TERRO'S REAL ESTATE AGENCY

BUSINESS REPORT:

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.

Objective:

To analyze the magnitude of each variable to which it can affect the price of a house in a particular locality.

Question 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	-1.189122464	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

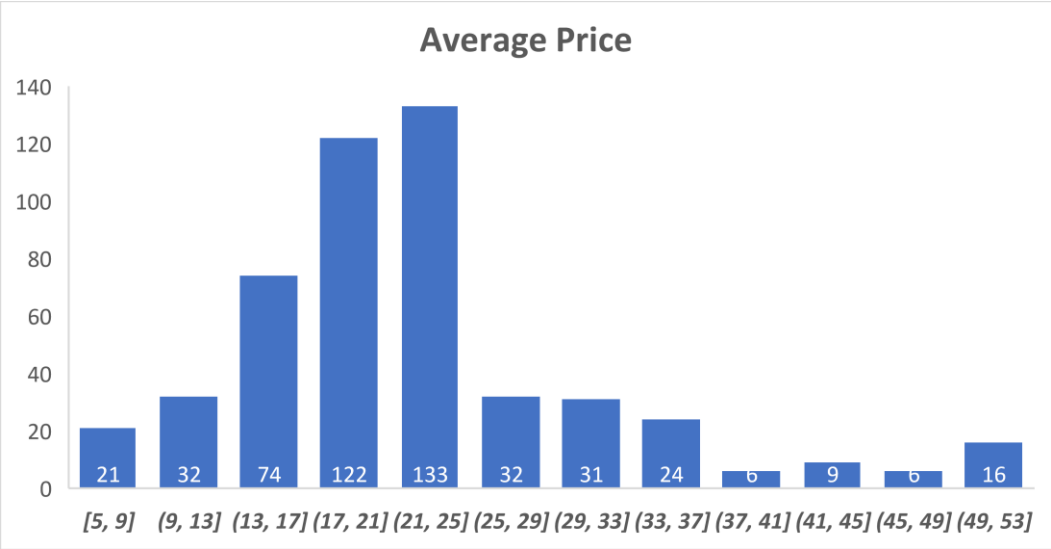
NOX		*DISTANCE*		*TAX*	
Mean	0.554695059	Mean	9.549407115	Mean	408.2371542
Standard Error	0.005151391	Standard Error	0.387084894	Standard Error	7.492388692
Median	0.538	Median	5	Median	330
Mode	0.538	Mode	24	Mode	666
Standard Deviation	0.115877676	Standard Deviation	8.707259384	Standard Deviation	168.5371161
Sample Variance	0.013427636	Sample Variance	75.81636598	Sample Variance	28404.75949
Kurtosis	-0.064667133	Kurtosis	-0.867231994	Kurtosis	-1.142407992
Skewness	0.729307923	Skewness	1.004814648	Skewness	0.669955942
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*		*AVG_PRICE*	
Mean	18.4555336	Mean	6.284634387	Mean	12.65306324	Mean	22.53280632
Standard Error	0.096243568	Standard Error	0.031235142	Standard Error	0.317458906	Standard Error	0.408861147
Median	19.05	Median	6.2085	Median	11.36	Median	21.2
Mode	20.2	Mode	5.713	Mode	8.05	Mode	50
Standard Deviation	2.164945524	Standard Deviation	0.702617143	Standard Deviation	7.141061511	Standard Deviation	9.197104087
Sample Variance	4.686989121	Sample Variance	0.49367085	Sample Variance	50.99475951	Sample Variance	84.58672359
Kurtosis	-0.28509138	Kurtosis	1.891500366	Kurtosis	0.493239517	Kurtosis	1.495196944
Skewness	-0.80232493	Skewness	0.403612133	Skewness	0.906460094	Skewness	1.108098408
Range	9.4	Range	5.219	Range	36.24	Range	45
Minimum	12.6	Minimum	3.561	Minimum	1.73	Minimum	5
Maximum	22	Maximum	8.78	Maximum	37.97	Maximum	50
Sum	9338.5	Sum	3180.025	Sum	6402.45	Sum	11401.6
Count	506	Count	506	Count	506	Count	506

From descriptive statistics of the given dataset, we can get few observations as:

- The number of records given in the dataset are 506.
- Firstly, if we consider Distance variable, we can analyse that maximum distance is 24 and has mode as 24. Which says that most of the houses are away from Highway.
- The average tax paid is 408.2 and tax range is 524.
- From the skewness of variables, we can say that dataset is highly skewed.
- And if we consider age variable the maximum age is 100 and mode is also 100 which says that most of the houses has age of 100.

