Assignment

Project Name : Boston Housing

Problem Statement: Predict the value of Prices of the house using the given features in Boston Housing dataset.

Pakage Required for this Assignment

1> numpy

2> pandas

3> matplotlib

4> seaborn

5> warnings

5> sklearn

6> pickle

Here I write the meaning of feature which is given in our dataset.

CRIM : per capital 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

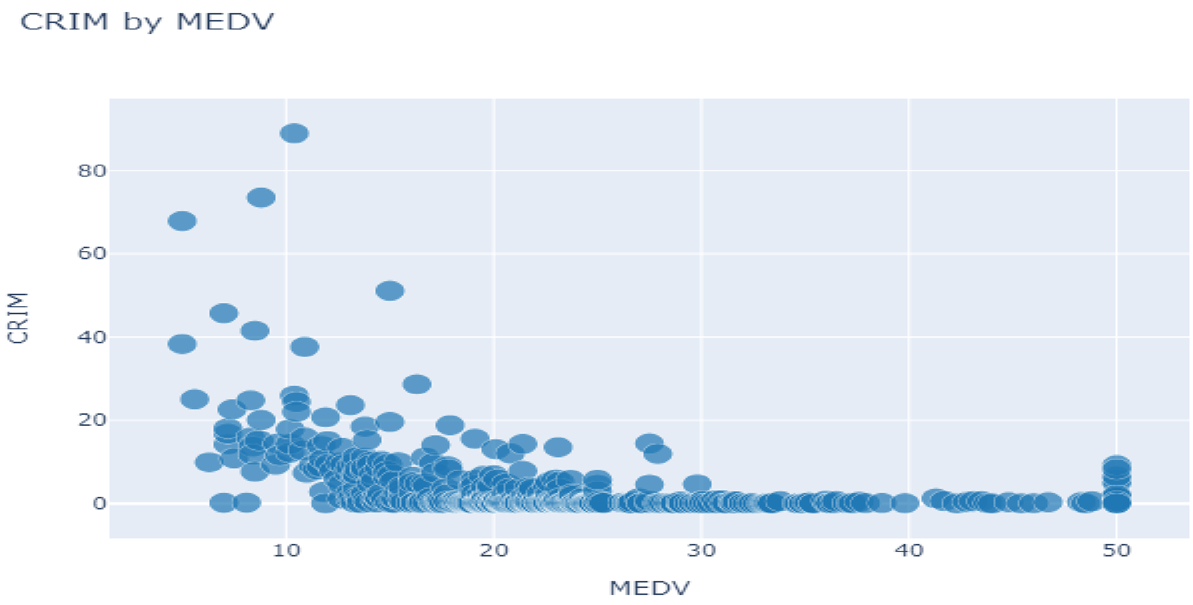
TAX : full-value property-tax rate per $10,000

PTRATIO : pupil-teacher ratio by town

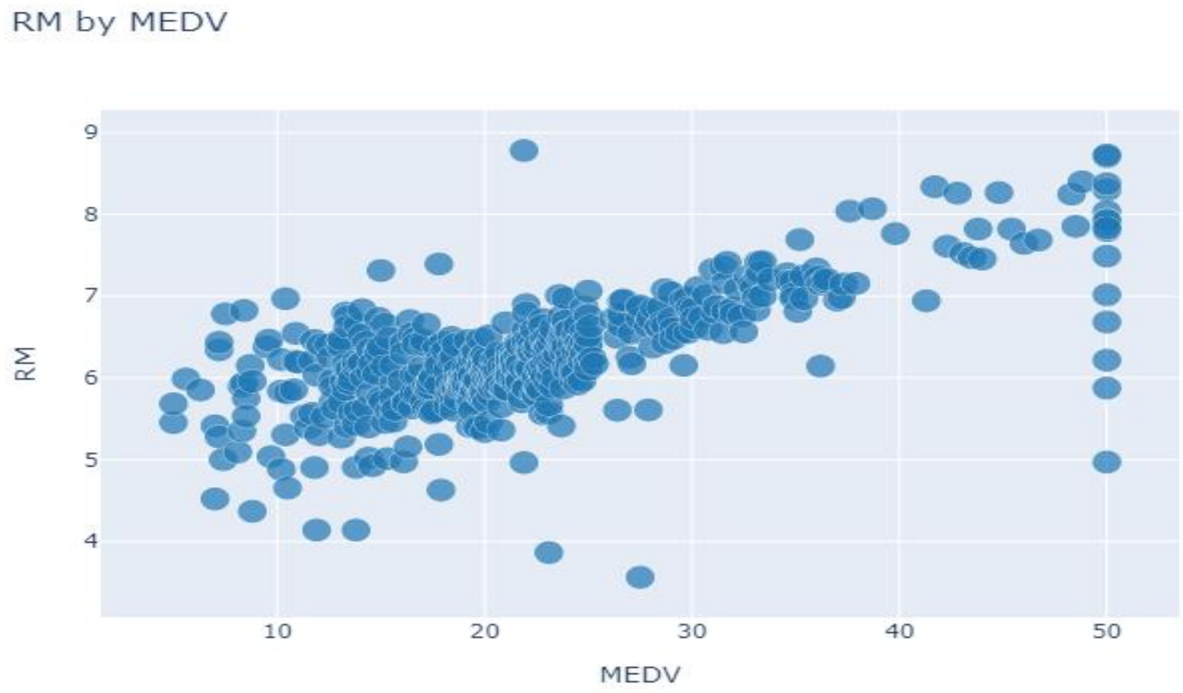
B : 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

LSTAT : % lower status of the population

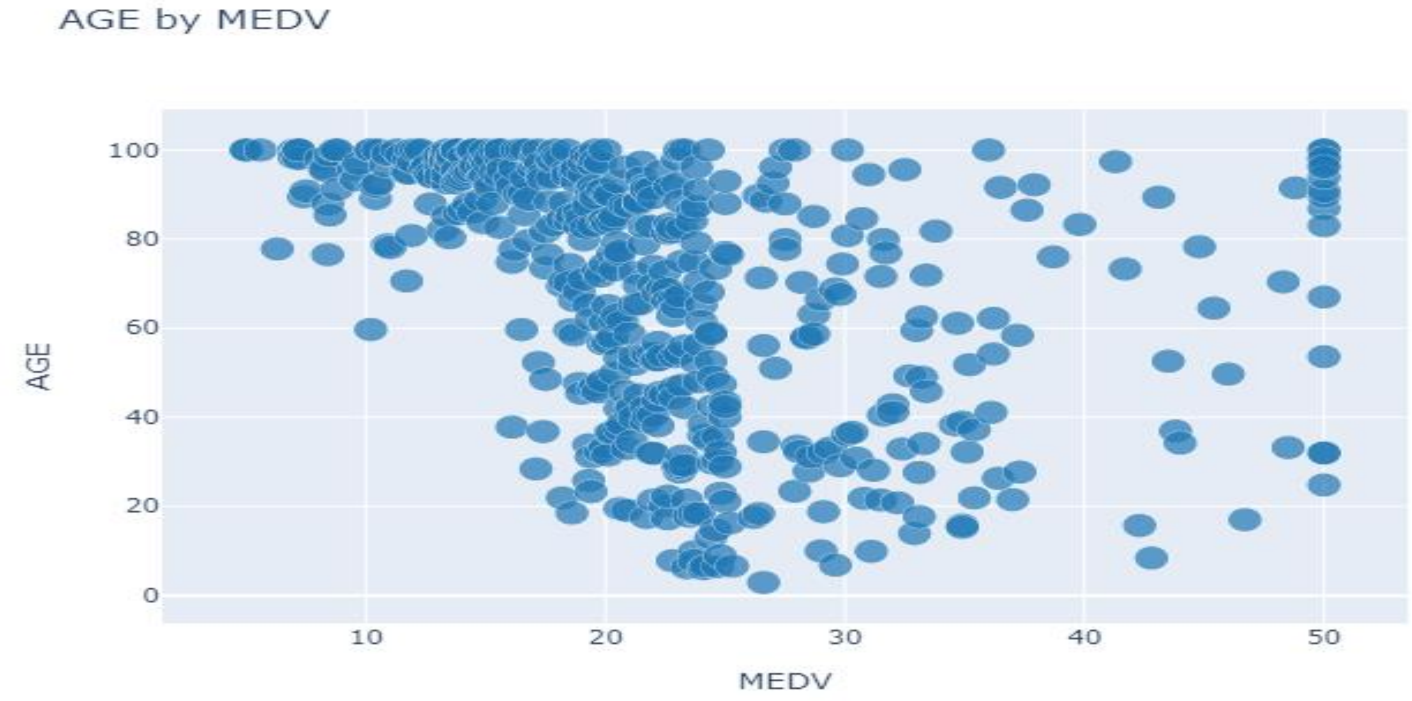
MEDV : Median value of owner-occupied homes in $1000's



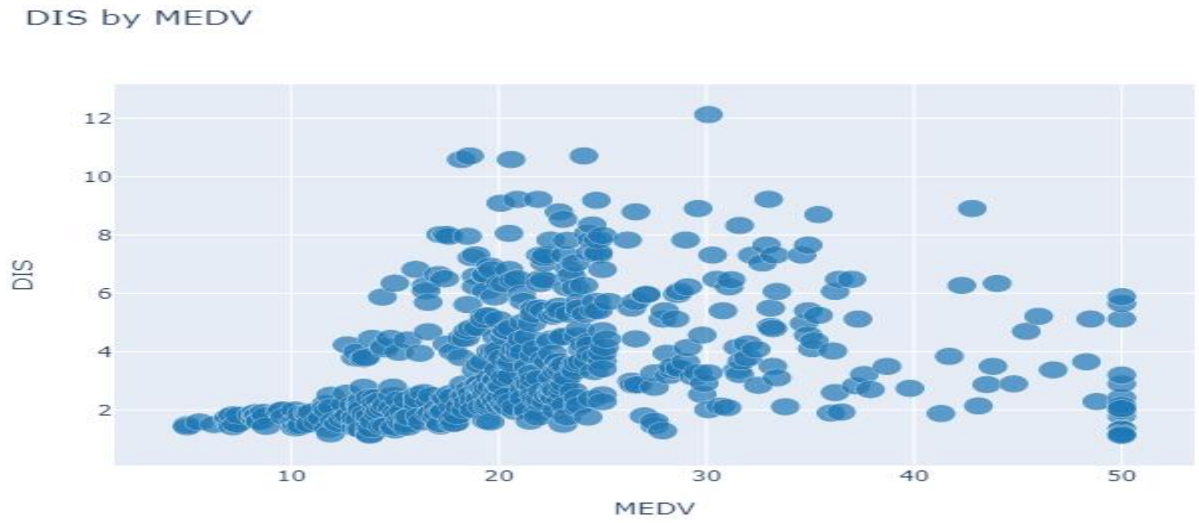
Here we can see that when price of house is increases than crime rate of that area is decreases. By this graph we conclude that when give more money for house then maximum chance that we get good house where crime rate is no more. In this graph we can see the crime rate is high in the range of home price 10000$ to 18000$ but when we increase the price of home then crime rate of that area is low.



In this graph we can see that when price of house is increases then number of rooms in house is increases. That mean when we give more money for house then maximum possibility that we get those house which in more rooms is present.

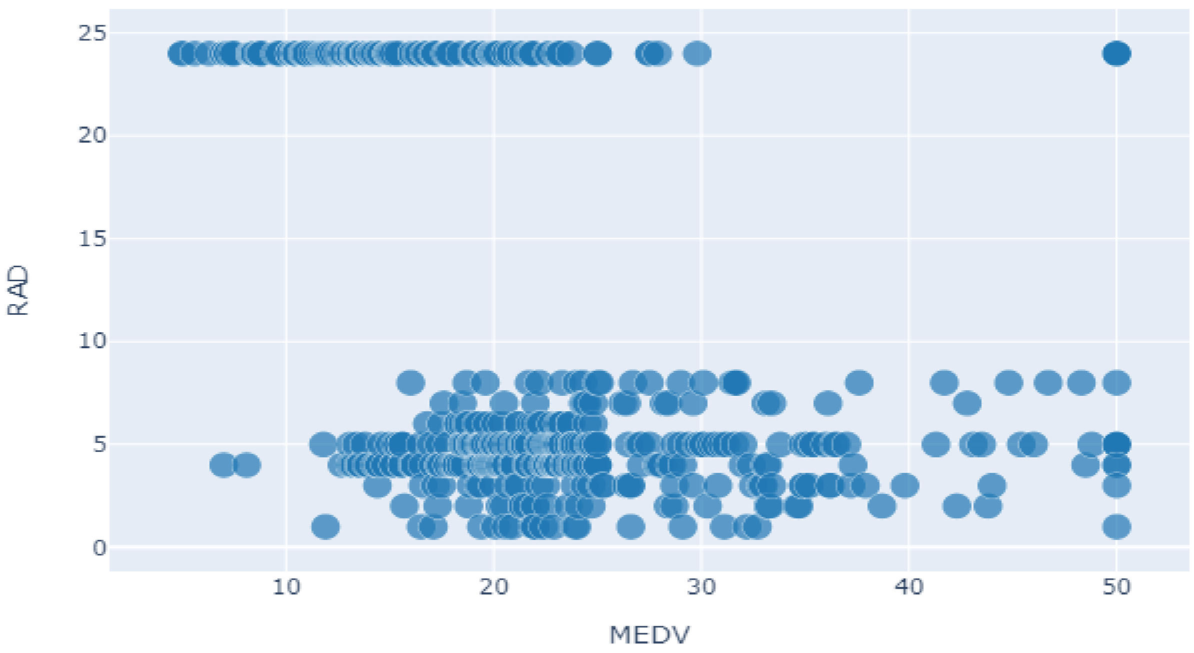


Here we can see that those home is most likely by person which price is lie between 15000$ to 25000$. In this home price range every person whose age is greater than 20 years want buy the more home. By this graph we conclude that mostly good house is get 15000$ to 25000$ and in this price range crime rate is also less. In above graph we see the crime rate is low when home price is lie between 15000$ to 25000$.



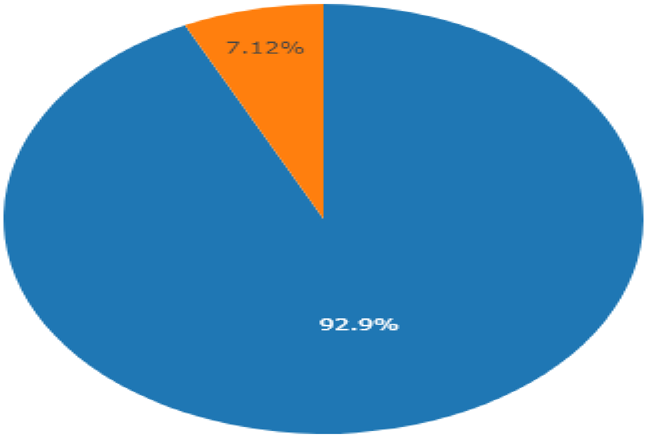
In this graph we see the weighted distances to five Boston employment center is high when price of home is greater than 15000$. That mean the mostly home is likely by people whose weighted distances to five Boston employment center is near by home and this type of home is mostly lie in between 12000$ to 26000$.

RAD by MEDV

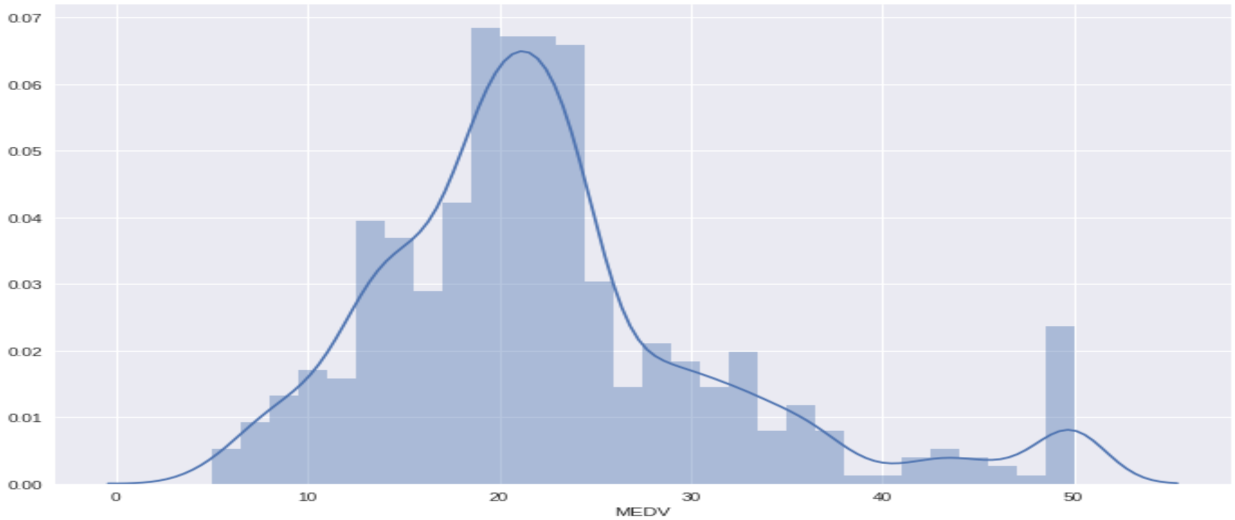


By using this graph we can see the radial highway is lie between 18000$ to 25000$. When we increase the price of home then after some price of home radial highway is aproximate constant with respect to price of home. In this maximum radial highway lie between 1 to 8.

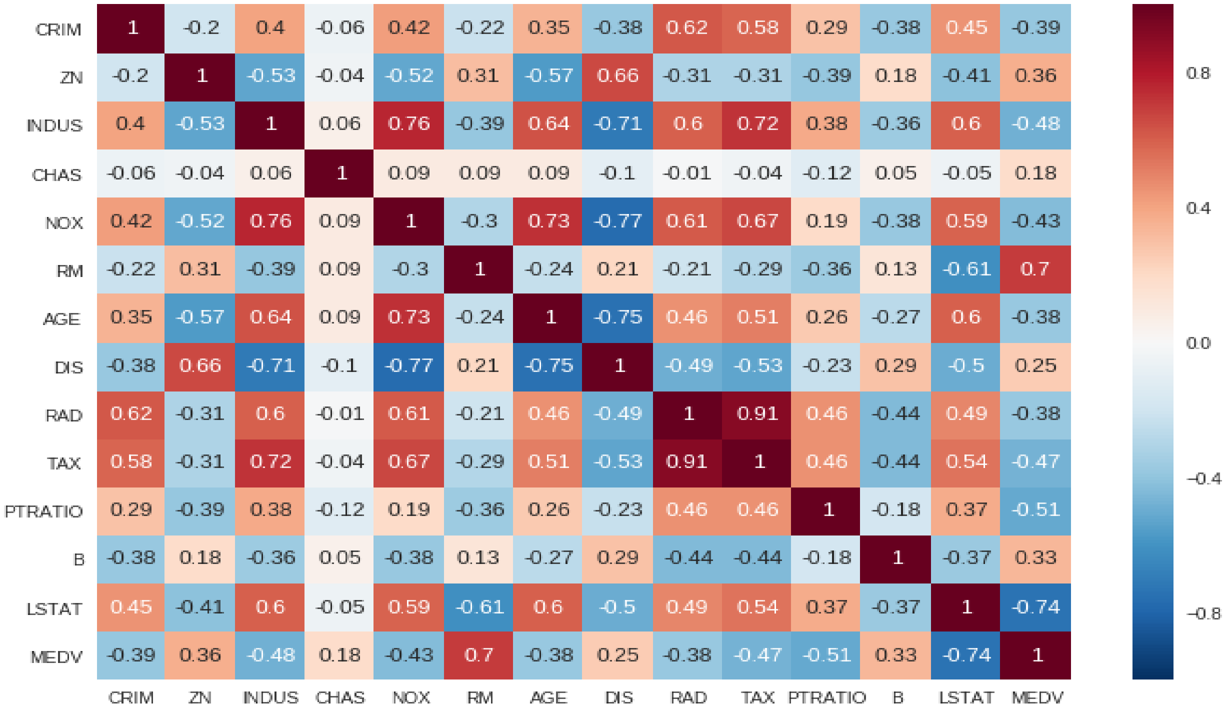
RM by CHAS



In this pie Chart we can see 92.9% river is lie near to Home and 7.12% river is not lie near to home. Which home near to river that means those home price is less compare to those home which is not near to river.



We see that the values of home are distributed normally with few outliers. In this graph we also see that maximum home is present in this dataset which price is lie between 18000 dollar to 27000 dollar of home.



Now we create a correlation matrix that measures the linear relationships between the variables. The correlation matrix can be formed by using the corr function from the pandas dataframe library. We will use the heatmap function from the seaborn library to plot the correlation matrix.

The correlation coefficient range from -1 to 1. If the value is close to 1, it means that there is a strong positive correlation between the two variables. When it is close to -1, the variables have a strong negative correlation.

Splitting the data into training and testing sests

We split the data into training and testing sets. We train the model with 75% of the samples and test with the remaining 25%. We do this to access the model's performance on unseen data. To split the data we use train_test_split function provided by sklearn library. We finally print the size of our training and test set to verify if the splitting has occurred properly.

Model Prediction

For model prediction I used RandomForestRegressor algorithm because when I used other algorithm like decision tree, linear regression or SVM then I get high mean absolute error while in RandomForestRegressor I get low mean absolute error.

Observations

1> To fit a RandomForest regressor model, we select those features which have high correlation with our target variable price of home (MEDV). By looking at the correlation matrix we can see that average number of rooms per dwelling has a strong positive correlation with price of home. Where as % lower status of the population has high negative correlation with price of home(MEDV).

2> An important point in selecting features for a RandomForest Regressor model is to check for multi-co-linearity. The features RAD,TAX have a correlation. These feature pairs are strongly correlated to each other. We should not select both these features together for training the model. Same goes for the features DTS and AGE which have a correlation of -0.75.

Based on the above observations we will RM and LSTAT as our features. Using a scatter plot let's see how these features vary with MEDV.