Terro's Real Estate Agency

Case Study report

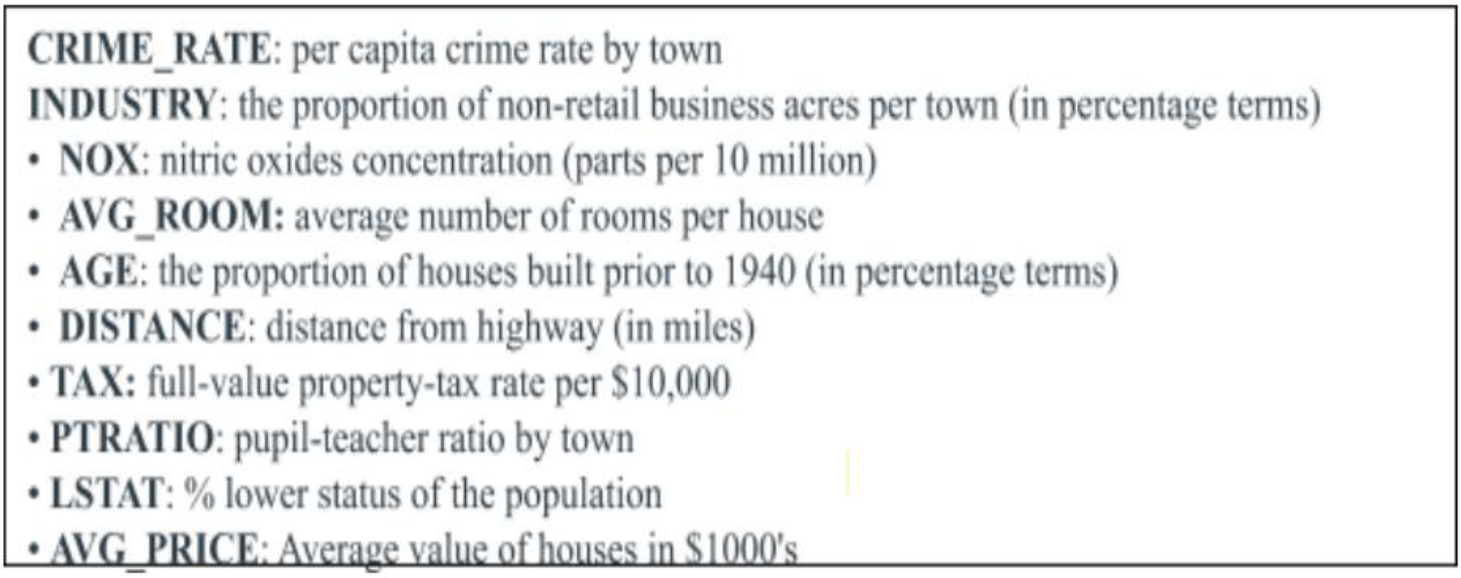
Objective: -

To analyse the extent and magnitude of each variable relative to the value of the house

Submitted by:

PRANAV.S

Data Dictionary:



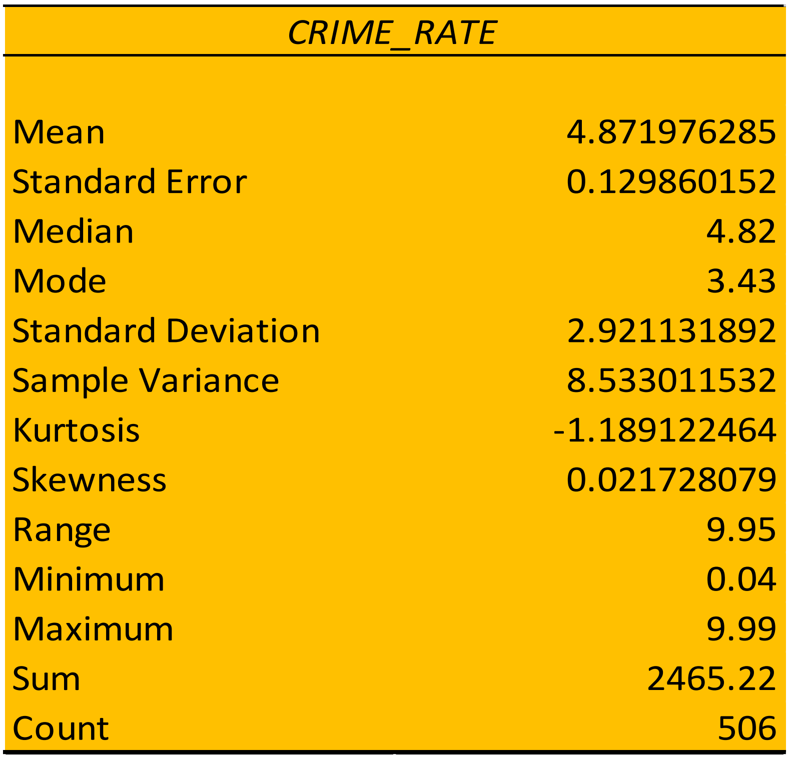
CRIME_RATE: per capita crime rate by town

INDUSTRY: the proportion of non-retail business acres per town (in percentage terms)

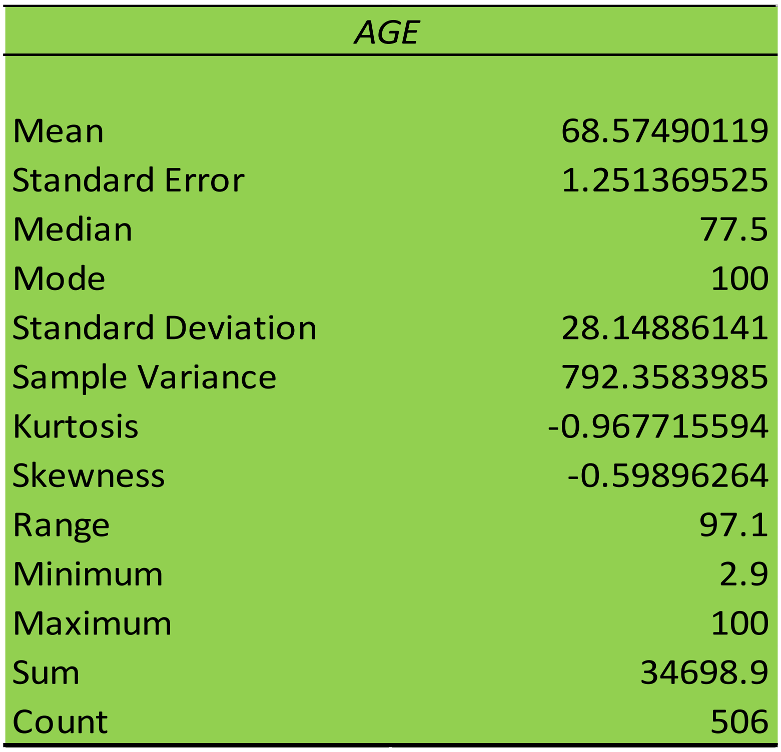
- **NOX**: nitric oxides concentration (parts per 10 million)
- **AVG_ROOM:** average number of rooms per house
- **AGE**: the proportion of houses built prior to 1940 (in percentage terms)
- **DISTANCE**: distance from highway (in miles)
- **TAX:** full-value property-tax rate per $10,000
- **PTRATIO**: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

1. The first step to any project is understanding the data. So, for this step, generate the summary statistics for each of the variables. What do you observe?

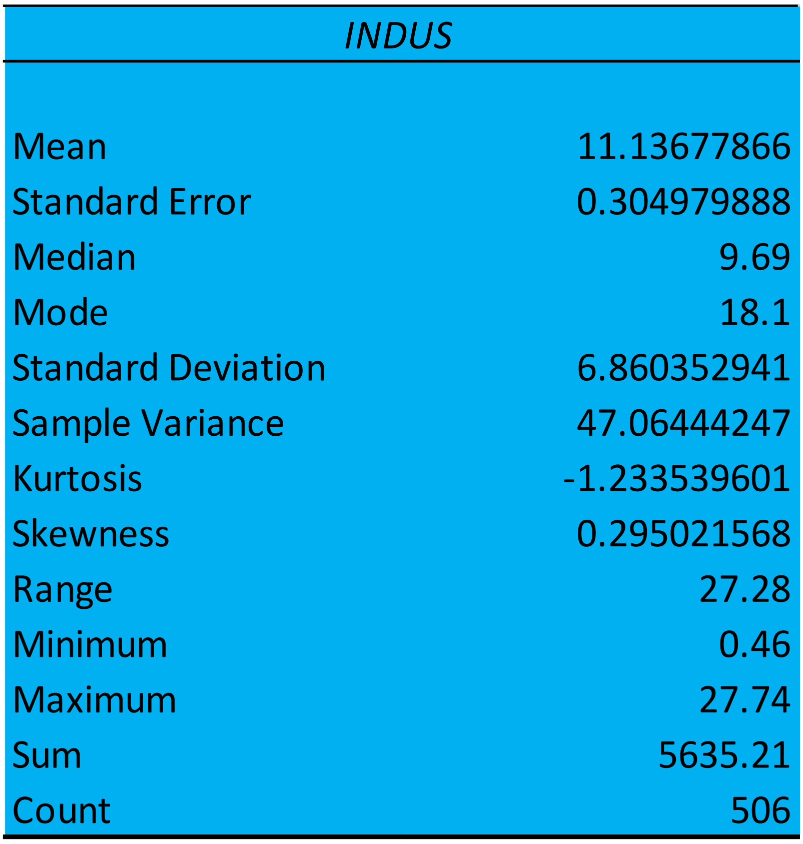
❖ Summary Statistics of the variables.



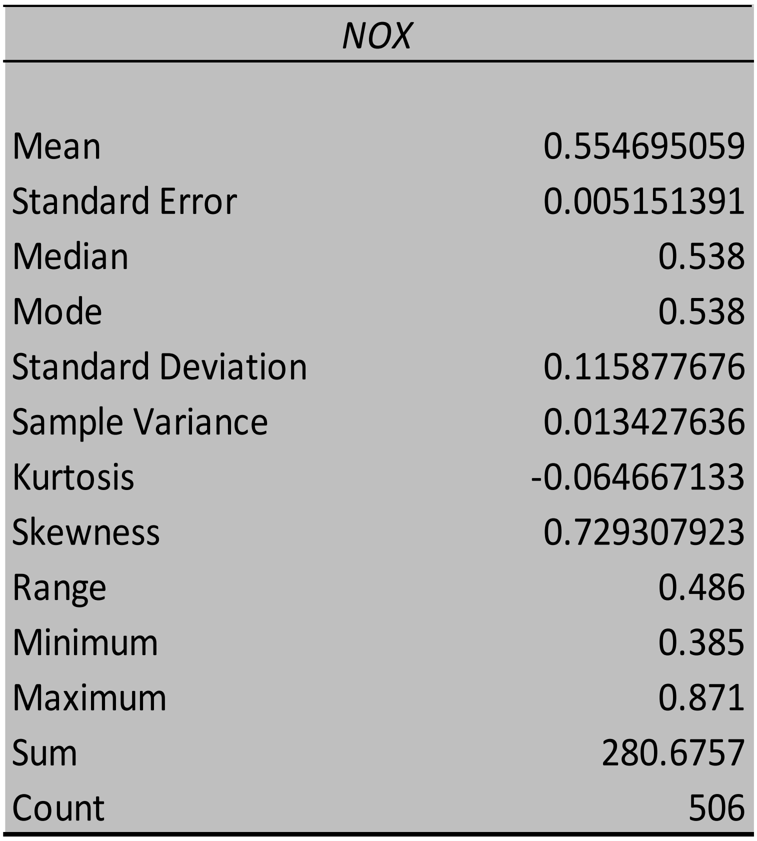
CRIME_RATE	
Mean	4.871976285
Standard Error	0.129860152
Median	4.82
Mode	3.43
Standard Deviation	2.921131892
Sample Variance	8.533011532
Kurtosis	-1.189122464
Skewness	0.021728079
Range	9.95
Minimum	0.04
Maximum	9.99
Sum	2465.22
Count	506



AGE	
Mean	68.57490119
Standard Error	1.251369525
Median	77.5
Mode	100
Standard Deviation	28.14886141
Sample Variance	792.3583985
Kurtosis	-0.967715594
Skewness	-0.59896264
Range	97.1
Minimum	2.9
Maximum	100
Sum	34698.9
Count	506



INDUS	
Mean	11.13677866
Standard Error	0.304979888
Median	9.69
Mode	18.1
Standard Deviation	6.860352941
Sample Variance	47.06444247
Kurtosis	-1.233539601
Skewness	0.295021568
Range	27.28
Minimum	0.46
Maximum	27.74
Sum	5635.21
Count	506



NOX	
Mean	0.554695059
Standard Error	0.005151391
Median	0.538
Mode	0.538
Standard Deviation	0.115877676
Sample Variance	0.013427636
Kurtosis	-0.064667133
Skewness	0.729307923
Range	0.486
Minimum	0.385
Maximum	0.871
Sum	280.6757
Count	506

DISTANCE	
Mean	9.549407115
Standard Error	0.387084894
Median	5
Mode	24
Standard Deviation	8.707259384
Sample Variance	75.81636598
Kurtosis	-0.867231994
Skewness	1.004814648
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

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

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

❖Observation from Summary Statistics of the variables

1. Mean Shows the arithmetic mean of the sample data
It's the sum of all observations divided by the number of observations.

- 4.871976285 is the mean or average per capita CRIME_RATE by town that's on average there is 4.871 crime rate in town

- 68.57490119 is the mean or average age that's the AVERAGE_AGE of people in this data .

- 11.13677866 is the mean or average INDUS that's on average the proportion of non-retail business acres per town.

- 0.554695059 is the mean or average NOX that's on average nitric oxide concentration.

- 9.549407115 is the mean or average distance that's the average distance from the highway

- 408.2371542 is the mean or average TAX that's the average full value property tax rate

- 18.4555336 is the mean or average PTRATIO that's the average pupil teacher ratio by town

- 6.284634387 is the mean or average AVG_ROOM that's the average no of rooms per house

- 12.65306324 is the mean or average LSTAT that's the average percent of lower status of the population

- 22.53280632 is the mean or average price that's the average value of the house.

2. Standard error

The "Standard Error" (SE) indicates how close the sample mean is from the "true" population mean. The standard error is calculated by dividing the standard deviation of the population (or the sample) by the square root of the total number of observations. The SE can be used to roughly define a range of certainty for the mean.

It tells you how much the sample mean would vary if you were to repeat a study using new samples from within a single population.

- Standard error for CRIME_RATE is 0.129860152. This value indicates that the sample we chose has a fairly less distribution of the population means

- Standard error for AGE is 1.251369525. This value indicates that the sample we chose has a less distribution of the population means

- Standard error for INDUS is 0.304979888 This value indicates that the sample we chose has a fairly less distribution of the population means

- Standard error for NOX is 0.005151391. This value indicates that the sample we chose has a very less distribution of the population means

- Standard error for DISTANCE is 0.387084894. This value indicates that the sample we chose has a fairly less distribution of the population means

- Standard error for TAX is 7.492388692 .This value indicates that the sample we chose has a fairly high distribution of the population means

- Standard error for PTRATIO is 0.096243568. This value indicates that the sample we chose has a less distribution of the population means

- Standard error for AVG_ROOM is 0.031235142. This value indicates that the sample we chose has a very less distribution of the population means

- Standard error for LSTAT is 0.317458906 .This value indicates that the sample we chose has a fairly less distribution of the population means

- Standard error for AVG_PRICE is 0.408861147. This value indicates that the sample we chose has a fairly less distribution of the population means

The highest standard error is of the TAX variable 7.49238869 and the least standard error is of NOX variable 0.005151391
The smaller the standard error, the less the spread and the more likely it is that any sample mean is close to the population mean. A small standard error is thus a Good Thing.

3. Median

The median is another measure of central tendency. To get the median you have to order the data from lowest to highest. The median is the number in the middle. If the number of cases is odd the median is the single value, for an even number of cases the median is the average of the two numbers in the middle.
The excel function is:
=MEDIAN(range of cells with the values of interest)

Median for CRIME_RATE variable is 4.82.

Median for AGE variable is 77.5

Median for INDUS variable is 9.69

Median for NOX variable is 0.538

Median for DISTANCE variable is 5.

Median for TAX variable is 330

Median for PTRATIO variable is 19.05

Median for AVG_ROOM variable is 6.2085

Median for LSTAT variable is 11.36

Median for AVG_PRICE variable is 21.2

This value indicates that the middle numbers of the sample we use

4. Mode

The mode refers to the most frequent, repeated or common number in the data. The excel function is:

=MODE (range of cells with the values of interest)

Mode for CRIME_RATE variable is 3.43.

Mode for AGE variable is 100

Mode for INDUS variable is 18.1

Mode for NOX variable is 0.538

Mode for DISTANCE variable is 24

Mode for TAX variable is 666

Mode for PTRATIO variable is 20.2

Mode for AVG_ROOM variable is 5.713

Mode for LSTAT variable is 8.05

Mode for AVG_PRICE variable is 50

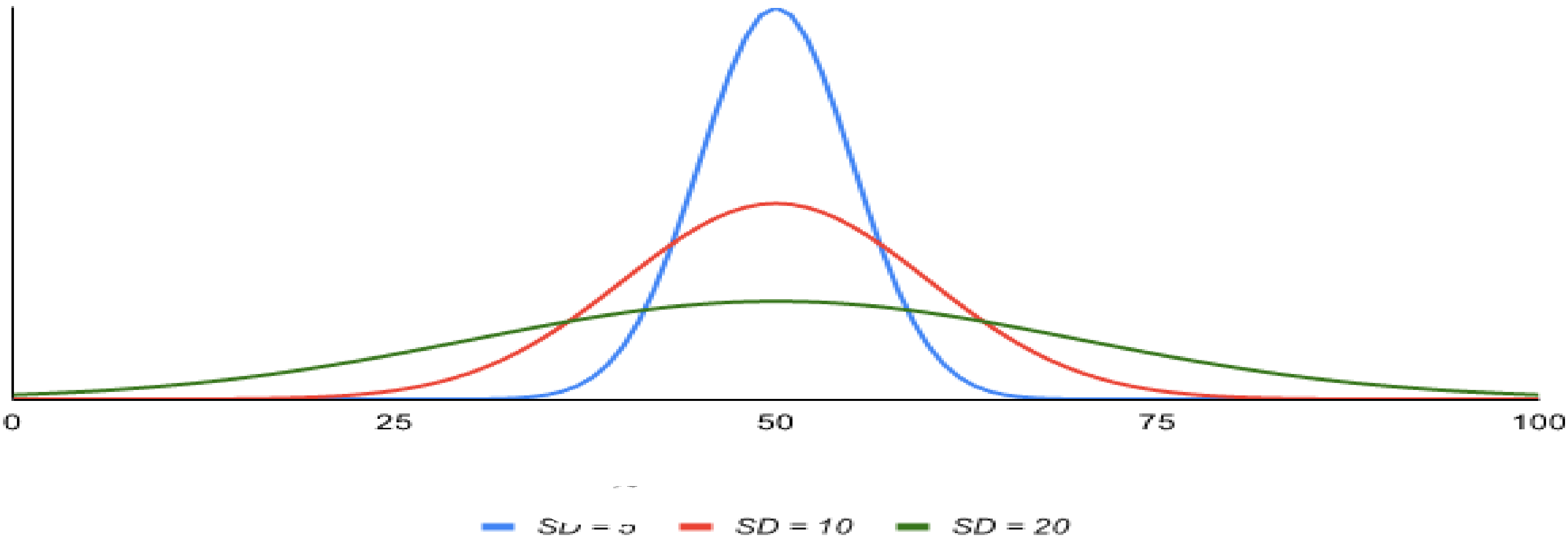
This value shows that the most repeated value on the sample we have of 506.

5. Standard deviation

The standard deviation is the squared root of the variance. Indicates how close the data is to the mean. Assuming a normal distribution, 68% of the values are within 1 sd from the mean, 95% within 2 sd and 99% within 3 sd. The excel formula is:

Standard deviation is a number used to tell how measurements for a group are spread out from the average (mean or expected value). A low standard deviation means that most of the numbers are close to the average, while a high standard deviation means that the numbers are more spread out.

=STDEV(range of cells with the values of interest)



- Standard deviation for CRIME_RATE variable is 2.921131892. This value indicates that the sample values that we use are not far enough from the mean value

- Standard deviation for AGE variable is 28.14886141. This value indicates that the sample values that we use are far from the mean value

- Standard deviation for INDUS variable is 6.860352941. This value indicates that the sample values that we use are not that far enough from the mean value

- Standard deviation for NOX variable is 0.115877676. This value indicates that the sample values that we use are very near to the the mean value

- Standard deviation for DISTANCE variable is 8.707259384. This value indicates that the sample values that we use are not far enough from the mean value

- Standard deviation for TAX variable is 168.5371161. This value indicates that the sample values that we use are very far enough from the mean value

- Standard deviation for PTRATIO variable is 2.164945524. This value indicates that the sample values that we use are near enough from the mean value

- Standard deviation for AVG_ROOM variable is 0.702617143. This value indicates that the sample values that we use are very near enough from the mean value

- Standard deviation for LSTAT variable is 7.141061511. This value indicates that the sample values that we use are not far enough from the mean value

- Standard deviation for AVG_PRICE variable is 9.197104087 . This value indicates that the sample values that we use are bit far enough from the mean value

The highest standard deviation is 168.5371161 of TAX variable this is very far away from the mean and 0.115877676 is the least standard deviation of variable NOX which is very near to the mean

6. sample variance

measures the dispersion of the data from the mean. It is the simple mean of the squared distance from the mean. It is calculated by:

SV = sum of (X-mean of X)^2 / Number of observations minus 1

Higher variance means more dispersion from the mean. The excel function is:
=VAR(range of cells with the values of interest)

- Sample variance for CRIME_RATE variable is 8.533011532. This value indicates that the sample values that we use are not very far enough from the mean value

- Standard variance for AGE variable is 792.3583985. This value indicates that the sample values that we use are far from the mean value

- Standard variance for INDUS variable is 47.06444247. This value indicates that the sample values that we use are not that far enough from the mean value

- Standard variance for NOX variable is 0.013427636. This value indicates that the sample values that we use are very near to the mean value

- Standard variance n for DISTANCE variable is 75.81636598. This value indicates that the sample values that we use are not far enough from the mean value

- Standard variance for TAX variable is 28404.75949 . This value indicates that the sample values that we use are very far enough from the mean value

- Standard variance for PTRATIO variable is 4.686989121. This value indicates that the sample values that we use are near enough from the mean value

- Standard variance for AVG_ROOM variable 0.49367085. This value indicates that the sample values that we use are very near enough from the mean value

- Standard variance for LSTAT variable is 50.99475951. This value indicates that the sample values that we use are not far enough from the mean value

- Standard variance for AVG_PRICE variable is 84.58672359
. This value indicates that the sample values that we use are bit far enough from the mean value

The highest standard variance is 28404.75949of TAX variable this is very far away from the mean and 0.013427636 is the least standard variance of variable NOX which is very near to the mean

7. Kurtosis

Kurtosis measures the peak of a distribution High kurtosis may suggest the presence of outlier’s kurtosis focuses more on the tails for the distribution than the peak, The excel function for kurtosis is:

=KURT (range of cells with the values of interest)

Different types of kurtosis

#1 – Mesokurtic

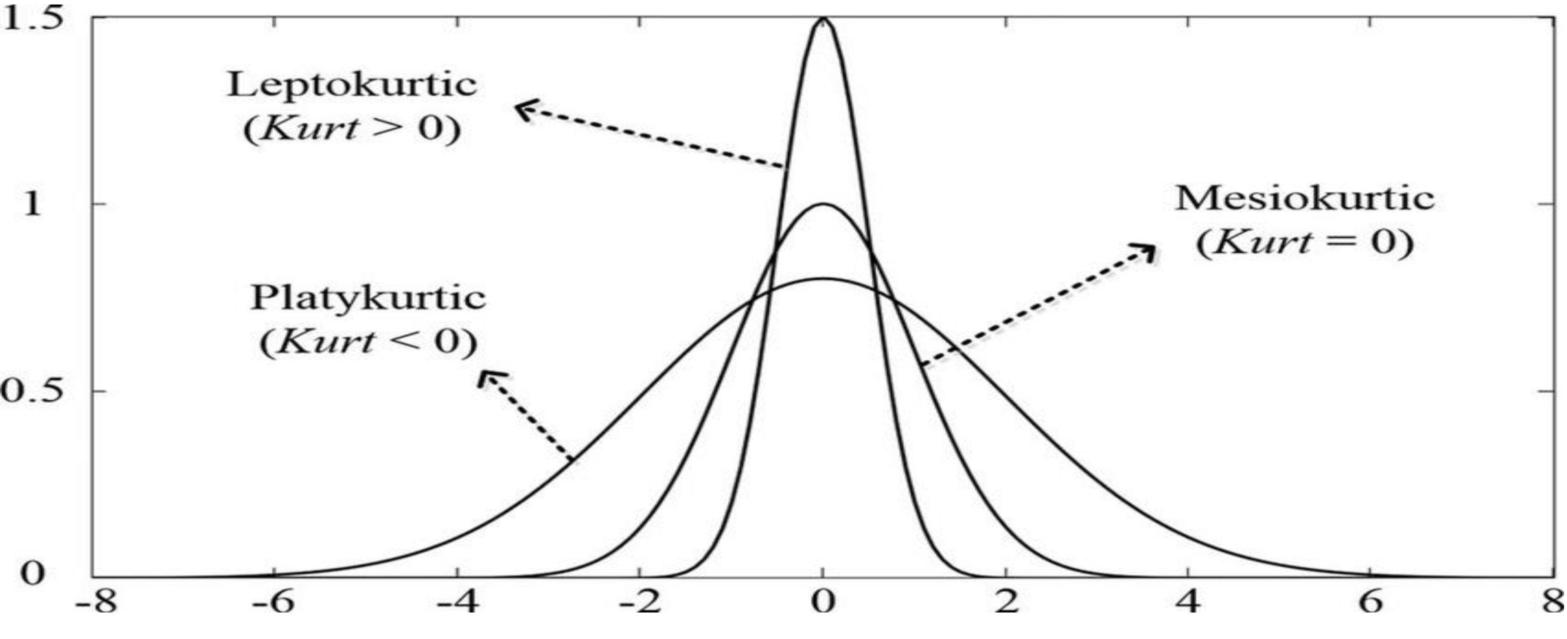
If the kurtosis of data falls close to zero or equal to zero, it is referred to as Mesokurtic. This means that the data set follows a normal distribution. such a pattern depicts risk at a moderate level.

