TERRO REAL ESTATE AGENCY

enerate the summary statistics for each variable in the
ble. (Use Data analysis tool pack). Write down your
bservation.

	4.871976	Mean	68.5749	Mean	11.13678	Mean	0.554695	Mean	9.549407	Mean	408.2372	Mean	18.45553	Mean	6.284634	Mean	12.65306	Mean	22.53281
lard I	0.12986	Standard I	1.25137	Standard I	0.30498	Standard I	0.005151	Standard I	0.387085	Standard I	7.492389	Standard I	0.096244	Standard I	0.031235	Standard I	0.317459	Standard I	0.408861
an	4.82	Median	77.5	Median	9.69	Median	0.538	Median	5	Median	330	Median	19.05	Median	6.2085	Median	11.36	Median	21.2
	3.43	Mode	100	Mode	18.1	Mode	0.538	Mode	24	Mode	666	Mode	20.2	Mode	5.713	Mode	8.05	Mode	50
lard I	2.921132	Standard I	28.14886	Standard I	6.860353	Standard I	0.115878	Standard I	8.707259	Standard I	168.5371	Standard I	2.164946	Standard I	0.702617	Standard I	7.141062	Standard I	9.197104
le Va	8.533012	Sample Va	792.3584	Sample Va	47.06444	Sample Va	0.013428	Sample Va	75.81637	Sample Va	28404.76	Sample Va	4.686989	Sample Va	0.493671	Sample Va	50.99476	Sample Va	84.58672
sis	-1.18912	Kurtosis	-0.96772	Kurtosis	-1.23354	Kurtosis	-0.06467	Kurtosis	-0.86723	Kurtosis	-1.14241	Kurtosis	-0.28509	Kurtosis	1.8915	Kurtosis	0.49324	Kurtosis	1.495197
ness	0.021728	Skewness	-0.59896	Skewness	0.295022	Skewness	0.729308	Skewness	1.004815	Skewness	0.669956	Skewness	-0.80232	Skewness	0.403612	Skewness	0.90646	Skewness	1.108098
e	9.95	Range	97.1	Range	27.28	Range	0.486	Range	23	Range	524	Range	9.4	Range	5.219	Range	36.24	Range	45
num	0.04	Minimum	2.9	Minimum	0.46	Minimum	0.385	Minimum	1	Minimum	187	Minimum	12.6	Minimum	3.561	Minimum	1.73	Minimum	5
mum	9.99	Maximum	100	Maximum	27.74	Maximum	0.871	Maximum	24	Maximum	711	Maximum	22	Maximum	8.78	Maximum	37.97	Maximum	50
	2465.22	Sum	34698.9	Sum	5635.21	Sum	280.6757	Sum	4832	Sum	206568	Sum	9338.5	Sum	3180.025	Sum	6402.45	Sum	11401.6
t	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506

EXPLANATION:

We can observe that the mean of age is high, in price the mode is $50.from the given data we can find the kurtosis values:

- CRIME_RATE is -1.1891, so the curve is not sharp, it is FLAT CURVE & POSITIVE SKEWNESS.
- AGE is-0.9677, so the curve is not sharp, it is FLAT CURVE& NEGATIVE SKEWNESS.
-
- INDUS is -1.23353, so the curve is not sharp, it is FLAT CURVE&POSITIVE SKEWNESS.
-
- NOX is -0.0646,so the curve is not sharp, it is FLAT CURVE & POSITIVE SKEWNESS.

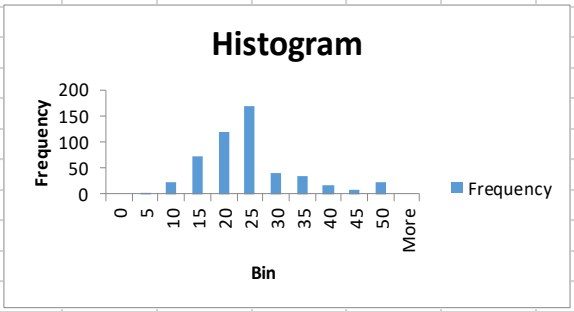
- DISTANCE is -0.86723,so the curve is not so sharp, it is FLAT CURVE &POSITIVE SKEWNESS.

- TAX is -1.1424, so the curve is not so sharp, it is FLAT CURVE & POSITIVE SKEWNESS.
- PTRATIO is -0.2850,so the curve is not sharp, it is FLAT CURVE & NEGATIVE SKEWNESS.

- AVG_ROOM is 1.8915,so the curve is not sharp, it is FLAT CURVE & POSITIVE SKEWNESS.
- LSTAT is 0.4932, so the curve is not sharp, it is FLAT CURVE & POSITIVE SKEWNESS.
- AVG_PRICE is 1.49519,so the curve is not sharp, it is FLAT CURVE & POSITIVE SKEWNESS.

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

Bin	Frequency
0	0
5	2
10	22
15	73
20	118
25	167
30	40
35	36
40	17
45	9
50	22
ore	0



XPLANATION:

rom the Histogram, it is inferred that Average Price has a positive skewers.

Compute the covariance matrix. Share your observations.

	CRIME_RAT	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	VG_ROOM	LSTAT	AVG_PRICE
IME_RA	8.516148									
E	0.562915	790.7925								
DUS	-0.11022	124.2678	46.97143							
X	0.000625	2.381212	0.605874	0.013401						
STANCE	-0.22986	111.55	35.47971	0.61571	75.66653					
X	-8.22932	2397.942	831.7133	13.0205	1333.117	28348.62				
RATIO	0.068169	15.90543	5.680855	0.047304	8.743402	167.8208	4.677726			
G_ROO	0.056118	-4.74254	-1.88423	-0.02455	-1.28128	-34.5151	-0.53969	0.492695		
TAT	-0.88268	120.8384	29.52181	0.48798	30.32539	653.4206	5.7713	-3.07365	50.89398	
G_PRIC	1.162012	-97.3962	-30.4605	-0.45451	-30.5008	-724.82	-10.0907	4.484566	-48.3518	84.41956

PLANATION:

gative value Denotes, both the X and Y values are mostly on opposite sides of their average.

eate a correlation matrix of all the variables (Use Data analysis tool pack).

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

Which are the top 3 positively correlated pairs

) 3 positively correlated pairs

(vs DISTANCE 0.910228

(vs INDUS 0.763651

X vs AGE 0.73147

Which are the top 3 negatively correlated pairs.

) 3 negatively correlated pairs

G_PRICE vs LSTAT -0.73766

AT vs AVG_ROOM -0.61381

G_PRICE vs PTRATIO -0.50779

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

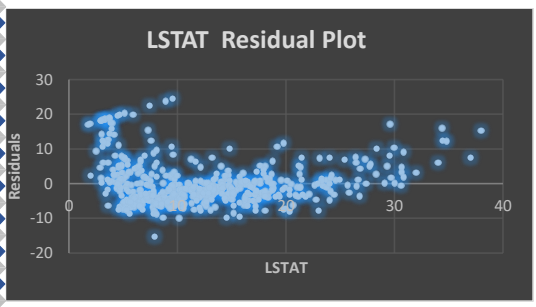
UMMARY OUTPUT

egression Statistics	
Multiple F	0.737663
Square	0.544146
djusted I	0.543242
tandard E	6.21576
)bservati	506

NOVA

	df	*SS*	*MS*	*F*
egressio	1	23243.91	23243.91	601.6179
esidual	504	19472.38	38.63568	
otal	505	42716.3		

	Coefficients	*tandard Err*	*t Stat*	*P-value*
ntercept	34.55384	0.562627	61.41515	3.7E-236
STAT	-0.95005	0.038733	-24.5279	5.08E-88



R Square 0.544146

Coefficient of LSTAT -0.95005

Intercept 34.55384

R square

R Square is just above 0.5 so this value is not significant. Square has to be near to 1.Coefficient of LSTAT

Coefficient of LSTAT is -0.95005. It is inferred that for each $1000 increase in Average price, there Will be a 0.95% decrease in population.

Intercept

It is inferred that the Intercept value is 34.5538.

Residual plot

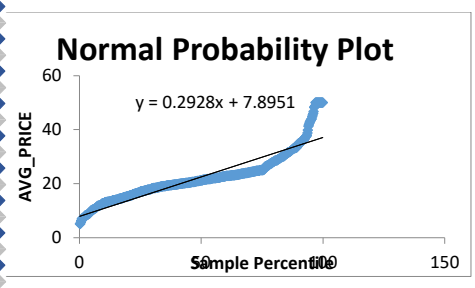
It is inferred that all the values are equally distributed

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

The p-value for LSTAT variable is 5.08110339438E-88. It is less than 0.05. So it is inferred that LSTAT variable is significant for the analysis.

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

UMMARY OUTPUT				
egression Statistics				
lultiple	0.7991			
Square	R Square			
djusted	0.637124			
tandard	5.540257			
bservati	506			
NOVA				
	df	*SS*	*MS*	*F*
egressic	2	27276.99	13638.49	444.3309
esidual	503	15439.31	30.69445	
otal	505	42716.3		
	Coefficients	*tandard Erı*	*t Stat*	*P-value*
ntercept	-1.35827	3.172828	-0.4281	0.668765
VG_ROO	5.094788	0.444466	11.46273	3.47E-27
STAT	-0.64236	0.043731	-14.6887	6.67E-41



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) AVG_PRICE = Intercept+ (Coefficient of AVG_ROOM * value of AVG_ROOM) +
c) (Coefficient of LSTAT * value of LSTAT)
d) AVG_PRICE = -1.35827281187456 + (5.09478798433655 * 7) + (-0.642358334244129 * 20)
e) AVG_PRICE = 21.4581
f) It is inferred that the Average price is $21.4581. But the company quoting a value of 30000 USD
g) for this locality. By the result, it is concluded that the company is overcharging.
h) R Square = 0.637124475470123 (Qn. 6)
i) R Square = 0.543241825954707 (Qn. 5)

c) 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.

It is inferred that the value of R Square is close to 1, if the count of independent variable Increases.

Based on the analysis, the performance of this model is better than the previous model.(Qn. 5)

Build another Regression model with all variables where AVG_PRICE alone be the Dependent Variable and all the other iables are independent. Interpret the output in terms of adjusted R Square, coefficient and Intercept values. Explain the nificance of each independent variable with respect to AVG_PRICE.

UMMARY OUTPUT				
egression Statistics				
ultiple	0.83298			
Square R Square				
djusted	0.6883			
tandard	5.13476			
bservat	506			
NOVA				
	df	*SS*	*MS*	*F*
egressi	9	29638.9	3293.21	124.905
esidual	496	13077.4	26.3658	
otal	505	42716.3		
	Coefficient	*tandard Err*	*t Stat*	*P-value*
ntercept	29.2413	4.81713	6.07028	2.5E-09
RIME_RA	0.04873	0.07842	0.62135	0.53466
GE	0.03277	0.0131	2.502	0.01267
NDUS	0.13055	0.06312	2.06839	0.03912
OX	-10.3212	3.89404	-2.65051	0.00829
ISTANCE	0.26109	0.06795	3.8426	0.00014
AX	-0.0144	0.00391	-3.68774	0.00025
TRATIO	-1.07431	0.1336	-8.0411	6.6E-15
VG_ROC	4.12541	0.44276	9.3175	3.9E-19
TAT	-0.60349	0.05308	-11.3691	8.9E-27

Pick out only the significant variables from the previous question. Make another instance of the Regression model using y the significant variables you just picked and answer the questions below:

Interpret the output of this model.

Compare the adjusted R-square value of this model with the model in the previous question, which model performs ter according to the value of adjusted R-square? c) Sort the values of the Coefficients in ascending order. What will ppen to the average price if the value of NOX is more in a locality in this town? d) Write the regression equation from s model.

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

Write the regression equation from this model.

gression Statistics	
ultiple F	0.8328
Square	0.6936
djusted	0.6887
tandarc	5.1316
bservat	506

NOVA

	df	SS	MS	F
egressi	8	29629	3703.6	140.64
esidual	497	13088	26.333	
otal	505	42716		

	Coefficient	andard Er	t Stat	P-value
tercept	29.428	4.8047	6.1249	2E-09
GE	0.0329	0.0131	2.5166	0.0122
DUS	0.1307	0.0631	2.0722	0.0388
OX	-10.27	3.8908	-2.64	0.0085
ISTANC	0.2615	0.0679	3.8512	0.0001
AX	-0.014	0.0039	-3.704	0.0002
TRATIC	-1.072	0.1335	-8.031	7E-15
VG_RC	4.1255	0.4425	9.3234	4E-19
STAT	-0.605	0.053	-11.42	5E-27

Adjusted R Square = 0.68868

usted R Square = 0.6886836818 (Qn.8)
usted R Square = 0.6882986468 (Qn.7)

the result, Adjusted R square for this model is greater comparing to the previous model.
it is concluded that this model performs better than previous model.

	Coefficients
OX	-10.2727
RATIO	-1.0717
TAT	-0.60516
X	-0.01445
GE	0.032935
DUS	0.13071
STANCE	0.261506
G_ROO	4.125469
tercept	29.42847

s inferred that if the value of NOX is more in a locality in this town, the value of the erage price will be reduced.

G_PRICE = Intercept + (coefficient of Age * value of Age) + (coefficient of Indus * value of Indus) +
efficient of NOX * value of NOX) + (coefficient of Distance * value of Distance) +
efficient of Tax * value of Tax) + (coefficient of PTRATIO * value of PTRATIO) +
efficient of Avg_room * value of Avg_room) +
efficient of LSTAT * value of LSTAT)

Commented [E1]: