Data Collection and Preprocessing Phase

Date	06 July 2024
Team ID	739915
Project Title	BlueBerry Yield Prediction
Maximum Marks	6 Marks

Data Exploration and Preprocessing Report

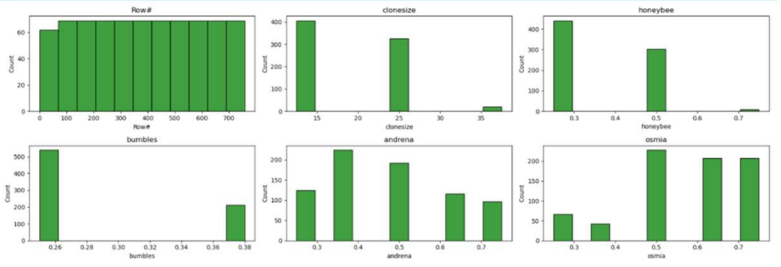
Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

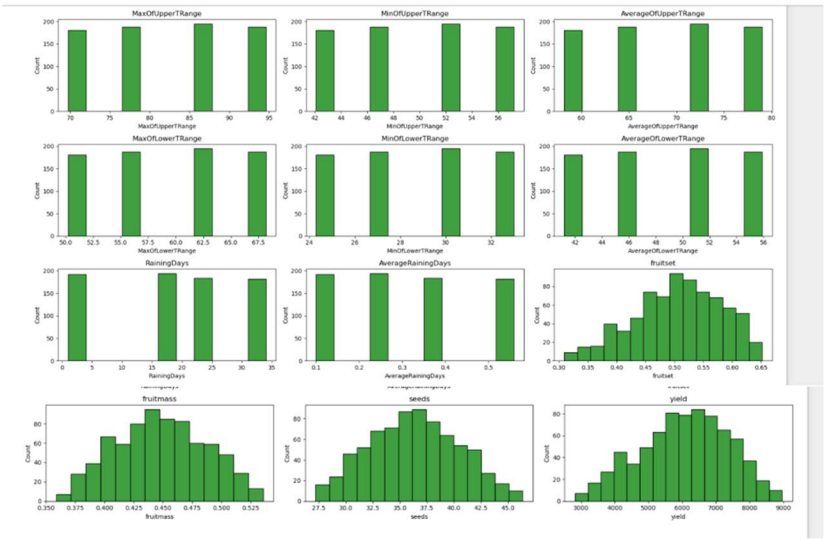
Section	Description
Data Overview	`[14]: p_d.describe()`

Univariate Analysis	`[15]: plt.figure(figsizes(18,18))`
	`for i,col in enumerate(data.columns):`
	`plt.subplot(6,3,i+1)`
	`sns.histplot(p_d[col],color='green')`
	`plt.xlabel(col)`
	`plt.title(col)`
	`plt.tight_layout()`

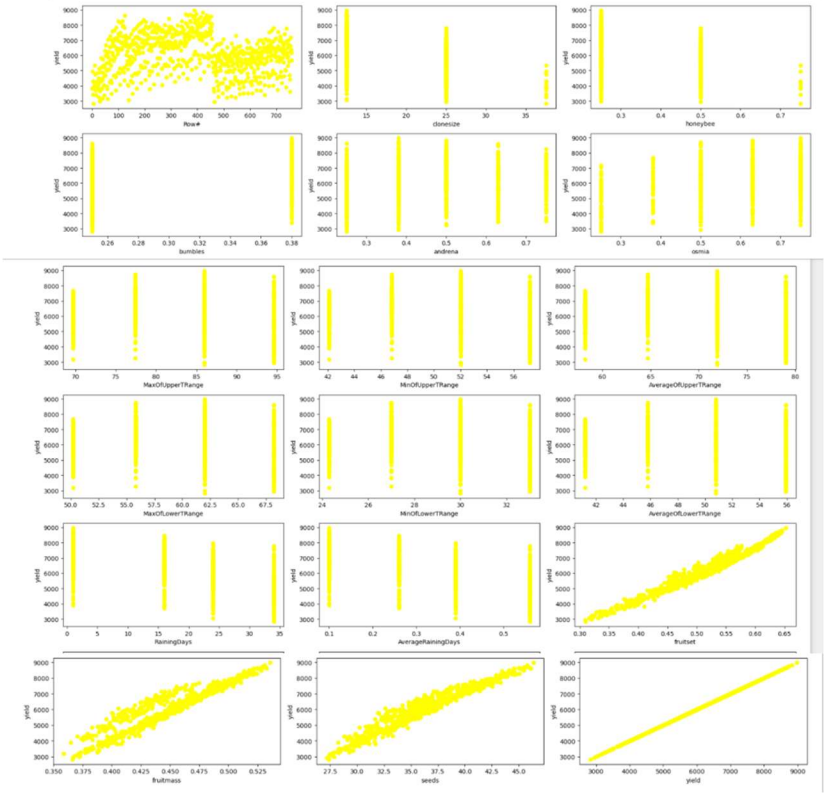
[14]:

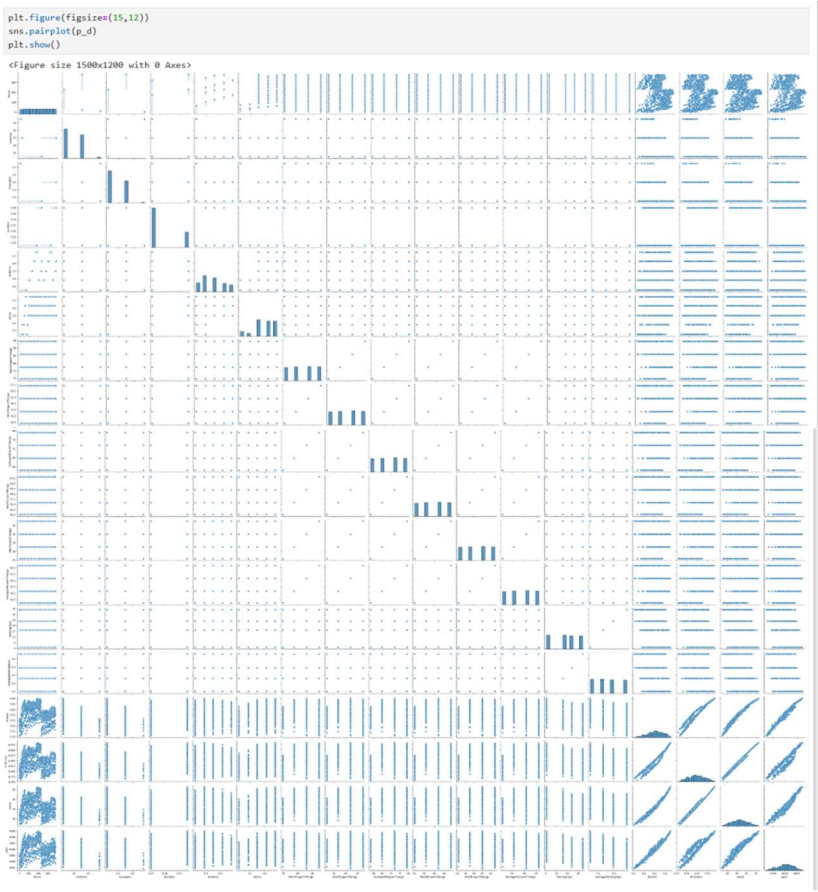
	Row#	clonesize	honeybee	bumbles	andrena	osmia	MaxOfUpperTRange	MinOfUpperTRange	AverageOfUpperTRange	MaxOfLowerTRange
count	752.000000	752.000000	752.000000	752.000000	752.000000	752.000000	752.000000	752.000000	752.000000	752.000000
mean	382.337766	18.583777	0.356383	0.286649	0.475000	0.576463	82.076729	49.617154	68.577527	59.159840
std	217.501250	6.885425	0.129602	0.058530	0.156807	0.149782	9.254791	5.610176	7.731659	6.687814
min	0.000000	12.500000	0.250000	0.250000	0.250000	0.250000	69.700000	42.100000	58.200000	50.200000
25%	194.750000	12.500000	0.250000	0.250000	0.380000	0.500000	77.400000	46.800000	64.700000	55.800000
50%	382.500000	12.500000	0.250000	0.250000	0.500000	0.630000	86.000000	52.000000	71.900000	62.000000
75%	570.250000	25.000000	0.500000	0.380000	0.630000	0.750000	88.150000	53.300000	73.675000	63.550000
max	758.000000	37.500000	0.750000	0.380000	0.750000	0.750000	94.600000	57.200000	79.000000	68.200000



<table>
<tr>
<td></td>
<td> 
</td>
</tr>
<tr>
<td>Bivariate Analysis</td>
<td>Bivariate analysis (scatter plot)
[14]:

```
plt.figure(figsize=(18,18))
for i,col in enumerate(data.columns):
    plt.subplot(6,3,i+1)
    plt.scatter(x=p_d[col],y=p_d['yield'],color='yellow')
    plt.xlabel(col)
    plt.ylabel('yield')
plt.tight_layout()
```


</td>
</tr>
</table>

Multivariate Analysis	

```
plt.figure(figsizes(15,12))
sns.pairplot(p_d)
plt.show()
```

```
<Figure size 1500x1200 with 0 Axes>
```

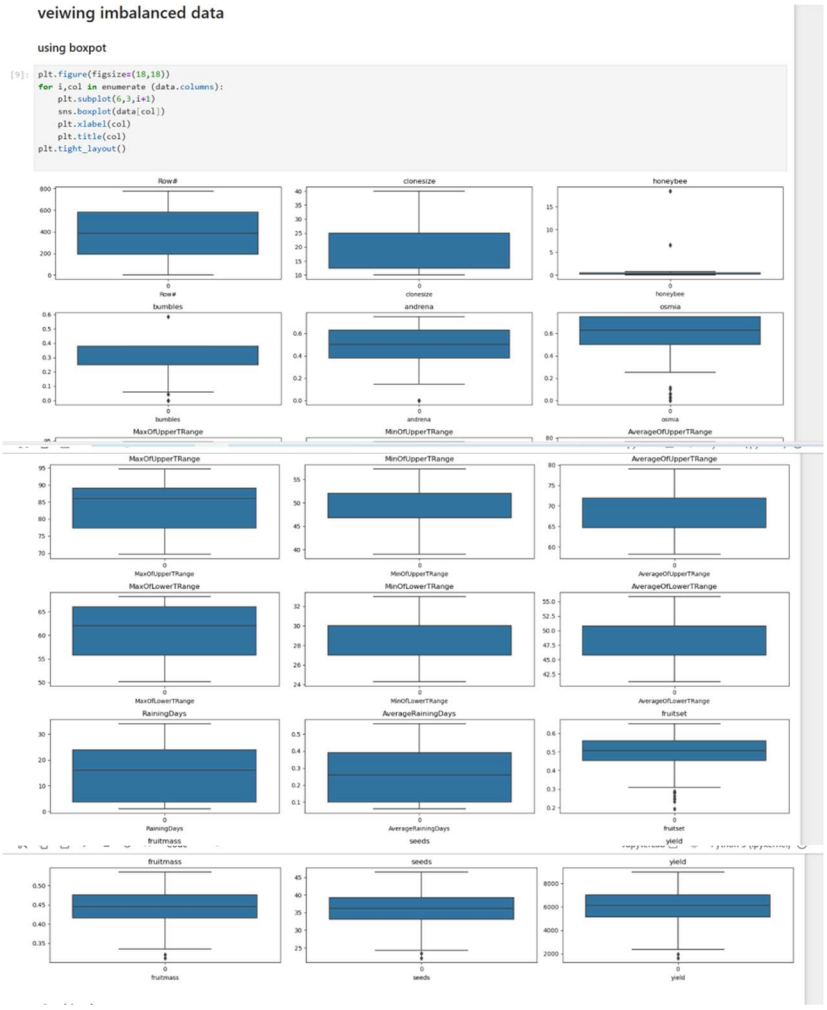
Data Preprocessing Code Screenshots

Loading Data

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

	Row#	clonesize	honeybee	bumbles	andrena	osmia	MaxOfUpperTRange	MinOfUpperTRange	AverageOfUpperTRange	MaxOfLowerTRange	MinOfLowerTRange
0	0	37.5	0.750	0.250	0.250	0.250	86.0	52.0	71.9	62.0	30.0
1	1	37.5	0.750	0.250	0.250	0.250	86.0	52.0	71.9	62.0	30.0
2	2	37.5	0.750	0.250	0.250	0.250	94.6	57.2	79.0	68.2	33.0

Handling Null Values	**Handling null values** [8]: `data.isnull().sum()` [8]: Row# 0 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

Viewing outliers	**veiwing imbalanced data**
	using boxpot
	[9]:
	`plt.figure(figsize=(18,18))`
	`for i,col in enumerate (data.columns):`
	`  plt.subplot(6,3,i+1)`
	`  sns.boxplot(data[col])`
	`  plt.xlabel(col)`
	`  plt.title(col)`
	`plt.tight_layout()`
	 

Handling outliers	**handling imbalance data** **by removing outliers** [223]: ``` x=data q1=x.quantile(0.25) q3=x.quantile(0.75) iqr=q3-q1 iqr ``` [223]: ``` Row# 388.000000 clonesize 12.500000 honeybee 0.250000 bumbles 0.130000 andrena 0.250000 osmia 0.250000 MaxOfUpperTRange 11.600000 MinOfUpperTRange 5.200000 AverageOfUpperTRange 7.200000 MaxOfLowerTRange 10.200000 MinOfLowerTRange 3.000000 AverageOfLowerTRange 5.000000 RainingDays 20.230000 AverageRainingDays 0.290000 fruitset 0.106571 fruitmass 0.059869 seeds 6.123577 yield 1897.334830 dtype: float64 ```
Saved Processed Data	``` p_d=data[~((data<(q1-1.5*iqr)) | (data>(q3+1.5*iqr))).any(axis=1)] p_d.shape ``` ``` (752, 18) ```