Terros Real Estate Agency

Mudassir Ahmed

BUSINESS REPORT

CONTENTS

PROBLEM STATEMENT

"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.

ABOUT 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)
AVG_ROOM	average number of rooms per house
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
PTRATIO	pupil-teacher ratio by town
AGE	proportion of houses built prior to 1940 (in percentage terms)
DISTANCE	distance from highway (in miles)
NOX	nitric oxides concentration (parts per 10 million)
AGE	proportion of houses built prior to 1940 (in percentage terms)

Q1)Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation.

	CRIME_RATE
Mean	4.871976285
Standard Error	0.129860152
Median	4.82
Mode	3.43
Standard Deviation	2.921131892
Sample Variance	8.533011532
Kurtosis	-1.189122464
Skewness	0.021728079
Range	9.95
Minimum	0.04
Maximum	9.99
Sum	2465.22
Count	506

1)The average per capita crime rate is 4.87 in boston in 506 houses
2)Since the Standard error is 0.129 we can say that there is not much variation in crime rate in the neighbourhood i.e it is close to 4.87
3)Half of the neighbourhood have crime rate below 4.82 and the other half above 4.87.
4)The most common crime rate is 3.43(The Frequency of crime rates is around 3.43)
5)Standard deviation of 2.92 indicates a moderate amount of variability in crime rates across the dataset.
6)The distribution is approximately symmetrical with a slight right skew
7)The kurtosis is negative (-1.19), indicating that the distribution has lighter tails and is less peaked than a normal distribution.

	AGE
Mean	68.57490119
Standard Error	1.251369525
Median	77.5
Mode	100
Standard Deviation	28.14886141
Sample Variance	792.3583985
Kurtosis	-0.967715594
Skewness	-0.59896264
Range	97.1
Minimum	2.9
Maximum	100
Sum	34698.9
Count	506

1)The Average Age is 68.5 that
built before 1940 in Boston
the age as the **standard error**

below and above 77.5
before 1940 in most of the

indicates **a moderate amount**
across the dataset.
suggests a concentration of
towards lower ages.
indicates that the distribution
tails compared to a normal

of age percentages, spanning

means 68.57% of the houses were
2)There is **not much Variation** in
is around **1.25**
3)Half of the People have the **age**
4**) 100% of the houses were built**
localities.
5)The **Standard Deviation of 28.15**
of variability in age percentages
6)The negative **skewness (-0.60**)
higher ages with a tail extending
7)The negative kurtosis (-0.97)
has less extreme values and lighter
distribution
8)The dataset covers a wide **range**
from **2.9% to 100%.**
9)The age distribution is diverse, with a concentration towards higher percentages, especially around the mode of 100

	INDUS		*NOX*		*DISTANCE*
Mean	11.13677866	Mean	0.554695059	Mean	9.549407115
Standard Error	0.304979888	Standard Error	0.005151391	Standard Error	0.387084894
Median	9.69	Median	0.538	Median	5
Mode	18.1	Mode	0.538	Mode	24
Standard Deviation	6.860352941	Standard Deviation	0.115877676	Standard Deviation	8.707259384
Sample Variance	47.06444247	Sample Variance	0.013427636	Sample Variance	75.81636598
Kurtosis	-1.233539601	Kurtosis	-0.064667133	Kurtosis	-0.867231994
Skewness	0.295021568	Skewness	0.729307923	Skewness	1.004814648
Range	27.28	Range	0.486	Range	23
Minimum	0.46	Minimum	0.385	Minimum	1
Maximum	27.74	Maximum	0.871	Maximum	24
Sum	5635.21	Sum	280.6757	Sum	4832
Count	506	Count	506	Count	506

INDUS

- The mean represents the average proportion of non-retail business acres per town in percentage terms.
- The median is 9.69%, suggesting that about half of the towns have value below this, and half have a value above.
- The mode of 18.1% indicates a concentration of towns with this specific percentage of non-retail business acres.
- The standard deviation of 6.86% indicates a moderate amount of variability in the percentage of non-retail business acres across towns.
- The range of 27.28% shows the spread from the minimum to the maximum value.
- The distribution characteristics suggest a moderate right skewness and a relatively flat distribution

NOX

- The mean represents the average nitric oxides concentration (parts per 10 million) in the dataset.
- The median is 0.538, suggesting that approximately half of the observations have a nitric oxides concentration below this value.
- The mode is 0.538, indicating a concentration of observations with this specific nitric oxides concentration

- The concentration of nitric oxides varies moderately across the dataset, with a range of 0.486.
- The distribution is slightly skewed to the right, indicating a concentration of lower nitric oxides concentrations. It has lighter tails and is less peaked.
- The nitric oxides concentration ranges from 0.385 to 0.871, capturing a diverse set of values.

DISTANCE

The **DISTANCE** variable in the dataset represents the distance from the highway in miles.

- The average distance is approximately 9.55 miles, with a median of 5 miles, and a mode of 24 miles, indicating a notable concentration of observations at the maximum distance.
- The distances exhibit a considerable variability, ranging from 1 to 24 miles, with a standard deviation of 8.71 miles.
- The distribution is positively skewed (skewness = 1.0048), suggesting a concentration of towns with shorter distances from the highway.
- The kurtosis value of -0.8672 indicates a distribution with slightly lighter tails than a normal distribution.

TAX		*PTRATIO*		*AVG_ROOM*	
Mean	408.2371542	Mean	18.4555336	Mean	6.284634387
Standard Error	7.492388692	Standard Error	0.096243568	Standard Error	0.031235142
Median	330	Median	19.05	Median	6.2085
Mode	666	Mode	20.2	Mode	5.713
Standard Deviation	168.5371161	Standard Deviation	2.164945524	Standard Deviation	0.702617143
Sample Variance	28404.75949	Sample Variance	4.686989121	Sample Variance	0.49367085
Kurtosis	-1.142407992	Kurtosis	-0.285091383	Kurtosis	1.891500366
Skewness	0.669955942	Skewness	-0.802324927	Skewness	0.403612133
Range	524	Range	9.4	Range	5.219
Minimum	187	Minimum	12.6	Minimum	3.561
Maximum	711	Maximum	22	Maximum	8.78
Sum	206568	Sum	9338.5	Sum	3180.025
Count	506	Count	506	Count	506

TAX

- The typical property-tax rate per $10,000 is around 408, with many towns clustering around this average. The most common tax rate observed is 666.
- Tax rates vary widely, spanning from 187 to 711. This means some towns have lower property taxes, while others have higher ones
- Most towns tend to have lower tax rates, as seen in the positive skewness (0.67). The distribution has a slight spread, with a tail on the right
- The dataset includes towns with tax rates as low as 187 and as high as 711. The most frequent tax rate is 666.

PTRATIO

- The average student-to-teacher ratio in these towns is about 18.46, but the actual ratios vary. Most towns have ratios around 20.2, but there are some with lower or higher ratios.
- Pupil-teacher ratios vary, ranging from a minimum of 12.6 to a maximum of 22, with a standard deviation of 2.16. This indicates diversity in the sizes of classes across different towns.
- The distribution of pupil-teacher ratios is slightly negatively skewed (-0.80), suggesting a tendency for higher ratios. The kurtosis of -0.29 indicates a distribution with less extreme values and lighter tails compared to a normal distribution

AVG_ROOM

- The average number of rooms in houses is approximately 6.28, but the actual counts vary. Most houses have around 5.713 rooms, but there are some with fewer or more rooms.
- The number of rooms in houses varies, ranging from approximately 3.561 to 8.78. This shows diversity in house sizes across different towns.
- Houses generally tend to have more rooms, as seen in the slightly positive skewness (0.40).
- The distribution has some houses with more extreme room counts, indicated by the kurtosis of 1.89.
- The most common number of rooms is around 5.713, and the median (middle value) is 6.2085. This suggests that many houses have similar room counts.

LSTAT		*AVG_PRICE*	
Mean	12.65306324	Mean	22.53280632
Standard Error	0.317458906	Standard Error	0.408861147
Median	11.36	Median	21.2
Mode	8.05	Mode	50
Standard Deviation	7.141061511	Standard Deviation	9.197104087
Sample Variance	50.99475951	Sample Variance	84.58672359
Kurtosis	0.493239517	Kurtosis	1.495196944
Skewness	0.906460094	Skewness	1.108098408
Range	36.24	Range	45
Minimum	1.73	Minimum	5
Maximum	37.97	Maximum	50
Sum	6402.45	Sum	11401.6
Count	506	Count	506

LSTAT

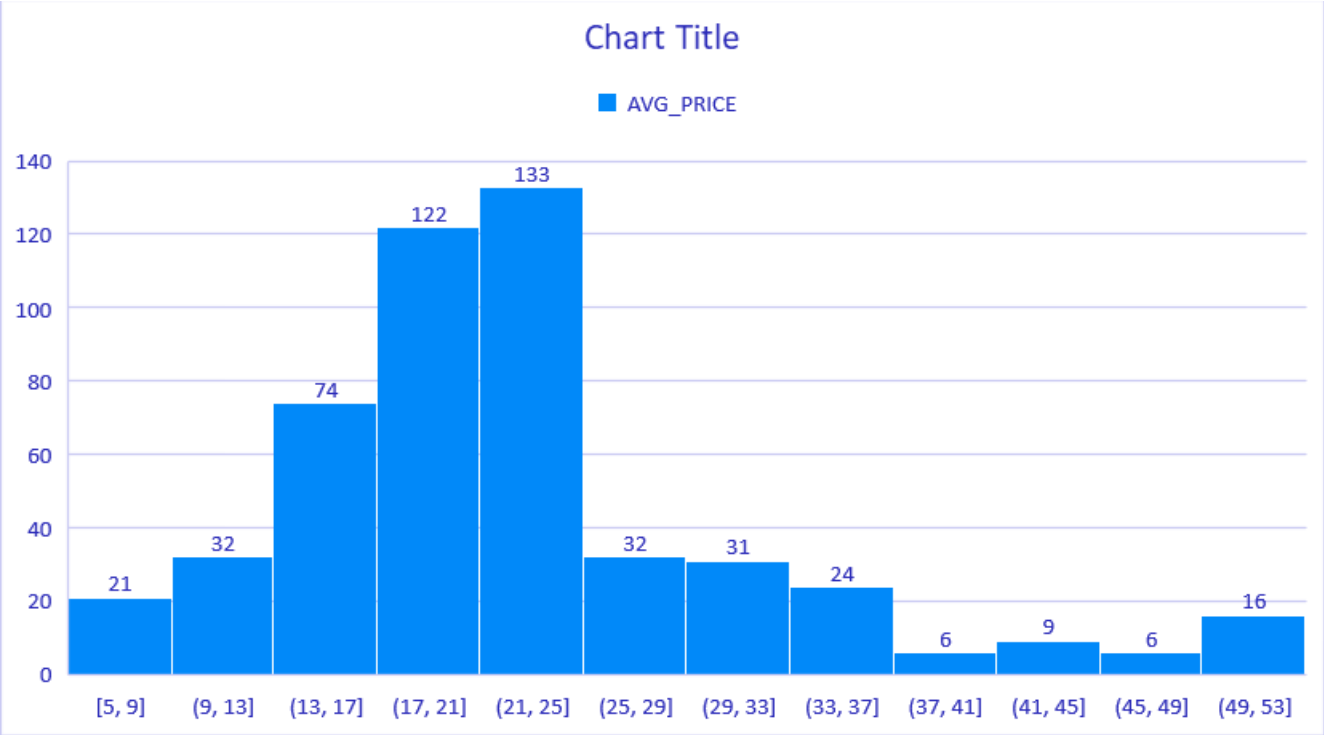
- On An average, about 12.65% of the population in these towns has a lower socioeconomic status.
- The percentage of lower status varies, ranging from a minimum of 1.73% to a maximum of 37.97%, with a standard deviation of 7.14
- The distribution of status percentages is positively skewed (0.91), suggesting a concentration of towns with lower socioeconomic status. The kurtosis of 0.49 indicates a distribution with moderately heavy tails.

- The most frequently occurring status percentage is around 8.05%, and the median (middle value) is 11.36%. This indicates a commonality in these values across the dataset.

AVG_PRICE

- The average house price is around $22,533, but the actual prices vary widely.
- Most towns have houses priced around $50,000, indicating a concentration of lower-priced homes. However, there is also diversity in housing costs, with some towns having higher-priced properties
- The distribution of house prices is positively skewed (1.11), suggesting a concentration of towns with lower-priced houses.
- The kurtosis of 1.50 indicates a distribution with moderately heavy tails.

2) Plot a histogram of the Avg_Price variable. What do you infer?



1) Most of the Houses In Boston Have The Average Price Between $21,000-$25000
2) 50% of Family that lives in Boston have house value under 25k and rest have value above 25k.
3) The Least Count Of the house is between The Price Range of $37000-$41000 &$45000-$49000

By observing the data in the histogram we can say that the data is more spread towards the left side of the histogram i.e, having a long left tail, thus we can say that the data is positively skewed.

3) Compute the covariance matrix. Share your observations.

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.516147873									
AGE	0.562915215	790.7924728								
INDUS	-0.110215175	124.2678282	46.97142974							
NOX	0.000625308	2.381211931	0.605873943	0.013401099						
DISTANCE	-0.229860488	111.5499555	35.47971449	0.615710224	75.66653127					
TAX	-8.229322439	2397.941723	831.7133331	13.02050236	1333.116741	28348.6236				
PTRATIO	0.068168906	15.90542545	5.680854782	0.047303654	8.74340249	167.8208221	4.677726296			
AVG_ROOM	0.056117778	-4.74253803	-1.88425427	-0.02454826	-1.28277391	-34.51510104	-0.53994518	0.492695216		
LSTAT	-0.882680362	120.8384405	29.52181125	0.487979871	30.32539213	653.4206174	5.771300243	-3.07354967	50.89397935	
AVG_PRICE	1.162021224	-97.39615288	-30.46050499	-0.454512407	-30.50083035	-724.8204284	-10.09067561	4.484565552	-48.35179219	84.41955616

1)There Is Highest positive covariance between The TAX and AGE.This Indicates that,If the AGE of the property(in percentage) Increases then the TAX increases.
2)There Is Negative Covariance Between AVG_PRICE and TAX.If the AVG_PRICE Decreases Then the TAX Increases.

- **Age,Indus,Distance vs Tax** -These data sets have more covariance thus we can say that they have a direct relationship to each other
- **Tax,Age, Lstat vs Avg Price**-These data sets have the least covariance thus they have an inverse relationship to each other,if one increases other data decreases.

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.

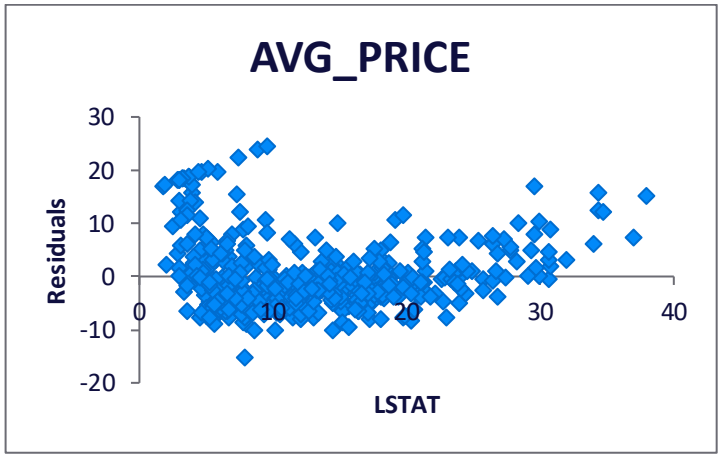
	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	1									
AGE	0.006859463	1								
INDUS	-0.005510651	0.644778511	1							
NOX	0.001850982	0.731470104	0.763651447	1						
DISTANCE	-0.009055049	0.456022452	0.595129275	0.611440563	1					
TAX	-0.016748522	0.506455594	0.720760 18	0.668023 2	0.910228189	1				
PTRATIO	0.010800586	0.261515012	0.383247556	0.188932677	0.464741179	0.460853035	1			
AVG_ROOM	0.02739616	-0.240264931	-0.391675853	-0.302188188	-0.209846668	-0.292047833	-0.3555015	1		
LSTAT	-0.042398321	0.602338529	0.603799716	0.590878921	0.488676335	0.543993412	0.374044432	-0.613808272	1	
AVG_PRICE	0.043337871	-0.376954565	-0.483725 16	-0.427320772	-0.381626231	-0.468535934	-0.5077867	0.695359947	-0.73766	1

Top 3 positively correlated pairs :
1)TAX-DISTANCE-0.910228189
2)NOX-INDUS(0.763651447)
3)NOX-AGE(0.731470104)
Top 3 negatively correlated pairs:
1)AVG_PRICE-LSTAT(-0.73766)
2)AVG_ROOM-LSTAT(-0.613808272)
3)AVG_PRICE-PTRATIO(-0.5077867)

5) Build an initial regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot. (8 marks)

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
Intercept	**34.55384088**
LSTAT	**-0.950049354**

a)

Multiple R (Correlation):

- The multiple correlation coefficient (R) is approximately 0.74, indicating a moderate positive correlation between predictor variables and the response variable.

R Square (Coefficient of Determination)

- The R-squared value is 0.54, suggesting that 54% of the variability in the response variable can be explained by the predictor variables in the model.

Adjusted R Square:

- The adjusted R-squared value is around 0.54, similar to R-squared, but it accounts for the number of predictors in the model.

Standard Error of the Estimate:

- The standard error is approximately 6.22, representing the average distance between observed values and values predicted by the regression model. It serves as a measure of the model's accuracy.

Coefficient & Intercept

- By checking the coefficient value and the intercept value we can say that the coefficient value increases by 1 the avg price decreases by 0.95, thus we can say that they are somewhat negatively(inversely) related
- While the intercept is a positive value which signifies that it will increase the price at all the instances.

Residual Plot

A residual plot is a graphical representation of the residuals (the differences between observed and predicted values) in a regression analysis. Residual plots are useful for understanding the goodness of fit of a regression model and for identifying patterns or trends in the residuals.

b)

As per the model **LSTAT variable** has **a p value of 5.0811*E^(-88)** which is way **less than 5%**, thus we can use it for further analysis. As checking the correlation we find that it is one of the variable which is mostly negative, hence an inverse relation. Thus it will affect the average price.

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

SUMMARY OUTPUT

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

	Coefficients	***P-value***
Intercept	-1.358272812	**0.668764941**
AVG_ROOM	5.094787984	**3.47226E-27**
LSTAT	-0.642358334	**6.66937E-41**

a)

y=$\alpha + \beta 1 * X0 + \beta 2 * X1$

Where

$$\alpha - Intercept, \beta - Coefficients, X0 - AVG_{ROOM}, X1 - LSTAT$$

$\alpha = -1.358272812, \beta 1 = 5.094789984, \beta 2 =$-0.642358334,X0-7,X1=20

$y = -1.358272812 + 5.094789984 * 7 -$0.642358334*20

=21.45≈$21450

=$21458<$30000

Therefore we can say that company is overcharging.

b)

Regression Statistics	
Multiple R	0.737662726
R Square	0.544146298
Adjusted R Square	0.543241826
Standard Error	6.215760405
Observations	506

Output of previous model

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

Output of this model

Therefore on comparing both we can say than this model performs better.

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 R²square, 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	***P-value***
Intercept	29.24131526	**2.53978E-09**
CRIME_RATE	0.048725141	**0.534657201**
AGE	0.032770689	**0.012670437**
INDUS	0.130551399	**0.03912086**
NOX	-10.3211828	**0.008293859**
DISTANCE	0.261093575	**0.000137546**
TAX	-0.01440119	**0.000251247**
PTRATIO	-1.074305348	**6.58642E-15**
AVG_ROOM	4.125409152	**3.89287E-19**
LSTAT	-0.603486589	**8.91071E-27**

- As we can see below R-Square (0.69) of this model is above 50% that indicates it is good model
- we need to see the coefficients of the independent variables used for creating this model. A positive coefficient indicates that if the values of the independent variable increases, the mean of the dependent variable also tend to increase vice versa.
- A negative coefficient indicates that if the values of the independent variables increase, the mean of the dependent variables tend to decreases.
- Variables Which have P-value less than 0.05 are significant variables those who have P-value greater than 0.05 are not significant variables.
- From this we can say that crime rate is not a significant variable for average price of an house as p-value is greater than 0.5.
- All the features combined explains 69% of variability for average price of a house.
- NOX, TAX, PTRATIO and LSTAT have negative coefficients which says that increase in these features will result decrease in price of the house and vice-versa.

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	0.832835773
R Square	0.693615426
Adjusted R Square	0.688683682
Standard Error	5.131591113
Observations	506

	Coefficients	*Standard Error*	***P-value***
Intercept	29.42847349	4.804728624	**1.84597E-09**
AGE	0.03293496	0.013087055	**0.012162875**
INDUS	0.130710007	0.063077823	**0.038761669**
NOX	-10.27270508	3.890849222	**0.008545718**
DISTANCE	0.261506423	0.067901841	**0.000132887**
TAX	-0.014452345	0.003901877	**0.000236072**
PTRATIO	-1.071702473	0.133453529	**7.08251E-15**
AVG_ROOM	4.125468959	0.44248544	**3.68969E-19**
LSTAT	-0.605159282	0.0529801	**5.41844E-27**

- **Since The p-value of all the variables are less than 0.05 we can say that they are significant.**

- **However we can see that the co-efficients of some some variables are negative this implies that an increase in value of these willd ecrease the AVG_PRICE.**
- **To Conclude We can say that ,*All the Factors Excluding CRIME RATE Contribute to the AVG_PRICE of the house in the Locality*.**
- **b)In the previous model the Adjusted R-square was 69.82% And in this Model it is 69.86%.We can Say that this model performs better**
- **If the NOX is More Then the AVG_PRICE of the House will Decrease By 10times**
- **y=29.428+0.03*X0+0.130*X1-10.27*X2+0.26*X3-0.01*X4-1.07*X5+4.125*X6-0.605*X7**
  X0-AGE,X1-INDUS,X2-NOX,X3-DISTANCE,X4-TAX,X5-PTRATIO,X6-AVG_ROOM,X7-LSTAT