Assignment – Terro's real estate agency

Real estate data analysis – Exploratory data analysis, Linear Regression.

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'

Submitted By

Bheemanagouda

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

Ans)

Summary Statistics

CRIME_RATE		AGE		INDUS		NOX		DISTANCE	
Mean	4.871976285	Mean	68.57490119	Mean	11.13677866	Mean	0.554695059	Mean	9.549407115
Standard Error	0.129860152	Standard Error	1.251369525	Standard Error	0.304979888	Standard Error	0.005151391	Standard Error	0.387084894
Median	4.82	Median	77.5	Median	9.69	Median	0.538	Median	5
Mode	3.43	Mode	100	Mode	18.1	Mode	0.538	Mode	24
Standard Deviation	2.921131892	Standard Deviation	28.14886141	Standard Deviation	6.860352941	Standard Deviation	0.115877676	Standard Deviation	8.707259384
Sample Variance	8.533011532	Sample Variance	792.3583985	Sample Variance	47.06444247	Sample Variance	0.013427636	Sample Variance	75.81636598
Kurtosis	-1.189122464	Kurtosis	-0.967715594	Kurtosis	-1.233539601	Kurtosis	-0.064667133	Kurtosis	-0.867231994
Skewness	0.021728079	Skewness	-0.59896264	Skewness	0.295021568	Skewness	0.729307923	Skewness	1.004814648
Range	9.95	Range	97.1	Range	27.28	Range	0.486	Range	23
Minimum	0.04	Minimum	2.9	Minimum	0.46	Minimum	0.385	Minimum	1
Maximum	9.99	Maximum	100	Maximum	27.74	Maximum	0.871	Maximum	24
Sum	2465.22	Sum	34698.9	Sum	5635.21	Sum	280.6757	Sum	4832
Count	506	Count	506	Count	506	Count	506	Count	506

Summary Statistics

TAX		PTRATIO		AVG_ROOM		LSTAT		AVG_PRICE	
Mean	408.2371542	Mean	18.4555336	Mean	6.284634387	Mean	12.65306324	Mean	22.53280632
Standard Error	7.492388692	Standard Error	0.096243568	Standard Error	0.031235142	Standard Error	0.317458906	Standard Error	0.408861147
Median	330	Median	19.05	Median	6.2085	Median	11.36	Median	21.2
Mode	666	Mode	20.2	Mode	5.713	Mode	8.05	Mode	50
Standard Deviation	168.5371161	Standard Deviation	2.164945524	Standard Deviation	0.702617143	Standard Deviation	7.141061511	Standard Deviation	9.197104087
Sample Variance	28404.75949	Sample Variance	4.686989121	Sample Variance	0.49367085	Sample Variance	50.99475951	Sample Variance	84.58672359
Kurtosis	-1.142407992	Kurtosis	-0.285091383	Kurtosis	1.891500366	Kurtosis	0.493239517	Kurtosis	1.495196944
Skewness	0.669955942	Skewness	-0.802324927	Skewness	0.403612133	Skewness	0.906460094	Skewness	1.108098408
Range	524	Range	9.4	Range	5.219	Range	36.24	Range	45
Minimum	187	Minimum	12.6	Minimum	3.561	Minimum	1.73	Minimum	5
Maximum	711	Maximum	22	Maximum	8.78	Maximum	37.97	Maximum	50
Sum	206568	Sum	9338.5	Sum	3180.025	Sum	6402.45	Sum	11401.6
Count	506	Count	506	Count	506	Count	506	Count	506

Here are some observations based on the summary statistics provided for the CRIME_RATE variable:

- The mean CRIME_RATE is 4.87, indicating that the average crime rate in the area is around 4.87.
- The median crime rate is slightly lower than the mean at 4.82, suggesting that the distribution of crime rates may be slightly skewed to the right.
- The standard deviation is relatively high at 2.92, which suggests that the crime rates in the area vary widely from the mean.
- The range of CRIME_RATE is quite large at 9.95, with the minimum rate being 0.04 and the maximum rate being 9.99.
- The mode of the CRIME_RATE is 3.43, which is lower than both the mean and median, indicating that the distribution may be slightly skewed to the left.
- The sample variance and standard error provide measures of the variability and precision of the sample mean, respectively.

