In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```
data=pd.read_csv(r"C:\Users\Fundemics\Downloads\banana_quality_dataset.csv")
```

In [3]:

```
data
```

Out[3]:

	sample_id	variety	region	quality_score	quality_category	ripeness_index	ripeness_category
0	1	Manzano	Colombia	1.88	Processing	2.11	Turning
1	2	Plantain	Guatemala	2.42	Processing	4.25	Ripe
2	3	Burro	Ecuador	3.57	Premium	6.24	Overripe
3	4	Manzano	Ecuador	2.21	Processing	5.39	Ripe
4	5	Red Dacca	Ecuador	2.35	Processing	5.84	Ripe
...	...	...	...	...	...	...	...
995	996	Burro	Ecuador	3.50	Good	4.94	Ripe
996	997	Cavendish	Philippines	2.38	Processing	6.74	Overripe
997	998	Plantain	Ecuador	1.68	Processing	1.41	Green
998	999	Fehi	Guatemala	2.02	Processing	1.34	Green
999	1000	Red Dacca	Ecuador	2.69	Good	2.69	Turning

1000 rows × 16 columns

In [4]:

```
'''1.How many total records are present in the dataset?
'''
data.size
```

Out[4]:

```
16000
```

In [5]:

```
'''2.How many unique banana varieties are there'''
data['variety'].nunique()
```

Out[5]:

```
8
```

In [6]:

```
'''3.How many unique regions are included in the dataset?
'''
data['region'].nunique()
```

Out[6]:

```
8
```

In [7]:

```
'''4.What are the different quality categories and their counts?
'''
data.quality_category.value_counts()
```

Out[7]:

```
Processing    506
Good          434
Unripe         35
Premium        25
Name: quality_category, dtype: int64
```

In [8]:

```
'''5.What are the different ripeness categories and their counts?'''
data.ripeness_category.value_counts()
```

Out[8]:

```
Ripe        349
Turning     310
Green       174
Overripe    167
Name: ripeness_category, dtype: int64
```

In [9]:

```
'''6.What is the average quality_score across all bananas?'''
data['quality_score'].mean()
```

Out[9]:

```
2.46516
```

In [10]:

```
'''7.What is the total weight_g of all bananas in the dataset?'''
data['weight_g'].sum()
```

Out[10]:

```
164738.93
```

In [11]:

```
'''8.What is the maximum sugar_content_brix recorded?'''
data['sugar_content_brix'].max()
```

Out[11]:

```
21.98
```

In [12]:

```
'''9.What is the minimum firmness_kgf recorded?'''
data['firmness_kgf'].min()
```

Out[12]:

```
0.5
```

In [13]:

```
'''10.How many bananas have a quality_category of 'Premium'?'''
data[data['quality_category']=='Premium']['quality_category'].value_counts()
```

Out[13]:

```
Premium    25
Name: quality_category, dtype: int64
```

In [14]:

```
'''11.How many bananas are in the 'Ripe' ripeness_category?
'''
data[data['ripeness_category']=='Ripe']['ripeness_category'].value_counts()
```

Out[14]:

```
Ripe    349
Name: ripeness_category, dtype: int64
```

In [15]:

```
'''12.What is the average length_cm for 'Plantain' variety bananas?
'''
data[data['variety']=='Plantain']['length_cm'].mean()
```

Out[15]:

```
18.927328767123285
```

In [16]:

```
'''13.What is the average weight_g for bananas from 'Ecuador'?'''
data[data['region']=='Ecuador']['weight_g'].mean()
```

Out[16]:

```
171.81306569343064
```

In [17]:

```
'''14.How many samples have a tree_age_years greater than 10?'''
data[data['tree_age_years']>10].shape[0]
```

Out[17]:

541

In [18]:

```
'''15.How many samples have rainfall_mm less than 1500?'''
data[data['rainfall_mm']<1500].shape[0]
```

Out[18]:

252

In [19]:

```
'''16.Show all records where quality_score is exactly 2.5.'''
data[data['quality_score']==2.5]
```

Out[19]:

	sample_id	variety	region	quality_score	quality_category	ripeness_index	ripeness_category
11	12	Manzano	Costa Rica	2.5	Processing	6.93	Overripe
25	26	Blue Java	Colombia	2.5	Processing	3.85	Turning
196	197	Lady Finger	Philippines	2.5	Processing	6.84	Overripe
227	228	Blue Java	Honduras	2.5	Processing	2.65	Turning
272	273	Fehi	India	2.5	Processing	5.68	Ripe
473	474	Plantain	Philippines	2.5	Processing	3.42	Turning
512	513	Manzano	Costa Rica	2.5	Processing	6.73	Overripe
785	786	Cavendish	Honduras	2.5	Processing	6.37	Overripe
930	931	Manzano	Honduras	2.5	Processing	4.48	Ripe

In [20]:

```
'''17.Show only sample_id, variety, and region for bananas with quality_category 'Unripe'.'''
data[data['quality_category']=='Unripe'][['sample_id','variety','region']]
```

Out[20]:

	sample_id	variety	region
10	11	Red Dacca	Costa Rica
13	14	Red Dacca	India
49	50	Blue Java	Ecuador
66	67	Plantain	Costa Rica
81	82	Plantain	Costa Rica
112	113	Fehi	Honduras
118	119	Fehi	Costa Rica
201	202	Fehi	Philippines
246	247	Cavendish	Colombia
283	284	Manzano	Ecuador
291	292	Red Dacca	Costa Rica
295	296	Plantain	Philippines
296	297	Blue Java	India
297	298	Manzano	Ecuador
338	339	Plantain	Colombia
349	350	Manzano	Costa Rica
397	398	Blue Java	Guatemala
402	403	Red Dacca	Ecuador
411	412	Fehi	Philippines
417	418	Lady Finger	Guatemala
420	421	Lady Finger	Ecuador
426	427	Cavendish	Guatemala
465	466	Manzano	Colombia
483	484	Blue Java	Ecuador
563	564	Manzano	Philippines
626	627	Burro	Ecuador
650	651	Manzano	India
666	667	Blue Java	Honduras
676	677	Plantain	Guatemala
714	715	Fehi	Colombia
725	726	Burro	Colombia
747	748	Manzano	Ecuador
780	781	Cavendish	Colombia
884	885	Fehi	Ecuador
905	906	Lady Finger	Philippines

In [21]:

```
'''18.What is the average sugar_content_brix for bananas in the 'Overripe' category?'''
data[data['ripeness_category']=='Overripe'][['sugar_content_brix']].mean()
```

Out[21]:

```
sugar_content_brix    18.773832
dtype: float64
```

In [22]:

```
'''19.What is the average firmness_kgf for bananas in the 'Green' ripeness category?'''
data[data['ripeness_category']=='Green']['firmness_kgf'].mean()
```

Out[22]:

```
2.6476436781609194
```

In [23]:

```
'''20.Which variety has the highest average weight_g?'''
data.groupby('variety')['weight_g'].mean().sort_values(ascending=False).index[0]
```

Out[23]:

```
'Cavendish'
```

In [24]:

```
'''21.Which region has the lowest average altitude_m?'''
data.groupby('region')['altitude_m'].min().sort_values().index[0]
```

Out[24]:

```
'Costa Rica'
```

In [25]:

```
'''22.What is the total sugar_content_brix for all 'Manzano' variety bananas?'''
data[data['variety']=='Manzano']['sugar_content_brix'].sum()
```

Out[25]:

```
2165.1
```

In [26]:

```
'''23.How many samples have a length_cm between 15 and 20 (inclusive)?'''
data[(data['length_cm']>=15) & (data['length_cm']<=20)].shape[0]
```

Out[26]:

```
248
```

In [27]:

```
'''24.How many samples have a weight_g between 15 and 20 (inclusive)?'''
data[(data['weight_g']>=15) & (data['weight_g']<=20)].shape[0]
```

Out[27]:

```
0
```

In [28]:

```
'''25.What is the average soil_nitrogen_ppm for bananas with 'Good' quality_category?'''
data[data['quality_category']=='Good']['soil_nitrogen_ppm'].mean()
```

Out[28]:

```
102.80990783410138
```

In [29]:

```
'''26.How many samples are from 'Colombia' and have a ripeness_category of 'Turning'?'''
data[(data['region']=='Colombia')& (data['ripeness_category']=='Turning')].shape[0]
```

Out[29]:

```
40
```

In [30]:

```
'''27.How many samples are 'Cavendish' variety and have a quality_category of 'Processing'?'''
data[(data['variety']=='Cavendish') & (data['quality_category']=='Processing')].shape[0]
```

Out[30]:

```
74
```

In [31]:

```
'''28.What is the average quality_score for each variety?'''
data.groupby('variety')['quality_score'].mean()
```

Out[31]:

```
variety
Blue Java      2.523009
Burro          2.507360
Cavendish      2.433008
Fehi           2.428582
Lady Finger    2.532051
Manzano        2.389160
Plantain       2.463425
Red Dacca      2.454867
Name: quality_score, dtype: float64
```

In [32]:

```
'''29.What is the average sugar_content_brix for each region?'''
data.groupby('region')['sugar_content_brix'].mean()
```

Out[32]:

```
region
Brazil         18.310577
Colombia       18.738897
Costa Rica     18.766308
Ecuador        18.093869
Guatemala      18.564683
Honduras       18.437290
India          18.762556
Philippines    18.411024
Name: sugar_content_brix, dtype: float64
```

In [33]:

```
'''30.What is the total weight_g for each ripeness_category?'''
data.groupby('ripeness_category')['weight_g'].sum()
```

Out[33]:

```
ripeness_category
Green       28509.73
Overripe    27891.94
Ripe        57900.80
Turning     50436.46
Name: weight_g, dtype: float64
```

In [42]:

```
'''31.Which region has the most 'Premium' quality bananas'''
data[data['quality_category']=='Premium']['region'].value_counts().index[0:3]
```

Out[42]:

```
Index(['Ecuador', 'Brazil', 'Costa Rica'], dtype='object')
```

In [38]:

```
'''32.Which variety has the most 'Ripe' bananas?'''
data[data['ripeness_category']=='Ripe']['variety'].value_counts().sort_values(ascending=False).inc
```

Out[38]:

```
'Fehi'
```

In [43]:

```
'''33.What is the maximum ripeness_index for bananas from 'India'?
'''
d=data[data['region']=='India']
d.groupby('region')['ripeness_index'].max()
```

Out[43]:

```
region
India    7.0
Name: ripeness_index, dtype: float64
```

In [44]:

```
'''34.What is the minimum sugar_content_brix for 'Burro' variety bananas?
'''
data[data['variety']=='Burro'].groupby('variety')['sugar_content_brix'].min()
```

Out[44]:

```
variety
Burro    15.01
Name: sugar_content_brix, dtype: float64
```

In [51]:

```
'''35.How many samples have a harvest_date in October 2023? (Assume '2023-10' prefix)
'''
data[(data['harvest_date']>='2023-10-01')&(data['harvest_date']<='2023-10-31')].shape[0]
```

Out[51]:

```
505
```

In [45]:

```
'''36.What is the average tree_age_years for each quality_category?'''
data.groupby('quality_category')['tree_age_years'].mean()
```

Out[45]:

```
quality_category
Good          11.063364
Premium       11.300000
Processing    10.687352
Unripe        11.080000
Name: tree_age_years, dtype: float64
```

In [46]:

```
'''37.How many samples have quality_score greater than 3 and firmness_kgf less than 2?
'''
data[(data['quality_score']>3) & (data['firmness_kgf']<2)].shape[0]
```

Out[46]:

```
55
```

In [47]:

```
'''38.What is the total rainfall_mm for all bananas with a ripeness_category of 'Green'?
'''
data[data['ripeness_category']=='Green']['rainfall_mm'].sum()
```

Out[47]:

```
341820.8999999999
```

In [48]:

```
'''39.Which variety has the highest average soil_nitrogen_ppm?'''
data.groupby('variety')['soil_nitrogen_ppm'].mean().sort_values(ascending=False).index[0]
```

Out[48]:

```
'Manzano'
```

In [49]:

```
'''40.What is the average length_cm for bananas with quality_category 'Processing' and ripeness_ca
data[(data['quality_category']=='Processing') & (data['ripeness_category']=='Turning')]['length_cm
```

Out[49]:

```
18.467581395348837
```

In [ ]: