INTRODUCTION:

Our client is Terror's Real Estate Agency. One of the jobs is to map all relevant properties and figure out all geographical points related. We must find the relationship between data and provide the best possible outcome to our clients.

FROM THE GIVEN DATA:

Question 1:

a)

1.We first find the descriptive data from the analysis toolPak.From this data we achieve major values like mean, mode, standard error, maximum and minimum values. The crime rate from analysis is 5% means its quite low and standard deviation is also about 3 which is very near to our mean. Skewness is positive peak but doesn't vary much. Hence it is a flat peaked and moderately distributed. It gives us a conclusion that the crime rate is quite low.

2.The average age of the houses is 68%.The maximum age occurred is 100 according to data. Standard deviation is more spread out. Since Skew is a negative number, it suggest that a lot of houses made after 1940 is new/more. Some houses are as new as 3 years and some are 100 years old as well.

3.The mean achieved is 11.15%. The mean is achieved is as early as there is no variation in the given data. The skew is positive but not much varied hence flat peak is obtained. The area range is between the 2-standard deviation.

4.This data is a good one since the data again is not varying much. Median and mode is 0.538 which depicts that the distribution is symmetric. The standard deviation 20% of the mean hence its not much spread(kurtosis is negative too).

5.Mode represents that most of the clusters of houses are 24km from the highway. Mean and standard deviation is almost similar so not much difference. Skewness is greater that 1 hence the distribution is very good. Range, max, and min lies between 2-standard deviation.

6.Mean obtained is 408.273 which is the average amount the people pay as tax. But many houses have tax more $666. Standard deviation is higher 168.5371 hence data is more spread-out. Positively skewed 0.66 but less than 1 hence its moderately skewed.

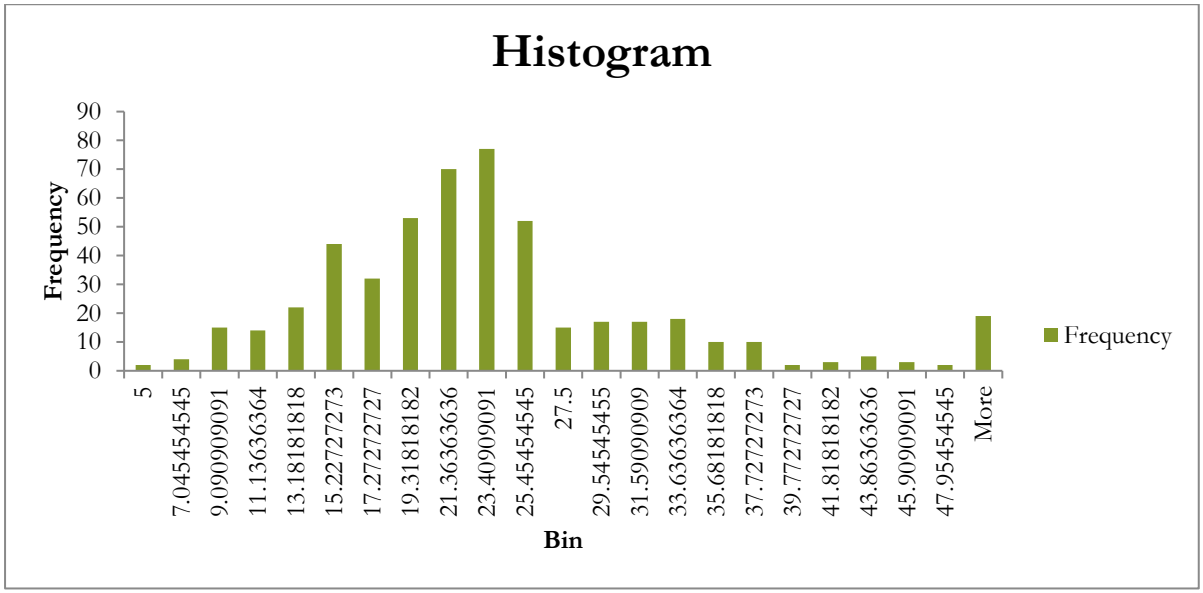
7.From the given data mean, median and mode is similar hence data is normally distributed.The data is negatively distributed which means that less number of students better quality of study. Standard deviation is less hence data is more clustered.

8.The average number of room per house is 6.28. Skewness is positive but not well distributed flat peak. Standard deviation is low hence most clustered. Skewness is less than 0.5 it is distributed somewhat symmetrical.

9.Lstat on an average is 12% hence much population does not lie in that category. Skewness is positive and well distributed and symmetrical.

10.Average prize is $22.53.10% of the house are $50,000(mode).Min to max is 5 to 50,000 with a range of 45,000.Skewness is really good since its above 1 hence its symmetrical. But this is 2-standard deviation is 95%.

b. Our Histogram is normally distributed and positively skewed.



c.)

Column1	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTR
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	4.677
AVG_ROOM	0.056117778	-4.74253803	-1.884225427	-0.024554826	-1.281277391	-34.51510104	-0.539
LSTAT	-0.882680362	120.8384405	29.52181125	0.487979871	30.32539213	653.4206174	5.771
AVG_PRICE	-1.16201224	-97.39615288	-30.46050499	-0.454512407	-30.50083035	-724.8204284	-10.09

This is the covariance matrix, we can see relationship between two variables. After conditional formatting the matrix which is red in color is positively correlated and green is negatively correlated.

d.)Correlation inferred from the data after conditional formatting.

Column1	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_RO
CRIME_RATE	1							
AGE	0.006859463	1						

INDUS	-0.005510651	0.644779	1					
NOX	0.001850982	0.73147	0.763651	1				
DISTANCE	-0.009055049	0.456022	0.595129	0.611441	1			
TAX	-0.016748522	0.506456	0.72076	0.668023	0.91022819	1		
PTRATIO	0.010800586	0.261515	0.383248	0.188933	0.46474118	0.460853	1	
AVG_ROOM	0.02739616	-0.24026	-0.39168	-0.30219	-0.20984667	-0.29205	-0.3555015	
LSTAT	-0.042398321	0.602339	0.6038	0.590879	0.48867633	0.543993	0.3740443	-0.61380
AVG_PRICE	0.043337871	-0.37695	-0.48373	-0.42732	-0.38162623	-0.46854	-0.5077867	0.69535

Top 3 correction = **0.910338 distance vs time**

0.763651 industry vs nox

0.73147 age vs nox

Bottom 3 correlation = **-0.73766 lstat vs avg prize**

-0.61381 avg room vs lstat

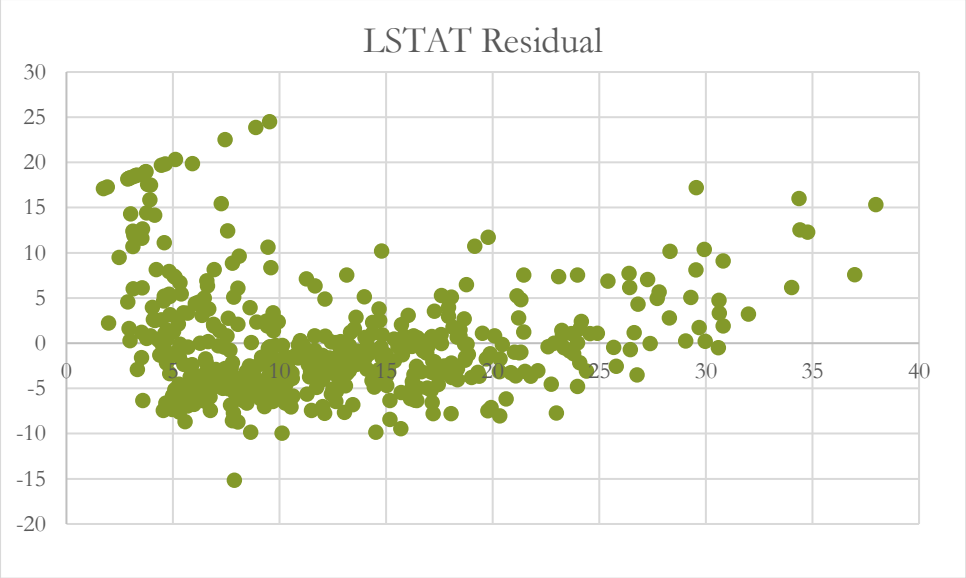
-0.50779 pt_ratio vs avg_price

Question2:

SIMPLE REGRESSION MODEL:

a) Y = -0.950049LSTAT + 34.55

b)



This is an interesting model, we achieved from regression function. The Lstat residual obtained from the residual data and predicted average price. As we can see the values are clustered around the zero showing that our predicted values are good.

The P value **5.08110339438785E-88** of LSTAT is less than 0.05 hence this model is significant and acceptable. Hence the values obtained is a null hypothesis which can be rejected and alternative hypothesis between LSTAT vs AVG_PRICE is acceptable. The average price of the property will decrease by 0.95 if the LSTAT value increases by 1%.

Question3:

MULTIVARIABLE REGRESSION:

a.) Y = -0.64AVG_ROOM + 5.09LSTAT – 1.35

b.) On interpreting all these values R- squared is 63% of the change in the average price and percentage of lstat is transferred to the average price of the house. From the p-value 3.47225760399802E-27, 6.66937E-41, is less than 0.05 hence the null hypothesis is seen to be false and alternative hypothesis is acceptable. Negative coefficient will decrease the value and positive coefficient will increase.

c.)

column	*lstat model 1*	*lstat,Avg_room Model2*
Multiple	0.737662726	0.799100498
R square	0.544146298	0.638561606
Adjusted R square	0.543241826	0.637124475

From the previous model we can compare R- squared, multiple R, Adjusted R square and we can check which is the best fit. The value of model 2 is better fit than model one as it approaches model 1.

R-squared is for model 2 is greater than model one

Adjusted R square model 2 is greater than model 1

4.MULTIVARIABLE REGRESSION 2:

a.) Model 2 regression model is created which gives us RMSE, R-squared, ADJUSTED R, intercept and coefficient, p-values.

b.)

Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372
Adjusted R Square	0.688298647

Interpreting the above values we get to know 69% of average price affects the any other change in values due to independent variables. If the coefficient value is positive like avg_room(>0) then the intercept value increases and if less than 0 the value tends to decrease as inferred from the given data. From the

data given we see that p-value is less than 0.05 the null hypothesis is rejected and hence relation is obtained(alternative hypothesis).

c.)

Column 1	slr_model_1	mlr_model_2	mlr_model_3
multiple R	0.737662726	0.799100498	0.832978824
R squared	0.544146298	0.638561606	0.69385372
adjusted R squared	0.543241826	0.637124475	0.688298647

Model three is a better fit as we compare multiple R value of 0.832978824 and that of model 1 and 2 does not have much of difference and also is greater value near 1.Since we take many independent variables we have a better chance that the model will perform well.

d.) list required

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

Question 5:

BEST FIT MODEL

a.) Table obtained from the given data and the independent variables.

Regression Statistics	
Multiple R	0.832835773
R Square	0.693615426
Adjusted R Square	0.688683682

b) Interpreting the above values, we get to know 69% of average price affects the any other change in values due to independent variables. If the coefficient value is positive like avg_room(>0) then the intercept value increases and if less than 0 the value tends to decrease as inferred from the given data. From the data given we see that p-value is less than 0.05 the null hypothesis is rejected and hence relation is obtained(alternative hypothesis).

c.)

column 1	slr_model_1	mlr_model_2	mlr_model_3	fit_model
multiple R	0.737662726	0.799100498	0.832978824	0.832835773
R squared	0.544146298	0.638561606	0.69385372	0.693615426
adjusted R squared	0.543241826	0.637124475	0.688298647	0.688683682

From the previous data we see that the CRIME_RATE has a p-value greater than 0.05 and hence we must omit this value so as to obtain the best fit model. Even though from the data table we see that model_3 has a better model we will still consider best fit model due to adjusted R value. Model_3 failed p-test hence not good fit and also it has many data points hence not very accurate as results. Hence the best model gives us better performance.

Question:6

a) Values of coefficient in ascending order:

	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

b) If the value of NOX is more in a locality, the property value should go down because nobody wants to live in a polluted area. This is also verified with the coefficient of NOX in the best fit model. If all values are constant only then the given value of the property will decrease if the NOX value increases.

Question 7:

The best fit model has the following equation:

AVG_PRICE = -10.27*NOX – 1.07*PTRATIO – 0.61*LSTAT 0.01*TAX +0.03*AGE +0.03*AGE + 0.13*INDUS + 0.26*DISTANCE +4.13*AVG_ROOM +29.43

CONCLUSION ON THE BEST FIT MODEL:

1. One big conclusion is that the crime rate in that area is it does not affect the prices of the area/houses/property
2. NOX, PTRATIO, LSTAT, TAX should never be more for a better property value.
3. AGE, INDUS, DISTANCE, AVG_ROOM should be higher for higher value of property.
4. Older the property value greater is the valuation.