EXCEL PROJECT

Terro's real estate agency

Statement:

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.

Data analysis:

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*		*DISTANCE*	
Mean	4.871976	Mean	68.5749	Mean	11.13678	Mean	0.554695	Mean	9.549407
Standard Erro	0.12986	Standard E	1.25137	Standard E	0.30498	Standard E	0.005151	Standard E	0.387085
Median	4.82	Median	77.5	Median	9.69	Median	0.538	Median	5
Mode	3.43	Mode	100	Mode	18.1	Mode	0.538	Mode	24
Standard Devi	2.921132	Standard D	28.14886	Standard D	6.860353	Standard D	0.115878	Standard D	8.707259
Sample Varian	8.533012	Sample Va	792.3584	Sample Va	47.06444	Sample Va	0.013428	Sample Va	75.81637
Kurtosis	-1.18912	Kurtosis	-0.96772	Kurtosis	-1.23354	Kurtosis	-0.06467	Kurtosis	-0.86723
Skewness	0.021728	Skewness	-0.59896	Skewness	0.295022	Skewness	0.729308	Skewness	1.004815
Range	9.95	Range	97.1	Range	27.28	Range	0.486	Range	23
Minimum	0.04	Minimum	2.9	Minimum	0.46	Minimum	0.385	Minimum	1
Maximum	9.99	Maximum	100	Maximum	27.74	Maximum	0.871	Maximum	24
Sum	2465.22	Sum	34698.9	Sum	5635.21	Sum	280.6757	Sum	4832
Count	506	Count	506	Count	506	Count	506	Count	506

TAX		PTRATIO	AVG_ROOM			LSTAT	AVG_PRICE		
Mean	408.2372	Mean	18.45553	Mean	6.284634	Mean	12.65306	Mean	22.53281
Standard E	7.492389	Standard E	0.096244	Standard E	0.031235	Standard E	0.317459	Standard E	0.408861
Median	330	Median	19.05	Median	6.2085	Median	11.36	Median	21.2
Mode	666	Mode	20.2	Mode	5.713	Mode	8.05	Mode	50
Standard D	168.5371	Standard D	2.164946	Standard D	0.702617	Standard D	7.141062	Standard D	9.197104
Sample Va	28404.76	Sample Va	4.686989	Sample Va	0.493671	Sample Va	50.99476	Sample Va	84.58672
Kurtosis	-1.14241	Kurtosis	-0.28509	Kurtosis	1.8915	Kurtosis	0.49324	Kurtosis	1.495197
Skewness	0.669956	Skewness	-0.80232	Skewness	0.403612	Skewness	0.90646	Skewness	1.108098
Range	524	Range	9.4	Range	5.219	Range	36.24	Range	45
Minimum	187	Minimum	12.6	Minimum	3.561	Minimum	1.73	Minimum	5
Maximum	711	Maximum	22	Maximum	8.78	Maximum	37.97	Maximum	50
Sum	206568	Sum	9338.5	Sum	3180.025	Sum	6402.45	Sum	11401.6
Count	506	Count	506	Count	506	Count	506	Count	506

Houses's number present in the dataset is 506 Average crime rate in the town is around 4.8

