Lab02

Amanda Mentesana

2025-02-07

Set up libraries and read dataset.

```
knitr::opts_chunk$set(echo = FALSE)

#install libraries
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.4.2
```

```
library(EnvStats)
```

```
## Warning: package 'EnvStats' was built under R version 4.4.2
```

```
##
## Attaching package: 'EnvStats'
```

```
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

```
library("ggplot2")
```

```
## Warning: package 'ggplot2' was built under R version 4.4.2
```

```
#read the dataset
#since we are making multiple models, we need to filter differently for different variables used
unfiltered_dataset <- read_csv("C:/Users/amanda/Downloads/NY-House-Dataset.csv")
```

```
## Rows: 4801 Columns: 17
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (11): BROKERTITLE, TYPE, ADDRESS, STATE, MAIN_ADDRESS, ADMINISTRATIVE_AR...
## dbl  (6): PRICE, BEDS, BATH, PROPERTYSQFT, LATITUDE, LONGITUDE
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Linear Model 1

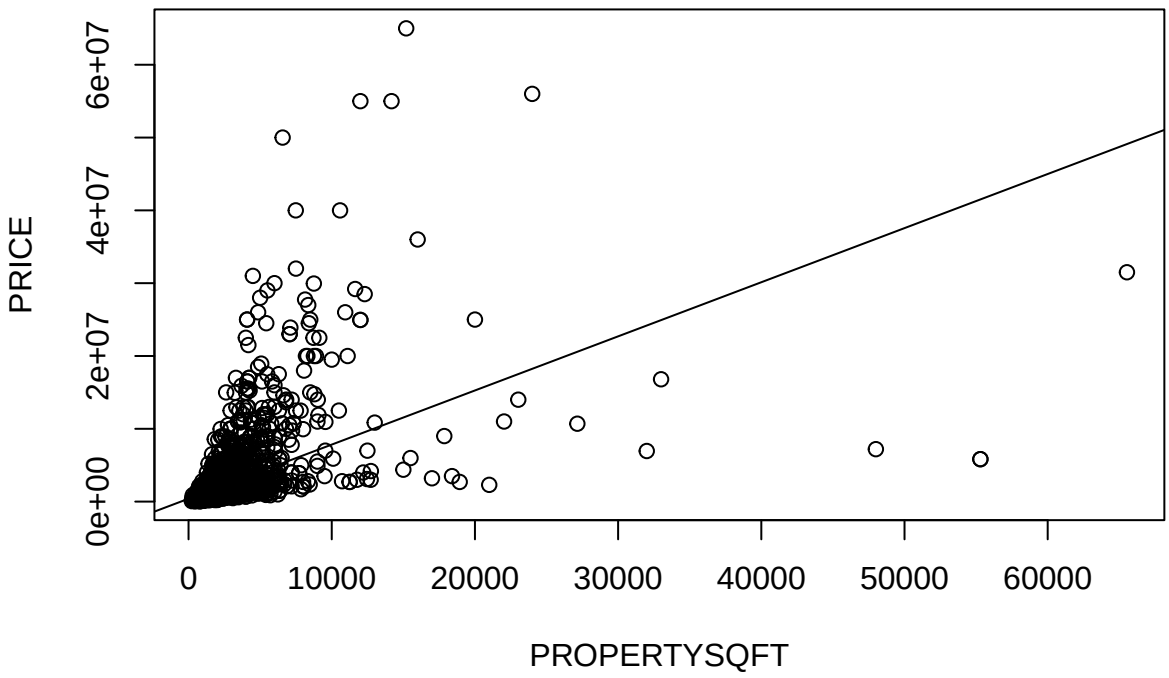
Using PROPERTYSQFT as a predictor for PRICE

From the following, we can see that the data must be log-fit in order to be interpretable

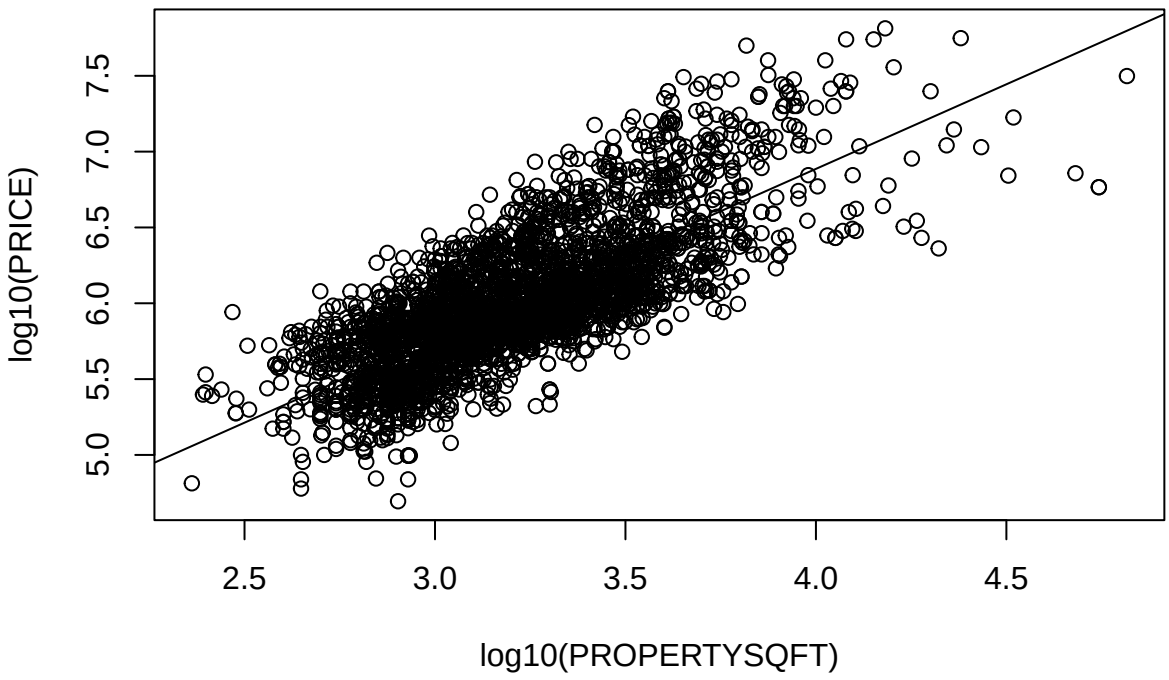
```
## [1] NA
```

```
##
## Call:
## lm(formula = PRICE ~ PROPERTYSQFT, data = dataset1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -35673821   -906581   -633994   -199729  53305104
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  396889.92   79835.49   4.971    7e-07 ***
## PROPERTYSQFT    743.29      21.99  33.801   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3602000 on 3176 degrees of freedom
## Multiple R-squared:  0.2646, Adjusted R-squared:  0.2643
## F-statistic:  1142 on 1 and 3176 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = log10(PRICE) ~ log10(PROPERTYSQFT), data = dataset1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96783 -0.19911 -0.05041  0.19131  1.01634
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.42346    0.05386   44.99   <2e-16 ***
## log10(PROPERTYSQFT)  1.11570    0.01673   66.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2887 on 3176 degrees of freedom
## Multiple R-squared:  0.5835, Adjusted R-squared:  0.5833
## F-statistic:  4449 on 1 and 3176 DF,  p-value: < 2.2e-16
```

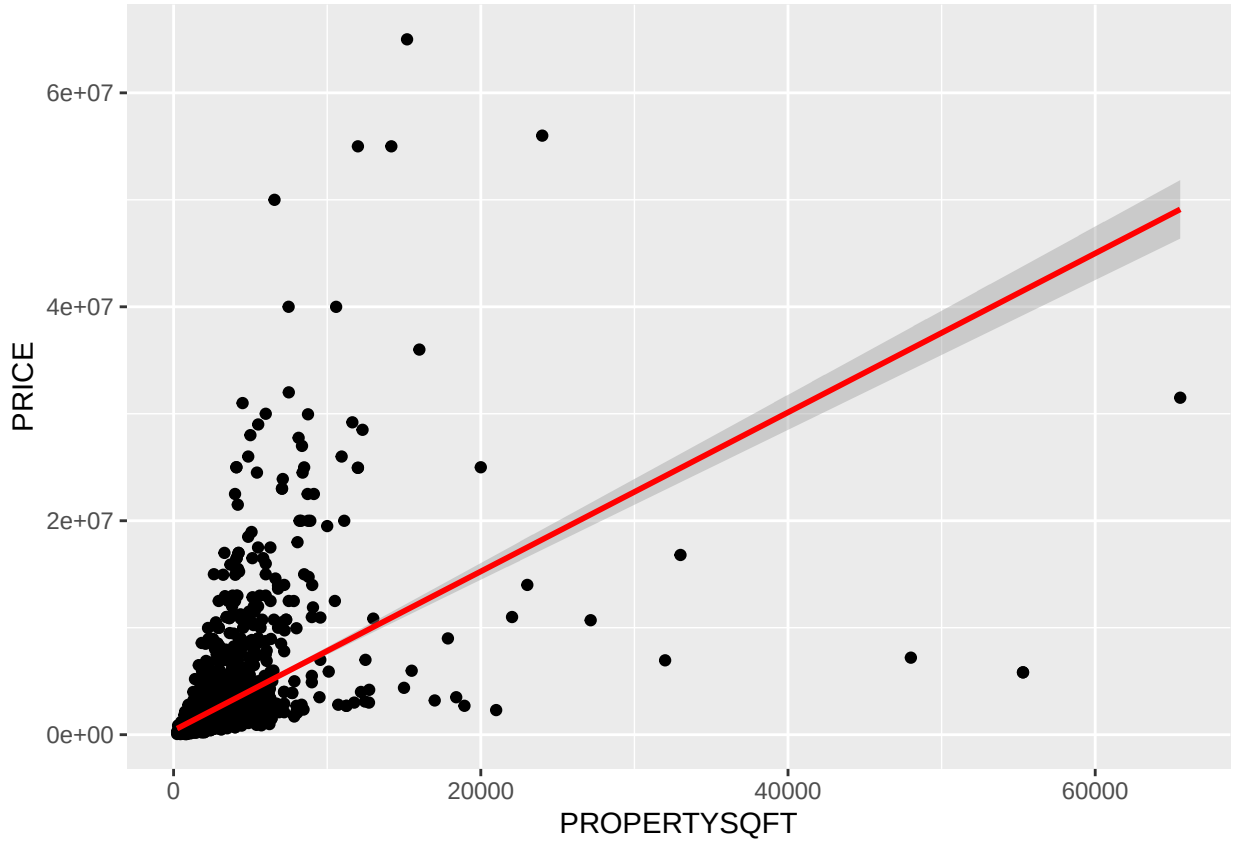

PRICE
0e+00
2e+07
4e+07
6e+07
0
10000
20000
30000
40000
50000
60000
PROPERTYSQFT

```
## integer(0)
```

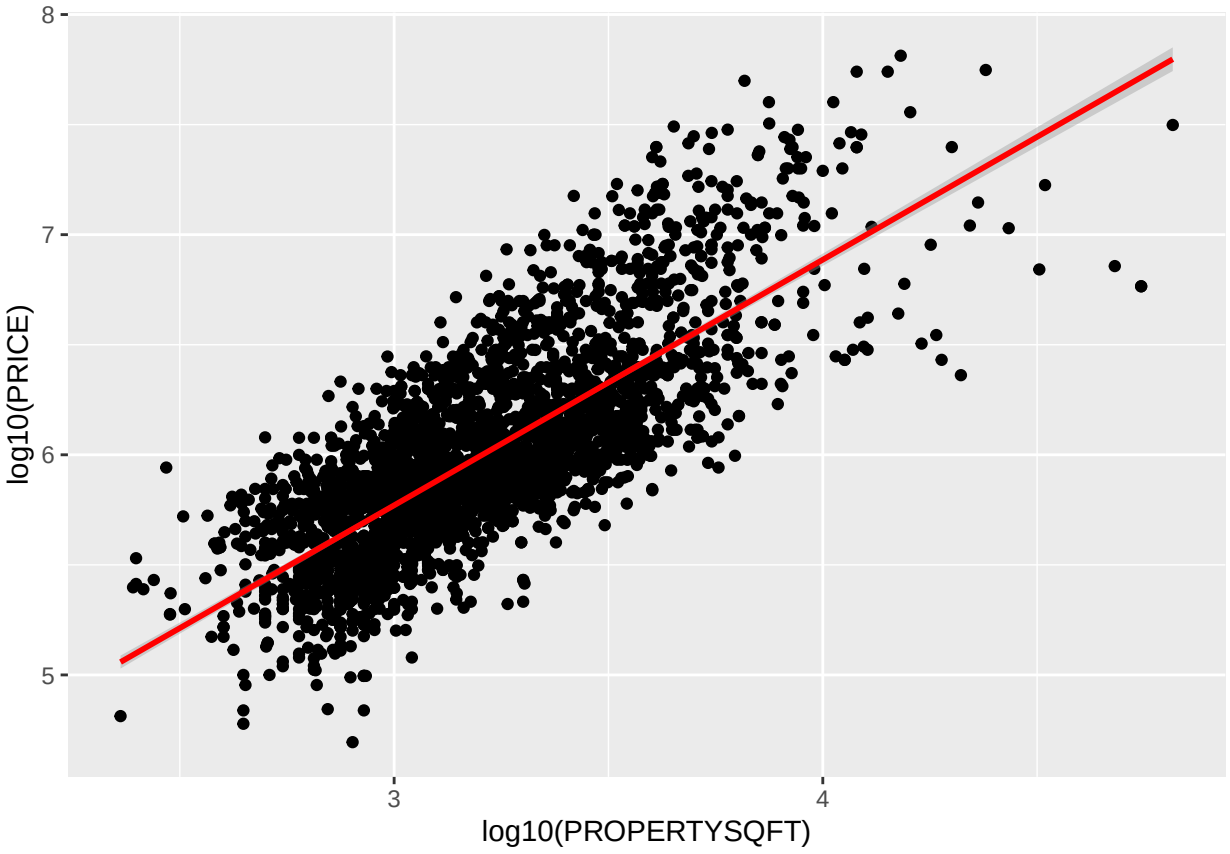


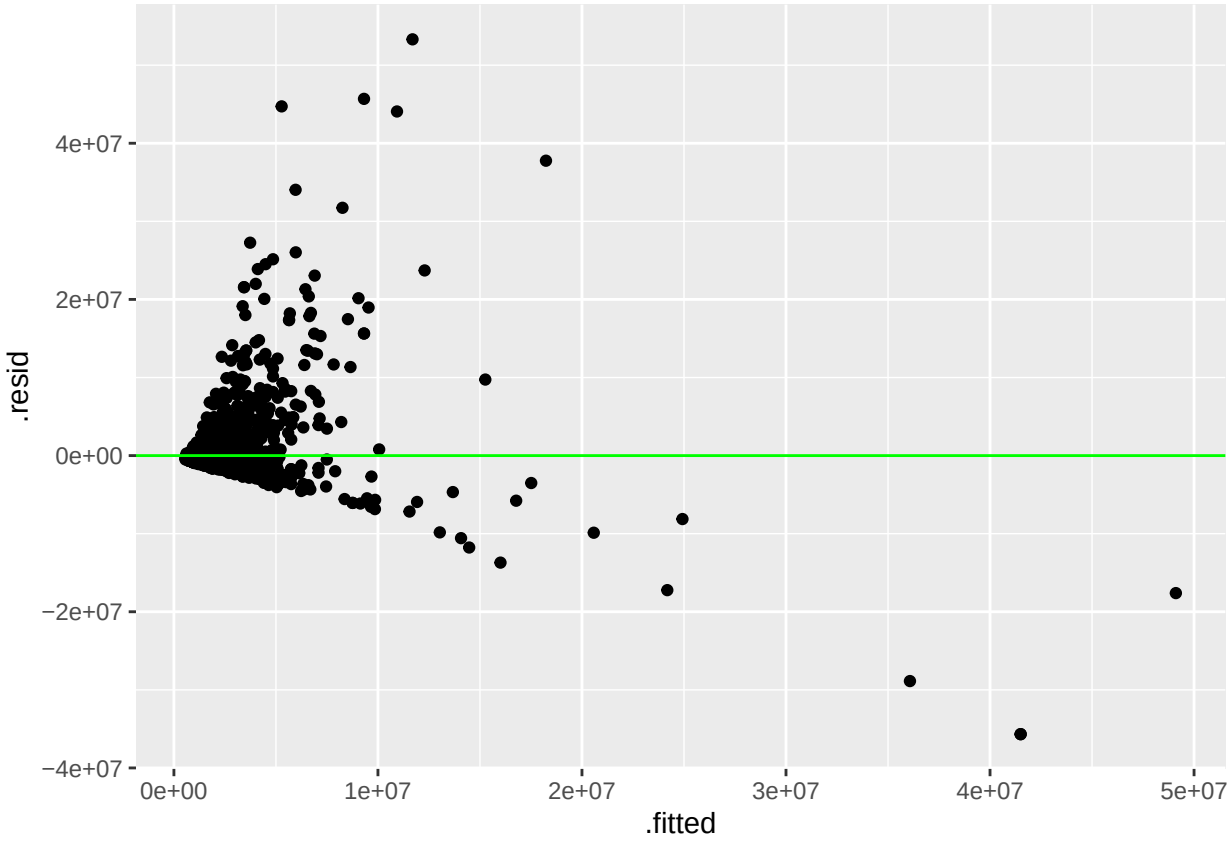

```
## integer(0)
```

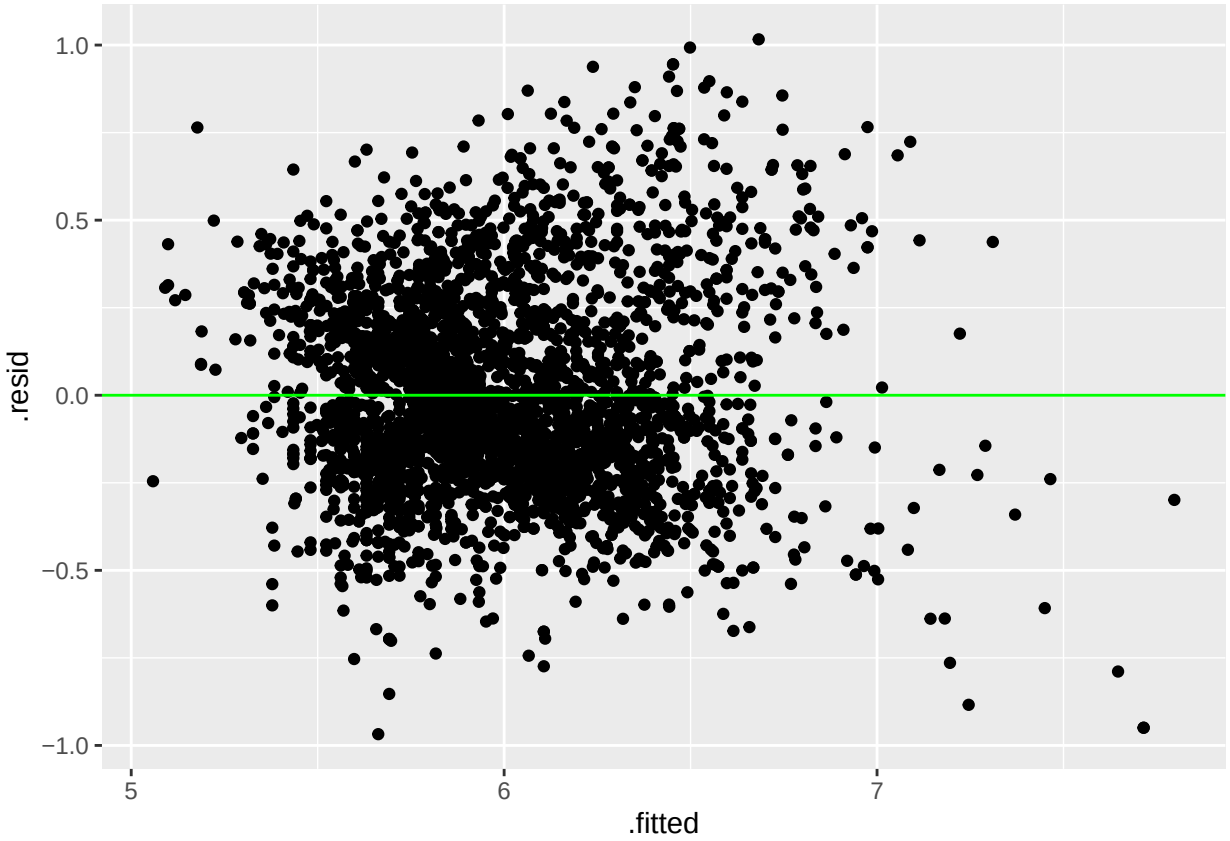
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
## `geom_smooth()` using formula = 'y ~ x'
```


8
7
6
5
log10(PRICE)
3
4
log10(PROPERTYSQFT)

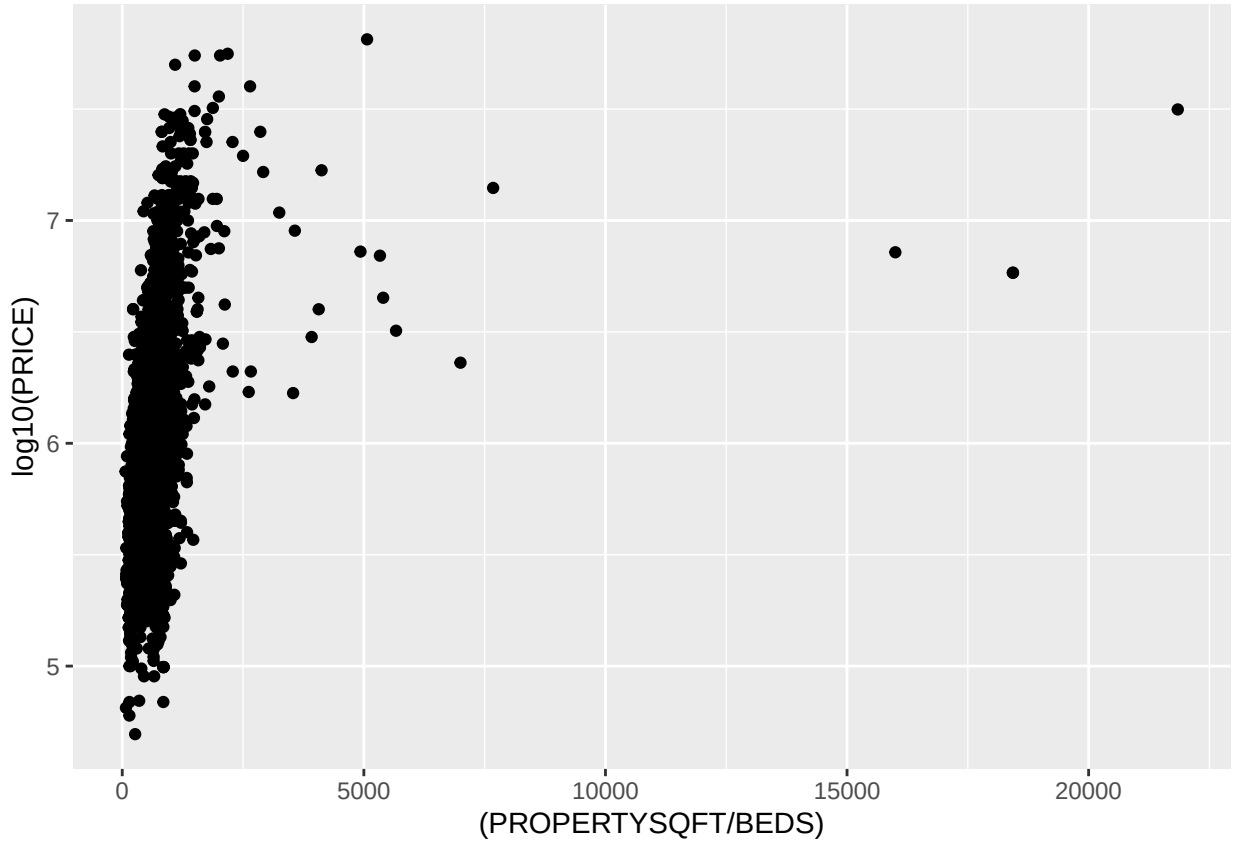

4e+07
2e+07
0e+00
−2e+07
−4e+07
.resid
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
.fitted



Linear Model 2

Using PROPERTYSQFT/BEDS as a predictor for PRICE

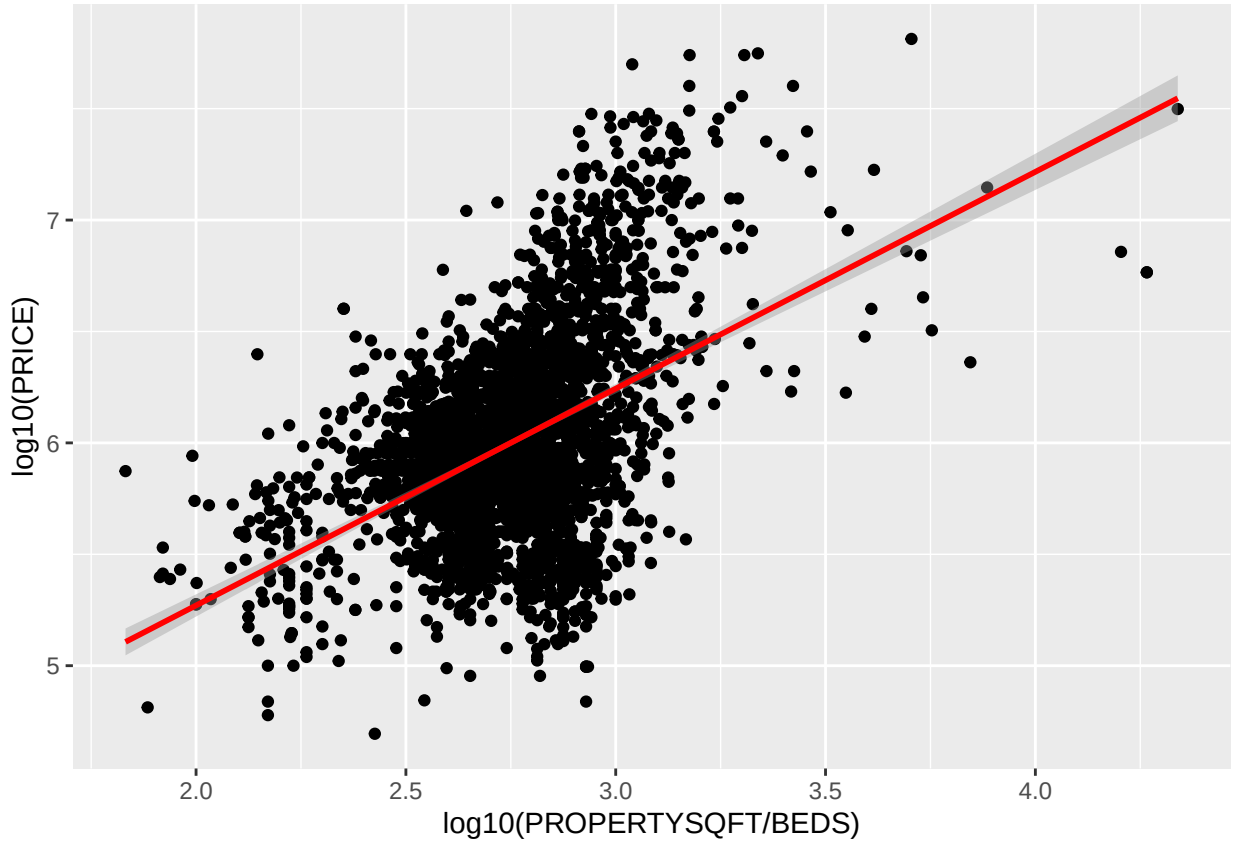
From the following, we can see that the data must be log-fit in order to be interpretative. Total square footage may not be an accurate way to depict multi-family homes. Assuming larger spaces cost more, this model may show that there is a relationship between price and the size of the property per bedroom

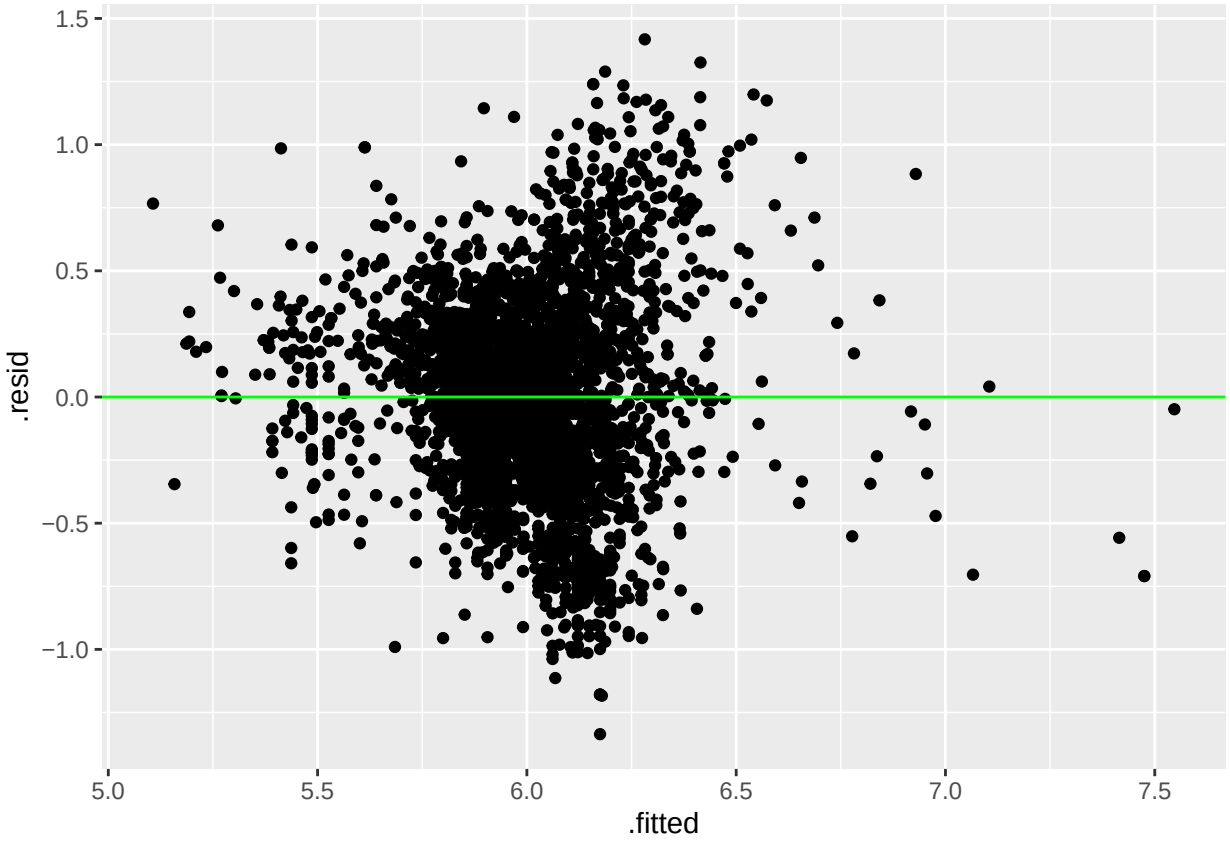


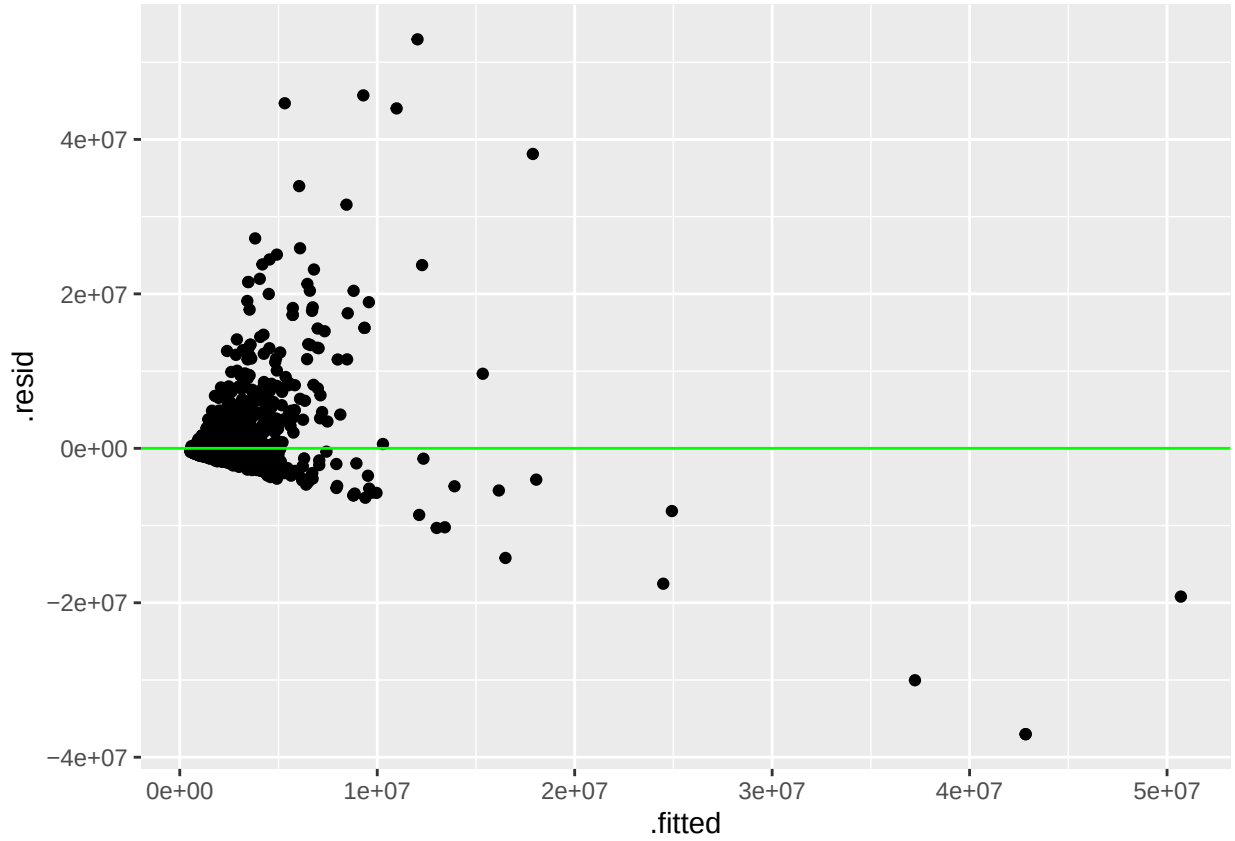

```
##
## Call:
## lm(formula = PRICE ~ PROPERTYSQFT/BEDS, data = dataset2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -37012341   -912933   -632930   -194285  52959144
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       366616.595  80790.682   4.538 5.89e-06 ***
## PROPERTYSQFT         782.381     27.462  28.489  < 2e-16 ***
## PROPERTYSQFT:BEDS     -4.780      2.014  -2.373   0.0177 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3599000 on 3175 degrees of freedom
## Multiple R-squared:  0.2659, Adjusted R-squared:  0.2654
## F-statistic: 574.9 on 2 and 3175 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = log10(PRICE) ~ log10(PROPERTYSQFT/BEDS), data = dataset2)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.33592 -0.24001 -0.01107  0.23548  1.41724
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3.32427    0.09000   36.94   <2e-16 ***
## log10(PROPERTYSQFT/BEDS)  0.97306    0.03263   29.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3954 on 3176 degrees of freedom
## Multiple R-squared:  0.2188, Adjusted R-squared:  0.2185
## F-statistic: 889.3 on 1 and 3176 DF,  p-value: < 2.2e-16
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




1.5
1.0
0.5
0.0
−0.5
−1.0
.resid
5.0
5.5
6.0
6.5
7.0
7.5
.fitted



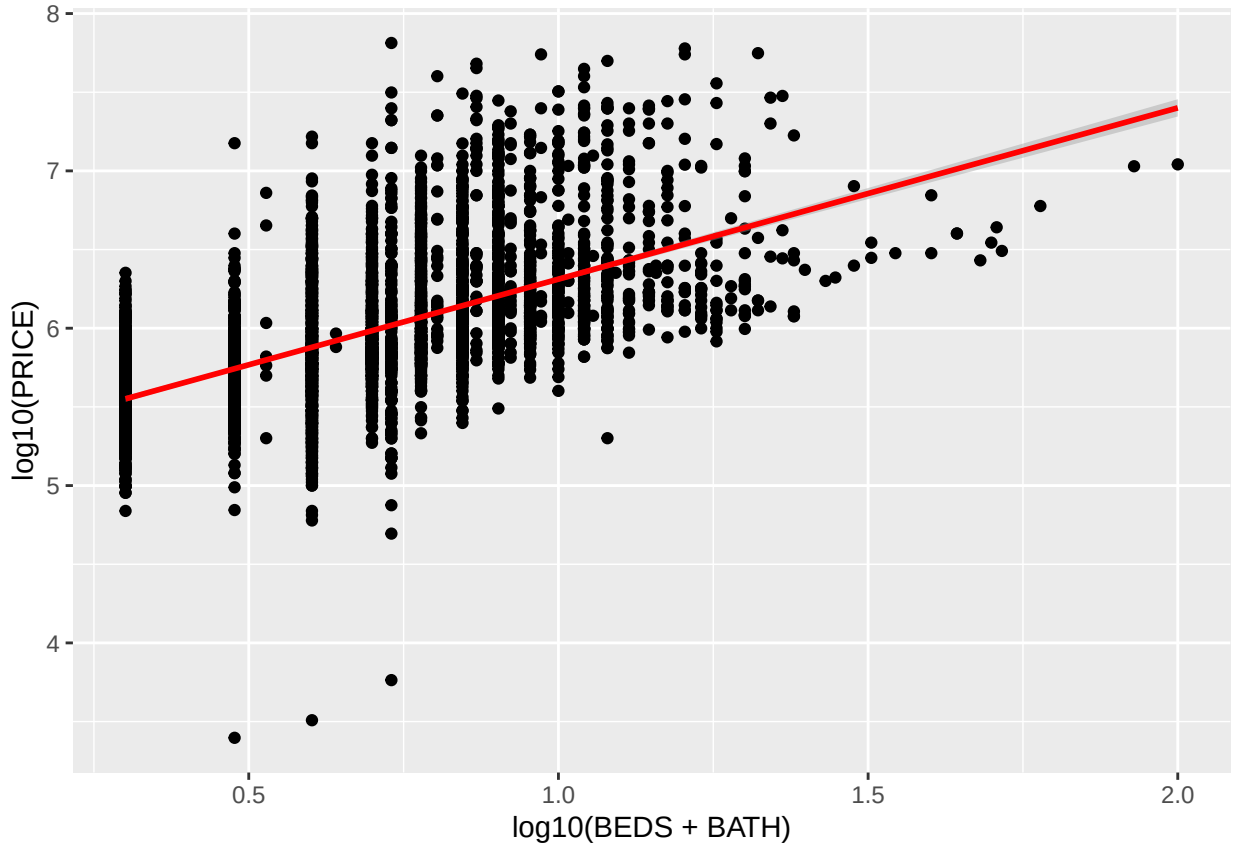
Linear Model 3

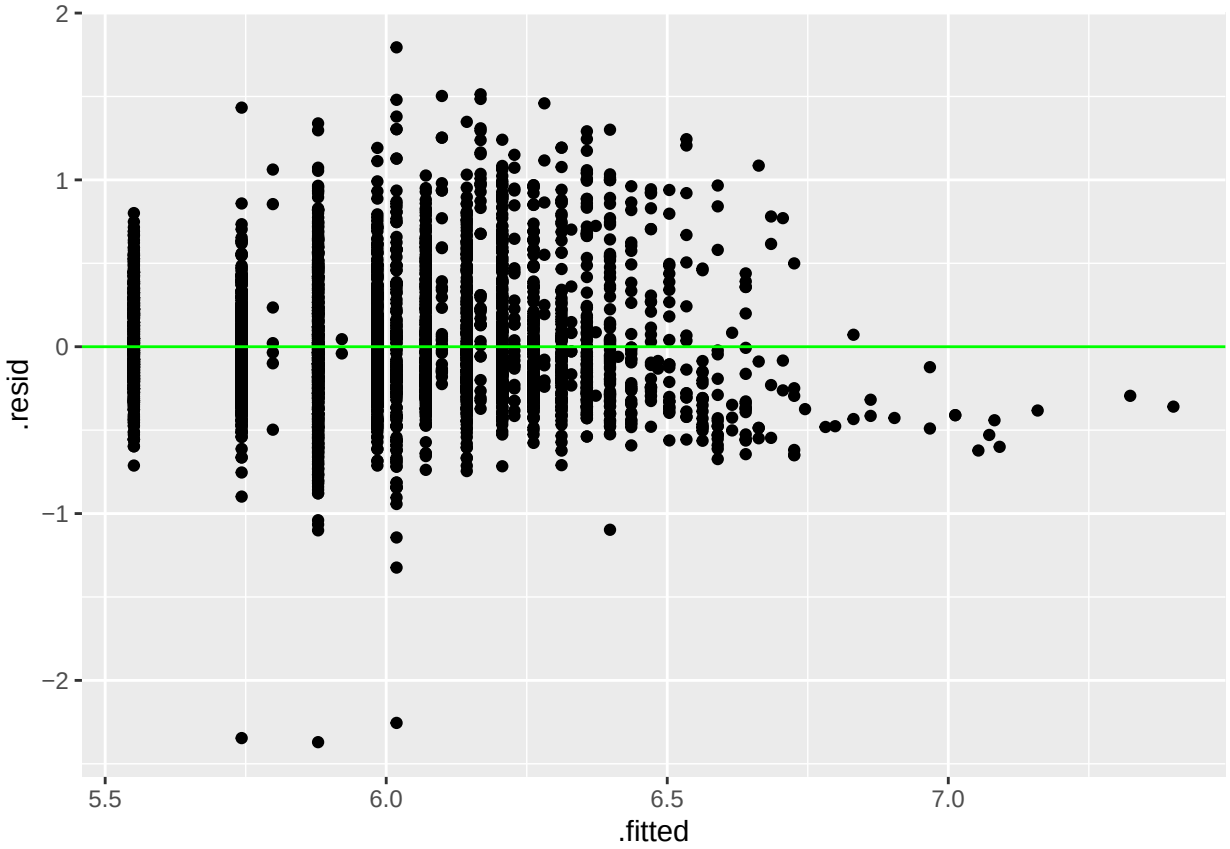
Using BEDS+BATH as a predictor for PRICE

From the following, we can see that the data must be log-fit in order to be interpretable.

```
##
## Call:
## lm(formula = log10(PRICE) ~ log10(BEDS + BATH), data = dataset3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.37029 -0.21924 -0.07257  0.18682  1.79452
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.22346    0.01523  343.05   <2e-16 ***
## log10(BEDS + BATH)   1.08851    0.02096   51.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3548 on 4797 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3597
## F-statistic:  2696 on 1 and 4797 DF,  p-value: < 2.2e-16
```

```
## `geom_smooth()` using formula = 'y ~ x'
```





Overall, I think the first linear model, comparing the square footage with price is the best indicator of fit, as that model yielded the largest r-squared value. This is not the only indicator, however, comparing values such as bed and bath number added countable numbers to the model, which caused clusters at specific values, for example, many properties were 1 bed but with varying prices.

Linear Model 4

Using zipcode as a predictor for PRICE

I wanted to look at the dataset in terms of location, but obviously, a lot of those are given in the address but we would need a way to simplify the address. The zipcode was compared to the log-form of price, but is not very descriptive in its meaning, likely because of how broad certain zipcodes in NYC may be.

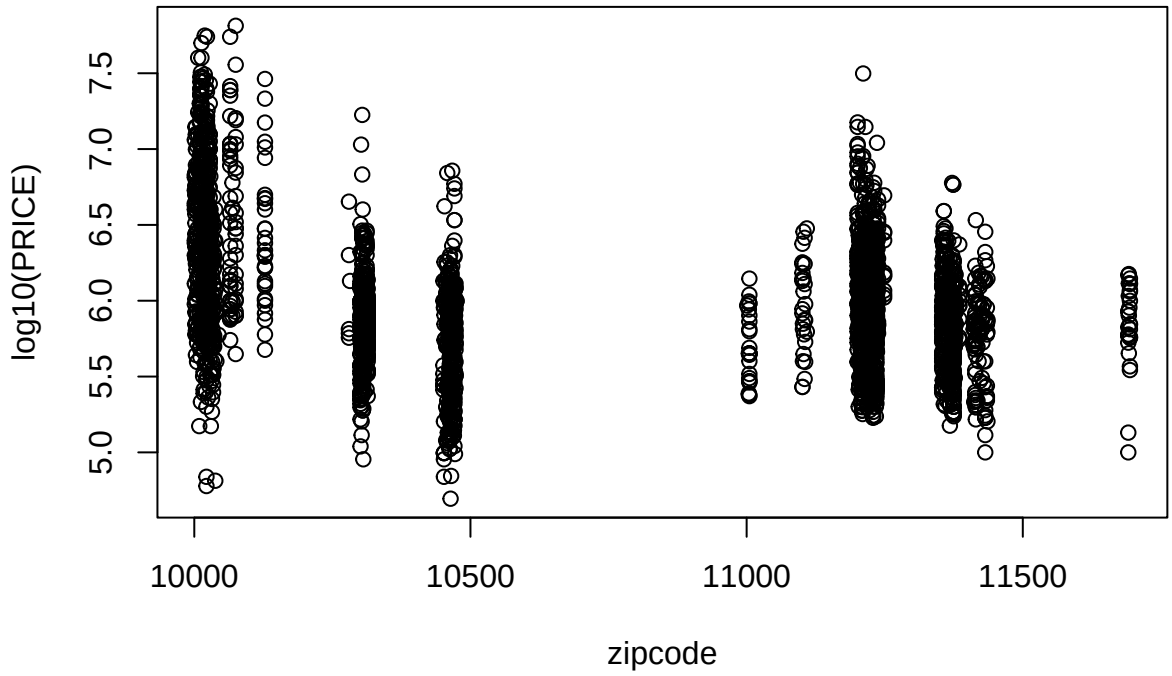
```
##
## Call:
## lm(formula = log10(PRICE) ~ (zipcode), data = dataset4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.40138 -0.21133 -0.00714  0.19120  1.32393
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.53942    0.09895  66.088  < 2e-16 ***
## zipcode10002 -0.14247    0.13722  -1.038 0.299230
```

```
## zipcode10003 -0.04831    0.11766  -0.411 0.681391
## zipcode10004 -0.40138    0.17139  -2.342 0.019247 *
## zipcode10005 -0.48209    0.14308  -3.369 0.000763 ***
## zipcode10006 -0.31328    0.22126  -1.416 0.156912
## zipcode10007  0.11611    0.12774   0.909 0.363463
## zipcode10009 -0.44584    0.17139  -2.601 0.009330 **
## zipcode10010  0.05542    0.13090   0.423 0.672063
## zipcode10011  0.24069    0.11467   2.099 0.035896 *
## zipcode10012  0.19439    0.12924   1.504 0.132648
## zipcode10013  0.33250    0.11426   2.910 0.003640 **
## zipcode10014  0.22853    0.12404   1.842 0.065514 .
## zipcode10016 -0.28254    0.11654  -2.424 0.015393 *
## zipcode10017 -0.48219    0.14677  -3.285 0.001030 **
## zipcode10018  0.14450    0.26180   0.552 0.581021
## zipcode10019 -0.11516    0.11111  -1.036 0.300063
## zipcode10021 -0.12572    0.11654  -1.079 0.280761
## zipcode10022 -0.38208    0.11191  -3.414 0.000648 ***
## zipcode10023 -0.09986    0.11282  -0.885 0.376168
## zipcode10024 -0.18555    0.11250  -1.649 0.099186 .
## zipcode10025 -0.33128    0.11555  -2.867 0.004172 **
## zipcode10026 -0.50736    0.17139  -2.960 0.003098 **
## zipcode10027 -0.41325    0.12404  -3.332 0.000874 ***
## zipcode10028  0.05963    0.12038   0.495 0.620410
## zipcode10029 -0.48000    0.12206  -3.932 8.60e-05 ***
## zipcode10030 -0.39480    0.13485  -2.928 0.003440 **
## zipcode10031 -0.30443    0.15115  -2.014 0.044089 *
## zipcode10032 -0.69538    0.14677  -4.738 2.26e-06 ***
## zipcode10033 -0.70254    0.15115  -4.648 3.50e-06 ***
## zipcode10034 -0.91652    0.18246  -5.023 5.38e-07 ***
## zipcode10035 -0.42126    0.13276  -3.173 0.001523 **
## zipcode10036 -0.49598    0.14308  -3.466 0.000535 ***
## zipcode10037 -0.92664    0.35677  -2.597 0.009442 **
## zipcode10038 -0.54741    0.18246  -3.000 0.002720 **
## zipcode10039 -0.61395    0.26180  -2.345 0.019086 *
## zipcode10040 -0.77423    0.26180  -2.957 0.003127 **
## zipcode10044 -0.76491    0.35677  -2.144 0.032115 *
## zipcode10065 -0.01521    0.11426  -0.133 0.894075
## zipcode10069 -0.22362    0.17139  -1.305 0.192069
## zipcode10075 -0.01178    0.12038  -0.098 0.922054
## zipcode10128 -0.12244    0.11555  -1.060 0.289406
## zipcode10280 -0.47775    0.18246  -2.618 0.008877 **
## zipcode10282 -0.40909    0.35677  -1.147 0.251620
## zipcode10301 -0.82375    0.10958  -7.517 7.34e-14 ***
## zipcode10302 -0.63096    0.13994  -4.509 6.77e-06 ***
## zipcode10303 -0.86254    0.12038  -7.165 9.72e-13 ***
## zipcode10304 -0.54223    0.10958  -4.948 7.90e-07 ***
## zipcode10305 -0.68977    0.11191  -6.164 8.06e-10 ***
## zipcode10306 -0.71806    0.10783  -6.659 3.26e-11 ***
## zipcode10307 -0.53311    0.11827  -4.508 6.81e-06 ***
## zipcode10308 -0.66467    0.11555  -5.752 9.69e-09 ***
## zipcode10309 -0.60035    0.11426  -5.254 1.59e-07 ***
## zipcode10310 -0.74525    0.12774  -5.834 5.99e-09 ***
## zipcode10312 -0.56926    0.10710  -5.315 1.14e-07 ***
## zipcode10314 -0.70809    0.10570  -6.699 2.50e-11 ***
```

