Analysis of house price prediction project by

using machine learning

* First thing first , we import our libraries and dataset and then we see the head of the data to know how the data looks like and use describe function to see the percentile’s and other key statistics.
* What can we infer from the above describe function ?

1. Look at the bedroom columns , the dataset has a house where the house has 33 bedrooms , seems to be a massive house and would be interesting to know more about it as we progress.
2. Maximum square feet is 13,450 where as the minimum is 290. we can see that the data is distributed.

* Similarly , we can infer so many things by just looking at the describe function.
* Now , we are going to see some visualization and also going to see how and what can we infer from visualization.

## Which is the most common house (Bedroom wise) ?

Sometimes it’s important for a builder to see which is the highest selling house type which enables the builder to make house based on that. Here in India , for a good locality a builder opts to make houses which are more than 3 bedrooms which attracts the higher middle class and upper class section of the society.

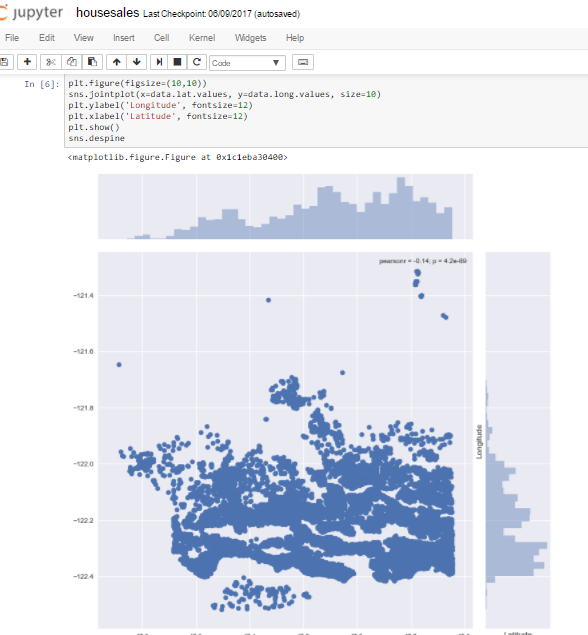


As we can see from the visualization 3 bedroom houses are most commonly sold followed by 4 bedroom. So how is it useful ? For a builder having this data , He can make a new building with more 3 and 4 bedroom’s to attract more buyers.

So now we know that 3 and 4 bedroom’s are highest selling

## Visualizing the location of the houses based on latitude and longitude

So according to the dataset , we have latitude and longitude on the dataset for each house. We are going to see the common location and how the houses are placed.

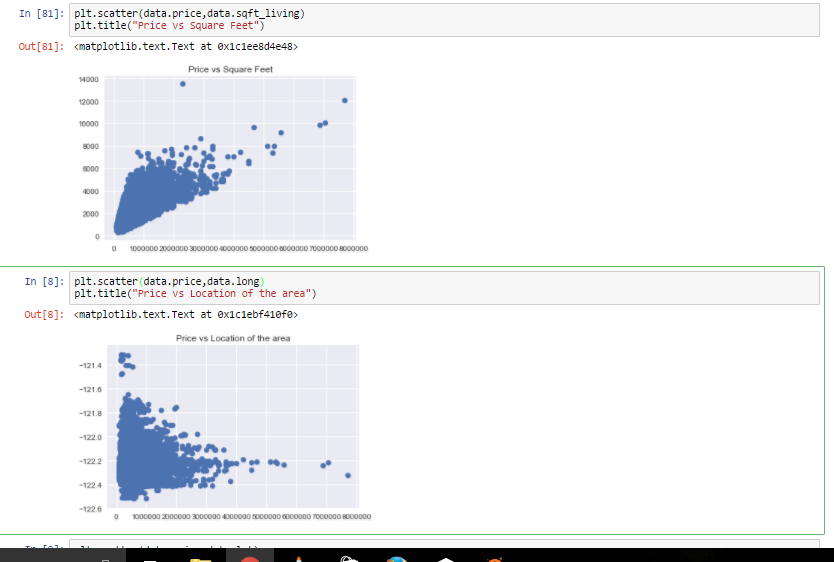


We use seaborn , and we get his beautiful visualization. Joinplot function helps us see the concentration of data and placement of data and can be really useful. Let us see what we can infer from this visualization. For latitude between -47.7 and -48.8 there are many houses , which would mean that maybe it’s an ideal location isn’t it ? But when we talk about longitude we can see that concentration is high between -122.2 to -122.4. Which would mean that most of the buy’s has been for this particular location.

* **How common factors are affecting the price of the houses ?**

We saw the common locations and now we’re going to see few common factors affecting the prices of the house and if so ? then by how much ?

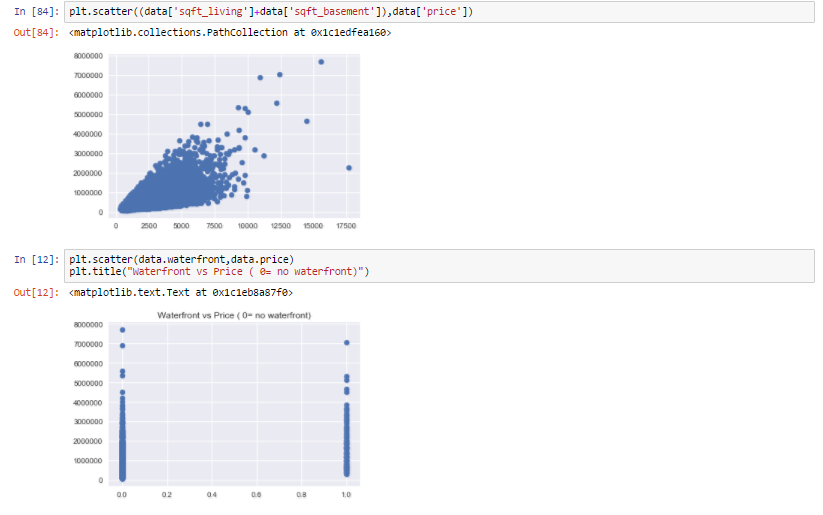
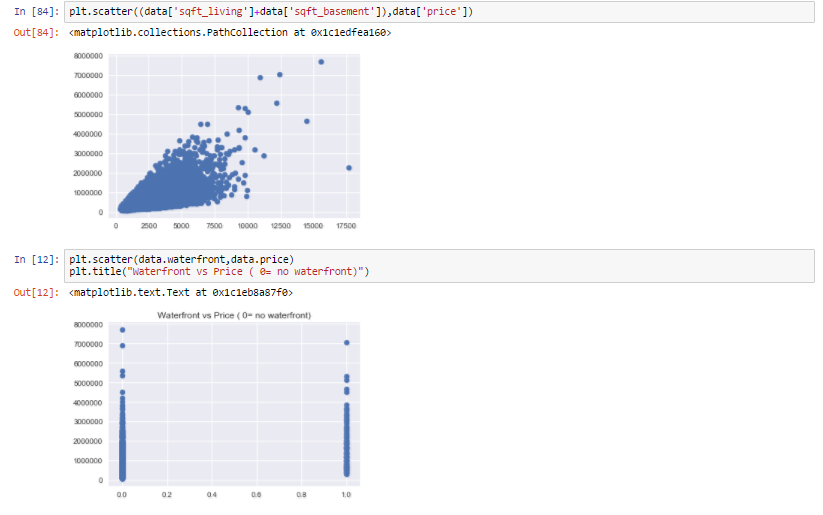
Let us start with , If price is getting affecting by living area of the house or not ?

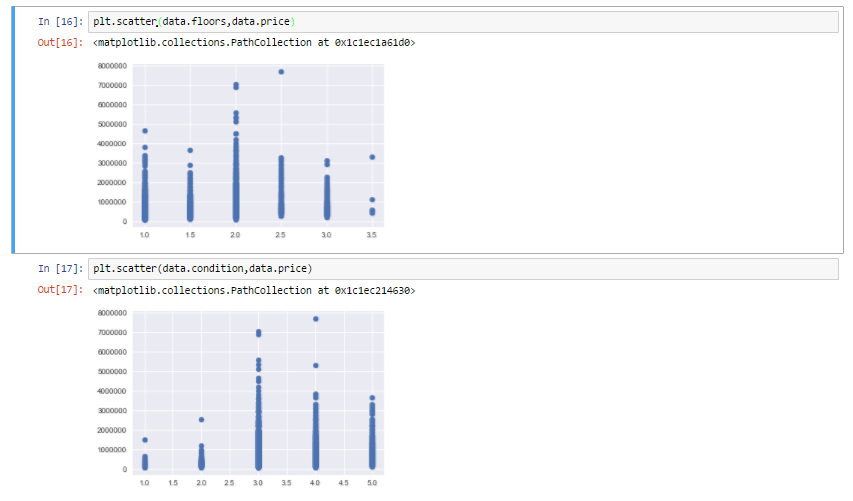
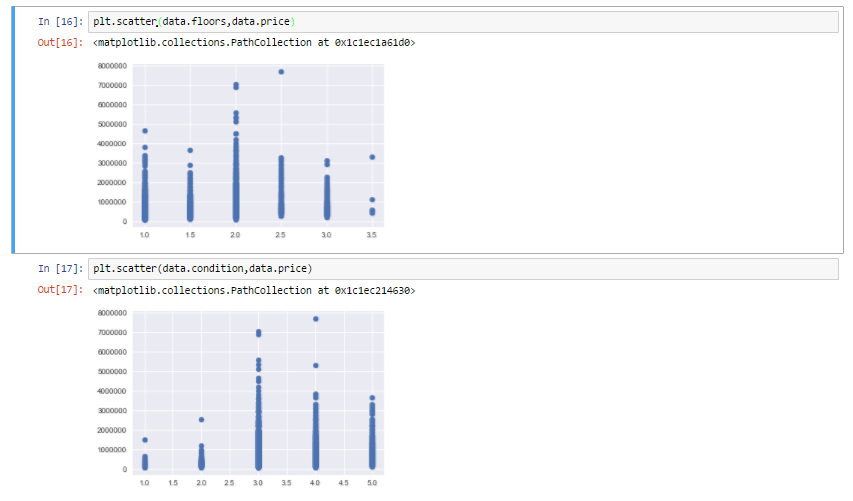


The plot that we used above is called scatter plot , scatter plot helps us to see how our data points are scattered and are usually used for two variables. From the first figure we can see that more the living area , more the price though data is concentrated towards a particular price zone , but from the figure we can see that the data points seem to be in linear direction. Thanks to scatter plot we can also see some irregularities that the house with the highest square feet was sold for very less , maybe there is another factor or probably the data must be wrong. The second figure tells us about the location of the houses in terms of longitude and it gives us quite an interesting observation that -122.2 to -122.4 sells houses at much higher amount.



We can see more factors affecting the price

Total sqft including basement vs price and waterfront vs price



Floors vs Price and condition vs Price

As we can see from all the above representation that many factors are affecting the prices of the house , like square feet which increases the price of the house and even location influencing the prices of the house.

* Linear regression on the data to predict prices

We use train data and test data , train data to train our machine and test data to see if it has learnt the data well or not.

1. We import our dependencies , for linear regression we use sklearn (built in python library) and import linear regression from it.
2. We then initialize Linear Regression to a variable reg.
3. Now we know that prices are to be predicted , hence we set labels (output) as price columns and we also convert dates to 1’s and 0’s so that it doesn’t influence our data much . We use 0 for houses which are new that is built after 2014.
4. We again import another dependency to split our data into train and test.
5. I’ve made my train data as 90% and 10% of the data to be my test data , and randomized the splitting of data by using random\_state.
6. So now , we have train data , test data and labels for both let us fit our train and test data into linear regression model.
7. After fitting our data to the model we can check the score of our data ie , prediction. in this case the prediction is **73%**

The accuracy of the model is lower than our aim of 85.So we use another model.That is gradient boosting regression.

* Gradient boosting regression on the data to predict prices

1. We first import the library from sklearn ( trust me , it is the best library for all statistical related models)
2. We create a variable where we define our gradient boosting regressor and set parameters to it , here

n\_estimator — The number of boosting stages to perform. We should not set it too high which would overfit our model.

max\_depth — The depth of the tree node.

learning\_rate — Rate of learning the data.

loss — loss function to be optimized. ‘ls’ refers to least squares regression

minimum sample split — Number of sample to be split for learning the data

3. We then fit our training data into the gradient boosting model and check for accuracy

4. We got an accuracy of**91.94%** .

Prepared by

D.Vijayabhanu(S180308)

K.Pravallika(S180514)