Property price prediction and visualization

1st Bhavesh Kumar V D
Vellore Institute of Technology
bhaveshkumar.vd2019@vitstudent.ac.in

2nd Kshitij Upadhyay
Vellore Institute of Technology
kshitij.upadhyay2019@vitstudent.ac.in

3rd Mausami Karogal
Vellore Institute of Technology
mausamimahesh.karogal2019@vitstudent.ac.in

4th Keshav Bhatia
Vellore Institute of Technology
keshav.bhatia2019@vitstudent.ac.in

***Abstract*—Data visualization can be complicated if there isn't sufficient data for analysis. Analyzing the real estate market is the essential need for the hour. Various machine learning models can be applied to the raw data to extract relevant knowledge to predict the house prices and key housing attributes. Fluctuations in house prices can be a real concern to the house owners or real estate. This project analyses a county property by building a Linear Regression model following the OSEMN process to solve a problem and create a model to predict the house pricing in King County. OSEMN is an acronym that represents Obtain, Scrub, Explore, Model, and Interpret. Using the many data points in the dataset like data, price, area (represented at latitudes and longitudes), floors, view, conditions of the house, EDA (exploratory data analysis)is performed to obtain insight from the data and create a model to predict the housing price in King County. The analysis will help understand the factors contributing to prices in a certain area and explain what fields are responsible for the rise in the overall value of the area, and the areas that are suitable for affordability, including the many conclusions that can be drawn in the paper further. This study can benefit a lot of people, especially housing developers, to ascertain the prominent factors to determine house prices and visualize them to get a better understanding of the problem statement**

***Index Terms*— OSEMN, EDA, property price, King County**

Introduction

Individuals have always had a basic demand for real estate property. As people's living circumstances increased, the demand for housing grew dramatically. While some people buy a house as an investment or as a piece of property, the majority of people buy a house as a shelter or as a source of income. The housing markets have a favorable impact on the economy of the country. The building industry's growth and customers' ability to make substantial investments are signs of the country's high level of real estate. The quality of life of residents, as well as the national economy, is influenced by the possibility of rising property prices. In the end, this issue will affect investors who are considering purchasing a home as an investment.

Every year, there is an increase in housing demand, which leads to an increase in house prices. Most stakeholders, including purchasers and developers, housebuilders, and the real estate business, would like to know the exact features or accurate aspects influencing the house price to help investors make judgments and help housebuilders establish the house price. The property-prediction model has numerous advantages for home purchasers, investors, and builders. This model can assist potential buyers in determining the features of a home that they desire based on their budget.

The core tenet of all business is real estate value. The analytical method of influencing the present worth of an asset or an organization is referred to as valuation. However, valuations are required for a wide variety of reasons. However, valuations are carried out in this case to determine the most efficient method of calculating a company's selling price. To forecast housing values in King County, we create a model based on the OSEMN technique.

.

Problem statement

The explosive growth in house prices in high-cost cities is fueled by factors like scarcity of housing units, a growing number of high-income families in the United States, and high-income families willing to out-bid lower-income families for scarce housing in preferred locationsThere are many disparities in real estate prices because of many aspects such as locations, land value per sq feet, moreover individuals are forced to pay much more than the actual price due to the brokerage system.

This study will assist people in determining the exact value of their homes in their unique neighborhood as well as other economical possibilities, allowing them to make cost-effective decisions.

The visualizations will portray a real estate viewpoint that includes all of the facilities and amenities that are important to consider when looking for affordable housing.

DATASET

The dataset is of King County’s Housing price.

The variables used in the dataset are :
id - uniquely identified for a house
date – Date house was sold
price – Price is the prediction target
bedrooms – number of bedrooms/house
bathrooms – number of bathrooms/house
sqft_living – square footage of the home
sqft_lot- square footage of the lot
floors – total floors(levels) in the house
waterfront – house which has a view to a waterfront
view – the house that has been viewed
condition – overall condition of the house
grade – overall grade is given to the housing unit (based on King County’s grading system
sqft_above – square footage of house apart from the basement
sqft_basement- square footage of the basement
yr_built – built year
yr_renovated – year when the house was renovated
zip code – zip code of the place
lat – latitude coordinate
long – longitude coordinate
sqft_living15 – square footage of interior housing living for the nearest 15 neighbors
sqft_lot15 - square footage of land lots of nearest 15 neighbors

RELATED WORKS

In the 21st century, real estate became more than a necessity, and now not only for those who are trying to buy more real estate but also for the companies that sell it. According to [4], real estate is not only a basic human need but also represents the wealth and reputation of today's people. Investing in real estate generally seems to be beneficial, as the value of the real estate does not decline immediately. Changes in real estate prices can affect a variety of household investors, bankers, policymakers, and more. Investing in the real estate sector seems to be an attractive option to invest in. Therefore, asset value forecasting is an important economic indicator. Supervised learning is learning that teaches or trains a machine using well-labeled data. In other words, some data already have the correct answer. A new set of examples will then be available on the machine, and the monitored learning algorithm will analyze the training data and produce the correct results from the marked data. Unsupervised learning is training a machine with unclassified and unlabelled information that allows the algorithm to operate on that information without guidance. The task of the machine here is to group information that is not sorted according to similarities, patterns, and differences without pre-training the data. In contrast to supervised learning, teachers are not offered. That is, the machine is not trained. Machine learning has many uses, one of which is to predict real estate. The real estate market is one of the most competitive in terms of pricing and tends to fluctuate significantly due to many factors in finding a strategy and deciding on the right strategy, so machines to maintain prices It will be one of the main areas to apply the concept of learning. High optimization and accuracy prediction. A study of land price trends is considered important to support city planning decisions. The real estate system is an unstable stochastic process. Investor decisions are based on market trends for maximum returns. Developers are interested in knowing future trends in decision-making. A large amount of data is required for analysis, modeling, and forecasting to accurately estimate real estate prices and future trends, which affects real estate prices.

The factors that influence the price of land need to be investigated, and the price impact needs to be modeled. You should consider analyzing historical data. Therefore, establishing a simple linear mathematical relationship to this time series data is not practical for prediction. Therefore, it has become essential to create nonlinear models that adapt well to the data characteristics in order to analyze and predict future trends. Due to the rapid development of the real estate sector, analyzing and forecasting real estate prices using mathematical modeling and other scientific methods is an urgent decision-making need for all involved.

Population growth and industrial activity are due to many factors, most notably the recent boom in the knowledge sector. Information Technology (IT) and Information Technology-based services. Land demand has been on the rise and housing and real estate activities have begun to flourish. All barren lands and rice fields are gone to pave the way for multiple stores and skyscrapers. Investment in the real estate industry has begun to flow, and over the years there has been no consistent pattern of land prices. Everyone in the industry felt the need to predict land price trends. Governments, regulators, credit institutions, developers, investors. Therefore, this article introduces some important features that can be used to accurately predict real estate prices. Regression models that use different functions can be used to reduce the error in the residual sum of squares. When using features in a regression model, some feature engineering is required to make better predictions. Often, a set of features with multiple regressions or polynomial regressions (features with different powers applied) is used to achieve a better model fit. These models are expected to tend to be overfitted to reduce ridge regression. Therefore, it is advisable to optimally use the regression model in addition to other techniques to optimize the results. The most widely used method for explaining and predicting property performance is multiple regression analysis. This technique is then extended to hedonic regression and used in simultaneous equation systems. Multiple regression analysis (MRA is based on correlation analysis. In general, you can use correlation and regression to perform multivariate analysis on relatively small samples. Steven (1992) points out that the strength of MRA is primarily in determining the relative importance of the independent variable to the dependent variable (the dependent variable is an explained phenomenon and the independent variable is used. It is a factor that explains the dependent variable to be used). The benefits of MRA can only be seen by understanding the basic concepts of MRA. The classic form of MRA is used for prediction. However, in more recent practice, MRA is usually used to explain the subject of research. They use the MRA as a causal model to explain the changes in the independent variables, explain the changes in the dependent variables, and evaluate the relative importance of each independent variable. Each regression coefficient estimates the amount of change that occurs in the dependent variable with respect to the unit change of the independent variable in the equation.

