TERRO'S REAL ESTATE AGENCY

Real estate data analysis-

Exploratory data analysis, Linear Regression

NAME – JAGRITI RAI

GLCA DA SEP 2023

DATE- 05-11-2023

jagriti rai

Contents:

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 **12**

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? **12**

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 **13**

Conclusion **13**

END **13**

List of Tables:

List of Charts:

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.

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):

Your job, as an auditor, is to analyse the magnitude of each variable to which it can affect the price of a house in a particular locality.

SAMPLE DATASET:

CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
6.32	65.2	2.31	0.538	1	296	15.3	6.575	4.98	24
4.31	78.9	7.07	0.469	2	242	17.8	6.421	9.14	21.6
7.87	61.1	7.07	0.469	2	242	17.8	7.185	4.03	34.7
6.47	45.8	2.18	0.458	3	222	18.7	6.998	2.94	33.4
5.24	54.2	2.18	0.458	3	222	18.7	7.147	5.33	36.2
9.75	58.7	2.18	0.458	3	222	18.7	6.43	5.21	28.7
9.42	66.6	7.87	0.524	5	311	15.2	6.012	12.43	22.9
2.76	96.1	7.87	0.524	5	311	15.2	6.172	19.15	27.1
7.66	100	7.87	0.524	5	311	15.2	5.631	29.93	16.5
1.12	85.9	7.87	0.524	5	311	15.2	6.004	17.1	18.9
7.52	94.3	7.87	0.524	5	311	15.2	6.377	20.45	15
1.55	82.9	7.87	0.524	5	311	15.2	6.009	13.27	18.9
3.7	39	7.87	0.524	5	311	15.2	5.889	15.71	21.7
7.14	61.8	8.14	0.538	4	307	21	5.949	8.26	20.4
0.21	84.5	8.14	0.538	4	307	21	6.096	10.26	18.2
8.6	56.5	8.14	0.538	4	307	21	5.834	8.47	19.9
6.95	29.3	8.14	0.538	4	307	21	5.935	6.58	23.1
0.8	81.7	8.14	0.538	4	307	21	5.99	14.67	17.5
8.5	36.6	8.14	0.538	4	307	21	5.456	11.69	20.2

Table 1 – dataset sample

This dataset has 10 variables with detail of 506 houses in Boston. It gives us all the information that may be used in predicting the right price of the room as per locality.

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

Summary Statistic:

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
Mean	4.871976285	68.57490119	11.13677866	0.554695059	9.549407115	408.2371542	18.4555336	6.284634387	12.65306324	22.53280632
Standard Error	0.129860152	1.251369525	0.304979888	0.005151391	0.387084894	7.492388692	0.096243568	0.031235142	0.317458906	0.408861147
Median	4.82	77.5	9.69	0.538	5	330	19.05	6.2085	11.36	21.2
Mode	3.43	100	18.1	0.538	24	666	20.2	5.713	8.05	50
Standard Deviation	2.921131892	28.14886141	6.860352941	0.115877676	8.707259384	168.5371161	2.164945524	0.702617143	7.141061511	9.197104087
Sample Variance	8.533011532	792.3583985	47.06444247	0.013427636	75.81636598	28404.75949	4.686989121	0.49367085	50.99475951	84.58672359
Kurtosis	-1.189122464	-0.967715594	-1.233539601	-0.064667133	-0.867231994	-1.142407992	-0.285091383	1.891500366	0.493239517	1.495196944
Skewness	0.021728079	-0.59896264	0.295021568	0.729307923	1.004814648	0.669955942	-0.802324927	0.403612133	0.906460094	1.108098408
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	34698.9	5635.21	280.6757	4832	206568	9338.5	3180.025	6402.45	11401.6
Count	506	506	506	506	506	506	506	506	506	506

Table 2. Descriptive Stats

- Summary statistics or descriptive statistic is performed to check the overall information about the dataset.
- It gives the information about where there's any missing value in the data or not'
- It gives all the statistical information such as: Mean, Median, Mode, Standard deviation, Kurtosis, Skewness, etc.
- Also, the basic calculations such as: Sum, Count, Min, Max, Range.
- So before moving to any calculation the first and most important thing is to get the statistical summary of the given the data.

Q2. Plot a histogram of the Avg. Price variable. What do you infer?

HISTOGRAM ON AVERAGE PRICE:

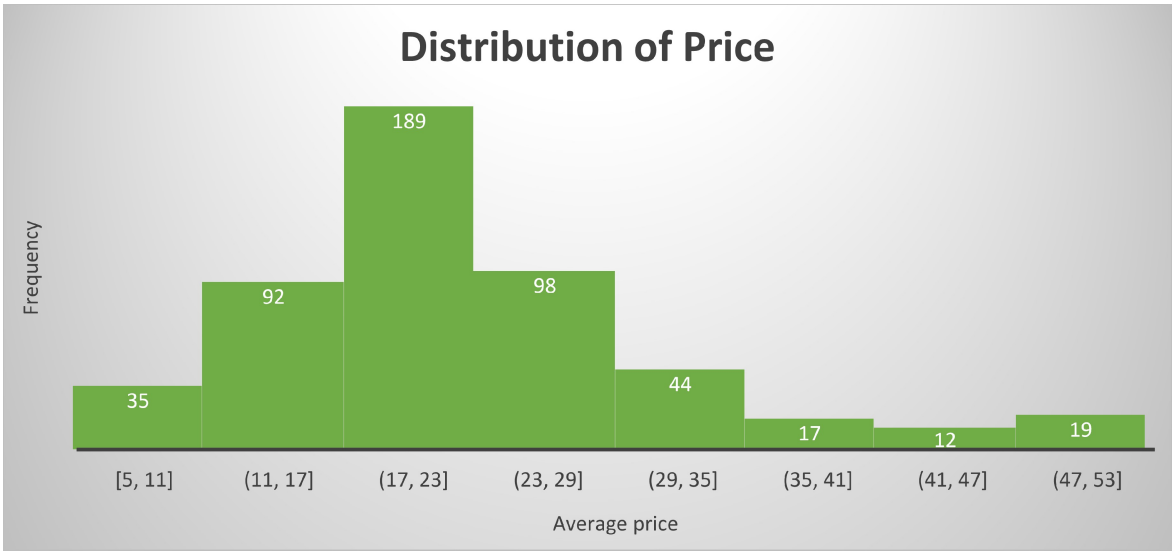

Chart 1. Histogram (dist. of price)

In this given chart, there is the distribution of Average price vary from locality to locality:

Considering this chart there are some inferences given below:

- This histogram is right-skewed, with a longer tail in right side.
- There are some houses with significantly higher prices as compared to the majority of houses.
- There are few houses with exceptionally high prices, they are considered as outliers.
- The range price varies from 5 to 53 depending on the locality and other variable.

As there are some outliers also in this histogram chart, so to know them more clearly, I have drawn a Box and Whisker chart for this same data.

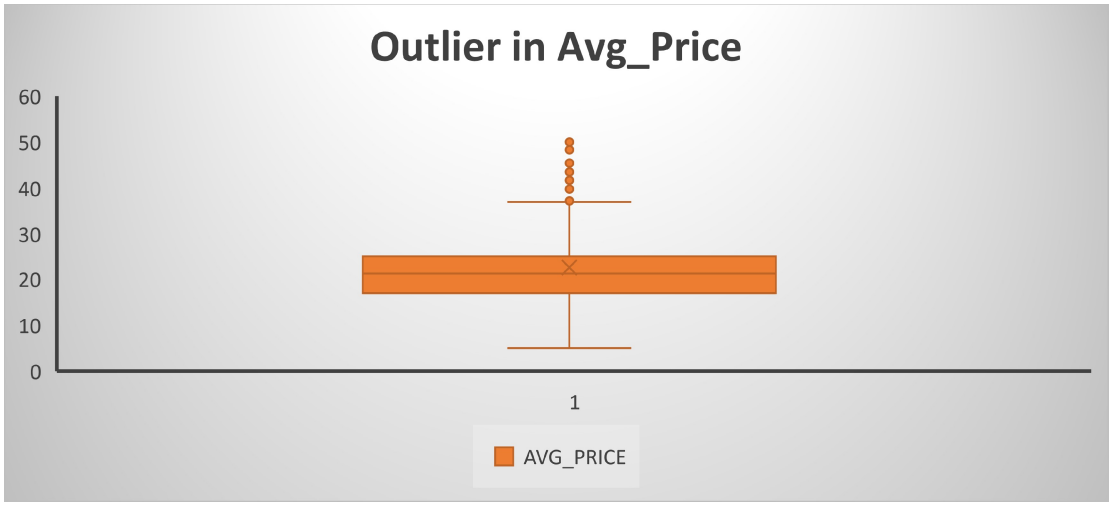

Chart 2. Box and Whisker

Q3. Compute the covariance matrix. Share your observations.

