Housing Price in King County, Washington

Present By
Haylee,Hanyue,Zhijian,Zhongyan

Agenda

- Background
- Economic significance
- Dataset description
- Data cleaning
- Data analysis
- Conclusion

Background

What is happening in the Real estate Market?

- Housing price has dropped by $70,000 in three months as market continues to cool
  - Higher Mortgage rate
  - Stable Rent Rate
  - Local population and job growth has also slowed
  - buyers from China, who have a strong presence in the Seattle market, have had trouble getting their money out of the country

Economic Significance

What we are interested to see?

- The median house last month sold for $760,000, a drop of $45,000 in just one month and $70,000 in three months
- What elements in housing drastically affect prices of sales
- What other variables will attract people to buy houses besides common known variables
- How regression model can be applied in Real estate industry

Data Description

Variables

variable.name	description
id	a notation for a house
date	Date house was sold
price	Price is prediction target
bedrooms	Number of Bedrooms/House
bathrooms	Number of bathrooms/bedrooms
sqft_living	square footage of the home
sqft_lot	square footage of the lot
floors	Total floors (levels) in house
waterfront	House which has a view to a waterfront
view	Has been viewed
condition	How good the condition is (Overall)
grade	overall grade given to the housing unit, based on King County grading system
sqft_above	square footage of house apart from basement
sqft_basement	square footage of the basement
yr_built	Built Year
yr_renovated	Year when house was renovated
lat	Latitude coordinate
long	Longitude coordinate

- dataset contains 19 house features plus the price and the id columns, along with 21613 observations including some expected variables (number of bedrooms,square feet,condition)
- Non-expected variables such as grade, latitude and longtitude
- dataset is tidy and contains float and integer variables

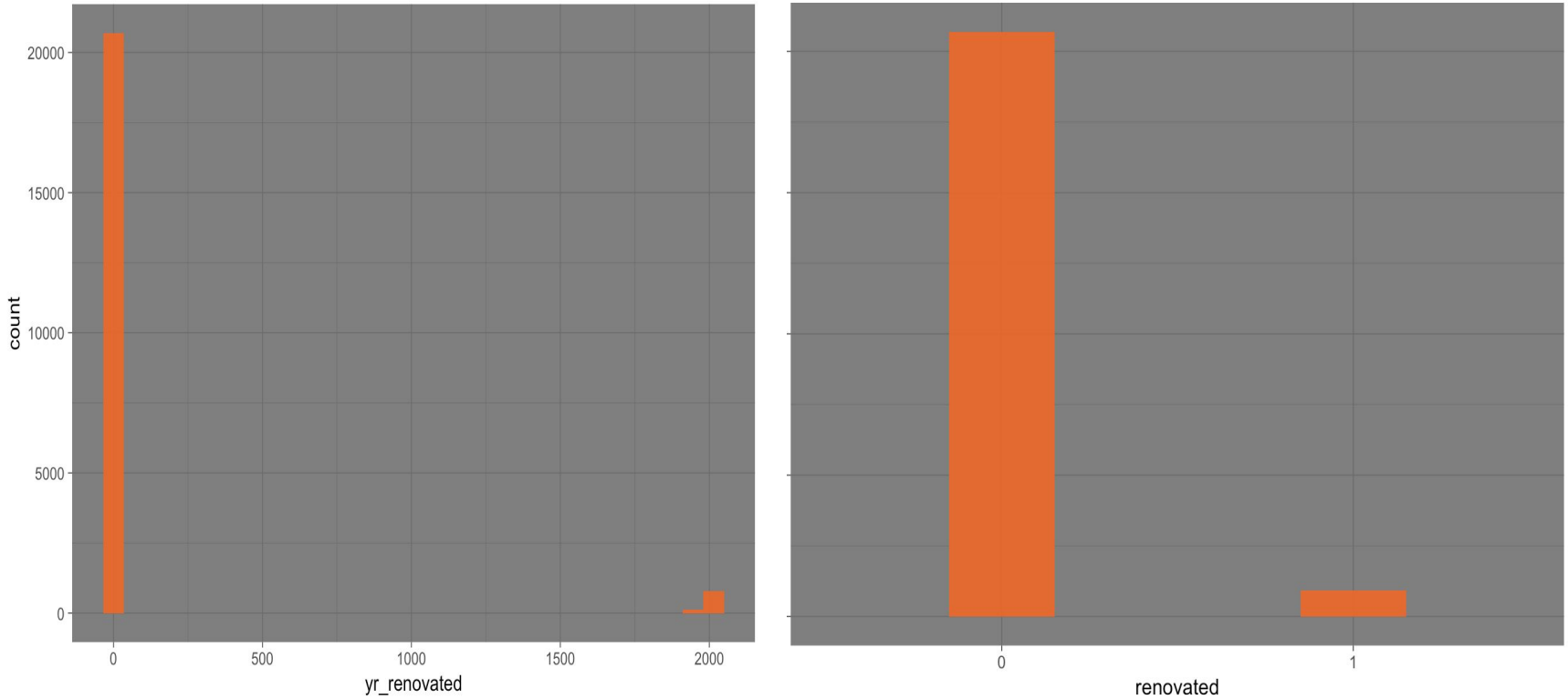
Data Manipulation

Age of the house

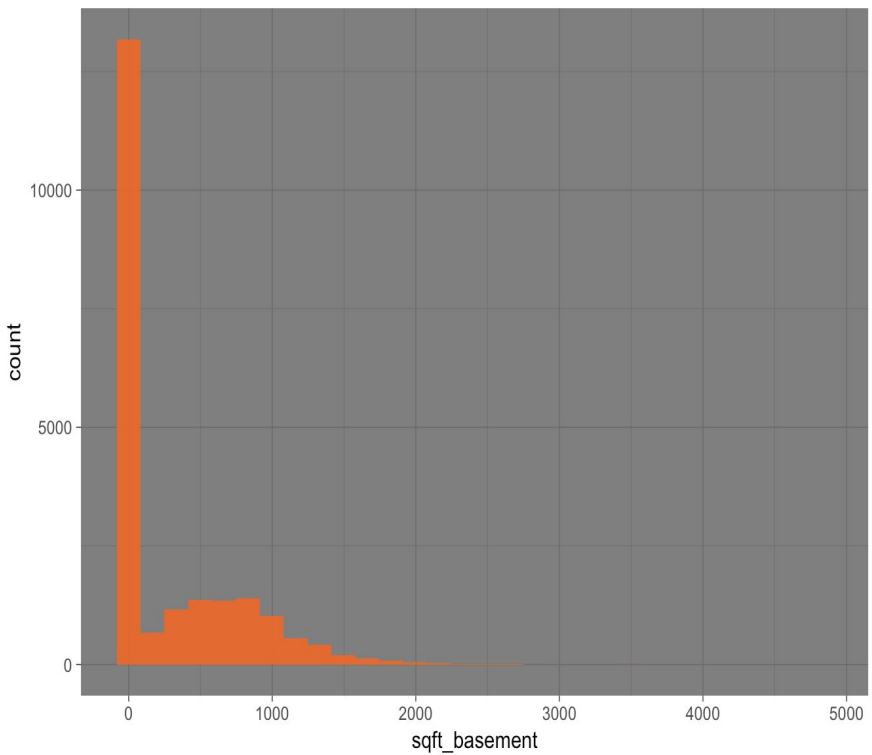
```
df <- df %>%
    mutate(yr_sold = year(date)) %>%
    mutate(house_age = yr_sold - yr_built)
```

