TERRO’S REAL ESTATE

Project submitted

By Madhusoodhan Ashok kumar

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

Crime Rate (CRIME_RATE):

- A lower crime rate is generally desirable for a residential area.
- The lower values of mean and median shows the crime rate is low .
- the range and standard deviation shows the variability in crime rates.

CRIME_RATE	
Mean	4,87197628
Standard Error	0,12986015
Median	4,82
Mode	3,43
Standard Deviation	2,92113189
Sample Variance	8,53301153
Kurtosis	-1,18912246
Skewness	0,02172808
Range	9,95
Minimum	0,04
Maximum	9,99
Sum	2465,22
Count	506

Age of the Population (AGE):

- The higher value of median age might indicate a more stable and settled community.

AGE	
Mean	68,57490119
Standard Error	1,251369525
Median	77,5
Mode	100
Standard Deviation	28,14886141
Sample Variance	792,3583985
Kurtosis	-0,967715594
Skewness	-0,59896264
Range	97,1
Minimum	2,9
Maximum	100
Sum	34698,9
Count	506

Nitric Oxide Concentration (NOX):

- Lower NOx levels are generally preferred for better air quality.
- The areas with lower mean and median values shows NOx concentration is better.

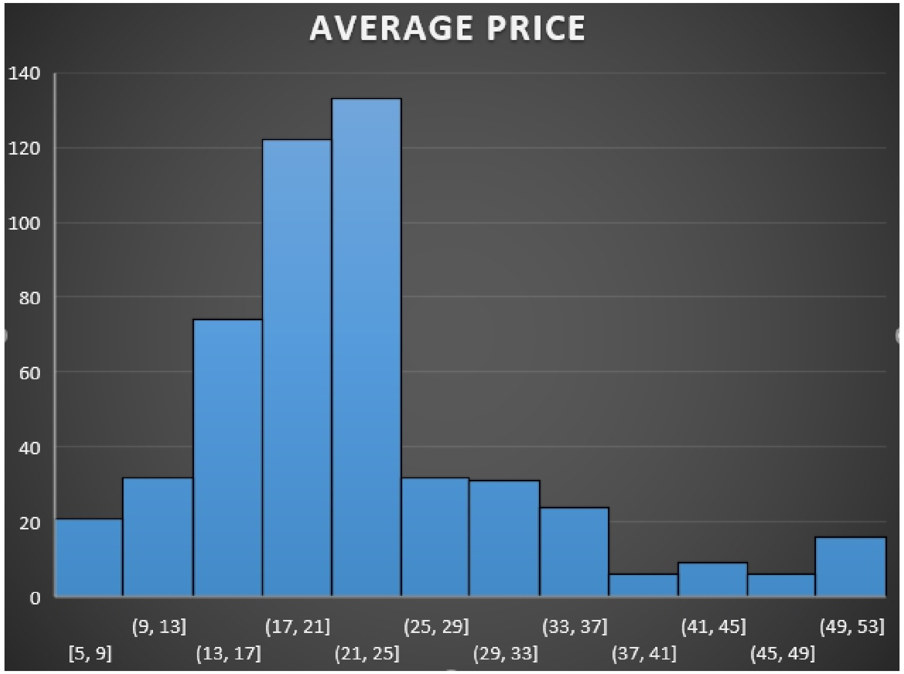
NOX	
Mean	0,55469506
Standard Error	0,00515139
Median	0,538
Mode	0,538
Standard Deviation	0,11587768
Sample Variance	0,01342764
Kurtosis	-0,06466713
Skewness	0,72930792
Range	0,486
Minimum	0,385
Maximum	0,871
Sum	280,6757
Count	506

Average Number of Rooms (AVG_ROOM):

- Considering the size of the houses in the area based on the average number of rooms.
- Due to positive skewness most of the houses should be less than 6 average rooms.

AVG_ROOM	
Mean	6,28463439
Standard Error	0,03123514
Median	6,2085
Mode	5,713
Standard Deviation	0,70261714
Sample Variance	0,49367085
Kurtosis	1,89150037
Skewness	0,40361213
Range	5,219
Minimum	3,561
Maximum	8,78
Sum	3180,025
Count	506

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



From the above Histogram chart we conclude that

- The highest tower shows the number of houses from a range of average price between 22 to 25 and in that range we have around 133 houses .
- The lowest tower shows the number of houses from a range of average price between 38 to 49 and we have less number of houses in that particular range.
- 10 to 13 and 26 to 29 range of average price have a same count of houses. Nearly 32 houses in that range.
- 38 to 41 and 46 to 49 range of average price also have a same count of houses . Nearly 6 houses in that range.

3) Compute the covariance matrix. Share your observations

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	8,51615									
AGE	0,56292	790,79247								
INDUS	-0,11022	124,26783	46,97143							
NOX	0,00063	2,38121	0,60587	0,01340						
DISTANCE	-0,22986	111,54996	35,47971	0,61571	75,66653					
TAX	-8,22932	2397,94172	831,71333	13,02050	1333,11674	28348,62360				
PTRATIO	0,06817	15,90543	5,68085	0,04730	8,74340	167,82082	4,67773			
AVG_ROOM	0,05612	-4,74254	-1,88423	-0,02455	-1,28128	-34,51510	-0,53969	0,49270		
LSTAT	-0,88268	120,83844	29,52181	0,48798	30,32539	653,42062	5,77130	-3,07365	50,89398	
AVG_PRICE	1,16201	-97,39615	-30,46050	-0,45451	-30,50083	-724,82043	-10,09068	4,48457	-48,35179	84,41956

Covariance:

- The off-diagonal elements represent the covariances between pairs of variables. For example:
- The covariance between Crime Rate (CRIME_RATE) and Age (AGE) is approximately 0.56.
- The covariance between Tax (TAX) and Industrial Areas Proportion (INDUS) is approximately 2397.94.

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	1,00000									
AGE	0,00686	1,00000								
INDUS	-0,00551	0,64478	1,00000							
NOX	0,00185	0,73147	0,76365	1,00000						
DISTANCE	-0,00906	0,45602	0,59513	0,61144	1,00000					
TAX	-0,01675	0,50646	0,72076	0,66802	0,91023	1,00000				
PTRATIO	0,01080	0,26152	0,38325	0,18893	0,46474	0,46085	1,00000			
AVG_ROOM	0,02740	-0,24026	-0,39168	-0,30219	-0,20985	-0,29205	-0,35550	1,00000		
LSTAT	-0,04240	0,60234	0,60380	0,59088	0,48868	0,54399	0,37404	-0,61381	1,00000	
AVG_PRICE	0,04334	-0,37695	-0,48373	-0,42732	-0,38163	-0,46854	-0,50779	0,69536	-0,73766	1,00000

A) Strong positive correlations (close to 1) are observed between:

- INDUS and TAX (0.72076)
- AGE and NOX (0.73147)
- AGE and INDUS (0.64478)

B) Strong negative correlations (close to -1) are observed between:

- AVG_ROOM and LSTAT (-0.61381)
- AVG_ROOM and AVG_PRICE (-0.73766)

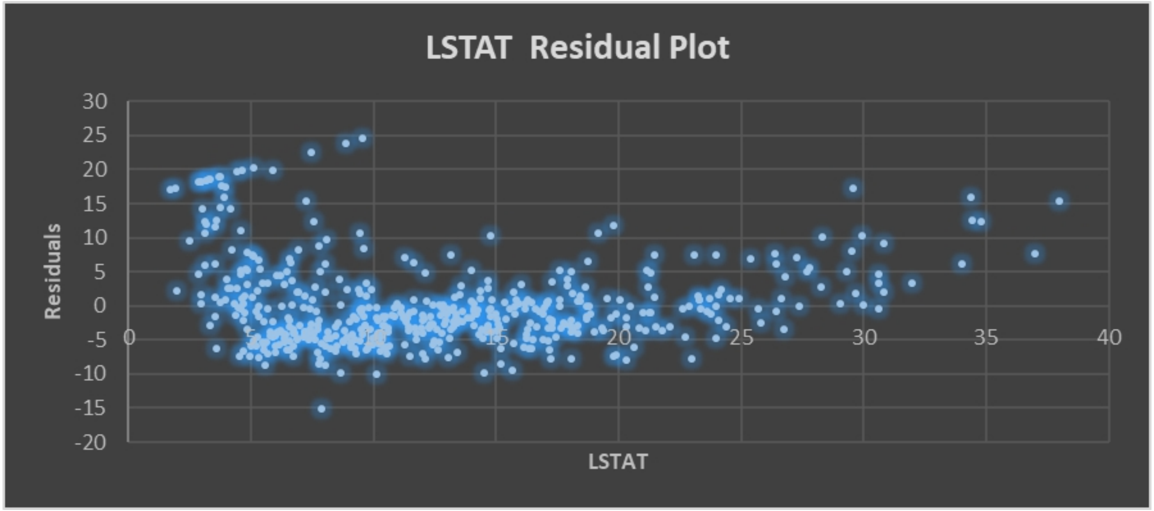
5) Build an initial regression model with AVG_PRICE as ‘y’ (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot.

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

Regression Statistics	
Multiple R	0,737662726
R Square	0,544146298
Adjusted R Square	0,543241826
Standard Error	6,215760405
Observations	506

R square value = 0.54
Adjusted R square value = 0.543

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	34,55384088	0,562627355	61,41514552	3,7431E-236	33,44845704	35,65922472
LSTAT	-0,950049354	0,038733416	-24,52789985	5,0811E-88	-1,0261482	-0,873950508



From the above chart we conclude that,

- After 25 and Below 5 the residual errors are to be Upper biased
- Between 5 to 25 the residual errors to be Biased and near to the linear line

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

	Coefficients	***Standard Error***	***t Stat***	***P-value***	***Lower 95%***	***Upper 95%***
Intercept	34,55384088	0,562627355	61,41514552	3,7431E-236	33,44845704	35,65922472
LSTAT	-0,950049354	0,038733416	-24,52789985	5,0811E-88	-1,0261482	-0,873950508

LSTAT is 5.0811E-88 which is less than the P value 0.05

6) Build a new Regression model including LSTAT and AVG_ROOM together as Independent variables and AVG_PRICE as dependent variable.

