PROJECT - TERRO'S REAL ESTATE DATA ANALYSIS

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:

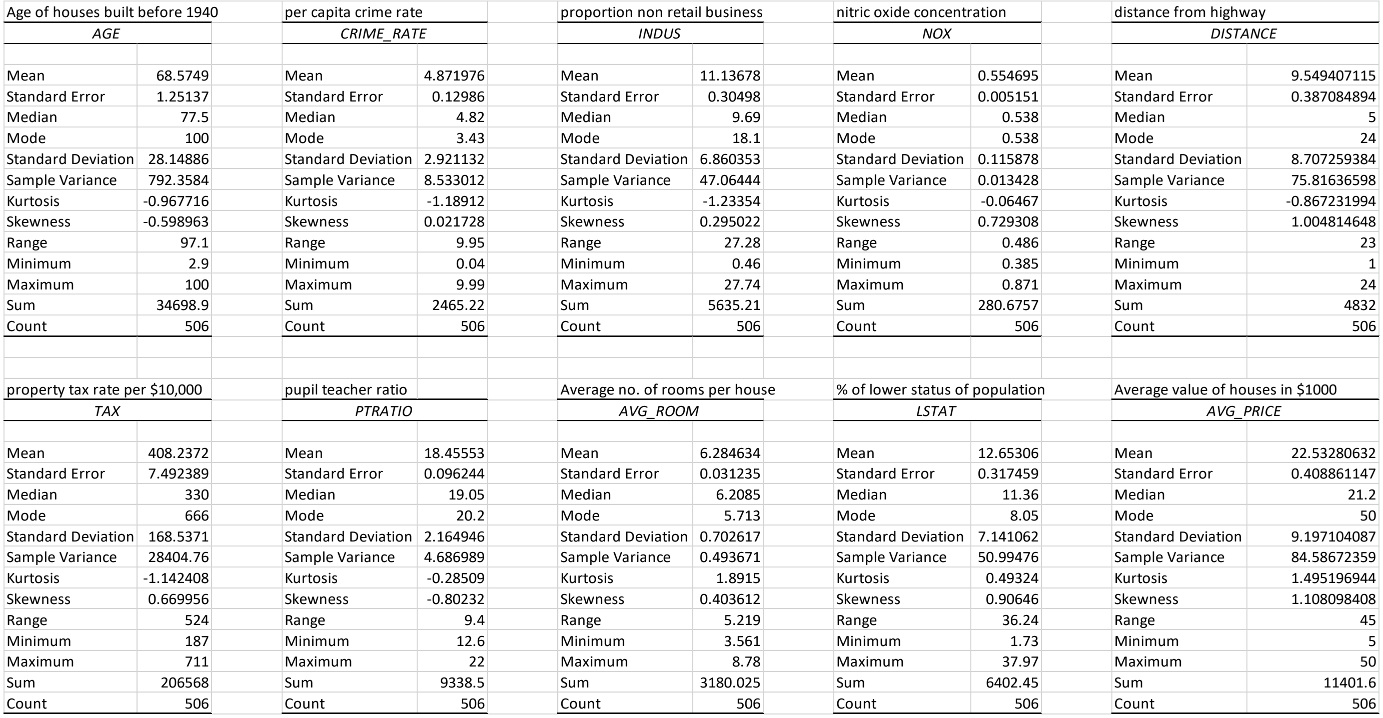
Attribute	Description
CRIME RATE	per capita crime rate by town
INDUSTRY	proportion of non-retail business acres per town (in percentage terms)
NOX	nitric oxides concentration (parts per 10 million)
AVG_ROOM	average number of rooms per house
AGE	proportion of houses built prior to 1940 (in percentage 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.

To do the analysis, you are expected to solve these questions:

1) Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation. (5 marks)



Age of houses built before 1940		per capita crime rate		proportion non retail business		nitric oxide concentration		distance from highway	
AGE		*CRIME_RATE*		*INDUS*		*NOX*		*DISTANCE*	
Mean	68.5749	Mean	4.871976	Mean	11.13678	Mean	0.554695	Mean	9.549407115
Standard Error	1.25137	Standard Error	0.12986	Standard Error	0.30498	Standard Error	0.005151	Standard Error	0.387084894
Median	77.5	Median	4.82	Median	9.69	Median	0.538	Median	5
Mode	100	Mode	3.43	Mode	18.1	Mode	0.538	Mode	24
Standard Deviation	28.14886	Standard Deviation	2.921132	Standard Deviation	6.860353	Standard Deviation	0.115878	Standard Deviation	8.707259384
Sample Variance	792.3584	Sample Variance	8.533012	Sample Variance	47.06444	Sample Variance	0.013428	Sample Variance	75.81636598
Kurtosis	-0.967716	Kurtosis	-1.18912	Kurtosis	-1.23354	Kurtosis	-0.06467	Kurtosis	-0.867231994
Skewness	-0.598963	Skewness	0.021728	Skewness	0.295022	Skewness	0.729308	Skewness	1.004814648
Range	97.1	Range	9.95	Range	27.28	Range	0.486	Range	23
Minimum	2.9	Minimum	0.04	Minimum	0.46	Minimum	0.385	Minimum	1
Maximum	100	Maximum	9.99	Maximum	27.74	Maximum	0.871	Maximum	24
Sum	34698.9	Sum	2465.22	Sum	5635.21	Sum	280.6757	Sum	4832
Count	506	Count	506	Count	506	Count	506	Count	506

property tax rate per $10,000		pupil teacher ratio		Average no. of rooms per house		% of lower status of population		Average value of houses in $1000	
TAX		*PTRATIO*		*AVG_ROOM*		*LSTAT*		*AVG_PRICE*	
Mean	408.2372	Mean	18.45553	Mean	6.284634	Mean	12.65306	Mean	22.53280632
Standard Error	7.492389	Standard Error	0.096244	Standard Error	0.031235	Standard Error	0.317459	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	168.5371	Standard Deviation	2.164946	Standard Deviation	0.702617	Standard Deviation	7.141062	Standard Deviation	9.197104087
Sample Variance	28404.76	Sample Variance	4.686989	Sample Variance	0.493671	Sample Variance	50.99476	Sample Variance	84.58672359
Kurtosis	-1.142408	Kurtosis	-0.28509	Kurtosis	1.8915	Kurtosis	0.49324	Kurtosis	1.495196944
Skewness	0.669956	Skewness	-0.80232	Skewness	0.403612	Skewness	0.90646	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

The summary statistics for each variable in the table is as shown above.

From the output it can be observed that:

- AGE Variable: The mean percentage of houses built before 1940 is 68.57 and the median is greater than mean, it shows that the distribution is left skewed and hence majority of the data is on the right side of mean. The mode is 100, which means most frequently the localities have all houses built before 1940. The distribution is moderately left skewed as the value is between 0.5 and 1. The kurtosis value is less than zero so it is platykurtic.

- CRIME_RATE Variable: The mean per capita crime rate by town is 4.87 and the median is 4.82, which shows that the distribution is almost symmetrical. So, half of the distribution has crime rate below 4.82 and half above, with the maximum being 9.99. The negative kurtosis indicates that the distribution is platykurtic.

- INDUS Variable: The mean percentage of proportion of non-retail business acres per town is 11.13 and the median is 9.69. The skewness is less than 0.5 hence it is almost symmetrical. The negative kurtosis indicates that the distribution is platykurtic.

- NOX Variable: The mean concentration of nitric oxide is 0.55 parts per 10 million. The median is less than mean, so the distribution is right skewed and the skewness is moderate. The mode and median are same, so most frequently the localities have nitric oxide concentration of 0.538 parts per 10 million. The kurtosis value is less than zero so it is platykurtic.

- DISTANCE Variable: The average distance from highway is 9.54 miles. The median is 5 with standard deviation of 8.7. The mode and maximum values are same and are equal to 24. So, the most frequent data is 24 miles. The value of skewness shows that the distribution is highly right skewed and the kurtosis value is less than zero so it is platykurtic. Most of the properties are situated at a distance of less than 9.54 miles from highway as the distribution is highly right skewed.

- TAX Variable: The mean full value property tax rate per 10,000 USD is 408.23 USD. The median is lesser than mean, so the distribution is moderately right skewed. The standard deviation is very high which shows that the data is more spread out. The mode is 666 USD which is the most frequent tax amount for the properties. It is platykurtic.

- PTRATIO Variable: The mean pupil teacher ratio by town is 18.45 and the median is greater than mean, so the distribution is moderately left skewed and is platykurtic. The standard deviation is 2.16, so 68% of the localities has pupil teacher ratio in the range of 16.29 and 20.61. The mode is 20.2, so most frequently the localities have pupil teacher ratio of 20.2.

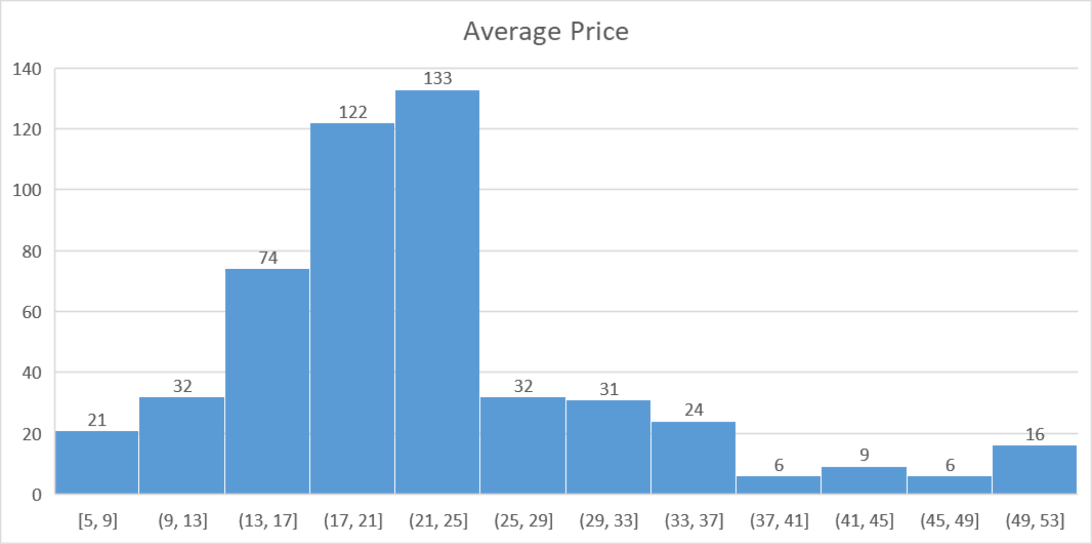
- AVG_ROOM Variable: The mean of average number of rooms per house is 6.28 and the median is 6.20. The mode is 5.71, so most frequently the houses have 5.71 average number of rooms. The distribution is moderately right skewed and is leptokurtic.

