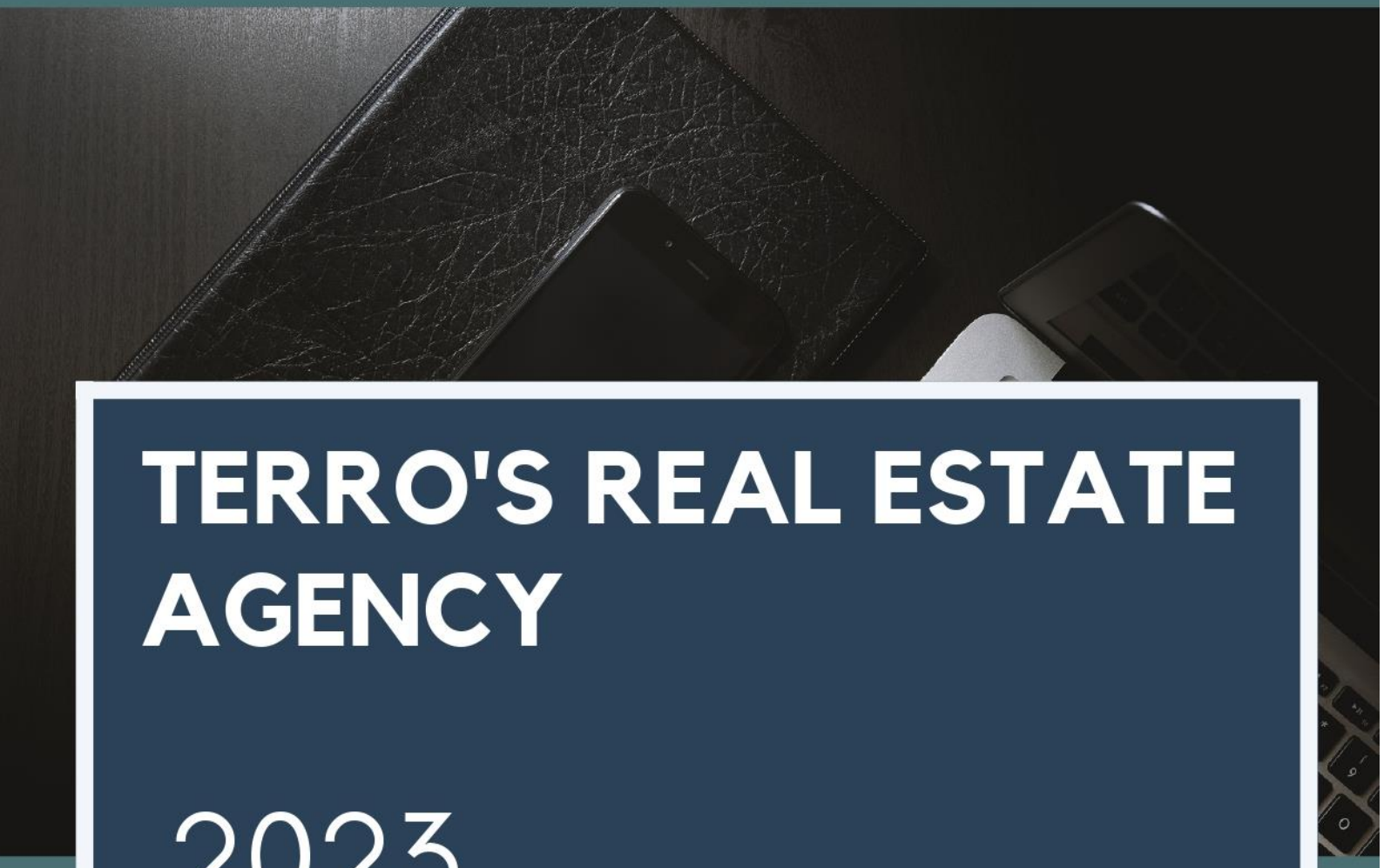
GREATLEARNING



TERRO'S REAL ESTATE AGENCY

2023

Real estate data analysis - Exploratory data analysis, Linear Regression

Prepared By:
Kishore kumar.D

Terro’s real estate agency

Table of contents:

Terro's real estate agency

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

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

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

Terro's real estate agency

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

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

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

Terro's real estate agency

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

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

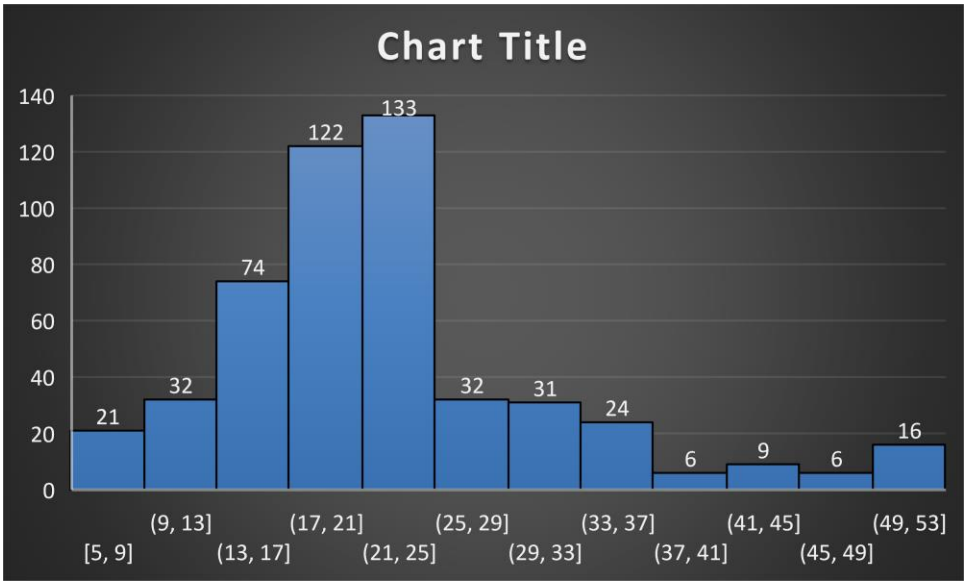
OBSERVATION:

1. We create a summary statistic with the help of the data analysis tool Pak.
2. In this statistic table they are several details but we take skewness and kurtosis for the observation.
3. In this table AGE and PTRATIO are negatively skewed.
4. Remaining variables are positively skewed.
5. Kurtosis of the Average Room has the highest positive value is 1.89
6. The 1.89 is the lepto Kurtic, -1.23 is my platy Kurtic.

7. Kurtosis lies between the -2 to 2.

Terro's real estate agency

Question: 2 Plot a histogram of the Avg_price variable. What do you infer?



OBSERVATION:

1. We create histogram chart using the average price variable data.
2. The average price of $21000 to $25000 has the highest preference.
3. This histogram chart shows they are 133 houses in Boston priced between $21000 to $25000.
4. The average price of $37000 to $41000 and $45000 to $49000 has the least preference.
5. This histogram chart shows they are 6 houses in Boston priced between $37000 to $41000.
6. And another 6 houses in Boston priced between $45000 to $49000.

Terro's real estate agency

Question:3 Compute the covariance matrix. Share your observations.

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RA	8.516147873									
AGE	0.562915215	791								
INDUS	-0.110215175	124	46.97							
NOX	0.000625308	2.38	0.606	0.013						
DISTANCE	-0.229860488	112	35.48	0.616	75.6665					
TAX	-8.229322439	2398	831.7	13.02	1333.12	28349				
PTRATIO	0.068168906	15.9	5.681	0.047	8.7434	167.8	4.67773			
AVG_ROO	0.056117778	-4.74	-1.88	-0.02	-1.28128	-34.5	-0.5397	0.49269522		
LSTAT	-0.882680362	121	29.52	0.488	30.3254	653.4	5.7713	-3.073655	50.894	
AVG_PRIC	1.16201224	-97.4	-30.5	-0.45	-30.5008	-725	-10.091	4.48456555	-48.35	84.419556

OBSERVATION:

1. We take all the variables and use to create the covariance matrix with the help of data analysis tool Pak.
2. In this covariance table the crime rate and average room are positive correlated.
3. A positive value indicates that two variables will decrease or increase in the same direction.
4. And the other all variables are negatively correlated.
5. A negative value indicates that if one variable decrease, the other increase, and an inverse relationship exist between them.
6. The upper part of the diagonal is empty as the covariance matrix is symmetric towards the diagonal.

Terro's real estate agency

Question: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

	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

The top 3 positively correlated pairs:

1. The relationship between Distance and Tax.
2. The relationship between Indus and Nox.
3. The relationship between Age and Nox.

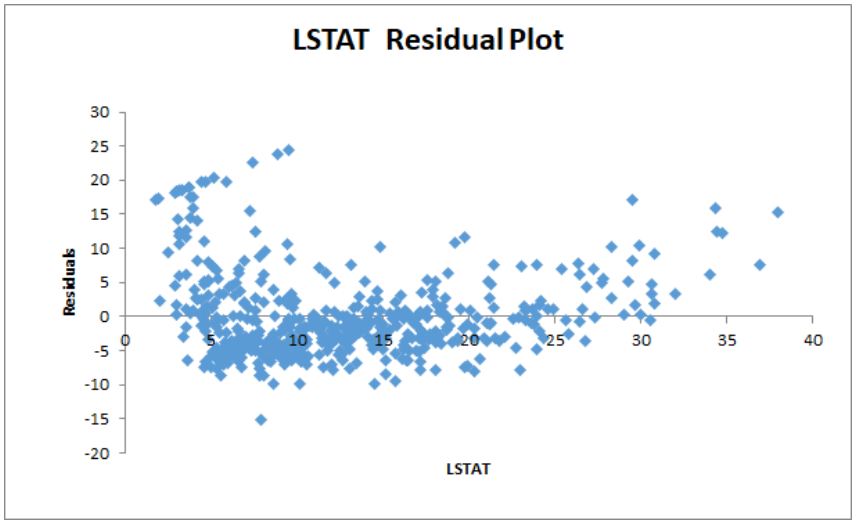
The top 3 Negatively correlated pairs:

1. The relationship between Lstat and Average price.
2. The relationship between Average room and Lstat.
3. The relationship between Ptratio and Average price.

Terro's real estate agency

Question :5A Build an initial regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot. What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?

OBSERVATION:



R square: 0.544146

Coefficients: - 0.95005

Intercept: 34.55384

1. The R square value is 54% it is slightly low value because if we get more percentage in the R square it will gives nearest answer.
2. The coefficient value is negative so that average price and Lstat are inversely proportional.
3. Lstat residual plot line facing downward, it means they should be a negative correlation.

Terro's real estate agency

Question :5B Is LSTAT variable significant for the analysis based on your model?

OBSERVATION:

P-value of the Lstat is 5.08E-88

1. In this question confidence level is not given, so we assume the confidence level is 95%.
2. Desired level of significance = 1-confidence level.
3. Then my significance level is 5%
4. The P value of Lstat is lesser than the significance level.
5. So Lstat variable is significant for analysis and it will reject the null hypothesis.

Question: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?

Terro's real estate agency

OBSERVATION:

6A.

1. In this question they gave the x value. The x1 is 7 and X2 is 20.
2. M value of average room is 5.094788(M1) and LSTAT is -0.64236(M2). And the value of intercept is -1.35827.
3. The formula for the regression is Y=m1x1+m2x2+b.
4. Y=5.094788*7+(-0.64236) *20+(-1.35827)
5. we can solve this we get 21458.
6. The company quoting value is 30000.
7. We can conclude with the help of this details; the company is overcharging.

6B.

1. The adjusted R square of this question 64%.
2. The adjusted R square of the previous question 54%

A. The adjusted R square of this question is better than the previous question, Because the adjusted value of this question is 64% and the previous question value is 54%.

B. So that the performance of Lstat, Average room, Average price model is best compared to the last question model.

Terro's real estate agency

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.

OBSERVATION:

7A.

Adjusted R Square	0.688298647
Intercept	29.24131526
CRIME_RATE	0.048725141
AGE	0.032770689
INDUS	0.130551399
NOX	-10.3211828
DISTANCE	0.261093575
TAX	-0.01440119
PTRATIO	-1.074305348
AVG_ROOM	4.125409152
LSTAT	-0.603486589

1. The Adjusted R square value is 69%, getting with the solving of all independent variable.
2. The adjusted R square value of this question is 69%, this is more effective compared to the previous question.
3. The coefficient of NOX, TAX, LSTAT, PTRATIO, are negatively correlated. So, they are inversely proportional.
4. The remaining variable are positively correlated. So, they are directly proportional.

Terro's real estate agency

7B.

	P-value	significant or not
CRIME_RATE	0.534657201	not significant
AGE	0.012670437	significant
INDUS	0.03912086	significant
NOX	0.008293859	significant
DISTANCE	0.000137546	significant
TAX	0.000251247	significant
PTRATIO	6.58642E-15	significant
AVG_ROOM	3.89287E-19	significant
LSTAT	8.91071E-27	significant

P-value <0.05 is significant.

P-value >0.05 is not significant.

1. In this P value table, the crime rate is greater than 0.05 it is not significant.
2. AGE, INDUS, DISTANCE, TAX, PTRATIO, AVERAGE ROOM, LSTAT variables p value is less than 0.05 it is significant.

Question: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:

Adjusted R Square	0.688684
Intercept	29.42847
AGE	0.032935
INDUS	0.13071
NOX	-10.2727
DISTANCE	0.261506
TAX	-0.01445
PTRATIO	-1.0717
AVG_ROOM	4.125469
LSTAT	-0.60516

Terro's real estate agency

a) Interpret the output of this model.

1. AGE, INDUS, DISTANCE, AVERAGE ROOM, has the positive coefficient value.
2. NOX, TAX, PTRATIO, LSTAT has the negative coefficient value.

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?

1. The Adjusted R square value of the previous question is 69% taking all significant and insignificant variables into regression.
2. The Adjusted R square value of the current question is also 69% taking only significant variables into regression.
3. So, this model is slightly better than the previous, because of we taking only significant variables in this question.

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?

Solution: The coefficient of NOX is negative, so the NOX is inversely proportional to the average price.

d) Write the regression equation from this model.

Y=(M1*X1+M2*X2+M3*X3+M4*X4+M5*X5+M6*X6+M7*X7+M8*X8) +B

Y= (0.0329349604286303) * Age (X1) + 0.130710006682182 * Indus (X2) + (- 10.2727050815094) * NOX (X3) + 0.261506423001819 * DISTANCE (X4) + (- 0.0144523450364819) * TAX (X5) + (-1.07170247269449) * PTRATIO (X6) + 4.12546895908474 * AVG_ROOM (X7) + (-0.605159282035406) * LSTAT (X8) + B