House Price Prediction

THE CAPSTONE PROJECT

Final Business Report

Submitted Anoop Raut

ABSTRACT

The Capstone Project "House Price Prediction"

In this capstone project, I have selected House Price Prediction as the problem statement as I had interest in housing market, geo-spatial data analysis and regression-based data science problem. In this business report, we are presenting the idea and understanding of the proposed business problem and methodology to solve the problem, initial insights from exploratory data analysis, using machine learning to create a predictive model that can calculate the house price based on the multiple input variables, and finally the insights and business recommendation. Let's discuss the overview in a brief passage.

The global real estate and house pricing index has skyrocketed over past two decades. The prices in US market have also gone up way higher over the years. Our problem statement revolves around the western USA focussed on Washington state. The dataset we have has 23 columns including the target variable "price" and 21613 entries. We spent a good amount of time to process the data and make it ready for modelling. Before data preparation we also performed EDA or exploratory data analysis. We checked how different variable have relation with price. How they have relation between each other. Which of them seems to be an important factor? Which categories of these variables/features are more popular or impactful? Then we moved to data prep part when we cleaned the data, change the variable type to their suitable data type, managed the missing values, treated the outliers, managed the multi-collinearity then split the data into train and test set along with the balancing of data before proceeding with the final modelling. For the modelling part we first started with basic regression modelling using stats-model. We generated a basic regression linear equation including all the variables. We then moved on to advanced modelling using machine learning models like linear regression, lasso regression, ridge regression, polynomial regression, elastic net regressor, support vector regressor, random forest regressor and finally xg-boost. To verify the performance of these models we used RMSE, RMSE cross-validation and R-square to measure the model performance. We got the best performance from XGB regressor, rest all performed lesser than XGB. R-square for XGB was the best among all which means the highest accuracy among all. Hence, we proceeded with that model and predicted the most importance features that are influencing the price of the house. We ended with the insights gained from EDA and the business recommendations that can be applied for a better pricing of the house so that all the stakeholders involved in the process may achieve the win-win situation.

TABLE OF CONTENT

CHAPTER 1

1) INTRODUCTION OF THE BUSINESS PROBLEM

a) Defining problem statement

A house value is simply more than location and square footage. Like the features that make up a person, an educated party would want to know all aspects that give a house its value. For example, you want to sell a house and you don't know the price which you may expect– it can't be too low or too high. To find house price you usually try to find similar properties in your neighbourhood and based on gathered data you will try to assess your house price.

While looking for a new house we rely on many factors that decide our dream house how it would be and where it would be. As a buyer our concern is the price, quality, location, reliability, space, sufficient room for all family members, and many other such factors. When a real estate business sells or buys a house, they also face this kind of issue as they are not sure just by looking at the property that how much price they should offer. Offering too low or high for the property price might create a financial blunder for both the parties. Hence, if we can anyhow create an analysis system using available data of the properties in the area, can help us predict a reasonable price, which can help both the parties to make a right deal. A prediction system is need to be developed that can evaluate that how the existing variables influence the house prices.

Here, we are provided with a dataset that has a record of house sold in a specific period of time. It covers different variables that affects the value of a property. Our task is to do an exploratory data analysis to analyse the data and come up with meaningful insights that tells the behaviour of sales pattern.

b) Need of the study/project

The project is a tour to give a great glimpse of how an industry-based data science project carries out. It has many learning lessons that covers up the most of the import skills needed for becoming a good data-analyst and data-scientist. It teaches us how to get empowered with power of data, how to turn numbers into meaningful insights that create guide to make right business decisions. This will be a great hands-on experience of working on a full-fledged data science project.

c) Understanding business/social opportunity

This can create an opportunity to create a win-win situation between the three points of triangle which holds the house buyer at one point who wish to buy their dream house at a suitable price without losing their precious money, the property dealer/owner would be at another point who wants to offer the best price without making any loss and third is us who wish to give them both the best service using the power of data analytics at a small cost of our effort to evaluate the right price. This can create an opportunity to create a 3rd party application/service for both the parties.

CHAPTER 2

2) DATA REPORT

a) Understanding how data was collected in terms of time, frequency and methodology

- The provided dataset is named as "innercity.xlsx".
- This is one year data from May/2014 to May/2015.
- The data is having frequency of day basis.
- The data is collected on date basis when any house is sold.
- Data is collected with 22 features along with date.

b) Visual inspection of data (rows, columns, descriptive details)

Brief glimpse of dataset:

	cid	dayhours	price	room_bed	room_bath	living_measure	lot_measure	ceil	coast	sight	condition	quality	ceil_measure	basement
0	3876100940	20150427T000000	600000	4.0	1.75	3050.0	9440.0	1	0	0.0	3	8.0	1800.0	1250.0
1	3145600250	20150317T000000	190000	2.0	1.00	670.0	3101.0	1	0	0.0	4	6.0	670.0	0.0
2	7129303070	20140820T000000	735000	4.0	2.75	3040.0	2415.0	2	1	4.0	3	8.0	3040.0	0.0
3	7338220280	20141010T000000	257000	3.0	2.50	1740.0	3721.0	2	0	0.0	3	8.0	1740.0	0.0
4	7950300670	20150218T000000	450000	2.0	1.00	1120.0	4590.0	1	0	0.0	3	7.0	1120.0	0.0

yr_built	yr_renovated	zipcode	lat	long	living_measure15	lot_measure15	furnished	total_area
1966	0	98034	47.7228	-122.183	2020.0	8660.0	0.0	12490
1948	0	98118	47.5546	-122.274	1660.0	4100.0	0.0	3771
1966	0	98118	47.5188	-122.256	2620.0	2433.0	0.0	5455
2009	0	98002	47.3363	-122.213	2030.0	3794.0	0.0	5461
1924	0	98118	47.5663	-122.285	1120.0	5100.0	0.0	5710

- The provided dataset has 22 features as given: *'cid', 'dayhours', 'price', 'room_bed', 'room_bath', 'living_measure', 'lot_measure', 'ceil', 'coast', 'sight', 'condition', 'quality', 'ceil_measure', 'basement', 'yr_built', 'yr_renovated', 'zipcode', 'lat', 'long', 'living_measure15', 'lot_measure15', 'furnished', 'total_area'.*
- Price is our target variable which we need to predict using modelling.
- Features like *'room_bed', 'room_bath', 'living_measure', 'lot_measure', 'ceil', 'condition', 'quality', 'ceil_measure', 'basement', 'yr_built', 'yr_renovated', 'furnished'* are property's internal features.
- Features like *'coast', 'sight', 'condition', 'quality', 'zipcode',* are external features.
- To analyse the zipcode data we have used the dataset of US zipcodes which we appended into the main dataset using groupby function in knime software.

Data Dictionary

1. cid: a notation or code for a house
2. dayhours: date and time house were sold (time was not recorded here)
3. price: Price of the house and this is the prediction target
4. room_bed: Number of bedrooms present in house
5. room_bath: Number of bathrooms present per bedrooms. (bathroom/bedroom)
6. living_measure: square footage of the home
7. lot_measure: square footage of the lot

8. ceil: Total floors (levels) in house
9. coast: House which has a view to a waterfront/seaface
10. sight: Sights that has been viewed
11. condition: How good the condition of house is (Overall condition of house)
12. quality: grade given to the housing unit, based on grading system
13. ceil_measure: square footage of house apart from basement
14. basement_measure: square footage of the basement
15. yr_built: Built Year of house when it was made
16. yr_renovated: Year when house was renovated
17. zipcode: zip code of the area
18. lat: Latitude coordinate
19. long: Longitude coordinate
20. living_measure15: Living room area in 2015 (implies-- some renovations) This might or might not have affected the lotsize area
21. lot_measure15: lotSize area in 2015 (implies-- some renovations)
22. furnished: Based on the quality of room
23. total_area: Measure of both living and lot

These are the data summary:

- The number of rows (observations) are 21613.
- The number of columns (variables) are 23.
- There are 23 variables. There are 12 float datatypes, 4 integer datatypes and 7 object datatypes.
- There are few time data columns like dayhours (day at which house was sold), which we have to change to time format.
- There are few datatypes that are object datatypes even if they have numeric values, this can be changed to float for seeing their correlation with other variables.

Statistical Summary of data:

	count	mean	std	min	25%	50%	75%	max
price	21613.0	540182.16	367362.23	75000.00	321950.00	450000.00	645000.00	7700000.00
bedroom	21505.0	3.37	0.93	0.00	3.00	3.00	4.00	33.00
bathroom	21505.0	2.12	0.77	0.00	1.75	2.25	2.50	8.00
living_area	21596.0	2079.86	918.50	290.00	1429.25	1910.00	2550.00	13540.00
lot_area	21571.0	15104.58	41423.62	520.00	5040.00	7618.00	10684.50	1651359.00
total_floors	21541.0	1.49	0.54	1.00	1.00	1.50	2.00	3.50
seaface	21582.0	0.01	0.09	0.00	0.00	0.00	0.00	1.00
sight_viewed	21556.0	0.23	0.77	0.00	0.00	0.00	0.00	4.00
condition	21528.0	3.41	0.65	1.00	3.00	3.00	4.00	5.00
quality_grade	21612.0	7.66	1.18	1.00	7.00	7.00	8.00	13.00
floor_area	21612.0	1788.37	828.10	290.00	1190.00	1560.00	2210.00	9410.00
basement_area	21612.0	291.52	442.58	0.00	0.00	0.00	560.00	4820.00
living_area_2015	21447.0	1987.07	685.52	399.00	1490.00	1840.00	2360.00	6210.00
lot_area_2015	21584.0	12766.54	27286.99	651.00	5100.00	7620.00	10087.00	871200.00
furnished	21584.0	0.20	0.40	0.00	0.00	0.00	0.00	1.00
total_area	21545.0	17192.04	41628.69	1423.00	7032.00	9575.00	13000.00	1652659.00
latitude	21613.0	47.56	0.14	47.15	47.45	47.57	47.68	47.76
longitude	21613.0	-122.20	0.16	-122.47	-122.32	-122.23	-122.12	-121.63
population	21613.0	34706.26	12704.49	3267.00	25474.00	35041.00	43471.00	70245.00
density	21613.0	1909.63	1400.18	17.60	892.50	1791.00	2649.40	7776.10
house_sold	21613.0	1.01	0.09	1.00	1.00	1.00	1.00	3.00
house_age	21598.0	44.99	29.37	1.00	19.00	41.00	65.00	116.00
renovation_yrs	21613.0	0.85	5.16	0.00	0.00	0.00	0.00	82.00
renovated	21613.0	0.04	0.20	0.00	0.00	0.00	0.00	1.00

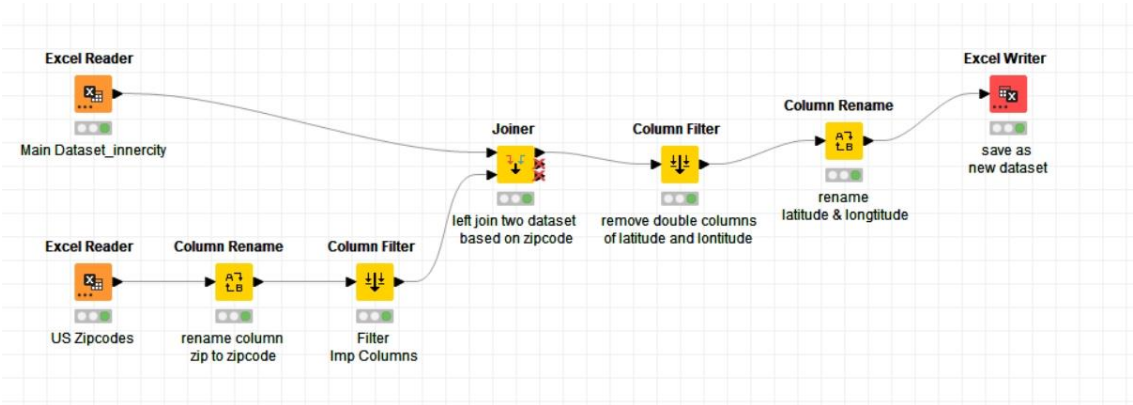
Key points from the statistical summary:

- CID: This is House id which we can use it to find out the property resold or not?
- price: this is our target column value. It falls in range of 75k - 7700k. Here, mean > median, hence right-skewed.
- room_bed: number of bedrooms range from 0 - 33. here, mean (slightly) > median hence slightly right-skewed.
- room_bath: number of bathrooms range from 0 - 8. mean (slightly) < median, hence, left-skewed.
- living_measure: square footage of house range from 290 - 13,540. as mean > median, it's right-skewed.
- lot_measure: square footage of lot range from 520 - 16,51,359. as mean almost double of median, it's hightly right-skewed.
- ceil: number of floors range from 1 - 3.5 as mean ~ median, it's almost normal distributed.
- coast: as this value represent whether house has waterfront view or not. it's categorical column. from above analysis we got know, very few houses has waterfront view.
- sight: value ranges from 0 - 4. as mean > median, it's right-skewed
- condition: represents rating of house which ranges from 1 - 5. as mean > median, it's right-skewed
- quality: representing grade given to house which range from 1 - 13. as mean > median, it's right-skewed.
- ceil_measure: square footage of house apart from basement ranges in 290 - 9,410. as mean > median, it's right-skewed.

- basement: square footage house basement ranges in 0 - 4,820. as mean highlty > median, it's highly right-skewed.
- yr_built: house-built year ranges from 1900 - 2015. as mean < median, it's left-skewed.
- yr_renovated: house renovation year only 2015. so, this column can be used as categorical variable for knowing whether house is renovated or not.
- zipcode: house zipcode ranges from 98001 - 98199. as mean > median, it's right-skewed.
- lat: lattitude ranges from 47.1559 - 47.7776 as mean < median, it's left-skewed.
- long: longittude ranges from -122.5190 to -121.315 as mean > median, it's right-skewed.
- living_measure15: value ranges from 399 to 6,210. as mean > median, it's right-skewed.
- lot_measure15: value ranges from 651 to 8,71,200. as mean highly > median, it's highly right-skewed.
- furnished: representing whether house is furnished or not. it's a categorical variable
- total_area total area of house ranges from 1,423 to 16,52,659. as mean is almost double of median, it's highly right-skewed. This means there are many houses with larger areas.

c) Understanding of attributes (variable info, renaming if required)

First, we will deal with location data. We will be using zipcodes to track city, county, state and population of the area. We have used knime to join the US zipcodes with the present zipcodes in the dataset.



The new updated dataset was renamed as "innercity_mod.xlsx". We will import that dataset and use the features of it like City name, county name, population and population density, for a better perspective of house price based on location.

	cid	dayhours	price	room_bed	room_bath	living_measure	lot_measure	ceil	coast	sight	condition	quality	ceil_measure	basement
0	3876100940	20150427T000000	600000	4.0	1.75	3050.0	9440.0	1	0	0.0	3	8.0	1800.0	1250.0
1	3145600250	20150317T000000	190000	2.0	1.00	670.0	3101.0	1	0	0.0	4	6.0	670.0	0.0
2	7129303070	20140820T000000	735000	4.0	2.75	3040.0	2415.0	2	1	4.0	3	8.0	3040.0	0.0
3	7338220280	20141010T000000	257000	3.0	2.50	1740.0	3721.0	2	0	0.0	3	8.0	1740.0	0.0
4	7950300670	20150218T000000	450000	2.0	1.00	1120.0	4590.0	1	0	0.0	3	7.0	1120.0	0.0

yr_built	yr_renovated	zipcode	living_measure15	lot_measure15	furnished	total_area	latitude	longitude	city	state_name	population	density
1966	0	98034	2020.0	8660.0	0.0	12490	47.71577	-122.21580	Kirkland	Washington	43471	1853.0
1948	0	98118	1660.0	4100.0	0.0	3771	47.54245	-122.26880	Seattle	Washington	49181	3037.8
1966	0	98118	2620.0	2433.0	0.0	5455	47.54245	-122.26880	Seattle	Washington	49181	3037.8
2009	0	98002	2030.0	3794.0	0.0	5461	47.30836	-122.21638	Auburn	Washington	33468	1797.1
1924	0	98118	1120.0	5100.0	0.0	5710	47.54245	-122.26880	Seattle	Washington	49181	3037.8

This is the new modified dataset which has actual city, state name, population, population density, county name as the new variables.

Further, we cleaned the data by changing the names for our ease and by changing datatypes which were incorrect/problematic.

- The new dataset has changed the data types. The highlighted variables were changed to numerical datatype from object datatype.
- We also dealt with special character '$' which were present in the dataset at many place. Now it is changed to NaN.
- We further renamed the dataset for our ease. These renames were done in the dataset.

  'cid':'house_id', 'dayhours':'date','room_bed':'bedroom', 'room_bath':'bathroom','ceil':'total_floors', 'coast':'seaface', 'sight':'sight_viewed','quality':'quality_grade','living_measure':'living_area', 'lot_measure':'lot_area','ceil_measure':'floor_area', 'basement':'basement_area', 'lat':'latitude', 'long':'longitude', 'living_measure15':'living_area_2015', 'lot_measure15':'lot_area_2015'

The new dataset after the renaming is as shown below:

	house_id	date	price	bedroom	bathroom	living_area	lot_area	total_floors	seaface	sight_viewed	condition	quality_grade	floor_area	basement_area
0	3876100940	2015-04-27	600000	4.0	1.75	3050.0	9440.0	1.0	0.0	0.0	3.0	8.0	1800.0	1250.0
1	3145600250	2015-03-17	190000	2.0	1.00	670.0	3101.0	1.0	0.0	0.0	4.0	6.0	670.0	0.0
2	7129303070	2014-08-20	735000	4.0	2.75	3040.0	2415.0	2.0	1.0	4.0	3.0	8.0	3040.0	0.0
3	7338220280	2014-10-10	257000	3.0	2.50	1740.0	3721.0	2.0	0.0	0.0	3.0	8.0	1740.0	0.0
4	7950300670	2015-02-18	450000	2.0	1.00	1120.0	4590.0	1.0	0.0	0.0	3.0	7.0	1120.0	0.0

yr_built	yr_renovated	zipcode	living_area_2015	lot_area_2015	furnished	total_area	latitude	longitude	city	state_name	population	density
1966.0	0	98034	2020.0	8660.0	0.0	12490.0	47.71577	-122.21580	Kirkland	Washington	43471	1853.0
1948.0	0	98118	1660.0	4100.0	0.0	3771.0	47.54245	-122.26880	Seattle	Washington	49181	3037.8
1966.0	0	98118	2620.0	2433.0	0.0	5455.0	47.54245	-122.26880	Seattle	Washington	49181	3037.8
2009.0	0	98002	2030.0	3794.0	0.0	5461.0	47.30836	-122.21638	Auburn	Washington	33468	1797.1
1924.0	0	98118	1120.0	5100.0	0.0	5710.0	47.54245	-122.26880	Seattle	Washington	49181	3037.8

We also converted three columns with date in it into some meaningful data

house_sold	house_age	renovation_yrs	renovated
1	62.0	0	0
1	29.0	0	0
1	38.0	0	0
1	15.0	0	0
1	40.0	0	0

Adding a new column 'house_sold' is generated from house_id. It shows house resold no of times.

- 1 mean not resold or sold the very first time. House sold 1 time
- 2 means resold one more time after the first purchase. House sold 2 times
- 3 shows resold two times after the first purchase. House sold 3 times

The reason behind doing this is because the price might increase or decrease if a property is resold multiple times.

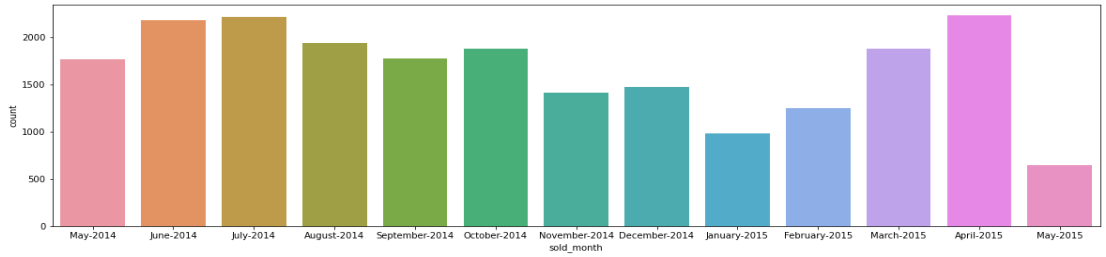
- 'house_age' represents the age of the house which is generated by substracting year with 2016 (additional year).
- 1 shows the the house is built in the present year. Further higher number shows that house is built that number of years back.
- 'renovation_yrs' shows the year before which the house was renovated.
- 'renovated' shows that whether house was renovated or not. 1 mean renovated, 0 mean not.

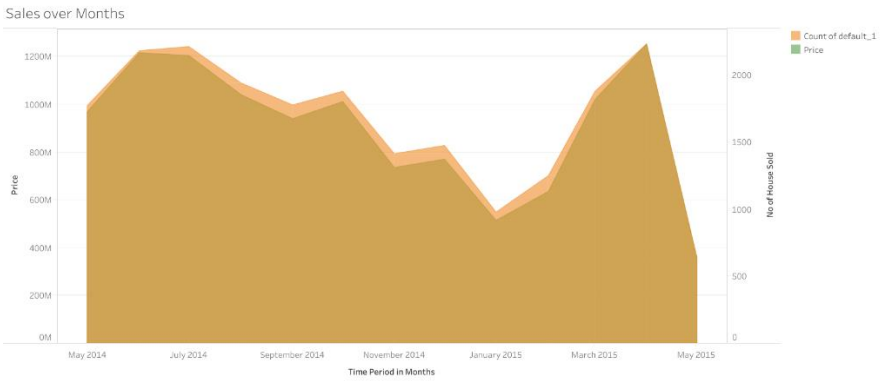
3) EXPLORATORY DATA ANALYSIS

3. a) Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)

3.a.1) Time Count plot

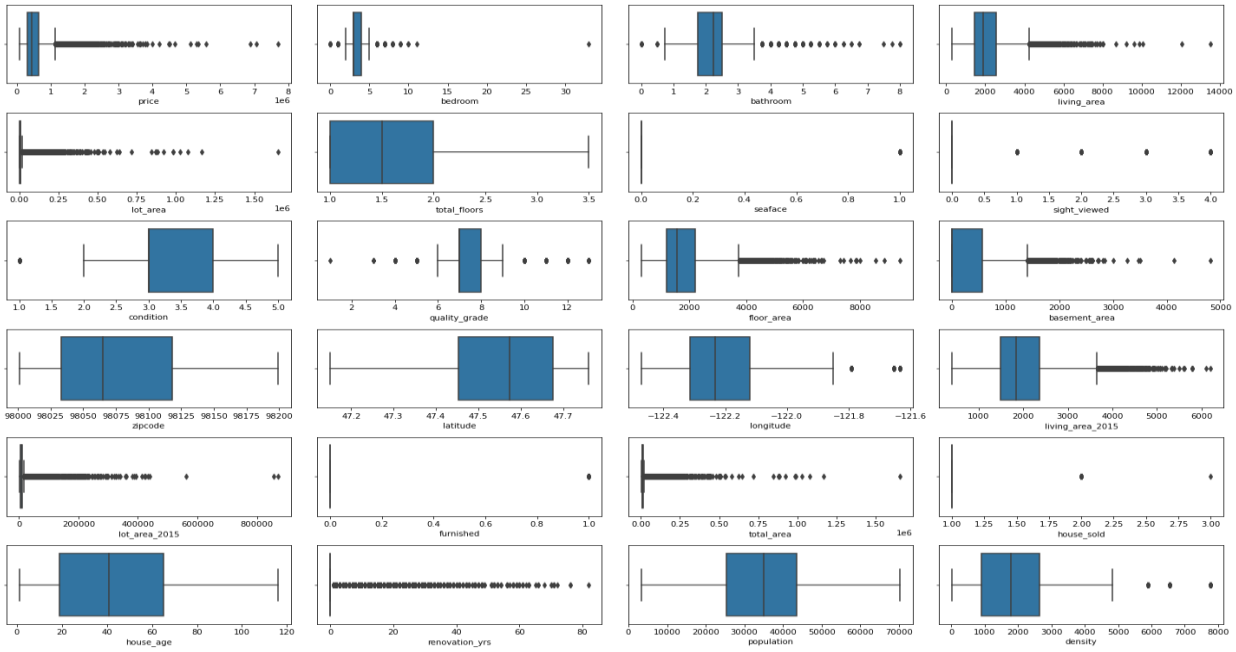
To get glimpse of the univariate analysis, first we will see the sales record of monthly sales made over a year shows that most of the house were sold in month of June, July of 2014 and in April 2015. During winters people prefer buying less may be because of discomfort to go out for sight-seeing.





