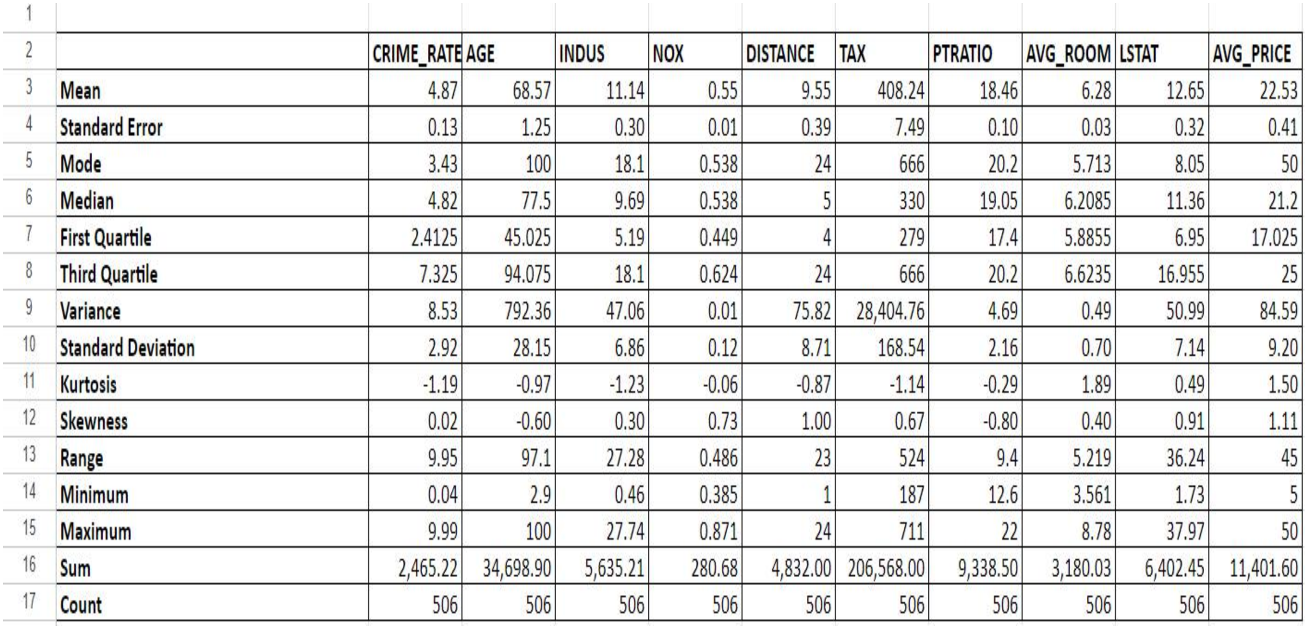
Assignment – Terro’s real estate agency

By- Tanishq Dalal

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

Ans-

- Open the data analysis tool pack by clicking on the "Data" tab in the top menu bar, then selecting "Data Analysis" in the "Analysis" section.
- Select the "Descriptive Statistics" for summary statistics.
- Enter the input range of data.
- Choose where you want to output your results, either in a new worksheet or in a new range of cells. So I select the new worksheet.
- Click "OK" to generate the summary statistics.



	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
Mean	4.87	68.57	11.14	0.55	9.55	408.24	18.46	6.28	12.65	22.53
Standard Error	0.13	1.25	0.30	0.01	0.39	7.49	0.10	0.03	0.32	0.41
Mode	3.43	100	18.1	0.538	24	666	20.2	5.713	8.05	50
Median	4.82	77.5	9.69	0.538	5	330	19.05	6.2085	11.36	21.2
First Quartile	2.4125	45.025	5.19	0.449	4	279	17.4	5.8855	6.95	17.025
Third Quartile	7.325	94.075	18.1	0.624	24	666	20.2	6.6235	16.955	25
Variance	8.53	792.36	47.06	0.01	75.82	28,404.76	4.69	0.49	50.99	84.59
Standard Deviation	2.92	28.15	6.86	0.12	8.71	168.54	2.16	0.70	7.14	9.20
Kurtosis	-1.19	-0.97	-1.23	-0.06	-0.87	-1.14	-0.29	1.89	0.49	1.50
Skewness	0.02	-0.60	0.30	0.73	1.00	0.67	-0.80	0.40	0.91	1.11
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	2,465.22	34,698.90	5,635.21	280.68	4,832.00	206,568.00	9,338.50	3,180.03	6,402.45	11,401.60
Count	506	506	506	506	506	506	506	506	506	506

Observations

1. **Crime Rate :**

- 50% of the crime rate is below 4.82 and 50% above this value.
- Maximum crime rate is around 3.4
- On an average , crime rate in town is 4.8
- Standard deviation of 2.9 says that data deviates from mean by this value.
- Negative kurtosis signifies flatter curve
- Skewness is nearly 0 which says curve follows normal distributuion.

2. Age :

- On an average , the houses built inn town is around 68 years.
- Negative skewness means that tail of the distribution points to the left or here we can say that more number of houses are built before 1940.
- Negative Kurtosis gives us a flatter distribution for this variable.

3. INDUS :

- On an average, 11.13 % of property belongs to non-retail business.

- Negative kurtosis specifies flatter curve – indicating values are spread across mean value.
- Positive skewness indicates that tail is towards right and most of the houses have less than 11.13% of land as non-retail business land.

4. NOX :

- On an average , nitric oxide concentration is around 0.55 ppm
- The data gives us slightly negative kurtosis but since the value is small so we can say follows normal distribution.
- Skewness of about 0.79 gives a right tailed distribution or can say more number of houses have NO concentration below 0.55 ppm.

5. DISTANCE :

- On an average, distance from highway is around 9.5 miles.
- Maximum houses have 24 miles of distance from highway.
- Negative kurtosis says that data gives us a flatter distribution – curve is not steep.
- Positive skewness indicates that more number of houses are less than 9.5 miles away from highway.

6. TAX :

- On an average , tax rate is $408 .
- The maximum number of houses have tax rate aroung $666.
- Negative kurtosis and positive skewness.

7. PTRATIO :

- On an average , pupil teacher ratio is 18 for 506 houses.
- Maximum houses give 20 as pupil teacher ratio.
- Negative skewness indicates that left tailed distribution : more number of houses have more than 19 as a pupil teacher ratio.

8. AVG_ROOM :

- On an average , 6 rooms are there.
- Positive kurtosis gives us a sharp curve than normal curve – saying more values are concentrated near to median.
- Positive skewness indicates right tailed distribution which says that more number of houses have less than 6 rooms.

9. LSTAT :

- On an average , 12% of population has lower status.
- Positive kurtosis gives us a sharp curve
- Positive skewness tells us that more number of houses have less than 12% lower status population.

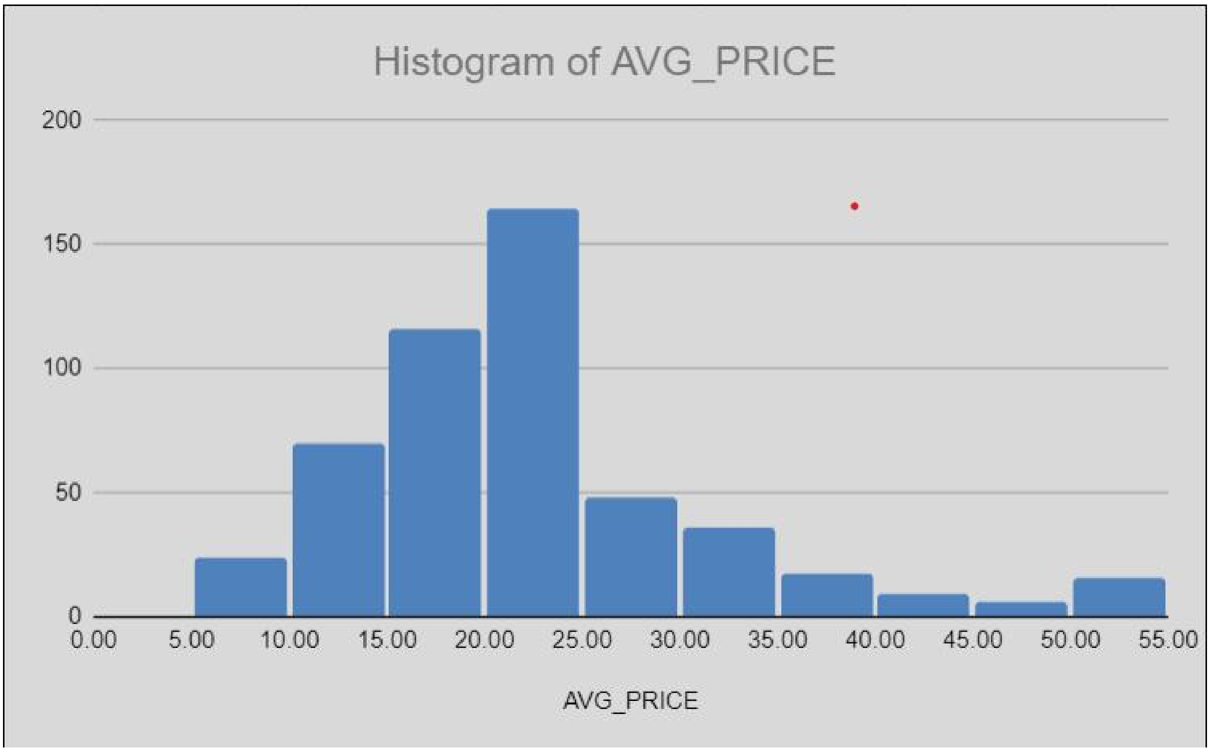
10. AVG_PRICE :

- Average value of price of house is around $22500.
- Positive kurtosis gives us a sharp curve and positive skewness tells us that more number of houses have price less than $22500.

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

Ans-

- Select the Avg_Price variable column in your dataset.
- Click on the "Insert" tab in the top ribbon.
- Click on "Histogram" in the "Charts" section.
- Choose the appropriate bin range and bin width for your histogram.
- Click "OK".



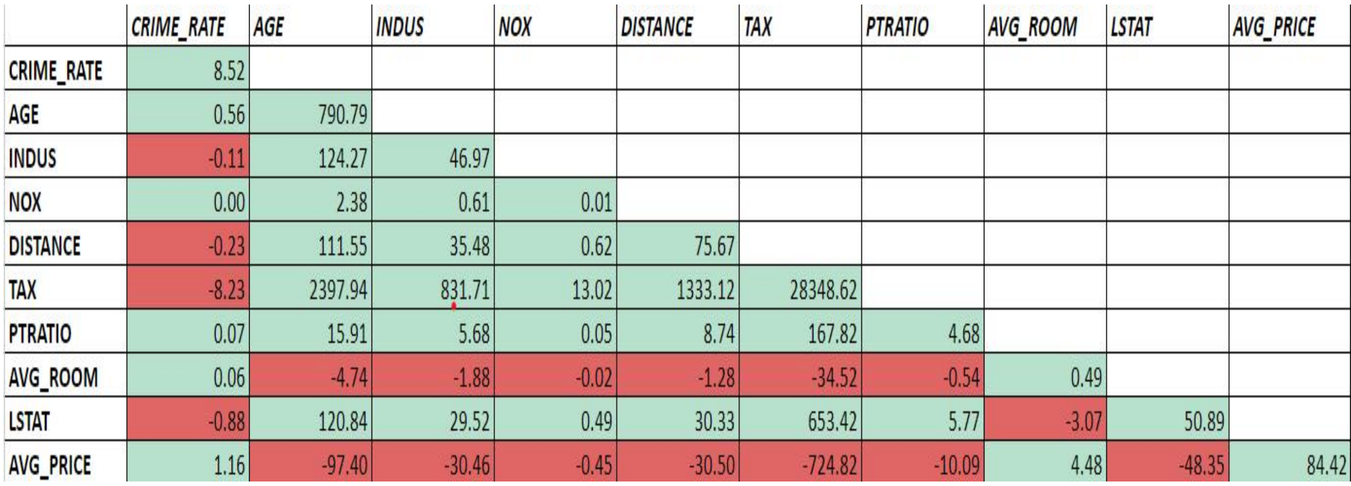
Observations

- We may deduce that there is a **greater count** of the average price between the range from the histogram that was plotted (15-25).
- The range contains the **minimum** count of average price (45-50).
- We can also see that the **average price** is somewhere in the middle of the range (20-30).
- The **standard deviation** of the data is likely to be moderate to high, given the spread of the data and the long tail on the right side of the distribution.

3) Compute the covariance matrix. Share your observations.

Ans-

- Open the data analysis tool pack by clicking on the "Data" tab in the top menu bar, then selecting "Data Analysis" in the "Analysis" section.
- Select the "Covariance Matrix".
- Enter the input range of data.
- Choose where you want to output your results, either in a new worksheet or in a new range of cells. So I select the new worksheet.
- Click "OK" to generate.



	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.52									
AGE	0.56	790.79								
INDUS	-0.11	124.27	46.97							
NOX	0.00	2.38	0.61	0.01						
DISTANCE	-0.23	111.55	35.48	0.62	75.67					
TAX	-8.23	2397.94	831.71	13.02	1333.12	28348.62				
PTRATIO	0.07	15.91	5.68	0.05	8.74	167.82	4.68			
AVG_ROOM	0.06	-4.74	-1.88	-0.02	-1.28	-34.52	-0.54	0.49		
LSTAT	-0.88	120.84	29.52	0.49	30.33	653.42	5.77	-3.07	50.89	
AVG_PRICE	1.16	-97.40	-30.46	-0.45	-30.50	-724.82	-10.09	4.48	-48.35	84.42

Observations

- Crime Rate has just one positive relation, with avg_price and that too not a significant one as per the value.
- Crime rate follows a highly negative relation with Tax – means the house which has high tax rate , their crime rate is low.
- The property tax rate is high for those houses who have been there for long since 1940. They share a positive relation.
- Non-retail business Industry , NOX , Dstance : they all share a positive relation with tax rate.
- Distance from highway shows a negative relation with average price of house.
- Tax and average price of house both share negative relation.
- The average price of house has a negative relation with pupil teacher ratio and LSTAT too.

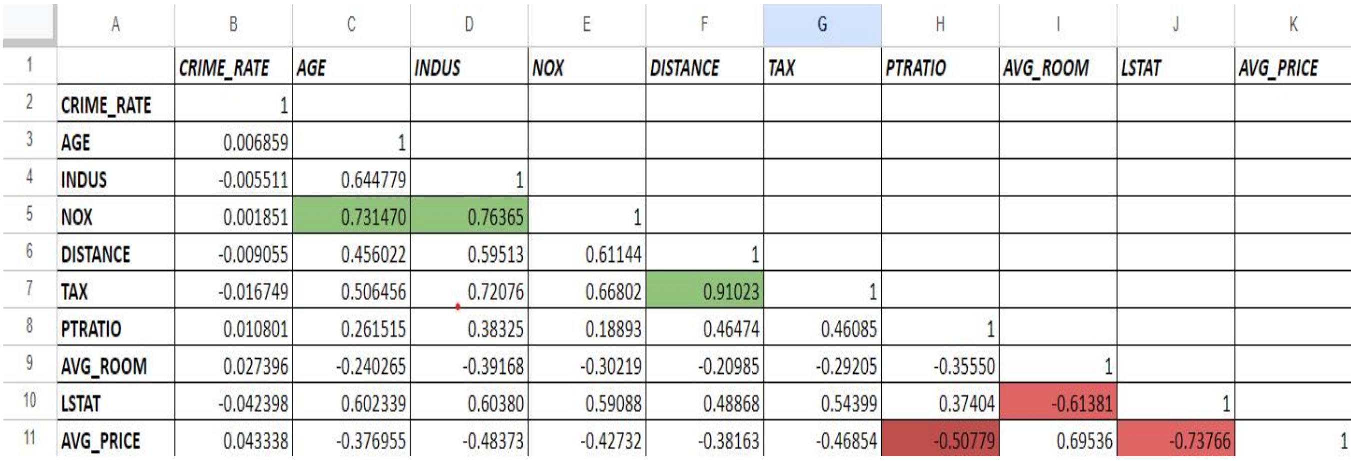
4) Create a correlation matrix of all the variables (Use Data analysis tool pack).

a) Which are the top 3 positively correlated pairs and

b) Which are the top 3 negatively correlated pairs.

Ans-

- Open the data analysis tool pack by clicking on the "Data" tab in the top menu bar, then selecting "Data Analysis" in the "Analysis" section.
- Select the "Correlation Matrix".
- Enter the input range of data.
- Choose where you want to output your results, either in a new worksheet or in a new range of cells. So I select the new worksheet.
- Click "OK" to generate.



	A	B	C	D	E	F	G	H	I	J	K
1		*CRIME_RATE*	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
2	CRIME_RATE	1									
3	AGE	0.006859	1								
4	INDUS	-0.005511	0.644779	1							
5	NOX	0.001851	0.731470	0.76365	1						
6	DISTANCE	-0.009055	0.456022	0.59513	0.61144	1					
7	TAX	-0.016749	0.506456	0.72076	0.66802	0.91023	1				
8	PTRATIO	0.010801	0.261515	0.38325	0.18893	0.46474	0.46085	1			
9	AVG_ROOM	0.027396	-0.240265	-0.39168	-0.30219	-0.20985	-0.29205	-0.35550	1		
10	LSTAT	-0.042398	0.602339	0.60380	0.59088	0.48868	0.54399	0.37404	-0.61381	1	
11	AVG_PRICE	0.043338	-0.376955	-0.48373	-0.42732	-0.38163	-0.46854	-0.50779	0.69536	-0.73766	1

a) Which are the top 3 positively correlated pairs

Looking at the table, we can see that the **top 3 positively correlated** pairs are:

- **DISTANCE AND TAX- 0.91023**
- **INDUS AND NOX- 0.76365**
- **AGE AND NOX- 0.731470**

b) Which are the top 3 negatively correlated pairs

Looking at the table, we can see that the **top 3 negatively correlated** pairs are:

- **AVG_PRICE AND LSTAT- -0.73766**
- **AVG_ROOM AND LSTAT- -0.61381**
- **PTRATIO AND AVG_PRICE- -0.50779**

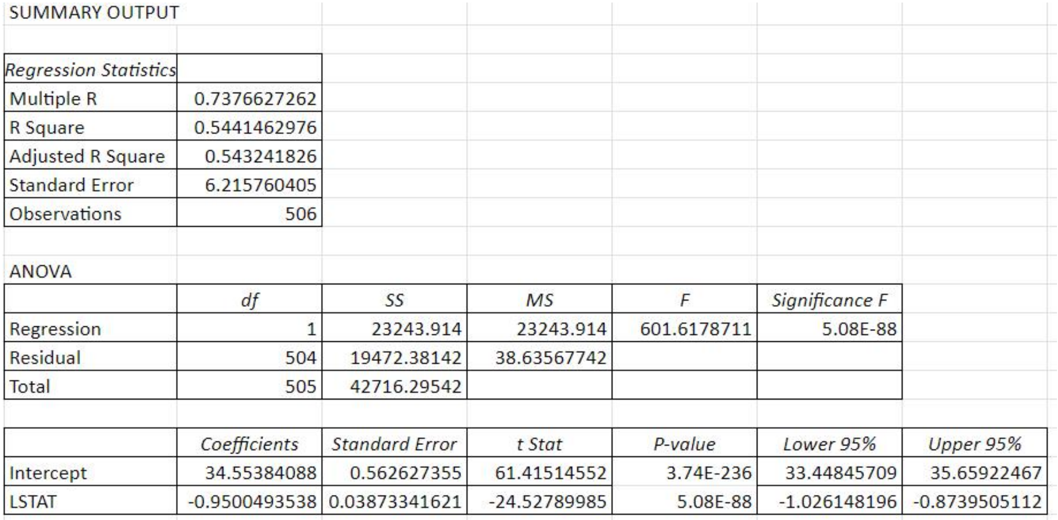
Q.5) Build an initial regression model with AVG_PRICE as ‘y’ (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot.

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

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

ANS-

- Open the data analysis tool pack by clicking on the "Data" tab in the top menu bar, then selecting "Data Analysis" in the "Analysis" section.
- Select the "Linear Regression".
- Enter the input range of data.
- Select AVG_PRICE as ‘y’and LSTAT as ‘x’.
- Select Residual and Residual plot.
- Choose where you want to output your results, either in a new worksheet or in a new range of cells. So I select the new worksheet.
- Click "OK" to generate.



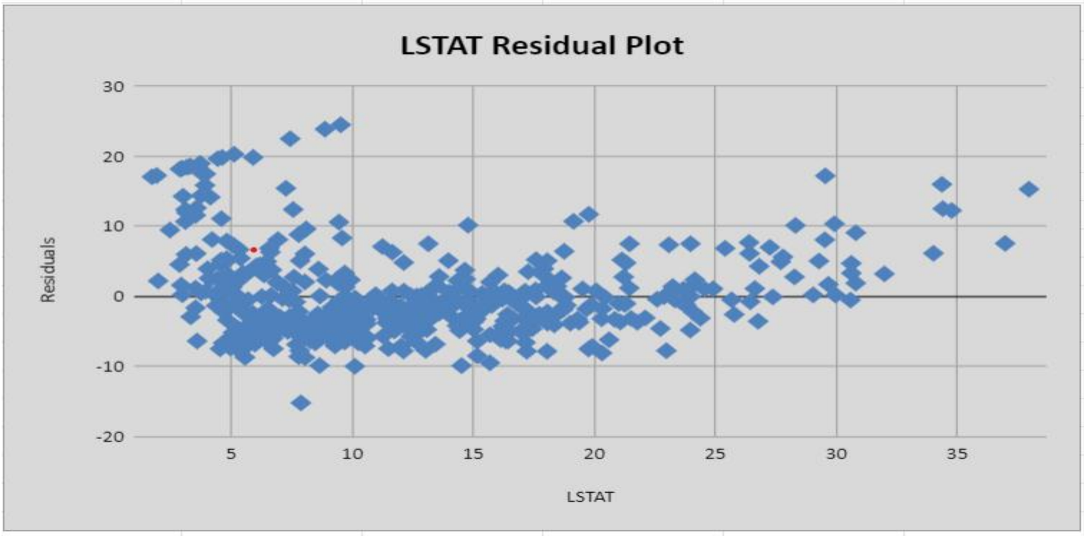
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.7376627262
R Square	0.5441462976
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	5.08E-88
Residual	504	19472.38142	38.63567742		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	34.55384088	0.562627355	61.41514552	3.74E-236	33.44845709	35.65922467
LSTAT	-0.9500493538	0.03873341621	-24.52789985	5.08E-88	-1.026148196	-0.8739505112



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

- Based on the regression summary output, we can see that the R-squared value is 0.544, which means that about 54.4% of the variance in the dependent variable (AVG_PRICE) is explained by the independent variable (LSTAT).

- The coefficient value for LSTAT is -0.95, which indicates that as the LSTAT value increases by one unit, the AVG_PRICE value decreases by 0.95 units

- Intercept: The intercept value of 34.553 is the predicted value of AVG_PRICE when LSTAT is equal to 0.
- The residual plot shows that there is a non-linear relationship between the independent and dependent variables, as there is a clear pattern in the residuals. This suggests that a linear regression model may not be the best fit for the data.

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

Checking the Regression summary for LSTAT variable

- p-value significantly less than 0.05 (level of significance)
- LSTAT is highly correlated with avg_price : 0.74
- R square : 0.54 == 54% of variation in avg_price can be explained by LSTAT. So we can say it is
  significant.
- Regression equation : Avg_price = 34.55 -0.95*(LSTAT). So , Yes LSTAT is significant variable for
  avg_price.

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

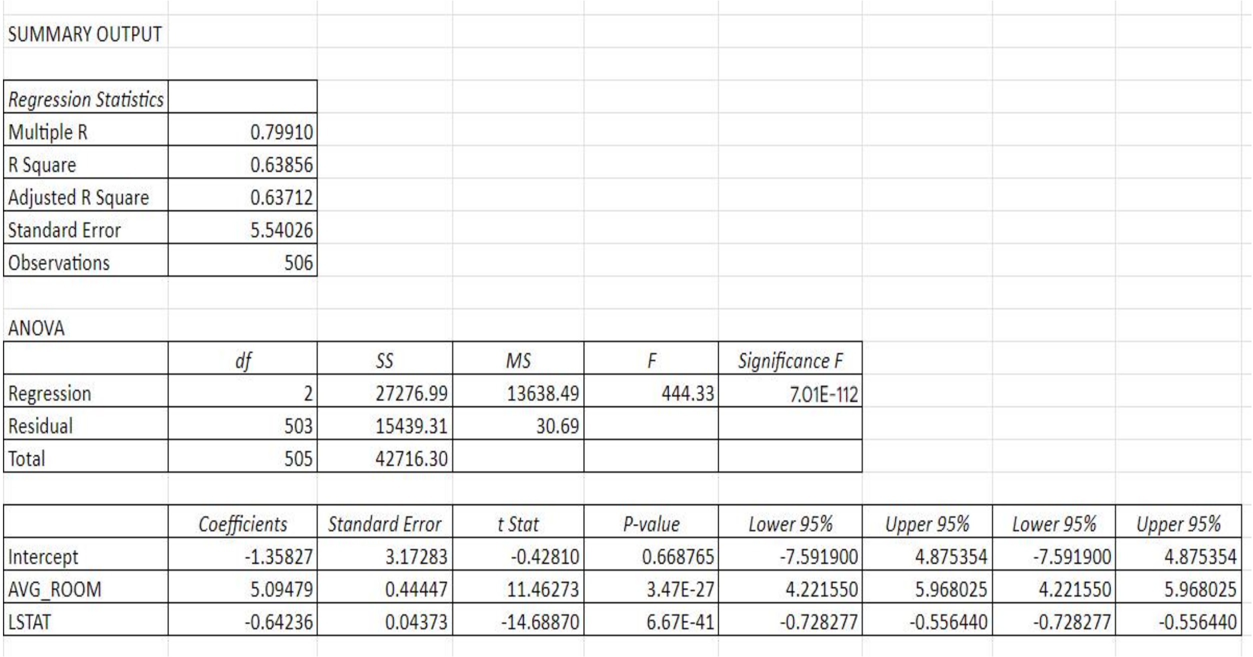
6) Build a new Regression model including LSTAT and AVG_ROOM together as independent variables and AVG_PRICE as dependent variable.

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?

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?

Ans-

- Open the data analysis tool pack by clicking on the "Data" tab in the top menu bar, then selecting "Data Analysis" in the "Analysis" section.
- Select the "Linear Regression".
- Enter the input range of data.
- Select AVG_PRICE as ‘y’and LSTAT and AVG_ROOM as ‘x’.
- Choose where you want to output your results, either in a new worksheet or in a new range of cells. So I select the new worksheet.
- Click "OK" to generate.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.79910
R Square	0.63856
Adjusted R Square	0.63712
Standard Error	5.54026
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	27276.99	13638.49	444.33	7.01E-112
Residual	503	15439.31	30.69		
Total	505	42716.30			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95%*	*Upper 95%*
Intercept	-1.35827	3.17283	-0.42810	0.668765	-7.591900	4.875354	-7.591900	4.875354
AVG_ROOM	5.09479	0.44447	11.46273	3.47E-27	4.221550	5.968025	4.221550	5.968025
LSTAT	-0.64236	0.04373	-14.68870	6.67E-41	-0.728277	-0.556440	-0.728277	-0.556440

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?

The regression equation with “LSTAT” and “AVG_ROOM” as independent variables and “AVG_PRICE” as the dependent variable is:

AVG_PRICE = β0 + β1*LSTAT + β2*AVG_ROOM

= -1.35 + (-0.64 * LSTAT) + (5.09 * AVG_ROOM)

= -1.35 + (-0.64 * 20) + (5.09 * 7)

= 21.5

So, AVG_PRICE IS 21,500 USD,
If company is charging 30000 USD it means they are Overcharging for this locality.

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-square for Q5 : 0.543241
- Adjusted R-square for this model: 0.63712

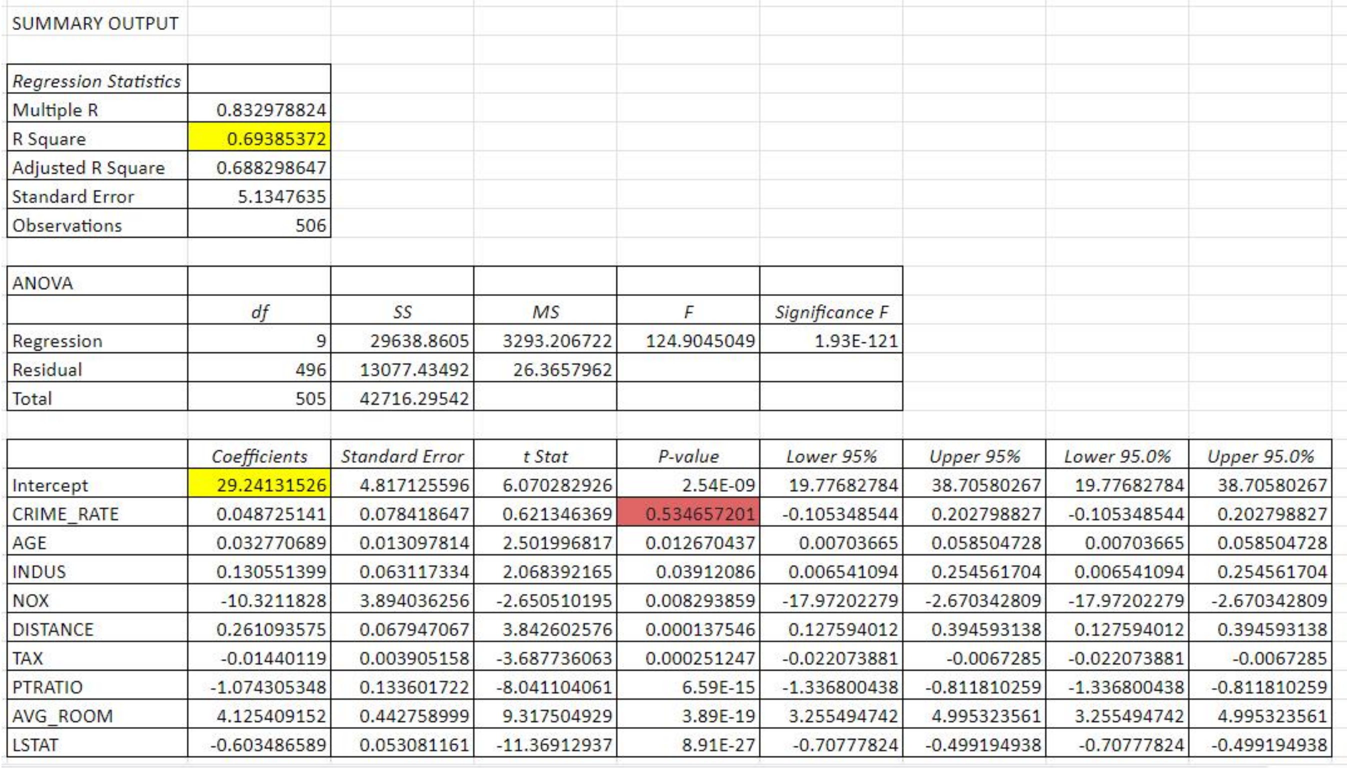
So, We can say that 63% of variation happened in Avg_Price is because of Avg_Room and LSTAT.
Also the standard error is less when compare the value from Q5

Overall, based on these metrics, we can say that the current model performs better than the previous model in explaining the variability in the dependent variable using the independent variables.

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 Rsquare, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE.

Ans-

- Open the data analysis tool pack by clicking on the "Data" tab in the top menu bar, then selecting "Data Analysis" in the "Analysis" section.
- Select the "Linear Regression".
- Enter the input range of data.
- Select AVG_PRICE as ‘y’and all others as ‘x’.
- Choose where you want to output your results, either in a new worksheet or in a new range of cells. So I select the new worksheet.
- Click "OK" to generate.



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	1.93E-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	2.54E-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	6.59E-15	-1.336800438	-0.811810259	-1.336800438	-0.811810259
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.89E-19	3.255494742	4.995323561	3.255494742	4.995323561
LSTAT	-0.603486589	0.053081161	-11.36912937	8.91E-27	-0.70777824	-0.499194938	-0.70777824	-0.499194938

- The multiple regression model has an adjusted R square of 0.688, which means that 68.8% of the variation in the dependent variable (AVG_PRICE) can be explained by the independent variables in the model.
- The Intercept value of 29.24 indicates that when all the independent variables are 0, the average price of a house in the area is $29,240.
- The coefficients are non zero, hence affects the avg_price value.
- Though INDUA and AGE's p-value less than 0.05, but there value is not very less so these variables have less significance on average price variable.
- CRIME_RATE has high positive significance on AVG_PRICE than other variables.
- LSTAT,PTRATIO,NOX, AND TAX variables follow negative relation with AVG_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: (8 marks)

a) Interpret the output of this model.

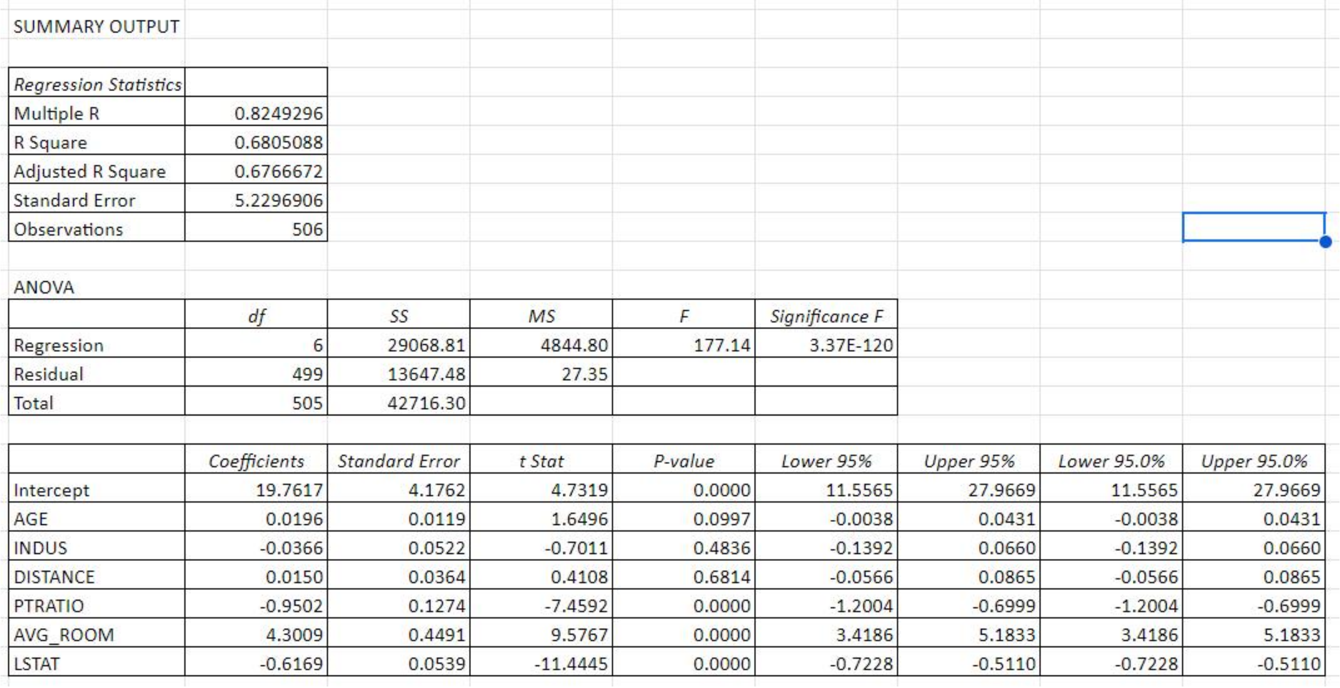
b) Compare the adjusted R-square value of this model with the model in the previous question, which model performs better according to the value of adjusted R-square?

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.

ANS-

- Open the data analysis tool pack by clicking on the "Data" tab in the top menu bar, then selecting "Data Analysis" in the "Analysis" section.
- Select the "Linear Regression".
- Enter the input range of data.
- Select AVG_PRICE as ‘y’and all others except CRIME_RATE as ‘x’.
- Choose where you want to output your results, either in a new worksheet or in a new range of cells. So I select the new worksheet.
- Click "OK" to generate.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8249296
R Square	0.6805088
Adjusted R Square	0.6766672
Standard Error	5.2296906
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	29068.81	4844.80	177.14	3.37E-120
Residual	499	13647.48	27.35		
Total	505	42716.30			

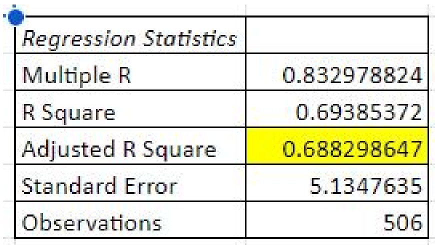
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	19.7617	4.1762	4.7319	0.0000	11.5565	27.9669	11.5565	27.9669
AGE	0.0196	0.0119	1.6496	0.0997	-0.0038	0.0431	-0.0038	0.0431
INDUS	-0.0366	0.0522	-0.7011	0.4836	-0.1392	0.0660	-0.1392	0.0660
DISTANCE	0.0150	0.0364	0.4108	0.6814	-0.0566	0.0865	-0.0566	0.0865
PTRATIO	-0.9502	0.1274	-7.4592	0.0000	-1.2004	-0.6999	-1.2004	-0.6999
AVG_ROOM	4.3009	0.4491	9.5767	0.0000	3.4186	5.1833	3.4186	5.1833
LSTAT	-0.6169	0.0539	-11.4445	0.0000	-0.7228	-0.5110	-0.7228	-0.5110

a) Interpret the output of this model.

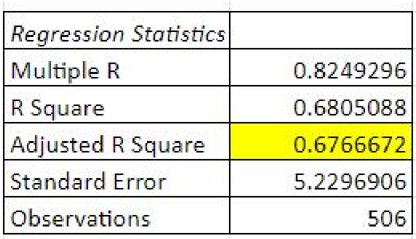
The new equation is:

Avg_Price=29.42 + 0.033*(AGE) + 0.13*(INDUS)-10.27*(NOX) + 0.26*(DISTANCE) - 0.014*(TAX)- 1.07*(PTRATIO) + 4.13*(AVG_ROOM) - 0.60*(LSTAT)

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?



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

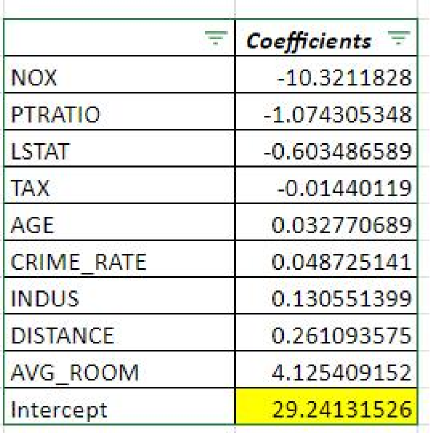


Regression Statistics	
Multiple R	0.8249296
R Square	0.6805088
Adjusted R Square	0.6766672
Standard Error	5.2296906
Observations	506

Comparing the adjusted R-squared values, we can see that the new model have a lower adjusted R-squared value of 0.676 compared to the previous model adjusted R-squared of

0.688. This indicates that the previous model is a better fit for the data than the original model.

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?



	Coefficients
NOX	-10.3211828
PTRATIO	-1.074305348
LSTAT	-0.603486589
TAX	-0.01440119
AGE	0.032770689
CRIME_RATE	0.048725141
INDUS	0.130551399
DISTANCE	0.261093575
AVG_ROOM	4.125409152
Intercept	29.24131526

If the value of NOX variable is more in the locality , the average price of house will decrease.

d) Write the regression equation from this model

Regression Equation:

Avg_Price=29.42 + 0.033*(AGE) + 0.13*(INDUS) - 10.27*(NOX) + 0.26*(DISTANCE) - 0.014*(TAX) -1.07*(PTRATIO) + 4.13*(AVG_ROOM) - 0.60*(LSTAT)