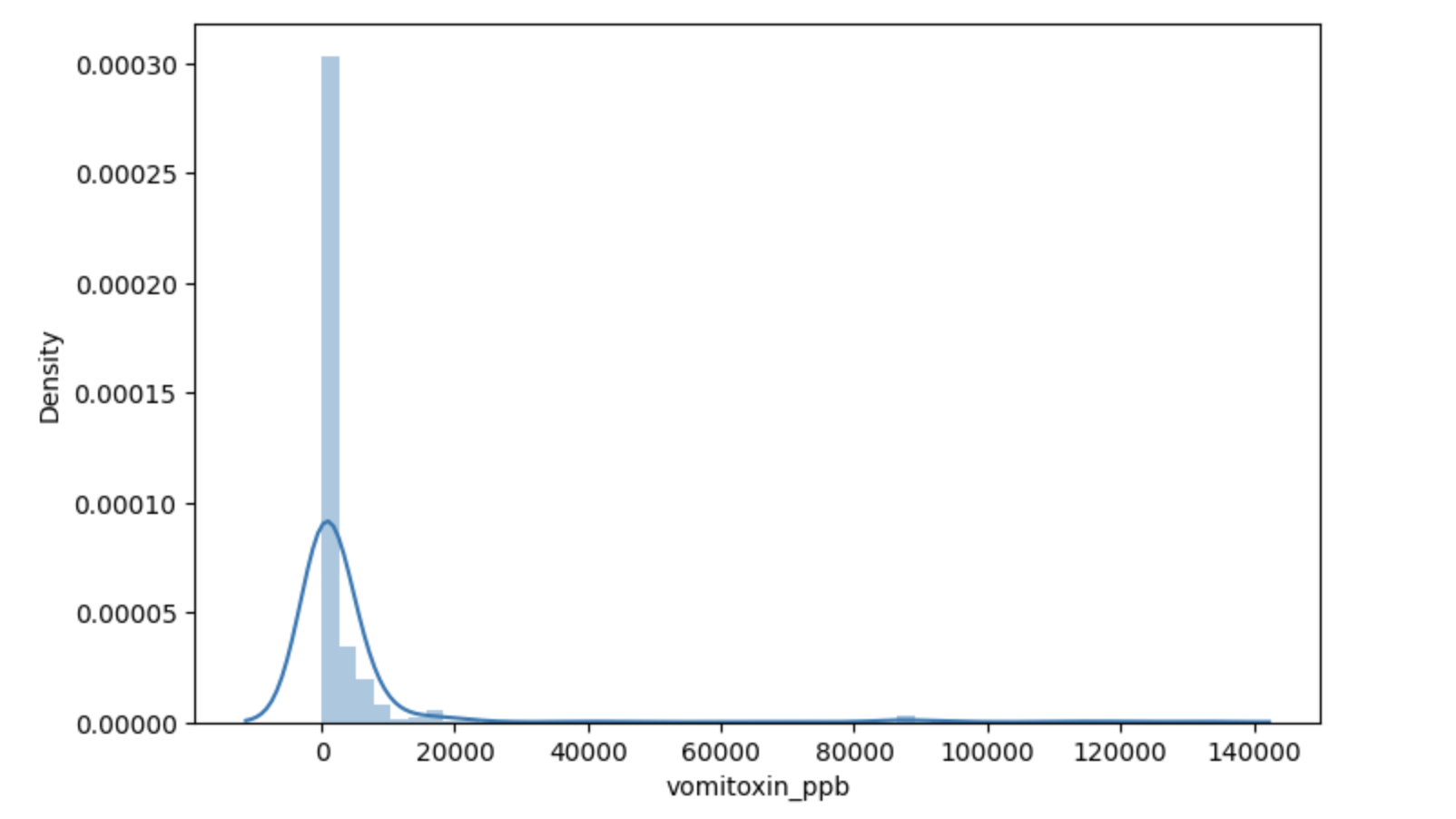
**Report : Analysis of Hyperspectral Data**

**1. Preprocessing Steps** :

**Data Inspection:**

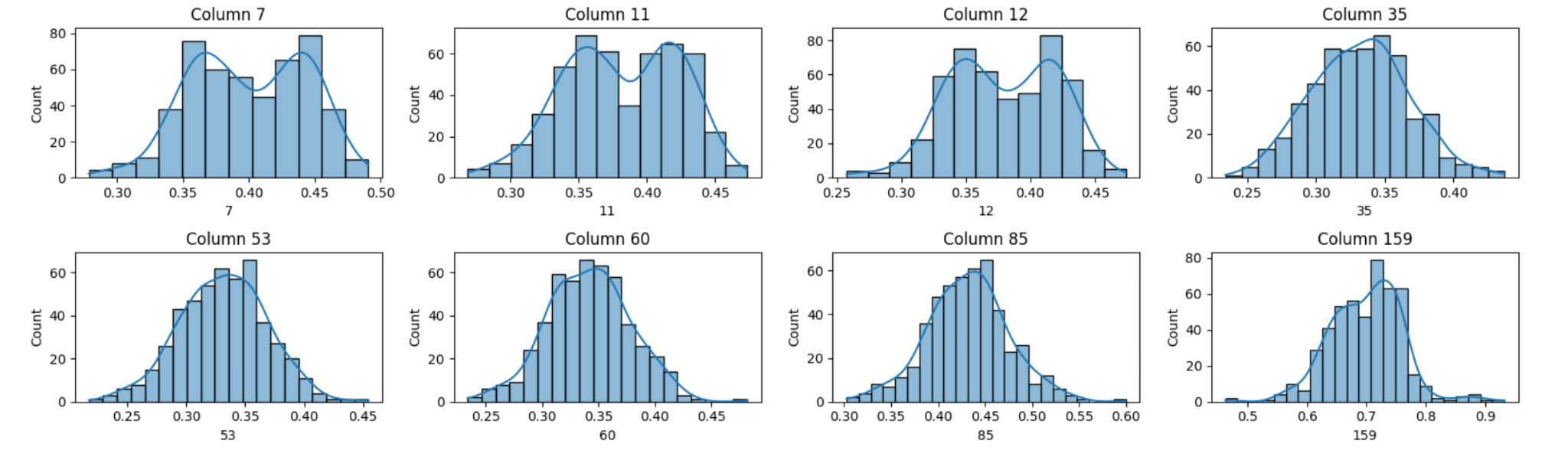
* **Missing Data and Duplicates:** A thorough check revealed no missing data or duplicate entries, ensuring that the dataset of size 500x450 was complete.

**Outlier Detection:**

****

* **Target Variable 'vomitoxin\_ppb':** Outlier analysis on the target variable identified 10 records with values greater than 20,000. These outliers were removed to prevent skewing the model training.

**Reflectance Data:**

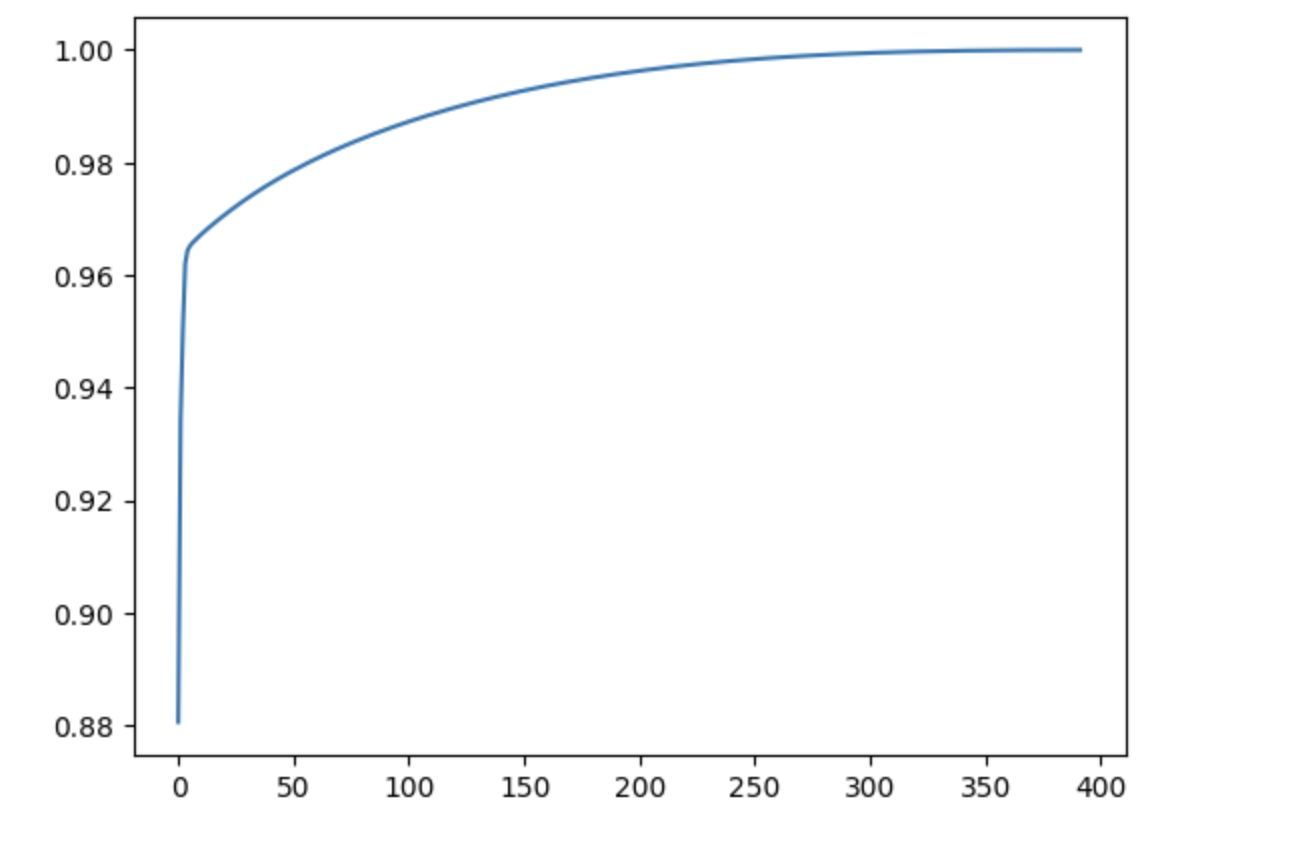
****

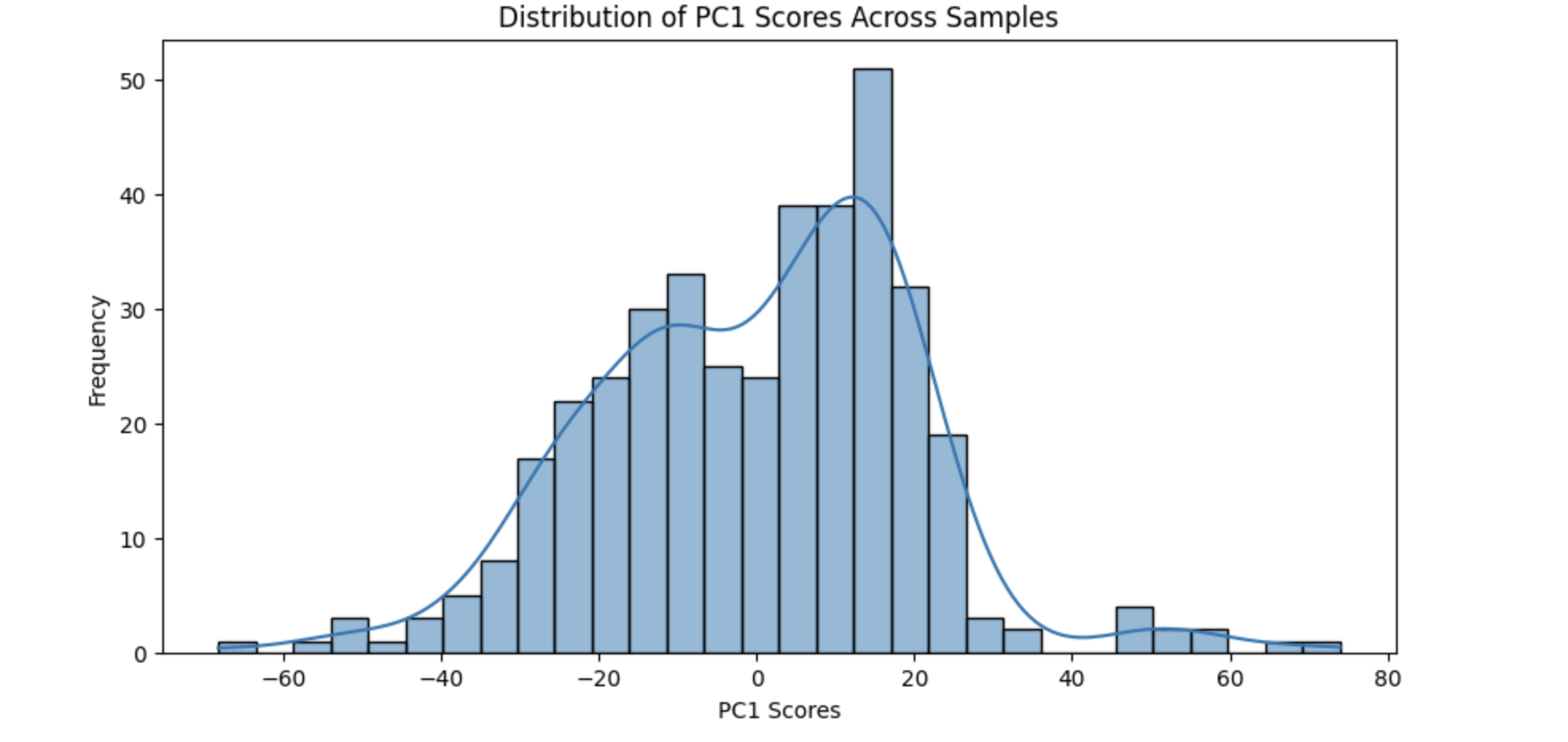
* The first 448 columns (indexed 0 to 447) contain reflectance data. Each spectral band exhibited a nearly normal distribution.

**Standardization:**

* Prior to applying dimensionality reduction techniques, standardization was performed on the reflectance data. This step ensured that each spectral band contributed equally during PCA, avoiding bias due to scale differences.

## **2. Insights from Dimensionality Reduction:**

* **PCA Application:  
  **
  + Principal Component Analysis (PCA) was applied to the standardized data with a variance retention threshold of 95%.
  + This process resulted in only 3 principal components, indicating a high level of correlation among the original spectral bands.
* **Observations from PCA:**
  + **Distribution of PCA1:** The distribution of the first principal component (PCA1) revealed two distinct peaks. This bimodal pattern suggests the possibility of different corn types or varieties within the dataset, which may contribute differently to the measured 'vomitoxin\_ppb' values.



## **3. Model Selection, Training, and Evaluation Details**

* **Model Exploration:**
  + Multiple models were considered, including:
    - Support Vector Regression (SVR)
    - Random Forest
    - Gradient Boosting
  + Hyperparameter tuning for these models was performed using **Optuna**, which streamlined the search process.
* **Selection and Tuning:**
  + **Random Forest** emerged as the best-performing model based on initial experiments.
  + Further hyperparameter tuning was conducted on the Random Forest model, using Root Mean Squared Error (RMSE) as the scoring metric.
* **Evaluation Metrics:**
  + **Mean Squared Error (MSE):** 4,036,476.5317
  + **Root Mean Squared Error (RMSE):** 2009.0984
  + **R² Score:** 0.4273

## **4. Key Findings and Suggestions for Improvement**

* **Key Findings:**
  + **High Feature Correlation:** PCA revealed that only 3 components were needed to capture 95% of the variance, highlighting significant redundancy among the spectral bands.
  + **Subgroup Identification:** The dual peaks in PCA1 suggest the presence of different corn types, which may influence the distribution of 'vomitoxin\_ppb'.
  + **Model Performance:** Despite the optimized Random Forest model performing best among the candidates, the R² score of 0.4273 indicates moderate explanatory power, leaving room for improvement.