King County House Price Prediction

Using Simple Regression Techniques

repo: https://github.com/YM88/king_county_house_price_linear_regression

Yashar Mansouri

STAT-615

American University

Data

This dataset contains house sale prices for King County, which includes Seattle.

It includes homes sold between May 2014 and May 2015.

Anomalies: Bedrooms, Bathrooms, Sold Multiple Times

Dropped: lat, long, id

```
      id                 date                          price           bedrooms       bathrooms      sqft_living      sqft_lot          floors
 Length:21613       Min.   :2014-05-02 00:00:00   Min.   :  75000   Min.   : 0.000   Min.   :0.000   Min.   :  290   Min.   :    520   Min.
:1.000
 Class :character   1st Qu.:2014-07-22 00:00:00   1st Qu.: 321950   1st Qu.: 3.000   1st Qu.:1.750   1st Qu.: 1427   1st Qu.:   5040   1st
Qu.:1.000
 Mode  :character   Median :2014-10-16 00:00:00   Median : 450000   Median : 3.000   Median :2.250   Median : 1910   Median :   7618   Median
:1.500
                    Mean   :2014-10-29 04:38:01   Mean   : 540088   Mean   : 3.371   Mean   :2.115   Mean   : 2080   Mean   :  15107   Mean
:1.494
                    3rd Qu.:2015-02-17 00:00:00   3rd Qu.: 645000   3rd Qu.: 4.000   3rd Qu.:2.500   3rd Qu.: 2550   3rd Qu.:  10688   3rd
Qu.:2.000
                    Max.   :2015-05-27 00:00:00   Max.   :7700000   Max.   :33.000   Max.   :8.000   Max.   :13540   Max.   :1651359   Max.
:3.500
   waterfront            view          condition         grade          sqft_above    sqft_basement      yr_built     yr_renovated       zipcode
 Min.   :0.000000   Min.   :0.0000   Min.   :1.000   Min.   : 1.000   Min.   : 290   Min.   :   0.0   Min.   :1900   Min.   :   0.0   Min.
:98001
 1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.: 7.000   1st Qu.:1190   1st Qu.:   0.0   1st Qu.:1951   1st Qu.:   0.0   1st
Qu.:98033
 Median :0.000000   Median :0.0000   Median :3.000   Median : 7.000   Median :1560   Median :   0.0   Median :1975   Median :   0.0   Median
:98065
 Mean   :0.007542   Mean   :0.2343   Mean   :3.409   Mean   : 7.657   Mean   :1788   Mean   : 291.5   Mean   :1971   Mean   :  84.4   Mean
:98078
 3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.: 8.000   3rd Qu.:2210   3rd Qu.: 560.0   3rd Qu.:1997   3rd Qu.:   0.0   3rd
Qu.:98118
 Max.   :1.000000   Max.   :4.0000   Max.   :5.000   Max.   :13.000   Max.   :9410   Max.   :4820.0   Max.   :2015   Max.   :2015.0   Max.
:98199
      lat             long        sqft_living15    sqft_lot15
 Min.   :47.16   Min.   :-122.5   Min.   : 399   Min.   :   651
 1st Qu.:47.47   1st Qu.:-122.3   1st Qu.:1490   1st Qu.:  5100
 Median :47.57   Median :-122.2   Median :1840   Median :  7620
 Mean   :47.56   Mean   :-122.2   Mean   :1987   Mean   : 12768
 3rd Qu.:47.68   3rd Qu.:-122.1   3rd Qu.:2360   3rd Qu.: 10083
 Max.   :47.78   Max.   :-121.3   Max.   :6210   Max.   :871200
```

Dataset can be found at kaggle and contains 21k+ records.

id: Unique ID for each home sold

date: Date of the home sale

price: Price of each home sold <– target variable

bedrooms: Number of bedrooms

bathrooms: Number of bathrooms, where .5 accounts for a room with a toilet but no shower

sqft_living: Square footage of the apartments interior living space

sqft_lot: Square footage of the land space

floors: Number of floors

waterfront: - A dummy variable for whether the apartment was overlooking the waterfront or not

view: An index from 0 to 4 of how good the view of the property was

condition: - An index from 1 to 5 on the condition of the apartment,

grade: An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.

sqft_above: The square footage of the interior housing space that is above ground level

sqft_basement: The square footage of the interior housing space that is below ground level

yr_built: The year the house was initially built

yr_renovated: The year of the house's last renovation

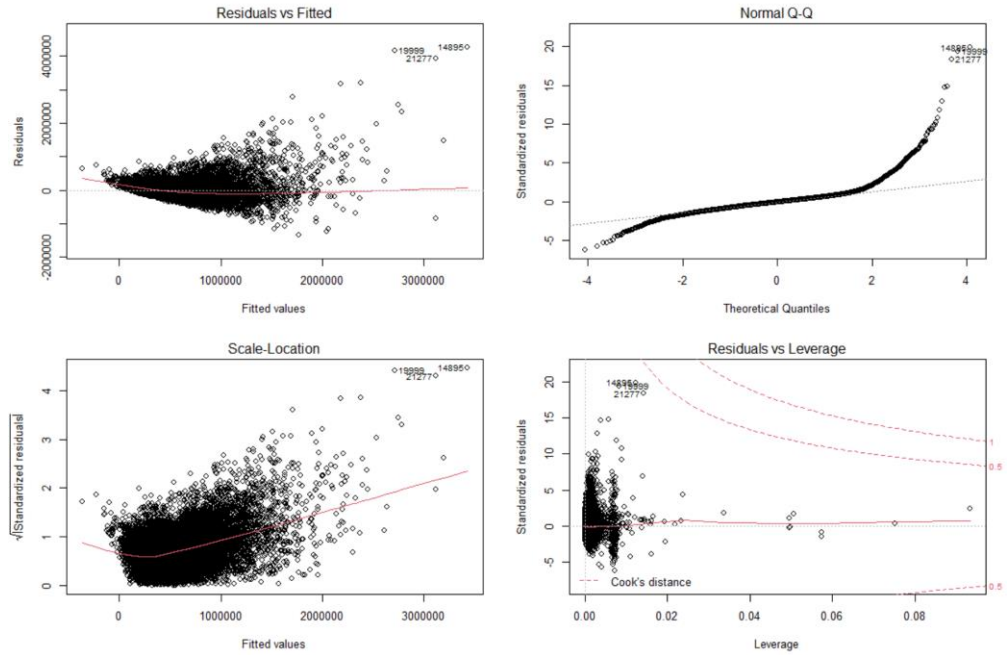
zipcode: What zipcode area the house is in

lat: Latitude

long: Longitude

sqft_living15: The square footage of interior housing living space for the nearest 15 neighbors

sqft_lot15: The square footage of the land lots of the nearest 15 neighbors

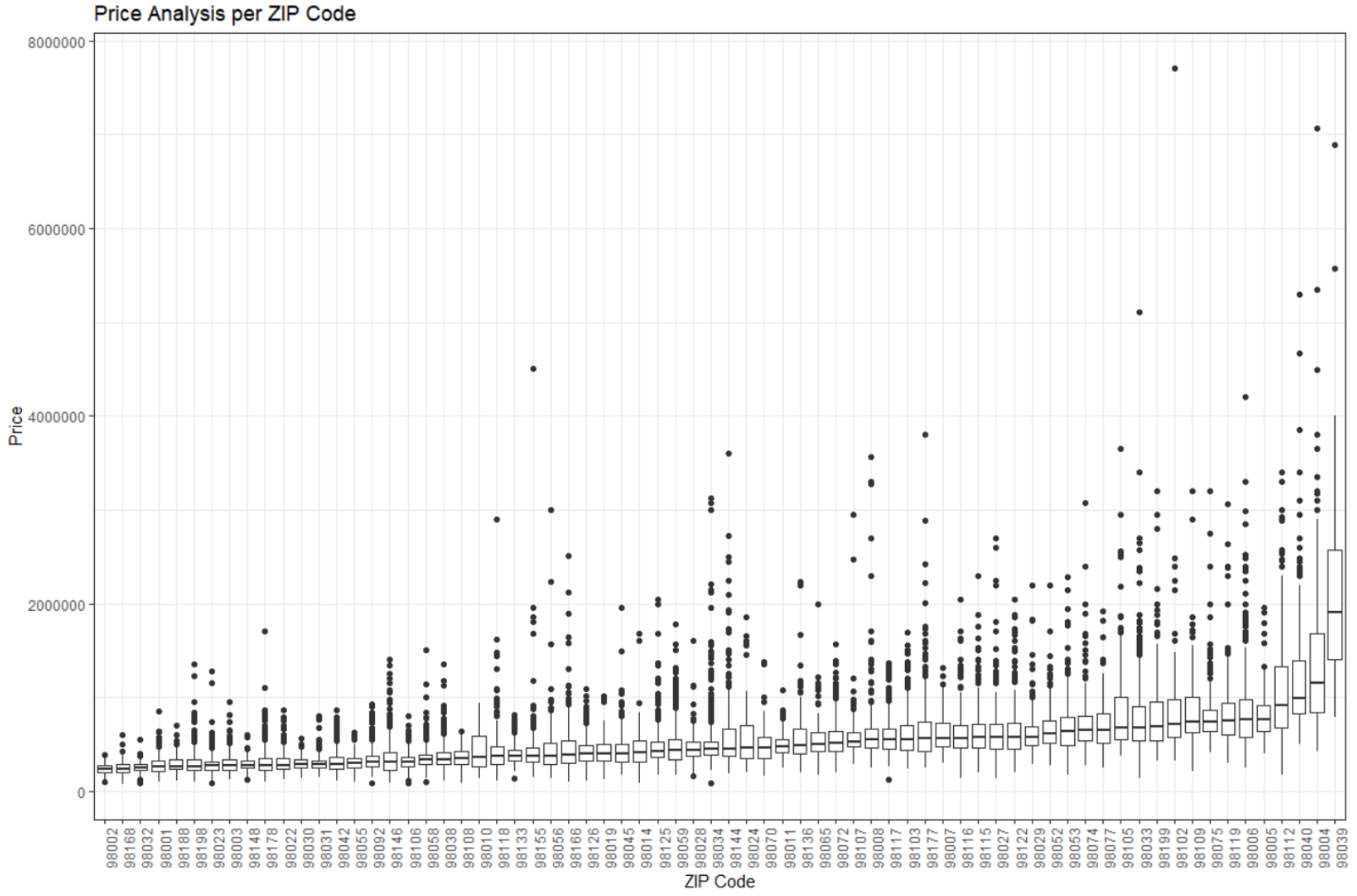



```
Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_living + sqft_lot +
    floors + waterfront + view + condition + grade + view + condition +
    grade + sqft_above + sqft_basement + yr_built + yr_renovated +
    sqft_living15 + sqft_lot15, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1344043  -109154    -9943    90129  4269962

Coefficients: (1 not defined because of singularities)
                 Estimate     Std. Error t value             Pr(>|t|)
(Intercept)   6261526.682177  139188.133402  44.986 < 0.0000000000000002 ***
bedrooms       -43146.723985    2128.483462 -20.271 < 0.0000000000000002 ***
bathrooms       47726.456518    3527.105375  13.531 < 0.0000000000000002 ***
sqft_living       169.177627       4.704802  35.959 < 0.0000000000000002 ***
sqft_lot           -0.007162       0.051302  -0.140               0.8890
floors          26086.546755    3797.959071   6.869   0.0000000000066626 ***
waterfront     577710.316865   18647.237975  30.981 < 0.0000000000000002 ***
view            42603.072519    2281.392397  18.674 < 0.0000000000000002 ***
condition       19419.593207    2514.158413   7.724   0.0000000000000118 ***
grade          120017.067370    2266.816734  52.945 < 0.0000000000000002 ***
sqft_above         -6.502103       4.558754  -1.426               0.1538
sqft_basement            NA             NA      NA                   NA
yr_built        -3600.414155      71.331436 -50.474 < 0.0000000000000002 ***
yr_renovated        9.662923       3.924296   2.462               0.0138 *
sqft_living15      23.973360       3.611435   6.638   0.0000000000325179 ***
sqft_lot15         -0.554540       0.078419  -7.072   0.0000000000015798 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 216100 on 21411 degrees of freedom
Multiple R-squared:  0.6542,    Adjusted R-squared:  0.654
F-statistic:  2893 on 14 and 21411 DF,  p-value: < 0.00000000000000022
```

