TERRO'S REAL ESTATE AGENCY

Project by,

S.Shanmugapriya

TABLE OF CONTENTS

INTRODUCTION

Terro's real-estate is an agency which estimates the pricing of houses in a certain locality. The pricing is concluded based on different features or factors of a property. **My job** as an **"auditor"** is to study and **analyze the various feature**s of a property like crime rate, pollution level, connectivity, education facilities etc on the provided dataset by the agency. This also helps in identifying the business value of a property and determine the price of a property by,

- Finding out the most relevant features for pricing of a house.
- Analyzing the magnitude of each variable to which it can affect the price of a house in a particular locality.

The dataset consists of 506 houses in Boston.The features provided for the analyze are,

1. CRIME RATE - per capita crime rate by town
2. INDUSTRY - proportion of non-retail business acres per town (in %)
3. NOX - nitric oxides concentration (parts per 10 million)
4. AVG_ROOM - average number of rooms per house
5. AGE - proportion of houses built prior to 1940 (in %)
6. DISTANCE - distance from highway (in miles)
7. TAX - full-value property-tax rate per $10,000
8. PTRATIO - pupil-teacher ratio by town
9. LSTAT - % lower status of the population
10. AVG_PRICE - Average value of houses in $1000's.

1. SUMMARY STATISTICS OF EACH VARIABLE

CRIME_RATE	
Mean	4.871976285
Standard Error	0.129860152
Median	4.82
Mode	3.43
Standard Deviation	2.921131892
Sample Variance	8.533011532
Kurtosis	-1.18912246
Skewness	0.021728079
Range	9.95
Minimum	0.04
Maximum	9.99
Sum	2465.22
Count	506

AGE	
Mean	68.574901
Standard Error	1.2513695
Median	77.5
Mode	100
Standard Deviation	28.148861
Sample Variance	792.3584
Kurtosis	-0.9677156
Skewness	-0.5989626
Range	97.1
Minimum	2.9
Maximum	100
Sum	34698.9
Count	506

INDUS	
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

The **Average per capita crime rate by town** is 4.87.

Here, it shows **negative kurtosis** which means a flat curve which is also called as "Platykurtic".

This summary shows a **positive skew** (mean>median)

The **Average proportion of houses built prior to 1940** in percentage is 68.5.

Here, it shows **negative kurtosis** which means a flat curve which is also called as "Platykurtic".

This summary shows a **negative skew** (mean<median).

The **Average proportion of non-retail business acres per town** in percentage is 11.13.

Here, it shows **negative kurtosis** which means a flat curve which is also called as "Platykurtic".

This summary shows a **positive skew** (mean>median).

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

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

TAX	
Mean	408.2371542
Standard Error	7.492388692
Median	330
Mode	666
Standard Deviation	168.5371161
Sample Variance	28404.75949
Kurtosis	-1.142407992
Skewness	0.669955942
Range	524
Minimum	187
Maximum	711
Sum	206568
Count	506

The **Average nitric oxides concentration** (parts per 10 million) is 0.554.

Here, it shows **negative kurtosis** which means a flat curve which is also called as "Platykurtic".

This summary shows a **positive skew** (mean>median).

The **Average distance from highway** (in miles) is 9.54.

Here, it **shows negative kurtosis** which means a flat curve which is also called as "Platykurtic".

This summary shows a **positive skew** (mean>median).

The **Average full-value property-tax rate per $10,000** is 408.23.

Here, it shows **negative kurtosis** which means a flat curve which is also called as "Platykurtic".

This summary shows a **positive skew** (mean>median).

PTRATIO	
Mean	18.4555336
Standard Error	0.096243568
Median	19.05
Mode	20.2
Standard Deviation	2.164945524
Sample Variance	4.686989121
Kurtosis	-0.285091383
Skewness	-0.802324927
Range	9.4
Minimum	12.6
Maximum	22
Sum	9338.5
Count	506

AVG_ROOM	
Mean	6.284634387
Standard Error	0.031235142
Median	6.2085
Mode	5.713
Standard Deviation	0.702617143
Sample Variance	0.49367085
Kurtosis	1.891500366
Skewness	0.403612133
Range	5.219
Minimum	3.561
Maximum	8.78
Sum	3180.025
Count	506

LSTAT	
Mean	12.65306324
Standard Error	0.317458906
Median	11.36
Mode	8.05
Standard Deviation	7.141061511
Sample Variance	50.99475951
Kurtosis	0.493239517
Skewness	0.906460094
Range	36.24
Minimum	1.73
Maximum	37.97
Sum	6402.45
Count	506

The **Average pupil – teacher ratio** by town is 18.45.

Here, it shows **negative kurtosis** which means a flat curve which is also called as "Platykurtic".

This summary shows a **negative skew** (mean<median).

The **Average of average number of rooms** per house is 6.28.

Here, it shows **negative kurtosis** which means a flat curve which is also called as "Platykurtic".

This summary shows a **positive skew** (mean>median).

The **Average % of lower status of the population** is 12.65.

Here, it shows **negative kurtosis** which means a flat curve which is also called as "Platykurtic".

This summary shows a **positive skew** (mean>median).

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

The **Average of Average value of houses in $1000's** is 22.53.

Here, it shows **negative kurtosis** which means a flat curve which is also called as "Platykurtic".

This summary shows **a positive skew** (mean>median).

OVERALL SUMMARY:

Using this statistical summary, I can infer the measures such as,

1) Measure of central tendency
    - Mean
    - Median
    - Mode

Where mean and median plays a major role.

2) Measure of dispersion
    - Range
    - Variance
    - Standard deviation
3) Measure of symmetry
    - Skewness

Skewness plays a major role as it affects mean the most due to outliers.

Here, variables like CRIME RATE, INDUS, NOX, DISTANCE, TAX, AVERAGE ROOM, LSTAT, AVERAGE PRICE has positive skew.

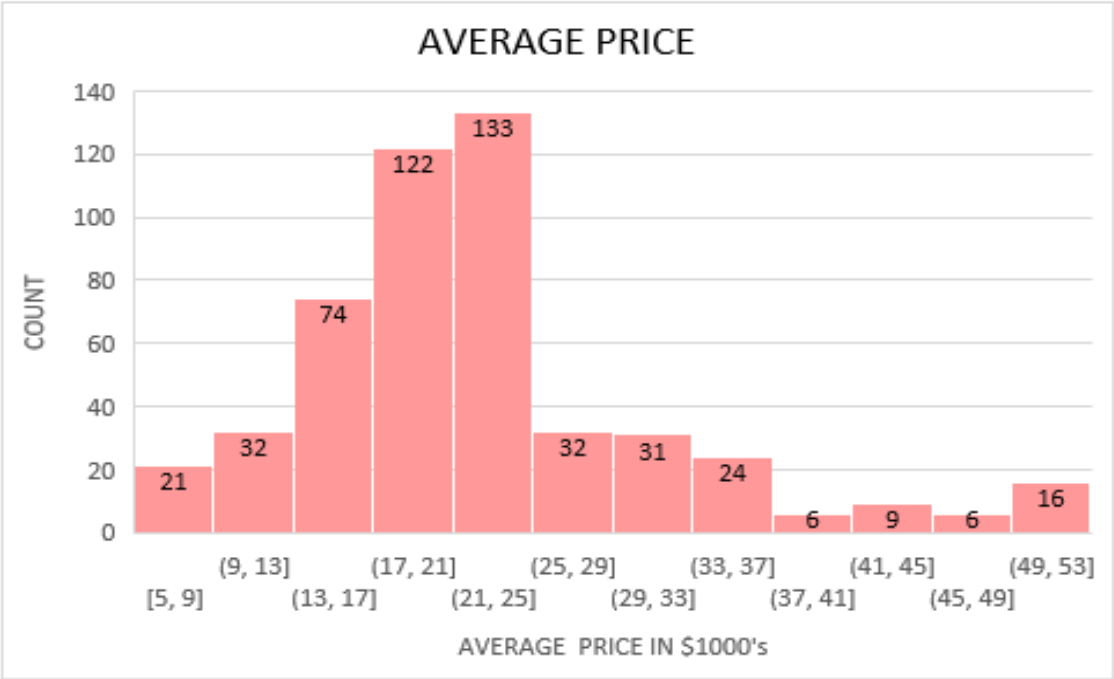
Whereas, AGE and PRATIO has negative skew.

4) Measure of peakedness
    - Kurtosis

Here, all the variables show negative kurtosis as the kurtosis value is below 3. It means a flat curve which is also called as "Platykurtic".

I can also infer the sum and count of each variable.

2. INFERENCE ON AVERAGE PRICE USING HISTOGRAM



- I can clearly see a higher spike in the bin [21,25], followed by [17,21]. There is least spike on two bins [37,41] and [45,49].
- Among 506 observations, a count of 133 is around $21000 to $25000.
- The average price of majority of the houses is around $21000 to $25000 and $17000 to $21000.

3. OBSERVATION ON COVARIANCE MATRIX

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.516147873									
AGE	0.562915215	790.7924728								
INDUS	-0.110215175	124.2678282	46.97142974							
NOX	0.000625308	2.381211931	0.605873943	0.013401099						
DISTANCE	-0.229860488	111.5499555	35.47971449	0.615710224	75.66653127					
TAX	-8.229322439	2397.941723	831.7133331	13.02050236	1333.116741	28348.6236				
PTRATIO	0.068168906	15.90542545	5.680854782	0.047303654	8.74340249	167.8208221	4.6777263			
AVG_ROOM	0.056117778	-4.74253803	-1.884225427	-0.02455483	-1.28127739	-34.51510104	-0.5396945	0.49269522		
LSTAT	-0.882680362	120.8384405	29.52181125	0.487979871	30.32539213	653.4206174	5.7713002	-3.073655	50.89397935	
AVG_PRICE	1.16201224	-97.3961529	-30.46050499	-0.45451241	-30.5008304	-724.8204284	-10.090676	4.48456555	-48.3517922	84.4195562

- Using Covariance we can measure the joint dispersion of two variables or their joint variation. In other words, how the variation in a independent variable(X) affects the variation in a dependent variable(Y).
- From this I can clearly infer that, crime rate and average room are directly proportional to average price of the house.
- If crime rate or average room increases, the average price also increases.
- Whereas the other variables such as age, industry, NOX, distance, tax, PRATIO, LSTAT are inversely proportional to average price.
- For instance, If age of the house increases, the price decreases.
- Likewise, PRATIO and distance are directly proportional to each other.
- LSTAT and average rooms are inversely proportional to each other.

4. CORRELATION MATRIX

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	1									
AGE	0.006859463	1								
INDUS	-0.005510651	0.644778511	1							
NOX	0.001850982	0.731470104	0.763651447	1						
DISTANCE	-0.009055049	0.456022452	0.595129275	0.611440563	1					
TAX	-0.016748522	0.506455594	0.72076018	0.6680232	0.910228189	1				
PTRATIO	0.010800586	0.261515012	0.383247556	0.188932677	0.464741179	0.460853035	1			
AVG_ROOM	0.02739616	-0.240264931	-0.391675853	-0.30218819	-0.20984667	-0.292047833	-0.35550149	1		
LSTAT	-0.042398321	0.602338529	0.603799716	0.590878921	0.488676335	0.543993412	0.374044317	-0.613808272	1	
AVG_PRICE	0.043337871	-0.376954565	-0.48372516	-0.42732077	-0.38162623	-0.468535934	-0.50778669	0.695359947	-0.737662726	1

a) TOP 3 POSITIVELY CORRELATED PAIRS

1) 0.910228189 - Tax and Distance
2) 0.763651447 - NOX and Indus
3) 0.731470104 - NOX and Age

b) TOP 3 NEGATIVELY CORRELATED PAIRS

1) -0.737662726 - Average price and LSTAT
2) -0.613808272 - LSTAT and Average room
3) -0.507786686 - Average price and PRATIO

5. INITIAL REGRESSION MODEL

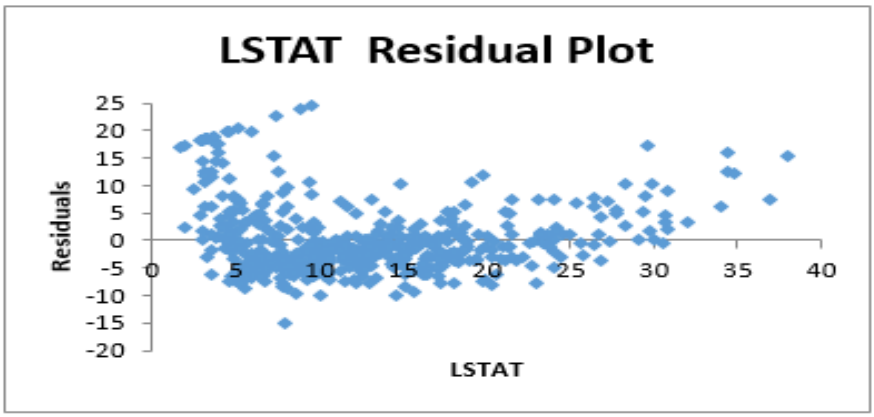
- INDEPENDENT VARIABLE (X) - LSTAT
- DEPENDENT VARIABLE (Y) – AVERAGE PRICE

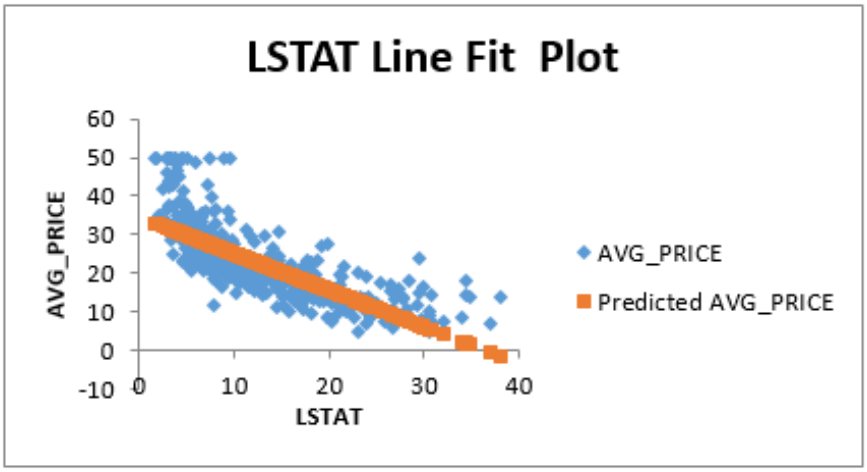
Regression Statistics	
Multiple R	0.737662726
R Square	0.544146298
Adjusted R Square	0.543241826
Standard Error	6.215760405
Observations	506

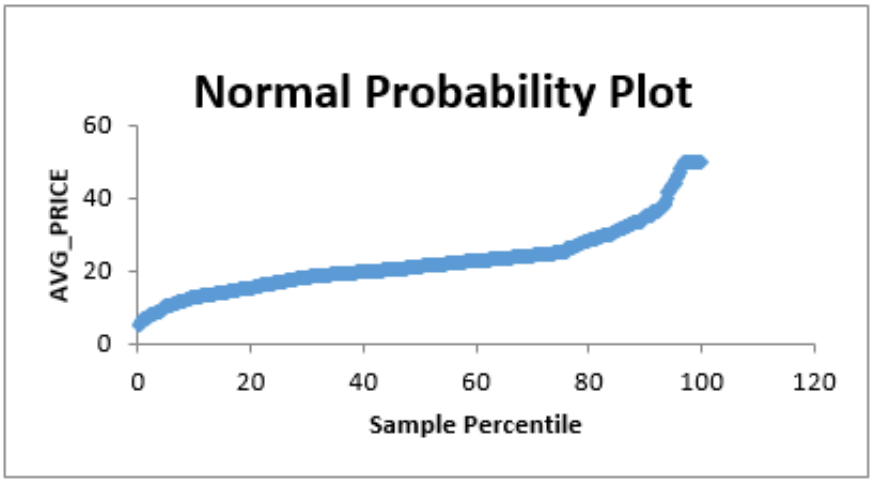
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	23243.914	23243.91	601.61787	5.0811E-88
Residual	504	19472.38142	38.63568		
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	61.41515	3.74E-236	33.44845704	35.659225	33.448457	35.6592247
LSTAT	-0.950049354	0.038733416	-24.5279	5.081E-88	-1.0261482	-0.8739505	-1.0261482	-0.87395051


LSTAT Residual Plot
Residuals
LSTAT
25
20
15
10
5
0
-5
-10
-15
-20
0
5
10
15
20
25
30
35
40


LSTAT Line Fit Plot
AVG_PRICE
LSTAT
AVG_PRICE
Predicted AVG_PRICE
60
50
40
30
20
10
0
-10
0
10
20
30
40


Normal Probability Plot
AVG_PRICE
Sample Percentile
60
40
20
0
0
20
40
60
80
100
120

a) INFERENCE FROM THE REGRESSION SUMMARY

R^2 is the square of correlation which is dimensionless. It is also known as coefficient of determination. R^2 measures how much of variance in independent variable (X) can affect the variance of dependent variable (Y). In other words, with the independent variable(X), I can predict the probability of the dependent variable(Y) using R^2.

R^2 is 54% which means that the LSTAT variable accounts for 54% of the variance in the average price of the houses(Y).

Coefficient shows the impact of a particular variable in the entire model. The coefficient value is -0.95 which indicates that if the independent variable(x) increases, the dependent variable (y)decreases as it has a negative coefficient. A negative coefficient shows inverse relationship

For instance, if LSTAT increases, the average price decreases.

The intercept indicates the amount of unexplainable behavior that are not given as input in the model. Here, the intercept is 34.55. It is the constant number.

For instance, when we predict the dependent variable(Y) with the independent variable(X), a constant error will be occurred because of the missing independent variables(x). This constant error can also be the intercept.

If independent variable(X) is able to predict 100% of the dependent variable(Y), then the intercept will be zero where X = Y.

From this residual plot, we can infer that there are more data points that deviate from the model from 0 to 5 and above 25 in x axis (LSTAT). In other words, there is a bias in 0 to 5 and above 25.

If the LSTAT value is less than 5 or above 25, then do not consider this model.

From 5 to 25, the data points are randomly distributed.

Line plot graph shows that a negative slope is present.

b) SIGNIFICANT VARIABLE

Based on my model, LSTAT is a significant variable as the P value is less than 0.05.

P value is also known as actual significance level. P value gives the actual risk or level of significance by which the null hypothesis is rejected.

Where, alpha = 1 - confidence level

confidence level = 95%

alpha = 1- 95%

alpha = 5%

Thus, the P value should be less than 5% or 0.05

Here, the P value of LSTAT is 0% (less than 5%) hence it is significant.

6. REGRESSION MODEL WITH 2 INDEPENDENT VARIABLES

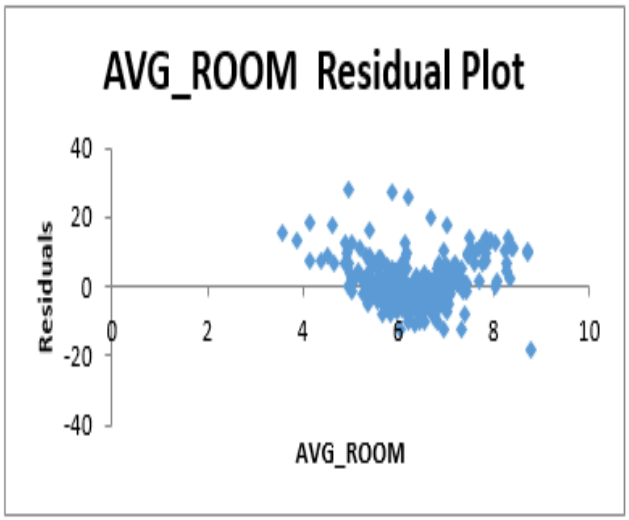
- INDEPENDENT VARIABLE (X1) – AVERAGE ROOM
- INDEPENDENT VARIABLE (X2) - LSTAT
- DEPENDENT VARIABLE (Y) – AVERAGE PRICE

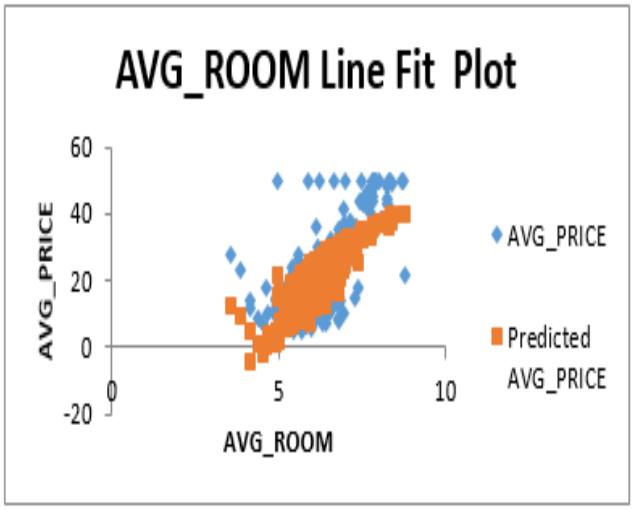
Regression Statistics	
Multiple R	0.799100498
R Square	0.638561606
Adjusted R Square	0.637124475
Standard Error	5.540257367
Observations	506

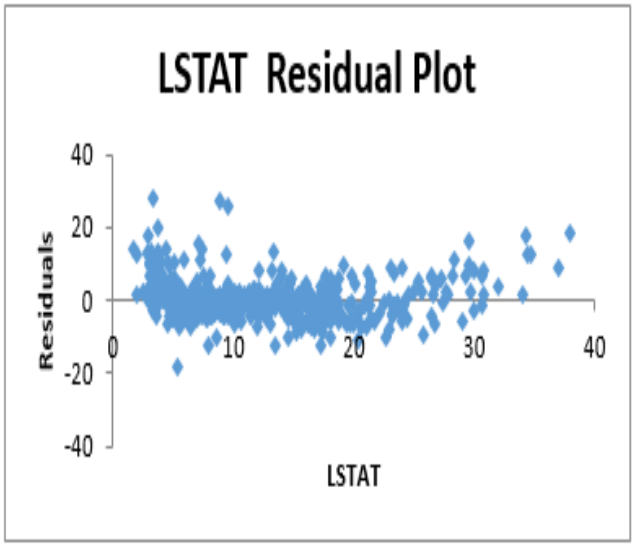
ANOVA

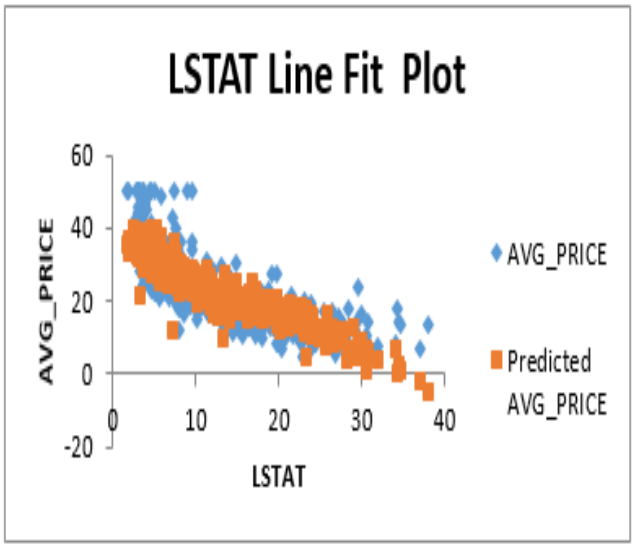
	df	*SS*	*MS*	*F*	*ignificance F*
Regression	2	27276.98621	13638.4931	444.330892	7.01E-112
Residual	503	15439.3092	30.6944517		
Total	505	42716.29542			

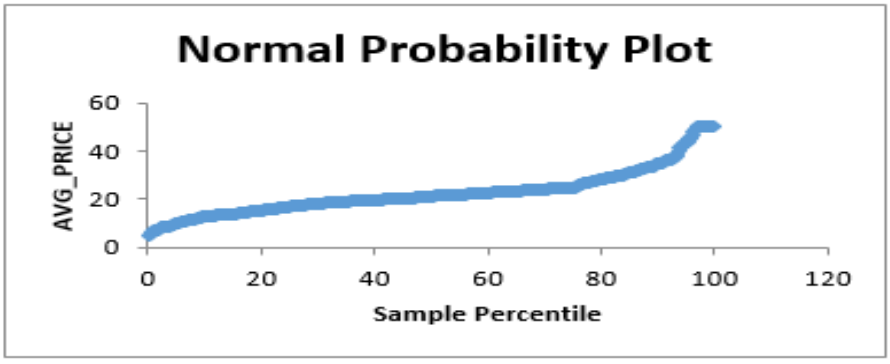
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.358272812	3.17282778	-0.4280953	0.66876494	-7.5919	4.87535466	-7.59190028	4.87535466
AVG_ROOM	5.094787984	0.4444655	11.4627299	3.4723E-27	4.2215504	5.96802553	4.22155044	5.96802553
LSTAT	-0.642358334	0.043731465	-14.688699	6.6694E-41	-0.728277	-0.5564395	-0.72827717	-0.5564395


