Report of Data Mining Course

House Sales Prices Prediction for King County

11720008

沈倪

1 Business Understanding

1.1Research background

Housing price is an everlasting and heated topic in the world. People have made great effort to explore it and there are already quantities of studies about this field. It is full of meaning and interest for us to study this topic. Moreover, the research plan to build some pricing models to analyze and dig out the relationship between the housing price and some relevant factors.

1.2Research target

This research aims to predict the housing price of house sale for King County, which includes Seattle. Our group plan to adopt the homes sold between May 2014 and May 2015.from Kaggle. It is a good dataset for evaluating simple regression models and to test our algorithms. The data have various dimensions, including bedrooms, bathrooms, floors and so on. These variables enable us to predict the housing price more accurate and explain the process and results more clearly.

1.3Research plan

In general, we will do some descriptive statistics to overview the whole data roughly and then do some regressions about the data to find some surface relationship between different variables. What's more, the research plan to utilize the supervised learning and k-nearest neighbors(KNN) algorithm to analyze and mining the data. In detail, the so-called K nearest neighbor algorithm, which is given a set of training data, when input a new instance, it can find nearest K examples of the new instance in the of training data (which is said the most K neighbor). These K instances belong to one class. In other words, The KNN algorithm aims to find the similar houses first and then study the dependent variable. KNN algorithm can be used not only for classification, but also for regression. By finding the nearest neighbors of a sample and assigning the average value of the attributes of these neighbors to the sample, the attributes of the K sample can be obtained. A more useful approach is to give different weights to the impact of the neighbor on the sample. The selection of K value, distance measure and classification decision rule are three basic elements of this algorithm.

In practical applications, a smaller K value is generally selected first and the cross validation method is usually used to select the optimal K value. By cross validation, the algorithm can be trained to predict the variable of numerical type based on the

samples previously seen, and they can also show the probability distribution of the prediction to help the user to explain the prediction process.
For example, you can decide which television to buy according to the relation of its size and prices easily. But it may be more difficult for us to explore the dataset whose price is not simply grows proportionally according to the growth of commodity size or characteristic quantity. Thus, the above algorithms can help us study these datasets better.

2 Data Understanding

2.1Dataset and variables

The dataset includes homes sold between May 2014 and May 2015 and is loaded from https://www.kaggle.com/harlfoxem/housesalesprediction/data
There are 19 house features plus the price and the id columns, along with 21613 observations.

variables	defination	variables	defination
Id	a notation for a house	sqft_living15	Living room area in 2015
Date	Date house was sold	sqft_lot15	lotSize area in 2015
Price	Price is prediction target	condition	How good the condition is
Bedrooms	Number of Bedrooms/Hous	waterfront	house has a view to waterfront
Bathrooms	Number of bathrooms/bedrooms	grade	overall grade given to the house
Floors	Total floors (levels) in house	view	Has been viewed
sqft_living	square footage of the home	yr_built	Built Year
sqft_lot	square footage of the lot	yr_renovated	Year when house was renovated
sqft_basement	square footage of the basement	zipcode	zip
sqft_above	square footage of house apart from basement	lat	Latitude coordinate

Table 1:definition of variables

2.2Descriptive statistics

Firstly, we do some descriptive statistics to overview the whole data roughly and then do some regressions about the data to find some surface relationship between different variables. In order to understand the data better, we can analyze each variable and try to understand what they mean and how relevant the relationship is. We first read the description of each variable. In general,it can be divided into four categories: Housing configuration，Housing condition, Housing location and Housing area. Then we should think about these three issues at the same time: When we buy a house will we consider this factor; If you take it into consideration, how important is this factor; Has the message from this factor appeared in other factors.

Based on experience and common sense, bearing these three questions in mind, we first define subjectively the most relevant variables. We think the housing area and location have a great impact on house prices. In the follow-up study, we will observe whether these variables have the same impact on house prices as we expected. We first analyze the given variables and find that there is no missing value for the data. There are 21613 original records in the original dataset.

We divide the variables into two categories roughly based on the statistics. One is numerical variable, the other is category variable.

numerical	sqft_living, sqft_lot, sqft_above, yr_built, yr_renovated,lat,long sqft_basement,sqft_living15,sqft_lot15,
category	bedrooms,bathrooms,waterfront,view,condition,grade,zipcode

Table 2:categories of variables

	sqft_living	sqft_lot	sqft_above	sqft_basement	sqft_living15
count	21613.000000	2.161300e+04	21613.000000	21613.000000	21613.000000
mean	2079.899736	1.510697e+04	1788.390691	291.509045	1986.552492
std	918.440897	4.142051e+04	828.090978	442.575043	685.391304
min	290.000000	5.200000e+02	290.000000	0.000000	399.000000
25%	1427.000000	5.040000e+03	1190.000000	0.000000	1490.000000
50%	1910.000000	7.618000e+03	1560.000000	0.000000	1840.000000
75%	2550.000000	1.068800e+04	2210.000000	560.000000	2360.000000
max	13540.000000	1.651359e+06	9410.000000	4820.000000	6210.000000

	sqft_lot15	yr_built	yr_renovated	lat	long
count	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000
mean	12768.455652	1971.005136	84.402258	47.560053	-122.213896
std	27304.179631	29.373411	401.679240	0.138564	0.140828
min	651.000000	1900.000000	0.000000	47.155900	-122.519000
25%	5100.000000	1951.000000	0.000000	47.471000	-122.328000
50%	7620.000000	1975.000000	0.000000	47.571800	-122.230000
75%	10083.000000	1997.000000	0.000000	47.678000	-122.125000
max	871200.000000	2015.000000	2015.000000	47.777600	-121.315000

Graph 1: descriptive statistic of numerical variables

The mean square footage of the living area of the house is 2079.90 while the mean square footage of the lot is 15106.97. The mean square footage of house apart from basement is 1788.39, which is near that of the living area. And the mean square footage of basement is about 291.51. In 2015, both the area of the living area and lot room have a slight decrease. It should be point that the year when house was

renovated is merely simply dealt with. It does not make sense to do descriptive statistics for it and we will conduct further analysis for it.

```
mean         3.370842
std          0.930062
min          0.000000
25%          3.000000
50%          3.000000
75%          4.000000
max         33.000000
Name: bedrooms, dtype: float64

mean         2.114757
std          0.770163
min          0.000000
25%          1.750000
50%          2.250000
75%          2.500000
max          8.000000
Name: bathrooms, dtype: float64

mean         1.494309
std          0.539989
min          1.000000
25%          1.000000
50%          1.500000
75%          2.000000
max          3.500000
Name: floors, dtype: float64

count     21613
unique        2
top           0
freq      21450
Name: waterfront, dtype: int64

count     21613
unique        5
top           3
freq      14031
Name: condition, dtype: int64

count     21613
unique       70
top       98103
freq        602
Name: zipcode, dtype: object

count     21613
unique        5
top           0
freq      19489
Name: view, dtype: int64

count     21613
unique       12
top           7
freq       8981
Name: grade, dtype: int64
```