METHODOLOGY

In this project we will follow the OSEMN process to solve a problem and create a model to predict the house pricing in King County. OSEMN is an acronym that represents Obtain, Scrub, Explore, Model, and iNterpret steps.

Obtain

This step requires understanding the problem, collecting and obtaining the data that you need.

Scrub

The main tasks at this stage are data cleansing and filtering.You might need to change some data formats to standardize all across the dataset. Take care of null values, placeholders etc. In this step, you might also derivate new columns from the current data to make better use of them in your model.In most cases, exploration and scrubbing are done at the same time.. After dealing with null values and placeholders, you wou;d want to do the exploring. Although scaling and normalisation is a part of the scrubbing step, it is more accurate to do the exploring part before doing all those. Feature engineering is also part of the scrubbing step which selects the columns that will be used in your model. Therefore, you should use them in the exploration step before eliminating the gaps.

Explore

This step focuses on studying the data you have. Understand each column, check the data type, and become familiar with the dataset. Searching for records raises some great business questions and you need to use the data to answer them. These questions may be asked to you, or it may be up to you to understand the dataset you have. You need to be able to get a visualization of your data that will help you see patterns and trends in your data.

Model

This step includes building your model, iterating with different features to get better accuracy, validate with different validation techniques and train your model.

Interpret

Interpreting data is one of the most important steps. It refers to presenting your data, the business question that needs to be answered, the answer with the visualisation and accountable insights that are found through the process. Interpreting data means creating a clear, easy to understand story from your results for your non-technical audience.

Solving the questions taken :
Here we take few questions and we'll try to solve them

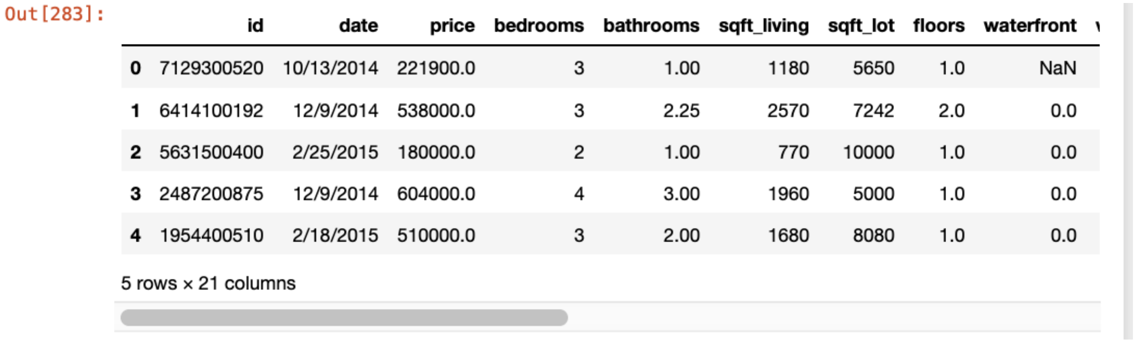
Correlation :
We check the correlation between the data we have for analysis in this part many correlations will be checked here

Scaling & Normalisation :
Here we scale and normalise lots of graphs for our analysis purposes

Risk Analysis :
With the huge amount of data in sales, increasing the efficiency of link prediction methods is a major challenge.To address this problem, I also cross checked the model

Analysis

Dataset Study



Out[283]:

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront
0	7129300520	10/13/2014	221900.0	3	1.00	1180	5650	1.0	NaN
1	6414100192	12/9/2014	538000.0	3	2.25	2570	7242	2.0	0.0
2	5631500400	2/25/2015	180000.0	2	1.00	770	10000	1.0	0.0
3	2487200875	12/9/2014	604000.0	4	3.00	1960	5000	1.0	0.0
4	1954400510	2/18/2015	510000.0	3	2.00	1680	8080	1.0	0.0

5 rows × 21 columns

Data Scrubbing

First thing I will check is the null values: I will use .isna() for that purpose. Before that, I will delete the id column which I do not need this feature for this project.

In [288]:
```
df.drop(['id'], axis=1, inplace=True)
```

In [289]:
```
df.isna().sum()
```

Out[289]:
```
date                0
price               0
bedrooms            0
bathrooms           0
sqft_living         0
sqft_lot            0
floors              0
waterfront       2376
view               63
condition           0
grade               0
sqft_above          0
sqft_basement       0
yr_built            0
yr_renovated     3842
zipcode             0
lat                 0
long                0
sqft_living15       0
sqft_lot15          0
dtype: int64
```

The view feature has only 63 missing value, however, waterfront has 2376, and yr_renovated has 3842 null values out of 21597 observations. I will check each one of them individually to see what I can do for those missing values.

Null values

waterfront

The discription of waterfront is "House which has a view to a waterfront". So it must be 1s and 0s. Lets see how the values are distributed. let's check the values and their counts.

In [300]:
```
df.waterfront.value_counts()
```

Out[300]:
```
0.0    19075
1.0      146
Name: waterfront, dtype: int64
```

Exploration

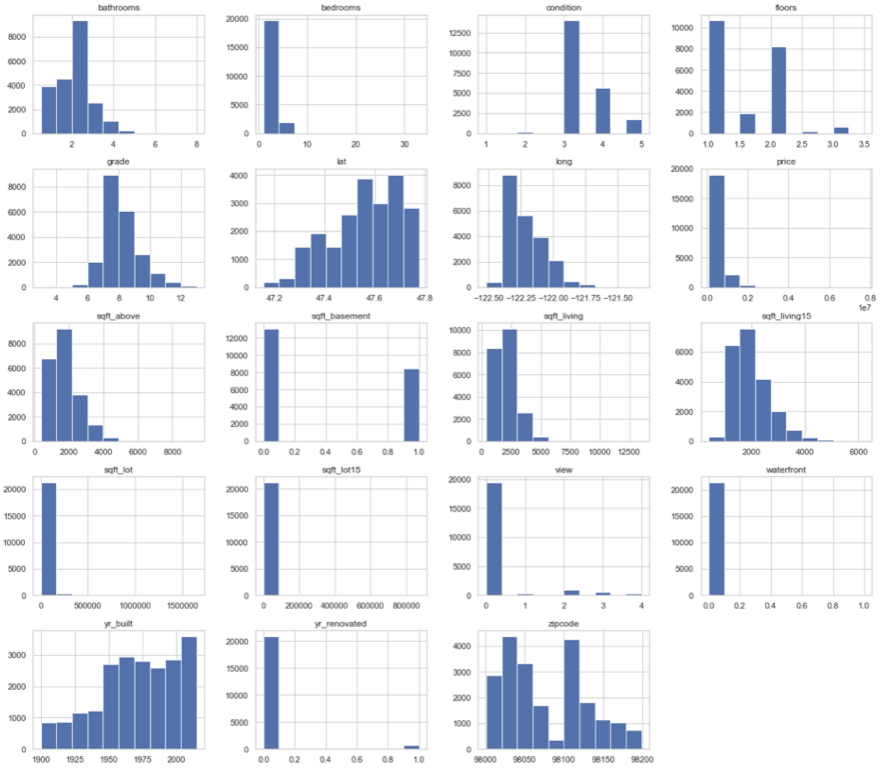
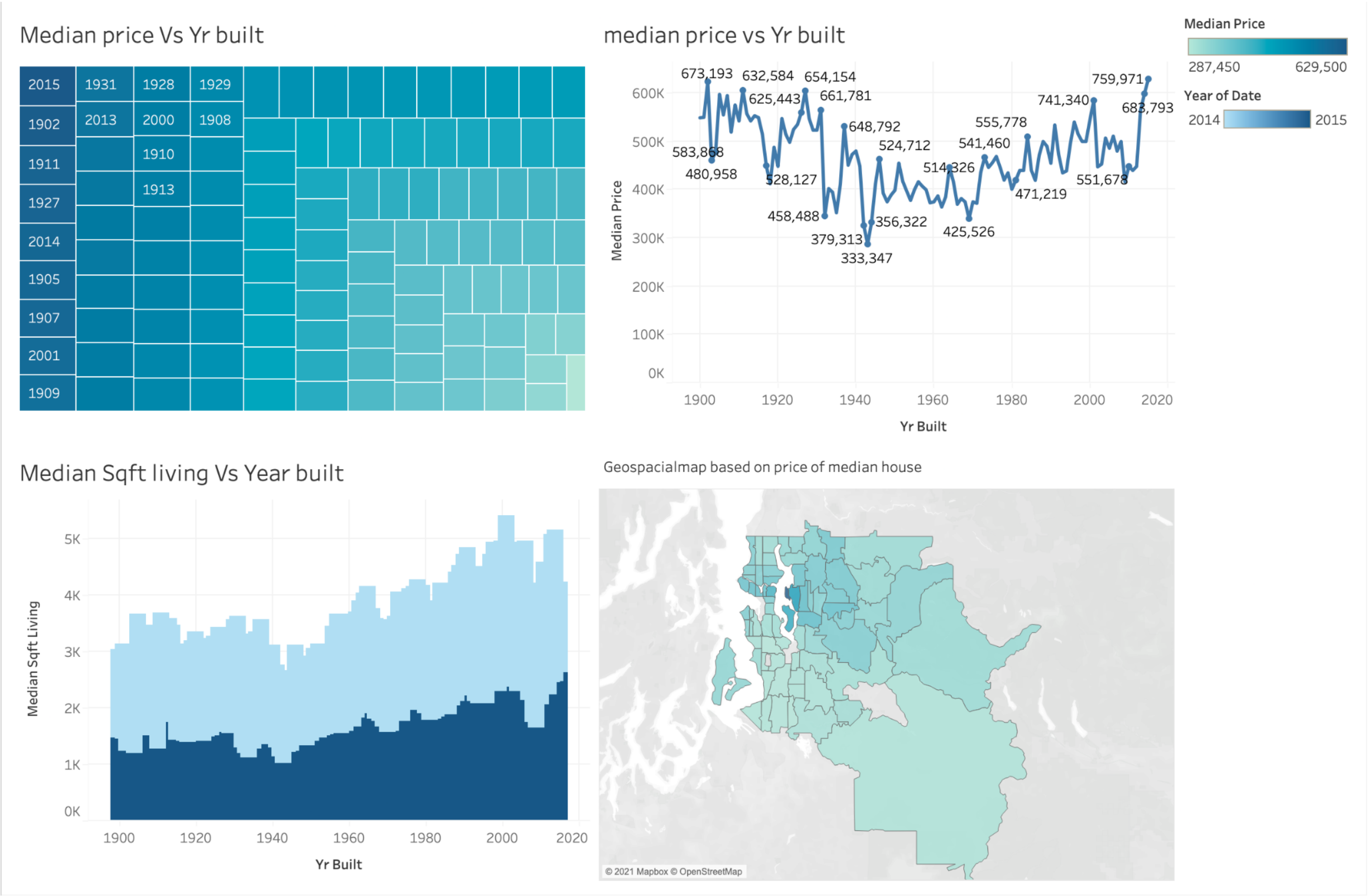
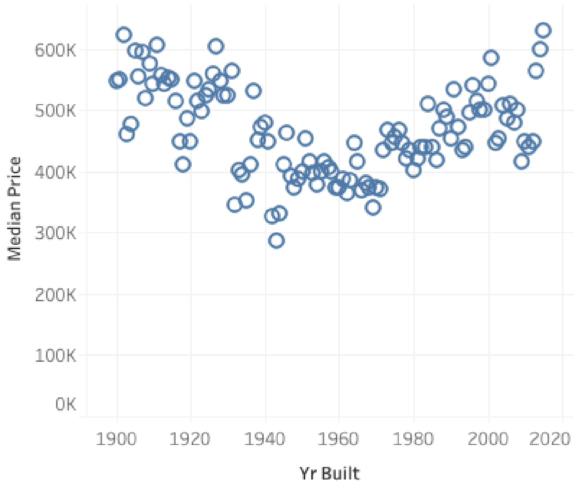


