**2.1. Analyze Concentrations of Metals (Au, Ag, Pb) by Purification Stage**

**Objective**: Understand how the concentrations of gold (Au), silver (Ag), and lead (Pb) change across different purification stages.

**Steps:**

1. **Identify the Relevant Columns**:
   * Look for columns related to the concentration of Au, Ag, and Pb at each stage (e.g., rougher.output.concentrate\_au, final.output.concentrate\_au, etc.).
2. **Visualize the Changes in Concentrations**:
   * Create line plots or bar charts to show how the concentrations of these metals change from the raw feed to the rougher concentrate to the final concentrate.
   * Use Seaborn or Matplotlib to plot these changes for each metal.
3. **Interpret the Results**:
   * Describe how the concentration levels change throughout the purification process. Are there stages where certain metals are more efficiently concentrated or lost?

**2.2. Compare Feed Particle Size Distributions**

**Objective**: Check if there are significant differences in feed particle size distributions between the training and test datasets.

**Steps:**

1. **Identify Feed Size Columns**:
   * Look for columns related to feed particle size (e.g., rougher.input.feed\_size).
2. **Plot the Distributions**:
   * Use histograms or KDE plots to visualize the feed particle size distribution in both the training and test sets.
   * Consider using a single plot with both distributions overlaid for direct comparison.
3. **Analyze the Distributions**:
   * Use statistical tests (like the Kolmogorov-Smirnov test) to check if there are significant differences between the distributions.
   * Assess if the model could be biased if the feed size distributions differ greatly between the training and test sets.

**2.3. Analyze Total Concentrations and Identify Anomalies**

**Objective**: Investigate the total concentrations of substances at different stages to detect any anomalies.

**Steps:**

1. **Calculate Total Concentrations**:
   * For each stage (raw feed, rougher concentrate, and final concentrate), calculate the total concentration by summing up the concentrations of all substances (Au, Ag, Pb, and others if relevant).
2. **Visualize the Total Concentrations**:
   * Create histograms or boxplots to visualize the total concentration distributions at each stage.
   * Look for outliers or abnormal values that do not fit the general pattern.
3. **Identify and Handle Anomalies**:
   * If you detect abnormal values (e.g., total concentration significantly above 100% due to measurement error or below expected levels), decide if these should be removed.
   * Document your criteria for removing anomalies and update your dataset accordingly.