**PROJECT**

**TERRO’S REAL ESTATE AGENCY.**

**MOHAMED FAIZAL BASEER M**

**INDEX**

|  |  |
| --- | --- |
| **QUESTIONS** | **PG NO** |
| Q1) SUMMARY STATISTICS | 2 |
| Q2) HISTOGRAM | 3 |
| Q3) COVARIANCE | 3 |
| Q4) CORRELATION | 4 |
| Q5) AVG PRICE VS LSTAT BASE MODEL | 5 |
| Q6) LSTAT & AVG ROOM VS AVG PRICE | 6 |
| Q7) ALL VARIABLE VS AVG PRICE | 7 |
| Q8) SIGNIFICANT VARIABLE VS AVG PRICE | 10 |

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

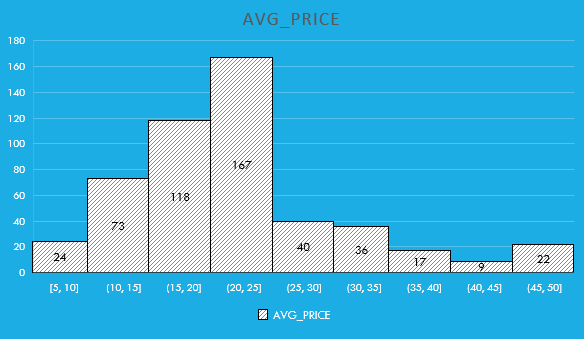
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SUMMARY | CRIME\_RATE | AGE | INDUS | NOX | DISTANCE |
| Mean | 4.871976285 | 68.57490119 | 11.13677866 | 0.554695059 | 9.549407115 |
| Standard Error | 0.129860152 | 1.251369525 | 0.304979888 | 0.005151391 | 0.387084894 |
| Median | 4.82 | 77.5 | 9.69 | 0.538 | 5 |
| Mode | 3.43 | 100 | 18.1 | 0.538 | 24 |
| Standard Deviation | 2.921131892 | 28.14886141 | 6.860352941 | 0.115877676 | 8.707259384 |
| Sample Variance | 8.533011532 | 792.3583985 | 47.06444247 | 0.013427636 | 75.81636598 |
| Kurtosis | -1.189122464 | -0.967715594 | -1.233539601 | -0.064667133 | -0.867231994 |
| Skewness | 0.021728079 | -0.59896264 | 0.295021568 | 0.729307923 | 1.004814648 |
| Range | 9.95 | 97.1 | 27.28 | 0.486 | 23 |
| Minimum | 0.04 | 2.9 | 0.46 | 0.385 | 1 |
| Maximum | 9.99 | 100 | 27.74 | 0.871 | 24 |
| Sum | 2465.22 | 34698.9 | 5635.21 | 280.6757 | 4832 |
| Count | 506 | 506 | 506 | 506 | 506 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SUMMARY | TAX | PTRATIO | AVG\_ROOM | LSTAT | AVG\_PRICE |
| Mean | 408.2371542 | 18.4555336 | 6.284634387 | 12.65306324 | 22.53280632 |
| Standard Error | 7.492388692 | 0.096243568 | 0.031235142 | 0.317458906 | 0.408861147 |
| Median | 330 | 19.05 | 6.2085 | 11.36 | 21.2 |
| Mode | 666 | 20.2 | 5.713 | 8.05 | 50 |
| Standard Deviation | 168.5371161 | 2.164945524 | 0.702617143 | 7.141061511 | 9.197104087 |
| Sample Variance | 28404.75949 | 4.686989121 | 0.49367085 | 50.99475951 | 84.58672359 |
| Kurtosis | -1.142407992 | -0.285091383 | 1.891500366 | 0.493239517 | 1.495196944 |
| Skewness | 0.669955942 | -0.802324927 | 0.403612133 | 0.906460094 | 1.108098408 |
| Range | 524 | 9.4 | 5.219 | 36.24 | 45 |
| Minimum | 187 | 12.6 | 3.561 | 1.73 | 5 |
| Maximum | 711 | 22 | 8.78 | 37.97 | 50 |
| Sum | 206568 | 9338.5 | 3180.025 | 6402.45 | 11401.6 |
| Count | 506 | 506 | 506 | 506 | 506 |

**OBSERVATIONS:**

* From distance table, nearly 50% of the houses from this given dataset has distance between from 1 to 5 miles.
* Except AGE & PTRATIO all other variables are positively skewed, which means mean>median.
* Except AVG\_ROOM, LSTAT & AVG\_PRICE all other variables have negative kurtosis (platykurtic) which means that more data values are located near mean & less data values are located on tails. And distribution produces fewer outliers than normal distribution.

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



**INFERENCE:**

* MAJORITY OF HOUSES HAS AVERAGE PRICE BETWEEN $16000s TO $25000s.
* AROUND 56.5% OF THE HOUSES IN THE DATASET PROVIDED BY THE COMPANY FALLS UNDER $16000s TO $25000s.

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



**OBSERVATIONS:**

* Covariance measures the direction of a relationship between two variables. When covariance is negative there is an inverse relationship between the variable.
* There is an inverse relationship between,  
   **\*** Avg\_price VS Age, Indus, NOX, Distance, Tax, PT Ratio, LSTAT.

**4) Create a correlation matrix of all the variables.**



**OBSERVATIONS:**

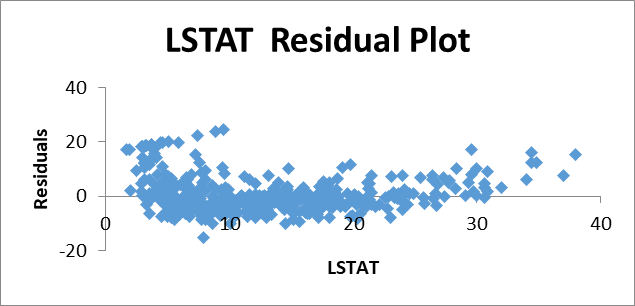
* Correlation = 1 means variables are perfectly correlated.
* There is an inverse relationship between,  
   **\*** Avg\_price VS Age, Indus, NOX, Distance, Tax, PT Ratio, LSTAT.
* The correlation between Crime\_rate VS Avg \_price is just 4.33% which is very low compared to other correlated pairs.

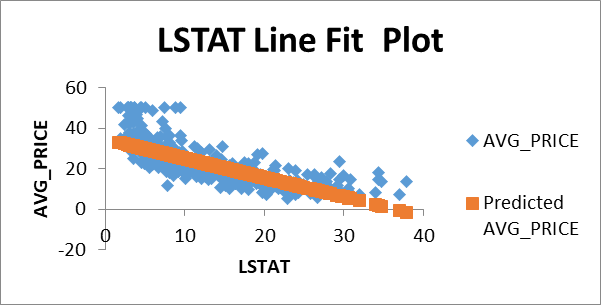
1. **Which are the top 3 positively correlated pairs?  
    \***Apart from 1, The next top three +vely correlated pairs are Tax VS Distance, Indus VS NOX, Age VS NOX.
2. **Which are the top 3 negatively correlated pairs.?  
    \*** Avg\_price VS LSTAT, Avg\_room VS LSTAT, Avg\_price VS PT Ratio.

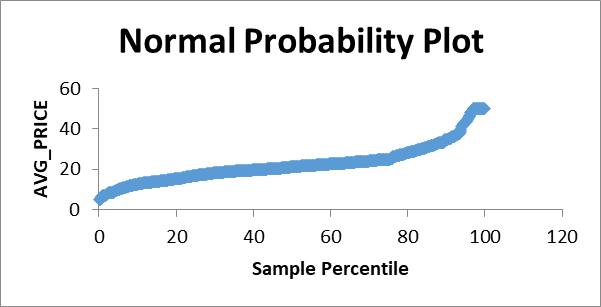
###### 5) Build an initial regression model with AVG\_PRICE as ‘y’ (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot.

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?

\* Model has been generated in excel. Kindly refer 5) LSTAT VS Avg\_price excel sheet.







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

**\*** The equation for this model is

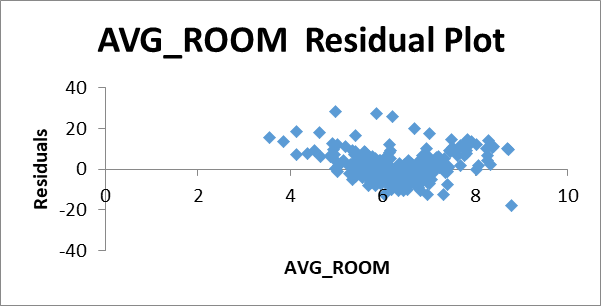
Avg\_price = (-0.950)\*LSTAT+34.553

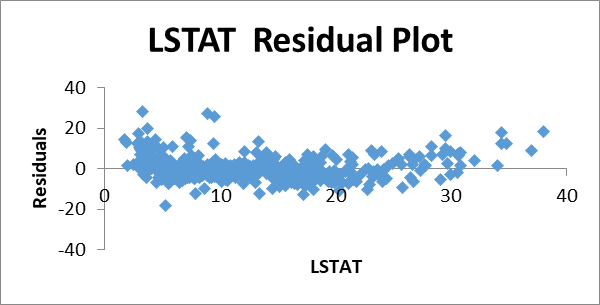
**\*** Mean Squared Error(MSE) = 38.482

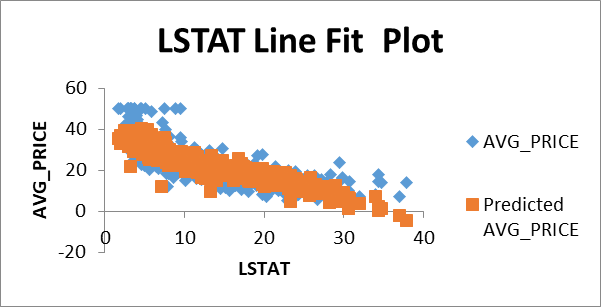
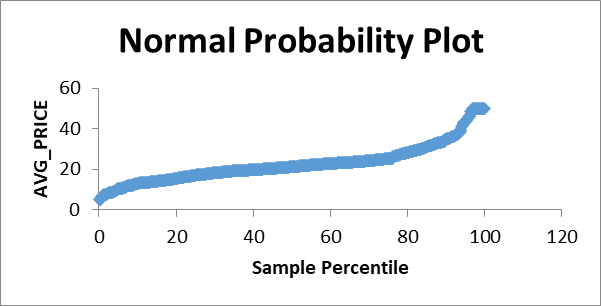
**\*** Significance Value is 5.0811E-88, means nearly 0.

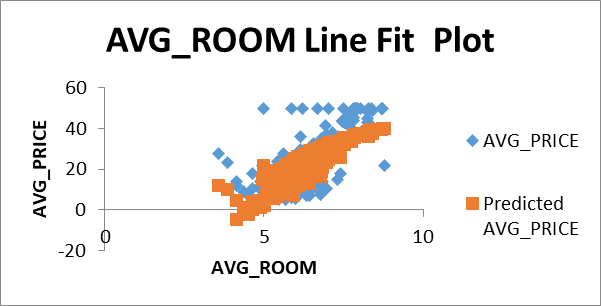
* From residual plot, The errors are uniformly distributed so this model is homoskedasticity.
* From R2, 54.41% of LSTAT variable is capable to determine Avg\_price from this regression model.
* As the coefficient of LSTAT is negative, both the variables are inversely proportional to each other.
* From intercept, we can know that if the value of LSTAT(independent) variable is zero then the Avg\_price will be USD34.553 (Intercept means value of “Y” at X=0).
* Yes, LSTAT is a significant variable as the significance level is lesser than 5%.

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

Regression model is created in Excel. Refer 6) LSTAT, ROOM VS AVG PRICE sheet.







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

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

Regression Equation:

**\*** Avg\_price=(5.09\*Avg\_room)+(-0.64\*LSTAT)-1.358

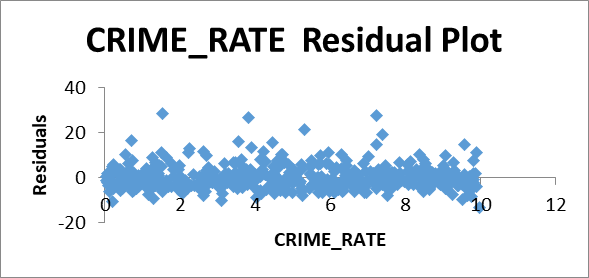
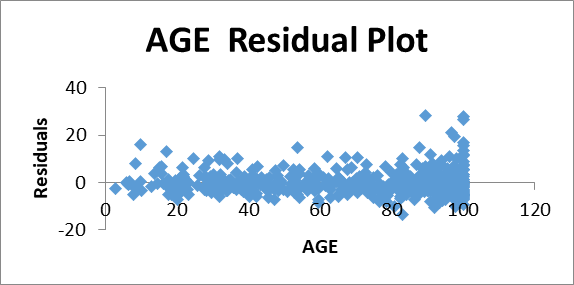
**\*** x1=7 , x2=20 , b = -1.358 , Avg\_Price = USD 21458.

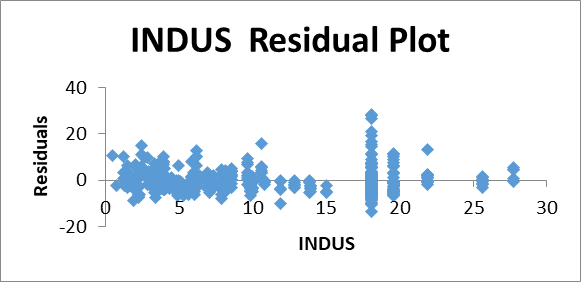
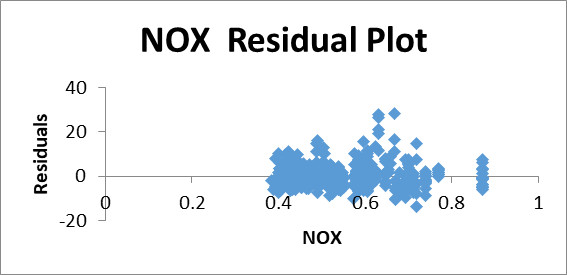
**\*** Company is overcharging.

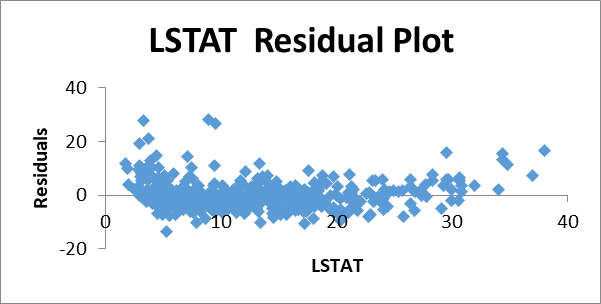
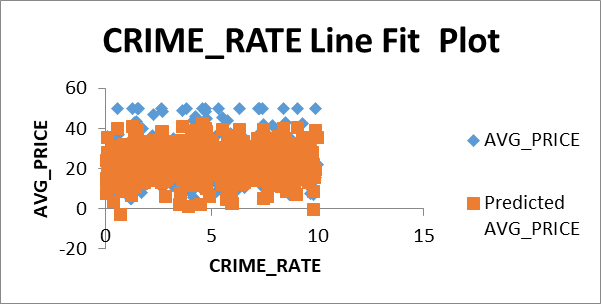
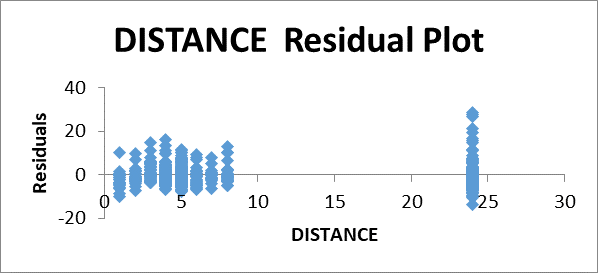
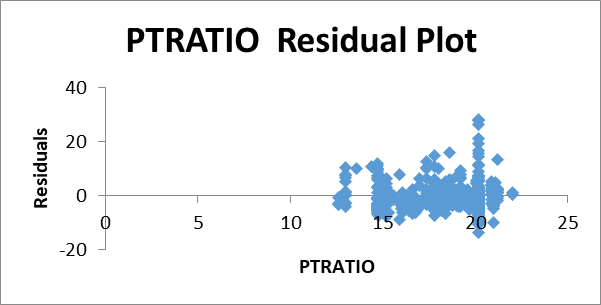
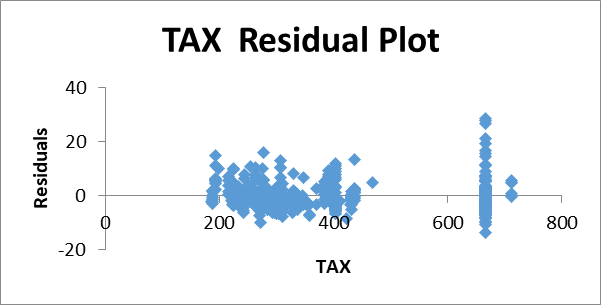
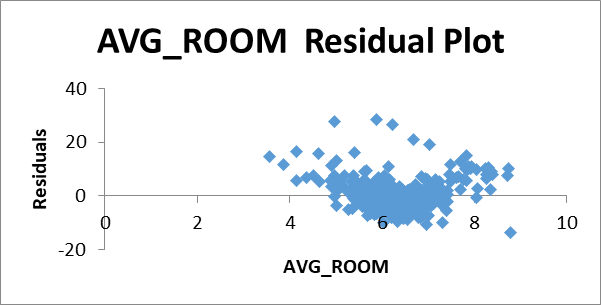
**\*** MSE = 30.512.

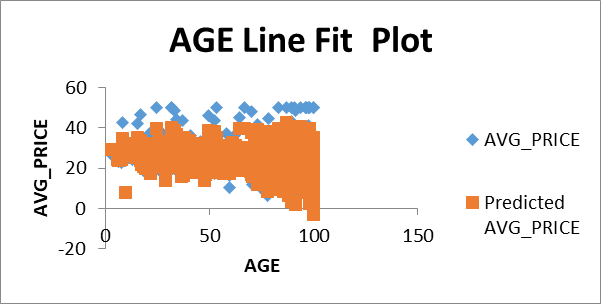
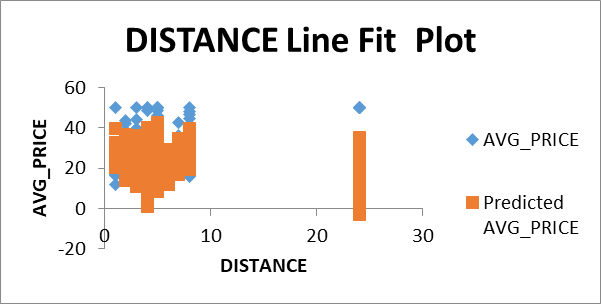
* Yes, performance of this model is better than the previous model as the value of R2 increases to 63.85% and mean squared error decreases to 30.512.

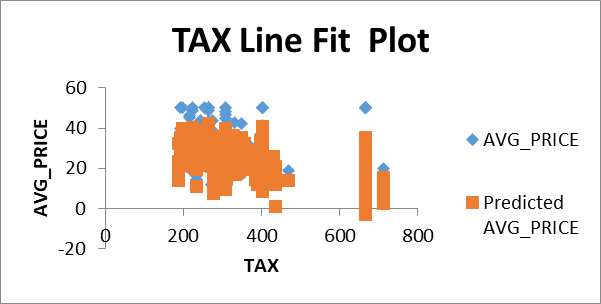
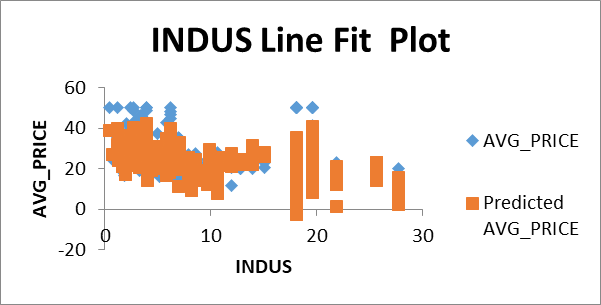
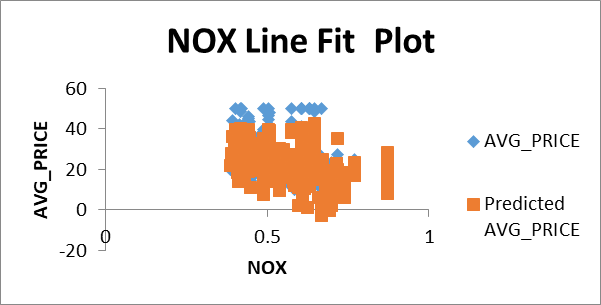
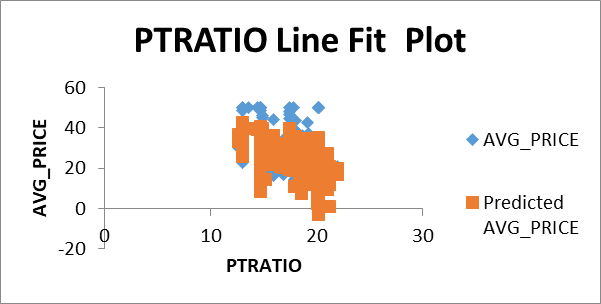
**7) Build another Regression model with all variables where AVG\_PRICE alone be the Dependent Variable and all the other variables are independent. Interpret the output in terms of adjusted R2, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG\_PRICE.**

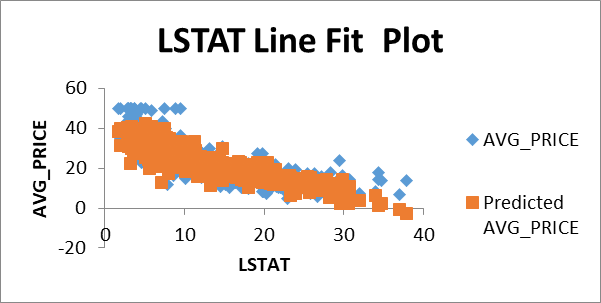
Refer 7) ALL VS AVG PRICE excel sheet for regression model.

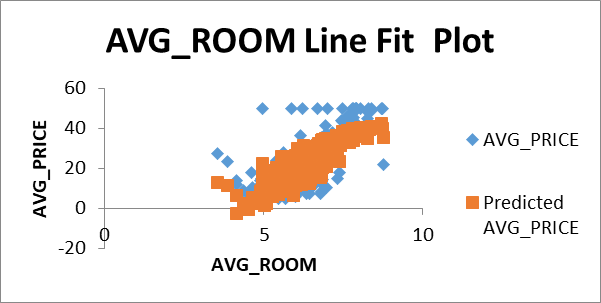


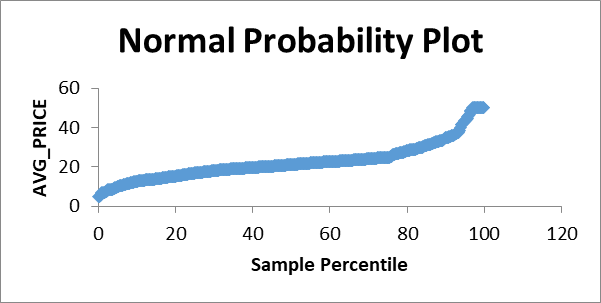












|  |  |
| --- | --- |
| **Regression Statistics** | |
| **Multiple R** | 0.832978824 |
| **R Square** | 0.69385372 |
| **Adjusted R Square** | 0.688298647 |
| **Standard Error** | 5.1347635 |
| **Observations** | 506 |

|  |  |  |
| --- | --- | --- |
|  | **Coefficients** | **P-value** |
| **Intercept** | 29.24132 | 2.54E-09 |
| **CRIME\_RATE** | 0.048725 | 0.534657 |
| **AGE** | 0.032771 | 0.01267 |
| **INDUS** | 0.130551 | 0.039121 |
| **NOX** | -10.3212 | 0.008294 |
| **DISTANCE** | 0.261094 | 0.000138 |
| **TAX** | -0.0144 | 0.000251 |
| **PTRATIO** | -1.07431 | 6.59E-15 |
| **AVG\_ROOM** | 4.125409 | 3.89E-19 |
| **LSTAT** | -0.60349 | 8.91E-27 |

* R2 = 69.38%, which is better than the previous model.
* Coefficients of NOX, Tax, PT Ratio, LSTAT are negative, which means these variables are inversely impacting the Avg\_price.
* Significance of crime rate is 53.46% which is more than 5%. Therefore, crime rate is insignificant variable.
* Significance level of Age, Indus, NOX, Distance, Tax, PT Ratio, Avg\_room, LSTAT are less than the 5%.
* MSE of this model is 25.84 which is lesser than the previous model.
* Intercept value is 29.24 which means when all the independent variables are zero then the value of Avg\_price will be USD 29241.

**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.**

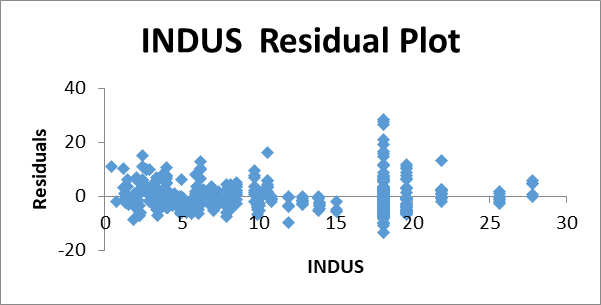
**b) Compare the adjusted R-square value of this model with the model in the previous question, which model performs better according to the value of adjusted R-square?**

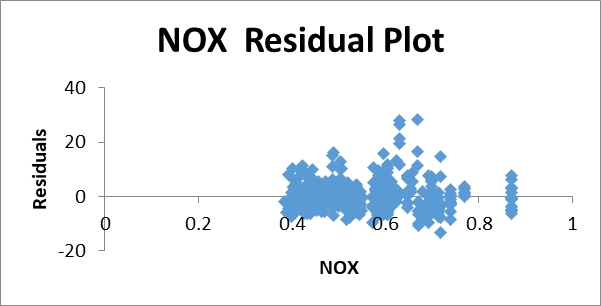
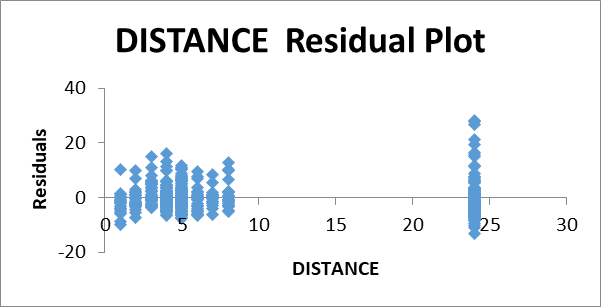
**c) Sort the values of the Coefficients in ascending order. What will happen to the average price if the value of NOX is more in a locality in this town?**

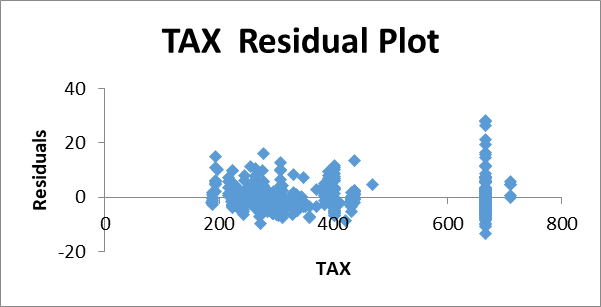
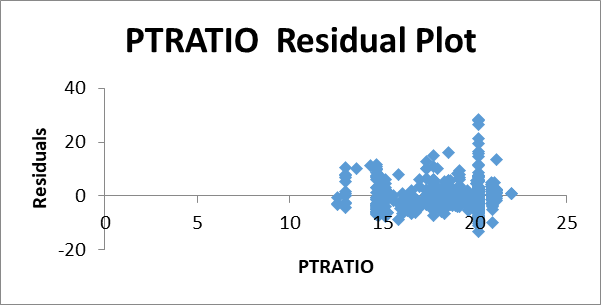
**d) Write the regression equation from this model.**

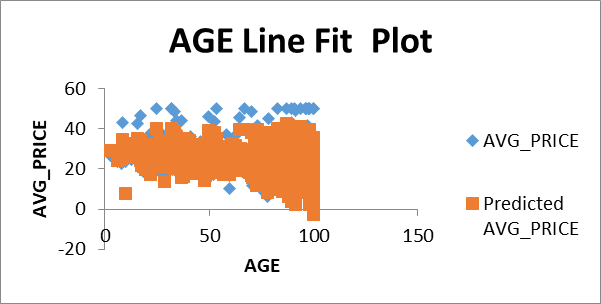
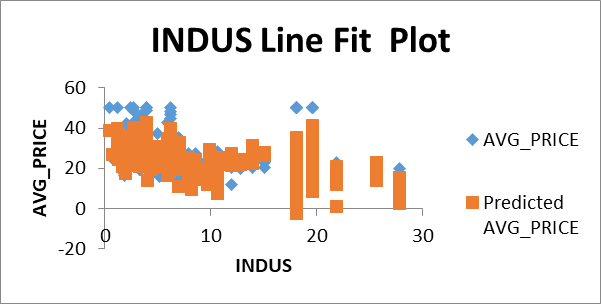
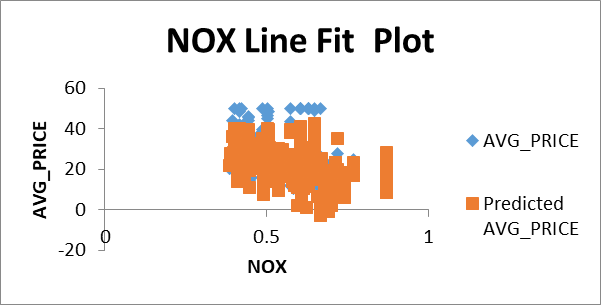
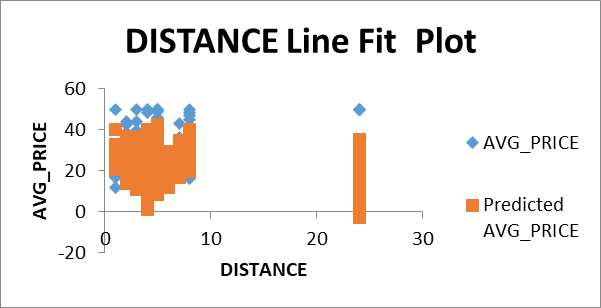
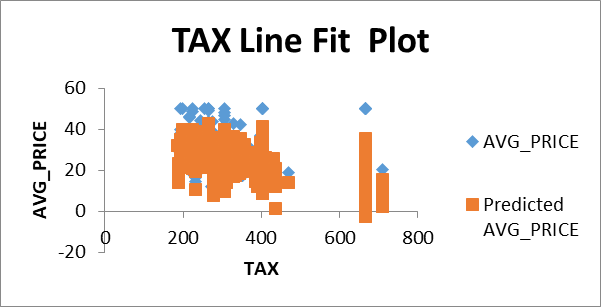
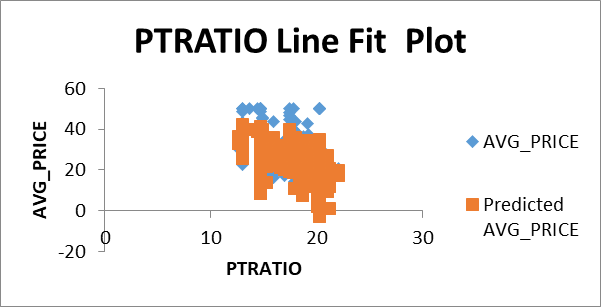
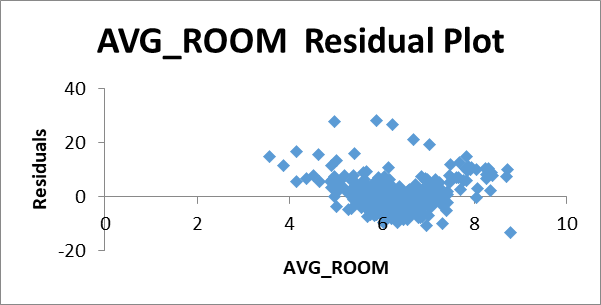
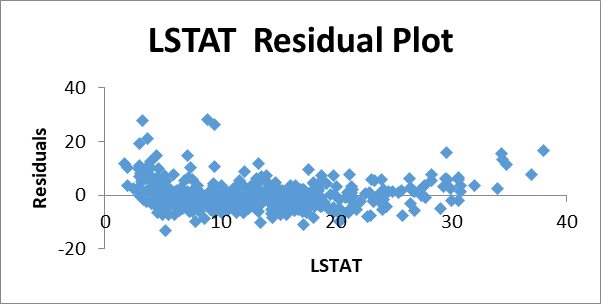
**\*** Significant (Independent) variable: Age, Indus, NOX, PT Ratio, Tax, LSTAT, Avg\_room, Distance.

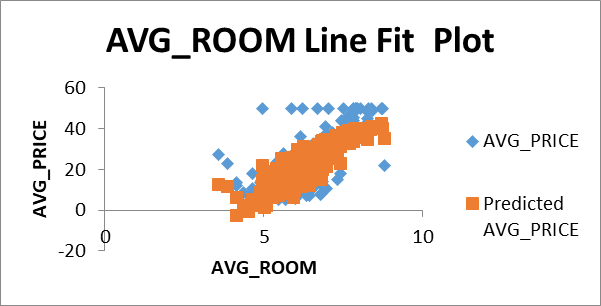
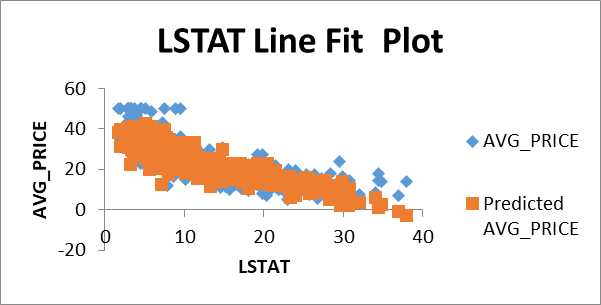
**\*** Dependent Variable: Avg\_price.

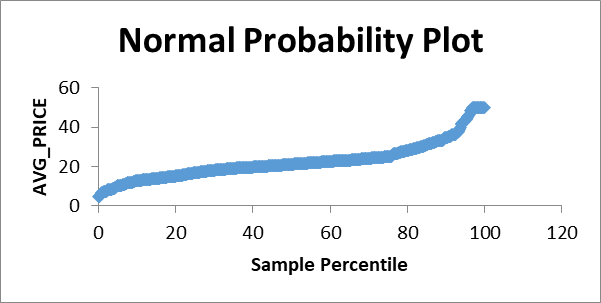
**\*** Refer 8) Significant VS price excel sheet for regression model.









|  |  |
| --- | --- |
| **Regression Statistics** | |
| **Multiple R** | 0.832835773 |
| **R Square** | 0.693615426 |
| **Adjusted R Square** | 0.688683682 |
| **Standard Error** | 5.131591113 |
| **Observations** | 506 |

**a) Interpret the output of this model.**

|  |  |  |
| --- | --- | --- |
|  | **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\_ROOM** | 4.125469 | 3.69E-19 |
| **LSTAT** | -0.60516 | 5.42E-27 |

* All the independent variables are significant variable.
* R2 = 69.36%.
* Intercept value is 29.42 which means when all the independent variables are zero then the value of Avg\_price will be USD 29428.

**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?**

* R2 of previous model is 69.38% and R2 of this model is 69.36% which is 0.02% lesser than the previous model.
* Previous model MSE = 25.84
* MSE of this model = 25.86
* So, changes in the MSE and R2 for this model is very less compared to the previous model.

**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?**

|  |  |
| --- | --- |
| **Variables** | **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 |

**\*** As the coefficient of NOX is negative, it is inversely proportional to Avg\_price. Therefore, if the value of NOX is more the Avg\_price will be decreased.

**d) Write the regression equation from this model.**

Avg\_price = (-10.272\*NOX) +(-1.071\* PT Ratio) +(-0.605\*LSTAT) +(-0.014\*Tax) +(0.032\*Age) + (0.130\*Indus) +(0.261\*Distance) +(4.125\*Avg\_room) +29.428.