- LSTAT Variable: The mean percentage of lower status of population is 12.65% and the median is lesser than mean. The mode is 8.05, so most frequently the localities have 8.05 percentage of lower status of population. The distribution is highly right skewed and is leptokurtic in nature. The maximum value is 37.97 and minimum is 1.37, so the range is large. Therefore, some localities have very low percentage of lower status population and some localities have very high percentage.

- AVG_PRICE Variable: The mean Average Price of houses is 22.53K USD and the median is lesser than mean. The distribution is highly right skewed, so average price of most houses lies to the left of mean value that is less than mean value. It is leptokurtic distribution as kurtosis value is positive.

So, it can be concluded that the summary statistics gives us information about the distribution of data in terms of central tendency (mean, median, mode), variability (standard deviation) and shape (symmetry and peak).

2) Plot a histogram of the Avg_Price variable. What do you infer? (5 marks)



From the histogram we can observe the shape of the distribution of Avg_Price variable. It can be observed from the histogram plot that most frequent Average price value is between 17K-25K USD. So, most of the houses have average price between 17,000-25,000 USD. These are the bins having high frequency of data distribution with 255 out of 506 observations, which indicates that half of the houses have Average Price in the range of 17K-25K USD. Thus, it becomes the most common value for average price of houses. The least common values are found at the higher values of over 37K-41K USD and 45K-49K USD.

From the histogram we can say that the distribution is positively skewed (right skewed) as most of values are present to the left of mean (near the lower end of range) and is right tailed (higher values are infrequent) and hence asymmetrical and has a sharp peak or is leptokurtic.

3) Compute the covariance matrix. Share your observations. (5 marks)

	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

Covariance basically signifies the direction of linear relationship between two variables. The direction usually refers to whether the variables vary directly or inversely to each other. The covariance matrix is as shown above. The observation from the covariance matrix is that the variables share either positive covariance or negative covariance between them. If covariance is zero it means that there is no directional relationship between the two variables. From the covariance matrix, to name a few, AGE and CRIME_RATE, INDUS and AGE, NOX and AGE, DISTANCE and AGE, TAX and AGE variables share a positive covariance relationship that is both the variables move in same direction. Also, to name a few, INDUS and CRIME_RATE, AVG_ROOM and AGE, LSTAT and AVG_ROOM variables share negative relationship or inverse relationship, that is if one variable increases, then the other decreases. There is no directional relationship between NOX and CRIME_RATE variables as covariance coefficient is equal to zero.

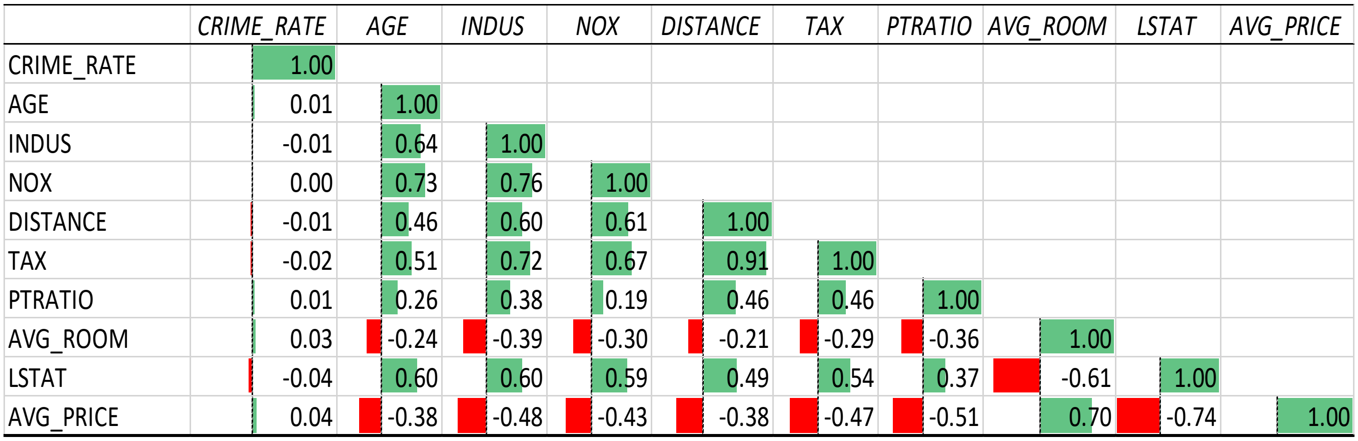
The AVG_PRICE variable has positive covariance with CRIME_RATE and AVG_ROOM variables.

The AVG_PRICE variable has negative covariance with AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO and LSTAT variables.

4) Create a correlation matrix of all the variables (Use Data analysis tool pack). (5 marks)

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	1.00									
AGE	0.01	1.00								
INDUS	-0.01	0.64	1.00							
NOX	0.00	0.73	0.76	1.00						
DISTANCE	-0.01	0.46	0.60	0.61	1.00					
TAX	-0.02	0.51	0.72	0.67	0.91	1.00				
PTRATIO	0.01	0.26	0.38	0.19	0.46	0.46	1.00			
AVG_ROOM	0.03	-0.24	-0.39	-0.30	-0.21	-0.29	-0.36	1.00		
LSTAT	-0.04	0.60	0.60	0.59	0.49	0.54	0.37	-0.61	1.00	
AVG_PRICE	0.04	-0.38	-0.48	-0.43	-0.38	-0.47	-0.51	0.70	-0.74	1.00

Correlation basically measures both strength and direction of a relationship between two variables. The larger the absolute value of coefficient, the stronger the relationship. Positive or negative sign indicates the direction of the relationship. If it is positive means, both variables move in same direction and if negative means they move in opposite directions. The correlation matrix is as shown above.

a) From the correlation matrix it can be observed that the top 3 positively correlated pairs are:

1. TAX and DISTANCE (91%)
2. NOX and INDUS (76%)
3. NOX and AGE (73%)

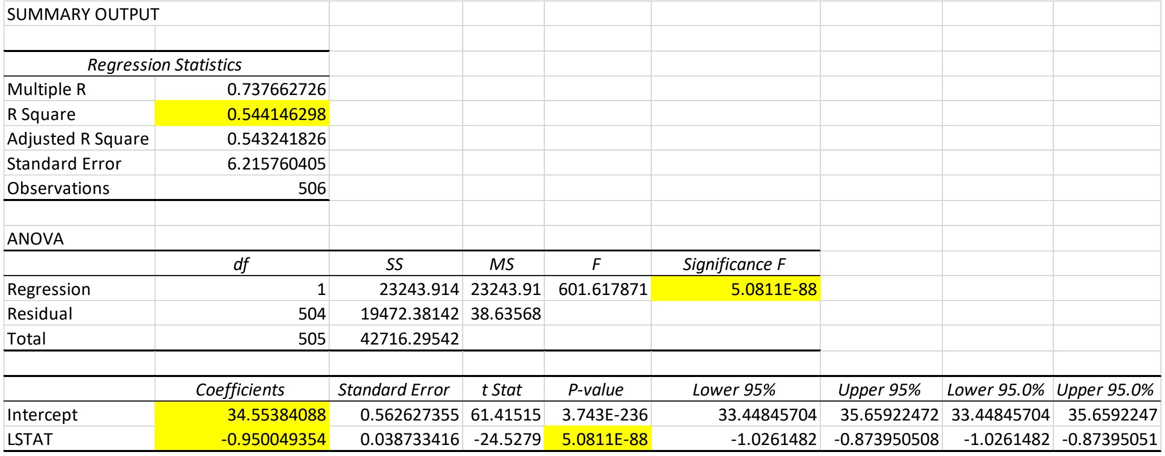
b) From the correlation matrix it can be observed that the top 3 negatively correlated pairs are:

1. AVG_PRICE and LSTAT (-74%)
2. LSTAT and AVG_PRICE (-61%)
3. AVG_PRICE and PTRATIO (-51%)

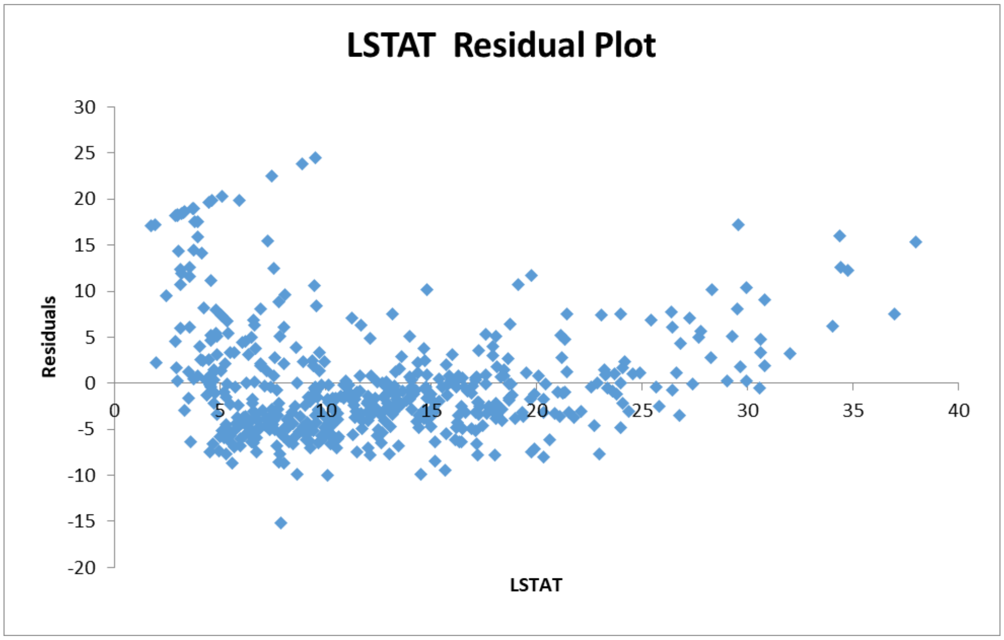
5) Build an initial regression model with AVG_PRICE as ‘y’ (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot. (8 marks)

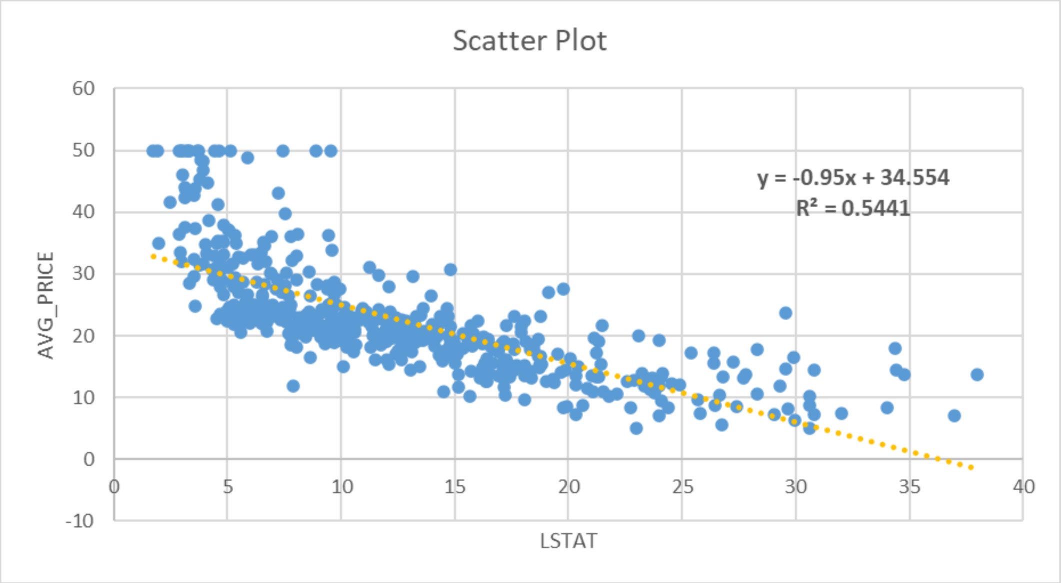
a) What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?

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



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	601.617871	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	61.41515	3.743E-236	33.44845704	35.65922472	33.44845704	35.6592247
LSTAT	-0.950049354	0.038733416	-24.5279	5.0811E-88	-1.0261482	-0.873950508	-1.0261482	-0.87395051





A Regression model is built with LSTAT as independent variable and AVG_PRICE as dependent variable. Also, a scatter plot is plotted for these two variables with linear trendline, equation and R square as shown in above image.

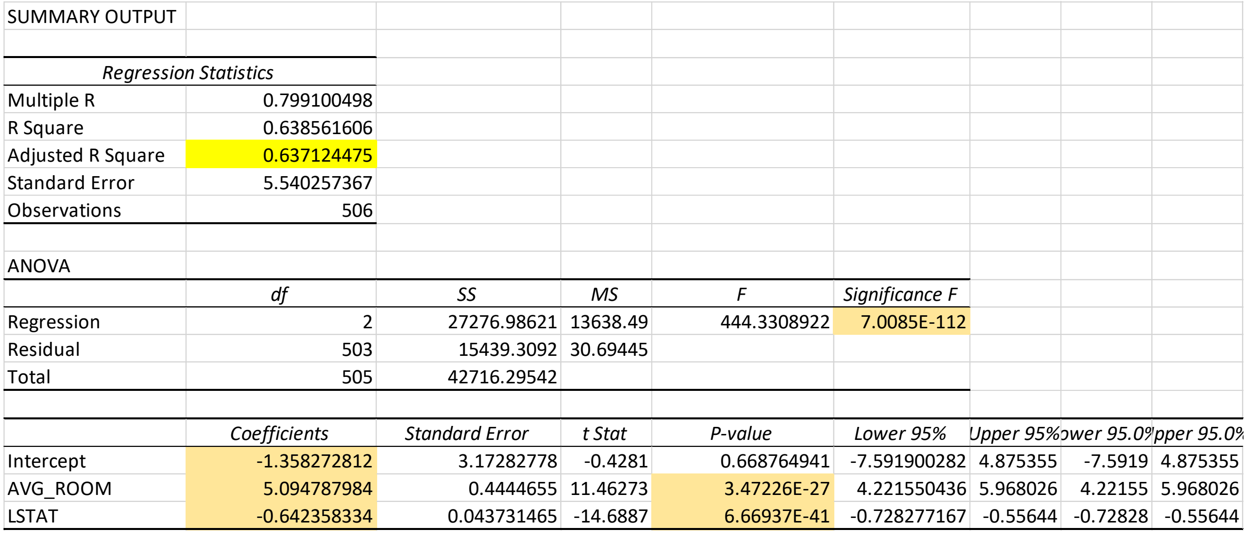
a) From the Regression summary output, it can be inferred that R Square value is 0.54 which means that 54% of variance in the dependent variable (AVG_PRICE) is explained by the independent variable (LSTAT) in this model. The coefficient value of LSTAT is -0.95 which means that with increase in value of LSTAT by 1 there is decrease in value of AVG_PRICE by 0.95. The intercept value is 34.55 which means that when LSTAT is zero, the value of AVG_PRICE is 34.55K USD. The residual plot indicates that the residual data points are randomly distributed around the x axis and are not following any particular pattern. So, the linear model is an appropriate model.

b) Yes, the LSTAT variable is significant for analysis as the p value which is 5.08E-88 is lower than 0.05, thus eliminating the null hypothesis for this variable.

6) Build a new Regression model including LSTAT and AVG_ROOM together as independent variables and AVG_PRICE as dependent variable. (6 marks)

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	13638.49	444.3308922	7.0085E-112			
Residual	503	15439.3092	30.69445					
Total	505	42716.29542						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-1.358272812	3.17282778	-0.4281	0.668764941	-7.591900282	4.875355	-7.5919	4.875355
AVG_ROOM	5.094787984	0.4444655	11.46273	3.47226E-27	4.221550436	5.968026	4.22155	5.968026
LSTAT	-0.642358334	0.043731465	-14.6887	6.66937E-41	-0.728277167	-0.55644	-0.72828	-0.55644

The regression model is built using LSTAT and AVG_ROOM as independent variables and AVG_PRICE as dependent variable. The summary output is as shown above.

a) From the Regression summary output, it can be inferred that the regression equation is

   AVG_PRICE = 5.09478 * AVG_ROOM - 0.64235 * LSTAT - 1.35827

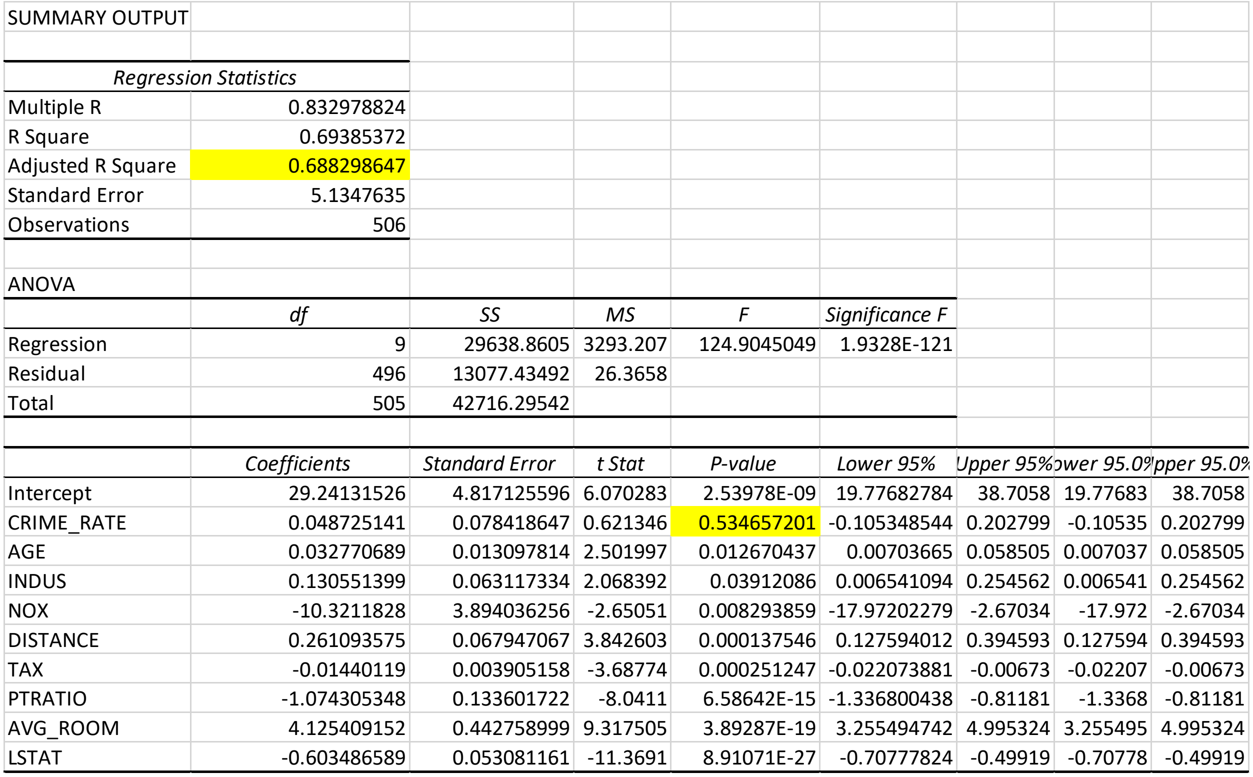
   If a new house has 7 rooms on average and LSTAT value of 20 means the AVG_PRICE will be

   AVG_PRICE = 5.09478 * **7** - 0.64235 * **20** - 1.35827 **= 21.4580K** or **21458 USD**

   If the company is quoting a value of 30000 USD for this locality, then the company is **overcharging** as the AVG_PRICE is 21458 USD as per this model.

b) Yes, the performance of this model is better than the previous model as the Adjusted R Square value of this model is 0.63 or 63% which is higher than 54% obtained in the previous model as this model can explain 63% of variance in AVG_PRICE by the independent variables AVG_ROOM and LSTAT. Also, the p value of both the variables are less than 0.05 and are significant variables for the model.

