House Price Prediction

rene_hiroki

2019/06/10

Contents

1. Introduction

In this report, we will build a couple of machine learning models to predict house price. We use the dataset that is provided by Shree from Kaggle Datasets(https://www.kaggle.com/shree1992/housedata). You can download the dataset from this link, or you can also download from *here* GitHub repository. Let's glance at the dataset structure:

```
## Observations: 4,600
## Variables: 18
## $ date          <fct> 2014-05-02 00:00:00, 2014-05-02 00:00:00, 2014-0...
## $ price         <dbl> 313000, 2384000, 342000, 420000, 550000, 490000,...
## $ bedrooms      <dbl> 3, 5, 3, 3, 4, 2, 2, 4, 3, 4, 3, 4, 3, 3, 5, 3, ...
## $ bathrooms     <dbl> 1.50, 2.50, 2.00, 2.25, 2.50, 1.00, 2.00, 2.50, ...
## $ sqft_living   <int> 1340, 3650, 1930, 2000, 1940, 880, 1350, 2710, 2...
## $ sqft_lot      <int> 7912, 9050, 11947, 8030, 10500, 6380, 2560, 3586...
## $ floors        <dbl> 1.5, 2.0, 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.5...
## $ waterfront    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ view          <int> 0, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ condition     <int> 3, 5, 4, 4, 4, 3, 3, 3, 4, 3, 3, 5, 3, 4, 3, 4, ...
## $ sqft_above    <int> 1340, 3370, 1930, 1000, 1140, 880, 1350, 2710, 1...
## $ sqft_basement <int> 0, 280, 0, 1000, 800, 0, 0, 0, 860, 0, 0, 1010, ...
## $ yr_built      <int> 1955, 1921, 1966, 1963, 1976, 1938, 1976, 1989, ...
## $ yr_renovated  <int> 2005, 0, 0, 0, 1992, 1994, 0, 0, 0, 2010, 1994, ...
## $ street        <fct> 18810 Densmore Ave N, 709 W Blaine St, 26206-262...
## $ city          <fct> Shoreline, Seattle, Kent, Bellevue, Redmond, Sea...
## $ statezip      <fct> WA 98133, WA 98119, WA 98042, WA 98008, WA 98052...
## $ country       <fct> USA, USA, USA, USA, USA, USA, USA, USA, USA, USA...
```

We can see how many observations and variables are in the dataset, and also see what data types they are. Again, we are going to build machine learning models to predict price with other variables.

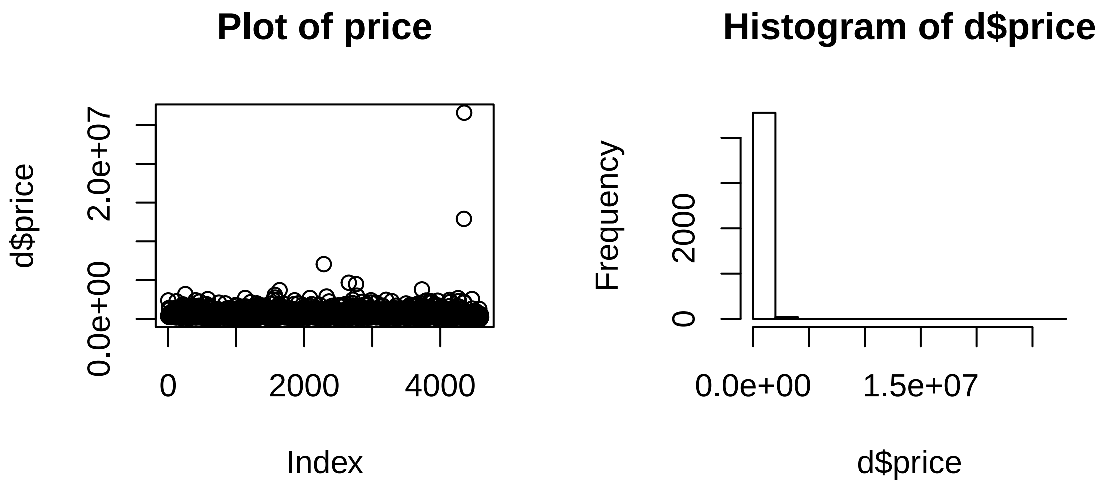
Our goal of this project is just predict house price with machine learning method. To evaluate our models, we define a loss function with RMSE. Thus, we should minimize RMSE as possible as we can.

Before moving on Analysis section, Exploratory Data Analysis(EDA) is coming next section. The more we understanding data, the more good models we can build. Because this task is regression, we try to build models with "multiple linear regression" and "random forests" in Analysis section. Then, We will evaluate our models in Result section and choose the best model in Conclusion section.

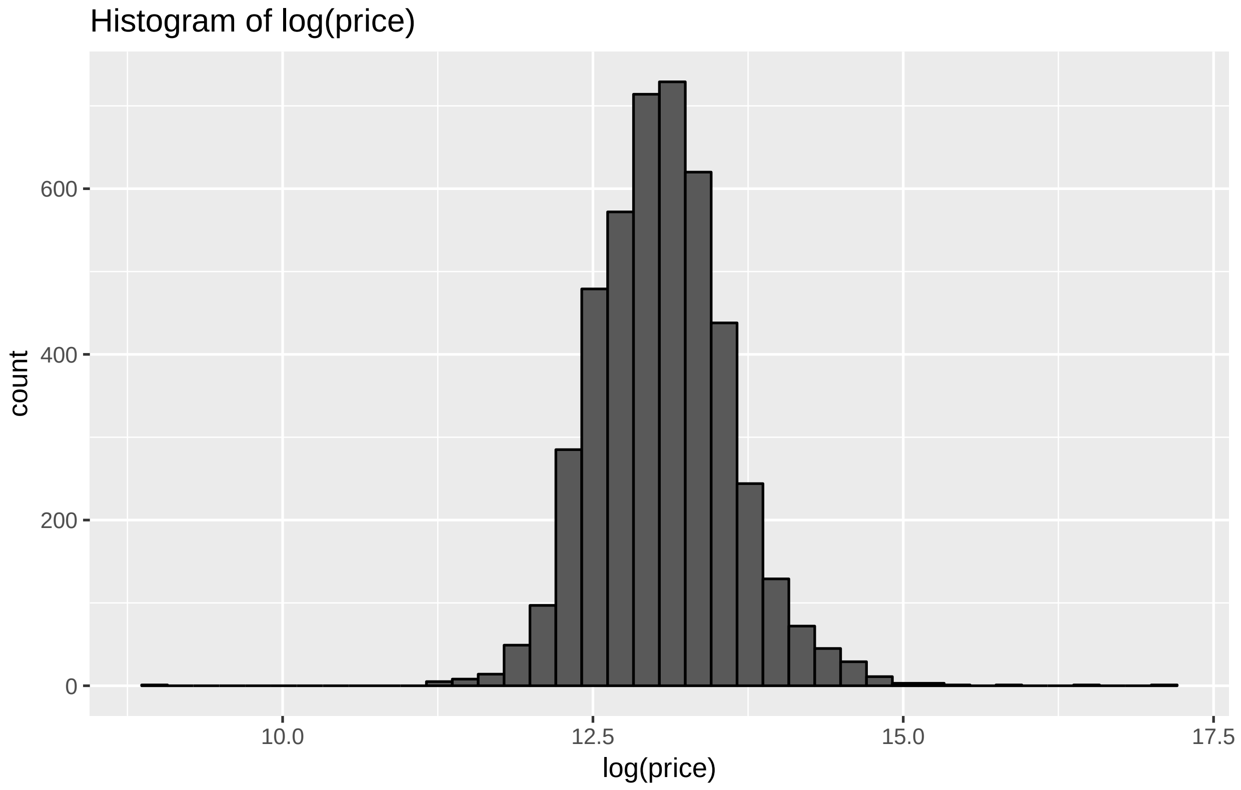
2. Exploratory Data Analysis

2.1 Response Variable

Our response variable is price. Let's look at distribution of price.



That's tough to see. Log transformation might help us.



That looks like a normal distribution. Thus, price might follow a log-normal distribution.

Then, let's see the top 5 and worst 5 prices.

top5	worst5
26590000	0
12899000	0
7062500	0
4668000	0
4489000	0

We can see 0s in worst 5 prices. Let's see the rows that price = 0.

```
## Observations: 49
## Variables: 18
## $ date          <fct> 2014-05-05 00:00:00, 2014-05-05 00:00:00, 2014-0...
## $ price         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ bedrooms      <dbl> 3, 4, 6, 5, 5, 4, 2, 4, 5, 5, 4, 4, 4, 5, 4, 5, ...
## $ bathrooms     <dbl> 1.75, 2.75, 2.75, 3.50, 1.50, 4.00, 2.50, 2.25, ...
## $ sqft_living   <int> 1490, 2600, 3200, 3480, 1500, 3680, 2200, 2170, ...
## $ sqft_lot      <int> 10125, 5390, 9200, 36615, 7112, 18804, 188200, 1...
## $ floors        <dbl> 1.0, 1.0, 1.0, 2.0, 1.0, 2.0, 1.0, 1.0, 2.0, 2.0...
## $ waterfront    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ view          <int> 0, 0, 2, 0, 0, 0, 3, 2, 0, 0, 2, 0, 0, 0, 4, 0, ...
## $ condition     <int> 4, 4, 4, 4, 5, 3, 3, 4, 3, 3, 3, 4, 3, 5, 3, 4, ...
## $ sqft_above    <int> 1490, 1300, 1600, 2490, 760, 3680, 2200, 1270, 3...
## $ sqft_basement <int> 0, 1300, 1600, 990, 740, 0, 0, 900, 1420, 0, 178...
## $ yr_built      <int> 1962, 1960, 1953, 1983, 1920, 1990, 2007, 1960, ...
## $ yr_renovated  <int> 0, 2001, 1983, 0, 0, 2009, 0, 2001, 0, 1923, 0, ...
## $ street        <fct> 3911 S 328th St, 2120 31st Ave W, 12271 Marine V...
## $ city          <fct> Federal Way, Seattle, Burien, Issaquah, Burien, ...
## $ statezip      <fct> WA 98001, WA 98199, WA 98146, WA 98075, WA 98166...
## $ country       <fct> USA, USA, USA, USA, USA, USA, USA, USA, USA, USA...
```

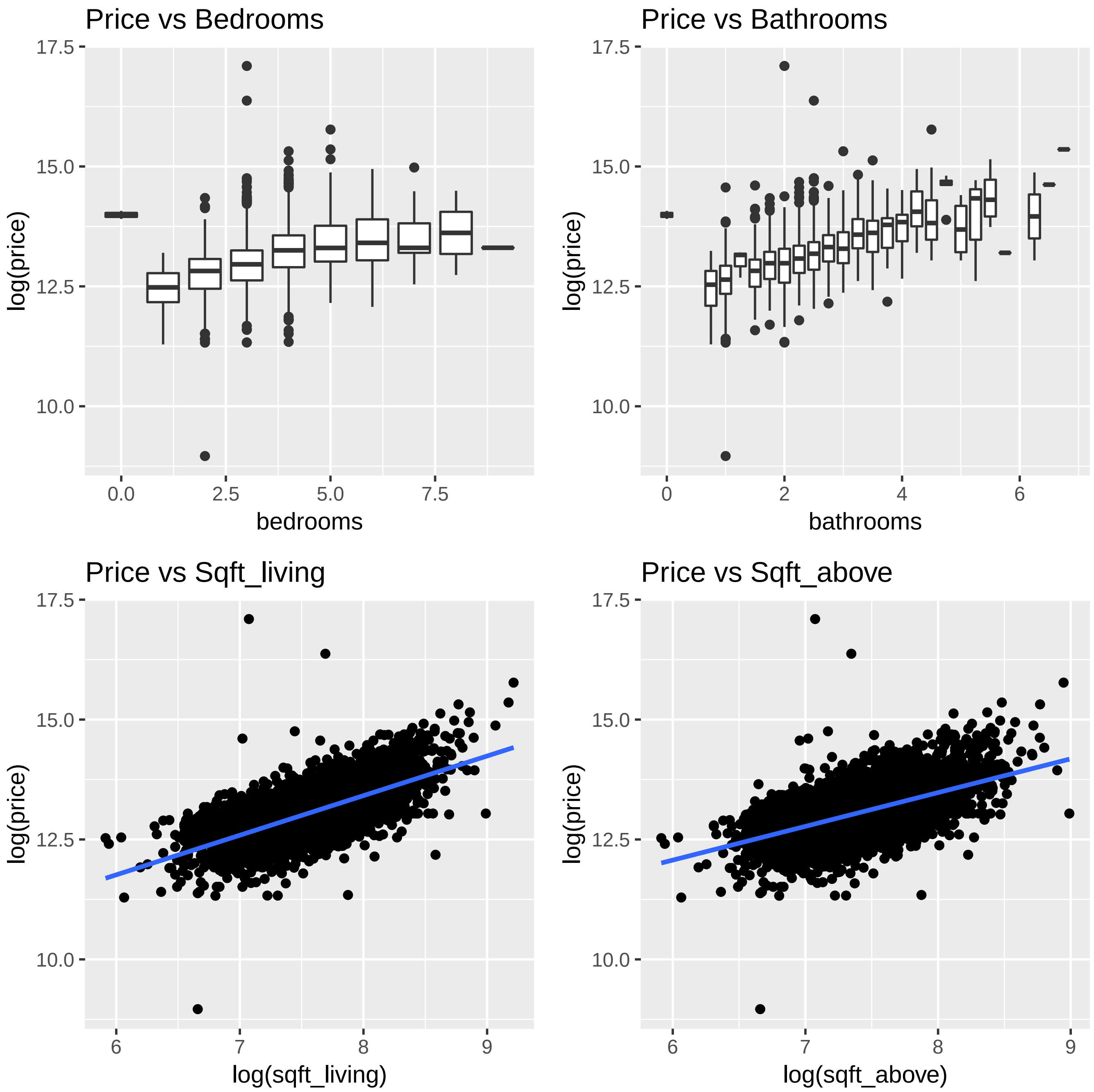
We can quickly know that there are 49 rows that price = 0, and these observations are not wrong. To improve our regression models, we replace these price 0 by median price 460943. Now, top 5 and worst 5 prices looks like this.

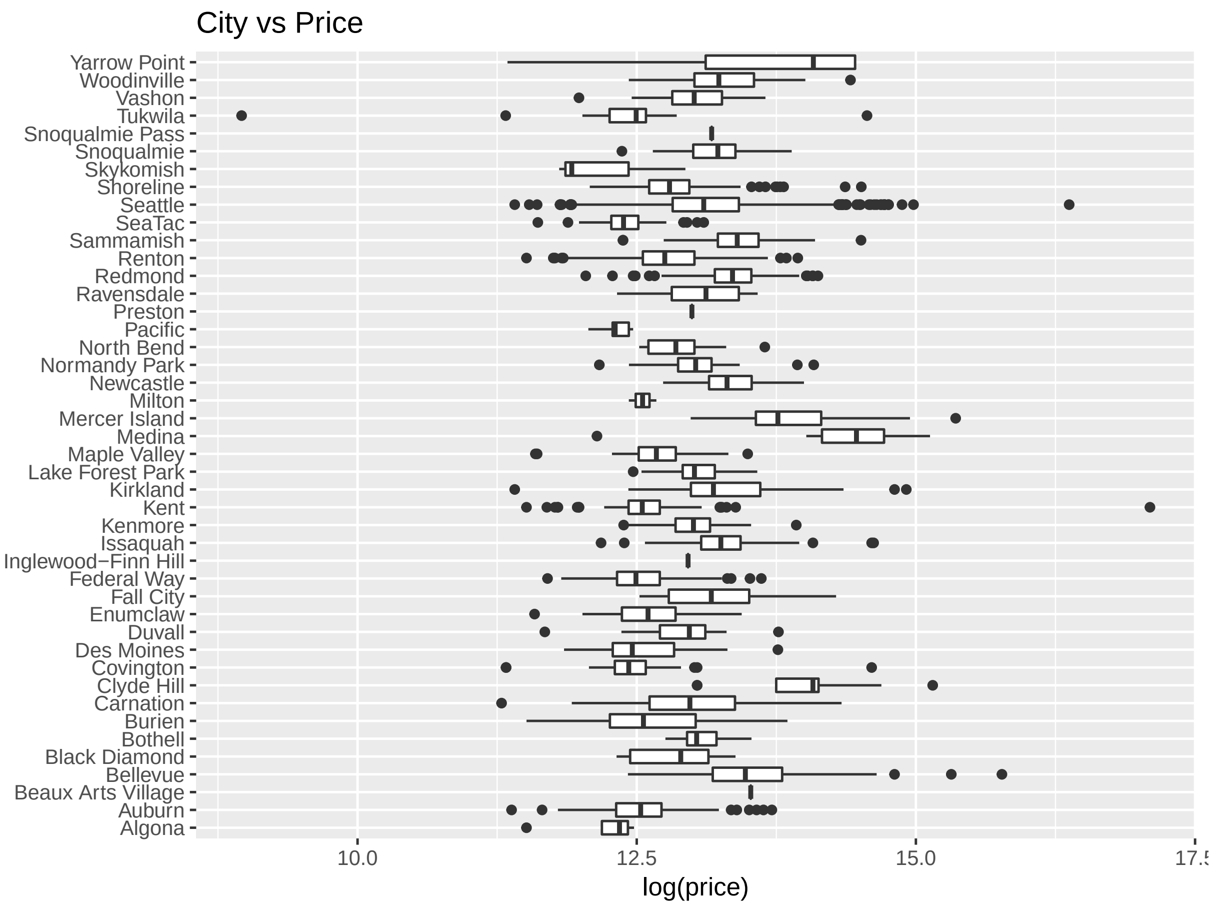
top5	worst5
26590000	83300
12899000	83300
7062500	83000
4668000	80000
4489000	7800

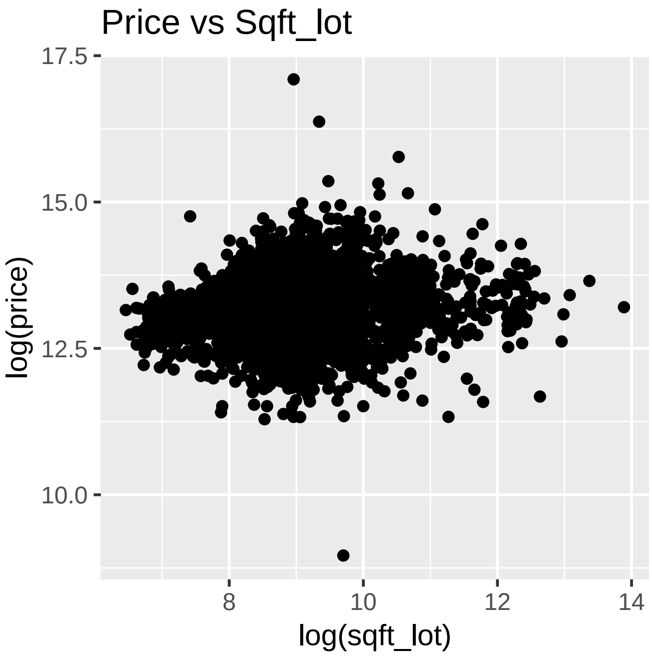
2.2 Predictor Variables

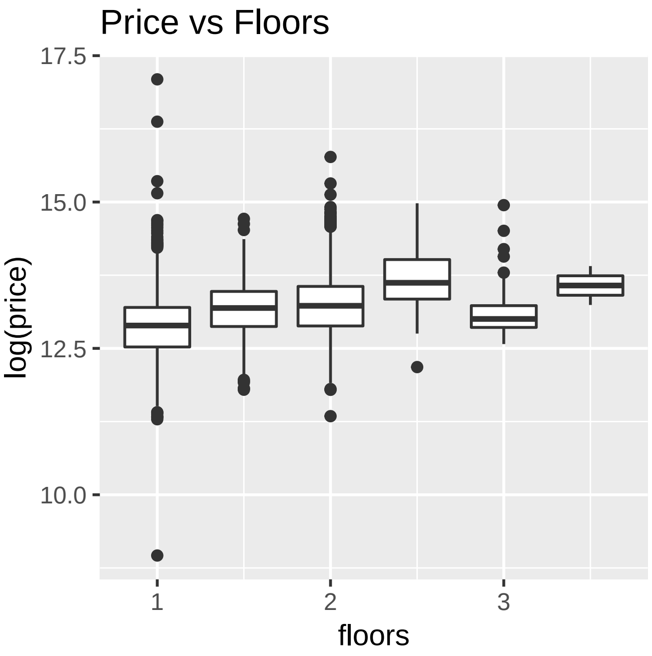
Before going forward EDA processes, we need to separate the dataset to trainset and testset. In this analysis, it is used that 90% of the dataset for the trainset and rest of dataset 10% is used for the testset.

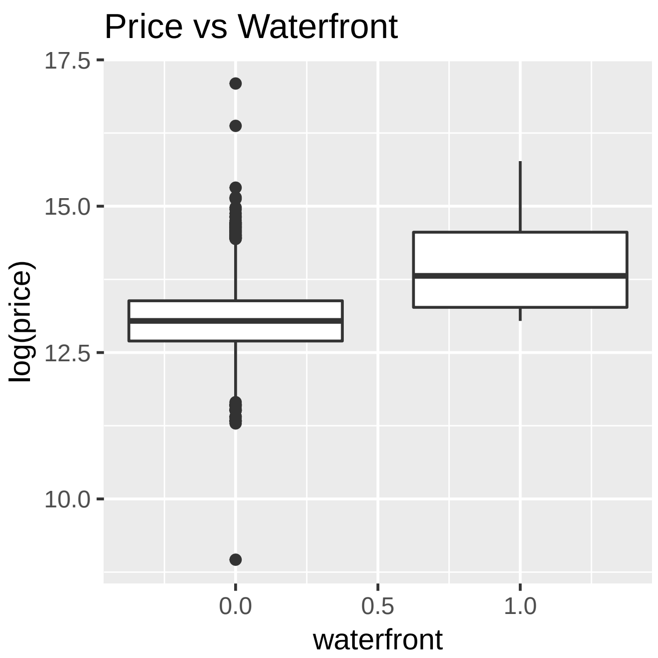
Then, we analyze relationships between response variable and predictor variables.

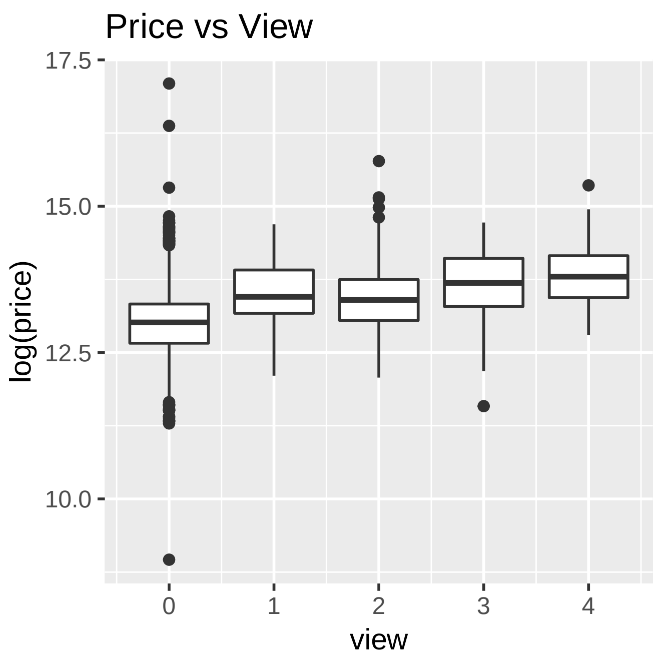


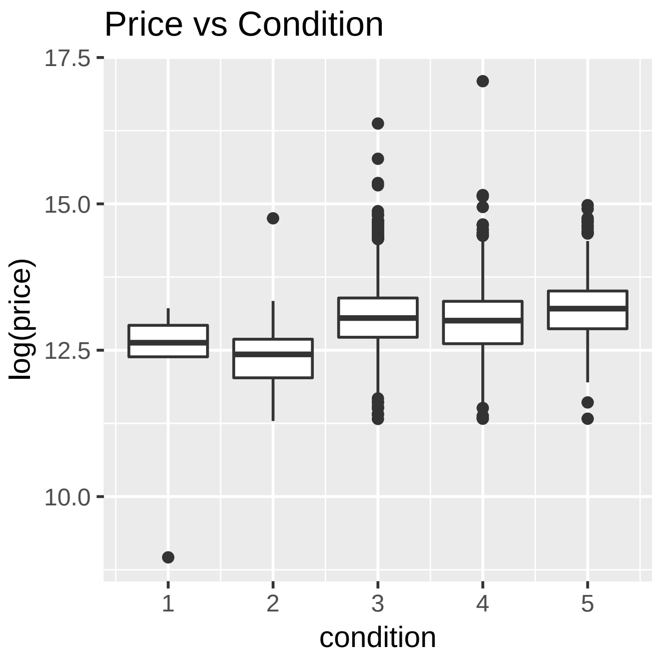

City vs Price
Yarrow Point
Woodinville
Vashon
Tukwila
Snoqualmie Pass
Snoqualmie
Skykomish
Shoreline
Seattle
SeaTac
Sammamish
Renton
Redmond
Ravensdale
Preston
Pacific
North Bend
Normandy Park
Newcastle
Milton
Mercer Island
Medina
Maple Valley
Lake Forest Park
Kirkland
Kent
Kenmore
Issaquah
Inglewood−Finn Hill
Federal Way
Fall City
Enumclaw
Duvall
Des Moines
Covington
Clyde Hill
Carnation
Burien
Bothell
Black Diamond
Bellevue
Beaux Arts Village
Auburn
Algona
10.0
12.5
15.0
17.5
log(price)

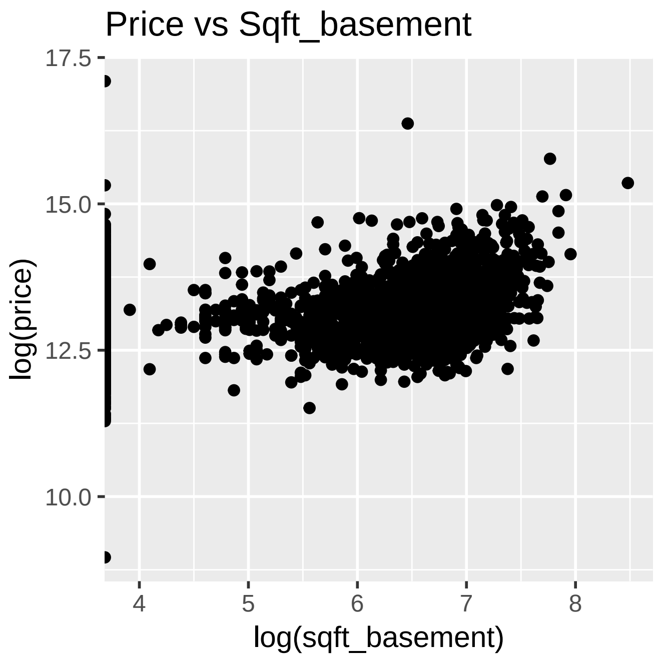

Price vs Sqft_lot
17.5
15.0
12.5
10.0
log(price)
8
10
12
14
log(sqft_lot)

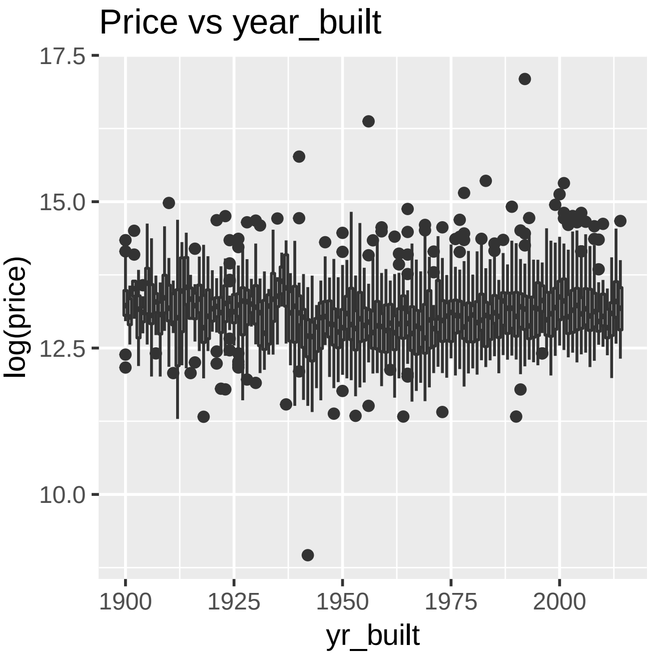

Price vs Floors
17.5
15.0
12.5
10.0
log(price)
1
2
3
floors


Price vs Waterfront
log(price)
17.5
15.0
12.5
10.0
0.0
0.5
1.0
waterfront

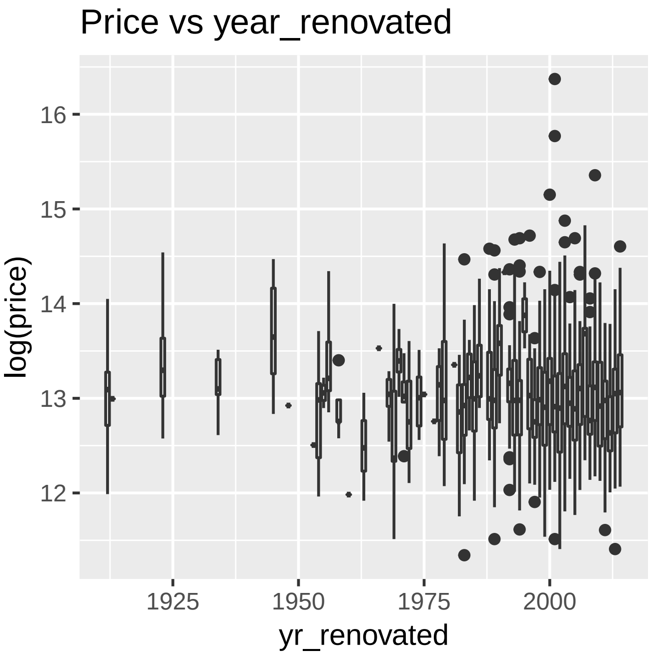

Price vs View
log(price)
17.5
15.0
12.5
10.0
0
1
2
3
4
view


Price vs Condition
log(price)
17.5
15.0
12.5
10.0
1
2
3
4
5
condition

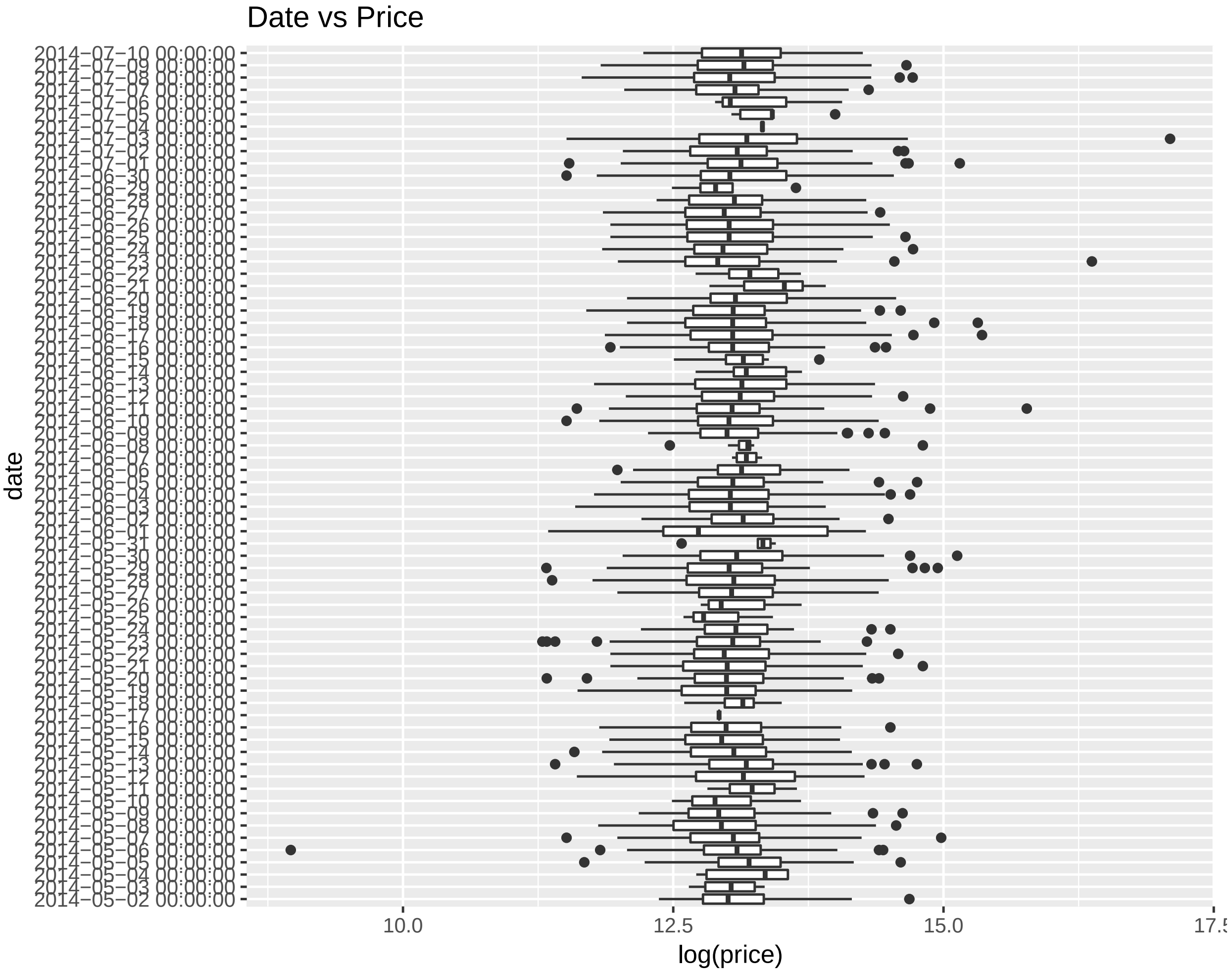

Price vs Sqft_basement
log(price)
17.5
15.0
12.5
10.0
4
5
6
7
8
log(sqft_basement)



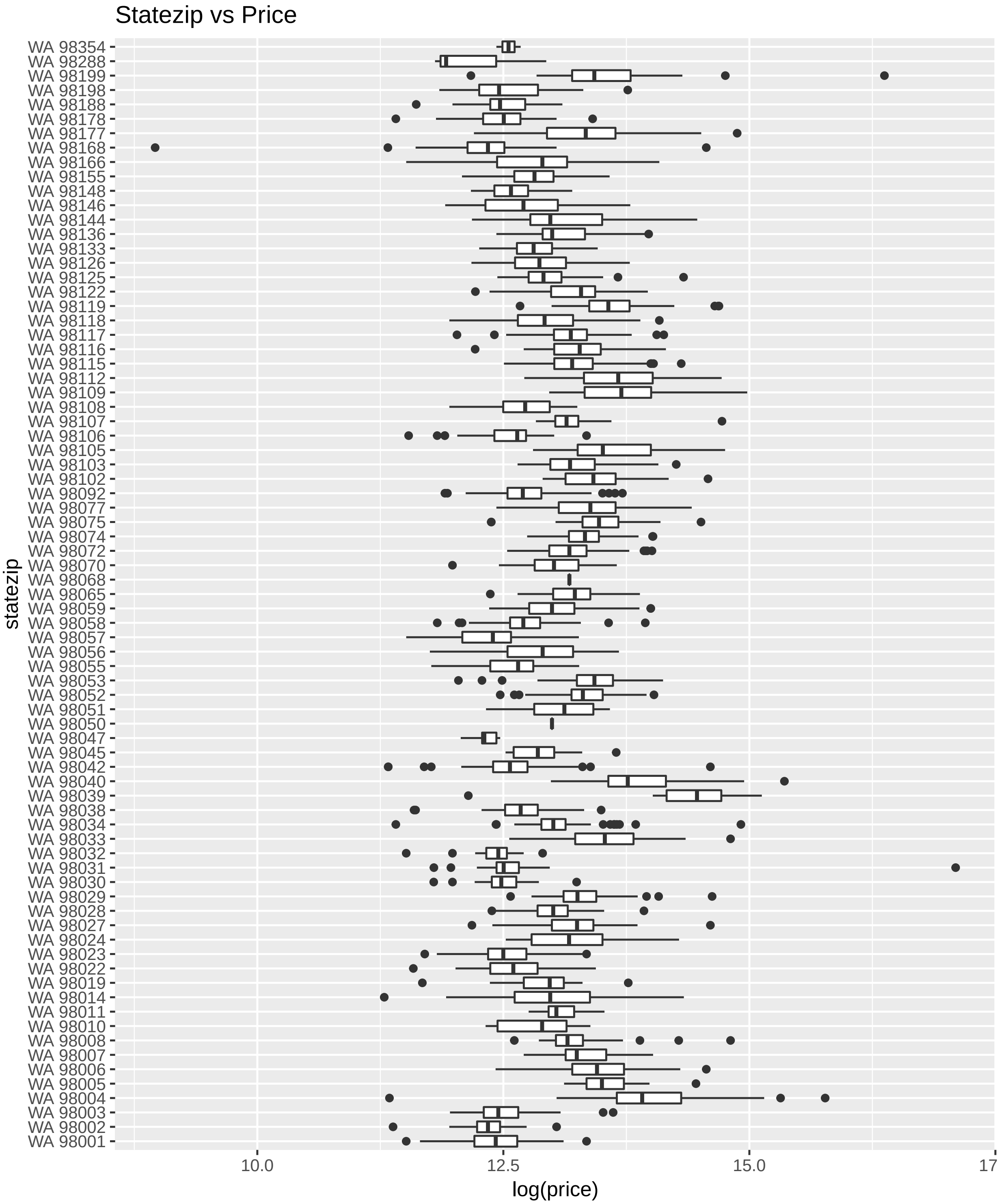
Price vs year_built



Price vs year_renovated



Date vs Price



From these graphs, we decide to use only 5 predictor variables, **bedrooms**, **bathrooms**, **sqft_living**, **sqft_above**, and **city**. These predictor variables might be able to predict price. On the other hand, rest of predictor variables looks like uncorrelated with price. (We can't use statezip for predictor variables because there too many categories.)

3. Analysis

3.1. Define loss function by RMSE

We use Root Mean Squared Error (RMSE) as a loss function. We define y_i as the price and denote our prediction with $\hat{y}_i$. The RMSE is then defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}(\hat{y}_i - y_i)^2}$$

3.2. Multiple Linear Regression

We can build multiple linear regression model with all predictor variables except city, because city is a categorical variable. If there are correlation among the predictor variables, it is inevitable to be a multicollinearity. We want to avoid this. Before building a model, let' look at correlation.

```
##              price bedrooms bathrooms sqft_living sqft_above
## price       1.0000   0.2013    0.3240      0.4269     0.3612
## bedrooms    0.2013   1.0000    0.5500      0.5961     0.4868
## bathrooms   0.3240   0.5500    1.0000      0.7604     0.6908
## sqft_living 0.4269   0.5961    0.7604      1.0000     0.8757
## sqft_above  0.3612   0.4868    0.6908      0.8757     1.0000
```

You can see that some variables are correlated. This implies that we have to pay attention to multicollinieearity. And it's important step to do log transformation to price, sqft_living, and sqft_above. This transformation makes residuals distribution normal.

Finally, we can build several multiple linear models. These equations are represent our models:

model1: $log(Y_{price}) = \alpha + \beta_{living}log(x_1) + \epsilon$
model2: $log(Y_{price}) = \alpha + \beta_{living}log(x_1) + \beta_{bath}x_2 + \epsilon$
model3: $log(Y_{price}) = \alpha + \beta_{living}log(x_1) + \beta_{bed}x_2 + \epsilon$
model4: $log(Y_{price}) = \alpha + \beta_{living}log(x_1) + \beta_{above}log(x_2) + \beta_{bath}x_3 + \epsilon$

Now we fit the model and then look the coefficients:

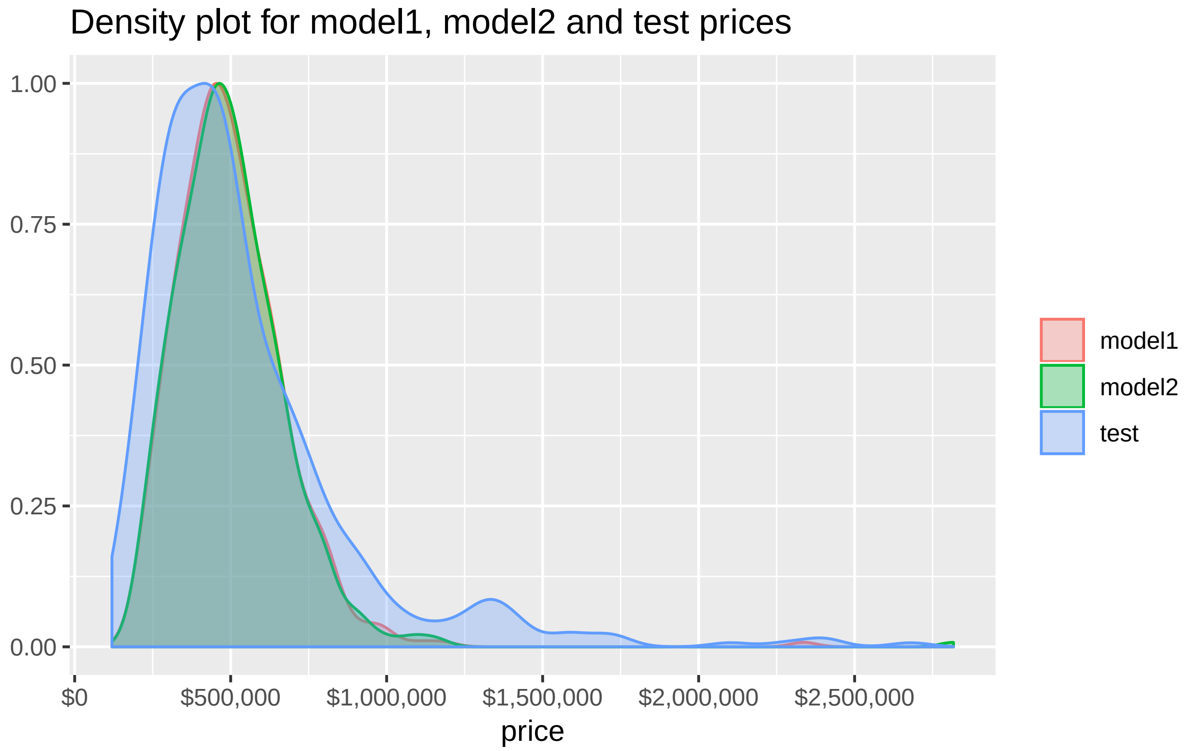
```
## (Intercept) sqft_living
##      6.8075      0.8259
```

```
## (Intercept) sqft_living   bathrooms
##     7.29732     0.74441     0.05899
```

```
## (Intercept) sqft_living    bedrooms
##     6.33369     0.91940    -0.06911
```

```
## (Intercept) sqft_living  sqft_above   bathrooms
##     7.32119     0.76011    -0.01956     0.06003
```

In model3, coefficient bedrooms is negative and also in model4 coefficient sqft_above is negative. These are opposite to our intuition. That is caused by multicollinearity so we don't use model3 and model4. Let's look at density plot for model1, model2, and testset and then we evaluate model_1 and model2 by RMSE with testset.



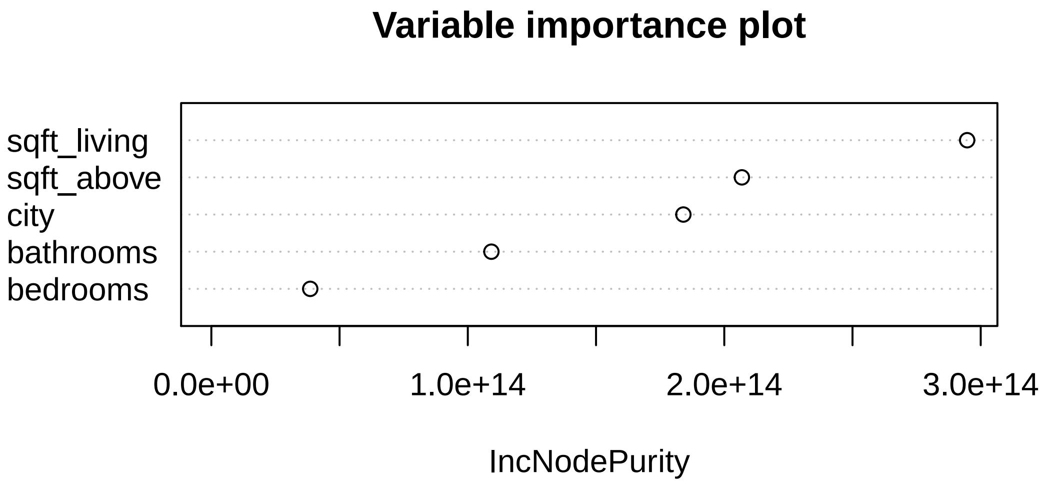
model1: RMSE = 264599

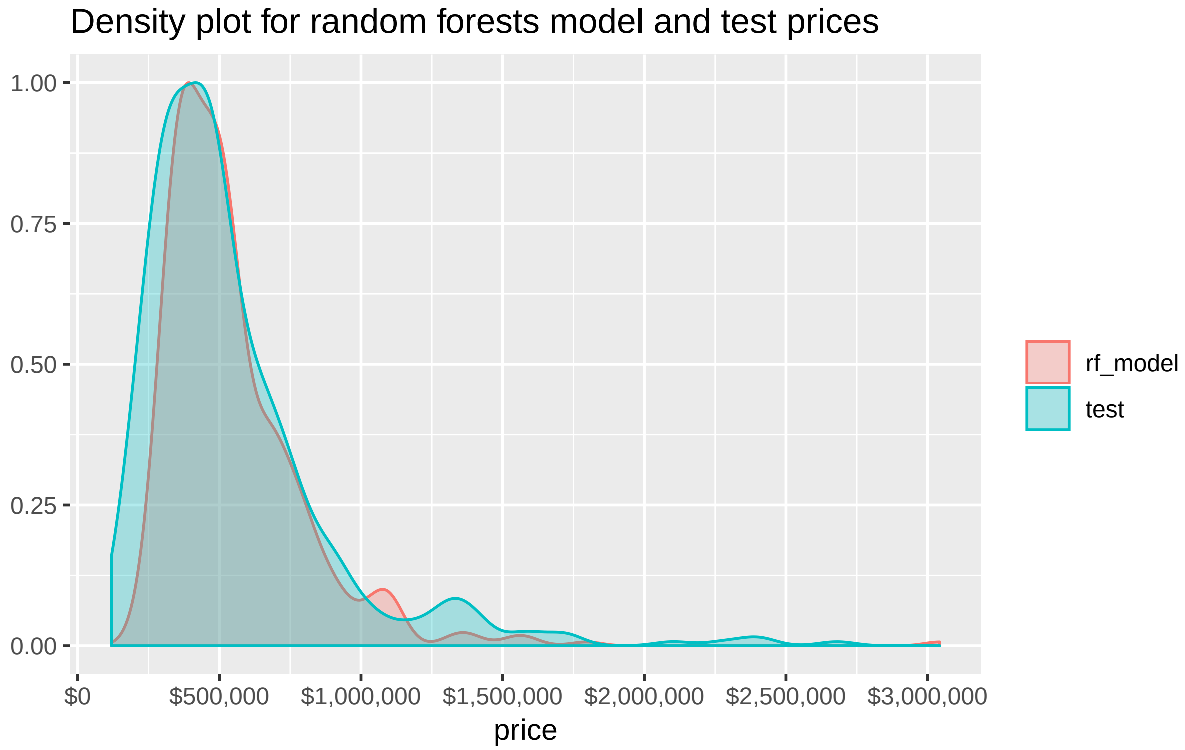
model2: RMSE = 264368

Although model1 and model2 are very close, model2 is slightly better than model1.

3.3. Random forests

We already to ready for building another model by random forests. We use predictor variables **bedrooms**, **bathrooms**, **sqft__living**, **sqft__above**, and **city**. The difference to the linear model is that city, categorical variable, is included in predictor variables. In section2 EDA, we saw that city has various median prices. This means that city is very useful to predict. Let's build a random forests model.





rf_model: RMSE = 201394

From variable importance plot, as we expected that city contributed to predict price. Density plot looks more better than linear models, and RMSE is minimized.

4. Results

model	RMSE
linear model1	264599
linear model2	264368
random forests model	201394

We built three models with multiple linear regression and random forests. The best model is random forests model with RMSE = 201394.

5. Conclusion

We can predict house price with random forests model with average error about $200,000. Sqft_living has the largest effect on house price. That is follow our intuition. And city has also some effect on price. This is why random forests model is better than linear model. Linear model can't use categorical variables(not 0 or 1). Random forests is so useful.