Project: Terro's Real Estate Agency

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. To do this activity the company employs an “Auditor”, who studies various geographic features of a property like pollution level (NOX), crime rate, education facilities (pupil to teacher ratio), connectivity (distance from highway), etc. This helps in determining the price of a property.

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

From the summary statistics we can find some observations as follows:

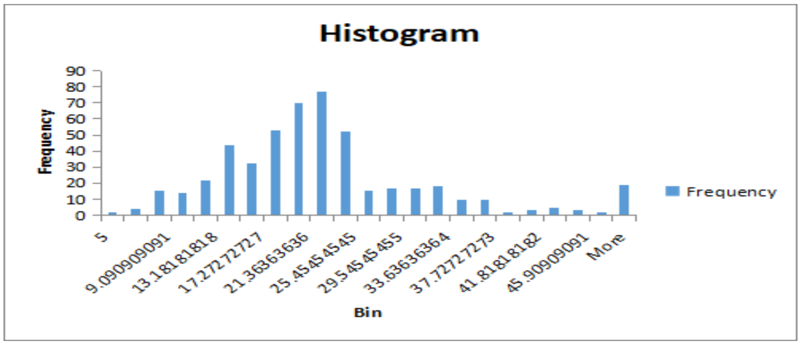
i)The crime rate and NOX is too low. So safe locality.

ii)The average age of house is 100.So the houses are old and strong

iii)The average price of house is 22.53 and the maximum price is 50.

iv)The average tax is 408.23 and the maximum tax is 711.

2) Plot a histogram of the Avg_Price variable. What do you infer?



The average price of the house ranges from $21 to $25

3. Compute the covariance matrix. Share your observations.

The INDUS and NOX has the highest covariance.

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

a) Which are the top 3 positively correlated pairs:

I)Distance – Tax

ii)NOX –Age

Iii)NOX and Indus

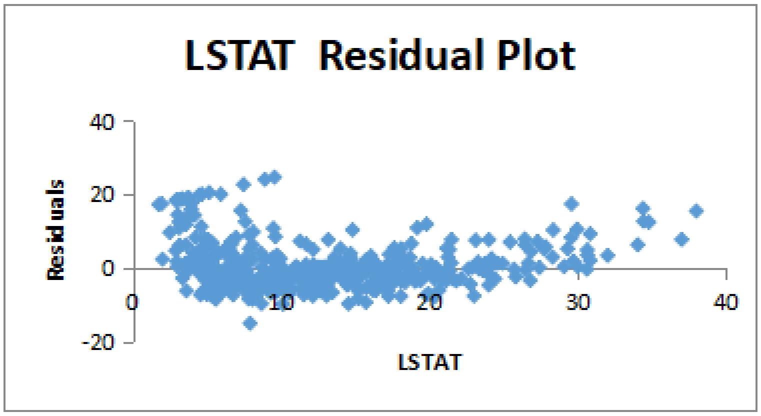
b) Which are the top 3 negatively correlated pairs:

I)LSTAT-Avg_Room

ii)Avg_price – PTRATIO

iii) Avg_price –LSTAT

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 Residual plot?

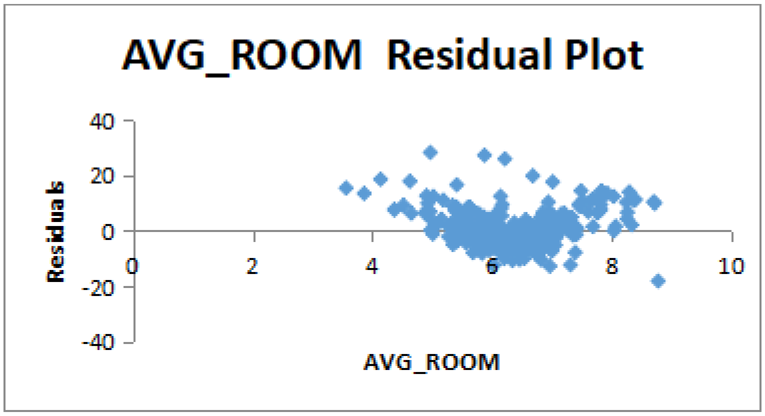
i)coefficient value is –0.950049354

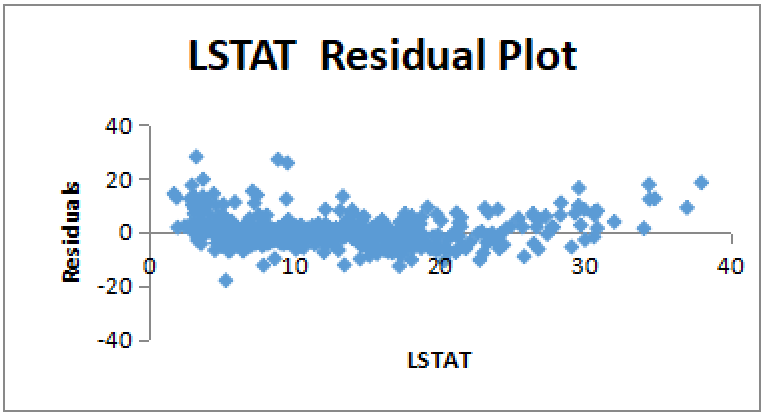
ii)Intercept value is 34.55384088

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

LSTAT is significant because the p value is less than 0.

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?

Regression equation is y=-1.358+5.09x0-0.642x1

Given,

X0=7 & x1 =20

Then the equation will be y=-1.358=5.09(7)-0.642(20)

y=21.44 which means $21440,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

In question 5 R value is 0.54,but now we have 0.638 after having the Avg_room .So this model is more signifivance than the previous model.

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.

	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**

Here the p value for LSTAT is too less than 0.5 but the p value for Crime rate is more than 0.5 ,so the model is not significant.We need to neglect some less coreelated depended variables to get the significant model.

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:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	29.24131526	4.817125596	6.070282926	2.53978E-09
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.534657201

AGE	0.032770689	0.013097814	2.501996817	0.012670437
INDUS	0.130551399	0.063117334	2.068392165	0.03912086
NOX	-10.3211828	3.894036256	-2.650510195	0.008293859
DISTANCE	0.261093575	0.067947067	3.842602576	0.000137546
TAX	-0.01440119	0.003905158	-3.687736063	0.000251247
PTRATIO	-1.074305348	0.133601722	-8.041104061	6.58642E-15
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.89287E-19
LSTAT	-0.603486589	0.053081161	-11.36912937	8.91071E-27

The p value for the Crime rate is more than 0.5, so the model is not signifcant.

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?

Multiple R	0.832978824
R Square	0.69385372
Adjusted R Square	0.688298647

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?

Intercept	-10.3211828
CRIME_RATE	-1.074305348
AGE	-0.603486589
INDUS	-0.01440119
NOX	0.032770689
DISTANCE	0.048725141
TAX	0.130551399
PTRATIO	0.261093575
AVG_ROOM	4.125409152
LSTAT	29.24131526

If NOX is more in locality ,the model average price will decreases by 10 times.

d) Write the regression equation from this model.

Y=0.03293496 (x0) + 0.130710007 (x1)-10.27270508 (x3)+0.261506423(x4)- 0.014452345(x5)-1.071702473(x6)+4.125468959(x7)-0.605159282 (x8)+29.42847349