**A Business Report on Terro’s Real State Agency**

**Problem Statement**

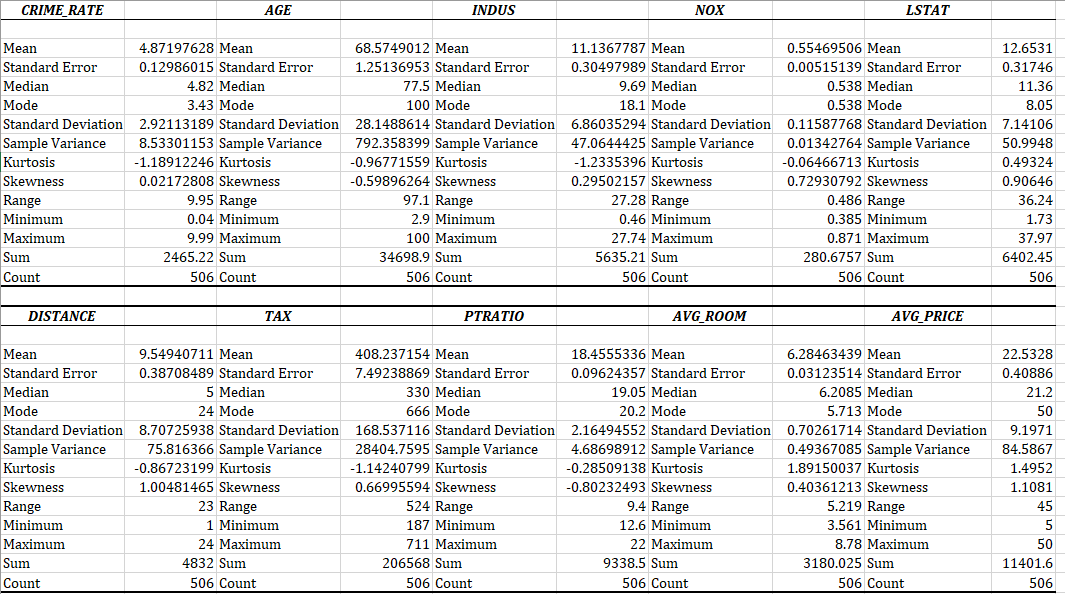
The problem under study is to find out the most relevant features for the pricing of a house. As an Auditor, there are various geographic features of a property like pollution level (NOX), crime rate, education facilities (pupil to teacher ratio), connectivity (distance from the highway), etc under study. This helps in determining the price of a property. The agency has provided a dataset of 506 houses in Boston.

**Objective**

* To analyze the magnitude of each variable under question to which it can affect the price of a house in a particular locality

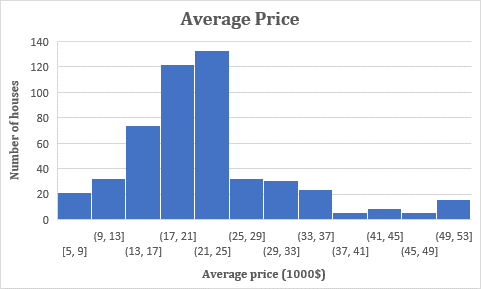
**Data analysis**

1. **Generate the summary statistics for each variable in the table. Write down your observation**

****

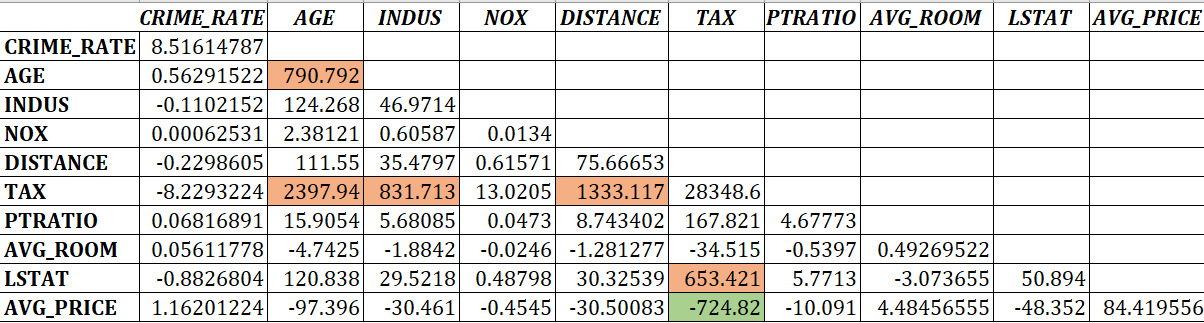
* The number of houses present in the dataset is 506
* Average crime rate in the town is around 4.8
* The house built in this locality is around 68 years old on average
* Most of the houses have 27% of land for non-retail business
* Most houses are 24 miles of distance from the highway
* The average tax rate is around 408$
* On average 12% of the population belongs to the lower status
* Maximum number of rooms is 8 in a house.

1. **Plot a histogram of the Average price variable. What do you infer?**



* In the Y-axis we have the number of houses and the X-axis shows the average price of houses at 1000$. From the graph, it can be inferred that it is positively skewed data. Here the average house prices are more concentrated between 21000USD and 25000USD.

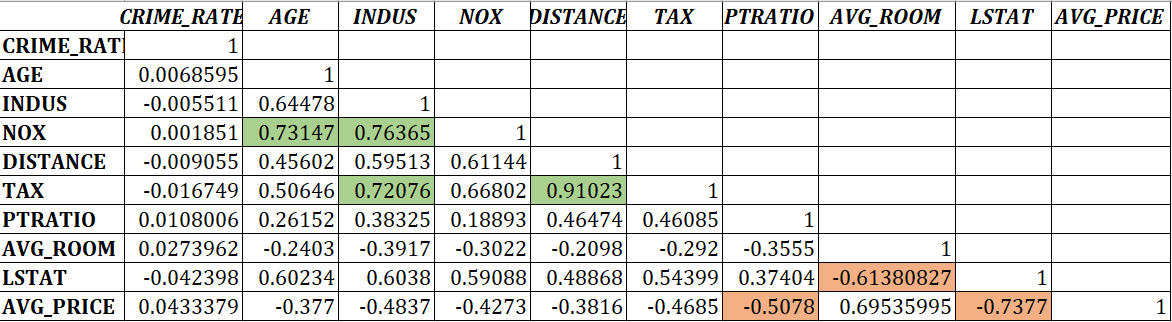
1. **Compute the covariance matrix. Share your observations.**

****

* The variance of the different variables is on the matrix diagonal. For example, the variance of crime rate is 8.5, Age is 790.7, etc.
* For average price and the rest of the independent variables, Positive covariance can be seen between average price and independent variables like crime rate and average room which implies that there is a positive relation between these variables that is, as the average number of rooms increases average price also increases
* As for variables like Age, Indus, Nox, Distance, Tax, Ptratio and Lstat have negative covariance with average price which means they are inversely proportional. Amongst this Tax variable has the highest negative covariance.
* With regards to independent variables, the given below variables have high positive covariance:

1. Tax and Age
2. Tax and Indus
3. Tax and Distance
4. Lstat and Tax

1. **Create a correlation matrix of all the variables**



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

**B) Which are the top 3 negatively correlated pairs**

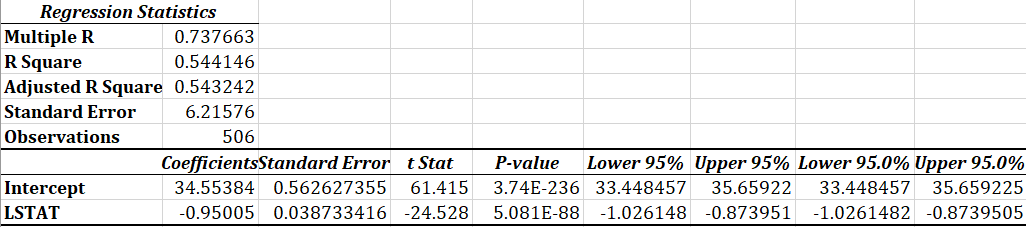
**b) 3 Negatively correlated variables**

* Average Price and Lstat -(-0.737)
* Lstat and Average room-(-0.61)
* Average price and Ptratio -(-0.503)

1. **3 Positively correlated variables**

* Tax and Distance - 0.91
* Nox and indus - 0.76
* Nox and Age – 0.73

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



* From the regression table, it can be inferred that the p-value of Lstat is significant and there is a significant relation between Average price and Lstat. Here the dependent variable is the Average price and the independent variable is Lstat.

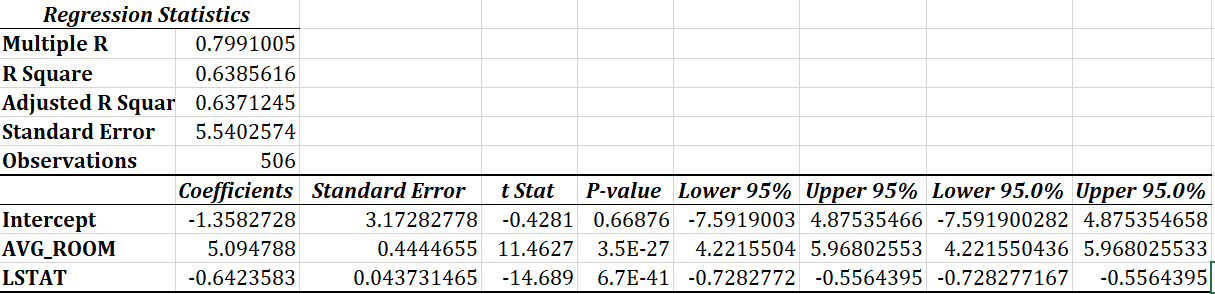
1. **What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?**

* The regression equation is Average price=-0.95\*Lstat + 34.55
* For a 1 unit increase in Lstat, there is a -0.95 dollars decrease in average price.
* The residual plot does not show patterns or trends so the model is pretty symmetrically distributed. If Lstat equals 0 then the expected value of the average price is 34.55.
* R-square is 0.54 which indicates that the Lstat in the regression model accounts for around 54% of the variation in the dependent variable. The remaining 46% is unexplained and perhaps attributable to other causes of variation.

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

Lstat is a significant variable for the model analysis as the p-value for the Lstat variable is less than 0.05. so this variable can be utilised for the model as it significantly impacts the average price.

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



Here the dependent variable is Average price and the independent variables are Average room and Lstat

**A) Write the Regression equation. If a new house in this locality has 7 rooms (on 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?**

* The regression equation is as follows:

Average Price=5.09\*Average room+ -0.642\*Lstat -1.35.

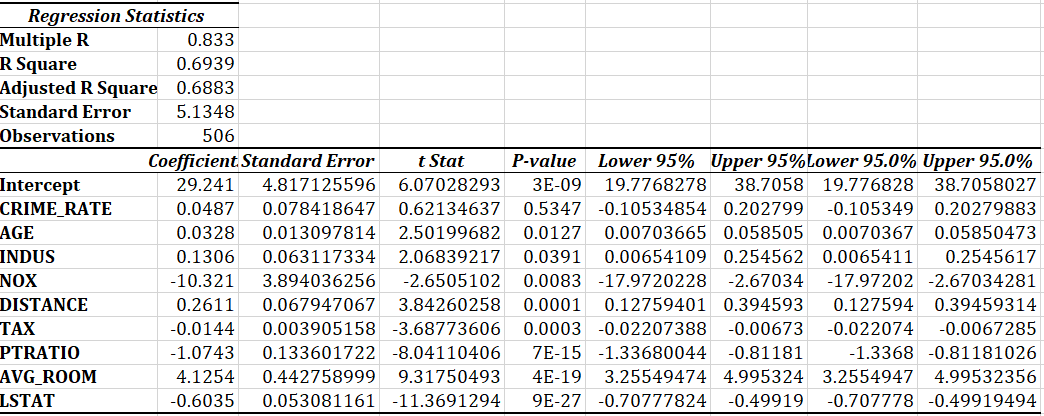
* Now by substituting the values in the given question the value of the average price is 21.44, which is 21,440 USD which is less than the value quoted by the company which is 30,000 USD. From this, we can infer that they are overcharging for the property. The calculation is as follows:

Average price =5.09\*7+(-0.642\*20)-1.35=21.44

**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.**

* This model is much better than the previous model in question 5 as the Adjusted r Square has increased from 0.54 to 0.63 with the inclusion of the average room. By looking at the p-value of the average room it is less than 0.05 which states that it is a significant variable.

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.**

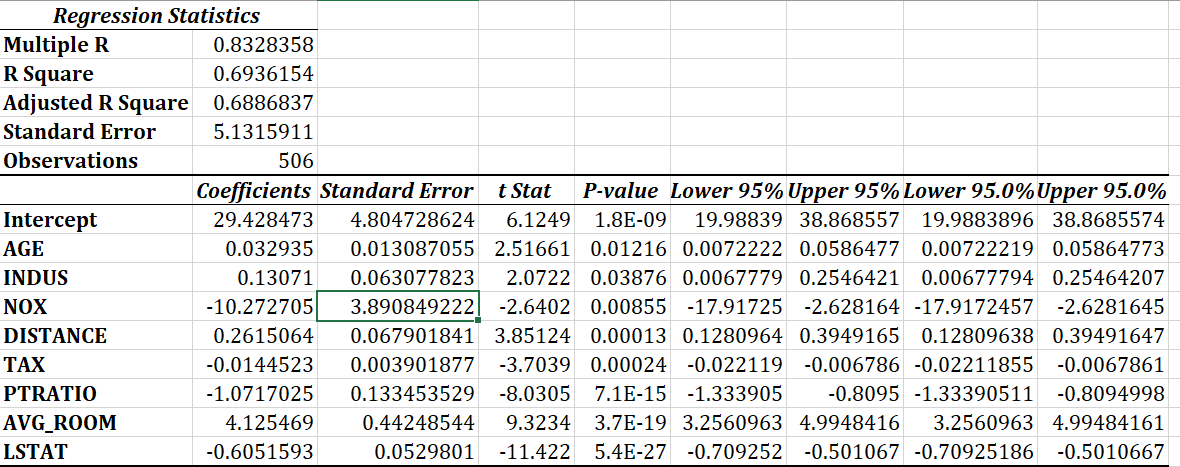


* From the adjusted R-square it can be inferred that it is a moderately good model. Compared to the previous models the addition of variables has improved the adjusted R-square from 0.63 to 0.68, so we can consider that this is the better model than the previous ones.
* As for the coefficients, Crime rate, Age, Indus, Distance and Average room are directly related to average price while Nox, Tax, Ptratio and Lstat are inversely related.
* The regression equation is

Average price = 0.048\*Crime rate + 0.03\*Age + 0.13\*Indus +(-10.32\*Nox) + 0.26\*Distance +(-0.014\*Tax) + (-1.07\*Ptratio) + 4.12\*Average room + (-0.6\*Lstat) + 29.241

* If any all the values of the explanatory variables are zero then the expected value of the average price is 29.241 dollars
* Variables which are highly significant for the model are Pratio, Average room and Lstat as its p-value are less than 0.05
* Variables like Tax, Distance, Nox, Indus and Age are moderately significant
* Crime rate is the only insignificant value which has a p value more than 0.05 which should be excluded from the model.

1. **Pick out only the significant variables from the previous question. Make another instance of the Regression model using only the significant variables**

****

1. **Interpret the output of this model.**

* Concerning R- square which is 0.68, this model is moderately good. 68% of the variation average variable can be attributed to the variation in other independent variables.
* The P-value of all independent variables is significant as all of them are below 0.05
* By looking at the coefficients, when variables like Age, Indus, Distance, and Average room increase by 1 unit then the average price also increases
* When other variables like Nox, Tax, Ptratio, and Lstat increase by 1 unit then the average price decreases.
* If any all the values of the explanatory variables are zero then the expected value of the average price is 29.42 dollars

1. **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?**

* If we compare the Adjusted r square, this model is performing much better than the previous model. In the previous model, the adjusted r square was 0.6883 and it has improved to 0.6886.

1. **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?**

* If Nox is more in the region, then the average price decreases by -10.27 so we can say that they are inversely related.

1. **Write the regression equation from this model**

The regression equation is as follows:

* Average price = 0.03\*Age + 0.13\*Indus + (-10.27\*Nox) + 0.26\*Distance +

(-0.01\*Tax)+ (-1.07\*Ptratio) + 4.12\*Average room + (-0.605\*Lstat)