Terro’s Real Estate Agency

Business Report

Prepared By:
Ezekiah Jabez

Ezekiah Jabez

ezekiahjabez2002@gmail.com

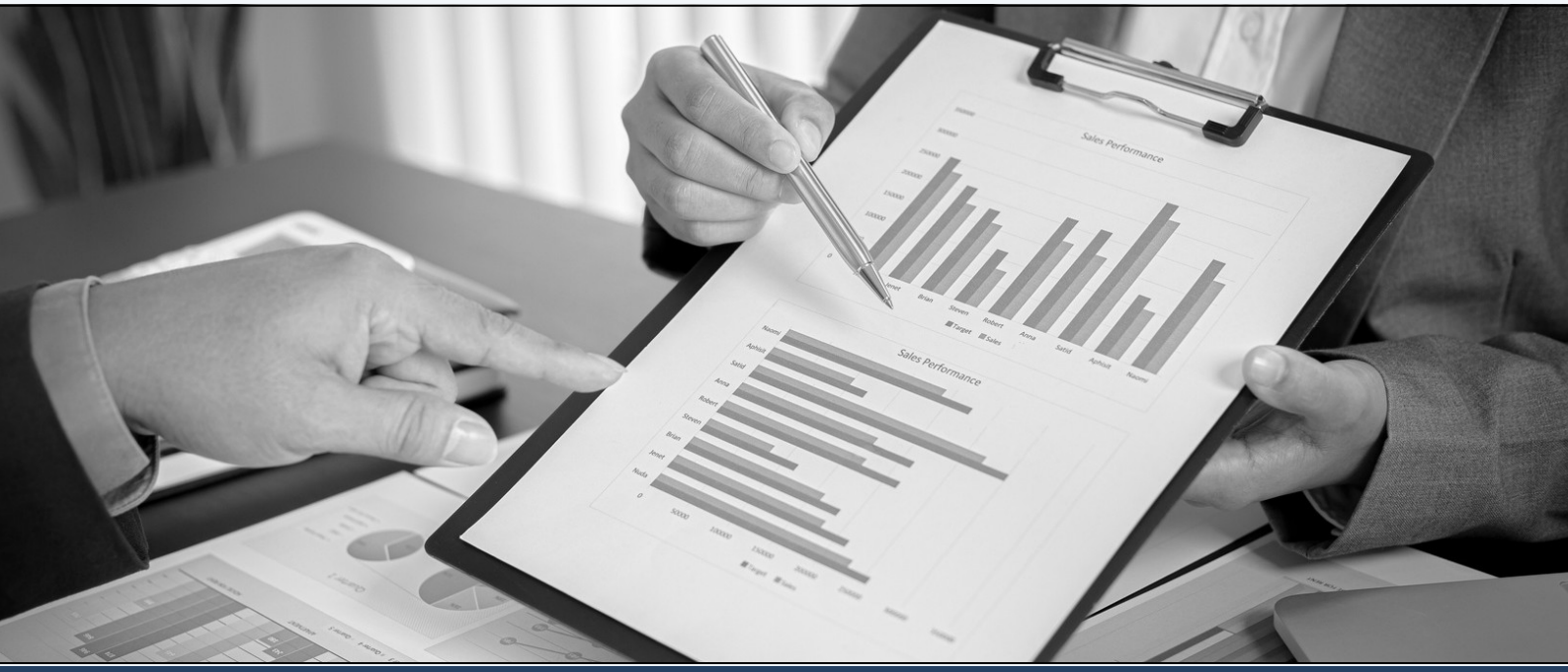


TABLE OF CONTENTS

INTRODUCTION

Terro's Real Estate is an agency that estimates the pricing of houses in a certain locality. They do this by analyzing different features/factors of a property, like pollution level (NOX), crime rate, education facilities (pupil-teacher ratio), and connectivity (distance from highway). This helps them determine the price of a property.

I have done exploratory data analysis and linear regression on this dataset, and I have observed many findings. For example, the crime rate and the proportion of lower-status people in the neighborhood are negatively correlated with house prices. The findings of this study will help Terro's Real Estate better understand the factors that affect house prices in Boston, and to develop more accurate pricing models.

DATA DICTIONARY

The Boston housing dataset is a well-known dataset that has been used to study the factors that affect house prices. The dataset includes 506 observations of houses in Boston, and it includes 13 variables that measure different aspects of the houses and their neighborhoods.

The data dictionary for the Boston housing dataset is as follows:

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

SUMMARY

This business report presents a comprehensive regression analysis aimed at understanding the factors that influence the average house price (AVG_PRICE) in a particular town. The analysis involved examining a dataset of various independent variables, such as crime rate, age, industrial land use, concentration of nitric oxides (NOX), distance to employment centers, property tax rate (TAX), pupil-teacher ratio (PTRATIO), average number of rooms (AVG_ROOM), and the percentage of lower-status population (LSTAT).

I've looked at a dataset of 506 houses and 13 different factors that could affect the price of a house, like the crime rate, the number of rooms, and the distance from the highway. We wanted to know which of these factors were most important in determining the price of a house.

KEY FINDINGS

- The average house price is $22,532. This is a relatively modest price, suggesting that Terro's real estate agency is giving in a relatively affordable price to live in.

- Houses in neighbourhoods with low crime rates and a high percentage of higher-status people are more expensive than houses in neighbourhoods with high crime rates and a low percentage of higher-status people. This is because people are willing to pay more for a house in a safe neighbourhood with good schools.

