TERRO’S REAL ESTATE AGENCY Real estate data analysis – Exploratory data analysis, Linear Regression

JIJo .G

JIjonobi@gmail.com

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

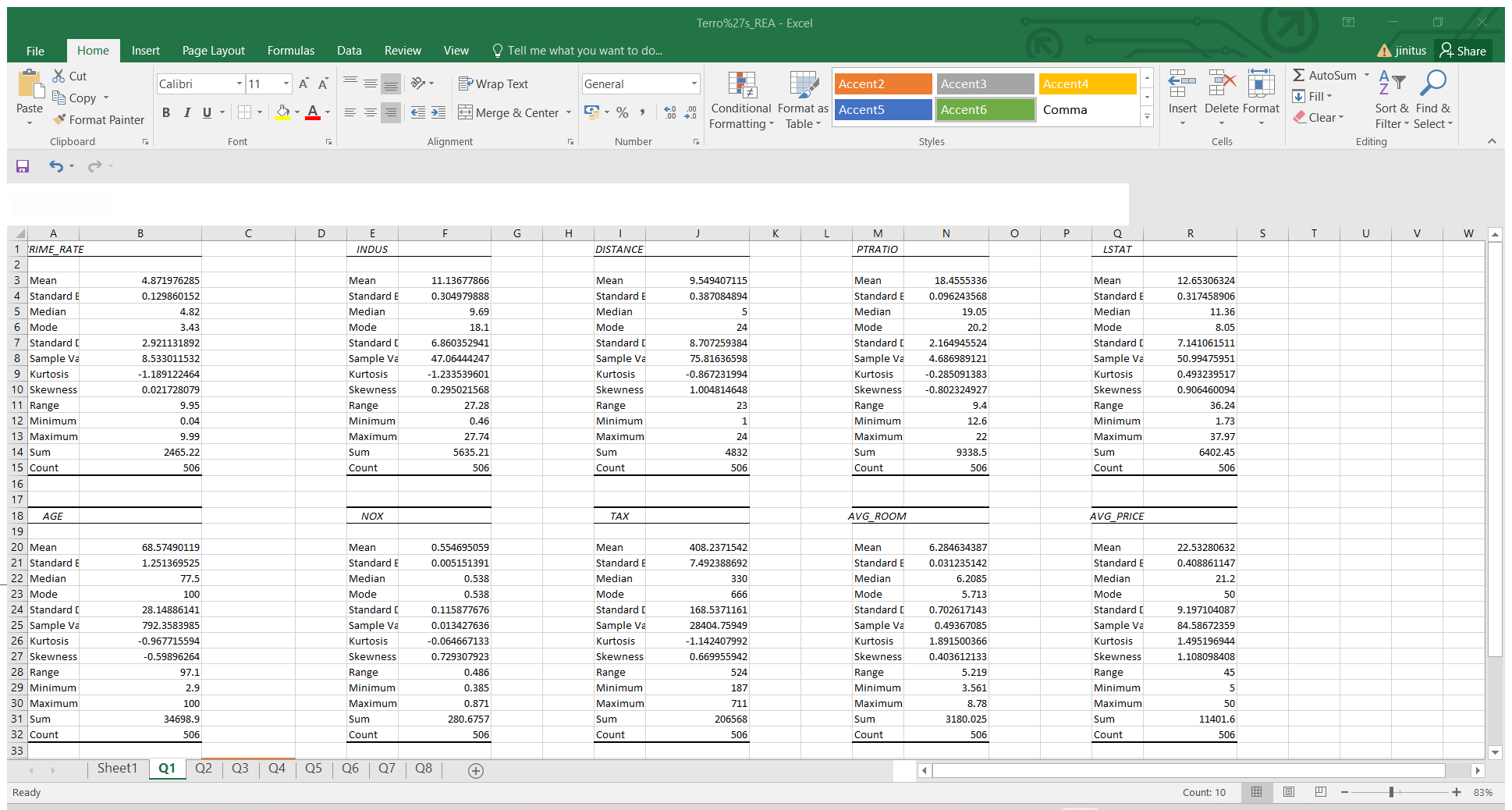
Ans:-

(i) Nitric oxide concentration (NOX) has the lowest standard error whereas TAX has the highest.

(ii) The highest Sample Variance is for TAX, while the lowest is for NOX.

(iii) The biggest standard deviation is 168.537 for TAX, and the lowest is for NOX.

(iv) While PTRATIO has a low skewness, AVG PRICE has a high skewness.

(v) AVG ROOM has the most capacity, whereas Kurtosis INDUS has the least. 

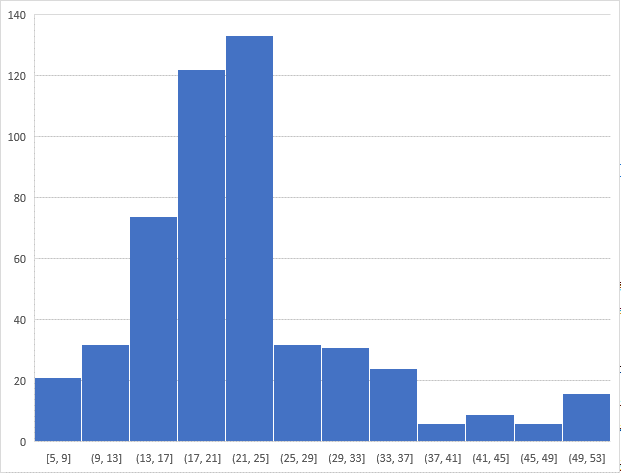
2) Plot a histogram of the Avg Price variable. What do you infer?

By altering a bin count, I created a compact graphic that groups data so that it is easier to compare price ranges.

(i) We may deduce that there is a greater count of the average price between the range from the histogram that was plotted (14,23).

(ii) THE RANGE CONTAINS THE MINIMUM COUNT OF AVERAGE PRICE (41,50).

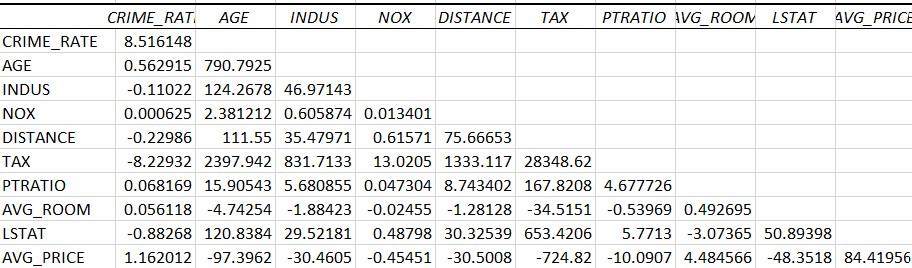
(iii) We can also see that the average price is somewhere in the middle of the range (23,32).



3) Compute the covariance matrix. Share your observations?

1) THE HIGHLY POSITIVE COVARIANCE BETWEEN AGE AND TAX.

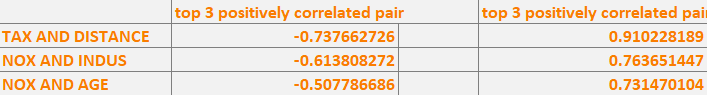
2) HIGH NEGATIVE COVARIANCE EXISTS BETWEEN AVG PRICE AND TAX.

