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

ANS: Summary statistics for each variable in the table

RIME_RATE		AGE		INDUS		NOX		DISTANCE	
Mean	4.871976	Mean	68.5749	Mean	11.13678	Mean	0.554695	Mean	9.549407
Standard	0.12986	Standard	1.25137	Standard	0.30498	Standard	0.005151	Standard	0.387085
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	2.921132	Standard	28.14886	Standard	6.860353	Standard	0.115878	Standard	8.707259
Sample V	8.533012	Sample V	792.3584	Sample V	47.06444	Sample V	0.013428	Sample V	75.81637
Kurtosis	-1.18912	Kurtosis	-0.96772	Kurtosis	-1.23354	Kurtosis	-0.06467	Kurtosis	-0.86723
Skewness	0.021728	Skewness	-0.59896	Skewness	0.295022	Skewness	0.729308	Skewness	1.004815
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	6.284634	Mean	12.65306	Mean	22.53281
Standard	7.492389	Standard	0.096244	Standard	0.031235	Standard	0.317459	Standard	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	168.5371	Standard	2.164946	Standard	0.702617	Standard	7.141062	Standard	9.197104
Sample V	28404.76	Sample V	4.686989	Sample V	0.493671	Sample V	50.99476	Sample V	84.58672
Kurtosis	-1.14241	Kurtosis	-0.28509	Kurtosis	1.8915	Kurtosis	0.49324	Kurtosis	1.495197
Skewness	0.669956	Skewness	-0.80232	Skewness	0.403612	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

Observation:

1.Age and PTratio having negative Skewness.
Which means they are Left Skew, there Peak lies on right side.

2. Avg_Room, LSTAT and Avg_Price having postive Kurtosis.
Which means they are having Sharp peakedness.

Q1 Observations:

Observation:

CRIME_RATE :

1. Average of the crime rate is 4.87 in this Area.
2. Meadian it is the middle value of the crime rate is 4.82 which means 50% of the crime rate is below 4.82.
3. Standard Deviation is the amount of the variations is 2.92 from the dataset.
5. Kurtosis is in negative . It is indicates that Flat curve.
6. Skewness is the near 0 that shows relatively symmetric.
7. Range is 9.95. it between maximum and minimum in the dataset.
8. The Minimum crime rate in the dataset is 0.04 and Maximum is 9.99.
9. The Sum of all Crime rate is 2465.

AGE :

- Average years of house built in area is 68 years.
- The Meadian Age is 77.5 years.
- The Most Frequent value Appears in the dataset is 100
- Skewness indicated that data distrubution is negatively skewed ,value is -0.59.
- kurtsis is measures the tailedness probability of the data , its in the negative as well as indicated flat compared normal distribution.
- Minimum Age in the dataset 2.9 year
- Maximum Age is 100 Years

INDUS:

- Average proportion of non-retail business acres per town is 11.13%
- Most of the houses have 18% of the property to no retail business.
- Most of the houses have 18% of the property to no retail business.
- skewness is in slightly going positively and it is showing tail is towards to right. Right skewed(0.29)
- Minimum value of this variable is 0.46
- Maximum value of this variable is 27.74

NOX:

- Average of nitric oxides concentration is around 0.55
- The data indicates that slightly negative kurtosis.(-0.06)
- Skewness is slightly positively indicates and its right tailed (0.79)
- Minimum value of this variable is 0.385
- Maximum value is 0.871
- NOX has the lowest standard error is 0.00515

DISTANCE:

- Distance from highway on an average is 9.54.
- Negative Kurtosis indicatea that flatter curve.
- Positive skewness indicates that more number of houses are less than 9.5 miles away from highway.
- Maximum houses have 24 miles of distance from highway.
- Only 1 miles Minimum houses have distance from highway.

TAX:

- On an average , tax rate is $408.
- The maximum number of houses have tax rate around $666.
- The maximum tax rate is $711
- Minimum is $187.
- Negative kurtosis and positive skewness
- TAX has the highest standard error 7.49

PTRATIO:

- On an average , pupil teacher ratio is 18 for 506 houses.
- Maximum houses give 20 as pupil teacher ratio.
- Negative skewness indicates that left tailed distribution , more number of houses have more than 19 as a pupil teacher ratio
- The minimum value of PTRATIO in the dataset is 12.6.
- The maximum value of PTRATIO in the dataset is 22

AVG_ROOM:

- On an average , 6 rooms are there.
- Positive kurtosis gives us a sharp curve than normal curve – saying more values are concentrated near to median
- Positive skewness indicates right tailed distribution which says that more number of houses have less than 6 rooms
- Minimum value of avg_room is 3.561
- Maximum Value of Avg_room is 8.78

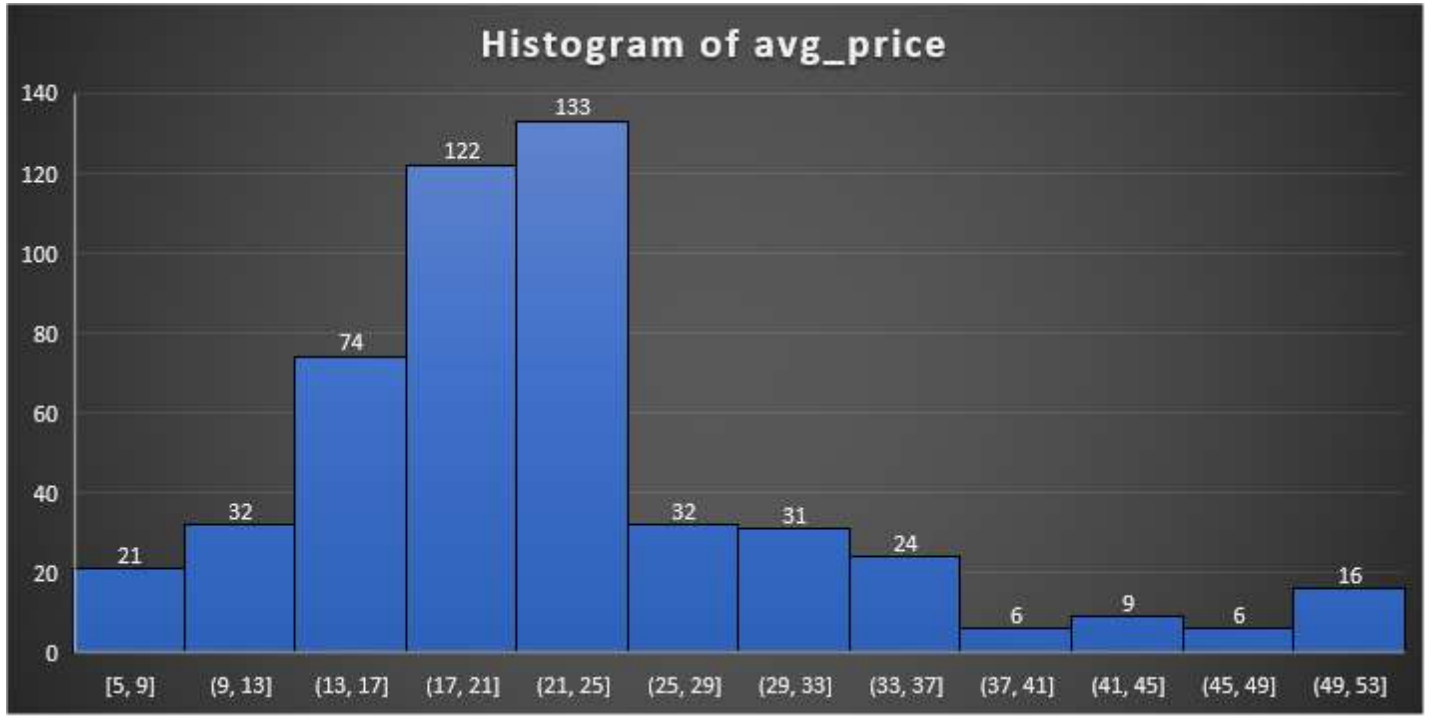
LSTAT:

- On an average , 12% of population has lower status.
- Positive kurtosis gives us a sharp curve
- Positive skewness tells us that more number of houses have less than 12% lower status population.
- Minimum value of 1.73
- Maximum is 37.97

AVG_PRICE:

- Average value of price of house is around $22500.
- Positive kurtosis gives us a sharp curve and positive skewness tells us that more number of houses have price less than $22500
- Maximum houseshave price aorund $50000

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



Observations:

1.The graph indicates that Right Tailed **Positively skewed.**

2. Majority of houses have average price less than the median value.

3. Its indicates that maximum point lie between 20 and 25.

Q3. Compute the covariance matrix. Share your observations.

Covariance	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	4.677726296			
AVG_ROOM	0.056117778	-4.74253803	-1.884225427	-0.024554826	-1.281277391	-34.51510104	-0.539694518	0.492695216		
LSTAT	-0.882680362	120.8384405	29.52181125	0.487979871	30.32539213	653.4206174	5.771300243	-3.073654967	50.89397935	
AVG_PRICE	1.16201224	-97.39615288	-30.46050499	-0.454512407	-30.50083035	-724.8204284	-10.09067561	4.484565552	-48.35179219	84.41955616

Observation:

1. Crime Rate has just one positive relation, with avg_price and that too not a significant one as per the value.
2. Crime rate follows a highly negative relation with tax means the house which has high tax rate , their crime rate is low.
3. The property tax rate is high for those houses who have been there for long since 1940. They share a positive relation
4. Non-retail business Industry , NOX , Distance : they all share a positive relation with tax rate.
5. Distance from highway shows a negative relation with average price of house.
6. Tax and average price of house both share negative relation.
7. The average price of house has a negative relation with pupil teacher ratio and LSTAT.
8. Age vs Tax, Indus vs Tax and Distance vs Tax are the data sets having more covariance thus we can say that they have a direct relation to each other, that is as one increases other data also increases.
9. Tax vs Avg price, Age vs Avg Price and Lstat vs Avg Price are the data sets having least covariance thus they have an inverse relation to each other, that is, if one increases other data decreases .

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.

Correlation	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

Observation:

A. Top 3 positively correlation pairs:
1 . DISTANCE and TAX : Correlation coefficient ≈ **0.9102** (strong positive correlation)
2. INDUS and NOX: Correlation coefficient ≈ **0.7636** (strong positive correlation)
3. NOX and AGE: Correlation coefficient ≈ **0.7314** (strong positive correlation)

B. Top 3 Negatively correlation pairs:
1. AVG_ROOM and LSTAT: Correlation coefficient ≈ **-0.6138** (strong negative correlation)
2. AVG_ROOM and LSTAT: Correlation coefficient ≈ **-0.7377** (strong negative correlation)
3. PTRAIO and AVG_PRICE: Correlation coefficient ≈ **-0.3916** (moderate negative correlation)

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?

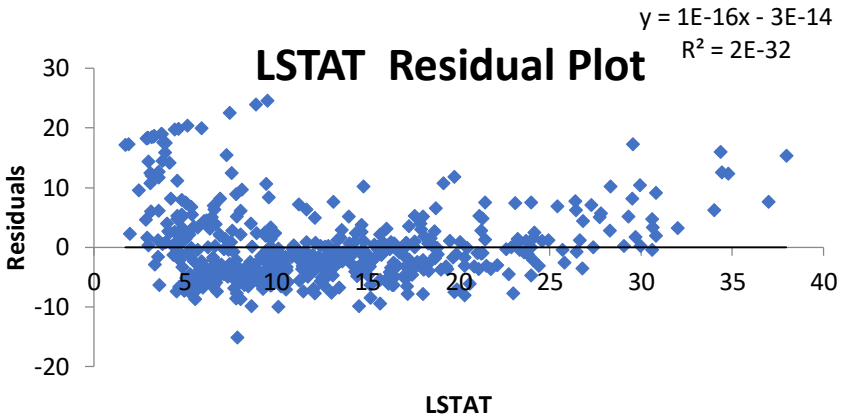
SUMMARY OUTPUT

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	23243.91	601.6178711	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	3.74308094E-236	33.44845704	35.65922472	33.44845704	35.65922472
LSTAT	-0.950049354	0.038733416	-24.5279	5.0811E-88	-1.0261482	-0.873950508	-1.0261482	-0.873950508



Ans Q5A.) By checking the coefficient value and the intercept value we can say that the coiefficent value increases by 1 the avg price decreases by 0.95, thus we can say that they are somewhat negatively(inversely) related, while the intercept is a positive value which signifies that it will increase the price at all the instances.

Q5 B.) As per the model LSTAT variable has a p value of 5.081*E^(-88) which is way less than 5%, thus we can use it for further analysis ,As checking the correlation we find that it is one of the variable which is mostly negative, hence an inverse relation Thus it will affect the average price:

It appears that the LSTAT variable is highly significant in the analysis and has a strong impact on the predicted AVG_PRICE.

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 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.

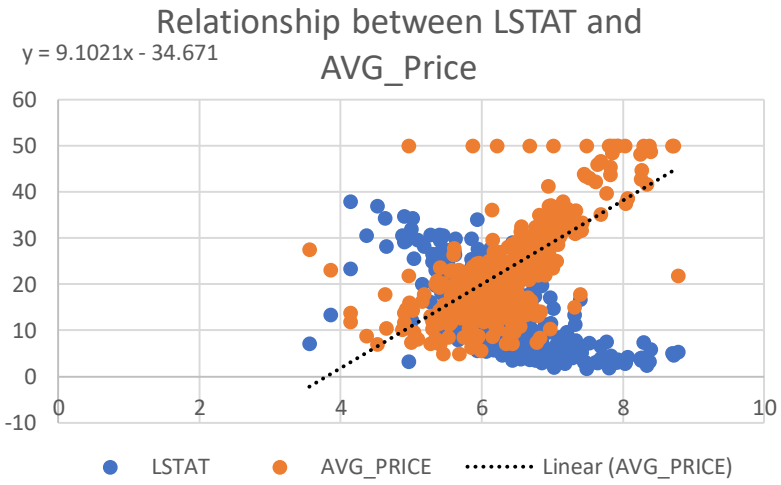
SUMMARY OUTPUT

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%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.358272812	3.17282778	-0.4281	0.668764941	-7.591900282	4.875354658	-7.591900282	4.875354658
AVG_ROOM	5.094787984	0.4444655	11.46273	3.47226E-27	4.221550436	5.968025533	4.221550436	5.968025533
LSTAT	-0.642358334	0.043731465	-14.6887	6.66937E-41	-0.728277167	-0.556439501	-0.728277167	-0.556439501



a) Regression Equation Y = 5.0948 X_1(Average room)-0.642 X_2(LSAT) -1.358

AVG_ROOM	7
L-STAT	20
AVG_PRICE	21.45807639

The company has quoted a value of 30000 USD for this locality, which is greater than our prediction 21458.076 USD. Hence The Company is Overchanging.

b)

Ans: While compating this model with the previous one, this model has an adjuster R square value of 0.637 and the previous one has the value of 0.543. Thus from these observations we can say, since the adjusted R square value of this model is high, the performance of this model is better than the previous model.

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.

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

a) While comparing the adjusted R square values of this model with the other values, we can observe that this model has the highest adjusted R square. value of 0.688 thus the performance of this model would be much better than the others.

b) Interpreting the coefficients and the intercept we can observe is that Crime rate, Age , Indus, Distance and Avg room are directly related with Average price and Nox, Tax, Ptratio and Lstat have are inversely related to Average price.

c) Observing the P values we can find that only insignificant variable in it is crime rate (P value 0.535) all others are significant.

d) While comparing the correlation table we can find that crime rate ana avg room are the ones having positive correlation and all the rest have negative correlation, thus we can conclude these are the significant variables.

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:

Ans:

A. While observing the Accuracy(81.52%) of this model.

B. By comparing the R square value(0.6886) of this model with the previous one(0.6882) we can interpret that this model has higher r square value than the previous.

C. The coefficient(-10.272) and correlation value(-0.427) of NOX is negative so we can summarise that as the value of NOX increases the value of AVG price decreases, thus they are inversely related.

D: REGRESSION EQUATION:

Y = 0.0329X_1(AGE) + 0.13X_2(INDUS) -10.272X_3(NOX) + 0.261X_4(DISTANCE) - 0.0144X_5(TAX) -1.0717X_6(PTRATIO) + 4.125 X_7(AVG ROOM) - 0.605 X_8(LSTAT) + 29.428