**Data Mining Project Report**

**House Price Prediction**

**CSCS 451**

**Miss Ayesha Khan**



**Presented By:**

Muhammad Muzammil 231485432

Danish Jamal 231466533

Muhammad Mustafa 231469864

**Abstract:**

People looking to buy a new home tend to be more conservative with their budgets and market strategies. The existing system involves calculation of house prices without the necessary prediction about future market trends and price increase. The goal of the paper is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. By analyzing previous market trends and price ranges, and upcoming developments, future prices will be predicted. This system will help the people in house prices based on many different factors. It will follow trends of the past.

**Introduction:**

This paper brings together the latest research on prediction markets to further their utilization by economic forecasters. Thus, there is a need to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. This paper efficiently analyses previous market trends and price ranges, to predict future prices. This topic brings together the latest research on prediction markets to further their utilization by economic forecasters. It provides a description of prediction markets, and also the current markets which are useful in understanding the market which helps in making useful predictions. Thus, there is a need to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. This paper uses linear regression algorithm to predict prices by analyzing current house prices, thereby forecasting the future prices according to the user’s requirements. The data for this system is taken from Kaggle.

**Literature Review:**

With a large amount of unstructured resources and documents, the Real estate industry has become a highly competitive business. The data mining process in such an industry provides an advantage to the developers by processing those data, forecasting future trends and thus assisting them to make favorable knowledge-driven decisions. In this paper, the focus is on data mining method and its approach to develop a model, which not only predicts the most suitable area for a customer according to his\her interests, and it also recognizes the most preferred location of real estate by its latitude and longitude and tells which location offers the best price.

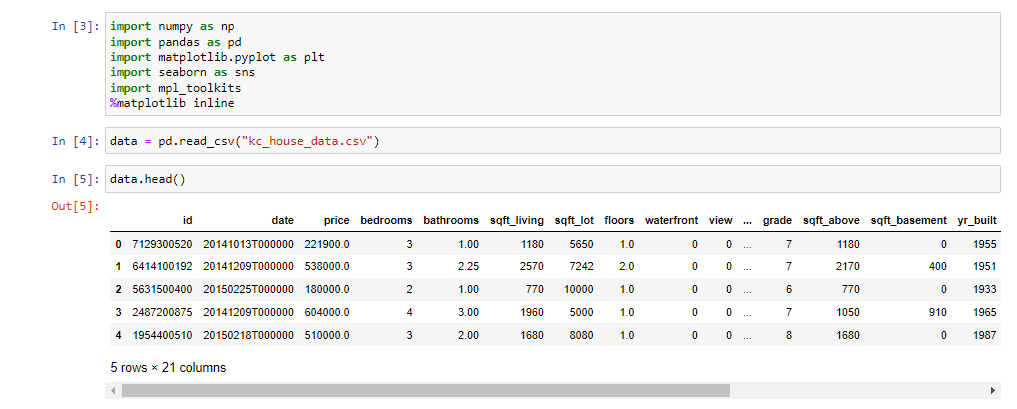
This prediction system takes into account the bedrooms in the house, floors in the house, size of the house and does it have a waterfront or not. Every attribute is compared with the price which helps us to know what are the customers looking for when buying a house and at what price. This allows the system to predict what will be the price of a certain house. . It uses a classical technique called linear regression and tries to give an analysis of the results obtained. It helps establishes the relationship strength between dependent variable and other changing independent variable known as label attribute and regular attribute respectively. Regression displays continuous value of the dependent variable i.e. label attribute that is used for prediction.

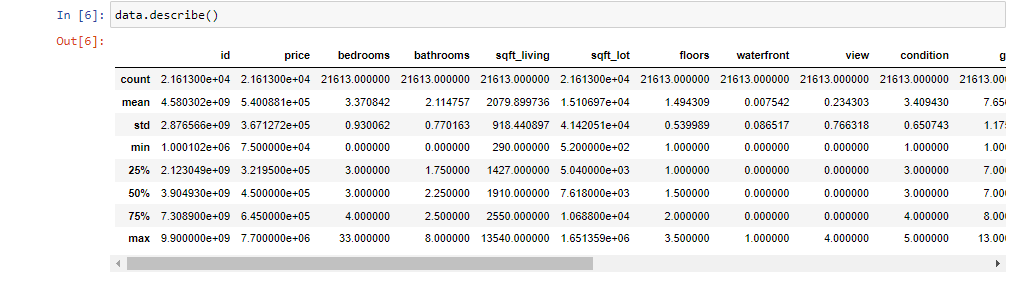
The major problem with this system is it does not predict future prices of the houses mentioned by the customer. Due to this, the risk in investment in an apartment or an area increases considerably. To minimize this error, customers tend to hire an agent, which again increases the cost of the process. This leads to the modification and development of the existing system.

**Proposed System:**

Nowadays, e-education and e-learning is highly influenced. Everything is shifting from manual to automated systems. The objective of this project is to predict the house prices so as to minimize the problems faced by the customer. The present method is that the customer approaches a real estate agent to manage his/her investments and suggest suitable estates for his investments. But this method is risky as the agent might predict wrong estates and thus leading to loss of the customer’s investments. The manual method which is currently used in the market is out dated and has high risk. So as to overcome this fault, there is a need for an updated and automated system. Data mining algorithms can be used to help investors to invest in an appropriate estate according to their mentioned requirements. Also the new system will be cost and time efficient. This will have simple operations. The proposed system works on Linear Regression Algorithm.

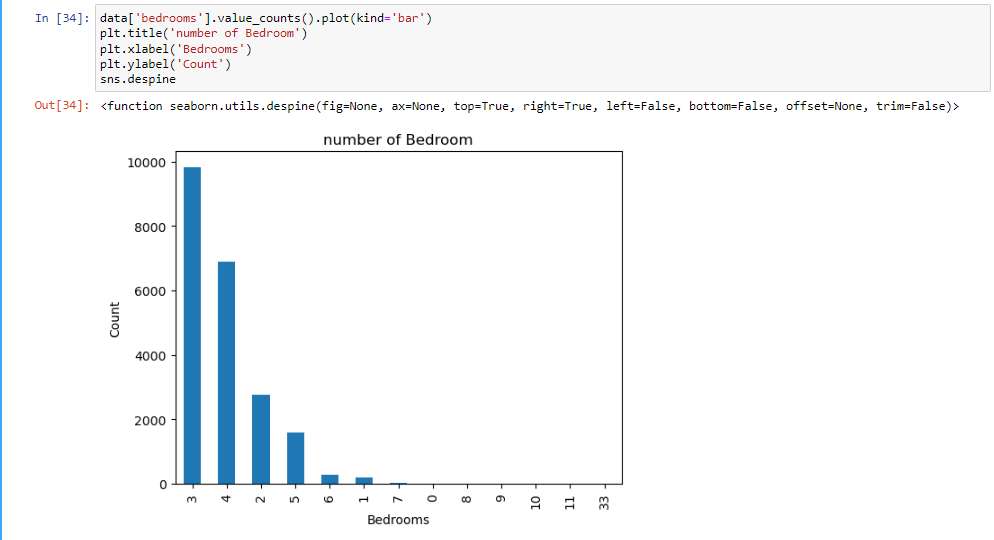
**Attributes:**

****

****

The first figure tells us the libraries used in the algorithm and the data being imported while second figure tells us the quartiles of the data and its minimum and maximum value of every column.

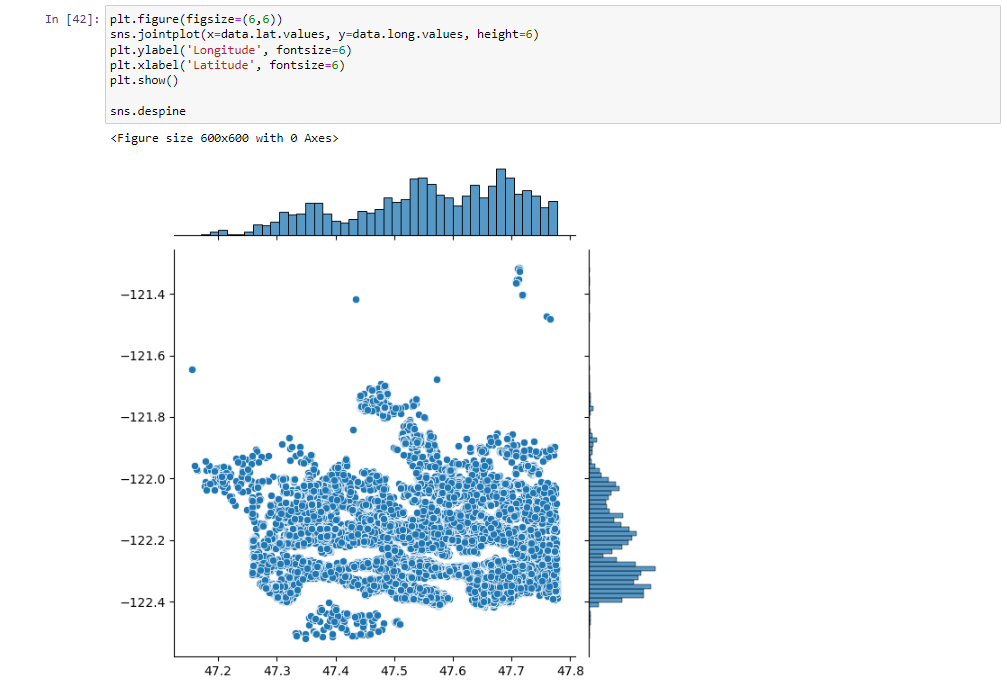
**Which is the most common house (Bedroom wise) ?**



Knowing about what buyers are looking for in a house is important. From a builder’s perspective, one should know how many bedrooms should be in a house that would sell quickly. 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/she can make a new building with more 3 and 4 bedroom’s to attract more buyers. But, with the right house place is also very important.

**Visualization of houses based on longitude and latitude:**

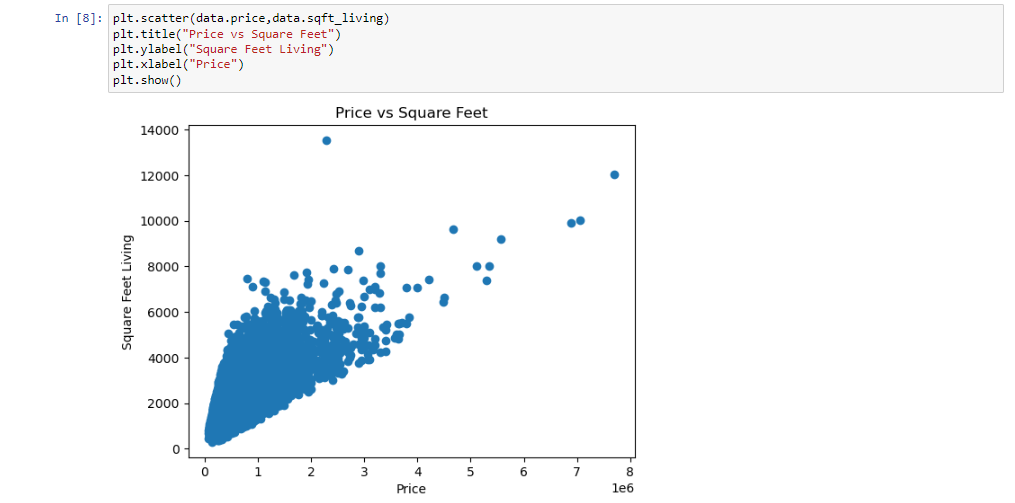
In the data, we are given longitude and latitude of the houses. Now we will see the common are where most of the houses are.



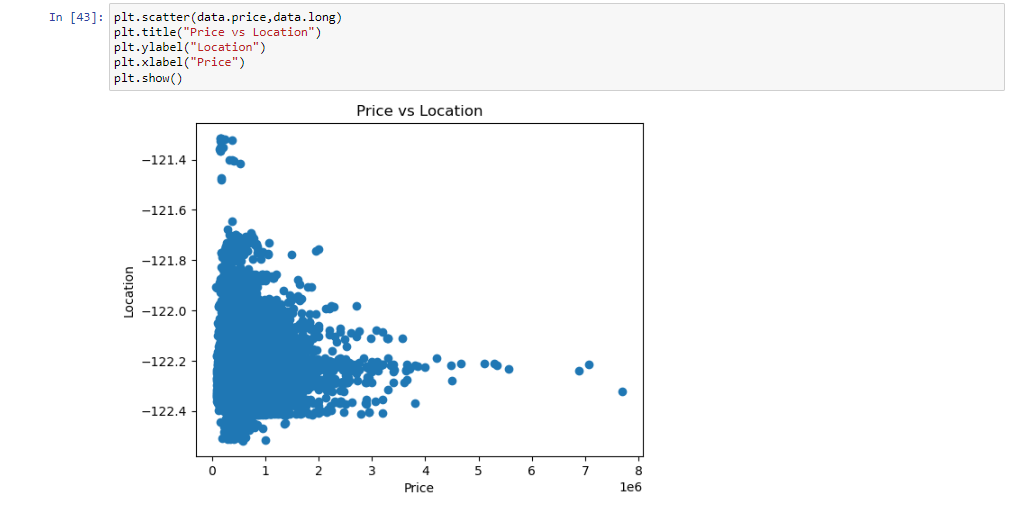
We use seaborn , and we get his visualization. Joinplot function helps us see the concentration of data and placement of data and can be useful. Let us see what we can infer from this visualization. The x-axis represent latitude and y-axis represent longitude. For latitude between 47.7 and 48.8, there are many houses, which would mean that maybe it is an ideal location to buy a house in. 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 buys has been for this particular location.

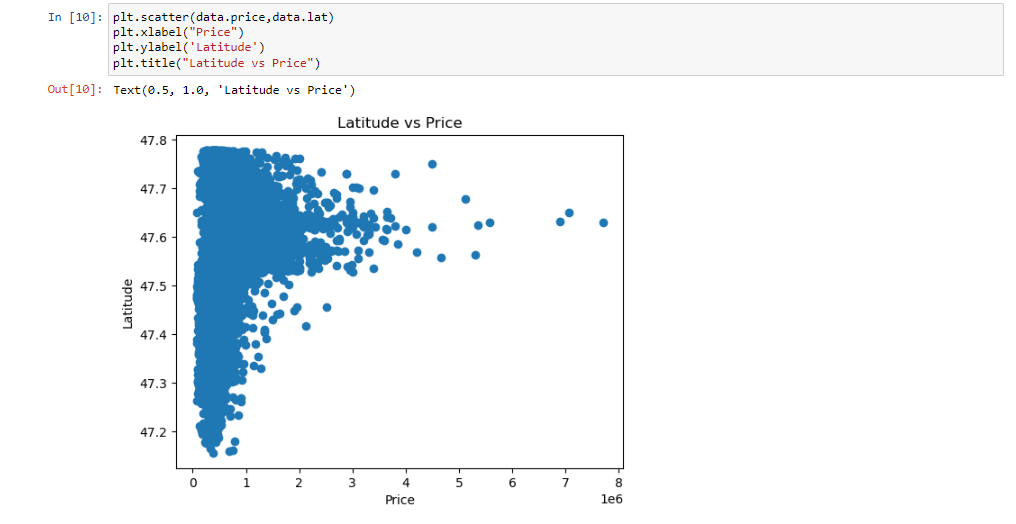
**How common factors are affecting the price of houses?**

At start we saw the common houses and their location and now we will see some common factors and how do they affect the price of the house.

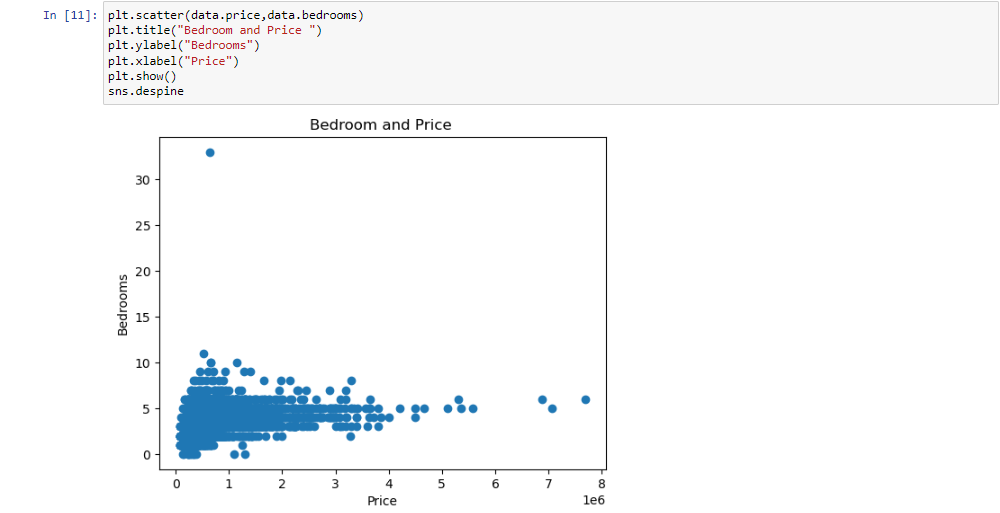


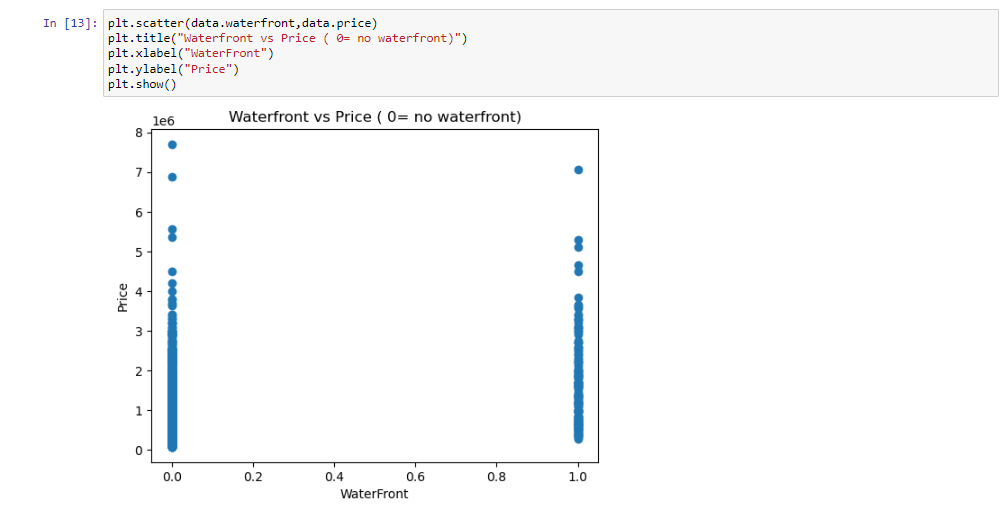
The plot that we used above is called scatter plot, Scatter plot helps us to see how our data points are scattered and are usually used for two variables. Here, the living area is in square feet on the x-axis while price is on x-axis. 1 unit of price is represented as 100,000. From the first figure we 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.



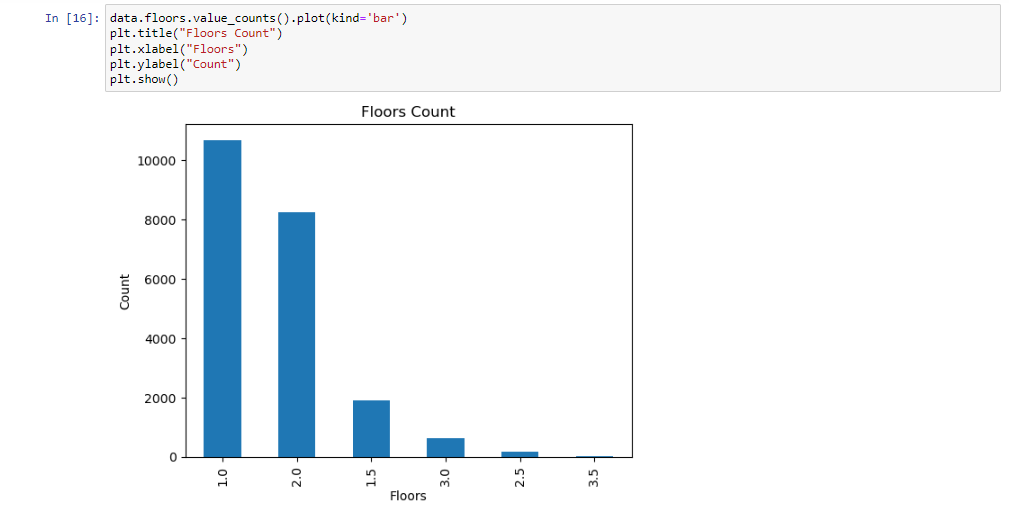


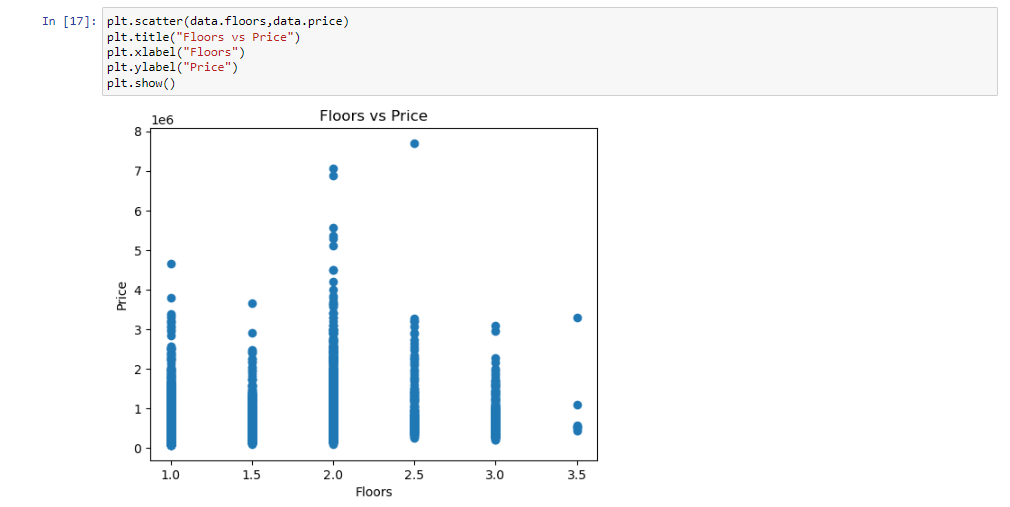
This is another attribute where price is compared with location. The first figure tells us about the location of the houses in terms of longitude and it gives us quite an interesting observation that -122.2 to -122.4 sells houses at much higher amount while the houses in the rest of the location is max priced at 400,000. The second figure tells us the prices based on the latitude and we can see that houses with latitude of 47.6 to 47.8 are in more demand and are sold at a high price also as compared to the other locations.





The first figure is a comparison between the number of bedrooms in a house and its price. It shows us that most houses have 3 or 4 bedrooms and their prices are at 200,000. This gives us the idea of what buyer is looking at. The second figure tells us about the prices houses which has waterfront and which do not have waterfront. This attribute is not a good attribute to judge prices of house as the graph shows the prices are almost the same although there are more houses with no waterfront.

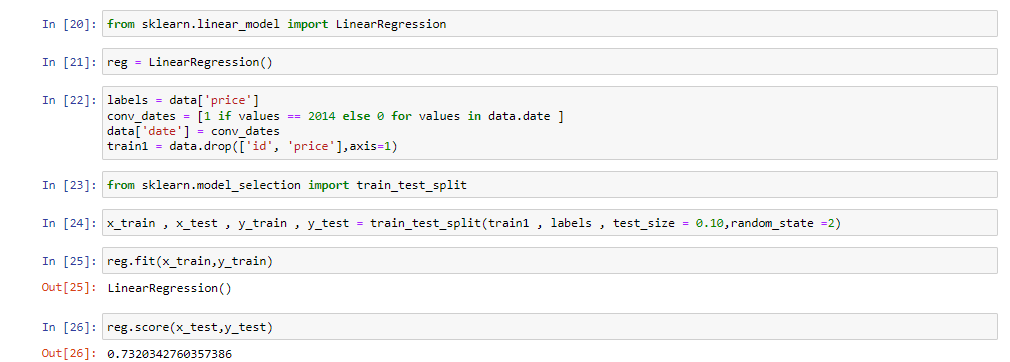




This attribute tells us how the houses are being made. How many floors are being preferred by the buyers which makes the job for a builder easy. The first figure tells us that majority of the houses have only one floor followed by houses with 2 floors. After that there are very less houses with 3 or more floors. The second figure tells us the varying of price according to the floors. Houses with two floors tend to have more price then houses with 1 floor.

**Linear regression:**

In easy words, a model in statistics, which helps us, predicts the future based upon past relationship of variables. 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.**



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 have 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%**

**Conclusion:**

In today’s real estate world, it has become tough to store such huge data and extract them for one’s own requirement. Also, the extracted data should be useful. The system makes optimal use of the Linear Regression Algorithm. The system makes use of such data in the most efficient way. The linear regression algorithm helps to fulfill customers by increasing the accuracy of estate choice and reducing the risk of investing in an estate. A lot’s of features that could be added to make the system more widely acceptable. One of the major future scopes is adding estate database of more cities which will provide the user to explore more estates and reach an accurate decision. More factors like recession that affect the house prices shall be added. In-depth details of every property will be added to provide ample details of a desired estate. This will help the system to run on a larger level.

**References:**

Dataset was extracted from here. [www.kaggle.com](http://www.kaggle.com)