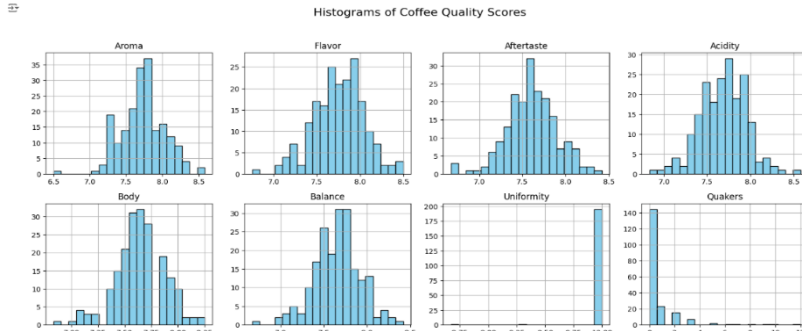


Data Collection and Preprocessing Phase

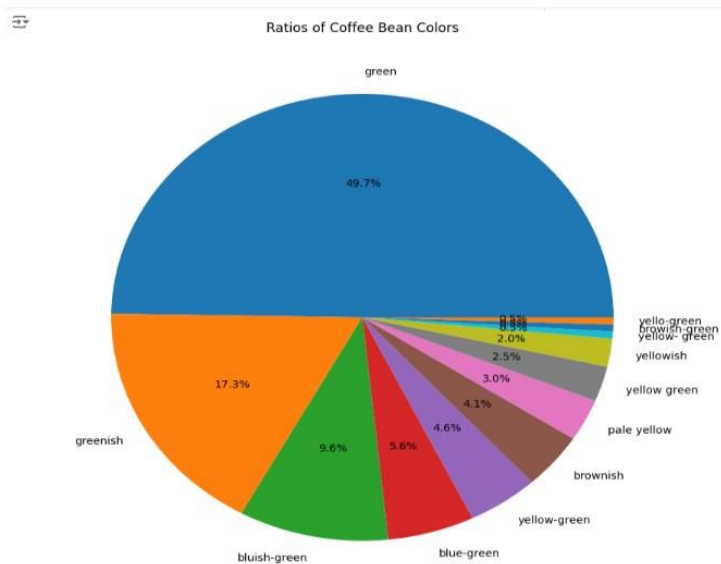
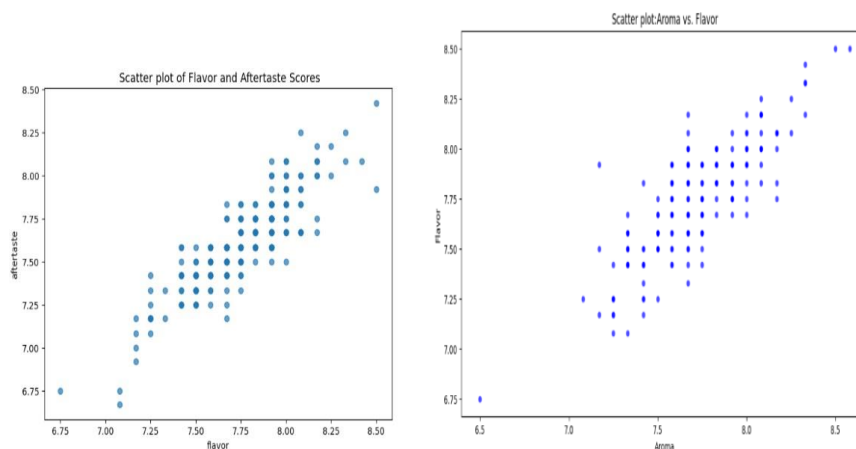
Date	9 July 2024
Team ID	team-739821
Project Title	Precise Coffee Quality Prediction
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

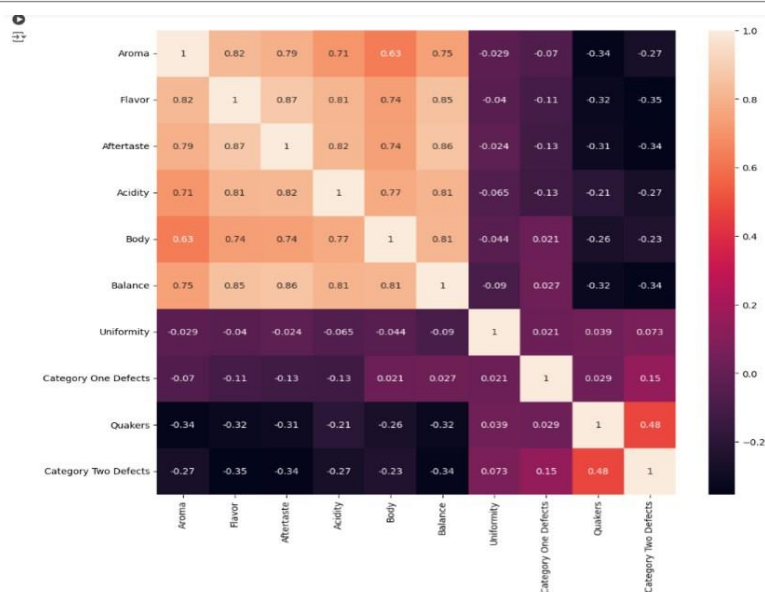
Dataset variables will be statistically analyzed to identify patterns and outliers, with python employed for preprocessing tasks like normalization and feature engineering .Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	Dimensions: 207 rows x 19 columns
	<u>Descriptive Statistics:</u>
Univariate Analysis	

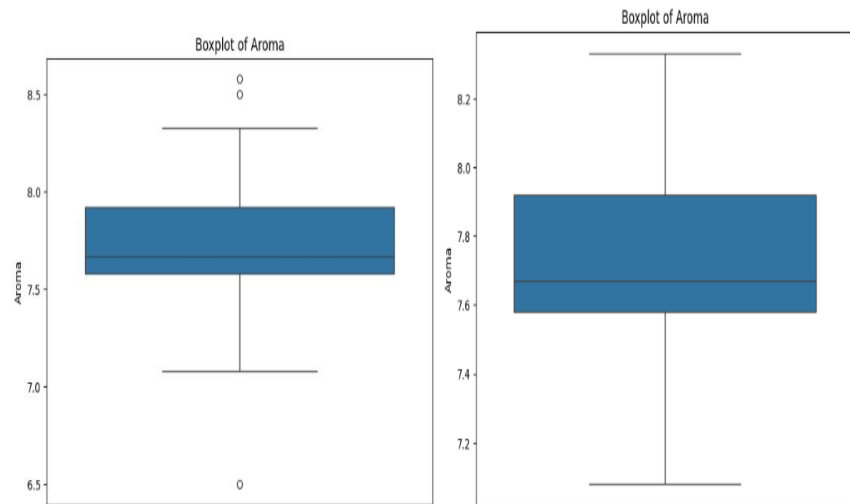
Bivariate Analysis



Multivariate Analysis



Outliers and Anomalies



Data Preprocessing Code Screenshots

Loading Data

```
[ ] Import numpy as np
Import pandas as pd
Import seaborn as sns
df = pd.read_csv("content/beers_data.csv")
df
```

ID	Number of Bags	Bag Weight	Variety	Processing Method	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Intensity	Overall	Total Cup Points	Moisture Percentage	Category One Defects	Quakers	Color	Category Two Defects	
0	0	1	35 kg	Castilla	Double Analeptic Washed	8.58	8.58	8.42	8.58	8.25	8.42	10.0	8.58	88.33	11.8	0	0	green	3
1	1	1	80 kg	Cashia	Washed / Wet	8.58	8.58	7.82	8.00	7.82	8.25	10.0	8.58	87.58	10.5	0	0	blue-green	0
2	2	19	25 kg	Jawa	Semi Washed	8.33	8.42	8.08	8.17	7.82	8.17	10.0	8.33	87.42	10.4	0	0	yellowish	2
3	3	1	22 kg	Cashia	Washed / Wet	8.08	8.17	8.17	8.25	8.17	8.08	10.0	8.25	87.17	11.0	0	0	green	0
4	4	2	24 kg	Real Bouillon	Honey Mashed	8.33	8.33	8.08	8.25	7.82	7.82	10.0	8.25	87.88	11.6	0	2	yellow-green	2
202	202	2240	80 kg	Mundo Novo	Natural / Dry	7.17	7.17	6.82	7.17	7.42	7.17	10.0	7.08	88.88	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	88.00	10.4	0	2	green	12
204	204	343	80 kg	Cashia	Washed / Wet	7.25	7.17	7.08	7.08	7.08	7.08	10.0	7.08	79.87	11.6	0	9	green	11
205	205	1	2 kg	Mangrove	Natural / Dry	6.50	6.75	6.75	7.17	7.08	7.08	10.0	6.83	78.88	11.0	0	12	blue-green	13
206	206	880	80 kg	Mundo Novo	SEM-LAMGO	7.25	7.08	6.87	6.83	6.83	6.87	10.0	6.87	78.00	11.3	0	0	green	1

207 rows x 19 columns

Handling Missing Data

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

```
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
dtype: int64
```

```
[ ] df.dropna(inplace=True)
```

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

```
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
dtype: int64
```

Data Transformation

```
[ ] from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
df1['Color_Encoded'] = label_encoder.fit_transform(df1['Color'])
df1 = df1.drop(['Color'],axis=1)
```

df1

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

Feature Engineering

```
[ ] 205 6.50 6.75 6.75 7.17 7.08 7.00 10.0 0 12 13 1
206 7.25 7.08 6.67 6.83 6.83 6.67 10.0 0 0 1 4
197 rows x 11 columns
```

```
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

Save Processed Data

```
[ ] import pickle
import warnings

[ ] with open("./coffee_quality_prediction(rfc).pkl","wb") as f:
    pickle.dump(RFC,f)
```