CASE: Terro's Real Estate Agency

(TOPICS COVERED: Descriptive Statistics, Covariance, Correlations, Simple Linear Regression, Multiple Linear Regression)

You have been hired at a Terro's Real Estate Agency in the capacity of an Auditor. One of the jobs that the auditors of this agency do is to map all the relevant features for the properties along with the information related to the geography around it. The agency wants to understand the relevance of the parameters that they collect in relation to the value of the house (MEDV).

You have been given a dataset of 506 houses of Boston. Please refer to the data dictionary below:

Data Dictionary:

- **CRIME_RATE**: per capita crime rate by town
- **INDUSTRY**: proportion of non-retail business acres per town (in percentage terms)
- **NOX**: nitric oxides concentration (parts per 10 million)
- **AVG_ROOM**: average number of rooms per house
- **AGE**: proportion of houses built prior to 1940 (in percentage terms)
- **DISTANCE**: distance from highway (in miles)
- **TAX**: full-value property-tax rate per $10,000
- **PTRATIO**: pupil-teacher ratio by town
- **LSTAT**: % lower status of the population
- **AVG_PRICE**: Average value of houses in $1000's

Your key job is to analyse the extent and magnitude of each variable relative to the value of the house. For this, you have the following deliverables to execute.

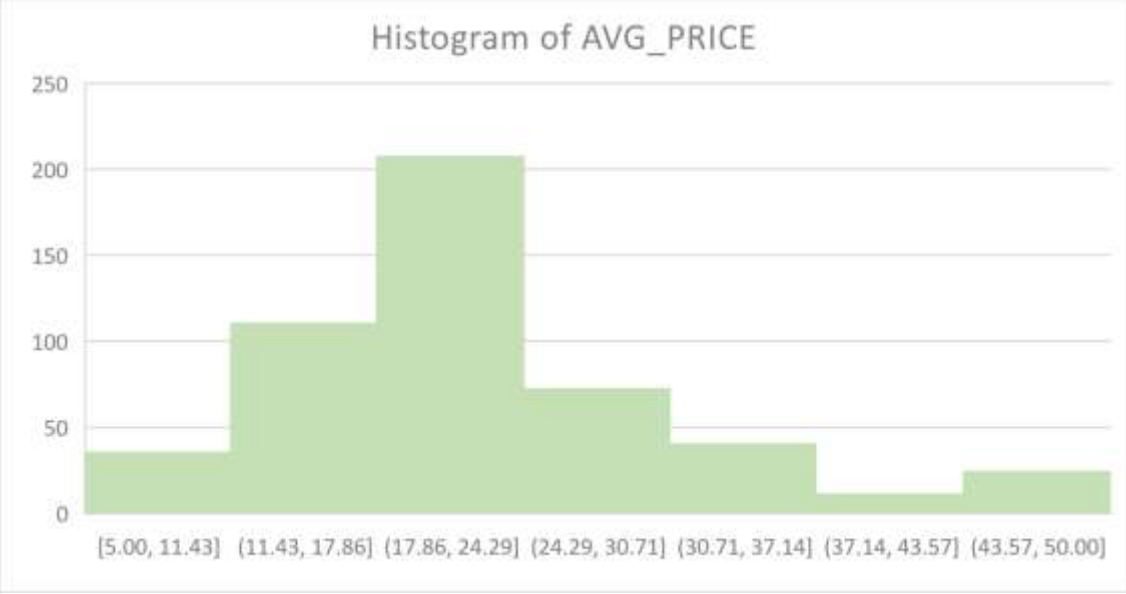
1. **The first step to any project is understanding the data. So for this step, generate the summary statistics for each of the variables. What do you observe? (5 marks)**

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
Mean	4.871976	68.5749	11.13678	0.554695	9.549407	408.2372	18.45553	6.284634	12.65306	22.53281
Standard Error	0.12986	1.25137	0.30498	0.005151	0.387085	7.492389	0.096244	0.031235	0.317459	0.408861
Median	4.82	77.5	9.69	0.538	5	330	19.05	6.2085	11.36	21.2
Mode	3.43	100	18.1	0.538	24	666	20.2	5.713	8.05	50
Standard Deviation	2.921132	28.14886	6.860353	0.115878	8.707259	168.5371	2.164946	0.702617	7.141062	9.197104

