ASSIGNMENT-II

1) Generate the summary statistics for each variable in the table.

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

Kurtosis value is -1.1891, so the curve is not so sharp. It's look like a flat curve.

It has positive skewness.

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

Kurtosis value is -0.9677155, so the curve is not so sharp. It's look like a flat curve.

It has negative skewness.

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

Kurtosis value is -1.23353, so the curve is not so sharp. It's look like a flat curve.

It has positive skewness.

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

Kurtosis value is -0.064667, so the curve is not so sharp. It's look like a flat curve.

It has positive skewness.

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

Kurtosis value is -0.86723, so the curve is not so sharp. It's look like a flat curve.

It has positive skewness.

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

Kurtosis value is -1.1424, so the curve is not so sharp. It's look like a flat curve.

It has positive skewness.

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

Kurtosis value is -0.285091, so the curve is not so sharp. It's look like a flat curve.

It has negative 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

Kurtosis value is 1.891500, so the curve is not so sharp. It's look like a flat curve.

It has positive skewness.

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

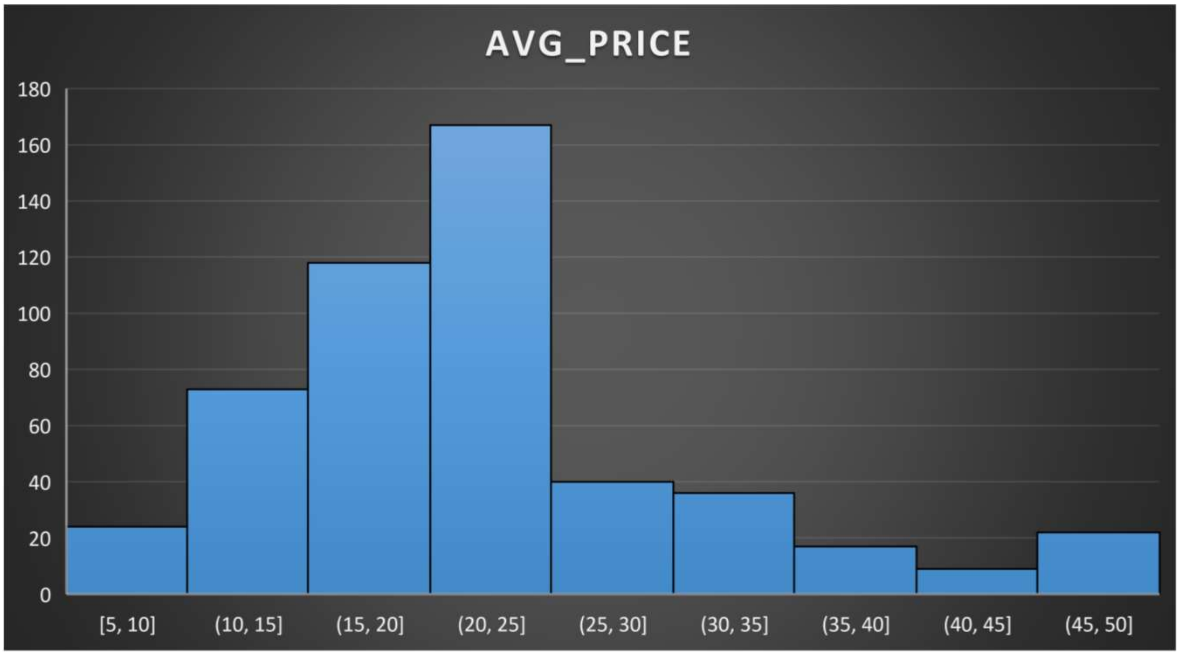
Kurtosis value is 0.493239, so the curve is not so sharp. It's look like a flat curve.

It has positive skewness.

Kurtosis value is 1.49519, so the curve is not so sharp. It's look like a flat curve.

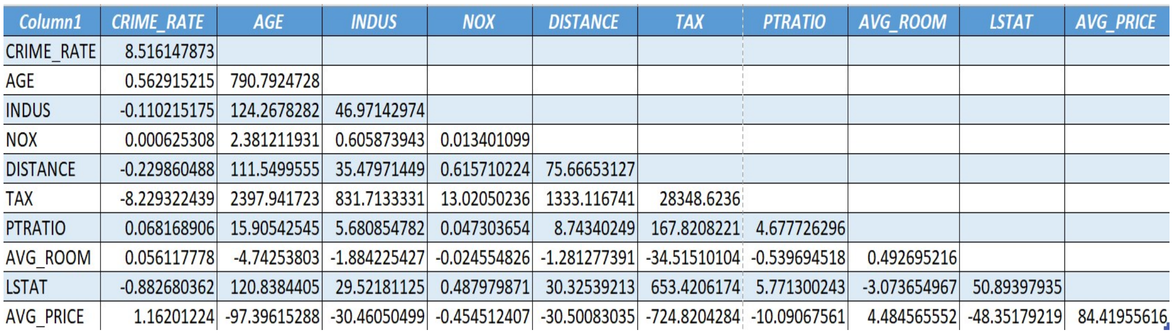
It has positive skewness.

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



From the Histogram, it is inferred that Average price has a positive skewness.

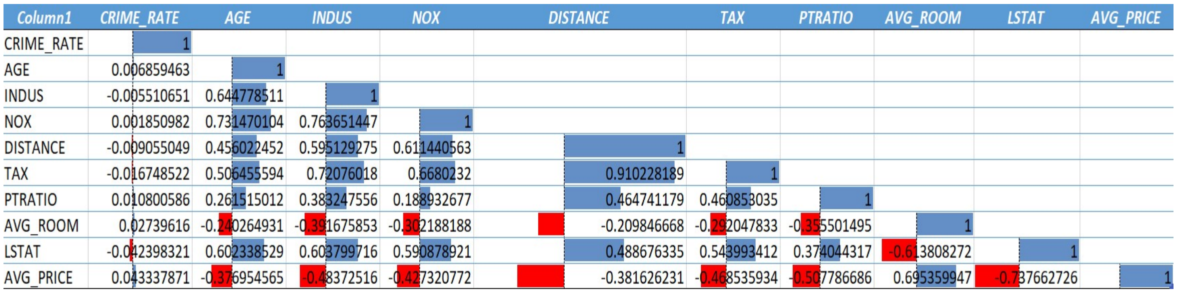
3) Compute the covariance matrix. Share your observations.



Column1	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

Positive value denotes, both the x and y values are above or below their averages.
Negative value denotes, both the x and y values are mostly on opposite sides of their averages.

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.



Column1	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

a)

Top 3 positively correlated pairs

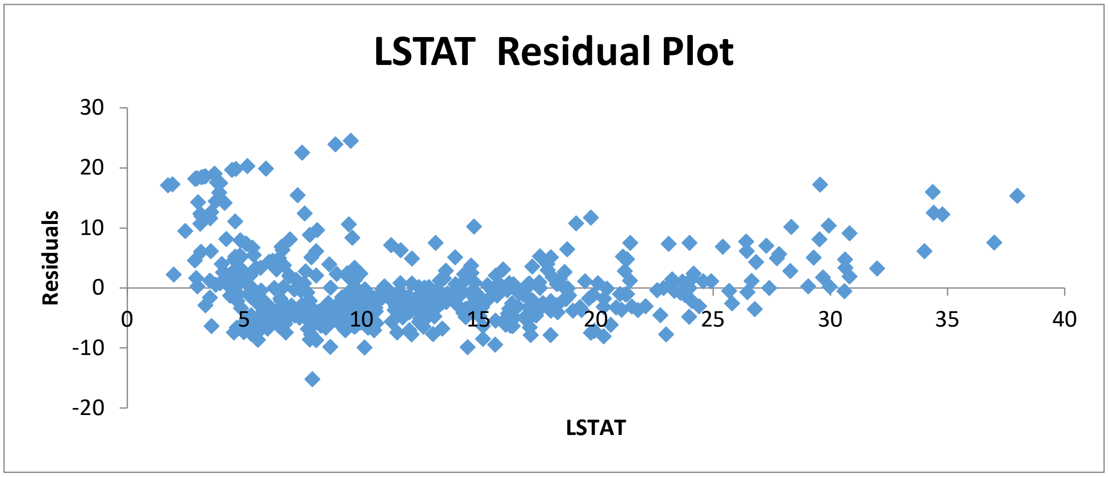
TAX vs DISTANCE	0.910228
NOX vs INDUS	0.763651
NOX vs AGE	0.73147

b)

Top 3 negatively correlated pairs

AVG_PRICE vs LSTAT	-0.73766
LSTAT vs AVG_ROOM	-0.61381
AVG_PRICE vs PTRATIO	-0.50779

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?



a)

R Square	0.544146
Coefficient of LSTAT	-0.95005
Intercept	34.55384

R square

R square is just above 0.5. So this value is not significant. R square has to be near to 1.

Coefficient of LSTAT

Coefficient of LSTAT is -0.95005. It is inferred that for each $1000 increase in Average price, there will be a 0.95% decrease in population.

Intercept

It is inferred that the Intercept value is 34.55384.

Residual plot

It is inferred that all the values are equally distributed.

b)

The p-value for LSTAT variable is 5.08110339438E-88. It is less than 0.05. So it is inferred that LSTAT variable is significant 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?
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.

a)
AVG_PRICE = Intercept + (Coefficient of AVG_ROOM * value of AVG_ROOM) + (Coefficient of LSTAT * value of LSTAT)

AVG_PRICE = -1.35827281187456 + (5.09478798433655 * 7) + (-0.642358334244129 * 20)
AVG_PRICE = 21.4581

It is inferred that the Average price is $21.4581. But the company quoting a value of 30000 USD for this locality. By the result, it is concluded that the company is overcharging.

Adjusted R Square = 0.637124475470123 (Qn. 6)

Adjusted R Square = 0.543241825954707 (Qn. 5)

b)
It is inferred that the value of R Square is close to 1, if the count of independent variable increases.
Based on the analysis, the performance of this model is better than the previous model.(Qn. 5)

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.

Adjusted R Square = 0.688298646855749

	Coefficients
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

** For every $1000 of avg. price of houses, per capita crime rate by town increases by 0.0487.
** For every $1000 of avg. price of houses, proportion of houses built prior to 1940 increases by 0.03%.
** For every $1000 of avg. price of houses, proportion of non-retail business acres per town increases by 0.13%.
** For every $1000 of avg. price of houses, nitric oxides concentration decreases by 10 million.
** For every $1000 of avg. price of houses, distance from highway increases by 0.2610 miles.
** For every $1000 of avg. price of houses, full-value property-tax rate decreases by 0.0144.
** For every $1000 of avg. price of houses, pupil-teacher ratio by town decreases by 1.0743.
** For every $1000 of avg. price of houses, average number of rooms per house increases by 4.12540.
** For every $1000 of avg. price of houses, lower status(LSTAT) of the population decreases by 0.603%.

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:

a) Interpret the output of this model.

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?

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?

d) Write the regression equation from this model.

a)

	Coefficients	*P-value*
Intercept	29.42847349	1.84597E-09
AGE	0.03293496	0.012162875
INDUS	0.130710007	0.038761669
NOX	-10.27270508	0.008545718
DISTANCE	0.261506423	0.000132887
TAX	-0.014452345	0.000236072
PTRATIO	-1.071702473	7.08251E-15
AVG_ROOM	4.125468959	3.68969E-19
LSTAT	-0.605159282	5.41844E-27

Adjusted R Square = 0.68868

b)

Adjusted R Square = 0.6886836818 (Qn.8)

Adjusted R Square = 0.6882986468 (Qn.7)

By the result, Adjusted R square for this model is greater comparing to the previous model.
So it is concluded that this model performs better than previous model.

c)

	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

It is inferred that if the value of NOX is more in a locality in this town, the value of the average price will be reduced.

d)

AVG_PRICE = Intercept + (coefficient of Age * value of Age) + (coefficient of Indus * value of Indus) +
(coefficient of NOX * value of NOX) + (coefficinet of Distance * value of Distance) +
(coefficient of Tax * value of Tax) + (coefficient of PTRATIO * value of PTRATIO) +
(coefficient of Avg_room * value of Avg_room) +
(coefficient of LSTAT * value of LSTAT)