Terro’s real estate agency

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 had employed me, where I studied 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.

Situation:

“Finding out the most relevant features for pricing of a house”

Task:

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

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

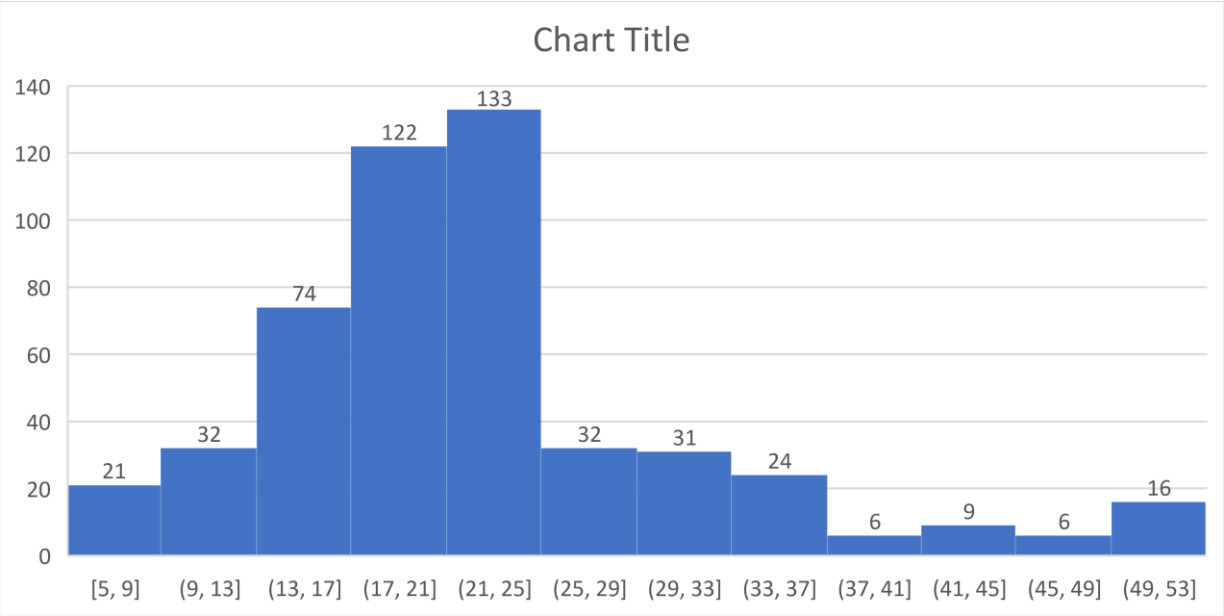
CRIME_RATE		*AGE*		*INDUS*		*NOX*	
Mean	4.871976285	Mean	68.57490119	Mean	11.13677866	Mean	0.554695059
Standard Error	0.129860152	Standard Error	1.251369525	Standard Error	0.304979888	Standard Error	0.005151391
Median	4.82	Median	77.5	Median	9.69	Median	0.538
Mode	3.43	Mode	100	Mode	18.1	Mode	0.538
Standard Deviation	2.921131892	Standard Deviation	28.14886141	Standard Deviation	6.860352941	Standard Deviation	0.115877676
Sample Variance	8.533011532	Sample Variance	792.3583985	Sample Variance	47.06444247	Sample Variance	0.013427636
Kurtosis	-1.189122464	Kurtosis	-0.967715594	Kurtosis	-1.233539601	Kurtosis	-0.064667133
Skewness	0.021728079	Skewness	-0.59896264	Skewness	0.295021568	Skewness	0.729307923
Range	9.95	Range	97.1	Range	27.28	Range	0.486
Minimum	0.04	Minimum	2.9	Minimum	0.46	Minimum	0.385
Maximum	9.99	Maximum	100	Maximum	27.74	Maximum	0.871
Sum	2465.22	Sum	34698.9	Sum	5635.21	Sum	280.6757
Count	506	Count	506	Count	506	Count	506

DISTANCE		*TAX*		*PTRATIO*		*AVG_ROOM*	
Mean	9.549407115	Mean	408.2371542	Mean	18.4555336	Mean	6.284634387
Standard Error	0.387084894	Standard Error	7.492388692	Standard Error	0.096243568	Standard Error	0.031235142
Median	5	Median	330	Median	19.05	Median	6.2085
Mode	24	Mode	666	Mode	20.2	Mode	5.713
Standard Deviation	8.707259384	Standard Deviation	168.5371161	Standard Deviation	2.164945524	Standard Deviation	0.702617143
Sample Variance	75.81636598	Sample Variance	28404.75949	Sample Variance	4.686989121	Sample Variance	0.49367085
Kurtosis	-0.867231994	Kurtosis	-1.142407992	Kurtosis	-0.285091383	Kurtosis	1.891500366
Skewness	1.004814648	Skewness	0.669955942	Skewness	-0.802324927	Skewness	0.403612133
Range	23	Range	524	Range	9.4	Range	5.219
Minimum	1	Minimum	187	Minimum	12.6	Minimum	3.561
Maximum	24	Maximum	711	Maximum	22	Maximum	8.78
Sum	4832	Sum	206568	Sum	9338.5	Sum	3180.025
Count	506	Count	506	Count	506	Count	506

LSTAT		*AVG_PRICE*	
Mean	12.65306324	Mean	22.53280632
Standard Error	0.317458906	Standard Error	0.408861147
Median	11.36	Median	21.2
Mode	8.05	Mode	50
Standard Deviation	7.141061511	Standard Deviation	9.197104087
Sample Variance	50.99475951	Sample Variance	84.58672359
Kurtosis	0.493239517	Kurtosis	1.495196944
Skewness	0.906460094	Skewness	1.108098408
Range	36.24	Range	45
Minimum	1.73	Minimum	5
Maximum	37.97	Maximum	50
Sum	6402.45	Sum	11401.6
Count	506	Count	506

- From Descriptive Statistical Summary Analysis, we can see that there are no varying data counts in all the data dictionaries the total count is about 506.
- From the Crime Rate analysis, we see the Max per capita crime rate by Town is 9.99 and mean rating is about 4.87.
- Mean Age of Houses is about 68.58 and many numbers of houses are of 100 years.
- Average proportion of non-retail business acres per town is about 11.14 acres.
- Average and median-mode NOX (nitric oxides concentration) is about 0.557 and 0.538 (parts per 10 million).
- The Average distance from highway is about 9.55 miles and the mode about 24 which is also the maximum miles.
- The average full-value property-tax rate per $10,000 is $408.23 minimum with 187 and maximum of 711.
- The average PTRATIO is about 18.45 so we can say that the availability of teacher for the pupil's education is good and the max is also about 22 which we can say a good ratio.
- The average number of rooms per house is 6.28 and with a mode of 5.7.
- The lower status of the population is about 12.65% and maximum about 37.97%
- The Average value of houses is about $21000 and max goes about $50000.

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



- The Histogram is clearly showing us that the most of the average price of the property is laying between $17k to $25k which is covering about 50.39% of the property, most at 21 to 25k.
- Least between 37-41k and 45-49k.

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	-48.35179219	84.41955616

- from the covariance matrix we can see that **tax and age** are having the stronger relation as compare to all other attributes.
- whereas the **avg_price and tax** are more negatively covariant.

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

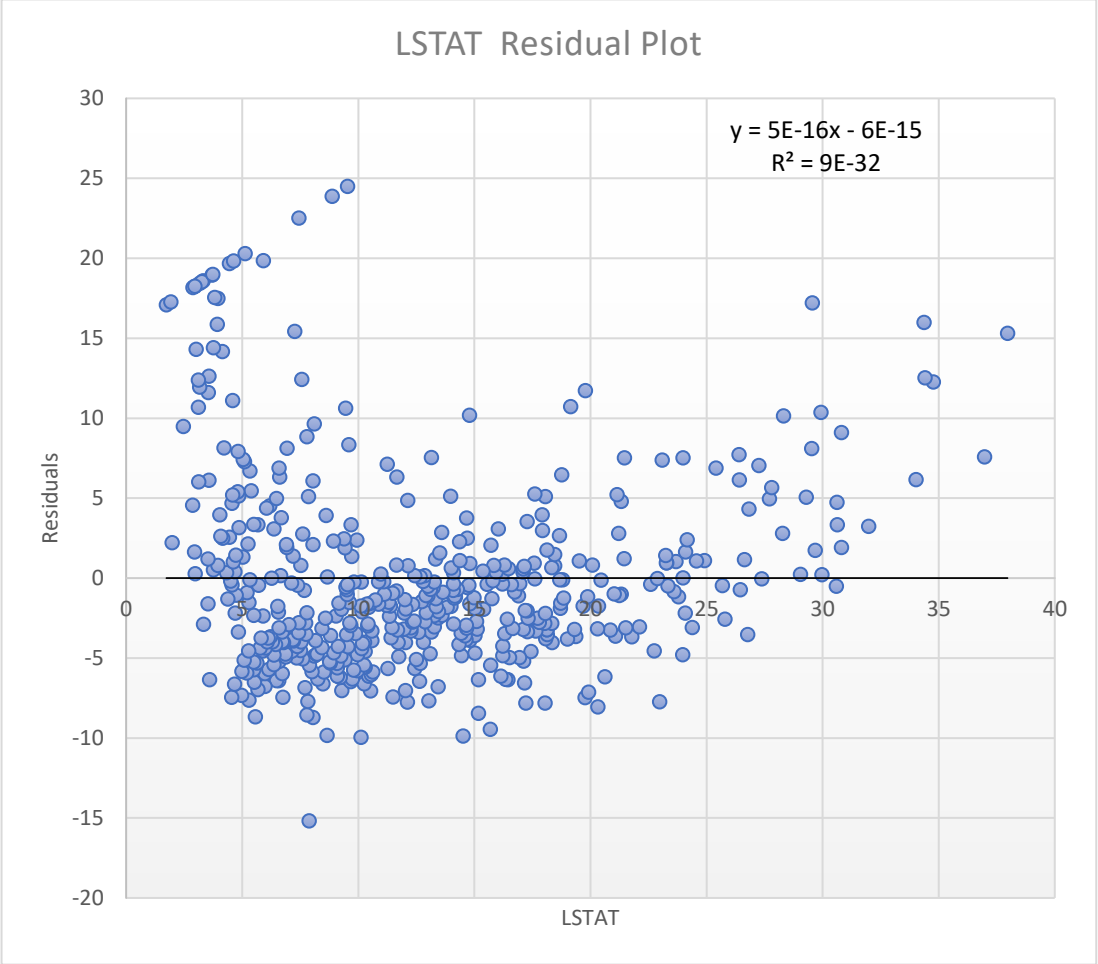
a) Which are the top 3 positively correlated pairs.

1) **TAX** and **DISTANCE** with correlation about **0.91**
2) **NOX** and **INDUS** with correlation about **0.76**
3) **NOX** and **AGE** with a correlation about **0.73**

b) Which are the top 3 negatively correlated pairs.

1) **AVG_PRICE** and **LSTAT** with a negative correlation about **-0.74**
2) **LSTAT** and **AVG_ROOM** with a negative correlation about **-0.61**
3) **AVG_PRICE** and **PTRATIO** with a negative correlation about **-0.51**

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?

- the variance is the R square value which is also known as the covariance of determination it is having the value of about **0.544** that is AVG_PRICE and LSTAT are **54%** dependable.
- the coefficient value is negative between AVG_PRICE and LSTAT that is as the average price and increases the lower status of the population decreases.
- the intercept value is the expected value of the response variable LSTAT is 34.5 but the predicted value is **-0.95**.
- the residual plot is randomly scattered over the plane with variance between each other which shows us that the difference between the observed and predicted values of the dependent variable for each observation.

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

From the summary output and with the least p-value (5.08E-88) we can say that LSTAT is significant about 54%

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	0.799100498							
R Square	0.638561606							
Adjusted R Square	0.637124475							
Standard Error	5.540257367							
Observations	506							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	27276.98621	13638.49311	444.3308922	7.0085E-112			
Residual	503	15439.3092	30.69445169					
Total	505	42716.29542						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.358272812	3.17282778	-0.428095348	0.668764941	-7.591900282	4.875354658	-7.591900282	4.875354658
AVG_ROOM	5.094787984	0.4444655	11.46272991	3.47226E-27	4.221550436	5.968025533	4.221550436	5.968025533
LSTAT	-0.642358334	0.043731465	-14.68869925	6.66937E-41	-0.728277167	-0.556439501	-0.728277167	-0.556439501

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=5.09478798433654*X1 - 0.642358334244129*X2 -1.35827281187446**

  Where Y=AVG_PRICE

  X1= AVG_ROOM

  X2= LSTAT

After calculating with **X1=7** and **X2=20** the value of **Y** is **21.458**, thus the company is overcharging about **$8541.92**

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 is better than that of previous comparing in terms of adjusted R-square value.