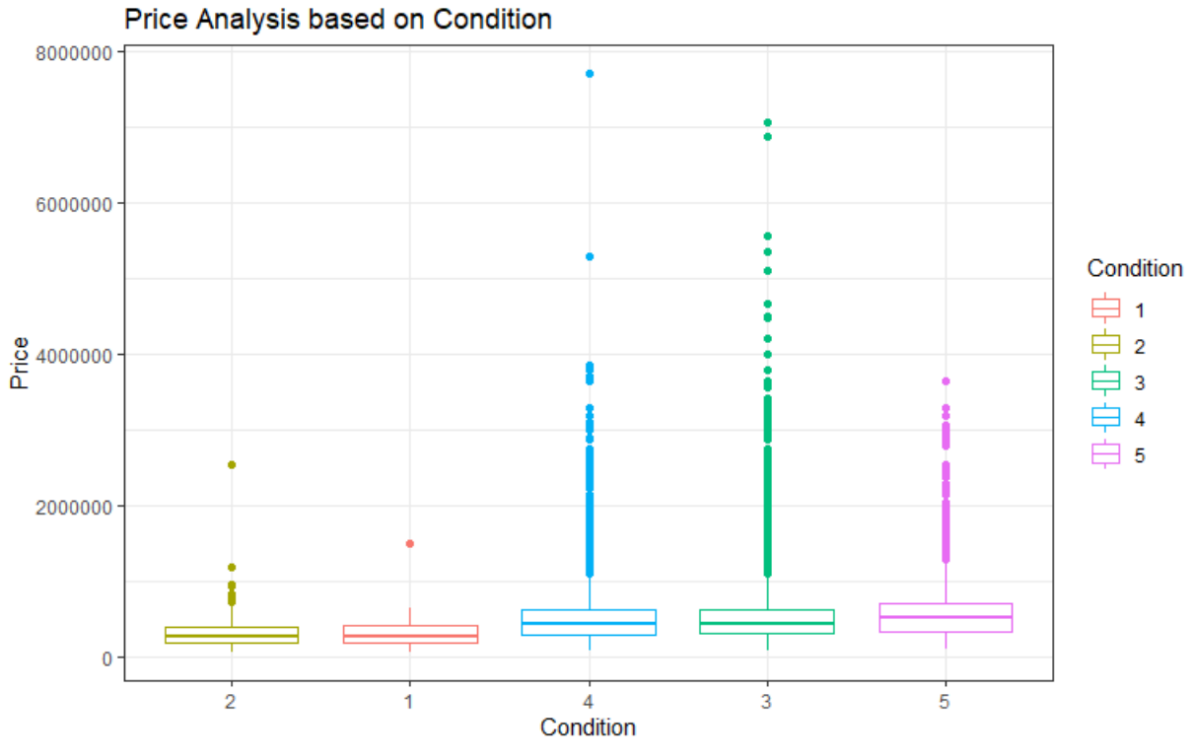
1st Reg: Base Score – without zipcode
Adjusted R-square 65%

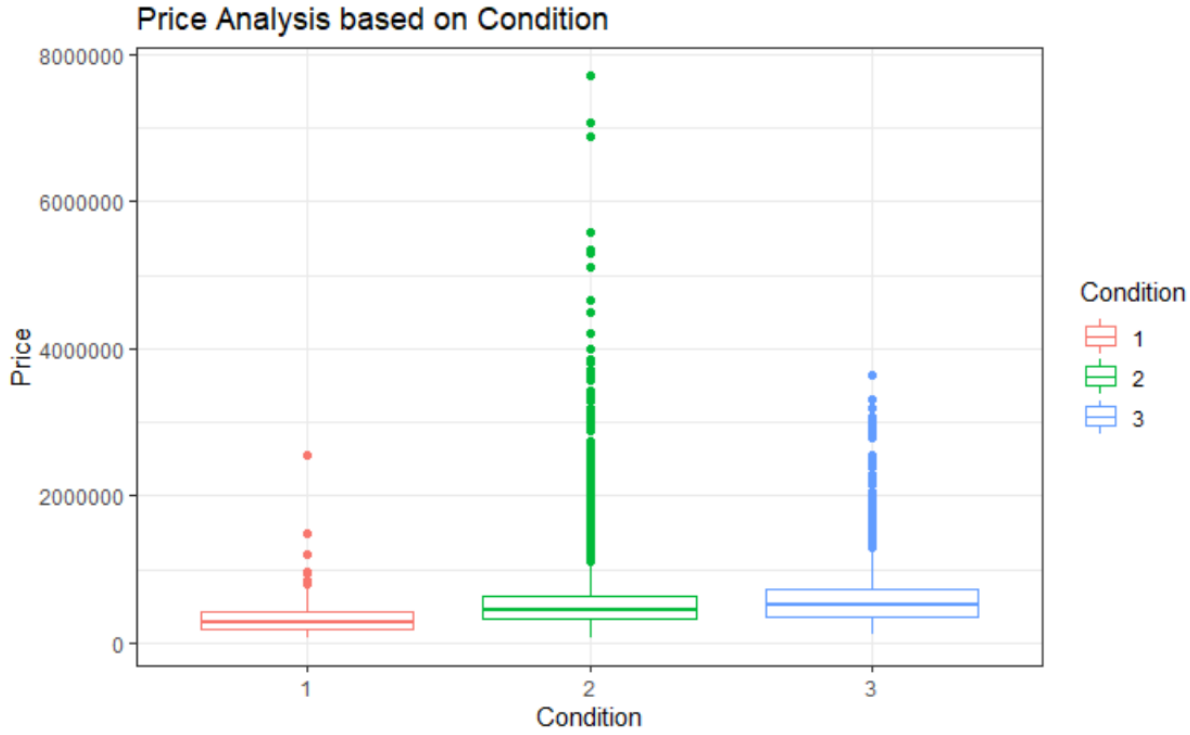
EDA - Waterfront

EDA – ZIP Code

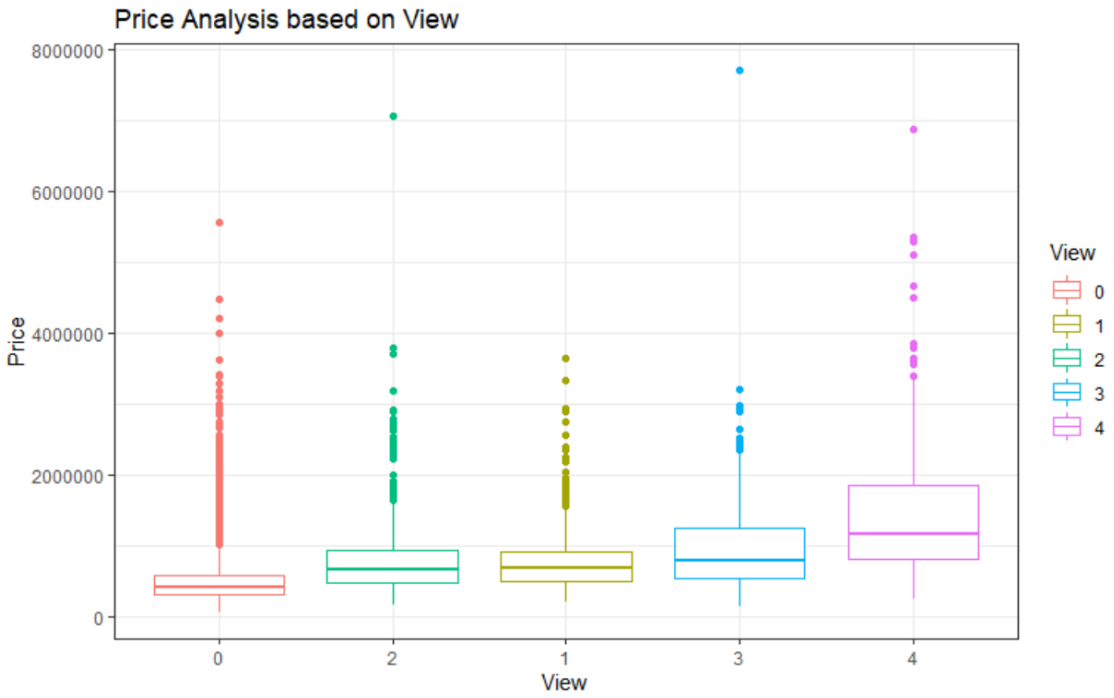

Price Analysis per ZIP Code
Price
8000000
6000000
4000000
2000000
0
ZIP Code
98002 98168 98032 98001 98188 98198 98023 98003 98148 98178 98022 98030 98031 98042 98055 98092 98146 98106 98058 98038 98108 98010 98118 98133 98155 98056 98166 98126 98019 98045 98014 98125 98059 98028 98034 98144 98024 98070 98011 98136 98065 98072 98107 98008 98117 98103 98177 98007 98116 98115 98027 98122 98029 98052 98053 98074 98077 98105 98033 98199 98102 98109 98075 98119 98006 98005 98112 98040 98004 98039

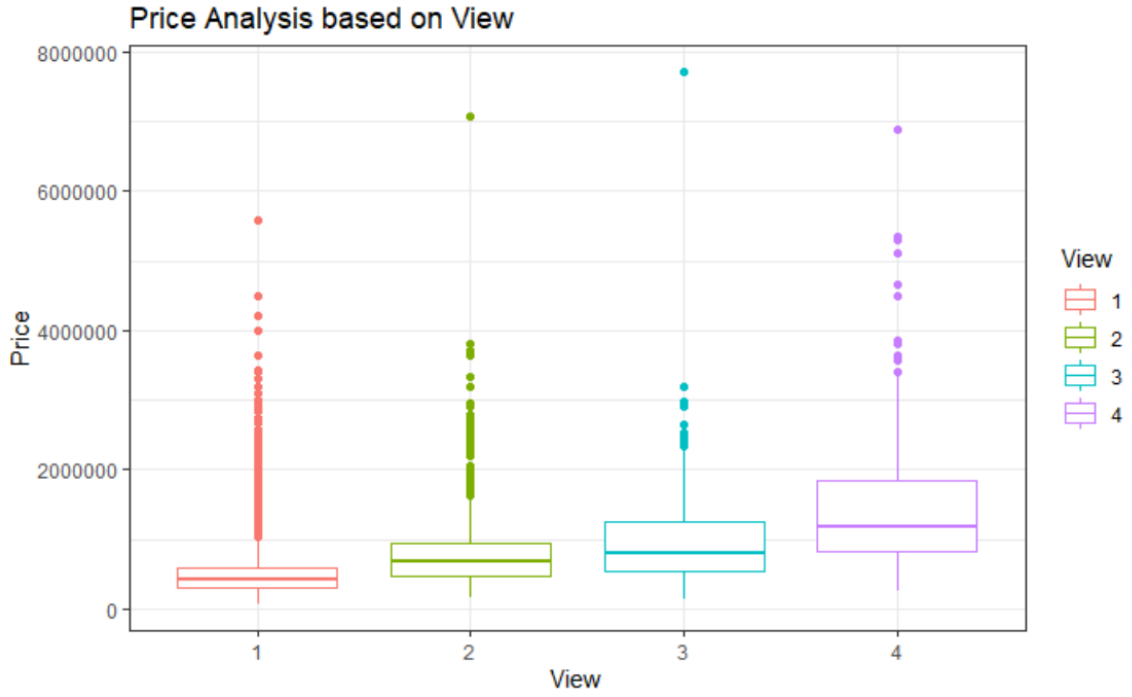
EDA – Condition



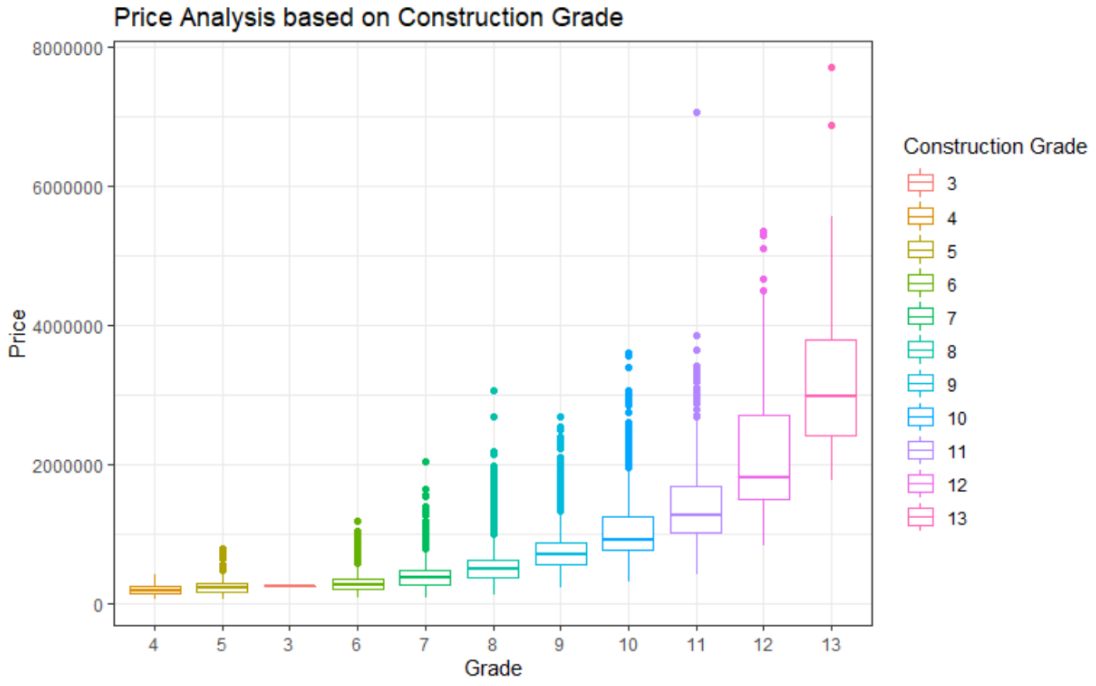


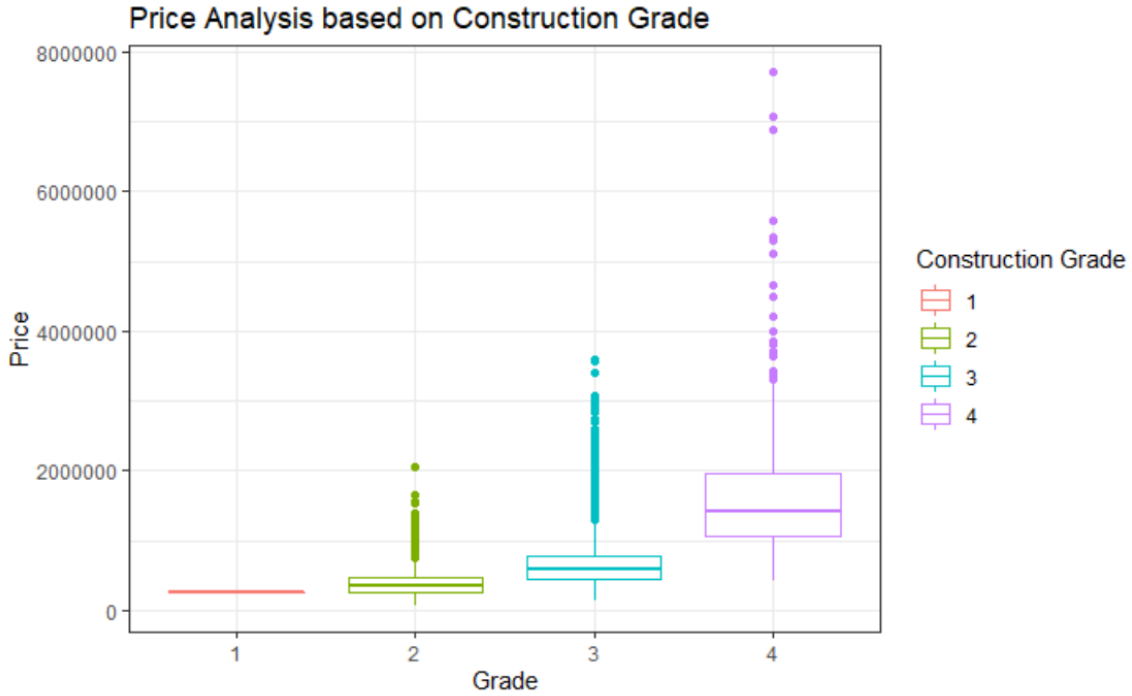
EDA – View


Price Analysis based on View
Price
8000000
6000000
4000000
2000000
0
0
2
1
3
4
View
View
0
1
2
3
4


Price Analysis based on View
Price
8000000
6000000
4000000
2000000
0
1
2
3
4
View
View
1
2
3
4

EDA – Grade


Price Analysis based on Construction Grade
8000000
6000000
4000000
2000000
0
Price
4
5
3
6
7
8
9
10
11
12
13
Grade
Construction Grade
3
4
5
6
7
8
9
10
11
12
13

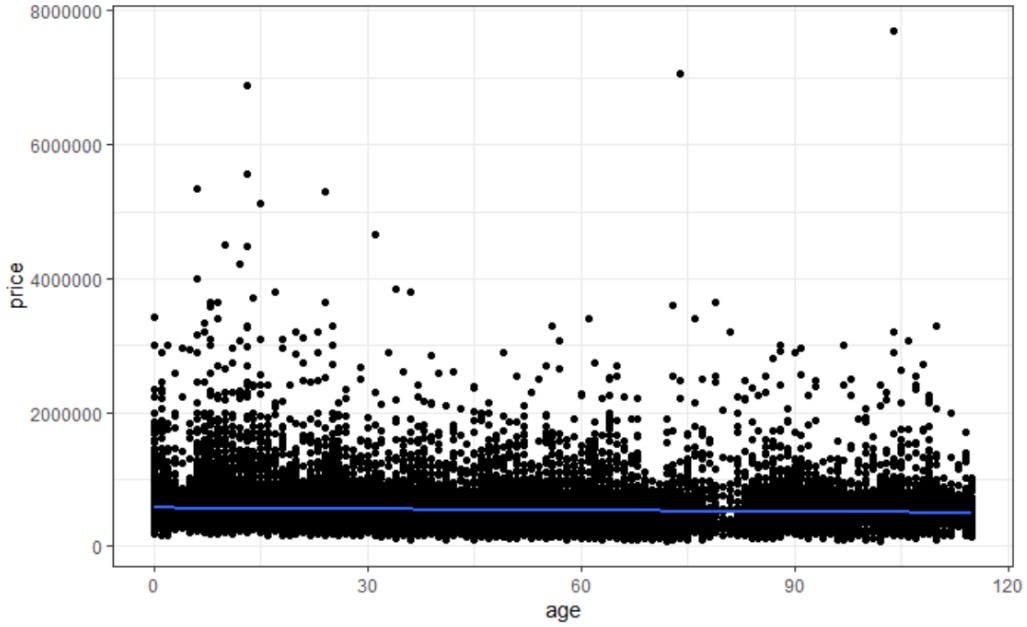

Price Analysis based on Construction Grade
8000000
6000000
4000000
2000000
0
Price
1
2
3
4
Grade
Construction Grade
1
2
3
4

EDA – Age and Renovated

Created age value from built year and sales year

Created categorical values if house was renovated

Some houses were sold before they were built or renovated.




```r
# getting house age
df %>%
  mutate(yr_sales = year(date),
         age = yr_sales - yr_built) ->
  df

df[,c("yr_sales", "age")] %>% glimpse()
```

```
Rows: 21,426
Columns: 2
$ yr_sales <dbl> 2015, 2014, 2014, 2015, 2015, 2015, 2015, 2014, 2014, 2014, 2015, 2014, 2014, 2014, 2014,
2014, 2014, 2014, 2014, 2014, 20~
$ age      <dbl> 24, 67, 62, 85, 64, 64, 55, 9, 69, 90, 90, 89, 112, 0, 113, 7, 73, 16, 16, 16, 9, 24, 24, 20,
20, 20, 21, 21, 17, 17, 19, ~
```

```r
# getting renovated before sales if there has been renovation
df %>%
  mutate(renovated = 0,
         renovated = if_else(yr_renovated != 0, 1, 0)) -> df

df[,c("age", "renovated")] %>% glimpse()
```

```
Rows: 21,426
Columns: 2
$ age       <dbl> 24, 67, 62, 85, 64, 64, 55, 9, 69, 90, 90, 89, 112, 0, 113, 7, 73, 16, 16, 16, 9, 24, 24, 20,
20, 20, 21, 21, 17, 17, 19,~
$ renovated <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 1, 0~
```

```r
summary(df$age)
df %>%
  filter(age<0) %>%
  select(date, yr_built, yr_sales, age)
```

R Console | tbl_df 12 x 4

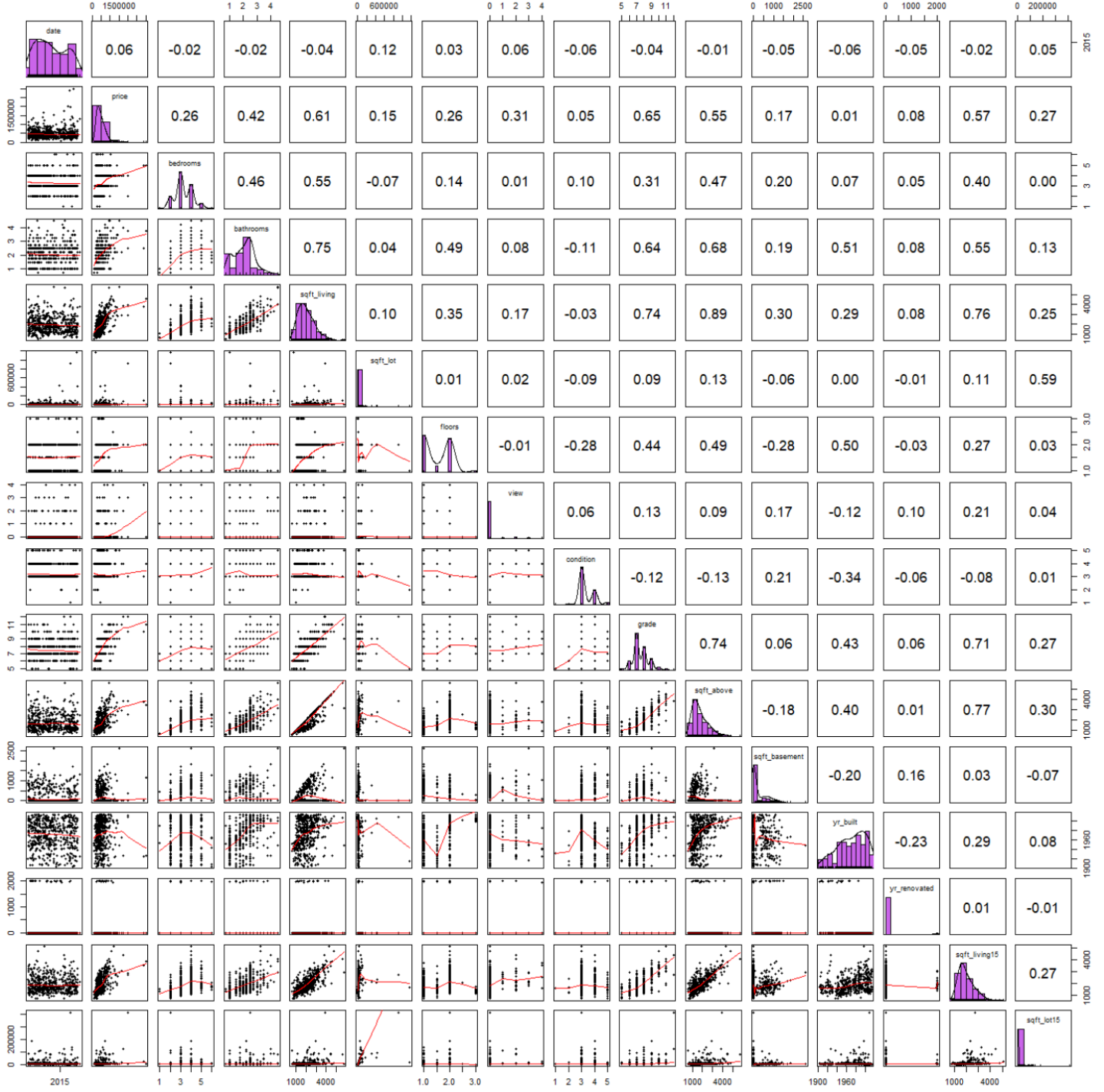
A tibble: 12 x 4

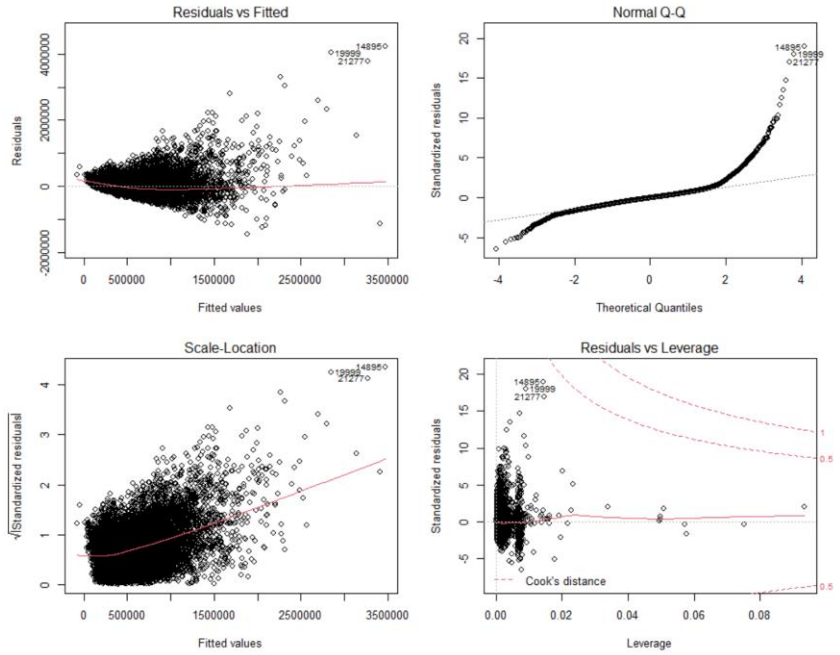
date <S3: POSIXct>	yr_built <dbl>	yr_sales <dbl>	age <dbl>
2014-06-24	2015	2014	-1
2014-07-09	2015	2014	-1
2014-06-25	2015	2014	-1
2014-08-26	2015	2014	-1
2014-10-29	2015	2014	-1
2014-11-25	2015	2014	-1
2014-07-31	2015	2014	-1
2014-05-20	2015	2014	-1
2014-06-17	2015	2014	-1
2014-08-01	2015	2014	-1

