Assignment – Terro's real estate agency Real estate data analysis – Exploratory data analysis, Linear Regression

Problem Statement:

"**Finding out the most relevant features for pricing of a house**"

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

. The agency has provided a dataset of 506 houses in Boston. Following are the details of the dataset:

Data Dictionary:

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

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

CRIME_RATE		*AGE*		*INDUS*	
Mean	4.871976285	Mean	68.57490119	Mean	11.13677866
Standard Error	0.129860152	Standard Error	1.251369525	Standard Error	0.304979888
Median	4.82	Median	77.5	Median	9.69
Mode	3.43	Mode	100	Mode	18.1
Standard Deviation	2.921131892	Standard Deviation	28.14886141	Standard Deviation	6.860352941
Sample Variance	8.533011532	Sample Variance	792.3583985	Sample Variance	47.06444247
Kurtosis	-1.189122464	Kurtosis	-0.967715594	Kurtosis	-1.233539601
Skewness	0.021728079	Skewness	-0.59896264	Skewness	0.295021568
Range	9.95	Range	97.1	Range	27.28
Minimum	0.04	Minimum	2.9	Minimum	0.46
Maximum	9.99	Maximum	100	Maximum	27.74
Sum	2465.22	Sum	34698.9	Sum	5635.21
Count	506	Count	506	Count	506

- ***CRIME_RATE***: 50percent of crime rated between 0.04 and 4.82 and rest of 50% crime rate have between 4.82 and 9.99,it seems top 50% crime rate is high. Probably few High Crime rate by towns are present in the dataset.
- ***AGE***: the range of values of the age deplicts the data is recorded for the most number houses means proportion houses built prior to 1940. and age data is showing negative skewness means, median is greater than the mean(median > mean
- ***INDUS:*** 50% of indus of non-retail business acres between 0.46 and 9.69 and rest of 50% non-retail business acres per town between 9.69 and 27.74,it seems 50% of non -retail business acres is quite high**.**

NOX		*DISTANCE*		*TAX*	
Mean	0.554695059	Mean	9.549407115	Mean	408.2371542
Standard Error	0.005151391	Standard Error	0.387084894	Standard Error	7.492388692
Median	0.538	Median	5	Median	330
Mode	0.538	Mode	24	Mode	666
Standard Deviation	0.115877676	Standard Deviation	8.707259384	Standard Deviation	168.5371161
Sample Variance	0.013427636	Sample Variance	75.81636598	Sample Variance	28404.75949
Kurtosis	-0.064667133	Kurtosis	-0.867231994	Kurtosis	-1.142407992
Skewness	0.729307923	Skewness	1.004814648	Skewness	0.669955942
Range	0.486	Range	23	Range	524
Minimum	0.385	Minimum	1	Minimum	187
Maximum	0.871	Maximum	24	Maximum	711
Sum	280.6757	Sum	4832	Sum	206568
Count	506	Count	506	Count	506

- **NOX**: minimum nitric oxides concentration is 0.385ppm and maximum nitric oxides concentration is 0.871ppm.Median Nitric oxides concentration is 0.538ppm suggesting that 50% of the data are less than 0.538ppm
- **Distance :** minimum distance for home to highways is 1miles and maximum distance for home to highways is 24miles. Median distance is 5miles suggesting that 50% of the distance for home to highways is 5miles.
- **Tax:** 50% of tax for full property value between 187 and 330 and rest of 50% of tax for full property value between 187 and 711.it seems 50% of full property value tax is quite high.

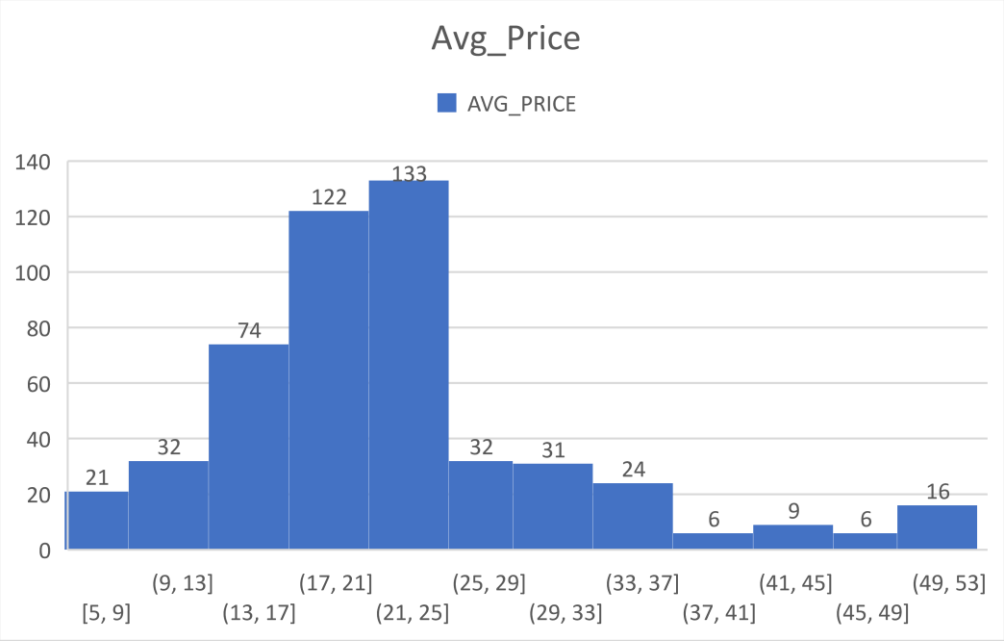
PTRATIO		*AVG_ROOM*		*LSTAT*	
Mean	18.4555336	Mean	6.284634387	Mean	12.65306324
Standard Error	0.096243568	Standard Error	0.031235142	Standard Error	0.317458906
Median	19.05	Median	6.2085	Median	11.36
Mode	20.2	Mode	5.713	Mode	8.05
Standard Deviation	2.164945524	Standard Deviation	0.702617143	Standard Deviation	7.141061511
Sample Variance	4.686989121	Sample Variance	0.49367085	Sample Variance	50.99475951
Kurtosis	-0.285091383	Kurtosis	1.891500366	Kurtosis	0.493239517
Skewness	-0.802324927	Skewness	0.403612133	Skewness	0.906460094
Range	9.4	Range	5.219	Range	36.24
Minimum	12.6	Minimum	3.561	Minimum	1.73
Maximum	22	Maximum	8.78	Maximum	37.97
Sum	9338.5	Sum	3180.025	Sum	6402.45
Count	506	Count	506	Count	506

- **Average room**: average minimum rooms are 3.56 and average maximum room is 8.78.Median of average room is 6.025 suggesting that 50% of average rooms are less than 6.025.
- **LSTAT**: minimum lowest status population is 1.73 and maximum lowest status population is 37.97. median of lowest status population is 11.36 it suggesting 50% of LSTST is less than 11.36.
- **PTRATIO**: pupil-teacher ratio by town its showing data is negative skewness ,means median is greater than the mean.

AVG_PRICE	
Mean	22.53280632
Standard Error	0.408861147
Median	21.2
Mode	50
Standard Deviation	9.197104087
Sample Variance	84.58672359
Kurtosis	1.495196944
Skewness	1.108098408
Range	45
Minimum	5
Maximum	50
Sum	11401.6
Count	506

- **Average price**: The prices of the house indicated by the variable average price is our target variable and remaining are the Independent variable based on which we will predict the value of houses.

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



We can see the highest spike in the 21-25 bin, and the lowest one in 37-41 and 45-49 bin suggesting that most of the houses have Average price between 21-25 average price, and few of the houses have Average price between 37-41 and 45-49 avg_price.

3) **Compute the covariance matrix. Share your observations**.

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*A*
CRIME_RATE	8.52									
AGE	0.56	790.79								
INDUS	-0.11	124.27	46.97							
NOX	0	2.3812	0.606	0.01						
DISTANCE	-0.23	111.55	35.48	0.62	75.7					
TAX	-8.23	2397.9	831.7	13	1333	28348.62				
PTRATIO	0.07	15.905	5.681	0.05	8.74	167.8208	4.6777			
AVG_ROOM	0.06	-4.743	-1.884	-0.02	-1.28	-34.5151	-0.54	0.493		
LSTAT	-0.88	120.84	29.52	0.49	30.3	653.4206	5.7713	-3.07	50.89	
AVG_PRICE	1.16	-97.4	-30.46	-0.45	-30.5	-724.82	-10.09	4.485	-48.4	

COV (X,Y) = ∑(x – x) (y – y) / n

- The covariance matrix is a square matrix to understand the relationships presented between the different variables in a dataset. It is easy and useful to show the covariance between two or more variables.

- The covariance will have both positive and negative values. A positive value indicates that two variables will decrease or increase in the same direction. A negative value indicates that if one variable decreases, the other increases, and an inverse relationship exist between them.

