Terro's Real Estate Agency

BUSINESS REPORT

Submitted by

Prasanna B

Problem Statement (Situation):

"Finding out the most relevant features for pricing of a house".

Terro's real-estate is an agency that estimates the price 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, studying various geographic features of a property like pollution level (NOX), crime rate, education facilities (pupil to teacher ratio), connectivity (distance from highway), etc. must be carried out. This helps in determining the price of a property.

The agency has provided a dataset of 506 houses in Boston. The following are the details of the dataset:

Attribution	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

INDEX

Table of contents

OBJECTIVES

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

Step by step approach:

- Go to the Data tab.
- Click data analysis.
- Click Descriptive statistics.
- Select the range of cells for which the summary statistics must be calculated and select a new worksheet.
- Enable the summary statistics checkbox.
- Click ok.
- Now, the required statistics appears on a new worksheet.

Summary statistics & observations:

A. CRIME RATE:

CRIME_RATE	
Mean	4.871976
Standard Error	0.12986
Median	4.82
Mode	3.43
Standard Deviation	2.921132
Sample Variance	8.533012
Kurtosis	-1.18912
Skewness	0.021728
Range	9.95
Minimum	0.04
Maximum	9.99
Sum	2465.22
Count	506

- Half of the households have a crime rate above 4.82. Even though the skewness is positive, but it just touches zero, the crime rate seems to follow the normal distribution and is almost fairly distributed around the average crime rate of 4.87.

B. AGE:

AGE	
Mean	68.5749
Standard Error	1.25137
Median	77.5
Mode	100
Standard Deviation	28.14886
Sample Variance	792.3584
Kurtosis	-0.96772
Skewness	-0.59896
Range	97.1
Minimum	2.9
Maximum	100
Sum	34698.9
Count	506

- Most of the households (i.e., 68.5%) are built prior to 1940, which ideally says that new households present is relatively less in number.
- The graph of this variable tends to be platykurtic & negatively skewed.

C. INDUSTRY:

INDUS	
Mean	11.13678
Standard Error	0.30498
Median	9.69
Mode	18.1
Standard Deviation	6.860353
Sample Variance	47.06444
Kurtosis	-1.23354
Skewness	0.295022
Range	27.28
Minimum	0.46
Maximum	27.74
Sum	5635.21
Count	506

- In areas where the industry value is above the median of 9.69, the prices of the households can be increased as the market value for the land can increase drastically in future.
- Proportion of non-retail business acres ranges between 0.46 – 27.74.

D. NOX:

NOX	
Mean	0.554695
Standard Error	0.005151
Median	0.538
Mode	0.538
Standard Deviation	0.115878
Sample Variance	0.013428
Kurtosis	-0.06467
Skewness	0.729308
Range	0.486
Minimum	0.385
Maximum	0.871
Sum	280.6757
Count	506

- Prices of houses in areas where the NOX value is between 0.385(minimum) and 0.538(median), can be increased since any customer would prefer a less polluted area for household.

E. DISTANCE:

DISTANCE	
Mean	9.549407
Standard Error	0.387085
Median	5
Mode	24
Standard Deviation	8.707259
Sample Variance	75.81637
Kurtosis	-0.86723
Skewness	1.004815
Range	23
Minimum	1
Maximum	24
Sum	4832
Count	506

- The positive skew says that most of the houses are nearby the highway road and their distance from the highway road is below the average value of 9.54 miles.
- Most of the houses are 24 miles far away from the highway road.

F. TAX:

TAX	
Mean	408.2372
Standard Error	7.492389
Median	330
Mode	666
Standard Deviation	168.5371
Sample Variance	28404.76
Kurtosis	-1.14241
Skewness	0.669956
Range	524
Minimum	187
Maximum	711
Sum	206568
Count	506

- The tax value ranges between 187-711.
- The positive skew indicates that most of the houses' tax values are less than the mean value of $408.23 (per $10,000)

G. PT-RATIO:

PTRATIO	
Mean	18.45553
Standard Error	0.096244
Median	19.05
Mode	20.2
Standard Deviation	2.164946
Sample Variance	4.686989
Kurtosis	-0.28509
Skewness	-0.80232
Range	9.4
Minimum	12.6
Maximum	22
Sum	9338.5
Count	506

- The negative skew indicates that most of the areas where the houses exist, have a pt-ratio value greater than 18.455 (mean value)

H. AVG_ROOM:

AVG_ROOM	
Mean	6.284634
Standard Error	0.031235
Median	6.2085
Mode	5.713
Standard Deviation	0.702617
Sample Variance	0.493671
Kurtosis	1.8915
Skewness	0.403612
Range	5.219
Minimum	3.561
Maximum	8.78
Sum	3180.025
Count	506

- The graph of this variable tends to be leptokurtic & positively skewed one.
- Almost 50% of the houses have an average of less than 6 rooms.

