Data Collection and Preprocessing Phase

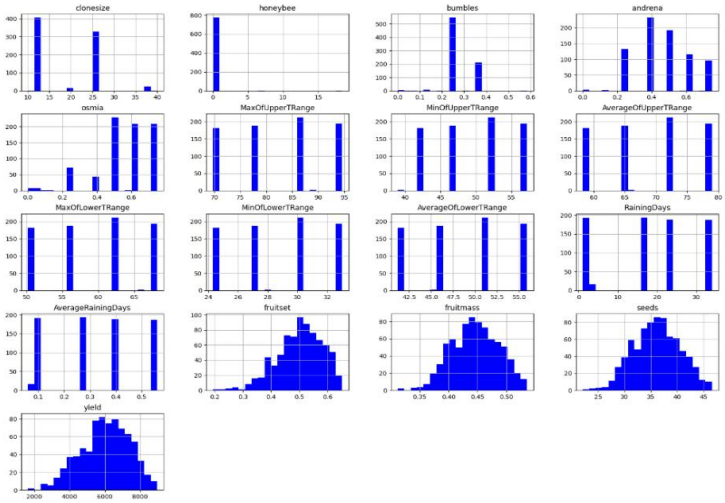
Date	12th June 2024
Team ID	SWTID1720109498
Project Title	Blueberry Yield Predictor
Maximum Marks	6 Marks

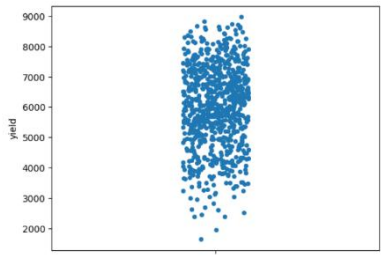
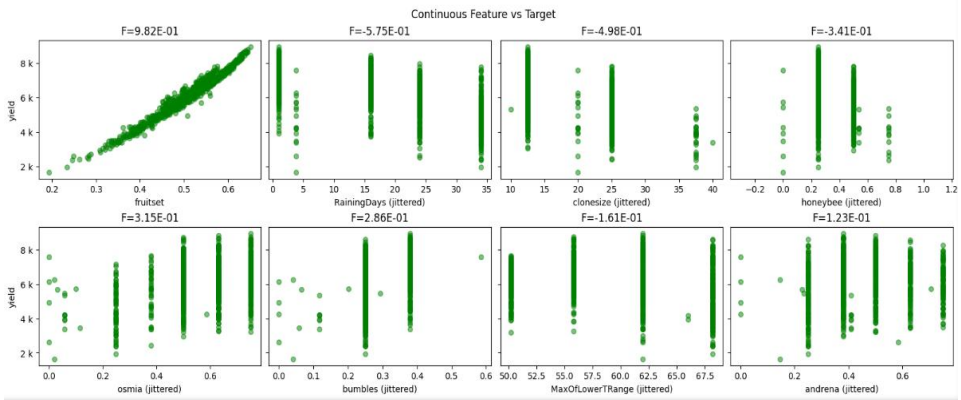
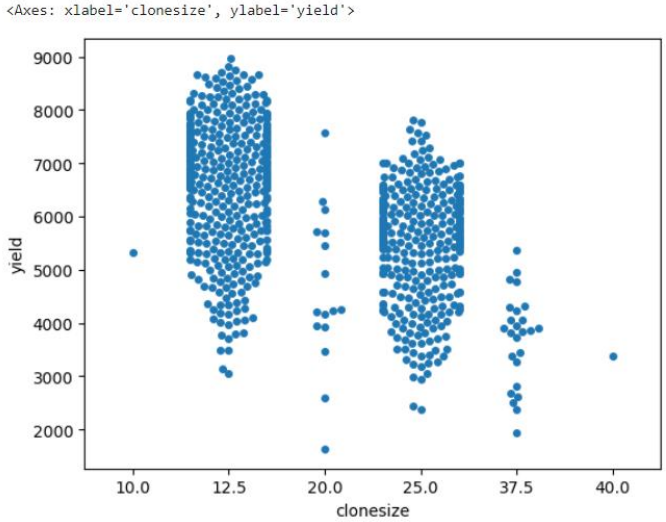
Data Exploration and Preprocessing

