Terro’s Real Estat Agency

Project Report by Sidhramappa.H

**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?**

To generate the summary statistics for each of the variables, first select the entire table then go to data and data analysis then select the descriptive statistics and also select the summary statistics.

Then we can observe the summary statistics for all the variables of the give data.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *CRIME\_RATE* |  | *AGE* |  | *INDUS* |  | *NOX* |  | *DISTANCE* |  |
|  |  |  |  |  |  |  |  |  |  |
| Mean | 4.871976285 | Mean | 68.57490119 | Mean | 11.13677866 | Mean | 0.554695059 | Mean | 9.549407115 |
| Standard Error | 0.129860152 | Standard Error | 1.251369525 | Standard Error | 0.304979888 | Standard Error | 0.005151391 | Standard Error | 0.387084894 |
| Median | 4.82 | Median | 77.5 | Median | 9.69 | Median | 0.538 | Median | 5 |
| Mode | 3.43 | Mode | 100 | Mode | 18.1 | Mode | 0.538 | Mode | 24 |
| Standard Deviation | 2.921131892 | Standard Deviation | 28.14886141 | Standard Deviation | 6.860352941 | Standard Deviation | 0.115877676 | Standard Deviation | 8.707259384 |
| Sample Variance | 8.533011532 | Sample Variance | 792.3583985 | Sample Variance | 47.06444247 | Sample Variance | 0.013427636 | Sample Variance | 75.81636598 |
| Kurtosis | -1.189122464 | Kurtosis | -0.967715594 | Kurtosis | -1.233539601 | Kurtosis | -0.064667133 | Kurtosis | -0.86723199 |
| Skewness | 0.021728079 | Skewness | -0.59896264 | Skewness | 0.295021568 | Skewness | 0.729307923 | Skewness | 1.004814648 |
| Range | 9.95 | Range | 97.1 | Range | 27.28 | Range | 0.486 | Range | 23 |
| Minimum | 0.04 | Minimum | 2.9 | Minimum | 0.46 | Minimum | 0.385 | Minimum | 1 |
| Maximum | 9.99 | Maximum | 100 | Maximum | 27.74 | Maximum | 0.871 | Maximum | 24 |
| Sum | 2465.22 | Sum | 34698.9 | Sum | 5635.21 | Sum | 280.6757 | Sum | 4832 |
| Count | 506 | Count | 506 | Count | 506 | Count | 506 | Count | 506 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *TAX* |  | *PTRATIO* |  | *AVG\_ROOM* |  | *LSTAT* |  | *AVG\_PRICE* |  |
|  |  |  |  |  |  |  |  |  |  |
| Mean | 408.2371542 | Mean | 18.4555336 | Mean | 6.284634387 | Mean | 12.65306324 | Mean | 22.53280632 |
| Standard Error | 7.492388692 | Standard Error | 0.096243568 | Standard Error | 0.031235142 | Standard Error | 0.317458906 | Standard Error | 0.408861147 |
| Median | 330 | Median | 19.05 | Median | 6.2085 | Median | 11.36 | Median | 21.2 |
| Mode | 666 | Mode | 20.2 | Mode | 5.713 | Mode | 8.05 | Mode | 50 |
| Standard Deviation | 168.5371161 | Standard Deviation | 2.164945524 | Standard Deviation | 0.702617143 | Standard Deviation | 7.141061511 | Standard Deviation | 9.197104087 |
| Sample Variance | 28404.75949 | Sample Variance | 4.686989121 | Sample Variance | 0.49367085 | Sample Variance | 50.99475951 | Sample Variance | 84.58672359 |
| Kurtosis | -1.142407992 | Kurtosis | -0.28509138 | Kurtosis | 1.891500366 | Kurtosis | 0.493239517 | Kurtosis | 1.495196944 |
| Skewness | 0.669955942 | Skewness | -0.80232493 | Skewness | 0.403612133 | Skewness | 0.906460094 | Skewness | 1.108098408 |
| Range | 524 | Range | 9.4 | Range | 5.219 | Range | 36.24 | Range | 45 |
| Minimum | 187 | Minimum | 12.6 | Minimum | 3.561 | Minimum | 1.73 | Minimum | 5 |
| Maximum | 711 | Maximum | 22 | Maximum | 8.78 | Maximum | 37.97 | Maximum | 50 |
| Sum | 206568 | Sum | 9338.5 | Sum | 3180.025 | Sum | 6402.45 | Sum | 11401.6 |
| Count | 506 | Count | 506 | Count | 506 | Count | 506 | Count | 506 |

**Mean:** which is the average number of the variables

**Standard Error:** which is errors of the variables because of their variation.

**Median:** which the middle value of the variables.

**Kurtosis:** which measures the how sharp the peak distribution values are.

**Skewness:** which determines asymmetry of the data.

Positive values indicate the right skewness and negative values indicate left skewness.

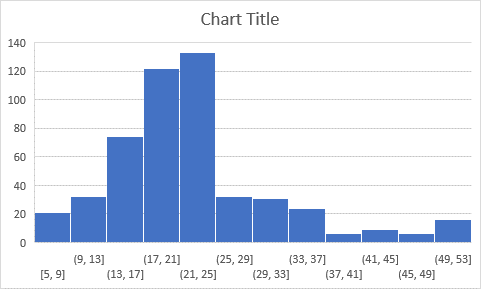
**Minimum:** which gives the minimum number of variables.

**Maximum:** which gives the maximum number of variables.

**Sum:** which calculate the sum of the all the variables.

**2.Plot the histogram of the Avg\_Price Variable. What do you infer?**

To plot the histogram select Avg\_price column then go to recommended charts then go all charts select histogram then click on ok button.



Here the histogram will plot the graph based on frequency range and make a particular number as bins.

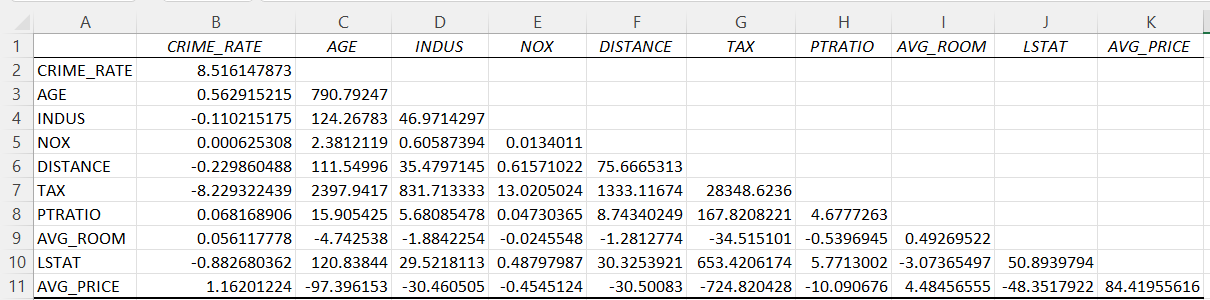
The avg\_price bin of 21 to 24 has more values i.e. 133(highest) =COUNTIFS(Table4[AVG\_PRICE],">=21",Table4[AVG\_PRICE],"<=25")

The avg\_price bin of 37 to 41 has less values i.e. 6(lowest).

=COUNTIFS(Table4[AVG\_PRICE],">=37",Table4[AVG\_PRICE],"<=41")

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

To compute the covariance matrix first we need to click on data then data analysis and select the covariance which gives the covariance matrix between the variables.



The covariance measures the variations between the two variables.

+ve covariance: points are sloping upwards.

-ve covariance: points are sloping downwards.

In the above covariance matrix CRIME\_RATE, AGE, NOX, PTRATIO, having the **positive covariance** between the variables and points are sloping upwards.

In the covariance matrix INDUS, DISTANCE, TAX, AVG\_ROOM, LSTAT, AVG\_PRICE are having the **negative covariance** between the variables and points are sloping upwards.

If the covariance between the any to variables are **zero**, we can say that there is **no relationship** between those variables.

**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.**

sss

We can correlation matrix by first click on data then click on data analysis later selecting the correlation final click on ok.

Table

Description automatically generated

Correlation finds the relationship between the variables.

Correlation is a number between -1 to +1

+1 indicates strong relationship between the two variables.

-1 indicates week relationship between the two variables.

0 indicates no relationship between the variables.

The correlations between Tax and Distance variables has high correlation or strong correlation. Which means Tax and Distance is directly proportional to each other, as distance increases the tax also increases

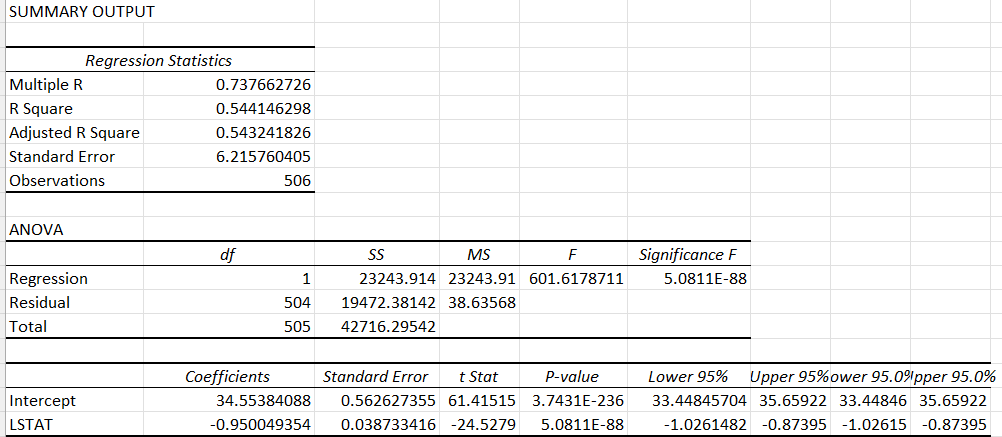
The correlation between Avg\_price and Lstat has less correlation or week correlation. Which means Avg\_price and Lstat are not interrelated each other.

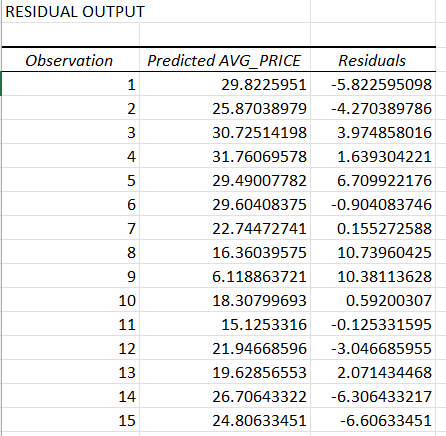
In the above correlation matrix we displayed the top3 positively correlated and top3 negatively correlated values.

**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.**

**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?**





Firstly I have taken the Avg\_price as dependent variable and Lstat as independent variable to build the regression model with data analysis tool and selecting the residual plot for the analysis.

**Variance** which shows the regression error and the residual error.

Significant F indicate that we should except or reject the model, if the significant F is greater than 0.05 we will reject model, if it is less than 0.05 we accept the model.

**Coefficient** values indicate the relationship between the dependent variables and the independent variables.

If the variables having negative values then it is negatively correlated.

If the variables having positive values the it positively correlated.

The size and sign of the values affect the result.

The coefficient value for the Lstat is -0.95004 which means that Lstat is negatively correlated to the Avg\_price.

Intercept this values are generated automatically.

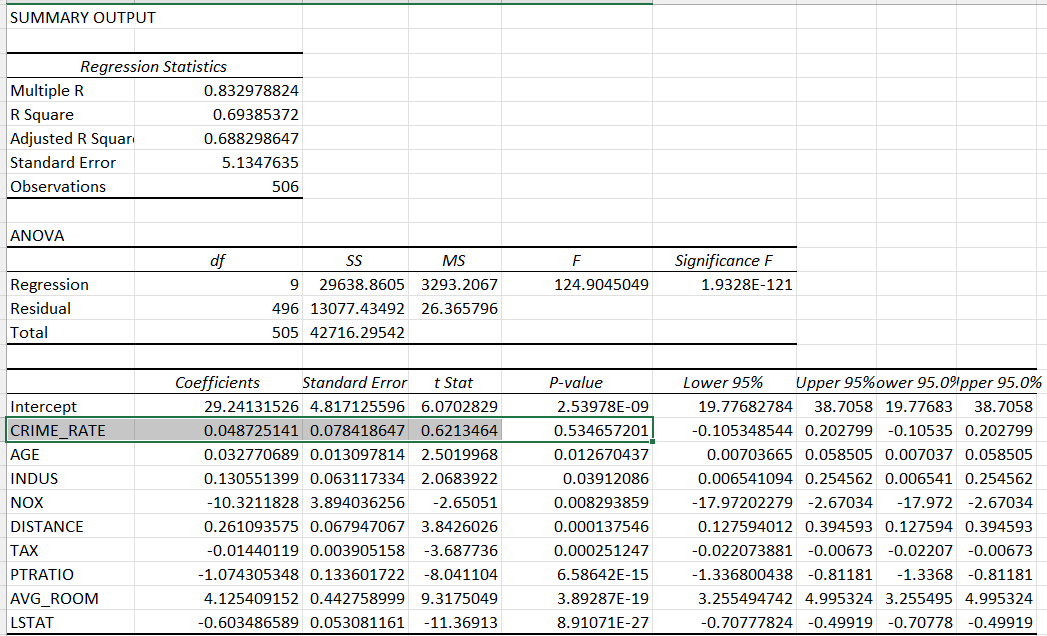
Residual Plot: it indicates the residual error like difference between the actual average price and predicated average price.