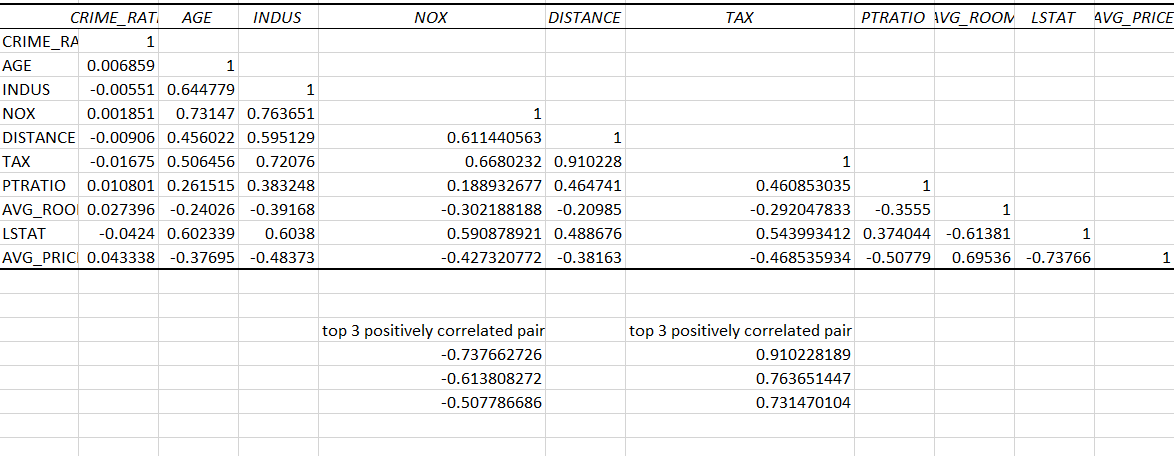


4) Create a correlation matrix of all the variables (Use Data analysis tool pack).

a) Which are the top 3 positively correlated pairs and

b) Which are the top 3 negatively correlated pairs.





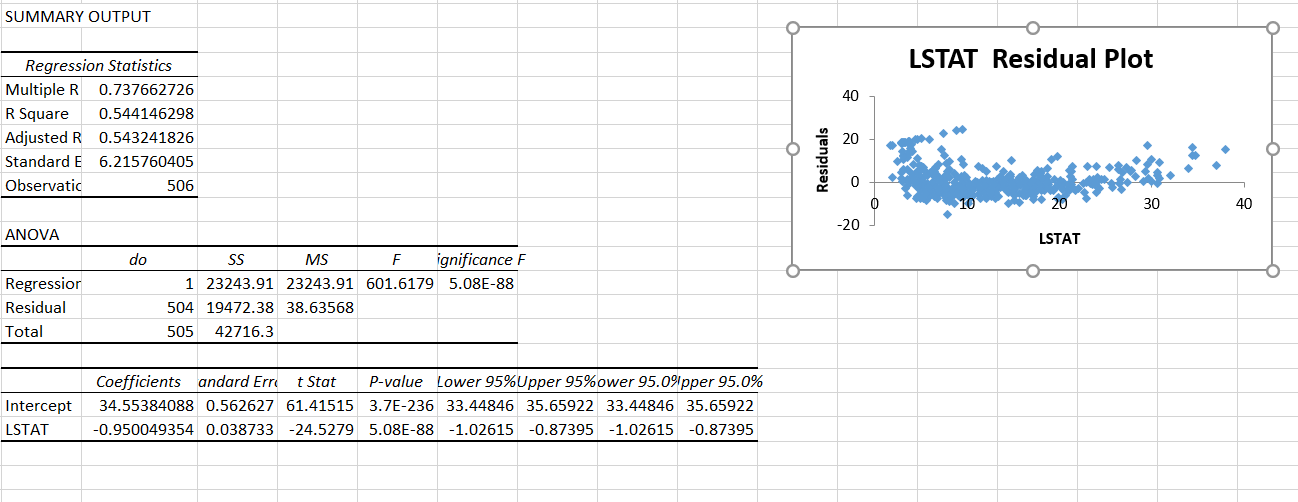
5) Build an initial regression model with AVG\_PRICE as ‘y’ (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot.

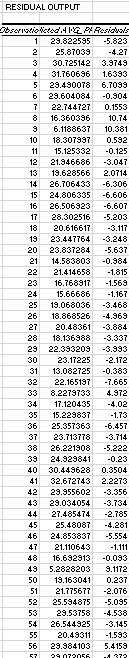
a) What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?

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

 Intercept and avg price is positively related and LSTAT and avg price is negatively related.

 LSTAT variable and avg price are positively corelated thus it is significant for our analysis.





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?

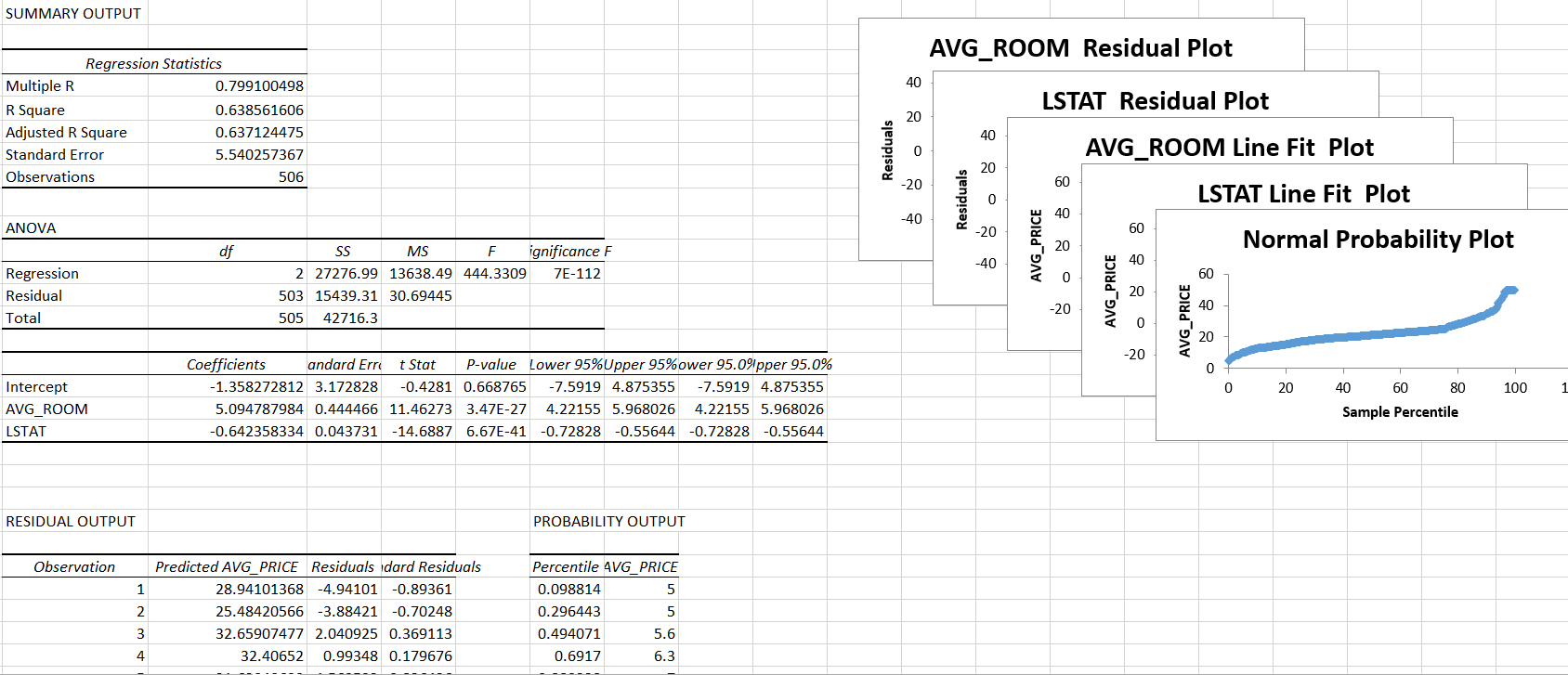
b) Is the performance of this model better than the previous model you built in Question 5? Compare in terms of adjusted R-square and explain?

Ans:-

1. REGRESSION EQUATION: y=(m1\*7)+(m2\*20)+b

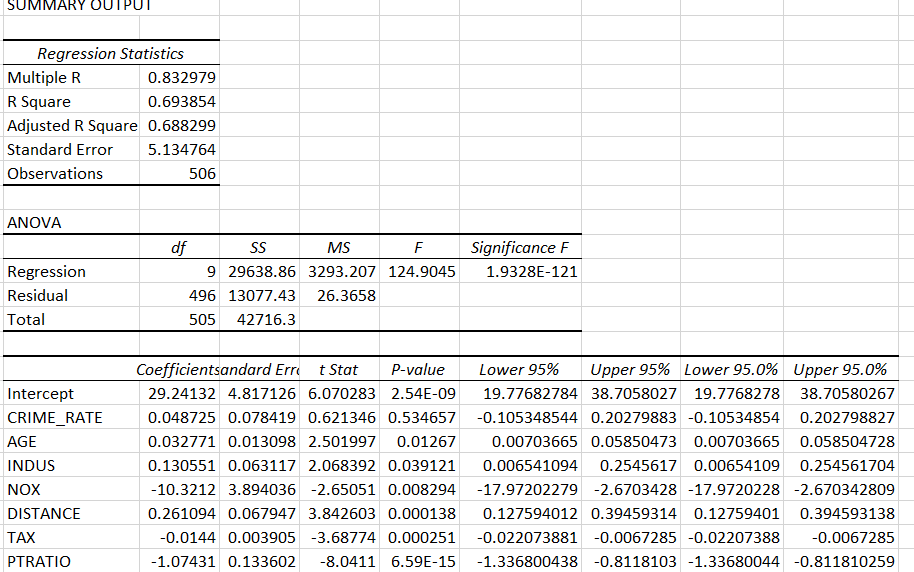
Yes, the company overcharges by 9,000 percent.

1. Yes, the performance is better, and the adjusted R square value has improved. Because in the previous question only one independent variable was taken into consideration, this question considers two independent variables, which has a lower P-value.



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.

|  |
| --- |
| Comparing this model to earlier question models, the adjusted R-square value increased, demonstrating that the variables taken into account in this model are highly significant with dependent variables. |
|  |
| AT THE POINT 29.24132, THE FUNCTION CROSSES, AS INDICATED BY THE INTERCEPT VALUE. |
| LSTAT, NOX, and PT Ratio all have negative slopes. |
| THIS MODEL'S STANDARD ERROR IS LESS. |
| WHERE WHICH CAN LEAD TO GOOD ACCURACY |
|  |
| Given that it has a high P-value, the crime rate is less significant in comparison to the average price. |
| Age can be significant because it has lower standard errors and P values. |
| Due to INDUS, DISTANCE, AVG ROOM'S LOW P-VALUE AND LESS STANDARD ERROR, IT IS VIEWED AS A SIGNIFICANT VARIABLE WITH RESPECT TO Y. |
| NOX HAS A SLIGHTLY MORE ERROR WITH RESPECT TO Y THAN NOX, TAX, PTRATIO, OR LSAT AND ALL HAVE A NEGATIVE SLOPE. |



8) B) THIS MODEL HAS SLIGHTLY MORE R QUARE VALUE THAN WITH THE PREVIOUS QUESTION MODEL.

IT ALSO GIVES GOOD ACCURACY RANGE.SO IT CAN PERFORM BETTER.

C) -10.27270508

-1.071702473

-0.605159282

-0.014452345

0.03293496

0.130710007

0.261506423

4.125468959

29.42847349

IF THE NO0X RATE IS HIGH IN THE LOCALITY ,THE AVG PRICE GETS REDUCED,BECAUSE PEOPLE WON'T

LIKE TO RESIDE IN HIGH POLLUTION AREA.

D) Y=(M1\*X1)+(M2\*X2)+(M3\*X3)+(M4\*X4)+(M5\*X5)+(M6\*X6)+(M7\*X7)+(M8\*X8)+B

