Springboard Capstone Project 1

Predicting the price of the house

Abhishek Kumar

Contents

1 Introduction ………………………………………………………………………… 2

2 Data Acquisition and Cleaning ……………………………………………………….. 3

3 Data Exploration ……………………………………………………………………… 4

3.1 CRIM…………………………………………………………………………….. 4
3.2 ZN …………………………………………………………………………………4
3.3 INDUS……………………………………………………………………………..5
3.4 CHAS…………………………………………………………………………...5
3.5 NOX……………………………………………………………………………….6
3.6 RM…………………………………………………………………………………6
3.7 AGE……………………………………………………………………………….7
3.8 DIS…………………………………………………………………………………7
3.9 TAX……………………………………………………………………………...8
3.10 PTRATIO…………………………………………………………………………..8
3.11 LSTAT……………………………………………………………………………...9
3.12 Correlation matrix………………………………………………………………….9

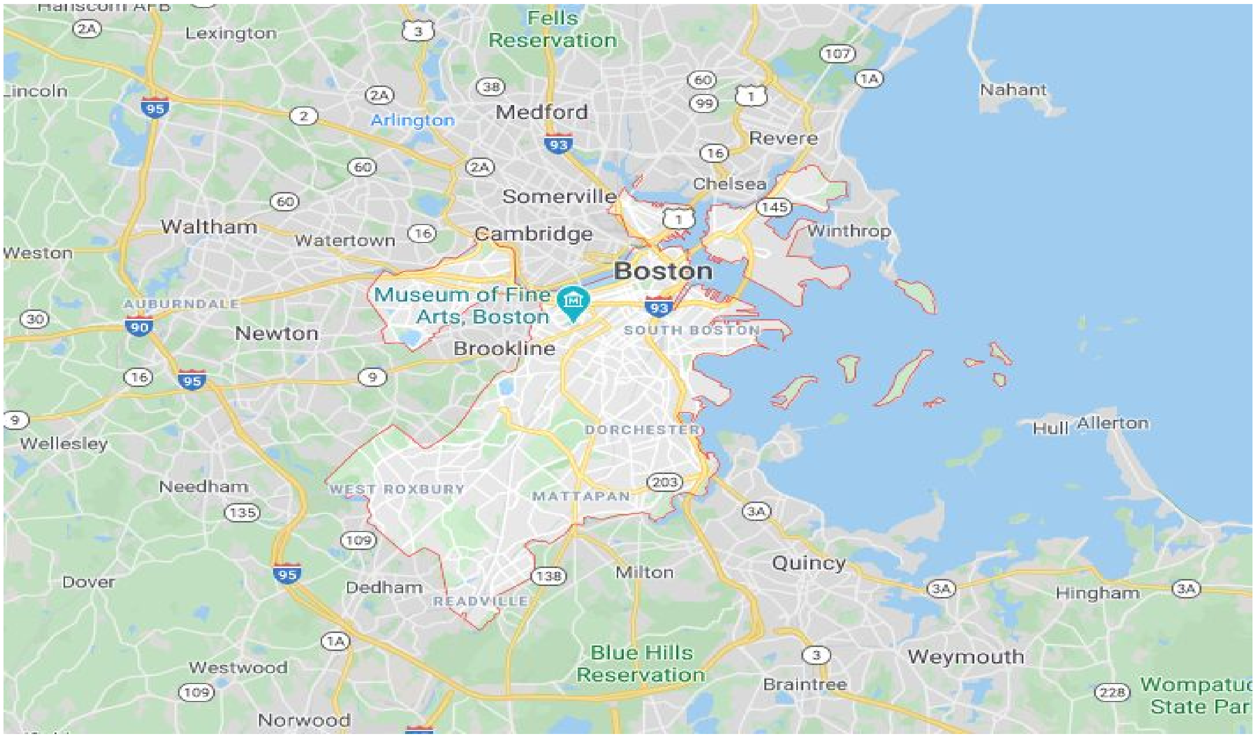
1 Introduction

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

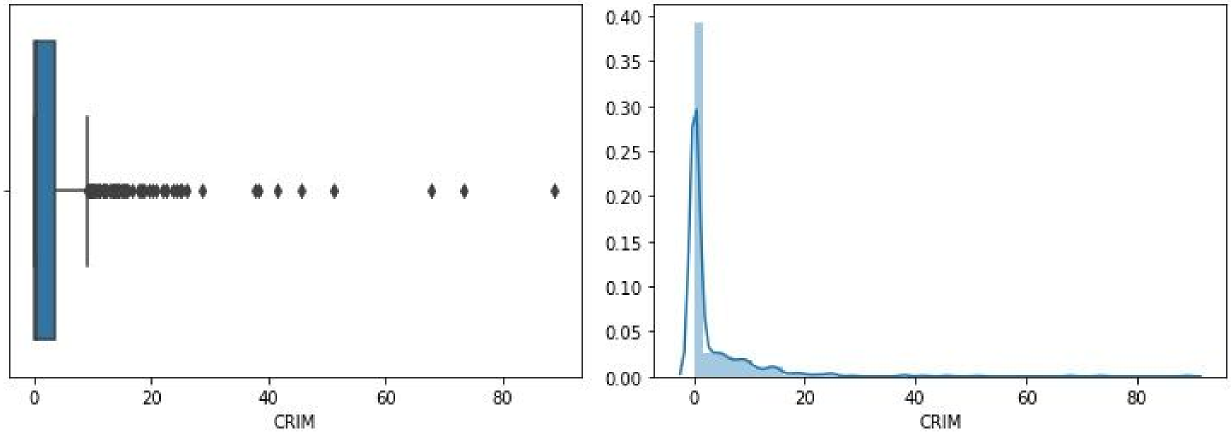
The downloaded data was needed to be converted to csv format, which was done through ms excel. The data was clean and no cleaning was required. The details of rows features are as follows

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

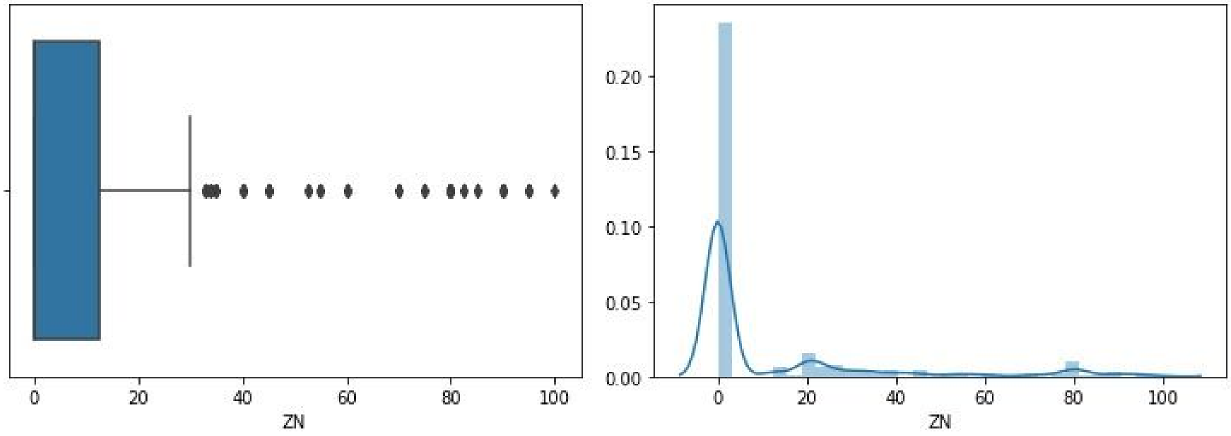
3.1 CRIM per capita crime rate by town

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.



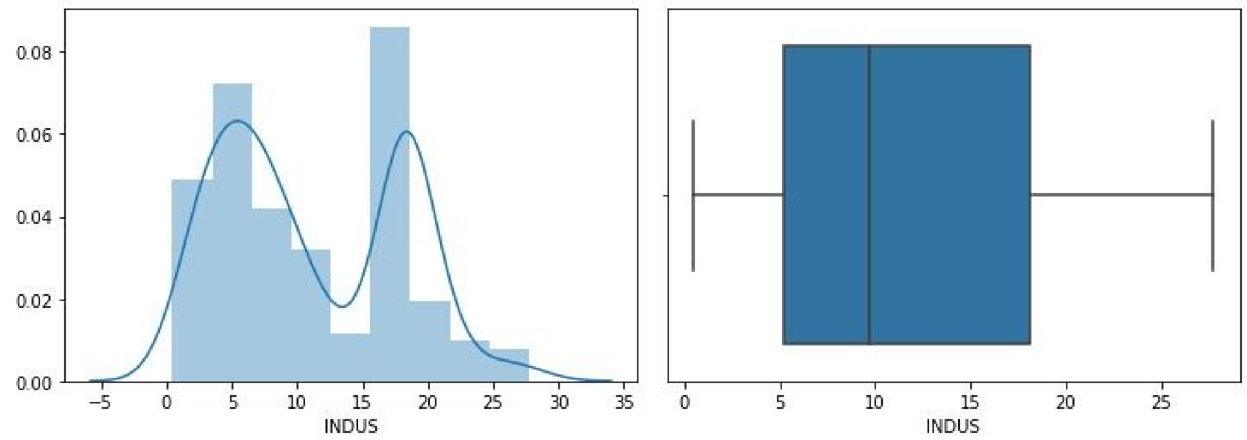
The mean of crime is 3.6 with a standard deviation of 8.6 and here we can see the crime doesn't have too many bars but there are a large number of outliers present.

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



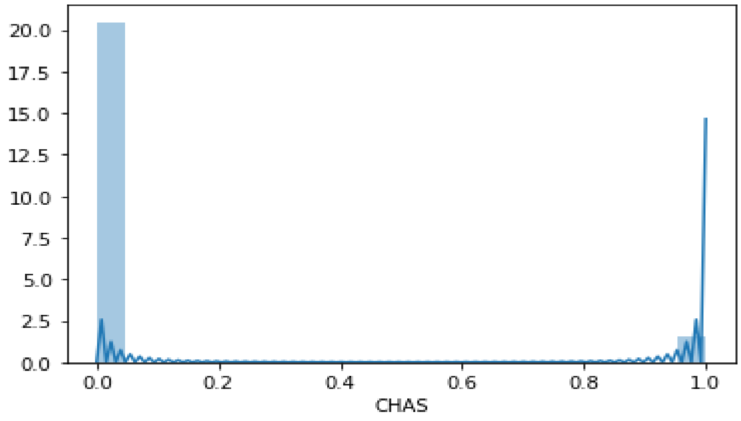
ZN too has lots of outliers and most of the values are close to zero.

3.3 INDUS   proportion of non-retail business acres per town



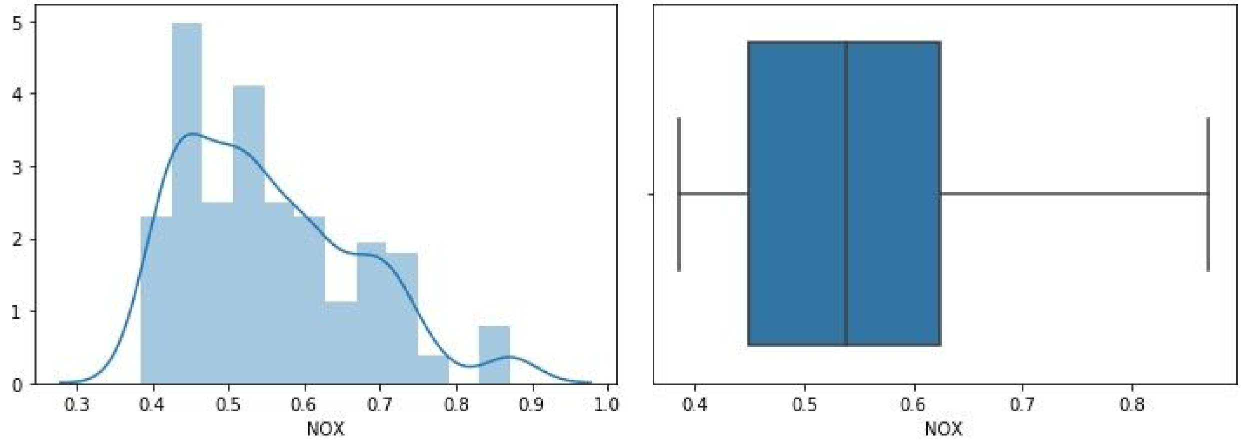
INDUS has no outliers and has two peaks .

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



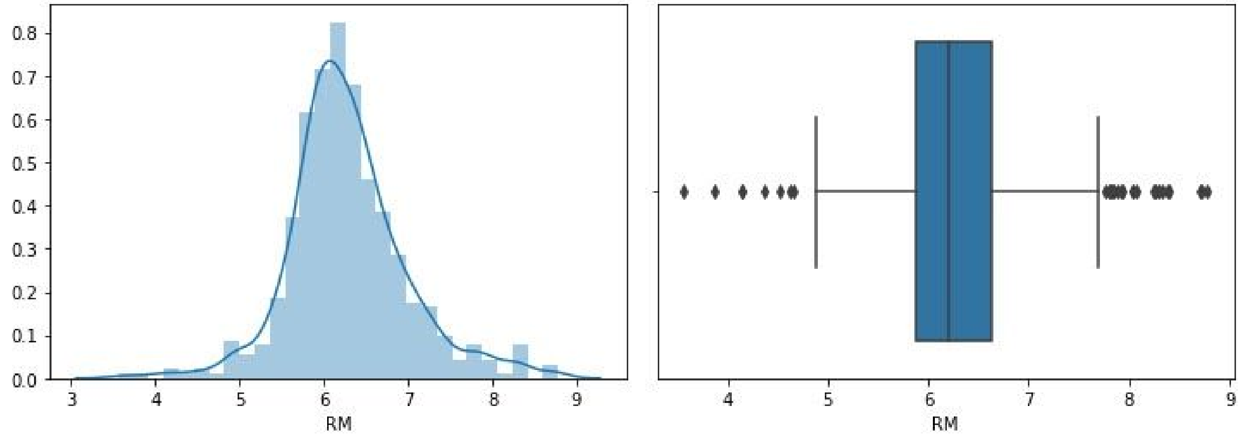
Charles is a river flowing in Boston. From the chat we can see that many people live in houses whose boundaries do not touch the charles river.

3.5 NOX nitric oxides concentration (parts per 10 million)



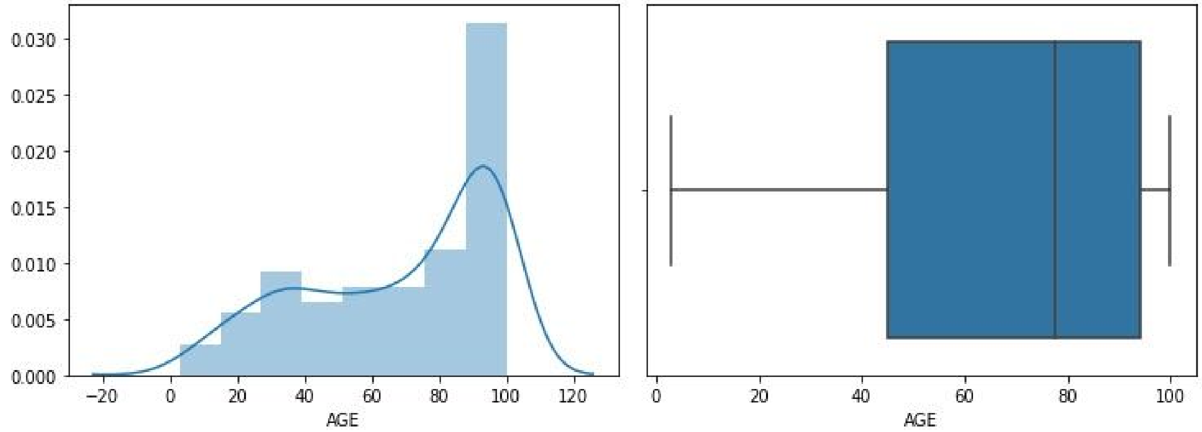
We can see from the plot that nitric oxide concentration has too many bars and the calculated mean is 0.55 The nitric oxide content is given in (parts per 10 million) and after looking for safe nitric oxide on wiki it was 250 (parts per 10 million) and from the data it was found to be safe with max value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

3.6 RM average number of rooms per dwelling



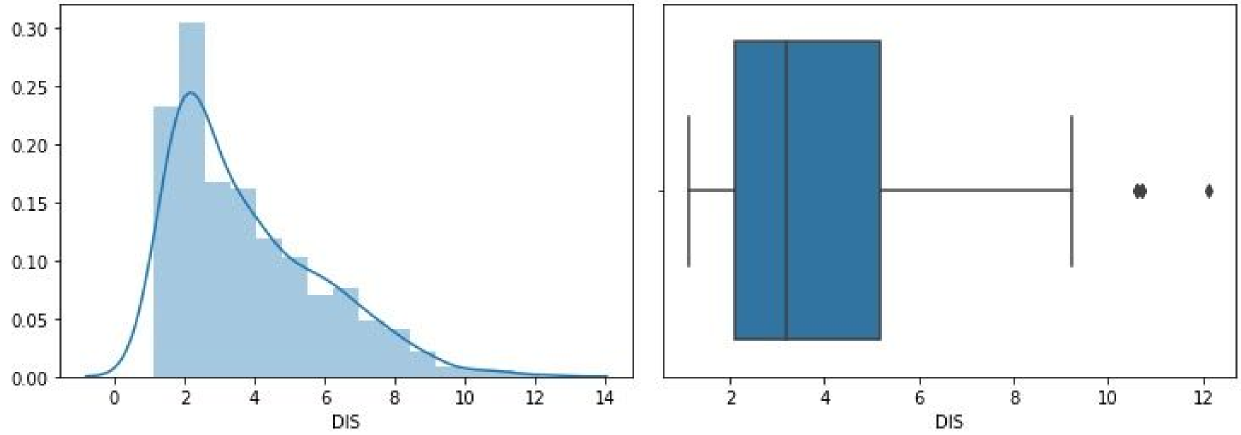
The mean number of rooms is 6.2 and also from the plot we can see frequency is most from 6 to 7 rooms. We can see that there are lots of outliers.

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



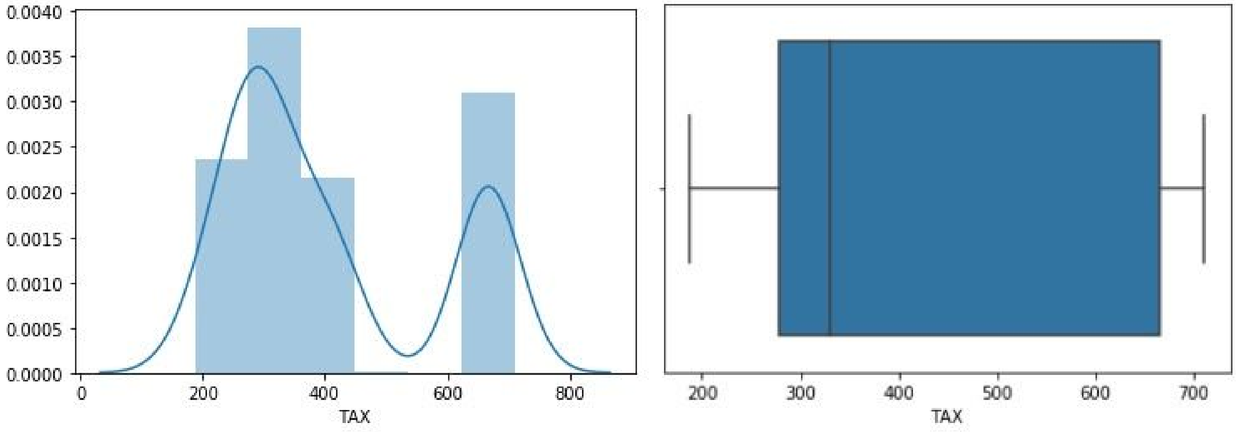
We can see that frequency is more at 100. So people like to live in buildings prior to 1940.

3.8 DIS weighted distances to five Boston employment centres



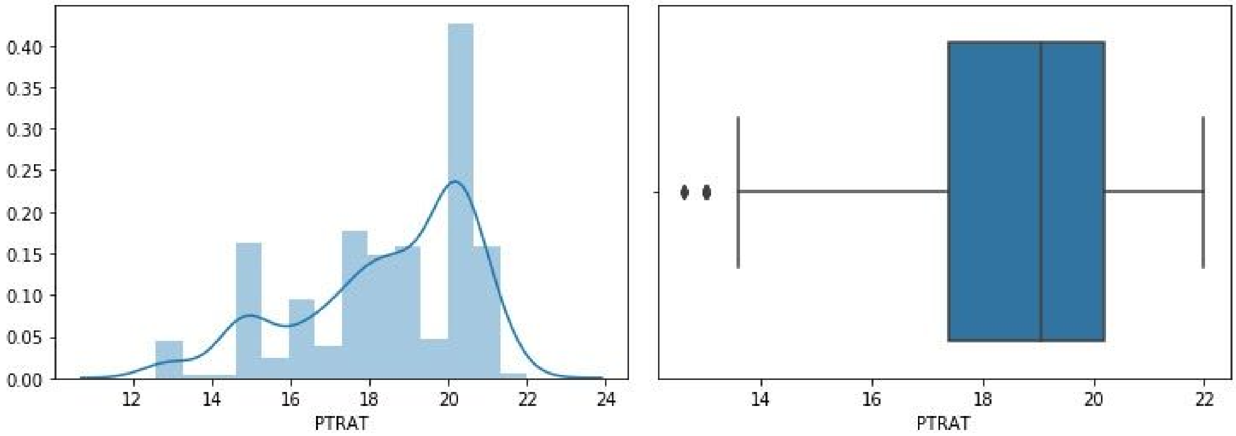
As we can expect people like to live near employment centers , this is also visible from this plot. Three outliers are present.

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



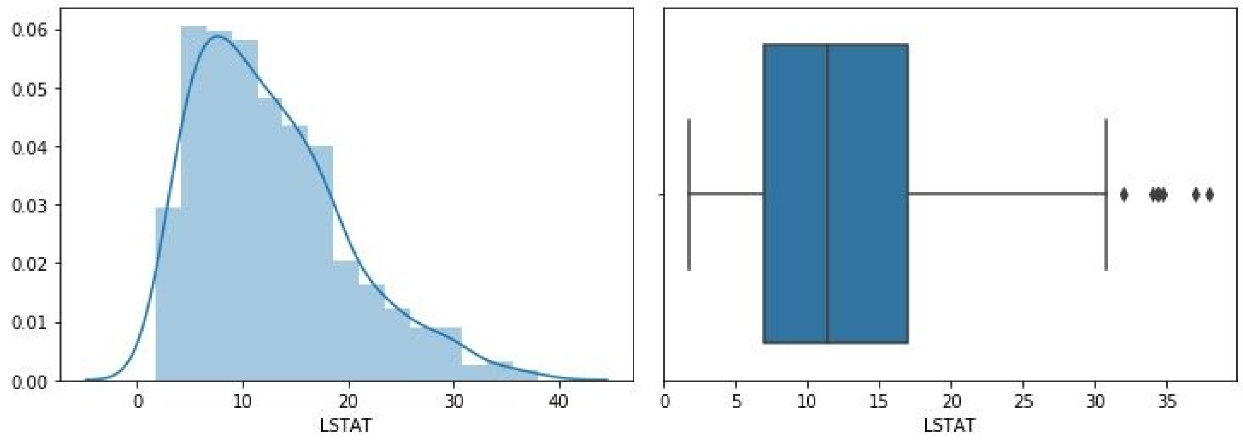
We can see people are willing to pay in the range $2000000 to $ 4000000 and we can see a surge at around $6000000 to $7000000.

3.10 PTRATIO pupil-teacher ratio by town



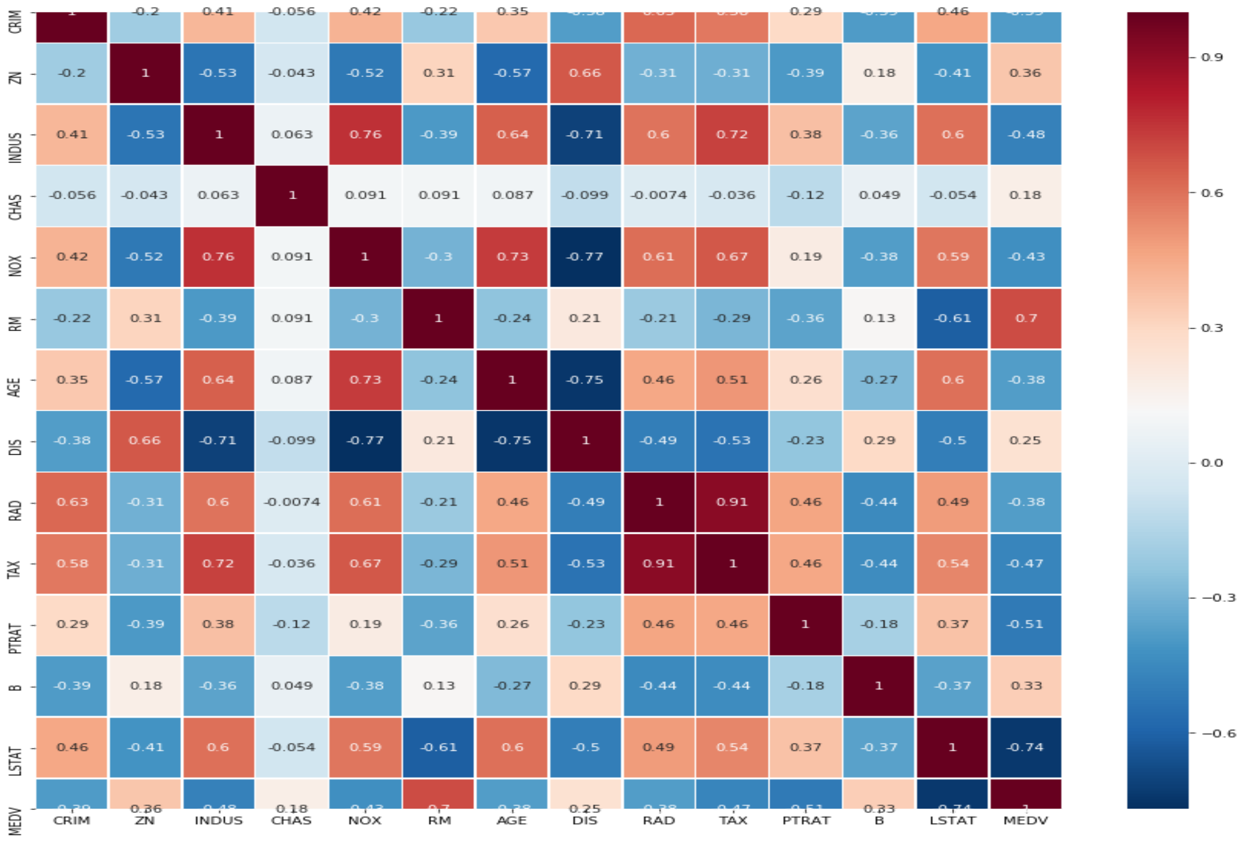
The average is 18.45 for pupil teacher ratio and the plot has many peaks. Two outliers are visible from the plots.

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.