MA 678 FINAL PROJECT

House Price Study

Luo, Xijia

luoxijia@bu.edu

List

1. Introduction:

Since I plan to enter REITs industry after graduation, it is necessary to understand how both internal and external factors may affect house price. This project can also help me to predict the house price when I am facing asset valuation of residential REITs in the future.

The original data is collected from Kaggle.com. There are 4551 samples from WA, U.S., with 17 independent variables and 1 response variable (Price) in this dataset. Besides these internal factors, I think per capita income and the quality of schools are also important. Thus, I search the Median Household Income in Washington and the school rate by zip code from the internet. The link is attached in the Appendix. Then, I join these three datasets together in order to get my analyze completer and more convincing.

After the data is ready, the analysis performed will help indicate whether the proposed treatment of additional stretching exercises results in significant improvement over the normal treatment. Although the data is only representative of the house price in WA, ideally the results would provide an indication of what to expect in other states, and perhaps lead to a larger study with greater representation.

1.1 - RESEARCH QUESTIONS

A. Which factors affect the house price most?
B. Whether the impact of external factors on housing prices is significant?
C. What, if any, area should we focus on for improvement? Is view and condition good indictors in predicting house price, etc?

1.2 - VARIABLES OF INTEREST

We analyzed ten of the variables collected in the study; the price served as the response variable, and the rest of them are all used as explanatory variables. Table1 provides the name and a brief description of each variable.

Variables	Description	Type	Level
price	the house price	continuous	7800-26590000 ($)
bedroom	# of bedroom	continuous	0-9 (unit)
bathroom	# of bathroom	continuous	0-8 (unit)
sqft_living	area of living space	continuous	370-13540 (sqft)
sqft_lot	area of lot space	continuous	638-1074218 (sqft)

floor	# of bathroom	continuous	1-3.5 (floor)
view	Viewing quantity categories	Ordinal	0 - worst 4 - best
condition	house condition quantity categories	Ordinal	1 - worst 5 - best
age	house age	continuous	0-114 (year)
ES	Elementary School Rate in the house location	Ordinal	There are 10 levels 1 - worst 10 - best
MS	Middle School Rate in the house location	Ordinal	There are 10 levels 1 - worst 10 - best
HS	High School Rate in the house location	Ordinal	There are 10 levels 1 - worst 10 - best
ncome	The median of the Household annual income in the house location	continuous	32085-132665 ($)

Table1 The table shows the variable name, description, type, and level for each variable. The first variables serve as response variable and the rest of variables is the explanatory variable.

2. Method:

In this section, I will describe in detail what I did in order. Include where I get the data and how I join these data together, the Exploratory data analysis, and the STATISTICAL ANALYSIS using multilevel regression model.

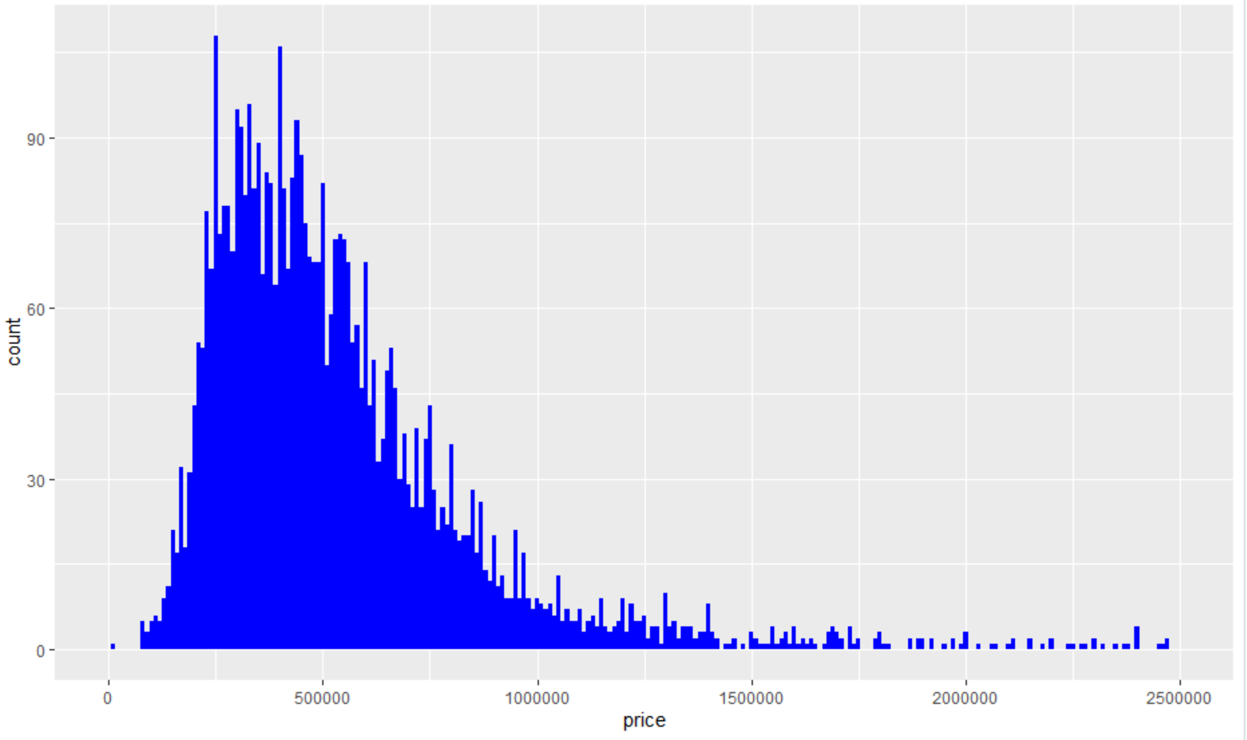
2.1 Data Collection and Data Mining

The data resource is from 3 different website. The links are attached in the Appendix. At first, I wanted to use per household income data and the school rate for the city where the house is located. But the decision was rejected because the error is not acceptable. For example, although 98112 and 98107 are both in the city Seattle, the household income and school rates are completely different. Therefore, using zip code is a better choice to obtain more accurate and convincing data. Then I search for the school's rate and income data one zip code by one zip code because both of them are online data. What's more, since the construction time is not a good indicator, I added a column of function in Excel to calculate the age of the house instead. Finally, I input data into R studio and use select function to complete the data mining work. At this point, data is ready to be analyze.

2.2 EDA

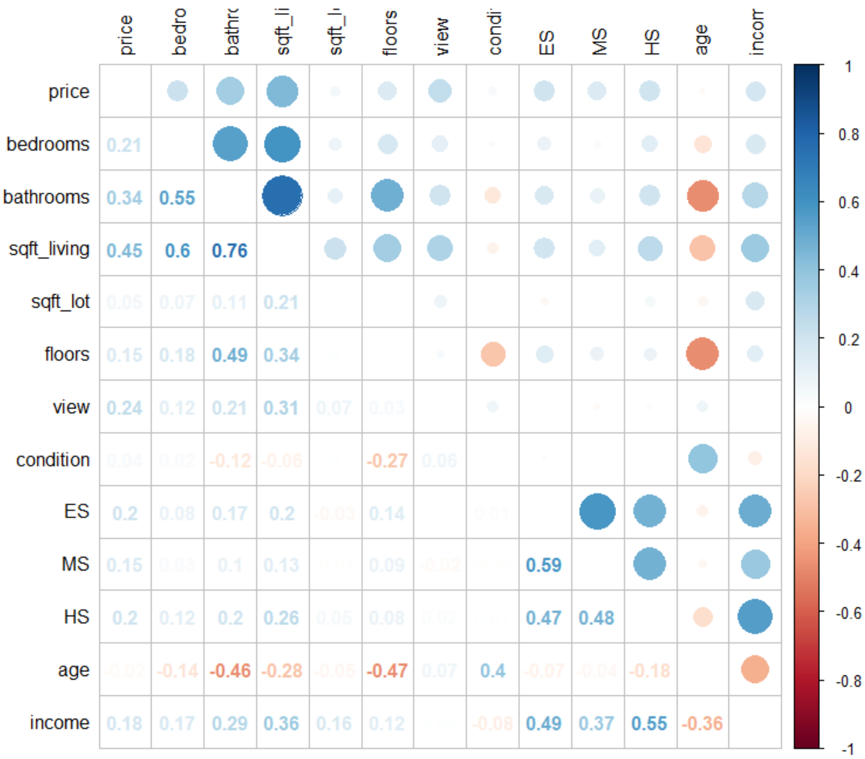
My EDA includes several data visualization work by R. These plots can give us a more intuitive view of the composition of the dataset, and can also help us choose which expression model to use.

2.2.1 House Price Distribution



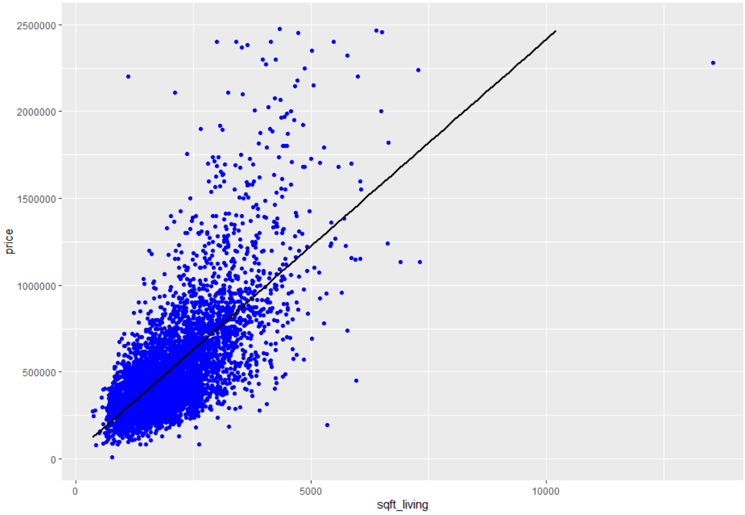
From this graph, we can find out the house prices are rights skewed. This is very consistent with the distribution of the gap between the rich and the poor. Only a few people can afford expensive houses. This graph is also helpful when we get house price data of other states in the future and make a comparation to check the impact of macroeconomic factors on housing prices.

2.2.2 Correlations visualization



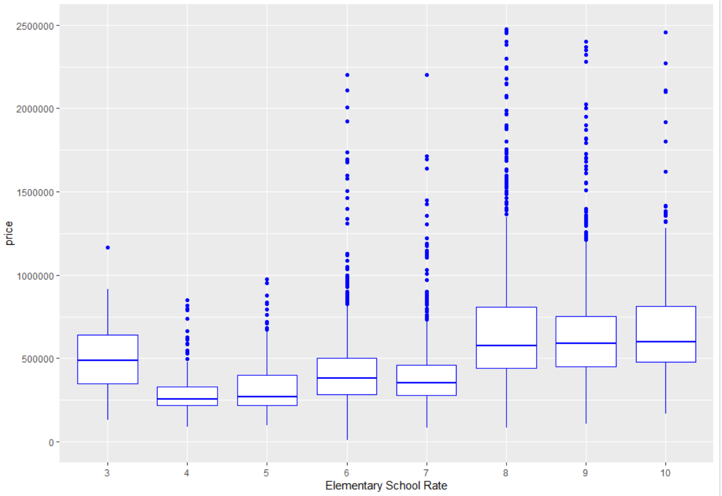
I made a matrix to display the correlation between these variables. The correlation matrix shows that all the correlations between price and other explanatory variable are positive except age. The living area and the elementary school rate have the highest correlation among the internal and external factors. Thus, I am going to visualize the relation between Price and these two indictors next.

2.2.3 Living area vs House Price



After eliminating outliers, I made the graph with a guide line shows the regression trend. From the plot we can see that the larger living area the house has, the more expensive the house will be. Therefore, there is a strong positive relationship between house price and living area.

2.2.4 Elementary School Rate vs House Price



Elementary school rate represents the quality of the best elementary school in the house's location on a scale from 1 (worst) to 10 (best). The reason why I choose the best elementary school instead of the average one is that agent will attract customers with better school data to sell houses. Thus, the local best elementary school rate can better reflect house prices. Except for the schools with three scores, we can find a positive correlation between them and house price. The house with high-rate(8-10) elementary school around are obviously more expensive that those with median-rate(4-7) elementary school. As for the outliers, 3 point schools, I think the reason may be that the developed tourist cities do not attach great importance to education.

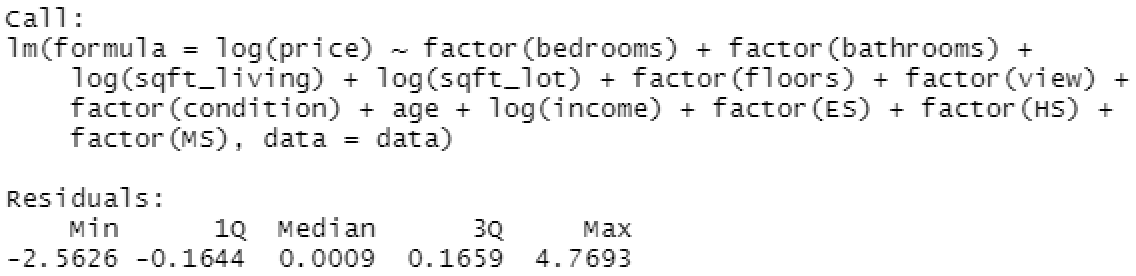
2.2.5 More EDA

Limited by the length of the article, I only select two of the most representative variables for visualization. You can find the visualization of all variables in my attach EDA.R file. You can use the link in appendix which lead you to my Github repository.

2.3 STATISTICAL ANALYSIS

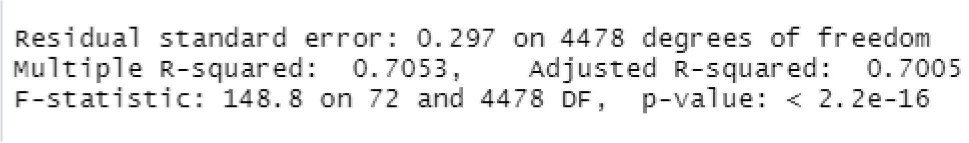
2.3.1 Simple Linear Regression

To answer the research questions, the best way is to set up a linear regression model to fit the dataset. Then we can make conclusion by the result of model. I begin to build the simple linear model in r. Since some of the variable had a huge value, I take log of them to let the model easier to read. The code and detailed result are attached in the Appendix and the fitness of the model is below:

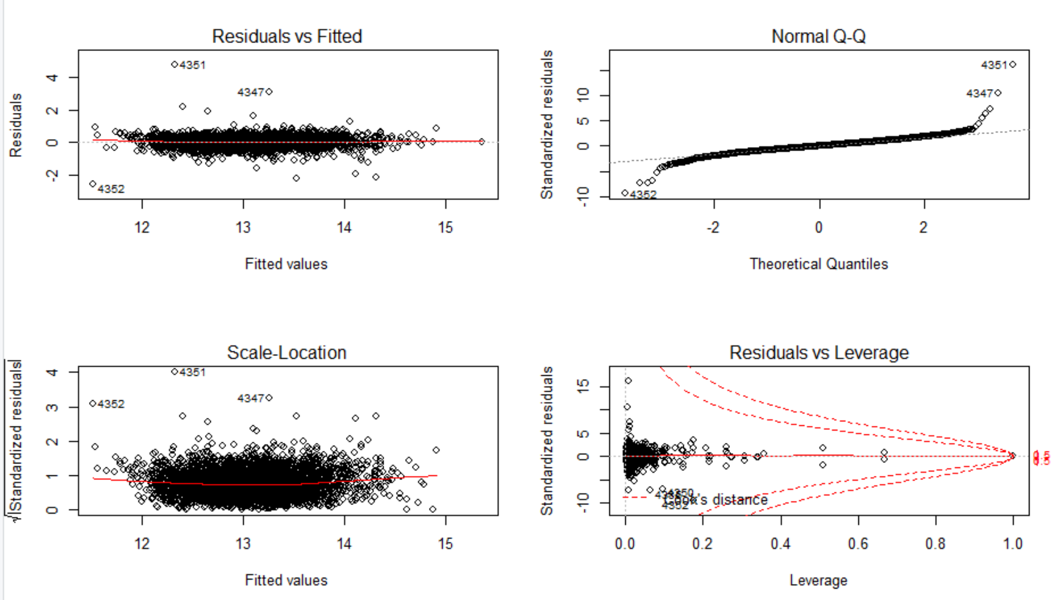


```
Call:
lm(formula = log(price) ~ factor(bedrooms) + factor(bathrooms) +
    log(sqft_living) + log(sqft_lot) + factor(floors) + factor(view) +
    factor(condition) + age + log(income) + factor(ES) + factor(HS) +
    factor(MS), data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5626 -0.1644  0.0009  0.1659  4.7693 
```



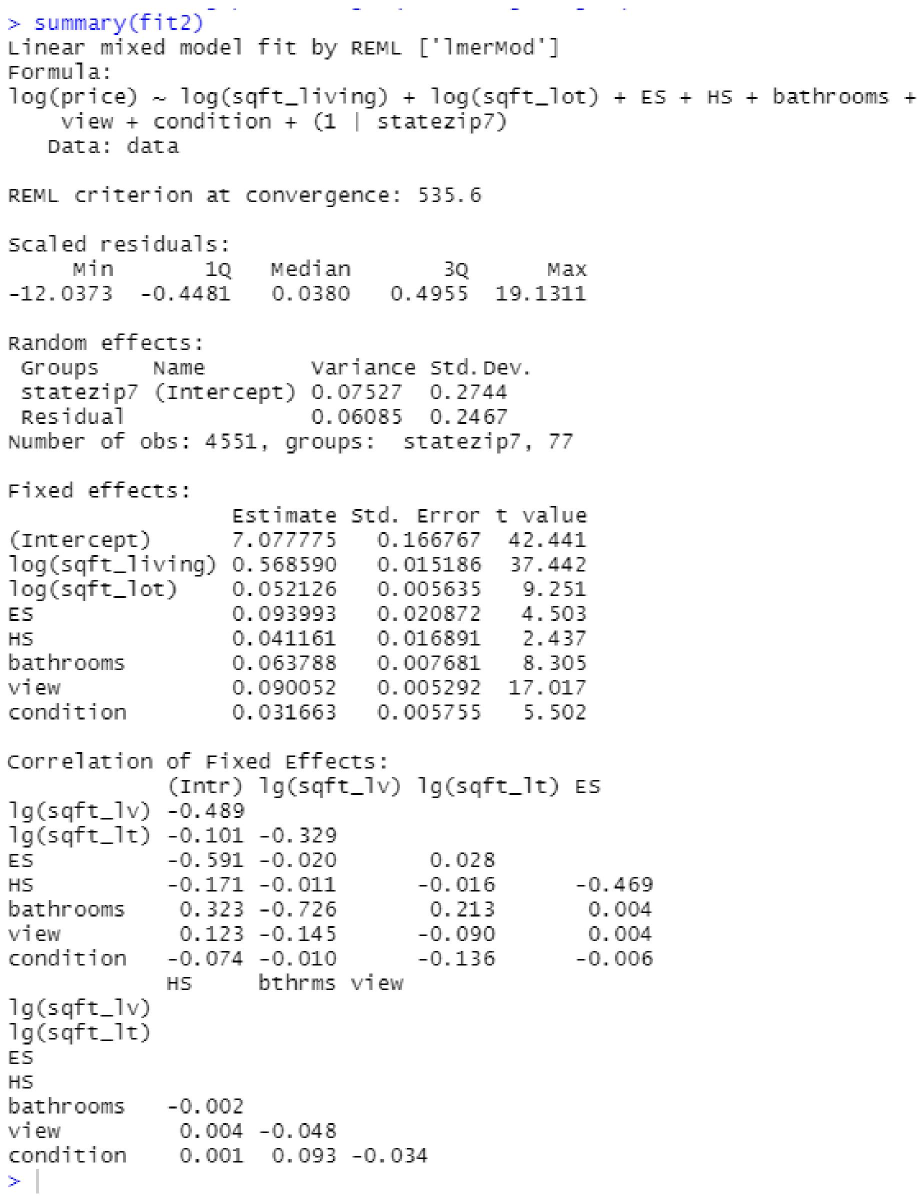
```
Residual standard error: 0.297 on 4478 degrees of freedom
Multiple R-squared:  0.7053,    Adjusted R-squared:  0.7005 
F-statistic: 148.8 on 72 and 4478 DF,  p-value: < 2.2e-16
```



The R Squared of this model is 0.7 and some of the indictors are not significant at all. From the residual plots we can find out that residuals are not randomly distributed. Therefore, we need use the multilevel linear regression to make some adjustments to our current model.

