Business Insights

Vinay Kumar Govardhanam

**Table of Contents**

1. **The first step to any project is understanding the data. So for this step, generate the summary statistics for each of the variables. What do you observe? ​(5 marks)**
2. **Plot the histogram of the Avg\_Price Variable. What do you infer? ​(5 marks)**
3. **Compute the covariance matrix. Share your observations. ​(5 marks)**
4. **Create a correlation matrix of all the variables as shown in the Videos and various case studies. State top 3 positively correlated pairs and top 3 negatively correlated pairs. ​(5 marks)**
5. **Build an initial regression model with AVG\_PRICE as the y or the Dependent variable and LSTAT variable as the Independent Variable. Generate the residual plot too. ​(8 marks)** 
   * 1. **What do you infer from the Regression Summary Output in terms of variance explained, coefficient value, Intercept and the Residual plot?**
     2. **Is LSTAT variable significant for the analysis based on your model?**

**6. Build another instance of the Regression model but this time including LSTAT and AVG\_ROOM together as Independent variables and AVG\_PRICE as the 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. Explain.**

**7. Now, build a Regression model with all variables. AVG\_PRICE shall be the Dependent Variable. Interpret the output in terms of adjusted R-square, coefficient and Intercept values, Significance of variables with respect to AVG\_price. Explain. ​(8 marks)**

**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. ​(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?**

* + 1. **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?**
    2. **Write the regression equation from this model.**

**9. Hypothesis Testing of the features (under assumptions where it is affecting the variable or any multicollinearity situations )**

**Crime\_Ratio Vs PTRatio** | **Crime\_Ratio Vs Indus** | **Age Vs Avg\_Price |** **Nox Vs Indus | Avg\_Room Vs Avg\_Price**

**Tax Vs Avg\_Price | LSTAT Vs Avg\_Price**

1. **The first step to any project is understanding the data. So for this step, generate the summary statistics for each of the variables. What do you observe? ​(5 marks)**

**SUMMARY STATISTICS**

**# Crime\_Rate**

In the Boston town, overall crime rate is 4.871976285

The total no of observation recorded so far are 506 and no missing values are found

The data shows that the total no of observations are on the right scale of the number line which is slightly right positive skewed of 0.021728079

The data shows that the structure of the observation is flat of 1.189122464 which is known as platykurtic and variation of the data is 2.921131892

The highest occurrence recorded is 3.43

The minimum rate is 0.04 and maximum rate is 9.99

The coefficient of variation of the data is 0.59957843

The most of data lies in the range of 4.9125

**# Age**

Overall proportion of houses prior to 1940 (in percentage terms) is 68.5749012

The total no of observation recorded so far are 506 and no missing values are found

The data shows that the total no of observations are on the left scale of the number line which is slightly left negative skewed of -0.5989626

The data shows that the structure of the observation is flat of -0.9677156 which is known as platykurtic and variation of the data is 6.86035294

The highest occurrence recorded is 100

The minimum rate is 2.9 and maximum rate is 100

The coefficient of variation of the data is 0.61600874

The most of data lies in the range of 12.91

**# Indus**

Overall proportion of non-retail business acres per town (in percentage terms) is 11.1367787

The total no of observation recorded so far are 506 and no missing values are found

The data shows that the total no of observations are on the right scale of the number line which is slightly right positive skewed of 0.29502157

The data shows that the structure of the observation is flat of -1.2335396 which is known as platykurtic which depicts the low areas of industry in the town and variation of the data is 6.86035294

The highest occurrence recorded is 18.1

The minimum rate is 0.46 and maximum rate is 27.74

The coefficient of variation of the data is 0.61600874

The most of data lies in the range of 12.91

**# Nox**

Overall proportion of nitric oxide concentration (parts per 10 million) is 0.554695059

The total no of observation recorded so far are 506 and no missing values are found

The data shows that the total no of observations are on the right scale of the number line which is slightly right positive skewed of 0.729307923

The data shows that the structure of the observation is flat of -0.064667133 which is known as platykurtic and variation of the data is 0.115877676

The highest occurrence recorded is 0.538

The minimum rate is 0.385 and maximum rate is 0.871

The coefficient of variation of the data is 0.208903385

The most of data lies in the range of 0.175

**# Distance**

The average distance covered from highway (in miles) is 9.549407115

The total no of observation recorded so far are 506 and no missing values are found

The data shows that the total no of observations are on the right scale of the number line which is right positive skewed of 1.004814648

The data shows that the structure of the observation is flat of

-0.867231994 which is known as platykurtic and variation of the data is 8.707259384

The highest occurrence recorded is 24

The minimum rate is 1 and maximum rate is 24

The coefficient of variation of the data is 0.911811517

The most of data lies in the range of 20

**# Tax**

On an average, the full-value property tax rate per ($10,000) is 408.237154

The total no of observation recorded so far are 506 and no missing values are found

The data shows that the total no of observations are on the right scale of the number line which is slightly right positive skewed of 0.66995594

The data shows that the structure of the observation is flat of -1.14240799247682 which is known as platykurtic and variation of the data is 168.537116

The minimum rate is 187 and maximum rate is 711

The highest occurrence recorded is 666

The coefficient of variation of the data is 0.4128412

The most of data lies in the range of 387

**# PTRatio**

On an average, pupil teacher ratio by town is 19.05

The total no of observation recorded so far are 506 and no missing values are found

The data shows that the total no of observations are on the left scale of the number line which is slightly left negative skewed of -0.8023249

The data shows that the structure of the observation is flat of -0.2850914 which is known as platykurtic and variation of the data is 2.16494552

The minimum rate is 12.6 and maximum rate is 22

The highest occurrence recorded is 20.2

The coefficient of variation of the data is 0.11730604

The most of data lies in the range of 2.8

There are exactly 15 outliers in the data which deviates the actual mean of the PTRatio

**# AVG\_ROOM**

On an average, the no of rooms per house is 6.2085

The total no of observation recorded so far are 506 and no missing values are found

The data shows that the total no of observations are on the right scale of the number line which is slightly right positive skewed of 0.66995594

The data shows that the structure of the observation is peak of 1.89150037 which is known as leptokurtic and variation of the data is 0.70261714

The minimum rate is 3.561 and maximum rate is 8.78

The highest occurrence recorded is 5.713

The coefficient of variation of the data is 0.11179921

The most of data lies in the range of 0.738

There are exactly 30 outliers in the data which deviates the actual mean of the AVG\_ROOM

**# LSTAT**

On an average, the lower status of the population is 11.36

The total no of observation recorded so far are 506 and no missing values are found

The data shows that the total no of observations are on the right scale of the number line which is slightly right positive skewed of 0.906460094

The data shows that the structure of the observation is peak of 0.493239517 which is known as leptokurtic and variation of the data is 7.141061511

The minimum rate is 1.73 and maximum rate is 37.97

The highest occurrence recorded is 8.05

The coefficient of variation of the data is 0.564374126

The most of data lies in the range of 10.005

There are exactly 7 outliers in the data which deviates the actual mean of the LSTAT

**# AVG\_PRICE**

On an average, the price of the houses is 21.2

The total no of observation recorded so far are 506 and no missing values are found

The data shows that the total no of observations are on the right scale of the number line which is slightly right positive skewed of 1.10809841

The data shows that the structure of the observation is peak of 1.49519694 which is known as leptokurtic and variation of the data is 9.19710409

The minimum rate is 5 and maximum rate is 50

The highest occurrence recorded is 50

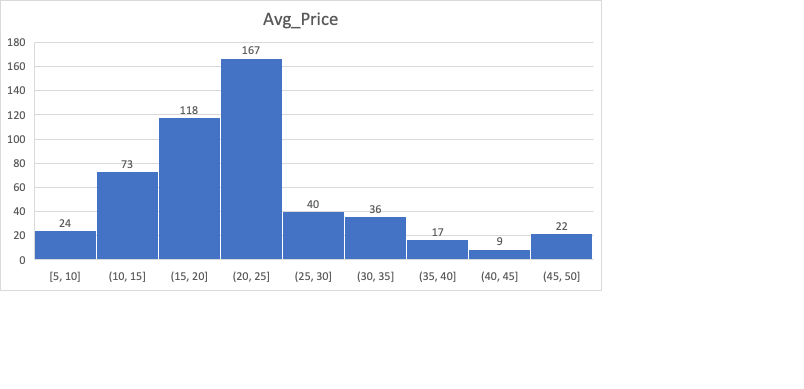
The coefficient of variation of the data is 0.4128412

The most of data lies in the range of 7.975

There are exactly 40 outliers in the data which deviates the actual mean of the Avg\_Price

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Features / Functions | **CRIME\_RATE** | **AGE** | **INDUS** | **NOX** | **DISTANCE** | **TAX** | **PTRATIO** | **AVG\_ROOM** | **LSTAT** | **AVG\_PRICE** |
| mean | 4.871976285 | 68.5749012 | 11.13677866 | 0.5546951 | 9.5494071 | 408.237154 | 18.4555336 | 6.284634387 | 12.653063 | 22.5328063 |
| median | 4.82 | 77.5 | 9.69 | 0.538 | 5 | 330 | 19.05 | 6.2085 | 11.36 | 21.2 |
| mode | 3.43 | 100 | 18.1 | 0.538 | 24 | 666 | 20.2 | 5.713 | 8.05 | 50 |
| sum | 2465.22 | 34698.9 | 5635.21 | 280.6757 | 4832 | 206568 | 9338.5 | 3180.025 | 6402.45 | 11401.6 |
| count | 506 | 506 | 506 | 506 | 506 | 506 | 506 | 506 | 506 | 506 |
| variance | 8.533011532 | 792.358399 | 47.06444247 | 0.0134276 | 75.816366 | 28404.7595 | 4.68698912 | 0.49367085 | 50.99476 | 84.5867236 |
| standard\_deviation | 2.921131892 | 28.1488614 | 6.860352941 | 0.1158777 | 8.7072594 | 168.537116 | 2.16494552 | 0.702617143 | 7.1410615 | 9.19710409 |
| Coefficient of Variation | 0.59957843 | 0.41048344 | 0.616008736 | 0.2089034 | 0.9118115 | 0.4128412 | 0.11730604 | 0.111799207 | 0.5643741 | 0.40816505 |
| minimum | 0.04 | 2.9 | 0.46 | 0.385 | 1 | 187 | 12.6 | 3.561 | 1.73 | 5 |
| maximum | 9.99 | 100 | 27.74 | 0.871 | 24 | 711 | 22 | 8.78 | 37.97 | 50 |
| range | 9.95 | 97.1 | 27.28 | 0.486 | 23 | 524 | 9.4 | 5.219 | 36.24 | 45 |
| Q1 | 2.4125 | 45.025 | 5.19 | 0.449 | 4 | 279 | 17.4 | 5.8855 | 6.95 | 17.025 |
| Q2 | 4.82 | 77.5 | 9.69 | 0.538 | 5 | 330 | 19.05 | 6.2085 | 11.36 | 21.2 |
| Q3 | 7.325 | 94.075 | 18.1 | 0.624 | 24 | 666 | 20.2 | 6.6235 | 16.955 | 25 |
| IQR | 4.9125 | 49.05 | 12.91 | 0.175 | 20 | 387 | 2.8 | 0.738 | 10.005 | 7.975 |
| Upperbound | 14.69375 | 167.65 | 37.465 | 0.8865 | 54 | 1246.5 | 24.4 | 7.7305 | 31.9625 | 36.9625 |
| Lowerbound | -4.95625 | -28.55 | -14.175 | 0.1865 | -26 | -301.5 | 13.2 | 4.7785 | -8.0575 | 5.0625 |
| skewness | 0.021728079 | -0.59896264 | 0.295021568 | 0.7293079 | 1.0048146 | 0.66995594 | -0.8023249 | 0.403612133 | 0.9064601 | 1.10809841 |
| kurtosis | -1.18912246 | -0.96771559 | -1.233539601 | -0.064667 | -0.867232 | -1.142408 | -0.2850914 | 1.891500366 | 0.4932395 | 1.49519694 |
| outliers | 0 | 0 | 0 | 0 | 0 | 0 | 15 | 30 | 7 | 40 |

* 1. **Plot the histogram of the Avg\_Price Variable. What do you infer? ​(5 marks)**



|  |  |
| --- | --- |
| *Bin* | *Frequency* |
| 5 | 2 |
| 7.04545455 | 4 |
| 9.09090909 | 15 |
| 11.1363636 | 14 |
| 13.1818182 | 22 |
| 15.2272727 | 44 |
| 17.2727273 | 32 |
| 19.3181818 | 53 |
| 21.3636364 | 70 |
| 23.4090909 | 77 |
| 25.4545455 | 52 |
| 27.5 | 15 |
| 29.5454545 | 17 |
| 31.5909091 | 17 |
| 33.6363636 | 18 |
| 35.6818182 | 10 |
| 37.7272727 | 10 |
| 39.7727273 | 2 |
| 41.8181818 | 3 |
| 43.8636364 | 5 |
| 45.9090909 | 3 |
| 47.9545455 | 2 |
| More | 19 |

**OBSERVATIONS**

* + - * we can say, the above distribution is rightly positive skewed of 1.10809841
      * from the middle value of the data is 21.2, there are 215 observations from the left and there are 214 observations from the right
      * from the above figure, the unimodal value of bin [20, 25] is 167
      * the least frequency occurrence of bin [40, 45] is 9
      * the most of observations lies in the range of 16.95 to 25 (Interquartile range)
      * there are exactly 40 outliers in the distribution

|  |  |
| --- | --- |
|  | *NOX* |
| CRIME\_RATE | 0.0006253 |
| AGE | 2.3812119 |
| INDUS | 0.6058739 |
| NOX | 0.0134011 |
| DISTANCE | 0.6157102 |
| TAX | 13.020502 |
| PTRATIO | 0.0473037 |
| AVG\_ROOM | -0.024555 |
| LSTAT | 0.4879799 |
| AVG\_PRICE | -0.454512 |

* 1. **Compute the covariance matrix. Share your observations. ​(5 marks)**

|  |  |
| --- | --- |
|  | *CRIME\_RATE* |
| CRIME\_RATE | 8.516147873 |
| AGE | 0.562915215 |
| INDUS | -0.11021518 |
| NOX | 0.000625308 |
| DISTANCE | -0.22986049 |
| TAX | -8.22932244 |
| PTRATIO | 0.068168906 |
| AVG\_ROOM | 0.056117778 |
| LSTAT | -0.88268036 |
| AVG\_PRICE | 1.16201224 |

|  |  |
| --- | --- |
|  | *AGE* |
| CRIME\_RATE | 0.56291522 |
| AGE | 790.792473 |
| INDUS | 124.267828 |
| NOX | 2.38121193 |
| DISTANCE | 111.549955 |
| TAX | 2397.94172 |
| PTRATIO | 15.9054254 |
| AVG\_ROOM | -4.74253803 |
| LSTAT | 120.838441 |
| AVG\_PRICE | -97.3961529 |

|  |  |
| --- | --- |
|  | *INDUS* |
| CRIME\_RATE | -0.110215175 |
| AGE | 124.2678282 |
| INDUS | 46.97142974 |
| NOX | 0.605873943 |
| DISTANCE | 35.47971449 |
| TAX | 831.7133331 |
| PTRATIO | 5.680854782 |
| AVG\_ROOM | -1.884225427 |
| LSTAT | 29.52181125 |
| AVG\_PRICE | -30.46050499 |

|  |  |
| --- | --- |
|  | *DISTANCE* |
| CRIME\_RATE | -0.2298605 |
| AGE | 111.54996 |
| INDUS | 35.479714 |
| NOX | 0.6157102 |
| DISTANCE | 75.666531 |
| TAX | 1333.1167 |
| PTRATIO | 8.7434025 |
| AVG\_ROOM | -1.2812774 |
| LSTAT | 30.325392 |
| AVG\_PRICE | -30.50083 |

|  |  |
| --- | --- |
|  | *TAX* |
| CRIME\_RATE | -8.2293224 |
| AGE | 2397.94172 |
| INDUS | 831.713333 |
| NOX | 13.0205024 |
| DISTANCE | 1333.11674 |
| TAX | 28348.6236 |
| PTRATIO | 167.820822 |
| AVG\_ROOM | -34.515101 |
| LSTAT | 653.420617 |
| AVG\_PRICE | -724.82043 |

|  |  |
| --- | --- |
|  | *PTRATIO* |
| CRIME\_RATE | 0.06816891 |
| AGE | 15.9054254 |
| INDUS | 5.68085478 |
| NOX | 0.04730365 |
| DISTANCE | 8.74340249 |
| TAX | 167.820822 |
| PTRATIO | 4.6777263 |
| AVG\_ROOM | -0.5396945 |
| LSTAT | 5.77130024 |
| AVG\_PRICE | -10.090676 |

|  |  |
| --- | --- |
|  | *AVG\_ROOM* |
| CRIME\_RATE | 0.056117778 |
| AGE | -4.74253803 |
| INDUS | -1.88422543 |
| NOX | -0.02455483 |
| DISTANCE | -1.28127739 |
| TAX | -34.515101 |
| PTRATIO | -0.53969452 |
| AVG\_ROOM | 0.492695216 |
| LSTAT | -3.07365497 |
| AVG\_PRICE | 4.484565552 |

|  |  |
| --- | --- |
|  | *LSTAT* |
| CRIME\_RATE | -0.8826804 |
| AGE | 120.83844 |
| INDUS | 29.521811 |
| NOX | 0.4879799 |
| DISTANCE | 30.325392 |
| TAX | 653.42062 |
| PTRATIO | 5.7713002 |
| AVG\_ROOM | -3.073655 |
| LSTAT | 50.893979 |
| AVG\_PRICE | -48.351792 |

|  |  |
| --- | --- |
|  | *AVG\_PRICE* |
| CRIME\_RATE | 1.16201224 |
| AGE | -97.396153 |
| INDUS | -30.460505 |
| NOX | -0.4545124 |
| DISTANCE | -30.50083 |
| TAX | -724.82043 |
| PTRATIO | -10.090676 |
| AVG\_ROOM | 4.48456555 |
| LSTAT | -48.351792 |
| AVG\_PRICE | 84.4195562 |

**# CRIME\_RATE**

* Covariance of CRIME\_RATE and CRIME\_RATE is 8.516147873 which depicts the magnitude of 9 and its states from Positive to Positive direction
* Covariance of CRIME\_RATE and Age is 0.562915215 which depicts the magnitude of 1 and its states from Positive to Positive direction
* Covariance of CRIME\_RATE and Indus is -0.110215175 which depicts the magnitude of 0 and its states from Negative to Zero direction
* Covariance of CRIME\_RATE and Nox is 0.000625308 which depicts the magnitude of 0 and its states from Positive to Zero direction
* Covariance of CRIME\_RATE and Distance is -0.229860488 which depicts the magnitude of 0 and its states from Negative to Zero direction
* Covariance of CRIME\_RATE and Tax is -8.229322439 which depicts the magnitude of -8 and its states from Negative to Negative direction
* Covariance of CRIME\_RATE and PTRatio is 0.068168906 which depicts the magnitude of 0 and its states from Positive to Zero direction
* Covariance of CRIME\_RATE and AVG\_ROOM is 0.056117778 which depicts the magnitude of 0 and its states from Positive to Zero direction
* Covariance of CRIME\_RATE and LSTAT is -0.882680362 which depicts the magnitude of -1 and its states from Negative to Negative direction
* Covariance of CRIME\_RATE and AVG\_PRICE is 1.16201224 which depicts the magnitude of 1 and its states from Positive to Positive direction

**# AGE**

* Covariance of AGE and CRIME\_RATE is 0.562915215 which depicts the magnitude of 1 and its states from Positive to Positive direction
* Covariance of AGE and Age is 790.7924728 which depicts the magnitude of 791 and its states from Positive to Positive direction
* Covariance of AGE and Indus is 124.2678282 which depicts the magnitude of 124 and its states from Positive to Positive direction
* Covariance of AGE and Nox is 2.381211931 which depicts the magnitude of 2 and its states from Positive to Positive direction
* Covariance of AGE and Distance is 111.5499555 which depicts the magnitude of 112 and its states from Positive to Positive direction
* Covariance of AGE and Tax is 2397.941723 which depicts the magnitude of 2398 and its states from Positive to Positive direction
* Covariance of AGE and PTRatio is 15.90542545 which depicts the magnitude of 16 and its states from Positive to Positive direction
* Covariance of AGE and AVG\_ROOM is -4.74253803 which depicts the magnitude of -5 and its states from Negative to Negative direction
* Covariance of AGE and LSTAT is 120.8384405 which depicts the magnitude of 121 and its states from Positive to Positive direction
* Covariance of AGE and AVG\_PRICE is -97.39615288 which depicts the magnitude of -97 and its states from Negative to Negative direction

**# INDUS**

* Covariance of INDUS and CRIME\_RATE is -0.110215175 which depicts the magnitude of 0 and its states from Negative to Zero direction
* Covariance of INDUS and Age is 124.2678282 which depicts the magnitude of 124 and its states from Positive to Positive direction
* Covariance of INDUS and Indus is 46.97142974 which depicts the magnitude of 47 and its states from Positive to Positive direction
* Covariance of INDUS and Nox is 0.605873943 which depicts the magnitude of 1 and its states from Positive to Positive direction
* Covariance of INDUS and Distance is 35.47971449 which depicts the magnitude of 35 and its states from Positive to Positive direction
* Covariance of INDUS and Tax is 831.7133331 which depicts the magnitude of 832 and its states from Positive to Positive direction
* Covariance of INDUS and PTRatio is 5.680854782 which depicts the magnitude of 6 and its states from Positive to Positive direction
* Covariance of INDUS and AVG\_ROOM is -1.884225427 which depicts the magnitude of -2 and its states from Negative to Negative direction
* Covariance of INDUS and LSTAT is 29.52181125 which depicts the magnitude of 30 and its states from Positive to Positive direction
* Covariance of INDUS and AVG\_PRICE is -30.46050499 which depicts the magnitude of -30 and its states from Negative to Negative direction

**# NOX**

* Covariance of NOX and CRIME\_RATE is 0.000625308 which depicts the magnitude of 0 and its states from Positive to Zero direction
* Covariance of NOX and Age is 2.381211931 which depicts the magnitude of 2 and its states from Positive to Positive direction
* Covariance of NOX and Indus is 0.605873943 which depicts the magnitude of 1 and its states from Positive to Positive direction
* Covariance of NOX and Nox is 0.013401099 which depicts the magnitude of 0 and its states from Positive to Zero direction
* Covariance of NOX and Distance is 0.615710224 which depicts the magnitude of 1 and its states from Positive to Positive direction
* Covariance of NOX and Tax is 13.02050236 which depicts the magnitude of 13 and its states from Positive to Positive direction
* Covariance of NOX and PTRatio is 0.047303654 which depicts the magnitude of 0 and its states from Positive to Zero direction
* Covariance of NOX and AVG\_ROOM is -0.024554826 which depicts the magnitude of 0 and its states from Negative to Zero direction
* Covariance of NOX and LSTAT is 0.487979871 which depicts the magnitude of 0 and its states from Positive to Zero direction
* Covariance of NOX and AVG\_PRICE is -0.454512407 which depicts the magnitude of 0 and its states from Negative to Zero direction

**# DISTANCE**

* Covariance of DISTANCE and CRIME\_RATE is -0.229860488 which depicts the magnitude of 0 and its states from Negative to Zero direction
* Covariance of DISTANCE and Age is 111.5499555 which depicts the magnitude of 112 and its states from Positive to Positive direction
* Covariance of DISTANCE and Indus is 35.47971449 which depicts the magnitude of 35 and its states from Positive to Positive direction
* Covariance of DISTANCE and Nox is 0.615710224 which depicts the magnitude of 1 and its states from Positive to Positive direction
* Covariance of DISTANCE and Distance is 75.66653127 which depicts the magnitude of 76 and its states from Positive to Positive direction
* Covariance of DISTANCE and Tax is 1333.116741 which depicts the magnitude of 1333 and its states from Positive to Positive direction
* Covariance of DISTANCE and PTRatio is 8.74340249 which depicts the magnitude of 9 and its states from Positive to Positive direction
* Covariance of DISTANCE and AVG\_ROOM is -1.281277391 which depicts the magnitude of -1 and its states from Negative to Negative direction
* Covariance of DISTANCE and LSTAT is 30.32539213 which depicts the magnitude of 30 and its states from Positive to Positive direction
* Covariance of DISTANCE and AVG\_PRICE is -30.50083035 which depicts the magnitude of -31 and its states from Negative to Negative direction

**# TAX**

* Covariance of TAX and CRIME\_RATE is -8.229322439 which depicts the magnitude of -8 and its states from Negative to Negative direction
* Covariance of TAX and Age is 2397.941723 which depicts the magnitude of 2398 and its states from Positive to Positive direction
* Covariance of TAX and Indus is 831.7133331 which depicts the magnitude of 832 and its states from Positive to Positive direction
* Covariance of TAX and Nox is 13.02050236 which depicts the magnitude of 13 and its states from Positive to Positive direction
* Covariance of TAX and Distance is 1333.116741 which depicts the magnitude of 1333 and its states from Positive to Positive direction
* Covariance of TAX and Tax is 28348.6236 which depicts the magnitude of 28349 and its states from Positive to Positive direction
* Covariance of TAX and PTRatio is 167.8208221 which depicts the magnitude of 168 and its states from Positive to Positive direction
* Covariance of TAX and AVG\_ROOM is -34.51510104 which depicts the magnitude of -35 and its states from Negative to Negative direction
* Covariance of TAX and LSTAT is 653.4206174 which depicts the magnitude of 653 and its states from Positive to Positive direction
* Covariance of TAX and AVG\_PRICE is -724.8204284 which depicts the magnitude of -725 and its states from Negative to Negative direction

**# PTRATIO**

* Covariance of PTRATIO and CRIME\_RATE is 0.068168906 which depicts the magnitude of 0 and its states from Positive to Zero direction
* Covariance of PTRATIO and Age is 15.90542545 which depicts the magnitude of 16 and its states from Positive to Positive direction
* Covariance of PTRATIO and Indus is 5.680854782 which depicts the magnitude of 6 and its states from Positive to Positive direction
* Covariance of PTRATIO and Nox is 0.047303654 which depicts the magnitude of 0 and its states from Positive to Zero direction
* Covariance of PTRATIO and Distance is 8.74340249 which depicts the magnitude of 9 and its states from Positive to Positive direction
* Covariance of PTRATIO and Tax is 167.8208221 which depicts the magnitude of 168 and its states from Positive to Positive direction
* Covariance of PTRATIO and PTRatio is 4.677726296 which depicts the magnitude of 5 and its states from Positive to Positive direction
* Covariance of PTRATIO and AVG\_ROOM is -0.539694518 which depicts the magnitude of -1 and its states from Negative to Negative direction
* Covariance of PTRATIO and LSTAT is 5.771300243 which depicts the magnitude of 6 and its states from Positive to Positive direction
* Covariance of PTRATIO and AVG\_PRICE is -10.09067561 which depicts the magnitude of -10 and its states from Negative to Negative direction

**# AVG\_ROOM**

* Covariance of AVG\_ROOM and CRIME\_RATE is 0.056117778 which depicts the magnitude of 0 and its states from Positive to Zero direction
* Covariance of AVG\_ROOM and Age is -4.74253803 which depicts the magnitude of -5 and its states from Negative to Negative direction
* Covariance of AVG\_ROOM and Indus is -1.884225427 which depicts the magnitude of -2 and its states from Negative to Negative direction
* Covariance of AVG\_ROOM and Nox is -0.024554826 which depicts the magnitude of 0 and its states from Negative to Zero direction
* Covariance of AVG\_ROOM and Distance is -1.281277391 which depicts the magnitude of -1 and its states from Negative to Negative direction
* Covariance of AVG\_ROOM and Tax is -34.51510104 which depicts the magnitude of -35 and its states from Negative to Negative direction
* Covariance of AVG\_ROOM and PTRatio is -0.539694518 which depicts the magnitude of -1 and its states from Negative to Negative direction
* Covariance of AVG\_ROOM and AVG\_ROOM is 0.492695216 which depicts the magnitude of 0 and its states from Positive to Zero direction
* Covariance of AVG\_ROOM and LSTAT is -3.073654967 which depicts the magnitude of -3 and its states from Negative to Negative direction
* Covariance of AVG\_ROOM and AVG\_PRICE is 4.484565552 which depicts the magnitude of 4 and its states from Positive to Positive direction

**# LSTAT**

* Covariance of LSTAT and CRIME\_RATE is -0.882680362 which depicts the magnitude of -1 and its states from Negative to Negative direction
* Covariance of LSTAT and Age is 120.8384405 which depicts the magnitude of 121 and its states from Positive to Positive direction
* Covariance of LSTAT and Indus is 29.52181125 which depicts the magnitude of 30 and its states from Positive to Positive direction
* Covariance of LSTAT and Nox is 0.487979871 which depicts the magnitude of 0 and its states from Positive to Zero direction
* Covariance of LSTAT and Distance is 30.32539213 which depicts the magnitude of 30 and its states from Positive to Positive direction
* Covariance of LSTAT and Tax is 653.4206174 which depicts the magnitude of 653 and its states from Positive to Positive direction
* Covariance of LSTAT and PTRatio is 5.771300243 which depicts the magnitude of 6 and its states from Positive to Positive direction
* Covariance of LSTAT and AVG\_ROOM is -3.073654967 which depicts the magnitude of -3 and its states from Negative to Negative direction
* Covariance of LSTAT and LSTAT is 50.89397935 which depicts the magnitude of 51 and its states from Positive to Positive direction
* Covariance of LSTAT and AVG\_PRICE is -48.35179219 which depicts the magnitude of -48 and its states from Negative to Negative direction