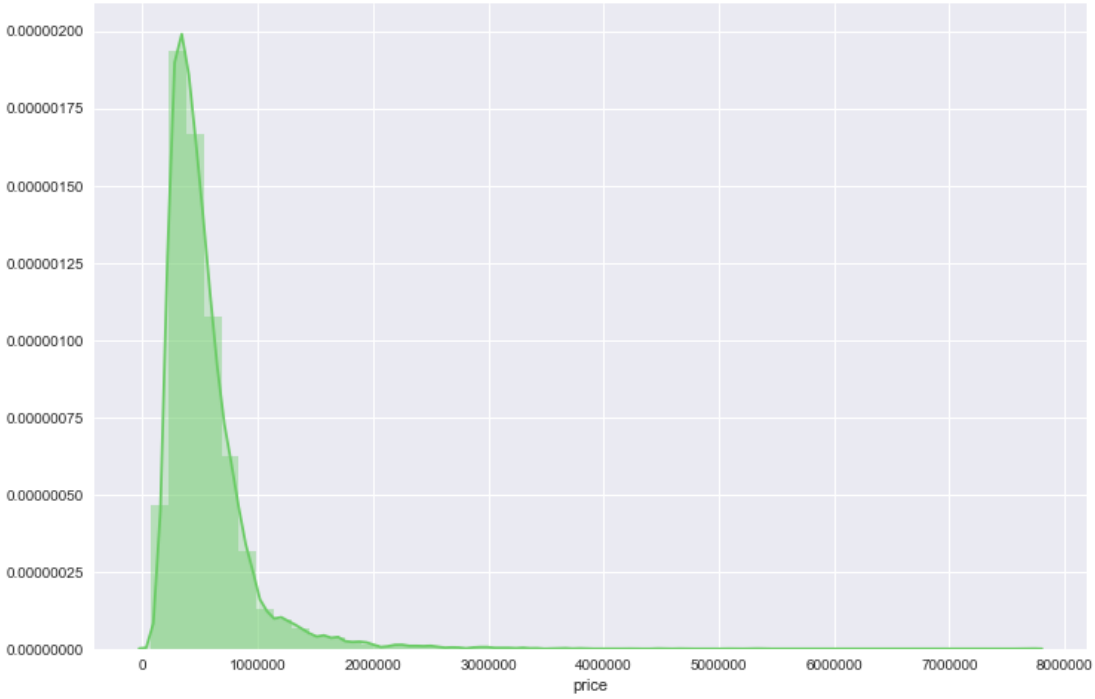
Graph 2: statistic graph of category variables

We can see that in average the houses have 2.11 bathrooms,3,37 bedrooms and 1.49 floors. Only several houses have waterfront.The condition of the houses is ranking from 1 to 5.Most houses get 3. The house has been view 0.23 times in average while the max is four. Interestingly, only taking this variable into consideration, if a house has been viewed more times, the price of it is significantly higher than other houses. The mean overall grade given to the housing unit, based on King County grading system is 7.65 and the mean condition is 3.41.

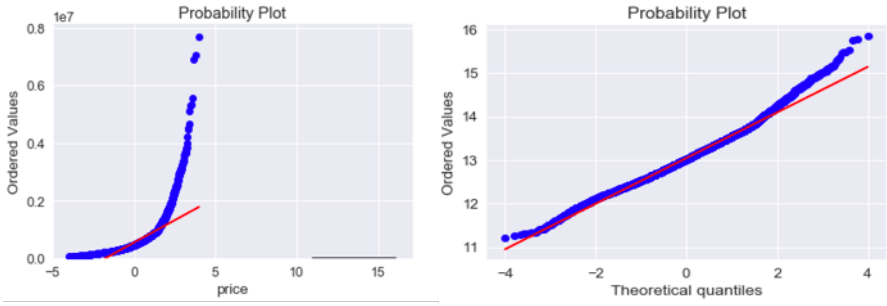
After study the variables, we turn to analyze the house price.The highest price is 7700000 dollars while the lowest house price is 75000 dollars. And the average price is 540088 dollars. The house price volatility is relatively large. In order to observe the pattern of price data distribution, we draw a histogram, we can see the price data deviation from the normal distribution, and the data has a peak. To solve this problem, we can take a logarithm of the price.

Name:Price		
Kurtosis: 34.585540	Std: 3.671272e+05	25%: 3.219500e+05
Skewness: 4.024069	Max: 7.700000e+06	50%:4.500000e+05
Mean: 5.400881e+05	Min: 7.500000e+04	70%:6.450000e+05

Table 3: descriptive statistic of house price



Graph 3：histogram of house price



Graph 4:Probability plot of house price

2.3Graphical description and Correlation

In order to observe the correlation between house prices and house features, we first plot the joint distribution between house prices and other variables. A joint plot is used to visualize the bivariate distribution. There are not only numeric variables but also some categorical variables, we measure the relationship between them using three different correlation coefficients.

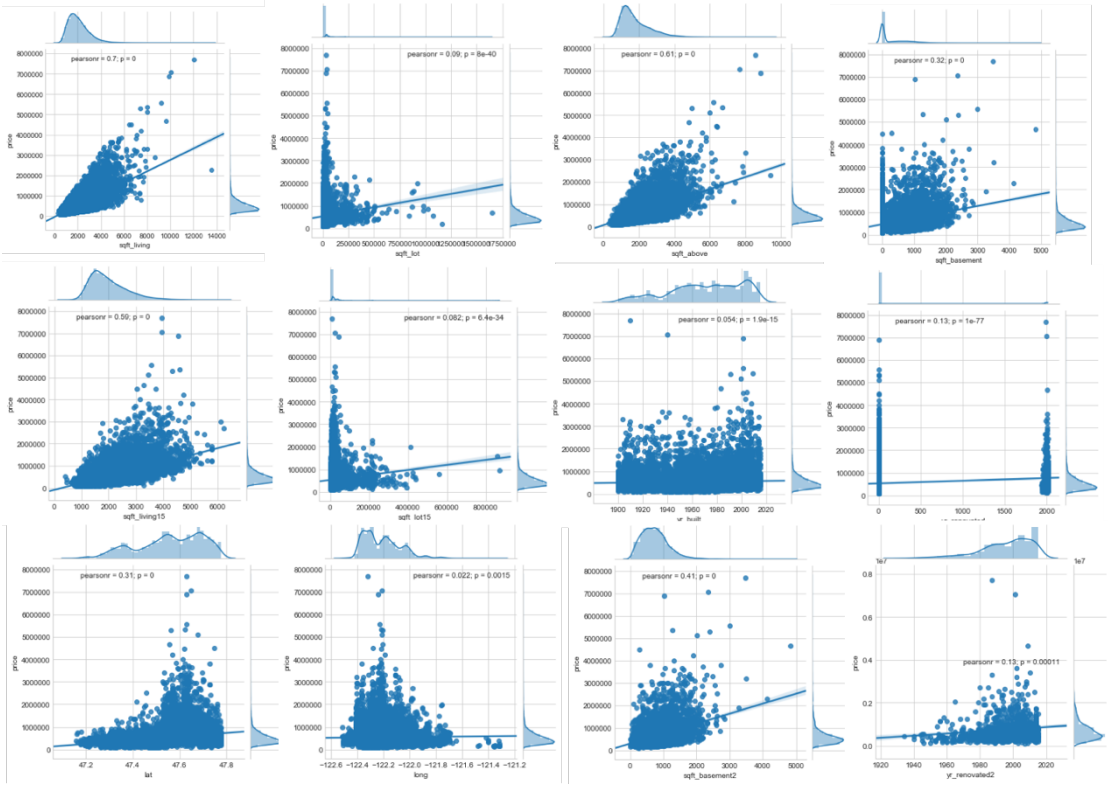
1. Pearson correlation coefficient, usually expressed as r or ρ, is a measure of the correlation between two variables X and Y in the range [-1, + 1]. When dealing with continuous data, normal distribution and linear relationship, it is most appropriate to use pearson correlation coefficient

2. Spearman's correlation coefficient for ranked data is mainly used to solve the problem of the nominal data and sequence data, the spearman correlation coefficient is also used for two sequenced measurement data.

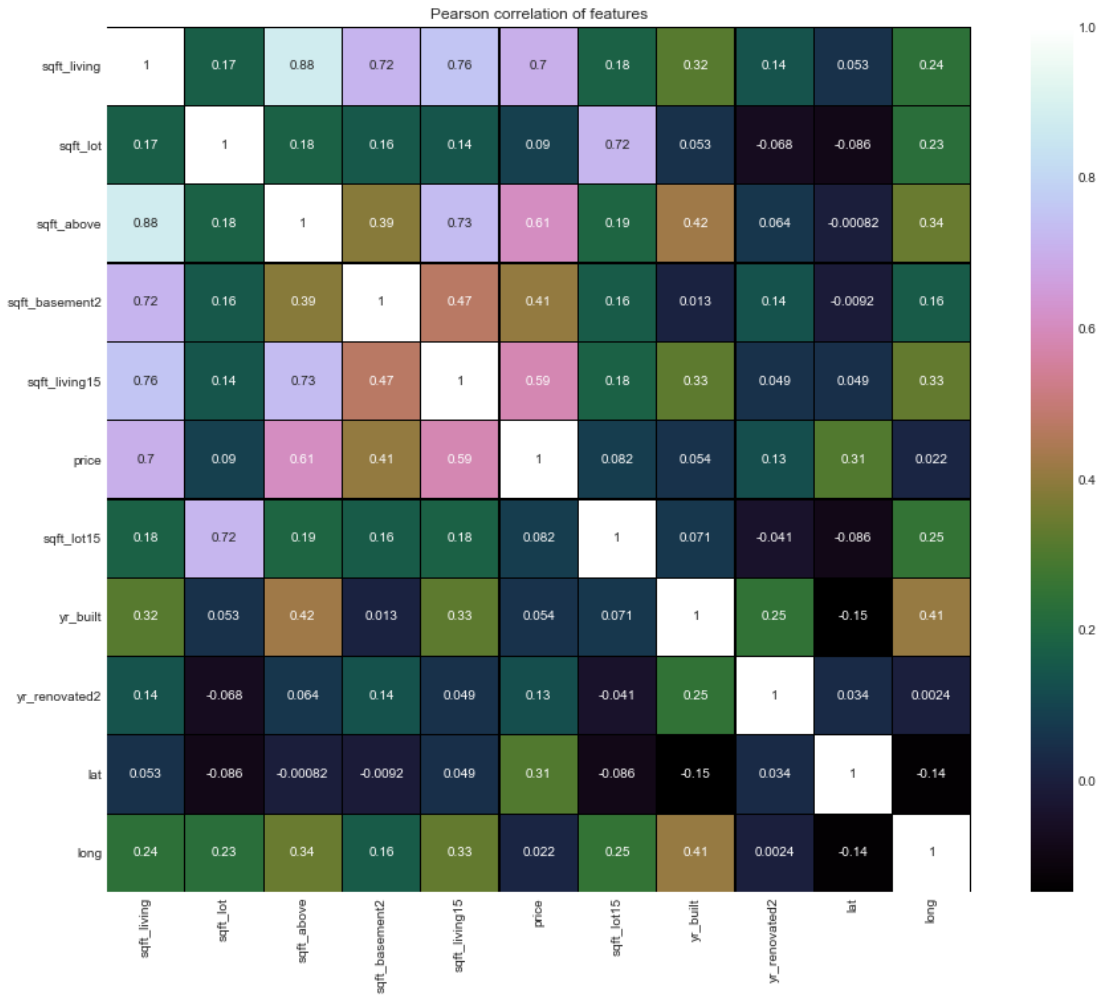
3. Point biserial correlation: measure the relationship between a binary variable, x, and a continuous variable, y.

2.3.1 Pearson correlation coefficient

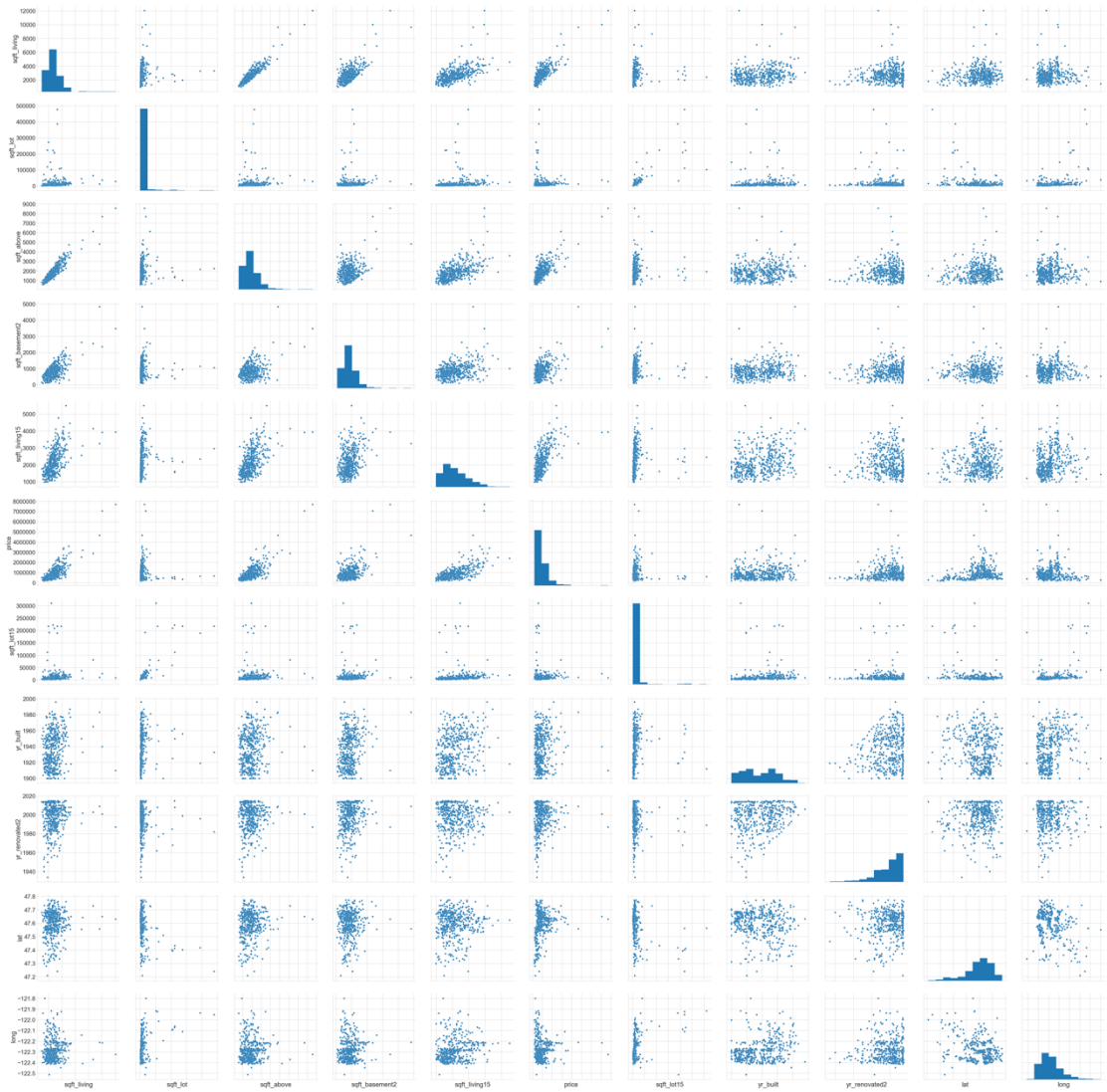
Firstly, we plot the joint distribution of price and sqft_living, sqft_lot, sqft_above, sqft_basement, sqft_living15, sqft_lot15, yr_built, yr_renovated, lat, long separately. It can be seen from the plot that only several houses were renovated and had basement. Therefore , we create two new columns for the analysis of these two features. We pick up those houses do have basement or were renovated to replot the joint distribution.



