Business Report :

Analysing Factors Affecting House Price in Boston.

NAME : Archana Ashok Malleshe.

Organization Name: Great Learning Pune.

Contents:

1. Introduction:

• Context:

The price of houses dependent on various factors like size or area, how many rooms, the price of other houses and many other factors. Real estate investors would like to find out the actual cost of the house in order to buy and sell real estate properties. Terro's real-estate is an agency that estimates the pricing of houses in a certain locality. To do this activity, the company employs an auditors who studies various geographic features and factors, including crime rates, education facilities, connectivity (distance from highway), etc. This helps in determining the price of a property.

• Problem Statement:

Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. In this project, house prices will be predicted given explanatory variables that cover many aspects of residential houses. The goal of this project is to provide invaluable insights in order to make the best possible, data-based decisions to optimise business performance. In addition, house price predictions are also beneficial for property investors to know the trend of housing prices in a certain location. So prediction of house prices are expected to help people who plan to buy a house so they can know the price range in the future, then they can plan their finance well.

2. About the Dataset:

The dataset consists of information related to 506 houses in Boston. It includes various attributes such as CRIME RATE, INDUSTRY, NOX AVG_ROOM, AGE, DISTANCE, TAX, PTRATIO, LSTAT, AVG_PRICE .

• Data Dictionary:

The variables given in the dataset are the following:

- 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 the highway (in miles).
- TAX: Full-value property-tax rate per $10,000.
- PTRATIO: Pupil-teacher ratio by town.
- LSTAT: % of lower status of the population.
- AVG_PRICE: Average value of houses in $1000.

3. Objective:

The objective of this study is to map all the relevant features for the properties along with the information related to the geography around it to estimate the value of a particular property/house. The agency wants to understand the relevance of the parameters that they collect in relation to the value of the house. The analysis involves doing EDA (Exploratory data analysis) and Regression analysis.

• Specific Objectives:

1. Generate summary statistics for key variables in the dataset.

2. Identify the most relevant features affecting house prices.

3. Build a predictive model for estimating house prices based on selected features.

4. Provide insights and recommendations for real estate stakeholders.

4. Scope and limitations:

• Scope:

This project primarily concentrates on the analysis of a specific dataset comprising 506 houses in Boston. The focus is to provide insights into the factors that influence house prices within this dataset.

• Limitations:

1. The analysis is constrained by the dataset's limited size and scope. It's important to note that the findings may not fully encapsulate all the real-world variables that influence house prices in Boston.

2. Outside economic or market features are not measured in this analysis.

5. Data Analysis

Task:

To do the analysis, you are expected to solve these questions:

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

Ans:

We use the descriptive statistics function inside the data analytics tool pack which can be found in the data ribbon for this task. The observation that we arrived at from the table are that:

AGE (Proportion of houses built prior to 1940).

- Negative skewness specifies that most of the houses are built before 1940 years.
- The average houses built in town is around 68 years.
- Negative Kurtosis gives us a flatter distribution for AGE.

INDUSTRY: Proportion of non-retail business acres per town.

- Maximum of the towns have 18% of land for non-retail business
- Positive skewness specifies that most of the towns have more than 11.13% of land as non-retail business land.
- Negative kurtosis indicating values are spread across mean value.
- On an average 11.13% of property belongs to non-retail business.

NOX: Nitric oxides concentration (parts per 10 million).

- Skewness is 0.72, indicates most of the houses have no Concentration below 0.55 ppm.
- On an average, nitric oxide concentration is about 0.55 parts per Million.

CRIME-RATE:

- Average crime-rate in town is 4.87 per capita.
- Skewness is 0.02, closely 0 which states curve follow normal distribution.
- 50% of the crime rate in town is under 4.82 per capita and 50% above this value.

- Maximum chance of having a crime in that town is 13.63 per capita.
- Standard deviation is 2.92 per capita, states that data deviates from mean by this value.
- Maximum crime-rate in town is around 3.43 per capita.
- Minimum chance of having a crime in that town is -3.89 per capita, that means zero chance of having a crime per capita.

DISTANCE: Distance from the highway (in miles).

- Positive skewness shows that most of the houses are more than 9.5 miles away from highway.
- Maximum houses have 24 miles of distance from highway.
- On an average, distance from highway is about 9.5 miles.

AVG_PRICE: Average value of houses in $1000.

- On an average, the cost of houses is about 22.53 thousand USD.
- Positive Skewness specifies that value of maximum of the houses have more than 22.53 thousand USD.
- Determined value of houses is 50 thousand USD.

AVG_ROOM: Average number of rooms per house.

- Positive skewness specifies that maximum of the houses have more than 6 rooms.
- The mean of AVG_ROOM is about 6.2, signifying about 6 rooms are there.

TAX: Full-value property-tax rate per $10,000.

- The maximum number of houses have tax rate about $666.
- On an average, tax-rate is $408.

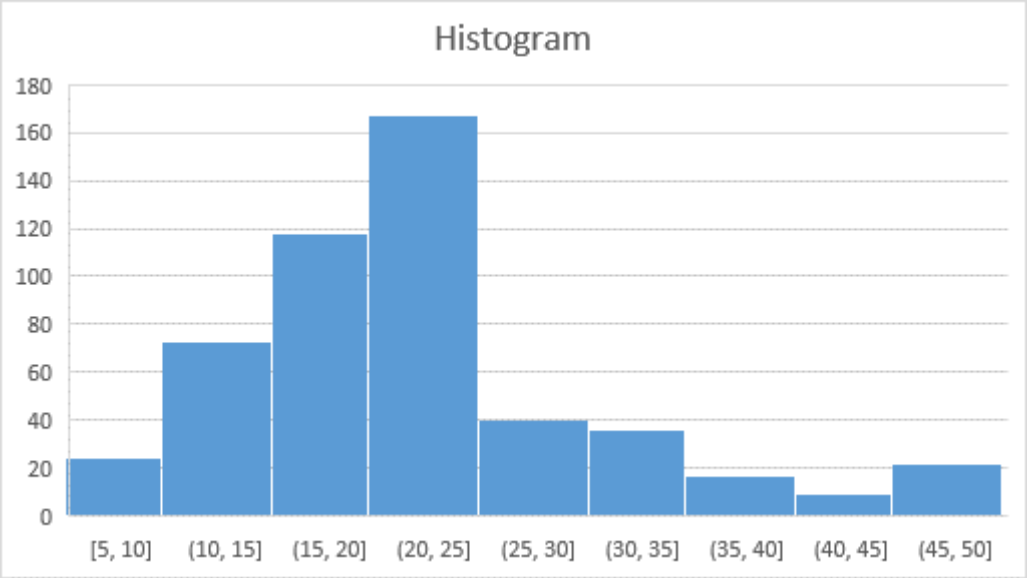
LSTAT: Percentage of lower status of the population.

- Positive Skewness specifies that most of the houses have more than 12% lower status population.
- On an average, 12% of population has lower status.

PTRATIO: Pupil-teacher ratio by town.

- On an average, Pupil-teacher ratio by town is 18 per town.
- Negative skewness specifies that most of the houses have more than 19 per town as a Pupil-teacher ratio.

2. Plot a histogram of the Avg_Price variable. What do you infer?

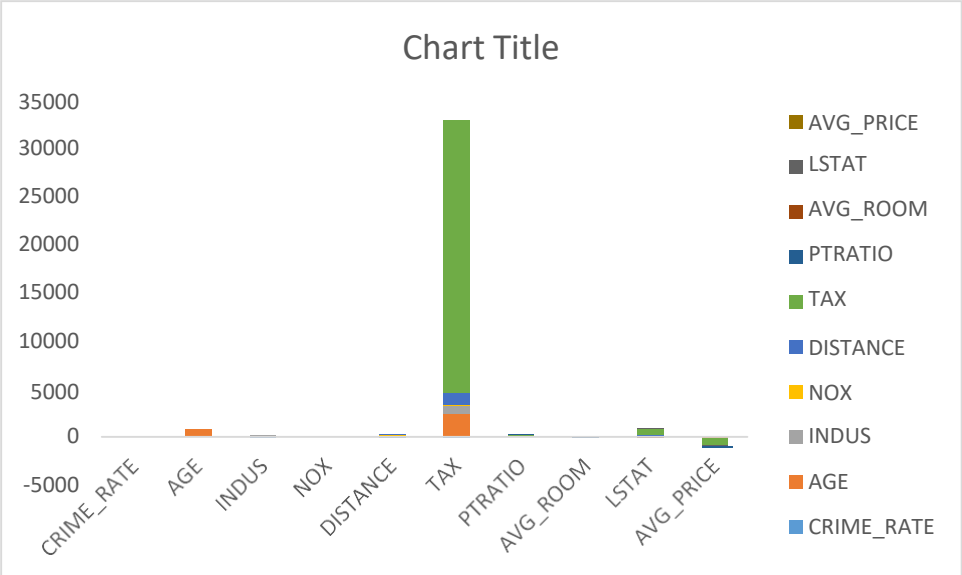


- By observing the data in the histogram, we can say that the demand of the houses with average price Between 20-25 thousand USD is higher and upto 25thousand USD demand is steadily increasing, but after 25 thousand USD demand is statistically decreasing.
- In housing business, houses of rates between 20-25thousand USD are more demandable than above 25thousand USD & below 20 thousand USD is moderate.
- Overall, the positively skewed distribution of AVG_PRICE highlights the falling diversity in house prices within the Boston locality, with the majority of houses within a moderate pricing range but some outliers commanding significantly higher values.
- The population can offer the houses which is having a range of price between 20-25 thousand USD more demandable as compared to above 25 thousand USD.

- This explanation provides a clear understanding of the distribution of house prices in the dataset, highlighting both the common price range and the presence of outliers with high prices.

- So, we can suggest that the person who involve in construction of houses try to make the houses below 25 thousand USD and construction of houses which cost below 25 thousand USD is more demandable than above 25 thousand USD.

3. Compute the covariance matrix. Share your observations.



In our analysis, we have observed distinct relationships between different variables and the target variable, AVG_PRICE.

- Average price has positive covariance with average room and crime rate. So, we can say that houses with a greater number of rooms tend to have higher average prices and areas with higher crime rates tend to have higher average house prices.

- And the remaining variables have negative relationships with AVG_PRICE.

- **NOX:** An increase in nitric oxides concentration corresponds to lower average house prices.

- **Industry:** A higher proportion of non-retail business acres per town is associated with lower house prices.

- **TAX:** A higher property-tax rate per $10,000 is linked to lower average house prices.

- AGE: The proportion of houses built prior to 1940 is inversely related to house prices, meaning that older houses tend to have lower prices.

- LSTAT: An increase in the percentage of lower-status population is negatively related to house prices.

- PTRATIO: A higher pupil-teacher ratio by town is correlated with lower house prices.

- DISTANCE: Greater distance from the highway is associated with lower house prices.

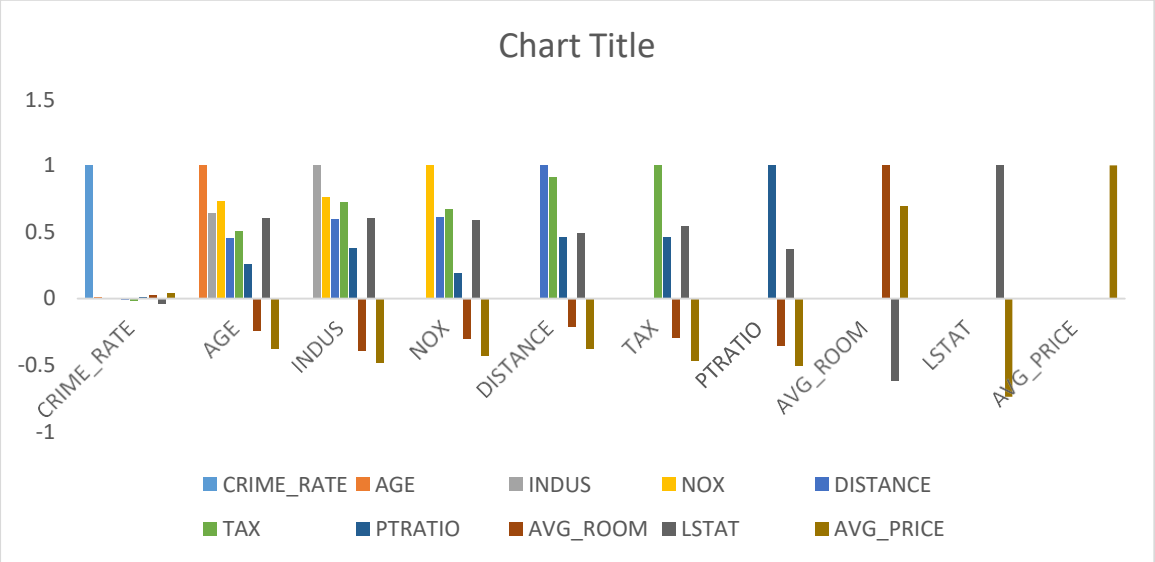
It is important to note that correlation does not imply causation, and further analysis and modeling will be necessary to establish causal relationships and make more exact predictions regarding house prices. This statement recapitulates the positive and negative relationships between different variables and AVG_PRICE, providing a clear overview of the findings from analysis. These relationships provide valuable insights into the potential factors influencing house prices in the Boston area.

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

a) Which are the top 3 positively correlated pairs and

b) Which are the top 3 negatively correlated pairs.

ANS:



• **Top 3 Negatively Correlated Pairs:**

1) PIRATIO-AVG PRICE: This specifies that areas with a higher pupil- teacher ratio are related with lower average house prices.

2) LSTAT-AVG PRICE : This specifies that as the percentage of lower- status population increases, the average house prices tend to decrease.

3) LSTAT-AVG ROOM: In areas with a higher percentage of lower- status population, houses tend to have less rooms on average.

• **Top 3 Positively Correlated Pairs:**

1) NOX-AGE: NOX-AGE pair have positive correlation, this indicates that areas with higher concentrations of nitric oxides often have a higher proportion of older houses built prior to 1940.

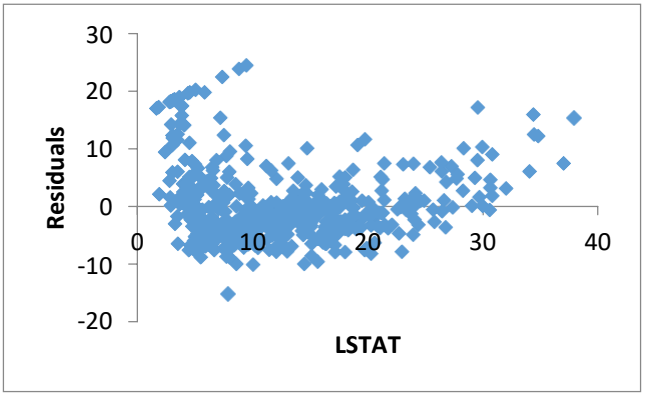
2) NOX-INDUS: NOX-INDUS pair have positive correlation, this indicates that areas with higher concentrations of nitric oxides also tend to have a greater proportion of non-retail business acres.

3) TAX-DISTANCE: TAX-DISTANCE pair have the highest positive correlation, suggests that areas with higher property-tax rates are typically located at greater distances from the highway within the Boston.

This statement providing insight into the relationships within the dataset. & summarizes the highest positive and negative correlations between pairs of variables.

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?

ANS:



a) Model Analysis:

• The residual plot has no patterns, representing no issues with the regression model. When examining the residual plot, we notice a concentration of points towards the lower values of LSTAT. This visual pattern suggests that there might be a non- linear or curvilinear relationship between LSTAT and AVG_PRICE.

• On an intercept of the model we can say that, even if the LSTAT variable is 0, the predicted value of AVG_PRICE is positive, starting at 34.55. This intercept value provides valuable insight, about house prices in the absence of lower-status population**.**

• The LSTAT variable has negative coefficients represents that price decreases as LSTAT increases.

• The model has an R-squared value of 0.544. This indicates that, model explains approximately 54.4% of the variance AVG_PRICE. It suggests that the model does not explain the variation in price very well, there may be other factors not considered in our model that also influence house prices.

b) Significance of LSTAT:

The significance value indicates whether the variable is statistically significant in explaining the variation in the target variable .The significance of predictor variables is crucial in regression analysis. In our case, the variable LSTAT (Percentage of Lower Status of the Population) has a significance value that is very close to 0, but not exactly 0.

Since the significance value for LSTAT is less than the commonly used significance level of 0.05 (typically indicating a 95%confidence level), we determine that LSTAT is indeed statistically significant and should be retained in our analysis. LSTAT has a significant variable. These implies that changes in LSTAT have a meaningful impact on predicting changes in house prices. These explanations provide an overview of model's performance and the significance of the LSTAT variable in predicting house prices, offering insights into regression analysis.

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) Regression Equation:

Regression Equation y = a+ b1X1 + b2X2

Where: X1 represents AVG_ROOM (Average Number of Rooms per House).

X2 : represents LSTAT (% of Lower Status of the Population).

y : represents the predicted AVG_PRICE.

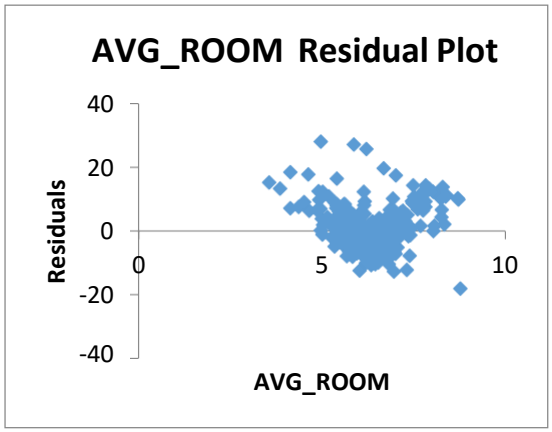
a : represents intercept.

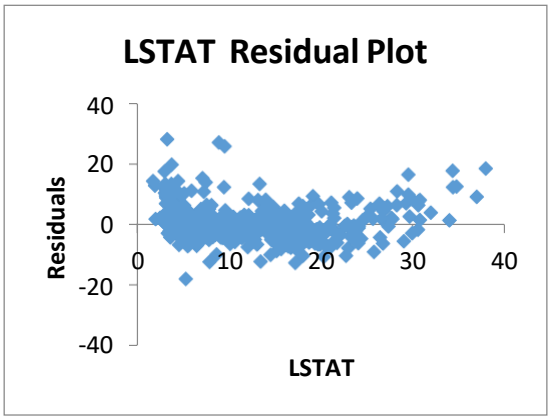
y = a+ X1b1 + X2b2

y = (-1.35827) + 7 * (5.09478) + 20 * (-0.64235)

y= 21.45819 i.e predicted price.

. Predicted price is 21.46K USD and quoted price is 30K USD, so we can interpret that companies are overcharging. This calculation suggests that, for a house with 7 rooms and a LSTAT value of 20, the predicted house price is approximately 21.46K USD.





b) Model Performance:

b) The R-square value here is 0.64 as compared to 0.54 of the previous model, this means by adding one more variable R-square increases, that says average room is also important factor to decide the price of the houses. So, by adding AVG_ROOMS to our existing model, we are able to capture additional 10% of the variance in AVG_Price, then we can say that this is a better model than the previous one.

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 R-squared value of this improved model stands at 0.6939. Comparatively, the previous model (with LSTAT and AVG_ROOM) had an R-squared value of 0.64. This indicates that our improved model captures a greater proportion of the variance in AVG_PRICE, highlighting its superior explanatory power.

The intercept value in our improved model is calculated as 29.24. This observation implies that even in scenarios where all the independent variables in our model are set to zero, the predicted value of AVG_PRICE remains at 29.24. This baseline value provides valuable insights into the starting point for house prices in the absence of specific predictors.

Among the variables considered, CRIM_RATE (Per Capita Crime Rate) shows a p-value that exceeds the common significance level of 0.05. Consequently, it is advisable to consider dropping CRIM_RATE from our model, as it does not appear to be statistically significant in predicting house prices. Conversely, the remaining variables exhibit p-values below 0.05, signifying their significance in the model.

Additionally, the adjusted R-squared value for the improved model is calculated as 0.6883. This statistic suggests that the significant variables within our model collectively account for approximately 68.83% of the variance in house prices. This underscores the importance and relevance of the selected predictor variables in explaining house price variability.

In summary, our improved regression model demonstrates superior performance, as evidenced by its higher R-squared value and adjusted R-squared value. The intercept value offers a valuable baseline reference, and the identification of CRIM_RATE as a potential candidate for removal highlights the importance of variable selection in refining our predictive model.

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.

Ans:

a) Output of the Model:

The intercept value is 29.42 signifying that if all independent variables are 0, then the value of the house would be 29.42. This model explains 69.36% of the variance in AVG_PRICE All variables are significant here. This model is acceptable as it has a decent R-square and all variables are significant.

b) Model Performance:

Adjusted R-square for this model is 0.6887 vs 0.6883 in the previous model. Although adjusted R-square value is not up drastically, but we have all significant variables here, so we consider these two factors together, then this model is a better model than the previous one.

c) NOX and AVG_Price are negatively related. If the value of NOX increases, then value of AVG_PRICE falls, more specifically every 1- unit increase in the value of NOX decreased the value of AVG_PRICE by 10.27.

d) Regression Equation:

AVG_PRICE = 29.4285 - 10.2727*NOX - 1.0717*PTRATIO - 0.6052* LSTAT

- 0.0145*TAX + 0.0329*AGE + 0.1307*INDUS + 0.2615*DISTANCE

+ 4.1255*AVG_ROOM

CONCLUSION:

In our analysis of Boston's housing market, we established analytical models that explain an important portion of house price differences. Significant variables like average room total and the percentage of lower-status population showed to be important interpreters. Our models offer valuable insights for pricing results, and stakeholders can advantage from considering these features. However, continuing investigation and model modification are key to staying competitive in the dynamic real estate market.