EXPLORATORY DATA ANALYSIS PROJECT

PROJECT TITLE: TERRO'S REAL ESTATE AGENCY

BY

MYADAM.SAI SUDHA

GLCA-MAR 2023

1. Observations:

CRIME_RATE	
Mean	4.871976285
Standard Error	0.129860152
Median	4.82
Mode	3.43
Standard Deviation	2.921131892
Sample Variance	8.533011532
Kurtosis	-1.189122464
Skewness	0.021728079
Range	9.95
Minimum	0.04
Maximum	9.99
Sum	2465.22
Count	506

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

INDUS	
Mean	11.13677866
Standard Error	0.304979888
Median	9.69
Mode	18.1
Standard Deviation	6.860352941
Sample Variance	47.06444247
Kurtosis	-1.233539601
Skewness	0.295021568
Range	27.28
Minimum	0.46
Maximum	27.74
Sum	5635.21
Count	506

PTRATIO	
Mean	18.4555336
Standard Error	0.096243568
Median	19.05
Mode	20.2
Standard Deviation	2.164945524
Sample Variance	4.686989121
Kurtosis	-0.285091383
Skewness	-0.802324927
Range	9.4
Minimum	12.6
Maximum	22
Sum	9338.5
Count	506

LSTAT	
Mean	12.65306324
Standard Error	0.317458906
Median	11.36
Mode	8.05
Standard Deviation	7.141061511
Sample Variance	50.99475951
Kurtosis	0.493239517
Skewness	0.906460094
Range	36.24
Minimum	1.73
Maximum	37.97
Sum	6402.45
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

DISTANCE	
Mean	9.549407115
Standard Error	0.387084894
Median	5
Mode	24
Standard Deviation	8.707259384
Sample Variance	75.81636598
Kurtosis	-0.867231994
Skewness	1.004814648
Range	23
Minimum	1
Maximum	24
Sum	4832
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

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

a) Crime Rate: Out of 506 Observations we have derived that the average crime rate of town is (4.87).

b) Industry: Average of 11.13% of Non-retail business is existing per acre.

c) NOX: In the complete observation we have Minimum of 0.385 and Maximum of 0.871 of NOX Concentration in environment in the town.

d) Average Room: In the 506 observations, we have average of 6 rooms and minimum of 3 rooms per each flat.

e) Age: Average age of observations or buildings that are built after the year 1940 is derived as 68.57%

f) Distance: Distance to the highways from the purchased venture will be the key factor for the customer in terms of purchasing property.
   - In terms of these 506 observations, we have the distance to highways are within the range 23 miles and average of 9.54 miles.

g) Tax: The average tax of the properties is evaluated as 408.237 in the town.

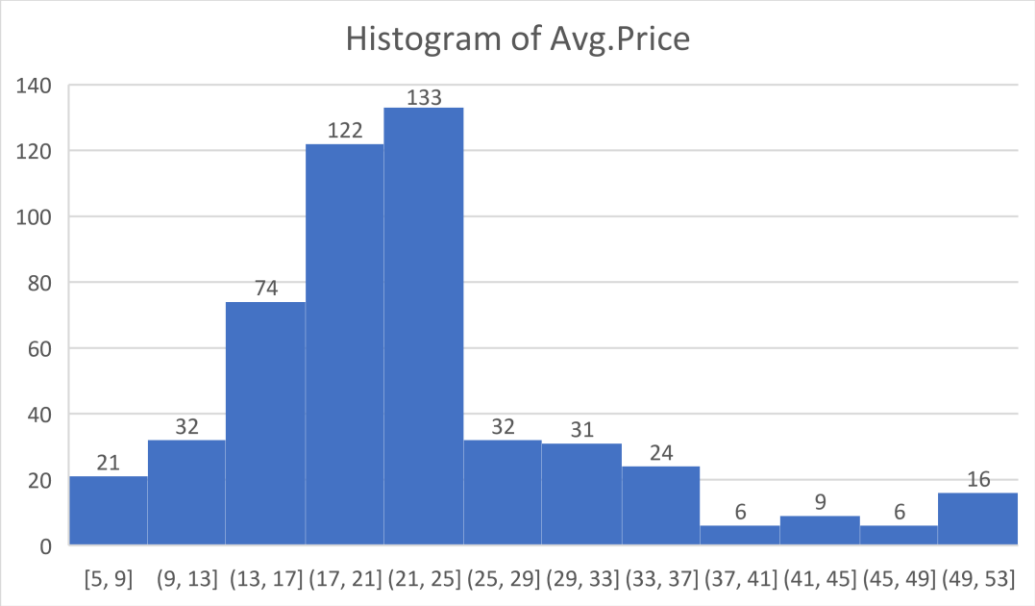
h) PT-Ratio: PT-Ratio is nothing but people to teacher ratio which basically states the availability of teachers to the total existing students.

- As per the observations made, we had evaluated that average of 18.45 teachers are available to the total students.

i) LSTAT: On an average 12% of population has lower status.

j) Average-Price: The average price of the properties has been observed 22,530$, where minimum and maximum values are 5000$ and 50,000$

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



1)The average price range is started from $5.
2)The most average price range falls ($21,$25)
3)The average price range started from $5 value 21 after that it rapidly increased to $25 value 133.But after 133 value suddenly decreasing.
4)The distribution appears to be slightly right skewed with a relatively longer tail on the right side.

Note: The values shown on x-axis of the histogram has to take in term's of 1000's $ as mentioned in the attachment.

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	790.7924728								
INDUS	-0.110215175	124.2678282	46.97142974							
NOX	0.000625308	2.381211931	0.605873943	0.013401099						
DISTANCE	-0.229860488	111.5499555	35.47971449	0.615710224	75.66653127					
TAX	-8.229322439	2397.941723	831.7133331	13.02050236	1333.116741	28348.6236				
PTRATIO	0.068168906	15.90542545	5.680854782	0.047303654	8.74340249	167.8208221	4.677726296			
AVG_ROOM	0.056117778	-4.74253803	-1.884225427	-0.024554826	-1.281277391	-34.51510104	-0.539694518	0.492695216		
LSTAT	-0.882680362	120.8384405	29.52181125	0.487979871	30.32539213	653.4206174	5.771300243	-3.073654967	50.89397935	
AVG_PRICE	1.16201224	-97.39615288	-30.46050499	-0.454512407	-30.50083035	-724.8204284	-10.09067561	4.484565552	-48.35179219	84.41955616

Covariance:

1)In terms of statistics covariance is nothing but finding relationship between two or more variables.

2)If the output value of 2 variables is positive then it is said to be both the variables are travelling in same direction, which means the variables are exhibiting similar behavior which is said to be positive covariance.

3)If the output value of two variables is negative then it is said to be both the variables are travelling in the opposite direction and exhibiting different behavior which is said to be negative covariance.

4)In the above generated covariance matrix TAX variable has positively strong covariance with all variables except for CRIME_RATE variable which has negative covariance.

5)Tax with Age variable has high variance in positive side.

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

A	B	C	D	E	F	G	H	I	J	K
	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	1									
AGE	0.006859463	1								
INDUS	-0.005510651	0.644778511	1							
NOX	0.001850982	0.731470104	0.763651447	1						
DISTANCE	-0.009055049	0.456022452	0.595129275	0.611440563	1					
TAX	-0.016748522	0.506455594	0.72076018	0.6680232	0.910228189	1				
PTRATIO	0.010800586	0.261515012	0.383247556	0.188932677	0.464741179	0.460853035	1			
AVG_ROOM	0.02739616	-0.240264931	-0.391675853	-0.302188188	-0.209846668	-0.292047833	-0.355501495	1		
LSTAT	-0.042398321	0.602338529	0.603799716	0.590878921	0.488676335	0.543993412	0.374044317	-0.613808272	1	
AVG_PRICE	0.043337871	-0.376954565	-0.48372516	-0.427320772	-0.381626231	-0.468535934	-0.507786686	0.695359947	-0.737662726	1

Correlation: It is a statistical measure which expresses the extension to which the variables are linearly inclined to each other and the value lies between -1 to +1.

a) Which are the top 3 positively correlated pairs?

- From the above generated correlation matrix output we have 3 strongly correlated variables as mentioned below.

1)Distance – Tax

2)NOX – Age

3)NOX – Indus

b) Which are the top 3 negatively correlated pairs?

- We have 3 negatively correlated pairs as mentioned below

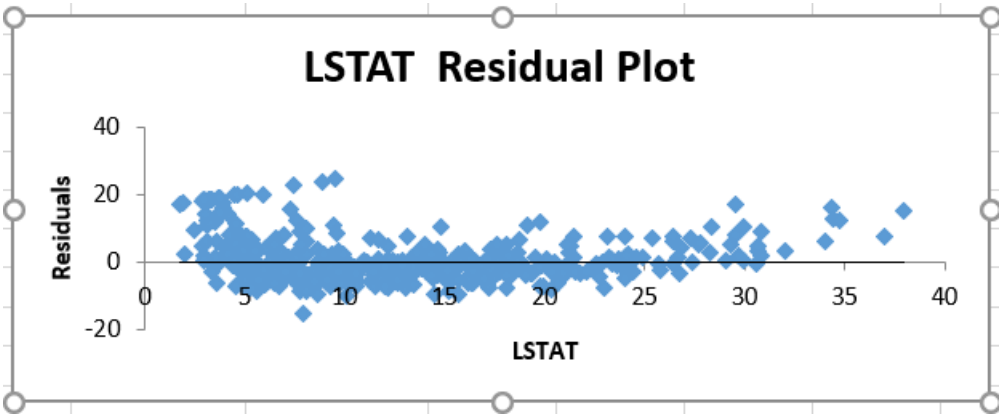
1)LSTAT – Avg-Room

2)Avg-Price – PTRATIO

3)Avg-Price – LSTAT

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

Generated Output:



1) The R-squared value is 0.544.
2) which means approximately 54.4% of the variance in AVG_PRICE is explained by the independent variable (LSTAT).
3) This indicates a moderate level of variance explained.
4) Adjusted R-squared: The adjusted R-squared value is 0.544,
5) which is slightly lower than the R-squared value.
6) The model's goodness of fit.

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

SUMMARY OUTPUT

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

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	34.55384088	0.562627355	61.41514552	3.7431E-236	33.44845704	35.65922472	33.44845704	35.65922472
LSTAT	-0.950049354	0.038733416	-24.52789985	5.0811E-88	-1.0261482	-0.873950508	-1.0261482	-0.873950508

1) R-squared (Multiple R-squared): The R-squared value is 0.544.
2) which means approximately 54.4% of the variance in AVG_PRICE is explained by the independent variable (LSTAT).
3) This indicates a moderate level of variance explained.
4) Adjusted R-squared: The adjusted R-squared value is 0.544, which is slightly lower than the R-squared value.
5) The model's goodness of fit.
6) From the above generated output, we see that there is 54% of the variation in the average price can be explained by the LSTAT.
7) The coefficient of LSTAT for this model is -0.950049354.
8) Intercept of LSTAT for this model is 34.55384088.

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

1)Yes, LSTAT is the significant variable of the avg-price for this model.
2)As the p-value (5.08E-88) we obtained from this model is very less than 0.05. and we already know the p-value is less than 0.05 it was significant to target variable.
3)By this way we can say that LSTAT is a significant variable according to this model.

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

SUMMARY OUTPUT

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

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.358272812	3.17282778	-0.428095348	0.668764941	-7.591900282	4.875354658	-7.591900282	4.875354658
AVG_ROOM	5.094787984	0.4444655	11.46272991	3.47226E-27	4.221550436	5.968025533	4.221550436	5.968025533
LSTAT	-0.642358334	0.043731465	-14.68869925	6.66937E-41	-0.728277167	-0.556439501	-0.728277167	-0.556439501

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?

Calculation:

A) The regression equation Y= A+BX
Avg-price= -3.37+5.094*avg_room-0.642*LSTAT

- 7 rooms and LSTAT value of 20 substitute value into the equation.
- AVG-Price = -3.37+5.09*7-0.642*20=21.472
- The predicted value of avg _price for this house is 19.42.
- Which is lesser then the company's quoted value of 30000.

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.

- The model with Avg-room, LSTAT is better than the previous model of only LSTAT.
- Because we clearly observed and understand that r square value of this model was 0.6385 it is greater than previous model with r square value of 0.5441.
- so, this model is more significant and have more variability.

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?

Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372
Adjusted R Square	0.688298647
Standard Error	5.1347635
Observations	506

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.24131526	4.817125596	6.070282926	2.53978E-09	19.77682784	38.70580267	19.77682784	38.70580267
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.534657201	-0.105348544	0.202798827	-0.105348544	0.202798827
AGE	0.032770689	0.013097814	2.501996817	0.012670437	0.00703665	0.058504728	0.00703665	0.058504728
INDUS	0.130551399	0.063117334	2.068392165	0.03912086	0.006541094	0.254561704	0.006541094	0.254561704
NOX	-10.3211828	3.894036256	-2.650510195	0.008293859	-17.97202279	-2.670342809	-17.97202279	-2.670342809
DISTANCE	0.261093575	0.067947067	3.842602576	0.000137546	0.127594012	0.394593138	0.127594012	0.394593138
TAX	-0.01440119	0.003905158	-3.687736063	0.000251247	-0.022073881	-0.0067285	-0.022073881	-0.0067285
PTRATIO	-1.074305348	0.133601722	-8.041104061	6.58642E-15	-1.336800438	-0.811810259	-1.336800438	-0.811810259
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.89287E-19	3.255494742	4.995323561	3.255494742	4.995323561
LSTAT	-0.603486589	0.053081161	-11.36912937	8.91071E-27	-0.70777824	-0.499194938	-0.70777824	-0.499194938

- From the above generated regression statistics model, we observed that 69% of the variation in the average price is explained by the all the independent variables in table.
- The intercept of this model is derived as 29.24131.
- The coefficient of CRIME_RATE is 0.048725141 this means that much times the dependent variable changes while changing it.
- The coefficient of variables is derived as below mentioned values from regression model.
- AGE is 0.032770689
- INDUS is 0.130551399
- NOX is -10.3211828
- DISTANCE is 0.261093575
- TAX is -0.01440119
- PTRATIO is -1.074305348
- AVG_ROOM is 4.125409152
- LSTAT is -0.603486589
- How the independent variable Significance to dependent variable is decided by p value. From this model the p value od every single variable must be less than 0.05 then only we decide that it is significant to the y variable(avg-price).

Variable	p-value	significant or non-significant
CRIME_RATE	0.534657201	non-significant variable
AGE	0.012670437	significant variable
INDUS	0.03912086	significant variable
NOX	0.008293859	significant variable
DISTANCE	0.000137546	significant variable
TAX	0.000251247	significant variable
PTRATIO	6.58642E-15	significant variable
AVG_ROOM	3.89287E-19	significant variable
LSTAT	8.91071E-27	significant variable

- Finally, I concluded that except crime-rate all the independent variable are significant to the average 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.

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372
Adjusted R Square	0.688298647
Standard Error	5.1347635
Observations	506

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*
Intercept	29.24131526	4.817125596	6.070282926	2.53978E-09	19.77682784	38.70580267	19.77682784
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.534657201	-0.105348544	0.202798827	-0.105348544
AGE	0.032770689	0.013097814	2.501996817	0.012670437	0.00703665	0.058504728	0.00703665
INDUS	0.130551399	0.063117334	2.068392165	0.03912086	0.006541094	0.254561704	0.006541094
NOX	-10.3211828	3.894036256	-2.650510195	0.008293859	-17.97202279	-2.670342809	-17.97202279
DISTANCE	0.261093575	0.067947067	3.842602576	0.000137546	0.127594012	0.394593138	0.127594012
TAX	-0.01440119	0.003905158	-3.687736063	0.000251247	-0.022073881	-0.0067285	-0.022073881
PTRATIO	-1.074305348	0.133601722	-8.041104061	6.58642E-15	-1.336800438	-0.811810259	-1.336800438
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.89287E-19	3.255494742	4.995323561	3.255494742
LSTAT	-0.603486589	0.053081161	-11.36912937	8.91071E-27	-0.70777824	-0.499194938	-0.70777824

- From this regression statistics we observed that 69% of the variation in the average price is explained by the all the independent variables in table.

- The intercept of this model is 29.24131.

8) 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?

- By the value of adjusted R-square value little more compared to previous question 7. So, current model performs better in these two models.

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?

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

- Average price will decrease if the value of NOX is increased in a locality of this town.

d)Write the regression equation from this model.

- Average Price = 29.4285 + 0.0329 * AGE + 0.1307 * INDUS - 10.2727 * NOX + 0.2615 * DISTANCE - 0.0145 * TAX - 1.0717 * PTRATIO + 4.1255 * AVG_ROOM - 0.6052 * LSTAT.
- 29.42847349+0.03293496(65.2)+ 0.130710007(2.31) -10.27270508(0.538)+ 0.261506423(1) -0.014452345(296) -1.071702473(15.3)+ 4.125468959(6.575) -0.605159282(4.98)

  =21.4581

Conclusion Statement:

- Terro's real-estate is an agency provided a table containing 10 different variables. Were the average price was the targeted variable and all other variables are independent variables.
- Using 506 observations, we can state that the average-price of houses was most frequently occurred in the price range of 21,000$ to 25,000$, most of the clients are more interested in purchasing the house falling in the medium price range.
- All the variables are significant to Average-Price except Crime-Rate.