- The diagonal elements of the matrix contain the variances of the variables, and the off-diagonal elements contain the covariances between all possible pairs of variables.

- The relationship presented between AVG_PRICE and TAX have negative values, it indicates that one variable is decreases and the other variable is increases an inverse relationship exist between them.
- The relationship presented between TAX and AGE,INDUS,DISTANCE have positive values, a positive value indicates that two variables will decreases or increases in the same direction.

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

a) Which are the top 3 positively correlated pairs and

b) Which are the top 3 negatively correlated pairs.

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_*
CRIME_RATE	1									
AGE	0.007	1								
INDUS	-0.01	0.645	1							
NOX	0.002	0.731	0.764	1						
DISTANCE	-0.01	0.456	0.595	0.611	1					
TAX	-0.02	0.506	0.721	0.668	0.91	1				
PTRATIO	0.011	0.262	0.383	0.189	0.46	0.461	1			
AVG_ROOM	0.027	-0.24	-0.39	-0.302	-0.21	-0.29	-0.356	1		
LSTAT	-0.04	0.602	0.604	0.591	0.49	0.544	0.374	-0.61	1	
AVG_PRICE	0.043	-0.38	-0.48	-0.427	-0.38	-0.47	-0.508	0.7	-0.738	

- we create a correlation matrix that measures the linear relationships between the variables.
- the correlation coefficient ranges from -1 to 1 values is close to 1, it means that there is a strong positive correlation between the two variables.
- when it is close to -1, the variables have a strong negative correlation.
- to fit linear regression model, we select those features which have a high correlation with our target variable AVG_PRICE
- by looking at the correlation matrix we can see that AVG_ROOM has a strong positive correlation with AVG_PRICE(0.7)
- where as LSTAT has a high negative correlation with AVG_PRICE(-0.7376)

- AGE ,INDUS,NOX,DISTANCE,TAX,PTRATIO,LSTAT variables has a negative correlation with AVG_PRICE
- CRIME_RATE,AVG_ROOM variable has a positive correlation with AVG_PRICE

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 the Residual plot?

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

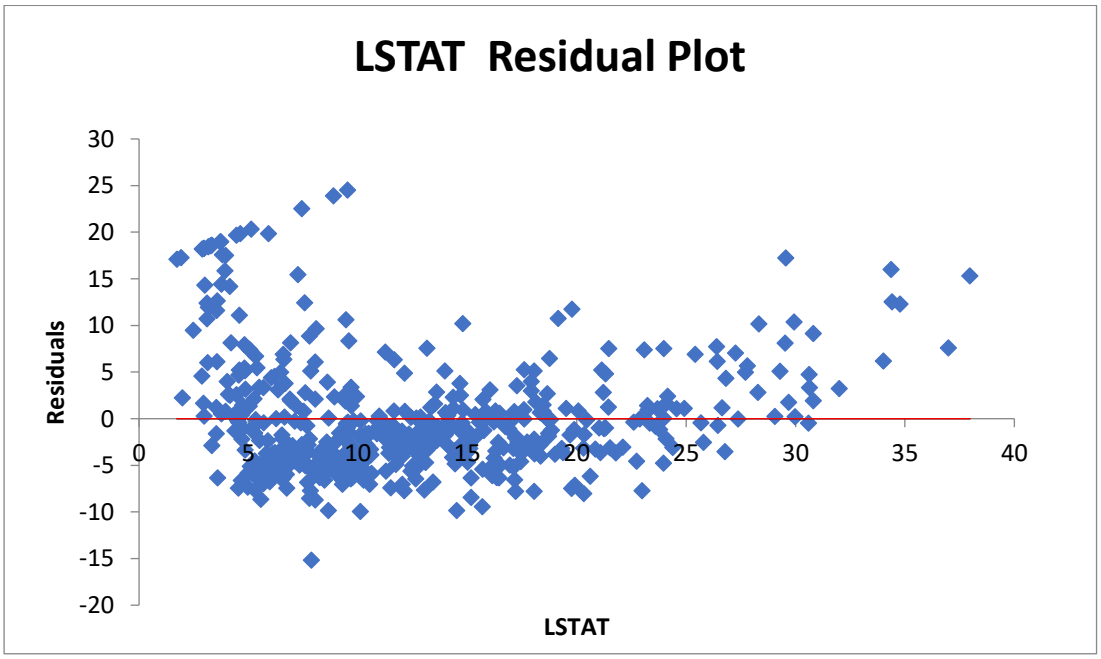
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.73766
R Square	0.54415
Adjusted R Square	0.54324
Standard Error	6.21576
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	23243.91	23243.914	601.618	5.1E-88
Residual	504	19472.38	38.635677		
Total	505	42716.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	34.5538	0.562627	61.415146	4E-236	33.4485	35.659	33.44846	35.65922
LSTAT	-0.95005	0.038733	-24.5279	5.1E-88	-1.02615	-0.874	-1.02615	-0.87395



- the multiple R is the correlation coefficient that measures the strength of a linear relationship between two variable.
- 0.74 Multiple R is the correlation coefficient its showing strong positive linear relationship between LATAT vs AVG_PRICE.
- R Square signifies the coefficient of determination, which shows the 0.55 percentage of variation in our AVG_PRICE that being explained by input variable LSTAT.
- Adjusted R square is the modified version of R Square that adjust for predictor that are not significant to the regression model.
- standard error is another goodness-of -fit measure that shows the precision of your regression analysis.
- anova stands for analysis of varience.it gives information about the levels of variability within regression model.
- The regression equation is a mathematical expression that represents the relationship between the dependent variable and independent variable .based on the estimation coefficients, the equation using: Y = -0.95X1 + 34.55
- we compare P-Value with significance value when P-value is lessthan significance value here is being reject null hypothesis,
- significance F is the P-value of F is lessthan significance value here is being reject null hypothesis and LATAT is significant.

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?

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.

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%*	*L*
Intercept	-1.358272812	3.17282778	0.428095348	0.668764941	-7.591900282	4.875354658	7
AVG_ROOM	5.094787984	0.4444655	11.46272991	3.47226E-27	4.221550436	5.968025533	4
LSTAT	-0.642358334	0.043731465	-14.68869925	6.66937E-41	-0.728277167	-0.556439501	-0

locality price
22.816352
20
7

- Compare to question 5 and question 6 the question 6 is more correlated and performance model also increases.

- the multiple R is the correlation coefficient that measures the strength of a linear relationship between two variable.
- 0.80 Multiple R is the correlation coefficient its showing strong positive linear relationship between LATAT ,AVG_ROOM vs AVG_PRICE.
- R Square signifies the coefficient of determination, which shows the 0.64 percentage of variation in our AVG_PRICE that being explained by input variable LSTAT and AVG_ROOM.
- Adjusted R square is the modified version of R Square that adjust for predictor that are not significant to the regression model.
- we can determine whether adding new independent variable AVG_ROOM we see a significant increases in R square value then the adjusted R square will also Increases.
- standard error is another goodness-of -fit measure that shows the precision of your regression analysis.
- anova stands for analysis of varience.it gives information about the levels of variability within regression model.
- The regression equation is a mathematical expression that represents the relationship between the dependent variable and independent variable .based on the estimation coefficients, the equation using: Y = -0.642X1 + 5.094X2 -1.358
- we compare P-Value with significance value when P-value is lessthan significance value here is being reject null hypothesis,
- significance F is the P-value of F is lessthan significance value here is being reject null hypothesis and LSTAT and AVG_ROOM is significant.
- the regression equation : Y = -0.642X1 + 5.094X2 - 1.358
- a new house in this locality has 7 rooms and has a value of 20 for L-STAT, the AVG_PRICE is 22.816,the company is **over charging** for this locality.

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.

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%*	*L*
Intercept	29.24131526	4.817125596	6.070282926	2.53978E-09	19.77682784	38.70580267	1
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.534657201	-0.105348544	0.202798827	0
AGE	0.032770689	0.013097814	2.501996817	0.012670437	0.00703665	0.058504728	
INDUS	0.130551399	0.063117334	2.068392165	0.03912086	0.006541094	0.254561704	0
NOX	-10.3211828	3.894036256	-2.650510195	0.008293859	-17.97202279	-2.670342809	1
DISTANCE	0.261093575	0.067947067	3.842602576	0.000137546	0.127594012	0.394593138	0
TAX	-0.01440119	0.003905158	-3.687736063	0.000251247	-0.022073881	-0.0067285	0
PTRATIO	-1.074305348	0.133601722	-8.041104061	6.58642E-15	-1.336800438	-0.811810259	1
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.89287E-19	3.255494742	4.995323561	3
LSTAT	-0.603486589	0.053081161	-11.36912937	8.91071E-27	-0.70777824	-0.499194938	-

