Problem: Terro’s real-estate is an agency that estimates the pricing of houses in a certain locality. The pricing is concluded based on different features / factors of a property. This also helps them in identifying the business value of a property. To do this activity the company employs an “Auditor”, who studies various geographic features of a property like pollution level (NOX), crime rate, education facilities (pupil to teacher ratio), connectivity (distance from highway), etc. This helps in determining the price of a property. Objective: To analyze the magnitude of each variable to which it can affect the price of a house in a particular locality.

Objective: To analyze the magnitude of each variable to which it can affect the price of a house in a particular locality.

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

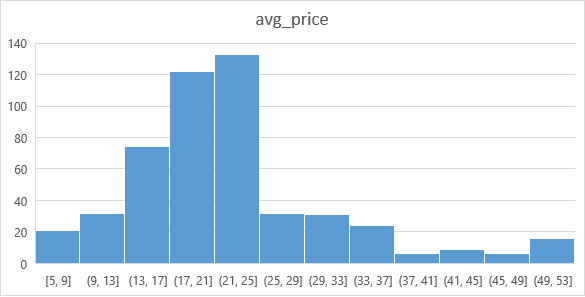
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *CRIME\_RATE* |  | *AGE* |  | *INDUS* |  |
|  |  |  |  |  |  |
| Mean | 4.871976285 | Mean | 68.57490119 | Mean | 11.13677866 |
| Standard Error | 0.129860152 | Standard Error | 1.251369525 | Standard Error | 0.304979888 |
| Median | 4.82 | Median | 77.5 | Median | 9.69 |
| Mode | 3.43 | Mode | 100 | Mode | 18.1 |
| Standard Deviation | 2.921131892 | Standard Deviation | 28.14886141 | Standard Deviation | 6.860352941 |
| Sample Variance | 8.533011532 | Sample Variance | 792.3583985 | Sample Variance | 47.06444247 |
| Kurtosis | -1.189122464 | Kurtosis | -0.967715594 | Kurtosis | -1.233539601 |
| Skewness | 0.021728079 | Skewness | -0.59896264 | Skewness | 0.295021568 |
| Range | 9.95 | Range | 97.1 | Range | 27.28 |
| Minimum | 0.04 | Minimum | 2.9 | Minimum | 0.46 |
| Maximum | 9.99 | Maximum | 100 | Maximum | 27.74 |
| Sum | 2465.22 | Sum | 34698.9 | Sum | 5635.21 |
| Count | 506 | Count | 506 | Count | 506 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *NOX* |  | *DISTANCE* |  | *TAX* |  |
|  |  |  |  |  |  |
| Mean | 0.554695059 | Mean | 9.549407115 | Mean | 408.2372 |
| Standard Error | 0.005151391 | Standard Error | 0.387084894 | Standard Error | 7.492389 |
| Median | 0.538 | Median | 5 | Median | 330 |
| Mode | 0.538 | Mode | 24 | Mode | 666 |
| Standard Deviation | 0.115877676 | Standard Deviation | 8.707259384 | Standard Deviation | 168.5371 |
| Sample Variance | 0.013427636 | Sample Variance | 75.81636598 | Sample Variance | 28404.76 |
| Kurtosis | -0.064667133 | Kurtosis | -0.867231994 | Kurtosis | -1.14241 |
| Skewness | 0.729307923 | Skewness | 1.004814648 | Skewness | 0.669956 |
| Range | 0.486 | Range | 23 | Range | 524 |
| Minimum | 0.385 | Minimum | 1 | Minimum | 187 |
| Maximum | 0.871 | Maximum | 24 | Maximum | 711 |
| Sum | 280.6757 | Sum | 4832 | Sum | 206568 |
| Count | 506 | Count | 506 | Count | 506 |
|  |  |  |  |  |  |
| PTRATIO |  | AVG\_ROOM |  | LSTAT |  |
|  |  |  |  |  |  |
| Mean | 18.4555336 | Mean | 6.284634 | Mean | 12.65306324 |
| Standard Error | 0.096243568 | Standard Error | e | Standard Error | 0.317458906 |
| Median | 19.05 | Median | 6.2085 | Median | 11.36 |
| Mode | 20.2 | Mode | 5.713 | Mode | 8.05 |
| Standard Deviation | 2.164945524 | Standard Deviation | 0.702617 | Standard Deviation | 7.141061511 |
| Sample Variance | 4.686989121 | Sample Variance | 0.493671 | Sample Variance | 50.99475951 |
| Kurtosis | -0.285091383 | Kurtosis | 1.8915 | Kurtosis | 0.493239517 |
| Skewness | -0.802324927 | Skewness | 0.403612 | Skewness | 0.906460094 |
| Range | 9.4 | Range | 5.219 | Range | 36.24 |
| Minimum | 12.6 | Minimum | 3.561 | Minimum | 1.73 |
| Maximum | 22 | Maximum | 8.78 | Maximum | 37.97 |
| Sum | 9338.5 | Sum | 3180.025 | Sum | 6402.45 |
| Count | 506 | Count | 506 | Count | 506 |

