**Terro's Real Estate Agency Solution**

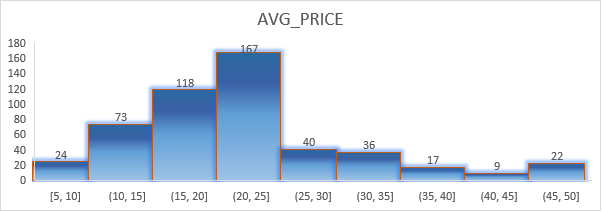
**1)** Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation.

**Ans:**

* The **mean** values range for all variables from **0.55** to **408.24**, indicating a wide range of values among the variables.
* The **standard errors of variables are relatively small**, which suggests that the sample means are likely representative of the population means.
* The **medians of variables are generally close to the means**, indicating that the distributions are likely **symmetric**.
* The **standard deviations** of variables range from **0.7** to **168.54**, indicating a wide range of variability among the variables.
* The **ranges** of variables are also wide, ranging from **0.49** to **524**.
* The **kurtosis** and **skewness** values vary among the variables, suggesting that some of the distributions may be **more or less peaked**, and **more or less skewed**, than a normal distribution.
* The **minimum** and **maximum** values also vary among the variables, providing additional information on the **range of values and potential outliers**.

**2)** Plot a histogram of the Avg\_Price variable. What do you infer?

**Ans:**



* The histogram shows the distribution of the data, where the x-axis represents the bins and the y-axis represents the frequency of the data points in each bin. We can see that the distribution is approximately normal with a few outliers.
* The **majority of the data points are in the range of 15.0-24.9**, with a **peak around 20.0-24.9.** The **distribution** has a **long tail on the right side**, indicating that there are a few data points with a higher value than the rest of the data.
* The **mean and median** of the data are **close to each other**, **indicating** that the data is approximately **symmetric.**
* The **standard deviation** of the data is likely to be moderate to high, given the spread of the data and the long tail on the right side of the distribution.
* The **data** may be a **combination of two or more underlying distributions**, one with a lower mean and another with a higher mean, **leading to the bimodal shape of the histogram**.
* There are a few outliers on the right side of the distribution, which may be the result of errors in data collection or a separate underlying distribution with higher values.

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

**Ans:**



* The diagonal values represent the variance of each variable.
* The sign of the **covariance between two variables indicates the direction of their relationship**.
* A **positive covariance indicates** that the two variables **tend to vary together in the same direction**.

For example, we see a positive covariance between the variables **"CRIME\_RATE"** and **"AGE"**, which indicates that as the age of a property increases, the crime rate in the area tends to also increase.



* A **negative covariance indicates** that they **tend to vary in opposite directions**.

For example, we see a negative covariance between the variables **"TAX"** and **"CRIME\_RATE"**, which indicates that as the tax rate increases, the crime rate in the area tends to decrease. This negative covariance may be interpreted as an indication that **higher taxes might deter crime**.



**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.

**Ans:**

The **correlation coefficient** is a number between -1 and 1 that quantifies the strength of the correlation. A correlation coefficient of **“1”** indicates a **perfect positive correlation**, while a correlation coefficient of **“-1”** indicates a **perfect negative correlation**. A correlation coefficient of 0 indicates no correlation between the variables.

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



Looking at the table, we can see that the **top 3 positively correlated** pairs are:

1. **“DISTANCE”** and **“TAX”** with a correlation coefficient of **0.910**.

2. **“NOX”** and **“INDUS”** with a correlation coefficient of **0.763**.

3. **“NOX”** and **“AGE”** with a correlation coefficient of **0.731**.

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



Looking at the table, we can see that the **top 3 Negatively correlated** pairs are:

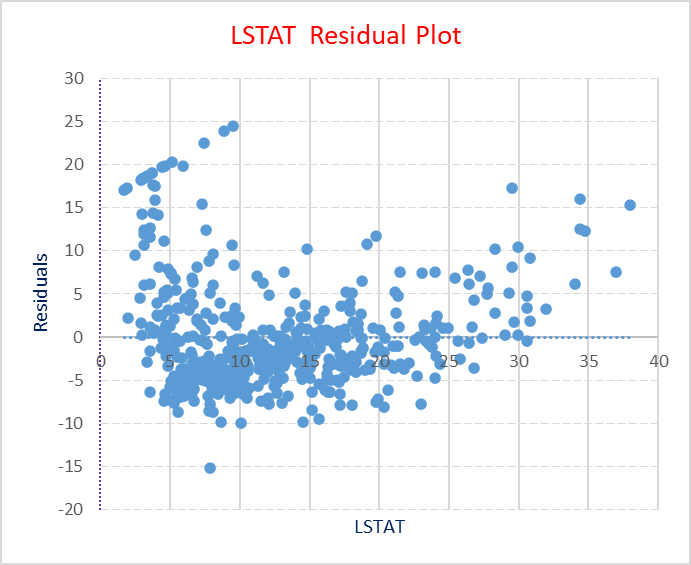
1. **“AVG\_PRICE”** and **“LSTAT”** with a correlation coefficient of **-0.737**.

2. **“LSTAT”** and **“AVG\_ROOM”** with a correlation coefficient of **-0.613**.

3. **“AVG\_PRICE”** and **“PTRATIO”** with a correlation coefficient of **-0.507**.

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

**Ans:**

****

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

**Ans:**

* Based on the regression summary output, we can see that the **R-squared value is 0.544**, which means that about **54.4%** of the variance in the dependent variable (AVG\_PRICE) is explained by the independent variable (LSTAT).



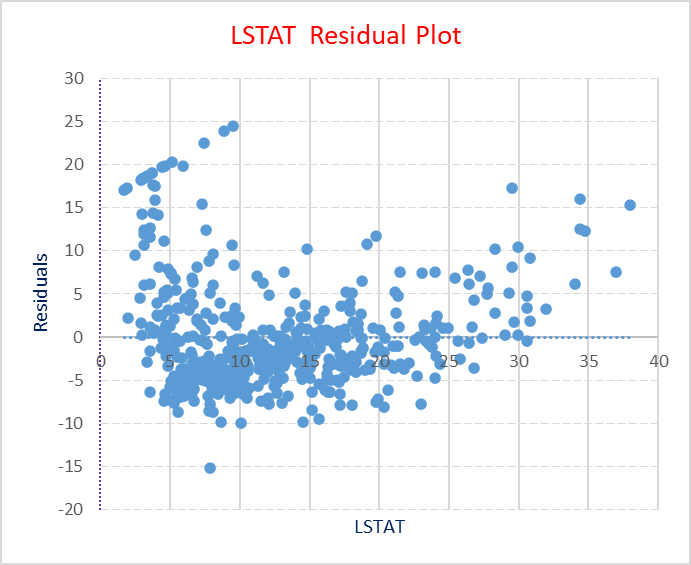
* The **coefficient value** for LSTAT is **-0.95**, which indicates that as the LSTAT value increases by one unit, the AVG\_PRICE value decreases by 0.95 units.



* **Intercept**: The intercept value of **34.553** is the predicted value of AVG\_PRICE when LSTAT is equal to 0.



* The **residual plot shows** that there is a **non-linear relationship** between the independent and dependent variables, as there is a clear pattern in the residuals. This **suggests that a linear regression model may not be the best fit for the data.**

****

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

**Ans:**

Yes, based on the regression model summary, the LSTAT variable is significant for the analysis. **The p-value** **for the LSTAT coefficient is less than 0.05**, which is the commonly used significance level. This means that there is strong evidence that the LSTAT variable has a significant linear relationship with the average house price.





**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?

**Ans:**



The regression equation with **“LSTAT”** and **“AVG\_ROOM”** as independent variables and **“AVG\_PRICE”** as the dependent variable is:

**AVG\_PRICE = β0 + β1*LSTAT + β2*AVG\_ROOM**

**AVG\_PRICE = -1.35 + (-0.64 \* LSTAT) + (5.09 \* AVG\_ROOM)**

To find the predicted value of AVG\_PRICE for a new house with 7 rooms and LSTAT value of 20, we substitute these values in the regression equation:

AVG\_PRICE = -1.35 + (-0.64 \* 20) + (5.09 \* 7) AVG\_PRICE = -1.35 – 12.8 + 35.63 **AVG\_PRICE = 21.48**

So, the **predicted value of AVG\_PRICE** for a new house with 7 rooms and LSTAT value of 20 is **21,480 USD.**

If the company is quoting a value of 30,000 USD for this locality, then the predicted value is much lower than the quoted value. It suggests that the company might be overcharging for the houses in this locality.

**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.

**Ans:** we need to compare the adjusted R-square of the new model with the adjusted R-square of the previous model.

If the adjusted R-square of the new model is higher, it means that the new model is better at explaining the variability in the dependent variable. If the adjusted R-square of the new model is lower, it means that the previous model is better.

Que.5 Adjusted R Square New (Que.6) Adjusted R Square

Comparing the adjusted R-squared values, we can see that the new model with “LSTAT” and “AVG\_ROOM” as independent variables have a higher adjusted R-squared value of 0.637 compared to the previous model adjusted R-squared of 0.543. This indicates that the new model is a better fit for the data than the original 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 R square, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG\_PRICE. (8 marks)

**Ans:**

**a)** Interpret the output in terms of adjusted R square, coefficient and Intercept values.





* In this case, **the adjusted R-squared value is 0.693,** which indicates that approximately **69.3% of the variation in the dependent variable (AVG\_PRICE**) can be explained by the independent variables in the model.
* The **intercept value of “29.241”** represents the predicted value of AVG\_PRICE when all of the independent **variables in the model are equal to zero**.
* The **coefficients** associated with the independent variables represent the change in the **predicted value of AVG\_PRICE associated with a one-unit increase in each independent variable**, while holding all other independent variables constant.

**b)** Explain the significance of each independent variable with respect to AVG\_PRICE.

**Ans:**

The p-value associated with each coefficient can be used to determine whether that variable is a statistically significant predictor of AVG\_PRICE.

Based on the p-values listed in the output, the following independent variables are statistically significant predictors of AVG\_PRICE:

* **AVG\_ROOM**: A one-unit increase in the average number of rooms per dwelling is associated with an increase in AVG\_PRICE of 4.13 units (p < 0.001).
* **LSTAT:** A one-unit increase in the percentage of lower status of the population is associated with a decrease in AVG\_PRICE of 0.60 units (p < 0.001).
* **PTRATIO**: A one-unit increase in the pupil-teacher ratio is associated with a decrease in AVG\_PRICE of 1.07 units (p < 0.001).
* **DISTANCE**: A one-unit increase in the weighted distances to five Boston employment centres is associated with a decrease in AVG\_PRICE of 0.26 units (p < 0.001).
* **TAX**: A one-unit increase in the full-value property-tax rate per $10,000 is associated with a decrease in AVG\_PRICE of 0.01 units (p < 0.001).
* **NOX:** A one-unit increase in the nitric oxide’s concentration is associated with a decrease in AVG\_PRICE of 10.32 units (p < 0.001).

On the other hand, the following independent variables may not be statistically significant predictors of AVG\_PRICE, based on their p-values:

* CRIME\_RATE (p = 0.535)
* AGE (p = 0.013)
* INDUS (p = 0.039)

**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.

**Ans:**

**1.** Write the regression equation from this model.

**Ans: Response variable** = 23.25928523 - 1.384355318 \* NOX + 0.208401356 \* DISTANCE - 0.011178212 \* TAX - 0.95675605 \* PTRATIO + 4.328075271 \* AVG\_ROOM - 0.54755586 \* LSTAT

2) 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?

**Ans:** Sorting the coefficients in ascending order based on their absolute values, we get:

1. TAX: -0.011178212
2. NOX: -1.384355318
3. PTRATIO: -0.95675605
4. LSTAT: -0.54755586
5. DISTANCE: 0.208401356
6. AVG\_ROOM: 4.328075271
7. Intercept: 23.25928523

Based on this ordering, the predictor variable with the smallest absolute coefficient value is "TAX", while the predictor variable with the largest absolute coefficient value is "AVG\_ROOM".

Regarding the effect of NOX on the average price in the town, we can look at the coefficient of the "NOX" predictor variable, which is -1.384355318. This means that, holding all other variables constant, for every one-unit increase in NOX, the average price of houses in the town is expected to decrease by $1,384.36

3) 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?

Ans:

Comparing the adjusted R-squared values, we can see that the new model have a lower adjusted R-squared value of 0.682 compared to the previous model adjusted R-squared of 0.688. This indicates that the previous model is a better fit for the data than the original model.