Terro's Real Estate Agency

Reetesh V

reeteshvenki@gmail.com

Title:- Terro's Real Estate Agency

Situation :- We have a terro's real estate agency and they need to earn a high profit margin by understanding the correct market price of all the property.

Task :- To analyze and understand the key characteristics of the property to get a correct estimation of the properties market price.

Action :-

1. **The first step to any project is understanding the data. So for this step, generate the summary statistics for each of the variables. What do you observe ?**

CRIM E_RATE		*AGE*		*IND US*	
Mean	4.871976	Mean	68.5749	Mean	11.13678
Standard Error	0.12986	Standard Error	1.25137	Standard Error	0.30498
Median	4.82	Median	77.5	Median	9.69
Mode	3.43	Mode	100	Mode	18.1
Standard Deviation	2.921132	Standard Deviation	28.14886	Standard Deviation	6.860353
Sample Variance	8.533012	Sample Variance	792.3584	Sample Variance	47.06444
Kurtosis	-1.18912	Kurtosis	-0.96772	Kurtosis	-1.23354
Skewness	0.021728	Skewness	-0.59896	Skewness	0.295022
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

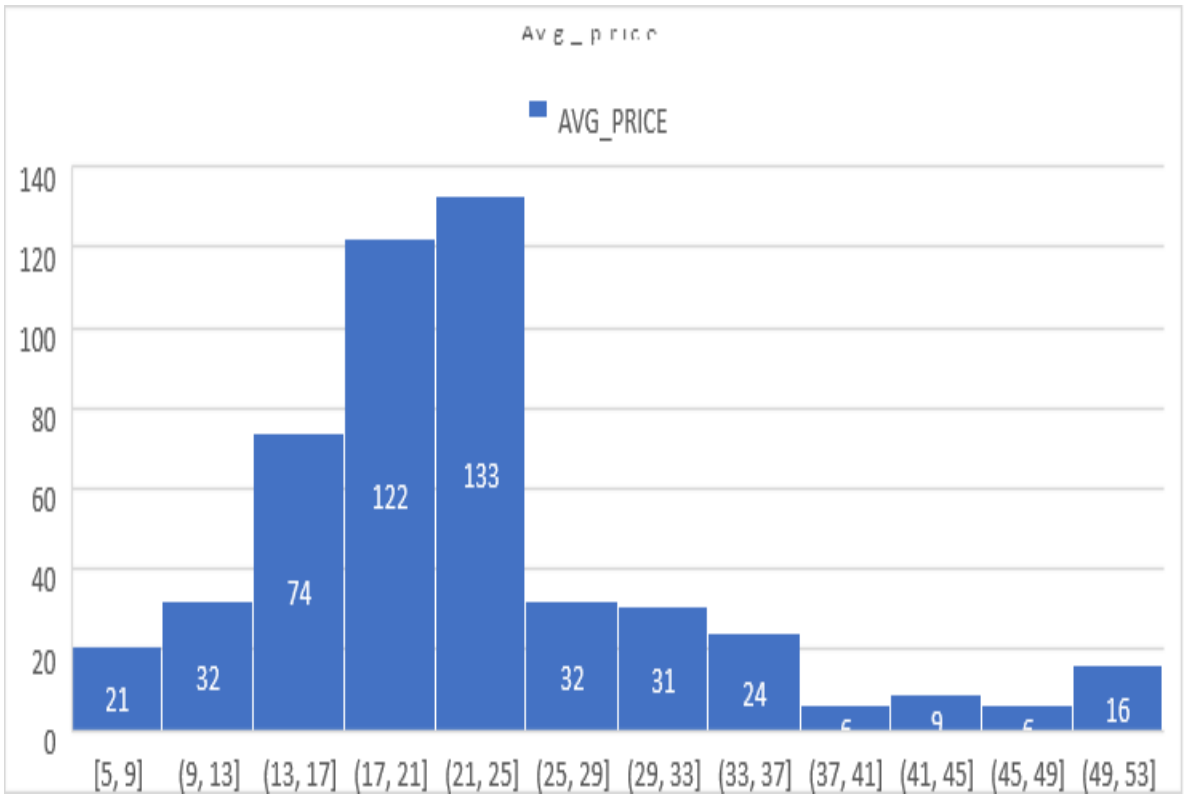
NOX		DISTANCE		TAX	
Mean	0.554695059	Mean	9.549407	Mean	408.2372
Standard Error	0.005151391	Standard Error	0.387085	Standard Error	7.492389
Median	0.538	Median	5	Median	330
Mode	0.538	Mode	24	Mode	666
Standard Deviation	0.115877676	Standard Deviation	8.707259	Standard Deviation	168.5371
Sample Variance	0.013427636	Sample Variance	75.81637	Sample Variance	28404.76
Kurtosis	-0.06466713	Kurtosis	-0.86723	Kurtosis	-1.14241
Skewness	0.729307923	Skewness	1.004815	Skewness	0.669956
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	18.45553	Mean	6.284634	Mean	12.65306	Mean	22.53281
Standard Error	0.096244	Standard Error	0.031235	Standard Error	0.317459	Standard Error	0.408861
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	2.164946	Standard Deviation	0.702617	Standard Deviation	7.141062	Standard Deviation	9.197104
Sample Variance	4.686989	Sample Variance	0.493671	Sample Variance	50.99476	Sample Variance	84.58672
Kurtosis	-0.28509	Kurtosis	1.8915	Kurtosis	0.49324	Kurtosis	1.495197
Skewness	-0.80232	Skewness	0.403612	Skewness	0.90646	Skewness	1.108098
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