COVARIANCE ANALYSIS:

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
2	8.516147873									
3	0.562915215	790.7924728								
4	-0.11021518	124.2678282	46.97142974							
5	0.000625308	2.381211931	0.605873943	0.013401099						
6	-0.22986049	111.5499555	35.47971449	0.615710224	75.66653127					
7	-8.22932244	2397.941723	831.7133331	13.02050236	1333.116741	28348.6236				
8	0.068168906	15.90542545	5.680854782	0.047303654	8.74340249	167.8208221	4.677726296			
9	0.056117778	-4.74253803	-1.88422543	-0.02455483	-1.28127739	-34.515101	-0.53969452	0.492695216		
10	-0.88268036	120.8384405	29.52181125	0.487979871	30.32539213	653.4206174	5.771300243	-3.07365497	50.89397935	
11	1.16201224	-97.3961529	-30.460505	-0.45451241	-30.5008304	-724.820428	-10.0906756	4.484565552	-48.3517922	84.4195562

Table 3. Covariance

Covariance is a statistical measure that quantifies the relationship between two variables. It is used to understand how changes in one variable are related to changes in another variable.

From the given table of covariance, the insights we get are:

- Understanding the Direction of Relationship - If there is positive covariance the variable tends to move in same direction, but if there is negative covariance the variable is tend to move in opposite direction.
- Quantifying the Strength of the Relationship - A higher absolute covariance value indicates a stronger relationship, while a lower value suggests a weaker relationship.
- Linear Regression Analysis - Covariance is used in linear regression analysis to determine the extent to which independent variables are related to the dependent variable.
- Multivariate Analysis - Covariance is crucial in multivariate analysis, where the relationships between multiple variables are examined simultaneously. It helps in understanding the interdependencies and interactions between different variables in a complex system.

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

CORRELATION ANALYSIS:

		CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
2	CRIME_RATE	1									
3	AGE	0.006859463	1								
4	INDUS	-0.005510651	0.644778511	1							
5	NOX	0.001850982	0.731470104	0.763651447	1						
6	DISTANCE	-0.009055049	0.456022452	0.595129275	0.611440563	1					
7	TAX	-0.016748522	0.506455594	0.72076018	0.6680232	0.910228189	1				
8	PTRATIO	0.010800586	0.261515012	0.383247556	0.188932677	0.464741179	0.460853035	1			
9	AVG_ROOM	0.02739616	-0.240264931	-0.391675853	-0.302188188	-0.209846668	-0.292047833	-0.355501495	1		
10	LSTAT	-0.042398321	0.602338529	0.603799716	0.590878921	0.488676335	0.543993412	0.374044317	-0.613808272	1	
11	AVG_PRICE	0.043337871	-0.376954565	-0.48372516	-0.427320772	-0.381626231	-0.468535934	-0.507786686	0.695359947	-0.737662726	1

Table 4. Correlation Matrix

Correlation is a statistical measure that quantifies the strength and direction of the linear relationship between two variables.

In this table of correlation matrix, we can easily say which variable is positively related to other and which one is negatively related to other.

Q4(a) Which are the top 3 positively correlated pairs and b) Which are the top 3 negatively correlated pairs.

There are top 3 positively correlated variables, which are:

- TAX and DISTANCE - 0.910228189
- NOX and INDUS – 0.763651447
- NOX and AGE – 0.731470104

There are top 3 negatively correlated variables, which are:

- AVG_PRICE and LSTAT – (-0.737662726)
- LSTAT and AVG_ROOM – (-0.61308272)
- AVG_PRICE and PTRATIO – (-0.507786686)

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 Residual plot? b) Is LSTAT variable significant for the analysis based on your model?

REGRESSION MODEL:

5	SUMMARY OUTPUT								
6									
7	*Regression Statistics*								
8	Multiple R	0.737662726							
9	R Square	0.544146298							
10	Adjusted R Squa	0.543241826							
11	Standard Error	6.215760405							
12	Observations	506							
13									
14									
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	34.55384088	0.562627355	61.41514552	3.7431E-236	33.44845704	35.65922472	33.44845704	35.65922472
18	LSTAT	-0.950049354	0.038733416	-24.52789985	5.0811E-88	-1.0261482	-0.873950508	-1.0261482	-0.873950508

Table 5. Regression analysis

This is a regression summary of only two variables, where AVG_PRICE is 'Y' (dependent variable) and LSTAT is 'X' (independent variable).

There is also a residual plot for this regression:

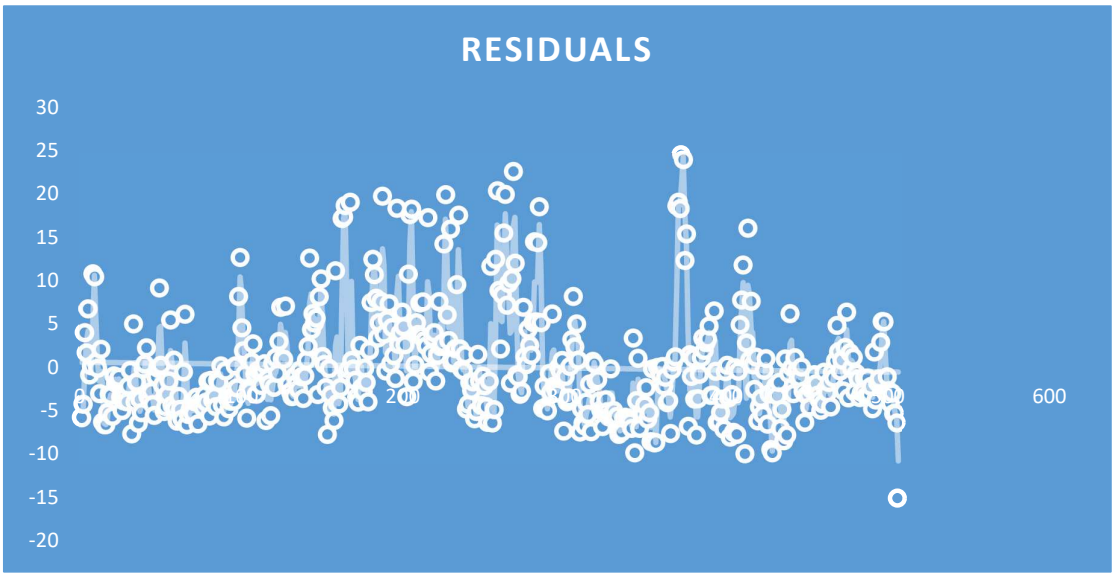

Chart 3. Residual plot.

By this regression model, we have gained some insights that are given below:

1. The regression model between AVG_PRICE and LSTAT is significant with a p-value of (5.0811E-88).
2.There is a negative linear relationship between the 'LSTAT' variable and the dependent variable.
>> Which means for every one-unit increase in 'LSTAT', the dependent variable decreases by approximately 0.95 units.
3. There is a lower standard error, which means the model has a good fit to the data.
4. Intercept value suggest that if the independent variable is zero, the estimate value of dependent variable will be 34.55.

After the calculation of RMSE which Is 28%, we can say that:

- The RMSE is 28%, which implies that, the model's predictions deviate by 28% of the mean value.
- The 'LSTAT' variable is significant for this model.
- The coefficient for the 'LSTAT' variable is -0.950049354, and it has a very low p-value of 5.0811E-88.

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) 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 MODEL 2:

SUMMARY OUTPUT

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	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

Table 6. Regression model2

This Regression model including LSTAT and AVG_ROOM together as independent variables and AVG_PRICE as dependent variable.

 # AVG_ROOM = 7
 # LSTAT = 20
Average price = 21.45807639
AFTER THE CALCULATION OF THE AVG PRICE, THE ACTUAL PRICE OF THIS ROOMS WIL BE AROUND 21000 AND LITTLE BIT MORE BUT THEY CHAREGES 30000 FOR THIS, THAT IS DIRECTLY OVERCHARGE.

COMPARISON:
Adjusted R Square of this data = 0.637124475
Adjusted R Square of previous model = 0.543241826

The closer Adjusted R square to 1, the higher it significant:
In this case Adjusted R Square of this data is closer to 1.
Hence, this regression model is more significant as compared to 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.

REGRESSION MODEL 3:

SUMMARY OUTPUT

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	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	29.24131526	4.817125596	6.07028293	2.5398E-09	19.7768278	38.7058027	19.7768278	38.7058027
CRIME_RATE	0.048725141	0.078418647	0.62134637	0.5346572	-0.10534854	0.20279883	-0.10534854	0.20279883
AGE	0.032770689	0.013097814	2.50199682	0.01267044	0.00703665	0.05850473	0.00703665	0.05850473
INDUS	0.130551399	0.063117334	2.06839217	0.03912086	0.00654109	0.2545617	0.00654109	0.2545617
NOX	-10.3211828	3.894036256	-2.6505102	0.00829386	-17.9720228	-2.67034281	-17.9720228	-2.67034281
DISTANCE	0.261093575	0.067947067	3.84260258	0.00013755	0.12759401	0.39459314	0.12759401	0.39459314
TAX	-0.01440119	0.003905158	-3.68773606	0.00025125	-0.02207388	-0.0067285	-0.02207388	-0.0067285
PTRATIO	-1.074305348	0.133601722	-8.04110406	6.5864E-15	-1.33680044	-0.81181026	-1.33680044	-0.81181026
AVG_ROOM	4.125409152	0.442758999	9.31750493	3.8929E-19	3.25549474	4.99532356	3.25549474	4.99532356
LSTAT	-0.603486589	0.053081161	-11.3691294	8.9107E-27	-0.70777824	-0.49919494	-0.70777824	-0.49919494

Table 7. Regression Model 3

With the help of above regression model, we get some important insights. That are as follows:

1.Adjusted R Square is 68.82%, that indicates approximately 68.83% of the variability in the average price
(AVG_PRICE) can be explained by the independent variables in the model.

2. The Coefficients indicate the direction and magnitude of the impact of each independent variable on the average price.

3.Intercept indicates value of dependent variable, when all the independent variable tends to be Zero.

4. While dealing with p-value every other variable is statistically significant except one that is CRIME_RATE.

The significance of each independent variable with respect to AVG. PRICE: -

1. **CRIME_RATE**: Coefficient = 0.048725141
   P-Value = 0.534657201, That shows it is statistically insignificant for average price.
2. **AGE:** Coefficient= 0.032770689
   P-Value = 0.012670437, The p-value is <0.05 so, this is statistically significant to average price.
3. **INDUS:** Coefficient = 0.130551399
   P-Value = 0.03912086, The p-value is <0.05 so. This is statistically significant to average price.
4. **NOX:** Coefficient = -10.3211828
   P-Value = 0.008293859, The p-value is <0.05 so, this is statistically significant to model.
5. **DISTANCE:** Coefficient = 0.261093575
   P-Value = 0.000137546, The p-value is <0.05 so, this is statistically significant.
6. **TAX:** Coefficient = -0.01440119
   P-Value = 0.000251247, The p-value is <0.05 so, this is statistically significant.
7. **PTRATIO:** Coefficient = -1.074305348
   P-Value = 6.58642E-15, The p-value is <0.05 so, this is statistically significant to model.
8. **AVG_ROOM:** Coefficient = 4.125409152
   P-Value = 3.89287E-19, The p-value is <0.05 so, this is statistically significant to the model.
9. **LSTAT:** Coefficient = -0.603486589
   P-Value = 8.91071E-27, The p-value is <0.05 so, this is statistically significant.

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

REGRESSION MODEL 4:

SUMMARY OUTPUT

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

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

Table 8. Regression Model 4

This is the final regression model after removing insignificant variable. There are some useful business insights we extract from this regression model. These are given below:

A) INTERPRETATION:

1. Adjusted R-square (68.86%) implies that the chosen independent variables collectively have a meaningful relationship with the average price.
2. Intercept value (29.428) shows the value of dependent variable if all independent variable is Zero.
3. P-value shows that all the independent variables are statistically significant because it implies (<0.05%)
4. Coefficients indicates the impact of each variable on average price whether it is positive/negative.

B) COMPARISON:

1. Multiple R in the both the cases are same (83% approx.)
2. For Adjusted R Square:

In this data, Adjusted R Square = 0.688683682

In Previous data, Adjusted R Square = 0.688298647

As we know, Adjusted R square is good when its closer to 1,
So, in this case the difference is very nominal but while doing the comparison we can say that:

Adjusted R Square of this data is more significant than the previous one because it is more likely to close to 1.

C) Regression Equation: Y = b0+(b1X1) +(b2X2) + …..+(bnXn)

where:

* Y is the dependent variable.
* b0 is the intercept.
* b1, b2, …, bn is the regression coefficient of the independent variables X1, X2, …, Xn

CONCLUSION:

While we conclude this project report, we can say that at last all the variables are perfectly significant, except CRIME_RATE. There are 506 houses with different price range and also have some outliers means there are some houses which are unexpectedly higher than other houses. The whole report helps us to find every answer related to buy a room in a locality while considering all other variables too.

REFERENCE:

There is an Excel File regarding this report referred for verification. In that excel sheet I have use many tools of Data Analysis to find the answer of all questions, such as, Descriptive Summary, Covariance, Correlation, And most important Regression Analysis.

THE END