R Notebook

Nelson Brown and Bryce Smith

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE)
```

Introduction

Summary Statistics and Graphics

Quantitative Values

Table 1: First Four Rows for Quantitative Values on Seattlle Housing Dataframe

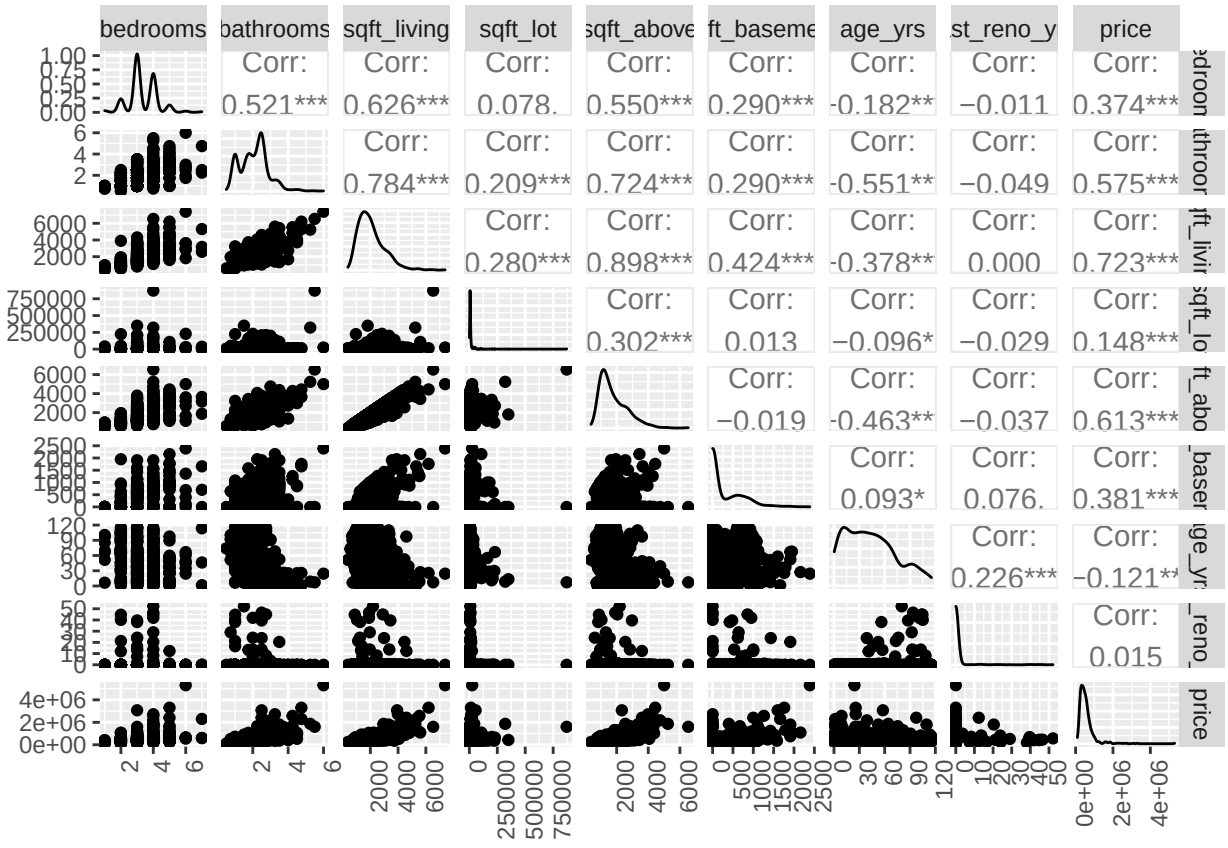
bedrooms	bathrooms	sqft_living	sqft_lot	sqft_above
3	1.750	1,570	6,975	1,040
5	3.750	3,050	8,972	3,050
3	1.750	1,570	12,506	1,570
4	1.750	1,390	10,660	1,030

Table 2: First Four Rows for Quantitative Values on Seattlle Housing Dataframe

sqft_basement	age_yrs	last_reno_yrs	price
530	35.712	0	359,950
0	1.288	0	909,950
0	56.118	0	318,000
360	54.751	0	272,000

Table 3: Summary Statistics for Values on Seattle Housing Dataframe

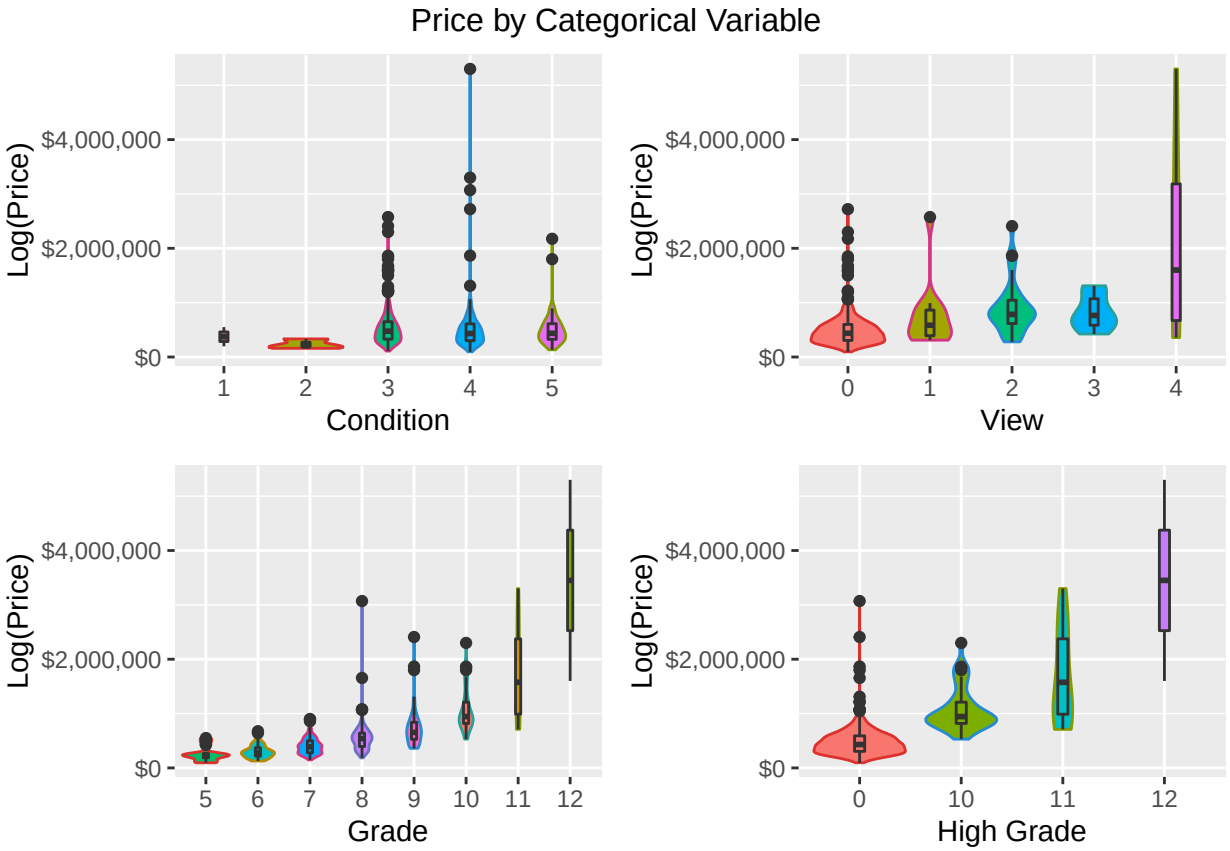
Statistic	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
price	545,427.700	408,545.900	95,000	315,000	631,500	5,300,000
bedrooms	3.352	0.876	1	3	4	7
bathrooms	2.092	0.805	0.500	1.500	2.500	6.000
sqft_living	2,073.669	963.763	380	1,370	2,550	7,390
sqft_lot	15,967.970	46,698.890	740	5,100	10,585	871,200
floors	1.479	0.535	1	1	2	3
waterfront	0.010	0.099	0	0	0	1
view	0.204	0.695	0	0	0	4
condition	3.388	0.641	1	3	4	5
grade	7.635	1.217	5	7	8	12
sqft_above	1,793.571	873.153	380	1,130	2,313	6,530
sqft_basement	280.098	424.835	0	0	570	2,390
yr_built	1,971.210	29.939	1,900	1,951	1,998	2,015
yr_renovated	84.527	401.973	0	0	0	2,014
age_yrs	43.669	29.943	0.274	16.830	64.219	115.381
last_reno_yrs	0.932	5.554	0	0	0	52

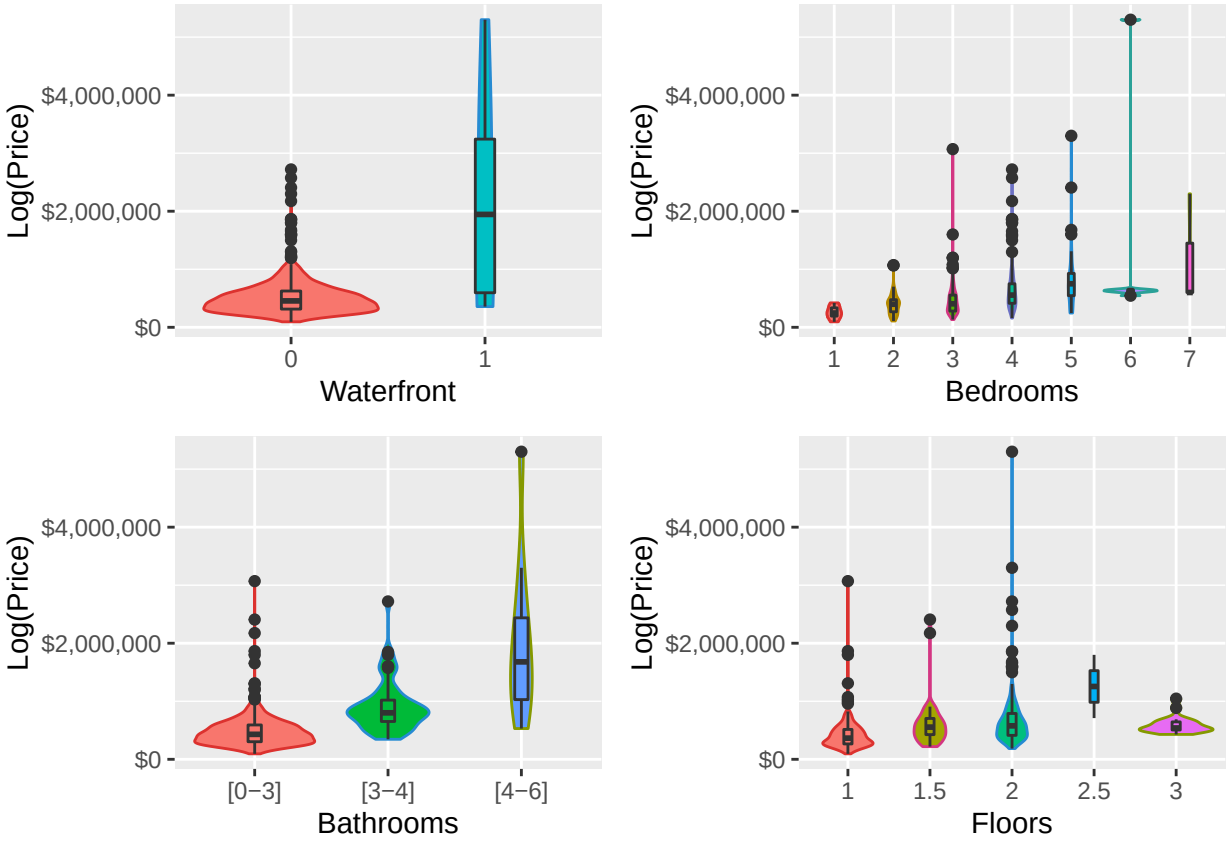


Discrete and Categorical Values

```
##        id                  date                price            bedrooms
```

```
##  Min.   :3.600e+06   Min.   :2014-05-02   Min.   :  95000   Min.   :1.000
##  1st Qu.:2.029e+09   1st Qu.:2014-07-28   1st Qu.: 315000   1st Qu.:3.000
##  Median :3.887e+09   Median :2014-10-29   Median : 459000   Median :3.000
##  Mean   :4.594e+09   Mean   :2014-11-07   Mean   : 545428   Mean   :3.352
##  3rd Qu.:7.385e+09   3rd Qu.:2015-03-02   3rd Qu.: 631500   3rd Qu.:4.000
##  Max.   :9.834e+09   Max.   :2015-05-13   Max.   :5300000   Max.   :7.000
##
##    bathrooms      sqft_living      sqft_lot          floors      waterfront
##  Min.   :0.500   Min.   : 380   Min.   :   740   Min.   :1.000   0:607
##  1st Qu.:1.500   1st Qu.:1370   1st Qu.:  5100   1st Qu.:1.000   1:  6
##  Median :2.250   Median :1890   Median :  7428   Median :1.000
##  Mean   :2.092   Mean   :2074   Mean   : 15968   Mean   :1.479
##  3rd Qu.:2.500   3rd Qu.:2550   3rd Qu.: 10585   3rd Qu.:2.000
##  Max.   :6.000   Max.   :7390   Max.   :871200   Max.   :3.000
##
##  view    condition     grade          sqft_above   sqft_basement
##  0:556   1:  2     Min.   : 5.000   Min.   : 380   Min.   :   0.0
##  1: 12   2:  3     1st Qu.: 7.000   1st Qu.:1130   1st Qu.:   0.0
##  2: 29   3:407     Median : 7.000   Median :1540   Median :   0.0
##  3:  9   4:157     Mean   : 7.635   Mean   :1794   Mean   : 280.1
##  4:  7   5: 44     3rd Qu.: 8.000   3rd Qu.:2313   3rd Qu.: 570.0
##                    Max.   :12.000   Max.   :6530   Max.   :2390.0
##
##     yr_built     yr_renovated       date_built            age_yrs
##  Min.   :1900   Min.   :   0.00   Min.   :1900-01-01   Min.   :  0.274
##  1st Qu.:1951   1st Qu.:   0.00   1st Qu.:1951-01-01   1st Qu.: 16.830
##  Median :1976   Median :   0.00   Median :1976-01-01   Median : 38.962
##  Mean   :1971   Mean   :  84.53   Mean   :1971-03-18   Mean   : 43.669
##  3rd Qu.:1998   3rd Qu.:   0.00   3rd Qu.:1998-01-01   3rd Qu.: 64.219
##  Max.   :2015   Max.   :2014.00   Max.   :2015-01-01   Max.   :115.381
##
##  last_reno_yrs     renovated has_basement high_grade  grade.factor view.factor
##  Min.   : 0.0000   0:587     0:371        0 :564     7      :234   0:556
##  1st Qu.: 0.0000   1: 26     1:242        10: 36     8      :168   1: 12
##  Median : 0.0000                          11: 11     9      : 78   2: 29
##  Mean   : 0.9316                          12:  2     6      : 71   3:  9
##  3rd Qu.: 0.0000                                     10     : 36   4:  7
##  Max.   :52.1836                                     5      : 13
##                                                      (Other): 13
##  multistory
##  0:314
##  1:299
##
##
##
##
##
```


Price by Categorical Variable
Log(Price)
$4,000,000
$2,000,000
$0
1
2
3
4
5
Condition
0
1
2
3
4
View
5
6
7
8
9
10
11
12
Grade
0
10
11
12
High Grade



Analysis

Initial Model

```
##
## Call:
## lm(formula = price ~ sqft_living, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -693119 -144669  -22906  107286 3125034
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -90183.99   27085.02   -3.33 0.000922 ***
## sqft_living    306.52      11.85   25.87  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 282400 on 611 degrees of freedom
## Multiple R-squared:  0.5228, Adjusted R-squared:  0.5221
## F-statistic: 669.5 on 1 and 611 DF,  p-value: < 2.2e-16

##
```

```
## Call:
## lm(formula = price ~ sqft_living + sqft_above, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -779573 -152200  -21701  109006 3027541
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -80830.85   27099.37  -2.983  0.00297 **
## sqft_living    377.49      26.71  14.133  < 2e-16 ***
## sqft_above     -87.28      29.48  -2.960  0.00319 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 280700 on 610 degrees of freedom
## Multiple R-squared:  0.5296, Adjusted R-squared:  0.528
## F-statistic: 343.4 on 2 and 610 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = price ~ sqft_living + sqft_above + sqft_basement,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -779573 -152200  -21701  109006 3027541
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -80830.85   27099.37  -2.983  0.00297 **
## sqft_living      377.49      26.71  14.133  < 2e-16 ***
## sqft_above       -87.28      29.48  -2.960  0.00319 **
## sqft_basement        NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 280700 on 610 degrees of freedom
## Multiple R-squared:  0.5296, Adjusted R-squared:  0.528
## F-statistic: 343.4 on 2 and 610 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = price ~ sqft_living + sqft_above + has_basement,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -824034 -153800  -23019  104748 2945567
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -57604.61   29882.09  -1.928  0.05435 .
## sqft_living      449.44      47.53   9.456  < 2e-16 ***
```

```
## sqft_above      -166.51      52.38  -3.179  0.00155 **
## has_basement1 -76778.69   41993.06  -1.828  0.06798 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 280100 on 609 degrees of freedom
## Multiple R-squared:  0.5322, Adjusted R-squared:  0.5299
## F-statistic: 230.9 on 3 and 609 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = price ~ sqft_living + sqft_above + age_yrs, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -767931 -140465  -17303  111826 2984774
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -241072.81   39685.96  -6.075 2.19e-09 ***
## sqft_living     363.90      26.23  13.874  < 2e-16 ***
## sqft_above      -37.65      30.24  -1.245    0.214
## age_yrs        2276.56     419.83   5.423 8.48e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 274400 on 609 degrees of freedom
## Multiple R-squared:  0.5513, Adjusted R-squared:  0.549
## F-statistic: 249.4 on 3 and 609 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = price ~ sqft_living + sqft_above + date_built, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -768339 -139472  -17403  112369 2986015
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.387e+05  2.858e+04  -4.853 1.55e-06 ***
## sqft_living  3.639e+02  2.623e+01  13.870  < 2e-16 ***
## sqft_above  -3.774e+01  3.025e+01  -1.248    0.213
## date_built  -6.212e+00  1.150e+00  -5.402 9.44e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 274400 on 609 degrees of freedom
## Multiple R-squared:  0.5511, Adjusted R-squared:  0.5489
## F-statistic: 249.2 on 3 and 609 DF,  p-value: < 2.2e-16

##
## Call:
```

```
## lm(formula = price ~ sqft_living + sqft_above + age_yrs + grade,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -595278 -136739   -8991  107606 2990896
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -955046.92  103637.05  -9.215  < 2e-16 ***
## sqft_living     301.26      26.53  11.358  < 2e-16 ***
## sqft_above      -88.47      29.78  -2.970  0.00309 **
## age_yrs        3286.90     424.91   7.735 4.30e-14 ***
## grade        116692.21   15756.21   7.406 4.36e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 263000 on 608 degrees of freedom
## Multiple R-squared:  0.5884, Adjusted R-squared:  0.5857
## F-statistic: 217.3 on 4 and 608 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = price ~ sqft_living + sqft_above + age_yrs + high_grade,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1397234  -134651   -20796   108534  1985030
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -84178.48   39237.16  -2.145 0.032319 *
## sqft_living      310.84      24.19  12.851  < 2e-16 ***
## sqft_above       -69.04      27.65  -2.497 0.012798 *
## age_yrs         1871.40     379.13   4.936 1.03e-06 ***
## high_grade10  175318.89   49565.17   3.537 0.000435 ***
## high_grade11  576221.69   84477.96   6.821 2.19e-11 ***
## high_grade12 1989546.00  182625.26  10.894  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 246500 on 606 degrees of freedom
## Multiple R-squared:  0.6395, Adjusted R-squared:  0.6359
## F-statistic: 179.1 on 6 and 606 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = price ~ sqft_living + sqft_above + age_yrs + high_grade +
##     view, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1064707  -122384   -19128   104309  1296242
```

```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -26399.81   37439.11  -0.705 0.480996
## sqft_living      254.32      23.64  10.757  < 2e-16 ***
## sqft_above       -36.33      26.23  -1.385 0.166508
## age_yrs         1376.16     361.17   3.810 0.000153 ***
## high_grade10  185171.59   46306.93   3.999 7.16e-05 ***
## high_grade11  544940.87   79318.77   6.870 1.60e-11 ***
## high_grade12 1714652.50  175095.97   9.793  < 2e-16 ***
## view1         207771.30   67908.43   3.060 0.002315 **
## view2         153658.02   45714.75   3.361 0.000825 ***
## view3         179076.06   78703.51   2.275 0.023237 *
## view4         815872.06   93750.56   8.703  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 230200 on 602 degrees of freedom
## Multiple R-squared:  0.6876, Adjusted R-squared:  0.6824
## F-statistic: 132.5 on 10 and 602 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = price ~ sqft_living + age_yrs + high_grade + view,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1085910  -123495   -20915   104233  1298809
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -42198.18   35686.77  -1.182 0.237489
## sqft_living      227.66      13.74  16.566  < 2e-16 ***
## age_yrs         1518.48     346.51   4.382 1.39e-05 ***
## high_grade10  177967.17   46048.96   3.865 0.000123 ***
## high_grade11  526657.52   78272.41   6.729 3.99e-11 ***
## high_grade12 1706267.26  175124.56   9.743  < 2e-16 ***
## view1         207317.33   67959.35   3.051 0.002384 **
## view2         155772.21   45724.06   3.407 0.000701 ***
## view3         193893.50   78032.55   2.485 0.013233 *
## view4         830740.34   93204.98   8.913  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 230400 on 603 degrees of freedom
## Multiple R-squared:  0.6866, Adjusted R-squared:  0.6819
## F-statistic: 146.8 on 9 and 603 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = price ~ sqft_living + bathrooms + age_yrs + high_grade +
##     view, data = df)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1079802  -120833   -16792    93170  1322772
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -139866.2    44943.1  -3.112 0.001946 **
## sqft_living      188.9       17.5  10.796  < 2e-16 ***
## bathrooms      73051.9    20754.4   3.520 0.000464 ***
## age_yrs         2124.8      384.1   5.532 4.72e-08 ***
## high_grade10  173829.0    45635.3   3.809 0.000154 ***
## high_grade11  514510.5    77620.3   6.629 7.54e-11 ***
## high_grade12 1673662.1   173740.9   9.633  < 2e-16 ***
## view1         199469.7    67363.4   2.961 0.003186 **
## view2         138700.3    45557.2   3.045 0.002432 **
## view3         200470.8    77328.5   2.592 0.009761 **
## view4         838835.9    92365.7   9.082  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 228300 on 602 degrees of freedom
## Multiple R-squared:  0.6929, Adjusted R-squared:  0.6878
## F-statistic: 135.8 on 10 and 602 DF,  p-value: < 2.2e-16


##
## Call:
## lm(formula = price ~ sqft_living + bedrooms + bathrooms + age_yrs +
##     high_grade + view, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1077345  -118302   -13162    93861  1310807
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -93666.2    49301.4  -1.900 0.057929 .
## sqft_living      210.4       19.9  10.575  < 2e-16 ***
## bedrooms      -32079.8    14301.0  -2.243 0.025248 *
## bathrooms      79468.3    20882.0   3.806 0.000156 ***
## age_yrs         2275.8      388.7   5.855 7.84e-09 ***
## high_grade10  155619.7    46201.9   3.368 0.000805 ***
## high_grade11  474688.3    79372.4   5.981 3.82e-09 ***
## high_grade12 1658297.6   173297.4   9.569  < 2e-16 ***
## view1         193374.8    67193.9   2.878 0.004146 **
## view2         122021.3    46010.2   2.652 0.008212 **
## view3         190287.4    77204.4   2.465 0.013991 *
## view4         811741.2    92846.9   8.743  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 227500 on 601 degrees of freedom
## Multiple R-squared:  0.6955, Adjusted R-squared:  0.6899
## F-statistic: 124.8 on 11 and 601 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = price ~ sqft_living + bedrooms + bathrooms + multistory +
##     age_yrs + high_grade + view, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1074210  -118395   -13127    98154  1308285
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -97341.87   49136.43  -1.981 0.048041 *
## sqft_living      203.86      20.01  10.186  < 2e-16 ***
## bedrooms      -31106.82   14251.92  -2.183 0.029449 *
## bathrooms      70452.60   21145.21   3.332 0.000916 ***
## multistory1    52545.16   22119.12   2.376 0.017836 *
## age_yrs         2425.39     392.27   6.183 1.16e-09 ***
## high_grade10  155281.76   46024.73   3.374 0.000789 ***
## high_grade11  481559.53   79120.43   6.086 2.06e-09 ***
## high_grade12 1669054.86  172691.20   9.665  < 2e-16 ***
## view1         187358.10   66983.74   2.797 0.005322 **
## view2         124296.51   45843.46   2.711 0.006893 **
## view3         212474.73   77472.94   2.743 0.006278 **
## view4         834597.76   92989.46   8.975  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 226600 on 600 degrees of freedom
## Multiple R-squared:  0.6983, Adjusted R-squared:  0.6923
## F-statistic: 115.7 on 12 and 600 DF,  p-value: < 2.2e-16


##
## Call:
## lm(formula = price ~ sqft_living + bedrooms + bathrooms + floors +
##     age_yrs + high_grade + view, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1073206  -115889   -10172    97669  1309318
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -170070.39   55900.31  -3.042 0.002450 **
## sqft_living      205.55      19.86  10.352  < 2e-16 ***
## bedrooms      -29060.35   14257.10  -2.038 0.041957 *
## bathrooms      66004.27   21293.65   3.100 0.002028 **
## floors         61101.20   21495.89   2.842 0.004629 **
## age_yrs         2584.93     401.42   6.440 2.46e-10 ***
## high_grade10  154575.55   45933.68   3.365 0.000814 ***
## high_grade11  482080.83   78951.83   6.106 1.84e-09 ***
## high_grade12 1667750.99  172317.74   9.678  < 2e-16 ***
## view1         190335.55   66810.17   2.849 0.004538 **
## view2         123355.82   45743.95   2.697 0.007201 **
## view3         212607.43   77154.29   2.756 0.006036 **
```