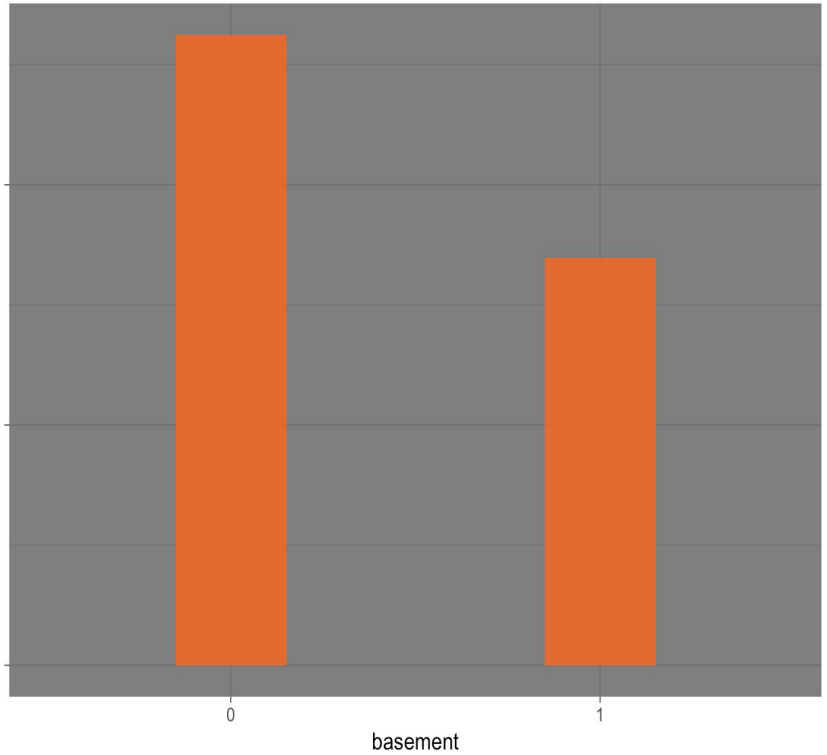
Age of the house = year it was sold - year it was built

Renovation(year of renovation -> renovated)

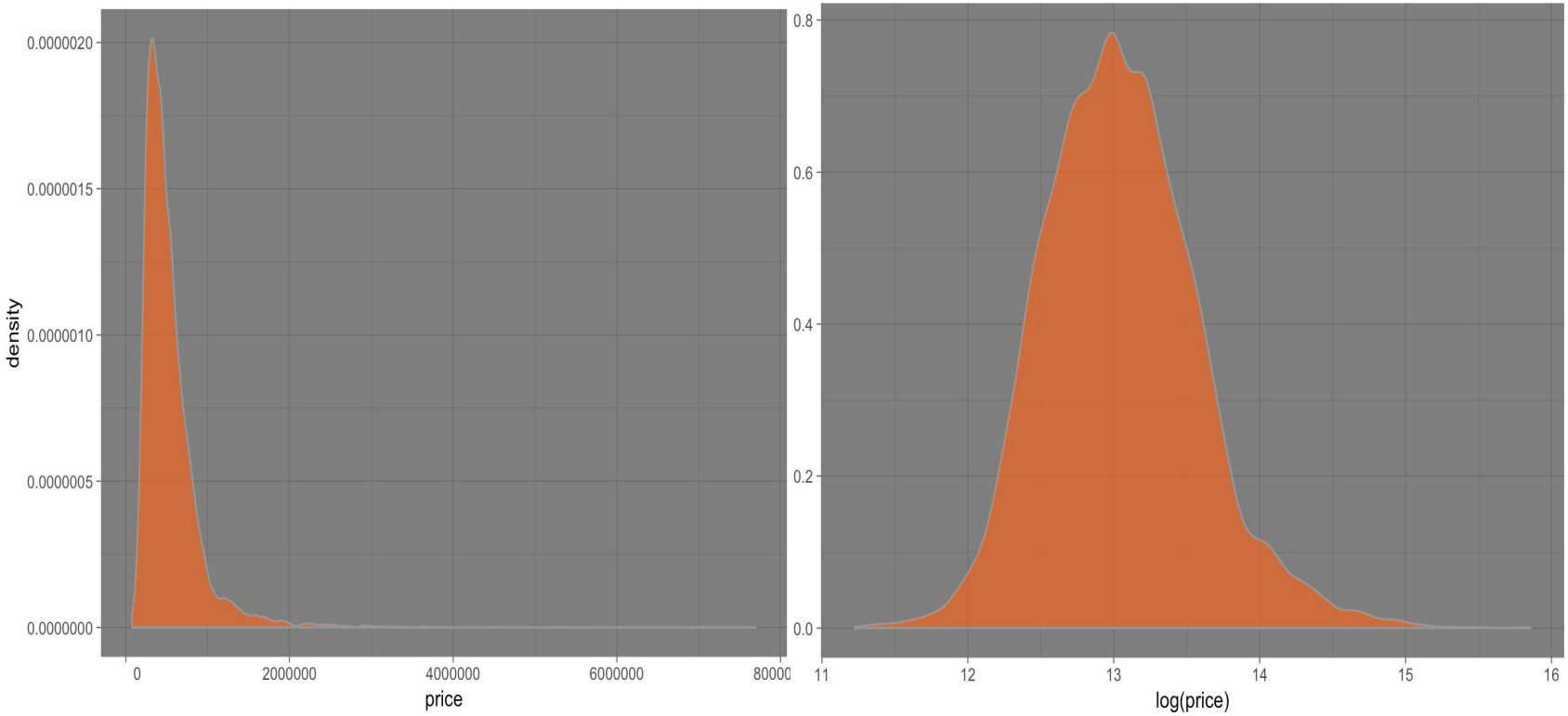


Basement (basement area -> basement)



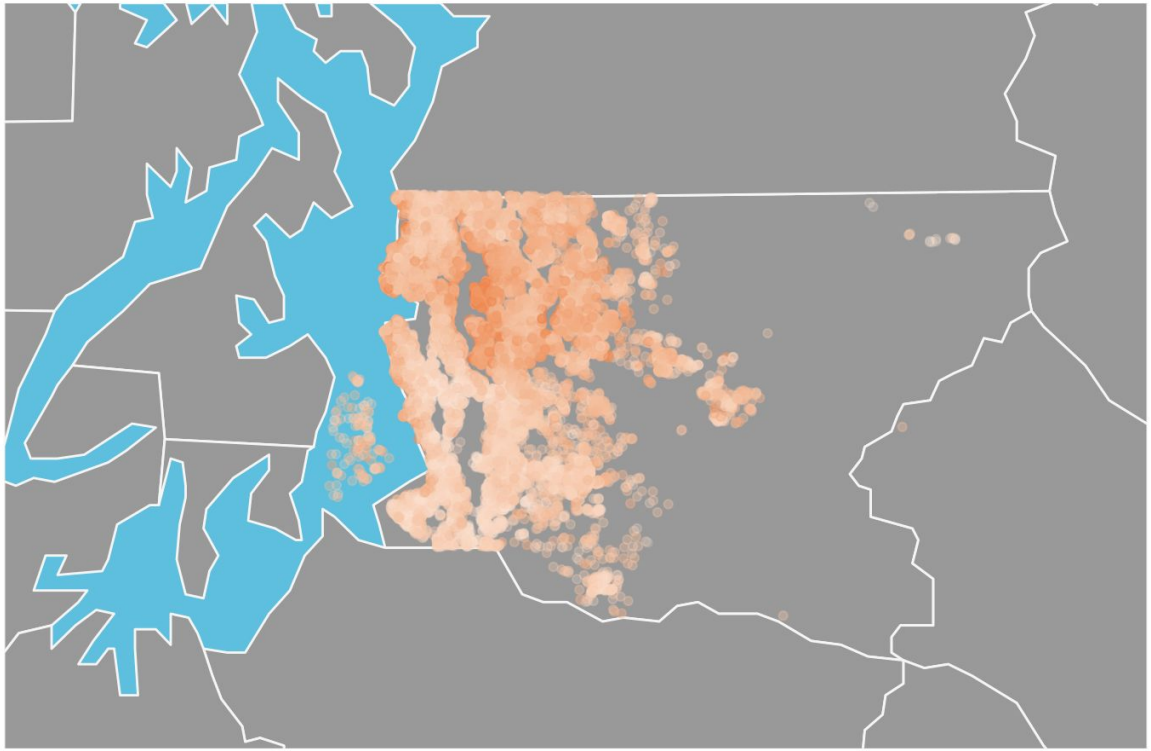


Price vs. log(Price)

