**Terro's Real Estate Agency**

**TABLE OF CONTENT**

|  |  |  |
| --- | --- | --- |
| **NO** | **TITLE** | **PAGE NO** |
| **1** | **Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation.** | **4** |
| **2** | **Plot a histogram of the Avg\_Price variable.** | **5** |
| **3** | **Compute the covariance matrix. Share your observations.** | **6** |
| **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.** | **7** |
| **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 Residual plot?**  **b) Is LSTAT variable significant for the analysis based on your model?** | **8** |
| **6** | **Build a new Regression model including LSTAT and AVG\_ROOM together as Independent variables and AVG\_PRICE as dependent variable** | **9** |
| **7** | **Build another Regression model with all variables where AVG\_PRICE alone be the Dependent Variable and all the other variables are independent.** | **10** |
| **8** | **Pick out only the significant variables from the previous question. Make another instance of the Regression mode** | **11** |

**Data given**

Terro’s real-estate is an agency that estimates the pricing of houses in a certain locality. The pricing is concluded based on different features / factors of a property. This also helps them in identifying the business value of a property. To do this activity the company employs an “Auditor”, who studies various geographic features of a property like pollution level (NOX), crime rate, education facilities (pupil to teacher ratio), connectivity (distance from highway), etc. This helps in determining the price of a property.

**Data Dictionary:**

|  |  |
| --- | --- |
| **ATRIBUTE** | **DISCRIPTION** |
| CRIME RATE | per capita crime rate by town |
| INDUSTRY | proportion of non-retail business acres per town (in percentage terms) |
| NOX | nitric oxides concentration (parts per 10 million) |
| AVG\_ROOM | average number of rooms per house |
| AGE | proportion of houses built prior to 1940 (in percentage terms) |
| DISTANCE | distance from highway (in miles) |
| TAX | full-value property-tax rate per $10,000 |
| PTRATIO | upil-teacher ratio by town |
| LSTAT | % lower status of the population |
| AVG\_PRICE | Average value of houses in $1000' |

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

**Observation**

a) Crime Rate: This area boasts an impressively low crime rate, with an average of just 4.87 and a maximum rate of only 0.04.

b) Age: The average age of residents here is 68.57 years.

c) Industrial Area (INDUS): The typical INDUS value is 11.14, and the middle value (median) is 9.69.

d) Nitrogen Oxide Levels (NOX): NOX levels are quite low, with an average of 0.55 and a median of 0.53

e) Distance to Highway: On average, the distance to the nearest highway is 9.55 miles.

f) Property Tax (TAX): The average property tax is $408.23, with a median of $330.

g) Pupil-Teacher Ratio (PTRATIO): The town's average PTRATIO is 18.46, with a median of 19.05. The kurtosis of -0.285 suggests a relatively flat distribution without significant peaks.

h) Average Room Size: Homes in this area have an average of 6.28 rooms, and the positive kurtosis of 1.891 indicates a peak in room availability.

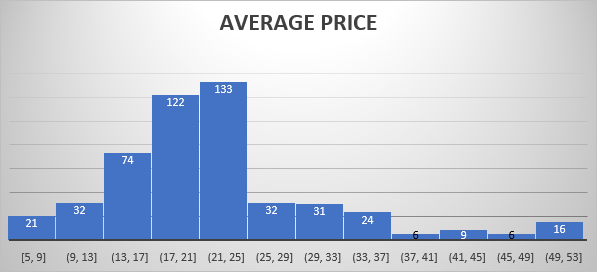
i) Average Price: The mean price for homes is $22.53, with a median of $21.2. There are 50 instances of the most frequently occurring price (mode) at this value.

**2) Plot a histogram of the Average Price variable. What do you infer?**

**HISTOGRAM**

* A histogram is a graphical representation of the distribution of a dataset. It divides the data into intervals or bins along the x-axis and displays the frequency or count of data points falling within each interval on the y-axis.
* Histograms provide insights into the shape, central tendency, and spread of data, making it easier to identify patterns, trends, and outliers in the dataset.
* They are commonly used in statistics and data analysis to visualize the underlying distribution of data. Histograms are especially useful for understanding data that may not follow a normal distribution.

**HISTOGRAM FOR AVERAGE PRICE**



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

**CONCLUSION**

Certainly, here are the conclusions based on the provided data:

**1**. CRIME\_RATE: The average crime rate in this area is 8.516, significantly higher than the maximum rate of 0.04, indicating a relatively high crime rate in this location.

**2**. AGE: The average age of people living in this area is 68.574 years.

**3**. INDUS: The average industrialization rate (INDUS) is 11.1367, with a median of 9.69.

**4**. NOX: The average level of nitrous oxide (NOX) is 0.55, and the median is 0.538, suggesting relatively low NOX levels.

**5**. DISTANCE: The average distance to highways is 9.549 miles.

**6**. TAX: The average property tax is 408.23, with a median of 330.

**7**. PTRATIO: The average student-teacher ratio (PTRATIO) is 18.455, with a median of 19.05. The kurtosis value of -0.285 indicates a relatively flat distribution with no pronounced peak.

**8**. AVG\_ROOM: The average number of rooms per house is 6.28, and the positive kurtosis of 1.891 suggests a peak in the distribution, indicating that there are more houses with a higher number of rooms available.

**9**. LSTAT: Details about LSTAT are missing from the table.

**10**. AVG\_PRICE: The mean average house price is 22.53380632, with a median of 21.2. There is a frequent mode for AVG\_PRICE at 50, indicating that certain house prices are repeated frequently in the dataset

**4.Create a correlation matrix of all the variables**

**a)Which are the top 3 positively correlated pairs**

* The top three correlated pairs are

1. The correlated pairs with the tax and distance is high at number 1

0.910228188533182.

1. The second top positively correlated pair is NOX with INDUS is 0.763651446920914
2. The top third positively correlated pair is NOX with AGE is

|  |  |  |  |
| --- | --- | --- | --- |
|  | DISTANCE | INDUS | AGE |
| TAX | 0.9102 | N/A | N/A |
| NOX | N/A | 0.763651 | N/A |
| NOX | N/A | N/A | 0.731 |

**b) Which are the top 3 negatively correlated pairs.**

**1.** The top negatively corelated pair is AVERAGE PRICE with LSTA

Is -0.737662726174014.

**2.** The top second negatively corelated pair is LSTAT and AVERAGE ROOMS is -0.613808271866395

**3.** The top third negatively corelated pair is AVERAGE PRICE with PTRATIO is -0.507786685537561

|  |  |  |  |
| --- | --- | --- | --- |
|  | DISTANCE | INDUS | AGE |
| TAX | 0.9102 | N/A | N/A |
| NOX | N/A | 0.763651 | N/A |
| NOX | N/A | N/A | 0.731 |

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

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

Y(predicted avg price) = -0.950049354 \* LSTAT + 34.55384088

* This equation represents a linear relationship between the predicted average price (Y) and the LSTAT variable. The coefficient of -0.950049354 indicates how much the predicted average price is expected to change for a one-unit increase in the LSTAT variable, assuming all other variables remain constant.
* So, if the LSTAT variable increases by 1 unit while keeping all other factors constant, the predicted average price is expected to decrease by approximately 0.950049354 units. This suggests that there is a negative relationship between LSTAT and the predicted average price. In other words, as the percentage of lower-status population (LSTAT) in an area increases, the average housing price is expected to decrease, according to this regression model.

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

As per my model the LAST variable is significant based on my analysis

Because the p value is 7.68275944181912E-88 as per the statistic model when the p-value is less than 0.05 then the data is significant data.

**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

**The regression equation is**

**equation**= intercept+(coefficient of Lstat\*value for lstat) + (coefficient of average room\*value for average room)

Intercept= -1.361465588

Coefficient of lstat= 5.095011657

Coefficient of average room= -0.642307146

value for lstat=20

value for average room=7

After substituting the values in the formula we have got **21.45747309**

**Average price of the house is 21.45747309**

* So price quoting value of 30000 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 is increased from the previous question

Adjustable R square value of pervious is 0.543204362

Adjustable R square value of this one is 0.6371

So by observing the above values model for this question is higher so this question model is better model

**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.**

a. The adjusted R^2 for the multiple regression model involving all variables is 0.693, which is significantly higher than zero. A higher adjusted R^2 indicates that the additional input variables are contributing valuable information to the model, making it a more robust and effective regression model.

b. The intercept value (c) is calculated as 29.2413152565006, and it represents the intercept for the model when all variables are held at zero.

c. The coefficient for the 'Crime Rate' variable (M) is 0.048725141.

d. The coefficient for the 'Age' variable (M) is 0.032770689.

e. The coefficient for the 'Industrial Area' variable (M) is 0.130551399.

f. The coefficient for the 'NOx' variable (M) is 10.3211828.

g. The coefficient for the 'Distance' variable (M) is 0.261093575.

h. The coefficient for the 'Tax' variable (M) is 0.01440119.

i. The coefficient for the 'Partio' variable (M) is 1.074305348.

j. The coefficient for the 'Average Room' variable (M) is 4.125409152.

k. The coefficient for the 'LSTAT' variable (M) is 0.603486589.

**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**

Regression Statistics:

**Multiple R** (Multiple Correlation Coefficient): 0.831276051

This value represents the correlation between the predicted values and the actual values of the dependent variable (Y). It indicates how well the independent variables collectively explain the variation in the dependent variable.

**R Square** (Coefficient of Determination): 0.691019874

R-squared measures the proportion of the variance in the dependent variable (Y) that can be explained by the independent variables in the model. In this case, approximately 69.10% of the variance in Y is explained by the independent variables.

**Adjusted R Square**: 0.686668041

Adjusted R-squared adjusts the R-squared value for the number of predictors in the model. It is slightly lower than R-squared, indicating that the model has a good fit but may not be overfitting.

**Standard Error**: 5.146443938

The standard error is a measure of the dispersion of the observed values around the regression line. In this case, it's approximately 5.15, which provides an estimate of the typical error in predicting Y.

**Observations**: 505

This is the number of data points or observations used in the regression analysis.

**ANOVA** (Analysis of Variance):

ANOVA is used to test the overall significance of the regression model.

The **F-statistic** is 158.79, and the associated p-value (Significance F) is very close to zero (2.1027E-122). This suggests that at least one of the independent variables in the model is statistically significant in explaining the variance in the dependent variable.

**Coefficients:**

* This section provides information about the coefficients (slopes) of the independent variables in the regression model.

**Intercept:** The intercept represents the estimated value of the dependent variable (Y) when all independent variables are zero. In this case, it's approximately 25.82.

* Coefficients for Independent Variables: Each independent variable has an associated coefficient.

**INDUS:** 0.15304344

**NOX**: -5.482039233

**DISTANCE:** 0.250267442

**TAX**: -0.015118139

**PTRATIO**: -1.000447003

**AVG\_ROOM:** 4.371826045

**LSTAT:** -0.56558964

* These coefficients represent the estimated change in the dependent variable (Y) for a one-unit change in the corresponding independent variable while holding all other variables constant.
* The p-values associated with each coefficient indicate whether each independent variable is statistically significant in explaining Y. If the p-value is small (typically less than 0.05), it suggests that the variable is significant.

The lower and upper 95% confidence intervals provide a range within which the true population parameter (coefficient) is likely to fall

**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?**

Adjusted R square value of this data is **0.686668041.**

Adjusted R square value of previous data is **0.688269237.**

* By comparing the two data the difference is 0.002642.

The better value is the data which removed the crime rate (This data)

**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?**

**ACCENDING ORDER OF COEFFICIENTS**

|  |  |
| --- | --- |
|  | *Coefficients* |
| NOX | -5.482039233 |
| PTRATIO | -1.000447003 |
| LSTAT | -0.56558964 |
| TAX | -0.015118139 |
| INDUS | 0.15304344 |
| DISTANCE | 0.250267442 |
| AVG\_ROOM | 4.371826045 |
| Intercept | 25.81504137 |

The R square value of the NOX is **0.182787255** the value is not closer to the 1 So when the average price of the NOX is increased then the average price will be decrease.

**d)Write the regression equation from this model.**

**The regression equation for this model is**

Y(predicted) = 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