Velammal Institute of Technology
Velammal Knowledge Park, Panchetti, Chennai.

# PREDICTING HOUSE PRICES USING MACHINE LEARNING

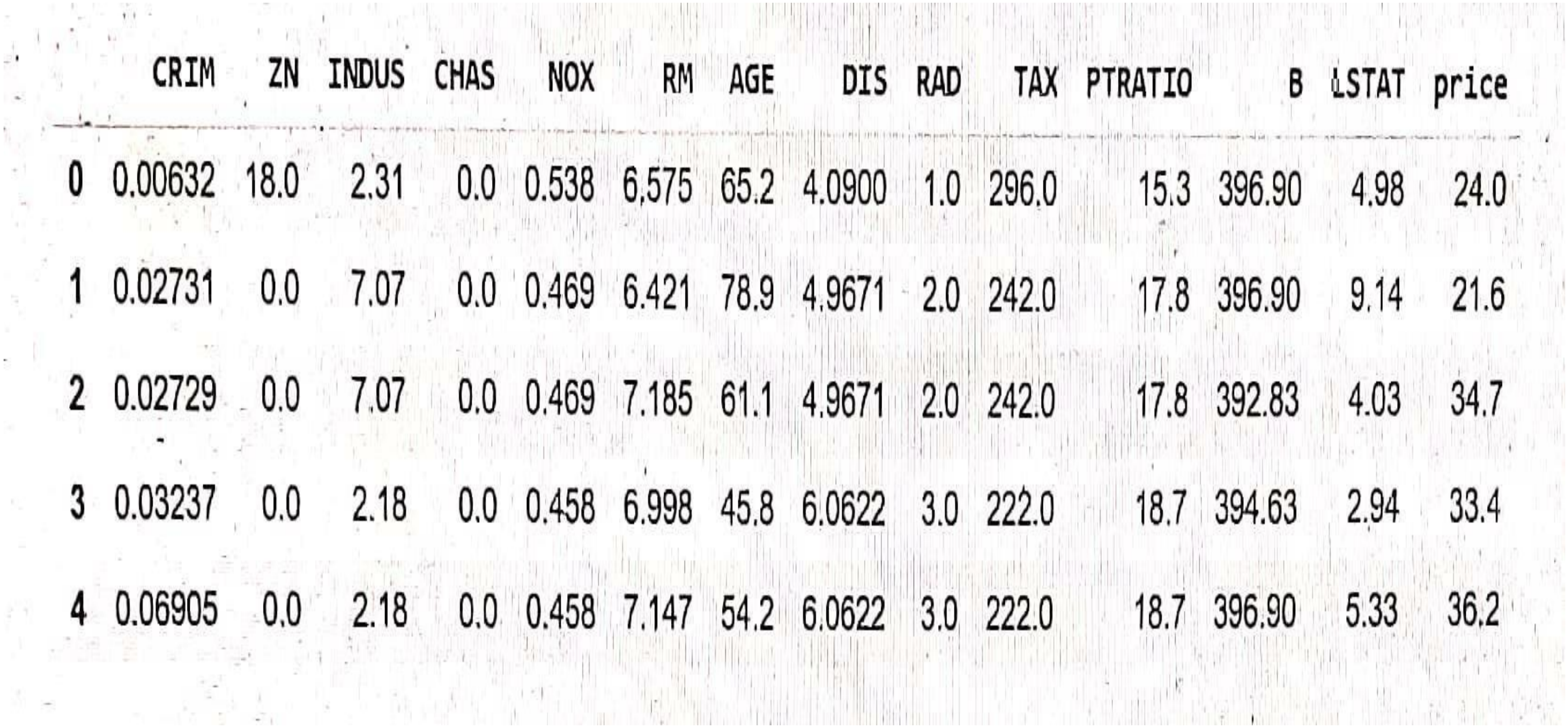
**IBM GROUP 1**
**ARITIFICIAL INTELLIGENCE – PHASE 3**

***Department of ECE:***
***Sai Preethi K P (113321106078)***
***Papita Biswas (113321106066)***
***S Sharada (113321106096)***
***Sandhiya B (113321106080)***
***Srinidhi M (113321106093)***

# EXPLANATION OF THE OUTPUT RESULTS AND THE DATASET

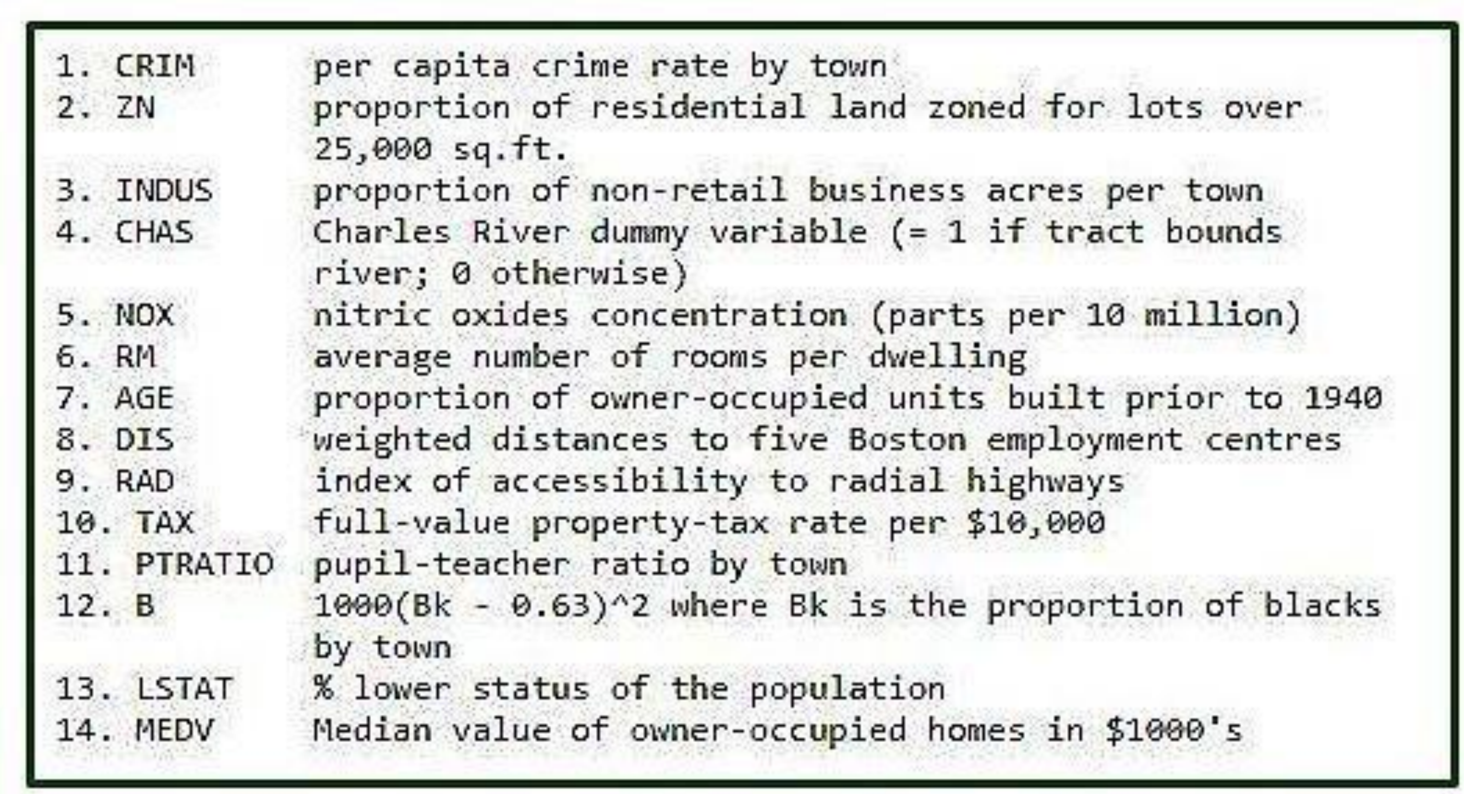
First we import a sample data from scleral library , you can get different types of sample data from Kaggle. The data taken here is the data of various parameters and the house prices in a given city called Boston in the year between 1970 to 2020.

Here the data parameters are explained as follows:



|   | CRIM | ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO | B | LSTAT | price |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.00632 | 18.0 | 2.31 | 0.0 | 0.538 | 6,575 | 65.2 | 4.0900 | 1.0 | 296.0 | 15.3 | 396.90 | 4.98 | 24.0 |
| 1 | 0.02731 | 0.0 | 7.07 | 0.0 | 0.469 | 6.421 | 78.9 | 4.9671 | 2.0 | 242.0 | 17.8 | 396.90 | 9.14 | 21.6 |
| 2 | 0.02729 | 0.0 | 7.07 | 0.0 | 0.469 | 7.185 | 61.1 | 4.9671 | 2.0 | 242.0 | 17.8 | 392.83 | 4.03 | 34.7 |
| 3 | 0.03237 | 0.0 | 2.18 | 0.0 | 0.458 | 6.998 | 45.8 | 6.0622 | 3.0 | 222.0 | 18.7 | 394.63 | 2.94 | 33.4 |
| 4 | 0.06905 | 0.0 | 2.18 | 0.0 | 0.458 | 7.147 | 54.2 | 6.0622 | 3.0 | 222.0 | 18.7 | 396.90 | 5.33 | 36,2 |

- Here the data parameters are explained as follows:



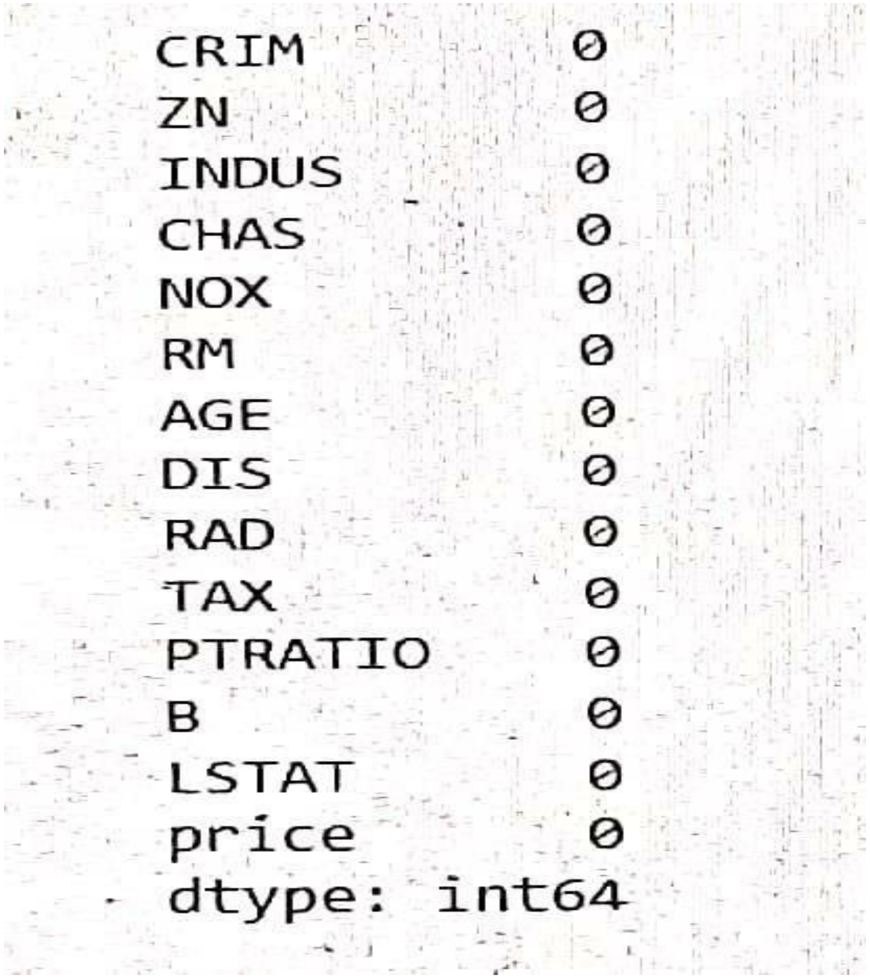
```
1. CRIM      per capita crime rate by town
2. ZN        proportion of residential land zoned for lots over
             25,000 sq.ft.
3. INDUS     proportion of non-retail business acres per town
4. CHAS      Charles River dummy variable (= 1 if tract bounds
             river; 0 otherwise)
5. NOX       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 centres
9. RAD       index of accessibility to radial highways
10. TAX      full-value property-tax rate per $10,000
11. PTRATIO  pupil-teacher ratio by town
12. B        1000(Bk - 0.63)^2 where Bk is the proportion of blacks
             by town
13. LSTAT    % lower status of the population
14. MEDV     Median value of owner-occupied homes in $1000's
```

- Here for understanding purpose we have taken first 5 index/instance of data and printed them.

- In total there are 506 rows of data from the dataset , of which we have printed first 5 rows using head() function.

- There are 14 columns in total, i.e., 13 columns containing data of the place, and the 14th column is the target column which contains the house prices.

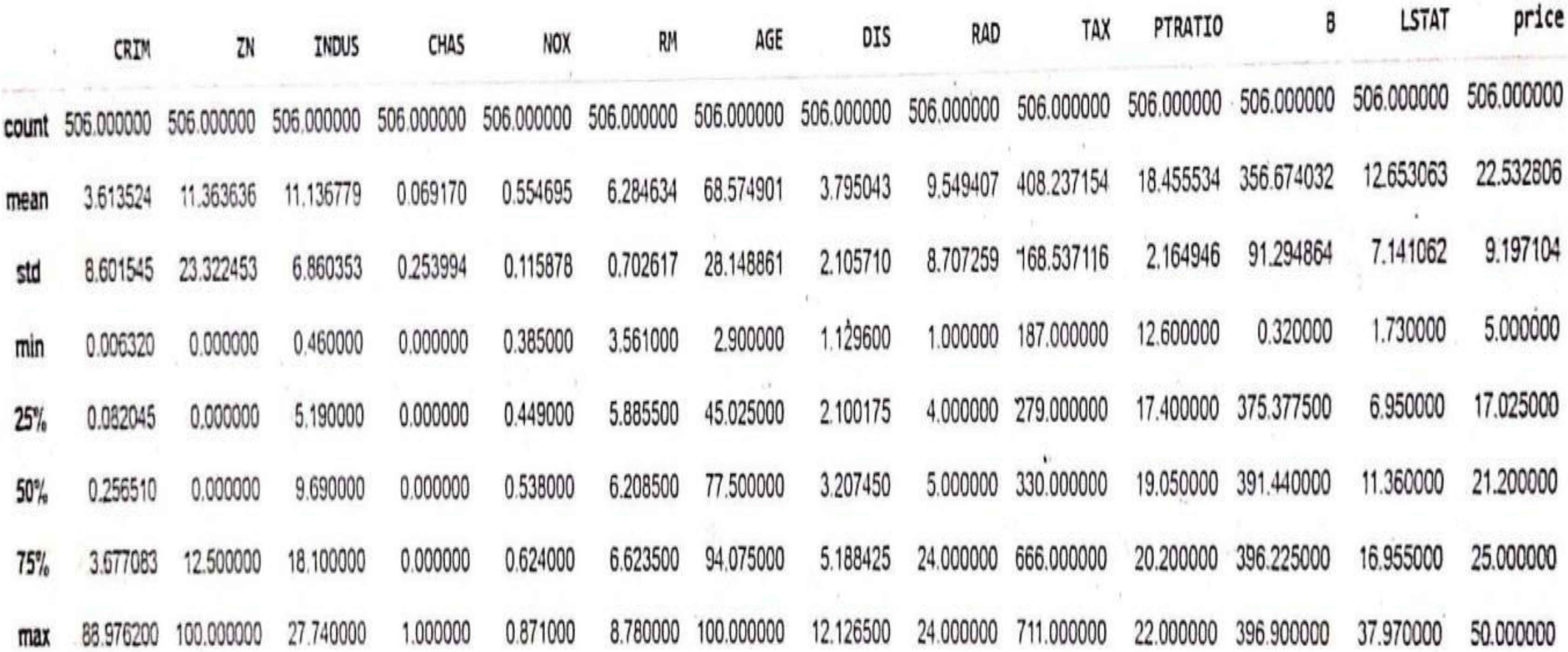
- Then we check if our data has some null values i.e. missing values. Since if the data is incomplete , then there will be error during processing state which may lead to loss of accuracy in predicting model. Here in our given data , there is no missing value as we can see.



```
CRIM       0
ZN         0
INDUS      0
CHAS       0
NOX        0
RM         0
AGE        0
DIS        0
RAD        0
TAX        0
PTRATIO    0
B          0
LSTAT      0
price      0
dtype: int64
```

- Since our data contains no missing value, the program will skip the dropping phase in data processing, where data is dropped to increase accuracy and fit missing values in a way so that it is suitable for modeling.

- Next we try to describe the data in such a way so that both people and machine find it easy to understand the given data . In order to do this we use the describe() function.



| | CRIM | ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO | B | LSTAT | price |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 506.000000 | 506.000000 | 506.000000 | 506.000000 | 506.000000 | 506.000000 | 506.000000 | 506.000000 | 506.000000 | 506.000000 | 506.000000 | 506.000000 | 506.000000 | 506.000000 |
| mean | 3.613524 | 11.363636 | 11.136779 | 0.069170 | 0.554695 | 6.284634 | 68.574901 | 3.795043 | 9.549407 | 408.237154 | 18.455534 | 356.674032 | 12.653063 | 22.532806 |
| std | 8.601545 | 23.322453 | 6.860353 | 0.253994 | 0.115878 | 0.702617 | 28.148861 | 2.105710 | 8.707259 | 168.537116 | 2.164946 | 91.294864 | 7.141062 | 9.197104 |
| min | 0.006320 | 0.000000 | 0.460000 | 0.000000 | 0.385000 | 3.561000 | 2.900000 | 1.129600 | 1.000000 | 187.000000 | 12.600000 | 0.320000 | 1.730000 | 5.000000 |
| 25% | 0.082045 | 0.000000 | 5.190000 | 0.000000 | 0.449000 | 5.885500 | 45.025000 | 2.100175 | 4.000000 | 279.000000 | 17.400000 | 375.377500 | 6.950000 | 17.025000 |
| 50% | 0.256510 | 0.000000 | 9.690000 | 0.000000 | 0.538000 | 6.208500 | 77.500000 | 3.207450 | 5.000000 | 330.000000 | 19.050000 | 391.440000 | 11.360000 | 21.200000 |
| 75% | 3.677083 | 12.500000 | 18.100000 | 0.000000 | 0.624000 | 6.623500 | 94.075000 | 5.188425 | 24.000000 | 666.000000 | 20.200000 | 396.225000 | 16.955000 | 25.000000 |
| max | 88.976200 | 100.000000 | 27.740000 | 1.000000 | 0.871000 | 8.780000 | 100.000000 | 12.126500 | 24.000000 | 711.000000 | 22.000000 | 396.900000 | 37.970000 | 50.000000 |

- Counts refers to the number of instances of data in each column i.e 506 since there are 506 rows of data for each column Mean refers to mean value of data in given colum.

- Std means the standard value  the most common value in given set of data for a particular column.

- Min refers the least data value in each column.

- Max refers to the maximum data value in each column.

- 25% refers that 25 percentile of the data in that column is equal to or below that value.

- Next we try to understand the correlation between the different values, in order to do that, the best way is by using heat map. Heat map is a representation of data in the form of a map or diagram in which data values are represented as colours.

- **Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate)**

- There are two types of correlation, they are:

- **Positive correlation:** A positive correlation is a relationship between two variables that move in tandem—that is, in the same direction. A positive correlation exists when one variable decreases as the other variable decreases, or one variable increases while the other increases.

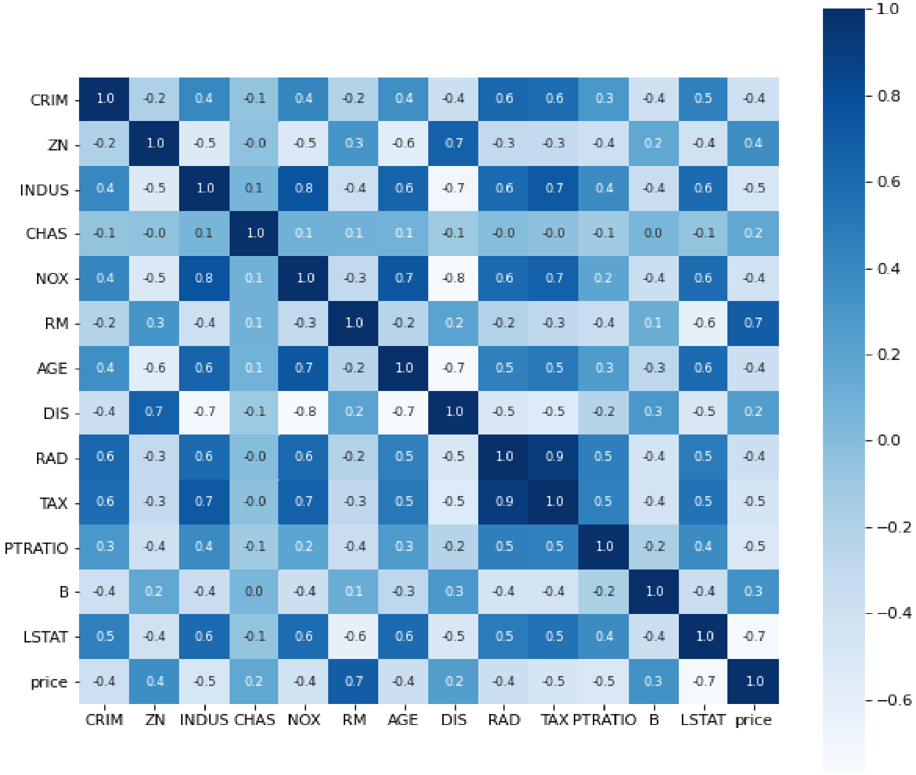
- **Negative correlation:** Negative correlation is a relationship between two variables in which one variable increases as the other decreases, and vice versa.

- In statistics, a perfect negative correlation is represented by the value -1.0, while a 0 indicates no correlation, and +1.0 indicates a perfect positive correlation.

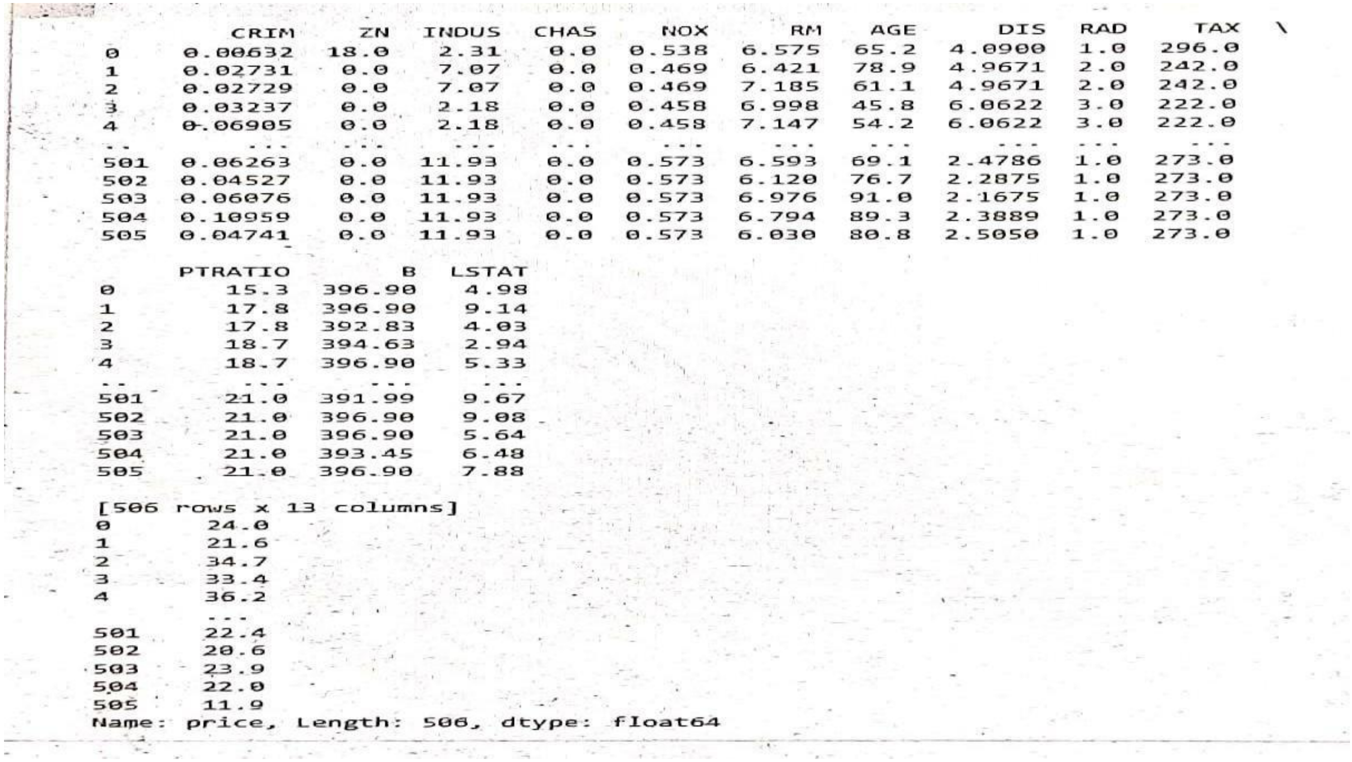
- A perfect negative correlation means the relationship that exists between two variables is exactly opposite all of the time.

- These are two types of correlation are represented numerically and as well as by shade of color in the heat map.

- HEATMAP – for better understanding of which place is best suited for individual personal preference based on given dataset. This uses correlation concept.



- Next we split our data into variables x and y , in order to train our model to predict data:

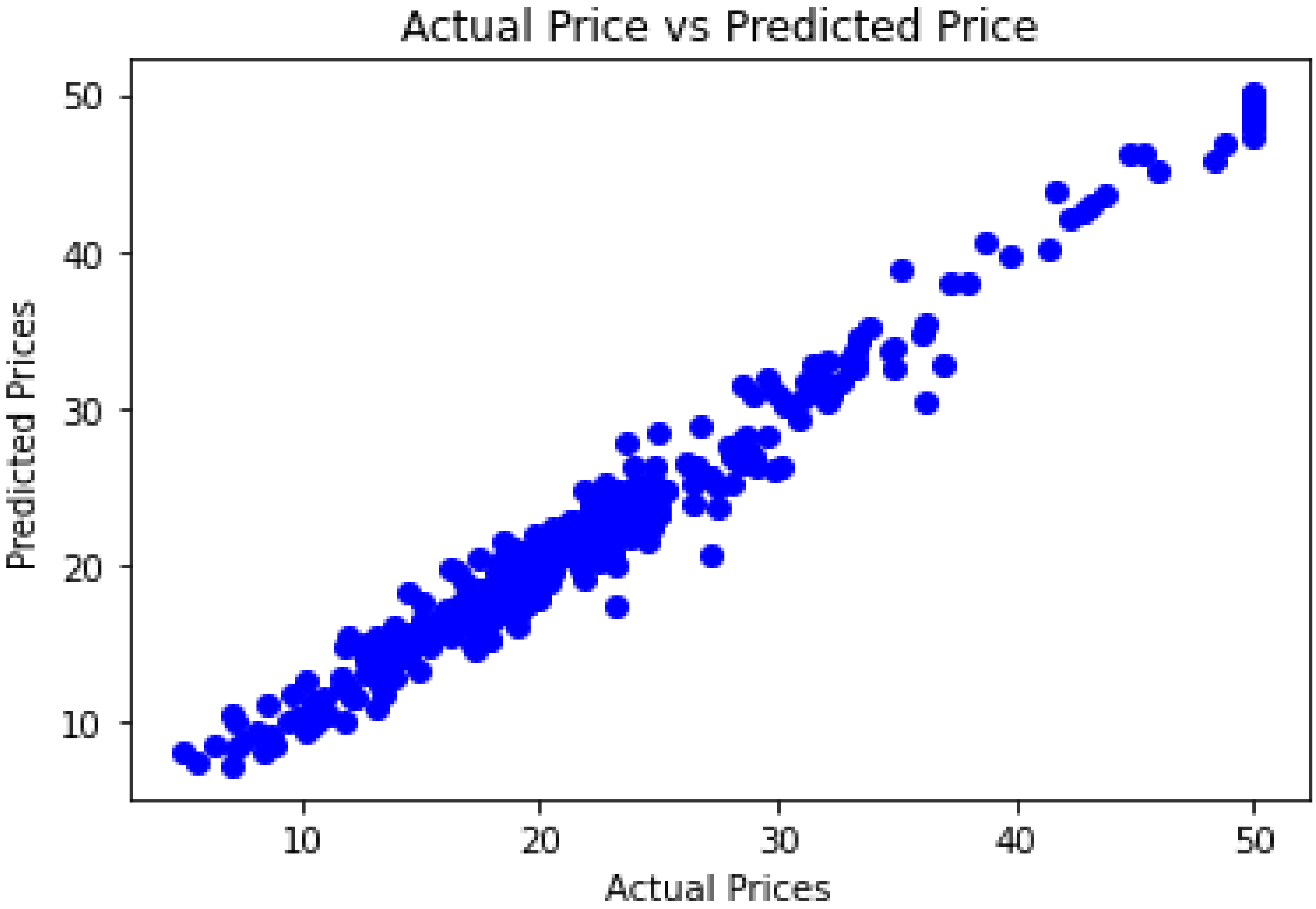


```
        CRIM    ZN  INDUS  CHAS    NOX     RM   AGE     DIS  RAD    TAX  \
0    0.00632  18.0   2.31   0.0  0.538  6.575  65.2  4.0900  1.0  296.0
1    0.02731   0.0   7.07   0.0  0.469  6.421  78.9  4.9671  2.0  242.0
2    0.02729   0.0   7.07   0.0  0.469  7.185  61.1  4.9671  2.0  242.0
3    0.03237   0.0   2.18   0.0  0.458  6.998  45.8  6.0622  3.0  222.0
4    0.06905   0.0   2.18   0.0  0.458  7.147  54.2  6.0622  3.0  222.0
..       ...   ...    ...   ...    ...    ...   ...     ...  ...    ...
501  0.06263   0.0  11.93   0.0  0.573  6.593  69.1  2.4786  1.0  273.0
502  0.04527   0.0  11.93   0.0  0.573  6.120  76.7  2.2875  1.0  273.0
503  0.06076   0.0  11.93   0.0  0.573  6.976  91.0  2.1675  1.0  273.0
504  0.10959   0.0  11.93   0.0  0.573  6.794  89.3  2.3889  1.0  273.0
505  0.04741   0.0  11.93   0.0  0.573  6.030  80.8  2.5050  1.0  273.0

     PTRATIO       B  LSTAT
0       15.3  396.90   4.98
1       17.8  396.90   9.14
2       17.8  392.83   4.03
3       18.7  394.63   2.94
4       18.7  396.90   5.33
..       ...     ...    ...
501     21.0  391.99   9.67
502     21.0  396.90   9.08
503     21.0  396.90   5.64
504     21.0  393.45   6.48
505     21.0  396.90   7.88

[506 rows x 13 columns]
0      24.0
1      21.6
2      34.7
3      33.4
4      36.2
       ...
501    22.4
502    20.6
503    23.9
504    22.0
505    11.9
Name: price, Length: 506, dtype: float64
```

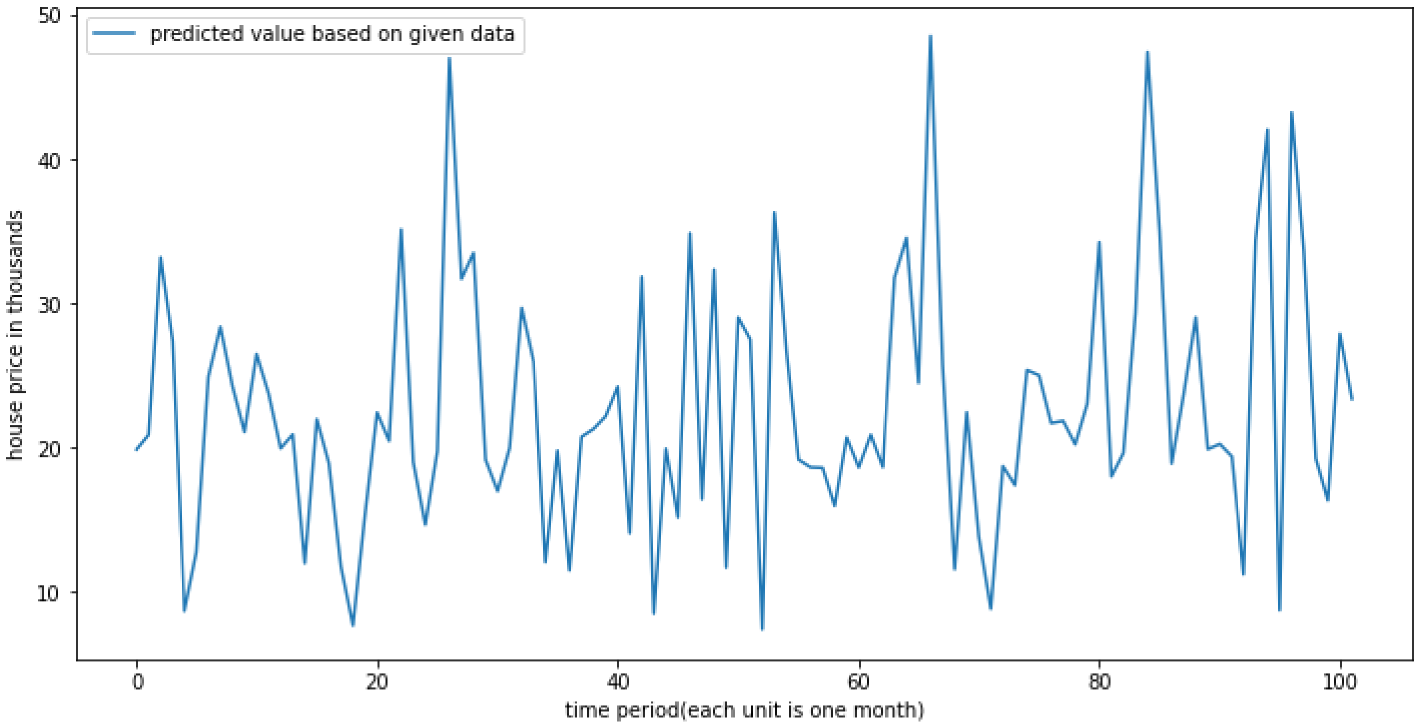
- Here the variable x contains the value of the first 13 columns i.e the parameters that are required for calculating and predicting the house prices. The variable y contains the 14th column values which are the house prices.
- First we predict the values in y using the values in x . Then we compare the actual prices and predicted prices by using scatter plot. Then we find the r square error and mean square error between them . If the errors is less enough then we proceed for testing of the model since the training phase is over. If the error is large , then we use optimizers like adam, and repeat drop and fitting process for a set number of epochs to reduce the error.
- The r square error or mean square error for good accuracy of the model in predicting the data is indicated numerically also.

- A model is good if these error values are less then 5.
- Then during testing process we predict the future house prices using present and past data parameters of houses in an location. Then we plot this graphically as a house price over time graph.
- For training the model , the error needs to be minimum for greater accuracy of model. The error between the actual and predicted price is plotted graphically using scatter plot. Here we can see that error is minimum sincethe data points of actual and predicted value are close to each other

# Graphical Analysis



# PREDICTED VALUE OF HOUSE PRICE BASED ON TEST SAMPLE DATA



THANK YOU