**BUSINESS REPORT**

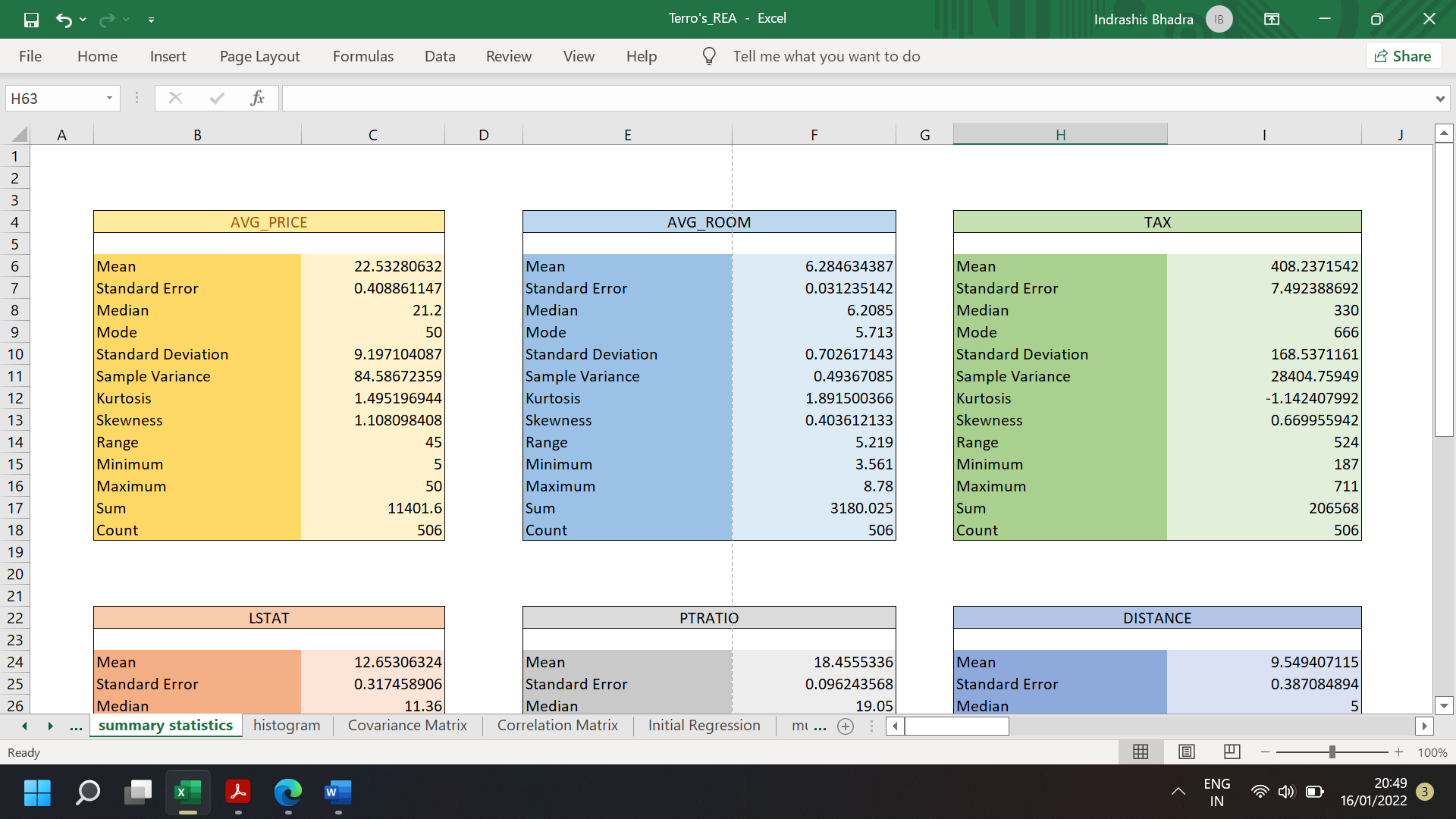
**TERRO’S REAL ESTATE AGENCY**

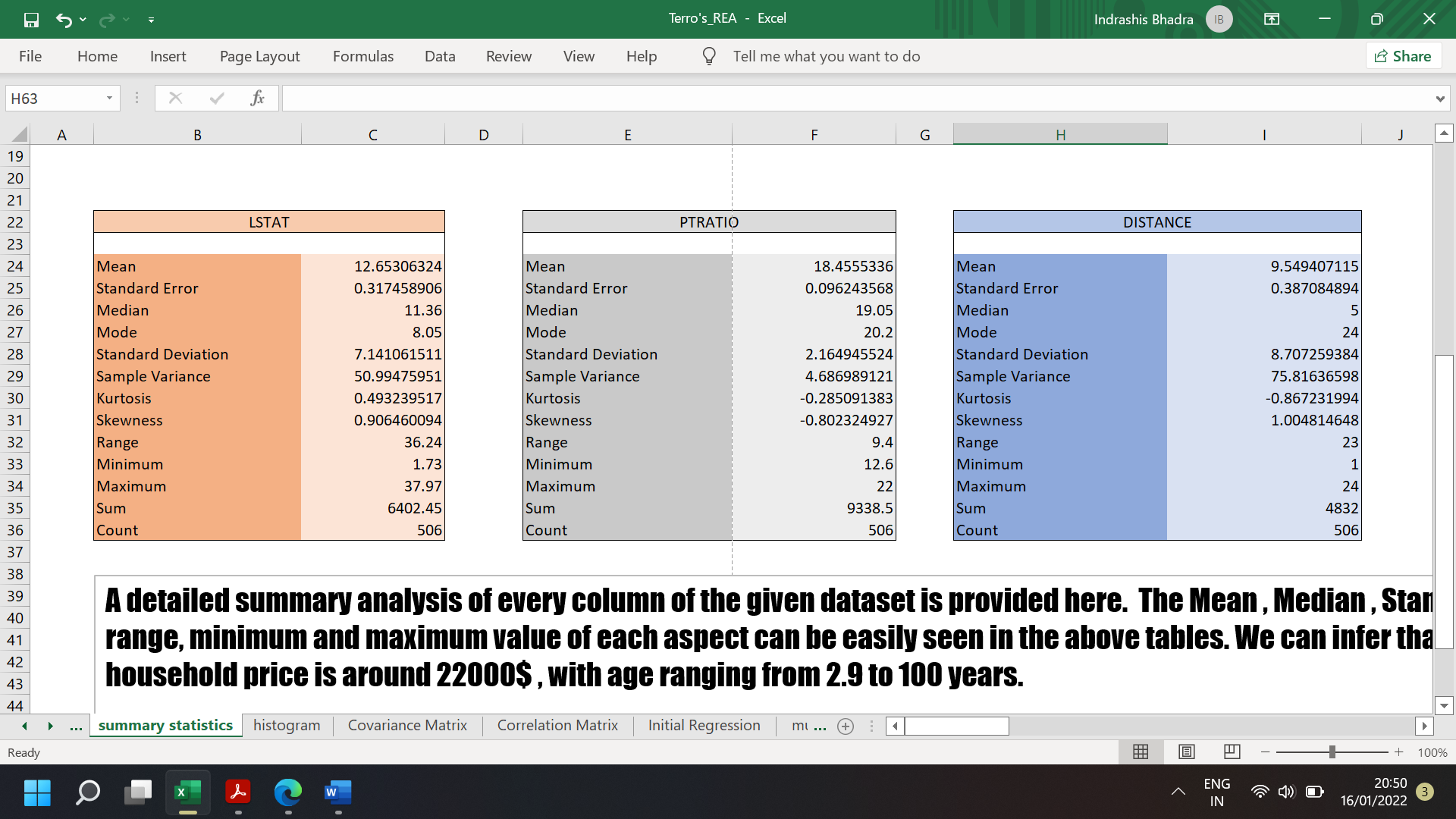
****

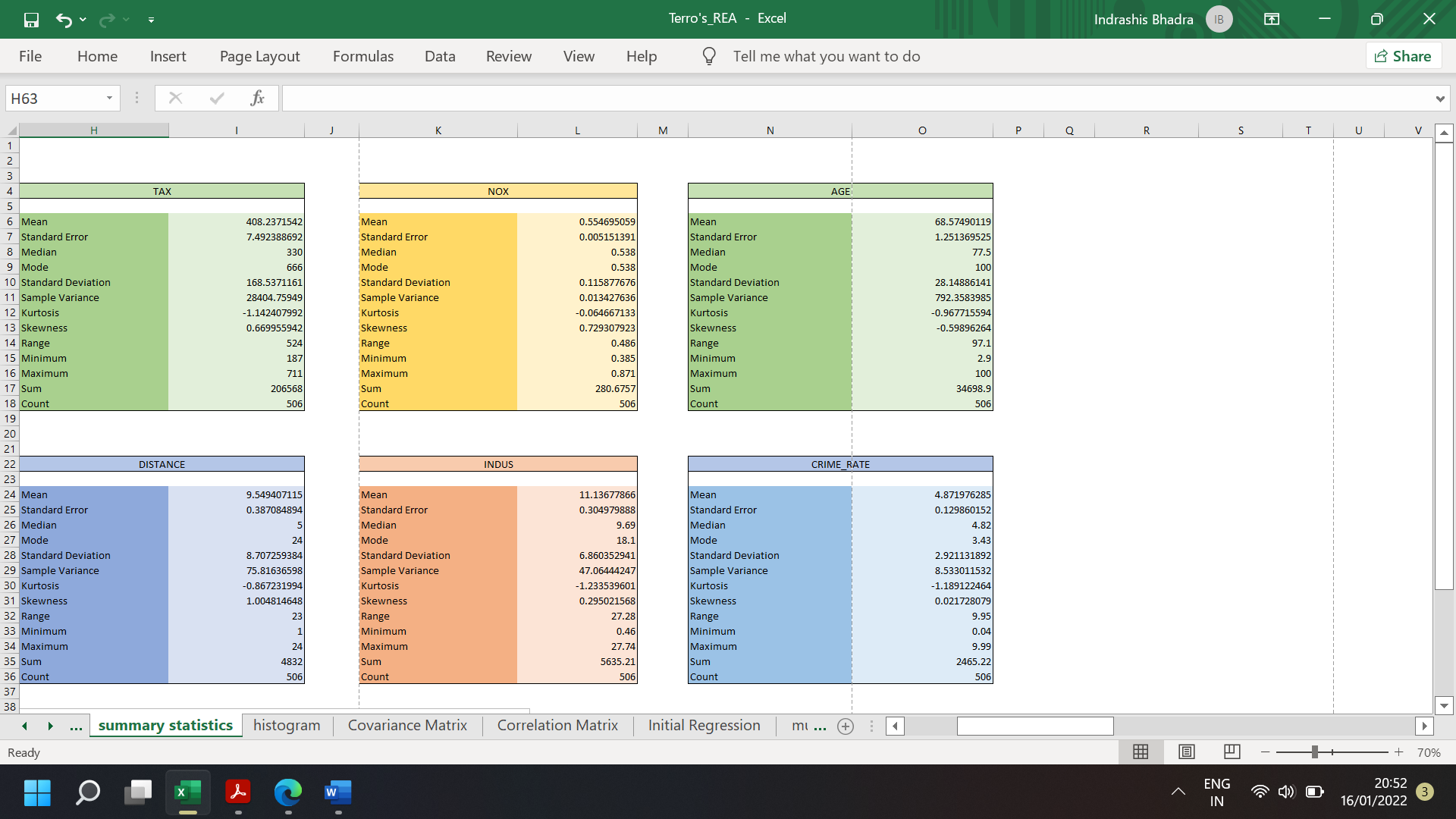
****

**INDRASHIS BHADRA (AUDITOR)** [**indrashisbhadra@gmail.com**](mailto:indrashisbhadra@gmail.com)

1. **Summary statistics for each of the variables. What do you observe?**



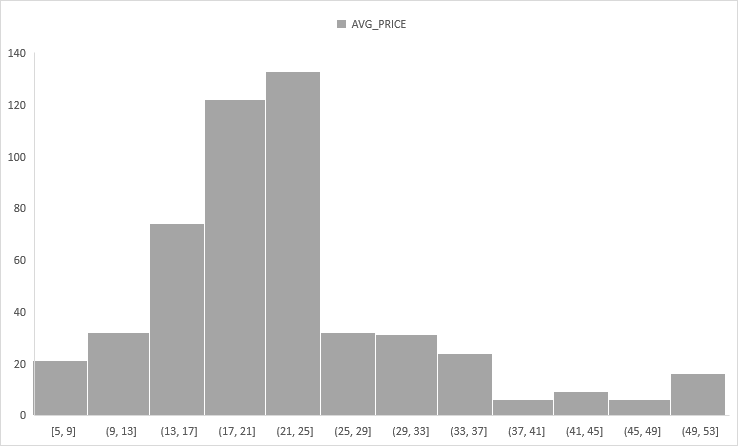




A detailed summary analysis of every column of the given dataset is provided here. The Mean, Median, Standard Deviation, range, minimum and maximum value of each aspect can be easily seen in the above tables. We can observe that an average household price is around 22000$, with age ranging from 2.9 to 100 years.

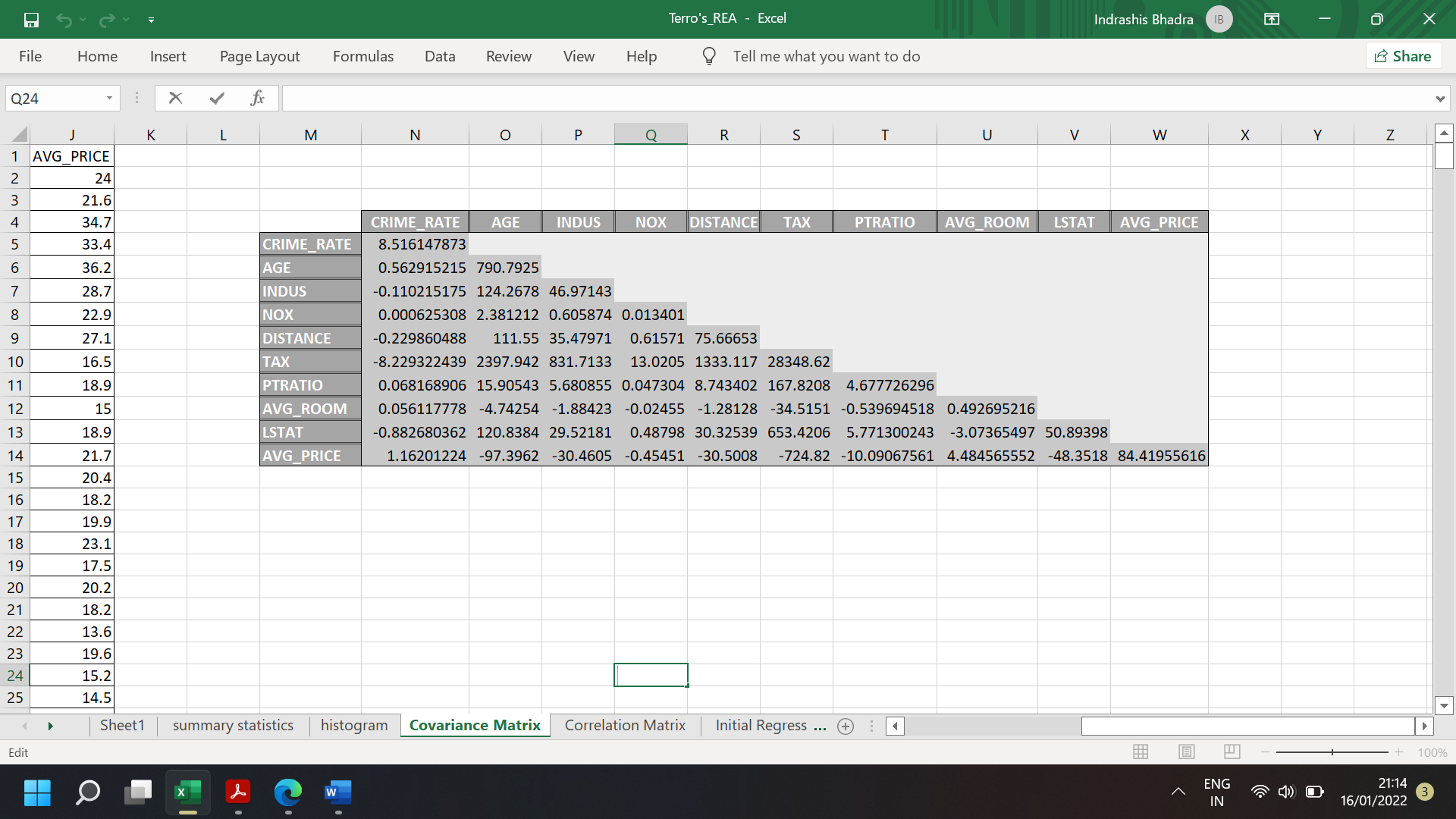
The count of 506 in every table shows that there is no excel dataset error present here. Thus, with a rough overview of the data, we can now proceed with further analysis.

1. **Histogram of the Avg\_Price Variable. What do you infer?**



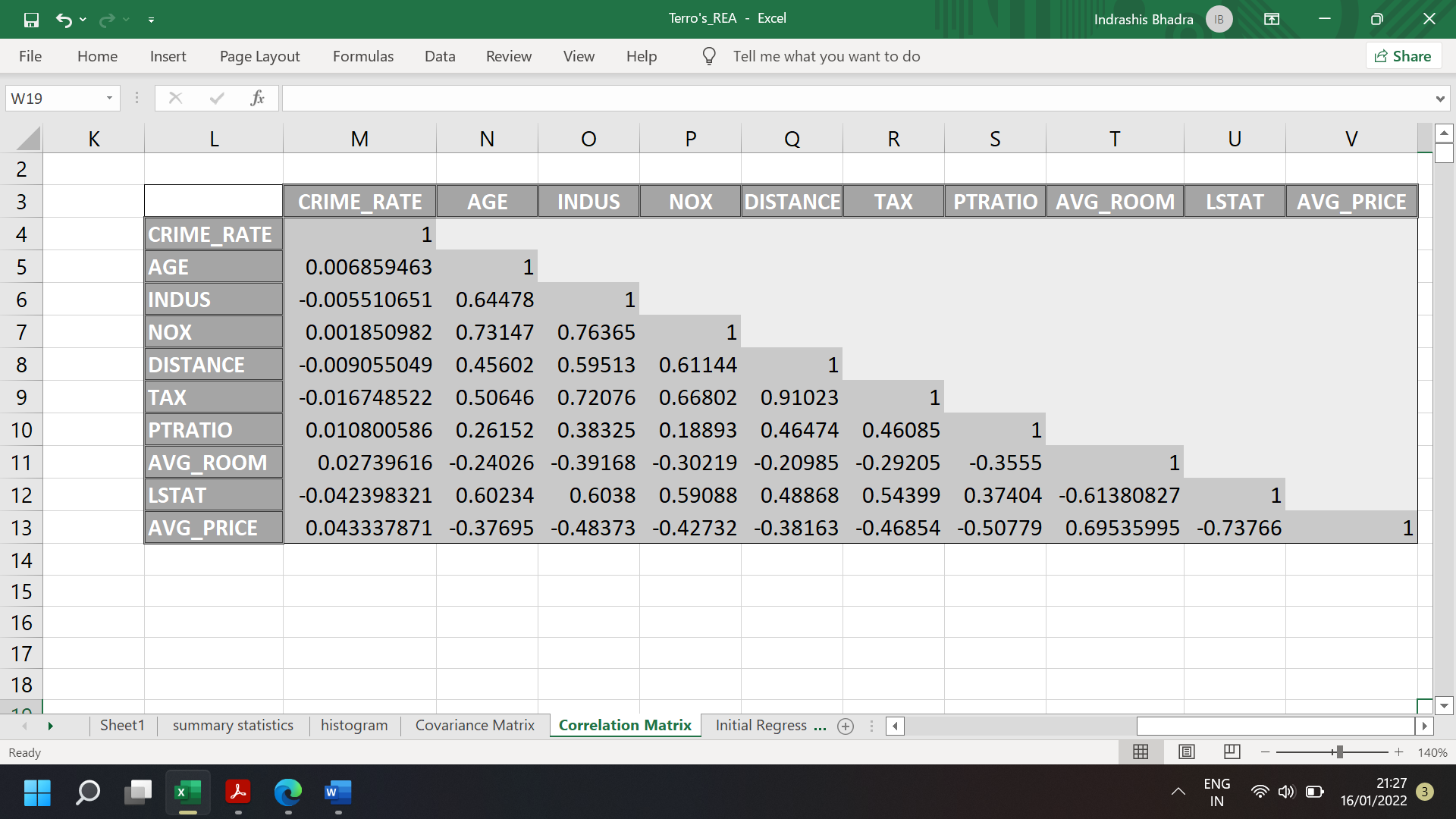
The most houses in this list have a price ranging from 21,000 to 25,000$. The number of houses in this range is 133. The next range is 17,000 to 21,000$ having 122 houses followed by 13,000 to 17,000$ range. A brief idea of average price range is inferred from this histogram.

1. **COVARIANCE MATRIX**



The above shown covariance matrix gives us an idea about symmetry and covariances of each pair of variables. Also, the covariance between same variables equals their variance. So, the diagonal shows the variance of each variable. We can interpret the direction of the multivariate data in the multidimensional space.

1. **CORRELATION MATRIX**

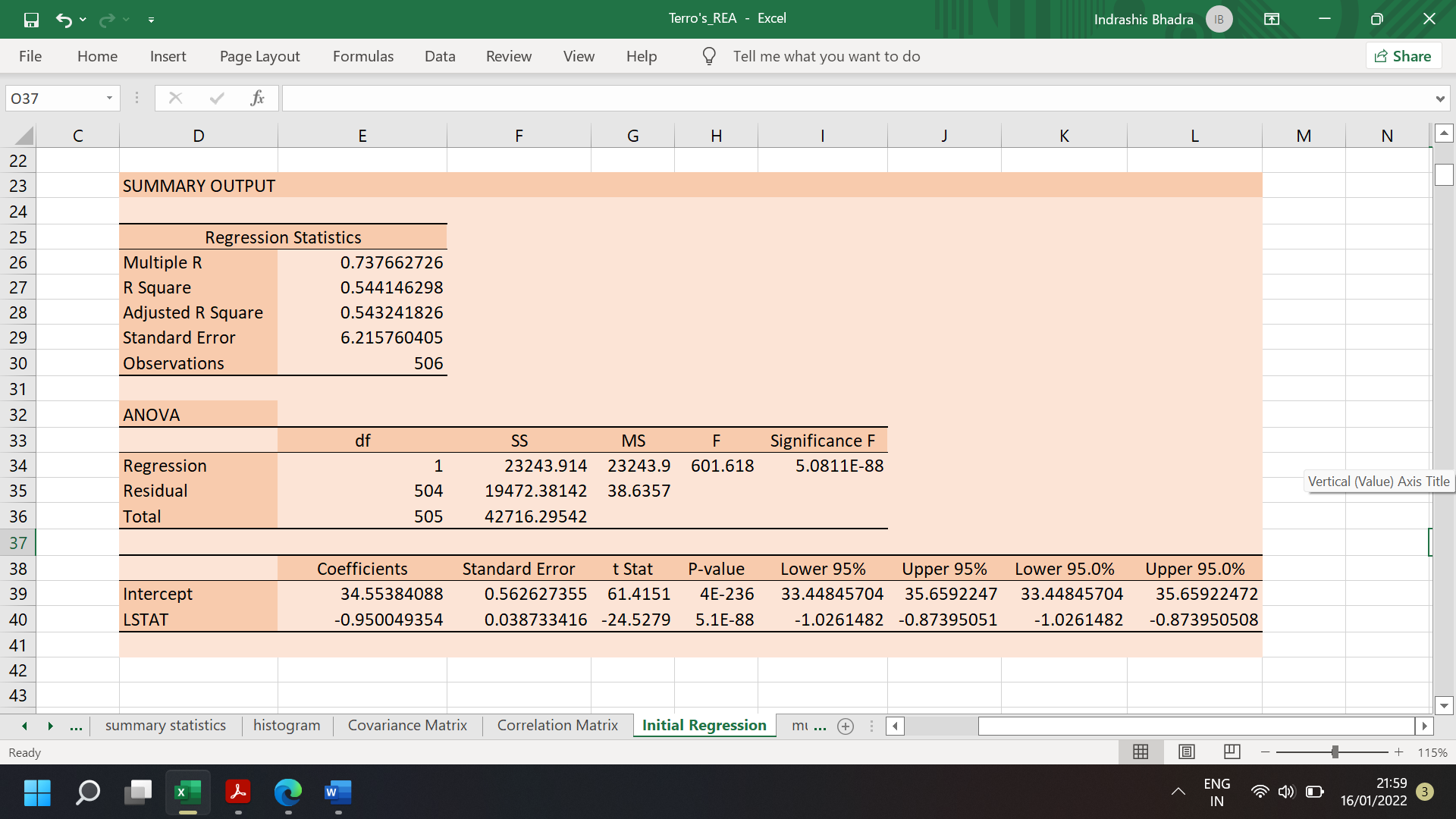


The above correlation matrix shows all the correlated pairs of the given dataset. The diagonal consists of 1 because the correlation of the same 2 variables is 1.

The 3 positively correlated pairs are (Tax, Distance) , (Nox, Indus) and (Nox ,Age) having correlation values 0.91023 , 0.76365 and 0.73147 respectively.

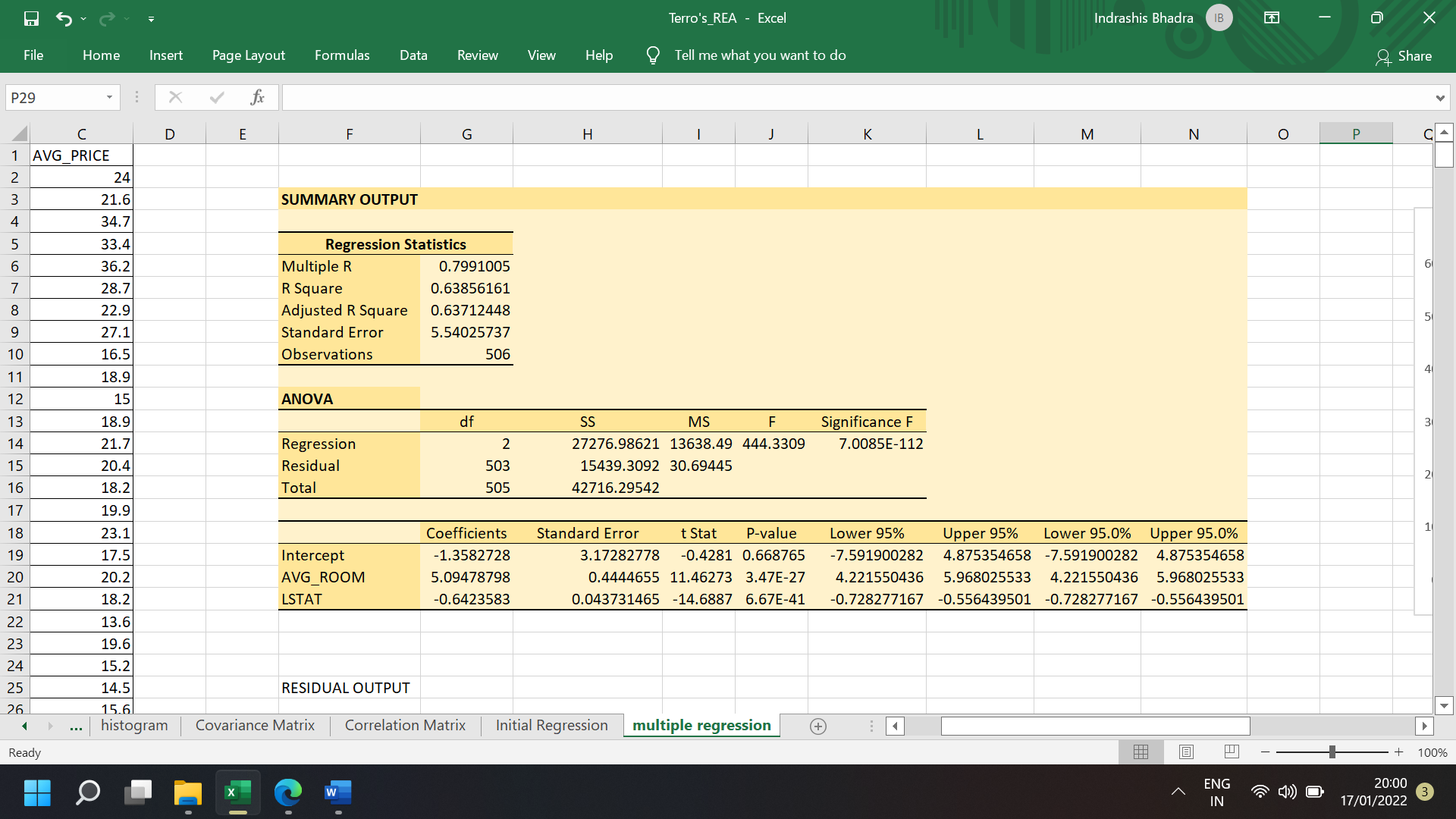
The 3 negatively correlated pairs are (Avg\_Price, LStat) , (LStat , Avg\_Room) and (Avg\_Price , PtRatio) having correlation values -0.73766 , -0.61380827 and -0.50779 respectively.

1. **INITIAL REGRESSION MODEL (WITH RESIDUAL PLOT)**

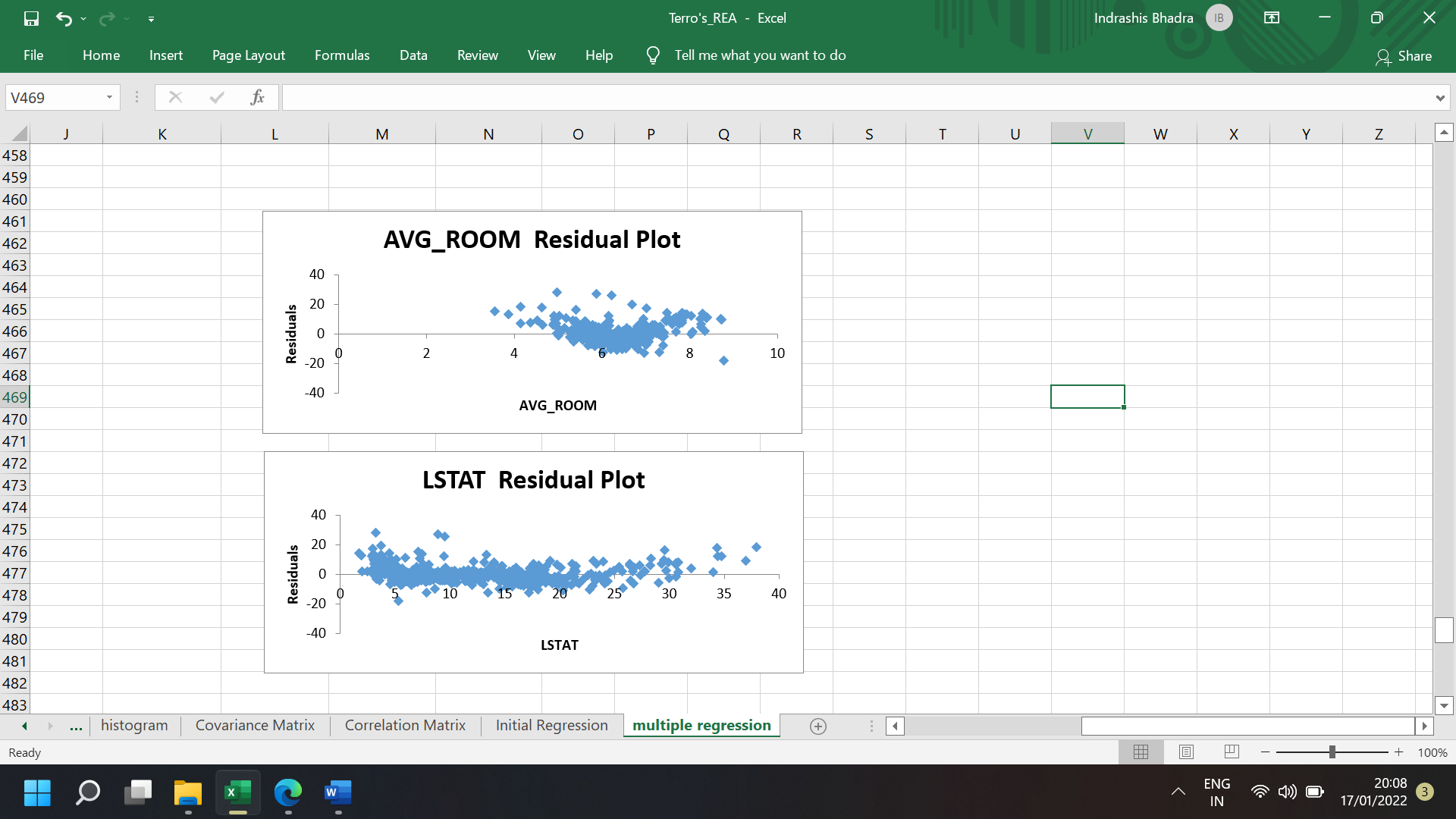
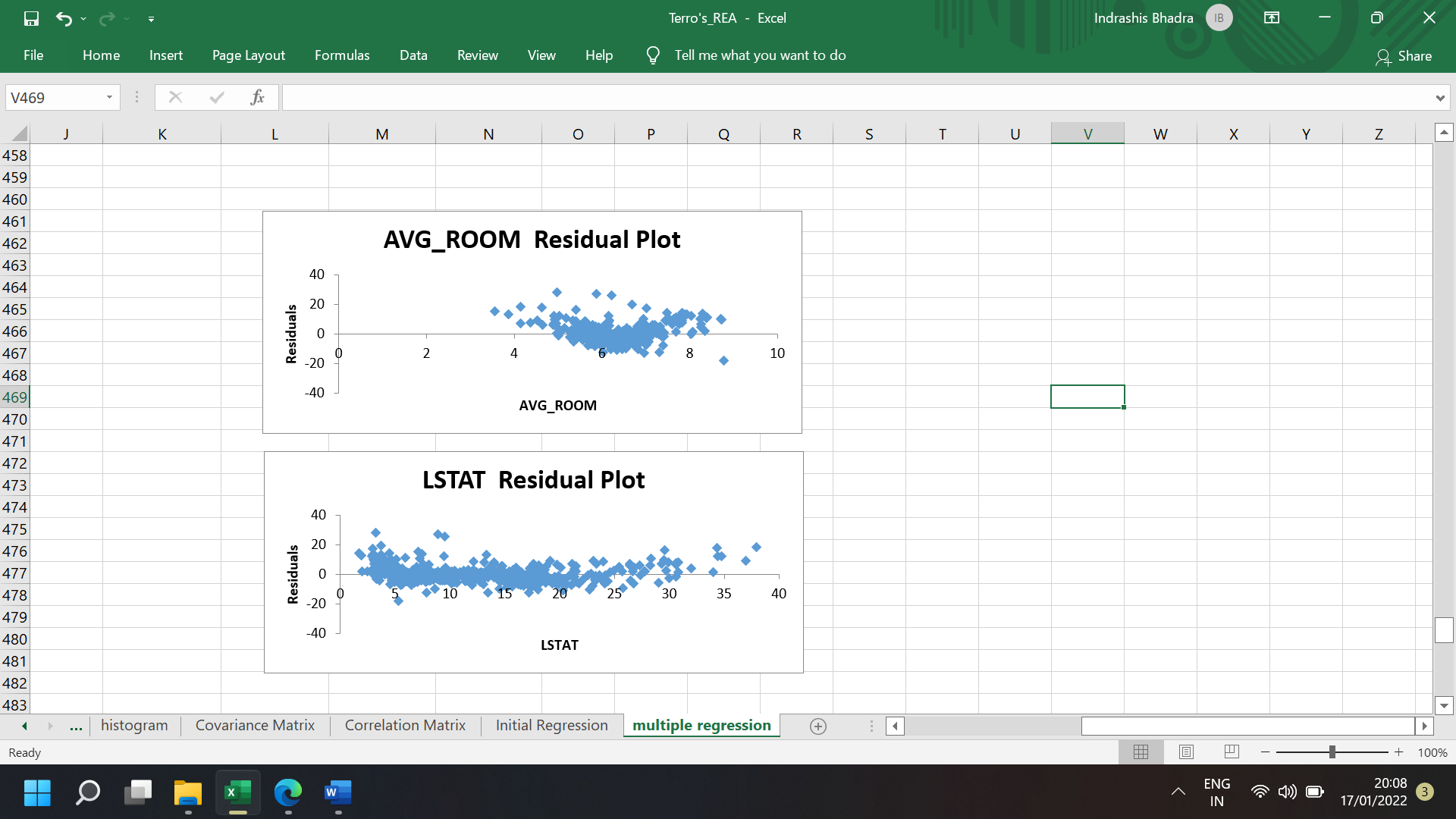


A detailed summary output along with Scatterplot, Regression Equation and Residual Plot is provided here. The coefficients give an estimate of least square values. The intercept value for coefficient is the constant of the regression equation. The p-value being much less than 0..05 shows that the LSTAT variable is significant.

1. **REGRESSION MODEL 2**



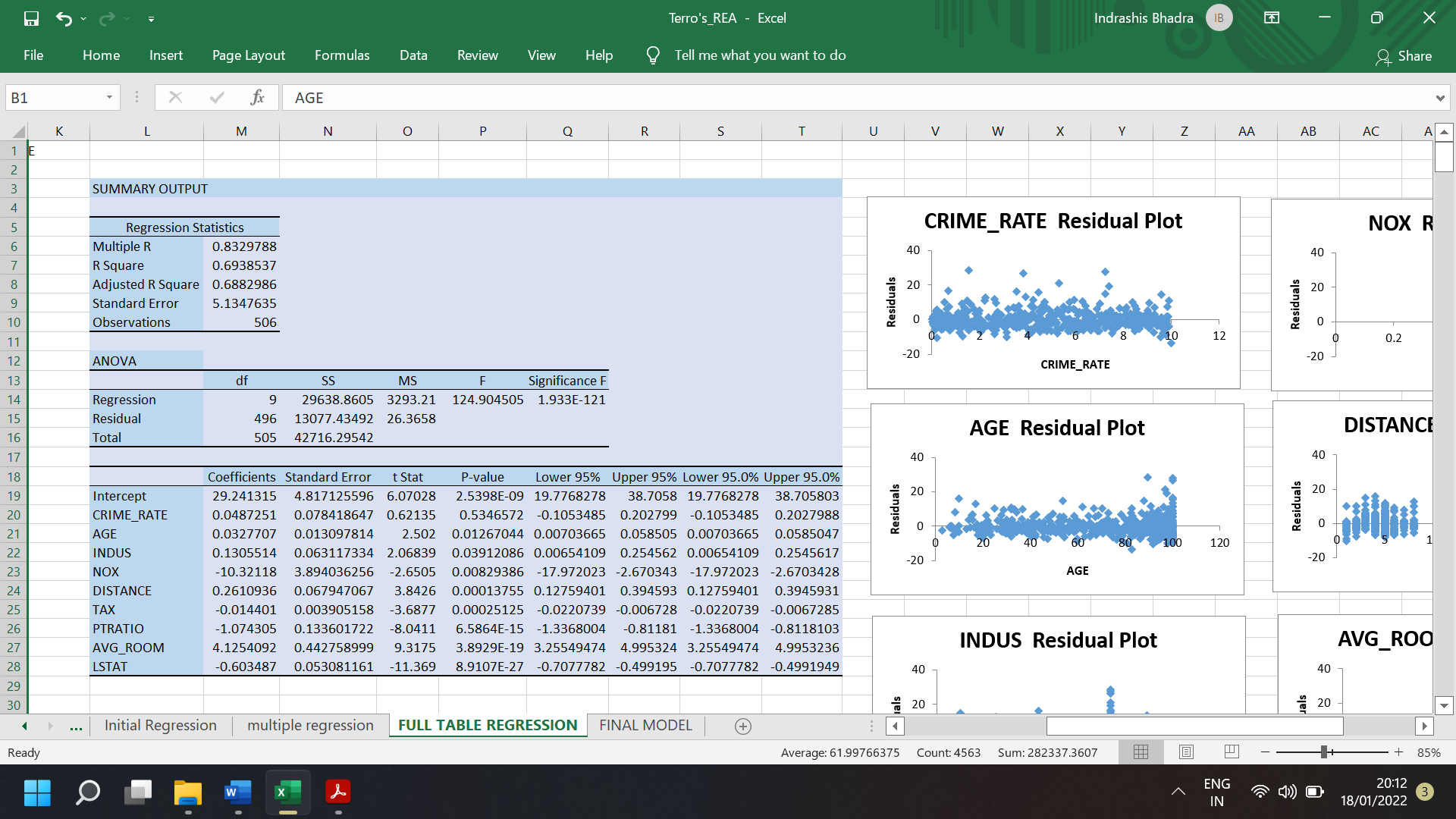
**Regression Equation: y= -0.64235833x1 +5.094787984x2 -1.35827281**



The Summary Output, Scatterplot, Regression Equation and Residual Plots are all provided above. It is a multiple regression model with AVG\_ROOM & LSTAT as independent variables. The Regression Equation is **y**= -0.64235833**x1** +5.094787984**x2** -1.35827281 . The value of AVERAGE\_PRICE based on the given information in question paper is **24.17462202 (**in terms of 1000 US. **$)**. A company that sells at an average of **30,000$** is clearly **overcharging**.

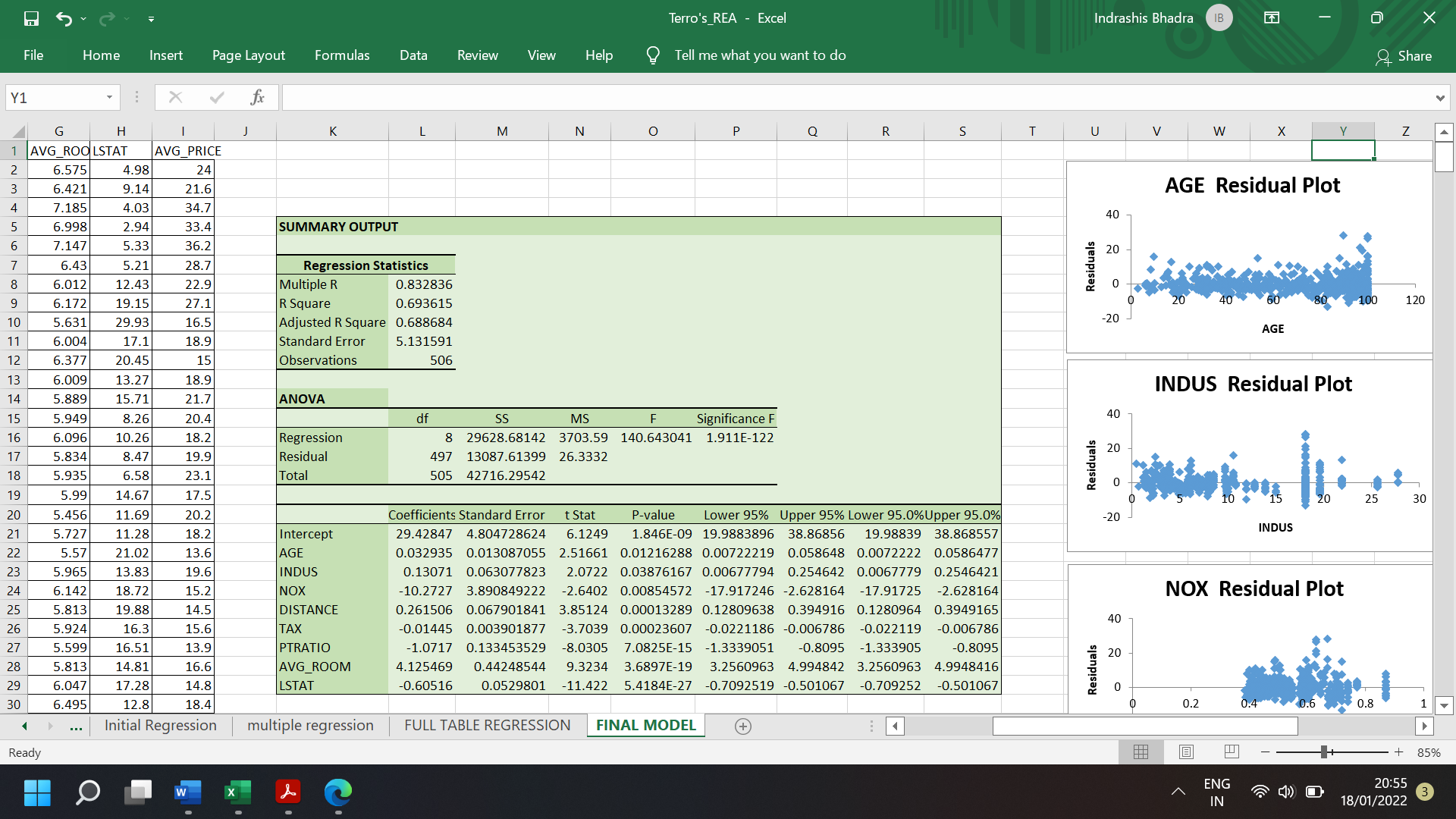
The performance of this model is no doubt better than the initial regression model as its adjusted R square value is **0.63712448** compared to **0.543242** of the other model. Higher the adjusted R-square value, better is the Regression Model.

1. **FULL TABLE REGRESSION MODEL**



The Summary Output and Residual Plots of given attributes is provided above. The **adjusted R-square** value is **0.688298646855749** which definitely shows that this model is better than all the previous models. The coefficients are the beta of the given variables. The significance of all the variables in comparison to the **AVG\_PRICE** can be measured from the **p-value**. Except for **CRIME\_RATE**, all the other variables have a **p-value** less than **0.05** which proves their significance.

1. **FINAL REGRESSION MODEL**



This is the Final Regression Model which consists of **only** the significant variables. The Summary Output and Residual Plots are provided above. The **adjusted R-square** value is **0.688683681872453**. Since we removed the non-significant **CRIME\_RATE** variable from this model , the adjusted R-square value **increased** by **0.00038504**. Thus, this is the most **successfu**l and **relevant** model that can be created from the given dataset.

The Regression Equation of this model is:

**y**= 29.42 + 0.03**x1** + 0.13**x2** – 10.27**x3** + 0.26**x4** - 0.01**x5** – 1.07**x6** + 4.12**x7** – 0.60**x8**

THE END