**# AVG\_PRICE**

* Covariance of AVG\_PRICE and CRIME\_RATE is 1.16201224 which depicts the magnitude of 1 and its states from Positive to Positive direction
* Covariance of AVG\_PRICE and Age is -97.39615288 which depicts the magnitude of -97 and its states from Negative to Negative direction
* Covariance of AVG\_PRICE and Indus is -30.46050499 which depicts the magnitude of -30 and its states from Negative to Negative direction
* Covariance of AVG\_PRICE and Nox is -0.454512407 which depicts the magnitude of 0 and its states from Negative to Zero direction
* Covariance of AVG\_PRICE and Distance is -30.50083035 which depicts the magnitude of -31 and its states from Negative to Negative direction
* Covariance of AVG\_PRICE and Tax is -724.8204284 which depicts the magnitude of -725 and its states from Negative to Negative direction
* Covariance of AVG\_PRICE and PTRatio is -10.09067561 which depicts the magnitude of -10 and its states from Negative to Negative direction
* Covariance of AVG\_PRICE and AVG\_ROOM is 4.484565552 which depicts the magnitude of 4 and its states from Positive to Positive direction
* Covariance of AVG\_PRICE and LSTAT is -48.35179219 which depicts the magnitude of -48 and its states from Negative to Negative direction
* Covariance of AVG\_PRICE and AVG\_PRICE is 84.41955616 which depicts the magnitude of 84 and its states from Positive to Positive direction
  1. **Create a correlation matrix of all the variables as shown in the Videos and various case studies. State top 3 positively correlated pairs and top 3 negatively correlated pairs. ​(5 marks)**

|  |  |
| --- | --- |
|  | *CRIME\_RATE* |
| CRIME\_RATE | 1 |
| AGE | 0.006859463 |
| INDUS | -0.00551065 |
| NOX | 0.001850982 |
| DISTANCE | -0.00905505 |
| TAX | -0.01674852 |
| PTRATIO | 0.010800586 |
| AVG\_ROOM | 0.02739616 |
| LSTAT | -0.04239832 |
| AVG\_PRICE | 0.043337871 |

|  |  |
| --- | --- |
|  | *AGE* |
| CRIME\_RATE | 0.006859463 |
| AGE | 1 |
| INDUS | 0.64477851 |
| NOX | 0.7314701 |
| DISTANCE | 0.45602245 |
| TAX | 0.50645559 |
| PTRATIO | 0.26151501 |
| AVG\_ROOM | -0.24026493 |
| LSTAT | 0.60233853 |
| AVG\_PRICE | -0.37695457 |

|  |  |
| --- | --- |
|  | *NOX* |
| CRIME\_RATE | 0.00185098 |
| AGE | 0.7314701 |
| INDUS | 0.76365145 |
| NOX | 1 |
| DISTANCE | 0.61144056 |
| TAX | 0.6680232 |
| PTRATIO | 0.18893268 |
| AVG\_ROOM | -0.30218819 |
| LSTAT | 0.59087892 |
| AVG\_PRICE | -0.42732077 |

|  |  |
| --- | --- |
|  | *INDUS* |
| CRIME\_RATE | -0.00551065 |
| AGE | 0.644778511 |
| INDUS | 1 |
| NOX | 0.763651447 |
| DISTANCE | 0.595129275 |
| TAX | 0.72076018 |
| PTRATIO | 0.383247556 |
| AVG\_ROOM | -0.39167585 |
| LSTAT | 0.603799716 |
| AVG\_PRICE | -0.48372516 |

|  |  |
| --- | --- |
|  | *TAX* |
| CRIME\_RATE | -0.0167485 |
| AGE | 0.50645559 |
| INDUS | 0.72076018 |
| NOX | 0.6680232 |
| DISTANCE | 0.91022819 |
| TAX | 1 |
| PTRATIO | 0.46085304 |
| AVG\_ROOM | -0.2920478 |
| LSTAT | 0.54399341 |
| AVG\_PRICE | -0.4685359 |

|  |  |
| --- | --- |
|  | *DISTANCE* |
| CRIME\_RATE | -0.009055 |
| AGE | 0.45602245 |
| INDUS | 0.59512927 |
| NOX | 0.61144056 |
| DISTANCE | 1 |
| TAX | 0.91022819 |
| PTRATIO | 0.46474118 |
| AVG\_ROOM | -0.2098467 |
| LSTAT | 0.48867633 |
| AVG\_PRICE | -0.3816262 |

|  |  |
| --- | --- |
|  | *PTRATIO* |
| CRIME\_RATE | 0.01080059 |
| AGE | 0.26151501 |
| INDUS | 0.38324756 |
| NOX | 0.18893268 |
| DISTANCE | 0.46474118 |
| TAX | 0.46085304 |
| PTRATIO | 1 |
| AVG\_ROOM | -0.3555015 |
| LSTAT | 0.37404432 |
| AVG\_PRICE | -0.5077867 |

|  |  |
| --- | --- |
|  | *AVG\_ROOM* |
| CRIME\_RATE | 0.02739616 |
| AGE | -0.2402649 |
| INDUS | -0.3916759 |
| NOX | -0.3021882 |
| DISTANCE | -0.2098467 |
| TAX | -0.2920478 |
| PTRATIO | -0.3555015 |
| AVG\_ROOM | 1 |
| LSTAT | -0.6138083 |
| AVG\_PRICE | 0.69535995 |

|  |  |
| --- | --- |
|  | *LSTAT* |
| CRIME\_RATE | -0.0423983 |
| AGE | 0.6023385 |
| INDUS | 0.6037997 |
| NOX | 0.5908789 |
| DISTANCE | 0.4886763 |
| TAX | 0.5439934 |
| PTRATIO | 0.3740443 |
| AVG\_ROOM | -0.6138083 |
| LSTAT | 1 |
| AVG\_PRICE | -0.7376627 |

|  |  |
| --- | --- |
|  | *AVG\_PRICE* |
| CRIME\_RATE | 0.04333787 |
| AGE | -0.3769546 |
| INDUS | -0.4837252 |
| NOX | -0.4273208 |
| DISTANCE | -0.3816262 |
| TAX | -0.4685359 |
| PTRATIO | -0.5077867 |
| AVG\_ROOM | 0.69535995 |
| LSTAT | -0.7376627 |
| AVG\_PRICE | 1 |

|  |  |  |
| --- | --- | --- |
| There are directly proportional(positively correlated) which are | | |
| Distance | Tax | 0.910228189 |
| Indus | Nox | 0.763651447 |
| Age | Nox | 0.73147010 |
| Indus | Tax | 0.72076018 |
| AVG\_Room | AVG\_Price | 0.695359947 |
| Nox | Tax | 0.66802320 |
| Age | Indus | 0.644778511 |
| Nox | Distance | 0.611440563 |
| Indus | LSTAT | 0.603799716 |
| Age | LSTAT | 0.602338529 |
| Indus | Distance | 0.595129275 |
| Nox | LSTAT | 0.590878921 |
| Tax | LSTAT | 0.543993412 |
| Age | Tax | 0.506455594 |
| Distance | LSTAT | 0.488676335 |
| Distance | PTRatio | 0.464741179 |
| Tax | PTRatio | 0.460853035 |
| Age | Distance | 0.456022452 |
| Indus | PTRatio | 0.383247556 |
| PTRatio | LSTAT | 0.374044317 |
| Age | PTRatio | 0.261515012 |
| Nox | PTRatio | 0.188932677 |

