BUSINESS REPORT

TERRO'S REAL ESTATE AGENCY

Siliveri Mohan

1. The first step to any project is understanding the data. So for this step, generate the summary statistics for each of the variables. What do you observe?

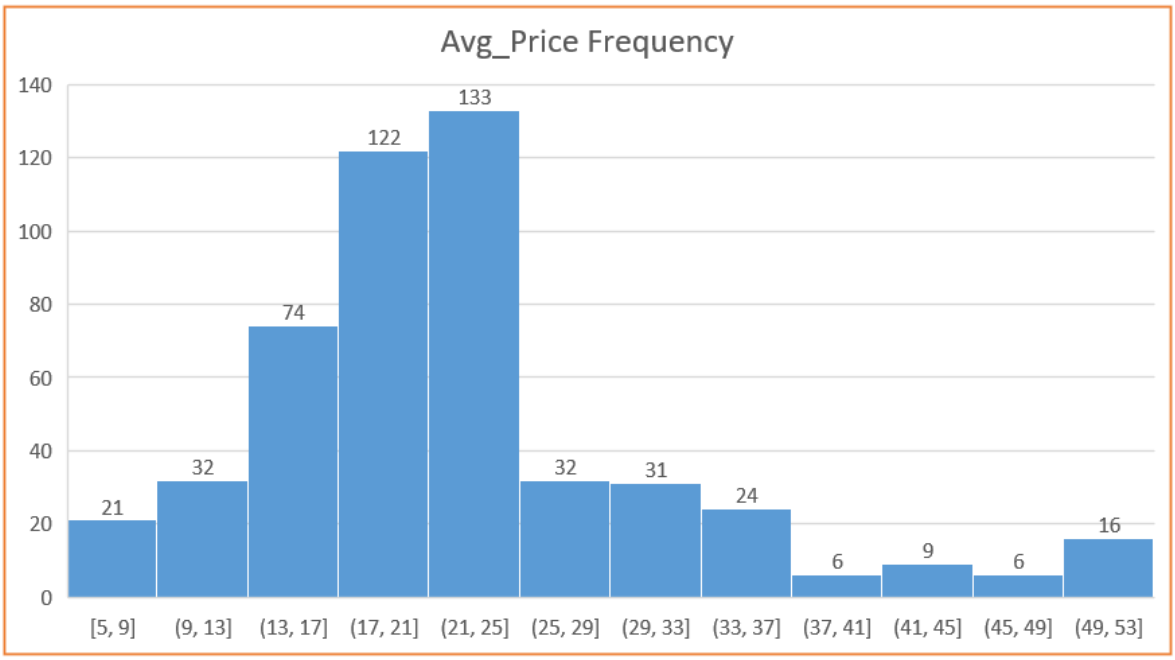
CRIME_RATE		AGE		INDUS		NOX		DISTANCE	
Mean	4.87198	Mean	68.57490119	Mean	11.1368	Mean	0.5547	Mean	9.549407115
Standard Error	0.12986	Standard	1.251369525	Standard	0.30498	Standard Error	0.00515	Standard	0.387084894
Median	4.82	Median	77.5	Median	9.69	Median	0.538	Median	5
Mode	3.43	Mode	100	Mode	18.1	Mode	0.538	Mode	24
Standard Deviation	2.92113	Standard	28.14886141	Standard	6.86035	Standard Deviation	0.11588	Standard	8.707259384
Sample Variance	8.53301	Sample Va	792.3583985	Sample Va	47.0644	Sample Variance	0.01343	Sample Va	75.81636598
Kurtosis	-1.1891	Kurtosis	-0.967715594	Kurtosis	-1.2335	Kurtosis	-0.0647	Kurtosis	-0.867231994
Skewness	0.02173	Skewness	-0.59896264	Skewness	0.29502	Skewness	0.72931	Skewness	1.004814648
Range	9.95	Range	97.1	Range	27.28	Range	0.486	Range	23
Minimum	0.04	Minimum	2.9	Minimum	0.46	Minimum	0.385	Minimum	1
Maximum	9.99	Maximum	100	Maximum	27.74	Maximum	0.871	Maximum	24
Sum	2465.22	Sum	34698.9	Sum	5635.21	Sum	280.676	Sum	4832
Count	506	Count	506	Count	506	Count	506	Count	506

