Terro’s real estate agency

NAME - PARTH MORI

DATE – 5/11/23

Contents

Problem Statement (Situation): 4

Data Dictionary: 4

1) Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation. 4

2) Plot a histogram of the AVG_PRICE variable. What do you infer? 6

3) Compute the covariance matrix. Share your observations. 7

4) Create a correlation matrix of all the variables (Use Data analysis tool pack). (5 marks) 8

a) Which are the top 3 positively correlated pairs and 8

b) Which are the top 3 negatively correlated pairs 8

5) Build an initial regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot. 9

a) What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot? 10

b) LSTAT variable significant for the analysis based on your model? 10

6) Build a new Regression model including LSTAT and AVG_ROOM together as independent variables and AVG_PRICE as dependent variable. 11

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 te company Overcharging/ Undercharging? 11

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. 12

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. 13

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: 14

a) Interpret the output of this model. 14

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? 15

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? 15

d) Write the regression equation from this model. 16

List of Tables

Table 1.Descriptive states....4
Table 2.coverience....7
Table 3.correlation....8
Table 4.AVG_PRICE VS LSTAT....9
Table 5. LSTAT and AVG_ROOM vs AVG_PRICE....11
Table 6. AVG_PRICE vs another variable....13
Table 7. coefficients and p value....14
Table 8.coefficient....15

List of Figure

Figure 1.Histogram....6
Figure 2. Residual plot LSTAT....9

Problem Statement (Situation):

"Finding out the most relevant features for pricing of a house" 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.

The agency has provided a dataset of 506 houses in Boston. Following are the details of the dataset:

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

Objective :

To analyse the magnitude of each variable to which it can affect the price of a house in a particular locality.

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

DESCRIPTIVE STATISTICS

STATISTICS	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
Mean	4.87	68.57	11.14	0.55	9.55	408.24	18.46	6.28	12.65	22.53
Standard Error	0.13	1.25	0.30	0.01	0.39	7.49	0.10	0.03	0.32	0.41
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 Deviatior	2.92	28.15	6.86	0.12	8.71	168.54	2.16	0.70	7.14	9.20
Sample Variance	8.53	792.36	47.06	0.01	75.82	28404.76	4.69	0.49	50.99	84.59
Kurtosis	-1.19	-0.97	-1.23	-0.06	-0.87	-1.14	-0.29	1.89	0.49	1.50
Skewness	0.02	-0.60	0.30	0.73	1.00	0.67	-0.80	0.40	0.91	1.11
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

Table 1.Descriptive states

By using data analyst tool pack we will be create a summery in statistics, we will be able to create a descriptive statistics for each variable, which tell us mean, standard error, median, mode, standard deviation, sample variance, kurtosis, skewness, range, minimum, maximum, sum, count .

CRIME_RATE:

1- Mean of CRIME_RATE is approximately 4.87 and SD is 2.92 that means spread of the values are around the mean.

2 - Average CRIME_RATE in the town is 4.82

3 - Distribution of the CRIME_RATE right skewed

4 - Kurtosis is negative (-1.19) that indicates distribution is flatter

AGE:

1- Average AGE of the house is 68.57
2- The maximum AGE of the houses is 100 and mode is also 100 which says that most of the houses has age of 100.
3- AGE of the skewed is negative (left tail) that indicates the most of the properties are old
4- Kurtosis is negative (-0.97) that indicates distribution is flatter
5- Most of the house age is around (86,99) years

INDUS:

1- Mean of the Indus is 11.14
2- Distribution of the Indus is right skewed(positive)
3- Kurtosis is negative (-1.23) that indicates the peak is flatter compared normal distribution

NOX:

1- Mean of the NOX is 0.55
2- Skewed of the NOX is 0.73 that indicates distributions is slightly towards right tail
3- Kurtosis is -0.06 it is close to Zero that indicates the distribution is relatively normal

DISTANCE:

1- Mean of the distance is 9.55
2- Mode of the distance and maximum distance are 24 that means most of the houses are away from the highway
3- Skewness of the distance is 1 that indicates distribution is right tail
4- Kurtosis is negative (-0.87) that means distribution is flat

TAX:

1- Average tax paid is 408.24
2- Bar graph indicates the most of the houses paid tax between (261,335)
3- Skewed of the tax is 0.67 that means positive skew distribution towards the right side
4- kurtosis of the tax is (-1.14) that means distribution is relatively flatter

PTRATIO:

1- Mean of the pupil teacher ratio is 18.46

2- mode of the pupil teacher is 20.20 and maximum pupil teacher is 22 that difference is small it's indicated the more school with higher pupil teacher
3- skewed of the pupil teacher is -0.80 that means negative so distribution is slightly left side
4- kurtosis is negative (-0.29) means distribution is relatively flatter

AVG_ROOM:

1- Average room of the houses is around 6
2- Kurtosis of the AVG_ROOM is 1.89 indicates the data is sharper peak compared to normal distribution
3- Given graph more houses are higher number of rooms
4- distribution of the AVG_ROOM is positive and we can see an outlier in the data

LSTAT:

1- The mean of LSTAT is 12.65
2- Distribution of the LSTAT is positive and we can see an outlier in the data
3- positive skewed indicates lower population in general
4- kurtosis is 0.49 indicate the distribution has relatively moderate peak

AVG_PRICE:

1- Average price of the houses is 22.53
2- Mode of the AVG_PRICE and max of the AVG_PRICE is 50 that means most of the house's price is 50
3- positive skewed indicates the more houses is lower price
4- There are some outliers in the data indicates the few houses with higher price

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

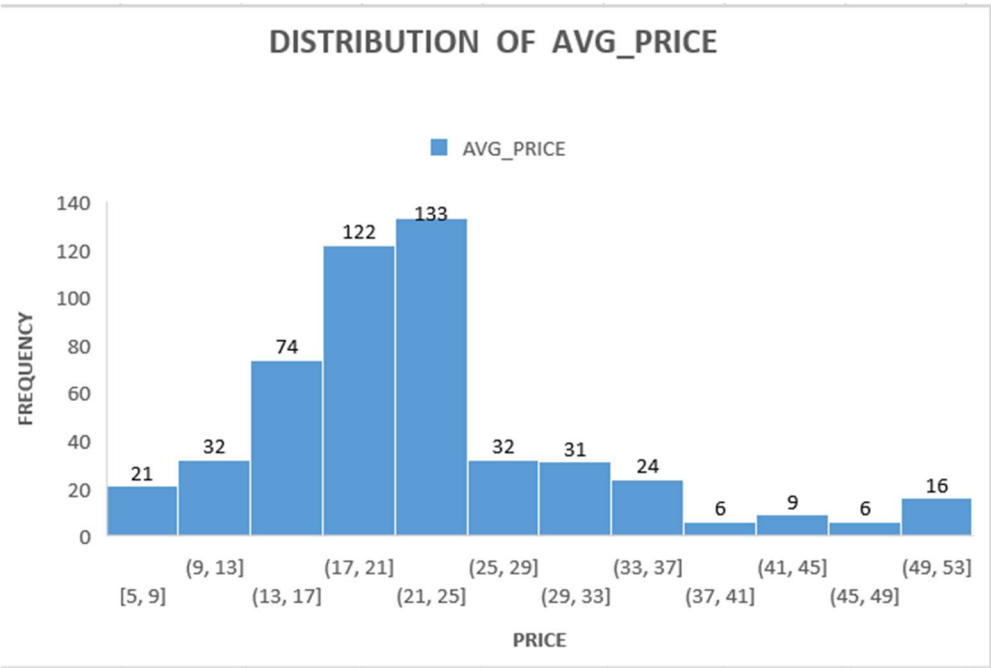

Figure 16.Histogram

From the above histogram we were able to observe that

1- Most of the house's price are range between ($21000, $25000)
2- we can see here some of the houses are with higher price followed by ($49000, $53000)
3- There are a greater number of houses are lower price and a smaller number of houses are higher price
4- Data is positive skewed (right tail)

3) Compute the covariance matrix. Share your observations.

covariance matrix is helps to understand whether two variables are directly proportional (positive covariance) or inversely proportional (negative covariance). positive value indicate directly proportional, and negative value indicates inversely proportional.

By observing the covariance matrix is

1- we can see tax is high covariance values with each other feature that means tax is a very good variability with other features
2- As per above data (age, tax), (Indus, tax), (distance, tax) have more covariance that direct relationship to each other one is increase other is also increase

CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
8.516									
0.563	790.792								
-0.110	124.268	46.971							
0.001	2.381	0.606	0.013						
-0.230	111.550	35.480	0.616	75.667					
-8.229	2397.942	831.713	13.021	1333.117	28348.624				
0.068	15.905	5.681	0.047	8.743	167.821	4.678			
0.056	-4.743	-1.884	-0.025	-1.281	-34.515	-0.540	0.493		
-0.883	120.838	29.522	0.488	30.325	653.421	5.771	-3.074	50.894	
1.162	-97.396	-30.461	-0.455	-30.501	-724.820	-10.091	4.485	-48.352	84.420

Table 2.coverience

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 and

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

Table 3.correlation

Top 3 positively correlated pairs

1- Distance – Tax	0.910228189
2- NOX – Indus	0.763651447
3- NOX – Age	0.731470104

Top 3 negatively correlated pairs

1.Avg_Price – LSTAT	- 0.737662726
2.LSTAT – AVG_ROOM	-0.613808272
3.Avg_Price – PTRATIO	-0.507786686

5) Build an initial regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as Independent Variable. Generate 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
RESIDUAL OUTPUT								
Observation	*Predicted AVG_PRICE*	*Residuals*						
1	29.8225951	-5.822595098						
2	25.87038979	-4.270389786						
3	30.72514198	3.974858016						
4	31.76069578	1.639304221						
5	29.49007782	6.709922176						
6	29.60408375	-0.904083746						
7	22.74472741	0.155272588						
8	16.36039575	10.73960425						

Table 4.AVG_PRICE VS LSTAT

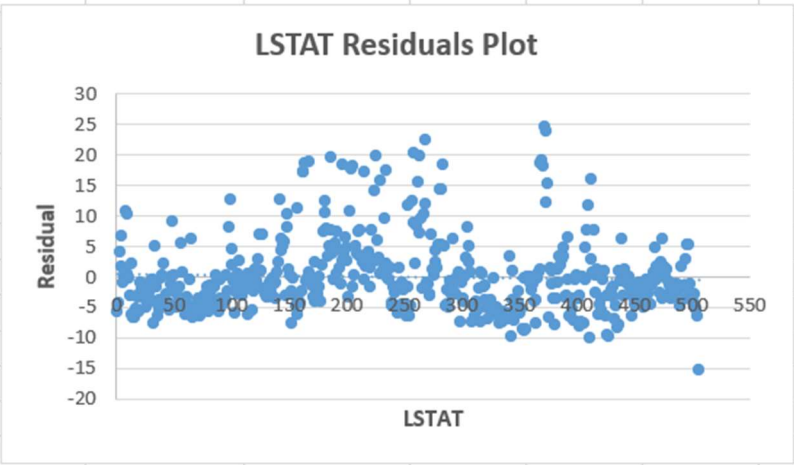

Figure 2. Residual plot LSTAT

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

Variance:

1- From this model 54% of the variation in the average price is explained by the LSTAT

Coefficient:

1- The coefficient of LSTAT for the model is -0.950049354.
2- LSTAT increase by 0.9 times then average price of house decreases 0.9 times

Intercept:

1- Intercept of LSTAT for the model is 34.55384088.

residual plot:

1- most of the plots are on upper side of the x axis

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

P value > Alpha = (is significant)

P value < Alpha = (not a significant)

P value is 0.05

p-value (5.08E-88) of this model is less than 0.05.

we can say that LSTAT is a significant variable according to this model.

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

SUMMARY OUTPUT

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

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	27276.98621	13638.49311	444.3308922	7.0085E-112
Residual	503	15439.3092	30.69445169		
Total	505	42716.29542			

	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

RESIDUAL OUTPUT

Observation	*Predicted AVG_PRICE*	*Residuals*
1	28.94101368	-4.941013681
2	25.48420566	-3.884205661
3	32.65907477	2.040925231
4	32.40652	0.99348

Table 5. LSTAT and AVG_ROOM vs AVG_PRICE

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?

Regression equation:

$$Y = MX + C$$

Where,

Y - dependent variable

X - independent variable

M - slope (coefficient of X)

C - constant (intercept)

here we can use multi linear regression as per summery output our equation is:

Y = M0X0 + M1X1 + C

Y = AVG_PRICE

X0 = AVG_ROOM M0 = 5.09

X1 = LSTAT M1 = -0.642

Above given Que-a X0 =7 and X1 = 20

So, put value in above equation

Y = (5.09) X0 + (-0.642) X1 + (-1.358)

Y =21.44

Multiply by 1000

So, the price for the new house is $21440

$21440 is lesser then the $30000 we can say that company is Overcharging

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.

Comparing the R-square

R- square of previous model(Q-5) = 0.54

R- square of this model = 0.63

We can say that this model is better than the previous model

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.

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	3293.206722	124.9045049	1.9328E-121
Residual	496	13077.43492	26.3657962		
Total	505	42716.29542			

	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

Table 6. AVG_PRICE vs another variable

Adjusted R-square value is 0.668

Adjusted R-square closer to zero indicates that model data is fit

P value > Alpha = (is significant)

P value < Alpha = (not a significant)

Level of significance can be denoted by alpha

Alpha = 1- confidence level

= 1-0.95

= 0.05

P value is greater than alpha is not significant less then alpha is significant

From the above model we can say that crime rate is not a significant variable for average price of a house as p-value is greater than 0.5.

NOX, TAX, PTRATIO and LSTAT have negative coefficients which says that increase in these features will result decrease in price of the house and vice-versa

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.

	Coefficients	*P-value*
Intercept	29.24131526	2.53978E-09
CRIME_RATE	0.048725141	0.534657201
AGE	0.032770689	0.012670437
INDUS	0.130551399	0.03912086
NOX	-10.3211828	0.008293859
DISTANCE	0.261093575	0.000137546
TAX	-0.01440119	0.000251247
PTRATIO	-1.074305348	6.58642E-15
AVG_ROOM	4.125409152	3.89287E-19
LSTAT	-0.603486589	8.91071E-27

Table 7. coefficients and p value

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?

Regression states from the previous model

Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372

Regression states for this model

Multiple R = 0.832835773

R square = 0.693615426

By comparing Multiple R and R square values for both the models we can conclude that both models perform well.

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?

	Coefficients
NOX	-10.3211828
PTRATIO	-1.074305348
LSTAT	-0.603486589
TAX	-0.01440119
AGE	0.032770689
CRIME_RATE	0.048725141
INDUS	0.130551399
DISTANCE	0.261093575
AVG_ROOM	4.125409152
Intercept	29.24131526

Table 8.coefficient

The coefficient of NOX is negative that means inversely proportional

NOX increase price is decrease

d) Write the regression equation from this model.

Y= 0.0327706 X0 + 0.1305513 X1 -10.27270508 X2 +0.261506423 X3 - 0.014452345 X4 -1.071702473 X5 + 4.125468959 X6 -0.605159282 X7 +29.42847349

Where Y = AVG_PRICE

X0 = Age

X1 = Indus

X2 = NOX

X3 = Distance

X4 = TAX

X5 = PTRATIO

X6 = AVG_ROOM

X7 = LSTAT

Summary:

From this Analysis, we can conclude that the average price of the house excluding crime rate

Negative coefficients which say that increase rate in those features will decrease the average price of the house like NOX, PTRATIO, TAX and LSTAT.