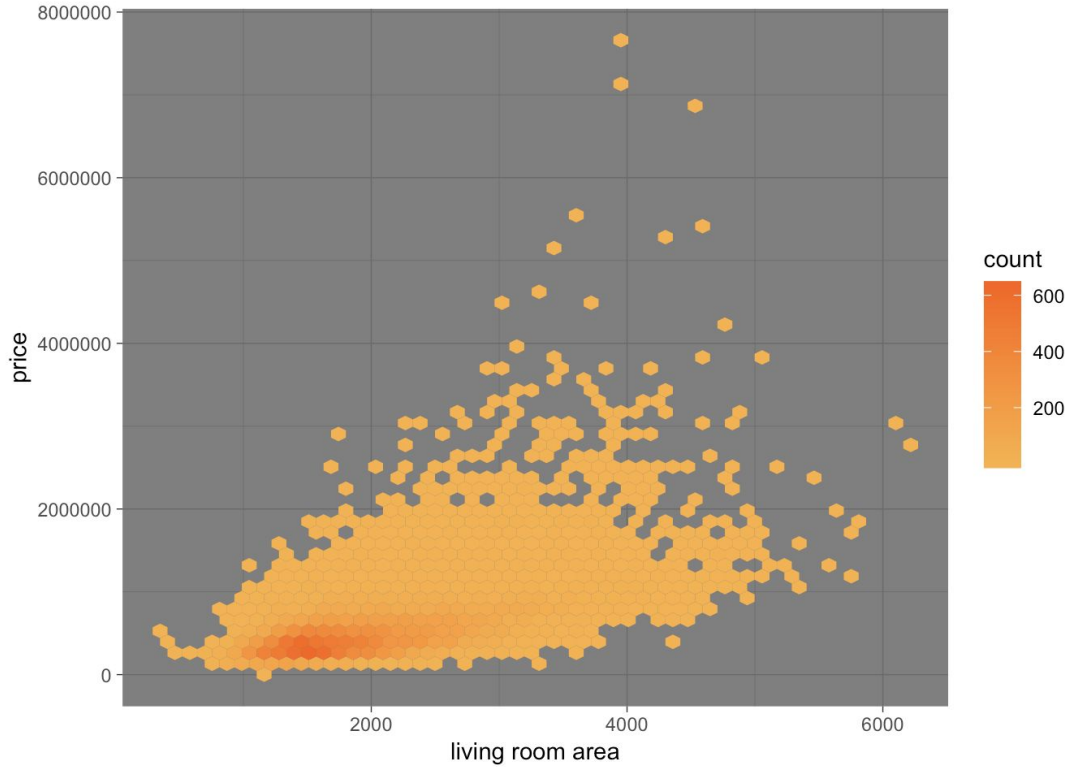


Visualization

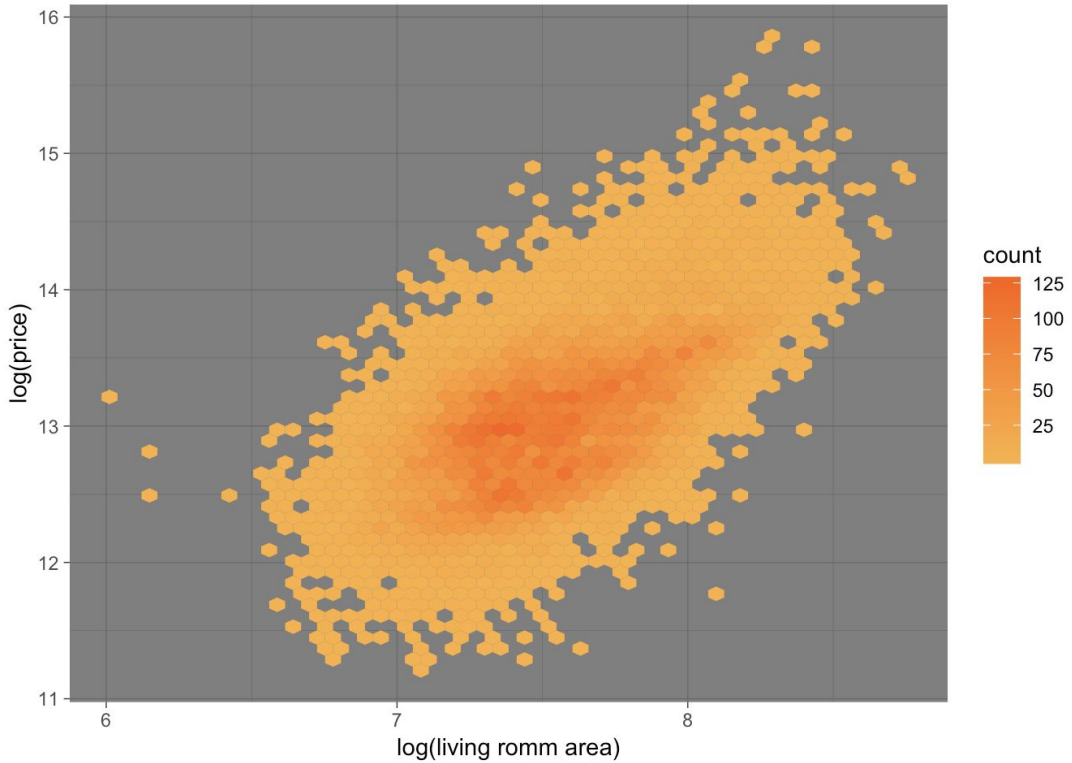
Map



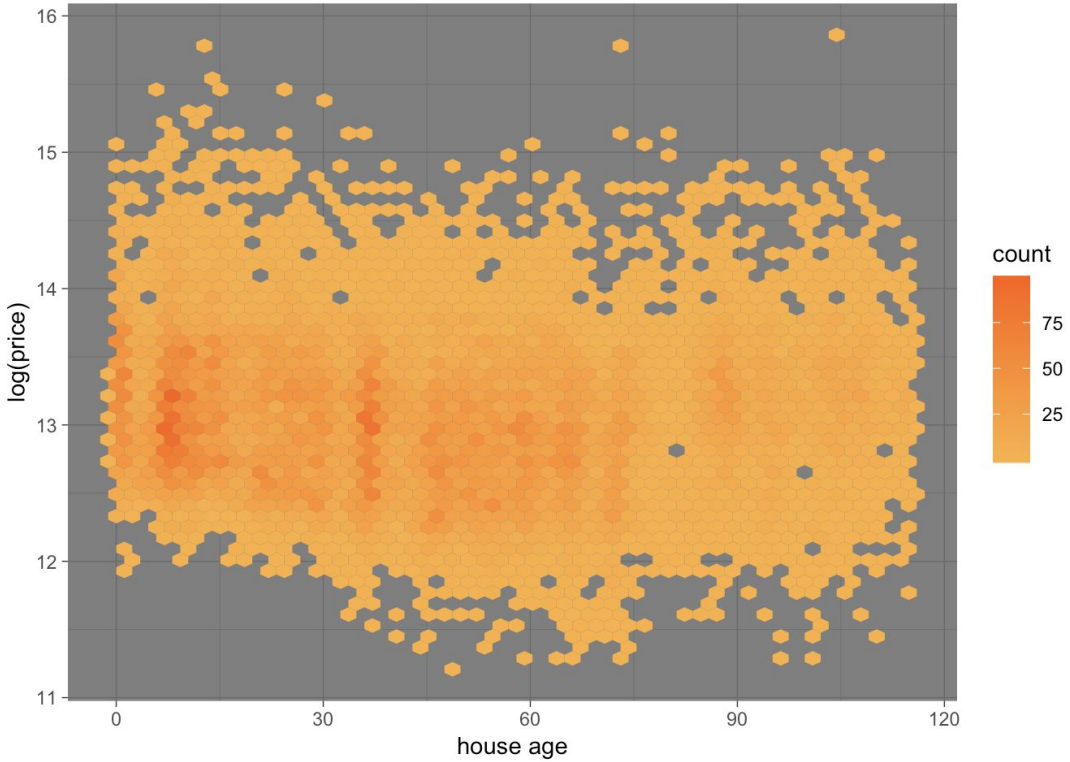
Price vs. Living room area



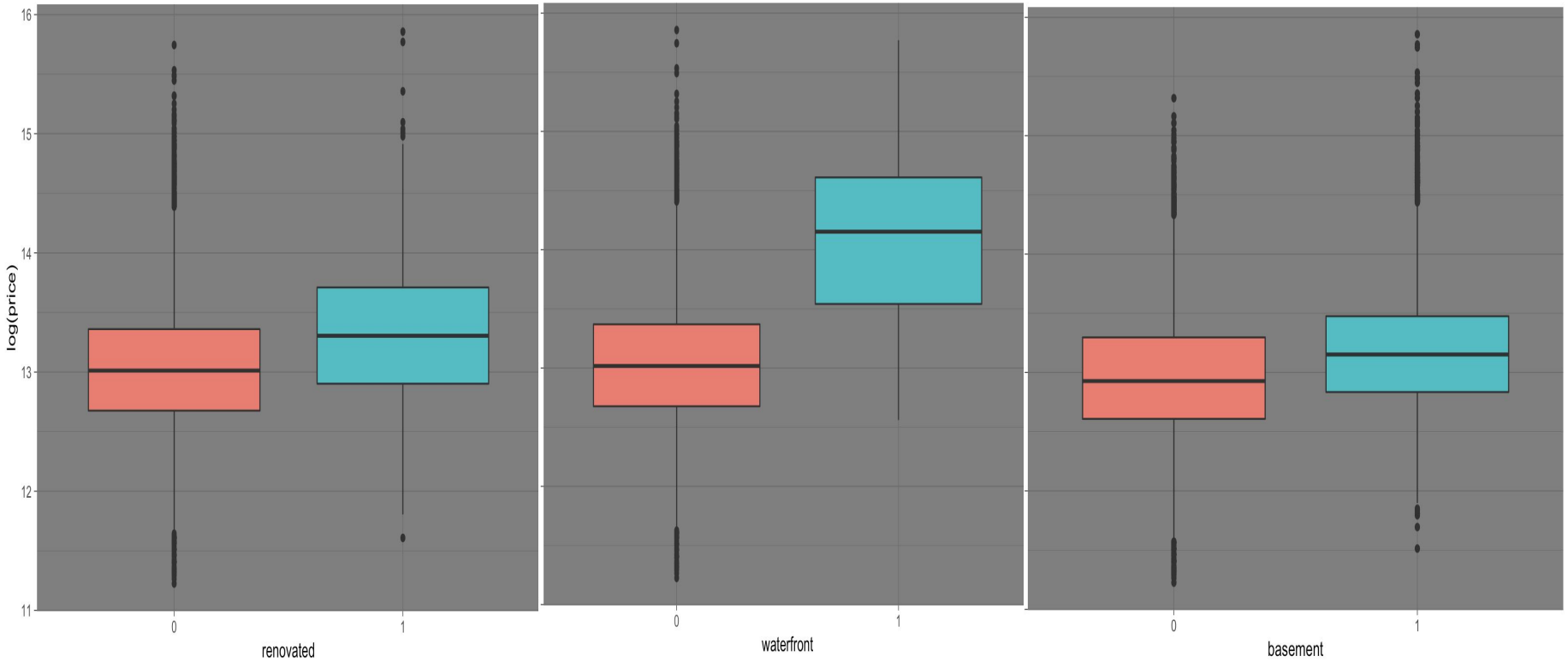
log(Price) vs. Living room area



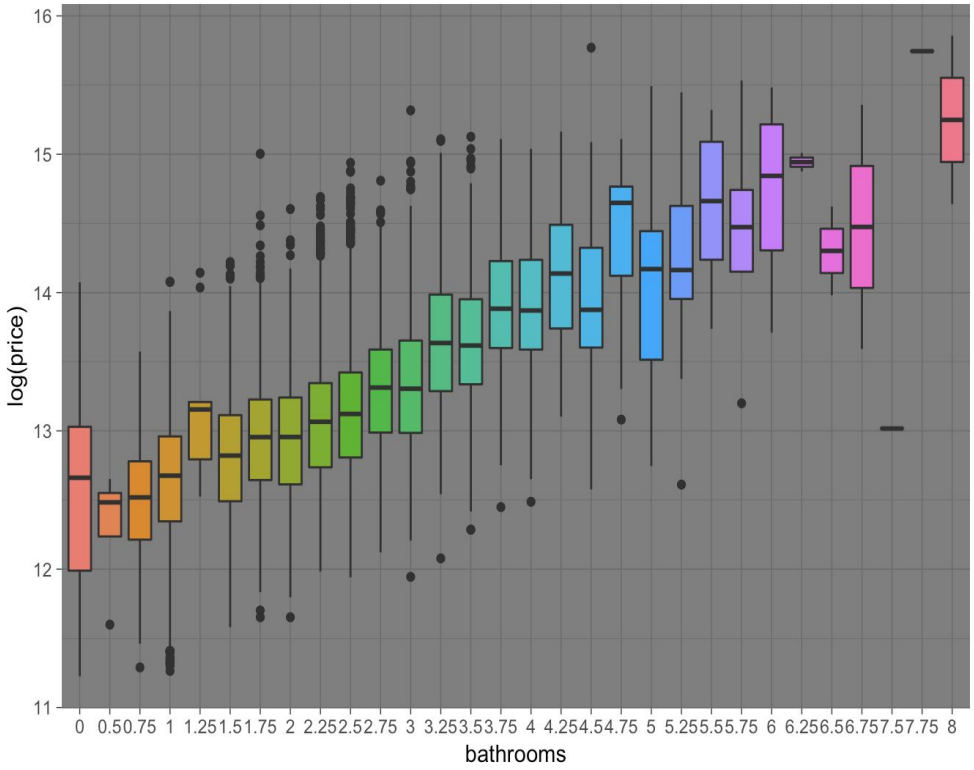
log(Price) vs. House age

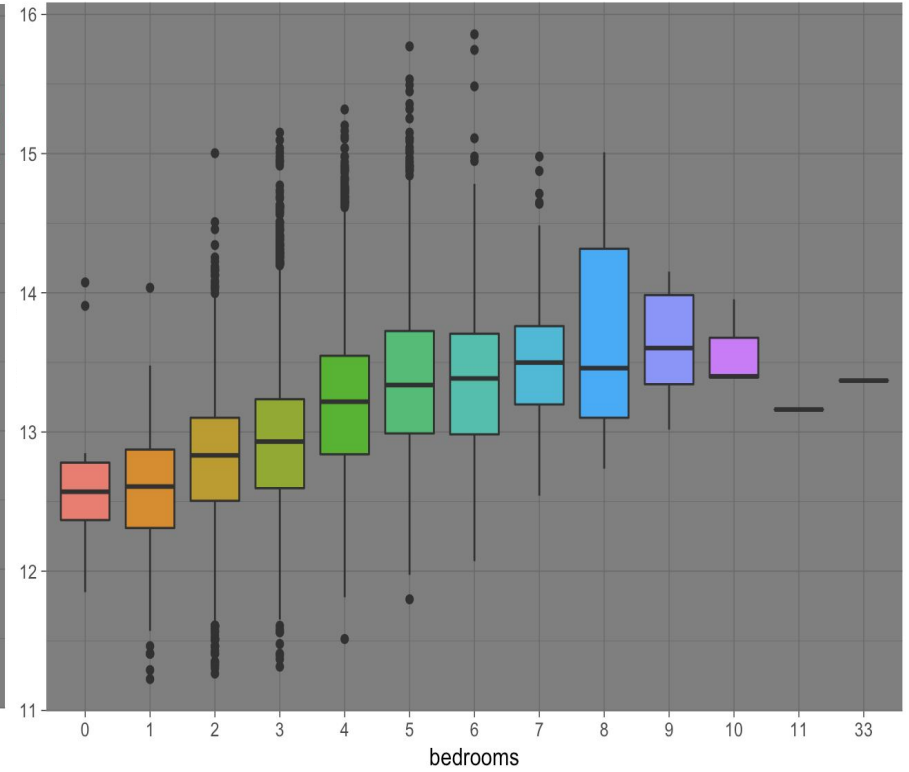


log(Price) vs. Renovation

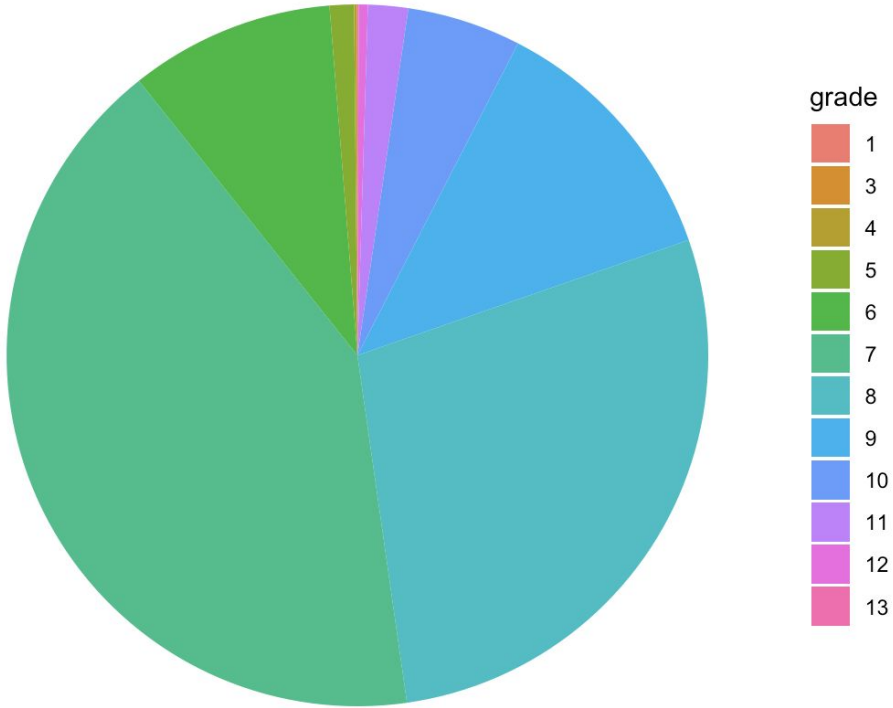


log(Price) vs. Number of bathrooms

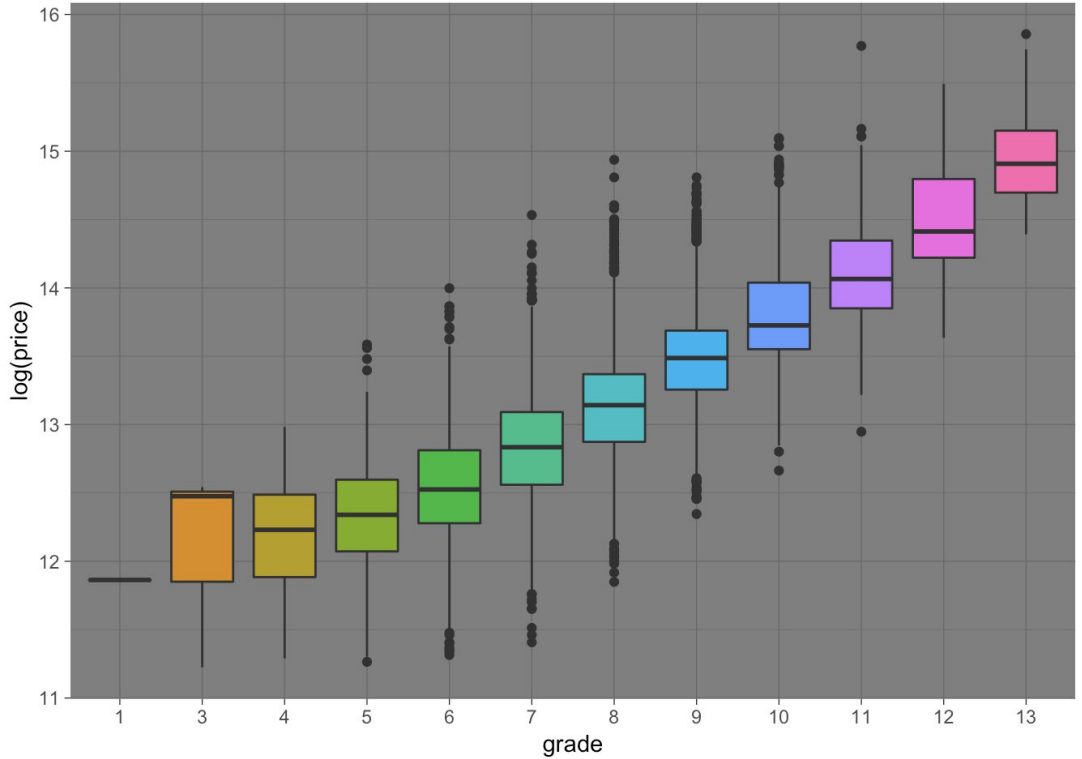




log(Price) vs. Grade



log(Price) vs. Grade



Data Analysis

Test Outliers

```
> fit11<-lm(price~date+bedrooms+bathrooms+floors+waterfront+condition+grade+ sqft_
living15+sqft_lot15+ house_age+renovated+basement,data=df2) #Without (lat&long|zip
code)
> abs(qt(0.1/2/21613,21602))
[1] 4.582199
> t <- rstudent(fit11)
> FF <- (t[ abs(t) > abs(qt(0.1/2/21613,21602))])
> FF
Price is prediction target
      269       311       516       653      1027      1159      1266      1275
 6.610098  4.746900  4.662947  6.732159  5.446636 12.152872  6.651493  5.487654
     1307      1424      1438      2029      2073      2254      2430      2611
11.977428  4.785484 16.251675  4.834841  6.908092  5.733216  6.307379  9.619468
     2848      2882      2956      3020      3728      3792      3849      3891
 6.376141  7.254814  6.768867  5.085909  6.644866  6.806783  6.758639 20.581171
     4011      4124      4164      4241      4311      4382      5419      5586
 5.249945 10.685872  6.719706  5.193250  4.897835 16.611261  6.522200  4.961215
```

Variable Selection

Variable Selection based on AIC Criteria and Stepwise Method

```
Step:  AIC=530680
price ~ grade + house_age + bathrooms + waterfront + sqft_living15 +
    basement + floors + condition + date + bedrooms + renovated +
    sqft_lot15


Call:
lm(formula = price ~ grade + house_age + bathrooms + waterfront +
    sqft_living15 + basement + floors + condition + date + bedrooms +
    renovated + sqft_lot15, data = df2)
```

The model include grade,house_age,bathrooms+waterfront,sqrt_living15,basement,floors,condition,bedrooms,renovated,date,sqft_lot15 variables has the smallest AIC.