**Results:**

|  |  |
| --- | --- |
| The Top 3 positively correlated pairs are |  |
| Distance | Tax | 0.91022819 |
| Indus | Nox | 0.76365145 |
| Age | Nox | 0.73147010 |

|  |  |
| --- | --- |
| There are inversely proportional(negatively correlated) which are | |
| Tax | AVG\_Room | -0.29204783 | |
| Tax | AVG\_Price | -0.46853593 | |
| PTRatio | AVG\_Room | -0.35550149 | |
| PTRatio | AVG\_Price | -0.50778669 | |
| Nox | AVG\_Room | -0.30218819 | |
| Nox | AVG\_Price | -0.42732077 | |
| LSTAT | AVG\_Price | -0.73766273 | |
| Indus | AVG\_Room | -0.39167585 | |
| Indus | AVG\_Price | -0.48372516 | |
| Distance | AVG\_Room | -0.20984667 | |
| Distance | AVG\_Price | -0.38162623 | |
| AVG\_Room | LSTAT | -0.61380827 | |
| Age | AVG\_Room | -0.24026493 | |
| Age | AVG\_Price | -0.37695457 | |

**Results:**

|  |  |
| --- | --- |
| The Top 3 negatively correlated pairs are |  |
| Tax | AVG\_Room | -0.2920478 |
| Tax | AVG\_Price | -0.4685359 |
| PTRatio | AVG\_Room | -0.3555015 |

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Regression Analysis | | |  | | |
|  | | |  | | |
| OVERALL FIT | | |  | | |
| Multiple R | | | 0.73766273 | | |
| R Square | | | 0.5441463 | | |
| Adjusted R Square | | | 0.54324183 | | |
| Standard Error | | | 6.21576041 | | |
| Observations | | | 506 | | |
| ANOVA |  |  | |  | Alpha | | 0.05 |  |  |
|  | *df* | *SS* | | *MS* | *F* | | *p-value* | *sig* |  |
| Regression | 1 | 23243.914 | | 23243.914 | 601.617871 | | 5.0811E-88 | yes |  |
| Residual | 504 | 19472.3814 | | 38.6356774 |  | |  |  |  |
| Total | 505 | 42716.2954 | |  |  | |  |  |  |
|  |  |  | |  |  | |  |  |  |
|  | *coeff* | *std err* | | *t stat* | *p-value* | | *lower* | *upper* | *sig* |
| Intercept | 34.5538409 | 0.56262735 | | 61.4151455 | 3.743E-236 | | 33.448457 | 35.6592247 | yes |
| LSTAT | -0.9500494 | 0.03873342 | | -24.5279 | 5.0811E-88 | | -1.0261482 | -0.8739505 | yes |
|  |  |  | |  |  | |  |  |  |
|  |  |  | |  |  | |  |  |  |



* + 1. **What do you infer from the Regression Summary Output in terms of variance explained, coefficient value, Intercept and the Residual plot?**

y = mx + c

Y = -0.950049354X + 34.55384088

The x is the independent variable(LSTAT) and y is the dependent variable(AVG\_Price)

|  |  |
| --- | --- |
| m | -0.95004935 |
| c | 34.5538409 |

|  |  |
| --- | --- |
| variance | 23243.914 |
| coefficient | -0.9500494 |
| Intercept | 34.5538409 |

* + 1. **Is LSTAT variable significant for the analysis based on your model?**

|  |  |
| --- | --- |
| The performance of model is | 54% |
| The unbiased estimation of the model is | 46% |

Therefore, we conclude that the LSTAT is significant under the rejection of null hypothesis

1. **Build another instance of the Regression model but this time including LSTAT and AVG\_ROOM together as Independent variables and AVG\_PRICE as the dependent variable.**

|  |  |
| --- | --- |
| Regression Analysis |  |
|  |  |
| OVERALL FIT |  |
| Multiple R | 0.799100498 |
| R Square | 0.638561606 |
| Adjusted R Square | 0.637124475 |
| Standard Error | 5.540257367 |
| Observations | 506 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  | Alpha | 0.05 |  |
|  | *df* | *SS* | *MS* | *F* | *p-value* | *sig* |
| Regression | 2 | 27276.98621 | 13638.4931 | 444.3308922 | 7.0085E-112 | yes |
| Residual | 503 | 15439.3092 | 30.6944517 |  |  |  |
| Total | 505 | 42716.29542 |  |  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *coeff* | *std err* | *t stat* | *p-value* | *lower* | *upper* | *vif* | *sig* |
| Intercept | -1.358272812 | 3.17282778 | -0.42809535 | 0.668764941 | -7.591900282 | 4.87535466 |  | no |
| AVG\_ROOM | 5.094787984 | 0.4444655 | 11.4627299 | 3.47226E-27 | 4.221550436 | 5.96802553 | 1.60451985 | yes |
| LSTAT | -0.642358334 | 0.043731465 | -14.6886992 | 6.66937E-41 | -0.728277167 | -0.5564395 | 1.60451985 | yes |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| variance | | | 27276.98621 | | | |  |  |  |  |
| coefficients | | | -0.642358334 | | | | 5.094787984 |  |  |  |
| Intercept | | | -1.358272812 | | | |  |  |  |  |
|  | | |  | | | |  |  |  |  |
|  | | |  | | | |  |  |  |  |
| We can say, the AVG\_Room and LSTAT are significant variables,  but y-intercept is not a significant variable interms of predicting AVG\_Price | | | | | | | | | |  |
|  | |  | | |
|  |  | | |  | |
| The performance of model is | | | 64% | | | |
| The unbiased estimation of the model is | | | 36% | | | |

* + 1. **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?**

**Given,**

|  |  |
| --- | --- |
| AVG\_Room (X2) = | 7 |
| LSTAT(X1) = | 20 |
| AVG\_Price(Y) = | ? |

**The multiple linear regression for the two independent & one dependent variable equation would be**

Y = M1X1 + M2X2 + C

Y = -0.642358334X1 + 5.094787984X2 -1.358272812

|  |  |
| --- | --- |
| m1 | -0.642358334 |
| m2 | 5.094787984 |
| c | -1.358272812 |

**By substitute X1, X2 in the above equation, we get**

|  |  |  |
| --- | --- | --- |
| AVG\_Price(Y) = | 21.4580764 | (on an average price of an house) |
|  | $21,458 | (on an average price of houses is calculated by 1000) |

**Therefore, the company quoting a value of 30000 USD for this locality is undercharging**

* + 1. **Is the performance of this model better than the previous model you built in Question 5? Compare in terms of adjusted R-square. Explain.**

**The performance of the model interms of**

|  |  |  |
| --- | --- | --- |
| LSTAT(independent) and AVG\_Price(dependent) | Q5 | 0.54324183 |
| LSTAT, AVG\_Room(independents) and AVG\_Price(dependent) | Q6 | 0.63712448 |

**Therefore, we conclude that previous model accuracy was 54% compared to this model is 64% which implies 9% performance increased compared from the previous model (interms of adjusted r-squared)**

