**Multivariate Analysis: Let Your Data Speak Before You Model!**

As part of my ongoing project on **Agricultural Product Distribution Forecasting**, I delved into **Multivariate Analysis** using Seaborn’s pairplot()—a powerful EDA tool that helps uncover hidden insights among key numeric variables such as **Revenue, Quantity Sold, and Quality Score**, segmented by **Product Type**.

🔍 **Why Pair Plots?** Pair plots allow you to visualize: ✅ **Distributions** of individual variables  
✅ **Relationships** between pairs of variables  
✅ **Group patterns** by categorical labels (in my case, product types like Apples, Tomatoes, Lettuce)

Here's what I discovered and how it shaped my modeling direction:

**💡 Key Insights from the Pair Plot**

1. 🎯 **Non-linear Relationships Detected**  
   The scatter between Quantity\_Sold and Revenue revealed that the relationship isn’t strictly linear across product categories. This insight directed me to consider **non-linear models** like:

* Random Forest Regressor
* XGBoost Regressor
* Polynomial Regression

1. 🌈 **Distinct Product Clustering**  
   Color-coded distributions suggested that each product category behaves differently. Instead of treating them uniformly, I chose to:

* Include **Product** as a feature via **one-hot encoding**, or
* Model each product type **independently** to enhance specificity.

1. 🔄 **Multicollinearity Warning**  
   Quantity\_Sold and Revenue were strongly correlated. Including both in a model could introduce redundancy. ➡️ Solution: I explored **Ridge and Lasso Regression** to mitigate multicollinearity or focused on predicting just one.
2. ⚙️ **Feature Engineering Boost**  
   Quality\_Score appeared bimodal—suggesting it's categorically influenced. I considered:

* Normalization
* Binning (e.g., low, medium, high quality)
* Feature scaling to enhance model interpretability

**🔍 Conclusion: Let Your EDA Drive Your ML**

Pair plots are **not just pretty visuals**—they’re **strategic tools** that guide: 📌 Feature selection  
📌 Model architecture  
📌 Preprocessing steps

🎓 **Tech Stack Used**:

* Python (Pandas, Seaborn, Matplotlib)
* Scikit-learn, XGBoost

👨🏽‍🔬 This is part of my broader Data Science journey through the #Amdari21DaysDataChallenge where I’m solving real-world problems using ML and Predictive Analytics.

💬 How do you use multivariate analysis to inform your modeling approach?

#DataScience #MachineLearning #EDA #PredictiveAnalytics #PairPlot #FeatureEngineering #Python #XGBoost #Amdari21DaysDataChallenge #DataVisualization #DataScienceProjects #LinkedInLearning #AgriTech #CohortAnalysis #DataStrategy #AnyaegbuCasmir