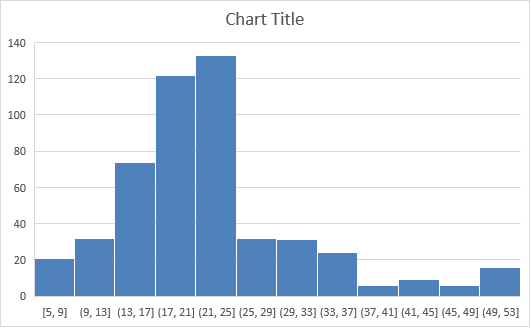
Terro's Real Estate Agency

Anirudh Singh | Data Analysis | 31-12-2023

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

|  |  |
| --- | --- |
| oldest age is 100 and newest age is 2.9 | |
| avg price has the highest skew |  |

Plot a histogram of the Avg\_Price variable. What do you infer?



By seeing this plot, we can say most of the houses are in the range of $17k - $25k

It has a positive skew for avg price, and the price of house starts from $5k till $50k

Compute the covariance matrix. Share your observations.

A white sheet with numbers and text

Description automatically generated with medium confidence

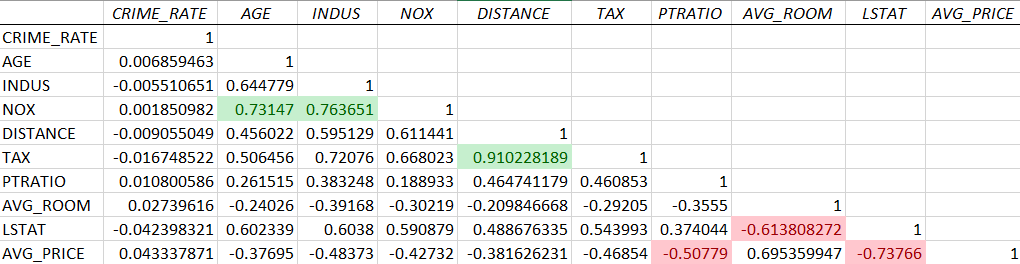
more distance from highway there is less crime rate

more distance from highway there will be less polluted locality

less number of poor people less crime rate

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

a) Which are the top 3 positively correlated pairs and



NOX-AGE, NOX-INDUS, TAX-DISTANCE are top 3 positively correlated pairs

b) Which are the top 3 negatively correlated pairs.

LSTAT-AVG\_ROOM, AVG\_PRICE-PTRATIO, AVG\_PRICE-LSTAT are the top 3 negatively correlated pairs

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?

|  |  |
| --- | --- |
| *Regression Statistics* | |
| Multiple R | 0.737662726 |
| R Square | 0.544146298 |
| Adjusted R Square | 0.543241826 |
| Standard Error | 6.215760405 |
| Observations | 506 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 34.55384088 | 0.562627355 | 61.41514552 | 3.7431E-236 |
| LSTAT | -0.950049354 | 0.038733416 | -24.52789985 | 5.0811E-88 |

The amount of poor people in locality is going to indicate that the houses are going to be less price in that locality

Intercept 34.55

AVG\_PRICE=-0.95\*LSTAT +34.55

b) Is LSTAT variable significant for the analysis based on your model?

Yes, the LSTAT variable is slightly significant

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?

|  |  |
| --- | --- |
| *Regression Statistics* | |
| Multiple R | 0.973885353 |
| R Square | 0.948452681 |
| Adjusted R Square | 0.946366278 |
| Standard Error | 5.53576654 |
| Observations | 506 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 0 | #N/A | #N/A | #N/A |
| AVG\_ROOM | 4.906906071 | 0.070193339 | 69.90557997 | 1.6137E-261 |
| LSTAT | -0.655739993 | 0.030558561 | -21.45847115 | 4.81185E-73 |

AVG\_PRICE = 4.9\*room - 0.655\*LSTAT

=4.9\*7-0.655\*20

$21.2K

Avg price of house is $21.2k and the company is asking for $30k so the company is overcharging.

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.

This model is better than previous model because its adjusted Rsq is 94% while the previous model adjusted Rsq is 54%

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 respect to AVG\_PRICE.

Crime rate is not significant

Increase in age is going to decrease price of house, close to industry increases price more the pollution lesser the price, more distance mean more price, more the tax lesser the price, more children lesser the price, number of rooms increases the price, more the rich people more the price.

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.

|  |  |
| --- | --- |
| *Regression Statistics* | |
| Multiple R | 0.832836 |
| R Square | 0.693615 |
| Adjusted R Square | 0.688684 |
| Standard Error | 5.131591 |
| Observations | 506 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 29.42847 | 4.804729 | 6.124898 | 1.85E-09 |
| AGE | 0.032935 | 0.013087 | 2.516606 | 0.012163 |
| INDUS | 0.13071 | 0.063078 | 2.072202 | 0.038762 |
| NOX | -10.2727 | 3.890849 | -2.64022 | 0.008546 |
| DISTANCE | 0.261506 | 0.067902 | 3.851242 | 0.000133 |
| TAX | -0.01445 | 0.003902 | -3.70395 | 0.000236 |
| PTRATIO | -1.0717 | 0.133454 | -8.03053 | 7.08E-15 |
| AVG\_ROOM | 4.125469 | 0.442485 | 9.3234 | 3.69E-19 |
| LSTAT | -0.60516 | 0.05298 | -11.4224 | 5.42E-27 |

Increase in age is going to decrease price of house, close to industry increases price more the pollution lesser the price, more distance mean more price, more the tax lesser the price, more children lesser the price, number of rooms increases the price, more the rich people more the price.

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?

This model is slightly better than previous model because its adjusted Rsq is 68.868% while the previous model adjusted Rsq is 68.830%

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?

|  |  |
| --- | --- |
|  | *Coefficients* |
| NOX | -10.2727 |
| PTRATIO | -1.0717 |
| LSTAT | -0.60516 |
| TAX | -0.01445 |
| AGE | 0.032935 |
| INDUS | 0.13071 |
| DISTANCE | 0.261506 |
| AVG\_ROOM | 4.125469 |
| Intercept | 29.42847 |

if the value of NOX is more in a locality, then the avg price will decrease

d) Write the regression equation from this model.

Avg\_Price= 0.0329\*Age+0.1307\*Indus-10.2727\*Nox+0.2615\*Distance-0.0144\*Tax-1.0717\*PTRatio+4.1254\*Avg\_Room-0.6051\*LSTAT+29.4284