I. LSTAT:

LSTAT	
Mean	12.65306
Standard Error	0.317459
Median	11.36
Mode	8.05
Standard Deviation	7.141062
Sample Variance	50.99476
Kurtosis	0.49324
Skewness	0.90646
Range	36.24
Minimum	1.73
Maximum	37.97
Sum	6402.45
Count	506

- 12.65% is the average lower status of the population but it is positively skewed.
- Lower status of the population ranges between 1.73% - 37.97%\

J. AVG_PRICE:

AVG_PRICE	
Mean	22.53281
Standard Error	0.408861
Median	21.2
Mode	50
Standard Deviation	9.197104
Sample Variance	84.58672
Kurtosis	1.495197
Skewness	1.108098
Range	45
Minimum	5
Maximum	50
Sum	11401.6
Count	506

- Average price of most of the houses is lesser than 22,500 USD.
- Average price of the houses ranges between 5,000 USD – 50,000 USD

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

Step by step approach:

- Select the range of cells (in a new sheet) in the Avg_Price column and go to insert tab.
- Click recommended charts and go to all charts.
- Select histogram.
- The histogram for the required data appears.

Histogram:

Inference:

- Most of the households hold the price range of 22-25k.
- The houses in the price range of 38-53k are comparatively less in number.
- Most of the households are suitable for middle class customers as their prices are of between 14-33k, which is again a middle range.

3) Compute the covariance matrix. Share your observations.

Step by step approach:

- Go to Data tab.
- Click data analysis.
- Click covariance.
- Select the range of cells for which the covariance matrix has to be calculated and select new worksheet.
- Click ok.
- Now, the required covariance matrix appears on a new worksheet.

Covariance matrix:

	CRIME_ RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ ROOM*	*LSTAT*	*AVG_ PRICE*
CRIME_ RATE	8.52									
AGE	0.56	790.79								
INDUS	-0.11	124.27	46.97							
NOX	0.00	2.38	0.61	0.01						
DISTANCE	-0.23	111.55	35.48	0.62	75.67					
TAX	-8.23	2397.94	831.71	13.02	1333.12	28348.62				
PTRATIO	0.07	15.91	5.68	0.05	8.74	167.82	4.68			
AVG_ ROOM	0.06	-4.74	-1.88	-0.02	-1.28	-34.52	-0.54	0.49		
LSTAT	-0.88	120.84	29.52	0.49	30.33	653.42	5.77	-3.07	50.89	
AVG_ PRICE	1.16	-97.40	-30.46	-0.45	-30.50	-724.82	-10.09	4.48	-48.35	84.42

NOTE - All the covariance values are rounded off to 2 decimals to fit the table in this sheet and the actual values are available in the excel workbook.

Observations:

- Covariance is a measure of how much two random variables change together, but it is not standardized, so it is difficult to compare covariances between different pairs of variables. However, the relationship between the variables can be determined using their signs.
- Negative sign indicate that the variables are inversely proportional to each other & positive sign indicate that the variables are directly proportional to each other.
- Negative relationship variables are,
    - INDUS & CRIME RATE
    - DISTANCE & CRIME RATE
    - TAX & CRIME RATE
    - AVG_ROOM & AGE
    - AVG_ROOM & INDUS
    - AVG_ROOM & NOX
    - AVG_ROOM & DISTANCE
    - AVG_ROOM & TAX
    - AVG_ROOM & PTRATIO
    - LSTAT & CRIME RATE
    - LSTAT & AVG_ROOM
    - AVG_PRICE & AGE
    - AVG_PRICE & INDUS
    - AVG_PRICE & NOX
    - AVG_PRICE & DISTANCE
    - AVG_PRICE & TAX
    - AVG_PRICE & PTRATIO
    - AVG_PRICE & LSTAT

- All other variables have a positive relationship between each other.

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.

Step by step approach:

- Go to Data tab.
- Click data analysis.
- Click correlation.
- Select the range of cells for which the correlation matrix must be calculated and select new worksheet.
- Click ok.
- Now, the required correlation matrix appears on a new worksheet.