Where the adjusted R-square value of this model is **0.637124475470123** whereas previous is **0.543241825954707**

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	0.832978824
R Square	0.69385372
Adjusted R Square	0.688298647
Standard Error	5.1347635
Observations	506

	Coefficients	*Standard Error*	*t Stat*	*P-value*
LSTAT	-0.603486589	0.053081161	-11.36912937	8.91071E-27
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.89287E-19
PTRATIO	-1.074305348	0.133601722	-8.041104061	6.58642E-15
Intercept	29.24131526	4.817125596	6.070282926	2.53978E-09
DISTANCE	0.261093575	0.067947067	3.842602576	0.000137546
TAX	-0.01440119	0.003905158	-3.687736063	0.000251247
NOX	-10.3211828	3.894036256	-2.650510195	0.008293859
AGE	0.032770689	0.013097814	2.501996817	0.012670437
INDUS	0.130551399	0.063117334	2.068392165	0.03912086
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.534657201

- Referring the above table p-value, we can say that LSTAT is most significant and CRIME_RATE is not significant.
- With adjusted R Square value of 0.688 the R Square is still 0.69 so we can say that the model is 69% significant.
- AVG_ROOM, DISTANCE, AGE, INDUS have positive coefficients so we can say that increase in these values will increase the value of AVG_PRICE.
- LSTAT, PTRATIO, TAX, NOX have negative coefficients so we can say that increase in these values will decrease the AVG_PRICE of the house.

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	*Standard Err*	*t Stat*	*P-value*
NOX	-10.27270508	3.890849222	-2.640221837	0.008545718
PTRATIO	-1.071702473	0.133453529	-8.030529271	7.08251E-15
LSTAT	-0.605159282	0.0529801	-11.42238841	5.41844E-27
TAX	-0.014452345	0.003901877	-3.703946406	0.000236072
AGE	0.03293496	0.013087055	2.516605952	0.012162875
INDUS	0.130710007	0.063077823	2.072202264	0.038761669
DISTANCE	0.261506423	0.067901841	3.851242024	0.000132887
AVG_ROOM	4.125468959	0.44248544	9.323400461	3.68969E-19
Intercept	29.42847349	4.804728624	6.124898157	1.84597E-09

From the previous model considering that all the significant variables that contribute for the effective 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?

Comparing Regression Statistics for current model with previous we can infer that Adjusted R-square value is nearly same for both model with value 0.688.

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	-10.27270508
PTRATIO	-1.071702473
LSTAT	-0.605159282
TAX	-0.014452345
AGE	0.03293496
INDUS	0.130710007
DISTANCE	0.261506423
AVG_ROOM	4.125468959
Intercept	29.42847349

From the model we can say that if the value of NOX is more in a locality in this town the average price of house gradually decreases as increase in NOX.

d) Write the regression equation from this model.

The Regression Equation as follows:

Y= -10.2727050815093*X1 -1.07170247269449*X2 -0.605159282035405*X3 -0.0144523450364819*X4 +0.0329349604286301*X5 +0.130710006682182*X6 + 0.261506423001821*X7 + 4.12546895908474*X8 + C

Where Y=AVG_PRICE

X1=NOX

X2=PTRATIO

X3=LSTAT

X4=TAX

X5=AGE

X6=INDUS

X7=DISTANCE

X8=AVG_ROOM

C=Intercept value (29.4284734939458)

Result:

From this Project, Implementation of Exploratory Data Analysis helped us to understand the nature of different data-attributes and understood how to use various statistical/analytical tools in MS Excel like Summary statistics, Histogram, correlation table, Regression analysis.

----------****----------

(Pavankumar Mudigoudra)