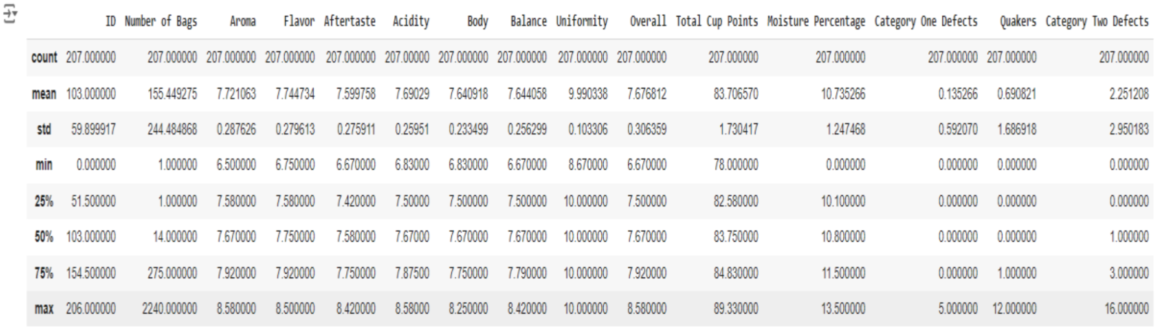
Data Collection and Preprocessing Phase

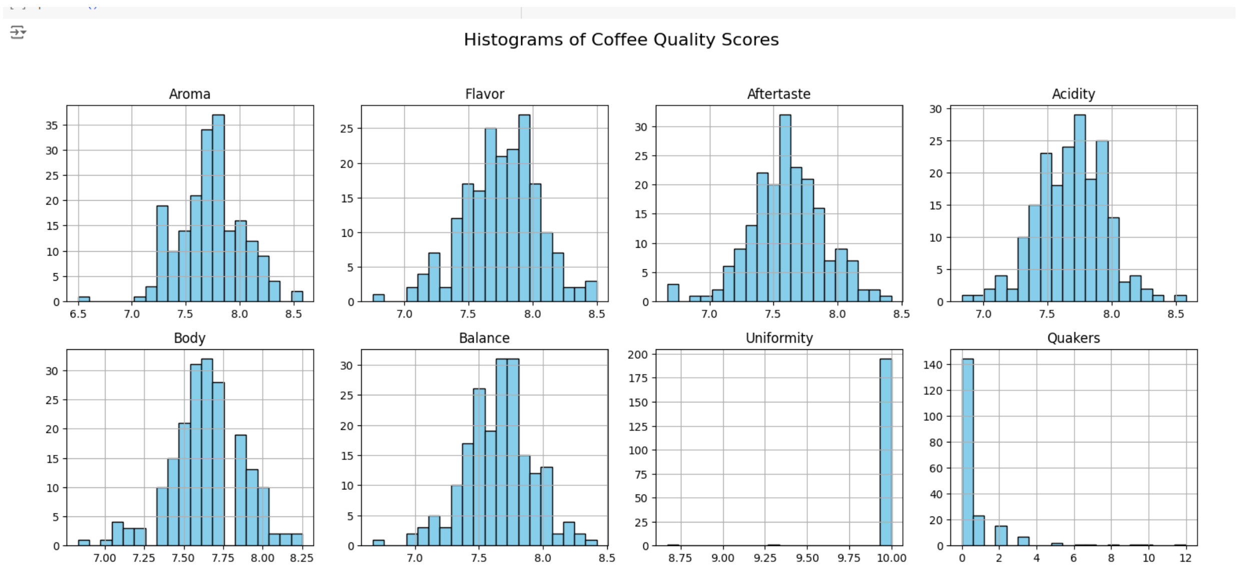
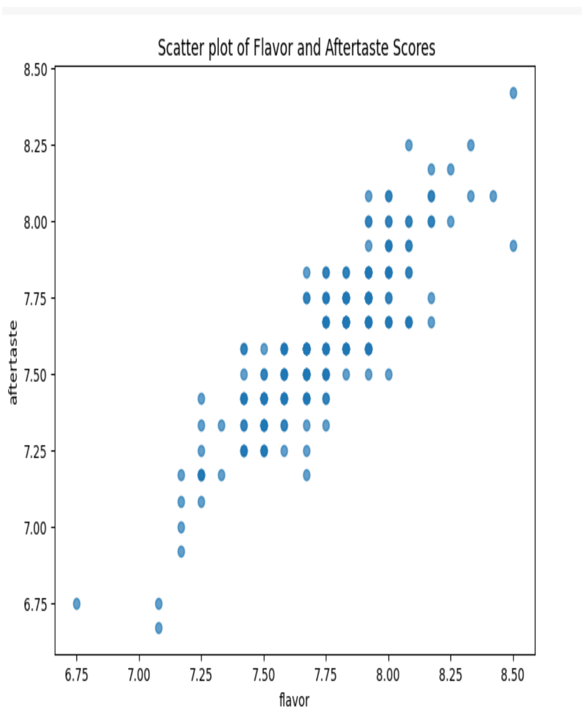
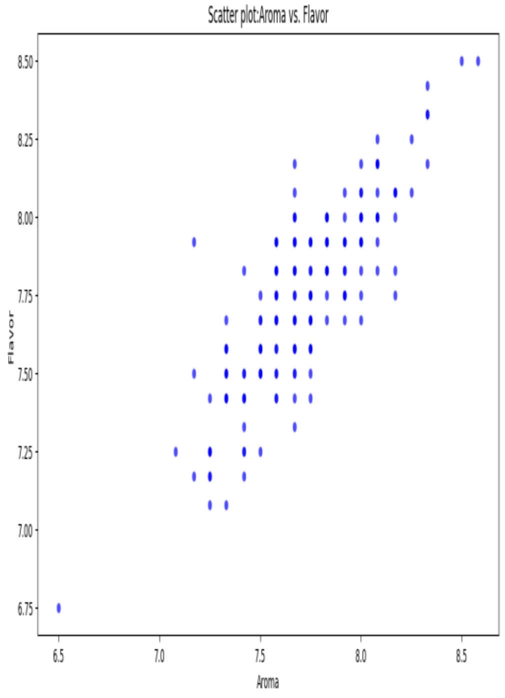
Date	10 July 2024
Team ID	Team-740058
Project Title	Masterful Machines: Precise Coffee Quality Predictions Throogh ML
Maximum Marks	6 Marks

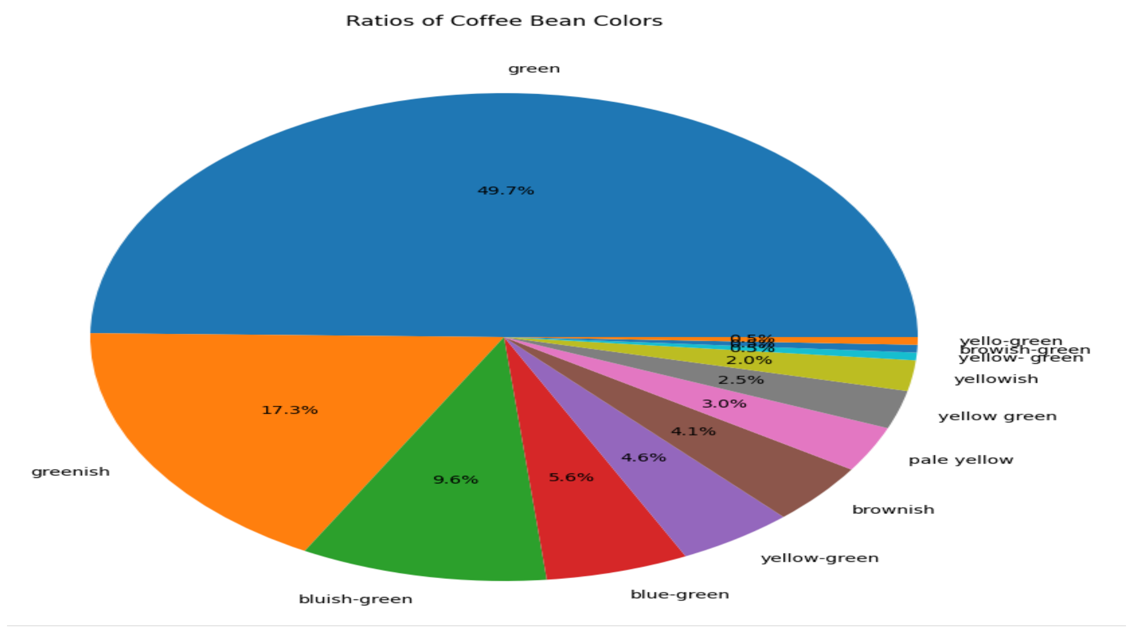
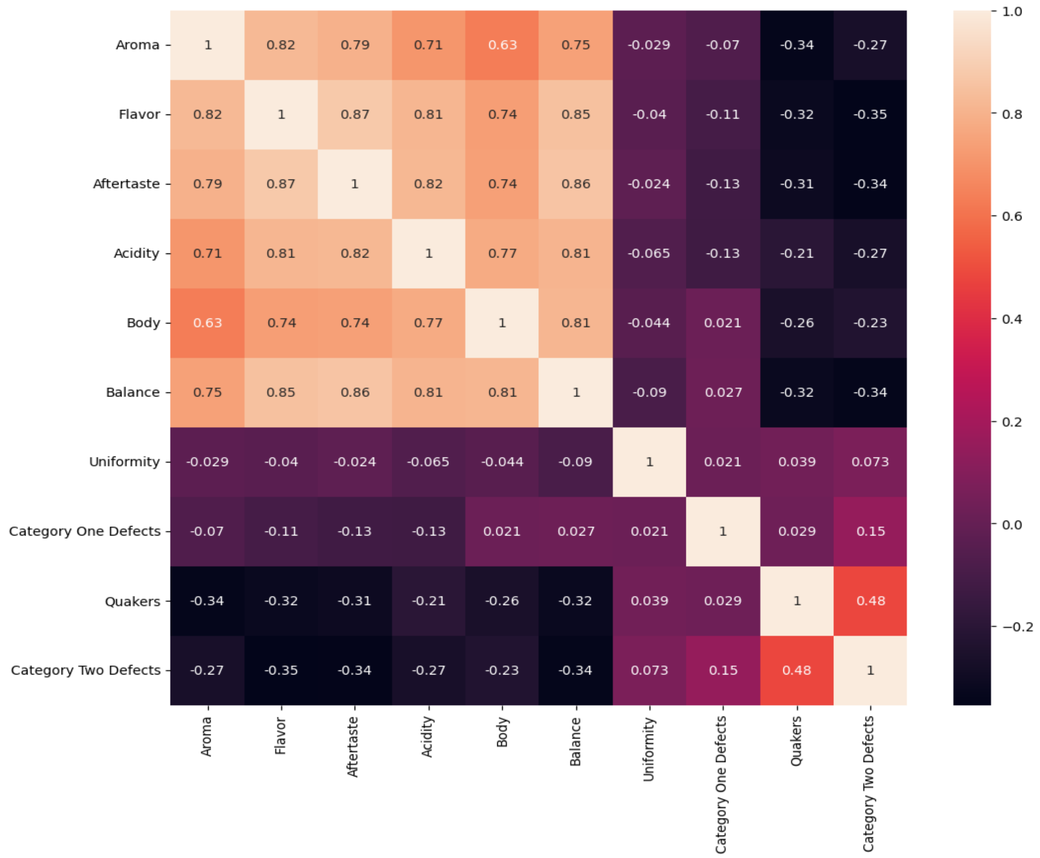
Data Exploration and Preprocessing Report

The coffee quality dataset was explored to understand distributions, correlations, and missing data patterns. Key preprocessing steps included handling missing values, outlier treatment, feature scaling, and encoding categorical variables. The data was then split into training and testing sets, ensuring readiness for precise coffee quality predictions using machine learning models.

<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>Data Overview</td><td>
<u>Dimension:</u>
207 rows × 19 columns
<u>Descriptive statistics:</u>

<table>
<tr><th></th><th>ID</th><th>Number of Bags</th><th>Aroma</th><th>Flavor</th><th>Aftertaste</th><th>Acidity</th><th>Body</th><th>Balance</th><th>Uniformity</th><th>Overall</th><th>Total Cup Points</th><th>Moisture Percentage</th><th>Category One Defects</th><th>Quakers</th><th>Category Two Defects</th></tr>
<tr><td>count</td><td>207.000000</td><td>207.000000</td><td>207.000000</td><td>207.000000</td><td>207.000000</td><td>207.00000</td><td>207.000000</td><td>207.000000</td><td>207.000000</td><td>207.000000</td><td>207.000000</td><td>207.000000</td><td>207.000000</td><td>207.000000</td><td>207.000000</td></tr>
<tr><td>mean</td><td>103.000000</td><td>155.449275</td><td>7.721063</td><td>7.744734</td><td>7.599758</td><td>7.69029</td><td>7.640918</td><td>7.644058</td><td>9.990338</td><td>7.676812</td><td>83.706570</td><td>10.735266</td><td>0.135266</td><td>0.690821</td><td>2.251208</td></tr>
<tr><td>std</td><td>59.899917</td><td>244.484868</td><td>0.287626</td><td>0.279613</td><td>0.275911</td><td>0.25951</td><td>0.233499</td><td>0.256299</td><td>0.103306</td><td>0.306359</td><td>1.730417</td><td>1.247468</td><td>0.592070</td><td>1.686918</td><td>2.950183</td></tr>
<tr><td>min</td><td>0.000000</td><td>1.000000</td><td>6.500000</td><td>6.750000</td><td>6.670000</td><td>6.83000</td><td>6.830000</td><td>6.670000</td><td>8.670000</td><td>6.670000</td><td>78.000000</td><td>0.000000</td><td>0.000000</td><td>0.000000</td><td>0.000000</td></tr>
<tr><td>25%</td><td>51.500000</td><td>1.000000</td><td>7.580000</td><td>7.580000</td><td>7.420000</td><td>7.50000</td><td>7.500000</td><td>7.500000</td><td>10.000000</td><td>7.500000</td><td>82.580000</td><td>10.100000</td><td>0.000000</td><td>0.000000</td><td>0.000000</td></tr>
<tr><td>50%</td><td>103.000000</td><td>14.000000</td><td>7.670000</td><td>7.750000</td><td>7.580000</td><td>7.67000</td><td>7.670000</td><td>7.670000</td><td>10.000000</td><td>7.670000</td><td>83.750000</td><td>10.800000</td><td>0.000000</td><td>0.000000</td><td>1.000000</td></tr>
<tr><td>75%</td><td>154.500000</td><td>275.000000</td><td>7.920000</td><td>7.920000</td><td>7.750000</td><td>7.87500</td><td>7.750000</td><td>7.790000</td><td>10.000000</td><td>7.920000</td><td>84.830000</td><td>11.500000</td><td>0.000000</td><td>1.000000</td><td>3.000000</td></tr>
<tr><td>max</td><td>206.000000</td><td>2240.000000</td><td>8.580000</td><td>8.500000</td><td>8.420000</td><td>8.58000</td><td>8.250000</td><td>8.420000</td><td>10.000000</td><td>8.580000</td><td>89.330000</td><td>13.500000</td><td>5.000000</td><td>12.000000</td><td>16.000000</td></tr>
</table>
</td></tr>
<tr><td>Univariate Analysis</td><td></td></tr>
</table>

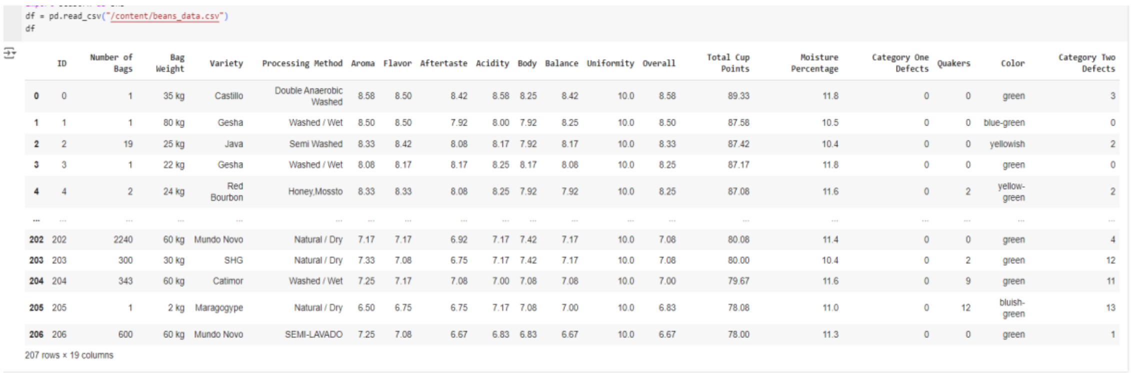
4o
.

	
Bivariate Analysis	 

Multivariate Analysis	
Outliers and Anomalies	

Data Preprocessing Code Screenshots

Loading Data

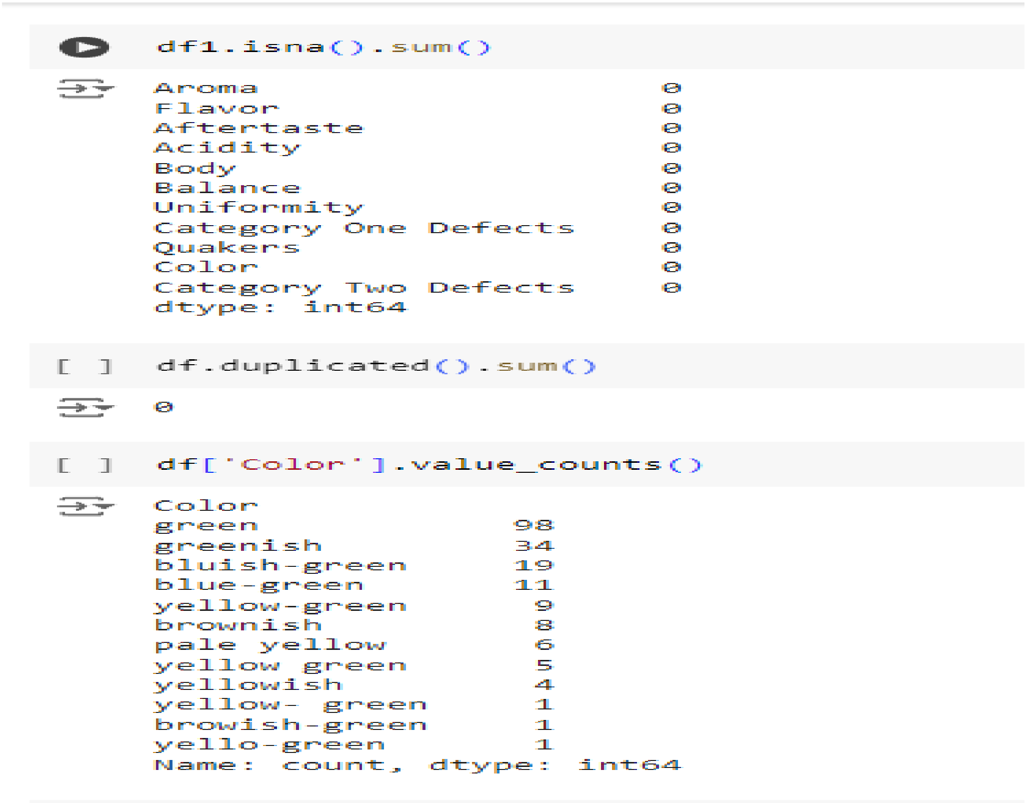


```
df = pd.read_csv("/content/beans_data.csv")
df
```

	ID	Number of Bags	Bag Weight	Variety	Processing Method	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Uniformity	Overall	Total Cup Points	Moisture Percentage	Category One Defects	Quakers	Color	Category Two Defects
0	0	1	35 kg	Castillo	Double Anaerobic Washed	8.58	8.50	8.42	8.58	8.25	8.42	10.0	8.58	89.33	11.8	0	0	green	3
1	1	1	80 kg	Gesha	Washed / Wet	8.50	8.50	7.92	8.00	7.92	8.25	10.0	8.50	87.58	10.5	0	0	blue-green	0
2	2	19	25 kg	Java	Semi Washed	8.33	8.42	8.08	8.17	7.92	8.17	10.0	8.33	87.42	10.4	0	0	yellowish	2
3	3	1	22 kg	Gesha	Washed / Wet	8.08	8.17	8.17	8.25	8.17	8.08	10.0	8.25	87.17	11.8	0	0	green	0
4	4	2	24 kg	Red Bourbon	Honey,Mossto	8.33	8.33	8.08	8.25	7.92	7.92	10.0	8.25	87.08	11.6	0	2	yellow-green	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
202	202	2240	60 kg	Mundo Novo	Natural / Dry	7.17	7.17	6.92	7.17	7.42	7.17	10.0	7.08	80.08	11.4	0	0	green	4
203	203	300	30 kg	SHG	Natural / Dry	7.33	7.08	6.75	7.17	7.42	7.17	10.0	7.08	80.00	10.4	0	2	green	12
204	204	343	60 kg	Catimor	Washed / Wet	7.25	7.17	7.08	7.00	7.08	7.08	10.0	7.00	79.67	11.6	0	9	green	11
205	205	1	2 kg	Maragogype	Natural / Dry	6.50	6.75	6.75	7.17	7.08	7.00	10.0	6.83	78.08	11.0	0	12	bluish-green	13
206	206	600	60 kg	Mundo Novo	SEMI-LAVADO	7.25	7.08	6.67	6.83	6.83	6.67	10.0	6.67	78.00	11.3	0	0	green	1

207 rows × 19 columns

Handling Missing Data



```
df1.isna().sum()
```

```
Aroma                    0
Flavor                   0
Aftertaste               0
Acidity                  0
Body                     0
Balance                  0
Uniformity               0
Category One Defects     0
Quakers                  0
Color                    0
Category Two Defects     0
dtype: int64
```

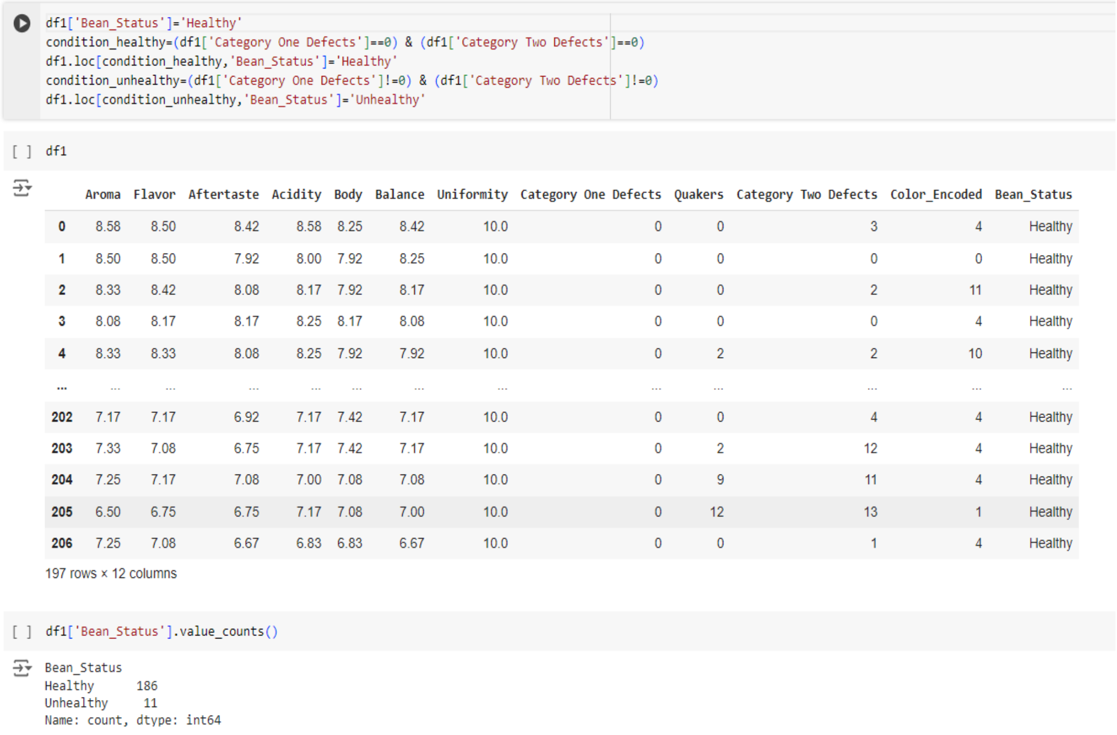
```
[ ]  df.duplicated().sum()
```

```
0
```

```
[ ]  df['Color'].value_counts()
```

```
Color
green              98
greenish           34
bluish-green       19
blue-green         11
yellow-green        9
brownish            8
pale yellow         6
yellow green        5
yellowish           4
yellow- green       1
browish-green       1
yello-green         1
Name: count, dtype: int64
```

Data Transformation



```
df1['Bean_Status']='Healthy'
condition_healthy=(df1['Category One Defects']==0) & (df1['Category Two Defects']==0)
df1.loc[condition_healthy,'Bean_Status']='Healthy'
condition_unhealthy=(df1['Category One Defects']!=0) & (df1['Category Two Defects']!=0)
df1.loc[condition_unhealthy,'Bean_Status']='Unhealthy'
```

```
[ ] df1
```

	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Uniformity	Category One Defects	Quakers	Category Two Defects	Color_Encoded	Bean_Status
0	8.58	8.50	8.42	8.58	8.25	8.42	10.0	0	0	3	4	Healthy
1	8.50	8.50	7.92	8.00	7.92	8.25	10.0	0	0	0	0	Healthy
2	8.33	8.42	8.08	8.17	7.92	8.17	10.0	0	0	2	11	Healthy
3	8.08	8.17	8.17	8.25	8.17	8.08	10.0	0	0	0	4	Healthy
4	8.33	8.33	8.08	8.25	7.92	7.92	10.0	0	2	2	10	Healthy
...	...	...	...	...	...	...	...	...	...	...	...	...
202	7.17	7.17	6.92	7.17	7.42	7.17	10.0	0	0	4	4	Healthy
203	7.33	7.08	6.75	7.17	7.42	7.17	10.0	0	2	12	4	Healthy
204	7.25	7.17	7.08	7.00	7.08	7.08	10.0	0	9	11	4	Healthy
205	6.50	6.75	6.75	7.17	7.08	7.00	10.0	0	12	13	1	Healthy
206	7.25	7.08	6.67	6.83	6.83	6.67	10.0	0	0	1	4	Healthy

197 rows × 12 columns

```
[ ] df1['Bean_Status'].value_counts()
```

```
Bean_Status
Healthy      186
Unhealthy     11
Name: count, dtype: int64
```

Feature Engineering

Attached the codes in final submission.

Save Processed Data

-