

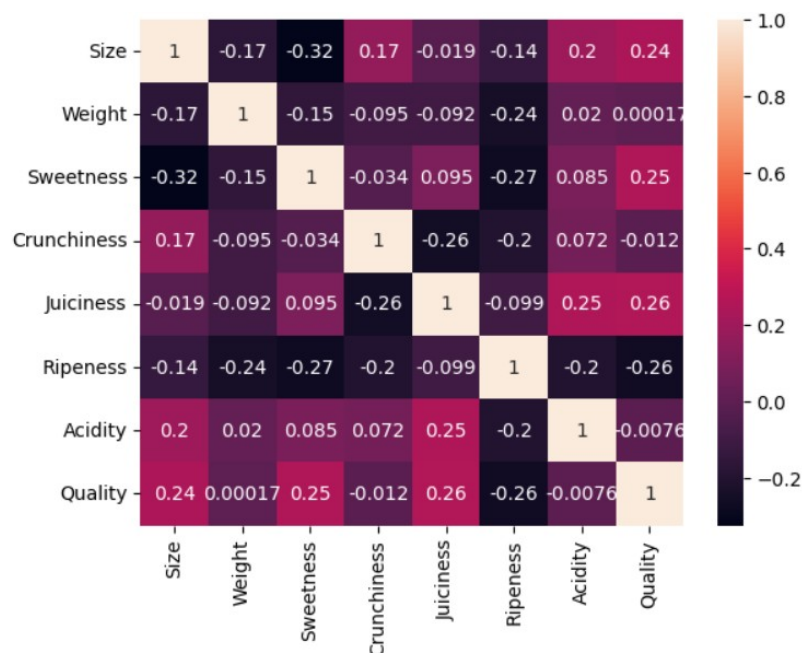
## Data Collection and Preprocessing Phase

Date	15 July 2024
Team ID	739885
Project Title	Golden Harvest: A Predictive Model for Apple Quality Assurance
Maximum Marks	6 Marks

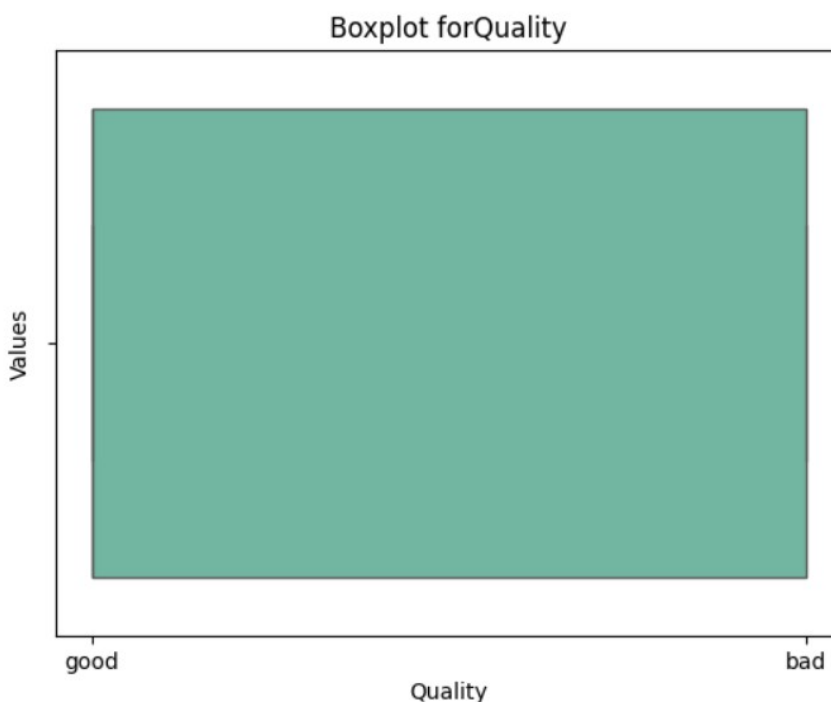
### Data Exploration and Preprocessing Template

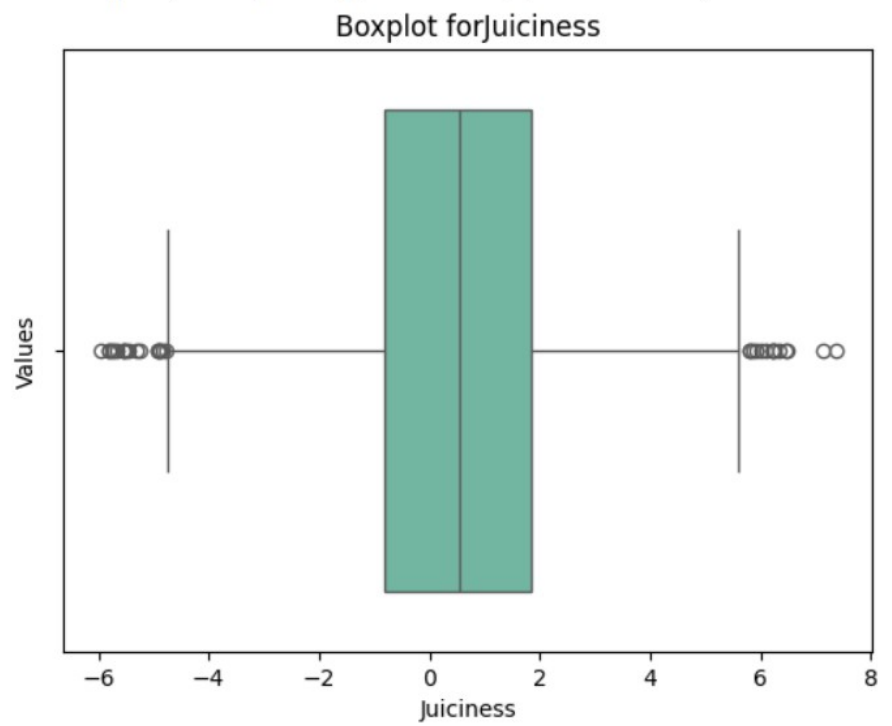
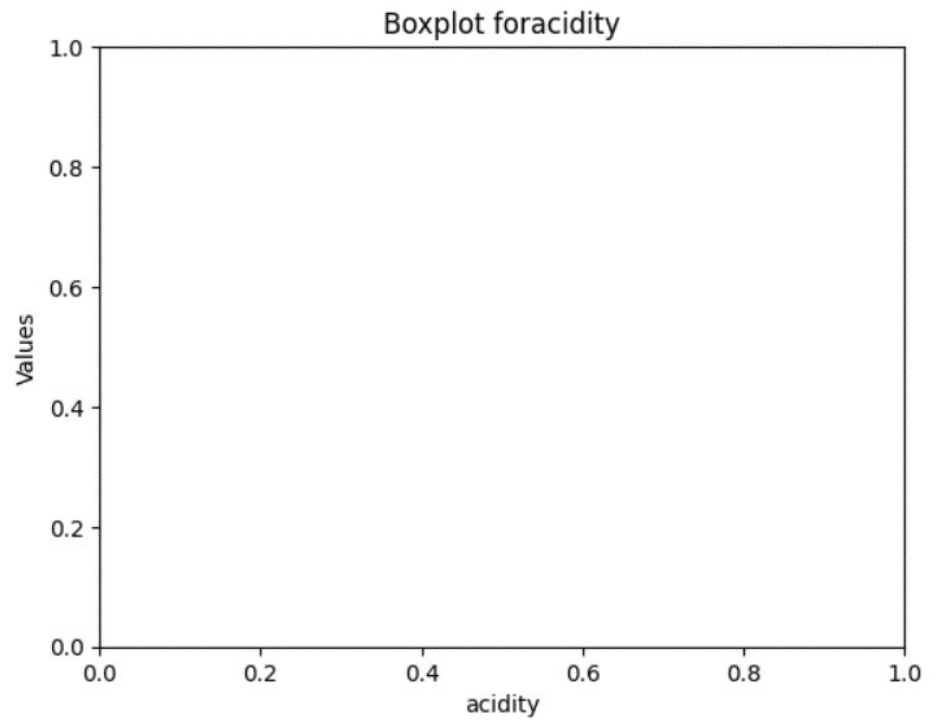
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

