### **TERRO’S REAL ESTATE AGENCY**

Real estate data analysis \_Exploratory data analysis,Linear regression

**1)Generate the summary statistics for each variable in the table.(use data analysis toolpack).Write down your observation?**

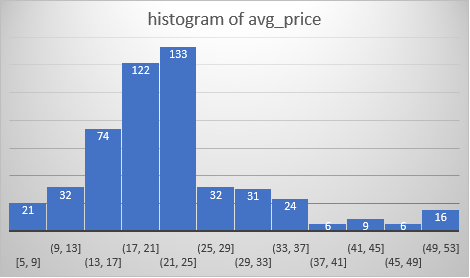
|  |  |  |
| --- | --- | --- |
| * **TAX has the highest mean and NOX has the lowest mean** |  |  |
| * **TAX holds the highest STANDARD DEVIATION and NOX holds the lowest value of standard deviation.** | | |
| * **TAX have the highest range and NOX have the smallest range.** |  |  |
| * **Avg\_room hold the maximum kurtosis INDUS holds the minimum.** | |  |
| * **AVG PRICE has the highest skewness PTRATIO has the lowest skewness.**  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | ***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 |  |  | |  |  | | |  |

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

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

**2)Plot a histogram of the Avg\_price Variable.What do you infer?**

* **This graph is positively skewed**



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

|  |  |
| --- | --- |
| **OBSERVATION:** |  |
| * Age VS Tax and Indus VS Tax have a direct relation to each other. |  |
| * Tax Vs Average price and Age vs Average have an inverse relation to each other. | |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ***CRIME\_RATE*** | ***AGE*** | ***INDUS*** | ***NOX*** | ***DISTANCE*** | ***TAX*** | ***PTRATIO*** | ***AVG\_ROOM*** | ***LSTAT*** | ***AVG\_PRICE*** |
| **CRIME\_RATE** | 8.516147873 |  |  |  |  |  |  |  |  |  |
| **AGE** | 0.562915215 | 790.7924728 |  |  |  |  |  |  |  |  |
| **INDUS** | -0.110215175 | 124.2678282 | 46.97142974 |  |  |  |  |  |  |  |
| **NOX** | 0.000625308 | 2.381211931 | 0.605873943 | 0.013401099 |  |  |  |  |  |  |
| **DISTANCE** | -0.229860488 | 111.5499555 | 35.47971449 | 0.615710224 | 75.66653127 |  |  |  |  |  |
| **TAX** | -8.229322439 | 2397.941723 | 831.7133331 | 13.02050236 | 1333.116741 | 28348.6236 |  |  |  |  |
| **PTRATIO** | 0.068168906 | 15.90542545 | 5.680854782 | 0.047303654 | 8.74340249 | 167.8208221 | 4.677726296 |  |  |  |
| **AVG\_ROOM** | 0.056117778 | -4.74253803 | -1.884225427 | -0.024554826 | -1.281277391 | -34.51510104 | -0.539694518 | 0.492695216 |  |  |
| **LSTAT** | -0.882680362 | 120.8384405 | 29.52181125 | 0.487979871 | 30.32539213 | 653.4206174 | 5.771300243 | -3.073654967 | 50.89397935 |  |
| **AVG\_PRICE** | 1.16201224 | -97.39615288 | -30.46050499 | -0.454512407 | -30.50083035 | -724.8204284 | -10.09067561 | 4.484565552 | -48.35179219 | 84.41955616 |

**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 |

|  |  |
| --- | --- |
| **b)Top 3 negative correlated pairs** | |
| Avg\_price\*Ptratio | -0.507786686 |
| Lstat\*Avg\_room | -0.613808272 |
| Avg\_price\*Lstat | -0.737662726 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *CRIME\_RATE* | *AGE* | *INDUS* | *NOX* | *DISTANCE* | *TAX* | *PTRATIO* | *AVG\_ROOM* | *LSTAT* | *AVG\_PRICE* |
| CRIME\_RATE | 1 |  |  |  |  |  |  |  |  |  |
| AGE | 0.006859463 | 1 |  |  |  |  |  |  |  |  |
| INDUS | -0.005510651 | 0.644778511 | 1 |  |  |  |  |  |  |  |
| NOX | 0.001850982 | 0.731470104 | 0.763651447 | 1 |  |  |  |  |  |  |
| DISTANCE | -0.009055049 | 0.456022452 | 0.595129275 | 0.611440563 | 1 |  |  |  |  |  |
| TAX | -0.016748522 | 0.506455594 | 0.72076018 | 0.6680232 | 0.910228189 | 1 |  |  |  |  |
| PTRATIO | 0.010800586 | 0.261515012 | 0.383247556 | 0.188932677 | 0.464741179 | 0.460853035 | 1 |  |  |  |
| AVG\_ROOM | 0.02739616 | -0.240264931 | -0.391675853 | -0.302188188 | -0.209846668 | -0.292047833 | -0.355501495 | 1 |  |  |
| LSTAT | -0.042398321 | 0.602338529 | 0.603799716 | 0.590878921 | 0.488676335 | 0.543993412 | 0.374044317 | -0.613808272 | 1 |  |
| AVG\_PRICE | 0.043337871 | -0.376954565 | -0.48372516 | -0.427320772 | -0.381626231 | -0.468535934 | -0.507786686 | 0.695359947 | -0.737662726 | 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 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?**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| * Intercept and Avg price is positively related . | | | | |
| * Lstat and avg price is negatively related . | | | | |
|  |  |  |  |  |
| * Lstat variable and avg price are positively correlated | | | | |
| * and thus it is the significant for our analysis | | | | |

|  |  |
| --- | --- |
|  | Coefficients |
| Intercept | 34.55384 |
| LSTAT | -0.95005 |

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

**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=5.095 \*(x)-0.642(x)-1.358 | | | | |  | | |  | | | | | |  | | |  | | | |  | | | |  | | |  |  |
|  | | | | |  | | |  | | | | | |  | | |  | | | |  | | | |  | | |  |  |
| INTERCEPT | | | | | 1 | | | -1.358272812 | | | | | |  | | |  | | | |  | | | |  | | |  |  |
| AVG\_ROOM | | | | | 7 | | | 5.094787984 | | | | | |  | | |  | | | |  | | | |  | | |  |  |
| L\_STAT | | | | | 20 | | | -0.642358334 | | | | | |  | | |  | | | |  | | | |  | | |  |  |
|  | | | | |  | | |  | | | | | |  | | |  | | | |  | | | |  | | |  |  |
|  | | | | |  | | |  | | | | | |  | | |  | | | |  | | | |  | | |  |  |
|  | | | | |  | | |  | | | | | |  | | |  | | | |  | | | |  | | |  |  |
| AVG\_PRICE | | | | | IS | | | 21.45807639 | | | | | |  | | |  | | | |  | | | |  | | |  |  |
|  | | | | |  | | |  | | | | | |  | | |  | | | |  | | | |  | | |  |  |
| 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) QN.5 R-SQUARE VALUE** |  | | |  | | | **VS** | |  | | | **QN.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 | | | | | |
|  |  | | |  | | |  | |  | | |  | | | | | | | |  | | | | | |
|  |  | | |  | | |  | |  | | |  | | | | | | | |  | | | | | |
|  | **The r-square model is better than the previous model value.** | | | | | | | | | | | | | | | | | | |  | | | | | |

**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, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG\_PRICE.**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **OBSERVATION:** | |  |  |  |  |  |  |  |  |  |
|  | a)Comparing the r-square values ,this model is better than others. | | | | | |  |  |  |  |
|  | b)Comparing the coefficient and intercepts of crime rate,age,Indus,distance, | | | | | | |  |  |  |
|  | avg room have direct relation with avg price and the others have inverse relation. | | | | | | | |  |  |
|  | Comparing the p value only crime rate has greater than 0.05 ,except that all are significant variables. | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |

|  |  |
| --- | --- |
|  | *Coefficients* |
| Intercept | 29.24131526 |
| CRIME\_RATE | 0.048725141 |
| AGE | 0.032770689 |
| INDUS | 0.130551399 |
| NOX | -10.3211828 |
| DISTANCE | 0.261093575 |
| TAX | -0.01440119 |
| PTRATIO | -1.074305348 |
| AVG\_ROOM | 4.125409152 |
| LSTAT | -0.603486589 |

|  |  |
| --- | --- |
| *Regression Statistics* | |
| Multiple R | 0.832978824 |
| R Square | 0.69385372 |
| Adjusted R Square | 0.688298647 |
| Standard Error | 5.1347635 |
| Observations | 506 |

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.

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.

This model has greater accuracy than others.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **QN.7 R\_SQUARE VALUE** |  |  | **QN.8 R-SQUARE VALUE** |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| Regression Statistics |  |  | Regression Statistics |  |  |  |  |  |  |  |
| Multiple R | 0.832979 |  | Multiple R | 0.832836 |  |  |  |  |  |  |
| **R Square** | **0.69385** | VS | R Square | 0.693615 |  |  |  |  |  |  |
| Adjusted R Square | 0.688299 |  | Adjusted R Square | 0.688684 |  |  |  |  |  |  |
| Standard Error | 5.134764 |  | Standard Error | 5.131591 |  |  |  |  |  |  |
| Observations | 506 |  | Observations | 506 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  | **THE R-SQUARE VALUE IS LESSER THAN THE PREVIOUS**  **(QN.7 R-SQUARE VALUE).** | | | | | | | |

|  |  |
| --- | --- |
| C) IF NOX IS MORE, THE AVG\_Price will decrease. | |
| IF Nox is less, the AVG\_Price will increase. |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **d) REGRESSION EQUATION** | | |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | Y=0.032935X\_1+0.13071X\_2-10.2727X\_3+0.261506X\_4-0.01445X\_5-1.0717X\_6+4.125469X\_7-0.60516X\_8+29.42847 | | | | | | | | | | |