

TERRO’S REAL ESTATE

AGENCY

BUISNESS REPORT 2023 REAL ESTATE DATA ANALYSIS

**ANALYSED BY**

**DHAYANAND.S**

**REPORT ON -Finding out the most relevant features for pricing of a house:**

**TERMS FOR REFERENCE:**

Terro’s real-estate is an agency that estimates the pricing of houses in a certain locality. The pricing is concluded based on different features / factors of a property. This also helps them in identifying the business value of a property. To do this activity the company employs an “Auditor”, who studies various geographic features of a property like pollution level (NOX), crime rate, education facilities (pupil to teacher ratio), connectivity (distance from highway), etc. This helps in determining the price of a property. The agency has provided a dataset of 506 houses in Boston. Following are the details of the dataset.Data Dictionary:

|  |  |
| --- | --- |
| Attribute | Description |
| CRIME RATE | per capita crime rate by town |
| INDUSTRY | proportion of non-retail business acres per town (in percentage terms) |
| NOX | nitric oxides concentration (parts per 10 million) |
| AVG\_ROOM | average number of rooms per house |
| AGE | proportion of houses built prior to 1940 (in percentage terms) |
| DISTANCE | distance from highway (in miles) |
| TAX | full-value property-tax rate per $10,000 |
| PTRATIO | pupil-teacher ratio by town |
| LSTAT | % lower status of the population |
| AVG\_PRICE | Average value of houses in $1000's |

**Objective (Task):**

To analyse the magnitude of each variable to which it can affect the price of a house in a particular locality.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *DESCRIPTIVE STATISTICS* | *CRIME\_ RATE* | *AG*  *E* | *IND US* | *NOX* | *DISTA NCE* | *TAX* | *PTRA*  *TIO* | *AVG\_R OOM* | *LST AT* | *AVG\_P*  *RICE* |
|  |  |  |  |  |  |  |  |  |  |  |
| Mean | 4.87 | 68.57 | 11.14 | 0.55 | 9.55 | 408.2  4 | 18.46 | 6.28 | 12.65 | 22.53 |
| Standard Error | 0.13 | 1.25 | 0.30 | 0.01 | 0.39 | 7.49 | 0.10 | 0.03 | 0.32 | 0.41 |
| Median | 4.82 | 77.50 | 9.69 | 0.54 | 5.00 | 330.0  0 | 19.05 | 6.21 | 11.36 | 21.20 |
| Mode | 3.43 | 100.0  0 | 18.10 | 0.54 | 24.00 | 666.0  0 | 20.20 | 5.71 | 8.05 | 50.00 |
| Standard Deviation | 2.92 | 28.15 | 6.86 | 0.12 | 8.71 | 168.5  4 | 2.16 | 0.70 | 7.14 | 9.20 |
| Sample Variance | 8.53 | 792.3  6 | 47.06 | 0.01 | 75.82 | 28404  .76 | 4.69 | 0.49 | 50.99 | 84.59 |
| Kurtosis | -1.19 | -0.97 | -1.23 | -0.06 | -0.87 | -1.14 | -0.29 | 1.89 | 0.49 | 1.50 |
| Skewness | 0.02 | -0.60 | 0.30 | 0.73 | 1.00 | 0.67 | -0.80 | 0.40 | 0.91 | 1.11 |
| Range | 9.95 | 97.1 | 27.28 | 0.486 | 23 | 524 | 9.4 | 5.219 | 36.24 | 45 |
| Minimum | 0.04 | 2.9 | 0.46 | 0.385 | 1 | 187 | 12.6 | 3.561 | 1.73 | 5 |
| Maximum | 9.99 | 100 | 27.74 | 0.871 | 24 | 711 | 22 | 8.78 | 37.97 | 50 |
| Sum | 2465.22 | 3469  8.9 | 5635.  21 | 280.6  757 | 4832 | 20656  8 | 9338.5 | 3180.025 | 6402.  45 | 11401.6 |
| Count | 506 | 506 | 506 | 506 | 506 | 506 | 506 | 506 | 506 | 506 |

# Question-1: DESCRIPTIVE STATISTICS ANSWER

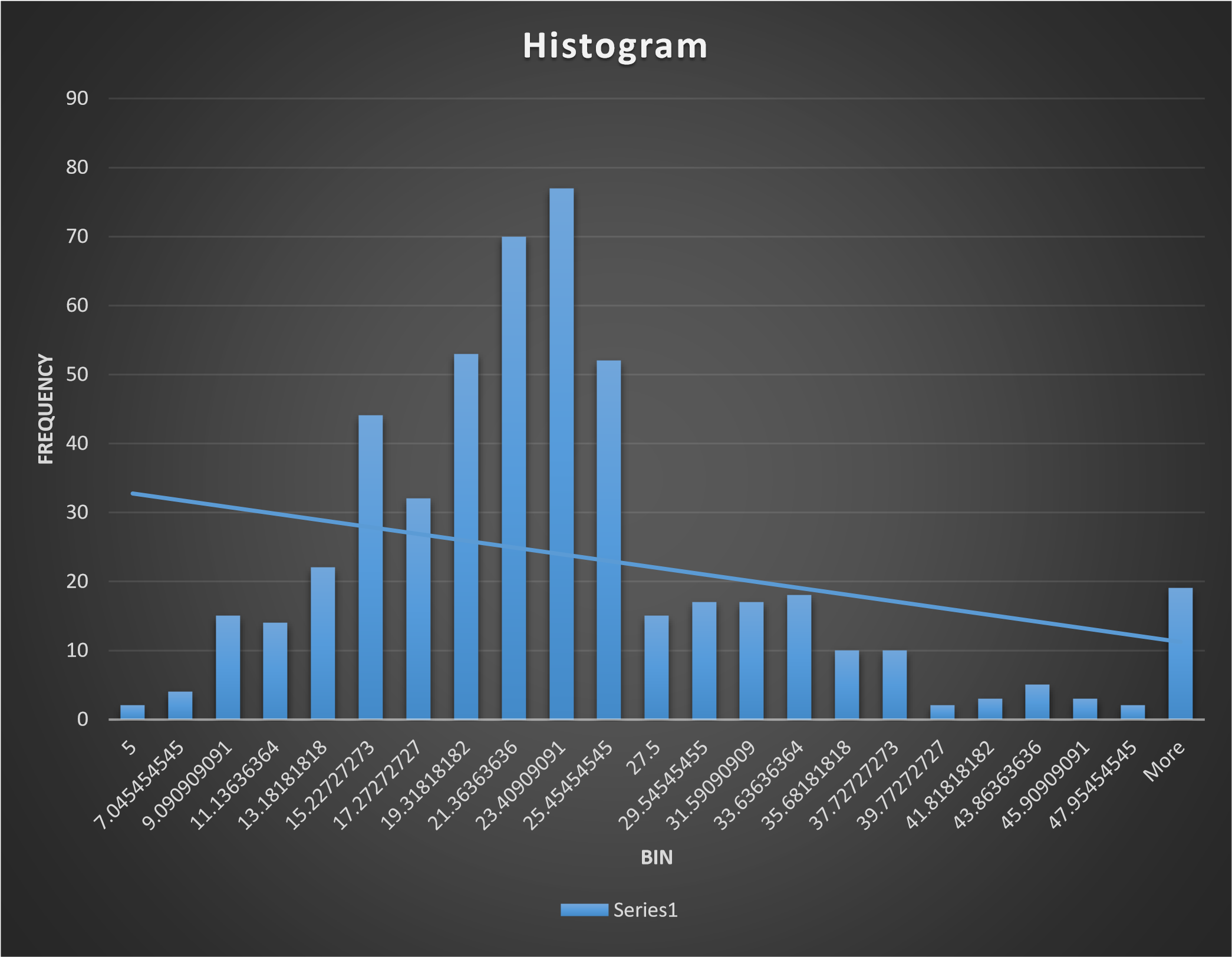
1. AGE and PTRATIO are negative skewed which show the MEAN<MEDIAN<MODE.
2. If the skewness is more than there will be more spread in the dataset.

3.AVG\_PRICE ,DISTANCE and LSTAT are the highest positively skewed in the dataset which shows that outliers are present in this variable.

4. The higher value of positive kurtosis of AVG\_ROOM>AVG\_PRICE , indicator that data has heavy tails or outliers.

# Question-2: HISTOGRAM-AVG\_PRICE

1. The histogram indicates that the AVG\_PRICE is positive skewed.
2. The average price of the house lies between $21$25k.



# Question-3: Covariance matrix

*CRIME\_ INDU DISTA PTRA AVG\_R AVG\_P*

*RATE AGE S NOX NCE TAX TIO OOM LSTAT RICE*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CRIME\_ RATE  AGE  INDUS  NOX  DISTAN  CE  TAX  PTRATIO  AVG\_RO OM  LSTAT  AVG\_PR  ICE | 8.51614  8  0.56291  5 |  | | | |  | | | | | | | | |
|  | 790.7  925  124.2  678  2.381  212  111.5 5 |  | |
| -  0.11022 |  |  | 46.97  143  0.605  874  35.47  971  831.7  133  5.680  855  -  1.884  23  29.52  181  -  30.46  05 |
| 0.00062  5 -  0.22986  -  8.22932  0.06816  9 |  | 0.013  401  0.615  71  13.02  05  0.047 304 |  | | | | | | | |
|  | 75.66 653 1333.  117  8.743  402  -  1.281  28  30.32  539  -  30.50  08 |  | | | | | |
| 2397.  942 | | | 28348  .62 |  | | | | |
|  | 15.90  543  -  4.742  54  120.8  384  -  97.39  62 | | 167.8  208  -  34.51  51  653.4  206 | 4.677  726  -  0.539  69  5.771  3  -  10.09  07 |  | | | |
|  |  | -  0.024  55 | |  | |  | |
| 0.05611  8 -  0.88268 |  | 0.49269  5 -  3.07365 |
|  | 0.487  98  -  0.454  51 | |  |  | 50.89  398  -  48.35 84.419  18 56 |
| 1.16201  2 | | -  724.8  2 | 4.48456  6 | | |

# OBSERVATION

1**.**If a covariance is positive it means both the variable tends to be high or low at the same time. Top 2 positive covariance:

* (Tax, age) o (Tax, Tax)

2**.**If a covariance is negative it means one variable is high and the other tend to low Top 2 negative covariance:

* (AVG\_ROOM, NOX) o (INDUS, CRIME\_RATE)

**QUESTION-4: CORRELATION MATRIX OF ALL VARIABLES:**

*CRIME\_RA AG IND NO DISTAN PTRATI AVG\_RO LSTA AVG\_PRI*

*TE E US X CE TAX O OM T CE*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CRIME\_RA  TE  AGE  INDUS  NOX  DISTANCE  TAX  PTRATIO  AVG\_ROO M  LSTAT  AVG\_PRIC  E | 1.00 |  | | |  |  |  |  | |  |
| 0.01 | 1.0 0 |  | |
| -0.01 | 0.6 4 | 1.00 |  |
| 0.00 | 0.7 3 | 0.76 | 1.0 0 |
| -0.01 | 0.4 6 | 0.60 | 0.6 1 | 1.00 |
| -0.02 | 0.5 1 | 0.72 | 0.6 7 | 0.91 | 1.0 0 |
| 0.01 | 0.2 6 | 0.38 | 0.1 9 | 0.46 | 0.4 6 | 1.00 |
| 0.03 | -  0.2  4 | -0.39 | -  0.3  0 | -0.21 | -  0.2  9 | -0.36 | 1.00 |  |
| -0.04 | 0.6 0 | 0.60 | 0.5 9 | 0.49 | 0.5 4 | 0.37 | -0.61 | 1.00 |
| 0.04 | -  0.3  8 | -0.48 | -  0.4  3 | -0.38 | -  0.4  7 | -0.51 | 0.70 | -  0.74 | 1.00 |

**OBSERVATION**

Top 3 positively correlated pairs (x, y moves in same direction)

* (TAX, DISTANCE)
* (NOX, INDUS)
* (NOX, AGE)

Top 3 negatively correlated pairs (x, y moves in opposite direction)

* (INDUS, Crime RATE)
* (DISTANCE, CRIME RATE)
* (TAX, CRIME RATE)

**QUESTION-5:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | *Regression Statistics* |  | | Multiple R | 0.737662726 | | R Square | 0.544146298 | | Adjusted R Square | 0.543241826 | | Standard Error | 6.215760405 | | Observations | 506 | | |  |  | | --- | --- | |  | *Coefficients* | | Intercept | 34.55384088 | | LSTAT | -0.950049354 | |

**OBSERVATION**

A. ADJUSTED R square value is 0.543241826 which indicates that this regression model is weak

B. Yes LSTAT VARIABLE is significant for the analysis in this module

**QUESTION-6:**

|  |  |
| --- | --- |
| *Regression Statistics* |  |
| Multiple R | 0.799100498 |
| R Square | 0.638561606 |
| Adjusted R Square | 0.637124475 |
| Standard Error | 5.540257367 |
| Observations | 506 |

|  |  |
| --- | --- |
|  | *Coefficients* |
| Intercept | -1.358272812 |
| AVG\_ROOM | 5.094787984 |
| LSTAT | -0.642358334 |

|  |  |
| --- | --- |
|  |  |

# OBSERVATION

**1.**Regression equation:

Y = (-1.358272812) + 5.094787984\*AVG\_ROOM+(0.642358334) \*LSTAT

2. Y = (-1.358272812) + 5.094787984\*7+(-0.642358334) \*20

|  |  |
| --- | --- |
| AVG PRICE OF HOUSE | $1,000.00 |
| AVG\_ROOM | 7 |
| LTSAT | 20 |
| PREDICTING AVG\_PRICE | $21,458.08 |