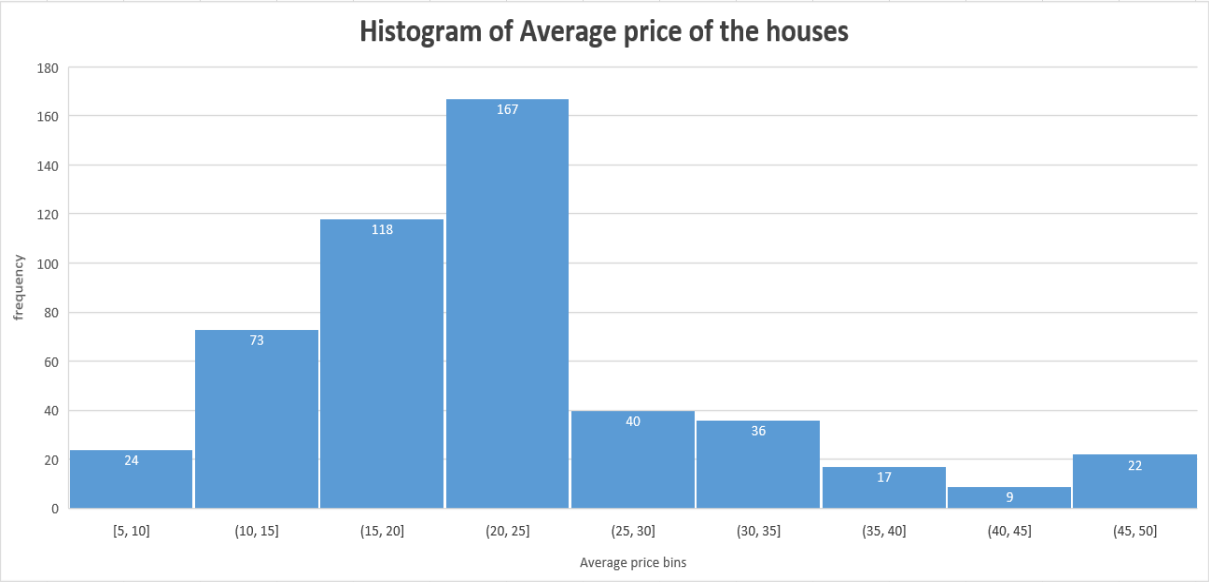
- The skewness value is positive but very close to 0, indicating that the distribution of crime rates is roughly symmetrical.
- The kurtosis value is negative, indicating that the distribution is slightly less peaked than a normal distribution.

Similarly, the summary statistics of all other variables can be interpreted as shown above .

Overall, these summary statistics provide insight into the central tendency, variability, and shape of the distribution of crime rates in the area.

Q2). Plot a histogram of the AVG_PRICE variable. What do you infer?

Ans)



Frequency Distribution Table

AVG_PRICE	Frequency
5-10	24
10-15	73
15-20	118
20-25	167
25-30	40
30-35	36
35-40	17
40-45	9
45-50	22
Grand Total	**506**

Observations:

•The x-axis represents the price ranges, while the y-axis represents the frequency of houses falling in that range.

•The most common price range for the houses is between $20K-$25K, with 164 house count falling in this range.

•There are very few products in the highest price range of $45K-$50K, with only 9 counts falling in this range.

•The majority of the products fall in the range of $10K-$30K, with fewer products on the lower and higher ends of the range.

Overall, the histogram provides a visual representation of the frequency distribution of the AVG_PRICE variable, allowing us to observe the range and distribution of prices of houses.

Q3). Compute the covariance matrix. Share your observations.

Ans)

Covariance Matrix										
	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	8.516									
AGE	0.563	790.792								
INDUS	-0.110	124.268	46.971							
NOX	0.001	2.381	0.606	0.013						
DISTANCE	-0.230	111.550	35.480	0.616	75.667					
TAX	-8.229	2397.942	831.713	13.021	1333.117	28348.624				
PTRATIO	0.068	15.905	5.681	0.047	8.743	167.821	4.678			
AVG_ROOM	0.056	-4.743	-1.884	-0.025	-1.281	-34.515	-0.540	0.493		
LSTAT	-0.883	120.838	29.522	0.488	30.325	653.421	5.771	-3.074	50.894	
AVG_PRICE	1.162	-97.396	-30.461	-0.455	-30.501	-724.820	-10.091	4.485	-48.352	84.420

Here are some observations based on the above covariance matrix:

•The diagonal elements represent the variance of each variable.

