PREPARED BY – UJJWAL AWANA
BATCH – GLDA SEP 2023

TABLE OF CONTENTS

Q1. The summary statistics for each variable in the table. Q1

Q2. A histogram of the Average Price variable. Q2

Q3. The covariance matrix. Q3

Q4. A correlation matrix of all the variables. Q4

Q5. Initial regression model with AVG_PRICE as **'y'** (Dependent variable) and LSTAT variable as Independent Variable. Q5

Q6. A new Regression model including LSTAT and AVG_ROOM together as Independent variables and AVG_PRICE as dependent variable. Q6

Q7. Another Regression model with all variables where AVG_PRICE alone be the Dependent Variable and all the other variables are independent. Q7

Q8. Regression Model with all Significant Variables. Q8

	CRIME_RATE	Age	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
Mean	4.871976285	68.5749	11.13678	0.554695	9.549407	408.2372	18.45553	6.2846344	12.65306	22.532806
Standard Error	0.129860152	1.25137	0.30498	0.005151	0.387085	7.492389	0.096244	0.0312351	0.317459	0.4088611
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	28.14886	6.860353	0.115878	8.707259	168.5371	2.164946	0.7026171	7.141062	9.1971041
Sample Variance	8.533011532	792.3584	47.06444	0.013428	75.81637	28404.76	4.686989	0.4936709	50.99476	84.586724
Kurtosis	-1.18912246	-0.96772	-1.23354	-0.06467	-0.86723	-1.14241	-0.28509	1.8915004	0.49324	1.4951969
Skewness	0.021728079	-0.59896	0.295022	0.729308	1.004815	0.669956	-0.80232	0.4036121	0.90646	1.1080984
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
CV	1.667838517	2.436152	1.623354	4.786902	1.096718	2.422239	8.52471	8.9446072	1.771874	2.4499893

OBSERVATION:

a) Summarizing the data based on our inputs we have Mean , median , mode , Standard Variation and skewness.
b) In terms of skewness we can find that Average Price is the most positively skewed and Crime rate has the least positively skewness.
c) Whereas PTRATIO is the most negative skewness.
d) As data was not on the same scale. In this case instead of checking SD , we have calculated Coefficient of Variation (CV).

$$CV = MEAN/SD$$

e) In this case Avg room has the highest CV and DISTANCE variable has the lowest. The lower the CV the better it is.

OBSERVATION:

By Observing from the above histogram we can figure it out about the

Highest and Lowest frequency of the Average Prices of houses.

a) The houses which are in bracket of 21000-25000 USD have the highest number of count or frequency.

b) Whereas the houses in the bracket of 37000-41000 USD and 45000-49000 USD have the lowest frequency.

	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	46.97143							
NOX	0.000625308	2.381211931	0.605874	0.013401						
DISTANCE	-0.229860488	111.5499555	35.47971	0.61571	75.66653					
TAX	-8.229322439	2397.941723	831.7133	13.0205	1333.117	28348.62				
PTRATIO	0.068168906	15.90542545	5.680855	0.047304	8.743402	167.8208	4.677726			
AVG_ROOM	0.056117778	-4.74253803	-1.88423	-0.02455	-1.28128	-34.5151	-0.53969	0.492695216		
LSTAT	-0.882680362	120.8384405	29.52181	0.48798	30.32539	653.4206	5.7713	-3.073654967	50.89398	
AVG_PRICE	1.16201224	-97.39615288	-30.4605	-0.45451	-30.5008	-724.82	-10.0907	4.484565552	-48.3518	84.41955616

a) Here we have observed that the Relationship which are highlighted in Green are the relationship which are having a Positive relationship / +ve relation.

b) Whereas the Relationship which are highlighted in Red are having a negative relationship/-ve relation.

c) Covariance gives the Direction of relationship but is not able to gives the strength like by correlation.

d) Having a +ve relation means both `X` and `Y` moves together in the same direction. So in this case relations which are highlighted in green moves in the same direction.

e) Having a -ve relation means both `X` and `Y` moves in opposite direction. In this case relations are highlighted by red colour.

	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	0.76365145	1						
DISTANCE	-0.009055049	0.456022452	0.59512927	0.6114406	1					
TAX	-0.016748522	0.506455594	0.72076018	0.6680232	0.910228189	1				
PTRATIO	0.010800586	0.261515012	0.38324756	0.1889327	0.464741179	0.460853	1			
AVG_ROOM	0.02739616	-0.24026493	-0.39167585	-0.3021882	-0.20984667	-0.29205	-0.3555	1		
LSTAT	-0.042398321	0.602338529	0.60379972	0.5908789	0.488676335	0.543993	0.374044	-0.613808272	1	
AVG_PRICE	0.043337871	-0.37695457	-0.48372516	-0.4273208	-0.38162623	-0.46854	-0.50779	0.695359947	-0.73766	1

a) Top 3 correlated pairs are highlighted in green font colour.

AGE-NOX INDUS-NOX DISTANCE-TAX

b) Bottom 3 correlated pairs are highlighted in red font colour

AVG ROOM-LSTAT PTRATIO-AVG PRICE LSTAT-AVG PRICE

c) In the TOP 3 correlated pairs

DISTANCE-TAX have the strongest +ve relationship.

d) In the Bottom 3 correlated pairs

LSAT-AVG PRICE have the weakest -ve relationship.

e) Correlation helps in finding out both the direction and strength.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.737662726
R Square	0.544146298 NOT MET
Adjusted R Square	0.543241826
Standard Error	6.215760405
Observations	506

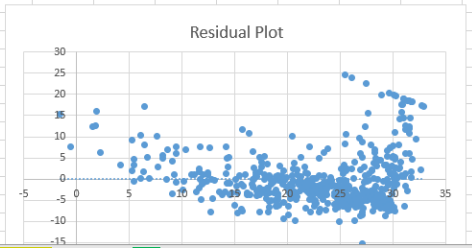
ANOVA

	df	SS	MS	F	Significance F
Regression	1	23243.91	23243.91	601.6178711	5.0811E-88
Residual	504	19472.38	38.63568		
Total	505	42716.3			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	pper 95.0%
Intercept	34.55384088	0.562627	61.41515	3.7431E-236	33.448457	35.65922472	33.44845704	35.65922
LSTAT	-0.950049354	0.038733	-24.5279	5.0811E-88	-1.0261482	-0.87395051	-1.0261482	-0.87395

RESIDUAL OUTPUT

Observation	Predicted AVG_PRICE	Residuals	RMSE SQ	MEAN	RMSE	MAX ERROR	%	
1	29.8225951	-5.8226	33.90261	38.48296723	6.20346413	22.53280632	27.53081015	Not Met
2	25.87038979	-4.27039	18.23623					
3	30.72514198	3.974858	15.7995					
4	31.76069578	1.639304	2.687318					
5	29.49007782	6.709922	45.02306		Assumptions of Residuals			
6	29.60408375	-0.90408	0.817367		1	mean	-2.7365E-14	Met
7	22.74472741	0.155273	0.02411		2	Distribution	1.457061987	Not Met
8	16.36039575	10.7396	115.3391		3	const. Var	no relation	Met
9	6.118863721	10.38114	107.768					
10	18.30799693	0.592003	0.350468					
11	15.1253316	-0.12533	0.015708					
12	21.94668596	-3.04669	9.282295					
13	19.62856553	2.071434	4.290841					



Observation:

a) Build a regression model with AVG_PRICE as ‘y’ (Dependent variable) and LSTAT variable as an Independent Variable.

b) The Steps to look that we can use this linear regression model has not been fulfilled.

c) R Square is less than 60%.

Also RMSE is 27.53% which is more than .10 or Max possible error.

And all assumptions of the Residual does not met.

1.Mean of the residual is 0

2.Residual have constant variance as there was no relation.

3.Residual was not normally distributed as skewness was more than .50

Residual plot was not having any relation to it. The graph was showing a parabolic curve. Trendline was also flat.

B) Yes LSTAT is a significant independent variable as Pvalue is less than 5%.

SUMMARY OUTPUT

Regression Statistics									
Multiple R	0.799100498					a)	Y = MX + C	21.45808	
R Square	0.638561606	Met					Company is charging 30000 USD. Whereas by using multiple linear regression we got 21000 USD.		
Adjusted R Square	0.637124475						So Company is Overcharging.		
Standard Error	5.540257367								
Observations	506					b)	Yes the performance is better than previous question as in that Q we got R Square value less thar		

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	2	27276.99	13638.49	444.3308922	7E-112
Residual	503	15439.31	30.69445		
Total	505	42716.3			

	Coefficients	*andard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.358272812	3.172828	-0.4281	0.668764941	-7.5919	4.875354658	-7.591900282	4.875355
AVG_ROOM	5.094787984	0.444466	11.46273	3.47226E-27	4.22155	5.968025533	4.221550436	5.968026
LSTAT	-0.642358334	0.043731	-14.6887	6.66937E-41	-0.72828	-0.5564395	-0.728277167	-0.55644

RESIDUAL OUTPUT

Observation	*Predicted AVG_PRICE*	*Residuals*	rmse *Sq*	*mean*	*RMSE*	*Actual Y*	*Max Possible Error*	
1	28.94101368	-4.94101	24.41362	30.51246878	5.523809	22.53280632	24.51451978	Not Met
2	25.48420566	-3.88421	15.08705					

Observations:

a) Y = MX + C

=(B18*7)+(B19*20)+B17 = 21.46 , here B18 = Coefficient of Avg Room

B19 = Coefficient of LSTAT

C = Value of Y when X is 0.

Company is charging 30000 USD. Whereas by using multiple linear regression we got around 21000 USD.

So Company is Overcharging

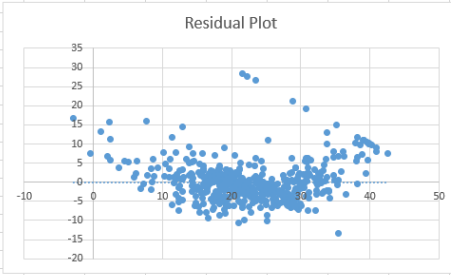
b) Yes the performance is better than previous question as in Q5 we got R Square value less than 60%. But in this case R Sqaure value is more than 60% and the difference between R Square and Adj R Square is less than 1%..

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

ANOVA

	df	SS	MS	F	gnificance F
Regression	9	29638.86	3293.207	124.9045049	1.9E-121
Residual	496	13077.43	26.3658		
Total	505	42716.3			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	29.24131526	4.817126	6.070283	2.53978E-09	19.77683	38.7058	19.77682784	38.70580267
CRIME_RATE	0.048725141	0.078419	0.621346	0.534657201	-0.10535	0.202799	-0.105348544	0.202798827
AGE	0.032770689	0.013098	2.501997	0.012670437	0.007037	0.058505	0.00703665	0.058504728
INDUS	0.130551399	0.063117	2.068392	0.03912086	0.006541	0.254562	0.006541094	0.254561704
NOX	-10.3211828	3.894036	-2.65051	0.008293859	-17.972	-2.67034	-17.97202279	-2.670342809
DISTANCE	0.261093575	0.067947	3.842603	0.000137546	0.127594	0.394593	0.127594012	0.394593138
TAX	-0.01440119	0.003905	-3.68774	0.000251247	-0.02207	-0.00673	-0.022073881	-0.0067285
PTRATIO	-1.074305348	0.133602	-8.0411	6.58642E-15	-1.3368	-0.81181	-1.336800438	-0.811810259
AVG_ROOM	4.125409152	0.442759	9.317505	3.89287E-19	3.255495	4.995324	3.255494742	4.995323561
LSTAT	-0.603486589	0.053081	-11.3691	8.91071E-27	-0.70778	-0.49919	-0.70777824	-0.499194938



RESIDUAL OUTPUT

Observation	Predicted AVG_PRICE	Residuals	RMSE Sq	Mean	rmse	Actual Y	Max possible error	
1	30.1153558	-6.11536	37.39758	25.84473304	5.083772	22.53281	22.56164396	Not Met
2	27.00714024	-5.40714	29.23717					
3	32.83291255	1.867087	3.486016					
4	31.20703392	2.192966	4.8091					
5	30.5947288	5.605271	31.41907					
6	28.07644731	0.623553	0.388818					
7	25.29985158	-2.39985	5.759288					

	Assumptions of Residual		
1	mean	-1.49586E-14	Yes
2	Normal Distribution	1.649649848	No
3	Constant Variance	no relation	Yes

Observations:

a) Regression model with all variables where AVG_PRICE alone be the Dependent Variable and all the other variables are independent.

b)R Square is 69% which is more than 60% which met our first step.

c) Also RMSE is 22.5% which is more than .10 or Max possible error.

d) And all assumptions of the Residual does not met.

1.Mean of the residual is 0

2.Residual have constant variance as there was no relation.

3.Residual was not normally distributed as skewness was more than .50

e) Crime rate variable is insignificant as it has Pvalue more than 0.05 , it means this independent variable does not related to the dependent variable ‘y’. These factors are also known as Noise or Redundant factors.

f) All other independent variables like AGE , INDUS , NOX , DISTANCE , TAX , PTRATIO , AVG ROOM , LSTAT are Significant variables and are relevant as each one of them has PVALUE lower than 0.05.

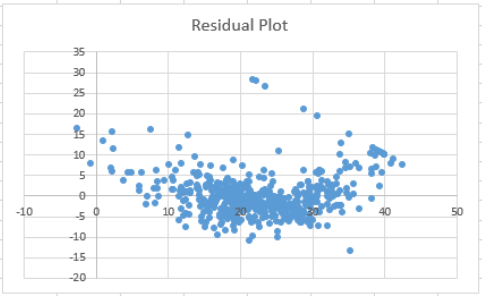
SUMMARY OUTPUT

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	gnificance F
Regression	8	29628.7	3703.59	140.643041	2E-122
Residual	497	13087.6	26.3332		
Total	505	42716.3			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	29.42847349	4.80473	6.1249	1.846E-09	19.9884	38.8686	19.9884	38.8686
AGE	0.03293496	0.01309	2.51661	0.01216288	0.00722	0.05865	0.00722	0.05865
INDUS	0.130710007	0.06308	2.0722	0.03876167	0.00678	0.25464	0.00678	0.25464
NOX	-10.27270508	3.89085	-2.6402	0.00854572	-17.917	-2.6282	-17.917	-2.6282
DISTANCE	0.261506423	0.0679	3.85124	0.00013289	0.1281	0.39492	0.1281	0.39492
TAX	-0.014452345	0.0039	-3.7039	0.00023607	-0.0221	-0.0068	-0.0221	-0.0068
PTRATIO	-1.071702473	0.13345	-8.0305	7.0825E-15	-1.3339	-0.8095	-1.3339	-0.8095
AVG_ROOM	4.125468959	0.44249	9.3234	3.6897E-19	3.2561	4.99484	3.2561	4.99484
LSTAT	-0.605159282	0.05298	-11.422	5.4184E-27	-0.7093	-0.5011	-0.7093	-0.5011



RESIDUAL OUTPUT

Observation	Predicted AVG_PRICE	Residuals	Sq	Mean	rmse Root	Actual Y	Max Error	
1	30.04888734	-6.0489	36.589	25.8648498	5.08575	22.5328	22.5704	Not Met
2	27.04098462	-5.441	29.6043					
3	32.69896454	2.00104	4.00414					
4	31.14306949	2.25693	5.09374					
5	30.58808735	5.61191	31.4936					
6	27.85095254	0.84905	0.72088					
7	25.07089688	-2.1709	4.71279					
8	22.63588287	4.46412	19.9283					
9	14.00883345	2.49117	6.20591					

Assumptions of Residual

1	Mean	-1.0395E-14	Yes
2	Distributi	1.64386951	No
3	Constant	no relation	Yes

Observations:

1)After putting only Significant Variables from the Q7 we have this Regression Model where all the independent variables are significant.

a) R Square is less than 60%.
b) Also RMSE is 22.57% which is more than 10% or Max possible error.
c) And all assumptions of the Residual does not met.
d) Mean of the residual is 0
e) Residual have constant variance as there was no relation.
f) Residual was not normally distributed as skewness was more than .50

2) By comparing R square & adj R square of this model with the model in the previous question. There was not much of a difference as both are same even if we have removed an insignificant variable which is CRIME RATE.

3)

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
NOX	-10.27270508	3.890849	-2.64022	0.008545718	-17.9172	-2.62816	-17.9172	-2.62816
PTRATIO	-1.071702473	0.133454	-8.03053	7.08251E-15	-1.33391	-0.8095	-1.33391	-0.8095
LSTAT	-0.605159282	0.05298	-11.4224	5.41844E-27	-0.70925	-0.50107	-0.70925	-0.50107
TAX	-0.014452345	0.003902	-3.70395	0.000236072	-0.02212	-0.00679	-0.02212	-0.00679
AGE	0.03293496	0.013087	2.516606	0.012162875	0.007222	0.058648	0.007222	0.058648
INDUS	0.130710007	0.063078	2.072202	0.038761669	0.006778	0.254642	0.006778	0.254642
DISTANCE	0.261506423	0.067902	3.851242	0.000132887	0.128096	0.394916	0.128096	0.394916
AVG_ROOM	4.125468959	0.442485	9.3234	3.68969E-19	3.256096	4.994842	3.256096	4.994842
Intercept	29.42847349	4.804729	6.124898	1.84597E-09	19.98839	38.86856	19.98839	38.86856

	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
AGE	1								
INDUS	0.644779	1							
NOX	0.73147	0.763651	1						
DISTANCE	0.456022	0.595129	0.611441	1					
TAX	0.506456	0.72076	0.668023	0.910228	1				
PTRATIO	0.261515	0.383248	0.188933	0.464741	0.460853	1			
AVG_ROOM	-0.24026	-0.39168	-0.30219	-0.20985	-0.29205	-0.3555	1		
LSTAT	0.602339	0.6038	0.590879	0.488676	0.543993	0.374044	-0.61381	1	
AVG_PRICE	-0.37695	-0.48373	-0.42732	-0.38163	-0.46854	-0.50779	0.69536	-0.73766	1

For finding out Strength and Direction in a relationship we have calculated Corelation.

Here , we can observe that NOX and Average Price have a negative relation.

So when NOX in a locality will increase it will decrease the AVERAGE PRICE in the Town.

Regression Equation of this model is

$$Y = MX + C$$

Y = {(-10.27270508*NOX) + (-1.071702473*PTRATIO) + (-0.605159282*LSTAT) + (-0.014452345*TAX) + (0.03293496*AGE) + (0.130710007*INDUS) + (0.261506423*DISTANCE) + (4.125468959*AVERAGE ROOM) + 29.42847349}