Correlation matrix:

	CRIME_ RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PT RATIO*	*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				
PT RATIO	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

COLOUR MAPPING	
TOP 3 POSITIVELY CORRELATED PAIRS	
TOP 3 NEGATIVELY CORRELATED PAIRS	

NOTE - All the correlation values are rounded off to 3 decimals to fit the table in this sheet and the actual values are available in the excel workbook.

a) **Top 3 positively correlated pairs:**

1) Correlation between TAX & DISTANCE **(0.910228189)**
2) Correlation between NOX & INDUS **(0.763651447)**
3) Correlation between NOX & AGE **(0.731470104)**

b) **Top 3 negatively correlated pairs:**

1) Correlation between AVG_PRICE & LSTAT **(-0.737662726)**
2) Correlation between LSTAT & AVG_ROOM **(-0.613808272)**
3) Correlation between AVG_PRICE & PTRATIO **(-0.507786686)**

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?**

Step by step approach:

- Go to Data tab.
- Click data analysis.
- Click regression.
- Select the Y & X range of cells for which the regression must be performed.
- Select new worksheet, enable residual plot checkbox.
- Click ok.
- Now, the required regression tables appear in a new worksheet.

Regression:

SUMMARY OUTPUT

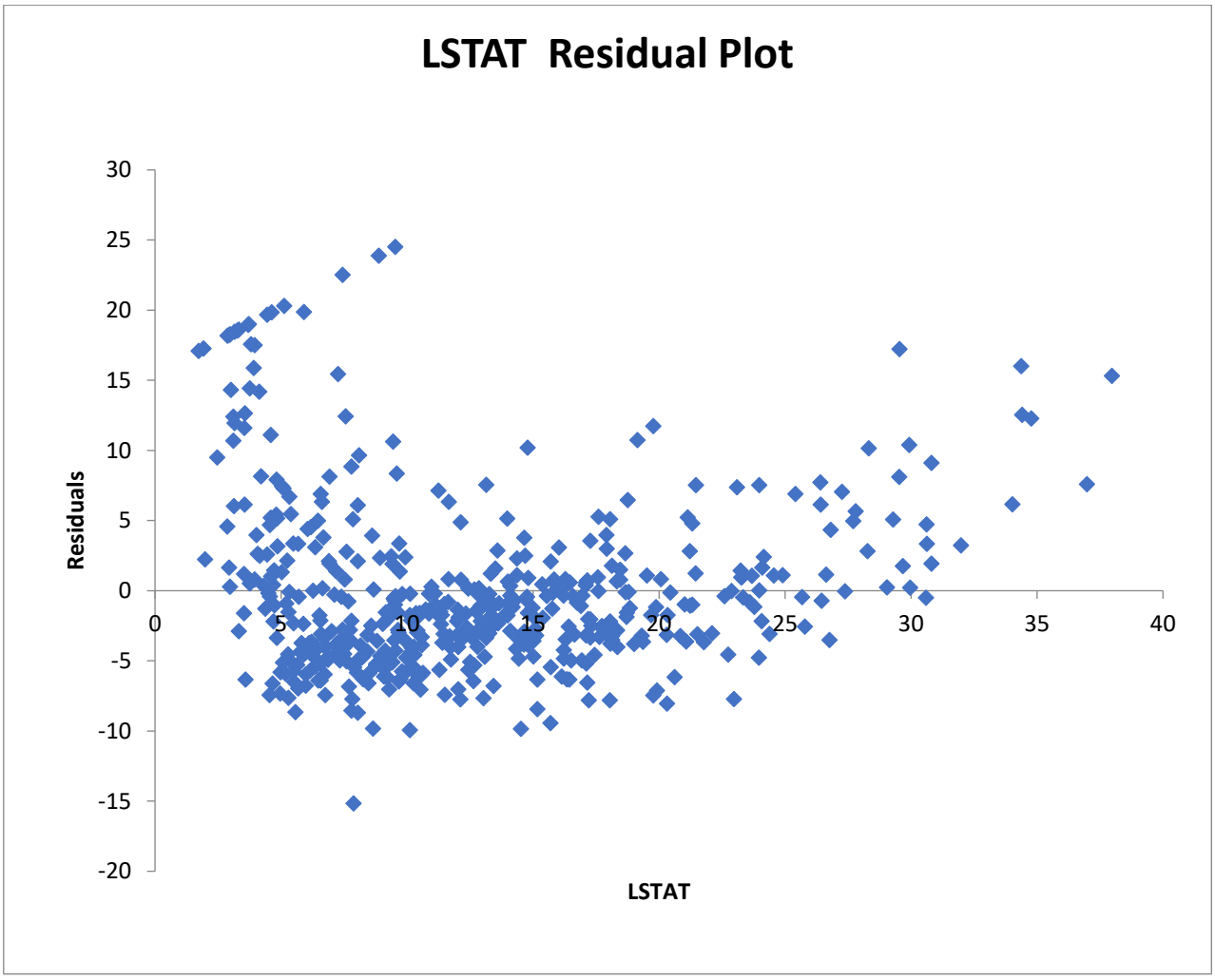
Regression Statistics	
Multiple R	0.737662726
R Square	0.544146298
Adjusted R Square	0.543241826
Standard Error	6.215760405
Observations	506

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	23243.914	23243.91	601.6179	5.0811E-88
Residual	504	19472.38142	38.63568		
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	61.4151	3.7E-236	33.4484	35.6592	33.4484	35.6592
LSTAT	-0.950049354	0.038733416	-24.5279	5.08E-88	-1.02615	-0.8739	-1.02615	-0.8739

NOTE – Few values are rounded off to 4 decimals to fit this table in this sheet and the actual values are available in the excel workbook.

Residual plot:



a) Inference:

- **Variance:** From ANOVA table, we can observe that F-value obtained from the table is lesser than the calculated F-value, hence we tend to observe the significance of LSTAT variable on the AVG_PRICE variable.
- **Intercept & coefficients values:** Intercept is however the average expected value for the response variable when the LSTAT values are equal to zero, but it's not true for every scenario. The negative coefficient indicates that as the value of LSTAT variable increase, the value of Avg_price gets decreased, which exhibits an inverse proportionality.
- **Residual plot:** Based on the residual plot, this regression model is not biased between the LSTAT values of 5-25. But below 5 & above 25, it is biased on the upper side.

b) Significance of LSTAT variable:

Chances of LSTAT variable can be either significant or insignificant. So, null & alternate hypothesis can be formulated with 95% confidence level.

H_0 : LSTAT is not significant
H_1 : LSTAT is significant

Significance level = 1 – confidence level = 1 – 0.95 = 0.05

Since the obtained p-value is lesser than the significance level, we reject the null hypothesis.
So, LSTAT variable is significant in determining 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.

Step by step approach:

- Go to Data tab.
- Click data analysis.
- Click regression.
- Select the Y & X range of cells for which the regression must be performed.
- Select new worksheet.
- Click ok.

- Now, the required regression tables appear on a new worksheet.

Regression:

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	-0.4281	0.6688	-7.5919	4.8754	-7.5919	4.8754
AVG_ROOM	5.094787984	0.4444655	11.4627	3.4E-27	4.2216	5.9680	4.2216	5.9680
LSTAT	-0.642358334	0.043731465	-14.6887	6.6E-41	-0.7283	-0.5564	-0.7283	-0.5564

NOTE – Few values are rounded off to 4 decimals to fit this table in this sheet and the actual values are available in the excel workbook.

a)

Regression equation:

AVG_PRICE =-1.358272812 + 5.094787984*(AVG_ROOM) - 0.642358334*(L_STAT)

AVG_ROOM =7, L-STAT =20, AVG_PRICE =?

Substituting these values in the regression equation, we get **$21,458.08,** which is lesser than $30,000.

In comparison with 30000 USD, the company is overcharging.

b) **Comparing the adjusted R-square value:** Adjusted R-square of current model (**0.638561606**) is greater than the adjusted R-square of previous model (**0.543241826**). Greater the R-square value, better is the model in predicting the response variable. So, the current regression model preforms better than the previous one due to the increased adjusted R-square value.

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.**

Step by step approach:

- Go to Data tab.
- Click data analysis.
- Click regression.
- Select the Y & X range of cells for which the regression must be performed.
- Select new worksheet.
- Click ok.
- Now, the required regression tables appear on a new worksheet.

