Data Analysis Report

Analyzed by: Kanaga Manikandan Gopal
Type of Analysis: Multiple Linear Regression
Dataset: Annual sales data of a sample company

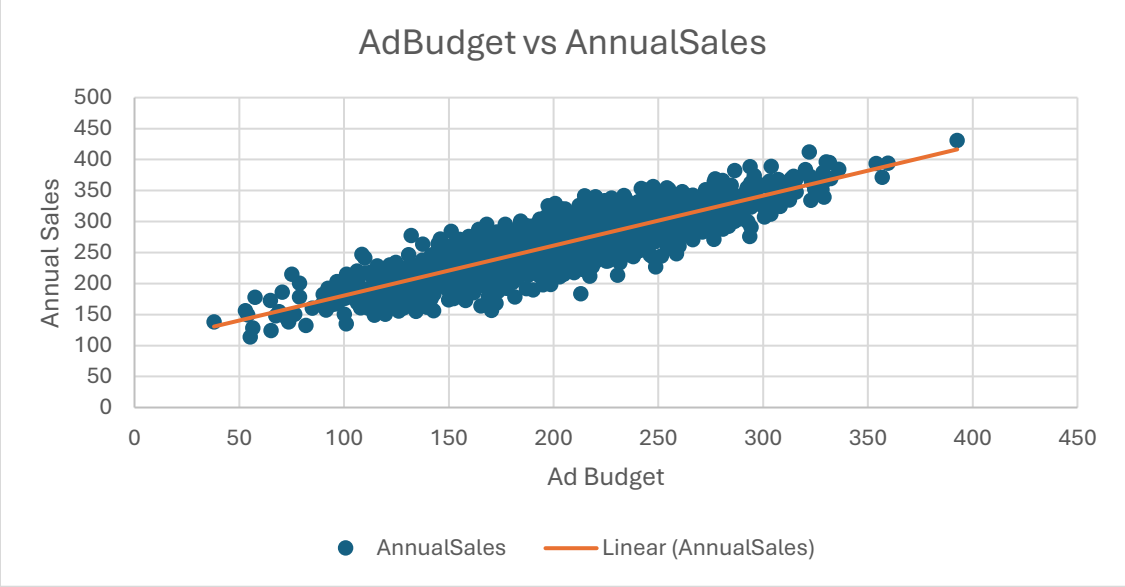
Overview of the Dataset:
The dataset is containing the data about the Annual Sales, Ad Budget, Price, Customer Visits, Competitor Price and Region.

Objective:
The goal is to explore the relationship between the response variable (Annual Sales) and the predictor variables (Ad Budget, Price, Customer Visits, Competitor Price and Region). With the Predicted outcomes and the assumptions for Linear Regression, the interpretations must be made.

Section 1: Data Visualization
1. Use appropriate plots to explore the relationships between the response variable and each of the predictors.
2. Interpretations on each graph.

