TERRO REAL ESTATE

Assignment – Terro’s Real Estate Agency

By : Loganathan S
Data Analytics Nov ‘05

Table of Contents

Executive Summary

"Finding out the most relevant features for pricing of a house" Terro's real-estate is an agency that estimates the pricing of houses in a certain locality. The pricing is concluded based on different features / factors of a property. This also helps them in identifying the business value of a property. To do this activity the company employs an "Auditor", who studies various geographic features of a property like pollution level (NOX), crime rate, education facilities (pupil to teacher ratio), connectivity (distance from highway), etc. This helps in determining the price of a property

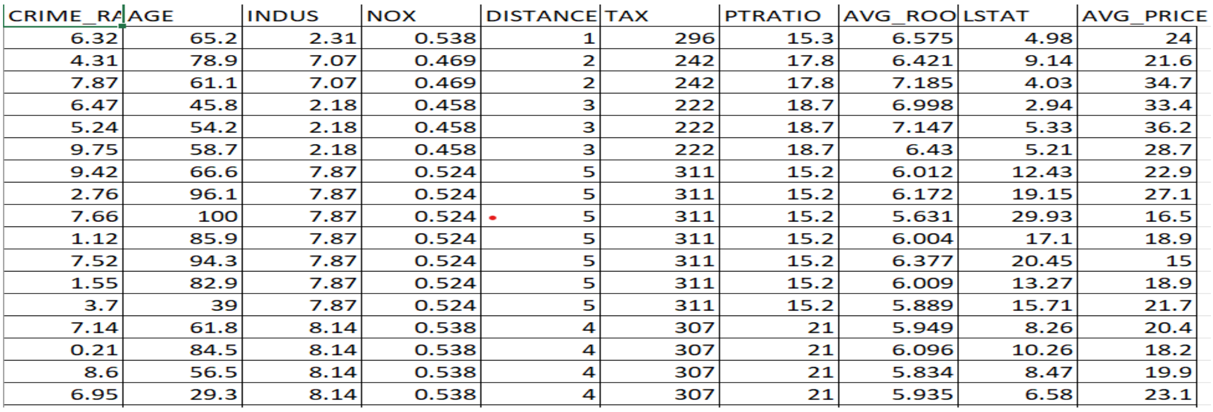
Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. The data consists of 506 different houses in Boston. Analyze the different attributes of the houses make which can help in analyzing the price of the houses. This assignment should help the student in exploring the summary statistics, contingency tables, conditional probabilities & hypothesis testing.

Data Description

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

Sample of the dataset:



CRIME_RA	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROO	LSTAT	AVG_PRICE
6.32	65.2	2.31	0.538	1	296	15.3	6.575	4.98	24
4.31	78.9	7.07	0.469	2	242	17.8	6.421	9.14	21.6
7.87	61.1	7.07	0.469	2	242	17.8	7.185	4.03	34.7
6.47	45.8	2.18	0.458	3	222	18.7	6.998	2.94	33.4
5.24	54.2	2.18	0.458	3	222	18.7	7.147	5.33	36.2
9.75	58.7	2.18	0.458	3	222	18.7	6.43	5.21	28.7
9.42	66.6	7.87	0.524	5	311	15.2	6.012	12.43	22.9
2.76	96.1	7.87	0.524	5	311	15.2	6.172	19.15	27.1
7.66	100	7.87	0.524	5	311	15.2	5.631	29.93	16.5
1.12	85.9	7.87	0.524	5	311	15.2	6.004	17.1	18.9
7.52	94.3	7.87	0.524	5	311	15.2	6.377	20.45	15
1.55	82.9	7.87	0.524	5	311	15.2	6.009	13.27	18.9
3.7	39	7.87	0.524	5	311	15.2	5.889	15.71	21.7
7.14	61.8	8.14	0.538	4	307	21	5.949	8.26	20.4
0.21	84.5	8.14	0.538	4	307	21	6.096	10.26	18.2
8.6	56.5	8.14	0.538	4	307	21	5.834	8.47	19.9
6.95	29.3	8.14	0.538	4	307	21	5.935	6.58	23.1

Table 1. Dataset Sample

This Dataset has collection of 18 attributes of houses in Boston. Each house make has different sets of attributes. Based on the characteristic price of the house is defined.

Exploratory Data Analysis

Let us check the types of variables in the data frame.

CRIME RATE	float64
INDUSTRY	float64
NOX	float64
AVG_ROOM	float64
AGE	float64
PTRATIO	float64
LSTAT	float64
AVG_PRICE	float64
DISTANCE	int64
TAX	int64

There is total 506 rows and 10 columns in the dataset. Out of 10, All the 10 are of either integer or float data type.

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
Mean	4.871976285	68.57490119	11.13677866	0.554695059	9.549407115	408.2371542	18.4555336	6.284634387	12.65306324	22.53280632
Standard Error	0.129860152	1.251369525	0.304979888	0.005151391	0.387084894	7.492388692	0.096243568	0.031235142	0.317458906	0.408861147
Median	4.82	77.5	9.69	0.538	5	330	19.05	6.2085	11.36	21.2
Mode	3.43	100	18.1	0.538	24	666	20.2	5.713	8.05	50
Standard Deviation	2.921131892	28.14886141	6.860352941	0.115877676	8.707259384	168.5371161	2.164945524	0.702617143	7.141061511	9.197104087
Sample Variance	8.533011532	792.3583985	47.06444247	0.013427636	75.81636598	28404.75949	4.686989121	0.49367085	50.99475951	84.58672359
Kurtosis	-1.189122464	-0.96771559	-1.2335396	-0.06466713	-0.86723199	-1.14240799	-0.28509138	1.891500366	0.493239517	1.495196944
Skewness	0.021728079	-0.59896264	0.295021568	0.729307923	1.004814648	0.669955942	-0.80232493	0.403612133	0.906460094	1.108098408
Range	9.95	97.1	27.28	0.486	23	524	9.4	5.219	36.24	45
Minimum	0.04	2.9	0.46	0.385	1	187	12.6	3.561	1.73	5
Maximum	9.99	100	27.74	0.871	24	711	22	8.78	37.97	50
Sum	2465.22	34698.9	5635.21	280.6757	4832	206568	9338.5	3180.025	6402.45	11401.6
Count	506	506	506	506	506	506	506	506	506	506

Table: Summary of the data

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

Inference:

From the above descriptive statistics, helps to describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. The most recognized types of descriptive statistics are measures of Centre: the mean, median, and mode, which are used at almost all levels of math and statistics.

- In the above descriptive statistics, we can see that there are CRIME_RATE has a wide range from 0.04 to 9.99 with a mean around 4.87, indicating variability in crime rates.
- AGE has a relatively high standard deviation so indicating that a wider distribution of age of properties.
- The average price of houses in Boston is $22.53
- Approximately 69% of the houses within a 10-kilometer radius were built before 1910.
- At a mean house value of $23, the entire property tax rate comes to 408 units.

Q2: Plot a histogram of the Avg_Price variable. What do you infer?

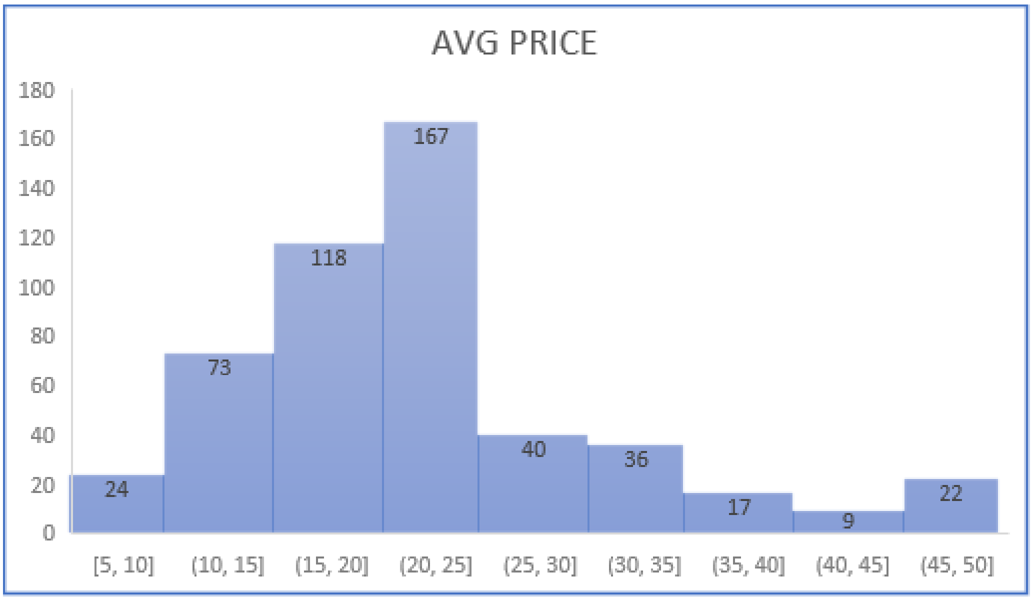


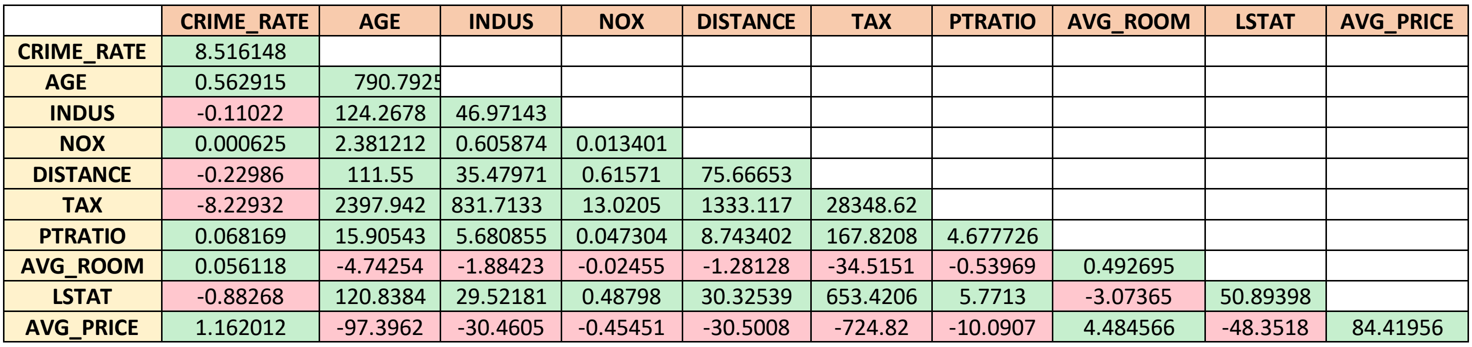
Fig.1 – Histogram of Avg Price

Inference:

After calculating the average price of the houses in Boston

- The maximum range of Average price between $20 to $25.
- The price above $40 are premium and luxuries houses in Boston.
- The price below $20 are affordable houses.

Q3: Compute the covariance matrix. Share your observations?



	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.516148									
AGE	0.562915	790.7925								
INDUS	-0.11022	124.2678	46.97143							
NOX	0.000625	2.381212	0.605874	0.013401						
DISTANCE	-0.22986	111.55	35.47971	0.61571	75.66653					
TAX	-8.22932	2397.942	831.7133	13.0205	1333.117	28348.62				
PTRATIO	0.068169	15.90543	5.680855	0.047304	8.743402	167.8208	4.677726			
AVG_ROOM	0.056118	-4.74254	-1.88423	-0.02455	-1.28128	-34.5151	-0.53969	0.492695		
LSTAT	-0.88268	120.8384	29.52181	0.48798	30.32539	653.4206	5.7713	-3.07365	50.89398	
AVG_PRICE	1.162012	-97.3962	-30.4605	-0.45451	-30.5008	-724.82	-10.0907	4.484566	-48.3518	84.41956

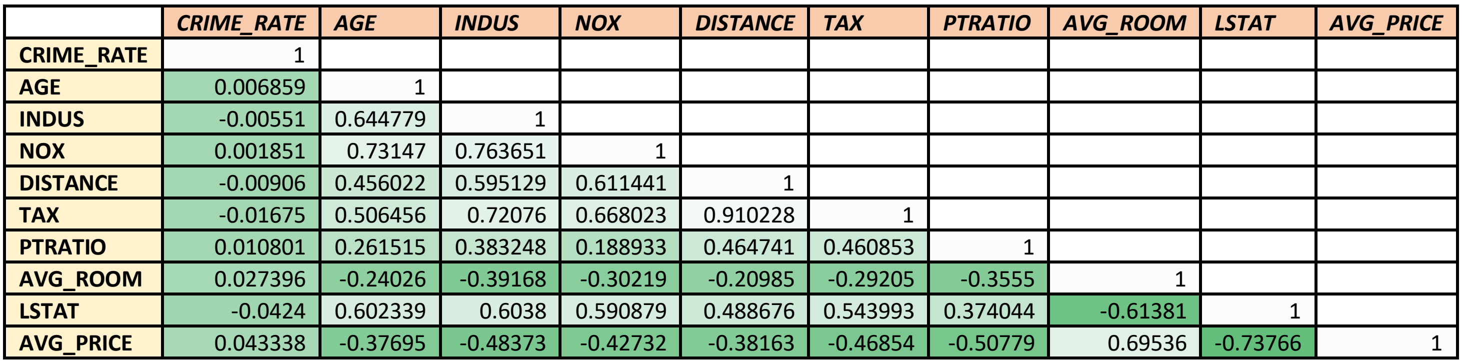
Inference:

- The largest covariances are between ("TAX" and "AGE") ,("AVG_ROOM" and "TAX") and (“DISTANCE” and “TAX”) These variables show relatively strong relationships

- The covariance between "CRIME_RATE" and "AGE" is approximately 0.563, indicating a positive covariance.

- Negative covariance is (AVG_PRICE, AGE), (AVG_PRICE, TAX), (AVG_LSTAT)

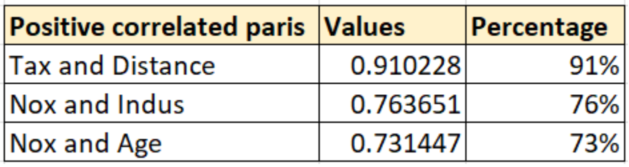
Q4: Create a correlation matrix of all the variables



	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: Correlation matrix

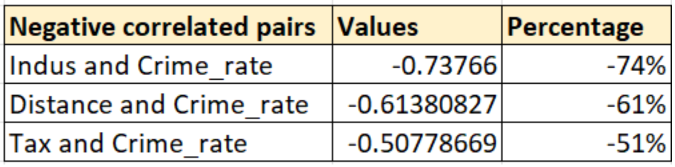
4.a) Which are the top 3 positively correlated pairs?



Positive correlated paris	Values	Percentage
Tax and Distance	0.910228	91%
Nox and Indus	0.763651	76%
Nox and Age	0.731447	73%

- "AVG_ROOM" and "AVG_PRICE" have the highest positive correlation with a coefficient of approximately 0.6954. If average room increases relatively house price increases.
- "NOX" and "INDUS" have the highest positive correlation with a coefficient of approximately 0.7637. The industrial land use increases, nitrogen oxide concentrations tend to increase.

4.b) Which are the top 3 negatively correlated pairs?

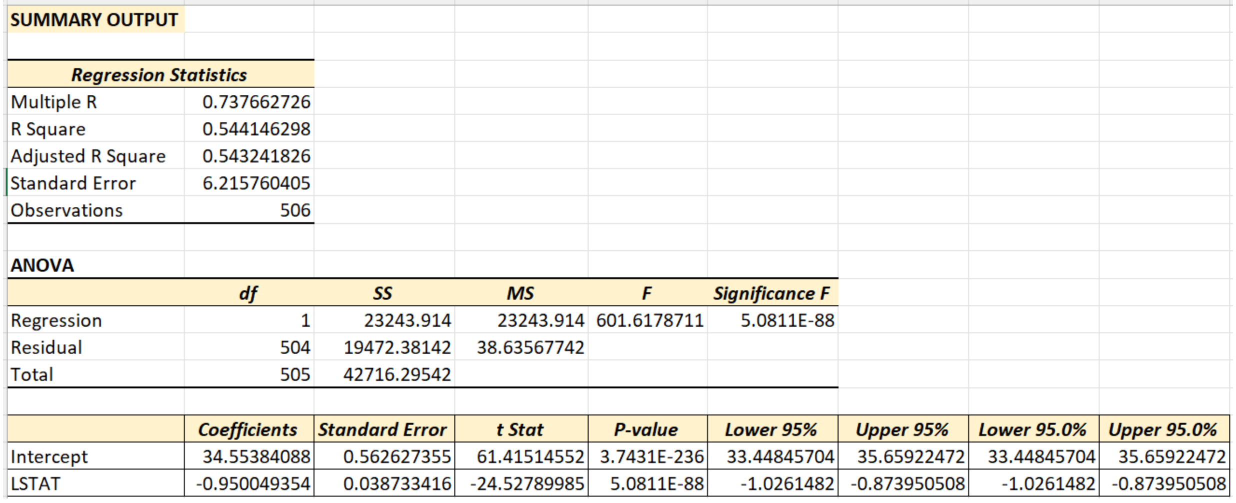


Negative correlated pairs	Values	Percentage
Indus and Crime_rate	-0.73766	-74%
Distance and Crime_rate	-0.61380827	-61%
Tax and Crime_rate	-0.50778669	-51%

- "AVG_ROOM" and "LSTAT" have the lowest negative correlation with a coefficient of approximately -0.6138.
- The average number of rooms per dwelling and the percentage of lower-income residents.
- "CRIME_RATE" and "AGE" have the lowest positive correlation with a coefficient of approximately 0.0069

- "DISTANCE" and "NOX" have the lowest positive correlation with a coefficient of approximately 0.4560.

Q5: Build an initial regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot:



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

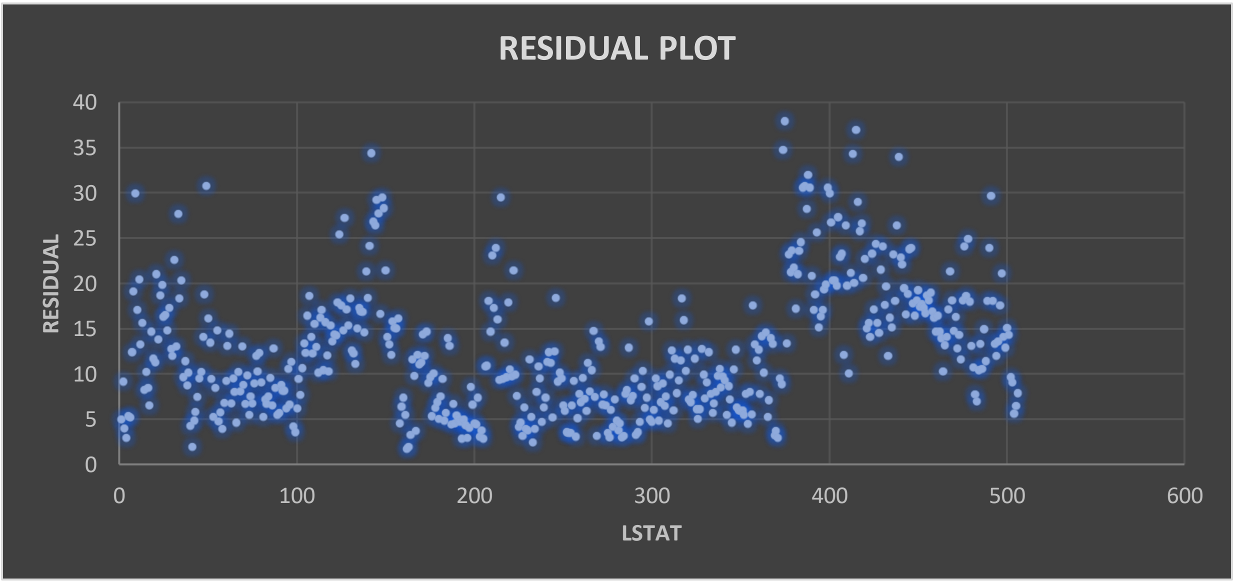


Fig 2 : Residual Plot

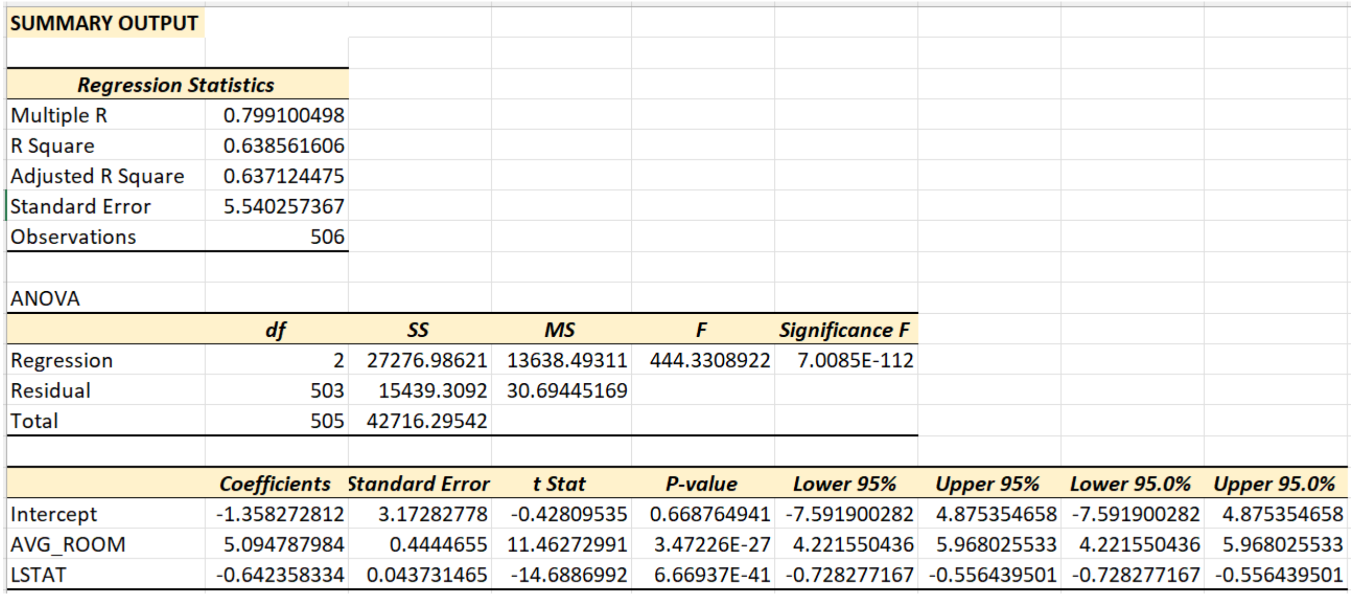
A) What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?

- Intercept 34.55384088
- Coefficient value -0.950049354
- The graph looks as scattered in plot of residual

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

- LSTAT value is insignificant, cause the adjusted R-value is seems low.

Q6: Build a new Regression model including LSTAT and AVG_ROOM together as independent variables and AVG_PRICE as dependent variable?



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	-0.42809535	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	-14.6886992	6.66937E-41	-0.728277167	-0.556439501	-0.728277167	-0.556439501

Lstat and avg_pri
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?

<u>REGRESSION EQUATION</u>

Y= (AVG ROOM *7 +(LSTAT*20)+INTERCEPTY

=(5.0947*7)+(-0.64236*20)+(-1.3582)

=21.45808 OF PREDICTED AVG PRICE

As the company is quoting for a value of 30000 USD for this locality by regression equationweget to know that the company is overcharging.

a. 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

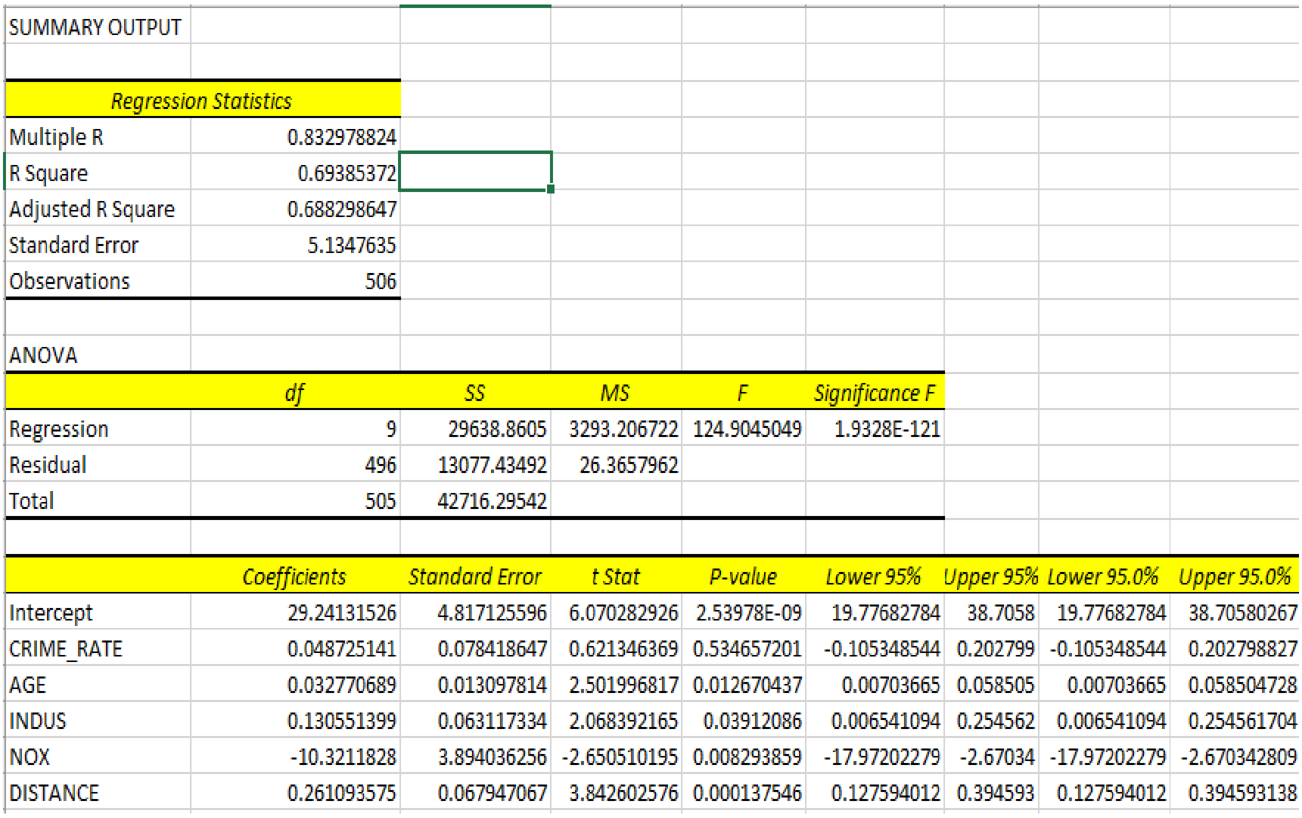
Yes, this model is better than the previous model because in the previous model the adjusted R-square is 54% and this model the adjusted R-square value is 0.63 which is independent variablethat explain 63% of the variation in the dependent variable Ideally this model performance well compare to this 5question.model.

- Adjusted R **0.637124** > Adjusted R **0.543242**

  Adjusted R value is giving better result then previous question.

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 Rsquare, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE The difference in means between two Normal distributions with unknown variance follows a student's t-distribution. The t-test is any statistical hypothesis test in which the test statistic follows a student's t-distribution under the null hypothesis. The student t-test is one of the oldest and widely used hypothesis test.



SUMMARY OUTPUT

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

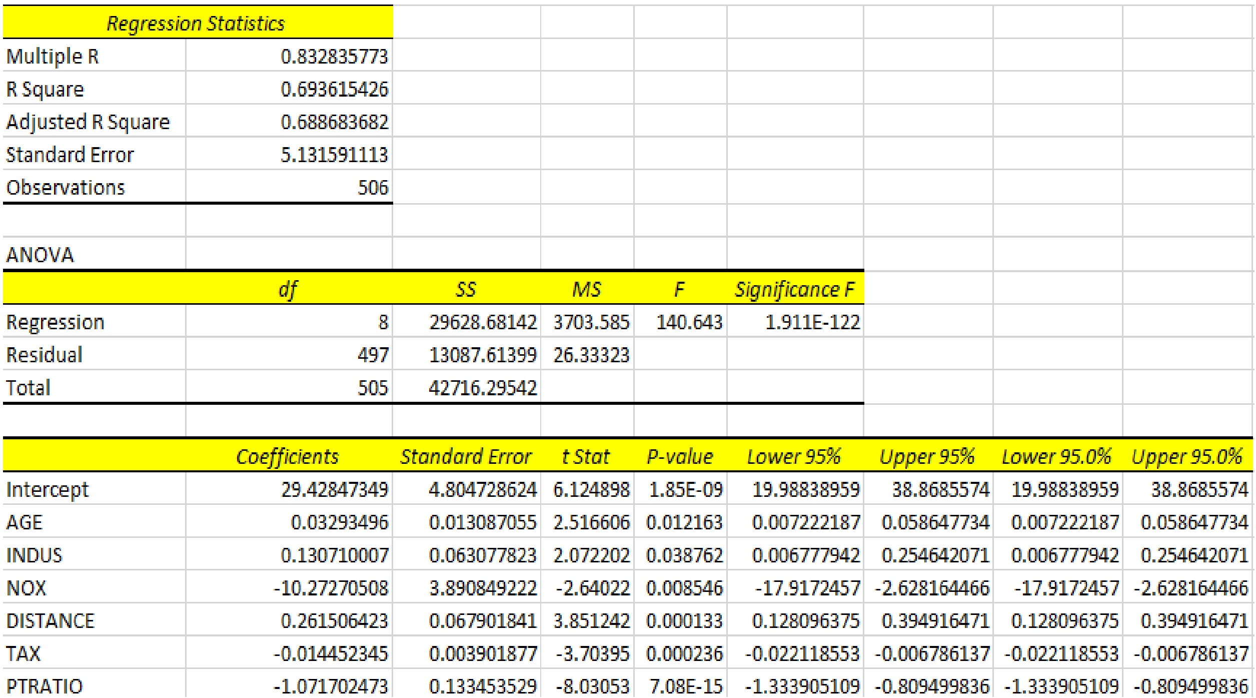
ANOVA

	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	38.7058	19.77682784	38.70580267
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.534657201	-0.105348544	0.202799	-0.105348544	0.202798827
AGE	0.032770689	0.013097814	2.501996817	0.012670437	0.00703665	0.058505	0.00703665	0.058504728
INDUS	0.130551399	0.063117334	2.068392165	0.03912086	0.006541094	0.254562	0.006541094	0.254561704
NOX	-10.3211828	3.894036256	-2.650510195	0.008293859	-17.97202279	-2.67034	-17.97202279	-2.670342809
DISTANCE	0.261093575	0.067947067	3.842602576	0.000137546	0.127594012	0.394593	0.127594012	0.394593138

- **Adjusted R Square:**
  As 0.688298647, The remarkable modified R-square value confirms that this model is appropriate for jobs involving prediction. It can be trusted to produce precise forecasts because of its capacity to efficiently account for data volatility. Consequently, this model is a promising contender for real-world use in a range of prediction scenarios.
- **Significant Variables:**
  AGE, INDUS, NOX, DISTANCE, LSTAT, PTRATIO, AVGROOM, and TAX are the factors that have shown statistical significance.
- **Insignificant Variable:**
  In contrast, there are no statistically significant correlations found in the model for the variable CRIME RATE.
- **High Coefficient for AVG ROOM:**
  AVG ROOM has a coefficient that is higher than all other variables', indicating that it has a dominant influence in the model.

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:



Regression Statistics	
Multiple R	0.832835773
R Square	0.693615426
Adjusted R Square	0.688683682
Standard Error	5.131591113
Observations	506

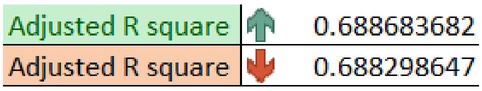
ANOVA

	df	SS	MS	F	Significance F
Regression	8	29628.68142	3703.585	140.643	1.911E-122
Residual	497	13087.61399	26.33323		
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	6.124898	1.85E-09	19.98838959	38.8685574	19.98838959	38.8685574
AGE	0.03293496	0.013087055	2.516606	0.012163	0.007222187	0.058647734	0.007222187	0.058647734
INDUS	0.130710007	0.063077823	2.072202	0.038762	0.006777942	0.254642071	0.006777942	0.254642071
NOX	-10.27270508	3.890849222	-2.64022	0.008546	-17.9172457	-2.628164466	-17.9172457	-2.628164466
DISTANCE	0.261506423	0.067901841	3.851242	0.000133	0.128096375	0.394916471	0.128096375	0.394916471
TAX	-0.014452345	0.003901877	-3.70395	0.000236	-0.022118553	-0.006786137	-0.022118553	-0.006786137
PTRATIO	-1.071702473	0.133453529	-8.03053	7.08E-15	-1.333905109	-0.809499836	-1.333905109	-0.809499836

a. Interpret the output of this model. We can utilize the 68% R-value to make predictions.

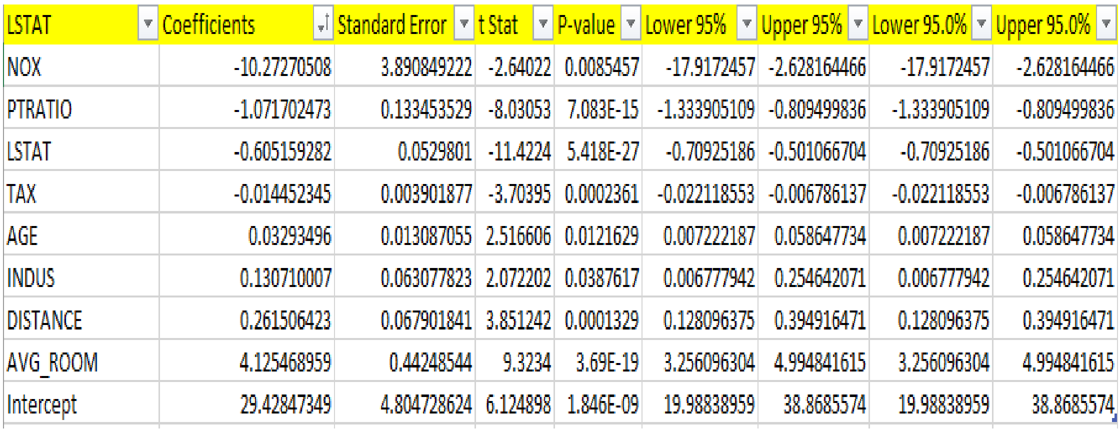
b. Compare the adjusted R-square value of this model with the model in the previous question, which model performs better according to the value of adjusted R-square?



Adjusted R square	↑	0.688683682
Adjusted R square	↓	0.688298647

As we can infer from this table that the adjusted R square from the previous table negative And the for this model the adjusted R Square is positive.

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?



LSTAT	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
NOX	-10.27270508	3.890849222	-2.64022	0.0085457	-17.9172457	-2.628164466	-17.9172457	-2.628164466
PTRATIO	-1.071702473	0.133453529	-8.03053	7.083E-15	-1.333905109	-0.809499836	-1.333905109	-0.809499836
LSTAT	-0.605159282	0.0529801	-11.4224	5.418E-27	-0.70925186	-0.501066704	-0.70925186	-0.501066704
TAX	-0.014452345	0.003901877	-3.70395	0.0002361	-0.022118553	-0.006786137	-0.022118553	-0.006786137
AGE	0.03293496	0.013087055	2.516606	0.0121629	0.007222187	0.058647734	0.007222187	0.058647734
INDUS	0.130710007	0.063077823	2.072202	0.0387617	0.006777942	0.254642071	0.006777942	0.254642071
DISTANCE	0.261506423	0.067901841	3.851242	0.0001329	0.128096375	0.394916471	0.128096375	0.394916471
AVG_ROOM	4.125468959	0.44248544	9.3234	3.69E-19	3.256096304	4.994841615	3.256096304	4.994841615
Intercept	29.42847349	4.804728624	6.124898	1.846E-09	19.98838959	38.8685574	19.98838959	38.8685574

Intercept 29.42847349

The correlation between NOX and AVG_PRICE is -0.42732, indicating an inverse relationship between the two variables. When NOX levels increase, AVG_PRICE tends to decrease. In other words, higher NOX values are associated with lower average prices.

d. Write the regression equation from this model.

Avg_price= coeffecient(age)*age)+ (coefficient(indus)*indus)+(coeffecient(nox)*nox)+ (coeffcient(distance)*distance)+ (coeffecient(tax)*tax)+ (coeffecient(ptratio)*ptratio)+ (coeffecient(avg room)*avg_room)+(coeffecient(lstat)*lstat)+ intercept

Conclusion & Recommendation -

Several independent variables, including AGE, INDUS, DISTANCE, TAX, PTRATIO, AVG_ROOM, and LSTAT, are statistically significant in predicting the dependent variable. This means that changes in these variables have a significant impact on the predicted outcome.

THE END!