EXPLORATORY DATA ANALYSIS PROJECT

Terro’s Real Estate Agency

PRASANNA YADAV

GLCA March 2023

Contents

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*		*NOX*	
Mean	4.871976	Mean	68.5749	Mean	11.13678	Mean	0.554695
Standard Error	0.12986	Standard Error	1.25137	Standard Error	0.30498	Standard Error	0.005151
Median	4.82	Median	77.5	Median	9.69	Median	0.538
Mode	3.43	Mode	100	Mode	18.1	Mode	0.538
Standard Deviatior	2.921132	Standard Deviation	28.14886	Standard Deviation	6.860353	Standard Deviation	0.115878
Sample Variance	8.533012	Sample Variance	792.3584	Sample Variance	47.06444	Sample Variance	0.013428
Kurtosis	-1.18912	Kurtosis	-0.96772	Kurtosis	-1.23354	Kurtosis	-0.06467
Skewness	0.021728	Skewness	-0.59896	Skewness	0.295022	Skewness	0.729308
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	9.549407	Mean	408.2372	Mean	18.45553	Mean	6.284634
Standard Error	0.387085	Standard Error	7.492389	Standard Error	0.096244	Standard Error	0.031235
Median	5	Median	330	Median	19.05	Median	6.2085
Mode	24	Mode	666	Mode	20.2	Mode	5.713
Standard Deviation	8.707259	Standard Deviation	168.5371	Standard Deviation	2.164946	Standard Deviation	0.702617
Sample Variance	75.81637	Sample Variance	28404.76	Sample Variance	4.686989	Sample Variance	0.493671
Kurtosis	-0.86723	Kurtosis	-1.14241	Kurtosis	-0.28509	Kurtosis	1.8915
Skewness	1.004815	Skewness	0.669956	Skewness	-0.80232	Skewness	0.403612
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	12.65306	Mean	22.53281
Standard Error	0.317459	Standard Error	0.408861
Median	11.36	Median	21.2
Mode	8.05	Mode	50
Standard Deviation	7.141062	Standard Deviation	9.197104
Sample Variance	50.99476	Sample Variance	84.58672
Kurtosis	0.49324	Kurtosis	1.495197
Skewness	0.90646	Skewness	1.108098
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

by using data analysis tool, we done the summary using descriptive statistics to each variable in given table.

1) CRIME_RATE:

- Maximum crime rate is around 9.99
- On an average crime rate is around 4.87.
- Standard deviation of 2.92 says that data deviates from mean by 4.87
- Negative Kurtosis signifies -1.18, It has flatter tails compared to a (Mesokurtic) normal distribution
- Skewness is nearly 0 which says curve follows normal distribution.
- 50% of crime rate lies below 4.82 and 50% lies above this value.
- the sum of crime rate in on locations was 2465.22.

2) AGE:

- Maximum age is around 100.
- On an average age in the given data is 68.574.
- Standard deviation of 28.148 says that data deviates from mean by 68.57
- Negative Kurtosis signifies -0.9677, It has flatter tails compared to a (Mesokurtic) normal distribution
- Skewness is nearly -0.598 which says curve goes on negative side.
- the sum of age of all locations was 34698.9

3) INDUS:

- Maximum no of industries are around 27.74.
- average of industries in all given locations was 11.13.
- Standard deviation of 6.86 says that data deviates from mean by 27.74
- Negative Kurtosis signifies -1.233, It has flatter tails compared to a (Mesokurtic) normal distribution
- Skewness is 0.29 which says curve skewed on positive side.

4) TAX:

- Maximum tax are around 711.
- average tax in all given locations was 408.23.
- Standard deviation of 168.53 says that data deviates from mean by 711
- Negative Kurtosis signifies -1.14, It has flatter tails compared to a (Mesokurtic) normal distribution
- Skewness is 0.66 which says curve skewed on positive side.

5) PTRATIO:

- Maximum of PTRATIO is around 22..
- average PTRATIO in all given locations was 18.455
- Standard deviation of 2.16 says that data deviates from mean by 22

- Negative Kurtosis signifies -0.28, It has flatter tails compared to a (Mesokurtic) normal distribution
- Skewness is -0.80 which says curve skewed on negative side.

6) AVG_ROOM:

- Maximum no of avg rooms are around 8.78
- average of avg_room in all given locations was 6.28
- Standard deviation of 0.702 says that data deviates from mean by 8.78
- Skewness is 0.403 which says curve skewed on positive side.

7) DISTANCE:

- On an average, the distance from highway around is 9.54 miles.
- Minimum houses distance from highway 1 mile.
- Maximum houses distance from high is 24 miles.

8) NOX:

- On an average, nitric oxides concentration is around 0.55 ppm
- Negative Kurtosis signifies -0.06, It has flatter tails compared to a (Mesokurtic) normal distribution.
- Skewness is positive 0.79 that indicates right side skewed.

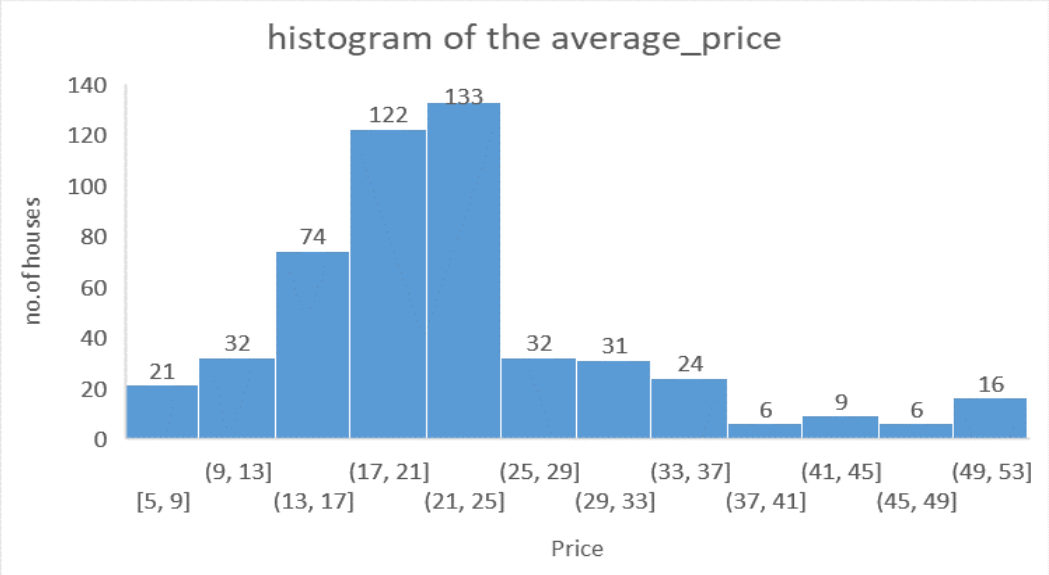
9) LSTAT:

- On an average,12% of population has lower status.
- Positive kurtosis gives us sharp curve.
- Positive skewness.

10) AVERAGE_PRICE:

- On an average price of house is $22,532
- Positive kurtosis gives us sharp curve.
- Positive skewness.

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



From the above histogram,we can get the information of average price of houses.

- In the above,label the x-axis as price and y-axis as number of houses.
- In that average price,the range of each interval is 4.
- The values starts from 5 and ends with 53 in ($1000's).
- Based on the histogram,most of the houses are between the interval of 21 to 25 i.e,133 and few of the houses are between the interval of 37 to 41 i.e,6 and 45 to 49 i.e,6.
- I concluded that the most of the houses are sold in the interval of (17,21) and (21,25).

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	790.7925								
INDUS	-0.110215175	124.2678	46.97143							
NOX	0.000625308	2.381212	0.605874	0.013401						
DISTANCE	-0.229860488	111.55	35.47971	0.61571	75.66653					
TAX	-8.229322439	2397.942	831.7133	13.0205	1333.117	28348.62				
PTRATIO	0.068168906	15.90543	5.680855	0.047304	8.743402	167.8208	4.677726			
AVG_ROOM	0.056117778	-4.74254	-1.88423	-0.02455	-1.28128	-34.5151	-0.53969	0.49269522		
LSTAT	-0.882680362	120.8384	29.52181	0.48798	30.32539	653.4206	5.7713	-3.07365497	50.893979	
AVG_PRICE	1.16201224	-97.3962	-30.4605	-0.45451	-30.5008	-724.82	-10.0907	4.48456555	-48.351792	84.41955616

Covariance: covariance is a measure of the relationship between two random variables. It describes how changes in one variable are associated with changes in another variable.

The covariance between two variables X and Y, denoted as Cov(X, Y), is calculated as the average of the products of the deviations of each variable from their respective means. Mathematically, the covariance formula is as follows:

$Cov(X, Y) = \Sigma[(X_i - \mu_x)(Y_i - \mu_y)] / (n - 1)$

- In the above covariance matrix we have values from negative to positive.
- Positive covariance ($Cov(X, Y) > 0$) indicates a direct or positive relationship, meaning that when one variable increases, the other tends to increase as well.
- Negative covariance ($Cov(X, Y) < 0$) indicates an inverse or negative relationship, meaning that when one variable increases, the other tends to decrease.
- In the above matrix TAX variable has positively strong covariance with all variables except for CRIME_RATE variable which has negative covariance.
- In above covariance matrix tax with age high variance in positive side.

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

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	VG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	1									
AGE	0.006859463	1								
INDUS	-0.005510651	0.644779	1							
NOX	0.001850982	0.73147	0.763651	1						
DISTANCE	-0.009055049	0.456022	0.595129	0.611441	1					
TAX	-0.016748522	0.506456	0.72076	0.668023	0.910228	1				
PTRATIO	0.010800586	0.261515	0.383248	0.188933	0.464741	0.460853	1			
AVG_ROOM	0.02739616	-0.24026	-0.39168	-0.30219	-0.20985	-0.29205	-0.3555	1		
LSTAT	-0.042398321	0.602339	0.6038	0.590879	0.488676	0.543993	0.374044	-0.61381	1	
AVG_PRICE	0.043337871	-0.37695	-0.48373	-0.42732	-0.38163	-0.46854	-0.50779	0.69536	-0.7376627	1

From 0 to +1: Positive correlation.

If 0: No correlation.

From 0 to -1: Negative correlation.

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

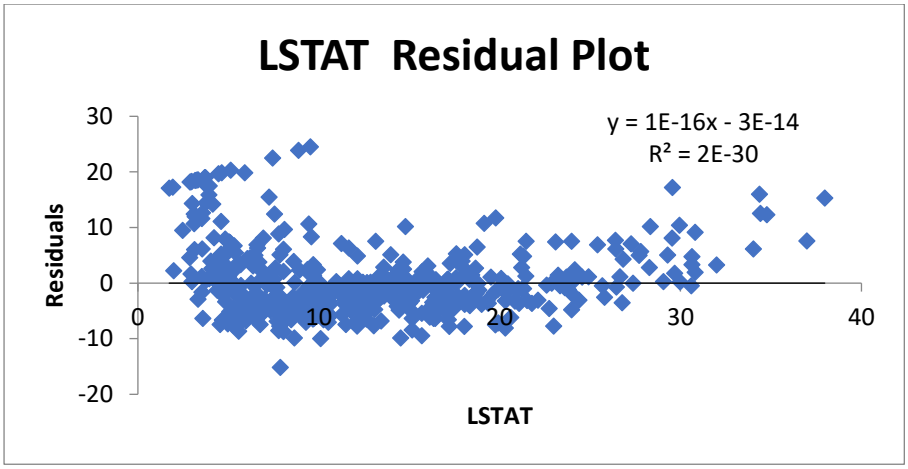
1.LSTAT – Avg_Room

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.

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	601.6179	5.0811E-88			
Residual	504	19472.38142	38.63568					
Total	505	42716.29542						
	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	33.448457	35.65922472
LSTAT	-0.950049354	0.038733416	-24.5279	5.08E-88	-1.0261482	-0.873950508	-1.0261482	-0.87395051



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 can be explained by the LSTAT.

Intercept of LSTAT for the model is 34.55384088.

The coefficient of LSTAT for the model is -0.950049354.

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.

SUMMARY OUTPUT								
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	-0.42809535	0.668764941	-7.591900282	4.875354658	-7.59190028	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	-14.6886992	6.66937E-41	-0.728277167	-0.556439501	-0.72827717	-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 formula:

$$Y = -1.358+5.09(AVG\_PRICE) + (-0.64)\ (LSTAT)$$

$$= -1.358+5.09(7) + (-0.64)\ (20)$$

$$= 21.472$$

The company is over charging,in this case as the value we got is less than charge that company is collecting.

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

The R-square value is 0.6385 as we can conclude that 63% of variability is explained by this model.

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

SUMMARY OUTPUT

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	3293.207	124.9045	1.9328E-121
Residual	496	13077.43492	26.3658		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.24131526	4.817125596	6.070283	2.54E-09	19.77682784	38.70580267	19.77682784	38.70580267
CRIME_RATE	0.048725141	0.078418647	0.621346	0.534657	-0.105348544	0.202798827	-0.10534854	0.202798827
AGE	0.032770689	0.013097814	2.501997	0.01267	0.00703665	0.058504728	0.00703665	0.058504728
INDUS	0.130551399	0.063117334	2.068392	0.039121	0.006541094	0.254561704	0.006541094	0.254561704
NOX	-10.3211828	3.894036256	-2.65051	0.008294	-17.97202279	-2.670342809	-17.9720228	-2.670342809
DISTANCE	0.261093575	0.067947067	3.842603	0.000138	0.127594012	0.394593138	0.127594012	0.394593138
TAX	-0.01440119	0.003905158	-3.68774	0.000251	-0.022073881	-0.0067285	-0.02207388	-0.0067285
PTRATIO	-1.074305348	0.133601722	-8.0411	6.59E-15	-1.336800438	-0.811810259	-1.33680044	-0.811810259
AVG_ROOM	4.125409152	0.442758999	9.317505	3.89E-19	3.255494742	4.995323561	3.255494742	4.995323561
LSTAT	-0.603486589	0.053081161	-11.3691	8.91E-27	-0.70777824	-0.499194938	-0.70777824	-0.499194938

From the above output,adjusted Rsquare is 0.69385372 and the intercept value is 29.24131526.

Where confidence level as 95%, then the p-value is 0.05.

Significant variables are those whose p-values are less than 0.05. If the p-value is greater than 0.05 then it is insignificant.

In the regression model,crime rate is less than p-value as highlighted in the above output.

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.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.832835773							
R Square	0.693615426							
Adjusted R Squar	0.688683682							
Standard Error	5.131591113							
Observations	506							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	8	29628.68142	3703.585178	140.643	1.911E-122			
Residual	497	13087.61399	26.33322735					
Total	505	42716.29542						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.42847349	4.804728624	6.124898157	1.85E-09	19.98838959	38.8685574	19.98838959	38.8685574
AGE	0.03293496	0.013087055	2.516605952	0.012163	0.007222187	0.058647734	0.007222187	0.058647734
INDUS	0.130710007	0.063077823	2.072202264	0.038762	0.006777942	0.254642071	0.006777942	0.254642071
NOX	-10.27270508	3.890849222	-2.640221837	0.008546	-17.9172457	-2.628164466	-17.9172457	-2.628164466
DISTANCE	0.261506423	0.067901841	3.851242024	0.000133	0.128096375	0.394916471	0.128096375	0.394916471
TAX	-0.014452345	0.003901877	-3.703946406	0.000236	-0.02211855	-0.006786137	-0.02211855	-0.006786137
PTRATIO	-1.071702473	0.133453529	-8.030529271	7.08E-15	-1.33390511	-0.809499836	-1.33390511	-0.809499836
AVG_ROOM	4.125468959	0.44248544	9.323400461	3.69E-19	3.256096304	4.994841615	3.256096304	4.994841615
LSTAT	-0.605159282	0.0529801	-11.42238841	5.42E-27	-0.70925186	-0.501066704	-0.70925186	-0.501066704

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?

By the value of adjusted R-square value little more compared to previous question 7. So, current model(only significant variable) performs better in these two models.

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

The average price will decrease if the value of NOX is more in a locality in that town.

D) Write the regression equation from this model.

Y(AVG_PRICE) = 29.42847349+0.03293496(65.2)+ 0.130710007(2.31) -10.27270508(0.538)+ 0.261506423(1) -0.014452345(296) -1.071702473(15.3)+ 4.125468959(6.575) - 0.605159282(4.98)

Y(AVG_PRICE) =21.4581.

Summary:

Terro's real-estate is an agency given a table of contents contain 10 different variables. In those average price is the targeted variable and all other variables are independent variables. According to my observation regarding whole table was average price of house was more in the interval of 21 to 25($1000's) most of them are willing to buy in medium price only. except crime rate remaining all the variables are significant to average price. And by increasing of NOX,LSTAT,PTRATIO there is change of decreasing of average price. So, u must need to decrease the tax then u get more number of people are buy for large price.

Thank You!!