TERRO’S REAL ESTATE AGENCY

(DATA ANALYSIS PROJECT)

SCHWARTZ A

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

CRIME RATE		*AGE*		*INDUS*	
Mean	4.871976285	Mean	68.57490119	Mean	11.13677866
Standard Error	0.129860152	Standard Error	1.251369525	Standard Error	0.304979888
Median	4.82	Median	77.5	Median	9.69
Mode	3.43	Mode	100	Mode	18.1
Standard Deviation	2.921131892	Standard Deviation	28.14886141	Standard Deviation	6.860352941
Sample Variance	8.533011532	Sample Variance	792.3583985	Sample Variance	47.06444247
Kurtosis	-1.18912246	Kurtosis	-0.967715594	Kurtosis	-1.233539601
Skewness	0.021728079	Skewness	-0.59896264	Skewness	0.295021568
Range	9.95	Range	97.1	Range	27.28
Minimum	0.04	Minimum	2.9	Minimum	0.46
Maximum	9.99	Maximum	100	Maximum	27.74
Sum	2465.22	Sum	34698.9	Sum	5635.21
Count	506	Count	506	Count	506

NOX		*DISTANCE*		*TAX*	
Mean	0.554695059	Mean	9.549407	Mean	408.23715
Standard Error	0.005151391	Standard Error	0.387085	Standard Error	7.4923887
Median	0.538	Median	5	Median	330
Mode	0.538	Mode	24	Mode	666
Standard Deviation	0.115877676	Standard Deviation	8.707259	Standard Deviation	168.53712
Sample Variance	0.013427636	Sample Variance	75.81637	Sample Variance	28404.759
Kurtosis	-0.06466713	Kurtosis	-0.86723	Kurtosis	-1.142408
Skewness	0.729307923	Skewness	1.004815	Skewness	0.6699559
Range	0.486	Range	23	Range	524
Minimum	0.385	Minimum	1	Minimum	187
Maximum	0.871	Maximum	24	Maximum	711
Sum	280.6757	Sum	4832	Sum	206568
Count	506	Count	506	Count	506

PTRATIO		*AVG ROOM*		*LSTAT*	
Mean	18.45553	Mean	6.284634	Mean	12.65306324
Standard Error	0.096244	Standard Error	0.031235	Standard Error	0.317458906
Median	19.05	Median	6.2085	Median	11.36
Mode	20.2	Mode	5.713	Mode	8.05
Standard Deviation	2.164946	Standard Deviation	0.702617	Standard Deviation	7.141061511
Sample Variance	4.686989	Sample Variance	0.493671	Sample Variance	50.99475951
Kurtosis	-0.28509	Kurtosis	1.8915	Kurtosis	0.493239517
Skewness	-0.80232	Skewness	0.403612	Skewness	0.906460094
Range	9.4	Range	5.219	Range	36.24
Minimum	12.6	Minimum	3.561	Minimum	1.73
Maximum	22	Maximum	8.78	Maximum	37.97
Sum	9338.5	Sum	3180.025	Sum	6402.45
Count	506	Count	506	Count	506

AVG PRICE	
Mean	22.53280632
Standard Error	0.408861147
Median	21.2
Mode	50
Standard Deviation	9.197104087
Sample Variance	84.58672359
Kurtosis	1.495196944
Skewness	1.108098408
Range	45
Minimum	5
Maximum	50
Sum	11401.6
Count	506

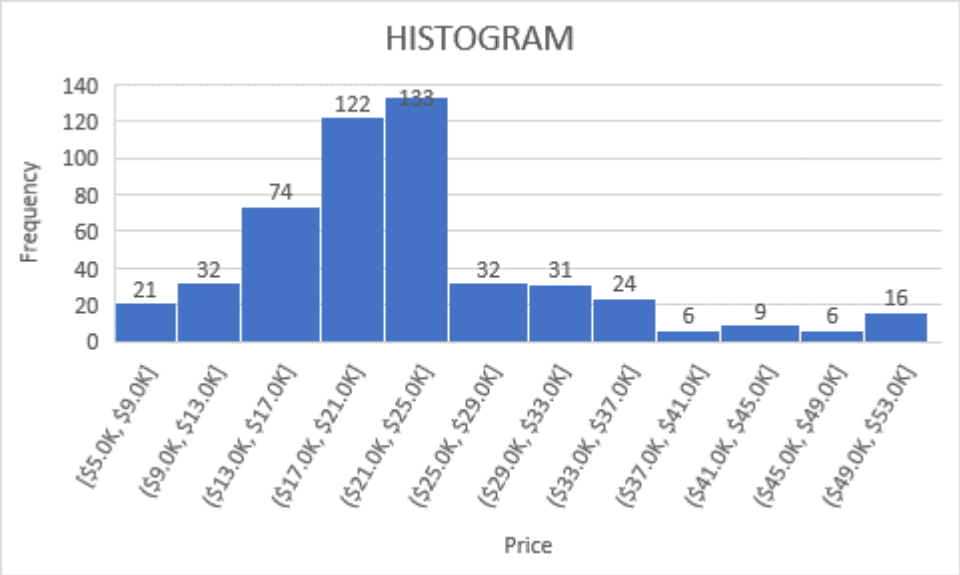
OBSERVATIONS :

In my perception **AVG_PRICE** of the house shows a positive skew indicating chances in slight presence of outliers.

TAX has the highest mean than any other variables or other column data and **NOX** has the lowest mode value

Highest variance is found in **TAX** as currency is involved here.

2) Plot a histogram of the AVG_PRICE variable. What do you infer?



OBSERVATIONS:

Most of the houses are having $21K TO $25K as their average price.

Very few houses are having $45K TO $49K as their average price.

It shows positive kurtosis.

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	0.562915	-0.11022	0.000625	-0.22986	-8.22932	0.068169	0.056117778	-0.88268	1.16201224
AGE	0.562915215	790.7925	124.2678	2.381212	111.55	2397.942	15.90543	-4.74253803	120.8384	-97.396153
INDUS	-0.110215175	124.2678	46.97143	0.605874	35.47971	831.7133	5.680855	-1.884225427	29.52181	-30.460505
NOX	0.000625308	2.381212	0.605874	0.013401	0.61571	13.0205	0.047304	-0.024554826	0.48798	-0.4545124
DISTANCE	-0.229860488	111.55	35.47971	0.61571	75.66653	1333.117	8.743402	-1.281277391	30.32539	-30.50083
TAX	-8.229322439	2397.942	831.7133	13.0205	1333.117	28348.62	167.8208	-34.51510104	653.4206	-724.82043
PTRATIO	0.068168906	15.90543	5.680855	0.047304	8.743402	167.8208	4.677726	-0.539694518	5.7713	-10.090676
AVG ROOM	0.056117778	-4.74254	-1.88423	-0.02455	-1.28128	-34.5151	-0.53969	0.492695216	-3.07365	4.48456555
LSTAT	-0.882680362	120.8384	29.52181	0.48798	30.32539	653.4206	5.7713	-3.073654967	50.89398	-48.351792
AVG PRICE	1.16201224	-97.3962	-30.4605	-0.45451	-30.5008	-724.82	-10.0907	4.484565552	-48.3518	84.4195562

OBSERVATIONS:

AVG_ROOM and CRIME_RATE shows positive co-variance with AVG_PRICE of the house which means they are directly proportional to each other.

Few pairs showing positive co-variance are (TAX,TAX),(AGE,TAX),(DISTANCE,TAX)

Few pairs showing negative co-variance are (CRIME RATE,INDUS),(CRIME RATE,DISTANCE) (DISTANCE,AVG_PRICE).

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

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	1.00	0.01	-0.01	0.00	-0.01	-0.02	0.01	0.03	-0.04	0.04
AGE	0.01	1.00	0.64	0.73	0.46	0.51	0.26	-0.24	0.60	-0.38
INDUS	-0.01	0.64	1.00	0.76	0.60	0.72	0.38	-0.39	0.60	-0.48
NOX	0.00	0.73	0.76	1.00	0.61	0.67	0.19	-0.30	0.59	-0.43
DISTANCE	-0.01	0.46	0.60	0.61	1.00	0.91	0.46	-0.21	0.49	-0.38
TAX	-0.02	0.51	0.72	0.67	0.91	1.00	0.46	-0.29	0.54	-0.47
PTRATIO	0.01	0.26	0.38	0.19	0.46	0.46	1.00	-0.36	0.37	-0.51
AVG_ROOM	0.03	-0.24	-0.39	-0.30	-0.21	-0.29	-0.36	1.00	-0.61	0.70
LSTAT	-0.04	0.60	0.60	0.59	0.49	0.54	0.37	-0.61	1.00	-0.74
AVG_PRICE	0.04	-0.38	-0.48	-0.43	-0.38	-0.47	-0.51	0.70	-0.74	1.00

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	100%	1%	-1%	0%	-1%	-2%	1%	3%	-4%	4%
AGE	1%	100%	64%	73%	46%	51%	26%	-24%	60%	-38%
INDUS	-1%	64%	100%	76%	60%	72%	38%	-39%	60%	-48%
NOX	0%	73%	76%	100%	61%	67%	19%	-30%	59%	-43%
DISTANCE	-1%	46%	60%	61%	100%	91%	46%	-21%	49%	-38%
TAX	-2%	51%	72%	67%	91%	100%	46%	-29%	54%	-47%
PTRATIO	1%	26%	38%	19%	46%	46%	100%	-36%	37%	-51%
AVG_ROOM	3%	-24%	-39%	-30%	-21%	-29%	-36%	100%	-61%	70%
LSTAT	-4%	60%	60%	59%	49%	54%	37%	-61%	100%	-74%
AVG_PRICE	4%	-38%	-48%	-43%	-38%	-47%	-51%	70%	-74%	100%

a) Which are the top 3 positively correlated pairs

(TAX , DISTANCE = 91%)
(INDUS , NOX = 76%)
(NOX , AGE = 73%)

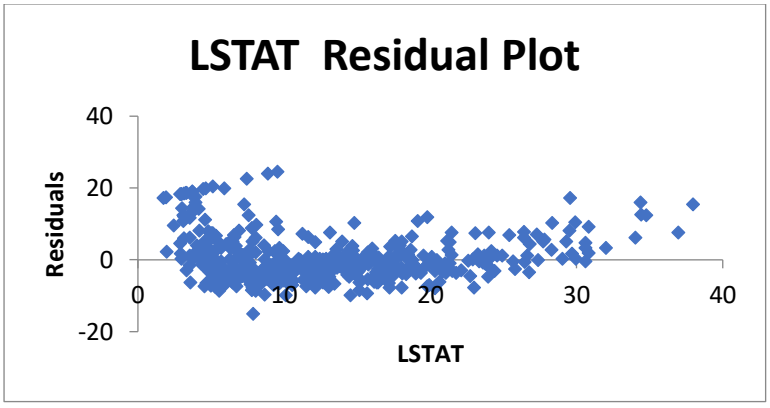
b) Which are the top 3 negatively correlated pairs.

(LSTAT, AVG_PRICE = -74%)
(LSTAT , AVG_ROOM = -61%)
(PTRATIO , AVG_PRICE =-51%)

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

Regression Statistics	
Multiple R	0.737662726
R Square	0.544146298
Adjusted R Square	0.543241826
Standard Error	6.215760405
Observations	506

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	34.55384088	0.562627355	61.41515	3.7E-236	33.44845704	35.65922472	33.44845704	35.65922472
X Variable 1	-0.950049354	0.038733416	-24.5279	5.08E-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 Residual plot?

INTERCEPT = 34.554

CO-EFFICIENT = -0.95004

The co-efficient of LSTAT is negative so we can say that LSTAT and AVG_PRICE are inversely proportional to each other and from the graph we can say that the plots are scattered and not in pattern.

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

Since the P value of this model is 3.7E-236 which is lesser than 0.05, we can say that LSTAT is a significant variable for this model.

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

Regression Statistics	
Multiple R	0.7991
R Square	0.638562
Adjusted R Square	0.637124
Standard Error	5.540257
Observations	506

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.35827	3.172828	-0.4281	0.668765	-7.5919	4.875355	-7.5919	4.875355
AVG-ROOM	5.094788	0.444466	11.46273	3.47E-27	4.22155	5.968026	4.22155	5.968026
LSTAT	-0.64236	0.043731	-14.6887	6.67E-41	-0.72828	-0.55644	-0.72828	-0.55644

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?

AVG_PRICE=CO-EFFICIENT OF AVG_ROOM(AVG_ROOM)+CO-EFFICIENT OF LSTAT(LSTAT)+INTERCEPT

= (5.094788 * 7) + (-0.64236*20)+(-1.35827)

= **21.45808** is the AVG_PRICE predicted

How does it compare to the company quoting a value of 30000 USD for this locality? Is the company Overcharging/ Undercharging?

so as company charges **$30k** for this locality , it is overcharging . Because AVG_PRICE predicted is **$21.45k** which is lesser than **30000 USD**

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

Yes , performance of this model is better than previous. Because, 63%>54%(ADJUSTED R-SQUARE).

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 R square, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE.

Regression Statistics	
Multiple R	0.832979
R Square	0.693854
Adjusted R Square	0.688299
Standard Error	5.134764
Observations	506

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.24132	4.817125596	6.070283	2.53978E-09	19.77682784	38.7058027	19.7768278	38.7058027
CRIME_RATE	0.048725	0.078418647	0.621346	0.534657201	-0.105348544	0.20279883	-0.1053485	0.20279883
AGE	0.032771	0.013097814	2.501997	0.012670437	0.00703665	0.05850473	0.00703665	0.05850473
INDUS	0.130551	0.063117334	2.068392	0.03912086	0.006541094	0.2545617	0.00654109	0.2545617
NOX	-10.3212	3.894036256	-2.65051	0.008293859	-17.97202279	-2.6703428	-17.972023	-2.67034281
DISTANCE	0.261094	0.067947067	3.842603	0.000137546	0.127594012	0.39459314	0.12759401	0.39459314
TAX	-0.0144	0.003905158	-3.68774	0.000251247	-0.022073881	-0.0067285	-0.0220739	-0.0067285
PTRATIO	-1.07431	0.133601722	-8.0411	6.58642E-15	-1.336800438	-0.8118103	-1.3368004	-0.81181026
AVG_ROOM	4.125409	0.442758999	9.317505	3.89287E-19	3.255494742	4.99532356	3.25549474	4.99532356
LSTAT	-0.60349	0.053081161	-11.3691	8.91071E-27	-0.70777824	-0.4991949	-0.7077782	-0.49919494

Adjusted R-SQUARE value is 68% (0.688299)

This model has good adjusted R_SQUARE . So, this could be a nice model and could be used in prediction

AVG_ROOM has the highest co-efficient value

Significant variables are AGE,INDUS,NOX,DISTANCE,LSTAT,PTRATIO,AVG_ROOM,TAX

Non Significant variable is CRIME_RATE.

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 and answer the questions below:

Regression Statistics	
Multiple R	0.832836
R Square	0.693615
Adjusted R Square	0.688684
Standard Error	5.131591
Observations	506

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.42847	4.804728624	6.124898	1.84597E-09	19.98838959	38.86856	19.9883896	38.8685574
AGE	0.032935	0.013087055	2.516606	0.012162875	0.007222187	0.058648	0.00722219	0.05864773
INDUS	0.13071	0.063077823	2.072202	0.038761669	0.006777942	0.254642	0.00677794	0.25464207
NOX	-10.2727	3.890849222	-2.64022	0.008545718	-17.9172457	-2.62816	-17.917246	-2.62816447
DISTANCE	0.261506	0.067901841	3.851242	0.000132887	0.128096375	0.394916	0.12809638	0.39491647
TAX	-0.01445	0.003901877	-3.70395	0.000236072	-0.022118553	-0.00679	-0.0221186	-0.00678614
PTRATIO	-1.0717	0.133453529	-8.03053	7.08251E-15	-1.333905109	-0.8095	-1.3339051	-0.80949984
AVG_ROOM	4.125469	0.44248544	9.3234	3.68969E-19	3.256096304	4.994842	3.2560963	4.99484161
LSTAT	-0.60516	0.0529801	-11.4224	5.41844E-27	-0.70925186	-0.50107	-0.7092519	-0.5010667

a) Interpret the output of this model

This model has R-SQUARE =0.693615 and adjusted R-SQUARE=0.688684

So this model can be used for prediction and all **p values** are less than 0.05 shows significance of the variables picked.

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 of this model = 0.688684

Adjusted R-SQUARE of previous model = 0.688299

Almost seems to have same adjusted R-SQUARE , but still **0.688684>0.688299**.So,we can say this model can perform better than previous one

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?

Coefficients in Ascending order,

-10.2727	NOX
-1.0717	PTRATIO
-0.60516	LSTAT
-0.01445	TAX
0.032935	AGE
0.13071	INDUS
0.261506	DISTANCE
4.125469	AVG_ROOM

-0.42732 is the correlation between NOX and AVG_PRICE . Since sign is negative indicates that both are inversly proportional to each other. So if NOX is more the AVG_PRICE gets decreased.

d) Write the regression equation from this model.

AVG_PRICE=(CO.EF(AGE)*AGE)+(C0.EFF(INDUS)*INDUS)+(COEFF(NOX)*NOX)+(COEFF(DISTANCE)*DISTANCE)+(COEFF(TAX)*TAX)+(COEFF(PTRATIO)*PTRATIO)+(COEFF(AVG ROOM)*AVG_ROOM)+(COEFF(LSTAT)*LSTAT)+INTERCCEPT