- The AVG_PRICE is right skewed distribution were as PTRATIO Is in Left skewed position.

- The Kurtosis of AVG_PRICE is maximum and at the Peak, and the kurtosis of INDUS is Minimum.
- The Maximum Crime_Rate is 9.99 with having Average room 6.2846.

2. **Plot the histogram of the Avg_Price Variable. What do you infer?**



From the above histogram we can infer that the average price ranges high at 21 – 25 and 133 houses lies at this price range category, and the minimum price ranges from 45-49 and 6 houses lies at this price range category.

3. Compute the covariance matrix. Share your observation ?

	CRIM E_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIM E_RATE	8.516147									
AGE	87	790.79								
INDUS	0.562915	25	46.971							
NOX	22	124.26	43	0.0134						
DISTANCE	-0.110215	78	0.6058	01	75.666					
TAX	0.000625	2.3812	74	0.6157	53	28348.				
PTRATIO	31	12	35.479	0.0473	1333.17	62	4.6777			
AVG_ROOM	-0.229860	111.55	71	1	8.7434	167.82	26	0.492695		
LSTAT	49	2397.9	831.71	13.020	02	08	-0.5396	22	50.893	
AVG_PRICE	-8.229322	15.905	5.6808	0.0245	1.2812	34.515	9	-3.073655	98	84.41955
	44	42	33	5	8	1				616
	-0.068168	4.7425	1.8842	0.0245	1.2812	34.515	0.5396	0.492695		
	91	43	55	04	02	08	26	22		
	-0.056117	4.7425	1.8842	0.0245	1.2812	34.515	0.5396	-0.492695		
	78	4	3	5	8	1	9			
	-0.882680	120.83	29.521	0.4879	30.325	653.42	-5.7713	3.073655	50.893	
	36	84	81	8	39	06		98		
	1.162012	97.396	30.460	0.4545	30.500	-724.82	10.090	4.484565	48.351	84.41955
	24	2	5	1	8		7	55	8	616

- The Tax and Distance are Positively related with each other.
- The Tax and Avg_Price is Negatively related with each other.

4. **Create a correlation matrix of all the variables as shown in the Videos and various case studies. State top 3 positively correlated pairs and top 3 negatively correlated pairs.**

	CRIM E_R ATE	*AGE*	*INDUS*	*NOX*	*DISTAN CE*	*TAX*	*PTRATI O*	*AVG_RO OM*	*LSTAT*	*AVG_PRI CE*
CRIM E_R ATE	1									
AGE	0.006859463	1								
INDUS	-0.005510651	0.64479	1							
NOX	0.001850982	0.73147	0.76351	1						
DISTANCE	-0.009055049	0.45622	0.59529	0.61141	1					
TAX	-0.016748522	0.50656	0.72076	0.66802	0.91028	1				
PTRATIO	0.010800586	0.26115	0.38332	0.18893	0.46473	0.46085	1			
AVG_ROOM	-0.027396 16	-0.2402 6	-0.3916 8	-0.3021 9	-0.2098 5	-0.2920 5	0.3555	1		
LSTAT	0.042398321	0.6023 39	0.6038 79	0.5908 79	0.4886 76	0.5439 93	-0.3740 44	-0.613808 3	1	
AVG_PRICE	-0.043337871	0.3769 5	0.4837 3	0.4273 2	0.3816 3	0.4685 4	-0.5077 9	0.695359 95	-0.737 66	1

- **Top 3 positively correlated pairs –**
-
    1. Distance vs Tax
    2. Indus vs NOX
    3. Age vs NOX

- **Top 3 negatively correlated pairs –**

    1. LStat vs Avg_Price

    2. Avg_Room vs Lstat

    3. PtRatio vs Avg_Price.

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

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

ANOVA

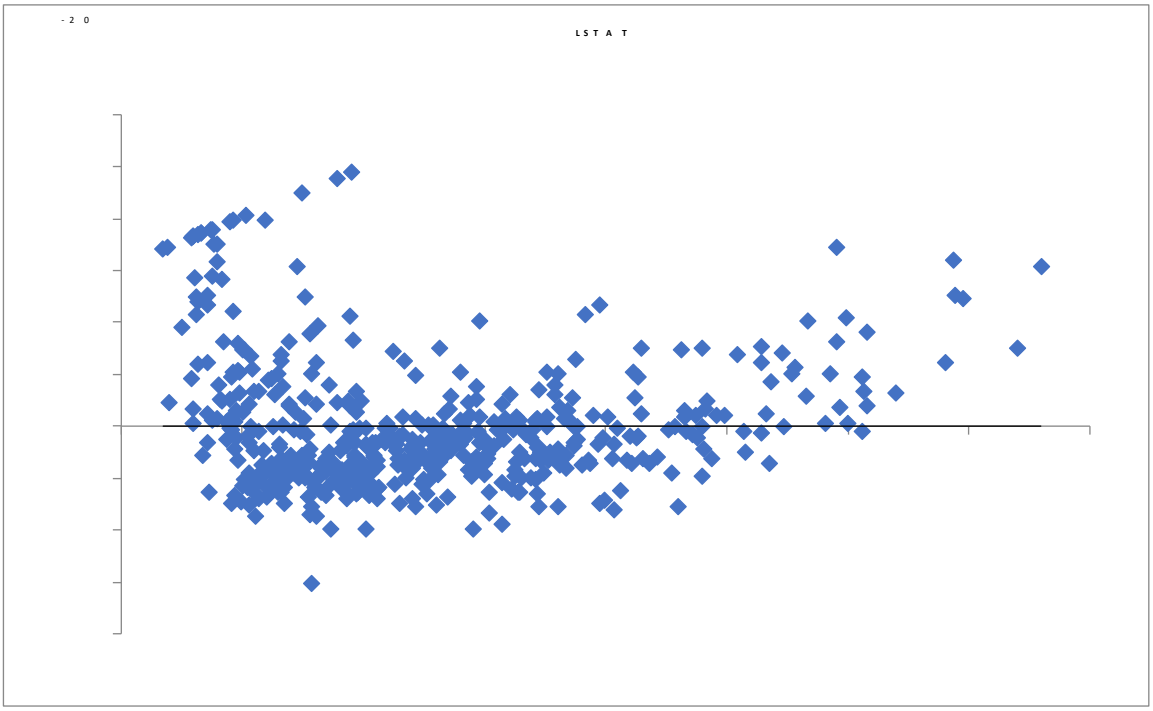
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	23243.914	23243.91	601.6	5.08E-81
Residual	504	19472.38142	38.63568		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	34.55384088	0.562627355	61.41515	3.74E-236
LSTAT	-0.950049354	0.038733416	-24.5279	5.081E-88

Lower 95%	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
33.448457	35.659225	33.44845704	35.65922471
-1.0261482	-0.873951	-1.0261482	-0.8739505

a.) What do you infer from the Regression Summary Output in terms of variance explained, coefficient value, Intercept and the Residual plot?

- The coefficient value is strong negative relationship between x and y.
- The P-Value is less than 0.05 and hence it is a good model.
- Here 54.4% of the variation in y is explained by x, 45.6% of the model is unexplained.
- Standard error is 6.21 and it is 6.21% far from the regression line.



- The Linear model is not appropriate, most of the values lies below the average line

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

- Yes, the LSTAT Variable is significant and it is having the significance F value less than 0.05.

6. Build another instance of the Regression model but this time including LSTAT and AVG_ROOM together as Independent variables and AVG_PRICE as the dependent variable.

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

ANOVA

	df	*SS*	*MS*	*F*
Regression	2	27276.99	13638.49	444.3309
Residual	503	15439.31	30.69445	
Total	505	42716.3		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-1.35827	3.172828	-0.4281	0.668765
AVG_ROOM	5.094788	0.444466	11.46273	3.47E-27
LSTAT	-0.64236	0.043731	-14.6887	6.67E-41

Lower 95%	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
-7.59190028	4.87535466	-7.591900282	4.87535466
4.221550044	5.96802553	4.221550436	5.96802553
-0.72827717	-0.5564395	-0.728277167	-0.5564395

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?

