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

|  |  |
| --- | --- |
| *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 |

* Descriptive statistics are used to describe or summarize data

In ways that are meaningful and useful.

* The Average of crime rate is 4.87.
* The standard error is dividing the standard deviation by the

Sample size square root, the standard deviation for crime rate

is 0.12986.

* The middle most value of Crime rate is 4.82.
* The most repeated in crime rate is 3.43.
* The Standard deviation of crime rate is 2.92 that is measure

Of how dispersed the data is in relation to the mean.

* The sample variance of crime rate is 8.55 that square of the

Difference from observation to the mean value

* The kurtosis value is -1.18 that is, it is a negative kurtosis

Having a flat peak so, it is called as “platykurtic”.

* The skewness for crime rate is positive that is 0.021 so it is right Skew.
* The range of Crime rate is 9.95 that is highest minus lowest value.
* The minimum crime rate is 0.04.
* The maximum crime rate is 9.99.
* Total sum of the crime rate is 2465.22.
* The number of crime rate is 506.
* The Mean value for Age is 68.57.
* The standard Error for age is 1.25.
* The middle most value of age variable is 77.5.
* The most repeated Age is 100.
* The standard deviation for Age is High comparing to the Crime rate that is the 28.14.

|  |  |
| --- | --- |
| *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 |

* The variance for the Age that is the more that is higher the variance higher the risk and the value of variance for age is 792.35.
* The kurtosis value is -0.96 that is, it is a negative kurtosis having a flat peak so it is called as platykurtic.
* The skewness for age is negative that is -0.59 so it is left skew.
* The range for Age variable is 97.1.
* The minimum value for age is 2.9.
* The maximum value for age is 100.
* The sum of the age is 34698.

|  |  |
| --- | --- |
| *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 |

* The kurtosis value is 1.2 that is a positive kurtosis haiving a sharp peak and it is called leptokutic.
* The skewness for the Indus is positive that is 0.29 so it is right skew.

|  |  |
| --- | --- |
| *NOX* | * The kurtosis value is 0.064 that is positive kurtosis having a sharp peak and it is called leaptokutic * The skewness for the Nox is positive that is 0.72 so it is right skew. |
|  |  |
| 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 |

|  |  |
| --- | --- |
| *DISTANCE* | * The kurtosis value is 0.86 that is positive having a sharp peak and it is called as leptokurtic. * Distance is positive and the extreamly skewed and the skews valu is greater than 1. |
|  |  |
| 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 |

|  |  |
| --- | --- |
| *TAX* | * The skewness value is 0.66 that is positive and it right skew. |
|  |  |
| 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 |

|  |  |
| --- | --- |
| *PTRATIO* | * P ratio is negative and right skewed. |
|  |  |
| 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 |

|  |  |
| --- | --- |
| *AVG\_ROOM* | * The avg room is symmetrical skewness. |
|  |  |
| 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 |

|  |  |
| --- | --- |
| *LSTAT* | * The kurtosis value is positive it has sharp peak and it is called as leaptokurtic. |
|  |  |
| 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 |

**2) Plot a histogram of the Avg\_Price variable. What do you infer?**

* A histogram shows how frequently a value falls into a particular bin
* From the above histogram plot we can say that there are only two houses where average price is in-between the 0-5.
* The highest average price is 40-50 where there are 22 houses.
* The most count of the house lies in between the average price 20-25,

When comparing with other average price bin.

* There is no house where the average price is more than 50.
* The peak stage of average price is in between 20-25, with increase in average price there is less count of house are left after the peak stage.
* With increasing the average price the number house cont increasing until the average price is 25, but with average price increasing there are decreasing in the count of the house as shown in the above histogram plot
* There are more than 50 percent of the houses lies where the average is less than 25.
* From above graph plot we can say that with less average price and highest the average prices there are less count of houses present.
* The second highest count of houses lies when the average price is in between 15-20.
* There is massive decrease in the count of the house when the average price is more than 25.

**3) Compute the covariance matrix. Share your observations.**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *CRIME\_RATE* | *AGE* | *INDUS* | *NOX* | *DISTANCE* | *TAX* | *PTRATIO* | *AVG\_ROOM* | *LSTAT* | *AVG\_PRICE* |
| CRIME\_RATE | 8.516148 |  |  |  |  |  |  |  |  |  |
| AGE | 0.562915 | 790.7925 |  |  |  |  |  |  |  |  |
| INDUS | -0.11022 | 124.2678 | 46.97143 |  |  |  |  |  |  |  |
| NOX | 0.000625 | 2.381212 | 0.605874 | 0.013401 |  |  |  |  |  |  |
| DISTANCE | -0.22986 | 111.55 | 35.47971 | 0.61571 | 75.66653 |  |  |  |  |  |
| TAX | -8.22932 | 2397.942 | 831.7133 | 13.0205 | 1333.117 | 28348.62 |  |  |  |  |
| PTRATIO | 0.068169 | 15.90543 | 5.680855 | 0.047304 | 8.743402 | 167.8208 | 4.677726 |  |  |  |
| AVG\_ROOM | 0.056118 | -4.74254 | -1.88423 | -0.02455 | -1.28128 | -34.5151 | -0.53969 | 0.492695 |  |  |
| LSTAT | -0.88268 | 120.8384 | 29.52181 | 0.48798 | 30.32539 | 653.4206 | 5.7713 | -3.07365 | 50.89398 |  |
| AVG\_PRICE | 1.162012 | -97.3962 | -30.4605 | -0.45451 | -30.5008 | -724.82 | -10.0907 | 4.484566 | -48.3518 | 84.41956 |

COVAIANCE MATRIX

* Covariance matrix say’s that weather there is positive relation between the two variables.
* The covariance is positive if x and y values are mostly both above or both below the averages
* With knowing weather there is positive or negative related between the two variables we can give a good amount of information to the agency.
* From the above covariance matrix the value which are greater than the zero is said to be positive(+ve) relationship where the value is less than zero than it is said to be negative relationship between the two variables.
* So from above covariance matrix we can say that the colored cell are all positive in relationship that is with increasing the x there is increasing in y or with increasing in y there is increasing in x.
* Let us see the relation between the crime rate and age, the relation between this two variables there is positive relation, the value is 0.56 where we can say that with increasing in crime rate there is increasing in Age.
* Let us see another positive relation in between two variables that is Age and distance. From this we can say that with increasing in Age of the house there will be far distance mile from the highway.
* we can say that with more the distance from the highway to the house than the age of the house will be more.
* There is positive relation with NOX and other variables that is crime rate, age, Indus, that is with increasing in nitric oxides concentration then there will increase in crime rate, age of the house, there will be more number of industries.
* Similarly as with other variables that is colored one are positive related.
* The negative values represent that with increasing in one variable the other variable will decrease.
* That is crime rate and Indus are negative related where with increasing in crime rate there will be decreasing in industry.
* Similarly with distance and crime rate with increasing in crime rate there will be decrease in distance from the high.

**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**.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *CRIME\_RATE* | *AGE* | *INDUS* | *NOX* | *DISTANCE* | *TAX* | *PTRATIO* | *AVG\_ROOM* | *LSTAT* | *AVG\_PRICE* |
| 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 | 0.611441 | 1 |  |  |  |  |  |
| TAX | -0.01675 | 0.506456 | 0.72076 | 0.668023 | 0.910228 | 1 |  |  |  |  |
| PTRATIO | 0.010801 | 0.261515 | 0.383248 | 0.188933 | 0.464741 | 0.460853 | 1 |  |  |  |
| AVG\_ROOM | 0.027396 | -0.24026 | -0.39168 | -0.30219 | -0.20985 | -0.29205 | -0.3555 | 1 |  |  |
| LSTAT | -0.0424 | 0.602339 | 0.6038 | 0.590879 | 0.488676 | 0.543993 | 0.374044 | -0.61381 | 1 |  |
| AVG\_PRICE | 0.043338 | -0.37695 | -0.48373 | -0.42732 | -0.38163 | -0.46854 | -0.50779 | 0.69536 | -0.73766 | 1 |

CORRELATION MATRIX

* From the above correlation matrix if the number is near to 1 then we can say that there is strong correlation.
* So from the above correlation matrix the strong top three correlation between the two variables are 1)(Tax,distance),2)(Nox,age)3)(Nox,indus)
* The top three positively correlated pairs values are

1)Tax and distance value is 0.910228 which is close to 1 value so they are

stronger correlated

2)Nox and age value is 0.73147 this is strong positively correlated.

3)Nox and Indus value is 0.763651.

* The top three negatively correlated pairs are

1)Average price and Lstat.

2)Lstat and Avg\_room.

3)Avg\_price and PTratio.

**5) Build an initial regression model with AVG\_PRICE as ‘y’ (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot. (8 marks)**

**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 |

* From the above regression statistics the multiple R value is correlation where which is close to 1 so we can say that the correlation of Lstat and Avg\_prive is very strong.
* R Square tells that 54% of the AVG\_Price variable was explained by the LSTAT in other words we can say that the variable Lstat explains theh 54% of the variability of AVG\_PRICE.

|  |  |
| --- | --- |
|  | *Coefficients* |
| Intercept | 34.55384088 |
| LSTAT | -0.950049354 |

* Coefficient value for Lstat is -0.950049354, where with every unit increase in LSTAT the AVG price is decreasing -0.9500499354.
* When the Percentage of lower status of the population is zero the minimum AVG\_PRICE is 34.55.
* From the Residuals plot the errors are in random then it is called as homoschedasticity.

**B)**According to my analysis LSTAT variable is significant in this model because the P-value is less than 0.05.

**6) Build a new Regression model including LSTAT and AVG\_ROOM together as Independent variables and AVG\_PRICE as dependent variable. (6 marks)**

**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.**

* Given that x1 and X2 values are 7 and 20 that is AVG\_rooms is 7 and L-STAT value is 20 and we got the coefficients of LSTAT and AVG\_rooom values from the regression model in excel that is -0.64,5.09 and intercept value is -1.35.
* We know that regression equation for multiple variables are Y=B0+B1X1+B2X2
* The equation is Y=-1.35+(-0.64\*7)+(5.09\*20)=21.45\*83=1781
* From the above value comparing to 30000USD we can say that the company is overcharging.

**B)**

|  |  |
| --- | --- |
| *Regression Statistics*   * Adjested R square Normalise the number of variables in the regression equation is using when there is two or more independent variables are there.Adjusted R square is very useful when there is two or more independent variables are there. The greater the adjusted R square the greater is the model. * When comparing the two adjusted in 5),6) Regression model the adjusted R square is more in the 6) Regression model so, this model is better than the previous model. | |
| Multiple R | 0.737662726 |
| R Square | 0.544146298 |
| Adjusted R Square | 0.543241826 |
| Standard Error | 6.215760405 |
| Observations | 506 |

**5)Regression model**

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

**6) Regression 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 Rsquare, 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 |

* From the above regression statistics he multiple of R is 0.83 so, correlation of the dependent variable and independent variable is strong.
* The adjusted R square is more when comparing with the other models so this regression model performance is more.

|  |  |
| --- | --- |
|  | *Coefficients*   * With every unit increase in the crime rate the avg\_price is increased by 0.04. * With every unit increase in the age the avg price is increases by 0.03. * With ever unit increase in the Indus the avg price is increased by 0.13. * With every unit increase in the nox the avg price is decreased by -10.32 |
| Intercept | 29.24132 |
| CRIME\_RATE | 0.048725 |
| AGE | 0.032771 |
| INDUS | 0.130551 |
| NOX | -10.3212 |
| DISTANCE | 0.261094 |
| TAX | -0.0144 |
| PTRATIO | -1.07431 |
| AVG\_ROOM | 4.125409 |
| LSTAT | -0.60349 |

* When the depended variables is zero the minimum AVG\_PRICE is29.24.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |  |
| Intercept | 29.24132 | 4.817126 | 6.070283 | 2.54E-09 |  |
| CRIME\_RATE | 0.048725 | 0.078419 | 0.621346 | 0.534657 |  |
| AGE | 0.032771 | 0.013098 | 2.501997 | 0.01267 |  |
| INDUS | 0.130551 | 0.063117 | 2.068392 | 0.039121 |  |
| NOX | -10.3212 | 3.894036 | -2.65051 | 0.008294 |  |
| DISTANCE | 0.261094 | 0.067947 | 3.842603 | 0.000138 |  |
| TAX | -0.0144 | 0.003905 | -3.68774 | 0.000251 |  |
| PTRATIO | -1.07431 | 0.133602 | -8.0411 | 6.59E-15 |  |
| AVG\_ROOM | 4.125409 | 0.442759 | 9.317505 | 3.89E-19 |  |
| LSTAT | -0.60349 | 0.053081 | -11.3691 | 8.91E-27 |  |

* From the above regression model we can the say that when the p value is less than the 0.05 then the dependent variable is significant of Avg price.
* So from the above table we can say that p-value is less for crime rate so we can say that crime rate is not significant variable of avg price.
* All the other variables of p-value is less than 0.05 so we can say that the variables are significant to the 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: (8 marks) 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.**

|  |  |
| --- | --- |
| *Regression Statistics* | |
| Multiple R | 0.832836 |
| R Square | 0.693615 |
| Adjusted R Square | 0.688684 |
| Standard Error | 5.131591 |
| Observations | 506 |

* From the above regression statistics the multiple of R value is closer to 1 so, the independent variable has strong relation with the Avg price.
* 69% of avg price variables are explained by the dependent variable.
* The adjusted R square form this model is more so we can say that this model performance is great.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 29.42847 | 4.804729 | 6.124898 | 1.85E-09 |
| AGE | 0.032935 | 0.013087 | 2.516606 | 0.012163 |
| INDUS | 0.13071 | 0.063078 | 2.072202 | 0.038762 |
| NOX | -10.2727 | 3.890849 | -2.64022 | 0.008546 |
| DISTANCE | 0.261506 | 0.067902 | 3.851242 | 0.000133 |
| TAX | -0.01445 | 0.003902 | -3.70395 | 0.000236 |
| PTRATIO | -1.0717 | 0.133454 | -8.03053 | 7.08E-15 |
| AVG\_ROOM | 4.125469 | 0.442485 | 9.3234 | 3.69E-19 |
| LSTAT | -0.60516 | 0.05298 | -11.4224 | 5.42E-27 |

* With every unit increase in the age the avg price is increased by 0.03.
* With every unit increase in the indus the avg price is increased by 0.13.
* With every unit increase in the nox the avg price is decreased by -10.27.
* With every unit increase in the distance the avg price is increased by 0.26.
* With every unit increase in the tax the avg price is decreased by -0.041.
* With every unit increase in the PTRATIO the avg price is decreased by -1.07.
* From the above table we can say that every independent variables are significant to avg price because the p-value is less than 0.05.

**B)**

|  |  |
| --- | --- |
|  | |
|  |  |
| *Regression Statistics*   * Comparing the two Adjusted R square for the 8) and 7) this adjusted R square is nearly close to each other. so both the models performance equally. | |
| Multiple R | 0.832836 |
| R Square | 0.693615 |
| Adjusted R Square | 0.688684 |
| Standard Error | 5.131591 |
| Observations | 506 |

8) regression statistics

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

7)regression statistics

**C)**

* If the NOX is more in a locality then there With every unit increase in the nox the avg price is decreased by -10.32.

D)The regression equation for this model is

**y=29.43+0.03\*age+0.13\*indus+(-10.27\*nox)+0.26\*distance+(-0.01\*tax)+(-1.07\*patratio)+4.13\*avg-room+(-0.61\*lstat)**