Assignment – Terro's real estate agency

Real estate data analysis – Exploratory data analysis, Linear Regression

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

Observation:

1) The number of records given in the dataset is 506.
2) Firstly if we consider the Distance variable we can analyze that the maximum distance is 24 and has a mode as 24.
3) The average tax paid is 408.2 and the tax range 524.
4) From the skewness of variables we can say that the dataset is highly skewed.
5) And if we consider the age variable the maximum age is 100 and the mode is also 100 which says that most of the houses has an age of 100.

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

observation:

1) We have least count of houses from range $37000 to $41000 and $45000 to $49000.

2) We can summarise that most of the houses are from range $21000 to $25000.We have least count of houses from range $37000 to $41000 and $45000 to$49000.

3) **Compute the covariance matrix. Share your observations**

observation:

1) We can see a highly positive covariance between Tax and Age in the table. So very Strong Relation b/w Them.
2) We have the lowest count of houses from range $37000 to $41000 and $45000 to $49000.

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

a) Which are the top 3 positively correlated pairs and

b) Which are the top 3 negatively correlated pairs.

a)

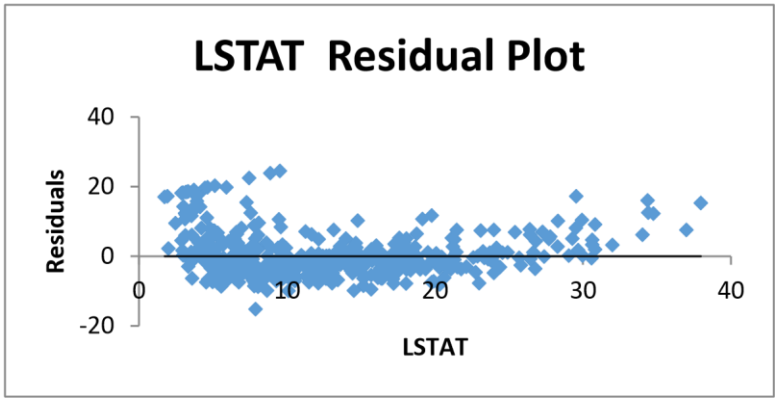
Top 3 Positively Correlated Pair	Column1
TAX VS DISTANCE	0.910228189
NOX VS INDUS	0.763651447
NOX VS AGE	0.731470104

b)

Column1	Column2	Column3
Top 3 Negatively		Correlated
Lstat vs Avg price		-0.7376627
Avg Room Vs Lstat		-0.6138083
PT-Ratio vs Avg Price		-0.5077867

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

a) **What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?**
b) **Is LSTAT variable significant for the analysis based on your model?**



a) observation:

1) From this model 54% of the variation in the average price is explained by the LSTAT.
2) The coefficient of LSTAT for the model is -0.950049354. This says that if LSTAT increases by 0.9 times then the average price of a house decreases 0.9 times. Intercept of LSTAT for the model is 34.55384088

b) observation:

1) Yes, LSTAT is a significant variable for the avg_price from this model
2) As the p-value(5.08E-88) we obtained from this model is less than 0.05.
3) By this we can say that LSTAT is a significant variable according to this model.

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

a) Write the Regression equation. If a new house in this locality has 7 rooms (on average) and has a value of 20 for L-STAT, then what will be the value of AVG_PRICE? How does it compare to a company quoting a value of 30000 USD for this locality? Is the company Overcharging/ Undercharging?

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.

a)

AVG_ROOM	5.094788	7
LSTAT	-0.64236	20
AVG_PRICE	21.45808	

b)

1) Yes, the performance of this model performs well compared to the previous model.
2) From this model the linear equation we obtained is y= -1.35 +5.09a -0.64b (Where a=Avg_room b=LSTAT) And the Value of R square = 0.638561606.

With this we can say that 63% of variability for average price is explained by Avg_room and LSTAT combined and we obtained multiple R values as 0.79 which says it is highly correlated. However in the previous model, LSTAT alone describes 54% of variability for average price.

7) Build another Regression model with all variables where AVG_PRICE alone is 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	29.24131526	2.53978E-09
CRIME_RATE	0.048725141	0.534657201
AGE	0.032770689	0.012670437
INDUS	0.130551399	0.03912086
NOX	-10.3211828	0.008293859
DISTANCE	0.261093575	0.000137546
TAX	-0.01440119	0.000251247
PTRATIO	-1.074305348	6.58642E-15
AVG_ROOM	4.125409152	3.89287E-19
LSTAT	-0.603486589	8.91071E-27

Observation:

1)From this we can say that the crime rate is not a significant variable for an average price of an house as p-value is greater than 0.5
2) All the features combined explain 69% of the variability for the average price of a house
3)NOX, TAX, PTRATIO, and LSTAT have negative coefficients which says that an increase in these features result decrease in the price of the house and vice versa

8) Pick out only the significant variables from the previous question. Make another an 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

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 NOX's value is higher in a locality in this town?

d) Write the regression equation from this model.

a)

	Coefficients	*P-value*
Intercept	29.42847349	1.84597E-09
AGE	0.03293496	0.012162875
INDUS	0.130710007	0.038761669
NOX	-10.27270508	0.008545718
DISTANCE	0.261506423	0.000132887
TAX	-0.014452345	0.000236072
PTRATIO	-1.071702473	7.08251E-15
AVG_ROOM	4.125468959	3.68969E-19
LSTAT	-0.605159282	5.41844E-27

b)

Regression Statistics	
Multiple R	0.832835773
R Square	0.693615426

R Square in value

Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372

c)

Column1	COEFFICIENT
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
INTERCEPT	29.42847349

d) Regression equation:

Y=0.03293496 X0 + 0.130710007 X1 -10.27270508 X3 +0.261506423 X4 -0.014452345 X5 -1.071702473 X6 + 4.125468959 X7 -0.605159282 X8+29.42847349