2023-2024

Assignment – Terro's real estate agency Real estate data analysis – Exploratory data analysis, Linear Re-gression

EXCEL PROJECT

Implementation of Exploratory Data Analysis using various statistical/analytical tools in MS Excel like Summary statistics, Histogram, correlation table, Regression analysis (using Data analysis tool pack)

Table of Contents

PREPARED BY: ROHIT CHETTRI

Assignment – Terro's real estate agency Real estate data analysis – Exploratory data analysis, Linear Regression

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 (Task):

PREPARED BY: ROHIT CHETTRI

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

INFERENCES

For CRIME_RATE

- Mean (Average): The mean crime rate is approximately 4.87, which gives an idea of the central tendency of the data. On average, the crime rate in the dataset is around 4.87.
- Median (Middle Value): The median crime rate is 4.82, which is very close to the mean. This suggests that the data is approximately symmetrically distributed.
- Mode (Most Common Value): The mode is 3.43, indicating that this value occurs most frequently in the dataset. It's useful for identifying a prominent peak or cluster in the data.
- Standard Deviation: The standard deviation is approximately 2.92. It measures the spread or dispersion of the data. A higher standard deviation suggests that the data points are more spread out from the mean.
- Kurtosis: The negative kurtosis value of approximately -1.19 suggests that the data is platykurtic. This means the distribution has thinner tails and is less peaked than a normal distribution.
- Skewness: The skewness value of approximately 0.02 is close to zero, indicating that the data is approximately symmetrically distributed. A positive skewness would indicate a right-skewed distribution, and negative skewness would indicate a left-skewed distribution.

The data appears to be roughly normally distributed or symmetric, as indicated by the mean, median, and skewness values being close to each other.

For AGE

- Mean (Average): The mean age is approximately 68.57, indicating the central tendency of the ages in the dataset.
- Median (Middle Value): The median age is 77.5, which is significantly higher than the mean. This suggests that the data might have a skewed distribution with a longer tail on the left side (negatively skewed).
- Mode (Most Common Value): The mode value is 100, indicating that 100 is the most frequently occurring age in the dataset.

PREPARED BY: ROHIT CHETTRI

- Standard Deviation: The standard deviation is approximately 28.15, which is relatively high. This suggests a significant amount of variation in the ages within the dataset.
- Kurtosis: The negative kurtosis value of approximately -0.97 suggests that the data is platykurtic. This means the distribution has thinner tails and is less peaked than a normal distribution.
- Skewness: The negative skewness value of approximately -0.60 indicates that the data is negatively skewed, meaning that the distribution is skewed to the left.

The data appears to be negatively skewed, with a longer tail on the left side of the distribution. There is a significant amount of variability in the ages, as indicated by the relatively high standard deviation and variance.

For INDUS

- Mean (Average): The mean proportion of non-retail business acres is approximately 11.14%, which represents the central tendency of this variable in the dataset.
- Median (Middle Value): The median value is 9.69%, which is lower than the mean. This suggests that the data might be positively skewed.
- Mode (Most Common Value): The mode is 18.1%, indicating that this value occurs most frequently in the dataset.
- Standard Deviation: The standard deviation is approximately 6.86, which is relatively high. This suggests a significant amount of variation in the proportion of non-retail business acres within the dataset.
- Kurtosis: The negative kurtosis value of approximately -1.23 suggests that the data is platykurtic, with thinner tails and being less peaked than a normal distribution.
- Skewness: The positive skewness value of approximately 0.30 indicates that the data is positively skewed, meaning that the distribution is skewed to the right.

The data appears to be positively skewed, with a longer tail on the right side of the distribution. There is a significant amount of variability in the proportion of non-retail business acres, as indicated by the relatively high standard deviation and variance

For NOX

- Mean (Average): The mean nitric oxides concentration is approximately 0.555 parts per 10 million, representing the central tendency of this variable in the dataset.
- Median (Middle Value): The median value is 0.538 parts per 10 million, which is close to the mean. This suggests that the data is approximately symmetrically distributed.

- Mode (Most Common Value): The mode is 0.538 parts per 10 million, indicating that this value occurs most frequently in the dataset.
- Standard Deviation: The standard deviation is approximately 0.116, which is relatively small. This suggests a limited amount of variation in nitric oxides concentration within the dataset.
- Kurtosis: The negative kurtosis value of approximately -0.065 suggests that the data is platykurtic, with thinner tails and being less peaked than a normal distribution.
- Skewness: The positive skewness value of approximately 0.729 indicates that the data is positively skewed, meaning that the distribution is skewed to the right.

The data appears to be positively skewed, with a longer tail on the right side of the distribution .There is relatively low variability in nitric oxides concentration, as indicated by the small standard deviation and variance.

For DISTANCE

- Mean (Average): The mean distance from the highway is approximately 9.55 miles, representing the central tendency of this variable in the dataset.
- Median (Middle Value): The median distance is 8.71 miles, which is slightly lower than the mean. This suggests that the data might have a positively skewed distribution.
- Mode (Most Common Value): The mode is 5 miles, indicating that this distance occurs most frequently in the dataset.
- Standard Deviation: The standard deviation is approximately 1.00, indicating a moderate amount of variation in the distances from the highway within the dataset.
- Kurtosis: The negative kurtosis value of approximately -0.867 suggests that the data is platykurtic, with thinner tails and being less peaked than a normal distribution.
- Skewness: The positive skewness value of approximately 1.005 indicates that the data is positively skewed, meaning that the distribution is skewed to the right.

The data appears to be positively skewed, with a longer tail on the right side of the distribution. There is moderate variability in the distances from the highway, as indicated by the standard deviation and variance.

For TAX

- Mean (Average): The mean property tax rate is approximately 408.24 per $10,000, representing the central tendency of this variable in the dataset.

- Median (Middle Value): The median property tax rate is 168.54 per $10,000, which is significantly lower than the mean. This suggests that the data might be positively skewed.
- Mode (Most Common Value): The mode is 666 per $10,000, indicating that this value occurs most frequently in the dataset.
- Standard Deviation: The standard deviation is approximately 284.76, indicating a substantial amount of variation in property tax rates within the dataset.
- Kurtosis: The negative kurtosis value of approximately -1.142 suggests that the data is platykurtic, with thinner tails and being less peaked than a normal distribution.
- Skewness: The positive skewness value of approximately 0.670 indicates that the data is positively skewed, meaning that the distribution is skewed to the right.

The data appears to be positively skewed, with a longer tail on the right side of the distribution. There is a substantial amount of variability in property tax rates, as indicated by the large standard deviation and variance.

For PTRATIO

- Mean (Average): The mean pupil-teacher ratio is approximately 18.46, representing the central tendency of this variable in the dataset.
- Median (Middle Value): The median pupil-teacher ratio is 19.05, which is slightly higher than the mean. This suggests that the data might have a negatively skewed distribution.
- Mode (Most Common Value): The mode is 20.2, indicating that this value occurs most frequently in the dataset.
- Standard Deviation: The standard deviation is approximately 2.165, indicating a moderate amount of variation in pupil-teacher ratios within the dataset.
- Kurtosis: The negative kurtosis value of approximately -0.285 suggests that the data is platykurtic, with thinner tails and being less peaked than a normal distribution.
- Skewness: The negative skewness value of approximately -0.802 indicates that the data is negatively skewed, meaning that the distribution is skewed to the left.

The data appears to be negatively skewed, with a longer tail on the left side of the distribution. There is a moderate amount of variability in pupil-teacher ratios, as indicated by the standard deviation and variance.

For AVG_ROOM

- Mean (Average): The mean average number of rooms per house is approximately 6.28, representing the central tendency of this variable in the dataset.
- Median (Middle Value): The median number of rooms is 6.21, which is close to the mean. This suggests that the data is approximately symmetrically distributed.
- Mode (Most Common Value): The mode is 5.713, indicating that this value occurs most frequently in the dataset.
- Standard Deviation: The standard deviation is approximately 0.703, indicating a moderate amount of variation in the number of rooms within houses in the dataset.
- Kurtosis: The positive kurtosis value of approximately 1.892 suggests that the data is leptokurtic, with fatter tails and a more peaked distribution compared to a normal distribution.
- Skewness: The positive skewness value of approximately 0.404 indicates that the data is positively skewed, meaning that the distribution is skewed to the right.

The data appears to be positively skewed, with a longer tail on the right side of the distribution. There is a moderate amount of variability in the number of rooms within houses, as indicated by the standard deviation and variance.

For AVG_PRICE

- Mean (Average): The mean average house price is approximately $22,532.81 (in $1000's), representing the central tendency of this variable in the dataset.
- Median (Middle Value): The median house price is $21.2 (in $1000's), which is lower than the mean. This suggests that the data might have a positively skewed distribution.
- Mode (Most Common Value): The mode is not provided, but it is not always easy to determine from summary statistics.
- Standard Deviation: The standard deviation is approximately $9.197 (in $1000's), indicating a moderate amount of variation in house prices within the dataset.
- Kurtosis: The positive kurtosis value of approximately 1.495 suggests that the data is somewhat leptokurtic, with fatter tails and a more peaked distribution compared to a normal distribution.
- Skewness: The positive skewness value of approximately 1.108 indicates that the data is positively skewed, meaning that the distribution is skewed to the right.

The data appears to be positively skewed, with a longer tail on the right side of the distribution. There is a moderate amount of variability in house prices, as indicated by the standard deviation and variance.

For AVG_PRICE

- Mean (Average): The mean average house price is approximately $22,532.81 (in $1000's), representing the central tendency of this variable in the dataset.
- Median (Middle Value): The median house price is $21.2 (in $1000's), which is lower than the mean. This suggests that the data might have a positively skewed distribution.
- Mode (Most Common Value): The mode is not provided, but it is not always easy to determine from summary statistics.
- Standard Deviation: The standard deviation is approximately $9.197 (in $1000's), indicating a moderate amount of variation in house prices within the dataset.
- Kurtosis: The positive kurtosis value of approximately 1.495 suggests that the data is somewhat leptokurtic, with fatter tails and a more peaked distribution compared to a normal distribution.
- Skewness: The positive skewness value of approximately 1.108 indicates that the data is positively skewed, meaning that the distribution is skewed to the right.

The data appears to be positively skewed, with a longer tail on the right side of the distribution. There is a moderate amount of variability in house prices, as indicated by the standard deviation and variance.

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

Inferences

- The Modal price (most common house price) lies in the range of (20,25] in $1000's.
- On an average the maximum value of houses is less than 25 thousand dollar's.
- The data appears to be positively skewed, with a longer tail on the right side of the distribution. (not normally distributed)
- The spread or variability in house prices is between 5 to 50 thousand dollar's.

3) Compute the covariance matrix. Share your observations?

	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

OBSERVATION

- Crime-rate: A positive covariance indicates a direct relationship with (AGE,XOX,PTRATIO,AVG_ROOM,AVG_PRICE) , while a negative covariance indicates an inverse relationship with (INDUS ,DISTANCE ,TAX and LSTAT) as observed .
- AGE :A positive covariance indicates a direct relationship with (CRIME_RATE ,INDUS ,NOX ,DISTANCE, TAX, PTRATIO,LSTAT) , while a negative covariance indicates an inverse relationship (AVG_ROOM,AVG_PRICE) as observed .
- INDUS:A positive covariance indicates a direct relationship with (NOX ,DISTANCE ,TAX, PTRATIO, LSTAT ,and AGE), while a negative covariance indicates an inverse relationship with (AVG_ROOM ,AVG_PRICE ,CRIME_RATE) as observed .
- NOX : A positive covariance indicates a direct relationship with (DISTANCE , TAX , PTRATIO , LSTAT , CRIME_RATE , AGE, and INDUS), while a negative covariance indicates an inverse relationship with (AVG_ROOM, AVG_PRICE) as observed .
- DISTANCE : A positive covariance indicates a direct relationship with (AGE, INDUS, NOX, TAX ,PTRATIO ,and LSTAT), while a negative covariance indicates an inverse relationship with (AVG_ROOM, AVG_PRICE) as observed .
- TAX : A positive covariance indicates a direct relationship with (AGE , INDUS , NOX , DISTANCE , PTRATIO , and LSTAT), while a negative covariance indicates an inverse relationship with (AVG_ROOM ,AVG_PRICE) as observed .
- For PTRATIO : A positive covariance indicates a direct relationship with (CRIME_RATE, AGE , INDUS , NOX , DISTANCE , TAX , and LSTAT), while a negative covariance indicates an inverse relationship with (AVG_ROOM, AVG_PRICE) as observed .
- For AVG_ROOM : A positive covariance indicates a direct relationship with (CRIME_RATE ,AVG_PRICE) , while a negative covariance indicates an inverse relationship with (AGE ,INDUS, NOX , DISTANCE, TAX , PTRATIO , and LSTAT) as observed .
- For LSTAT : A positive covariance indicates a direct relationship with (AGE ,INDUS ,NOX ,DISTANCE, TAX , and PTRATIO), while a negative covariance indicates an inverse relationship with (AVG_ROOM, AVG_PRICE, CRIME_RATE).
- For AVG_PRICE : A positive covariance indicates a direct relationship with (CRIME_RATE , AVG_ROOM), while a negative covariance indicates an inverse relationship with (AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, and LSTAT).

A positive covariance suggests that the two variables tend to increase or decrease together, while a negative covariance suggests an inverse relationship where one tends to increase as the other decreases.

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

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

OBSERVATION

Finding the Correlation between the data.

(r > 0): When the correlation coefficient is positive (closer to 1), it indicates a positive linear relationship. This means that as one variable increases, the other tends to increase as well.(r < 0): When the correlation coefficient is negative (closer to -1), it indicates a negative linear relationship. This means that as one variable increases, the other tends to decrease.(r = 0): When the correlation coefficient is zero, it suggests no linear relationship between the variables.

a) The top 3 positively correlated pairs are.

1. *DISTANCE* -TAX : with (0.910228189)
2. *INDUS* -NOX : with (0.763651447)

3. *AGE* -NOX : with (0.731470104)

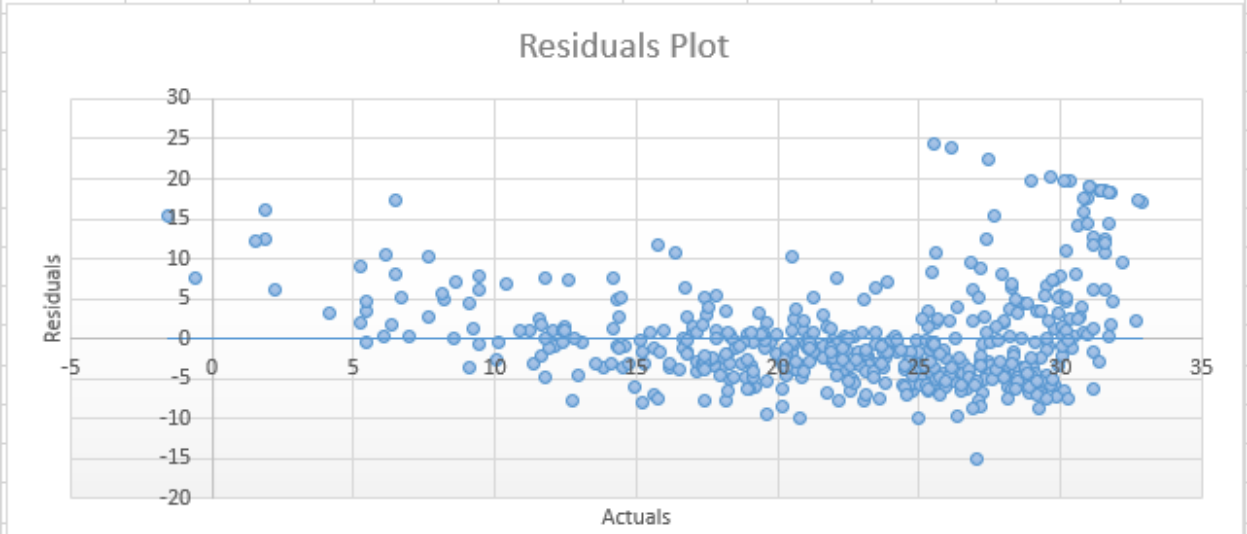
b) The top 3 negatively correlated pairs are.

1. *LSTAT* -AVG_PRICE : with (-0.737662726)
2. *AVG_ROOM* -LSTAT: with (-0.613808272)
3. *PTRATIO*-AVG_PRICE : with (-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.

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?**



INFERENCES

Variance Explained:

- R-squared (R^2): This is a measure of how well the independent variables explain the variation in the dependent variable. It ranges from 0 to 1 .Here R-squared of 0.544 means that 54% of the variance in the dependent variable (AVG_PRICE) is explained by the independent variables (LSTAT).

Coefficient Values:

- Coefficients: a coefficient of -0.950049354, it indicates that as the independent variable increases(LSTAT), the dependent variable is expected to decrease (AVG_PRICE). There is a negative linear relationship between the two variables. A coefficient of approximately -0.95 suggests a relatively *strong negative relationship* between the independent and dependent variables, *assuming all other factors remain constant.*

Intercept:

- The estimated value of the dependent(AVG_PRICE) variable when the independent variable is zero(LSTAT).It suggests that, when there are no lower-status residents in the area (LSTAT is zero), the average house price is estimated to be $34,553.84 (in $1000's).

Residual Plot:

- The trendline suggests that the spread of the data is minimum , indicating that we are violating homogeneity of variance assumption.

It appears that the LSTAT variable is significant for explaining house prices, as it has a strong negative relationship with house prices, and it contributes to explaining a substantial portion of the variance in the data. Considering the statistical significance of the LSTAT coefficient (p-value) and address the issue of violating the homogeneity of variance assumption for a more comprehensive analysis.

Model Evaluation

Model Fit.

R Square	0.544146298

- The rule suggests that whenever we build the Regression model the R square should be greater than 0.6. Here in our case the R Square vale is 0.544 which is less than 0.6 thus we can say that although the model captures a significant proportion of the variation in the dependent variable, it might not be a perfect fit.

Prediction Accuracy

					#RMSE					
Observation	Predicted AVG_PRICE	Residuals		Residuals^2	Mean	Root				
1	29.8225951	-5.822595098		33.90261367	38.48297	6.203464	Average error in the model			
2	25.87038979	-4.270389786		18.23622892						
3	30.72514198	3.974858016		15.79949625		Minimum error in the model		0		
4	31.76069578	1.639304221		2.687318328		Maximum error in the model		22.53281		
5	29.49007782	6.709922176		45.02305561		% of error in the model		0.275308	28%	RMSE

RMSQ(Root Mean Square Error) .

- *Here 28% indicated the error percentage and these percentages help us in finding the Prediction Accuracy based on the defined Business Model*
- The 28% RMSE error suggests that the model's predictions, on average, deviate by 28% from the actual observed values

	#MAPE								
Observation	Predicted AVG_PRICE	Residuals	AVG_PRICE	Pe(Residuals/Actuals)	Absolute(Pe)	Mean			
1	29.8225951	-5.822595098	24	-0.242608129	0.242608129	0.213521	21%	MAPE	
2	25.87038979	-4.270389786	21.6	-0.197703231	0.197703231				

MAPE(Mean Absolute Percentage Error)

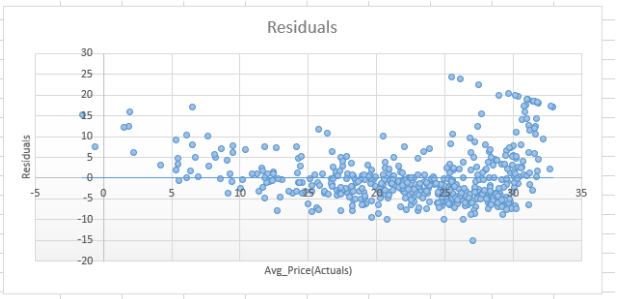
- A MAPE of 21% suggests that, on average, the model's predictions deviate from the actual values by approximately 21%. This can give us an idea of the typical size of errors in our model's predictions.
- A 21% MAPE indicates that, on average, the model's predictions have an error of around 21% concerning the actual values

Assumption Check

#Assumption Check

Predicted AVG_PRICE	Residuals	AVG_PRICE
29.8225951	-5.8226	24
25.87038979	-4.27039	21.6
30.72514198	3.974858	34.7
31.76069578	1.639304	33.4
29.49007782	6.709922	36.2
29.60408375	-0.90408	28.7
22.74472741	0.155273	22.9
16.36039575	10.7396	27.1
6.118863721	10.38114	16.5
18.30799693	0.592003	18.9
15.1253316	-0.12533	15
21.94668596	-3.04669	18.9
19.62856553	2.071434	21.7

Assumption Check	Predicted	Rresiduals
mean	22.53281	-2.7365E-14
Skewness	-0.90377	1.452739072
Variance	45.93659	38.48296723



Interpretation

Mean of Residuals: -2.7365E-14 (essentially zero):

- The mean of the residuals being extremely close to zero suggests that, on average, the model's predictions are nearly unbiased. The sum of the residuals is very close to zero, indicating that, on the whole, the model's predictions are relatively accurate.

Skewness of Residuals (1.452739072):

- Skewness measures the symmetry of the distribution of residuals. A positive skewness value (1.45 in this case) suggests that the residuals are somewhat right-skewed. This means the distribution of residuals may have a tail on the right side, indicating that there might be some outliers or the presence of some larger positive residuals. It implies that the residuals might not be normally distributed.

Variance of Residuals (38.48296723):

- The variance of the residuals measures their spread or dispersion. A higher variance (38.48 in this case) means that the residuals are more spread out from the mean. This suggest greater variability in the errors around the regression line.

Interpretation regarding Assumption Check:

- Mean: A mean close to zero is in line with the assumption of the residuals having an average of zero, supporting the validity of the linear regression model.
- Skewness: The positive skewness indicate that the residuals are not normally distributed. This violates the assumption of normality. It could imply the presence of outliers or certain patterns not captured by the model.
- Variance: A high variance suggest heteroscedasticity (unequal variance) in the residuals, violating the assumption of homoscedasticity. This indicates that the variability of the errors might not be constant across all values of the predictors.

Residual Plot:

- The trendline suggests that the spread of the data is minimum , indicating that we are violating homogeneity of variance assumption.

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

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.

REGRESSION EQUATION

AVG_PRICE= β0 + β1 × LSTAT + β2 × AVG_ROOM

β0 = −1.358272812

Coefficient for LSTAT (β1) is -0.642358334.

Coefficient for AVG_ROOM (β2) is 5.094787984.

AVG_PRICE= β0 + β1 × LSTAT + β2 × AVG_ROOM
β0 = −1.358272812
Coefficient for LSTAT (β1) is -0.642358334.
Coefficient for AVG_ROOM (β2) is 5.094787984.

Intercept	-1.358272812
β1	-0.642358334
β2	5.094787984
LSTAT	20
AVG_ROOM	7
AVG_PRICE	21.45807639 *(in $1000's)*
	$21,458.08

a)

Comparing the predicted value from the model ($21.46 thousand) with the company's quoted value ($30,000), it seems that the company is overcharging for the house. The company's quote of $30,000 is significantly higher than the predicted price derived from the regression model.

The model's prediction suggests that the house's value should be around $21.46 thousand, whereas the company is quoting $30,000, which indicates a substantial difference. Hence, based on the model's estimation, the company is overcharging for the house in this locality.

b)

Model 1 (With both LSTAT & AVG_ROOM):

R-Squared (R^2): 0.638561606

This means approximately 63.86% of the variance in AVG_PRICE is explained by the independent variables (LSTAT and AVG_ROOM) in this model.

Model 2 (With only LSTAT):

R-Squared (R^2): 0.544146298

This model explains around 54.41% of the variance in AVG_PRICE using the LSTAT variable alone.

Comparing Adjusted R-squared:

The adjusted R-squared is higher in Model-1,indicating that the model accounts for a greater portion of the variance in the dependent variable while considering the number of predictors involved. So, in this case, the model with both LSTAT and AVG_ROOM has more predictors but still maintains a higher adjusted R-squared, is generally considered a better model as it explains more variance relative to the number of variables used in Model-2.

Thus we can say that the performance of this model better than the previous model we built in Question 5.

Model Evaluation

Model Fit

R Square	0.638561606

- The rule suggests that whenever we build the Regression model the R square should be greater than 0.6. Here in our case the R Square vale is 0.638 which is greater than 0.6 thus we can say that the model captures a significant proportion of the variation in the dependent variable, and it might be a perfect fit.

Prediction Accuracy

#Precaution Accuracy

#RMSE

bservatio	Predicted AVG_PRICE	Residuals	Residuals^2	Mean	Root				
1	28.94101368	-4.941013681	24.41361619	30.51247	5.523809	Average error in the model			
2	25.48420566	-3.884205661	15.08705361						
3	32.65907477	2.040925231	4.1653758			Minimum error in the model	0		
4	32.40652	0.99348	0.987002511			Maximum error in the model	22.53281		
5	31.63040699	4.569593009	20.88118027			% of error in the model	0.245145	25%	RMSE
6	28.05452701	0.645472994	0.416635386						

RMSQ(Root Mean Square Error) .

- *Here 25% indicated the error percentage and these percentages help us in finding the Prediction Accuracy based on the defined Business Model*

- The 25% RMSE error suggests that the model's predictions, on average, deviate by 25% from the actual observed values

#MAPE

Observation	Predicted AVG_	Residuals	AVG_PRICE	Pe(Residuals/Actuals)	Absolute(Pe)	Mean		
1	28.94101	-4.941013681	24	-0.20587557	0.20587557	0.207538	21%	MAPE
2	25.48421	-3.884205661	21.6	-0.179824336	0.179824336			

MAPE(Mean Absolute Percentage Error)

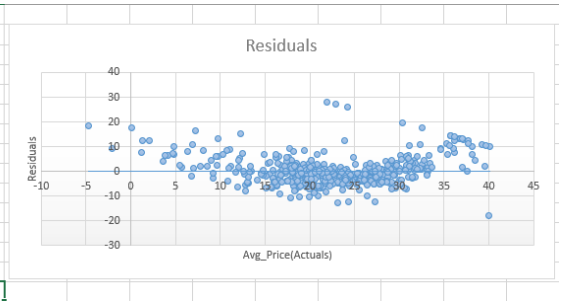
- A MAPE of 21% suggests that, on average, the model's predictions deviate from the actual values by approximately 21%. This can give us an idea of the typical size of errors in our model's predictions.
- A 21% MAPE indicates that, on average, the model's predictions have an error of around 21% concerning the actual values

Assumption Check

#Assumption Check

Predicted AVG_PRICE	Residuals	AVG_PRICE
28.94101368	-4.941013681	24
25.48420566	-3.884205661	21.6
32.65907477	2.040925231	34.7
32.40652	0.99348	33.4
31.63040699	4.569593009	36.2
28.05452701	0.645472994	28.7
21.28707846	1.612921545	22.9
17.78559653	9.314403473	27.1
8.104693384	8.395306616	16.5
18.24650673	0.653493269	18.9
17.99496223	-2.994962229	15
20.73221309	-1.832213091	18.9

Assumption Check	Predicted	Rresiduals
mean	22.53281	1.44741E-14
Skewness	-0.345117	1.34323094
Variance	53.90709	30.51246878



Interpretation

Mean of Residuals (1.44741E-14):

- The mean of the residuals is extremely close to zero (approximately 1.44741×10−141.44741×10−14). This near-zero mean indicates that, on average, the residuals are centered around zero, suggesting an unbiased model.

Skewness of Residuals (1.34323094):

- The skewness value of 1.34 suggests a degree of right-skewness in the distribution of the residuals. This indicates that the residuals are positively skewed, meaning there might be a tail on the right side of the distribution, which implies the presence of some larger positive residuals.

Variance of Residuals (30.51246878):

- The variance of 30.51 indicates the spread or dispersion of the residuals. A relatively moderate variance suggests that the residuals have a reasonable amount of variability around the mean.

Interpretation regarding Assumption Check:

- Mean: The near-zero mean aligns with the assumption of an unbiased model, supporting the validity of the linear regression model.
- Skewness: The positive skewness indicates that the residuals are not normally distributed. This deviation from normality may suggest potential outliers or patterns not captured by the model.
- Variance: The moderate variance implies a moderate amount of variability in the errors around the regression line.

The mean close to zero is consistent with an unbiased model, but the positive skewness and the moderate variance suggest deviations from the assumptions of normality and homoscedasticity. These issues may affect the model's performance and suggest the need for further investigation or potential adjustments to improve the model's reliability.

Residual Plot:

- The trendline suggests that the spread of the data is minimum , indicating that we are violating homogeneity of variance assumption.

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

Interpretation

Adjusted R-squared (R^2): 0.69385372

- This suggests that approximately 69.39% of the variance in AVG_PRICE is explained by the independent variables (CRIME_RATE, AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG_ROOM, and LSTAT) in this model. The adjusted R-squared accounts for the number of predictors in the model.

Intercept: 29.24131526

- This is the estimated value of AVG_PRICE when all the independent variables are zero. In this case, it suggests that if all the factors (CRIME_RATE, AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG_ROOM, LSTAT) are absent or have no effect, the average house price would be approximately $29,241 (in $1000's).

Significance based on P-values::

- Significant variables are those whose p-values are less than 0.05. If the p-value is greater than 0.05 then it is insignificant. Thus based on the p-value we can conclude:
- Significant-Variables- AGE INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG_ROOM,and LSTAT
- Insignificant-Variables-CRIME_RATE

Coefficients:

Each coefficient represents the change in AVG_PRICE for a one-unit change in the respective independent variable, assuming all other variables remain constant.

- *CRIME_RATE*: 0.0487 - For every one-unit increase in the crime rate, the AVG_PRICE is expected to increase by $48.7 (in $1000's).
- *AGE*: 0.0328 - For every one-unit increase in the proportion of units built prior to 1940, the AVG_PRICE is expected to increase by $32.8 (in $1000's).
- *INDUS*: 0.1306 - For every one-unit increase in the proportion of non-retail business acres per town, the AVG_PRICE is expected to increase by $130.6 (in $1000's).
- *NOX*: -10.3212 - For every one-unit increase in nitric oxides concentration, the AVG_PRICE is expected to decrease by $10,321.2 (in $1000's).
- *DISTANCE*: 0.2611 - For every one-unit increase in the weighted distances to five Boston employment centers, the AVG_PRICE is expected to increase by $261.1 (in $1000's).
- *TAX*: -0.0144 - For every one-unit increase in the full-value property-tax rate, the AVG_PRICE is expected to decrease by $14.4 (in $1000's).
- *PTRATIO*: -1.0743 - For every one-unit increase in pupil-teacher ratio by town, the AVG_PRICE is expected to decrease by $1,074.3 (in $1000's).
- *AVG_ROOM*: 4.1254 - For every one-unit increase in the average number of rooms per dwelling, the AVG_PRICE is expected to increase by $4,125.4 (in $1000's).
- *LSTAT*: -0.6035 - For every one-unit increase in the percentage of lower status of the population, the AVG_PRICE is expected to decrease by $603.5 (in $1000's).

Model Evaluation

Model Fit

R Square	0.69385372

- The rule suggests that whenever we build the Regression model the R square should be greater than 0.6. Here in our case the R Square vale is 0.693 which is greater than 0.6 thus we can say that the model captures a significant proportion of the variation in the dependent variable, and it might be a perfect fit.

Prediction Accuracy

#Precaution Accuracy

#RMSE

Observation	*Predicted AVG_PRICE*	*Residuals*	*Residuals^2*	*Mean*	*Root*				
1	30.1153558	-6.11536	37.39757659	25.84473	5.083772	Average error in the model			
2	27.00714024	-5.40714	29.23716562						
3	32.83291255	1.867087	3.486015563		Minimum error in the model		0		
4	31.20703392	2.192966	4.809100243		Maximum error in the model		22.53281		
5	30.5947288	5.605271	31.41906527		% of error in the model		0.225616	23%	**RMSE**
6	28.07644731	0.623553	0.388817954						

RMSQ(Root Mean Square Error) .

- *Here 23% indicated the error percentage and these percentages help us in finding the Prediction Accuracy based on the defined Business Model*
- The 23% RMSE error suggests that the model's predictions, on average, deviate by 23% from the actual observed values

#MAPE

bservatio	*Predicted AVG_PRICE*	*Residuals*	**AVG_PRICE**	*Pe(Residuals/Actuals)*	*Absolute(Pe)*	*Mean*		
1	30.1153558	-6.115355802	24	-0.254806492	0.254806492	0.184517	18%	**MAPE**
2	27.00714024	-5.407140244	21.6	-0.250330567	0.250330567			
3	32.83291255	1.867087455	34.7	0.053806555	0.053806555			

MAPE(Mean Absolute Percentage Error)

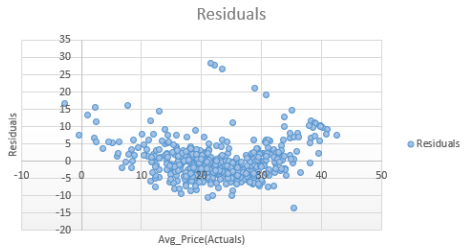
- A MAPE of 18% suggests that, on average, the model's predictions deviate from the actual values by approximately 18%. This can give us an idea of the typical size of errors in our model's predictions.
- A 18% MAPE indicates that, on average, the model's predictions have an error of around 18% concerning the actual values

Assumption Check

#Assumption Check

Observation	Predicted AVG_PRICE	Residuals	AVG_PRICE
1	30.1153558	-6.115355802	24
2	27.00714024	-5.407140244	21.6
3	32.83291255	1.867087455	34.7
4	31.20703392	2.192966083	33.4
5	30.5947288	5.605271204	36.2
6	28.07644731	0.623552688	28.7
7	25.29985158	-2.39985158	22.9
8	22.54671305	4.553286952	27.1
9	14.17584015	2.324159854	16.5
10	22.67662156	-3.776621559	18.9
11	22.78083379	-7.780833791	15

Assumption Check	Predicted	Rresiduals
mean	22.53281	-1.49586E-14
Skewness	-0.064369	1.644755549
Variance	58.57482	25.84473304



Interpretation

Mean of Residuals (-1.49586E-14):

- The mean of the residuals is extremely close to zero (approximately −1.49586×10−14−1.49586×10−14). This near-zero mean suggests that, on average, the residuals are centered around zero, indicating an unbiased model.

Skewness of Residuals (1.644755549):

- The skewness value of 1.64 indicates a considerable right-skew in the distribution of the residuals. A positive skewness implies a tail on the right side of the distribution, suggesting the presence of larger positive residuals.

Variance of Residuals (25.84473304):

- The variance of 25.84 indicates the spread or dispersion of the residuals. A moderate variance implies a moderate amount of variability in the errors around the mean.

Interpretation regarding Assumption Check:

- Mean: The near-zero mean aligns with the assumption of an unbiased model, supporting the validity of the linear regression model.
- Skewness: The positive skewness strongly suggests that the residuals are not normally distributed. This deviation from normality could indicate potential outliers or patterns not captured by the model.

- Variance: The moderate variance implies a reasonable amount of variability in the errors around the regression line.

Residual Plot:

- The trendline suggests that the spread of the data is minimum , indicating that we are violating homogeneity of variance assumption.

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.

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**.**

a)

Interpretation

Adjusted R-squared (R^2): 0.693615

- This suggests that approximately 69.39% of the variance in AVG_PRICE is explained by the independent variables (AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG_ROOM, and LSTAT) in this model. The adjusted R-squared accounts for the number of predictors in the model.

Intercept:

- 29.42847 This is the estimated value of AVG_PRICE when all the independent variables are zero. In this case, it suggests that if all the factors (AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG_ROOM, LSTAT) are absent or have no effect, the average house price would be approximately $29,428 (in $1000's).

Significance based on P-values::

- Signific ant variables are those whose p-values are less than 0.05

Coefficients:

- AGE: For every one-unit increase in the proportion of units related to age (AGE), the AVG_PRICE is expected to increase by $32.93 (in $1000's).
- INDUS: For every one-unit increase in the proportion of non-retail business acres per town (INDUS), the AVG_PRICE is expected to increase by $130.71 (in $1000's).
- NOX: For every one-unit increase in nitric oxides concentration (NOX), the AVG_PRICE is expected to decrease by $10,272.71 (in $1000's).
- DISTANCE: For every one-unit increase in the weighted distances to five Boston em-ployment centers (DISTANCE), the AVG_PRICE is expected to increase by $261.51 (in $1000's).
- TAX: For every one-unit increase in the full-value property-tax rate (TAX), the AVG_PRICE is expected to decrease by $14.45 (in $1000's).
- PTRATIO: For every one-unit increase in pupil-teacher ratio by town (PTRATIO), the AVG_PRICE is expected to decrease by $1,071.70 (in $1000's).
- AVG_ROOM: For every one-unit increase in the average number of rooms per dwelling (AVG_ROOM), the AVG_PRICE is expected to increase by $4,125.47 (in $1000's).
- LSTAT: For every one-unit increase in the percentage of lower status of the popula-tion (LSTAT), the AVG_PRICE is expected to decrease by $605.16 (in $1000's).

b)

Compare the adjusted R-square value

For the revised model with a different set of significant variables:

Revised Model:

- Adjusted R-squared (R^2): 0.693615426
- Significant variables used in this model: AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG_ROOM, LSTAT

Interpreting the output and compare it with the previous model:

Adjusted R-squared (R^2) Comparison:

- The adjusted R-squared value for the revised model is 0.693615426.

- In the previous model, the adjusted R-squared was 0.69385372.

Comparing Models:

- The adjusted R-squared in the revised model (0.693615426) is slightly lower than the adjusted R-squared in the original model (0.69385372).

Which model performs better according to the adjusted R-square value?

- The model with the higher adjusted R-squared value is considered to explain a greater proportion of the variance in the dependent variable relative to the number of predictors involved.
- In this case, although the difference is very small, the original model (R Square 0.69385372) performs slightly better in explaining the variance in AVG_PRICE compared to the model with only the significant variables (R Square 0.693615426).

It's essential to consider that while one model may have a slightly higher adjusted R-squared, it's also important to balance that with the inclusion of only significant variables to avoid overfitting or unnecessary complexity in the model.

c)

Negative coefficients (NOX)

Sorting the coefficients in ascending order:

- NOX (-10.27270508)
- LSTAT (-0.605159282)
- PTRATIO (-1.071702473)
- TAX (-0.014452345)
- AGE (0.03293496)
- INDUS (0.130710007)
- DISTANCE (0.261506423)
- AVG_ROOM (4.125468959)

If the value of NOX (nitric oxides concentration) is higher in a locality in this town, based on the provided coefficients, it would lead to a decrease in the average price of houses (AVG_PRICE). The model suggests that as NOX increases, the average house price decreases significantly.

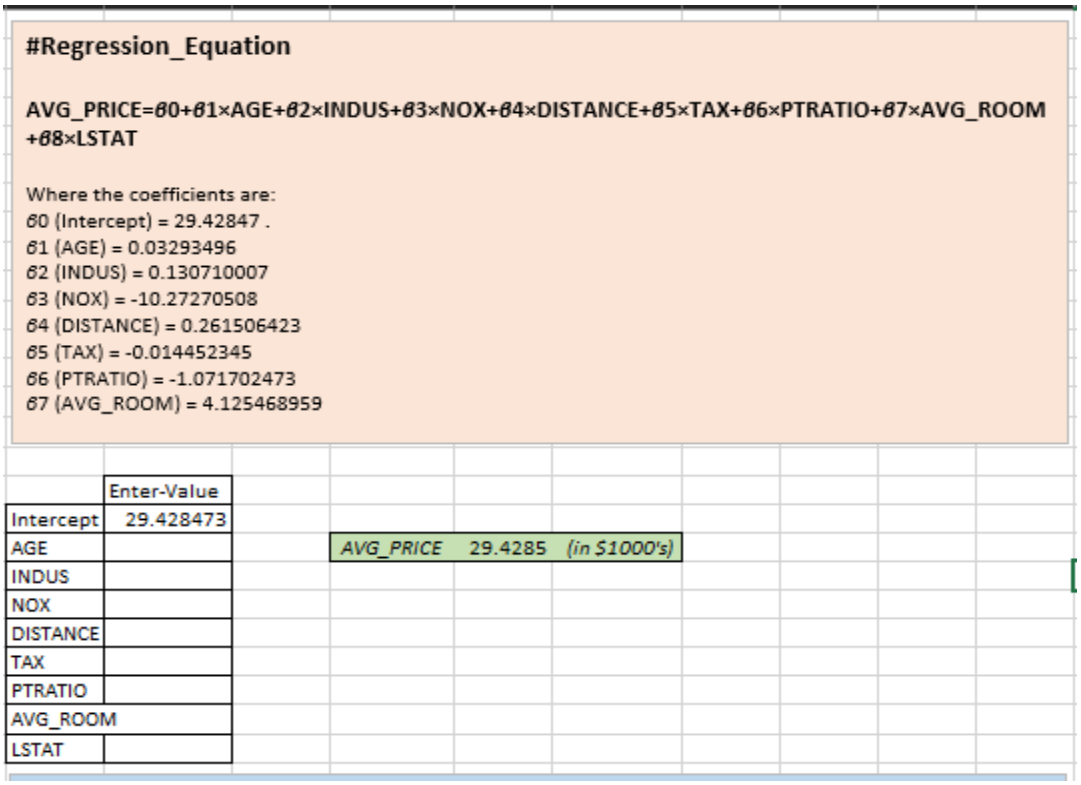
d)

Regression Equation

AVG_PRICE=β0+β1×AGE+β2×INDUS+β3×NOX+β4×DISTANCE+β5×TAX+β6×PTRATIO+β7×AVG_ROOM+β8×LSTAT

Where the coefficients are:

- β0 (Intercept) = 29.42847 .
- β1 (AGE) = 0.03293496
- β2 (INDUS) = 0.130710007
- β3 (NOX) = -10.27270508
- β4 (DISTANCE) = 0.261506423
- β5 (TAX) = -0.014452345
- β6 (PTRATIO) = -1.071702473
- β7 (AVG_ROOM) = 4.125468959
- β8 (LSTAT) = -0.605159282



Model Evaluation

Model Fit

R Square	0.693615426

- The rule suggests that whenever we build the Regression model the R square should be greater than 0.6. Here in our case the R Square vale is 0.693 which is greater than 0.6 thus we can say that the model captures a significant proportion of the variation in the dependent variable, and it might be a perfect fit.

Prediction Accuracy

	#Precaution Accuracy									
				#RMSE						
	RESIDUAL OUTPUT									
bservatic	Predicted AVG_PRICE	Residuals	Residuals^2	Mean	Root					
1	30.04888734	-6.048887337	36.58903801	25.8648498	5.08575	Average error in the model				
2	27.04098462	-5.440984617	29.60431361							
3	32.69896454	2.001035462	4.004142921			Minimum error in the model	0			
4	31.14306949	2.256930513	5.093735341			Maximum error in the model	22.53281	(average of y variable)		
5	30.58808735	5.611912655	31.49356364			% of error in the model	0.225704	23%	RMSE	
6	27.85095254	0.849047463	0.720881594							

RMSQ(Root Mean Square Error) .

- *Here 23% indicated the error percentage and these percentages help us in finding the Prediction Accuracy based on the defined Business Model*
- The 23% RMSE error suggests that the model's predictions, on average, deviate by 23% from the actual observed values

				#MAPE					
	RESIDUAL OUTPUT								
			Actuals						
bservatio	Predicted AVG_PRICE	Residuals	AVG_PRICE	Pe(Residuals/Actuals)	Absolute(Pe)	Mean			
1	30.04888734	-6.04889	24	-0.252036972	0.252036972	0.184787	18%	MAPE	
2	27.04098462	-5.44098	21.6	-0.251897436	0.251897436				

MAPE(Mean Absolute Percentage Error)

- A MAPE of 18% suggests that, on average, the model's predictions deviate from the actual values by approximately 18%. This can give us an idea of the typical size of errors in our model's predictions.
- A 18% MAPE indicates that, on average, the model's predictions have an error of around 18% concerning the actual values

Assumption Check

	#Assumption Check		
			Actuals
Observation	Predicted AVG_PRICE	Residuals	AVG_PRICE
1	30.04888734	-6.04889	24
2	27.04098462	-5.44098	21.6
3	32.69896454	2.001035	34.7
4	31.14306949	2.256931	33.4
5	30.58808735	5.611913	36.2
6	27.85095254	0.849047	28.7
7	25.07089688	-2.1709	22.9
8	22.63588287	4.464117	27.1
9	14.00883345	2.491167	16.5
10	22.84744402	-3.94744	18.9

Assumption Check	Predicted	Rresiduals
mean	22.53280632	-1.03948E-14
Skewness	-0.06656939	1.638992365
Variance	58.55470637	25.86484979

Interpretation

Mean of Residuals: -1.03948E-14 (essentially zero):

- The mean of the residuals being extremely close to zero suggests that, on average, the model's predictions are nearly unbiased. The sum of the residuals is very close to zero, indicating that, on the whole, the model's predictions are relatively accurate.

Skewness of Residuals: 1.638992365 (positive skew):

- A positive skewness of 1.64 suggests that the distribution of the residuals is skewed to the right (positively skewed). The positive skewness imply that there are more extreme positive residuals in the data, indicating potential issues or outliers with the model's predictions in overestimating the outcomes more than underestimating.

Variance of Residuals: 25.86484979 (larger variance):

- The variance of 25.86 is a measure of the spread or dispersion of the residuals around their mean. A larger variance indicates that the residuals are more widely spread out from the mean.
- A high variance suggest that the residuals have a wide range of values, signifying higher variability in the accuracy of the model's predictions.

The near-zero mean suggests the model's predictions are, on average, relatively accurate.
The positive skewness indicates that the residuals are not normally distributed, with a tendency for more extreme positive residuals.
The larger variance suggests wider variability in the model's prediction accuracy.

Residual Plot:

- The trendline suggests that the spread of the data is minimum , indicating that we are violating homogeneity of variance assumption.