Yi - Dependent variable, Xi – Independent variable

Y = b0 + b1x1 +b2x2

Y = -1.35 + 5.09 * 7 + (-0.642 * 20)

Y = 21.45808 = $21000

- The company is overcharging for the Locality.
- The Avg_price cost around $21000

b.) Is the performance of this model better than the previous model you built in Question 5? Compare in terms of adjusted R-square. Explain.

- Yes, The performance of this model is better than the previous model.
- The adjusted R Square of the previous model is 54.32%.
- In current model the adjusted R Square is 63.71% this is greater than the previous model.
- Hence the performance of the current model is better than the previous.

7. Now, build a Regression model with all variables. AVG_PRICE shall be the Dependent Variable. Interpret the output in terms of adjusted R-square, coefficient and Intercept values, Significance of variables with respect to AVG_price. Explain.

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	3293.20672	124.9045	1.9327E-121
Residual	496	13077.43492	26.36657962		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	29.24131526	4.817125596	6.07028293	2.54E-09
CRIM E_RATE	0.048725141	0.078418647	0.62134637	0.534657
AGE	0.032770689	0.013097814	2.50199682	0.01267
INDUS	0.130551399	0.063117334	2.06839217	0.039121
NOX	-10.3211828	3.894036256	-2.65051102	0.008294
DISTANCE	0.261093575	0.067947067	3.84260258	0.000138
TAX	-0.01440119	0.003905158	-3.6877361	0.000251
PTRATIO	-1.07430535	0.133601722	-8.0411041	6.59E-15
AVG_ROOM	4.125409152	0.442758999	9.31750493	3.89E-19
LSTAT	-0.60348659	0.053081161	-11.369129	8.91E-27

- In this model about 68.82% of data can be explained.
- The coefficient and Intercept values are in positively related.

- LSTAT, AVG_ROOM, PTRATIO, TAX, DISTANCE, NOX, INDUS, AGE are significant Variables with respect to AVG_PRICE, and this values are less than 0.05 and the null hypothesis is rejected alternate hypothesis is accepted.
- CRIME_RATE is not significant with respect to AVG_PRICE, here Alternate hypothesis is rejected and null hypothesis is accepted.

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.

a.) Interpret the output of this model.

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	3703.585	140.643	1.911E-122
Residual	497	13087.61399	26.33323		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	29.42847349	4.804728624	6.124898	1.85E-09	19.98838959
AGE	0.03293496	0.013087055	2.516606	0.012163	0.007222187
INDUS	0.130710007	0.063077823	2.072202	0.038762	0.006777942
NOX	-10.27270508	3.890849222	-2.64022	0.008546	-17.9172457
DISTANCE	0.261506423	0.067901841	3.851242	0.000133	0.128096375
TAX	-0.014452345	0.003901877	-3.70395	0.000236	-0.022118553
PTRATIO	-1.071702473	0.133453529	-8.03053	7.08E-15	-1.333905109
AVG_ROOM	4.125468959	0.442485544	9.3234	3.69E-19	3.256096304
LSTAT	-0.605159282	0.0529801	-11.4224	5.42E-27	-0.70925186

Upper 95%	Lower 95.0%	Upper 95.0%
38.86855574	19.98838959	38.86855574
0.058647734	0.007222187	0.058647734
0.254642071	0.006777942	0.254642071
-2.628164466	-17.91724577	-2.628164466
0.394916471	0.128096375	0.394916471
-0.006786137	-0.022118553	-0.006786137

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 the adjusted R square with the previous model the previous model 68.82% can be explained.
- In the present model the Adjusted R square value is 68.86% can be explained.
- Therefore the adjusted R square value is greater in the present model and the present model performs better.

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?

Column1	Column2
NOX	-10.272705
PTRATIO	-1.0717025
LSTAT	-0.6051593
TAX	-0.0144523
AGE	0.032935
INDUS	0.13071
DISTANCE	0.2615064
AVG_ROOM	4.125469
Intercept	29.428473

- If NOX increases than the Avg_Price decreases and if NOX decreases Avg_Price increases

D.) Write the regression equation from this model.

$$Y = b0 + b1x1 + b2x2 + b3x3 - - - - - bnxn$$

Y = 29.4284 + (-10.272705 * NOX) + (-1.0717025 * PTRATIO) + (-0.6051593 * LSTAT) +

(- 0.0144523 * TAX) + (0.032935 * AGE)+ (0.13071 * INDUS)

+ (0.2615064 * DISTANCE) + (4.125469 * AVG_ROOM)

Thank you