•The covariance between CRIME_RATE and TAX is negative (-8.229), which suggests a negative linear relationship between these two variables. This means that areas with higher crime rates tend to have lower property taxes.

•The covariance between AGE and TAX is positive (2397.942), which suggests a strong positive linear relationship between these two variables. This means that older houses tend to have higher property taxes.

•The covariance between LSTAT and AVG_PRICE is negative (-48.352), which suggests a strong negative linear relationship between these two variables. This means that areas with higher percentages of lower-status population tend to have lower average house prices.

•The covariance of 13.021 between NOX and TAX suggests a positive linear relationship between these two variables. This means that as the level of nitric oxide (NOX) increases in a particular area, the taxes (TAX) tend to be higher in that same area.

It is important to note that the covariance matrix only provides information about the linear relationships between variables, and it does not indicate causality, and additional analysis may be required to determine true relationship between variables.

Q4). Create a correlation matrix of all the variables (Use Data analysis tool pack).

a) Which are the top 3 positively correlated pairs.

b) Which are the top 3 negatively correlated pairs.

Ans)

Correlation Matrix										
	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	1.000									
AGE	0.007	1.000								
INDUS	-0.006	0.645	1.000							
NOX	0.002	0.731	0.764	1.000						
DISTANCE	-0.009	0.456	0.595	0.611	1.000					
TAX	-0.017	0.506	0.721	0.668	0.910	1.000				
PTRATIO	0.011	0.262	0.383	0.189	0.465	0.461	1.000			
AVG_ROOM	0.027	-0.240	-0.392	-0.302	-0.210	-0.292	-0.356	1.000		
LSTAT	-0.042	0.602	0.604	0.591	0.489	0.544	0.374	-0.614	1.000	
AVG_PRICE	0.043	-0.377	-0.484	-0.427	-0.382	-0.469	-0.508	0.695	-0.738	1.000

a) The top 3 positively correlated pairs are:	b) The top 3 negatively correlated pairs are:
DISTANCE & TAX = 0.910	LSTAT & AVG_PRICE = -0.738
INDUS & NOX = 0.764	AVG_ROOM & LSTAT = -0.614
AGE & NOX = 0.731	PTRATIO & AVG_PRICE = -0.508

a) The top 3 positively correlated pairs are:

- DISTANCE and TAX (correlation coefficient of 0.910)
- INDUS and NOX (correlation coefficient of 0.764)
- AGE and NOX (correlation coefficient of 0.731)

b) The top 3 negatively correlated pairs are:

- AVG_PRICE and LSTAT (correlation coefficient of -0.738)
- AVG_ROOM and LSTAT (correlation coefficient of -0.614)
- PTRATIO and AVG_PRICE (correlation coefficient of -0.508)

Q5). 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?

Ans) Regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as Independent Variable.

SUMMARY OUTPUT

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

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	23243.914	23243.914	601.6179	5.0811E-88
Residual	504	19472.38142	38.6356774		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	34.55384088	0.562627355	61.4151455	3.7E-236	33.44845704	35.65922472	33.448457	35.65922472
LSTAT	-0.950049354	0.038733416	-24.5278999	5.08E-88	-1.0261482	-0.873950508	-1.0261482	-0.873950508

a). The above observation is the result of a linear regression between the dependent variable AVG_PRICE and independent variable LSTAT.

The Multiple R value of 0.737 indicates a strong positive correlation between LSTAT and AVG_PRICE. The R-squared value of 0.544 indicates that LSTAT explains 54.4% of the variance in AVG_PRICE.

The coefficient value for LSTAT is -0.950, which indicates that there is a negative relationship between LSTAT and AVG_PRICE. In other words, as LSTAT increases, AVG_PRICE tends to decrease.

The Intercept value of 34.554 represents the estimated AVG_PRICE when LSTAT is 0.

The p-value for LSTAT is very small (5.08E-88), indicating that LSTAT is a significant predictor of AVG_PRICE. Overall, this regression model suggests that LSTAT is a strong predictor of AVG_PRICE and has a negative impact on house prices.

b). Yes, the LSTAT variable is significant for the analysis based on the data provided. This is because the p-value associated with the LSTAT coefficient is very small (5.08E-88), which is less than the conventional significance level of 0.05.

Q6). 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) s the performance of this model better than the previous model you built in Question 5? Compare in terms of adjusted R-square and explain.

Ans) Regression model including LSTAT and AVG_ROOM together as Independent variables and AVG_PRICE as dependent variable.

SUMMARY OUTPUT

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

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	27276.98621	13638.49311	444.3308922	7.0085E-112
Residual	503	15439.3092	30.69445169		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.358272812	3.17282778	-0.428095348	0.668764941	-7.591900282	4.875354658	-7.591900282	4.875354658
AVG_ROOM	5.094787984	0.4444655	11.46272991	3.47226E-27	4.221550436	5.968025533	4.221550436	5.968025533
LSTAT	-0.642358334	0.043731465	-14.68869925	6.66937E-41	-0.728277167	-0.556439501	-0.728277167	-0.556439501

A). The regression equation for the given data is:

AVG_PRICE = coefficient of intercept + (Coefficient of AVG_ROOM * AVG_ROOM Value) + (Coefficient of LSTAT * LSTAT Value)

AVG_PRICE = -1.358 +(5.095 * AVG_ROOM) + (- 0.642 * LSTAT)

If a new house in this locality has 7 rooms (on an average) and has a value of 20 for LSTAT, then its AVG_PRICE will be:

AVG_PRICE = -1.358 + (5.095 * 7) + (- 0.642 * 20) = -1.358 + 35.665 - 12.84 = 21.45807

AVG_PRICE (in USD) = 21.45807 * 1000 = 21,458.07 USD

The predicted value of AVG_PRICE for this house is 21,458 USD. This value is less than the company quoting a value of 30,000 USD for this locality. Hence, the company is overcharging for the house.

B). The previous model had an adjusted R-square value of 0.543, which means that it was able to explain 54.3% of the variation in the dependent variable using the independent variable(s). On the other hand, the current model has an adjusted R-square value of 0.637, which means that it can explain 63.7% of the variation in the dependent variable

As we can see, the adjusted R-square value of the current model is higher than the adjusted R-square value of the previous model. Thus, We can say that the current model is a better fit for the data compared to the previous model.

Q7). 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.

Ans) Regression model with all variables where AVG_PRICE alone be the Dependent Variable and all the other variables are independent.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372
Adjusted R Square	0.688298647
Standard Error	5.1347635
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	29638.8605	3293.206722	124.9045049	1.9328E-121
Residual	496	13077.43492	26.3657962		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.24131526	4.817125596	6.070282926	2.53978E-09	19.77682784	38.70580267	19.77682784	38.70580267
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.534657201	-0.105348544	0.202798827	-0.105348544	0.202798827
AGE	0.032770689	0.013097814	2.501996817	0.012670437	0.00703665	0.058504728	0.00703665	0.058504728
INDUS	0.130551399	0.063117334	2.068392165	0.03912086	0.006541094	0.254561704	0.006541094	0.254561704
NOX	-10.3211828	3.894036256	-2.650510195	0.008293859	-17.97202279	-2.670342809	-17.97202279	-2.670342809
DISTANCE	0.261093575	0.067947067	3.842602576	0.000137546	0.127594012	0.394593138	0.127594012	0.394593138
TAX	-0.01440119	0.003905158	-3.687736063	0.000251247	-0.022073881	-0.0067285	-0.022073881	-0.0067285
PTRATIO	-1.074305348	0.133601722	-8.041104061	6.58642E-15	-1.336800438	-0.811810259	-1.336800438	-0.811810259
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.89287E-19	3.255494742	4.995323561	3.255494742	4.995323561
LSTAT	-0.603486589	0.053081161	-11.36912937	8.91071E-27	-0.70777824	-0.499194938	-0.70777824	-0.499194938

The given regression model shows an adjusted R-square of 0.688, which indicates that 68.8% of the variation in the dependent variable AVG_PRICE can be explained by the independent variables included in the model.

The intercept value is 29.24, which represents the predicted value of AVG_PRICE when all the independent variables are equal to zero.

The coefficients of the independent variables can be interpreted as the change in the dependent variable for one unit change in independent variable.

In this model, the independent variable CRIME_RATE has a coefficient of 0.0487, which means that for every one-unit increase in CRIME_RATE, the AVG_PRICE is predicted to increase by 0.0487.

Similarly, the independent variables AGE, INDUS, DISTANCE, and AVG_ROOM have coefficients of 0.0327, 0.1305, 0.2610 and 4.125 respectively. This means that for every one-unit increase in INDUS, NOX, and LSTAT, the AVG_PRICE is predicted to increase by of 0.0327, 0.1305, 0.2610 and 4.125 respectively.

The independent variable NOX has a coefficient of -10.321 implies that for every one-unit increase in NOX, the AVG_PRICE is estimated to decrease by 10.321.

Similarly, the independent variables TAX, PTRATIO, and LSTAT have coefficients of -0.144, -1.0743 and -0.6034 respectively. This means that for every one-unit increase in TAX, PTRATIO, and LSTAT is predicted to decrease AVG_PRICE by of -0.14, -1.0743 and -0.6034 respectively.

Overall, In this model, AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG_ROOM, and LSTAT have p-values less than 0.05, which indicates that these variables are significant predictors of AVG_PRICE. CRIME_RATE has a p-value of 0.535, which is greater than 0.05, indicating that it is not a significant predictor of AVG_PRICE in this model.

Q8). 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:

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.

Ans)

SUMMARY OUTPUT

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

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	29628.68142	3703.585178	140.6430411	1.911E-122
Residual	497	13087.61399	26.33322735		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.42847349	4.804728624	6.124898157	1.84597E-09	19.98838959	38.8685574	19.98838959	38.8685574
AGE	0.03293496	0.013087055	2.516605952	0.012162875	0.007222187	0.058647734	0.007222187	0.058647734
INDUS	0.130710007	0.063077823	2.072202264	0.038761669	0.006777942	0.254642071	0.006777942	0.254642071
NOX	-10.27270508	3.890849222	-2.640221837	0.008545718	-17.9172457	-2.628164466	-17.9172457	-2.628164466
DISTANCE	0.261506423	0.067901841	3.851242024	0.000132887	0.128096375	0.394916471	0.128096375	0.394916471
TAX	-0.014452345	0.003901877	-3.703946406	0.000236072	-0.022118553	-0.006786137	-0.022118553	-0.006786137
PTRATIO	-1.071702473	0.133453529	-8.030529271	7.08251E-15	-1.333905109	-0.809499836	-1.333905109	-0.809499836
AVG_ROOM	4.125468959	0.44248544	9.323400461	3.68969E-19	3.256096304	4.994841615	3.256096304	4.994841615
LSTAT	-0.605159282	0.0529801	-11.42238841	5.41844E-27	-0.70925186	-0.501066704	-0.70925186	-0.501066704

a) Above is the multiple linear regression model with AVG_PRICE as the dependent variable and AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG_ROOM, and LSTAT as the independent variables.

In this model, the coefficients for AGE, INDUS, DISTANCE, and AVG_ROOM are positive, indicating that as these variables increase, the AVG_PRICE also tends to increase. The coefficients for NOX, TAX, PTRATIO and LSTAT are negative, indicating that as these variables increase, the AVG_PRICE tends to decrease.

Notably, the coefficient for NOX is quite large, indicating that air pollution (measured by NOX) has a particularly strong negative impact on AVG_PRICE.

b).The adjusted R-square value of the first model is 0.688298646855749, while the adjusted R-square value of the second model is 0.6886836.

However, the difference in adjusted R-square values between the two models is relatively small still, Comparing the adjusted R-square values, we see that the present model (adjusted R-square = 0.6886836) has a slightly higher value than the previous model (adjusted R-square = 0.688298646855749). This suggests that the second model is a slightly better fit for the data, as it explains a slightly larger proportion of the variability in the dependent variable (AVG_PRICE).

c). <u>coefficients in ascending order</u>

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

From the sorted coefficients, we can see that the NOX variable has the largest negative coefficient (-10.27270508), which suggests that an increase in the level of nitric oxide (NOX) concentration in a locality would lead to a decrease in the average price of houses in that locality.

d) AVG_PRICE = Coefficient of Intercept + (Coefficient of AGE * AGE Value) + (Coefficient of INDUS * INDUS Value) + (Coefficient of NOX * NOX Value) + (Coefficient of DISTANCE * DISTANCE Value) + (Coefficient of TAX * TAX Value) + (Coefficient of PTRATIO * PTRATIO Value) + (Coefficient of AVG_ROOM * AVG_ROOM Value) + (Coefficient of LSTAT * LSTAT Value).

i.e, AVG_PRICE = 29.4284 + (0.0329 * AGE) + (0.1307 * INDUS) + (-10.2727 * NOX) + (0.2615 * DISTANCE) + (-0.0144 * TAX) + (-1.0717 * PTRATIO) + (4.1254 * AVG_ROOM) + (-0.6051 * LSTAT).