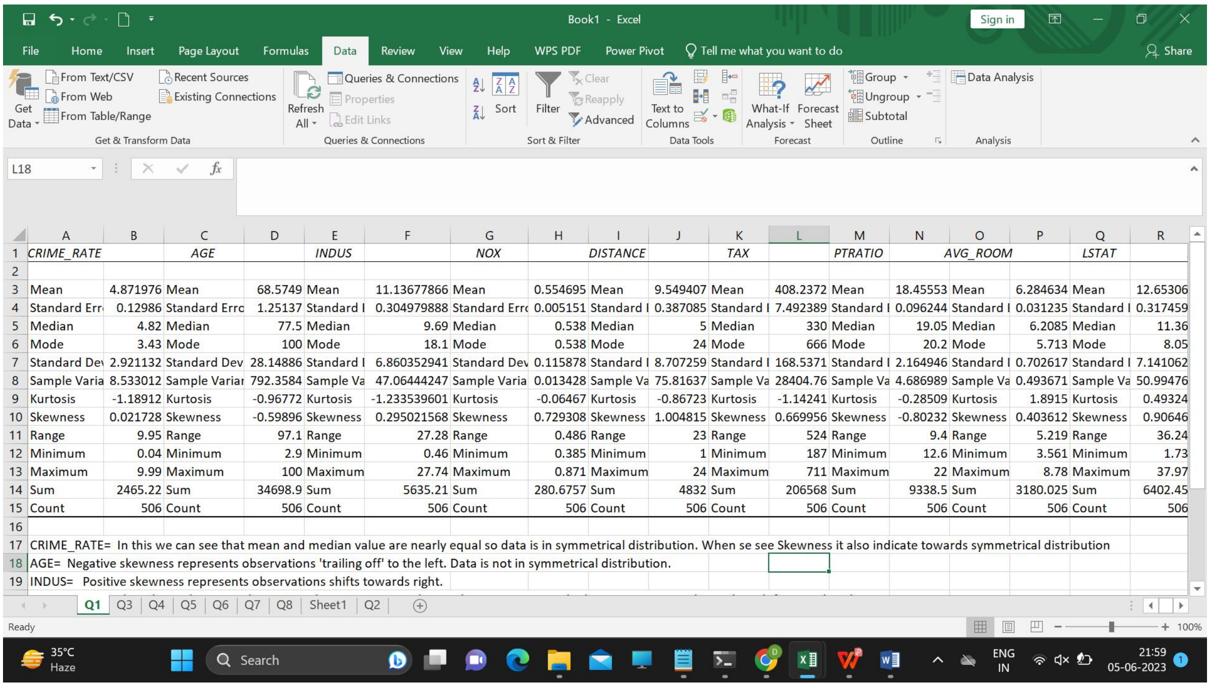
PROJECT REPORT

1. The first step to any project is understanding the data. So for this step, generate the summary statistics for each of the variables. What do you observe?

