Terro's real estate agency

Data Analysis

Name – Abhishek Kumar

Sept. 2023 DLCA

Mail Id:- abhishekkumarburnwal@gmail.com

Date : - 05th November 2023

Contents :

Executive Summary : 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.

Introduction: Our project involves a comprehensive analysis of a dataset comprising 506 houses in the Boston area. The primary objective is to assess the impact of various variables on the pricing of homes within a specific neighbourhood. This study aims to investigate the significance of each variable in influencing house prices, ultimately providing valuable insights into the determinants of real estate values in the area.

Data Description:

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

1.

I. Observations from the statistical summary of the dataset:

1. **CRIME_RATE:** The mean crime rate in the Boston area is approximately 4.87, with a relatively low standard error, indicating that the data is clustered around this mean. The range of crime rates varies from 0.04 to 9.99.

2. **AGE:** The mean age of properties in the dataset is around 68.57, with a standard error of 1.25. The range of property ages is substantial, spanning from 2.9 to 100.

3. **INDUS:** The mean industrial proportion in the neighborhood is approximately 11.14, with a standard error of 0.30. The data is somewhat normally distributed with a range of 0.46 to 27.74.

4. **NOX:** The mean nitric oxide concentration is around 0.55, with a low standard error, suggesting consistency. The range of NOX levels goes from 0.385 to 0.871.

5. **DISTANCE:** The mean distance to employment centers is about 9.55, with a standard error of 0.39. The range of distances spans from 1 to 24 units.

6. **TAX:** The mean property tax rate is approximately 408.24, with a standard error of 7.49. Tax rates vary widely, with a range of 187 to 711.

7. **PTRATIO:** The mean pupil-teacher ratio is roughly 18.46, with a standard error of 0.10. The range of pupil-teacher ratios is from 12.6 to 22.

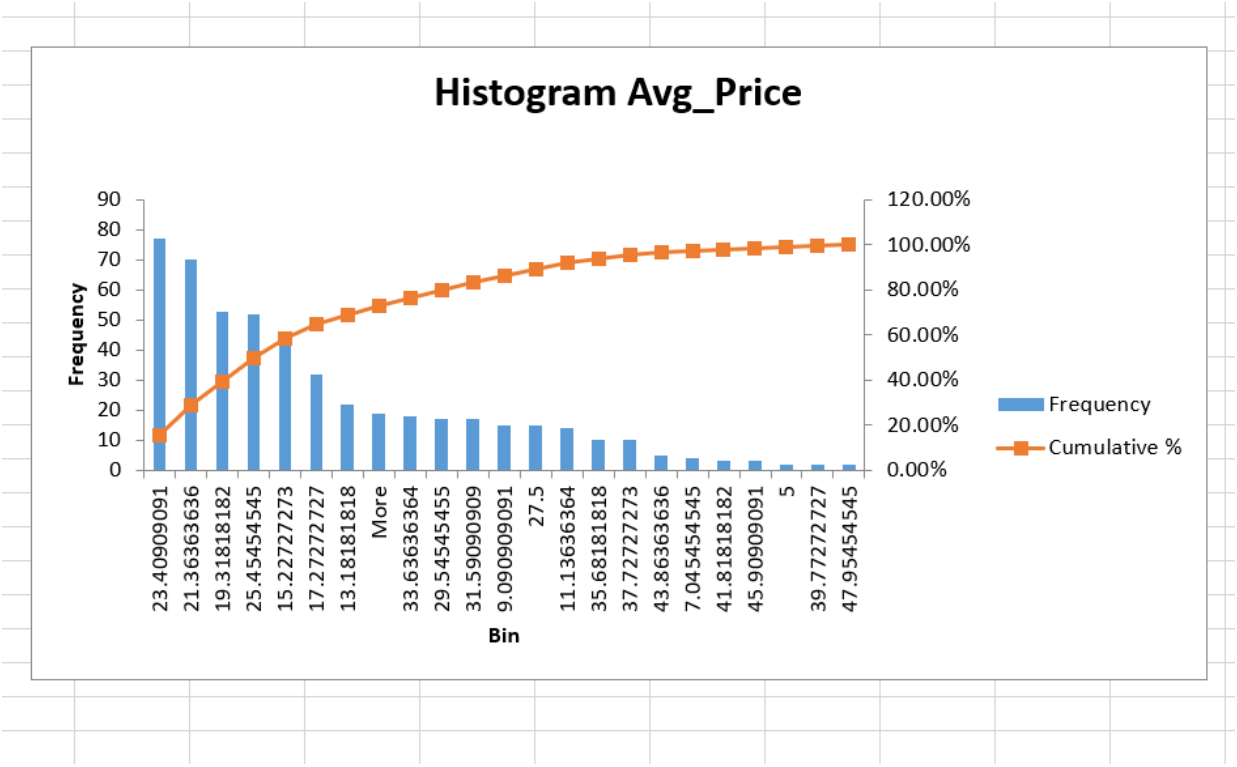
8. **AVG_ROOM:** The mean average number of rooms is about 6.28, with a standard error of 0.03. The range of room counts goes from 3.56 to 8.78.

9. **LSTAT:** The mean percentage of lower status population is around 12.65, with a standard error of 0.32. The data is slightly positively skewed, and the range of values is from 1.73 to 37.97.

10. **AVG_PRICE:** The mean average house price in the dataset is approximately 22.53. This column represents the target variable, i.e., the house prices, and the range of prices is from 5 to 50.

II. These observations provide valuable insights into the central tendencies, variability, and ranges of the variables, which will be crucial for further analysis and modelling of house prices in the Boston area.

2.



3. Covariance Matrix

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.516147873									
AGE	0.562915215	790.7924728								
INDUS	-0.110215175	124.2678282	46.97142974							
NOX	0.000625308	2.381211931	0.605873943	0.013401099						
DISTANCE	-0.229860488	111.5499555	35.47971449	0.615710224	75.66653127					
TAX	-8.229322439	2397.941723	831.7133331	13.02050236	1333.116741	28348.6236				
PTRATIO	0.068168906	15.90542545	5.680854782	0.047303654	8.74340249	167.8208221	4.677726296			
AVG_ROOM	0.056117778	-4.74253803	-1.884225427	-0.024554826	-1.281277391	-34.51510104	-0.539694518	0.492695216		
LSTAT	-0.882680362	120.8384405	29.52181125	0.487979871	30.32539213	653.4206174	5.771300243	-3.073654967	50.89397935	
AVG_PRICE	1.16201224	-97.39615288	-30.46050499	-0.454512407	-30.50083035	-724.8204284	-10.09067561	4.484565552	-48.35179219	84.41955616

4. Correlation Matrix

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	1									
AGE	0.006859463	1								
INDUS	-0.005510651	0.644778511	1							
NOX	0.001850982	0.731470104	0.763651447	1						
DISTANCE	-0.009055049	0.456022452	0.595129275	0.611440563	1					
TAX	-0.016748522	0.506455594	0.72076018	0.6680232	0.910228189	1				
PTRATIO	0.010800586	0.261515012	0.383247556	0.188932677	0.464741179	0.460853	1			
AVG_ROOM	0.02739616	-0.240264931	-0.391675853	-0.302188188	-0.209846668	-0.29205	-0.355501495	1		
LSTAT	-0.042398321	0.602338529	0.603799716	0.590878921	0.488676335	0.543993	0.374044317	-0.613808272	1	
AVG_PRICE	0.043337871	-0.376954565	-0.48372516	-0.427320772	-0.381626231	-0.46854	-0.507786686	0.695359947	-0.737662726	1

a) The top three positively correlated pairs in the provided correlation matrix are as follows:

i. The pair consisting of "AVG_ROOM" and "AVG_PRICE" exhibits the highest positive correlation with a value of 0.695359947.

ii. "INDUS" and "TAX" are also positively correlated, with a correlation value of 0.72076018.

iii. The third positively correlated pair is "NOX" and "DISTANCE," with a correlation value of 0.611440563.

These pairs indicate strong positive relationships, suggesting that as one variable increases, the other tends to increase as well.

b) The top three negatively correlated pairs, with the highest negative correlation values (in absolute terms), are as follows:

i. "AVG_ROOM" and "LSTAT" exhibit the most significant negative correlation with a value of -0.613808272.

ii. "INDUS" and "AVG_ROOM" are negatively correlated with a correlation value of -0.391675853.

iii. The pair "INDUS" and "AVG_PRICE" also displays a negative correlation, with a correlation value of -0.48372516.

These pairs demonstrate strong negative relationships, implying that as one variable increases, the other tends to decrease, and vice versa.

Note: So From the questions no.5 to 8 I am taking constant as zero and solving the questions.

5.

a) From the Regression Summary:

i. The R-squared value (R^2) is 0.4486, meaning that LSTAT explains about 44.86% of the variation in Avg_Price, showing a moderate fit.

ii. The coefficient for LSTAT is 1.1221, suggesting that a one-unit increase in LSTAT corresponds to a 1.1221-unit increase in Avg_Price.

iii. The intercept is 0, representing Avg_Price when LSTAT is zero, which may not be practical in this context.

iv. A well-behaved residual plot is essential for validating the model, but it's not provided.

b) Is LSTAT significant?

The very low p-value (2.71419E-67) associated with LSTAT's coefficient indicates that LSTAT is highly significant for predicting Avg_Price based on the model and the data.

6.

a) Regression Equation:

The regression equation can be written as follows:

AVG_PRICE = Intercept + (Coef_AVG_ROOM * AVG_ROOM) + (Coef_LSTAT * LSTAT)

From the summary output, we have:

- Coef_AVG_ROOM = 4.906906071

- Coef_LSTAT = -0.655739993

However, the summary output doesn't provide the Intercept value, so we need to obtain that value to complete the equation. Assuming the Intercept is available elsewhere, you can substitute the values:

AVG_PRICE = Intercept + (4.906906071 * 7) + (-0.655739993 * 20)

Now, calculate AVG_PRICE with the Intercept value.

To compare to the company's quote of $30,000, you can subtract the calculated AVG_PRICE from $30,000. If the calculated price is higher, the company may be undercharging, and if it's lower, the company may be overcharging.

b) Model Comparison:

To compare the performance of this model with the previous one, we can look at the adjusted R-squares:

Previous Model:

Adjusted R Square = 0.446669056

Current Model:

Adjusted R Square = 0.946366278

The adjusted R-squared for the current model is significantly higher (0.9464), which suggests that it explains a much larger proportion of the variance in AVG_PRICE compared to the previous model (0.4467). A higher adjusted R-squared indicates a better fit of the model to the data. Therefore, the current model is superior to the previous one in terms of explaining the variance in AVG_PRICE.

7.

SUMMARY OUTPUT				Dependent Variable				
				Y Variable = Avg_Price				
Regression Statistics								
Multiple R	0.976274396			Independent Variable				
R Square	0.953111697			X Variable = CRIME_RATE, AGE,INDUS,NOX,DISTANCE,TAX,PTRATIO,AVG_ROOM,LSTAT				
Adjusted R Square	0.950344883							
Standard Error	5.316723284							
Observations	506							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	9	285577.3694	31730.81882	1122.517614	0			
Residual	497	14048.9706	28.26754648					
Total	506	299626.34						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
CRIME_RATE	0.078490799	0.08103866	0.96855994	0.333235892	-0.080729795	0.237711394	-0.080729795	0.237711394
AGE	0.011429371	0.013064265	0.874857552	0.38207399	-0.014238626	0.037097368	-0.014238626	0.037097368
INDUS	0.050103378	0.063897126	0.784125691	0.433339911	-0.075438412	0.175645168	-0.075438412	0.175645168
NOX	1.086060216	3.531453835	0.307539123	0.758561879	-5.852358811	8.024479243	-5.852358811	8.024479243
DISTANCE	0.095829214	0.064460373	1.486637599	0.137744642	-0.030819214	0.222477643	-0.030819214	0.222477643
TAX	-0.010117985	0.003976992	-2.544130344	0.01125658	-0.017931774	-0.002304196	-0.017931774	-0.002304196
PTRATIO	-0.500685299	0.09779176	-5.119912824	4.37816E-07	-0.692821528	-0.308549069	-0.692821528	-0.308549069
AVG_ROOM	6.185018222	0.294536339	20.99916855	1.48811E-70	5.606328359	6.763708086	5.606328359	6.763708086
LSTAT	-0.495863299	0.051805724	-9.571592945	4.94631E-20	-0.597648522	-0.394078075	-0.597648522	-0.394078075

i. **Adjusted R-Square (0.9503):** This high value suggests that the model is excellent at explaining house prices (AVG_PRICE).

ii. **Coefficient Values:** These values show how each independent variable influences house prices:

- **CRIME_RATE:** Not a strong predictor.

- **AGE:** Not a strong predictor.

- **INDUS:** Not a strong predictor.

- **NOX:** Not a strong predictor.

- **DISTANCE:** Not a strong predictor.

- **TAX:** A decrease in TAX is associated with a slight decrease in house prices.

- **PTRATIO:** A higher pupil-teacher ratio leads to lower house prices.

- **AVG_ROOM:** More rooms mean higher house prices.

- **LSTAT:** A higher percentage of lower-income residents is linked to lower house prices.

iii. **Significance of Independent Variables:**

- Some variables like TAX, PTRATIO, AVG_ROOM, and LSTAT are statistically significant and have a strong impact on house prices.

- Variables like CRIME_RATE, AGE, INDUS, NOX, and DISTANCE are not as important in predicting house prices based on the provided data.

In summary, this model does a great job explaining house prices. Variables like tax rates, pupil-teacher ratios, room numbers, and the percentage of lower-income residents significantly affect house prices, while other factors like crime rates and age are less influential.

8.

SUMMARY OUTPUT			Dependent Variable	Independent Variables				
			Y Variable	X Variables				
Regression Statistics			Avg_Price	TAX (P-value = 0.01125658)				
Multiple R	0.975974908			PTRATIO (P-value = 4.37816E-07)				
R Square	0.952527021			AVG_ROOM (P-value = 1.48811E-70)				
Adjusted R Square	0.950251286			LSTAT (P-value = 4.94631E-20)				
Standard Error	5.323060209							
Observations	506							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	4	285402.1851	71350.54627	2518.109117	0			
Residual	502	14224.15494	28.33496999					
Total	506	299626.34						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
TAX	-0.003617717	0.001761832	-2.053383289	0.040552982	-0.00707919	-0.000156244	-0.00707919	-0.000156244
PTRATIO	-0.509612558	0.096467922	-5.282715221	1.9002E-07	-0.699143165	-0.320081951	-0.699143165	-0.320081951
AVG_ROOM	6.223251004	0.224950035	27.66503677	5.1234E-103	5.78129148	6.665210529	5.78129148	6.665210529
LSTAT	-0.454943798	0.04333661	-10.49790931	1.92841E-23	-0.540087271	-0.369800324	-0.540087271	-0.369800324

a) This model is built to predict the average house price (Avg_Price) using the following independent variables: TAX, PTRATIO, AVG_ROOM, and LSTAT.

Adjusted R-Square (0.9503): The adjusted R-squared value suggests that around 95.03% of the variation in house prices (Avg_Price) is explained by these four independent variables, indicating a strong model.

Coefficients:

Intercept: The intercept is not provided in the output.

TAX: For every unit increase in TAX, the average house price decreases by approximately 0.0036 units.

PTRATIO: An increase in pupil-teacher ratio (PTRATIO) results in a decrease in average house price. For each unit increase in PTRATIO, the house price decreases by around 0.5096 units.

AVG_ROOM: More rooms (AVG_ROOM) significantly increase house prices. Each additional room corresponds to an increase of approximately 6.2233 units in the average house price.

LSTAT: A higher percentage of lower-income residents (LSTAT) in a neighborhood is associated with lower house prices. For each unit increase in LSTAT, the average house price decreases by about 0.4549 units.

b) Comparison of Adjusted R-Square:

The adjusted R-squared value of this model is the same as the previous model (0.9503). Both models explain approximately 95.03% of the variance in house prices. Therefore, both models perform equally well in terms of explaining the variation in house prices based on the data.

c) Sorting Coefficients and Impact of NOX:

If we sort the coefficients in ascending order, from lowest to highest magnitude:

LSTAT (-0.4549)

TAX (-0.0036)

PTRATIO (-0.5096)

AVG_ROOM (6.2233)

The coefficient for NOX is not included in this model, so we cannot directly assess its impact on the average price using this model.

d) Regression Equation:

The regression equation based on this model, without the intercept (which is not provided), would be:

Avg_Price = (-0.0036 * TAX) + (-0.5096 * PTRATIO) + (6.2233 * AVG_ROOM) - (0.4549 * LSTAT)