Terro’s Real Estate Agency

# Business Analysis Report



Tadimarri Mahammad Aqib

INDEX

# S.No Content Page

# 1 Problem Statement 3

# 2 Data Dictionary 3

# 3 Objective 3

# 4 Questions 4

# 1) Descriptive Statistics 4

# 2) Histogram 9

# 3) Covariance Matrix 10

# 4) Correlation Matrix 11

# 5) Initial Linear Regression 12

# 6) Regression Model 14

# 

# 7) Regression Model 15

# 8) regression Model with Significant 17

# values

# 

Real estate data analysis : Exploratory data analysis, Linear Regression

**1 .Problem Statement (Situation):**

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.

**2 .Data Dictionary:**

**Attribute Description**

* **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** : pupil-teacher ratio by town
* **LSTAT** : % lower status of the population
* **AVG**\_**PRICE** : Average value of houses in $1000's

**3 .Objective (Task):**

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

**4 .Questions:**

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

The summary statistics tells us various measures i.e., mean, median, standard deviation, skewness etc. that are calculated based on the data points present in the variable. It can be helped to understand how the data points are distributed.

|  |  |
| --- | --- |
| ***CRIME\_RATE*** | |
|  |  |
| **Mean** | 4.871976285 |
| **Standard Error** | 0.129860152 |
| **Median** | 4.82 |
| **Mode** | 3.43 |
| **Standard Deviation** | 2.921131892 |
| **Sample Variance** | 8.533011532 |
| **Kurtosis** | -1.189122464 |
| **Skewness** | 0.021728079 |
| **Range** | 9.95 |
| **Minimum** | 0.04 |
| **Maximum** | 9.99 |
| **Sum** | 2465.22 |
| **Count** | 506 |

CRIME RATE\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The Crime rate variable contains data whose mean is 4.87 and median is 4.82 which is almost similar so we can assume that the data points do not contain any outliers and the most frequent data point is 3.43 as shown in mode. The above measures tell about the measures of central tendency and now lets see about the measures of dispersion which tell how the values as dispersed from the centre. The Dispersion measures are Standard Deviation, Range, Variance and their values are 2.92, 9.95 and 8.53 respectively.

There are two measures which tell us about the symmetry of the data points when plotted in a graph those are called measures of symmetry, they are Kurtosis and Skewness whose values are -1.189 and 0.021 respectively. The Kurtosis tells us the frequency distribution of values in the graph, if there is a value with large frequency compared to other values it is depicted with sharp peak and it is known as Leptokurtic and outliers are highly frequent in such dataset. If the dataset has infrequent outliers and the distribution is thin tailed then it is called Platykurtic. If the distribution is medium tailed and outliers are neither highly frequent nor highly infrequent then it is Mesokurtic.

The kurtosis is in negative so it is Platykurtic. Skewness tells us the symmetry of the distribution is it to left or right it can be zero also. Since crime rate skewness is near to zero so it can be said zero skewed or approximately symmetric. Descriptive Statistics also tells us the Maximum, Minimum, Sum and Count of the dataset.

|  |  |
| --- | --- |
| ***AGE*** | |
|  |  |
| **Mean** | 68.57490119 |
| **Standard Error** | 1.251369525 |
| **Median** | 77.5 |
| **Mode** | 100 |
| **Standard Deviation** | 28.14886141 |
| **Sample Variance** | 792.3583985 |
| **Kurtosis** | -0.967715594 |
| **Skewness** | -0.59896264 |
| **Range** | 97.1 |
| **Minimum** | 2.9 |
| **Maximum** | 100 |
| **Sum** | 34698.9 |
| **Count** | 506 |

AGE

The Age Datasets measure of central tendency i.e., Mean is 68.57, Median is 77.5 and Mode is 100. We can infer that mean is less than median so there is a chance of outliers being present. The mid value in the dataset is 77.5, The data value with most frequency is 100.

The Measures of Dispersion i.e., Range, Standard Deviation, Variance are 97.1, 28.15, and 792.358 respectively. Maximum is 100 and Minimum is 2.9 so the difference between Max and Min is the range. Variance is the sum of squares of each data points distance from the mean by total number of variables. Standard deviation is the root of variance.

The Measures of Symmetry i.e., Skewness and Kurtosis are -0.59 and -0.96 respectively. Since skewness is between -0.5 and -1 so its moderately skewed and kurtosis is less than 3 so its Platykurtic. The Sum and count of the Age dataset are 34698.9 and 506 respectively.

|  |  |
| --- | --- |
| ***INDUS*** | |
|  |  |
| **Mean** | 11.13677866 |
| **Standard Error** | 0.304979888 |
| **Median** | 9.69 |
| **Mode** | 18.1 |
| **Standard Deviation** | 6.860352941 |
| **Sample Variance** | 47.06444247 |
| **Kurtosis** | -1.233539601 |
| **Skewness** | 0.295021568 |
| **Range** | 27.28 |
| **Minimum** | 0.46 |
| **Maximum** | 27.74 |
| **Sum** | 5635.21 |
| **Count** | 506 |

INDUS

The INDUS Datasets measure of central tendency i.e., Mean is 11.136, Median is 9.69 and Mode is 18.1. We can infer that mean is higher than median so there is a chance of outliers being present. The mid value in the dataset is 9.69. The data value with most frequency is 18.1.

The Measures of Dispersion i.e., Range, Standard Deviation, Variance are 27.28, 6.86, and 47.06 respectively. Maximum is 27.74 and Minimum is 0.46 so the difference between max and min is the range. Variance is the sum of squares of each data points distance from the mean by total number of variables. Standard deviation is the root of variance.

The Measures of Symmetry i.e., Skewness and Kurtosis are 0.295 and -1.233 respectively. Since skewness is between -0.5 and 0.5 so it’s approximately symmetric and kurtosis is less than 3 so it’s Platykurtic. The Sum and count of the INDUS dataset are 5635.21 and 506 respectively.

|  |  |
| --- | --- |
| ***NOX*** | |
|  |  |
| **Mean** | 0.554695059 |
| **Standard Error** | 0.005151391 |
| **Median** | 0.538 |
| **Mode** | 0.538 |
| **Standard Deviation** | 0.115877676 |
| **Sample Variance** | 0.013427636 |
| **Kurtosis** | -0.064667133 |
| **Skewness** | 0.729307923 |
| **Range** | 0.486 |
| **Minimum** | 0.385 |
| **Maximum** | 0.871 |
| **Sum** | 280.6757 |
| **Count** | 506 |

NOX

The NOX Datasets measure of central tendency i.e., Mean is 0.55, Median is 0.538 and Mode is 0.538. We can infer that mean is almost same as median so there is a less chance of outliers being present. The mid value in the dataset is 0.538. The data value with most frequency is 0.538. Median, Mode and Mean are almost same so the data form a symmetric frequency distribution.

The Measures of Dispersion i.e., Range, Standard Deviation, Variance are 0.486, 0.116, and 0.013 respectively. Maximum is 0.871 and Minimum is 0.385 so the difference between max and min is the range. Variance is the sum of squares of each data points distance from the mean by total number of variables. Standard deviation is the root of variance.

The Measures of Symmetry i.e., Skewness and Kurtosis are 0.729 and -0.065 respectively. Since skewness is between 0.5 and 1 so it’s moderately skewed and kurtosis is less than 3 so it’s Platykurtic. The Sum and count of the NOX dataset are 280.68 and 506 respectively.

|  |  |
| --- | --- |
| ***DISTANCE*** | |
|  |  |
| **Mean** | 9.549407115 |
| **Standard Error** | 0.387084894 |
| **Median** | 5 |
| **Mode** | 24 |
| **Standard Deviation** | 8.707259384 |
| **Sample Variance** | 75.81636598 |
| **Kurtosis** | -0.867231994 |
| **Skewness** | 1.004814648 |
| **Range** | 23 |
| **Minimum** | 1 |
| **Maximum** | 24 |
| **Sum** | 4832 |
| **Count** | 506 |

DISTANCE

The Distance Datasets measure of central tendency i.e., Mean is 9.549, Median is 5 and Mode is 24. We can infer that mean is higher than median so there is a chance of outliers being present. The mid value in the dataset is 5 and most of the data points lie near median. The data value with most frequency is 24.

