Data Collection and Preprocessing Phase

Date	15 July 2024
Team ID	740662
Project Title	Golden Harvest: A Predictive Model for Apple Quality Assurance
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

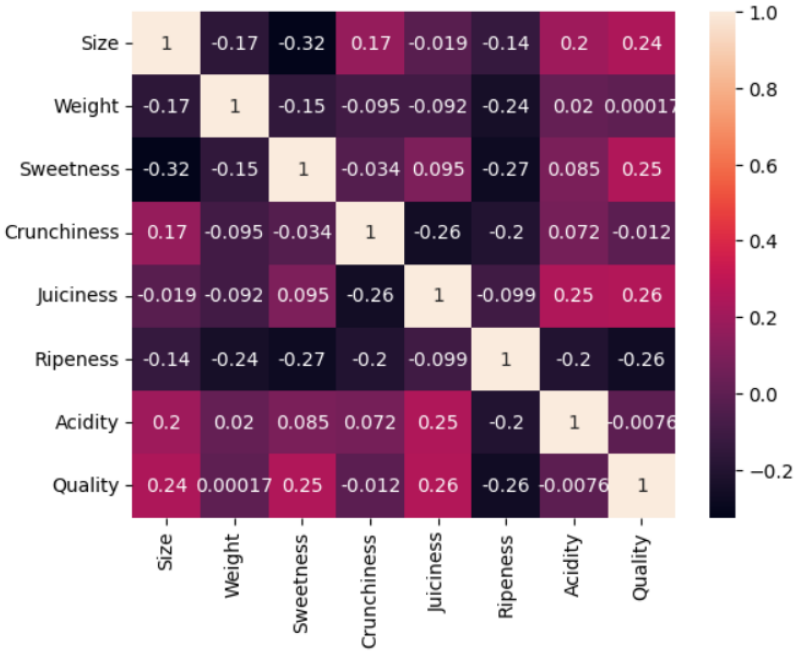
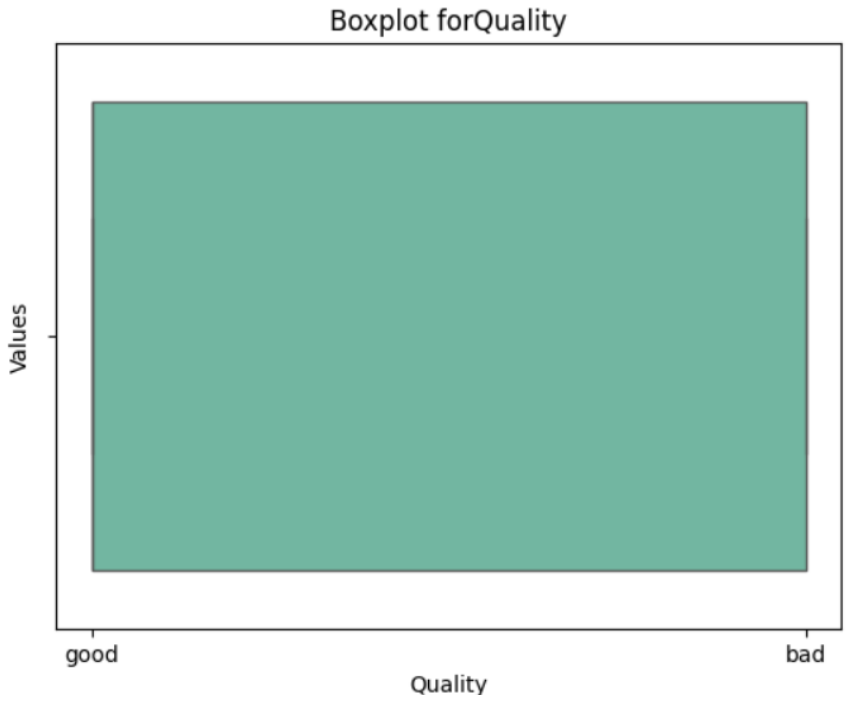
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

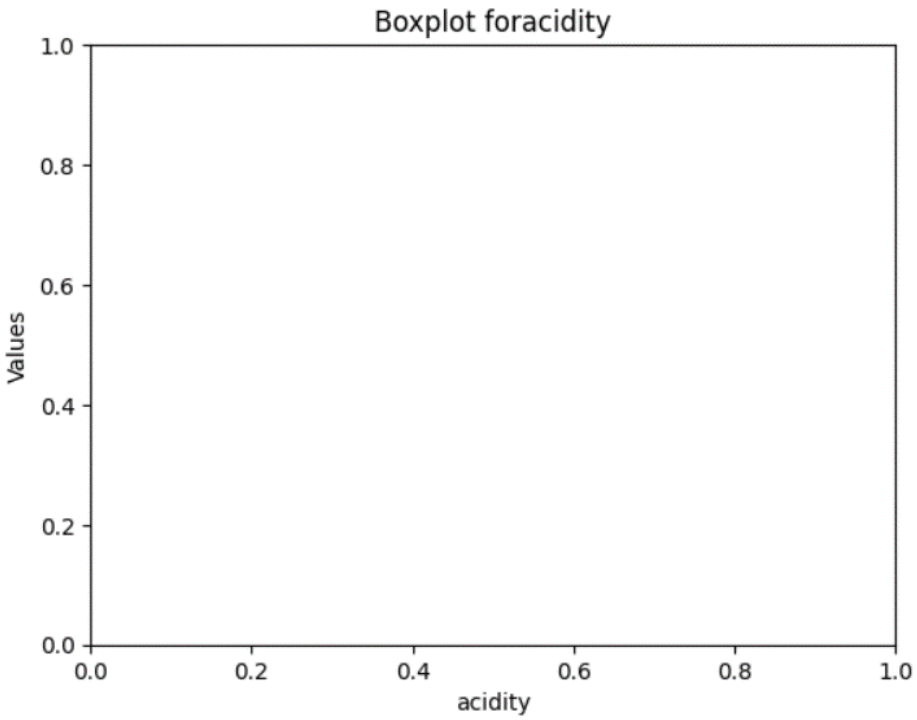
Section	Description

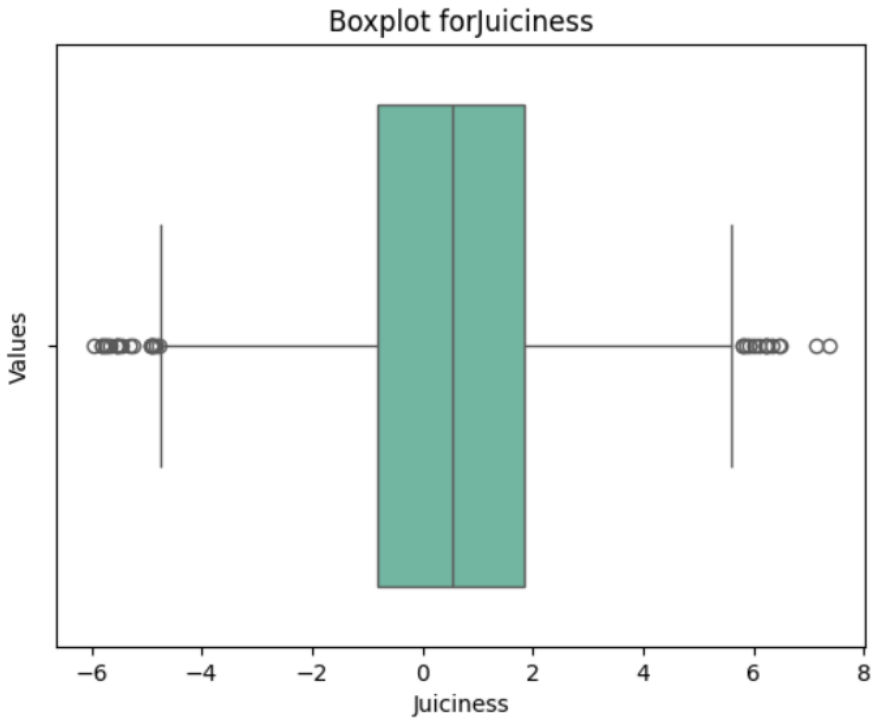
Data Overview	**Descriptive Statistics**
	`data.describe()`
	`data.shape`
	`(4001, 9)`

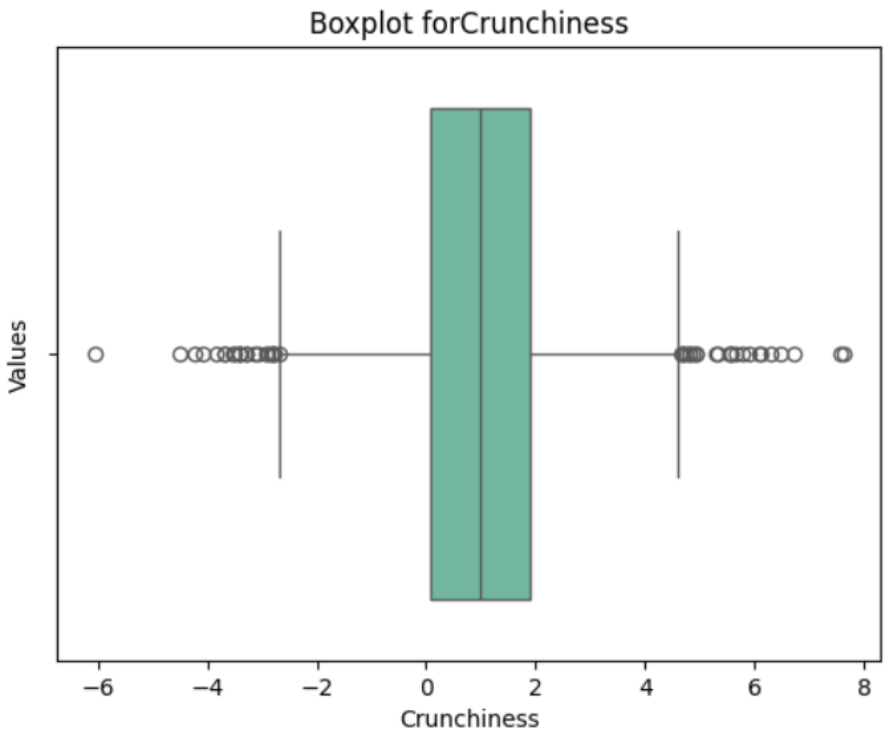
Univariate Analysis	-
Bivariate Analysis	-

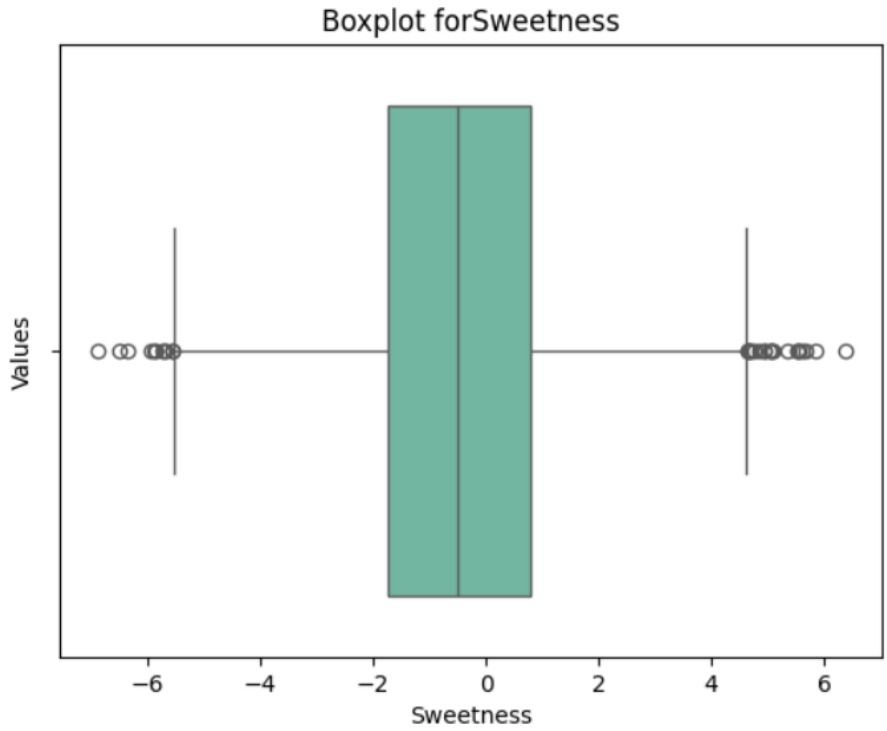
	A_id	Size	Weight	Sweetness	Crunchiness	Juiciness	Ripeness	Acidity
count	4000.000000	4000.000000	4000.000000	4000.000000	4000.000000	4000.000000	4000.000000	4000.000000
mean	1999.500000	-0.502695	-0.991229	-0.472248	0.984194	0.513127	0.498102	0.076639
std	1154.844867	1.917446	1.574517	1.931684	1.369437	1.917024	1.866614	2.101441
min	0.000000	-5.750201	-5.075890	-5.548946	-2.684440	-4.757179	-4.578510	-5.709299
25%	999.750000	-1.816765	-2.011770	-1.738425	0.062764	-0.801286	-0.771677	-1.377424
50%	1999.500000	-0.513703	-0.984736	-0.504758	0.998249	0.534219	0.503445	0.022609
75%	2999.250000	0.805526	0.030976	0.801922	1.894234	1.835976	1.766212	1.510493
max	3999.000000	4.738963	3.095097	4.612442	4.641439	5.791870	5.573044	5.842368

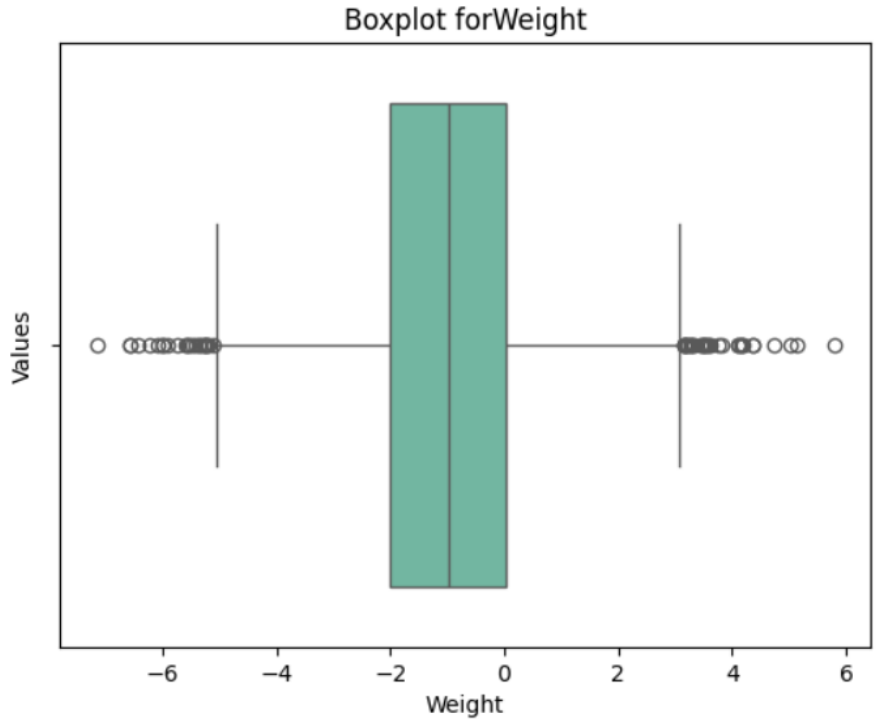
Multivariate Analysis	
Outliers and Anomalies	

