BUISNESS REPORT 2023 REAL ESTATE

DATA ANALYSIS

TERRO'S REAL ESTATE AGENCY

ANALYSED BY
B. GUNAL

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

Summary statistics of crime rate

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

In the crime rate variables, the skewness is positively skewed. So, the most of the outliers falls on positive side.

The kurtosis is negative. So, this is platy Kurtic.

Summary statistics of distance

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

In the distance variable the skewness is positively skewed. So, the most of the houses have a distance of less than 9.54 kilo meters form the highway road.

The kurtosis is negative so this is a "platykurtic".

Summary statistics of Average room

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

\

Due to positive skewness most of the house should be less than 6 average rooms

The kurtosis is positive so this is a "Leptokurtic".

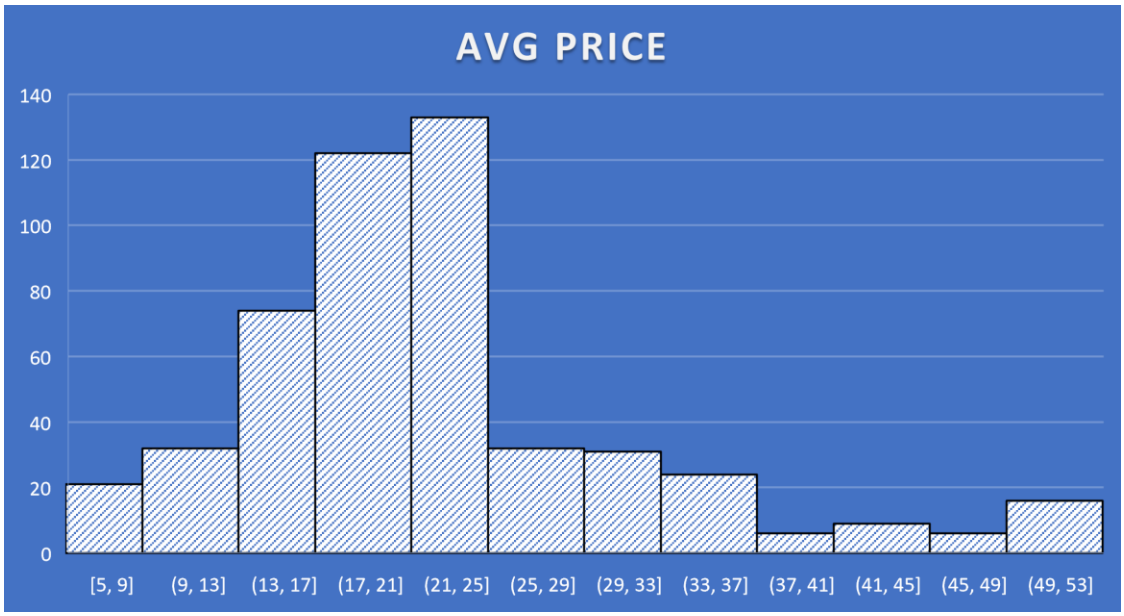
Summary statistics of LSTAT

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

Due to positively skewed the average lower status of the population must be less than 12.6%

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



From the above Histogram chart, we conclude that,

- The highest number of houses from a range of average price between 22 to 25 and in that range we have around 133 houses.
- The lowest number of houses from a range of average price between 38 to 49 and we have a smaller number of houses in that particular range.
- 10 to 13 and 26 to 29 range of average price have a same count of houses. Nearly 32 houses in that range.

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	-34.515100104	-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	-97.396152880	-30.460500499	-0.454512407	-30.500833035	-724.8204284	-10.090675561	4.484565552	-48.351779219	84.419556616

- I have computed the covariance matrix for all the variables and I observe that if we got a positive value then that is a directly proportional and if we got a negative value then that is an inversely proportional.

From the above covariance matrix table,

- Distance is inversely proportional to average price
- Average room is directly proportional to average price
- PTRATIO is directly proportional to LSTAT

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.

From the above correlation matrix table these are the top 3 positively correlated pairs

- 0.910228 positively correlated between Tax and Distance
- 0.763651 positively correlated between Indus and NOx
- 0.73147 positively correlated between Age and NOx

	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

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?

Regression Statistics	
Multiple R	0.737663
R Square	0.544146
Adjusted R Square	0.543242
Standard Error	6.21576
Observations	506

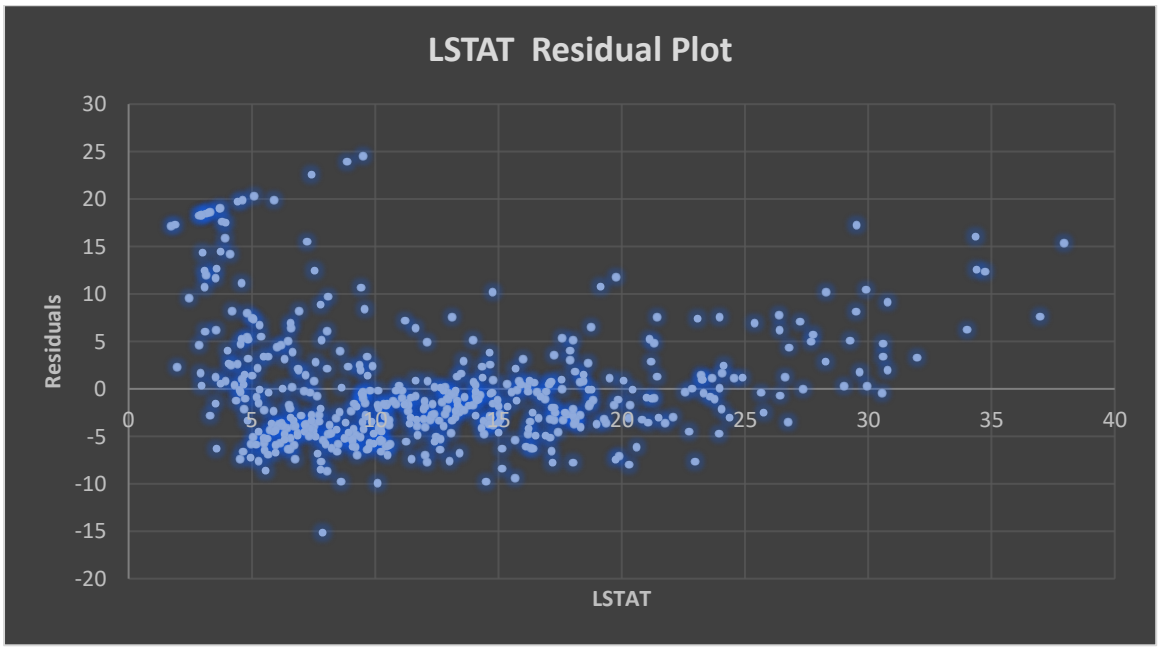
From the above regression summary output table

R square value is 0.54 and Adjusted R square value is 0.543

From the above table

LSTAT coefficient value is negative. So, if the LSTAT is increased then the Average price is decreased.

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



From the above chart we conclude that,

- After 25 and Below 5 the residual errors are to be Upper biased
- Between 5 to 25 the residual errors to be Biased and near to the linear line

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

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

Yes, LSTAT VARIABLE is significant for the analysis in this module.

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 statistics for previous model

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

	Coefficients
Intercept	-1.358272812
AVG_ROOM	5.094787984
LSTAT	-0.642358334

OBSERVATION

1.Regression equation:

Y = (-1.358272812) + 5.094787984*AVG_ROOM+(0.642358334) *LSTAT

2. Y = (-1.358272812) + 5.094787984*7+(-0.642358334) *20

AVG PRICE OF HOUSE	$1,000.00
AVG_ROOM	7
LTSAT	20
PREDICTING AVG_PRICE	$21,458.08

* The company has quoted a value of 30000 USD for this locality

* In this model, we used the regression equation to predict the AVG_PRICE for 7 rooms (on an average) and has a value of 20 for L-STAT, the value is $21,458.08, which the company is overcharging.

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.

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

Since the Adjusted R Square value is 0.688299 it is above 50%. So, we conclude that this Regression model to be a good one.

Intercept	29.24131526

Adjusted R Square	0.688298647

variable	Insignificant
CRIME_RATE	0.534657201

variable	*significant*
Intercept	2.53978E-09
AGE	0.012670437
INDUS	0.03912086
DISTANCE	0.008293859
TAX	0.000137546
NOX	0.000251247
PTRATIO	6.58642E-15
AVG_ROOM	3.89287E-19
LSTAT	8.91071E-27

variable	*significant*
Intercept	2.53978E-09
AGE	0.012670437
INDUS	0.03912086
DISTANCE	0.008293859
TAX	0.000137546
NOX	0.000251247
PTRATIO	6.58642E-15
AVG_ROOM	3.89287E-19
LSTAT	8.91071E-27

OBSERVATION

* In this model, adjusted R square is 68.82% which will give more impact to the AVG_PRICE.

*CRITERIA: In this model, the significance should be less than 0.05. And the insignificance should be greater than 0.05.

* Except CRIME_RATE(INSIGNIFICANCE), all other independent variable is significance.

* INDUS has more significance compared to other independent variables.

* LSTAT has less significance compared to other independent variables.

* In this model, AVG_ROOM has higher coefficients (4.125) compared to other variables which means AVG_ROOM has higher weightage for predicting 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:

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

B) OBSERVATION

In this model, Adjusted R square value is 68.86 %which gives more impact in AVG_PRICE predicting.

* In this model, R square value is 68.86% which gives more impact to AVG_PRICE compared to previous model, because in previous model more variables are taken for predicting the value of AVG_PRICE, but in this model less variable are taken.

* This model has better performance compared to previous model

variable	***Coefficients***
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

- The Coefficient of NOX value is negative, which tends to decrease in AVG_PRICE

- D)REGRESSION EQUATION

Y =INTERCEPT + 0.03*AGE+0.13*INDUS-10*NOX+0.26*DISTANCE *0.014*TAX+1.07*PTRATIO+4.12*AVG_ROOM-0.6*LSTAT

RECOMMENDATION:

- The most relevant price for the house should be lies between $21- 25k.

- AVG_PRICE has high positive correlation with the AVG_ROOM. so if the AVG_ROOM increases AVG_PRICE will increase.

  LSTAT has low negative correlation with AVG_PRICE

- REGRESSION EQUATION: Y = (-1.358272812) + 5.094787984*AVG_ROOM+(-0.642358334) *LSTAT

- So, if LSTAT decreases the AVG_PRICE will increase.

THANK YOU