* The company has quoting a value of 30000 USD for this locality
* In this model, we used the regression equation to predict the AVG\_PRICE for 7 rooms (on an average) and has a value of 20 for L-STAT, the value is $21,458.08, which the company is overcharging**.**

**QUESTION-7:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Intercept | 29.24131526 |  | Adjusted R Square | 0.688298647 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | *variable* | *significant* | | Intercept | 2.53978E-09 | | AGE | 0.012670437 | | INDUS | 0.03912086 | | DISTANCE | 0.008293859 | | TAX | 0.000137546 | | NOX | 0.000251247 | | PTRATIO | 6.58642E-15 | | AVG\_ROOM | 3.89287E-19 | | LSTAT | 8.91071E-27 | | |  |  | | --- | --- | | variable | Insignificant | | CRIME\_RATE | 0.534657201 | |
| |  |  | | --- | --- | | *Regression Statistics* |  | | Multiple R | 0.832978824 | | R Square | 0.69385372 | | Adjusted R Square | 0.688298647 | | Standard Error | 5.1347635 | | Observations | 506 | |

|  |  |
| --- | --- |
|  | *Coefficients* |
| Intercept | 29.24131526 |
| CRIME\_RATE | 0.048725141 |
| AGE | 0.032770689 |
| INDUS | 0.130551399 |
| NOX | -10.3211828 |
| DISTANCE | 0.261093575 |
| TAX | -0.01440119 |
| PTRATIO | -1.074305348 |
| AVG\_ROOM | 4.125409152 |
| LSTAT | -0.603486589 |

# OBSERVATION

* In this model, adjusted R square is 68.82% which will give more impact to the AVG\_PRICE.

\*CRITERIA: In this model, the significance should be less than 0.05. And the insignificance should be greater than 0.05.

* Except CRIME\_RATE(INSIGNIFICANCE) , all other independent variable are significance.
* INDUS has more significance compared to other independent variables.
* LSTAT has less significance compared to other independent variables.
* In this model, AVG\_ROOM has higher coefficients (4.125) compared to other variables which means AVG\_ROOM has higher weightage for predicting AVG\_PRICE.

**QUESTION-8:**

|  |  |
| --- | --- |
| *Regression Statistics* |  |
| Multiple R | 0.832835773 |
| R Square | 0.693615426 |
| Adjusted R Square | 0.688683682 |
| Standard Error | 5.131591113 |
| Observations | 506 |

# B) OBSERVATION

In this model, Adjusted R square value is 68.86 %which gives more impact in AVG\_PRICE predicting.

* In this model, R square value is 68.86% which gives more impact to AVG\_PRICE compared to previous model, because in previous model more variables are taken for predicting the value of AVG\_PRICE, but in this model less variable are taken.
* This model has better performance compared to previous model

|  |  |
| --- | --- |
| ***variable*** | ***Coefficients*** |
| NOX | -10.27270508 |
| PTRATIO | -1.071702473 |
| LSTAT | -0.605159282 |
| TAX | -0.014452345 |
| AGE | 0.03293496 |
| INDUS | 0.130710007 |
| DISTANCE | 0.261506423 |
| AVG\_ROOM | 4.125468959 |
| Intercept | 29.42847349 |
|  |  |

 The Coefficient of NOX value is negative, which tends to decrease in AVG\_PRICE

D)REGRESSION EQUATION

Y =INTERCEPT + 0.03\*AGE+0.13\*INDUS-10\*NOX+0.26\*DISTANCE \*0.014\*TAX+1.07\*PTRATIO+4.12\*AVG\_ROOM-0.6\*LSTAT

**RECOMMENDATION:**

* The most relevant price for the house should be lies between $21- 25k.
* AVG\_PRICE has high positive correlation with the

AVG\_ROOM. so if the AVG\_ROOM increases AVG\_PRICE will increase.

* LSTAT has low negative correlation with AVG\_PRICE.
* So, if LSTAT decreases the AVG\_PRICE will increase.
* For the relevant price of house (AVG\_PRICE) The AVG\_ROOM should be increases and LSTAT should be decreases.
* REGRESSION EQUATION: Y = (-1.358272812) +

5.094787984\*AVG\_ROOM+(-0.642358334) \*LSTAT

* This model has adjusted R square 68.82% with less variables which is more compared to other model in terms of predicting price of house
* In this model, we used the regression equation to predict the AVG\_PRICE for 7 rooms (on an average) and has a value of 20 for L-STAT, the value of house is $21,458.08