Regression:

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%*
Inter-cept	29.24131526	4.817125596	6.0703	2.5E-09	19.7768	38.7058	19.7768	38.7058
CRIME_ RATE	0.048725141	0.078418647	0.6213	0.5347	-0.1053	0.2028	-0.1053	0.2028
AGE	0.032770689	0.013097814	2.5020	0.0127	0.0070	0.0585	0.0070	0.0585
INDUS	0.130551399	0.063117334	2.0684	0.0391	0.0065	0.2546	0.0065	0.2546
NOX	-10.3211828	3.894036256	-2.6505	0.0083	-17.9720	-2.6703	-17.9720	-2.6703
DIST-ANCE	0.261093575	0.067947067	3.8426	0.0001	0.1276	0.3946	0.1276	0.3946
TAX	-0.01440119	0.003905158	-3.6877	0.0003	-0.0221	-0.0067	-0.0221	-0.0067
PT RATIO	-1.074305348	0.133601722	-8.0411	6.5E-15	-1.3368	-0.8118	-1.3368	-0.8118
AVG_ ROOM	4.125409152	0.442758999	9.3175	3.8E-19	3.2555	4.9953	3.2555	4.9953
LSTAT	-0.603486589	0.053081161	-11.3691	8.9E-27	-0.7078	-0.4992	-0.7078	-0.4992

NOTE – Few values are rounded off to 4 decimals to fit this table in this sheet and the actual values are available in the excel workbook.

Explanation:

- **R-Square:** As R-square value is above 50% (0.69), this regression model can be considered a good one.
- **Intercept & coefficients values:** Intercept is however the average expected value for the response variable when all of the predictor variables are equal to zero, but its not true for every scenario. The positive coefficients indicate that as the value of predictor variable increase, the value of Avg_price also gets increased, which exhibits a direct proportionality. The negative coefficients indicate that as the value of predictor variable increase, the value of Avg_price gets decreased and vice versa, which exhibits an inverse proportionality.
- **Significance:** According to the p-value, except the CRIME_RATE variable (for which the p-value is 0.53), all other predictor variables are significant as they have a p-value lesser than 0.05.

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:

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.

Approach used:

- Go to Data tab.
- Click data analysis.
- Click regression.
- Select the Y & X range of cells for which the regression must be performed.
- Select new worksheet.
- Click ok.
- Now, the required regression tables appear on a new worksheet.

Regression:

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%*
Inter-cept	29.42847349	4.804728624	6.1249	1.8E-09	19.9884	38.8686	19.9884	38.8686
AGE	0.03293496	0.013087055	2.5166	0.0122	0.0072	0.0586	0.0072	0.0586
INDUS	0.130710007	0.063077823	2.0722	0.0388	0.0068	0.2546	0.0068	0.2546
NOX	-10.27270508	3.890849222	-2.6402	0.0085	-17.9172	-2.6282	-17.9172	-2.6282
DIST-ANCE	0.261506423	0.067901841	3.8512	0.0001	0.1281	0.3949	0.1281	0.3949
TAX	-0.014452345	0.003901877	-3.7039	0.0002	-0.0221	-0.0068	-0.0221	-0.0068
PT RATIO	-1.071702473	0.133453529	-8.0305	7.08E-15	-1.3339	-0.8095	-1.3339	-0.8095
AVG_ ROOM	4.125468959	0.44248544	9.3234	3.68E-19	3.2561	4.9948	3.2561	4.9948
LSTAT	-0.605159282	0.0529801	-11.4224	5.41E-27	-0.7093	-0.5011	-0.7093	-0.5011

NOTE – Few values are rounded off to 4 decimals to fit this table in this sheet and the actual values are available in the excel workbook.

a) Interpreting the output:

- **Difference from the previous model:** After removing the CRIME_RATE variable, there is a slight difference of significance in this regression model in terms of
  - **Decreased Standard error (individual as well as combined)**
  - **Increased R-square**
  - **Increased intercept & coefficients values.**
  - **Decreased p-values**

b) Comparing the adjusted R-square value:

Adjusted R-square of current model (**0.688683682**) is greater than the adjusted R-square of previous model (**0.688298647**). Greater the R-square value, better is the model in predicting the response variable. So, the current regression model preforms better than the previous one due to the increased adjusted R-square value.

c) Coefficients in ascending order

Variable	Co-efficient
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

If the value of NOX is more in a locality in this town, then the average price will get decreased since it has a negative coefficient value in this regression model, which exhibits an inverse proportionality.

d) Regression equation

AVG_PRICE = 29.42847349 + 0.03293496*(AGE) + 0.130710007*(INDUS) - 10.27270508*(NOX) + 0.261506423*(DISTANCE) - 0.014452345*(TAX) - 1.071702473*(PTRATIO) + 4.125468959*(AVG_ROOM) - 0.605159282*(LSTAT)

****

NOTE:

Kindly refer to the **Terro's_REA.xlsx (excel)** file for detailed statistics, submitted along with this report.