INTRODUCTION:

Today we are going to work on a dataset which consists information about the location of the house, price and other aspects such as square feet etc.

When we work on these sort of data, we need to see which column is important for us and which is not.

Our main aim today is to make a model which can give us a good prediction on the price of the house based on other variables.

We are going to use Linear Regression for this dataset and see if it gives us a good accuracy or not.

STEPS THAT WE FOLLOWED IN THIS:

STEP 1:

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.

How the describe function is useful:

Look at the bedroom columns, the dataset has a house where the house has 33 bed rooms , seems to be a massive house and would be interesting to know more about it as we progress.

Maximum square feet is 13,450 where the minimum is 290. we can see that the data is distributed.

So we can get much information by using describe function

STEP 2:

After the above next we are going to see some visualization and also going to see how and what can we get from visualization.

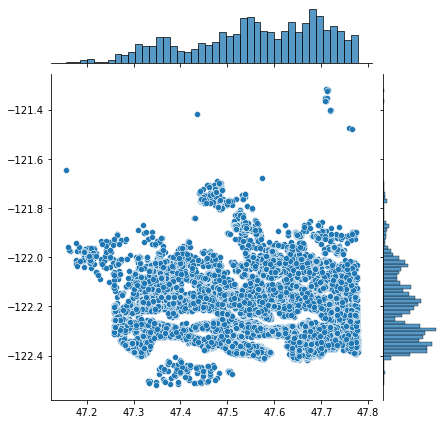
Let’s see which is most common bed room number . You may wonder why is it Important ?

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.

So now we know that 3 and 4 bedroom’s are highest selling. But at which locality ?

Next one more thing that effecting

**Visualizing the location of the houses based on latitude and longitude.**



We saw the common locations and now we’re going to see few common factors affecting the prices of the house ?

If price is getting affecting by living area of the house or not ?

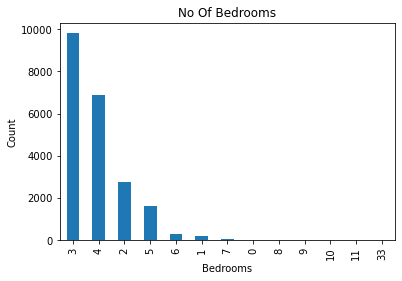
The plot that we used there is called scatter plot , scatter plot helps us to see how our data points are scattered and are usually used for two variables. From the first figure we can see that more the living area , more the price though data is concentrated towards a particular price zone .

We can see more factors affecting the price

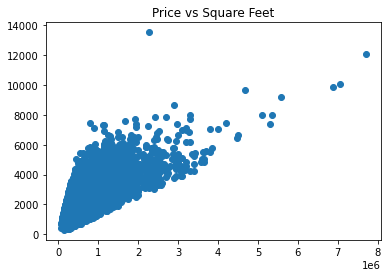
As we can see from all the there representation that many factors are affecting the prices of the house , like square feet which increases the price of the house and even location influencing the prices of the house.

Now that we are familiar with all these representation and can tell our own story let us move and create a model to which would predict the price of the house based upon the other factors such as square feet , water front etc . We are going to see what is linear regression and how do we do it ?

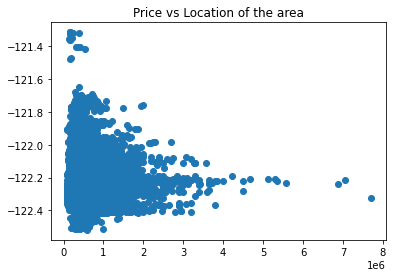
Bedroom Counts



Price vs Square Feet

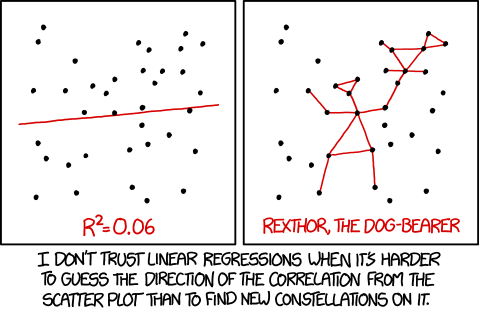


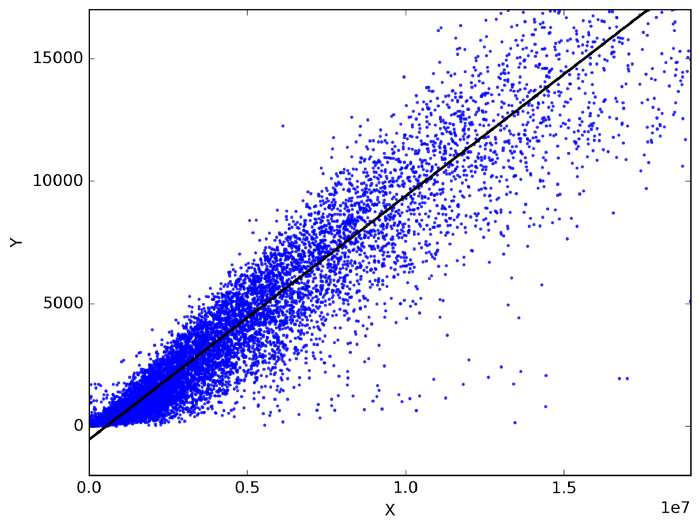
Price vs Location of area



Linear Regression :

In easy words a model in statistics which helps us predicts the future based upon past relationship of variables. So when you see your scatter plot being having data points placed linearly you know regression can help you



regression works on the line equation , y = mx +c , trend line is set through the data points to predict the outcome.

The variable we are predicting is called the criterion variable and is referred to as Y.

The variable we are basing our predictions on is called the predictor variable and is referred to as X. When there is only one predictor variable, the prediction method is called **Simple Regression. and if multiple predictor variable are present then multiple regression.**

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.

Before anything , I want everyone to *remember that the machine is the student and train data is the syllabus and test data is the exam. we see how much the machine has scored and if it scores well are model is successful****.***

**What we gonna do we will go step by step:**

We import our dependencies , for linear regression we use sklearn (built in python library) and import linear regression from it.

We then initialize Linear Regression to a variable reg.

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.

We again import another dependency to split our data into train and test.

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.

So now , we have train data , test data and labels for both let us fit our train and test data into linear regression model.

After fitting our data to the model we can check the score of our data ie , prediction. in this case the prediction is **73%**

**Conclusion:**

**We are concluding that our prediction that testing data score gives 73% accuracy,we can say that anything greater than 70%** **is a great model performance**.

In this we use multilinear Regression to predict the the Values

**UPDATION IN PROJECT**

For building a prediction model , many experts use **gradient boosting regression ,**so what is gradient boosting ? It is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

By using gradient boosting regression we can archive 92 % in predicting score

TEAM:

D.CHARAN

V.S.S.SAI

TEAM BLACK SQUAD