TITLE: Coffee Quality Database Documentation

Introduction

The Coffee Quality Database aims to provide a comprehensive analysis of various coffee samples from around the world. This dataset contains detailed information on coffee characteristics, including the species, farm details, and quality metrics. The primary objective is to facilitate research and insights into factors influencing coffee quality, such as geographic origin, altitude, and processing methods.

Aim

The aim of this dataset is to analyze and understand the factors that influence coffee quality, such as geographical location, altitude, and farm practices. This data can help improve coffee cultivation and processing methods, benefiting researchers, producers, and coffee enthusiasts.

Business Problem / Problem Statement

The coffee industry faces challenges in maintaining and improving the quality of coffee due to varying factors like climate, altitude, and farming practices. This dataset aims to provide insights into these factors, helping stakeholders identify key areas for improvement to enhance coffee quality consistently.

Project Workflow

The project workflow involves several key steps:

1. **Data Collection**: Gathering the coffee production dataset.
2. **Data Understanding**: Exploring and summarizing the dataset.
3. **Data Cleaning**: Addressing missing values, outliers, and inconsistencies.
4. **Data Transformation**: Creating derived metrics and filtering data for analysis.
5. **Exploratory Data Analysis (EDA)**: Conducting univariate, bivariate, and multivariate analyses.
6. **Insights and Conclusions**: Summarizing findings and providing recommendations.

Data Understanding

The dataset contains various columns such as `Species`, `Owner`, `Country.of.Origin`, `Farm.Name`, `Lot.Number`, `Mill`, `ICO.Number`, `Company`, `Altitude`,`Region`,`Producer`,`Number.of.Bags`,`Bag.Weight`,`In.Country.Partner`,`Harvest.Year`,`Grading.Date`,`Owner.1`,`Variety`,`Processing.Method`, `Aroma`, `Flavor`, `Aftertaste`, `Acidity`, `Body`, `Balance`, `Uniformity`, `Clean.Cup`, `Sweetness`,

`Cupper.Points`, `Total.Cup.Points`, `Moisture`, `Category.One.Defects`, `Quakers`, `Color`, `Category.Two.Defects`, `Expiration`, `Certification.Body`, `Certification.Address`, `Certification.Contact`, `unit_of_measurement`, `altitude_low_meters`, `altitude_high_meters`, and `altitude_mean_meters`.

Column Descriptions

- **Unnamed: 0**: Index of the dataset.
- **Species**: Type of coffee (e.g., Arabica, Robusta).
- **Owner**: Owner of the coffee sample.
- **Country.of.Origin**: Country where the coffee was grown.
- **Farm.Name**: Name of the coffee farm.
- **Lot.Number**: Lot number of the coffee sample.
- **Mill**: Name of the mill where the coffee was processed.
- **ICO.Number**: International Coffee Organization number.
- **Company**: Company associated with the coffee sample.
- **Altitude**: Altitude range of the farm (meters).
- **Color**: Color of the coffee beans.
- **Category.Two.Defects**: Number of category two defects in the sample.
- **Expiration**: Expiration date of the certification.
- **Certification.Body**: Organization that certified the coffee.
- **Certification.Address**: Address of the certification body.
- **Certification.Contact**: Contact information for the certification body.
- **unit_of_measurement**: Unit of measurement for altitude.
- **altitude_low_meters**: Lowest altitude of the farm (meters).
- **altitude_high_meters**: Highest altitude of the farm (meters).
- **altitude_mean_meters**: Mean altitude of the farm (meters).

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

df=pd.read_csv("E:/kgm/coffeeQuality.csv")
df
```

```
      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]
```

```
print("Number of Rows: ",df.shape[0])
print("Number of Columns: ",df.shape[1])
```

```
Number of Rows:  1339
Number of Columns:  44
```

```
print("Data types: ", df.dtypes)
```

```
Data types:  Unnamed: 0                int64
Species                   object
Owner                     object
Country.of.Origin         object
Farm.Name                 object
Lot.Number                object
Mill                      object
ICO.Number                object
Company                   object
Altitude                  object
Region                    object
Producer                  object
```

