PROJECT REPORT

TERRO'S REAL ESTATE AGENCY

By

~SUNILKUMAR A

DATA ANALYTICS JULY '22

Problem Statement (Situation):

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

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 SET GIVEN.

TASK TO REPORT

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

CRIME RATE	
Mean	4.871976
Standard Er	0.12986
Median	4.82
Mode	3.43
Standard De	2.921132
Sample Vari	8.533012
Kurtosis	-1.18912
Skewness	0.021728
Range	9.95
Minimum	0.04
Maximum	9.99
Sum	2465.22
Count	506

AGE	
Mean	68.5749
Standard E	1.25137
Median	77.5
Mode	100
Standard D	28.14886
Sample Var	792.3584
Kurtosis	-0.96772
Skewness	-0.59896
Range	97.1
Minimum	2.9
Maximum	100
Sum	34698.9
Count	506

INDUSTRY	
Mean	11.13678
Standard Err	0.30498
Median	9.69
Mode	18.1
Standard Dev	6.860353
Sample Varia	47.06444
Kurtosis	-1.23354
Skewness	0.295022
Range	27.28
Minimum	0.46
Maximum	27.74
Sum	5635.21
Count	506

NOX	
Mean	0.554695
Standard Err	0.005151
Median	0.538
Mode	0.538
Standard De	0.115878
Sample Varia	0.013428
Kurtosis	-0.06467
Skewness	0.729308
Range	0.486
Minimum	0.385
Maximum	0.871
Sum	280.6757
Count	506

DISTANCE	
Mean	9.549407
Standard Er	0.387085
Median	5
Mode	24
Standard De	8.707259
Sample Varia	75.81637
Kurtosis	-0.86723
Skewness	1.004815
Range	23
Minimum	1
Maximum	24
Sum	4832
Count	506

TAX	
Mean	408.2372
Standard Er	7.492389
Median	330
Mode	666
Standard De	168.5371
Sample Varia	28404.76
Kurtosis	-1.14241
Skewness	0.669956
Range	524
Minimum	187
Maximum	711
Sum	206568
Count	506

PTRATIO	
Mean	18.45553
Standard Er	0.096244
Median	19.05
Mode	20.2
Standard De	2.164946
Sample Varia	4.686989
Kurtosis	-0.28509
Skewness	-0.80232
Range	9.4
Minimum	12.6
Maximum	22
Sum	9338.5
Count	506

AVERAGE ROOT	
Mean	6.284634
Standard Er	0.031235
Median	6.2085
Mode	5.713
Standard De	0.702617
Sample Varia	0.493671
Kurtosis	1.8915
Skewness	0.403612
Range	5.219
Minimum	3.561
Maximum	8.78
Sum	3180.025
Count	506

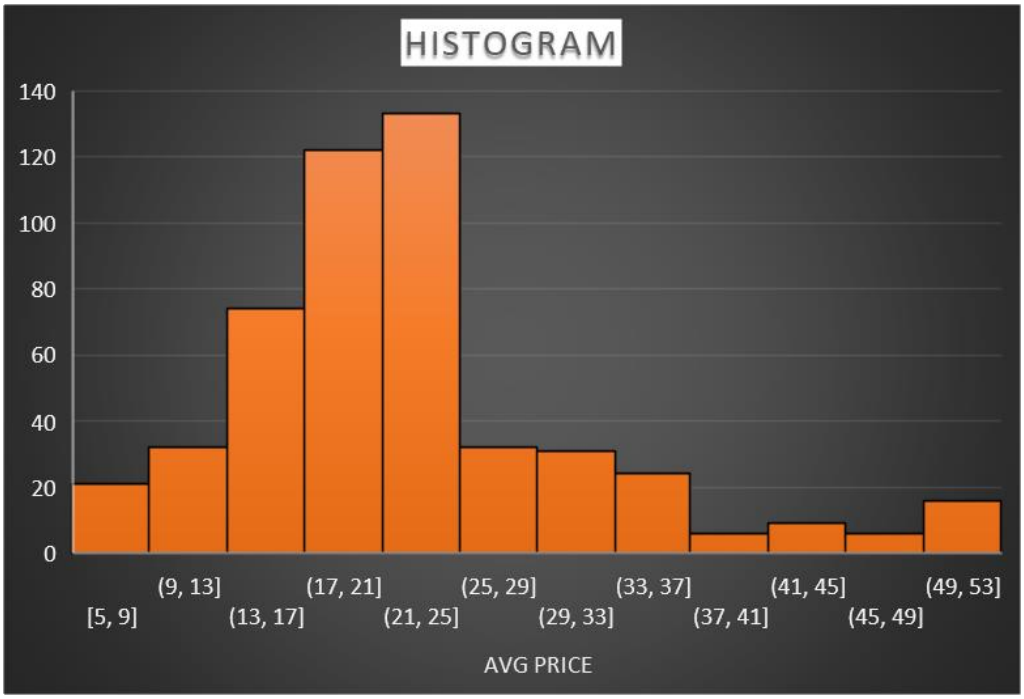
LSTAT	
Mean	12.65306
Standard Err	0.317459
Median	11.36
Mode	8.05
Standard Dev	7.141062
Sample Varia	50.99476
Kurtosis	0.49324
Skewness	0.90646
Range	36.24
Minimum	1.73
Maximum	37.97
Sum	6402.45
Count	506

AVG PRICE	
Mean	22.53281
Standard E	0.408861
Median	21.2
Mode	50
Standard D	9.197104
Sample Va	84.58672
Kurtosis	1.495197
Skewness	1.108098
Range	45
Minimum	5
Maximum	50
Sum	11401.6
Count	506

IMPRESSION:

- As per the data set for task number 1, By comparing the all variable we can say that Mean(Average) of TAX is higher then all other variables and NOX shows lesser Mean.
- Also we can say standard deviation of TAX is higher then all other variables.

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



- Here as per point of view in Histogram, The range of 21-25 shows highest peak range and 37-49 shows lowest peak range in Avg_price. The Avg price represented in ascending order.
- And the concept as Measure Of Peakedness(Kurtosis), Hence the Highest peak mentions as Leptokurtic.

3. Compute the covariance matrix. Share your observations.

	CRIME_RA	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROO	LSTAT	AVG_PRIC
CRIME_RA	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_ROO	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_PRIC	1.162012	-97.3962	-30.4605	-0.45451	-30.5008	-724.82	-10.0907	4.484566	-48.3518	84.41956

- According to the table we can see some positive covariance here, they are… (DISTANCE,TAX) (TAX,TAX)

- Negative covariance are (DISTANCE,CRIME RATE)

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

	CRIME_RA	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROO	LSTAT	AVG_PRIC
CRIME_RA	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_ROO	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_PRIC	0.043338	-0.37695	-0.48373	-0.42732	-0.38163	-0.46854	-0.50779	0.69536	-0.73766	1

IMPRESSION:

A)Top 3 positively correlated pairs		
0.910228	TAX AND DISTANCE	91%
0.763651	NOX AND INDUS	77%
0.73147	NOX AND AGE	73%

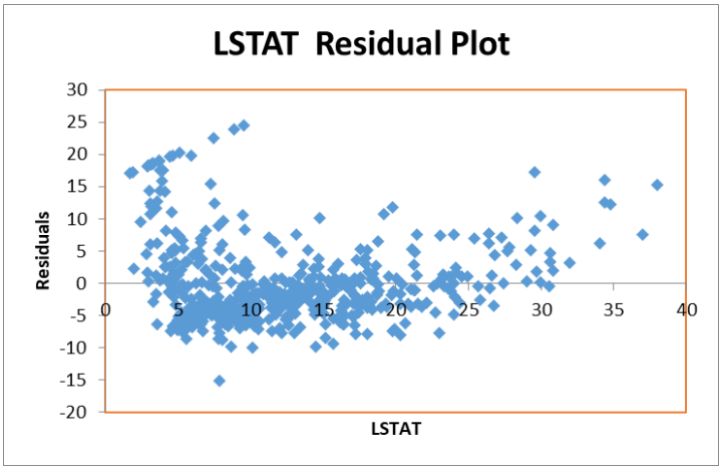
B)Top 3 negatively correlated pairs		
-0.73766273	AVG PRICE AND LSTA	(-74%)
-0.61380827	LSTAT AND AVG ROO	T(-61%)
-0.50778669	AVG PRICE AND PTRA	(-50%)

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 Residual plot?

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

Regression Statistics	
Multiple R	0.737663
R Square	0.544146
Adjusted F	0.543242
Standard E	6.21576
Observatic	506



	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	34.55384	0.562627	61.41515	3.7E-236	33.448457	35.65922	33.44846	35.65922
X Variable	-0.95005	0.038733	-24.5279	5.08E-88	-1.0261482	-0.87395	-1.02615	-0.87395

A)

- intercept 34.55384088
- Coefficient -0.950049354
- The graph looks as scattered in plot of residual.

B) LSTAT value is insignificant, cause the adjusted R-value is seems low.

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	0.7991
R Square	0.638562
Adjusted R	0.637124
Standard Er	5.540257
Observatior	506

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.35827	3.172828	-0.4281	0.668765	-7.5919	4.875355	-7.5919003	4.875354658
AVG ROOM	5.094788	0.444466	11.46273	3.47E-27	4.22155	5.968026	4.22155044	5.968025533
LSTAT	-0.64236	0.043731	-14.6887	6.67E-41	-0.72828	-0.55644	-0.7282772	-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?

- FORMULA:

Y= (AVG ROOM*7+(LSTAT*20)=INTERCEPT

=(5.0947*7)+(-0.64236*20)+(-1.3582)

=21.45808 OF PREDICTED AVG PRICE

How does it compare to the company quoting a value of 30000 USD for this locality? Is the company Overcharging/ Undercharging?

- The company is charging an exorbitant amount of $30,000 for services in this locality, which is clearly an overcharge. The predicted average price (AVG_PRICE) stands at a reasonable $21.45k, significantly lower than the company's asking price. By this company is Overchanging.

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

- | Adjusted R | 0.637124 | > | Adjusted R | 0.543242 |

Adjusted R value is giving better results then previous 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 R square, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE.

Regression Statistics	
Multiple R	0.832979
R Square	0.693854
Adjusted R Square	0.688299
Standard Error	5.134764
Observations	506

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	29.24132	4.817126	6.070283	2.54E-09	19.77682784	38.7058	19.77683	38.7058
CRIME_RATE	0.048725	0.078419	0.621346	0.534657	-0.10534854	0.202799	-0.10535	0.202799
AGE	0.032771	0.013098	2.501997	0.01267	0.00703665	0.058505	0.007037	0.058505
INDUS	0.130551	0.063117	2.068392	0.039121	0.006541094	0.254562	0.006541	0.254562
NOX	-10.3212	3.894036	-2.65051	0.008294	-17.9720228	-2.67034	-17.972	-2.67034
DISTANCE	0.261094	0.067947	3.842603	0.000138	0.127594012	0.394593	0.127594	0.394593
TAX	-0.0144	0.003905	-3.68774	0.000251	-0.02207388	-0.00673	-0.02207	-0.00673
PTRATIO	-1.07431	0.133602	-8.0411	6.59E-15	-1.33680044	-0.81181	-1.3368	-0.81181
AVG_ROOM	4.125409	0.442759	9.317505	3.89E-19	3.255494742	4.995324	3.255495	4.995324
LSTAT	-0.60349	0.053081	-11.3691	8.91E-27	-0.70777824	-0.49919	-0.70778	-0.49919

- **Adjusted R Square** 0.688298647, The impressive adjusted R-square value validates the suitability of this model for prediction tasks. Its ability to effectively account for the variance in the data indicates that it can be relied upon to make accurate predictions. As a result, this model is a strong candidate for practical use in various prediction scenarios.
- **Significant variables:** AGE, INDUS, NOX, DISTANCE, LSTAT, PTRATIO, AVGROOM, TAX
- **Insignificant variables:** CRIME RATE
- **Coefficient** of AVG ROOM is higher then all other variables.

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	0.832836
R Square	0.693615
Adjusted F	0.688684
Standard E	5.131591
Observatio	506

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	29.42847	4.804729	6.124898	1.85E-09	19.98839	38.86856	19.98839	38.86856
AGE	0.032935	0.013087	2.516606	0.012163	0.007222	0.058648	0.007222	0.058648
INDUS	0.13071	0.063078	2.072202	0.038762	0.006778	0.254642	0.006778	0.254642
NOX	-10.2727	3.890849	-2.64022	0.008546	-17.9172	-2.62816	-17.9172	-2.62816
DISTANCE	0.261506	0.067902	3.851242	0.000133	0.128096	0.394916	0.128096	0.394916
TAX	-0.01445	0.003902	-3.70395	0.000236	-0.02212	-0.00679	-0.02212	-0.00679
PTRATIO	-1.0717	0.133454	-8.03053	7.08E-15	-1.33391	-0.8095	-1.33391	-0.8095
AVG_ROO	4.125469	0.442485	9.3234	3.69E-19	3.256096	4.994842	3.256096	4.994842
LSTAT	-0.60516	0.05298	-11.4224	5.42E-27	-0.70925	-0.50107	-0.70925	-0.50107

a) Interpret the output of this model

The R-value is 68% so we can use to predictions.

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?

The Adjusted r value is better then previous model.

Adjusted R Sq	0.688684	>	Adjusted R Square	0.688299

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 in Ascending order,

-10.2727	NOX
-1.0717	PTRATIO
-0.60516	LSTAT
-0.01445	TAX
0.032935	AGE
0.13071	INDUS
0.261506	DISTANCE
4.125469	AVG_ROO

Intercept 29.42847349

The correlation between NOX and AVG_PRICE is -0.42732, indicating an inverse relationship between the two variables. When NOX levels increase, AVG_PRICE tends to decrease. In other words, higher NOX values are associated with lower average prices.

d) Write the regression equation from this model.

avg_price=(coeffecient(age)*age)+
(coefficient(indus)*indus)+(coeffecient(nox)*nox)+
(coeffcient(distance)*distance)+ (coeffecient(tax)*tax)+
(coeffecient(ptratio)*ptratio)+ (coeffecient(avg room)*avg_room)+
(coeffecient(lstat)*lstat)+ intercept