Amazon Sales ML Project

So here’s what I did in this notebook:

Step 1: Understanding the data

- First I checked the dataset, looked at the columns, types, and missing values.
- Found the top 5 cities by total sales.
- Looked at monthly sales trends and visualized them to see seasonality.
- Checked which product categories bring the highest revenue.
- Calculated average order value per customer.
- Detected outliers in the sales column using IQR and made a boxplot.
- Also made a heatmap to check correlations between numeric fields.

Step 2: Classification

- I wanted to predict returns, so I set up a classification task.
- The data was imbalanced, so I used **SMOTE** to fix that.
- Tried **Logistic Regression** first, then added multi-class models.
- Checked accuracy, precision, recall, F1 — the usual stuff — and confusion matrices.

Step 3: Regression

- Here the goal was to predict sales based on other numeric features.
- Picked the numeric columns as features, used `Sales` as the target.
- Split the data into training and testing sets with `train_test_split`.
- Trained a few regression models (linear, trees, maybe forests).
- Measured them with MSE and R^2.
- Compared models visually using bar plots.

Step 4: Visualizations

- Heatmaps for correlations.
- Trend charts for monthly sales.
- Boxplots for outliers.
- Bar plots to compare models.

Tools I used

`pandas`, `numpy`, `matplotlib`, `seaborn`, `scikit-learn`, `imblearn` (for SMOTE).

Project Summary – Amazon Sales Analysis

In this project, I worked through the full data science pipeline using Amazon sales data. Here's what I accomplished:

🧹 Data Understanding & Cleaning

I explored the dataset structure, handled missing values, and checked for data quality.
I detected and treated outliers (IQR method) and examined correlations to guide feature selection.

📈 Exploratory Data Analysis (EDA)

I analyzed top cities by total sales, identified high-revenue product categories, and studied monthly sales trends.
I calculated average order values per customer and created visualizations (heatmaps, boxplots, and trend charts) for deeper insights.

🤖 Machine Learning – Classification

I set up a classification task to predict product returns.
Using SMOTE to fix class imbalance, I trained Logistic Regression and other multi-class classifiers, then evaluated them using accuracy, precision, recall, F1-scores, and confusion matrices.

📊 Machine Learning – Regression

I built regression models to predict sales amounts.
I selected numeric features, split the data into training/testing sets, trained multiple regressors (linear and tree-based), and compared models using MSE and R^2 scores.

🖼️ Visualization & Insights

I created plots to visualize sales patterns, category performance, and customer behavior.
I summarized model performance with bar plots, making it easy to compare results.

🛠️ Tools & Techniques

- **Python libraries**: Pandas, NumPy, Matplotlib, Seaborn, scikit-learn, imblearn (SMOTE)
- **Techniques**: Data cleaning, EDA, outlier detection, classification & regression modeling, model evaluation, visualization

---

🚀 End-to-End Data Science Workflow

This project demonstrates the full journey from raw data → cleaning → analysis → modeling → evaluation → visualization.
It's structured for sharing insights and could easily evolve into a production dashboard or ML pipeline.