

**Project Report: Terro’s Real estate Agency**

**Report for assessment of Data Analyst Project**

**Submitted by:**

**Nithin S Shetty**

**Academic Year: 2023-24**

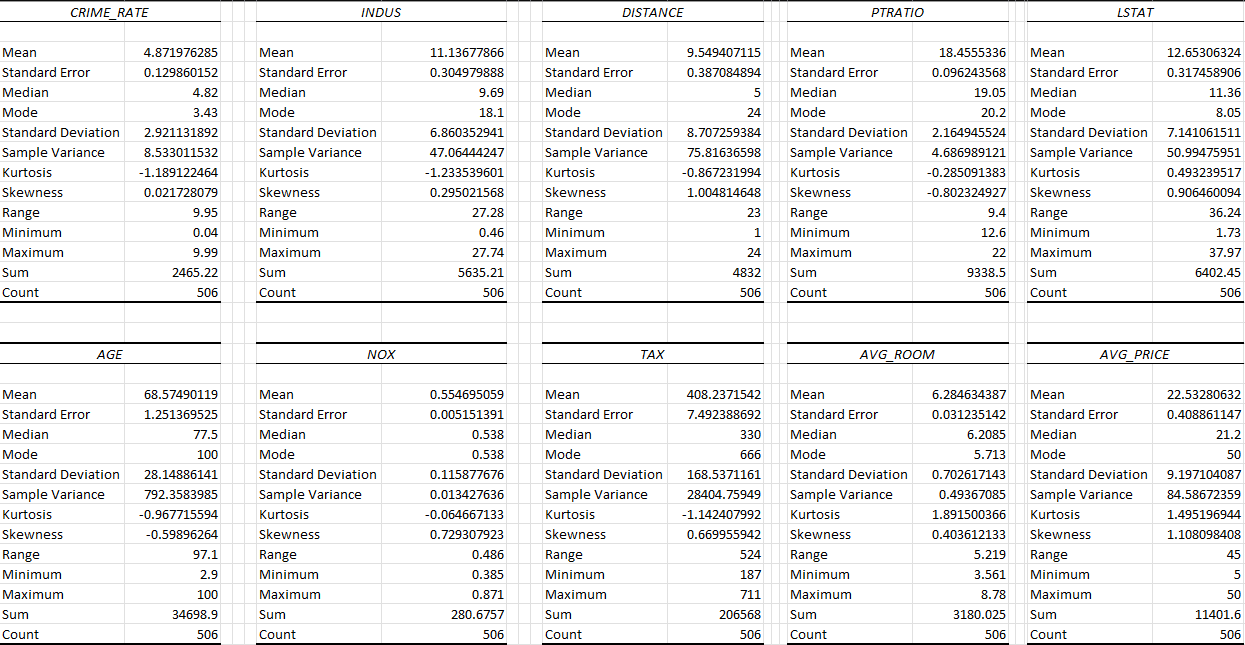
**Department of Data Analyst**

**Great Learning Academy**

**Plot No.758-759,1st Floor,19th Main Rd, Sector 2, HSR Layout**

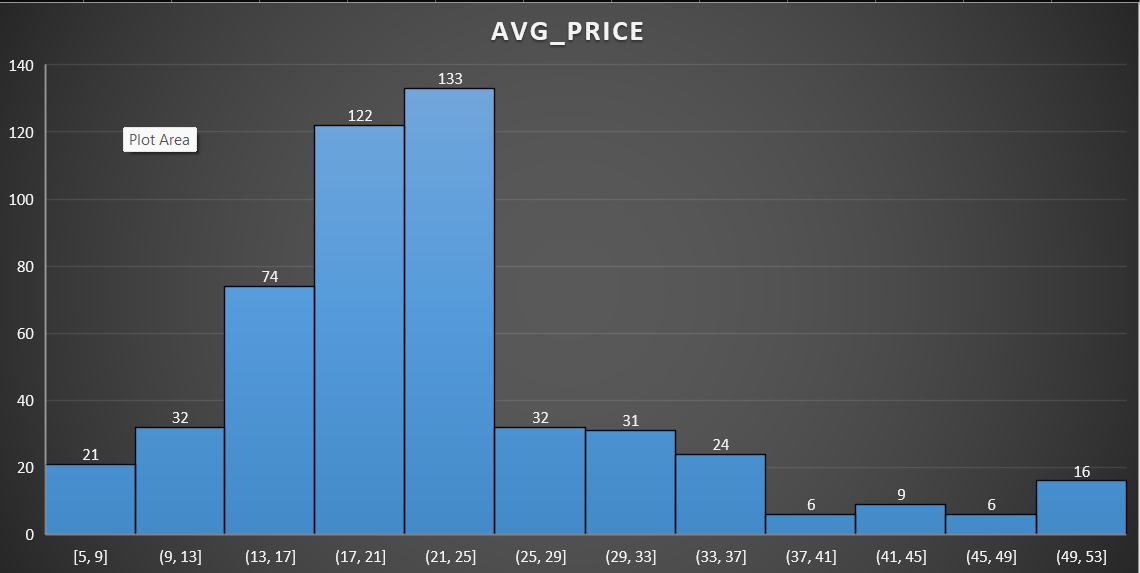
**Bengaluru-560** **102**

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



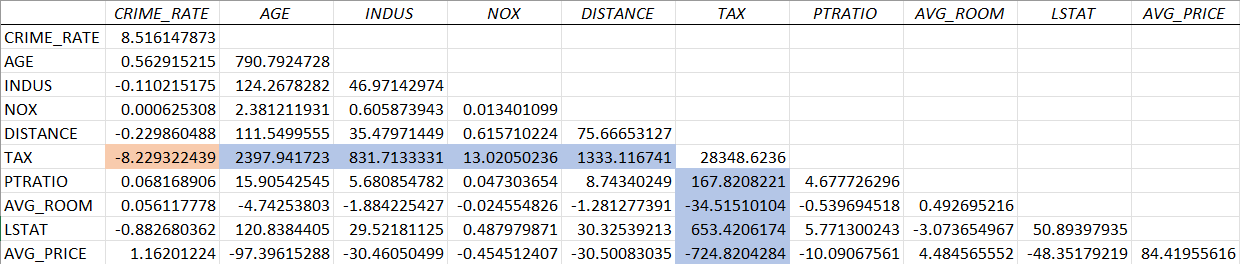
* The number of records given in the dataset are 506.
* The average tax paid is 408.2 and tax range is 524.
* The maximum age is 100 and mode is also 100 which says that most of the houses has age of 100
* The maximum AVG\_PRICE is 50 and mode is also 50 which says that most of the houses has AVG\_PRICE of 50

Question 2. Plot a histogram of the AVG\_PRICE variable. What do you infer?



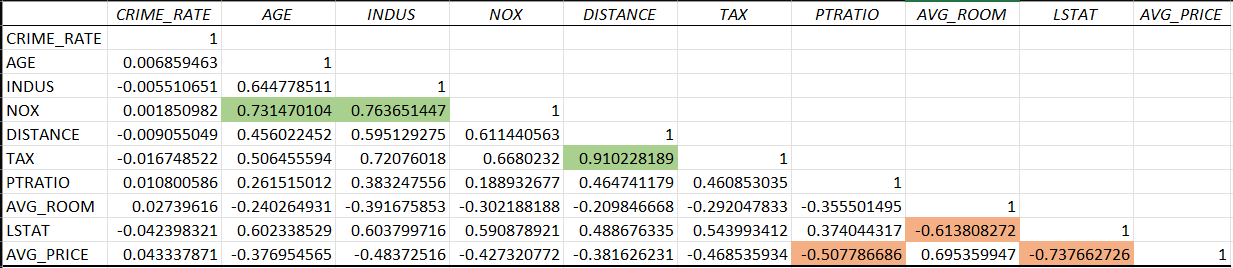
* We can summarise that the majority of houses fall within the price range of 21,000 to 25,000, while there are fewer houses in the price ranges of 37,000 to 41,000 and 45,000 to 49,000.

Question 3. Compute the covariance matrix. Share your observations?



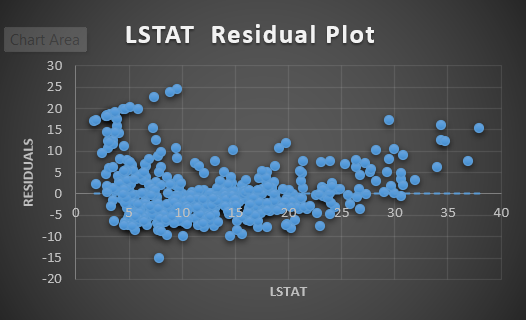
* We can see that tax variable has high covariance values with each other feature except crime rate.

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



* Which are the top 3 positively correlated pairs?
* From above correlation matrix we can analyse the top 3 positively correlated pairs as.
  + 1.NOX - AGE.
  + 2.NOX - INDUS.
  + 3.TAX - DISTANCE.
* Which are the top 3 negatively correlated pairs?
* From above correlation matrix we can analyse the top 3 negatively correlated pairs as.
  + 1.AVG\_PRICE - PTRATIO.
  + 2.LSTAT - AVG\_ROOM.
  + 3.AVG\_PRICE – LSTAT.