From Data Analysis select Descriptive
Statistic



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	CRIME_RATE		AGE		INDUS		NOX		DISTANCE		TAX		PTRATIO		AVG_ROOM		LSTAT	
2																		
3	Mean	4.871976	Mean	68.5749	Mean	11.13677866	Mean	0.554695	Mean	9.549407	Mean	408.2372	Mean	18.45553	Mean	6.284634	Mean	12.65306
4	Standard Err	0.12986	Standard Errc	1.25137	Standard I	0.304979888	Standard Err	0.005151	Standard I	0.387085	Standard I	7.492389	Standard I	0.096244	Standard I	0.031235	Standard I	0.317459
5	Median	4.82	Median	77.5	Median	9.69	Median	0.538	Median	5	Median	330	Median	19.05	Median	6.2085	Median	11.36
6	Mode	3.43	Mode	100	Mode	18.1	Mode	0.538	Mode	24	Mode	666	Mode	20.2	Mode	5.713	Mode	8.05
7	Standard Dev	2.921132	Standard Dev	28.14886	Standard I	6.860352941	Standard Dev	0.115878	Standard I	8.707259	Standard I	168.5371	Standard I	2.164946	Standard I	0.702617	Standard I	7.141062
8	Sample Varia	8.533012	Sample Variar	792.3584	Sample Va	47.06444247	Sample Varia	0.013428	Sample Va	75.81637	Sample Va	28404.76	Sample Va	4.686989	Sample Va	0.493671	Sample Va	50.99476
9	Kurtosis	-1.18912	Kurtosis	-0.96772	Kurtosis	-1.233539601	Kurtosis	-0.06467	Kurtosis	-0.86723	Kurtosis	-1.14241	Kurtosis	-0.28509	Kurtosis	1.8915	Kurtosis	0.49324
10	Skewness	0.021728	Skewness	-0.59896	Skewness	0.295021568	Skewness	0.729308	Skewness	1.004815	Skewness	0.669956	Skewness	-0.80232	Skewness	0.403612	Skewness	0.90646
11	Range	9.95	Range	97.1	Range	27.28	Range	0.486	Range	23	Range	524	Range	9.4	Range	5.219	Range	36.24
12	Minimum	0.04	Minimum	2.9	Minimum	0.46	Minimum	0.385	Minimum	1	Minimum	187	Minimum	12.6	Minimum	3.561	Minimum	1.73
13	Maximum	9.99	Maximum	100	Maximum	27.74	Maximum	0.871	Maximum	24	Maximum	711	Maximum	22	Maximum	8.78	Maximum	37.97
14	Sum	2465.22	Sum	34698.9	Sum	5635.21	Sum	280.6757	Sum	4832	Sum	206568	Sum	9338.5	Sum	3180.025	Sum	6402.45
15	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506
16																		

17 CRIME_RATE= In this we can see that mean and median value are nearly equal so data is in symmetrical distribution. When se see Skewness it also indicate towards symmetrical distribution

18 AGE= Negative skewness represents observations 'trailing off' to the left. Data is not in symmetrical distribution.

19 INDUS= Positive skewness represents observations shifts towards right.

CRIME_RATE= In this we can see that mean and median value are nearly equal so data is in symmetrical distribution. When se see Skewness it also indicate towards symmetrical distribution.

AGE= Negative skewness represents observations 'trailing off' to the left. Data is not in symmetrical distribution.

INDUS= Positive skewness represents observations shifts towards right.

NOX= Mean and median value is nearly same so data is symmetrical. But when we see towards skewness i.e. 0.7 it shows data shift towards right.

DISTANCE= In this median is 5 which shows half of the houses are just 5miles away from highway. Positive value of skewness shows data shift towards right.

TAX= Through range(524) we can see that there is large difference in tax.

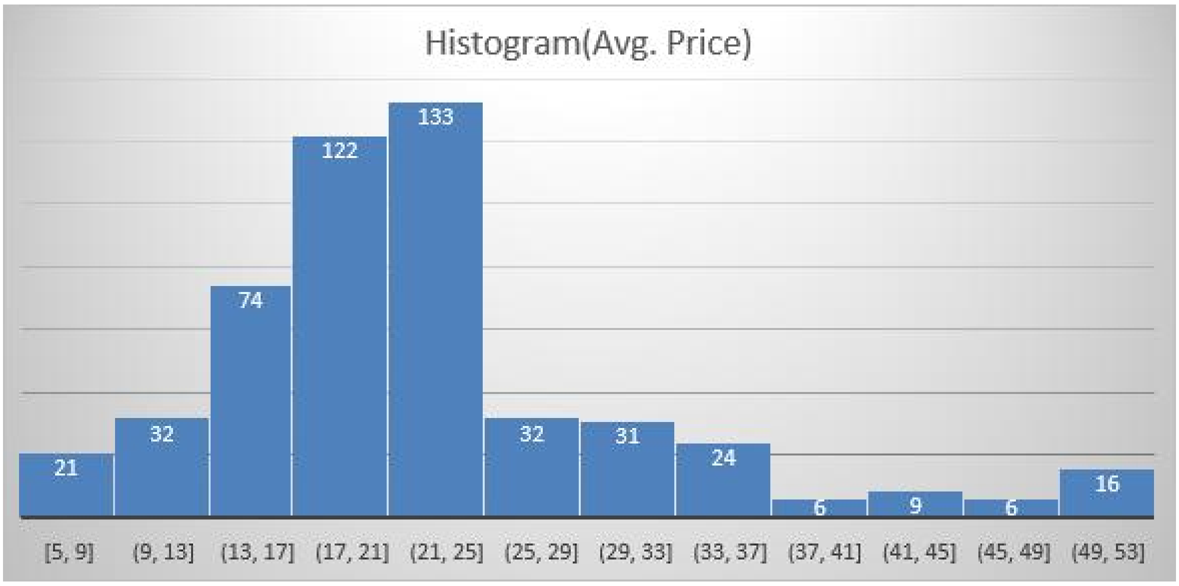
PTRATIO= Pupil teacher ratio is max(22), which is nearly equal to mean(18.5), median(19) and mode(20). So the positive skewness shows data shift highly towards right.

AVG_ROOM= In this mean and median are nearly equal. Median(6) we can conclude that 50% of the house have room more then 6.

LSTAT= Positive skewness 0.9 data observations shifts towards right.

AVG_PRICE= Average price of house is 22 which is nearly equal to median 21. Range(45) which is high in Avg_Price of house, 50% of the houses are 21000+USD.

2. Plot the histogram of the Avg_Price Variable. What do you infer?



Most of the houses average price in between $17000-$25000.
Through this histogram we can clearly see that there only some houses that are expensive.

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	790.7924728								
INDUS	-0.110215175	124.2678282	46.97142974							
NOX	0.000625308	2.381211931	0.605873943	0.013401099						
DISTANCE	-0.229860488	111.5499555	35.47971449	0.615710224	75.66653127					
TAX	-8.229322439	2397.941723	831.7133331	13.02050236	1333.116741	28348.6236				
PTRATIO	0.068168906	15.90542545	5.680854782	0.047303654	8.74340249	167.8208221	4.677726			

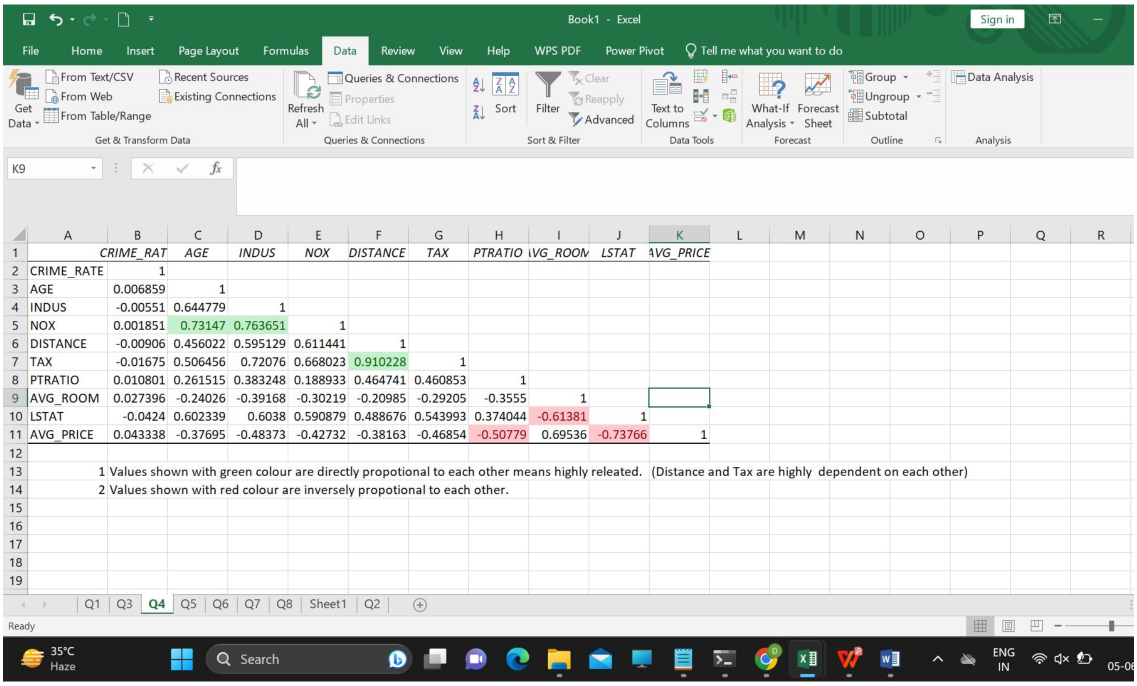
AVG_ROOM	0.056117778	-4.74253803	-1.88425427	-0.024554826	-1.28127739	-34.51510104	-0.53969	0.492695216		
LSTAT	-0.882680362	120.8384405	29.52181125	0.487979871	30.32539213	653.4206174	5.7713	-3.073654967	50.89397935	
AVG_PRICE	1.16201224	-97.39615288	-30.46050499	-0.454512407	-30.50083035	-724.8204284	-10.0907	4.484565552	-48.35179219	84.41956

A high covariance basically indicates there is a strong relationship between the variables. A low value means there is a weak relationship.

1 Distance and Tax are highly relatable.

2 Tax and Avg_Price are inversely relatable.
Values which are nearly equal to zero means there is no relation between them.

4.Create a correlation matrix of all the variables as shown in the Videos and various case studies. State top 3 positively correlated pairs and top 3 negatively correlated pairs.



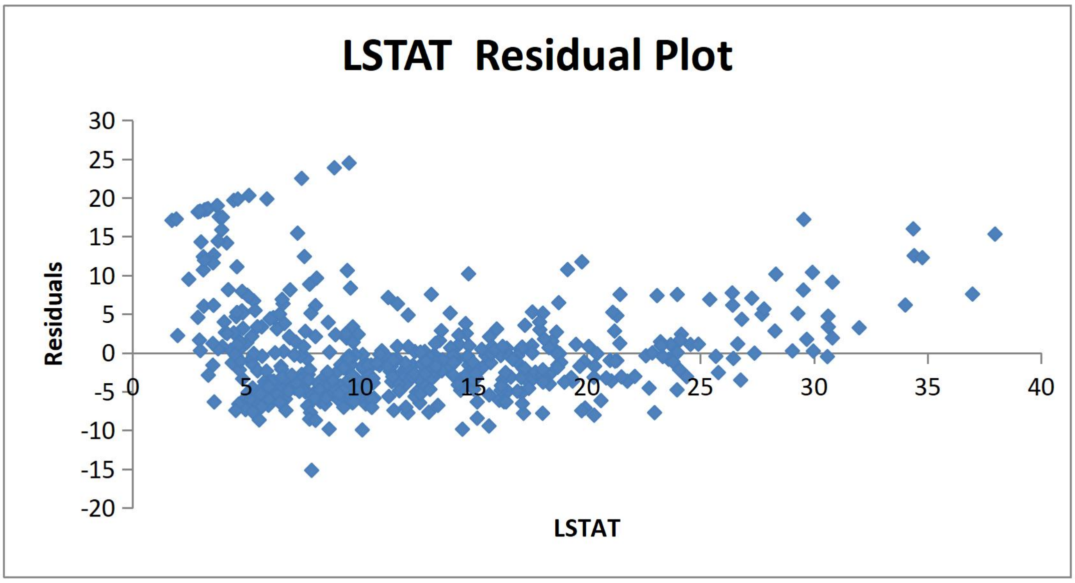
	CRIME_RAT	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	VG_ROOM	LSTAT	VG_PRICE
CRIME_RATE	1									
AGE	0.006859	1								
INDUS	-0.00551	0.644779	1							
NOX	0.001851	0.73147	0.763651	1						
DISTANCE	-0.00906	0.456022	0.595129	0.611441	1					
TAX	-0.01675	0.506456	0.72076	0.668023	0.910228	1				
PTRATIO	0.010801	0.261515	0.383248	0.188933	0.464741	0.460853	1			
AVG_ROOM	0.027396	-0.24026	-0.39168	-0.30219	-0.20985	-0.29205	-0.3555	1		
LSTAT	-0.0424	0.602339	0.6038	0.590879	0.488676	0.543993	0.374044	-0.61381	1	
AVG_PRICE	0.043338	-0.37695	-0.48373	-0.42732	-0.38163	-0.46854	-0.50779	0.69536	-0.73766	1

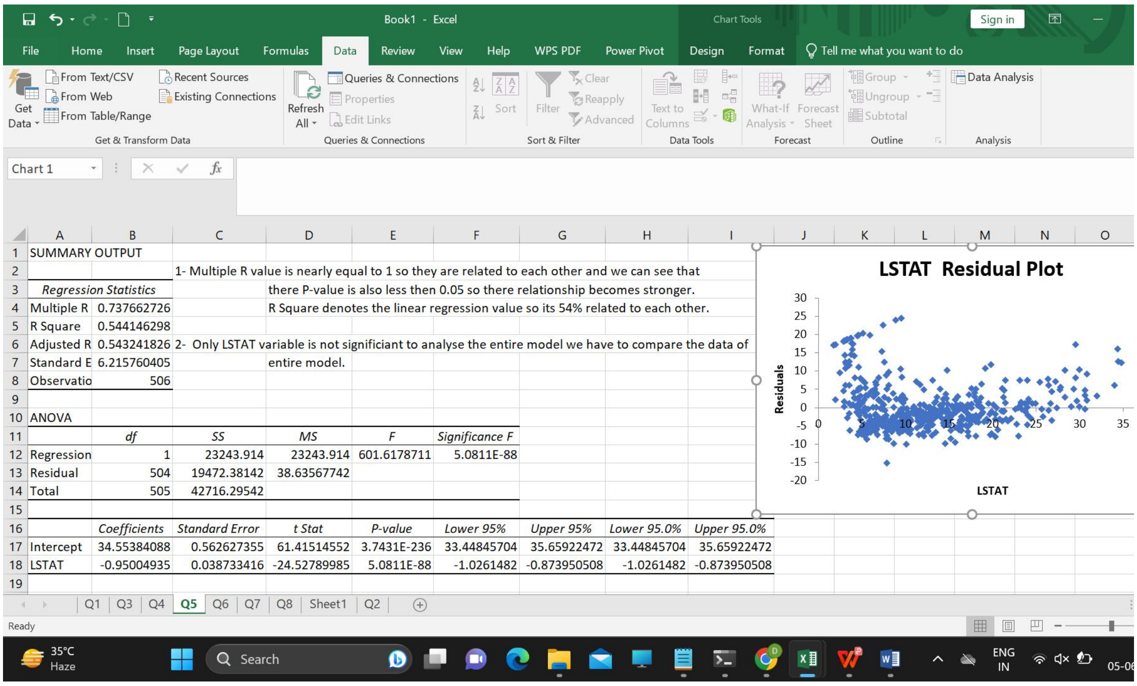
1 Values shown with green colour are directly propotional to each other means highly releated. (Distance and Tax are highly dependent on each other)

2 Values shown with red colour are inversely propotional to each other.

1- Values shown with green colour are directly propotional to each other means highly related.
2- Values shown with red colour are inversely propotional to each other.

5.Build an initial regression model with AVG_PRICE as the y or the Dependent variable and LSTAT variable as the Independent Variable. Generate the residual plot too.





SUMMARY OUTPUT

1- Multiple R value is nearly equal to 1 so they are related to each other and we can see that
there P-value is also less then 0.05 so there relationship becomes stronger.
R Square denotes the linear regression value so its 54% related to each other.

2- Only LSTAT variable is not significiant to analyse the entire model we have to compare the data of entire model.

Regression Statistics	
Multiple R	0.737662726
R Square	0.544146298
Adjusted R	0.543241826
Standard E	6.215760405
Observatio	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	33.44845704	35.65922472
LSTAT	-0.95004935	0.038733416	-24.52789985	5.0811E-88	-1.0261482	-0.873950508	-1.0261482	-0.873950508

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

Multiple R value is nearly equal to 1 so they are related to each other and we can see that

