Question 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?





The dataset contains the following parameters such as crime rate, age, Indus, Nox, Distance, Tax, PT Ratio, Avg\_Room, LSTAT, Avg\_Price among the houses in Boston.

CRIME\_RATE (percentage of crime rate):

* Minimum: 0.04 %
* Maximum: 9.99 %
* Average: 4.87 %

AGE:

* Minimum: 2.9
* Maximum: 100
* Average: 68.57

INDUS:

* Minimum: 0.46 % of the town’s area per acre is used for non- retail business purpose.
* Maximum: 27.74% of the town’s area per acre is used for non- retail business purpose.
* Average: 11.13 % of the town’s area per acre is used for non- retail business purpose.

NOX (Nitric Oxide concentration):

* Minimum: 0.385 parts per million.
* Maximum: 0.871 parts per million.
* Average: 0.55 parts per million.

DISTANCE (distance from the highway to the house):

* Minimum: 1 mile
* Maximum 24 miles
* Average: 9.55 miles

TAX (Tax for a house per $10000)

* Minimum: $187
* Maximum: $711
* Average: $408.24

PTRATIO (People to Teacher ratio):

* Minimum: 12.6
* Maximum: 22
* Average: 18.45

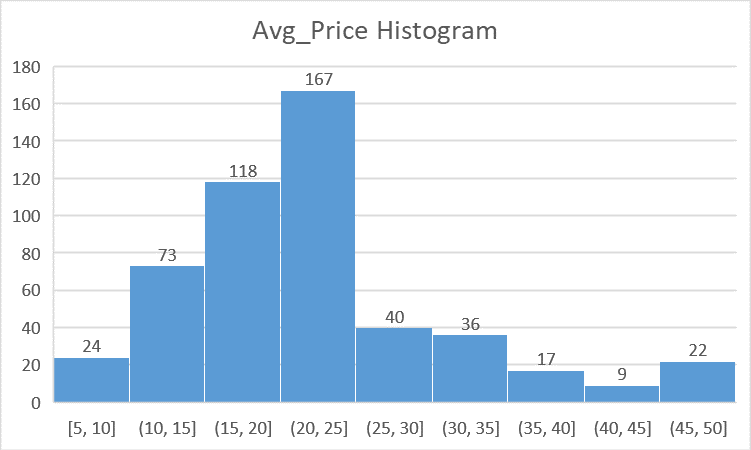
AVG\_ROOM (average number rooms)

* Minimum: 3.561
* Maximum: 8.78
* Average: 6.28

LSTAT (percentage of lower status of population):

* Minimum: 1.73 %
* Maximum: 37.97 %
* Average: 12.65 %

Question 2. Plot the histogram of the Avg\_Price Variable. What do you infer?



The above Histogram plot depicts the distribution of Avg\_Price data of the given dataset.

There were a greater number of Avg\_Price found between the 20 to 25 bins (i.e., 167 houses).

Question 3. Compute the covariance matrix. Share your observations

Question 4. 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.

Top 3 Positively corelated pairs are 0.910, 0.763 and 0.731.

Top 3 Negatively corelated pairs are -0.737, -0.614 and -0.508.

Question 5. Build an initial regression model with AVG\_PRICE as the y or the Dependent variable and LSTAT variable as the Independent Variable. Generate the residual plot too. (8 marks)

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 Equation:** y = mx + b

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

PRICE = (-0.95005 \* LSTAT) + 34.55384

1. **R Square:** We can see that 54% of the variance in the Price is defined by the LSTAT

**Co-efficient Value**: Our co-efficient value is negative (i.e., -0.95005), so if the LSTAT value increases the PRICE decreases.

**Intercept**: The intercept from the regression is 34.55384. When the LSTAT becomes

**Residual Plot:** The LSTAT residual plot has uniform variance except a few points are deviating from them.

1. The regression has been done with 5% significance and the result that we obtained i.e.; p-value is 5.08E-88(less than 5%). Hence, the LSTAT variable is significant for the analysis.

Question 6. Build another instance of the Regression model but this time including 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.

**Regression Equation:**

y = m1x1 + m2x2 + b

PRICE = (m1\*LSTAT) + (m2\*ROOM) + b

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

1. Regression Equation: PRICE = (m1\*LSTAT) + (m2\*ROOM) + b

If the new house has 7 more rooms on average and has 20 as L-STAT value, then price of the new house is,

PRICE = (m1\*LSTAT) + (m2\*ROOM) + b

= (m1\*20) + (m2\*7) + b

= (-0.64236\*20) + (5.094788\*7) + (-1.35827)

= -12.8472 + 35.663516 -1.35827

PRICE = 21.45805 K USD

The average price of the house when it has 7 rooms and has 20 as L-STAT value will be 21458.05 USD.

If the company quotes 30000 USD for this house, then the company is overcharging for the house (i.e., the company quotes 8541.95 USD additionally)



|  |  |  |
| --- | --- | --- |
| ***Regression Statistics*** | X Var - (L-STAT) | X Var – (L-STAT & Avg\_Room) |
| R Square | 0.54414629758648 | 0.638561606260341 |
| Adjusted R Square | 0.543241825954707 | 0.637124475470123 |

Both the R Square and Adjusted R Square of L-STAT and Avg\_Room is greater than that of the L-STAT. So, *the performance of this model is better than that of the previous model*.

Question 7. 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.

**Adjusted R-Square:** The adjusted R-square that we get is 0.688 or 68.8%, which means 68.8% of the Avg\_Price depends on the other factors. This is greater and much reliable than that of the previous models.

**Coefficient and Intercept:**



Regression Equation 🡪

Intercept = 29.2413

PRICE = m1x1 + m2x2 +m3x3 +m4x4 +m5x5 + m6x6 + m7x7 + m8x8 +m9x9 +b

PRICE = (-0.6034\*LSTAT) + (4.1254\*Avg\_Room) + (-1.0743\*PTRatio) + (-0.0144\* Tax) + (0.2610\*Distance) + (-10.3211\*NOX) + (0.1305\*INDUS) + (0.03277 \* Age) + (0.0487\*Crime\_Rate) + 29.2413

From the above table and equation, we could see that when the variables Lstat, PTRatio, Tax and NOX increases the price decreases. So, the real estate agency should consider these variables and maintain these as low as possible.

**Significance of variables with respect to Avg\_Price:**



It is obvious that except Crime\_Rate (which is 53.46%), all the other variables are statistically significant (i.e., lower than 5% level of significance).

So, except Crime\_Rate all the other variables can be considered for the Avg\_Price calculation.

Question 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. (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 the value of NOX is more in a locality in this town?

d. Write the regression equation from this model.

* From the below mentioned table, all the variables are statistically significant (i.e., below 5% level of significance)



* Adjusted R-Square:

The above images show us that after neglecting the Crime\_Rate, we get similar Adjusted R-Square value in both the previous model and current model

* NOX variable:



🡪If the NOX value increases further, the price of the house becomes cheaper.

* Regression Equation:



PRICE = m1x1 + m2x2 +m3x3 +m4x4 +m5x5 + m6x6 + m7x7 + m8x8 + b

PRICE = (-0.6051\*LSTAT) + (4.1254\*Avg\_Room) + (-1.0717\*PTRatio) + (-0.0144\* Tax) + (0.2615\*Distance) + (-10.2727\*NOX) + (0.1307\*INDUS) + (0.03293 \* Age) + 29.4285