Assignment – Terro’s real estate agency

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

• From the given data, the mean=68.575 shows the average age value

in the town.

• From the median=77.5, it clearly shows that most of them are aged

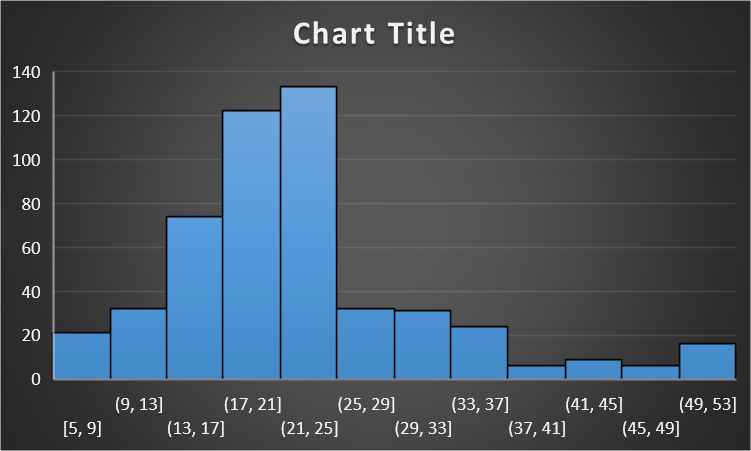
persons in the town.

• All the given variables except the PT-ratio have a Positive skewness.

• There is minimum outlier data in Crime rate, NOX and

Avg. Room

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



• AVERAGE PRICE is a dependent variable, which has a very few outlier

data and it is Rightly skewed with a leptokurtic kurtosis.

• There are 133 Houses with the price range of 21000 dollars to 25000

dollars.

3) Compute the covariance matrix. Share your observations.

There are both Positive covariance and Negative covariance in the above matrix.



The most positive covariance value is 2397.942 (Tax and Age)

The most Negative covariance value is -724.82 (Tax and Avg\_price)

4) Create a correlation matrix of all the variables (Use Data analysis tool pack).



a) Which are the top 3 positively correlated pairs and

Top 3 Positive correlated values

• Tax and Distance

• Indus and Nox

• Age and Nox

b) Which are the top 3 negatively correlated pairs.

Top 3 Negative Correlated va

• Avg\_price and Lstat

• Avg\_room and Lstat

• Avg\_price and PT ratio

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?

• The value of the variance is less than 0.05, so the model is

highly significant.

• From the residual plot, we can see that there is some

outlier in a data

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

Yes, LSTAT variable is significant for the analysed mode

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?

By using a regression equation, the AVG\_PRICE value is

21.4581. Comparing to the quoting value, the company is

overcharging for the property.

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

• Previous model Adj. R-square= 0.543241825954707

• New model Adj. R-square= 0.637124475470123

• Yes, the performance of this new model is better than the previous model by comparing to the value of adjusted R-square

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 withrespect to AVG\_PRICE.



By seeing the coefficient from the regression data, we can say that

• Avg. price and Avg. room variables are directly proportional.

• NOX and Avg. Price are inversely proportional to each

other. The significance variables in this model are:

• AGE

• INDUS

• NOX

• DISTANCE

• TAX

• PTRATIO

• AVG\_ROOM

• LSTAT

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. This model has a Better Adjusted R-square value of 68.86%

This model has a Better Adjusted R-square value of 68.86%.

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 Rsquare?

The New model performs well than the previous model by

comparing the value of an Adjusted R-square value

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 AVERAGE\_PRICE is HIGH, if the NOX is LOW

d. Write the regression equation from this model.

Multi-Linear regression equation:

Y=m1x1+ m2x2+ m3x3+ m4x4+ m5x5+ m6x6+ m7x7+ m8x8 + c

Multi-linear regression equation of this model:

Y= 4.515421 X AVG\_ROOM - 0.57181 X LSTAT – 0.93072 X PTRATIO

Y= Predicted Variable

m= weight of the variable

Xn= Independent variable

C= Intercept