

# Treo's Real estate agency Project

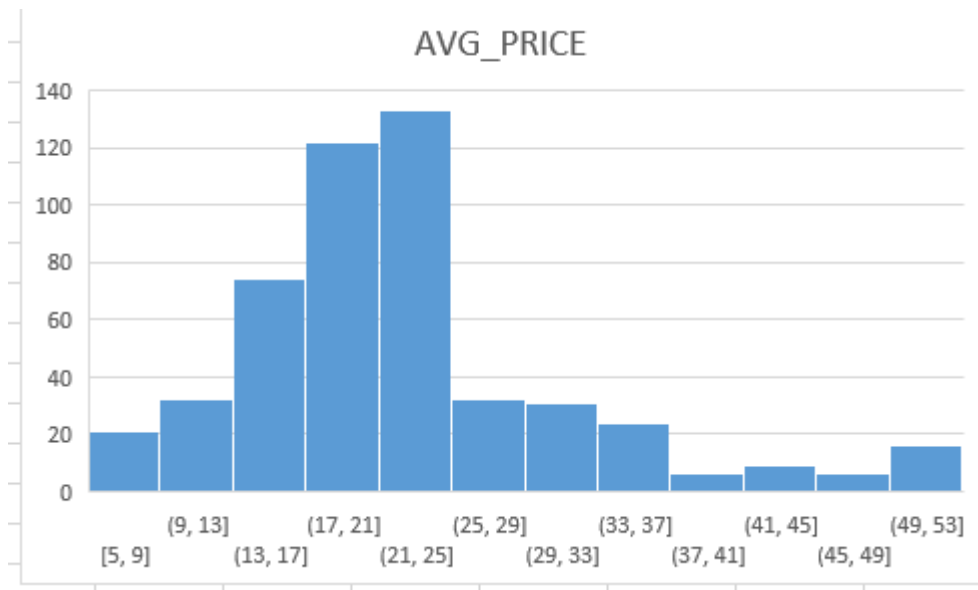
By GOWTHAM SINGARAM

**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?**

CRIME_RATE		AGE		INDUS		NOX		DISTANCE		TAX	
Mean	4.872	Mean	68.575	Mean	11.137	Mean	0.5547	Mean	9.5494	Mean	408.24
Standard Error	0.1293	Standard Error	1.2514	Standard Error	0.305	Standard Error	0.0052	Standard Error	0.3871	Standard Error	7.4924
Median	4.82	Median	77.5	Median	9.69	Median	0.538	Median	5	Median	330
Mode	3.43	Mode	100	Mode	18.1	Mode	0.538	Mode	24	Mode	666
Standard Deviation	2.3211	Standard Deviation	28.149	Standard Deviation	6.8604	Standard Deviation	0.1159	Standard Deviation	8.7073	Standard Deviation	168.54
Sample Variance	8.533	Sample Variance	792.36	Sample Variance	47.064	Sample Variance	0.0134	Sample Variance	75.816	Sample Variance	28405
Kurtosis	-1.189	Kurtosis	-0.968	Kurtosis	-1.234	Kurtosis	-0.065	Kurtosis	-0.867	Kurtosis	-1.142
Skewness	0.0217	Skewness	-0.539	Skewness	0.295	Skewness	0.7293	Skewness	1.0048	Skewness	0.67
Range	9.95	Range	97.1	Range	27.28	Range	0.486	Range	23	Range	524
Minimum	0.04	Minimum	2.9	Minimum	0.46	Minimum	0.385	Minimum	1	Minimum	187
Maximum	9.99	Maximum	100	Maximum	27.74	Maximum	0.871	Maximum	24	Maximum	711
Sum	2465.2	Sum	34639	Sum	5635.2	Sum	280.68	Sum	4832	Sum	206568
Count	506	Count	506	Count	506	Count	506	Count	506	Count	506
PT-RATIO		AVG_ROOM		LSTAT		AVG_PRICE					
Mean	18.456	Mean	6.2846	Mean	12.653	Mean	22.533				
Standard Error	0.0962	Standard Error	0.0312	Standard Error	0.3175	Standard Error	0.4089				
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	2.1649	Standard Deviation	0.7026	Standard Deviation	7.1411	Standard Deviation	9.1971				
Sample Variance	4.687	Sample Variance	0.4937	Sample Variance	50.995	Sample Variance	84.587				
Kurtosis	-0.285	Kurtosis	1.8915	Kurtosis	0.4932	Kurtosis	1.4952				
Skewness	-0.802	Skewness	0.4036	Skewness	0.9065	Skewness	1.1081				
Range	9.4	Range	5.219	Range	36.24	Range	45				
Minimum	12.6	Minimum	3.561	Minimum	1.73	Minimum	5				
Maximum	22	Maximum	8.78	Maximum	37.97	Maximum	50				
Sum	9338.5	Sum	3180	Sum	6402.5	Sum	11402				
Count	506	Count	506	Count	506	Count	506				

- From the given data, the mean=68.575 shows the average age value in the town.
- From the median=77.5, it clearly shows that most of them are aged persons in the town.
- All the given variables except the PT-ratio have a Positive skewness.
- There is minimum outlier data in Crime rate, NOX and Avg. Room.

## 2. Plot the histogram of the Avg\_Price Variable. What do you infer?



- AVERAGE PRICE is a dependent variable, which has a very few outlier data and it is Rightly skewed with a leptokurtic kurtosis.
- There are 133 Houses with the price range of 21000 dollars to 25000 dollars.

## 3. 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	790.7925								
INDUS	-0.110215175	124.2678	46.97143							
NOX	0.000625308	2.381212	0.605874	0.013401						
DISTANCE	-0.229860488	111.55	35.47971	0.61571	75.66653					
TAX	-8.229322439	2397.942	831.7133	13.0205	1333.117	28348.62				
PTRATIO	0.068168906	15.90543	5.680855	0.047304	8.743402	167.8208	4.677726			
AVG_ROOM	0.056117778	-4.74254	-1.88423	-0.02455	-1.28128	-34.5151	-0.53969	0.492695216		
LSTAT	-0.882680362	120.8384	29.52181	0.48798	30.32539	653.4206	5.7713	-3.07365497	50.89398	
AVG_PRICE	1.16201224	-97.3962	-30.4605	-0.45451	-30.5008	-724.82	-10.0907	4.484565552	-48.3518	84.41955616

There are both Positive covariance and Negative covariance in the above matrix.

The most positive covariance value is 2397.942 (Tax and Age)

The most Negative covariance value is -724.82 (Tax and Avg\_price)

**4. Create a correlation matrix of all the variables as shown in the Videos and various case studies. State top 3 positively correlated pairs and top 3 negatively correlated pairs.**

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	1									
AGE	0.006859463	1								
INDUS	-0.005510651	0.644779	1							
NOX	0.001850982	0.73147	0.763651	1						
DISTANCE	-0.009055049	0.456022	0.595129	0.611441	1					
TAX	-0.016748522	0.506456	0.72076	0.668023	0.91022819	1				
PTRATIO	0.010800586	0.261515	0.383248	0.188933	0.46474118	0.460853	1			
AVG_ROOM	0.02739616	-0.24026	-0.39168	-0.30219	-0.2098467	-0.29205	-0.3555	1		
LSTAT	-0.042398321	0.602339	0.6038	0.590879	0.48867633	0.543993	0.374044	-0.61380827	1	
AVG_PRICE	0.043337871	-0.37695	-0.48373	-0.42732	-0.3816262	-0.46854	-0.50779	0.695359947	-0.73766	1

**Top 3 Positive correlated values**

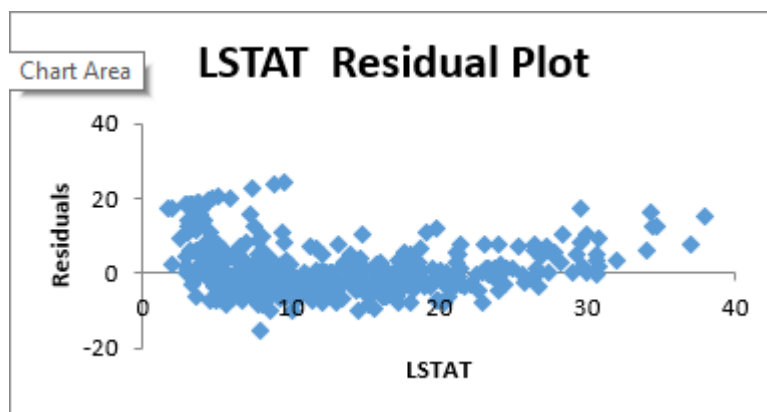
- Tax and Distance
- Indus and Nox
- Age and Nox