Here the residual plot for the Lstat values are positively correlated and which are good to fit the model.

b. **Yes**, LSTAT is significant for the analysis based on the model has its **P-value is less than 0.05**.

**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.**

Here we building the regression model with Avg\_price as dependent variable and remaining all other variables as independent variables.



Adjusted R square values are always less than the R square values, as the model increases the(by adding the new variables) the Adjusted R values also increases.

When the insignigicate variables are added to the model the adjusted r square value decreases.

The Adjusted R square value considers only those independent variables which actually affect the model, and leaving the variables which does not effect the model.

Here Adjusted R square value 0.688298647 significance the moderate effect.

**Coefficient** is the value of of the variables which affect the model. Coefficient changes for each of the variables.

**Intercept** values are the constant values, the slop is zero when the dependent variable is equal to intercept.

**Significance** is used find whether the model is good fit or not. The model is good fit, if the significance value is less than 0.05.

And the P-value is used find whether the particular variables are significant or not, if the P-value is less than 0.05 than it’s a significant variable if it is grater than the 0.05 then its not a significant.

**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.**

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

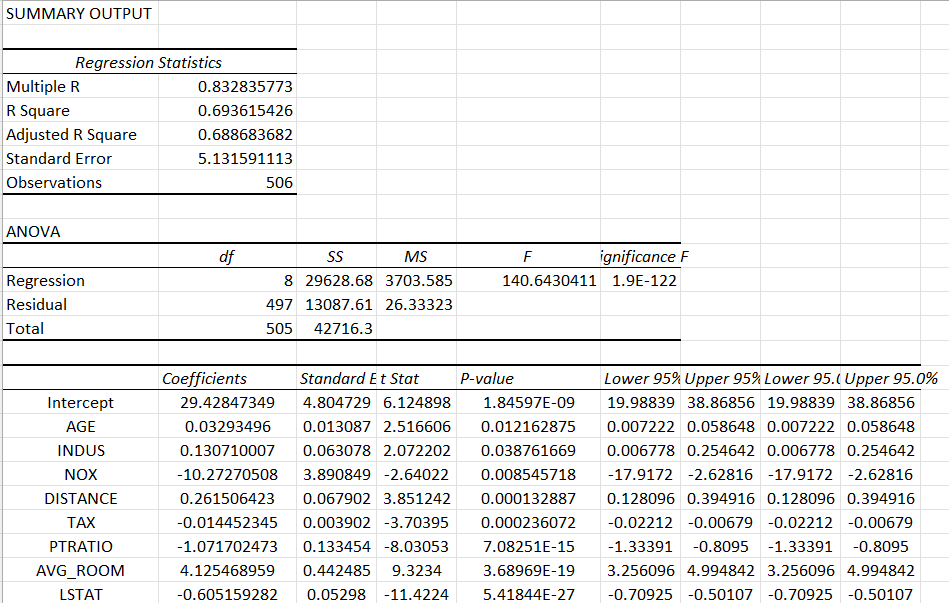
**b. Compare the adjusted R-square value of this model with the model in the previous question, which model performs better according to the value of adjusted R-square?**

**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.**

Now building the regression model with only significant variables, from the previous question analysis based on Adjusted R square, Coefficient, Intercept, Significance and with P-value we can say that **expect CRIME\_RATE** remaining all other variables are significant as there P-value is less than 0.05.

The below table shows the Regression model with only the significant variables.

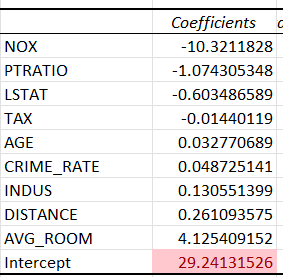
****

Here in this table, we are seeing only the significant variables, here we can observer that the model performance is better than the previous question model as the previous model contains all significant and insignificant variable and here the adjusted R square value is slightly increased compared to previous model.

The Adjusted R squared values are always less than the R square values as the R squared values assumes that all the independent variables affect the result of the model whereas the Adjusted R squared value considers only those independent variables which actually effect the result.

Compared to this model Adjusted R square value with the previous model Adjusted R square value which slightly increased which means by removing the insignificant values the Adjusted R square increase.

* **Sorting the coefficient in ascending order**



The average price decreases as the NOX value increases in locality in this town.

* The Regression equation for the model is

Y=mX + B

Where,

Y is predicated variable

m is the slope of the variable

X is the independent variable

B is the constant.

Y = 29.42 + 0.03\* AGE + 0.12\* INDUS +(-10.27)\*NOX + 0.26\*DISTANCE + (-0.01)\*TAX + (-1.07)\*PTRATIO + 4.12\*AVG\_ROOM + (-0.60)\*LSTAT