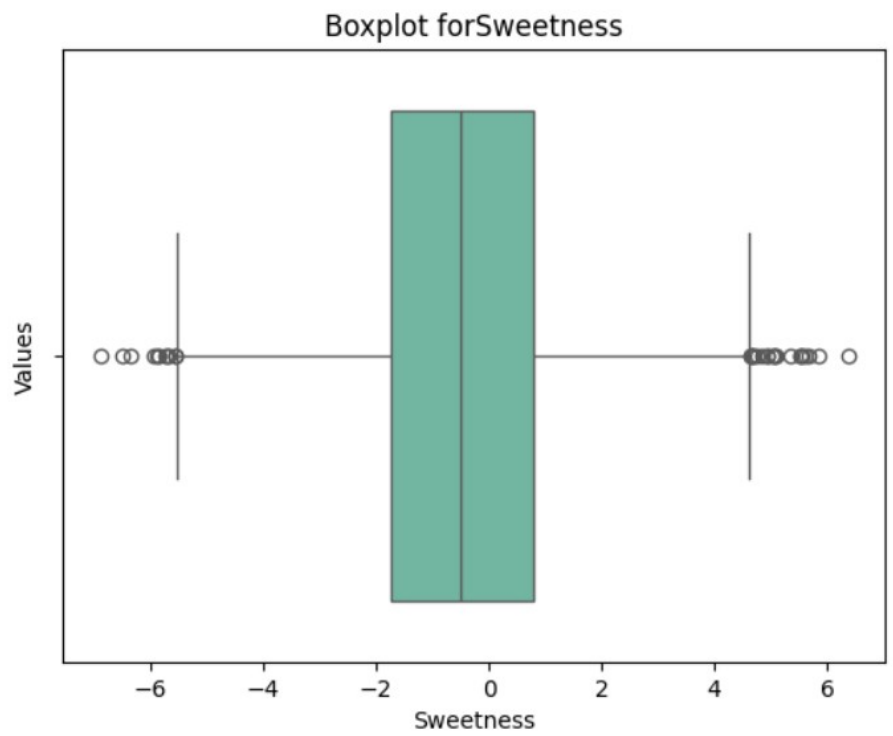
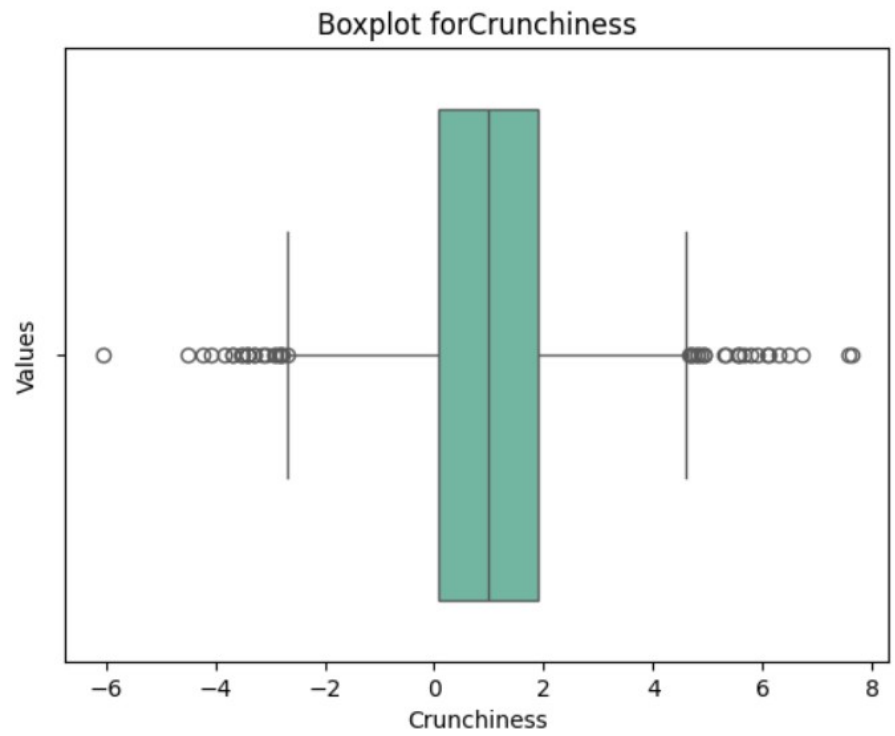
Section	Description																																																																																	
Data Overview	<div>Descriptive Statistics</div> <div><div><div>▶</div><div>data.describe()</div></div><div><div><div>↕</div><div></div></div><table><thead><tr><th></th><th>A_id</th><th>Size</th><th>Weight</th><th>Sweetness</th><th>Crunchiness</th><th>Juiciness</th><th>Ripeness</th><th>Acidity</th></tr></thead><tbody><tr><td>count</td><td>4000.000000</td><td>4000.000000</td><td>4000.000000</td><td>4000.000000</td><td>4000.000000</td><td>4000.000000</td><td>4000.000000</td><td>4000.000000</td></tr><tr><td>mean</td><td>1999.500000</td><td>-0.502695</td><td>-0.991229</td><td>-0.472248</td><td>0.984194</td><td>0.513127</td><td>0.498102</td><td>0.076639</td></tr><tr><td>std</td><td>1154.844867</td><td>1.917446</td><td>1.574517</td><td>1.931684</td><td>1.369437</td><td>1.917024</td><td>1.866614</td><td>2.101441</td></tr><tr><td>min</td><td>0.000000</td><td>-5.750201</td><td>-5.075890</td><td>-5.548946</td><td>-2.684440</td><td>-4.757179</td><td>-4.578510</td><td>-5.709299</td></tr><tr><td>25%</td><td>999.750000</td><td>-1.816765</td><td>-2.011770</td><td>-1.738425</td><td>0.062764</td><td>-0.801286</td><td>-0.771677</td><td>-1.377424</td></tr><tr><td>50%</td><td>1999.500000</td><td>-0.513703</td><td>-0.984736</td><td>-0.504758</td><td>0.998249</td><td>0.534219</td><td>0.503445</td><td>0.022609</td></tr><tr><td>75%</td><td>2999.250000</td><td>0.805526</td><td>0.030976</td><td>0.801922</td><td>1.894234</td><td>1.835976</td><td>1.766212</td><td>1.510493</td></tr><tr><td>max</td><td>3999.000000</td><td>4.738963</td><td>3.095097</td><td>4.612442</td><td>4.641439</td><td>5.791870</td><td>5.573044</td><td>5.842368</td></tr></tbody></table></div></div> <div><div>{x}</div><div><div>▶</div><div>data.shape</div></div><div><div>🔑</div><div><div>↕</div><div>(4001, 9)</div></div></div></div>		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
		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																																																																									
Univariate Analysis	-																																																																																	
Bivariate Analysis	-																																																																																	
Multivariate Analysis																																																																																		

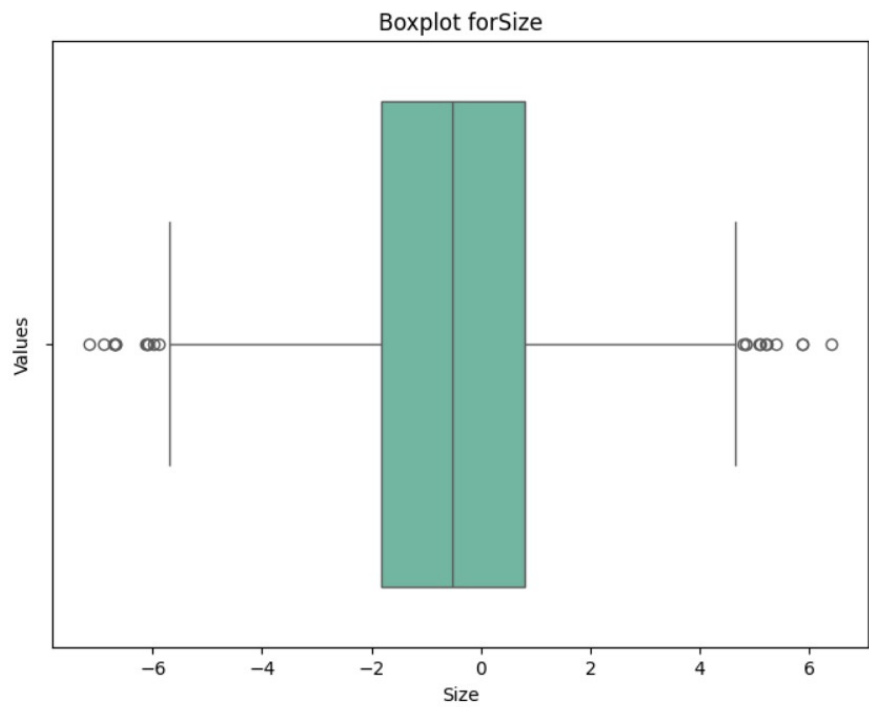
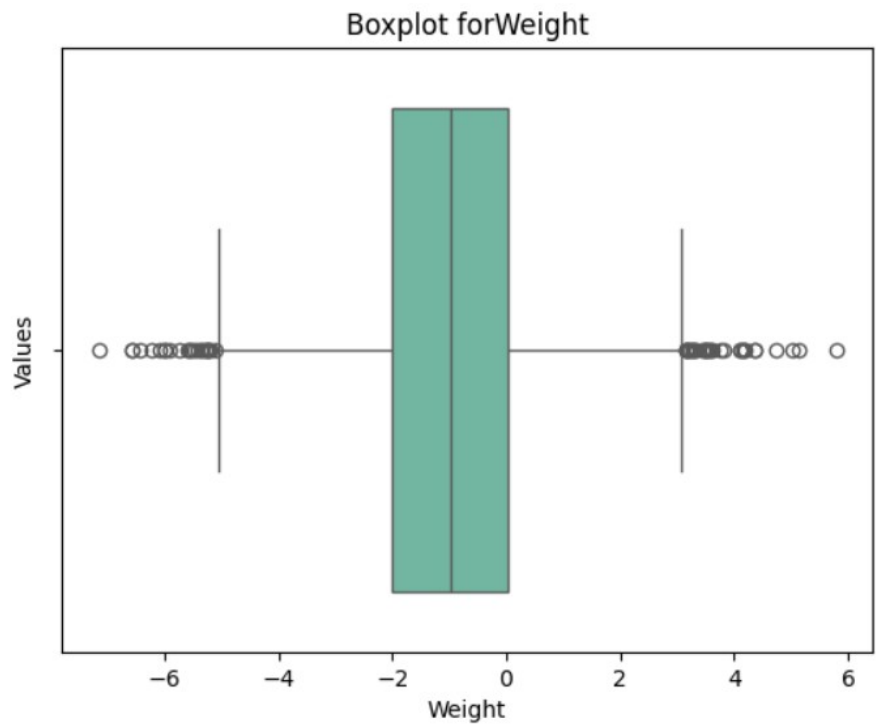


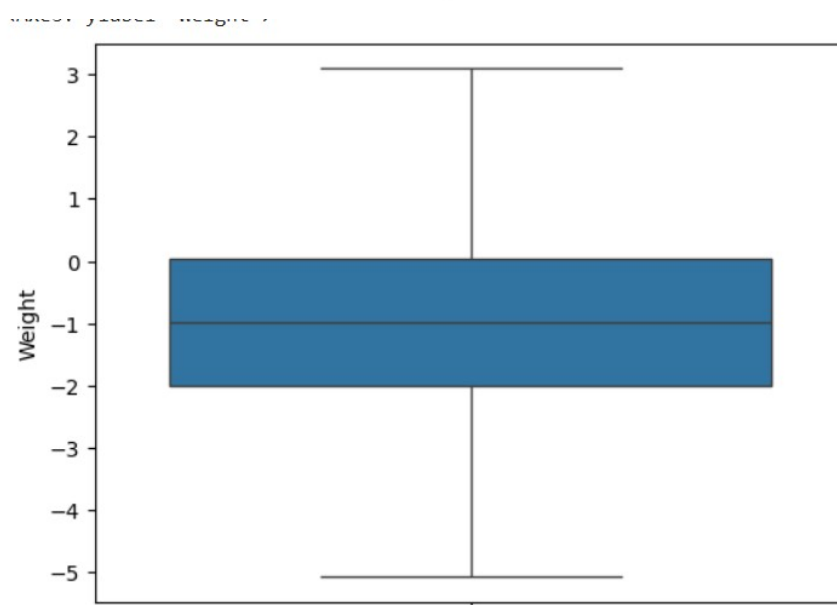
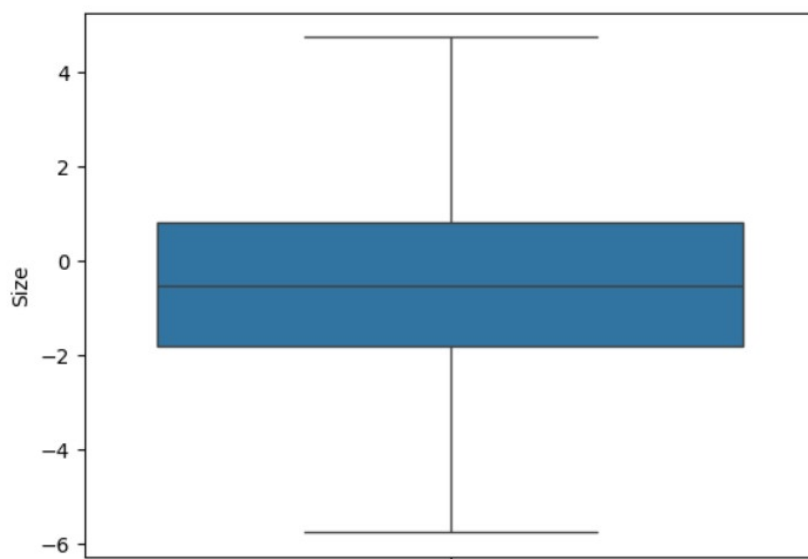
## Outliers and Anomalies

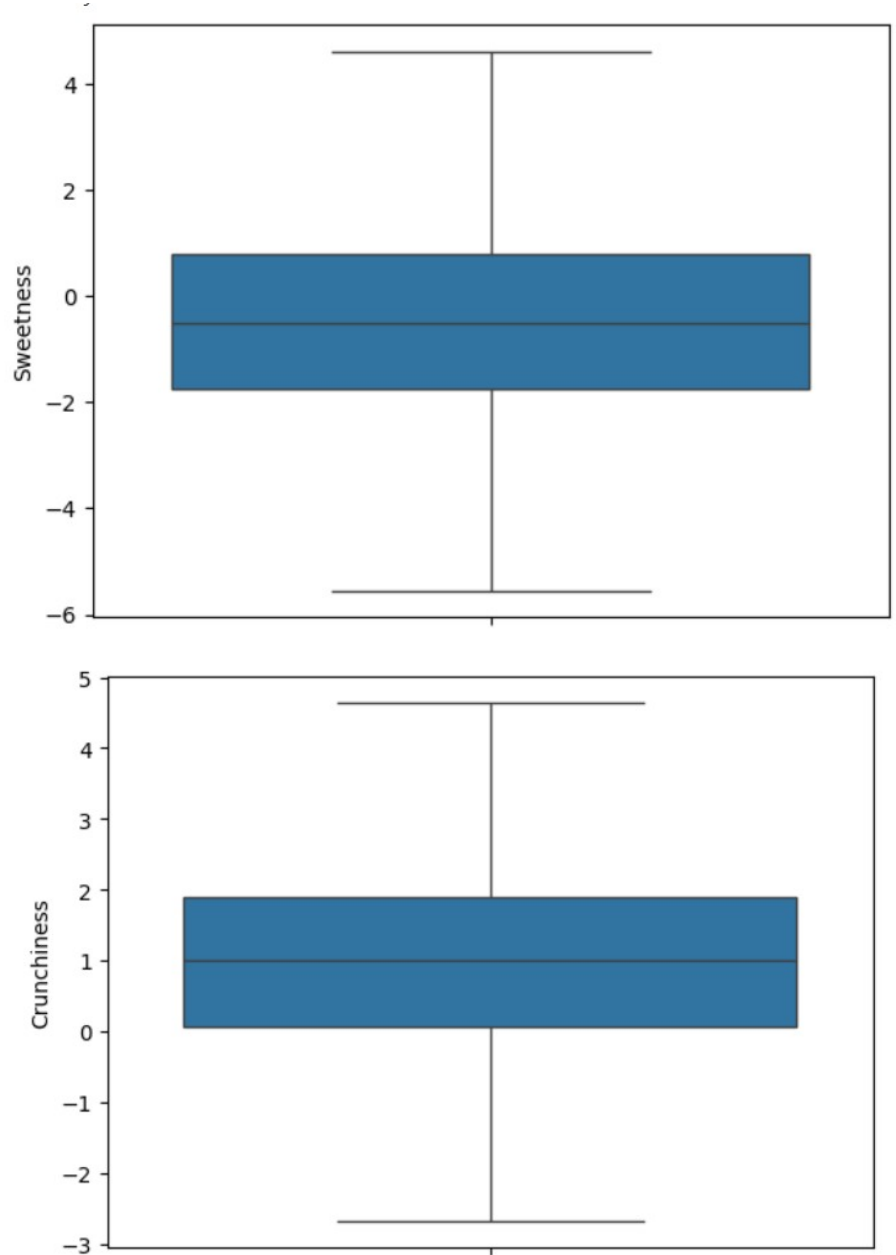


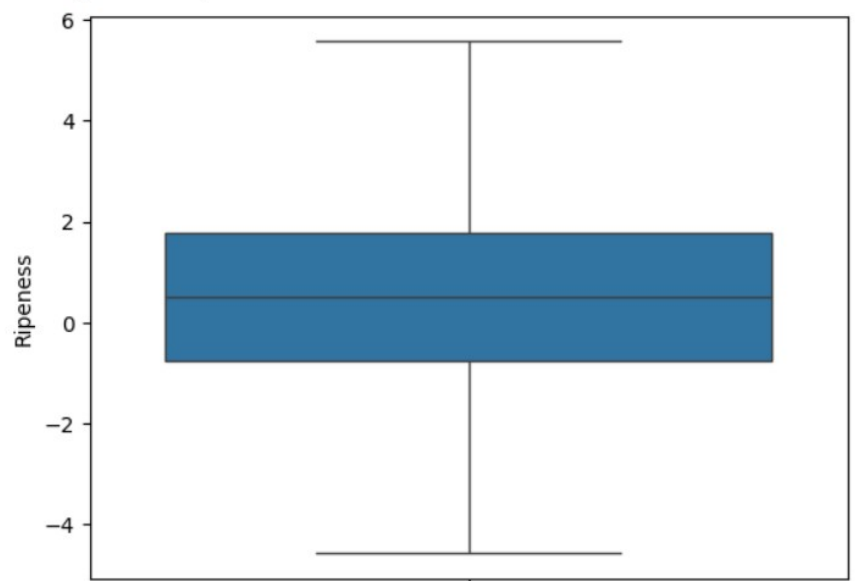
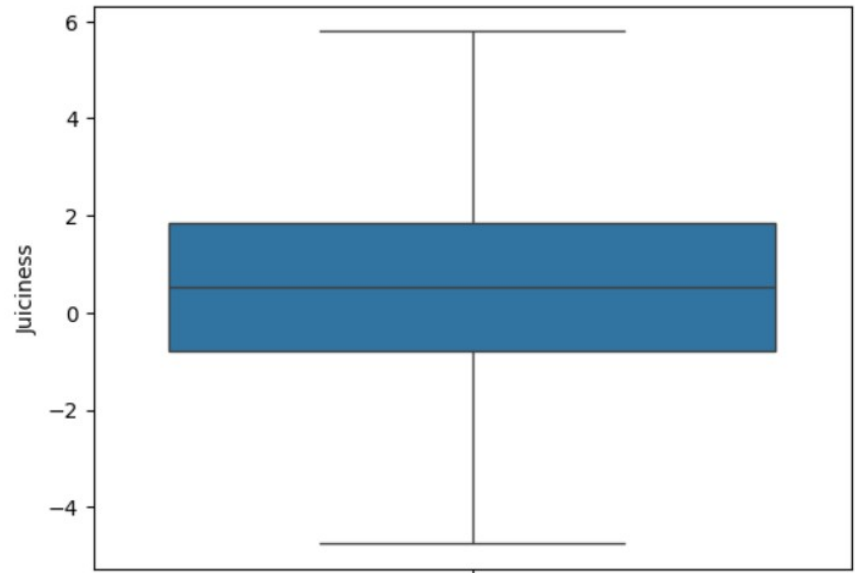




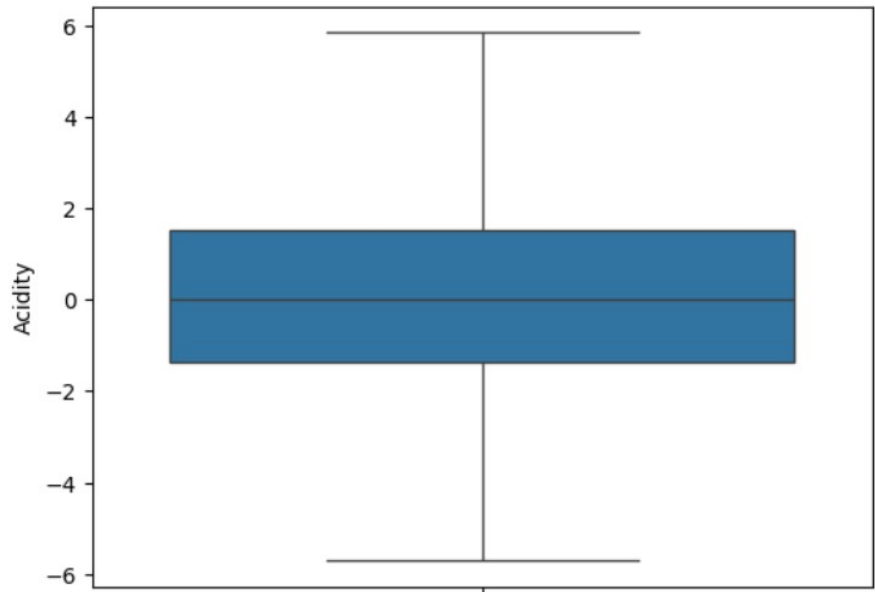












## Data Preprocessing Code Screenshots

### 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
```

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