Assignment – 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 tool pack). Write down your observation. **(5 marks)**

ANS

I. Crime Rate

CRIME_RATE	
Mean	4.87197628
Standard Error	0.12986015
Median	4.82
Mode	3.43
Standard Deviation	2.92113189
Sample Variance	8.53301153
Kurtosis	-1.1891225
Skewness	0.02172808
Range	9.95
Minimum	0.04
Maximum	9.99
Sum	2465.22
Count	506

- Central Tendency
    - Mean and Median are closer together there not much deference
- Dispersion
    - Coefficient of Variance = 0.59957843
- Its slightly higher that 0.5 so the Spread is more
- Symmetry
    - Skewness is closer to 0 its almost symmetric and its positive more data to the left of the mean witch implies median is lesser than mean
- Kurtosis
    - The Kurtosis is negative so its flat.

II. Age

AGE	
Mean	68.5749012
Standard Error	1.25136953
Median	77.5
Mode	100
Standard Deviation	28.1488614
Sample Variance	792.358399
Kurtosis	-0.9677156
Skewness	-0.5989626
Range	97.1
Minimum	2.9
Maximum	100
Sum	34698.9
Count	506

- Central Tendency
  - Median is slightly Higher than Mean
- Dispersion

  Coefficient of Variance = 0.41048344

  It's between 0.2-0.5, So the Spread is Normal

- Symmetry
  - Skewness is Negative more data to the Right of the mean witch implies median is Grater than mean. That means tail to the right

- Kurtosis
  - The Kurtosis is negative so its flat.

III. Industry

INDUSTRY	
Mean	11.1367787
Standard Error	0.30497989
Median	9.69
Mode	18.1
Standard Deviation	6.86035294
Sample Variance	47.0644425
Kurtosis	-1.2335396
Skewness	0.29502157
Range	27.28
Minimum	0.46

Maximum	27.74
Sum	5635.21
Count	506

- Central Tendency
  - Mean is slightly Higher than Median so we take Mean as a center
- Dispersion

  Coefficient of Variance = 0.61600874

  Its higher that 0.5 so the Spread is more

- Symmetry
  - Skewness is Positive more data to the left of the mean witch implies median is lesser than mean. That means tail to the left
- Kurtosis
  - The Kurtosis is negative so its flat.

IV. NOX

NOX	
Mean	0.55469506
Standard Error	0.00515139
Median	0.538
Mode	0.538
Standard Deviation	0.11587768
Sample Variance	0.01342764
Kurtosis	-0.0646671
Skewness	0.72930792
Range	0.486
Minimum	0.385
Maximum	0.871
Sum	280.6757
Count	506

- Central Tendency
  - Mean and Median are closer together there not much deference
- Dispersion
  - Coefficient of Variance = 0. 20890339
- It's between 0.2-0.5, So the Spread is Normal
- Symmetry
  - Skewness is Positive more data to the left of the mean witch implies median is lesser than mean. That means tail to the left
- Kurtosis
  - The Kurtosis is negative so its flat.

V. Distance

DISTANCE	
Mean	9.54940711
Standard Error	0.38708489
Median	5
Mode	24
Standard Deviation	8.70725938
Sample Variance	75.816366
Kurtosis	-0.867232
Skewness	1.00481465
Range	23
Minimum	1
Maximum	24
Sum	4832
Count	506

- Central Tendency
  - Mean is Higher than Median so Median will be center
- Dispersion

  Coefficient of Variance = 0. 91181152

  It's higher that 0.5 so the Spread is more
- Symmetry
  - Skewness is Positive more data to the left of the mean witch implies median is lesser than mean. That means tail to the left
- Kurtosis
  - The Kurtosis is negative so its flat.

VI. Tax

TAX	
Mean	408.237154
Standard Error	7.49238869
Median	330
Mode	666
Standard Deviation	168.537116
Sample Variance	28404.7595
Kurtosis	-1.142408
Skewness	0.66995594
Range	524
Minimum	187
Maximum	711
Sum	206568
Count	506
	0.4128412

- Central Tendency

    - Mean is Higher than Median so Median will be center
- Dispersion

  Coefficient of Variance = 0. 4128412

  It's between 0.2-0.5, So the Spread is Normal
- Symmetry
    - Skewness is Positive more data to the left of the mean witch implies median is lesser than mean. That means tail to the left
- Kurtosis
    - The Kurtosis is negative so its flat.

VII. Ptratio

PTRATIO	
Mean	18.4555336
Standard Error	0.09624357
Median	19.05
Mode	20.2
Standard Deviation	2.16494552
Sample Variance	4.68698912
Kurtosis	-0.2850914
Skewness	-0.8023249
Range	9.4
Minimum	12.6
Maximum	22
Sum	9338.5
Count	506

- Central Tendency
    - Mean and Median are closer together there not much deference

- Dispersion
    - Coefficient of Variance = 0. 11730604

      Its lower than 0.2 so the Spread is less.

- Symmetry
    - Skewness is Negative more data to the Right of the mean witch implies median is Greater than mean. That means tail to the right

- Kurtosis
    - The Kurtosis is negative so its flat.

VIII. Avg_Room

AVG_ROOM	
Mean	6.28463439
Standard Error	0.03123514
Median	6.2085

Mode	5.713
Standard Deviation	0.70261714
Sample Variance	0.49367085
Kurtosis	1.89150037
Skewness	0.40361213
Range	5.219
Minimum	3.561
Maximum	8.78
Sum	3180.025
Count	506

- Central Tendency
  - Mean and Median are similar

- Dispersion
  - Coefficient of Variance = 0. 11179921
    Its lower than 0.2 so the Spread is less.

- Symmetry
  - Skewness is Positive more data to the left of the mean witch implies median is lesser than mean. That means tail to the left
- Kurtosis
  - The Kurtosis is postive so its slight peak.

IX. Lstat

LSTAT	
Mean	12.6530632
Standard Error	0.31745891
Median	11.36
Mode	8.05
Standard Deviation	7.14106151
Sample Variance	50.9947595
Kurtosis	0.49323952
Skewness	0.90646009
Range	36.24
Minimum	1.73
Maximum	37.97
Sum	6402.45

- Central Tendency
  - Mean is Higher than Median so Median will be center
- Dispersion

  Coefficient of Variance = 0. 56437413

  It's higher that 0.5 so the Spread is more
- Symmetry
  - Skewness is Positive more data to the left of the mean witch implies median is lesser than mean. That means tail to the left

- Kurtosis
  - The Kurtosis is postive so its slight sharp peak.

X. Avg_Price

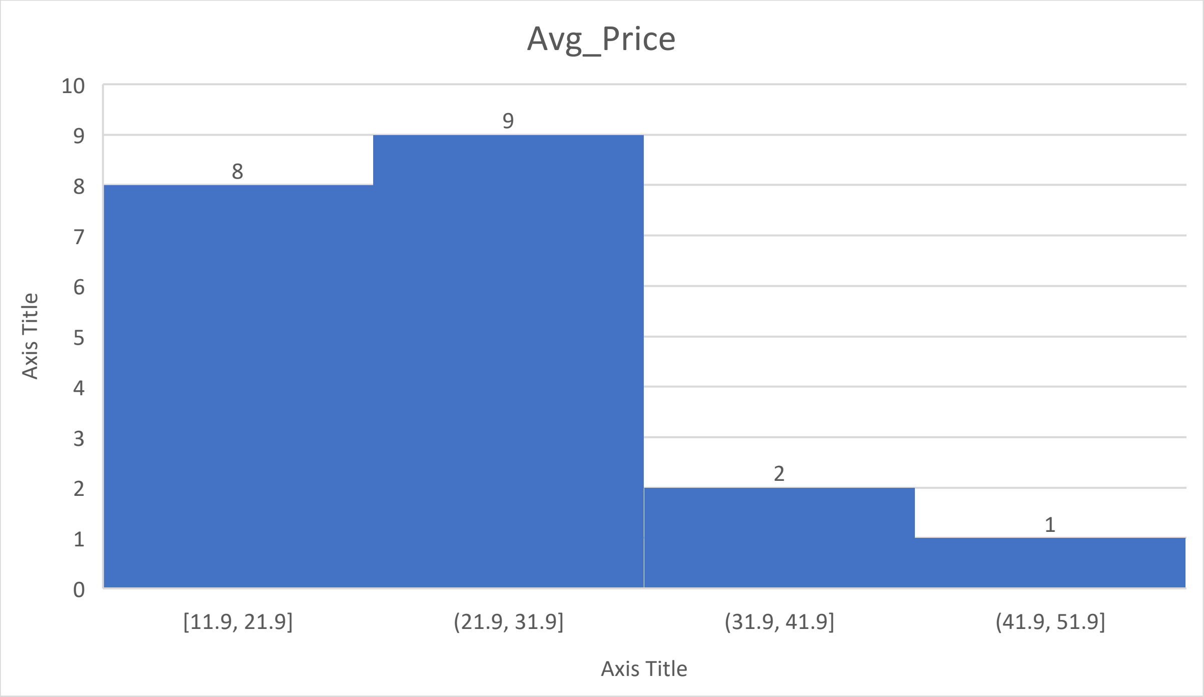
AVG_PRICE	
Mean	22.5328063
Standard Error	0.40886115
Median	21.2
Mode	50
Standard Deviation	9.19710409
Sample Variance	84.5867236
Kurtosis	1.49519694
Skewness	1.10809841
Range	45
Minimum	5
Maximum	50
Sum	11401.6
Count	506

- Central Tendency
  - Mean is Higher than Median so Median will be center
- Dispersion
  - Coefficient of Variance = 0. 40816505

  It’s between 0.2-0.5, So the Spread is Normal
- Symmetry
  - Skewness is Positive more data to the left of the mean witch implies median is lesser than mean. That means tail to the left
- Kurtosis
  - The Kurtosis is postive so its slight sharp peak

2. Plot a histogram of the Avg_Price variable. What do you infer? **(5 marks)**

Ans)



Based on the information provided, it appears that the distribution of the data has a positive skewness, as there are more values on the left side of the median. This implies that the tail of the distribution extends towards the right. Additionally, the presence of a peak suggests that the kurtosis is positive.

3. Compute the covariance matrix. Share your observations. **(5 marks)**

Ans)

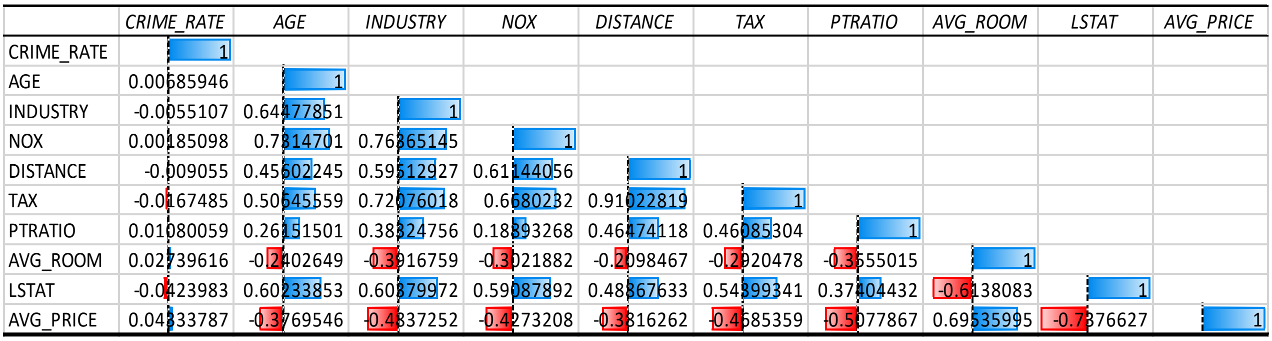
	CRIME_RATE	*AGE*	*INDUSTRY*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	8.51614787									
AGE	0.56291522	790.792473								
INDUSTRY	-0.1102152	124.267828	46.9714297							
NOX	0.00062531	2.38121193	0.60587394	0.0134011						
DISTANCE	-0.2298605	111.549955	35.4797145	0.61571022	75.6665313					
TAX	-8.2293224	2397.94172	831.713333	13.0205024	1333.11674	28348.6236				
PTRATIO	0.06816891	15.9054254	5.68085478	0.04730365	8.74340249	167.820822	4.6777263			
AVG_ROOM	0.05611778	-4.742538	-1.8842254	-0.0245548	-1.2812774	-34.515101	-0.5396945	0.49269522		
LSTAT	-0.8826804	120.838441	29.5218113	0.48797987	30.3253921	653.420617	5.77130024	-3.073655	50.8939794	
AVG_PRICE	1.16201224	-97.396153	-30.460505	-0.4545124	-30.50083	-724.82043	-10.090676	4.48456555	-48.351792	84.4195562

Based on the covariance matrix provided, there are both positive and negative values, indicating that the data is spread in multiple dimensions with a variety of relationships between variables.

4) Create a correlation matrix of all the variables (Use Data analysis tool pack). **(5 marks)**

Ans)

a) Which are the top 3 positively correlated pairs and
b) Which are the top 3 negatively correlated pairs.



	CRIME_RATE	*AGE*	*INDUSTRY*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	1									
AGE	0.00685946	1								
INDUSTRY	-0.0055107	0.64477851	1							
NOX	0.00185098	0.7314701	0.76365145	1						
DISTANCE	-0.009055	0.45602245	0.59512927	0.61144056	1					
TAX	-0.0167485	0.50645559	0.72076018	0.6680232	0.91022819	1				
PTRATIO	0.01080059	0.26151501	0.38324756	0.18893268	0.46474118	0.46085304	1			
AVG_ROOM	0.02739616	-0.2402649	-0.3916759	-0.3021882	-0.2098467	-0.2920478	-0.3555015	1		
LSTAT	-0.0423983	0.60233853	0.60379972	0.59087892	0.48867633	0.54399341	0.37404432	-0.6138083	1	
AVG_PRICE	0.04333787	-0.3769546	-0.4837252	-0.4273208	-0.3816262	-0.4685359	-0.5077867	0.69535995	-0.7376627	1

- Top 3 positively correlated pairs
  - TAX & DISTANCE 0.91022819
  - NOX & INDUSTRY 0.76365145
  - NOX & AGE 0.7314701

- Top 3 negatively correlated pairs
  - AVG_PRICE & LSTAT -0.7376627
  - LSTAT & AVG_ROOM -0.6138083
  - AVG_PRICE & PTRATIO -0.5077867

5) Build an initial regression model with AVG_PRICE as **'y'** (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot. **(8 marks)**

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?

ANS) a) Condition 1

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	34.5538409	0.56262735	61.4151455	3.743E-236	33.448457	35.6592247	33.448457	35.6592247
LSTAT	-0.9500494	0.03873342	-24.5279	5.0811E-88	-1.0261482	-0.8739505	-1.0261482	-0.8739505

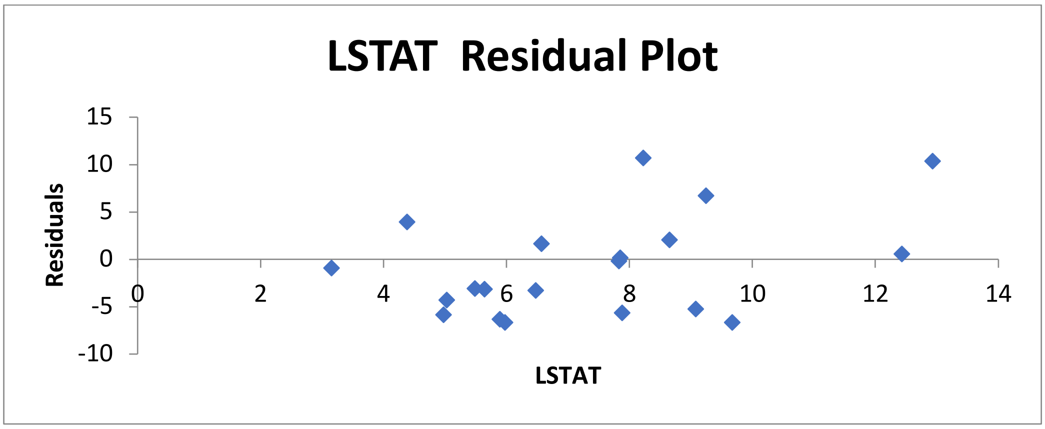
From the Regression Summary in this model the P-Value is lesser that 0.05 So, Alternate Hypothesis is True and Null Hypothesis is False.

Condition 2

Regression Statistics	
Multiple R	0.73766273
R Square	0.5441463
Adjusted R Square	0.5441463
Standard Error	6.21576041
Observations	506

The Adjusted R Square is 0.5441463
In this model X Explain 54% times of Y

Condition 3



The Error is Random its Scatter around the 0.

b) This model is satisfied 1 & 3 Condition ,but in the 2 Condition Adjusted R Square is low so it’s not the perfect model for Analysis.

6. Build a new Regression model including LSTAT and AVG_ROOM together as Independent variables and AVG_PRICE as dependent variable. **(6 marks)**

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.

Ans) a) The coefficients are

	Coefficients
Intercept	-1.3582728
AVG_ROOM	5.09478798
LSTAT	-0.6423583

Regression Equation = Intercept + AVG_ROOM Coefficients * AVG_ROOM Value + LSTAT Coefficients * LSTAT Value

Predict value = -1.3582728+5.09478798*7+-0.6423583*20
= 21.4580771
As we can se that from our Predicted Price 21000 USD so Company is Overcharging

b)

Regression Statistics	
Multiple R	0.7991005
R Square	0.63856161
Adjusted R Square	0.63712448
Standard Error	5.54025737
Observations	506

In this model Adjusted R Square is increased as compare to the previous model So this will perform better than 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. **(8 marks)**

Ans)

Regression Statistics	
Multiple R	0.83297882
R Square	0.69385372
Adjusted R Square	0.68829865
Standard Error	5.1347635
Observations	506

The Adjusted R Square is 0. 68829865

In this model all the other variables explain 68% times of Avg_price

	Coefficients
Intercept	29.2413153
CRIME_RATE	0.04872514
AGE	0.03277069
INDUSTRY	0.1305514
NOX	-10.321183
DISTANCE	0.26109357
TAX	-0.0144012
PTRATIO	-1.0743053
AVG_ROOM	4.12540915
LSTAT	-0.6034866

- As per this model
  - The Intercept value for this model is 29241 USD
  - For every Crime Rate our Price is increasing 48 USD
  - For every 1 year of increase the price decrease by 32 USD
  - For every present of increase in industry the price increase by 130 USD
  - For every per 10 millions of NOX increase the decrease by 10321 USD
  - For every miles increase the price decreases 261 USD
  - For every 10000 USD increase in Tax the price decrease by 14 USD
  - For every unit Ptration increase the price decrease by 1074 USD
  - For every Room increase the price increase by 4125 USD
  - For ever percentage of LSTAT increase the price of decrease by 603 USD

	P-value
Intercept	2.5398E-09
CRIME_RATE	0.5346572
AGE	0.01267044
INDUSTRY	0.03912086
NOX	0.00829386
DISTANCE	0.00013755
TAX	0.00025125
PTRATIO	6.5864E-15
AVG_ROOM	3.8929E-19
LSTAT	8.9107E-27

- As per the P-value all the independent variable is less than 0.05 except Crime Rate So we can say that Alternate Hypothesis is false and null hypothesis true in Crime rate But in other variables that is Age, Industry, NOX, Distance, Tax, Ptratio, Avg_Room & LSTAT Alternate Hypothesis is true And Null hypothesis is false.

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: **(8 marks)**
   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.

Ans) a) From this model

- Condition 1

	P-value
Intercept	1.846E-09
AGE	0.01216288
INDUSTRY	0.03876167
NOX	0.00854572
DISTANCE	0.00013289
TAX	0.00023607
PTRATIO	7.0825E-15
AVG_ROOM	3.6897E-19
LSTAT	5.4184E-27

In this model all P-values is lower than 0.05,So Alternate Hypothesis is true and Null Hypothesis is False.

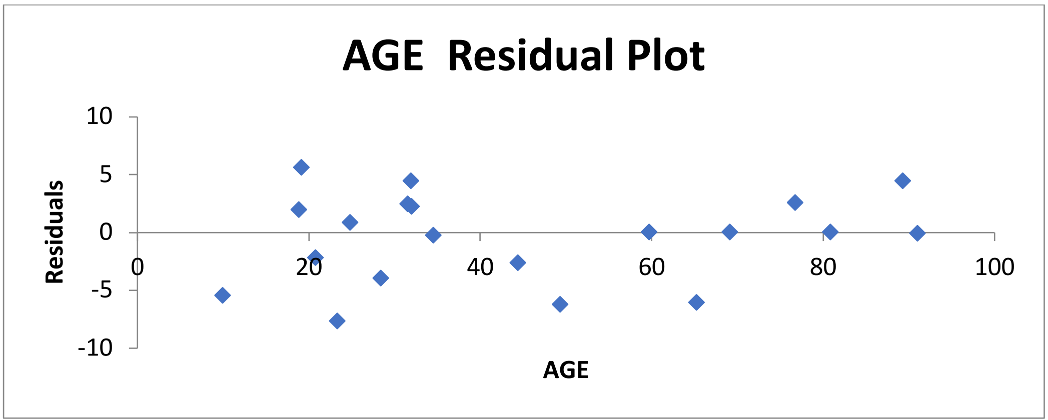
- Condition 2

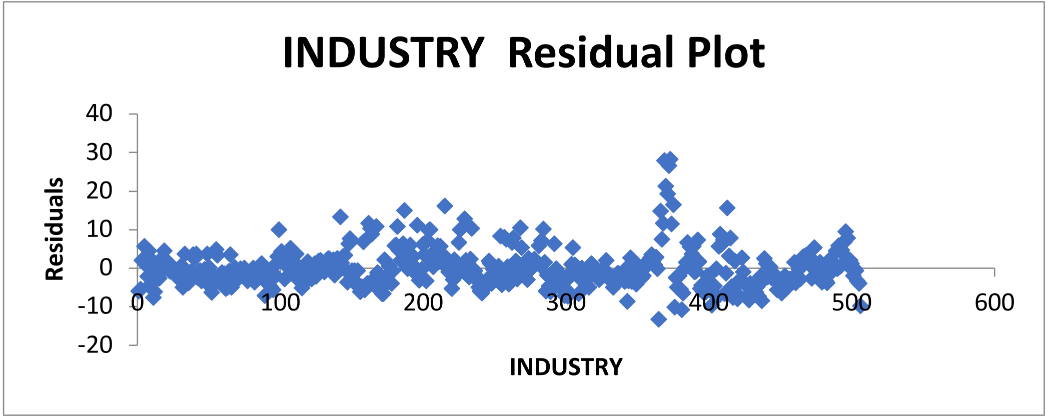
Regression Statistics	
Multiple R	0.83283577
R Square	0.69361543
Adjusted R Square	0.68868368
Standard Error	5.13159111
Observations	506

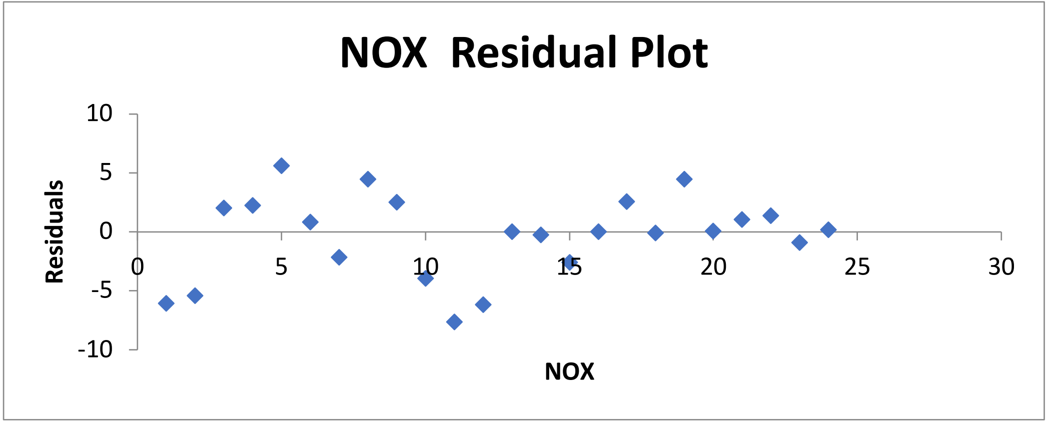
The Adjusted R Square is 0. 68868368

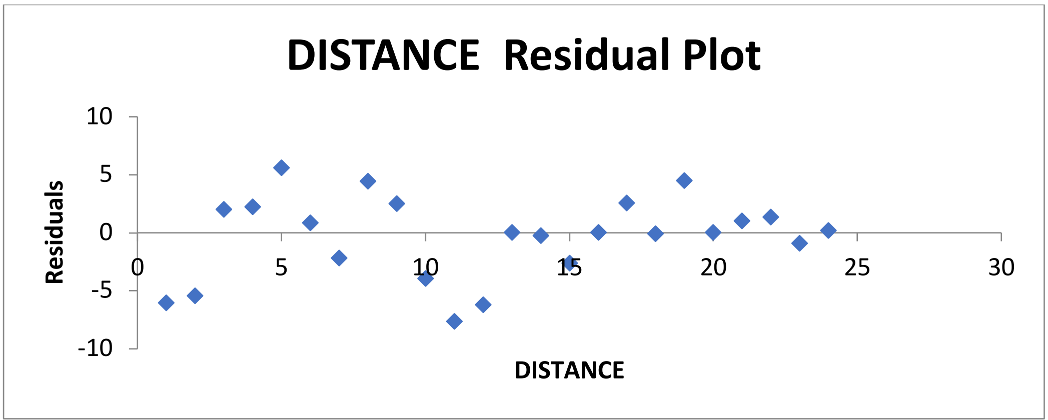
In this model all the significant variables Explain 68% times of Avg_Price

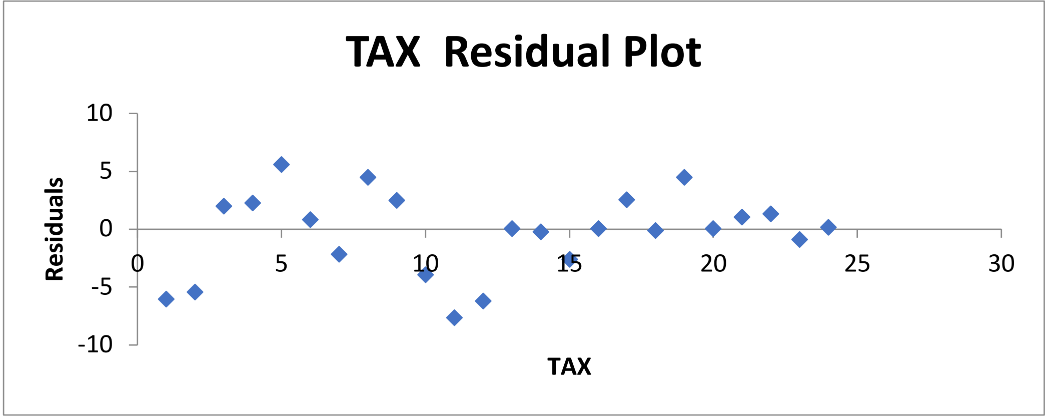
- Condition 3

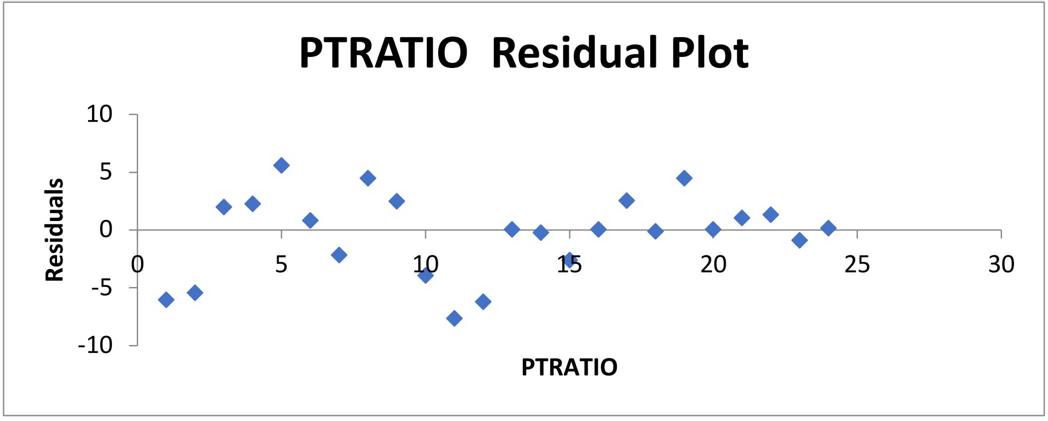


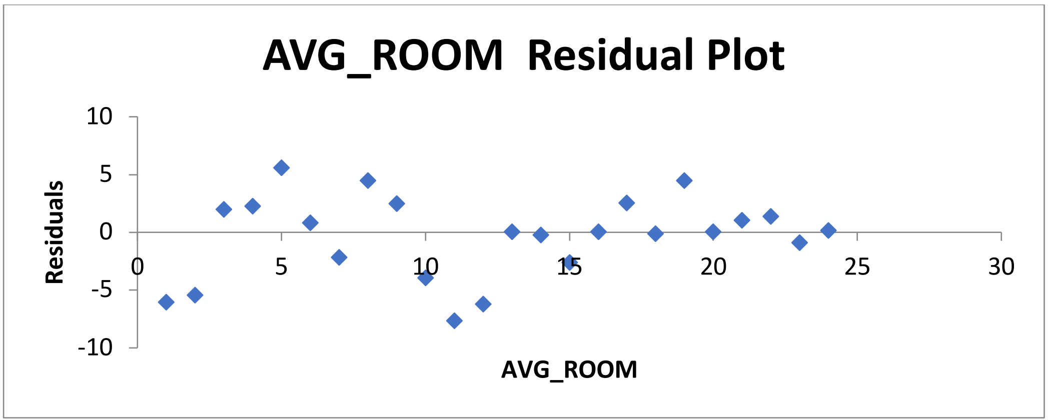


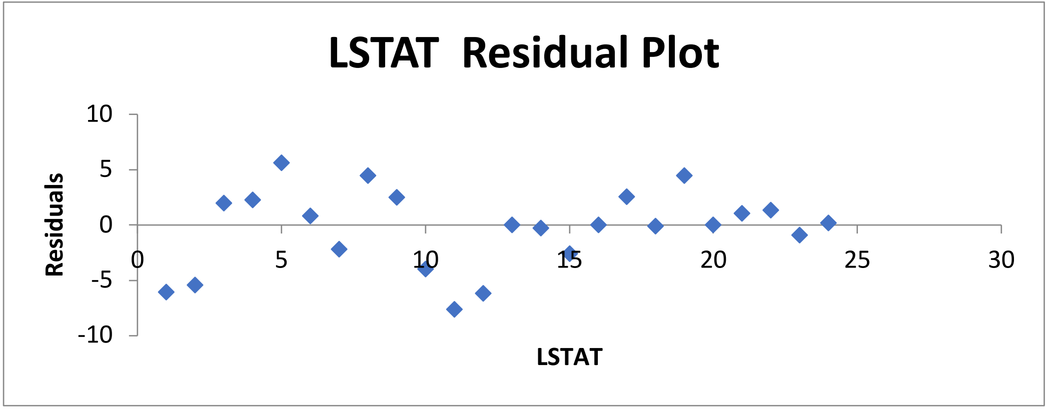



DISTANCE Residual Plot
Residuals
10
5
0
-5
-10
0
5
10
15
20
25
30
DISTANCE


TAX Residual Plot
Residuals
10
5
0
-5
-10
0
5
10
15
20
25
30
TAX


PTRATIO Residual Plot
Residuals
10
5
0
-5
-10
0
5
10
15
20
25
30
PTRATIO





- After analysing all Residual Plot of Significant variables above
  The Error are Random they are Scattered Around 0.So this condition Succeeded.

- This model Satisfied all the three condition. We can use this model for Prediction
  But The Adjusted R square not that much Strong if we get Some more Data it might help us improve the Adjusted R Square

b) In this model Adjusted R Square is slightly Increased as compare to the Previous model. So we can clearly say that this model will more effective than previous model.

c)

lable	*Coefficients*
Intercept	29.4284735
NOX	-10.272705
PTRATIO	-1.0717025
LSTAT	-0.6051593
TAX	-0.0144523
AGE	0.03293496
INDUSTRY	0.13071001
DISTANCE	0.26150642
AVG_ROOM	4.12546896

- If the NOX value increase in the locality the Avg_Price will decrease.10.272 USD for Unit of NOX

d) Regression Equation = Intercept + Coefficient of NOX*NOX value +Coefficient of PTratio*PTratio value + Coefficient of LSTAT*LSTAT value +Coefficient of TAX*TAX Value + Coefficient of Age*Age Value + Coefficient of Industry*Industry Value + Coefficient of Distance*Distance value + Coefficient of Avg_Room*Avg_Room value