- Houses with a high number of rooms and that are located close to highways are more expensive than houses with a low number of rooms and that are located away from highways, because people want houses that are big enough for their families and that are convenient to get to work or school.

ANALYSIS

Objective (Task):

1) Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation. (5 marks)

Variable	Mean	Median	Mode	Standard Deviation	Variance	Kurtosis	Skewness	Range	Minimum	Maximum	Sum
CRIME_RATE	4.872	4.82	3.43	2.9211	8.5330	-1.1891	0.0217281	9.95	0.04	9.99	2465.22
AGE	68.57	77.5	100	28.1489	792.3584	-0.9677	-0.598963	97.1	2.9	100	34698.9
INDUSTRY	11.14	9.69	18.1	6.8604	47.0644	-1.2335	0.2950216	27.28	0.46	27.74	5635.21
NOX	0.555	0.538	0.538	0.1159	0.0134	-0.0647	0.7293079	0.486	0.385	0.871	280.6757
DISTANCE	9.549	5	24	8.7073	75.8164	-0.8672	1.0048146	23	1	24	4832
TAX	408.2	330	666	168.5371	28404.76	-1.1424	0.6699559	524	187	711	206568
PTRATIO	18.46	19.05	20.2	2.1649	4.6870	-0.2851	-0.802325	9.4	12.6	22	9338.5
AVG_ROOM	6.285	6.2085	5.713	0.7026	0.4937	1.8915	0.4036121	5.219	3.561	8.78	3180.025
LSTAT	12.65	11.36	8.05	7.1411	50.9948	0.4932	0.9064601	36.24	1.73	37.97	6402.45
AVG_PRICE	22.53	21.2	50	9.1971	84.5867	1.4952	1.1080984	45	5	50	11401.6

- The average crime rate is 4.87(mean), with a standard deviation of 2.92. The data shows a slightly positive skewness, indicates that most areas have relatively lower crime rates, but there are some areas with higher crime rates and the variation is notable as it ranges from 0.04 to 9.99.

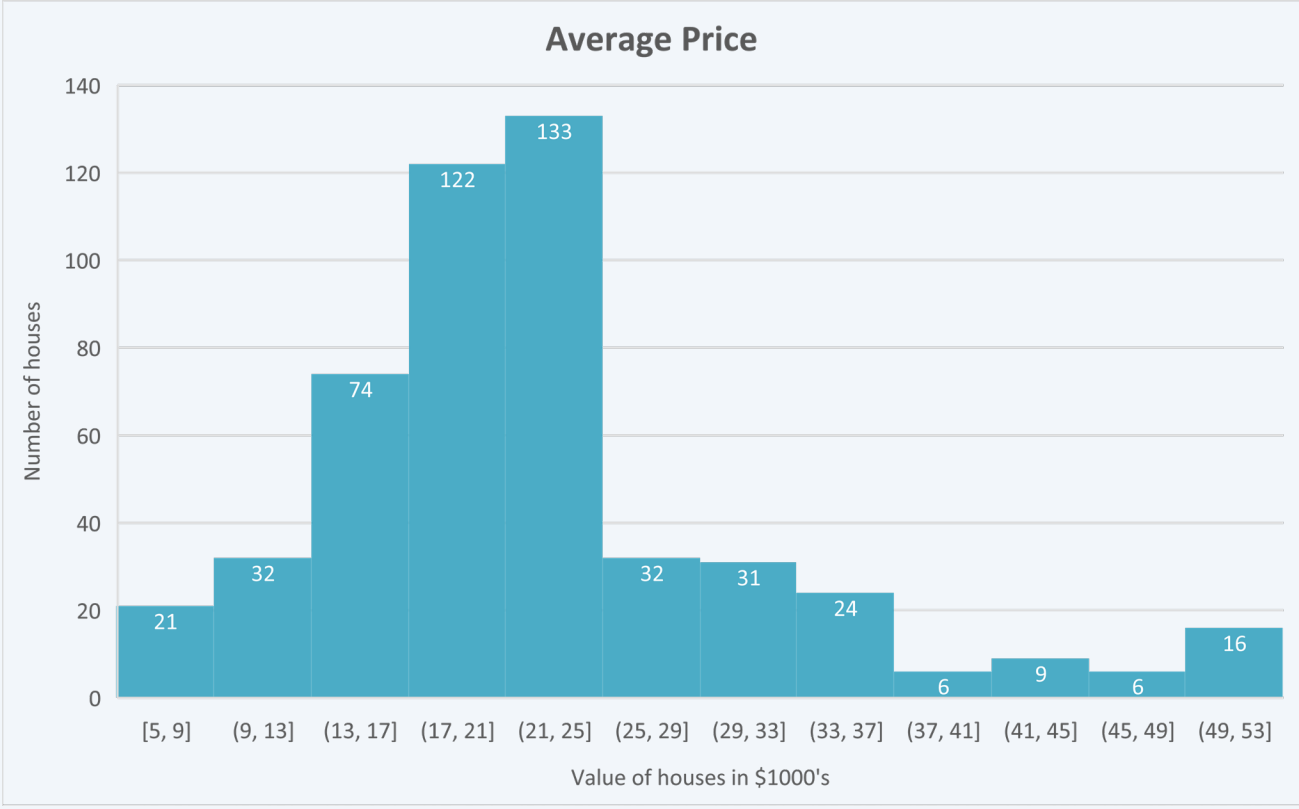
- The "AGE" variable represents the age of properties in the dataset. Average age of properties is 68.57, with a standard deviation of 28.15 which indicates variability in ages around the average. The distribution is slightly negatively skewed (-0.60), which says that most houses are older, but there are a few locations with newer properties. The range of ages is from 2.9 to 100, a mix of older and newer buildings.

- The "INDUS" variable represents the proportion of non-retail business acres per town, the median value of 9.69 is lower than the mean, and the mode of 18.1 is higher than both the mean and median. From this we can see that the data might be slightly positively skewed, indicating that more places have lower non-retail business proportions, and there are some with higher proportions.

- The average concentration of nitric oxides is 0.555, with a standard deviation of 0.116. The data has a positive skewness (0.72), indicating that most locations have lower nitric oxide concentrations, but some areas have higher concentrations. The kurtosis (-0.064) value suggests that extreme nitric oxide levels are less likely (i.e) the distribution of nitric oxide levels is platykurtic. The range from 0.385 to 0.871 shows some variation in air pollution levels.

- The average distance to employment centres is 9.55. The data shows positive skewness, which states that most locations are closer to employment centres, but a few places are more distant.
- The pupil-teacher's average ratio is 18.46, which means that there are 18.46 students for every teacher in the city. And, there is a standard deviation of 2.16, which means there is some variation in the pupil-teacher ratio across different places. The data shows negative skewness (-0.802), which means that the distribution is skewed to the right. This suggests that most places have higher pupil-teacher ratios, but there are a few locations with lower ratios.
- The percentage of lower-status population ranges from 1.73% to 37.97%, with an average of 12.65%. This says that there is diversity in socio-economic status. Some locations may be very wealthy, with a low percentage of lower-status population, while other places may be very poor, with a high percentage of lower-status population.
- The average house price is $22,530. The positive skewness value of 1.108 says that the data is skewed towards the lower end. This means that there are relatively more locations with lower house, with some locations having much higher prices. The range of $5,000 to $50,000 shows that there is a significant difference in the cost of housing.

2) Plot a histogram of the Avg_Price variable. What do you infer? (5 marks)



- There are a cluster of 21 houses between $5,000 and $9,000 suggests that there is a strong demand for affordable housing. This could be due to a number of factors.
- The affordable housing in this price range is likely to be in less desirable neighbourhoods, as it is priced lower. This mean that the homes are in need of repairs, or that they are located in areas with high crime rates or poor schools.
- Most of the houses are between $21,000 and $25,000, which is the average price. This suggests that there is a good market for mid-range housing.
- There are some minimum numbers of houses between $37,000 and $41,000 and another six houses between $45,000 and $49,000. This shows that there is a limited demand for high-end housing.
- There are only 16 houses between the highest prices of $49,000 and $53,000 suggests that there is a limited demand for high-end housing. This could be due to a number of factors, such as the high cost of living or limited availability of high-end housing.

3) Compute the covariance matrix. Share your observations. (5 marks)

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*
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

- The values in the covariance like (CRIME_RATE, AGE, INDUS, DISTANCE, NOX, and LSTAT) are mostly positive. And this shows positive co-variance.
- This shows that these variables will increase or decrease together. For example, when CRIME_RATE increases, AGE, INDUS, DISTANCE, NOX, and LSTAT also tend to increase, and vice versa.
- Some of the values like (TAX, PTRATIO, and AVG_ROOM) are mostly negative, which is negative co-variance. This indicates that these variables tend to have an inverse relationship. For example, when the tax rate increases, the pupil-teacher ratio and the average number of rooms per house tend to decrease. This is because higher tax rates are associated with lower-quality schools and smaller houses.
- Some values close to zero (NOX, AGE, AVG_ROOM and LSTAT) suggest weak or no relationship between these variables.

4) Create a correlation matrix of all the variables (Use Data analysis tool pack). (5 marks)
a) Which are the top 3 positively correlated pairs
b) Which are the top 3 negatively correlated pairs.

	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	0.460853035	1			
AVG_ROOM	0.02739616	-0.240264931	-0.391675853	-0.302188188	-0.209846668	-0.292047833	-0.355501495	1		
LSTAT	-0.042398321	0.602338529	0.603799716	0.590878921	0.488676335	0.543993412	0.374044317	-0.613808272	1	
AVG_PRICE	0.043337871	-0.376954565	-0.48372516	-0.427320772	-0.381626231	-0.468535934	-0.507786686	0.695359947	-0.737662726	1

- The cells which are highlighted in green indicates the top three positively correlated pairs. And they are TAX and DISTANCE (correlation coefficient = 0.9102), NOX and INDUS (correlation coefficient = 0.76360), NOX and AGE (correlation coefficient = 0.7314)

- The cells which are highlighted in red represent the top three negatively correlated pairs of variables. They are AVG_PRICE and LSTAT (correlation coefficient = -0.7376), LSTAT and AVG_ROOM (correlation coefficient = -0.6138), AVG_PRICE and PTRATIO (correlation coefficient = -0.5077)

5) Build an initial regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT variable an Independent Variable. Generate the residual plot. (8 marks)

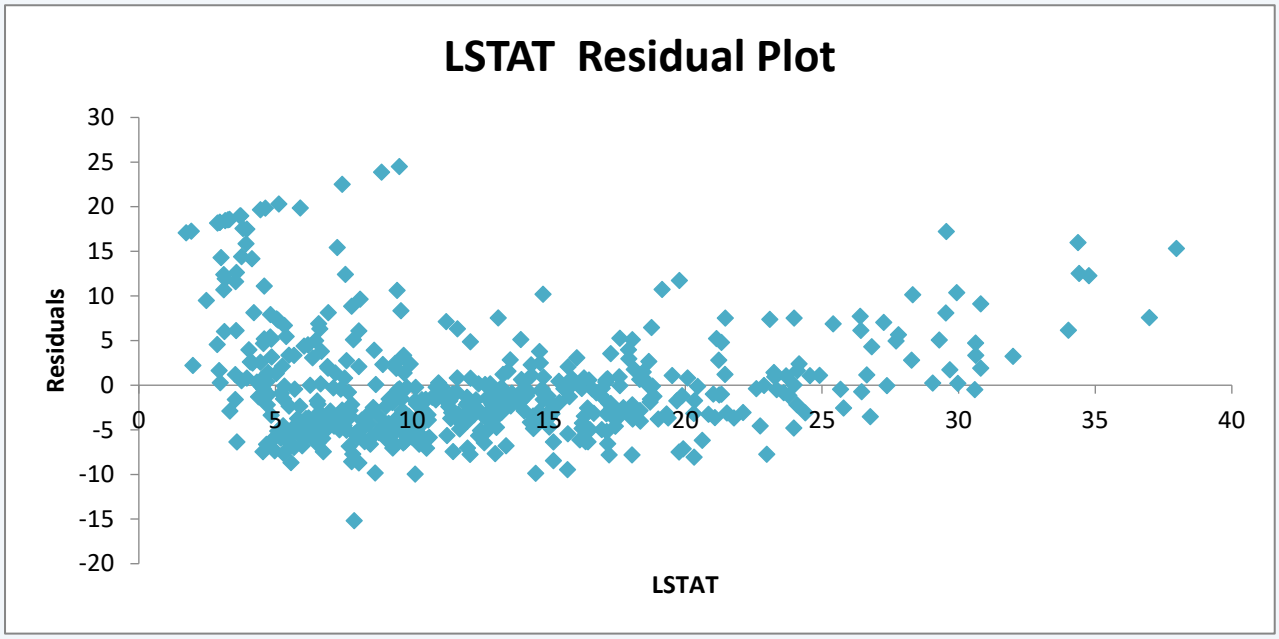
a) What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?

SUMMARY OUTPUT:

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

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	23243.914	23243.914	601.6178711	5.0811E-88
Residual	504	19472.38142	38.63567742		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	34.55384088	0.562627355	61.41514552	3.7431E-236	33.44845704	35.65922472	33.44845704	35.65922472
LSTAT	-0.950049354	0.038733416	-24.52789985	5.0811E-88	-1.0261482	-0.873950508	-1.0261482	-0.873950508



- The R Square value is 0.5441, indicating that approximately 54.41% of the total variance in the dependent variable (AVG_PRICE) is explained by the independent variable (LSTAT). This value represents the good fit of the regression model, and a higher R Square indicates a better fit.

- The coefficient value for the independent variable LSTAT is -0.9500. This means that for every one unit increase in percentage of lower-status population (LSTAT), the average house price (AVG_PRICE) to decrease by approximately $950.05. To better understand this, as the percentage of lower-status people in a neighbourhood increases, the average house price in that neighbourhood is expected to decrease this is negatively correlated.

- The intercept is a value which will be added to the product of the independent variables to get the predicted value of the dependent variable. In this case, the intercept is 34.5538, which means that if the percentage of lower-status population (LSTAT) is zero, the predicted average house price (AVG_PRICE) is 34,553.80 USD.

- This value is not very meaningful in reality, the percentage of lower-status population (LSTAT) cannot be zero. there are always going to be some people in a neighbourhood who are of lower status. So the intercept is not a very useful measure here.

- The residual plot analysis reveals that the regression model performs well and exhibits lower bias (i.e., more accurate predictions) for LSTAT values between 5 and 25. But, for LSTAT values below 5 and above 25, the model shows more error and bias so this model would not be a great fit for LSTAT values below 5 and above 25, indicating limitations in accurately estimating average house prices in those specific ranges.

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

- The LSTAT variable is significant for the analysis based on the model. In the Regression Summary output, the coefficient for the LSTAT variable is -0.9500, and the associated P-value is very close to zero (5.0811E-88) and it is less than or equal to α so it is Alternate Hypothesis. Based on the model we are rejecting the Null Hypothesis; from this we can conclude that the LSTAT variable is significant and has impact on predicting the average house prices.

6) Build a new Regression model including LSTAT and AVG_ROOM together as independent variables and AVG_PRICE as dependent variable. (6 marks)

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?

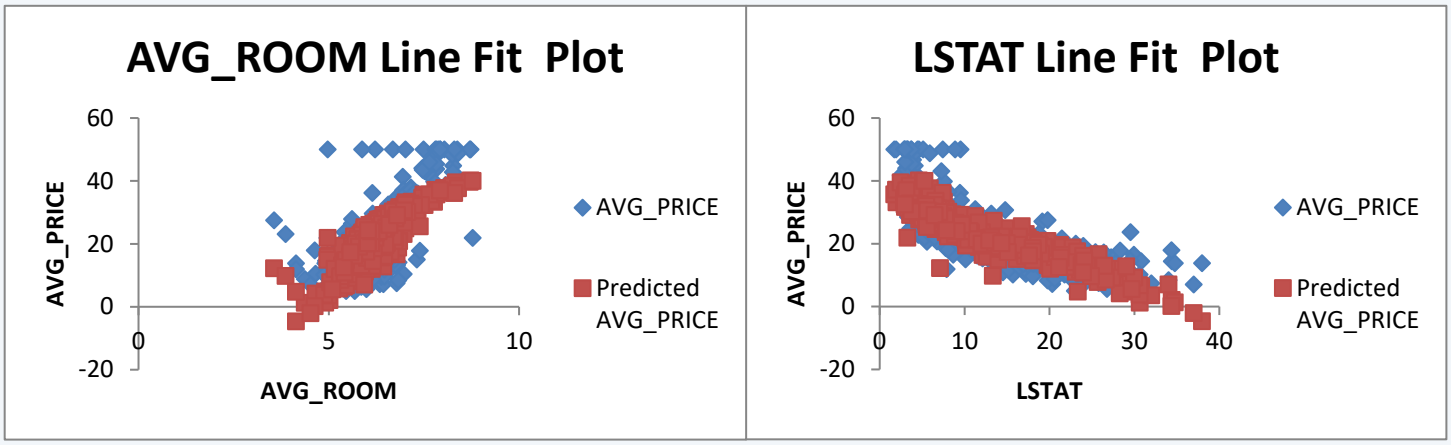
Here's the regression model including LSTAT and AVG_ROOM together as independent variables and AVG_PRICE as dependent variable.

SUMMARY OUTPUT:

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

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	27276.99	13638.49	444.3309	7.0085E-112
Residual	503	15439.31	30.69445		
Total	505	42716.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.358272812	3.17282778	-0.42809535	0.668764941	-7.591900282	4.875354658	-7.591900282	4.875354658
AVG_ROOM	5.094787984	0.4444655	11.4627299	3.47226E-27	4.221550436	5.968025533	4.221550436	5.968025533
LSTAT	-0.642358334	0.043731465	-14.6886992	6.66937E-41	-0.728277167	-0.556439501	-0.728277167	-0.556439501



- The model demonstrates a reasonably strong relationship, with a multiple R of 0.7991, indicating a good fit. The R-squared value is 0.6386, implying that approximately 63.86% of the variability in the average house price is explained by the model.
- The coefficients of the independent variables are statistically significant (p-values < 0.05), suggesting that both LSTAT and AVG_ROOM have a significant impact on predicting the average house price.
- AVG_ROOM (Average Number of Rooms): The coefficient for AVG_ROOM is positive (5.094787984). This indicates that there is a positive correlation between the average number of rooms in houses and the average house price.
- LSTAT (Percentage of Lower Status Population): The coefficient for LSTAT is negative (-0.642358334). This indicates that there is a negative correlation between the percentage of lower status population in the locality and the average house price.

Regression Equation $y=m_1*x_1+m_2*x_2+b$

b(Intercept) = -1.3582

AVG_ROOM Co-efficient (m_1) = 5.0947

LSTAT Coefficient (m_2) = -0.6423

$X_1=7$

$X_2=20$

Average Price = 5.0947*7+-0.6423*20-1.3582

Average Price = 21.458

Now, let's compare this predicted value to the company's quote of 30000 USD for this locality:

Company's quote = 30000 USD

Predicted AVG_PRICE = 21.458

The company is overcharging as they are quoting 30000 USD for the average price of the house, whereas the predicted value based on the regression model is only around 21.458 USD. The predicted value is significantly lower than the company's quote, indicating that the company's quote is much higher than what the regression model suggests.

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.

In question 5 (previous model):
Adjusted R Square: 0.543241826

In question 6 (current model):
Adjusted R Square: 0.637124475

- The adjusted R-square is a measure of how well the independent variables explain the variability in the dependent variable, taking into account the number of variables and the sample size. When we compare those two models, the performance of the model built in question 6 is better than the previous model built in question 5.

- The adjusted R-square of the model in question 6 (0.6371) is higher than that of the model in question 5 (0.5432). Higher adjusted R-square says that a larger proportion of total variance in the dependent variable (AVG_PRICE) is explained by both the variables AVG_ROOM and LSTAT in the current model compared to the previous model with only LSTAT in question 5.

- A model with more variables can generally be more robust and reliable. The inclusion of AVG_ROOM alongside LSTAT in Question 6 makes the model more robust, whereas the model of question 5 has only one variable LSTAT. Therefore, the model in question 6 is better at explaining the variation in average house prices and is considered to have a better fit to the data than the model in question 5.

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 with respect to AVG_PRICE. (8 marks)

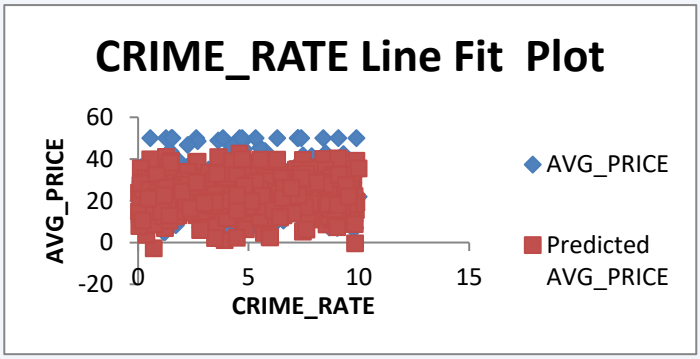
SUMMARY OUTPUT:

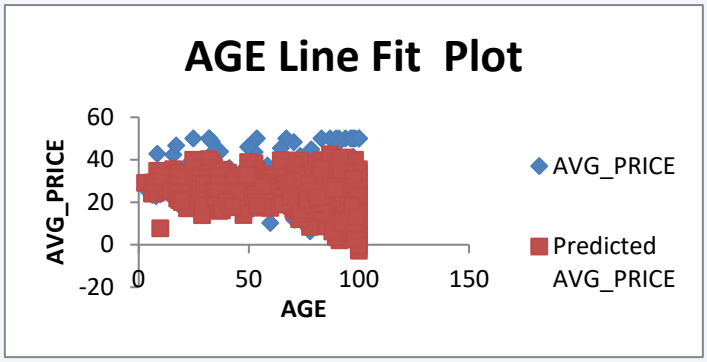
Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372
Adjusted R Square	0.688298647
Standard Error	5.1347635
Observations	506

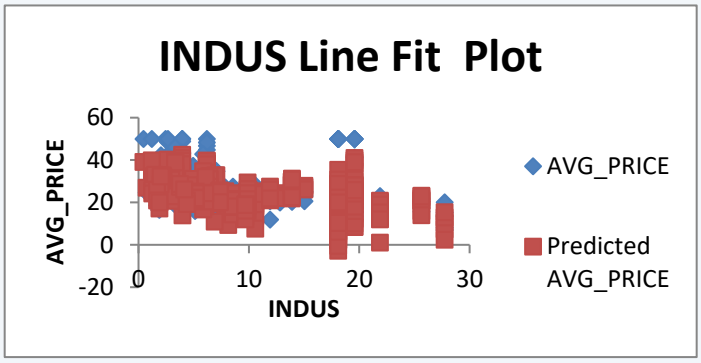
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	29638.8605	3293.21	124.9045049	1.9328E-121
Residual	496	13077.43492	26.3658		
Total	505	42716.29542			

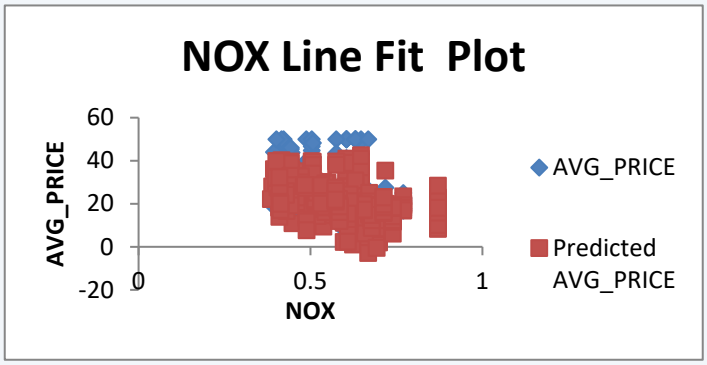
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.241315	4.817125596	6.07028	2.53978E-09	19.776828	38.70580267	19.77682784	38.7058027
CRIME_RATE	0.0487251	0.078418647	0.62135	0.534657201	-0.1053485	0.202798827	-0.10534854	0.20279883
AGE	0.0327707	0.013097814	2.502	0.012670437	0.0070367	0.058504728	0.00703665	0.05850473
INDUSTRY	0.1305514	0.063117334	2.06839	0.03912086	0.0065411	0.254561704	0.006541094	0.2545617
NOX	-10.32118	3.894036256	-2.6505	0.008293859	-17.972023	-2.670342809	-17.9720228	-2.67034281
DISTANCE	0.2610936	0.067947067	3.8426	0.000137546	0.127594	0.394593138	0.127594012	0.39459314
TAX	-0.014401	0.003905158	-3.6877	0.000251247	-0.0220739	-0.0067285	-0.02207388	-0.0067285
PTRATIO	-1.074305	0.133601722	-8.0411	6.58642E-15	-1.3368004	-0.811810259	-1.33680044	-0.81181026
AVG_ROOM	4.1254092	0.442758999	9.3175	3.89287E-19	3.2554947	4.995323561	3.255494742	4.99532356
LSTAT	-0.603487	0.053081161	-11.369	8.91071E-27	-0.7077782	-0.499194938	-0.70777824	-0.49919494

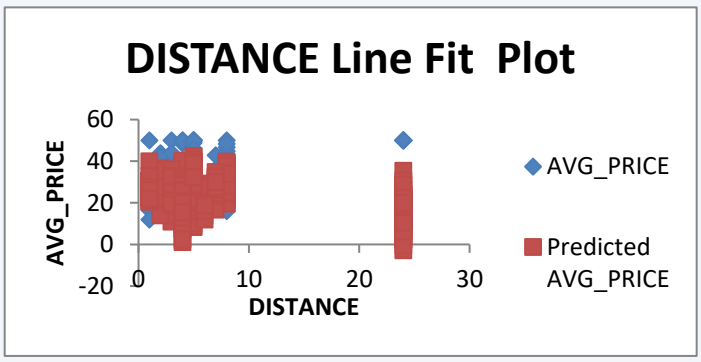
- The adjusted R-square of 0.6882 tells us that around 68.82% of the variation in house prices (AVG_PRICE) is shown by various factors considered in the model. This higher value suggests that the model is quite effective in making predictions for house prices. Nearly 69% of the fluctuations in house prices can be explained by the combination of independent variables used in the model, such as crime rate, age of houses, proximity to industries, air quality (NOX), distance to essential amenities, property tax, student-to-teacher ratio, average number of rooms, and the percentage of lower-status population. Having a strong adjusted R-square is good because it indicates that chosen independent variables are meaningful predictors of house prices.

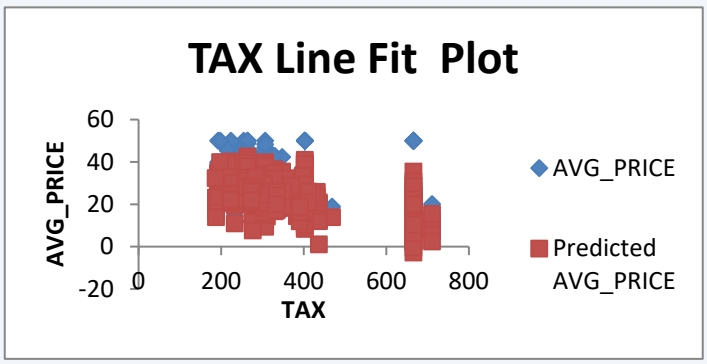


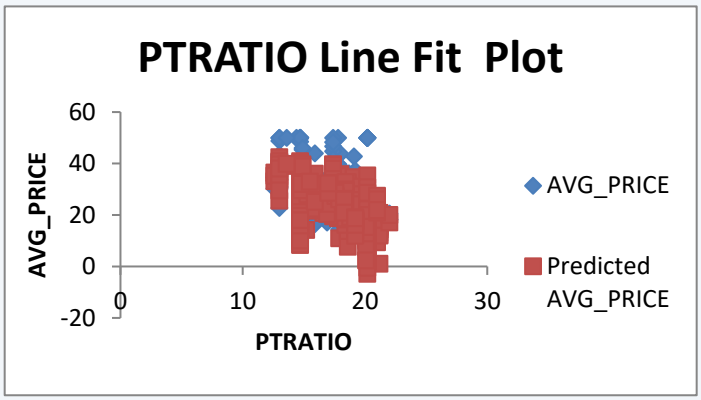


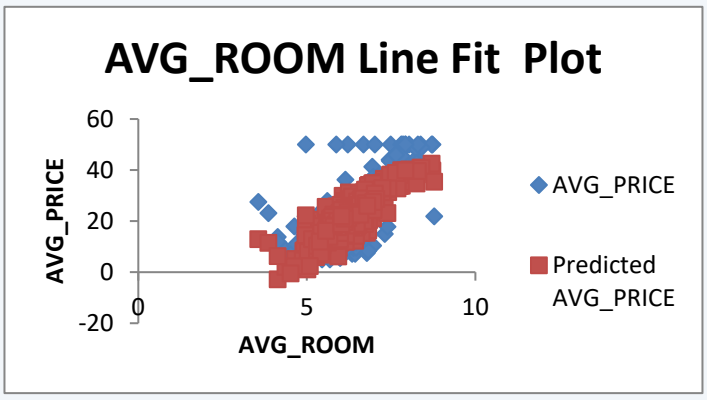


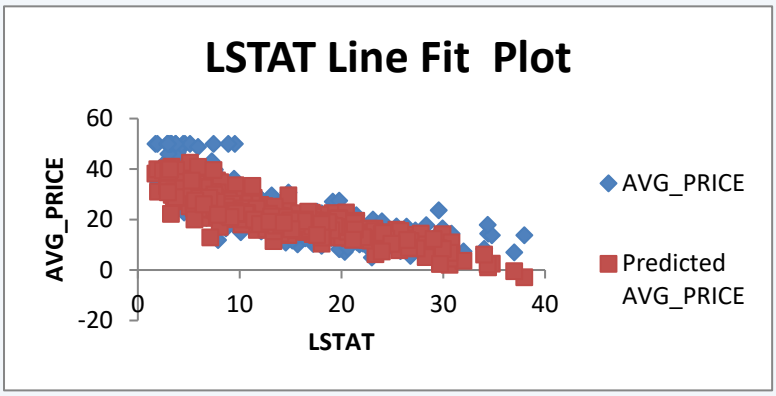












The coefficients represent the impact of each independent variable on the dependent variable (AVG_PRICE)

Crime Rate (CR): Slightly influences house prices (+0.0487).

Age of Homes (AGE): Older homes, increase house prices (+0.0328).

Industrial Proportion (INDUSTRY): More industry, increase in house prices (+0.1306).

Nitric Oxide Levels (NOX): High NOX decreases house prices (-10.3212).

Distance to Employment Centres (DISTANCE): Closer means higher house prices (+0.2611).

Property Tax Rate (TAX): Higher tax rate, lower house prices (-0.0144).

Pupil-Teacher Ratio (PTRATIO): Lower ratio increases house prices (-1.0743).

Average Number of Rooms (AVG_ROOM): More rooms, higher house prices (+4.1254).

% Lower Status (LSTAT): Increase in Lower status decreases house prices (-0.6035).

- The intercept of 29,241.30 USD is like a fantasy house price. It is the price of a house in a neighbourhood where the percentage of lower-status population is zero, the average number of rooms is zero, the distance to employment centres is zero, and other independent variables are zero. This is a theoretical construct that does can't be applied in the real world.
- In reality, there will always be some percentage of lower-status population in a neighbourhood, and the average number of rooms will never be zero. Therefore, the intercept of 29,241.30 USD is not a very useful measure.

The significance of each independent variable with respect to AVG_PRICE can be interpreted as follows:

	P-value
Intercept	2.53978E-09
CRIME_RATE	0.534657201
AGE	0.012670437
INDUSTRY	0.03912086
NOX	0.008293859
DISTANCE	0.000137546
TAX	0.000251247
PTRATIO	6.58642E-15
AVG_ROOM	3.89287E-19
LSTAT	8.91071E-27

AGE, INDUSTRY, NOX, DISTANCE, TAX, PTRATIO, AVG_ROOM, and LSTAT have extremely low p-values, all less than 0.01, indicating highly statistically significant relationships with AVG_PRICE. These variables are likely to have a significant impact on the average house price.

CRIME_RATE has a p-value of 0.5347, which is greater than 0.05, indicating that it is not statistically significant in explaining the variability in average house price.

Overall, the model shows that most of the independent variables are highly significant, except for CRIME_RATE. This means that the majority of the variables have a meaningful influence on the average house price, while crime rate does not have a impact in this particular 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: (8 marks)

a) Interpret the output of this model.

SUMMARY OUTPUT:

Regression Statistics	
Multiple R	0.832836
R Square	0.693615
Adjusted R Square	0.688684
Standard Error	5.131591
Observations	506

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regressior	8	29628.68	3703.585	140.643	1.911E-122
Residual	497	13087.61	26.33323		
Total	505	42716.3			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	29.4284735	4.804728624	6.124898	1.85E-09	19.98838959	38.868557	19.98838959	38.8685574
AGE	0.03293496	0.013087055	2.516606	0.012163	0.007222187	0.0586477	0.007222187	0.058647734
INDUS	0.13071001	0.063077823	2.072202	0.038762	0.006777942	0.2546421	0.006777942	0.254642071
NOX	-10.272705	3.890849222	-2.64022	0.008546	-17.9172457	-2.628164	-17.9172457	-2.628164466
DISTANCE	0.26150642	0.067901841	3.851242	0.000133	0.128096375	0.3949165	0.128096375	0.394916471
TAX	-0.0144523	0.003901877	-3.70395	0.000236	-0.02211855	-0.006786	-0.02211855	-0.006786137
PTRATIO	-1.0717025	0.133453529	-8.03053	7.08E-15	-1.33390511	-0.8095	-1.33390511	-0.809499836
AVG_ROOM	4.12546896	0.44248544	9.3234	3.69E-19	3.256096304	4.9948416	3.256096304	4.994841615
LSTAT	-0.6051593	0.0529801	-11.4224	5.42E-27	-0.70925186	-0.501067	-0.70925186	-0.501066704

- The regression model has an overall multiple R-squared value of 0.6936, indicating that approximately 69.36% of the variability in AVG_PRICE is explained by the combination of independent variables in the model. Suggesting that the model is a good fit for predicting average house prices.
- Overall, the regression output indicates that the model is statistically significant and can be used to predict the average house prices effectively. The significant independent variables (AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG_ROOM, and LSTAT) have meaningful impacts on the average house prices and should be considered when making predictions or decisions based on this model. The non-significant variable (CRIME_RATE) is not useful in predicting AVG_PRICE.
- The standard error is 5.1316. It measures the average deviation between the actual values of AVG_PRICE and the predicted values from the model. A lower standard error indicates that the model's predictions are closer to the actual data points.

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?

In question 7 (previous model):
Adjusted R Square: 0.6882

In question 8 (current model):
Adjusted R Square: 0.6886

- In the current question, the adjusted R-square value is 0.6886, and in the previous question, the adjusted R-square value is 0.6882. Both models have similar adjusted R-square values, but the current question's model (adjusted R-square: 0.6886) has a slightly higher value compared to the previous question's model (adjusted R-square: 0.6882).
- Since both models have a high adjusted R-square value, it indicates that they both explain a portion of the variability in the dependent variable (AVG_PRICE) using the selected independent variables. The current question's model performs slightly better in terms of the adjusted R-square value because this current model selects only the significant variables, suggesting that it might have a slightly better fit for the data compared to the model in the previous question.

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?

Here's the sorted values of the Coefficients in ascending order.

	Coefficients
NOX	-10.2727
PTRATIO	-1.0717
LSTAT	-0.60516
TAX	-0.01445
AGE	0.032935
INDUS	0.13071
DISTANCE	0.261506
AVG_ROOM	4.125469
Intercept	29.42847

- In the regression model, we have a coefficient for the variable NOX, which is approximately -10.27. This means that for every one-unit increase in the concentration of nitric oxides (NOX) in a locality, the average price of houses is estimated to decrease by approximately $10.27.

- If the value of NOX is higher in a locality in this town, it is associated with lower average house prices. This suggests that higher levels of nitric oxides in the air may have a negative impact on property values in area. Homebuyers might perceive higher NOX levels as an undesirable factor, leading to lower demand and lower average house prices in such locations.

d) Write the regression equation from this model.

$y=m_1x_1+m_2x_2+m_3x_3+m_4x_4+m_5x_5+m_6x_6+m_7x_7+m_8x_8+b$

y - Dependent Variable (Target Variable)
m - Slope
x - Independent Variable (Predictor Variable)
b - Intercept

CONCLUSION

- In conclusion, we conducted a comprehensive regression analysis to understand the relationship between various independent variables and the average house price (AVG_PRICE) in a particular town. We started by exploring the dataset, checking for missing values, and identifying potential outliers.

- Next, we performed a correlation analysis to identify any significant correlations between the independent variables and AVG_PRICE. Based on this analysis, we selected the most relevant independent variables, including AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG_ROOM, and LSTAT, to build our regression model.

- We built the regression model and evaluated its performance using various statistical metrics such as R-squared, adjusted R-squared, and significance tests. The model exhibited a reasonably high R-squared value, indicating that the selected independent variables collectively explain a substantial portion of the variability in AVG_PRICE.