**Task 2: Optimizing Regression Model Performance**

**Objective:**

The goal of this task was to apply **Lasso regression** to the **100\_Sales dataset** and calculate key evaluation metrics — **Mean Squared Error (MSE)**, **Mean Absolute Error (MAE)**, and **Root Mean Squared Error (RMSE)**. The performance of the model was then optimized to ensure that these values fall within the expected range of **0.111 to 1-12**.

**Dataset:**

The **100\_Sales dataset** consists of various sales-related features such as **Total Revenue**, **Total Profit**, **Unit Cost**, and **Item Type**. The dataset includes both **numerical features** (e.g., Total\_Revenue, Unit\_Cost) and **categorical features** (e.g., Region, Sales\_Channel, Order\_Priority). The goal was to predict **Total Revenue** using these features.

**Data Preparation and Preprocessing:**

1. **Handling Missing Values**:
   * Columns with missing data were removed to ensure that only complete data was used for training the model. This step was crucial for ensuring the quality and integrity of the dataset.
2. **Date Handling**:
   * The Ship\_Date column, which contained shipping dates in string format, was cleaned by replacing the / with - and then converted to a **datetime format**.
   * From this date, new features such as Ship\_Year, Ship\_Month, and Ship\_Day were extracted, which provided useful numerical information for the model.
3. **Feature Encoding**:
   * Categorical columns such as Region, Sales\_Channel, and Item\_Type were transformed into numerical representations using **one-hot encoding**. This method helps represent categorical data in a way that machine learning models can understand.
   * The Order\_Priority column, which represents the priority of orders (e.g., Low, Medium, High, Critical), was transformed using **ordinal encoding** to reflect the inherent ranking in the data.
4. **Outlier Detection and Handling**:
   * Outliers in the numerical columns (Total\_Profit, Total\_Revenue, Unit\_Cost) were detected using the **Interquartile Range (IQR)** method. Any values outside of the IQR bounds were capped to the nearest acceptable value to prevent them from disproportionately affecting the regression model.

**Modeling and Evaluation:**

1. **Lasso Regression**:
   * **Lasso regression** was applied to the dataset to predict Total\_Revenue. Lasso is a form of linear regression that uses **L1 regularization** to reduce overfitting by penalizing large coefficients, which helps in selecting the most important features.
   * The model was trained on the **training set** and evaluated on the **test set**.
2. **Evaluation Metrics**:
   * Three key evaluation metrics were calculated to assess the model's performance:
     + **Mean Absolute Error (MAE)**: Measures the average magnitude of the errors in the predictions, without considering their direction. A lower MAE indicates better performance.
     + **Mean Squared Error (MSE)**: Measures the average of the squared differences between predicted and actual values. Like MAE, lower values are better, but MSE penalizes larger errors more significantly.
     + **Root Mean Squared Error (RMSE)**: The square root of MSE, which brings the error metric back to the original units of the target variable. Again, a lower RMSE indicates a better model.

The goal was to achieve **low values** for these metrics ensuring that the model performs well without overfitting or underfitting.