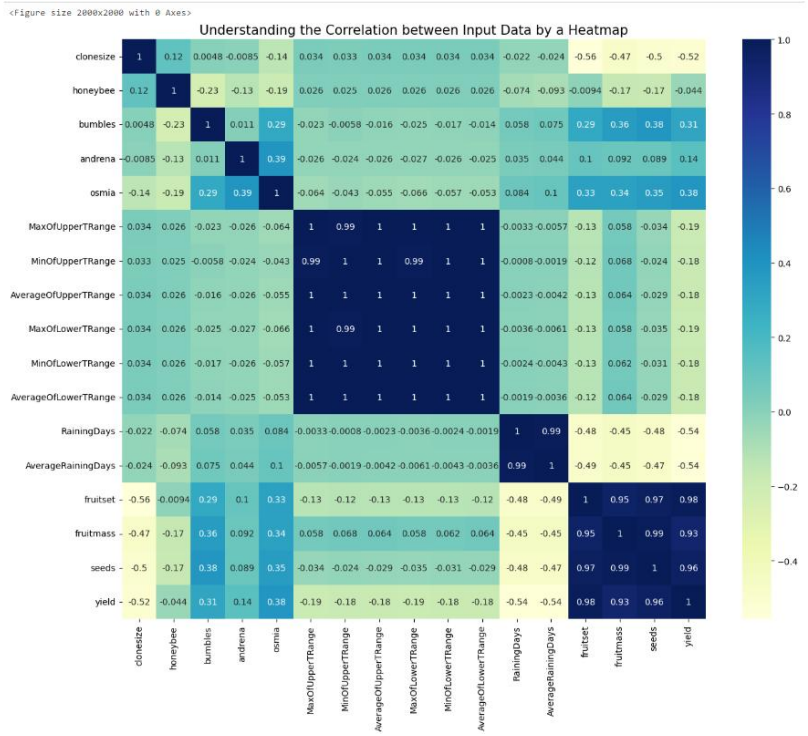
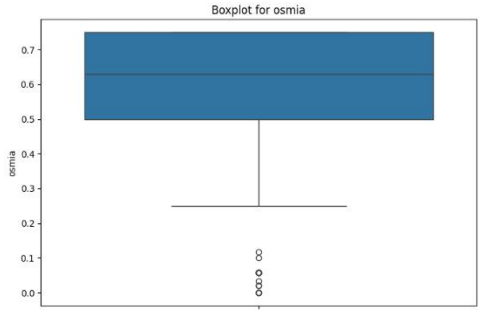
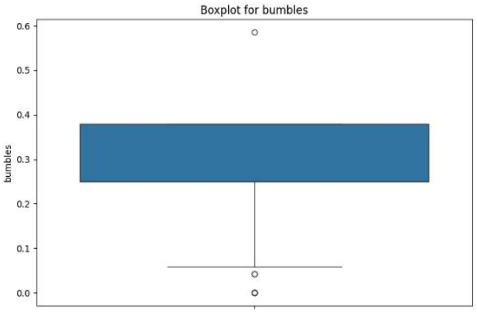
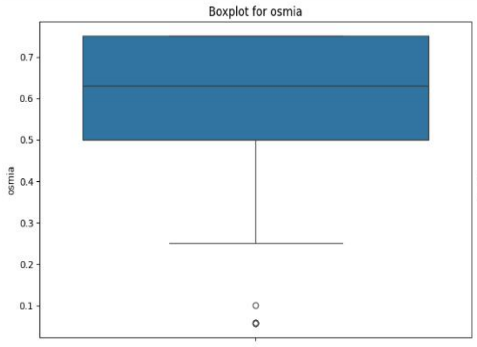
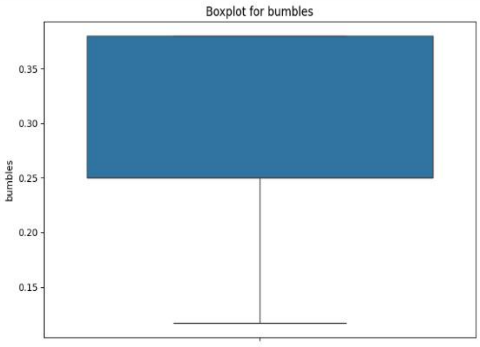
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	
Univariate Analysis	

	clonesize	honeybee	bumbles	andrena	osmia	MaxOfUpperTRange	MinOfUpperTRange	AverageOfUpperTRange	MaxOfLowerTRange
count	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000
mean	18.767696	0.417133	0.282389	0.468817	0.562062	82.277091	49.700515	68.723037	59.309395
std	6.999063	0.978904	0.066343	0.161052	0.169119	9.193745	5.595769	7.676984	6.647760
min	10.000000	0.000000	0.000000	0.000000	0.000000	69.700000	39.000000	58.200000	50.200000
25%	12.500000	0.250000	0.250000	0.380000	0.500000	77.400000	46.800000	64.700000	55.800000
50%	12.500000	0.250000	0.250000	0.500000	0.630000	86.000000	52.000000	71.900000	62.000000
75%	25.000000	0.500000	0.380000	0.630000	0.750000	89.000000	52.000000	71.900000	66.000000
max	40.000000	18.430000	0.585000	0.750000	0.750000	94.600000	57.200000	79.000000	68.200000



	
Bivariate Analysis	 

Multivariate Analysis	
Outliers and Anomalies	BEFORE REMOVAL   AFTER REMOVAL  

Data Preprocessing Code Screenshots

Loading Data

```
df = pd.read_csv('WildBlueberryPollinationSimulationData (1).csv')
df
```

	Row#	clonesize	honeybee	bumbles	andrena	osmia	MaxOfUpperTRange	MinOfUpperTRange	AverageOfUpperTRange	MaxOfLowerTRange	MinOfLowerTRa
0	0	37.5	0.750	0.250	0.250	0.250	86.0	52.0	71.9	62.0	
1	1	37.5	0.750	0.250	0.250	0.250	86.0	52.0	71.9	62.0	
2	2	37.5	0.750	0.250	0.250	0.250	94.6	57.2	79.0	68.2	
3	3	37.5	0.750	0.250	0.250	0.250	94.6	57.2	79.0	68.2	
4	4	37.5	0.750	0.250	0.250	0.250	86.0	52.0	71.9	62.0	
...	...	...	...	...	...	...	...	...	...	...	
772	772	10.0	0.537	0.117	0.409	0.058	86.0	52.0	71.9	62.0	
773	773	40.0	0.537	0.117	0.409	0.058	86.0	52.0	71.9	62.0	
774	774	20.0	0.537	0.117	0.409	0.058	86.0	52.0	71.9	62.0	

Handling Missing Data

No missing values in the dataset –

```
df.isna().sum()
```

```
clonesize               0
honeybee                0
bumbles                 0
andrena                 0
osmia                   0
MaxOfUpperTRange        0
MinOfUpperTRange        0
AverageOfUpperTRange    0
MaxOfLowerTRange        0
MinOfLowerTRange        0
AverageOfLowerTRange    0
RainingDays             0
AverageRainingDays      0
fruitset                0
fruitmass               0
seeds                   0
yield                   0
dtype: int64
```

Data Transformation

Removing outliers –

```
from scipy import stats
# Removing outliers
z_scores = np.abs(stats.zscore(df.select_dtypes(include=[np.number])))
threshold = 3
outliers = (z_scores > threshold).any(axis=1)
new_df = df[~outliers]
num_outliers_removed = outliers.sum()
print(f"Number of outliers removed: {num_outliers_removed}")
```

```
Number of outliers removed: 13
```

Feature Engineering

Removing unwanted columns after visualizing the dataset –

```
new_df = new_df.drop(columns=['bumbles','fruitmass','AverageRainingDays','fruitset','MaxOfUpperTRange','MaxOfLowerTRange', 'MinOfLowerTRange'])
new_df.head()

# ALL THE ABOVE COLUMNS HAVE HIGH CORRELATION WITH OTHER COLUMNS, SO THEY ARE BEING REMOVED
```

	clonesize	honeybee	andrena	osmia	MinOfUpperTRange	AverageOfUpperTRange	AverageOfLowerTRange	RainingDays	seeds	yield
0	37.5	0.75	0.25	0.25	52.0	71.9	50.8	16.0	31.678898	3813.165795
1	37.5	0.75	0.25	0.25	52.0	71.9	50.8	1.0	33.449385	4947.605663
2	37.5	0.75	0.25	0.25	57.2	79.0	55.9	16.0	30.546306	3866.798965
3	37.5	0.75	0.25	0.25	57.2	79.0	55.9	1.0	31.562586	4303.943030
4	37.5	0.75	0.25	0.25	52.0	71.9	50.8	24.0	28.873714	3436.493543

Save Processed Data

-