1. Preprocessing Steps and Rationale

- **Outlier Removal:** Samples with `vomitoxin_ppb` > 8000 were removed to mitigate the effect of extreme outliers on model training.
- **Missing Values:** Checked across all spectral bands; any missing values were imputed with the **mean** of the respective columns.
- **Standardization:** Spectral data was normalized using `StandardScaler` to center the data and ensure unit variance, a necessary step for PCA and CNN training.

2. Insights from Dimensionality Reduction

- **Principal Component Analysis (PCA):** Applied to reduce the dimensionality of spectral data to the top 10 components.
- The **explained variance** ratio showed that the first few PCs captured significant variation in the dataset.
- A scatter plot of **PC1 vs PC2** showed a relatively smooth and consistent distribution of data, indicating some underlying structure.

3. Model Selection, Training, and Evaluation

- A **1D Convolutional Neural Network (CNN)** was used for regression, leveraging spatial locality in spectral features.
- **Architecture Summary:**
    - Conv1D → MaxPooling → Flatten → Dense layers.
    - **Dropout** was used to prevent overfitting.
- **Training Strategy:**
    - 80-20 train-test split.
    - EarlyStopping used to halt training when validation loss stopped improving.
- **Evaluation Metrics:**

    - Metrics such as MAE and RMSE (assumed, but not seen yet) would provide insight into regression performance.

4. Key Findings and Suggestions

- CNN performed well with normalized and filtered spectral data.
- PCA aided in understanding data variance but the model was trained on the full set of spectral features.
- **Improvements:**
    - Consider comparing PCA-reduced input vs. full-spectrum input for CNN.
    - Hyperparameter tuning (batch size, learning rate) can be applied for optimization.
    - Implement cross-validation for more robust evaluation.