Exploratory Data Analysis (EDA) — Lettuce Growth Optimization Project

1. Overview

This analysis explores environmental and operational factors influencing lettuce growth in a controlled environment. The dataset includes daily measurements of temperature, humidity, pH, TDS, rolling averages, optimal condition flags, and a composite environmental score (Env_Score). The goal of this EDA is to understand patterns, detect anomalies, and identify which conditions most strongly relate to plant growth duration.

---

2. Data Quality and Missingness

A full missing-value audit was performed across all variables.
The dataset shows **minimal missingness**, indicating strong data collection consistency. No variables required imputation at this stage, and the dataset was suitable for direct exploration analysis.

---

3. Distribution of Raw Environmental Variables

Histograms of **Temperature**, **Humidity**, **TDS (ppm)**, and **pH** reveal the natural operating ranges of the system.

Key observations:

- Temperature and humidity show relatively tight distributions, suggesting stable climate control.
- pH and TDS exhibit wider variability, which may indicate more frequent adjustments or sensor fluctuations.
- No extreme outliers were detected, but pH shows slight skewness that may influence plant stress.

These distributions help establish baseline environmental behavior before evaluating optimality.

---

4. Growth Duration Distribution

Growth_Days follows a moderately right-skewed distribution.
Most plants reach maturity within a similar time window, but a small number take significantly longer.

This suggests:

- A consistent growing process for most plants.
- A subset of plants may be affected by environmental deviations or biological variability.

Understanding these longer-growth cases is important for operational optimization.

---

5. Rolling 3-Day Environmental Trends

Time-series plots of all `*_Roll3` variables show how environmental conditions evolve over time.

Insights:

- Rolling averages smooth out daily noise and reveal broader climate patterns.
- Periods of instability (e.g., dips in humidity or spikes in TDS) are clearly visible.
- These fluctuations may align with longer Growth_Days or lower optimal condition rates.

This view is essential for diagnosing operational issues.

---

6. Correlation Structure of Numeric Variables

A correlation matrix was generated for all numeric variables.

Notable relationships:

- Env_Score shows meaningful correlation with Growth_Days, validating its usefulness as a composite KPI.
- Temperature and humidity correlate moderately with their respective optimality scores.
- Rolling averages correlate strongly with their raw counterparts, confirming correct feature engineering.

This matrix helps identify which variables may be redundant or predictive.

---

7. Environmental Score vs Growth Days

A scatterplot with a linear trendline highlights the relationship between **Env_Score** and **Growth_Days**.

Interpretation:

- Higher environmental scores generally correspond to shorter growth durations.
- The negative trend suggests that maintaining optimal conditions accelerates plant development.
- The relationship is not perfectly linear, indicating other factors may also contribute.

This is a strong justification for using Env_Score as a KPI in the dashboard.

---

8. Optimal Condition Flags Over Time

Stacked proportion charts show how often each environmental variable was within its optimal range.

Patterns observed:

- Some conditions (e.g., temperature) remain optimal, more consistent than others (e.g., TDS).
- Periods of low optimality align with dips in rolling averages.
- These patterns help identify operational bottlenecks.

This visualization is ideal for communicating system performance to stakeholders.

---

9. Composite Optimal Score Breakdown

Boxplots of the four optimal condition numeric scores reveal their variability.

Findings:

- Some optimality components are more volatile than others.
- This helps explain fluctuations in the overall Env_Score.
- It also informs weight decisions if the composite score is refined later.

---

10. Growth Patterns by Plant

Density curves of Growth_Days grouped by Plant_ID show plant-level variation.

Insights:

- Most plants follow similar growth patterns.
- A few plants show noticeably longer or shorter growth durations.
- These differences may reflect micro-environmental variation or biological differences.

This sets the stage for deeper plant-level analysis or clustering.

---

11. Summary Statistics for Reporting

A summary table of means, standard deviations, and ranges was exported for use in Tableau or reporting.

This table provides:

- Quick reference KPIs
- Baseline operational benchmarks
- Inputs for dashboard cards and tooltips
