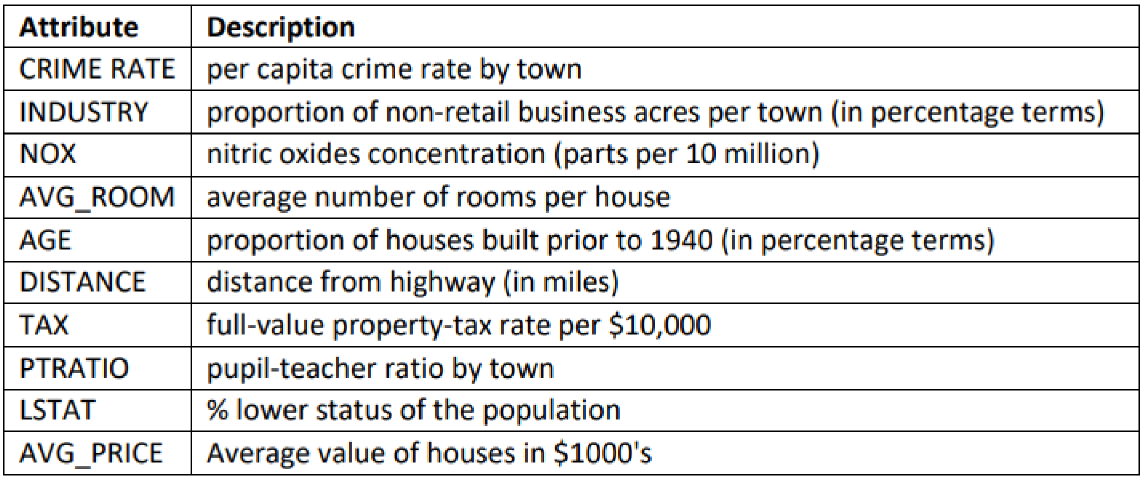
Terro's real estate agency

Problem Statement:

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.

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

Objective:

Your job, as an auditor, is to analyse the magnitude of each variable to which it can affect the price of a house in a particular locality.

Q1) Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation.

CRIME_RATE		*AGE*		*INDUS*	
Mean	4.871976	Mean	68.5749	Mean	11.13678
Standard Error	0.12986	Standard Error	1.25137	Standard Error	0.30498
Median	4.82	Median	77.5	Median	9.69
Mode	3.43	Mode	100	Mode	18.1
Standard Deviation	2.921132	Standard Deviation	28.14886	Standard Deviation	6.860353
Sample Variance	8.53301	Sample Variance	792.3584	Sample Variance	47.06444
Kurtosis	1.189122	Kurtosis	-0.96772	Kurtosis	-1.23354
Skewness	0.021728	Skewness	-0.59896	Skewness	0.295022
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

.

NOX		*DISTANCE*		*TAX*	
Mean	0.554695	Mean	9.549407	Mean	408.2372
Standard Error	0.005151	Standard Error	0.387085	Standard Error	7.492389
Median	0.538	Median	5	Median	330
Mode	0.538	Mode	24	Mode	666
Standard Deviation	0.115878	Standard Deviation	8.707259	Standard Deviation	168.5371
Sample Variance	0.013428	Sample Variance	75.81637	Sample Variance	28404.76
Kurtosis	-0.06467	Kurtosis	-0.86723	Kurtosis	-1.14241
Skewness	0.729308	Skewness	1.004815	Skewness	0.669956
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

PTRATIO		*AVG_ROOM*		*LSTAT*		*AVG_PRICE*	
Mean	18.45553	Mean	6.284634	Mean	12.65306	Mean	22.53281
Standard Error	0.096244	Standard Error	0.031235	Standard Error	0.317459	Standard Error	0.408861
Median	19.05	Median	6.2085	Median	11.36	Median	21.2
Mode	20.2	Mode	5.713	Mode	8.05	Mode	50
Standard Deviation	2.164946	Standard Deviation	0.702617	Standard Deviation	7.141062	Standard Deviation	9.197104
Sample Variance	4.686989	Sample Variance	0.493671	Sample Variance	50.99476	Sample Variance	84.58672
Kurtosis	-0.28509	Kurtosis	1.8915	Kurtosis	0.49324	Kurtosis	1.495197
Skewness	-0.80232	Skewness	0.403612	Skewness	0.90646	Skewness	1.108098
Range	9.4	Range	5	Range	36.24	Range	45
Minimum	12.6	Minimum	4	Minimum	1.73	Minimum	5
Maximum	22	Maximum	9	Maximum	37.97	Maximum	50
Sum	9338.5	Sum	3180.025	Sum	6402.45	Sum	11401.6
Count	506	Count	506	Count	506	Count	506

CRIME_RATE:

Minimum per capita crime rate by town starts from 0.04 to maximum crime rate of 9.99 with an average of 4.18, we can that the data is positively skewed 0.02, the data is negative kurtosis -1.18 indicates a relatively flat distribution of data.

AGE:

The age of houses are starts with a minimum of 2.9 to maximum of 100 with an average age house approximately 68.57, the Data is negatively skewed -0.59. the data is negative kurtosis -0.96 indicates a relatively flat distribution of data.

INDUS:

The proportion of non-retail business acres per town starts from 0.46 and end in 27.74 the average is 11.13, the data is positively skewed 0.29 and it is having a negative kurtosis -1.23 indicates a relatively flat distribution of data.

NOX:

The nitric oxides concentration is starts from 0.38 and ends in 0.87 with an average of 0.55, the data is positively skewed 0.72 and there is a negative kurtosis -0.06 indicates a relatively flat distribution of data.

DISTANCE:

The Distance of houses from the high-way starts from 1 and ends in 24 with an average of 9.54, the data is positively skewed 1.004 and there is a negative kurtosis -0.06 indicates a relatively flat distribution of data, this says maximum number of houses are away from high-way.

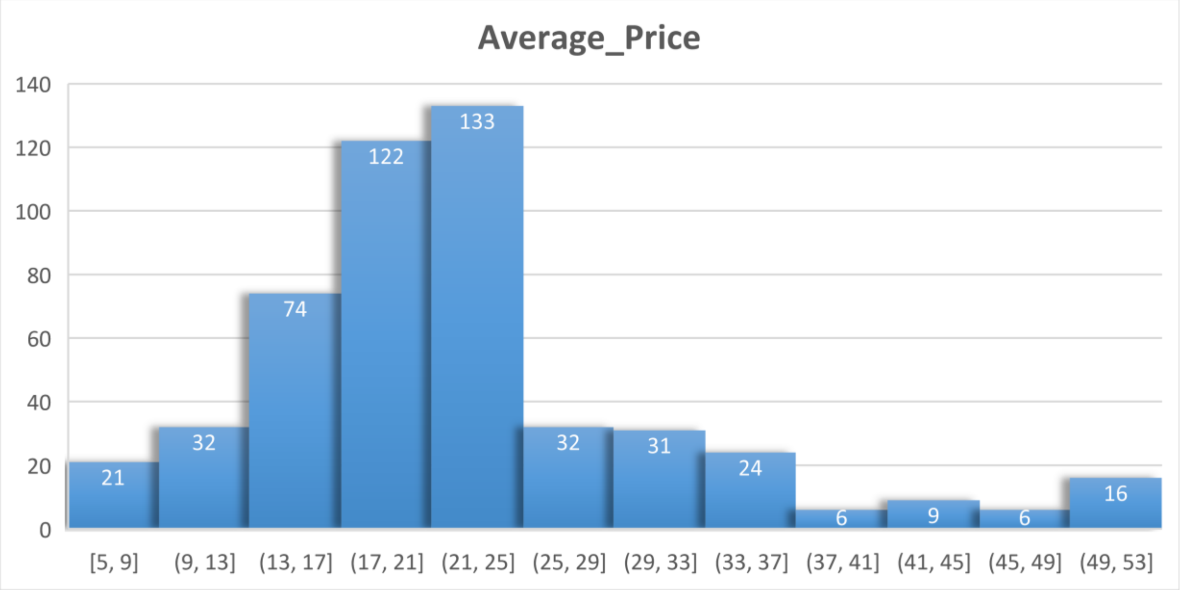
AVG_ROOM:

The average number of rooms per house starts from 5 and ends in 9 with an average of 6, the data is positively skewed 0.40 and there is a positive kurtosis 1.89 indicates a relatively peaked distribution of data.

AVG_PRICE:

The average price of houses starts from 5 and ends in 50 with an average of 22.53, the data is positively skewed 1.10 and there is a positive kurtosis 1.49 indicates a relatively peaked distribution of Data.

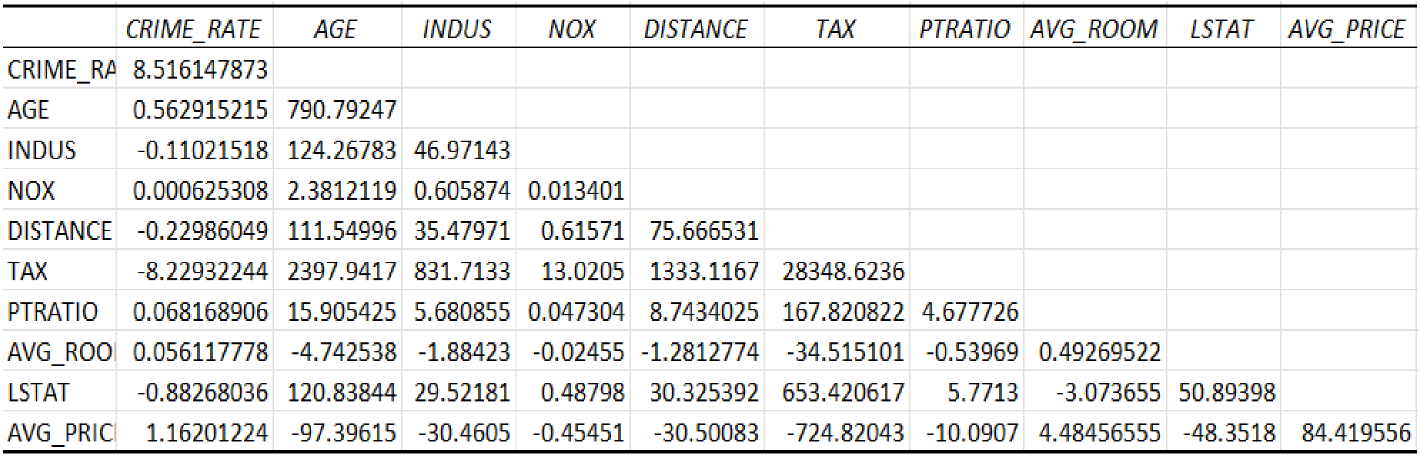
Q2) Plot a histogram of the AVG_PRICE variable. What do you infer?



Inference:

- The houses range from **$5000** to **$50000**.
- most of the houses in this dataset Ranges from **$17000** to **$25000.**
- the least price of the houses ranges from **$37000** to **$41000** and **$45000** to **$49000**.

Q3) Compute the covariance matrix. Share your observations.



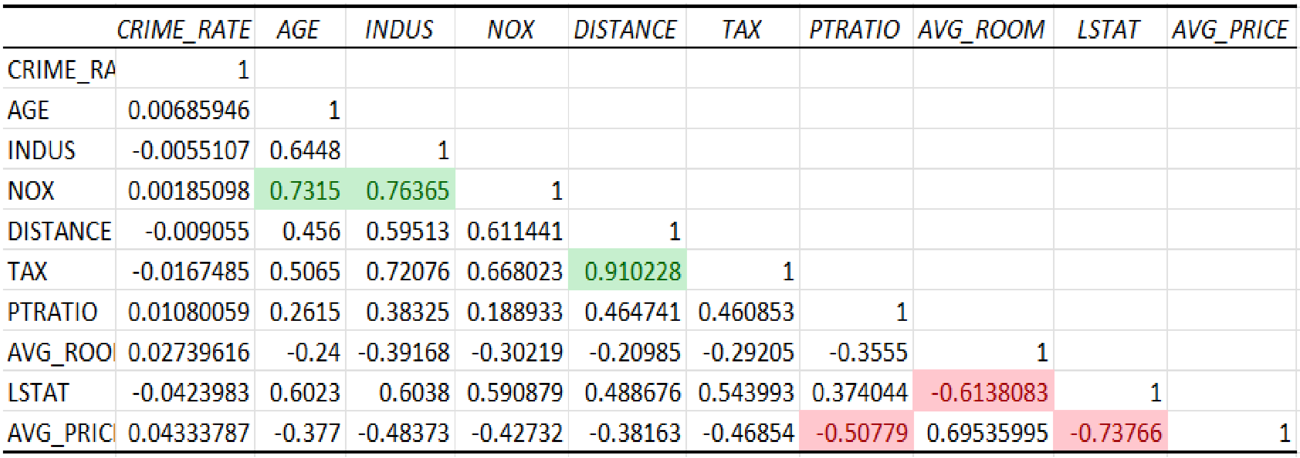
	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RA	8.516147873									
AGE	0.562915215	790.79247								
INDUS	-0.11021518	124.26783	46.97143							
NOX	0.000625308	2.3812119	0.605874	0.013401						
DISTANCE	-0.22986049	111.54996	35.47971	0.61571	75.666531					
TAX	-8.22932244	2397.9417	831.7133	13.0205	1333.1167	28348.6236				
PTRATIO	0.068168906	15.905425	5.680855	0.047304	8.7434025	167.820822	4.677726			
AVG_ROO	0.056117778	-4.742538	-1.88423	-0.02455	-1.2812774	-34.515101	-0.53969	0.49269522		
LSTAT	-0.88268036	120.83844	29.52181	0.48798	30.325392	653.420617	5.7713	-3.073655	50.89398	
AVG_PRIC	1.16201224	-97.39615	-30.4605	-0.45451	-30.50083	-724.82043	-10.0907	4.48456555	-48.3518	84.419556

Observation:

In the above covariance matrix, Positive covariance values that indicate high relationship of two variables, Negative covariance values that indicates low relationship of two variables.

	Positive Variables		Negative Variables	
1)	AGE vs TAX	2397.94	DISTANCE vs AVG_PRICE	-30.50
2)	DISTANCE vs TAX	1333.11	TAX vs AVG_ROOM	-34.52
3)	INDUS vs TAX	831.71	LSTAT vs AVG_PRICE	-48.35
4)	TAX vs LSTAT	653.42	AGE vs AVG_PRICE	-97.40
5)	TAX vs PTRATIO	167.82	TAX vs AVG_PRICE	-724.82

Q4) Create a correlation matrix of all the variables (Use Data analysis tool pack).



	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RA	1									
AGE	0.00685946	1								
INDUS	-0.0055107	0.6448	1							
NOX	0.00185098	0.7315	0.76365	1						
DISTANCE	-0.009055	0.456	0.59513	0.611441	1					
TAX	-0.0167485	0.5065	0.72076	0.668023	0.910228	1				
PTRATIO	0.01080059	0.2615	0.38325	0.188933	0.464741	0.460853	1			
AVG_ROOI	0.02739616	-0.24	-0.39168	-0.30219	-0.20985	-0.29205	-0.3555	1		
LSTAT	-0.0423983	0.6023	0.6038	0.590879	0.488676	0.543993	0.374044	-0.6138083	1	
AVG_PRICI	0.04333787	-0.377	-0.48373	-0.42732	-0.38163	-0.46854	-0.50779	0.69535995	-0.73766	1

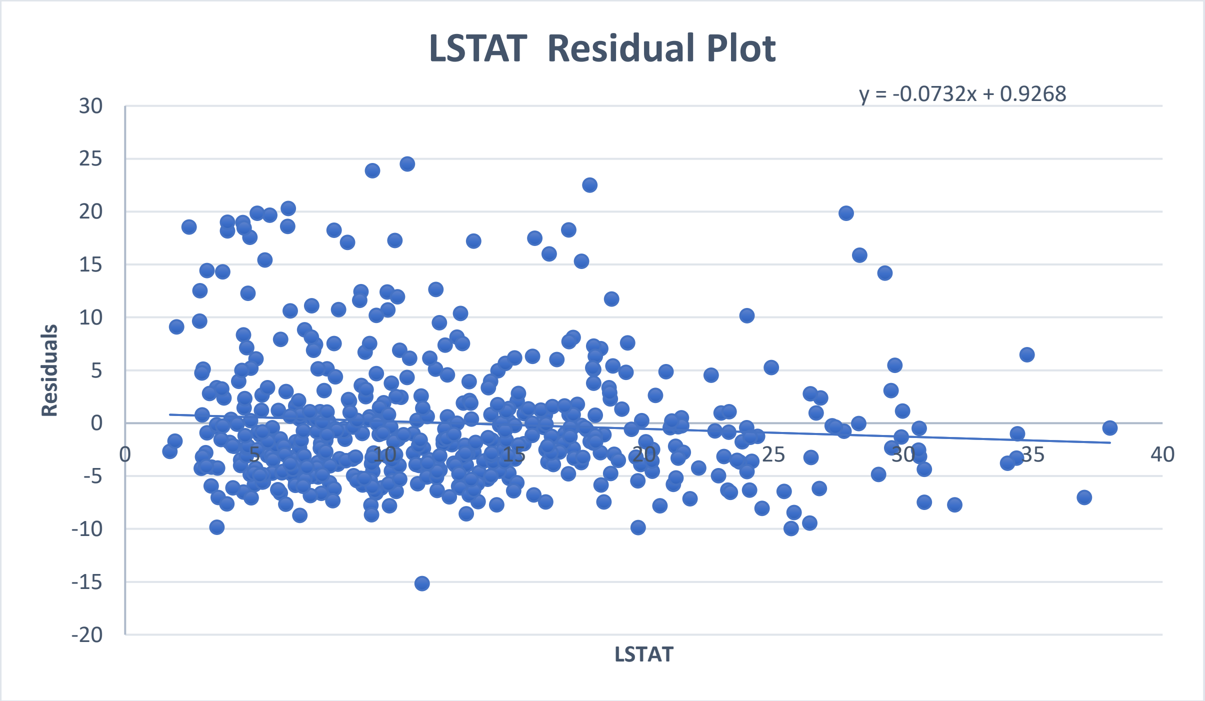
a) **Which are the top 3 positively correlated pairs.**

1) AGE vs NOX: **0.73147**
2) INDUS vs NOX: **0.76365**
3) DISTANCE vs TAX: **0.91023**

b) **Which are the top 3 negatively correlated pairs**.

1) AVG_ROOM vs LSTAT: **-0.61381**
2) PTRATIO vs AVG_PRICE: **-0.50779**
3) LSTAT vs AVG_PRICE: **-0.73766**

Q5) 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?**

inference:

According to this If LSTAT is increases then the AVG_PRICE decreases.

Variance:

The R-squared value 0.544146298 indicates that the approximately 54.41% of the variance in the dependent variable can be explained by the independent variables included in this LSTAT model.

Coefficient Values:

The coefficient of LSTAT for this model this -0.95005. According to this If LSTAT is increases then the AVG_PRICE decreases.

Intercept:

In this LSTAT Model Intercept is 34.55384.

residual:

A plot between the independent variable on the x axis (LSTAT) and residuals on the Y axis, there is a random residual plot.

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

Yes, LSTAT is significant for AVG_PRICE variable for this model**.**
As we can see the P-value is (5.08E-88) this concludes the P-value is below 5% according to this we LSTAT is Significant for the Analysis for this model**.**

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

a) **Write the Regression equation. If a new house in this locality has 7 rooms (on an average) and has a value of 20 for L-STAT, then what will be the value of AVG_PRICE? How does it compare to the company quoting a value of 30000 USD for this locality? Is the company Overcharging/Undercharging?**

Regression Equation:

Price =4.9 * AVG_ROOM-0.65574 * LSTAT

= 4.9 * 7 - 0.65574 * 20

AVG_PRICE = 21.19

the Price of the new house is **$21000**, while the company charging **$30000** so we can say that the company is Overcharging.

b) **Is the performance of this model better than the previous model you built in Question 5? Compare in terms of adjusted R-square and explain.**

Previous Model

Regression Statistics	
Multiple R	0.737663
R Square	0.544146
Adjusted R Square	0.543242
Standard Error	6.21576
Observations	506

Current Model

Regression Statistics	
Multiple R	0.973885
R Square	0.948453
Adjusted R Square	0.946366
Standard Error	5.535767
Observations	506

Comparison:

The previous model's adjusted R-squared value is **0.543242** indicates that **54.32%** of the variance in the dependent variable can be accounted for by the independent variables in the previous model.

while in this model's adjusted R-squared value is **0.94637** showing increase of independent variable approximately **94%**. here after making intercept Zero the adjusted R Square is also increased.

increasing in the number of independent variable penalty is also increase.

Q7) 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.

Adjusted R Square:

The Adjusted R Square Value is 0.688298646855749 which is Approximately 68.83%. The adjusted R Square take place by the number of independent variables.

Coefficient:

	Coefficients	*P-value*
Intercept	29.24132	2.54E-09
CRIME_RATE	0.048725	0.534657
AGE	0.032771	0.01267
INDUS	0.130551	0.039121
NOX	-10.3212	0.008294
DISTANCE	0.261094	0.000138
TAX	-0.0144	0.000251
PTRATIO	-1.07431	6.59E-15
AVG_ROOM	4.125409	3.89E-19
LSTAT	-0.60349	8.91E-27

the positive Coefficients tells us a positive relationship, And Negative Coefficients tells us a negative relationship.

Intercept:

the intercept value is **29.24132.** it represents the value of the dependent variable when all independent variables are set to be zero.

Explain the significance of each independent variable with respect to AVG_PRICE.

CRIME_RATE:

The coefficient of CRIME_RATE is **0.048725..** The positive coefficient represents the positive relationship between CRIME_RATE and AVG_PRICE. But CRIME_PRICE is not Significant variable with respect to P-Value.

AGE:

The Coefficient of AGE is **0.032771..** The Positive Coefficient represents the positive relationship between Age and AVG_PRICE. It is Significance variable for AVG_PRICE with respect to P-value.

INDUS:

The Coefficient of INDUS is **0.130551..** The positive Coefficient represents the positive relationship between Indus and AVG_PRICE. It is Significance variable for AVG_PRICE with respect to P-value.

NOX:

The Coefficient of NOX is **-10.3212..** The negative Coefficient represents the negative relationship between NOX and AVG_PRICE. It is Significance variable for AVG_PRICE with respect to P-value.

DISTANCE:

The Coefficient of DISTANCE is **0.261094..** The positive Coefficient represents the positive relationship between Distance and AVG_PRICE. It is Significance variable for AVG_PRICE with respect to P-value. According to this the most houses are away from highway.

TAX:

The Coefficient of TAX is **-0.0144..** The negative Coefficient represents the negative relationship between Tax and AVG_PRICE. It is Significance variable for AVG_PRICE with respect to P-value.

PTRATIO:

The Coefficient of PTRATION is **-1.07431..** The negative Coefficient represents the negative relationship between PTRATIO and AVG_PRICE. It is Significance variable for AVG_PRICE with respect to P-value.

AVG_ROOM:

The Coefficient of AVG_ROOM is **4.125409..** The positive Coefficient represents the positive relationship between AVG_PRICE and AVG_PRICE. It is Significance variable for AVG_PRICE with respect to P-value. according to this house are more price with more rooms.

LSTAT:

The Coefficient of LSTAT is **-0.60349..** The negative Coefficient represents the negative relationship between LSTAT and AVG_PRICE. It is Significance variable for AVG_PRICE with respect to P-value. According to this If LSTAT is increases then the AVG_PRICE decreases.

Q8) 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.

Multiple R:

Multiple R value is **0.832836..** this represents the multiple correlation between independent and dependent variable. The value representing a strong positive correlation between independent and dependent variable.

R Square:

The R Square value is **0.69361..** Which is approximately 69.36% of the variance in the dependent variable can be explained by the independent variables included in the model.

Adjusted R Square:

The adjusted R square value is **0.68868..** This represent s the R-Square value adjusted for the number of independent variables in this model, the adjusted R-Square value is always less than R-Square Value, increasing in the number of independent variable penalty is also increase.

Standard Error:

the value of standard Error is **5.131..**the differences between the actual and predicted values of the dependent variable.

Observations:

it represents the number of data points uses to produce the regression model, in this model there are **506** observations.

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?

Previous Model

Regression Statistics	
Multiple R	0.83298
R Square	0.69385
Adjusted R Square	0.6883
Standard Error	5.13476
Observations	506

Current Model

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

Comparison:

The previous model's adjusted R-squared value is **0.6882..** indicates that 68.83% of the variance in the dependent variable can be accounted for by the independent variables in the previous model.

while in this model's adjusted R-squared value is **0.68868..**showing decrease of independent variable approximately 68.87%. here in this model after removing CRIME_RATE the adjusted R Square is also increased. *this model is better compared to the previous model.*

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?

	Coefficients
NOX	-10.273
PTRATIO	-1.0717
LSTAT	-0.6052
TAX	-0.0145
AGE	0.03293
INDUS	0.13071
DISTANCE	0.26151
AVG_ROOM	4.12547
Intercept	29.4285

If NOX is more in locality, according to this model AVG_PRICE of houses will decrease.

d) Write the regression equation from this model.

AVG_PRICE = 0.03293 * AGE + 0.13071 * INDUS - 10.2727 * NOX + 0.26151 * DISTANCE - 0.01445 * TAX- 1.0717 * PTRATIO + 4.12547 * AVG_ROOM - 0.6051 * LSTAT + 29.4285