ASSIGNMENT – TERRO’S REAL ESTATE AGENCY

Real estate data analysis – Exploratory data analysis, Linear Regression

PROBLEM STATEMENT: Terro’s real-estate is an agency that estimates the pricing of houses in a certain locality. The pricing is concluded based on different features / factors of a property. This also helps them in identifying the business value of a property. To do this activity the company employs an “Auditor”, who studies various geographic features of a property like pollution level (NOX), crime rate, education facilities (pupil to teacher ratio), connectivity (distance from highway), etc. This helps in determining the price of a property

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

Solution:

1.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

From the above data of crime rate

- Mean value is 4.87, Median is 4.82 and Mode is 3.43.
- According to the data of crime rate,mean is greater than median hence it is slightly right-skewed.
- The range of crime rate is from 0.04 to 9.99.
- The variable has a moderate amount of variability, with a standard deviation of 2.92.

2.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

From the above data of age

- Mean value is 68.87 it means average life of a house is 68.87 Median is 77.5 and Mode is 100 it states that most of the age of house is 100.
- According to the data of age mean, is lesser than median hence it is slightly left-skewed.
- The range of age of house is between 2.9 and 100.
- The variable has a moderate amount of variability, with a standard deviation of 28.14.

3.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

From the above data of indus

- Mean value is 11.13, Median is 9.69 and Mode is 18.1.
- According to the data of indus, mean is greater than median hence it is slightly right-skewed.
- There is a wide range of industrial proportions, from 0.46 to 27.74.
- The variable has a moderate amount of variability,with standard deviation of 6.86.

4.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

From the above data of nox

- Mean value is 0.55, Median and Mode is 0.538.
- According to the data of nox, mean is greater than median hence it is slightly right-skewed.
- The range of nox is from 0.04 to 9.99.
- The variable has a moderate amount of variability, with a standard deviation of 0.115.

5.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

From the above data of distance

- Mean value is 9.5 it states the average distance from highway, Median is 5 and Mode is 24 it states that most of the houses are far from highway.
- According to the data of distance, mean is greater than median hence it is right-skewed.

- The range of distance lies between 1 to 24.
- The variable has a moderate amount of variability, with a standard deviation of 8.70.

6.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

From the above data of tax

- Mean tax value is 408.23 , Median is 330 and Mode is 666.
- According to the data of tax ,mean is greater than median hence it is right-skewed.
- The range of tax rate is from 187 to 524.
- The variable has a moderate amount of variability, with a standard deviation of 168.53.

7.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

From the above data of ptratio

- Mean value is 18.45, Median is 19.05 and Mode is 20.2.

- According to the data of ptratio ,mean is greater than median hence it is slightly left-skewed..
- The range of ptratio is from 12.6 to 22.
- The variable has a moderate amount of variability, with a standard deviation of 2.16.

8.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

From the above data of average room

- Mean value is 6.28, Median is 6.20 and Mode is 5.73.
- According to the data of avg_room, mean is greater than median hence it is slightly right-skewed.
- The range of avg room is from 3.56 to 8.78.
- The variable has a moderate amount of variability, with a standard deviation of 0.70.

.

9.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

From the above data of lstat

- Mean value is 12.65, Median is 11.36 and Mode is 8.05.
- According to the data of ltsat, mean is greater than median hence it is slightly right-skewed.
- The range of lstat is from 1.73 to 37.97.
- The variable has a moderate amount of variability, with a standard deviation of 7.14.

10.AVG_PRICE

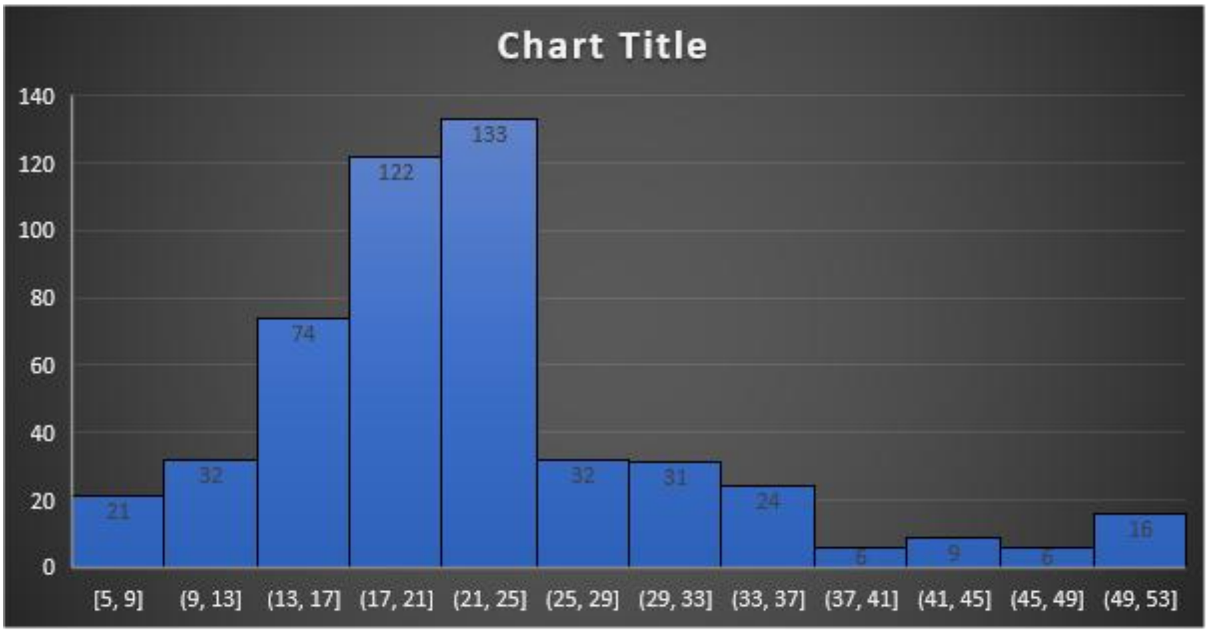
Mean	22.53280632
Standard Error	0.408861147
Median	21.2
Mode	50
Standard Deviation	9.197104087
Sample Variance	84.58672359
Kurtosis	1.495196944
Skewness	1.108098408
Range	45
Minimum	5
Maximum	50
Sum	11401.6
Count	506

From the above data of average price

- Mean value is 22.53, Median is 21.2 and Mode is 50.
- According to the data of average price, mean is greater than median hence it is slightly right-skewed.
- The range of crime rate is from 5 to 50.
- The variable has a moderate amount of variability, with a standard deviation of 9.19.

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

Solution:



From the above histogram we can see that most of the houses are in range between $21000 to $25000 that is 133 and least in range between $37000 to $41000 and $45000 to $49000 that is 6

QUESTION 3: Compute the covariance matrix. Share your observations

Solution :

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	8.516									
AGE	0.563	790.792								
INDUS	-0.110	124.268	46.971							
NOX	0.001	2.381	0.606	0.013						
DISTANCE	-0.230	111.550	35.480	0.616	75.667					
TAX	-8.229	2397.942	831.713	13.021	1333.117	28348.624				
PTRATIO	0.068	15.905	5.681	0.047	8.743	167.821	4.678			
AVG_ROOM	0.056	-4.743	-1.884	-0.025	-1.281	-34.515	-0.540	0.493		
LSTAT	-0.883	120.838	29.522	0.488	30.325	653.421	5.771	-3.074	50.894	
AVG_PRICE	1.162	-97.396	-30.461	-0.455	-30.501	-724.820	-10.091	4.485	-48.352	84.420

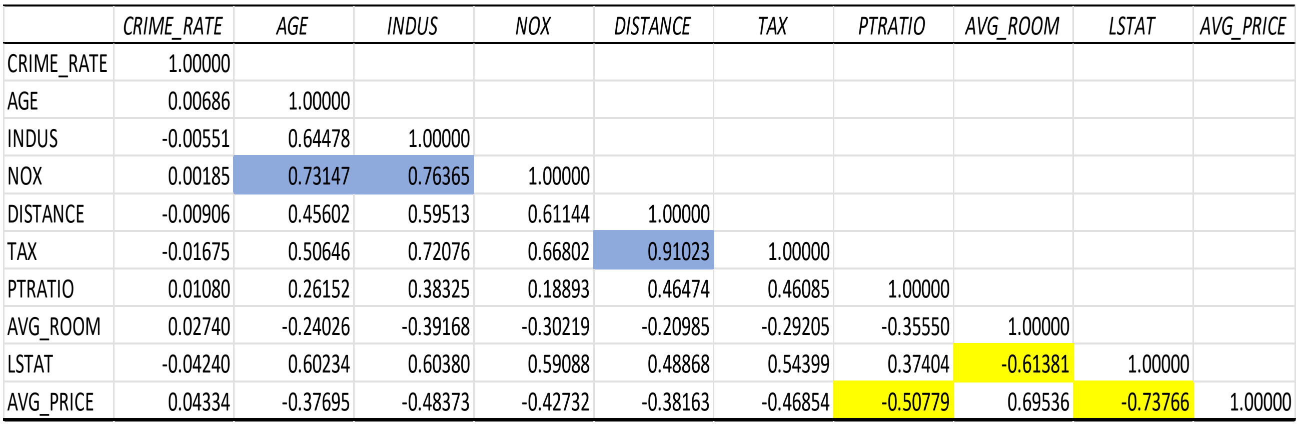
__from the above covariance table there are positive as well as negative covariance between two variable which also will have impact on the prices . however from the covariance table we can say that the tax has a high covariance with other variables except crime rates .hence it states that tax is a good variable with other features

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.

SOLUTION:



	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

From the above corelation table we can analyse that

a.)top 3 positively correlated pairs

1.The high positively correlated value is between distance and tax

2.The second highest positively correlated value is between indus and nox

3.The third highest positively correlated value is between age and nox

b.)top 3 negatively correlated pairs

1. The high negatively correlated value is between lstat and average price

2.The second highest negatively correlated value is between average room and lstat

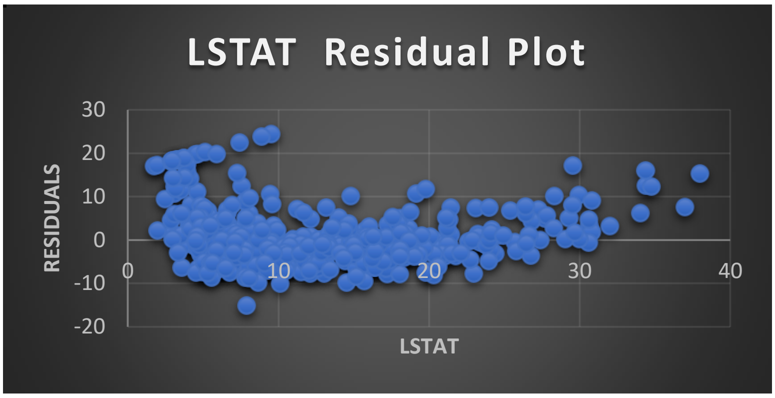
3.The third highest negatively correlated value is between ptratio and average price

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

Solution:

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

- The R-squared value in this regression model of LSTAT is 0.5441. it means of 54% of variance in dependent variable AVG_PRICE is explained by independent variable LSTAT in model
- The coefficient value of the variable LSTAT is -0.9500493
- The intercept of model LSTAT is 34.553840
- Residual plot



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

Yes, according to the regression model of LSTAT we can say that the LSTAT variable is significant for the analysis as the p-valve of the LSTAT model is less than 0.5 that is 5.08E-88

By this we can say the model is significant for analysis

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

Solution:

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?

- The regression equation is
  $Y=B_0+B_1X_1+B_2X_2.......B_n+X_n$
  Where y= average price
  X1=average room
  X2=LSTAT
  B_0=intercept(constant)
  Hence equation is,
  $Y= y = -1.358 +5.09\ X_1\ -0.642\ X_2$
- From the model, average price for the new model can be calculated as
  Y=-1.358+5.09(7)-0.642(20)
  =21.432 that is 21432USD
  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.

The adjusted R-square of previous model is 0.543241826

The adjusted R-square of this model is 0.637124

As the adjusted R-square is greater in this model compared to previous model it indicates that this model which includes average room and LSTAT explain the better variance in average price compared to previous model. Hence this model has a better performance than previous model which consider only LSTAT.

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.

Solution:

Regression Statistics	
Multiple R	0.832979
R Square	0.693854
Adjusted R Square	0.688299
Standard Error	5.134764
Observations	506

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	29.24131526	4.817125596	6.070283	2.54E-09
CRIME_RATE	0.048725141	0.078418647	0.621346	0.534657
AGE	0.032770689	0.013097814	2.501997	0.01267
INDUS	0.130551399	0.063117334	2.068392	0.039121
NOX	-10.3211828	3.894036256	-2.65051	0.008294
DISTANCE	0.261093575	0.067947067	3.842603	0.000138
TAX	-0.01440119	0.003905158	-3.68774	0.000251
PTRATIO	-1.074305348	0.133601722	-8.0411	6.59E-15
AVG_ROOM	4.125409152	0.442758999	9.317505	3.89E-19
LSTAT	-0.603486589	0.053081161	-11.3691	8.91E-27

- From the above details of model it suggest that all variable are significant predictor of dependent variables but not crime rate as the p- value of the crime rate is above 0.5 it is not a significant predictor for the dependent variable
- All the variable combinely explain 69% of variability for average price of a house
- Here nox, tax, ptratio, lstat has negative coefficient it says that increase in this variables will result in decrease in price of a house and decrease in variable will result in increase in price of a house .

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.

Solution:

a) Interpret the output of this model

R Square	0.693615426

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	29.42847349	4.804728624	6.124898157	1.85E-09
AGE	0.03293496	0.013087055	2.516605952	0.012163
INDUS	0.130710007	0.063077823	2.072202264	0.038762
NOX	-10.27270508	3.890849222	-2.640221837	0.008546
DISTANCE	0.261506423	0.067901841	3.851242024	0.000133
TAX	-0.014452345	0.003901877	-3.703946406	0.000236
PTRATIO	-1.071702473	0.133453529	-8.030529271	7.08E-15
AVG_ROOM	4.125468959	0.44248544	9.323400461	3.69E-19
LSTAT	-0.605159282	0.0529801	-11.42238841	5.42E-27

from the above data the R-square_explain 69% of variability for average price of a house

after excludeing the crime rate variable now the remaining variable in this model is a significant predictors for the dependent variable as the p-value of these variable are lesser than 0.5

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

Regression statistics of pervious model

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

Regression statistics for this model

Regression Statistics	
Multiple R	0.832835773
R Square	0.693615426
Adjusted R Square	0.688683682
Standard Error	5.131591113
Observations	506

According to the above details of the models the current models adjusted r square is slightly better than previous model . hence there is a small difference in value both the models performs better according to adjusted R squares

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?

Coefficient in ascending order:

	Coefficients
NOX	-10.2727
PTRATIO	-1.0717
LSTAT	-0.60516
TAX	-0.01445
AGE	0.032935
INDUS	0.13071
DISTANCE	0.261506
AVG_ROOM	4.125469
Intercept	29.42847

If the value of NOX is more in a locality in the town ,according to the coefficient of the NOC that is -10.2727 ,the price of the house in that locality will decrease by 10.27 units. It suggests that higher level of NOX will decrease the price of the house.

d) Write the regression equation from this model.

Equation: $y=B_0+B_1X_1+B_2X_2+B_3X_3+B_4X_4+B_5X_5+B_6X_6+B_7X_7+B_8X_8\ldots\ldots\ldots\ldots\ldots B_nX_n$

Where y=AVG_PRICE

B_0= intercept(constant)

X_1 =AGE

X_2 =INDUS

X_3=NOX

X_4=DISTANCE

X_5 =TAX

X_6 =PTRATIO

X_7 =AVG_ROOM

X_8 =LSTAT

Hence,

$$Y=29.42+0.032X_1+0.130X_2-10.27X_3+0.26X_4-0.014X_5-1.07X_6+4.12X_7-0.60X_8$$