The Measures of Dispersion i.e., Range, Standard Deviation, Variance are 23, 8.707 and 75.816 respectively. Maximum is 24 and Minimum is 1 so the difference between max and min is the range. Variance is the sum of squares of each data points distance from the mean by total number of variables. Standard deviation is the root of variance.

The Measures of Symmetry i.e., Skewness and Kurtosis are 1.005 and -0.867 respectively. Since skewness is above 1 so it’s highly skewed and kurtosis is less than 3 so it’s Platykurtic. The Sum and count of the Distance dataset are 4832 and 506 respectively.

|  |  |
| --- | --- |
| ***TAX*** | |
|  |  |
| **Mean** | 408.2371542 |
| **Standard Error** | 7.492388692 |
| **Median** | 330 |
| **Mode** | 666 |
| **Standard Deviation** | 168.5371161 |
| **Sample Variance** | 28404.75949 |
| **Kurtosis** | -1.142407992 |
| **Skewness** | 0.669955942 |
| **Range** | 524 |
| **Minimum** | 187 |
| **Maximum** | 711 |
| **Sum** | 206568 |
| **Count** | 506 |

TAX

The TAX Datasets measure of central tendency i.e., Mean is 408.237, Median is 330 and Mode is 666. We can infer that mean is higher than median so there is a chance of outliers being present. The mid value in the dataset is 330 and most of the data points lie near median. The data value with most frequency is 666.

The Measures of Dispersion i.e., Range, Standard Deviation, Variance are 524, 168.537 and 28404.76 respectively. Maximum is 711 and Minimum is 187 so the difference between max and min is the range. Variance is the sum of squares of each data points distance from the mean by total number of variables. Standard deviation is the root of variance.

The Measures of Symmetry i.e., Skewness and Kurtosis are 0.6699 and -1.142 respectively. Since skewness is between 0.5 and 1 so it’s moderately right skewed which is also known as Positively Skewed and kurtosis is less than 3 so it’s Platykurtic. The Sum and count of the TAX dataset are 206568 and 506 respectively.

|  |  |
| --- | --- |
| ***PTRATIO*** | |
|  |  |
| **Mean** | 18.4555336 |
| **Standard Error** | 0.096243568 |
| **Median** | 19.05 |
| **Mode** | 20.2 |
| **Standard Deviation** | 2.164945524 |
| **Sample Variance** | 4.686989121 |
| **Kurtosis** | -0.285091383 |
| **Skewness** | -0.802324927 |
| **Range** | 9.4 |
| **Minimum** | 12.6 |
| **Maximum** | 22 |
| **Sum** | 9338.5 |
| **Count** | 506 |

PTRATIO

The PTRATIO Datasets measure of central tendency i.e., Mean is 18.45, Median is 19.05 and Mode is 20.2. We can infer that mean is only slightly less than median so there is a less chance of outliers being present. The mid value in the dataset is 19.05 and most of the data points lie near median. The data value with most frequency is 20.2.

The Measures of Dispersion i.e., Range, Standard Deviation, Variance are 9.4, 2.165 and 4.687 respectively. Maximum is 22 and Minimum is 12.6 so the difference between max and min is the range. Variance is the sum of squares of each data points distance from the mean by total number of variables. Standard deviation is the root of variance.

The Measures of Symmetry i.e., Skewness and Kurtosis are -0.802 and -0.285 respectively. Since skewness is between -0.5 and -1 so it’s moderately skewed and kurtosis is less than 3 so it’s Platykurtic. The Sum and count of the PTRATIO dataset are 9338.5 and 506 respectively.

|  |  |
| --- | --- |
| ***AVG\_ROOM*** | |
|  |  |
| **Mean** | 6.284634387 |
| **Standard Error** | 0.031235142 |
| **Median** | 6.2085 |
| **Mode** | 5.713 |
| **Standard Deviation** | 0.702617143 |
| **Sample Variance** | 0.49367085 |
| **Kurtosis** | 1.891500366 |
| **Skewness** | 0.403612133 |
| **Range** | 5.219 |
| **Minimum** | 3.561 |
| **Maximum** | 8.78 |
| **Sum** | 3180.025 |
| **Count** | 506 |

AVG\_ROOM

The Avg\_Room Datasets measure of central tendency i.e., Mean is 6.284, Median is 6.208 and Mode is 5.713. We can infer that mean is almost same as median so there is a less chance of outliers being present. The mid value in the dataset is 6.208 and most of the data points lie near median. The data value with most frequency is 5.713.

The Measures of Dispersion i.e., Range, Standard Deviation, Variance are 5.219, 0.703 and 0.494 respectively. Maximum is 8.78 and Minimum is 3.561 so the difference between max and min is the range. Variance is the sum of squares of each data points distance from the mean by total number of variables. Standard deviation is the root of variance.

The Measures of Symmetry i.e., Skewness and Kurtosis are 0.404 and 1.89 respectively. Since skewness is between 0.5 and 1 so it’s moderately skewed and kurtosis is less than 3 so it’s Platykurtic. The Sum and count of the Avg\_Room dataset are 3180.025 and 506 respectively.

LSTAT

|  |  |
| --- | --- |
| ***LSTAT*** | |
|  |  |
| **Mean** | 12.65306324 |
| **Standard Error** | 0.317458906 |
| **Median** | 11.36 |
| **Mode** | 8.05 |
| **Standard Deviation** | 7.141061511 |
| **Sample Variance** | 50.99475951 |
| **Kurtosis** | 0.493239517 |
| **Skewness** | 0.906460094 |
| **Range** | 36.24 |
| **Minimum** | 1.73 |
| **Maximum** | 37.97 |
| **Sum** | 6402.45 |
| **Count** | 506 |

The LSTAT Datasets measure of central tendency i.e., Mean is 12.653, Median is 11.36 and Mode is 8.05. We can infer that mean is slightly greater than median so there is a less chance of outliers being present. The mid value in the dataset is 11.36 and most of the data points lie near median. The data value with most frequency is 8.05.

The Measures of Dispersion i.e., Range, Standard Deviation, Variance are 36.24, 7.141 and 50.995 respectively. Maximum is 37.97 and Minimum is 1.73 so the difference between max and min is the range. Variance is the sum of squares of each data points distance from the mean by total number of variables. Standard deviation is the root of variance.

The Measures of Symmetry i.e., Skewness and Kurtosis are 0.906 and 0.493 respectively. Since skewness is between 0.5 and 1 so it’s moderately skewed and kurtosis is less than 3 so it’s Platykurtic. The Sum and count of the LSTAT dataset are 6402.45 and 506 respectively.

|  |  |
| --- | --- |
| ***AVG\_PRICE*** | |
|  |  |
| **Mean** | 22.53280632 |
| **Standard Error** | 0.408861147 |
| **Median** | 21.2 |
| **Mode** | 50 |
| **Standard Deviation** | 9.197104087 |
| **Sample Variance** | 84.58672359 |
| **Kurtosis** | 1.495196944 |
| **Skewness** | 1.108098408 |
| **Range** | 45 |
| **Minimum** | 5 |
| **Maximum** | 50 |
| **Sum** | 11401.6 |
| **Count** | 506 |

AVG\_PRICE

The Avg\_Price Datasets measure of central tendency i.e., Mean is 22.53, Median is 21.2 and Mode is 50. We can infer that mean is slightly greater than median so there is a less chance of outliers being present. The mid value in the dataset is 21.2 and most of the data points lie near median. The data value with most frequency is 50.