```
Number.of.Bags            float64
Bag.Weight                 object
In.Country.Partner         object
Harvest.Year               object
Grading.Date               object
Owner.1                    object
Variety                    object
Processing.Method          object
Aroma                     float64
Flavor                    float64
Aftertaste                float64
Acidity                   float64
Body                      float64
Balance                   float64
Uniformity                float64
Clean.Cup                 float64
Sweetness                 float64
Cupper.Points             float64
Total.Cup.Points          float64
Moisture                  float64
Category.One.Defects        int64
Quakers                   float64
Color                      object
Category.Two.Defects        int64
Expiration                 object
Certification.Body         object
Certification.Address      object
Certification.Contact      object
unit_of_measurement        object
altitude_low_meters       float64
altitude_high_meters      float64
altitude_mean_meters      float64
dtype: object
```

The dataset comprises 1,339 entries, each representing a coffee sample with attributes such as species, owner, country of origin, and various quality metrics. The dataset includes both quantitative and qualitative data, offering a comprehensive view of coffee characteristics.

```
df = df.drop(columns=['Unnamed: 0'])
df
```

```
      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]
```

Summary Statistics and Insights

- **Numerical Columns**: For numerical columns such as `Number.of.Bags`, `Aroma`, `Flavor`, `Aftertaste`, `Acidity`, `Body`, `Balance`, `Uniformity`, `Clean.Cup`, `Sweetness`, `Cupper.Points`, `Total.Cup.Points`, `Moisture`, `Category.One.Defects`, `Quakers`, `Category.Two.Defects`, `altitude_low_meters`, `altitude_high_meters`, and `altitude_mean_meters`, summary statistics like mean, median, standard deviation, and range will be calculated.
- **Categorical Columns**: For categorical columns such as `Species`, `Owner`, `Country.of.Origin`, `Farm.Name`, `Lot.Number`, `Mill`, `ICO.Number`, `Company`, `Altitude`, `Region`, `Producer`, `Bag.Weight`, `In.Country.Partner`, `Harvest.Year`, `Grading.Date`, `Owner.1`, `Variety`, `Processing.Method`, `Color`, `Expiration`, `Certification.Body`, `Certification.Address`, `Certification.Contact`, and `unit_of_measurement`, frequency counts and mode will be analyzed.

```
df.describe()
```

```
       Number.of.Bags        Aroma       Flavor   Aftertaste
Acidity  \
count     1338.000000  1339.000000  1339.000000  1339.000000
1339.000000
mean       159.085202     7.770187     7.520426     7.401083
7.535706
std        173.698167     5.534440     0.398442     0.404463
0.379827
min          0.000000     0.000000     0.000000     0.000000
0.000000
25%         14.000000     7.420000     7.330000     7.250000
7.330000
50%        175.000000     7.580000     7.580000     7.420000
7.580000
75%        275.000000     7.750000     7.750000     7.580000
7.750000
max       3200.000000   200.000000     8.830000     8.670000
8.750000

              Body      Balance   Uniformity    Clean.Cup    Sweetness
\
count  1339.000000  1339.000000  1339.000000  1339.000000  1339.000000

mean      7.517498     7.518013     9.834877     9.835108     9.856692

std       0.370064     0.408943     0.554591     0.763946     0.616102

min       0.000000     0.000000     0.000000     0.000000     0.000000

25%       7.330000     7.330000    10.000000    10.000000    10.000000

50%       7.500000     7.500000    10.000000    10.000000    10.000000

75%       7.670000     7.750000    10.000000    10.000000    10.000000

max       8.580000     8.750000    10.000000    10.000000    10.000000


       Cupper.Points  Total.Cup.Points     Moisture
Category.One.Defects  \
count    1339.000000       1339.000000  1339.000000
1339.000000
mean        7.503376         82.089851     0.088379
0.479462
std         0.473464          3.500575     0.048287
2.549683
min         0.000000          0.000000     0.000000
0.000000
```

```
25%          7.250000          81.080000     0.090000
0.000000
50%          7.500000          82.500000     0.110000
0.000000
75%          7.750000          83.670000     0.120000
0.000000
max         10.000000          90.580000     0.280000
63.000000

            Quakers  Category.Two.Defects  altitude_low_meters  \
count  1338.000000           1339.000000          1109.000000
mean      0.173393              3.556385          1750.713315
std       0.832121              5.312541          8669.440545
min       0.000000              0.000000             1.000000
25%       0.000000              0.000000          1100.000000
50%       0.000000              2.000000          1310.640000
75%       0.000000              4.000000          1600.000000
max      11.000000             55.000000        190164.000000

       altitude_high_meters  altitude_mean_meters
count           1109.000000           1109.000000
mean            1799.347775           1775.030545
std             8668.805771           8668.626080
min                1.000000              1.000000
25%             1100.000000           1100.000000
50%             1350.000000           1310.640000
75%             1650.000000           1600.000000
max           190164.000000         190164.000000
```

Data Cleaning

Data cleaning involves several steps:

- **Missing Values Imputation**: Handling missing values by imputing with mean, median, or mode, or removing rows with significant missing data.
- **Outliers**: Identifying and treating outliers using statistical methods or domain knowledge.
- **Inconsistent Values**: Standardizing text entries and correcting inconsistencies in categorical data.

```
# finding missing values in the data set
df.isnull()

      Species  Owner  Country.of.Origin  Farm.Name  Lot.Number   Mill
\
0       False  False              False      False        True  False

1       False  False              False      False        True  False

2       False  False              False      False        True   True
```

```
3       False  False               False      False        True  False

4       False  False               False      False        True  False

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

1334    False  False               False      False       False  False

1335    False  False               False      False       False  False

1336    False  False               False      False        True  False

1337    False  False               False       True        True   True

1338    False  False               False       True        True   True


      ICO.Number  Company  Altitude  Region  ...  Color
Category.Two.Defects  \
0          False    False     False   False  ...  False
False
1          False    False     False   False  ...  False
False
2           True     True     False    True  ...   True
False
3           True    False     False   False  ...  False
False
4          False    False     False   False  ...  False
False
...          ...      ...       ...     ...  ...    ...
...
1334        True    False      True   False  ...  False
False
1335        True    False     False   False  ...  False
False
1336        True    False     False   False  ...   True
False
1337       False    False      True    True  ...  False
False
1338        True    False      True    True  ...   True
False

      Expiration  Certification.Body  Certification.Address  \
0          False               False                  False
1          False               False                  False
2          False               False                  False
3          False               False                  False
4          False               False                  False
...          ...                 ...                    ...
```

```
1334       False                False                   False
1335       False                False                   False
1336       False                False                   False
1337       False                False                   False
1338       False                False                   False

      Certification.Contact  unit_of_measurement  altitude_low_meters
\
0                     False                False                False

1                     False                False                False

2                     False                False                False

3                     False                False                False

4                     False                False                False

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

1334                  False                False                 True

1335                  False                False                False

1336                  False                False                False

1337                  False                False                 True

1338                  False                False                 True


      altitude_high_meters  altitude_mean_meters
0                    False                 False
1                    False                 False
2                    False                 False
3                    False                 False
4                    False                 False
...                    ...                   ...
1334                  True                  True
1335                 False                 False
1336                 False                 False
1337                  True                  True
1338                  True                  True

[1339 rows x 43 columns]

df.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
```

```
#missing value for 'Lot.Number is 1063' so dropping this column

data=df.drop(columns=['Lot.Number','Harvest.Year'])
data
```

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

```
0                       guji-hambela                                  METAD
PLC
1                       guji-hambela                                  METAD
PLC
2                                NaN                                      
NaN
3                            oromia  Yidnekachew Dabessa Coffee          
Plantation
4                       guji-hambela                                  METAD
PLC
...                              ...                                     
...
1334               san juan, playas          Café Robusta del Ecuador    
S.A.
1335               san juan, playas          Café Robusta del Ecuador    
S.A.
1336  kwanza norte province, angola                           Cafe       
Cazengo
1337                             NaN                                     
NaN
1338                             NaN                                     
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 41 columns]
```

```
data.isnull().sum()
```

```
Species                     0
Owner                       7
Country.of.Origin           1
Farm.Name                 359
Mill                      318
ICO.Number                159
Company                   209
Altitude                  226
Region                     59
Producer                  232
Number.of.Bags              1
Bag.Weight                  0
In.Country.Partner          0
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
```

```
numerical_columns = ['Altitude', 'altitude_low_meters',
'altitude_high_meters', 'altitude_mean_meters', 'Number.of.Bags']

for col in numerical_columns:
    data[col] = pd.to_numeric(data[col], errors='coerce')
```

```
for i in numerical_columns:
    data[i].fillna(data[i].median(), inplace=True)

data.isnull().sum()
```

```
Species                     0
Owner                       7
Country.of.Origin           1
Farm.Name                 359
Mill                      318
ICO.Number                159
Company                   209
Altitude                    0
Region                     59
Producer                  232
Number.of.Bags              0
Bag.Weight                  0
In.Country.Partner          0
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         0
altitude_high_meters        0
altitude_mean_meters        0
dtype: int64
```

```
categorical_columns = ['Owner', 'Country.of.Origin', 'Farm.Name',
'Mill', 'ICO.Number', 'Company', 'Region', 'Producer', 'Variety',
'Processing.Method', 'Color']
for col in categorical_columns:
    data[col].fillna(data[col].mode()[0], inplace=True)


data.isnull().sum()
```

```
Species                    0
Owner                      0
Country.of.Origin          0
Farm.Name                  0
Mill                       0
ICO.Number                 0
Company                    0
Altitude                   0
Region                     0
Producer                   0
Number.of.Bags             0
Bag.Weight                 0
In.Country.Partner         0
Grading.Date               0
Owner.1                    7
Variety                    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
Color                      0
Category.Two.Defects       0
Expiration                 0
Certification.Body         0
Certification.Address      0
Certification.Contact      0
unit_of_measurement        0
altitude_low_meters        0
altitude_high_meters       0
```

```
altitude_mean_meters     0
dtype: int64
```

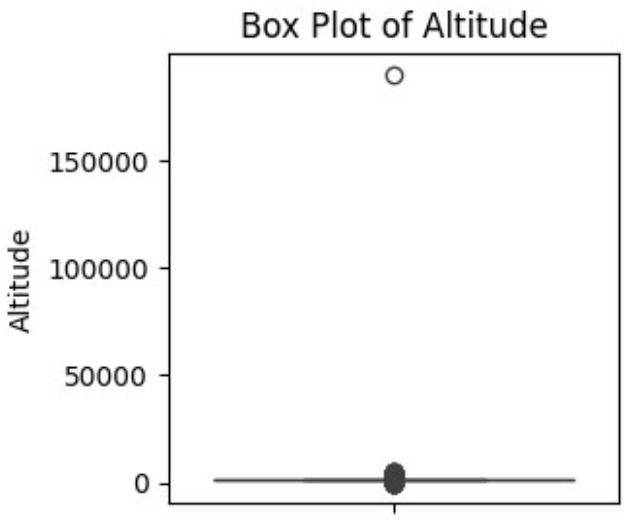
```
data["Owner.1"].unique()
```

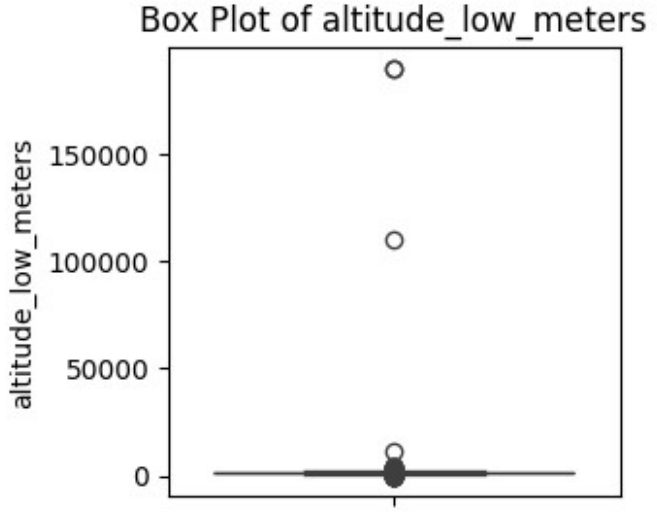
```
array(['metad plc', 'Grounds for Health Admin', 'Yidnekachew Dabessa',
       'Ji-Ae Ahn', 'Hugo Valdivia', 'Ethiopia Commodity Exchange',
       'Diamond Enterprise Plc', 'Mohammed Lalo',
       'CQI Q Coffee Sample Representative', 'Yunnan Coffee Exchange',
       'EssenceCoffee', 'The Coffee Source Inc.', 'ROBERTO LICONA
FRANCO',
       'NUCOFFEE', 'Kabum Trading company', 'Bismarck Castro',
       'Lin, Che-Hao Krude 林哲豪', 'Nora Zeas', 'Specialty Coffee-
Korea',
       'