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

CRIME_RATE		*AGE*		*INDUS*	
Mean	4.871976285	Mean	68.57490119	Mean	11.13677866
Standard Error	0.129860152	Standard Error	1.251369525	Standard Error	0.304979888
Median	4.82	Median	77.5	Median	9.69
Mode	3.43	Mode	100	Mode	18.1
Standard Deviation	2.921131892	Standard Deviation	28.14886141	Standard Deviation	6.860352941
Sample Variance	8.533011532	Sample Variance	792.3583985	Sample Variance	47.06444247
Kurtosis	1.189122464	Kurtosis	0.967715594	Kurtosis	1.233539601
Skewness	0.021728079	Skewness	-0.59896264	Skewness	0.295021568
Range	9.95	Range	97.1	Range	27.28
Minimum	0.04	Minimum	2.9	Minimum	0.46
Maximum	9.99	Maximum	100	Maximum	27.74
Sum	2465.22	Sum	34698.9	Sum	5635.21
Count	506	Count	506	Count	506

NOX		*DISTANCE*		*TAX*	
Mean	0.554695059	Mean	9.549407115	Mean	408.2371542
Standard Error	0.005151391	Standard Error	0.387084894	Standard Error	7.492388692
Median	0.538	Median	5	Median	330
Mode	0.538	Mode	24	Mode	666
Standard Deviation	0.115877676	Standard Deviation	8.707259384	Standard Deviation	168.5371161
Sample Variance	0.013427636	Sample Variance	75.81636598	Sample Variance	28404.75949
Kurtosis	0.064667133	Kurtosis	0.867231994	Kurtosis	1.142407992
Skewness	0.729307923	Skewness	1.004814648	Skewness	0.669955942
Range	0.486	Range	23	Range	524
Minimum	0.385	Minimum	1	Minimum	187
Maximum	0.871	Maximum	24	Maximum	711
Sum	280.6757	Sum	4832	Sum	206568
Count	506	Count	506	Count	506

PTRATIO		*AVG_ROOM*		*LSTAT*	
Mean	18.4555336	Mean	6.28463438 7	Mean	12.6530632 4
Standard Error	0.09624356 8	Standard Error	0.03123514 2	Standard Error	0.31745890 6
Median	19.05	Median	6.2085	Median	11.36
Mode	20.2	Mode	5.713	Mode	8.05
Standard Deviation	2.16494552 4	Standard Deviation	0.70261714 3	Standard Deviation	7.14106151 1
Sample Variance	4.68698912 1	Sample Variance	0.49367085	Sample Variance	50.9947595 1
Kurtosis	- 0.28509138 3	Kurtosis	1.89150036 6	Kurtosis	0.49323951 7
Skewness	- 0.80232492 7	Skewness	0.40361213 3	Skewness	0.90646009 4
Range	9.4	Range	5.219	Range	36.24
Minimum	12.6	Minimum	3.561	Minimum	1.73
Maximum	22	Maximum	8.78	Maximum	37.97
Sum	9338.5	Sum	3180.025	Sum	6402.45
Count	506	Count	506	Count	506

AVG_PRICE	
Mean	22.53280632
Standard Error	0.408861147
Median	21.2
Mode	50
Standard Deviation	9.197104087
Sample Variance	84.58672359
Kurtosis	1.495196944
Skewness	1.108098408
Range	45
Minimum	5
Maximum	50
Sum	11401.6
Count	506

The total count of each attribute is 506 and hence there are no empty.

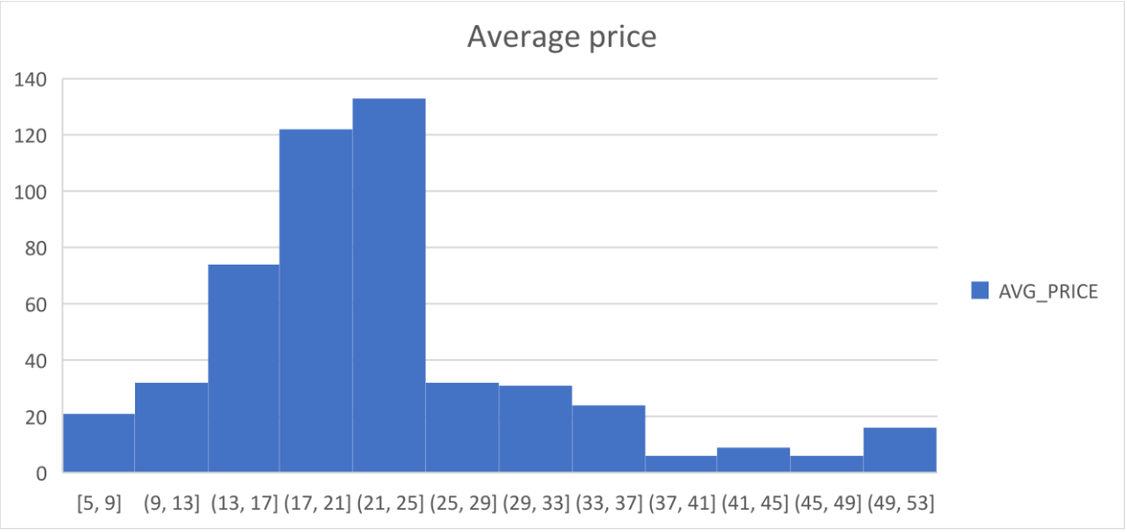
The average crime rate per town is 4.87 and the skewness is very minimum

