Real Estate Report

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

CRIME_RATE		***AGE***	
Mean	4.871976285	Mean	68.57490119
Standard Error	0.129860152	Standard Error	1.251369525
Median	4.82	Median	77.5
Mode	3.43	Mode	100
Standard Deviation	2.921131892	Standard Deviation	28.14886141
Sample Variance	8.533011532	Sample Variance	792.3583985
Kurtosis	-1.189122464	Kurtosis	-0.967715594
Skewness	0.021728079	Skewness	-0.59896264
Range	9.95	Range	97.1
Minimum	0.04	Minimum	2.9
Maximum	9.99	Maximum	100
Sum	2465.22	Sum	34698.9
Count	506	Count	506
INDUS		***NOX***	
Mean	11.13677866	Mean	0.554695059
Standard Error	0.304979888	Standard Error	0.005151391
Median	9.69	Median	0.538
Mode	18.1	Mode	0.538
Standard Deviation	6.860352941	Standard Deviation	0.115877676
Sample Variance	47.06444247	Sample Variance	0.013427636
Kurtosis	-1.233539601	Kurtosis	-0.064667133
Skewness	0.295021568	Skewness	0.729307923
Range	27.28	Range	0.486
Minimum	0.46	Minimum	0.385
Maximum	27.74	Maximum	0.871
Sum	5635.21	Sum	280.6757
Count	506	Count	506

DISTANCE		***TAX***	
Mean	9.549407115	Mean	408.2371542
Standard Error	0.387084894	Standard Error	7.492388692

Median	5	Median	330
Mode	24	Mode	666
Standard Deviation	8.707259384	Standard Deviation	168.5371161
Sample Variance	75.81636598	Sample Variance	28404.75949
Kurtosis	-0.867231994	Kurtosis	-1.142407992
Skewness	1.004814648	Skewness	0.669955942
Range	23	Range	524
Minimum	1	Minimum	187
Maximum	24	Maximum	711
Sum	4832	Sum	206568
Count	506	Count	506

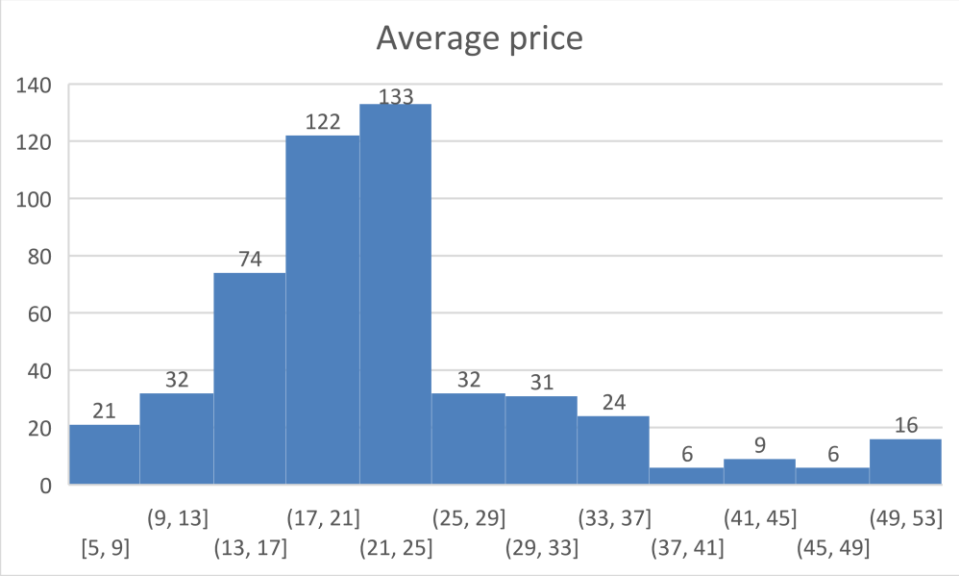
PTRATIO		***AVG_ROOM***	
Mean	18.4555336	Mean	6.284634387
Standard Error	0.096243568	Standard Error	0.031235142
Median	19.05	Median	6.2085
Mode	20.2	Mode	5.713
Standard Deviation	2.164945524	Standard Deviation	0.702617143
Sample Variance	4.686989121	Sample Variance	0.49367085
Kurtosis	-0.285091383	Kurtosis	1.891500366
Skewness	-0.802324927	Skewness	0.403612133
Range	9.4	Range	5.219
Minimum	12.6	Minimum	3.561
Maximum	22	Maximum	8.78
Sum	9338.5	Sum	3180.025
Count	506	Count	506

LSTAT		***AVG_PRICE***	
Mean	12.65306324	Mean	22.53280632
Standard Error	0.317458906	Standard Error	0.408861147
Median	11.36	Median	21.2
Mode	8.05	Mode	50
Standard Deviation	7.141061511	Standard Deviation	9.197104087
Sample Variance	50.99475951	Sample Variance	84.58672359
Kurtosis	0.493239517	Kurtosis	1.495196944

Skewness	0.906460094	Skewness	1.108098408
Range	36.24	Range	45
Minimum	1.73	Minimum	5
Maximum	37.97	Maximum	50
Sum	6402.45	Sum	11401.6
Count	506	Count	506

- The mean and median values of tax is farther as the outskirts values are high it can impact mean value. So here we can go with median value for central tendency,
- Skewness for average price is 1.108 so it depicts that most of the values are present to the left of the average values.
- Avg_room has kurtosis value of 1.891 which represents high peakedness and indus has low value of -1.23 which represents its platy kurtosis.
- Range of tax is 524 which shows the high outskirts value.
- Average price has standard deviation of 9.19 where the spread is high.

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



AVG_PRICE	
Mean	22.53280632
Standard Error	0.408861147
Median	21.2

Mode	50
Standard Deviation	9.197104087
Sample Variance	84.58672359
Kurtosis	1.495196944
Skewness	1.108098408
Range	45
Minimum	5
Maximum	50
Sum	11401.6
Count	506

Here we can see that mean and median are almost closer to each other.Also we can see that most of the data present to the left of average value, so positive skewness and also shows high peakedness.

3) Compute the covariance matrix. Share your observations.

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	8.5161479									
AGE	0.5629152	790.79247								
INDUS	-0.110215	124.26783	46.97143							
NOX	0.0006253	2.3812119	0.6058739	0.0134011						
DISTANCE	-0.22986	111.54996	35.479714	0.6157 1022	75.666531					
TAX	-8.229322	2397.9417	831.71333	13.0205024	1333.1167	28348.6236				
PTRATIO	0.0681689	15.905425	5.6808548	0.0473 0365	8.7434025	167.820822	4.6777263			
AVG_ROOM	0.0561178	-4.742538	-1.882254	-0.024554 83	-1.281277	-34.515101	-0.539695	0.4926952		
LSTAT	-0.88268	120.83844	29.521811	0.4879 7987	30.325392	653.420617	5.7713002	-3.073655	50.893979	
AVG_PRICE	1.1620122	-97.39615	-30.460505	-0.454512 41	-30.50083	-724.82043	-10.09068	4.4845656	-48.351792	84.41956

Here in the above table, average price vs crime rate has positive relationship and **average price vs average** room has **positive relationship**. **Tax vs average price** has **strong negative relationship**. The **other X variables** with average price has **negative relationship.**

4) Create a correlation matrix of all the variables (Use Data analysis tool pack).
a) Which are the top 3 positively correlated pairs and

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	1									
AGE	0.006859946	1								
INDUS	-0.005510778	0.644778511	1							
NOX	0.001805098	0.731470104	0.763651447	1						
DISTANCE	-0.009055	0.456022452	0.595129275	0.611440563	1					
TAX	-0.016785485	0.506455594	0.720766018	0.668085232	0.910228189	1				
PTRATIO	0.010860059	0.261515012	0.383247556	0.188932677	0.464741179	0.460853035	1			
AVG_ROOM	-0.027369616	-0.240264931	-0.391675853	-0.302188188	-0.209846668	-0.292047833	-0.355901495	1		
LSTAT	0.042383983	-0.602338529	0.603799716	0.590878921	-0.488676335	0.543993412	0.374044317	-0.613808272	1	
AVG_PRICE	-0.043373787	0.376954565	-0.483752516	-0.427320772	0.381626231	-0.468535934	-0.507786686	0.695359947	-0.737662726	1

b) Which are the top 3 negatively correlated pairs.

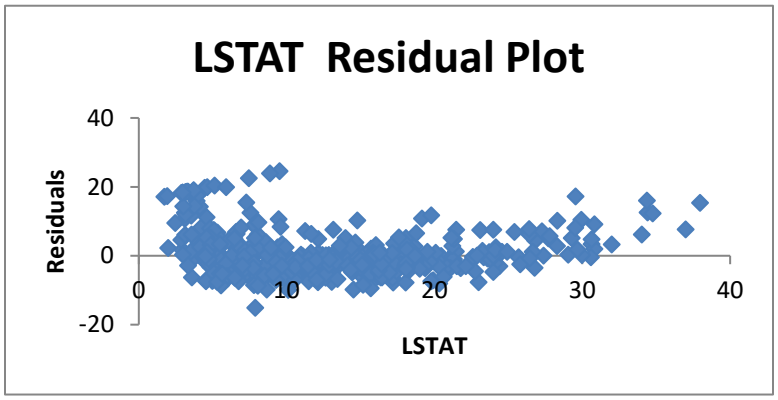
Top 3 positive		Top 3 negative	
Tax Vs Distance	0.910228189	Avg prive vs LSTAT	-0.737662726
NOX vs Indus	0.763651447	LSTAT vs average room	-0.613808272
NOX vs Age	0.731470104	Avgprice vs Ptratio	-0.507786686

From the above table we can see that age vs other X variables are highly correlated.

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?

	Coefficients	Standard Error	t Stat	P-value
Intercept	34.55384088	0.562627355	61.41514552	3.7431E-236
LSTAT	-0.950049354	0.038733416	-24.52789985	5.0811E-88

Adjusted R Square	0.543241826



From the above regression, we get adjusted R^2 of 0.54, which is somewhat low value to get explained Y by X values.

The coefficient of LSTAT is in **negative** which means Y value decreases with LSTAT.

Intercept is 34.558 which indicates that when X is 0, Y takes the respective value.

Residual plot shows that no traces of patterns.

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

The regression model gives a **p-value** of **5.08110339438785E-88** which is significant for 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?

	Coefficients	Standard Error	t Stat	P-value
Intercept	-1.358272812	3.17282778	-0.428095348	0.668764941
AVG_ROOM	5.094787984	0.4444655	11.46272991	3.47226E-27
LSTAT	-0.642358334	0.043731465	-14.68869925	6.66937E-41
Adjusted R Square	0.637124475			

Average price Y=-1.358272812 +5.094787984*7(Avg_room) + -0.642358334*20(LSTAT)

Average price
21.45808
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.

From the Adjusted R^2 we get to know that previous model has value of 0.543 which is less than 0.637

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.

	Coefficients	Standard Error	t Stat	P-value
Intercept	29.24131526	4.817125596	6.070282926	2.53978E-09
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.534657201
AGE	0.032770689	0.013097814	2.501996817	0.012670437
INDUS	0.130551399	0.063117334	2.068392165	0.03912086
NOX	-10.3211828	3.894036256	-2.650510195	0.008293859
DISTANCE	0.261093575	0.067947067	3.842602576	0.000137546
TAX	-0.01440119	0.003905158	-3.687736063	0.000251247

PTRATIO	-1.074305348	0.133601722	-8.041104061	6.58642E-15
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.89287E-19
LSTAT	-0.603486589	0.053081161	-11.36912937	8.91071E-27

Adjusted R Square	0.688298647

- From the above analysis we can see that adjusted R^2 value is nearer to 1.
- We can also infer that Avg_room has highest coefficient of 4.125 and Lstat has lowest coefficient of -0.603.
- Intercept value is 29.241 which means average price will be 29.241 when all X variables are zero.
- The p-value of crime rate is 0.534 which is not significant.
- The other X variable such as Age,Indus,NOX,distance,tax,ptratio,avg_room are **significant** as their values are less than 0.05 .

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.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	29.42847349	4.804728624	6.124898157	1.84597E-09
AGE	0.03293496	0.013087055	2.516605952	0.012162875
INDUS	0.130710007	0.063077823	2.072202264	0.038761669
NOX	-10.27270508	3.890849222	-2.640221837	0.008545718
DISTANCE	0.261506423	0.067901841	3.851242024	0.000132887
TAX	-0.014452345	0.003901877	-3.703946406	0.000236072
PTRATIO	-1.071702473	0.133453529	-8.030529271	7.08251E-15
AVG_ROOM	4.125468959	0.44248544	9.323400461	3.68969E-19
LSTAT	-0.605159282	0.0529801	-11.42238841	5.41844E-27

Adjusted R Square	0.688683682

Here the above model explains 68% of variance and has intercept values of 29.428 which describes that if all independent values are zero average price will be 29.428.Ultimately this model is acceptable as it has good r^2 value. The residual plot depicts no traces of pattern.

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?

Adjusted R Square 1	0.688299
Adjusted R Square 2	0.688684

There is no major difference in Adjusted R^2 of both models. But the value of F in previous model is 124.90 and 140.64 in recent model. So, from this we can say that the last model is better.

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?

NOX	-10.27270508
PTRATIO	-1.071702473
LSTAT	-0.605159282
TAX	-0.014452345
AGE	0.03293496
INDUS	0.130710007
DISTANCE	0.261506423
AVG_ROOM	4.125468959

From the above table we can predict that if NOX value increases average price will decrease as it is negatively correlated. So, for every one unit increase in NOX will show significant decrease of 10.27 in average price.

d) Write the regression equation from this model.

Y =29.428+0.03*age+0.13*indus+(-10.27*NOX) +0.261*distance+(-0.01*tax) +(-1.071*ptratio)+4.12*avgroom+(-0.605*ltstat)