TERRO'S REAL ESTATE AGENCY

Real estate data analysis – Exploratory data analysis , Linear Regression .

TABLE OF CONTENTS

QUESTION – 1:

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

CRIME_RATE	
Mean	4.871976
Standard Error	0.12986
Median	4.82
Mode	3.43
Standard Deviation	2.921132
Sample Variance	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 Error	1.25137
Median	77.5
Mode	100
Standard Deviation	28.14886
Sample Variance	792.3584
Kurtosis	-0.96772
Skewness	-0.59896
Range	97.1
Minimum	2.9
Maximum	100
Sum	34698.9
Count	506

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

INDUS	
Mean	11.13678
Standard Error	0.30498
Median	9.69
Mode	18.1
Standard Deviation	6.860353
Sample Variance	47.06444
Kurtosis	-1.23354
Skewness	0.295022
Range	27.28
Minimum	0.46
Maximum	27.74
Sum	5635.21
Count	506

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

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

AVG_ROOM	
Mean	6.284634
Standard Error	0.031235
Median	6.2085
Mode	5.713
Standard Deviation	0.702617
Sample Variance	0.493671
Kurtosis	1.8915
Skewness	0.403612
Range	5.219
Minimum	3.561
Maximum	8.78
Sum	3180.025
Count	506

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

AVG_PRICE	
Mean	22.53281
Standard Error	0.408861
Median	21.2
Mode	50
Standard Deviation	9.197104
Sample Variance	84.58672
Kurtosis	1.495197
Skewness	1.108098
Range	45
Minimum	5
Maximum	50
Sum	11401.6
Count	506

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

Obsevations

Crime_Rate is least Negligibly Skewed.

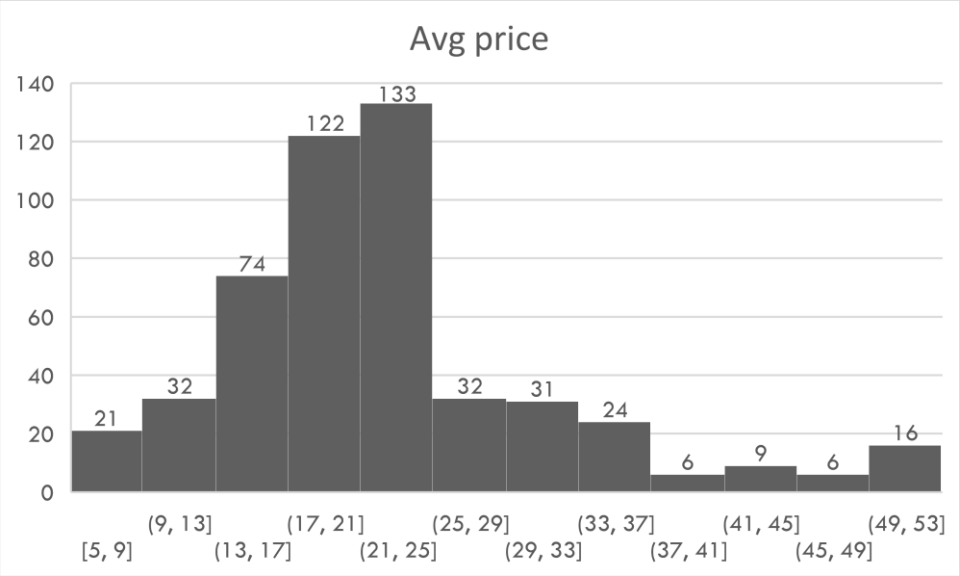
Age and PT Ratio are Negatively Skewed.

Indus, Nox, Distance, Tax, Avg_Room, Lstat and Avg_Price are Positively skewed

Crime_Rate,age,PT ratio,Indus, Nox, Distance, Tax, Avg_Room, Lstat and Avg_Price does not have a flat distribution as kurtosis is between -2 to 2

QUESTION 2:

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



Inference

The avg price range between 21-25 has the highest count

The avg price range between 37-41 and 45-49 has the least count

QUESTION 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

Inference

The Avg_price is directly proportional to Crime_Rate and Avg_Room.

The Avg_price is inversely proportional to age,indus,nox,distance,tax,ptratio,lstat and Avg_Room.

