# TERRO'S REAL ESTATE AGENCY

## Problem Statement (Situation):

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 analyze the magnitude of each variable to which it can affect the price of a house in a particular locality.

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

1)Column Age:

- The mode of age is 100, as mode is 100 we suggest you to go for the model which is around that Age.
- And I see that data set has high standard deviation with high variance also mean the spread of the data is too high
  i) it mean the data is widely spread there wide spread of age also.
- The range is 97.1, from a minimum age of 2.9 and maximum age of 100.

- Conclusion:

we understand this from the data set is widely distributed from the mean and the mode is 100 we should go for a model which near around that age.

2) **Indus**:

- The mode of the Indus is = 18.1
- Std deviation is = 6.86
- Variance is = 47.3

3) NOX (Nitrogen Oxide Concentration):

- The mean NOX concentration is = 0.55

## 4) Distance:

- The mean distance is 9.55 and Median is 5 by this the data seems to the Right-Skewed because the mean greater than median.

## 5) TAX :

- The mean tax rate is 408.24.The data is right skewed, with a median tax rate of 330.

## 6) PT Ratio (Pupil Teacher Ratio):

- The mean of PT Ratio is 18 and the median of PT Ratio is 19.05 which more than the mean .
- Mean the graph is slightly left-skewed.

## 7) AVG_ROOM(avg number of rooms):

- The average number of rooms are the 6 .

## 8) LSTAT (Percentage age of lower status population):

- The mean age of Percentage age lower status population is 12.65 .

## 9) Average Price (avg home price):

- The mean of avg_price of the house is 22.5 and the mode is 21.2 .
- Which mean the graph of this slightly right skewed.

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

- From the histogram we understand that the more number of variable are present in the

  range between 21-25
- From this we understand the people are more intrested to buy the avg_price between 21-25.
- from the histogram of the average price we can understand the less variable are in the (37,41) and (45,49) of bin range.

## Q.3) Compute the covariance matrix. Share your observations.

- In this covariance we have to understand the relation between the two columns if the covariance value is positive we can say that both columns have positive relation , if the both column have negative value mean they have negative relation means.
- Positive relation means :: if one column value is increasing another column value should also increase.
- Negetive relation means :: if one column value is increasing another column value should also Decrease.

**Q.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) b) Which are the top 3 negatively correlated pairs.**

**The top 3 positive co-relation values are .**

- Distnace and tax.
- Indus and Nox.
- Age and Nox.

**The top 3 negative co-relation values are .**

- **Avg_price and LSTAT.**
- **Avg_room and LSTAT.**
- **PT Ratio and avg_price.**

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

- After plotting the regetion model of avg_price and LSTAT as independent variable by seeing the R Square values Is 0.544 mean that it has an relation between two columns are of 54.4% we can say.
- Means if LSTAT value is changing there might be chances of the value of avg_price also get change.

a) Coefficient value. The coefficient for the LSTAT variable indicates the change in the predicted average price of each one unit change in the LSTAT. Variable in the coefficient is -0.955 suggest that as percentage of lower status population increases the average price range to decrease

b) Intercept the intercept is the predicted average price when the LSTAT percentage is zero. In the most cases, this value me not have meaningful interpretation.

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

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?

This is the formula of straight line with multiple columns .
(Y=m1x1+m2x2+m3x3…......+Mnxn+c)
Y=5.094787984×7+(-0.642358334) ×20+(-1.35827812)
y=21.45807639, y=$21,450
the com pany is quoting $30,000.
The company is overcharging the customer.

a) 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.

- Yes, is model performance is better than the previous model because the adjusted r square value is greater in this model. The adjusted r value of is the pervious 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.**

The adjusted R square value of this model is 0.688298647.

The intercept value is 29.24131526

The coefficient value of CRIME RATE IS 0.04872514

The coefficient Value of AGE is 0.032770689

The coefficient Value of INDUS is 0.130551399

The coefficient Value of NOX is - 10.3211828.

The coefficient Value of DISTANCE is 0.261093575

The coefficient Value of TAX is -0.01440119

The coefficient Value of PTRATIO is - 1.074305348.

The coefficient Value of LSTAT is -0.603486589

The coefficient Value of AVG-PRICE is 4.125409152

**Conclution :**

the significant of independence variables with respect to avg-price,

not all the independent variables affect the outcome with respect to avgprice. there 4 independent variables that affect the outcome of the avg-price are nox, ptratio, avg-room, lstat. the variables that does not affect the avg-price are crime rate, age, indus, tax, distance.

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:

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.

HINT: Significant variables are those whose p-values are less than 0.05. If the p-value is greater than 0.05 then it is insignificant

a)

The output of those are Adjusted R value is 0.688683682.

The intercept value is 29.42847349.

The significant F (P value) is 1.911-122 So, this model is liner equation because the P value is less than 0.05 (P value<0.05)

b)

The adjusted R value of pervious model is 0.688298647.

The adjusted R value of is model is 0.688683682.

TERRO PROJECT Current model performance better than the previous model because Current model have larger P value compare to this model.

**c)**

The coefficient of independent variables is ascending order are.

NOX is -10.27276698

2) PTRATIO is -1.071282734

3) LSTAT is -0.6015159282

4)TAX is -0.015009898

5)AGE IS 0.03293496

6)INDUS is 0.130701866

7)DISTANCE is 0.261529529

8) AVG- ROOM IS 4.12451678461 If the value of NOX increases on the locality the price of AVG-PRICE of a property decreases because NOX have negative relationship with the AVG-PRICE.

d)

The regression equation from this model is

y=m1x1+m2x2+m2x2+m3x3+m4x4+m5x5+6x6+m7x7+c

this is the regression equation of this model.