AVG_ROOM Residual Plot
Residuals
AVG_ROOM
40
20
0
-20
-40
0
2
4
6
8
10


AVG_ROOM Line Fit Plot
AVG_PRICE
AVG_ROOM
60
40
20
0
-20
0
5
10
AVG_PRICE
Predicted AVG_PRICE


LSTAT Residual Plot
Residuals
LSTAT
40
20
0
-20
-40
0
10
20
30
40


LSTAT Line Fit Plot
AVG_PRICE
LSTAT
60
40
20
0
-20
0
10
20
30
40
AVG_PRICE
Predicted AVG_PRICE


Normal Probability Plot
AVG_PRICE
Sample Percentile
60
40
20
0
0
20
40
60
80
100
120

a) REGRESSION EQUATION

Y = m1*X1 + m2*X2 + b

Where m is Slope,

X is Independent variable,

Y is Dependent variable,

b is Intercept,

Y = 5.094 * Average number of rooms – 0.642 * LSTAT – 1.358

By Given condition, if

Average rooms	7	Intercept	-1.35827
LSTAT	20	Coefficient of Average rooms	5.094788
Average price	?	Coefficient of LSTAT	-0.64236
Predicted Y	5.094*7 - 0.642*20 - 1.358		
Average price (predicted Y)	21.458		

Here, when we compare our predicted average price of $21000 with the company pricing $30000, we can infer that there is a big difference if $9000.

The company which is pricing $30000 is clearly OVERCHARGING. Hence, I would recommend not to accept this quotation of $30000 since it is highly priced.

b) COMPARISON

Yes, this model (Built using avg room, LSTAT, avg price) is better than the previous model (Built using LSTAT and average price)

Adjusted R^2 changes based on three important values namely,

- R^2,
- Number of independent variables and
- Number of samples (observation).

	LSTAT(X) and AVERAGE PRICE(Y)	AVERAGE ROOMS (X1), LSTAT(X2) and AVERAGE PRICE(Y)
R^2	54.4%	63.8%
NO. OF INDEPENDENT VARIABLE	1	2
NO. OF SAMPLES	506	506

From the above table, it is clearly seen that the R^2 and number of independent variable changes and hence adjusted R^2 also increases.

Here, the Adjusted R^2 of the previous model is 54.4% whereas the Adjusted R^2 of this model is 63.8 %. Higher the Adjusted R^2, more better the model as it has higher impact on the model.

Using Adjusted R^2, more accurate answers can be predicted.

7. REGRESSION MODEL WITH ALL THE INDEPENDENT VARIABLES

- DEPENDENT VARIABLE (Y) – AVERAGE PRICE
- INDEPENDENT VARIABLE (X) – ALL OTHER VARIABLES

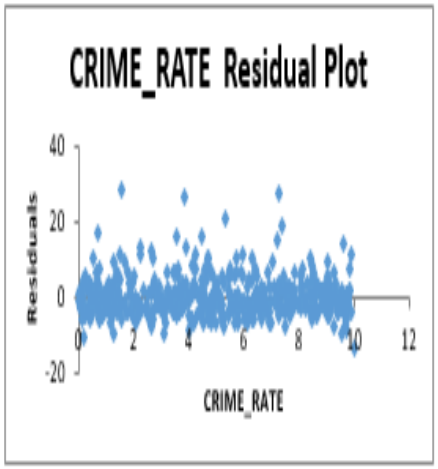
Regression Statistics	
Multiple R	0.83297882
R Square	0.69385372
Adjusted R Squa	0.68829865
Standard Error	5.1347635
Observations	506

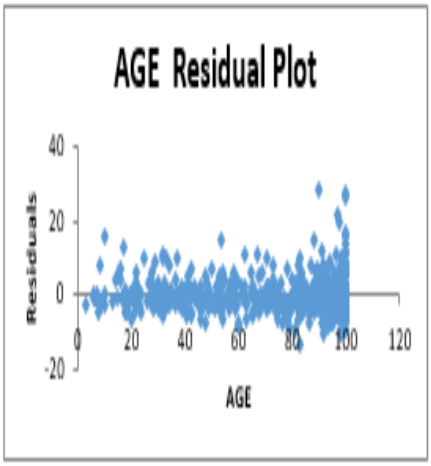
ANOVA

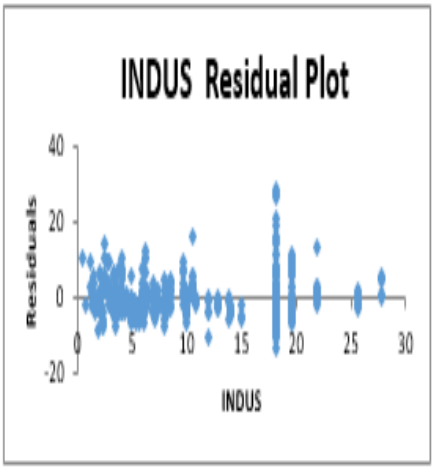
	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	29638.8605	3293.21	124.905	1.933E-121
Residual	496	13077.43492	26.3658		
Total	505	42716.29542			

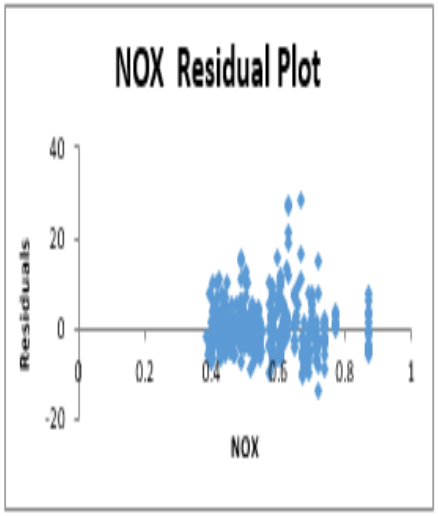
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.2413153	4.817125596	6.07028	2.54E-09	19.7768278	38.7058	19.7768278	38.705803
CRIME_RATE	0.04872514	0.078418647	0.62135	0.53466	-0.10534854	0.202799	-0.1053485	0.2027988
AGE	0.03277069	0.013097814	2.502	0.01267	0.00703665	0.058505	0.00703665	0.0585047
INDUS	0.1305514	0.063117334	2.06839	0.03912	0.00654109	0.254562	0.00654109	0.2545617
NOX	-10.321183	3.894036256	-2.6505	0.00829	-17.9720228	-2.67034	-17.972023	-2.6703428
DISTANCE	0.26109357	0.067947067	3.8426	0.00014	0.12759401	0.394593	0.12759401	0.3945931
TAX	-0.0144012	0.003905158	-3.6877	0.00025	-0.02207388	-0.00673	-0.0220739	-0.0067285
PTRATIO	-1.0743053	0.133601722	-8.0411	6.6E-15	-1.33680044	-0.81181	-1.3368004	-0.8118103
AVG_ROOM	4.12540915	0.442758999	9.3175	3.9E-19	3.25549474	4.995324	3.25549474	4.9953236
LSTAT	-0.6034866	0.053081161	-11.369	8.9E-27	-0.70777824	-0.49919	-0.7077782	-0.4991949

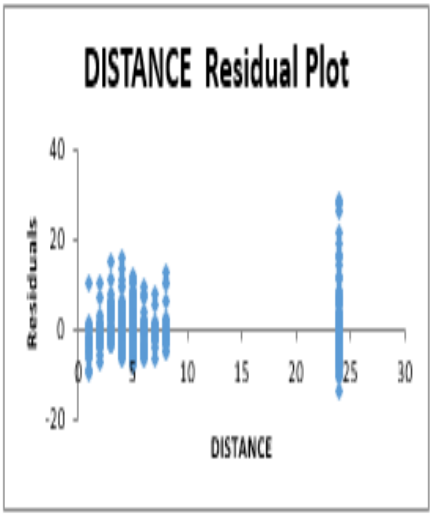
RESIDUAL PLOTS

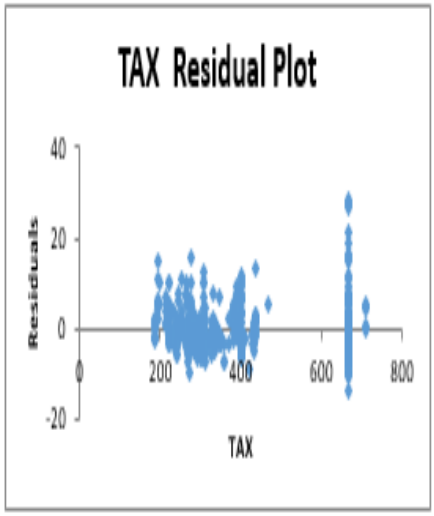


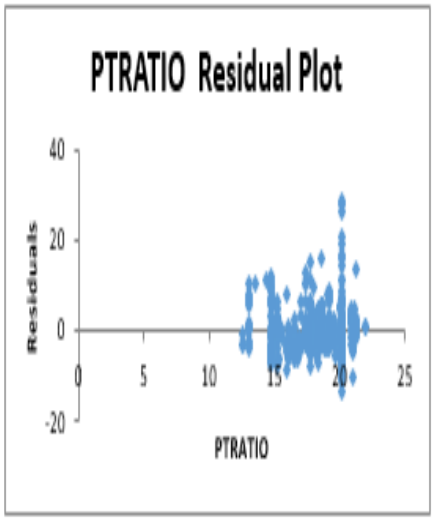


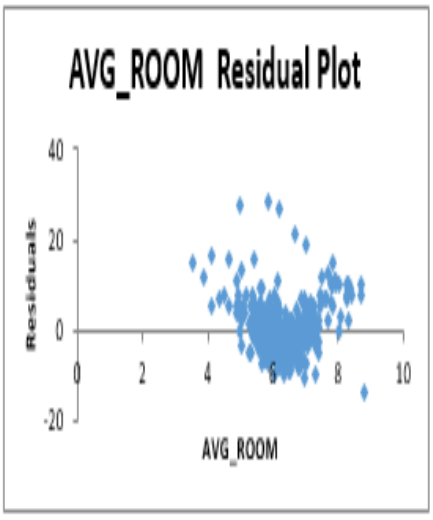


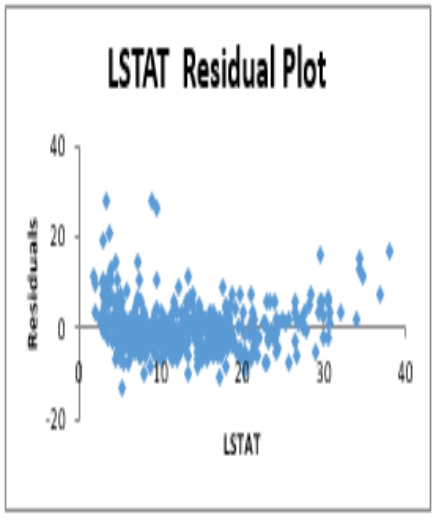




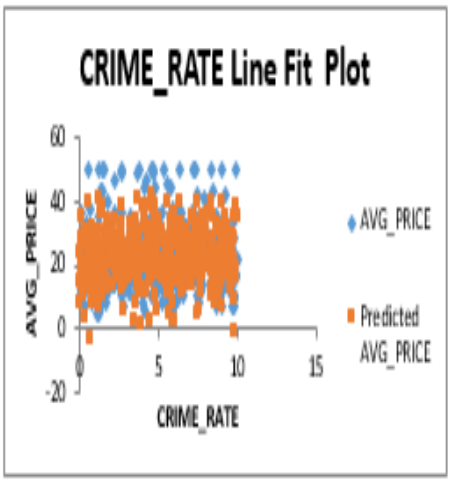


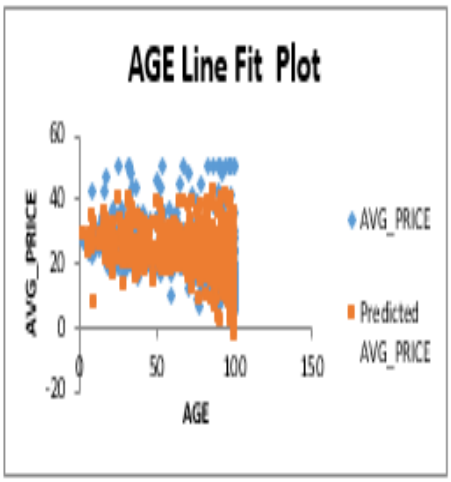


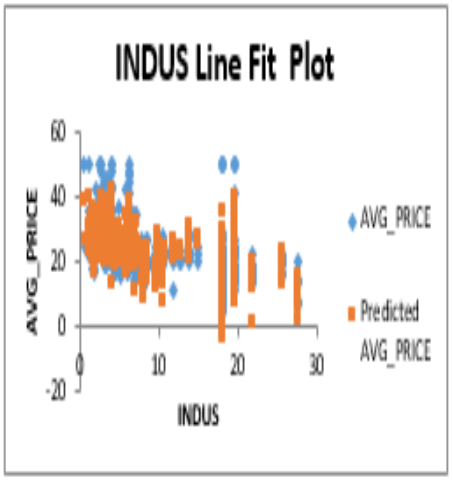


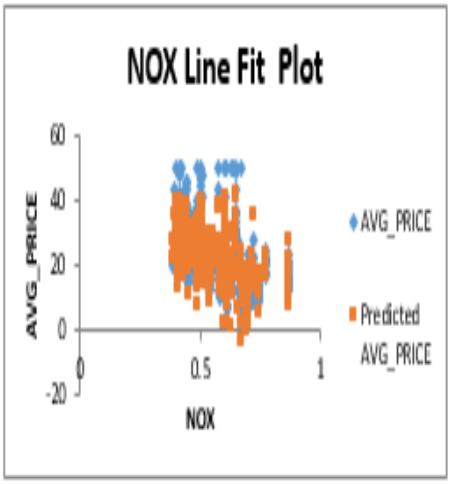


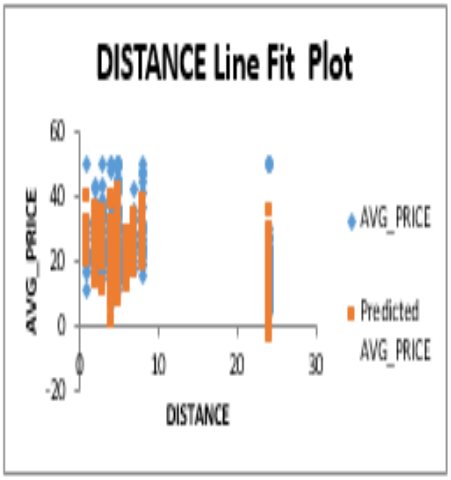
LINE FIT PLOT

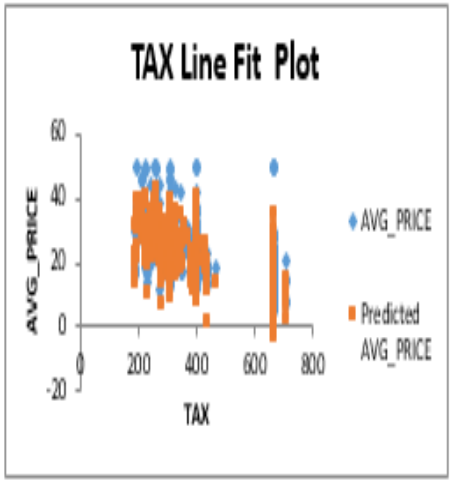


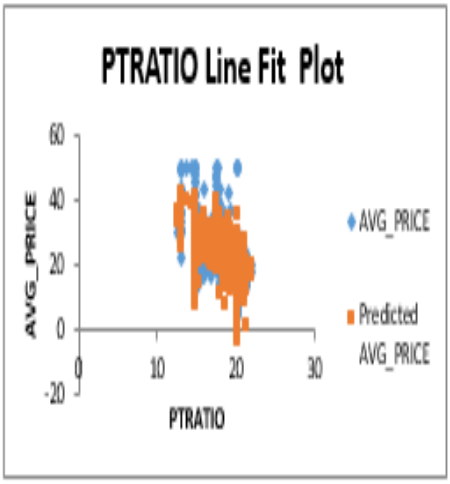


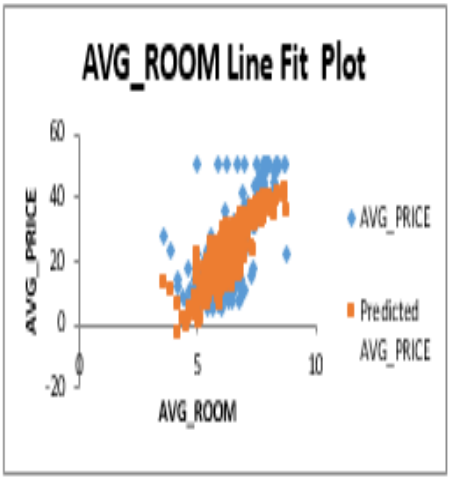


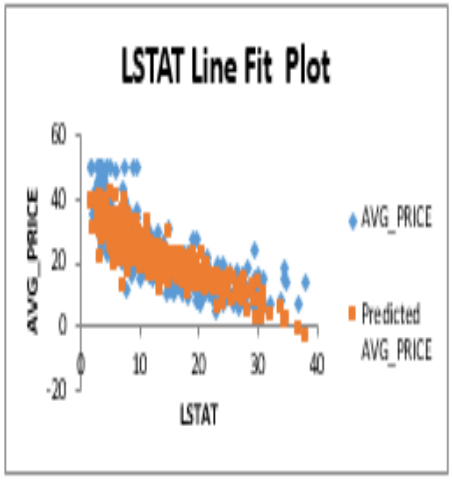




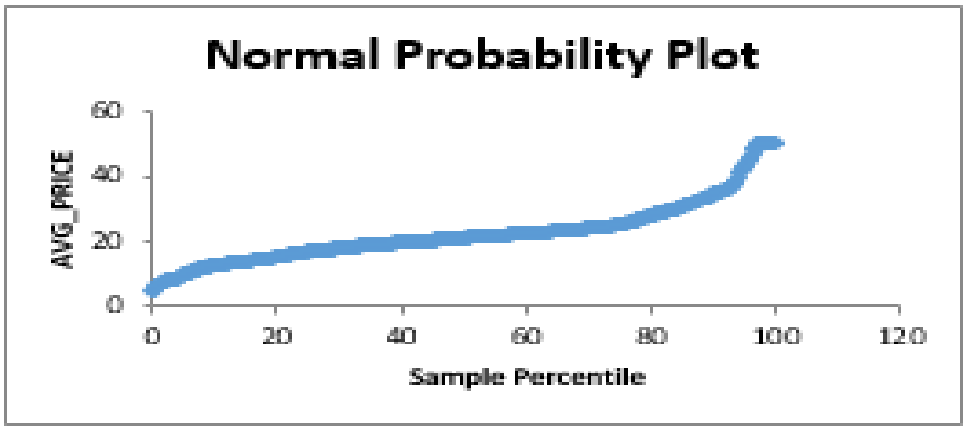








NORMAL PROBABILITY PLOT



INFERENCE

The adjusted R^2 is 68.8% which is higher than the previously built models.

The coefficients of the independent variable shows the impact that each particular variable has on the dependent variable and in the built model.

Here, the crime rate should be excluded, as it as a p value of 53% which is more than the 5% and hence it is insignificant.

though the coefficient of crime rate shows a positive relationship, it cannot be considered as it is insignificant.

the other coefficients of independent variable such as age, industry, distance, average room shows positive coefficients and hence if the independent variable(X) increases, dependent variable (Y) also increases .

Independent variables such as NOX, tax, PRATIO, LSTAT shows negative coefficients which means if the independent variable(X) increases, dependent variable (Y) decreases.

.

For instance, if there is a increase in these variables, then our average price automatically increases

- Age
- Indus
- Distance
- Average room

if there is a increase in these variables, then our average price automatically decreases,

- NOX
- Tax
- PRATIO
- LSTAT

The intercept is the point where the trendline cuts the y axis. Intercept indicates the amount of unexplainable behavior that are not given as input in the model. Here, the intercept is 29.24. It is the constant number.

For instance, when we predict the dependent variable(Y) with the independent variable(X), a constant error will be occurred because of the missing independent variables(x). This constant error can also be the intercept.

NOX has the highest negative impact on the average price.

Average room has the highest positive impact on the average price.

8. FINAL REGRESSION MODEL

- DEPENDENT VARIABLE (Y) – AVERAGE PRICE
- INDEPENDENT VARIABLE (X) – ONLY THE SIGNIFICANT VARIABLES

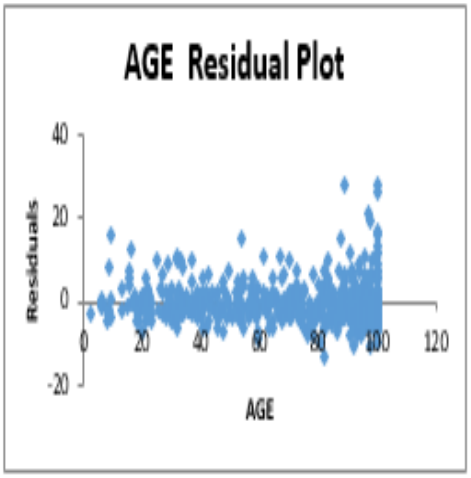
Regression Statistics	
Multiple R	0.832836
R Square	0.693615
Adjusted R Square	0.688684
Standard Error	5.131591
Observations	506

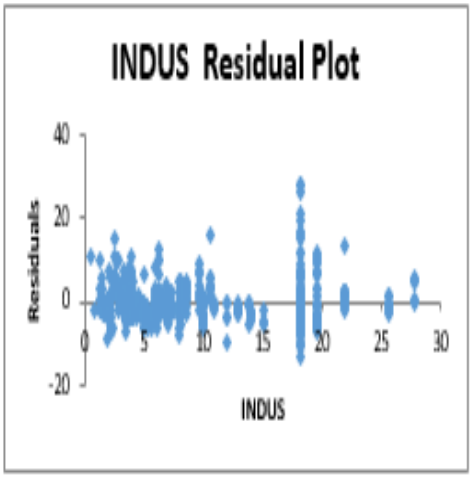
ANOVA

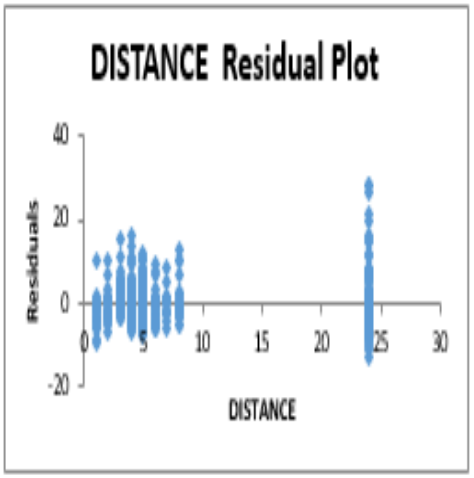
	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	29628.68142	3703.5852	140.64304	1.91E-122
Residual	497	13087.61399	26.333227		
Total	505	42716.29542			

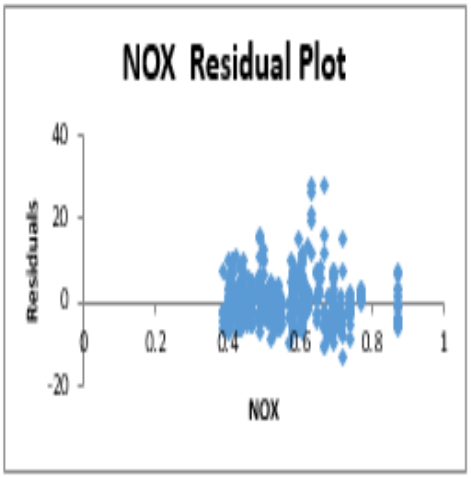
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.42847	4.804728624	6.1248982	1.846E-09	19.98839	38.868557	19.98839	38.8685574
AGE	0.032935	0.013087055	2.516606	0.0121629	0.0072222	0.0586477	0.0072222	0.058647734
INDUS	0.13071	0.063077823	2.0722023	0.0387617	0.0067779	0.2546421	0.0067779	0.254642071
NOX	-10.27271	3.890849222	-2.6402218	0.0085457	-17.917246	-2.628164	-17.91725	-2.62816447
DISTANCE	0.261506	0.067901841	3.851242	0.0001329	0.1280964	0.3949165	0.1280964	0.394916471
TAX	-0.014452	0.003901877	-3.7039464	0.0002361	-0.0221186	-0.006786	-0.022119	-0.00678614
PTRATIO	-1.071702	0.133453529	-8.0305293	7.083E-15	-1.3339051	-0.8095	-1.333905	-0.80949984
AVG_ROOM	4.125469	0.44248544	9.3234005	3.69E-19	3.2560963	4.9948416	3.2560963	4.994841615
LSTAT	-0.605159	0.0529801	-11.422388	5.418E-27	-0.7092519	-0.501067	-0.709252	-0.5010667