2.3.2 Multilevel Linear Regression

We use lmer() fuction to set up a multilevel linear model. Since the houses are located all over the WA, we can use zipcode to classify them. There are 77 groups in total. We will consider them as random indictor using (1|zipcode7) or (ES|zipcode7) in the regression function. Then, I eliminate some unsignificant variable such as age so that the model will be fitter to the data. After many attempts, I finally decided to use living area, lot area, ES, HS, MS, bathrooms, view, and condition with the random factor (1|zipcode7) since this model is the most fitted one. The result is at below.



```
> summary(fit2)
Linear mixed model fit by REML ['lmerMod']
Formula:
log(price) ~ log(sqft_living) + log(sqft_lot) + ES + HS + bathrooms +
    view + condition + (1 | statezip7)
   Data: data

REML criterion at convergence: 535.6

Scaled residuals:
     Min       1Q   Median       3Q      Max
-12.0373  -0.4481   0.0380   0.4955  19.1311

Random effects:
 Groups    Name        Variance Std.Dev.
 statezip7 (Intercept) 0.07527  0.2744
 Residual              0.06085  0.2467
Number of obs: 4551, groups:  statezip7, 77

Fixed effects:
                 Estimate Std. Error t value
(Intercept)      7.077775   0.166767  42.441
log(sqft_living) 0.568590   0.015186  37.442
log(sqft_lot)    0.052126   0.005635   9.251
ES               0.093993   0.020872   4.503
HS               0.041161   0.016891   2.437
bathrooms        0.063788   0.007681   8.305
view             0.090052   0.005292  17.017
condition        0.031663   0.005755   5.502

Correlation of Fixed Effects:
            (Intr) lg(sqft_lv) lg(sqft_lt) ES
lg(sqft_lv) -0.489
lg(sqft_lt) -0.101 -0.329
ES          -0.591 -0.020       0.028
HS          -0.171 -0.011      -0.016      -0.469
bathrooms    0.323 -0.726       0.213       0.004
view         0.123 -0.145      -0.090       0.004
condition   -0.074 -0.010      -0.136      -0.006
            HS     bthrms view
lg(sqft_lv)
lg(sqft_lt)
ES
HS
bathrooms   -0.002
view         0.004 -0.048
condition    0.001  0.093 -0.034
> |
```

The results show that house prices in any region will be affected by the following explanatory variables. Here is the interruption of these fixed effects:
Sqft_living: 1% increase will cause the house price increase 0.568% on average. (p=0)
Sqft_log:1% increase will cause the house price increase 0.052% on average. (p=0)
ES: each rate score augment will cause the house price increase 0.093% on average. (p=0)
HS: each rate score augment will cause the house price increase 0.041% on average. (p=0.014<0.05)
Bathrooms: every extra bathroom will cause the house price increase 0.063% on average. (p=0)
View: each rate score augment will cause the house price increase 0.09% on average. (p=0)
Condition: each rate score augment will cause the house price increase 0.031% on average. (p=0)

3. Result:

Based on the two linear regression model, living area size is the most statistically significant variable in predicting house price. It's very common sense, because the first thing people think about when they buy a house is size. All these internal factors can be connected to the concern of size. For example, the number of bedrooms is larger, the living area will be bigger, thus the house price goes up.

For external factors, the quality of the school district directly determines the quality of housing prices. The higher the school score, the higher the house price. Among them, the primary school rate accounts for the highest proportion in determine the value of a house. However, the interesting thing is that annual family income is not significant in predicting house price. Apart from the fact that the model may be wrong, I think the reason for this phenomenon is that the local people's income source does not necessarily come from the local. In other words, some people who live in the suburbs may earn money in the city, so their annual income reflects the economic situation of the city rather than the local area. Moreover, in order to make more money, people are more willing to go to the city to earn money, so the median annual income of different regions has little difference, so it is impossible to accurately predict local house prices.

In conclusion, both internal and external factors are very important for forecasting house prices. Although some of them have been proved not to be significant, most of them are useful for us to better understand the house price model.

4. Discussion

4.1 What I learn from this project

The biggest benefit is that I learned to think comprehensively. In my earliest version of proposal, I only mentioned considering internal factors, and my main goal was the accuracy of prediction. In my original plan, I didn't need to know what each variable meant. I just needed to know the independent variable and dependent variable, and then get the result through statistical method. But I find it more important to understand this project than to do code all the time because this project is meaningful only if you understand the whole research. That's also the most helpful lesson in this class, in this semester.

4.2 Consideration (include next step)

There are several considerations to help understand and improve the study of this data. Firstly, the sample range and timeliness are limited. The dataset was created in 2014, 6 years from now. And all the houses are located in WA. It is not difficult to solve this problem, but rather tedious. I plan to collect the data by myself from Zillow group, INC. From there, I can get detailed data, including all the internal factors that interest me such as the swimming pool and garbage size. However, Limited by time and energy, I chose the one sorted by others on Kaggle.com. If I want to continue to study this topic, I have plans and ways to get better datasets.

Secondly, the choice of external factors is not perfect. For example, as I mentioned earlier, the median household income may not reflect the regional economy well. Local GDP might be a better choice which is available in some professional paid databank. On the other hand, the school rating system only includes public schools. If the house is close to some top private schools, I think the house price will also rise. Then, this study did not include the quality of nearby universities, the convenience of public transportation, the impact of climate and other external factors. To improve this research, a lot of time and effort is needed to build and refine the dataset.

Finally, we only roughly analyze the factors that affect house prices through different kinds of regression, and the accuracy of regression model prediction is not enough. In order to better predict house prices, I'm going to break the data I've collected into two parts- train&test. I'll use xgboost method to train the data from the train section, and then bring the machine learning model back to the test section for prediction. Then I will compare the predict data and the real data to see the accuracy. This part of work probably happen in next semester.

Reference:

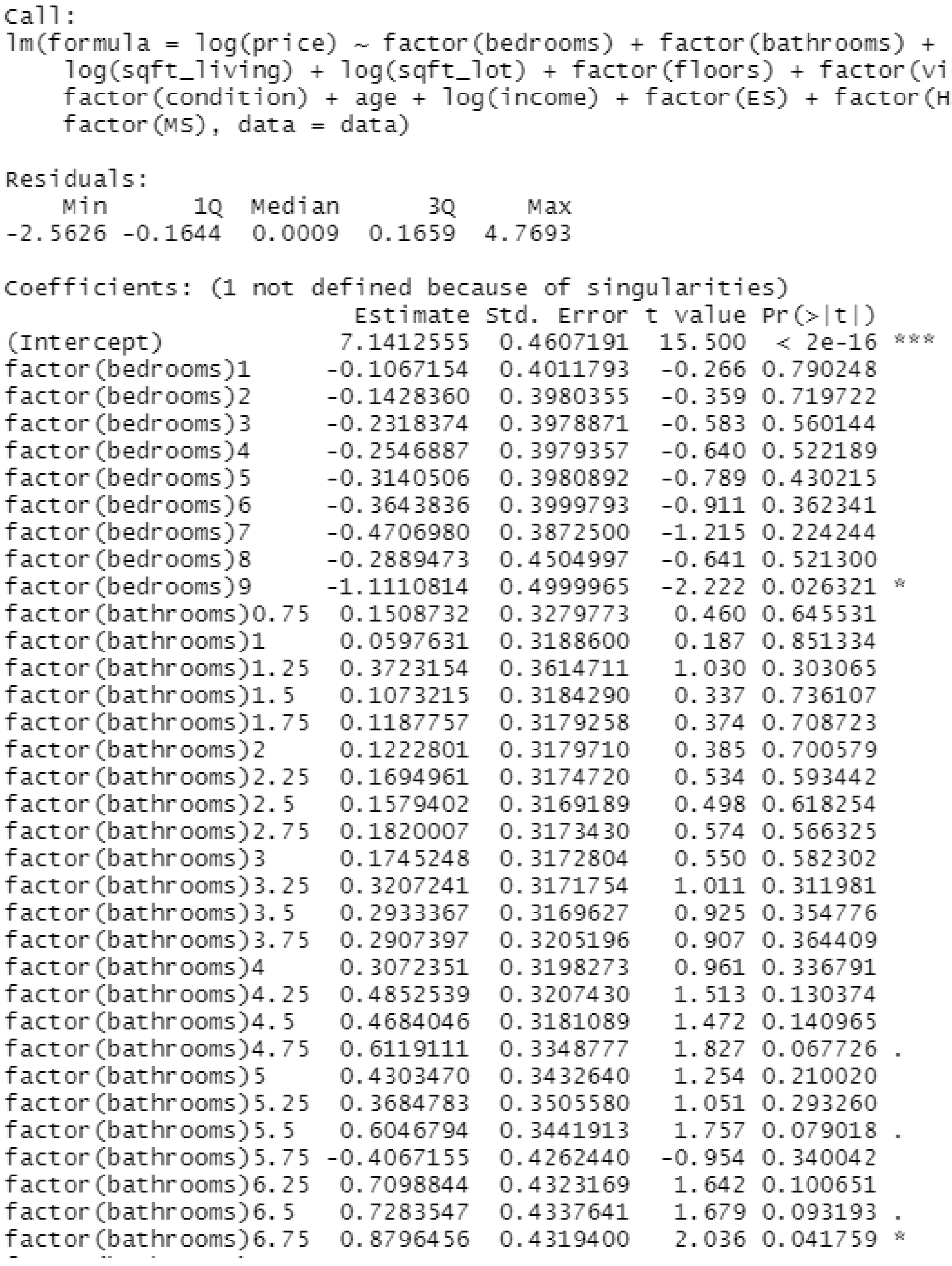
https://quantdev.ssri.psu.edu/tutorials/r-bootcamp-introduction-multilevel-model-and-interactions
https://zhuanlan.zhihu.com/p/150878441
https://www.kaggle.com/erikbruin/house-prices-lasso-xgboost-and-a-detailed-eda/log

Appendix:

Data Source: https://www.kaggle.com/shree1992/housedata
Github Link: https://github.com/Jack557557/MA678FinalProject
School rate link: https://www.greatschools.org/
Household Income data source link: http://zipatlas.com/us/wa/zip-code-comparison/median-

household-income.2.htm

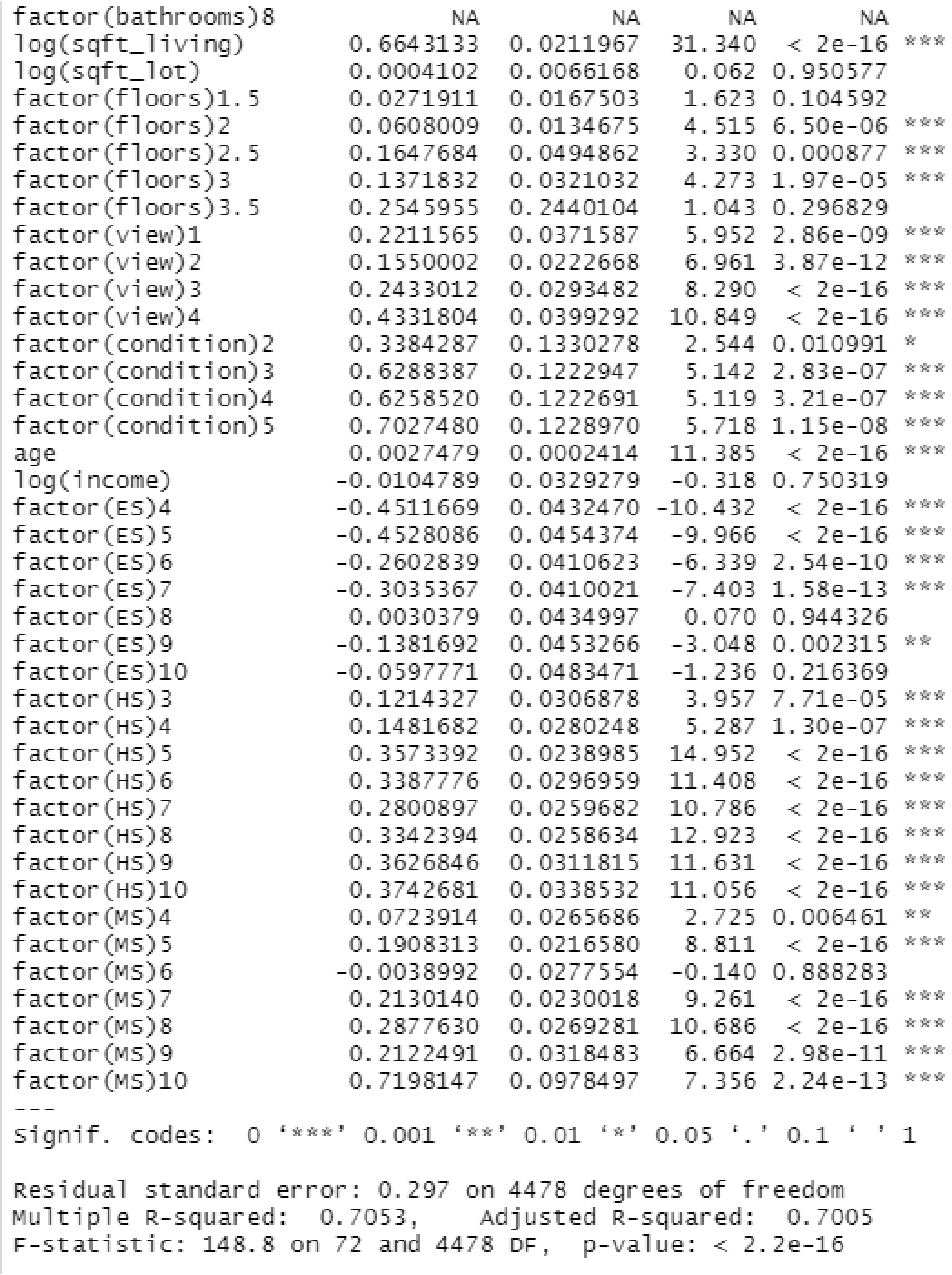
Simple linear regression model:



```
Call:
lm(formula = log(price) ~ factor(bedrooms) + factor(bathrooms) +
    log(sqft_living) + log(sqft_lot) + factor(floors) + factor(view) +
    factor(condition) + age + log(income) + factor(ES) + factor(HS) +
    factor(MS), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5626 -0.1644  0.0009  0.1659  4.7693

Coefficients: (1 not defined because of singularities)
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            7.1412555  0.4607191  15.500  < 2e-16 ***
factor(bedrooms)1     -0.1067154  0.4011793  -0.266 0.790248
factor(bedrooms)2     -0.1428360  0.3980355  -0.359 0.719722
factor(bedrooms)3     -0.2318374  0.3978871  -0.583 0.560144
factor(bedrooms)4     -0.2546887  0.3979357  -0.640 0.522189
factor(bedrooms)5     -0.3140506  0.3980892  -0.789 0.430215
factor(bedrooms)6     -0.3643836  0.3999793  -0.911 0.362341
factor(bedrooms)7     -0.4706980  0.3872500  -1.215 0.224244
factor(bedrooms)8     -0.2889473  0.4504997  -0.641 0.521300
factor(bedrooms)9     -1.1110814  0.4999965  -2.222 0.026321 *
factor(bathrooms)0.75  0.1508732  0.3279773   0.460 0.645531
factor(bathrooms)1     0.0597631  0.3188600   0.187 0.851334
factor(bathrooms)1.25  0.3723154  0.3614711   1.030 0.303065
factor(bathrooms)1.5   0.1073215  0.3184290   0.337 0.736107
factor(bathrooms)1.75  0.1187757  0.3179258   0.374 0.708723
factor(bathrooms)2     0.1222801  0.3179710   0.385 0.700579
factor(bathrooms)2.25  0.1694961  0.3174720   0.534 0.593442
factor(bathrooms)2.5   0.1579402  0.3169189   0.498 0.618254
factor(bathrooms)2.75  0.1820007  0.3173430   0.574 0.566325
factor(bathrooms)3     0.1745248  0.3172804   0.550 0.582302
factor(bathrooms)3.25  0.3207241  0.3171754   1.011 0.311981
factor(bathrooms)3.5   0.2933367  0.3169627   0.925 0.354776
factor(bathrooms)3.75  0.2907397  0.3205196   0.907 0.364409
factor(bathrooms)4     0.3072351  0.3198273   0.961 0.336791
factor(bathrooms)4.25  0.4852539  0.3207430   1.513 0.130374
factor(bathrooms)4.5   0.4684046  0.3181089   1.472 0.140965
factor(bathrooms)4.75  0.6119111  0.3348777   1.827 0.067726 .
factor(bathrooms)5     0.4303470  0.3432640   1.254 0.210020
factor(bathrooms)5.25  0.3684783  0.3505580   1.051 0.293260
factor(bathrooms)5.5   0.6046794  0.3441913   1.757 0.079018 .
factor(bathrooms)5.75 -0.4067155  0.4262440  -0.954 0.340042
factor(bathrooms)6.25  0.7098844  0.4323169   1.642 0.100651
factor(bathrooms)6.5   0.7283547  0.4337641   1.679 0.093193 .
factor(bathrooms)6.75  0.8796456  0.4319400   2.036 0.041759 *