Sample Variance	8.533012	792.3584	47.06444	0.013428	75.81637	28404.76	4.686989	0.493671	50.99476	84.58672
Kurtosis	-1.18912	-0.96772	-1.23354	-0.06467	-0.86723	-1.14241	-0.28509	1.8915	0.49324	1.495197
Skewness	0.021728	-0.59896	0.295022	0.729308	1.004815	0.669956	-0.80232	0.403612	0.90646	1.108098
Range	9.95	97.1	27.28	0.486	23	524	9.4	5.219	36.24	45
Minimum	0.04	2.9	0.46	0.385	1	187	12.6	3.561	1.73	5
Maximum	9.99	100	27.74	0.871	24	711	22	8.78	37.97	50
Sum	2465.22	34698.9	5635.21	280.6757	4832	206568	9338.5	3180.025	6402.45	11401.6
Count	506	506	506	506	506	506	506	506	506	506

The average value of owner occupied houses (AVG_Price) in the data set is 22.53 ('000 USD), and the median value is 21.20. The range of values is between 5 to 50. The data has a bit of skewness. The average AGE is 68.5 and the median age is 77.5 suggesting negative skewness in this variable, it means that there are extreme values towards the lower end of the spectrum. Mean of AVG_Rooms is 6.28 and median is 6.2 suggesting that this variable could be normally distributed (more analysis would be required to know the exact picture). Most frequent value of AVG_ROOMS is 5.7.

2. **Plot the histogram of the Avg_Price Variable. What do you infer? (5 marks)**



The histogram shows that the variable, AVG_PRICE Is positively skewed as there is a tail towards the right, indicating the presence of some very high values in the data. Most of the houses are priced in the 17.88 to 24.49 bracket (prices are in '000 USD).

3. **Compute the covariance matrix. Share your observations. (5 marks)**

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	8.516148									
AGE	0.562915	790.7925								
INDUS	-0.11022	124.2678	46.97143							
NOX	0.000625	2.381212	0.605874	0.013401						
DISTANCE	-0.22986	111.55	35.47971	0.61571	75.66653					
TAX	-8.22932	2397.942	831.7133	13.0205	1333.117	28348.62				
PTRATIO	0.068169	15.90543	5.680855	0.047304	8.743402	167.8208	4.677726			
AVG_ROOM	0.056118	-4.74254	-1.88423	-0.02455	-1.28128	-34.5151	-0.53969	0.492695		
LSTAT	-0.88268	120.8384	29.52181	0.48798	30.32539	653.4206	5.7713	-3.07365	50.89398	
AVG_PRICE	1.162012	-97.3962	-30.4605	-0.45451	-30.5008	-724.82	-10.0907	4.484566	-48.3518	84.41956

CRIM_RATE and AVG_PRICE, AVG_ROOM and AVG_PRICE are positively related, rest all variables are negatively related with AVG_PRICE.

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. (5 marks)**

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	1									
AGE	0.006859	1								
INDUS	-0.00551	0.644779	1							
NOX	0.001851	0.73147	0.763651	1						
DISTANCE	-0.00906	0.456022	0.595129	0.611441	1					
TAX	-0.01675	0.506456	0.72076	0.668023	0.910228	1				
PTRATIO	0.010801	0.261515	0.383248	0.188933	0.464741	0.460853	1			
AVG_ROOM	0.027396	-0.24026	-0.39168	-0.30219	-0.20985	-0.29205	-0.3555	1		
LSTAT	-0.0424	0.602339	0.6038	0.590879	0.488676	0.543993	0.374044	-0.61381	1	
AVG_PRICE	0.043338	-0.37695	-0.48373	-0.42732	-0.38163	-0.46854	-0.50779	0.69536	-0.73766	1

The red cells suggest high positive correlations. Top 3 positively correlated pairs – TAX and Distance (0.891), NOX and INDUS (0.76) and NOX and AGE (0.73). Top 3 negative correlations – LSTAT and AVG_Price (-0.74), LSTAT and AVG_ROOM (-0.61) and PTRATIO and AVG_PRICE (-0.51).

5. **Build an initial regression model with AVG_PRICE as the y or the Dependent variable and LSTAT variable. Generate the residual plot too. (8 marks)**

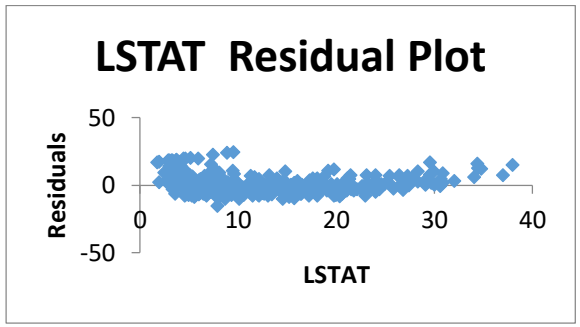
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.73766
R Square	0.54415
Adjusted R Square	0.54324
Standard Error	6.21576
Observations	506.00000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1.00000	23243.91400	23243.91400	601.61787	0.00000
Residual	504.00000	19472.38142	38.63568		
Total	505.00000	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	34.55384	0.56263	61.41515	0.00000	33.44846	35.65922	33.44846	35.65922
LSTAT	-0.95005	0.03873	-24.52790	0.00000	-1.02615	-0.87395	-1.02615	-0.87395



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

   The model has a R-squared value of 0.544 which suggests that this explains 54.4% of the variance in the AVG_PRICE. The intercept is 34.55 which suggests that even if LSTAT is 0, the value of AVG_PRICE will be positive, i.e. 34.55. Looking t the residual plot, we see more concentration of points towards the lower values of LSTAT, visually it suggests that there might be a pattern here, so we should explore more models we could get a better model.

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

   LSTAT has a significance value very close to 0, but it cannot be absolute 0. Since it is less than the significance level of 0.05, this variable LSTAT is significant and should be retained in our analysis.

6. **Build another instance of the Regression model but this time include LSTAT and AVG_ROOM variable together viz a viz AVG_PRICE as the dependent variable. (6 marks)**

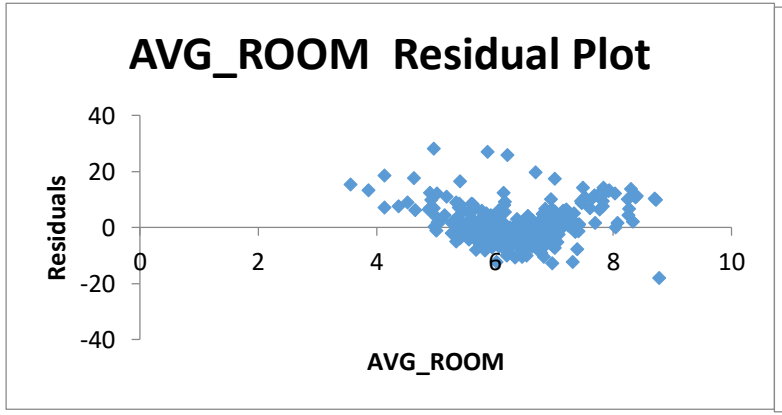
SUMMARY OUTPUT

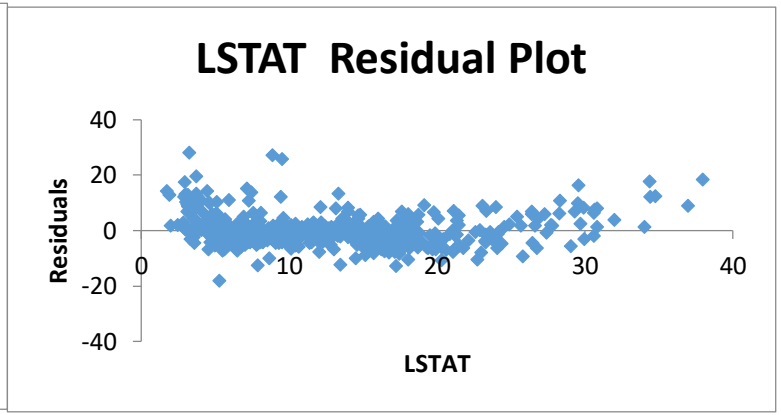
Regression Statistics	
Multiple R	0.80
R Square	0.64
Adjusted R Square	0.64
Standard Error	5.54
Observations	506.00

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2.00	27276.99	13638.49	444.33	0.00
Residual	503.00	15439.31	30.69		
Total	505.00	42716.30			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.36	3.17	-0.43	0.67	-7.59	4.88	-7.59	4.88
AVG_ROOM	5.09	0.44	11.46	0.00	4.22	5.97	4.22	5.97
LSTAT	-0.64	0.04	-14.69	0.00	-0.73	-0.56	-0.73	-0.56





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

$y = B_1X_1 + B_2X_2 + c$

$y = (5.09 * X_1) – (0.64 * X_2) - 1.36$

$X_1 = 7$ (AVG_ROOM)

$X_2 = 20$ (LSTAT)

Y = (5.09 * 7) – (0.64 * 20) – 1.36

Y = 35.63 – 12.8 – 1.36 = 21.47

The company is quoting a value of 30 against a prediction of 21.47, which suggests that the 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. Explain.**

   The R-squared value here is 0.64 as compared to 0.54 of the previous model, this means by adding AVG_ROOMS to our existing model, we are able to capture additional 10% of the variance in AVG_Price, because of which this is a better model than the previous one.

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. (8 marks)**

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8330
R Square	0.6939
Adjusted R Square	0.6883
Standard Error	5.1348
Observations	506.0000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	9.0000	29638.8605	3293.2067	124.9045	0.0000
Residual	496.0000	13077.4349	26.3658		

Total	505.0000	42716.2954

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.2413	4.8171	6.0703	0.0000	19.7768	38.7058	19.7768	38.7058
CRIME_RATE	0.0487	0.0784	0.6213	0.5347	-0.1053	0.2028	-0.1053	0.2028
AGE	0.0328	0.0131	2.5020	0.0127	0.0070	0.0585	0.0070	0.0585
INDUS	0.1306	0.0631	2.0684	0.0391	0.0065	0.2546	0.0065	0.2546
NOX	-10.3212	3.8940	-2.6505	0.0083	-17.9720	-2.6703	-17.9720	-2.6703
DISTANCE	0.2611	0.0679	3.8426	0.0001	0.1276	0.3946	0.1276	0.3946
TAX	-0.0144	0.0039	-3.6877	0.0003	-0.0221	-0.0067	-0.0221	-0.0067
PTRATIO	-1.0743	0.1336	-8.0411	0.0000	-1.3368	-0.8118	-1.3368	-0.8118
AVG_ROOM	4.1254	0.4428	9.3175	0.0000	3.2555	4.9953	3.2555	4.9953
LSTAT	-0.6035	0.0531	-11.3691	0.0000	-0.7078	-0.4992	-0.7078	-0.4992

This particular model has an R-squared value of 0.6939 against a R-squared value of 0.64 in the previous model (with LSTAT and AVG_ROOM), this model captures more variance as compared to the previous model. Also, here the adjusted R-square value is 0.6883 suggesting that the significant variables are contributing to 68.83% of the variance. The intercept value is 29.24, suggesting that even if all the independent variables were zero, the AVG_PRICE would be 29.24. Looking at the p-values, CRIM_RATE should be dropped as its p-value is more than 0.05. Rest all variables are significant.

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. (8 marks)**

   (HINT: 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)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8328
R Square	0.6936

Adjusted R Square	0.6887
Standard Error	5.1316
Observations	506.0000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8.0000	29628.6814	3703.5852	140.6430	0.0000
Residual	497.0000	13087.6140	26.3332		
Total	505.0000	42716.2954			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.4285	4.8047	6.1249	0.0000	19.9884	38.8686	19.9884	38.8686
AGE	0.0329	0.0131	2.5166	0.0122	0.0072	0.0586	0.0072	0.0586
INDUS	0.1307	0.0631	2.0722	0.0388	0.0068	0.2546	0.0068	0.2546
NOX	-10.2727	3.8908	-2.6402	0.0085	-17.9172	-2.6282	-17.9172	-2.6282
DISTANCE	0.2615	0.0679	3.8512	0.0001	0.1281	0.3949	0.1281	0.3949
TAX	-0.0145	0.0039	-3.7039	0.0002	-0.0221	-0.0068	-0.0221	-0.0068
PTRATIO	-1.0717	0.1335	-8.0305	0.0000	-1.3339	-0.8095	-1.3339	-0.8095
AVG_ROOM	4.1255	0.4425	9.3234	0.0000	3.2561	4.9948	3.2561	4.9948
LSTAT	-0.6052	0.0530	-11.4224	0.0000	-0.7093	-0.5011	-0.7093	-0.5011

Answer the questions below:

a. **Interpret the output of this model.**

This model explains 69.36% of the variance in AVG_PRICE. The intercept value is 29.42 suggesting that if all independent variables are 0, then the value of the house would be 29.42. All variables are significant here. This model is acceptable as it has a decent R-square and all variables are significant.

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?

Adjusted R-square for this model is 0.6887 vs 0.6883 in the previous model. Although adjusted R-square value is not up drastically, but we have all significant variables here, so is we consider these two factors together, then this model is a better model than the previous one.

c. Sort the values of the Coefficients in ascending order. What will happen to the average price if value of NOX is more in a locality in this town?

NOX	-10.2727
PTRATIO	-1.0717
LSTAT	-0.6052
TAX	-0.0145
AGE	0.0329
INDUS	0.1307
DISTANCE	0.2615
AVG_ROOM	4.1255

NOX and AVG_Price are negatively related. If the value of NOX increases then value of AVG_PRICE falls, more specifically every 1-unit increase in the value of NOX decreased the value of AVG_PRICE by 10.27.

d. Write the regression equation from this model.

Intercept (c)	29.4285
AGE	0.0329
INDUS	0.1307
NOX	-10.2727
DISTANCE	0.2615
TAX	-0.0145

PTRATIO	-1.0717
AVG_ROOM	4.1255
LSTAT	-0.6052

$$Y = 29.4285 + 0.0329 * X_1 + 0.1307 * X_2 + -10.2727 * X_3 + 0.2615 * X_4 – 0.0145 * X_5 – 1.0717 * X_6 + 4.1255 * X_7 – 0.6052 * X_8$$