1. **Now, build a Regression model with all variables. AVG\_PRICE shall be the Dependent Variable. Interpret the output in terms of adjusted R-square, coefficient and Intercept values, Significance of variables with respect to AVG\_price. Explain. ​(8 marks)**

|  |  |
| --- | --- |
| Regression Analysis |  |
|  |  |
| OVERALL FIT |  |
| Multiple R | 0.832978824 |
| R Square | 0.69385372 |
| Adjusted R Square | 0.688298647 |
| Standard Error | 5.1347635 |
| Observations | 506 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  | Alpha | 0.05 |  |
|  | *df* | *SS* | *MS* | *F* | *p-value* | *sig* |
| Regression | 9 | 29638.8605 | 3293.20672 | 124.9045 | 1.933E-121 | yes |
| Residual | 496 | 13077.43492 | 26.3657962 |  |  |  |
| Total | 505 | 42716.29542 |  |  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *coeff* | *std err* | *t stat* | *p-value* | *lower* | *upper* | *vif* | *sig* |
| Intercept | 29.241315 | 4.817125596 | 6.07028293 | 2.54E-09 | 19.7768278 | 38.7058027 |  | yes |
| CRIME\_RATE | 0.0487251 | 0.078418647 | 0.62134637 | 0.534657 | -0.10534854 | 0.20279883 | 1.00505888 | no |
| AGE | 0.0327707 | 0.013097814 | 2.50199682 | 0.01267 | 0.00703665 | 0.05850473 | 2.60357319 | yes |
| INDUS | 0.1305514 | 0.063117334 | 2.06839217 | 0.039121 | 0.00654109 | 0.2545617 | 3.59120914 | yes |
| NOX | -10.321183 | 3.894036256 | -2.6505102 | 0.008294 | -17.9720228 | -2.6703428 | 3.899869 | yes |
| DISTANCE | 0.2610936 | 0.067947067 | 3.84260258 | 0.000138 | 0.12759401 | 0.39459314 | 6.70432226 | yes |
| TAX | -0.0144012 | 0.003905158 | -3.68773606 | 0.000251 | -0.02207388 | -0.0067285 | 8.29695485 | yes |
| PTRATIO | -1.0743053 | 0.133601722 | -8.04110406 | 6.59E-15 | -1.33680044 | -0.8118103 | 1.60239117 | yes |
| AVG\_ROOM | 4.1254092 | 0.442758999 | 9.31750493 | 3.89E-19 | 3.25549474 | 4.99532356 | 1.85362878 | yes |
| LSTAT | -0.6034866 | 0.053081161 | -11.3691294 | 8.91E-27 | -0.70777824 | -0.4991949 | 2.75205346 | yes |

The independent variables are X1= CRIME\_RATE, X2 = AGE, X3 = INDUS, X4 = NOX, X5 = DISTANCE, X6 = TAX, X7 = PTRATIO, X8 = AVG\_ROOM, X9 = LSTAT and Y is the dependent variable(AVG\_PRICE)

Y = M1X1 + M2X2 + M3X3 + M4X4 + M5X5 + M6X6 + M7X7 + M8X8 + M9X9 + C

Y = 0.048725141X1 + 0.032770689X2 + 0.130551399X3 -10.3211828X4 + 0.261093575X5 -0.01440119X6 -1.074305348X7 + 4.125409152X8 -0.603486589X9 + 29.24131526

|  |  |
| --- | --- |
| M1 | 0.0487251 |
| M2 | 0.0327707 |
| M3 | 0.1305514 |
| M4 | -10.321183 |
| M5 | 0.2610936 |
| M6 | -0.0144012 |
| M7 | -1.0743053 |
| M8 | 4.1254092 |
| M9 | -0.6034866 |
| C | 29.241315 |

We can say, the CRIME\_RATE is not significant and remaining independent variables are significant enough to predicting AVG\_Price

|  |  |
| --- | --- |
| The performance of model is | 69% |
| The unbiased estimation of the model is | 31% |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| variance | 29638.8605 |  |  |  |  |  |  |  |  |
| coefficients | 0.04872514 | 0.03277069 | 0.1305514 | -10.321183 | 0.26109357 | -0.0144012 | -1.0743053 | 4.12540915 | -0.6034866 |
| Intercept | 29.2413153 |  |  |  |  |  |  |  |  |

1. **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.**

**​(8 marks)**

|  |  |
| --- | --- |
| After dropping the insignificant variable which is CRIME\_RATE, we get regression model | |
| Regression Analysis |  |
|  |  |
| OVERALL FIT |  |
| Multiple R | 0.8328358 |
| R Square | 0.6936154 |
| Adjusted R Square | 0.6886837 |
| Standard Error | 5.1315911 |
| Observations | 506 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  | Alpha | 0.05 |  |
|  | *df* | *SS* | *MS* | *F* | *p-value* | *sig* |
| Regression | 8 | 29628.68142 | 3703.58518 | 140.643 | 1.911E-122 | yes |
| Residual | 497 | 13087.61399 | 26.3332274 |  |  |  |
| Total | 505 | 42716.29542 |  |  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *coeff* | *std err* | *t stat* | *p-value* | *lower* | *upper* | *vif* | *sig* |
| Intercept | 29.428473 | 4.804728624 | 6.12489816 | 1.85E-09 | 19.9883896 | 38.8685574 |  | yes |
| AGE | 0.032935 | 0.013087055 | 2.51660595 | 0.012163 | 0.00722219 | 0.05864773 | 2.6025124 | yes |
| INDUS | 0.13071 | 0.063077823 | 2.07220226 | 0.038762 | 0.00677794 | 0.25464207 | 3.5911504 | yes |
| NOX | -10.272705 | 3.890849222 | -2.64022184 | 0.008546 | -17.9172457 | -2.6281645 | 3.89830346 | yes |
| DISTANCE | 0.2615064 | 0.067901841 | 3.85124202 | 0.000133 | 0.12809638 | 0.39491647 | 6.70368116 | yes |
| TAX | -0.0144523 | 0.003901877 | -3.70394641 | 0.000236 | -0.02211855 | -0.0067861 | 8.29326725 | yes |
| PTRATIO | -1.0717025 | 0.133453529 | -8.03052927 | 7.08E-15 | -1.33390511 | -0.8094998 | 1.6008158 | yes |
| AVG\_ROOM | 4.125469 | 0.44248544 | 9.32340046 | 3.69E-19 | 3.2560963 | 4.99484161 | 1.8536287 | yes |
| LSTAT | -0.6051593 | 0.0529801 | -11.4223884 | 5.42E-27 | -0.70925186 | -0.5010667 | 2.74497496 | yes |

We can say, by dropping the significant variable (CRIME\_RATE) there is no change in the adjusted r-squared and we can conclude enough to predicting AVG\_Price

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| variance | 29628.68142 |  |  |  |  |  |  |  |
| coefficients | 0.03293496 | 0.130710007 | -10.2727051 | 0.261506 | -0.01445235 | -1.0717025 | 4.12546896 | -0.6051593 |
| Intercept | 29.42847349 |  |  |  |  |  |  |  |

* + 1. **Interpret the output of this model.**

|  |  |  |
| --- | --- | --- |
| we can interpret output of the model by | |  |
|  | The performance of model is | 69% |
|  | The unbiased estimation of the model is | 31% |

* + 1. **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?**

The performance of the model interms of

Crime\_rate, Age, Indus, Nox, Distance, Tax, PTRatio, Avg\_Room, LSTAT(independent variables) and AVG\_Price(dependent)

|  |  |
| --- | --- |
| Q7 (Adjusted R-squared) | 69% |
| Q7 (R-square) | 0.6938537 |

Age, Indus, Nox, Distance, Tax, PTRatio, Avg\_Room, LSTAT(independent variables) and AVG\_Price(dependent)

|  |  |
| --- | --- |
| Q8 (Adjusted R-squared) | 69% |
| Q8 (R-square) | 0.6936154 |

Therefore, we conclude that previous model accuracy was 69% compared to this model is 69% there is no change(0.000385035) in comparison from the previous model (interms of adjusted r-squared)

By the dropping the feature (CRIME\_RATE) is no change in the orginal model and there is no change in r-square since if we add or remove feature from the model the r-square will be increasing or decreasing and that too we can conclude there is enough significant variables to predict y variable

* + 1. **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?**

Sorted the coefficient values in a ascending order

|  |  |
| --- | --- |
| M3 for the NOX and AVG\_PRICE | -10.27270508 |
| M6 for the PTRATIO and AVG\_PRICE | -1.071702473 |
| M8 for the LSTAT and AVG\_PRICE | -0.605159282 |
| M5 for the TAX and AVG\_PRICE | -0.014452345 |
| M1 for the AGE and AVG\_PRICE | 0.03293496 |
| M2 for the INDUS and AVG\_PRICE | 0.130710007 |
| M4 for the DISTANCE and AVG\_PRICE | 0.261506423 |
| M7 for the AVG\_ROOM and AVG\_PRICE | 4.125468959 |

the value of NOX is more in this locality of this town then the average price will be less which tends to negative high values

# Hypothesis Testing

H0 : NOX is less in a locality in this town

H1 : NOX is more in a locality in this town

|  |  |
| --- | --- |
| Regression Analysis | |
|  |  |
| OVERALL FIT |  |
| Multiple R | 0.427320772 |
| R Square | 0.182603043 |
| Adjusted R Square | 0.180981223 |
| Standard Error | 8.323347578 |
| Observations | 506 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  | Alpha | 0.05 |  |
|  | *df* | *SS* | *MS* | *F* | *p-value* | *sig* |
| Regression | 1 | 7800.125507 | 7800.12551 | 112.5914803 | 7.065E-24 | yes |
| Residual | 504 | 34916.16991 | 69.2781149 |  |  |  |
| Total | 505 | 42716.29542 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *coeff* | *std err* | *t stat* | *p-value* | *lower* | *upper* |
| Intercept | 41.345874 | 1.811191646 | 22.8279954 | 9.86625E-80 | 37.787459 | 44.90429 |
| NOX | -33.91606 | 3.196337032 | -10.6109133 | 7.06504E-24 | -40.19584 | -27.636269 |

# Y = -33.91605501X + 41.34587447

where Y = Avg\_Price & X = Nox

if the nox value is 0.871 which is max value in the feature 11.8049906

then we reject the null hypothesis

p-value of the nox value is less than 0.05

then we reject the null hypothesis, The average price is greater than nox in the locality of the town

* + 1. **Write the regression equation from this model.**

we can interpret the model by the regression equation

The independent variables are X1 = AGE, X2 = INDUS, X3 = NOX, X4 = DISTANCE, X5 = TAX, X6 = PTRATIO, X7 = AVG\_ROOM, X8 = LSTAT and Y is the dependent variable(AVG\_PRICE)

Y = M1X1 + M2X2 + M3X3 + M4X4 + M5X5 + M6X6 + M7X7 + M8X8 + C

Y = 0.03293496X1 + 0.130710007X2 -10.27270508X3 + 0.261506423X4 -0.014452345X5 -1.071702473X6 + 4.125468959X7 -0.605159282X8 + 29.42847349

|  |  |
| --- | --- |
| M1 for the AGE and AVG\_PRICE | 0.03293496 |
| M2 for the INDUS and AVG\_PRICE | 0.130710007 |
| M3 for the NOX and AVG\_PRICE | -10.2727051 |
| M4 for the DISTANCE and AVG\_PRICE | 0.261506423 |
| M5 for the TAX and AVG\_PRICE | -0.01445235 |
| M6 for the PTRATIO and AVG\_PRICE | -1.07170247 |
| M7 for the AVG\_ROOM and AVG\_PRICE | 4.125468959 |
| M8 for the LSTAT and AVG\_PRICE | -0.60515928 |
| C | 29.42847349 |

1. **Hypothesis Testing of the features (under assumptions where it is affecting the variable or any multicollinearity situations )**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Crime\_Ratio Vs PTRatio |  |  |  |  |  |  |  |
|  | The null hypothesis states that the Teacher to the pupil ratio is low then crime rate is high | | | | | | |
|  |  | H0 : Crime\_Rate ∝ 1 / PTRatio | | | (crime rate is inversely proportional to teacher pupil ratio) | | |
|  |  |  |  |  |  |  |  |
|  | The alternative hypothesis states that the Teacher to the pupil ratio is high then crime rate is low | | | | | | |
|  |  | H1 : PTRatio ∝ 1 / Crime\_Rate | | | (Teacher pupil is inversely proportional to crime rate) | | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Crime\_Ratio Vs Indus |  |  |  |  |  |  |  |
|  | The null hypothesis states that the Crime\_Rate is high then Industry is low | | | | | |  |
|  |  | H0 : Crime\_Rate ∝ 1 / Indus | | | (crime rate is inversely proportional to Industry) | | |
|  |  |  |  |  |  |  |  |
|  | The alternative hypothesis states that the Crime\_Rate is low then Industry is high | | | | | |  |
|  |  | H1 : Indus ∝ 1 / Crime\_Rate | | | (Industry is inversely proportional to Crime\_Rate) | | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Age Vs Avg\_Price |  |  |  |  |  |  |  |
|  | The null hypothesis states that the higher age of the house built prior to 1940 (in percentage) then lower the Avg\_Price | | | | | | |
|  |  | H0 : Age ∝ 1 / Avg\_Price | |  | (age is inversely proportional to Avg\_Price) | | |
|  |  |  |  |  |  |  |  |
|  | The alternative hypothesis states that the higher age of the house built prior to 1940 (in percentage) then higher the Avg\_Price | | | | | | |
|  |  | H1 : Age ∝ Avg\_Price | |  | (age is directly proportional to Avg\_Price) | | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Nox Vs Indus | The null hypothesis states that the higher nox then higher the industry | | | | |  |  |
|  |  | H0 : Nox ∝ Indus | |  | (Nitric oxide is directly proportional to Industry) | | |
|  |  |  |  |  |  |  |  |
|  | The alternative hypothesis states there is no significant difference on nox by the industry | | | | | | |
|  |  | H1 : Indus ∝ 1/ Nox | |  | (Nitric oxide is inversely proportional to Industry) | | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Avg\_Room Vs Avg\_Price |  |  |  |  |  |  |  |
|  | The null hypothesis states that the higher avg\_room then higher the avg\_price | | | | | |  |
|  |  | H0 : Avg\_Room ∝ Avg\_Price | | | (Avg\_Room is directly proportional to Avg\_Price) | | |
|  |  |  |  |  |  |  |  |
|  | The alternative hypothesis states the higher Avg\_Room then Avg\_Price is also higher | | | | | |  |
|  |  | H1 : Avg\_Room ∝ 1 / Avg\_Price | | | (Avg\_Room is inversely proportional to Avg\_Price) | | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Tax Vs Avg\_Price |  |  |  |  |  |  |  |
|  | The null hypothesis states that the higher Tax then lower the avg\_price | | | | |  |  |
|  |  | H0 : Tax ∝ 1/ Avg\_Price | |  | (Tax is inversely proportional to Avg\_Price) | | |
|  |  |  |  |  |  |  |  |
|  | The alternative hypothesis states that the higher Tax then higher the Avg\_Price | | | | | |  |
|  |  | H1 : Tax ∝ Avg\_Price | |  | (Tax is directly proportional to Avg\_Price) | | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| LSTAT Vs Avg\_Price |  |  |  |  |  |  |  |
|  | The null hypothesis states that the higher LSTAT then lower the avg\_price | | | | | |  |
|  |  | H0 : LSTAT ∝ 1/ Avg\_Price | |  | (LSTAT is inversely proportional to Avg\_Price) | | |
|  |  |  |  |  |  |  |  |
|  | The alternative hypothesis states that the higher LSTAT then higher the Avg\_Price | | | | | |  |
|  |  | H1 : LSTAT ∝ Avg\_Price | |  | (LSTAT is directly proportional to Avg\_Price) | | |