Final Project – Exploratory Data Analysis (EDA)

Authors: Menashe Lorenzi & Ofek Avraham

Course: 88-6970-01

1 - Introduction

This project explores a dataset related to tomato yield. The goal is to identify the most influential factors affecting tomato production, using statistical and visual analysis.

2 - Data Preparation

Data was loaded, missing values were handled, outliers were treated, and numerical features were normalized to allow meaningful comparisons and modeling.

3 - General Overview

The dataset includes 18 columns and over 15,000 rows. Initial exploration involved checking data types and basic statistics to understand the structure and quality of the data.

4 - Feature Classification & Distributions

Variables were classified as numerical or categorical. Histograms and countplots showed how features are distributed, confirming preprocessing worked (e.g., normalization and outlier handling).

(Index A)

5 - Correlation Analysis

Pearson correlations revealed strong relationships between *tomato yield* and variables like *fruit set*, *mass of fruit*, and *average seeds*.

A very strong positive correlation (r = 0.86) was found between *Bee_1* Pollination Activity and *Average Plant Size*, indicating that higher levels of bee activity are closely associated with increased plant growth. This relationship may reflect a direct effect of pollination on plant development or a shared dependence on favorable environmental conditions.

(Index B)

6 - Feature-Level Insights

Regression plots and R^2 values helped rank variables by predictive strength. *Temperatures* features showed weaker trends , while some features like *fruit_set, mass_of_fruit and average_seeds* were shown strong connection to *tomato yield*.

(Index C)

7 - Hypothesis Testing for Categorical Variables

Although ANOVA and Kruskal-Wallis tests did not yield statistically significant p-values, the bar plots revealed interesting patterns worth further investigation.
Notably, specific levels of *bee pollination* activity were associated with distinct decreases in *average tomato yield*:

Bee_1 activity at 0.75 (106 cases)

Bee_3 activity at 0.56 (1 case)

Bee_4 activity at 0.606 and 0.62 (2 cases)

Furthermore, a *Maximum Lower Bloom Temperature* of 52 was linked to higher *average tomato yield*, suggesting a possible threshold effect. (1 case)

Regarding *rainfall, average tomato yield* was highest with only 1 *rainy day* (3509 cases), and began to decrease as the number of *rainy days* increased.
A sharp decline was observed around 26 (1 case) *rainy days*, suggesting that excessive rain may severely hinder productivity, possibly due to plant stress or reduced pollination efficiency.
This non-linear pattern highlights the importance of maintaining an optimal weather balance to maximize yield.

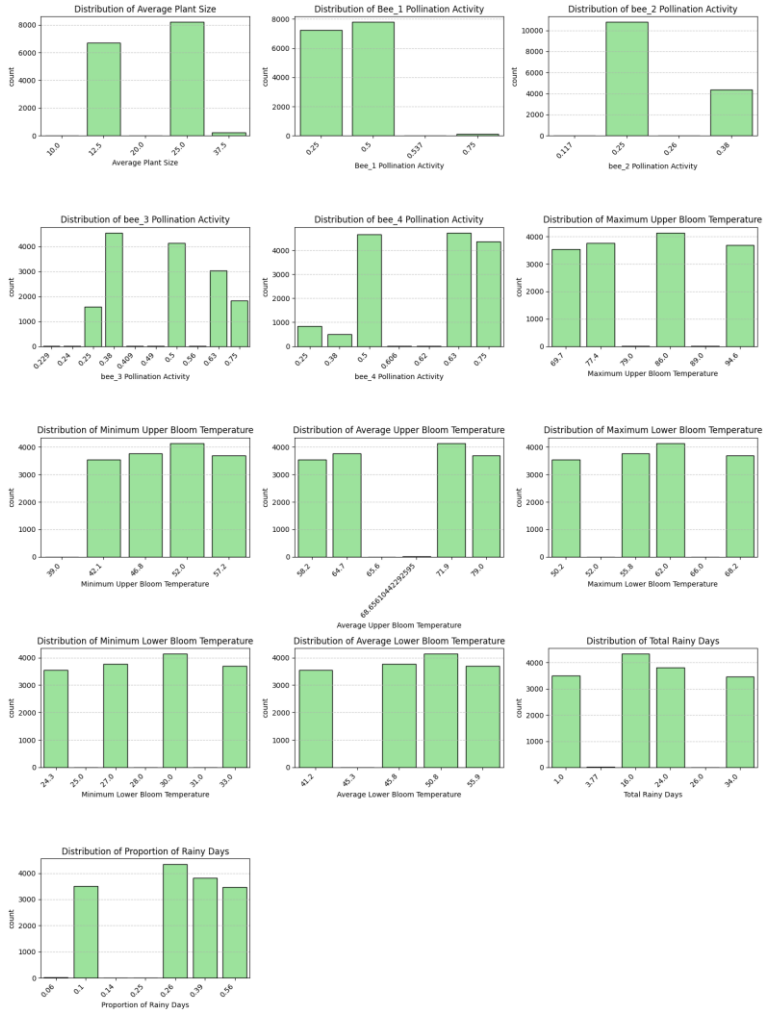
Finally, an inverse relationship was observed between *average plant size* and *average tomato yield*, where larger plants tended to produce lower yields on average — potentially due to resource competition or overgrowth effects. (25 – 8188 cases, 12.5 - 6697 cases, 37.5 – 247 cases, 20 – 8 cases, 10 – one case)

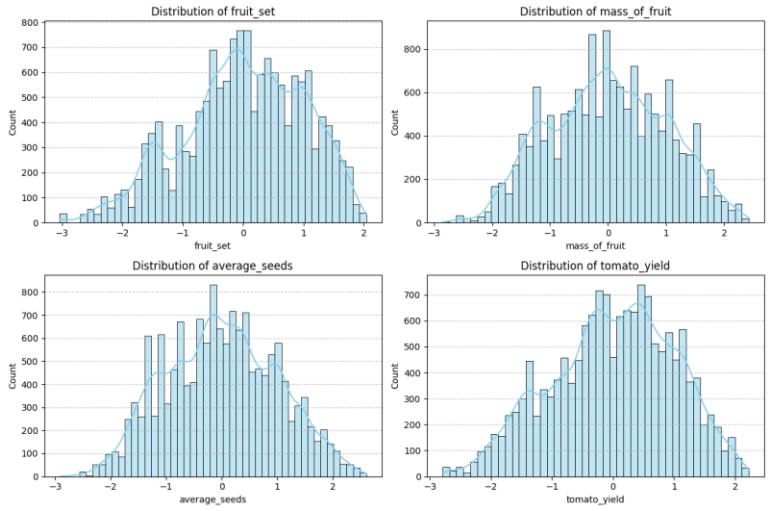
9 - Conclusions

This analysis showed that ***fruit set***, ***mass of fruit***, and ***average seeds*** are key predictors of tomato yield.
In contrast, ***larger plant size*** appears to have a ***negative influence*** on yield, we suggest that {12.5} is the ideal plant size.
Bee activity and ***rainfall*** exhibited more complex or weaker effects.
Interestingly, **a very strong positive correlation (r = 0.86)** was observed between ***Bee_1 Pollination Activity*** and ***Average Plant Size***, suggesting that increased bee activity may promote plant growth — a potentially actionable agricultural insight.

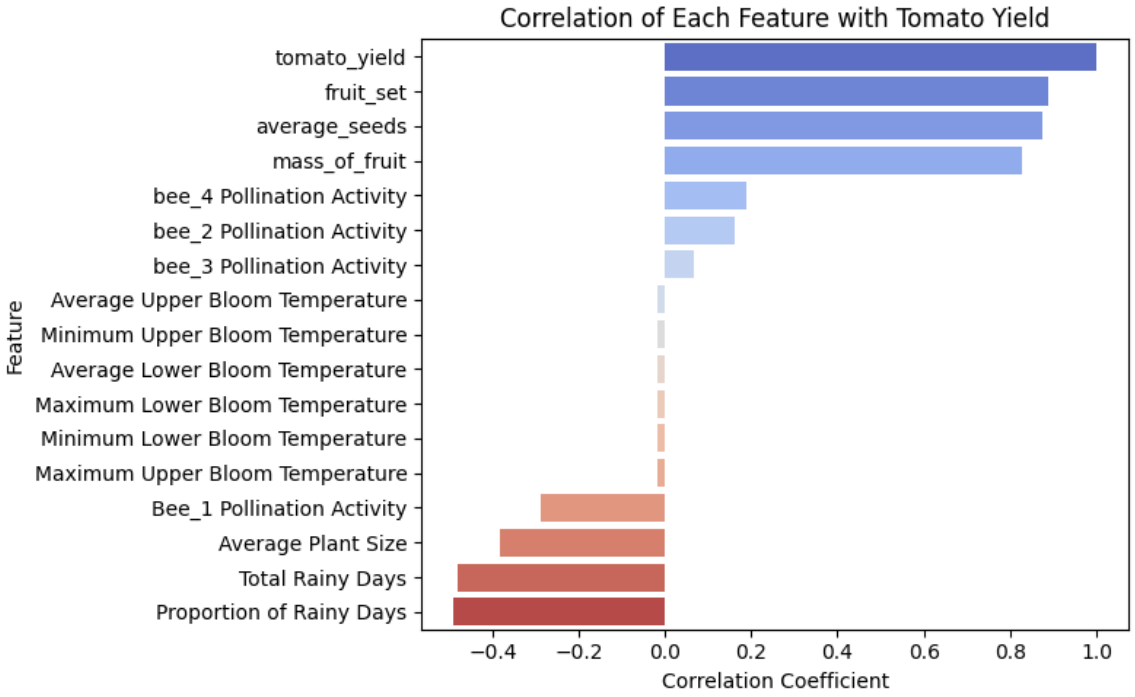
We suggest that future work should include **predictive modeling** and explore potential **interaction effects** between pollination and environmental conditions, such as rainfall and temperature.

Index A: Feature Classification & Distributions





Index B: Correlation Analysis



fruit_set 0.887666

average_seeds 0.873695

mass_of_fruit 0.828310

bee_4 Pollination Activity 0.187060

bee_2 Pollination Activity 0.160939

bee_3 Pollination Activity 0.066978

Average Upper Bloom Temperature -0.016546

Minimum Upper Bloom Temperature -0.016805

Average Lower Bloom Temperature -0.016966

Maximum Lower Bloom Temperature -0.017086

Minimum Lower Bloom Temperature -0.017204

Maximum Upper Bloom Temperature -0.017404

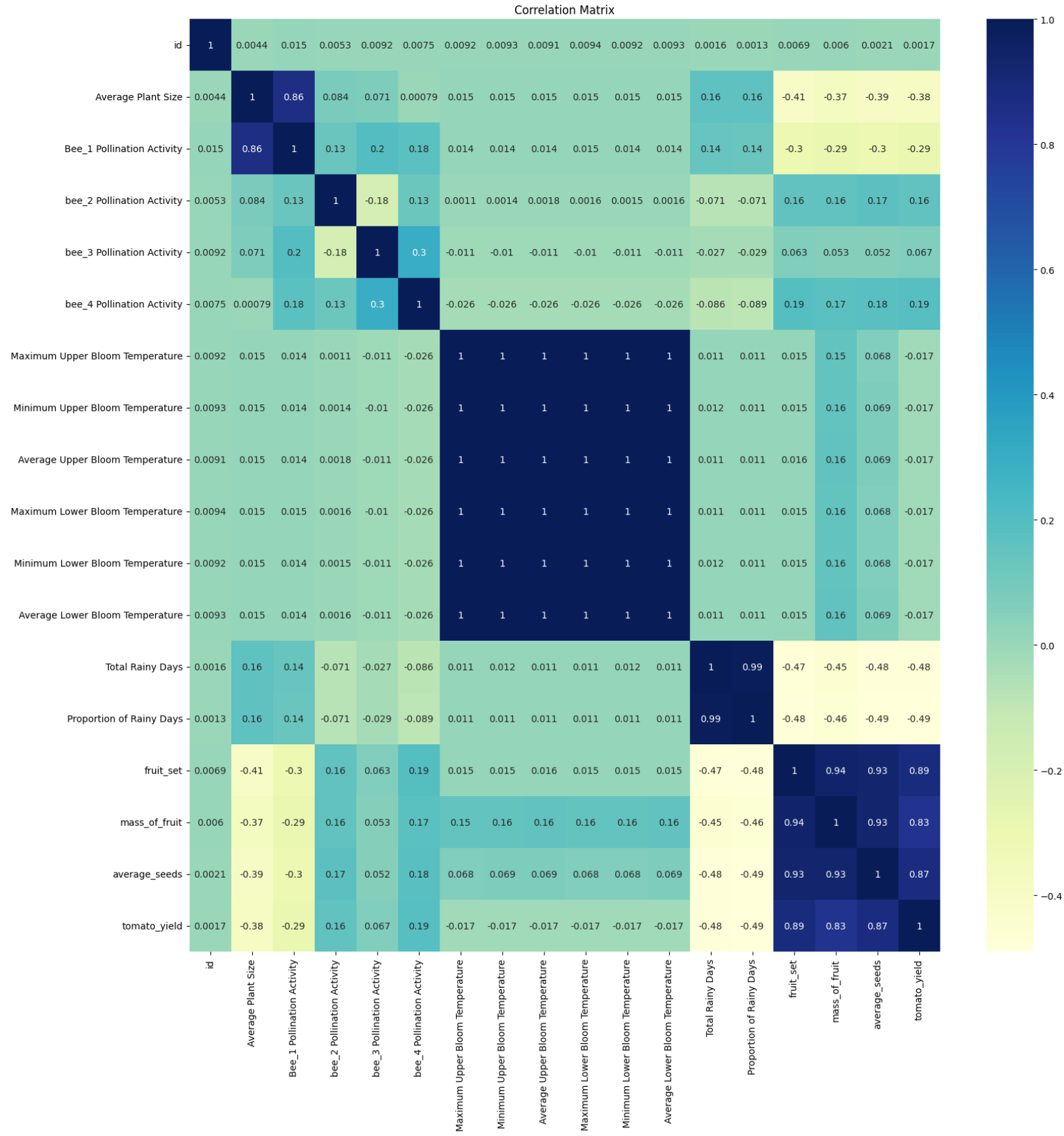
Bee_1 Pollination Activity -0.287213

Average Plant Size -0.381544

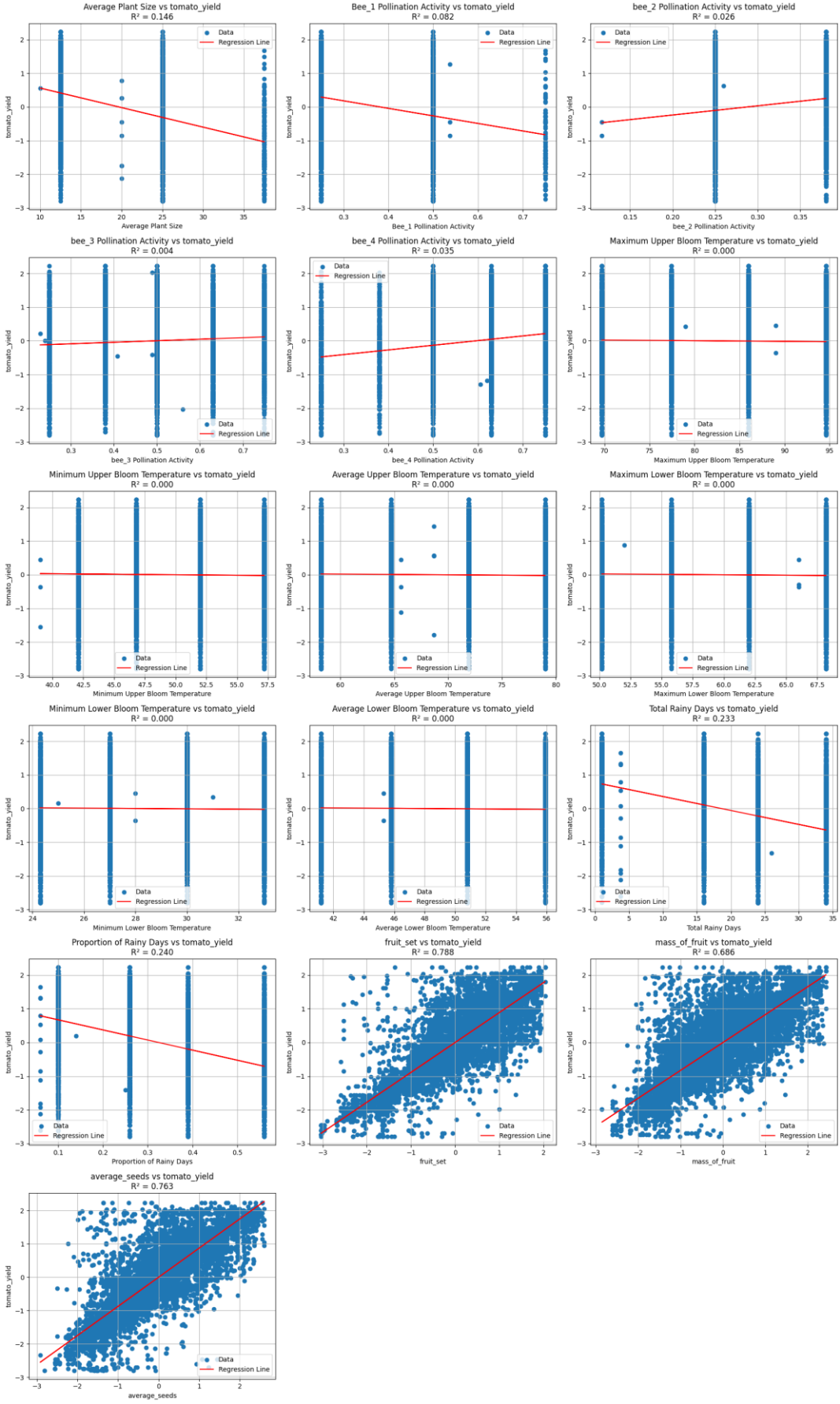
Total Rainy Days -0.482227

Proportion of Rainy Days -0.490211

Correlation Matrix to Explore All Pairwise Relationships



Index C: Regression plots



Index D: Hypothesis Testing for Categorical Variables