TERRO'S REAL ESTATE AGENCY

Data Analysis Project

Project Done by

Ram K

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

AGE		*INDUS*		*NOX*		*DISTANCE*	
Mean	68.5749	Mean	11.13678	Mean	0.554695	Mean	9.549407
Standard Error	1.25137	Standard Error	0.30498	Standard Error	0.005151	Standard Error	0.387085
Median	77.5	Median	9.69	Median	0.538	Median	5
Mode	100	Mode	18.1	Mode	0.538	Mode	24
Standard Deviation	28.14886	Standard Deviation	6.860353	Standard Deviation	0.115878	Standard Deviation	8.707259
Sample Variance	792.3584	Sample Variance	47.06444	Sample Variance	0.013428	Sample Variance	75.81637
Kurtosis	-0.96772	Kurtosis	-1.23354	Kurtosis	-0.06467	Kurtosis	-0.86723
Skewness	-0.59896	Skewness	0.295022	Skewness	0.729308	Skewness	1.004815
Range	97.1	Range	27.28	Range	0.486	Range	23
Minimum	2.9	Minimum	0.46	Minimum	0.385	Minimum	1
Maximum	100	Maximum	27.74	Maximum	0.871	Maximum	24
Sum	34698.9	Sum	5635.21	Sum	280.6757	Sum	4832
Count	506	Count	506	Count	506	Count	506

TAX		*PTRATIO*		*AVG_ROOM*		*LSTAT*		*AVG_PRICE*	
Mean	408.2372	Mean	18.45553	Mean	6.284634	Mean	12.65306	Mean	22.53281
Standard Error	7.492389	Standard Error	0.096244	Standard Error	0.031235	Standard Error	0.317459	Standard Error	0.408861
Median	330	Median	19.05	Median	6.2085	Median	11.36	Median	21.2
Mode	666	Mode	20.2	Mode	5.713	Mode	8.05	Mode	50
Standard Deviation	168.5371	Standard Deviation	2.164946	Standard Deviation	0.702617	Standard Deviation	7.141062	Standard Deviation	9.197104
Sample Variance	28404.76	Sample Variance	4.686989	Sample Variance	0.493671	Sample Variance	50.99476	Sample Variance	84.58672
Kurtosis	-1.14241	Kurtosis	-0.28509	Kurtosis	1.8915	Kurtosis	0.49324	Kurtosis	1.495197
Skewness	0.669956	Skewness	-0.80232	Skewness	0.403612	Skewness	0.90646	Skewness	1.108098
Range	524	Range	9.4	Range	5.219	Range	36.24	Range	45
Minimum	187	Minimum	12.6	Minimum	3.561	Minimum	1.73	Minimum	5
Maximum	711	Maximum	22	Maximum	8.78	Maximum	37.97	Maximum	50
Sum	206568	Sum	9338.5	Sum	3180.025	Sum	6402.45	Sum	11401.6
Count	506	Count	506	Count	506	Count	506	Count	506

- According to my perception from the dataset **TAX** has the highest average around (**408.2372**) comparing others and comparing every column data the median of **TAX** is higher than other data and value of mode is very low in **NOX**

- From the data,

AGE		INDUS		NOX		DISTANCE	
Skewness	-0.59896	Skewness	0.295022	Skewness	0.729308	Skewness	1.004815

TAX		PTRATIO		AVG_ROOM		LSTAT		AVG_PRICE	
Skewness	0.669956	Skewness	-0.80232	Skewness	0.403612	Skewness	0.90646	Skewness	1.108098

Highly Skewed

+ve

AVG_PRICE - **1.108098**

DISTANCE - **1.004815**

LSTAT - **0.90646**

Here there are top 3 positively skewed values which lies towards left

-ve

PTRATIO - **-0.80232**

AGE - **-0.59896**

Here there are negatively skewed values which lies towards Right

- **Measure of Dispersion**

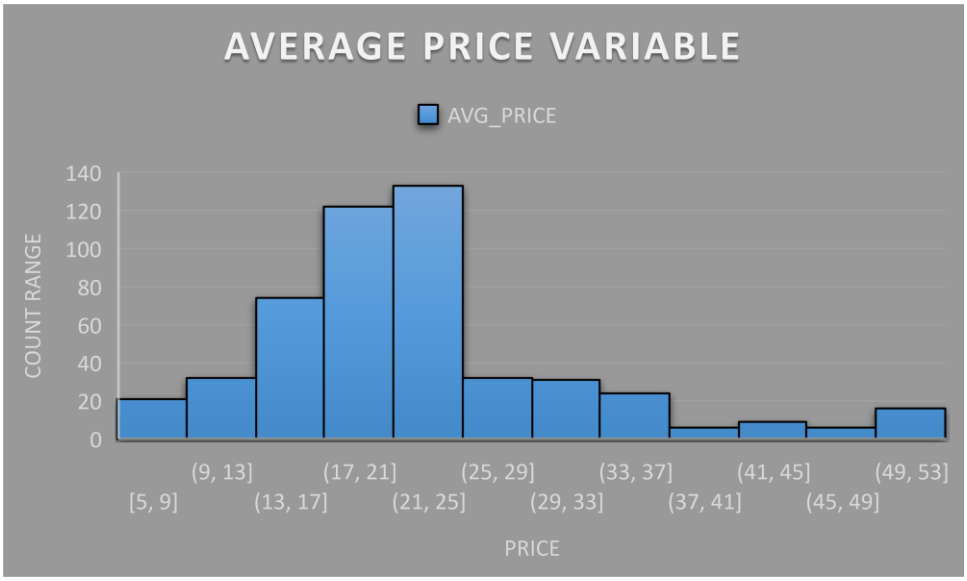
AGE		*INDUS*		*NOX*		*DISTANCE*	
Sample Variance	792.3584	Sample Variance	47.06444	Sample Variance	0.013428	Sample Variance	75.81637
Standard Deviation	28.14886	Standard Deviation	6.860353	Standard Deviation	0.115878	Standard Deviation	8.707259

TAX		*PTRATIO*		*AVG_ROOM*		*LSTAT*		*AVG_PRICE*	
Sample Variance	28404.76	Sample Variance	4.686989	Sample Variance	0.493671	Sample Variance	50.99476	Sample Variance	84.58672
Standard Deviation	168.5371	Standard Deviation	2.164946	Standard Deviation	0.702617	Standard Deviation	7.141062	Standard Deviation	9.197104

TAX FACTOR HAS THE HIGHEST VARIANCE (28404.76) SINCE MONEY INVOLVED IN THIS DATA SO WE CAN SEE HIGHER VARIENCE HERE

AGAIN THE TAX HAS THE STANDARD DEVIATION (168.5371) .

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



- From above Histogram around (21-25)price has highest value
- From above Histogram around (37-41)price has very low value
- Here in average price histogram representation the avg_price value is arranged in ascending order range and then plotted it initially starts in low range reaches a peak point and again goes back to low range
- Since it has peak point the kurtosis is positive and which can be called as leptokurtic

3) Compute the covariance matrix. Share your observations.

	CRIME_RAT	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RA	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	75.66653					
TAX	-8.22932	2397.942	831.7133	13.0205	1333.117	28348.62				
PTRATIO	0.068169	15.90543	5.680855	0.047304	8.743402	167.8208	4.677726			
AVG_ROOI	0.056118	-4.74254	-1.88423	-0.02455	-1.28128	-34.5151	-0.53969	0.492695		
LSTAT	-0.88268	120.8384	29.52181	0.48798	30.32539	653.4206	5.7713	-3.07365	50.89398	
AVG_PRICI	1.162012	-97.3962	-30.4605	-0.45451	-30.5008	-724.82	-10.0907	4.484566	-48.3518	84.41956

Here

Positive Covariance means both X and Y increases or decreases at same time.

Negative covariance means when X/Y increases the other variable decreases, X/Y is decreased then other variable will be increased

These are few Positive covariance in the above covariance table

(TAX, TAX)

(DISTANCE, TAX)

(AGE, TAX)

These are few Negative covariance in the above covariance table

(CRIME RATE, INDUS)

(CRIME RATE, DISTANCE)

(DISTANCE, AVG PRICE)

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

	CRIME_RAT	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RA	1									
AGE	0.006859	1								
INDUS	-0.00551	0.644779	1							
NOX	0.001851	0.73147	0.763651	1						
DISTANCE	-0.00906	0.456022	0.595129	0.611441	1					
TAX	-0.01675	0.506456	0.72076	0.668023	0.910228	1				
PTRATIO	0.010801	0.261515	0.383248	0.188933	0.464741	0.460853	1			
AVG_ROO	0.027396	-0.24026	-0.39168	-0.30219	-0.20985	-0.29205	-0.3555	1		
LSTAT	-0.0424	0.602339	0.6038	0.590879	0.488676	0.543993	0.374044	-0.61381	1	
AVG_PRICI	0.043338	-0.37695	-0.48373	-0.42732	-0.38163	-0.46854	-0.50779	0.69536	-0.73766	1

a) Which are the top 3 positively correlated pairs

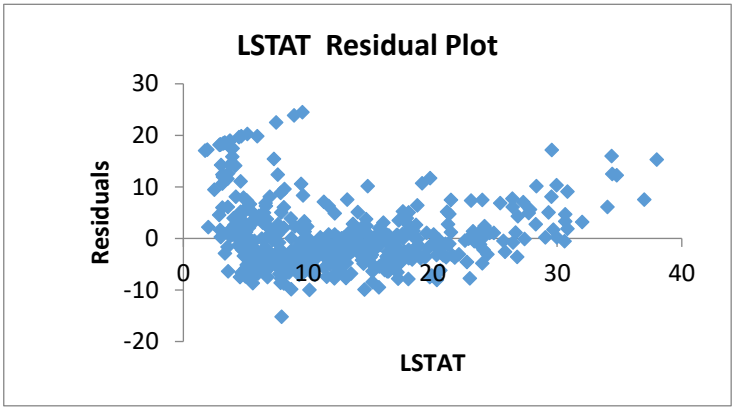
1ST	DISTANCE AND TAX	0.910228
2ND	INDUS AND NOX	0.763651
3RD	AGE AND NOX	0.73147

b) Which are the top 3 negatively correlated pairs

1ST	LSTAT AND AVG PRICE	-0.73766
2ND	AVG ROOM AND LSTAT	-0.61381
3RD	PTRATIO AND AVG PRICE	-0.50779

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

Regression Statistics	
Multiple R	0.737663
R Square	0.544146
Adjusted R Square	0.543242
Standard Error	6.21576
Observations	506



	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	34.55384	0.562627	61.41515	3.7E-236	33.44846	35.65922	33.44846	35.65922
LSTAT	-0.95005	0.038733	-24.5279	5.08E-88	-1.02615	-0.87395	-1.02615	-0.87395

a) What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?
In Residual Plot
According to my view the graph is not in pattern
The values are scattered
The Coefficient of LSTAT is in negative but the value is near to -1

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

For preparing a **good model**
The adjusted r value should be **0.8**

If adjusted r value is **0.9** it is **better**

If the adjusted r value is **1** it is **accurate**

Since the Adjusted R is Very Low (0.543242)
So LSTAT variable is not significant

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

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

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.35827	3.172828	-0.4281	0.668765	-7.5919	4.875355	-7.5919	4.875355
AVG_ROOM	5.094788	0.444466	11.46273	3.47E-27	4.22155	5.968026	4.22155	5.968026
LSTAT	-0.64236	0.043731	-14.6887	6.67E-41	-0.72828	-0.55644	-0.72828	-0.55644

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?

Regression Equation

$Y=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}$

Y=Intercept+(AVG_ROOM*7)+(LSTAT*20)

Y=-1.35827+(5.094588*7)+(-0.64236*20)

Y=21.45808

The company has coted value of 30000USD

But according to this question the value is around 21458.08

Hence the company is **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.

Previous Question Adjusted R square Value

Adjusted R Square	0.543242

Adjusted R Square value of this Question

Adjusted R Square	0.637124

Comparing both the Adjusted R square value

The performance of this model (0.637124) is better than previous question (0.543242)

Since the Adjusted value of this part increases comparing previous one so this give better performance

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.

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

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.24132	4.817126	6.070283	2.54E-09	19.77683	38.7058	19.77683	38.7058
CRIME_RATE	0.048725	0.078419	0.621346	0.534657	-0.10535	0.202799	-0.10535	0.202799
AGE	0.032771	0.013098	2.501997	0.01267	0.007037	0.058505	0.007037	0.058505
INDUS	0.130551	0.063117	2.068392	0.039121	0.006541	0.254562	0.006541	0.254562
NOX	-10.3212	3.894036	-2.65051	0.008294	-17.972	-2.67034	-17.972	-2.67034
DISTANCE	0.261094	0.067947	3.842603	0.000138	0.127594	0.394593	0.127594	0.394593
TAX	-0.0144	0.003905	-3.68774	0.000251	-0.02207	-0.00673	-0.02207	-0.00673
PTRATIO	-1.07431	0.133602	-8.0411	6.59E-15	-1.3368	-0.81181	-1.3368	-0.81181
AVG_ROOM	4.125409	0.442759	9.317505	3.89E-19	3.255495	4.995324	3.255495	4.995324
LSTAT	-0.60349	0.053081	-11.3691	8.91E-27	-0.70778	-0.49919	-0.70778	-0.49919

- The adjusted R square Value is 0.693854
- This has good adjusted R square value so this model can be used for prediction
- The coefficient of this avg_room is higher comparing others then the distance, indus, crime_rate, age ,tax , LSTAT, PTRATIO, nox respectively
- The Coefficient Value of intercept is 29.24132
- Significant Variable

  Distance

  Indus

  Age

  Tax

  LSTAT

  PTRATIO

  Nox

  Avg_room

- Insignificant Variable

  Crime_Rate (since it has higher p value 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

- Significant Variable

  Distance

  Indus

  Age

  Tax

  LSTAT

  PTRATIO

  Nox

  Avg_room

a) Interpret the output of this model.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.42847	4.804729	6.124898	1.85E-09	19.98839	38.86856	19.98839	38.86856
AVG_ROOM	4.125469	0.442485	9.3234	3.69E-19	3.256096	4.994842	3.256096	4.994842
DISTANCE	0.261506	0.067902	3.851242	0.000133	0.128096	0.394916	0.128096	0.394916
INDUS	0.13071	0.063078	2.072202	0.038762	0.006778	0.254642	0.006778	0.254642
AGE	0.032935	0.013087	2.516606	0.012163	0.007222	0.058648	0.007222	0.058648
TAX	-0.01445	0.003902	-3.70395	0.000236	-0.02212	-0.00679	-0.02212	-0.00679
LSTAT	-0.60516	0.05298	-11.4224	5.42E-27	-0.70925	-0.50107	-0.70925	-0.50107
PTRATIO	-1.0717	0.133454	-8.03053	7.08E-15	-1.33391	-0.8095	-1.33391	-0.8095
NOX	-10.2727	3.890849	-2.64022	0.008546	-17.9172	-2.62816	-17.9172	-2.62816

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

Adjusted R Square Value is 0.688684

This model can be used to get good results

Every P value of this model are significant

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?

The Adjusted R Square of this model is 0.688684

The Adjusted R Square of Previous Model is 0.688299

Since

0.688684>0.688299

This model performs better than the previous one

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?

Values in Accending order

	Coefficients
NOX	-10.2727
PTRATIO	-1.0717
LSTAT	-0.60516
TAX	-0.01445
AGE	0.032935
INDUS	0.13071
DISTANCE	0.261506
AVG_ROOM	4.125469
Intercept	29.42847

Since Nox is negative If Nox increases the value of average price will get decreased

d) Write the regression equation from this model.

The Regression Equation is

$Y=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}\ldots..\ \beta_nX_{ni}$

Y=Intercept + co.eff of age*65.2+co.eff of INDUS*2.31+co.eff of nox*0.538+co.eff of Distance*1+co.eff of tax*296+co.eff of PTRATIO*15.3+co.eff of avg_room*6.575+co.eff of LSTAT*4.98

Y=29.42847+0.032935*65.2+0.13071*2.31-10.2727*0.538+0.261506*1-0.01445*296-1.0717*15.3+4.125469*6.575-0.60516*4.98

Y=30.04961738