TERRO'S REAL ESTATE AGENCY Real estate data analysis – Exploratory data analysis, Linear Regression

-V. Apekshaa

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

Observation

1) The standard deviation of TAX is 7.49 which is the highest of all in the given table
2)The average of each variable is observed accordingly.
3)Kurtosis is Maximum for Avg_room and Minimum for INDUS
4)Skewness is maximum for Avg_price and minimum for PT Ratio
5) The standard error and standard deviation holds the lowest for NOX

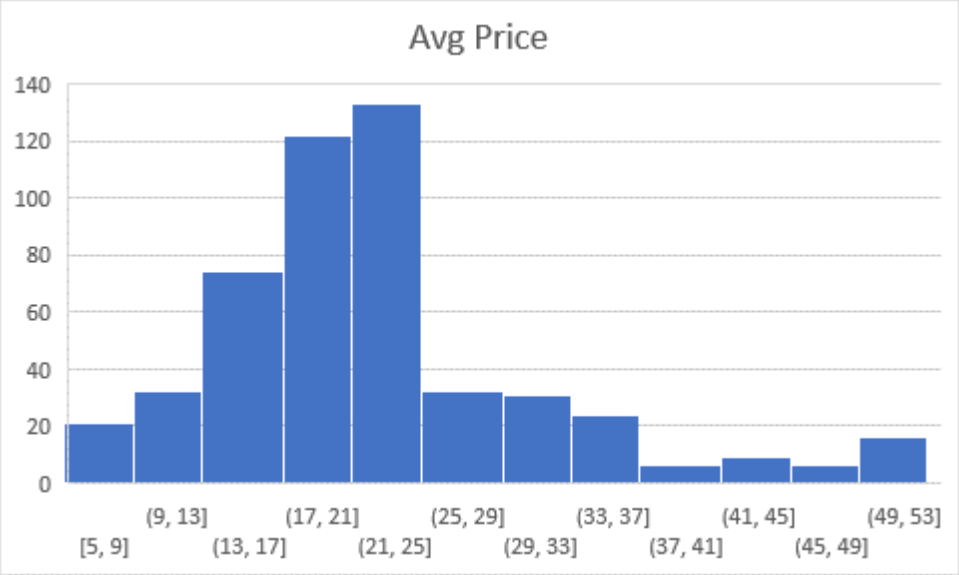
AVG ROOM	
Mean	6.284634387
Standard Error	0.031235142
Median	6.2085
Mode	5.713
Standard Deviation	0.702617143
Sample Variance	0.49367085
Kurtosis	1.891500366
Skewness	0.403612133
Range	5.219
Minimum	3.561
Maximum	8.78
Sum	3180.025
Count	506

AVG PRICE	
Mean	22.53280632
Standard Error	0.408861147
Median	21.2
Mode	50
Standard Deviation	9.197104087
Sample Variance	84.58672359
Kurtosis	1.495196944
Skewness	1.108098408
Range	45
Minimum	5
Maximum	50
Sum	11401.6
Count	506

NOX	
Mean	0.554695059
Standard Error	0.005151391
Median	0.538
Mode	0.538
Standard Deviation	0.115877676
Sample Variance	0.013427636
Kurtosis	-0.064667133
Skewness	0.729307923
Range	0.486
Minimum	0.385
Maximum	0.871
Sum	280.6757
Count	506

TAX	
Mean	408.2371542
Standard Error	7.492388692
Median	330
Mode	666
Standard Deviation	168.5371161
Sample Variance	28404.75949
Kurtosis	-1.142407992
Skewness	0.669955942
Range	524
Minimum	187
Maximum	711
Sum	206568
Count	506

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



OBSERVATION:

From the above histogram it is visually represented that the Avg_Price is distributed more on the left side and it is positive.

3)Compute the covariance matrix. Share your observations?

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	8.5161									
AGE	0.5629	790.8								
INDUS	-0.11	124.3	46.971							
NOX	0.0006	2.381	0.6059	0.013						
DISTANCE	-0.23	111.5	35.48	0.616	75.67					
TAX	-8.229	2398	831.71	13.02	1333	28349				
PTRATIO	0.0682	15.91	5.6809	0.047	8.743	167.82	4.6777			
AVG_ROOM	0.0561	-4.743	-1.884	-0.025	-1.281	-34.52	-0.54	0.493		
LSTAT	-0.883	120.8	29.522	0.488	30.33	653.42	5.7713	-3.07	50.894	
AVG_PRICE	1.162	-97.4	-30.46	-0.455	-30.5	-724.8	-10.09	4.485	-48.35	84.42

Observation:

According to the above covariance matrix the AGE and the CRIME RATE and co-related, NOX and INDUS are co-relate positively co-related with each other

The LSTAT Vs CRIME RATE, AVG PRICE Vs NOX and AVG Room VS PTRatio are inversely co-related

4) Create a correlation matrix of all the variables (Use Data analysis tool pack).
a) Which are the top 3 positively correlated pairs and
b) Which are the top 3 negatively correlated pairs.

Top 3 positively correlated pairs

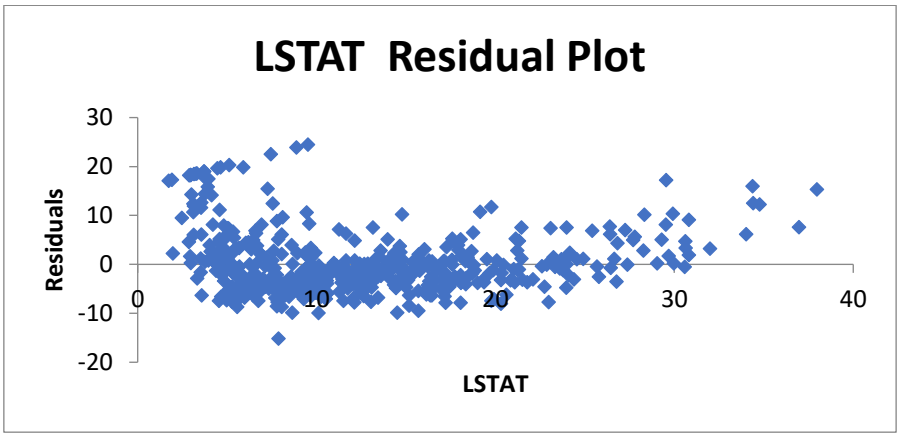
- Distance vs Tax
- NOX vs INDUS
- NOX vs AGE

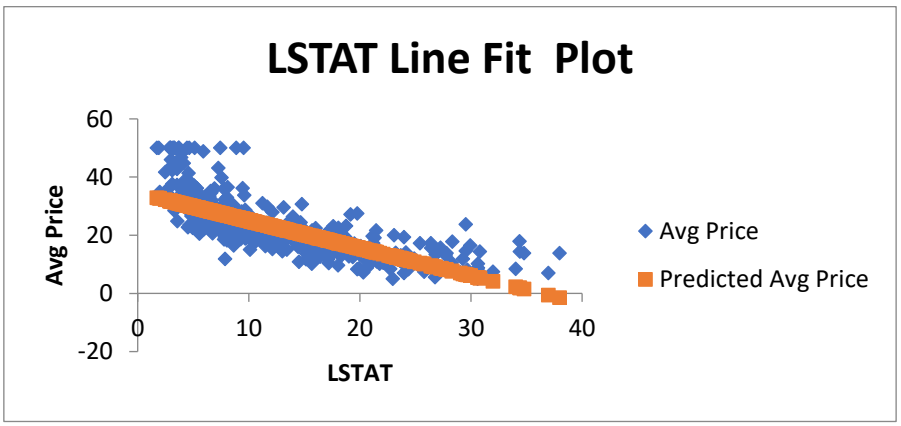
Top 3 negatively correlated pairs

- Avg price vs LSTAT
- Avg price vs LSTAT
- Avg price vs PTRATIO

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	8.5161									
AGE	0.5629	790.8								
INDUS	-0.11	124.3	46.971							
NOX	0.0006	2.381	0.6059	0.013						
DISTANCE	-0.23	111.5	35.48	0.616	75.67					
TAX	-8.229	2398	831.71	13.02	1333	28349				
PTRATIO	0.0682	15.91	5.6809	0.047	8.743	167.82	4.6777			
AVG_ROOM	0.0561	-4.743	-1.884	-0.025	-1.281	-34.52	-0.54	0.493		
LSTAT	-0.883	120.8	29.522	0.488	30.33	653.42	5.7713	-3.07	50.894	
AVG_PRICE	1.162	-97.4	-30.46	-0.455	-30.5	-724.8	-10.09	4.485	-48.35	84.42

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





Observation

1) The slope of LSTAT is -0.95 and intercept with 34.553. The residual plot shows that the LSTAT and the AVG PRICE are inversely proportional.

2) Yes, LSTAT variable is significant for analysis as they are co-related to each other.

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?

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?

Observation

Y=Mx+B **Y = 5.095*X1+(- 0.642)*X2+ (- 1.358)**

INTERCEPT	1	-1.358272812		
AVG_ROOM	7	5.094787984	**AVG PRICE**	21.45807639
L-STAT	20	-0.642358334		

1) Based on the observation the company is overcharging.
2)The performace of this model is better than the previos model as the R-square value increase from 0.5 to 0.6

Previous Model

Regression Statistics	
Multiple R	0.737662726
R Square	0.544146298
Adjusted R Square	0.543241826
Standard Error	6.215760405
Observations	506

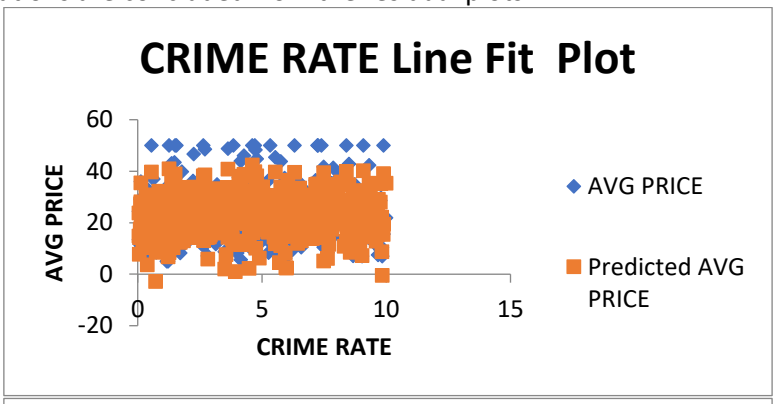
Current Model

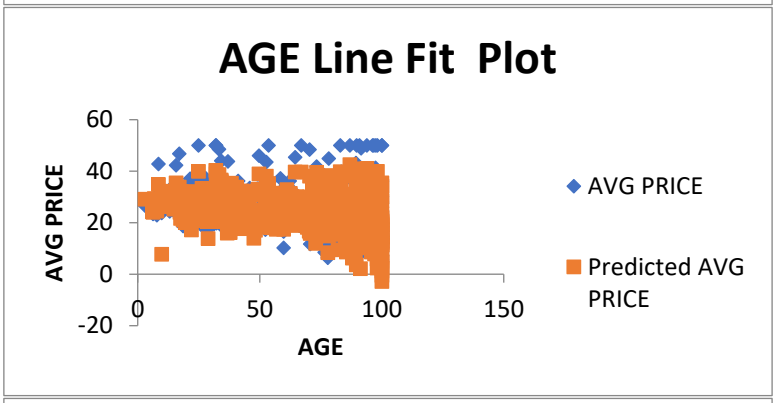
Regression Statistics	
Multiple R	0.799100498
R Square	0.638561606
Adjusted R Square	0.637124475
Standard Error	5.540257367
Observations	506

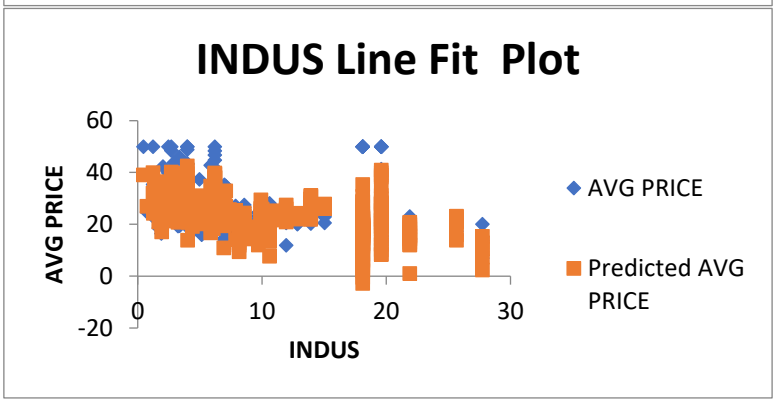
7) Build another Regression model with all variables where AVG_PRICE alone be the Dependent Variable and all the other variables are independent. Interpret the output in terms of adjusted Rsquare, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE.

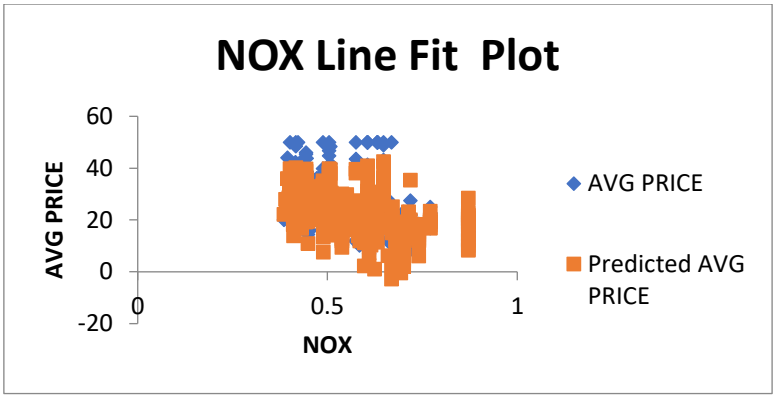
Observation

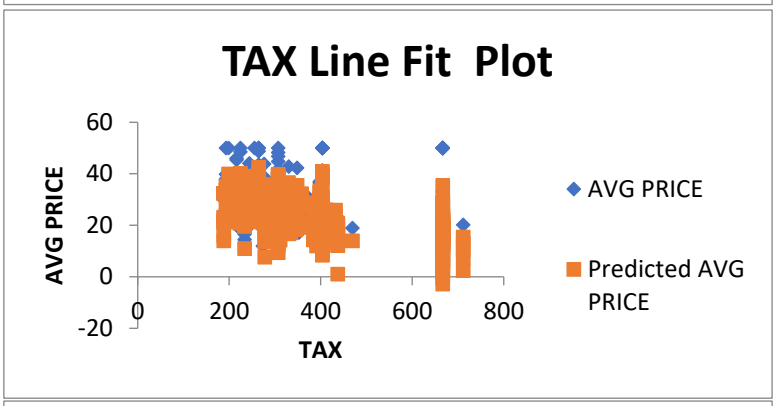
1) The avg price and the Crime rate are unifromly distributed and directly proportional to each other.
2)The avg price and the INDUS are randomly skewed
3) The avg price with Age and NOX is directly proportional to each other.
4) The DISTANCE and TAX are randomly disturbted with Avg price
5)the PTRATIO, LSTAT and AVG ROOM are all directly proportional with the avg_price.
The above all observations are concluded from the residual plots.

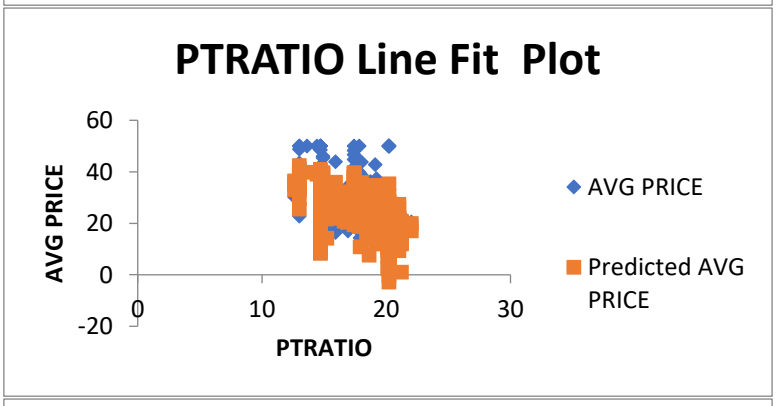


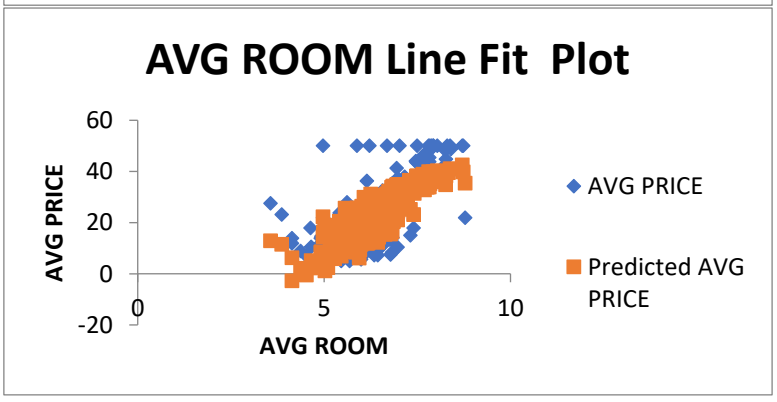


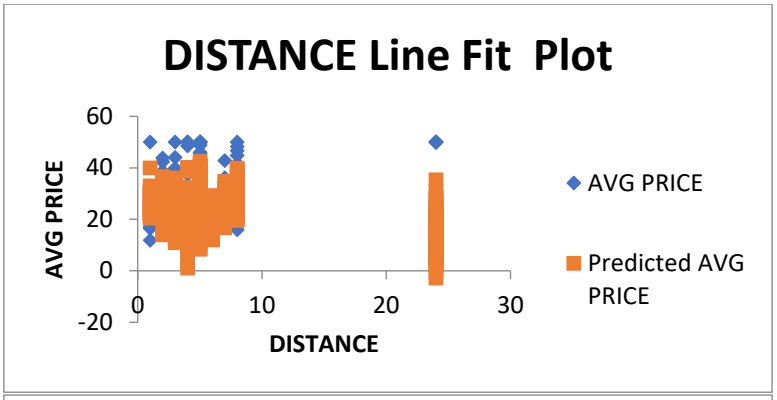


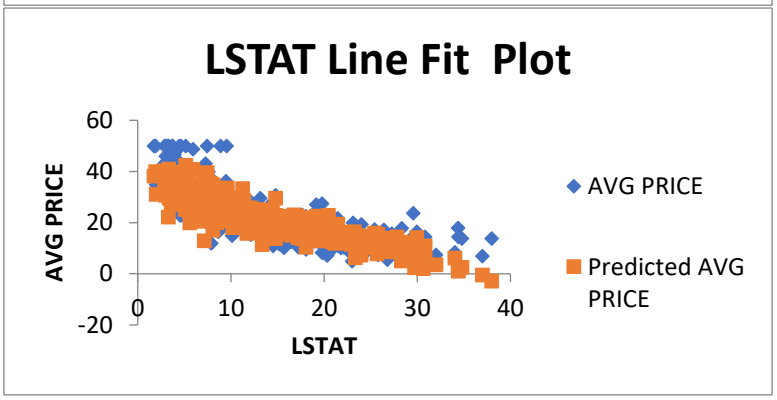

NOX Line Fit Plot
AVG PRICE
60
40
20
0
-20
0
0.5
1
NOX
AVG PRICE
Predicted AVG PRICE


TAX Line Fit Plot
AVG PRICE
60
40
20
0
-20
0
200
400
600
800
TAX
AVG PRICE
Predicted AVG PRICE


PTRATIO Line Fit Plot
AVG PRICE
60
40
20
0
-20
0
10
20
30
PTRATIO
AVG PRICE
Predicted AVG PRICE


AVG ROOM Line Fit Plot
AVG PRICE
60
40
20
0
-20
0
5
10
AVG ROOM
AVG PRICE
Predicted AVG PRICE





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.

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?

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?

d) Write the regression equation from this model.

a) This model has accuracy than others.
b) This model has the greater adjusted R square thus this is a better model.
c) If NOX is more avg price will decrease
d)y=(0.02 X1+0.017 X2+0.03 X3+0.66 X4+0.13 X5+(-0.014)X6+4.93 X7+(-0.61)X8)+1.88