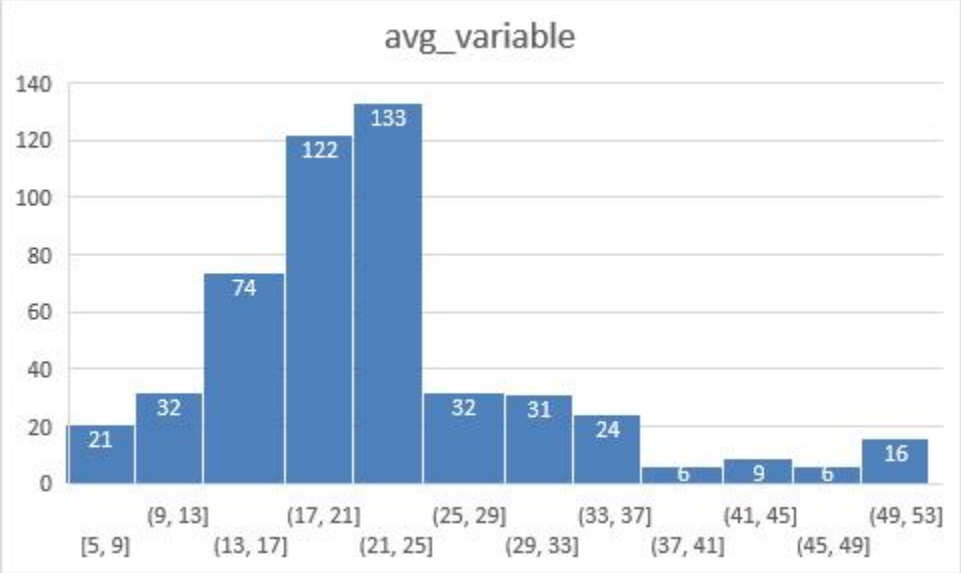
Assignment – Terro's real estate agency

1) Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation
   - CRIME RATE its positively skewed with big tail to the right , indicating the presence of extreme values in the data. kurtosis is negative so flat peak in this plot.
   - The AVG,INDUS,NOX, DISTANCE ,TAX,PTRATIO have kurtosis also negative so flat peak in these variable plots.
   - The AVG_ROOM,AVG PRICE,LSTAT have positive kurtosis, so sharp peak in these variable plots
   - The PTRATIO and AGE have negative skewness, so these are reprecents big tail to the left ,remaining variables have positive skewness with big tail to the right.
   - The standard deviation is more than half of mean is represents positive skewness othervise negative skewness

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



AVG_PRICE	
Mean	22.532806
Standard Error	0.4088611
Median	21.2
Mode	50
Standard Deviation	9.1971041
Sample Variance	84.586724
Kurtosis	1.4951969
Skewness	1.1080984
Range	45
Minimum	5
Maximum	50

Sum	11401.6
Count	506

- In the first glance avg price value look like a normal distribution. The median is 21 and mean of avg price is 50. The sample consists of ranges from 5 to 50,mean and median is approx.
- price 21 with a standard deviation of 9.1971.

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DIS*
CRIME_RATE	8.516148				
AGE	0.562915	790.7925			
INDUS	-0.11022	124.2678	46.97143		
NOX	0.000625	2.381212	0.605874	0.013401	
DISTANCE	-0.22986	111.55	35.47971	0.61571	
TAX	-8.22932	2397.942	831.7133	13.0205	
PTRATIO	0.068169	15.90543	5.680855	0.047304	
AVG_ROOM	0.056118	-4.74254	-1.88423	-0.02455	
LSTAT	-0.88268	120.8384	29.52181	0.48798	
AVG_PRICE	1.162012	-97.3962	-30.4605	-0.45451	

3) Compute the covariance matrix. Share your observations.

- Gendraly It is positive if X and Y value are mostly both above or both below their averages. It is negative if X and Y values are mostly on opposite sides of their averages.
- The most maximum of negative covariance relation value is -724.82 and positive value is 283448.62

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	-0.2098467	-0.29205	-0.355501	1	
LSTAT	-0.042398321	0.602339	0.6038	0.590879	0.48867633	0.543993	0.3740443	0.613808272	-1
AVG_PRICE	0.043337871	-0.37695	-0.48373	-0.42732	-0.3816262	-0.46854	-0.507787	0.695359947	-0.73766

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

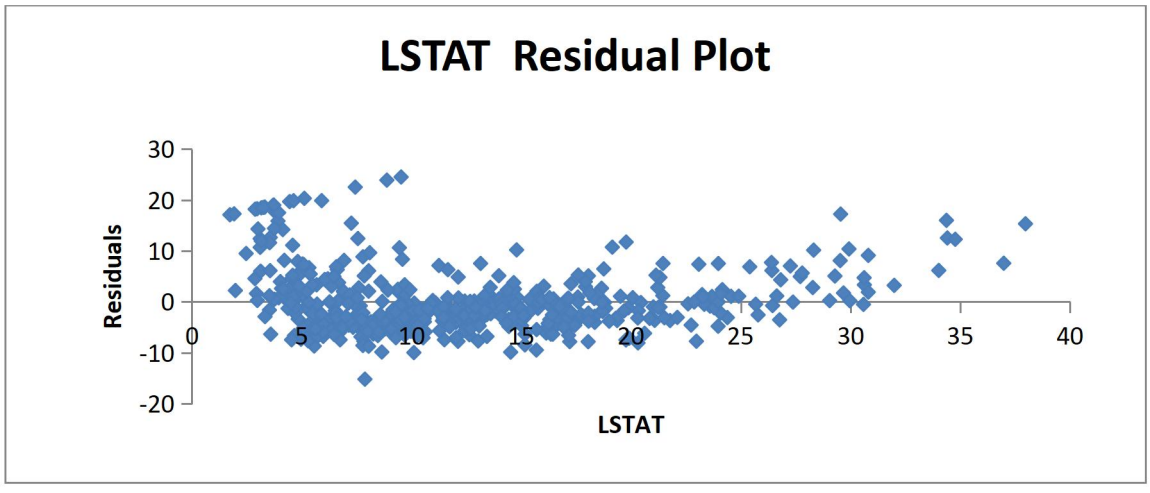
- 0.910228
- 0.763651
- 0.73147

b) Which are the top 3 negatively correlated pairs.

- -0.50778
- -0.48373
- -0.42732

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

6)



	Coefficients	*P-value*
Intercept	34.55384	3.7E-236
LSTAT	-0.95005	5.08E-88

Regression Statistics	
Multiple R	0.737663

R Square	0.544146
Adjusted R Square	0.543242
Standard Error	6.21576
Observations	506

a) What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?
   - 0.544146 of variation of LSTAT can be explained by average price
   - Remaining 46 % of variation of sales cannot be explained by average price

b) Is LSTAT variable significant for the analysis based on your model?
   - If p value is less than 0.05,our regression value is correct
   - The LSAT value is -0.95005, so its true

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

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	*P-value*
Intercept	-1.35827	0.668765
AVG_ROOM	5.094788	3.47E-27
LSTAT	-0.64236	6.67E-41

=-1.35827+(5.0948)*avg room+(-0.6424)*LSTAT

5th ques Regression Statistics	
Multiple R	0.737663
R Square	0.544146
Adjusted R Square	0.543242
Standard Error	6.21576

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

Observations	506

6th ques Regression Statistics	
Multiple R	0.7991
R Square	0.638562
Adjusted R Square	0.637124
Standard Error	5.540257
Observations	506

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	*P-value*
Intercept	29.24132	2.54E-09
CRIME_RATE	0.048725	0.534657
AGE	0.032771	0.01267
INDUS	0.130551	0.039121
NOX	-10.3212	0.008294
DISTANCE	0.261094	0.000138
TAX	-0.0144	0.000251
PTRATIO	-1.07431	6.59E-15
AVG_ROOM	4.125409	3.89E-19
LSTAT	-0.60349	8.91E-27

- If p value is less than .05,our regression value is correct
- But the crime rate p value is greater than 0.05
- The crime rate p value is 0.5347

Regression Statistics	
Multiple R	0.832979
R Square	0.693854
Adjusted R Square	0.688299
Standard Error	5.134764
Observations	506

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)

Regression Statistics	
Multiple R	0.832836
R Square	0.693615
Adjusted R Square	0.688684
Standard Error	5.131591
Observations	506

	Coefficients	*P-value*
Intercept	29.42847	1.85E-09
AGE	0.032935	0.012163
INDUS	0.13071	0.038762
NOX	-10.2727	0.008546
DISTANCE	0.261506	0.000133
TAX	-0.01445	0.000236
PTRATIO	-1.0717	7.08E-15
AVG_ROOM	4.125469	3.69E-19
LSTAT	-0.60516	5.42E-27

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.