- the multiple R is the correlation coefficient that measures the strength of a linear relationship between two variable.
- 0.83 Multiple R is the correlation coefficient its showing strong positive linear relationship between all the independent variable vs AVG_PRICE.
- R Square signifies the coefficient of determination, which shows the 0.69 percentage of variation in our AVG_PRICE that being explained by all the independent variable.
- Adjusted R square is the modified version of R Square that adjust for predictor that are not significant to the regression model.
- we can determine whether adding new independent variable to the model we see a significant increases in R square value then the adjusted R square will also Increases.
- standard error is another goodness-of -fit measure that shows the precision of your regression analysis.
- anova stands for analysis of varience.it gives information about the levels of variability within regression model.
- The regression equation is a mathematical expression that represents the relationship between the dependent variable and independent variable .based on the estimation coefficients, the equation using:
  :Y = 29.24 + 0.048X1 + 0.032X2 + 0.1305X3 - 10.32X4 + 0.26X5 - 0.014X6 - 1.074X7 + 4.125X8 -0.603X9
- we compare P-Value with significance value when P-value is lessthan significance value here is being reject null hypothesis,
- we compare P- value with significance value, when P-value is greater than significance value here we fail to reject null hypothesis,
- significance F is the P-value of F is lessthan significance value here is being reject null hypothesis and NOX,DISTANCE,TAX,PTRATIO,LSTAT and AVG_ROOM is significant.
- significance F is the P-value of F is Grete than significance value here we fail to reject null hypothesis and CRIME_RATE,AGE,INDUS is insignificant
- the regression equation :
  Y = 29.24 + 0.048X1 + 0.032X2 + 0.1305X3 - 10.32X4 + 0.26X5 - 0.014X6 - 1.074X7 + 4.125X8 -0.603X9

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:

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	0.828268
R Square	0.686028
Adjusted R Square	0.682252
Standard Error	5.184325
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	29304.56	4884.093	181.7186	4.4E-122
Residual	499	13411.74	26.87723		
Total	505	42716.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	23.25929	4.517194	5.149056	3.77E-07	14.38422	32.13435	14.38422	32.13435
NOX	-1.38436	2.981328	-0.46434	0.642605	-7.24186	4.473148	-7.24186	4.473148
DISTANCE	0.208401	0.065817	3.166373	0.001638	0.079089	0.337714	0.079089	0.337714
TAX	-0.01118	0.003603	-3.10253	0.002028	-0.01826	-0.0041	-0.01826	-0.0041
PTRATIO	-0.95676	0.130552	-7.32852	9.44E-13	-1.21326	-0.70026	-1.21326	-0.70026
AVG_ROOM	4.328075	0.433739	9.978528	1.67E-21	3.475896	5.180255	3.475896	5.180255
LSTAT	-0.54756	0.050117	-10.9255	4.64E-25	-0.64602	-0.44909	-0.64602	-0.44909

- Compare to question 7 and question 8 the question 8 is slightly less correlated and performance model also decreases.

- the multiple R is the correlation coefficient that measures the strength of a linear relationship between two variable.
- 0.82 Multiple R is the correlation coefficient its showing strong positive linear relationship between all the independent variable vs AVG_PRICE.
- R Square signifies the coefficient of determination, which shows the 0.68 percentage of variation in our AVG_PRICE that being explained by input variable NOX,DISTANCE,TAX,PTRATIO,AVG_ROOM,LSTAT.
- Adjusted R square is the modified version of R Square that adjust for predictor that are not significant to the regression model.

- If R Square does not increases significantly on the addition of new independent variable than the adjusted R square will actually Decreases.
- we can determine whether adding new independent variable to the model we see a significant increases in R square value then the adjusted R square will also Increases.
- standard error is another goodness-of -fit measure that shows the precision of your regression analysis.
- anova stands for analysis of varience.it gives information about the levels of variability within regression model.
- The regression equation is a mathematical expression that represents the relationship between the dependent variable and independent variable .based on the estimation coefficients, the equation using:
  Y = 23.25 - 1.38X1 + 0.208X2 - 0.011X3 - 0.956X4 + 4.328X5 - 0.547X6
- we compare P-Value with significance value when P-value is lessthan significance value here is being reject null hypothesis,
- we compare P- value with significance value, when P-value is greater than significance value here we fail to reject null hypothesis,
- significance F is the P-value of F is lessthan significance value here is being reject null hypothesis and DISTANCE,TAX,PTRATIO,LSTAT and AVG_ROOM is significant.
- significance F is the P-value of F is Grete than significance value here we fail to reject null hypothesis and NOX is insignificant
- the regression equation :
  Y = 23.25 - 1.38X1 + 0.208X2 - 0.011X3 - 0.956X4 + 4.328X5 - 0.547X6