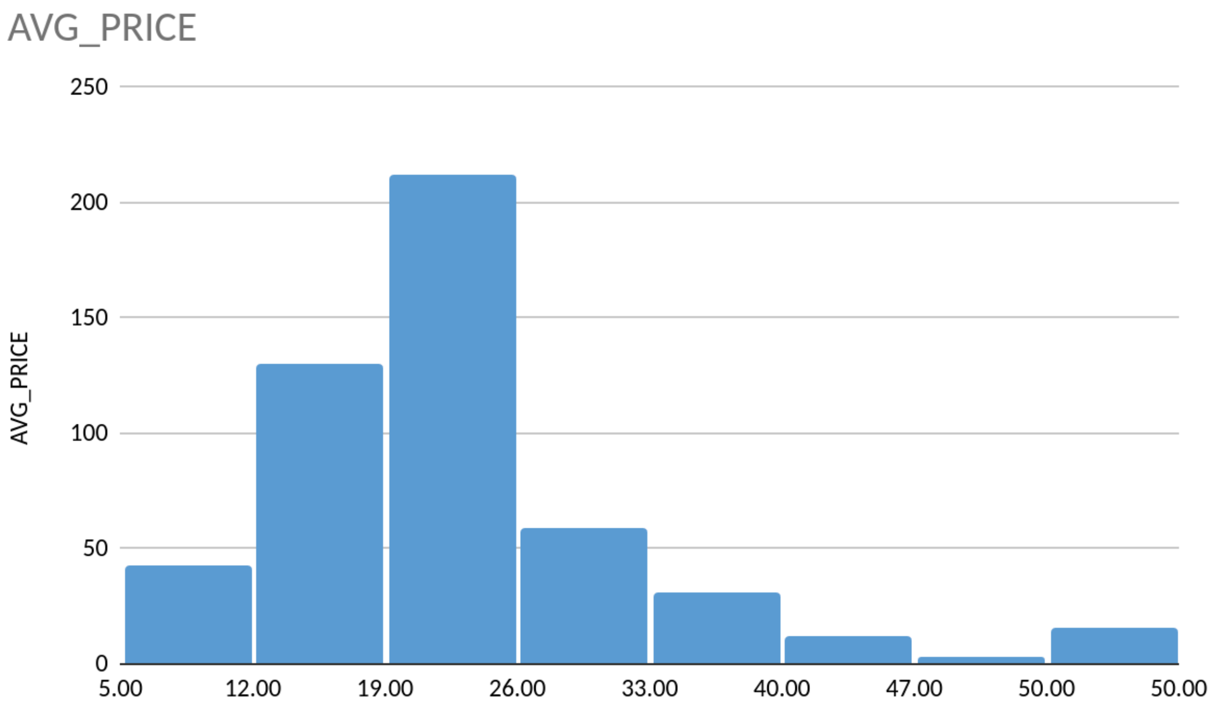
1. The first step to any project is understanding the data. So for this step, generate the summary statistics for each of the variables. What do you observe?

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
Mean	4.87198	68.57490	11.13678	0.55470	9.54941	408.23715	18.45553	6.28463	12.65306	22.53281
Standard Error	0.12986	1.25137	0.30498	0.00515	0.38708	7.49239	0.09624	0.03124	0.31746	0.40886
Median	4.82	77.5	9.69	0.538	5	330	19.05	6.2085	11.36	21.2
Mode	3.43	100	18.1	0.538	24	666	20.2	5.713	8.05	50
Standard Deviation	2.92113	28.14886	6.86035	0.11588	8.70726	168.53712	2.16495	0.70262	7.14106	9.19710
Sample variance	8.53301	792.35840	47.06444	0.01343	75.81637	28404.75949	4.68699	0.49367	50.99476	84.58672
Kurtosis	-1.18912	-0.96772	-1.23354	-0.06467	-0.86723	-1.14241	-0.28509	1.89150	0.49324	1.49520
Skewness	0.02173	-0.59896	0.29502	0.72931	1.00481	0.66996	-0.80232	0.40361	0.90646	1.10810
Range	9.95	97.1	27.28	0.486	23	524	9.4	5.219	36.24	45
Minimum	0.04	2.9	0.46	0.385	1	187	12.6	3.561	1.73	5
Maximum	9.99	100	27.74	0.871	24	711	22	8.78	37.97	50
Sum	2465.22	34698.9	5635.21	280.6757	4832	206568	9338.5	3180.025	6402.45	11401.6
Count	506	506	506	506	506	506	506	506	506	506

The average value of owner occupied houses (AVG_Price) in the data set is 22.53 ('000 USD), and the median value is 21.20. The range of values is between 5 to 50. The data has a bit of skewness. The average AGE is 68.5 and the median age is 77.5 suggesting negative skewness in this variable, it means that there are extreme values towards the lower end of the spectrum. Mean of AVG_Rooms is 6.28 and median is 6.2 suggesting that this variable could be normally distributed (more analysis would be required to know the exact picture). Most frequent value of AVG_ROOMS is 5.7.

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



The histogram of the ‘avg_price’ variable reveals the following insights:

1. **Distribution Shape**: The histogram shows a bell-shaped curve, indicating a roughly normal distribution of average prices, centered around the mean of approximately 22.5.
2. **Central Tendency**: Most of the data points are concentrated around the mean (22.5) and median (21.6), indicating that a large number of properties have average prices within this range.
3. **Spread**: The spread of the data is fairly wide, ranging from the minimum value of 5.0 to the maximum value of 50.0. The standard deviation of 9.188 also suggests significant variability in the prices.
4. **Skewness**: There is a slight right skew, as the tail on the right side is a bit longer, suggesting a presence of higher-priced properties.

Overall, the histogram provides a clear visualization of the distribution of average property prices, confirming the summary statistics and indicating a predominantly normal distribution with a slight right skew.

3. Compute the covariance matrix. Share your observations.

Covariance Matrix

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.516148									
AGE	0.562915	790.792473								
INDUS	-0.110215	124.267828	46.971430							
NOX	0.000625	2.381212	0.605874	0.013401						
DISTANCE	-0.229860	111.549955	35.479714	0.615710	75.666531					
TAX	-8.229322	2397.941723	831.713333	13.020502	1333.116741	28348.623600				
PTRATIO	0.068169	15.905425	5.680855	0.047304	8.743402	167.820822	4.677726			
AVG_ROOM	0.056118	-4.742538	-1.884225	-0.024555	-1.281277	-34.515101	-0.539695	0.492695		
LSTAT	-0.882680	120.838441	29.521811	0.487980	30.325392	653.420617	5.771300	-3.073655	50.893979	
AVG_PRICE	1.162012	-97.396153	-30.460505	-0.454512	-30.500830	-724.820428	-10.090676	4.484566	-48.351792	84.419556

CRIM_RATE and AVG_PRICE, AVG_ROOM and AVG_PRICE are positively related, rest all variables are negatively related with AVG_PRICE.

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_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	1									
AGE	0.006859	1								
INDUS	-0.005511	0.644779	1							
NOX	0.001851	0.731470	0.763651	1						
DISTANCE	-0.009055	0.456022	0.595129	0.611441	1					
TAX	-0.016749	0.506456	0.720760	0.668023	0.910228	1				
PTRATIO	0.010801	0.261515	0.383248	0.188933	0.464741	0.460853	1			
AVG_ROOM	0.027396	-0.240265	-0.391676	-0.302188	-0.209847	-0.292048	-0.355501	1		
LSTAT	-0.042398	0.602339	0.603800	0.590879	0.488676	0.543993	0.374044	-0.613808	1	
AVG_PRICE	0.043338	-0.376955	-0.483725	-0.427321	-0.381626	-0.468536	-0.507787	0.695360	-0.737663	1

Top 3 positively correlated pairs (Highlighted with yellow color) – TAX and Distance (0.91), NOX and INDUS (0.76) and NOX and AGE (0.73). Top 3 negative correlations(Highlighted with red color)– LSTAT and AVG_Price (-0.74), LSTAT and AVG_ROOM (-0.61) and PTRATIO and AVG_PRICE (-0.51).

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

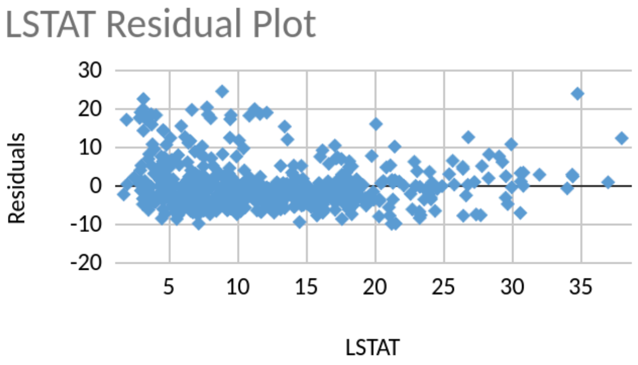
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.7376627262
R Square	0.5441462976
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	0
Residual	504	19472.38142	38.63567742		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95%*	*Upper 95%*
Intercept	34.55384	0.56263	61.41515	0	33.44846	35.65922	33.44846	35.65922
LSTAT	-0.95005	0.03873	-24.52790	0	-1.02615	-0.87395	-1.02615	-0.87395



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

The model has a R-squared value of 0.544 which suggests that this explains 54.4% of the variance in the AVG_PRICE. The intercept is 34.55 which suggests that even if LSTAT is 0, the value of AVG_PRICE will

be positive, i.e. 34.55. Looking t the residual plot, we see more concentration of points towards the lower values of LSTAT, visually it suggests that there might
be a pattern here, so we should explore more models so we could get a better model.

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

LSTAT has a significance value very close to 0, but it cannot be absolute 0. Since it is less than the significance level of 0.05,
this variable LSTAT is significant and should be retained in our analysis.

6. Build another instance of the Regression model but this time include LSTAT and AVG_ROOM variable together viz a viz AVG_PRICE as the dependent variable.

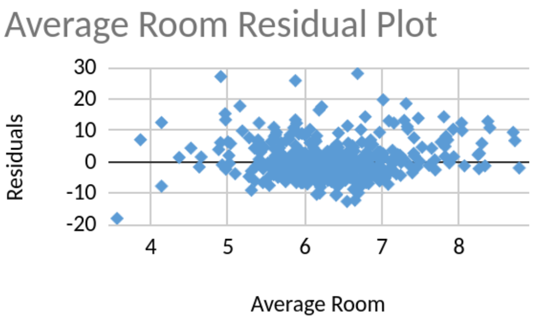
SUMMARY OUTPUT

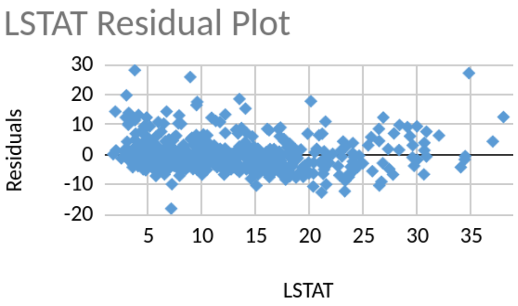
Regression Statistics	
Multiple R	0.7991
R Square	0.6386
Adjusted R Square	0.6371
Standard Error	5.5403
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	27276.986	13638.493	444.331	0
Residual	503	15439.309	30.694		
Total	505	42716.295			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95%*	*Upper 95%*
Intercept	-1.358273	3.172828	-0.428095	0.668765	-7.591900	4.875354	-7.591900	4.875354
Avg_Room	5.094788	0.444466	11.462730	0	4.221550	5.968025	4.221550	5.968025
LSTAT	-0.642358	0.043731	-14.688699	0	-0.728277	-0.556440	-0.728277	-0.556440





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?

Regression Equation
y = B1X1 + B2X2 + c
y = (5.09 * X1) – (0.64 * X2) - 1.36
X1 = 7 (AVG_ROOM)
X2 = 20 (LSTAT)Y = (5.09 * 7) – (0.64 * 20) – 1.36
Y = 35.63 – 12.8 – 1.36 = 21.47
The company is quoting a value of 30 against a prediction of 21.47, which suggests that the 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.

The R-squared value here is 0.64 as compared to 0.54 of the previous model, this means by adding AVG_ROOMS to our existing model, we are able to capture additional 10% of the variance in AVG_Price, because of which this is a better model than the previous one.

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	0.8330
R Square	0.6939
Adjusted R Square	0.6883
Standard Error	5.1348
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	29638.8605	3293.2067	124.9045	0
Residual	496	13077.4349	26.3658		
Total	505	42716.2954			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95%*	*Upper 95%*
Intercept	29.241300	4.817100	6.070300	0.000000	19.7768	38.7058	19.776800	38.7058
CRIME_RATE	0.048700	0.078400	0.621300	0.5347	-0.1053	0.2028	-0.105300	0.2028
AGE	0.032800	0.013100	2.502000	0.0127	0.007	0.0585	0.007000	0.0585
INDUS	0.1306	0.0631	2.0684	0.0391	0.0065	0.2546	0.0065	0.2546

NOX	-10.3212	3.894	-2.6505	0.0083	-17.972	-2.6703	-17.972	-2.6703
DISTANCE	0.2611	0.0679	3.8426	0.0001	0.1276	0.3946	0.1276	0.3946
TAX	-0.0144	0.0039	-3.6877	0.0003	-0.0221	-0.0067	-0.0221	-0.0067
PTRATIO	-1.0743	0.1336	-8.0411	0	-1.3368	-0.8118	-1.3368	-0.8118
AVG_ROOM	4.1254	0.4428	9.3175	0	3.2555	4.9953	3.2555	4.9953
LSTAT	-0.6035	0.0531	-11.3691	0	-0.7078	-0.4992	-0.7078	-0.4992

This particular model has an R-squared value of 0.6939 against a R-squared value of 0.64 in the previous model (with LSTAT and AVG_ROOM), this model captures more variance as compared to the previous model. Also, here the adjusted R-square value is 0.6883 suggesting that the significant variables are contributing to 68.83% of the variance. The intercept value is 29.24, suggesting that even if all the independent variables were zero, the AVG_PRICE would be 29.24. Looking at the p-values, CRIM_RATE should be dropped as its p-value is more than 0.05. Rest all variables are significant.

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) Answer the questions below:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8328
R Square	0.6936
Adjusted R Square	0.6887
Standard Error	5.1316
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	29628.6814	3703.5852	140.643	0
Residual	497	13087.614	26.3332		
Total	505	42716.295			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95%*	*Upper 95%*
Intercept	29.428500	4.804700	6.124900	0.000000	19.9884	38.8686	19.988400	38.8686
AGE	0.032900	0.013100	2.516600	0.0122	0.0072	0.0586	0.007200	0.0586
INDUS	0.130700	0.063100	2.072200	0.0388	0.0068	0.2546	0.006800	0.2546
NOX	-10.2727	3.8908	-2.6402	0.0085	-17.9172	-2.6282	-17.9172	-2.6282
DISTANCE	0.2615	0.0679	3.8512	0.0001	0.1281	0.3949	0.1281	0.3949
TAX	-0.0145	0.0039	-3.7039	0.0002	-0.0221	-0.0068	-0.0221	-0.0068
PTRATIO	-1.0717	0.1335	-8.0305	0	-1.3339	-0.8095	-1.3339	-0.8095
AVG_ROOM	4.1255	0.4425	9.3234	0	3.2561	4.9948	3.2561	4.9948

LSTAT	-0.6052	0.053	-11.4224	0	-0.7093	-0.5011	-0.7093	-0.5011

a. Interpret the output of this model.

This model explains 69.36% of the variance in AVG_PRICE. The intercept value is 29.42 suggesting that if all independent variables are 0, then the value of the house would be 29.42. All variables are significant here. This model is acceptable as it has a decent R-square and all variables 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?

Adjusted R-square for this model is 0.6887 vs 0.6883 in the previous model. Although adjusted R-square value is not up drastically, but we have all significant variables here, so is we consider these two factors together, then this model is a better model than the previous one.

c. Sort the values of the Coefficients in ascending order. What will happen to the average price if value of NOX is more in a locality in this town?

NOX and AVG_Price are negatively related. If the value of NOX increases then value of AVG_PRICE falls, more specifically every 1-unit increase in the value of NOX decreased the value of AVG_PRICE by 10.27.

d. Write the regression equation from this model.

$Y = 29.4285 + 0.0329 * X1 + 0.1307 * X2 + -10.2727 * X3 + 0.2615 * X4 – 0.0145 * X5 – 1.0717 * X6 + 4.1255 * X7 – 0.6052 * X8$