BDM Assignment Report:

By: Sijal Shrestha, C0910639

1. Dataset Overview

The dataset used for this project consists of transaction data from an online retail sales platform. It includes several attributes such as Country, Description, Quantity, UnitPrice, and InvoiceDate, with the objective to predict the total sales (total_sales) for each transaction.

2. Data Cleaning and Preprocessing

The data cleaning process involved the following steps:

- **Handling Missing Values**:
    - Rows with missing or NULL values in Description, CustomerID, or InvoiceDate were identified and handled by either imputing the missing data or removing the rows where necessary.
    - The Country column was also checked for any null values and handled accordingly.
- **Data Transformation**:
    - **Feature Engineering**:
        - We created a new column, total_sales, which was calculated as Quantity * UnitPrice.
    - **One-Hot Encoding**:
        - Categorical variables such as Country were encoded using StringIndexer followed by OneHotEncoder to represent them as binary features for the model.
    - **Dropping Unnecessary Columns**:

        - Columns like InvoiceDate, Description, and CustomerID were either excluded or not used as features for the model due to their lack of relevance for predicting total_sales.

3. Key Insights from Data Analysis

During the exploratory data analysis (EDA), several key insights were obtained:

- **Sales Distribution**:
    - The dataset shows a significant variation in sales across countries. The United Kingdom had the highest total sales, followed by the Netherlands and EIRE.
    - Some countries had very low total sales, indicating that the data may be skewed or have outliers.
- **Quantity and Unit Price Correlation**:
    - There is a moderate correlation between Quantity and total_sales, while UnitPrice shows a higher positive correlation with total_sales.
- **Outliers**:
    - Outliers were detected in the UnitPrice and Quantity columns, which could have an impact on the model performance. These were handled by removing extreme outliers and normalizing the data when necessary.

4. Machine Learning Model

A **DecisionTreeRegressor** was chosen to predict total_sales based on the features:

- **Features Used**:
    - Country (encoded using One-Hot Encoding)
    - Quantity
    - UnitPrice
- **Model**:
    - **DecisionTreeRegressor**: A regression tree algorithm was selected due to its ability to model non-linear relationships between features and target variable. It is also interpretable and capable of handling both categorical and continuous data.

- **Pipeline**:
    - A pipeline was used to streamline the process, combining preprocessing steps (One-Hot Encoding, VectorAssembler) and model training (DecisionTreeRegressor) into one unified process.

5. Model Evaluation and Performance

The model performance was evaluated using the following metrics:

- **Root Mean Squared Error (RMSE)**: RMSE measures the model's prediction error. Lower RMSE indicates better performance.
    - **RMSE on test data**: 436.77

6. Model Tuning and Hyperparameter Optimization

- **Cross-Validation**: The model was trained using 5-fold cross-validation to ensure generalization and avoid overfitting.
- **Hyperparameter Tuning**:
    - The hyperparameters such as maxDepth and maxBins for the DecisionTreeRegressor were optimized using grid search within a defined parameter grid.

After implementing these tuning techniques, we observed significant improvements in our model's performance:

1. RMSE (Root Mean Square Error) Reduction:
    - The RMSE on the test data decreased, indicating that our model's predictions are closer to the actual total sales values.
    - A lower RMSE suggests improved accuracy in our sales predictions.
2. R^2 (R-squared) Improvement:
    - We saw an increase in the R^2 value on the test data.
    - This improvement indicates that our tuned model explains a larger proportion of the variance in total sales, demonstrating better fit and predictive power.

Output Interpretation:

1. Model Performance:

- RMSE: 65.46
- R-squared: 0.3708

The RMSE of 65.46 indicates the average deviation of our predictions from the actual total_sales values. The lower the RMSE, the better the model's performance.The R-squared value of 0.3708 suggests that our model explains about 37.08% of the variance in the total_sales. This indicates a moderate fit, as there's still a significant portion of the variance unexplained by our model.

2. Feature Importances:
    - Country: 0.0337 (3.37%)
    - UnitPrice: 0.0002 (0.02%)
    - Quantity: 0.0093 (0.93%)

These values show the relative importance of each feature in predicting total_sales.

- The 'Country' feature has the highest importance, suggesting that the location of the sale has the most significant impact on the total sales.
- 'Quantity' is the second most important feature.
- 'UnitPrice' has very low importance in this model.

7. Conclusion

In conclusion:

- The **DecisionTreeRegressor** model, with hyperparameter tuning and feature engineering, showed promising results with an RMSE of **436.77**.
- The analysis of sales by country and other features has provided valuable insights for potential business decisions.