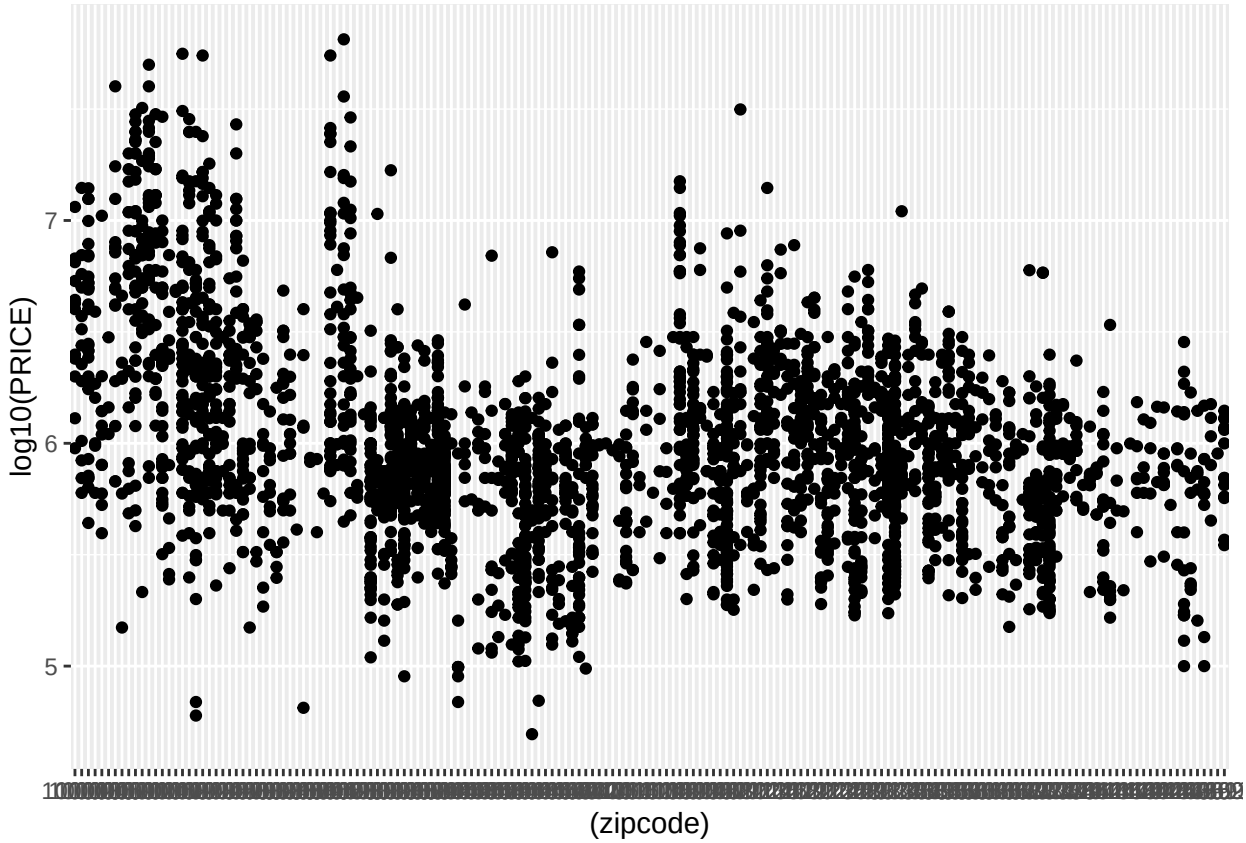
```
## zipcode10451 -0.93445    0.15115  -6.182 7.17e-10 ***
## zipcode10452 -1.42267    0.15645  -9.093  < 2e-16 ***
## zipcode10453 -0.49338    0.16302  -3.026 0.002495 **
## zipcode10454 -0.59618    0.26180  -2.277 0.022842 *
## zipcode10456 -0.79701    0.15645  -5.094 3.72e-07 ***
## zipcode10457 -0.53956    0.16302  -3.310 0.000945 ***
## zipcode10458 -0.94252    0.14308  -6.587 5.27e-11 ***
## zipcode10459 -0.96976    0.17139  -5.658 1.67e-08 ***
## zipcode10460 -0.67428    0.14308  -4.713 2.56e-06 ***
## zipcode10461 -0.69291    0.11766  -5.889 4.31e-09 ***
## zipcode10462 -0.97803    0.11603  -8.429  < 2e-16 ***
## zipcode10463 -0.90715    0.10810  -8.392  < 2e-16 ***
## zipcode10464 -0.97092    0.17139  -5.665 1.61e-08 ***
## zipcode10465 -0.75798    0.11220  -6.756 1.70e-11 ***
## zipcode10466 -0.67206    0.11963  -5.618 2.11e-08 ***
## zipcode10467 -0.83692    0.12924  -6.476 1.10e-10 ***
## zipcode10468 -1.01931    0.16302  -6.253 4.61e-10 ***
## zipcode10469 -0.70444    0.11827  -5.956 2.88e-09 ***
## zipcode10470 -1.07111    0.13276  -8.068 1.02e-15 ***
## zipcode10471 -0.81434    0.11315  -7.197 7.75e-13 ***
## zipcode10472 -0.68181    0.14677  -4.646 3.54e-06 ***
## zipcode10473 -0.76111    0.12516  -6.081 1.35e-09 ***
## zipcode10474 -0.54621    0.26180  -2.086 0.037027 *
## zipcode10475 -0.53986    0.35677  -1.513 0.130340
## zipcode11001 -0.57099    0.35677  -1.600 0.109605
## zipcode11004 -0.86773    0.18246  -4.756 2.07e-06 ***
## zipcode11005 -0.84746    0.12301  -6.889 6.80e-12 ***
## zipcode11101 -0.53417    0.14677  -3.640 0.000278 ***
## zipcode11102 -0.78562    0.22126  -3.551 0.000390 ***
## zipcode11103 -0.46199    0.19790  -2.334 0.019637 *
## zipcode11104 -0.76135    0.35677  -2.134 0.032923 *
## zipcode11105 -0.51411    0.18246  -2.818 0.004868 **
## zipcode11106 -0.74579    0.19790  -3.769 0.000167 ***
## zipcode11109 -0.40307    0.26180  -1.540 0.123761
## zipcode11201 -0.18306    0.11163  -1.640 0.101142
## zipcode11203 -0.62225    0.13276  -4.687 2.89e-06 ***
## zipcode11204 -0.51565    0.11426  -4.513 6.64e-06 ***
## zipcode11205 -0.32481    0.14677  -2.213 0.026968 *
## zipcode11206 -0.41621    0.18246  -2.281 0.022608 *
## zipcode11207 -0.74986    0.12404  -6.045 1.68e-09 ***
## zipcode11208 -0.74226    0.11963  -6.205 6.23e-10 ***
## zipcode11209 -0.67834    0.10548  -6.431 1.47e-10 ***
## zipcode11210 -0.65290    0.12639  -5.166 2.55e-07 ***
## zipcode11211 -0.18518    0.13485  -1.373 0.169769
## zipcode11212 -0.71483    0.16302  -4.385 1.20e-05 ***
## zipcode11213 -0.69852    0.14677  -4.759 2.03e-06 ***
## zipcode11214 -0.53922    0.11603  -4.647 3.51e-06 ***
## zipcode11215 -0.23995    0.11510  -2.085 0.037172 *
## zipcode11216 -0.38143    0.12774  -2.986 0.002850 **
## zipcode11217 -0.03535    0.15115  -0.234 0.815079
## zipcode11218 -0.67894    0.12924  -5.253 1.60e-07 ***
## zipcode11219 -0.47820    0.12774  -3.743 0.000185 ***
## zipcode11220 -0.50103    0.12301  -4.073 4.76e-05 ***
## zipcode11221 -0.37988    0.11892  -3.194 0.001416 **
```

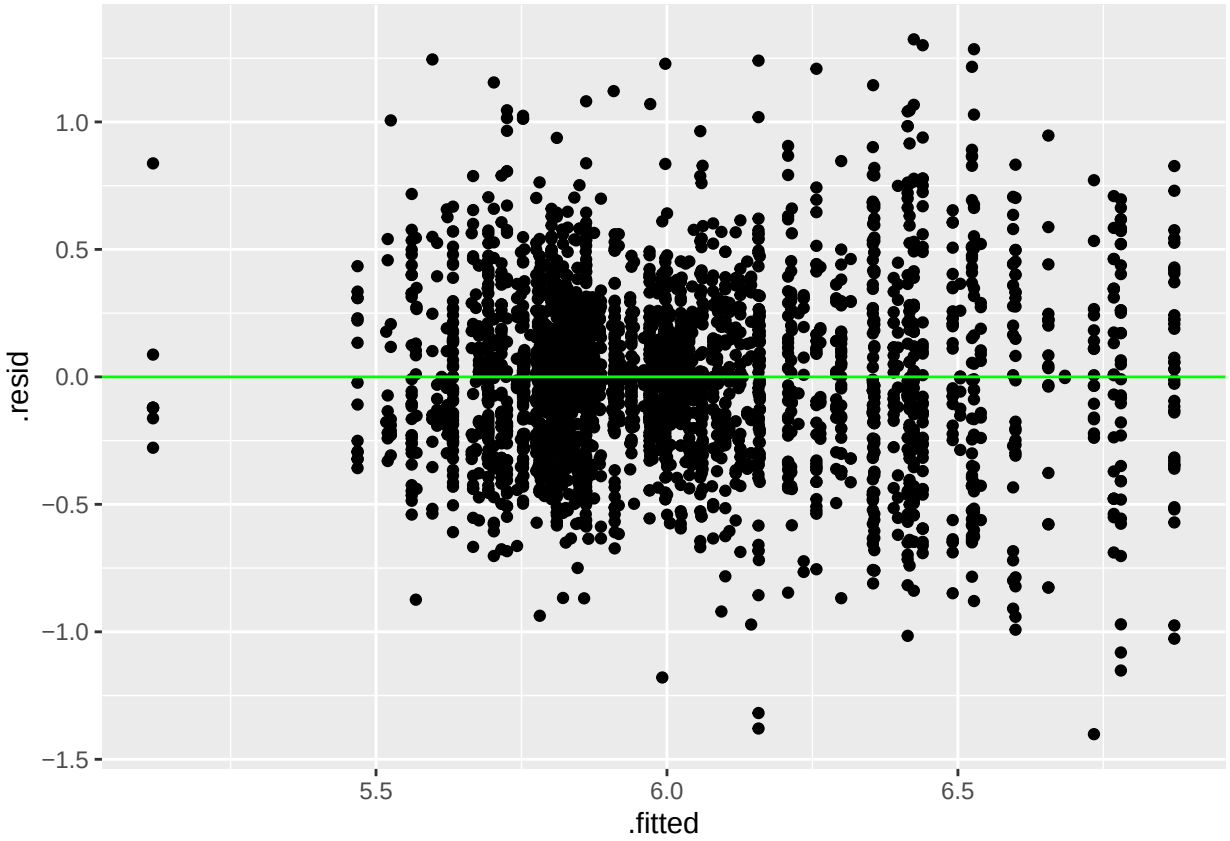
```
## zipcode11222 -0.24861    0.13276  -1.873 0.061213 .
## zipcode11223 -0.72888    0.11766  -6.195 6.62e-10 ***
## zipcode11224 -0.76408    0.11766  -6.494 9.73e-11 ***
## zipcode11225 -0.55491    0.14308  -3.878 0.000107 ***
## zipcode11226 -0.50795    0.15115  -3.361 0.000787 ***
## zipcode11228 -0.46009    0.11963  -3.846 0.000123 ***
## zipcode11229 -0.72865    0.11063  -6.586 5.30e-11 ***
## zipcode11230 -0.75830    0.11827  -6.412 1.67e-10 ***
## zipcode11231 -0.14975    0.12516  -1.196 0.231621
## zipcode11232 -0.60256    0.15115  -3.987 6.86e-05 ***
## zipcode11233 -0.45339    0.13276  -3.415 0.000646 ***
## zipcode11234 -0.62919    0.10721  -5.869 4.87e-09 ***
## zipcode11235 -0.73781    0.10563  -6.985 3.49e-12 ***
## zipcode11236 -0.56813    0.13485  -4.213 2.59e-05 ***
## zipcode11237 -0.45169    0.15645  -2.887 0.003916 **
## zipcode11238 -0.32599    0.12639  -2.579 0.009951 **
## zipcode11249 -0.27570    0.14677  -1.879 0.060409 .
## zipcode11354 -0.71938    0.11467  -6.274 4.03e-10 ***
## zipcode11355 -0.69312    0.12639  -5.484 4.51e-08 ***
## zipcode11356 -0.52213    0.13722  -3.805 0.000145 ***
## zipcode11357 -0.43908    0.11892  -3.692 0.000226 ***
## zipcode11358 -0.52120    0.16302  -3.197 0.001402 **
## zipcode11360 -0.71005    0.11708  -6.065 1.49e-09 ***
## zipcode11361 -0.43209    0.13994  -3.088 0.002035 **
## zipcode11362 -0.79036    0.17139  -4.612 4.16e-06 ***
## zipcode11363 -0.44369    0.17139  -2.589 0.009677 **
## zipcode11365 -0.59601    0.16302  -3.656 0.000261 ***
## zipcode11366 -0.55454    0.16302  -3.402 0.000679 ***
## zipcode11367 -0.91774    0.14308  -6.414 1.64e-10 ***
## zipcode11368 -0.71335    0.12774  -5.584 2.56e-08 ***
## zipcode11369 -0.78603    0.18246  -4.308 1.70e-05 ***
## zipcode11370 -0.61935    0.22126  -2.799 0.005156 **
## zipcode11372 -0.78676    0.12038  -6.536 7.41e-11 ***
## zipcode11373 -0.78437    0.12038  -6.516 8.44e-11 ***
## zipcode11374 -0.78674    0.11766  -6.687 2.71e-11 ***
## zipcode11375 -0.84632    0.10958  -7.723 1.53e-14 ***
## zipcode11377 -0.66660    0.13722  -4.858 1.25e-06 ***
## zipcode11378 -0.54430    0.16302  -3.339 0.000852 ***
## zipcode11379 -0.57896    0.16302  -3.551 0.000389 ***
## zipcode11385 -0.54963    0.14677  -3.745 0.000184 ***
## zipcode11411 -0.79736    0.22126  -3.604 0.000319 ***
## zipcode11412 -0.73611    0.15645  -4.705 2.65e-06 ***
## zipcode11413 -0.74280    0.18246  -4.071 4.80e-05 ***
## zipcode11414 -0.75614    0.13090  -5.776 8.41e-09 ***
## zipcode11415 -1.01401    0.15115  -6.709 2.34e-11 ***
## zipcode11417 -0.70331    0.22126  -3.179 0.001495 **
## zipcode11418 -1.02190    0.26180  -3.903 9.69e-05 ***
## zipcode11419 -0.54029    0.35677  -1.514 0.130031
## zipcode11420 -0.66724    0.17139  -3.893 0.000101 ***
## zipcode11421 -0.63398    0.19790  -3.204 0.001372 **
## zipcode11422 -0.71101    0.19790  -3.593 0.000332 ***
## zipcode11423 -0.54441    0.26180  -2.079 0.037657 *
## zipcode11426 -0.67630    0.15115  -4.474 7.95e-06 ***
## zipcode11427 -0.81597    0.35677  -2.287 0.022260 *
```

```
## zipcode11429 -0.68944    0.16302  -4.229 2.42e-05 ***
## zipcode11432 -0.87259    0.13090  -6.666 3.11e-11 ***
## zipcode11435 -0.87492    0.15645  -5.592 2.44e-08 ***
## zipcode11436 -0.76328    0.18246  -4.183 2.95e-05 ***
## zipcode11691 -0.83695    0.14677  -5.703 1.29e-08 ***
## zipcode11692 -0.56404    0.18246  -3.091 0.002011 **
## zipcode11693 -0.58518    0.35677  -1.640 0.101066
## zipcode11694 -0.65476    0.13485  -4.856 1.26e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3428 on 3006 degrees of freedom
## Multiple R-squared:  0.4444, Adjusted R-squared:  0.4128
## F-statistic: 14.06 on 171 and 3006 DF,  p-value: < 2.2e-16

## Warning in abline(lmod_log): only using the first two of 172 regression
## coefficients
```




```
## integer(0)

## `geom_smooth()` using formula = 'y ~ x'
```


log10(PRICE)
7
6
5
(zipcode)


1.0
0.5
0.0
−0.5
−1.0
−1.5
.resid
5.5
6.0
6.5
.fitted