BUSINESS REPORT OF THE PROJECT WEEK-2

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

ANSWER:

- The mean value of variables are [4.871, 68.514, 11.136, 0.554, 9.549, 408.237, 18.455, 6.284, 12.653, 22.532], the median value of variables are [4.82, 77.5, 9.69, 0.538, 5, 330, 19.05, 6.208, 11.36, 21.2], the mode value of variables are [3.43, 100, 18.1, 0.538, 24, 666, 20.2, 5.713, 8.05, 50], the standard deviation of variables are [2.92, 28.148, 6.86, 0.11, 8.7,168.53 , 2.16, 0.70, 7.14, 9.19], the variance of variables are [8.53, 792.35, 47.06, 0.01, 75.81, 28404.75, 4.68, 0.49, 50.99, 84.98] out of which the highest value is for TAX and lowest is for NOX.
- The kurtosis of variables are [-1.18, -0.96, -1.23, -0.06, -0.86, -1.14, -0.28, 1.89, 0.49, 1.49] out of which 1.89, 0.49, 1.49 are positive kurtosis indicates heavier tails and a more peaked distribution. -1.18, -0.96, -1.23, -0.86, -1.14 are negative kurtosis indicates lighter tails and a flatter distribution. -0.28, -0.06 are close to zero suggests a normal distribution.
- The skewness values are [0.02, -0.59, 0.29, 0.72, -0.08, -1.14, -0.80,0.40,0.90,1.10] out of which 0.72,0.40,0.90,1.10 are positive skewness indicates a longer right tail, meaning the data is skewed towards higher values. -0.59, -1.14, -0.80 are negative skewness indicates a longer left tail, meaning the data is skewed towards lower values. 0.02, 0.29, -0.08 are close to zero suggests a relatively symmetric distribution.

QUESTION-2 Plot a histogram of the AVG_PRICE variable. What do you infer?

ANSWER:

- Histograms are used when you want to study the frequency distribution of a variable.
- The above histogram of the Avg_Price variable is positively skewed and represents observations as ‘trailing off’ to the right.
- Histogram sharp peak of the distribution is flat peak which represents negative kurtosis.

QUESTION-3 Compute the covariance matrix. Share your observations.

ANSWER:

The positive Covariance pairs are:

1. Covariance between TAX and TAX is 28348.62.
2. Covariance between AGE and TAX is 2397.94.
3. Covariance between DISTANCE and TAX is 1333.11.
4. Covariance between INDUS and TAX is 831.71.
5. Covariance between AGE and AGE is 790.79.

The Negatively Covariance pairs are;

1.Covariance between TAX and AVG_PRICE is -724.82

2.Covariance between AGE and AVG_PRICE is -97.39.

A negative sign of covariance value represents that two variables move to the opposite directions.

QUESTION-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.

ANSWER:

a)The top 3 positively correlated pairs are;

1. Correlation between DISTANCE and TAX is 0.91.
2. Correlation between INDUS and NOX is 0.76.
3. Correlation between INDUS and TAX is 0.72.

All the above top 3 correlated pairs are positively correlated and are close to +1,

which indicates a positive linear relationship between all the 3 correlated pairs of variables.

As the variable of one magnitude increases, the variable of the other one magnitude also increases.

b)The top 3 Negatively Correlated pairs are;

1.Correlation between LSTAT and AVG_PRICE is -0.73.

2.Correlation between AVG_ROOM and LSTAT is -0.61.

3.Correlation between PTRATIO and AVG_PRICE is -0.50.

All the above top 3 correlated pairs are negatively correlated and are close to -1,
which indicates a negative linear relationship between all the 3 correlated pairs of variables.
As the variable of one magnitude decreases, the variable of the other one magnitude also decreases.

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

ANSWER:

a)

- R-Square: the value of R-Square is 0.55 which indicates that 55% of the variability is explained by this model. It is quite low to the proportion of variance between Avg_Price variable and LSTAT. It indicates positive but not high enough.
- Co-efficient value of LSTAT is -0.95. It is a negative co-efficient which indicates that the Avg_Price(y) variable decreases with an decrease in the predictor.
- Intercept value of LSTAT is 34.55. It is a positive value which indicates the variable is high, so it shifts the regression line upwards.

b)

- P-value is less than 0.05. P-value is low which suggests strong evidence against null hypothesis, supporting the conclusion that the regression model is highly significant.
- Standard error value of LSTAT is 6.21 which is quite low. Looking at the standard error value, there is uncertainty in the relationship between Avg_Price and LSTAT to be very low.
- F-value: The F-statistic value is 601.61, which means that the explained variance is 601.61 of unexplained variance. The probability of trusting the goodness of this model is very low. It doesn't have significant impact on the regression model. We can't trust this model.

Hence upon analysing, we can say that this regression model is not a good fit to predict and it is a weak model. Hence, we reject this regression model.

QUESTION-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) 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.

ANSWER:

A) Regression Equation:

Avg_Price= (-1.35)+(5.09 * Avg_Room)+(-0.64 * LSTAT)

A new house in the locality has 7 rooms and has a value of 20 for LSTAT.

Avg_Price= (-1.35)+(5.09 * 7)+(-0.64 * 20)

Avg_Price= 21.48

Based on the regression equation predicted Avg_Price is 21.48 for these two values of independent variables Avg_Room and LSTAT.

Our predicted value has + or – standard deviation range in our prediction and standard deviation is 5.5.

Average deviation of actual value and predicted value is

21.48 + 5.5 = 26.48

21.48 – 5.5 = 15.98

According to Standard Deviation the maximum company should be quoted somewhere about 27,000 USD. As per the question the company is quoting a value of 30,000 USD which is overcharging.

B) In the previous question the value of adjusted R-Square is 0.54 which indicates that 54% of the variability is explained by this model.

In this question the value of adjusted R-Square is 0.63 which indicates that 63% of the variability is explained by this model.

Hence this question model of adjusted R-square is great model but better than previous question adjusted R-square model.

QUESTION-7 Build another Regression model with all variables where AVG_PRICE alone be the Dependent Variable and 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.

ANSWER:

1. The intercept value is 29.24 represents the predicated value of the variable is positive indicates the variable is high, so it shifts the regression line upward.
2. Co-efficient of NOX is very low when compared to all remaining independent variables, negative co-efficient indicates a decrease in the predicted value.
3. P-value is less than 0.05 which conclude that there is a relationship between all the independent variable and the probability of error is low. P-value of CRIME_RATE (0.53) is higher than 0.05 is somewhat statistically significant.
4. Multiple R value is 0.83, which represents the correlation between the predicted values and the observed values of the variable ranges very close to +1 indicating a perfect positive correlation.
5. R-square value is 0.69 measures the proportion of variance that ranges less than 1 indicates that the model explains all the variability in the dependent variable.
6. Adjusted R-square value is 0.68, which is a low value and it is not significant to this regression model.
7. Standard Error is 5.13 represents the average deviation between the variables.
8. F-STAT value is 124.9, which means that the explained variance is 124.9 times of unexplained variance. The probability of trusting the goodness of this model is very low in number and close to zero. It doesn't have significant Impact on the model.

   Hence, we can trust this model and we can consider co-efficient of this model to access predicted value, it is a good regression model if we exclude CRIME_RATE it will be more significant.

Regression Equation:

AVG PRICE= (29.42)+(0.048*CRIME RATE) +(0.03*AGE)+(0.13*INDUS)+(-10.27)*NOX)+(0.26*DISTANCE)+(-0.14)*TAX) +(-1.07)*PTRATIO) +(4.12*AVG_ ROOM)+(-0.60)*LSTAT)

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 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.

ANSWER:

a)

- The intercept value is 29.42 represents the predicated value of the variable is positive indicates the variable is high, so it shifts the regression line upward.
- P-value is less than 0.05, which conclude that there is a relationship between all the independent variables. Hence it indicates statistically significant relationship.
- The Multiple R value is 0.83 which is a high positive correlation between the actual and predicted values.
- The R-square value of this model is 0.69 which is not a good proportion of variance that is explained by this model. This gives us an idea that we need to look into our independent variables and optimize the model further.
- The Adjusted R-square is almost the same as R-square which indicates that the current is not affected by the number of independent variables.
- The Standard Error of the model seems to be on the higher side which indicates that we may want to optimize this model further to reduce the Standard Error.
- F-STAT value is 5.13, which means that the explained variance is 5.13 times of unexplained variance. The probability of trusting the goodness of this model is very low. It doesn't have significant Impact on the model.

b)

In previous question Adjusted R-square value is 0.68, which is a low value and it is not significant to this regression model.

The Adjusted R-square is 0.68 almost the same as R-square which indicates that the current is not affected by the number of independent variables

Both the models same with respect to adjusted R-square.

c)

After sorting the values of co-efficient in ascending order the Avg_Price value of the NOX is -3.22 which is negative and is less in a locality in this town.

d)

Regression Equation:

AVG PRICE= (29.42) +(0.03*AGE) +(0.13 *INDUS)

(-10.27)*NOX)+(0.26" DISTANCE)+(-0.14)*TAX) +(-1.07)"PTRATIO) +(4.12*AVG_ ROOM)+(-0.60)*LSTAT)