1-10 of 12 rows

Previous 1 2 Next

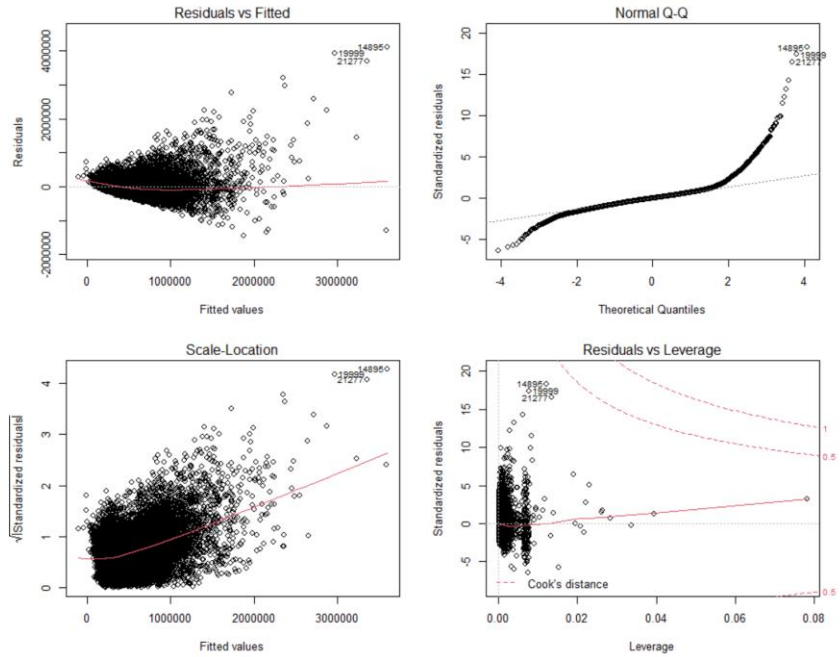
EDA – Quantitative





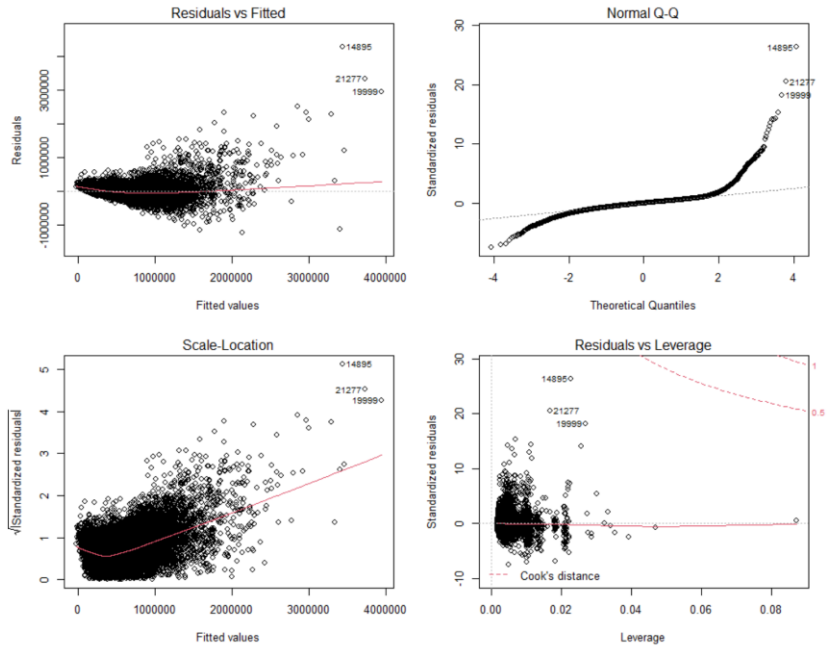

```
Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_living + sqft_lot +
    floors + waterfront + view + condition + grade + sqft_above +
    basement + sqft_living15 + sqft_lot15 + age + renovated,
    data = df)

Residuals:
      Min       1Q   Median       3Q      Max
-1458032  -116072    -8332    93749  4230250

Coefficients:
                Estimate    Std. Error t value             Pr(>|t|)
(Intercept)   -620722.96542   15922.66647 -38.984 < 0.0000000000000002 ***
bedrooms       -51118.03250    2206.74128 -23.164 < 0.0000000000000002 ***
bathrooms       51444.22018    3711.53066  13.861 < 0.0000000000000002 ***
sqft_living       175.01726       6.92817  25.262 < 0.0000000000000002 ***
sqft_lot           -0.01602       0.05341  -0.300                0.7642
floors          29578.52338    3967.99607   7.454  0.00000000000009380 ***
waterfront     530808.28493   19858.68481  26.729 < 0.0000000000000002 ***
view            80013.16644    3616.31435  22.126 < 0.0000000000000002 ***
condition       43734.25741    5540.56463   7.893  0.00000000000000308 ***
grade          116960.70944    4136.56571  28.275 < 0.0000000000000002 ***
sqft_above         48.50637       7.54640   6.428  0.00000000013223512 ***
basement        33333.18302    5830.44143   5.717  0.00000001098025551 ***
sqft_living15      51.57963       3.71216  13.895 < 0.0000000000000002 ***
sqft_lot15         -0.63881       0.08160  -7.829  0.00000000000000516 ***
age              3300.30120      72.05518  45.802 < 0.0000000000000002 ***
renovated       17948.05954    8082.44614   2.221                0.0264 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 224900 on 21410 degrees of freedom
Multiple R-squared:  0.6253,    Adjusted R-squared:  0.625
F-statistic:  2381 on 15 and 21410 DF,  p-value: < 0.00000000000000022
```

2nd Reg: without zipcode
Adjusted R-square 62%




```
Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_living + sqft_lot +
    floors + waterfront + view + condition + grade + sqft_above +
    basement + age + renovated, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1453465  -115383    -8420    93617  4108437

Coefficients:
               Estimate   Std. Error t value             Pr(>|t|)
(Intercept) -596416.3671   15818.3478 -37.704 < 0.0000000000000002 ***
bedrooms     -50938.1654    2216.4821 -22.982 < 0.0000000000000002 ***
bathrooms     50724.7202    3729.3407  13.602 < 0.0000000000000002 ***
sqft_living     186.3387       6.9002  27.005 < 0.0000000000000002 ***
sqft_lot         -0.3157       0.0383  -8.243 < 0.0000000000000002 ***
floors        22595.4265    3932.9725   5.745   0.0000000093097437 ***
waterfront   518344.3391   19951.9186  25.980 < 0.0000000000000002 ***
view          87075.7337    3599.1941  24.193 < 0.0000000000000002 ***
condition     41619.1592    5569.6416   7.473   0.0000000000000817 ***
grade        130501.9026    4051.9392  32.207 < 0.0000000000000002 ***
sqft_above       63.2726       7.5022   8.434 < 0.0000000000000002 ***
basement      35115.0167    5861.5720   5.991   0.0000000021225606 ***
age            3252.5134      72.1851  45.058 < 0.0000000000000002 ***
renovated     12556.9947    8119.8142   1.546                0.122
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 226200 on 21412 degrees of freedom
Multiple R-squared:  0.621,     Adjusted R-squared:  0.6208
F-statistic:  2699 on 13 and 21412 DF,  p-value: < 0.00000000000000022
```

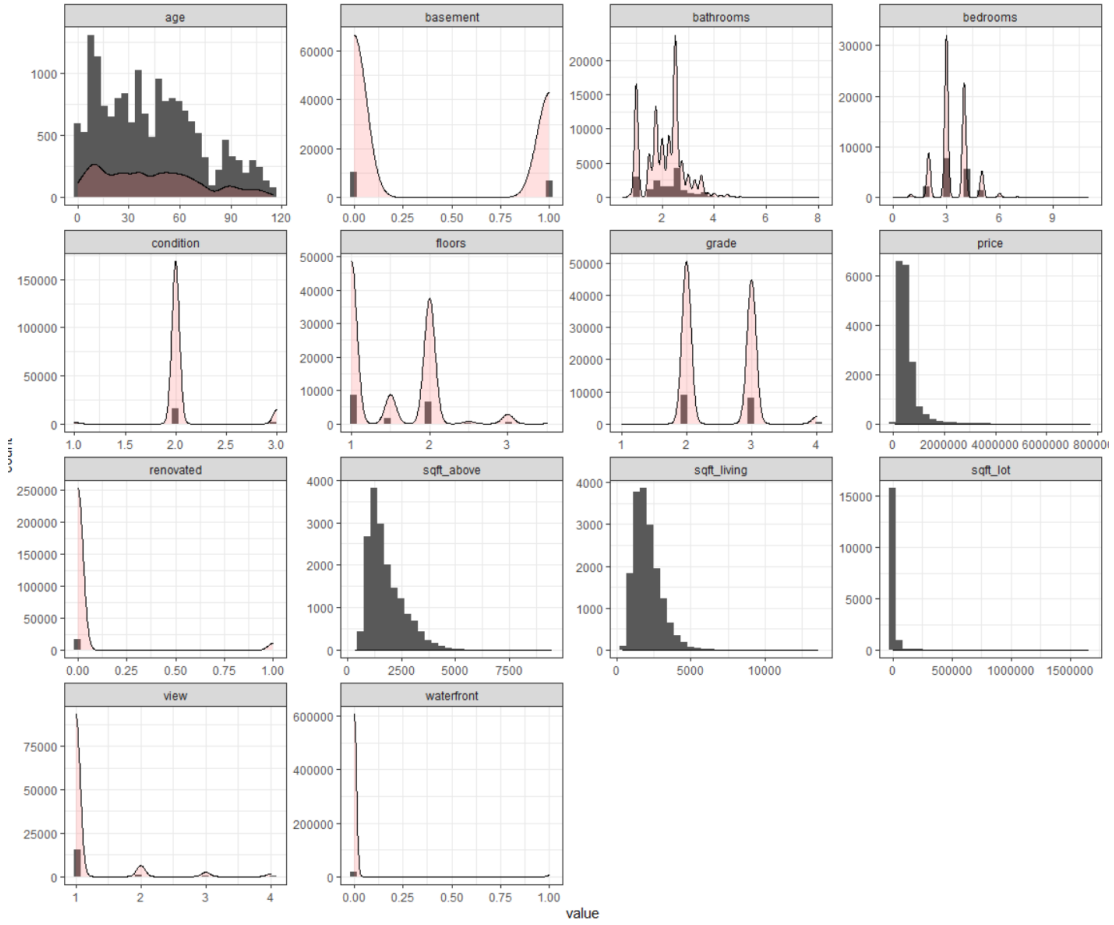
3rd Reg: without zipcode
Adjusted R-square 62%
Removing `sqft_living15` and `sqft_lot15` is not affecting the model significantly

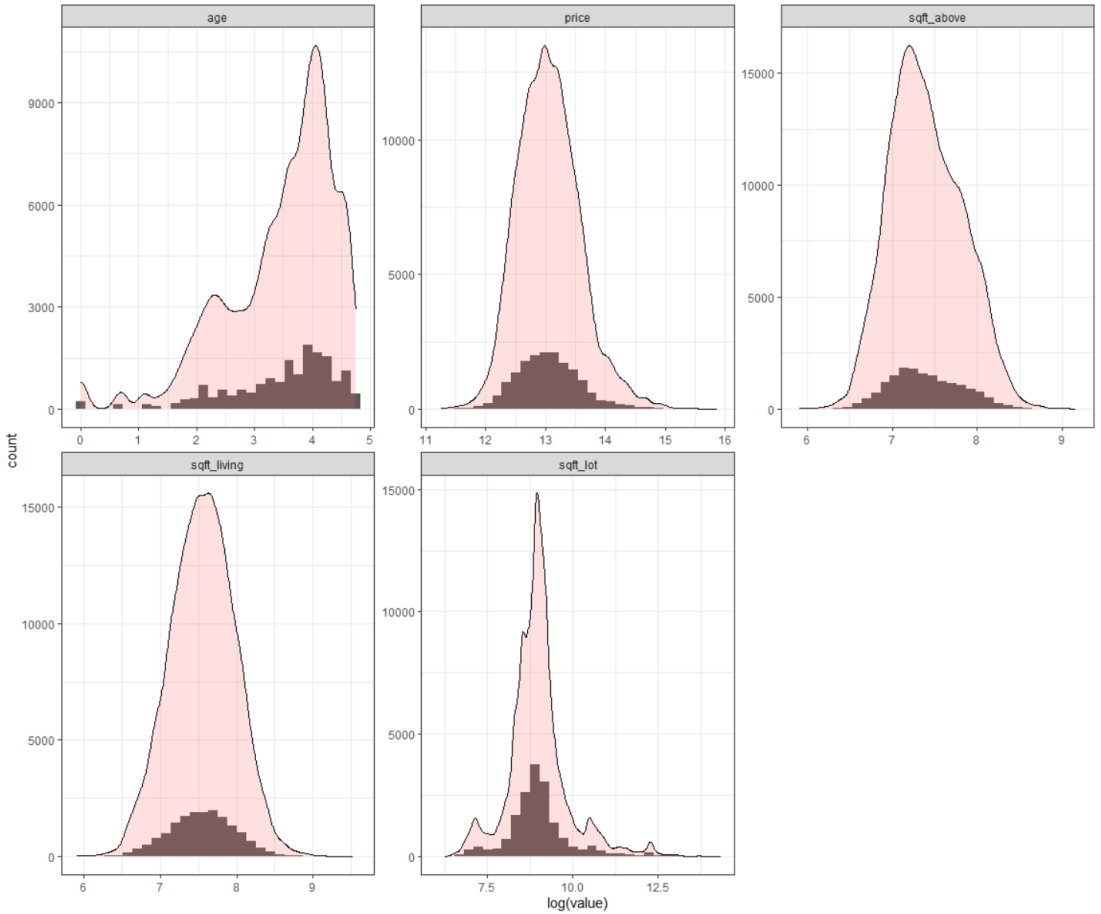


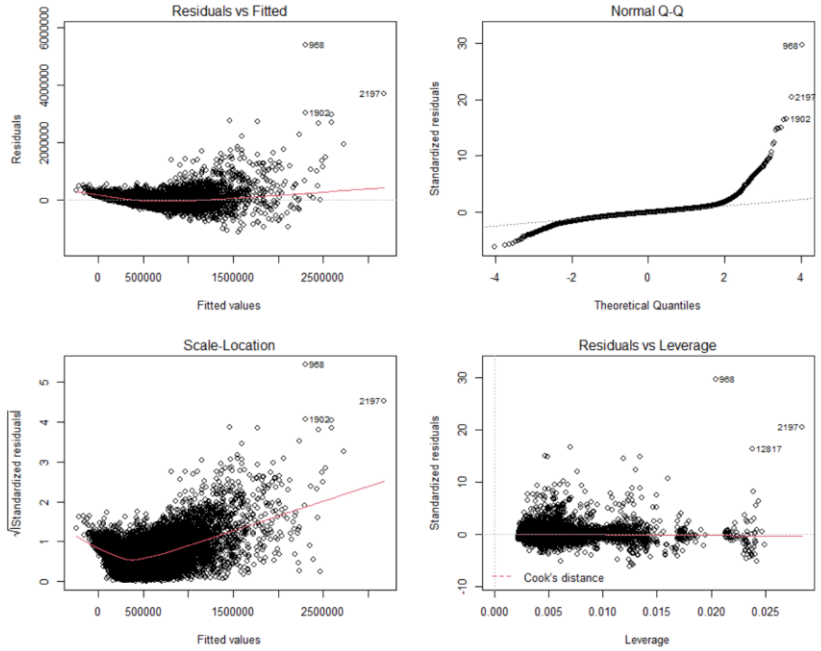

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 164100 on 21343 degrees of freedom
Multiple R-squared:  0.8012,    Adjusted R-squared:  0.8005
F-statistic:  1049 on 82 and 21343 DF,  p-value: < 0.00000000000000022
```

4th Reg: with zipcode
Adjusted R-square 80%

Feature Normalization

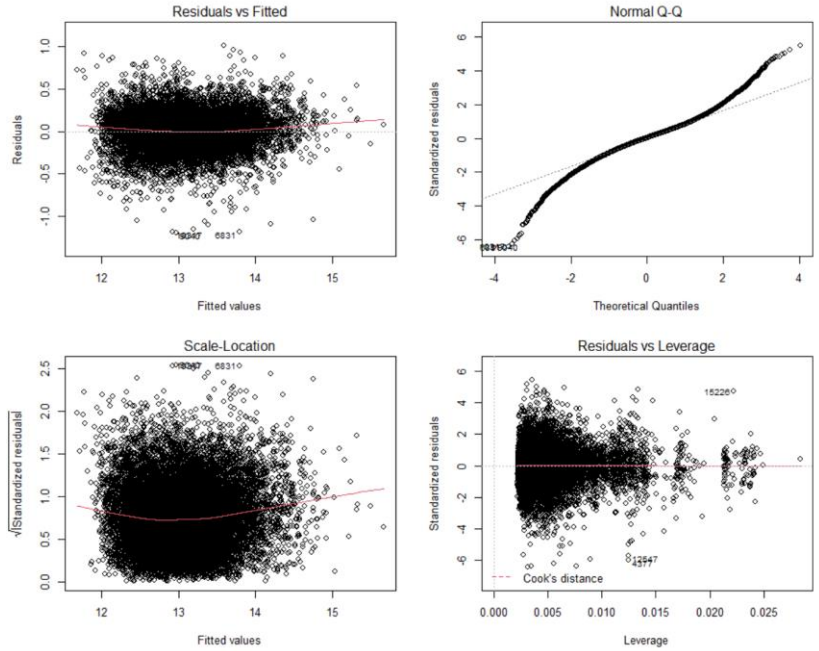






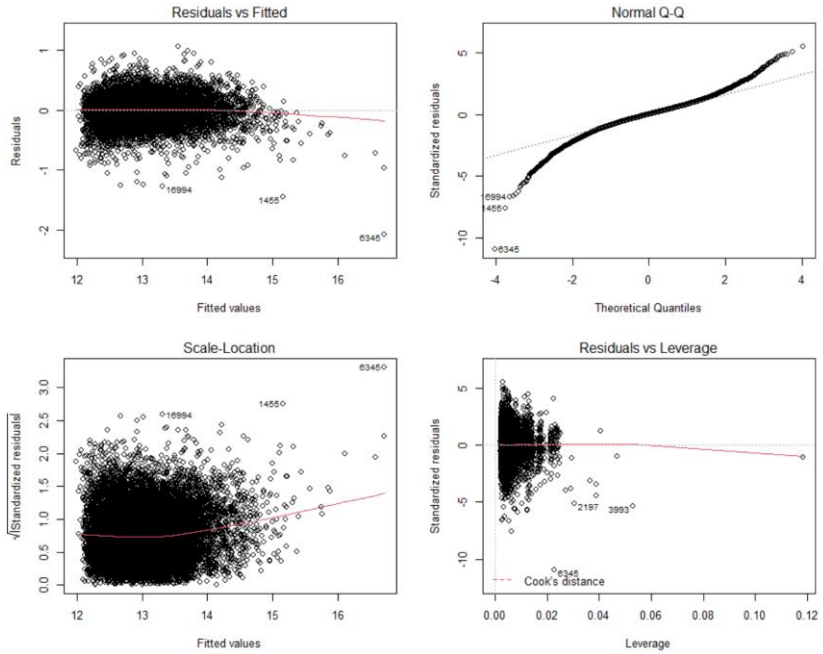

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 184100 on 17057 degrees of freedom
Multiple R-squared:  0.752,     Adjusted R-squared:  0.7508
F-statistic: 630.8 on 82 and 17057 DF,  p-value: < 0.00000000000000022
```

5th Reg: log(predictors) Adjusted R-square 75%!



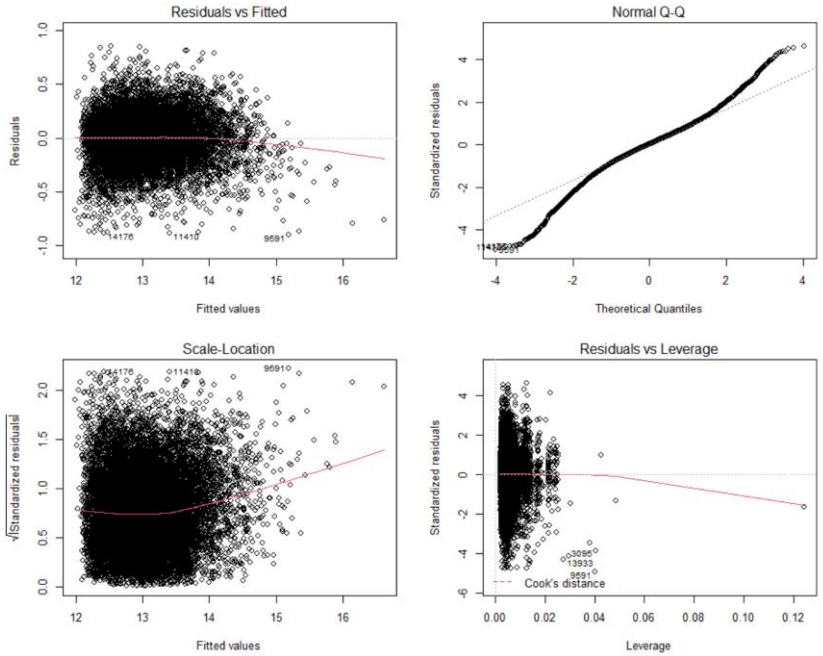

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1855 on 17057 degrees of freedom
Multiple R-squared:  0.8748,    Adjusted R-squared:  0.8742
F-statistic:  1454 on 82 and 17057 DF,  p-value: < 0.00000000000000022
```

6th Reg: log(response)~ log(predictors) Adjusted R-square 87%!



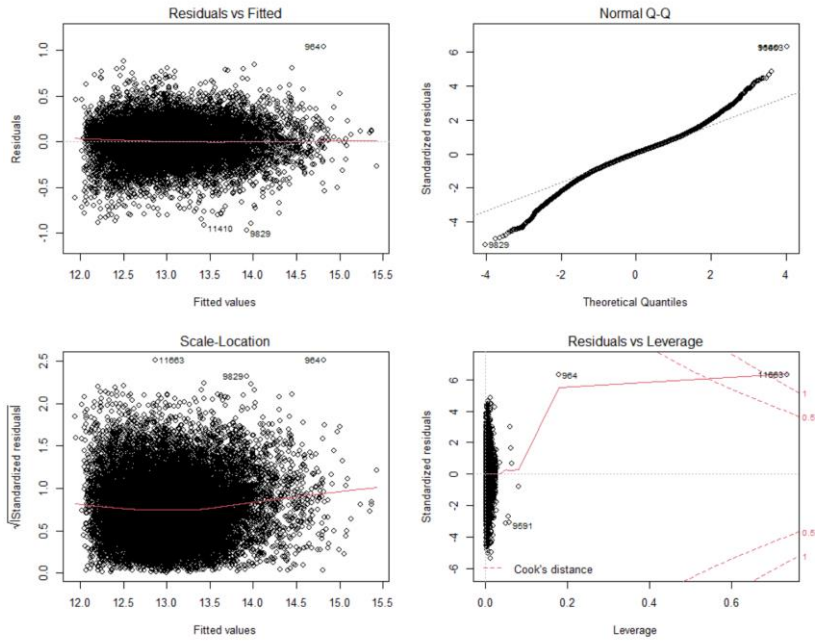

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1911 on 17057 degrees of freedom
Multiple R-squared:  0.8671,    Adjusted R-squared:  0.8665
F-statistic:  1358 on 82 and 17057 DF,  p-value: < 0.00000000000000022
```