Model 1 Summary

Summary the Model

P-value of condition

Variable >0.05, therefore,

Remove the condition

variable.

```
Call:
lm(formula = price ~ grade + house_age + bathrooms + waterfront +
    sqft_living15 + basement + floors + condition + date + bedrooms +
    renovated + sqft_lot15)

Residuals:
Price is prediction target
     Min       1Q   Median       3Q      Max
-1334111  -117276   -14081    90165  5168186

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.789e+06  2.299e+05 -12.131  < 2e-16 ***
grade         1.609e+05  2.196e+03  73.276  < 2e-16 ***
house_age     4.168e+03  7.460e+01  55.878  < 2e-16 ***
bathrooms     1.057e+05  3.456e+03  30.579  < 2e-16 ***
waterfront1   7.544e+05  1.818e+04  41.504  < 2e-16 ***
sqft_living15 1.007e+02  3.403e+00  29.592  < 2e-16 ***
basement      4.211e+04  3.605e+03  11.679  < 2e-16 ***
floors        3.195e+04  3.866e+03   8.266  < 2e-16 ***
condition2   -3.955e+04  4.510e+04  -0.877   0.3805
condition3   -4.840e+04  4.181e+04  -1.158   0.2469
condition4   -3.037e+04  4.181e+04  -0.726   0.4676
condition5    1.012e+04  4.206e+04   0.241   0.8098
date          9.170e+01  1.377e+01   6.658 2.85e-11 ***
bedrooms     -1.141e+04  2.008e+03  -5.683 1.34e-08 ***
renovated     3.228e+04  8.263e+03   3.907 9.38e-05 ***
sqft_lot15   -1.472e-01  5.825e-02  -2.528   0.0115 *
```

F-test & R-square

```
Residual standard error: 227600 on 21494 degrees of freedom
Multiple R-squared:  0.616,     Adjusted R-squared:  0.6158
F-statistic:  2299 on 15 and 21494 DF,  p-value: < 2.2e-16
```

From the model summary, we can see the p-value of F-test <2.2e-16, which means the Xi and y are linear relationships.

R-square=0.6158, which means all X variables explain 61.58% information of Y.

Model2 without Condition Variable

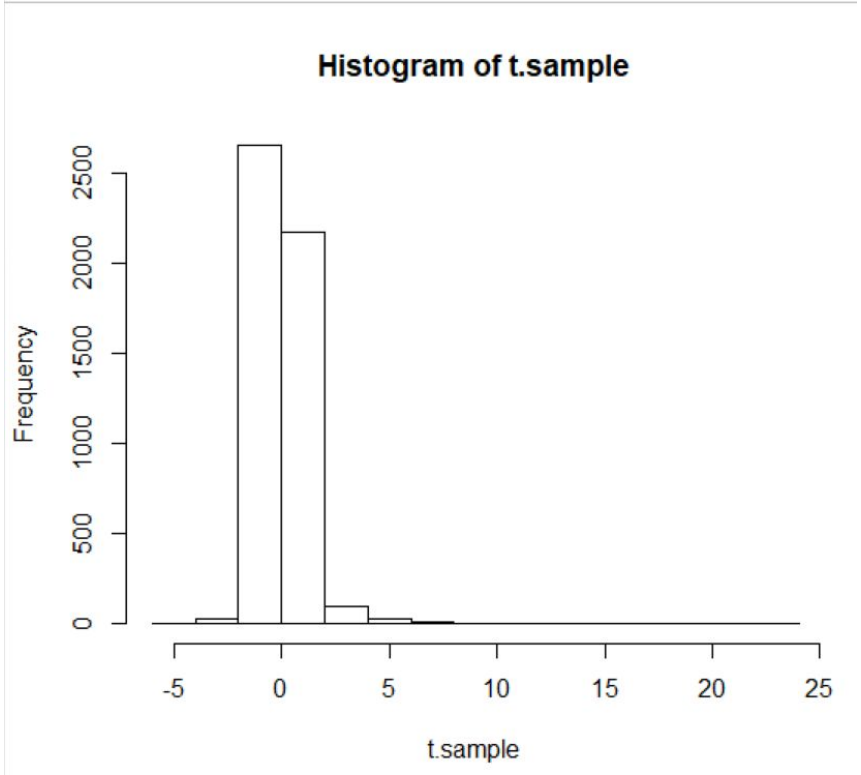
```
Call:
lm(formula = price ~ grade + house_age + bathrooms + waterfront +
    sqft_living15 + basement + floors + date + bedrooms + renovated +
    sqft_lot15, data = df2)

Residuals:
Price is prediction target
     Min       1Q   Median       3Q      Max
-1347936  -117649   -14403    90284  5159907

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.697e+06  2.263e+05 -11.919  < 2e-16 ***
grade          1.605e+05  2.194e+03  73.149  < 2e-16 ***
house_age      4.401e+03  7.027e+01  62.630  < 2e-16 ***
bathrooms      1.090e+05  3.443e+03  31.665  < 2e-16 ***
waterfront1    7.582e+05  1.821e+04  41.642  < 2e-16 ***
sqft_living15  1.001e+02  3.403e+00  29.414  < 2e-16 ***
basement       4.236e+04  3.611e+03  11.730  < 2e-16 ***
floors         2.874e+04  3.834e+03   7.496 6.85e-14 ***
date           8.307e+01  1.377e+01   6.032 1.64e-09 ***
bedrooms      -1.074e+04  2.009e+03  -5.346 9.09e-08 ***
renovated      1.919e+04  8.145e+03   2.356   0.0185 *
sqft_lot15    -1.375e-01  5.830e-02  -2.359   0.0183 *
---
Signif. codes:  0 ¡®***¡¯ 0.001 ¡®**¡¯ 0.01 ¡®*¡¯ 0.05 ¡®.¡¯ 0.1 ¡® ¡¯

Residual standard error: 228100 on 21498 degrees of freedom
Multiple R-squared:  0.6144,    Adjusted R-squared:  0.6142
F-statistic:  3114 on 11 and 21498 DF,  p-value: < 2.2e-16
```

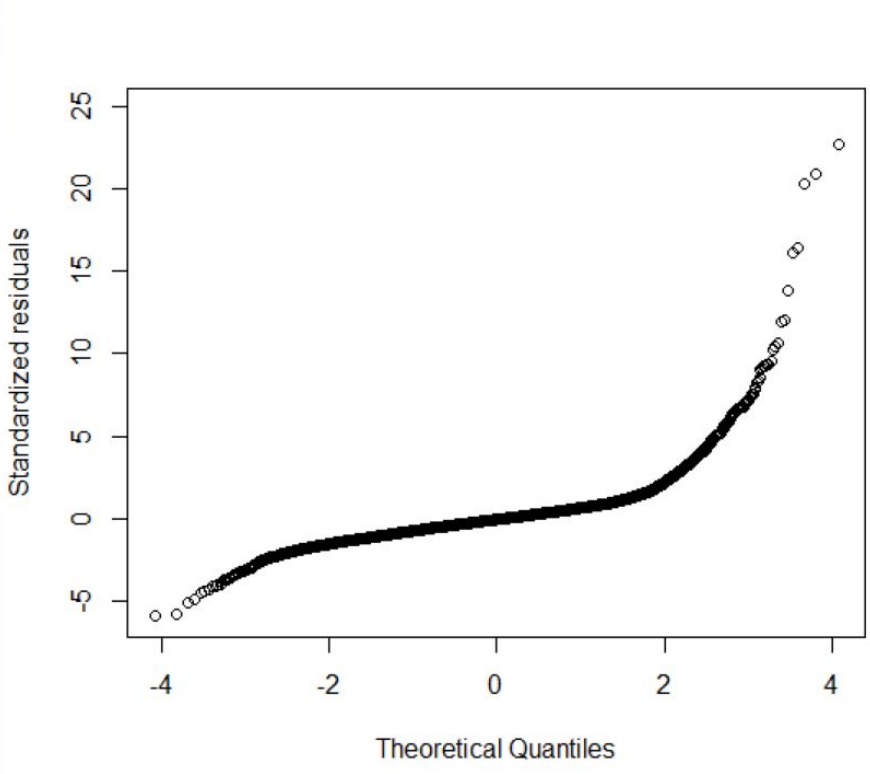
The p-value of all variables are less than 0.05, therefore, all variables has significant effect on Y.

P-value of F-test<2.2e-16, conclude that all Xi has linear relationship with Y.

R-square=0.6142,which means all X variables explain 61.42% information of Y. (total variance of Y)

Test Normal Distribution of Residuals





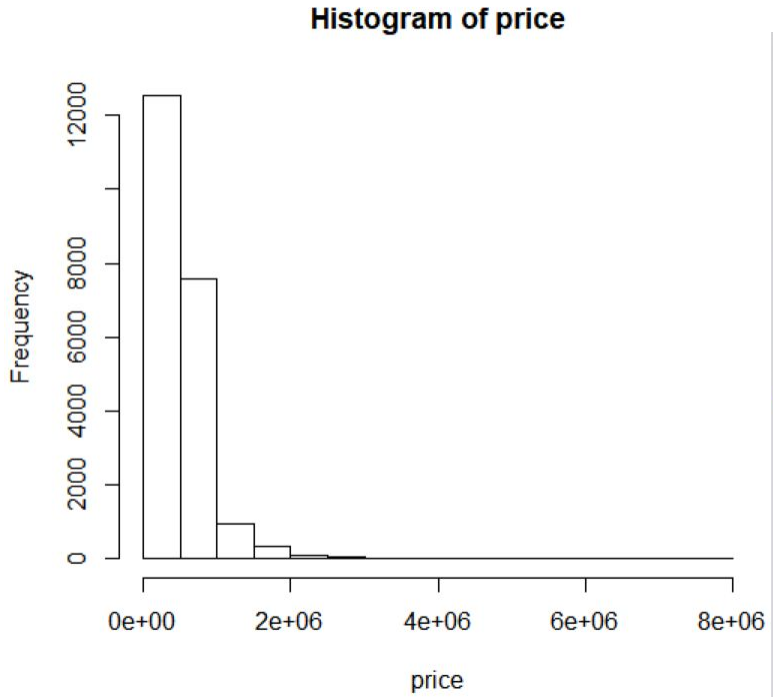
Test Normal Distribution

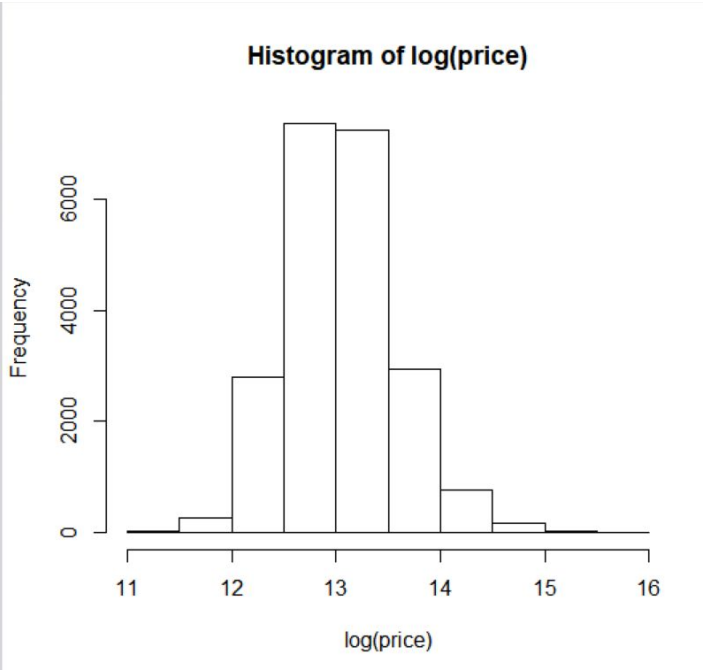
```
> shapiro.test(t.sample)

        Shapiro-Wilk normality test

data:  t.sample
W = 0.84745, p-value < 2.2e-16
```

Dataset is too large to do Shapiro-Test. We need to do sampling.

Reject null-Hypothesis, which means it is non-normal distribution.

Test Normal Distribution.



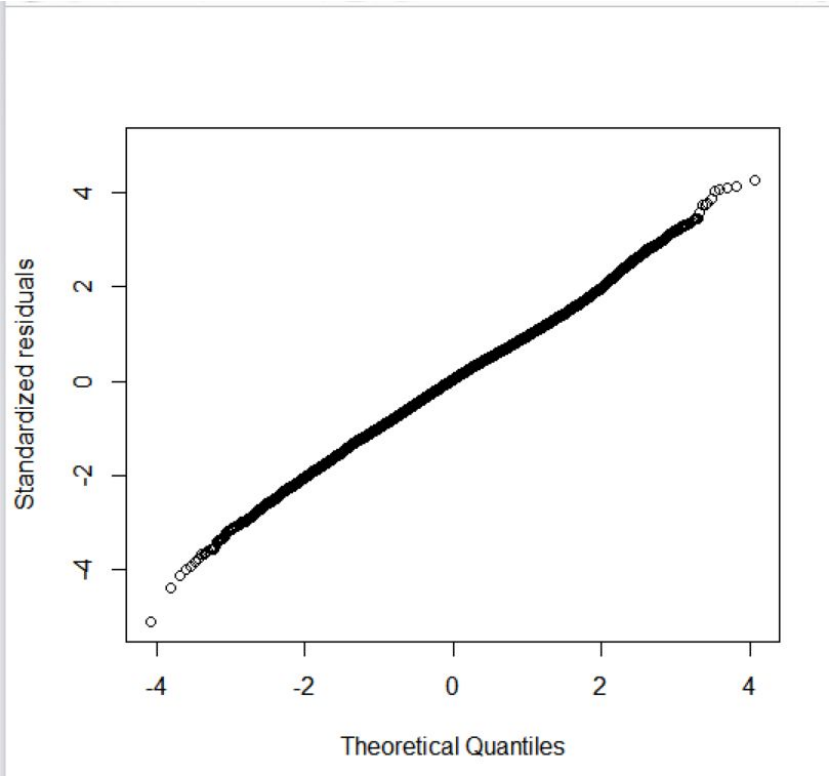


Check Independent variable(Price)

Need to do transformation

Test Normal Distribution

Plot the model with log(price) variable.



Test Normal Distribution

```
> shapiro.test(t.sample3)

        Shapiro-Wilk normality test

data:  t.sample3
W = 0.99873, p-value = 0.000594
```

Model3 Summary with Log(Price) Variable

```
Call:
lm(formula = log(price) ~ grade + house_age + bathrooms + waterfront +
    sqft_living15 + basement + floors + date + bedrooms + renovated +
    sqft_lot15, data = df2)

Residuals:
Price is prediction target
    Min      1Q  Median      3Q     Max
-1.5989 -0.2090  0.0102  0.2062  1.3175

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.150e+00  3.108e-01  26.218  < 2e-16 ***
grade          2.337e-01  3.013e-03  77.555  < 2e-16 ***
house_age      6.214e-03  9.662e-05  64.314  < 2e-16 ***
bathrooms      1.165e-01  4.731e-03  24.632  < 2e-16 ***
waterfront1    5.145e-01  2.493e-02  20.638  < 2e-16 ***
sqft_living15  1.670e-04  4.671e-06  35.743  < 2e-16 ***
basement       1.286e-01  4.962e-03  25.915  < 2e-16 ***
floors         1.221e-01  5.262e-03  23.202  < 2e-16 ***
date           1.239e-04  1.892e-05   6.550  5.9e-11 ***
bedrooms      -1.523e-03  2.760e-03  -0.552    0.581
renovated      1.771e-02  1.120e-02   1.582    0.114
sqft_lot15     5.944e-08  7.998e-08   0.743    0.457
---
Signif. codes:  0 ¡®***¡¯ 0.001 ¡®**¡¯ 0.01 ¡®*¡¯ 0.05 ¡®.¡¯ 0.1 ¡® ¡¯

Residual standard error: 0.3132 on 21485 degrees of freedom
Multiple R-squared:  0.6466,    Adjusted R-squared:  0.6464
F-statistic:  3573 on 11 and 21485 DF,  p-value: < 2.2e-16
```

There are three variables insignificant, which are bedrooms,renovated and sqft_lot15.

