Predicting House Price with History Data

Ver 1.0

Tao Chen

2019/04

Revisions

Date	Reviser	Comment
2019-4-26	Tao Chen	Initial release.

Table of Content

1. Introduction

1.1. Background

The real estate market in The United States is very open and vigorous , which is reflected in the data in some news and websites. Many real estate investors have a strong sense of risk, they are trying to use a scientific way to lower the risk. One of the most common ways is to use data science to help do some predictions of the real estate market, the more insights we get from the data, the more accurate the prediction would be, the lower risk would be. For example, It is necessary to study the house sales data, on the basis of the data science methodology, the prediction model of house price is designed, which will help give a reasonable price when the decision-making of sale/buy is needed.

1.2. Problem

If we collect some history data of house sales, for example , the location of the house , the size of the house ,number of rooms of the house, the neighborhood of the house etc. , and we also know the sales price of the house, can find out some relationships between house price and house properties with the help of modern data science? The aim is to get some insights from the history data, and the, when we want to buy a house , we can come up with a reasonable price according to the features we have for the house.

1.3. Analytic Approach

In the context of the business background and business problem definition, we can identify the type of patterns will be needed to address the question most effectively. The question is to determine price of a house, so a predictive model might be used. We can use the history sales data to build a predictive model , as the target of this model is the house price , which is a continuous variable , a regression approach is selected.

2. Data acquisition

In order to build a predictive model , first, we need some house sales data, luckily there is an open dataset that we can download from the web, we will use it as a base dataset for this assignment.

Also, I think the neighborhood of the house is very important to the sales price, a good neighborhood might contribute to a good price.So Foursquare API is used to get neighborhood data as a complement for this dataset.

2.1. House Sales Dataset

In this dataset that I downloaded from the web , I will predict the sales price of houses in King County, Seattle. It includes homes sold between May 2014 and May 2015.The dataset is in CSV file format. Before doing anything we should first know about the dataset to get what it contains and the structure of data.

As described in the dataset introduction web page, the dataset contains 20 house features plus the price, along with 21613 observations.

The description for the 20 features is in the table below:

#	Feature	Description
1	id	It is the unique numeric number assigned to each house being sold.
2	date	It is the date on which the house was sold out.
3	price	It is the price of house which we have to predict so this is our target variable and apart from it are our features.
4	bedrooms	It determines number of bedrooms in a house.
5	bathrooms	It determines number of bathrooms in a bedroom of a house.
6	sqft_living	It is the measurement variable which determines the measurement of house in square foot.
7	sqft_lot	It is also the measurement variable which determines square foot of the lot.
8	floors	It determines total floors means levels of house.
9	waterfront	This feature determines whether a house has a view to waterfront 0 means no 1 means yes.
10	view	This feature determines whether a house has been viewed or not 0 means no 1 means yes.
11	condition	It determines the overall condition of a house on a scale of 1 to 5.
12	grade	It determines the overall grade given to the housing unit, based on King County grading system on a scale of 1 to 11.
13	sqft_above	It determines square footage of house apart from basement.
14	sqft_basement	It determines square footage of the basement of the house.
15	yr_built	It determines the date of building of the house.
16	yr_renovated	It determines year of renovation of house.
17	zipcode	It determines the zipcode of the location of the house.
18	lat	It determines the latitude of the location of the house.
19	long	It determines the longitude of the location of the house.
20	sqft_living15	Living room area in 2015(implies-- some renovations)
21	sqft_lot15	lotSize area in 2015(implies-- some renovations)

2.2. Get Neighborhood Data with Foursquare API

Foursquare provides detailed location data , we can leverage it to explore the neighborhood that the house located, this might help to determine the house price. Since we have the zipcode and lat and long location in the house price dataset , it’s easy and straightforward to get neighborhood data through Foursquare API.

2.2.1. Foursquare API introduction

We used “Get Venue Recommendations” API to get the neighborhood data for the house, which returned a list of recommended venues near the house location, the name,location,categories of the venues was included.

2.2.2. Get and Preprocess Foursquare location data

As I have sandbox account for Foursquare API only which limits 950 regular calls/pay , and we have more than 20,000 records for this dataset, I can not proceed with the API by one call per record for this dataset, as a workaround , I use the latitude and longitude of average center of the zipcode to get the location data for each house.

After get the venues nearby , we added a new feature VenueCount to the dataset , which stands for the number of venues near the house. might help predict the sales price.

3. Data Cleaning and Transformation

First , I dropped id and date column, it’s useless to the model for now.

Second, the yr_built and yr_renovated columns can tell us how old is the house, to make these two features more easy to understand ans use, I did a feature transformation , convert them to another two features: house_age and house_renew_age, the computation logic is use the year of sale minus yr_built and yr_renovated..

Third, for the house location features: lat,long and zipcode, the combination of latitude and longitude has almost the same impact as zipcode on the price prediction, we can use either of them, for this lab , zipcode is used.

Last, I found some NaN values in the derived feature: VenueCount, that may be the case that the Foursquare API return no venue data for the location, so I did data cleaning to set these NaN values to 0.

4. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is very important for us to understand the data and select the model features. I used some python tools such as pandas, matplotlib, and seaborn to do data distribution analysis, data visualization and data relationship analysis, which would summarize the main characteristics of the dataset and help select the features to build the model.