The Measures of Dispersion i.e., Range, Standard Deviation, Variance are 45, 9.197 and 84.586 respectively. Maximum is 50 and Minimum is 5 so the difference between max and min is the range. Variance is the sum of squares of each data points distance from the mean by total number of variables. Standard deviation is the root of variance.

The Measures of Symmetry i.e., Skewness and Kurtosis are 1.108 and 1.495 respectively. Since skewness is above so it’s highly skewed and kurtosis is less than 3 so it’s Platykurtic. The Sum and count of the AVG\_PRICE dataset are 11401.6 and 506 respectively.

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

The Observations we can infer from the above Histogram is the frequency of the Avg\_Price. It depicts the data values present in the Avg\_Price variable which are dispersed from 5K USD to 45K USD. The Data values are present in different range bins and the range with more number of data points is 18K - 22.5K. The histogram depicts that there are no outliers present in the data points. The Avg\_Price of the houses are mostly between 13.5K to 27K. Median and mean are also present in this range. The graph is slightly positively skewed because the symmetry is falling to the left. The Avg\_Price is Leptokurtic as the graph shows a sharp peak in the graph.

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

The Covariance Matrix tells us about joint dispersion of two variables i.e., product of summation of distances of data values of both variables from their mean by total number of data values. We can check the covariance of any two variables from the matrix. The Covariance tells us the direction.

The Matrix is symmetrical and the diagonal values are always the variance of the variable. In the matrix if the x and y values are both above average then the value will be +ve. If x and y are mostly on opposite sides of their averages then covariance is –ve.

We can see that the diagonal values are all the variances of the variables.

**Covariance.s (array1, array2) == Variance.s (array)**

There are both positive and negative values in the matrix .The Positive Values i.e., Avg\_Price & Crime\_Rate etc., which tells us that the positives are values whose x and y values are above their mean , whereas The Negatives i.e., Avg\_Price & Age etc., have their x and y in opposite sides of their average. Since the matrix is symmetric so only lower triangle will be filled with values and the above values are the mirror image of the lower triangle.

**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.**

Using the Data analysis the correlation matrix was built. It is also similar to the covariance matrix. In covariance matrix we get to know the direction but in this we get to know magnitude. The value of the magnitude is always between -1 and 1. It is a dimension less variable. If the value of any two variables is +ve, that means the relationship is positive otherwise negative.

The correlation coefficient is +ve then that means with increase in x there is also increase in y. We can say x is directly proportional to y. If the coefficient is –ve that means with increase in x, y is gradually decreasing. It can be said x and y are inversely proportional. The value being close to 0 then it means there is no correlation between the pair, if value near +1 then the pair has positively linear relationship. If value is near -1 then the pair has negatively liner relationship.

The correlation of a variable with itself is always 1. So, we get 1 in the diagonal as the pairs are in correlation with themselves. The absolute value of the correlation coefficient tells us how strong the relation is between the two variables i.e.,

r = 0 - 0.19 (Very Weak)

r = 0.2 - 0.39 (Weak)

r = 0.4 - 0.59 (Moderate)

r = 0.6 - 0.79 (Strong)

r = 0.8 - 1 (Very Strong)

The top three positively correlated pairs and the top three negatively correlated pairs from the above correlation matrix are as below.

|  |
| --- |
| **Top 3 Positively Correlated** |
| 0.910228189 |
| 0.763651447 |
| 0.731470104 |

|  |
| --- |
| **Top 3 Negatively Correlated** |
| -0.737662726 |
| -0.613808272 |
| -0.507786686 |

**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?**



**Regression Equation:** *AVG\_PRICE = 34.554 + (-0.950)\*LSTAT*

**a)** Regression Summary tells us how well the linear regression model we built fits the data given. It contains different measures that help us in determining the acceptance of the model and also to compare between two or more models. The summary contains various variables such as Multiple R, R2, Adjusted R2, Coefficients, T-stat, P-value, Standard Error etc.

The inference of above summary can be explained in terms of variance, coefficient value, intercept, and residual plot. The summary contains a measure called Multiple R which is the correlation coefficient which tells us how strong the linear relationship is. The R Square is the coefficient of determination which tells us the proportion of variation of Y values around mean which are explained by X. The Adjusted R square adjusts for the number of terms in model and it’s mostly instead of R square if more than one independent variable is present. The values of the measures that are generated by the regression model are 0.7377, 0.5441, and 0.5432 for Multiple R, R Square and Adjusted R Square respectively.

|  |  |
| --- | --- |
| ***Regression Statistics*** | |
| ***Multiple R*** | 0.737662726 |
| ***R Square*** | 0.544146298 |
| ***Adjusted R Square*** | 0.543241826 |

Coefficient values interpret the values by which the dependent variable is getting affected by a unit change in the respective independent variable. Since the model we built is linear regression so we have only two coefficient values one is Intercept whose value is 34.553 and other one is LSTAT coefficient i.e., X coefficient whose value is -0.95. The Coefficient values indicate us about the relation between Y and X. If coefficient is -ve then X and Y negatively related its otherwise if its +ve. The Intercept tells us Y value if the X is 0 or the X coefficient is 0, this value is the point where the line touches the y-axis in the graph.

|  |  |
| --- | --- |
|  | *Coefficients* |
| ***Intercept*** | 34.55384088 |
| ***LSTAT*** | -0.950049354 |

A residual plot is a scatter plot that displays the residuals on the vertical axis and the independent variable on the horizontal axis. Residual plots help us to determine whether a linear model is appropriate in modelling the given data. Since a residual is the "leftover" value after subtracting the expected value from the actual value and the expected value is obtained through a linear model such as a line of best fit, a residual plot shows how the data points deviate from the model.

If the residuals are randomly scattered around, it means that a linear model approximates the data points well without favouring certain inputs. In such a case, we conclude that a linear model is appropriate. If the residuals show a curved pattern, it indicates that a linear model captures the trend of some data points better than that of others. In such a case, we should consider using a model other than a linear model. Since the below residual plot is not showing any pattern we can conclude that the linear regression model we built is appropriate.

**b)** Yes, LSTAT variable is significant as per the model built since its p-value is less than 0.05. A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true. The lower the p-value, the greater the statistical significance of the observed difference. A p-value of 0.05 or lower is generally considered statistically significant.

|  |  |
| --- | --- |
|  | *P-value* |
| ***Intercept*** | 3.7431E-236 |
| ***LSTAT*** | 5.0811E-88 |

**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?**

**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.**



**a)** From the regression summary, the coefficient of the intercept gives us the alpha value, coefficient of avg\_room is beta one and coefficient of LSTAT is beta two. We now substitute these coefficients and get the equation.



The values of Avg\_Room and LSTAT are given as 7 and 20 respectively, we now substitute in the equation.

*AVG\_PRICE = (-1.358272812) + 5.094787984 \* 7 + (-0.642358334) \* 20*

*AVG\_PRICE = 21.458076396*

We observe that the avg\_price of the house in Boston as 21.45K but the company is quoting 30K. So the company is overcharging for the houses.

**b)** The Adjusted R square adjusts for the number of terms in model and it’s mostly instead of R square if more than one independent variable is present. So, we use adjusted R square to compare this model and previous model as there is a difference in the independent variables used. The model with best adjusted R square has better performance. The previous model has the adjusted R square value of 54% and the current model has the adjusted R square of 63%. There is clear difference between the two models and the current regression model has better performance.

 

**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 each independent variable with**

**respect to AVG\_PRICE.**





The inference of above summary can be explained in terms of variance, coefficient value, intercept, and residual plot. The summary contains a measure called Multiple R which is the correlation coefficient which tells us how strong the linear relationship is. The R Square is the coefficient of determination which tells us the proportion of variation of Y values around mean which are explained by X.



The Adjusted R square adjusts for the number of terms in model and it’s mostly instead of R square if more than one independent variable is present. The values of the measures that are generated by the regression model are 0.8329, 0.6938, and 0.6883 for Multiple R, R Square and Adjusted R Square respectively.

Coefficient values interpret the values by which the dependent variable is getting affected by a unit change in the respective independent variable. Since the model we built is multi linear regression with more than one independent variable so we have more coefficient values, one is Intercept whose value is 29.241. The Intercept tells us Y value if the all X values are 0 or the X coefficients are 0, this value is the point where the line touches the y-axis in the graph. The Coefficient values indicate us about the relation between Y and X. The coefficient is -ve then X and Y negatively related it’s otherwise its +ve. The variables which have positive relation with Avg\_Price are Crime\_Rate, Age, Indus, Distance and Avg\_Room. The variables which are negatively affecting are NOX, Tax, PTRATIO and LSTAT.

The p-value for each independent variable tests the null hypothesis that the variable has no [correlation](https://statisticsbyjim.com/glossary/correlation/) with the dependent variable. If there is no correlation, there is no change in dependent variable with increase or decrease in independent variable. If the [p-value](https://statisticsbyjim.com/glossary/p-value/) for a variable is less than your [significance level](https://statisticsbyjim.com/glossary/significance-level/), your sample data provide enough evidence to reject the null hypothesis for the entire population. This variable is statistically significant and probably a worthwhile addition to your regression model. On the other hand, when a p value in regression is greater than the significance level, it indicates there is insufficient evidence in your sample to conclude that a correlation exists.

Since all the independent variables except Crime\_Rate variable have p-value less than 0.05 which is significance, we can consider adding them in the regression model and remove Crime\_Rate variable to increase the model performance.

**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.**

**a)** The summary contains a measure called Multiple R which is the correlation coefficient which tells us how strong the linear relationship is. The R Square is the coefficient of determination which tells us the proportion of variation of Y values around mean which are explained by X.

The Adjusted R square adjusts for the number of terms in model and it’s mostly instead of R square if more than one independent variable is present. The values of the measures that are generated by the regression model are 0.8328, 0.6936, and 0.6887 for Multiple R, R Square and Adjusted R Square respectively.



Coefficient values interpret the values by which the dependent variable is getting affected by a unit change in the respective independent variable. Since the model we built is multi linear regression with more than one independent variable so we have more coefficient values, one is Intercept whose value is 29.248. The Intercept tells us Y value if the all X values are 0 or the X coefficients are 0, this value is the point where the line touches the y-axis in the graph. The Coefficient values indicate us about the relation between Y and X. The coefficient is -ve then X and Y negatively related it’s otherwise its +ve. The variables which have positive relation with Avg\_Price are Crime\_Rate, Age, Indus, Distance and Avg\_Room. The variables which are negatively affecting are NOX, Tax, PTRATIO and LSTAT.

**b)** The Adjusted R square adjusts for the number of terms in model and it’s mostly instead of R square if more than one independent variable is present. So, we use adjusted R square to compare this model and previous model as there is a difference in the independent variables used. The model with best adjusted R square has better performance. The previous model has the adjusted R square value of 68.83% and the current model has the adjusted R square of 68.87%. There is slight difference between the two models and the current regression model has better performance.

 

**c)** 

The Coefficient values indicate us about the relation between Y and X. The coefficient is -ve then X and Y negatively related it’s otherwise its +ve. The NOX variable and Avg\_Price have a negative relation as we can observe from the above table. In other words, with a unit value increase in NOX variable the Avg\_Price value is decreasing by 10.272.

**d)**



The Regression Equation above is generated after performing the regression analysis on the data. It gives us the success rate of 69%.