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

OBSERVATIONS:

: CRIME_RATE: The per capita crime rate in town is measured in a scale of 0 to 10. In which some area had minimum crime rate of 0.04 and maximum of 9.99. And mean of the crime rate is 4.87197628458498 so the most of the values are centred around this value. And the 50% of the values are in between 0.04 and 4.82, and balance 50% having crime rate between 4.82 to 9.99. And the range and Standard Deviation shows there are some areas have high crime rate and some of them have low rate. And there is a small positive skewness of 0.02172

:. AGE: The proportion of houses built prior to 1940 (in percentage terms) have a minimum percentage of 2.9 and maximum of 100.And most of the values centred on a value of 68.5749011857. The 50% of values are in between 2.9 and 77.5 . And the standard deviation shows large deviation with mean. And the most of them are in age of 100(mode=100). The skewness shows a negative value -0.5989626.

INDUS: The proportion of non-retail business acres per town having a minimum percentage of 0.46 and maximum of 27.74. And there is a mean of 11.1367 and median 9.69, therefore 50% of value in between 0.46 and 9.69. And there is a deviation of 6.8603.

NOX: The fourth variable is nitric oxides concentration (parts per 10 million). The minimum NOX value is 0.385 and maximum value is 0.871. And it has an average of 0.5546. And there is a small standard deviation of 0.11587767. Almost 50% of the values are in between 0.385 and 0.538 and remaining values in the range of 0.538 to 0.871. And there is a small positive skewness.

DISTANCE: The distance from highway (in miles) varies greatly, from minimum of 1 mile to maximum of 24 miles. Almost 50% of values are in between 1 and 5. It shows most of the houses have less distance from highway. The standard deviation is 8.7 . And there is a positive skewness of 1.004.

TAX: The full-value property-tax rate per $10,000 is starting from value of 187 and have maximum value of 711. The 50% of values are in between 187 and 330, remaining value under 711. And most of the houses have tax rate of 666.There is a positive skewness of 0.6699.

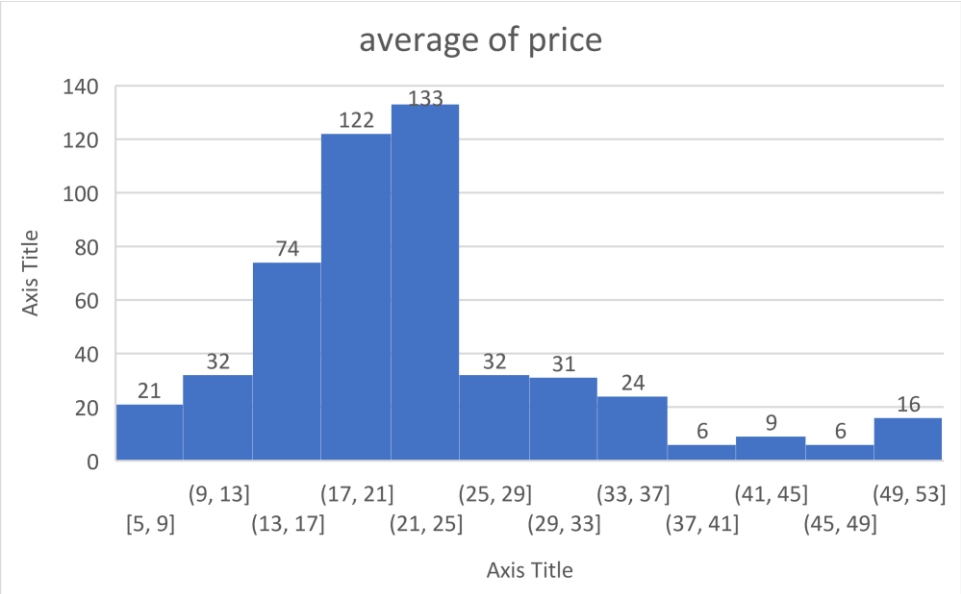
PTRATIO: It gives the pupil-teacher ratio by town. From this we can understand the education facilities in the locality. There are localities of PTRATIO of minimum 12.6 and maximum of 22. Most of the values are centred on 18.455. And 50% of the localities having PTRATIO of less than 19.5. Which means most of the areas have PTRATIO above average. And there is a negative skewness of -0.8.

AVG_ROOM: It gives the details about number of rooms in a house. The average number of rooms per house varies from 3 to 8. Which means there are houses having 3 rooms and houses having maximum of 8 rooms. Most of the house contain 6 rooms. 50% of houses have below 6 rooms. And there is positive skewness.

LSTST: Percentage of lower status of the population. It have a minimum value of 1.73 and maximum value of 37.97. Most of the values are centred on 12.65. And 50% of areas have percentage of lower status of population as value below 11.36. There is a skewness of 0.9064. Most of the areas have LSTAT value as 8.05.

AVG_PRICE: There are houses of average price of 5 (minimum) and average price of 50(maximum). Which means there are cheaper and expensive houses. More values are centred at 22.53. Almost 50% of houses have rate of below 21.2 . There is a deviation of9.197. And there is high positive skewness of 1.108

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



Observation:

1)The most land prize in between 21-25,i.e 133 houses.
2)The lowest land price between 31-31 and 45-49 ,I.e 6 houses .
3)It is a positive skewness.

3) Compute the covariance matrix. Share your observation.

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	VG_ROOM	LSTAT	VG_PRICE
CRIME_RATE	8.516147873									
AGE	0.562915215	790.792								
INDUS	-0.11021518	124.268	46.9714							
NOX	0.000625308	2.38121	0.60587	0.0134						
DISTANCE	-0.22986049	111.55	35.4797	0.61571	75.6665					
TAX	-8.22932244	2397.94	831.713	13.0205	1333.12	28348.6				
PTRATIO	0.068168906	15.9054	5.68085	0.0473	8.7434	167.821	4.67773			
AVG_ROOM	0.056117778	-4.74254	-1.88423	-0.02455	-1.28128	-34.5151	-0.53969	0.4927		
LSTAT	-0.88268036	120.838	29.5218	0.48798	30.3254	653.421	5.7713	-3.07365	50.894	
AVG PRICE	1.16201224	-97.3962	-30.4605	-0.45451	-30.5008	-724.82	-10.0907	4.48457	-48.3518	84.4196

Variables having top covariance:

- 28348.623599 is the covariance coefficients for the TAX variable when predicting the TAX.
- 2397.94172 is the coefficients of covariance for the AGE variable when predicting the TAX.
- 1333.11674 is the coefficients of covariance for the DISTANCE variable when predicting the TAX.

Variables having less covariance:

- 724.8204 is the covariance coefficients for the TAX variable when predicting the AVG_PRICE.

-97.396152is the covariance coefficients for the AGE variable when predicting the AVG_PRICE.

--48.35179 is the covariance coefficients for the LSTAT variable when predicting the AV_AVG.

4) Create a correlation matrix of all the variables?

a) Which are the top 3 positively correlated pairs.

b) Which are the top 3 negatively correlated pairs.

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM
CRIME_RATE	1							
AGE	0.006859	1						
INDUS	-0.00551	0.644779	1					
NOX	0.001851	0.73147	0.763651	1				
DISTANCE	-0.00906	0.456022	0.595129	0.611440563	1			
TAX	-0.01675	0.506456	0.72076	0.6680232	0.910228	1		
PTRATIO	0.010801	0.261515	0.383248	0.188932677	0.464741	0.460853	1	
AVG_ROOM	0.027396	-0.24026	-0.39168	-0.302188188	-0.20985	-0.29205	-0.3555	1
LSTAT	-0.0424	0.602339	0.6038	0.590878921	0.488676	0.543993	0.374044	-0.61381
AVG_PRICE	0.043338	-0.37695	-0.48373	-0.427320772	-0.38163	-0.46854	-0.50779	0.69536

a)Which are the top 3 positively correlate pairs .

Top positive:

1-distance-tax (0.91)

2-index – Nox (0.76)

3-age – Nox (0.73)

b)Which are top 3 positively correlate pairs.