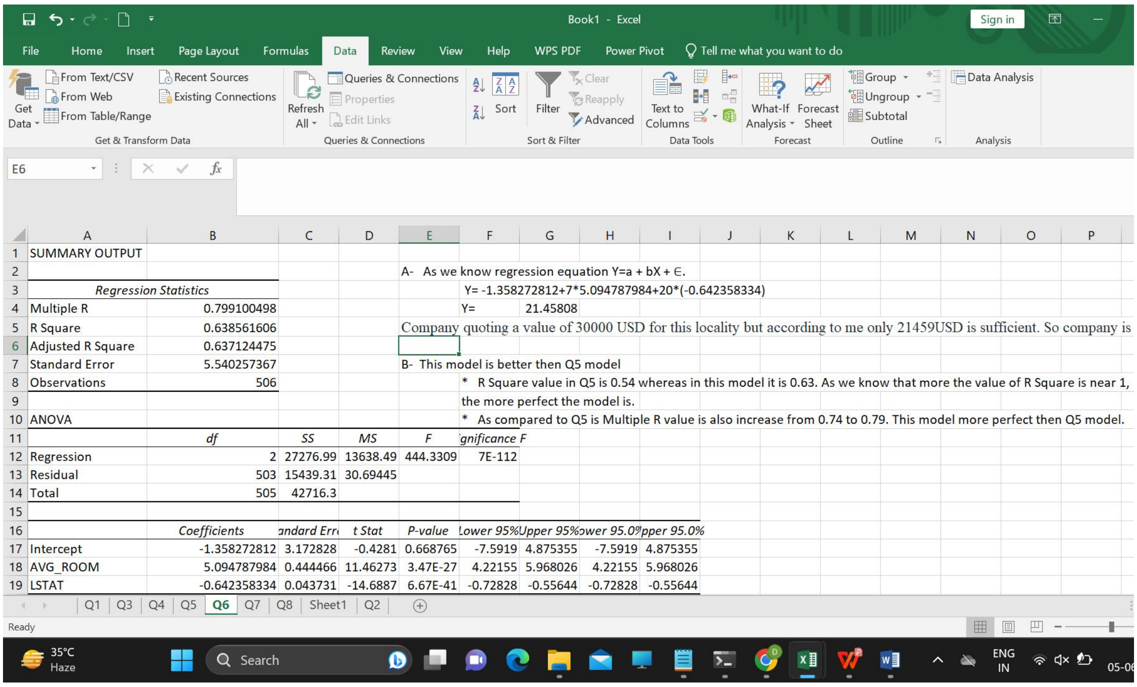
there P-value is also less than 0.05 so there relationship becomes stronger.

R Square denotes the linear regression value so its 54% related to each other.

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

Only LSTAT variable is not significant to analyse the entire model. We have to compare the data of entire model.

6. Build another instance of the Regression model but this time including LSTAT and AVG_ROOM together as Independent variables and AVG_PRICE as the dependent variable.



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	gnificance F
Regression	2	27276.99	13638.49	444.3309	7E-112
Residual	503	15439.31	30.69445		
Total	505	42716.3			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	-1.358272812	3.172828	-0.4281	0.668765	-7.5919	4.875355	-7.5919	4.875355
AVG_ROOM	5.094787984	0.444466	11.46273	3.47E-27	4.22155	5.968026	4.22155	5.968026
LSTAT	-0.642358334	0.043731	-14.6887	6.67E-41	-0.72828	-0.55644	-0.72828	-0.55644

A- As we know regression equation Y=a + bX + ∈.

Y= -1.358272812+7*5.094787984+20*(-0.642358334)

Y= 21.45808

Company quoting a value of 30000 USD for this locality but according to me only 21459USD is sufficient. So company is

B- This model is better then Q5 model

* R Square value in Q5 is 0.54 whereas in this model it is 0.63. As we know that more the value of R Square is near 1,

the more perfect the model is.

* As compared to Q5 is Multiple R value is also increase from 0.74 to 0.79. This model more perfect then Q5 model.

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?

A- As we know regression equation Y=a + bX + ∈.

Y= -1.358272812+7*5.094787984+20*(-0.642358334)

Y= 21.45808

Company quoting a value of 30000 USD for this locality but according to me only 21459USD is sufficient. So company is overcharging.

b. Is the performance of this model better than the previous model you built in Question 5? Compare in terms of adjusted R-square. Explain.

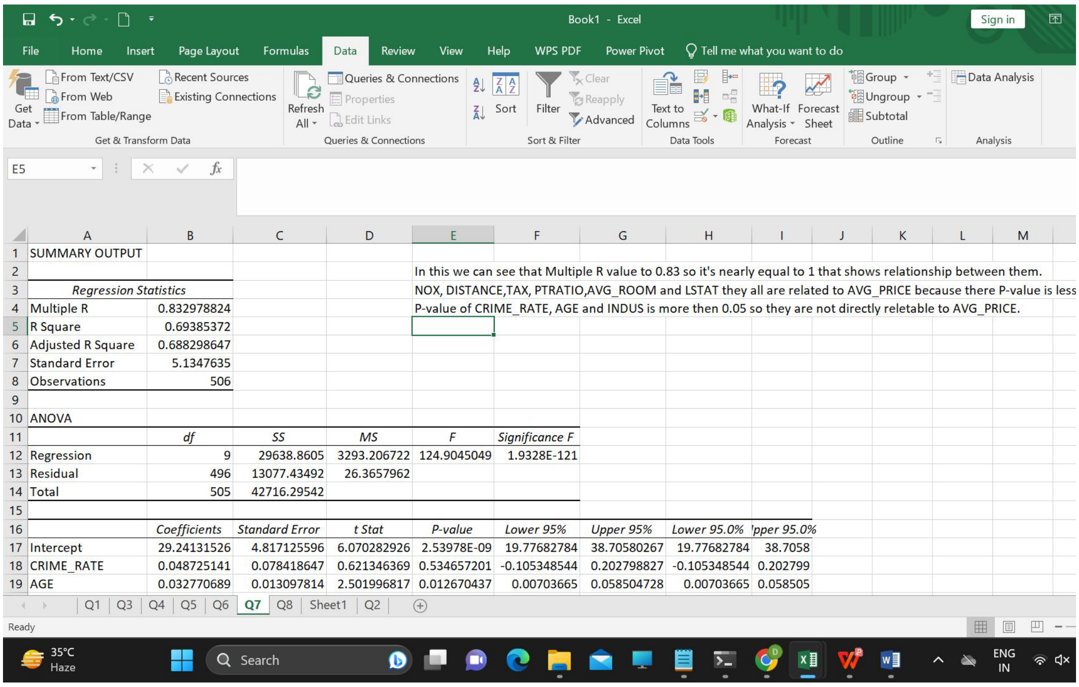
B- This model is better then Q5 model

* R Square value in Q5 is 0.54 whereas in this model it is 0.63. As we know that more the value of R Square is near 1,

the more perfect the model is.

* As compared to Q5 is Multiple R value is also increase from 0.74 to 0.79. This model more perfect then Q5 model.

7. Now, build a Regression model with all variables. AVG_PRICE shall be the Dependent Variable. Interpret the output in terms of adjusted R-square, coefficient and Intercept values, Significance of variables with respect to AVG_price. Explain.



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			

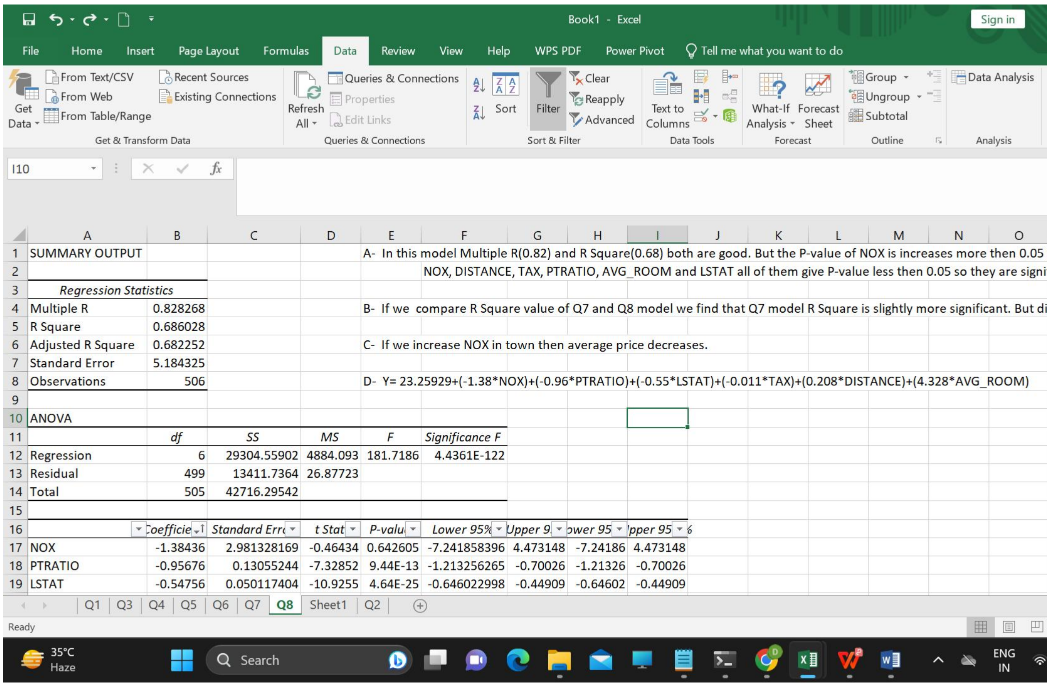
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*'pper 95.0%*
Intercept	29.24131526	4.817125596	6.070282926	2.53978E-09	19.77682784	38.70580267	19.77682784	38.7058
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.534657201	-0.105348544	0.202798827	-0.105348544	0.202799
AGE	0.032770689	0.013097814	2.501996817	0.012670437	0.00703665	0.058504728	0.00703665	0.058505

In this we can see that Multiple R value to 0.83 so it's nearly equal to 1 that shows relationship between them.
NOX, DISTANCE,TAX, PTRATIO,AVG_ROOM and LSTAT they all are related to AVG_PRICE because there P-value is less
P-value of CRIME_RATE, AGE and INDUS is more then 0.05 so they are not directly reletable to AVG_PRICE.

In this we can see that Multiple R value to 0.83 so it's nearly equal to 1 that shows relationship between them.
NOX, DISTANCE,TAX, PTRATIO,AVG_ROOM and LSTAT they all are related to AVG_PRICE because there P-value is less then 0.05.
P-value of CRIME_RATE, AGE and INDUS is more then 0.05 so they are not directly reletable to AVG_PRICE.

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.

(HINT: Significant variables are those whose p-values are less than 0.05. If the p-value is greater than 0.05 then it is insignificant)



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.828268
R Square	0.686028
Adjusted R Square	0.682252
Standard Error	5.184325
Observations	506

ANOVA

	df	SS	MS	F	Significance F
Regression	6	29304.55902	4884.093	181.7186	4.4361E-122
Residual	499	13411.7364	26.87723		
Total	505	42716.29542			

	Coefficie	Standard Err	t Stat	P-valu	Lower 95%	Upper 9	ower 95	pper 95
NOX	-1.38436	2.981328169	-0.46434	0.642605	-7.241858396	4.473148	-7.24186	4.473148
PTRATIO	-0.95676	0.13055244	-7.32852	9.44E-13	-1.213256265	-0.70026	-1.21326	-0.70026
LSTAT	-0.54756	0.050117404	-10.9255	4.64E-25	-0.646022998	-0.44909	-0.64602	-0.44909

A- In this model Multiple R(0.82) and R Square(0.68) both are good. But the P-value of NOX is increases more then 0.05
NOX, DISTANCE, TAX, PTRATIO, AVG_ROOM and LSTAT all of them give P-value less then 0.05 so they are signi

B- If we compare R Square value of Q7 and Q8 model we find that Q7 model R Square is slightly more significant. But di

C- If we increase NOX in town then average price decreases.

D- Y= 23.25929+(-1.38*NOX)+(-0.96*PTRATIO)+(-0.55*LSTAT)+(-0.011*TAX)+(0.208*DISTANCE)+(4.328*AVG_ROOM)

Answer the questions below:

a. Interpret the output of this model.

A- In this model Multiple R(0.82) and R Square(0.68) both are good. But the P-value of NOX is increases more then 0.05 so it is insignificant.

NOX, DISTANCE, TAX, PTRATIO, AVG_ROOM and LSTAT all of them give P-value less then 0.05 so they are significant.

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?

B- If we compare R Square value of Q7 and Q8 model we find that Q7 model R Square is slightly more significant. But difference is just 0.07.

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?

C- If we increase NOX in town then average price decreases.

d. Write the regression equation from this model.

D- Y= 23.25929+(-1.38*NOX)+(-0.96*PTRATIO)+(-0.55*LSTAT)+(-0.011*TAX)+(0.208*DISTANCE)+(4.328*AVG_ROOM)