The house built in this locality is around 68 years old on average

Most of the houses have 27% of land for non-retail business

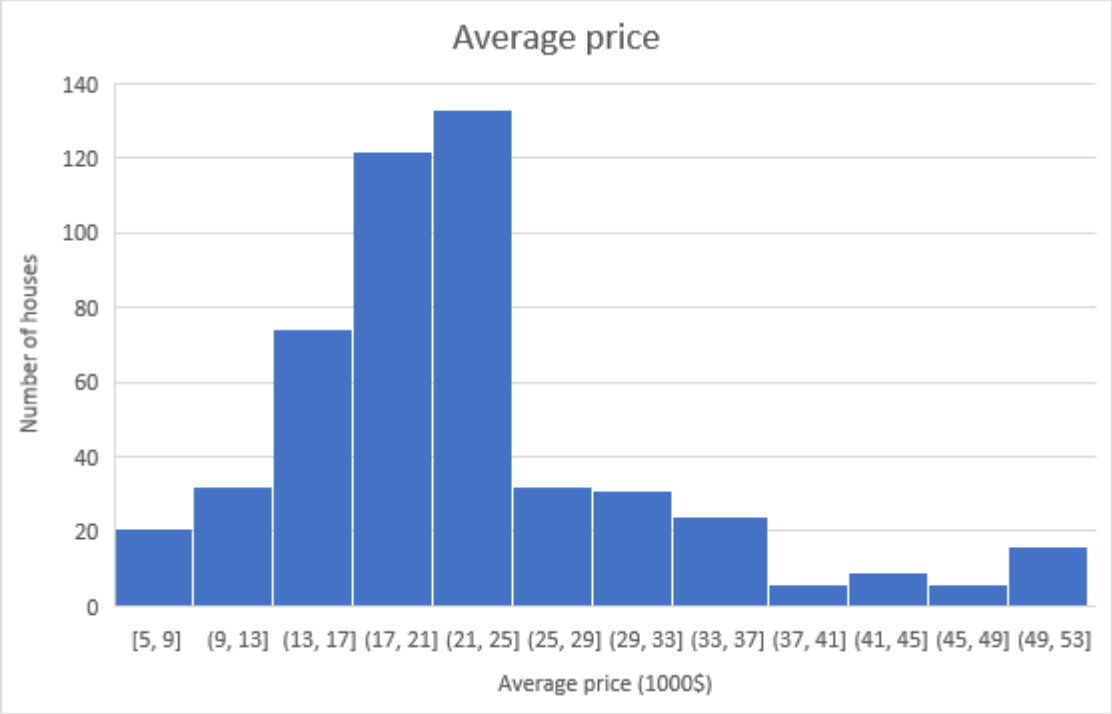
Most houses have 24 miles of distance from the highway

The average tax rate is around 408$

On average 12% of the population belongs to the lower status

Maximum number of rooms is 8 in a house.

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



In the Y-axis we have the number of houses and the x-axis shows the average Price of houses at 1000$ .from the graph , it can be inferred that it is positively Skewed data . Here the average house price are more concentrated between 21000USD and 25000USD.

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	0.492695		
LSTAT	-0.882680362	120.8384	29.52181	0.48798	30.32539	653.4206	5.7713	-3.07365	50.89398	
AVG_PRICE	1.16201224	-97.3962	-30.4605	-0.45451	-30.5008	-724.82	-10.0907	4.484566	-48.3518	84.41956

- The variance of the different variables is on the matrix diagonal. For example, the variance of crime rate is 8.5, Age is 790.7, etc.
- For average price and the rest of the independent variables, Positive covariance can be seen between average price and independent variables like crime rate and average room which implies that there is a positive relation between these variables that is, as the average number of rooms increases average price also increases
- As for variables like Age, Indus, Nox, Distance, Tax, Ptratio and Lstat have negative covariance with average price which means they are inversely proportional. Amongst this Tax variable has the highest negative covariance.
- With regards to independent variables, the given below variables have high positive covariance:
    1. Tax and Age
    2. Tax and Indus
    3. Tax and Distance

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	0.644779	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	-0.6138083	1	
AVG_PRICE	0.043337871	-0.37695	-0.48373	-0.42732	-0.38163	-0.46854	-0.50779	0.69535995	-0.73766	1

A) Which are the top 3 positively correlated pairs
B) Which are the top 3 negatively correlated pairs

a) 3 Positively correlated variables

- Tax and Distance - 0.91
- Nox and indus - 0.76
- Nox and Age – 0.73

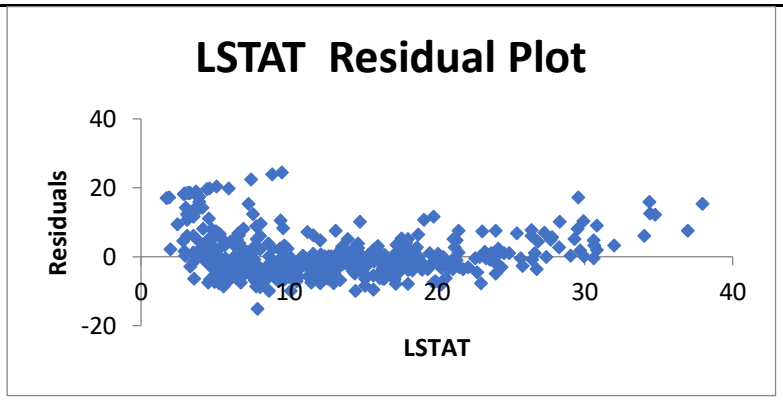
b) 3 Negatively correlated variables

- Average Price and Lstat -(-0.737)
- Lstat and Average room-(-0.61)
- Average price and Ptratio -(-0.503)

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	0.737663
R Square	0.544146
Adjusted R Square	0.543242
Standard Error	6.21576
Observations	506

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	34.55384	0.562627	61.41515	3.7E-236	33.44846	35.65922	33.44846	35.65922
LSTAT	-0.95005	0.038733	-24.5279	5.08E-88	-1.02615	-0.87395	-1.02615	-0.87395



Above given LSTAT residual plot from the regression table, it can be it can be inferred that the p-value of Lstat is significant relation between average price and Lstat.

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

- Average price=-0.95*Lstat + 34.55 (Regression equation)
- For a 1 unit increase in Lstat, there is a -0.95 dollars decrease in average price.

- The residual plot does not show patterns or trends so the model is pretty symmetrically distributed. If Lstat equals 0 then the expected value of the average price is 34.55.
- R-square is 0.54 which indicates that the Lstat in the regression model accounts for around 54% of the variation in the dependent variable. The remaining 46% is unexplained and perhaps attributable to other causes of variation.

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

Lstat is a significant variable for the model analysis as the p-value for the Lstat variable is less than 0.05. so this variable can be utilised for the model as it significantly impacts the average price

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	0.7991005							
R Square	0.6385616							
Adjusted R Squar	0.6371245							
Standard Error	5.5402574							
Observations	506							
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.3582728	3.17282778	-0.4281	0.66876	-7.5919003	4.87535466	-7.591900282	4.875354658
AVG_ROOM	5.094788	0.4444655	11.4627	3.5E-27	4.2215504	5.96802553	4.221550436	5.968025533
LSTAT	-0.6423583	0.043731465	-14.689	6.7E-41	-0.7282772	-0.5564395	-0.728277167	-0.5564395

Above given tables show the dependent variable is average price and independent variable are average room and Lstat.

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?

Average price = 5.09*Average room+ -o.642*Lstat-1.35

Now by substituting the values in the given question the value of the average price is 21.44, which is 21,440 USD which is less than the value quoted by the company Which is 30000 USD . from this , we can infer that they are overcharging for the Property.

Average price = 5.09*7+(-0.642*20)-1.35=21.44

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.

- This model is very easy understanding than previous model because previous model as r square has increase from 0.54 to 0.63 with the inclusion of the average room. By looking at the p-value of the average room it is less than 0.05 which states that it is a significant variable.

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	0.833							
R Square	0.6939							
Adjusted R Square	0.6883							
Standard Error	5.1348							
Observations	506							
	Coefficient	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.241	4.817125596	6.07028293	3E-09	19.7768278	38.7058	19.776828	38.7058027
CRIME_RATE	0.0487	0.078418647	0.62134637	0.5347	-0.10534854	0.202799	-0.105349	0.20279883
AGE	0.0328	0.013097814	2.50199682	0.0127	0.00703665	0.058505	0.0070367	0.05850473
INDUS	0.1306	0.063117334	2.06839217	0.0391	0.00654109	0.254562	0.0065411	0.2545617
NOX	-10.321	3.894036256	-2.6505102	0.0083	-17.9720228	-2.67034	-17.97202	-2.67034281
DISTANCE	0.2611	0.067947067	3.84260258	0.0001	0.12759401	0.394593	0.127594	0.39459314
TAX	-0.0144	0.003905158	-3.68773606	0.0003	-0.02207388	-0.00673	-0.022074	-0.0067285
PTRATIO	-1.0743	0.133601722	-8.04110406	7E-15	-1.33680044	-0.81181	-1.3368	-0.81181026
AVG_ROOM	4.1254	0.442758999	9.31750493	4E-19	3.25549474	4.995324	3.2554947	4.99532356
LSTAT	-0.6035	0.053081161	-11.3691294	9E-27	-0.70777824	-0.49919	-0.707778	-0.49919494

- As for the coefficients, Crime rate, Age, Indus, Distance and Average room are directly related to average price while Nox, Tax, Ptratio and Lstat are inversely related.
- The regression equation is
  Average price = 0.048*Crime rate + 0.03*Age + 0.13*Indus +(-10.32*Nox) + 0.26*Distance +(-0.014*Tax) + (-1.07*Ptratio) + 4.12*Average room + (-0.6*Lstat) + 29.241

- If any all the values of the explanatory variables are zero then the expected value of the average price is 29.241 dollars
- Variables which are highly significant for the model are Pratio, Average room and Lstat as its p-value are less than 0.05
- Variables like Tax, Distance, Nox, Indus and Age are moderately significant
- Crime rate is the only insignificant value which has a p value more than 0.05 which should be excluded from the model.

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	0.8328358							
R Square	0.6936154							
Adjusted R Square	0.6886837							
Standard Error	5.1315911							
Observations	506							
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.428473	4.804728624	6.1249	1.8E-09	19.98839	38.868557	19.9883896	38.8685574
AGE	0.032935	0.013087055	2.51661	0.01216	0.0072222	0.0586477	0.00722219	0.05864773
INDUS	0.13071	0.063077823	2.0722	0.03876	0.0067779	0.2546421	0.00677794	0.25464207
NOX	-10.272705	3.890849222	-2.6402	0.00855	-17.91725	-2.628164	-17.9172457	-2.6281645
DISTANCE	0.2615064	0.067901841	3.85124	0.00013	0.1280964	0.3949165	0.12809638	0.39491647
TAX	-0.0144523	0.003901877	-3.7039	0.00024	-0.022119	-0.006786	-0.02211855	-0.0067861
PTRATIO	-1.0717025	0.133453529	-8.0305	7.1E-15	-1.333905	-0.8095	-1.33390511	-0.8094998
AVG_ROOM	4.125469	0.44248544	9.3234	3.7E-19	3.2560963	4.9948416	3.2560963	4.99484161
LSTAT	-0.6051593	0.0529801	-11.422	5.4E-27	-0.709252	-0.501067	-0.70925186	-0.5010667

a) Interpret the output of this model.

- 68% of the variation average variable can be attributed to the variation in other independent variables.

- The P-value of all independent variables is significant as all of them are below 0.05
- By looking at the coefficients, when variables like Age, Indus, Distance, and Average room increase by 1 unit then the average price also increases
- When other variables like Nox, Tax, Ptratio, and Lstat increase by 1 unit then the average price decreases.

a) 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?
   - If we compare the Adjusted r square, this model is performing much better than the previous model. In the previous model, the adjusted r square was 0.6883 and it has improved to 0.6886.

b) 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?
   - If Nox is more in the region, then the average price decreases by -10.27 so we can say that they are inversely related.

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?
   - If Nox is more in the region, then the average price decreases by -10.27 so we can say that they are inversely related.

d) Write the regression equation from this model

The regression equation is as follows:

- Average price = 0.03*Age + 0.13*Indus + (-10.27*Nox) + 0.26*Distance + (-0.01*Tax)+ (-1.07*Ptratio) + 4.12*Average room + (-0.605*Lstat)