### REAL ESTATE DATA ANALYSIS

PROJECT REPORT

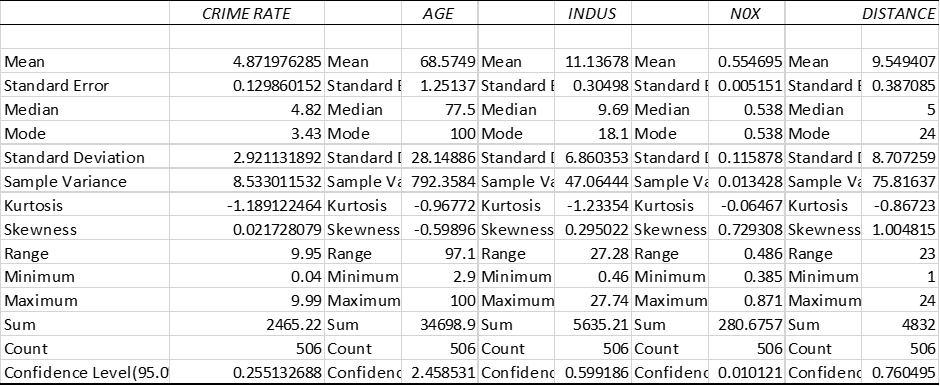
**Problem Statement (Situation):** **“Finding out the most relevant features for pricing of a house”**

**objectives**

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.

* To analyse the magnitude of each variable to which it can affect the price of a house in a particular locality using exploratory data analysis and linear regression.

**FINDINGS**

1. Generate the summary statistics for each variable in the table. Write down your observation



* **CRIME RATE**

Since mean and median values are same, there’s no outliers present in the data. Skewness is almost zero and kurtosis is -1.189, hence platykurtic.

Range of the crime rate is 9.95.

* Age

The mean of age is 68.5749 and median 77.5, which means there’s outliers are present. The curve is platykurtic and negatively skewed. Range is 97.1

* Industry

The mean is 11.13 and median is 9.69. the curve is platykurtic and positively skewed.

* NOX

The values of mean, median and mode are equal, hence the data is normally distributed.

* Distance

Mean and median values shows that there are outliers present in the data. The data is platykurtic and positively skewed.

* Tax

The mean is 408 and median 330, which shows there is outliers present in the data. Positively skewed and platykurtic.

* People teacher ratio

Mean is 18 and median is 19, hence no outliers are present. Negatively skewed and platykurtic

* Average room

Mean and median are same, hence normally distributed. the data is positively skewed and leptokurtic.

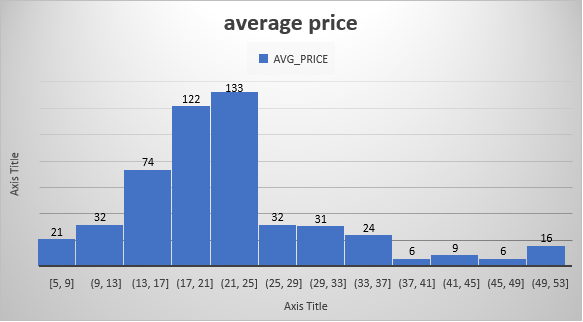
* Lower status

Mean is 12 and median is 11, leptokurtic and positively skewed data and range is 36.

* Average price

Mean and median are almost same, hence no presence of outliers. 45 is the average price range.

1. Plot a histogram of the Avg-Price variable. What do you infer?



* The average prices of 133 houses are in between 21 to 25.
* Among the total houses the average prices of 122 houses are in between 17 to 21.
* The least average prices of 6 houses are between 37 to 41 and 45 to 49.

1. Compute the covariance matrix. Share your observations



|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

The covariance measures joint dispersion of two variables or their joint variation.

Age and tax, industry and tax, distance and tax have the highest joint dispersion. Average price and age, distance and average price, tax and average price have the lowest joint dispersion.

1. Create a correlation matrix of all the variable



The correlation or correlation coefficient is a dimensionless version of the covariance.

Correlation is a number between -1 and +1.

+1 indicates a perfect positive linear relationship between X and Y.

-1 indicates a perfect negative linear relationship between X and Y.

0 indicates either no relationship or a complex nonlinear relationship

1. Which are the top 3 positively correlated pairs

* Distance and tax
* Age and NOX
* Industry and NOX

1. Which are the top 3 negatively correlated pairs

* Lower status population and average price
* Average room and LSTAT
* People teacher ratio and average price

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



1. do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?

The linear regression model is given by y=34.5-0.95x

where y – average price (dependent variable) and x – lower status population (independent variable)

The overall quality of fit in multiple linear regression can be assessed using R-square (or Multiple R-square). R-square measures the percentage of the variation in Y that is collectively explained by the X variables in the model.

Here the R square value is 0.5449 , which is good, that is around 55%. Then we can say that the model is good.

The intercept of the model is 34.5

A residual plot shows the difference between the observed response and the fitted response values.

The plots showing the random scatter of points , there is no linear pattern present in the model, hence model is significant.

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

Regression model: y=34.5-0.95X1

Y: average price

X1: lower status population

Here, p value is 5.08\*e^10 that is less than the significant value 0.05. Therefore, reject the null hypothesis. Lower status population variable is significant for the model.

6) Build a new Regression model including LSTAT and AVG\_ROOM together as independent variables and AVG\_PRICE as dependent variable



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?

The regression model is given by y=(-1.29)+5.09x1-0.64x2

Where y – average price (dependent variable)

X1 – average room

X2 – lower status

Now put x1=7 and x2 = 20, the average price ,

y= (-1.29)+(5.09\*7)+(-0.64\*20) = 21.54

the value of 30000 USD that the company quoting is not significant to this model. The predicted average price from the regression model is 21,540, hence the company is 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 and explain.

Since all p values are significant and R square is 63%, this model is better than the previous model.

Here the adjusted R square is 63.76, that is 64% of the dependent variable being studied is explained by the variance of the independent variable. Here the model can say a good fitted 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



The regression model is

Y=29.34+0.05x1+0.03x2+0.12x3-10.15x4+.026x5-0.014x6-1.07x7+4.12x8-0.61x9

Where y – average price

x1 – crime rate

X2 – age

X3 – industry

X4 – nitric oxide

X5 - distance

X6 – tax

X7 – people teacher ratio

X8 – average room

X9 – lower status population

* The intercept of the model is 29.34 , that is the mean value of the response variable when all of the predicter variables in the model are zero.
* The adjusted R square is 0.689, that is 68.9% of the variance of the dependent variable being studied is explained by the independent variables, hence here the model is a good fit,
* X1 - crime rate : if one-unit changes in crime rate holding other variables constant 0.05 changes happen in average price
* X2 – age : if one-unit changes in age holding other variables constant 0.03 changes happen in average price
* X3 – industry : if one-unit changes in industry 0.12 changes happens in average price, holding other variables constant.
* X4- nitric oxide : if one-unit changes in nitric oxide -10.15 changes happens in average price, holding other variables constant.
* X5 – distance : if one- unit changes in distance, 0.26 changes happen in average price, holding other variables constant.
* X6 –tax: if one-unit changes in tax holding other variables constant -0.01 changes happen in average price
* X7 – people teacher ratio: if one-unit changes in ptratio holding other variables constant -1.07 changes happen in average price
* X8 – average room: if one-unit changes in average room 4.12 changes happens in average price, holding other variables constant.
* X9 – lower status population : if one-unit changes in lower status population -0.61 changes happens in average price, holding other variables constant.
* The p value of crime rate is 0.518 which is greater than the significance level 0.05 , hence we accept the null hypothesis that the crime rate is not significant to the average price rate.
* The p value for all other variables is less than 0.05, hence we accept the alternate hypothesis. That is except crime rate all other variables are significant to the average price rate of the locality.

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.

* Now the model contains the significant variables only. That is their p values is less than 0.05 and hence we accept the alternate hypothesis that all the variables age, industry, nox, distance, tax, ptratio, lstat and average room are significant.

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?

* Since the crime rate doesn’t make any impact on the model , the adjusted R square value doesn’t change a lot , after removing it also.

For the previous model it was 68.91% and for the new model it is 68.95%, only 0.4% difference. Hence comparing with the previous model new model is a good fit.

1. 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?



Here the coefficient of nitric oxide is -10.103 , that is one unit increase in the nox -10.103-unit changes in the average price of the property holding other variables constant. So, we can say that one percent increase in the nitric oxide concentration causes 10.103 percentage decrease in the average price of the property.

d) Write the regression equation from this model

The regression model is

Y=29.53-10.103X1-1.076X2-0.607X3 - 0.014X4+0.033X5+0.124X6+0.257X7+4.109X8

X1- NOX

X2- people teacher ratio

X3- lower status population rate

X4- age

X5- tax

X6- industry

X7- distance

X8 – average room

SUMMARY

* The higher average price of houses is between 21 to 25 and the least average price are between 45 to 49.
* The highly positively correlated variables are Distance and tax, age and nox, industry and nox.
* The highly negatively correlated variables are Lower status population and average price, average room and lstat , ptratio and avg price.
* Among the variables crime rate is not significant to the average price of the house. All other variables are significant to the average price of the houses.
* The fitted regression model is

y=29.53-10.103X1-1.076X2-0.607X3 - 0.014X4+0.033X5+0.124X6+0.257X7+4.109X8

where X1- NOX,X2- people teacher ratio,X3- lower status population rate ,X4- age,X5- tax,X6- industry,X7- distance,X8 – average room