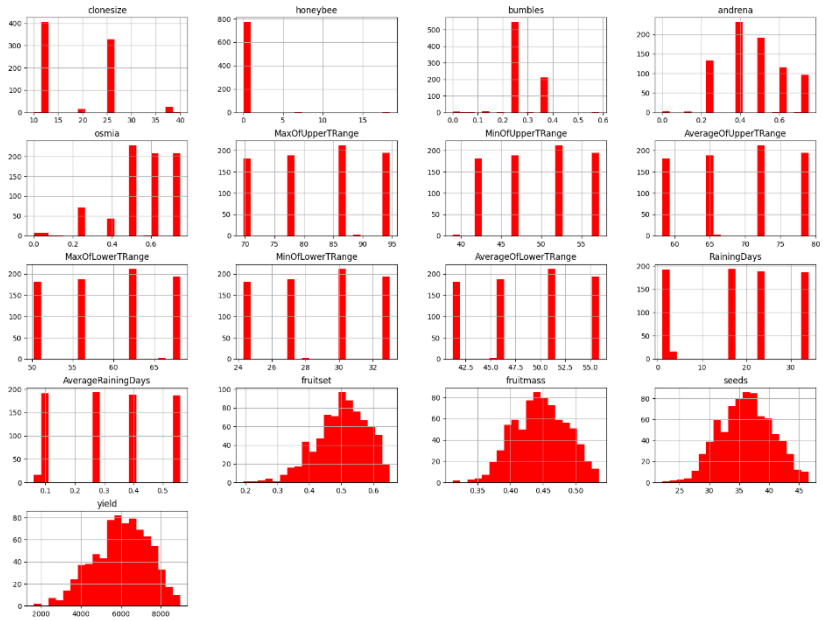
Data Collection and Preprocessing Phase

Date	12 July 2024
Team ID	SWTID1720077079
Project Title	Wild Blueberry Yield Prediction
Maximum Marks	6 Marks

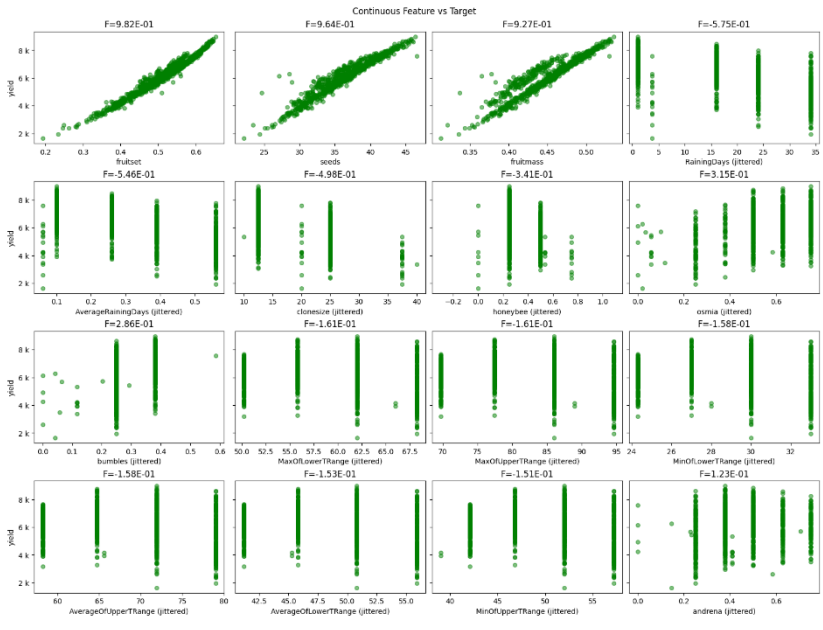
Section	Description
Data Overview	Data columns (total 17 columns): # Column Non-Null Count Dtype --- ------ -------------- ----- 0 clonesize 777 non-null float64 1 honeybee 777 non-null float64 2 bumbles 777 non-null float64 3 andrena 777 non-null float64 4 osmia 777 non-null float64 5 MaxOfUpperTRange 777 non-null float64 6 MinOfUpperTRange 777 non-null float64 7 AverageOfUpperTRange 777 non-null float64 8 MaxOfLowerTRange 777 non-null float64 9 MinOfLowerTRange 777 non-null float64 10 AverageOfLowerTRange 777 non-null float64 11 RainingDays 777 non-null float64 12 AverageRainingDays 777 non-null float64 13 fruitset 777 non-null float64 14 fruitmass 777 non-null float64 15 seeds 777 non-null float64 16 yield 777 non-null float64 dtypes: float64(17)

Univariate Analysis

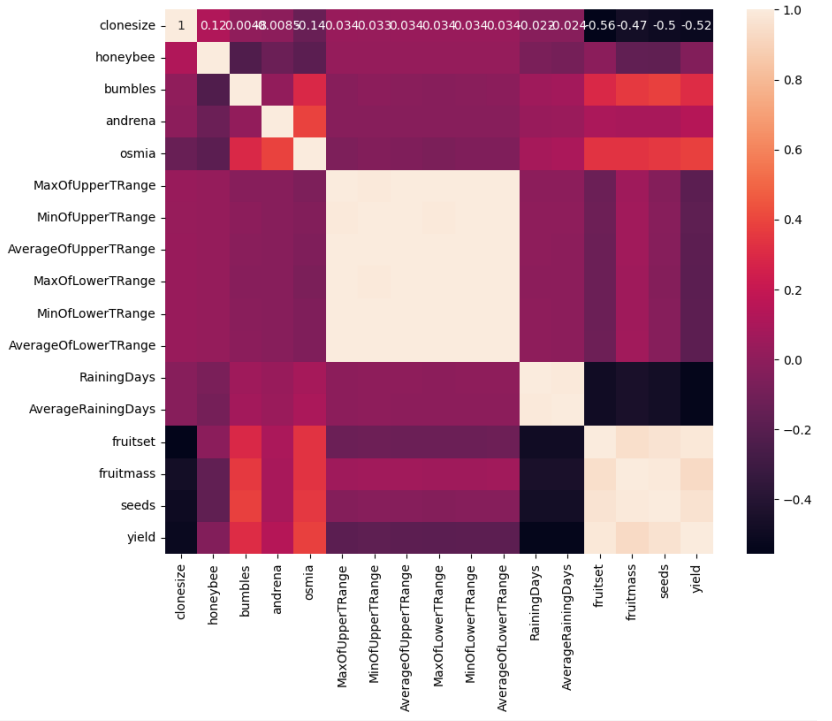
	count	mean	std	min	25%	50%	75%	max
clonesize	773.0	18.761320	7.016608	10.000000	12.500000	12.500000	25.000000	40.000000
honeybee	773.0	0.354427	0.135574	0.000000	0.250000	0.250000	0.500000	0.750000
bumbles	773.0	0.283712	0.063874	0.000000	0.250000	0.250000	0.380000	0.585000
andrena	773.0	0.470757	0.159029	0.000000	0.380000	0.500000	0.630000	0.750000
osmia	773.0	0.564900	0.164867	0.000000	0.500000	0.630000	0.750000	0.750000
MaxOfUpperTRange	773.0	82.257827	9.213615	69.700000	77.400000	86.000000	94.600000	94.600000
MinOfUpperTRange	773.0	49.688616	5.607792	39.000000	46.800000	52.000000	57.200000	57.200000
AverageOfUpperTRange	773.0	68.706598	7.693431	58.200000	64.700000	71.900000	79.000000	79.000000
MaxOfLowerTRange	773.0	59.295472	6.662130	50.200000	55.800000	62.000000	68.200000	68.200000
MinOfLowerTRange	773.0	28.683441	3.216463	24.300000	27.000000	30.000000	33.000000	33.000000
AverageOfLowerTRange	773.0	48.601811	5.428795	41.200000	45.800000	50.800000	55.900000	55.900000
RainingDays	773.0	18.384528	12.110224	1.000000	3.770000	16.000000	24.000000	34.000000
AverageRainingDays	773.0	0.321345	0.170694	0.060000	0.100000	0.260000	0.390000	0.560000
yield	773.0	6014.170238	1359.823169	1637.704022	5124.854901	6107.382466	7024.748359	8969.401842



Bivariate Analysis

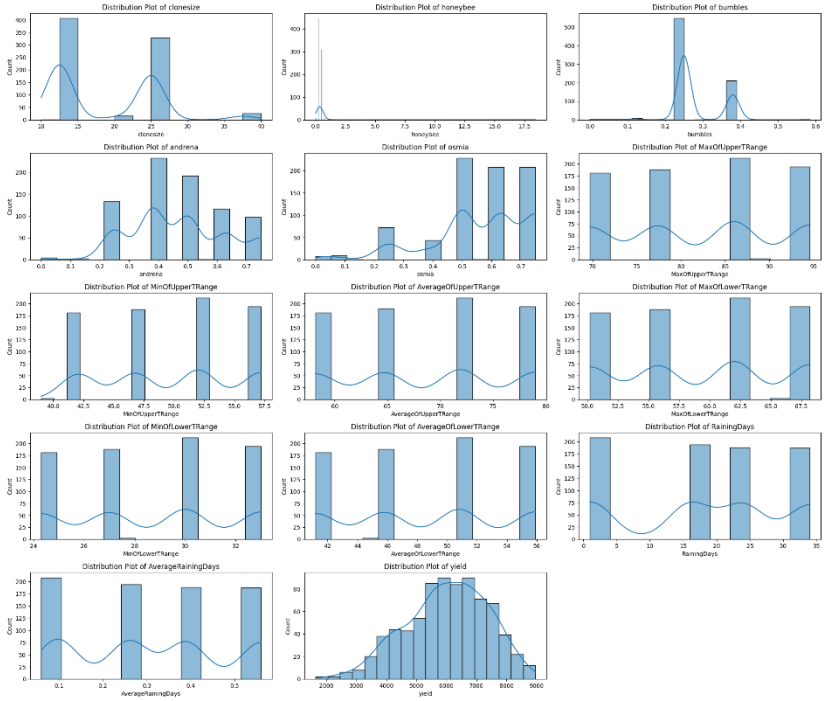


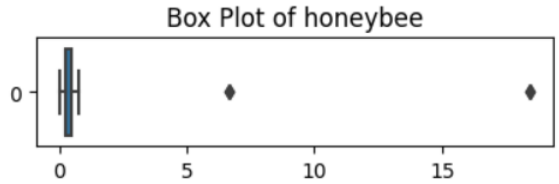
Multivariate Analysis



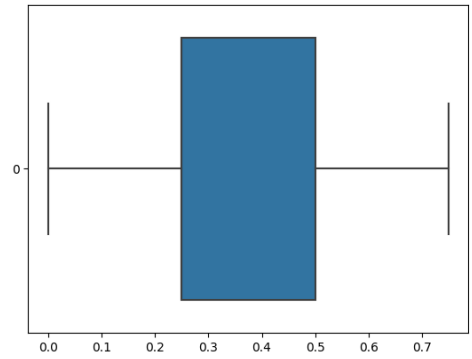
	clonesize	honeybee	bumbles	andrena	osmia	MaxOfUpperTRange	MinOfUpperTRange	AverageOfUpperTRange	MaxOfLowerTRange	MinOfLowerTRange	AverageOfLowerTRange	RainingDays	AverageRainingDays	yield
clonesize	1.000000	0.785162	0.008667	-0.006426	-0.139572	0.033804	0.032496	0.033392	0.033943	0.033412	0.033215	-0.020680	-0.023211	-0.516851
honeybee	0.785162	1.000000	0.119736	0.200733	0.145459	0.002798	-0.006444	-0.001290	0.003744	-0.000513	-0.001786	0.022507	0.028300	-0.307027
bumbles	0.008667	0.119736	1.000000	-0.039157	0.244957	-0.015396	0.002504	-0.007875	-0.017274	-0.009015	-0.006173	0.035665	0.047216	0.318185
andrena	-0.006426	0.200733	-0.039157	1.000000	0.363390	-0.021955	-0.019231	-0.020913	-0.022239	-0.021028	-0.020566	0.021236	0.025980	0.139833
osmia	-0.139572	0.145459	0.244957	0.363390	1.000000	-0.058594	-0.037381	-0.049888	-0.060815	-0.051123	-0.047695	0.065564	0.081402	0.388608
MaxOfUpperTRange	0.033804	0.002798	-0.015396	-0.021955	-0.058594	1.000000	0.990953	0.998313	0.999900	0.998841	0.997656	-0.000773	-0.002508	-0.186848
MinOfUpperTRange	0.032496	-0.006444	0.002504	-0.019231	-0.037381	0.990953	1.000000	0.997053	0.998979	0.998264	0.997796	0.001763	0.001352	-0.175643
AverageOfUpperTRange	0.033392	-0.001290	-0.007875	-0.020913	-0.049888	0.998313	0.997053	1.000000	0.997411	0.998944	0.998924	0.000247	-0.000946	-0.181537
MaxOfLowerTRange	0.033943	0.003744	-0.017274	-0.022239	-0.060815	0.999900	0.998979	0.997411	1.000000	0.998870	0.998606	-0.001047	-0.002922	-0.187213
MinOfLowerTRange	0.033412	-0.000513	-0.009015	-0.021028	-0.051123	0.998841	0.998264	0.998944	0.998870	1.000000	0.998787	0.000133	-0.001132	-0.183107
AverageOfLowerTRange	0.033215	-0.001786	-0.006173	-0.020566	-0.047695	0.997656	0.997796	0.998924	0.998606	0.998787	1.000000	0.000586	-0.000454	-0.181061
RainingDays	-0.020680	0.022507	0.035665	0.021236	0.065564	-0.000773	0.001763	0.000247	-0.001047	0.000133	0.000586	1.000000	0.991579	-0.543522
AverageRainingDays	-0.023211	0.028300	0.047216	0.025980	0.081402	-0.002508	0.001352	-0.000946	-0.002922	-0.001132	-0.000454	0.991579	1.000000	-0.546222
yield	-0.516851	-0.307027	0.318185	0.139833	0.388608	-0.186848	-0.175643	-0.181537	-0.187213	-0.183107	-0.181061	-0.543522	-0.546222	1.000000

Outliers and Anomalies





Outlier in feature 'honeybee' found using boxplot



Handled outlier

Data Preprocessing Code Screenshots

Loading Data

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import dabl
```

```
data=pd.read_csv('WildBlueberryPollinationSimulationData.csv')
```

Handling Missing Data

```
data.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

```
from sklearn.preprocessing import StandardScaler
scale = StandardScaler()
```

```
X_scaled=scale.fit_transform(X)
```

```
X_scaled
```

```
array([[ 2.67234719,  2.91964747, -0.52812593, ...,  0.40517505,
        -0.19702952, -0.35962034],
       [ 2.67234719,  2.91964747, -0.52812593, ...,  0.40517505,
        -1.43645427, -1.29757569],
       [ 2.67234719,  2.91964747, -0.52812593, ...,  1.34521837,
        -0.19702952, -0.35962034],
       ...,
       [ 0.17664981,  1.34753621, -2.61170931, ...,  0.40517505,
         0.46399701,  0.40246839],
       [ 0.17664981,  1.34753621, -2.61170931, ..., -0.60859715,
        -1.20757383, -1.53206453],
       [ 0.17664981,  1.34753621, -2.61170931, ..., -0.60859715,
         0.46399701,  0.40246839]])
```

Feature Engineering

For handling outlier

	`` Q1 = data['honeybee'].quantile(0.25) Q3 = data['honeybee'].quantile(0.75) IQR=Q3-Q1 lower_limit = Q1 - 1.5 * IQR upper_limit = Q3 + 1.5 * IQR print('lower_limit: ',lower_limit) print('upper_limit: ',upper_limit) ``` ``` data = data[(data.honeybee>lower_limit)&(data.honeybee<upper_limit)] ```
Save Processed Data	``` X=pd.DataFrame(X_scaled, columns=names) ``` ``` X ``` **Saving the scaler** ``` with open('standard_scaler.pkl', 'wb') as file: pickle.dump(scale, file) ```