7th Reg: log(response)~ predictors
Adjusted R-square 86%



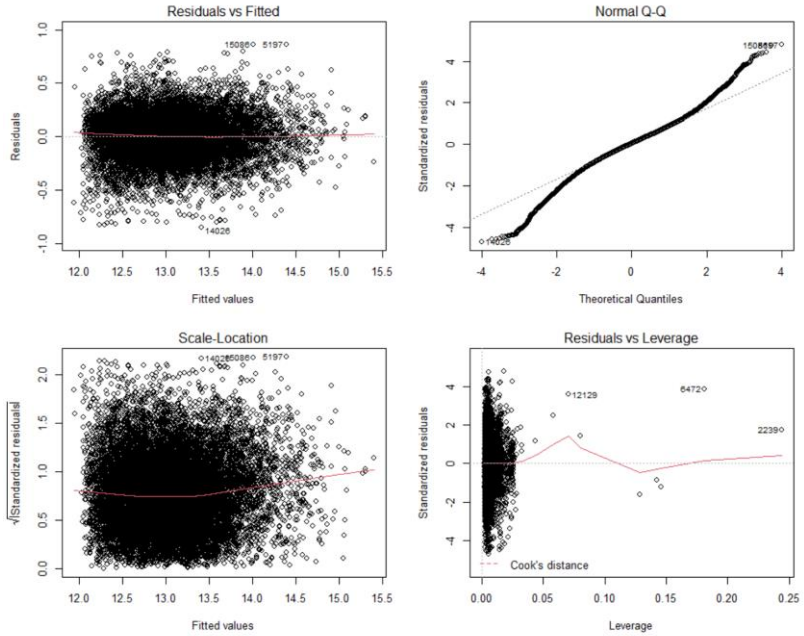

```
Residual standard error: 0.1862 on 17031 degrees of freedom
Multiple R-squared:  0.873,     Adjusted R-squared:  0.8724
F-statistic:  1427 on 82 and 17031 DF,  p-value: < 0.00000000000000022
```

8th Reg: log(response)~ predictors
Outlier Removal
Adjusted R-square 87%
There are more than 1000 high leverage points,
too many data points to lose



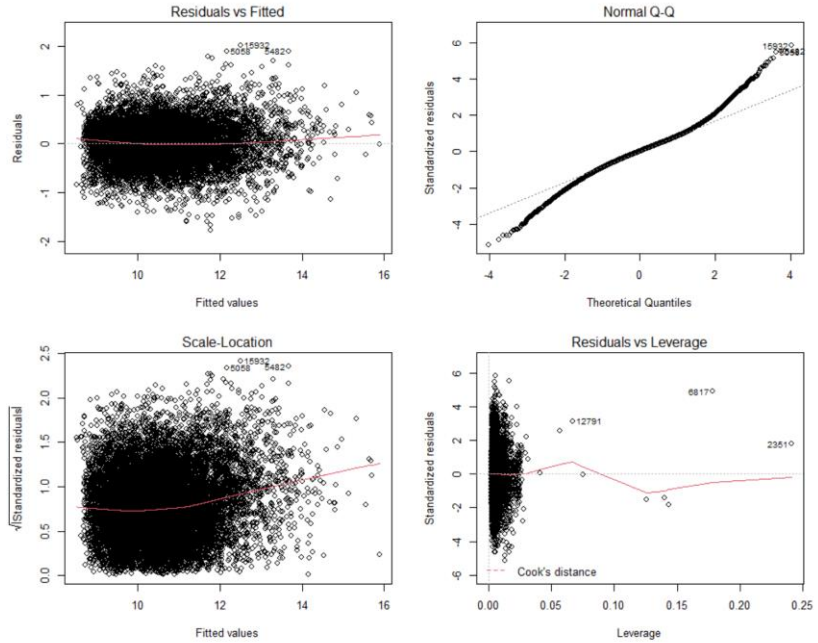

```
Call:
lm(formula = price_log ~ bedrooms + bathrooms + poly(sqft_living,
    2) + poly(sqft_lot, 2) + floors + waterfront + view + condition +
    grade + sqft_above + basement + poly(age, 3) + renovated +
    zipcode, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.96956 -0.09974  0.00674  0.10595  1.03721

Coefficients:
                          Estimate  Std. Error t value             Pr(>|t|)
(Intercept)           11.76941734  0.02412929 487.765 < 0.0000000000000002 ***
bedrooms              -0.01256648  0.00212115  -5.924  0.00000000319515003 ***
bathrooms              0.03964475  0.00344867  11.496 < 0.0000000000000002 ***
poly(sqft_living, 2)1 18.72926891  0.76567782  24.461 < 0.0000000000000002 ***
poly(sqft_living, 2)2 -5.06919145  0.20013787 -25.328 < 0.0000000000000002 ***
poly(sqft_lot, 2)1     4.18543718  0.20383474  20.533 < 0.0000000000000002 ***
poly(sqft_lot, 2)2    -1.30437377  0.19064286  -6.842  0.00000000000807676 ***
floors                -0.06997575  0.00427314 -16.376 < 0.0000000000000002 ***
waterfront             0.42766305  0.01831526  23.350 < 0.0000000000000002 ***
view                   0.11655785  0.00335931  34.697 < 0.0000000000000002 ***
condition              0.08811516  0.00508249  17.337 < 0.0000000000000002 ***
grade                  0.10650952  0.00387182  27.509 < 0.0000000000000002 ***
sqft_above             0.00014735  0.00000688  21.418 < 0.0000000000000002 ***
basement               0.02735434  0.00551305   4.962  0.00000070536258077 ***
poly(age, 3)1          1.25395790  0.31208436   4.018  0.00005894409110261 ***
poly(age, 3)2          2.41957603  0.24121027  10.031 < 0.0000000000000002 ***
poly(age, 3)3         -1.06066597  0.20374218  -5.206  0.00000019527917715 ***
renovated              0.04266421  0.00740087   5.765  0.00000000831915104 ***
zipcode98002          -0.04244527  0.01821871  -2.330              0.019830 *
zipcode98003           0.03945288  0.01656568   2.382              0.017248 *
zipcode98004           1.16235049  0.01597535  72.759 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1821 on 17027 degrees of freedom
Multiple R-squared:  0.8785,    Adjusted R-squared:  0.8779
F-statistic:  1432 on 86 and 17027 DF,  p-value: < 0.00000000000000022
```

9th and 10th Reg: log(response)~ Polynomial
Exhaustive Search and Stepwise
Adjusted R-square 87%

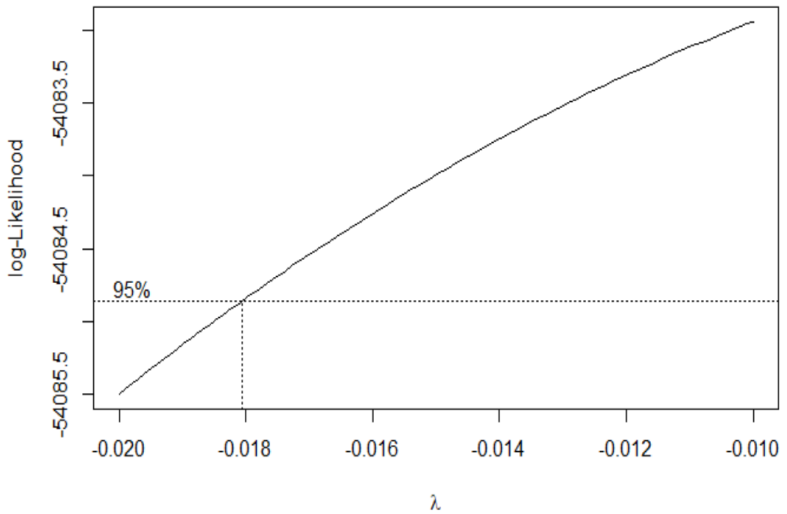



```
Call:
lm(formula = price_log ~ bedrooms + bathrooms + poly(sqft_living,
    2) + poly(sqft_lot, 2) + floors + waterfront + view + condition +
    grade + sqft_above + basement + poly(age, 3) + renovated +
    zipcode, data = train_nolev)

Residuals:
     Min       1Q   Median       3Q      Max
-0.85133 -0.10007  0.00655  0.10627  0.85999

Coefficients:
                          Estimate   Std. Error t value             Pr(>|t|)
(Intercept)           11.766446068  0.024652678 477.289 < 0.0000000000000002 ***
bedrooms              -0.012890268  0.002170869  -5.938  0.00000000294740665 ***
bathrooms              0.040362873  0.003516287  11.479 < 0.0000000000000002 ***
poly(sqft_living, 2)1 17.902949182  0.757720431  23.627 < 0.0000000000000002 ***
poly(sqft_living, 2)2 -4.977921606  0.199907541 -24.901 < 0.0000000000000002 ***
poly(sqft_lot, 2)1     4.255708546  0.204503488  20.810 < 0.0000000000000002 ***
poly(sqft_lot, 2)2    -1.710217331  0.191611919  -8.925 < 0.0000000000000002 ***
floors                -0.071283611  0.004366329 -16.326 < 0.0000000000000002 ***
waterfront             0.412488872  0.019468385  21.188 < 0.0000000000000002 ***
view                   0.118359776  0.003425588  34.552 < 0.0000000000000002 ***
condition              0.089064477  0.005200054  17.128 < 0.0000000000000002 ***
grade                  0.105905090  0.003952972  26.791 < 0.0000000000000002 ***
sqft_above             0.000148550  0.000007041  21.098 < 0.0000000000000002 ***
basement               0.025307365  0.005648996   4.480  0.00000751650775539 ***
poly(age, 3)1          1.124764544  0.310846386   3.618             0.000297 ***
poly(age, 3)2          2.504588150  0.240329081  10.421 < 0.0000000000000002 ***
poly(age, 3)3         -1.156290408  0.202609147  -5.707  0.00000001170072109 ***
renovated              0.044951340  0.007553131   5.951  0.00000000271450892 ***
zipcode98002          -0.041078101  0.018482662  -2.223             0.026262 *
zipcode98003           0.046168008  0.016899758   2.732             0.006304 **
zipcode98004           1.165825843  0.016254559  71.723 < 0.0000000000000002 ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1809 on 16111 degrees of freedom
Multiple R-squared:  0.879,     Adjusted R-squared:  0.8783
F-statistic:  1360 on 86 and 16111 DF,  p-value: < 0.00000000000000022
```

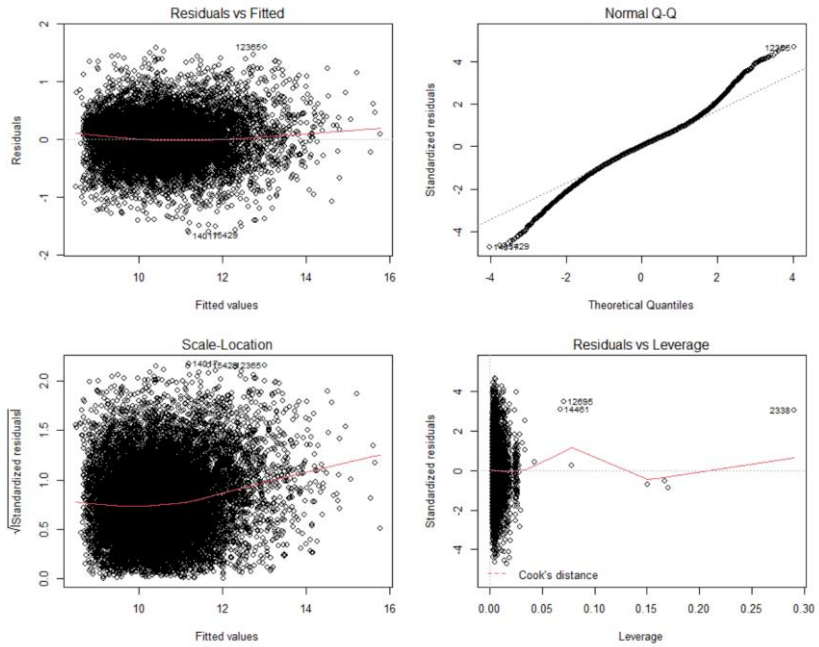
11th Reg: log(response)~ Polynomial
Outlier and High Leverage Removal (900 points)
Adjusted R-square 87%




```
Call:
lm(formula = price^(0.18) ~ bedrooms + bathrooms + poly(sqft_living,
    2) + poly(sqft_lot, 2) + floors + waterfront + view + condition +
    grade + sqft_above + basement + poly(age, 3) + renovated +
    zipcode, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.77878 -0.19799  0.00758  0.19853  2.01939

Coefficients:
                          Estimate  Std. Error t value             Pr(>|t|)
(Intercept)           8.08676657  0.04608429 175.478 < 0.0000000000000002 ***
bedrooms             -0.02879308  0.00405593  -7.099      0.00000000000131 ***
bathrooms             0.07912180  0.00658643  12.013 < 0.0000000000000002 ***
poly(sqft_living, 2)1 37.59053725  1.45832315  25.777 < 0.0000000000000002 ***
poly(sqft_living, 2)2 -5.45566704  0.38322289 -14.236 < 0.0000000000000002 ***
poly(sqft_lot, 2)1     7.65729234  0.39313691  19.477 < 0.0000000000000002 ***
poly(sqft_lot, 2)2    -3.02967878  0.36825968  -8.227 < 0.0000000000000002 ***
floors                -0.13687668  0.00815872 -16.777 < 0.0000000000000002 ***
waterfront             0.94923630  0.03510478  27.040 < 0.0000000000000002 ***
view                   0.23973345  0.00641680  37.360 < 0.0000000000000002 ***
condition              0.16962769  0.00970183  17.484 < 0.0000000000000002 ***
grade                  0.20691415  0.00738999  27.999 < 0.0000000000000002 ***
sqft_above             0.00029134  0.00001313  22.184 < 0.0000000000000002 ***
basement               0.04936626  0.01052651   4.690      0.00000275721480 ***
poly(age, 3)1          3.09606610  0.59630998   5.192      0.00000021039467 ***
poly(age, 3)2          4.48283531  0.46103944   9.723 < 0.0000000000000002 ***
poly(age, 3)3         -2.33445349  0.38899185  -6.001      0.00000000199723 ***
renovated              0.08325353  0.01415074   5.883      0.00000000409627 ***
zipcode98002          -0.05515363  0.03477331  -1.586              0.112737
zipcode98003           0.05936170  0.03161744   1.877              0.060467 .
zipcode98004           2.28078263  0.03050052  74.778 < 0.0000000000000002 ***
zipcode98005           1.45559186  0.03760165  38.711 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3476 on 17019 degrees of freedom
Multiple R-squared:  0.8825,    Adjusted R-squared:  0.8819
F-statistic:  1487 on 86 and 17019 DF,  p-value: < 0.00000000000000022
```



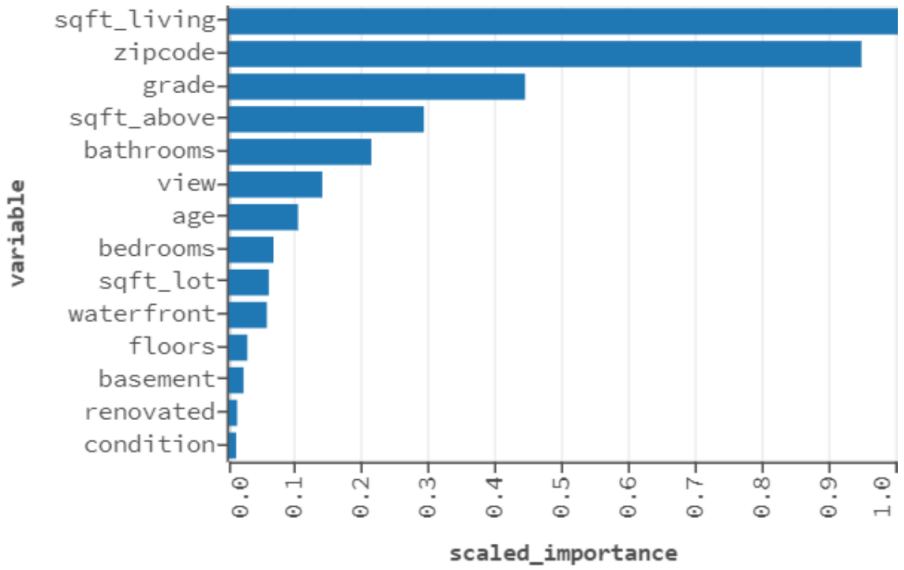
12th Reg: BOXCOX(response)~ Polynomial
Adjusted R-square 88%




```
Call:
lm(formula = price^(0.18) ~ bedrooms + bathrooms + poly(sqft_living,
    2) + poly(sqft_lot, 2) + floors + waterfront + view + condition +
    grade + sqft_above + basement + poly(age, 3) + renovated +
    zipcode, data = train_nolev)

Residuals:
     Min       1Q   Median       3Q      Max
-1.61530 -0.19695  0.00797  0.19809  1.59877

Coefficients:
                      Estimate  Std. Error t value             Pr(>|t|)
(Intercept)           8.08492012  0.04584602 176.349 < 0.0000000000000002 ***
bedrooms             -0.02750955  0.00403656  -6.815      0.00000000000974 ***
bathrooms             0.07785822  0.00655666  11.875 < 0.0000000000000002 ***
poly(sqft_living, 2)1 36.77546219  1.43966445  25.544 < 0.0000000000000002 ***
poly(sqft_living, 2)2 -5.76394354  0.38054801 -15.146 < 0.0000000000000002 ***
poly(sqft_lot, 2)1     7.47916121  0.39055943  19.150 < 0.0000000000000002 ***
poly(sqft_lot, 2)2    -3.45557714  0.36423553  -9.487 < 0.0000000000000002 ***
floors                -0.13695541  0.00811959 -16.867 < 0.0000000000000002 ***
waterfront             0.99375792  0.03504688  28.355 < 0.0000000000000002 ***
view                   0.23685392  0.00640692  36.968 < 0.0000000000000002 ***
condition              0.16980471  0.00966074  17.577 < 0.0000000000000002 ***
grade                  0.20483377  0.00734925  27.871 < 0.0000000000000002 ***
sqft_above             0.00029367  0.00001308  22.458 < 0.0000000000000002 ***
basement               0.05111047  0.01048667   4.874      0.00000110438040 ***
poly(age, 3)1          3.02986325  0.59097237   5.127      0.00000029778829 ***
poly(age, 3)2          4.58193398  0.45660466  10.035 < 0.0000000000000002 ***
poly(age, 3)3         -2.23213109  0.38534083  -5.793      0.00000000705319 ***
renovated              0.07701452  0.01408597   5.467      0.00000004630001 ***
zipcode98002          -0.05456904  0.03453796  -1.580              0.114132
zipcode98003           0.06022302  0.03138478   1.919              0.055019 .
zipcode98004           2.28791499  0.03032386  75.449 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3443 on 16877 degrees of freedom
Multiple R-squared:  0.8835,    Adjusted R-squared:  0.8829
F-statistic:  1488 on 86 and 16877 DF,  p-value: < 0.00000000000000022
```

13th Reg: BOXCOX(response)~ Polynomial
Outlier Removal
Adjusted R-square 88%
118 leverage points
Performs 87% R-Square on Test DATASET

▾ VARIABLE IMPORTANCES



	mean	*sd*	*cv_1_valid*	*cv_2_valid*	*cv_3_valid*	*cv_4_valid*	*cv_5_valid*
mae	75994.95	2334.9062	75622.1	75988.51	77046.67	72470.016	78847.46
mean_residual_deviance	2.05737226E10	3.60085504E9	1.83930225E10	2.32981811E10	2.34940068E10	1.52949688E10	2.2388439E10
mse	2.05737226E10	3.60085504E9	1.83930225E10	2.32981811E10	2.34940068E10	1.52949688E10	2.2388439E10
r2	0.8492802	0.016802097	0.8642381	0.82674164	0.8367997	0.8638124	0.8548091
residual_deviance	2.05737226E10	3.60085504E9	1.83930225E10	2.32981811E10	2.34940068E10	1.52949688E10	2.2388439E10
rmse	142967.27	12946.2295	135620.88	152637.42	153277.55	123672.83	149627.67
rmsle	0.18672153	0.0030231932	0.18507983	0.18807924	0.18926743	0.18220085	0.18898031

14th Model: Random Forest

Thank You