Model4 Summary without Insignificant Variables

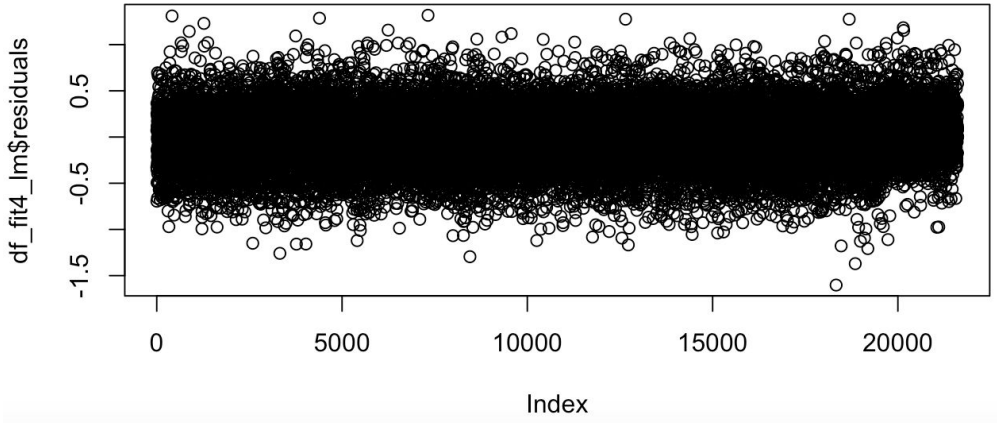
```
Call:
lm(formula = log(price) ~ grade + house_age + waterfront + bathrooms +
    date + sqft_living15 + basement + floors, data = df2)

Residuals:
Price is prediction target
    Min      1Q  Median      3Q     Max
-1.6016 -0.2089  0.0105  0.2067  1.3176

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.153e+00  3.108e-01   26.23  < 2e-16 ***
grade         2.339e-01  3.010e-03   77.69  < 2e-16 ***
house_age     6.251e-03  9.162e-05   68.22  < 2e-16 ***
waterfront1   5.187e-01  2.481e-02   20.91  < 2e-16 ***
bathrooms     1.166e-01  4.347e-03   26.82  < 2e-16 ***
date          1.233e-04  1.891e-05    6.52 7.21e-11 ***
sqft_living15 1.667e-04  4.580e-06   36.40  < 2e-16 ***
basement      1.280e-01  4.943e-03   25.89  < 2e-16 ***
floors        1.223e-01  5.226e-03   23.40  < 2e-16 ***
---
Signif. codes:  0 ¡®***¡¯ 0.001 ¡®**¡¯ 0.01 ¡®*¡¯ 0.05 ¡®.¡¯ 0.1 ¡® ¡¯ 1

Residual standard error: 0.3132 on 21488 degrees of freedom
Multiple R-squared:  0.6465,    Adjusted R-squared:  0.6464
F-statistic:  4912 on 8 and 21488 DF,  p-value: < 2.2e-16
```

Compare 4 Models

```
> selcri(fit1)
           rsq   adj.rsq    aic      bic         press
[1,] 0.6160375 0.6157696 530680 530807.6 1.117563e+15
> selcri(fit2)
           rsq  adj.rsq      aic      bic         press
[1,] 0.6143993 0.614202 530763.6 530859.3 1.121854e+15
> selcri(fit3)
           rsq   adj.rsq       aic      bic    press
[1,] 0.6465549 0.6463739 -49899.81 -49804.1 2110.227
> selcri(fit4)
           rsq   adj.rsq      aic       bic    press
[1,] 0.6464971 0.6463655 -49902.3 -49830.52 2109.896
>
```

Test Constant Variance of Residuals

```
> ncvTest(fit4)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.487297, Df = 1, p = 0.22264
>
```



Test Multicollinearity

```
> vif(fit4)
        grade     house_age    waterfront     bathrooms          date
     2.743104      1.585819      1.014823      2.456618      1.001746
sqft_living15      basement        floors
     2.159156      1.276826      1.745112
```

Test Autocorrelation

```
> dwtest(fit4)

        Durbin-Watson test

data:  fit4
DW = 1.9684, p-value = 0.01015
alternative hypothesis: true autocorrelation is greater than 0
```

DW close to 2. Residual do not has serial correlated

K-fold for 4 models

```
library(boot)

df_fit1_glm <- glm(price ~  grade + house_age + bathrooms + waterfront + sqft_living15 +
                      basement + floors + condition + date + bedrooms + renovated +
                      sqft_lot15, data=df)

df_fit2_glm <- glm(price ~ grade + house_age + bathrooms + waterfront +
                      sqft_living15 + basement + floors + date + bedrooms +
                      renovated + sqft_lot15,data=df)

df_fit3_glm <- glm(log(price) ~ grade + house_age + bathrooms + waterfront +
              sqft_living15 + basement + floors + date + bedrooms +
              renovated + sqft_lot15, data = df)
df_fit4_glm <- glm(log(price) ~ grade + house_age  + waterfront + bathrooms+date+
              sqft_living15 + basement + floors, data=df)

cv_10K_1 <- cv.glm(df, df_fit1_glm, K=10)
cv_10K_2 <- cv.glm(df, df_fit2_glm, K=10)
cv_10K_3 <- cv.glm(df, df_fit3_glm, K=10)
cv_10K_4 <- cv.glm(df, df_fit4_glm, K=10)

mse_all <- c("MSE Fit1"=cv_10K_1$delta[1], "MSE Fit2"=cv_10K_2$delta[1],
             "MSE Fit3"=cv_10K_3$delta[1],"MSE Fit4"=cv_10K_4$delta[1])

mse_all
```

```
> mse_all
    MSE Fit1     MSE Fit2     MSE Fit3     MSE Fit4
5.195074e+10 5.208202e+10 9.819972e-02 9.818896e-02
```

Estimated Regression Equation

log(Y) = 2.339e-01(grade) + 6.251e-03(house_age) + 5.187e-01(waterfront) + 1.166e-01(bathrooms) + 1.233e-04(date) + 1.667e-04(sqft_living15) + 1.280e-01(basement) + 1.223e-01(floors)

With a log - linear model, when x1 is increase by 1 unit, y will increase by (100*Beta1) %

Example :
When grade goes up by 1 unit, price will increase by 100* 0.2339 % = 23.39 %, keeping other variables constant.

Reference

- Harlfoxem. “House Sales in King County, USA.” *RSNA Pneumonia Detection Challenge | Kaggle*, 25 Aug. 2016, www.kaggle.com/harlfoxem/housesalesprediction.
- Rosenberg, Mike. “Seattle Home Prices Drop by $70,000 in Three Months as Market Continues to Cool.” The Seattle Times, The Seattle Times Company, 9 Sept. 2018, www.seattletimes.com/business/real-estate/seattle-home-prices-drop-by-70000-in-three-months-as-market-cooldown-continues/