rgrergg

**DATA ANALYSIS**

Gurvinder Singh

**Terro’s Real Estate Agency**

**About data set:**

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. The agency has provided a dataset of 506 houses in Boston. There are 10 columns out of that Avg Price is dependent variable and all other variables are independent variable.

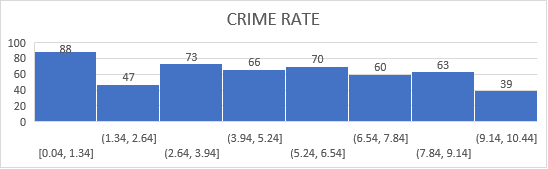
**Problem Statement:**

Summary statistics for each variable in the data.

**Solution:**

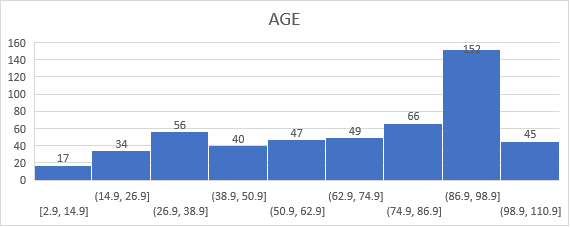
Summary statistics has been done for each variable with the help of Data Analysis Tool pack. The Histograms are also plotted for each variable for better understanding.

**CRIME RATE:** The mean and median are close to each other indicates data has symmetrical distributed. The standard deviation value indicates that values are not widely dispersed from mean. The value of skewness indicates that there are no outliers in data. The value of kurtosis indicates that relatively flat distribution or normal distribution.



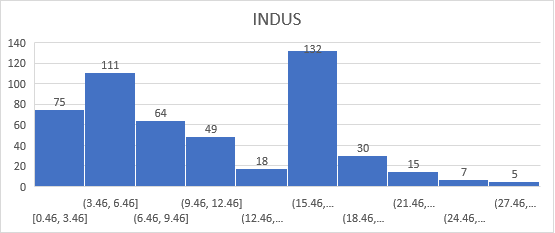
|  |  |
| --- | --- |
| *CRIME\_RATE* |  |
|  |  |
| Mean | 4.871976285 |
| Standard Error | 0.129860152 |
| Median | 4.82 |
| Mode | 3.43 |
| Standard Deviation | 2.921131892 |
| Sample Variance | 8.533011532 |
| Kurtosis | -1.189122464 |
| Skewness | 0.021728079 |
| Range | 9.95 |
| Minimum | 0.04 |
| Maximum | 9.99 |
| Sum | 2465.22 |

**AGE:** The value of variance indicates that data is far away from mean. The data is highly skewed towards left side on graph. The standard deviation indicates that values are widely dispersed from mean. The value of kurtosis indicates relatively flat distribution. The mean and median are far away from each other indicates data is not symmetrical distributed. Although the range of data is relatively good.



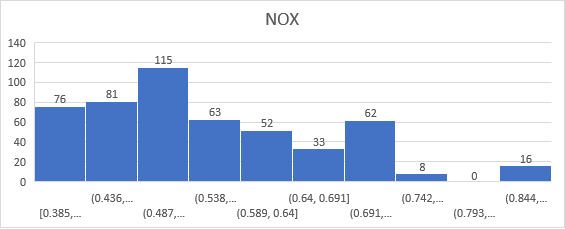
|  |  |
| --- | --- |
| *AGE* |  |
|  |  |
| Mean | 68.57490119 |
| Standard Error | 1.251369525 |
| Median | 77.5 |
| Mode | 100 |
| Standard Deviation | 28.14886141 |
| Sample Variance | 792.3583985 |
| Kurtosis | -0.967715594 |
| Skewness | -0.59896264 |
| Range | 97.1 |
| Minimum | 2.9 |
| Maximum | 100 |
| Sum | 34698.9 |
| Count | 506 |

**INDUS:** The mean and median are close to each other indicates data has symmetrical distributed. The standard deviation value indicates that values are not widely dispersed from mean. The variance indicates data is far away from mean. The mode indicates 18.1 is most repeated value in data. The kurtosis indicates flat distribution. The data is skewed towards right side.



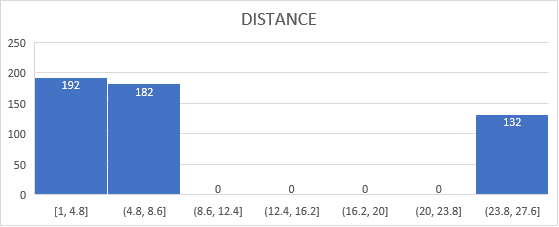
|  |  |
| --- | --- |
| *INDUS* |  |
|  |  |
| Mean | 11.13677866 |
| Standard Error | 0.304979888 |
| Median | 9.69 |
| Mode | 18.1 |
| Standard Deviation | 6.860352941 |
| Sample Variance | 47.06444247 |
| Kurtosis | -1.233539601 |
| Skewness | 0.295021568 |
| Range | 27.28 |
| Minimum | 0.46 |
| Maximum | 27.74 |
| Sum | 5635.21 |
| Count | 506 |

**NOX:** The mean and median are close to each indicates it is symmetrical distributed. The 0.53 is the most repeated value in the data. The range of the data is also good. There are no outliers in data. The data is highly skewed towards right side. The kurtosis indicates flat distribution.



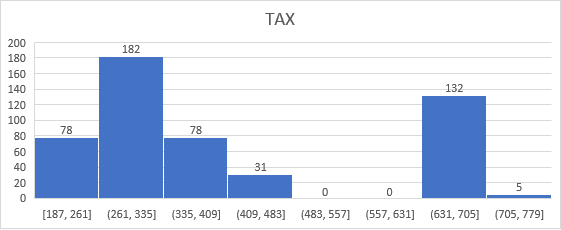
|  |  |
| --- | --- |
| *NOX* |  |
|  |  |
| Mean | 0.554695059 |
| Standard Error | 0.005151391 |
| Median | 0.538 |
| Mode | 0.538 |
| Standard Deviation | 0.115877676 |
| Sample Variance | 0.013427636 |
| Kurtosis | -0.064667133 |
| Skewness | 0.729307923 |
| Range | 0.486 |
| Minimum | 0.385 |
| Maximum | 0.871 |
| Sum | 280.6757 |
| Count | 506 |

**DISTANCE:** The mean and median are not close to each other but not much far away from each other. The value of standard deviation indicates that values are not widely dispersed from mean. The variance indicates data is far away from mean. The data is skewed towards right side. The kurtosis indicates flat distribution.



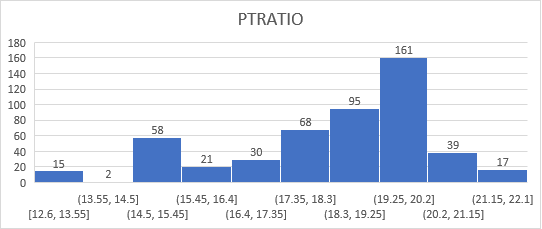
|  |  |
| --- | --- |
| *DISTANCE* |  |
|  |  |
| Mean | 9.549407115 |
| Standard Error | 0.387084894 |
| Median | 5 |
| Mode | 24 |
| Standard Deviation | 8.707259384 |
| Sample Variance | 75.81636598 |
| Kurtosis | -0.867231994 |
| Skewness | 1.004814648 |
| Range | 23 |
| Minimum | 1 |
| Maximum | 24 |
| Sum | 4832 |
| Count | 506 |

**TAX:** The mean and median are far away from each other indicates is not symmetrical distributed. The variance indicates data is too far away from mean. The data is having a good range. The data is skewed towards right side. The kurtosis indicates flat distribution. There are no outliers in data.



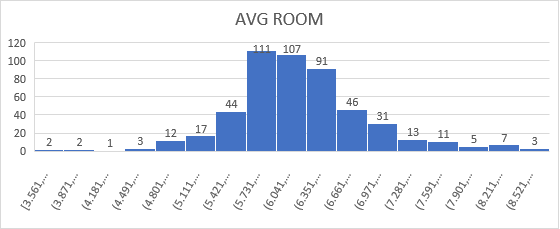
|  |  |
| --- | --- |
| *TAX* |  |
|  |  |
| Mean | 408.2371542 |
| Standard Error | 7.492388692 |
| Median | 330 |
| Mode | 666 |
| Standard Deviation | 168.5371161 |
| Sample Variance | 28404.75949 |
| Kurtosis | -1.142407992 |
| Skewness | 0.669955942 |
| Range | 524 |
| Minimum | 187 |
| Maximum | 711 |
| Sum | 206568 |
| Count | 506 |

**PTRATIO:** The mean and median are close to each other indicates data is symmetrical distributed. The mode indicates 20.2 is the most repeated value in data. The standard indicates values are not widely dispersed from mean. The variance indicates data is not far away from mean. The data is highly skewed towards left side in graphs. The kurtosis indicates flat distribution. There are outliers in data.



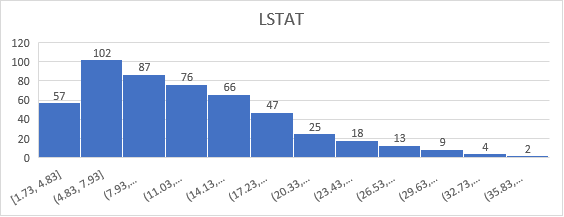
|  |  |
| --- | --- |
| *PTRATIO* |  |
|  |  |
| Mean | 18.4555336 |
| Standard Error | 0.096243568 |
| Median | 19.05 |
| Mode | 20.2 |
| Standard Deviation | 2.164945524 |
| Sample Variance | 4.686989121 |
| Kurtosis | -0.285091383 |
| Skewness | -0.802324927 |
| Range | 9.4 |
| Minimum | 12.6 |
| Maximum | 22 |
| Sum | 9338.5 |
| Count | 506 |

**AVG ROOM:** The mean and median are very close to each other indicates data is symmetrical distributed. The variance indicates data is not far away from the mean. The data is skewed right side on graphs. The kurtosis indicates sharp peak edge. There are many outliers in data. The data is having good range.



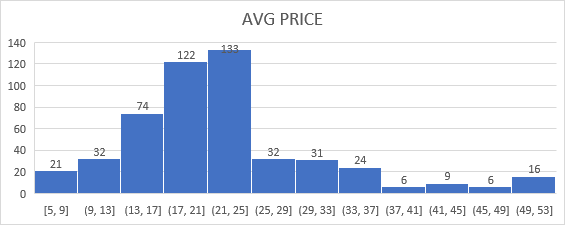
|  |  |
| --- | --- |
| *AVG\_ROOM* |  |
|  |  |
| Mean | 6.284634387 |
| Standard Error | 0.031235142 |
| Median | 6.2085 |
| Mode | 5.713 |
| Standard Deviation | 0.702617143 |
| Sample Variance | 0.49367085 |
| Kurtosis | 1.891500366 |
| Skewness | 0.403612133 |
| Range | 5.219 |
| Minimum | 3.561 |
| Maximum | 8.78 |
| Sum | 3180.025 |
| Count | 506 |

**LSTAT:** The mean and median are close to each other indicates symmetrical distributed. The variance indicates data is away from mean. The standard deviation indicates values are not widely dispersed from mean. The data is highly skewed towards right side. The kurtosis indicates not too sharp peak. There are outliers in data. The data is having a good range.



|  |  |
| --- | --- |
| *LSTAT* |  |
|  |  |
| Mean | 12.65306324 |
| Standard Error | 0.317458906 |
| Median | 11.36 |
| Mode | 8.05 |
| Standard Deviation | 7.141061511 |
| Sample Variance | 50.99475951 |
| Kurtosis | 0.493239517 |
| Skewness | 0.906460094 |
| Range | 36.24 |
| Minimum | 1.73 |
| Maximum | 37.97 |
| Sum | 6402.45 |
| Count | 506 |

**AVG PRICE:** The mean and median are close to each other indicates symmetrical distributed. The mode indicates 50 is the most repeated value in data. The sample variance indicates data is far away from mean. The data is highly skewed towards right side. The kurtosis indicates not too sharp peak edge. There are many outliers in data.



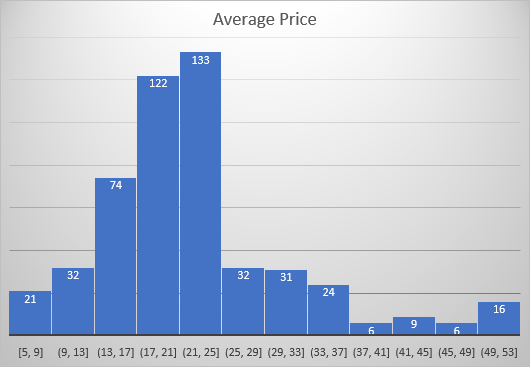
|  |  |
| --- | --- |
| *AVG\_PRICE* |  |
|  |  |
| Mean | 22.53280632 |
| Standard Error | 0.408861147 |
| Median | 21.2 |
| Mode | 50 |
| Standard Deviation | 9.197104087 |
| Sample Variance | 84.58672359 |
| Kurtosis | 1.495196944 |
| Skewness | 1.108098408 |
| Range | 45 |
| Minimum | 5 |
| Maximum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

**Problem Statement:**

Plot a histogram of the Avg\_Price variable and write down observation.

**Solution:**

The Histogram is used for numerical data used to check the shape and spread of data and for better observation in case of single variable.



**Observation:**

* The majority of the prices that we have is in the range of 21-25($1000) followed by 17-21($1000) . The 50% of the property prices are between 17-25($1000).
* The property prices are having a good range from $5000 to $53000.
* The prices of houses depend on -:

- Age of Houses

-Industries near to Town

-Nox (Pollution)

-Distance from highways

-Tax Rate

-Pupil Ratio Ratio by Town

-Status of population

-Crime Rate in Town

-Average Rooms in Houses

**Problem Statement:**

Compute the covariance matrix and write down observation.

**Solution:**

A covariance matrix is square matrix that shows the covariance between many different variables. The covariance has done with the help of Data Analysis Tool pack and in it we have select range of the data.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *ST. CR* | *ST. AGE* | *ST. INDUS* | *ST. NOX* | | *ST. DISTANCE* | *ST. TAX* | *ST. PTRATIO* | *ST. ROOM* | *ST. LSTAT* | *ST. PRICE* |
| ST. CR | 0.998024 |  |  | |  |  |  |  |  |  |  |
| ST. AGE | 0.006846 | 0.998024 |  | |  |  |  |  |  |  |  |
| ST. INDUS | -0.0055 | 0.643504 | 0.998024 | |  |  |  |  |  |  |  |
| ST. NOX | 0.001847 | 0.730025 | 0.762142 | | 0.998024 |  |  |  |  |  |  |
| ST. DISTANCE | -0.00904 | 0.455121 | 0.593953 | | 0.610232 | 0.998024 |  |  |  |  |  |
| ST. TAX | -0.01672 | 0.505455 | 0.719336 | | 0.666703 | 0.908429 | 0.998024 |  |  |  |  |
| ST. PTRATIO | 0.010779 | 0.260998 | 0.38249 | | 0.188559 | 0.463823 | 0.459942 | 0.998024 |  |  |  |
| ST. ROOM | 0.027342 | -0.23979 | -0.3909 | | -0.30159 | -0.20943 | -0.29147 | -0.3548 | 0.998024 |  |  |
| ST. LSTAT | -0.04231 | 0.601148 | 0.602606 | | 0.589711 | 0.487711 | 0.542918 | 0.373305 | -0.6126 | 0.998024 |  |
| ST. PRICE | 0.043252 | -0.37621 | -0.48277 | | -0.42648 | -0.38087 | -0.46761 | -0.50678 | 0.693986 | -0.7362 | 0.998024 |

**Observation:**

* The positive number indicates that the two variables tend to increase or decrease in tandem.
* The negative number indicates that as one variable increases other variable decreases and vice versa.
* Avg Price of houses is our primary data and other are secondary data.
* When the Crime Rate increases by 0.04 the Avg Price of Houses increases. When the Crime Rate decreases the Avg Price of Houses decreases. Same in the case when Avg Rooms increases by 0.69 the Avg Price of Houses increases.
* When the Age of Houses increases by 0.47 the Avg Price of Houses decreases. When the Age of Houses decreases the Avg Price of Houses increases. Same in the case when Indus increases by 0.48 Avg Price of Houses decreases, Nox increases by 0.42 Avg Price of Houses decreases, Distance increases by 0.38 Avg Price of Houses decreases, Tax Rate increases by 0.46 Avg Price of Houses decreases, PTRatio increases by 0.50 Avg Price of Houses decreases and LStat increases by 0.73 Avg Price of Houses decreases.

**Problem Statement:**

Create a correlation matrix of all the variables.

a) Which are the top 3 positively correlated pairs

b) Which are the top 3 negatively correlated pairs.

**Solution:**

A correlation matrix is simply a table which displays the correlation coefficients for different variables. A correlation has done with the help of Data Analysis Tool pack and in it we must select range of data.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *ST. CR* | *ST. AGE* | *ST. INDUS* | *ST. NOX* | *ST. DISTANCE* | *ST. TAX* | *ST. PTRATIO* | *ST. ROOM* | *ST. LSTAT* | *ST. PRICE* |
| ST. CR | 1 |  |  |  |  |  |  |  |  |  |
| ST. AGE | 0.006859 | 1 |  |  |  |  |  |  |  |  |
| ST. INDUS | -0.00551 | 0.644779 | 1 |  |  |  |  |  |  |  |
| ST. NOX | 0.001851 | 0.73147 | 0.763651 | 1 |  |  |  |  |  |  |
| ST. DISTANCE | -0.00906 | 0.456022 | 0.595129 | 0.611441 | 1 |  |  |  |  |  |
| ST. TAX | -0.01675 | 0.506456 | 0.72076 | 0.668023 | 0.910228 | 1 |  |  |  |  |
| ST. PTRATIO | 0.010801 | 0.261515 | 0.383248 | 0.188933 | 0.464741 | 0.460853 | 1 |  |  |  |
| ST. ROOM | 0.027396 | -0.24026 | -0.39168 | -0.30219 | -0.20985 | -0.29205 | -0.3555 | 1 |  |  |
| ST. LSTAT | -0.0424 | 0.602339 | 0.6038 | 0.590879 | 0.488676 | 0.543993 | 0.374044 | -0.61381 | 1 |  |
| ST. PRICE | 0.043338 | -0.37695 | -0.48373 | -0.42732 | -0.38163 | -0.46854 | -0.50779 | 0.69536 | -0.73766 | 1 |

|  |  |
| --- | --- |
| Top 3 Positive Correlated Pairs | Value |
| DISTANCE AND TAX | 0.91 |
| INDUS and NOX | 0.76 |
| AGE and NOX | 0.73 |

|  |  |
| --- | --- |
| Top 3 Negative Correlated Pairs | Value |
| LSTAT and AVG PRICE | -0.74 |
| ROOM and LSTAT | -0.61 |
| PTRATIO and AVG PRICE | -0.51 |

Note- (Positive relation indicates that as one variable increases the other value also increases

and vice versa.)

(Negative relation indicates that as one variable increases the other variable decreases

and vice versa.)

**Problem Statement:**

Build an initial regression model with Avg Price (y) and Lstat (x) and write infer from the Regression Summary output. Also, state that Lstat variable significant for the analysis.

**Solution:**

A linear regression model shows relationship between a dependent variable (x) and one independent variable (y). A linear regression is done with the help of Data Analysis Tool pack. The range must be selected for (y) variable and (x) variable. It is better to display model on new sheet.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Regression Statistics* | | | | | | | | | |
| Multiple R | | | | | 0.737662726 | | | | |
| R Square | | | | | 0.544146298 | | | | |
| Adjusted R Square | | | | | 0.543241826 | | | | |
| Standard Error | | | | | 6.215760405 | | | | |
| Observations | | | | | 506 | | | | |
| ANOVA | |  | |  | | |  | |  | | |  | |
|  | | *df* | | *SS* | | | *MS* | | *F* | | | *Significance F* | |
| Regression | | 1 | | 23243.91 | | | 23243.91 | | 601.6179 | | | 5.08E-88 | |
| Residual | | 504 | | 19472.38 | | | 38.63568 | |  | | |  | |
| Total | | 505 | | 42716.3 | | |  | |  | | |  | |
|  | *Coefficients* | | *Standard Error* | | | *t Stat* | | *P-value* | | | *Lower 95%* | | *Upper 95%* | | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 34.55384 | | 0.562627 | | | 61.41515 | | 3.7E-236 | | | 33.44846 | | 35.65922 | | 33.44846 | 35.65922 |
| LSTAT | -0.95005 | | 0.038733 | | | -24.5279 | | 5.08E-88 | | | -1.02615 | | -0.87395 | | -1.02615 | -0.87395 |

Summary Output

**Infer:**

* This is the regression model of AVG PRICE and LSTAT. We can predict that 54% of the data is accurate by checking R Square value. The P-value for LSTAT is lesser then 0.05 so this mean LSTAT is significant predictor of AVG PRICE. The negative values of residuals indicate that predicted value is high and the positive values indicates that predicted value is low. The negative coefficient of LSTAT suggests that as the LSTAT increases, the AVG PRICE decreases and vice versa. In graph shown above the linear trendline passes through origin.
* LSTAT is significant variable for analysis by checking the P-value in the regression model which is less than 0.05 along with that we are confident that 54% of data is accurate.

Note – (When the sum of residuals is done it give us errors in model.)

**Problem Statement:**

Build a regression model with Avg Price dependent variable and Lstat & Avg Room independent variable. Write down the equation and if there are 7 Rooms and 20 Lstat what will be the value of Avg Price. Compare it with company charging USD 30000. Also, compare the performance of this model with earlier done model.

**Solution:**

The regression model is built where Avg Price is (y) and Lstat is (x1) & Avg Room is (x2). The regression model is made with the help of Data Analysis Tool pack. Then, we must select range for dependent variable (Avg Price) and also for independent variable (Lstat, Avg Room). In the Data Analysis Tool pack along with the above data we can go for residuals, residuals plot in case of need.

Summary Output

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



|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -1.35827 | 3.172828 | -0.4281 | 0.668765 | -7.5919 | 4.875355 | -7.5919 | 4.875355 |
| AVG\_ROOM | 5.094788 | 0.444466 | 11.46273 | 3.47E-27 | 4.22155 | 5.968026 | 4.22155 | 5.968026 |
| LSTAT | -0.64236 | 0.043731 | -14.6887 | 6.67E-41 | -0.72828 | -0.55644 | -0.72828 | -0.55644 |





This model is definitely better than the previous model because in this model we can predict that our 64% of the data is accurate as compared with 54% of the data in earlier model.

**Problem Statement:**

Build a regression model with dependent variable and all independent variables and write infer from the Regression Summary output. Explain the significance of each independent variable with respect to dependent variable.

**Solution:**

The multi linear regression model is build with the help of Data Analysis Tool pack. In the model one variable is dependent variable and other variables are independent variables.

Summary Output

|  |  |  |
| --- | --- | --- |
| *Regression Statistics* | | |
| Multiple R | 0.832979 | |
| R Square | 0.693854 | |
| Adjusted R Square | 0.688299 | |
| Standard Error | 5.134764 | |
| Observations | 506 | |
| ANOVA | |  | |  |  |  |  |
|  | | *df* | | *SS* | *MS* | *F* | *Significance F* |
| Regression | | 9 | | 29638.8605 | 3293.206722 | 124.9045049 | 1.9328E-121 |
| Residual | | 496 | | 13077.43492 | 26.3657962 |  |  |
| Total | | 505 | | 42716.29542 |  |  |  |



**Infer:**

The adjusted R Square is basically modification of R Square. The adjusted R Square takes into account the independent variables used for predicting target variables. We can predict that 68% of the data is accurate by checking Adjusted R Square. The positive coefficient indicates that Crime Rate, Age, Indus, Distance, Avg Room increases the Avg Price increases. The negative coefficient indicates that Nox, Tax, Ptratio, Lstat increases the Avg Price decreases. Intercept is point were function crosses y axis.

By checking the P-value we can conclude that Age, Indus, Nox, Distance, Tax, Ptratio, Avg Room, Lstat is significant predictor of Avg Room. The Crime Rate is not the significant predictor of Avg Price.

**Problem Statement:**

Make another regression model with only significant variables.

1. Write down the infer
2. Compare the adjusted R Square of this model with the model done above.
3. Write the equation of model and sort the values of the Coefficients in ascending order.
4. What will happen to the average price if the value of NOX is more in a locality in this town.

**Solution:**

The multi linear regression model is build with the dependent variable and all the independent variable which are significant predictor of dependent variable (Avg Price). The range must be selected for dependent variable (y) and all the independent variables (x1,x2,x3,\_ \_ \_ \_ xn). We can also plot residuals in summary output.

Summary Output

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Regression Statistics* | | | | |
| Multiple R | | 0.832836 | | |
| R Square | | 0.693615 | | |
| Adjusted R Square | | 0.688684 | | |
| Standard Error | | 5.131591 | | |
| Observations | | 506 | | |
| ANOVA | | |  | | | |  | | |  | |  | | |  | |
|  | | | *df* | | | | *SS* | | | *MS* | | *F* | | | *Significance F* | |
| Regression | | | 8 | | | | 29628.68142 | | | 3703.585178 | | 140.6430411 | | | 1.911E-122 | |
| Residual | | | 497 | | | | 13087.61399 | | | 26.33322735 | |  | | |  | |
| Total | | | 505 | | | | 42716.29542 | | |  | |  | | |  | |
|  | *Coefficients* | | | *Standard Error* | | *t Stat* | | *P-value* | *Lower 95%* | | *Upper 95%* | | *Lower 95.0%* | *Upper 95.0%* | |
| Intercept | 29.42847 | | | 4.804729 | | 6.124898 | | 1.85E-09 | 19.98839 | | 38.86856 | | 19.98839 | 38.86856 | |
| AGE | 0.032935 | | | 0.013087 | | 2.516606 | | 0.012163 | 0.007222 | | 0.058648 | | 0.007222 | 0.058648 | |
| INDUS | 0.13071 | | | 0.063078 | | 2.072202 | | 0.038762 | 0.006778 | | 0.254642 | | 0.006778 | 0.254642 | |
| NOX | -10.2727 | | | 3.890849 | | -2.64022 | | 0.008546 | -17.9172 | | -2.62816 | | -17.9172 | -2.62816 | |
| DISTANCE | 0.261506 | | | 0.067902 | | 3.851242 | | 0.000133 | 0.128096 | | 0.394916 | | 0.128096 | 0.394916 | |
| TAX | -0.01445 | | | 0.003902 | | -3.70395 | | 0.000236 | -0.02212 | | -0.00679 | | -0.02212 | -0.00679 | |
| PTRATIO | -1.0717 | | | 0.133454 | | -8.03053 | | 7.08E-15 | -1.33391 | | -0.8095 | | -1.33391 | -0.8095 | |
| AVG\_ROOM | 4.125469 | | | 0.442485 | | 9.3234 | | 3.69E-19 | 3.256096 | | 4.994842 | | 3.256096 | 4.994842 | |
| LSTAT | -0.60516 | | | 0.05298 | | -11.4224 | | 5.42E-27 | -0.70925 | | -0.50107 | | -0.70925 | -0.50107 | |

* We can predict that 69% of the data is accurate by checking R Square value. All the independent variables are significant predictor of Avg Price. The negative values of residuals indicate that predicted value is high and the positive values indicates that predicted value is low. The negative coefficient of suggests that as the independent variable increases, the AVG PRICE decreases and vice versa.
* The adjusted R Square of both the models are similar. There is very less difference in their values (in decimals). This model performs better than last model in 7th question if we calculate approx. difference.
* Values of coefficients in ascending order:



* If the value of Nox increases in town, then the Avg Price decreases because there is inverse relation between Nox and Avg Price by checking the value of coefficient of Nox

|  |
| --- |
| Regression equation is |
| **y=(m1\*x1) + (m2\*x2) + (-m3\*x3) + (m4\*x4) + (-m5\*x5) + (-m6\*x6) + (m7\*x7) + (-m8\*x8) + c** |

**Thank You**