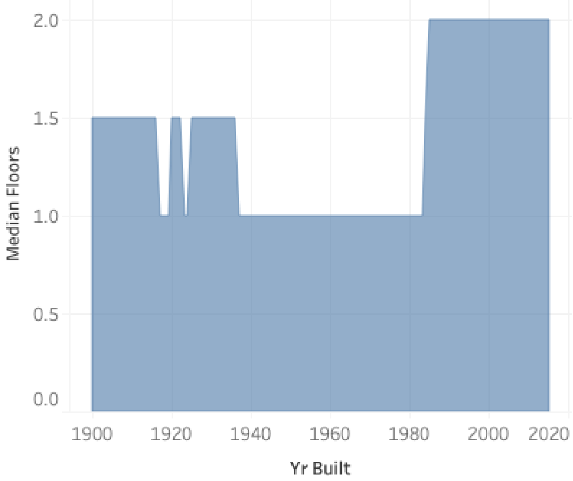
Tableau Dashboards



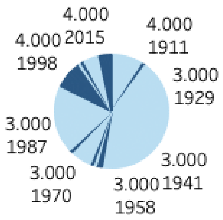
Price vs Year Built



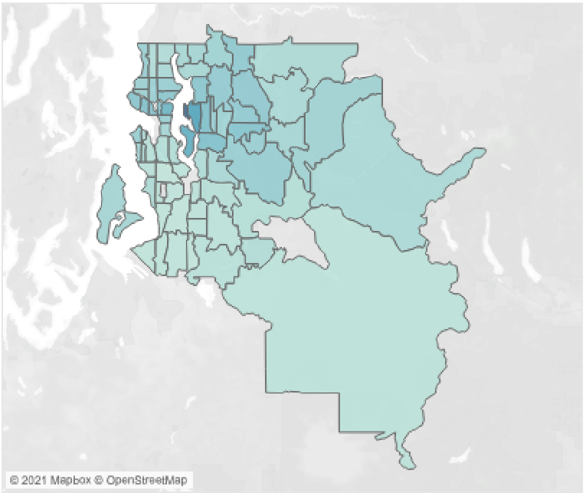
Floors vs Year Built

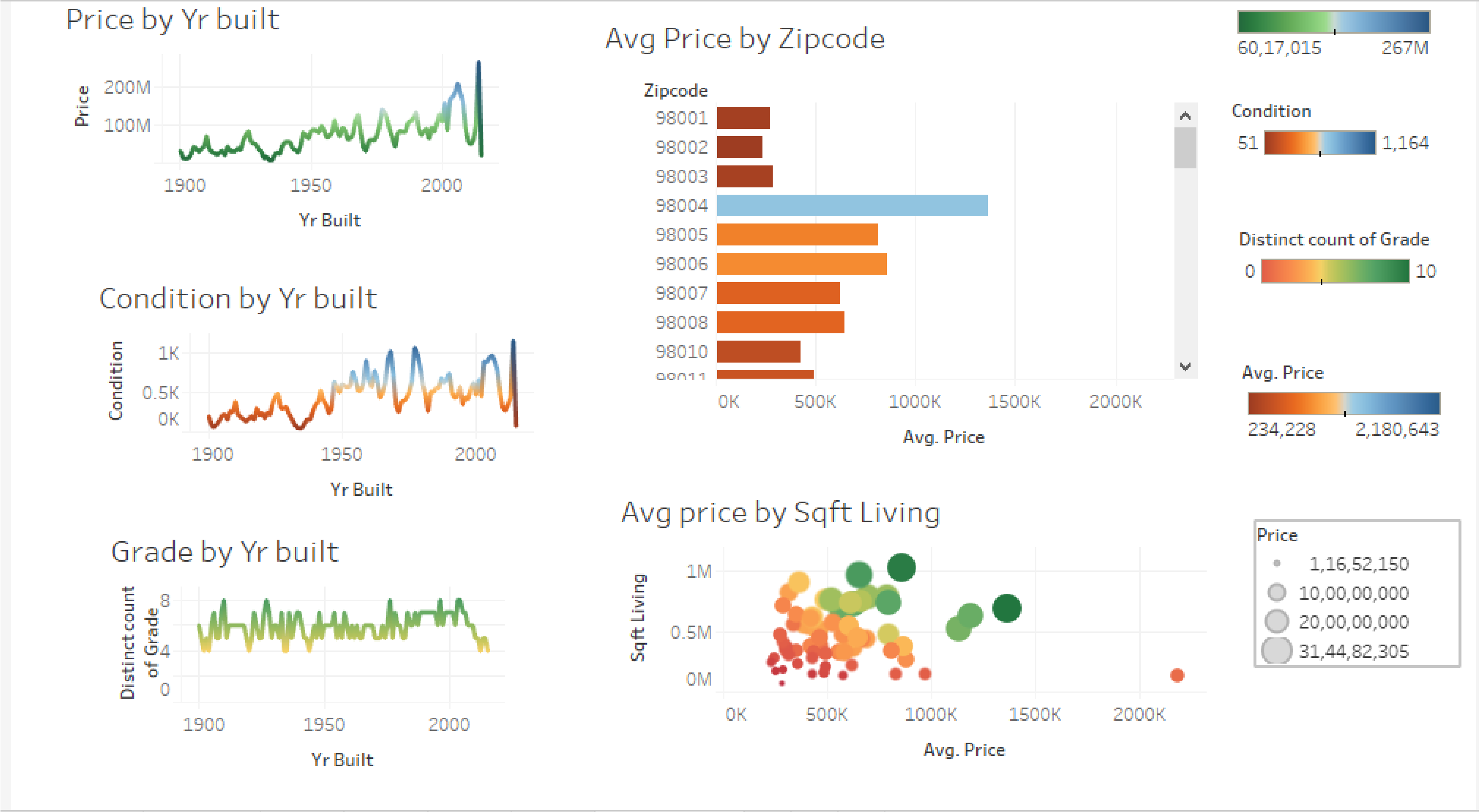


Number of Bedrooms Year Built

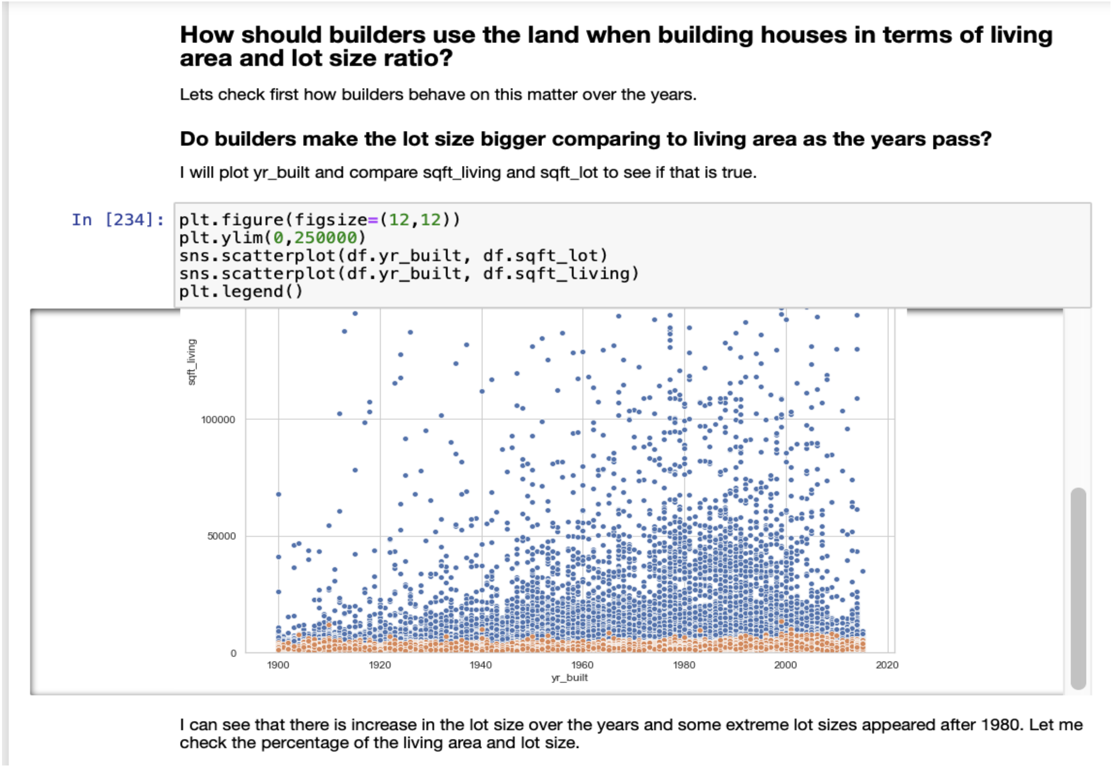


Price According to Zipcode




Price by Yr built
Price
200M
100M
1900
1950
2000
Yr Built
Condition by Yr built
Condition
1K
0.5K
0K
1900
1950
2000
Yr Built
Grade by Yr built
Distinct count of Grade
8
4
0
1900
1950
2000
Yr Built
Avg Price by Zipcode
Zipcode
98001
98002
98003
98004
98005
98006
98007
98008
98010
98011
0K
500K
1000K
1500K
2000K
Avg. Price
Avg price by Sqft Living
Sqft Living
1M
0.5M
0M
0K
500K
1000K
1500K
2000K
Avg. Price
60,17,015
267M
Condition
51
1,164
Distinct count of Grade
0
10
Avg. Price
234,228
2,180,643
Price
1,16,52,150
10,00,00,000
20,00,00,000
31,44,82,305

Inference



How should builders use the land when building houses in terms of living area and lot size ratio?

Lets check first how builders behave on this matter over the years.

Do builders make the lot size bigger comparing to living area as the years pass?

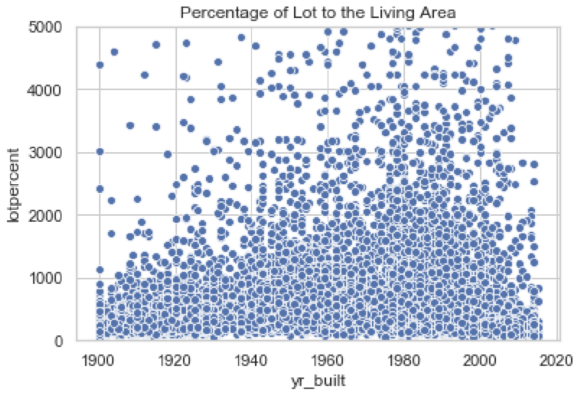
I will plot yr_built and compare sqft_living and sqft_lot to see if that is true.

In [234]:

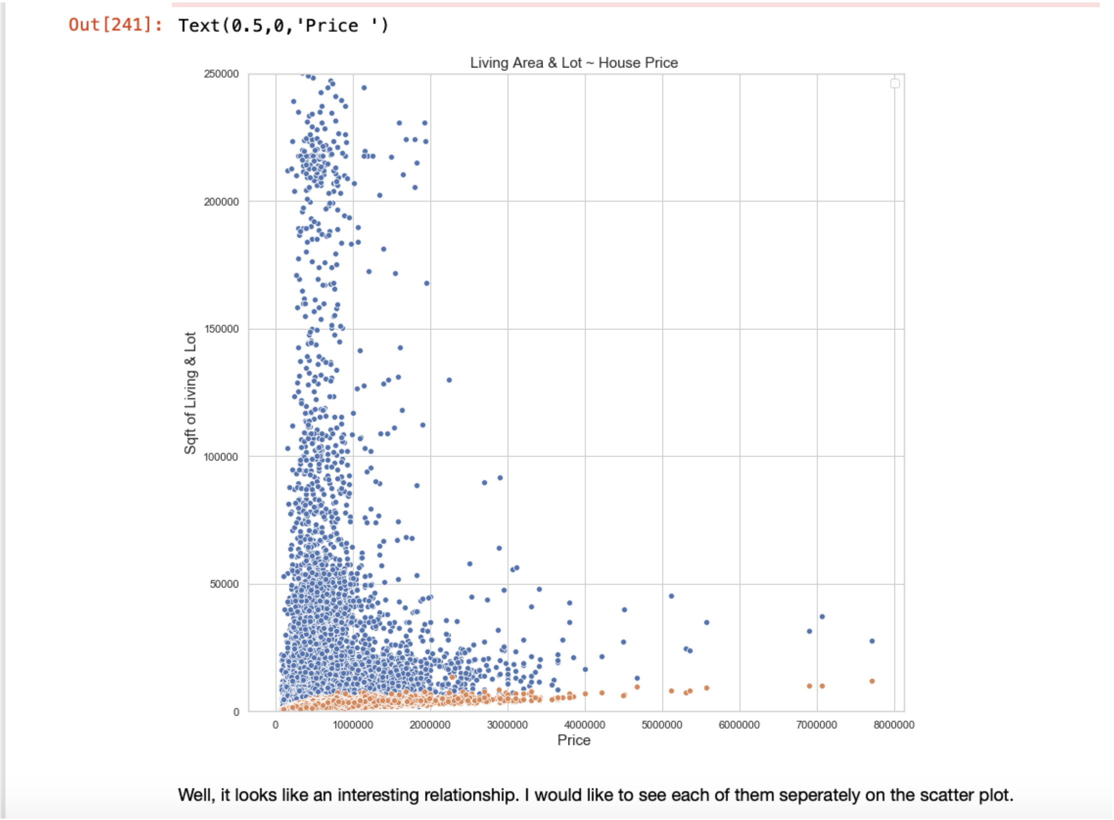
```
plt.figure(figsize=(12,12))
plt.ylim(0,250000)
sns.scatterplot(df.yr_built, df.sqft_lot)
sns.scatterplot(df.yr_built, df.sqft_living)
plt.legend()
```

I can see that there is increase in the lot size over the years and some extreme lot sizes appeared after 1980. Let me check the percentage of the living area and lot size.

```
Out[236]: Text(0.5,1,'Percentage of Lot to the Living Area')
```



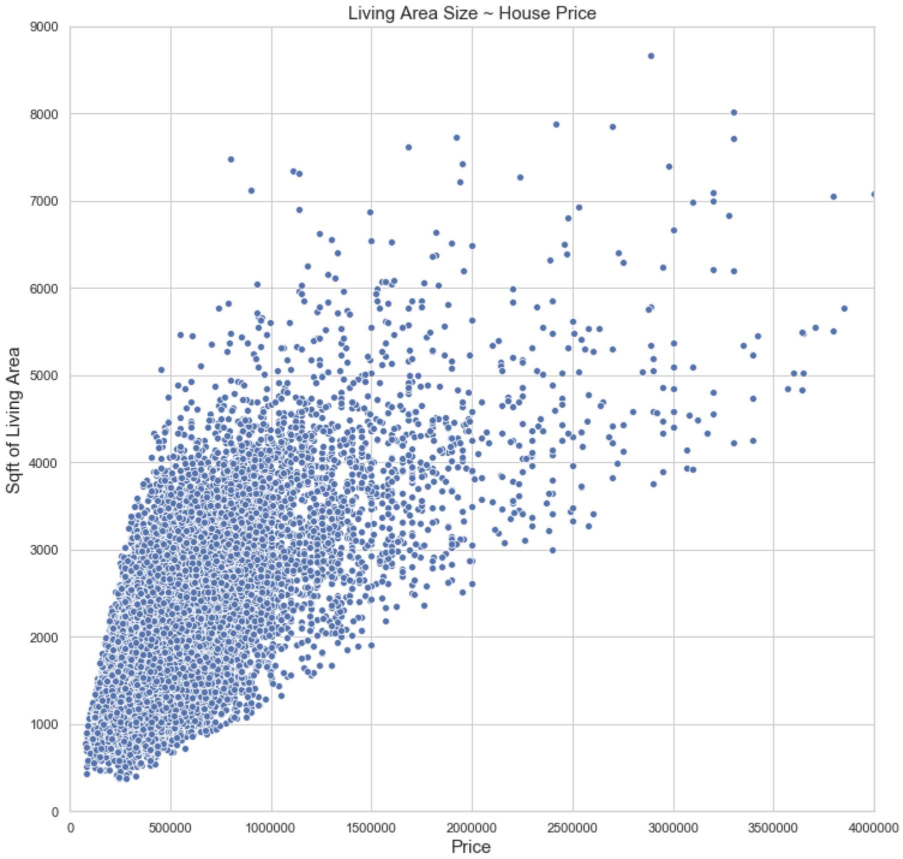
According to this scatter plot, the number of houses that has large lots has clearly increased over the years. The ratio of lot size to living area has also increased. The increase we see might be a real "Yes " to my question. But I want to see what is the ratio of the houses that have large lots to the moderate ones. Or, is this an increase of overal number of houses built that year? I would like to check the number of the houses that lot percent is bigger than %500 of the living area. Here is my next question;



```
Out[241]: Text(0.5,0,'Price ')
```

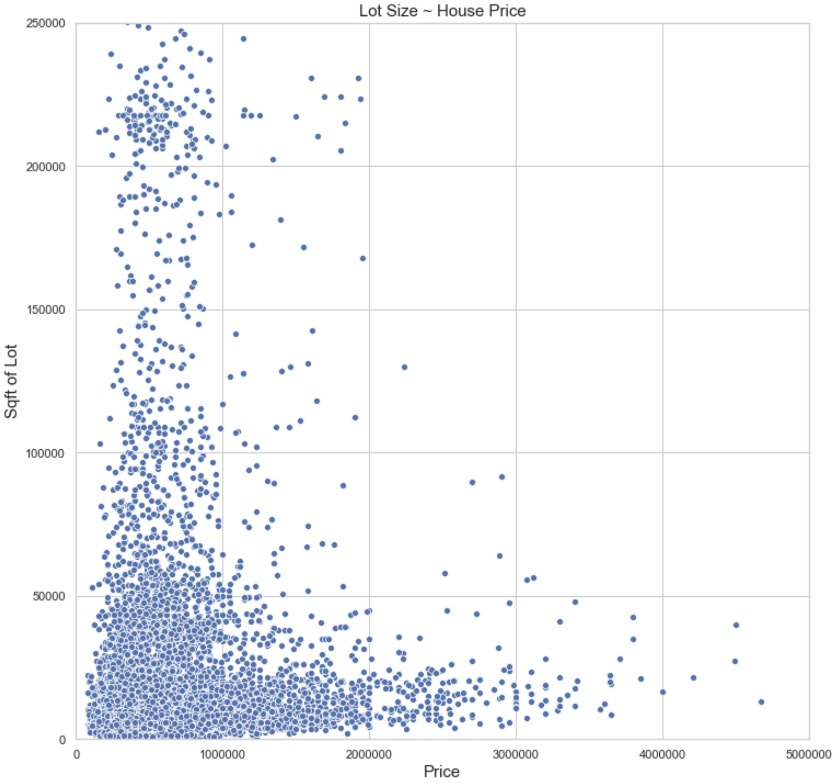
Well, it looks like an interesting relationship. I would like to see each of them seperately on the scatter plot.

```
Out[242]: Text(0.5,0,'Price')
```



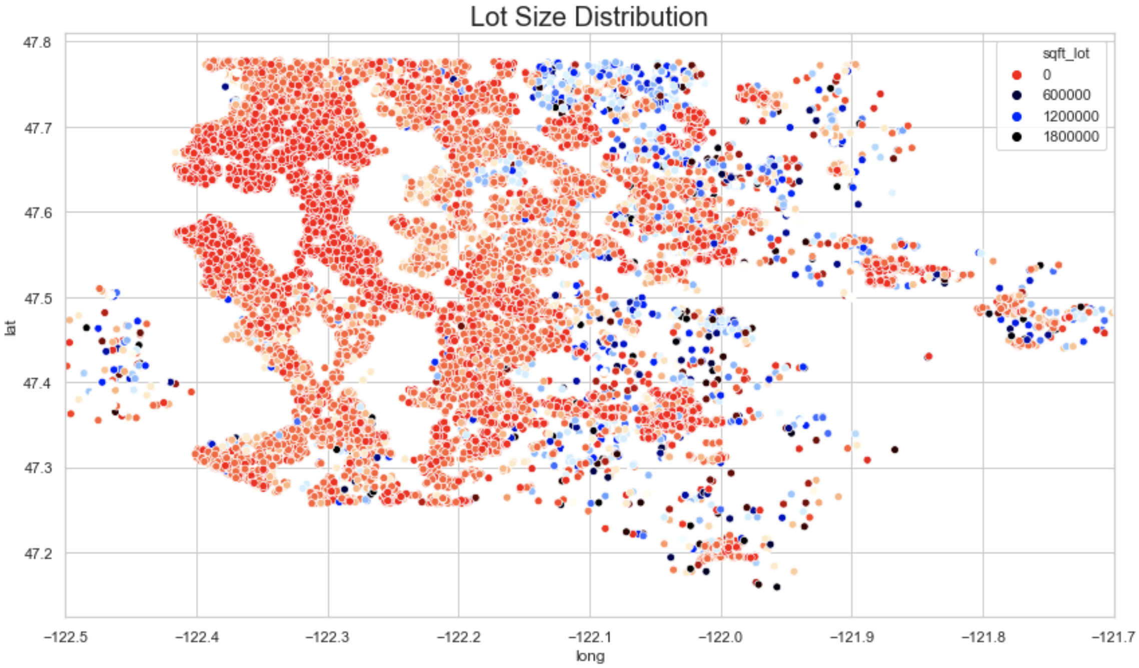
Price and sqft_living have pozitif and a kind of linear relationship.

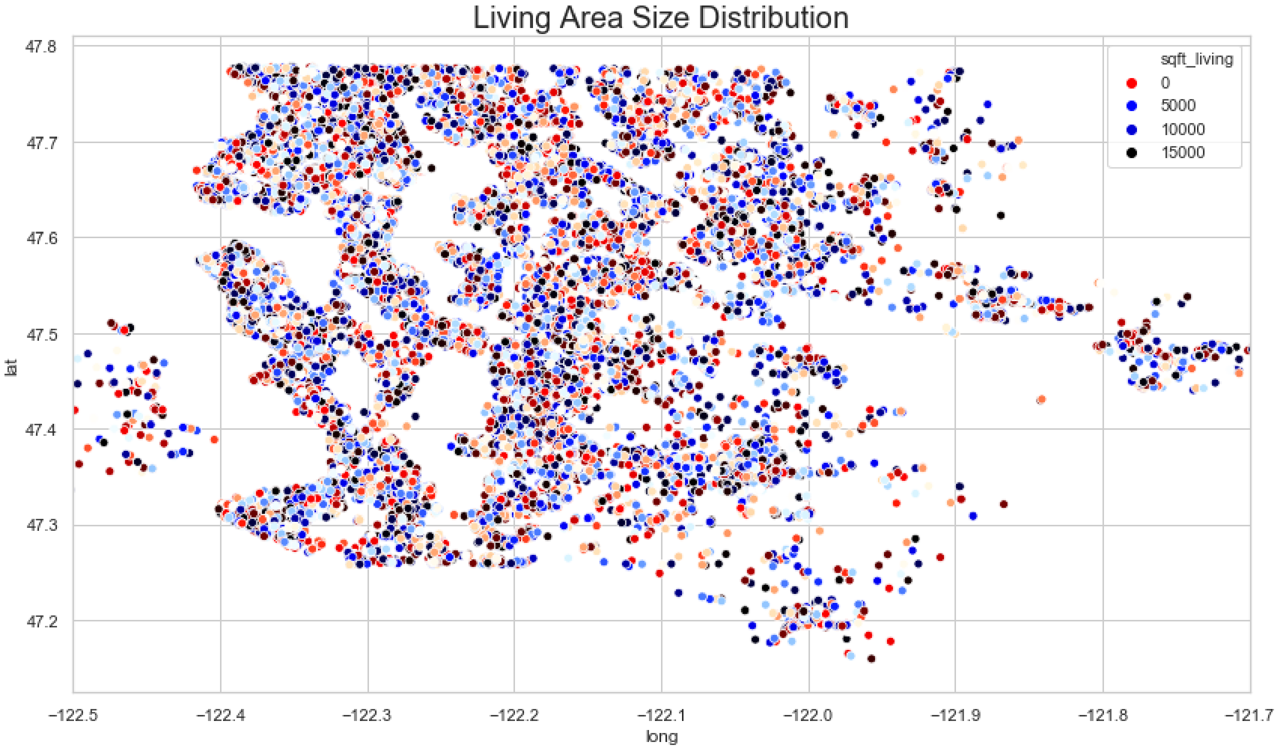
```
Out[243]: Text(0.5,0,'Price')
```

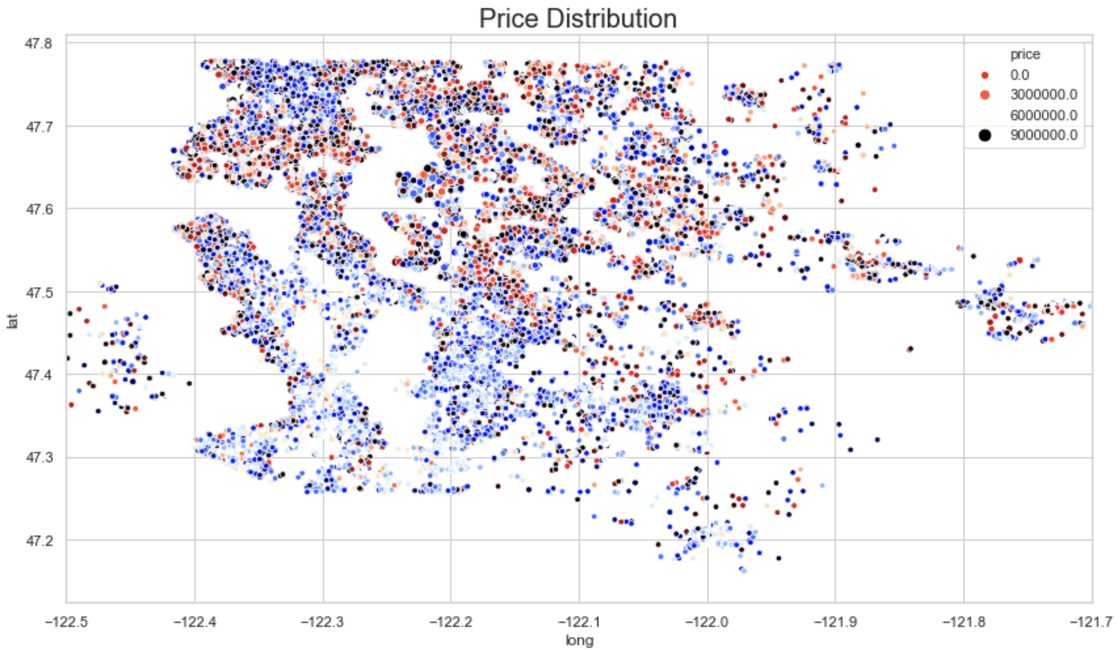


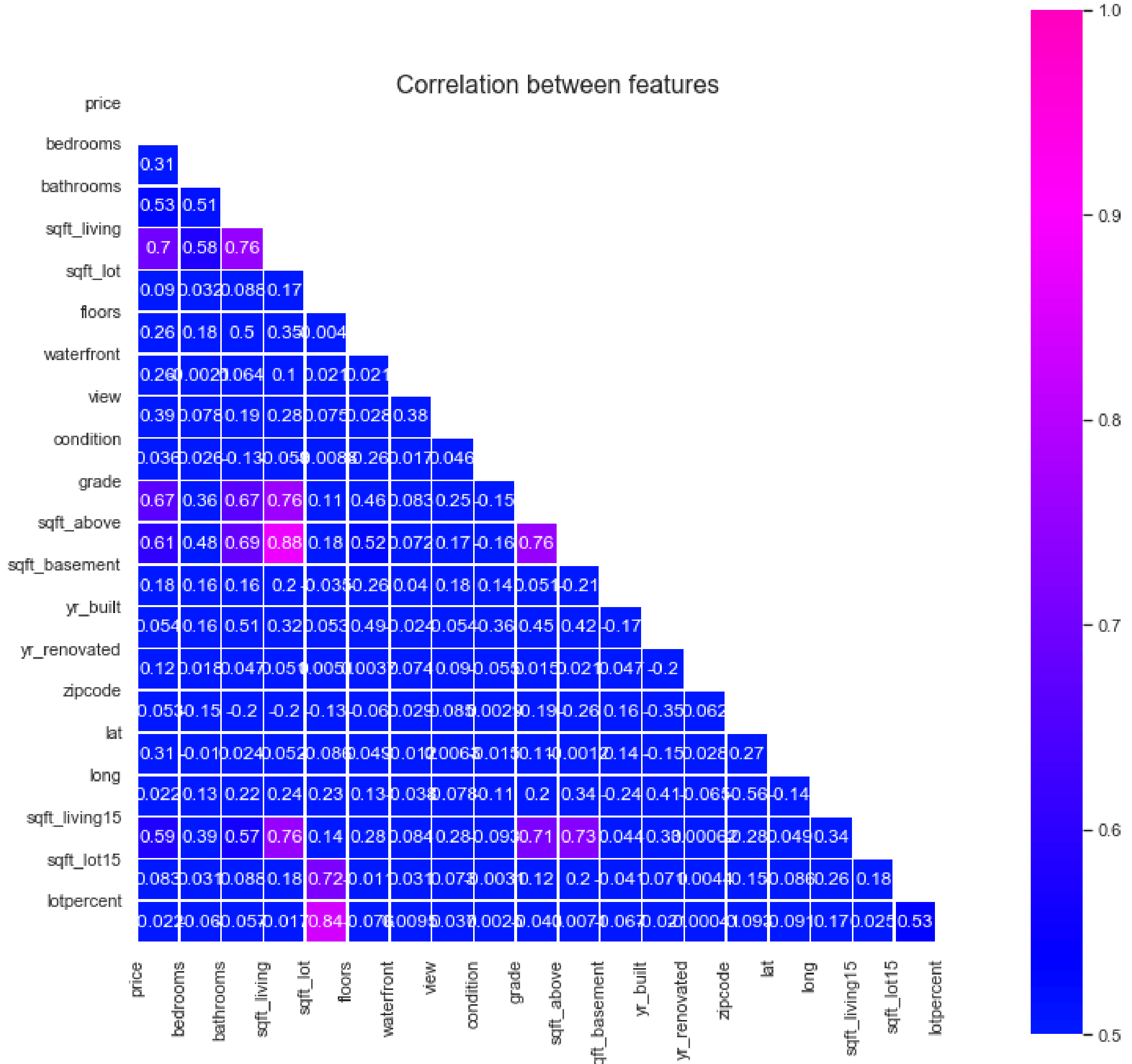
From these plots, I can say that there is a positif and more linear relationship between price and living area size.But for lot size and price, the relationship is different. For most of the houses have large lots, price is in 1 million range. And there are many houses with small lots price range is very high. That sounds like there is not significant direct relationship between price and lot size. So, having larger lot might not be everyones demand, but larger living area size seems preferable for most of the buyers. In this case the next question would be where might be a good choise to built a house with bigger lot.

House Distribution






Price Distribution
price
0.0
3000000.0
6000000.0
9000000.0
lat
47.8
47.7
47.6
47.5
47.4
47.3
47.2
long
−122.5
−122.4
−122.3
−122.2
−122.1
−122.0
−121.9
−121.8
−121.7


Correlation between features
price
bedrooms 0.31
bathrooms 0.53 0.51
sqft_living 0.7 0.58 0.76
sqft_lot 0.09 0.032 0.088 0.17
floors 0.26 0.18 0.5 0.35 0.004
waterfront 0.26 0.002 0.064 0.1 0.021 0.021
view 0.39 0.078 0.19 0.28 0.075 0.028 0.38
condition 0.036 0.026 -0.13 0.059 0.0089 -0.26 0.017 0.046
grade 0.67 0.36 0.67 0.76 0.11 0.46 0.083 0.25 -0.15
sqft_above 0.61 0.48 0.69 0.88 0.18 0.52 0.072 0.17 -0.16 0.76
sqft_basement 0.18 0.16 0.16 0.2 -0.035 -0.26 0.04 0.18 0.14 0.051 -0.21
yr_built 0.054 0.16 0.51 0.32 0.053 0.49 -0.024 -0.054 -0.36 0.45 0.42 -0.17
yr_renovated 0.12 0.018 0.047 0.051 0.005 0.0037 0.074 0.09 -0.059 0.015 0.021 0.047 -0.2
zipcode -0.053 -0.15 -0.2 -0.2 -0.13 -0.06 0.029 0.085 0.0029 -0.19 -0.26 0.16 -0.35 0.062
lat 0.31 -0.01 0.024 0.052 -0.086 0.049 -0.012 -0.0063 -0.015 0.11 -0.0012 0.14 -0.15 0.028 0.27
long 0.022 0.13 0.22 0.24 0.23 0.13 -0.038 -0.078 -0.11 0.2 0.34 -0.24 0.41 -0.065 -0.56 -0.14
sqft_living15 0.59 0.39 0.57 0.76 0.14 0.28 0.084 0.28 -0.093 0.71 0.73 0.044 0.33 -0.00062 -0.28 0.049 0.34
sqft_lot15 0.083 0.031 0.088 0.18 0.72 -0.011 0.031 0.073 -0.0031 0.12 0.2 -0.04 0.071 0.0044 -0.15 -0.086 0.26 0.18
lotpercent -0.022 -0.06 -0.057 -0.017 0.84 -0.076 0.0099 0.037 0.0024 -0.04 -0.007 -0.067 0.02 -0.00041 -0.093 -0.091 0.17 0.025 0.53
price
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
lotpercent
1.0
0.9
0.8
0.7
0.6
0.5

