Project Terro's Real Estate Agency

Q1. The first step to any project is understanding the data. So, for this step, generate the summary statistics for each of the variables. What do you observe?

A1.

- Mean of crime rate is 4.87 that show average crime rate in Boston, as skewness is positive and b/w -0.5 to 0.5 indicates that the distribution is fairly symmetrical. 50% of the households have crime rate above 4.82

CRIME_RATE	
Mean	4.871976
Standard Error	0.12986
Median	4.82
Mode	3.43
Standard Deviation	2.921132
Sample Variance	8.533012
Kurtosis	-1.18912
Skewness	0.021728
Range	9.95
Minimum	0.04
Maximum	9.99
Sum	2465.22
Count	506

- 68% of houses in Boston were built prior to 1940, the number of households relatively younger is real less as the data is negatively skewed.

AGE	
Mean	68.5749
Standard Error	1.25137
Median	77.5
Mode	100
Standard Deviation	28.14886
Sample Variance	792.3584
Kurtosis	-0.96772
Skewness	-0.59896
Range	97.1
Minimum	2.9
Maximum	100
Sum	34698.9
Count	506

- Non-retail business acres in Boston are 11%. Skewness is Positive that most of that houses have land for non-retail business

INDUS	
Mean	11.13678
Standard Error	0.30498
Median	9.69
Mode	18.1
Standard Deviation	6.860353
Sample Variance	47.06444
Kurtosis	-1.23354
Skewness	0.295022
Range	27.28
Minimum	0.46
Maximum	27.74
Sum	5635.21
Count	506

- Skewness is positive that indicates that most of the houses have No concentration. Negative kurtosis suggests a normal curve with no real peaks. Mean of NOX is around 0.55

NOX	
Mean	0.554695
Standard Error	0.005151
Median	0.538
Mode	0.538
Standard Deviation	0.115878
Sample Variance	0.013428
Kurtosis	-0.06467
Skewness	0.729308
Range	0.486
Minimum	0.385
Maximum	0.871
Sum	280.6757
Count	506

- Most of the houses are 24miles away from highways, which is also the maximum distance. Negative Kurtosis indicates flatter curve and no/short tail. Average distance from Highway is around 9.55miles.

DISTANCE	
Mean	9.549407
Standard Error	0.387085
Median	5
Mode	24
Standard Deviation	8.707259
Sample Variance	75.81637
Kurtosis	-0.86723
Skewness	1.004815
Range	23
Minimum	1
Maximum	24
Sum	4832
Count	506

- Skewness is positive that indicates most of Tax Rate lies under mean (408.23), Highly negative Kurtosis suggests no tail, most of the Tax rate is close to 666

TAX	
Mean	408.2372
Standard Error	7.492389
Median	330
Mode	666
Standard Deviation	168.5371
Sample Variance	28404.76
Kurtosis	-1.14241
Skewness	0.669956
Range	524
Minimum	187
Maximum	711
Sum	206568
Count	506

- Mean of PT-Ratio is 18.5 and it is near to the maximum value (22), Negative Skewness Indicates most of house's PT-Ratio is over the mean

PTRATIO	
Mean	18.45553
Standard Error	0.096244
Median	19.05
Mode	20.2
Standard Deviation	2.164946
Sample Variance	4.686989
Kurtosis	-0.28509
Skewness	-0.80232
Range	9.4
Minimum	12.6
Maximum	22
Sum	9338.5
Count	506

- Average No. of Room per house is 6 it also shows 50% of houses have above 6 rooms, Positive Kurtosis gives sharp curve that most of houses is near mean

AVG_ROOM	
Mean	6.284634
Standard Error	0.031235
Median	6.2085
Mode	5.713
Standard Deviation	0.702617
Sample Variance	0.493671
Kurtosis	1.8915
Skewness	0.403612
Range	5.219
Minimum	3.561
Maximum	8.78
Sum	3180.025
Count	506

- Nearly 13% of population has lower status, Positive kurtosis indicates a shape curve, Positive Skewness indicates that most of the houses have lower status