**Top 3 Negative Correlated values**

- Avg\_price and Lstat
- Avg\_room and Lstat
- Avg\_price and PT ratio

**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.**

SUMMARY OUTPUT								
<b>Regression Statistics</b>								
Multiple R	0.7376627							
R Square	0.5441463							
Adjusted R	0.5432418							
Standard Error	6.2157604							
Observations	506							
<b>ANOVA</b>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	23243.914	23243.91	601.6179	5.0811E-88			
Residual	504	19472.38142	38.63568					
Total	505	42716.29542						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	34.553841	0.562627355	61.41515	3.7E-236	33.44845704	35.6592247	33.44845704	35.65922472
LSTAT	-0.9500494	0.038733416	-24.5279	5.08E-88	-1.0261482	-0.8739505	-1.0261482	-0.873950508



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

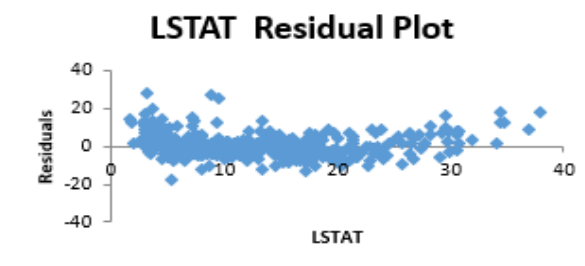
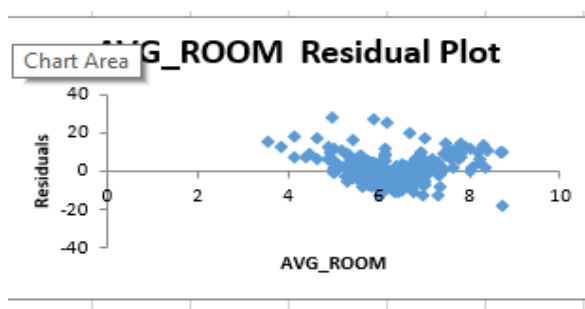
- The value of the variance is less than 0.05, so the model is highly significant.
- From the residual plot, we can see that there is some outlier in a data.

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

Yes, LSTAT variable is significant for the analysed model.

**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?**

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.7991005							
R Square	0.63856161							
Adjusted R Square	0.63712448							
Standard Error	5.54025737							
Observations	506							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	2	27276.98621	13638.49311	444.331	7.01E-112			
Residual	503	15439.30912	30.69445					
Total	505	42716.29542						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1.3582728	3.17282778	-0.428095	0.66876	-7.5919003	4.8753547	-7.591900282	4.875354658
AVG_ROOM	5.09478798	0.4444655	11.46273	3.5E-27	4.2215504	5.9680255	4.221550436	5.968025533
LSTAT	-0.6423583	0.043731465	-14.6887	6.7E-41	-0.7282772	-0.55644	-0.728277167	-0.556439501



- 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?

BY using a regression equation, the AVG\_PRICE value is 21.4581. Comparing to the quoting value, the company is overcharging for the property.

- 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.

- Previous model Adj. R-square= 0.543241825954707
- New model Adj. R-square= 0.637124475470123
- Yes, the performance of this new model is better than the previous model by comparing to the value of adjusted R-square.

