TERRO’S REAL ESTATE AGENCY

Business Report

Abstract

”This report investigates the impact of various factors on house pricing in a Boston locality for Terro's real-estate agency”

Mahamad Jameer Makandar

Auditor

Table of Contents

Problem Statement and Objective:

Situation:

Terro's Real-Estate, a prominent real estate agency, specializes in estimating property prices within specific localities. They rely on a comprehensive analysis of various property features and factors to determine the market value of a property. To facilitate this assessment, the agency employs an "Auditor" who conducts in-depth studies of various geographic features that may impact property prices. These features encompass parameters such as crime rates, pollution levels (NOX), education facilities (pupil to teacher ratio), connectivity (distance from the highway), and more.

The ability to accurately assess these factors aids in the precise pricing of properties, thereby enabling Terro's Real-Estate to provide valuable insights to their clients.

Objective:

As the appointed Auditor for Terro's Real-Estate, the primary task is to analyse the magnitude and relevance of each variable in the dataset and understand their potential impact on the pricing of houses within a particular locality. The dataset in question comprises 506 houses in Boston, and each house is characterized by several attributes, including:

Data Dictionary:

- 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: Percentage of the lower status of the population.
- AVG_PRICE: Average value of houses in $1000's.

1. Summary Statistics for Each Variable (Using Data Tool Pack):

Observations:

- **CRIME_RATE:** The average crime rate of 4.87 incidents per unit, exhibiting significant variability (standard deviation of approximately 2.92), with crime rates ranging from 0.04 to 9.99, highlighting a wide diversity of crime levels.

- **AGE:** This represents a crucial average age of approximately 68.57, with a substantial age variation (standard deviation around 28.15), and an extensive age range from 2.9 to 100 (minimum to maximum), showcasing a diverse spectrum of ages within the population.

- **INDUSTRY:** The average industrial proportion of 11.14, with variability (standard deviation ~6.86), and a relatively flattened distribution (negative kurtosis -1.23), indicating the extent and characteristics of industrial activity.

- **NOX:** The mean NOX concentration of 0.5547, with consistent levels (low standard deviation ~0.116), a right-skewed distribution (positive skewness 0.73), and a significant range from 0.385 to 0.871 (minimum to maximum), facilitating assessment of air quality disparities.
  This range is essential for evaluating air quality disparities across different areas or time periods.

- **DISTANCE:** The dataset presents a critical mean distance to employment centres of 9.55 miles, reflecting transportation and accessibility factors, along with notable variability (standard deviation ~8.71) and a range extending from 1 to 24 miles, highlighting accessibility disparities.

- **TAX:** It indicates a crucial average property tax rate of 408.24, accompanied by significant variation (substantial standard deviation ~168.54) and a wide range spanning from 187 to 711, showcasing diverse taxation levels for property owners.

- **PTRATIO:** Influential mean pupil-teacher ratio of 18.46 impacting education quality, accompanied by modest variability (standard deviation ~2.16) and a positively skewed distribution (positive kurtosis 1.89) with heavier tails.

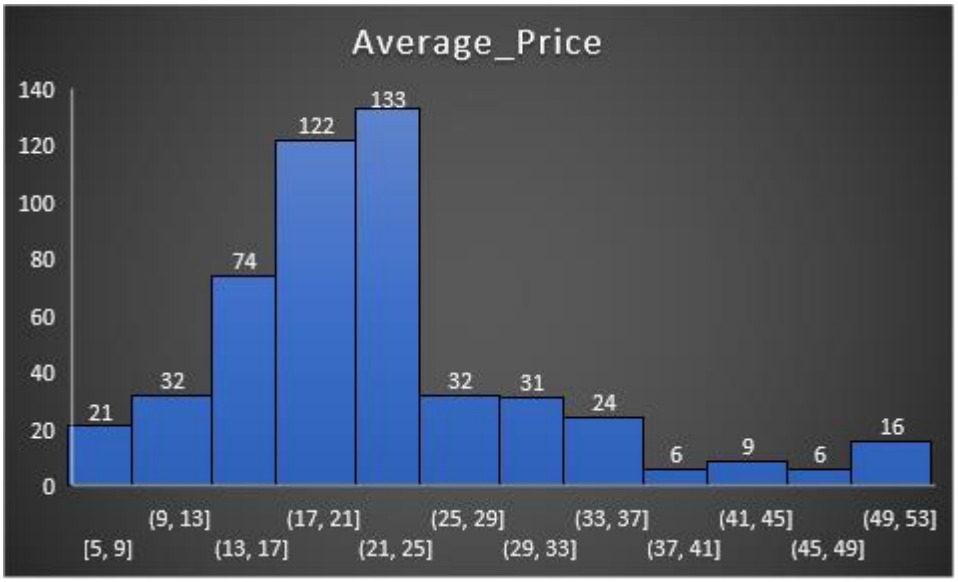
- **AVG_ROOM:** The mean **average number of rooms per dwelling is 6.28**, a critical housing metric.

  The right-skewed distribution (positive skewness 0.40) for the average number of rooms per dwelling, indicating a tendency towards larger dwellings.

- **LSTAT:** The **mean percentage of lower-status population is 12.65**, influencing socioeconomic characteristics.
  - **Standard Deviation:** A **standard deviation of approximately 7.14** indicates variability in socioeconomic status.
  - **Kurtosis: Positive kurtosis (0.49)** suggests a distribution with heavier tails.

- **AVG_PRICE:** The **mean average house price is $22.53**, a critical indicator for the real estate market.
  - **Standard Deviation:** With a **standard deviation of around $9.20**, there's variability in house prices.
  - **Skewness: Positive skewness (1.11)** indicates a right-skewed distribution.

2. **Histogram of the Avg_Price variable:**



The interval (21-25] displayed the highest frequency count, with 133 data points falling within this range. This finding suggests a substantial concentration of properties with average prices falling into this specific bin, potentially indicating a significant category.

3. Covariance Matrix:

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	8.52									
AGE	0.56	790.79								
INDUS	-0.11	124.27	46.97							
NOX	0.00	2.38	0.61	0.01						
DISTANCE	-0.23	111.55	35.48	0.62	75.67					
TAX	-8.23	2397.94	831.71	13.02	1333.12	28348.62				
PTRATIO	0.07	15.91	5.68	0.05	8.74	167.82	4.68			
AVG_ROOM	0.06	-4.74	-1.88	-0.02	-1.28	-34.52	-0.54	0.49		
LSTAT	-0.88	120.84	29.52	0.49	30.33	653.42	5.77	-3.07	50.89	
AVG_PRICE	1.16	-97.40	-30.46	-0.45	-30.50	-724.82	-10.09	4.48	-48.35	84.42

A covariance matrix provides information on how multiple variables change together but can be challenging to interpret due to its dependence on variable units. For more meaningful insights, analysts often prefer using correlation matrices, where correlation coefficients offer standardized, unit-independent measures of the strength and direction of linear relationships between variables.

4. Corelation Matrix:

	CRIME_ RATE	*AGE*	*INDUS*	*NOX*	*DIS -TANCE*	*TAX*	*PT RATIO*	*AVG_ ROOM*	*LSTAT*	*AVG_ PRICE*
CRIME_ RATE	1									
AGE	0.00686	1								
INDUS	-0.00551	0.644779	1							
NOX	0.001851	0.73147	0.763651	1						
DISTANCE	-0.00906	0.456022	0.595129	0.611441	1					
TAX	-0.01675	0.506456	0.72076	0.668023	0.910228	1				
PTRATIO	0.010801	0.261515	0.383248	0.188933	0.464741	0.460853	1			
AVG_ROOM	0.027396	-0.24026	-0.39168	-0.30219	-0.20985	-0.29205	-0.3555	1		
LSTAT	-0.0424	0.602339	0.6038	0.590879	0.488676	0.543993	0.374044	-0.61381	1	
AVG_PRICE	0.043338	-0.37695	-0.48373	-0.42732	-0.38163	-0.46854	-0.50779	0.69536	-0.7376	1

a. 3 positively correlated pairs:

- **TAX vs. DISTANCE:** The correlation coefficient is approximately **0.91023**, indicating a very **strong positive** linear relationship.
- **NOX vs. INDUS:** The correlation coefficient is approximately **0.76365**, indicating a **strong positive** linear relationship.
- **NOX vs. AGE:** The correlation coefficient is approximately **0.73147**, indicating a **strong positive** linear relationship.

b. 3 negatively correlated pairs:

- **AVG_PRICE vs. LSTAT:** The correlation coefficient is approximately -**0.73766, indicating a strong negative** linear relationship.
- **LSTAT vs. AVG_ROOM:** The correlation coefficient is approximately -**0.61381, indicating a strong negative** linear relationship.
- **AVG_PRICE vs. PTRATIO:** The correlation coefficient is approximately **-0.50779, indicating a moderate negative linear** relationship.

5. Initial Regression Model With AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as Independent Variable:

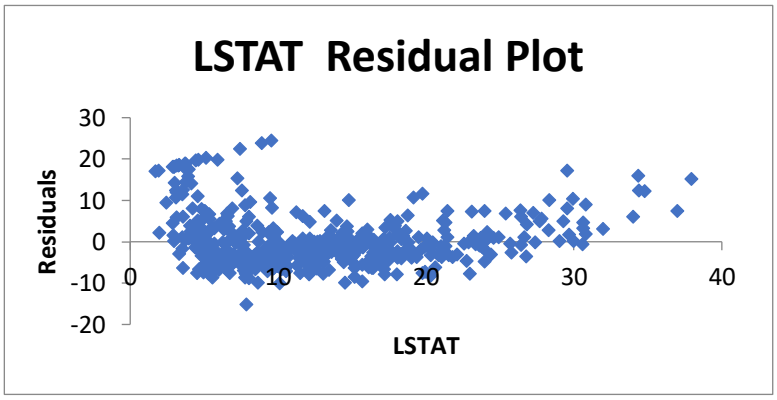
Regression Statistics	
Multiple R	0.737662726
R Square	0.544146298
Adjusted R Square	0.543241826
Standard Error	6.215760405
Observations	506

The R-squared (R^2) value is used to assess model improvement. In our case, it's 0.5441, indicating that approximately 54.41% of the variation in the

dependent variable is explained by the independent variables in the regression model.

We consider an R-squared greater than 50% as a good improvement, but the threshold may vary from one company to another; some companies might require an R-squared of 80% or higher to consider the model strong.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	34.55384088	0.562627355	61.41514552	3.7E-236	33.44845704	35.65922	33.44846	35.65922
LSTAT	-0.950049354	0.038733416	-24.5278999	5.08E-88	-1.0261482	-0.87395	-1.02615	-0.87395



a. Regression Summary output:

- **Variance:** The R-squared (R^2) value of approximately 0.5441 suggests that about 54.41% of the variance in the dependent variable is explained by the model.

- **Coefficient Value:** The coefficient for the Intercept is approximately 34.55, representing the estimated value of the dependent variable when all independent variables are zero. The coefficient for LSTAT is approximately -0.9501, indicating that for each unit increase in LSTAT, the dependent variable is estimated to decrease by approximately 0.9501 units.

- **Intercept:** The intercept of 34.55 signifies the estimated value of the dependent variable when all independent variables are zero.

- **Residual Plot:** A residual plot assesses a regression model's fit. In this plot for LSTAT and Average_Price shows no patterns or trends, it indicates the model's assumptions are met, suggesting a good fit between these variables.

b. Assessing the Significance of the LSTAT Variable:

A p-value of 0.05 (or 5%) is a common significance level used in statistical tests; if the p-value is less than 0.05, it's often considered statistically significant, indicating strong evidence against the null hypothesis.

For the above model the p-valve **of LSTAT is "5.0811E-88,"** which is very close to zero and which is **also less than 0.05(5%)**, For Our model, the LSTAT variable appears to be **significant for the analysis**.

6. Regression model including LSTAT and AVG_ROOM together as Independent variables and AVG_PRICE as dependent variable:

Regression Statistics	
Multiple R	0.7991
R Square	**0.638562**
Adjusted R Square	0.637124
Standard Error	5.540257
Observations	506

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.35827	3.172828	-0.4281	0.6687649	-7.5919	4.875355	-7.5919	4.875355
AVG_ROOM	5.094788	0.444466	11.46273	3.472E-27	4.22155	5.968026	4.22155	5.968026
LSTAT	-0.64236	0.043731	-14.6887	6.669E-41	-0.72828	-0.55644	-0.72828	-0.55644

a. The Regression equation for the above model is can be written as;

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\ldots\ldots+\beta_nX_n$$

Where:
Y = Dependent Variable
β_0 = Intercept
$\beta_1, \beta_2 \ldots \beta_n$=coefficients associated with the independent variables
$X_1, X_2 \ldots .X_n$ = are the independent variables (AVG_ROOM and LSTAT in the question), $X_1 = 7,\ X_1 = 20$

$$Y = -1.358272812 + (5.094787984 * X_1) - (0.642358334 * X_2)$$

$$Y = -1.358272812 + (5.094787984 * 7) - (0.642358334 * 20)$$

$$\underline{Y = 21.45}$$

The predicted AVG_PRICE for a house with 7 rooms and an L-STAT value of 20 is approximately 21.46.

The company's quote of 30,000 USD for this locality is reasonable, we compare it to the predicted price. The **company's quote is much higher (30,000 USD)** than the **predicted price (approximately 21.46 USD)**.

Therefore, based on this regression model, it appears that the company is **overcharging for houses in this locality**.

b. Comparing the Adjusted R Square:

The Adjusted R Square values tell us how well our model fits the data(or Improvement). A higher Adjusted R Square (0.6371) means our model does a better job of explaining the house prices compared to a lower value (0.5432). This improvement likely came from using additional factors or making adjustments in the model to make it more accurate.

7. AVG_PRICE alone be the Dependent Variable and all the other variables are independent:

Regression Statistics	
Multiple R	0.83297882
R Square	0.69385372
Adjusted R Square	**0.68829865**
Standard Error	5.1347635
Observations	506

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.2413153	4.8171256	6.070283	2.54E-09	19.77682784	38.7058	19.77683	38.7058
CRIME_RATE	0.04872514	0.0784186	0.621346	0.5346572	-0.105348544	0.202799	-0.10535	0.202799
AGE	0.03277069	0.0130978	2.501997	0.0126704	0.00703665	0.058505	0.007037	0.058505
INDUS	0.1305514	0.0631173	2.068392	0.0391209	0.006541094	0.254562	0.006541	0.254562
NOX	-10.3211828	3.8940363	-2.65051	0.0082939	-17.97202279	-2.67034	-17.972	-2.67034
DISTANCE	0.26109357	0.0679471	3.842603	0.0001375	0.127594012	0.394593	0.127594	0.394593
TAX	-0.01440119	0.0039052	-3.68774	0.0002512	-0.022073881	-0.00673	-0.02207	-0.00673

PTRATIO	-1.07430535	0.1336017	-8.0411	6.586E-15	-1.336800438	-0.81181	-1.3368	-0.81181
AVG_ROOM	4.12540915	0.442759	9.317505	3.893E-19	3.255494742	4.995324	3.255495	4.995324
LSTAT	-0.60348659	0.0530812	-11.3691	8.911E-27	-0.70777824	-0.49919	-0.70778	-0.49919

- **Adjusted R Square** (0.6883): Indicates that 68.83% of house price variability is explained by the model, suggesting its effectiveness. A higher Adjusted R Square suggests a good fit, meaning the model is effective in explaining variations in house prices.

- The **coefficients associated with each independent variable** provide insights into their impact on house prices.

  Eg: A positive coefficient for AVG_ROOM (4.1254) suggests that an increase in the average number of rooms tends to increase house prices.

  Conversely, a negative coefficient for LSTAT (-0.6035) implies that an increase in the percentage of lower-income residents tends to decrease house prices.

- The intercept(29.2413) represents the estimated house price when all independent variables are zero. In this context, it's the baseline price of a house when other factors are not considered.

As seen in the table above, all variables are significant (p-value < 0.05 or 5%), except for Crime_Rate (0.5346572), which is not statistically significant.

8. Creating a Regression Model with Significant Variables:

Regression Statistics	
Multiple R	0.832835773
R Square	**0.693615426**
Adjusted R Square	0.688683682
Standard Error	5.131591113
Observations	506

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.42847349	4.804729	6.124898	1.85E-09	19.98839	38.86856	19.98839	38.86856
AGE	0.03293496	0.013087	2.516606	0.012163	0.007222	0.058648	0.007222	0.058648
INDUS	0.130710007	0.063078	2.072202	0.038762	0.006778	0.254642	0.006778	0.254642
NOX	-10.27270508	3.890849	-2.64022	0.008546	-17.9172	-2.62816	-17.9172	-2.62816
DISTANCE	0.261506423	0.067902	3.851242	0.000133	0.128096	0.394916	0.128096	0.394916
TAX	-0.014452345	0.003902	-3.70395	0.000236	-0.02212	-0.00679	-0.02212	-0.00679
PTRATIO	-1.071702473	0.133454	-8.03053	7.08E-15	-1.33391	-0.8095	-1.33391	-0.8095
AVG_ROOM	4.125468959	0.442485	9.3234	3.69E-19	3.256096	4.994842	3.256096	4.994842
LSTAT	-0.605159282	0.05298	-11.4224	5.42E-27	-0.70925	-0.50107	-0.70925	-0.50107

a. Output of this Model:

The **R-squared (R^2)** value for the provided regression model is **approximately 0.6936, indicating that about 69.36%** of the variability in house prices is explained by the significant independent variables in the model.

b. Comparing the Adjusted R-Square with the previous Adjusted R-Square Value:

The **second model with an adjusted R-squared value of 0.688683682 performs slightly better than the first model with an adjusted R-squared value of 0.688298647**. This suggests that the second model, which includes the same significant independent variables as the first model, explains a slightly higher proportion of the variability in house prices.

c. Sorting coefficients in ascending order and checking e average price if the value of NOX is more:

Variables	*Coefficients*
NOX	-10.27270508
PTRATIO	-1.071702473
LSTAT	-0.605159282
TAX	-0.014452345
AGE	0.03293496
INDUS	0.130710007
DISTANCE	0.261506423
AVG_ROOM	4.125468959
Intercept	29.42847349

The coefficient **for NOX is -10.27270508**, which is the **most negative among these coefficients**. In the context of the regression model, this means that if the value of NOX is higher in a locality within this town, it would lead to a decrease in the average price of houses in that locality. In simpler terms, higher levels of NOX are associated with lower house prices in the model.

d. Regression Equation:

$$\boldsymbol{Y=\beta_0+\beta_1X_1+\beta_2X_2+\ldots\ldots+\beta_8X_8}$$

Where:

Y= AVG_PRICE (Dependent Variable)
β_0 = Intercept (29.42847349)
$\beta_1, \beta_2 \ldots \beta_n$=coefficients associated with the independent variables
$X_1, X_2, \ldots X_8$ = are the independent variables (Age, Indus, NOX, Distance, Tax, PTratio, Avg_Room and LSTAT)

AVG_PRICE=β_0+(β_1∗AGE)+(β_2∗INDUS)+(β_3∗NOX)+(β_4∗DISTANCE)+ (β_5∗TAX)+(β_6∗PTRATIO)+(β_7∗AVG_ROOM)+(β_8∗LSTAT)

AVG_PRICE=29.42847349+(0.03293496∗AGE)+(0.130710007∗INDUS)−(10 .27270508∗NOX)+(0.261506423∗DISTANCE)−(0.014452345∗TAX)−(1.0717 02473∗PTRATIO)+(4.125468959∗AVG_ROOM)−(0.605159282∗LSTAT)