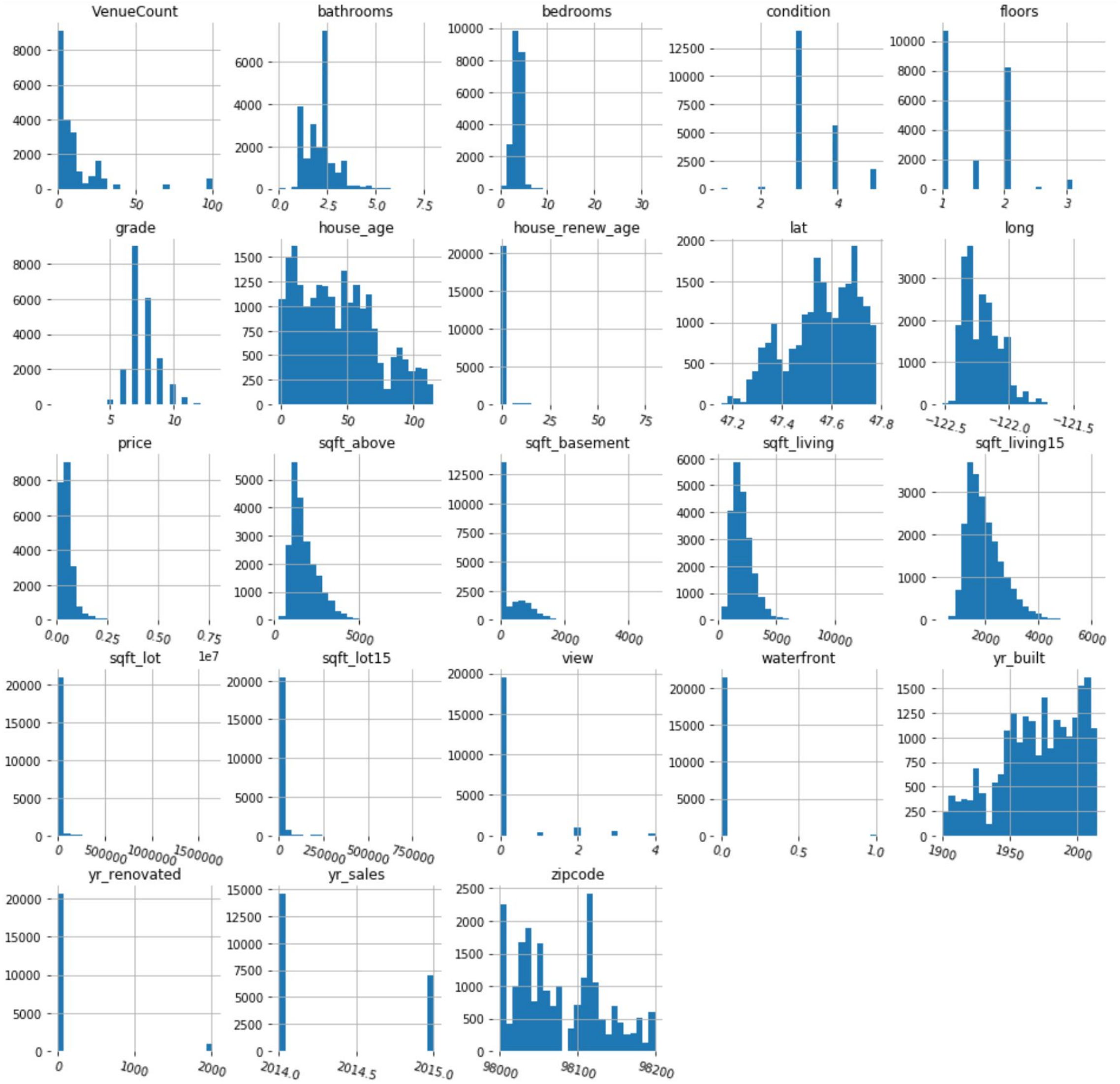
4.1. Overall data description

First ,I used pandas describe function to examine the numeric columns of the dataset.

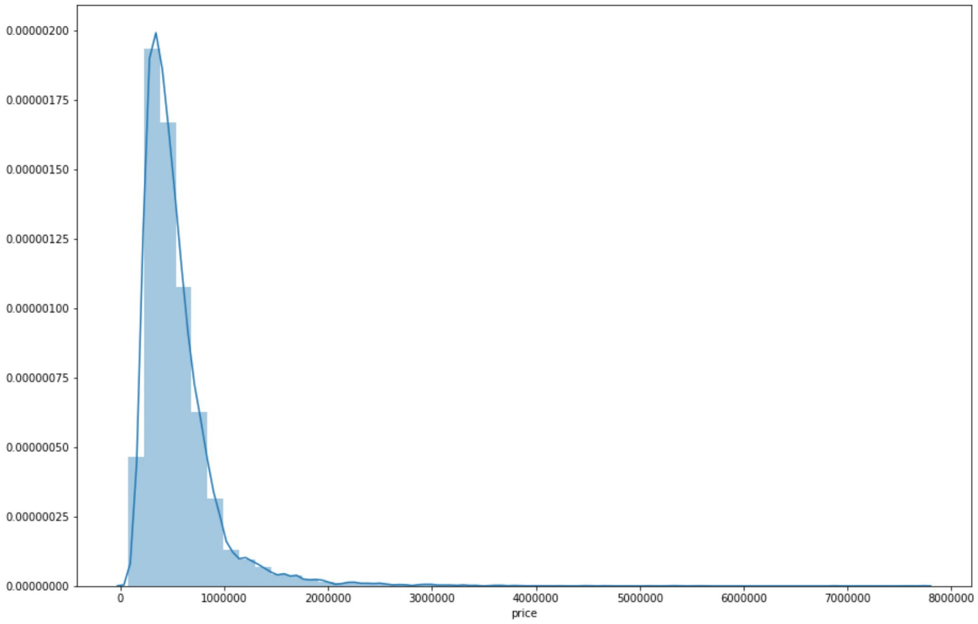
Feature	count	mean	std	min	25%	50%	75%	max

Feature	count	mean	std	min	25%	50%	75%	max
price	21613	540088.142	367127.196	75000.000	321950.000	450000.000	645000.000	7700000.000
bedrooms	21613	3.371	0.930	0.000	3.000	3.000	4.000	33.000
bathrooms	21613	2.115	0.770	0.000	1.750	2.250	2.500	8.000
sqft_living	21613	2079.900	918.441	290.000	1427.000	1910.000	2550.000	13540.000
sqft_lot	21613	15106.968	41420.512	520.000	5040.000	7618.000	10688.000	1651359.000
floors	21613	1.494	0.540	1.000	1.000	1.500	2.000	3.500
waterfront	21613	0.008	0.087	0.000	0.000	0.000	0.000	1.000
view	21613	0.234	0.766	0.000	0.000	0.000	0.000	4.000
condition	21613	3.409	0.651	1.000	3.000	3.000	4.000	5.000
grade	21613	7.657	1.175	1.000	7.000	7.000	8.000	13.000
sqft_above	21613	1788.391	828.091	290.000	1190.000	1560.000	2210.000	9410.000
sqft_basement	21613	291.509	442.575	0.000	0.000	0.000	560.000	4820.000
yr_built	21613	1971.005	29.373	1900.000	1951.000	1975.000	1997.000	2015.000
yr_renovated	21613	84.402	401.679	0.000	0.000	0.000	0.000	2015.000
zipcode	21613	98077.940	53.505	98001.000	98033.000	98065.000	98118.000	98199.000
lat	21613	47.560	0.139	47.156	47.471	47.572	47.678	47.778
long	21613	-122.214	0.141	-122.519	-122.328	-122.230	-122.125	-121.315
sqft_living15	21613	1986.552	685.391	399.000	1490.000	1840.000	2360.000	6210.000
sqft_lot15	21613	12768.456	27304.180	651.000	5100.000	7620.000	10083.000	871200.000
VenueCount	21613	11.198	18.393	0.000	2.000	5.000	10.000	100.000
yr_sales	21613	2014.323	0.468	2014.000	2014.000	2014.000	2015.000	2015.000
house_age	21613	43.318	29.375	-1.000	18.000	40.000	63.000	115.000
house_renew_age	21613	0.780	4.895	-1.000	0.000	0.000	0.000	80.000

then I drew an histogram diagram for all features as below to visualize the overall data distribution.

