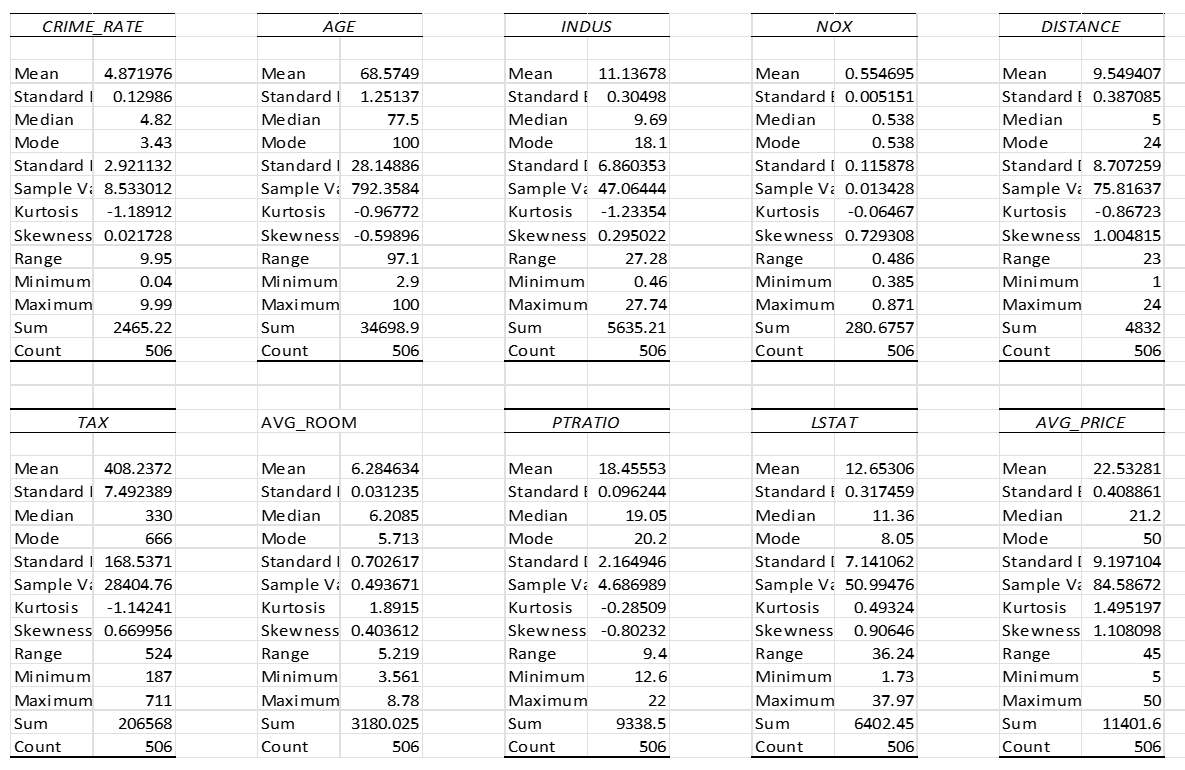
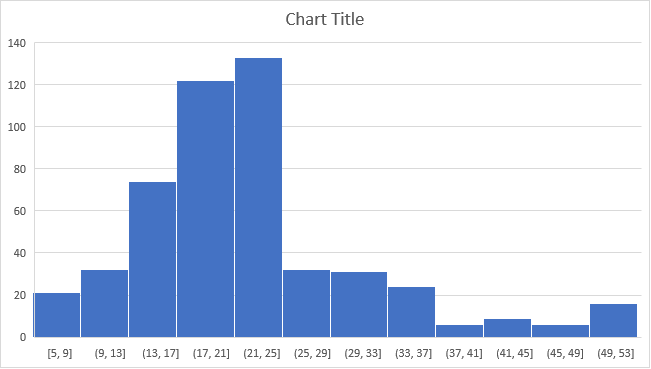
1. **The first step to any project is understanding the data. So, for this step, generate the summary statistics for each of the variables. What do you observe?**



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



We can see from the above histogram that the data is right skewed, and thus the mean is greater than the median, indicating that the data is not symmetric. Then the mean may not provide a good estimate for the centre of the data and represent where most of the data falls. In this case, we should consider using the median rather than the mean to determine the centre of the data.

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

By looking atthe above matrix, we can observe that the cells highlighted in green colour have positive co variance, it means that both the variables covariance are either positive or negative, they return a positive covariance.

The cells which are highlighted with red colour have Negative covariance, it means either one of variables have negative or positive covariance, if one variable has positive covariance and the other one has negative covariance or vice versa it give a negative covariance.

1.  **Create a correlation matrix of all the variables as shown in the Videos and various case studies. State top 3 positively correlated pairs and top 3 negatively correlated pairs.**

From the above Correlation matrix, the top 3 positively correlated variables are

TAX vs DISTANCE = 0.910

NOX vs INDUS = 0.763

NOX vs AGE = 0.731

and the top 3 negatively correlated variables are

AVG\_PRICE vs LSTAT = -0.738

LSTATvsAVG\_ROOM = -0.614

AVG\_PRICE vs PTRATIO = -0.508

1. **Build an initial regression model with AVG\_PRICE as the y or the Dependent variable and LSTAT as the Independent variable. Generate the residual plot too. 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?**



1. The R-squared value is 0.5441, indicating that the LSTAT can explain 54.41% of the variance in the average price.

the Adjusted R-squared is 0.5432.

Standard Error is 6.2157, the observed values fall an average of 6.2157 units from the regression line. The p-value is 5.0811E-88, which is less than the commonly accepted significance level of 0.05. This indicates that the regression model is statistically significant. i.e., the model fits the data better than the model with no predictor variables. When all the independent variables are zero. the intercept is 34.55384088. i.e., When LSTAT is zero, the Avg\_price Will be 34.55384088 (in $1000), the LSTAT coefficient is -0.950049354, which indicates that the Avg\_Price tends to decrease as the STAT variable increases.

1. The Adjusted R-squared is 0.5432 so the LSTAT can explain 54.30% of variance in average price and p-value is 5.0811E-88, which is less than the commonly accepted significance level of 0.05. This indicates that the regression model is statistically significant.
2. **Build another instance of the Regression model but this time include LSTAT and AVG\_ROOM together as independent variables and AVG\_PRICE as the 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. Explain**



* 1. Regression Equation

Y= b0+b1x1+b2x2+e

Avg\_Price = -1.358+5.095\*(Avg\_Rooms)+ (-0.642)\*(L-STAT)

= -1.358+5.095\*(7)+ (-0.642)\*(20)

= 21.45807639

The value of Avg\_price when the locality has 7 rooms (on an average) and value of 20 for L-STAT, then the Avg\_Price of house will be $ 21458.07.

We can observe that the business is overcharging for this location when we compare the above-average pricing with the $30000 charged by the company.

* 1. Both the R-squared and the Adjusted R-squared values are 0.637. demonstrating that 63.7% of the variance in the average price can be explained by this model. Compared to the prior model, which we constructed using L-STAT, the adjusted R-squared value is 0.5432, and the R-squared value is 0.5441. demonstrating that 54.32% of the variance in the average price can be explained by the LSTAT. By comparing these two modules, we can see that the model we created using L-STAT and Avg Price has a higher Adjusted R-Squared value than the module we created before. Consequently, we can say that the model we built using L-STAT and Avg Price is better than The earlier model we built.

1. **Now, build a Regression model with all variables. AVG\_PRICE shall be the Dependent Variable. Interpret the output in terms of adjusted R-square, coefficient and Intercept values, Significance of variables with respect to AVG\_Price. Explain.**



The R-squared value is 0.6938 and the Adjusted R-squared is 0.6882indicating that the variables can explain 68.82% of the variance in the average price.

Standard Error is 5.134 the observed values fall an average of 5.134units from the regression line. The p-value is 1.9328E-121, which is less than the commonly accepted significance level of 0.05. This indicates that the regression model is statistically significant. i.e., the model fits the data better than the model with no predictor variables. When all the independent variables are zero. the intercept is 29.24131526. i.e., When all the variables value is zero, the Avg\_price Will be 29.2413 (in $1000), the coefficient of Crime\_Rate 0.04872514, Age 0.032770689, Indus 0.130551399, and Avg\_Room 4.1254 are positively related which indicates that the Avg\_Price tends to increase as the value of these variables increases.

the coefficient of L-STAT is -0.950049354, NOX -10.3211828, TAX -0.01440119, PTRATIO -1.074305348 are inversely related which indicates that the Avg\_Price tends to decrease as the value of these variables increases.

1. **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. (HINT: 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) 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 value of NOX is more in a locality in this town? d. Write the regression equation from this model**.
2. The R-squared value is 0.6936 and the Adjusted R-squared is 0.6886 indicating that the variables can explain 68.86% of the variance in the average price.

Standard Error is 5.131 the observed values fall an average of 5.131units from the regression line. The p-value is 1.911E-122, which is less than the commonly accepted significance level of 0.05. This indicates that the regression model is statistically significant.

1. Adjusted R Squared Value 0.6886 ( this Model )

Adjusted R Squared Value 0.6882 ( previous model)

By comparing both model we can observe that the

1. 

1. As the value of NOX increases in a locality, the AVG\_PRICE decreases, or the quoting price reduces significantly.
2. Regression equation

AVG\_PRICE = 29.43 + (AVG\_ROOM\*4.13) + (DISTANCE\*0.26) + (INDUS\*0.13) + (AGE\*0.03) + (TAX\*-0.01) + (LSTAT\*-0.61) + (PTRATIO\*-1.07) + (NOX\*-10.27)