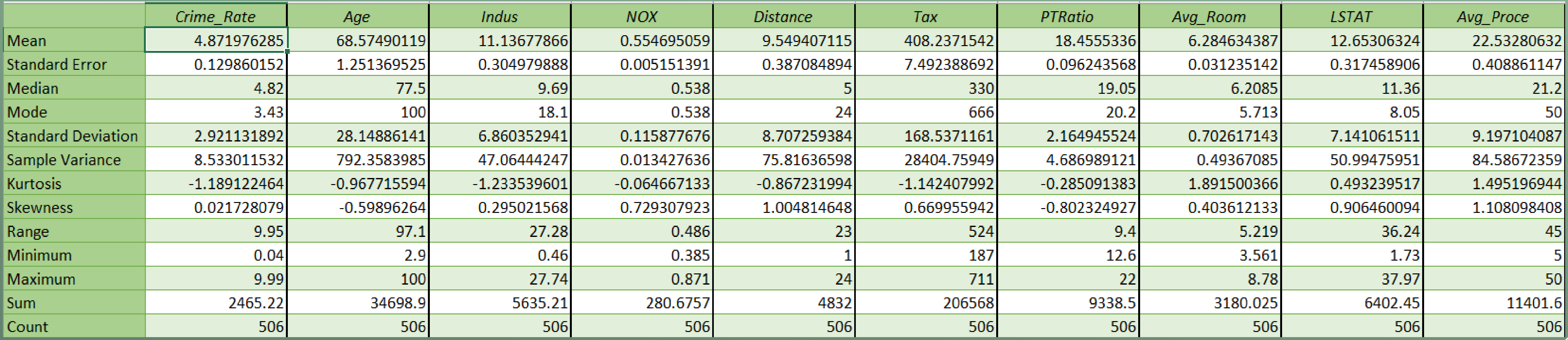
Task 1



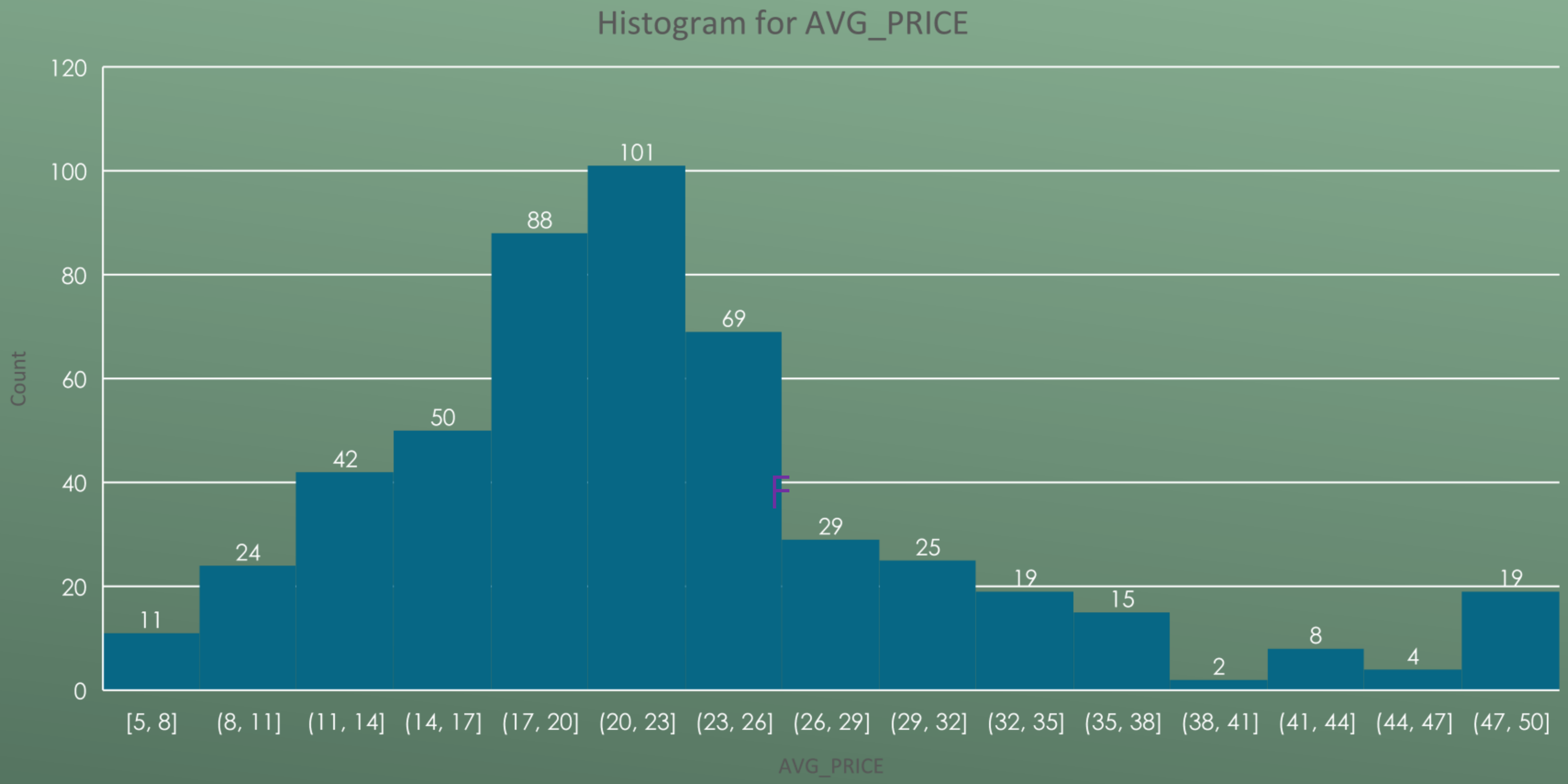
	Crime_Rate	Age	Indus	NOX	Distance	Tax	PTRatio	Avg_Room	LSTAT	Avg_Proce
Mean	4.871976285	68.57490119	11.13677866	0.554695059	9.549407115	408.2371542	18.4555336	6.284634387	12.65306324	22.53280632
Standard Error	0.129860152	1.251369525	0.304979888	0.005151391	0.387084894	7.492388692	0.096243568	0.031235142	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	-0.967715594	-1.233539601	-0.064667133	-0.867231994	-1.142407992	-0.285091383	1.891500366	0.493239517	1.495196944
Skewness	0.021728079	-0.59896264	0.295021568	0.729307923	1.004814648	0.669955942	-0.802324927	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

