Business report

Question 1

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

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
Mean	4.872	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

To generate statistical analysis I had move into data-data analysis then I move on to descriptive analysis then I input all the variable in to range and tick on the summary statistics the below is the interpretation my observation given below overall observation is 506

a) Crime rate =The crime rate mean is 4.872 standard deviation is 2.92 whereas sample variance is 8.53 and the range is from 0.04-0.99, and the data is skewed sightly towards right(0.021)

b) Age = the average age of houses is 68.5 years with standard deviation 28.14, whereas sample variance is 792.35 with the range 2.9 to 100 the data is slightly skewed towards left (skewness=-0.59)

c) Indus = the mean proportion of non-retail business acres per town is 11.13 with standard deviation 6.86, whereas sample variance is 47.06 with the range 0.46 to 27.74 the data is slightly skewed towards right (skewness=-0.29)

d)NOX= the average of nitric oxide concentration is 0.555 with standard deviation 0.11, whereas sample variance is 0.01 with the range 0.385 to 0.871 the data is slightly skewed towards right (skewness=0.72)

e) Distance= the highway mean distance is 9.54 with standard deviation 8.70, whereas sample variance is 75.81 with the range 1 to 24 the data is slightly skewed towards right (skewness = 1.00)

f) Tax = the mean full value- property-tax rate$10000 is 408.23 with standard deviation 168.53, whereas sample variance is 28404.759 with the range 187 to 711 the data is slightly skewed towards right (skewness= 0.66)

g) PT ratio= the mean pupil-teacher ratio by town is 18.46, with standard deviation 2.16, whereas sample variance is 4.68 with the range 12.6 to 22 the data is slightly skewed towards left (skewness = -0.802)

h) Average Room = the average number of rooms per dwelling is 6.285 with standard deviation 0.703, whereas sample variance is 0.493 with the range 3.569 to 8.78 the data is slightly skewed towards right (skewness = 0.40)

i) LSTAT= the percentage of the lower status of the population is 12.56 with standard deviation 7.14, whereas sample

variance is 50.99 with the range 1.73 to 37.47 the data is slightly skewed towards right (skewness = 0.90)

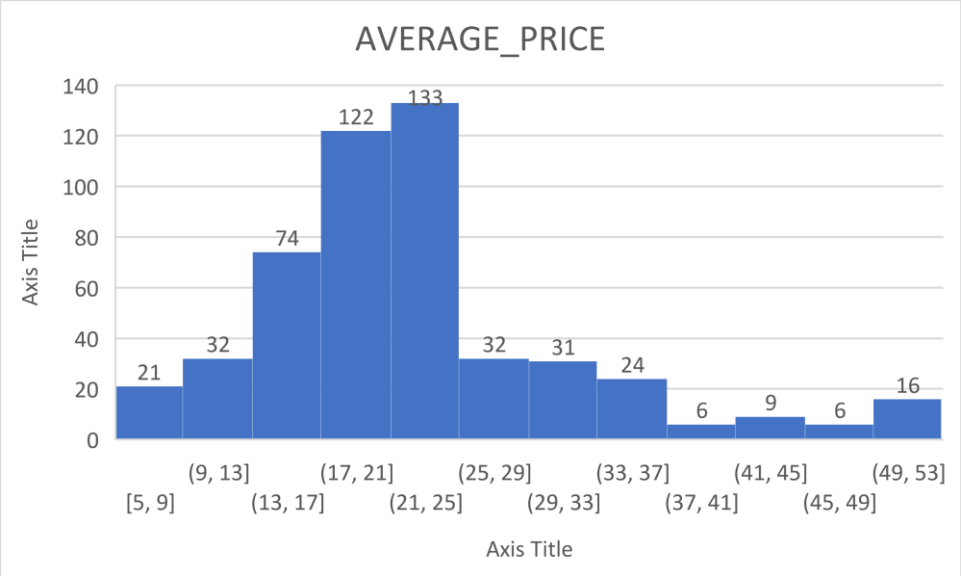
j) Average price= the mean value of owner-occupied homes in $1000's is 22.533 with standard deviation 9.19, whereas sample variance is 84.58 with the range 5 to 50 the data is slightly skewed towards right (skewness = 1.10)

these are the overall of my observation explained each variable.

QUESTION 2

Plot a histogram of the Avg _Price variable. What do you infer?

- To generate histogram
- Click on insert button.
- Click on the chart and click into histogram and plot the chart ok on average_ price.



The above graph observation is x axis is average price whereas y axis is frequency by interpreting the histogram we

can see that data there is overall 133 average prices as more in graph which is price falling under 20000 to 25000.which is maximum number of prices which is 133.

QUESTION 3

Compute the covariance matrix. Share your observations.

- Select the data and data analysis.
- Click on the covariance option.
- Select all range for the input range and click on labels in first row and where to put output range then click ok.

	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	-34.515101	-0.539694518	0.492695216		
LSTAT	-0.882680362	120.8384405	29.52181125	0.487979871	30.32539213	653.4206174	5.771300243	-3.073654967	50.893979	
AVG_PRICE	1.16201224	-97.39615288	-30.46050499	-0.454512407	-30.50083035	-724.820428	-10.09067561	4.484565552	-48.35179	84.4195562

The diagonal values of all are positive, its indicating that all covariance is positive variance.

The diagonal values shows that tax rate as highest variance whereas Age, Indus, NOX, distance as lowest variance.

The off-diagonal values shows that the covariance between the two pairs for example crime rate and LSTAT as negatively

correlated, age and average room is negatively correlated while tax and LSTAT as positively correlated.

The magnitude of the off diagonal values indicates the strength of the correlation between two variables for example the covariance between tax and Indus is much higher than the average room and pt ratio, indicating that tax and Indus are highly correlated whereas average room and pt ratio are poorly correlated

The covariance of average price and average room are highly correlated whereas average room and LSTAT as weakly correlated.

QUESTION 4

Create a correlation matrix of all the variables (Use Data analysis tool pack). (5 marks)

a) Which are the top 3 positively correlated pairs and

b) Which are the top 3 negatively correlated pairs

- Select the data and data analysis.
- Click on the correlation option.
- Select all range for the input range and click on labels in first row and where to put output range then click ok.

Whereas correlation of coefficient ranges from is

+1 is perfect positively correlated.

-1 is negatively correlated.

0 is no correlation which is not correlated with each other

	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	-0.20984667	-0.292047833	-0.355501495	1		
LSTAT	-0.042398321	0.602338529	0.603799716	0.590878921	0.488676335	0.543993412	0.374044317	-0.61380827	1	
AVG_PRICE	0.043337871	-0.376954565	-0.48372516	-0.427320772	-0.38162623	-0.468535934	-0.507786686	0.695359947	-0.73766273	1

a) the top 3 positively correlated pairs are

- Tax and Distance is 0.9102
- NOX and Indus are 0.7636.
- NOX and Age is 0.7314

b) the top 3 negatively correlated pairs

- average price and LSTAT is -0.7376
- LSTAT and average room is -0.6138
- Average price and pt ratio are -0.5077

The graph I represent that +1 is highlighted as green colour.

0 is highlighted as white colour -1 is highlighted as red colour.

I represent graph in such a way that I click on home tab.

Then click on conditional formatting.

Then click on the option new rule

I format 3 colour cell option and gave a number for each cell.

QUESTION 5

Build an initial regression model with AVG_PRICE as ‘y’ (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot. (8 marks)

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?

- Select the data move on to data analysis tool
- Select the option regression analysis
  ‘y’ as dependent variable is average price ‘x’ as independent variable is LSTAT.
- Then select ranges for x and y.
- Click on labels in row.
- Click on residual plot the ok then analysis report will be displayed.

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			

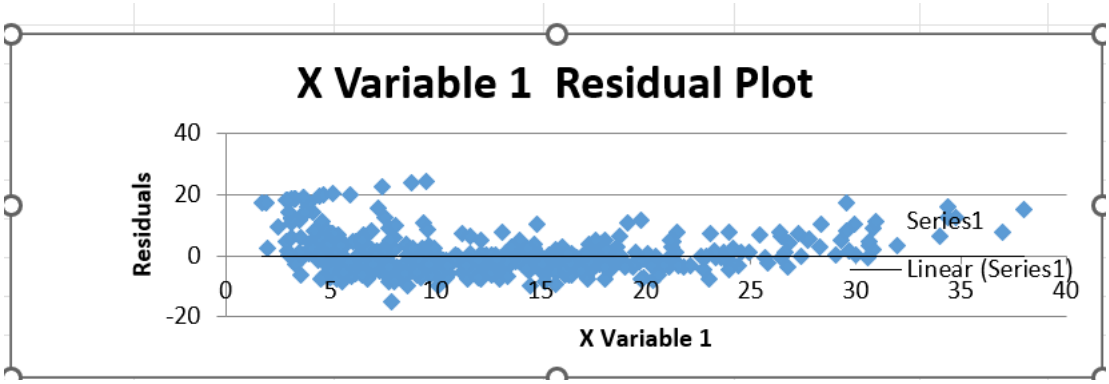
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	34.55384088	0.562627355	61.41514552	3.7431E-236	33.44845704	35.65922472	33.44845704	35.65922472
X Variable 1	-0.950049354	0.038733416	-24.52789985	5.0811E-88	-1.0261482	-0.873950508	-1.0261482	-0.873950508

a) From The regression summary output in terms of variance explained, coefficient value, intercept and the residual plot are

Variance explained: the R squared value is indicates as 0.547 That is 54.7% of the variance is average price is explained by LSTAT variable.

Coefficient value: the coefficient values is -0.95 indicates that every one unit is increase in LSTAT. There is one unit decrease in average price.

Intercept: the intercept value is 34.55 when LSTAT is zero the predicted average price is 34.55



Residual plot: the above graph represents that the residuals are randomly distributed around the zero. That indicates that assumptions of linearity, independence, normality and equality in variance of residuals are satisfied in all the observations.

b) Yes LSTAT is significant based on the model because its as its p values is less than 0.05,which is based on statistical significance the p value for LSTAT is 5.081103394387 85E-88 (i.e,0.00) which is of less than 0.05, which is LSTAT variable has a significant effect on the average variable.

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.

- Select the data move on to data analysis tool.
- Select the option regression analysis
  'y' as dependent variable is average price 'x' as independent variable is LSTAT and average room.
- Then select ranges for x and y.
- Click on labels in row.
- Click on the ok then regression analysis report will be displayed.

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	-0.428095348	0.668764941	-7.591900282	4.875354658	-7.59190028	4.875354658
AVG_ROOM	5.094787984	0.4444655	11.46272991	3.47226E-27	4.221550436	5.968025533	4.22155044	5.968025533
LSTAT	-0.642358334	0.043731465	-14.68869925	6.66937E-41	-0.728277167	-0.556439501	-0.72827717	-0.5564395

a) The regression equation is

Y= a+bX

Y = b1x1+b2x2…….bnxn

- 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?

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

=21.45

The company is quoting in 30000USD for this locality which is much higher than the predicted value 21.45USD. therefore, company is overcharging.

b)to compare this model and previous model we are taking the adjusted R as comparison which is of previous model contains the adjusted R squared value is 0.54 & this model contains the adjusted R squared value is 0.63 where as this model contains higher than the previous model where as this indicates that the new model is better explaining of the variance of the average price than the previous model.

QUESTION7

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.

- Select the data move on to data analysis tool.
- Select the option regression analysis
  'y' as dependent variable is average price 'x' as independent variable is Crime Rate, Age, Indus, NOX, Distance, Tax, Pt ratio, Average Room & LSTAT
- Then select ranges for x and y.
- Click on labels in row.
- Click on the ok then regression analysis report will be displayed.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372
Adjusted R Squar	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	38.70580267	19.77682784	38.70580267
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.534657201	-0.105348544	0.202798827	-0.10534854	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	-17.9720228	-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	-0.02207388	-0.0067285
PTRATIO	-1.074305348	0.133601722	-8.041104061	6.58642E-15	-1.336800438	-0.811810259	-1.33680044	-0.811810259
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.89287E-19	3.255494742	4.995323561	3.255494742	4.995323561
LSTAT	-0.603486589	0.053081161	-11.36912937	8.91071E-27	-0.70777824	-0.499194938	-0.70777824	-0.499194938

The adjusted R square of this model is 0.688, which is the model explain about 68.8% this is slight improvement over the previous model.

The intercept of the value is 29.24 which is all independent variable is zero then the predicted value of y=average price is 29.24.

The coefficient represents the impact on each independent variable on the dependent variable when other variable is constant held up.

- The crime rate of coefficient is 0.049 which is crime rate is leads 1 unit which is increase by 0.049 in predicted value of average price, how ever p value is greater than 0.05, indicating this variable is not statistically significant.
- Age has coefficient is 0.03 which is leads to increase in age 1year which is increase by 0.03 in predicted value of average price, however the p-value is less than 0.05, indicating this variable which is statistically significant.
- Indus has coefficient of 0.131 which indicates that increase in proportion to the non-retail business acres per town by 1 lead to 0.131 in predicted value of average price, however the p value is less than 0.05, indicating this variable which is statistically significant.
- NOX has coefficient of -10.321 which is leads to increase nitric oxide of the concentration by 1 unit leads to -10.321 which decrease in predicted value of average price, however the p value is less than 0.05, indicating this variable which Is statistically significant.

- Distance of coefficient 0.261 which is increase in weighted distance from highway Boston employment by 1 unit leads to 0.261 is increase in predicted value of average price however the p value is less than 0.05, indicating this variable which is statistically significant.
- Tax of the coefficient -0.014 indicating that increase in full-value property rate per $10000 by 1unit leads to -0.014 is decrease in predicted value of average price, however the p value is less than 0.05, indicating that variable which is statistically significant.
- Pt ratio has a coefficient -1.074 indicating that increase in the pupil teacher ratio by 1 unit leads to -1.074 is decrease in predicted value of average price, however the p value is less than 0.05, indicating that variable is statistically significant.
- Average room of coefficient is 4.124 indicating that increase in the overall average rooms are dwelling by 1 unit leads to 4.125 is increase in predicted value of average price, however the p value is less than 0.05, indicating that variables is statistically significant.
- LSTAT has coefficient of -0.603 indicating that increase in percentage of lower status population by 1 unit which is leads to -0.603 which is decrease in predicated value of average price, whereas p value is less than 0.05, indicating that variables is statistically significant.

Overall the model indicates that average room as strongest impact on an average price where as the NOX as the lowest

impact on an average price in the model. The crime rate is not significant rest of others are statistically significant.

QUESTION 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.

- Select the data move on to data analysis tool.
- Select the option regression analysis
  'y' as dependent variable is average price 'x' as independent variable is age, Indus, NOX, distance, tax, pt ratio, LSTAT and average room
- Then select ranges for x and y.
- Click on labels in row.
- Click on the ok then regression analysis report will be displayed.

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	3703.585178	140.6430411	1.911E-122			
Residual	497	13087.61399	26.33322735					
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	6.124898157	1.84597E-09	19.98838959	38.8685574	19.98838959	38.8685574
AGE	0.03293496	0.013087055	2.516605952	0.012162875	0.007222187	0.058647734	0.007222187	0.058647734
INDUS	0.130710007	0.063077823	2.072202264	0.038761669	0.006777942	0.254642071	0.006777942	0.254642071
NOX	-10.27270508	3.890849222	-2.640221837	0.008545718	-17.9172457	-2.628164466	-17.9172457	-2.628164466
DISTANCE	0.261506423	0.067901841	3.851242024	0.000132887	0.128096375	0.394916471	0.128096375	0.394916471
TAX	-0.014452345	0.003901877	-3.703946406	0.000236072	-0.022118553	-0.006786137	-0.022118553	-0.006786137
PTRATIO	-1.071702473	0.133453529	-8.030529271	7.08251E-15	-1.333905109	-0.809499836	-1.333905109	-0.809499836
AVG_ROOM	4.125468959	0.44248544	9.323400461	3.68969E-19	3.256096304	4.994841615	3.256096304	4.994841615
LSTAT	-0.605159282	0.0529801	-11.42238841	5.41844E-27	-0.70925186	-0.501066704	-0.70925186	-0.501066704

- Interpretation of the output
  - Multiple R is 0.833 which is strong positive in correlation between the independent variable and dependent variable
  - The R squared value is 0.693 which is 69.3% which is variance is dependent variable explained by the independent variable
  - The adjusted R squared model is 0.688 for this previous model as same as this model.
  - The standard error is 5.13 which means that the predicted value on an average of 5.13 units are away from the actual values.
  - The anova model shows that regression model is significant ,with an f statistics 140.64 then the significance p value is 122 which is low.

- The coefficient represents impact on each independent variables on the dependent variable which is of interpretation is below

- Age has coefficient is 0.03 which is leads to increase in age 1year which is increase by 0.03 in predicted value of average price
- Indus has coefficient of 0.131 which indicates that increase in proportion to the non-retail business acres per town by 1 lead to 0.131 in predicted value of average price
- NOX has coefficient of -10.321 which leads to increase nitric oxide of the concentration by 1 unit leads to -10.321 which decrease in predicted value of average price.
- Distance of coefficient 0.261 which is increase in weighted distance from highway Boston employment by 1 unit leads to 0.261 is increase in predicted value of average price
- Tax of the coefficient -0.014 indicating that increase in full-value property rate per $10000 by 1unit leads to -0.014 is decrease in predicted value of average price.
- Pt ratio has a coefficient -1.074 indicating that increase in the pupil teacher ratio by 1 unit leads to -1.074 decrease in predicted value of average price.
- Average room of coefficient is 4.124 indicating that increase in the overall average rooms is dwelling by 1 unit leads to 4.125 is increase in predicted value of average price
- LSTAT has coefficient of -0.603 indicating that increase in percentage of lower status population by 1 unit which is leads to -0.603 which is decrease in predicated value of average price.

- the adjusted R-square value of this model is 0.688 with the model in the previous question adjusted R-Square value of this model is 0.688, the model performs better according to the value of adjusted R-square because both adjusted R value is same so there is no difference in value so performance better.

c)

- Firstly, I selected the coefficient value and copied and paste to another worksheet.
- Then I selected one cell right click select sort option and select smallest to highest. The coefficient of ascending order table is shown below.

NOX	-10.2727
PTRATIO	-1.0717
LSTAT	-0.60516
TAX	-0.01445
AGE	0.032935
INDUS	0.13071
DISTANCE	0.261506
AVG_ROO	4.125469

If the value of NOX is more in locality in this town is the average of house is decrease by 10.27 per unit increase in NOX

d)the regression equation in this model is.

Y=intercept+b1x1+b2x2+b3x3……..bnxn

Average Price = 29.428+0.033(age)+0.131(indus)+10.27(nox)+0.26(distance)-0.014(tax)-1.0717(Ptratio)+4.125(average room)-0.605(LSTAT)

Thank you!