QUESTION 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									
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

Which are the top 3 positively correlated pairs

The top 3 positively correlated pairs

Tax/Distance	0.910228189
Nox/indus	0.763651447
Nox/age	0.731470104

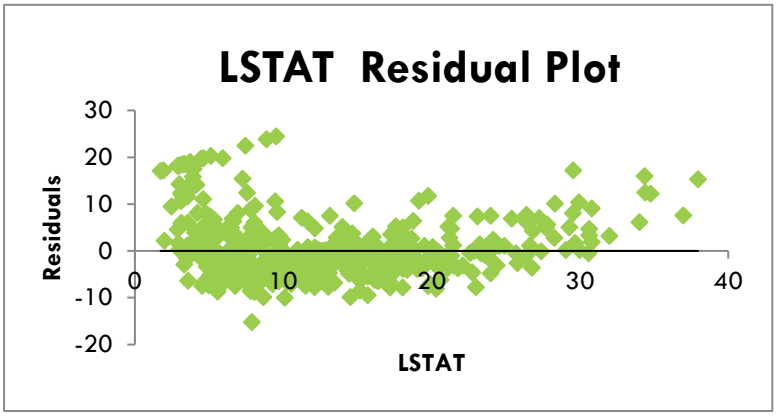
Which are the top 3 negatively correlated pairs.

The top 3 negatively correlated pairs

INDUS/CRIME_RATE	-0.00551
DISTANCE/CRIME_RATE	-0.00906
DISTANCE/CRIME_RATE	-0.01675

QUESTION 5:

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



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

R square	0.544146
Intercept	34.55384
coefficient	-0.95005

The R Square of this model is 54% thus it can be improved and the coefficient is negative insist that Avg price and Lstat are inversely proportional

From this Residual Plot we are unable to see any Patterns and we can able to perform Linear regression. Hence it is known as Homoskedasticity.

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

Significant Value (P-Value)	5.08E-88

Yes, Lstat Variable is Significant for analysis, because P Value of Lstat is lesser than 0.05

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?

Regression Equation

Y = M1*X1+M2*X2+B M1 = 5.09478798433655 , M2 = -0.642358334244129, B = -1.35827281187456

Y = 5.09478798433655 * X1(AVG_ROOM) + (-0.642358334244129) * X2(LSTAT)+ (- 1.35827281187456)

X1(AVG_ROOM) = 7, X2(LSTAT) = 20

Y = 5.09478798433655 * 7 + (-0.642358334244129) * 20+ (-1.35827281187456)

Y = 21.4580763935987 * 1000 USD = 21458 USD

AVG_PRICE = 21458 USD

Terro'S Real Estate Agency Company Quoting Price 30000 USD.

From This Calculation We Concluded That Company Quoting Price Is Greater Than Avg_Price. Hence Company Is Overcharging.

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 Square of Lstat, Avg_Room and Avg_price	0.637124	64%
Adjusted R Square for Lstat and Avg_Price	0.543242	54%

The Performance of LSTAT, Avg_Room and Avg_Price Model is better than LSTAT and Avg_Price as we have high Rsquare or Adjusted R Square Value

Here we have High Adjusted R Square value for LSTAT, Avg_Room and Avg_Price Model Compared to LSTAT and Avg_Price Model.

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 Rsquare, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE*.**

Adjusted R Square	0.688299
Intercept	29.24132
coefficients	
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

The Adjusted R Square is 69% comparing to the before model this is more effective
As coefficients NOX,PTRATIO,LSTAT,TAX are negatively correlated so they inversely proportional to AVG_PRICE whereas the other coefficient such as CRIME_RATE,AGE,INDUS,DISTANCE,AVG_ROOM are directly proportional to AVG_PRICE

SIGNIFICANCE OF ALL VARIABLE WITH Y:

CRIME_RATE	0.534657
AGE	0.01267
INDUS	0.039121
NOX	0.008294
DISTANCE	0.000138
TAX	0.000251
PTRATIO	6.59E-15
AVG_ROOM	3.89E-19
LSTAT	8.91E-27

P Value < 0.05 is Significant .
P Value > 0.05 is Insignificant
The Age, Indus, Nox, Distance, Tax, PTRatio, Avg_Room, Lstat are Significant with Avg_Price as P Values are less tha
Crime Rate is Insignificant with Avg_Price as P Value is greater than 0.05.

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:

Interpret the output of this model.

Avg_Room increases, the mean of Avg_Price also Increases as it has Positive Coefficient value.
If LSTAT, Age, Nox, Distance, Tax, PTRatio, Increases Avg_Price Decreases as it as Negative Coefficient value.

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 FOR AGE,INDUS,NOX,DISTANCE,TAX,PTRATIO,AVG_ROOM,LSTAT AND AVG_PRICE 0.688683682 69%
ADJUSTED R SQUARE FOR ALL VARIABLE VS AVG_PRICE 0.688298647 69% 0.688298647 69%
Comparing two adjusted R Square Values, there is very Slight difference in adjusted R square Value but it can be Noticed

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	10.27271
PTRATIO	1.071702
LSTAT	0.605159
TAX	0.014452
AGE	0.032935
INDUS	0.13071
DISTANCE	0.261506
AVG_ROOM	4.125469

Th NOX variable is inversely proportional to Avg_Price

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