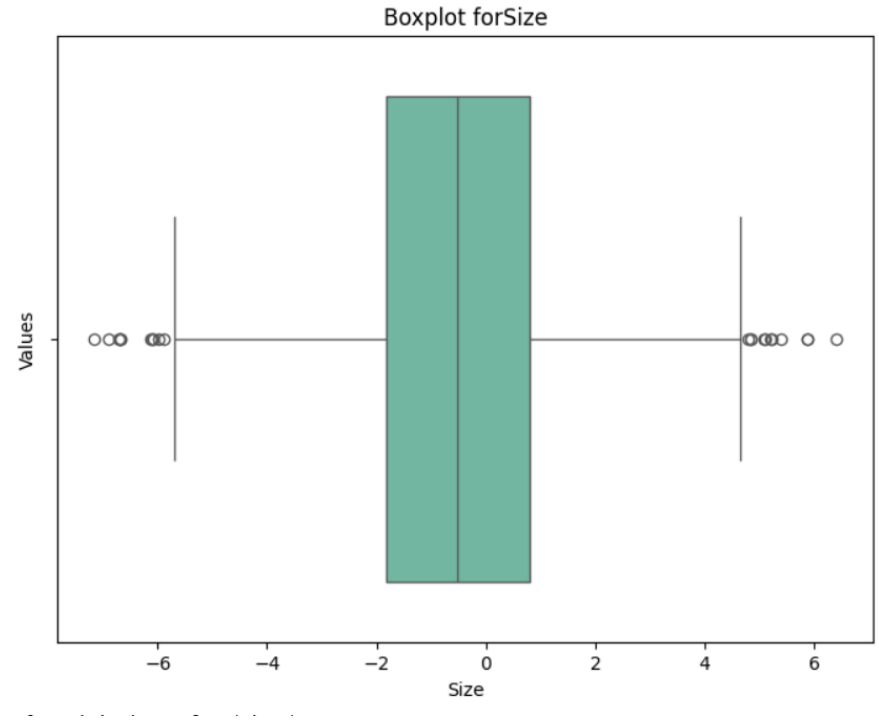

Boxplot foracidity
1.0
0.8
0.6
0.4
0.2
0.0
Values
0.0
0.2
0.4
0.6
0.8
1.0
acidity

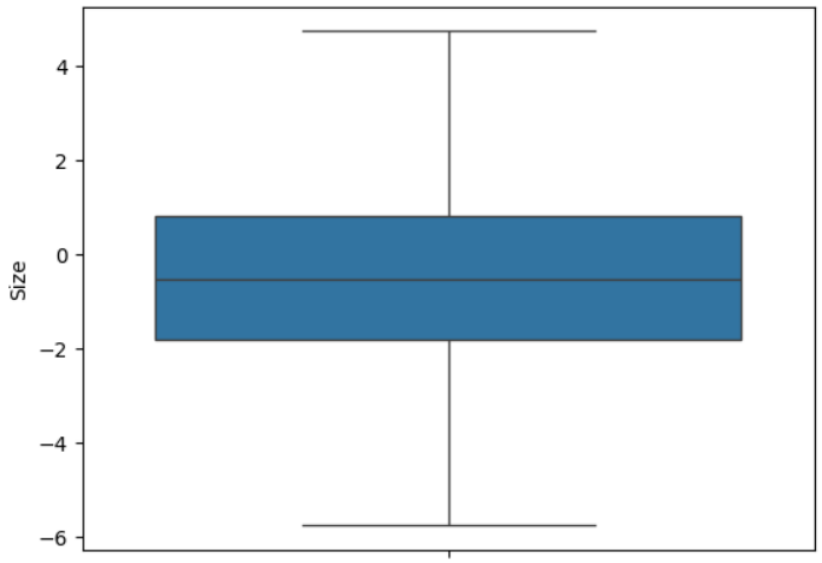

Boxplot forJuiciness
Values
−6
−4
−2
0
2
4
6
8
Juiciness

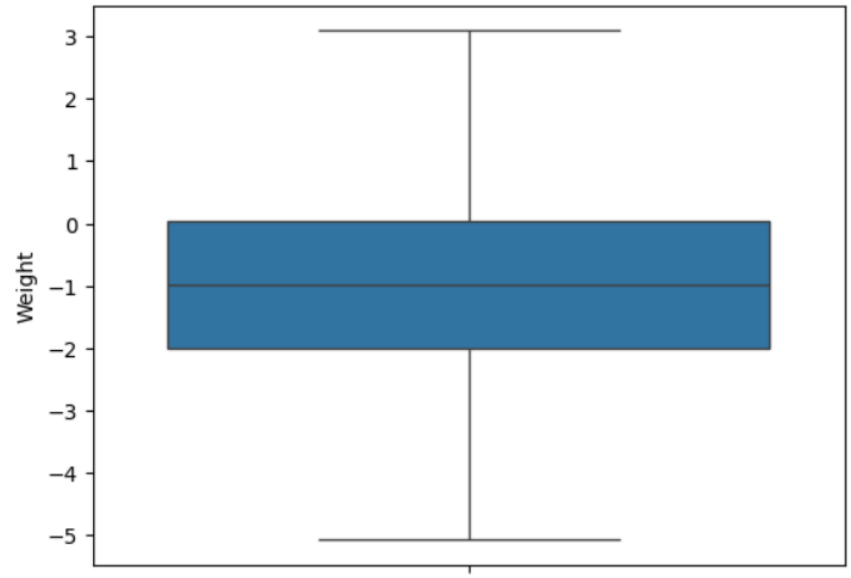

Boxplot forCrunchiness
Values
−6
−4
−2
0
2
4
6
8
Crunchiness

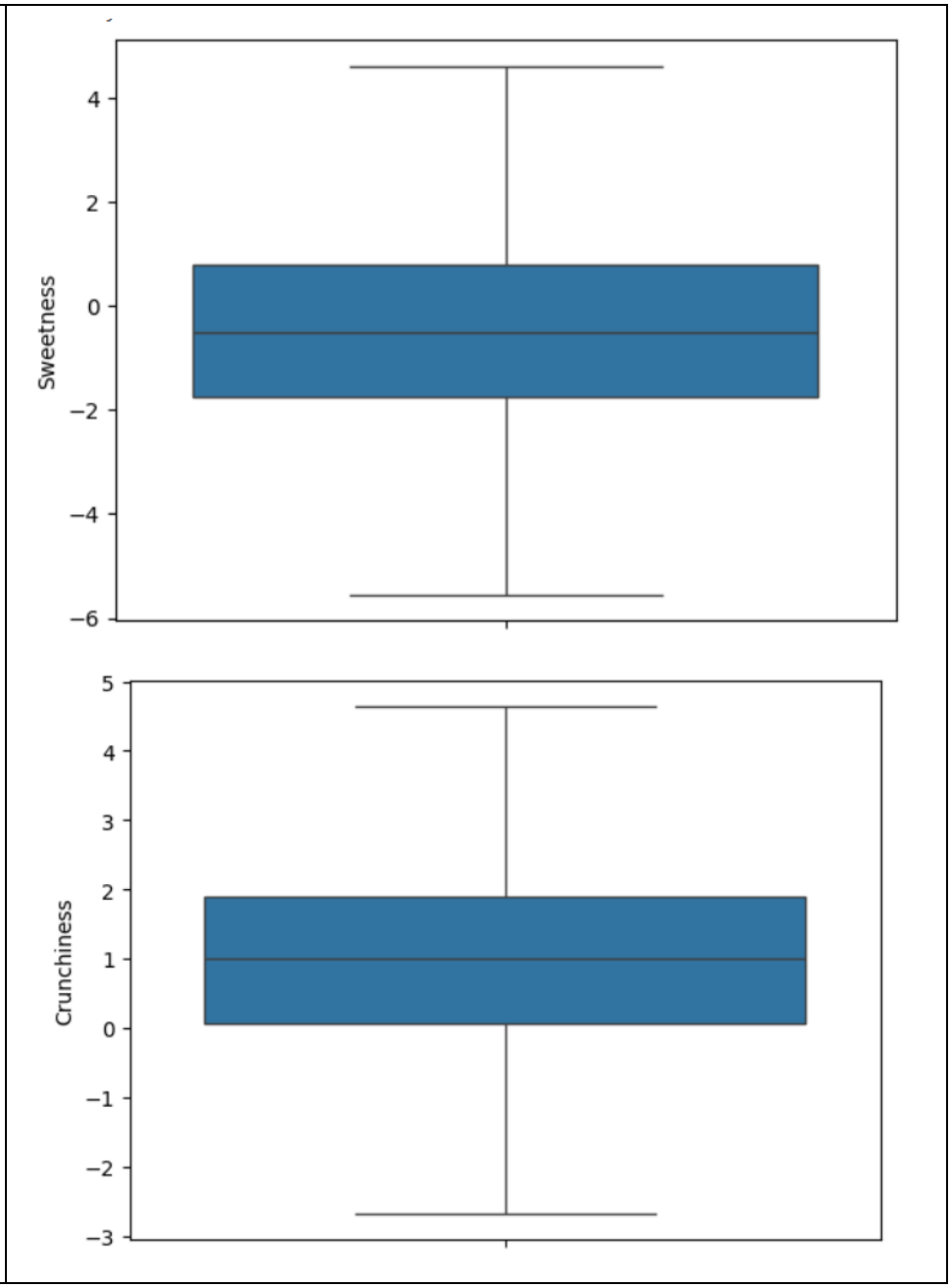

Boxplot forSweetness
Values
−6
−4
−2
0
2
4
6
Sweetness

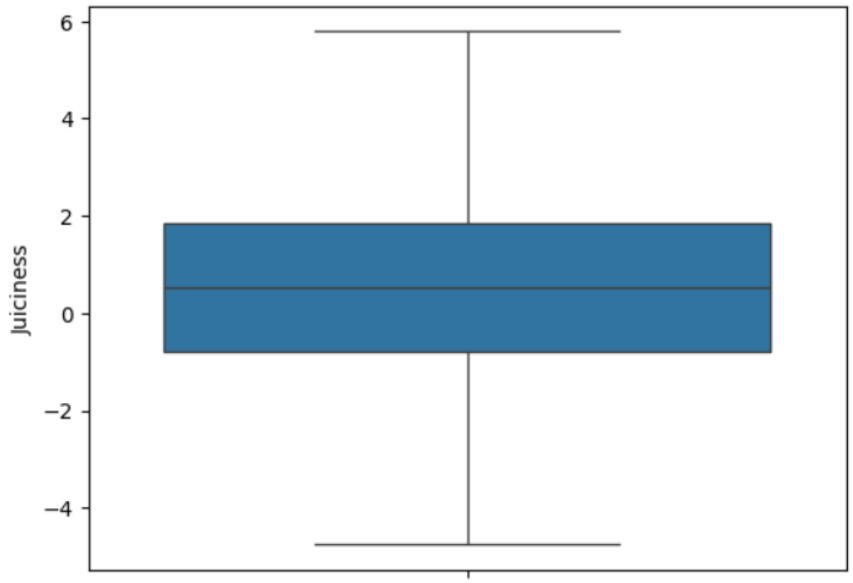

Boxplot forWeight
Values
−6
−4
−2
0
2
4
6
Weight

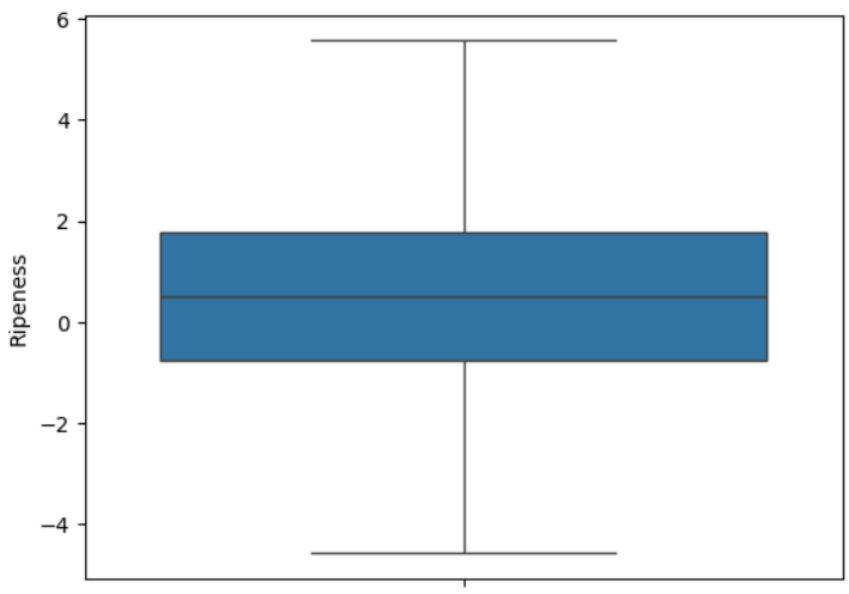

Boxplot forSize
Values
−6
−4
−2
0
2
4
6
Size

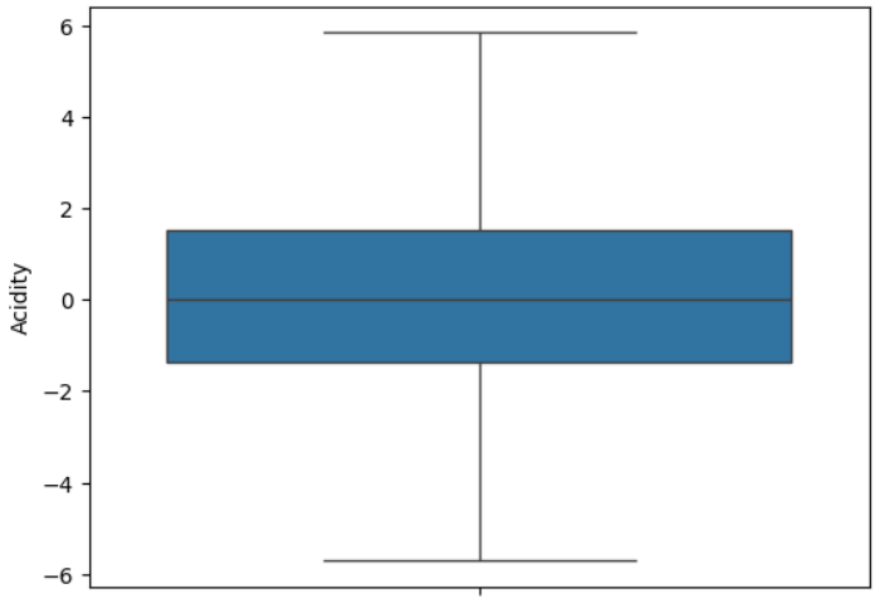

Size
4
2
0
−2
−4
−6


Weight
3
2
1
0
−1
−2
−3
−4
−5


Sweetness
4
2
0
−2
−4
−6
Crunchiness
5
4
3
2
1
0
−1
−2
−3


6
4
2
0
−2
−4
Juiciness


6
4
2
0
−2
−4
Ripeness



Data Preprocessing Code Screenshots

Loading Data

```
[ ] data=pd.read_csv('/content/apple_quality.csv')
```

```
data.head()
```

	A_id	Size	Weight	Sweetness	Crunchiness	Juiciness	Ripeness	Acidity	Quality
0	0.0	-3.970049	-2.512336	5.346330	-1.012009	1.844900	0.329840	-0.491590483	good
1	1.0	-1.195217	-2.839257	3.664059	1.588232	0.853286	0.867530	-0.722809367	good
2	2.0	-0.292024	-1.351282	-1.738429	-0.342616	2.838636	-0.038033	2.621636473	bad
3	3.0	-0.657196	-2.271627	1.324874	-0.097875	3.637970	-3.413761	0.790723217	good
4	4.0	1.364217	-1.296612	-0.384658	-0.553006	3.030874	-1.303849	0.501984036	good

Handling Missing Data

```
[ ] data.shape
```

```
(4001, 9)
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4001 entries, 0 to 4000
Data columns (total 9 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   A_id         4000 non-null   float64
 1   Size         4000 non-null   float64
 2   Weight       4000 non-null   float64
 3   Sweetness    4000 non-null   float64
 4   Crunchiness  4000 non-null   float64
 5   Juiciness    4000 non-null   float64
 6   Ripeness     4000 non-null   float64
 7   Acidity      4001 non-null   object
 8   Quality      4000 non-null   object
dtypes: float64(7), object(2)
memory usage: 281.4+ KB
```

	[ ] data.isnull().sum() A_id 1 Size 1 Weight 1 Sweetness 1 Crunchiness 1 Juiciness 1 Ripeness 1 Acidity 0 Quality 1 dtype: int64 [ ] data.dropna(inplace=True)
Data Transformation	-
Feature Engineering	Code for creating new features or modifying existing ones.
Save Processed Data	[ ] import pickle pickle.dump(model1,open("model.pkl","wb"))