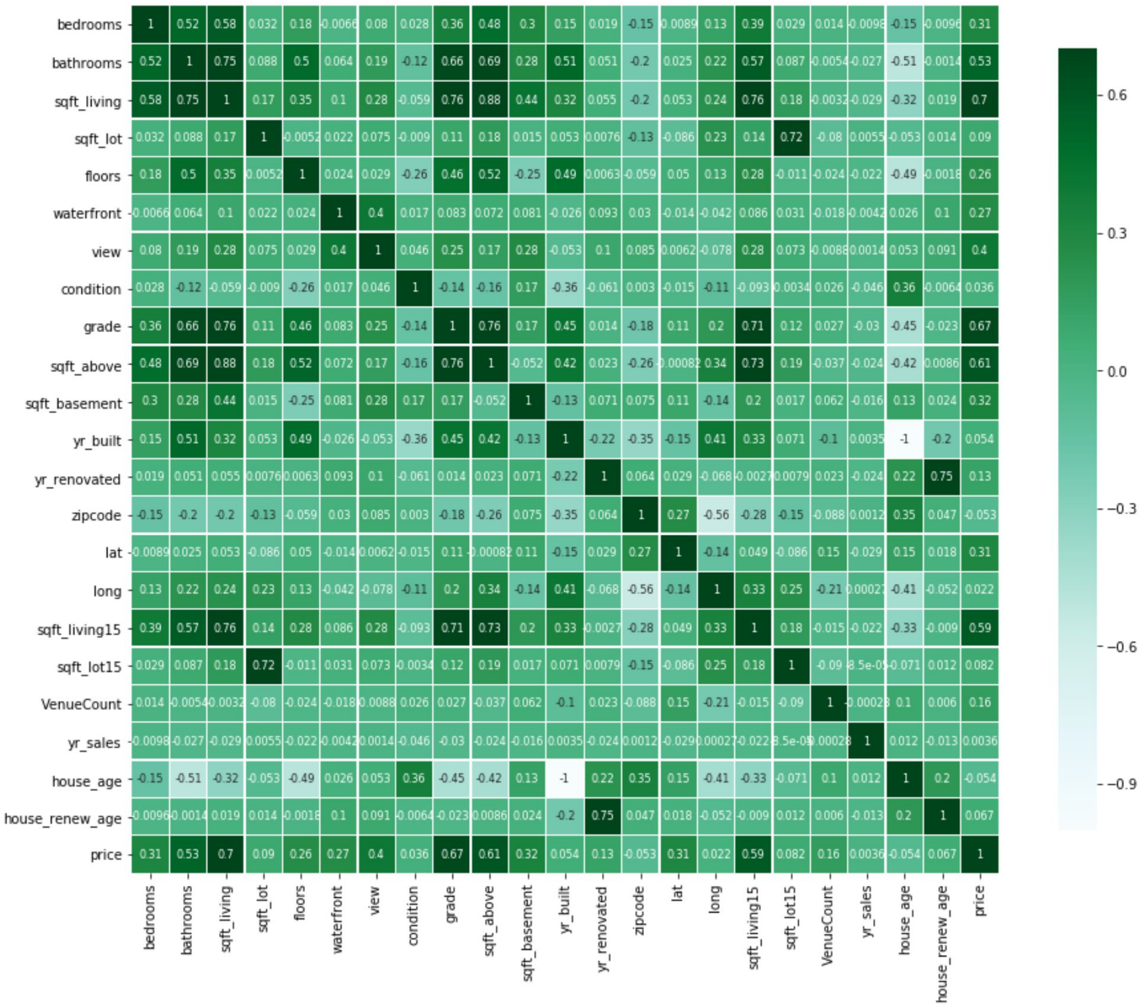


I plot a histogram to see the distribution of the target variable: price , it looks like it’s a skewed normal distribution.



4.2. Relationship Analysis

Data relationship analysis help us understand the relationship between features or feature and target variable.First , I used a heatmap to visualize the correlation matrix, as the dark boxes in the diagram demonstrates, grade, sqft_living, sqft_living15, bedrooms have very strong correlations with target variable price. Some features also have strong correlations with other features, e.g. sqft_above and grade, sqft_living and bathrooms ,and so on.