Out[266]:

OLS Regression Results

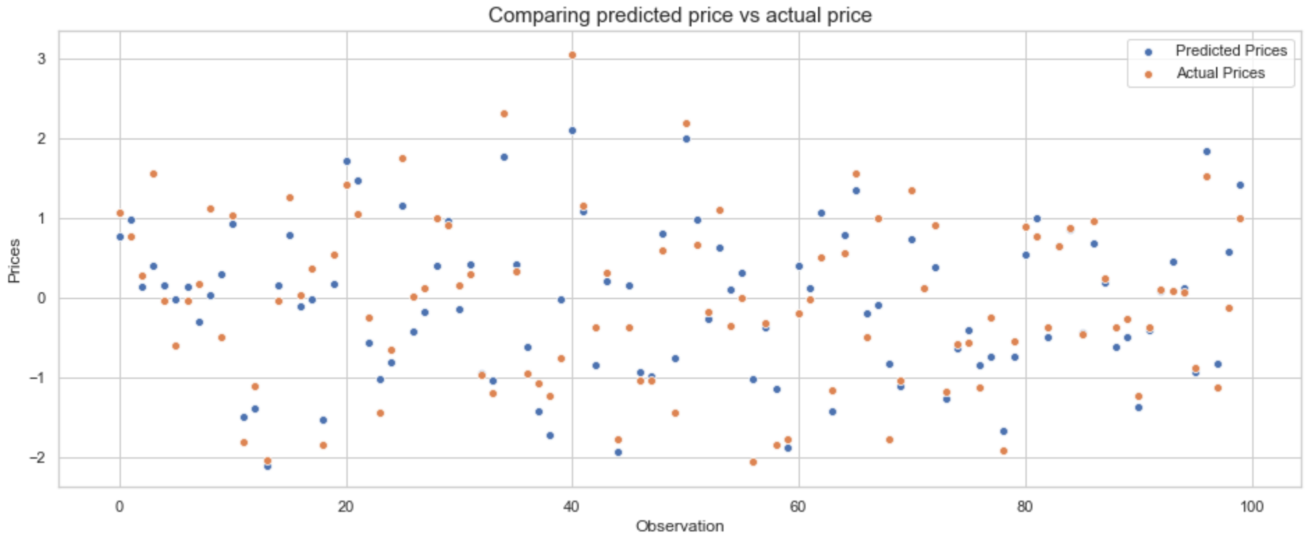
Dep. Variable:	price	**R-squared:**	0.810
Model:	OLS	**Adj. R-squared:**	0.809
Method:	Least Squares	**F-statistic:**	1950.
Date:	Tue, 09 Jul 2019	**Prob (F-statistic):**	0.00
Time:	20:01:25	**Log-Likelihood:**	-12732.
No. Observations:	21597	**AIC:**	2.556e+04
Df Residuals:	21549	**BIC:**	2.594e+04
Df Model:	47		
Covariance Type:	nonrobust		

The model looks pretty good for an initial fit! We get a R-squared of 0.810 and an adjusted R-squared of 0.809. The contribution of the categorical featues is totaly positive to the model.

The p-values for the continuous variables look pretty good except sqft_lot15. Only a few categorical variables have p-values greater than the 0.05 confidence threshold. I will eliminate some of these values in the next iteration of the model by using only the features has p values smaller than 0.05 and observe how it affects the goodness of fit.

```
R^2 Score: 0.81
Train Mean Squarred Error: 0.1904857891530859
Test Mean Squarred Error: 0.19074674220485718
```

Comparing predicted price vs actual price



Conclusion

This model can make 81% accurate prediction for a house priceFeatures that go through the model are Location (latitude and longitude), Number of bedrooms, Living, lot and basement size, Number of views the house get,Year of built, Number of floors, condition

Most significant features are

Location, condition, living area and waterfront effects a house price more than any other features.

How does that work?

Each features effect the model result based on their coefficents. Coefficient define how many unit a feature change the target as it increase one unit. According to this model, Latitude and longitude ~ upto 1.46 unit)Condition grade higher than 3 ~ (3 :0.8 unit, 4: 0.9 unit, 5:1.0 unit,)View (the most significant view is waterfront ~ 0.63 unit, Living area size ~ 0.48 unit

Findings

Larger lots does not increase the price of a house directly. But the larger house size always increase the house price. Choose Seattle, East Urban and East Rural areas to built houses. Always keep the living area large. Keep the lot size minimum in Seattle and East Urban area.Built relatively larger lots in South Urban, west of East Rural and west of South Rural area. Have options for huge lots for farm houses in the east side of the county.

Limitations

requires a lot of data

only 81% efficiency

specific to a particular location

References

1 A. S. Temür, M. Akgün, and G. Temür, “Predicting Housing Sales in Turkey Using Arima, Lstm and Hybrid Models,” J. Bus. Econ. Manag., vol. 20, no. 5, pp. 920–938, 2019, doi: 10.3846/jbem.2019.10190..

2 G. Gao et al., “Location-Centered House Price Prediction: A Multi-Task Learning Approach,” pp. 1–14, 2019

3 T. D. Phan, “Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia,” Proc. - Int. Conf. Mach. Learn. Data Eng. iCMLDE 2018, pp. 8–13, 2019, doi: 10.1109/iCMLDE.2018.00017.

4 Mark, A.S., & John, W.B. Estimating Price Paths for Residential Real Estate. Journal of Real Estate Research; 2003: 25, 277–300.

5 Zhangming, H. Research on Forecasting Real Estate Price Index Based on Neural Networks. Journal of the Graduates Sun Yat Sen University, 2006;27.

6 W. T. Lim, L. Wang, Y. Wang, and Q. Chang, “Housing price prediction using neural networks,” 2016 12th Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov. ICNC-FSKD 2016, pp. 518–522, 2016, doi: 10.1109/FSKD.2016.7603227.

7 T. D. Phan, “Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia,” Proc. - Int. Conf. Mach. Learn. Data Eng. iCMLDE 2018, pp. 8–13, 2019, doi: 10.1109/iCMLDE.2018.00017.

8 A. Yusof and S. Ismail, “Multiple Regressions in Analysing House Price Variations,” Commun. IBIMA, vol. 2012, pp. 1–9, 2012, doi: 10.5171/2012.383101.Sample

9 A. Varma, A. Sarma, S. Doshi, and R. Nair, “House Price Prediction Using Machine Learning and Neural Networks,” Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2018, pp. 1936–1939, 2018, doi: 10.1109/ICICCT.2018.8473231

10 R. Reed, “The relationship between house prices and demographic variables: An Australian case study,” Int. J. Hous. Mark. Anal., vol. 9, no. 4, pp. 520–537, 2016, doi: 10.1108/IJHMA-02-2016-0013.

11 V. Limsombunchai, ―House price prediction: Hedonic price model vs. artificial neural network, Am. J , 2004

12 R Manjula - Real estate value prediction using multivariate regression models Materials Science and Engineering Conference Series, volume 263, issue 4,2017

13 H Yu, J Wu- Real estate price prediction with regression and classification CS 229 Autumn, 2016

14 Nihar Bhagat, Ankit Mohokar, Shreyash Mane - House Price Forecasting using Data Mining International Journal of Computer Applications, 2016

15 Debanjan Banerjee, Suchibrota Dutta- Predicting the housing price direction using machine learning techniques, 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering