TERRO'S REAL ESTATE AGENCY

BUSINESS REPORT

BIBEK KUMAR PATEL
Analyst

Index:

Introduction

In the dynamic and competitive real estate market, Terro's real-estate agency employs a systematic approach to estimate property pricing, leveraging a comprehensive set of features and factors. The company's dedicated auditor conducts a thorough analysis of geographic attributes such as pollution levels, crime rates, education facilities, and connectivity to derive accurate property valuations. This business report aims to unveil the most relevant features influencing house pricing, providing valuable insights for strategic decision-making in the real estate industry.

Data Dictionary:

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

Explanation of Data analysis:

1. Observations of descriptive statistics of each variable given in the table.

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
Mean	4.871976285	68.57490119	11.13677866	0.554695059	9.549407115	408.2371542	18.4555336	6.284634387	12.65306324	22.53280632
Standard Error	0.129860152	1.251369525	0.304979888	0.005151391	0.387084894	7.492388692	0.096243568	0.031235142	0.317458906	0.408861147
Median	4.82	77.5	9.69	0.538	5	330	19.05	6.2085	11.36	21.2
Mode	3.43	100	18.1	0.538	24	666	20.2	5.713	8.05	50
Standard Deviation	2.921131892	28.14886141	6.860352941	0.115877676	8.707259384	168.5371161	2.164945524	0.702617143	7.141061511	9.197104087
Sample Variance	8.533011532	792.3583985	47.06444247	0.013427636	75.81636598	28404.75949	4.686989121	0.49367085	50.99475951	84.58672359
Kurtosis	-1.189122464	-0.967715594	-1.233539601	-0.064667133	-0.867231994	-1.142407992	-0.285091383	1.891500366	0.493239517	1.495196944
Skewness	0.021728079	-0.59896264	0.295021568	0.729307923	1.004814648	0.669955942	-0.802324927	0.403612133	0.906460094	1.108098408
Range	9.95	97.1	27.28	0.486	23	524	9.4	5.219	36.24	45
Minimum	0.04	2.9	0.46	0.385	1	187	12.6	3.561	1.73	5
Maximum	9.99	100	27.74	0.871	24	711	22	8.78	37.97	50
Sum	2465.22	34698.9	5635.21	280.6757	4832	206568	9338.5	3180.025	6402.45	11401.6
Count	506	506	506	506	506	506	506	506	506	506

i. CRIME_RATE:

CRIME_RATE	
Mean	4.871976285
Standard Error	0.129860152
Median	4.82
Mode	3.43
Standard Deviation	2.921131892
Sample Variance	8.533011532
Kurtosis	-1.189122464
Skewness	0.021728079
Range	9.95
Minimum	0.04
Maximum	9.99
Sum	2465.22
Count	506

- In this database count of 506 the avg crime rate of Boston city is 4.87 per 1000 people in a town.
- In this dataset 50% observed crime rate are below 4.82, while remaining 50% are above this threshold.
- The std deviation of this dataset is 2.921 which measures the variation in dataset.
- Kurtosis which is -1.18 refers to slightly platykurtic means crime may be in some places of Boston.
- Skewness is 0.02 where data distribution is trailing off to right means crime rate is nearby 0 i.e. crime rate is equally distributed in Boston city.
- Maximum crime rate in city is 9.90.

ii. AGE:

AGE	
Mean	68.57490119
Standard Error	1.251369525
Median	77.5
Mode	100
Standard Deviation	28.14886141
Sample Variance	792.3583985
Kurtosis	-0.96771559
Skewness	-0.59896264
Range	97.1
Minimum	2.9
Maximum	100
Sum	34698.9
Count	506

- The avg of house built before 1940 is 68.57% and having std deviation of 28.14 which measures the variance of the data.
- In Boston city 50% of house having below 77.5% and above.
- And kurtosis is negative her (-0.96) which refers to having flat distribution around the Boston city. And skewness is slightly trialed left (-0.59) which is lesser than the median which refers that having more old houses built prior 1940.
- The min age of house is 2.9 and max is of 100 having huge difference of range.

iii. INDUS:

INDUS	
Mean	11.13677866
Standard Error	0.304979888
Median	9.69
Mode	18.1
Standard Deviation	6.860352941
Sample Variance	47.06444247
Kurtosis	-1.233539601
Skewness	0.295021568
Range	27.28
Minimum	0.46
Maximum	27.74
Sum	5635.21
Count	506

- The average non retail business acres per town is 11.13%
- In this 50% of dataset below 9.69% and above non retail business acres.
- The negative kurtosis (-1.23) suggests a flatter distribution and positive skewness (0.29) indicates a rightward tail which means most of the non-retail lands are near the Boston city and other are outliers.
- The range from 0.46 to 27.74 suggests few instances of lower industry levels and more instances of higher ones in each locality.

iv. NOX:

NOX	
Mean	0.554695059
Standard Error	0.005151391
Median	0.538
Mode	0.538
Standard Deviation	0.115877676
Sample Variance	0.013427636
Kurtosis	-0.06466713
Skewness	0.729307923
Range	0.486
Minimum	0.385
Maximum	0.871
Sum	280.6757
Count	506

- The average nitrogen oxide concentration is about 0.554 (parts per 10 million particles), and it can vary around this value by approximately 0.116.
- Most concentrations are close to 0.538, but some are higher, making the distribution slightly spread out to the right.
- The range of concentrations goes from 0.385 to 0.871, showing the extent of variation in the dataset.

v. DISTANCE:

DISTANCE	
Mean	9.549407115
Standard Error	0.387084894
Median	5
Mode	24
Standard Deviation	8.707259384
Sample Variance	75.81636598
Kurtosis	-0.86723199
Skewness	1.004814648
Range	23
Minimum	1
Maximum	24
Sum	4832
Count	506

- The average distance from highway is 9.54 and std deviation is 8.70 which calculates the variance of dataset.
- In this dataset 50% of house are below 5 miles away from the highway and above.
- Here kurtosis is in negative (-0.86) which is having flat distribution and skewness is 1.0 which is positive in nature trailing off to right which explains that there are some localities more far than median of the data.
- Some houses in Boston city having minimum 1 miles away from the highway & maximum is 24 miles.

vi. TAX:

TAX	
Mean	408.2371542
Standard Error	7.492388692
Median	330
Mode	666
Standard Deviation	168.5371161
Sample Variance	28404.75949
Kurtosis	-1.142407992
Skewness	0.669955942
Range	524
Minimum	187
Maximum	711
Sum	206568
Count	506

- In this dataset average full-tax rate per $10000, in Boston city is 408.23.
- Half of the houses in dataset has full value property-tax rate below 330 per 10000$ and half of the houses has tax rate above 330
- Tax rate differs in different locality by 168.5.
- Here kurtosis is in negative (-1.14) and skewness is in positive which means there will be higher and lower tax localities in Boston city.
- The maximum tax rate in Boston locality is 711 & minimum is around 187

vii. PTRATO:

PTRATIO	
Mean	18.4555336
Standard Error	0.096243568
Median	19.05
Mode	20.2
Standard Deviation	2.164945524
Sample Variance	4.686989121
Kurtosis	-0.28509138
Skewness	-0.80232493
Range	9.4
Minimum	12.6
Maximum	22
Sum	9338.5
Count	506

- In this dataset the average PTRATIO is 18.455 with the std deviation of 2.16 that measure the variation in dataset.
- Half of the dataset has student-teacher ratio below 19.05 and remaining half of dataset has student-teacher-ratio above 19.05.
- Here kurtosis and skewness are both in negative which refers that it has flat distribution and trailed off to the left indicates that having less Pupil teacher ratio in the Boston City.
- A town in Boston has minimum student-teacher ratio of 12.6 (i.e. for 1 teacher 12 students are there). And maximum student-teacher ratio is 22(i.e. for 1 teacher 22 students are there).

viii. AVG ROOM:

AVG ROOM	
Mean	6.284634387
Standard Error	0.031235142
Median	6.2085
Mode	5.713
Standard Deviation	0.702617143
Sample Variance	0.49367085
Kurtosis	1.891500366
Skewness	0.403612133
Range	5.219
Minimum	3.561
Maximum	8.78
Sum	3180.025
Count	506

- There is average 6.28 rooms are there.
- The std deviation is 0.70 which measures the variation of dataset.
- In this dataset kurtosis is in positive (1.89) which refers that there is a strong relation around the median data and skewness is slightly positive where Rooms are more than average in Boston city.
- We have min as 3.55 which is found less in Boston city according to skewness and kurtosis and we have max of 8.78 per capita of average room.

ix. LSTAT:

| LSTAT	
Mean	12.65306324
Standard Error	0.317458906
Median	11.36
Mode	8.05
Standard Deviation	7.141061511
Sample Variance	50.99475951
Kurtosis	0.493239517
Skewness	0.906460094
Range	36.24
Minimum	1.73
Maximum	37.97
Sum	6402.45
Count	506

- The average of the % of lower status population in Boston city is 12.65 where std deviation is 7.141 which is the variance of the dataset.
- Half of the data set has lower status population below 11.36% per town and remaining half dataset has lower status population above 11.36% per town.
- Kurtosis (0.49) & skewness (0.90) are in positive nature which refers it having slight peak than normal distribution and trailed to the right defining that Boston city is having less %of lower status of the population.

x. AVG_PRICE:

AVG_PRICE	
Mean	22.53280632
Standard Error	0.408861147
Median	21.2
Mode	50
Standard Deviation	9.197104087
Sample Variance	84.58672359
Kurtosis	1.495196944
Skewness	1.108098408
Range	45
Minimum	5
Maximum	50
Sum	11401.6
Count	506

- The average of this dataset is 22530$ and std deviation is of 21.2 which measures the variation of dataset.
- 50% of houses in the dataset has average price below 21200$ per town and remaining half of the houses has average price above 21200$ per town.
- Here kurtosis is more than normal distribution (1.49) which is a concentrated distribution with positive skewness (1.10) trailing right refers that the price is more than the median in the locality of Boston city.
- We have min of 5 and max of 50 according to the data distribution.

2. Histogram inference of AVG_PRICE variable.
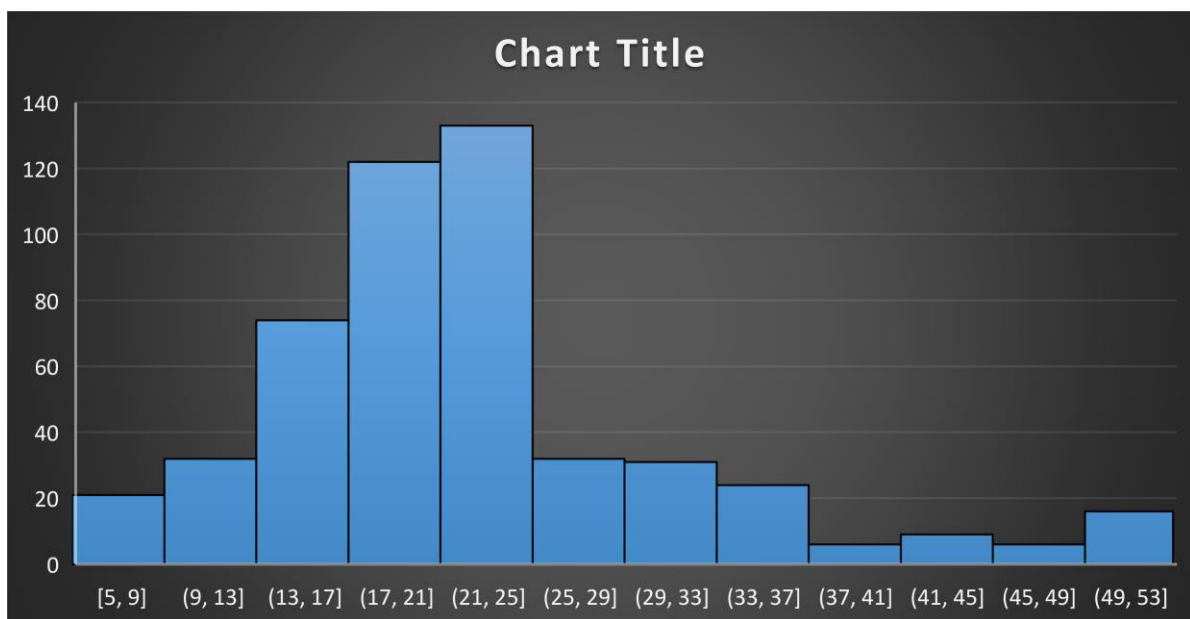
Inferences:

- 33% of houses (167) has average price between 20000 to 25000$ and around 23% of houses
- (118) average price is between 15000 to 20000$.
- Only 5% of houses (24 houses) has avg_price between 5000$ to 10000$.
- Majority of houses (56%) has average price between 15000$ to 25000$.
- 24% of houses has average price more than 25000$.
- Here kurtosis is more than normal distribution which is a concentrated distribution with positive skewness trailing right refers that the price is more than the median in the locality of Boston city.

3. Observation on the computed 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

- In this covariance we observe a strong positive covariance of 2397.94 between Tax Rate and Age suggesting that when one variable increases the other variable tends to increase simultaneously.
- And Average Price and TAX has strong negative covariance (-724.82042).
- AVG_PRICE column is negatively related to almost all columns except 3 columns
- AVG_ROOM column is negatively related to almost all columns except 2 columns.
- PTRATIO, NOX and AGE columns are positively related with all the columns.
- AVG_PRICE of house increases with increase in crime rate.
- AVG_PRICE of house increases with increase in AVG_ROOM.
- With increase in NOX concentration the AVG_PRICE of house decreases.

4. Correlation Matrix Analysis:

- Top 3 positively correlated pairs.
- 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

a) Top 3 positively correlated pairs are TAX & Distance (0.91), NOX & INDUS (0.76), NOX & Age (0.73).

b) Top 3 negatively correlated pairs are AVG Price & LSTAT (-0.737), LSTAT & AVG_ROOM (-0.61), AVG_PRICE & PTRATIO (-0.50).

5. Regression Model Analysis:
   - Inference on Variance Explained, Coefficient Value, Intercept, and Residual Plot.
   - Significance of LSTAT Variable in the Analysis.

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

	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

LSTAT Residual Plot
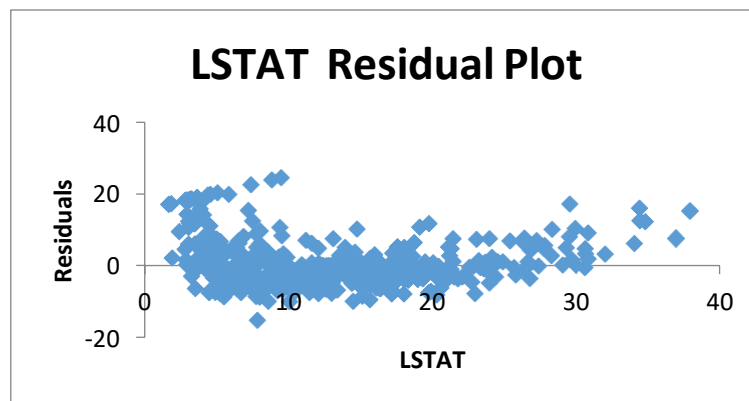
- In this regression model 54.41% of the independent variable explains variance in dependent variable (average price).
- The intercept is the starting point of prediction when all the variables are zero the estimated value of dependent variable is 34.55.
- And P-value of LSTAT is below 0.005, so we are rejecting the null hypothesis and accepting the alternative hypothesis.

6. Regression Model with LSTAT and AVG_ROOM:
   - Regression Equation and Price Comparison for a House with 7 Rooms and LSTAT of 20.
   - Performance Comparison with Previous Model Based on Adjusted R-square.

	Regression Statistics	
Multiple R		0.799100498
R Square		0.638561606
Adjusted R Square		0.637124475
Standard Error		5.540257367
Observations		506

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1.358272812	3.17282778	-0.428095348	0.668764941	-7.591900282	4.875354658	-7.591900282	4.875354658
AVG_ROOM	5.094787984	0.4444655	11.46272991	3.47226E-27	4.221550436	5.968025533	4.221550436	5.968025533
LSTAT	-0.642358334	0.043731465	-14.68869925	6.66937E-41	-0.728277167	-0.556439501	-0.728277167	-0.556439501

a) Regression equation= Intercept+ AVG_ROOM *(x1) + LSTAT *(x2)

= (-1.3582) +5.0947(7) +(-0.64235) (20)

value of average price = 21.45807639

Here the value of AVG_PRICE is $21458.07 and with the comparison of company quoting a value 30000 USD for this locality they are overcharging.

b) Yes, the performance of the present model is better than that of the previous model. We can observe that the adjusted R-squared of the previous model is 0.543, while the current adjusted R-square is 0.637 This indicates that a higher proportion of the variance in the dependent variable is explained by the independent variable in the current mode.

7. Regression Model with All Variables:
   ▪ Interpretation of Output in Terms of Adjusted R-square, Coefficient, and Intercept.
   ▪ Significance Explanation for Each Independent Variable with Respect to AVG_PRICE.

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

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	29.24131526	4.817125596	6.070282926	2.53978E-09	19.77682784	38.70580267	19.77682784	38.70580267
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.534657201	-0.105348544	0.202798827	-0.105348544	0.202798827
AGE	0.032770689	0.013097814	2.501996817	0.012670437	0.00703665	0.058504728	0.00703665	0.058504728
INDUS	0.130551399	0.063117334	2.068392165	0.03912086	0.006541094	0.254561704	0.006541094	0.254561704
NOX	-10.3211828	3.894036256	-2.650510195	0.008293859	-17.97202279	-2.670342809	-17.97202279	-2.670342809
DISTANCE	0.261093575	0.067947067	3.842602576	0.000137546	0.127594012	0.394593138	0.127594012	0.394593138
TAX	-0.01440119	0.003905158	-3.687736063	0.000251247	-0.022073881	-0.0067285	-0.022073881	-0.0067285
PTRATIO	-1.074305348	0.133601722	-8.041104061	6.58642E-15	-1.336800438	-0.811810259	-1.336800438	-0.811810259
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.89287E-19	3.255494742	4.995323561	3.255494742	4.995323561
LSTAT	-0.603486589	0.053081161	-11.36912937	8.91071E-27	-0.70777824	-0.499194938	-0.70777824	-0.499194938

Explanation:

In this regression model the adjusted R-square value of 0.6882 indicated that 68.82% of the variance in the dependent variable (AVG_PRICE) is explained by the independent variables.

There is positive coefficient for an independent variable which suggests that as that variable increases, the AVG PRICE is expected to increase. And there are some negative coefficients which suggest that as the variable increases, the AVG_PRICE is expected to decrease here.

In this dataset the intercept value is 29.24 which represents the estimated average value of the dependent variable when all independent variables are zero.

Significance of each independent variable with respect to AVG_PRICE are:

Independent variables such as crime rate, age, INDUS, distance, and AVG room have positive coefficients, indicating a direct positive influence on AVG price. This means that as these independent variables increase, the AVG price is also expected to increase.

An increase in NOX, TAX, PTRATIO, and LSTAT is associated with a decrease in AVG price.

8. Regression Model with Significant Variables:
   - Interpretation of Output.
   - Comparison of Adjusted R-square with Previous Model.
   - Coefficient Sorting and Impact of NOX on Average Price.
   - Regression Equation for the Model.

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	4.804728624	6.124898157	1.84597E-09	19.98838959	38.8685574	19.98838959	38.8685574
AGE	0.03293496	0.013087055	2.516605952	0.012162875	0.007222187	0.058647734	0.007222187	0.058647734
INDUS	0.130710007	0.063077823	2.072202264	0.038761669	0.006777942	0.254642071	0.006777942	0.254642071
NOX	-10.27270508	3.890849222	-2.640221837	0.008545718	-17.9172457	-2.628164466	-17.9172457	-2.628164466
DISTANCE	0.261506423	0.067901841	3.851242024	0.000132887	0.128096375	0.394916471	0.128096375	0.394916471
TAX	-0.014452345	0.003901877	-3.703946406	0.000236072	-0.022118553	-0.006786137	-0.022118553	-0.006786137
PTRATIO	-1.071702473	0.133453529	-8.030529271	7.08251E-15	-1.333905109	-0.809499836	-1.333905109	-0.809499836
AVG_ROOM	4.125468959	0.44248544	9.323400461	3.68969E-19	3.256096304	4.994841615	3.256096304	4.994841615
LSTAT	-0.605159282	0.0529801	-11.42238841	5.41844E-27	-0.70925186	-0.501066704	-0.70925186	-0.501066704

a) Interpretation of the model:

In this statistics Multiple R is 0.83 which indicates a strong positive correlation between the dependent and independent variables. And R-square is 0.69 which explains 69% of dependent variable by the independent variable.

b) Comparison of adjusted R-square value:

The adjusted R-square value in this model is 0.6887, whereas the adjusted R-square in the previous model was 0.6882. The difference is very small, and both models perform quite similarly in explaining the variance in AVG_PRICE.

c) Sorting the values of the coefficient values in ascending order:

NOX (-10.27): Negative coefficient, a higher NOX value is associated with a decrease in AVG_PRICE.

TAX (-0.0145): Negative coefficient, a higher TAX value is associated with a decrease in AVG_PRICE.

PTRATIO (-1.072): Negative coefficient, a higher PTRATIO is associated with a decrease in AVG_PRICE.

LSTAT (-0.605): Negative coefficient, a higher LSTAT is associated with a decrease in AVG_PRICE.

AGE (0.0329): Positive coefficient, a higher AGE is associated with an increase in AVG_PRICE.

INDUS (0.131): Positive coefficient, a higher INDUS is associated with an increase in AVG_PRICE.

DISTANCE (0.2615): Positive coefficient, a higher DISTANCE is associated with an increase in AVG_PRICE.

AVG_ROOM (4.125): Positive coefficient, a higher AVG_ROOM is associated with an increase in AVG_PRICE.

d) Regression Equation:

$AVG\_PRICE = 29.43 + 0.0329 \times AGE + 0.131 \times INDUS - 10.27 \times NOX + 0.2615 \times DISTANCE - 0.0145 \times TAX - 1.072 \times PTRATIO + 4.125 \times AVG\_ROOM - 0.605 \times LSTAT$.

Conclusion:

In conclusion, the analysis conducted provides valuable insights into the factors influencing house pricing in the specified locality. Through regression models, we identified key variables and their significance, aiding Terro's real estate in making more informed decisions for property valuation. This data-driven approach enhances the agency's competitiveness and strategic positioning in the real estate mark.