From the above table which depicts the summary statistics for each variable available we can infer how the given data's central tendency is situated by calculating the values of Mean, Median, Mode

We can infer how is our data dispersed around our mean or simply say how far our data are spread out from our mean by using the values of variance, Standard deviation, range.

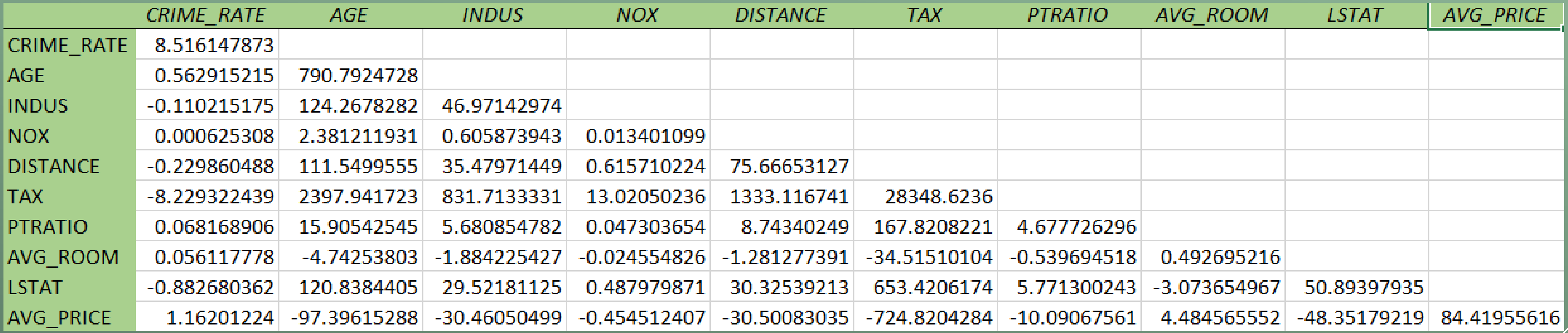
We can also see some other stats like how the distribution of the data is using the skewness and the Kurtosis, Max and Min of each variable.

TASK 2



From the histogram above we can infer the distribution of the data. By omitting the outlier (i.e.) the last 4 on the right side of the plot, we can assume that the data Is having the Normal distribution and by doing so we can predict the future values. For example if an another measure is to be taken it will most probabily lie with in the range 14000 to 26000 USD.

TASK 3

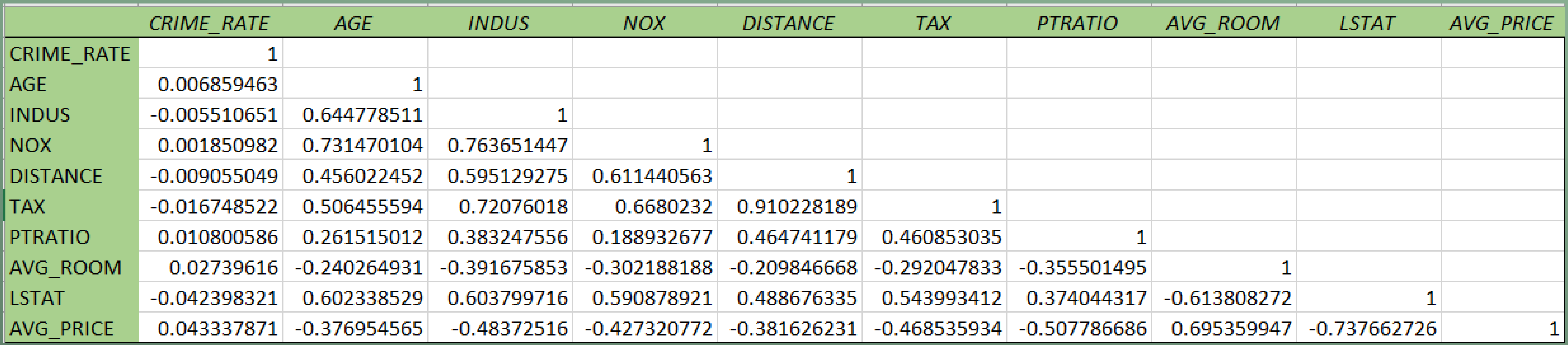


	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.516147873									
AGE	0.562915215	790.7924728								
INDUS	-0.110215175	124.2678282	46.97142974							
NOX	0.000625308	2.381211931	0.605873943	0.013401099						
DISTANCE	-0.229860488	111.5499555	35.47971449	0.615710224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

Covariance is a measure of the relationship between two random variables and to what extent they change together. A positive covariance indicates that the variables are positively related, meaning that when one variable increases, the other variable also increases. A negative covariance indicates that the variables are negatively related, meaning that when one variable increases, the other variable decreases. A high covariance indicates a strong relationship where as the low covariance means the relationship is weak.

As we can see the covariance between the Distance and Tax is(1333) very high which means a strong positive relation and the covariance between Avg_Price and Tax is(-724) very low and it means there is a weak relationship.

TASK 4



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

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two specific variables. The correlation coefficient 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, and 1 indicates a perfectly positive linear correlation between two variables.

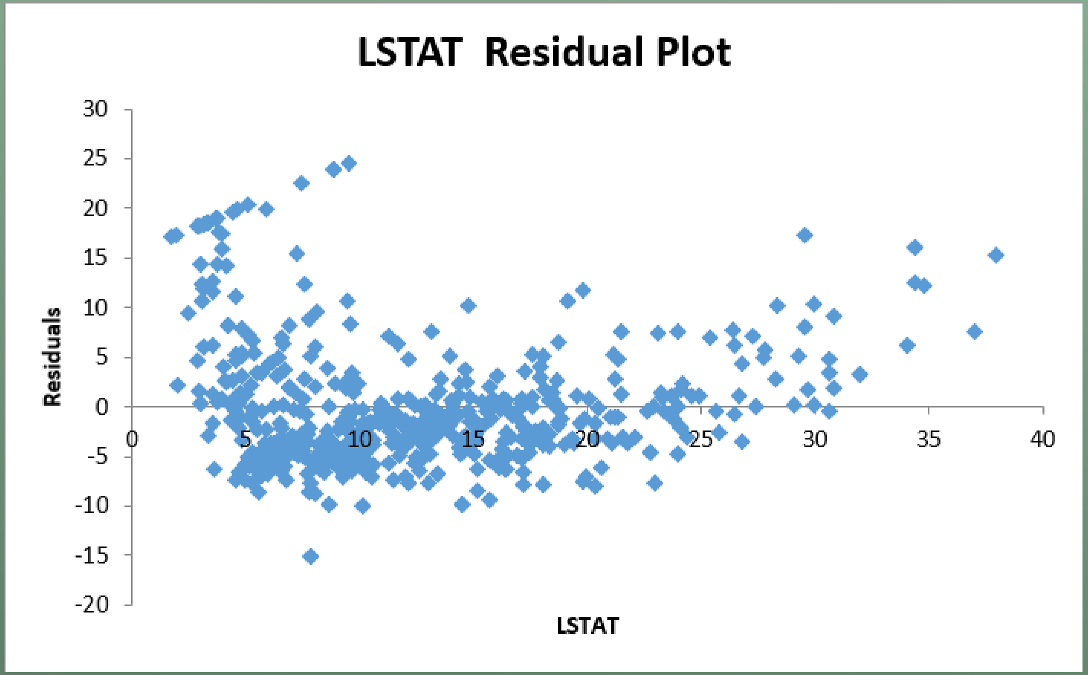
Top 3 positively correlated pairs are

- Tax and Distance(0.9102) which means there is a 91% chance if the Tax increases the Distance increases.
- NOX and Age(0.7315) which means there is a 73% chance if the NOX increases the Age increases.
- Tax and INDUS(0.7208) which means there is a 72% chance if the Tax increases the INDUS increases.

Top 3 negatively correlated pairs

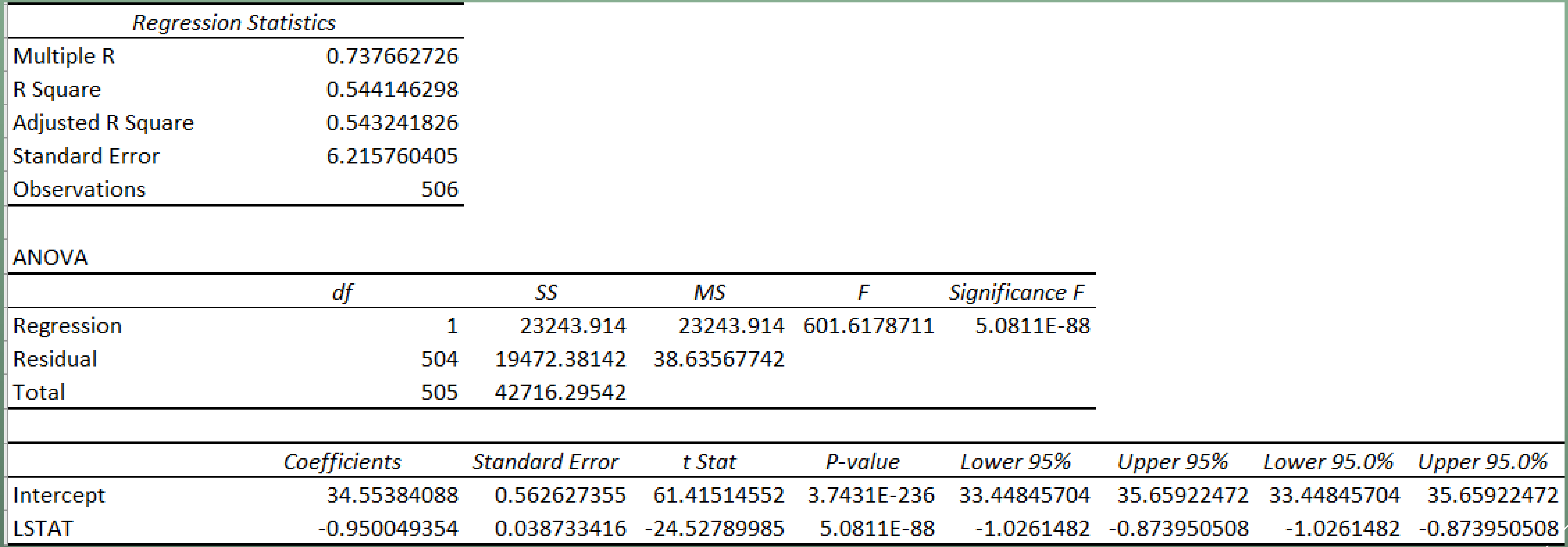
- Avg_Price and LSTAT(-0.7377) which means there is a 72% chance if the Avg_Price increases the LSTAT decreases.
- LSTAT and Avg_Room (-0.6138) which means there is a 61% chance if the LSTAT increases the Avg_Room decreases.
- Avg_Price and PT-Ratio (-0.5077) which means there is a 51% chance if the Avg_Priceincreases the PT-Ratio decreases

TASK 5



A residual plot is a graph that displays the residuals on the vertical axis and the independent variable on the horizontal axis. Residuals are the difference between observed values and predicted values. A residual plot is typically used to find problems with regression.

As we can see that the residual values are randomly distributed around the horizontal axis without any pattern. The linear regression model will be appropriate for the data.



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

From model we infer that variance explained value tells that 54.41% of the dependent variable is explained by other independent variable.

The independent value can be found using the formula.

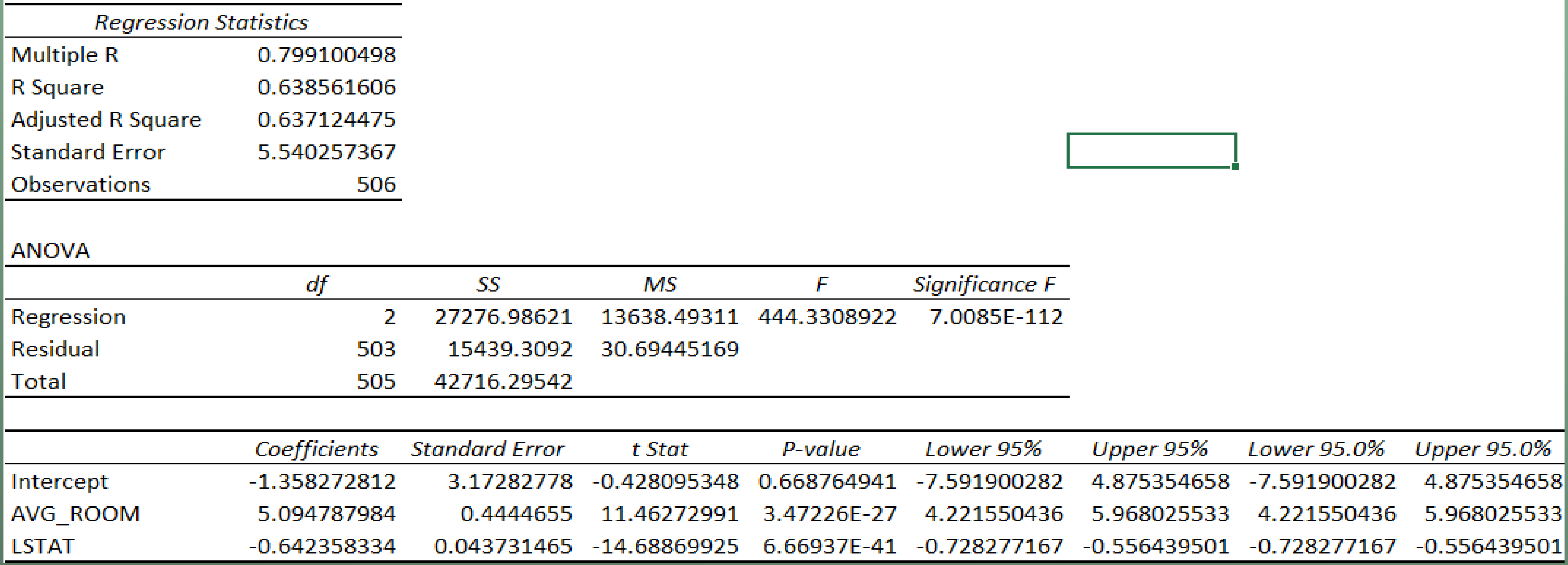
Y= **α** **+** **βX + ε** where **α** is Intercept, **β** is LSTAT coefficient and **ε** is Standard error.

In this Model

Y=34.5538+(-0.95*X)

TASK 5 B. This regression model has the significance to predict the dependent variable but the value of significance is only 54.41%

TASK 6



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

The Regression equation for the above model is given as

$$\mathbf{Y= \alpha + \beta X_1 + \beta X_2 + \varepsilon}$$

If a new house in this lobby has a 7 rooms(on a average) and has a value of 20 for L-STAT, then the Avg-Price(Y) is

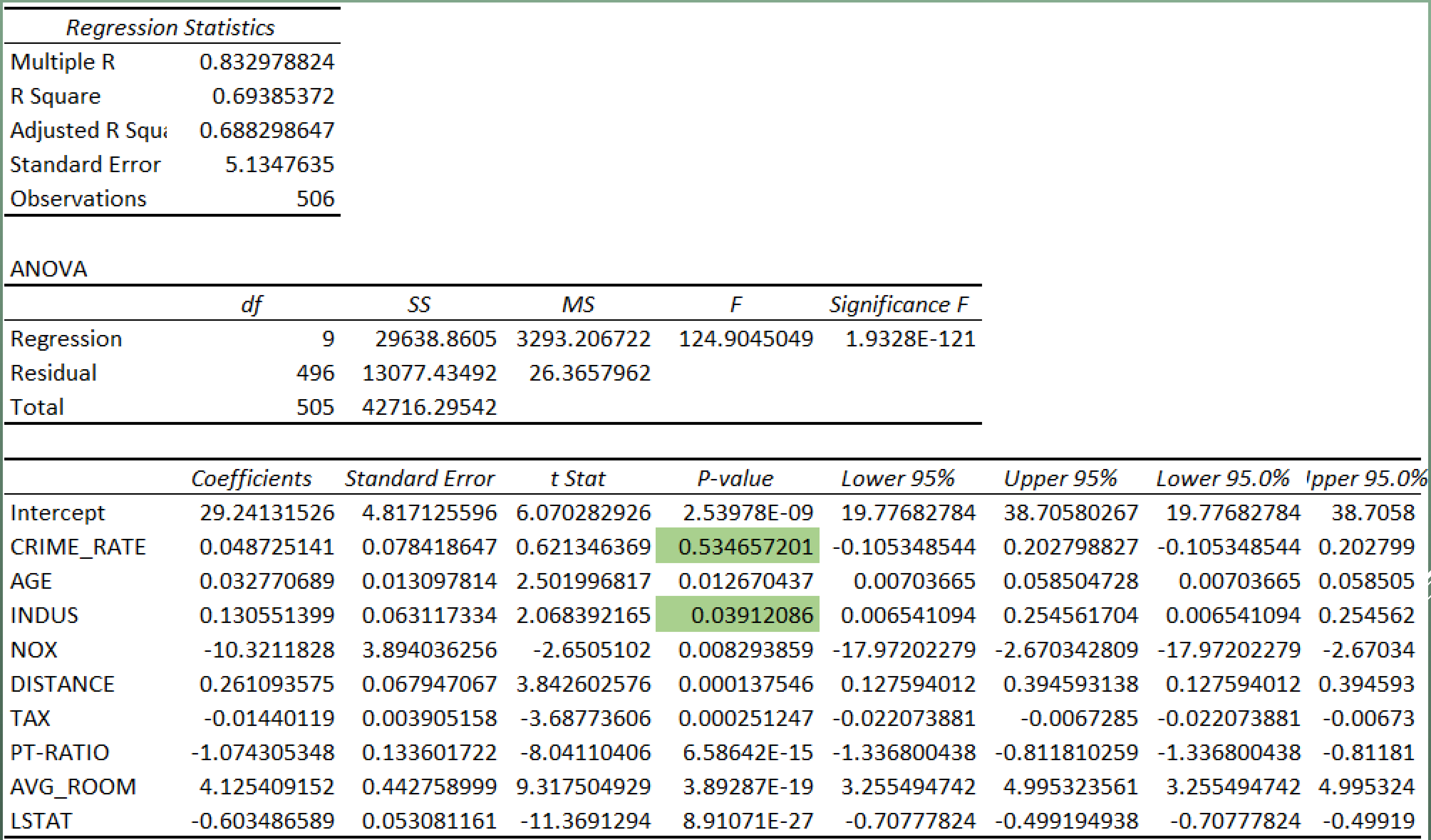
Y= (-1.358277) + (5.0947984*7) + (-0.64235834*20) + 0.445 + 0.0437

Y= 21,947 USD.

From the results we can clearly see that the company is overcharging.

TASK 6.B By comparing the R2 valued we can infer that this regression model is better than the model built in TASK 5

TASK 7



Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372
Adjusted R Squa	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			

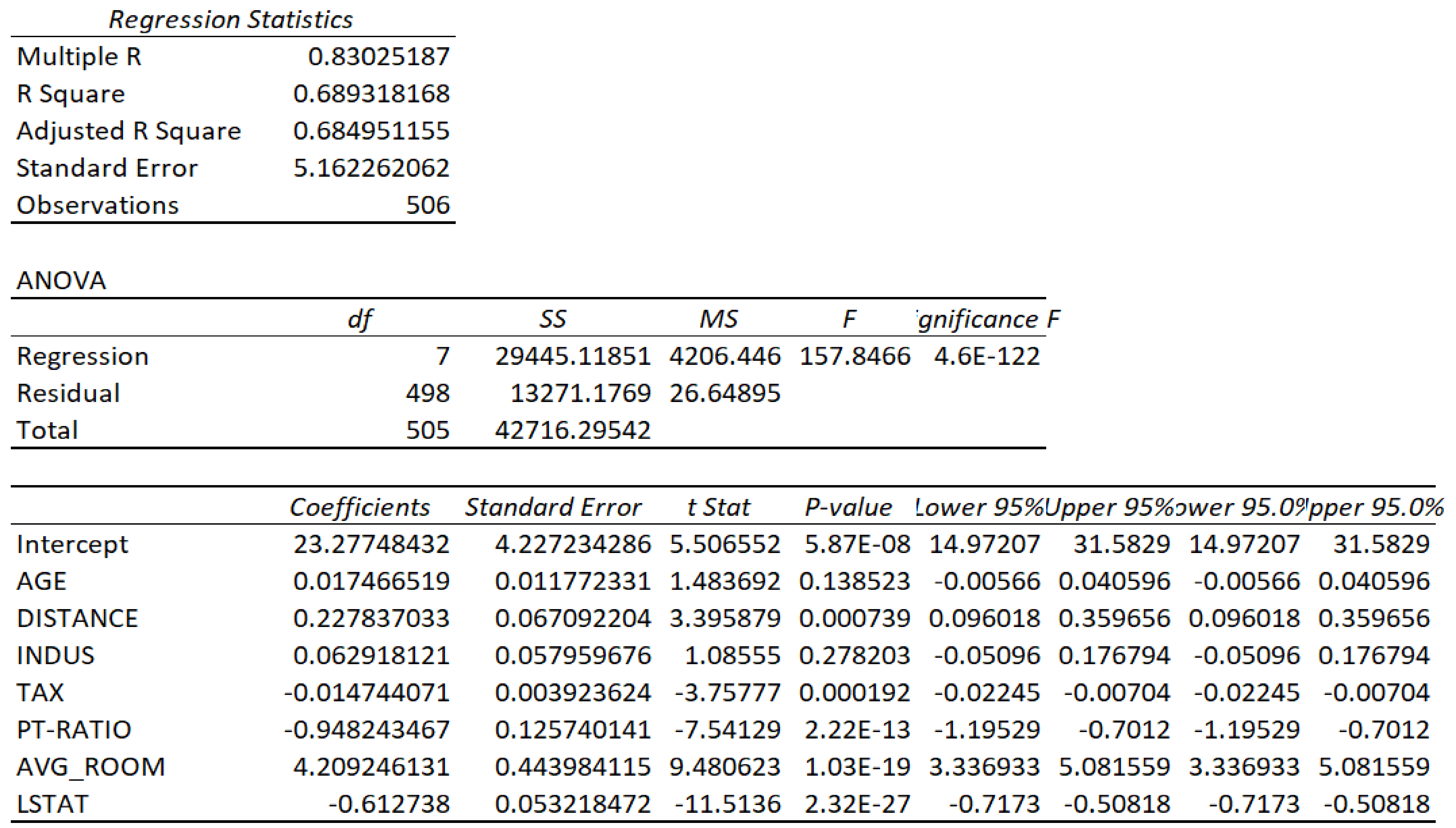
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*pper 95.0%*
Intercept	29.24131526	4.817125596	6.070282926	2.53978E-09	19.77682784	38.70580267	19.77682784	38.7058
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.534657201	-0.105348544	0.202798827	-0.105348544	0.202799
AGE	0.032770689	0.013097814	2.501996817	0.012670437	0.00703665	0.058504728	0.00703665	0.058505
INDUS	0.130551399	0.063117334	2.068392165	0.03912086	0.006541094	0.254561704	0.006541094	0.254562
NOX	-10.3211828	3.894036256	-2.6505102	0.008293859	-17.97202279	-2.670342809	-17.97202279	-2.67034
DISTANCE	0.261093575	0.067947067	3.842602576	0.000137546	0.127594012	0.394593138	0.127594012	0.394593
TAX	-0.01440119	0.003905158	-3.68773606	0.000251247	-0.022073881	-0.0067285	-0.022073881	-0.00673
PT-RATIO	-1.074305348	0.133601722	-8.04110406	6.58642E-15	-1.336800438	-0.811810259	-1.336800438	-0.81181
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.89287E-19	3.255494742	4.995323561	3.255494742	4.995324
LSTAT	-0.603486589	0.053081161	-11.3691294	8.91071E-27	-0.70777824	-0.499194938	-0.70777824	-0.49919

From the regression model for TASK 7 we can infer that the R2 tells that 69.39% of the dependent variable is explained by other 9 independent variables

By looking closely at the model. We can see P-values which describes how significant the variable is in deciding the Avg_Price.

- We can say that the Crime_rate has a P – value of 0.5346 which is way higher than that the prescribed value of 0.05. this means that the Crime_ratio affect the Avg_price a little only.
- We can say that the Age has a P – value of 0.0126 which is lower than that the prescribed value of 0.05. this means that the Age has a high significance on deciding the Avg_price.
- We can say that the INDUS has a P – value of 0.0391 which is lower than that the prescribed value of 0.05. this means that the INDUS has a high significance on deciding the Avg_price.
- We can say that the NOX has a P – value of 0.008293859 which is lower than that the prescribed value of 0.05. this means that the NOX has a high significance on deciding the Avg_price.
- We can say that the Distance has a P – value of 0.000137546 which is lower than that the prescribed value of 0.05. this means that the Distance has a high significance on deciding the Avg_price.
- We can say that the Tax has a P – value of 0.000251247 which is lower than that the prescribed value of 0.05. this means that the Tax has a high significance on deciding the Avg_price.
- We can say that the PT-Ratio has a P – value of 6.58642E-15which is lower than that the prescribed value of 0.05. this means that the PT-Ratio has a high significance on deciding the Avg_price.
- We can say that the Avg_Room has a P – value of 3.89287E-19 which is way higher that the prescribed value of 0.05. this means that the Avg_room has a high significance on deciding the Avg_price
- We can say that the LSTAT has a P – value of 8.91071E-27 which is way higher that the prescribed value of 0.05. this means that the LSTAT has a high significance on deciding the Avg_price

TASK 8



Regression Statistics	
Multiple R	0.83025187
R Square	0.689318168
Adjusted R Square	0.684951155
Standard Error	5.162262062
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	7	29445.11851	4206.446	157.8466	4.6E-122
Residual	498	13271.1769	26.64895		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	23.27748432	4.227234286	5.506552	5.87E-08	14.97207	31.5829	14.97207	31.5829
AGE	0.017466519	0.011772331	1.483692	0.138523	-0.00566	0.040596	-0.00566	0.040596
DISTANCE	0.227837033	0.067092204	3.395879	0.000739	0.096018	0.359656	0.096018	0.359656
INDUS	0.062918121	0.057959676	1.08555	0.278203	-0.05096	0.176794	-0.05096	0.176794
TAX	-0.014744071	0.003923624	-3.75777	0.000192	-0.02245	-0.00704	-0.02245	-0.00704
PT-RATIO	-0.948243467	0.125740141	-7.54129	2.22E-13	-1.19529	-0.7012	-1.19529	-0.7012
AVG_ROOM	4.209246131	0.443984115	9.480623	1.03E-19	3.336933	5.081559	3.336933	5.081559
LSTAT	-0.612738	0.053218472	-11.5136	2.32E-27	-0.7173	-0.50818	-0.7173	-0.50818

The regression equation for the ATSK 8 can be written as

$$Y= 23.27748432 + (0.017466519 * X1) + (0.227837033 *X2)+(0.062918121 *X3)+(-0.014744071 *X4)+(-0.948243467 *X5)+(4.209246131 *X6)+(-0.612738 *X7)$$

From the R2 value we can say that this model is slightly more significant than the last model