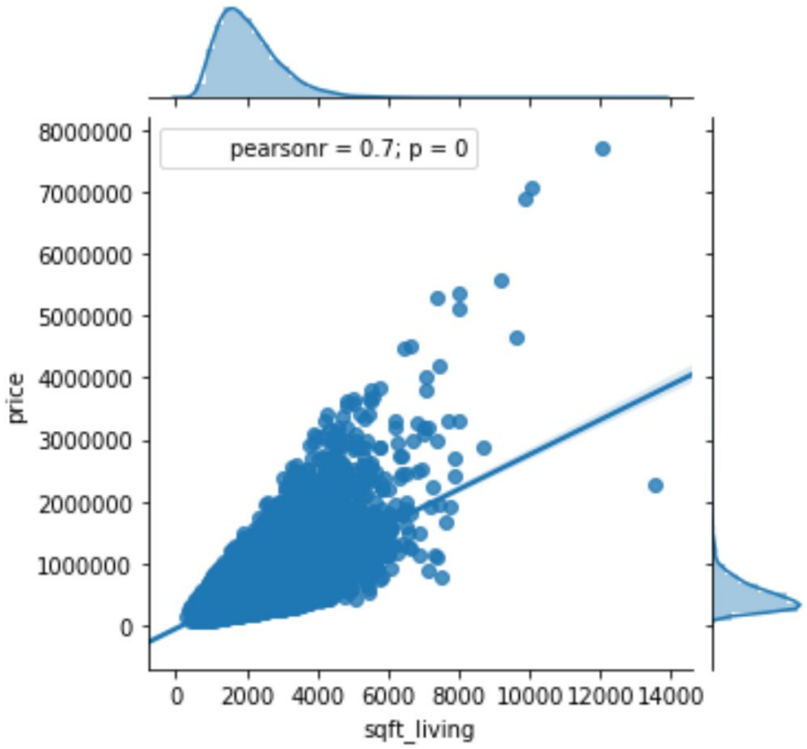


	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15	VenueCount	yr_sales	house_age	house_renew_age	price
bedrooms	1	0.52	0.58	0.032	0.18	-0.0066	0.08	0.028	0.36	0.48	0.3	0.15	0.019	-0.15	-0.0089	0.13	0.39	0.029	0.014	-0.0098	-0.15	-0.0096	0.31
bathrooms	0.52	1	0.75	0.088	0.5	0.064	0.19	-0.12	0.66	0.69	0.28	0.51	0.051	-0.2	0.025	0.22	0.57	0.087	-0.0054	-0.027	-0.51	-0.0014	0.53
sqft_living	0.58	0.75	1	0.17	0.35	0.1	0.28	-0.059	0.76	0.88	0.44	0.32	0.055	-0.2	0.053	0.24	0.76	0.18	-0.0032	-0.029	-0.32	0.019	0.7
sqft_lot	0.032	0.088	0.17	1	-0.0052	0.022	0.075	-0.009	0.11	0.18	0.015	0.053	0.0076	-0.13	-0.086	0.23	0.14	0.72	-0.08	0.0055	-0.053	0.014	0.09
floors	0.18	0.5	0.35	-0.0052	1	0.024	0.029	-0.26	0.46	0.52	-0.25	0.49	0.0063	-0.059	0.05	0.13	0.28	-0.011	-0.024	-0.022	-0.49	-0.0018	0.26
waterfront	-0.0066	0.064	0.1	0.022	0.024	1	0.4	0.017	0.083	0.072	0.081	-0.026	0.093	0.03	-0.014	-0.042	0.086	0.031	-0.018	-0.0042	0.026	0.1	0.27
view	0.08	0.19	0.28	0.075	0.029	0.4	1	0.046	0.25	0.17	0.28	-0.053	0.1	0.085	0.0062	-0.078	0.28	0.073	-0.0088	0.0014	0.053	0.091	0.4
condition	0.028	-0.12	-0.059	-0.009	-0.26	0.017	0.046	1	-0.14	-0.16	0.17	-0.36	-0.061	0.003	-0.015	-0.11	-0.093	-0.0034	0.026	-0.046	0.36	-0.0064	0.036
grade	0.36	0.66	0.76	0.11	0.46	0.083	0.25	-0.14	1	0.76	0.17	0.45	0.014	-0.18	0.11	0.2	0.71	0.12	0.027	-0.03	-0.45	-0.023	0.67
sqft_above	0.48	0.69	0.88	0.18	0.52	0.072	0.17	-0.16	0.76	1	-0.052	0.42	0.023	-0.26	-0.00082	0.34	0.73	0.19	-0.037	-0.024	-0.42	0.0086	0.61
sqft_basement	0.3	0.28	0.44	0.015	-0.25	0.081	0.28	0.17	0.17	-0.052	1	-0.13	0.071	0.075	0.11	-0.14	0.2	0.017	0.062	-0.016	0.13	0.024	0.32
yr_built	0.15	0.51	0.32	0.053	0.49	-0.026	-0.053	-0.36	0.45	0.42	-0.13	1	-0.22	-0.35	-0.15	0.41	0.33	0.071	-0.1	0.0035	-1	-0.2	0.054
yr_renovated	0.019	0.051	0.055	0.0076	0.0063	0.093	0.1	-0.061	0.014	0.023	0.071	-0.22	1	0.064	0.029	-0.068	-0.0027	0.0079	0.023	-0.024	0.22	0.75	0.13
zipcode	-0.15	-0.2	-0.2	-0.13	-0.059	0.03	0.085	0.003	-0.18	-0.26	0.075	-0.35	0.064	1	0.27	-0.56	-0.28	-0.15	-0.088	0.0012	0.35	0.047	-0.053
lat	-0.0089	0.025	0.053	-0.086	0.05	-0.014	0.0062	-0.015	0.11	-0.00082	0.11	-0.15	0.029	0.27	1	-0.14	0.049	-0.086	0.15	-0.029	0.15	0.018	0.31
long	0.13	0.22	0.24	0.23	0.13	-0.042	-0.078	-0.11	0.2	0.34	-0.14	0.41	-0.068	-0.56	-0.14	1	0.33	0.25	-0.21	0.00027	-0.41	-0.052	0.022
sqft_living15	0.39	0.57	0.76	0.14	0.28	0.086	0.28	-0.093	0.71	0.73	0.2	0.33	-0.0027	-0.28	0.049	0.33	1	0.18	-0.015	-0.022	-0.33	-0.009	0.59
sqft_lot15	0.029	0.087	0.18	0.72	-0.011	0.031	0.073	-0.0034	0.12	0.19	0.017	0.071	0.0079	-0.15	-0.086	0.25	0.18	1	-0.09	8.5e-05	-0.071	0.012	0.082
VenueCount	0.014	-0.0054	-0.0032	-0.08	-0.024	-0.018	-0.0088	0.026	0.027	-0.037	0.062	-0.1	0.023	-0.088	0.15	-0.21	-0.015	-0.09	1	-0.00028	0.1	0.006	0.16
yr_sales	-0.0098	-0.027	-0.029	0.0055	-0.022	-0.0042	0.0014	-0.046	-0.03	-0.024	-0.016	0.0035	-0.024	0.0012	-0.029	0.00027	-0.022	8.5e-05	-0.00028	1	0.012	-0.013	0.0036
house_age	-0.15	-0.51	-0.32	-0.053	-0.49	0.026	0.053	0.36	-0.45	-0.42	0.13	-1	0.22	0.35	0.15	-0.41	-0.33	-0.071	0.1	0.012	1	0.2	-0.054
house_renew_age	-0.0096	-0.0014	0.019	0.014	-0.0018	0.1	0.091	-0.0064	-0.023	0.0086	0.024	-0.2	0.75	0.047	0.018	-0.052	-0.009	0.012	0.006	-0.013	0.2	1	0.067
price	0.31	0.53	0.7	0.09	0.26	0.27	0.4	0.036	0.67	0.61	0.32	0.054	0.13	-0.053	0.31	0.022	0.59	0.082	0.16	0.0036	-0.054	0.067	1

4.2.1. Relationship between house area and price

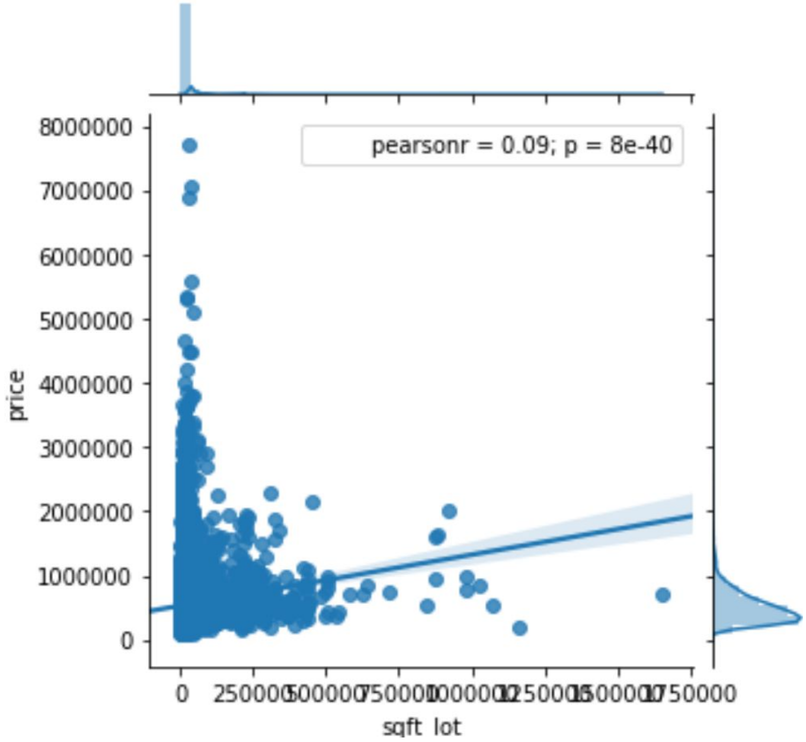
There are three features which describe the house area that we can use for daily life, they are: sqft_living, sqft_above, sqft_basement, apparently, sqft_living is the sum of sqft_above and sqft_basement, which shows that sqft_living has strong correlation with sqft_above and sqft_basement, this kind of strong correlation between independent variables should be avoid , So, we will drop sqft_above and sqft_basement and use feature sqft_living only.

We plot a scatter diagram as below to show the relationship between sqft_living and price , it’s easy to see that the price goes up while the house area goes up.



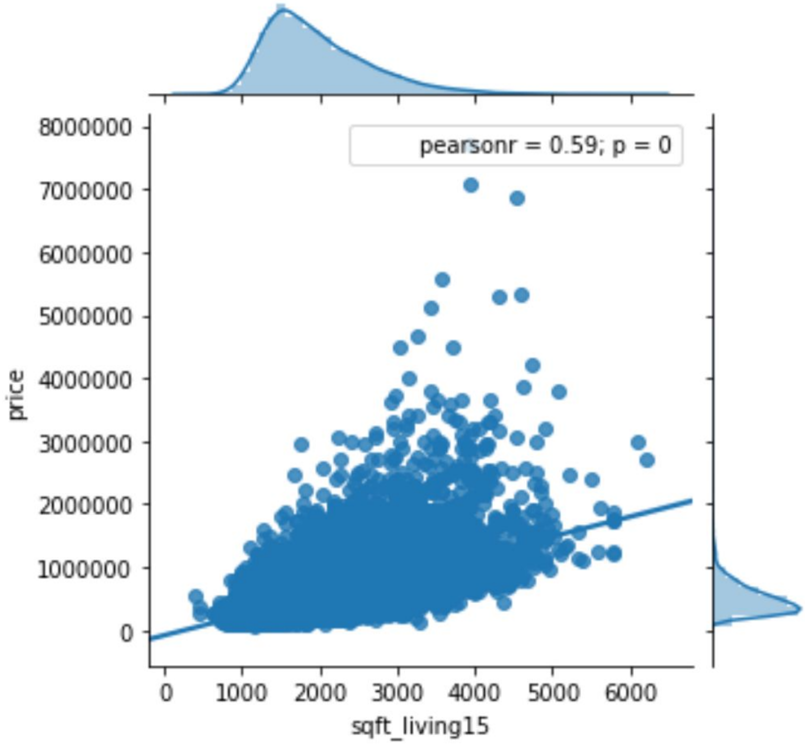
4.2.2. Relationship between sqft_lot and price

We plot a scatter diagram to show the relationship between sqft_lot and price, which tell us that most houses have low sqft_lot ,and seems sqft_lot doesn’t have strong correlation with price.



