**Terro’s Real Estate Agency**

**EDA Report**

1. Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation. (5 marks)

The Summary statistics can be found in the excel sheet Q1 and the observations are as follows:

1.CRIME\_RATE:

Mean = 4.871976285

Median = 4.82

Skewness = 0.021728079

Standard Deviation = 2.921131892

Kurtosis = -1.189122464

Since mean is almost equal to median I would say that the distribution would be more like a symmetrical distribution and we can also see that we have little positive skewness which means there are slightly more towns which have lesser CRIME\_RATE compared to the rest of the distribution but it is not significant since the skewness is very low and we would have little tail on the right side of the distribution which is above the average side. Since the Standard deviation is very large compared to mean so the distribution would be flat, and we can confirm that by seeing the kurtosis which is negative.

2.AVG\_ROOM:

Mean = 6.284634387

Median = 6.2085

Standard Deviation = 0.702617143

Skewness = 0.403612133

Kurtosis = 1.891500366

Since the mean and the median is almost equal to, I would say that the distribution is symmetrical, and we can see that the standard deviation Is very low compared to mean so that more data will be around mean and we will be having a distribution which will be very sharp around the mean and we can confirm that by seeing kurtosis which is positive and also as high as 1.89+. The standard deviation also says that there are more house around 6 number of rooms which is the mean. We can also see that we have little positive skewness which means there are more houses with less number of rooms compared to rest of the distribution, and we can also see that the distribution is having a little tail in the right of the distribution which is above the average side.

1. Plot a histogram of the Avg\_Price variable. What do you infer? (5 marks)

A picture containing text, screenshot, plot, line

Description automatically generated

The histogram can also be found in the excel sheet Q2 and the insights are as follows:

We can see that there is more data clustered below the average and it says that there are less number of costly houses which are outliers and majority of the distribution is having the normal house price range between 13k and 25k and there are few houses which are very costly and those are the ones which are outliers and they make the distribution asymmetrical. The costlier houses might be overpriced, or it could be due to various other factors like number of rooms so the major factors could be the facilities.

1. Compute the covariance matrix. Share your observations. (5 marks)

The covariance matrix can be found in the excel sheet Q3 and the observations are as follows:

Positive Covariance:

When two variable move in the same direction then there is positive covariance.

Example: CRIME\_RATE and AGE, INDUS and AGE

Negative Covariance:

When two variable move in opposite direction then there is negative covariance.

Example: INDUS and CRIME\_RATE, TAX and CRIME\_RATE

4) Create a correlation matrix of all the variables (Use Data analysis tool pack). (5 marks)

a) Which are the top 3 positively correlated pairs

1.TAX and DISTANCE

2.INDUS and NOX

3.AGE and NOX

b) Which are the top 3 negatively correlated pairs.

1.LSTAT and AVG\_PRICE

2.AVG\_ROOM and LSTAT

3.PRATIO and AVG\_PRICE

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

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

The regression model is in the excel sheet Q5 and the observations are as follows:

Adjusted R^2 value is very low which is 0.543241826 so that LSTAT is only 54% account for the variability in the AVG\_PRICE so that the remaining 46% is explained by many other variables. And if there is one unit increase in LSTAT then there will be -0.950049354 decrease in the AVG\_PRICE which means when the frequency of the lower socio-economic status people increases then the AVG\_PRICE will decrease and the intercept is 34.55384088 which says that if there is a situation where there is LSTAT is 0 then the AVG\_PRICE will be of 34.55384088 so it is constant for all the house prices and with a variability of -0.950049354\*LSTAT. And there is no pattern in the residual plot.

b) Is LSTAT variable significant for the analysis based on your model?

Adjusted R^2 value is very low which is 0.543241826 so that LSTAT is only 54% account for the variability in the AVG\_PRICE so that the remaining 46% is explained by many other variables.

So, I would say that LSTAT variable is significant and the model is insignificant because the LSTAT variable alone contributed for 54% of variability in AVG\_PRICE but we need to consider more variable for the model to predict the AVG\_PRICE correctly.

6) Build a new Regression model including LSTAT and AVG\_ROOM together as independent variables and AVG\_PRICE as dependent variable. (6 marks)

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?

Pred AVG\_PRICE = -1.358272812 + 5.094787984\*AVG\_ROOM-0.642358334\*LSTAT

Pred AVG\_PRICE = -1.358272812 + 5.094787984\*7-0.642358334\*20

Pred AVG\_PRICE = $21.458K (or) $21458

In the question they have mentioned that the company is charging $30K for the same house but according to our model it is $21.458K so that the company is overcharging.

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.

Yes, the performance of this model is better than the previous model because the adjusted R^2 value has increase from 0.543241826 to 0.637124475.

So, the previous model is only account for the 54% variability in AVG\_PRICE but the new model is account for 63% variability in the AVG\_PRICE so, the new model has improved.

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. (8 marks)

The adjusted R^2 have increased from 0.637124475 to 0.688298647 which is slight improvement because the previous model is account for 63% variability in AVG\_PRICE but the new model is account for 68% variability.

If there is unit change (increase) in CRIME\_RATE then there will be increase of 0.048725141 In AVG\_PRICE and if there is unit change(increase) in AGE then there will be increase of 0.032770689 In AVG\_PRICE and if there is unit change (increase) in INDUS then there will be increase of 0.130551399 in AVG\_PRICE and if there is unit change (increase) in NOX then there will be decrease of -10.3211828 and if there is unit change (increase) in DISTANCE then there will be increase of 0.261093575 in AVG\_PRICE and if there is unit change (increase) in TAX then there will be decrease of -0.01440119 in the AVG\_PRICE and if there is unit change (increase) in PTRATIO then there will be decrease of -1.074305348 in the AVG\_PRICE and if there is unit change (increase) in AVG\_ROOM then there will be increase of 4.125409152 in the AVG\_PRICE and if there is unit change (increase) in LSTAT then there will be decrease of -0.603486589 in the AVG\_PRICE.

And if there is no change in any of the variables then there will be increase of 29.24131526 in the AVG\_PRICE so it will be constant.

And Predicted AVG\_PRICE = 0.048725141\*CRIME\_RATE+0.032770689\*AGE+0.130551399\*INDUS-10.3211828\*NOX+0.261093575\*DISTANCE-0.01440119\*TAX-1.074305348\*PTRATIO+4.125409152\*AVG\_ROOM-0.603486589\*LSTAT + 29.24131526.

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: (8 marks)

a) Interpret the output of this model.

There is very negligible increase in adjusted R^2 as compared to the previous model and this model is account for the 68% variability in the AVG\_PRICE. And the p-value have improved significantly.

If there is unit change(increase) in AGE then there will be increase of 0.03293496 In AVG\_PRICE and if there is unit change (increase) in INDUS then there will be increase of 0.130710007 in AVG\_PRICE and if there is unit change (increase) in NOX then there will be decrease of -10.27270508 and if there is unit change (increase) in DISTANCE then there will be increase of 0.261506423 in AVG\_PRICE and if there is unit change (increase) in TAX then there will be decrease of -0.014452345 in the AVG\_PRICE and if there is unit change (increase) in PTRATIO then there will be decrease of -1.071702473 in the AVG\_PRICE and if there is unit change (increase) in AVG\_ROOM then there will be increase of 4.125468959 in the AVG\_PRICE and if there is unit change (increase) in LSTAT then there will be decrease of -0.605159282 in the AVG\_PRICE.

And if there is no change in any of the variables then there will be increase of 29.42847349 in the AVG\_PRICE so it will be constant.

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?

There is very negligible increase in adjusted R^2 as compared to the previous model and this model is account for the 68% variability in the AVG\_PRICE since there is very little increase in the adjusted R^2 we could say both the model are equal but considering the increase in adjusted R^2 and the decrease in p-value I could say that this model is better than the previous model.

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?

Sorted the values of coefficients in ascending order:

A picture containing text, screenshot, font, number

Description automatically generated

If NOX is more in a locality in this town, then the AVG\_PRICE will decrease by a factor of -10.27270508 for every unit increase in NOX.

d) Write the regression equation from this model.

Predicted AVG\_PRICE = 29.42847349 +0.03293496\*AGE  
+0.130710007\*INDUS-10.27270508\*NOX+0.261506423\*DISTANCE  
-0.014452345\*TAX-1.071702473 \*PTRATIO +4.125468959\* AVG\_ROOM-0.605159282\*LSTAT.