Assignment – Terro's real estate agency

Real estate data analysis – Exploratory data analysis, Linear Regression

Problem Statement (Situation):

"Finding out the most relevant features for pricing of a house" Terro's real-estate is an agency that estimates the pricing of houses in a certain locality. The pricing is concluded based on different features / factors of a property. This also helps them in identifying the business value of a property. To do this activity the company employs an "Auditor", who studies various geographic features of a property like pollution level (NOX), crime rate, education facilities (pupil to teacher ratio), connectivity (distance from highway), etc. This helps in determining the price of a property.

Objective (Task):

Your job, as an auditor, is to analyse the magnitude of each variable to which it can affect the price of a house in a particular locality.

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

OBSERVATION:

- The difference between the mean and the median for TAX, DIATANCE and INDHUS is huge which means there are many outliers.
- AGE and PT_RATIO are negatively skewed. (MEAN<MEDIAN)
- The mode and maximum value is equal for AGE and DISTANCE.
- The variable with highest range and standard error is TAX with the count 554 and 7.49 respectively.
- AVG_ROOM, LSTAT ,AVG_PRICE are PLATYKURTIC. While the remaining are leptokurtic.

Q-2 Plot a histogram of the Average price variable. What do you infer?

OBSERVATION:

- From the above histogram we can clearly predict that majority of the houses price range from $21000 to $25000.
- There are 12 houses with the minimal average price ($37000 to $41000 and $45000 to $4900).

Q-3 Compute the covariance matrix. Share your observations.

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.516147873									
AGE	0.562915215	790.7925								
INDUS	-0.110215175	124.2678	46.97143							
NOX	0.000625308	2.381212	0.605874	0.013401						
DISTANCE	-0.229860488	111.55	35.47971	0.61571	75.66653					
TAX	-8.229322439	2397.942	831.7133	13.0205	1333.117	28348.62				
PTRATIO	0.068168906	15.90543	5.680855	0.047304	8.743402	167.8208	4.677726296			
AVG_ROOM	0.056117778	-4.74254	-1.88423	-0.02455	-1.28128	-34.5151	-0.539694518	0.492695216		
LSTAT	-0.882680362	120.8384	29.52181	0.48798	30.32539	653.4206	5.771300243	-3.073654967	50.89398	
AVG_PRICE	1.16201224	-97.3962	-30.4605	-0.45451	-30.5008	-724.82	-10.09067561	4.484565552	-48.3518	84.41955616

OBSERVATION:

- The data which is highlighted in red indicates the variance while the remaining measures the covariance.
- The average price of the house has a negative covariance with maximum number of variables (age, industries, nitric acid concentration, distance from highway, tax, pupil-teacher ratio and % lower status of the population).
- Whereas, crime rate and average rooms are heading in the same direction with average price.

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

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	1									
AGE	0.006859463	1								
INDUS	-0.00551065	0.64478	1							
NOX	0.001850982	0.73147	0.76365	1						
DISTANCE	-0.00905505	0.45602	0.59513	0.61144	1					
TAX	-0.01674852	0.50646	0.72076	0.66802	0.91023	1				
PTRATIO	0.010800586	0.26152	0.38325	0.18893	0.46474	0.46085	1			
AVG_ROOM	0.02739616	-0.24026	-0.39168	-0.30219	-0.20985	-0.29205	-0.35550149	1		
LSTAT	-0.04239832	0.60234	0.6038	0.59088	0.48868	0.54399	0.374044317	-0.61380827	1	
AVG_PRICE	0.043337871	-0.37695	-0.48373	-0.42732	-0.38163	-0.46854	-0.50778669	0.695359947	-0.73766273	1

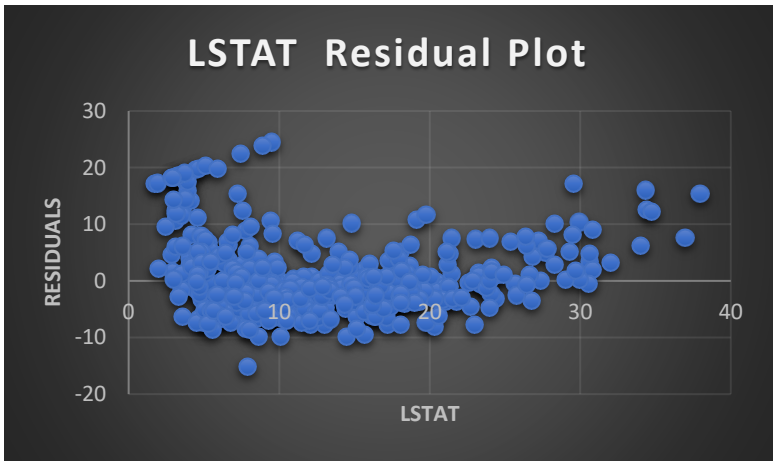
a)Which are the top 3 positively correlated pairs.

1. DISTANCE – TAX (0.91023)
2. NOX – INDUS (0.76365)
3. NOX – AGE (0.73147)

b)Which are the top 3 negatively correlated pairs.

1. AVG – PRICE-LSTAT (-0.73766273)
2. LSTAT –AVG_ROOM (-0.61380827)
3. AVG_PRICE –PTRATIO (-0.50778669)

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



1)What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?

- Significant F (5.0811E-88) is less than 0.05 (It denotes strong evidence against the null hypothesis, since there is below 5% probability of the null being correct).
- **y=a+bx**
  a is the intercept
  b is the coefficient

  Avg_price=34.55384088+ (-0.950049354* LSTAT)
- The coefficient of LSTAT is -0.950049354. (The average price of house decreases 0.9 times if the LSTAT increases).
- Residual Plot:
  The data are randomly distributed on both side of the line(there is no trend identified).Hence, the regression line is a good model for the data.

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

LSTAT variable is significant for the analysis:

- Coefficient of determination is 0.544146298. So,54% of the variation in the average price is explained by the LSTAT.
- Significant F (5.0811E-88) is less than 0.05.

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

1)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?

Given,

AVG_ROOMS (x1) =7

L-STAT(x2) =20

y= dependent variable (AVG_PRICE)

x1= Independent variable (AVG_ROOM)

x2= Independent variables (LSTAT)

Regression Equation:

$$y = a+bx1 + bx2$$

y= -1.358 +5.09 (x1) -0.642 (x2)

y= -1.358 + 5.09(7) -0.642(20) = 21.44

The price for the new house is $ 21440.

Comparing to the company quoting a value of 3000 USD is clearly overcharging.

2) *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.*

- Coefficient of determination of this model is 0.638562.Which means 63% of variability for average price is explained by the variables Avg_room and LSTAT combined.

- On the other hand, only 54% of the variation in the average price is explained by the LSTAT.
- Hence, the performance of this model is better than the previous model.

Q-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*
CRIME_RATE	**0.048725141**	**0.534657**
AGE	**0.032770689**	**0.01267**
INDUS	**0.130551399**	**0.039121**
NOX	**-10.3211828**	**0.008294**
DISTANCE	**0.261093575**	**0.000138**
TAX	**-0.01440119**	**0.000251**
PTRATIO	**-1.07430534**	**6.59E-15**
AVG_ROOM	**4.125409152**	**3.89E-19**
LSTAT	**-0.60348658**	**8.91E-27**

- Adjusted R Square is 69%. Which means all the variable present in the above mentioned table are 69% responsible for the variation in average house price.
- When the crime rate, age, indus, distance from highway ,number of room increases the average price also increases.
- All the other variable increases with the decrease in the average price.
- Except Crime rate all the other variables are significant as the p-values are less than 0.05.

Q-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:

1)Interpret the output of this model.

- The model is good as the Significance F value is below 0.05.
- Apart from this the individualS p-value of each independent variable is also less than 0.05.

2)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 value of this model is greater than the previous one. Therefore, the latest model which does not include the crime_rate performs better.

3)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?

Coefficients in ascending order:

If NOX increases the average price of the house will decrease by 10 times.

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

4) Write the regression equation from this model.

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

	Coefficients
Intercept	29.42847349
AGE	0.03293496
INDUS	0.130710007
NOX	-10.2727050
DISTANCE	0.261506423
TAX	-0.01445234
PTRATIO	-1.07170247
AVG_ROOM	4.125468959
LSTAT	-0.60515928

Conclusion:

From the above analysis we conclude that all the variables except crime rate plays a vital role in pricing the house**.**