Graph 5：Pairplot of variables



Graph 6:Pearson Correlation table of features

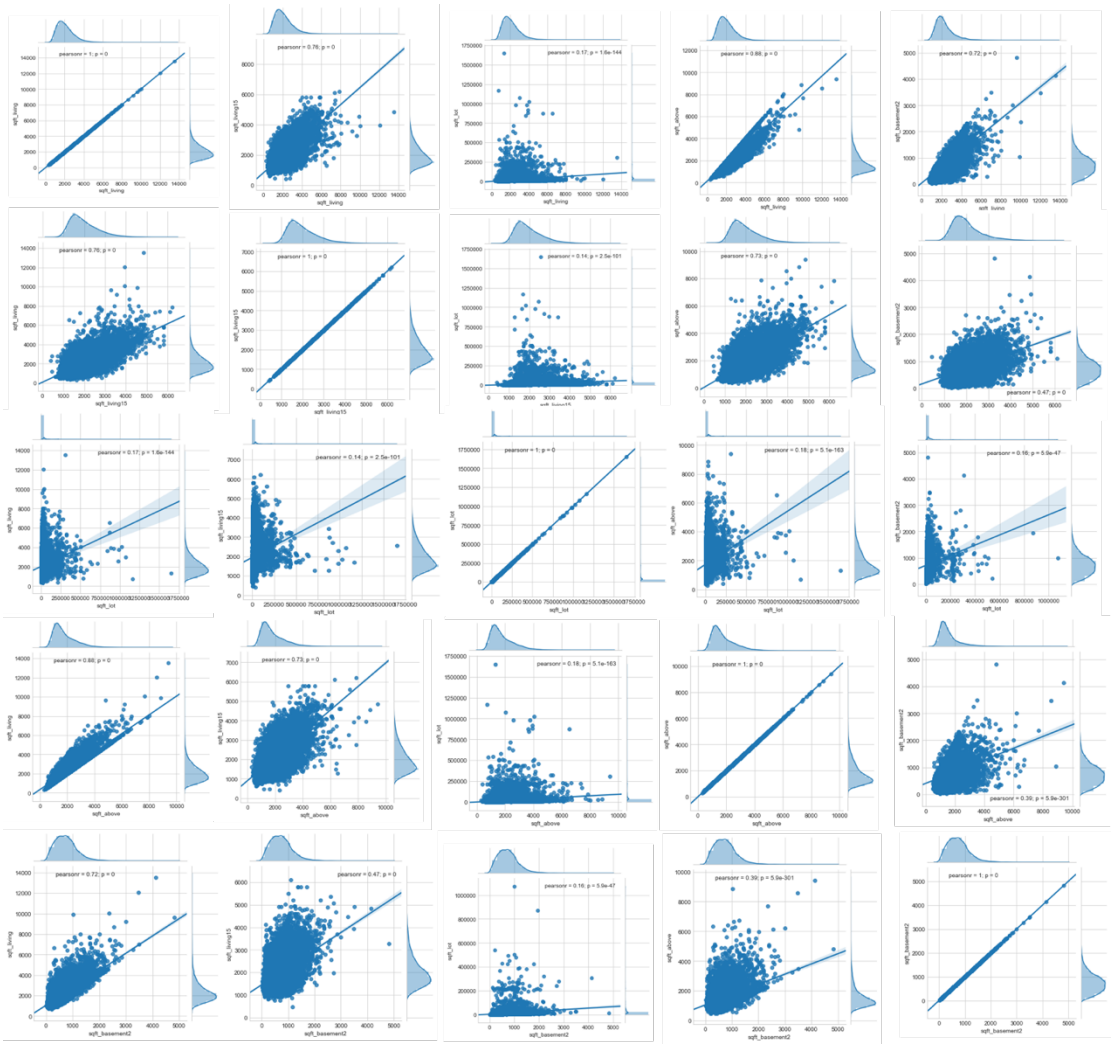


Graph 7:scatter plot of variables

Through observing the coefficient, we initially found that the living area and latitude are relatively strongly related with price, which is consistent with the cognitive, in general, the larger the house and the more convenient the location, the higher the price.

After analyzing the scatter plot and the Pearson Correlation table, we can see the correlation coefficients between sqft_living, sqft_above, sqft_living15, sqft_lot, sqft_basement2, these variables are also high, thus there may be multiple collinearity between these variables. In order to better observe these five variables, we plot the joint distribution for them. The plots show a strong correlation between sqft_living and sqft_above variables. In fact, the degree of relevance reaches a multicollinearity. We can conclude that these variables contain almost the same information. The housing area and the aboveground area are not much different in real life. They are like twins. There is no need to distinguish between them, so we only need one of them for further study.From the plot it can also be inferred that there may not be too many changes for these houses over the time.The housing area in 2015 is similar with the

original. The coefficient is 0.76. Therefore, it is of great chance that there are multi-collinearity among these three variables.



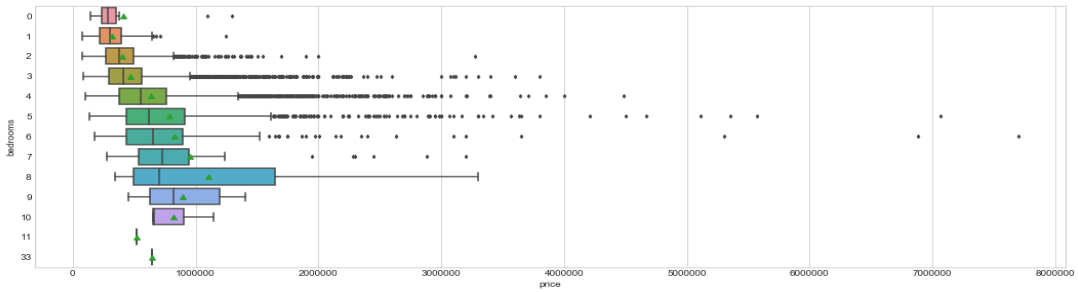
Graph 8: Pairplot of selected variables

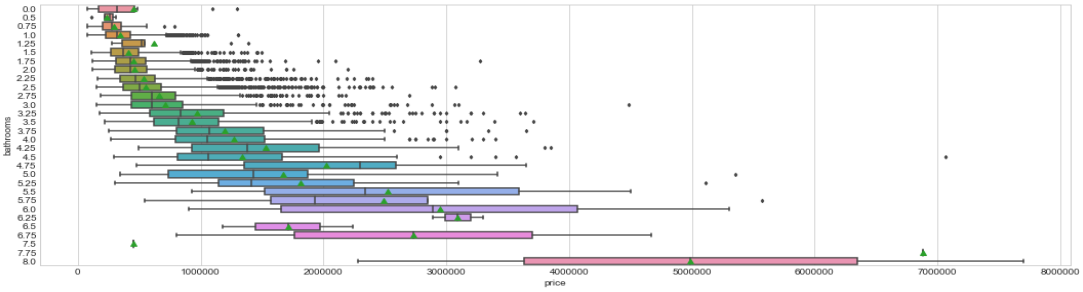
2.3.2 Spearman's correlation coefficient

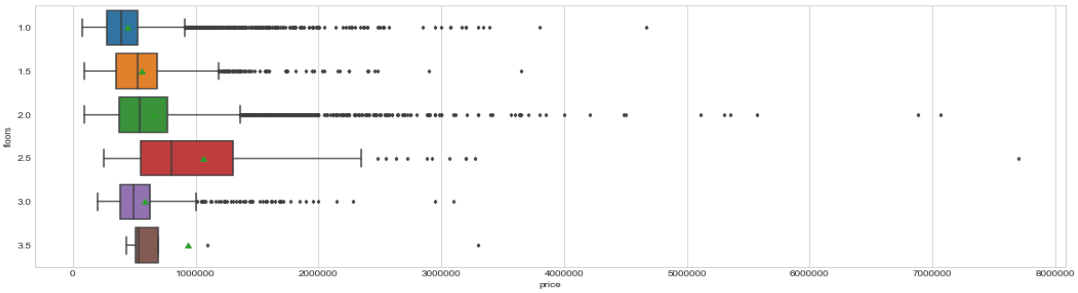
As for the nominal data and sequence data, such as the bedrooms, bathrooms, floors, view, condition and grade. We use Spearman's correlation coefficient to measure their relationship with the house price. Here are the results.

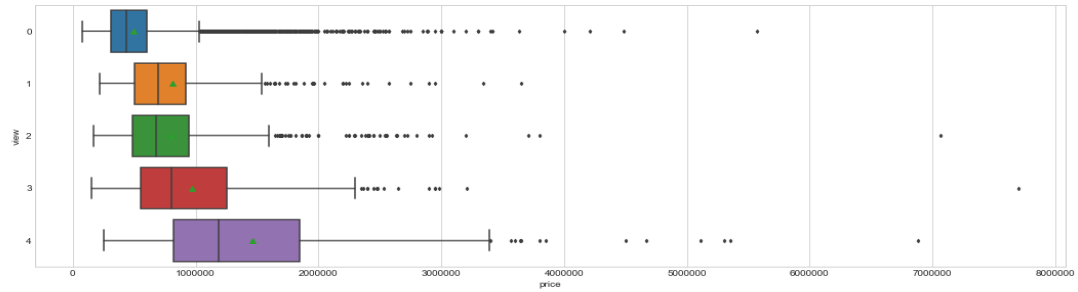
	spearman correlation r	with p
bedrooms	0.34465237096	0.0
bathrooms	0.497160350811	0.0
floors	0.322346550036	0.0
view	0.29393116417	0.0
condition	0.0184899583013	0.00656082840655
grade	0.658215221426	0.0

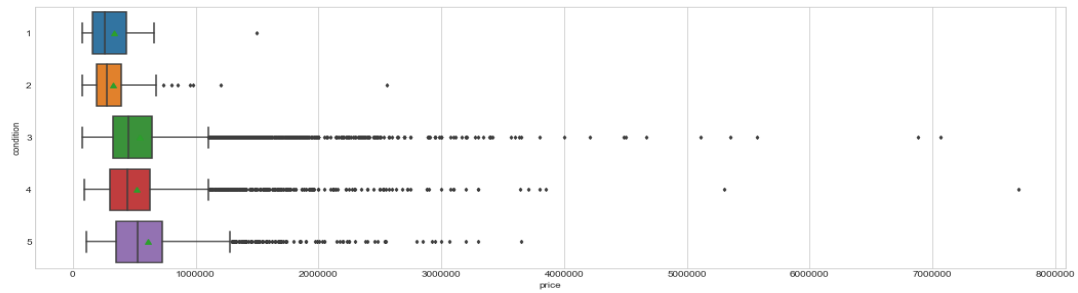
Table 4: Spearman’s correlation coefficient of variables

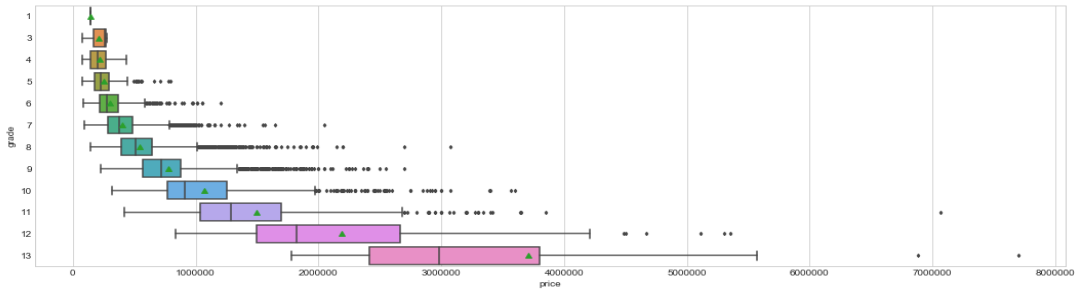

bedrooms
0
1
2
3
4
5
6
7
8
9
10
11
33
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
8000000
price


bathrooms
0.0
0.5
0.75
1.0
1.25
1.5
1.75
2.0
2.25
2.5
2.75
3.0
3.25
3.5
3.75
4.0
4.25
4.5
4.75
5.0
5.25
5.5
5.75
6.0
6.25
6.5
6.75
7.5
7.75
8.0
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
8000000
price


floors
1.0
1.5
2.0
2.5
3.0
3.5
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
8000000
price


view
0
1
2
3
4
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
8000000
price


condition
1
2
3
4
5
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
8000000
price


grade
1
3
4
5
6
7
8
9
10
11
12
13
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
8000000
price

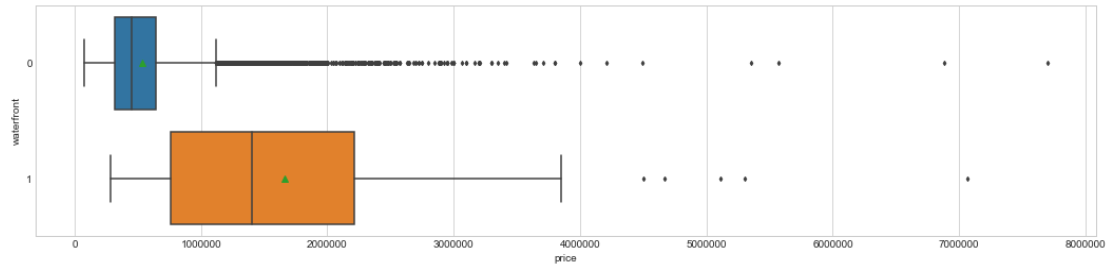
Graph 9:boxplot of category variables

We also plot the boxplot for these variables. It can be seen that the house price and bedrooms roughly has the same distribution trend. The price increase with the number of the bedrooms. It can also be seen that the house price and bathrooms roughly has the same distribution trend. The price increase with the number of the bathrooms. But when the number exceed 6,the price decrease. We do need bathrooms but we don't need too many. When there are about 2.5 floors (levels) in the house, the price is the highest. The remaining number of the floors has little difference. As the view times increase for a house, it was more likely to be sold at a high price. The box diagram shows the relationship between price and condition. It can be seen that the price of a house whose condition is graded in 1 and 2 points is lower than that in condition 3, 4 and 5 points, but the gap in their prices is not very big. As for the grade, it is obviously that the grade has a same trend with the price. The higher the grade, the higher the price. It is reasonable that the price get higher with the grade.

2.3.3 point biserial correlation

Waterfront is a binary variable. Thus we use point biserial correlation to measure the relationship between it and the continuous variable, house price.

The boxplot reveals that the average price of houses who have a view to waterfront is obviously higher. The point biserial correlation r is 0.266369434031 with p = 0.0.



Graph 10:boxplot of waterfront

3 Data Preparation

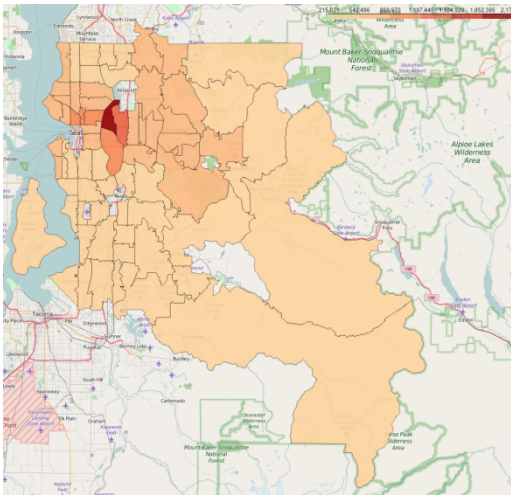
1.1 General Preprocessing

- Firstly, let's have a general data inspection. After dropping repeated samples, the data is as follows. We can see that the data is pretty clean. There are no pesky nulls which we need to treat and most of the features are in numeric format. Let's go ahead and make some transformations on the variables to make our model more reasonable.

```
Data columns (total 21 columns):
id               21613 non-null int64
date             21613 non-null datetime64[ns]
price            21613 non-null float64
bedrooms         21613 non-null int64
bathrooms        21613 non-null float64
sqft_living      21613 non-null int64
sqft_lot         21613 non-null int64
floors           21613 non-null float64
waterfront       21613 non-null int64
view             21613 non-null int64
condition        21613 non-null int64
grade            21613 non-null int64
sqft_above       21613 non-null int64
sqft_basement    21613 non-null int64
yr_built         21613 non-null int64
yr_renovated     21613 non-null int64
zipcode          21613 non-null int64
lat              21613 non-null float64
long             21613 non-null float64
sqft_living15    21613 non-null int64
sqft_lot15       21613 non-null int64
dtypes: datetime64[ns](1), float64(5), int64(15)
```

1.2 Data transformation

- In order to make use of date information, we calculate the years houses have been built, which using variable "house_age" to represent. And also we drop variable "yr_built". Also, we transform"date"into"year" and "month", then drop "date".
- Next we consider to deal with the location information. As we know the location of houses is an important factor for prices. From this heat map we can see that the prices are not linearly related to "longitude" and "latitude". So we replace variables "long" and "lat" with the square of them (sqrt(long), sqrt(lat))and also a product of them(long*lat).
- Not all houses have basement and have been renovated, so variables "basement_present" and "yr_renovated" have many 0 in samples. To make regression performs better, we generate two boolean variables "basement_present" and "renovated" to show is one house has basement and if has been renovated.



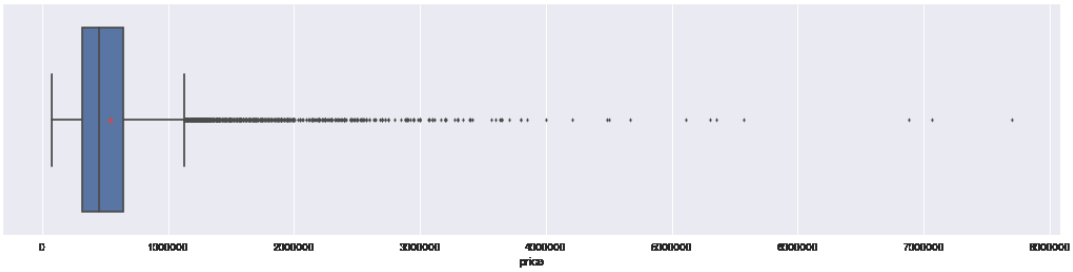
1.3 Encode categorical variables using dummies

- For regression model, we shall encode categorical variables using dummies. Here we choose "floors" and "condition".
- We have 70 zipcodes, so encoded all zipcodes will add 70 dummies variables. Instead, we will only encode the 6 most expensive zipcodes as shown in the map.

```
'zipcode#98004','zipcode#98102','zipcode#98109','zipcode#98112','zipcode#98039','zipcode#98040'
```

1.4 Extract outliers

There seems to be a lot of outliers at the top of the distribution, with a few houses above the 5000000$ value. Here we extract the samples whose prices are more than 1130000(which is calculated according to (Q3-Q1)*1.5+Q3). Now the size of sample is 20467.

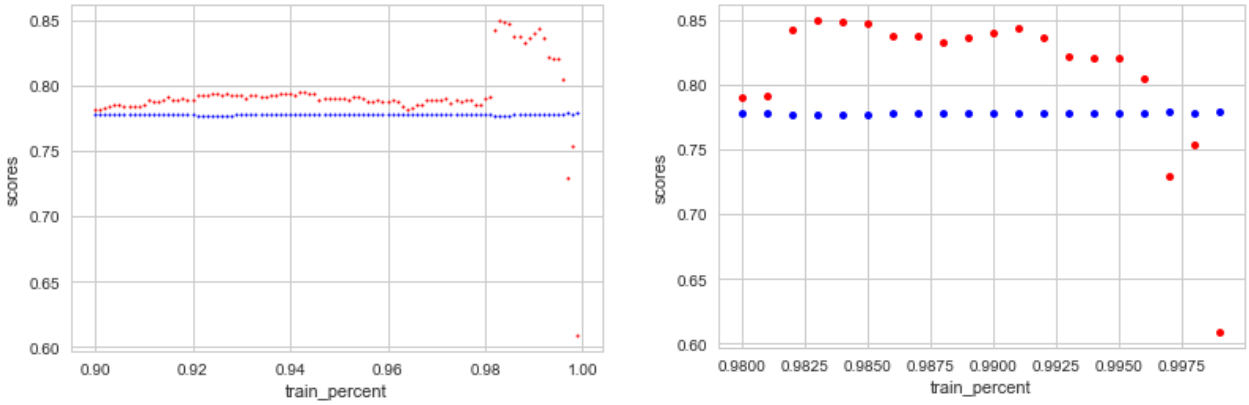


Graph 11:boxplot of houseprice

4 Modeling and Evaluation

4.1 Linear Regression

In order to make a general exploration on this data, let's firstly do linear regression using all features in the data. In order to get best result, here we experiment different size of training data. The result is as follows.



Graph 12:result of the linear regression

We can see that for this data, best percent of training data lies on 0.983, so we set alpha as this value and below is the result of linear regression.

```
||------------linear regression----------------||
MSE: 22697853365.0
RMSE: 150658.067706
score= 0.8496098813
```

The coefficients are:

```
bedrooms -28031.6072187
bathrooms 28037.4697631
sqft_living 129040.852482
sqft_lot 8121.34828578
waterfront 52165.720375
view 41439.4494204
grade 94141.7071762
sqft_above 35359.5309503
lat -794583.331503
long 13056747.9645
sqft_living15 8756.20338297
sqft_lot15 -6625.10406222
month 5684.0465391
year 18380.5261651
house_age 43556.2584002
lat_2 -20355439.6821
long_2 4671870.73157
lat_long -23867488.7695
basement_present -5474.70049139
renovated 6175.47489571
```

```
floors#1.0 -9.50705021616e+15
floors#1.5 -5.39662179902e+15
floors#2.0 -9.23505265824e+15
floors#2.5 -1.63886477962e+15
floors#3.0 -3.15778095931e+15
floors#3.5 -3.68925806883e+14
condition#1 -1.42790518083e+16
condition#2 -3.35807320649e+16
condition#3 -1.81511178333e+17
condition#4 -1.67407420251e+17
condition#5 -1.02531420693e+17
zipcode#98004 66234.0507647
zipcode#98102 15207.2870254
zipcode#98109 14277.6958736
zipcode#98112 36944.4706851
zipcode#98039 51841.7777203
zipcode#98040 33292.1513258
```

Because we train all features, some coefficients seem to be out of our common sense, like “sqrt_lot15”, “basement_present”. This may be caused by some variables who have strong correlation with each other.

According to the ‘Pearson Correlation of features’ graph, we can find that variables “sqrt_living”, “sqrt_living15”, “sqrt_above” and “bathrooms” has relatively strong relation with each other. Also we surveying the result of linear regression we did before, we guess possible combinations of variables to do regression to make a better performance.

After many tries, we find that when we drop variables “sqft_living15”, “sqft_lot15” and “basement_present”, the test score can increase from 0.849 to 0.851, little improvement, however.

But when we try to transform all the zipcodes to dummy variables, we excitedly find that the test score can achieve 0.869! Which means zipcode is an important vector for price prediction. That is make sense according to our common sense. Location plays a main role in price dependence.

4.2 Feature Elimination

After we encode “zipcode” using dummies, there are totally 99 features. In order to make a feature selection, we try two method: RFECV(recursive feature elimination with cross validation) and LassoCV (lasso regression with cross validation).

a) RFECV(Recursive feature elimination with cross validation)

When we set the number of features to remove at each iteration larger than 1, no features will be eliminated. When it equals 1, we can see below that feature “condition#5” is extracted. However, that doesn’t make sense for us. So we try another method, Lasso regression.

b) LassoCV (Lasso regression with cross validation)

LassoCV’s result suggests us to eliminate following 7 features. It seems a little better. But the result can’t offer us the more understanding on the data.

```
||-----------RFECV----------------||
Total number of features:  99
Optimal number of features : 98
test score=  0.869328212814
new_col= ['condition#5']
```

```
||-----------lasso cv regression---------------
MSE: 19893217470.3
RMSE: 141043.317709
test score=  0.868192674938
best alpha= 257.435167368

Features to be eliminated:
long
long_2
```

4.3After encoding “zipcode” using dummies

Since the feature elimination doesn’t work perfectly, we still use current 99 features to do normal linear regression and ridge regression with cross validation.

- Linear regression

```
||------------linear regression----------------||
MSE: 19711901559.0
RMSE: 140399.079623
test score= 0.869394027373
```

- RidgeCV(Ridge regression with cross validation)

It shows that when alpha=0.00168, test score can reach 0.869.

```
||------------Ridge CV regression ----------------||
MSE: 19726975383.3
RMSE: 140452.751427
test score=  0.869294152104
best alpha= 0.000168
```

Conclusion: According to the result of linear regression, we eliminate some strong related features. And we also find that encode all zipcodes using dummies will improve model’s performance dramatically. That tells us maybe we can try some other models such as decision trees, combining ensemble method.

4.4 Decision tree regression

Here we do decision tree regression for the following two reasons:

- When we do linear regression, we find that many categorical variables contribute much to the prediction, such as “zipcode”. And decision tree regression allows categorical ones as input.
- After the decision tree regression, we can get the importance of these features, thus can get to know the structure of data better.

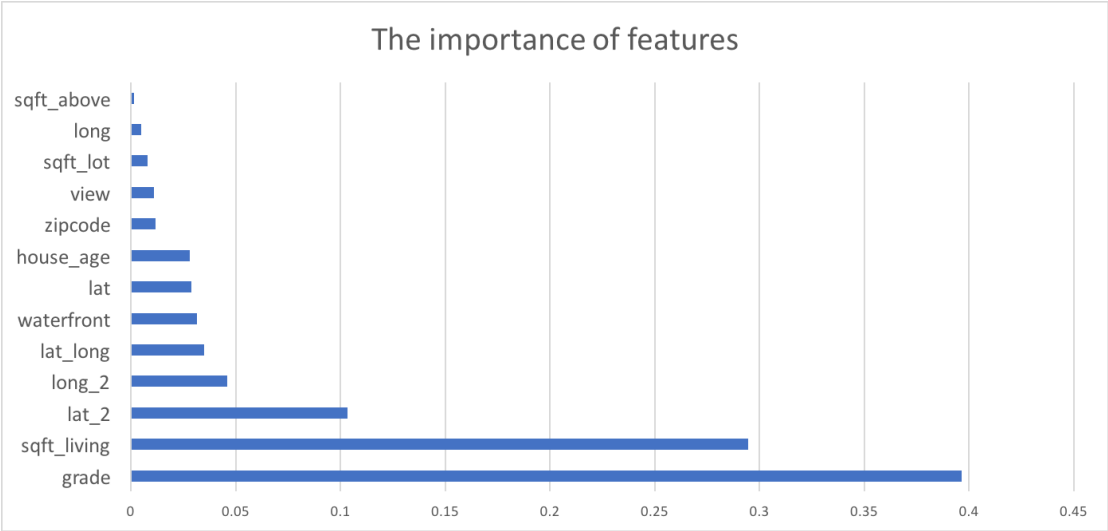
4.4.1 Decision tree regression

Firstly, we just do the decision tree regression without ensemble method. Here we use method GridSearchCV() from sklearn.model_selection to choose a proper parameter. After our adjustment, we get the following result:

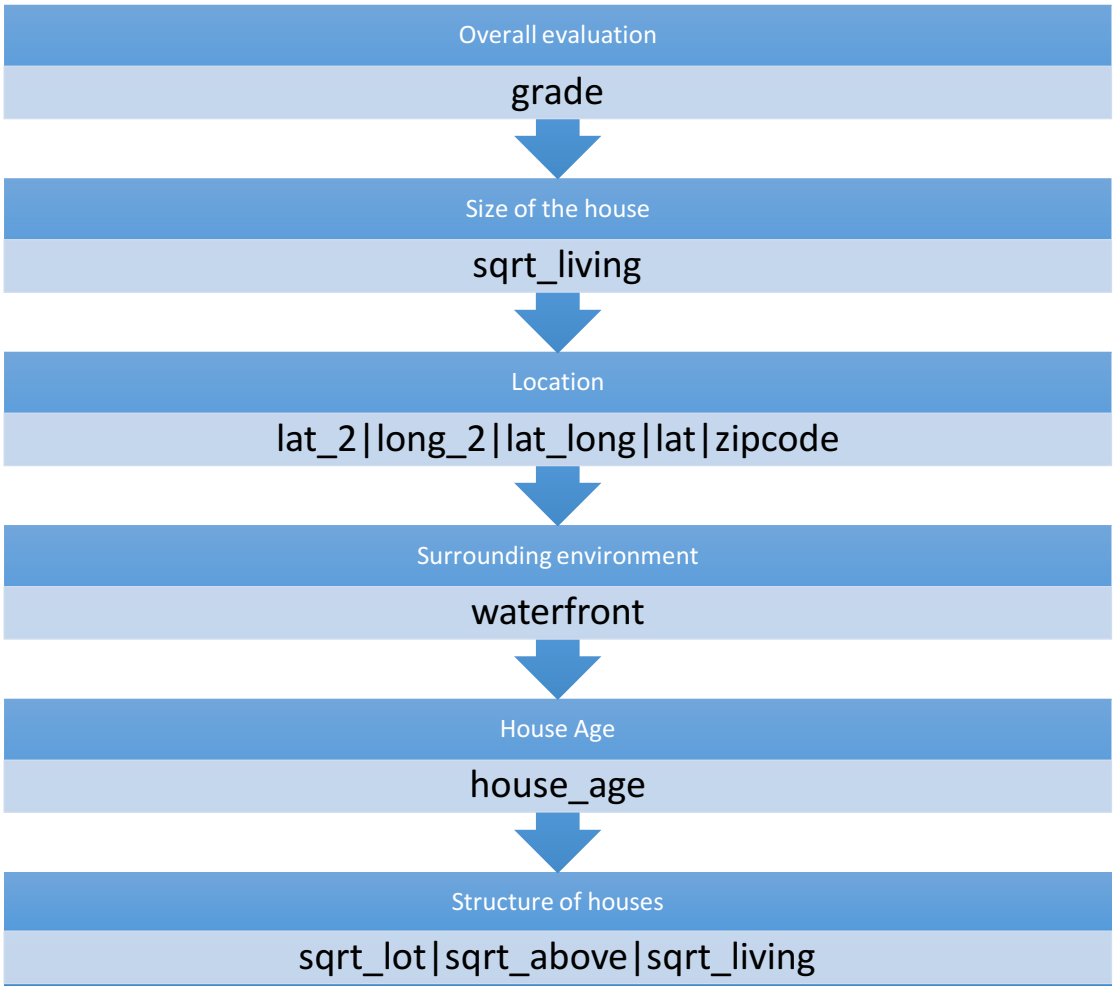
```
_depth=10,max_features='auto',max_leaf_nodes=88,min_impurity_decrease= 83,min_impurity_split= 72, min_samples_leaf=9, min_samples_sp
```

```
||------------decision tree regression ----------------||
MSE: 24190863018.8
RMSE: 155534.121719
test score=  0.796585985671

('grade', 0.39636300602571539)
('sqft_living', 0.29452633157175478)
('lat_2', 0.095014185641427717)
('long', 0.059876587416615994)
('lat_long', 0.046021692416278355)
('lat', 0.03676490655160207)
('waterfront', 0.031602964331509768)
('house_age', 0.011744422747740429)
('zipcode', 0.010939740619591201)
('view', 0.007772542339825249)
('sqft_lot', 0.0048511975700312833)
('long_2', 0.0030340103967464393)
('sqft_above', 0.0014884123711613941)
('bedrooms', 0.0)
('bathrooms', 0.0)
('floors', 0.0)
('condition', 0.0)
('month', 0.0)
('year', 0.0)
('renovated', 0.0)
```



Although the test score(R^2) is not as high as linear regression, this result helps us to identify the important features. And we can consider there are 7 features who has the parameter 0 seems like to be less important. Given by this graph above, we can also draw a factor flow which can represent the dependent level for house prices.



It seems to make sense, right? Next, in order to make a more precise prediction, we choose to try two ensemble methods, bagging and random forest.

4.4.2 Bagging regression (base estimator=decision tree regressor)

To enhance our prediction performance, we firstly try to do regression using bagging. In order to verify that the flow analysis we do in decision tree regression makes sense, we do bagging with and without dropping the 7 less important variables.

For parameter adjustment, we focus on the parameter we adjust in decision tree regression and also the 'n_estimators' for bagging. We do the adjustment separately for two sets of features. Below is the result.

```
||-----------bagging regression---------------||
MSE: 12177293978.1
RMSE: 110350.776971
test_score= 0.897604634864
```

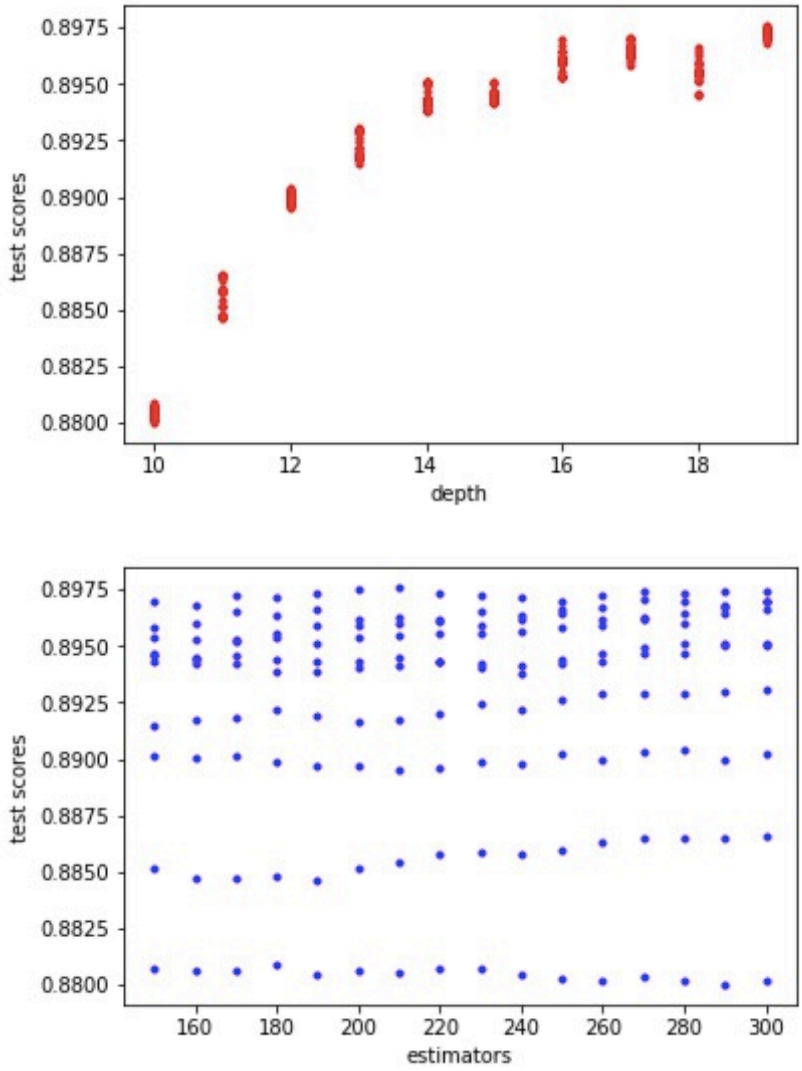
```
||------------bagging regression dropping 7 features----------------||
MSE: 11853012860.4
RMSE: 108871.542932
test_score= 0.90033142158
```

After dropping the 7 less important features we analyzed before, the test score increases to 0.9, which shows that the analysis before seems reasonable.

4.4.3 Random forest regression

Because the bagged trees are correlated, when bootstrap samples are correlated, the benefit of bagging decreases. So here we try to use random forest regression to avoid this situation.
We mainly focus on two parameters for this algorithm, and finally we get a best test score of 0.897, when n_estimators=200, and max_depth=19. (Here we only 4 of 7 variables:month, year, condition, renovated)

```
||------------random forest regression----------------
('test score= ', 0.89618162391314349, 'n_estimators= ', 200, 'max_depth= ', 17)
('MSE:', 12346524515.694073)
('RMSE:', 111114.91581103805)
||------------random forest regression----------------
('test score= ', 0.89535610109698893, 'n_estimators= ', 200, 'max_depth= ', 18)
('MSE:', 12444699213.392963)
('RMSE:', 111555.81210045922)
||------------random forest regression----------------
('test score= ', 0.89747748398047278, 'n_estimators= ', 200, 'max_depth= ', 19)
('MSE:', 12192415304.076235)
('RMSE:', 110419.2705286366)
```



Graph 13：result of random forest regression

4.5 ANN and kNN

Finally, we also try to train our data with ANN(Artificial neural network) and kNN(k-nearest neighbors algorithm).

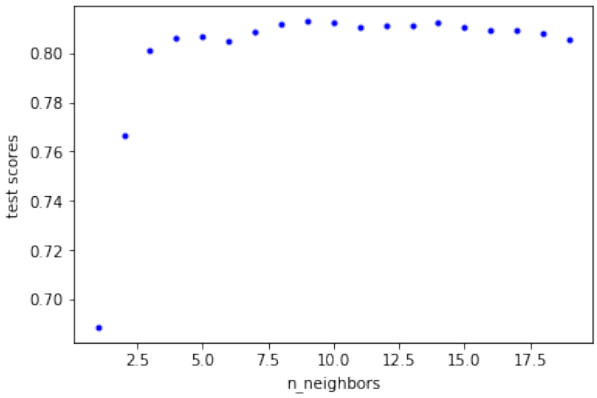
4.5.1 ANN

There are too many parameters for us to adjust. Also, we observe that if we want to get a not bad prediction outcome, it will spend too much time compared with other method. (hidden layer>1) So finally, we give up the parameter adjustment of ANN.

```
||------------MLP regression----------------
MSE: 32731957135.7
RMSE: 180919.753304
test score=  0.724766380073
```

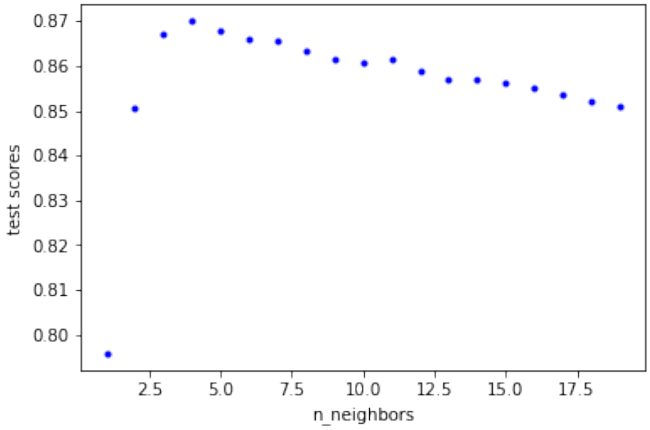
4.5.2 k-nearest neighbors algorithm

kNN is sort of non-parametric method for regression. Here we only consider the parameter “n_neighbors” and “weight” to find a best model.




```
n_neighbors= 9
RMSE: 149228.965536
test score=  0.812743990047
```

However, when we drop the 7 less important features, we can get a much better result.




```
n_neighbors= 4
RMSE: 124364.234653
test score=  0.869946959082
```

We guess it’s performance will get better when the amount of samples grows up.

5 Conclusion

For the group assignments of this course, we first identify the problems to be studied. This is the most important task. It is the first step to solve the problems by clearly defining the problems. The last thing we need to do is predict the price. The dataset comes from the Kaggle website.

After going through the dataset, we first make a philosophical analysis of the variables, and combine the practical meaning to get a business understanding of the variables. We classify them roughly according to the characteristics of the houses, and then combine the attributes of different variables with the statistical correlation knowledge, dividing them into numerical variables, categorical variables and so on. Not only did we conduct a separate analysis of Price, we also analyzed the most relevant variables. A preliminary analysis of the relationship between the independent variables was also made, and some eigenvalues were selected for the subsequent training of the prediction model. After pre-processing the data, the next step is to set up a model to predict the price of the house, and to establish a certain index for evaluation. We did not have any benchmark for initial training, so we decided to use as many models we studied to predict this dataset. These methods include simple linear regression, Lasso regression, decision tree, clustering algorithm and so on. We judge the merits and demerits of our model by distinguishing the training set and the forecasting set, observing the fitting score and the error of the prediction.

Through this assignment, we first realized that it is very important for mutual trust and interaction among team members to work together to get process. Second, for the job itself, the knowledge learned in the classroom can be applied to practice. And we can also learn how to launch and complete an analysis of the dataset by setting up a scientific framework. We should make reasonable plans. As this data analysis will inevitably have a lot of charts, we should make reasonable arrangements for the use of visual tools. The data can offer information and support to us, thus it is important to learn to extract practical significance from the data. Our mastery of data analysis tools also gets improved. We also have in-depth study of some algorithms. Pretreatment of data is very important, we should try different algorithms and models step by step in order to gradually get better results. We should be patient and always eager to learn new method and knowledge

In conclusion: It was an exciting time and an era of big data. Social media has given us more and more insight into the complicated patterns of human behavior from the data. Data-based technologies determine the future of people, but not the data itself has changed our world, and it is the decisive part of our growth in available knowledge