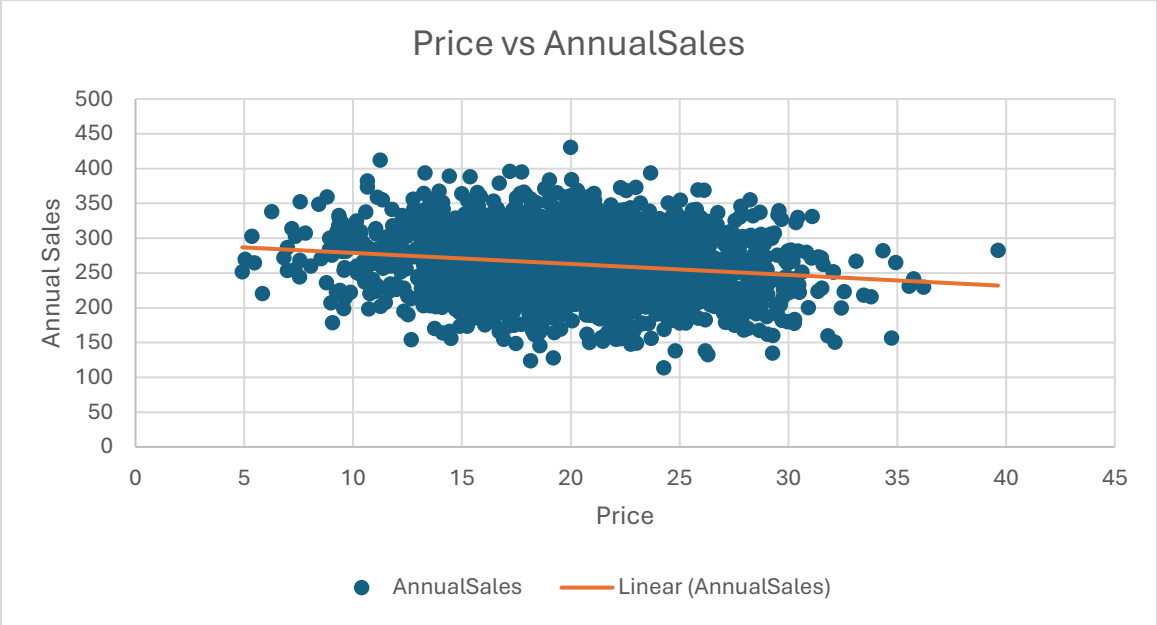
Charts & Interpretations:

1. **Ad Budget vs Annual Sales**



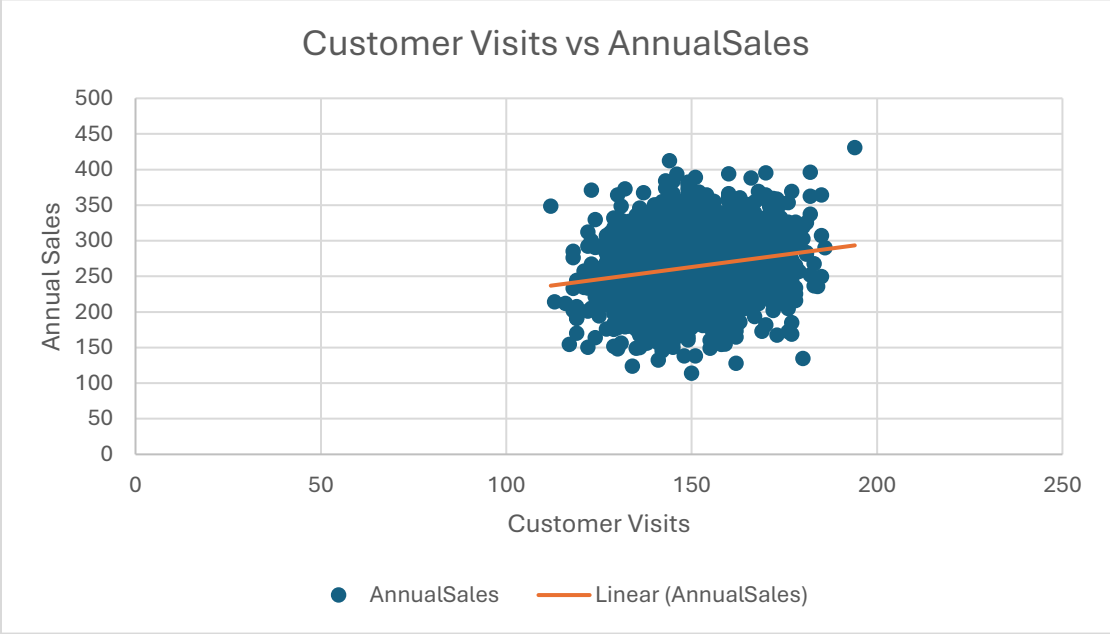
Interpretation:
Increase in Ad Budget, increase in Annual Sales.

2. **Price vs Annual Sales**



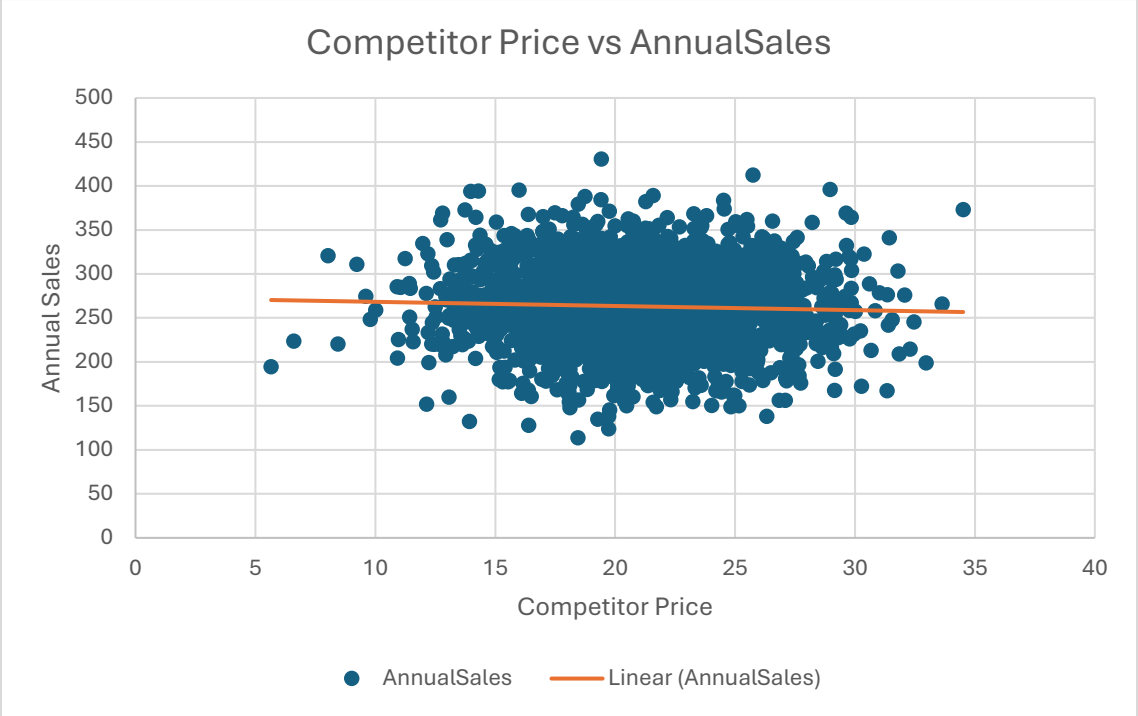
Interpretation:
Increase in Price, slight decrease in Annual Sales.

3. **Customer Visits vs Annual Sales**



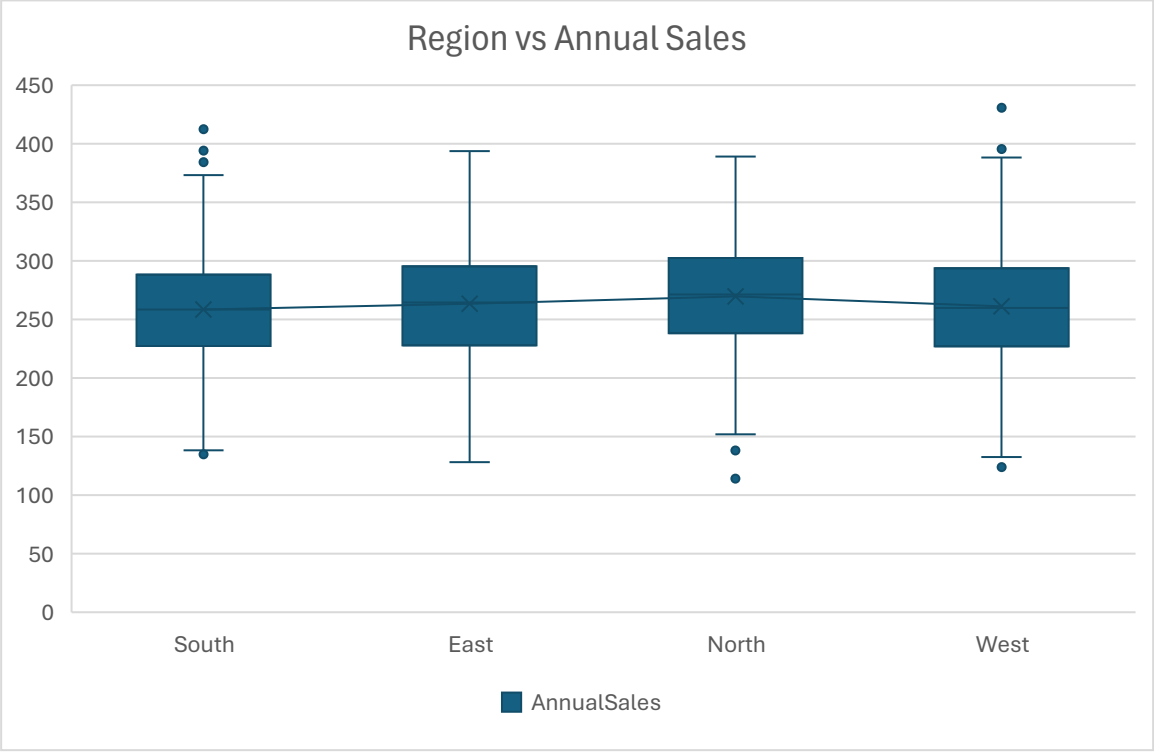
Interpretation:
Increase in Customer Visits, slight increase in Annual Sales.

4. **Competitor Price vs Annual Sales**



Interpretation:
Increase in Competitor Price, minor decrease in Annual Sales.

5. **Region vs Annual Sales**



Interpretations:

a. The median sales for all regions moving around 250, showing a relatively consistent central tendency.
b. The South region has significant outliers, with sales above 350 and dropping below 150. Similarly, the West region has extreme values above 400 and below 100.
c. The East and North regions show a narrower distribution, with fewer outliers and a more concentrated sales range.
d. The highest sales are achieved in the East and West regions, peaking close to 400.

Section 2: Estimation and Interpretation of Regression Models

1. Fit a multiple linear regression model with all five predictors adequately, performing the necessary transformations.
    a. Provide the estimated regression equation.
    b. Interpret all the coefficients of the predictor variables.
2. Discuss how the inclusion of the categorical variable affects the interpretation of the intercept.

Summary Output of the Whole Regression Model:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.902567343
R Square	0.814627809
Adjusted R Square	0.8139764
Standard Error	19.84024759
Observations	2000

Whole Regression Model

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	3445864.185	492266.3121	1250.564054	0
Residual	1992	784121.7656	393.6354245		
Total	1999	4229985.95			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	38.62915269	6.56689376	5.882408656	4.73103E-09	25.75045225	51.50785313	25.75045225	51.50785313
AdBudget	0.802006584	0.008980594	89.30440437	0	0.784394242	0.819618925	0.784394242	0.819618925
Price	-1.460722138	0.088445199	-16.51556169	1.58358E-57	-1.634176935	-1.287267342	-1.634176935	-1.287267342
CustomerVisits	0.656700992	0.036520981	17.98147195	4.1257E-67	0.585077666	0.728324318	0.585077666	0.728324318
CompetitorPrice	-0.389754271	0.110696781	-3.520917828	0.000439725	-0.606847883	-0.172660659	-0.606847883	-0.172660659
North	7.575518647	1.264965013	5.988717926	2.50278E-09	5.094725434	10.05631186	5.094725434	10.05631186
South	-3.514140139	1.262485864	-2.783508504	0.005428216	-5.990071356	-1.038208923	-5.990071356	-1.038208923
East	0.587346811	1.255107972	0.467965166	0.639860694	-1.874115212	3.048808833	-1.874115212	3.048808833

Estimated Regression Equation:

Y cap = 38.62915269 + 0.802006584*(AdBudget) +
(-1.460722138)*(Price) + 0.656700992*(Customer Visits) +
(-0.389754271)*(Competitor Price) +
7.575518647*(North) +
(-3.514140139)*(South) +
0.587346811*(East)

Interpretations of the Coefficients:

- **Intercept (38.63)** represents the expected value of the annual sales when all predictor variables are set to zero.
- **AdBudget (0.802)** represents a strong positive impact, meaning that increasing the advertising budget significantly boosts the annual sales.
- **Price (-1.461)** represents a negative impact, meaning that increasing the price leads to lower Annual Sales.
- **CustomerVisits (0.657)** shows a strong positive effect on annual sales which means more customer visits lead to higher sales.
- **CompetitorPrice (-0.390)** shows a negative but relatively small impact; higher competitor prices reduce the annual sales but not as strongly as other variables.
- **North (7.576)** represents that locations in the North region tend to perform better.
- **South (-3.514)** represents a Negative impact, showing worse performance in the South.
- **East (0.587, p-value 0.64)** represents an insignificant effect, meaning that the East region does not strongly impact the annual sales.

Section 3: Prediction of Annual Sales for a sample input

For a new predictor observed:

AdBudget = 200, Price = 20, CustomerVisits = 150, CompetitorPrice = 21, Region = "East"

Compute:

a. The point prediction for Annual Sales

b. A 95% confidence interval for the expected value of sales

c. A 95% prediction interval for a new observation

Summary Output of the Scenario Regression Model:

SUMMARY OUTPUT

Scenario Regression Model

Regression Statistics	
Multiple R	0.898372076
R Square	0.807072386
Adjusted R Square	0.806588616
Standard Error	20.23038192
Observations	2000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	3413904.855	682780.971	1668.296526	0
Residual	1994	816081.0953	409.2683527		
Total	1999	4229985.95			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	260.7298042	0.894989905	291.321503	0	258.9745908	262.4850176	258.9745908	262.4850176
AdBudget1	0.80225747	0.009156918	87.61216905	0	0.78429934	0.820215601	0.78429934	0.820215601
Price2	-1.436942686	0.090105447	-15.94734548	5.59012E-54	-1.613653378	-1.260231993	-1.613653378	-1.260231993
CustomerVisits	0.664560108	0.037211538	17.85898001	2.71901E-66	0.591582537	0.737537678	0.591582537	0.737537678
CompetitorPrice1	-0.430063636	0.112781039	-3.813261881	0.000141312	-0.651244667	-0.208882606	-0.651244667	-0.208882606
East1	-0.77032001	1.03755647	-0.742436708	0.457910233	-2.805128446	1.264488426	-2.805128446	1.264488426

Given Scenario	
AdBudget	200
Price	20
CustomerVisits	150
CompetitorPrice	21
Region (East)	1

Point Prediction of the Annual Sales:

- Y cap = 38.62915269 + 0.802006584*200 + (-1.460722138)*20 + 0.656700992*150 + (-0.389754271)*21 + 0.587346811*1 + 7.575518647*0 + (-3.514140139)*0
- Y cap = **260.7236825**

Confidence Value -> t (alpha/2)	1.961155595	T.INV.2T(0.05,2000-7-1), where degree_of_freedom = n-k-1
Standard Deviation from Scenario Regression Model (SD)	0.894989905	
Standard Error from Whole Regression Model (SE)	19.84024759	
Margin of Error	38.94938116	Conf. Value* SQRT(SD^2 + SE^2)

A 95% confidence interval for the expected value of sales:

Lower Confidence Interval	258.9684681	Point Prediction - Conf. Value * SD
Upper Confidence Interval	262.478897	Point Prediction + Conf. Value * SD

A 95% prediction interval for a new observation:

Lower Prediction Interval	221.7743014	Point Prediction - Conf. Value * SQRT(SD^2 + SE^2)
Upper Prediction Interval	299.6730637	Point Prediction + Conf. Value * SQRT(SD^2 + SE^2)

Section 4: Goodness-of-Fit

Report and interpret:

a. R-squared

b. Adjusted R-squared

c. RMSE or standard error of the regression

Regression Statistics of the Whole Regression Model

Regression Statistics	
Multiple R	0.902567343
R Square	0.814627809
Adjusted R Square	0.8139764
Standard Error	19.84024759
Observations	2000

Interpretations:

1. **R Square**
    a. There is 81.4 percent variation in the output which is explained by the input variables.
    b. The rest of the precent in variation is due to the underlying hidden factors and couldn't be explained by the model.

2. **Adjusted R Square**
    a. There is a small difference between R^2 (0.81463) and Adjusted R^2 (0.81398).
    b. Most of the predictor variables are contributing significantly to predict the response variable which means that most predictors are relevant.
    c. No strong signs of overfitting.

3. **Standard Error or RMSE**
    The actual value is about 19.84 units far from the model's predicted value.

Section 5: Tests of Significance

Conduct:

a. An overall F-test for model significance

b. t-tests for each predictor.

Interpret the results and state which predictors are statistically significant at the 5% level.

F-test data:

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	3445864.185	492266.3121	1250.564054	0
Residual	1992	784121.7656	393.6354245		
Total	1999	4229985.95			

Interpretations:

- This test is to check if at least one predictor variable is statistically significant to predict the dependent variable in the model.
- Since p value (Significance F) < 0.05, the expected condition becomes true.

t-test Data:

	Coefficients	**Standard Error**	**t Stat**	**P-value**	**Lower 95%**	**Upper 95%**	**Lower 95.0%**	**Upper 95.0%**
Intercept	38.62915269	6.56689376	5.882408656	4.73103E-09	25.75045225	51.50785313	25.75045225	51.50785313
AdBudget	0.802006584	0.008980594	89.30440437	0	0.784394242	0.819618925	0.784394242	0.819618925
Price	-1.460722138	0.088445199	-16.51556169	1.58358E-57	-1.634176935	-1.287267342	-1.634176935	-1.287267342
CustomerVisits	0.656700992	0.036520981	17.98147195	4.1257E-67	0.585077666	0.728324318	0.585077666	0.728324318

Competitor Price	-0.389754271	0.110696781	-3.520917828	0.000439725	-0.606847883	-0.172660659	-0.606847883	-0.172660659
North	7.575518647	1.264965013	5.988717926	2.50278E-09	5.094725434	10.05631186	5.094725434	10.05631186
South	-3.514140139	1.262485864	-2.783508504	0.005428216	-5.990071356	-1.038208923	-5.990071356	-1.038208923
East	0.587346811	1.255107972	0.467965166	0.639860694	-1.874115212	3.048808833	-1.874115212	3.048808833

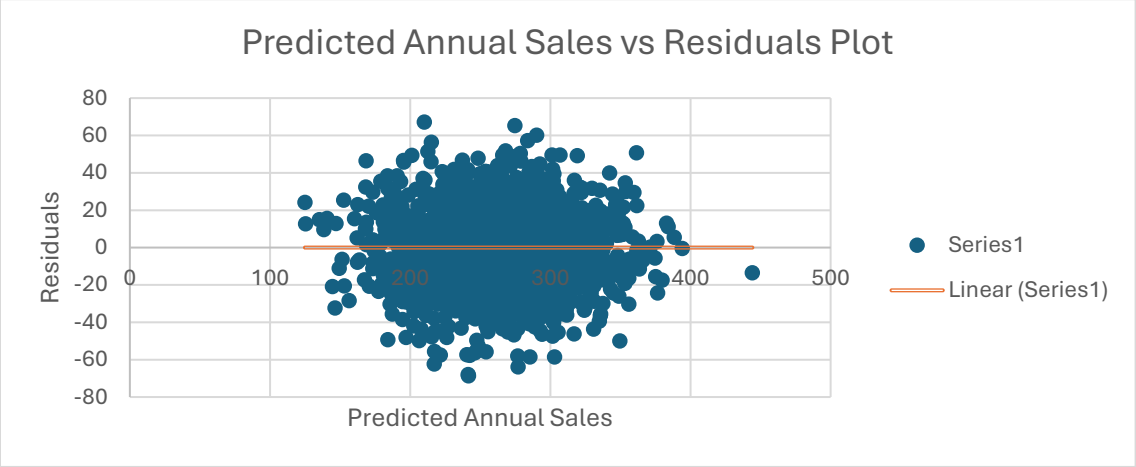
Interpretation:
This test is to find how each predictor is helpful for predicting the response variable.

Independent Variables	P value	Statistically Significant	Interpretations
Intercept	4.73103E-09	YES	When all X values = 0 (independent variables), the Annual Sales is 38.63
Ad Budget	0	YES	Sales increase significantly (very strong), when each unit in Ad Budget increases.
Price	1.58358E-57	YES	Higher the price, lower the sales (Strong Negative Relationship)
CustomerVisits	4.1257E-67	YES	More customer visits, more sales (highly significant), as p value < 0.05
CompetitorPrice	0.000439725	YES	Higher competitor prices slightly reduce our sales, but statistically significant.
North	2.50278E-09	YES	p < 0.05, the region dummy variable for North contributes meaningfully to the model.
South	0.005428216	YES	p < 0.05, the region dummy variable for South contributes meaningfully to the model.
East	0.639860694	NO	p > 0.05, the region dummy variable for East doesn't contribute meaningfully to the model.

Section 6: Assumption Diagnostics
Evaluate the assumptions of the OLS model:
• Linearity
• Constant variance (Homoscedasticity)
• Normality of residuals
• Independence of errors
• No multicollinearity
Briefly explain how violations of assumptions, if any, might affect your analysis.

Interpretation of Assumptions using Scatter Plots, Histogram and VIF

Predicted Annual Sales vs Residuals Plot:

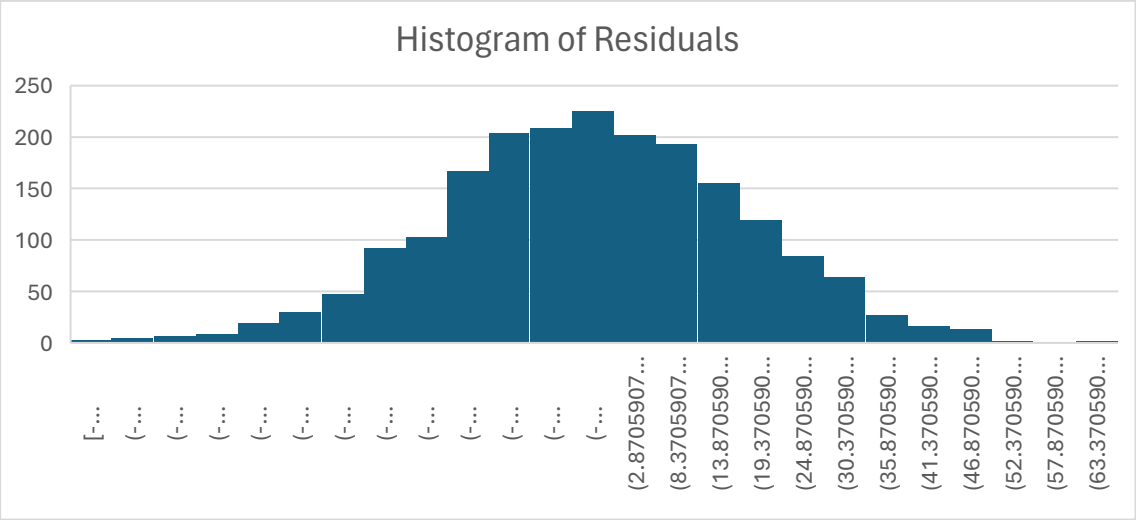


Assumptions 1 & 2: Linearity and Homoscedasticity

Interpretations:

a. There is no unusual pattern identified in Predicted Annual Sales vs Residuals Plot.
b. We could see that the residuals are randomly scattered around the horizontal axis. So, linearity assumption is fulfilled.
c. There is no visible cone shape which reflects an increase or a decrease in variance, so homoscedasticity assumption is fulfilled.

Histogram of Residuals:

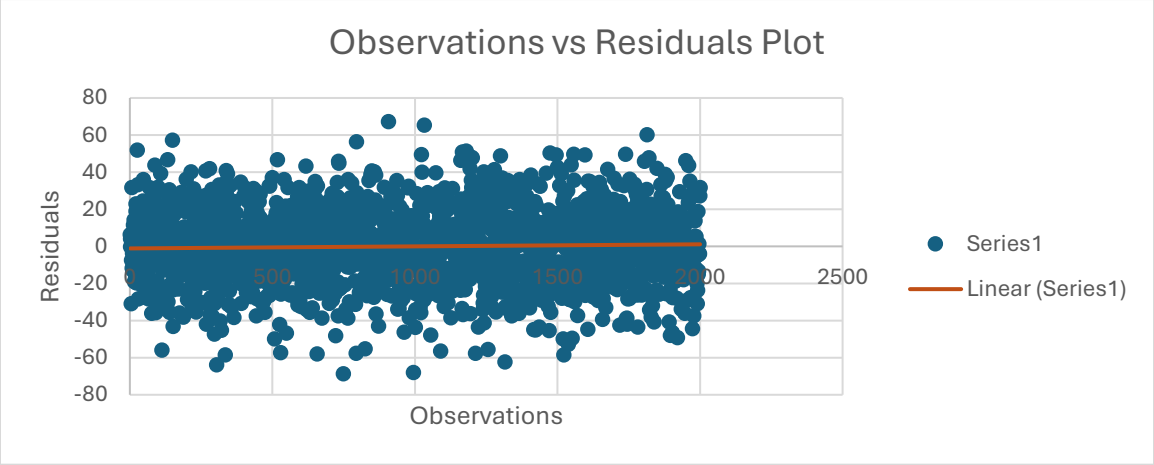


Assumption 3: Normality of Residuals (Errors)

Interpretations:

a. The Errors are normally distributed. The histogram is bell-shaped and symmetric.
b. So, the normality of residuals (errors) assumption is fulfilled.

Observations vs Residuals Plot:



Assumption 4: Independence of Errors

Interpretations:

a. Randomly scattered data in observations vs residuals plot, without any patterns (curves or funnel shapes) or trends. So, no errors are correlated with each other.
b. This says that Independence of Errors assumption is fulfilled.

Variance Inflation Factor (VIF):

This factor explains how much the variance of a predictor's coefficient is inflated due to correlation with other predictors in a regression model. This inflation in variance is caused by multicollinearity, when predictors are linearly related to each other.

Assumption 5: No Multicollinearity

- To find if there is any multicollinearity occurrence in the dataset, we need to calculate the **Variance Inflation Factor (VIF).**
- We need to create Regression Model, keeping each predictor as dependent variable against all other predictors and fetch the R square value from each regression model.
- We need to calculate VIF using R square value of each regression model.
- **Formula:** VIF = 1 / (1-R Square)

Calculation Results:

Predictor	R^2	VIF
AdBudget	0.000441788	1.000441984
Price	0.003781088	1.003795439
CustomerVisits	0.001861995	1.001865469
CompetitorPrice	0.00286244	1.002870657
North	0.343120202	1.522348538
South	0.341417996	1.518413796
East	0.34402226	1.524441973

Scale for Multicollinearity:

If VIF = 1	No multicollinearity (Perfect Full Rank)
If VIF < 2	Very Low multicollinearity (Still Acceptable)
If 2 < VIF < 5	Moderate multicollinearity (needs attention)
If VIF > 5	High multicollinearity (problematic)

Interpretations:

Predictor	**R^2**	**VIF**	**Interpretation**
AdBudget	0.000441788	1.0004	VIF ~ 1. So, no multicollinearity with other predictors.
Price	0.003781088	1.0038	VIF ~ 1. So, no multicollinearity with other predictors.
CustomerVisits	0.001861995	1.0019	VIF ~ 1. So, no multicollinearity with other predictors.
CompetitorPrice	0.00286244	1.0029	VIF ~ 1. So, no multicollinearity with other predictors.
North	0.343120202	1.5223	VIF < 2. Very low multicollinearity.
South	0.341417996	1.5184	VIF < 2. Very low multicollinearity.
East	0.34402226	1.5244	VIF < 2. Very low multicollinearity.

So, no multicollinearity assumption is fulfilled.