BOE – WEEK 2 ASSIGNMENT

Terro's real estate agency

Objective (Task):

Your job, as an auditor, is to analyse the magnitude of each variable to which it can affect the price of a house in a particular locality.

To do the analysis, you are expected to solve these questions:

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

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

By taking descriptive Statistics (Summary Statistics) we can able to get the summary of the data given significant to the price of the house:

1) The average (Mean) price of the house is $ 22,532.
2) Median value (mid value) is $ 21,200.
3) Price range of the house lies between 5 to 50.
4) The Skewness is 1.108 which indicates it's a positive skewness.

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

1) The Average (Mean) of the average room is 6.284.
2) Median value is 6.208.
3) Mode (Most frequent value) of the Average_room is 5.71.

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

1) The Average (Mean) of the Tax is 408.27.
2) Median value is 330.
3) The tax range varies between 187 to 711.

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

1) The Average (Mean) of Age is 68.574.
2) Median value is 77.5.
3) The Skewness is -0.598 Negative skewness.

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

1) The average crime rate is 4.87.
2) Median value is 4.82.
3) The Mode value is 3.43.

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

Inference: By plotting a histogram for the avg_Price variable, we can stat that the average_Price of the house is Positively skewed, the most of the houses price ranges between 17 to 25**.**

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

Observations:

Positively related values

1) Crime_rate & Avg_Price 1.162.
2) Avg_Room & Avg_Price 4.484.

By taking Covariance matrix, we can observe that only these 2 relations mentioned above has positive relation with each other rest of the variable have negative relation with Avg_Price.

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

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

	Cells	**Top 3 +vely correlation**
1)	Tax & Distance	0.910228189
2)	NOX & Indus	0.763651447
3)	NOX & Age	0.731470104
	Cells	**Top 3 -vely Correlation**
1)	Avg_Price & Lstat	-0.737662726
2)	Lstat & Avg_Room	-0.613808272
3)	Avg_Price & PTRatio	-0.507786686

- The cells which are highlighted in Yellow are top 3 positively correlated.
- The cells which are highlighted in Green are top 3 negatively correlated.

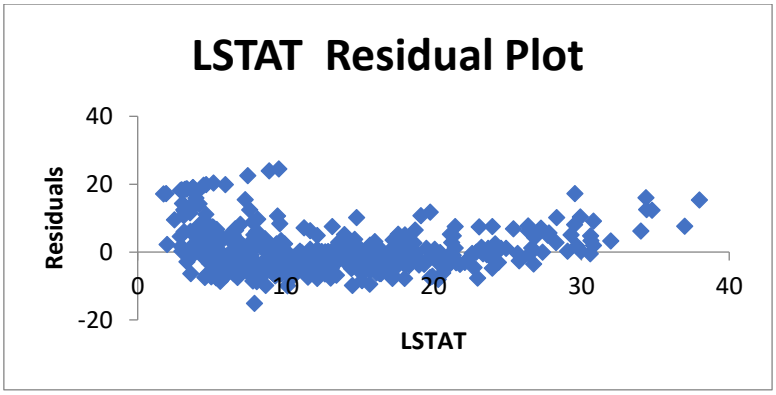
5) Build an initial regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot.

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

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

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

	Coefficients	*P-value*
Intercept	34.55384088	3.7431E-236
LSTAT	-0.95004935	5.0811E-88



5a)

- In regression statistics we get the Adjusted R Square value as 0.5432 which stats that the value is lesser than 1 but not closer to 1.
- The R Square value is 0.54 which stats that the variance is 54.5% in the Avg_Price.
- The LSTAT Residual plot is formed in a randomized manner and
- The Coefficient for the intercept is 34.55 and Lstat is -0.95
- The P-value for regression statistics for Avg_Price is less than 0.05 which stats that we can consider this as a pattern for further purpose.

5b)

- The significance of the LSTAT is near to zero and not 0, The P-value of the Lstat is less than 0.05.
- So, the LSTAT variable is significant and retained for the analysis.

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

	Coefficients	*P-value*
Intercept	-1.358272812	0.668764941
AVG_ROOM	5.094787984	3.47226E-27
LSTAT	-0.642358334	6.66937E-41

Regression Equation

Y = MX + C

y = M1X1 + M2X2 + c

y = (5.09 * X1) – (0.64 * X2) - 1.36

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

y = (5.09 * 7) – (0.64 * 20) - 1.36

Y = 35.63 – 12.8 – 1.36 = 21.47

The average price of the new House is **$21,470**

- The company coated price value is 30000, but the average price according to the predicted value is 21470.
- which shows a massive difference between prices.
- Therefore, The 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.

- The Adjusted R Square value is 0.6371.
- The Adjusted R Square value of the previous model is 0.5432.
- This shows that while adding Avg_Room in the model, there is a 10% variance in the Avg_Price.

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	0.832978824
R Square	0.69385372
Adjusted R Square	0.688298647
Standard Error	5.1347635
Observations	506

	Coefficients	*P-value*
Intercept	29.24131526	2.53978E-09
CRIME_RATE	0.048725141	0.534657201
AGE	0.032770689	0.012670437
INDUS	0.130551399	0.03912086
NOX	-10.3211828	0.008293859
DISTANCE	0.261093575	0.000137546
TAX	-0.01440119	0.000251247
PTRATIO	-1.074305348	6.58642E-15
AVG_ROOM	4.125409152	3.89287E-19
LSTAT	-0.603486589	8.91071E-27

- The Adjusted R Square value of this model has 0.6882
- The Adjusted R Square value of the previous model has 0.6371 with LSTAT and AVG_ Room.
- Comparatively this model has more variance to analyse the data and the adjusted r square of this model has more.
- The P-value of the Crime Rate is only more than 0.05, other than that every other p-values are significant.
- The intercept value is 29.241.

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

	Coefficients	*P-value*
Intercept	29.42847	1.85E-09
AGE	0.032935	0.012163
INDUS	0.13071	0.038762
NOX	-10.2727	0.008546
DISTANCE	0.261506	0.000133
TAX	-0.01445	0.000236
PTRATIO	-1.0717	7.08E-15
AVG_ROO	4.125469	3.69E-19
LSTAT	-0.60516	5.42E-27

a) Interpret the output of this model.

- The intercept value is 29.42
- The Adjusted R Square of this model is 0.688 with the variance of Avg_Price.
- All the variables in this model are significant. The Adjusted R Square is closer to 1 (With a decent).
- The final value of the Avg_house will be 29.42, when all the independent variable is 0.

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?

- The Adjusted R Square of this model is 0.6886 and for the previous model is 0.6882.
- Comparatively the value of the Adjusted R Square is just slightly up and not highly varied, and the significant variable stats to consideration with the P-value.
- On considering both the models, this model performs better according to the value of the Adjusted R Square.

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?

- While sorting the values of the coefficient in ascending order, The Nox & Avg_Price are negatively Related to each other.
- This states that is the value of the NOX Increases, then the Avg_Price of the house decreases.
- Every 1 unit of Nox value increases, the value of Avg_price decreases to 10.27.

d) Write the regression equation from this model.

- Y = 29.4285 + 0.0329 * X1 + 0.1307 * X2 + -10.2727 * X3 + 0.2615 * X4 – 0.0144 * X5 – 1.0717 * X6 + 4.1254 * X7 – 0.6051 * X8.
- Y = Interface + Age * X1 + Indus * X2 + NOX * X3 + Distance * X4 + Tax * X5 + PTRatio * X6 + Avg_Room * X7 + LSTAT * X8.