More than 50 percent of the houses are older than 77.5 years which is more than mean hence a negative skew can be observed. More number of houses are 100 years

Considering distance, it is extremely skewed.

The tax is observed to be $524 full-value property-tax rate per $10,000

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



From the above histogram

we can clearly see that most of the houses range from $21000 to $25000 and very less number of houses are in the range of $45000 to $49000

3) Compute the covariance matrix. Share your observations.

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.5161478730									
AGE	0.5629152150	790.7924728163								
INDUS	-0.1102151752	124.2678282390	46.9714297415							
NOX	0.0006253082	2.3812119313	0.6058739426	0.0134010989						
DISTANCE	-0.2298604884	111.5499554750	35.4797144933	0.6157102243	75.6665312690					
TAX	-8.2293224390	2397.9417230390	831.7133331250	13.0205023575	1333.1167413957	28348.6235998063				
PTRATIO	0.0681689059	15.9054254480	5.6808547821	0.0473036538	8.7434024903	167.8208220719	4.6777262963			
AVG_ROOM	0.0561177779	-4.7425380302	-1.8842254268	-0.0245548261	-1.2812773907	-34.5151010405	-0.5396945183	0.4926952161		
LSTAT	-0.8826803621	120.8384405201	29.5218112512	0.4879798709	30.3253921324	653.4206174132	5.7713002429	-3.0736549670	50.8939793517	
AVG_PRICE	1.1620122405	-97.3961528848	-30.4605049915	-0.4545124071	-30.5008303520	-724.8204283773	-10.0906756081	4.4845655517	-48.3517921933	84.4195561562

From the above table we can conclude that

Tax parameter has high covariance among most of parameters like age of the house, industry, distance from highway, tax, % lower status of the population

Average price has very less covariance between other parameters

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	0.644778511	1							
NOX	0.001850982	0.731470104	0.763651447	1						
DISTANCE	-0.009055049	0.456022452	0.595129275	0.611440563	1					
TAX	-0.016748522	0.506455594	0.72076018	0.6680232	0.910228189	1				
PTRATIO	0.010800586	0.261515012	0.383247556	0.188932677	0.464741179	0.460853035	1			
AVG_ROOM	0.02739616	-0.240264931	-0.391675853	-0.302188188	-0.209846668	-0.292047833	-0.355501495	1		
LSTAT	-0.042398321	0.602338529	0.603799716	0.590878921	0.488676335	0.543993412	0.374044317	-0.613808272	1	
AVG_PRICE	0.043337871	-0.376954565	-0.48372516	-0.427320772	-0.381626231	-0.468535934	-0.507786686	0.695359947	-0.737662726	1

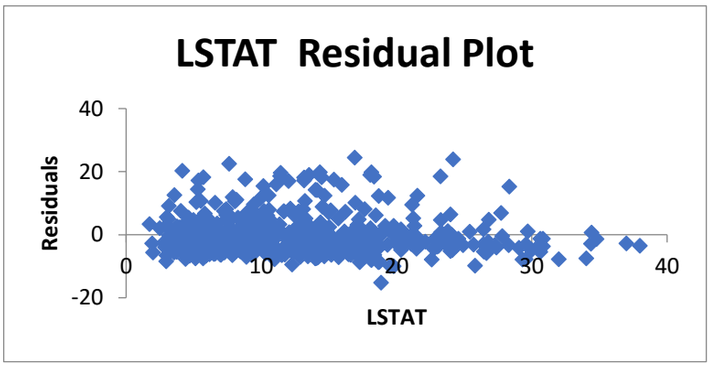
a) Top 3 positively correlated pairs are

   i. NOX - Age
   ii. NOX - Age
   iii. Tax - Distance

b) Top 3 negatively correlated pairs are
   Avg price – PTRATIO
   Avg price – LSTAT
   Avg room – LSTAT

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?

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



SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.737663							
R Square	0.544146							
Adjusted R	0.543242							
Standard E	6.21576							
Observatic	506							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regressior	1	23243.91	23243.91	601.6179	5.08E-88			
Residual	504	19472.38	38.63568					
Total	505	42716.3						
	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	34.55384	0.562627	61.41515	3.7E-236	33.44846	35.65922	33.44846	35.65922
LSTAT	-0.95005	0.038733	-24.5279	5.08E-88	-1.02615	-0.87395	-1.02615	-0.87395

The coefficient of LSTAT is -0.95005. It says that if LSTAT increases by 0.9 times, then the average house price falls by 0.9 times

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

Yes, LSTAT is a significant parameter for the average price as the p value is 5.08E-88 which is very much less than 0.05

Hence LSTAT is significant for the analysis based on the 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.

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?

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

The regression equation is as follows:

$y = -1.358 + 5.09x_0 - 0.642x_1$

where,

y – average price

x_0 – Average room

x_1 = LSTAT

For 7 rooms and a value of LSTAT = 20

$y = -1.358 + 5.09*7 - 0.642*20$

$= 21.44$

Hence, price for new house is $21,440

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.

The linear equation for the above model is

$y = -1.358 + 5.09a - 0.642b$

The value of R square in this model = 63.85%

The value of R square in the previous model = 54.4%

Hence, we can clearly say that the performance of this model better than the previous model as R square value is more than the previous 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.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.832978824							
R Square	0.69385372							
Adjusted R Square	0.688298647							
Standard Error	5.1347635							
Observations	506							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	9	29638.8605	3293.206722	124.9045049	1.9328E-121			
Residual	496	13077.43492	26.3657962					
Total	505	42716.29542						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.24131526	4.817125596	6.070282926	2.53978E-09	19.77682784	38.70580267	19.77682784	38.70580267
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.534657201	-0.105348544	0.202798827	-0.105348544	0.202798827
AGE	0.032770689	0.013097814	2.501996817	0.012670437	0.00703665	0.058504728	0.00703665	0.058504728
INDUS	0.130551399	0.063117334	2.068392165	0.03912086	0.006541094	0.254561704	0.006541094	0.254561704
NOX	-10.3211828	3.894036256	-2.650510195	0.008293859	-17.97202279	-2.670342809	-17.97202279	-2.670342809
DISTANCE	0.261093575	0.067947067	3.842602576	0.000137546	0.127594012	0.394593138	0.127594012	0.394593138
TAX	-0.01440119	0.003905158	-3.687736063	0.000251247	-0.022073881	-0.0067285	-0.022073881	-0.0067285
PTRATIO	-1.074305348	0.133601722	-8.041104061	6.58642E-15	-1.336800438	-0.811810259	-1.336800438	-0.811810259
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.89287E-19	3.255494742	4.995323561	3.255494742	4.995323561
LSTAT	-0.603486589	0.053081161	-11.36912937	8.91071E-27	-0.70777824	-0.499194938	-0.70777824	-0.499194938

From the above table we can see that crime rate does not play a significant role in the average price of the house as the p-value is less than 0.05

R-square = 0.693

R-square of 69.3% reveals that 69.3% of the variability observed in the target variable is seen by the regression model.

NOX, TAX, PTRATIO, and LSTAT have negative coefficients that say soan increase in these attributes will result in a decrease in the price of the house and vice versa.

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.

SUMMARY OUTPUT

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

ANOVA

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

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

a) Interpret the output of this model.

All the parameters play a significant role in the average price of the house as we can see the p-value is less than 0.05

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?

Multiple R	0.832978824
R Square	0.69385372

Multiple R	0.832835773
R Square	0.693615426

By comparing Multiple R and R square of the models, we can see that there is negligible difference in the values. Hence, we can conclude that both the models perform well

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?

	Coefficients
NOX	-10.27270508
PTRATIO	-1.071702473
LSTAT	-0.605159282
TAX	-0.014452345
AGE	0.03293496
INDUS	0.130710007
DISTANCE	0.261506423
AVG_ROOM	4.125468959
Intercept	29.42847349

If the value of NOX is more in a locality in this town, then the average price of the decreases by 10.27 times

d) Write the regression equation from this model.

The regression equation from this model is

$y = 0.03293496\ X_0 + 0.130710007\ X_1 - 10.27270508\ X_2 + 0.261506423\ X_3 - 0.014452345\ X_4 - 1.071702473\ X_5 + 4.125468959\ X_6 - 0.605159282\ X_7 + 29.42847349$

X_0 = Age

X_1 = Indus

X_2 = NOX

X_3 = Distance

X_4 = Tax

X_5 = PTRATIO

X_6 = Average room

X_7 = LSTAT