2023

TERRO’S REAL ESTATE AGENCY`

BY

KIRAN AIGALI

Table of Contents

Abstract: 1

1. Generate the summary statistics for each variable in the table. Write down your observation. 2

- Observations 2

2. Plot a histogram of the Average Price variable. What do u infer? 5

3. Compute the covariance matrix, share your observations. 6

4. Create a correlation matrix of all the variables (Use Data Analysis tool pack). 7

a. Which are the top 3 positively correlated pairs? 8

b. Which are the top 3 negatively correlated pairs? 8

5. Build an initial regression module with Avg price as ‘y’ (Dependent variable) and LSTAT variable as independent variable. Generate the residual plot. 8

a. What do you infer from the regression summary output in terms of variance explained, coefficient value, intercept, and residual plot? 8

b. Is LSTST variable significant for the analysis based on your model? 10

6. Build a new Regression model including LSTAT and Avg Room together as independent variable and Avg Price as dependent variable. 10

a. write a regression equation. If a new house in this locality has 7 rooms and has a value of 20 for LSTAT, 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? 10

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. 11

7. Build another Regression model with all variables where Avg Price be the dependent variable and all other variables are dependent. Interpret the output in terms of adjusted R square, coefficient and interpret values. Explain the significance of each independent variable with respect to Avg Price. 11

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: 12

a. Interpret the output of this model. 13

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? 13

c. Sort the values of the coefficient in ascending order. What will happen to the average price if the value of NOx is more in locality in town? 13

d. Write the regression equation from this model. 14

List of Tables

Table 1: Descriptive Analysis -- 2
Table 2: Covariance Matrix -- 6
Table 3: Correlation Matrix -- 7
Table 4: Initial Regression (Avg Price & LSTST)-- 8
Table 5: Regression Module with All Variables -- 11
Table 6: Regression Module Using Significant Variables -- 12
Table 7: Sorting of Coefficient values From the Regression Module -- 13

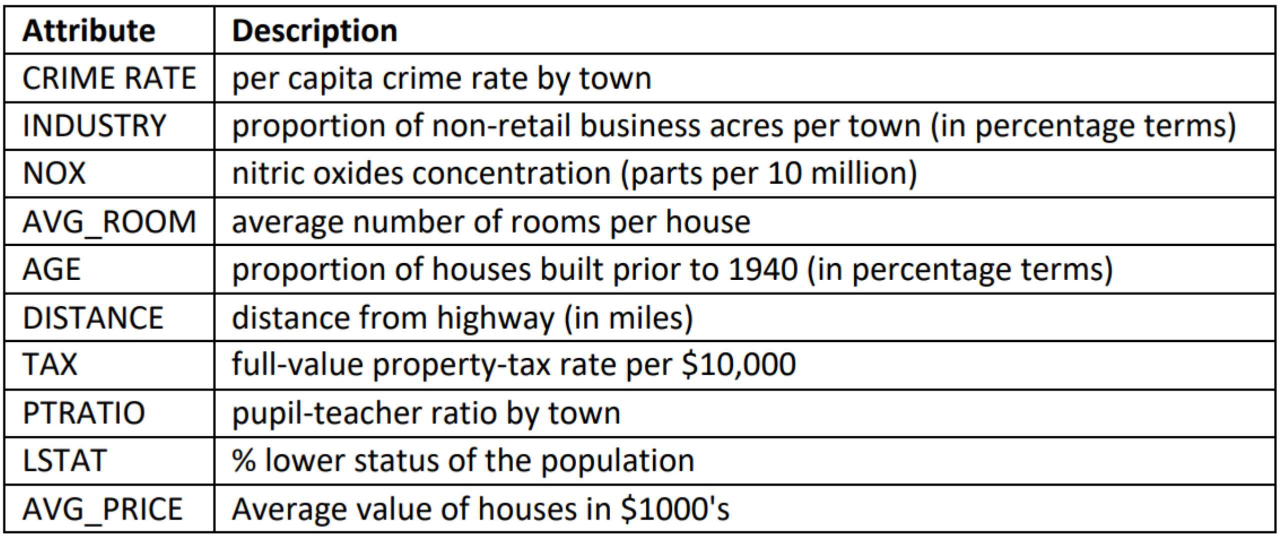
List of Figures

Figure 1: Histogram Plot-- 5
Figure 2: Residual Plot --- 9

Abstract:

“Find 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. Data Set is Given



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

- Generate the summary statistics for each variable in the table. Write down your observation.

Descriptive Statistics

Statistics	*CRIME_RATE*	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
Mean	**4.87**	**68.57**	**11.14**	**0.55**	**9.55**	**408.24**	**18.46**	**6.28**	**12.65**	**22.53**
Standard Error	**0.13**	**1.25**	**0.30**	**0.01**	**0.39**	**7.49**	**0.10**	**0.03**	**0.32**	**0.41**
Median	**4.82**	**77.50**	**9.69**	**0.54**	**5.00**	**330.00**	**19.05**	**6.21**	**11.36**	**21.20**
Mode	**3.43**	**100.00**	**18.10**	**0.54**	**24.00**	**666.00**	**20.20**	**5.71**	**8.05**	**50.00**
Standard Deviation	**2.92**	**28.15**	**6.86**	**0.12**	**8.71**	**168.54**	**2.16**	**0.70**	**7.14**	**9.20**
Sample Variance	**8.53**	**792.36**	**47.06**	**0.01**	**75.82**	**28404.76**	**4.69**	**0.49**	**50.99**	**84.59**
Kurtosis	**-1.19**	**-0.97**	**-1.23**	**-0.06**	**-0.87**	**-1.14**	**-0.29**	**1.89**	**0.49**	**1.50**
Skewness	**0.02**	**-0.60**	**0.30**	**0.73**	**1.00**	**0.67**	**-0.80**	**0.40**	**0.91**	**1.11**
Range	**9.95**	**97.10**	**27.28**	**0.49**	**23.00**	**524.00**	**9.40**	**5.22**	**36.24**	**45.00**
Minimum	**0.04**	**2.90**	**0.46**	**0.39**	**1.00**	**187.00**	**12.60**	**3.56**	**1.73**	**5.00**
Maximum	**9.99**	**100.00**	**27.74**	**0.87**	**24.00**	**711.00**	**22.00**	**8.78**	**37.97**	**50.00**
Sum	**2465.22**	**34698.90**	**5635.21**	**280.68**	**4832.00**	**206568.00**	**9338.50**	**3180.03**	**6402.45**	**11401.60**
Count	**506.00**	**506.00**	**506.00**	**506.00**	**506.00**	**506.00**	**506.00**	**506.00**	**506.00**	**506.00**

Table 1: Descriptive Analysis

- Observations

1. Crime Rate

The mean of the crime rate is 4.87, with standard deviation of 2.92, indicating a wide spread of values around the mean. The value ranges from 0.04 to 9.99. The data seems to be slightly positively skewed. The skewness value is 0.02, indicating that are lower crime instances. The kurtosis is -1.19, indicates a relatively flat distribution.

2. **Age**

   The mean age is 68.57, with a standard deviation of 28.15. The age value ranges from 2.9 to 100 with a slightly negatively skewed. The skewness value is -0.6, indicating that there are more properties with higher age values. The kurtosis value of -0.97 indicates the age distribution has a relatively flat peak.

3. **Indus**

   The mean Indus is about 11.14, with a standard deviation of 6.86. The values range from 0.46 to 27.74, it is slightly positively skewed. The skewness value is 0.73, indicating skewness towards higher values. The distribution has a negative kurtosis of -1.23, indicating a slightly flatter peak compared to normal distribution.

4. **NOx (Nitrogen Oxide Concentration)**

   The mean value is 0.55, with a standard deviation of 0.12. The NOX values range from 0.385 to 0.871, with a positive skew. The skewness value is 0.73, indicating the distribution is slightly towards right. The kurtosis value is -0.06 which is close to zero, indicating that the distribution is relatively normal.

5. **Distance**

The mean distance is about 9.55, with a standard deviation of 8.71. The data ranges from 1 to 24, and it is positively skewed. The Skewness value is 1.00, suggesting that more properties are closer to the highway. The kurtosis value of approximately -0.87 indicates that the distribution is relatively flattered.

6. **Tax**

The mean property tax rate is around 408.24, with a standard deviation of 168.54. The tax rates range from 187 to 711. And it positively skewed. The skewness value is 0.67, indicating that the data is slightly skewed to the right. The kurtosis value of -1.14 indicates that the distribution is relatively flattered.

7. **PT Ratio**

The mean pupil-teacher ratio is approximately 18.46, with a standard deviation of 2.16. The values range from 12.6 to 22, and it is negatively skewed. The skewness value is -0.80, indicating that there are more schools with higher pupil-teacher. The kurtosis value of -0.29.

8. **Average Number of Rooms**

The mean number of rooms is about 6.28, with a standard deviation of 0.70. The values range from 3.56 to 8.78. and it is positively skewed. The skewness value is 0.40. Indicating that there are slightly more houses with a higher number of rooms. The kurtosis value of approximately 1.89 indicates that the distribution of data has heavier tails and a sharper peak compared to a normal distribution.

9. **% Lower Status Population**

The mean LSTAT is approximately 12.65, with a standard deviation of 7.14. The values range from 1.73 to 37.97. It is positively skewed. The skewness value is 0.91, indicating a lower-status population in general. The kurtosis value is 0.49, indicating that the distribution has a relatively moderate peak.

10. **Average House Price**

The mean house price is around 22.53, with a standard deviation of 9.20. The prices range from 5 to 50. The price appears to be positively skewed. The skewness value is 1.11, indicating that there are more houses with lower prices and a few with higher prices. The kurtosis value is approximately 1.50, indicates that the distribution of house prices has a taller and sharper peak.

- Plot a histogram of the Average Price variable. What do u infer?

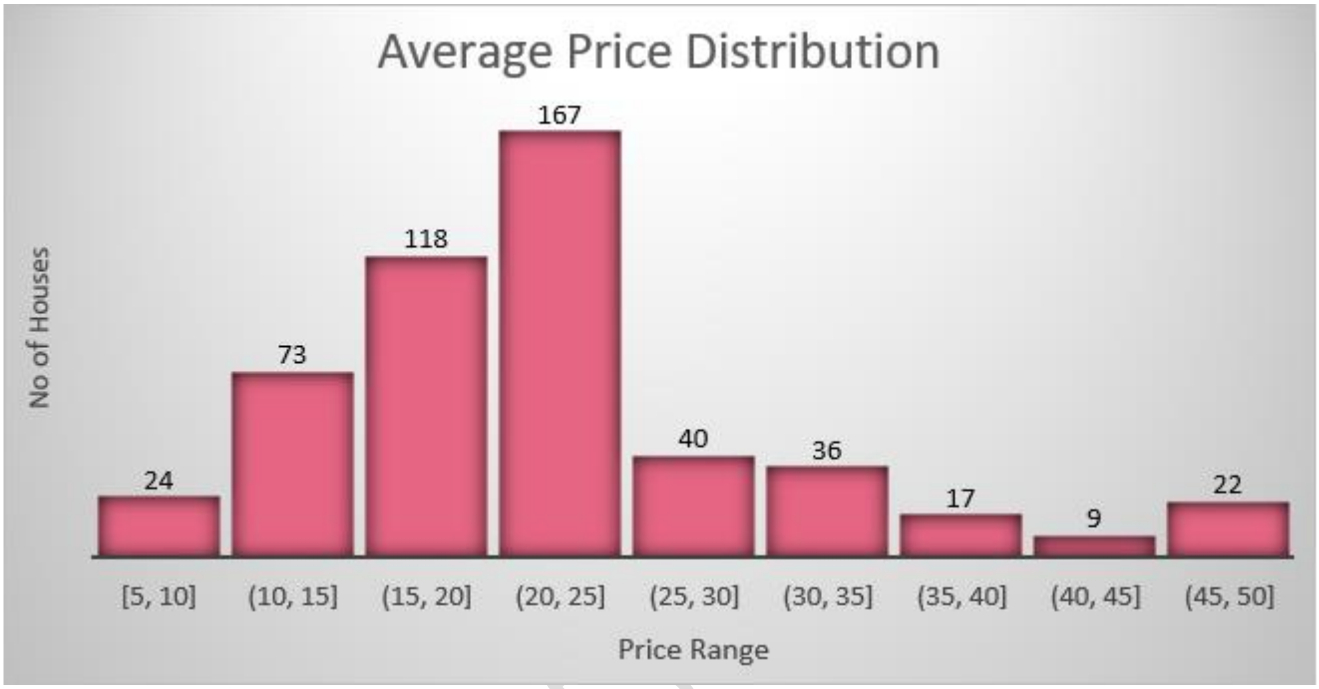


Figure 1: Histogram Plot

By observing the histogram, we were able to observe that horizontal axis line indicates the price range and vertical axis line indicates number of houses.

From the above histogram we were able to observe that,

- Most number of the houses ranges between the price range of $20,000 to $25,000, with 167 houses.
- Less number of houses ranges between the price range of $40,000 to $45,000, with 9 houses.
- By observing the histogram, we were able to conclude that there are more number of houses in lower price range and less number of houses for higher price range.

- Compute the covariance matrix, share your observations.

	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**	**4.677726296**			
AVG_ROOM	**0.056117778**	**-4.74253803**	**-1.884225427**	**-0.024554826**	**-1.281277391**	**-34.51510104**	**-0.539694518**	**0.492695216**		
LSTAT	**-0.882680362**	**120.8384405**	**29.52181125**	**0.487979871**	**30.32539213**	**653.4206174**	**5.771300243**	**-3.073654967**	**50.89397935**	
AVG_PRICE	**1.16201224**	**-97.39615288**	**-30.46050499**	**-0.454512407**	**-30.50083035**	**-724.8204284**	**-10.09067561**	**4.484565552**	**-48.35179219**	**84.41955616**

Table 2: Covariance Matrix

Covariance matrix helps understand whether two variables are directly proportional or inversely proportional. A positive value indicates directly proportional, and the negative value indicates the inversely proportional.

By observing the covariance matrix,

- Age, Indus, NOx, Distance, Tax, PT ratio, LSTAT are inversely proportional indicating that when these values tend to increase the average price decreases.
- Crime rate and Avg price are directly proportional indicating that when these values tend to increase the average price also increases.

- 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_RATE	**1**									
AGE	**0.006859**	**1**								
INDUS	**-0.00551**	**0.644779**	**1**							
NOX	**0.001851**	**0.73147**	**0.763651**	**1**						
DISTANCE	**-0.00906**	**0.456022**	**0.595129**	**0.611441**	**1**					
TAX	**-0.01675**	**0.506456**	**0.72076**	**0.668023**	**0.910228**	**1**				
PTRATIO	**0.010801**	**0.261515**	**0.383248**	**0.188933**	**0.464741**	**0.460853**	**1**			
AVG_ROOM	**0.027396**	**-0.24026**	**-0.39168**	**-0.30219**	**-0.20985**	**-0.29205**	**-0.3555**	**1**		
LSTAT	**-0.0424**	**0.602339**	**0.6038**	**0.590879**	**0.488676**	**0.543993**	**0.374044**	**-0.61381**	**1**	
AVG_PRICE	**0.043338**	**-0.37695**	**-0.48373**	**-0.42732**	**-0.38163**	**-0.46854**	**-0.50779**	**0.69536**	**-0.73766**	**1**

Table 3: Correlation Matrix

a. Which are the top 3 positively correlated pairs?

1. 0.910228 between Distance and Tax

2. 0.763651 between Indus and NOx

3. 0.73147 between Age and NOx

b. Which are the top 3 negatively correlated pairs?

1. – 0.73766 between LSTAT and Avg price.

2. – 0.61381 between Avg Room and LSTAT.

3. – 0.5077 between PT Ratio and Avg price.

- Build an initial regression module 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 residual plot?

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

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

Table 4: Initial Regression (Avg Price & LSTST)

- **Variance**

  The R-Square value is 0.5441, which means that about 54% of the variance in the dependent variable (Average Price) can be explained by the independent variable (LSTAT). The model is effective in predicting the house price based on the % lower status of the population.

- **Co-efficient**

The coefficient value for the independent variable LSTAT is -0.95. This indicates that for each one-unit increase in the % lower status of the population, the average house price decreased by approximately $950. The negative sign indicates an inverse relationship between house price and the % lower status population.

- **Intercept**

The intercept value is 34.55. it represents the predicted average house price when the % lower status population (LSTST) is zero.

Residual Plot

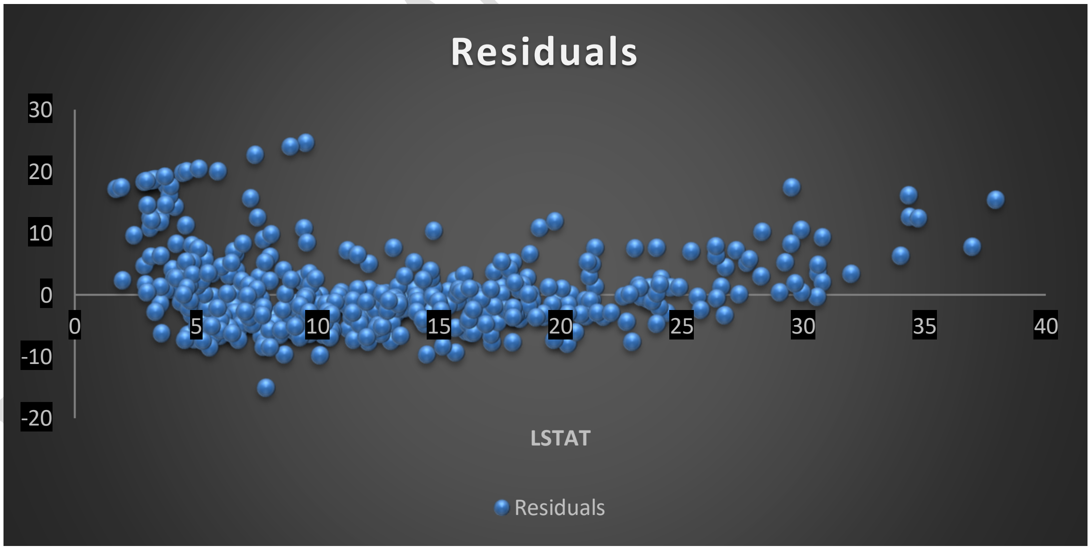


Figure 2: Residual Plot

By observing the residual plot, we were able to observe,

- A good residual plot should be symmetric along the X axis, but the above plot is scattered randomly.
- In LSTAT less than 10 the plots are scattered on the upper side.
- Whereas LSTST 10 to 25 the plots are somewhat evenly distributed on both sides.
- In LSTAT more than 25 have plots scattered on the upper side.

b. Is LSTST variable significant for the analysis based on your model?

- LSTAT value is insignificant, because the adjusted R-value is low.

6. Build a new Regression model including LSTAT and Avg Room together as independent variable and Avg Price as dependent variable.

a. write a regression equation. If a new house in this locality has 7 rooms and has a value of 20 for LSTAT, 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
- Formula, Y= (Avg Room*7) +(LSTAT*20) = Intercept
  = (5.0947*7) +(-0.64236*20) +(-1.3582)
  = 21.45807639 Of Predicted Avg Price

- The company is charging $30,000 which is clearly an overcharge. The predicted average price stands at a reasonable $21,458, significantly lower than the company's asking price. By using this 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.

Adjusted R Square	0.543241826	<	Adjusted R Square	0.637124475

- Adjusted R value gives better results than previous question.

7. Build another Regression model with all variables where Avg Price be the dependent variable and all other variables are dependent. Interpret the output in terms of adjusted R square, coefficient and interpret values. Explain the significance of each independent variable with respect to Avg Price.

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

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.24131526	4.817125596	6.070282926	2.53978E-09	19.77682784	38.70580267	19.77682784	38.70580267
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.534657201	-0.105348544	0.202798827	-0.105348544	0.202798827
AGE	0.032770689	0.013097814	2.501996817	0.012670437	0.00703665	0.058504728	0.00703665	0.058504728
INDUS	0.130551399	0.063117334	2.068392165	0.03912086	0.006541094	0.254561704	0.006541094	0.254561704
NOX	-10.3211828	3.894036256	-2.650510195	0.008293859	-17.97202279	-2.670342809	-17.97202279	-2.670342809
DISTANCE	0.261093575	0.067947067	3.842602576	0.000137546	0.127594012	0.394593138	0.127594012	0.394593138
TAX	-0.01440119	0.003905158	-3.687736063	0.000251247	-0.022073881	-0.0067285	-0.022073881	-0.0067285
PTRATIO	-1.074305348	0.133601722	-8.041104061	6.58642E-15	-1.336800438	-0.811810259	-1.336800438	-0.811810259
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.89287E-19	3.255494742	4.995323561	3.255494742	4.995323561
LSTAT	-0.603486589	0.053081161	-11.36912937	8.91071E-27	-0.70777824	-0.499194938	-0.70777824	-0.499194938

Table 5: Regression Module with All Variables

- The adjusted R square value is 0.68829, The adjusted R square value validates the suitability of this model for prediction task. Its ability to effectively account for the variance in the data indicates that it can be relied upon to make accurate predictions. As a result, this model is a strong candidate for practical use in various prediction scenarios.
- Significant variables are Age, Indus, NOx, Distance, LSTAT, PT Ratio, Vg Room, Tax. P-Values is less than 0.05.
- An insignificant variable is Crime Rate, whose p-Value is greater than 0.05.

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:

	Coefficients	***Standard Error***	***t Stat***	***P-value***	***Lower 95%***	***Upper 95%***	***Lower 95.0%***	***Upper 95.0%***
Intercept	**29.42847**	**4.804729**	**6.124898**	**1.85E-09**	**19.98839**	**38.86856**	**19.98839**	**38.86856**
AGE	**0.032935**	**0.013087**	**2.516606**	**0.012163**	**0.007222**	**0.058648**	**0.007222**	**0.058648**
INDUS	**0.13071**	**0.063078**	**2.072202**	**0.038762**	**0.006778**	**0.254642**	**0.006778**	**0.254642**
NOX	**-10.2727**	**3.890849**	**-2.64022**	**0.008546**	**-17.9172**	**-2.62816**	**-17.9172**	**-2.62816**
DISTANCE	**0.261506**	**0.067902**	**3.851242**	**0.000133**	**0.128096**	**0.394916**	**0.128096**	**0.394916**
TAX	**-0.01445**	**0.003902**	**-3.70395**	**0.000236**	**-0.02212**	**-0.00679**	**-0.02212**	**-0.00679**
PTRATIO	**-1.0717**	**0.133454**	**-8.03053**	**7.08E-15**	**-1.33391**	**-0.8095**	**-1.33391**	**-0.8095**
AVG_ROOM	**4.125469**	**0.442485**	**9.3234**	**3.69E-19**	**3.256096**	**4.994842**	**3.256096**	**4.994842**
LSTAT	**-0.60516**	**0.05298**	**-11.4224**	**5.42E-27**	**-0.70925**	**-0.50107**	**-0.70925**	**-0.50107**

Regression Statistics

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

Table 6: Regression Module Using Significant Variables

a. Interpret the output of this model.

- The R-Value is 68%, so we can use to predictions.

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?

- The adjusted R value is better than the previous model.

Adjusted R Square	0.688684

>

Adjusted R Square	0.688298647

By comparing Adjusted R square value, we can see that this model has slightly higher R square value compared to the previous model. Hence, this model is better than the previous model.

c. Sort the values of the coefficient in ascending order. What will happen to the average price if the value of NOx is more in locality in town?

By sorting the coefficient values in ascending order,

	Coefficients
NOX	-10.3211828
PTRATIO	-1.074305348
LSTAT	-0.603486589
TAX	-0.01440119
AGE	0.032770689
CRIME_RATE	0.048725141
INDUS	0.130551399
DISTANCE	0.261093575
AVG_ROOM	4.125409152
Intercept	29.24131526

Table 7: Sorting of Coefficient values From the Regression Module

The coefficient of NOx is negative which implies that it is inversely proportional. So, if the NOx increases the average price tends to decrease.

d. Write the regression equation from this model.

Multi Linear Regression Equation [Y = (m1x1 + m2x2 + …………) + Intercept]

Where,

Y = Dependent Variable (to be predicted value)

m1, m2 = Slopes

x1, x2 = Independent variables

By substituting the coefficients value provided in the table, the regression equation for this model,

Avg Price = (Coefficient (Age) * Age) + (Coefficient (Indus) * Indus) + (Coefficient (NOx) * NOx) + (Coefficient (Distance) * Distance) + (Coefficient (Tax) * Tax) + (Coefficient (PT ratio) * PT ratio) + (Coefficient (Avg Room) * Avg Room) + (Coefficient (LSTAT) * LSTAT

By substituting the values in the Equation, we can find the Dependent variable.