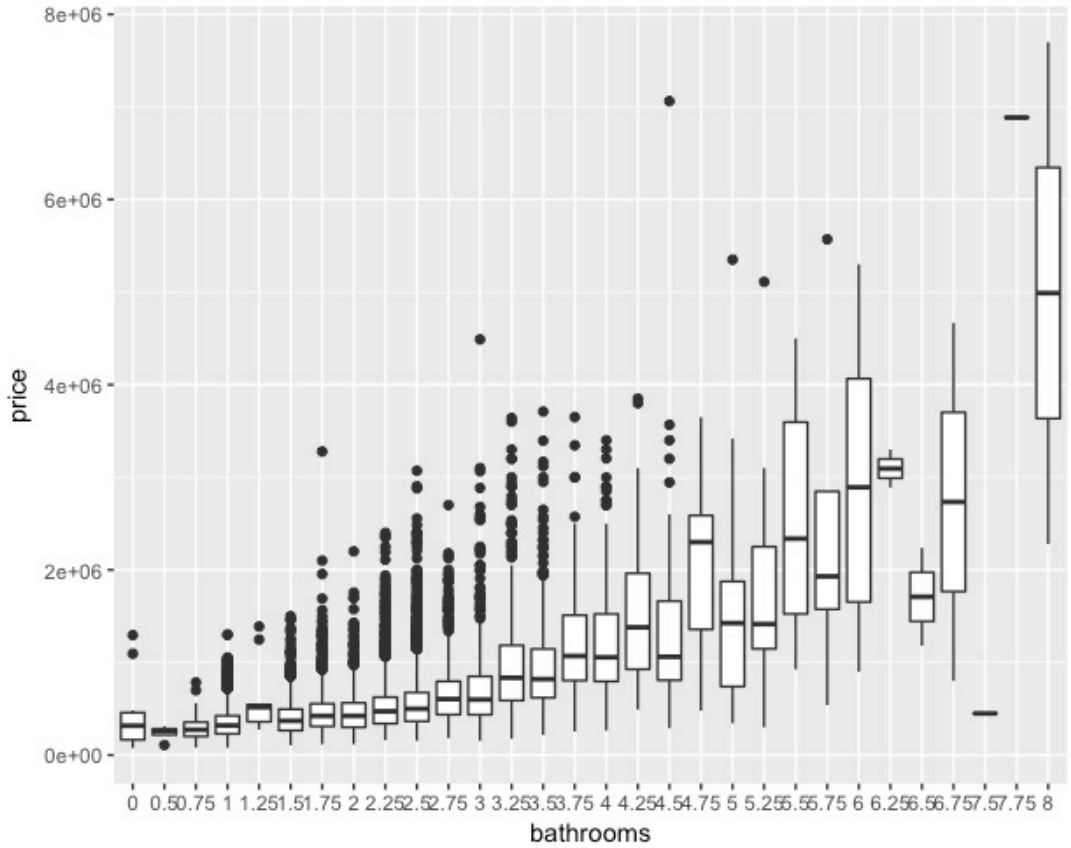
Question 1: descriptive statistics (25 points)
Of the variables in the dataset, which variables are interesting for predict house price? The questions you should address are:

1. **(presentation point) Give a clear list of the interesting variables.**
   Interesting variables:
   bathrooms, bedrooms, view, grade, sqft_living, sqft_living15, sqft_above, sqft_basement

   not interested variables:
   id, date, sqft_lot, floors, condition, waterfront, yr_built, yr_renovated, zipcode, long, lat

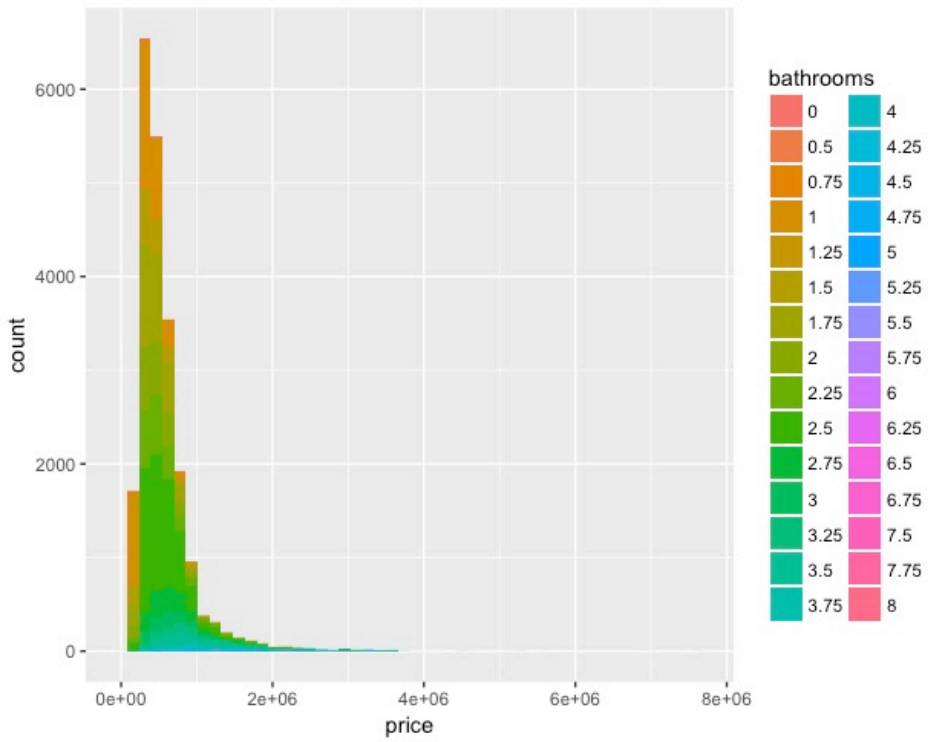
2. **(10pt) For each of the interesting variable, explain by plot(s) and num- ber(s) its relationship with price. Write a brief summary.**

   BATHROOMS
   Plot – Bathrooms vs Price Boxplot



   Their relationship is positive, and the correlation value between bathroom number and price is 0.5251. The more bathrooms a house have, the average of the price tends to be higher.

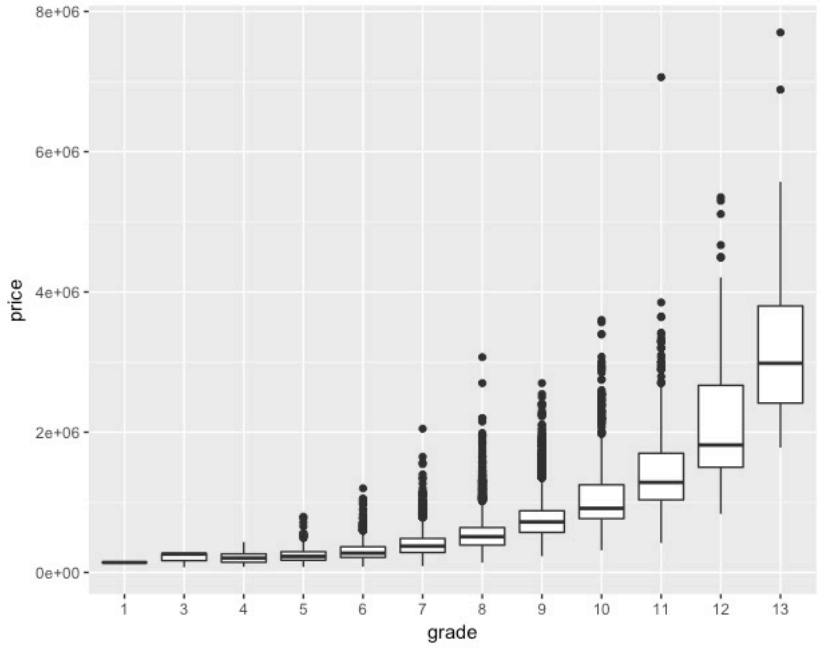
Plot – Bathrooms vs Price Histogram



Most of the houses have less than three bathrooms, and the majority of the houses have price lower than 2e+06 dollars, which is mainly constructed with about two bathrooms.

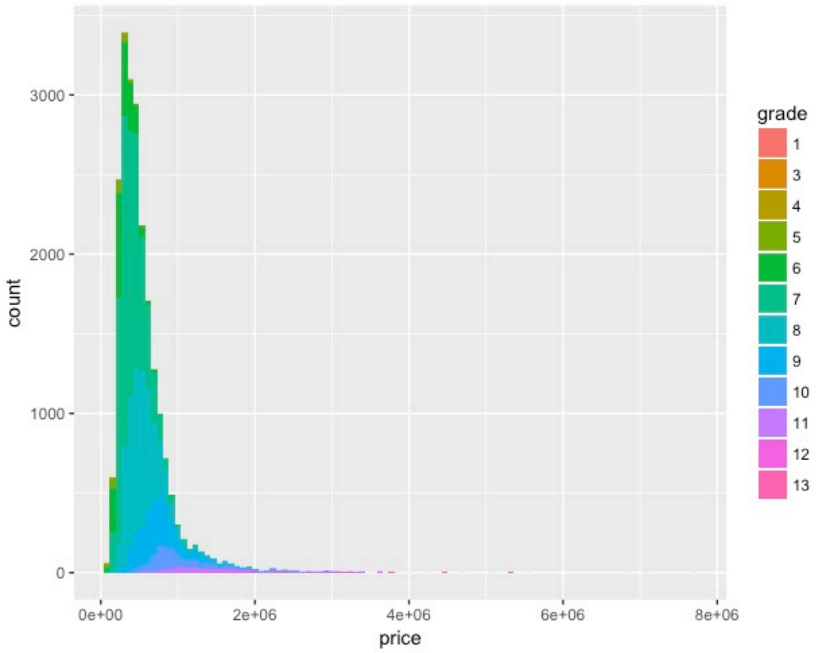
GRADE

Plot – Grade vs Price Boxplot



The relationship between grade and price is positive with the correlation value of 0.6674. The higher the grade of the house is, the higher the average price would be.

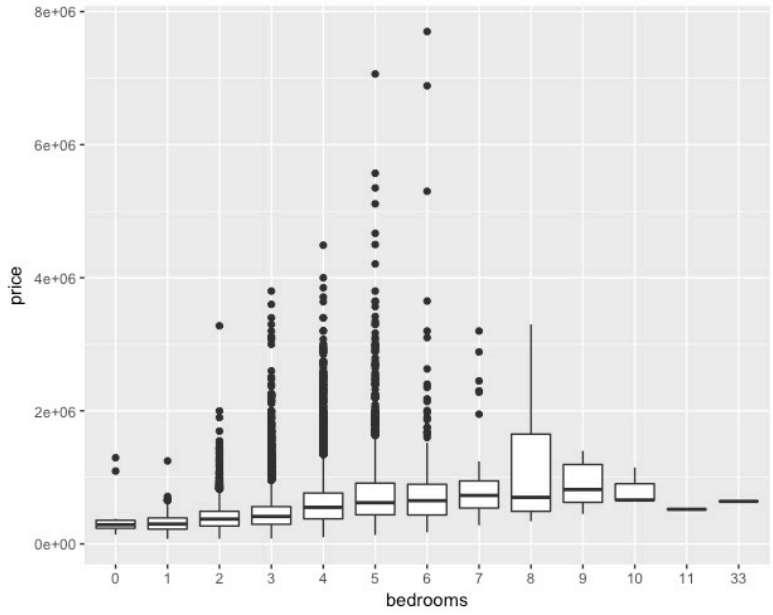
Plot – Grade vs Price Histogram



From this plot, we can see most houses with less than 2e+06 dollars would be around grade 7 or 8. Rarely do we see houses with grade less than 5 or higher than 10. And the price is relative low for houses with grades less than 5 and relatively high for houses with grade more and 10. We could see from the graph that houses priced more than 2e+06 is all greater or equal to grade 10.

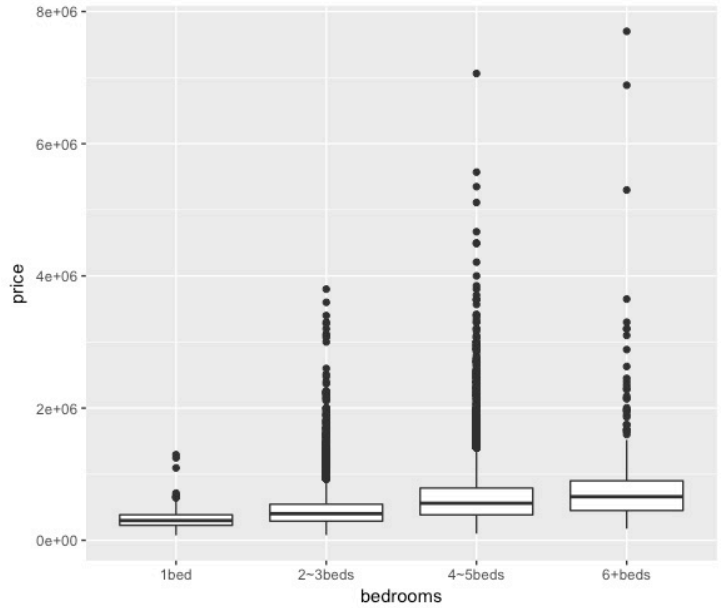
BEDROOMS

Plot – Bedrooms vs Price with 14 categories



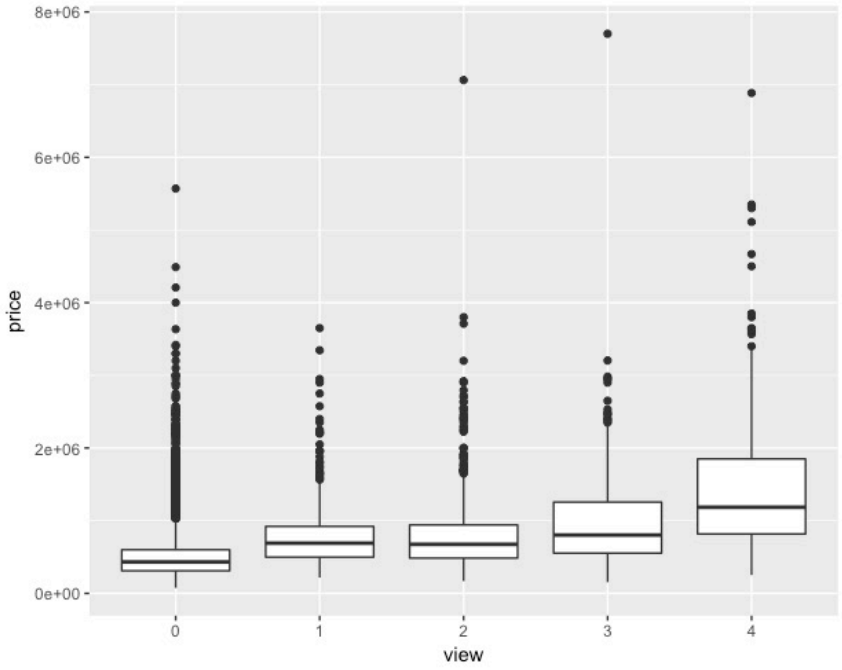
The relationship between bedrooms and price is positive with a correlation value of 0.3083496. Although houses with 4-6 bedrooms seems to contain many houses with extremely high prices, but the average price value does increase as bedroom numbers increase little by little.
To eliminate the extreme values larger than 6 bedrooms, I group them into four major categories: 0-1beds, 2-3beds, 4-5beds, and 6+beds so that the extreme values such as 11bedrooms or 33bedrooms are all included in 6+beds category. The graph is also easier for interpretation, which is shown as the following.

Plot – Bedrooms vs Price with 4 categories



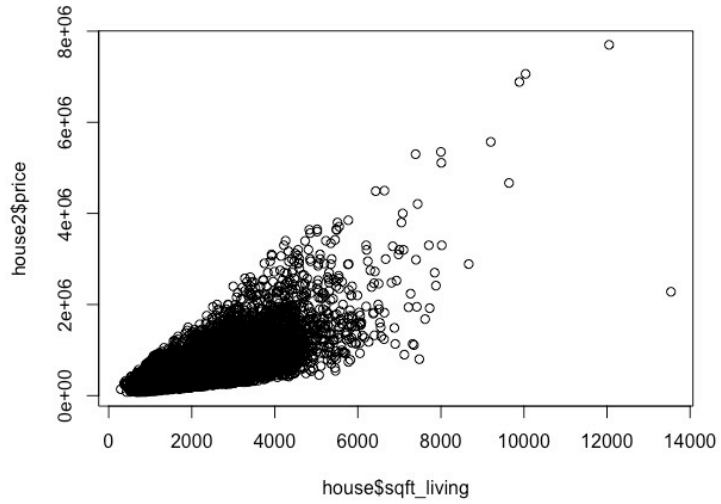
VIEW

Plot – View vs Price Boxplot



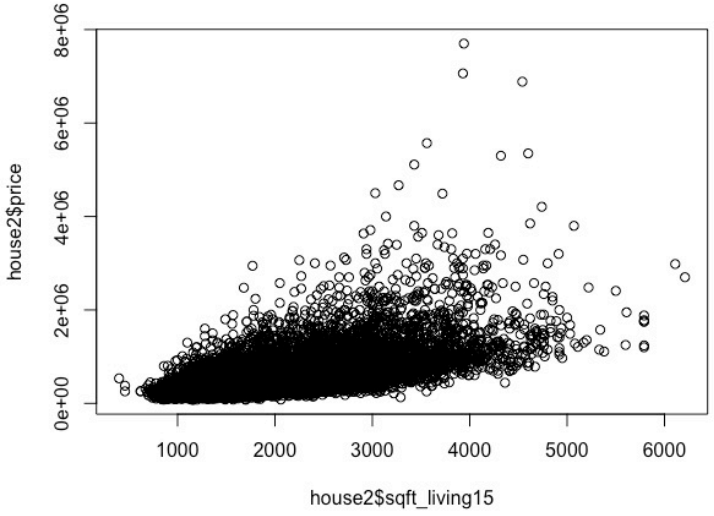
From boxplot, we notice the average of price increase as number of view increase. This is a positive relationship with a correlation value of 0.3972935. The correlation value is so low because there are many points outside the 3rd-quatile. Those points make the relationship not obvious.

SQFT_LIVING Plot – Sqft_living vs Price



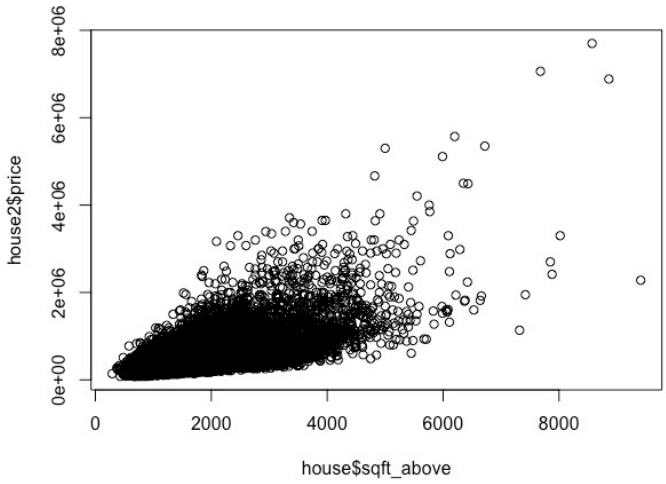
The correlation value is 0.7020, which shows that their relationship is rather close. Obviously, the area of living area has a positive relation with the price, but with a broad range. But at least this would be helpful to determine the possible price range. For living area larger than 8000sqft, it shows a more obvious trend of "the larger living area is, the higher price it can be," except one house of about 13000sqft living area.

SQFT_LIVING15 Plot – Sqft_living15 vs Price



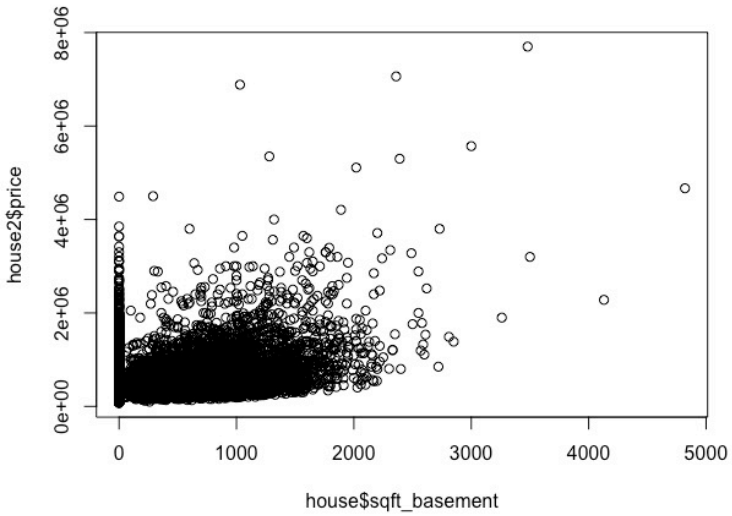
The correlation value of sqft_living15 and price is 0.5853789, which shows that this is a positive relationship with some variation. Since Sqft_living15 measures the square footage of living area for its closest neighbors, its price range becomes larger than that of sqft_living. However, the relationship between it and the price is still positive with no obvious violation. Another interesting point comparing to sqft_living is that the range of sqft_living15 domain only covers area less than 6500sqft where the maximum sqft_living area is almost 14000sqft. Thus, by averaging the neighborhood living area, the outliers are eliminated.

SQFT_ABOVE Plot – Sqft_above vs Price



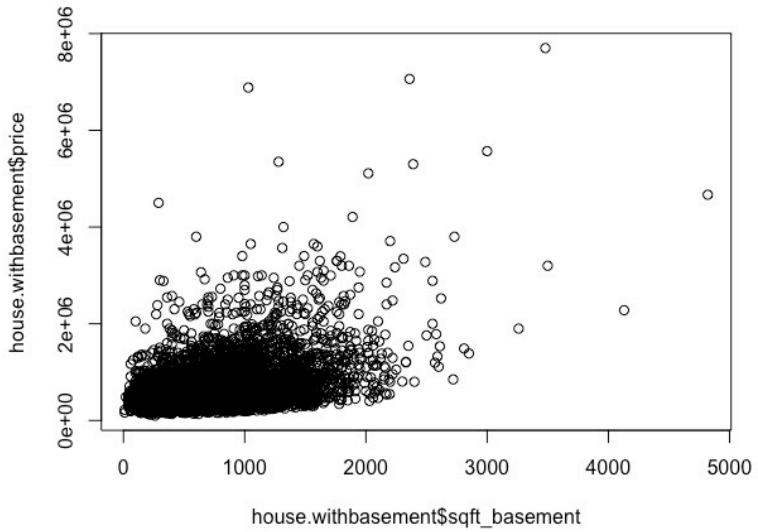
Similar to sqft_living15, sqft_above also shows a positive relationship between itself and price with a correlation value of 0.6055673.

SQFT_BASEMENT Plot – Sqft_basement vs Price



The positive relationship between the area of basement and price is a little bit blurred with a correlation value of 0.323816. If eliminating the houses with no basement, the correlation value would be higher, which becomes 0.4073082, and the graph looks better as shown in below.

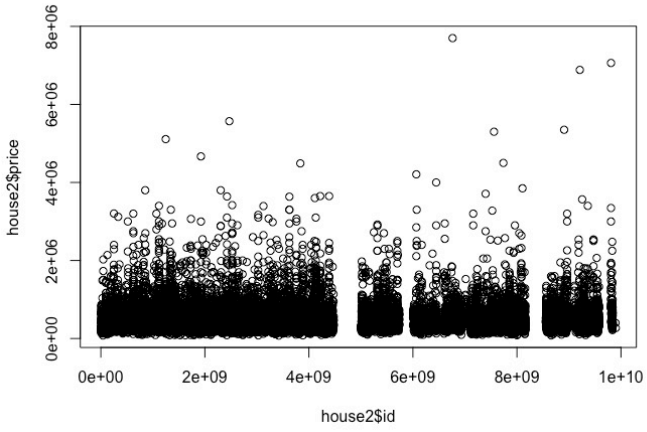
Plot – Sqft_basement vs Price excluding houses without basement



3. **(10pt) For each of the uninteresting variable, explain by plot(s) and number(s) why it is uninteresting. Write a brief summary.**

ID

Plot – House ID vs Price

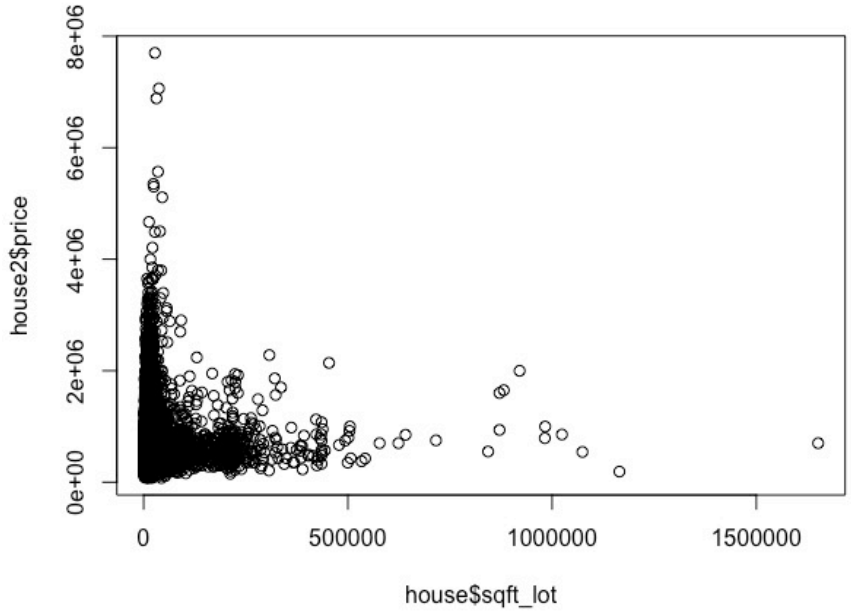


Since ID is for identifying, the price has little relation with the id number, except there are some number that is not assigned with a price, other id all have randomly distributed price. Thus， we could not use ID to predict the price of the house.

DATE Plot – Date vs Price

Date has no relationship to price, since we can observe all prices are randomly distributed. Not an interesting variable for predicting price.

SQFT_LOT
Plot – Sqft_lot vs Price

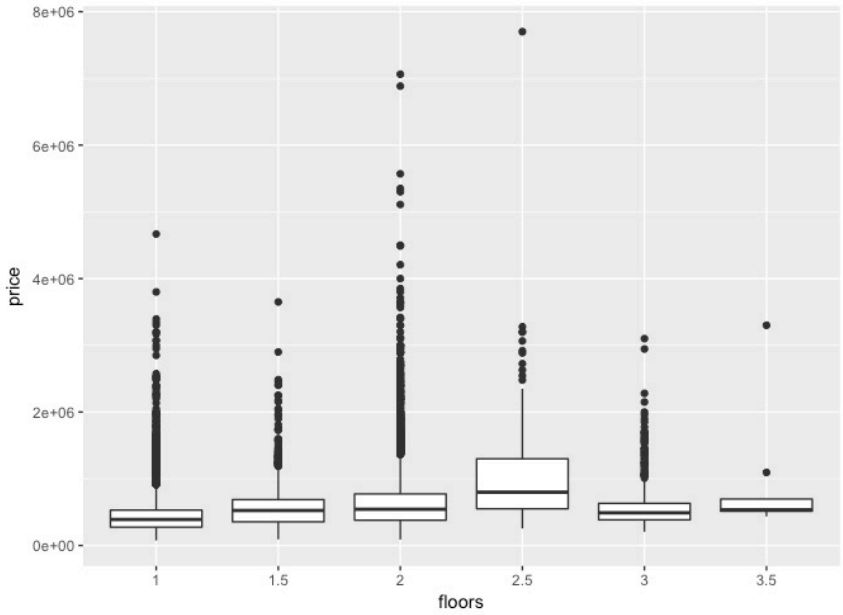


Although the graphs seems to tell us that the sqft_lot variable has an inverse relationship with price, the price seems to be distributed randomly within the range of 2,000,000 dollars after the area of lot getting bigger than 100,000sqft and. And for houses with lot less than 100,000sqft, the price spread out in a larger range. This

distribution also are not helpful for predicting the price. We could give it a try when producing the regression model, but this is a variable that could be eliminated since other variables probably have a stronger influence on the final price.

FLOORS

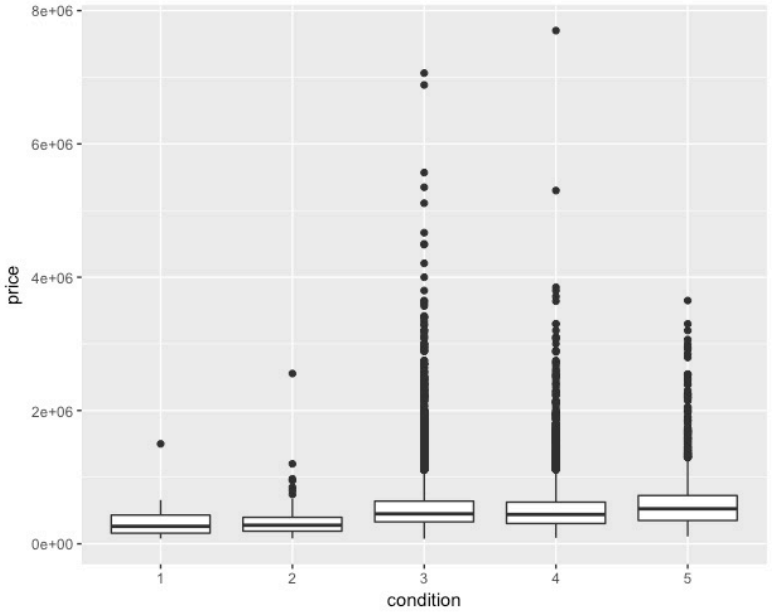
Plot – Floors vs Price



The relationship between floors and price are not obvious. We can see the average price for 2.5 floor is the highest and lowest for 1 floor. The mean price for each category increase from 1 to 2.5 but decrease as floor number gets higher than 2.5. The range of the price is also very large. Due to the large oscillation of price for different floors, number of floors is not useful to predict price of the house.

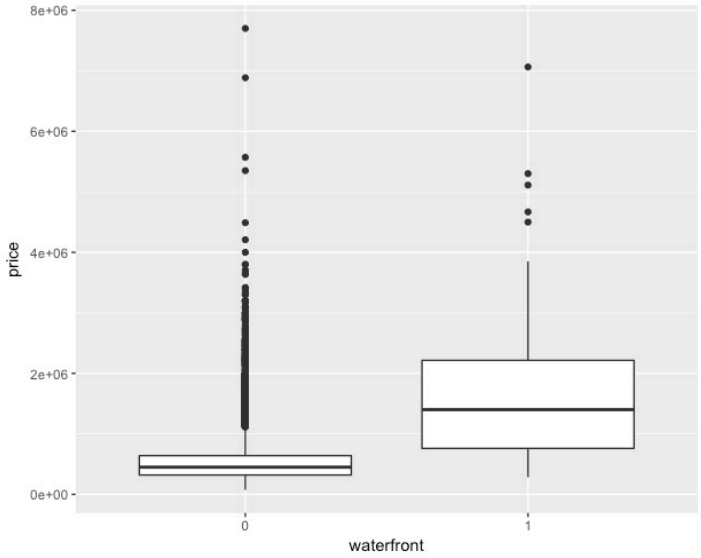
CONDITION

Plot – Condition vs Price



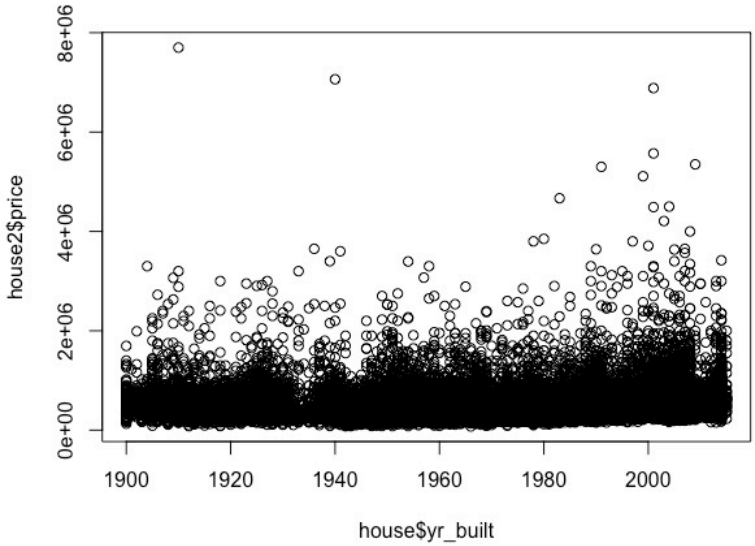
The relationship between condition and price is not obvious, and the correlation value is also close to zero (0.03636179) which means the price is pretty random by condition. Thus, this is not an interesting value for predicting house price.

WATERFRONT Plot – Waterfront vs Price



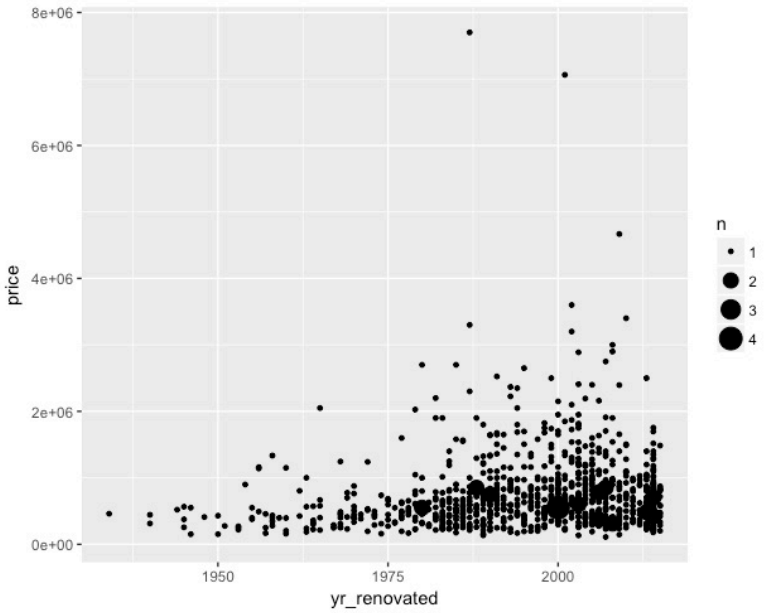
Since the waterfront only have two categories, the correlation seems to be high by number (0.2663694), but the price range for house with no waterfront is very large that almost also cover the price range for house with one waterfront. Hence, even though most houses with no waterfront would averagely have lower price, but it doesn't mean that the price is lower for sure. Thus, to make the model simple, we would not consider this variable as an interesting one.

YR_BUILT Plot – Year Built vs Price



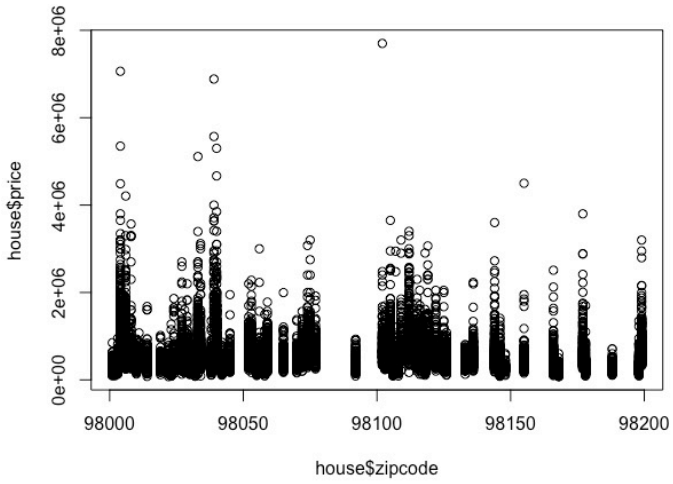
Almost no relationship exists between year built and price. Even though higher price houses with price between 4,000,000 and 6,000,000 were built around year 2000, there are still high prices exist for 1910 and 1940. Hence this is not a variable that could help us predict house price.

YR_RENOVATED Plot – Year Renovated vs Price



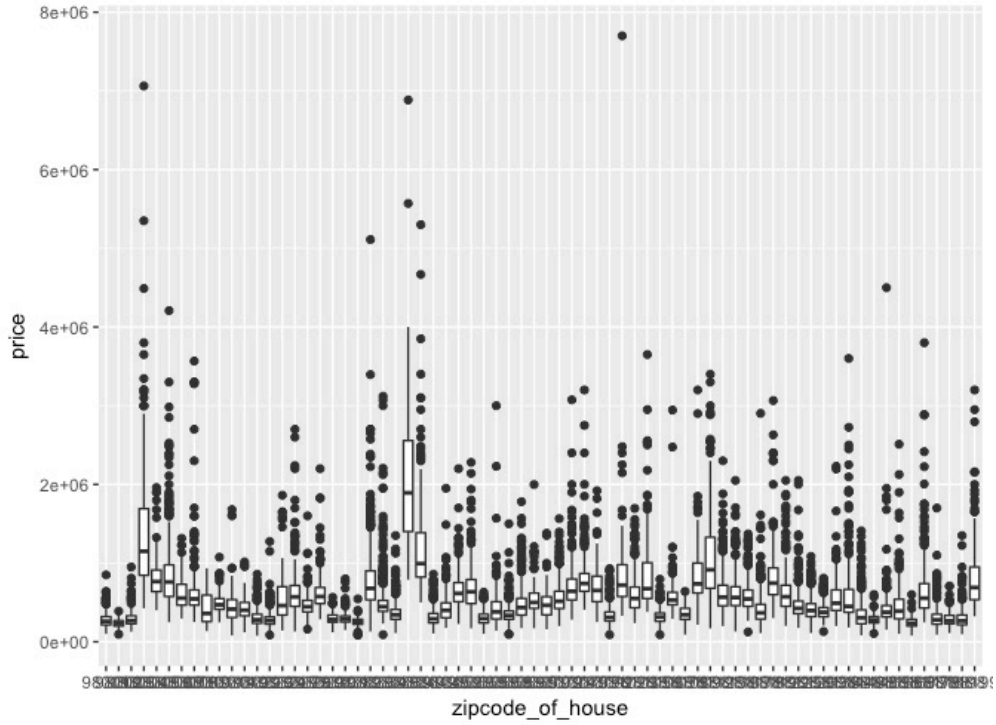
We can observe that before 2010, the more recently the house is renovated, the larger range of the price is. Even though the variable gives a trend of the price range, there is low price houses no matter where the house it. The correlation value is also very low (0.1264338), which indicate they are less possibly associated.

ZIPCODE Plot – Zipcode vs Price



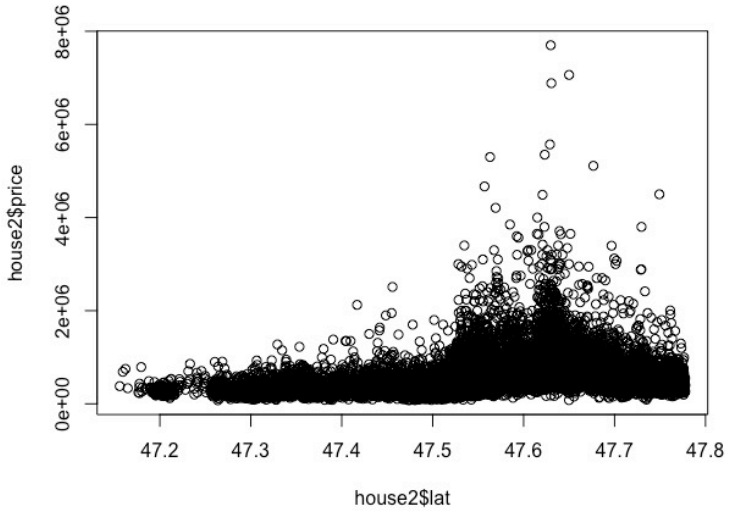
A general plot shows two variables are not very related. To view if there are some houses in certain zipcode, I generate a boxplot in below.

Plot – Boxplot of Zipcode vs Price



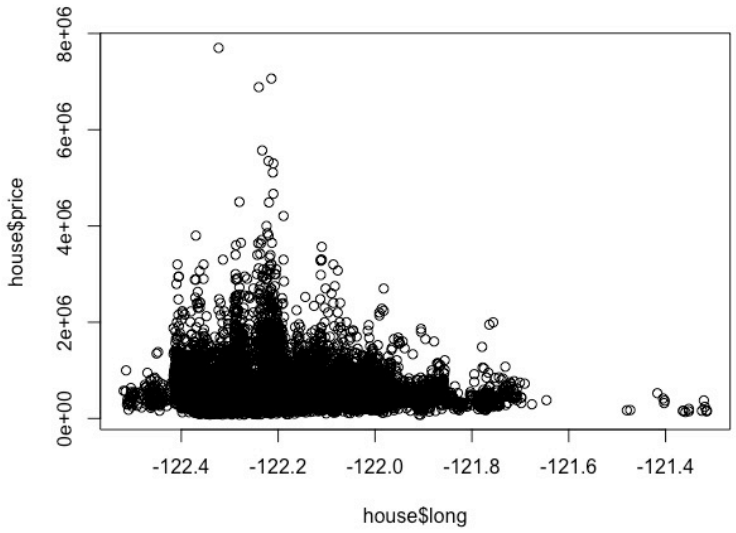
Generally, the average price in each zipcode is less than 1,000,000 dollars, which vibrates a lot without patterns. Only three zipcode have a obviously higher average price, but also it has a extreme large price range that we cannot count on the mean only for the regression model later. Hence, zipcode is not a good reference for price of the house.

LATITUE Plot – Latitude vs Price



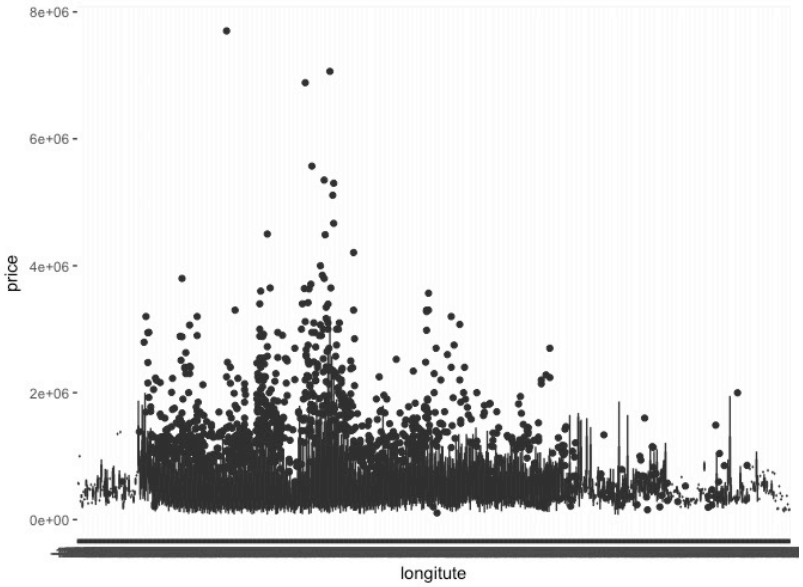
The relationship between latitude and price is not as obvious and indeed we could only see that many expensive houses located around latitude 47.6.

LONGITUDE Plot – Longitute vs Price



We could observe that most houses are located between -122.5 and -121.7 in terms of longitude. Although from above plot, the house around -122.2 would possibly have a higher price, the boxplot tell us the average price or even the 3rd-quatile is not much larger than other.

Plot – Longitude vs Price Boxplot

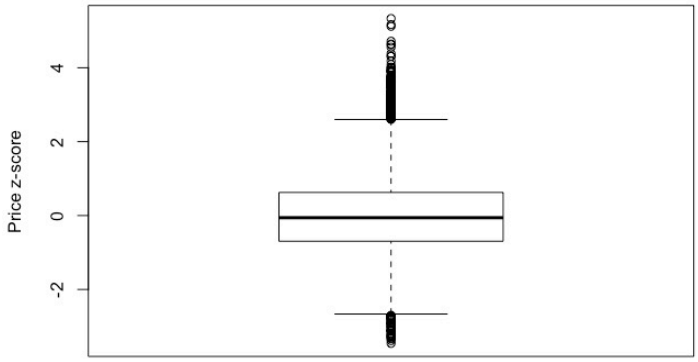


All the lines are representing the length from 1st-quatile to 3rd-quatiles and points are values outside the IQR. We could see many dots are around -122.2 but the lines are generally the same as other part of the graph. Thus, the price of the house in such longitude are not necessarily higher. This variable is not interesting for predicting the price.

4. **(5pt) What is an outlier? Are there outliers in your data? Should you exclude such outliers from the regression model below, and why?**

   Outliers are observation points that is far from other observation points.

   From the first sight, the price of house spread out to a very large range, thus we need some investigation to determine if there are some outliers. Thus, we use log to narrow it down, and the price now become more normally distributed. I rescale the price by its mean and standard deviation and get the corresponding z-score. Then plot the scaled price using boxplot. The result is the following:



   Three points are very far from other parts and the value is more than 3 standard deviation and IQR. Thus, we want to find out which three points are they and see if they are really outliers and whether we need to eliminate them or not. The three points are with z-score 5.120786 5.169115 5.333200, and the corresponding to actual value 6885000 7062500 7700000. The rest of the price value are less than 5570000, with a mean of 540088. Hence, we would eliminate above three data points.

   Additionally, when looking at the interesting variables categories, one observation point has an extremely high number of bedrooms and another observation point has extremely large area of living that is distant from other observation point values. The value of the number of bedrooms is 33, which is not realistic for a normal house, which could be consider as a typo. Since we have no access to the original data entry person, the value would be eliminated from the dataset. Then for the point with high living area, it does not follow the other trend and is distant from other observation point, thus we would also eliminate this point.

5. **(Bonus: 5pt): Are there interesting relationships amongst the input variables? (input variables are all variables except price).**
   Bedrooms vs Bathrooms; Bedrooms vs sqft_living; Bathrooms vs sqft_living; Bathrooms vs grade; grade vs sqft_living, etc.

Question 2: inferential statistics (30 points)
Run a linear regression model of price vs other input variables. The questions you should address are:

1. **(20pt) Detail how you did variable selection: which models did you run, why did you discard certain models or variables, any variable transfor- mations you did and why, which diagnostic tests did you run and what they showed, justifications if you removed outliers.**

   The variables selected are what's in the list of interesting variables related to price. In case if I miss some important variables, I also use R to help me select the variables that are significant for building a regression model for predicting price.

   The result of selecting variables by AIC is bathrooms, bedrooms, grade, view, sqft_living, sqft_lot, floors, waterfront, condition, sqft_above, yr_built, yr_renovated, zipcode, lat, long, sqft_lot15, and sqft_living, which include all interesting variables. For simplicity, just use the interesting variables.

   To improve the result R-square, p-value for each coefficient and simplicity, I eliminate sqft_living15, bathrooms, and use log for sqft_living and price. Putting log on price largely improved the diagnostic tests and give a very good result.

2. **(10pt) Call your final regression model model.lm. Clearly show your final regression model: the R command, and the R output summary. Write down the equation that R gives you. Interpret all the coefficients and the p-values associated with the coefficients. Report the R2 and adjusted R2 of your model. What are the meaning of these values? Run a diagnostic plot for your model, and explain if your model is a good fit.**

   Code:

```
model.lm <- lm(log(price) ~ bedrooms + grade + view + log(sqft_living) + sqft_above +
sqft_basement, data = house[!select,])
summary(model.lm)
plot(model.lm)
```

   Result:

```
Call:
lm(formula = log(price) ~ bedrooms + grade + view + log(sqft_living) +
   sqft_above + sqft_basement, data = house[!select, ])

Residuals:
    Min      1Q  Median      3Q     Max
-1.41794 -0.24418  0.00611  0.22896  1.33597

Coefficients:
```

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.039e+01  1.296e-01  80.215  < 2e-16 ***
bedrooms        -2.248e-02  3.397e-03  -6.618 3.74e-11 ***
grade            1.888e-01  3.270e-03  57.722  < 2e-16 ***
view             9.078e-02  3.256e-03  27.878  < 2e-16 ***
log(sqft_living)  1.278e-01  2.001e-02   6.385 1.75e-10 ***
sqft_above       1.306e-04  9.368e-06  13.946  < 2e-16 ***
sqft_basement    2.257e-04  1.042e-05  21.663  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3408 on 21601 degrees of freedom
Multiple R-squared:  0.5799,          Adjusted R-squared:  0.5797
F-statistic:  4969 on 6 and 21601 DF,  p-value: < 2.2e-16.
```

Given equation by R:

$$\log(Y_{Price}) = 10.39 - 0.02248X_{Bedrooms} + 0.1888X_{Grade} + 0.09078X_{View} + 0.1278 \times \log(X_{sqft-living}) + 1.306\times10^{-4}X_{sqft-above} + 2.257\times10^{-4}X_{\text{sqft-basement}}$$

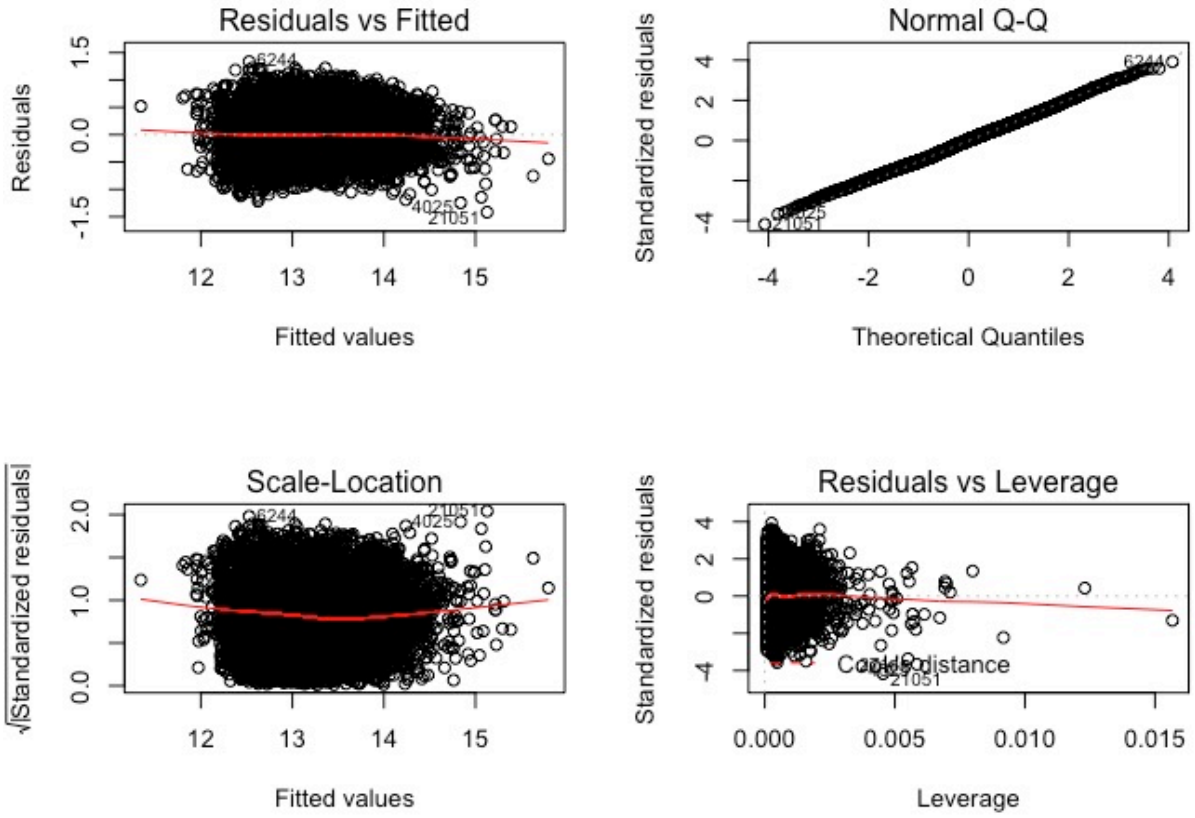
Thus

$$Y_{Price} = \exp(10.39 - 0.02248X_{Bedrooms} + 0.1888X_{Grade} + 0.09078X_{View} + 0.1278 \times \log(X_{sqft-living}) + 1.306\times10^{-4}X_{sqft-above} + 2.257\times10^{-4}X_{\text{sqft-basement}})$$

Since we put log on price to hold all of the assumptions for regression (a) independence (no mean trend) (b) normal distribution and (c) constant variance assumptions), the log of price has an average of 13.05, thus the intercept rise the value close to the number and all other coefficients are small.

Since all variables in this regression model are numeric, the coefficient for each variable means how much the price would vary if increase one unit of such variable. The p-value for each variable is less than the significant level of 0.01.

R-squared value is 0.5797. Even though using all variable would give an adjusted value around 0.7, but simplicity is also an important factor to consider. By previous analysis, we choose to use the interesting variables and then the adjusted R-squared value become 0.582. From this point, we tried to improve by add log to variables to narrow the distance between each point and the regression equation line. Although adding log to sqft_living improved r-squared, but log(price) lower the r-squared, however significantly improve the diagnostics test.



Without putting log on price, both normal distribution and constant variance assumptions would be violated. Therefore, the current r-squared value is sufficient for predicting the price without overfitting and too many variables.

Question 3: sampling, error and other topics (20 points)
When searched for products online (eg: Google shopping, Amazon), some users claim that the prices can change depending on their browsing data (such as their location, inferred through their IP addresses, how long they stay at a particular page product, how many times they rephrase their search queries, etc). You want to decide which, if any of such factors, affect the items' prices.

1. **(10 points) Detail how you would design an experiment to answer the above question. Clearly list: the research question, the population, how you sample, and the variables you collect.**

   The research question:

   Are the prices for online products in Amazon affected by the browsing data such as location, IP addresses, how long they stay at a particular page product, how many times they rephrase their search queries, etc? If yes, to what extend and which factors are more influential?

The population:

Customer who shop in Amazon in Austin.

How you sample

Set up a few research locations in Austin, and use different computers or devices to connect for different IP. Do experiments by different length of time and number of refresh times and collect what price they see. The length of time would have few categories for comparison (ex. 1min, 5min, 20min, 40min, 1h) and each categories applies to different location and devices. Similar for number of times for refresh. Several categories exists (eg, refresh 5, 10, 20, 40, 80 times). We have these four main factors that we want to research, and we would do experiments that generate data that controls variables. For example, for different location, we would have a subset of data that have same device, same number of refresh times, and same time length staying on the page. Although we change device and other two variable, but when comparing, we choose the columns while controlling all others except the one that we want to value if the variable affect the final price.

Variables you collect:

their location, IP addresses, how long they stay at the page product, and how manytimes rephrase their search queries

Assumptions needed to make your sample reasonable.

Assume deleting the browsing data would reset the status of the customers.
Assume IP address determined by the device in use
Assume the price setting does not change
Assume the testing time of the day or dates does not affect the price.

2. **(5 points) List at least TWO significant biases your experiment may have, and how your design mitigates them.**
    a) Although we do experiment in Austin only, but since we set up experiments in all kinds of place such as downtown area, UT area, east Austin, West Austin, South Austin, Round Rock, etc. Then th type of location changed, and we could see if urban area are given different price from suburban and rural areas.
    b) We focus on the online shopping in Amazon to narrow the range down. Although the research only applies to one website, Amazon is representative for online retail shopping. Thus the result of the experiment should be representative for general online shopping strategy.

A pollster surveyed 1500 people, and reported that 45% of them buy more than half of their household products online, while the rest buy the majority of their household products in shops.

1. **(3 points) Compute a 95% confidence interval for p0, the true proportion of Americans who buy the majority of their household products online. Do a hypothesis test for whether p0 is significantly different from 0.5. What is the meaning of the p-value of this test?**

   The Confidence interval for p0 is 42.43095% to 47.56905%.
   Set null hypothesis to be that p0 is no significantly different from 0.5. Then set significant level to be 0.01. The p-value we get is 0.9999504 > 0.01. The p-value show that the percentage p0 is not significantly different from 0.5.

2. **(2 points) Another pollster reported that their poll gives 47%, with a 95%-confidence interval of 42% to 52%. Why is it WRONG to say that "there is a 95% chance that between 42% and 49% of Americans buy the majority of their household products online"? Give the correct interpretation of this confidence interval.**

   Because the chance that Americans buy the majority of their household products online is fixed, a reality, so we cannot say this fixed number have a change to fall into this 95%-confidence interval.

   We can only say we are confident that the 95%-confidence interval contains the actual percent of Americans buy the majority online.