**Terro’s Real Estate Agency**

**Main report**

1. **The first step to any project is understanding the data. So for this step, generate the summary statistics for each of the variables. What do you observe?**

* You can see the mean median mode to have an idea about the data is distributed centrally. Thus it is used to determine the “centre” of the distribution of the data.

Mean: average of all values

Median: it is the middle values of data after it has been sorted.

Mode : the most common value.

* Variance: It is a measure of how far the observation from the mean (i.e. average area of the mean)
* Standard deviation : it is used to quantify the amount of variation (square root of variation)

The greater will be the standard deviation the greater will be the magnetite of dispersion.so we can check each variable for further study.

* **The “shape” refers to how the data values are distributed across the range of values in the sample**

**kurtosis**

* **Leptokurtic**: Sharply peaked with fat tails (these are CRIMERATE, AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO
* **Platykurtic**: Flattest peak and highly dispersed (these are AVG ROOM, LSTAT, AVG PRICE)

**Skewness**

* Positive: The distribution is right-skewed, which means the tail extends to the right side of the distribution. (They are CRIMERATE, INDUS, NOX, DISTANCE, TAX, LSTAT, AVG PRICE, AVG ROOM)
* Negative: The distribution is right-skewed, which means the tail extends to the right side of the distribution (they are AGE AND PT RATIO)

**RANGE**

The range is simply the difference between the maximum and the minimum values. SO individually we can see the range and above this we can find the outliers in the further study.

1. **Plot the histogram of the Avg\_Price Variable. What do you infer?**

I observe that the AVG PRICE is positively skewed ,it means that most of the data lie on left side of the distribution.

The distribution is right-skewed, which means the tail extends to the right side of the distribution.

In a positively skewed distribution, mode < median < mean.

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

The diagonal of covariance matrix provides the variance of each individual variable, covariance with itself.

The metric evaluates how much – to what extent – the variables change together.  Thus, **Covariance** is a measure of the relationship between two random variables. So it measures the joint dispersion/variation of various variables.

* **Positive covariance (>0)**: Indicates that two variables tend to move in the same direction. I can observe that the cell that are in green in colour are positive covariance. So crime rate and average room are positive covariance with avg price. Thus both crime rate and avg room moves in same direction with avg price.
* **Negative covariance(<0)**: Reveals that two variables tend to move in inverse directions. Other than crime rate and avg room all the variables are negative covariance with avg price. (Moves in the opposite direction)

1. **Create a correlation matrix of all the variables as shown in the Videos and various case studies. State top 3 positively correlated pairs and top 3 negatively correlated pairs.**

**top 3 positively correlated pairs** – 1) DIS-TAX

2)INDUS-NOX

3)AGE-NOX

4)AVG ROOM-AVG PRICE

**top 3 negatively correlated pairs-** – 1) LSTAT-AVG PRICE

2)AVG ROOM-LSTAT

3)PT RATIO-AVG PRICE

1. **Build an initial regression model with AVG\_PRICE as the y or the Dependent variable and LSTAT variable as the Independent Variable. Generate the residual plot too. (8 marks)**

Y= -0.95X+34.55 AVG PRICE = -0.95\*LSTAT +34.55

* 1. **What do you infer from the Regression Summary Output in terms of variance explained, coefficient value, Intercept and the Residual plot?**
* R-Squared indicates how well the model or regression line “fits” the data. It indicates the proportion of variance in the dependent variable (Y) explained by the independent variable (X). The R-square indicates the percentage of variation in AVG PRICES explained by the LSTAT. It tells you the percentage of change in AVG PRICES that is caused by varying the LSTAT. **Our regression output indicates that 54% of the variation in AVG PRICE is explained by LSTAT. And 46% (100%-54%) of the variation is caused by factors other than LSTAT.**
* **The intercept of 34.55 indicates that avg prices will be 34.55 if there is no LSTAT(% lower status of population)**. This is because LSTAT is zero, it (zero) is multiplied by the slope or m (here -0.95), resulting in a zero. This is added to your intercept, leaving you only the intercept value 34.55
* The coefficient m (here -0.95) indicates that for every unit increase in the X variable (here LSTAT), the Y variable (here AVG PRICE) will change by the amount of the coefficient -0.95.  It is also referred to as the slope of the line in a simple linear equation.

HERE the coefficient of the independent variable LSTAT is negative, for every unit increase in the independent variable, the dependent variable (AVG PRICE) will decrease by the value of the coefficient. Correspondingly, for every unit decrease in the independent variable, the dependent variable will increase by the value of the coefficient. **LSTAT show negative correlations with house prices.**

* Residual refers to the difference between observed value(y) vs predicted value (y^).

residual=observed Value−predicted Value =y−y^

The data points form a curved pattern, a U-shaped pattern. Since there is a detectable pattern in the residual plot, **we conclude that a linear model is not a right fit for the data.**

* 1. **Is LSTAT variable significant for the analysis based on your model?**

Though LSTAT is a significant variable as p value is less than 0.05 AS P-value of the LSTAT variable in our model is very small and This indicates that this is a ‘significant variable’ and that the LSTAT is likely to impacts AVG PRICES. BUT R SQUARE is not good thus LSTAT is not able to explain the dependent variable fully. It tells you the percentage of change in AVG PRICES that is caused by varying the LSTAT.

**Our regression output indicates that 54% of the variation in AVG PRICE is explained by LSTAT. And 46%(100%-54%) of the variation is caused by factors other than LSTAT.**

If the percentage of lower working-class people is higher, it is likely that they have low purchasing power and therefore, they houses will cost less. They are inversely proportional variables.

1. **Build another instance of the Regression model but this time including LSTAT and AVG\_ROOM together as Independent variables and AVG\_PRICE as the dependent variable.**

Y=M1X1 +M2X2 +C

1. **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?**

Y=M1X1 +M2X2 +C AVGPRICE=M1AVGROOM+M2LSTAT+C

THE VALUE OF AVG PRICE IS 21.45 $ (i.e. 21458 USD)

IF the company quoting a value of 30000 USD for this locality then they are overcharging.

1. **Is the performance of this model better than the previous model you built in Question 5? Compare in terms of adjusted R-square. Explain.**

More is the value of r-square near to 1, better is the model. But the problem lies in the fact that the value of r-square always increases as new variables(attributes) are added to the model, no matter that the newly added attributes have a positive impact on the model or not. also, it can lead to overfitting of the model if there are large no. of variables

**Adjusted r-square is a modified form of r-square whose value increases if new predictors tend to improve model’s performance and decreases if new predictors do not improve performance as expected.** Adjusted R² will only increase when the added variable is relevant.

**YES** the performance of this model better than the previous model I built in Question 5 because here both the R square the adjusted R Square tends to increase from 54% to 63%. So here **adjusted R-squared says additional input variable ( i.e. AVG ROOM) are contributing to the model more efficiently.**

1. **Now, build a Regression model with all variables. AVG\_PRICE shall be the Dependent Variable. Interpret the output in terms of adjusted R-square, coefficient and Intercept values, Significance of variables with respect to AVG\_price. Explain**

* The adjusted R-squared is a modified version of R-squared that adjusts for predictors that are not significant in a regression model.
* Compared to a model with additional input variables, a lower adjusted R-squared indicates that the additional input variables are not adding value to the model.
* Compared to a model with additional input variables, a higher adjusted R-squared indicates that the additional input variables are adding value to the model.

So in our case this regression model explains that the ADJUSTED R SQUARE is quite good contributing to 68%. So this shows that adding all the new variables has increased the efficiency of the regression model. Thus mostly the variables are relevant to the model.

* If the coefficient of the independent variable X is positive, it indicates for every unit increase in the independent variable; the dependent variable will increase by the value of the coefficient. CRIMERATE, AGE, INDUS, NOX, DISTANCE AND AVG ROOM all are positive.

On the other hand, if the coefficient of the independent variable X is negative, for every unit increase in the independent variable, the dependent variable will decrease by the value of the coefficient. NOX,TAX,PTRATIO all are negative coefficient.

* The intercept of 29.24 indicates that AVG PRICES will be 29.24 if there is no independent variable. This is because all the x values, i.e (zero) is multiplied by the slope or m , resulting in a zero. This is added to your intercept, leaving you only the intercept value 29.24
* The P-value indicates the probability that the estimated coefficient is wrong or unreliable. The best way to understand the P-value is as the “probability of an error.”  We want the P-value to be as small as possible.

How small should the P-value be? That depends on a cut-off level that we decide on separately. This cut-off level is called the significance level. The cut-off selected depends on the nature of the data studied and the different error types. The cut-off or significance level is usually 1%, 5%, or 10%. **Generally, a cut-off point of 5% is used.**

I have noticed that the P-value of the independent variable in our model is very small (i.e less than 5%) except CRIMERATE. This indicates that this is a ‘significant variable’ and that the all the independent variable (X)is likely to impacts AVG PRICE (Y).

**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. (8 marks) (HINT: Significant variables are those whose p-values are less than 0.05. If the p-value is greater than 0.05 then it is insignificant) Answer the questions below**:

**a. Interpret the output of this model**.

This model is best as compared to other model.

* **Multiple R: 0.832**. This represents the multiple correlation between the dependent variable (AVG PRICE) and the independent variables (others).
* **R Square: 0.693**: This is known as the coefficient of determination. It is the proportion of the variance in the dependent variable (AVG PRICE) that can be explained by the independent variables. **Our regression output indicates that 69% of the variation in AVG PRICE is explained by other independent variables. And 31% (100%-69%) of the variation is caused by factors other than these independent variables.**
* **Adjusted r-square (0.688):** This is a modified form of r-square whose value increases if new predictors tend to improve model’s performance and decreases if new predictors do not improve performance as expected**. SO adjusted R-squared says additional input variables are contributing to the model more efficiently.**
* **P-values.**The individual p-values tell us whether or not each independent variable is statistically significant.

So here all the independent variables are below 0.05. This indicates that these independent variables are ‘significant variables’ and all the independent variables (X)is likely to impacts AVG PRICE (Y).

**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?**

According to the data this question reg model is slightly good as their adjusted r square is (0.6886) more than this question 7 model (.6882). By removing the insignificant variable our new model has been increased.

**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?**

The coefficient of the independent variable NOX is negative, for every unit increase in the independent variable NOX, the dependent variable AVG PRICE will decrease by the value of the coefficient(i.e. -10.27)

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

AVGPRICE=(-10.27NOX) +(-1.07PTRATIO) +(-0.60LSTAT) +(0.01TAX)+0.03AGE+0.13INDEX+0.25DISTANCE+4.12AVGROOM+29.42