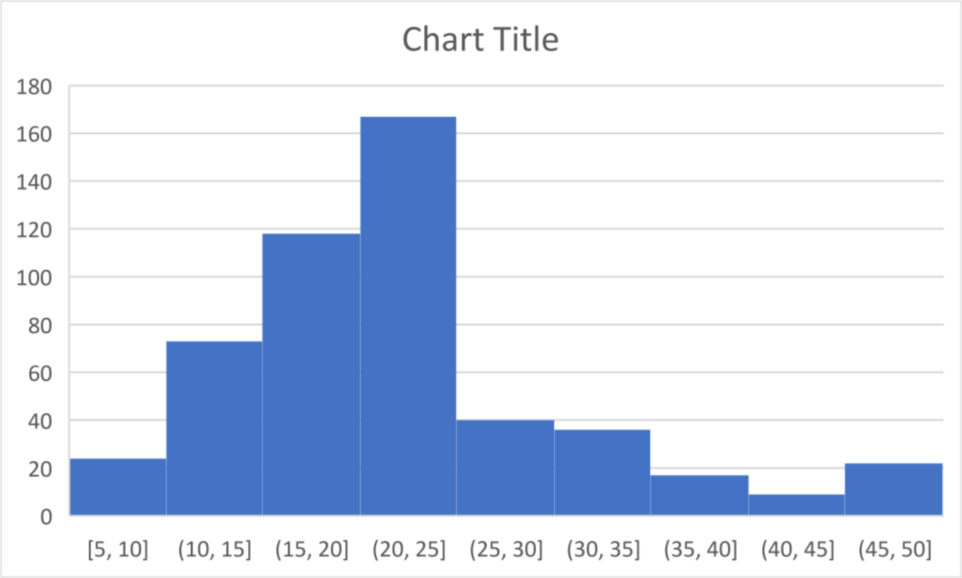
LSTAT	
Mean	12.65306
Standard Error	0.317459
Median	11.36
Mode	8.05
Standard Deviation	7.141062
Sample Variance	50.99476
Kurtosis	0.49324
Skewness	0.90646
Range	36.24
Minimum	1.73
Maximum	37.97
Sum	6402.45
Count	506

- Average value of houses is 22.5k,50% of houses are values more than 21.2k, Positive Kurtosis indicates High peaked curve, Positive skewness indicates maximum values lying under the mean

AVG_PRICE	
Mean	22.53281
Standard Error	0.408861
Median	21.2
Mode	50
Standard Deviation	9.197104
Sample Variance	84.58672
Kurtosis	1.495197
Skewness	1.108098
Range	45
Minimum	5
Maximum	50
Sum	11401.6
Count	506

Q2. Plot the histogram of the AvgPrice Variable. What do you infer?

A2. Avg price of the houses in Boston is b/w 20-25k. 50% of Family that lives in Boston have house value under 25k and rest have value above 25k.



Q3. Compute the covariance matrix. Share your observations.

A3. Covariance measures the direction of the relationship between two variables. It indicates the relationship of two variables whenever one variable changes. If an increase in one variable results in an increase in the other variable, both variables are said to have a positive covariance. Decreases in one variable also cause a decrease in the other.

- We can see highly positive covariance between Tax and Age in below table that both fields increase and decrease together that indicates they have very strong relation between them.
- We can see highly negative covariance between AVG-Price and Tax in below table that if one of the fields increase the other one decreases vice versa that indicates they have very weak relation between them.

	CRIME RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG ROOM	LSTAT	AVG PRICE
CRIME RATE	8.52									
AGE	0.56	790.79								
INDUS	-0.11	124.27	46.97							
NOX	0.00	2.38	0.61	0.01						
DISTANCE	-0.23	111.55	35.48	0.62	75.67					
TAX	-8.23	2397.94	831.71	13.02	1333.12	28348.62				
PTRATIO	0.07	15.91	5.68	0.05	8.74	167.82	4.68			
AVG ROOM	0.06	-4.74	-1.88	-0.02	-1.28	-34.52	-0.54	0.49		
LSTAT	-0.88	120.84	29.52	0.49	30.33	653.42	5.77	-3.07	50.89	
AVG PRICE	1.16	-97.40	-30.46	-0.45	-30.50	-724.82	-10.09	4.48	-48.35	84.42

Q4. Create a correlation matrix of all the variables. State top 3 positively correlated pairs and top 3 negatively correlated pairs.

A4.

	CRIME RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG ROOM*	*LSTAT*	*AVG PRICE*
CRIME RATE	1									
AGE	0.01	1								
INDUS	-0.01	0.64	1							
NOX	0.00	0.73	0.76	1						
DISTANCE	-0.01	0.46	0.60	0.61	1					
TAX	-0.02	0.51	0.72	0.67	0.91	1				
PTRATIO	0.01	0.26	0.38	0.19	0.46	0.46	1			
AVG_ROOM	0.03	-0.24	-0.39	-0.30	-0.21	-0.29	-0.36	1		
LSTAT	-0.04	0.60	0.60	0.59	0.49	0.54	0.37	-0.61	1	
AVG_PRICE	0.04	-0.38	-0.48	-0.43	-0.38	-0.47	-0.51	0.70	-0.74	1

Top 3 positively correlated pairs

- Distance by Tax (0.91)
- Indus by Nox (0.76)
- Age by Nox (0.73)

Top 3 negatively correlated pairs

- Lstat by Avg Price (-0.74)
- Avg Room by Lstat (-0.61)
- PT-Ratio by Avg Price (-0.51)

Q5. Build an initial regression model with AVG_PRICE as the y or the Dependent variable and LSTAT variable as the Independent Variable. Generate the residual plot too.

- **What do you infer from the Regression Summary Output in terms of variance explained, coefficient value, Intercept and the Residual plot?**
- **Is LSTAT variable significant for the analysis based on your model?**

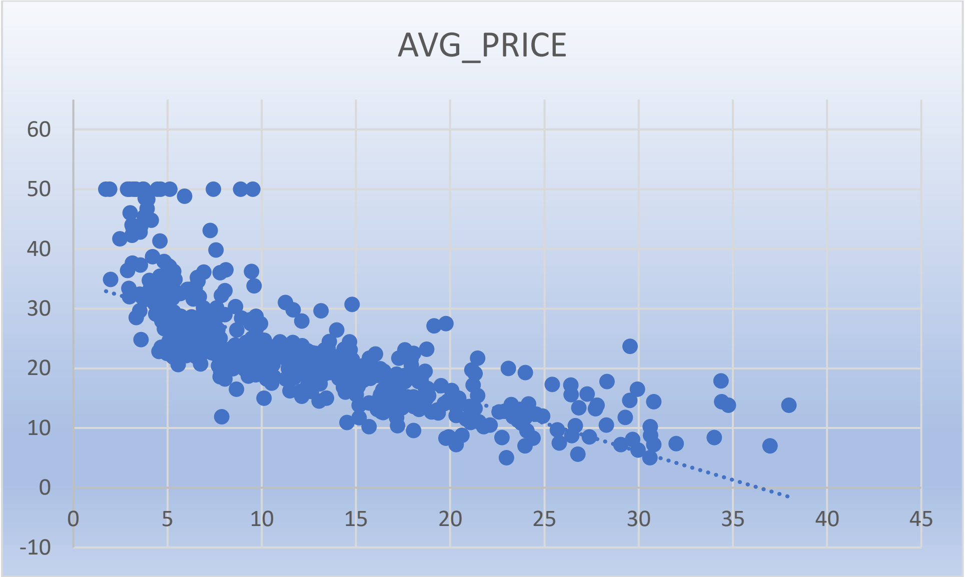
A5.

- ANOVA, which stands for Analysis of Variance, is a statistical test used to analyse the difference between the means of more than two groups. A one-way ANOVA uses one independent variable, while a two-way ANOVA uses two independent variables.
  - DF means the degrees of freedom in the source.
  - SS means the sum of squares due to the source.
  - MS means the mean sum of squares due to the source.
  - F means the F-statistic
  - Significance F is the probability that the null hypothesis in our regression model cannot be rejected.
- The intercept is simply the mean of the reference group, Managers. The coefficients for the other two groups are the differences in the mean between the reference group and the other groups. The intercept is the estimate of the dependent variable when all the independent variables are 0.
- A residual plot shows the difference between the observed response and the fitted response values. The ideal residual plot, called the null residual plot, shows a random scatter of points forming an approximately constant width band around the identity line. In below output we can see that the best fit line is inversely linear but this relation is quite close and approaches a linear relation.
- Yes, LSTAT is a significant variable for the analysis Lstat and Avg-Price are negatively correlated. The Lstat coefficient is negative to negate the overall value of average price corresponding to the Lstat Variable. as it has a highly negative correlation with Avg Price. R-Square of this regression model is 54% (0.54) that indicates how significant this variable is

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

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	23243.91	23243.91	601.6179	5.08E-88
Residual	504	19472.38	38.63568		
Total	505	42716.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	34.55384	0.562627	61.41515	3.7E-236	33.44846	35.65922	33.44846	35.65922
LSTAT	-0.95005	0.038733	-24.5279	5.08E-88	-1.02615	-0.87395	-1.02615	-0.87395



Q6. Build another instance of the Regression model but this time including LSTAT and AVG_ROOM together as independent variables and AVG_PRICE as the dependent variable.

- **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?**
- **Is the performance of this model better than the previous model you built in Question 5? Compare in terms of adjusted R-square. Explain.**

A6.

- Regression Equation
  Y=m1x1+m2x2+b
  Y=Avg-Price
  M1=5.094
  X1=Avg-Room (Given 7)
  M2= (-0.642)
  X2=Lstat (Given 20)
  B= (-1.358)

  Y=5.094*7+(-0.642) *20+(-1.358)
  Y=21.46

  Avg-Price is 21460$ and company quoting a value of 30000$ that clearly shows company is overcharging this household
- Yes, this model is better than the previous one made in above question(Q5), in pervious question we only choose one variable (LSTAT), where the R-Square is 0.54, where here we choose two variables (LSTAT and AVG_ROOM), where R-Square is 0.64 which is a better accurate model Previous One.
  R-Square is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model

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

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	27276.99	13638.49	444.3309	7E-112
Residual	503	15439.31	30.69445		
Total	505	42716.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.35827	3.172828	-0.4281	0.668765	-7.5919	4.875355	-7.5919	4.875355
AVG_ROOM	5.094788	0.444466	11.46273	3.47E-27	4.22155	5.968026	4.22155	5.968026
LSTAT	-0.64236	0.043731	-14.6887	6.67E-41	-0.72828	-0.55644	-0.72828	-0.55644

Q7. Now, build a Regression model with all variables. AVG_PRICE shall be the Dependent Variable. Interpret the output in terms of adjusted R-square, coefficient and Intercept values, Significance of variables with respect to AVG_price. Explain.

A7.

- As we can see below R-Square (0.69) of this model is above 50% that indicates it is good model
- we need to see the coefficients of the independent variables used for creating this model. A positive coefficient indicates that if the values of the independent variable increases, the mean of the dependent variable also tend to increase vice versa. A negative coefficient indicates that if the values of the independent variables increase, the mean of the dependent variables tend to decreases.
- Variables Which have P-value less than 0.05 are significant variables those who have P-value greater than 0.05 are not significant variables.

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

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	29638.86	3293.207	124.9045	1.9E-121
Residual	496	13077.43	26.3658		
Total	505	42716.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.24132	4.817126	6.070283	2.54E-09	19.77683	38.7058	19.77683	38.7058
CRIME_RATE	0.048725	0.078419	0.621346	0.534657	-0.10535	0.202799	-0.10535	0.202799
AGE	0.032771	0.013098	2.501997	0.01267	0.007037	0.058505	0.007037	0.058505
INDUS	0.130551	0.063117	2.068392	0.039121	0.006541	0.254562	0.006541	0.254562
NOX	-10.3212	3.894036	-2.65051	0.008294	-17.972	-2.67034	-17.972	-2.67034
DISTANCE	0.261094	0.067947	3.842603	0.000138	0.127594	0.394593	0.127594	0.394593
TAX	-0.0144	0.003905	-3.68774	0.000251	-0.02207	-0.00673	-0.02207	-0.00673
PTRATIO	-1.07431	0.133602	-8.0411	6.59E-15	-1.3368	-0.81181	-1.3368	-0.81181
AVG_ROOM	4.125409	0.442759	9.317505	3.89E-19	3.255495	4.995324	3.255495	4.995324

LSTAT	-0.60349	0.053081	-11.3691	8.91E-27	-0.70778	-0.49919	-0.70778	-0.49919

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.

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.**

A8.

- Interpreting the output of the model we need to see the coefficients of the independent variables used for creating this model. A positive coefficient indicates that if the values of the independent variable increases, the mean of the dependent variable also tend to increase vice versa. A negative coefficient indicates that if the values of the independent variables increase, the mean of the dependent variables tend to decreases.
- By comparing R-Square value of this model with previous model we find that this model is better then previous one because R-Square of this model is higher than previous model R-Square. Error is Also less in this model compare to previous one
- As we can see NOX has negative coefficient, If the value of NOX is increases, then the Average Price will go downward

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

- Regression Equation
  Y=m1x1+m2x2+m3x3+m4x4+m5x5+m6x6+m7x7+m8x8+b
  Y=Avg-Price
  M1=0.032 M2= 0.130 M2= 0.130 M3= (-10.272) M4=0.261 M5= (-0.014) M6= (-1.071) M7=4.125 M8= (-0.605)

  X1=AGE X2=INDUS X3=NOX X4=DISTANCE X5=TAX X6=PT-RATIO X7=AVG-ROOM X8=LSTAT
  B= 29.428

  Y = 0.032 * AGE + 0.130 * INDUS + (-10.272) * NOX + 0.261 * DISTANCE + (-0.014) * TAX + (-1.071) * PT-RATIO + 4.125 * AVG-ROOM + (-0.605) * LSTAT + 29.428

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.381626
R Square	0.145639
Adjusted R Square	0.143943
Standard Error	8.509467
Observations	506

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	6221.141	6221.141	85.91428	5.47E-19
Residual	504	36495.15	72.41102		
Total	505	42716.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	26.38213	0.561757	46.96362	3.3E-186	25.27845	27.4858	25.27845	27.4858
DISTANCE	-0.4031	0.043489	-9.269	5.47E-19	-0.48854	-0.31765	-0.48854	-0.31765