Business Report

Assignment – Terro's real estate agency

Real estate data analysis – Exploratory data analysis, Linear Regression

BY

K S MUGILAN

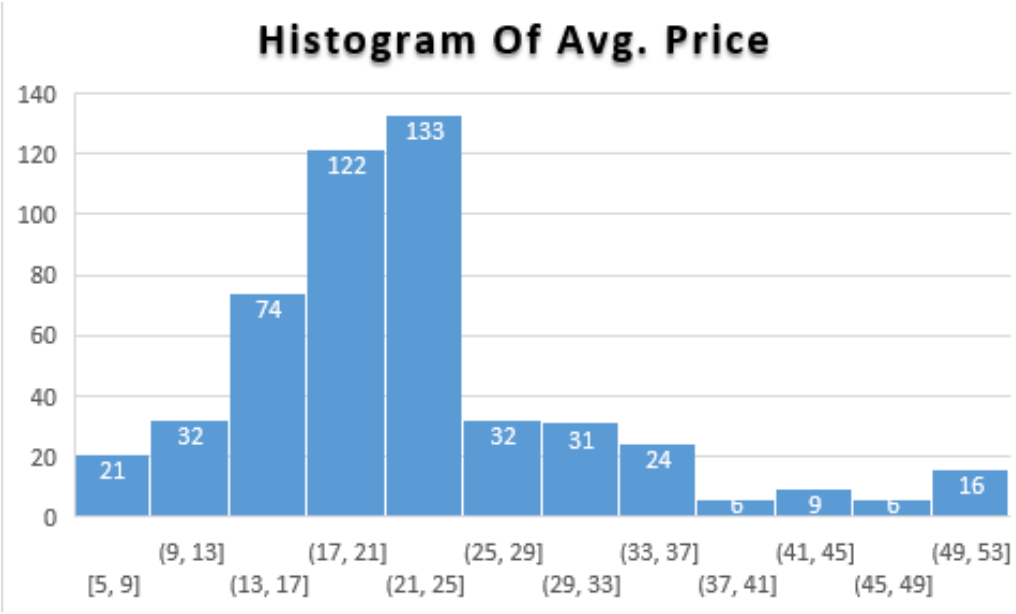
1) **Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation.**

Ans: Used the Descriptive Statistics tool in the Data Analysis Tool pack for summary statistics for each variable. The observations that are interpret from the table.

- Avg_room has the highest kurtosis among all other variables means that its distribution is more peaked and has heavier tails compared to the rest. This suggests that the data for Avg_room is more occures around its mean, the positive skewness value of 0.4036 suggests that the data is slightly skewed to the right.
- The average price of the mean and median are 22.53 and 21.20, The presence of outliers, not far from the mean. Both the maximum price sold and the mode are 50 suggests that many purchases around this price, indicating a common preference or popular pricing point.
- The standard deviation of PTRATIO and Avg_room indicates that most values are close to the average, with less variability or spread in the data. As mean and median mostly same so that outliers are not much deviate from the mean.
- In a dataset with a mean distance of 9.8 units and a standard deviation of 8.4 units, the wide spread in distances indicates significant variability around the average distance traveled, suggesting diverse travel patterns among the data points.
- The maximum age is 100 years, and the most frequently (mode) is also 100. The negative kurtosis indicates that the distribution is less peaked than a normal distribution, with fewer extreme values. Similarly, the negative skewness suggests that the data is skewed to the left, meaning there are more houses with ages below the average.
- In the given dataset , crime rate with the mean and median are close at 4.87 and 4.82, that indicating no outliers. The highest crime rate is 9.99, with the most common rate being 3.43 (mode). The negative skewness suggests that the data is skewed to the left. Additionally, the positive kurtosis indicates that the distribution has heavier tails and is more peaked around the mean.

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

Ans:



By observing the histogram, we notice that the data is skewed towards the left, indicated by a longer tail on the left side. This suggests a positive skewness, indicates that there are more data occurrences of lower values in the dataset**.**

3) Compute the covariance matrix. Share your observations.

Ans:

	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

a) Age vs Tax, Indus vs Tax, and Distance vs Tax has higher covariance, suggesting a direct relationship between them. This means that as one variable increases, the other variable also tends to increase.

b) Tax vs Avg Price, Age vs Avg Price, and LSTAT vs Avg Price has the negative covariance, indicating an inverse relationship. This means that when one variable increases, the other variable tends to decrease.

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**.

Ans: We use correlation tool in the data analysis tool pack for correlation matrix

CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
1									
0.006859463	1								
-0.005510651	0.644778511	1							
0.001850982	0.731470104	0.763651447	1						
-0.009055049	0.456022452	0.595129275	0.611440563	1					
-0.016748522	0.506455594	0.72076018	0.6680232	0.910228189	1				
0.010800586	0.261515012	0.383247556	0.188932677	0.464741179	0.460853035	1			
0.02739616	-0.240264931	-0.391675853	-0.302188188	-0.209846668	-0.292047833	-0.355501495	1		
-0.042398321	0.602338529	0.603799716	0.590878921	0.488676335	0.543993412	0.374044317	-0.613808272	1	
0.043337871	-0.376954565	-0.48372516	-0.427320772	-0.381626231	-0.468535934	-0.507786686	0.695359947	-0.737662726	1

Top 3 positively correlated pairs	
1.Distance vs Tax	0.910228189
2.Indus vs Nox	0.763651447
3.Age vs Nox	0.731470104

a) Distance vs tax , Indus vs Nox, Age vs Nox has higher correlation, suggesting a direct relationship between them. This means that as one variable increases, the other variable also tends to increase.

Top 3 negatively correlated pairs	
1.LSTAT vs Avg_Price	-0.73766273
2.LSTAT vs Avg_room	-0.61380827
3.PTRATIO vs Avg_price	-0.50778669

b) LSTAT vs Avg_Price, LSTAT vs Avg_room, PTRATIO vs Avg_price has the negative correlation, indicating an inverse relationship. This means that when one variable increases, the other variable tends to decrease.

5) **Build an initial regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot. a) What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot? b) Is LSTAT variable significant for the analysis based on your model?**

Ans: We use Regression tool in the data analysis tool pack for correlation matrix

Regression Statistics	
Multiple R	0.737663
R Square	0.544146
Adjusted R Square	0.543242
Standard Error	6.21576
Observations	506

	Coefficients	*andard Err*	*t Stat*	*P-value*
Intercept	34.55384	0.562627	61.41515	3.7E-236
LSTAT	-0.95005	0.038733	-24.5279	5.08E-88

a) By analyzing the coefficient value and intercept, we can say coefficient value increase by 1, the average price decreases by 0.95 , indicating a negative relation. However, the positive intercept suggests that increase the price. The residual plot is randomly scattered around all axis and no pattern, it indiactes that regression model is appropriate.

b) As per the regression model, the p value of LSTAT is 5.06E-88 is less than 0.05 value and it is significant. The coefficient value is negative and it is inversely proportion to the avg_price.

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?**

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.

Ans: We use Regression tool in the data analysis tool pack for correlation matrix

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

	Coefficients	*andard Err*	*t Stat*	*P-value*
Intercept	-1.35827	3.172828	-0.4281	0.668765
AVG_ROOM	5.094788	0.444466	11.46273	3.47E-27
LSTAT	-0.64236	0.043731	-14.6887	6.67E-41

a) Regression Equation:

Y = (5.094788)*(AVG_ROOM) + (-0.64236)*(LSTAT) -1.35827

Avg_room = 7, LSTAT= 20

Then **Avg_Price = 21.454**

The company's quoted value for this locality is 30,000 USD, which higher than our predicted value of 21,458.076 USD. Therefore, it appears that the company is **overcharging**.

b) While comparing this model with previous model, this model adjusted R square value is 0.637 and the previous model is 0.543.The Adjusted R square value of this model is high, the performance of 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 R square, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE.**

Ans: We use Regression tool in the data analysis tool pack for correlation matrix

Regression Statistics	
Multiple R	0.832979
R Square	0.693854
Adjusted R Square	0.688299
Standard Error	5.134764
Observations	506

	Coefficients	*andard Err*	*t Stat*	*P-value*
Intercept	29.24132	4.817126	6.070283	2.53978E-09
CRIME_RATE	0.048725	0.078419	0.621346	0.534657201
AGE	0.032771	0.013098	2.501997	0.012670437
INDUS	0.130551	0.063117	2.068392	0.03912086
NOX	-10.3212	3.894036	-2.65051	0.008293859
DISTANCE	0.261094	0.067947	3.842603	0.000137546
TAX	-0.0144	0.003905	-3.68774	0.000251247
PTRATIO	-1.07431	0.133602	-8.0411	6.58642E-15
AVG_ROOM	4.125409	0.442759	9.317505	3.89287E-19
LSTAT	-0.60349	0.053081	-11.3691	8.91071E-27

- The Adjusted R Square value of 0.6882 suggests a strong relationship between the independent variables and the average price in the model.
- The intercept value is 29.241.A positive intercept value suggests that even in the absence of any independent variable, there is a starting or baseline value for the dependent variable.
- Observing the P-values, we find that the only insignificant variable in the model is crime rate (P-value 0.535), while all other variables are significant**.**
- The Coefficient value of Avg_room, distance, Indus, Age are positively corelated with the avg_price, where the coefficient of Nox, Tax, PRATIO , LSTAT are negatively co related with the avg_price.

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. 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? 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? d) Write the regression equation from this model

Ans: We use Regression tool in the data analysis tool pack for correlation matrix

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

	Coefficients	*ndard Err*	*t Stat*	*P-value*
Intercept	29.42847	4.804729	6.124898	1.85E-09
AGE	0.032935	0.013087	2.516606	0.012163
INDUS	0.13071	0.063078	2.072202	0.038762
NOX	-10.2727	3.890849	-2.64022	0.008546
DISTANCE	0.261506	0.067902	3.851242	0.000133
TAX	-0.01445	0.003902	-3.70395	0.000236
PTRATIO	-1.0717	0.133454	-8.03053	7.08E-15
AVG_ROOM	4.125469	0.442485	9.3234	3.69E-19
LSTAT	-0.60516	0.05298	-11.4224	5.42E-27

a) This is regression model is only taken the significant variable of the model, the multiple R value is 0.8328 , which indicates the strong relationship with the avg price.

b) By comparing the R square value (0.6886) of this model with the previous one (0.6882) we can interpret that this model has higher R square value than the previous, thus this is a better model for our prediction.

c) The positive coefficient value in ascending order Age(0.0329) , Indus(0.130), Distance(0.26), Avg room(4.125) and the negative coefficient value in ascending order tax(-0.0144), LSTAT(-0.60516), PTRATIO(-1.0717) and NOX(-10.27). If NOX value increase the avg price decrease by 10.27.

d) **Y=(0.0329)*(AGE)+(0.13071)*(INDUS)+(-10.2727)*(NOX)+(0.261)*(DISTANCE)+(-0.0144)*(TAX)+(-1.0717)*(PTRATIO)+(4.125)*(AVG_ROOM)+(-0.605)*(LSTAT) + 29.42**