**BUSINESS REPORT**

TERRO’S REAL ESTATE AGENCY

By SONAL KUMARI

**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.

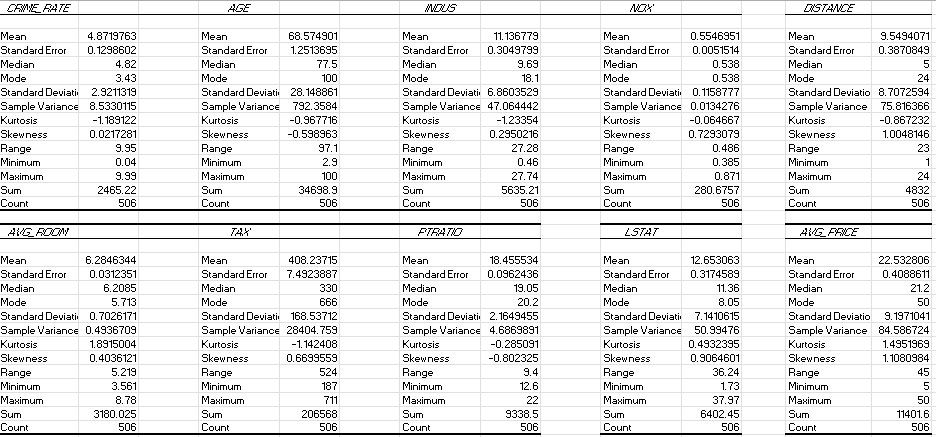
**Objective (Task):**

Our 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.

**To do the analysis, we are expected to solve these questions**:

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

**-** summary statistics for each table:-

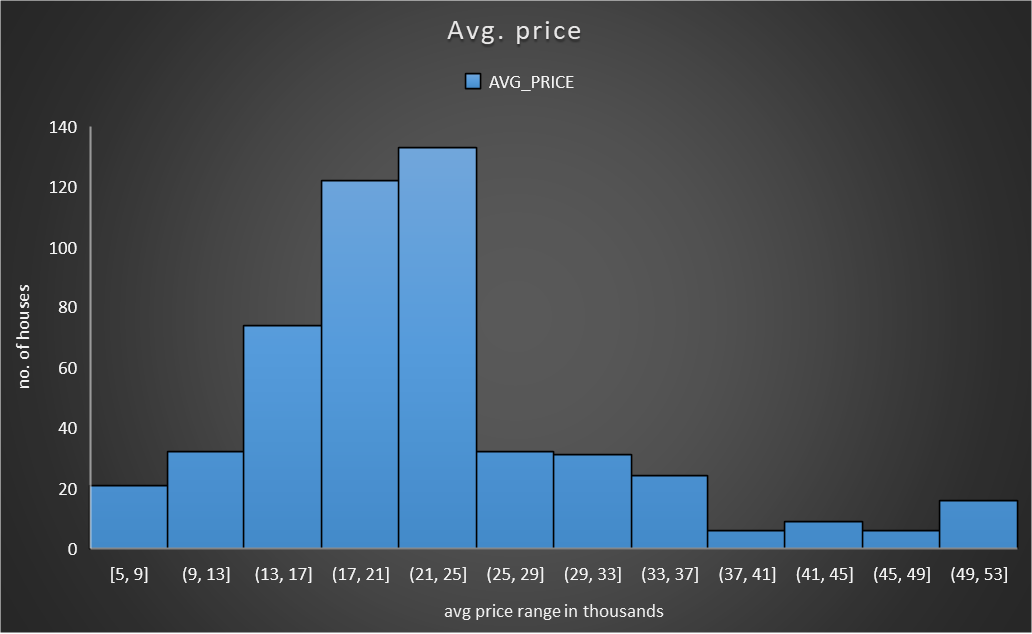
****

From descriptive statistics of the given dataset we can get few observations as:

* The number of records given in the dataset are 506.
* Firstly if we consider Distance variable we can analyse that maximum distance is 24 and has mode as 24. Which says that most of the houses are away from “highway”.
* The average tax paid is 408.2 and tax range is 524.
* From the skewness of variables we can say that dataset is highly skewed.
* And if we consider age variable the maximum age is 100 and mode is also 100 which says that most of the houses has age of 100.
* All the given variables except the PT-ratio have a Positive skewness.
* There is minimum outlier data in Crime rate, NOX and Avg. Room

**2. Plot a histogram of the Avg\_Price variable. What do you infer?**

**-**histogram of avg\_price-



From above Histogram, we can infer:

* 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.

* AVERAGE PRICE is a dependent variable, which has a very few outlier data and it is Rightly skewed with a leptokurtic kurtosis.
* There are 133 Houses with the price range of $ 21000 to $ 25000.

1. **Compute the covariance matrix. Share your observations**

Covariance matrix

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *CRIME\_RATE* | *AGE* | *INDUS* | *NOX* | *DISTANCE* | *TAX* | *PTRATIO* | *AVG\_ROOM* | *LSTAT* | *AVG\_PRICE* |
| 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 | 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\_ROOM | 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\_PRICE | 1.162012 | -97.3962 | -30.4605 | -0.45451 | -30.5008 | -724.82 | -10.0907 | 4.484566 | -48.3518 | 84.41956 |

Observations:-

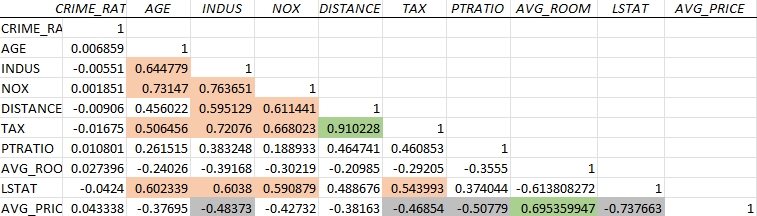
* As we can see there is high covariance value for some of the features which tells that they are highly correlated and explains the variability of the other features.
* We can see that tax variable has high covariance values with each other feature except crime rate. That means tax explains a very good variability with other features.

**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**.

Correlation matrix

****

1. Top 3 positively correlated pairs:-

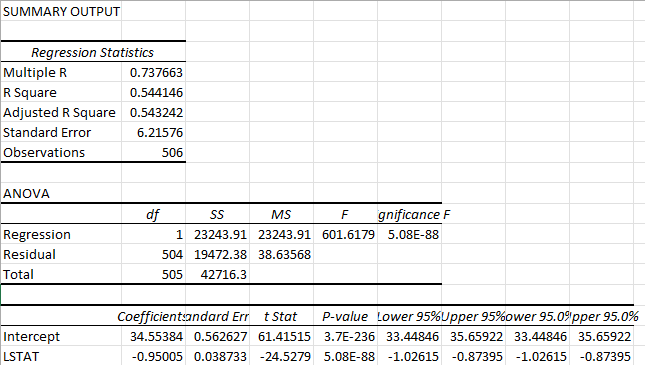
* Tax and Distance
* Indus and Nox
* Age and Nox

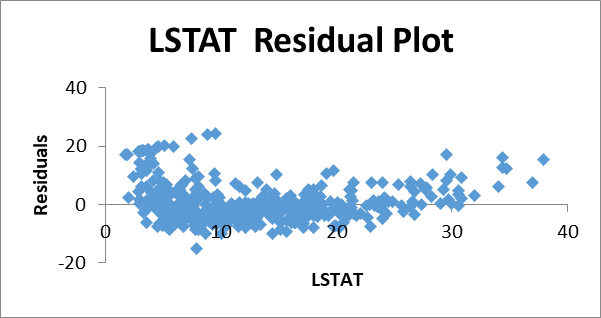
1. Top3 negatively correlated pairs:-

* Avg\_price and Lstat
* Avg\_room and Lstat
* Avg\_price and PT ratio

1. **Build an initial regression model with AVG\_PRICE as ‘y’ (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot.**
2. **What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?**
3. **Is LSTAT variable significant for the analysis based on your model?**

Regression model:

****



**A.**

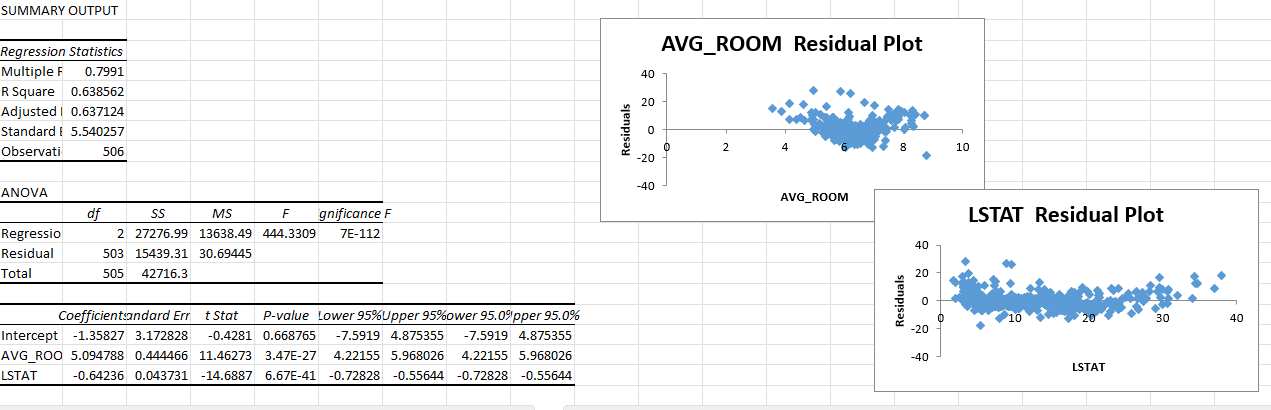
* The value of the variance is less than 0.05, so the model is highly significant.
* From the residual plot, we can see that there is some outlier in a data.
* From this model 54% of the variation in the average price is explained by the LSTAT.
* The coefficient of LSTAT for the model is -0.950049354.This says that if LSTAT increases by 0.9 times the averageprice of house decreases 0.9times.

**B.**

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

1. **Build a new Regression model including LSTAT and AVG\_ROOM together as Independent variables and AVG\_PRICE as dependent variable.**
2. **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?**
3. **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 .**

Regression model :

****

**A)**

BY using a regression equation, the AVG\_PRICE value is 21.4581. Comparing to the quoting value, the company is overcharging for the property.

* Regression Equation we obtained for this model is :

y = -1.358 +5.09 X0 -0.642 X1

Where y=Avg\_price

X0 = avg\_room

X1 = LSTAT

As per the model, avg\_price for new house can be calculated as

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

So ,the price for the new house is $21440 .

**B)**

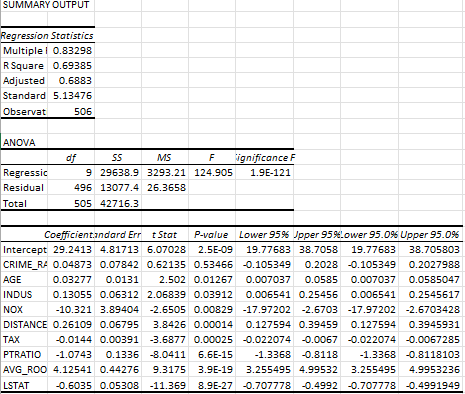
* Previous model Adj. R-square= 0.543241825954707

New model Adj. R-square= 0.637124475470123

* Yes, the performance of this new model is better than the previous model by comparing to the value of adjusted R-square.

1. **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 model:

****

By seeing the coefficient from the regression data, we can say that:-

• Avg. price and Avg. room variables are directly proportional.

• NOX and Avg. Price are inversely proportional to each other.

The significance variables in this model are:

• AGE

• INDUS

• NOX

• DISTANCE

• TAX

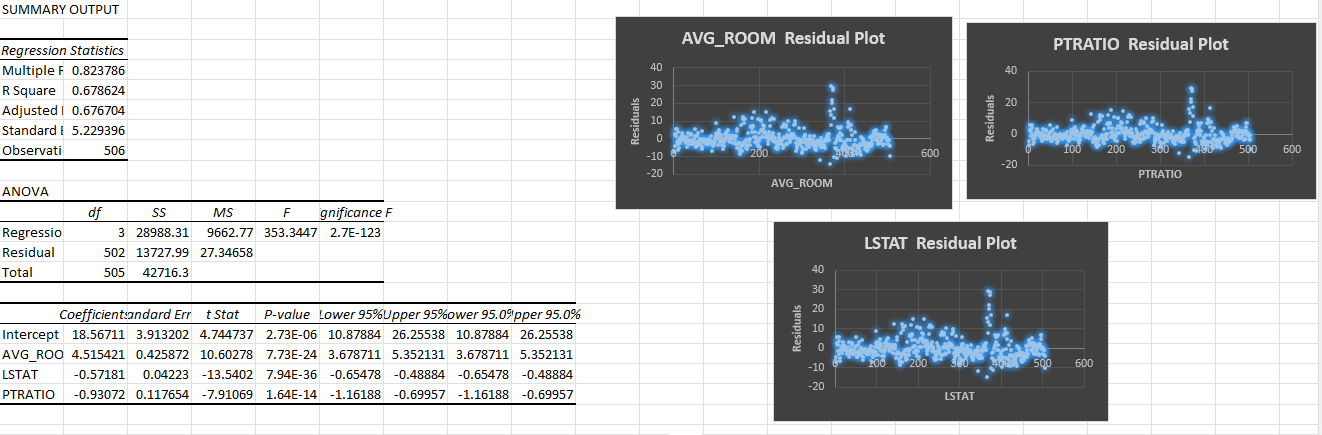
• PTRATIO

• AVG\_ROOM

• LSTAT

1. **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:**
2. **Interpret the output of this model.**
3. **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?**
4. **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?**
5. **Write the regression equation from this model.**

Regression model:

****

**A)** This model has a Better Adjusted R-square value of 67.67.

1. The New model performs well than the previous model by comparing the value of an Adjusted R-square value.

|  |  |
| --- | --- |
| **Column1** | **Column2** |
| 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 |

**C**)

**D)** Multi-Linear regression equation: Y=m1x1+ m2x2+ m3x3 + c

The AVERAGE\_PRICE is HIGH, if the NOX is LOW

Multi-linear regression equation of this model:

Y= 4.515420944 X AVG\_room - 0.571805688 X LSTAT - 0.930722555 X PTRATIO

Y= Predicted Variable

m= weight of the variable

x= Independent variable

c= Intercept