**Assignment – Terro’s real estate agency**

**Problem Statement (Situation):**

“Finding out the most relevant features for pricing of a house”

Assignment – Terro’s real estate agency Real estate data analysis – Exploratory data analysis, Linear Regression Problem Statement (Situation): “Finding out the most relevant features for pricing of a house” 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.

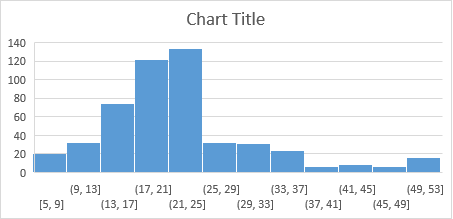
**Question Number 1:**

Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | *AGE* |  | | *INDUS* | |  | | *NOX* |  | | | |  |  | |  | |  | |  |  | | | | Mean | 68.5749 | | Mean | | 11.136779 | | Mean | 0.5546951 | | | | Standard Error | 1.25137 | | Standard Error | | 0.3049799 | | Standard Error | 0.0051514 | | | | Median | 77.5 | | Median | | 9.69 | | Median | 0.538 | | | | Mode | 100 | | Mode | | 18.1 | | Mode | 0.538 | | | | Standard Deviation | 28.14886 | | Standard Deviation | | 6.8603529 | | Standard Deviation | 0.1158777 | | | | Sample Variance | 792.3584 | | Sample Variance | | 47.064442 | | Sample Variance | 0.0134276 | | | | Kurtosis | -0.96772 | | Kurtosis | | -1.23354 | | Kurtosis | -0.064667 | | | | Skewness | -0.59896 | | Skewness | | 0.2950216 | | Skewness | 0.7293079 | | | | Range | 97.1 | | Range | | 27.28 | | Range | 0.486 | | | | Minimum | 2.9 | | Minimum | | 0.46 | | Minimum | 0.385 | | | | Maximum | 100 | | Maximum | | 27.74 | | Maximum | 0.871 | | | | Sum | 34698.9 | | Sum | | 5635.21 | | Sum | 280.6757 | | | | Count | 506 | | Count | | 506 | | Count | 506 | | | |  |  | |  | |  | |  |  | | | | *DISTANCE* | |  | | *TAX* | |  | *PTRATIO* | |  | |  | |  | |  | |  |  | |  | | Mean | | 9.549407 | | Mean | | 408.2372 | Mean | | 18.45553 | | Standard Error | | 0.387085 | | Standard Error | | 7.492389 | Standard Error | | 0.096244 | | Median | | 5 | | Median | | 330 | Median | | 19.05 | | Mode | | 24 | | Mode | | 666 | Mode | | 20.2 | | Standard Deviation | | 8.707259 | | Standard Deviation | | 168.5371 | Standard Deviation | | 2.164946 | | Sample Variance | | 75.81637 | | Sample Variance | | 28404.76 | Sample Variance | | 4.686989 | | Kurtosis | | -0.86723 | | Kurtosis | | -1.14241 | Kurtosis | | -0.28509 | | Skewness | | 1.004815 | | Skewness | | 0.669956 | Skewness | | -0.80232 | | Range | | 23 | | Range | | 524 | Range | | 9.4 | | Minimum | | 1 | | Minimum | | 187 | Minimum | | 12.6 | | Maximum | | 24 | | Maximum | | 711 | Maximum | | 22 | | Sum | | 4832 | | Sum | | 206568 | Sum | | 9338.5 | | Count | | 506 | | Count | | 506 | Count | | 506 | |  |  |  |  |  |  |  |  |
|  | |  |  |  |  |  |  |  |
|  | | | | | | | | |
|  | | | | | | |  |  |
|  |  |  |  |  |  |  |  |  |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | *AVG\_ROOM* |  | *LSTAT* |  | *AVG\_PRICE* |  | |  |  |  |  |  |  | | Mean | 6.284634 | Mean | 12.65306 | Mean | 22.53281 | | Standard Error | 0.031235 | Standard Error | 0.317459 | Standard Error | 0.408861 | | Median | 6.2085 | Median | 11.36 | Median | 21.2 | | Mode | 5.713 | Mode | 8.05 | Mode | 50 | | Standard Deviation | 0.702617 | Standard Deviation | 7.141062 | Standard Deviation | 9.197104 | | Sample Variance | 0.493671 | Sample Variance | 50.99476 | Sample Variance | 84.58672 | | Kurtosis | 1.8915 | Kurtosis | 0.49324 | Kurtosis | 1.495197 | | Skewness | 0.403612 | Skewness | 0.90646 | Skewness | 1.108098 | | Range | 5.219 | Range | 36.24 | Range | 45 | | Minimum | 3.561 | Minimum | 1.73 | Minimum | 5 | | Maximum | 8.78 | Maximum | 37.97 | Maximum | 50 | | Sum | 3180.025 | Sum | 6402.45 | Sum | 11401.6 | | Count | 506 | Count | 506 | Count | 506 |  |  |  | | --- | --- | | **Observations:** |  | | |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | 1.The total count of values is 506 |  |  |  |  |  |  |  | | 2. Tax and Distance median is less than mean meaning that most the values falling below 50% are lower values  and has a very high values after 50% | | | | | | | | | 3.In Age category median is high compared to mean meaning 50% of the values after 50% will be very high | | | | | |  |  | | | |  | | |  |  |  |  |  |  |  |  |

**Question Number 2:**

Plot a histogram of the Avg\_Price variable. What do you infer?



|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | 1. From the chart we observe that most of the average houses lies in between price value of 21-25k and the second most between is between 17-21k price. | | | | | | 1. We can clearly state that the average price is up to 50k but most of the houses are lying below the average of 25k. |  |  |  |  | | | | |
| **Question Number 3:**  Compute the covariance matrix. Share your observations.   |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  | *CRIME\_RATE* | *AGE* | *INDUS* | *NOX* | *DISTANCE* | *TAX* | *PTRATIO* | *AVG\_ROOM* | *LSTAT* | *AVG\_PRICE* | | CRIME\_RATE | 8.516148 |  |  |  |  |  |  |  |  |  | | AGE | 0.562915 | 790.7925 |  |  |  |  |  |  |  |  | | INDUS | -0.110215 | 124.2678 | 46.97143 |  |  |  |  |  |  |  | | NOX | 0.000625 | 2.381212 | 0.605874 | 0.013401 |  |  |  |  |  |  | | DISTANCE | -0.22986 | 111.55 | 35.47971 | 0.61571 | 75.66653 |  |  |  |  |  | | TAX | -8.229322 | 2397.942 | 831.7133 | 13.0205 | 1333.117 | 28348.62 |  |  |  |  | | PTRATIO | 0.068169 | 15.90543 | 5.680855 | 0.047304 | 8.743402 | 167.8208 | 4.67773 |  |  |  | | AVG\_ROOM | 0.056118 | -4.742538 | -1.88423 | -0.024555 | -1.28128 | -34.5151 | -0.53969 | 0.492695 |  |  | | LSTAT | -0.88268 | 120.8384 | 29.52181 | 0.48798 | 30.32539 | 653.4206 | 5.7713 | -3.073655 | 50.894 |  | | AVG\_PRICE | 1.162012 | -97.39615 | -30.4605 | -0.454512 | -30.5008 | -724.8204 | -10.0907 | 4.484566 | -48.352 | 84.41956 |   Observation-  We can clearly state that tax has a wide spread of variables when compared with other features as we can see from the descriptive stats that it has a range of 524 and std dev of 168.53 meaning 67% of the data includes too many variables | | | |
|  |  |  |  |
| **Question Number 4:**  Create a correlation matrix of all the variables (Use Data analysis tool pack).  a) Which are the top 3 positively correlated pairs and  b) Which are the top 3 negatively correlated pairs.   |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  | *CRIME\_RATE* | *AGE* | *INDUS* | *NOX* | *DISTANCE* | *TAX* | *PTRATIO* | *AVG\_ROOM* | *LSTAT* | *AVG\_PRICE* | | CRIME\_RATE | 1 |  |  |  |  |  |  |  |  |  | | AGE | 0.0069 | 1 |  |  |  |  |  |  |  |  | | INDUS | -0.006 | 0.6448 | 1 |  |  |  |  |  |  |  | | NOX | 0.0019 | 0.7315 | 0.764 | 1 |  |  |  |  |  |  | | DISTANCE | -0.009 | 0.456 | 0.595 | 0.6114 | 1 |  |  |  |  |  | | TAX | -0.017 | 0.5065 | 0.721 | 0.668 | 0.9102 | 1 |  |  |  |  | | PTRATIO | 0.0108 | 0.2615 | 0.383 | 0.1889 | 0.4647 | 0.461 | 1 |  |  |  | | AVG\_ROOM | 0.0274 | -0.24 | -0.392 | -0.302 | -0.2098 | -0.292 | -0.356 | 1 |  |  | | LSTAT | -0.042 | 0.6023 | 0.604 | 0.5909 | 0.4887 | 0.544 | 0.374 | -0.614 | 1 |  | | AVG\_PRICE | 0.0433 | -0.377 | -0.484 | -0.427 | -0.3816 | -0.469 | -0.508 | 0.6954 | -0.738 | 1 | |  |  |  |
|  | | | |