#2 – Leptokurtic

When kurtosis is positive on in other terms, more than zero, the data falls under leptokurtic. Leptokurtic has heavy steep curves on both sides, indicating the heavy population of outliers in the data set leptokurtic distribution is said to be a risky investment

#3 – Platykurtic

Whenever the kurtosis is less than zero or negative, it refers to Platykurtic. The distribution set follows the subtle or pale curve, and that curve indicates the small number of outliers in a distribution. Also, the small outliers and flat tail indicate the less risk involved.



- Kurtosis for CRIME_RATE variable is -**1.189122464**. Because the value of kurtosis is less than zero or negative, it refers to Platykurtic, we can conclude that the sample used is platykurtic distribution (tends to be flat).

- Kurtosis for AGE variable is -0.967715594. Because the value of kurtosis is less than zero or negative, it refers to Platykurtic, we can conclude that the sample used is platykurtic distribution (tends to be flat).

- Kurtosis for INDUS variable is -1.233539601. Because the value of kurtosis is less than zero or negative, it refers to Platykurtic, we can conclude that the sample used is platykurtic distribution (tends to be flat).

- Kurtosis for NOX variable is 0.064667133. Because the value of kurtosis is more than zero or positive, it refers to leptokurtic, we can conclude that the sample used is leptokurtic distribution (tends to be steep).

- Kurtosis for DISTANCE variable is -0.867231994. Because the value of kurtosis is less than zero or negative, it refers to Platykurtic, we can conclude that the sample used is platykurtic distribution (tends to be flat).

- Kurtosis for TAX variable is -1.142407992 . Because the value of kurtosis is less than zero or negative, it refers to Platykurtic, we can conclude that the sample used is platykurtic distribution (tends to be flat).

- Kurtosis for PTRATIO variable is -0.285091383 . Because the value of kurtosis is less than zero or negative, it refers to Platykurtic, we can conclude that the sample used is platykurtic distribution (tends to be flat).

- Kurtosis for AVG_ROOM variable is 1.891500366. Because the value of kurtosis is more than zero or positive, it refers to leptokurtic, we can conclude that the sample used is leptokurtic distribution (tends to be steep).

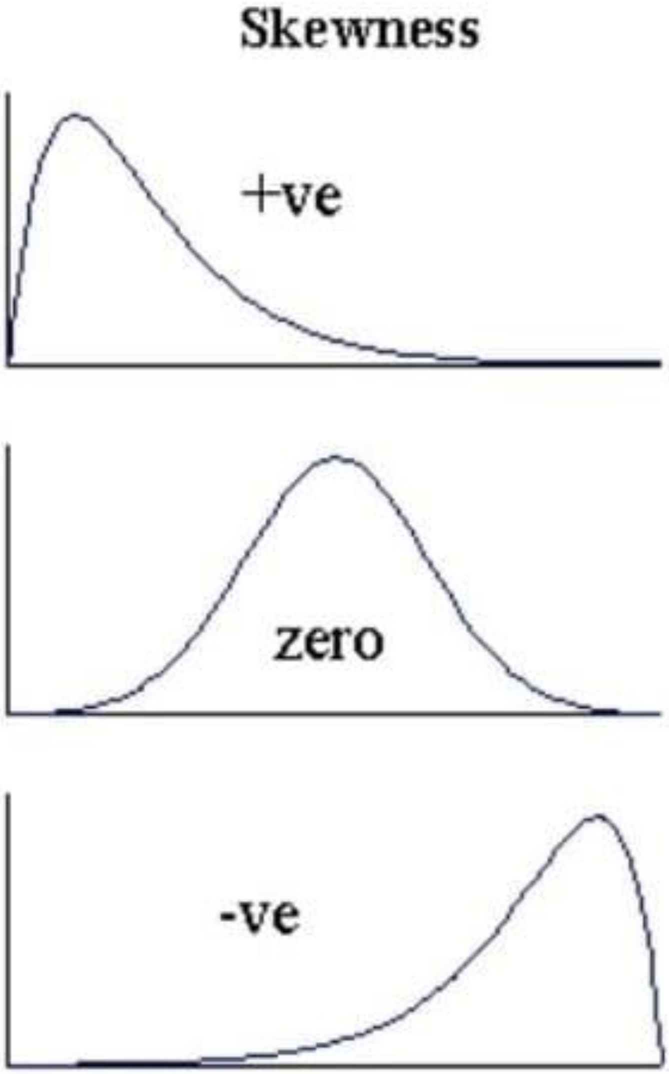
- Kurtosis for LSTAT variable is 0.493239517. Because the value of kurtosis is more than zero or positive, it refers to leptokurtic, we can conclude that the sample used is leptokurtic distribution (tends to be steep).

- Kurtosis for AVG_PRICE variable is 1.495196944. Because the value of kurtosis is more than zero or positive, it refers to leptokurtic, we can conclude that the sample used is leptokurtic distribution (tends to be steep).

8. Skewness

Skewness measures the asymmetry of the data, when in an otherwise normal curve one of the tails is longer than the other. It is a roughly test for normality in the data (by dividing it by the SE). If it is positive there is more data on the left side of the curve (right skewed, the median and the mode are lower than the mean). A negative value indicates that the mass of the data is concentrated on the right of the curve (left tail is longer, left skewed, the median and the mode are higher than the mean). A normal distribution has a skew of 0. Skewness can also be estimated with the following function:

=SKEW (range of cells with the values of interest)



- Skewness for CRIME_RATE variable is **0.021728079**. Because the skewness value is grater than zero, we can conclude that the data tends to be right inclined or positively skewed.

- Skewness for AGE variable is -0.59896264. Because the skewness value is less than zero, we can conclude that the data tends to be left inclined or negative skewed.

- Skewness for INDUS variable is 0.295021568 Because the skewness value is grater than zero, we can conclude that the data tends to be right inclined or positively skewed.

- Skewness for NOX variable is 0.729307923. Because the skewness value is grater than zero, we can conclude that the data tends to be right inclined or positively skewed.

- Skewness for DISTANCE variable is 1.004814648. Because the skewness value is grater than zero, we can conclude that the data tends to be right inclined or positively skewed.

- Skewness for TAX variable is 0.669955942. Because the skewness value is grater than zero, we can conclude that the data tends to be right inclined or positively skewed.

- Skewness for PTRATIO variable is -0.802324927 Because the skewness value is less than zero, we can conclude that the data tends to be left inclined or negative skewed.

- Skewness for AVG_ROOM variable is 0.403612133 Because the skewness value is grater than zero, we can conclude that the data tends to be right inclined or positively skewed.

- Skewness for LSTAT variable is 0.906460094. Because the skewness value is grater than zero, we can conclude that the data tends to be right inclined or positively skewed.

- Skewness for AVG_PRICE variable is 1.108098408. Because the skewness value is grater than zero, we can conclude that the data tends to be right inclined or positively skewed.

9. Range

Range is a measure of dispersion. It is simple the difference between the largest and smallest value, "max"--"min".

These value indicates that the difference between the highest number and the lowest number.

Range for CRIME_RATE variable is 9.95.

Range for AGE variable is 97.1

Range for INDUS variable is 27.28

Range for NOX variable is 0.486

Range for DISTANCE variable is 23

Range for TAX variable is 524

Range for PTRATIO variable is 9.4

Range for AVG_ROOM variable is 5.219

Range for LSTAT variable is 36.24

Range for AVG_PRICE variable is 45

10. Minimum

This shows the lowest value of the variable.

11. Maximum

This shows the highest value of the variable .

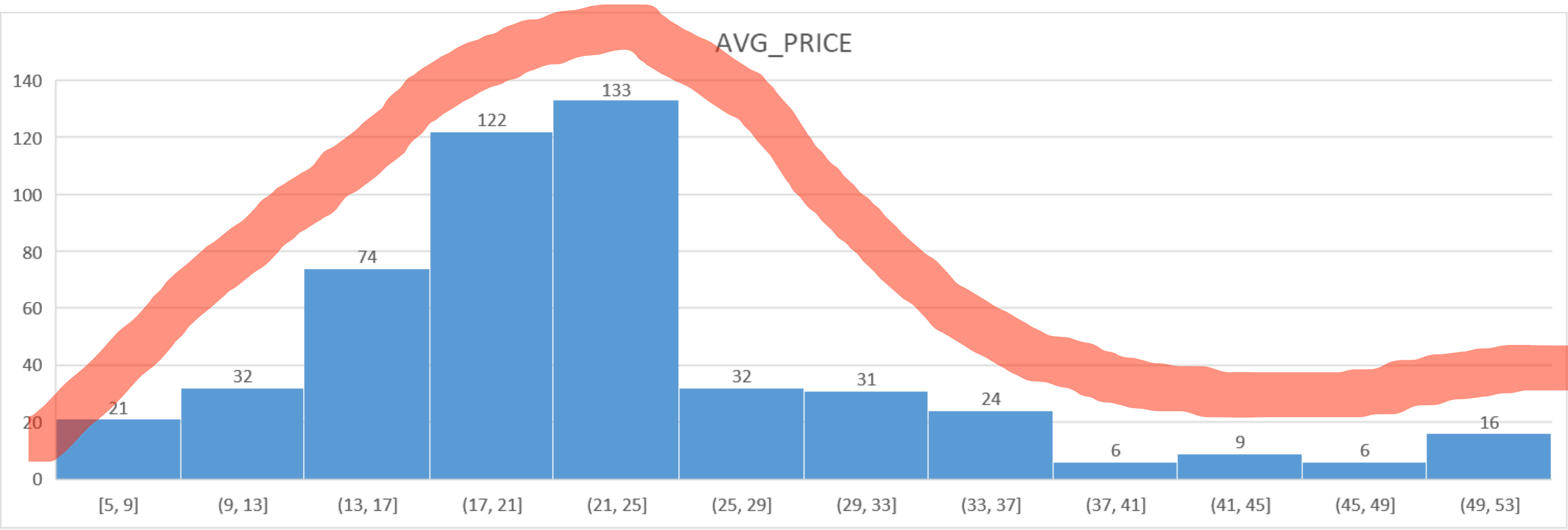
12. Sum

This shows the sum of all value of that particular variable

13. Count

This value indicates the amount of data used
506 Is the count of data used.

2. Plot the histogram of the Average Price Variable. What do you infer?



❖ Histogram observation

- ✓ X-axis: The X-axis are intervals that show the scale of values which the measurements fall under.

- ✓ Y-axis: The Y-axis shows the number of times that the values occurred within the intervals set by the X-axis.

- ✓ The bars: The height of the bar shows the number of times that the values occurred within the interval, while the width of the bar shows the interval that is covered. For a histogram with equal bins, the width should be the same across all bars.

- There is total 12 classes in the histogram of avenge price and the class width is 4(5,9) and the lower limit is 5 and upper limit is 9 that is (5,9)

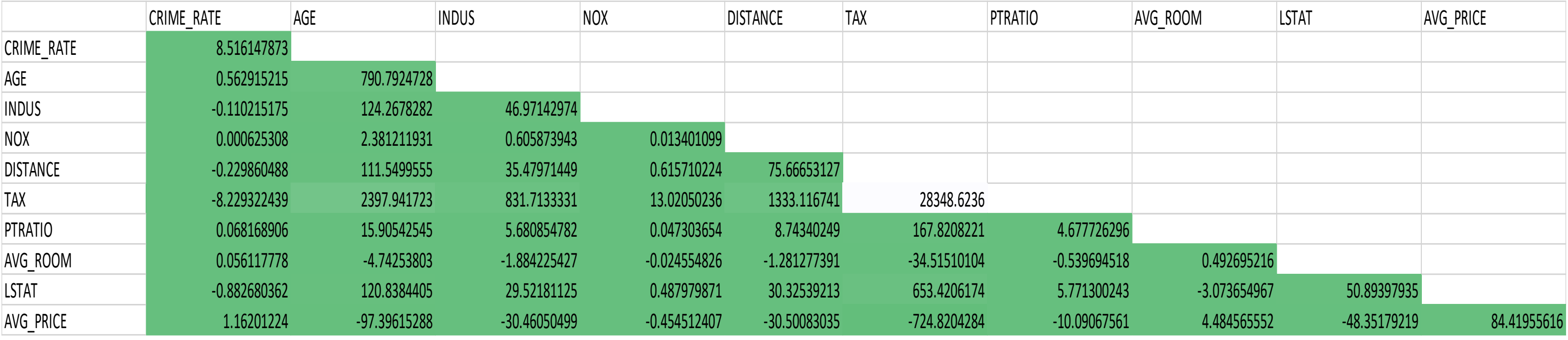
- From 5-25 range there is a constant increases in the average price in a increasing order that no of values that lie in between 5-25 that is left half of the distribution are 382 values and from 25 -29 there is a sudden decrease in the average price from 29-49 there is constant decrease in average price the no of values that lie in between 25-53 are 125 values

- As we can absorb from the above histogram of average price most no of distribution lie in the right side of the histogram, so we can say the data is right-skewed or positively skewed, then the mean is typically GREATER THAN the median

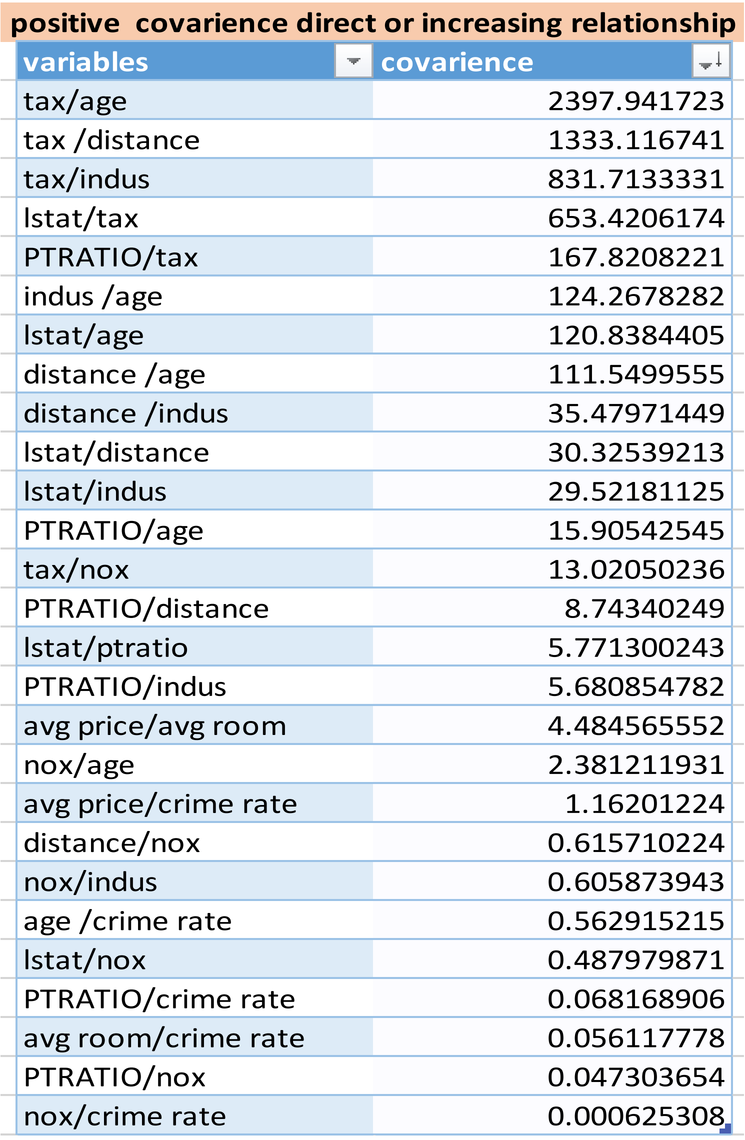
- From the above histogram of average price we can observe the (49,53) is a outlier frequency there are 16 Outliers in average price and these outliers can be described as extremely low or high values that do not fall near any other data points. Sometimes outliers represent unusual cases. Other times they represent data entry errors, or perhaps data that does not belong with the other data of interest. outliers can easily be identified using a histogram

3. Compute the covariance matrix. Share your observations.

covariance matrix

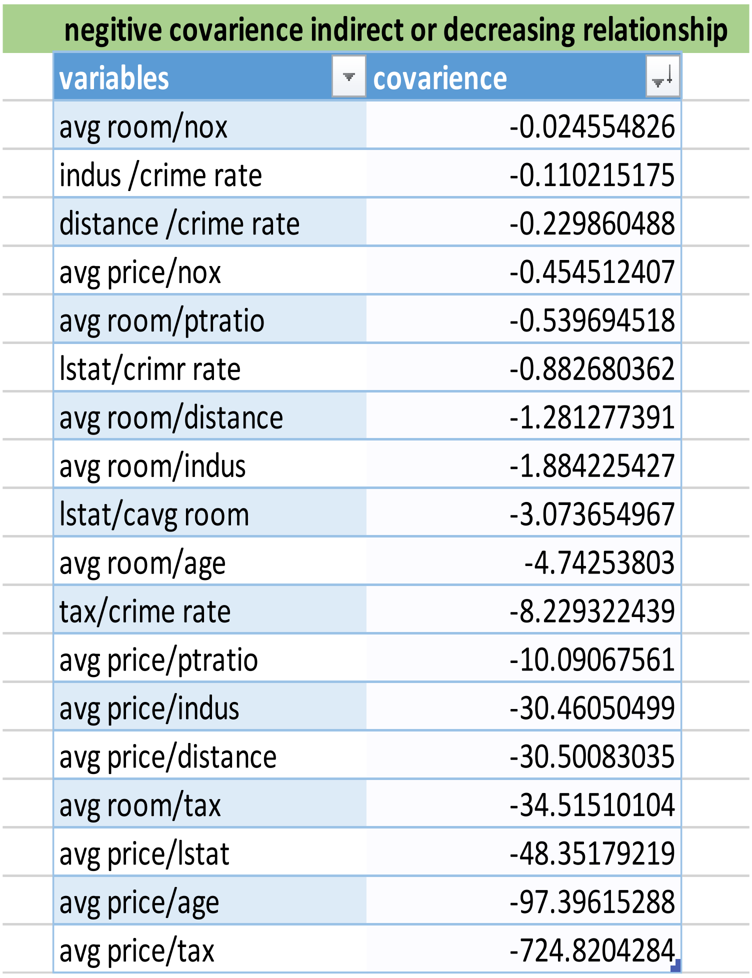


	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.516147873									
AGE	0.562915215	790.7924728								
INDUS	-0.110215175	124.2678282	46.97142974							
NOX	0.000625308	2.381211931	0.605873943	0.013401099						
DISTANCE	-0.229860488	111.5499555	35.47971449	0.615710224	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	4.677726296			
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



positive covarience direct or increasing relationship

variables	covarience
tax/age	2397.941723
tax /distance	1333.116741
tax/indus	831.7133331
lstat/tax	653.4206174
PTRATIO/tax	167.8208221
indus /age	124.2678282
lstat/age	120.8384405
distance /age	111.5499555
distance /indus	35.47971449
lstat/distance	30.32539213
lstat/indus	29.52181125
PTRATIO/age	15.90542545
tax/nox	13.02050236
PTRATIO/distance	8.74340249
lstat/ptratio	5.771300243
PTRATIO/indus	5.680854782
avg price/avg room	4.484565552
nox/age	2.381211931
avg price/crime rate	1.16201224
distance/nox	0.615710224
nox/indus	0.605873943
age /crime rate	0.562915215
lstat/nox	0.487979871
PTRATIO/crime rate	0.068168906
avg room/crime rate	0.056117778
PTRATIO/nox	0.047303654
nox/crime rate	0.000625308



negitive covarience indirect or decreasing relationship

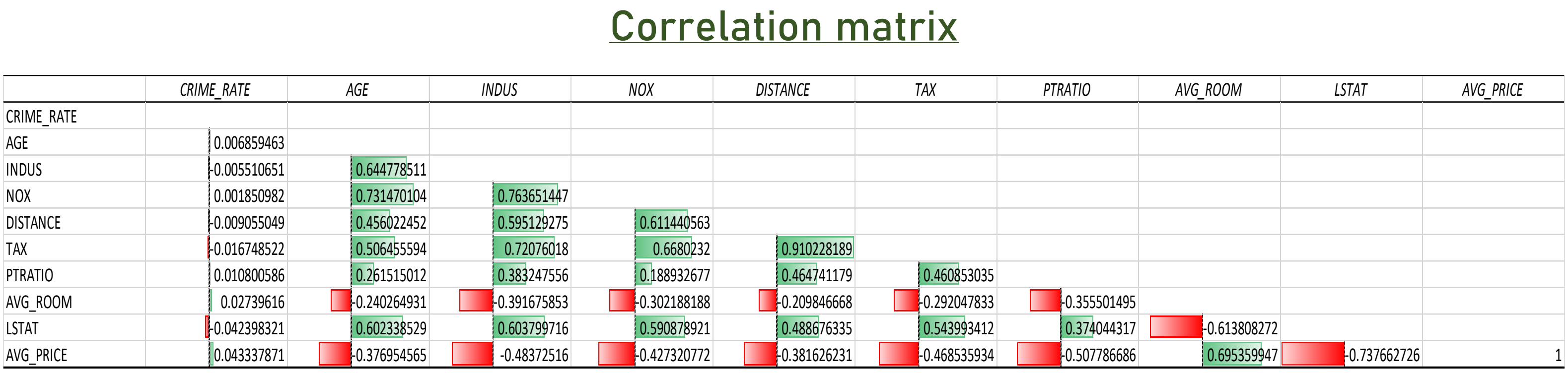
variables	covarience
avg room/nox	-0.024554826
indus /crime rate	-0.110215175
distance /crime rate	-0.229860488
avg price/nox	-0.454512407
avg room/ptratio	-0.539694518
lstat/crimr rate	-0.882680362
avg room/distance	-1.281277391
avg room/indus	-1.884225427
lstat/cavg room	-3.073654967
avg room/age	-4.74253803
tax/crime rate	-8.229322439
avg price/ptratio	-10.09067561
avg price/indus	-30.46050499
avg price/distance	-30.50083035
avg room/tax	-34.51510104
avg price/lstat	-48.35179219
avg price/age	-97.39615288
avg price/tax	-724.8204284

- From the above table we can absorb the variable with positive covariance and variable with negative covariance
- The positive covariance in the above table are arranged according to highest to lowest positive covariance
- The negative covariance in the above table are arranged according to lowest to highest negative covariance
- covariance matrix is used to analyse linear relationship between two variables
- covariance denotes the trend between the variable, is it a upward trend or downward trend
- how do two variables behave as a pair
- positive value indicates direct or increasing relationship
- negative value indicates indirect or decreasing relationship
- zero no relation
- the diagonals of the covariance matrix provide the values of each individual variable, covariance it self
- the off-diagonal entries in the matrix provides the covariance between each variable pair

Key Result: Covariance

In these results, the covariance between tax and age is 2397.941, which indicates that the relationship is highly positive. The covariance between AVG_PRICE and TAX is about −724.820 and These values indicate highly negative relationships.

4. **Create a correlation matrix of all the variables. State top 3 positively correlated pairs and top 3 negatively correlated pairs.**



Correlation matrix

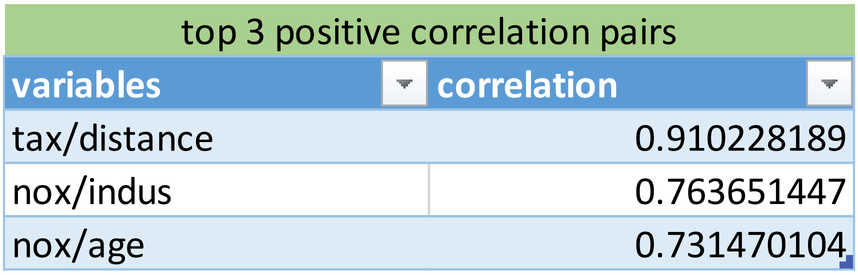
	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE										
AGE	0.006859463									
INDUS	-0.005510651	0.644778511								
NOX	0.001850982	0.731470104	0.763651447							
DISTANCE	-0.009055049	0.456022452	0.595129275	0.611440563						
TAX	-0.016748522	0.506455594	0.72076018	0.6680232	0.910228189					
PTRATIO	0.010800586	0.261515012	0.383247556	0.188932677	0.464741179	0.460853035				
AVG_ROOM	0.02739616	-0.240264931	-0.391675853	-0.302188188	-0.209846668	-0.292047833	-0.355501495			
LSTAT	-0.042398321	0.602338529	0.603799716	0.590878921	0.488676335	0.543993412	0.374044317	-0.613808272		
AVG_PRICE	0.043337871	-0.376954565	-0.48372516	-0.427320772	-0.381626231	-0.468535934	-0.507786686	0.695359947	-0.737662726	1

correlation has a value between -1 and 1 where:

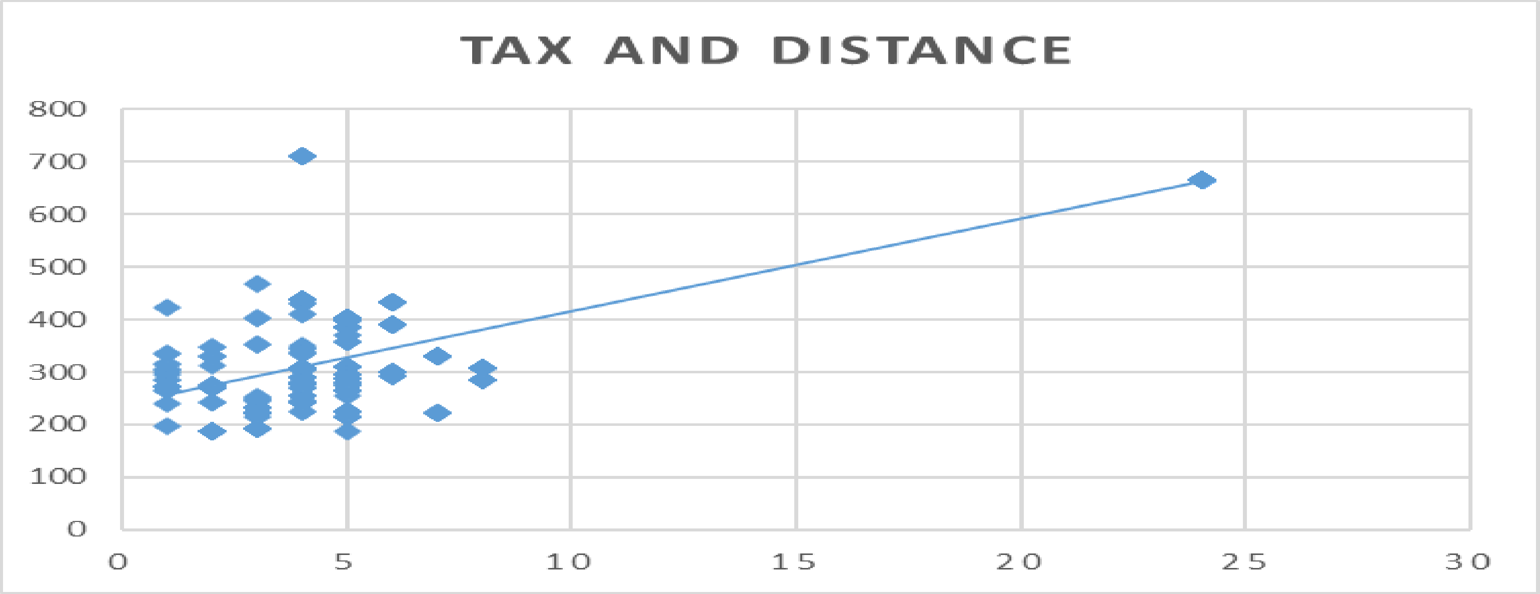
- -1 indicates a perfectly negative linear correlation between two variables
- 0 indicates no linear correlation between two variables
- 1 indicates a perfectly positive linear correlation between two variables

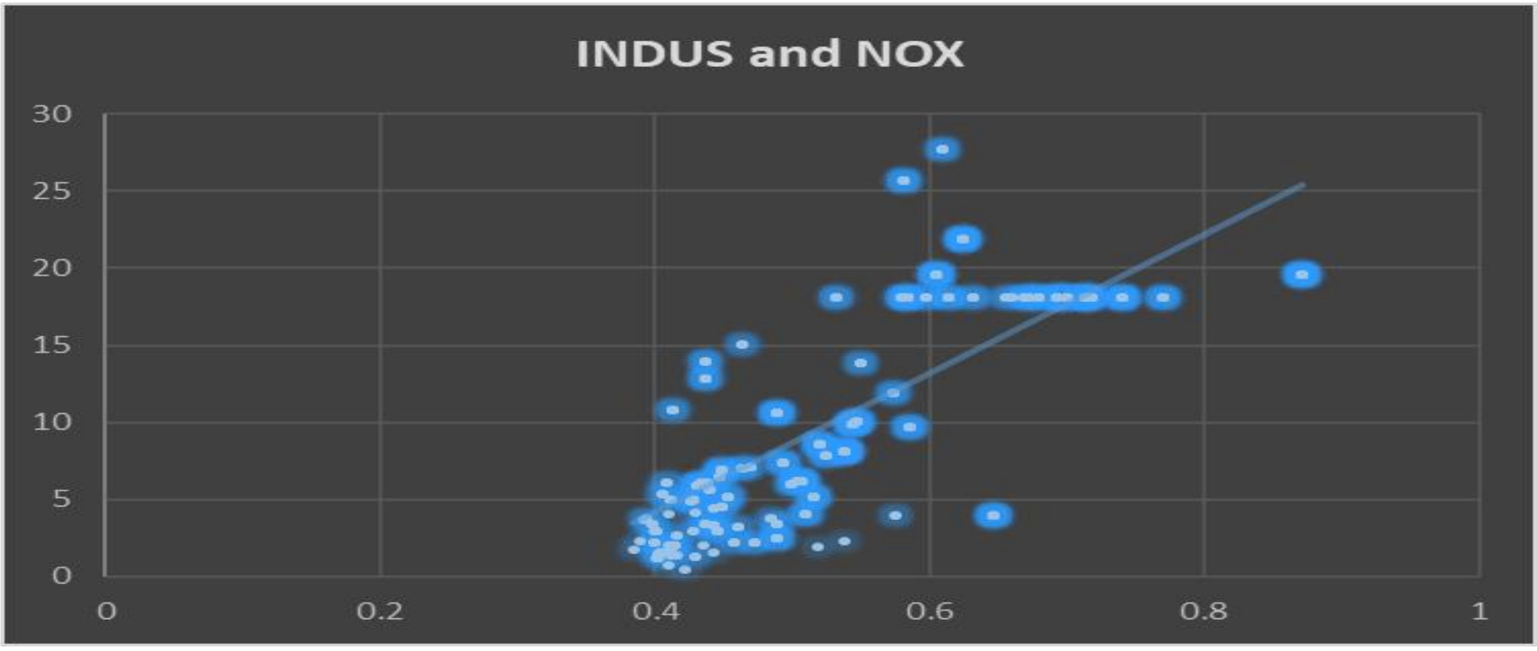
The further away the correlation coefficient is from zero, the stronger the relationship between the two variables.

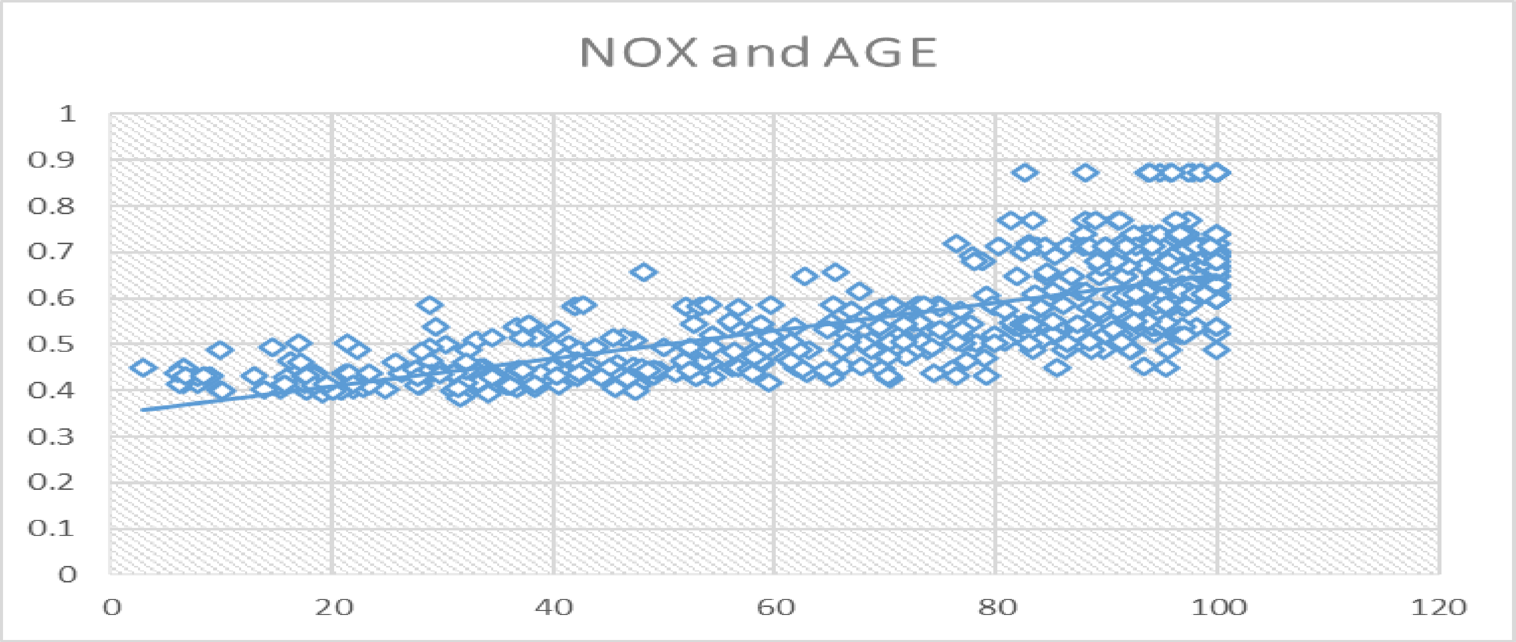
The values in the individual cells of the correlation matrix tell us the Pearson Correlation Coefficient between each pairwise combination of variables.

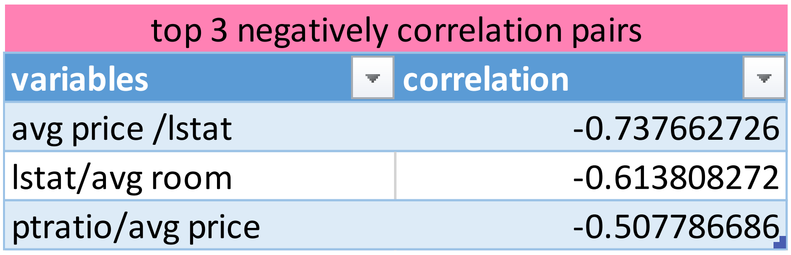


top 3 positive correlation pairs	
variables	correlation
tax/distance	0.910228189
nox/indus	0.763651447
nox/age	0.731470104



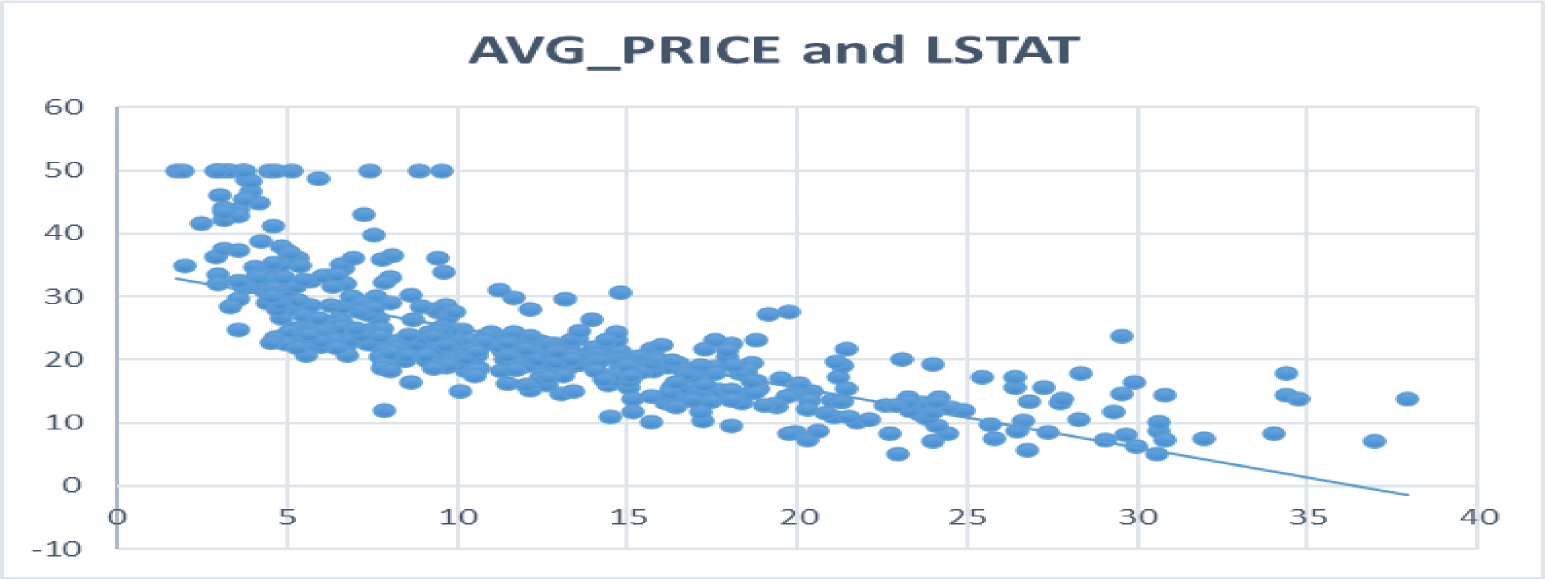




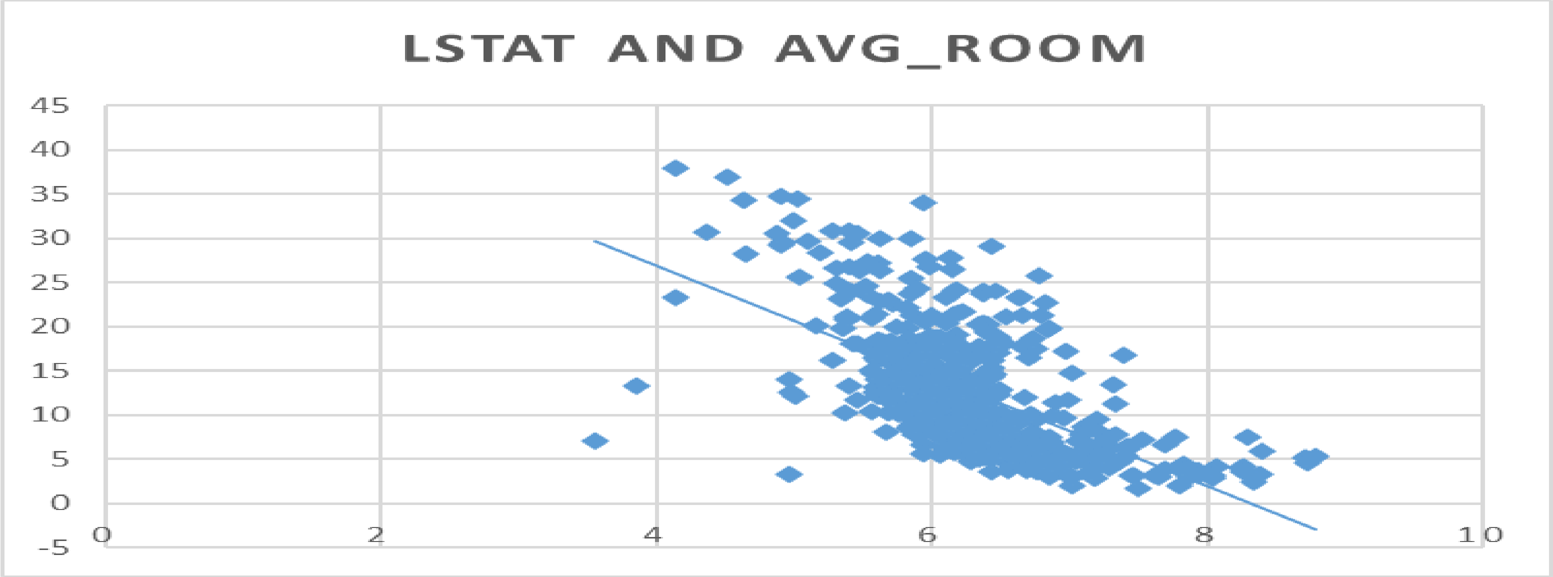


top 3 negatively correlation pairs

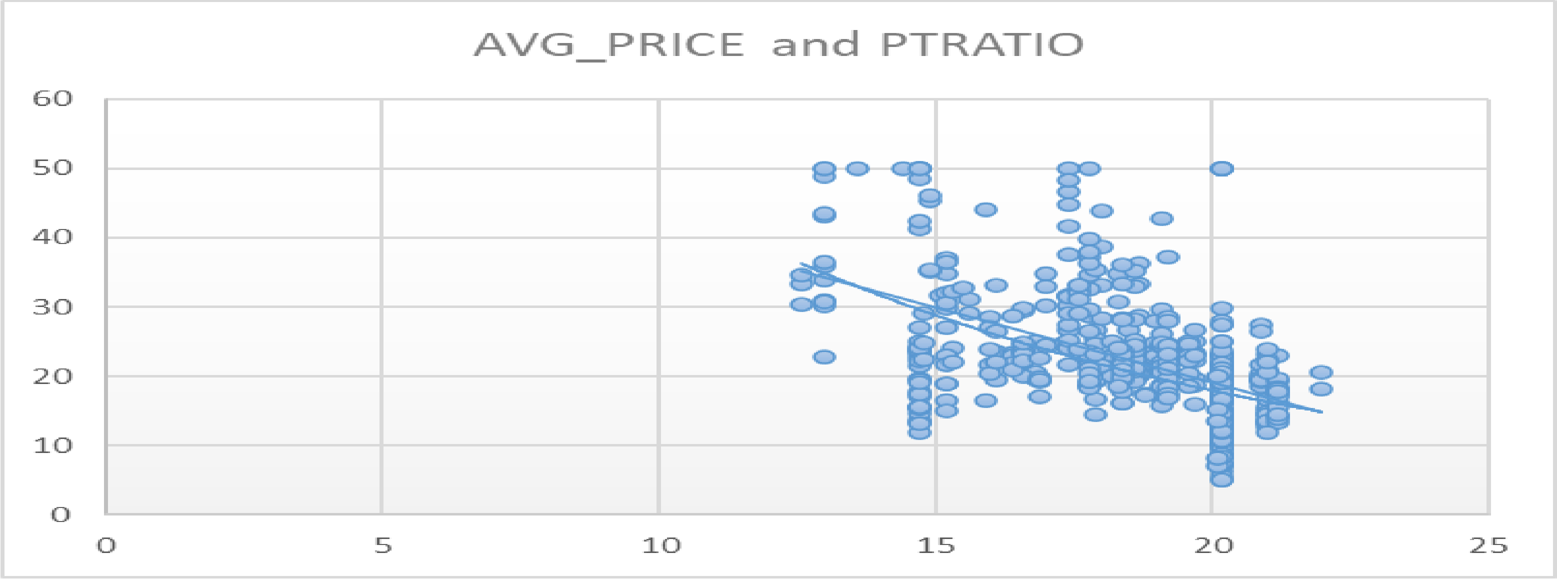
variables	correlation
avg price /lstat	-0.737662726
lstat/avg room	-0.613808272
ptratio/avg price	-0.507786686



AVG_PRICE and LSTAT



LSTAT AND AVG_ROOM



AVG_PRICE and PTRATIO

Direction
The correlation coefficient sign denotes the direction of the relationship between two variables.

- A positive value of the coefficient denotes a direct relationship. On a graph, it provides an upward slope.
- A negative value of the coefficient denotes a reverse relationship. On a graph, it provides a downward slope.

As we can see from the above scatter plot the top 3 positive correlation pairs have high correlation or positive relationship and the slop is upward sloping from left to right

Correlation between TAX/DISTANCE is 0.910228189

NOX/INDUS is 0.763651447

NOX/AGE is 0.731470104

are strongly positively correlated.

As we can see from the above scatter plot the top 3 negative correlation pairs have low correlation or negative relationship and the slop is downward sloping from right to left

Correlation between AVG PRICE /LSTAT is -0.737662726

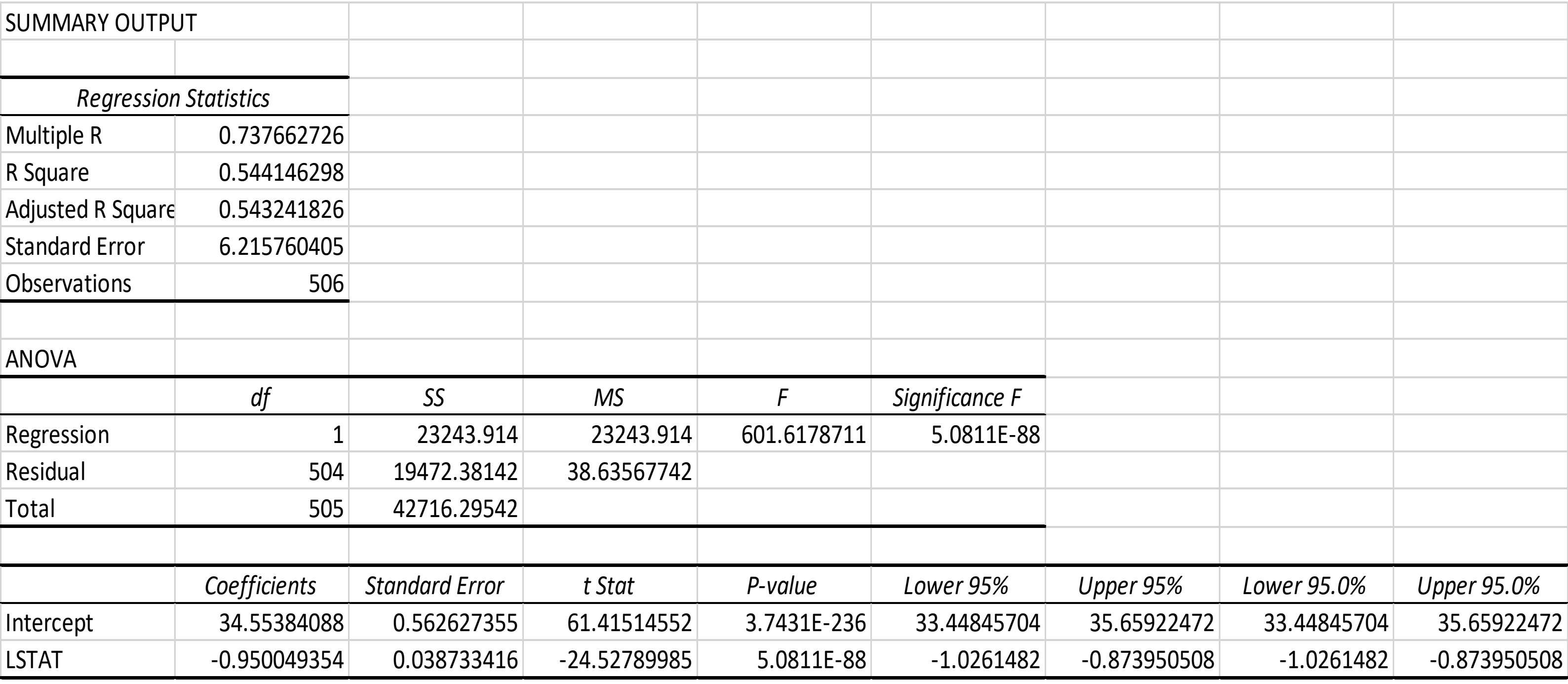
LSTAT/AVG ROOM is -0.613808272

PTRATIO/AVG PRICE is -0.507786686

these variable are highly negatively correlated.

5. **Build an initial regression model with AVG_PRICE as the y or the Dependent variable and LSTAT variable as the Independent Variable. Generate the residual plot too**

Initial regression model



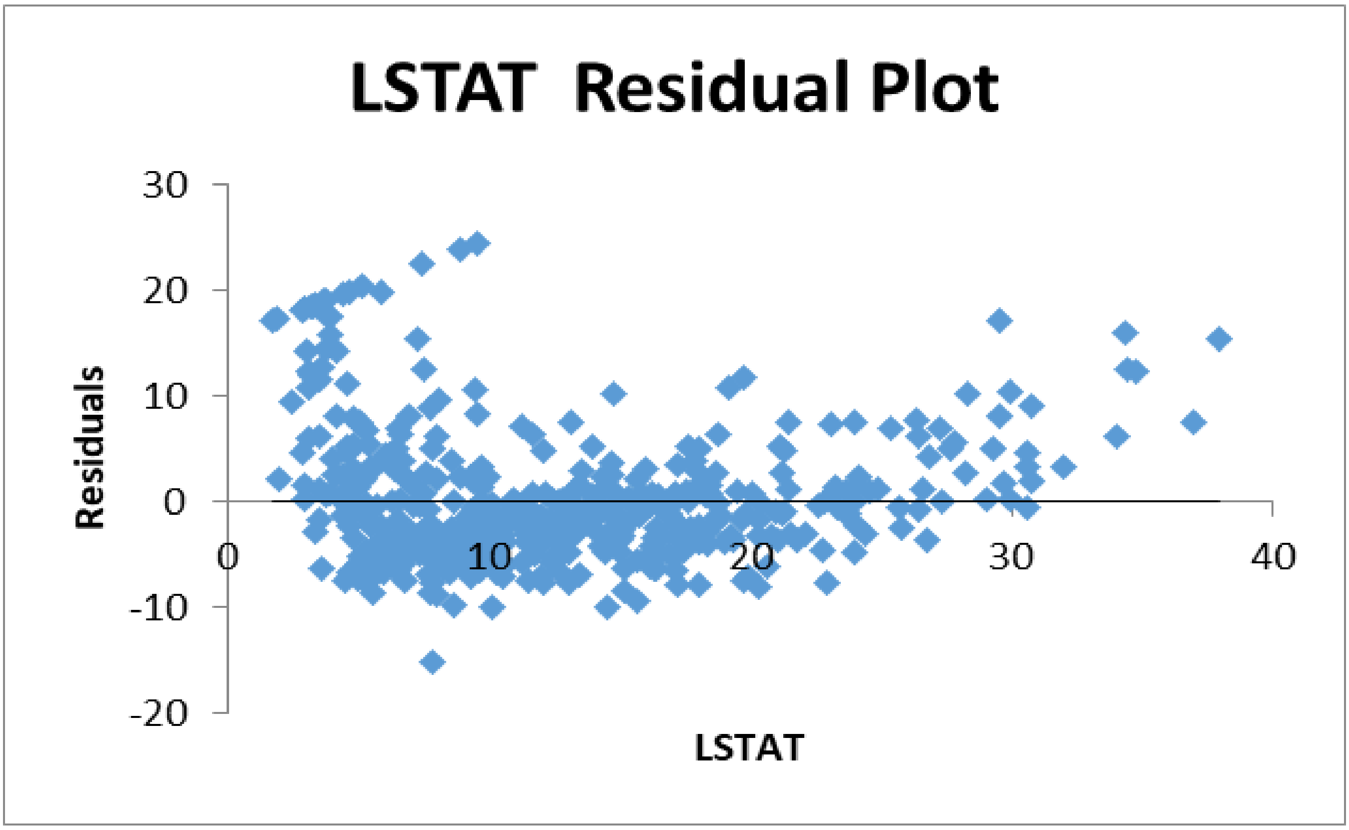
SUMMARY OUTPUT

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	23243.914	601.6178711	5.0811E-88
Residual	504	19472.38142	38.63567742		
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	61.41514552	3.7431E-236	33.44845704	35.65922472	33.44845704	35.65922472
LSTAT	-0.950049354	0.038733416	-24.52789985	5.0811E-88	-1.0261482	-0.873950508	-1.0261482	-0.873950508



5.a. What do you infer from the Regression Summary Output in terms of variance explained, coefficient value, Intercept and the Residual plot?

Multiple R. is the correlation coefficient. It tells you how strong the linear relationship is. The Multiple R is the Correlation Coefficient that measures the strength of a linear relationship between two variables. The larger the absolute value, the stronger is the relationship.

It is the square root of r squared

- 1 means a strong positive relationship
- -1 means a strong negative relationship
- 0 means no relationship at all

Multiple R between LSTAT and AVG_PRICE is 0.737662726 ,it tell us that its more than zero and close to 1 that mean it has a positive relationship but not very strong. indicates that the variables move in perfect tandem and in the same direction

What does the intercept indicate?

The intercept of 34.553 indicates that AVG_PRICE will be 34.553 if we do not spend any money on LSTAT. This is because when spend on percent of lower status of population is zero, it (zero) is multiplied by the slope o (here -0.950), resulting in a zero. This is added to your intercept, leaving you only the intercept value 34.533.

