**Business Report**

Real Estate Data Analysis- Exploratory Data Analysis, Linear Regression

**Terro’s Real Estate Agency**

By

**RAJASEKAR S**

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

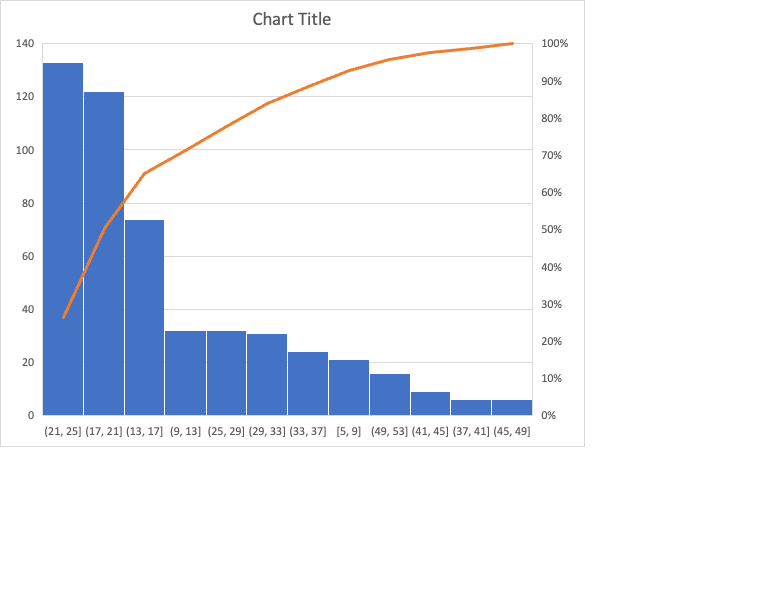
Ans: We use the Descriptive Statistics function inside the Data Analysis tool pack which can be found in the data ribbon for this task. The observation that we have arrived at from the table are that,

* Among the given data TAX has the highest Mean as**(408.23)**and NOX(Nitric oxides concentration) has the lowest Mean of **(0.55)**, thus sample means are widely spread around the population mean for NOX and the sample means are closely distributed around the population mean for TAX.
* Among the given data the Average Price has the Highest Skewness value of (1.10),and the lowest skewness value is obtained for PTRatio which has **(-0.11)**,thus the data spread out around the Average Price, and the data are clustered around the mean for PTRatio.
* TAX holds the highest **(168.537)** standard deviation while NOX holds the lowest **(0.116)**,thus the data are more spread out around the mean for TAX, and the data are clustered around the mean for NOX.
* TAX has the highest Maximum value as **(711)** and the Smallest minimum tax value which has **(0.48),**thus the more data spread out around the TAX, and the data are clustered around the mean for NOX.
* TAX has the highest range value as **(524)** and the smallest range value is NOX which has **(0.48),**thus the more data spread out around the Tax, and the data are Clustered around the mean for NOX.



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

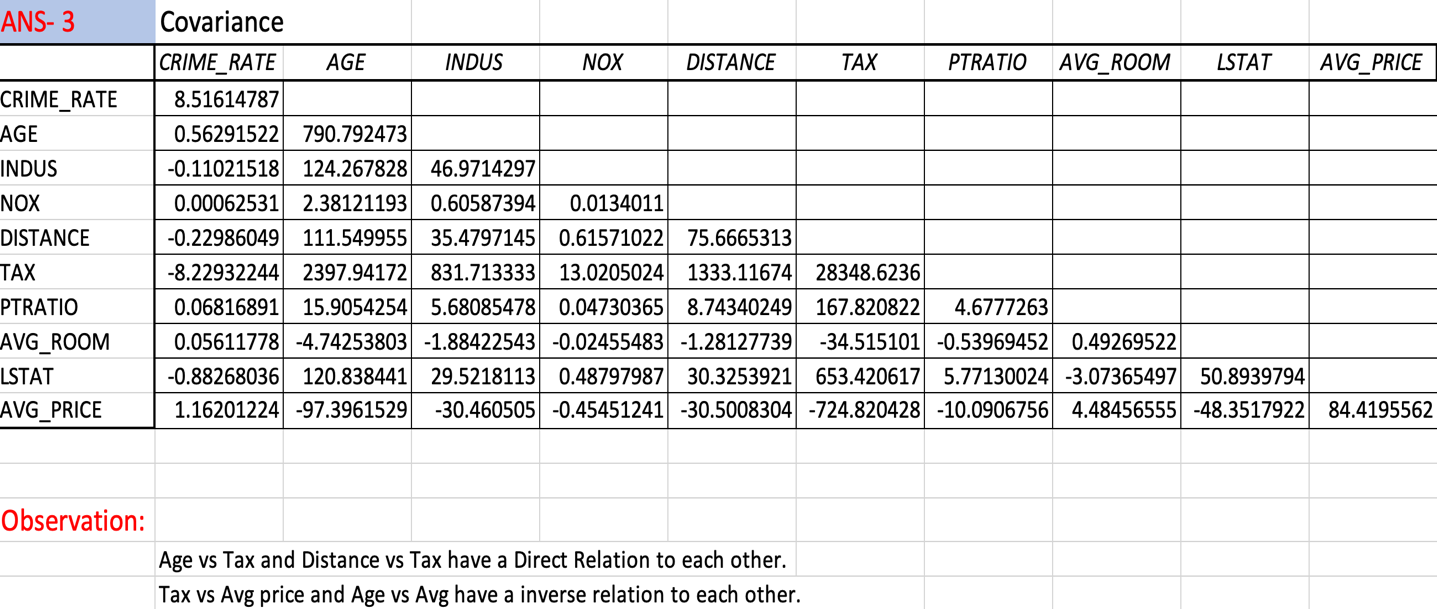
Ans: To find the Histogram for the Average Price Variable. We have to Copy the data which is given from the original data, after that select the insert from the menu bar. Select the Histogram for this Chart.



By observing the data in the histogram chart we can say that the data is more spread towards the left side of the histogram. The Data is positively skewed from the data to analysis which the plot are represent in detail manner.

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

Ans: For the Covariance Matrix. Select data which is in menu bar and within it click Data Analysis option. In that we have to select the Covariance option. When we get the covariance table we have to select the respective X and Y range and also enable the Label option to prior the headlines which is given from the original data. After we select all the features we get the covariance table.



After the Covariance table we are able to get Age vs Tax, Distance vs Tax, they have a Direct relation to Each other.

Also infer the Tax vs Average Price, Age vs Average Price have an Inverse Relation to Each other.

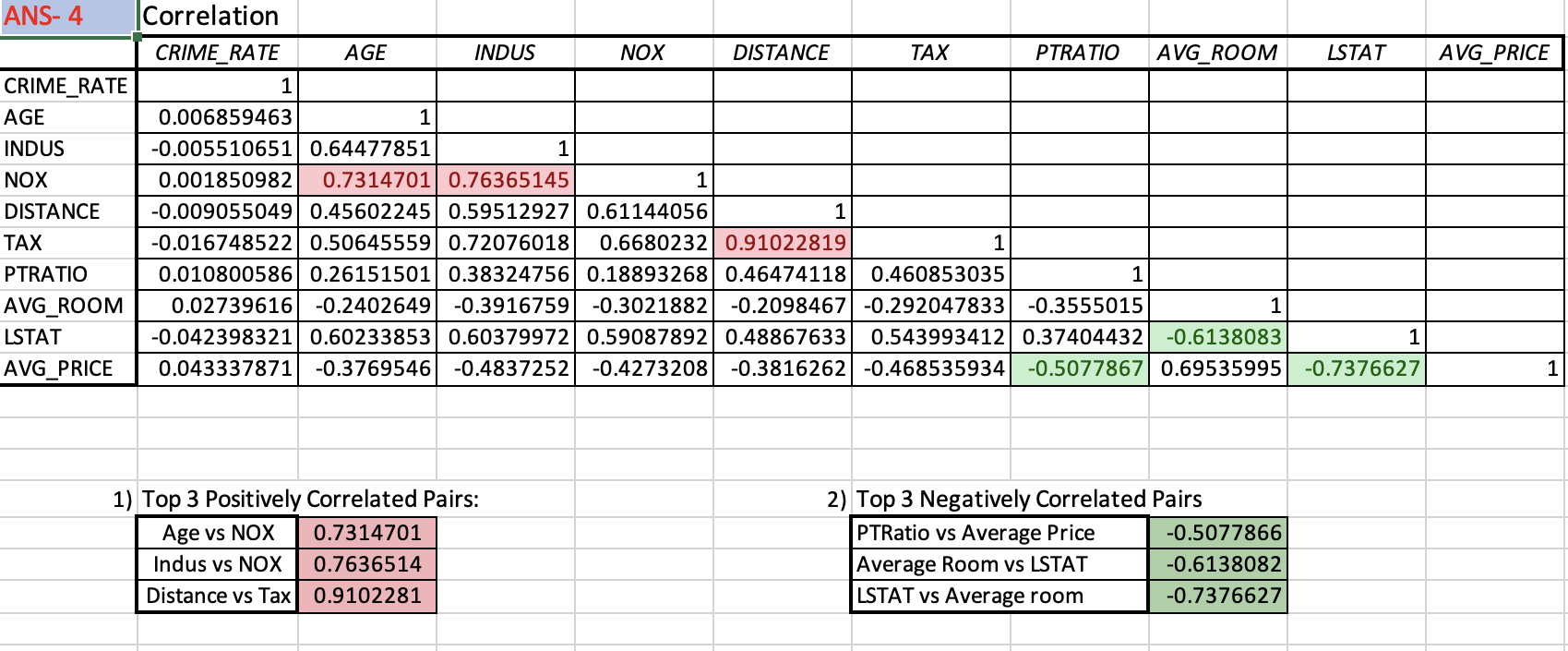
From the Covariance table we conclude that we able to find the relation between for each other from the given data.

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

**a) Which are the top 3 positively correlated pairs**

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

From the given data we want to find the correlation matrix of all the variable. To find that we want to select the Data from the menu bar option and within that we select Data Analayis. We will find further option, in that we have to select the correlation option. After that we have to select the respective range of X and Y to the table extend. Enable the Label option to Mark the prior Headlines to Correlate the table by the given data.



We want to select the as Slope direct which is releated to the the table eg; Crime rate vs Crime Rate. After that we want to select the Conditional Formatting which is in the home page of Excel. We want to select the Top to bottom and that click more option and select the range from the table. After that we have to click over whether the option are given to us. In that table we can select top and bottom. And also select top 3 positively Correlation and Top 3 Negative Correlation.

A) Top 3 Positively Correlated Pairs

Age vs NOX and we get the value as 0.7314701

Indus vs NIX we get the value as 0.7636514

Distance vs Tax 0.9102281

B) Top 3 Negatively Correlated Pairs

PTRatio vs Average Price we get the value as -0.5077866

Average room vs LSTAT we get the value as -0.6138082

LSTAT vs Average room we get the value as -0.7376627

**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?**

Regression is a methodology by which an output can be derived from the set of Population and Sample’s.

For finding the regression model we have to select the data from the menu bar and click the option as Data Analysis. From that search and click the regression process, while after that select the respective range of X and Y. Enable to Label for headline prior and select all the negative and residual plot and also independent variable for the Summary Output.

We will get the summary output with Coefficient, Significant, P value etc.

|  |  |
| --- | --- |
|  | *Coefficients* |
| Intercept | 34.55384088 |
| LSTAT | -0.950049354 |
|  |  |

From the Given data the inference:

The 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.**

a) Regression Equation

y=5.095x\_1-0.642x\_2-1.358

|  |  |  |
| --- | --- | --- |
| Intercept | 1 | -1.35827 |
| Average Room | 7 | 5.09478 |
| LSTAT | 20 | -0.64236 |

|  |  |
| --- | --- |
| Average Price | 21.45799 |

The Company Quoting a value of 30000USD but the calculated value of Avg price is lesser than the company quotation.

Hence we conclude that the company is OVERCHARGING.

Coefficient:

|  |  |
| --- | --- |
|  | *Coefficients* |
| Intercept | -1.358272812 |
| AVG\_ROOM | 5.094787984 |
| LSTAT | -0.642358334 |

b) R Square for Better value

|  |  |
| --- | --- |
| *Regression Statistics* |  |
| Multiple R | 0.73766273 |
| R Square | 0.5441463 |
| Adjusted R Square | 0.54324183 |
| Standard Error | 6.21576041 |
| Observations | 506 |

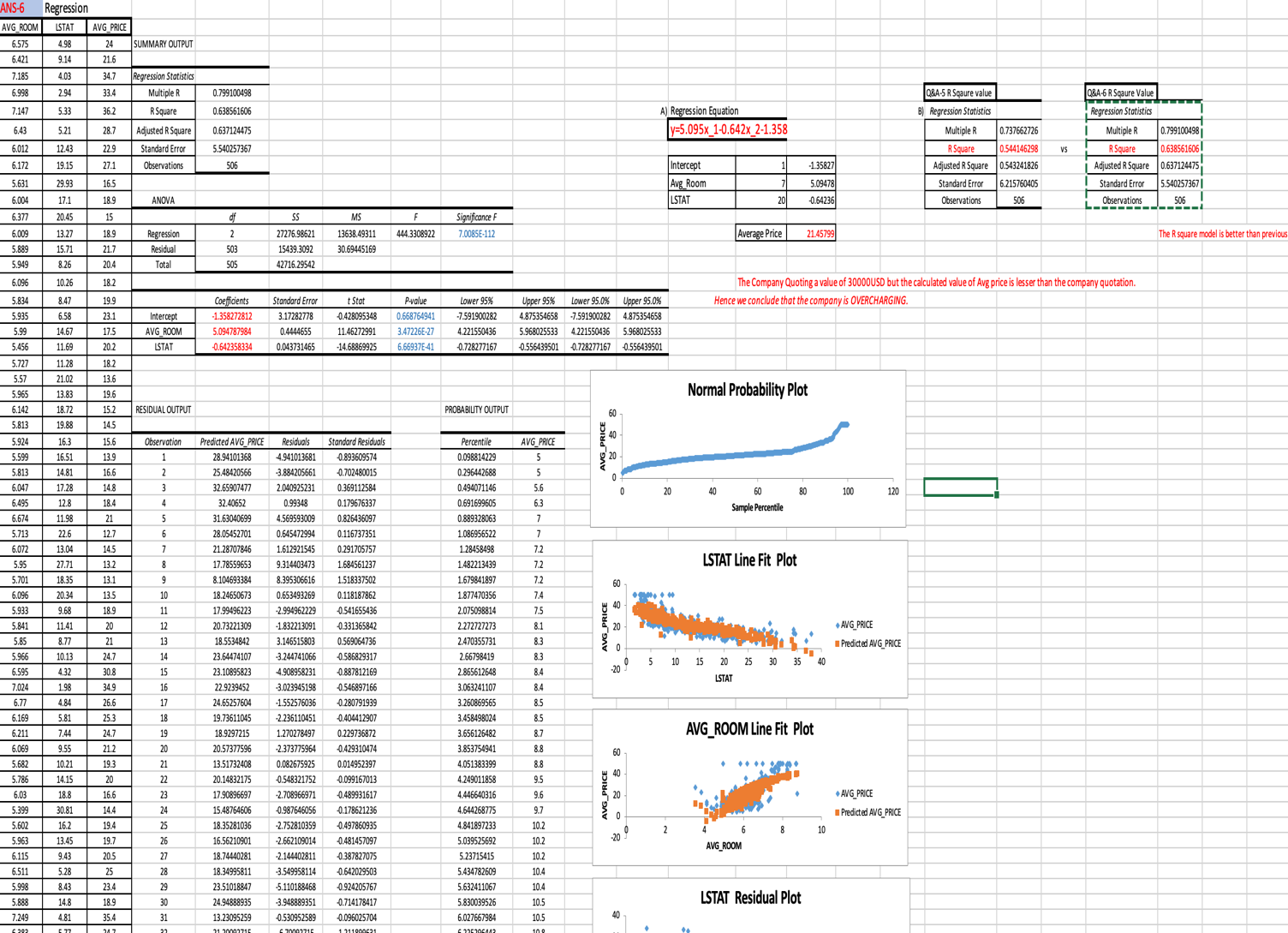
This Regression Statistics is from the Previous Question.

|  |  |
| --- | --- |
| *Regression Statistics* |  |
| Multiple R | 0.7991005 |
| R Square | 0.63856161 |
| Adjusted R Square | 0.63712448 |
| Standard Error | 5.54025737 |
| Observations | 506 |

This Regression Statistics is from the same question.

For comparing the value of R Square

0.63856161 is greater value than the 0.5441463 so we get better value from the previous question.



We will get the Normal Probability test, residual plot and also LSTAT line fit plot etc on the Summary Output of Regression Process.

**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: Comparing the R Square values this model is better than others and also comparing the coefficients and intercept

Crime rate and intercept Crime rate, age, Indus, distance, avg room, have the direct relation with avg price and the others have inverse relation.

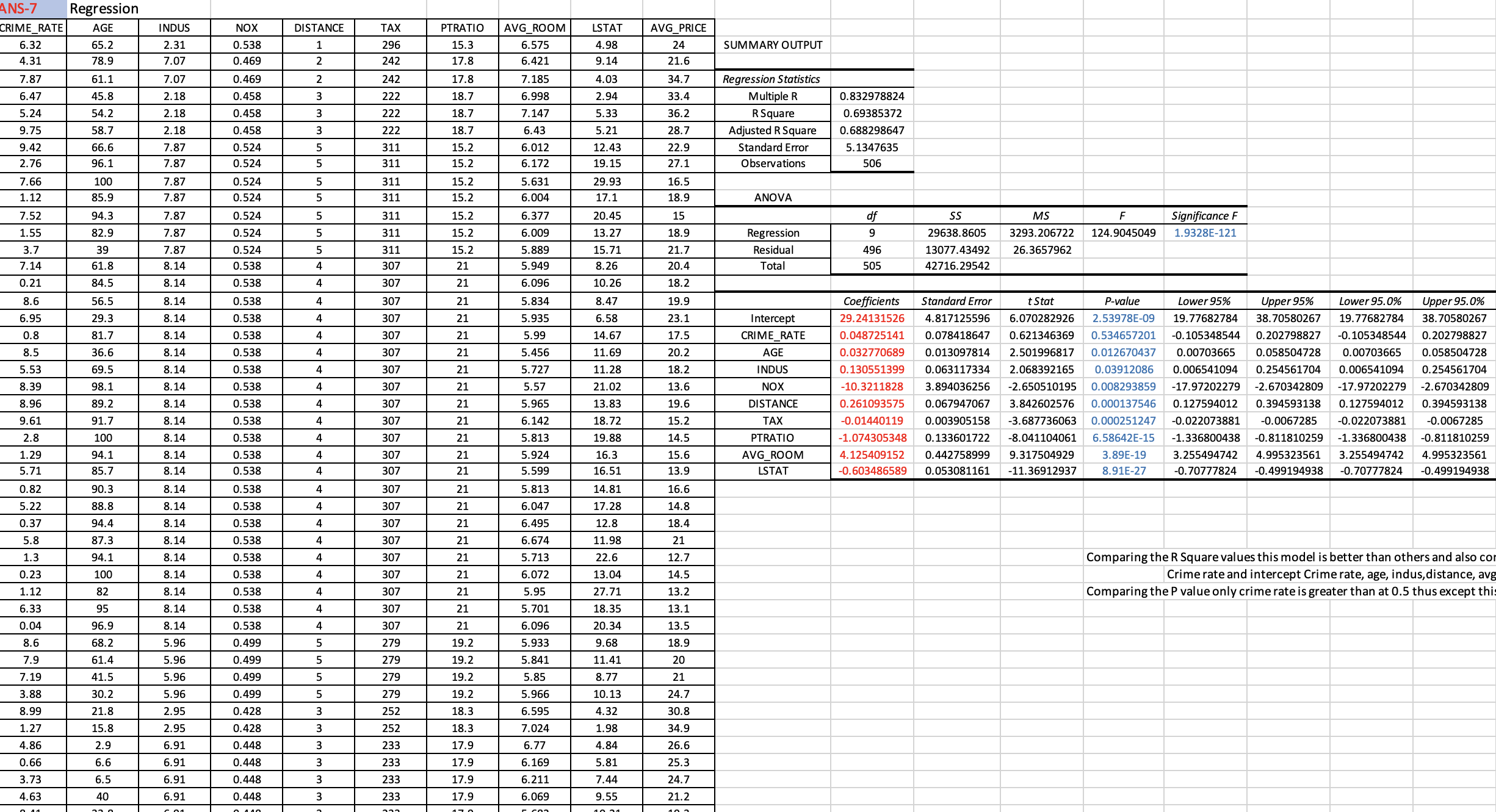
Comparing the P value only crime rate is greater than at 0.5 thus except this all the others are Significant.

|  |  |
| --- | --- |
| *Regression Statistics* |  |
| Multiple R | 0.832978824 |
| R Square | 0.69385372 |
| Adjusted R Square | 0.688298647 |
| Standard Error | 5.1347635 |
| Observations | 506 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 29.24131526 | 4.817125596 | 6.070282926 | 2.53978E-09 |
| CRIME\_RATE | 0.048725141 | 0.078418647 | 0.621346369 | 0.534657201 |
| AGE | 0.032770689 | 0.013097814 | 2.501996817 | 0.012670437 |
| INDUS | 0.130551399 | 0.063117334 | 2.068392165 | 0.03912086 |
| NOX | -10.3211828 | 3.894036256 | -2.650510195 | 0.008293859 |
| DISTANCE | 0.261093575 | 0.067947067 | 3.842602576 | 0.000137546 |
| TAX | -0.01440119 | 0.003905158 | -3.687736063 | 0.000251247 |
| PTRATIO | -1.074305348 | 0.133601722 | -8.041104061 | 6.58642E-15 |
| AVG\_ROOM | 4.125409152 | 0.442758999 | 9.317504929 | 3.89E-19 |
| LSTAT | -0.603486589 | 0.053081161 | -11.36912937 | 8.91E-27 |

We want to find the regression process from the given data so we get summary output and values of the plots by the data with respective X and Y.

R square value is labeled.



**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) This Regression model has Greater Accuracy than Others.

we want to find the *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*

|  |  |
| --- | --- |
| *Regression Statistics* |  |
| Multiple R | 0.832835773 |
| R Square | 0.693615426 |
| Adjusted R Square | 0.688683682 |
| Standard Error | 5.131591113 |
| Observations | 506 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 29.42847349 | 4.804728624 | 6.124898157 | 1.84597E-09 |
| AGE | 0.03293496 | 0.013087055 | 2.516605952 | 0.012162875 |
| INDUS | 0.130710007 | 0.063077823 | 2.072202264 | 0.038761669 |
| NOX | -10.27270508 | 3.890849222 | -2.640221837 | 0.008545718 |
| DISTANCE | 0.261506423 | 0.067901841 | 3.851242024 | 0.000132887 |
| TAX | -0.014452345 | 0.003901877 | -3.703946406 | 0.000236072 |
| PTRATIO | -1.071702473 | 0.133453529 | -8.030529271 | 7.08251E-15 |
| AVG\_ROOM | 4.125468959 | 0.44248544 | 9.323400461 | 3.68969E-19 |
| LSTAT | -0.605159282 | 0.0529801 | -11.42238841 | 5.41844E-27 |

b)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Q&A-7 R-Sqaure |  |  | Q&A-8 Rsqaure |  |
| *Regression Statistics* |  |  | *Regression Statistics* |  |
| Multiple R | 0.832978824 |  | Multiple R | 0.83283577 |
| R Square | 0.69385372 | vs | R Square | 0.69361543 |
| Adjusted R Square | 0.688298647 |  | Adjusted R Square | 0.68868368 |
| Standard Error | 5.1347635 |  | Standard Error | 5.13159111 |
| Observations | 506 |  | Observations | 506 |

**The R Square value is lesser than the previous value**

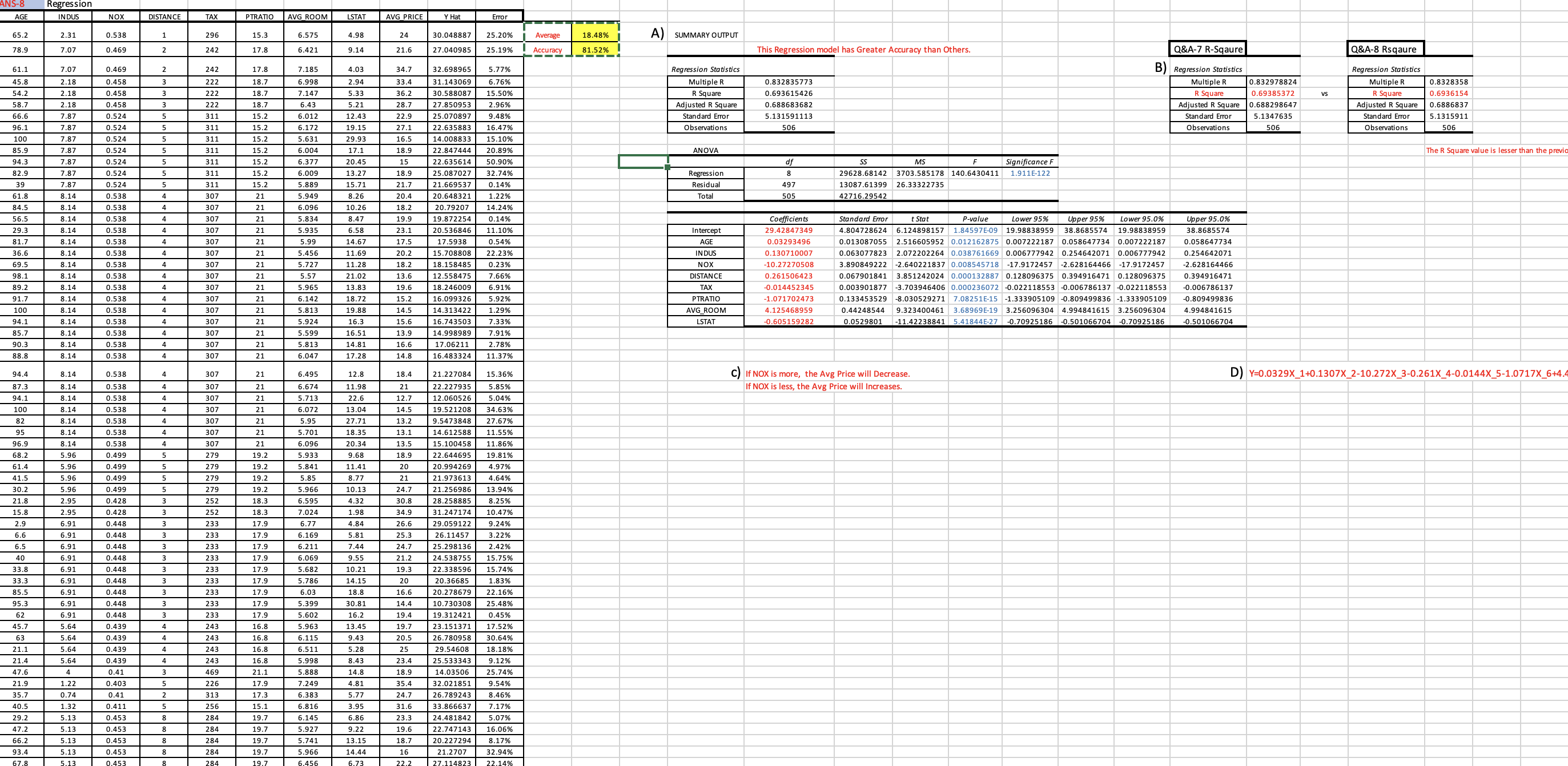
**c)** If NOX is more, the Avg Price will Decrease.

If NOX is less, the Avg Price will Increases.

**d)** **Y=0.0329X\_1+0.1307X\_2-10.272X\_3-0.261X\_4-0.0144X\_5-1.0717X\_6+4.414X\_7-0.0605X\_8+29.428**

Average and Accuracy

|  |  |
| --- | --- |
| Average | 18.48% |
| Accuracy | 81.52% |
|  |  |
|  |  |



Business Report

By

**Rajasekar S**