**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?**

SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.8329788							
R Square	0.6938537							
Adjusted R Square	0.6882986							
Standard Error	5.1347635							
Observations	506							
ANOVA								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	9	29638.8605	3293.21	124.905	2E-121			
Residual	496	13077.43492	26.3658					
Total	505	42716.29542						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	29.241315	4.817125596	6.07028	2.5E-09	19.7768	38.70580267	19.77682784	38.705803
CRIME_RATE	0.0487251	0.078418647	0.62135	0.53466	-0.10535	0.202798827	-0.10534854	0.2027988
AGE	0.0327707	0.013097814	2.502	0.01267	0.00704	0.058504728	0.00703665	0.0585047
INDUS	0.1305514	0.063117334	2.06839	0.03912	0.00654	0.254561704	0.006541094	0.2545617
NOX	-10.32118	3.894036256	-2.6505	0.00829	-17.972	-2.67034281	-17.9720228	-2.6703428
DISTANCE	0.2610936	0.067947067	3.8426	0.00014	0.12759	0.394593138	0.127594012	0.3945931
TAX	-0.014401	0.003905158	-3.6877	0.00025	-0.02207	-0.0067285	-0.02207388	-0.0067285
PTRATIO	-1.074305	0.133601722	-8.0411	6.6E-15	-1.3368	-0.81181026	-1.33680044	-0.8118103
AVG_ROOM	4.1254092	0.442758999	9.3175	3.9E-19	3.25549	4.995323561	3.255494742	4.9953236
LSTAT	-0.603487	0.053081161	-11.369	8.9E-27	-0.70778	-0.49919494	-0.70777824	-0.4991949

By seeing the coefficient from the regression data, we can say that

- Avg. price and Avg. room variables are directly proportional.
- NOX and Avg. Price are inversely proportional to each other.

The significance variables in this model are:

- AGE
- INDUS
- NOX
- DISTANCE
- TAX
- PTRATIO
- AVG\_ROOM
- LSTAT

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.

SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.83283577							
R Square	0.69361543							
Adjusted R Square	0.68868368							
Standard Error	5.13159111							
Observations	506							
ANOVA								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	8	29628.68142	3703.59	140.643	1.911E-122			
Residual	497	13087.61399	26.3332					
Total	505	42716.29542						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	29.4284735	4.804728624	6.1249	1.8E-09	19.98838959	38.86856	19.98839	38.8685574
AGE	0.03293496	0.013087055	2.51661	0.01216	0.007222187	0.058648	0.0072222	0.05864773
INDUS	0.13071001	0.063077823	2.0722	0.03876	0.006777942	0.254642	0.0067779	0.25464207
NOX	-10.272705	3.890849222	-2.6402	0.00855	-17.9172457	-2.62816	-17.917246	-2.6281645
DISTANCE	0.26150642	0.067901841	3.85124	0.00013	0.128096375	0.394916	0.1280964	0.39491647
TAX	-0.0144523	0.003901877	-3.7039	0.00024	-0.02211855	-0.00679	-0.0221186	-0.0067861
PTRATIO	-1.0717025	0.133453529	-8.0305	7.1E-15	-1.33390511	-0.8095	-1.3339051	-0.8094998
AVG_ROOM	4.12546896	0.44248544	9.3234	3.7E-19	3.256096304	4.994842	3.2560963	4.99484161
LSTAT	-0.6051593	0.0529801	-11.422	5.4E-27	-0.70925186	-0.50107	-0.7092519	-0.5010667

a. Interpret the output of this model.

This model has a Better Adjusted R-square value of 68.86%.

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 New model performs well than the previous model by comparing the value of an Adjusted R-square value.



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?

Column1	Column2
NOX	-10.272705
PTRATIO	-1.0717025
LSTAT	-0.6051593
TAX	-0.0144523
AGE	0.03293496
INDUS	0.13071001
DISTANCE	0.26150642
AVG_ROOM	4.12546896

The **AVERAGE\_PRICE** is HIGH, if the **NOX** is LOW

d. Write the regression equation from this model.

Multi-Linear regression equation:

$$Y = m_1X_1 + m_2X_2 + m_3X_3 + m_4X_4 + m_5X_5 + m_6X_6 + m_7X_7 + m_8X_8 + c$$

Multi-linear regression equation of this model:

$$Y = -10.2727050815094x_1 - 1.07170247269449x_2 - 0.605159282035406x_3 - 0.0144523450364819x_4 + 0.0329349604286303x_5 + 0.130710006682182x_6 + 0.261506423001819x_7 + 4.12546895908474x_8 + 29.4284734939458$$

Y= Predicted Variable

m= weight of the variable

$X_n$ = Independent variable

C= Intercept