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

Observations:-

1. THE “CRIME _RATE” variable has a wide range. Some values are very close to 0 and some other values are high as 100. The crime rate range was 9.99.
2. The “AGE” variable also shows a wide range. The minimum age 2.9 to maximum age is 100 to indicating the diversity ages in the dataset.
3. The “INDUS” variable has a moderate mean and standard deviation. The mean of this data is 11.1368 and range is 27.28. This suggesting moderate variability in the proportion of non-retail business acres per town.
4. The “NOX” variable representing nitric oxides concentration has relatively narrow range with a mean close to 0.5547 and standard deviation is close to 0 and the range of NOX is also close to 0.
5. The “DISTANCE” variable appears to be representing distance to various locations and has a mean around 9.549.
6. The “AVERAGE PRICE” variable, the total rage of average price is 45 and the values are close to 5 and has high is 50. The total kurtosis is 1.495 and Skewness is 1.108 of given dataset.
7. Considering the “TAX” variable the average tax is 408.2 and range of tax is 524.
8. The “Avg_room” variable representing the average number of rooms per dwelling shows that the maximum is 8.78.

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

Observations:-

1. In this histogram of the AVG_PRICE the first variable bin range of (5,9) which is 21.
2. In Histogram We can understand the high number of variables are in the (2,25) which is 133.
3. In Histogram we can understand the less number of variables are in the (37,41) and (45,49) which is 6.

Q.3. Compute the covariance matrix. Share your observations.

Observations:-

1. In the covariance we can understand about the positive and the negative relation of the variables. For the given data, covariance formula is applied to all the variables.
2. If we get the positive then it is positive relationship with variable. And the highest positive variable is 28348.62(TAX).
3. If we get the negative then it is negative relationship with variable. And the lowest negative variable is -724.82.

Q.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.

Observations:-

1. In the correlation chat we can understand the relationship of the variable. It shows the how independent variable have significant impact on the dependent variable.
2. In this correlation chat the top three positive correlation variables are:
    a. Distance & Tax - 0.910228
    b. Indus & NOX - 0.763651
    c. Age & KNOX - 0.73147
3. In this correlation chat the top three negative correlation variables are:
    a. Lstat & Avg_price – (-0.73766)
    b. Avg_price & Lstat – (-0.61381)
    c. Ptration & Avg_price – (-0.50779)

4. The correlation measures the linear relationship between two variables.

Q.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 Residual plot?**
b) **Is LSTAT variable significant for the analysis based on your model?**

Observations:-

a).

1. I observed that, the R-square and adjusted R-square are having similar values.
2. Variance explained the value indicates the proportion of the variance in the dependent variable that's explained by the independent variable. A higher R-squared suggest that the "LSTAT" variable is better at explaining the variability in "Avg_price".
3. Intercept is the predicted average price when the "LSTAT" percentage is 0. In the most cases, this value might not have a meaningful interpretation.
4. If coefficient is negative, it suggests that as the percentage of lower status population increases, the average price tends to decrease.
5. The coefficient for the "LSTAT" variable indicates the change in the predicted average price for each one-unit change in the "LSTAT" variable.

b).

Yes, LSTAT is significant variable for the AVG_PRICE from the model, as the P-value (5.0811E-88) is less than 0.05. So, we can say that LSTAT is a significant variable according to the model.

Q.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

Observations:-

A). 1. Y=M1X1+M2X2+M3X3+……+MnXn+C

where, y=avg_price

x0= Avg_room

x1=Avg_LSTAT

Y= 5.094787984+(-.642358334)*20+(-1.35827812)

Y=21.45807639, Y=$21,450

The company quoting $30000

The company is over charging the customer.

B).

1. Here, R-square and Adjusted R-square is having similar values.

2. Yes, the model performance is better than previous model because the adjusted R-square value is greater in this model. The adjusted R-value of this model is 0.637124475. The adjusted R value of the previous model is 0.54.

Q.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.

Observations :-

1. From the regression model the crime rate is not significant variable for average price of the house as P- value is greater than 0.05.
2. NOX,TAX,PTRATION and LSTAT have negative coefficient which says that increase in these features results in decrease I price of the house.

Q.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.

Observations:-

a). the output of thus are adjusted R- value is 0.688683682

The intercept value is 29.42847349

The significant F(PVALUE) is 1.911E-122

So, this model is linear equation because the P value is less than 0.05(p value <0.05)

b). The adjusted value R value of previous model is 0.688298647

The adjusted value R value of is model is 0.688683682

The value of the previous model and present model are almost similar.

c). The coefficient of independent variable is ascending order are

1) TAX -0.01445
2) PTRATION -1.0717
3) NOX -10.2727
4) LSTAT -0.60516
5) INDUS 0.13071
6) DISTANCE 0.261506
7) AVG_ ROOM 4.125469
8) AGE 0.032935

If the value of NOX increases on the locality the price of AVG_Price of a property decreases because NOX have negative relationship with the AVG_Price.

d). The regression equation from this model is :

Y= M1X1+M2X2+M3X3+M4X4+M5X5+M6X6+M7X7+C

This is the regression equation of this model.