**Boston House Price Prediction**

**By SABAL SHARMA**

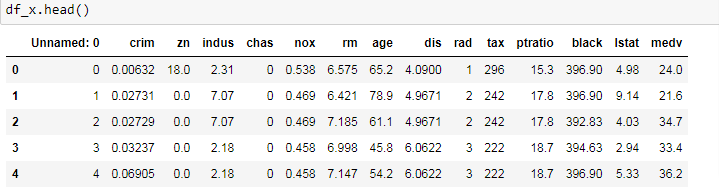
**THE PROBLEM**

Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. Ask a home buyer to describe their dream house, and they probably won’t begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition’s data-set proves that much more influences price negotiations than the number of bedrooms or a white-picket fence

**ABOUT THE DATASET**

Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. In this project, house prices will be predicted given explanatory variables that cover many aspects of residential houses. The goal of this project is to create a regression model that is able to accurately estimate the price of the house given the features. In this dataset made for predicting the Boston House Price Prediction. Here I just show the all of the feature for each house separately. Such as Number of Rooms, Crime rate of the House’s Area and so on. We’ll show in the upcoming part.

**DATA OVERVIEW**

****

1. **CRIM** per capital crime rate by town

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

3. **INDUS** proportion of non-retail business acres per town

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

5. **NOX** nitric oxides concentration (parts per 10 million)

6. **RM** average number of rooms per dwelling

7. **AGE** proportion of owner-occupied units built prior to 1940

8. **DIS** weighted distances to five Boston employment centers

9. **RAD** index of accessibility to radial highways

10.**TAX** full-value property-tax rate per 10,000 USD

11. **PTRATIO** pupil-teacher ratio by town

12. **Black** 1000(Bk — 0.63)² where Bk is the proportion of blacks by town

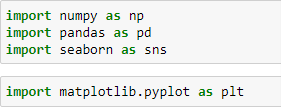
13. **LSTAT** % lower status of the population

**About the Algorithms used in**

The major aim of in this project is to predict the house prices based on the features using some of the regression techniques and algorithms.

1. **Linear Regression**

**Machine Learning Packages are used for in this Project**

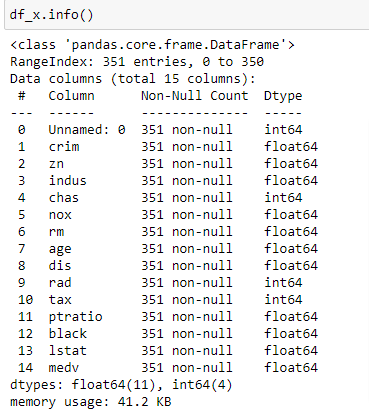


This Dataset consist several features such as Number of Rooms, Crime Rate, and Tax and so on. Let’s know about how to read the dataset into the Jupyter Notebook. You can download the dataset from [Kaggle](https://www.kaggle.com/altavish/boston-housing-dataset" \t "_blank) in csv file format.

As well we can also able to get the dataset from the sklearn datasets.

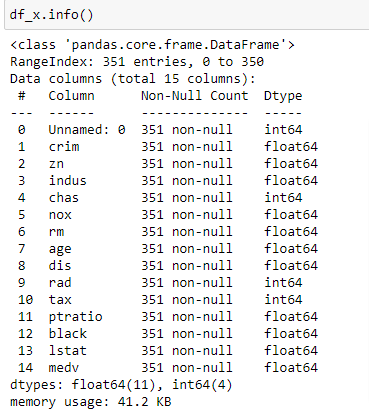
House Price Pred (2).png

**Learning about the dataset**

****

**Data Preprocessing**

**Checking for null values:**

****

We find that the dataset is already cleaned

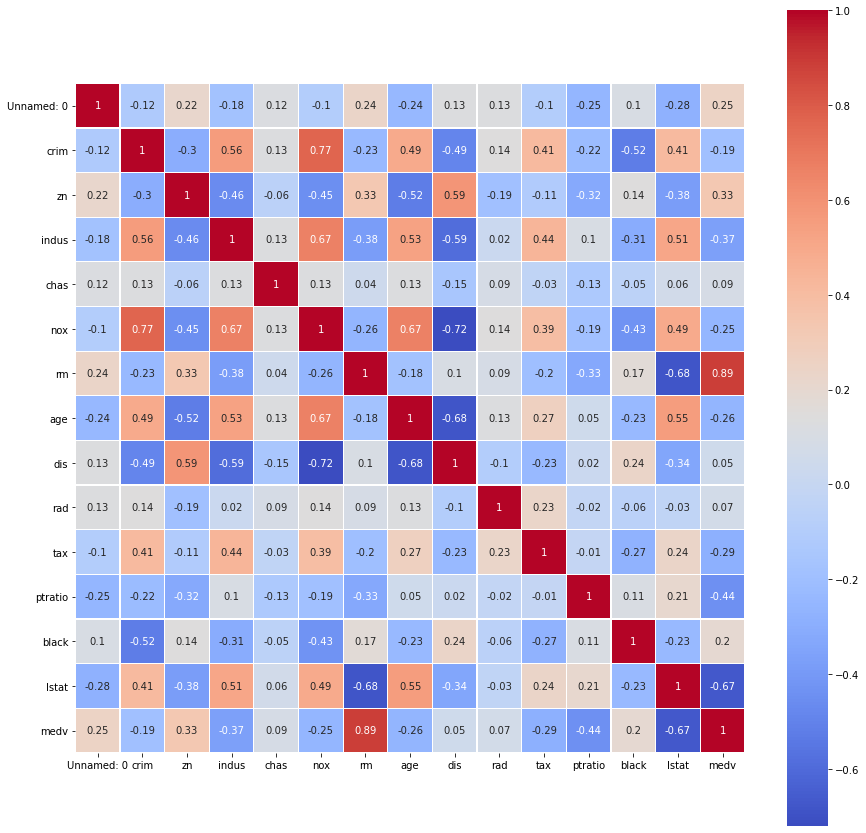
Also Here our target variable is medv

**Exploratory Data Analysis**

In statistics, exploratory data analysis (**EDA**) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily **EDA** is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

First Understanding the correlation of features between target and other features.

House Price Pred (4).png



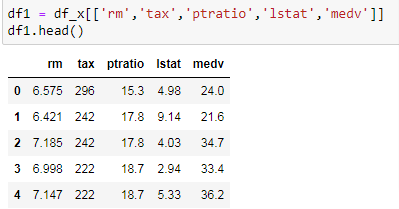
The Big colorful picture above which is called Heatmap helps us to understand how features are correlated to each other.

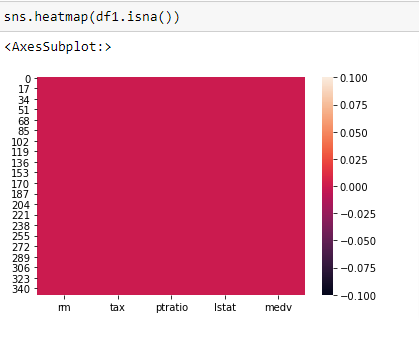
1.Postive sign implies postive correlation between two features whereas Negative sign implies negative correlation between two features.

2.I am here interested to know which features have good correlation with our dependent variable MEDV and can help in having good predictions.

3.I observed that INDUS, RM, TAX, PTRATIO and LSTAT shows some good correaltion with MEDV and I am interested to know more about them.

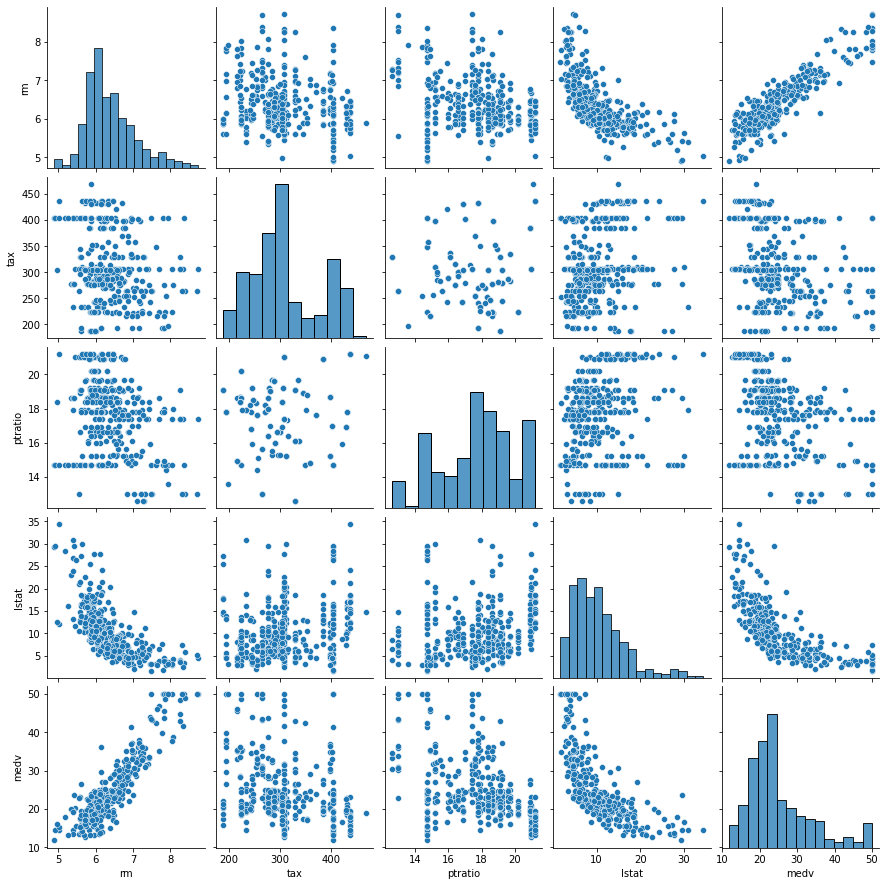
4.However I noticed that INDUS shows good correlation with TAX and LSAT which is a pain point because it leads to Multicollinearity. So I decided NOT to consider this feature and do further analysis with other 5 remaining features.

****

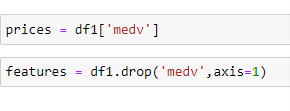
Checking null values in df1

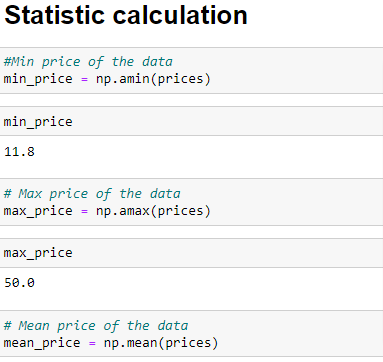
Let us try and plot all the co-relation

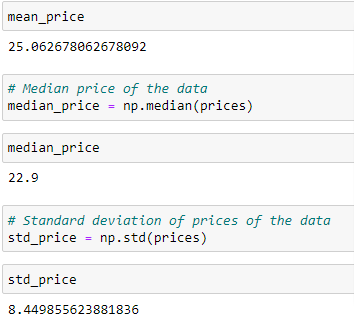
House Price Pred (7).png



Now since MEDV is our target variable let’s explore it more with statistical analysis.







**Model Fitting**



Here the model gave us an accuracy of 0.7967322690509783

i.e. 79.67%

**Output & Conclusion**

From the Exploratory Data Analysis, we could generate insight from the data. How each of the features relates to the target. Also, the evalution of the linear model can be seen.

**Business Recommendation**

Using this idea we can create a user friendly application to deduce the required capital amount for a dream house based on its features.