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



From above Histogram,

- 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.

Question 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									
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	-48.3517922	84.41956

From above matrix we can get assumptions as :

- As we can see there is high covariance value for some of the features which tells that they are higly correlated and explains the variability of the other features.
- 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

Question 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									
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

a) Which are the top 3 positively correlated pairs.

From above correlation matrix we can analyse the top 3 positively correlated pairs as

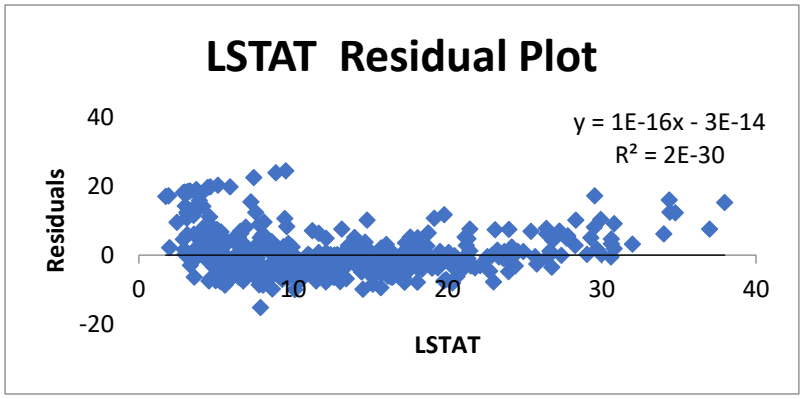
- 1.Distance – Tax
- 2.NOX – Age
- 3.NOX – Indus

b) Which are the top 3 negatively correlated pairs.

From above correlation matrix we can analyse the top 3 negatively correlated pairs as

- 1.LSTAT – Avg_Room
- 2.Avg_Price – PTRATIO
- 3.Avg_Price – LSTAT

Question 5 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?

- From this model 54% of the variation in the average price is explained by the LSTAT.
- The coefficient of LSTAT for the model is -0.950049354. This says that if LSTAT increases by 0.9 times then average price of house decreases 0.9times.
- Intercept of LSTAT for the model is 34.55384088.

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

- 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.

Question 6 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. 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 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

- So , the price for the new house is $21440 .
- we can say that 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 and explain.

- Yes, the performance of this model performs well compared to previous model.
- From this model the linear equation we obtained is

  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 combinely 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.

Question 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.

From the model we can obtain coefficients and p values as below:

	Coefficients	*P-value*
Intercept	29.24131526	2.53978E-09
CRIME_RATE	0.048725141	0.534657201
AGE	0.032770689	0.012670437
INDUS	0.130551399	0.03912086
NOX	-10.3211828	0.008293859
DISTANCE	0.261093575	0.000137546
TAX	-0.01440119	0.000251247
PTRATIO	-1.074305348	6.58642E-15
AVG_ROOM	4.125409152	3.89287E-19
LSTAT	-0.603486589	8.91071E-27

- From this we can say that crime rate is not a significant variable for average price of an house as p-value is greater than 0.5.
- All the features combinely 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.

Question 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:

a) Interpret the output of this model.

	Coefficients	*P-value*
Intercept	22.52588301	1.51363E-07
DISTANCE	0.206811735	0.00173655
TAX	-0.011644432	0.000815915
PTRATIO	-0.941134259	3.68488E-13
AVG_ROOM	4.325996678	1.61122E-21
LSTAT	-0.555793401	8.63707E-29

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

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?

- Regression stats from previous model

Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372

- Regression stats for this model.

Regression Statistics	
Multiple R	0.82818595
R Square	0.685891968

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

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
NOX	-10.3211828
PTRATIO	-1.07430535
LSTAT	-0.60348659
TAX	-0.01440119
AGE	0.032770689
CRIME_RATE	0.048725141
INDUS	0.130551399
DISTANCE	0.261093575
AVG_ROOM	4.125409152
Intercept	29.24131526

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

d)Write the regression equation from this model.

- Y^=29.24131526−10.3211828·NOX−1.07430535·PTRATIO−0.60348659·LSTAT−0.0144 0119·TAX+0.032770689·AGE+0.048725141·CRIME_RATE+0.130551399·IN-DUS+0.261093575·DISTANCE+4.125409152·AVG_ROOM
- Where Y = average Price

Summary:

- From this Analysis, we can conclude that all the features play a vital role in estimating the average price of the house excluding crime rate.
- And few features have negative coefficients which say that increase rate in those features will decrease the average price of the house like NOX, PTRATIO, TAX and LSTAT.