COFFEE QUALITY

1 Milestone-1 Evaluation Project

Project Documentation: Exploratory Data Analysis of Coffee quality Dataset :

Title : Data Analysis on Coffee quality

Name : Sreehari A

DA/DS : Data Analytics (DA)

Batch number : B4 (June - Online)(M) - DA & DS

Online/Offline : Online

Roll Number : 60624OL004OLR018

Table of Contents :

NOTE : All the codes used for this are given after the documentation and displaying of results.

1) Introduction :

The coffee quality dataset comprises various attributes related to coffee beans, including Species , flavor , aroma , etc. The goal of this project is to conduct a comprehensive analysis of the dataset to derive insights into coffee quality as to what actually derives the quality of a coffee which can be helpful for our clients whether they are thinking of starting a new coffee brand or even for normal consumers to choose the best brand of coffee to have the best cup of coffee..

Columns in the dataset related to Coffee quality:

- *Species*:Species of the coffee plant , here, only arabica and robusta is present.
- *Owner*:The one who farmed the coffee plants.
- *Country.of.Origin*:From which country it comes
- *Farm.Name*: The farm the coffee was grown
- *Lot.Number*: The number of the lot
- *Mill*: The mill it was grown
- *ICO.Number*: It is unique identifier asssigned to each bag of coffee bean
- *Company*: The company that imports and export the coffee beans.
- *Altitude*: How much above the sea level the coffee farm is situated
- *Region*: The region where the coffee plants were grown.
- *Producer*: The one who produces which all coffee.
- *Number.of.Bags*: Number of bags produced.
- *Bag.Weight*: Weight of 1 bag.
- *In.Country.Partner*: The country in which the supplier is for specific coffee company.
- *Harvest.Year*: The year it was harvested.
- *Grading.Date*: The date it was graded.O
- *Owner.1*: The person who got it right after first exporting.
- *Variety*: Variety of the coffee bean
- *Processing.Method*: The method coffee bean was processed(like washed , semi-washed , natural ,etc)
- *Aroma*: How good the smell is.
- *Flavor*: How good the flavor of the coffee is.
- *Aftertaste*: The aftertaste that the coffee leaves in your mouth
- *Acidity*: How low the pH of the coffee is.
- *Body*: Refers to the texture and weight of the coffee in your mouth.
- *Balance*: Refers to harmony and equilibrium of flavors,acidity and body in a cup of coffee.
- *Uniformity*: Refers to the consistency of flavor, quality, and appearance of the coffee beans
- *Clean.Cup*: Refers to a cup of coffee that is free from defects, impurities, and off-flavors
- *Sweetness*: How sweet the coffee is
- *Cupper.Points*: Refer to a standardized system used to evaluate and score the quality of coffee

- *Total.Cup.Points*: Refers to the final score assigned to a coffee based on the evaluation of its various attributes and using the cupper point
- *Moisture*: Amount of moisture the coffee has retained.
- *Category.One.Defects*: More severe defects that affect coffee quality(like, Moldy, Skunky, fermented, etc)
- *Quakers*: Refer to a type of defective coffee bean that is lighter in color and has a distinct flavor.
- *Color*: Refers to the visual appearance of the coffee beans which can indicate various aspects of coffee.
- *Category.Two.Defects*: Less severe defects that affect coffee quality(like, woody, nutty, papery, etc)
- *Expiration*: When the coffee becomes unable to consume.
- *Certification.Body*: Organisation that ensure that coffee beans, farms, or production processes meet certain standards.
- *Certification.Address*: Address of the certification body that certified the specific bag of coffee beans.
- *Certification.Contact*: Contact method and info for the certification body.
- *unit_of_measurement*: Units that help coffee professionals and enthusiasts measure, communicate, and perfect their coffee-related tasks.
- *altitude_low_meters*: Refers to a measurement of altitude (height above sea level) that is relatively low, in meters.
- *altitude_high_meters*: Refers to a measurement of altitude (height above sea level) that is relatively high, in meters.
- *altitude_mean_meters*: Altitude mean meters refers to the average height of a location or area above sea level, measured in meters. **

2) Aim :

The aim of this project is to conduct a comprehensive analysis of the dataset to derive insights into overall coffee quality, catering to both consumers and manufacturers in the computer industry.

Problem Statement:

The coffee market is highly competitive, as almost everyone consumes coffee, most of us need a cup of coffee to even function properly.So it is essential for someone trying to make their own coffee brand to aware what all to focus on to make sure that they have the best quality of coffee in the current market and what all to keep an eye on to stay ahead of the competition. This can also be useful for coffee enthusiasts to make sure that they start their day with the best quality of coffee.

Specifically, the problem is:

1) What makes the best cup of coffee possible to start your everday? 2) What are the most important values that will give you the best quality of coffee? 3) What values should we concentrate for a best flavor?

4) Project Workflow :

Overview of the project workflow or methodology followed.

- Data Cleaning
- Exploratory Data Analysis (EDA)
- Data Visualization
- Analysis and Interpretation
- Documentation

5) Data Understanding :

➢ Description of the dataset, including structure, dimensions, and data types. ➢ Summary statistics and insights gained from initial data exploration. **Insights gained from initial data exploration**

- There are 1339 rows and 44 columns in the Dataset.
- From the info we conclude that out of the 44 columns, 24 were object type, 17 were float and 3 were integer.
- Unnamed: 0 column should be dropped

```
import numpy as np
import pandas as pd
z=pd.read_csv("C:/Users/Administrator/Desktop/coffeeQuality.csv")
#Loading data from a CSV file into a Pandas DataFrame
z
```

```
      Unnamed: 0  Species                    Owner Country.of.Origin
\
0              0  Arabica                metad plc          Ethiopia

1              1  Arabica                metad plc          Ethiopia

2              2  Arabica  grounds for health admin         Guatemala

3              3  Arabica       yidnekachew dabessa          Ethiopia

4              4  Arabica                metad plc          Ethiopia

...          ...      ...                      ...               ...

1334        1334  Robusta              luis robles           Ecuador

1335        1335  Robusta              luis robles           Ecuador

1336        1336  Robusta              james moore     United States

1337        1337  Robusta           cafe politico             India

1338        1338  Robusta           cafe politico           Vietnam
```

```
                                    Farm.Name Lot.Number
Mill  \
0                                   metad plc        NaN       metad
plc
1                                   metad plc        NaN       metad
plc
2     san marcos barrancas "san cristobal cuch        NaN
NaN
3        yidnekachew dabessa coffee plantation        NaN
wolensu
4                                   metad plc        NaN       metad
plc
...                                       ...        ...
...
1334                                robustasa   Lavado 1     our own
lab
1335                                robustasa   Lavado 3  own
laboratory
1336                          fazenda cazengo        NaN    cafe
cazengo
1337                                      NaN        NaN
NaN
1338                                      NaN        NaN
NaN

             ICO.Number                                  Company
Altitude  \
0              2014/2015      metad agricultural developmet plc
1950-2200
1              2014/2015      metad agricultural developmet plc
1950-2200
2                    NaN                                   NaN  1600 -
1800 m
3                    NaN  yidnekachew debessa coffee plantation
1800-2200
4              2014/2015      metad agricultural developmet plc
1950-2200
...                  ...                                   ...
...
1334                 NaN                              robustasa
NaN
1335                 NaN                              robustasa
40
1336                 NaN                global opportunity fund     795
meters
1337  14-1118-2014-0087                         cafe politico
NaN
1338                 NaN                         cafe politico
```

```
NaN

      ...       Color Category.Two.Defects           Expiration  \
0     ...       Green                    0      April 3rd, 2016
1     ...       Green                    1      April 3rd, 2016
2     ...         NaN                    0       May 31st, 2011
3     ...       Green                    2     March 25th, 2016
4     ...       Green                    2      April 3rd, 2016
...   ...         ...                  ...                  ...
1334  ...  Blue-Green                    1   January 18th, 2017
1335  ...  Blue-Green                    0   January 18th, 2017
1336  ...         NaN                    6  December 23rd, 2015
1337  ...       Green                    1    August 25th, 2015
1338  ...         NaN                    9    August 25th, 2015

                      Certification.Body  \
0     METAD Agricultural Development plc
1     METAD Agricultural Development plc
2           Specialty Coffee Association
3     METAD Agricultural Development plc
4     METAD Agricultural Development plc
...                                  ...
1334        Specialty Coffee Association
1335        Specialty Coffee Association
1336        Specialty Coffee Association
1337        Specialty Coffee Association
1338        Specialty Coffee Association

                         Certification.Address  \
0     309fcf77415a3661ae83e027f7e5f05dad786e44
1     309fcf77415a3661ae83e027f7e5f05dad786e44
2     36d0d00a3724338ba7937c52a378d085f2172daa
3     309fcf77415a3661ae83e027f7e5f05dad786e44
4     309fcf77415a3661ae83e027f7e5f05dad786e44
...                                       ...
1334  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1335  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1336  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1337  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1338  ff7c18ad303d4b603ac3f8cff7e611ffc735e720

                         Certification.Contact unit_of_measurement  \
0     19fef5a731de2db57d16da10287413f5f99bc2dd                    m
1     19fef5a731de2db57d16da10287413f5f99bc2dd                    m
2     0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m
3     19fef5a731de2db57d16da10287413f5f99bc2dd                    m
4     19fef5a731de2db57d16da10287413f5f99bc2dd                    m
...                                       ...                  ...
1334  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
1335  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
```

```
1336  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
1337  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
1338  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m

     altitude_low_meters altitude_high_meters altitude_mean_meters
0                 1950.0                2200.0                2075.0
1                 1950.0                2200.0                2075.0
2                 1600.0                1800.0                1700.0
3                 1800.0                2200.0                2000.0
4                 1950.0                2200.0                2075.0
...                  ...                   ...                   ...
1334                 NaN                   NaN                   NaN
1335                40.0                  40.0                  40.0
1336               795.0                 795.0                 795.0
1337                 NaN                   NaN                   NaN
1338                 NaN                   NaN                   NaN

[1339 rows x 44 columns]
```

```
#Displays a concise summary of the DataFrame's structure, content, and
memory usage
z.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1339 entries, 0 to 1338
Data columns (total 44 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Unnamed: 0             1339 non-null   int64
 1   Species                1339 non-null   object
 2   Owner                  1332 non-null   object
 3   Country.of.Origin      1338 non-null   object
 4   Farm.Name              980 non-null    object
 5   Lot.Number             276 non-null    object
 6   Mill                   1021 non-null   object
 7   ICO.Number             1180 non-null   object
 8   Company                1130 non-null   object
 9   Altitude               1113 non-null   object
 10  Region                 1280 non-null   object
 11  Producer               1107 non-null   object
 12  Number.of.Bags         1338 non-null   float64
 13  Bag.Weight             1339 non-null   object
 14  In.Country.Partner     1339 non-null   object
 15  Harvest.Year           1292 non-null   object
 16  Grading.Date           1339 non-null   object
 17  Owner.1                1332 non-null   object
 18  Variety                1113 non-null   object
 19  Processing.Method      1169 non-null   object
 20  Aroma                  1339 non-null   float64
 21  Flavor                 1339 non-null   float64
```

```
 22  Aftertaste              1339 non-null   float64
 23  Acidity                 1339 non-null   float64
 24  Body                    1339 non-null   float64
 25  Balance                 1339 non-null   float64
 26  Uniformity              1339 non-null   float64
 27  Clean.Cup               1339 non-null   float64
 28  Sweetness               1339 non-null   float64
 29  Cupper.Points           1339 non-null   float64
 30  Total.Cup.Points        1339 non-null   float64
 31  Moisture                1339 non-null   float64
 32  Category.One.Defects    1339 non-null   int64
 33  Quakers                 1338 non-null   float64
 34  Color                   1069 non-null   object
 35  Category.Two.Defects    1339 non-null   int64
 36  Expiration              1339 non-null   object
 37  Certification.Body      1339 non-null   object
 38  Certification.Address   1339 non-null   object
 39  Certification.Contact   1339 non-null   object
 40  unit_of_measurement     1339 non-null   object
 41  altitude_low_meters     1109 non-null   float64
 42  altitude_high_meters    1109 non-null   float64
 43  altitude_mean_meters    1109 non-null   float64
dtypes: float64(17), int64(3), object(24)
memory usage: 460.4+ KB
```

I) Data Cleaning

```
#Removing the unnamed column
d=z.iloc[:,1:]

d
```

```
      Species                     Owner Country.of.Origin  \
0     Arabica                 metad plc          Ethiopia
1     Arabica                 metad plc          Ethiopia
2     Arabica  grounds for health admin         Guatemala
3     Arabica       yidnekachew dabessa          Ethiopia
4     Arabica                 metad plc          Ethiopia
...       ...                       ...               ...
1334  Robusta              luis robles           Ecuador
1335  Robusta              luis robles           Ecuador
1336  Robusta              james moore     United States
1337  Robusta            cafe politico             India
1338  Robusta            cafe politico           Vietnam

                                  Farm.Name Lot.Number
Mill  \
0                                 metad plc        NaN       metad
plc
1                                 metad plc        NaN       metad
```

```
plc
2     san marcos barrancas "san cristobal cuch        NaN
NaN
3        yidnekachew dabessa coffee plantation        NaN
wolensu
4                                   metad plc        NaN       metad
plc
...                                       ...        ...
...
1334                                robustasa   Lavado 1     our own
lab
1335                                robustasa   Lavado 3  own
laboratory
1336                          fazenda cazengo        NaN    cafe
cazengo
1337                                      NaN        NaN
NaN
1338                                      NaN        NaN
NaN

             ICO.Number                                  Company
Altitude  \
0             2014/2015      metad agricultural developmet plc
1950-2200
1             2014/2015      metad agricultural developmet plc
1950-2200
2                   NaN                                   NaN  1600 -
1800 m
3                   NaN  yidnekachew debessa coffee plantation
1800-2200
4             2014/2015      metad agricultural developmet plc
1950-2200
...                 ...                                   ...
...
1334                NaN                            robustasa
NaN
1335                NaN                            robustasa
40
1336                NaN                global opportunity fund     795
meters
1337  14-1118-2014-0087                          cafe politico
NaN
1338                NaN                          cafe politico
NaN

                          Region  ...       Color
Category.Two.Defects  \
0                     guji-hambela  ...       Green
0
```

```
1                       guji-hambela  ...       Green
1
2                                NaN  ...         NaN
0
3                             oromia  ...       Green
2
4                       guji-hambela  ...       Green
2
...                              ...  ...         ...
...
1334               san juan, playas  ...  Blue-Green
1
1335               san juan, playas  ...  Blue-Green
0
1336  kwanza norte province, angola  ...         NaN
6
1337                             NaN  ...       Green
1
1338                             NaN  ...         NaN
9

                Expiration                   Certification.Body  \
0         April 3rd, 2016  METAD Agricultural Development plc
1         April 3rd, 2016  METAD Agricultural Development plc
2          May 31st, 2011        Specialty Coffee Association
3        March 25th, 2016  METAD Agricultural Development plc
4         April 3rd, 2016  METAD Agricultural Development plc
...                   ...                                 ...
1334   January 18th, 2017        Specialty Coffee Association
1335   January 18th, 2017        Specialty Coffee Association
1336  December 23rd, 2015        Specialty Coffee Association
1337    August 25th, 2015        Specialty Coffee Association
1338    August 25th, 2015        Specialty Coffee Association

                          Certification.Address  \
0     309fcf77415a3661ae83e027f7e5f05dad786e44
1     309fcf77415a3661ae83e027f7e5f05dad786e44
2     36d0d00a3724338ba7937c52a378d085f2172daa
3     309fcf77415a3661ae83e027f7e5f05dad786e44
4     309fcf77415a3661ae83e027f7e5f05dad786e44
...                                        ...
1334  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1335  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1336  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1337  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1338  ff7c18ad303d4b603ac3f8cff7e611ffc735e720

                          Certification.Contact unit_of_measurement  \
0     19fef5a731de2db57d16da10287413f5f99bc2dd                    m
1     19fef5a731de2db57d16da10287413f5f99bc2dd                    m
```

```
2     0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m
3     19fef5a731de2db57d16da10287413f5f99bc2dd                    m
4     19fef5a731de2db57d16da10287413f5f99bc2dd                    m
...                                       ...                   ...
1334  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
1335  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
1336  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
1337  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
1338  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
```

	altitude_low_meters	altitude_high_meters	altitude_mean_meters
0	1950.0	2200.0	2075.0
1	1950.0	2200.0	2075.0
2	1600.0	1800.0	1700.0
3	1800.0	2200.0	2000.0
4	1950.0	2200.0	2075.0
...	...	...	...
1334	NaN	NaN	NaN
1335	40.0	40.0	40.0
1336	795.0	795.0	795.0
1337	NaN	NaN	NaN
1338	NaN	NaN	NaN

```
[1339 rows x 43 columns]
```

➽ Handle missing values:

```
#Returns the count of missing values in each column, helping identify
data quality issues.
d.isnull().sum()
```

```
Species                     0
Owner                       7
Country.of.Origin           1
Farm.Name                 359
Lot.Number               1063
Mill                      318
ICO.Number                159
Company                   209
Altitude                  226
Region                     59
Producer                  232
Number.of.Bags              1
Bag.Weight                  0
In.Country.Partner          0
Harvest.Year               47
Grading.Date                0
Owner.1                     7
Variety                   226
Processing.Method         170
```

```
Aroma                       0
Flavor                      0
Aftertaste                  0
Acidity                     0
Body                        0
Balance                     0
Uniformity                  0
Clean.Cup                   0
Sweetness                   0
Cupper.Points               0
Total.Cup.Points            0
Moisture                    0
Category.One.Defects        0
Quakers                     1
Color                     270
Category.Two.Defects        0
Expiration                  0
Certification.Body          0
Certification.Address       0
Certification.Contact       0
unit_of_measurement         0
altitude_low_meters       230
altitude_high_meters      230
altitude_mean_meters      230
dtype: int64
```

```
#Removing columns with more than 15% misssing value
df=d.drop(columns=['Lot.Number','Owner','Owner.1','Region','Farm.Name'
,'Mill','Color','ICO.Number','Producer','Altitude','altitude_low_meter
s','altitude_high_meters','altitude_mean_meters','Company','Variety'])

df.isnull().sum()
```

```
Species                   0
Country.of.Origin         1
Number.of.Bags            1
Bag.Weight                0
In.Country.Partner        0
Harvest.Year             47
Grading.Date              0
Processing.Method       170
Aroma                     0
Flavor                    0
Aftertaste                0
Acidity                   0
Body                      0
Balance                   0
Uniformity                0
Clean.Cup                 0
Sweetness                 0
```

```
Cupper.Points               0
Total.Cup.Points            0
Moisture                    0
Category.One.Defects        0
Quakers                     1
Category.Two.Defects        0
Expiration                  0
Certification.Body          0
Certification.Address       0
Certification.Contact       0
unit_of_measurement         0
dtype: int64
```

```
#Checking for all the unique values
df['Processing.Method'].unique()
```

```
array(['Washed / Wet', nan, 'Natural / Dry', 'Pulped natural / honey',
       'Semi-washed / Semi-pulped', 'Other'], dtype=object)
```

```
#Getting the mode for the column & storing it in a variable
a=df['Processing.Method'].mode()[0]
a
```

```
'Washed / Wet'
```

```
df['Processing.Method'].fillna(a,inplace=True) #Missing value
imputation
df.isnull().sum()
```

```
Species                    0
Country.of.Origin          1
Number.of.Bags             1
Bag.Weight                 0
In.Country.Partner         0
Harvest.Year              47
Grading.Date               0
Processing.Method          0
Aroma                      0
Flavor                     0
Aftertaste                 0
Acidity                    0
Body                       0
Balance                    0
Uniformity                 0
Clean.Cup                  0
Sweetness                  0
Cupper.Points              0
Total.Cup.Points           0
Moisture                   0
Category.One.Defects       0
Quakers                    1
```

```
Category.Two.Defects      0
Expiration                0
Certification.Body        0
Certification.Address     0
Certification.Contact     0
unit_of_measurement       0
dtype: int64
```

```
df['Harvest.Year'].unique()
```

```
array(['2014', nan, '2013', '2012', 'Mar-10', 'Sept 2009 - April
2010',
       'May-August', '2009/2010', '2015', '2011', '2016', '2015/2016',
       '2010', 'Fall 2009', '2017', '2009 / 2010', '2010-2011',
       '2009-2010', '2009 - 2010', '2013/2014', '2017 / 2018', 'mmm',
       'TEST', 'December 2009-March 2010', '2014/2015', '2011/2012',
       'Jan-11', '4T/10', '2016 / 2017', '23-Jul-10',
       'January Through April', '1T/2011', '4t/2010', '4T/2010',
       'August to December', 'Mayo a Julio', '47/2010', 'Abril -
Julio',
       '4t/2011', 'Abril - Julio /2011', 'Spring 2011 in Colombia.',
       '3T/2011', '2016/2017', '1t/2011', '2018', '4T72010', '08/09
crop'],
      dtype=object)
```

```
a=df['Harvest.Year'].mode()[0]
a
```

```
'2012'
```

```
df['Harvest.Year'].fillna(a,inplace=True)
df.isnull().sum()
```

```
Species                   0
Country.of.Origin         1
Number.of.Bags            1
Bag.Weight                0
In.Country.Partner        0
Harvest.Year              0
Grading.Date              0
Processing.Method         0
Aroma                     0
Flavor                    0
Aftertaste                0
Acidity                   0
Body                      0
Balance                   0
Uniformity                0
Clean.Cup                 0
Sweetness                 0
Cupper.Points             0
```

```
Total.Cup.Points          0
Moisture                  0
Category.One.Defects      0
Quakers                   1
Category.Two.Defects      0
Expiration                0
Certification.Body        0
Certification.Address     0
Certification.Contact     0
unit_of_measurement       0
dtype: int64
```

```
df['Quakers'].unique()
```

```
array([ 0.,  1.,  4.,  2.,  5.,  6.,  3., 11.,  7., nan,  9.,  8.])
```

```
df.Quakers.mode()[0]
```

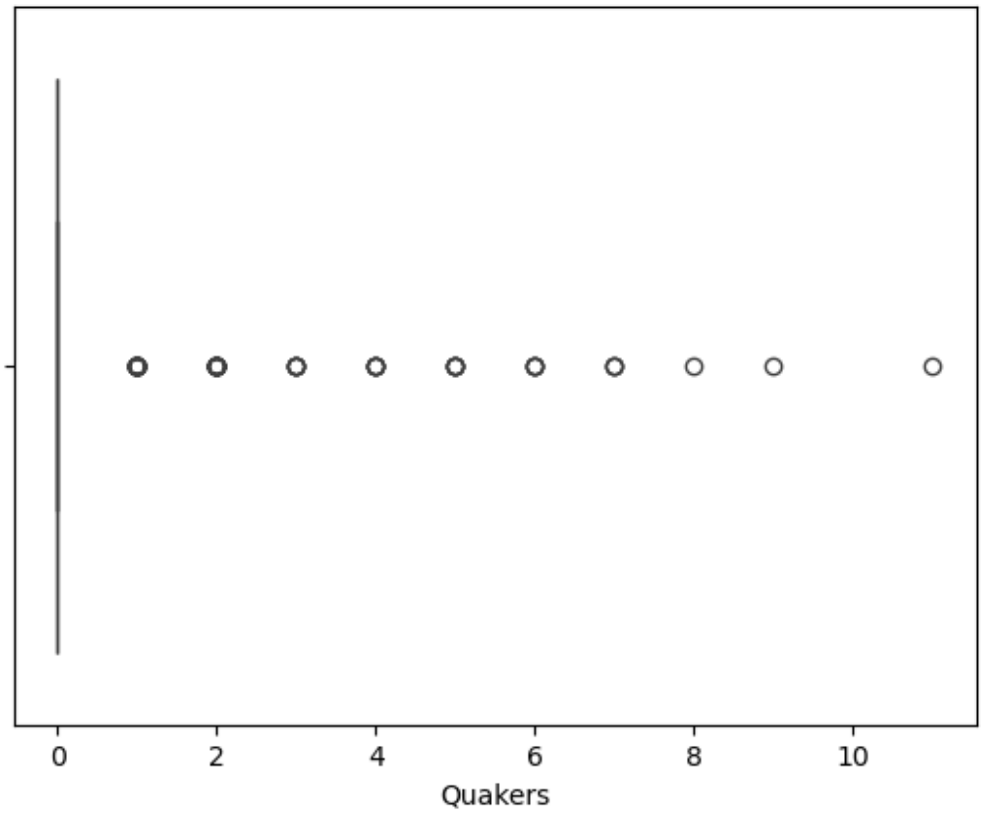
```
0.0
```

```
df['Quakers'].dtype #checking for the type
```

```
dtype('float64')
```

```
import seaborn as sns

#Plotting a boxplot for the column
sns.boxplot(x=df['Quakers'])
```

```
<Axes: xlabel='Quakers'>
```




```
#Sorting the values in ascending order to find median
df['Quakers'].sort_values(ascending=True)
```

```
0         0.0
892       0.0
891       0.0
890       0.0
889       0.0
         ...
1260      7.0
1186      8.0
637       9.0
241      11.0
366       NaN
Name: Quakers, Length: 1339, dtype: float64
```

```
#Finding median & storing it in a variable
a=df.Quakers.median()
a
```

```
0.0
```

```
df.Quakers.fillna(a,inplace=True)
df.isnull().sum()
```

```
Species                   0
Country.of.Origin         1
Number.of.Bags            1
Bag.Weight                0
In.Country.Partner        0
Harvest.Year              0
Grading.Date              0
Processing.Method         0
Aroma                     0
Flavor                    0
Aftertaste                0
Acidity                   0
Body                      0
Balance                   0
Uniformity                0
Clean.Cup                 0
Sweetness                 0
Cupper.Points             0
Total.Cup.Points          0
Moisture                  0
Category.One.Defects      0
Quakers                   0
Category.Two.Defects      0
Expiration                0
Certification.Body        0
Certification.Address     0
Certification.Contact     0
unit_of_measurement       0
dtype: int64
```

```
df['Country.of.Origin'].unique()
```

```
array(['Ethiopia', 'Guatemala', 'Brazil', 'Peru', 'United States',
       'United States (Hawaii)', 'Indonesia', 'China', 'Costa Rica',
       'Mexico', 'Uganda', 'Honduras', 'Taiwan', 'Nicaragua',
       'Tanzania, United Republic Of', 'Kenya', 'Thailand',
'Colombia',
       'Panama', 'Papua New Guinea', 'El Salvador', 'Japan',
'Ecuador',
       'United States (Puerto Rico)', 'Haiti', 'Burundi', 'Vietnam',
       'Philippines', 'Rwanda', 'Malawi', 'Laos', 'Zambia', 'Myanmar',
       'Mauritius', 'Cote d?Ivoire', nan, 'India'], dtype=object)
```

```
a=df['Country.of.Origin'].mode()[0]
a
```

```
'Mexico'
```

```
df['Country.of.Origin'].fillna(a,inplace=True)
df.isnull().sum()
```

```
Species                   0
Country.of.Origin         0
Number.of.Bags            1
Bag.Weight                0
In.Country.Partner        0
Harvest.Year              0
Grading.Date              0
Processing.Method         0
Aroma                     0
Flavor                    0
Aftertaste                0
Acidity                   0
Body                      0
Balance                   0
Uniformity                0
Clean.Cup                 0
Sweetness                 0
Cupper.Points             0
Total.Cup.Points          0
Moisture                  0
Category.One.Defects      0
Quakers                   0
Category.Two.Defects      0
Expiration                0
Certification.Body        0
Certification.Address     0
Certification.Contact     0
unit_of_measurement       0
dtype: int64
```

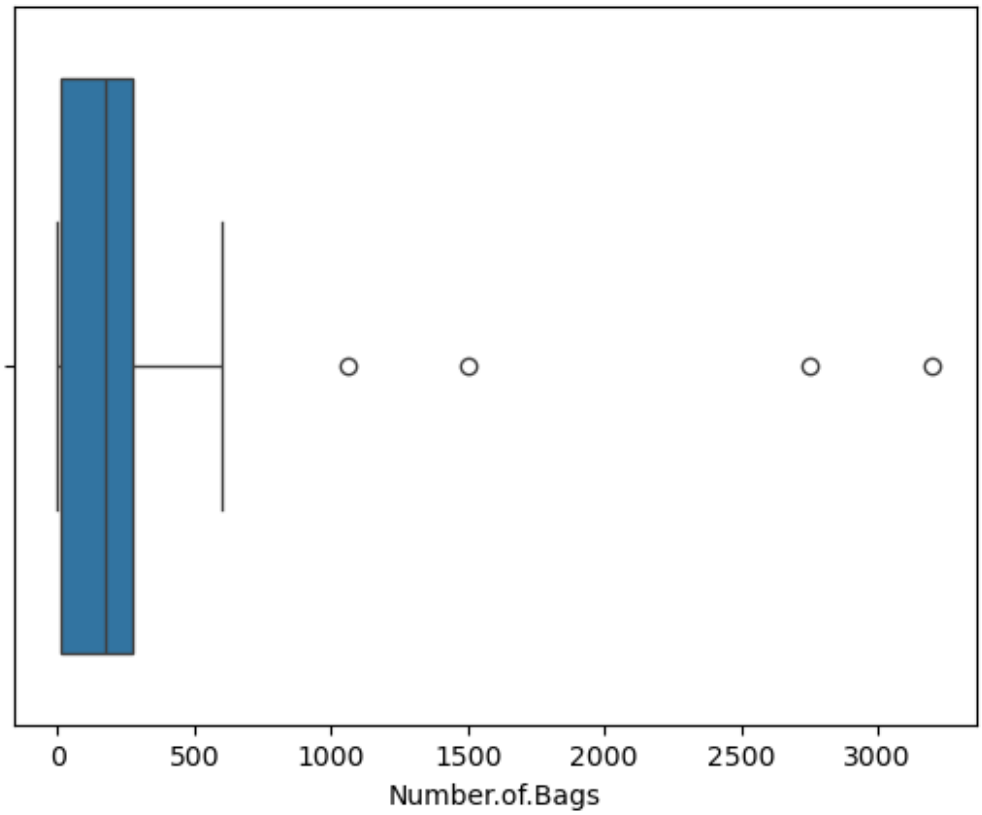
```
df['Number.of.Bags'].unique()
```

```
array([3.000e+02, 5.000e+00, 3.200e+02, 1.000e+02,       nan,
5.000e+01,
       1.000e+01, 1.000e+00, 1.500e+02, 3.000e+00, 2.500e+02,
1.400e+01,
       2.750e+02, 2.000e+01, 2.900e+01, 2.500e+01, 5.300e+01,
1.200e+01,
       7.000e+00, 8.000e+01, 3.700e+01, 2.800e+02, 1.900e+01,
8.000e+00,
       1.600e+01, 2.000e+00, 3.600e+01, 3.600e+02, 5.400e+01,
1.300e+01,
       2.700e+01, 2.000e+02, 1.350e+02, 1.700e+02, 3.800e+01,
3.100e+01,
       1.500e+01, 2.430e+02, 2.520e+02, 1.340e+02, 4.000e+00,
1.200e+02,
       2.750e+03, 2.350e+02, 1.250e+02, 6.600e+01, 7.500e+01,
1.100e+01,
       3.500e+01, 5.600e+01, 3.040e+02, 6.900e+01, 1.500e+03,
2.300e+02,
```

```
       2.480e+02, 6.500e+01, 3.770e+02, 1.300e+02, 3.050e+02,
3.200e+03,
       1.380e+02, 2.700e+02, 4.500e+01, 2.260e+02, 4.800e+01,
1.670e+02,
       1.750e+02, 1.800e+01, 2.850e+02, 3.300e+01, 2.450e+02,
1.800e+02,
       6.000e+02, 5.000e+02, 3.900e+01, 6.000e+00, 2.200e+02,
2.600e+01,
       3.000e+01, 2.320e+02, 8.400e+01, 9.000e+01, 3.100e+02,
3.250e+02,
       1.700e+01, 1.210e+02, 2.300e+01, 1.290e+02, 4.000e+01,
3.200e+01,
       2.100e+01, 6.000e+01, 9.300e+01, 7.700e+01, 2.880e+02,
1.980e+02,
       7.000e+01, 4.200e+01, 2.800e+01, 4.300e+01, 4.900e+01,
7.400e+01,
       5.100e+01, 0.000e+00, 4.400e+01, 1.062e+03, 1.490e+02,
2.740e+02,
       1.140e+02, 4.500e+02, 6.200e+01, 1.660e+02, 2.400e+01,
3.020e+02,
       5.800e+01, 1.650e+02, 5.500e+02, 1.230e+02, 2.400e+02,
1.600e+02,
       9.400e+01, 4.400e+02, 2.200e+01, 2.560e+02, 4.000e+02,
8.200e+01,
       2.090e+02, 3.800e+02, 2.530e+02, 2.230e+02, 1.270e+02,
2.020e+02,
       9.000e+00, 8.500e+01, 1.400e+02])
```

```
sns.boxplot(x=df['Number.of.Bags'])
```

```
<Axes: xlabel='Number.of.Bags'>
```




```
df['Number.of.Bags'].sort_values(ascending=True).head()
```

```
704     0.0
1206    1.0
379     1.0
1188    1.0
444     1.0
Name: Number.of.Bags, dtype: float64
```

```
a=df['Number.of.Bags'].median()
a
```

```
175.0
```

```
df['Number.of.Bags'].fillna(a,inplace=True)
df.isnull().sum()
```

```
Species                   0
Country.of.Origin         0
Number.of.Bags            0
Bag.Weight                0
In.Country.Partner        0
Harvest.Year              0
Grading.Date              0
Processing.Method         0
```

```
Aroma                     0
Flavor                    0
Aftertaste                0
Acidity                   0
Body                      0
Balance                   0
Uniformity                0
Clean.Cup                 0
Sweetness                 0
Cupper.Points             0
Total.Cup.Points          0
Moisture                  0
Category.One.Defects      0
Quakers                   0
Category.Two.Defects      0
Expiration                0
Certification.Body        0
Certification.Address     0
Certification.Contact     0
unit_of_measurement       0
dtype: int64
```

```
#Removing all the objects to only have numerical columns.
df1=df.select_dtypes(exclude=['object'])
df1
```

```
      Number.of.Bags  Aroma  Flavor  Aftertaste  Acidity  Body
Balance  \
0              300.0   8.67    8.83        8.67     8.75  8.50
8.42
1              300.0   8.75    8.67        8.50     8.58  8.42
8.42
2                5.0   8.42    8.50        8.42     8.42  8.33
8.42
3              320.0   8.17    8.58        8.42     8.42  8.50
8.25
4              300.0   8.25    8.50        8.25     8.50  8.42
8.33
...              ...    ...     ...         ...      ...   ...       ..
.
1334             1.0   7.75    7.58        7.33     7.58  5.08
7.83
1335             1.0   7.50    7.67        7.75     7.75  5.17
5.25
1336             1.0   7.33    7.33        7.17     7.42  7.50
7.17
1337             1.0   7.42    6.83        6.75     7.17  7.25
7.00
1338             1.0   6.75    6.67        6.50     6.83  6.92
6.83
```

```
      Uniformity  Clean.Cup  Sweetness  Cupper.Points
Total.Cup.Points  \
0           10.00      10.00      10.00           8.75
90.58
1           10.00      10.00      10.00           8.58
89.92
2           10.00      10.00      10.00           9.25
89.75
3           10.00      10.00      10.00           8.67
89.00
4           10.00      10.00      10.00           8.58
88.83
...           ...        ...        ...            ...                  ..
.
1334        10.00      10.00       7.75           7.83
78.75
1335        10.00      10.00       8.42           8.58
78.08
1336         9.33       9.33       7.42           7.17
77.17
1337         9.33       9.33       7.08           6.92
75.08
1338         9.33       9.33       6.67           7.92
73.75

      Moisture  Category.One.Defects  Quakers  Category.Two.Defects
0         0.12                     0      0.0                     0
1         0.12                     0      0.0                     1
2         0.00                     0      0.0                     0
3         0.11                     0      0.0                     2
4         0.12                     0      0.0                     2
...        ...                   ...      ...                   ...
1334      0.00                     0      0.0                     1
1335      0.00                     0      0.0                     0
1336      0.00                     0      0.0                     6
1337      0.10                    20      0.0                     1
1338      0.12                    63      0.0                     9

[1339 rows x 16 columns]
```

```
q1=df1.quantile(0.25)
q3=df1.quantile(0.75)
q1
```

```
Number.of.Bags          14.00
Aroma                    7.42
Flavor                   7.33
Aftertaste               7.25
Acidity                  7.33
```

```
Body                      7.33
Balance                   7.33
Uniformity               10.00
Clean.Cup                10.00
Sweetness                10.00
Cupper.Points             7.25
Total.Cup.Points         81.08
Moisture                  0.09
Category.One.Defects      0.00
Quakers                   0.00
Category.Two.Defects      0.00
Name: 0.25, dtype: float64
```

```
q3
```

```
Number.of.Bags          275.00
Aroma                     7.75
Flavor                    7.75
Aftertaste                7.58
Acidity                   7.75
Body                      7.67
Balance                   7.75
Uniformity               10.00
Clean.Cup                10.00
Sweetness                10.00
Cupper.Points             7.75
Total.Cup.Points         83.67
Moisture                  0.12
Category.One.Defects      0.00
Quakers                   0.00
Category.Two.Defects      4.00
Name: 0.75, dtype: float64
```

```
iqr=q3-q1
iqr
```

```
Number.of.Bags          261.00
Aroma                     0.33
Flavor                    0.42
Aftertaste                0.33
Acidity                   0.42
Body                      0.34
Balance                   0.42
Uniformity                0.00
Clean.Cup                 0.00
Sweetness                 0.00
Cupper.Points             0.50
Total.Cup.Points          2.59
Moisture                  0.03
Category.One.Defects      0.00
```

```
Quakers                     0.00
Category.Two.Defects        4.00
dtype: float64

#Finding the outliers & storing it in a variable
b=(df1<(q1-1.5*iqr))|(df1>(q3+1.5*iqr))
b
```

```
      Number.of.Bags  Aroma  Flavor  Aftertaste  Acidity   Body
Balance  \
0              False   True    True        True     True   True
True
1              False   True    True        True     True   True
True
2              False   True    True        True     True   True
True
3              False  False    True        True     True   True
False
4              False   True    True        True     True   True
False
...              ...    ...     ...         ...      ...    ...
..
1334           False  False   False       False    False   True
False
1335           False  False   False       False    False   True
True
1336           False  False   False       False    False  False
False
1337           False  False   False        True    False  False
False
1338           False   True    True        True    False  False
False

      Uniformity  Clean.Cup  Sweetness  Cupper.Points
Total.Cup.Points  \
0          False      False      False           True
True
1          False      False      False           True
True
2          False      False      False           True
True
3          False      False      False           True
True
4          False      False      False           True
True
...          ...        ...        ...            ...
.
1334       False      False       True          False
False
1335       False      False       True           True
```

```
False
1336        True       True       True          False
True
1337        True       True       True          False
True
1338        True       True       True          False
True

      Moisture  Category.One.Defects  Quakers  Category.Two.Defects
0        False                 False    False                 False
1        False                 False    False                 False
2         True                 False    False                 False
3        False                 False    False                 False
4        False                 False    False                 False
...        ...                   ...      ...                   ...
1334      True                 False    False                 False
1335      True                 False    False                 False
1336      True                 False    False                 False
1337     False                  True    False                 False
1338     False                  True    False                 False

[1339 rows x 16 columns]
```

```
#Removing the outliers from the original data frame
filter=df[~(b).any(axis=1)]
filter
```

```
      Species Country.of.Origin  Number.of.Bags Bag.Weight  \
21    Arabica        Costa Rica           250.0      3 lbs
30    Arabica         Nicaragua           275.0          6
34    Arabica          Ethiopia           320.0      60 kg
35    Arabica             Kenya           320.0       1 kg
43    Arabica            Taiwan            10.0      15 kg
...       ...               ...             ...        ...
1167  Arabica          Colombia           250.0      70 kg
1182  Arabica            Taiwan            50.0      20 kg
1183  Arabica            Mexico            12.0       1 kg
1205  Arabica            Mexico            14.0       1 kg
1209  Arabica            Mexico            20.0       1 kg

                      In.Country.Partner Harvest.Year
Grading.Date  \
21          Specialty Coffee Association         2014       April 2nd,
2014
30          Specialty Coffee Association         2012        May 18th,
2010
34    METAD Agricultural Development plc         2014      March 27th,
2015
35      Kenya Coffee Traders Association         2013        May 30th,
2014
```

```
43          Specialty Coffee Association         2015       June 10th,
2015
...                                    ...          ...
...
1167                             Almacafé        4T/10    February 9th,
2011
1182        Specialty Coffee Association         2014    November 7th,
2014
1183                              AMECAFE         2012  September 10th,
2012
1205                              AMECAFE         2012  September 17th,
2012
1209                              AMECAFE         2012      August 1st,
2012

              Processing.Method  Aroma  Flavor  ...  Total.Cup.Points
\
21                 Washed / Wet   8.08    8.25  ...             87.17

30                 Washed / Wet   7.92    8.25  ...             86.58

34                Natural / Dry   8.00    8.08  ...             86.25

35                 Washed / Wet   8.08    8.00  ...             86.25

43    Semi-washed / Semi-pulped   8.08    8.17  ...             86.08

...                          ...    ...     ...  ...               ...

1167               Washed / Wet   7.25    7.17  ...             79.58

1182               Washed / Wet   7.08    6.83  ...             79.25

1183               Washed / Wet   7.00    7.00  ...             79.25

1205               Washed / Wet   7.50    7.00  ...             78.92

1209               Washed / Wet   7.25    6.83  ...             78.75


      Moisture  Category.One.Defects  Quakers  Category.Two.Defects  \
21        0.11                     0      0.0                     2
30        0.08                     0      0.0                     2
34        0.10                     0      0.0                     3
35        0.12                     0      0.0                     1
43        0.12                     0      0.0                     0
...        ...                   ...      ...                   ...
1167      0.10                     0      0.0                     4
1182      0.11                     0      0.0                     0
1183      0.13                     0      0.0                    10
```

```
1205      0.16                        0      0.0                        0
1209      0.14                        0      0.0                        0

                Expiration                   Certification.Body  \
21         April 2nd, 2015        Specialty Coffee Association
30           May 18th, 2011       Specialty Coffee Association
34        March 26th, 2016  METAD Agricultural Development plc
35           May 30th, 2015    Kenya Coffee Traders Association
43           June 9th, 2016       Specialty Coffee Association
...                     ...                                ...
1167    February 9th, 2012                          Almacafé
1182    November 7th, 2015       Specialty Coffee Association
1183  September 10th, 2013                           AMECAFE
1205  September 17th, 2013                           AMECAFE
1209      August 1st, 2013                           AMECAFE

                          Certification.Address  \
21    36d0d00a3724338ba7937c52a378d085f2172daa
30    36d0d00a3724338ba7937c52a378d085f2172daa
34    309fcf77415a3661ae83e027f7e5f05dad786e44
35    ccba45b89d859740b749878be8c6d16fbdb96c2e
43    36d0d00a3724338ba7937c52a378d085f2172daa
...                                        ...
1167  e493c36c2d076bf273064f7ac23ad562af257a25
1182  36d0d00a3724338ba7937c52a378d085f2172daa
1183  59e396ad6e22a1c22b248f958e1da2bd8af85272
1205  59e396ad6e22a1c22b248f958e1da2bd8af85272
1209  59e396ad6e22a1c22b248f958e1da2bd8af85272

                          Certification.Contact  unit_of_measurement
21    0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m
30    0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m
34    19fef5a731de2db57d16da10287413f5f99bc2dd                    m
35    d752c909a015f3c76224b3c5cc520f8a67afda74                    m
43    0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m
...                                        ...                  ...
1167  70d3c0c26f89e00fdae6fb39ff54f0d2eb1c38ab                    m
1182  0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m
1183  0eb4ee5b3f47b20b049548a2fd1e7d4a2b70d0a7                    m
1205  0eb4ee5b3f47b20b049548a2fd1e7d4a2b70d0a7                    m
1209  0eb4ee5b3f47b20b049548a2fd1e7d4a2b70d0a7                    m

[592 rows x 28 columns]
```

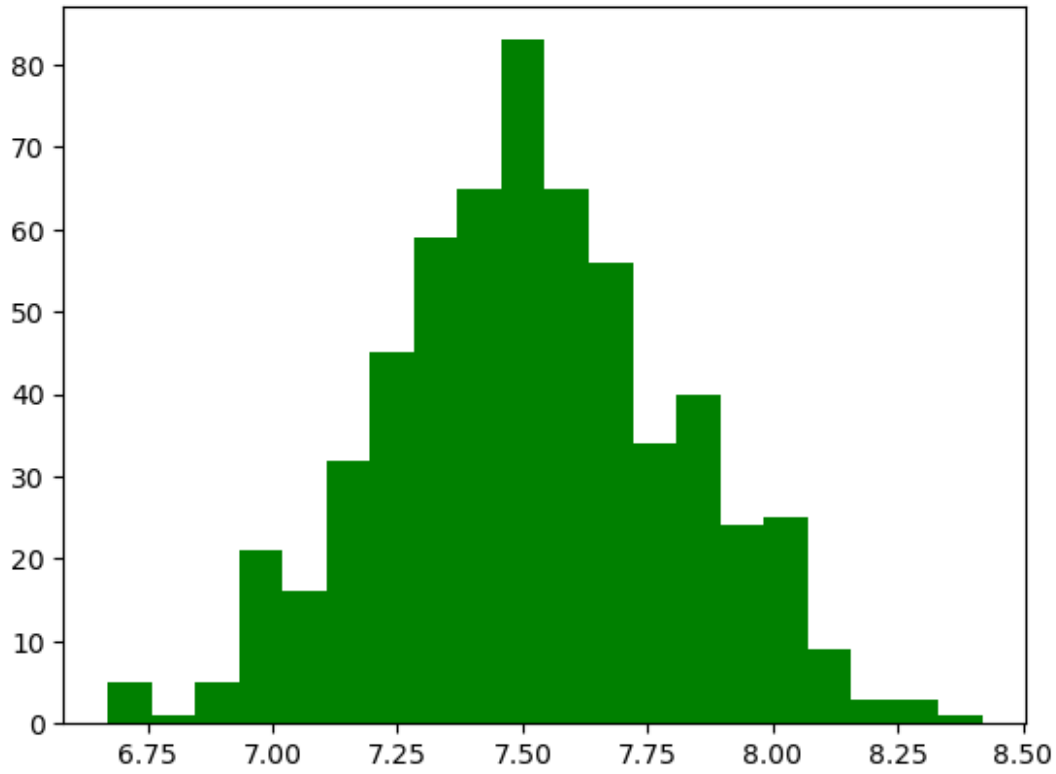
```
filter.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 592 entries, 21 to 1209
Data columns (total 28 columns):
 #   Column                 Non-Null Count  Dtype
```

```
---  ------                  --------------  -----
 0   Species                 592 non-null    object
 1   Country.of.Origin       592 non-null    object
 2   Number.of.Bags          592 non-null    float64
 3   Bag.Weight              592 non-null    object
 4   In.Country.Partner      592 non-null    object
 5   Harvest.Year            592 non-null    object
 6   Grading.Date            592 non-null    object
 7   Processing.Method       592 non-null    object
 8   Aroma                   592 non-null    float64
 9   Flavor                  592 non-null    float64
 10  Aftertaste              592 non-null    float64
 11  Acidity                 592 non-null    float64
 12  Body                    592 non-null    float64
 13  Balance                 592 non-null    float64
 14  Uniformity              592 non-null    float64
 15  Clean.Cup               592 non-null    float64
 16  Sweetness               592 non-null    float64
 17  Cupper.Points           592 non-null    float64
 18  Total.Cup.Points        592 non-null    float64
 19  Moisture                592 non-null    float64
 20  Category.One.Defects    592 non-null    int64
 21  Quakers                 592 non-null    float64
 22  Category.Two.Defects    592 non-null    int64
 23  Expiration              592 non-null    object
 24  Certification.Body      592 non-null    object
 25  Certification.Address   592 non-null    object
 26  Certification.Contact   592 non-null    object
 27  unit_of_measurement     592 non-null    object
dtypes: float64(14), int64(2), object(12)
memory usage: 134.1+ KB
```

```
#Univariate analysis

import matplotlib.pyplot as plt

plt.hist(filter['Cupper.Points'],bins=20,color='green') #histogram to
plot cupper points
```

```
(array([ 5.,  1.,  5., 21., 16., 32., 45., 59., 65., 83., 65., 56.,
34.,
        40., 24., 25.,  9.,  3.,  3.,  1.]),
 array([6.67  , 6.7575, 6.845 , 6.9325, 7.02  , 7.1075, 7.195 ,
7.2825,
        7.37  , 7.4575, 7.545 , 7.6325, 7.72  , 7.8075, 7.895 ,
7.9825,
        8.07  , 8.1575, 8.245 , 8.3325, 8.42  ]),
 <BarContainer object of 20 artists>)
```




```
df.groupby(['Species']).count()
```

```
         Country.of.Origin  Number.of.Bags  Bag.Weight
In.Country.Partner  \
Species

Arabica                1311            1311        1311
1311
Robusta                  28              28          28
28

         Harvest.Year  Grading.Date  Processing.Method  Aroma  Flavor
\
Species

Arabica          1311          1311               1311   1311    1311

Robusta            28            28                 28     28      28


         Aftertaste  ...  Total.Cup.Points  Moisture
Category.One.Defects  \
Species              ...

Arabica        1311  ...              1311      1311
1311
```

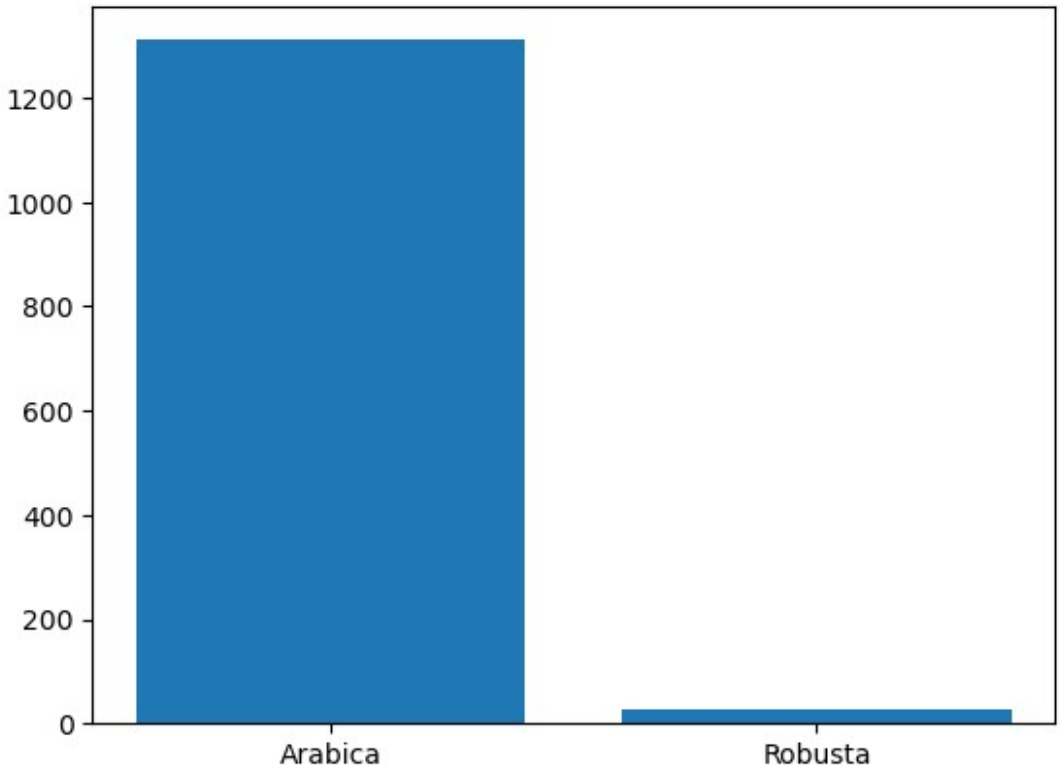
```
Robusta          28  ...                  28        28
28

         Quakers  Category.Two.Defects  Expiration  Certification.Body
\
Species

Arabica     1311                  1311        1311                1311

Robusta       28                    28          28                  28


         Certification.Address  Certification.Contact
unit_of_measurement
Species

Arabica                   1311                   1311
1311
Robusta                     28                     28
28

[2 rows x 27 columns]
```

```
a=df.groupby(['Species']).size().reset_index(name="count").rename(colu
mns={"Species":"sp"})
a
```

```
        sp  count
0  Arabica   1311
1  Robusta     28
```

```
plt.bar(a['sp'],a['count']) #bar graph plotting species
```

```
<BarContainer object of 2 artists>
```




```
#Adding an extra column "count%" to show the percentage of each
species present
a['count%']=a['count']/sum(a['count'])*100
a
```

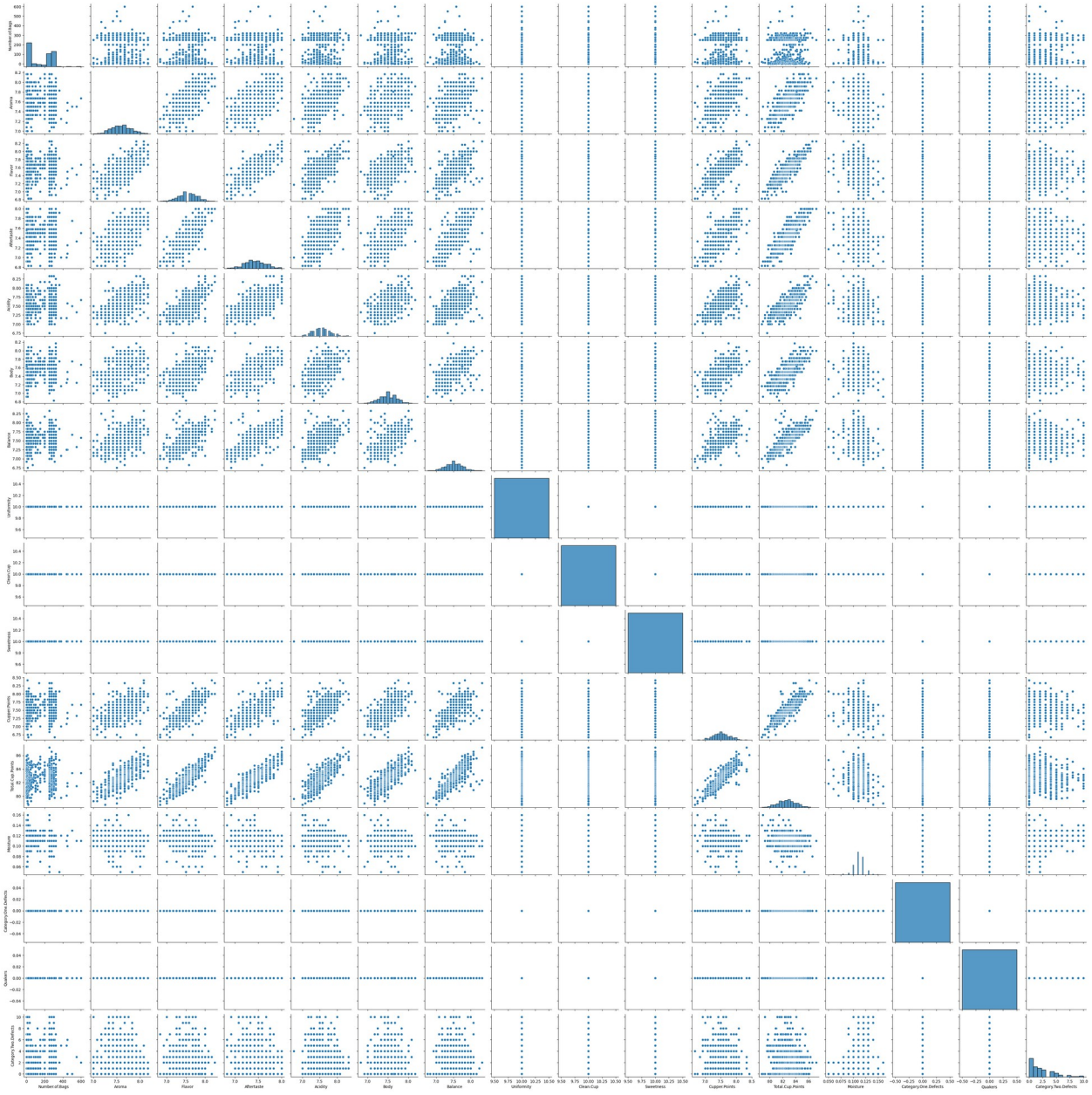
```
        sp  count     count%
0  Arabica   1311  97.908887
1  Robusta     28   2.091113
```

```
#Bivariate analysis
```

A pairplot is typically used to visualize relationships between multiple numerical variables in a dataset by creating scatter plots for each pair of variables. In our case we have only one numerical column, so creating a pairplot doesn't make sense since there are no pairs of variables to plot.

```
a=sns.pairplot(filter)
a
```

```
<seaborn.axisgrid.PairGrid at 0x1dbdfed1c90>
```



```
filter
```

	Species	Country.of.Origin	Number.of.Bags	Bag.Weight	\
21	Arabica	Costa Rica	250.0	3 lbs	
30	Arabica	Nicaragua	275.0	6	
34	Arabica	Ethiopia	320.0	60 kg	
35	Arabica	Kenya	320.0	1 kg	
43	Arabica	Taiwan	10.0	15 kg	
...	...	...	...	...	
1167	Arabica	Colombia	250.0	70 kg	
1182	Arabica	Taiwan	50.0	20 kg	
1183	Arabica	Mexico	12.0	1 kg	

```
1205  Arabica            Mexico             14.0       1 kg
1209  Arabica            Mexico             20.0       1 kg

                     In.Country.Partner Harvest.Year
Grading.Date  \
21          Specialty Coffee Association         2014       April 2nd,
2014
30          Specialty Coffee Association         2012        May 18th,
2010
34    METAD Agricultural Development plc         2014      March 27th,
2015
35      Kenya Coffee Traders Association         2013        May 30th,
2014
43          Specialty Coffee Association         2015       June 10th,
2015
...                                  ...          ...
...
1167                           Almacafé        4T/10    February 9th,
2011
1182        Specialty Coffee Association         2014    November 7th,
2014
1183                            AMECAFE         2012  September 10th,
2012
1205                            AMECAFE         2012  September 17th,
2012
1209                            AMECAFE         2012      August 1st,
2012

              Processing.Method  Aroma  Flavor  ...  Total.Cup.Points
\
21                 Washed / Wet   8.08    8.25  ...             87.17

30                 Washed / Wet   7.92    8.25  ...             86.58

34                Natural / Dry   8.00    8.08  ...             86.25

35                 Washed / Wet   8.08    8.00  ...             86.25

43    Semi-washed / Semi-pulped   8.08    8.17  ...             86.08

...                          ...    ...     ...  ...               ...

1167               Washed / Wet   7.25    7.17  ...             79.58

1182               Washed / Wet   7.08    6.83  ...             79.25

1183               Washed / Wet   7.00    7.00  ...             79.25

1205               Washed / Wet   7.50    7.00  ...             78.92
```

```
1209               Washed / Wet   7.25    6.83  ...             78.75


      Moisture  Category.One.Defects  Quakers  Category.Two.Defects  \
21        0.11                     0      0.0                     2
30        0.08                     0      0.0                     2
34        0.10                     0      0.0                     3
35        0.12                     0      0.0                     1
43        0.12                     0      0.0                     0
...        ...                   ...      ...                   ...
1167      0.10                     0      0.0                     4
1182      0.11                     0      0.0                     0
1183      0.13                     0      0.0                    10
1205      0.16                     0      0.0                     0
1209      0.14                     0      0.0                     0

                Expiration                  Certification.Body  \
21         April 2nd, 2015        Specialty Coffee Association
30          May 18th, 2011        Specialty Coffee Association
34        March 26th, 2016  METAD Agricultural Development plc
35          May 30th, 2015    Kenya Coffee Traders Association
43          June 9th, 2016        Specialty Coffee Association
...                    ...                                 ...
1167   February 9th, 2012                           Almacafé
1182   November 7th, 2015        Specialty Coffee Association
1183  September 10th, 2013                            AMECAFE
1205  September 17th, 2013                            AMECAFE
1209      August 1st, 2013                            AMECAFE

                         Certification.Address  \
21    36d0d00a3724338ba7937c52a378d085f2172daa
30    36d0d00a3724338ba7937c52a378d085f2172daa
34    309fcf77415a3661ae83e027f7e5f05dad786e44
35    ccba45b89d859740b749878be8c6d16fbdb96c2e
43    36d0d00a3724338ba7937c52a378d085f2172daa
...                                       ...
1167  e493c36c2d076bf273064f7ac23ad562af257a25
1182  36d0d00a3724338ba7937c52a378d085f2172daa
1183  59e396ad6e22a1c22b248f958e1da2bd8af85272
1205  59e396ad6e22a1c22b248f958e1da2bd8af85272
1209  59e396ad6e22a1c22b248f958e1da2bd8af85272

                         Certification.Contact  unit_of_measurement
21    0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m
30    0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m
34    19fef5a731de2db57d16da10287413f5f99bc2dd                    m
35    d752c909a015f3c76224b3c5cc520f8a67afda74                    m
43    0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m
...                                        ...                  ...
1167  70d3c0c26f89e00fdae6fb39ff54f0d2eb1c38ab                    m
```

```
1182  0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m
1183  0eb4ee5b3f47b20b049548a2fd1e7d4a2b70d0a7                    m
1205  0eb4ee5b3f47b20b049548a2fd1e7d4a2b70d0a7                    m
1209  0eb4ee5b3f47b20b049548a2fd1e7d4a2b70d0a7                    m

[592 rows x 28 columns]
```

```
filter.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 592 entries, 21 to 1209
Data columns (total 28 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   Species                592 non-null    object 
 1   Country.of.Origin      592 non-null    object 
 2   Number.of.Bags         592 non-null    float64
 3   Bag.Weight             592 non-null    object 
 4   In.Country.Partner     592 non-null    object 
 5   Harvest.Year           592 non-null    object 
 6   Grading.Date           592 non-null    object 
 7   Processing.Method      592 non-null    object 
 8   Aroma                  592 non-null    float64
 9   Flavor                 592 non-null    float64
 10  Aftertaste             592 non-null    float64
 11  Acidity                592 non-null    float64
 12  Body                   592 non-null    float64
 13  Balance                592 non-null    float64
 14  Uniformity             592 non-null    float64
 15  Clean.Cup              592 non-null    float64
 16  Sweetness              592 non-null    float64
 17  Cupper.Points          592 non-null    float64
 18  Total.Cup.Points       592 non-null    float64
 19  Moisture               592 non-null    float64
 20  Category.One.Defects   592 non-null    int64  
 21  Quakers                592 non-null    float64
 22  Category.Two.Defects   592 non-null    int64  
 23  Expiration             592 non-null    object 
 24  Certification.Body     592 non-null    object 
 25  Certification.Address  592 non-null    object 
 26  Certification.Contact  592 non-null    object 
 27  unit_of_measurement    592 non-null    object 
dtypes: float64(14), int64(2), object(12)
memory usage: 134.1+ KB
```

```
a=filter.select_dtypes(exclude=['object'])
a
```

```
      Number.of.Bags  Aroma  Flavor  Aftertaste  Acidity  Body  
Balance  \
```

```
21              250.0   8.08    8.25        8.00     8.17  8.00
8.33
30              275.0   7.92    8.25        8.00     8.33  8.00
8.08
34              320.0   8.00    8.08        7.92     8.00  8.08
8.08
35              320.0   8.08    8.00        8.00     8.25  7.92
7.92
43               10.0   8.08    8.17        7.75     8.08  7.75
7.83
...               ...    ...     ...         ...      ...   ...       ..
.
1167            250.0   7.25    7.17        7.00     6.75  7.17
7.33
1182             50.0   7.08    6.83        6.83     7.25  7.42
7.08
1183             12.0   7.00    7.00        6.92     7.17  7.17
7.08
1205             14.0   7.50    7.00        6.92     7.08  6.92
6.75
1209             20.0   7.25    6.83        6.83     7.00  7.17
7.00

      Uniformity  Clean.Cup  Sweetness  Cupper.Points
Total.Cup.Points  \
21          10.0       10.0       10.0           8.33
87.17
30          10.0       10.0       10.0           8.00
86.58
34          10.0       10.0       10.0           8.08
86.25
35          10.0       10.0       10.0           8.08
86.25
43          10.0       10.0       10.0           8.42
86.08
...          ...        ...        ...            ...                ..
.
1167        10.0       10.0       10.0           6.92
79.58
1182        10.0       10.0       10.0           6.75
79.25
1183        10.0       10.0       10.0           6.92
79.25
1205        10.0       10.0       10.0           6.75
78.92
1209        10.0       10.0       10.0           6.67
78.75

      Moisture  Category.One.Defects  Quakers  Category.Two.Defects
```

```
21         0.11                        0       0.0                       2
30         0.08                        0       0.0                       2
34         0.10                        0       0.0                       3
35         0.12                        0       0.0                       1
43         0.12                        0       0.0                       0
...         ...                      ...       ...                     ...
1167       0.10                        0       0.0                       4
1182       0.11                        0       0.0                       0
1183       0.13                        0       0.0                      10
1205       0.16                        0       0.0                       0
1209       0.14                        0       0.0                       0

[592 rows x 16 columns]
```

```
#Dropping columns that has only a single value repeating
b=a.drop(columns=['Uniformity','Clean.Cup','Sweetness','Category.One.D
efects','Category.One.Defects','Quakers'])
b
```

```
      Number.of.Bags  Aroma  Flavor  Aftertaste  Acidity  Body
Balance  \
21             250.0   8.08    8.25        8.00     8.17  8.00
8.33
30             275.0   7.92    8.25        8.00     8.33  8.00
8.08
34             320.0   8.00    8.08        7.92     8.00  8.08
8.08
35             320.0   8.08    8.00        8.00     8.25  7.92
7.92
43              10.0   8.08    8.17        7.75     8.08  7.75
7.83
...              ...    ...     ...         ...      ...   ...      ..
.
1167           250.0   7.25    7.17        7.00     6.75  7.17
7.33
1182            50.0   7.08    6.83        6.83     7.25  7.42
7.08
1183            12.0   7.00    7.00        6.92     7.17  7.17
7.08
1205            14.0   7.50    7.00        6.92     7.08  6.92
6.75
1209            20.0   7.25    6.83        6.83     7.00  7.17
7.00

      Cupper.Points  Total.Cup.Points  Moisture  Category.Two.Defects

21             8.33             87.17      0.11                     2

30             8.00             86.58      0.08                     2
```

```
34              8.08              86.25      0.10                      3
35              8.08              86.25      0.12                      1
43              8.42              86.08      0.12                      0
...              ...                ...       ...                    ...
1167            6.92              79.58      0.10                      4
1182            6.75              79.25      0.11                      0
1183            6.92              79.25      0.13                     10
1205            6.75              78.92      0.16                      0
1209            6.67              78.75      0.14                      0

[592 rows x 11 columns]
```

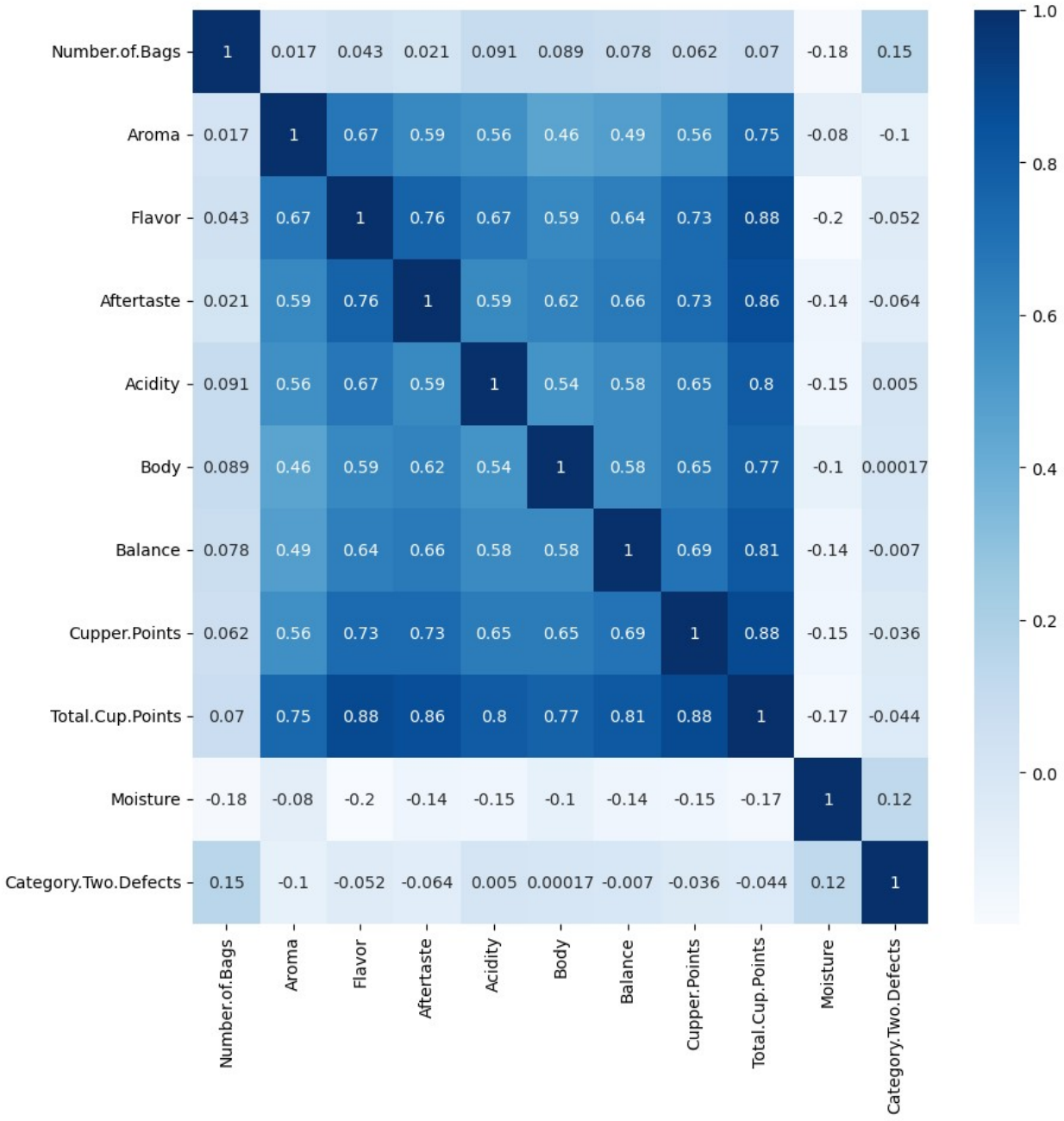
```
#Calculates the correlation coefficients between columns, measuring the strength of linear relationships.
b.corr()
```

```
                      Number.of.Bags     Aroma    Flavor  Aftertaste  \
Number.of.Bags              1.000000  0.016714  0.043065    0.021425
Aroma                       0.016714  1.000000  0.670133    0.593727
Flavor                      0.043065  0.670133  1.000000    0.761918
Aftertaste                  0.021425  0.593727  0.761918    1.000000
Acidity                     0.091179  0.558948  0.671595    0.585304
Body                        0.089487  0.456227  0.594085    0.618701
Balance                     0.077773  0.488713  0.635438    0.662380
Cupper.Points               0.061557  0.556487  0.730047    0.730731
Total.Cup.Points            0.069869  0.746388  0.880839    0.861924
Moisture                   -0.180161 -0.080026 -0.198329   -0.143079
Category.Two.Defects        0.152571 -0.101692 -0.051657   -0.064072

                       Acidity      Body   Balance  Cupper.Points  \
Number.of.Bags        0.091179  0.089487  0.077773       0.061557
```

```
Aroma                    0.558948  0.456227  0.488713        0.556487
Flavor                   0.671595  0.594085  0.635438        0.730047
Aftertaste               0.585304  0.618701  0.662380        0.730731
Acidity                  1.000000  0.537732  0.582737        0.649262
Body                     0.537732  1.000000  0.580287        0.649705
Balance                  0.582737  0.580287  1.000000        0.688170
Cupper.Points            0.649262  0.649705  0.688170        1.000000
Total.Cup.Points         0.799773  0.766142  0.809651        0.878995
Moisture                -0.154790 -0.102323 -0.142152       -0.154504
Category.Two.Defects     0.004964  0.000167 -0.006968       -0.035668
```

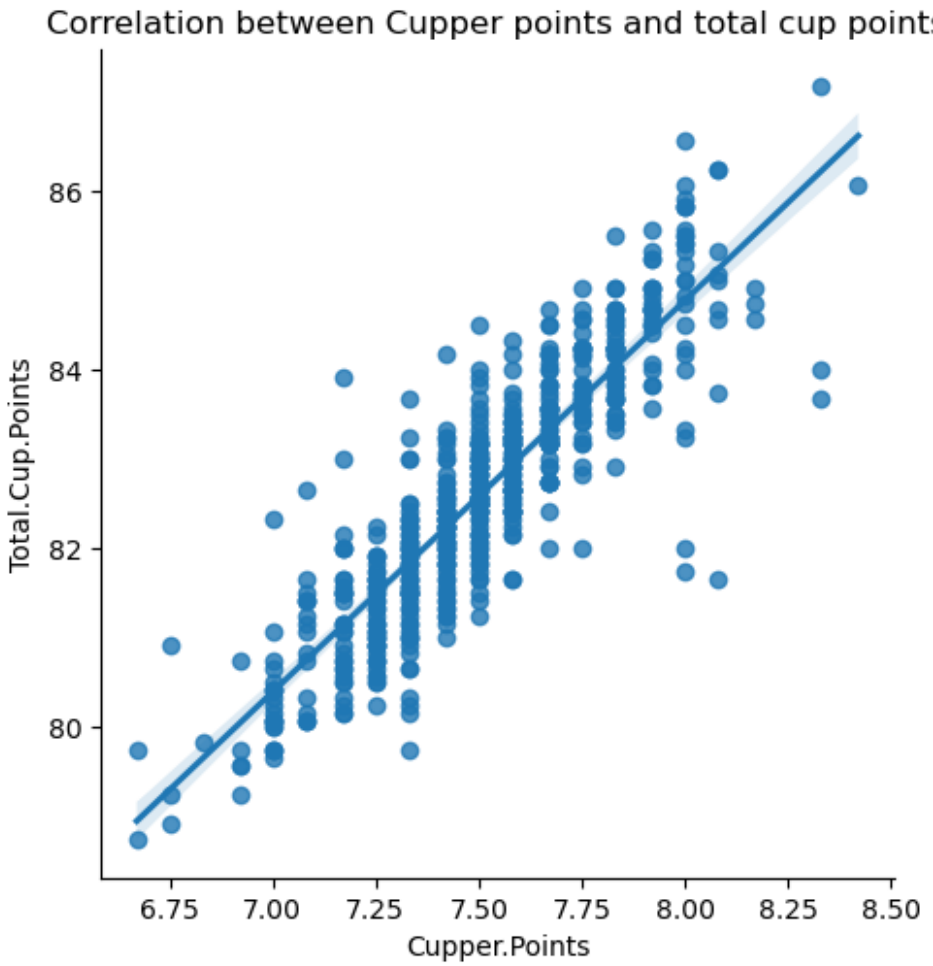
	Total.Cup.Points	Moisture	Category.Two.Defects
Number.of.Bags	0.069869	-0.180161	0.152571
Aroma	0.746388	-0.080026	-0.101692
Flavor	0.880839	-0.198329	-0.051657
Aftertaste	0.861924	-0.143079	-0.064072
Acidity	0.799773	-0.154790	0.004964
Body	0.766142	-0.102323	0.000167
Balance	0.809651	-0.142152	-0.006968
Cupper.Points	0.878995	-0.154504	-0.035668
Total.Cup.Points	1.000000	-0.170731	-0.043774
Moisture	-0.170731	1.000000	0.124455
Category.Two.Defects	-0.043774	0.124455	1.000000

```
#Plotting a heatmap for better visualization of the correlation
matrix, to get an intuitive representation of relationships between
variables
plt.figure(figsize=(10,10))
sns.heatmap(b.corr() ,annot=True,cmap='Blues')
plt.show()
```



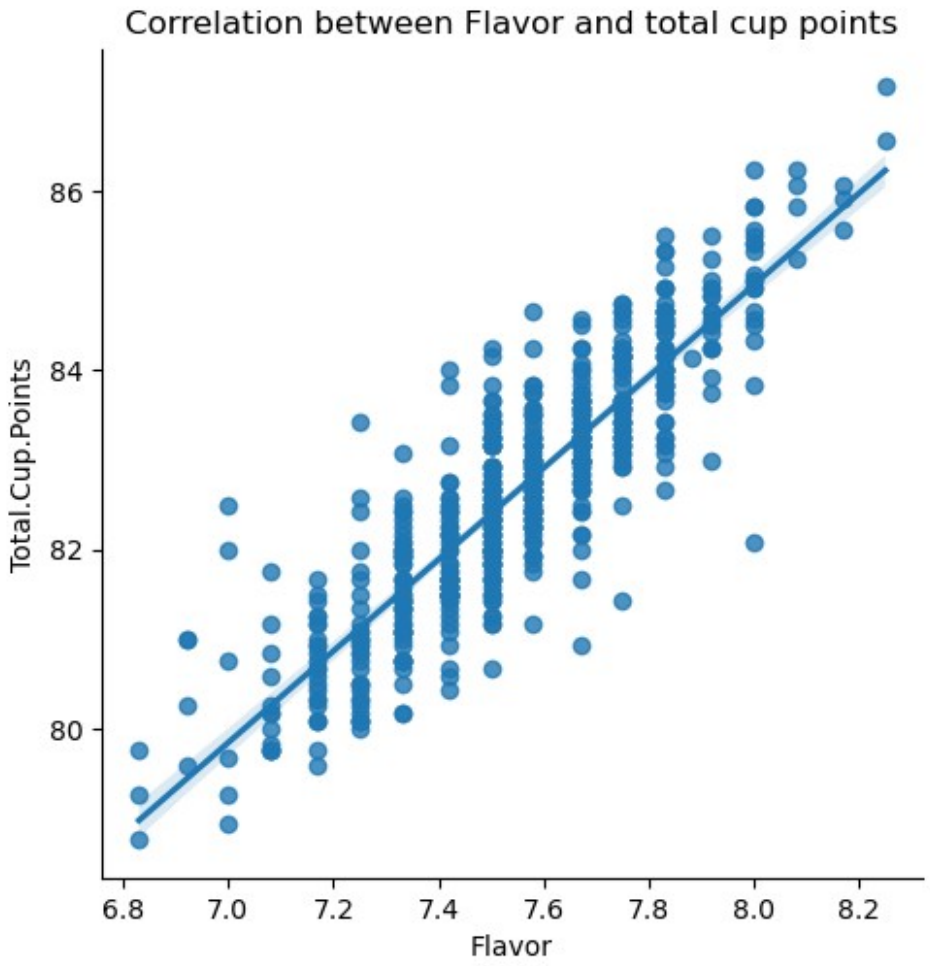

```
sns.lmplot(x='Cupper.Points',y='Total.Cup.Points',data=b)
plt.title("Correlation between Cupper points and total cup points")
```

```
Text(0.5, 1.0, 'Correlation between Cupper points and total cup
points')
```




```
sns.lmplot(x='Flavor',y='Total.Cup.Points',data=b)
plt.title("Correlation between Flavor and total cup points")
```

```
Text(0.5, 1.0, 'Correlation between Flavor and total cup points')
```




```
#Assign the independent variables to x and the dependent variable
(total cup points) to y for performing ANOVA test
x=filter[['Flavor','Acidity','Aftertaste','Cupper.Points']]
y=filter['Total.Cup.Points']
x
```

	Flavor	Acidity	Aftertaste	Cupper.Points
21	8.25	8.17	8.00	8.33
30	8.25	8.33	8.00	8.00
34	8.08	8.00	7.92	8.08
35	8.00	8.25	8.00	8.08
43	8.17	8.08	7.75	8.42
...	...	...	...	...
1167	7.17	6.75	7.00	6.92
1182	6.83	7.25	6.83	6.75
1183	7.00	7.17	6.92	6.92
1205	7.00	7.08	6.92	6.75
1209	6.83	7.00	6.83	6.67

```
[592 rows x 4 columns]
```

```
y
```

```
21      87.17
30      86.58
34      86.25
35      86.25
43      86.08
        ...
1167    79.58
1182    79.25
1183    79.25
1205    78.92
1209    78.75
Name: Total.Cup.Points, Length: 592, dtype: float64
```

```
from sklearn.feature_selection import f_classif
e=f_classif(x,y)
e
```

```
(array([25.53697691, 14.16599345, 21.39500164, 25.4364055 ]),
 array([1.03140812e-134, 2.76042872e-088, 6.44670987e-120,
2.24140767e-134]))
```

```
#So we can see that Quality of the coffee is more dependable on the
Flavour and Cupper p
```