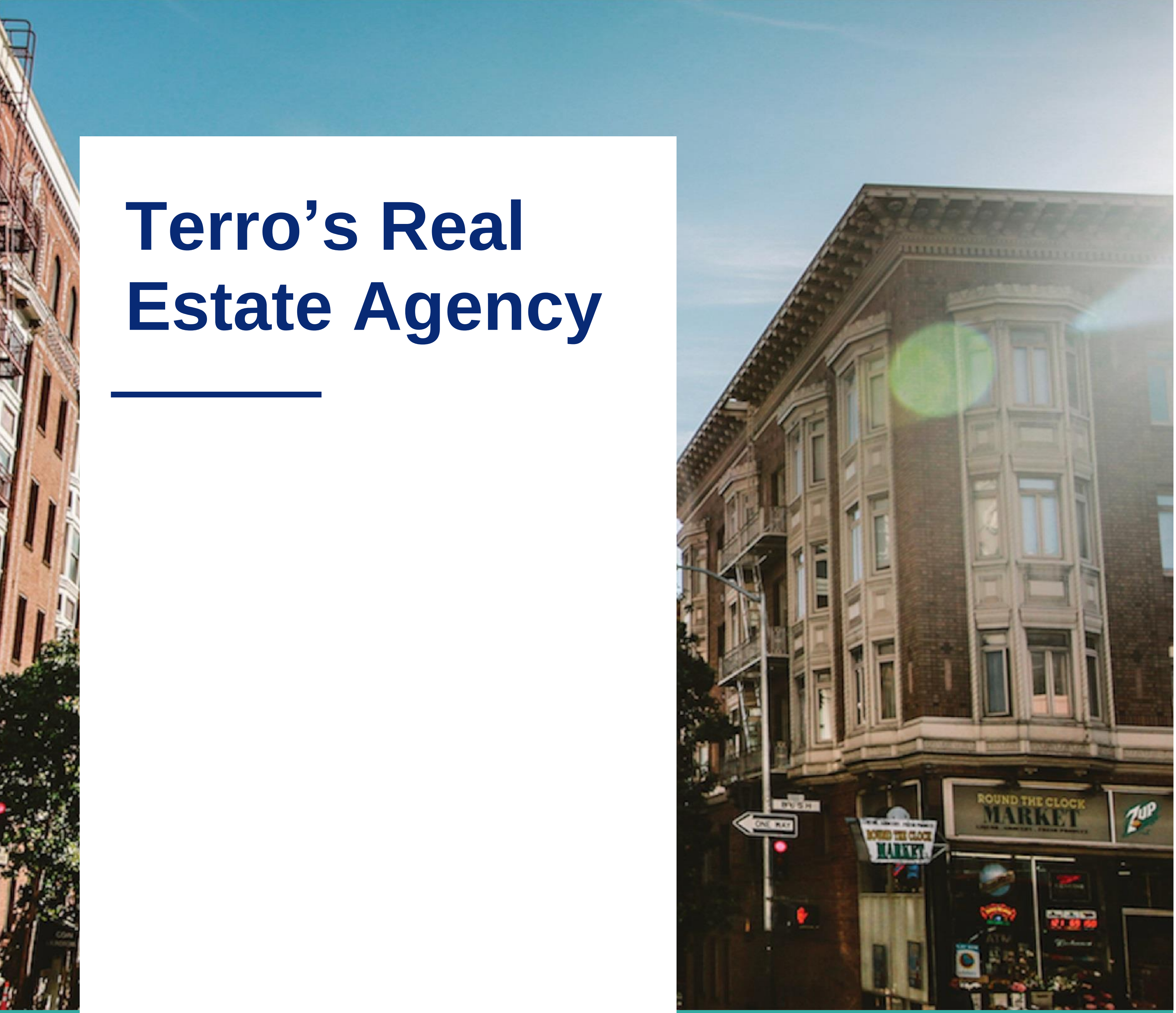


Terro's Real Estate Agency

SEPTEMBER,2023

BUSINESS REPORT

Authored by: EKTA SINGH

Project Description:

TOPICS COVERED: (Descriptive Statistics, Covariance, Correlations, Simple Linear Regression, Multiple Linear Regression)

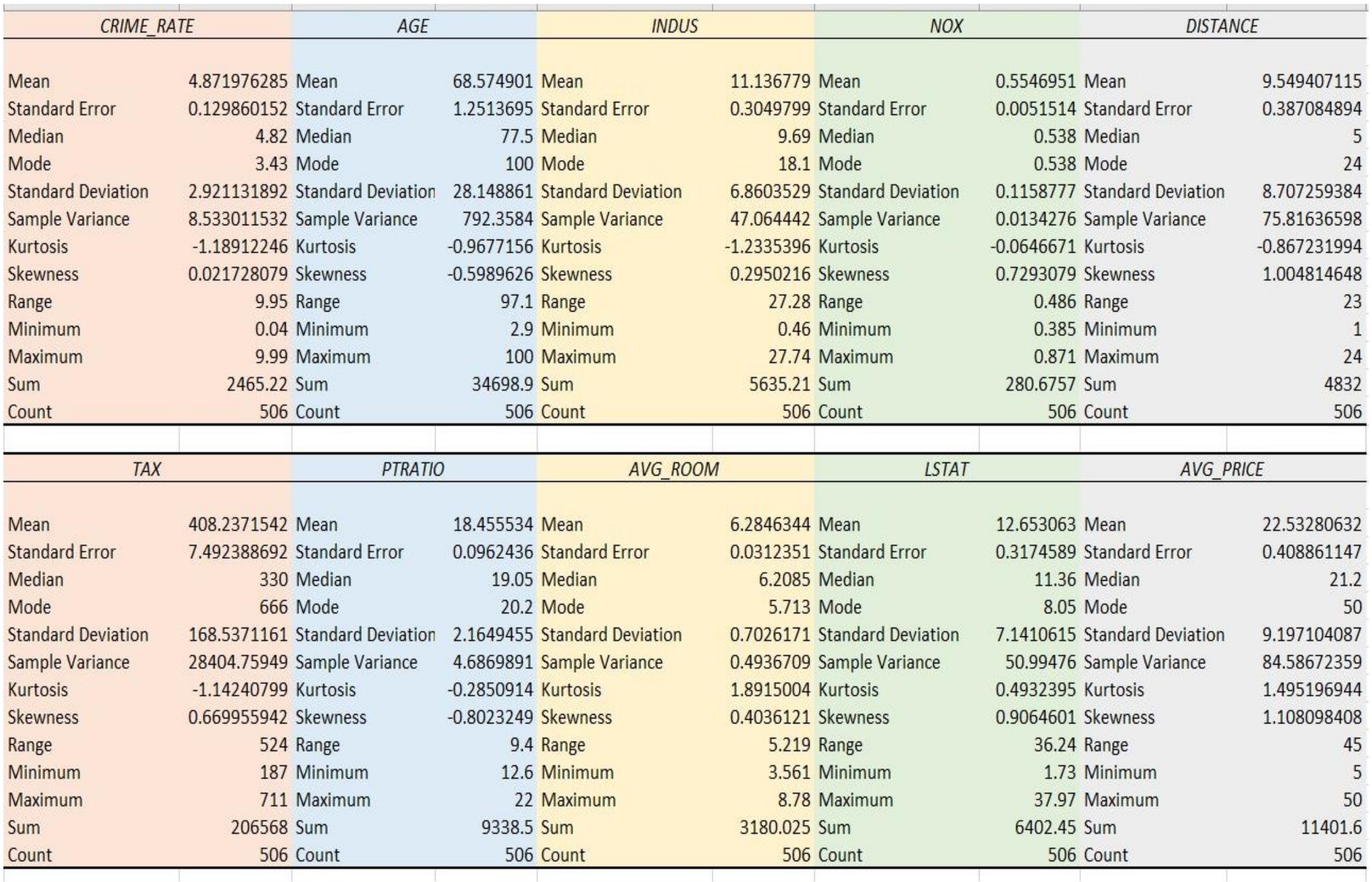
Terro's real-estate is an agency that estimates the pricing of houses in a certain locality. The pricing is concluded based on different features / factors of a property. This also helps them in identifying the business value of a property. To do this activity the company employs an "Auditor", who studies 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. The agency has provided a dataset of 506 houses in Boston. Following are the details of the dataset:

DATA DICTIONARY:

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

1). Generate the summary statistics for each variable in the table. (Use Data analysis toolpak). Write down your observations.

Summary statistics for each of the variables-



CRIME_RATE		AGE		INDUS		NOX		DISTANCE	
Mean	4.871976285	Mean	68.574901	Mean	11.136779	Mean	0.5546951	Mean	9.549407115
Standard Error	0.129860152	Standard Error	1.2513695	Standard Error	0.3049799	Standard Error	0.0051514	Standard Error	0.387084894
Median	4.82	Median	77.5	Median	9.69	Median	0.538	Median	5
Mode	3.43	Mode	100	Mode	18.1	Mode	0.538	Mode	24
Standard Deviation	2.921131892	Standard Deviation	28.148861	Standard Deviation	6.8603529	Standard Deviation	0.1158777	Standard Deviation	8.707259384
Sample Variance	8.533011532	Sample Variance	792.3584	Sample Variance	47.064442	Sample Variance	0.0134276	Sample Variance	75.81636598
Kurtosis	-1.18912246	Kurtosis	-0.9677156	Kurtosis	-1.2335396	Kurtosis	-0.0646671	Kurtosis	-0.867231994
Skewness	0.021728079	Skewness	-0.5989626	Skewness	0.2950216	Skewness	0.7293079	Skewness	1.004814648
Range	9.95	Range	97.1	Range	27.28	Range	0.486	Range	23
Minimum	0.04	Minimum	2.9	Minimum	0.46	Minimum	0.385	Minimum	1
Maximum	9.99	Maximum	100	Maximum	27.74	Maximum	0.871	Maximum	24
Sum	2465.22	Sum	34698.9	Sum	5635.21	Sum	280.6757	Sum	4832
Count	506	Count	506	Count	506	Count	506	Count	506

TAX		PTRATIO		AVG_ROOM		LSTAT		AVG_PRICE	
Mean	408.2371542	Mean	18.455534	Mean	6.2846344	Mean	12.653063	Mean	22.53280632
Standard Error	7.492388692	Standard Error	0.0962436	Standard Error	0.0312351	Standard Error	0.3174589	Standard Error	0.408861147
Median	330	Median	19.05	Median	6.2085	Median	11.36	Median	21.2
Mode	666	Mode	20.2	Mode	5.713	Mode	8.05	Mode	50
Standard Deviation	168.5371161	Standard Deviation	2.1649455	Standard Deviation	0.7026171	Standard Deviation	7.1410615	Standard Deviation	9.197104087
Sample Variance	28404.75949	Sample Variance	4.6869891	Sample Variance	0.4936709	Sample Variance	50.99476	Sample Variance	84.58672359
Kurtosis	-1.14240799	Kurtosis	-0.2850914	Kurtosis	1.8915004	Kurtosis	0.4932395	Kurtosis	1.495196944
Skewness	0.669955942	Skewness	-0.8023249	Skewness	0.4036121	Skewness	0.9064601	Skewness	1.108098408
Range	524	Range	9.4	Range	5.219	Range	36.24	Range	45
Minimum	187	Minimum	12.6	Minimum	3.561	Minimum	1.73	Minimum	5
Maximum	711	Maximum	22	Maximum	8.78	Maximum	37.97	Maximum	50
Sum	206568	Sum	9338.5	Sum	3180.025	Sum	6402.45	Sum	11401.6
Count	506	Count	506	Count	506	Count	506	Count	506

Observations-

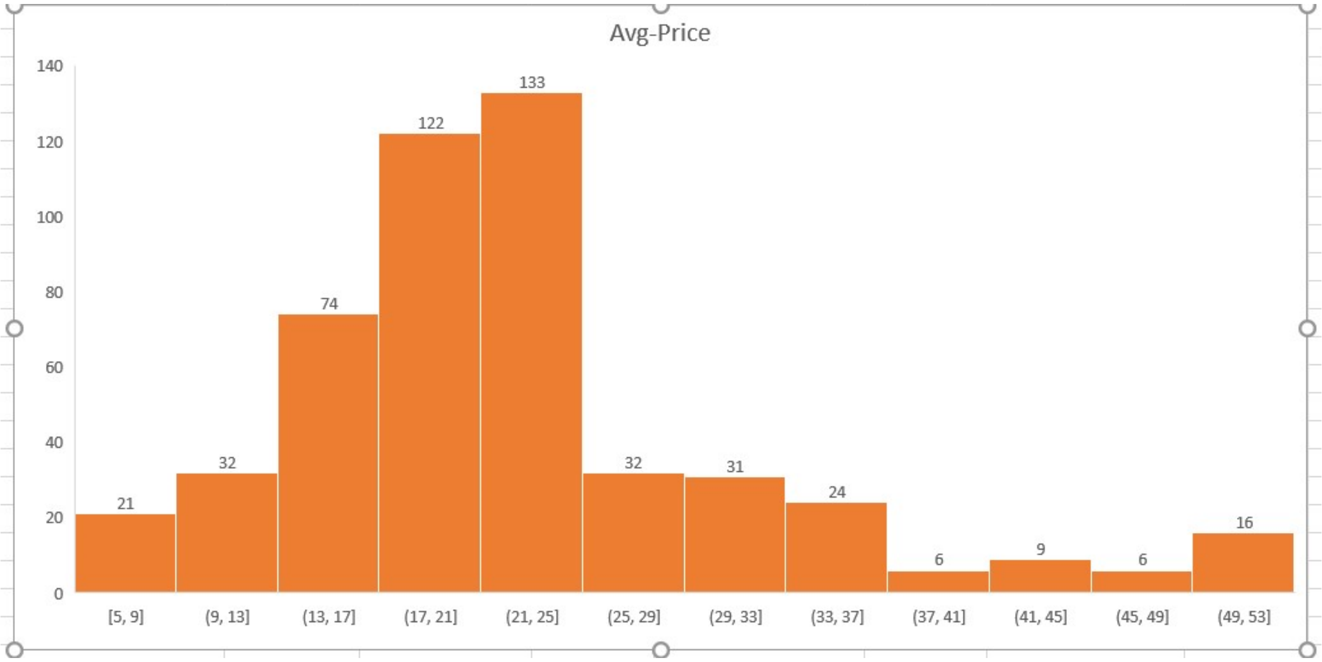
- > Based on Measures of Symmetry, we can say that 'AVG_ROOM' has the sharpest peak as it has the highest kurtosis

-> 'AVG_PRICE' is the most positively skewed variable.

-> Based on Measures of variability, it can be inferred that the Standard deviation for 'TAX' variable is the highest, indicating that its data is more spread out.

-> Based on minimum and maximum values, we can say that a lot of outliers are present in 'TAX' and 'AGE' variables.

2) Plot a histogram of the Avg_Price variable. What do you infer?



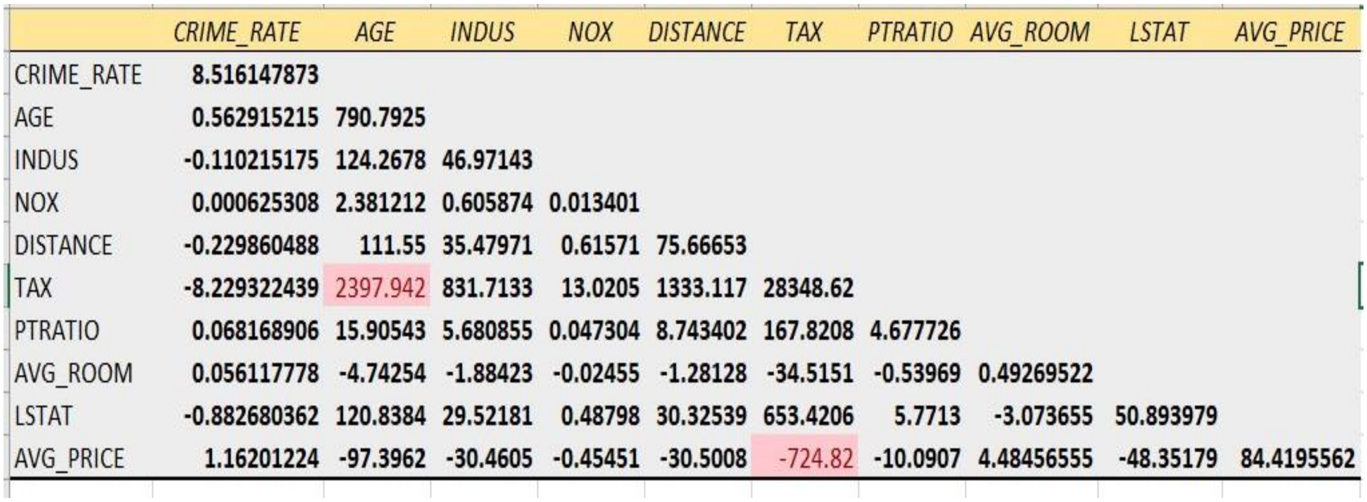
OBSERVATIONS-

-> Based on the shape of distribution of data, we can say that the AVG_PRICE variable has a positive skew meaning most of the values occur before the mean.

-> Since, most of data points falls on the left side of the mean then it is called Right Skewed data or Positive Skewed data.

-> The general relationship among the central tendency measures in a positively skewed distribution may be expressed using the following in equality:-
Mean > Median > Mode

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	0.49269522		
LSTAT	-0.882680362	120.8384	29.52181	0.48798	30.32539	653.4206	5.7713	-3.073655	50.893979	
AVG_PRICE	1.16201224	-97.3962	-30.4605	-0.45451	-30.5008	-724.82	-10.0907	4.48456555	-48.35179	84.4195562

OBSERVATIONS-

From the above covariance matrix, we can infer that the variables:-

-> TAX and AGE have the highest covariance, which means as the age of the house proportion built prior to 1940 increases, the tax also increases.

-> TAX and DISTANCE have the second highest covariance.

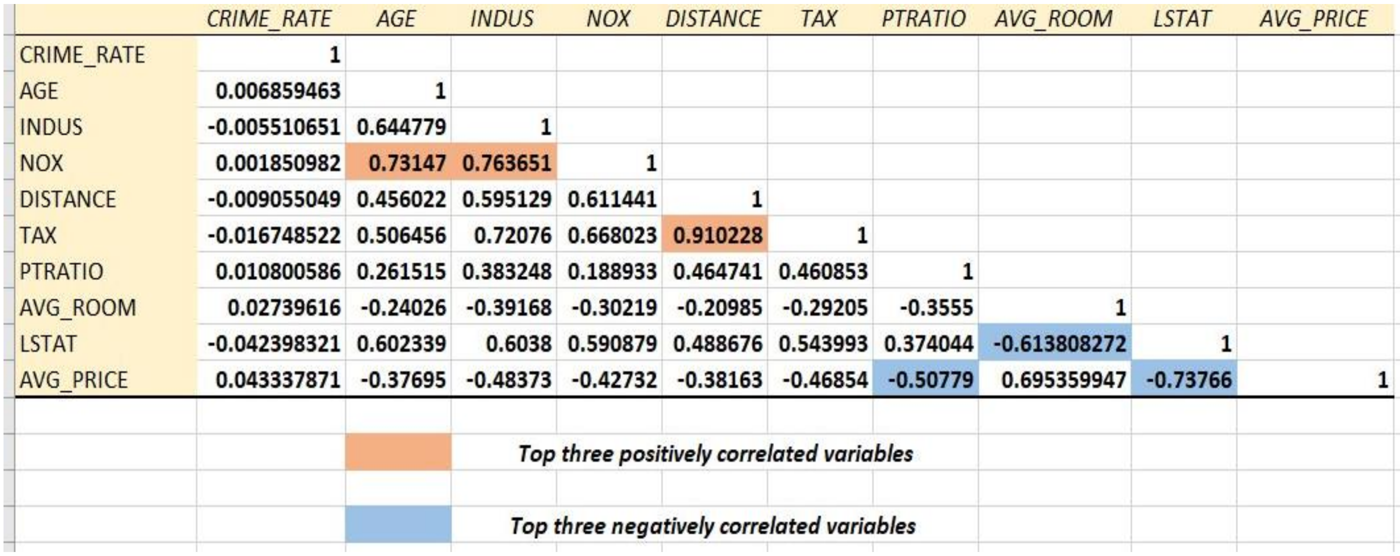
->TAX and AVG_PRICE have the least covariance

-> TAX and CRIME_RATE have high negative covariance value, which means the houses which pays more tax have less crime rate.

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

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

b). Which are the 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	0.910228	1				
PTRATIO	0.010800586	0.261515	0.383248	0.188933	0.464741	0.460853	1			
AVG_ROOM	0.02739616	-0.24026	-0.39168	-0.30219	-0.20985	-0.29205	-0.3555	1		
LSTAT	-0.042398321	0.602339	0.6038	0.590879	0.488676	0.543993	0.374044	-0.613808272	1	
AVG_PRICE	0.043337871	-0.37695	-0.48373	-0.42732	-0.38163	-0.46854	-0.50779	0.695359947	-0.73766	1

🟧 *Top three positively correlated variables*

🟦 *Top three negatively correlated variables*

OBSERVATIONS-

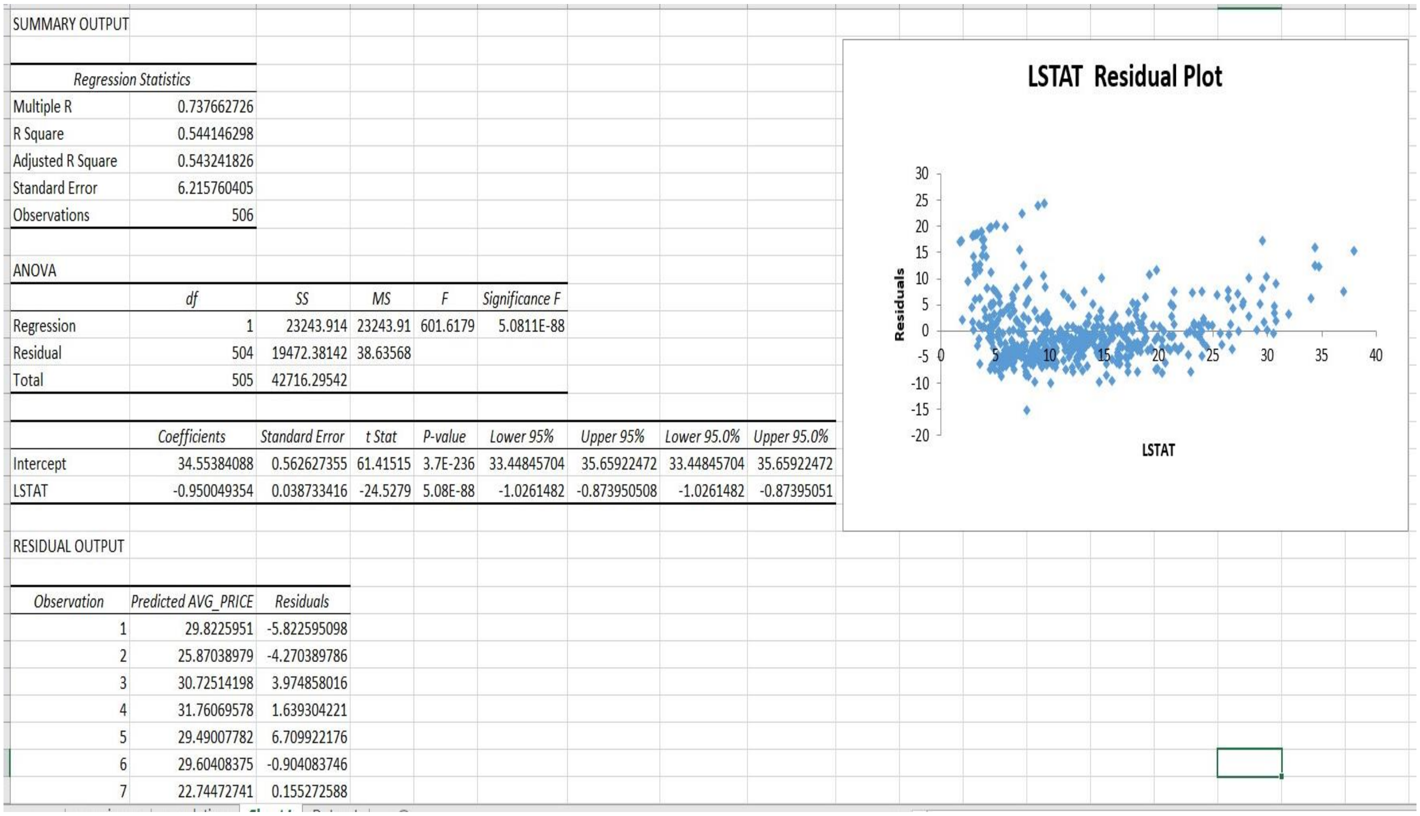
-> Top 3 positively correlated pairs are -

TAX & DISTANCE, NOX & INDUS, NOX & AGE.

-> Top 3 negatively correlated pairs are -

AVG_PRICE& LSTAT, AVG_ROOM& LSTAT, AVG_PRICE& PTRATIO.

5) 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	23243.91	601.6179	5.0811E-88
Residual	504	19472.38142	38.63568		
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	61.41515	3.7E-236	33.44845704	35.65922472	33.44845704	35.65922472
LSTAT	-0.950049354	0.038733416	-24.5279	5.08E-88	-1.0261482	-0.873950508	-1.0261482	-0.87395051

RESIDUAL OUTPUT

Observation	*Predicted AVG_PRICE*	*Residuals*
1	29.8225951	-5.822595098
2	25.87038979	-4.270389786
3	30.72514198	3.974858016
4	31.76069578	1.639304221
5	29.49007782	6.709922176
6	29.60408375	-0.904083746
7	22.74472741	0.155272588

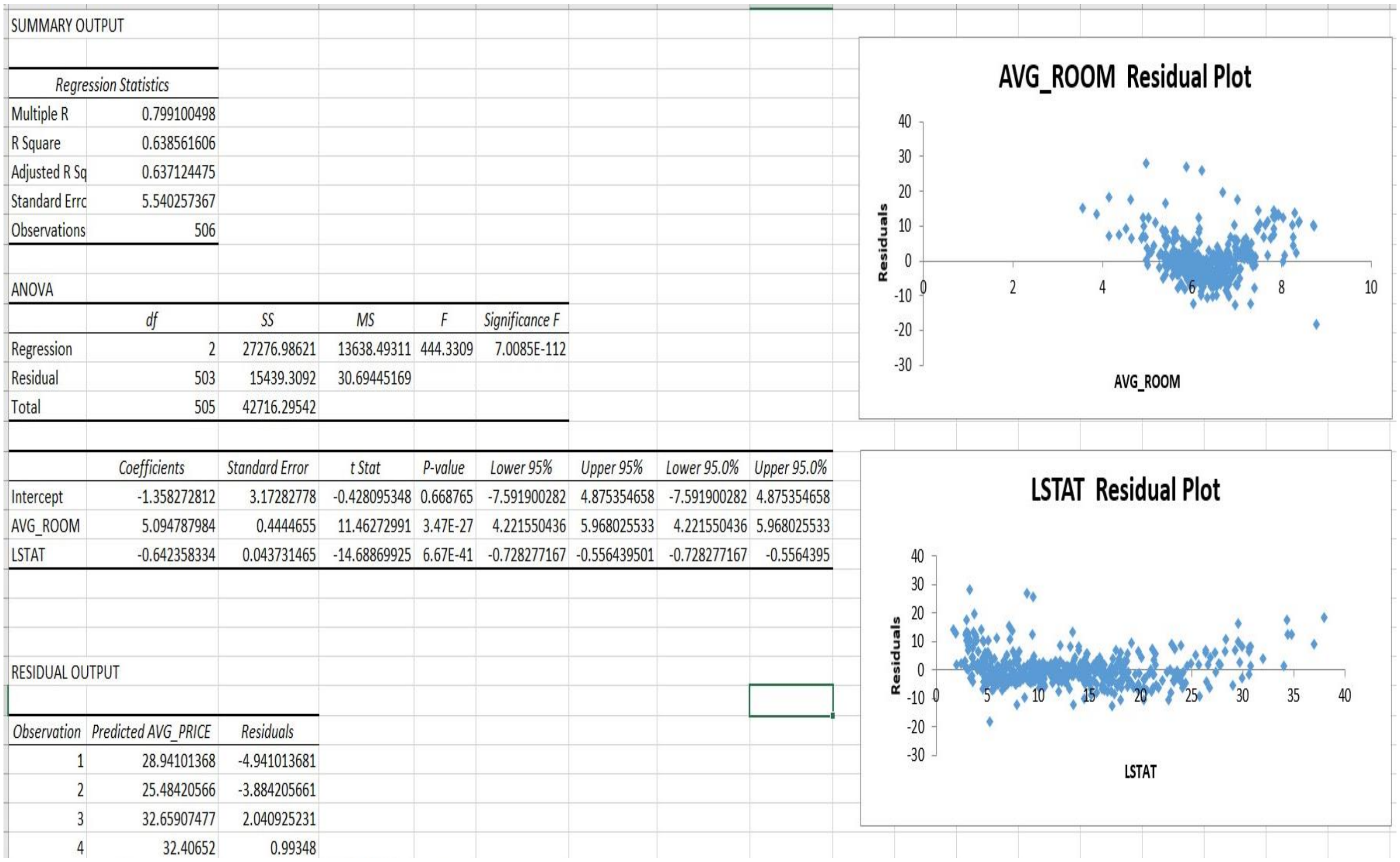
a). What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?

Regression Summary output provides information on how well the model describes the data and the relationships between the independent and dependent variables.
-> Since the R square value is low, the model does not explain the variation in price very well.
-> A negative value for the coefficient of LSTAT variable represents that the price goes down as LSTAT goes up.
-> The intercept represents the estimated value of the dependent variable when all independent variables are set to zero.
-> The residual plot has no patterns, representing no issues with the regression model. In a well-fitted model, residuals are randomly scattered around zero without any discernible pattern.

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

-> If p ≤ significance level (e.g., 0.05), the coefficient is statistically significant.
-> If p > significance level, the coefficient is not statistically significant.
P-value for LSTAT variable is less than 0.05, so it is considered as a significant variable.

6). 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 Sq	0.637124475
Standard Errc	5.540257367
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	27276.98621	13638.49311	444.3309	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	0.668765	-7.591900282	4.875354658	-7.591900282	4.875354658
AVG_ROOM	5.094787984	0.4444655	11.46272991	3.47E-27	4.221550436	5.968025533	4.221550436	5.968025533
LSTAT	-0.642358334	0.043731465	-14.68869925	6.67E-41	-0.728277167	-0.556439501	-0.728277167	-0.5564395

RESIDUAL OUTPUT

Observation	*Predicted AVG_PRICE*	*Residuals*
1	28.94101368	-4.941013681
2	25.48420566	-3.884205661
3	32.65907477	2.040925231
4	32.40652	0.99348

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-> -1.3582 + AVG_ROOM*5.0947 - LSTAT*0.6423

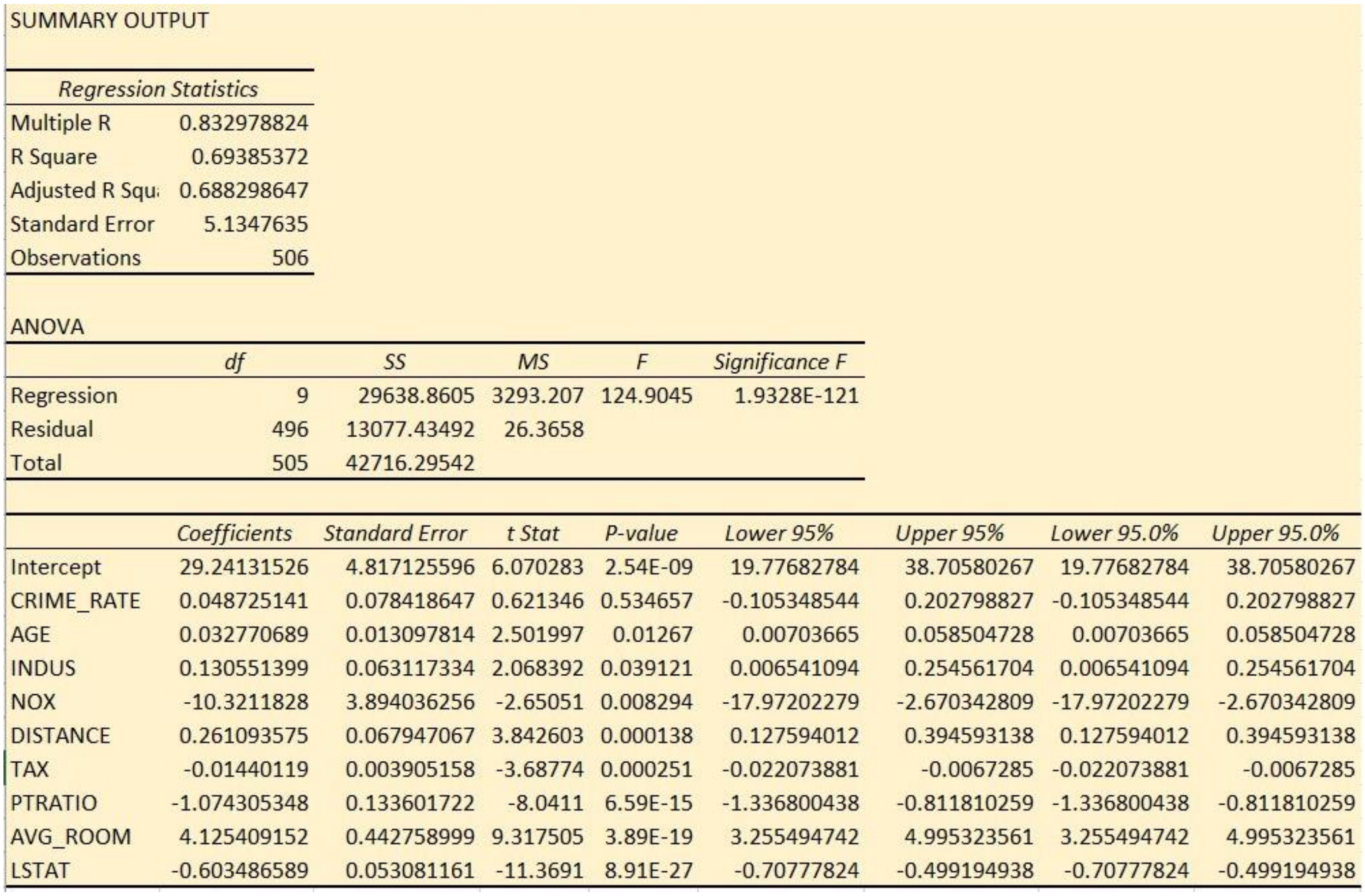
Avg_price = -1.3582 + 7*5.0947 - 20*0.6423 = 21.4k USD

Predicted price is 21.4k USD and the company is quoting 30k USD. Thus, they are 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-squared (Adj. R^2): This is a modified version of R-squared that adjusts for the number of predictors in the model. It penalizes the inclusion of unnecessary variables in the model. A higher adjusted R-squared suggests that the model is a better fit, especially if you're comparing models with different numbers of predictors.
Since the adjusted R square value is higher than the previous model, this model is better at explaining the dependent variable than the previous model (5th question).

7) Build another Regression model with all variables where AVG_PRICE alone be the Dependent Variable and all the other variables are independent. Interpret the output in terms of adjusted R-square, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE.



SUMMARY OUTPUT

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

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	29638.8605	3293.207	124.9045	1.9328E-121
Residual	496	13077.43492	26.3658		
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	6.070283	2.54E-09	19.77682784	38.70580267	19.77682784	38.70580267
CRIME_RATE	0.048725141	0.078418647	0.621346	0.534657	-0.105348544	0.202798827	-0.105348544	0.202798827
AGE	0.032770689	0.013097814	2.501997	0.01267	0.00703665	0.058504728	0.00703665	0.058504728
INDUS	0.130551399	0.063117334	2.068392	0.039121	0.006541094	0.254561704	0.006541094	0.254561704
NOX	-10.3211828	3.894036256	-2.65051	0.008294	-17.97202279	-2.670342809	-17.97202279	-2.670342809
DISTANCE	0.261093575	0.067947067	3.842603	0.000138	0.127594012	0.394593138	0.127594012	0.394593138
TAX	-0.01440119	0.003905158	-3.68774	0.000251	-0.022073881	-0.0067285	-0.022073881	-0.0067285
PTRATIO	-1.074305348	0.133601722	-8.0411	6.59E-15	-1.336800438	-0.811810259	-1.336800438	-0.811810259
AVG_ROOM	4.125409152	0.442758999	9.317505	3.89E-19	3.255494742	4.995323561	3.255494742	4.995323561
LSTAT	-0.603486589	0.053081161	-11.3691	8.91E-27	-0.70777824	-0.499194938	-0.70777824	-0.499194938

-> R squared value is0.69 or 69%which indicates a proper fit model for the data.

-> Except for NOX, TAX, PTRATIO, LSTAT which have negative coefficients, indicating that increase in those variables results in a decrease in the average price.

-> All other variables have positive coefficients, which means they have linear relationship with the average price.

-> Crime rate is the only variable whose p-value is not less than 0.05. Therefore, all variables except for ‘crime rate’ are significant for the prediction of average price.

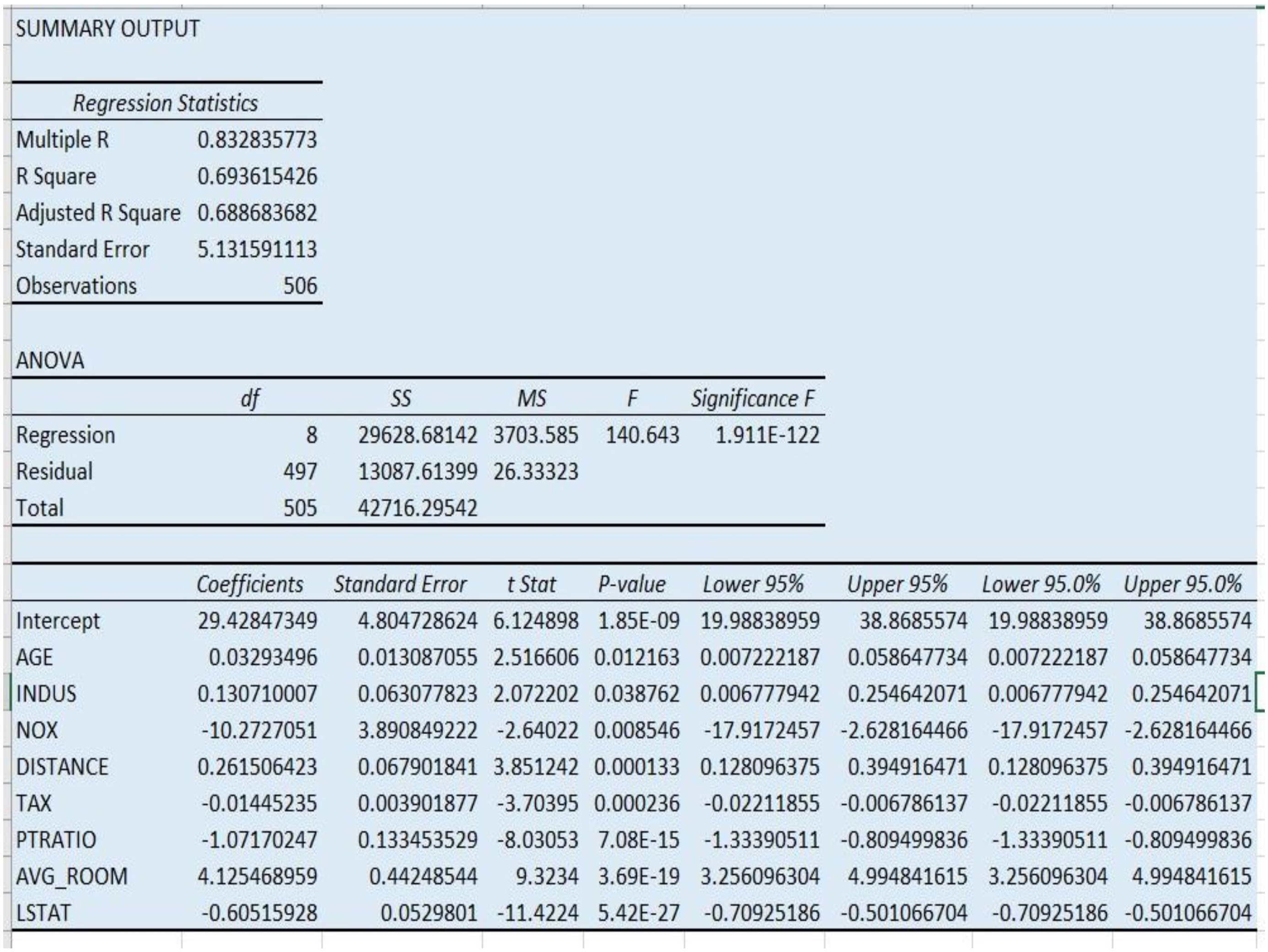
8). Pick out only the significant variables from the previous question. Make another instance of the Regression model using only the significant variables you just picked and answer the questions below:

a). Interpret the output of this model.

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?

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?

d). Write the regression equation from this model.



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	3703.585	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	-10.2727051	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	-0.01445235	0.003901877	-3.70395	0.000236	-0.02211855	-0.006786137	-0.02211855	-0.006786137
PTRATIO	-1.07170247	0.133453529	-8.03053	7.08E-15	-1.33390511	-0.809499836	-1.33390511	-0.809499836
AVG_ROOM	4.125468959	0.44248544	9.3234	3.69E-19	3.256096304	4.994841615	3.256096304	4.994841615
LSTAT	-0.60515928	0.0529801	-11.4224	5.42E-27	-0.70925186	-0.501066704	-0.70925186	-0.501066704

a). This model has an R squared value very similar to the previous model but an adjusted R square value that is slightly higher. All the p values are also less than 0.05 making all the variables significant.

b). The value of adjusted R in previous model is 0.6882 and in this model it is equal to 0.6886.Since this model has a slightly higher value of adjusted R, it explains the output variable better.

c). Values of the coefficients in the ascending order-

NOX, PTRATIO, LSTAT, TAX, AGE, INDUS, DISTANCE, AVG_ROOM

Since NOX variable has a negative coefficient, higher value of NOX leads to a decrease in price.

d). Regression Equation: 29.42 - 10.27*NOX - 1.07*PTRATIO - 0.60*LSTAT - 0.01*TAX+0.03*AGE + 0.13*INDUS + 0.26*DISTANCE + 4.12*AVG_ROOM.

END OF THE REPORT