# 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<sup>2</sup> 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.

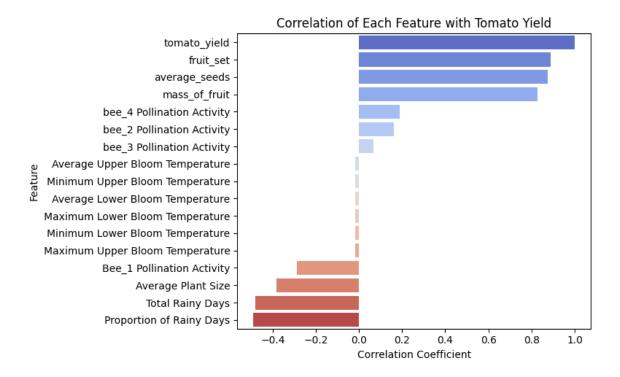
Interestingly, a very strong positive correlation (r = 0.86) was observed between *Bee\_1 Pollination Activity* and *Average Plant Size*, suggesting that increased bee activity may promote plant growth — a potentially actionable agricultural insight.

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

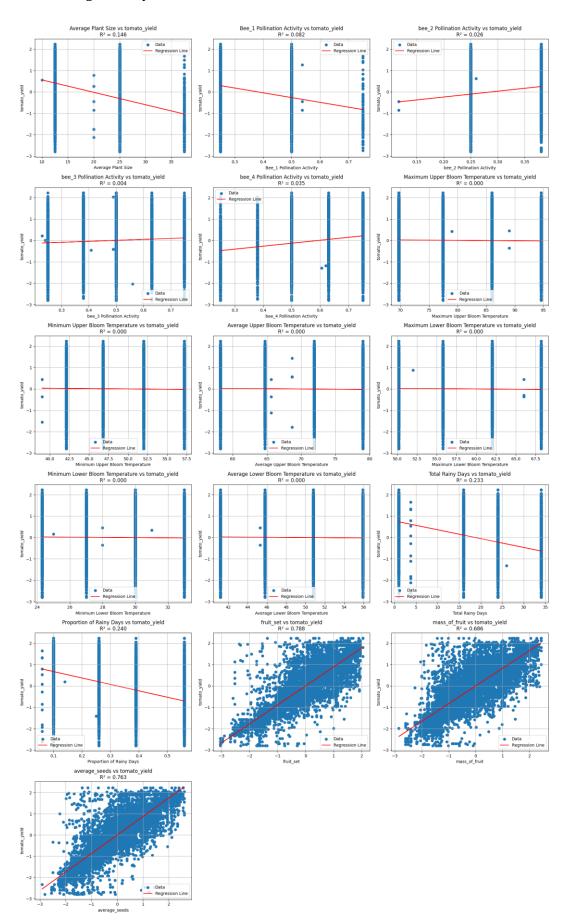
Proportion of Rainy Days -0.490211

# Correlation Matrix to Explore All Pairwise Relationships

		Correlation Matrix																
id -	- 1	0.0044	0.015	0.0053	0.0092	0.0075	0.0092	0.0093	0.0091	0.0094	0.0092	0.0093	0.0016	0.0013	0.0069	0.006	0.0021	0.0017
Average Plant Size -	- 0.0044	1	0.86	0.084	0.071	0.00079	0.015	0.015	0.015	0.015	0.015	0.015	0.16	0.16	-0.41	-0.37	-0.39	-0.38
Bee_1 Pollination Activity	- 0.015	0.86	1	0.13	0.2	0.18	0.014	0.014	0.014	0.015	0.014	0.014	0.14	0.14	-0.3	-0.29	-0.3	-0.29
bee_2 Pollination Activity	- 0.0053	0.084	0.13	1	-0.18	0.13	0.0011	0.0014	0.0018	0.0016	0.0015	0.0016	-0.071	-0.071	0.16	0.16	0.17	0.16
bee_3 Pollination Activity	- 0.0092	0.071	0.2	-0.18	1	0.3	-0.011	-0.01	-0.011	-0.01	-0.011	-0.011	-0.027	-0.029	0.063	0.053	0.052	0.067
bee_4 Pollination Activity	- 0.0075	0.00079	0.18	0.13		1	-0.026	-0.026	-0.026	-0.026	-0.026	-0.026	-0.086	-0.089	0.19	0.17	0.18	0.19
Maximum Upper Bloom Temperature -	- 0.0092	0.015	0.014	0.0011	-0.011	-0.026	1	1	1	1	1	1	0.011	0.011	0.015	0.15	0.068	-0.017
Minimum Upper Bloom Temperature	- 0.0093	0.015	0.014	0.0014	-0.01	-0.026	1	1	1	1	1	1	0.012	0.011	0.015	0.16	0.069	-0.017
Average Upper Bloom Temperature	- 0.0091	0.015	0.014	0.0018	-0.011	-0.026	1	1	1	1	1	1	0.011	0.011	0.016	0.16	0.069	-0.017
Maximum Lower Bloom Temperature	- 0.0094	0.015	0.015	0.0016	-0.01	-0.026	1	1	1	1	1	1	0.011	0.011	0.015	0.16	0.068	-0.017
Minimum Lower Bloom Temperature	- 0.0092	0.015	0.014	0.0015	-0.011	-0.026	1	1	1	1	1	1	0.012	0.011	0.015	0.16	0.068	-0.017
Average Lower Bloom Temperature	- 0.0093	0.015	0.014	0.0016	-0.011	-0.026	1	1	1	1	1	1	0.011	0.011	0.015	0.16	0.069	-0.017
Total Rainy Days	- 0.0016	0.16	0.14	-0.071	-0.027	-0.086	0.011	0.012	0.011	0.011	0.012	0.011	1	0.99	-0.47	-0.45	-0.48	-0.48
Proportion of Rainy Days	- 0.0013	0.16	0.14	-0.071	-0.029	-0.089	0.011	0.011	0.011	0.011	0.011	0.011	0.99	1	-0.48	-0.46	-0.49	-0.49
fruit_set -	- 0.0069	-0.41	-0.3	0.16	0.063	0.19	0.015	0.015	0.016	0.015	0.015	0.015	-0.47	-0.48	1	0.94	0.93	0.89
mass_of_fruit -	- 0.006	-0.37	-0.29	0.16	0.053	0.17	0.15	0.16	0.16	0.16	0.16	0.16	-0.45	-0.46	0.94	1	0.93	0.83
average_seeds -	- 0.0021	-0.39	-0.3	0.17	0.052	0.18	0.068	0.069	0.069	0.068	0.068	0.069	-0.48	-0.49	0.93	0.93	1	0.87
tomato_yield -	- 0.0017	-0.38	-0.29	0.16	0.067	0.19	-0.017	-0.017	-0.017	-0.017	-0.017	-0.017	-0.48	-0.49	0.89	0.83	0.87	1
	.9	Average Plant Size -	Bee_1 Pollination Activity -	bee_2 Pollination Activity -	bee_3 Pollination Activity -	bee_4 Pollination Activity -	Maximum Upper Bloom Temperature -	Minimum Upper Bloom Temperature –	Average Upper Bloom Temperature –	Maximum Lower Bloom Temperature -	Minimum Lower Bloom Temperature -	Average Lower Bloom Temperature -	Total Rainy Days -	Proportion of Rainy Days -	fruit_set -	mass_of_fruit -	average_seeds -	tomato_yield -

- 0.0

**Index C:** Regression plots



**Index D:** Hypothesis Testing for Categorical Variables