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?

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-1,358272812	3,17282778	-0,428095348	0,668764941	-7,591900282	4,875354658
AVG_ROOM	5,094787984	0,4444655	11,46272991	3,47226E-27	4,221550436	5,968025533
LSTAT	-0,642358334	0,043731465	-14,68869925	6,66937E-41	-0,728277167	-0,556439501

Company Quoting value = 30000 USD

Regression equation Y = m1x1+ m2x2 +b

Slope of Average room = m1 = 5.094787984

Slope of LSTAT= m2 = -0.642358334

x1 = 7

x2 = 20

Intercept = b = -1.358272812

Regression equation Y = (5.094787984)*7 + (-0.642358334)*20 + (-1.358272812)

= 21.4580764

21.4580764 is like 21,000 USD so its below to the company quoting value 30,000 USD. Hence, the Company is over charging.

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.

Regression statistics for previous model

Regression Statistics	
Multiple R	0,737662726
R Square	0,544146298
Adjusted R Square	0,543241826
Standard Error	6,215760405
Observations	506

Regression statistics for this model

Regression Statistics	
Multiple R	0,799100498
R Square	0,638561606
Adjusted R Square	0,637124475
Standard Error	5,540257367
Observations	506

Adjusted R-square of Previous model = 0.543241826

Adjusted R-square of This model = 0.637124475

This model has the highest value of adjusted R-square. So, this model is better than the previous model

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

Since the Adjusted R Square value is 0.688299 it is above 50%. So, we conclude that this Regression model to be a good one.

	Coefficients
Intercept	29.24132
Crime rate	0.048725
Age	0.032771
Indus	0.130551
Nox	-10.3212
Distance	0.261094
Tax	-0.0144
Ptratio	-1.07431
Avg room	4.125409
Lstat	-0.60349

From the above table,

The coefficient values of Crime rate, Age, Indus, Distance, and Avg room is positive. So these variables are Directly proportional to Average price.

The coefficient values of Nox ,Tax, Ptratio, Lstat is Negative. So these variables are Inversely proportional to Average price.

The Intercept value is positive.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	29.24132	4.817126	6.070283	2.54E-09
Crime rate	0.048725	0.078419	0.621346	0.534657
Age	0.032771	0.013098	2.501997	0.01267
Indus	0.130551	0.063117	2.068392	0.039121
Nox	-10.3212	3.894036	-2.65051	0.008294
Distance	0.261094	0.067947	3.842603	0.000138
Tax	-0.0144	0.003905	-3.68774	0.000251
Ptratio	-1.07431	0.133602	-8.0411	6.59E-15
Avg room	4.125409	0.442759	9.317505	3.89E-19
Lstat	-0.60349	0.053081	-11.3691	8.91E-27

Crime rate

The level of significance of Crime rate is 0.534657 which is greater than P-value 0.05. Hence Crime rate is not significant

Age

The level of significance of Age is 0.01267 Which is less than P-value 0.05. Hence Age is significant

Indus

The level of significance of Indus is 0.039121 Which is less than P-value 0.05. Hence Indus is significant

Distance

The level of significance of Distance is 0.000138 Which is less than P-value 0.05. Hence Distance is significant

Tax

The level of significance of Tax is 0.000251 Which is less than P-value 0.05. Hence Tax is significant

PTRatio

The level of significance of Ptratio is 6.59E-15 Which is less than P-value 0.05. Hence Ptratio is significant

Avg room

The level of significance of Avg room is 3.89E-19 Which is less than P-value 0.05. Hence Avg room is significant

LSTAT

The level of significance of Lstat is 8.19E-27 Which is less than P-value 0.05. Hence Lstat is 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 and answer the questions below:

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

From the above table we conclude that,

After removing Crime rate variable there is no huge difference in Multiple R , R Square and Standard error . there is some slight difference in all values**.**

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 of Previous model = 0.688299
Adjusted R Square of This model = 0.688683

This model has the highest value of adjusted R-square so this model is better than the previous model.

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?

Predictor variable	Coefficients
Nox	-10.27270508
Ptratio	-1.071702473
Lstat	-0.605159282
Tax	-0.014452345
Age	0.03293496
Indus	0.130710007
Distance	0.261506423
Avg Room	4.125468959
Intercept	29.42847349

From the table ,

The coefficient value of Nox is Negative. So, Nox is inversely proportional to the Average price .

So If the Nox is increased then the Average price will decreased

d) Write the regression equation from this model.

	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

Multi linear regression equation is,

Y = m1x1 + m2x2 + m3x3+ m4x4 +………b

The Regression equation is,

Y = (0.03293496)*Age + (0.130710007)*Indus+(- 10.27270508)*Nox + (0.261506423)*Distance+(- 0.014452345)*Tax+(-1.071702473)*Ptratio + (4.125468959)*Avg room+(-0.605159282)*Lstat+29.42847349