SOULPAGE

IT SOLUTIONS

House Price Sales Prediction

House Price Prediction Analysis

Problem:-

Saratoga County in NewYork is one of the lowest county tax rates in the state. Saratoga County is an excellent place to live, work and visit. The cultural venues, excellent schools and colleges along with trails, parks and recreation programs make Saratoga County a great place to call home.

We are using various features like lot size, water front, age, land value, construction, air condition etc of Saratoga County to predict the price of the house. This model can be very useful to the Real estate agents, Home sellers, Home buyers to decide the Prices of the houses while buying and selling.

Opportunity:-

We can use all the available data to perform some analysis on the features we have and also we can predict the price of the house with respect to different features availability and their values. We can perform analysis like

- How many houses has waterfront?
- How many have air-condition facility?
- What type of fuel is used to produce heat in-house?
- What is sewage system available in house?
- How many fireplaces we have in each house?
- How many bathrooms and living rooms we have?

Tools and techniques used:-

We have several tools and methods available to perform exploratory data analysis on all the features we have. We have also used regression models in machine learning to see which model has performed well in our dataset.

- Python 3.6
- Numpy
- Pandas
- Matplotlib

And machine learning models like

- Linear regression

- Ridge regression
- Lasso regression
- Random forest

Dataset:-

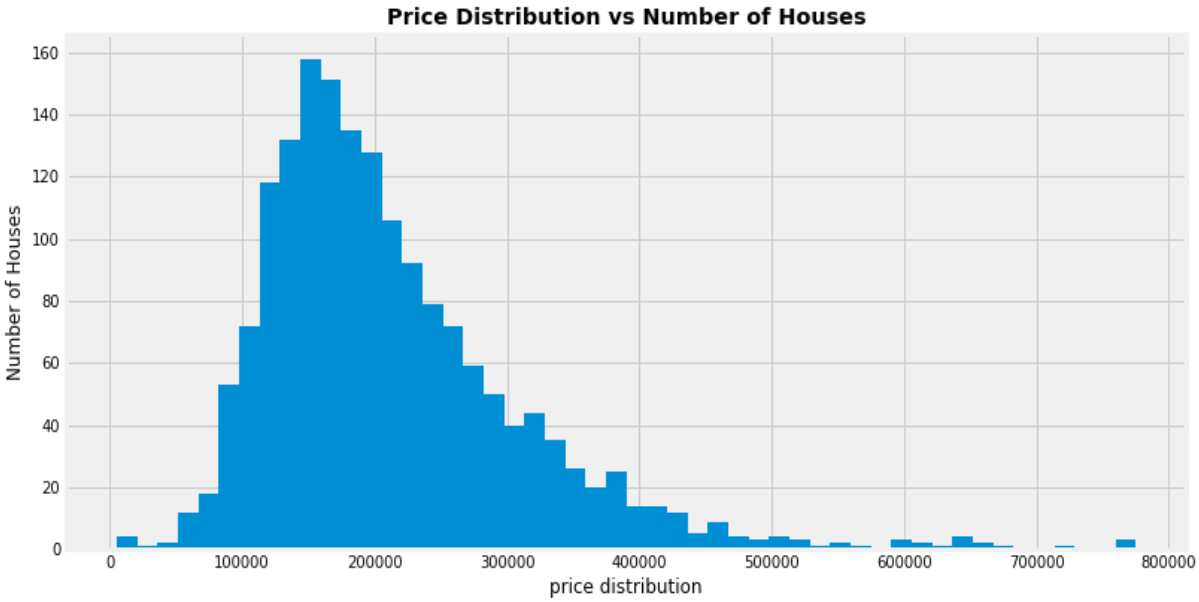
The dataset consists of 1700 records of houses with 14 features (201KB)

Process:-

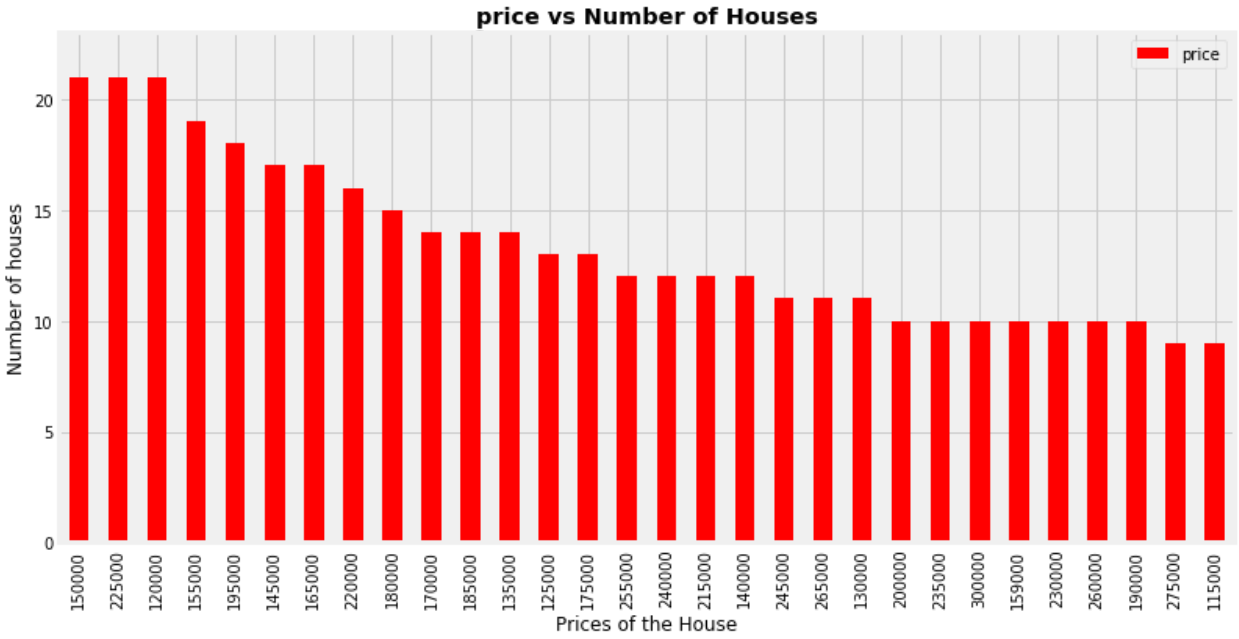
We have performed some analysis which provides an answer to the above mentioned questions with the help of visualisation and Machine learning models.

Price in Dollars:-

We have plotted the price feature that lies in the range of 5000$ to 775000$ and can infer that the data follows a normal distribution. (Refer to the below graph)

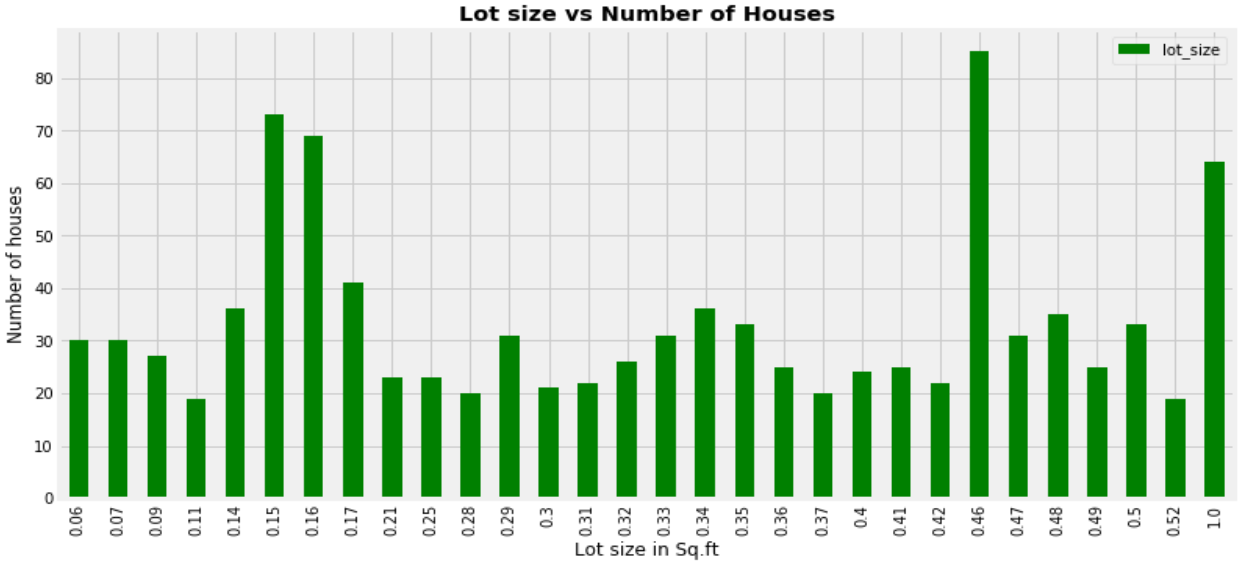


Further, (from the below graph) we can observe that around 20 houses are in the range of 120000$ to 150000$.



Lot size:-

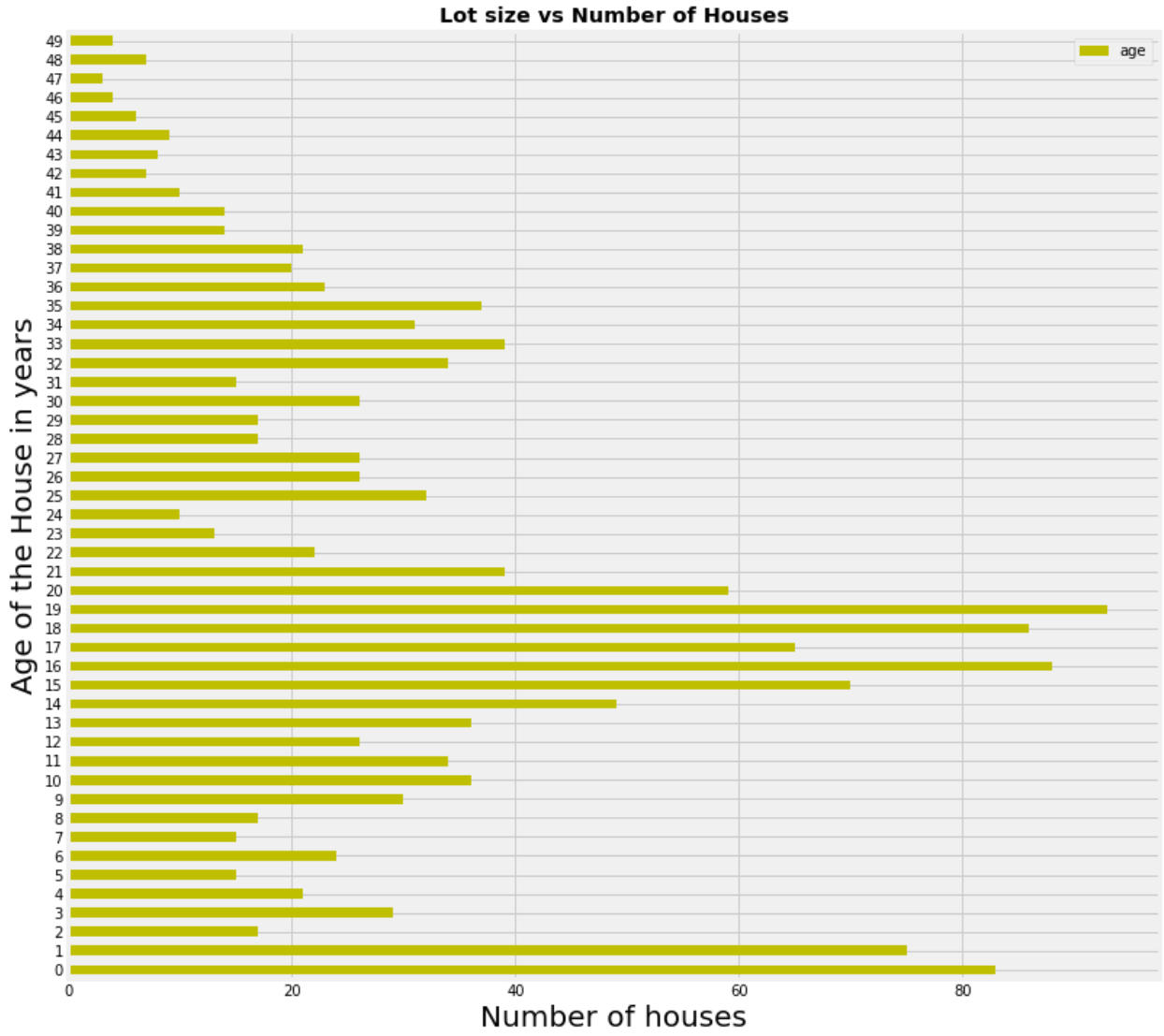
The feature - lot size, tells about the total land size.



We have around 80 houses with lot size 0.46Sq.ft and 70 houses with lot size of 0.15 Sq.ft.

Age of House (in years):-

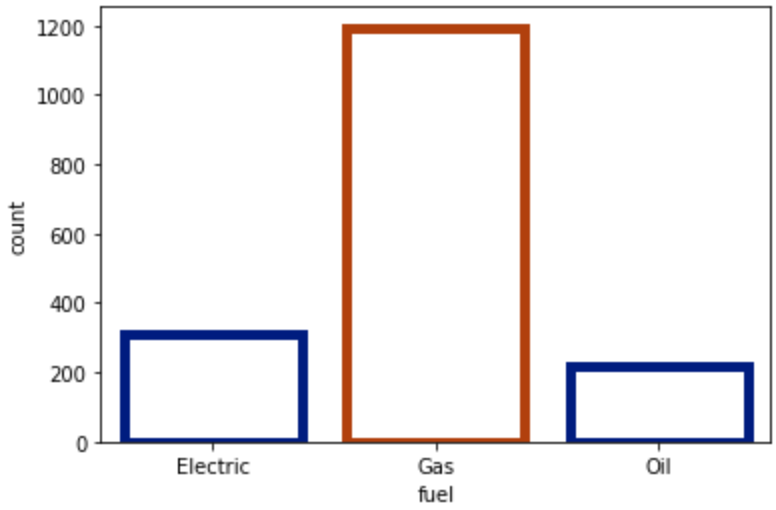
Houses are aged from 0 to 255 years.



From the graph we can see that we have 93 houses are 19 years old.and 88 houses are 18 years old.

Fuel:-

Gas, Electric and Oil are the 3 types of fuel which produces heat in the house.

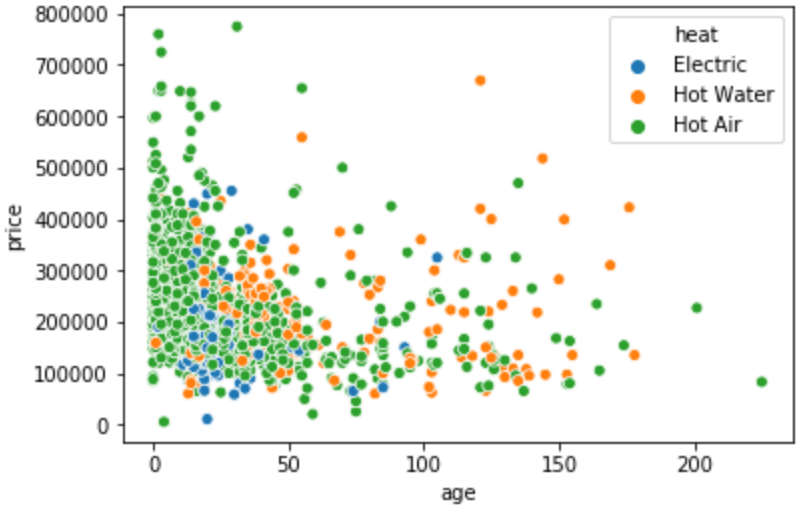


Around 1200 houses uses Gas as fuel, 300 houses use electric facility where 200 houses use oil as fuel to produce heat.

Age of the House vs Price of House:-

From the graph below we can see that a drop in sales price with increase in Age of building. Also,the recent constructed houses have heating system of hot air and old houses have heating system of Hot water.

Age of House in Years vs Price of House with Heat system



Price vs Living area with Heat System, Rooms, Ac

We can understand from the graph that houses with air conditioning in all rooms and with more bedrooms are priced higher.

Model building:-

We have 14 features and in them we have 6 features with categorical values in it. Hence, one-hot encoding and converting them to numerical data.

Data is split into train and test split with 70% in train and 30% in test set.

We have applied the regression algorithms like Linear Regression, Ridge Regression, Lasso Regression, and Random Forest to predict the price of the house on the features.

Their accuracies are captured below.

Linear regression - 65.88%
Random forests - 65.09%

Ridge regression - 65.82%
Lasso regression - 65.88%

So hence fit the Linear regression model that gave the best accuracy.

Conclusion:

We have got the best accuracy of 65% for the data. We can say that the prices of the house can be predicted with 65% accuracy which isn't sufficient . We need to consider more featured which affects the prices to get better accuracy.