```



```
factor(bathrooms)8             NA         NA      NA       NA    
log(sqft_living)        0.6643133  0.0211967  31.340  < 2e-16 ***
log(sqft_lot)           0.0004102  0.0066168   0.062 0.950577    
factor(floors)1.5       0.0271911  0.0167503   1.623 0.104592    
factor(floors)2         0.0608009  0.0134675   4.515 6.50e-06 ***
factor(floors)2.5       0.1647684  0.0494862   3.330 0.000877 ***
factor(floors)3         0.1371832  0.0321032   4.273 1.97e-05 ***
factor(floors)3.5       0.2545955  0.2440104   1.043 0.296829    
factor(view)1           0.2211565  0.0371587   5.952 2.86e-09 ***
factor(view)2           0.1550002  0.0222668   6.961 3.87e-12 ***
factor(view)3           0.2433012  0.0293482   8.290  < 2e-16 ***
factor(view)4           0.4331804  0.0399292  10.849  < 2e-16 ***
factor(condition)2      0.3384287  0.1330278   2.544 0.010991 *  
factor(condition)3      0.6288387  0.1222947   5.142 2.83e-07 ***
factor(condition)4      0.6258520  0.1222691   5.119 3.21e-07 ***
factor(condition)5      0.7027480  0.1228970   5.718 1.15e-08 ***
age                     0.0027479  0.0002414  11.385  < 2e-16 ***
log(income)            -0.0104789  0.0329279  -0.318 0.750319    
factor(ES)4            -0.4511669  0.0432470 -10.432  < 2e-16 ***
factor(ES)5            -0.4528086  0.0454374  -9.966  < 2e-16 ***
factor(ES)6            -0.2602839  0.0410623  -6.339 2.54e-10 ***
factor(ES)7            -0.3035367  0.0410021  -7.403 1.58e-13 ***
factor(ES)8             0.0030379  0.0434997   0.070 0.944326    
factor(ES)9            -0.1381692  0.0453266  -3.048 0.002315 ** 
factor(ES)10           -0.0597771  0.0483471  -1.236 0.216369    
factor(HS)3             0.1214327  0.0306878   3.957 7.71e-05 ***
factor(HS)4             0.1481682  0.0280248   5.287 1.30e-07 ***
factor(HS)5             0.3573392  0.0238985  14.952  < 2e-16 ***
factor(HS)6             0.3387776  0.0296959  11.408  < 2e-16 ***
factor(HS)7             0.2800897  0.0259682  10.786  < 2e-16 ***
factor(HS)8             0.3342394  0.0258634  12.923  < 2e-16 ***
factor(HS)9             0.3626846  0.0311815  11.631  < 2e-16 ***
factor(HS)10            0.3742681  0.0338532  11.056  < 2e-16 ***
factor(MS)4             0.0723914  0.0265686   2.725 0.006461 ** 
factor(MS)5             0.1908313  0.0216580   8.811  < 2e-16 ***
factor(MS)6            -0.0038992  0.0277554  -0.140 0.888283    
factor(MS)7             0.2130140  0.0230018   9.261  < 2e-16 ***
factor(MS)8             0.2877630  0.0269281  10.686  < 2e-16 ***
factor(MS)9             0.2122491  0.0318483   6.664 2.98e-11 ***
factor(MS)10            0.7198147  0.0978497   7.356 2.24e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.297 on 4478 degrees of freedom
Multiple R-squared:  0.7053,    Adjusted R-squared:  0.7005 
F-statistic: 148.8 on 72 and 4478 DF,  p-value: < 2.2e-16
```