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

Observation;

1. if the mean is significantly different from the median, it may indicate skewness in the data.

2.Positive Kurtosis (Leptokurtic): A positive kurtosis value indicates that the distribution has heavier tails and a sharper peak compared to a normal distribution. This suggests that the data has more extreme values or outliers than would be expected under a normal distribution. It indicates a higher propensity for large deviations from the mean.

Negative Kurtosis (Platykurtic): A negative kurtosis value indicates that the distribution has lighter tails and a flatter peak compared to a normal distribution. This suggests that the data has fewer extreme values or outliers than would be expected under a normal distribution. It indicates a lower propensity for large deviations from the mean.

3.A smaller standard error typically results in a narrower confidence interval and increases the chances of finding a statistically significant result.

2) Plot a histogram of the Avg_Price variable. What do you infer?

From the Histogram we are able to observe that,

It has a trailing off effect to the right captured by positive skewness.

It is not a proper bell-shaped curve.

3) Compute the covariance matrix. Share your observations.

The diagonal elements of the matrix represent the variances of the variables, while the off-diagonal elements represent the covariances between pairs of variables.

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

a) Which are the top 3 positively correlated pairs and

b) Which are the top 3 negatively correlated pairs.

Solution:

A) Top 3 positively correlated pairs are

1) DISTANCE and TAX with (0.91) value.

2) INDUSTRY and NOX- Nitric oxide concentration (0.76) Value.

3) AGE and NOX- Nitric oxide concentration (0.73) value.

B) Top Three Negatively Correlated Pairs are

1) LSTAT and AVG_PRICE (-0.73)

2) AVG_ROOM and LSTAT (-0.61)

3) PTRATIO (pupil-teacher ratio by town) and AVG_PRICE (-0.50).

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

Solution (a)

1.R-Square - The R-square value is 54.4%, which indicates that approximately 54.4% of the variation in Average price can be explained by the variation LSTAT.But by Looking at P-Value is close to 0 Therefore, It seems safe to conclude TV seems to be a significant predictor of Sales.

2. Upon looking at the Intercept and Slope Coefficient, we can say that the equation of the best fit line is,

Average price = 34.55 + (-0.95) * LSTAT + e. Intercept 34.55 gives an indication that if there is 0% LSTAT, the expected Average price is 34.55.

3. The coefficient of Average price is -0.95 which indicates that for every unit increase in the LSTAT, we can expect Average price to go down by -0.95 units.

Q5. (b) Is LSTAT variable significant for the analysis based on your model?

SOLUTION: b

1. In this regression model, the F-stat is 601.61 which is the ratio between the explained and unexplained variance in the Average price by this model. It means that the variance explained by the model is 601.61 times the variance not explained by the model. This is a very High ratio indicating that this regression model can be a good predictor of the Average price.

2. This is further validated by looking at R square and Adjusted R square values. The R-square value is 54.41%, which indicates that approximately 54.41% of the variation in AVG_PRICE can be explained by the variation in LSTAT. The adjusted R square is same , which means that the standardized explained varianced, adjusted to the number of independent variables is also same.

3. Standard error - In this case, the standard error is approximately 6.21 units, indicating that on an average, the predicted AVG_PRICE may deviate from the actual AVG_PRICE value by about 6.21 units.

4. Looking at the P-values of then LSTAT variable, we can see that this variable has a p-value lesser than 0.05 . It indicates that this variable are good predictor of Average price

5. F-Statistic - The F-statistic tests the overall significant of the regression model. It compares the variation explained by the regression to the unexplained variation. A high F-value indicates that the regression model as a whole is statistically significant. In this case, the F-value is very high, which indicates that this regression model is

highly significant.

Upon Analysing we can conclude that LSAT variable is significant for the analysis of our model

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?

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.

Solution (a) ;

The Regression equation will be :- AVG_PRICE= (-1.35)+(5.09*AVG_ROOM)+(-0.64*LSTAT)

If a new house in this locality has 7 rooms (on an average) and has a value of 20 for L-STAT, by substituting the values in Regression equation. The value of AVG_PRICE will be 21.48.

The other company is clearly overcharging. Because if we consider standard error its value is 5.54, if we add 5.54 with 21.48 we get 27.02, while the company quoted value is 30,000 USD more than our Predicted value. If we subtract 5.54 also we get 15.94, there is absolutely no chance of quoted value falling in that range, so we conclude that the company is overcharging.

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 Rsquare, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE.

Solution :

Adjusted R-squared: The adjusted R-squared value of 0.688 is slightly lower than the R-squared value. Adjusted R-squared takes into account the number of predictors in the model and penalizes for the inclusion of additional variables. The adjusted R-squared value of 0.688 indicates that approximately 68.83% of the variance in the dependent variable is explained by the independent variables while considering model complexity.

Comparing the R-squared and adjusted R-squared values, the adjusted R-squared is lower because it accounts for the number of predictors in the model. This suggests that the inclusion of additional predictors in the model may not contribute significantly to improving the model's explanatory power. Therefore, the adjusted R-squared provides a more conservative estimate of the model's effectiveness in explaining the variance in the dependent variable.

Overall, with an R-squared value of around 69.39% and an adjusted R-squared value of approximately 68.83%, the model shows a moderate level of explanatory power, indicating that the included predictors have some ability to explain the variation observed in the Average price

interpreting in term of coefficient

CRIME_RATE :(per capita crime rate by town)

For every unit increase in crime rate, AVG_PRICE will increase by 0.048 units. and P-value for crime rate tells us that , with a p-value of 0.534, it indicates that the observed result is not statistically significant at a conventional significance level of 0.05. In other words, there is insufficient evidence to reject the null hypothesis and conclude that there is a significant effect or relationship between the variables.

AGE : (proportion of houses built prior to 1940 (in percentage terms))

For every unit increase in AGE, AVG_PRICE will increase by 0.032 units. and P-value for AGE tells us that , with a pvalue of 0.01, is considered statistically significant. It suggests that the observed result is unlikely to have occurred by chance alone, assuming the null hypothesis (no effect or relationship) is true.

INDUSTRY ;(proportion of non-retail business acres per town (in percentage terms))

For every unit increase in INDUSTRY, AVG_PRICE will increase by 0.13 units. and P-value for INDUSTRY tells us that , with a pvalue of 0.03, is considered statistically significant. It suggests that

the observed result is unlikely to have occurred by chance alone, assuming the null hypothesis (no effect or relationship) is true.

NOX : (nitric oxides concentration (parts per 10 million)

For every unit increase in NOX, AVG_PRICE will decrease by 10.32 units. and P-value for NOX tells us that , with a pvalue of 0.008, is considered statistically significant. It suggests that the observed result is unlikely to have occurred by chance alone, assuming the null hypothesis (no effect or relationship) is true.

DISTANCE : (distance from highway (in miles)

For every unit increase in DISTANCE, AVG_PRICE will increase by 0.26 units. and P-value for INDUSTRY tells us that , with a pvalue of 0.0001, is considered statistically significant. It suggests that the observed result is unlikely to have occurred by chance alone, assuming the null hypothesis (no effect or relationship) is true.

TAX : (full-value property-tax rate per $10,000)

For every unit increase in TAX, AVG_PRICE will decrease by 0.01 units. and P-value for INDUSTRY tells us that , with a pvalue of 0.0002, is considered statistically significant. It suggests that the observed result is unlikely to have occurred by chance alone, assuming the null hypothesis (no effect or relationship) is true.

PTRATIO : (pupil-teacher ratio by town)

For every unit increase in PTRATIO, AVG_PRICE will decrease by 1.07 units. and P-value for PTRATIO tells us that , with a pvalue of closer to 0 , is considered statistically significant. It suggests that the observed result is unlikely to have occurred by chance alone, assuming the null hypothesis (no effect or relationship) is true.

AVG_ ROOM : (average number of rooms per house)

For every unit increase in AVG_ROOM, AVG_PRICE will increase by 4.12 units. and P-value for AVG_ROOM tells us that , with a pvalue of closer to 0 , is considered statistically significant. It suggests that the observed result is unlikely to have occurred by chance alone, assuming the null hypothesis (no effect or relationship) is true.

LSTAT : (% lower status of the population)

For every unit increase in LSTAT, AVG_PRICE will decrease by 0.60 units. and P-value for PTRATIO tells us that , with a pvalue of closer to 0 , is considered statistically significant. It suggests that the observed result is unlikely to have occurred by chance alone, assuming the null hypothesis (no effect or relationship) is true.

Upon looking at the Intercept and Slope Coefficient, . Intercept 29.24 gives an indication that if there are 0units of independent factors, the expected sales is 29.24.

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.

SOL (A)

1. In this regression model, the F-stat is 140.63 which is the ratio between the explained and unexplained variance in the Average price by this model. It means that the variance explained by the model is 140.63 times the variance not explained by the model. This is a moderate value indicating that this regression model can be a good predictor of the Average price.

SOL (B)

The multiple R, R square and adjusted R-square seems to have not increased much from the previous model.When the R-squared and multiple R-squared increase, it suggests that the additional independent variables or changes made to the model have enhanced its ability to explain or predict the variation in the dependent variable. This increase indicates a better fit of the model to the data and stronger relationship between the independent and dependent variables.

It's important to note that an increase in R-squared or multiple R-squared doesn't necessarily mean the model is perfect or that it includes all relevant variables. Other factors, such as model complexity, sample size, and potential multicollinearity, should also be considered when interpreting the results and assessing the overall quality of the model.

After comparing adjusted Squared, we can see that this model performs better with respect to previous model

SOLUTION (C)

All the values are sorted in ascending order, For every unit increase in NOX, AVG_PRICE will DECREASE.

SOLUTION D :

AVG_PRICE= 29.42+((-10.27)*NOX)+((-1.07)*PTRATIO)+((-0.60)*LSTAT))+((-0.014)*TAX)+(0.03*AGE)+(0.13*INDUS)+(0.26*DISTANCE)+(4.12*AVG_ROOM))

Solution (b) :

I. In this model, the F-stat is 444.33 which is a lower compared to the F-stat of the previous model which was about 601.61. This indicates that this regression model is not a better fit to predicted AVG_PRICE than the previous one. It also validates our assumption that AVG_ROOM (average number of rooms per house) do not have a significant impact on the AVG_PRICE.

2. The multiple R, R square and adjusted R-square seems to have increased much from the previous model. When the R-squared and multiple R-squared increase, it suggests that the additional independent variables or changes made to the model have enhanced its ability to explain or predict the variation in the dependent variable. This increase indicates a better fit of the model to the data and stronger relationship between the independent and dependent variables.

It's important to note that an increase in R-squared or multiple R-squared doesn't necessarily mean the model is perfect or that it includes all relevant variables. Other factors, such as model complexity, sample size, and potential multicollinearity, should also be considered when interpreting the results and assessing the overall quality of the model.

3. We can see that the p-value of LSTAT and AVG_ROOM is less than 0.05. it indicates both of these variables are good predictors of AVG_PRICE.

4.This model has Significance F value Lesser than the previous model, it tells us the probability of this model being wrong is lesser than previous model.