<u>Linear Regression on Ecommerce Data</u>

First few rows of the dataset:

Ecommerce_Customers

Email	Address	Avatar	Avg. Session Length	Time on App	Time on Website	Length of Membership	Yearly Amount Spent
mstephenson@fernandez.com	835 Frank Tunnel Wrightmouth, MI 82180-9605	Violet	34.49726772511230	12.655651149166800	39.57766801952620	4.082620632952960	587.9510539684010
hduke@hotmail.com	4547 Archer Common Diazchester, CA 06566-8576	DarkGreen	31.926272026360200	11.109460728682600	37.268958868297700	2.66403418213262	392.2049334443260
pallen@yahoo.com	24645 Valerie Unions Suite 582 Cobbborough, DC 99414-7564	Bisque	33.000914755642700	11.330278057777500	37.11059744212090	4.104543202376420	487.54750486747200
riverarebecca@gmail.com	1414 David Throughway Port Jason, OH 22070-1220	SaddleBrown	34.30555662975550	13.717513665142500	36.72128267790310	3.1201787827480900	581.8523440352180
mstephens@davidson-herman.com	14023 Rodriguez Passage Port Jacobville, PR 37242-1057	MediumAquaMarine	33.33067252364640	12.795188551078100	37.53665330059470	4.446308318351440	599.4060920457630
alvareznancy@lucas.biz	645 Martha Park Apt. 611 Jeffreychester, MN 67218-7250	FloralWhite	33.87103787934200	12.026925339755100	34.47687762925050	5.493507201364200	637.102447915074
katherine20@yahoo.com	68388 Reyes Lights Suite 692 Josephbury, WV 92213-0247	DarkSlateBlue	32.02159550138700	11.366348309710500	36.683776152869600	4.6850172465709100	521.5721747578270
awatkins@yahoo.com	Unit 6538 Box 8980 DPO AP 09026-4941	Aqua	32.739142938380300	12.35195897300290	37.373358858547600	4.4342734348999400	549.9041461052940
vchurch@walter-martinez.com	860 Lee Key West Debra, SD 97450-0495	Salmon	33.98777289568560	13.386235275676400	37.534497341555700	3.2734335777477100	570.2004089636200
bonnie69@lin.biz	PSC 2734, Box 5255 APO AA 98456-7482	Brown	31.936548618448900	11.814128294972200	37.14516822352820	3.202806071553460	427.19938489532800

The goal is to predict the <u>yearly spend of customers</u> on an e-commerce website based on features present in the data, using linear regression.

To implement linear regression from scratch, we'll do the following:

1. **Split the data into training and testing sets in a 75:25 ratio.**
2. **Implement the formula for linear regression to calculate the coefficients.**
3. **Use the coefficients to make predictions on the testing set.**
4. **Compare the predicted values with the actual values to evaluate the model.**

We'll use the 'Yearly Amount Spent' column as the target variable and the rest of the columns – 'Avg. Session Length', 'Time on App', 'Time on Website', and 'Length of Membership' – as features (explanatory variables).

1) The data has been split into training and testing sets:

- Training set: 375 samples, each with 4 features
- Testing set: 125 samples, each with 4 features

2) Next, we'll implement the formula for linear regression to calculate the coefficients. The formula for a simple linear regression model is **y = b0 + b1*x**, where **b0** is the y-intercept and **b1**

is the slope of the line. In the context of multiple linear regression, **b0** is still the y-intercept but **b1** becomes a vector of coefficients corresponding to each feature in the data.
The formula to calculate the coefficients in a multiple linear regression model is **b = (X^T * X)^-1 * X^T * y**, where **X** is the matrix of feature values, **y** is the vector of target values, **X^T** is the transpose of **X**, and **^-1** denotes the matrix inverse.

The coefficients for the linear regression model are:
[-1.04054902e+03, 2.57239444e+01, 3.81644545e+01, 3.67064559e-01, 6.12482943e+01]

These coefficients correspond to the intercept term and the features 'Avg. Session Length', 'Time on App', 'Time on Website', and 'Length of Membership', respectively.

Therefore, the model is as follows,
Yearly Amount Spent *= -1.04054902e+03 + 2.57239444e+01*****Avg. Session Length*** *+ 3.81644545e+01*****Time on App*** *+ 3.67064559e-01*****Time on Website*** *+ 6.12482943e+01*****Length of Membership***

3) Next, we'll use these coefficients to make predictions on the testing set. The formula for making predictions with a multiple linear regression model is **y_pred = b0 + b1*x1 + b2*x2 + ... + bn*xn**, where **b0** is the y-intercept, **b1** to bn are the coefficients, and **x1** to **xn** are the feature values. In matrix form, this can be simplified to **y_pred = X*b**.

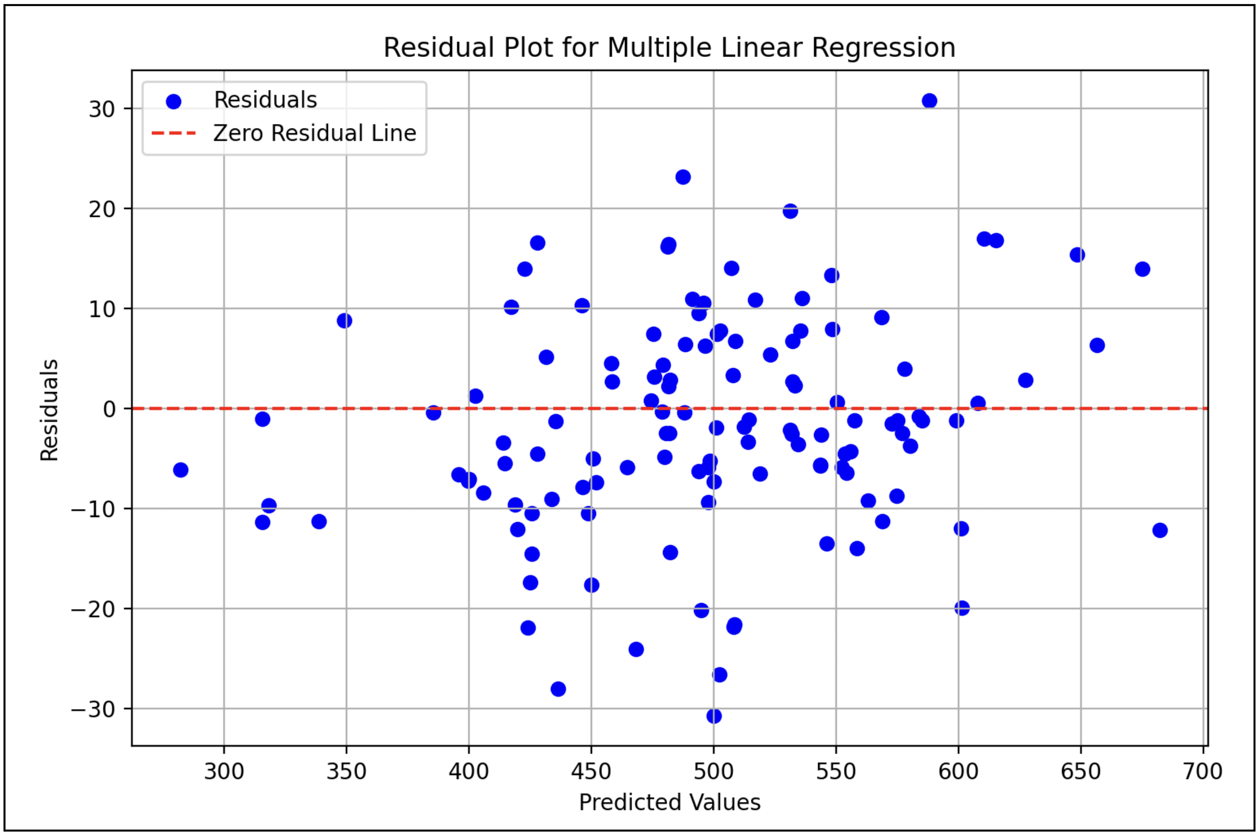
The first few predicted yearly amounts spent by customers are:
[**475.77674039**, **481.55472257**, **532.15691422**, **508.0227986**, **385.46444218**, **516.89782874**, **536.14385339**, **414.00190028**, **585.175746**, **494.69848602**]

4) Finally, we'll compare these predicted values with the actual values to evaluate the performance of our model. We'll calculate the mean squared error (MSE), which is a common metric for regression problems. The MSE is the average of the squared differences between the predicted and actual values. The lower the MSE, the better the model's performance.

The Mean Squared Error (MSE) of our model is: **120.45208751395882**

This value represents the average squared difference between the predicted and actual yearly amounts spent by customers. The lower the MSE, the better the model's predictions match the actual values.

Lastly, we'll plot a **residual plot**. A residual plot is a graph that shows the residuals (the differences between the predicted and actual values) on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.



The residuals are the differences between the actual and predicted yearly amounts spent by customers. In a well-performing model, we would expect the residuals to be randomly and evenly distributed around the horizontal axis. If there are any patterns in the residuals, it suggests that our model is not capturing some aspect of the data.

In this case, the residuals seem to be randomly distributed around the horizontal axis, suggesting that our linear regression model is a good fit for the data. However, there are a few outliers, which could be due to noise in the data or non-linear relationships that our model is not capturing.