**BUSINESS REPORT**

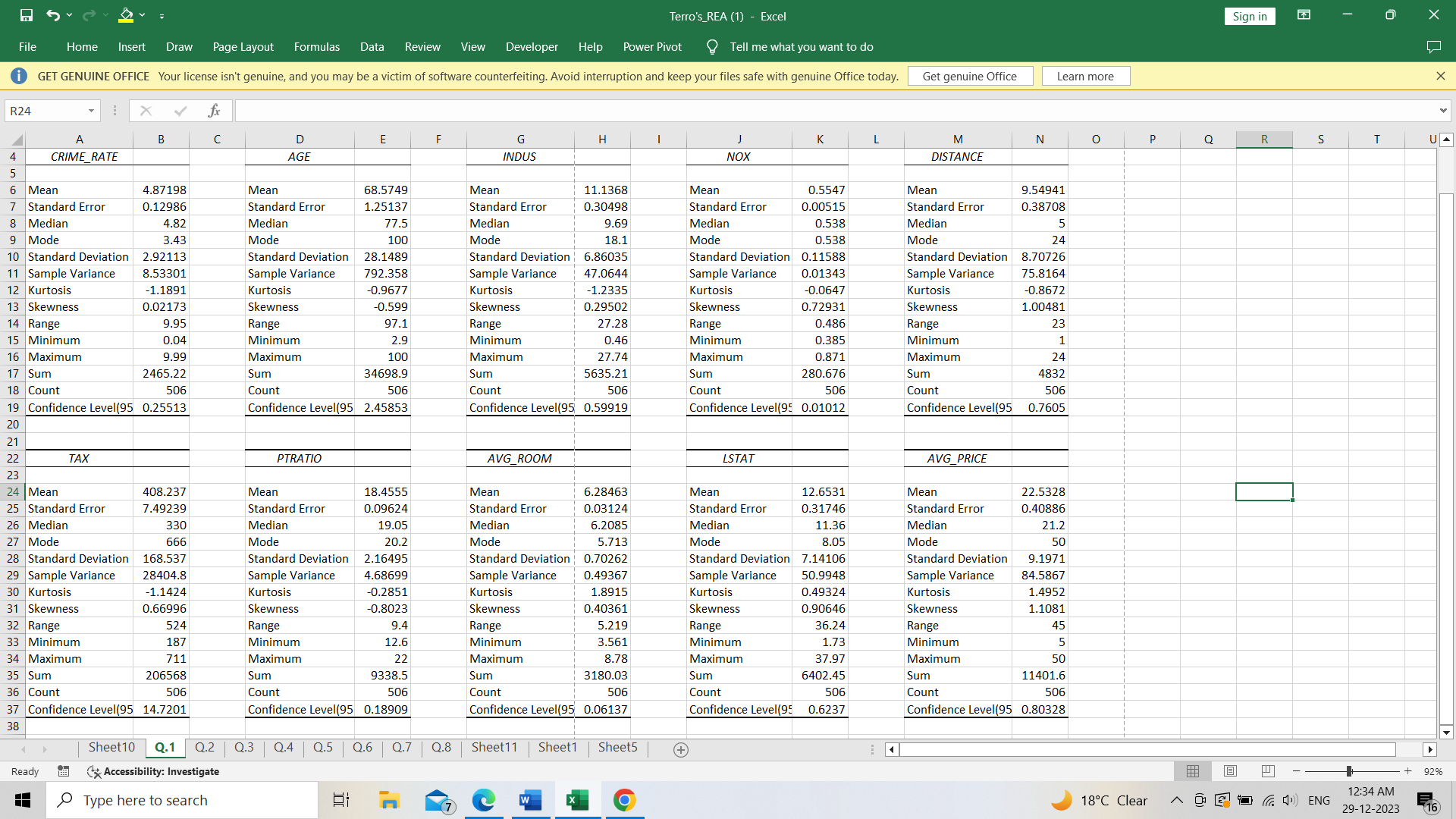
***Terro’s Real Estate Agency***

By,

Aniket Gurubaxani

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

Ans.

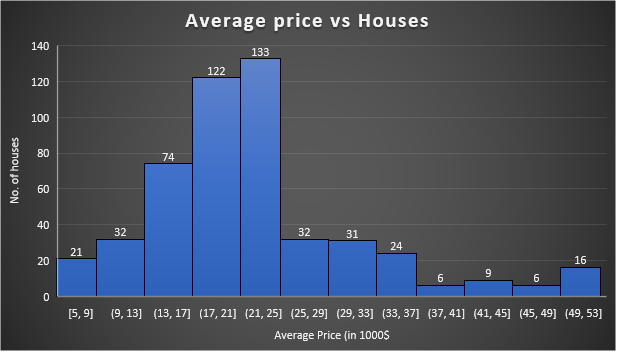


Observations:

* From above summary statistics I observed that TAX variable has highest standard deviation (168.53) which means TAX variable has greater dispersion of data points from its mean than any other variable.
* Comparing with the normal distribution AVG\_ROOM has the sharp peak because it has higher positive kurtosis (1.8915) and INDUS variable has flatter peak having negative kurtosis (-1.2335).
* AVG\_PRICE is most positively skewed variable with skewness of 1.1081. So, most of distribution of this variable concentrated on the left side of the mean. PTRATIO is the most negatively skewed variable with skewness of -0.8023. So, most of distribution of this variable concentrated on the left right of the mean.
* NOX variable has smallest standard error (0.00515) which indicates higher precision and TAX variable has larger standard error (7.49239) which indicates lower precision and shows greater variability in distribution.

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

Ans.



* From above histogram of Avg\_Price variable I concluded that graph showing longer tail of data points on the right side which means Avg\_Price is the positively skewed variable with skewness of 1.1081.
* The most of data points are concentrated on the left side of the mean.

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

Ans.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *AGE* | *INDUS* | *NOX* | *DISTANCE* | *TAX* | *PTRATIO* | *AVG\_ROOM* | *LSTAT* | *AVG\_PRICE* |
| AGE | 790.7925 |  |  |  |  |  |  |  |  |
| INDUS | 124.2678 | 46.97143 |  |  |  |  |  |  |  |
| NOX | 2.381212 | 0.605874 | 0.013401 |  |  |  |  |  |  |
| DISTANCE | 111.55 | 35.47971 | 0.61571 | 75.66653 |  |  |  |  |  |
| TAX | 2397.942 | 831.7133 | 13.0205 | 1333.117 | 28348.62 |  |  |  |  |
| PTRATIO | 15.90543 | 5.680855 | 0.047304 | 8.743402 | 167.8208 | 4.677726 |  |  |  |
| AVG\_ROOM | -4.74254 | -1.88423 | -0.02455 | -1.28128 | -34.5151 | -0.53969 | 0.492695 |  |  |
| LSTAT | 120.8384 | 29.52181 | 0.48798 | 30.32539 | 653.4206 | 5.7713 | -3.07365 | 50.89398 |  |
| AVG\_PRICE | -97.3962 | -30.4605 | -0.45451 | -30.5008 | -724.82 | -10.0907 | 4.484566 | -48.3518 | 84.41956 |

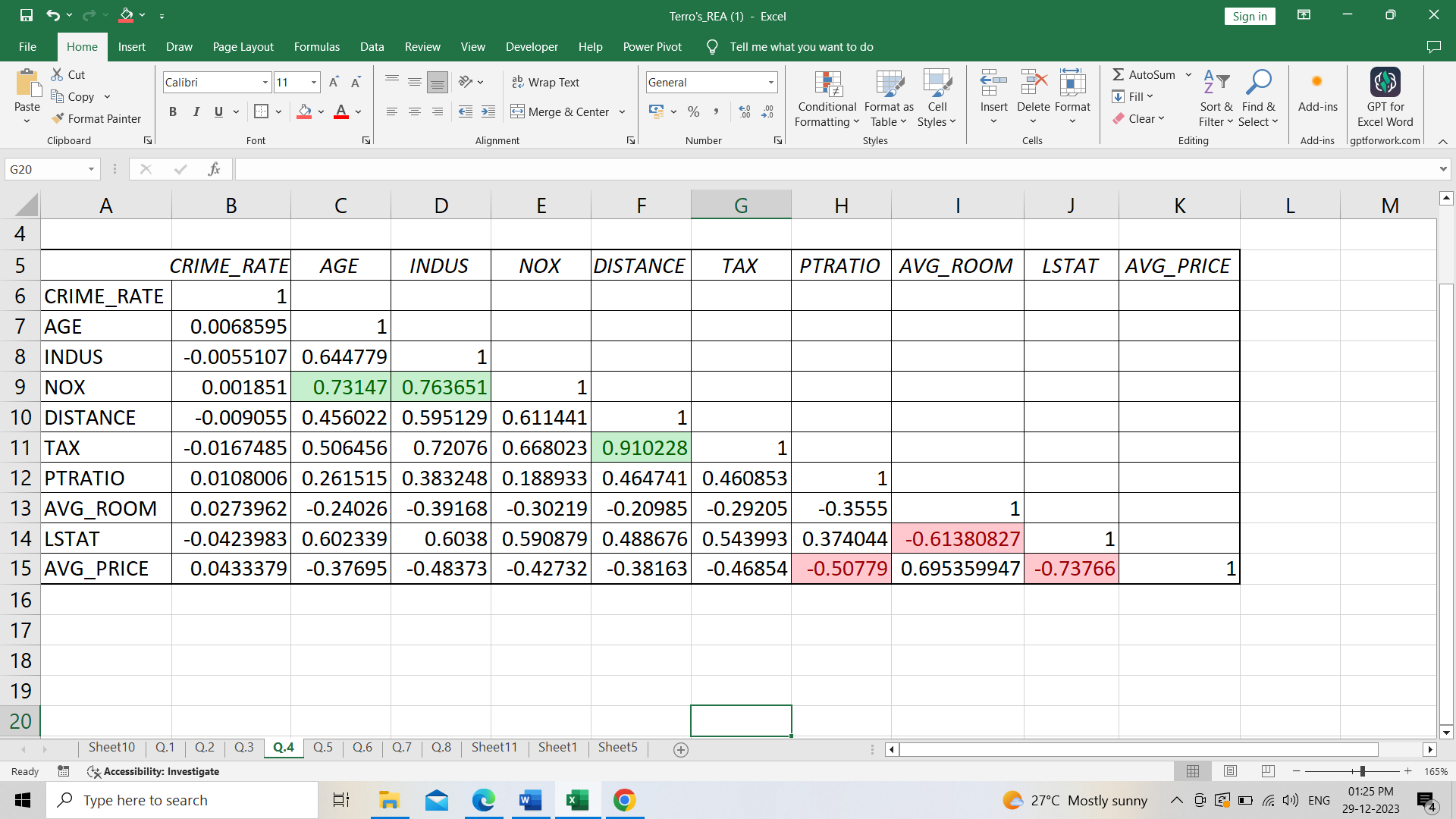
* From covariance matrix of all variables, I observed that the TAX and AGE has the highest positive covariance which means that both variables are moving in the same direction.
* TAX and AVG\_PRICE has the negative covariance which indicates variables are moving in inverse direction. If TAX will increase AVG\_PRICE will decrease.
* From covariance we can indicate only direction of variables not dependency between the variables.

**Q.4)** Create a correlation matrix of all the variables.

a) Which are the top 3 positively correlated pairs and

b) Which are the top 3 negatively correlated pairs.

Ans.

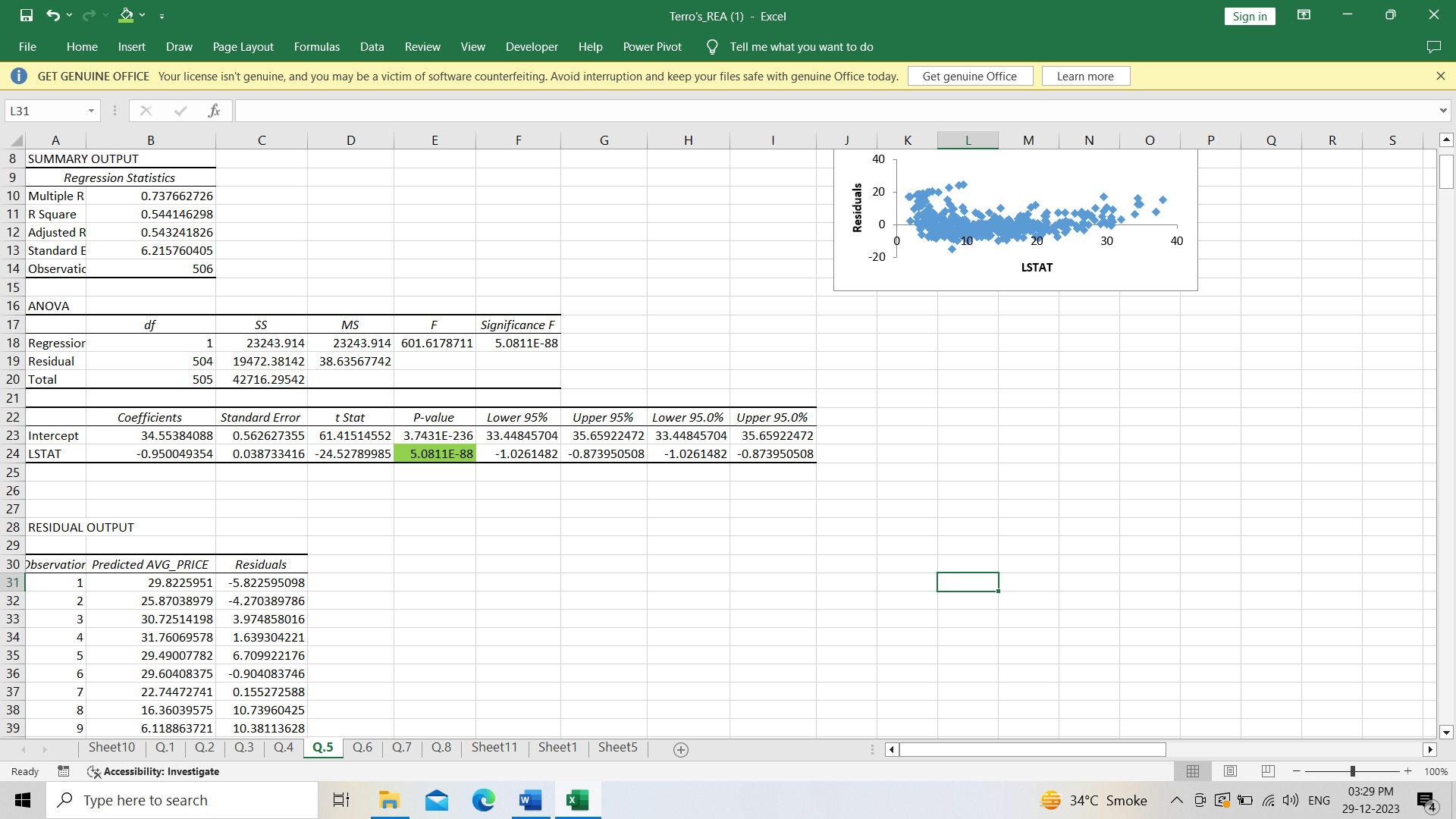


a) Top 3 positively correlated pairs are Distance & Tax (0.910228), Indus & NOX (0.763651) and AGE & NOX (0.73247).

b) Top 3 negatively correlated pairs are LSTAT & AVG\_PRICE (-0.73766), AVG\_ROOM & LSTAT (-0.6138) and PRATIO & AVG\_PRICE (-0.50779).

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

Ans.



Above regression model determines the strength and relationship between one dependent and one independent variable.

a) On the basis of R2, the value of R2 is 0.54414 which is not so good value. 55% variation of dependent variable values around the mean are explained by the values of independent variables. It means only 55% of the values are fitting in model. The coefficient value of intercept is positive (34.55) but it is far away from the 0. The coefficient value of LSTAT is negative (-0.95) it means both variables are moving in opposite directions. If the value of LSTAT increase, AVG\_PRICE will get decreased. Also, residual plot is not showing any pattern.

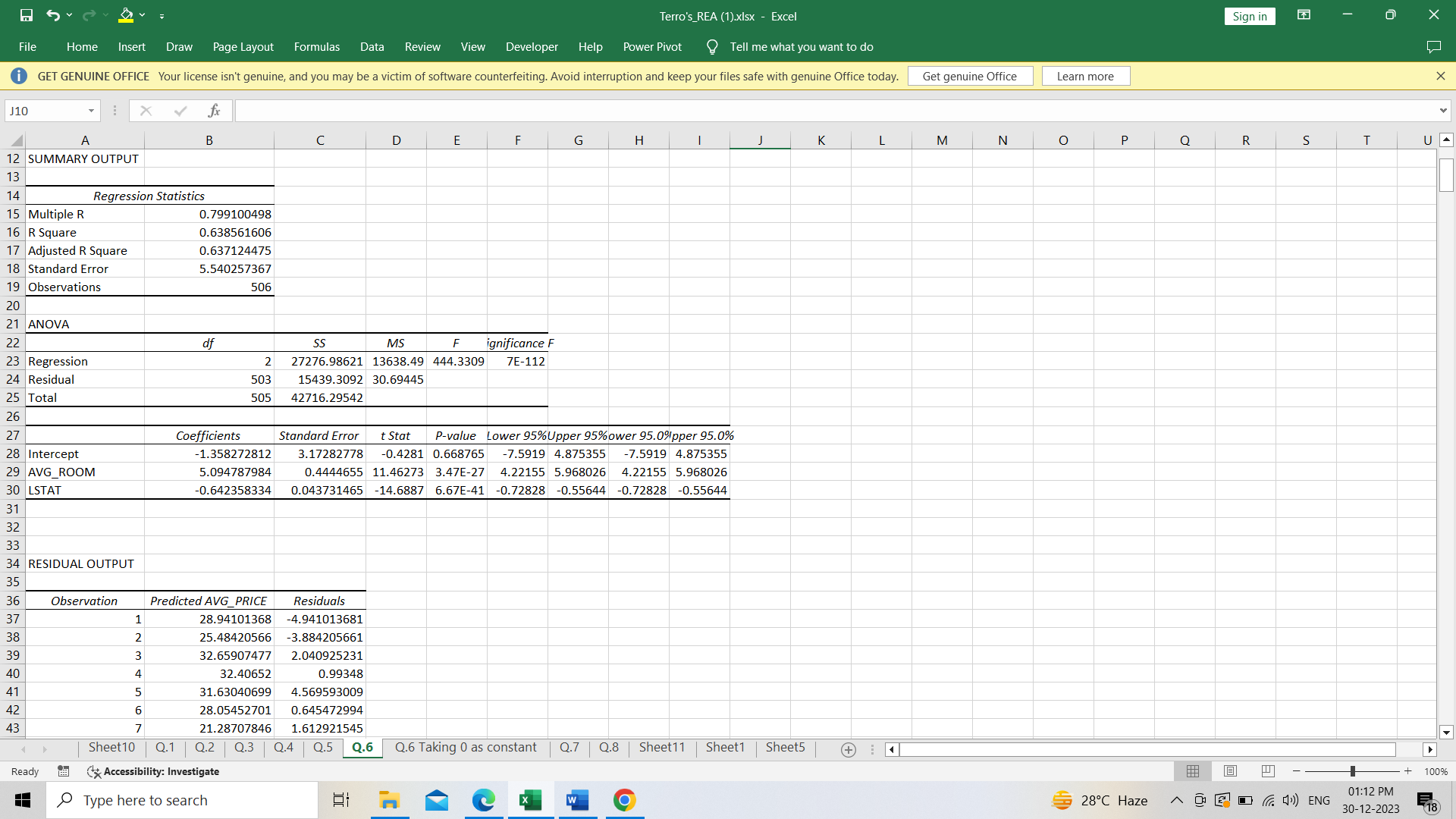
b) LSTAT value is statistically significant variable for analysis because the P-value is less than 0.05.

**Q.6)** Build a new Regression model including LSTAT and AVG\_ROOM together as independent variables and AVG\_PRICE as dependent variable.

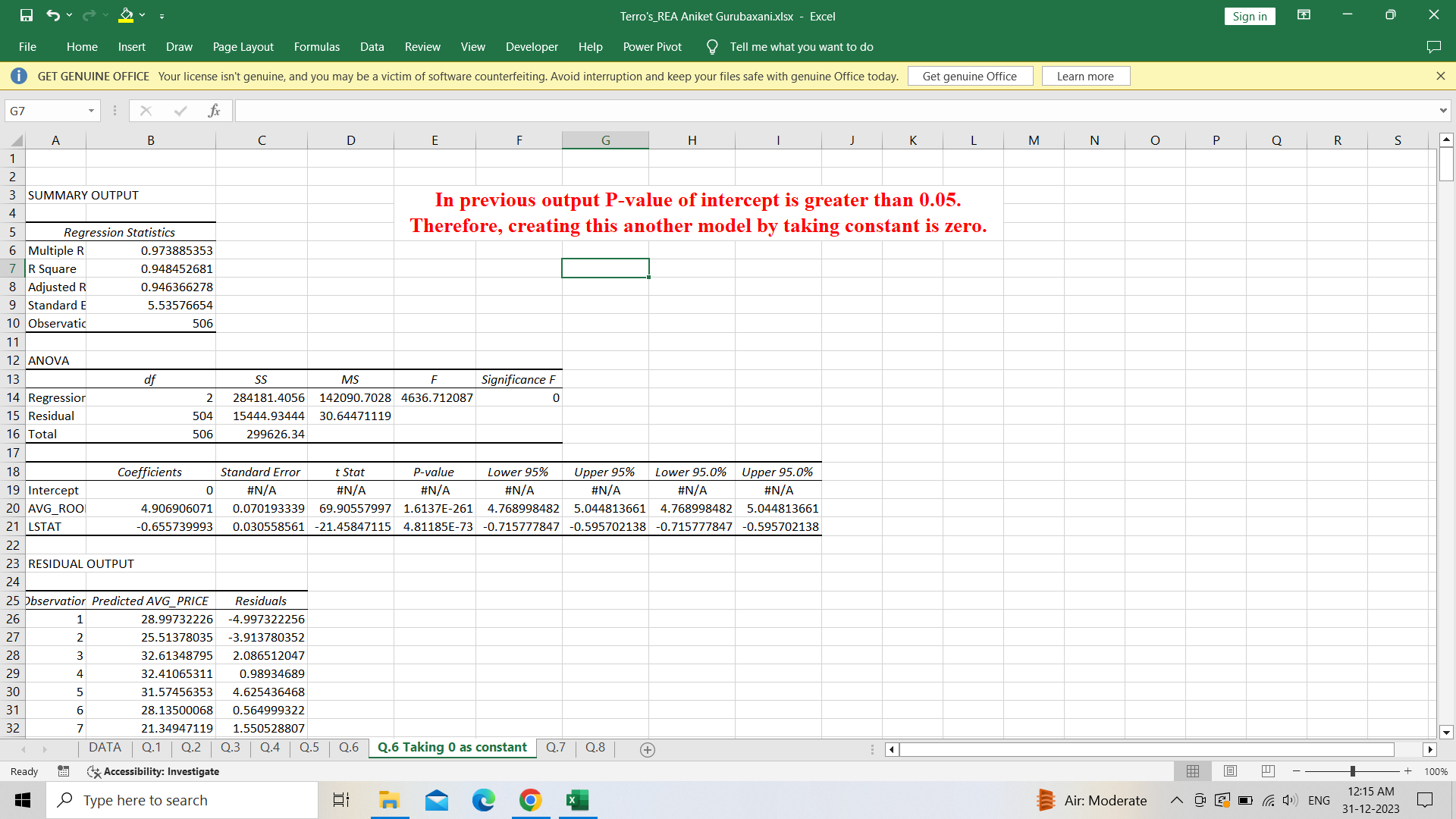
a) Write the Regression equation. If a new house in this locality has 7 rooms (on an average) and has a value of 20 for L-STAT, then what will be the value of AVG\_PRICE? How does it compare to the company quoting a value of 30000 USD for this locality? Is the company Overcharging/ Undercharging?

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.



In above model the P-value of intercept is greater than 0.05. So, I made another model by taking constant is zero.



a) Regression equation: **AVG\_PRICE = 4.9069 \* AVG\_ROOM – 0.6557 \* LSTAT + 0**

If a new house in this locality has 7 rooms (on an average) and has a value of 20 for L-STAT, then average price is

Avg\_Price = 21.234 K $ = 21234 USD

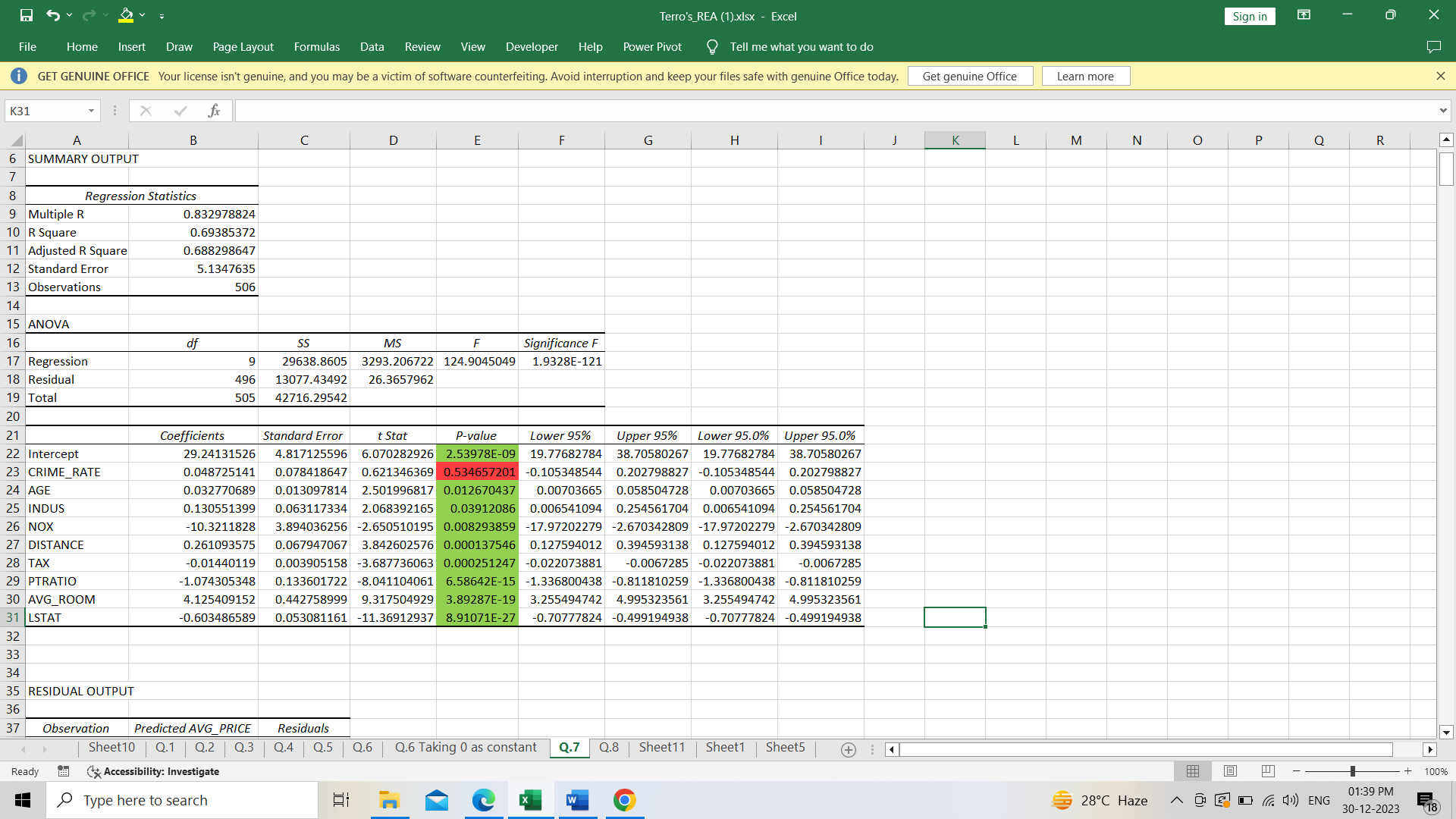
So, the predicted Avg\_Price is less than company quoting price which is 30000 USD.

Therefore, the company is overcharging.

b) The value of adjusted R2 of this model is greater than previous model. So, this model explains better character of relationship between dependent and independent variables.

**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 R square, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG\_PRICE.

Ans.

****

* The variables which have negative coefficients are NOX, TAX, PTRATIO, LSTAT, if the coefficient value of these variables increases then AVG\_PRICE will get decrease. Except these variables remaining other variables have positive coefficient which are directly proportional to dependent variable. The value adjusted R2 is 0.6882, which means approximately 69% of values fits the model.
* From all the variables, CRIME\_RATE is the only variable having P-value greater than 0.05 which is not statistically significant to AVG\_PRICE. Remaining other variables have P-value less than 0.05 so they are statistically significant for predicting AVG\_PRICE.

**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:

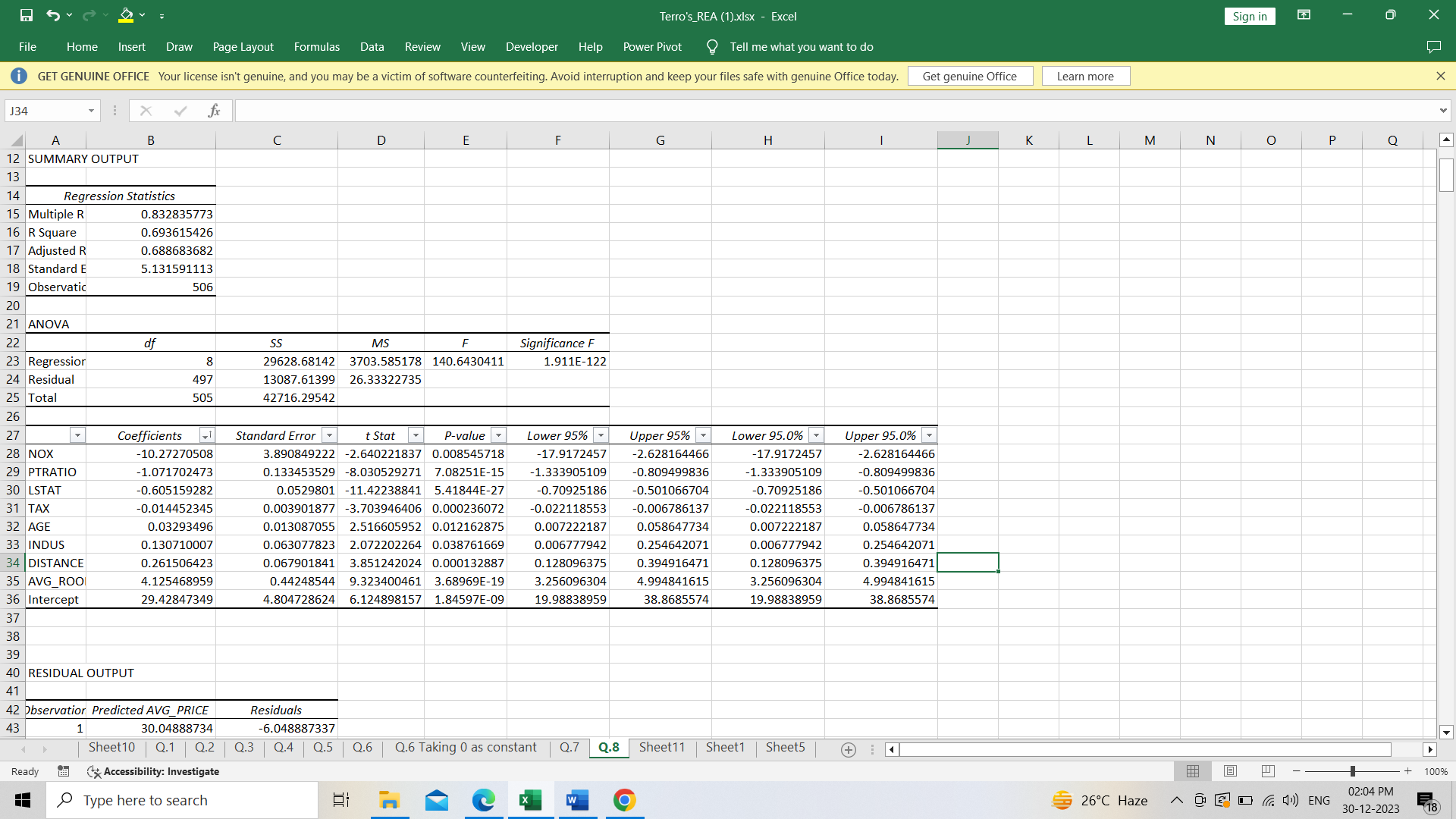
a) Interpret the output of this model.

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.

Ans.



a) Interpretation made from this model is that all the variables are statistically significant for the predicting the value for dependent variable.

b) The value of R2 is similar to the previous model but the value of adjusted R2 is little higher than previous model. So, this model will perform better than previous.

c) Sorted the values of the coefficients in ascending order, if the value of NOX increases in locality in this town, the value of AVG\_PRICE will decrease, because the value of coefficient of NOX is negative.

d) Regression equation:

**AVG\_PRICE = -10.272 \* NOX - 1.071 \* PTRATIO - 0.605 \* LSTAT - 0.014 \* TAX + 0.032 \* AGE + 0.13 \* INDUS + 0.261 \* DISTANCE + 4.125\* AVG\_ROOM + 29.428**.