**TERRO’S REAL ESTATE AGENCY**

Real estate data analysis \_Exploratory analysis, Linear regression

1). **Generate the Summary Statistics for each variable in the table.(Use data analysis tool pack). Write down your observation?**

|  |  |  |
| --- | --- | --- |
| * *1.From the summary Statistics Table having the largest Mean value in Tax column* * *and smallest in NOx column.* | | |
| * *2. From the above table Tax, NOx having the largest, smallest Median value respectively.* |  |  |
| * *3. when compared to all Standard deviation value Tax having the largest among them.* * *NOx having the lowest one.* | | |
| * *4.Skewness having highest value in Average price column and smallest value in Age column.* |  |  |
| * *5.Range of the above column having largest value in Tax column, lowest value in NOx column.* | |  |

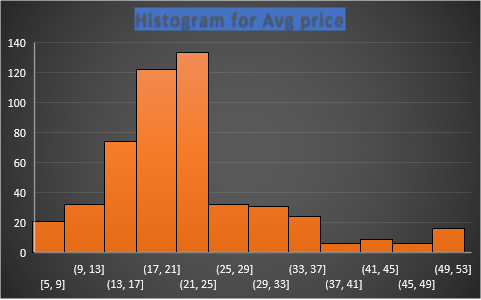
|  |  |  |
| --- | --- | --- |
|  | ***AGE*** |  |
|  |  |  |
| 4.871976285 | Mean | 68.57490119 |
| 0.129860152 | Standard Error | 1.251369525 |
| 4.82 | Median | 77.5 |
| 3.43 | Mode | 100 |
| 2.921131892 | Standard Deviation | 28.14886141 |
| 8.533011532 | Sample Variance | 792.3583985 |
| -1.189122464 | Kurtosis | -0.967715594 |
| 0.021728079 | Skewness | -0.59896264 |
| 9.95 | Range | 97.1 |
| 0.04 | Minimum | 2.9 |
| 9.99 | Maximum | 100 |
| 2465.22 | Sum | 34698.9 |
| 506 | Count | 506 |

|  |  |
| --- | --- |
| ***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 |

|  |  |
| --- | --- |
| ***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 |
| ***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 |

**2) Plot the Histogram of the average Price variable. What do you infer?**

|  |  |
| --- | --- |
|  | **\*). This Histogram of the Average price chart shows that data is positively skewed.** |



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

|  |  |  |
| --- | --- | --- |
| *We can get this Covariance table through Data analysis in Data Bar.* |  |  |
| *This Covariance table shows that Age vs Tax and Indus vs Tax have a direct relation to each other.* | | |
| *Tax vs Average price and Age vs Average price have an inverse relation to each other.* |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | **Covariance Matrix** | | | |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  | ***CRIME\_RATE*** | ***AGE*** | ***INDUS*** | ***NOX*** | ***DISTANCE*** | ***TAX*** | ***PTRATIO*** | ***AVG\_ROOM*** | ***LSTAT*** | ***AVG\_PRICE*** |  |
| **CRIME\_RATE** | 8.516 |  |  |  |  |  |  |  |  |  |  |
| **AGE** | 0.563 | 790.8 |  |  |  |  |  |  |  |  |  |
| **INDUS** | -0.11 | 124.3 | 46.97 |  |  |  |  |  |  |  |  |
| **NOX** | 6E-04 | 2.381 | 0.606 | 0.013 |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **DISTANCE** | -0.23 | 111.5 | 35.48 | 0.616 | 75.67 |  |  |  |  |  |  |
| **TAX** | -8.23 | 2398 | 831.7 | 13.02 | 1333 | 28349 |  |  |  |  |  |
| **PTRATIO** | 0.068 | 15.91 | 5.681 | 0.047 | 8.743 | 167.8 | 4.678 |  |  |  |  |
| **AVG\_ROOM** | 0.056 | -4.74 | -1.88 | -0.02 | -1.28 | -34.5 | -0.54 | 0.493 |  |  |  |
| **LSTAT** | -0.88 | 120.8 | 29.52 | 0.488 | 30.33 | 653.4 | 5.771 | -3.07 | 50.89 |  |  |
| **AVG\_PRICE** | 1.162 | -97.4 | -30.5 | -0.45 | -30.5 | -725 | -10.1 | 4.485 | -48.4 | 84.42 |  |
|  |  |  |  |  |  |  |  |  |  |  |  |

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

a). Which are the top 3 positively correlated pairs and

b). which are the top 3 negatively correlated pairs.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | |  |  |  | | |  |  |  |  |  |  |  |
| **Top 3 positively correlated pairs:** | | | | | | |
|  | |  | | | |  |
| **Tax \*Distance** | | 0.910228189 | | | |  |
| **NOx\*Indus** | | 0.763651447 | | | |  |
| **NOx\*age** | | 0.731470104 | | | |  |
|  | |  | | | |  |

|  |  |  |
| --- | --- | --- |
| **Top 3 Negatively correlated pairs:** | | |
|  |  |  |
| **Av price\*PTRATIO** | -0.507786686 |  |
| **LSTAT\*Av Room** | -0.613808272 |  |
| **Av Price\*LSTAT** | -0.737662726 |  |
|  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  | *Correlation Table* | | |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  | ***CRIME\_RATE*** | ***AGE*** | ***INDUS*** | ***NOX*** | ***DISTANCE*** | ***TAX*** | ***PTRATIO*** | ***AVG\_ROOM*** | ***LSTAT*** | ***AVG\_PRICE*** |
| **CRIME\_RATE** | 1 |  |  |  |  |  |  |  |  |  |
| **AGE** | 0.007 | 1 |  |  |  |  |  |  |  |  |
| **INDUS** | -0.006 | 0.645 | 1 |  |  |  |  |  |  |  |
| **NOX** | 0.002 | 0.731 | 0.76365 | 1 |  |  |  |  |  |  |
| **DISTANCE** | -0.009 | 0.456 | 0.59513 | 0.611441 | 1 |  |  |  |  |  |
| **TAX** | -0.017 | 0.506 | 0.72076 | 0.668023 | 0.9 | 1 |  |  |  |  |
| **PTRATIO** | 0.011 | 0.262 | 0.38325 | 0.188933 | 0.5 | 0.460853035 | 1 |  |  |  |
| **AVG\_ROOM** | 0.027 | -0.24 | -0.3917 | -0.30219 | -0 | -0.292047833 | -0.355501 | 1 |  |  |
| **LSTAT** | -0.042 | 0.602 | 0.6038 | 0.590879 | 0.5 | 0.543993412 | 0.374044 | -0.61381 | 1 |  |
| **AVG\_PRICE** | 0.043 | -0.377 | -0.4837 | -0.42732 | -0 | -0.468535934 | -0.507787 | 0.69536 | -0.73766 | 1 |

***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 outputs in terms of various explained, coefficient value, intercept, and the residual plot?***

***b) is LSTAT variable significant for the analysis based on your model.***

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| a. | From this output summary the variance, coefficient, intercept shows that Intercept and  Average price is positively related | | | | | | | | | | |
|  | and LSTAT and Average price is negatively related. | | | | |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
| b. | The LSTTAT variable and Average price are positively correlated so, it is significant for our analysis. | | | | | | | | | |  |

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

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?

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?

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| a) | ***Regression Equation*** |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  | y=5.0947\*(x1)-0.6423\*(x2-1.3582 | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  | Intercept | 1 | -1.358272812 |  |  |  |  |  |  |
|  | AVG\_ROOM | 7 | 5.094787984 |  |  |  |  |  |  |
|  | LSTAT | 20 | -0.642358334 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  | Average Price is |  | 24.17462202 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  | The company quoting a value of 30000 USD but the calculated value of Average  Price is lesser than the company quotation. | | | | | | | | |

***Hence, we conclude that the company is Overcharging****.*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | | | | |  | |  | |  | |  |
| B) | |  |  |  | **Comparing R Square Value** | | | | |  | |
|  | |  |  |  |  |  | |  | |  | |
| **Qtn 5 R Square Value** | | |  |  |  |  | | **Qtn 6 R square value** | | | |
|  | |  |  |  |  |  | |  | |  | |
| Regression Statistics | |  |  |  |  | Regression Statistics | |  | |  | |
| Multiple R | | 0.737662726 |  |  |  | Multiple R | | 0.799100498 | |  | |
| R Square | | 0.544146298 |  |  |  | R Square | | 0.638561606 | |  | |
| Adjusted R Square | | 0.543241826 |  |  |  | Adjusted R Square | | 0.637124475 | |  | |
| Standard Error | | 6.215760405 |  |  |  | Standard Error | | 5.540257367 | |  | |
| Observations | | 506 |  |  |  | Observations | | 506 | |  | |
|  | |  |  |  |  |  | |  | |  | |
|  | |  |  |  |  |  | |  | |  | |
|  | | *Therefore, when compared to other r square value this is better than others.* | | | | | | | | | |

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 Average price

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 29.24131526 | 4.817125596 | 6.070282926 | 2.53978E-09 | 19.77682784 | 38.70580267 | 19.77682784 | 38.70580267 |
| CRIME\_RATE | 0.048725141 | 0.078418647 | 0.621346369 | 0.534657201 | -0.105348544 | 0.202798827 | -0.105348544 | 0.202798827 |
| AGE | 0.032770689 | 0.013097814 | 2.501996817 | 0.012670437 | 0.00703665 | 0.058504728 | 0.00703665 | 0.058504728 |
| INDUS | 0.130551399 | 0.063117334 | 2.068392165 | 0.03912086 | 0.006541094 | 0.254561704 | 0.006541094 | 0.254561704 |
| NOX | -10.3211828 | 3.894036256 | -2.650510195 | 0.008293859 | -17.97202279 | -2.670342809 | -17.97202279 | -2.670342809 |
| DISTANCE | 0.261093575 | 0.067947067 | 3.842602576 | 0.000137546 | 0.127594012 | 0.394593138 | 0.127594012 | 0.394593138 |
| TAX | -0.01440119 | 0.003905158 | -3.687736063 | 0.000251247 | -0.022073881 | -0.0067285 | -0.022073881 | -0.0067285 |
| PTRATIO | -1.074305348 | 0.133601722 | -8.041104061 | 6.58642E-15 | -1.336800438 | -0.811810259 | -1.336800438 | -0.811810259 |
| AVG\_ROOM | 4.125409152 | 0.442758999 | 9.317504929 | 3.89287E-19 | 3.255494742 | 4.995323561 | 3.255494742 | 4.995323561 |
| LSTAT | -0.603486589 | 0.053081161 | -11.36912937 | 8.91071E-27 | -0.70777824 | -0.499194938 | -0.70777824 | -0.499194938 |

* Comparing the R square values this model is better than the others.
* While Comparing the coefficients and Intercept of Crime rate, age, Indus, distance ,avg room, having the direct relation and others having the inverse relation.

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: (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? 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

|  |  |
| --- | --- |
|  | *Coefficients* |
| Intercept | 29.42847349 |
| AGE | 0.03293496 |
| INDUS | 0.130710007 |
| NOX | -10.27270508 |
| DISTANCE | 0.261506423 |
| TAX | -0.014452345 |
| PTRATIO | -1.071702473 |
| AVG\_ROOM | 4.125468959 |
| LSTAT | -0.605159282 |

*Interpretation of the Significant variable regression output*

|  |  |  |
| --- | --- | --- |
| *This regression value is having greater accuracy than* | | |
| *the other regression value.* |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | *Comparing Adjusted Rsquare* | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
| Qtn 5 |  |  | Qtn 6 |  |  | Qtn 7 |  |  |  | Qtn 8 |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
| Regression Statistics |  |  | Regression Statistics |  |  | Regression Statistics |  |  |  | Regression Statistics |  |
| Multiple R | 0.737662726 |  | Multiple R | 0.799100498 |  | Multiple R | 0.832978824 |  |  | Multiple R | 0.832835773 |
| R Square | 0.544146298 |  | R Square | 0.638561606 |  | R Square | 0.69385372 |  |  | R Square | 0.693615426 |
| Adjusted R Square | 0.543241826 |  | Adjusted R Square | 0.637124475 |  | Adjusted R Square | 0.688298647 |  |  | Adjusted R Square | 0.688683682 |
| Standard Error | 6.215760405 |  | Standard Error | 5.540257367 |  | Standard Error | 5.1347635 |  |  | Standard Error | 5.131591113 |
| Observations | 506 |  | Observations | 506 |  | Observations | 506 |  |  | Observations | 506 |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
| |  |  | | --- | --- | | Qtn 8 |  | |  |  | | Regression Statistics |  | | Multiple R | 0.832835773 | | R Square | 0.693615426 | | Adjusted R Square | 0.688683682 | | Standard Error | 5.131591113 | | Observations | 506 | | **Hence, while comparing the other Adjusted R Square Value**  **This value is more significant with Average Price.** | | | | | | | | |  |  |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***C. In this case, If Nox is increasing the average price will*** | | | | | |  |  |  |  |  |
| ***Decreasing.*** | |  |  |  |  |  |  |  |  |  |
| ***If Nox is reduce, The Average price will Increasing.*** | | | | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| ***D.*** |  | ***Regression Equation*** | | |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  | **y= 0.0329\*(x1)+0.130\*(x2)-10.2727\*(x3)+0.261\*(x4)-0.01445\*(x5)-1.0717\*(x5)+4.125\*(x6)-0.60516** | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |