Springboard Capstone Project 1

Predicting the price of the house

Abhishek Kumar

02 May 2020

Contents

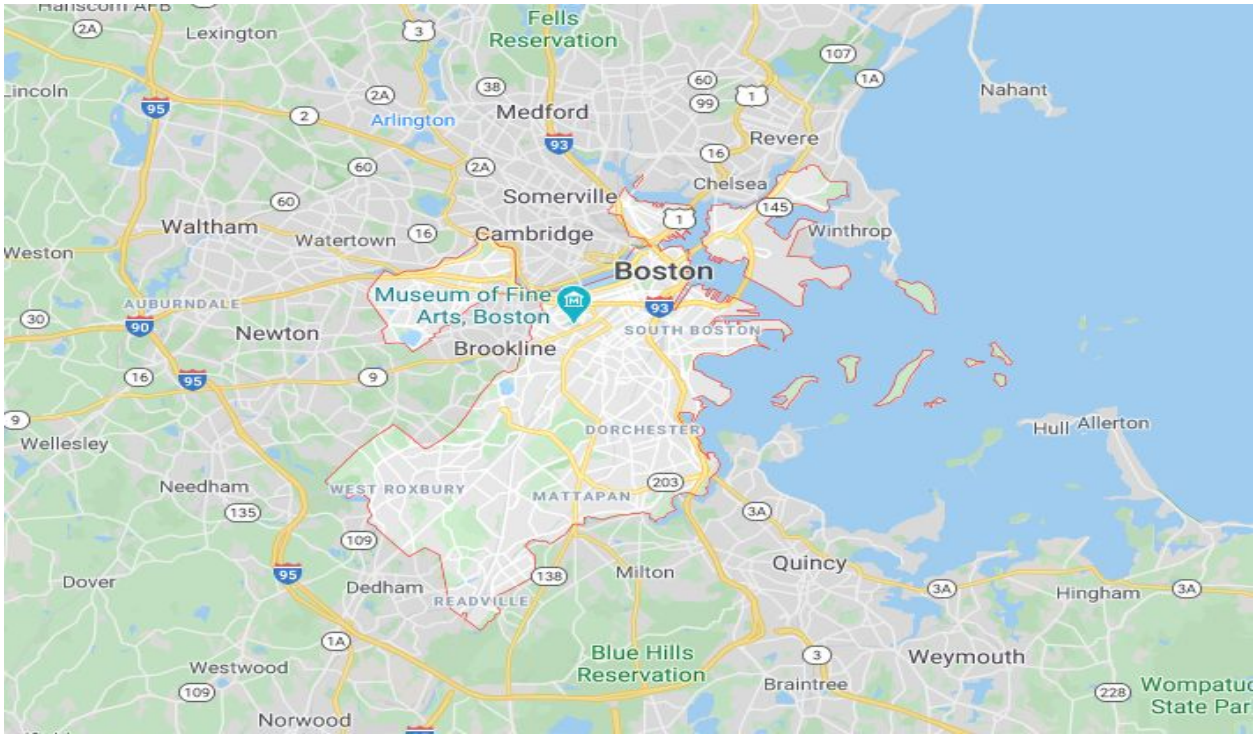
1 Introduction

This project is about predicting the price of a house. As it is well known that the price of houses are in the *thousands of dollars* (basically lots of money is at stake). So the decision of owning a house of an average family or a person is a *big one* , for most it involves saving of lifetime or to have a huge burden of loan. If real estate companies dealing in houses, apartments,flats etc are able to offer a house at a reasonable rate, it would *immensely* help their clients.Most of the customers take a good amount of time (sometimes upto a year or more) in deciding which property to buy, where to buy etc. So here the role of the real estate companies becomes crucial, if they have a good idea of what a property would cost over a period of time they can pass that info to their clients. Though the price of a property is determined by many factors such as location, connectivity, amenities, size etc. Here I am using data from UCI machine learning repository (Boston housing data) and build a supervised learning model for predicting the price of the house.

Companies dealing in buying and selling of property such as housing.com, 99 acres.com and more can use this model or similar model in serving their clients better, helping them in decision making and providing attractive offers. This model can increase the customer base and even enhance the reputation of the company. Companies can even build an application which would show how much the cost of property would increase over time, how many people are interested in a particular property.

2 Data Acquisition and Cleaning

The data was collected from *here* , it contains data of Boston.



Courtesy: Google maps

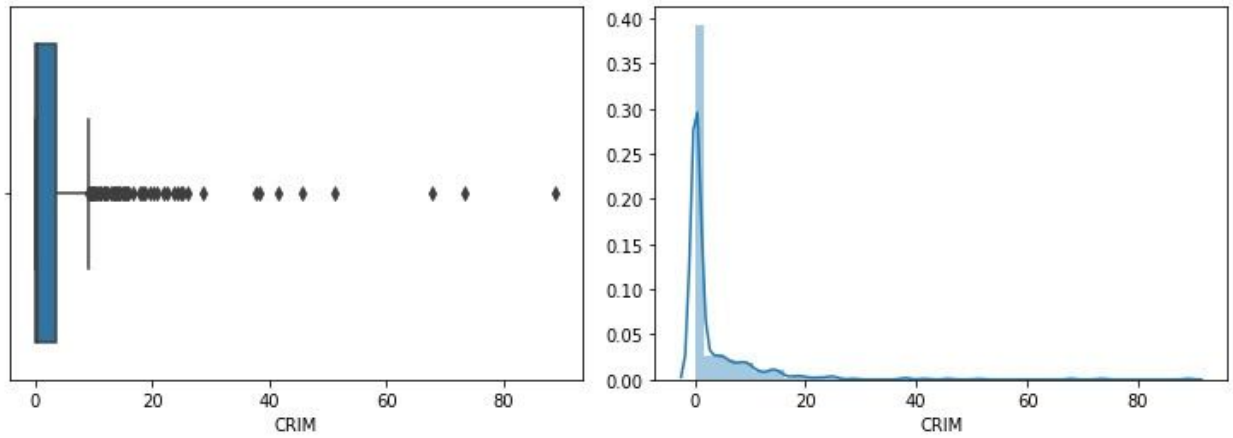
The downloaded data was needed to be converted to csv format, which was done through ms excel. The data was clean and no cleaning was required. The details of rows features are as follows

CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000's

3 Data Exploration

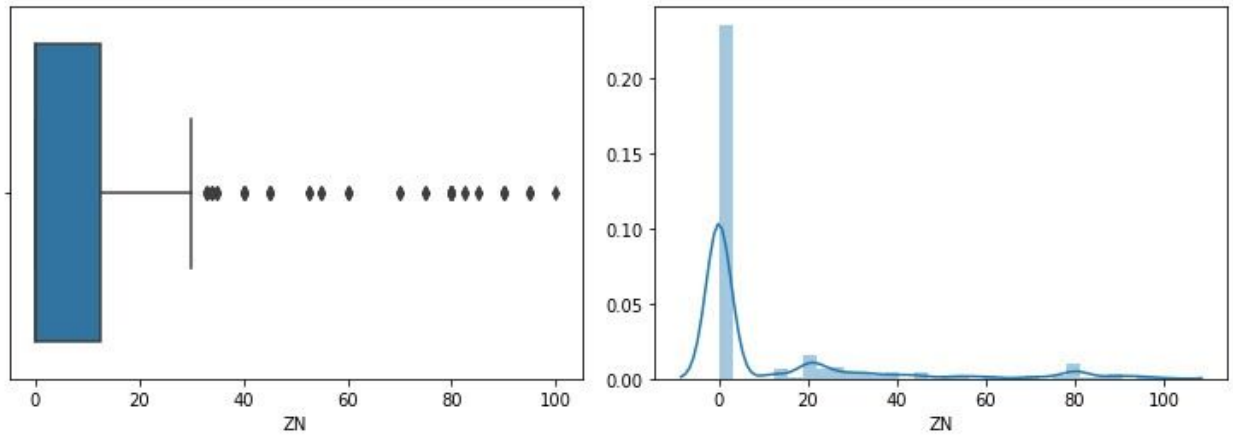
3.1 CRIM per capita crime rate by town

Crime in an area is an important aspect for deciding the price of the house, if the crime is lower the price will be high.



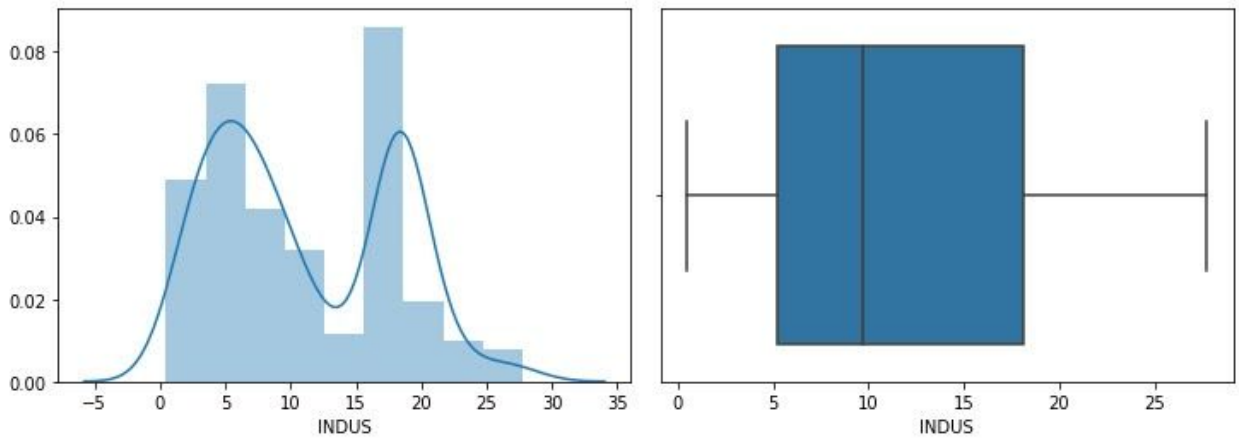
The mean of crime is 3.6 with a standard deviation of 8.6 and here we can see the crime doesn't have too many bars but there are a large number of outliers present.

3.2 ZN proportion of residential land zoned for lots over 25,000 sq.ft.



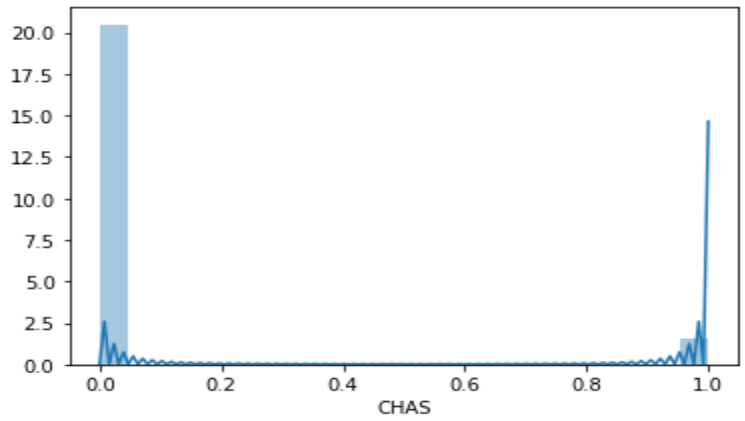
ZN too has lots of outliers and most of the values are close to zero.

3.3 INDUS proportion of non-retail business acres per town



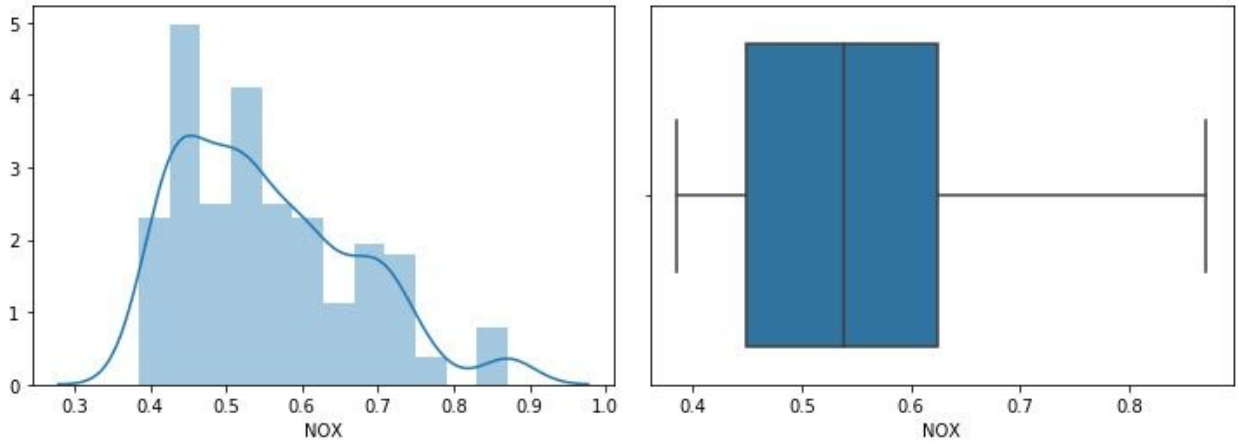
INDUS has no outliers and has two peaks .

3.4 CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)



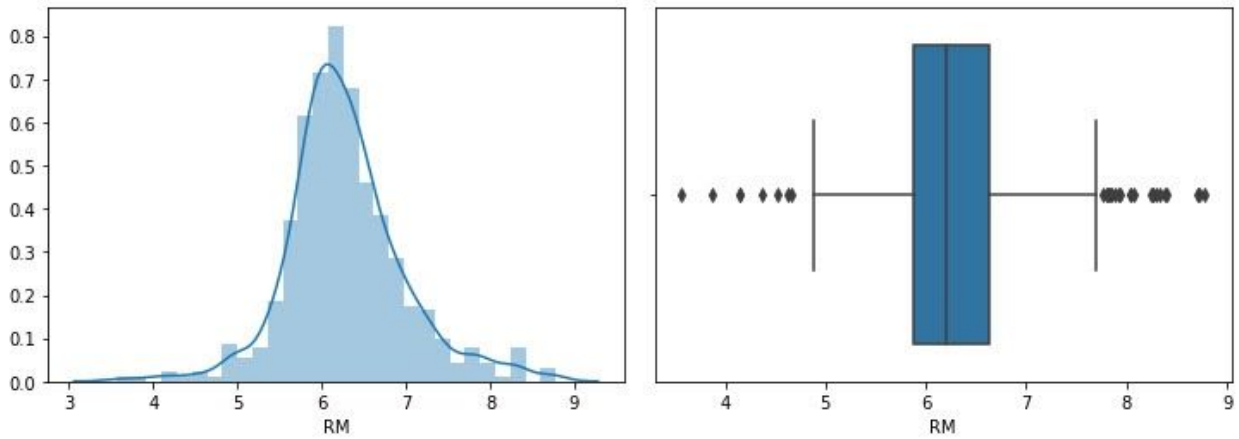
Charles is a river flowing in Boston. From the chat we can see that many people live in houses whose boundaries do not touch the charles river.

3.5 NOX nitric oxides concentration (parts per 10 million)



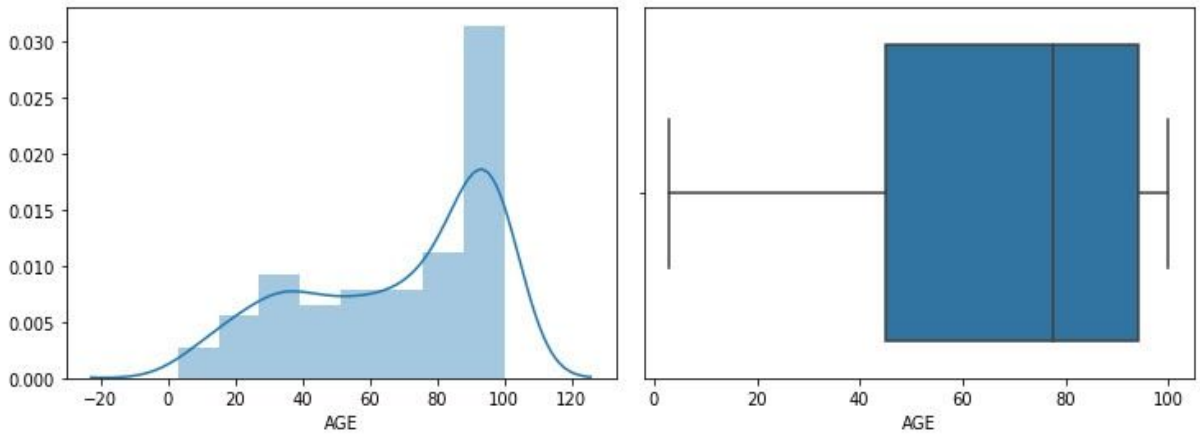
We can see from the plot that nitric oxide concentration has too many bars and the calculated mean is 0.55 The nitric oxide content is given in (parts per 10 million) and after looking for safe nitric oxide on wiki it was 250 (parts per 10 million) and from the data it was found to be safe with max value close to 9 ppm.

3.6 RM average number of rooms per dwelling



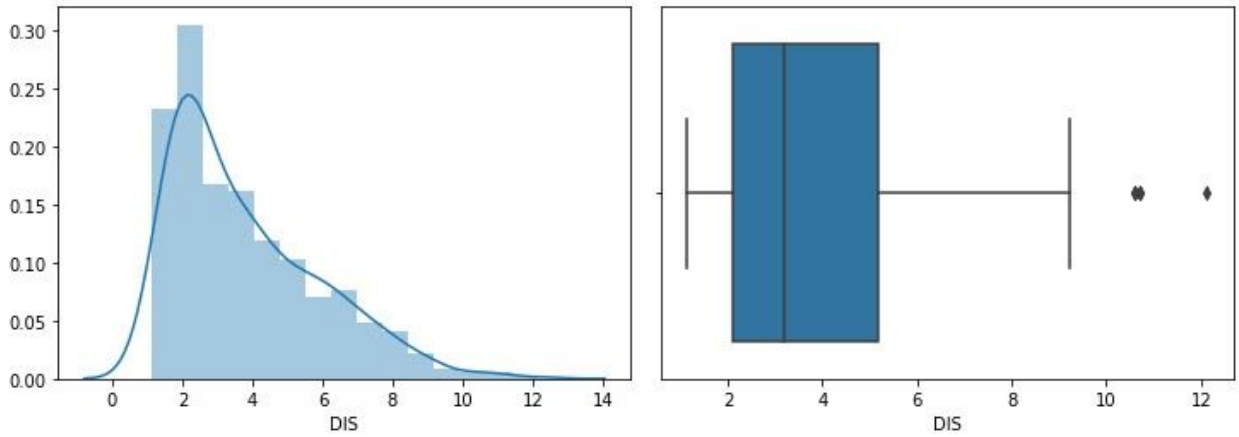
The mean number of rooms is 6.2 and also from the plot we can see frequency is most from 6 to 7 rooms. We can see that there are lots of outliers.

3.7 AGE proportion of owner-occupied units built prior to 1940



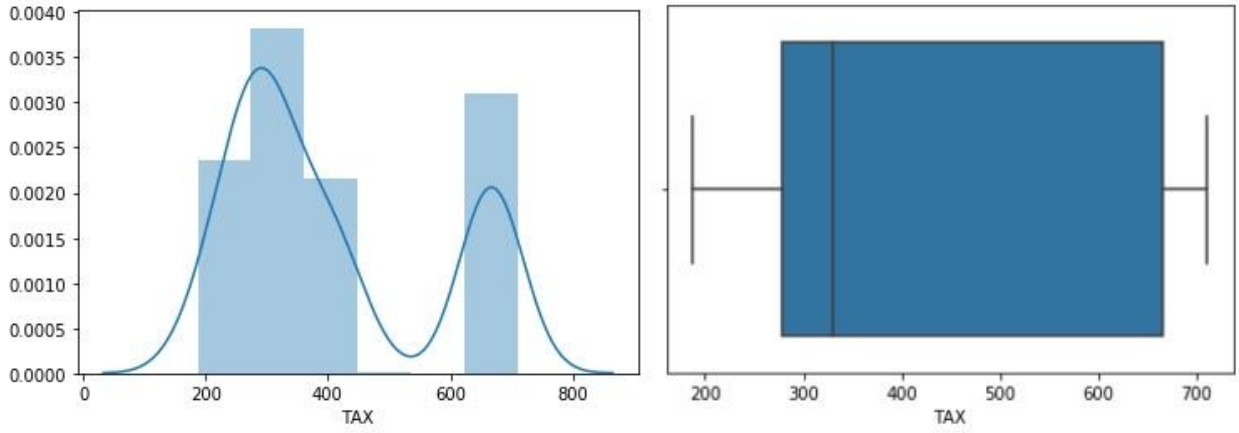
We can see that frequency is more at 100. So people like to live in buildings prior to 1940.

3.8 DIS weighted distances to five Boston employment centres



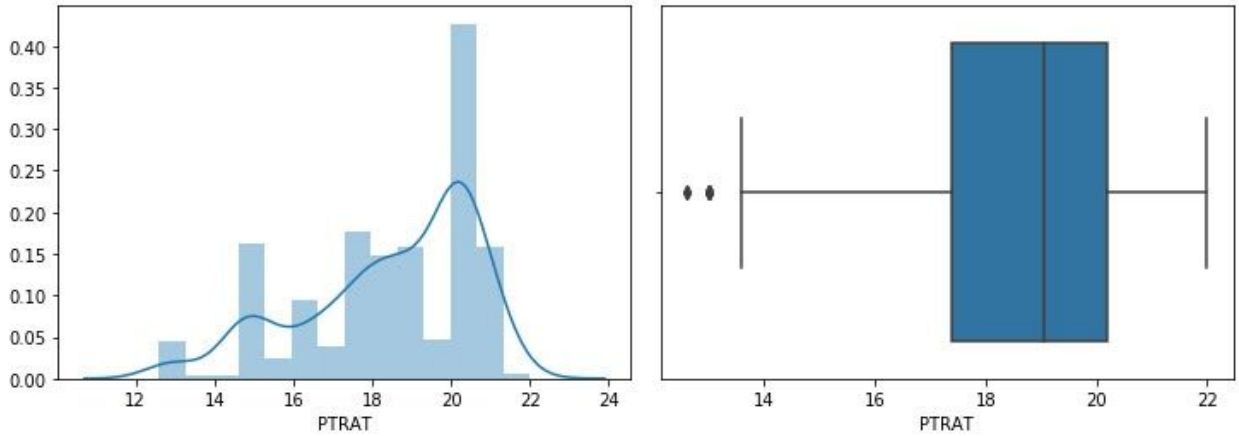
As we can expect people like to live near employment centers , this is also visible from this plot. Three outliers are present.

3.9 TAX full-value property-tax rate per $10,000



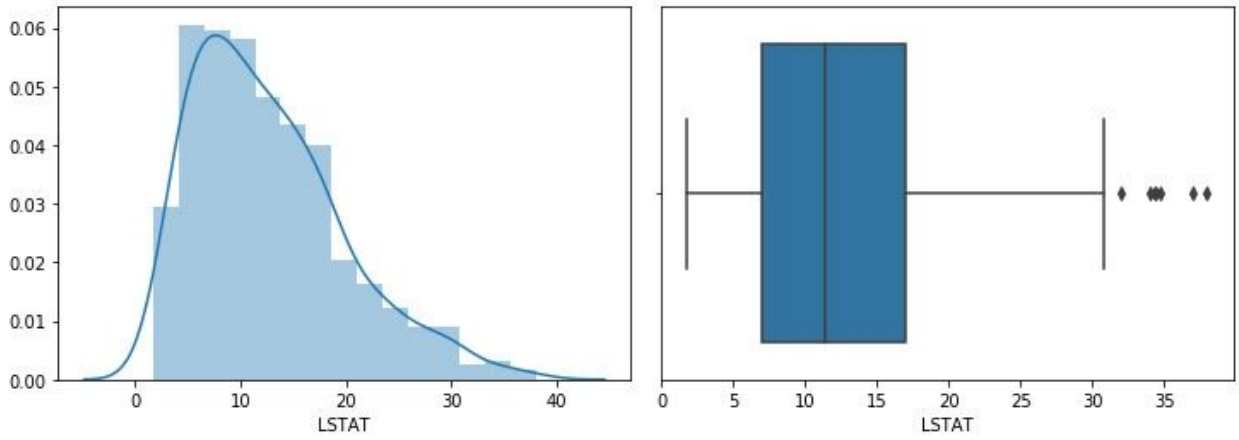
We can see people are willing to pay in the range $2000000 to $ 4000000 and we can see a surge at around $6000000 to $7000000.

3.10 PTRATIO pupil-teacher ratio by town



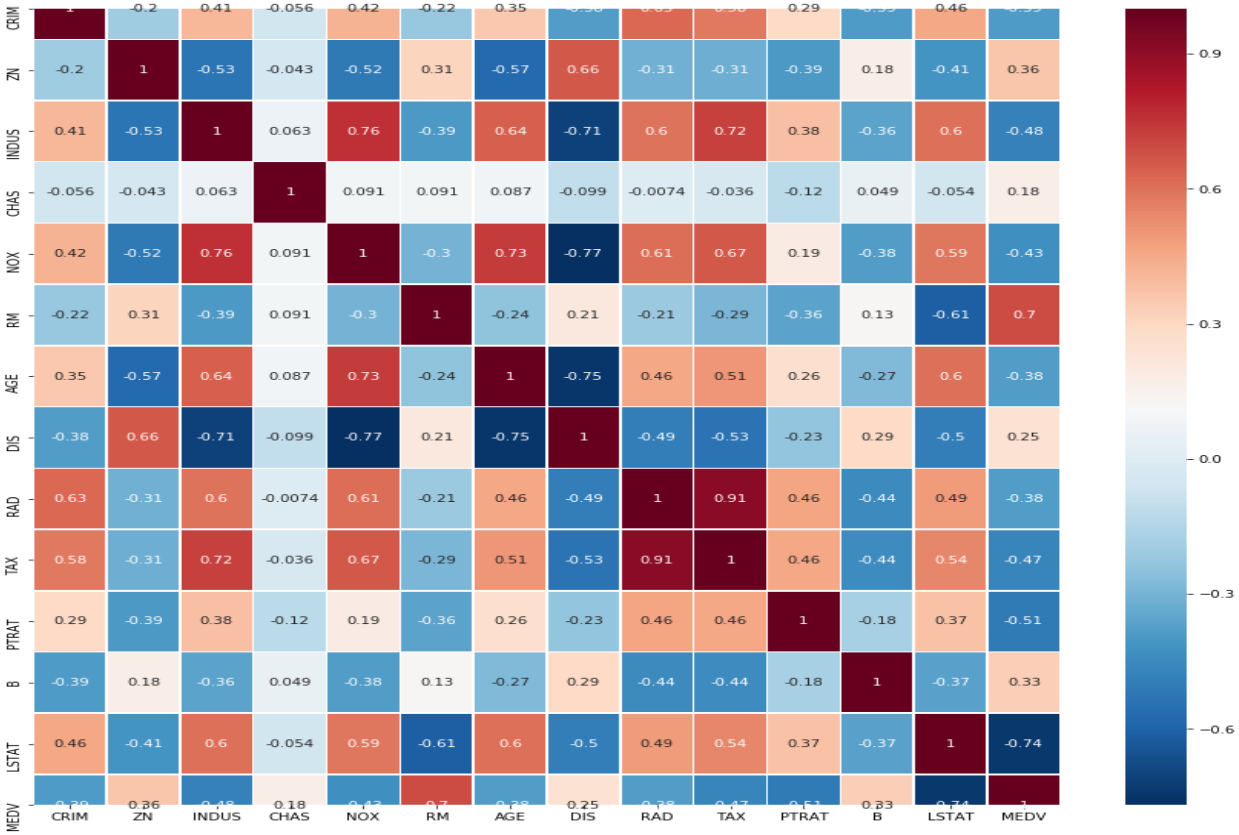
The average is 18.45 for pupil teacher ratio and the plot has many peaks. Two outliers are visible from the plots.

3.11 LSTAT % lower status of the population



We can say that in a place where financially weak people leave there the price of the property would be low. The mean of LSTAT is 12.6 with standard deviation of 7.14.

3.12 Correlation matrix



From the correlational matrix it was found that RM(average number of rooms per dwelling) had a very strong relationship with the dependent variables, which is obvious. There's no strong relationship of and other independent variables with the dependent variable.

4 Modeling

After cleaning the data and analysing the various features, we can proceed with model building. Here we will use a supervised machine learning model and since the target variable is continuous we will use linear regression, decision tree and random forest for model building.

4.1 Linear Regression

For linear regression models we need to work with features that are strongly related with features , they may be proportionally or inversely related. We will use the value of correlational matrix for setting the limit on features which we want to use for model building and calculate the errors.

	Feature Name	Total Feature	Corr value	r2 score	MAE	MSE
0	['RM' 'PTRAT' 'LSTAT' 'MEDV']	3	0.5	0.488164	4.40443	41.678
1	['RM' 'LSTAT' 'MEDV']	2	0.6	0.488164	4.14244	37.3831
2	['LSTAT' 'MEDV']	1	0.7	0.488164	4.86401	46.3363
3	['INDUS' 'NOX' 'RM' 'TAX' 'PTRAT' 'LSTAT' 'MEDV']	6	0.4	0.488164	4.3945	42.6519

As we can see from the table above we are selecting values of correlational matrix which are 0.5, 0.6, 0.7,0.4. Basically selecting features which are strongly related and values which are not so strongly related and comparing their errors. R2 score remains constant throughout but for MAE and MSE we can see that they are optimum around the corr value of 0.6. If we select features which are highly correlated and features which are highly non correlated the errors increase. So it would be wise to use corr value of 0.6 to build a linear model

4.2 Decision Tree

We can also build a decision tree model on this problem. For the decision tree model we made a random search for the hyperparameters of the decision tree.. After getting the parameters values from random search,we did a grid search and passed every value to parameters of the decision tree. And got the best parameters which gives the lowest error. Then we compared the test results with the default values of parameters of the decision tree with the best parameters and saw that the default parameter performed well. There may be various reasons for this one could be we did not give a broad range in random search as it would be too much computational load for the local machine.

4.3 Random Forest

In the random forest model we did the same thing as the above like the decision tree. We performed a random search then a grid search extracted the best parameters and compared the result with the default parameters. Here our random forest model performed better than the base model and we got an improvement of 2.18%.

4.4 Random Forest with feature improvement

In the above random forest model we used all the features.
But feature importance is a method where we can look at which features are important for model building in random forest.
From the figure we can see what weights each features. So after having a look at the importance value we decided to use the top 3 features to build a random forest.

We used the best value parameters from the above model and applied that to this model and we got a 0.58% from the base model.

	importance
RM	0.475831
LSTAT	0.337513
DIS	0.058791
CRIM	0.038557
NOX	0.018237
TAX	0.016091
PTRAT	0.015945
AGE	0.014654
B	0.012604
INDUS	0.006015
RAD	0.003753
ZN	0.001184
CHAS	0.000824

5 Future Work

In future we can add more features like location, population density and many more. We can also add more data. We can even deploy deep learning models for better accuracy.

6 Conclusion

The above models can be deployed by real estate companies, hotel booking companies for having a good estimate of the price of the property. As we were trying to build models on the local systems some the base models showed better performance, which could reverse when the model is built on cloud.

Thank You