| |  |  | | --- | --- | | Top 3 positively correlated : | | | 1. Tax - Distance | | | 2. Nox - Indus |  | | 3. Nox - Age   |  | | --- | | Top 3 negatively correlated: | | 1. Avg\_price – LSTAT | | 2. LSTAT - Avg\_room | | 3. Avg\_price - PTRATIO | |  | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Question Number 5:**  Build an initial regression model with AVG\_PRICE as ‘y’ (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot.  a) What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?  b) Is LSTAT variable significant for the analysis based on your model?   |  |  | | --- | --- | | *Regression Statistics* | | | Multiple R | 0.737662726 | | R Square | 0.544146298 | | Adjusted R Square | 0.543241826 | | Standard Error | 6.215760405 | | Observations | 506 |  |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | | Intercept | 34.55384088 | 0.562627355 | 61.41514552 | 3.7431E-236 | | LSTAT | -0.950049354 | 0.038733416 | -24.52789985 | 5.0811E-88 |   **Observation-**  a)  1.The R square here is 0.544 meaning nearly 54.4% of the Avg\_price is explained by LSTAT.  2. With the help of graph we can observe that LSTAT is negatively correlated to Avg\_price.  3. P-value of the variable LSTAT is below 0.05 which means we are able to reject the null hypothesis and allowed for further analysis.  4. The range doesn’t include zero in between them.  5. Intercept is 34.55 and -0.95 is coefficient of LSTAT meaning higher the LSTAT lower will be the Y.  b)  1. LSTAT is explaining a large amount of Avg\_Price but some other variables might also explain at significant value.  2. As per the observation of correlation few other variables were also correlated to Avg\_Price.  **Question Number 6:**  Build a new Regression model including LSTAT and AVG\_ROOM together as Independent variables and AVG\_PRICE as dependent variable.  a) Write the Regression equation. If a new house in this locality has 7 rooms (on an average) and has a value of 20 for L-STAT, then what will be the value of AVG\_PRICE? How does it compare to the company quoting a value of 30000 USD for this locality? Is the company Overcharging/ Undercharging?  b) Is the performance of this model better than the previous model you built in Question 5? Compare in terms of adjusted R-square and explain.   |  |  | | --- | --- | | *Regression Statistics* | | | Multiple R | 0.799100498 | | R Square | 0.638561606 | | Adjusted R Square | 0.637124475 | | Standard Error | 5.540257367 | | Observations | 506 |  |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | | Intercept | -1.358272812 | 3.17282778 | -0.428095348 | 0.668764941 | | AVG\_ROOM | 5.094787984 | 0.4444655 | 11.46272991 | 3.47226E-27 | | LSTAT | -0.642358334 | 0.043731465 | -14.68869925 | 6.66937E-41 |   **Observation:**  a)   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | Y= Avg\_Price | |  |  |  |  |  | | X1=LSTAT=20 | |  |  |  |  |  | | X2=Avg\_Room =7 | |  |  |  |  |  | |  |  |  |  |  |  |  | | Y= (-1.358) +(-0.6423)X1+(5.094)X2 | | | |  |  |  | | Y= (-1.358)+(-0.6423)\*20+(5.094)\*7 | | | |  |  |  | | Y=22.812 |  |  |  |  |  |  | | We have average value of 22.812k for above given variable value | | | | | | | | with this model | |  |  |  |  |  |   b)  1. In the previous answer we got 54% correlation but the R value has significantly increased as another independent variable Avg\_Room is provided.  2. Fischers value is lower that before and both the variables have P-values less than 5%, so we can reject null hypothesis.  3. Avg\_Room is positively correlated and LSTAT is negatively correlated meaning decrease in LSTAT and increase in Avg\_Room values will increase the Avg\_Price.  **Question Number 7:**  Build another Regression model with all variables where AVG\_PRICE alone be the Dependent Variable and all the other variables are independent. Interpret the output in terms of adjusted R\_square, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG\_PRICE.   |  |  | | --- | --- | | *Regression Statistics* | | | Multiple R | 0.832978824 | | R Square | 0.69385372 | | Adjusted R Square | 0.688298647 | | Standard Error | 5.1347635 | | Observations | 506 |  |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | | Intercept | 29.24131526 | 4.817125596 | 6.070282926 | 2.53978E-09 | | CRIME\_RATE | 0.048725141 | 0.078418647 | 0.621346369 | 0.534657201 | | AGE | 0.032770689 | 0.013097814 | 2.501996817 | 0.012670437 | | INDUS | 0.130551399 | 0.063117334 | 2.068392165 | 0.03912086 | | NOX | -10.3211828 | 3.894036256 | -2.650510195 | 0.008293859 | | DISTANCE | 0.261093575 | 0.067947067 | 3.842602576 | 0.000137546 | | TAX | -0.01440119 | 0.003905158 | -3.687736063 | 0.000251247 | | PTRATIO | -1.074305348 | 0.133601722 | -8.041104061 | 6.58642E-15 | | AVG\_ROOM | 4.125409152 | 0.442758999 | 9.317504929 | 3.89287E-19 | | LSTAT | -0.603486589 | 0.053081161 | -11.36912937 | 8.91071E-27 |   **Observation:**  1. We can clearly state that Crime\_Rate has P-value more than 5% so it is not related.  2. Age, Indus, Nox, Distance , Tax are all less than 5% and can reject null hypothesis but not with a significant value.  3. PTRATIO is very low value when compared with other variables.  4. PTRATIO is the one can be another independent variable which can be significant in explaining in Avg\_Price.  **Question 8:**  Pick out only the significant variables from the previous question. Make another instance of the Regression model using only the significant variables you just picked and answer the questions below: (8 marks)  a) Interpret the output of this model.  b) Compare the adjusted R-square value of this model with the model in the previous question, which model performs better according to the value of adjusted R-square?  c) Sort the values of the Coefficients in ascending order. What will happen to the average price if the value of NOX is more in a locality in this town?  d) Write the regression equation from this model.   |  |  | | --- | --- | | *Regression Statistics* | | | Multiple R | 0.823786477 | | R Square | 0.67862416 | | Adjusted R Square | 0.676703587 | | Standard Error | 5.229396169 | | Observations | 506 |  |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | | Intercept | 18.56711151 | 3.913201635 | 4.744736724 | 2.72581E-06 | | PTRATIO | -0.930722555 | 0.117653724 | -7.910693569 | 1.64466E-14 | | AVG\_ROOM | 4.515420944 | 0.425871541 | 10.60277692 | 7.73479E-24 | | LSTAT | -0.571805688 | 0.042230224 | -13.54020033 | 7.94421E-36 |   **Observation:**  a) The R square value is significantly better with PTRATIO combined as independent variable  b) The R square value compared to the previous que is nearly the same with only three variables compared to all the nine variables  c) There is no certain correlation in between Avg\_Price and Nox values. So the Avg\_Price cannot be explained with increase or decrease in Nox value  d) The regression model is:  Y=18.567+(-0.9307)\*X1+(4.5154)\*X2+(-0.5718)\*X3  Y=Avg\_price  X1=PTRATIO  X2=Avg\_Room  X3=LSTAT |