DATA ANALYSIS REPORT

TERRO'S REAL ESTATE

Name: **Venkatesh Kulkarni**

Batch: **Data Analysis Sept 23**

TABLE OF CONTENTS

	b) Is LSTAT variable significant for the analysis based on your model?	
10	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 14(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	13
11	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 Rsquare, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE.	15
12	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.	16
13	Summary	18

List of Figures

List of Tables

Executive Summary:

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.

Introduction :

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. This exercise involves implementation of exploratory data analysis helps understand the nature of different data-attributes. This exercise helps to understand how to use various statistical/analytical tools in MS Excel like Summary statistics, Histogram, correlation table, Regression analysis (using Data analysis tool pack).

Data Description

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

Table 1: Data Description

Sample of the Data Set

The Data Set contains description of 506 houses in Boston with 10 attributes. And each attributes description is mentioned above in data description

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
5.53	69.5	8.14	0.538	4	307	21	5.727	11.28	18.2
8.39	98.1	8.14	0.538	4	307	21	5.57	21.02	13.6
8.96	89.2	8.14	0.538	4	307	21	5.965	13.83	19.6
9.61	91.7	8.14	0.538	4	307	21	6.142	18.72	15.2
2.8	100	8.14	0.538	4	307	21	5.813	19.88	14.5

Table 2: Sample of dataset

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

CRIME_RATE		AGE		INDUS		NOX		DISTANCE	
Mean	4.871976285	Mean	68.5749	Mean	11.13678	Mean	0.554695	Mean	9.549407115
Standard Error	0.129860152	Standard Error	1.25137	Standard Error	0.30498	Standard Error	0.005151	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	28.14886	Standard Deviation	6.860353	Standard Deviation	0.115878	Standard Deviation	8.707259384
Sample Variance	8.533011532	Sample Variance	792.3584	Sample Variance	47.06444	Sample Variance	0.013428	Sample Variance	75.81636598
Kurtosis	-1.189122464	Kurtosis	-0.96772	Kurtosis	-1.23354	Kurtosis	-0.06467	Kurtosis	-0.867231994
Skewness	0.021728079	Skewness	-0.59896	Skewness	0.295022	Skewness	0.729308	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

TAX		PTRATIO		AVG_ROOM		LSTAT		AVG_PRICE	
Mean	408.2372	Mean	18.45553	Mean	6.284634387	Mean	12.65306	Mean	22.53281
Standard Error	7.492389	Standard Error	0.096244	Standard Error	0.031235142	Standard Error	0.317459	Standard Error	0.408861
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	168.5371	Standard Deviation	2.164946	Standard Deviation	0.702617143	Standard Deviation	7.141062	Standard Deviation	9.197104
Sample Variance	28404.76	Sample Variance	4.686989	Sample Variance	0.49367085	Sample Variance	50.99476	Sample Variance	84.58672
Kurtosis	-1.14241	Kurtosis	-0.28509	Kurtosis	1.891500366	Kurtosis	0.49324	Kurtosis	1.495197
Skewness	0.669956	Skewness	-0.80232	Skewness	0.403612133	Skewness	0.90646	Skewness	1.108098
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

Table 3: Summery Statistics

Inferences:

- The number of records in given data set are 506, count for all variables in data set are also 506 which is there are no missing values in any variable.
- From the skewness variables most of the variables are highly skewed but Crime_rate is very nearly to 0 so we can say that it is normally distributed.
- Negative skewness indicates that left tailed distribution, Positive skewness indicates right tailed distribution. Here in above table Age and PTRATIO has negative skewness we can say they are left tailed, aside from these all are right tailed distributions.
- Negative kurtosis signifies flatter curve in all variables.
- Positive kurtosis says it's a sharp curve than normal curve saying more values are concentrated near to median.
- Crime Rate 50% of the crime rate is below 4.82 and 50% above this value.
- Age of houses on an average of 68 years from the built year.
- INDUS On an average of11.13 % of property belongs to non-retail business
- NOX On an average nitric oxide concentration around 0.55 ppm
- Distance from highway is around average of 9.5 miles.
- Tax rate is average of $408
- PTRATIO on an average, pupil teacher ratio is 18 for 506 houses.
- In a house there are average of 6 rooms are in there.
- From LSTAT 12% of the population has lower status
- Average price of houses values around 22.5k USD.

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

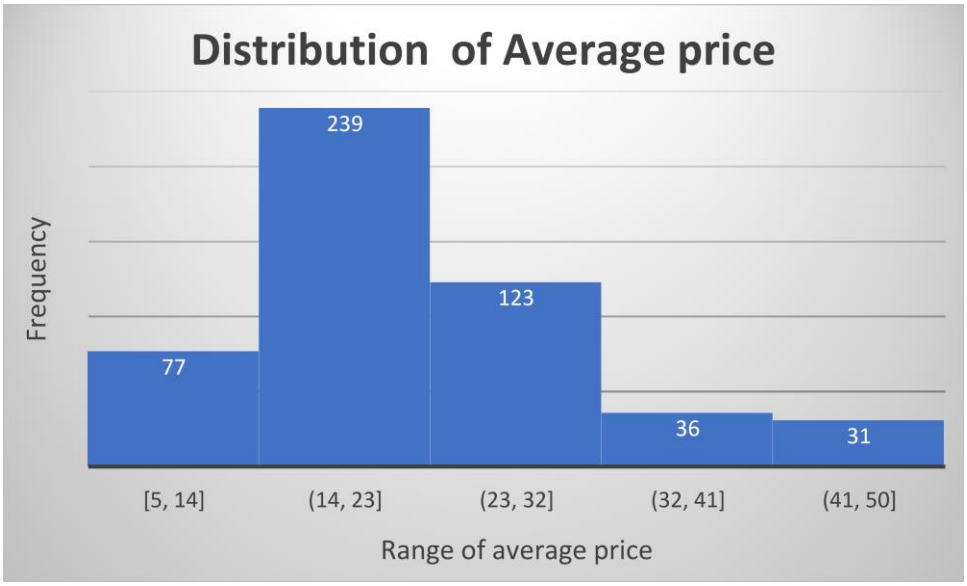

Fig.1: Histogram for Avg_Price

Inferences:

- From above Histogram the data is not a perfect bell we can say that the data is positively skwed or right skewed data.
- Most of the houses are from range of $14000 to $23000.
- The price in between 14k to 23k is the affordable average price for a most people.
- We have Least count of house ranges from $41000 to $50000.

Q.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	790.7925								
INDUS	-0.110215175	124.2678	46.97143							
NOX	0.000625308	2.381212	0.605874	0.013401						
DISTANCE	-0.229860488	111.55	35.47971	0.61571	75.66653					
TAX	-8.229322439	2397.942	831.7133	13.0205	1333.117	28348.62				
PTRATIO	0.068168906	15.90543	5.680855	0.047304	8.743402	167.8208	4.677726296			
AVG_ROOM	0.056117778	-4.74254	-1.88423	-0.02455	-1.28128	-34.5151	-0.539694518	0.492695216		
LSTAT	-0.882680362	120.8384	29.52181	0.48798	30.32539	653.4206	5.771300243	-3.073654967	50.89398	
AVG_PRICE	1.16201224	-97.3962	-30.4605	-0.45451	-30.5008	-724.82	-10.09067561	4.484565552	-48.3518	84.41955616

Table 4: Covariance Matrix

Inferences:

- From the above matrix diagonal numbers are Variabilities and others are co variables between features.
- Nagative value between two variables means they are inversely proportional to each other.
- Positive value between two variables means they are directly proportional to each other.
- In above table all cell filled with green are directly proportional to each other, and filled with red are inversely proportional to each other.
- From above chart or matrix we can assume that Tax variable has high covariance between all features except Crime rate .
- That means Tax has a good variability between other features.

Q.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	0.644779	1							
NOX	0.001850982	0.73147	0.763651	1						
DISTANCE	-0.009055049	0.456022	0.595129	0.611441	1					
TAX	-0.016748522	0.506456	0.72076	0.668023	0.910228	1				
PTRATIO	0.010800586	0.261515	0.383248	0.188933	0.464741	0.460853	1			
AVG_ROOM	0.02739616	-0.240265	-0.391676	-0.302188	-0.209847	-0.292048	-0.355501	1		
LSTAT	-0.042398321	0.602339	0.6038	0.590879	0.488676	0.543993	0.374044	-0.613808272	1	
AVG_PRICE	0.043337871	-0.376955	-0.483725	-0.427321	-0.381626	-0.468536	-0.507787	0.695359947	-0.737662726	1

Table 5 : Correlation Matrix

a) From above correlation Matrix top 3 positively correlated pairs are

- DISTANCE – TAX
- INDUS – NOX
- AGE – NOX

b) From above correlation Matrix top 3 Negatively correlated pairs are

- LSTAT - AVG_PRICE
- AVG_ROOM – LSTAT
- PTRATIO - AVG_PRICE

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?

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	601.6178711	5.0811E-88			
Residual	504	19472.38142	38.63567742					
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	61.41514552	3.7431E-236	33.44845704	35.65922472	33.448457	35.65922472
LSTAT	-0.950049354	0.038733416	-24.52789985	5.0811E-88	-1.0261482	-0.873950508	-1.0261482	-0.87395051

Table 6: Regression model for LSTAT

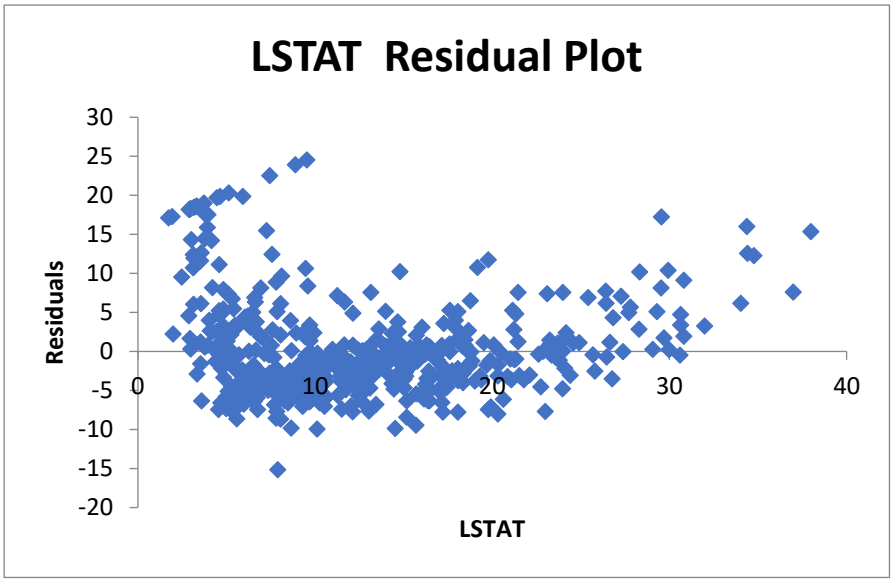

Fig 2: LSTAT Residual plot

Inferences:

A)

- R is correlation coefficient value greater the value greater the correlation between Avg_Price and LSTAT.
- Adjusted R square = 54% of variance of avg_Price is explained by the LSTAT.
- The coefficient of LSTAT for the model is -0.950049354.That means LSTAT and AVG_price are inversely proportional to each other.
- If LSTAT increases 0.9 times Avg_Price decreases by 0.9 times.
- Intercept of LSTAT for the model is 34.55384088.

B)

- Residual is nothing but a difference between actual Avg_Price and the predicted Avg_Price.
- For LSTAT P value is (5.08E-88) less than alpha (0.05) so the we reject null hypothesis.
- By this we can say that LSTAT is a significant variable according to this model.

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							
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

Table 7: Regression model for LSTAT and Avg_Room

a)

Regression Equation we obtained for this model is Y = -1.358 +5.09X0 -0.642 X1

Where Y=Avg_price

X0 = avg_room

X1 = LSTAT As per the model,

avg_price for new house can be calculated as

Y= -1.358 + 5.09(7) -0.642(20) = 21.44

From the above solved problem predicted price is 21.4K USD compare to the company quoting a value of 30K USD for this locality they are overcharging.

b)

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

Regression statistics of Q5

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

Regression statistics of Q6

- With this model we can say that nearly 64% of variability for AVG_Price is explained by Avg_Room and LSTAT.
- In previous model 54% of variability is explained by LSTAT alone, So we can say that this model is better at explaining Avg_price or dependent variable than previous one(Q5)

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.

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	124.904505	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	2.5398E-09	19.77682784	38.70580267	19.77682784	38.70580267
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.5346572	-0.10534854	0.202798827	-0.10534854	0.202798827
AGE	0.032770689	0.013097814	2.501996817	0.01267044	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	-2.6505102	0.00829386	-17.9720228	-2.67034281	-17.9720228	-2.67034281
DISTANCE	0.261093575	0.067947067	3.842602576	0.00013755	0.127594012	0.394593138	0.127594012	0.394593138
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	9.317504929	3.8929E-19	3.255494742	4.995323561	3.255494742	4.995323561
LSTAT	-0.603486589	0.053081161	-11.3691294	8.9107E-27	-0.70777824	-0.49919494	-0.70777824	-0.49919494

Table 8: Regression model for all variables

Inferences:

- From this regression model we can say that Crime_rate is not a significant variable for Avg_Price since its P-value is greater than 0.05.
- Excluding Crime_Rate all other variables are Significant variables for Avg_Price since there P-value is less than alpha(0.05).
- From adjusted R square we can say that all features together explains 69% of variability for Avg_Price.
- For coefficients and intercepts NOX,TAX,PTRATIO,LSTAT have negative value they are inversely proportional to Avg_price, if Avg_Price increases they decreases and vice-versa.

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.

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	140.643041	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%*
NOX	-10.27270508	3.890849222	-2.64022184	0.00854572	-17.9172457	-2.62816447	-17.9172457	-2.62816447
PTRATIO	-1.071702473	0.133453529	-8.03052927	7.0825E-15	-1.33390511	-0.80949984	-1.33390511	-0.80949984
LSTAT	-0.605159282	0.0529801	-11.4223884	5.4184E-27	-0.70925186	-0.5010667	-0.70925186	-0.5010667
TAX	-0.014452345	0.003901877	-3.70394641	0.00023607	-0.02211855	-0.00678614	-0.02211855	-0.00678614
AGE	0.03293496	0.013087055	2.516605952	0.01216288	0.007222187	0.058647734	0.007222187	0.058647734
INDUS	0.130710007	0.063077823	2.072202264	0.03876167	0.006777942	0.254642071	0.006777942	0.254642071
DISTANCE	0.261506423	0.067901841	3.851242024	0.00013289	0.128096375	0.394916471	0.128096375	0.394916471
AVG_ROOM	4.125468959	0.44248544	9.323400461	3.6897E-19	3.256096304	4.994841615	3.256096304	4.994841615
Intercept	29.42847349	4.804728624	6.124898157	1.846E-09	19.98838959	38.8685574	19.98838959	38.8685574

Table 9: Regression model for all significant variables

Inferences:

A) From this regression model we can say that all the variables are significant for Avg_price of house as there all P-value are less than 0.05

B) By comparing this model and previous one(Q7) adjusted R square value neary same for both that is 69%, SO we can conclude that both the models perform well.

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

Regression statistics of Q7

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

Regression statistics of Q8

C) By sorting values of coefficients in ascending order if NOX is more in the locality Avg_Price of the houses will decrease by 10 times.

D) Regression equation:

Y=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 +29.42847349

Summary:

The above analysis concludes that all the features play a vital role in estimating the average price of the house excluding crime rate and few features have negative coefficients which say that increase rate in those features will decrease the average price of the house like NOX, PTRATIO, TAX and LSTAT.