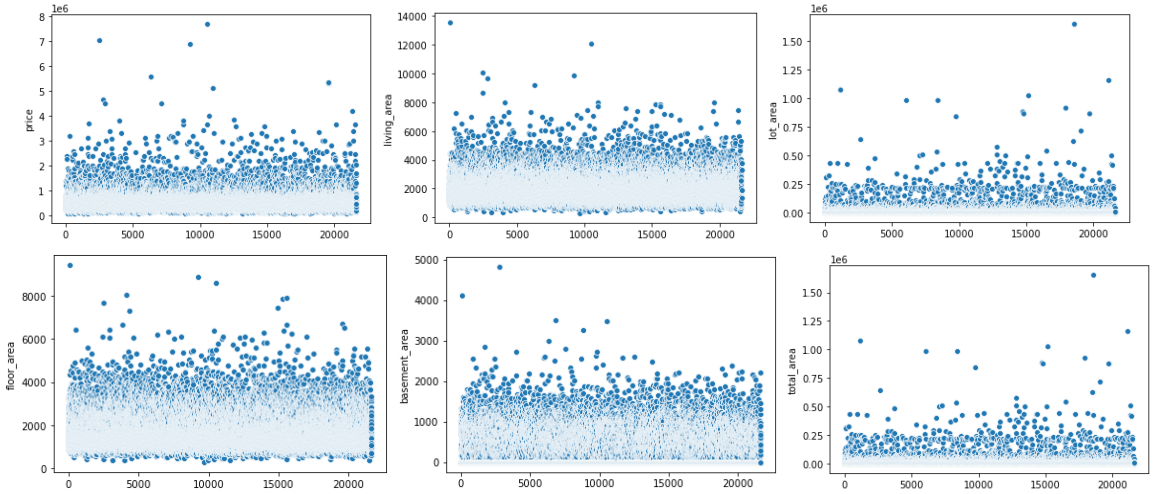
Next, we will look at the boxplots of all the variables to find the outliers in the dataset.

3.a.2) Boxplot of all the numerical variables:



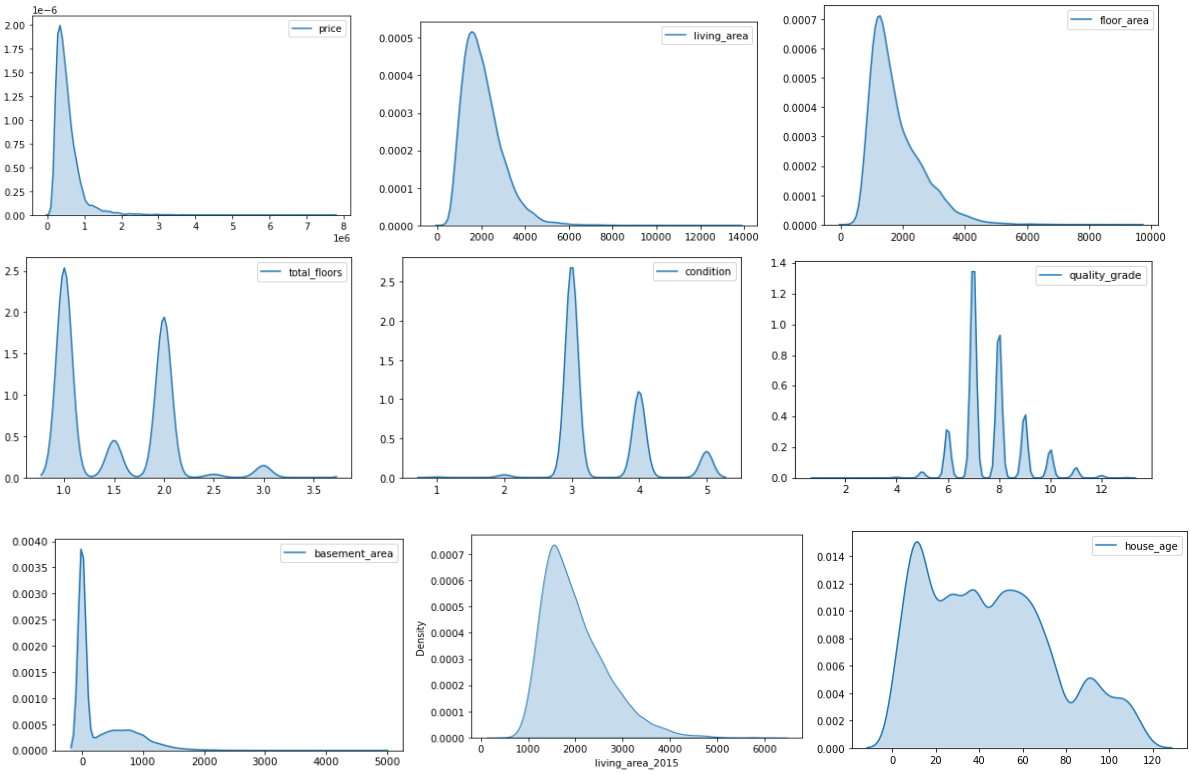
Here, we can see that there are many outliers present in many of the variables. But we will avoid removing outliers randomly because we do not want to lose data. Like the area values has many outliers. The bedroom seems to have a very high outlier of 33 which we will remove separately. Even price has outliers but we will not touch price as this is our target variable. Lot area shows many outliers because most of the house do not have lot available like house made in apartment building where there won't be any lot area available.

3.a.3) Scatter Plots of important variables:



The scatter plot for price shows that most of the house has price between 1 to 2 million. Beyond 2 million, price fluctuates and might have add values of its luxury. Same goes with living area and lot area, floor area, basement area and total area which all might be correlated to each other. Lot area seems to be correlated with total area as both behaves similarly.

3.a.4) Density Plots for variables:

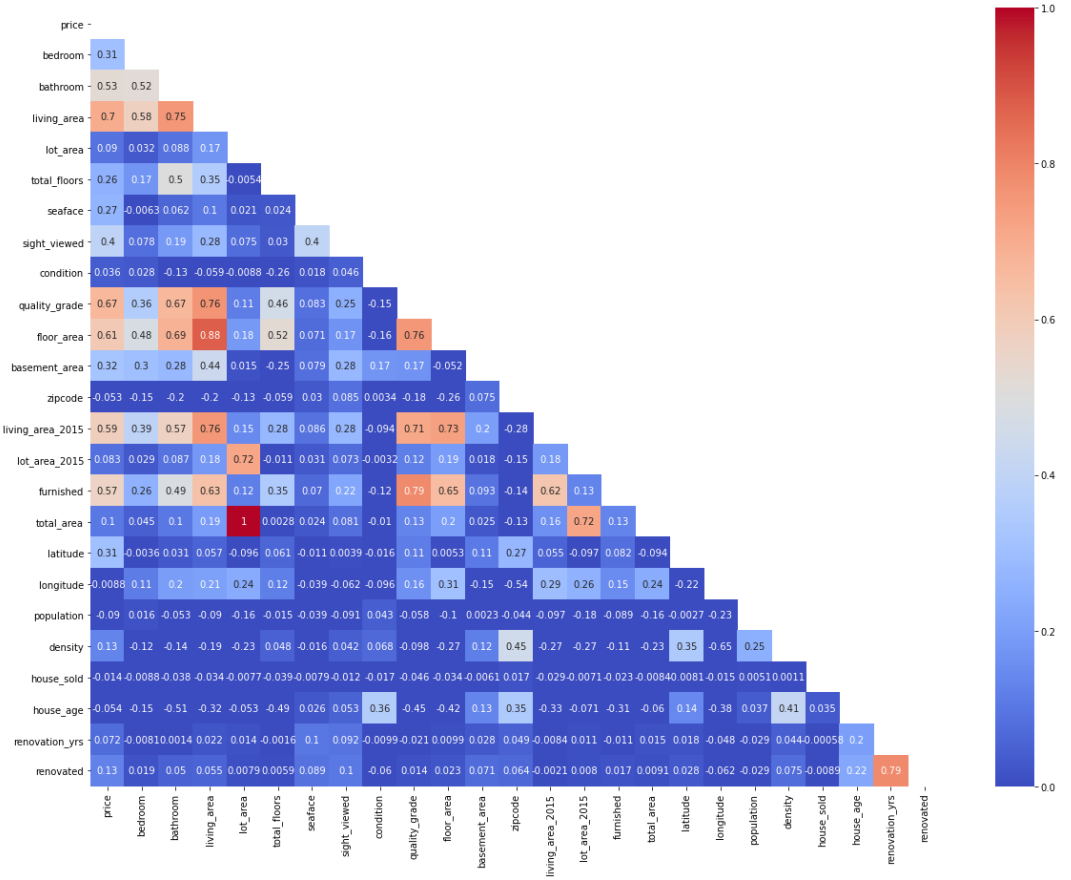


The density plot here shows all the skewness of data that we predicted earlier. Price, floor area, living area all are right-skewed. This shows that major of the property has a particular range of price. There are few luxury houses also available on the chart which are responsible for the skewness.

House age shows that most of the house are within 20 years and 20 to 60 years. There are also many houses that built in range of 80 to 120 years ago. These might be large mansions and ancestral properties. Most of the house has average condition, then better and then best. Hardly few properties are not maintained and has bad condition. The grading system also strengthen the point. Major of the house 1 floor then upto 2 floors. Very less house has beyond that number of floors. Which means we do not have any apartment house here as guessed earlier.

3.b) Bivariate analysis (relationship between different variables, correlations)

3.b.1) Correlation Plot:

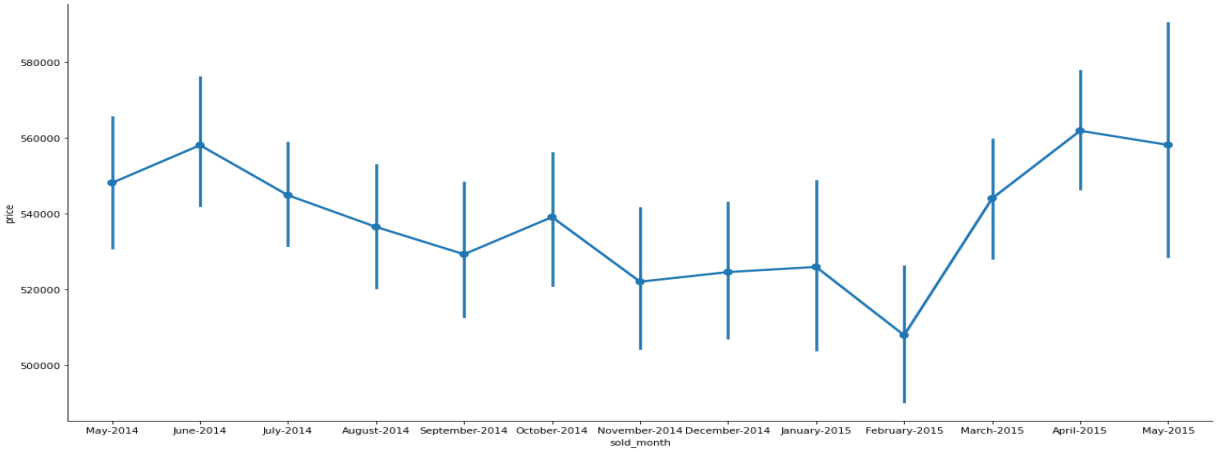


The correlation plot shows a lot about the behaviour of data and interdependency. The price is surely dependent on living area, floor area, quality grade, and furnished status. Total area is highly correlated to lot area. Living area is correlated to bathroom ratio. Quality of house depends on bathroom and living area. Floor area is highly correlated to living area. Quality grade depends on furnished status of house.

Now, let's see some bivariate analysis between variables and target variable 'price'.

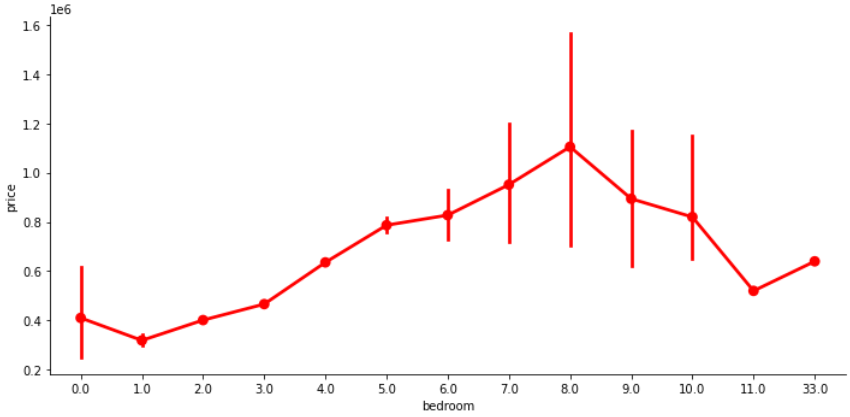
3.b.2) Some bi-variate plots

1. Plot between Time of House Sell vs Price:



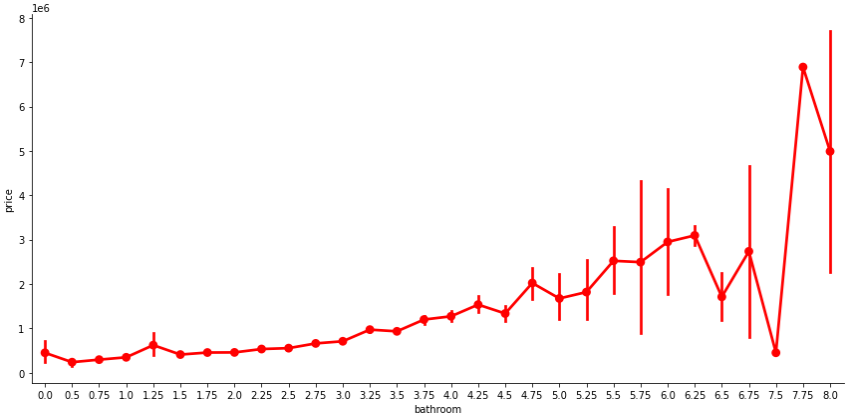
The plot between price and timestamp justifies the statement stated earlier that during winters the price of house goes down whereas during summer it rises up.

2. No of bedrooms vs Price:



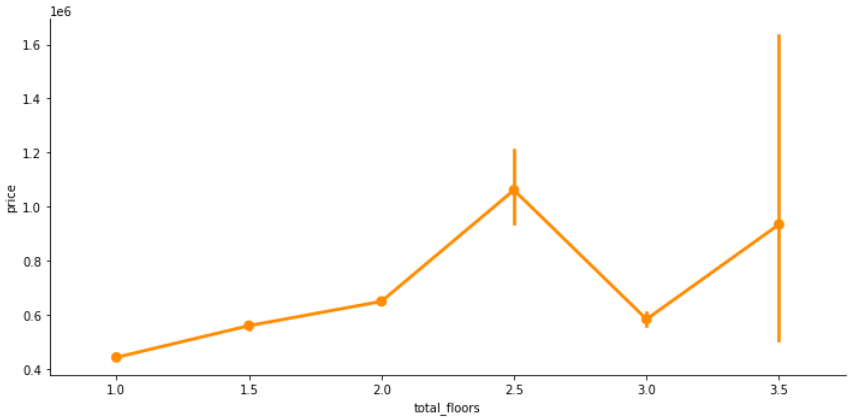
The plot shows that house with 8 bedrooms has high price, but beyond that it decreases.

3. No. of Bathrooms/Bedroom vs Price:



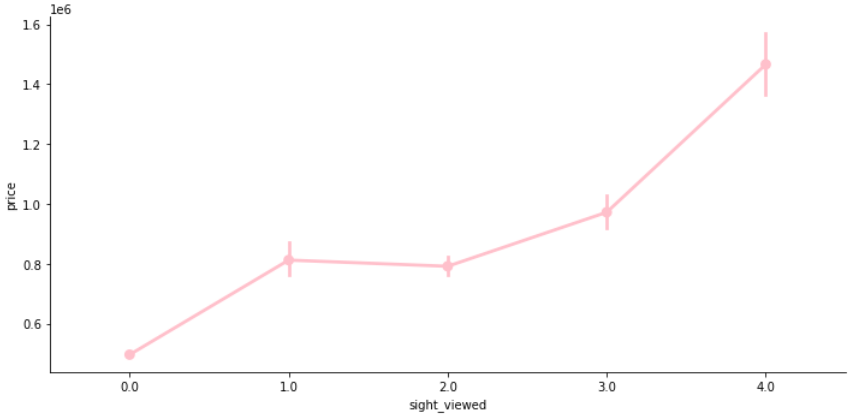
It grows steadily which means price rises up as the number of bathrooms increases in house.

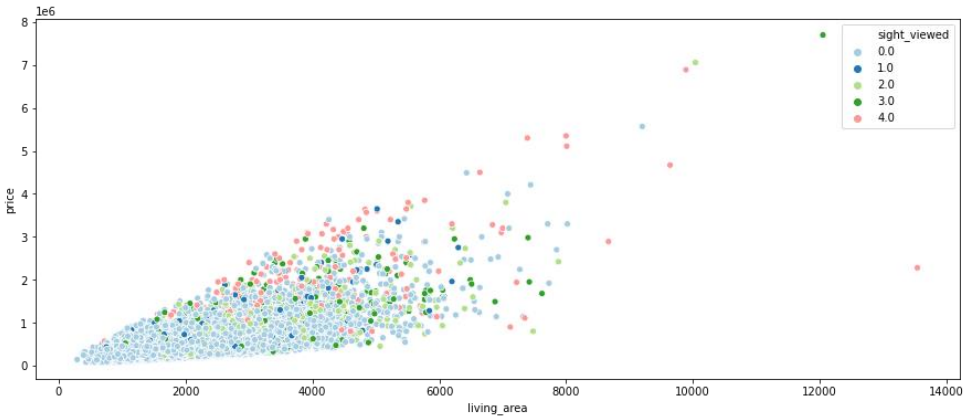
5. No. of Floors vs Price:



This shows that most of the house price increases as the number of floor increases till 2.5 floors. But beyond that at 3.5 price fluctuates a lot.

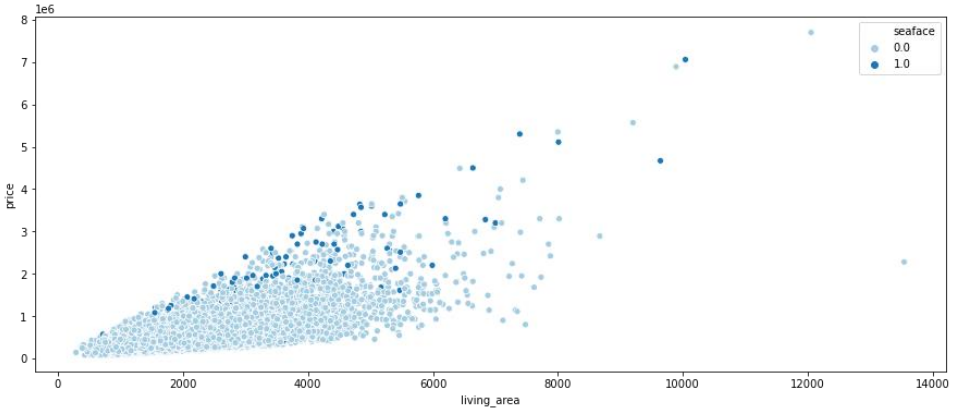
6. No. of sights viewed vs Price:





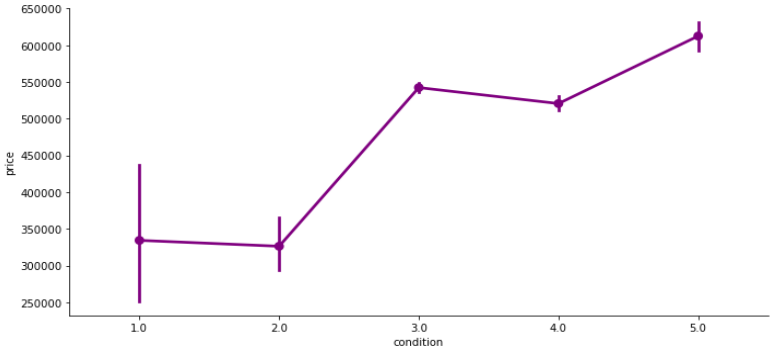
The plot shows that house which has more sights views has higher price. However, scatter plot shows that most of the house has less than 1 sight views which mean people buy at first view of the house.

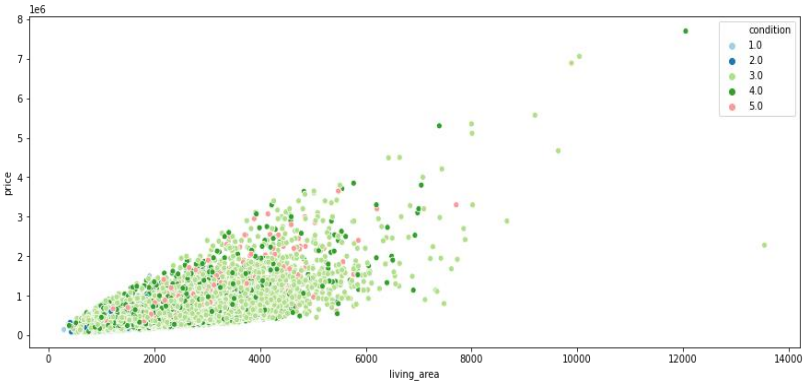
7. Effect of seaface view on Price:



The houses with seaface views are scattered over high price range as shown in the plot. Few of them also has very high living area.

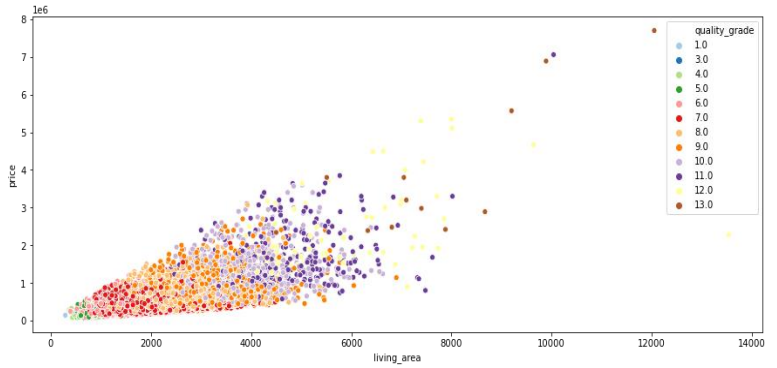
8. Effect of condition of house on price:

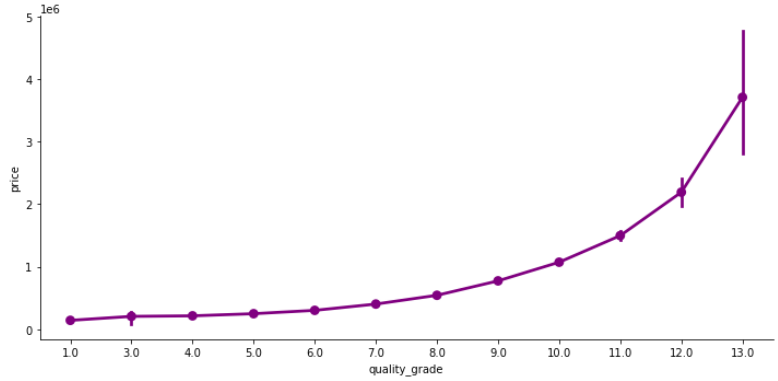




The properties that have better condition are priced more. Best condition properties are at higher price range.

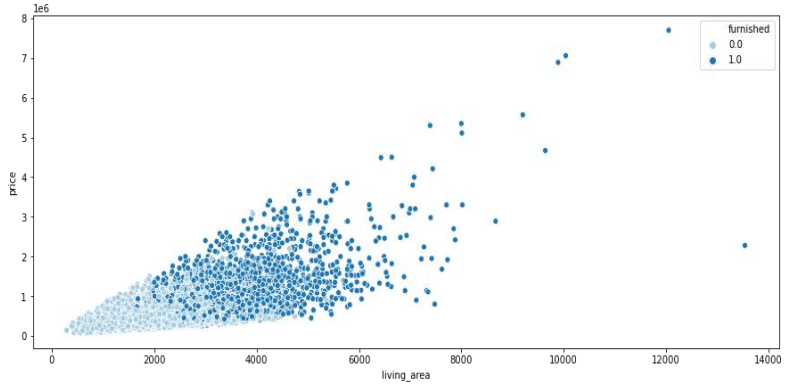
9. Effect of quality grade of house on its price:





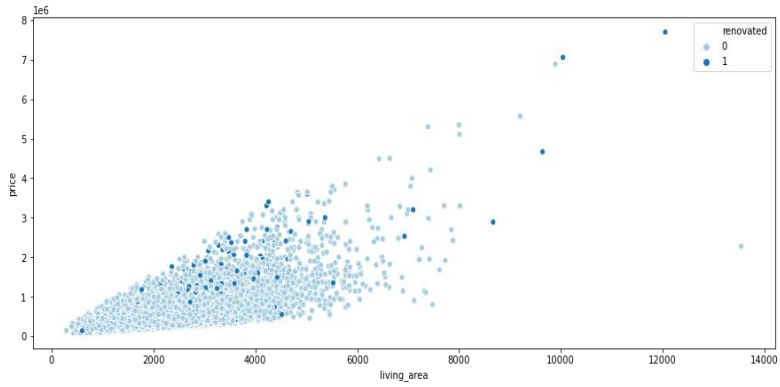
The quality grading depends a lot on size of house which can be seen through the scatter plot. House with higher grading also cost higher price.

10. Effect of house furnishing on price:



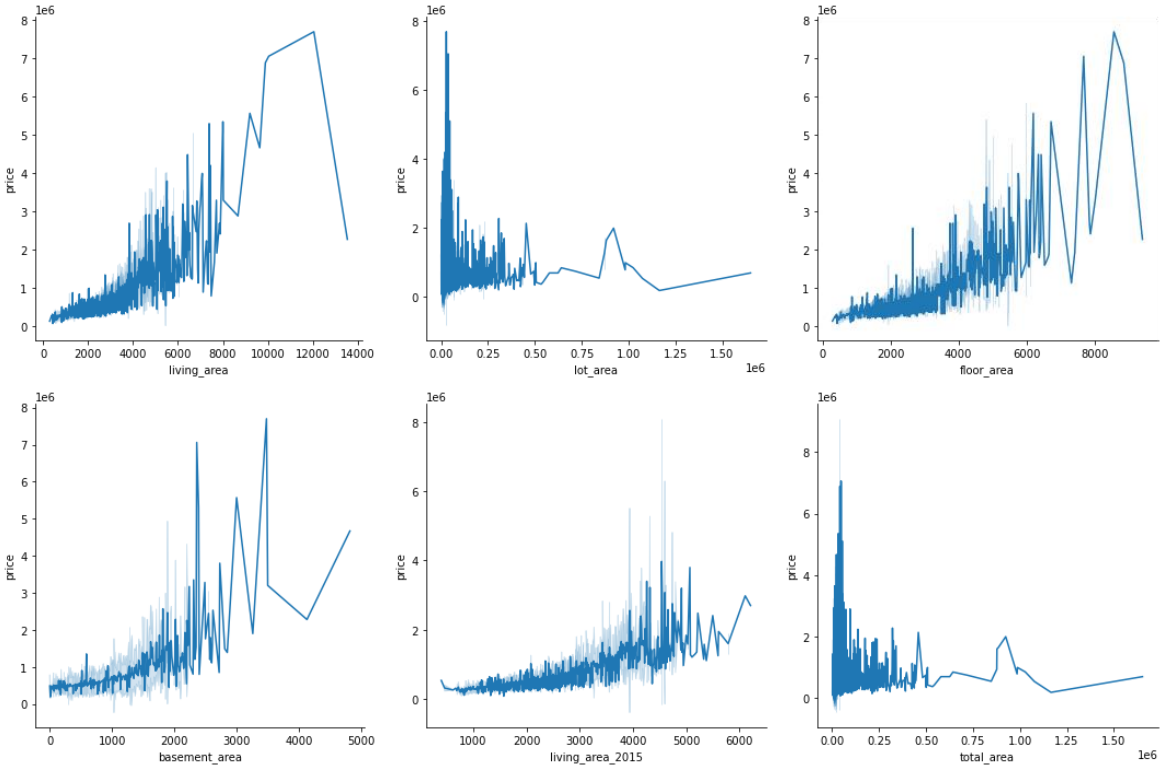
Furnished house also have larger area and also cost more.

11. Effect of renovation on price:



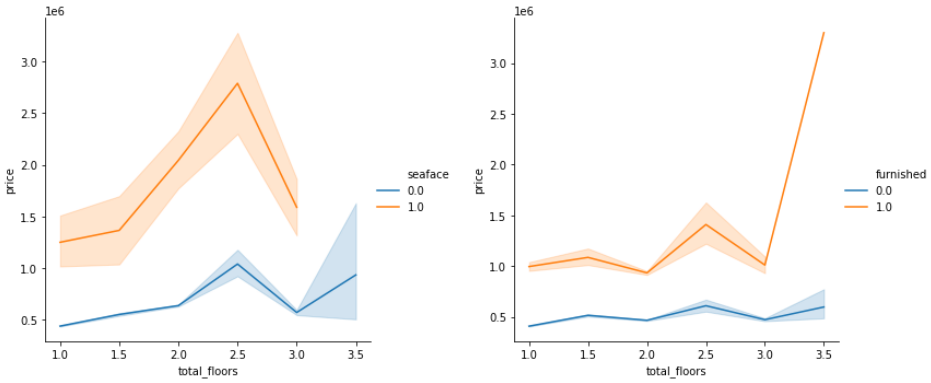
From the above plot it can be seen that house which are renovated has a higher cost.

12. Effect on different areas on price:



The above series of plots shows that with increase in area of house the price also shoots up. But as we can see that most of the house have average house areas. Only few has higher area which also cost higher price.

13. How seaface and furnished property are priced as per number floors

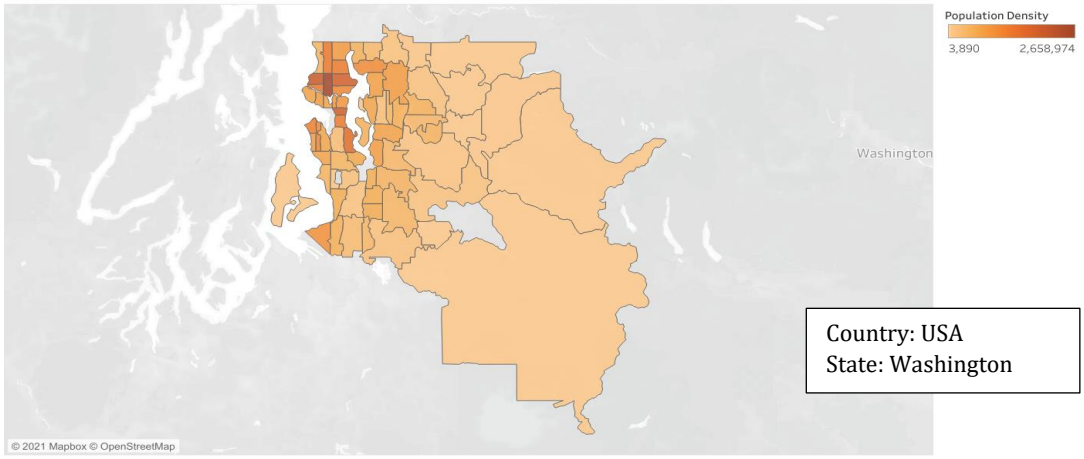


Here we can see that most of seaface house has higher price even very spike in price is seen with increase in number of floors as compared to non-seafacing houses. The furnished houses shows higher price with increase in number of floors.

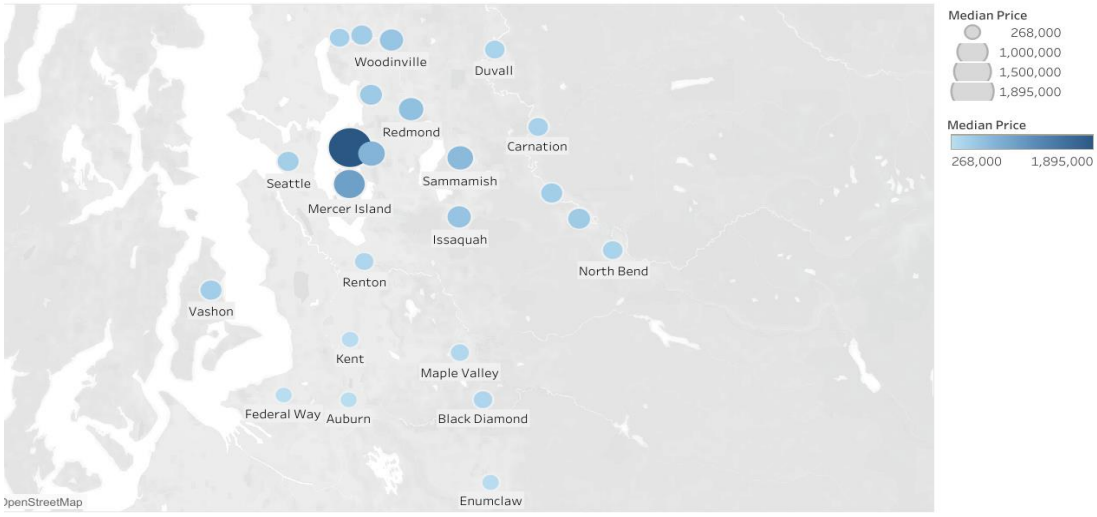
3.b.3) Location plots

Population density across zipcodes:

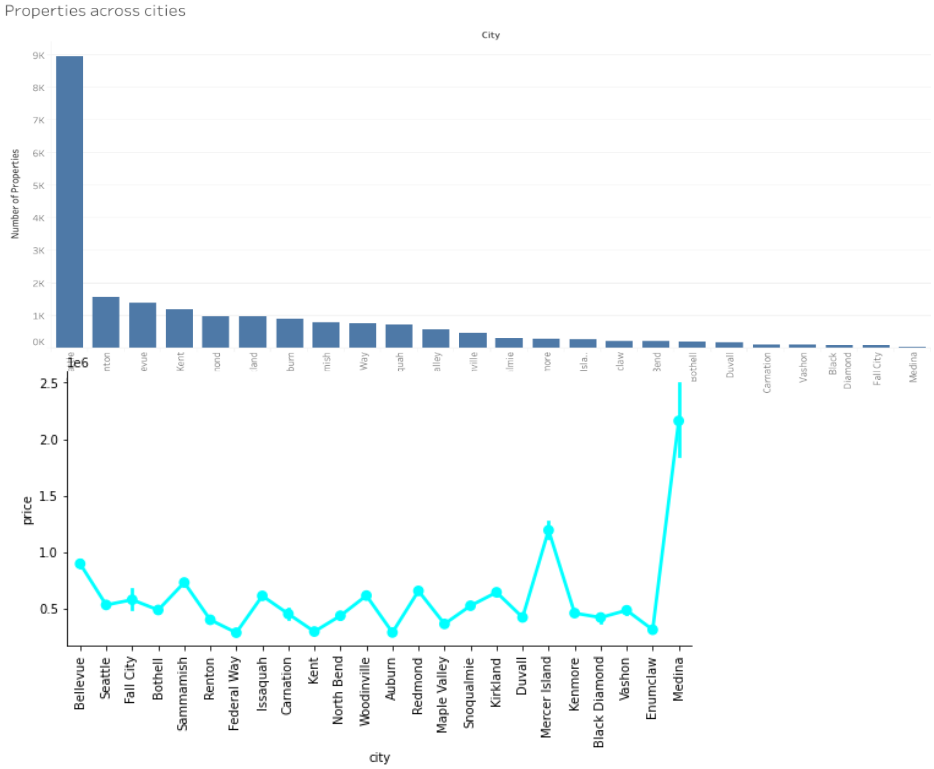
Population Density



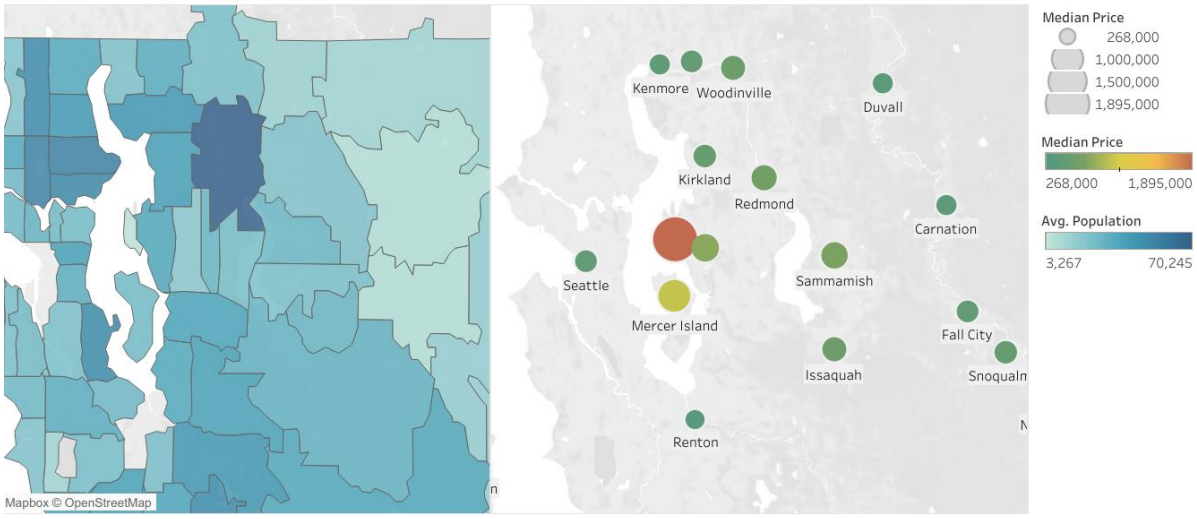
Country: USA
State: Washington



Median price variation across cities:



The zipcode map shows that how population density has spread over the map of Washington. We can see that few areas in top are overly populated and they might have a chance of high price distribution because of densely populated area. Zipcodes of Medina are the most populated area in the Washington. The same can be seen in the city map based on price variation that medina.



In the next plot, we can see that instead of higher population in Redmond, the price of house are relatively very low as compared to Medina. The reason might be medina is coastal city and must have many coast view or sea facing properties which might be costing more. It also has highly dense population. Whereas if we see the number of properties available then Seattle has the highest number of properties that were sold.

3.c) Removal of unwanted variables

we have removed the unwanted variables like house_id, year_built which is converted to house_age, year_renovation which is converted into renovated or not and renovation years ago.

3.d) Missing Value treatment

The missing values in the dataset is treated carefully and all the missing values were imputed with the median values.

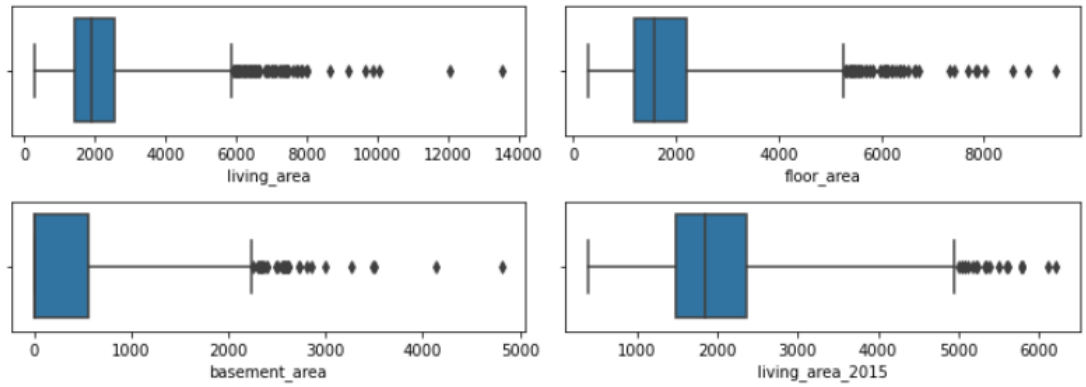
Before and after is shown below.

date	0
price	0
bedroom	108
bathroom	108
living_area	17
lot_area	42
total_floors	72
seaface	31
sight_viewed	57
condition	85
quality_grade	1
floor_area	1
basement_area	1
zipcode	0
living_area_2015	166
lot_area_2015	29
furnished	29
total_area	68
latitude	0
longitude	0
city	0
state_name	0
population	0
density	0
county_name	0
house_sold	0
house_age	15
renovation_yrs	0
renovated	0

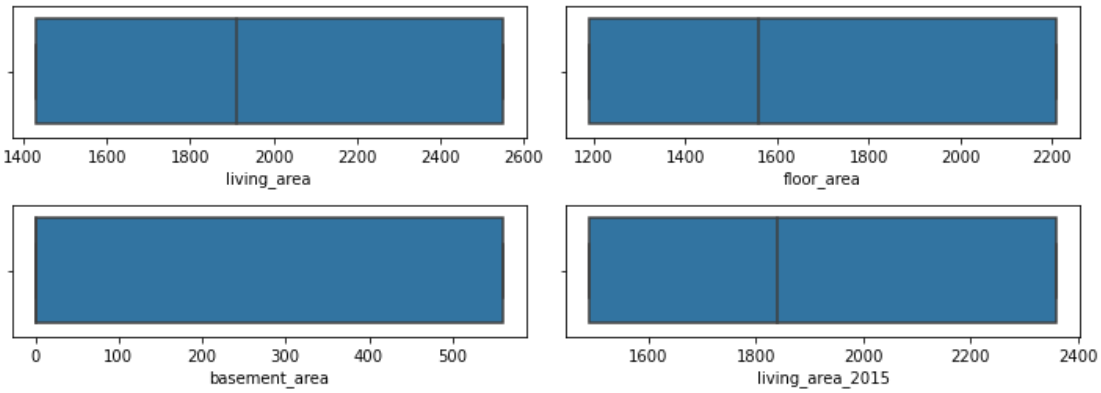
date	0
price	0
bedroom	0
bathroom	0
living_area	0
lot_area	0
total_floors	0
seaface	0
sight_viewed	0
condition	0
quality_grade	0
floor_area	0
basement_area	0
zipcode	0
living_area_2015	0
lot_area_2015	0
furnished	0
total_area	0
latitude	0
longitude	0
city	0
state_name	0
population	0
density	0
county_name	0
house_sold	0
house_age	0
renovation_yrs	0
renovated	0

3.e) Outlier treatment

Outliers were treated in the dataset. The doubtful outliers were spotted in the 5 columns. Bedroom has 33 beds as an outlier which was removed. The rest risky outliers were present in living_area, floor_area, basement_area and living_area_2015. These were treated by capping with 25 and 75 percentile. We avoided any removal of data.



After outlier treatment;



e) Variable transformation

The encoding was done. Zipcode was earlier float then was converted into object type to use for encoding. City, state and county was encoded.

zipcode	city	state_name
5	1	0
68	20	0
14	7	0
45	20	0
9	3	0

f) Addition of new variables

There were many new variables were introduced into the original dataset. There were two types of new introduction was made. First was location based new variables and the second was year based.

- Location based variables: city, state_name, county, population, density
- Year based variables: house_sold, house_age. renovation_years, renovated(yes or not)

Addition of new varibales were discussed earlier in 2.C.

1). Model building and interpretation

a. Build various models (You can choose to build models for either or all of descriptive, predictive or prescriptive purposes)

b. Test your predictive model against the test set using various appropriate performance metrics

1. The first model which we tried is the basic **Linear Regression model using stats model**.

A. Feature selection: The best features were selected using correlation of variables wrt price when the correlation exist is either greater than 10% (positive correlation) or less than -10% (negative correlation). These were the best features: 'bedroom', 'living_area', 'total_floors', 'quality_grade', 'floor_area', 'basement_area', 'latitude', 'living_area_2015', 'total_area', 'population', 'population_density', 'bathroom', 'basement_orNot' and 'zipcode'.

B. zipcodes were converted into numeric integers and not encodes as its already an encoded figure.

C. linear regression using stats model were used to produce a linear regression model.

Logistic regression results were:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.635
Model:	OLS	Adj. R-squared:	0.635
Method:	Least Squares	F-statistic:	2124.
Date:	Mon, 17 Jan 2022	Prob (F-statistic):	0.00
Time:	09:24:07	Log-Likelihood:	-2.1730e+05
No. Observations:	17109	AIC:	4.346e+05
Df Residuals:	17094	BIC:	4.348e+05
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.304e+07	1.41e+06	-9.272	0.000	-1.58e+07	-1.03e+07
bedroom	3758.5219	1960.847	1.917	0.055	-84.939	7601.983
living_area	74.8280	3.841	19.482	0.000	67.300	82.356
total_floors	-1484.7348	1945.650	-0.763	0.445	-5298.409	2328.939
quality_grade	5.129e+04	1701.727	30.138	0.000	4.8e+04	5.46e+04
floor_area	44.3043	4.039	10.968	0.000	36.387	52.222
basement_area	3.9452	9.041	0.436	0.663	-13.776	21.666
latitude	5.922e+05	7956.080	74.440	0.000	5.77e+05	6.08e+05
living_area_2015	61.1251	2.706	22.590	0.000	55.821	66.429
total_area	0.7467	0.334	2.233	0.026	0.091	1.402
population	-2.6624	0.088	-30.417	0.000	-2.834	-2.491
population_density	37.1702	1.259	29.517	0.000	34.702	39.639
bathroom	-2866.8453	575.095	-4.985	0.000	-3994.091	-1739.599
basement_orNot	1.918e+04	4136.395	4.636	0.000	1.11e+04	2.73e+04
zipcode	-156.6977	13.829	-11.331	0.000	-183.803	-129.592

Omnibus:	13.504	Durbin-Watson:	2.001
Prob(Omnibus):	0.001	Jarque-Bera (JB):	14.555
Skew:	0.035	Prob(JB):	0.000691
Kurtosis:	3.124	Cond. No.	2.42e+08

D. various changes were tried, but almost all of them were giving an adjusted R-square in range of 0.63-0.65. With this range the prediction was not good.

The linear regression equation that was developed was:

*Price = (-13035181.5) * Intercept + (3758.52) * bedroom + (74.83) * living_area + (-1484.73) * total_floors + (51286.05) * quality_grade + (44.3) * floor_area + (3.95) * basement_area + (592247.9) * latitude + (61.13) * living_*

*area_2015 + (0.75) * total_area + (-2.66) * population + (37.17) * population_density + (-2866.85) * bathroom + (19175.76) * basement_orNot + (-156.7) * zipcode*

2. Now, we moved on to advanced modelling techniques.

Advanced modelling was done in Phase 1 and Phase 2.

In phase 1. The data was not treated for outliers or multi-collinearity. This means we did not clean the anomalies present in the data to check the predictions without any tempering.

Categorical encoding was done for zipcode and city.

Data was split into 80:20, 80% for train and 20% for test.

Following model scores were used to predict the efficiency of models: MAE, MSE, RMSE, R2 Score, RMSE

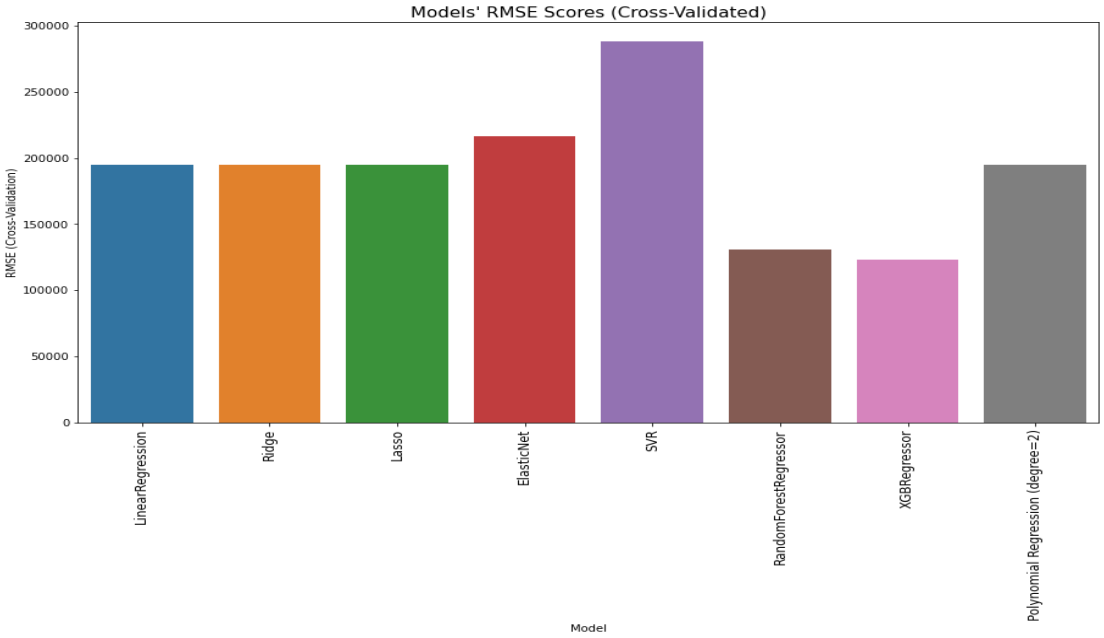
7 different models were used: Linear Regression, Ridge Regression, Lasso Regression, Elastic Regression, support Vector Machine, Random Forest Regressor, XG Boost Regressor, and Polynomial Regression of degree2.

Here is the model comparison of all the models:

	Model	MAE	MSE	RMSE	R2 Score	RMSE (Cross-Validation)
6	XGBRegressor	66778.233436	1.632371e+10	127764.282859	0.884653	123340.524755
5	RandomForestRegressor	69893.210793	1.813834e+10	134678.646427	0.871830	130744.062354
2	Lasso	120999.417531	3.851481e+10	196251.910285	0.727845	195105.910072
1	Ridge	121001.703622	3.852493e+10	196277.682617	0.727774	195129.586948
0	LinearRegression	121010.031719	3.851637e+10	196255.867875	0.727834	195130.606017
7	Polynomial Regression (degree=2)	95224.245472	2.444431e+10	156346.748958	0.827271	195130.606017
3	ElasticNet	134349.792172	4.810076e+10	219318.845637	0.660109	216292.839518
4	SVR	159320.250834	8.839376e+10	297310.881938	0.375389	288353.453162

Here, we have sorted the model based on the RMSE score for cross validation.

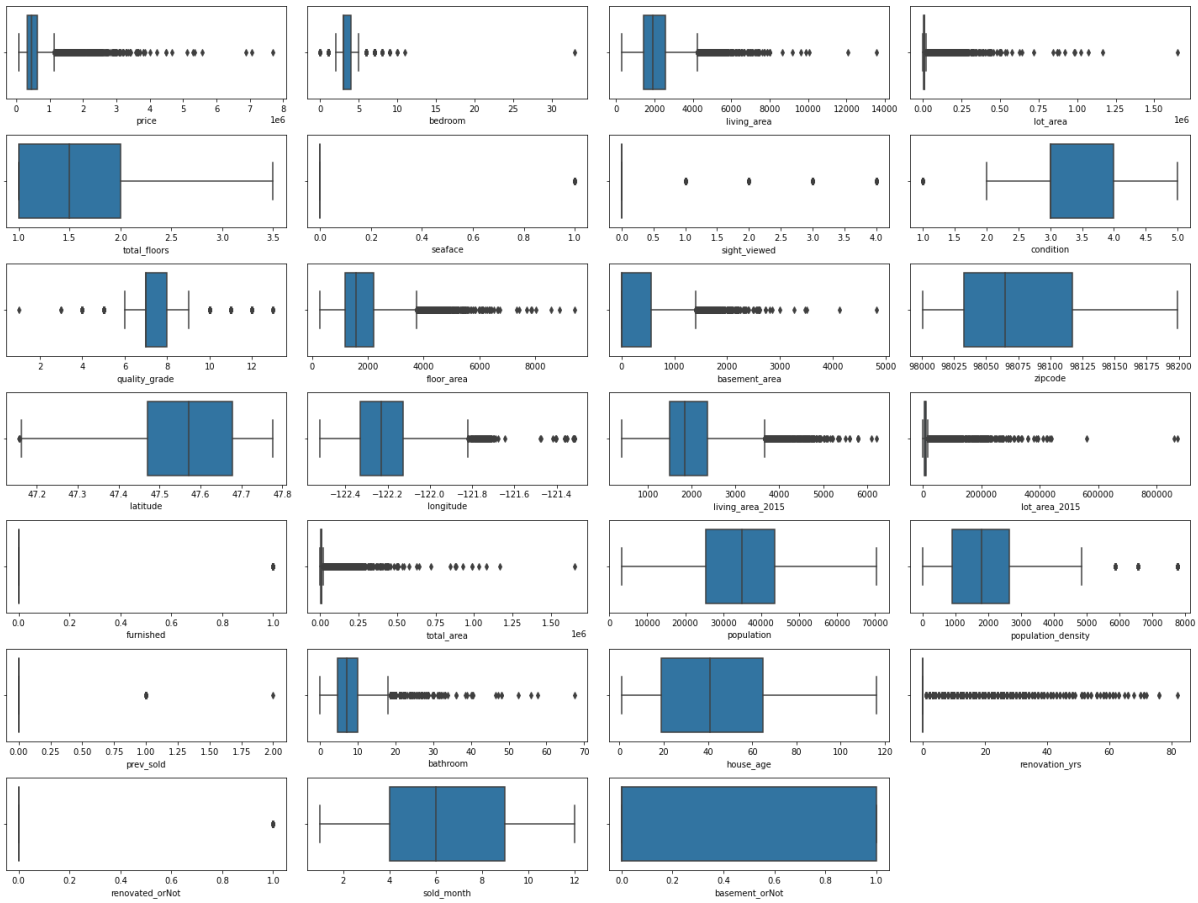
The least RMSE score is given by XGB Regressor which means is producing the best model with 88% of accuracy.



Now, we will move on to phase 2, where we will play with data to make a more affordable and efficient model.

In first part of phase.2 modelling, we treated the outliers from the important variables.

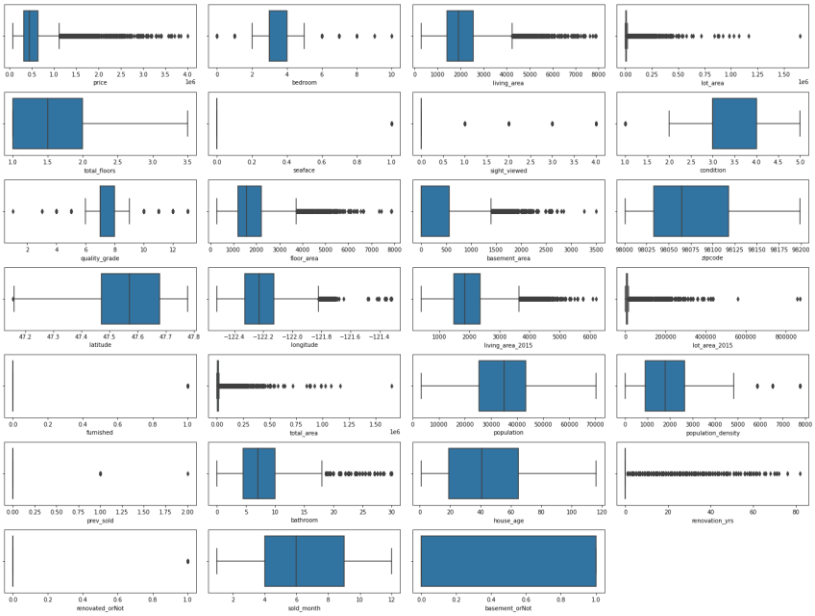
Here is the box plot before the outlier treatment



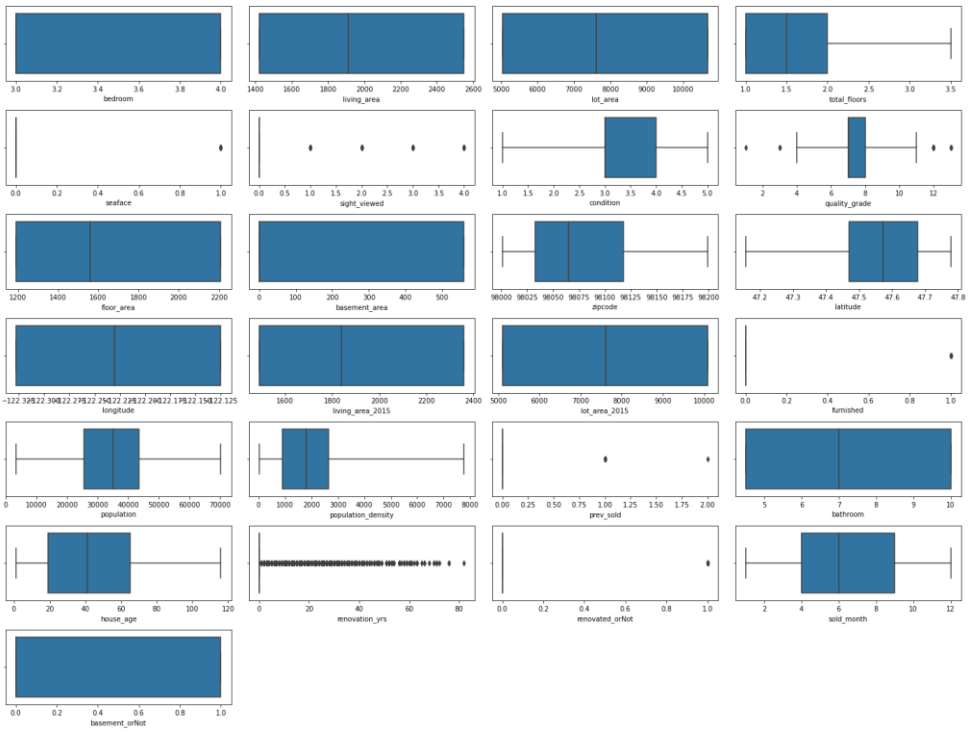
Important variables that are important and affected by severe outliers are: price, bedroom, bathroom, living area, lot area, floor area, basement_area, total area, living area 2015, lot area 2015.

Here total area and lot area are completely correlated variables.
First we tried hand with trimming the data for bedroom, living area, bathroom, price without any major loss to data. The entire loss while this trimming was only 0.17%.

We still had the outliers as seen in plot



Hence, we decided to treat the outliers to avoid a low performance model.
Here, we decided to avoid the data loss. Hence, we treated the outliers by limiting it. The loss of data was still 0.17 percent.