Question 5. Build an initial regression model with AVG\_PRICE as ‘y’ (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot?

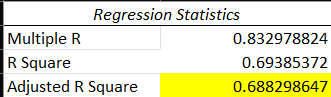


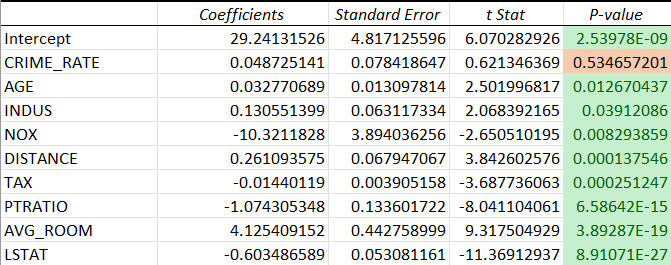
* What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?
* In this model 54% of the variation in the average price is explained by the LSTAT.
* The coefficient of LSTAT for this model is -0.950049354.
* Intercept of LSTAT for the model is 34.55384088.
* Is LSTAT variable significant for the analysis based on your model?
* Yes, LSTAT is significant variable for this model.
* As the p-value we obtained from this model is less than 0.05.

Question 6. Build a new Regression model including LSTAT and AVG\_ROOM together as Independent variables and AVG\_PRICE as dependent variable?

* 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?
* Regression Equation:
* y = -1.358 +5.09 X1 -0.642 X2
* Where y = AVG\_PRICE
* X1 = AVG\_ROOM
* X2 = LSTAT
* AVG\_PRICE for new house can be calculated as:
* Y= -1.358 + 5.09(7) -0.642(20) = 21.45
* The company is Overcharging.
* 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?
* Yes, the performance of this model performed well compared to previous model.
* we can say that 63% of the variance for average price is explained by AVG\_ROOM and LSTAT
* But in previous model LSTAT alone describes 54% of variability for average price.

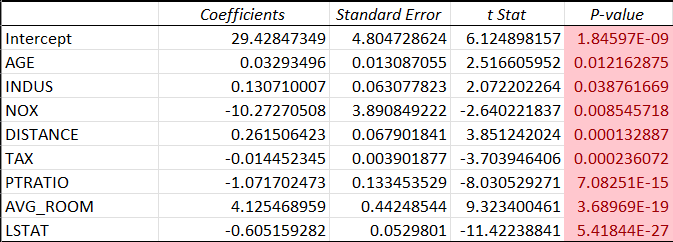
Question 7. Build another Regression model with all variables where AVG\_PRICE alone be the Dependent Variable and all the other variables are independent. Interpret the output in terms of adjusted R square, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG\_PRICE?

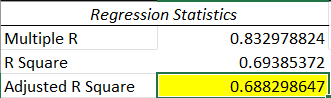




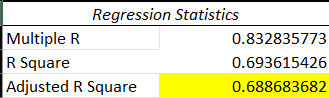
* From this we can say that crime rate is not a significant variable. As the p-value we obtained from this model is greater than 0.05.
* All the multiple features explains 69% of variability for average price of a house.
* In this model where NOX, TAX, PTRATIO and LSTAT have negative coefficients.

Question 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:

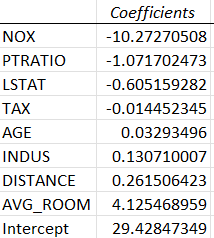
* Interpret the output of this model?
* we can conclude that all the variable are significant variables.
* Compare the adjusted R-square value of this model with the model in the previous question, which model performs better according to the value of adjusted R-square?
* Previous model:



* This model:



* By comparing the adjusted R square values for both the models, we can conclude that the both models are performing well.
* 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?



* The coefficient of NOX is negative, which means that an increases in the value of NOX will lead to a decrease in the average price of houses in a locality.
* Write the regression equation from this model?
* Y=29.42847349 + 0.03293496 X1 + 0.130710007 X2 -10.27270508 X3 +0.261506423 X4 - 0.014452345 X5 -1.071702473 X6 + 4.125468959 X7 -0.605159282 X8
* Where Y = AVG\_PRICE
* X1 = AGE
* X2 = INDUS
* X3 = NOX
* X4 = DISTANCE
* X5 = TAX
* X6 = PTRATIO
* X7 = AVG\_ROOM
* X8 = LSTAT