```
## view4         834292.10   92645.20   9.005  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 226200 on 600 degrees of freedom
## Multiple R-squared:  0.6995, Adjusted R-squared:  0.6935
## F-statistic: 116.4 on 12 and 600 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = price ~ sqft_living + sqft_lot + bedrooms + bathrooms +
##     floors + age_yrs + high_grade + view, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1041741  -115156   -10373    96822  1326038
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.569e+05  5.569e+04  -2.818  0.00499 **
## sqft_living   2.164e+02  2.004e+01  10.796  < 2e-16 ***
## sqft_lot     -6.392e-01  2.108e-01  -3.033  0.00253 **
## bedrooms     -3.349e+04  1.424e+04  -2.353  0.01896 *
## bathrooms     6.608e+04  2.115e+04   3.125  0.00187 **
## floors        5.461e+04  2.146e+04   2.545  0.01118 *
## age_yrs       2.543e+03  3.989e+02   6.375 3.66e-10 ***
## high_grade10  1.482e+05  4.567e+04   3.244  0.00124 **
## high_grade11  5.168e+05  7.925e+04   6.521 1.48e-10 ***
## high_grade12  1.655e+06  1.712e+05   9.667  < 2e-16 ***
## view1         1.828e+05  6.641e+04   2.752  0.00609 **
## view2         1.427e+05  4.588e+04   3.111  0.00196 **
## view3         2.105e+05  7.664e+04   2.747  0.00620 **
## view4         8.197e+05  9.214e+04   8.895  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 224700 on 599 degrees of freedom
## Multiple R-squared:  0.704,  Adjusted R-squared:  0.6976
## F-statistic: 109.6 on 13 and 599 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = price ~ sqft_living + sqft_lot + bedrooms + high_bath +
##     floors + age_yrs + high_grade + view, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -983585 -113099  -12662   99002 1386577
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -9.032e+04  5.246e+04  -1.722 0.085648 .
## sqft_living   2.370e+02  1.794e+01  13.211  < 2e-16 ***
## sqft_lot     -8.073e-01  2.127e-01  -3.796 0.000162 ***
```

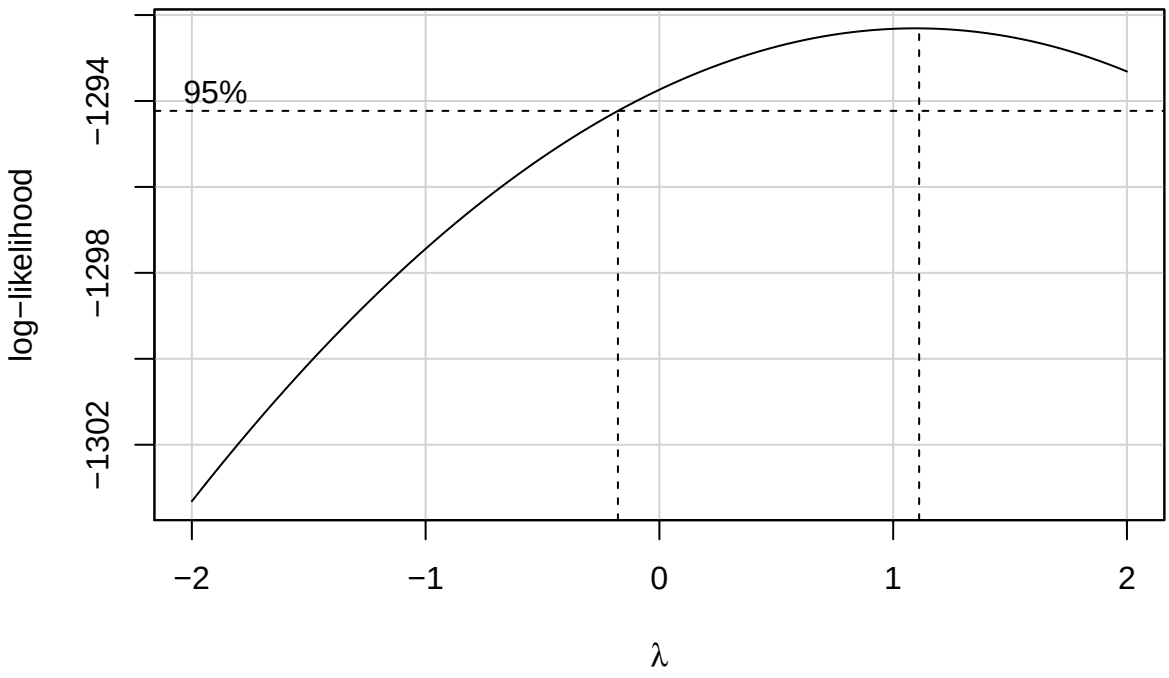
```
## bedrooms     -2.803e+04  1.397e+04  -2.006 0.045259 *
## high_bath1    3.625e+05  8.290e+04   4.373 1.44e-05 ***
## floors        7.398e+04  2.079e+04   3.558 0.000403 ***
## age_yrs       2.141e+03  3.677e+02   5.822 9.49e-09 ***
## high_grade10  1.120e+05  4.632e+04   2.417 0.015934 *
## high_grade11  4.450e+05  8.113e+04   5.485 6.10e-08 ***
## high_grade12  1.561e+06  1.720e+05   9.076  < 2e-16 ***
## view1         1.639e+05  6.613e+04   2.479 0.013468 *
## view2         1.461e+05  4.530e+04   3.224 0.001331 **
## view3         2.224e+05  7.609e+04   2.923 0.003599 **
## view4         7.738e+05  9.214e+04   8.398 3.30e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 222900 on 599 degrees of freedom
## Multiple R-squared:  0.7085, Adjusted R-squared:  0.7022
## F-statistic:   112 on 13 and 599 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = price ~ age_yrs + sqft_living + sqft_lot + bedrooms +
##     grade + high_bath + floors + view, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1113908  -108971    -4655    93838  1932185
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.169e+05  1.005e+05  -7.136 2.78e-12 ***
## age_yrs      3.130e+03  3.966e+02   7.893 1.41e-14 ***
## sqft_living  2.024e+02  2.077e+01   9.747  < 2e-16 ***
## sqft_lot    -6.483e-01  2.195e-01  -2.953  0.00327 **
## bedrooms    -3.130e+04  1.422e+04  -2.201  0.02810 *
## grade        9.601e+04  1.422e+04   6.752 3.45e-11 ***
## high_bath1   5.657e+05  8.179e+04   6.916 1.19e-11 ***
## floors       3.925e+04  2.222e+04   1.767  0.07776 .
## view1        1.245e+05  6.875e+04   1.811  0.07070 .
## view2        1.005e+05  4.663e+04   2.155  0.03158 *
## view3        1.515e+05  7.912e+04   1.914  0.05605 .
## view4        9.203e+05  9.358e+04   9.835  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 231500 on 601 degrees of freedom
## Multiple R-squared:  0.6847, Adjusted R-squared:  0.6789
## F-statistic: 118.6 on 11 and 601 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = price ~ age_yrs + sqft_living + sqft_lot + bedrooms +
##     high_grade + high_bath + floors + view.factor, data = df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -983585 -113099  -12662   99002 1386577
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -9.032e+04  5.246e+04  -1.722 0.085648 .
## age_yrs       2.141e+03  3.677e+02   5.822 9.49e-09 ***
## sqft_living   2.370e+02  1.794e+01  13.211  < 2e-16 ***
## sqft_lot     -8.073e-01  2.127e-01  -3.796 0.000162 ***
## bedrooms     -2.803e+04  1.397e+04  -2.006 0.045259 *
## high_grade10  1.120e+05  4.632e+04   2.417 0.015934 *
## high_grade11  4.450e+05  8.113e+04   5.485 6.10e-08 ***
## high_grade12  1.561e+06  1.720e+05   9.076  < 2e-16 ***
## high_bath1    3.625e+05  8.290e+04   4.373 1.44e-05 ***
## floors        7.398e+04  2.079e+04   3.558 0.000403 ***
## view.factor1  1.639e+05  6.613e+04   2.479 0.013468 *
## view.factor2  1.461e+05  4.530e+04   3.224 0.001331 **
## view.factor3  2.224e+05  7.609e+04   2.923 0.003599 **
## view.factor4  7.738e+05  9.214e+04   8.398 3.30e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 222900 on 599 degrees of freedom
## Multiple R-squared:  0.7085, Adjusted R-squared:  0.7022
## F-statistic:   112 on 13 and 599 DF,  p-value: < 2.2e-16
```

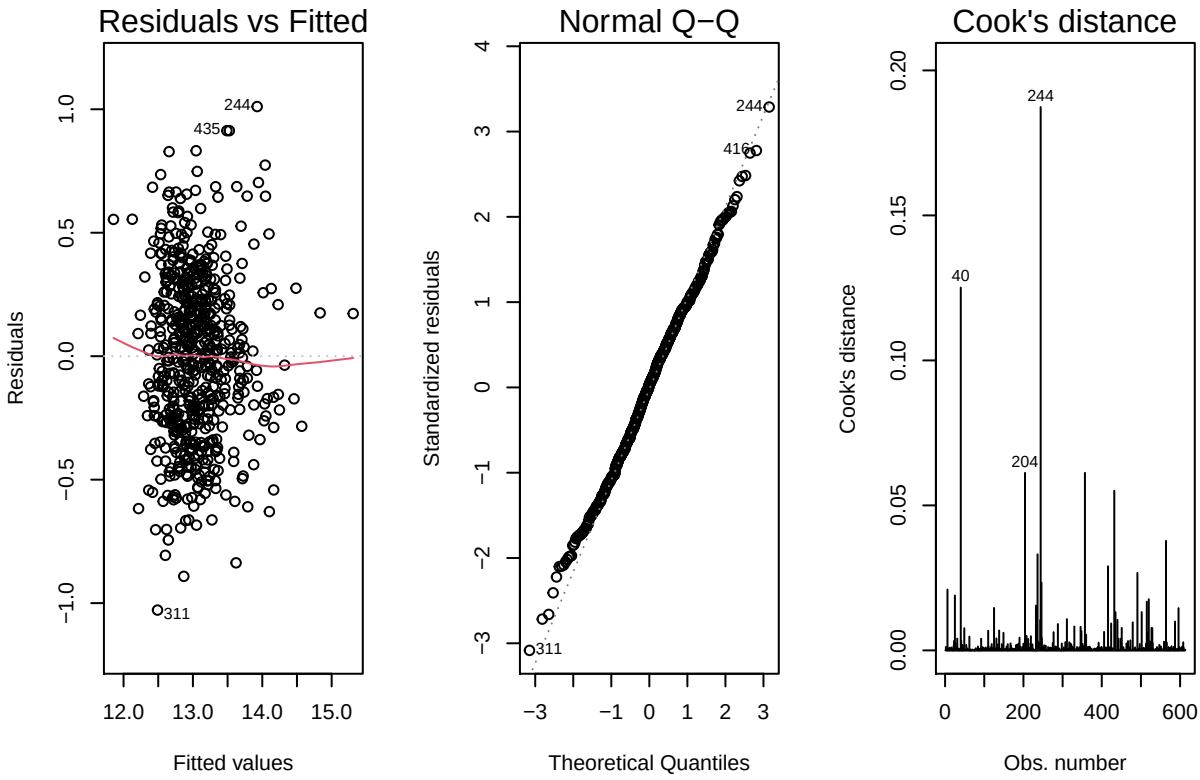
```
##
## Call:
## lm(formula = log(price) ~ age_yrs + log(sqft_living) + sqft_lot +
##     bedrooms + high_grade + high_bath + floors + view.factor,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0654 -0.2480  0.0078  0.2322  1.0217
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       7.220e+00  3.525e-01  20.485  < 2e-16 ***
## age_yrs           3.679e-03  5.577e-04   6.597 9.21e-11 ***
## log(sqft_living)  7.180e-01  5.311e-02  13.518  < 2e-16 ***
## sqft_lot         -5.250e-07  3.173e-07  -1.655 0.098503 .
## bedrooms         -3.124e-02  2.188e-02  -1.428 0.153890
## high_grade10      2.802e-01  6.635e-02   4.223 2.79e-05 ***
## high_grade11      4.841e-01  1.170e-01   4.137 4.03e-05 ***
## high_grade12      8.418e-01  2.541e-01   3.313 0.000978 ***
## high_bath1        1.401e-01  1.243e-01   1.127 0.260028
## floors            2.092e-01  3.131e-02   6.681 5.42e-11 ***
## view.factor1      2.662e-01  9.945e-02   2.677 0.007634 **
## view.factor2      2.674e-01  6.730e-02   3.974 7.93e-05 ***
## view.factor3      3.984e-01  1.145e-01   3.479 0.000540 ***
## view.factor4      4.538e-01  1.371e-01   3.309 0.000992 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.335 on 599 degrees of freedom
## Multiple R-squared:  0.6165, Adjusted R-squared:  0.6082
## F-statistic: 74.09 on 13 and 599 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = log(price) ~ age_yrs + log(sqft_living) + high_grade +
##     floors + view.factor, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.02818 -0.24274  0.01217  0.23231  1.01101
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.5194436  0.2964223  25.367  < 2e-16 ***
## age_yrs          0.0035900  0.0005519   6.504 1.64e-10 ***
## log(sqft_living) 0.6625296  0.0396348  16.716  < 2e-16 ***
## high_grade10     0.3068780  0.0635911   4.826 1.77e-06 ***
## high_grade11     0.5065818  0.1079796   4.691 3.36e-06 ***
## high_grade12     0.8716502  0.2491442   3.499 0.000502 ***
## floors           0.2151399  0.0311641   6.903 1.29e-11 ***
## view.factor1     0.2868164  0.0990096   2.897 0.003906 **
## view.factor2     0.2670116  0.0656841   4.065 5.44e-05 ***
## view.factor3     0.4083527  0.1143149   3.572 0.000382 ***
## view.factor4     0.5002904  0.1348985   3.709 0.000228 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3356 on 602 degrees of freedom
## Multiple R-squared:  0.6134, Adjusted R-squared:  0.607
## F-statistic: 95.52 on 10 and 602 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = price ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -913051 -105734   -3368   92435 1333304
##
## Coefficients: (13 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.864e+10  2.303e+10   1.243 0.214210
## id               8.197e-07  3.122e-06   0.263 0.792965
## date             1.104e+02  8.014e+01   1.378 0.168693
## bedrooms         6.754e+03  2.652e+04   0.255 0.799051
## bathrooms        1.972e+04  2.477e+04   0.796 0.426188
## sqft_living      2.367e+02  4.235e+01   5.589 3.53e-08 ***
## sqft_lot        -5.580e-01  2.151e-01  -2.594 0.009739 **
## floors           5.930e+04  4.701e+04   1.261 0.207733
## waterfront1      2.059e+05  2.449e+05   0.841 0.400854
## view1            1.274e+05  6.515e+04   1.955 0.051065 .
```

```
## view2           1.287e+05  4.464e+04   2.882 0.004096 **
## view3           1.745e+05  7.618e+04   2.290 0.022383 *
## view4           6.121e+05  2.226e+05   2.749 0.006160 **
## condition2     -1.187e+05  2.033e+05  -0.584 0.559533
## condition3     -2.710e+04  1.580e+05  -0.172 0.863868
## condition4     -2.342e+04  1.580e+05  -0.148 0.882190
## condition5      4.810e+04  1.607e+05   0.299 0.764842
## grade           7.425e+04  2.088e+04   3.555 0.000409 ***
## sqft_above     -8.526e+01  4.688e+01  -1.819 0.069472 .
## sqft_basement          NA         NA      NA       NA
## yr_built       -1.454e+07  1.169e+07  -1.244 0.214174
## yr_renovated    1.257e+05  1.490e+05   0.844 0.399245
## date_built      3.980e+04  3.201e+04   1.243 0.214255
## age_yrs                NA         NA      NA       NA
## last_reno_yrs   1.200e+05  1.485e+05   0.808 0.419595
## renovated1     -2.532e+08  3.003e+08  -0.843 0.399401
## has_basement1  -7.983e+03  3.435e+04  -0.232 0.816337
## high_grade10    2.981e+04  5.556e+04   0.537 0.591782
## high_grade11    3.338e+05  9.323e+04   3.580 0.000373 ***
## high_grade12    1.286e+06  1.863e+05   6.901 1.37e-11 ***
## grade.factor6  -5.069e+04  5.650e+04  -0.897 0.370035
## grade.factor7  -4.217e+04  3.984e+04  -1.058 0.290297
## grade.factor8  -3.717e+04  3.104e+04  -1.198 0.231550
## grade.factor9          NA         NA      NA       NA
## grade.factor10         NA         NA      NA       NA
## grade.factor11         NA         NA      NA       NA
## grade.factor12         NA         NA      NA       NA
## view.factor1           NA         NA      NA       NA
## view.factor2           NA         NA      NA       NA
## view.factor3           NA         NA      NA       NA
## view.factor4           NA         NA      NA       NA
## multistory1     3.550e+03  4.929e+04   0.072 0.942609
## bath_group[3-4] 3.956e+02  4.602e+04   0.009 0.993144
## bath_group[4-6] 3.444e+05  9.721e+04   3.542 0.000429 ***
## bed_group2      9.726e+03  7.641e+04   0.127 0.898748
## bed_group3     -1.011e+05  7.190e+04  -1.407 0.160082
## bed_group4     -1.197e+05  7.844e+04  -1.526 0.127520
## bed_group5     -1.988e+05  9.752e+04  -2.039 0.041942 *
## bed_group6      3.259e+04  1.424e+05   0.229 0.819106
## bed_group7             NA         NA      NA       NA
## high_bath1             NA         NA      NA       NA
## high_bed1              NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 215700 on 574 degrees of freedom
## Multiple R-squared:  0.7385, Adjusted R-squared:  0.7212
## F-statistic: 42.66 on 38 and 574 DF,  p-value: < 2.2e-16
```

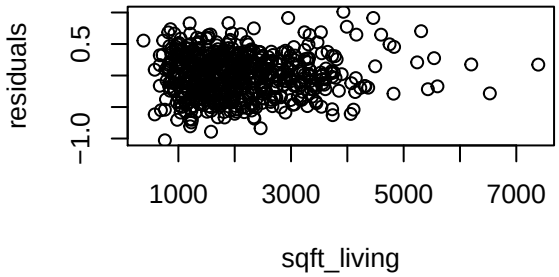


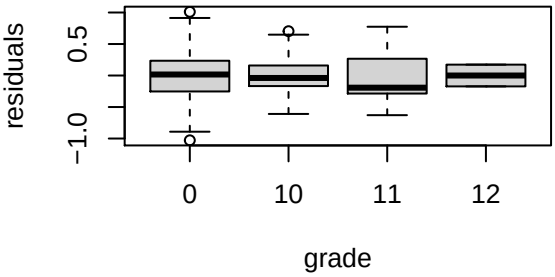

```
par(mfrow = c(1,3))
plot(lm.price,c(1,2,4))
```

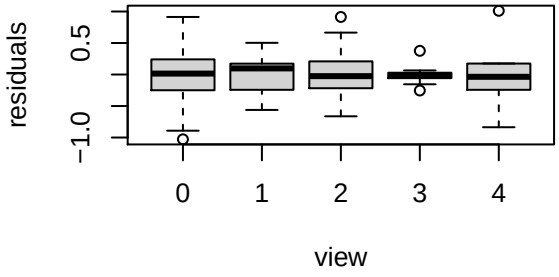



```
##                       GVIF Df GVIF^(1/(2*Df))
## age_yrs           1.484488  1        1.218396
## log(sqft_living) 1.709663  1        1.307541
## high_grade        1.442258  3        1.062936
## floors            1.511292  1        1.229346
## view.factor       1.231720  4        1.026394
```

```
par(mfrow = c(2,2))
plot(df$sqft_living ,lm.price$residuals, xlab = "sqft_living", ylab = "residuals")
plot(df$high_grade,lm.price$residuals, xlab = "grade", ylab = "residuals")
plot(df$view ,lm.price$residuals, xlab = "view", ylab = "residuals")
```







Results and Conclusions

Appendix: All Code for This Report

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE)
library(dplyr)
library(car)
library(ggplot2)
library(grid)
library(gridExtra)
library(ggthemes)
library(lubridate)
library(GGally)
library(scales)
library(stargazer) # Used for latex tables to summarize the data and models

# Read the Data
df <- read.csv('Seattle.csv', strip.white = TRUE, stringsAsFactors = FALSE)
# Clean the Data
df$date <- ymd(substr(df$date,1,nchar(df$date) - 7)) # Convert string to date object
df$date_built <- as.Date(ISOdate(df$yr_built,1,1))
df$age_yrs <- as.double(df$date - df$date_built)/365.
df$last_reno_yrs <- with(df,
     ifelse(yr_renovated == 0,
            0,
            as.double(date-as.Date(ISOdate(df$yr_renovated,1,1)))/365
          )
 )
df$renovated <- as.factor(with(df,
     ifelse(yr_renovated > 0,
            1,
            0
          )
 ))
df$has_basement <- as.factor(with(df,
     ifelse(sqft_basement > 0,
            1,
            0
          )
 ))

# Quantitative Values Section
quant.columns <- c(4:7,13:14, 18:19, 3)

# Print head of initial dataframe
stargazer(df[1:4,quant.columns[1:5]],
          rownames=FALSE,
          summary=FALSE,
          header=FALSE,
          title="First Four Rows for Quantitative Values on Seattlle Housing Dataframe")

# Print head of initial dataframe
stargazer(df[1:4,quant.columns[6:length(quant.columns)]],
          rownames=FALSE,
          summary=FALSE,
```

```
          header=FALSE,
          title="First Four Rows for Quantitative Values on Seattlle Housing Dataframe")

# Summarize initial dataframe
stargazer(df[,-c(1)],
          header=FALSE,
          omit.summary.stat=c('N'),
          title="Summary Statistics for Values on Seattle Housing Dataframe")

# Quantitative Data Pairs
ggpairs(df[,quant.columns], progress=FALSE) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Discrete and Categorical Values Section
df$condition <- as.factor(df$condition)
df$high_grade <- (as_tibble(
  select(df, grade, price)) %>%
  mutate(tag=case_when(
    grade < 10 ~ 0,
    grade == 10 ~ 10,
    grade == 11 ~ 11,
    grade == 12 ~ 12
    )))$tag
df$grade.factor <- as.factor(df$grade)
df$high_grade <- as.factor(df$high_grade)
df$waterfront <- as.factor(df$waterfront)
df$view <- as.factor(df$view)
df$view.factor <- as.factor(df$view)
#df$bedrooms <- as.factor(df$bedrooms)
df$multistory <- as.factor(with(df, ifelse(floors==1,0,1)))
#df$floors <- as.factor(df$floors)
cat.discrete.price.columns <- c(3,9,10,11)
summary(df)

tags <- c("[0-3]", "[3-4]", "[4-6]")
bgroup <- as_tibble(select(df, price, bathrooms)) %>%
  mutate(tag = case_when(
  bathrooms == 0.25|bathrooms == 0.5|
    bathrooms == 0.75|bathrooms == 1.00 |
  bathrooms == 1.25|bathrooms == 1.5|
    bathrooms == 1.75|bathrooms == 2.00|
  bathrooms == 2.25|bathrooms == 2.5|
    bathrooms == 2.75|bathrooms == 3.00 ~ tags[1],
  bathrooms == 3.25|bathrooms == 3.5|
    bathrooms == 3.75|bathrooms == 4.00 ~ tags[2],
  bathrooms > 4.00 & bathrooms <= 6.00 ~ tags[3],
))
df$bath_group <- as.factor(bgroup$tag)
df$bed_group <- as.factor(df$bedrooms)
df$high_bath <- as.factor(with(df, ifelse(bathrooms>4,1,0)))
df$high_bed <- as.factor(with(df, ifelse(bedrooms>4,1,0)))

plot_price_by_cat <- function(df, cat_var, cat_var_name) {
```

```
  ggplot(df, aes(x=cat_var, y=price, fill=cat_var)) +
    scale_colour_solarized("red") +
    geom_violin(aes(color=cat_var)) +
    geom_boxplot(width=0.1) +
    xlab(cat_var_name) +
    ylab("Log(Price)") +
    scale_y_continuous(labels = scales::dollar_format(scale = 1)) +
    theme(legend.position="none")
}

p <- plot_price_by_cat(df, df$condition, "Condition")
q <- plot_price_by_cat(df, df$view, "View")
r <- plot_price_by_cat(df, df$grade.factor, "Grade")
s <- plot_price_by_cat(df, df$high_grade, "High Grade")
t <- plot_price_by_cat(df, df$waterfront, "Waterfront")
u <- plot_price_by_cat(df, df$bed_group, "Bedrooms")
v <- plot_price_by_cat(df, df$bath_group, "Bathrooms")
w <- plot_price_by_cat(df, as.factor(df$floors), "Floors")

grid.arrange(grobs=list(p, q,
                        r, s),
             ncol=2,
             top="Price by Categorical Variable")

grid.arrange(grobs=list(t, u,
                        v, w),
             ncol=2,
             top=NULL)

# Initial Model
lm.initial <- lm(price ~ sqft_living, data=df)
summary(lm.initial,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')

lm.initial <- lm(price ~ sqft_living + sqft_above, data=df)
summary(lm.initial,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')

lm.initial <- lm(price ~ sqft_living + sqft_above + sqft_basement, data=df)
summary(lm.initial,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')

lm.initial <- lm(price ~ sqft_living + sqft_above + has_basement, data=df)
summary(lm.initial,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')
```

```
lm.initial <- lm(price ~ sqft_living + sqft_above + age_yrs, data=df)
summary(lm.initial,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')

lm.initial <- lm(price ~ sqft_living + sqft_above + date_built, data=df)
summary(lm.initial,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')

lm.price <- lm(price ~ sqft_living + sqft_above + age_yrs + grade, data=df)
summary(lm.price,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')

lm.price <- lm(price ~ sqft_living + sqft_above +  age_yrs + high_grade, data=df)
summary(lm.price,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')

lm.price <- lm(price ~ sqft_living + sqft_above + age_yrs + high_grade + view, data=df)
summary(lm.price,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')

lm.price <- lm(price ~ sqft_living + age_yrs + high_grade + view, data=df)
summary(lm.price,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')

lm.price <- lm(price ~ sqft_living + bathrooms + age_yrs + high_grade + view, data=df)
summary(lm.price,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')

lm.price <- lm(price ~ sqft_living + bedrooms + bathrooms + age_yrs + high_grade + view, data=df)
summary(lm.price,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')

lm.price <- lm(price ~ sqft_living + bedrooms + bathrooms + multistory + age_yrs + high_grade + view, da
summary(lm.price,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')
```

```r
lm.price <- lm(price ~ sqft_living + bedrooms + bathrooms + floors + age_yrs + high_grade + view, data=
summary(lm.price,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')

lm.price <- lm(price ~ sqft_living + sqft_lot + bedrooms + bathrooms + floors + age_yrs + high_grade +
summary(lm.price,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')

lm.price <- lm(price ~ sqft_living + sqft_lot + bedrooms + high_bath + floors + age_yrs + high_grade +
summary(lm.price,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')

lm.price <- lm(price ~ age_yrs + sqft_living + sqft_lot + bedrooms + grade + high_bath + floors + view,
summary(lm.price,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')

lm.price <- lm(price ~ age_yrs + sqft_living + sqft_lot + bedrooms + high_grade + high_bath + floors +
summary(lm.price,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')

lm.price <- lm(log(price) ~ age_yrs + log(sqft_living) + sqft_lot + bedrooms + high_grade + high_bath +
summary(lm.price,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')

lm.price <- lm(log(price) ~ age_yrs + log(sqft_living) + high_grade + floors + view.factor, data=df)
summary(lm.price,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')

lm.full <- lm(price ~ ., data=df)
summary(lm.full,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')
boxCox(lm.price)
par(mfrow = c(1,3))
plot(lm.price,c(1,2,4))
vif(lm.price)
par(mfrow = c(2,2))
plot(df$sqft_living ,lm.price$residuals, xlab = "sqft_living", ylab = "residuals")
```

```r
plot(df$high_grade,lm.price$residuals, xlab = "grade", ylab = "residuals")
plot(df$view ,lm.price$residuals, xlab = "view", ylab = "residuals")
```