Encoding The categorical variables like city and zipcode were encoded ordinally

Dealing with Multi-collinearity

Best features based on correlation with price after the treatment were in the decreasing order of importance:

```
quality_grade
furnished
living_area
floor_area
living_area_2015
bathroom
sight_viewed
latitude
bedroom
total_floors
seaface
basement_area
basement_orNot
population_density
lot_area
```

Best features based on Variance Inflation Factor:

	variables	VIF
6	house_age	3.946816
4	population_density	3.319328
9	sold_month	2.967401
8	renovated_orNot	2.780598
7	renovation_yrs	2.710621
2	basement_area	1.694152
1	sight_viewed	1.431715
3	furnished	1.327381
0	seaface	1.206317
5	prev_sold	1.010255

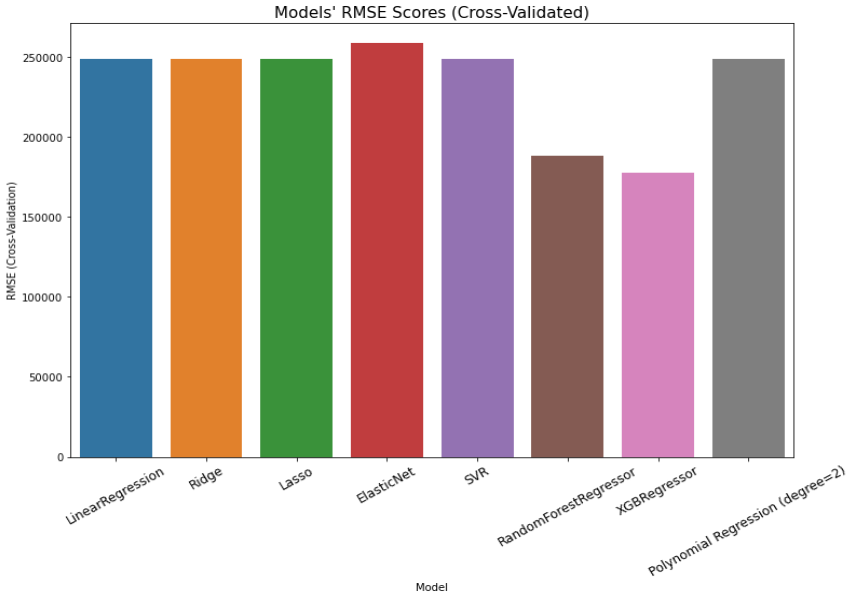
Here, we will go with the VIF features and create our model.

data was scaled after slitting into X and y

The 7 models as earlier were model on this data and we got the following table having the metric results:

	Model	MAE	MSE	RMSE	R2 Score	RMSE (Cross-Validation)
6	XGBRegressor	109588.810407	3.388545e+10	184080.015062	0.721648	177662.842134
5	RandomForestRegressor	113827.825063	3.621693e+10	190307.474485	0.702496	188572.932944
4	SVR	153611.600318	6.350519e+10	252002.363955	0.478337	248661.920421
1	Ridge	164076.783812	6.364604e+10	252281.663818	0.477180	248840.402085
2	Lasso	164077.460363	6.364602e+10	252281.634957	0.477180	248840.438676
0	LinearRegression	164077.427094	6.364604e+10	252281.672118	0.477180	248840.472700
7	Polynomial Regression (degree=2)	157616.532787	6.021923e+10	245396.076806	0.505330	248840.472700
3	ElasticNet	168166.954505	6.905008e+10	262773.822645	0.432789	258668.262934

Here also we can see that XGB Regressor has performed best and has given the better model. But we can see that there is decrease in accuracy score (72%) after treatment of data and feature selection based on VIF.



Phase3. Modelling again with one hot encoding instead of ordinal encoding

zipcode_98178	zipcode_98188	zipcode_98198	zipcode_98199	city_0	city_1	city_2	city_3	city_4	city_5	city_6	city_7
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

After encoding the shape of the dataset became (21350, 175). Now we have 175 variables. This might help or become a case of high cardinality.

We divided the data into train and validation (test) into 80:20.

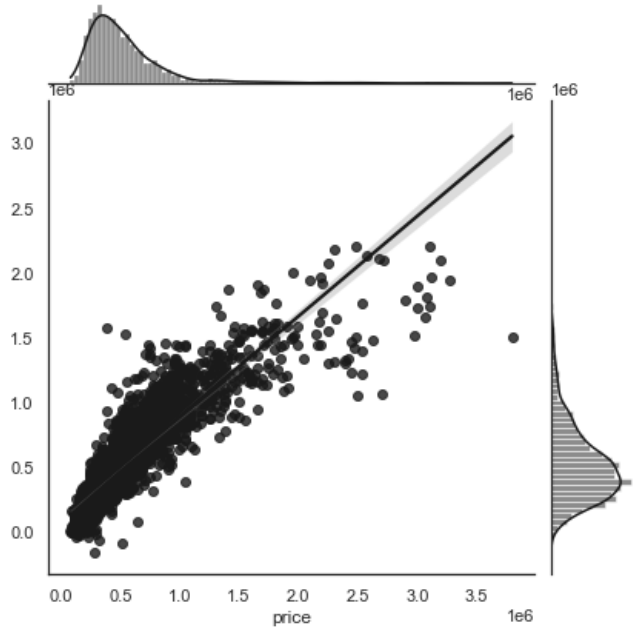
The 7 models were used here: Linear Regression (Linear-Reg-Model), Lasso Regression (Linear-Reg Lasso), Ridge Regression (Linear-Reg Ridge), K-Nearest Neighbour (KNN), Support Vector Machine (SVR), and Decision Tree (DT).

Here are the scores of the models.

	Method	Val Score	RMSE_vl	MSE_vl	MAE_vl	train Score	RMSE_tr	MSE_tr	MAE_tr
0	Linear Reg Model1	0.772993	169223.094598	2.863646e+10	101175.175211	0.781587	161062.731159	2.594120e+10	99077.558390
0	Linear-Reg Lasso1	0.772993	169223.167826	2.863648e+10	101174.520648	0.781587	161062.773865	2.594122e+10	99075.817159
0	Linear-Reg Ridge1	0.773021	169212.736153	2.863295e+10	101178.275603	0.781584	161063.691658	2.594151e+10	99077.677508
0	knn1	0.689777	197823.382224	3.913409e+10	104967.421384	0.999999	262.033321	6.866146e+04	8.910246
0	SVR1	-0.061997	366017.647093	1.339689e+11	220520.407253	-0.062606	355256.316521	1.262071e+11	218338.256863
0	DT1	0.740507	180926.847031	3.273452e+10	101866.435363	1.000000	81.545198	6.649619e+03	1.376698

Now, we look at the scores of these models, Decision Tree and KNN both are overfitting, SVR is performing poorly. Linear Regression models are performing good and the best score is Lasso Regression based on lowest RMSE in the top performing model scores.

Prediction curve based on Lasso Regression model



2). Model Tuning and Business Implication

a. Ensemble modelling (if necessary)

as we saw earlier case that XGB was performing well as compared to linear regression model, hence we will try Ensemble methods like bagging and boosting for a better model.

After bagging and boosting we can see that updated table of score:

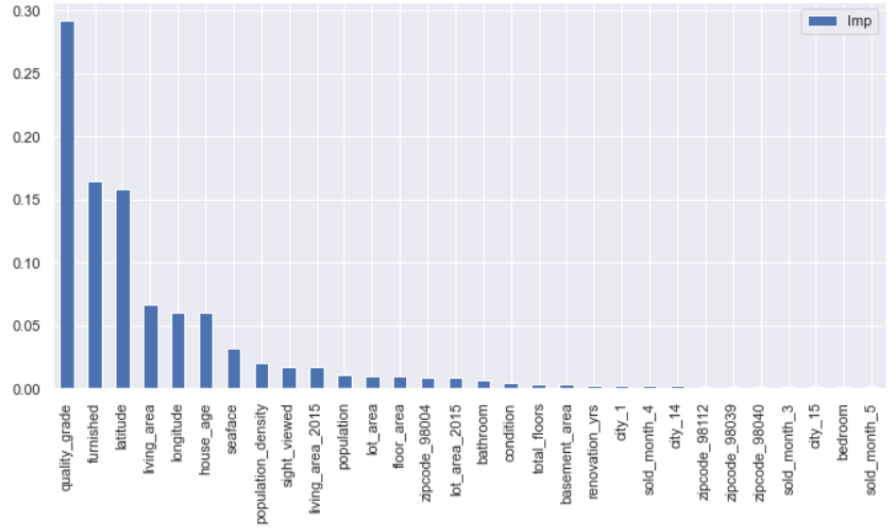
The new ensemble models were Gradient Boosting, Bagging Gradient Bossting, Random Forest.

The updated scores of the tables can be seen here:

	Method	Val Score	RMSE_vl	MSE_vl	MAE_vl	train Score	RMSE_tr	MSE_tr	MAE_tr
0	Linear Reg Model1	0.772993	169223.094598	2.863646e+10	101175.175211	0.781587	161062.731159	2.594120e+10	99077.558390
0	Linear-Reg Lasso1	0.772993	169223.167826	2.863648e+10	101174.520648	0.781587	161062.773865	2.594122e+10	99075.817159
0	Linear-Reg Ridge1	0.773021	169212.736153	2.863295e+10	101178.275603	0.781584	161063.691658	2.594151e+10	99077.677508
0	knn1	0.689777	197823.382224	3.913409e+10	104967.421384	0.999999	262.033321	6.866146e+04	8.910246
0	SVR1	-0.061997	366017.647093	1.339689e+11	220520.407253	-0.062606	355256.316521	1.262071e+11	218338.256863
0	DT1	0.740507	180926.847031	3.273452e+10	101866.435363	1.000000	81.545198	6.649619e+03	1.376698
0	GB1	0.879764	123156.738557	1.516758e+10	73105.326965	0.904692	106394.784858	1.131985e+10	67573.380525
0	BGG1	0.861565	132148.859318	1.746332e+10	73443.426515	0.980068	48655.579184	2.367365e+09	27075.424393
0	RF1	0.868449	128821.089452	1.659487e+10	72342.369321	0.980914	47612.257771	2.266927e+09	26521.537028

Here, we can see that in gradient boosting model has performing well has a good score among all. Even better than regression model score. The fitting of data between train and test (val) is also good as compared to others. Even the RMSE score is also lower than previous models. This shows that model is better than previous models.

The first 30 important features which add significant value (97%) to model are:



4) BUSINESS INSIGHTS FROM EDA

a) Is the data unbalanced? If so, what can be done? Please explain in the context of the business

Data imbalance can attack the prediction model by weaking it, hence it should be addressed before modeling. Here, our data is unbalanced for few of the variables. Like, seaface and house renovated are imbalanced variable.

These variables are not our target variables; hence we can ignore this imbalance. If we wish to consider these variables and wish to balance the data then we can use Random under/over-sampling with imblearn. RandomUnder/OverSampler is a fast and easy way to balance the data by randomly selecting a subset of data for the targeted classes. Under/Over-sample the majority/minority class(es) by randomly picking samples with or without replacement.

b) Any business insights using clustering (if applicable)

There is bit multi-collinearity between few variables. Total area depends on lot area. Hence, we can eliminate one of it before modeling. We can use VIF factor to see the multi-collinearity or to choose the best variables that contributes towards the output.

c) Any other business insights

Through EDA, we can observe that house price depends on variables. Price of premium houses with better conditions and better-quality grading are costlier. The house at premium locations and seafront locating houses are costlier. Cities like Seattle are over populated because of high density of population and hence price and also high compared to other cities. The average price is high at medina but median price are high at Seattle. The costlier cities have lower floor area as compared to cheaper cities where big house cost the less for the same floor area in over populated cities. The house those are furnished cost more. House with a greater number of floors cost more. The lot area has not much impact on price yet the large the size of lot area mean higher would-be land and higher would be the house price.

CONCLUSION

From the results we can conclude the following points:

- Quality grade matters the most for the price of house. Better the quality, higher would be the price.
- Furnishing status of house, whether it is furnished or not, affect the price of house. If house is furnished it increases the price of house and affects it very significantly.
- As latitude and longitude also affects the price, which means location also has a great affect on the price. The locality of the house matters which is defined by the latitude.
- Living square feet area of the house also matter for house price.
- House age matters for the price.
- Sea facing or Water front view houses are costlier than normal houses without ant such luxury.
- Population density also affects the model. Higher the population density, higher would be the price of house.
- Then sight viewed also affect the price. The higher number of sight viewed mean the price might cost more or is of higher price.
- Then living area after renovation is affects the price.
- Then floor area also adds its importance to price.
- Zipcode 98004, which is in city Bellevue Square of state Washington has highest contribution to price means price will be higher here.
- Then bathroom and conditions will affect the price.
- Important features in descending order.

	Imp
quality_grade	0.29212
furnished	0.16474
latitude	0.15766
living_area	0.06638
longitude	0.05998
house_age	0.05984
seaface	0.03223
population_density	0.02009
sight_viewed	0.01730
living_area_2015	0.01725
population	0.01047
lot_area	0.00951
floor_area	0.00943
zipcode_98004	0.00917
lot_area_2015	0.00863
bathroom	0.00666
condition	0.00493
total_floors	0.00335
basement_area	0.00326
renovation_yrs	0.00276

C. THE BUSINESS RECOMMENDATION:

Recommendation:

The models can be used to predict the price of the property with great efficiency and can lead to better financial decisions by the investor/seller.

The customers/sellers should focus on these variables in the decreasing order to decide for a suitable price. They should focus on quality, furnished, then location, sea facing house. The variables importance in order will let the customer play with their choices to get a better deal instead of random sacrifices.