TERRO'S REAL ESTATE AGENCY

Real estate data analysis – Exploratory data analysis, Linear Regression

Objective (Task):

Your job, as an auditor, is to analyze the magnitude of each variable to which it can affect the price of a house in a particular locality.

To do the analysis, you are expected to solve these questions:

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

Answer :-

(CRIME_RATE)

1. **Average Crime Rate**: The average crime rate in the dataset is around 4.87, indicating a moderate level of crime.
2. **Variability**: Crime rates vary notably, with a range from 0.04 to 9.99 and a standard deviation of approximately 2.92.
3. **Distribution**: The data has a roughly symmetric distribution with a slight right skew and is less peaked compared to a normal distribution.

(AGE)

1. **Average Age**: The average age of houses in the dataset is about 68.57. This suggests that, on average, houses in the represented areas are relatively old.
2. **Variability in Age**: The ages of houses vary significantly, with the dataset ranging from a minimum of 2.9 years to a maximum of 100 years. The standard deviation of approximately 28.15 indicates this notable variability.
3. **Distribution Shape**: The age distribution is negatively skewed, with a skewness of approximately -0.60. This means that more houses tend to be older, pulling the distribution towards the younger side. However, there are still some relatively young houses in the dataset.

(INDOS)

1. **Average Industrialization**: The average level of industrialization in these areas is approximately 11.14.
2. **Variability**: The extent of industrialization varies considerably, ranging from as low as 0.46 to as high as 27.74. This means that some areas are much more industrialized than others.
3. **Distribution Shape**: The data is somewhat skewed to the right, suggesting that more areas have moderate industrialization levels, but there are still some with very high industrialization levels.

(NOX)

1. **Average Nitrogen Oxide Level (NOX)**: On average, the level of nitrogen oxide (NOX) in these areas is around 55%.
2. **Variability**: The NOX levels vary quite a bit, with a range from 39% to 87%, indicating some areas have much higher levels of NOX pollution than others.
3. **Distribution Shape**: The data is heavily skewed to the right, with a high positive skewness of 73%. This means that there are a few areas with extremely high NOX levels, which pull the average higher.
4. **Kurtosis**: The data has negative kurtosis, indicating that it has thinner tails and is less peaked compared to a normal distribution.

(DISTANCE)

1. **Average Distance**: On average, the distance to various amenities or locations in these areas is approximately 9.55 units.
2. **Variability**: The distances vary notably, ranging from 1 unit to 24 units, indicating that the proximity to different amenities or locations can greatly differ in these areas.
3. **Distribution Shape**: The data is positively skewed with a skewness value of approximately 1. This suggests that most areas are relatively close to amenities, but there are some outliers where the distances are much greater.
4. **Kurtosis**: The data has negative kurtosis, indicating that it is less peaked and has thinner tails compared to a normal distribution.

(TAX)

1. **Average Tax Rate**: The average tax rate in these areas is approximately 408.24.
2. **Variability**: Tax rates vary significantly, with the lowest being 187 and the highest being 711. This indicates a wide range of tax rates in these areas.
3. **Distribution Shape**: The data is positively skewed (skewness of approximately 0.67), which means that more areas have higher tax rates than lower ones, but there are still some areas with relatively lower tax rates.
4. **Kurtosis**: The data has negative kurtosis, which suggests that it is less peaked and has thinner tails compared to a normal distribution.

(PTRATIO)

1. **Average Pupil-Teacher Ratio**: The average pupil-teacher ratio in the dataset is approximately 18.46.
2. **Variability**: Pupil-teacher ratios vary across the dataset, ranging from a minimum of 12.6 to a maximum of 22. This indicates differences in class sizes in various areas.

3. **Distribution Shape**: The data is negatively skewed (skewness of approximately -0.80), which means that more areas tend to have higher pupil-teacher ratios, but there are still some areas with lower ratios.

4. **Kurtosis**: The data has negative kurtosis, suggesting that it is less peaked and has thinner tails compared to a normal distribution.

(AVG ROOM)

1. **Average Number of Rooms**: The average number of rooms in the houses in the dataset is approximately 6.28 rooms.

2. **Variability**: The number of rooms in the houses varies, ranging from a minimum of 3.561 rooms to a maximum of 8.78 rooms.

3. **Distribution Shape**: The data is slightly positively skewed (skewness of approximately 0.40), indicating that more houses tend to have a higher number of rooms, but there are still some houses with fewer rooms.

4. **Kurtosis**: The data has positive kurtosis, suggesting that it is somewhat peaked compared to a normal distribution, meaning that there is a concentration of data around the average number of rooms.

(LSTAT)

1. **Average Lower Status Percentage**: On average, the percentage of lower-status population in the neighbourhoods is approximately 12.65%.

2. **Variability**: The percentage of lower-status population in neighbourhoods varies, with a range from a minimum of 1.73% to a maximum of 37.97%.

3. **Distribution Shape**: The data is positively skewed (skewness of approximately 0.91), indicating that most neighbourhoods have a lower percentage of lower-status population, but there are some neighbourhoods with higher percentages.

4. **Kurtosis**: The data has positive kurtosis, suggesting that it is somewhat peaked compared to a normal distribution, meaning that there is a concentration of data around the average percentage of lower-status population

(AVG PRICE)

1. **Average House Price**: The average house price in the dataset is approximately $22,533, indicating the typical price of houses in the neighbourhoods.

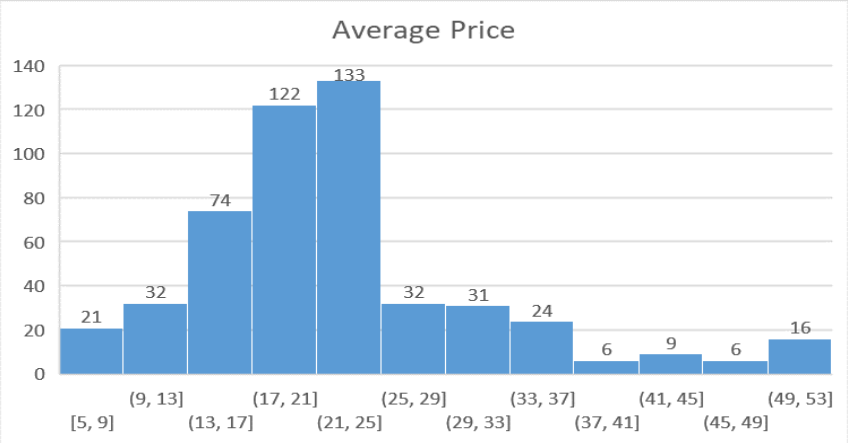
2. **Variability**: House prices vary, with the least expensive house priced at $5 and the most expensive at $50. The data has a standard deviation of about $9.20, indicating how much prices deviate from the average.

3. **Distribution Shape**: The data is positively skewed (skewness of approximately 1.11), which means that there are relatively few neighbourhoods with very high house prices, pulling the average price upward.

4. **Kurtosis**: The data has positive kurtosis, suggesting that it is somewhat peaked compared to a normal distribution, indicating a concentration of house prices around the average.

Q.2) Plot a histogram of the Avg-Price variable. What do you infer? (5 marks)

Answer:-



1. **Most Houses Are Priced Similarly**: Most neighbourhoods have houses that are priced similarly. This means that in these areas, the prices of houses are close to each other.
2. **Few Very Expensive Houses**: In a few neighbourhoods, there are some houses that are extremely expensive compared to others. These neighbourhoods have a few houses that are way more costly.
3. **Some Very Affordable Houses**: On the other hand, there might be some neighbourhoods where the houses are quite affordable compared to the rest.
4. **Variation in Prices**: Overall, the prices of houses vary across all neighbourhoods, but in many places, they tend to be around a certain price range.

Q.3) Compute the covariance matrix. Share your observations. (5 marks)

Answer:-

- Covariance values indicate how two columns in the table are related.
- A positive covariance suggests a positive relationship between the columns.
    - This means when one column's value increases, the other column's value also tends to increase.
- A negative covariance implies a negative relationship between the columns.
    - This indicates that when one column's value increases, the other column's value tends to decrease.
- The sign of the covariance value helps determine whether the two columns move together or in opposite directions.

Q.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 and

b) Which are the top 3 negatively correlated pairs

Answer:-

Top 3 positively correlated pairs:

1. Positive correlation between **TAX** and **DISTANCE**.
2. Positive correlation between **INDUS** and **NOX**.
3. Positive correlation between **NOX** and **AGE**.

Top 3 negatively correlated pairs:

1. Negative correlation between **LSTAT** and **AVG PRICE**.
2. Negative correlation between **AVG ROOM** and **LSTAT**.
3. Negative correlation between **AVG PRICE** and **PTRATIO**.

Q.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?

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

Answer:-

a):-- Infer from the Regression Summary output,

1. **Variance Explained:** The R-squared value of **0.544** suggests that approximately **54.4%** of the variance in the dependent variable can be explained by the independent variable(s) in the model.
2. **Coefficient Value:** For the variable **LSTAT**, the coefficient value is **-0.950**. This means that for every one-unit increase in **LSTAT**, the dependent variable decreases by **0.950 units.**
3. **Intercept:** The intercept value is **34.5538**. In the context of the model, this suggests that when the independent variable(s) are at 0, the estimated average value of the dependent variable is **34.5538.**
4. **Residual Plot:** Without the specific plot, it's not possible to provide a direct interpretation. However, in general, a good model would have residuals that are randomly scattered around the horizontal axis. Patterns in the residual plot could indicate that the model does not capture all the explanatory information.

b):-- Yes, the LSTAT variable is significant for the analysis based on the model. This is evident from the following observations:

1. The coefficient for **LSTAT** is **-0.950049354**, indicating that for every one-unit increase in LSTAT, the dependent variable decreases by approximately **0.95 units**, holding other variables constant.
2. The t-statistic value for **LSTAT** is **-24.52789985**, suggesting that the coefficient is statistically significant. A higher absolute t-value, combined with a small p-value **(5.0811E-88)**, indicates that the effect of LSTAT on the dependent variable is not likely due to random chance.
3. The **95%** confidence interval for the coefficient does not include zero, further supporting the conclusion that the LSTAT variable is significant for the analysis.

Q.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.

Answer:-

a)

The regression equation is *straight line with multiple columns*:

(Y=m1x1+m2x2+m3x3……..+Mnxn+c)

Given that the new house has 7 rooms and an LSTAT value of 20, we can calculate the predicted AVG_PRICE:

Y=5.094787984×7+(-0.642358334) ×20+(-1.35827812) y=21.45807639, y=$21,450

The predicted value of AVG_PRICE is approximately **$21,450.**

Comparing this value to the company's quote of 30,000 USD, the model's prediction suggests that the company is overcharging for the house with 7 rooms and an LSTAT value of 20 since the model's predicted price is significantly lower.

b)

Comparing the adjusted R-squared values of the two models:

1. For the previous model (Question 5):

- Adjusted R-squared: **0.543241826**

2. For the current model:
    - Adjusted R-squared: **0.637124475**

The adjusted R-squared value for the current model **(0.637124475)** is higher than the adjusted R-squared value for the previous model **(0.543241826).** This indicates that the current model explains a larger proportion of the variance in the dependent variable compared to the previous model.

Therefore, based on the adjusted R-squared values, we can conclude that the performance of the current model is better than the previous model. The higher adjusted R-squared value suggests that the current model is a better fit for the data and is more effective in explaining the variance in the dependent variable.

Q.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. (8 marks)

Answer:-

1. **Adjusted R-squared :** The value of **0.688** indicates that about **68.8%** of the changes in the house prices (AVG_PRICE).
2. **Intercept (Intercept Value):** When all the other factors are zero, the model predicts the house price to be approximately **29.24 units**.
3. **Coefficient Values:**
    - A one-unit increase in **CRIME_RATE** leads to an increase in house price by **0.0487** units.
    - A one-unit increase in **AGE** results in an increase in house price by **0.0328 units.**
    - A one-unit increase in **INDUS** leads to an increase in house price by **0.1306 units**.
    - A one-unit increase in **NOX** leads to a decrease in house price by **10.3212 units.**
    - A one-unit increase in **DISTANCE** results in an increase in house price by **0.2611 units.**
    - A one-unit increase in **TAX** results in a decrease in house price by **0.0144 units.**
    - A one-unit increase in **PTRATIO** leads to a decrease in house price by **1.0743 units**.
    - A one-unit increase in **AVG_ROOM** leads to an increase in house price by **4.1254 units.**
    - A one-unit increase in **LSTAT** results in a decrease in house price by **0.6035 units.**

4. **Significance of Independent Variables:** Except for **CRIME_RATE** and **DISTANCE**, all the other variables, including **AGE**, **INDUS**, **NOX**, **TAX**, **PTRATIO, AVG_ROOM**, and **LSTAT**, are statistically significant in explaining the changes in the house prices. This means that these factors have a significant impact on determining house prices.

Q.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.

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:-

Certainly, here are the answers to the given questions based on the output summary provided:

a) Interpretation of the Output:

- The model shows an adjusted R-squared value of **0.688683682**, indicating that approximately **68.9%** of the variability in the AVG_PRICE can be explained by the significant variables included in the model.
- Among the significant variables, **NOX, PTRATIO, LSTAT, TAX, AGE, INDUS, DISTANCE**, and **AVG_ROOM** have coefficients that indicate their impact on the **AVG_PRICE**. The intercept value is **29.42847349**.

b) Comparison of Adjusted R-square:

The adjusted R-square value for this model is **0.688683682**. Comparing this with the previous model's adjusted R-square value of **0.688298647**, this model performs slightly better in explaining the variation in the **AVG_PRICE**.

c) Sorting of Coefficients:

Sorting the coefficients in ascending order:

- NOX: -10.27270508
- PTRATIO: -1.071702473
- LSTAT: -0.605159282
- TAX: -0.014452345
- AGE: 0.03293496
- INDUS: 0.130710007
- DISTANCE: 0.261506423
- AVG_ROOM: 4.125468959

If the value of **NOX** increases in a locality in this town, based on the negative coefficient, the **AVG_PRICE** is expected to decrease. This implies that an increase in **NOX** levels is associated with a decrease in the average house price in the locality.

d) Regression Equation:

The regression equation based on the model is as follows:

y=m1x1+m2x2+m2x2+m3x3+m4x4+m5x5+6x6+m7x7+c

Thank you!