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. (8 marks)



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	3293.207	124.9045049	1.9328E-121
Residual	496	13077.43492	26.3658		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	29.24131526	4.817125596	6.070283	2.53978E-09	19.77682784	38.7058	19.77683	38.7058
CRIME_RATE	0.048725141	0.078418647	0.621346	0.534657201	-0.105348544	0.202799	-0.10535	0.202799
AGE	0.032770689	0.013097814	2.501997	0.012670437	0.00703665	0.058505	0.007037	0.058505
INDUS	0.130551399	0.063117334	2.068392	0.03912086	0.006541094	0.254562	0.006541	0.254562
NOX	-10.3211828	3.894036256	-2.65051	0.008293859	-17.97202279	-2.67034	-17.972	-2.67034
DISTANCE	0.261093575	0.067947067	3.842603	0.000137546	0.127594012	0.394593	0.127594	0.394593
TAX	-0.01440119	0.003905158	-3.68774	0.000251247	-0.022073881	-0.00673	-0.02207	-0.00673
PTRATIO	-1.074305348	0.133601722	-8.0411	6.58642E-15	-1.336800438	-0.81181	-1.3368	-0.81181
AVG_ROOM	4.125409152	0.442758999	9.317505	3.89287E-19	3.255494742	4.995324	3.255495	4.995324
LSTAT	-0.603486589	0.053081161	-11.3691	8.91071E-27	-0.70777824	-0.49919	-0.70778	-0.49919

The regression model is built using CRIME_RATE, AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG_ROOM and LSTAT as independent variables and AVG_PRICE as dependent variable.

From the Regression summary output, it can be inferred that the Adjusted R Square value is 0.6882 or 68.82% which is higher than the previous two regression models. Thus 68% of variance in AVG_PRICE is explained by the independent variables in this model.

The intercept value is 29.24 which means that the AVG_PRICE is 29.24 USD when the value of all the independent variables is zero.

It is observed that the coefficients of CRIME_RATE, AGE, INDUS, DISTANCE and AVG_ROOM are positive so whenever there is increase in value of these variables, the AVG_PRICE also increases accordingly. On the other hand, it is observed that the coefficients of NOX, TAX, PTRATIO and LSTAT are negative and hence the AVG_PRICE decreases with increase in the value of these variables.

It is observed from the p values of variables that except for CRIME_RATE all the other independent variables have a value less than 0.05 and hence are significant variables with respect to AVG_PRICE. The p value of CRIME_RATE is 0.5346 which is greater than 0.05, so the null hypothesis cannot be rejected for this particular variable and thus becomes an insignificant variable in this model.

So, the significant variables with respect to AVG_PRICE are AGE, INDUS, DISTANCE, NOX, TAX, PTRATIO, LSTAT and AVG_ROOM as per this model as in case of these variables the null hypothesis is rejected and hence, they are significant variables.

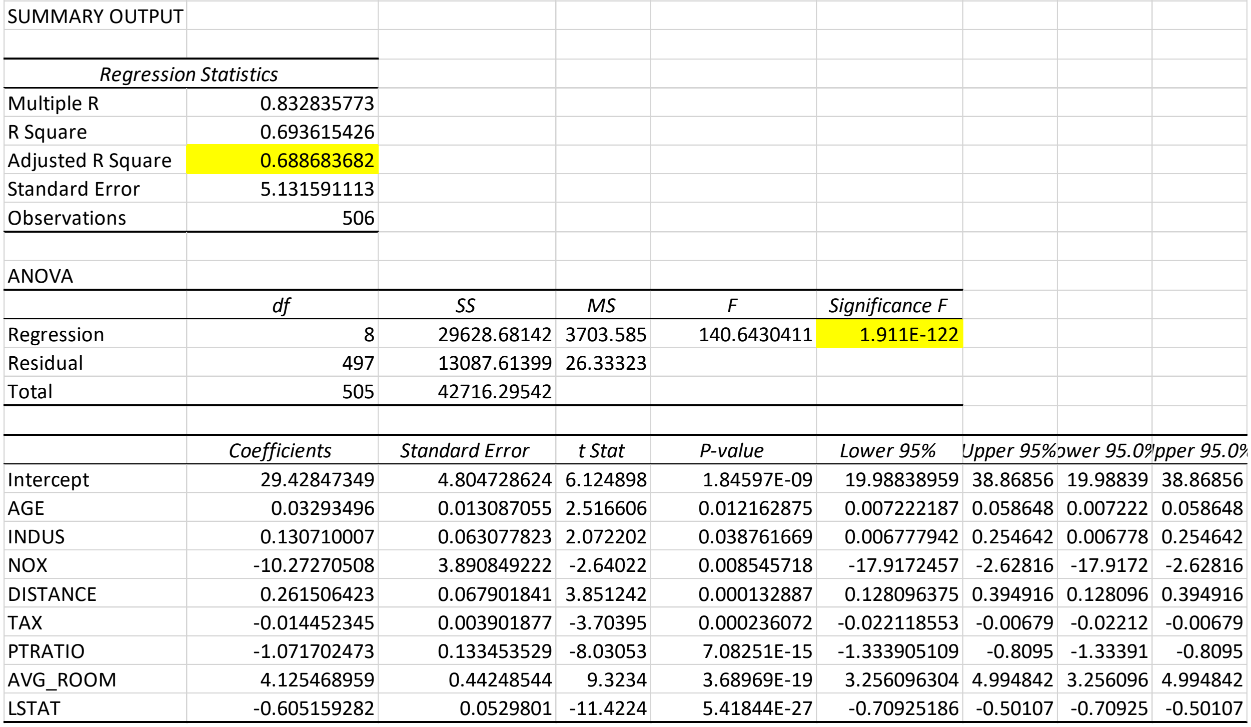
8) Pick out only the significant variables from the previous question. Make another instance of the Regression model using only the significant variables you just picked and answer the questions below: (8 marks)

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.



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	3703.585	140.6430411	1.911E-122			
Residual	497	13087.61399	26.33323					
Total	505	42716.29542						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	29.42847349	4.804728624	6.124898	1.84597E-09	19.98838959	38.86856	19.98839	38.86856
AGE	0.03293496	0.013087055	2.516606	0.012162875	0.007222187	0.058648	0.007222	0.058648
INDUS	0.130710007	0.063077823	2.072202	0.038761669	0.006777942	0.254642	0.006778	0.254642
NOX	-10.27270508	3.890849222	-2.64022	0.008545718	-17.9172457	-2.62816	-17.9172	-2.62816
DISTANCE	0.261506423	0.067901841	3.851242	0.000132887	0.128096375	0.394916	0.128096	0.394916
TAX	-0.014452345	0.003901877	-3.70395	0.000236072	-0.022118553	-0.00679	-0.02212	-0.00679
PTRATIO	-1.071702473	0.133453529	-8.03053	7.08251E-15	-1.333905109	-0.8095	-1.33391	-0.8095
AVG_ROOM	4.125468959	0.44248544	9.3234	3.68969E-19	3.256096304	4.994842	3.256096	4.994842
LSTAT	-0.605159282	0.0529801	-11.4224	5.41844E-27	-0.70925186	-0.50107	-0.70925	-0.50107

From the previous model it is observed that CRIME_RATE, is an insignificant variable. So, in this model the independent variables used for regression analysis are AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG_ROOM and LSTAT and AVG_PRICE is the dependent variable.

a) From the output it is inferred that all the variables are significant as the p values of all the independent variables is less than 0.05, hence rejecting the null hypothesis. The Adjusted R Square value is 68.86% which means 68.86% of variance in AVG_PRICE is explained by the independent variables in this model. It is also inferred that the coefficients of AGE, INDUS, DISTANCE and AVG_ROOM are positive so whenever there is increase in value of these variables, the AVG_PRICE also increases accordingly. On the other hand, it is observed that the coefficients of NOX, TAX, PTRATIO and LSTAT are negative and hence the AVG_PRICE decreases with increase in value of these variables.

b) From the Regression summary output, it can be inferred that the Adjusted R Square value is 68.86%, which is the highest when compared to the previous regression models. So, this model performs better than previous models.

c) After sorting, the values of coefficients in ascending order are, as follows:

NOX (-10.27) **<** PTRATIO (-1.07) **<** LSTAT (-0.60) **<** TAX (-0.01) **<** AGE (0.03) **<** INDUS (0.13) **<** DISTANCE (0.26) **<** AVG_ROOM (4.12)

If the value of NOX is more in a locality, then the AVG_PRICE decreases by 10.27 points for 1 point increase in NOX. So, basically the AVG_PRICE of house will be lesser in a locality having higher concentration of nitic oxides (pollutants). More the pollution, lesser will be the average price of the property.

d) The Regression equation for this model is

AVG_PRICE = 0.03 * AGE + 0.13 * INDUS - 10.27 * NOX + 0.26 * DISTANCE - 0.01 * TAX - 1.07 * PTRATIO + 4.12 * AVG_ROOM - 0.605 * LSTAT + 29.42

CONCLUSION OF TERRO'S REAL ESTATE DATA ANALYSIS:

From the analysis, it is concluded that the localities having higher proportion of houses built prior to 1940, having higher proportion of non- retail business acres, away from the highway, houses having more average number of rooms, low pollution, low property taxes, low pupil teacher ratio, low percentage of lower status of population are priced higher for the properties. The crime rate in a locality does not have any impact on the value of the property. Thus, it can be said that Average Price of a property depends on the above said variables (AGE, INDUSTRY, DISTANCE, TAX, NOX, PTRATIO, AVG_ROOM, LSTAT) from the multi linear regression model developed for this analysis.

MOST RELEVANT FEATURES FOR PRICING OF A HOUSE IN A LOCALITY

FEATURES WHICH INCREASE THE PRICE OF A HOUSE IN A LOCALITY ARE:

- Higher proportion of houses built prior to 1940
- Higher proportion of non-retail business acres per town
- Property away from the highway
- More average number of rooms in the house
- Lower pollution in the locality
- Lower property tax in the locality
- Lower pupil teacher ratio in the town
- Lower percentage of lower status population in the locality