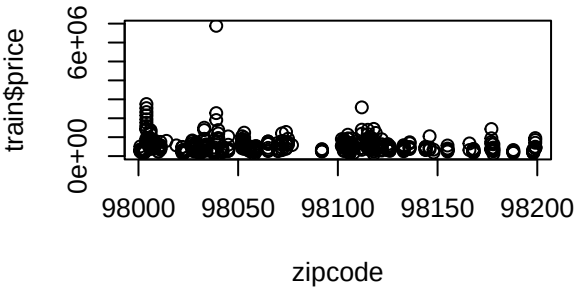

train$log10price
5.0
6.0
1900
1940
1980
yr_built

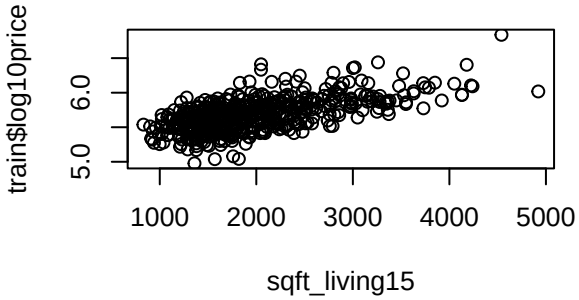

train$price
0e+00
6e+06
1900
1940
1980
yr_built

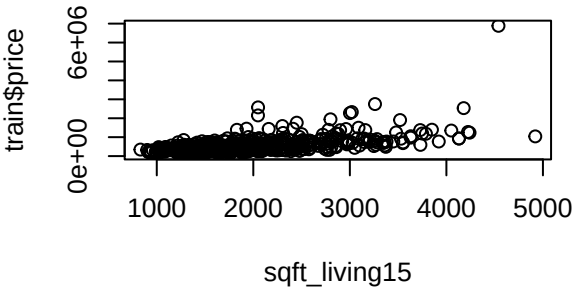

train$log10price
5.0
6.0
0
500
1000
1500
2000
yr_renovated

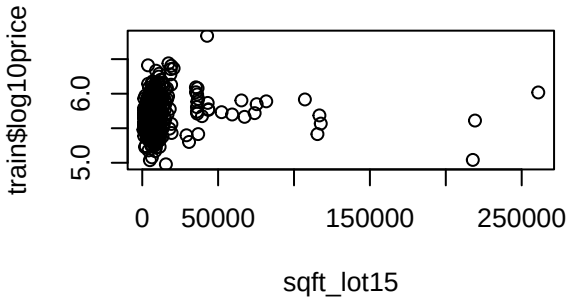

train$price
0e+00
6e+06
0
500
1000
1500
2000
yr_renovated

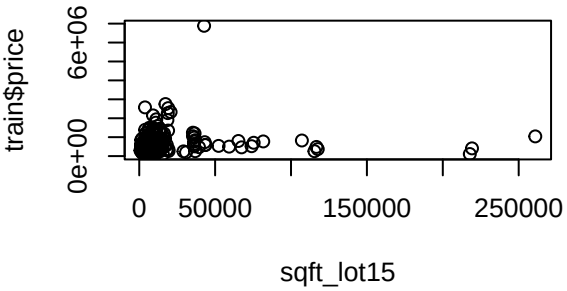

train$log10price
5.0
6.0
98000
98050
98100
98150
98200
zipcode


train$price
0e+00
6e+06
98000
98050
98100
98150
98200
zipcode


train$log10price
5.0
6.0
1000
2000
3000
4000
5000
sqft_living15


train$price
0e+00
6e+06
1000
2000
3000
4000
5000
sqft_living15


train$log10price
5.0
6.0
0
50000
150000
250000
sqft_lot15


train$price
0e+00
6e+06
0
50000
150000
250000
sqft_lot15