A

REPORT

ON

TOPIC: Terro's real estate agency

Real estate data analysis – Exploratory data analysis, Linear Regression

Submitted by,

Muthamma P V

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Introduction:

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.

Problem Statement:

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

Data Dictionary:

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

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

Table 1: Data Description

Objective:

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

Q1: DESCRIPTIVE STATISTICS

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
Mean	4.872	68.575	11.137	0.555	9.549	408.237	18.456	6.285	12.653	22.533
Standard Error	0.130	1.251	0.305	0.005	0.387	7.492	0.096	0.031	0.317	0.409
Median	4.820	77.500	9.690	0.538	5.000	330.000	19.050	6.209	11.360	21.200
Mode	3.430	100.000	18.100	0.538	24.000	666.000	20.200	5.713	8.050	50.000
Standard Deviation	2.921	28.149	6.860	0.116	8.707	168.537	2.165	0.703	7.141	9.197
Sample Variance	8.533	792.358	47.064	0.013	75.816	28404.759	4.687	0.494	50.995	84.587
Kurtosis	-1.189	-0.968	-1.234	-0.065	-0.867	-1.142	-0.285	1.892	0.493	1.495
Skewness	0.022	-0.599	0.295	0.729	1.005	0.670	-0.802	0.404	0.906	1.108
Range	9.950	97.100	27.280	0.486	23.000	524.000	9.400	5.219	36.240	45.000
Minimum	0.040	2.900	0.460	0.385	1.000	187.000	12.600	3.561	1.730	5.000
Maximum	9.990	100.000	27.740	0.871	24.000	711.000	22.000	8.780	37.970	50.000
Sum	2465.220	34698.900	5635.210	280.676	4832.000	206568.000	9338.500	3180.025	6402.450	11401.600
Count	506.000	506.000	506.000	506.000	506.000	506.000	506.000	506.000	506.000	506.000

Fig 1.1: Descriptive Statistics

CRIME_RATE:

- The average crime rate around the town's in the locality is 4.872.
- Standard deviation of crime rate is 2.92 indicates that crime rate varies from town to town.
- The skewness value of crime rate is 0.022 which indicates that the data points are symmetric in nature or they are said to have a normal distribution.

AGE:

- The average age resulted is 68.5%, which indicates that proportion of houses built prior to 1940 is more.
- The skewness of the ages of the houses is -0.599.

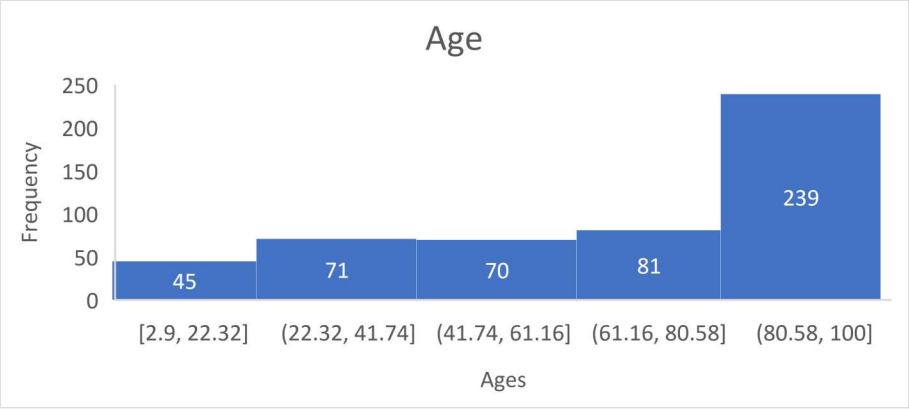

Fig 1.2: Histogram of age

It indicates that the data has negative skewness or the data distribution is said to be trailing off to the left that indicates that most of the houses in the town are aged.

- The range of age of houses is 97.1, which indicates that there exist a wide range of ages of houses in the town. The houses in the town seems to have a maximum age of 100 years, and there are also new houses with the minimum of age of 2.9 years.

INDUSTRY:

- The average percentage of Industry is 11.137% , which indicates that average proportion of non-retail business acres of 506 towns is about 11.137%.
- Standard deviation is about 0.30 which indicates that there is no much variation the data points, they are closer to mean.
- The range of industries in town is 27.28, which indicates that some town areas have more number of non-retail business area while some has less non-retail business area.

NOX:

- The average nitric oxides concentration is 0.555 parts per 10 million. The increase in nitric oxide concentration indicates that there exist pollution in the locality which will reduce the business value of the property.
- The nitric oxides concentration is skewness value is 0.729.

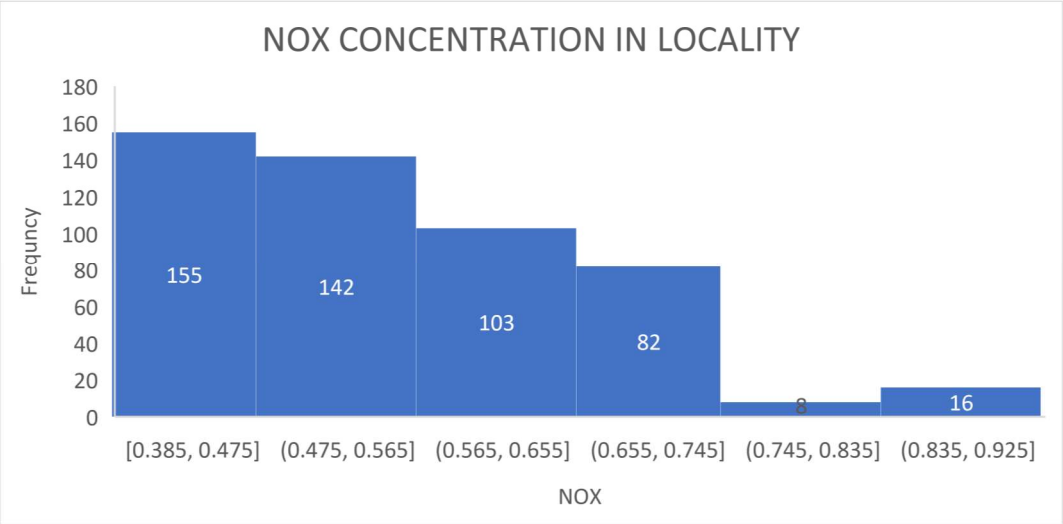

Fig 1.3: Histogram of NOX

It indicates that the data has positive skewness or the data distribution is said to be trailing off to the right that indicates that most of the houses in the town had a nitric oxide concentration to be lower than the average NOX value. We can say that air was not polluted.

DISTANCE

- The average distance of the houses in the locality from the highway is 9.549 miles.

- Skewness of the distance is 1.005 which indicates that there exist positive skewness, the data distribution is trailing off to the right.

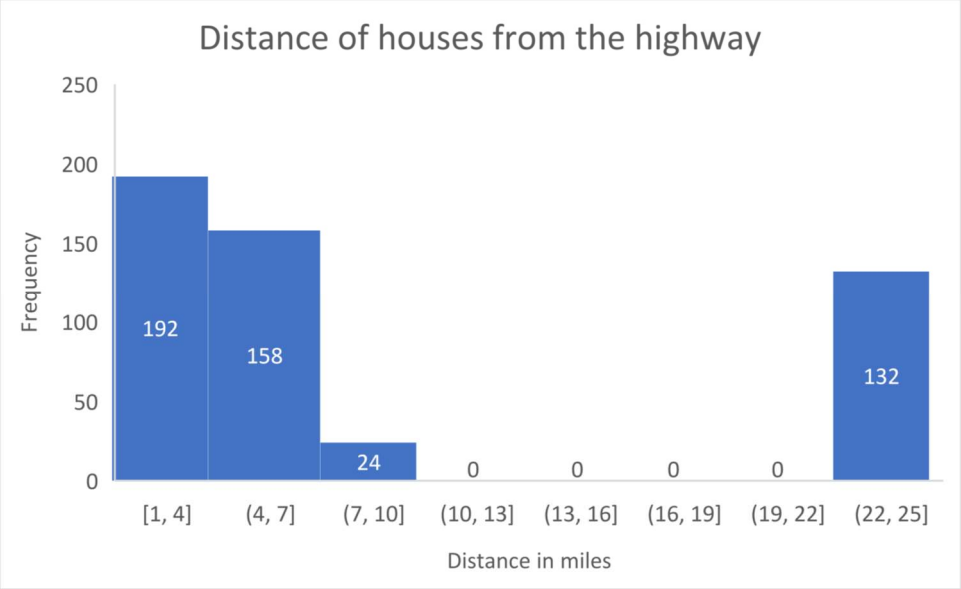

Fig 1.4: Histogram of Distance

- Out of the 520 houses nearly 375 houses lies within a distance of 1-10miles and only132 houses lies within a distance 22-25 houses which indicates that most of the house owners preferred a lesser distance from highway.
- The range of the distance is 23miles which indicates that the distances from the highway to the houses vary and the longest distance is said to be 24miles while the least distance is observed to be 1miles from the highway.

TAX

- The average property-tax rate is 408.237 per $10,000.
- The standard deviation is 168.537, which indicates that there exist variation in the tax rates applied on the property across the different towns. Some areas might have high tax rates while some might have lower tax rates.
- The kurtosis value is -1.142, indicating that it's platykurtic, we can say if a curve is drawn across data points it will have a flat peak.

PTRATIO:

- The average pupil-teacher ratio is 18.456 which indicates on an average, there are 18.456 students for every teacher who are utilizing the education facility in the local schools.

- The range of pupil-teacher ratio is 9.400 which indicates that the strength of the schools vary from in town to town. Minimum pupil-teacher ratio is 12.600, while the maximum is 22.000.
- The pupil-teacher ratio of 20.200 is the most commonly observed ratio among the towns.

AVERAGE_ROOMS:

- On an average, houses in the locality have around 6.285 rooms.
- The median of the average rooms is 6.209 which means that half of the houses have more than 6.209 rooms, and the other half have lesser than 6.209 rooms.

LSTAT:

- On average 12.653% of the population are of lower status .
- The percentage of lower status population of 8.050% is the most frequently observed percentage in the locality.

AVERAGE PRICE:

- The mean of average price of the houses in the locality is $22,533.
- The average price has a standard deviation of 9.197 which suggests that average price of the houses vary in the locality.
- The skewness value of 1.495 indicates that it's a positive skewness, the distribution is trailing off to the right.

Q2: HISTOGRAM OF AVG_PRICE

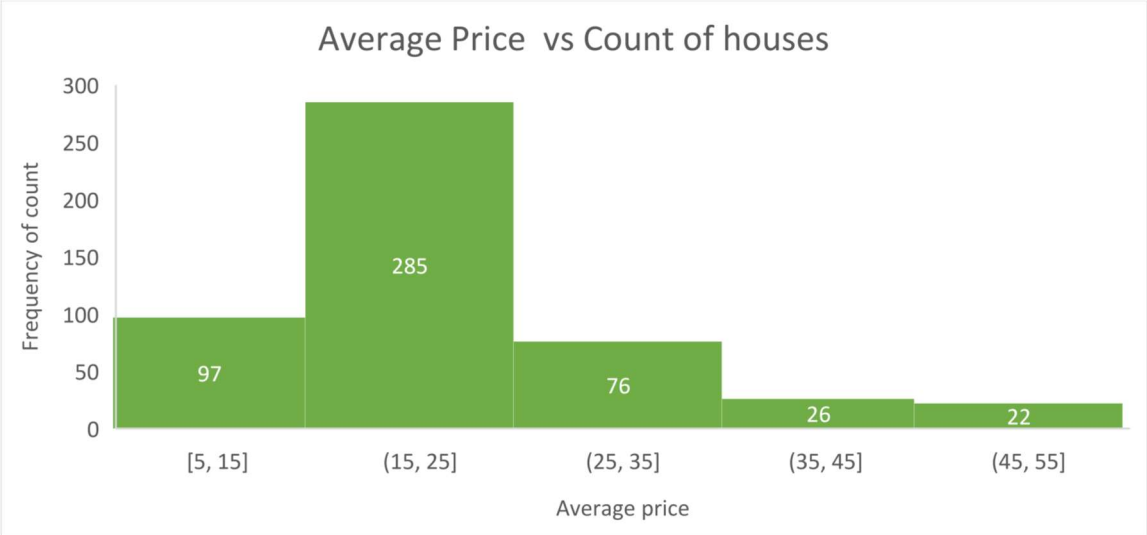

Fig 2.1: Histogram of AVG_PRICE

- When the data is univariant and numeric and are continuous then histogram is used.
- In the x-axis we have average price, y-axis indicates the count of houses.
- Maximum number of houses has lower average price. Only few houses have a higher average price value.
- Nearly out of 520houses 285 houses in the locality have their house property value to be in range of $15000 to $25000.
- Only few of the property buyers have purchased house between $35000 and $55000.
- The data points is not Normally distributed. There exist irregularity in the data points otherwise the data points appear to be asymmetric in nature.
- By looking at the histogram we can say that data points appear to be right skewed or positive skewness, a large number of data values occur on the left side and fewer data on the right side.
- The mean value will be affected by the fewer data points that is present on the right side.
- From histogram we can infer that the mean of the average price is greater than median of average price.

Q3: COVARIENCE

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

Fig 3.1: Covariance Matrix

Covariance is a measure of how one variable change when the other variable change. It mainly tells us about the direction of relationship among the variables. The value of the covariance could be positive or negative and it could be zero.

Positive covariance(>0): Two variables tend to change in the same direction. In other words, meaning if one variable increase the other one also increase or if one variable decrease the other one will also decrease.

In covariance matrix generated the values having positive covariance is formatted in 'green'. For example the Tax and Age here have the positive covariance(covariance=2397.942). We can say the higher the property-tax rate, higher was the proportion of houses built prior to 1940.

Negative Covariance(<0): Two variables tend to change in the opposite direction, one variable increase the other one will decrease or vice versa.

In the covariance matrix generated the values having negative covariance is formatted in 'red'. For example the lower status population (LSTAT) and the average home price(covariance= -

48.35). This indicates that areas with a higher percentage of lower status population will have lower average home prices or vice-versa.

No Covariance(=0): If covariance is 0 indicates that changes in one variable are not associated with changes in the other variable.

In the covariance matrix generated the values having covariance near to zero is formatted in 'white'. For example the NOX and the crime rate (covariance= 0.001). Since the value is close to zero, we can say that changes in NOX are not strongly associated with the changes in the crime rate.

Q4: CORRELATION

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	1.0000									
AGE	0.0069	1.0000								
INDUS	-0.0055	0.6448	1.0000							
NOX	0.0019	0.7315	0.7637	1.0000						
DISTANCE	-0.0091	0.4560	0.5951	0.6114	1.0000					
TAX	-0.0167	0.5065	0.7208	0.6680	0.9102	1.0000				
PTRATIO	0.0108	0.2615	0.3832	0.1889	0.4647	0.4609	1.0000			
AVG_ROOM	0.0274	-0.2403	-0.3917	-0.3022	-0.2098	-0.2920	-0.3555	1.0000		
LSTAT	-0.0424	0.6023	0.6038	0.5909	0.4887	0.5440	0.3740	-0.6138	1.0000	
AVG_PRICE	0.0433	-0.3770	-0.4837	-0.4273	-0.3816	-0.4685	-0.5078	0.6954	-0.7377	1.0000

Fig 4.1: Correlation Matrix

Correlation analysis is a used to determine the strength and direction of the relationship between the variables, where Strength indicates whether the variables have strong linear relationship.

The correlation values ranges from -1 to +1, where '-' and '+' indicates the direction of relationship among variables. '-'ve indicates that if one increase other decreases or vice-versa while '+'ve indicates either both the variables will increase together or decrease together.

If value of correlation is greater than zero then we can say that it has positive linear relationship. If the correlation value is less than zero we say that variables have a negative linear relationship. If it is equal to zero then we say that there exist no relationship among the variables.

Below are the few observations from the correlation matrix generated:

1. Top 3 positively correlated pairs:

i) TAX, DISTANCE: 0.9102

Here the correlation value is closer to 1, from which can say that there exist significantly strong positive linear relationship between Tax and Distance. Both Tax and Distance increase or decrease together.

ii) NOX, INDUSTRY: 0.7637

Here the correlation value is 0.7637, from which can say that there exist almost a strong positive linear relationship between NOX and Industry. Both NOX and Industry increase or decrease together.

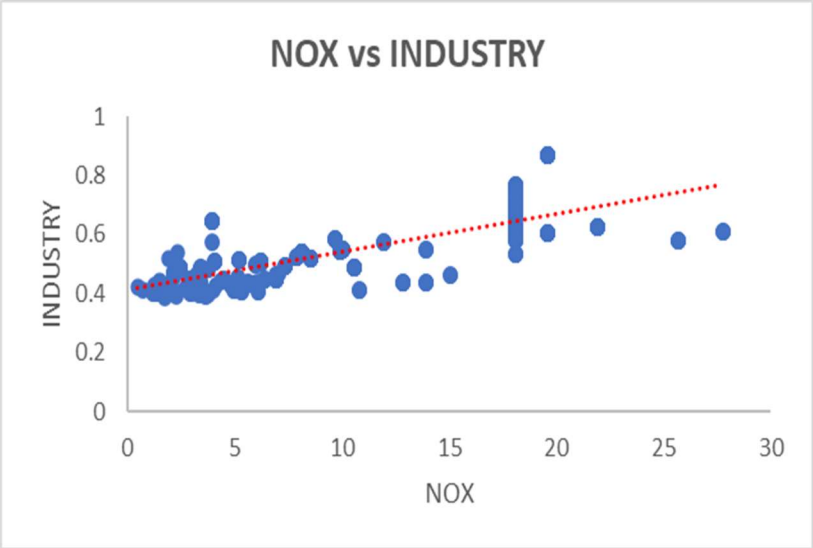

Fig 4.2: Scatter plot of NOX vs Industry

From the above Scatter plot we can visualize the existence of positive Linear relationship between NOX and Industry.

iii)NOX, AGE: 0.7315

Here the correlation value is 0.7317, from which can say that there exist approximately strong positive linear relationship between NOX and Age. Both NOX and Age increase or decrease together.

1. <u>Top 3 negatively correlated pairs:</u>
   i) AVG_PRICE, LSTAT: -0.7377

Here the correlation value is -0.7377, from which can say that there exist very strong negative linear relationship between Average price and LSTAT. Here if Avg_price increase, LSTAT will decrease or vice-versa.

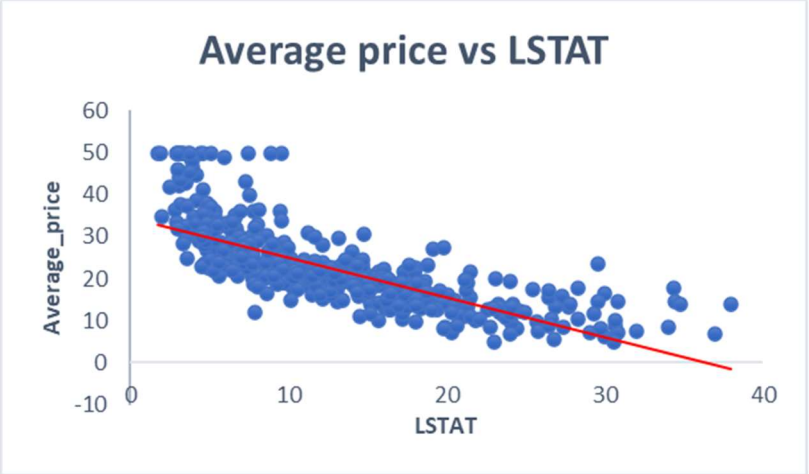

Fig 4.3: Scatter plot of AVG_PRICE vs LSTAT

From the above Scatter plot we can visualize the existence of negative Linear relationship between AVG_PRICE and LSTAT.

ii) LSTAT, AVG_ROOM: -0.6138

Here the correlation value is -0.6138, from which can say that there exist strong negative linear relationship between LSTAT and Average room. Here if AVG_ROOM increase, LSTAT will decrease or vice-versa.

iii) AVG_PRICE, PRATIO: -0.5078

Here the correlation value is -0.7377, from which can say that there exist strong negative linear relationship between Average price and LSTAT. Here if AVG_PRICE increase, PRATIO will decrease or vice-versa.

Q5: Regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as Independent Variable.

a) Inferences from the Regression Summary

Regression is a statistical analysis method for modeling the relationship between dependent variable and one or more independent variables. It is used to predict the dependent variable based on the independent variables.

Here we have considered AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as 'x'(Independent Variable). A regression with one variable is referred to as Simple Linear Regression. Below is the summary output of the regression.

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	23243.91	601.6178711	5.0811E-88			
Residual	504	19472.38142	38.63568					
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	61.41515	3.7431E-236	33.44845704	35.65922472	33.44845704	35.65922472
LSTAT	-0.950049354	0.038733416	-24.5279	5.0811E-88	-1.0261482	-0.873950508	-1.0261482	-0.873950508

Fig 5.1: Model 1- Regression Statistics of LSTAT vs AVG_PRICE

Evaluating Fit of the Model:

Multiple R : Multiple R stands for Multiple correlation coefficient. It measures the strength and direction of the linear relationship between variables. The multiple R is 0.7376, which indicates small positive linear relationship between the AVG_PRICE and LSTAT.

R Square: R Square value is 0.5441 which indicates that 54.4% of the variation in AVG_PRICE can be explained by the LSTAT alone.

Adjusted R square: It indicates weather the number of independent variable included in model, and tell if it's useful or irrelevant. Adjusted R square value here is 0.543. Since it is a simple linear regression, there exist one variable and adjusted R square does not provide more insight.

Intercept: The intercept value is 34.5538 which means that when LSTAT have a value of zero, the estimated Average price value is 34.5538.

Coefficient of LSTAT : The coefficient value obtained is -0.9500 which indicates for each one-unit increase in the LSTAT, the estimated average value of the house price decreases by 0.9500 units.

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

P-value: The p-value indicates the weather the independent variable is significant or not. Here the p-value of LSTAT appears to be 5.0811E-88 which is very close to zero, LSTAT is highly significant.

RESIDUAL SUMMARY

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

Fig 5.2: Model 1- Residual output from regression

Predicted AVG_PRICE: Indicates the predicted average price for each observation.

Residuals: Residuals are the differences between the actual and the predicted value.

Residual=Actual_Value – Predicted_Value.

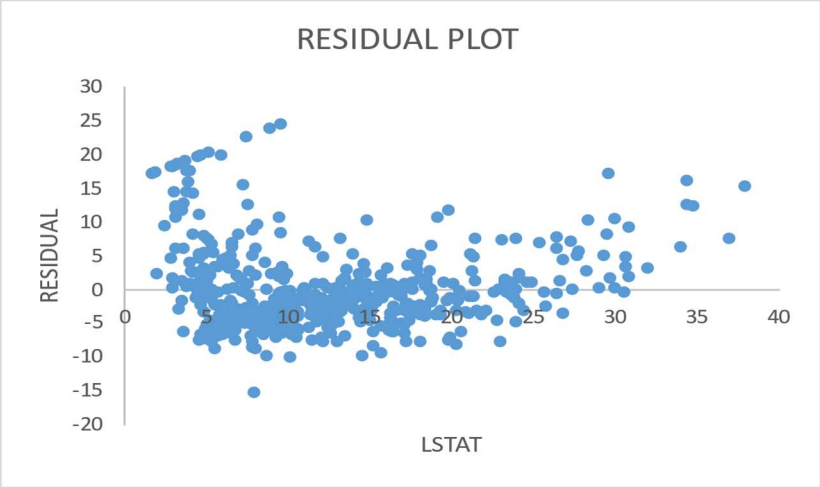

Fig 5.3: Residual Plot

The points above zero in the plot represents a positive residual. This means the observed value for AVERAGE_PRICE is greater than the value predicted by the regression model.

The points below zero represents a negative residual. This means the observed value for AVERAGE_PRICE is less than the value predicted by the regression model.

Since the points in the plot are scattered around a residual value of 0 with no clear pattern, this indicates that the model prediction is almost accurate.

PREDICTION ACCURACY: MAPE AND RMSE:

ROOT MEAN SQUARED ERROR(RMSE)					
R^2	*Mean(R^2)*	*Sq.Root(Mean)*	*Max Error*	*Min Error*	***RMSE***
33.90261367	38.48296723	6.203464131	22.53	0	**28%**
18.23622892					
15.79949625					

MEAN ABSOLUTE PERCENTAGE ERROR(MAPE)			
AVG_PRICE	*Percentage Error(PE)*	*Absolute PE*	***MAPE***
24	-0.242608129	0.242608129	**21%**
21.6	-0.197703231	0.197703231	
34.7	0.114549222	0.114549222	

Fig 5.4: Model 1-RMSE and MAPE calculation

RMSE (Root Mean Square Error)**:** RMSE of 28% indicates that, the predictions made by model have an error of 28% when compared to the actual values.

MAPE (Mean Absolute Percentage Error): MAPE of 21% is the percentage difference between the predicted and actual values.

We can say that predictions made are somehow close to the actual values.

ASSUMPTION CHECK

ASSUMPTION CHECK		
1.Residual Mean 0	-2.7365E-14	MET
2.Residual Skew <0.5	1.457061987	NOT MET
3.Constant Varience		MET

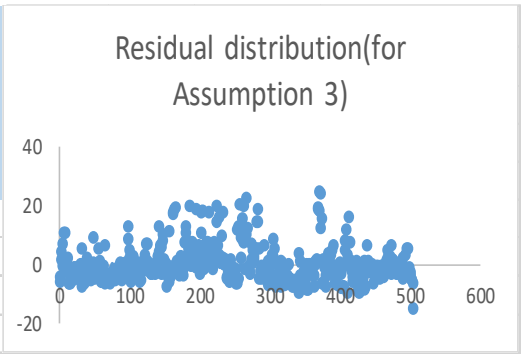

Fig 5.5: Model 1-Calculation for Assumption check

1.Residual Mean obtained was Zero.

2.Based on the value obtained from skewness we can say that Residual was not normally distributed.

3.Varience of residual is approximately constant.

From which we can say that model generated is not a best fit model.

Q6: Regression model including LSTAT and AVG_ROOM together as Independent variables and AVG_PRICE as dependent variable.

Here we have considered AVG_PRICE as 'y' (Dependent variable) and LSTAT and AVG_ROOM variable as 'x'(Independent Variables). A regression with two independent variable is referred to as Multiple Linear Regression. Below is the summary output of the regression.

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	-0.42809535	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	-14.6886992	6.66937E-41	-0.728277167	-0.556439501	-0.728277167	-0.556439501

Fig 6.1: Model 2- Regression Statistics of AVG_ROOM, LSTAT vs AVG_PRICE

Multiple R: Multiple R of 0.7991, means there is positive linear relationship between independent variables (LSTAT and AVG_ROOM) and the AVG_PRICE.

R Square: R square of 0.638561 meaning 63.86% of the variation in the AVG_PRICE is explained by the LSTAT and AVG_ROOM.

P-value: P-value for both LSTAT AND AVG_ROOM is closer to zero indicating that both of these variable's are highly significantly.

Intercept: The intercept value is 34.5538 which means that when LSTAT and AVG_ROOM have a value of zero, the estimated Average price value is 34.5538.

Coefficient of AVG_ROOM :The coefficient value obtained is 5.0947 which indicates for each one-unit increase in the AVG_ROOM, the estimated average value of the house price increases by 5.0947.

Coefficient of LSTAT :The coefficient value obtained is -0.6423 which indicates for each one-unit increase in the LSTAT, the estimated average value of the house price decreases by 0.6423.

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?

We can write the Regression equation as:

$Y_{AVERAGE\_PRICE} = Intercept + 5.0947 * X_{AVG\_ROOM} - 0.6423 * X_{LSAT}$

If a new house in this locality has 7 rooms (on an average) and has a value of 20 for L-STAT, then the value of AVG_PRICE will be,

AVG_PRICE = -1.3582 + 5.09 * 7 - 0.64 * 20
AVG_PRICE = 21.45807639

If a company is quoting a value of 30000 USD for this locality, We can see that the predicted value for average price of a house is $21,458.07 which is lesser then the price quoted by the company. We can say that the 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.

Adjusted R-square comparison:

Model	Adjusted R Square
LSTAT vs AVG_PRICE	0.543241826
LSTAT, AVG_ROOM vs AVG_PRICE	0.637124475

TABLE 2: Adjusted R-square comparison

We can see that the Adjusted R square value of LSTAT vs AVG_PRICE is lower when compared to the Adjusted R square value of LSTAT, AVG_ROOM vs AVG_PRICE. The adjusted R-squared values tell that about 54.32% of the variation in AVG_PRICE can be explained by the LSTAT and about 63.71% of the variation in AVG_PRICE can be explained by LSTAT and AVG_ROOM together which indicates that the performance of this model is better than the previous model that contained only LSTAT.

Q7. 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 R square, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372
Adjusted R Squar	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

Fig 7.1: Model 3- Regression Statistics of all independent vs AVG_PRICE

Adjusted R square: Adjusted R square value obtained is 0.6882 that is about 68.82% of the variability in AVG_PRICE is explained by the independent variables considered. The value of Adjusted R square obtained for this model is slightly higher than the previous two model indicating that it's performance is better than the last two models.

Intercept: The intercept value obtained is 29.24131526, which indicates the value of average price when rest of values of variables becomes zero.

Coefficient values:

	Coefficients	*Inferences*
CRIME_RATE	0.048725141	For every one-unit increase in crime rate, the average house price increases by $0.0487
AGE	0.032770689	For every one-unit increase in proportion of houses built prior to 1940 , the average house price increases by $0.0327
INDUS	0.130551399	For every one-unit increase in non retail business acres, the average house price increases by $0.1305
NOX	-10.3211828	For every one-unit increase in nitic oxide concentration, the average house price decreases by $10.3211
DISTANCE	0.261093575	For every one-unit increase in distance from highway, the average house price increases by $0.2610
TAX	-0.01440119	For every one-unit increase in property-tax rate , the average house price decreases by $0.0144
PTRATIO	-1.07430534	For every one-unit increase in pupil-teacher ratio , the average house price decreases by $1.0743
AVG_ROOM	4.125409152	For every one-unit increase in average number of rooms per house, the average house price increases by $4.1254
LSTAT	-0.60348658	For every one-unit increase in % lower status of the population, the average house price decreases by $0.6034

TABLE 3: Inferences on Coefficient

P-VALUE: P-value<0.05 indicates that variable is significant, while a high p-value>0.05 suggests the variable is not be significant.

	P-VALUE	***INFERENCES***
CRIME_RATE	0.534657201	P-value is >0.05, Crime rate is not significant in predicting house prices.
AGE	0.012670437	P-value is <0.05, AGE is significant factor in predicting house prices.
INDUS	0.03912086	P-value is <0.05, Industry is significant factor in predicting house prices.
NOX	0.008293859	P-value is <0.05, NOX is significant factor in predicting house prices.
DISTANCE	0.000137546	P-value is <0.05, Distance is significant factor in predicting house prices.
TAX	0.000251247	P-value is <0.05, TAX is significant factor in predicting house prices.
PTRATIO	6.58642E-15	P-value is <0.05, PRATIO is significant factor in predicting house prices.
AVG_ROOM	3.89287E-19	P-value is <0.05, AVG_ROOM is significant factor in predicting house prices.
LSTAT	8.91071E-27	P-value is <0.05, LSTAT is significant factor in predicting house prices.

TABLE 4: Inferences on P-value

Based on the value of p-value of significance we can discard the variable crime_rate.

Q8: 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:

Independent variables : NOX, PTRATIO, LSTAT, TAX, AGE, INDUS, DISTANCE, AVG_ROOM.

Dependent variable: AVG_PRICE

a) Interpret the output of this model.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.832835773							
R Square	0.693615426							
Adjusted R Square	0.688683682							
Standard Error	5.131591113							
Observations	506							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	8	29628.68142	3703.585178	140.6430411	1.911E-122			
Residual	497	13087.61399	26.33322735					
Total	505	42716.29542						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
NOX	-10.27270508	3.890849222	-2.640221837	0.008545718	-17.9172457	-2.628164466	-17.9172457	-2.6281645
PTRATIO	-1.071702473	0.133453529	-8.030529271	7.08251E-15	-1.333905109	-0.809499836	-1.33390511	-0.8094998
LSTAT	-0.605159282	0.0529801	-11.42238841	5.41844E-27	-0.70925186	-0.501066704	-0.70925186	-0.5010667
TAX	-0.014452345	0.003901877	-3.703946406	0.000236072	-0.022118553	-0.006786137	-0.02211855	-0.0067861
AGE	0.03293496	0.013087055	2.516605952	0.012162875	0.007222187	0.058647734	0.007222187	0.05864773
INDUS	0.130710007	0.063077823	2.072202264	0.038761669	0.006777942	0.254642071	0.006777942	0.25464207
DISTANCE	0.261506423	0.067901841	3.851242024	0.000132887	0.128096375	0.394916471	0.128096375	0.39491647
AVG_ROOM	4.125468959	0.44248544	9.323400461	3.68969E-19	3.256096304	4.994841615	3.256096304	4.99484161
Intercept	29.42847349	4.804728624	6.124898157	1.84597E-09	19.98838959	38.8685574	19.98838959	38.8685574

Fig 8.1: Model 4: Regression Statistics of all significant independent variables vs AVG_PRICE

Evaluating Fit of the Model:

Multiple R :The Multiple R value obtained is 0.8328,it indicates a strong positive linear relationship between the significant independent variables and the dependent variable.

R Square: R Square value obtained is 0.6936,meaning 69.36% of the variability in average house prices is explained by the combination of significant independent variables.

Adjusted R Square: Adjusted R square value obtained is 0.6886 that is about 68.86% of the variability in AVG_PRICE is explained by significant independent variables, which means the number of significant variables chosen are useful in making predictions.

P-value: The p-value obtained for all less than 0.05, indicates all of them are significant in predicted house price.

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?

Adjusted R Square comparison:

Model	Adjusted R Square
All independent variables vs AVG_PRICE	0.688298647
Significant independent variable vs AVG_PRICE	0.688683682

TABLE 5: Adjusted R Square comparison

The adjusted R-squared for the 'Significant independent variable vs AVG_PRICE' model is slightly higher than the adjusted R-square value of 'All independent variables vs AVG_PRICE' model. The difference in the adjusted R-squared values between these two models is very small.

On comparison , we can say that a regression model with only significant independent variables performs better.

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?

Coefficient: The coefficient value for NOX, PRATIO, LSTAT and TAX obtained is negative which mean for every unit increase in these variables the AVERAGE_PRICE tend to decrease. The coefficient value for AGE, INDUS, DISTANCE and AVG_ROOM obtained is positive which mean for every unit increase in these variables the AVERAGE_PRICE tend to increase.

On sorting the Coefficients in ascending order, the result observed shown below.

	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

Fig 8.2: Sorted Coefficient

We can say that based on NOX coefficient that for every unit increase in the NOX the average price tend to decrease by $10.2727. Assuming Nitric oxide concentration is more in the locality we can say that the average price tend to decrease.

d) Write the regression equation from this model.

Regression Equation:

$$Y_{AVG\_PRICE} = -10.2727 * X_{NOX} - 1.0717 * X_{PTRATIO} - 0.6051 * X_{LSTAT} - 0.014 * X_{TAX} + 0.0329 * X_{AGE} + 0.1307 * X_{INDUS} + 0.2615 * X_{DISTANCE} + 4.1254 * X_{AVG\_ROOM}.$$

Residual Summary:

RESIDUAL OUTPUT		
Observation	*Predicted AVG_PRICE*	*Residuals*
1	30.04888734	-6.048887337
2	27.04098462	-5.440984617
3	32.69896454	2.001035462

Fig 8.3: Residual Summary

Predicted AVG_PRICE: Indicates the predicted average price for each observation.

Residuals: Residuals are the differences between the actual and the predicted values. Residual=Actual_Value – Predicted_Value.

PREDICTION ACCURACY: MAPE AND RMSE:

ROOT MEAN SQUARED ERROR(RMSE)					
R^2	*Mean(R^2)*	*Sq.Root(Mean)*	*Max Error*	*Min Error*	***RMSE***
36.58903801	25.86484979	5.085749678	22.3	0	23%

MEAN ABSOLUTE PERCENTAGE ERROR(MAPE)			
AVG_PRICE	*Percentage Error(PE)*	*Absolute PE*	***MAPE***
24	-0.252036972	0.252036972	18%
21.6	-0.251897436	0.251897436	
34.7	0.057666728	0.057666728	
33.4	0.06757277	0.06757277	

Fig 8.4: Calculation of RMSE and MAPE

RMSE (Root Mean Square Error)**:** RMSE of 23% indicates that, the predictions made by model have an error of 23% when compared to the actual values.

MAPE (Mean Absolute Percentage Error): MAPE of 18% is the percentage difference between the predicted and actual values.

We can say that predictions made for average price are close to the actual values.

ASSUMPTION CHECK

ASSUMPTION CHECK		
1.Residual Mean 0	-1.03948E-14	MET
2.Residual Skew <0.5	1.643869514	NOT MET
3.Constant Varience		MET

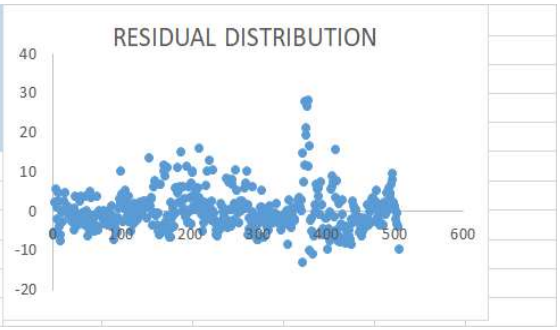

Fig 8.6: Model 4-Calculation for Assumption check

1.Residual Mean obtained was Zero.

2.Based on the value obtained from skewness we can say that Residual was not normally distributed.

3.Varience of residual is approximately constant.

From which we can say that model generated is not a best fit model.

CONCLUSION

In this business report, we conducted a comprehensive analysis of both descriptive statistics and regression using Data analysis Tool Pak. Decriptive statistics helped in understanding of Terro's Real Estate Agency dataset by including statistical measures like mean, median, mode, standard deviation, skewness, kurtosis, minimum, maximum, range, count. Multiple Regression models helped in identifying relationship between Averge_price and all other independent variables, out of which we understood that all variables except crime_rate can be considered a significant in determining the average house price and based on parameters like Adjusted R square we can conclude that prediction of house prices made by the model is somewhat accurate.