TERRO'S REAL ESTATE AGENCY
BUSINESS REPORT

Prepared by: G Khushal Sai
Date: 15th Jan 2023

OBSERVATIONS:

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

Ans: For Summary statistics refer to the excel workbook, the mean, median, mode tells the measure of location.

i. In CRIME_RATE, the variance is 8.53 tells us how far the observations are from the average. The skewness value is 0.021 and it is said to be "Fairly symmetrical". As the kurtosis value is negative, then distribution is called "Platykurtic".

ii. In AGE, the variance is 792.35 tells us how far the observations are from the average. The skewness value is -0.598 and it is said to be "Negatively Skewed". As the kurtosis value is negative, then distribution is called "Platykurtic".

iii. In INDUS, the variance is 47.06 tells us how far the observations are from the average. The skewness value is 0.29 and it is said to be "Positively Skewed". As the kurtosis value is negative, then distribution is called "Platykurtic".

iv. In NOX, the variance is 0.013 tells us how far the observations are from the average. The skewness value is 0.72 and it is said to be "Positively Skewed". As the kurtosis value is negative, then distribution is called "Platykurtic".

v. In DISTANCE, the variance is 75.81 tells us how far the observations are from the average. The skewness value is 1.00 and it is said to be "Positively Skewed". As the kurtosis value is negative, then distribution is called "Platykurtic".

vi. In TAX, the variance is 28404.75 tells us how far the observations are from the average. The skewness value is 0.66 and it is said to be "Positively Skewed". As the kurtosis value is negative, then distribution is called "Platykurtic".

vii. In PTRATIO, the variance is 4.68 tells us how far the observations are from the average. The skewness value is -0.80 and it is said to be "Negatively Skewed". As the kurtosis value is negative, then distribution is called "Platykurtic".

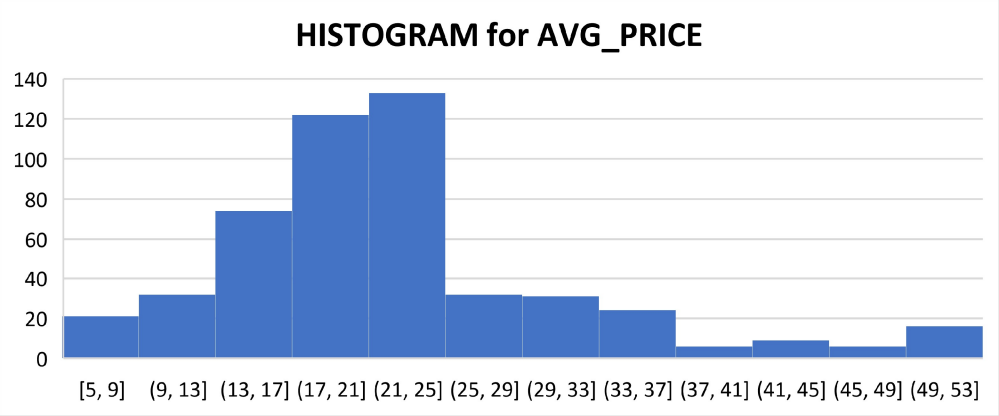
viii. In AVG_ROOM, the variance is 0.49 tells us how far the observations are from the average. The skewness value is 0.40 and it is said to be "Fairly Symmetrical". As the kurtosis value is positive, then distribution is called "Leptokurtic".

ix. In, LSTAT, the variance is 50.99 tells us how far the observations are from the average. The skewness value is 0.90 and it is said to be "Positive Skewed". As the kurtosis value is positive, then distribution is called "Leptokurtic".

x. In AVG_PRICE, the variance is 84.58 tells us how far the observations are from the average. The skewness value is 1.10 and it is said to be "Positive Skewed". As the kurtosis value is positive, then distribution is called "Leptokurtic".

2. Plot a histogram of the AVG_PRICE variable. What do you infer? (5 marks)

Ans:



From the above histogram plotted, I can infer that in the interval 21 to 25 there are more data points and the shape of the distribution is "Skewed Right". It is observed as "Positive Skewness".

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

Ans: Refer to the excel workbook for covariance matrix. As we know, Covariance tells the direction.

- if X and Y are mostly both above or below their average then it is ” Positive”.
- if X and Y are mostly on opposite of their average then it is “Negative”.
- From the excel workbook, in the matrix the diagonal values are the variance of their respective variables.
- The green highlighted cells show that the covariance is positive and there is a “Positive relationship” between two variables.
- The Red highlighted cells shows that the covariance is negative and there is a “Negative relationship” between two variables.

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: Refer to the excel workbook for correlation matrix.
From the Excel workbook**,**

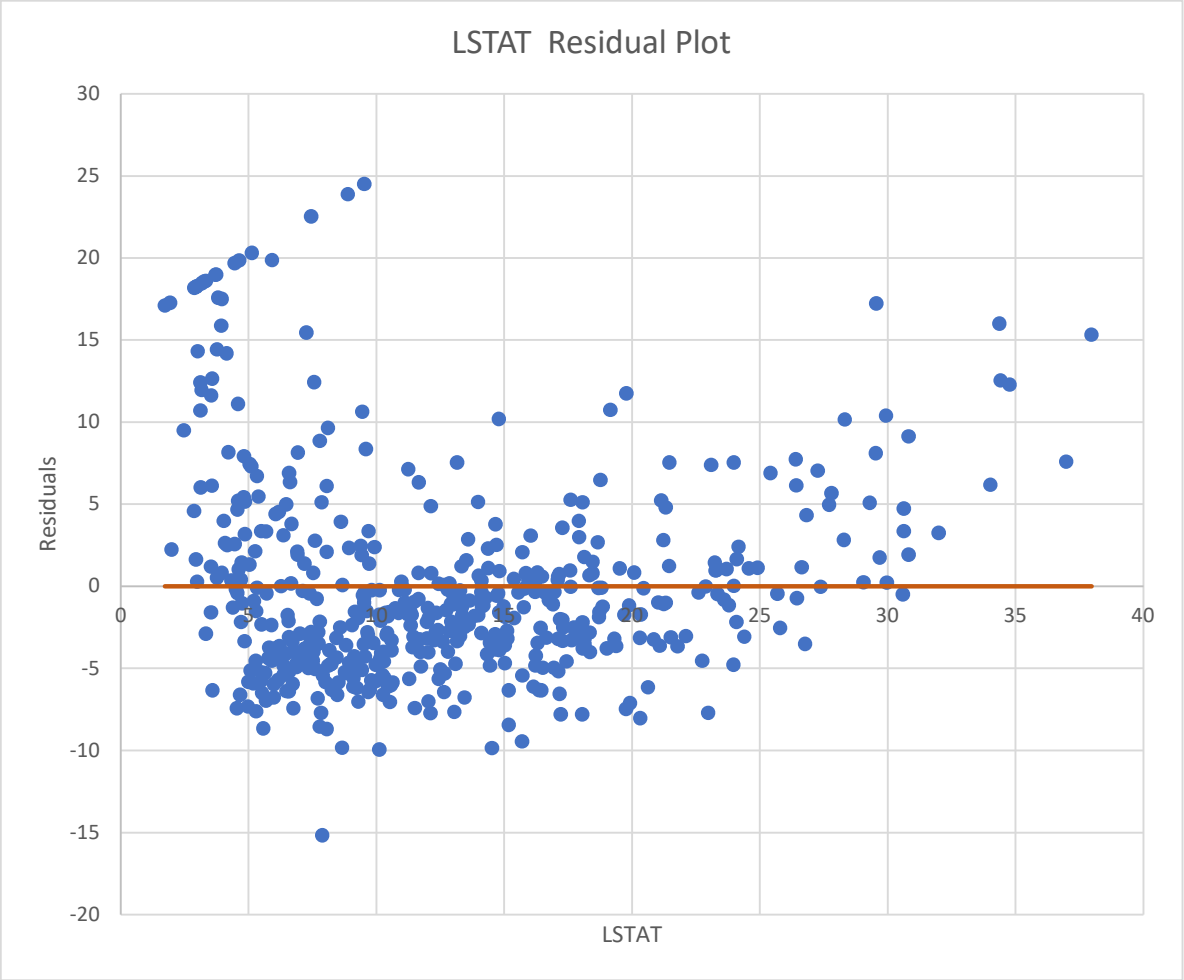
Top Positively Correlated Pairs	**Top Negatively Correlated Pairs**
TAX and DISTANCE	INDUS and CRIME_RATE
NOX and INDUS	DISTANCE and CRIME_RATE
NOX and AGE	TAX and CRIME_RATE

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?

Ans: Refer to the excel workbook.
From Regression Summary output,

a) 54.4% of the variance of AVG_PRICE values around the mean are explained by the LSTAT values. This means 54.4% of the values fit the model.

- As the coefficient value of LSTAT value is Negative it tells that, decrease in 'X' variable (LSTAT) there is an increase in 'Y' variable (AVG_PRICE).
- The intercept value is the constant value is a point where the function crosses the y axis.
- From the residual plot we can observe that the data points are random in nature. There is no linear relationship between the two variables.
- The following chart shows the residual plot,



b) From the summary output, p-value for LSTAT is less than 0.05 which makes the variable significant for the analysis.

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.

Ans: Refer to the excel workbook,

a) Multi Linear Regression Equation:

Given, x1 =7, x2 = 20

Equation = B0+(B1*x1) +(B2+x2)

Here,

- B0 is the Intercept value
- B1and B2 are the coefficients of the 'X' variables which we will get from the summary output
- x1 and x2 are the averages (given in the question).

Now substituting the values in the equation,

= -1.35827281187456+(5.09478798433655*7) + (-0.642358334244129*20)

=> AVG_PRICE = 21.45807639 (in $1000's)

By comparing the AVG_PRICE with the company quoting a value of 30000 USD for this locality, the company is overcharging.

b) The adjusted R square value for previous model is 0.543241825 and the adjusted R square value for this model is 0.63712447547 by comparing the values we can say that this model performance is better than the previous model because the number of variables is more i.e., two variable (AVG_ROOM and LSTAT) and it tends to increases the model accuracy.

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: From the model generated in excel workbook, I can infer that 68.8% of variance of AVG_PRICE values are explained by all independent variables**.**

- In the coefficient column it can be inferred that the variables CRIME_RATE, AGE, INDUS, DISTANCE and AVG_ROOM are positive and it states that if there is an increase in independent variable(X) then there is an increase in dependent variable (AVG_PRICE).
- The variables NOX, TAX, PTRATIO and LSTAT are negative and states that if there is an increase in independent variable(X) there is a decrease in dependent variable (AVG_PRICE).
- The intercept value is constant value and it is a point where the function crosses the y axis.
- CRIME_RATE with respect to AVG_PRICE, the p-value is 0.53 which is greater than 0.05 and this variable is not statistically significant with AVG_PRICE.

- AGE with respect to AVG_PRICE, the p-value is 0.012 which is lesser than 0.05 and this variable is significant with AVG_PRICE.
- INDUS with respect to AVG_PRICE, the p-value is 0.093 which is lesser than 0.05 and this variable is significant with AVG_PRICE.
- NOX with respect to AVG_PRICE, the p-value is 0.0082 which is lesser than 0.05 and this variable is significant with AVG_PRICE.
- DISTANCE with respect to AVG_PRICE, the p-value is 0.00013 which is lesser than 0.05 and this variable is significant with AVG_PRICE.
- PTRATIO with respect to AVG_PRICE, the p-value is 6.5E-15 which is lesser than 0.05 and this variable is significant with AVG_PRICE.
- LSTAT with respect to AVG_PRICE, the p-value is 8.910E-27 which is lesser than 0.05 and this variable is significant with AVG_PRICE.

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: Refer to the excel workbook,

a)

- From the previous model we would consider removing "Crime Rate" because the value of p is not less than the usual significant value 0.05. Keeping variables that are not statistically significant can reduce the model's precision. From the above created model we can observe that the variables AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG_ROOM, LSTAT are statistically significant variables for the dependent variable "AVG_PRICE".
- As the p-value for the above variables are less than the significance level, we can say that the data is providing enough evidence to reject the null hypothesis for this model.
- Thus, I can infer that in spite of having positive and negative values in coefficient column for the variables, states that based on p-value these are statistically significant with dependent variable "AVG_PRICE".

b)

	Q7 Regression Model	Q8 Regression Model
Variables Used	CRIME RATE, AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG ROOM, LSTAT	AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG ROOM, LSTAT
Adjusted R Square	0.688298647 = 68.82%	0.688683682 = 68.86%

- Q8 Model performs better than the previous model because the value of adjusted R square is more compared to the value of previous model.

c) As we know, A positive coefficient indicates that as the value of the independent variable increases, then the dependent variable also tends to increase. A negative coefficient suggests that as the independent variable increases, the dependent variable tends to decrease. After sorting the coefficients in ascending order, if the "NOX" is more, then the "AVG_PRICE" decreases.

d) Regression: It is a Multi Linear Regression

Equation=

B0+(B1*X1) +(B2*X2)+(B3*X3)+(B4*X4)+(B5*X5)+(B6*X6)+(B7*X7)+(B8*X8)

Here,

- B0 is the Intercept Value
- B1, B2, B3, B4, B5, B6, B7, B8 are the Coefficients of the Independent Variables
- X1, X2, X3, X4, X5, X6, X7, X8 are the values of independent variables