Problem: 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.

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

Answer

SUMMERY STATISTIC FOR EACH VARIABLE

CRIME_RATE	
Column1	Column2
Mean	4.87197628
Standard Error	0.129860152
Median	4.82
Mode	3.43
Standard Deviation	2.921131892
Sample Variance	8.533011532
Kurtosis	-1.189122464
Skewness	0.021728079
Range	9.95
Minimum	0.04
Maximum	9.99
Sum	2465.22
Count	506

AGE	
Column1	Column2
Mean	68.57490119
Standard Error	1.251369525
Median	77.5
Mode	100
Standard Deviatior	28.14886141
Sample Variance	792.3583985
Kurtosis	-0.967715594
Skewness	-0.59896264
Range	97.1
Minimum	2.9
Maximum	100
Sum	34698.9
Count	506

INDUS	
Column1	Column2
Mean	11.1367787
Standard Error	0.30497989
Median	9.69
Mode	18.1
Standard Deviatior	6.86035294
Sample Variance	47.0644425
Kurtosis	-1.2335396
Skewness	0.29502157
Range	27.28
Minimum	0.46
Maximum	27.74
Sum	5635.21
Count	506

NOX	
Column1	Column2
Mean	0.5547
Standard Error	0.00515
Median	0.538
Mode	0.538
Standard Deviatior	0.11588
Sample Variance	0.01343
Kurtosis	-0.0647
Skewness	0.72931
Range	0.486
Minimum	0.385
Maximum	0.871
Sum	280.676
Count	506

DISTANCE	
Column1	Column2
Mean	9.549407
Standard Error	0.387085
Median	5
Mode	24
Standard Deviatior	8.707259
Sample Variance	75.81637
Kurtosis	-0.867232
Skewness	1.004815
Range	23
Minimum	1
Maximum	24
Sum	4832
Count	506

TAX	
Column1	Column2
Mean	408.2371542
Standard Error	7.492388692
Median	330
Mode	666
Standard Deviatior	168.5371161
Sample Variance	28404.75949
Kurtosis	-1.142407992
Skewness	0.669955942
Range	524
Minimum	187
Maximum	711
Sum	206568
Count	506

PTRATIO	
Column1	Column2
Mean	18.4555336
Standard Error	0.096243568
Median	19.05
Mode	20.2
Standard Deviatior	2.164945524
Sample Variance	4.686989121
Kurtosis	-0.285091383
Skewness	-0.802324927
Range	9.4
Minimum	12.6
Maximum	22
Sum	9338.5
Count	506

AVG_ROOM	
Column1	Column2
Mean	6.28463439
Standard Error	0.03123514
Median	6.2085
Mode	5.713
Standard Deviatior	0.70261714
Sample Variance	0.49367085
Kurtosis	1.89150037
Skewness	0.40361213
Range	5.219
Minimum	3.561
Maximum	8.78
Sum	3180.025
Count	506

LSTAT	
Column1	Column2
Mean	12.6531
Standard Error	0.31746
Median	11.36
Mode	8.05
Standard Deviatior	7.14106
Sample Variance	50.9948
Kurtosis	0.49324
Skewness	0.90646
Range	36.24
Minimum	1.73
Maximum	37.97
Sum	6402.45
Count	506

AVG_PRICE	
Column1	Column2
Mean	22.53281
Standard Error	0.408861
Median	21.2
Mode	50
Standard Deviatior	9.197104
Sample Variance	84.58672
Kurtosis	1.495197
Skewness	1.108098
Range	45
Minimum	5
Maximum	50
Sum	11401.6
Count	506

Observation

1.The number of the house in dataset is given 506.

2.If we consider the DISTANCE, we can analyse that the maximum distance between the houses is 24 miles and has the same mode as 24. that the house is far away from each other.

3.If we consider the TAX, we can see that the average tax pay is 408.23 and the rang is 524.

4. From the skewness, we can see that the dataset is highly skewed.

5. if we see the AGE, the maximum age and mode is 100 which means that the age of the house is 100 years.

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

Answer

Observation

We can summarise that most of the houses are from range $21000 to $25000.
We have least count of houses from range $37000 to $41000 and $45000 to$49000.

Que3 Compute the covariance matrix. Share your observations

Ans

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	4.677726			
AVG_ROOM	0.056117778	-4.74253803	-1.884225427	-0.024554826	-1.281277391	-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	84.41955616

Observation

We can see that tax variable has high covariance values with each other feature except crime rate. That means tax explains a very good variability with other features

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

Correlation for each variable

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	1									
AGE	0.006859463	1								
INDUS	-0.005510651	0.644779	1							
NOX	0.001850982	0.73147	0.763651	1						
DISTANCE	-0.009055049	0.456022	0.595129	0.611441	1					
TAX	-0.016748522	0.506456	0.72076	0.668023	0.910228	1				
PTRATIO	0.010800586	0.261515	0.383248	0.188933	0.464741	0.460853	1			
AVG_ROOM	0.02739616	-0.24026	-0.39168	-0.30219	-0.20985	-0.29205	-0.3555	1		
LSTAT	-0.042398321	0.602339	0.6038	0.590879	0.488676	0.543993	0.374044	-0.613808272	1	
AVG_PRICE	0.043337871	-0.37695	-0.48373	-0.42732	-0.38163	-0.46854	-0.50779	0.695359947	-0.737662726	1

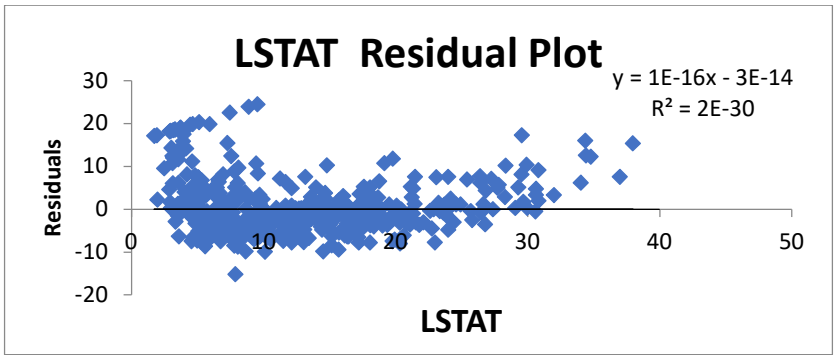
ans A	top 3 positively correlated pairs	0.73147	0.763651	0.910228
ans B	top 3 negatively correlated pairs	-0.50779	-0.613808272	-0.737662726

A.	top 3 positive correlated pairs	Distance --Tax
		NOX--Age
		NOX--Indus

B.	top 3 negatively correlated pairs	LSTAT--Avg_Room
		AVG_Price--LSTAT
		AVG_Price--PTRATIO

Que5 Build an initial regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot.

Ans



ANSWER (a)

the coefficient of LSTAT is -0.950049354.
This means if LSTAT increases by 0.9 times then average price of the house decreases 0.9 times.
Intercept for the model is 34.55384088

ANSWER (b)

Yes, LSTAT is significant variable for the avg_price from this model.

As the p-value (5.08E-88) we obtained from this model is away less than 0.05. By this we can say that LSTAT is a significant variable according to this model

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

Ans

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	27276.99	13638.49	444.3309	7.0085E-112
Residual	503	15439.31	30.69445		
Total	505	42716.3			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1.358272812	3.172828	-0.4281	0.668765	-7.591900282	4.875354658	-7.591900282	4.875354658
AVG_ROOM	5.094787984	0.444466	11.46273	3.47E-27	4.221550436	5.968025533	4.221550436	5.968025533
LSTAT	-0.642358334	0.043731	-14.6887	6.67E-41	-0.728277167	-0.556439501	-0.728277167	-0.556439501

A. Regression Equation we obtained for this model is :

y = -1.358 +5.09 X0 -0.642 X1
Where y=Avg_price, X0 = avg_room ,X1 = LSTAT
As per the model, avg_price for new house can be calculated as
Y= -1.358 + 5.09(7) -0.642(20) = 21.44
The price for the new house is $21440 .
we can say that company is Overcharging

B. y= -1.35 +5.09a -0.64b (Where a=Avg_room, b=LSTAT)
And Value of R square = 0.638561606 .
With this we can say that 63% of variability for average price is explained by
Avg_room and LSTAT combinedly and we obtained multiple R value as 0.79 which says it is highly correlated. But in previous model LSTAT alone describes 54% of variability for average price.

Que 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 Rsquare, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE.
Ans

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	19.7768278	38.70580267
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.534657201	-0.105348544	0.202798827	-0.10534854	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	0.00654109	0.254561704
NOX	-10.3211828	3.894036256	-2.650510195	0.008293859	-17.97202279	-2.670342809	-17.9720228	-2.670342809
DISTANCE	0.261093575	0.067947067	3.842602576	0.000137546	0.127594012	0.394593138	0.12759401	0.394593138
TAX	-0.01440119	0.003905158	-3.687736063	0.000251247	-0.022073881	-0.0067285	-0.02207388	-0.0067285
PTRATIO	-1.074305348	0.133601722	-8.041104061	6.58642E-15	-1.336800438	-0.811810259	-1.33680044	-0.811810259
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.89287E-19	3.255494742	4.995323561	3.25549474	4.995323561
LSTAT	-0.603486589	0.053081161	-11.36912937	8.91071E-27	-0.70777824	-0.499194938	-0.70777824	-0.499194938

Observation

From this we can say that crime rate is not a significant variable for average price of a house as p-value is greater than 0.5. All the features combinedly explains 69% of variability for average price of a house. NOX, TAX, PTRATIO and LSTAT have negative coefficients which says that increase in these features will result decrease in price of the house and vice versa.

Que8 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)

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	3703.58518	140.643041	1.911E-122
Residual	497	13087.61399	26.3332274		
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	6.12489816	1.846E-09	19.9883896	38.8686	19.9883896	38.8685574
AGE	0.03293496	0.013087055	2.51660595	0.01216288	0.00722219	0.05865	0.00722219	0.05864773
INDUS	0.130710007	0.063077823	2.07220226	0.03876167	0.00677794	0.25464	0.00677794	0.25464207
NOX	-10.27270508	3.890849222	-2.64022184	0.00854572	-17.9172457	-2.6282	-17.9172457	-2.62816447
DISTANCE	0.261506423	0.067901841	3.85124202	0.00013289	0.12809638	0.39492	0.12809638	0.39491647
TAX	-0.014452345	0.003901877	-3.70394641	0.00023607	-0.02211855	-0.0068	-0.02211855	-0.00678614
PTRATIO	-1.071702473	0.133453529	-8.03052927	7.0825E-15	-1.33390511	-0.8095	-1.33390511	-0.80949984
AVG_ROOM	4.125468959	0.44248544	9.32340046	3.6897E-19	3.2560963	4.99484	3.2560963	4.99484161
LSTAT	-0.605159282	0.0529801	-11.4223884	5.4184E-27	-0.70925186	-0.5011	-0.70925186	-0.5010667

From this we can conclude that all the features are significant variables for average price of the house.

(B)

Regression stats from previous model

Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372

Regression stats for this model.

Regression Statistics	
Multiple R	0.832835773
R Square	0.693615426

By comparing Multiple R and R square values for both the models we can conclude that both models perform well.

(C)

	Coefficients
Intercept	29.4284735
AVG_ROOM	4.12546896
DISTANCE	0.26150642
INDUS	0.13071001
AGE	0.03293496
TAX	-0.01445235
LSTAT	-0.60515928
PTRATIO	-1.07170247
NOX	-10.2727051

If NOX is more in the locality, according to this model average price of the house will decrease by 10 times.

(D)

Y=0.03293496 X0 + 0.130710007 X1 -10.27270508 X3 +0.261506423 X4 -0.014452345 X5 -1.071702473 X6 + 4.125468959 X7 -0.605159282 X8
+29.42847349
Where Y = average_Price
X0 = Age
X1 = Indus
X2 = NOX
X3 = Distance
X4 = TAX
X5 = PTRATIO
X6 = Avg_room
X7 = LSTAT