If I spend $0 on LSTAT, I can expect to have AVG_PRICE of $34.553

What do the coefficients indicate?

If the coefficient of the independent variable X is positive, it indicates for every unit increase in the independent variable; the dependent variable will increase by the value of the coefficient. This also means that for every unit decrease in the independent variable, the dependent variable will decrease by the value of the coefficient.

On the other hand, if the coefficient of the independent variable X is negative, for every unit increase in the independent variable, the dependent variable will decrease by the value of the coefficient. Correspondingly, for every unit decrease in the independent variable, the dependent variable will increase by the value of the coefficient.

We have only one independent variable in this example. Since you have only one independent variable, it is called simple linear regression. When you have more than one independent variable, it will be called multiple regression. Therefore, you will see a coefficient for every independent variable in the multiple regression output. The interpretation of these coefficients will be the same.

The coefficient (here -0.950) indicates that for every unit increase in the X variable (here LSTAT), the Y variable (here AVG_PRICE) will change by the amount of the coefficient. It is also referred to as the slope of the line in a simple linear equation.

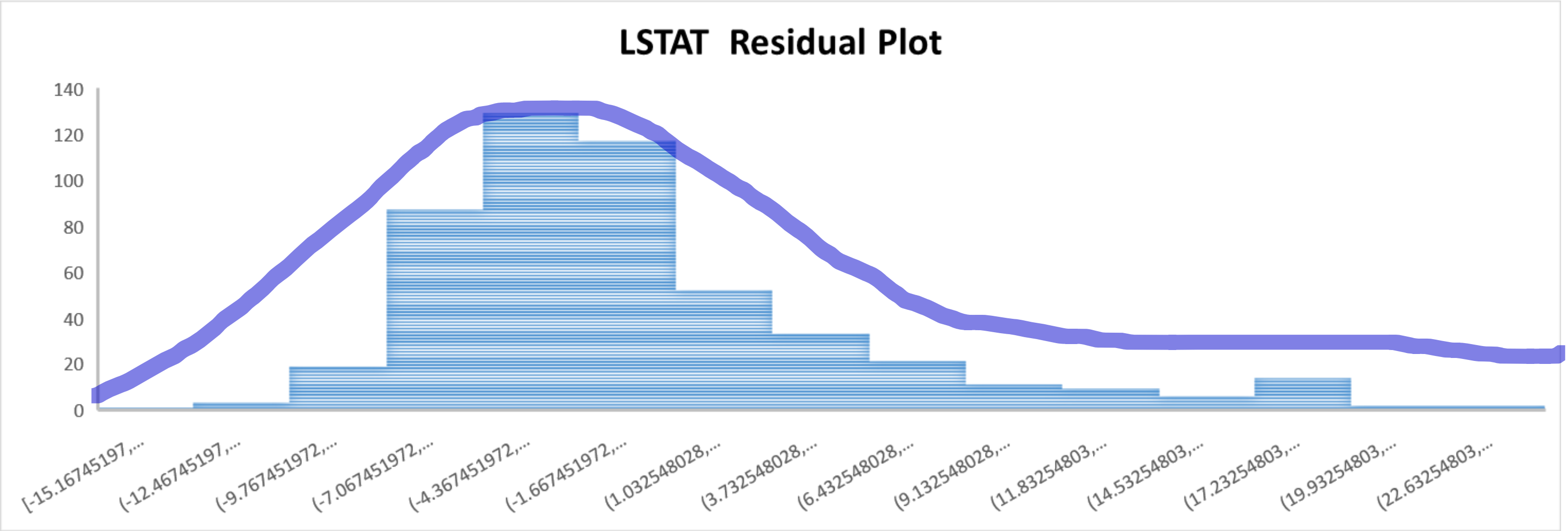
Residual Plot Analysis

A residual plot is a scatterplot that displays the residuals on the vertical axis and the independent variable on the horizontal axis. Residual plots help us to determine whether a linear model is appropriate in modelling the given data. Residual plots can be used to assess the quality of a regression .Residuals the difference between the observed value of the dependent variable (y) and the predicted value (ŷ) is called the residual (e). Each data point has one residual.

Residual = Observed value - Predicted value (e = y – ŷ)
The residuals show a curved pattern, it indicates that a linear model captures the trend of some data points better than that of others.

The data points are above the residual=0 line near Then, we detect all of the data points under the residual=0 line near The next data points are again clustered on or above the residual line=0. The data points form a curved pattern, a U-shaped pattern. Since there is a detectable pattern in the residual plot, we conclude that a linear model is not a right fit for the data.

Histogram plot of the residuals



- The Histogram of the Residual can be used to check how the variance of distributed. we can absorb from the above histogram of LSTAT residual plot most no of distribution lie in the right side of the histogram, so we can say the data is right-skewed or positively skewed.

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

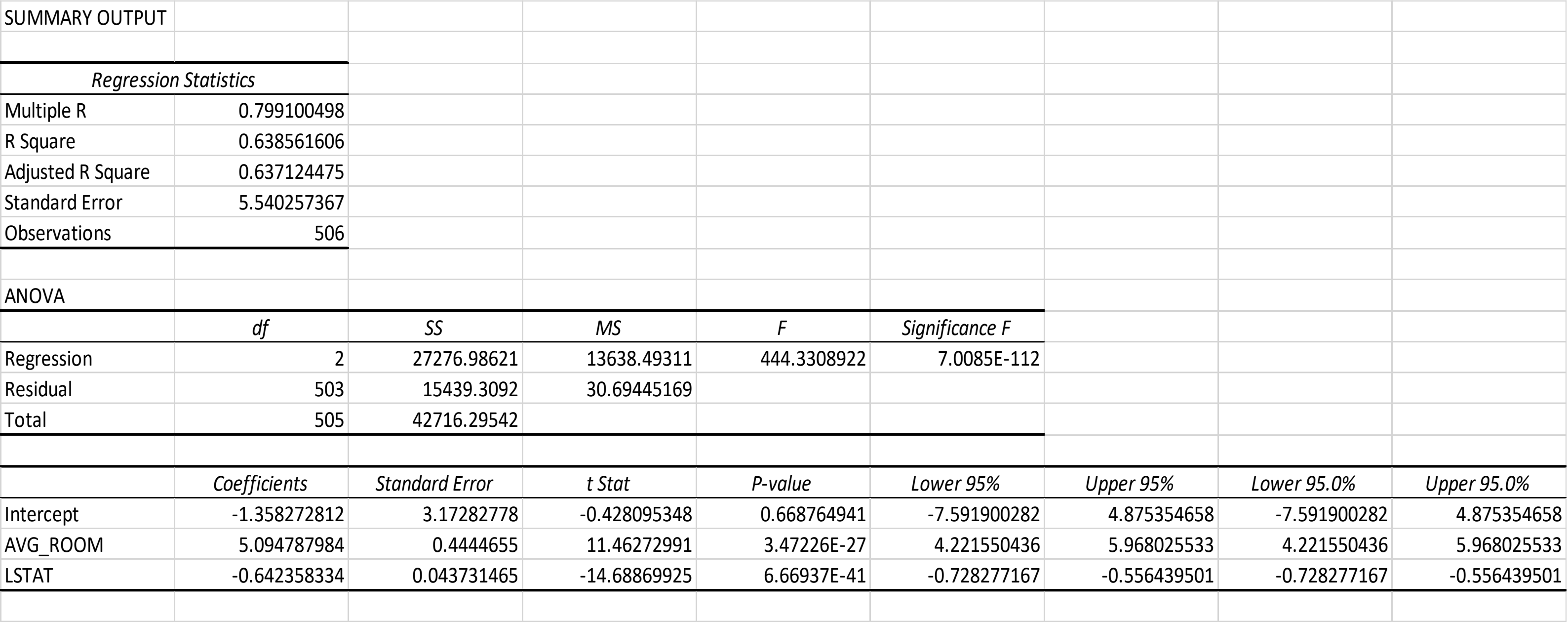
Yes, the LSTAT variable is significant for the analysis based on model

Because LSTAT p-value is 5.0811E-88 and Significant variables are those whose p-values are less than 0.05.

Standard Error of LSTAT is 0.038733416 another goodness-of-fit measure that shows the precision of regression analysis.

6. **Build another instance of the Regression model but this time including LSTAT and AVG_ROOM together as Independent variables and AVG_PRICE as the dependent variable.**

Initial regression model



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%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.358272812	3.17282778	-0.428095348	0.668764941	-7.591900282	4.875354658	-7.591900282	4.875354658
AVG_ROOM	5.094787984	0.4444655	11.46272991	3.47226E-27	4.221550436	5.968025533	4.221550436	5.968025533
LSTAT	-0.642358334	0.043731465	-14.68869925	6.66937E-41	-0.728277167	-0.556439501	-0.728277167	-0.556439501

6.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*:-**

Y=m1x1+m2x2+C

(here- m1 is AVG_ROOM coefficient & m2 is LSTAT coefficient, X1 is 7 & x2 is 20, C is intercept)

Y= 5.094787984*7+(-0.642358334*20) +(-1.358272812)

Y=21.4580764

There for the value of AVG_PRICE = 21.4580764 * 1000

AVG_PRICE=$21458.0764

- When we compare to the company quoting a value of $30000 for this locality, this company is charging only $21458.0764 this means the company is Undercharging compared to other company .
- The company is undercharging or charging $8541.92 less than the other company.

6.b. Is the performance of this model better than the previous model you built in Question 5? Compare in terms of adjusted R-square. Explain.

Yes, the performance of this model 6question is better than the previous model built in Question 5 because there is increase in the Adjuster R square .

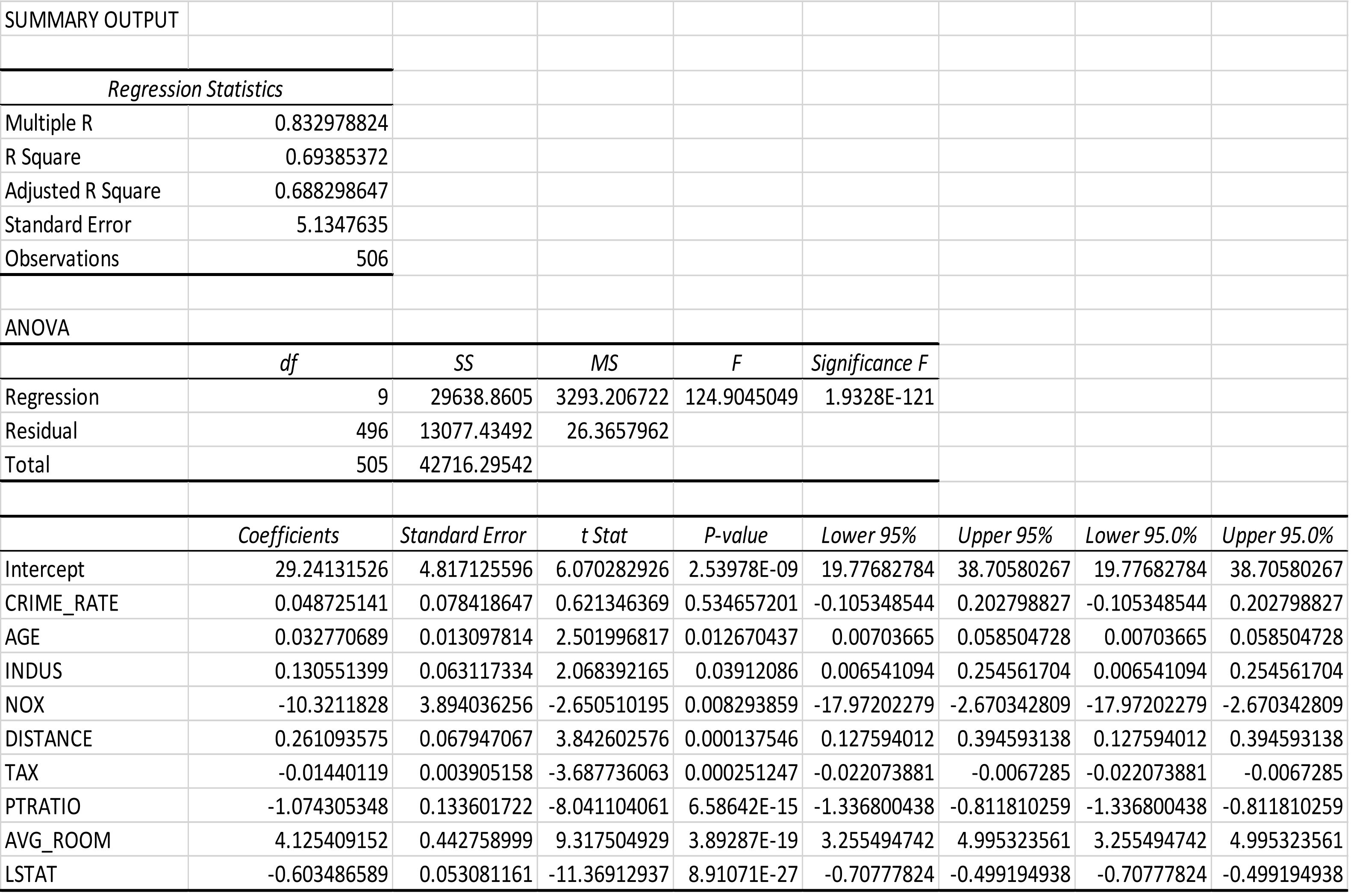
 Adjusted R Square is the modified version of R square that adjusts for predictors that are not significant to the regression model, which mean there is good correlation or significant between variables in this model .

The Adjusted R Square for the 5question model is 0.543241826 and the Adjusted R Square for this model for question 6 is 0.637124475.

We can see that there is a increase is Adjusted R square by 0.09388265 , which mean the variables in this model have better significance.

7. **Now, build a Regression model with all variables. AVG_PRICE shall be the Dependent Variable. Interpret the output in terms of adjusted R-square, coefficient and Intercept values, Significance of variables with respect to AVG_price. Explain**

Initial regression model



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%*	*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

Adjusted R-square

Adjusted R Square is the modified version of R square that adjusts for predictors that are not significant to the regression model, The Adjusted R Square for this model is 0.688298647 which mean there is good correlation or significant between variables in this model .

Intercept values

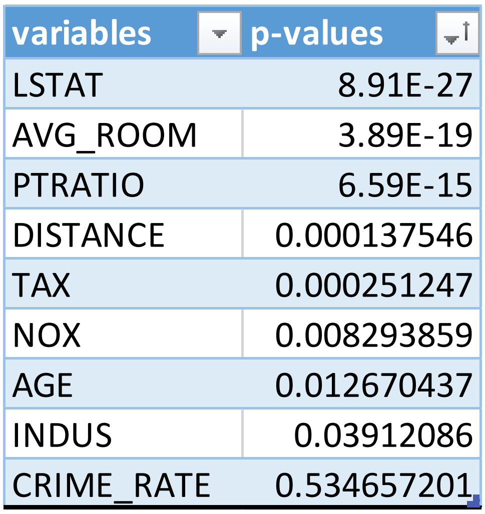
The intercept of 29.241 indicates that AVG_PRICE will be 29.241 if we do not spend on another independent variable. This is because when spend on other independent variables is zero, it (zero) is multiplied by the slope o ,resulting in a zero. This is added to your intercept, leaving you only the intercept value 29.241.

If I spend $0 on all other independent variable, I can expect to have AVG_PRICE of $29.241

coefficient

The coefficient (here -10.27270508,-1.071702473,-0.605159282,-0.014452345,0.03293496,0.130710007,0.261506423,4.125468959) indicates that for every unit increase in the X variable (here,NOX,PTRATIO,LSTAT,TAX,AGE,INDUS,DISTANCE,AVG_ROOM), the Y variable (here AVG_PRICE) will change by the amount of the coefficient. It is also referred to as the slope of the line in a simple linear equation.

Significance of variables with respect to AVG_price



variables	p-values
LSTAT	8.91E-27
AVG_ROOM	3.89E-19
PTRATIO	6.59E-15
DISTANCE	0.000137546
TAX	0.000251247
NOX	0.008293859
AGE	0.012670437
INDUS	0.03912086
CRIME_RATE	0.534657201

The P-value indicates the probability that the estimated coefficient is wrong or unreliable. We want the P-value to be as small as possible

Significant variables are those whose p-values are less than 0.05. If the p-value is greater than 0.05 then it is insignificant.

According to the table of variables and their p-values with respect to AVG_PRICE we absorber :-

- The p-value of LSTAT is 8.91E-27 that's is least p-value compared to all variables its less than 0.05 which means LSTAT is a significant variable

- The p-value of AVG_ROOM is 3.89E-19 that's its p-value is less than 0.05 which means AVG_ROOM is a significant variable

- The p-value of PTRATIO is 0.0001375 that's its p-value is less than 0.05 which means PTRATIO is a significant variable

- The p-value of DISTANCE is 0.0002512 that’s its p-value is less than 0.05 which means DISTANCE is a significant variable

- The p-value of TAX 0.0002512 that’s is p-value its less than 0.05 which means TAX is a significant variable

- The p-value of NOX is 0.008293 that’s its p-value is less than 0.05 which means NOX is a significant variable

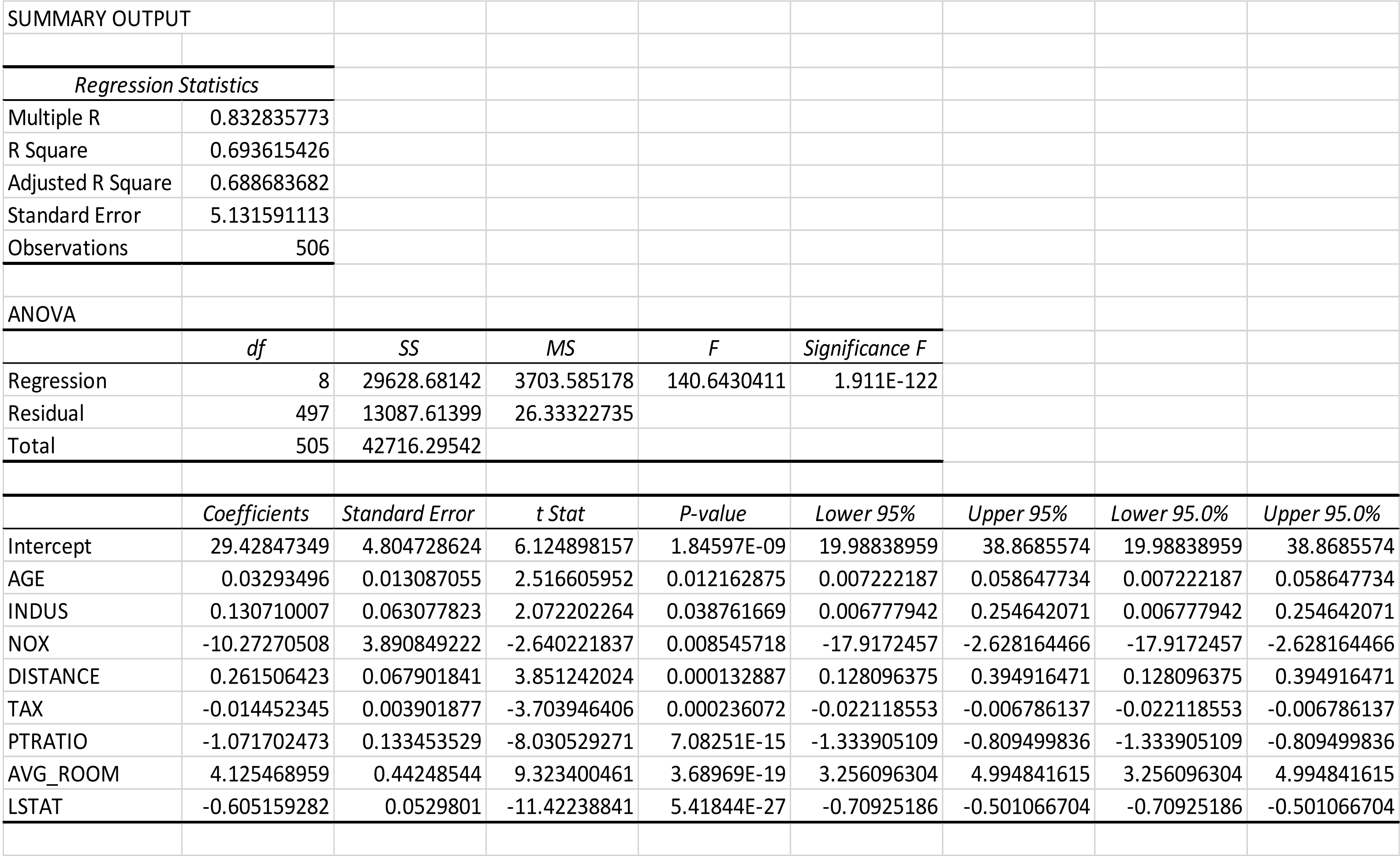
- The p-value of AGE is 0.01267 that’s its p-value is less than 0.05 which means AGE is a significant variable

- The p-value of INDUS is 0.03912 that’s its p-value is close to 0.05 but less than 0.05 which means INDUS is a significant variable

- The p-value of CRIMRATE is 0.5346 that’s is highest p-value compared to all variables its more than 0.05 which means CRIME_RATE is a insignificant variable.

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**

Initial regression model (**significant variables)**



SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.832835773							
R Square	0.693615426							
Adjusted R Square	0.688683682							
Standard Error	5.131591113							
Observations	506							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	8	29628.68142	3703.585178	140.6430411	1.911E-122			
Residual	497	13087.61399	26.33322735					
Total	505	42716.29542						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.42847349	4.804728624	6.124898157	1.84597E-09	19.98838959	38.8685574	19.98838959	38.8685574
AGE	0.03293496	0.013087055	2.516605952	0.012162875	0.007222187	0.058647734	0.007222187	0.058647734
INDUS	0.130710007	0.063077823	2.072202264	0.038761669	0.006777942	0.254642071	0.006777942	0.254642071
NOX	-10.27270508	3.890849222	-2.640221837	0.008545718	-17.9172457	-2.628164466	-17.9172457	-2.628164466
DISTANCE	0.261506423	0.067901841	3.851242024	0.000132887	0.128096375	0.394916471	0.128096375	0.394916471
TAX	-0.014452345	0.003901877	-3.703946406	0.000236072	-0.022118553	-0.006786137	-0.022118553	-0.006786137
PTRATIO	-1.071702473	0.133453529	-8.030529271	7.08251E-15	-1.333905109	-0.809499836	-1.333905109	-0.809499836
AVG_ROOM	4.125468959	0.44248544	9.323400461	3.68969E-19	3.256096304	4.994841615	3.256096304	4.994841615
LSTAT	-0.605159282	0.0529801	-11.42238841	5.41844E-27	-0.70925186	-0.501066704	-0.70925186	-0.501066704

a. **Interpret the output of this model.**

Multiple R between all variables in the regression model and AVG_PRICE is 0.83283 ,it tell us that its more than zero and close to 1 that mean it has a positive relationship. Indicates that the variables move in perfect tandem and in the same direction

R squared. This is r2, the Coefficient of Determination. it is the sum of the squared deviations of the original data from the mean. R Square signifies the Coefficient of Determination, which shows the goodness of fit. It shows how many points fall on the regression line. In our model, the value of R square is 0.693, which is an average fit. In other words, 69.3% of the dependent variables (y-values) are explained by the independent variables (x-values).

Adjusted R-square is the modified version of R square that adjusts for predictors that are not significant to the regression model, The Adjusted R Square for this model is 0.688683682 which mean there is good correlation or significant between variables in this model

Standard Error of the regression is An estimate of the standard deviation of the error μ ,it represents the average distance that the observed values fall from the regression line. Conveniently, it tells you how wrong the regression model is on average using the units of the response variable. This is not the same as the standard error in descriptive statistics. The standard error of the regression is the precision that the regression coefficient is measured.

Standard Error is another goodness-of-fit measure that shows the precision of your regression analysis. The standard error for this model is 5.131591113

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?**

When we Compare the adjusted R-square value of this model with the model in the previous question seven, we can see this model question number eight performs better according to the value of adjusted R-square.

The Adjusted R Square for the 7question model is **0.688298647**and the Adjusted R Square for this model for question 8 is **0.688683682**.

We can see that there is a increase is Adjusted R square by **0.000385035**, which mean the variables in this model have better significance.

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?**

If the value of NOX is more in a locality in this town the AVG_PRICE will decrease

values of the Coefficients in ascending order

	Coefficients
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
Intercept	29.42847349

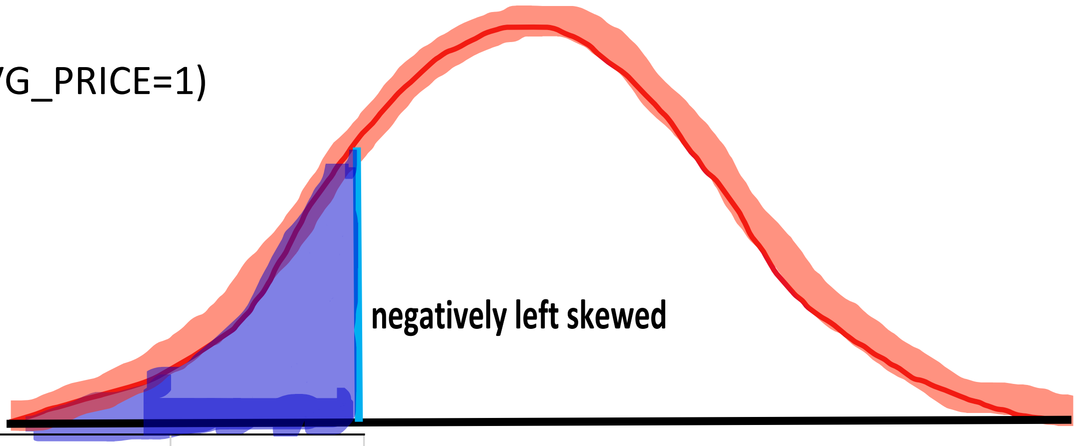
Performing hypothesis testing:-

Null hypothesis(Ho)- AVG_PRICE > NOX

Alternative hypothesis(H1)- AVG_PRICE < NOX

(Left tailed hypothesis test)

(AVG_PRICE=1)



NOX	-10.27270508

Accept null hypothesis(Ho) and reject alternative hypothesis (H1)

d. Write the regression equation from this model.

Regression equation

$Y=m1x1+m2x2+mnxn+C$

$$Y=0.033x+0.130x+0.130x+(-10.28)x+0.261x+(-0.014)x+(-1.071)x+4.125x+(-0.605)x+29.43$$