TAX		PTRATIO		AVG_ROOM		LSTAT		AVG_PRICE	
Mean	408.237	Mean	18.4555	Mean	6.28463	Mean	12.6531	Mean	22.5328
Standard	7.49239	Standard Error	0.09624	Standard	0.03124	Standard	0.31746	Standard	0.40886
Median	330	Median	19.05	Median	6.2085	Median	11.36	Median	21.2
Mode	666	Mode	20.2	Mode	5.713	Mode	8.05	Mode	50
Standard	168.537	Standard Deviation	2.16495	Standard	0.70262	Standard	7.14106	Standard	9.1971
Sample Va	28404.8	Sample Variance	4.68699	Sample Va	0.49367	Sample Va	50.9948	Sample Va	84.5867
Kurtosis	-1.1424	Kurtosis	-0.2851	Kurtosis	1.8915	Kurtosis	0.49324	Kurtosis	1.4952
Skewness	0.66996	Skewness	-0.8023	Skewness	0.40361	Skewness	0.90646	Skewness	1.1081
Range	524	Range	9.4	Range	5.219	Range	36.24	Range	45
Minimum	187	Minimum	12.6	Minimum	3.561	Minimum	1.73	Minimum	5
Maximum	711	Maximum	22	Maximum	8.78	Maximum	37.97	Maximum	50
Sum	206568	Sum	9338.5	Sum	3180.03	Sum	6402.45	Sum	11401.6
Count	506	Count	506	Count	506	Count	506	Count	506

Here, we generate summary statistics of every column of the given dataset. The Mean, Median, Standard Deviation, range, minimum, maximum, Skewness ad Kurtosis values of each aspect can be seen in the above tables. We can observe that, an average household price is around 22000$, with age ranging from 2.9 to 100 years.

2. Plot the histogram of the Avg_Price Variable. What do you infer?

Here, we plot the histogram to find the avg_price and frequency of the houses that are shown in below graph. The price of the houses ranging from $21000 to $25000 and the number of houses in this range i.e; frequency is 133. The range between $17000 to $21000 with frequency of 122 houses. The least frequency of the houses is 6 and the price range between $37000 to $41000 and $45000 and $49000.



3. Compute the covariance matrix. Share your observations.

Covariance Matrix

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.516147873									
AGE	0.562915215	790.7925								
INDUS	-0.110215175	124.2678	46.97143							
NOX	0.000625308	2.381212	0.605874	0.013401						
DISTANCE	-0.229860488	111.55	35.47971	0.61571	75.66653					
TAX	-8.229322439	2397.942	831.7133	13.0205	1333.117	28348.62				
PTRATIO	0.068168906	15.90543	5.680855	0.047304	8.743402	167.8208	4.677726			
AVG_ROOM	0.056117778	-4.74254	-1.88423	-0.02455	-1.28128	-34.5151	-0.53969	0.492695		
LSTAT	-0.882680362	120.8384	29.52181	0.48798	30.32539	653.4206	5.7713	-3.07365	50.89398	
AVG_PRICE	1.16201224	-97.3962	-30.4605	-0.45451	-30.5008	-724.82	-10.0907	4.484566	-48.3518	84.41956

The above table represents the covariance of the matrix. Covariance measures the direction of relationship between two variables. The positive covariance means the both variables are trend to high or low at the same time. The negative covariance represents that one variable is high and another variable is low. Here, the tax vs tax increases by 28348 and the Avg_price vs tax goes decreases by -725.

4. Create a correlation matrix of all the variables as shown in the Videos and various case studies. State top 3 positively correlated pairs and top 3 negatively correlated pairs.

The below table represents the correction of the matrix. Correlation is statistical relationship between two entities or variables.

Correlation Matrix

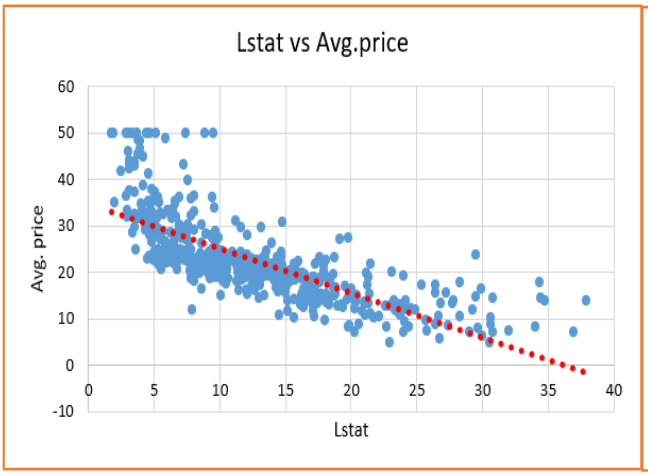
	CRIME_RAT	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	1									
AGE	0.006859	1								
INDUS	-0.00551	0.644779	1							
NOX	0.001851	0.73147	0.763651	1						
DISTANCE	-0.00906	0.456022	0.595129	0.611441	1					
TAX	-0.01675	0.506456	0.72076	0.668023	0.910228	1				
PTRATIO	0.010801	0.261515	0.383248	0.188933	0.464741	0.460853	1			
AVG_ROOM	0.027396	-0.24026	-0.39168	-0.30219	-0.20985	-0.29205	-0.3555	1		
LSTAT	-0.0424	0.602339	0.6038	0.590879	0.488676	0.543993	0.374044	-0.613808272	1	
AVG_PRICE	0.043338	-0.37695	-0.48373	-0.42732	-0.38163	-0.46854	-0.50779	0.695359947	-0.73766	1

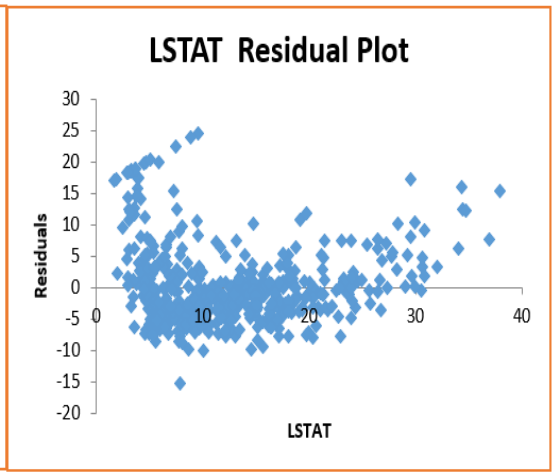
	Top 3 positive	Top 3 negative
1	tax vs distance	avg. price vs Lstat
2	nox vs indus	Lstat vs avg. room
3	nox vs age	avg. prive vs ptratio

The 3 positive correlation pairs are Tax vs Distance, Nox vs Indus and Nox vs Age having correction values are 0.91, 0.76 and 0.73 respectively.

The 3 negative correction pairs are Avg.price vs Lstat, Lstat vs Avg.room and Avg.price vs Ptratio having correction values are -0.73, -0.61 and -0.507 respectively.

5. Build an initial regression model with AVG_PRICE as the y or the Dependent variable and LSTAT variable as the Independent Variable. Generate the residual plot too.
    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?





SUMMARY OUTPUT

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

ANOVA

	df	SS	MS	F	Significance F
Regression	1	23243.914	23243.914	601.6178711	5.0811E-88
Residual	504	19472.38142	38.63567742		
Total	505	42716.29542			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	34.55384088	0.562627355	61.41514552	3.7431E-236	33.44845704	35.65922472	33.44845704	35.65922472
LSTAT	-0.950049354	0.038733416	-24.52789985	5.0811E-88	-1.0261482	-0.873950508	-1.0261482	-0.873950508

Here we use regression model for Avg.price and Lstat variables. The regression summary output along with Scatterplot, Regression Equation and Residual Plot is provided. The coefficient of Lstat is very less ie; -0.95. The intercept value for coefficient is constant for regression equation. If the p-value is less than 0.05, it is significant variable and the p-value is greater than 0.05, it is insignificant variable. In the above regression table, the p-value much less than 0.05, it shows that the LSTAT variable is significant.

6. Build another instance of the Regression model but this time including LSTAT and AVG_ROOM together as independent variables and AVG_PRICE as the 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. Explain.

SUMMARY OUTPUT

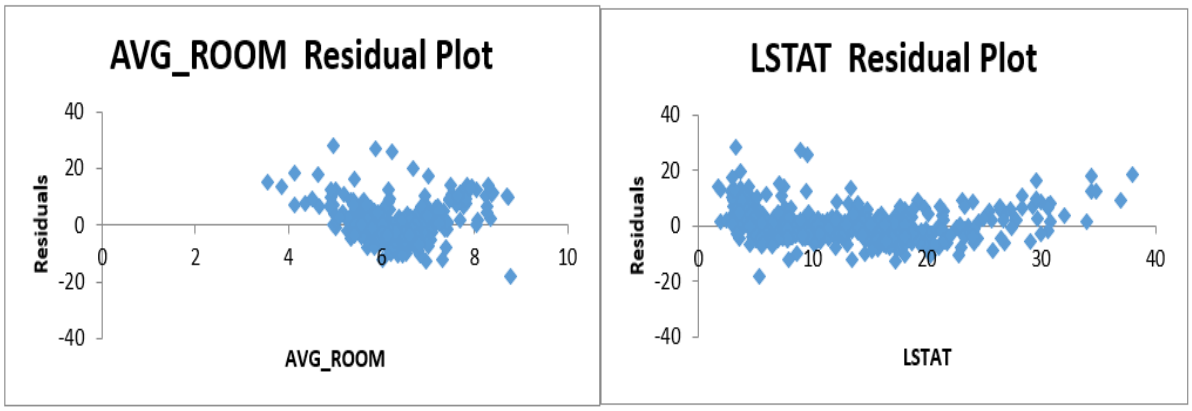
Regression Statistics	
Multiple R	0.799100498
R Square	0.638561606
Adjusted R Square	0.637124475
Standard Error	5.540257367
Observations	506

ANOVA

	df	SS	MS	F	Significance F
Regression	2	27276.98621	13638.49311	444.3308922	7.0085E-112
Residual	503	15439.3092	30.69445169		
Total	505	42716.29542			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1.358272812	3.17282778	-0.428095348	0.668764941	-7.591900282	4.875354658	-7.591900282	4.875354658
AVG_ROOM	5.094787984	0.4444655	11.46272991	3.47226E-27	4.221550436	5.968025533	4.221550436	5.968025533
LSTAT	-0.642358334	0.043731465	-14.68869925	6.66937E-41	-0.728277167	-0.556439501	-0.728277167	-0.556439501

Regression Equation = -1.35827+(5.0947879*7)+(-0.642358*20)



Here we use multiple regression model with Avg.room and Lstat as independent varables and compare with Avg.price as dependent variable. We provide scatter plots, regression equation and residual plot.

The Regression equation is Y=-1.3582+(5.09478*7) + (-0.64235*20)

The value of Avg. price based on data is **21.45 (**in terms of 1000 US. **$)**. A company that sells at an average of **30,000$** is clearly **Overcharging**.

The performance of this model is better than the previous regression model as the adjusted R square value is **0.63712448** compared to **0.543242** of the other models. Higher the adjusted R-square value, better for the Regression Model.

7. Now, build a Regression model with all variables. AVG_PRICE shall be the Dependent Variable. Interpret the output in terms of adjusted R-square, coefficient and Intercept values, Significance of variables with respect to AVG_price. Explain.

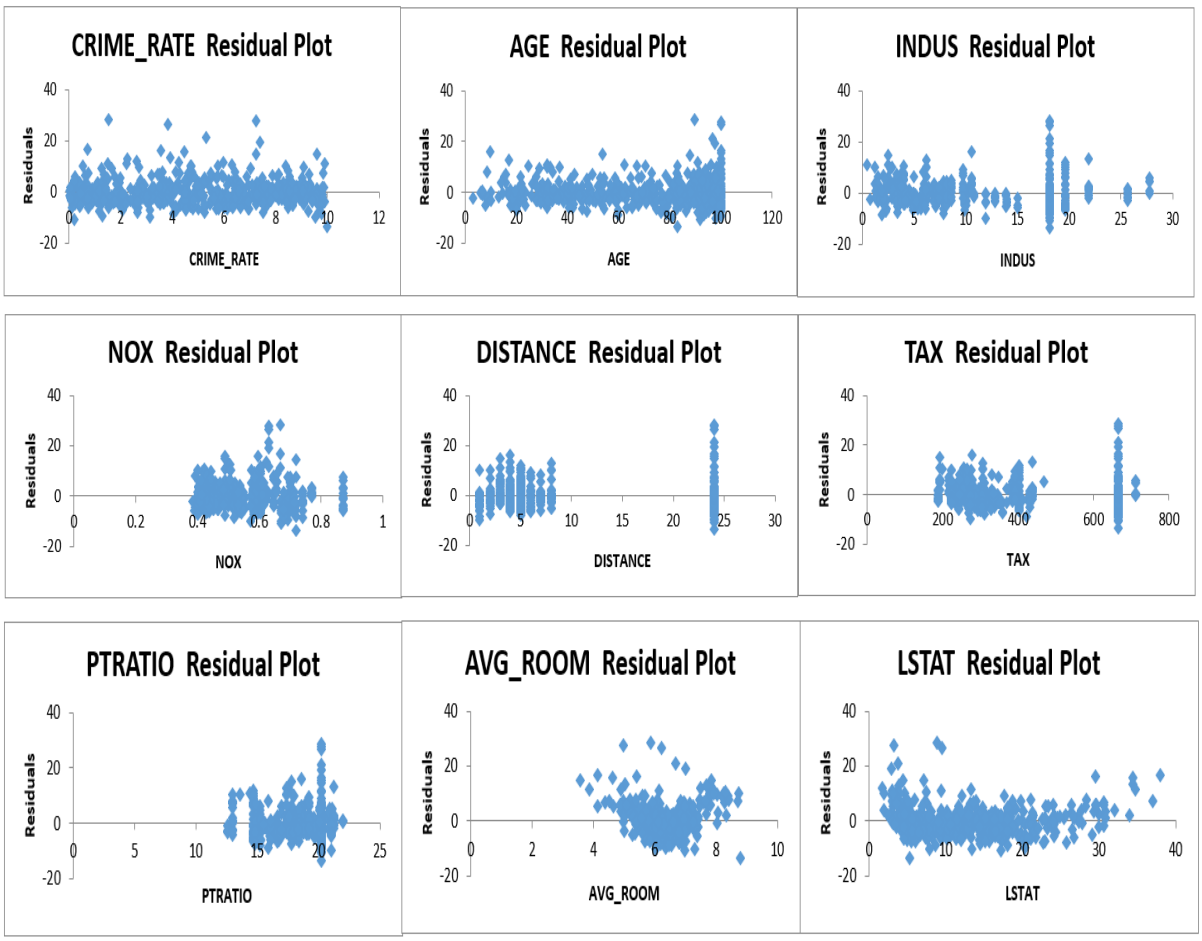
SUMMARY OUTPUT

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

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	29638.8605	3293.206722	124.9045049	1.9328E-121
Residual	496	13077.43492	26.3657962		
Total	505	42716.29542			

	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



The Summary Output and Residual Plots of given attributes are provided above. The **adjusted R-square** value is **0.68829** which definitely shows that this model is better than all the previous models. The coefficients are the beta of the given variables. The significance of all the variables in comparison to the **AVG_PRICE** can be measured from the **p-value**. Except for **CRIME_RATE**, all the other variables have a **p-value** less than **0.05** which proves that it is a significance.

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.
   (HINT: Significant variables are those whose p-values are less than 0.05. If the p-value is greater than 0.05 then it is insignificant)
   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.

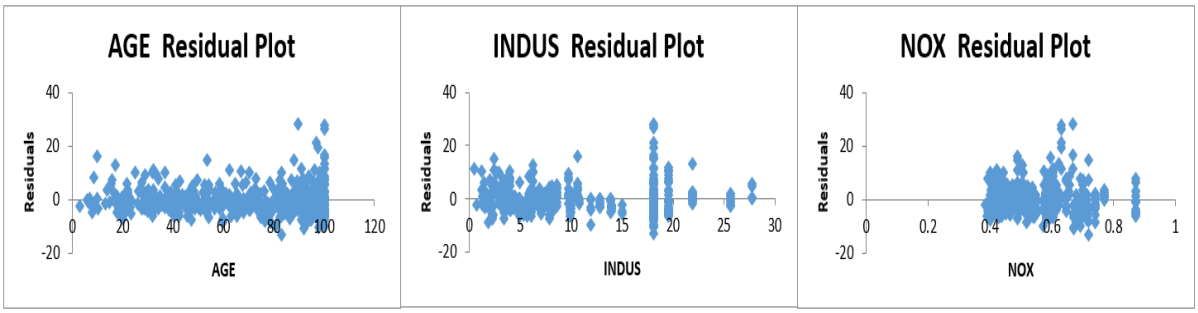
SUMMARY OUTPUT

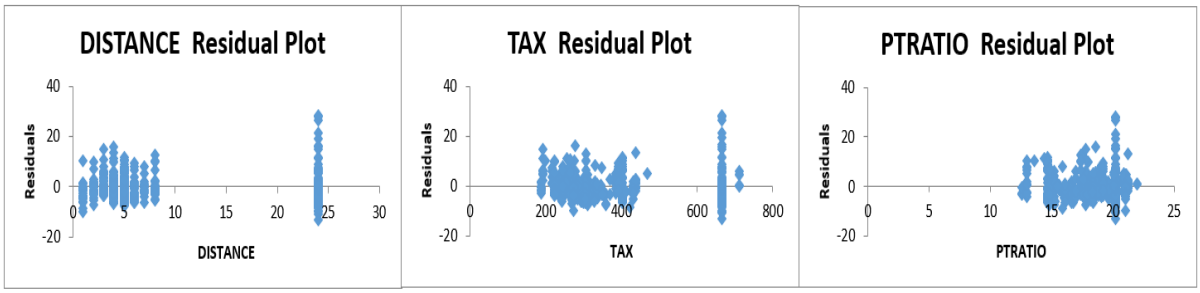
Regression Statistics	
Multiple R	0.832835773
R Square	0.693615426
Adjusted R Square	0.688683682
Standard Error	5.131591113
Observations	506

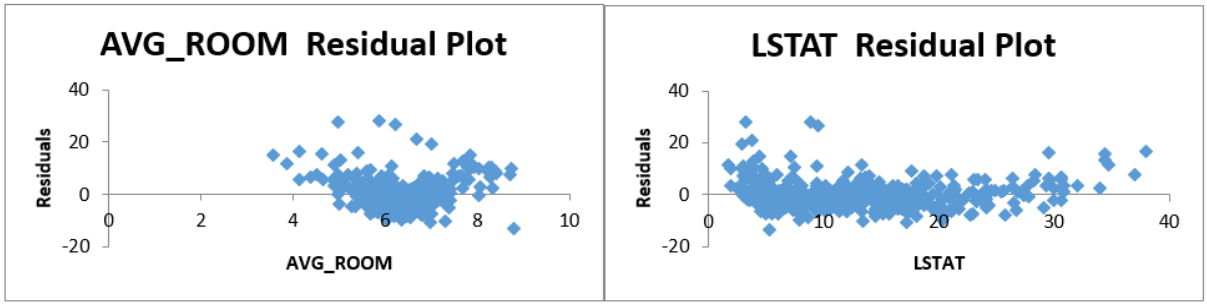
ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	8	29628.68142	3703.585178	140.6430411	1.9E-122
Residual	497	13087.61399	26.33322735		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	29.42847349	4.804728624	6.124898157	1.84597E-09	19.98839	38.86856	19.98839	38.86856
AGE	0.03293496	0.013087055	2.516605952	0.012162875	0.007222	0.058648	0.007222	0.058648
INDUS	0.130710007	0.063077823	2.072202264	0.038761669	0.006778	0.254642	0.006778	0.254642
NOX	-10.27270508	3.890849222	-2.640221837	0.008545718	-17.9172	-2.62816	-17.9172	-2.62816
DISTANCE	0.261506423	0.067901841	3.851242024	0.000132887	0.128096	0.394916	0.128096	0.394916
TAX	-0.014452345	0.003901877	-3.703946406	0.000236072	-0.02212	-0.00679	-0.02212	-0.00679
PTRATIO	-1.071702473	0.133453529	-8.030529271	7.08251E-15	-1.33391	-0.8095	-1.33391	-0.8095
AVG_ROOM	4.125468959	0.44248544	9.323400461	3.68969E-19	3.256096	4.994842	3.256096	4.994842
LSTAT	-0.605159282	0.0529801	-11.42238841	5.41844E-27	-0.70925	-0.50107	-0.70925	-0.50107







This is the Final Regression Model which consists of only the significant variables i.e; the p-value is less than 0.05. The Summary Output and Residual Plots are provided above. The **adjusted R-square** value is **0.68868**. Since we removed the non-significant **CRIME_RATE** variable from this model , the adjusted R-square value **increased** by **0.00038504**. Thus, this is the most **successfu**l and **relevant** model that can be created from the given dataset.

The Regression Equation of this model is

Y=29.428+0.033(Age)+0.131(Indus)-10.273(Nox)+0.2615(Distance)-0.015(Tax)-1.072(PtRatio)+4.125(Avg_Room)-0.605(Lstat).

Terro's Real Estate Agency

Done by

Siliveri Mohan