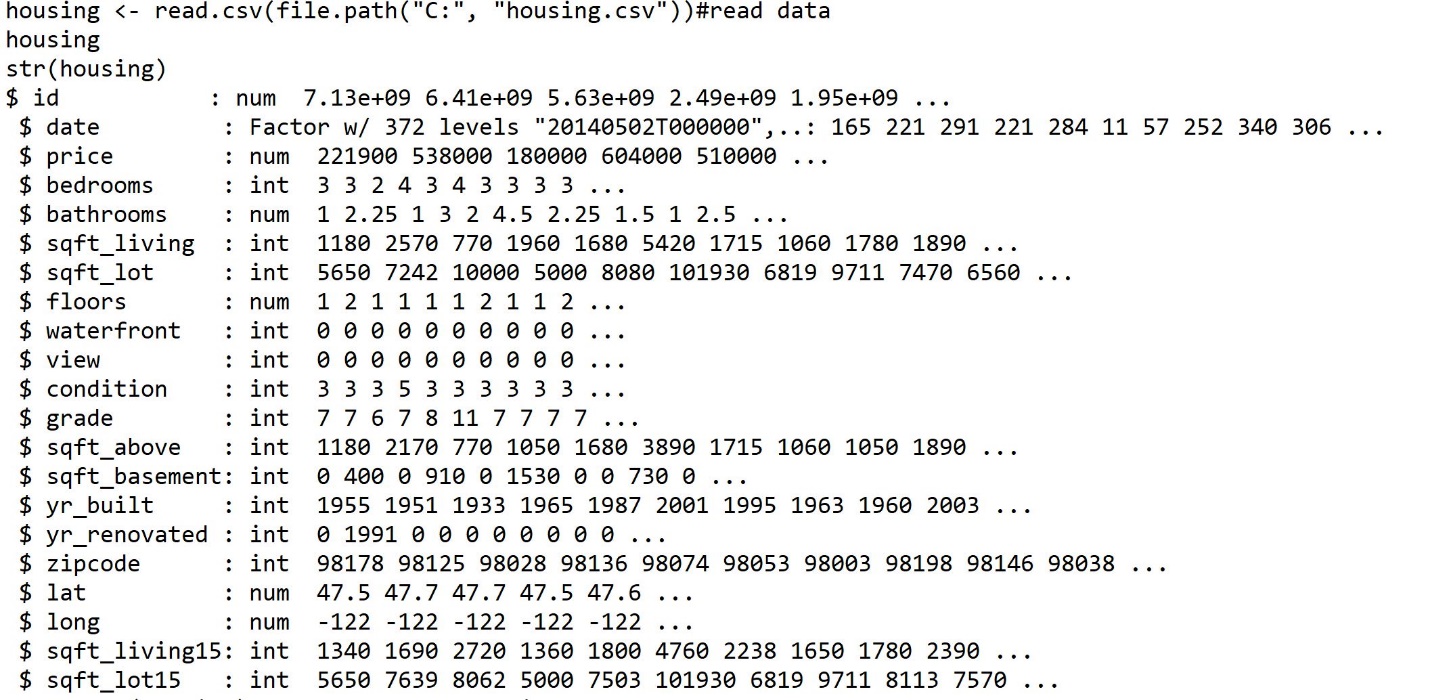
**Data visualizations and Price prediction of Houses Sold in King County:**

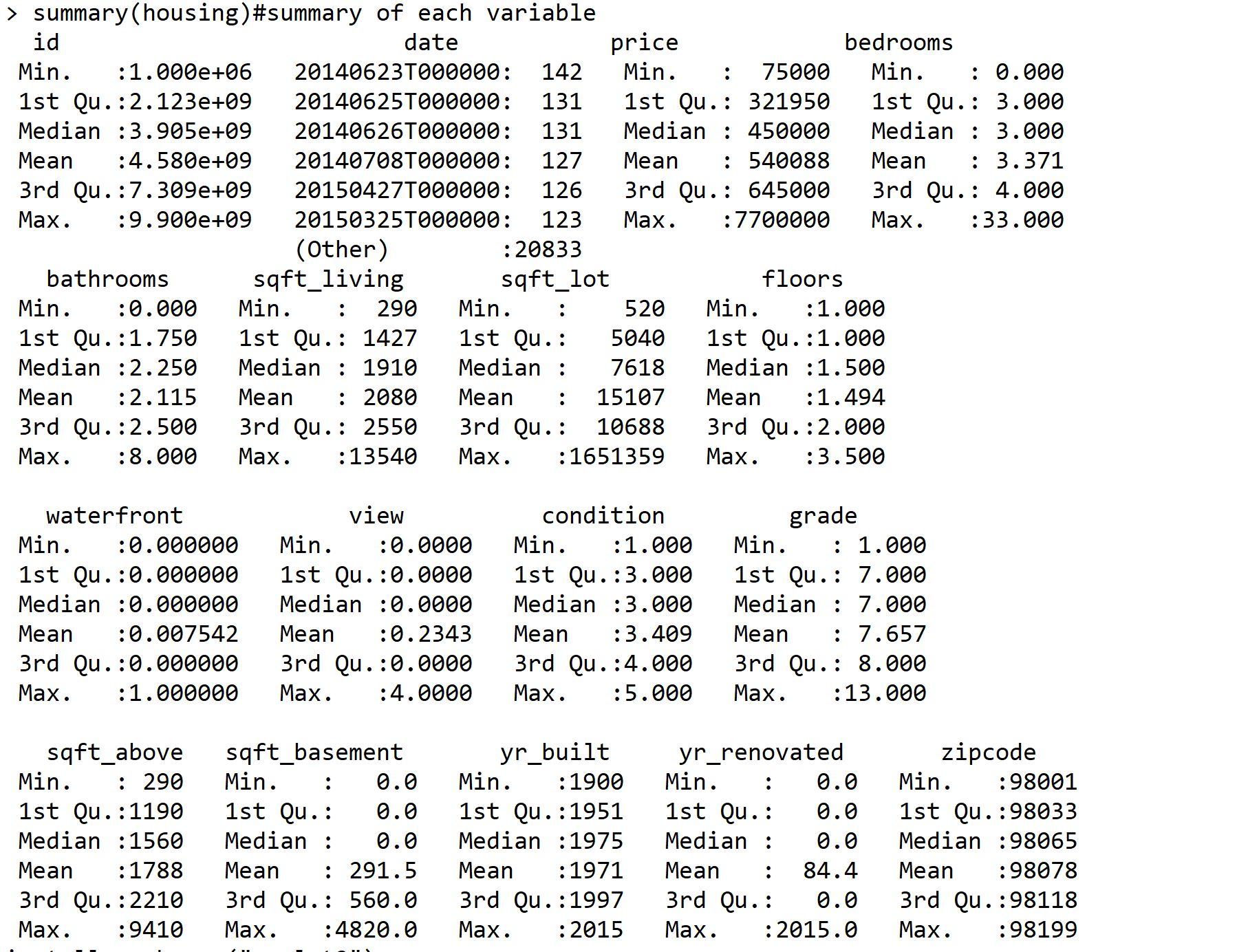
**About this Dataset:**

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.



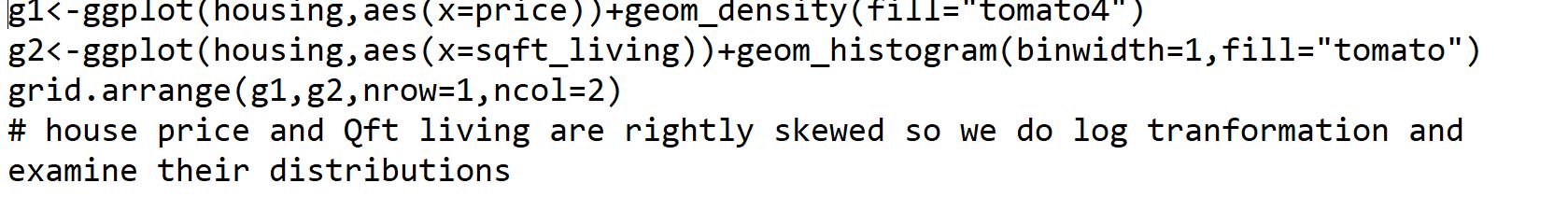
**summary(housing):**

#summary of each variable



**Exploratory Data Analysis:**

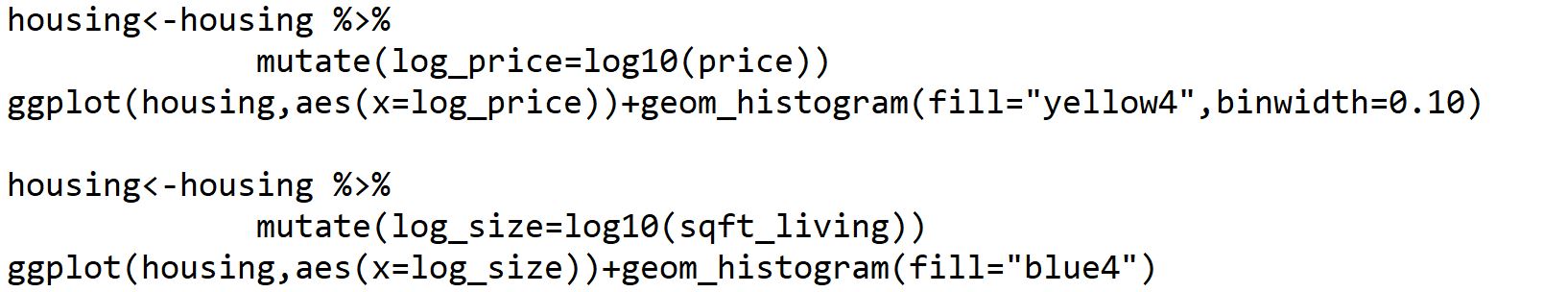
****

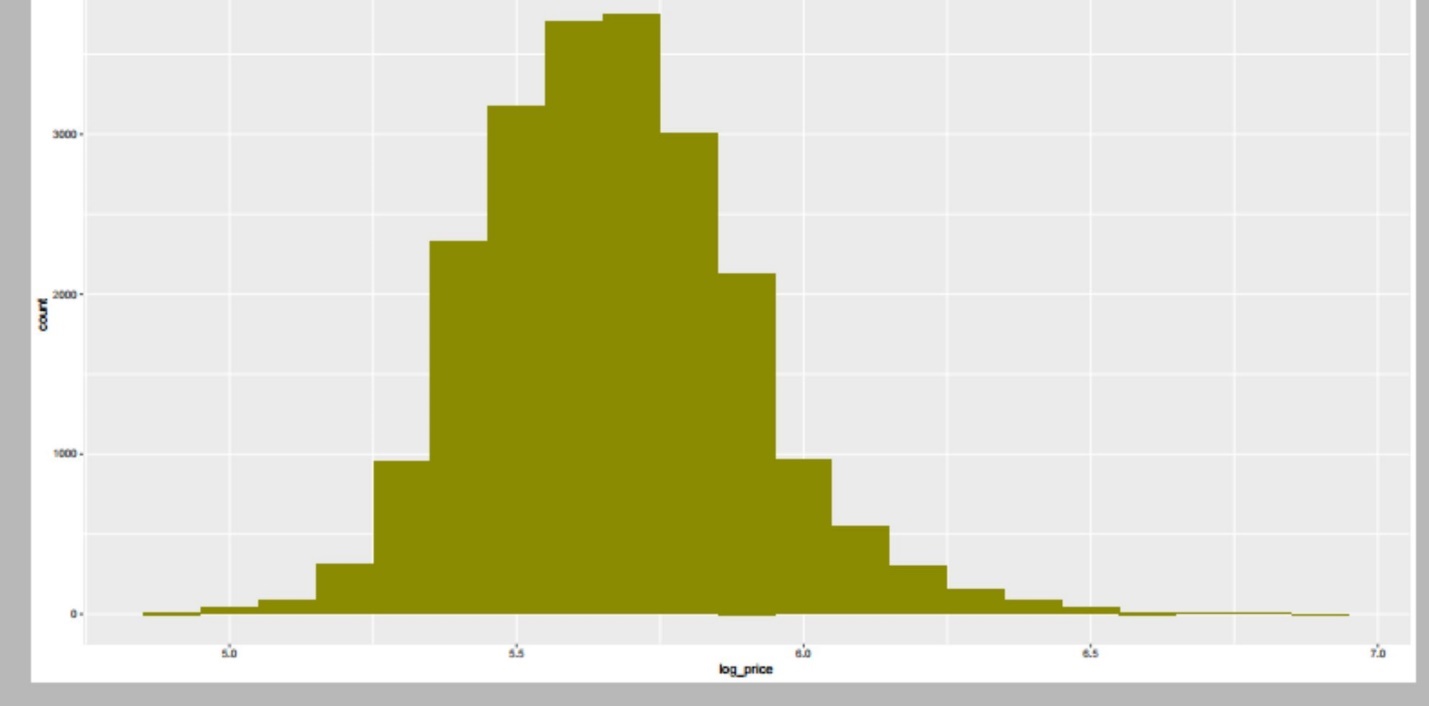
****

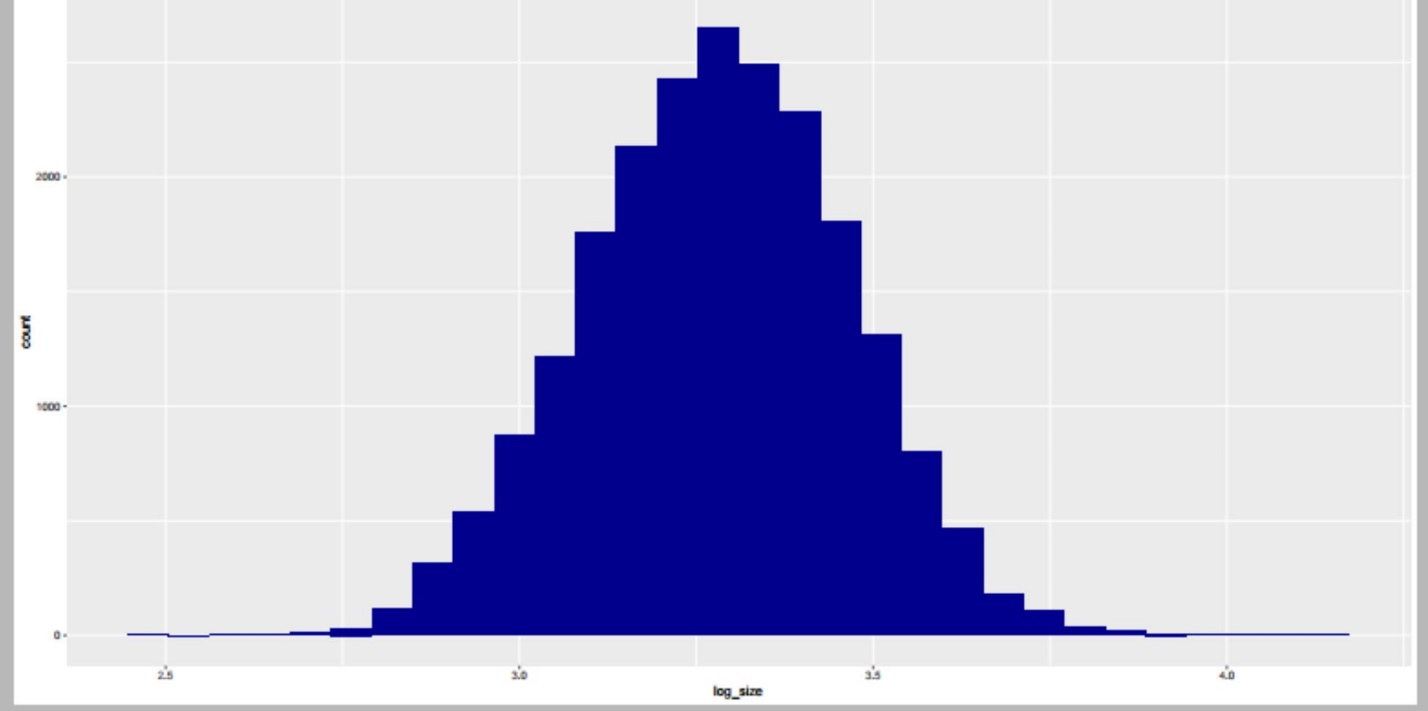


House Price Distribution:

Distribution of house prices & Sqft living was right skewed, so lets apply log() and then plot the distribution

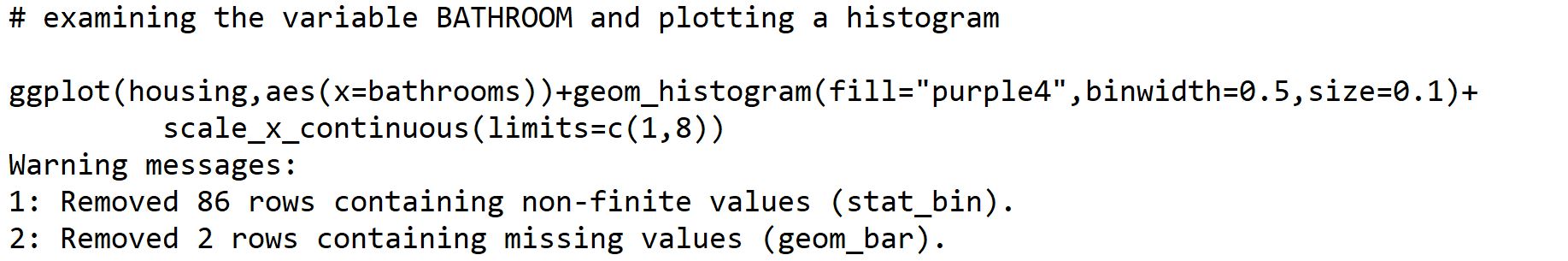


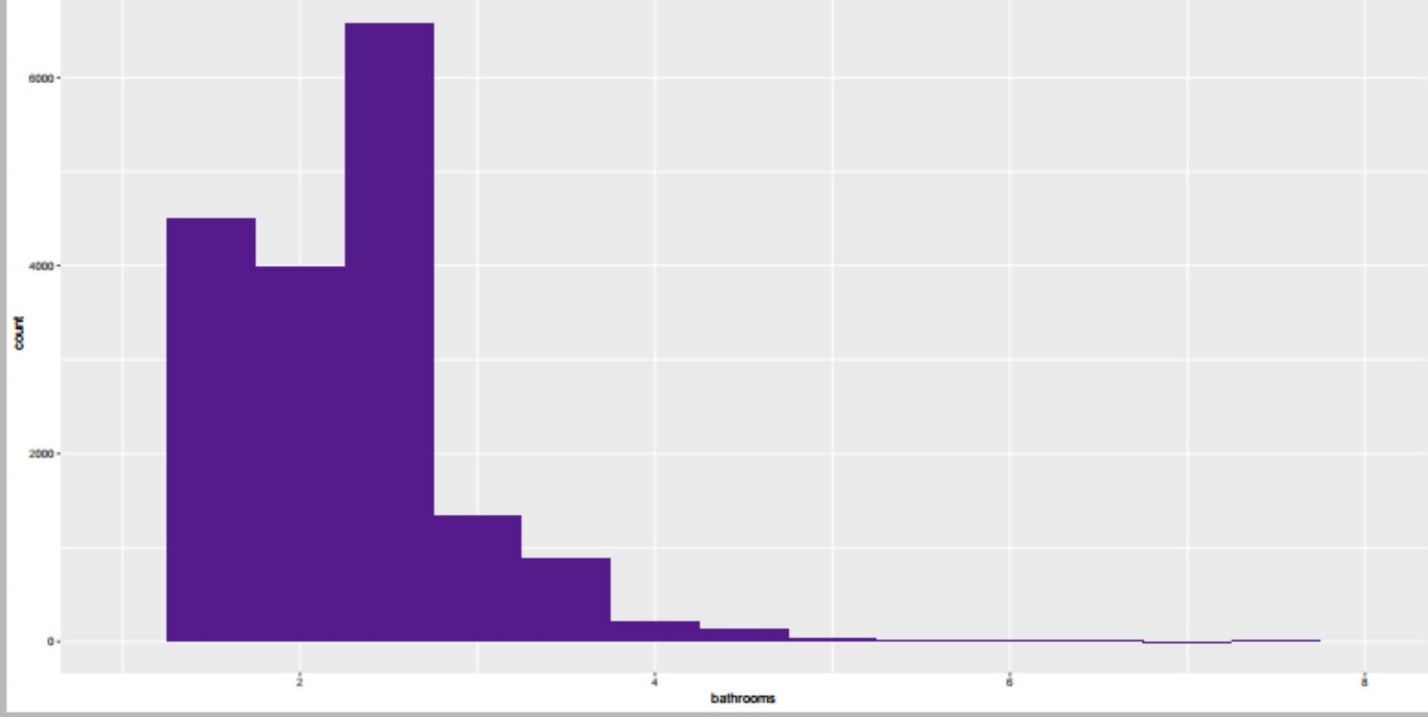




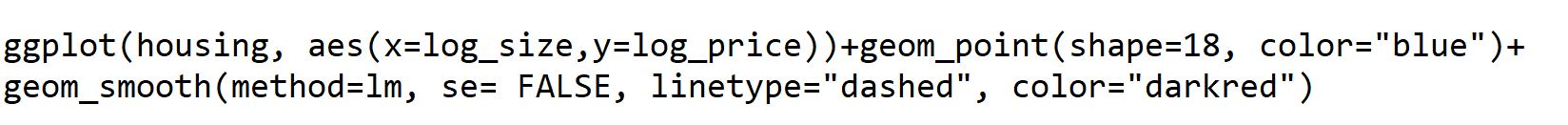
Most of house price lies between 5.4 to 6 million.

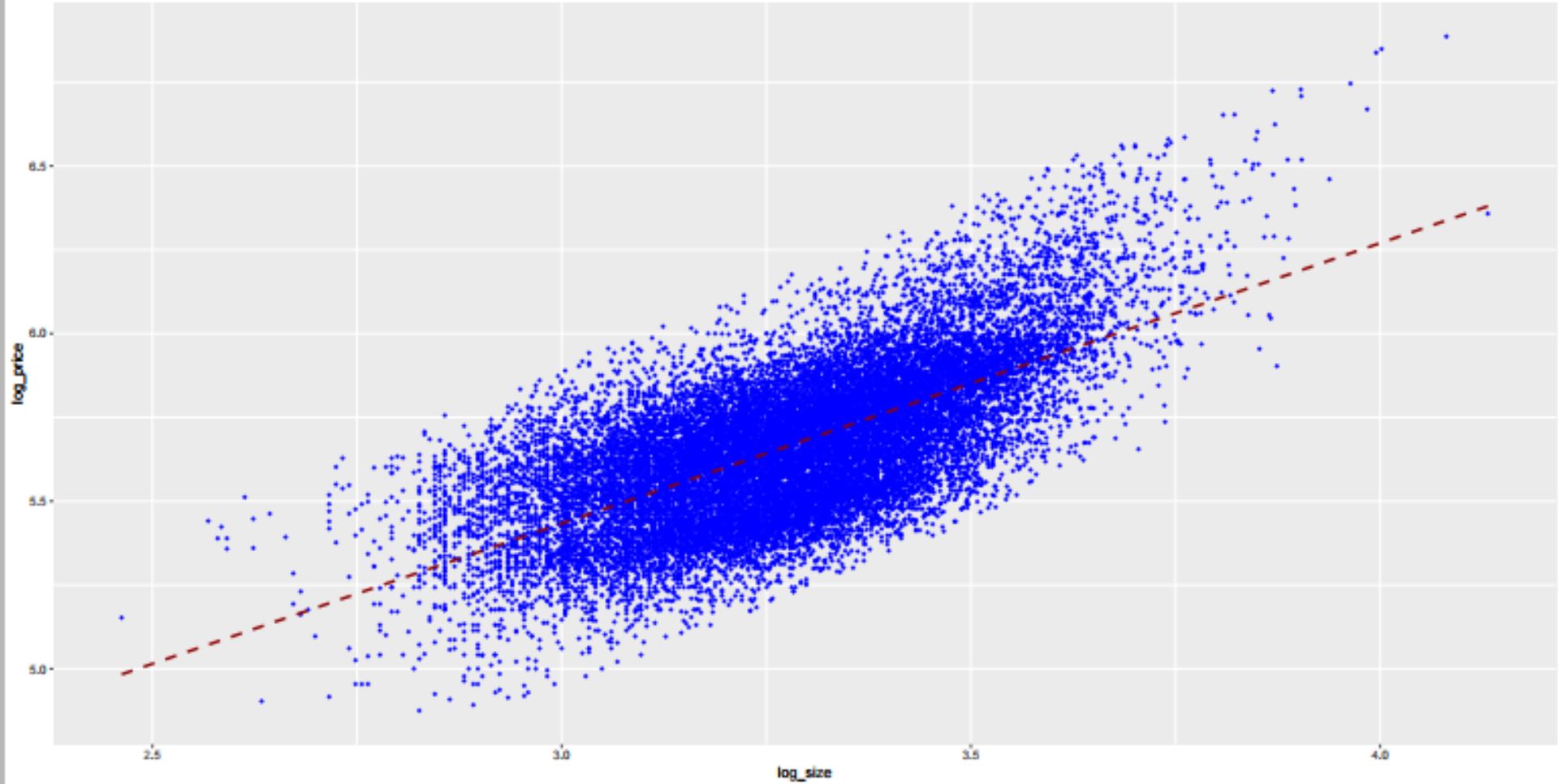
**Bathrooms:**



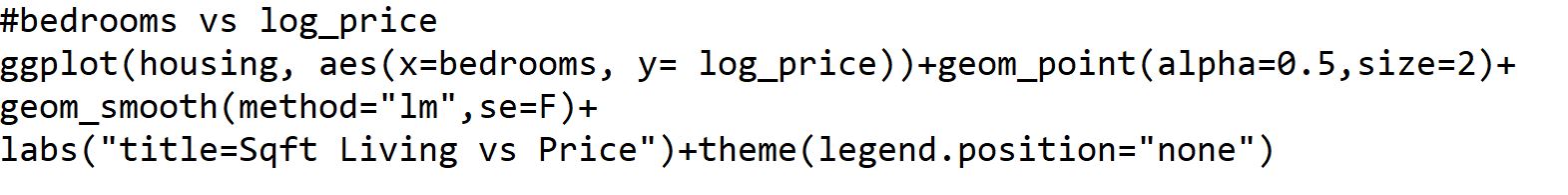


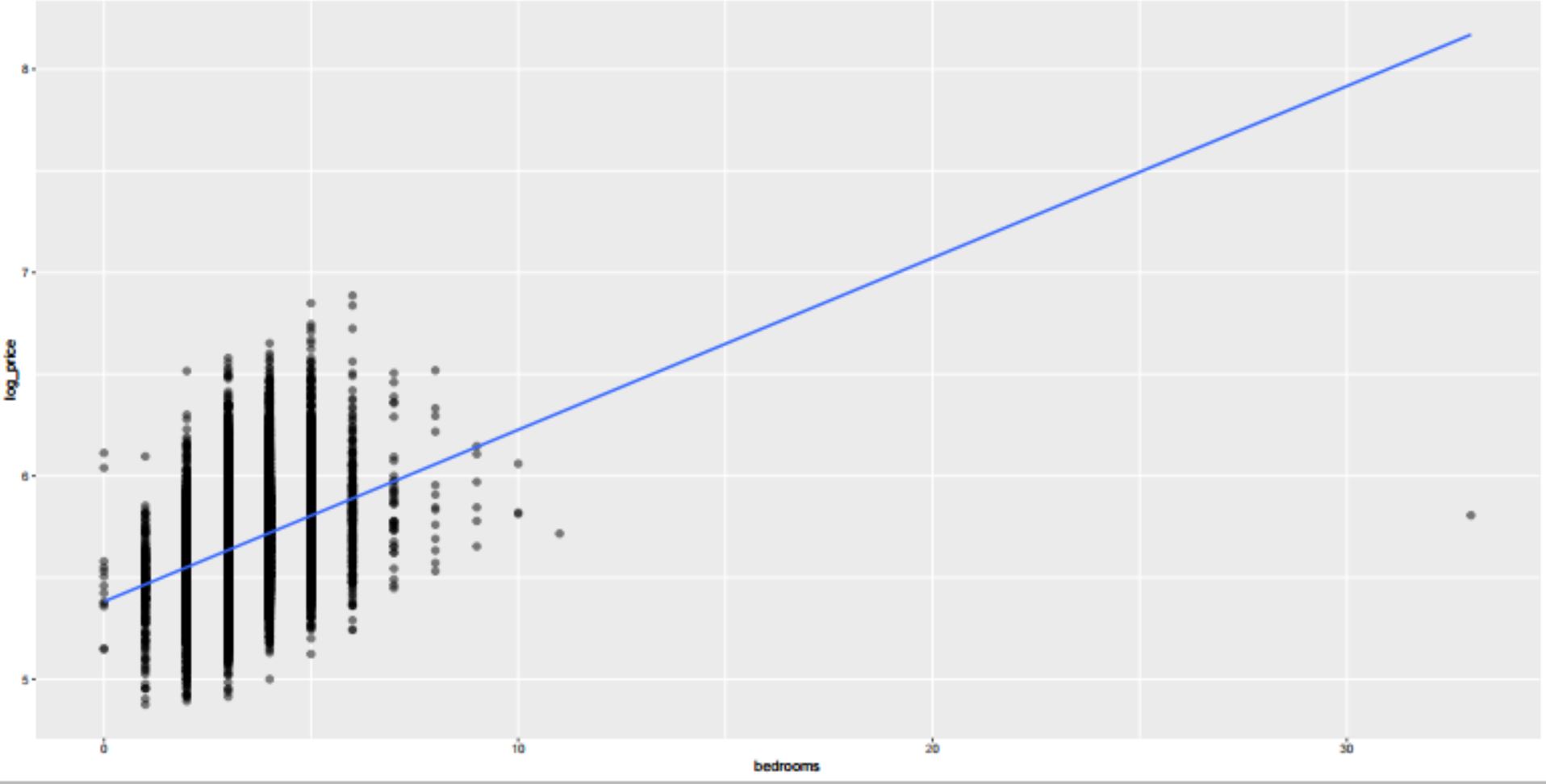
**Examining the relationship between log\_price and log\_size:**



****

**Relationship between bedrooms and log\_price:**

****

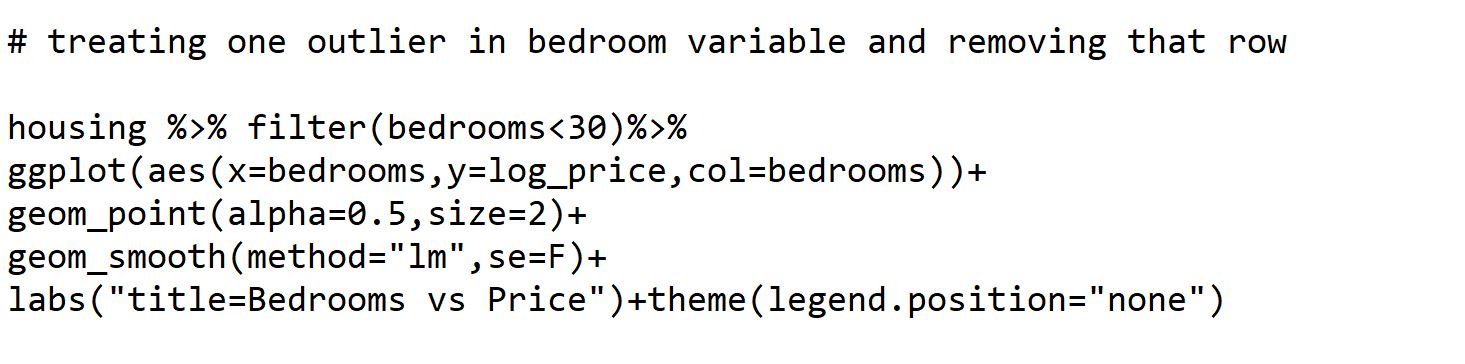
****

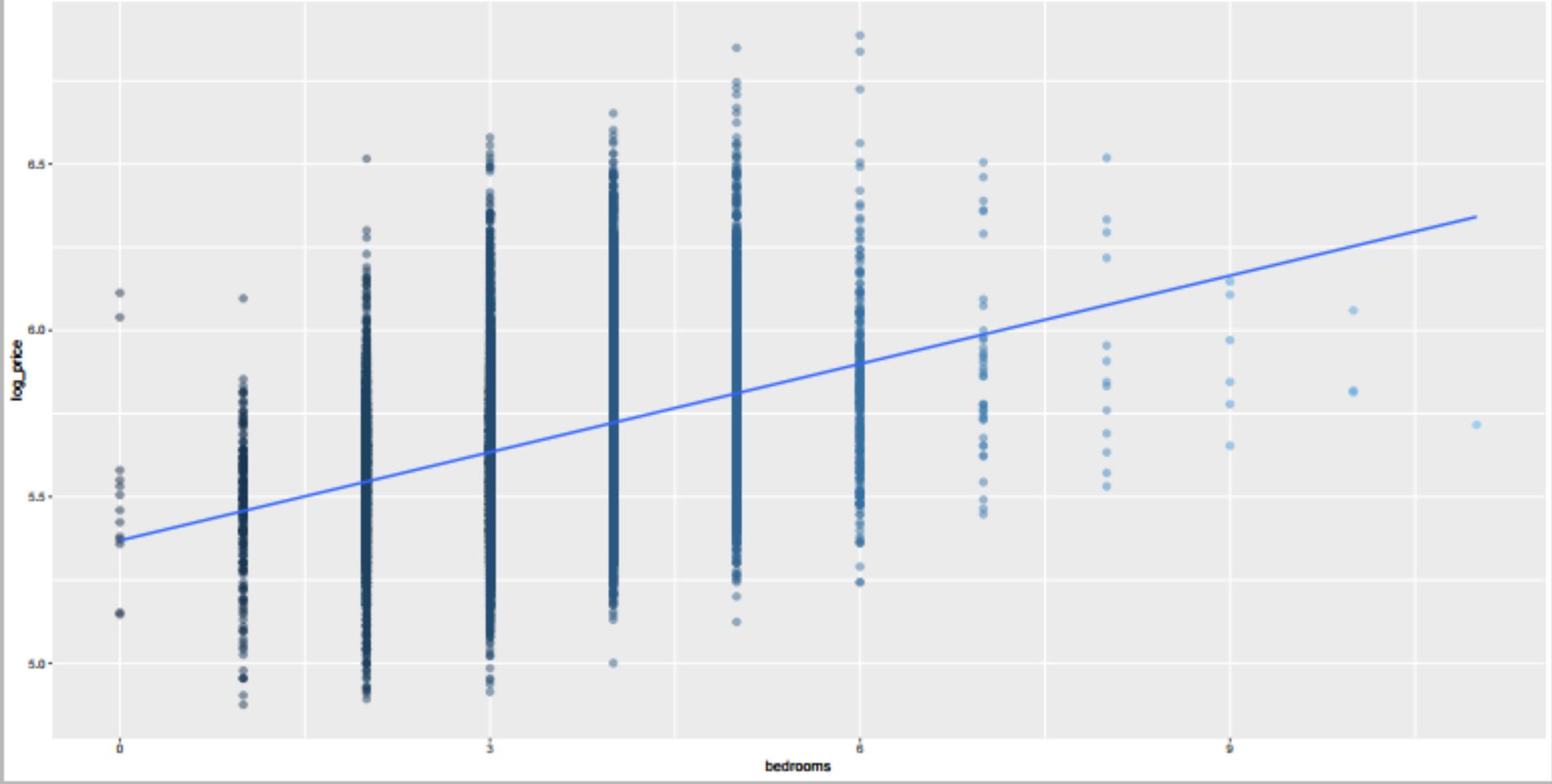
There exists one outlier in the bedrooms variable, which shows house with 33 bedrooms, which may not be the case, so let’s remove that row and plot the graph again.

5, 6-bedroom house price seems to high

**Treating an outlier in variable bedrooms:**

There exists one outlier in the bedrooms variable, which shows house with 33 bedrooms, which may not be the case, so let’s remove that row. and plot the graph again. 5, 6-bedroom house price seems to high.

****



**House condition and prices:**

table(housing$condition)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1  30 | 2  172 | 3  14031 | 4  5679 | 5  1701 |

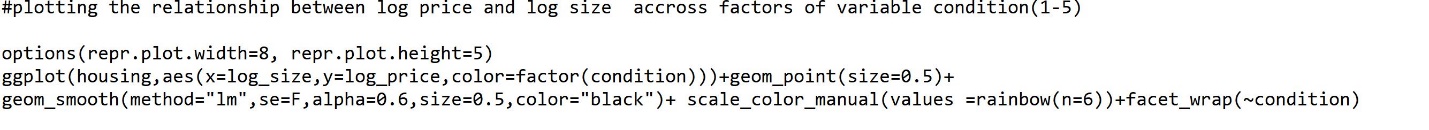
Number 1 being the worst and 5 being the best condition house and most of the houses are of condition 3 (14031)

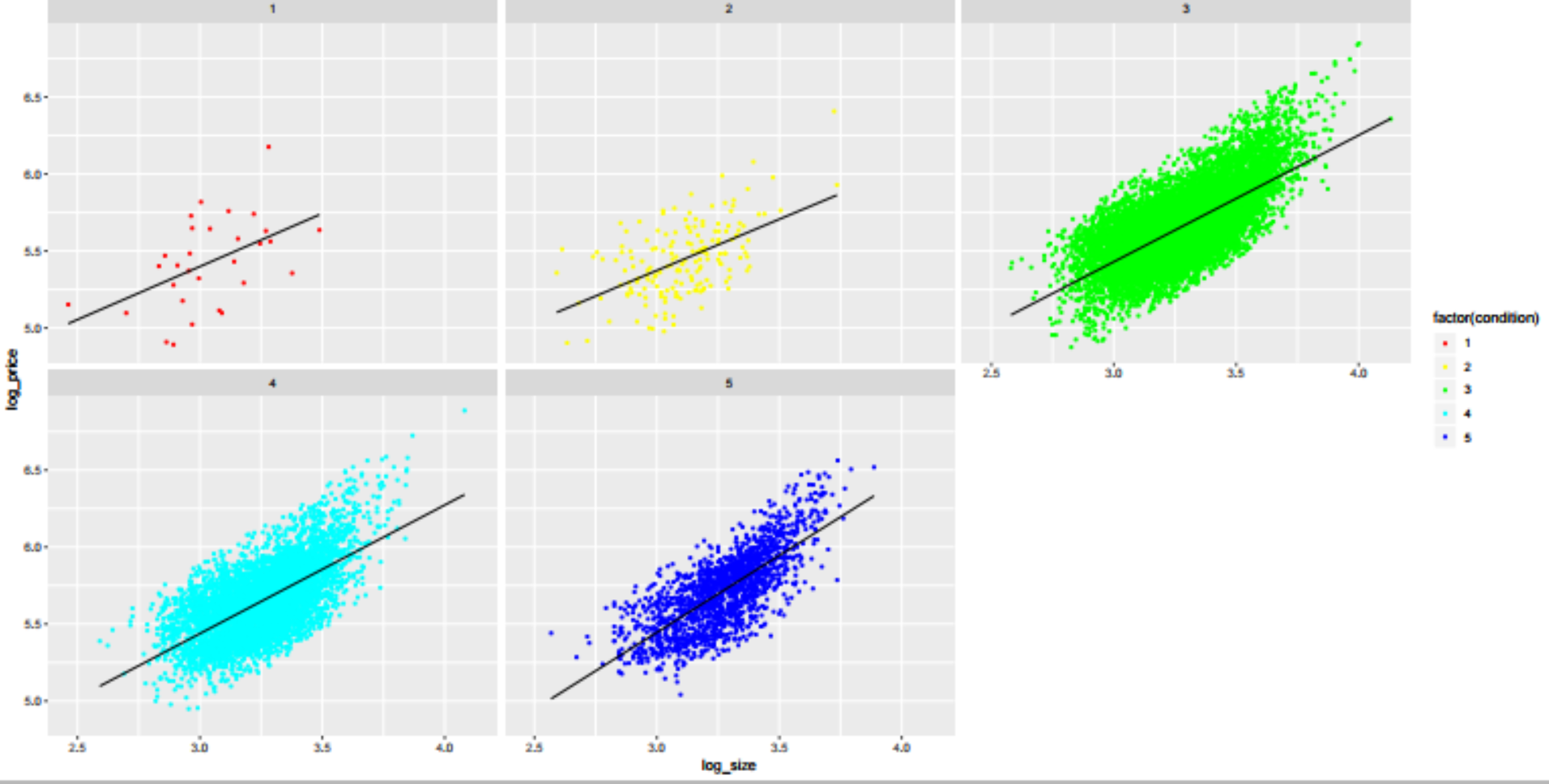
**creating a table of relative mean prices of house according to their conditions:**

housing %>%group\_by(factor(condition))%>%summarise(mean\_price=mean(log\_price),sd=sd(log\_price),count=n())

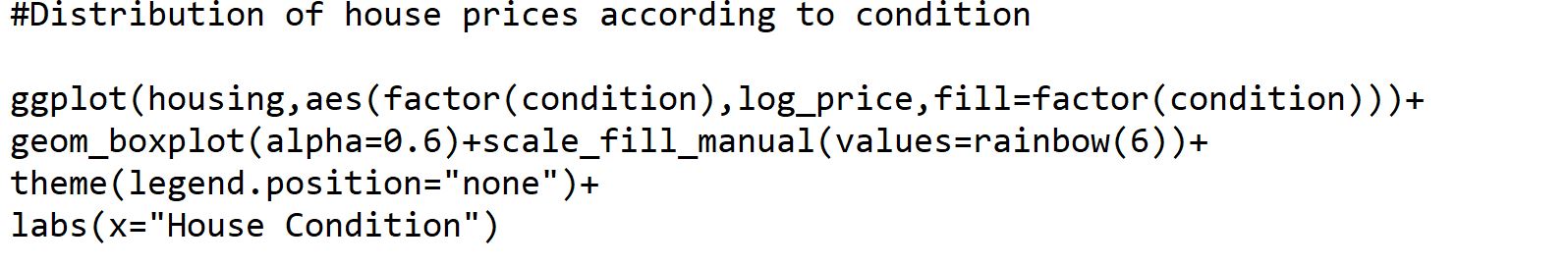
|  |  |  |  |
| --- | --- | --- | --- |
| Factor  condition | Mean\_price | sd | count |
| 1  2  3  4  5 | 5.42  5.45  5.76  5.65  5.71 | 0.293  0.233  0.224  0.228  0.244 | 30  172  14031  5679  1704 |

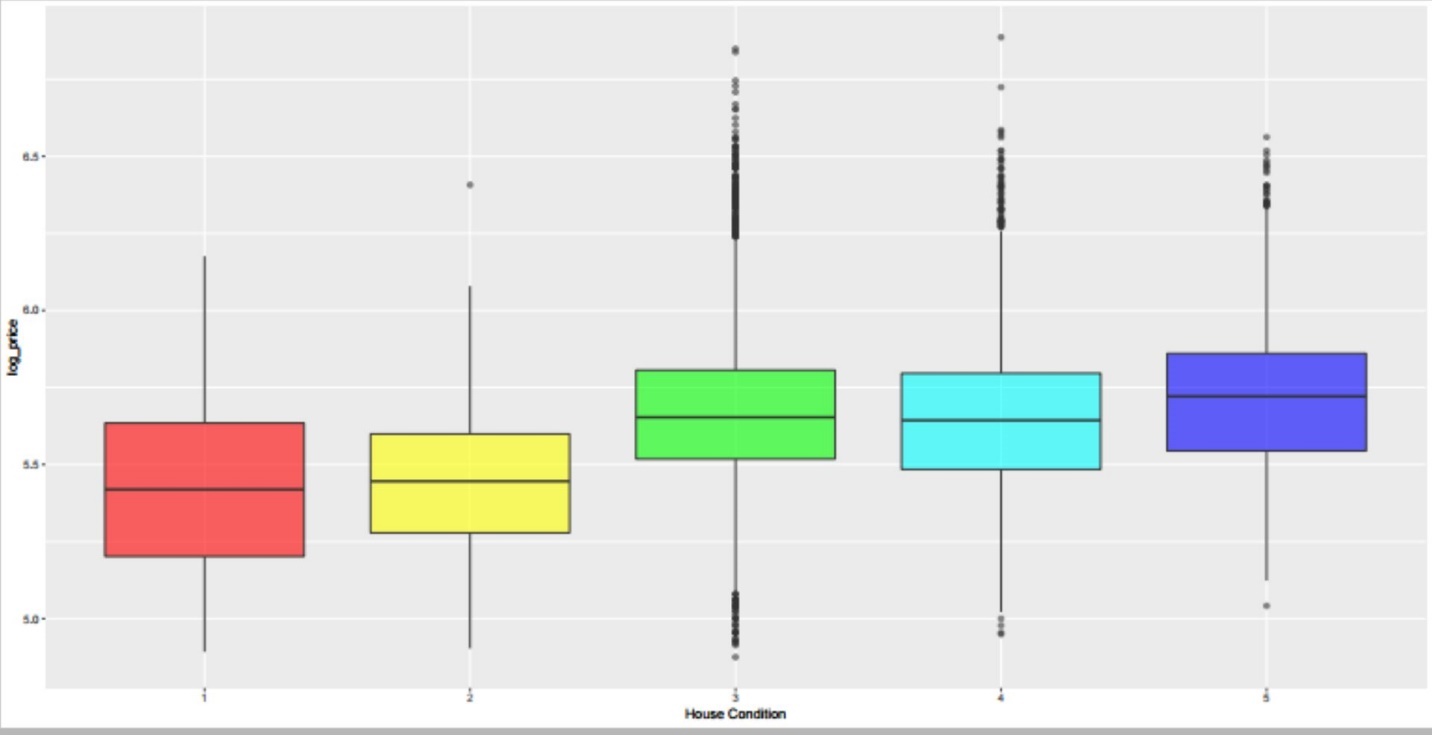
**Relationship between sqft\_living, price and the condition of the house:**





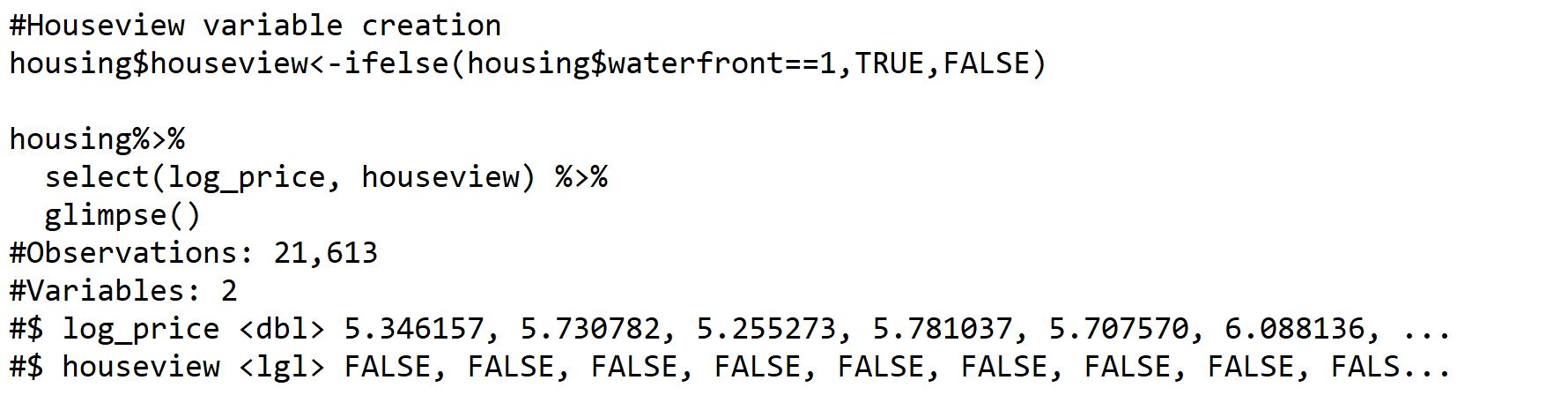
**Distribution of house prices according to condition:**

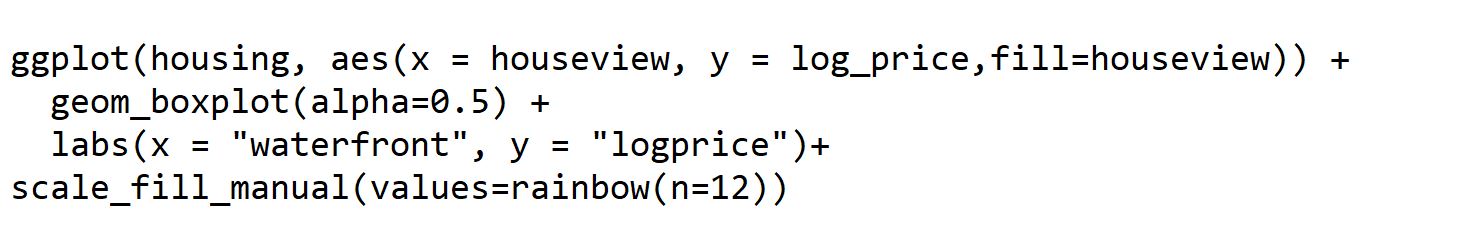


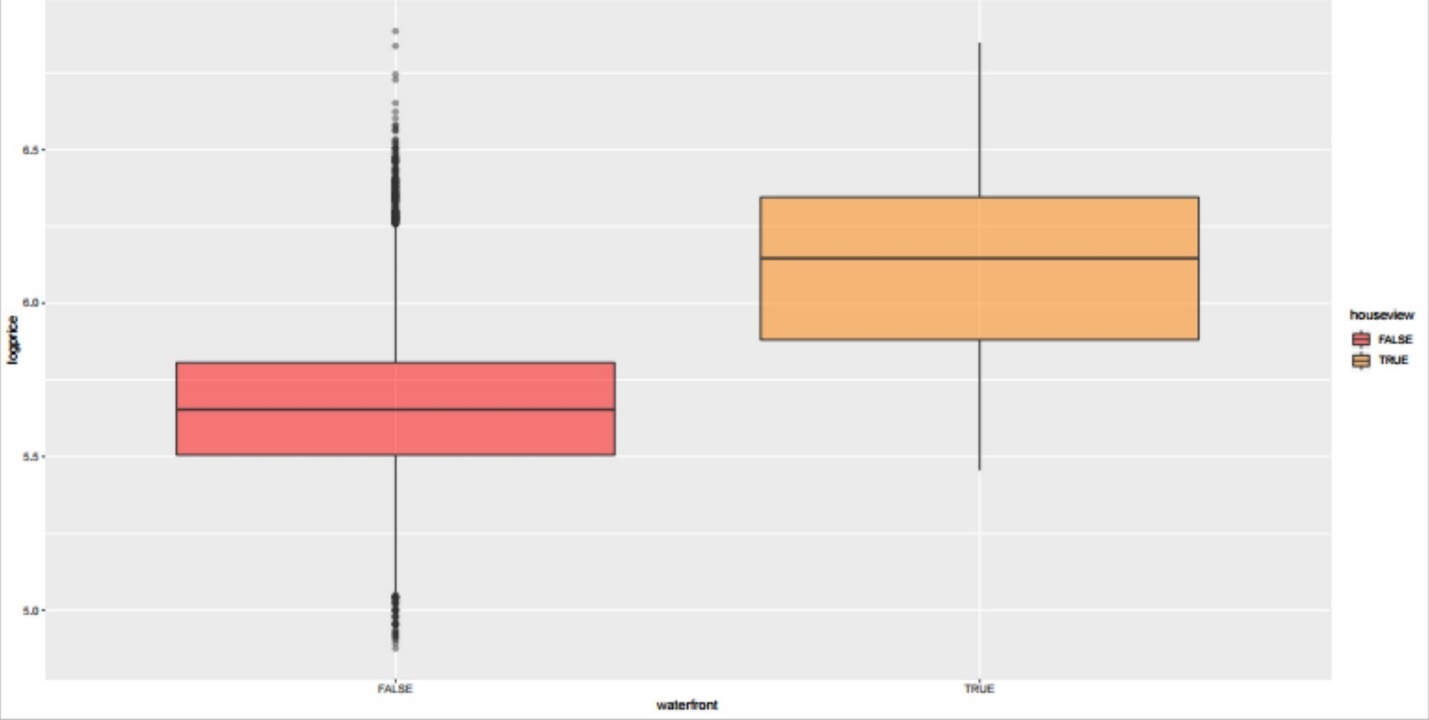


From the above plot its very clear that house prices were high if the condition was good.

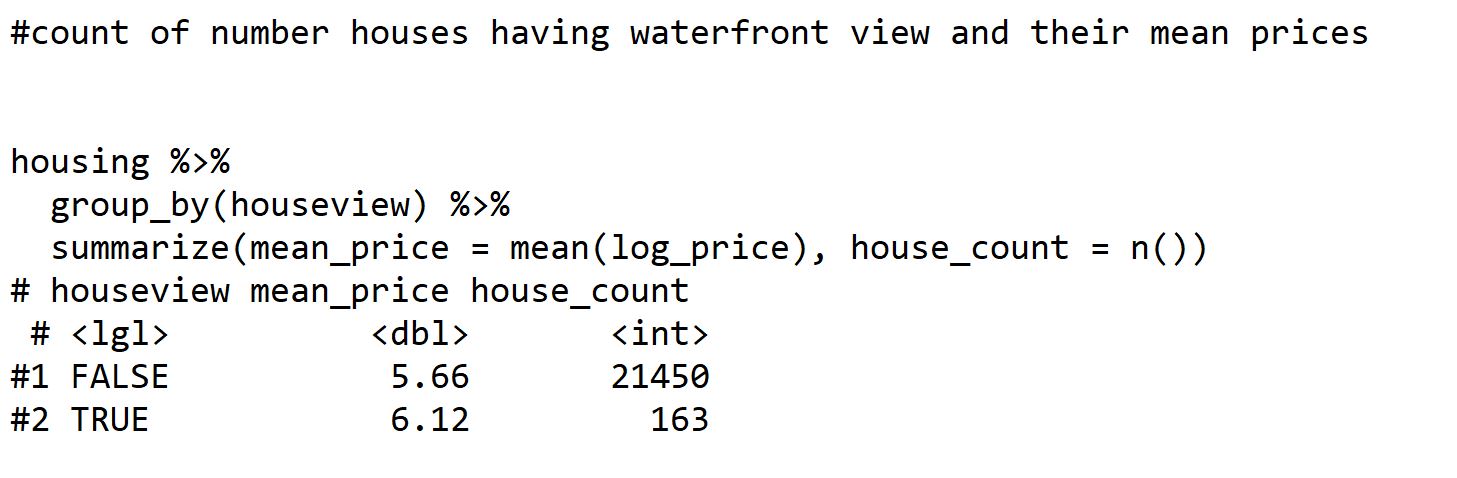
**Houseview variable creation**:







Houses that have view of waterfront tend to be much expensive than house not having a view. Most of the houses doesn't have water front, see below the price difference between those.



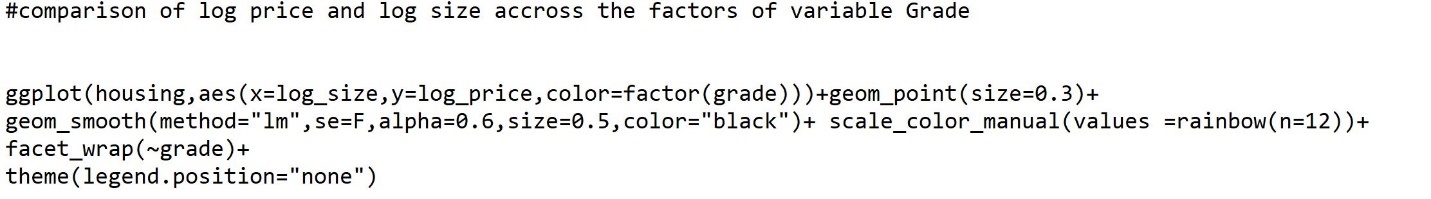
**Grade vs price:**

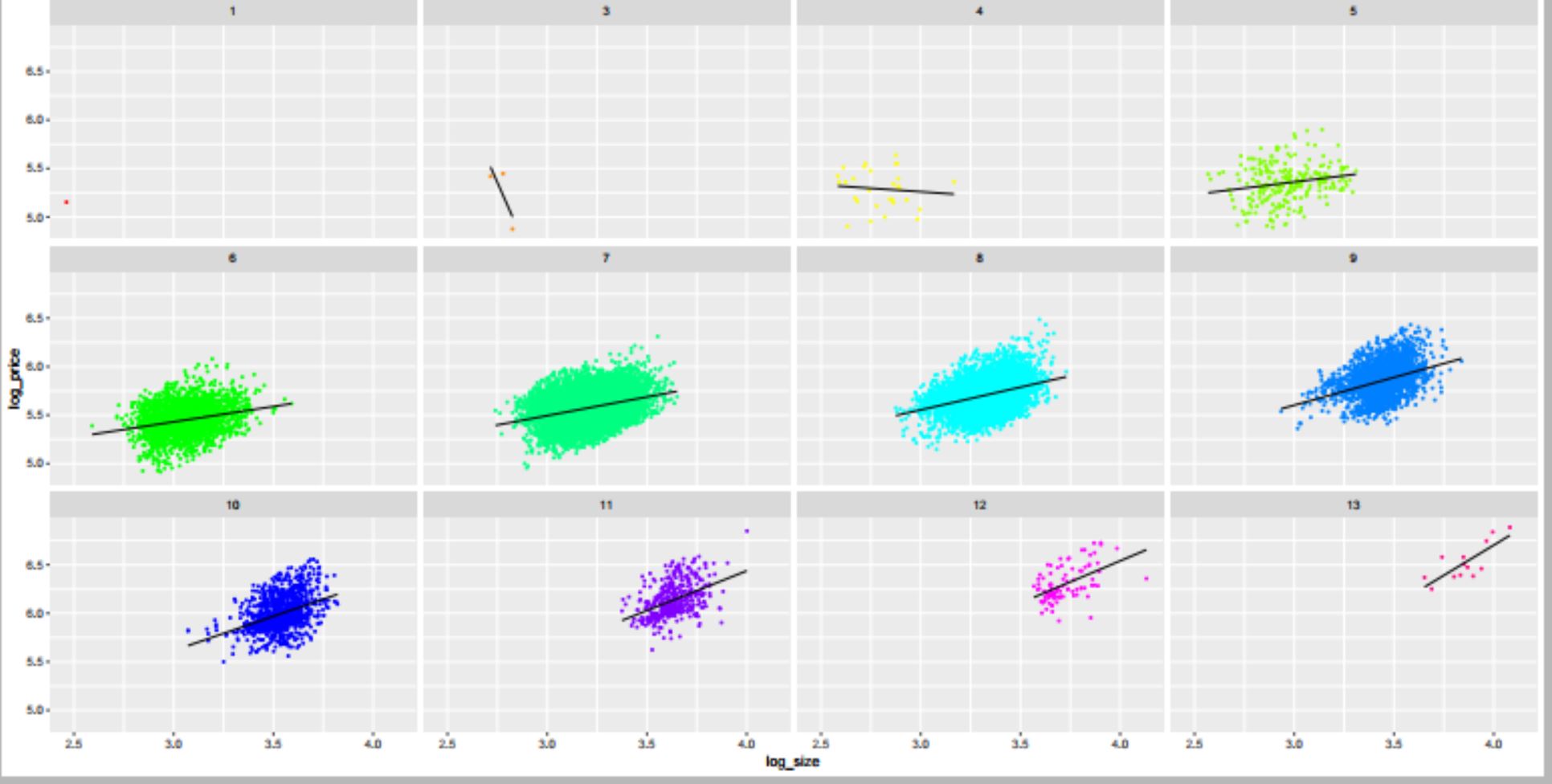
table(housing$grade)

1 3 4 5 6 7 8 9 10 11 12 13

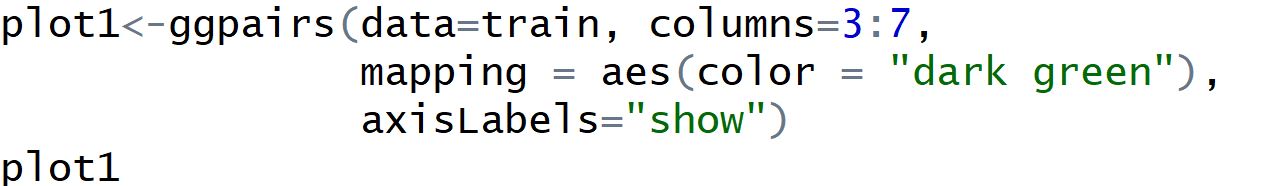
1 3 29 242 2038 8981 6068 2615 1134 399 90 1

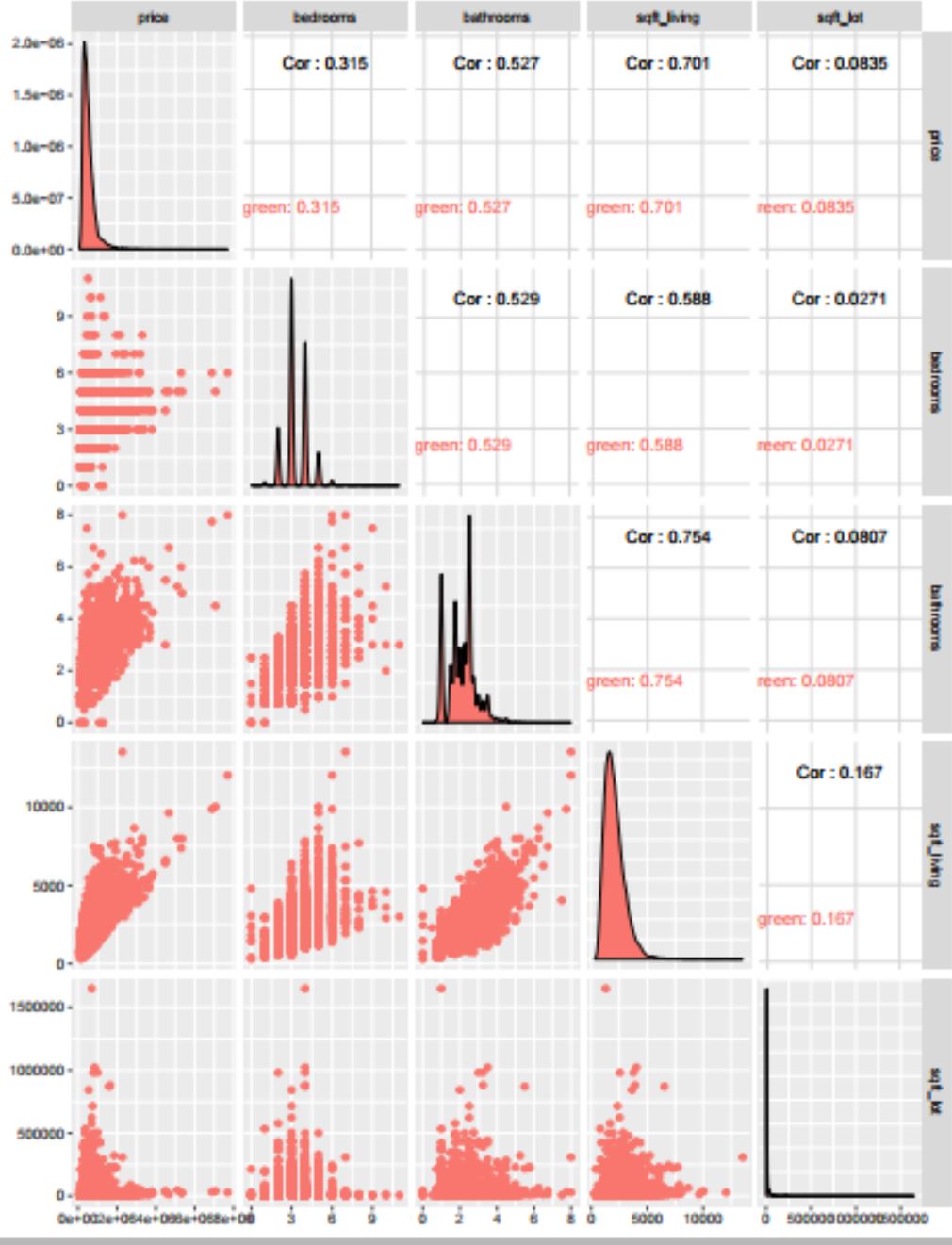
**comparison of log price and log size accross the factors of variable Grade:**



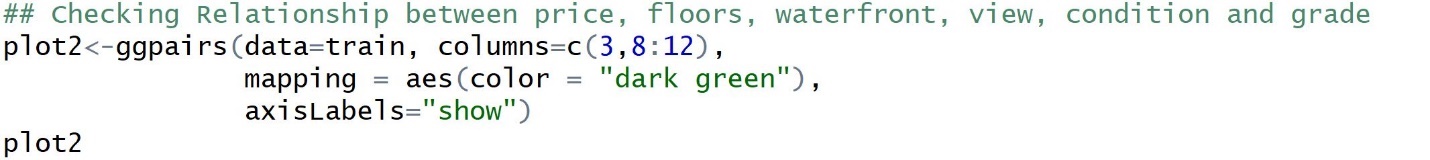


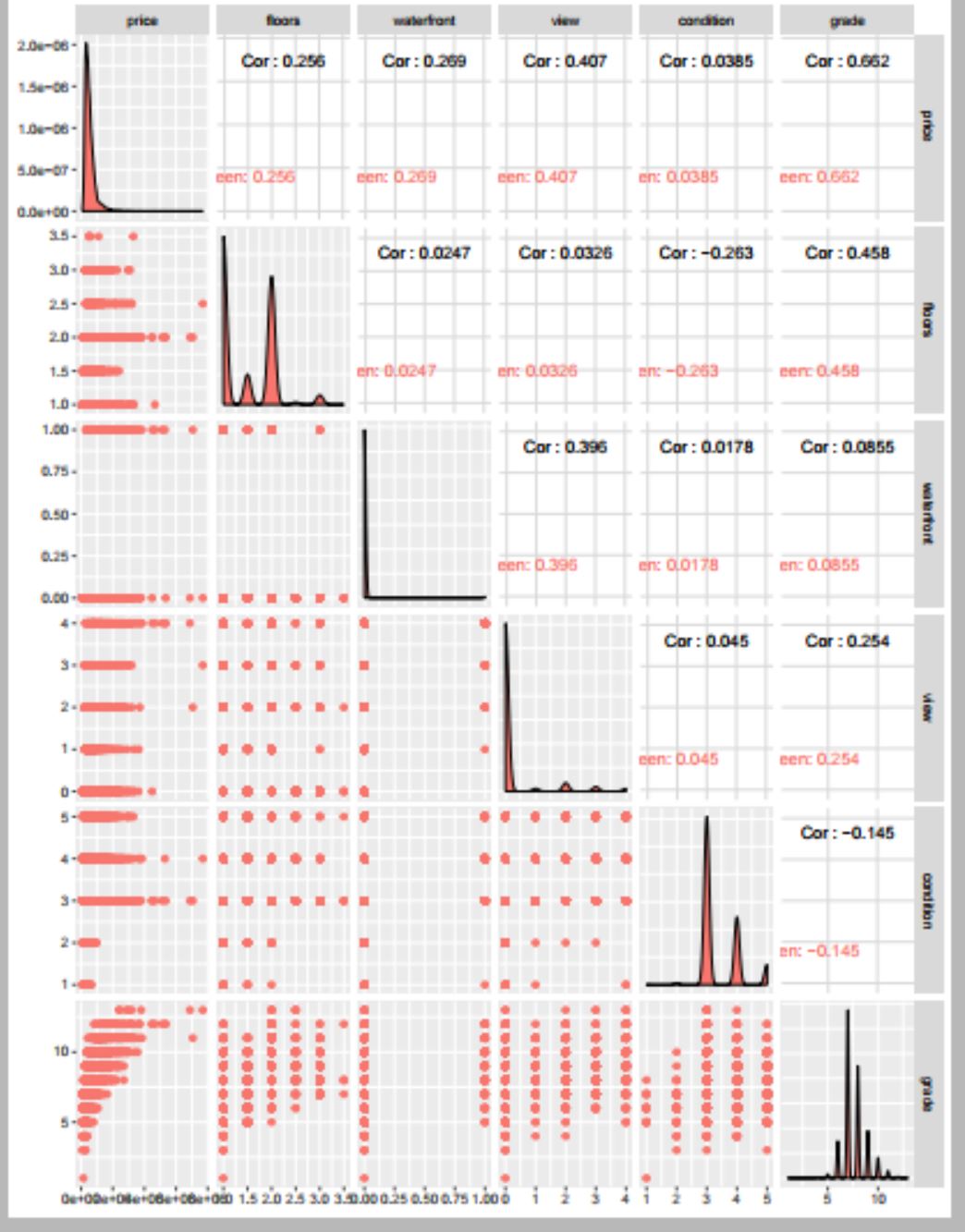
### **Correlation Matrix Among Variables:**



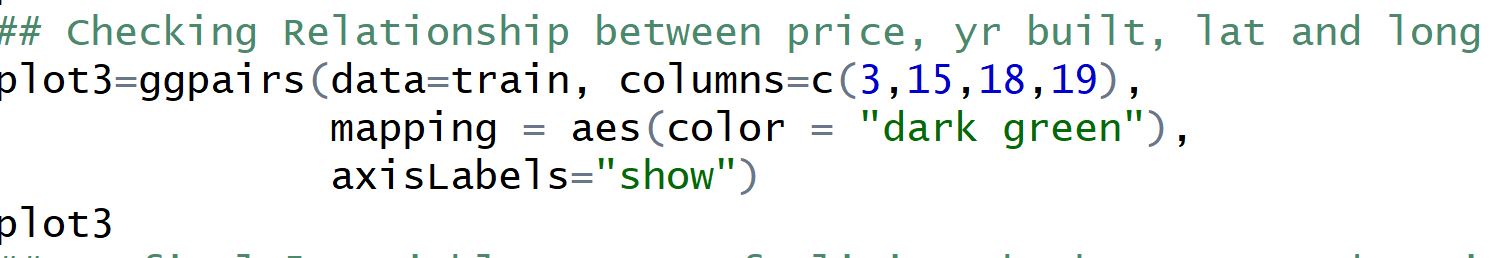


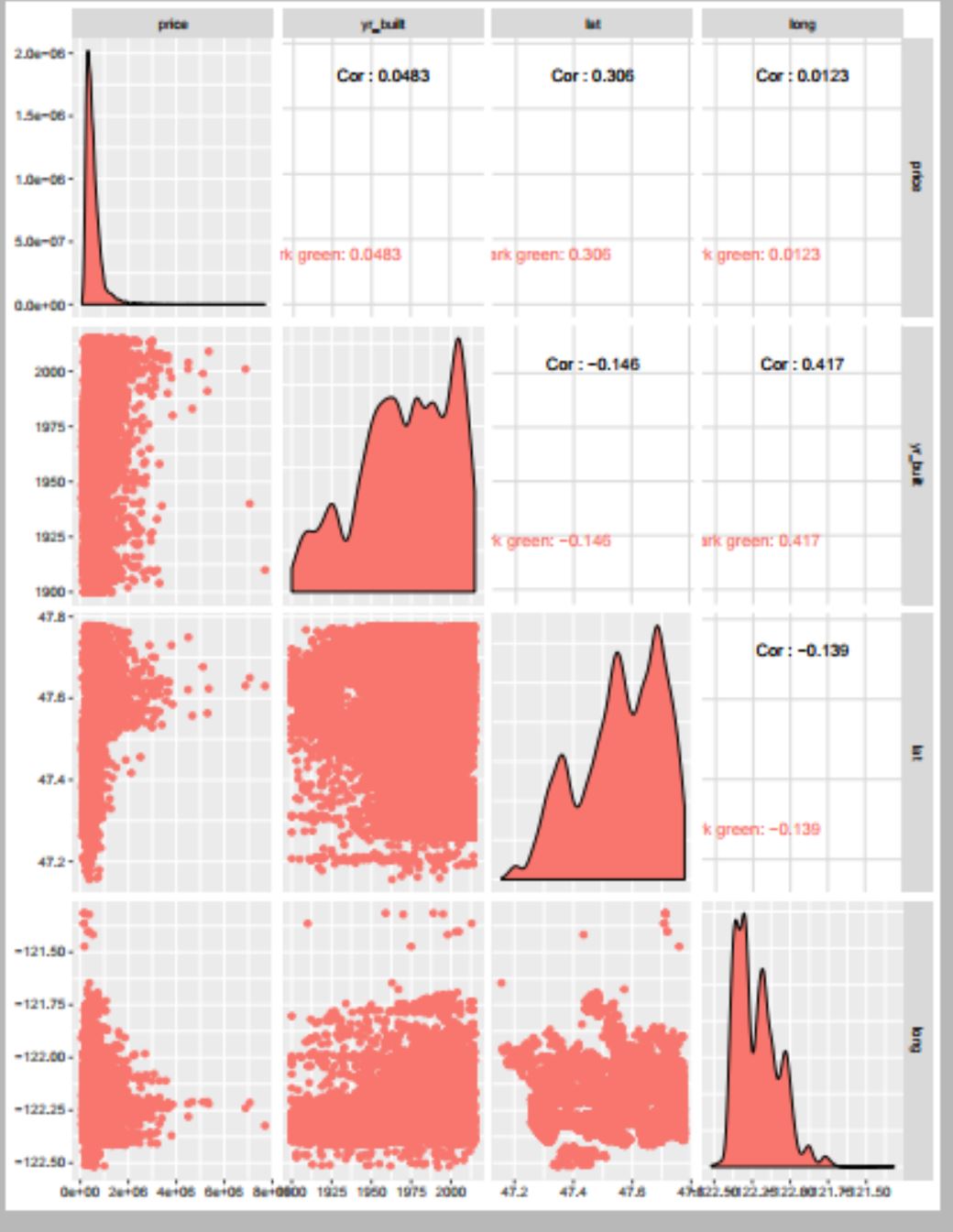
**Checking Relationship between price, floors, waterfront, view, condition and grade:**

****

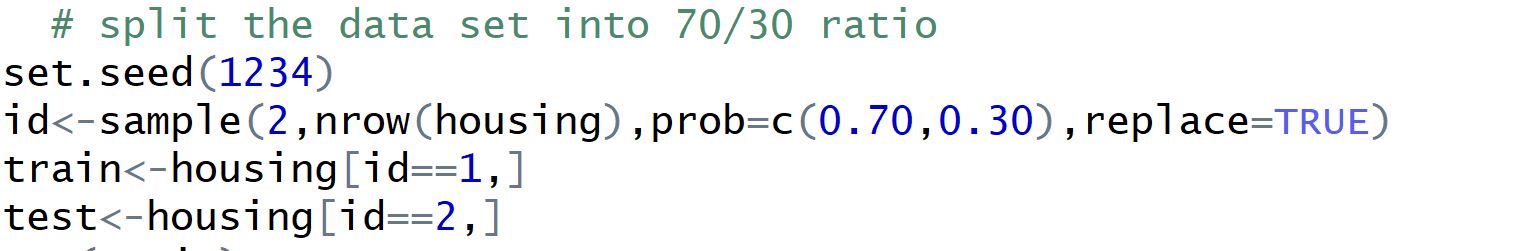


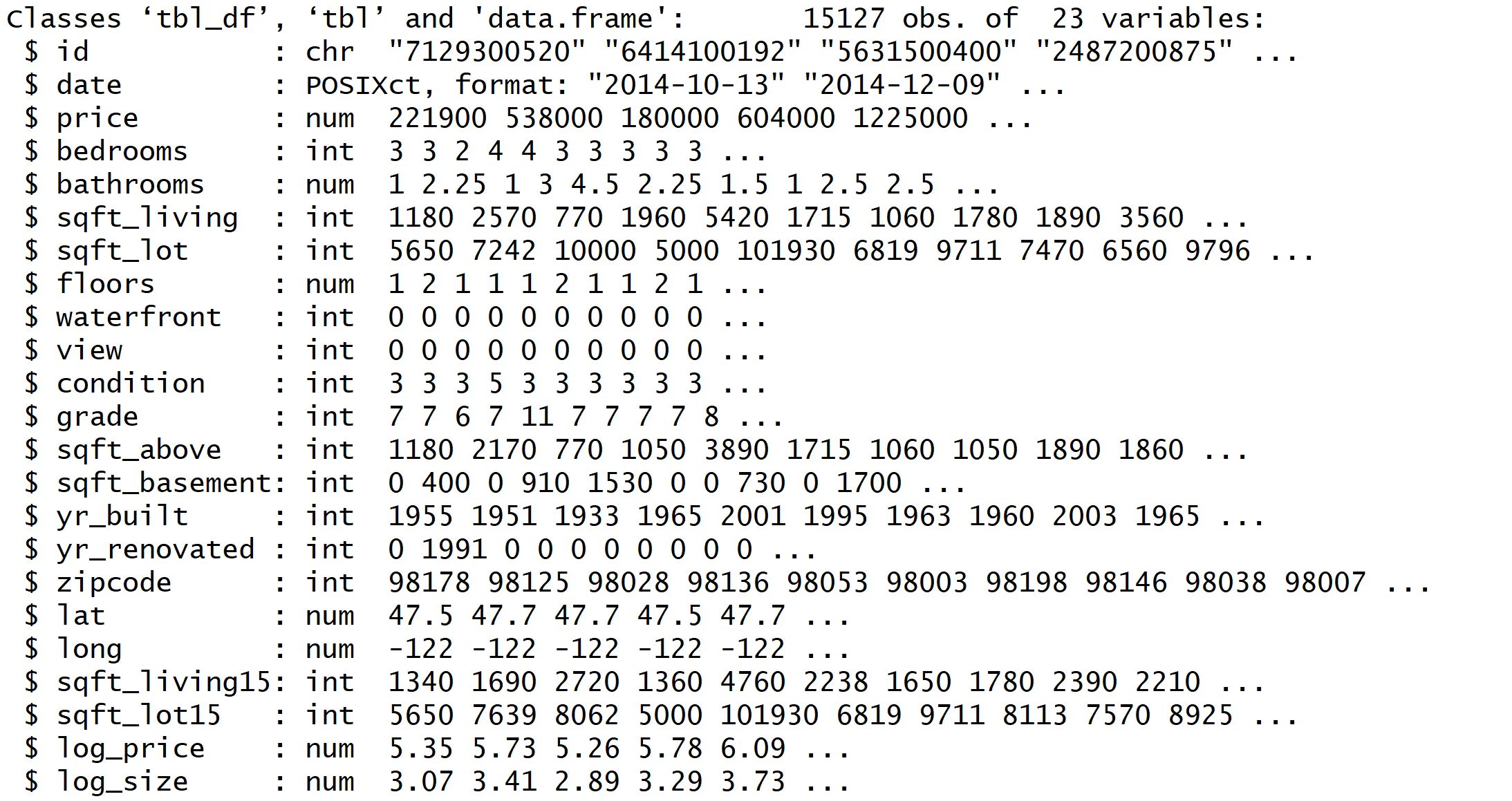
**Checking Relationship between price, yr built, lat and long:**





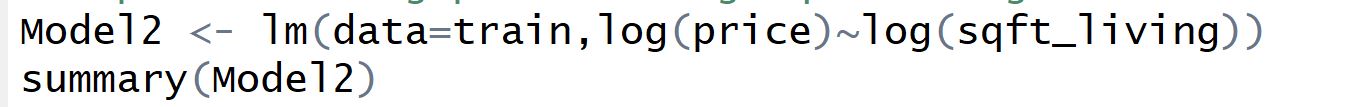
**Splitting the data into train and test data sets with 70/30 ratio:**

****

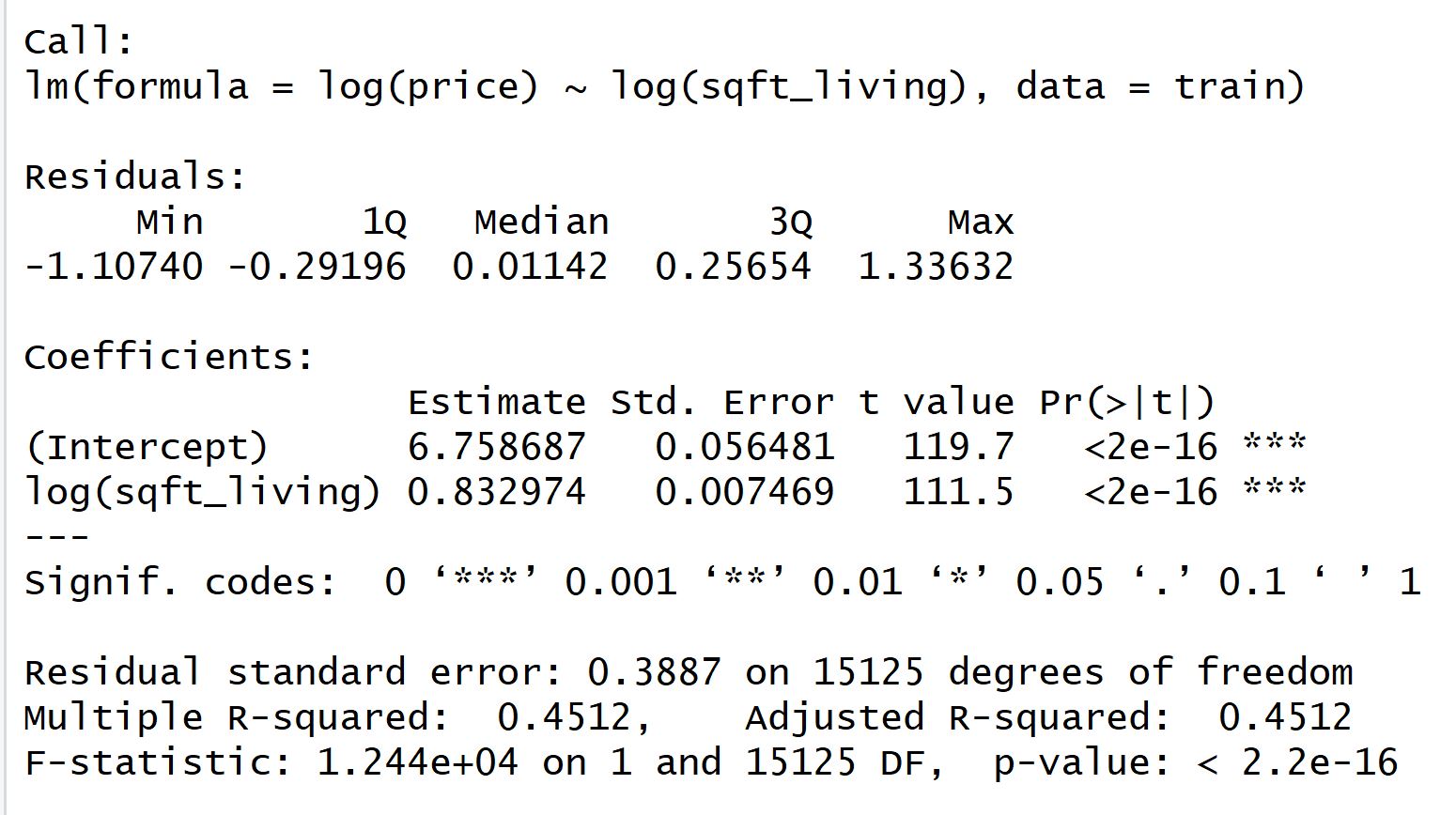
****

**Simple Regression Model:**

Let’s build a simple regression model with the variable having highest correlation with Price which is sqft\_living.

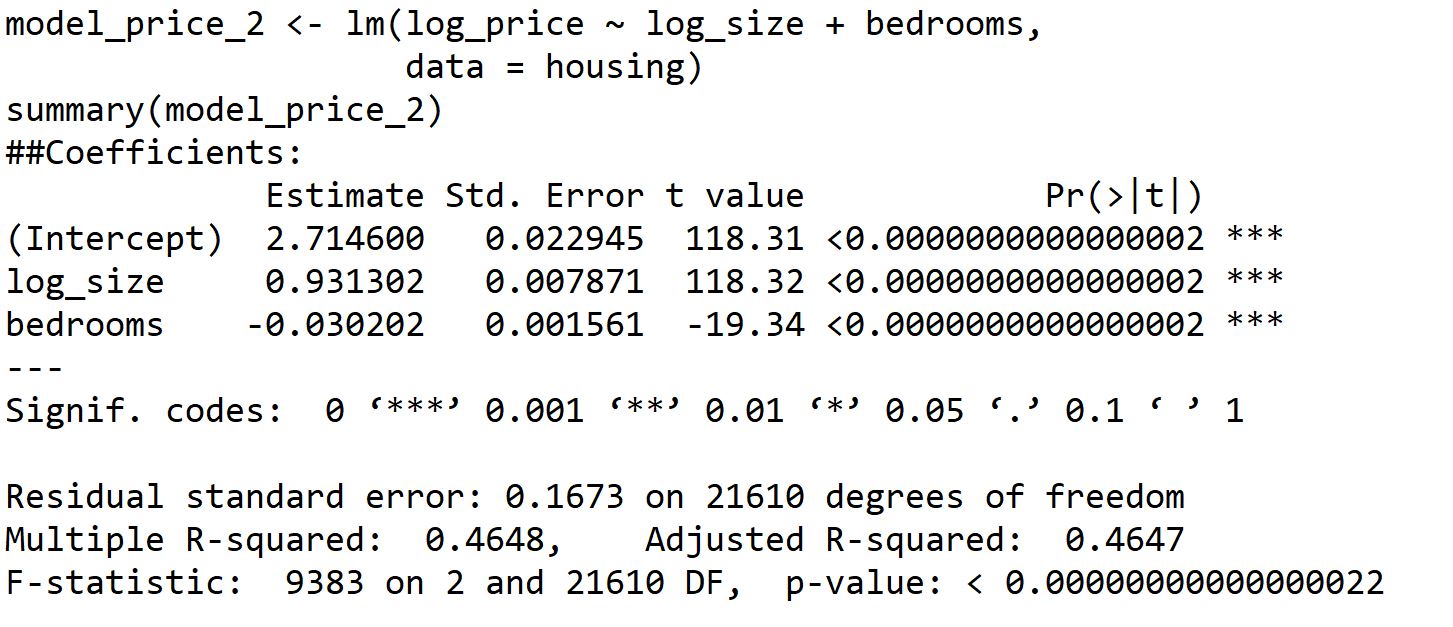
****

By looking at the regression output, we can see the values of Rsq and adjusted Rsq. The values indicate that the model is not a perfect fit which is normally not an easy task.

****

**Multiple regression Model with two independent variables:**

Now let’s build a multiple regression model by including another independent variable Bedrooms that is associated with the Price variable.

****

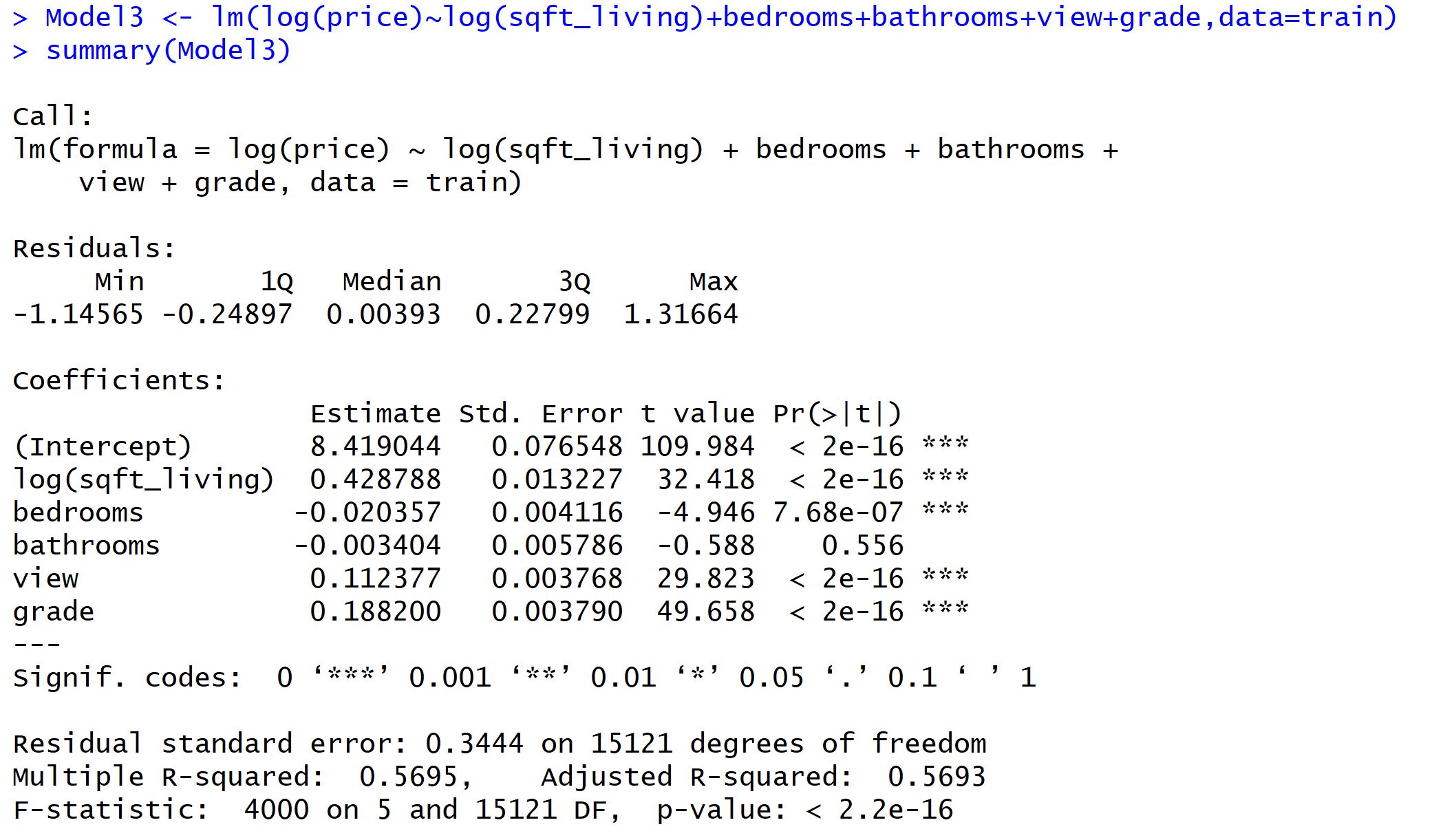
The above regression output shows that the value of adjusted Rsq is slightly higher than the simple regression model but not significantly higher. The Estimate of Bedrooms -0.030 shows that for every unit increase in bedrooms the Price will decline by 0.030 units.

**Multiple regression Model 2:**

In the previous section I used a simple linear regression and found a poor fit. In order to improve this model, I am planning to add more features, but we should be careful about the overfit which can be seen by the difference between the training and test evaluation metrics. When we have more than one feature in a linear regression, it is defined as multiple regression.

Another important thing is correlation, if there is very high correlation between two features, keeping both of them is not a good idea most of the time. For instance, sqt\_above and sqt\_living is highly correlated. This can be estimated when you look at the definitions at the dataset and check to be sure by looking at the correlation matrix.

Now, let’s include few more variables that have significant association with the variable Price. This model includes Sqft\_living, Bedrooms, Bathrooms, View and Grade.

****

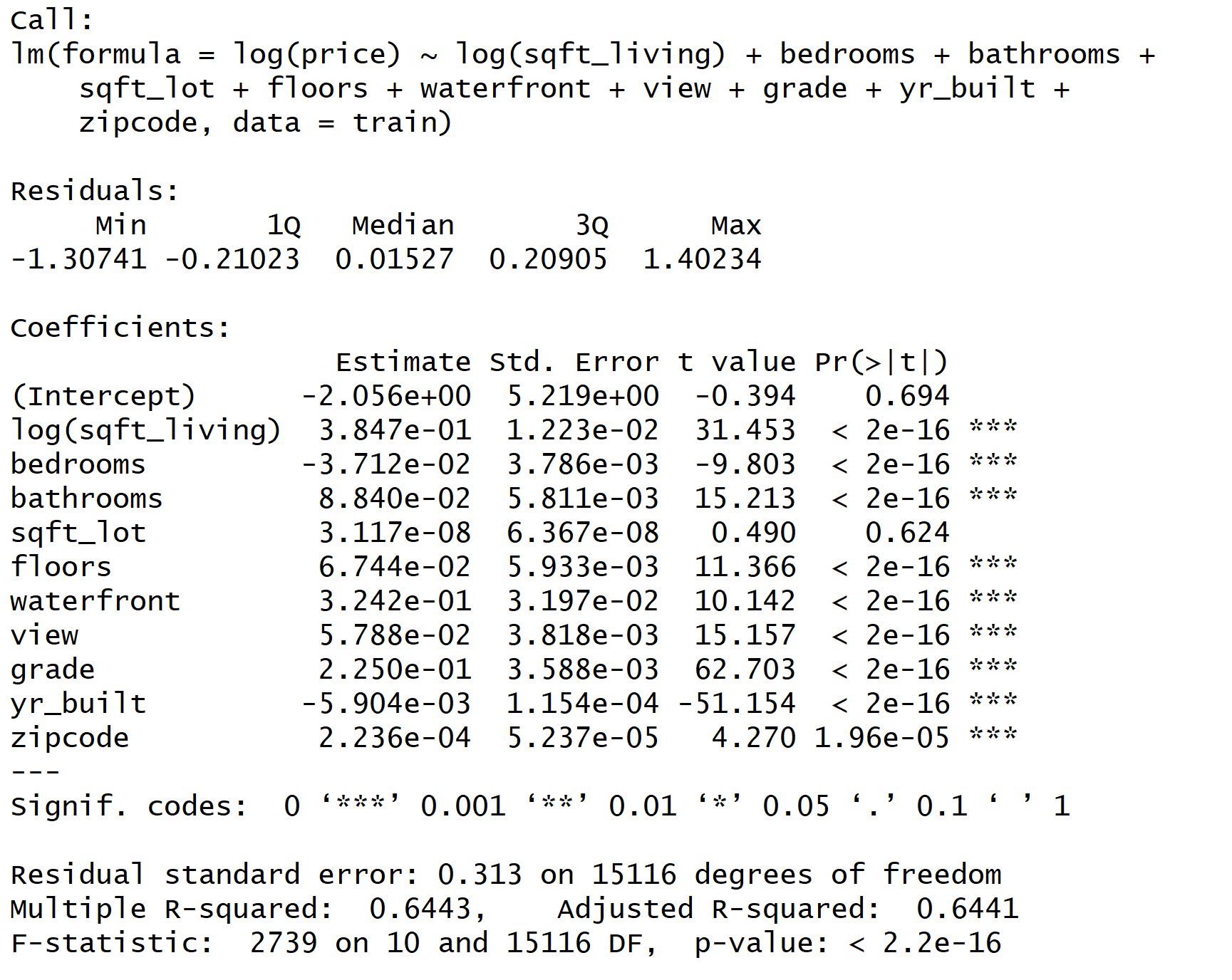
The value of adjusted Rsq is higher than the previous models so this model might be better.

**Final Model:**

Now, lets include the variables ***sqft living, sqft lot, bedrroms,bathrooms,floors,waterfront,view, grade, yr buit, zipcode***.

****

**Output:**

****

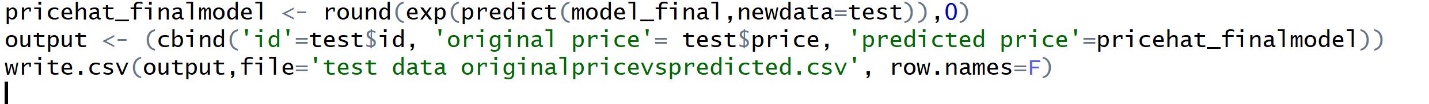
**Examining the Results of Final Model:**

By looking at the output, It is clear that the adjusted R sq is the highest among all the models created. This model might be a

Better choice among all.

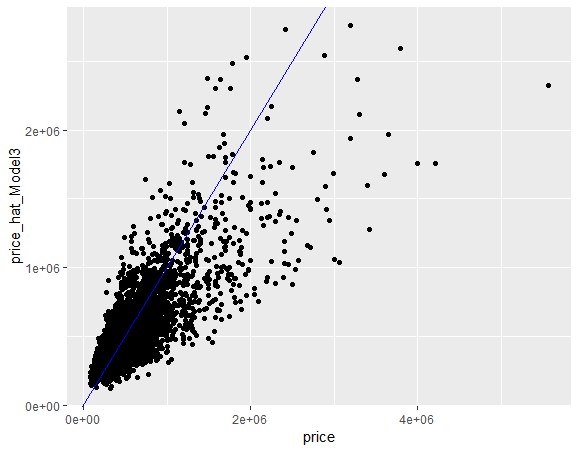
**Predicted prices from test data:**

Checking the predicted prices from selected model:

****

**Plotting predicted values vs Actual values:**

****

****