**GREAT LEARNING**

**ASSIGNMENT 2**

**J RANJITH KUMAR**

A picture containing text, screenshot, font, number

Description automatically generated

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *CRIME\_RATE* |  |  | *AGE* |  |
|  |  |  |  |  |
| Mean | 4.871976 |  | Mean | 68.5749 |
| Standard Error | 0.12986 |  | Standard Error | 1.25137 |
| Median | 4.82 |  | Median | 77.5 |
| Mode | 3.43 |  | Mode | 100 |
| Standard Deviation | 2.921132 |  | Standard Deviation | 28.14886 |
| Sample Variance | 8.533012 |  | Sample Variance | 792.3584 |
| Kurtosis | -1.18912 |  | Kurtosis | -0.96772 |
| Skewness | 0.021728 |  | Skewness | -0.59896 |
| Range | 9.95 |  | Range | 97.1 |
| Minimum | 0.04 |  | Minimum | 2.9 |
| Maximum | 9.99 |  | Maximum | 100 |
| Sum | 2465.22 |  | Sum | 34698.9 |
| Count | 506 |  | Count | 506 |

i. Kurtosis value is -1.18912, the curve I. Kurtosis value is -0.96772, the curve is not sharp but represent flat curve. is not sharp but represent flat curve

ii. Skewness is positive ii. Skewness is negative

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| DISTANCE |  |  | *TAX* |  |
|  |  |  |  |  |
| Mean | 9.549407 |  | Mean | 408.2372 |
| Standard Error | 0.387085 |  | Standard Error | 7.492389 |
| Median | 5 |  | Median | 330 |
| Mode | 24 |  | Mode | 666 |
| Standard Deviation | 8.707259 |  | Standard Deviation | 168.5371 |
| Sample Variance | 75.81637 |  | Sample Variance | 28404.76 |
| Kurtosis | -0.86723 |  | Kurtosis | -1.14241 |
| Skewness | 1.004815 |  | Skewness | 0.669956 |
| Range | 23 |  | Range | 524 |
| Minimum | 1 |  | Minimum | 187 |
| Maximum | 24 |  | Maximum | 711 |
| Sum | 4832 |  | Sum | 206568 |
| Count | 506 |  | Count | 506 |

i. Kurtosis value is -0.86723, the curve i. Kurtosis value is -1.14241, the curve is not sharp but represent flat curve. is not sharp but represent flat curve.

ii. Skewness is positive. ii. Skewness is positive.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *LSTAT* |  |  | *AVG\_PRICE* |  |
|  |  |  |  |  |
| Mean | 12.65306 |  | Mean | 22.53281 |
| Standard Error | 0.317459 |  | Standard Error | 0.408861 |
| Median | 11.36 |  | Median | 21.2 |
| Mode | 8.05 |  | Mode | 50 |
| Standard Deviation | 7.141062 |  | Standard Deviation | 9.197104 |
| Sample Variance | 50.99476 |  | Sample Variance | 84.58672 |
| Kurtosis | 0.49324 |  | Kurtosis | 1.495197 |
| Skewness | 0.90646 |  | Skewness | 1.108098 |
| Range | 36.24 |  | Range | 45 |
| Minimum | 1.73 |  | Minimum | 5 |
| Maximum | 37.97 |  | Maximum | 50 |
| Sum | 6402.45 |  | Sum | 11401.6 |
| Count | 506 |  | Count | 506 |

i. Kurtosis value is 0.49324, the curve i. Kurtosis value is 1.495197, the curve is not sharp but represent flat curve. is not sharp but represent flat curve.

ii. Skewness is positive ii. Skewness is positive

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *INDUS* |  |  | *NOX* |  |
|  |  |  |  |  |
| Mean | 11.13678 |  | Mean | 0.554695 |
| Standard Error | 0.30498 |  | Standard Error | 0.005151 |
| Median | 9.69 |  | Median | 0.538 |
| Mode | 18.1 |  | Mode | 0.538 |
| Standard Deviation | 6.860353 |  | Standard Deviation | 0.115878 |
| Sample Variance | 47.06444 |  | Sample Variance | 0.013428 |
| Kurtosis | -1.23354 |  | Kurtosis | -0.06467 |
| Skewness | 0.295022 |  | Skewness | 0.729308 |
| Range | 27.28 |  | Range | 0.486 |
| Minimum | 0.46 |  | Minimum | 0.385 |
| Maximum | 27.74 |  | Maximum | 0.871 |
| Sum | 5635.21 |  | Sum | 280.6757 |
| Count | 506 |  | Count | 506 |

i. Kurtosis value is -1.23354, the curve i. Kurtosis value is -0.06467, the curve is not sharp but represent flat curve. is not sharp but represent flat curve.

ii. Skewness is positive ii. Skewness is positive

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *PTRATIO* |  |  | *AVG\_ROOM* |  |
|  |  |  |  |  |
| Mean | 18.45553 |  | Mean | 6.284634 |
| Standard Error | 0.096244 |  | Standard Error | 0.031235 |
| Median | 19.05 |  | Median | 6.2085 |
| Mode | 20.2 |  | Mode | 5.713 |
| Standard Deviation | 2.164946 |  | Standard Deviation | 0.702617 |
| Sample Variance | 4.686989 |  | Sample Variance | 0.493671 |
| Kurtosis | -0.28509 |  | Kurtosis | 1.8915 |
| Skewness | -0.80232 |  | Skewness | 0.403612 |
| Range | 9.4 |  | Range | 5.219 |
| Minimum | 12.6 |  | Minimum | 3.561 |
| Maximum | 22 |  | Maximum | 8.78 |
| Sum | 9338.5 |  | Sum | 3180.025 |
| Count | 506 |  | Count | 506 |

i. Kurtosis value is -0.28509, the curve i. Kurtosis value is 1.8915, the curve is not sharp but represent flat curve. is not sharp but represent flat curve.

ii. Skewness is negative ii. Skewness is positive

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



* Average price has a positive skewness.

1. Compute the covariance matrix. Share your observations.



* The random values display in covariance matrix. In that values are below or above the average is positive.
* In that another average values is negative

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.



1. Top 3 positive correlated pairs

Tax vs Distance 0.910228

Nox vs INDUS 0.7636651

Nox vs AGE 0.73147

1. Top 3 Negative correlated pairs

AVG\_PRICE vs LSTAT -0.73766

LSTAT vs AVG\_ROOM -0.61381

Avg\_PRICE vs PTRATIO -0.50779

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

1. R square is above 0.5

Coefficient of LSTAT is -0.95005, $1000 increase in average price, 0.95% decrease population.

Intercept value 34.553

Residual plot is equally distributed

|  |  |
| --- | --- |
|  | *P-value* |
| Intercept | 3.7E-236 |
| LSTAT | 5.08E-88 |

1. 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.

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?

AVG\_PRICE = Intercept+ (Coefficient of LSTAT \* value of LSTAT)+ (Coefficient of AVG\_ROOM \* value of AVG\_ROOM)

AVG\_PRICE = -1.35827281187456 +(-0.642358334244129 \* 20) + (5.09478798433655 \* 7) **AVG\_PRICE = 21.4581**

**The Average price is $21.4581 it is over charging, because the company quoted value 30000 USD.**

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.

|  |  |
| --- | --- |
| R Square | 0.544146298 |

R Square 0.637124475

The count of values Independent variable increases.

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.

|  |  |
| --- | --- |
|  | *Coefficients* |
| 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 |

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.

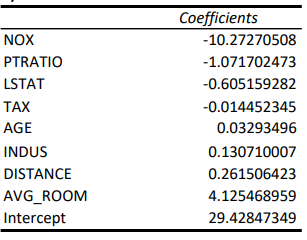
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 29.24132 | 4.817126 | 6.070283 | 2.54E-09 | 19.77683 | 38.7058 | 19.77683 | 38.7058 |
| CRIME\_RATE | 0.048725 | 0.078419 | 0.621346 | 0.534657 | -0.10535 | 0.202799 | -0.10535 | 0.202799 |
| AGE | 0.032771 | 0.013098 | 2.501997 | 0.01267 | 0.007037 | 0.058505 | 0.007037 | 0.058505 |
| INDUS | 0.130551 | 0.063117 | 2.068392 | 0.039121 | 0.006541 | 0.254562 | 0.006541 | 0.254562 |
| NOX | -10.3212 | 3.894036 | -2.65051 | 0.008294 | -17.972 | -2.67034 | -17.972 | -2.67034 |
| DISTANCE | 0.261094 | 0.067947 | 3.842603 | 0.000138 | 0.127594 | 0.394593 | 0.127594 | 0.394593 |
| TAX | -0.0144 | 0.003905 | -3.68774 | 0.000251 | -0.02207 | -0.00673 | -0.02207 | -0.00673 |
| PTRATIO | -1.07431 | 0.133602 | -8.0411 | 6.59E-15 | -1.3368 | -0.81181 | -1.3368 | -0.81181 |
| AVG\_ROOM | 4.125409 | 0.442759 | 9.317505 | 3.89E-19 | 3.255495 | 4.995324 | 3.255495 | 4.995324 |
| LSTAT | -0.60349 | 0.053081 | -11.3691 | 8.91E-27 | -0.70778 | -0.49919 | -0.70778 | -0.49919 |

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?

R Square = 0.68868

R Square = 0.68829

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.

AVG\_PRICE = Intercept + (coefficient of Age \* value of Age) + ( coefficient of Indus \* value of Indus) + (coefficient of NOX \* value of NOX) + (coefficinet of Distance \* value of Distance) + (coefficient of Tax \* value of Tax) + (coefficient of PTRATIO \* value of PTRATIO) + (coefficient of Avg\_room \* value of Avg\_room) + (coefficient of LSTAT \* value of LSTAT)