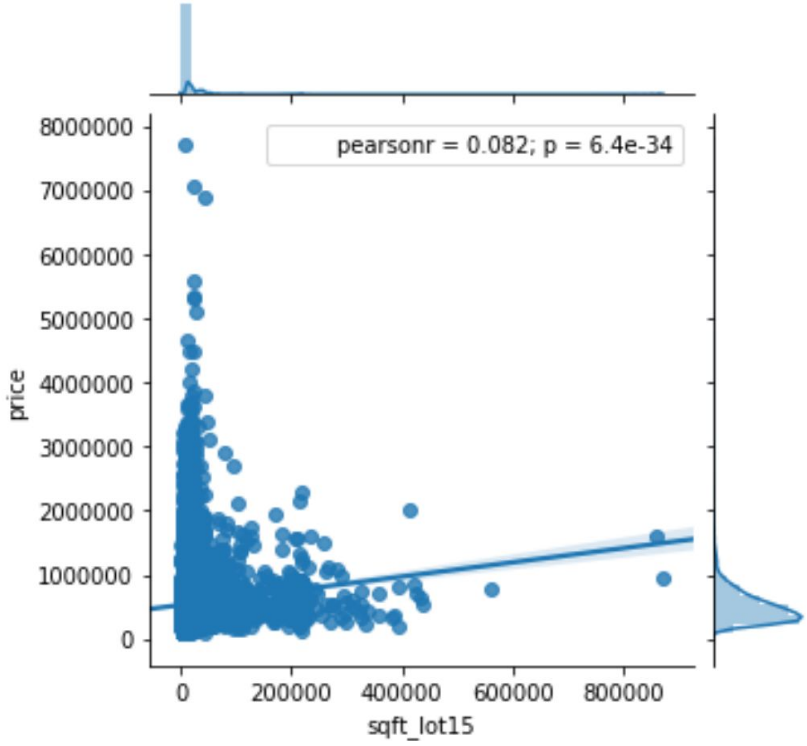
4.2.3. Relationship between sqft_living15 and price

We plot a scatter diagram to show the relationship between sqft_living15 and price, it shows sqft_living15 have strong correlation with price.



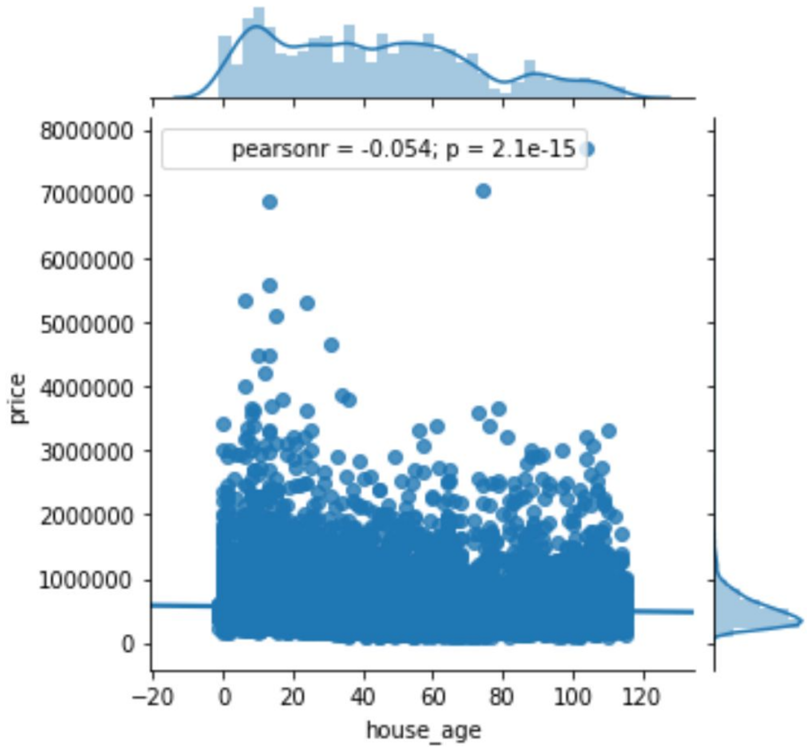
4.2.4. Relationship between sqft_lot15 and price

We plot a scatter diagram to show the relationship between sqft_lot15 and price, it shows sqft_lot15 doesn't have strong correlation with price.



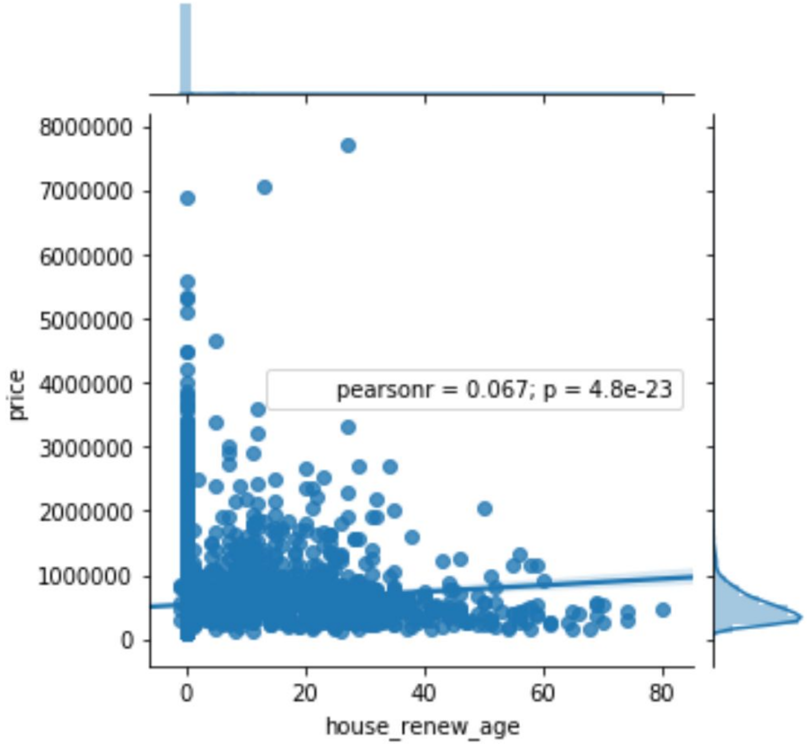
4.2.5. Relationship between house_age and price

We plot a scatter diagram to show the relationship between house_age and price, it shows house_age doesn't have strong correlation with price.



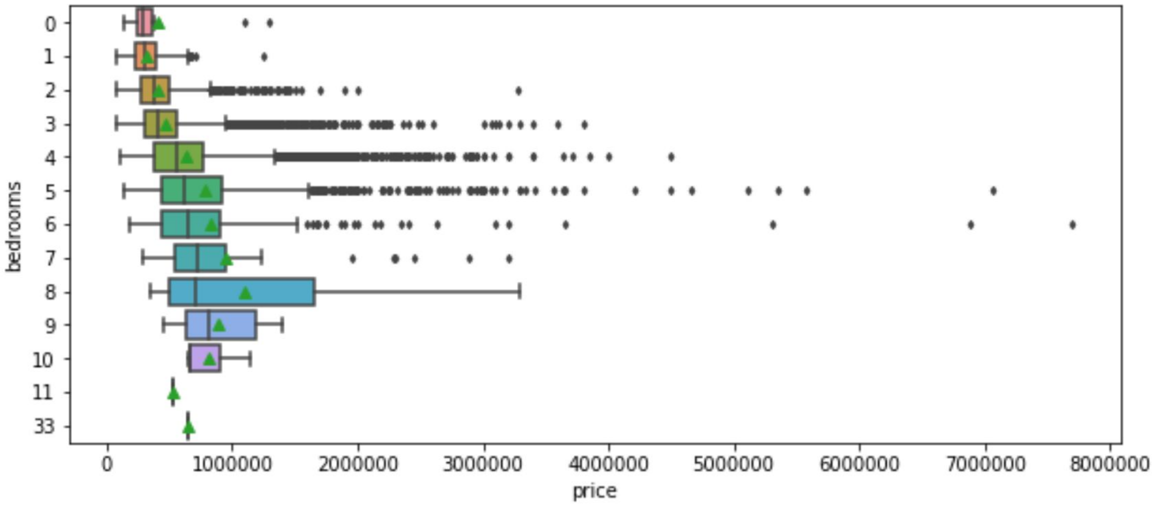
4.2.6. Relationship between house_renew_age and price

We plot a scatter diagram to show the relationship between house_renew_age and price, it shows house_renew_age doesn't have strong correlation with price.



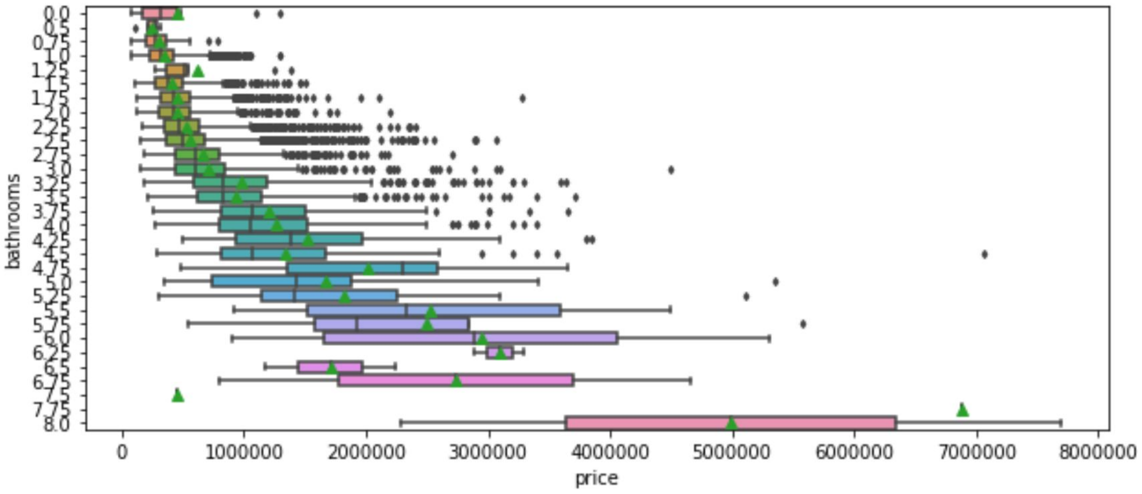
4.2.7. Relationship between bedrooms and price

We plot a boxplot diagram to show the relationship between bedrooms and price, it shows more bedrooms go with higher price.



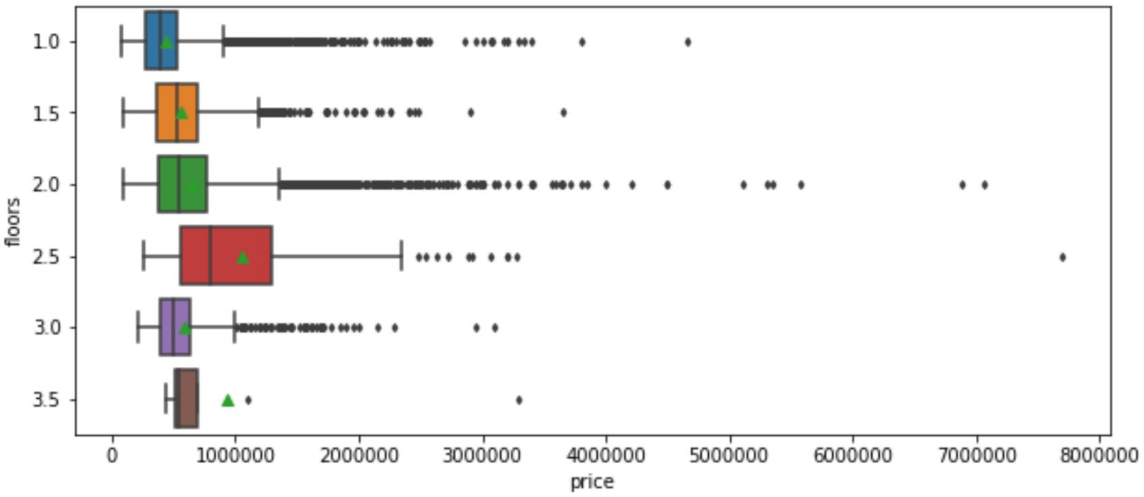
4.2.8. Relationship between bathrooms and price

We plot a boxplot diagram to show the relationship between bathrooms and price, it shows more bathrooms go with higher price.



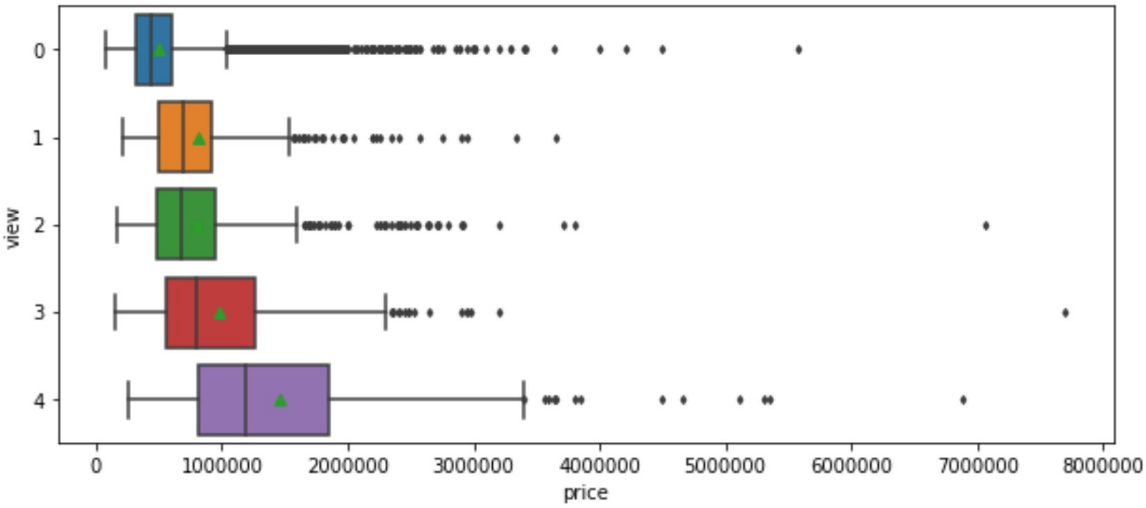
4.2.9. Relationship between floors and price

We plot a boxplot diagram to show the relationship between floors and price, it shows floors doesn’t have strong relationship with price.



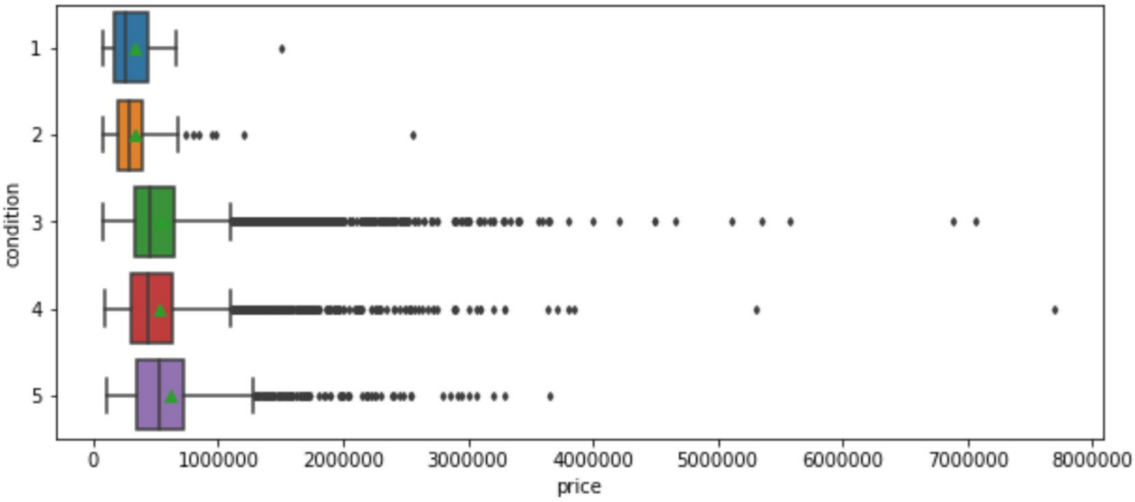
4.2.10. Relationship between view and price

We plot a boxplot diagram to show the relationship between view and price, it shows better view go with higher price.



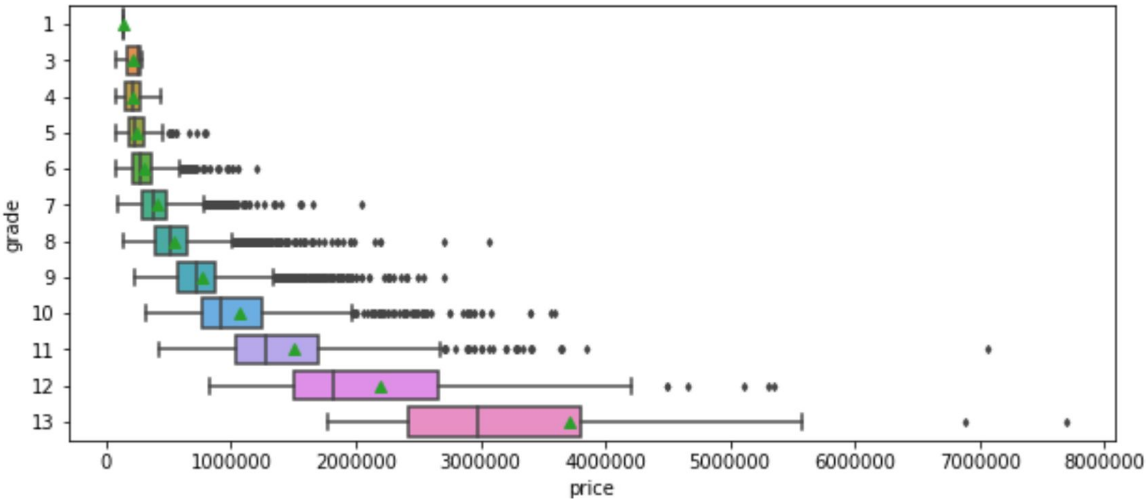
4.2.11. Relationship between condition and price

We plot a boxplot diagram to show the relationship between condition and price, it shows better condition goes with higher price.



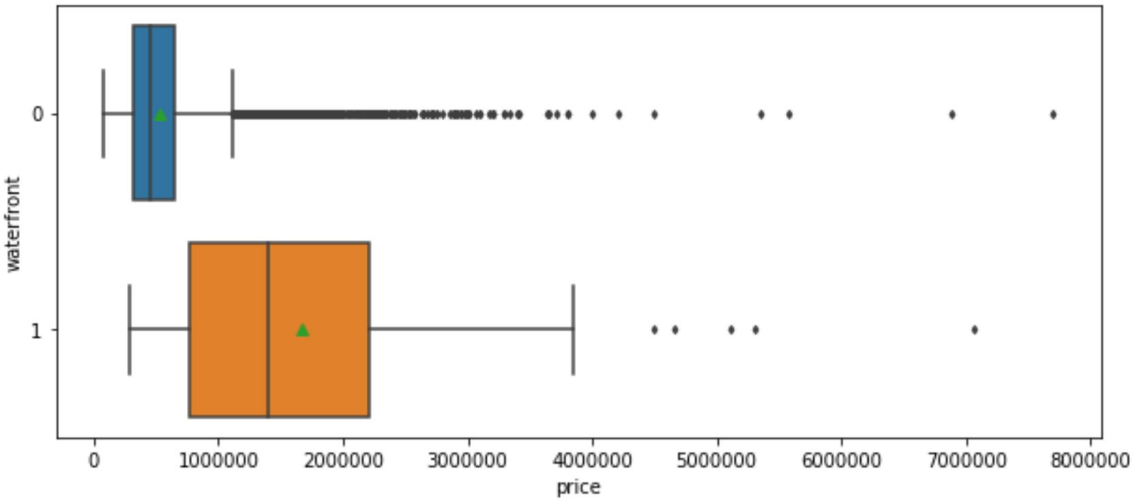
4.2.12. Relationship between grade and price

We plot a boxplot diagram to show the relationship between grade and price, it shows better grade goes with higher price.



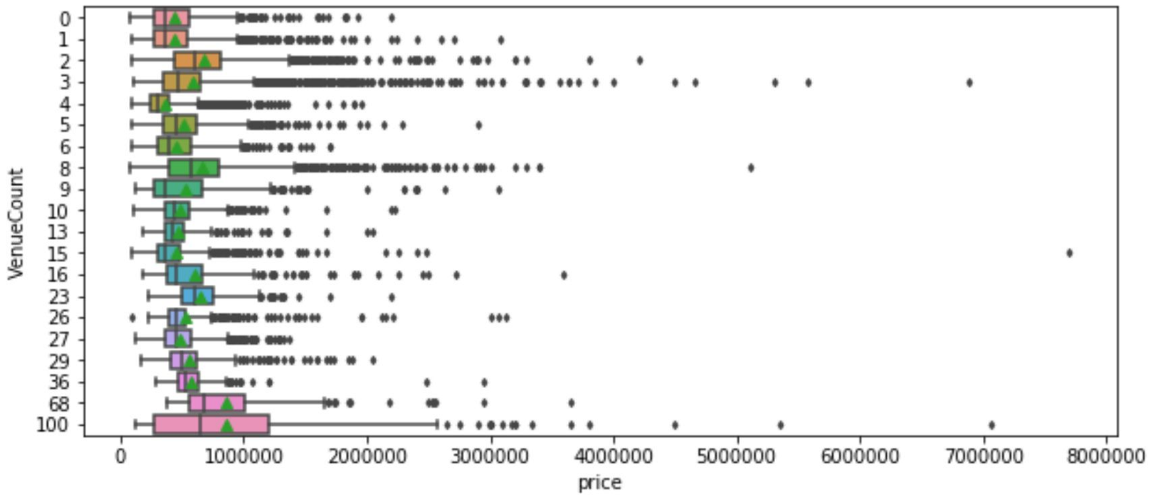
4.2.13. Relationship between waterfront and price

We plot a boxplot diagram to show the relationship between waterfront and price, it shows house with waterfront often has higher price.



4.2.14. Relationship between nearby venues count and price

We plot a boxplot diagram to show the relationship between venues count and price, to my surprise, venues count doesn't have straightforward and strong relationship with price.



4.3. Feature selection

After the exploratory data analysis, we had an deep understanding of all the features that would help predict the house price ,we discovered the data distributions, we exposed the relationships between the features and the target variable, we also did some feature transformation and derivations.

Now ,we can choose some features to build our predictive model, the following features are chose:

#	Feature	Selected(Y/N)	Reason
1	id	N	Record key, useless.
2	date	N	Useless for now.
3	price	N	Target Variable.
4	bedrooms	Y	Contribute to price.
5	bathrooms	Y	Contribute to price.
6	sqft_living	Y	Contribute to price.

7	sqft_lot	N	Similar feature selected and no strong correlation.
8	floors	N	No strong relationship.
9	waterfront	Y	Contribute to price.
10	view	Y	Contribute to price.
11	condition	Y	Contribute to price.
12	grade	Y	Contribute to price.
13	sqft_above	N	Similar feature selected
14	sqft_basement	N	Similar feature selected and no strong correlation.
15	yr_built	N	Transformed to another feature.
16	yr_renovated	N	Transformed to another feature.
17	zipcode	Y	Contribute to price.
18	lat	N	Similar feature selected
19	long	N	Similar feature selected
20	sqft_living15	N	Similar feature selected
21	sqft_lot15	N	Similar feature selected and no strong correlation.
22	house_age	Y	Contribute to price.
23	house_renew_age	Y	Contribute to price.
24	VenueCount	Y	Contribute to price.

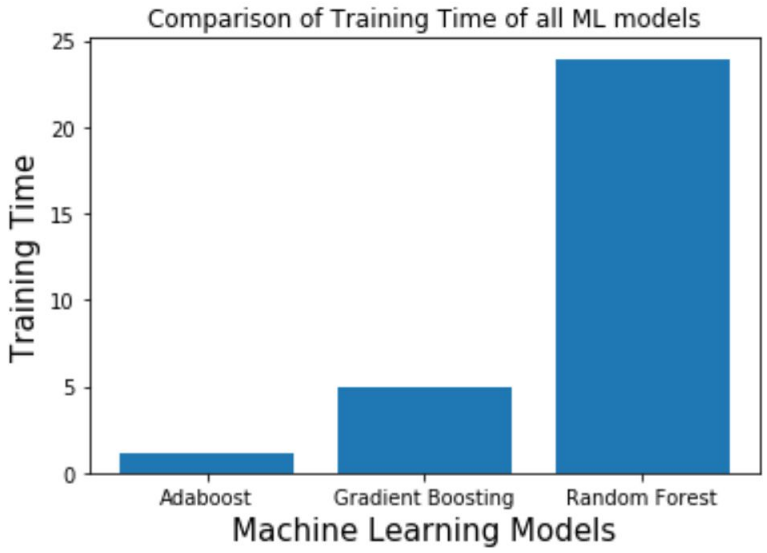
5. Predictive Modeling

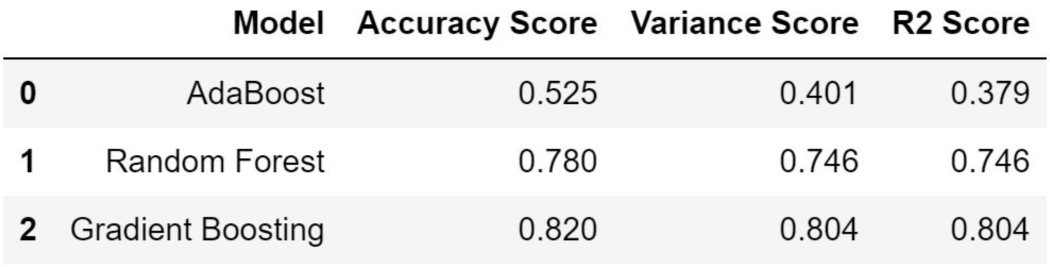
As described in the analytic approach part, we are building a predictive model to predict house price, which is a continuous value ,so I used some regression algorithm to build the model.

AdaBoostRegressor, RandomForestRegressor and GradientBoostingRegressor were used to fit the model. As usual, I split the dataset into 2 parts, the training dataset and test dataset , these 2 datasets are similar, I fit the model on the training dataset, and test the model with the test dataset. As the fitted model doesn't see the test dataset before, if it provides a good prediction on the test dataset,we have more confidence it will work on the new data to predict. That's the way we are trying to avoid overfitting.

In this case, I am splitting dataset into 20% of test data and remaining 80% will used for training the model.

I used R2-score , Accuracy Score and Explained Variance Score to evaluate model performance, the training time and performance score are listed as below, Gradient Boosting Regressor got the best performance.





	Model	Accuracy Score	Variance Score	R2 Score
0	AdaBoost	0.525	0.401	0.379
1	Random Forest	0.780	0.746	0.746
2	Gradient Boosting	0.820	0.804	0.804

6. Discussion

Because of the limitation of sandbox account of Foursquare API , I can not get more detailed information of the nearby venues of a house , so this kind of neighborhood information contribute less to the model for now. But intuitively, mature neighborhood with consummate supportive commercial and residential facilities will help a house get higher price, later, if I have the chance to get more data ,this could be a good direction to study further.

7. Conclusions

In this report, I described and explained how we can use data science to help predict a the sale price of a house. I went through the process of data science methodology, during the lab , I analyzed the relationship between the house price and the independent variables in our dataset.I identified sqft_living, grade, view, bathrooms , bedrooms among the most important features that affect a house's sale price.

I built three regression models to predict the house price. A house buyer can use this model to select house with competitive price , a house seller or agent can use this mode to evaluate house price, which can be very useful in helping house sales or real estate market.

8. References

[1] House Sales in King County, USA

[2] Foursquare API Documents : https://developer.foursquare.com/docs

[3] Pandas API Reference : http://pandas.pydata.org/pandas-docs/stable/reference/index.html