Top negative:

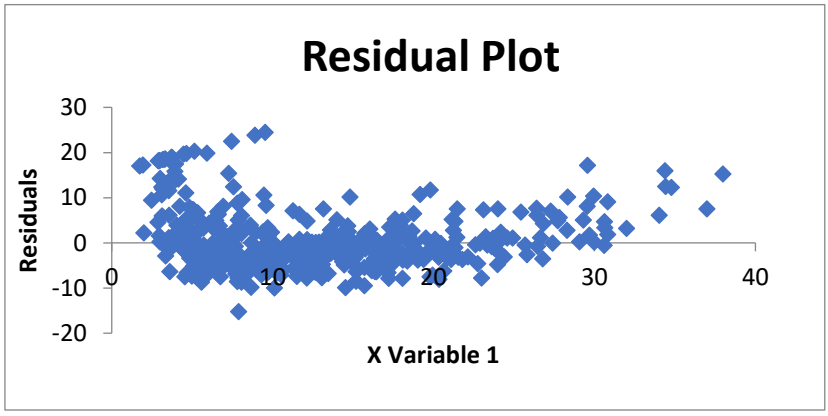
1-lstat – Average price (-73)

2-avg room – Lstat (-61)

3-pratario – Average price (-50)

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?



*The R-squared value is approximately 0.544,indicating that the regression model explains about 54.4% of the
variance in the dependent variable.

*The coefficient value for the LSTAT variable is -0.95. This coefficient represents the change in the dependent
variable for a one-unit change in the LSTAT variable .

* Here the coefficient of LSTAT is negative , so as the LSTAT value increases, the predicted value of the dependent
variable decreases. ϖ The intercept value is 34.5538 .

*The intercept represents the predicted value of the dependent variable when all independent variables are set to zero.

* The p-value is 5.08110339438785E-88. It is a very low p-value ,shows that the regression model is statistically significant.

6) Build a new Regression model including LSTAT and AVG_ROOM together as Indepen 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.

(Average rooms * 7+ LSTAT *20) + intercept
=21.45807639

Intercept	-1.35827	*7
AVG_ROOM	5.094788	*20
LSTAT	-0.64236	

By this we can assuming its over charging.
* The performance of this model is better than previous model because R-square value is high and this model is fit compare to previous model the R-Square value is lessar than 60% and that model is not fit.

* And this model is fit because grater than 60%. In this regression analysis the adjusted R-square value is 0.63712.
Means that around 63.71% of the variability in the dependent variable can be explained by the independent
variables in the model.

*When we compare with adjusted R square of previous model, it's R square value 0.5432418(54.3%).Which is less than value of this model.

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 ?

* The value of adjusted R-square is 0.688, means that about 68.8% of the variability in the average housing price can be explained by the dependent variables included in the model. This indicates a reasonably good fit of the model to the data.
* For each unit increase in the CRIME_RATE variable the average housing price will increase by approximately 0.04872514.
* For each unit increase in the LSTAT variable the average housing price will decrease by approximately 0.603486589.
* For each unit increase in the DISTANCE variable the average housing price will increase by approximately 0.26109357.
* For each unit increase in the DISTANCE variable the average housing price will increase by approximately 0.26109357.

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

b) 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) In this regression model all variables are significant because the P-Value of these variables is

less than 0.05 as shown in figure below.

b) Adjusted R Square:

*In this model the Adjusted R-square value is 0.68 and it is greater than 0.6(60%) so we consider that the model is fit.

compare the r square value.

**Current model:

Adjusted R Square 0.688684

**Previous model:

Adjusted R Square 0.688299

c)

	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

*When sorting the data it is clear that NOX has greater negative impact in average housing price. In this case NOX have negative coefficient, which indicates a negative relationship with AVG_PRICE. So as the NOX value increases the value of AVG_PRICE decreases by approximately 10.2727 units.

D) Regration equation:

AVG_PRICE =29.42847349+0.03293496*AGE+0.130710007*INDUS10.27270508*NOX+0.261506423*

DISTANCE- 0.014452345*TAX1.071702473*PTRATIO+4.125468961*AVG_ROOM-0.605159282*LSTA