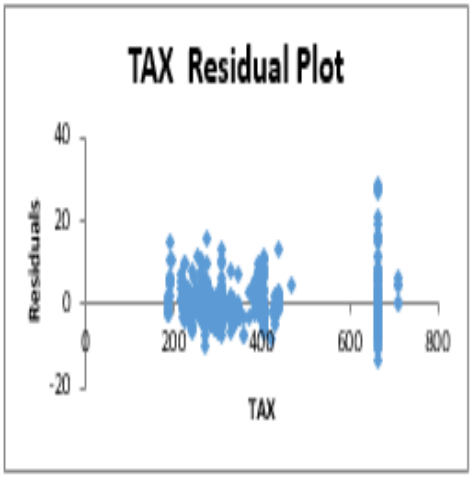
RESIDUAL PLOT

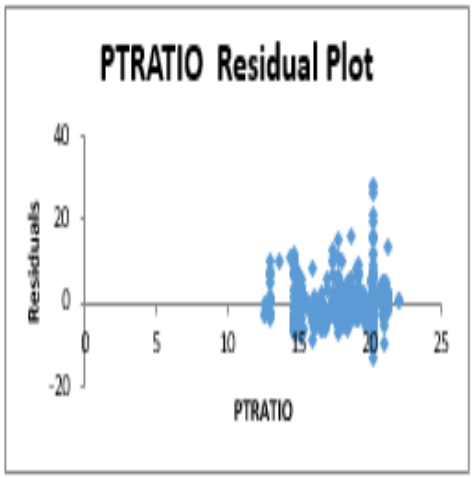


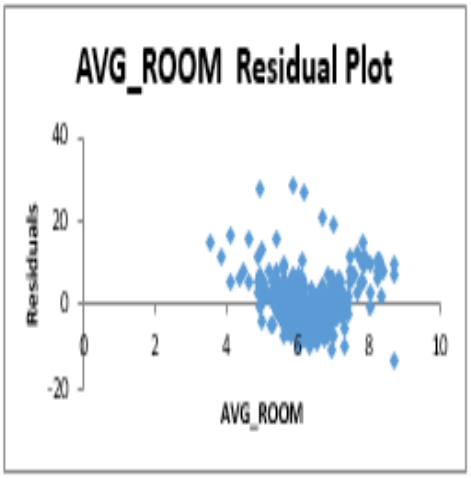


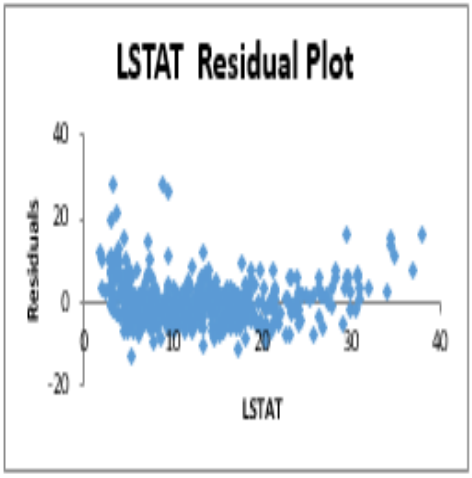




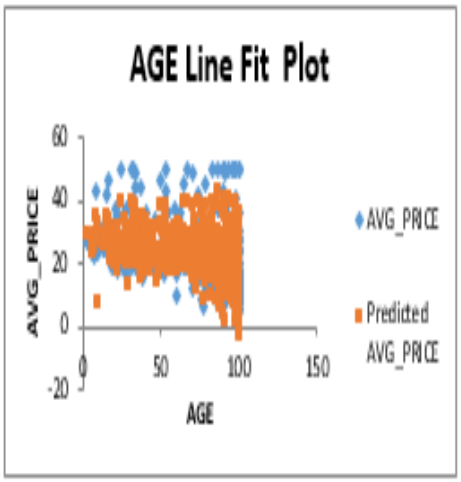


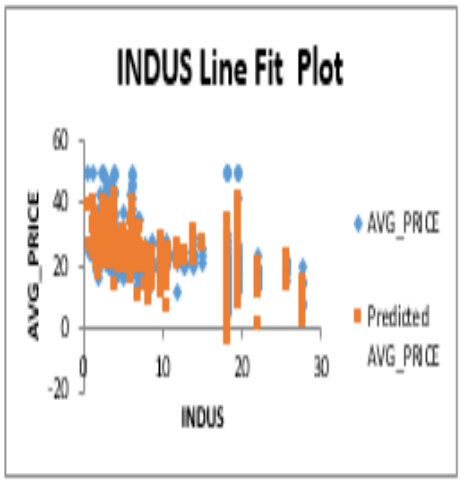


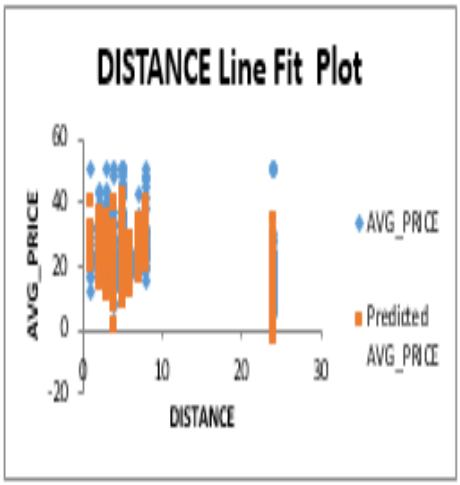


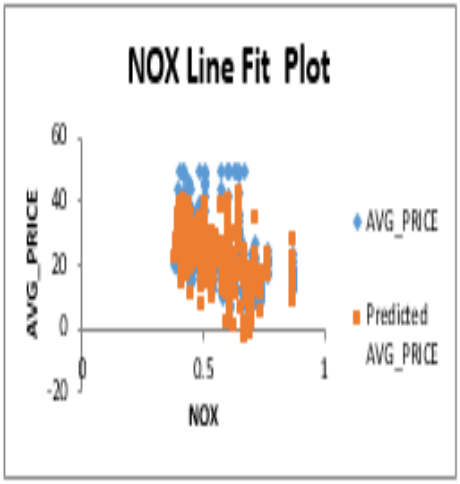


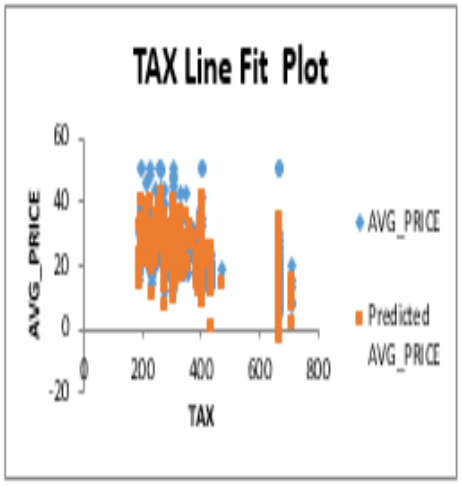
LINE FIT PLOT

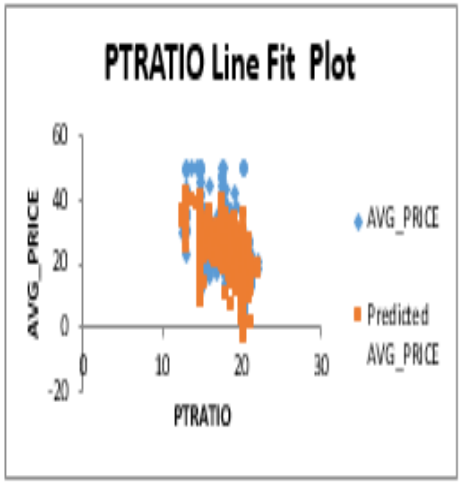


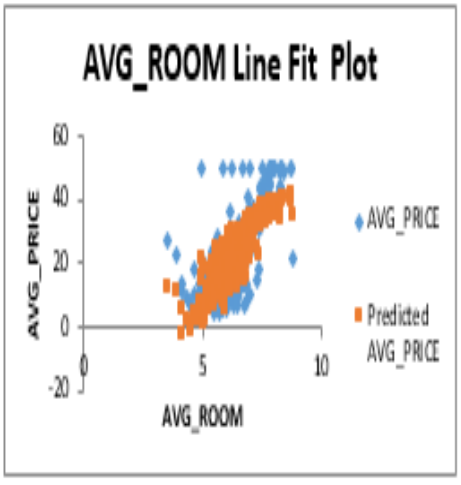


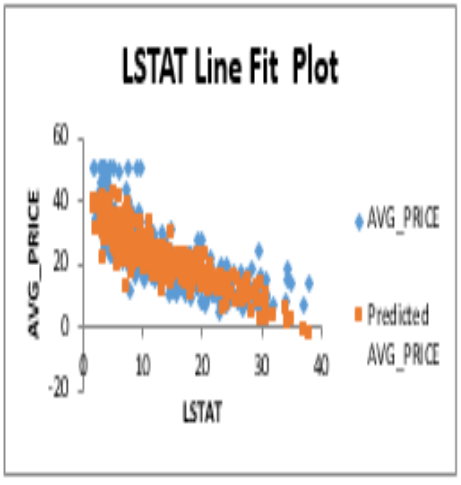


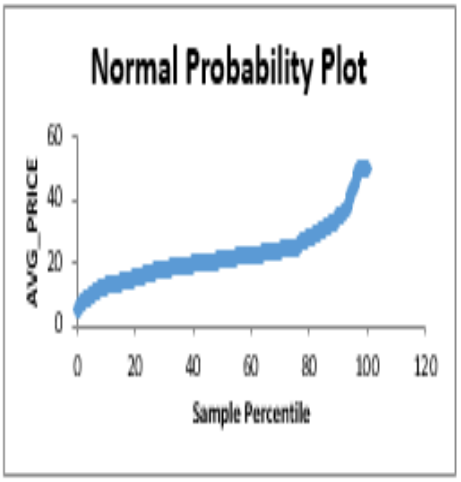












INFERENCE

a) In this model, the crime rate has been removed since p value is above 0.05 which makes it insignificant. Compared to previous models, this model gives better prediction as the adjusted R^2 is high. It shows a slight less R^2 in this model compared to the previous model as the previous model has more number of independent variables but it cannot be considered as a better model as it contains a insignificant variable.

b)

Previous model (without removing the insignificant values)	This model (with removing the insignificant values)
Adjusted R^2 = 0.6882 (68.83%)	Adjusted R^2 = 0.6886 (68.87%)
Lesser adjusted R^2 comparatively	Slightly higher Adjusted R^2 comparatively
Less impact on average price Y comparatively	High impact on average price Y and hence Performs better than previous model

The adjusted r2 is slightly higher after removing the insignificant values.

Hence the model in which the insignificant values are removed perform better comparatively.

c) COEFFICIENTS IN ASCENDING ORDER

INDEPENDENT VARIABLE	Coefficient
NOX	-10.27270508
PTRATIO	-1.071702473
LSTAT	-0.605159282
TAX	-0.014452345
AGE	0.03293496
INDUS	0.130710007
DISTANCE	0.261506423
AVG_ROOM	4.125468959

If the value of NOX is increased then the average price will decrease which will impact a huge loss.

since NOX has a negative coefficient, it has a inverse relation with the average price (Y). Increasing the independent variable, decreases the dependent variable.

d) REGRESSION EQUATION OF THIS MODEL:

Y = m1*X1 + m2*X2 + m3*X3 + m4*X4 + m5*X5 + m6*X6 + m7*X7 + m8*X8 + b

Y = -10.272 * NOX(X1) -1.071 * PRATIO(X2) - 0.605 * LSTAT(X3) - 0.014 * TAX(X4) + 0.032 * AGE(X5) + 0.1307 * INDUS(X6) + 0.261 * DISTANCE(X7) + 4.125 * AVERAGE ROOM(X8) + 29.428

Using this formula, the predicted Y is calculated. For your reference, the average price (predicted Y) of first 25 houses are calculated.

S.NO.	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE (ACTUAL Y)	PREDICTED Y
1	65.2	2.31	0.538	1	296	15.3	6.575	4.98	24	30.0488873
2	78.9	7.07	0.469	2	242	17.8	6.421	9.14	21.6	27.0409846
3	61.1	7.07	0.469	2	242	17.8	7.185	4.03	34.7	32.6989645
4	45.8	2.18	0.458	3	222	18.7	6.998	2.94	33.4	31.1430695
5	54.2	2.18	0.458	3	222	18.7	7.147	5.33	36.2	30.5880873
6	58.7	2.18	0.458	3	222	18.7	6.43	5.21	28.7	27.8509525
7	66.6	7.87	0.524	5	311	15.2	6.012	12.43	22.9	25.0708969
8	96.1	7.87	0.524	5	311	15.2	6.172	19.15	27.1	22.6358829
9	100	7.87	0.524	5	311	15.2	5.631	29.93	16.5	14.0088334
10	85.9	7.87	0.524	5	311	15.2	6.004	17.1	18.9	22.847444
11	94.3	7.87	0.524	5	311	15.2	6.377	20.45	15	22.635614
12	82.9	7.87	0.524	5	311	15.2	6.009	13.27	18.9	25.0870265
13	39	7.87	0.524	5	311	15.2	5.889	15.71	21.7	21.6695368
14	61.8	8.14	0.538	4	307	21	5.949	8.26	20.4	20.6483212
15	84.5	8.14	0.538	4	307	21	6.096	10.26	18.2	20.7920702
16	56.5	8.14	0.538	4	307	21	5.834	8.47	19.9	19.8722535
17	29.3	8.14	0.538	4	307	21	5.935	6.58	23.1	20.536846
18	81.7	8.14	0.538	4	307	21	5.99	14.67	17.5	17.5938001
19	36.6	8.14	0.538	4	307	21	5.456	11.69	20.2	15.7088076
20	69.5	8.14	0.538	4	307	21	5.727	11.28	18.2	18.1584852
21	98.1	8.14	0.538	4	307	21	5.57	21.02	13.6	12.5584751
22	89.2	8.14	0.538	4	307	21	5.965	13.83	19.6	18.2460094
23	91.7	8.14	0.538	4	307	21	6.142	18.72	15.2	16.0993259
24	100	8.14	0.538	4	307	21	5.813	19.88	14.5	14.313422
25	94.1	8.14	0.538	4	307	21	5.924	16.3	15.6	16.743503

9. CONCLUSION

In this project, I can clearly infer that the Adjusted R^2 is slowly increasing from the initial regression model to final regression model.

- Initial regression model has a R^2 of 54.4% and Adjusted R^2 of 54.3%.
- Regression model using two independent variable(X) and a dependent variable(Y) has a R^2 of 63.8% and Adjusted R^2 of 63.7%.
- Regression model using all the independent and dependent variables has a R^2 of 69.38% and Adjusted R^2 of 68.82%.
- Final regression model (removed insignificant variable) has a R^2 of 69.36% and Adjusted R^2 of 68.86%.

The final regression model shows a slight less R^2 compared to the previous model as the previous model has more number of independent variables but it cannot be considered as a better model as it contains a insignificant variable.

Based this R^2 and Adjusted R^2, I can conclude that the final regression model built is better where the crime rate (one of the independent variables) is removed as it is insignificant. This shows that the final regression model has % of impact on the average price of the houses (dependent variable)

The features that are to be taken into major consideration are,

1) Average number of rooms – which has the highest positive coefficient and hence if independent variable(X) increases, average price (dependent variable(Y)) also increases. Hence, it has high magnitude on affecting the average price of the house.

   Distance, Indus, Age also shows similar positive coefficient.

2) NOX (Nitrogen Oxide Concentration) – which has the highest negative coefficient and hence if independent variable(X) increases, average price (dependent variable(Y)) also decreases. Hence, it has high magnitude on affecting the average price by causing loss.

PRATIO, LSTAT, Tax shows similar negative coefficient.

THANK YOU