---
title: "Plant Health Data Analysis"
author: "Ryan Evrist"
format: html
editor: visual

Analysis of Plant Health Status

1. [Introduction]{.underline}

-   **Objective:** To explore the intricate relationship between healthy plant growth and the key nutritional factors and environmental conditions that support it.

-   **Context**: Plant life is the most abundant and essential form of life on Earth. Conducting a detailed analysis of the factors that contribute to healthy plant growth can provide valuable insights for nurturing and sustaining our chlorophyll-based companions.

Key Questions: Which nutrients play the most critical role in promoting healthy plant status? How do elevated levels of soil moisture and humidity impact plant health?

2. [Data Description]{.underline}

-   **Source:\
    ** The dataset used for this analysis is the Plant Health Dataset from Kaggle.\
    [Access the dataset here](https://www.kaggle.com/datasets/ziya07/plant-health-data).

-   **Structure:\
    ** The dataset contains 1200 rows and 14 columns.

-   **Summary:\
    ** Key variables include:

    -   Numerical variables: Nutrient content in soil, ambient and soil temperatures, humidity, light intensity, and soil pH.

    -    Exploratory summaries were conducted to identify missing values, outliers, and overall trends.

-   **Tools Used:\
    ** The following R packages and functions were utilized for data loading and inspection:

    -   readr: For efficient data import (read_csv()).

    -   dplyr: For data wrangling and summarization (select(), summarize()).

    -   skimr: For detailed exploratory summaries (skim()).

    -   ggplot2: For initial visualizations of variable distributions.

-   **3. [Methodology]{.underline}**

    -   **Approach**: Begin by cleaning the dataset, including correcting any spelling errors for consistency. Analyze various combinations of nutrient levels, temperature, pH, humidity, and other environmental factors to uncover potential correlations with plant health categories: Healthy, Moderate Stress, and High Stress.

    -   **Techniques/Algorithms**: Utilize mean statistics, various ggplot visuals, and summaries of different factors.

    -   **R Code Examples:**

```{r}
```

````
#Load libraries
library(tidyverse)
library(ggplot2)
library(dplyr)
library(tidyr)
```

```{r}
#Load dataset from .csv file
plant_data <- read_csv("plant_health_data.csv")
```

```{r}
#Preview data
head(plant_data)
glimpse(plant_data)
```

```{r}
# Correct column name spelling
plant_data <- plant_data %>%
  rename(Phosphorous_Level = Phosphorus_Level)
```

```{r}
#Check for any missing data
plant_data %>%
  summarize(across(everything(), ~ sum(is.na(.)))) %>%
  rename_with(~ paste("Missing", ., sep = "_"))
```

```{r}
# Reorder Plant_Health_Status
plant_data <- plant_data %>%
  mutate(Plant_Health_Status = factor(
    Plant_Health_Status,
    levels = c("Healthy", "Moderate Stress", "High Stress")
  ))
```

```{r}
#Summarize Total Status of Each Plant(Healthy, Moderate Stress, High Stress)
status_counts <- plant_data %>%
  count(Plant_Health_Status) %>%
  mutate(percentage = n / sum(n) * 100)
```

-   299 Healthy Plants

-   401 Moderate Stress Plants

-   500 High Stress Plants

```{r}
#Pie Chart Showing Percentages of Plant Status
ggplot(status_counts, aes(x = "", y = percentage, fill = Plant_Health_Status)) +
  geom_col(width = 1, color = "white") +
  coord_polar(theta = "y") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_stack(vjust = 0.5),
            color = "white", size = 4) +
  labs(
    title = "Distribution of Plant Health Status",
    fill = "Health Status"
  ) +
````

  theme_void() +
  theme(plot.title = element_text(hjust = 0.5))
```

**Here, we will explore how nutrient and chlorophyll levels influence plant health status.**

```{r}
# Summarize data for each nutrient
nutrient_summary <- plant_data %>%
  group_by(Plant_Health_Status) %>%
  summarize(
    Mean_Nitrogen_Level = mean(Nitrogen_Level, na.rm = TRUE),
    Mean_Phosphorous_Level = mean(Phosphorous_Level, na.rm = TRUE),
    Mean_Potassium_Level = mean(Potassium_Level, na.rm = TRUE),
    Mean_Chlorophyll_Content = mean(Chlorophyll_Content, na.rm = TRUE)
  )
```

```{r}
#Reshape the Data
nutrient_summary_long <- nutrient_summary %>%
  pivot_longer(
    cols = starts_with("Mean_"),
    names_to = "Nutrient",
    values_to = "Mean_Value"
  )
```

```{r}
#Create a Faceted Line Chart
ggplot(nutrient_summary_long, aes(x = Plant_Health_Status, y = Mean_Value, group = 1)) +
  geom_line(color = "blue", size = 1) +
  geom_point(color = "red", size = 3) +
  facet_wrap(~ Nutrient, scales = "free_y", labeller = labeller(Nutrient = label_wrap_gen(10))) +
  labs(
    title = "Mean Nutrient Levels by Plant Health Status",
    x = "Plant Health Status",
    y = "Mean Value"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 45, hjust = 1),
    strip.text = element_text(size = 10)
  )
```

4. Results:

Insights from the Faceted Line Chart

-   **Overall Observation**:

    -   Healthy plants consistently exhibit the **highest mean nutrient levels** across all four nutrient categories.

-   **Phosphorous and Nitrogen Levels**:

    -   A clear trend is observed: **lower nutrient content correlates with higher plant stress**.

    -   This pattern highlights the importance of maintaining adequate Phosphorous and Nitrogen levels for plant health.
```

- **Potassium Levels**:

    - The trend does not apply to Potassium content.

    - Plants under **moderate stress have lower Potassium levels** than those under high stress, suggesting a different relationship for this nutrient.

This chart provides actionable insights into the role of nutrients in plant health, with some exceptions that merit further investigation.

Next, we will examine how environmental conditions impact plant health.

```{r}
# Summarize environmental conditions by plant health status
environment_summary <- plant_data %>%
  group_by(Plant_Health_Status) %>%
  summarize(
    Mean_Soil_Moisture = mean(Soil_Moisture, na.rm = TRUE),
    Mean_Ambient_Temperature = mean(Ambient_Temperature, na.rm = TRUE),
    Mean_Soil_Temperature = mean(Soil_Temperature, na.rm = TRUE),
    Mean_Humidity = mean(Humidity, na.rm = TRUE),
    Mean_Light_Intensity = mean(Light_Intensity, na.rm = TRUE),
    Mean_Soil_pH = mean(Soil_pH, na.rm = TRUE)
  )
```

```{r}
# Reshape the Data
environment_long <- environment_summary %>%
  pivot_longer(
    cols = starts_with("Mean_"),
    names_to = "Condition",
    values_to = "Value"
  )
```

```{r}
# Faceted line chart for environmental conditions
ggplot(environment_long, aes(x = Plant_Health_Status, y = Value, group = 1)) +
  geom_line(color = "blue", size = 1) +
  geom_point(color = "red", size = 3) +
  facet_wrap(~ Condition, scales = "free_y", ncol = 2) +
  labs(
    title = "Environmental Conditions by Plant Health Status",
    x = "Plant Health Status",
    y = "Mean Value"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```

This faceted line chart reveals distinct trends in environmental factors affecting plant health:

1. **Soil Moisture**: Healthy plants display moderate levels of soil moisture, with increasing or decreasing levels correlating with stress.

2. **Ambient Temperature**: Optimal temperature ranges are linked to healthy plants, while deviations align with moderate and high-stress conditions.

3. **Soil Temperature**: Similar to ambient temperature, soil temperature also follows a

clear pattern where extreme values are associated with plant stress.

4. **Humidity**: Healthy plants are associated with balanced humidity levels, whereas overly high or low humidity results in stress.

5. **Light Intensity**: Stress levels appear to increase when light intensity deviates significantly from optimal levels observed in healthy plants.

6. **Soil pH**: Healthy plants are associated with near-neutral soil pH levels, while high stress coincides with more acidic or alkaline conditions.

5. Discussion

Insights

- The analysis reveals a strong correlation between nutrient levels and plant health status:

    - Healthy plants consistently exhibit the highest mean levels of Nitrogen, Phosphorous, and Chlorophyll content.

    - Phosphorous and Nitrogen levels show a clear trend: lower levels are associated with increased plant stress, highlighting their critical role in maintaining plant health.

    - Potassium levels deviate from this trend, as plants under moderate stress show lower levels than those under high stress, suggesting Potassium may not be as straightforward a determinant of plant health.

- Environmental conditions such as soil moisture, temperature, and pH also significantly impact plant health:

    - Healthy plants are associated with optimal soil pH and balanced soil moisture levels.

    - High stress is linked to extreme variations in ambient temperature and soil moisture, emphasizing the need for stable environmental conditions for optimal plant growth.

Challenges

- **Missing Data**: Some variables had missing or incomplete data, requiring imputation or exclusion, which might have influenced the results.

- **Complex Interactions**: The interplay between multiple nutrients and environmental factors is complex, and simple summaries might not fully capture these dynamics.

- **Computational Limitations**: Visualizations involving large datasets required efficient resource handling and processing time adjustments.

Limitations

- **Data Quality**: The accuracy of the results heavily depends on the quality of the data collected. Any errors or inconsistencies in data collection could affect the conclusions.

- **Scope of Analysis**: The study focused on averaged data, which may overlook local or specific trends within subsets of the data.

- **Generalizability**: Findings are specific to the dataset analyzed and may not apply universally to all plant species or environmental contexts.

- **Unmeasured Variables**: Other unmeasured factors, such as pest presence or soil microbiota, could also influence plant health and were not accounted for in this analysis.

By addressing these challenges and limitations, future work can aim for more detailed modeling and broader datasets to refine these insights further.

6. Conclusion

Summary

- The analysis highlights the importance of nutrient levels and environmental conditions in determining plant health:

    - Healthy plants have the highest levels of key nutrients like Nitrogen, Phosphorous, and Chlorophyll.

    - Stress levels increase as nutrient content decreases, with clear trends for Nitrogen and Phosphorous.

    - Environmental factors, including soil moisture, temperature, and pH, play a critical role, with extremes correlating to plant stress.

- Potassium levels showed a unique pattern, with moderate-stress plants exhibiting lower levels than those under high stress, warranting further investigation.

Impact

- These findings can inform agricultural practices aimed at optimizing plant health and yield:

    - **Nutrient Management**: Insights into nutrient trends can guide fertilization strategies to maintain optimal levels of Nitrogen and Phosphorous.

    - **Environmental Monitoring**: Identifying the effects of soil moisture, temperature, and pH can aid in creating stable growing conditions for crops.

    - **Stress Mitigation**: Early detection of stress conditions through environmental monitoring and nutrient testing can prevent plant health deterioration.

Next Steps

- **Expand Data Collection**: Include more plant species, additional environmental variables, and longitudinal data to assess changes over time.

- **Detailed Modeling**: Employ advanced statistical or machine learning models to explore complex interactions between nutrients and environmental factors.

- **Targeted Experiments**: Conduct controlled studies to isolate the effects of individual nutrients and environmental factors on plant health.

- **Integration with Technology**: Develop tools or systems, such as sensors or AI-based monitoring, to apply these findings in real-time agricultural settings.

By addressing these next steps, this work can contribute to improving agricultural efficiency and ensuring sustainable plant growth practices.