|  |  |
| --- | --- |
| *AVG\_PRICE* |  |
|  |  |
| Mean | 22.53280632 |
| Standard Error | 0.408861147 |
| Median | 21.2 |
| Mode | 50 |
| Standard Deviation | 9.197104087 |
| Sample Variance | 84.58672359 |
| Kurtosis | 1.495196944 |
| Skewness | 1.108098408 |
| Range | 45 |
| Minimum | 5 |
| Maximum | 50 |
| Sum | 11401.6 |
| Count | 506 |
|  |  |
|  |  |

|  |  |
| --- | --- |
| OBSERVATIONS |  |
| 1.The total number of data set we are having is 506 | | |  |  |  |  |  |  |  |  |  |
| 2.we observed that negative skewness in age and ptratio, that means observations are trailing off to the left. | | | | | | |  |  |  |  |  |
| 3. we observed that most of them are having negative kurtosis that means sharp peak but in avg\_room,lstat  And avg\_price are having positive kurtosis that means they are having sharp peak. | | | | | | | | | | |  |
| 4. even we can observe the change of mean,mode,standard devition,range,min,max. | | | | |  |  |  |  |  |  |  |

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



OBSERVATIONS

|  |  |  |  |
| --- | --- | --- | --- |
| 1. From histogram we can observe that the maximum houses are from the range $21k to $25k | | | |
| 2. least or minimum range is $37k ,$41k and $45k,$49k |  |  |  |

Question 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 | 790.792 |  |  |  |  |  |  |  |  |
| INDUS | -0.11022 | 124.268 | 46.9714 |  |  |  |  |  |  |  |
| NOX | 0.000625 | 2.38121 | 0.60587 | 0.013401 |  |  |  |  |  |  |
| DISTANCE | -0.22986 | 111.55 | 35.4797 | 0.61571 | 75.6665 |  |  |  |  |  |
| TAX | -8.22932 | 2397.94 | 831.713 | 13.0205 | 1333.12 | 28348.6 |  |  |  |  |
| PTRATIO | 0.068169 | 15.9054 | 5.68085 | 0.047304 | 8.7434 | 167.821 | 4.67773 |  |  |  |
| AVG\_ROOM | 0.056118 | -4.7425 | -1.8842 | -0.02455 | -1.2813 | -34.515 | -0.5397 | 0.4927 |  |  |
| LSTAT | -0.88268 | 120.838 | 29.5218 | 0.48798 | 30.3254 | 653.421 | 5.7713 | -3.07365 | 50.89 |  |
| AVG\_PRICE | 1.162012 | -97.396 | -30.461 | -0.45451 | -30.501 | -724.82 | -10.091 | 4.48457 | -48.35 | 84.4196 |

OBSERVATIONS

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.we can observe that tax variable has high covariance value with each other feature except crime rate.  that means tax explains a very good variability with other features. | | | | | | | | | | | |  |

Question 4 Create a correlation matrix of all the variables (Use Data analysis tool pack).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *CRIME\_RATE* | *AGE* | *INDUS* | *NOX* | *DISTANCE* | *TAX* | *PTRATIO* | *AVG\_ROOM* | *LSTAT* | *AVG\_PRICE* |
| CRIME\_RATE | 1 |  |  |  |  |  |  |  |  |  |
| AGE | 0.0068595 | 1 |  |  |  |  |  |  |  |  |
| INDUS | -0.005511 | 0.644779 | 1 |  |  |  |  |  |  |  |
| NOX | 0.001851 | 0.73147 | 0.76365 | 1 |  |  |  |  |  |  |
| DISTANCE | -0.009055 | 0.456022 | 0.59513 | 0.611441 | 1 |  |  |  |  |  |
| TAX | -0.016749 | 0.506456 | 0.72076 | 0.668023 | 0.91023 | 1 |  |  |  |  |
| PTRATIO | 0.0108006 | 0.261515 | 0.38325 | 0.188933 | 0.46474 | 0.4609 | 1 |  |  |  |
| AVG\_ROOM | 0.0273962 | -0.24026 | -0.3917 | -0.30219 | -0.2098 | -0.292 | -0.3555 | 1 |  |  |
| LSTAT | -0.042398 | 0.602339 | 0.6038 | 0.590879 | 0.48868 | 0.544 | 0.374044 | -0.613808 | 1 |  |
| AVG\_PRICE | 0.0433379 | -0.37695 | -0.4837 | -0.42732 | -0.3816 | -0.4685 | -0.50779 | 0.6953599 | -0.7377 | 1 |

1. Which are the top 3 positively correlated pairs

1.Distance – Tax

2.NOX – Age

3.NOX – Indus

.

b) Which are the top 3 negatively correlated pairs.

1.LSTAT – Avg\_Room

2.Avg\_Price – PTRATIO

3.Avg\_Price – LSTAT

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

OBSERVATIONS

a) What do you infer from the Regression Summary output in terms of variance explained, coefϐicient value, Intercept, and the Residual plot?

|  |
| --- |
| From this model 50 to 55% of the variation in the avg price is explained by the LSTAT plot.  b) Is LSTAT variable signiϐicant for the analysis based on your model? |
| Yes, LSTAT Is significant variable for the avg\_price from this model, P value is less than 0.05 i.e(5.08E-88) |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |

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

A) Write the Regression equation. If a new house in this locality has 7 rooms (on an average) and has a value of 20 for L-STAT, then what will be the value of AVG\_PRICE? How does it compare to the company quoting a value of 30000 USD for this locality? Is the company Overcharging/ Undercharging

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Linear Equation we obtained is [ y = 9.1021x - 34.671 ] |  |  |  |  |  |  |  |  |
| and regression equation is y= -1.358+5.09 X0-0.642 X1  where, y = avg price  X0 = avg\_room,  X1= LSTAT  after calculating the value ,we can say that company is Overcharging  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.  YES, the performance of this model performs well compared to previous model.  where we had R square value is 0.638561 and previous model we had R square value is 0.544 | | | | | | | | |

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 29.24131526 | 4.817125596 | 6.070282926 | 2.53978E-09 |
| CRIME\_RATE | 0.048725141 | 0.078418647 | 0.621346369 | 0.534657201 |
| AGE | 0.032770689 | 0.013097814 | 2.501996817 | 0.012670437 |
| INDUS | 0.130551399 | 0.063117334 | 2.068392165 | 0.03912086 |
| NOX | -10.3211828 | 3.894036256 | -2.650510195 | 0.008293859 |
| DISTANCE | 0.261093575 | 0.067947067 | 3.842602576 | 0.000137546 |
| TAX | -0.01440119 | 0.003905158 | -3.687736063 | 0.000251247 |
| PTRATIO | -1.074305348 | 0.133601722 | -8.041104061 | 6.58642E-15 |
| AVG\_ROOM | 4.125409152 | 0.442758999 | 9.317504929 | 3.89287E-19 |
| LSTAT | -0.603486589 | 0.053081161 | -11.36912937 | 8.91071E-27 |

OBSERVATIONS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| From this we can say that crime rate is not a significant variable for avg price of an house as  P value is greater than 0.5 | | | | |
| NOX TAX, PTRATIO,and LSTAT are negative coefficients. |  |  |  |  |

Question 8) Pick out only the signiϐicant variables from the previous question. Make another instance of the Regression model using only the signiϐicant variables you just picked and answer the questions below:

1. Interpret the output of this model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 29.42847349 | 4.804728624 | 6.124898157 | 1.84597E-09 |
| AGE | 0.03293496 | 0.013087055 | 2.516605952 | 0.012162875 |
| INDUS | 0.130710007 | 0.063077823 | 2.072202264 | 0.038761669 |
| NOX | -10.27270508 | 3.890849222 | -2.640221837 | 0.008545718 |
| DISTANCE | 0.261506423 | 0.067901841 | 3.851242024 | 0.000132887 |
| TAX | -0.014452345 | 0.003901877 | -3.703946406 | 0.000236072 |
| PTRATIO | -1.071702473 | 0.133453529 | -8.030529271 | 7.08251E-15 |
| AVG\_ROOM | 4.125468959 | 0.44248544 | 9.323400461 | 3.68969E-19 |
| LSTAT | -0.605159282 | 0.0529801 | -11.42238841 | 5.41844E-27 |
| from this we can conclude that all the features are significant variables for avg price of the house.  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?   |  |  | | --- | --- | | *Regression Statistics* | | | Multiple R | 0.832835773 | | R Square | 0.693615426 | | Adjusted R Square | 0.688683682 | | Standard Error | 5.131591113 | | Observations | 506 | | | | | | |  |
| By comparing multiple r and r square values for both the models we can conclude that both models perform well | | | | | |  |

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

If NOX is more in the locality , according to this model avg\_price of the house will decrease by 10 times

d) Write the regression equation from this model.

Y=0.03293496 X0 + 0.130710007 X1 -10.27270508 X3 +0.261506423 X4 - 0.014452345 X5 -1.071702473 X6 + 4.125468959 X7 -0.605159282 X8 +29.42847349

Where Y = average\_Price

X0 = Age

X1 = Indus

X2 = NOX

X3 = Distance

X4 = TAX

X5 = PTRATIO

X6 = Avg\_room

X7 = LSTAT

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |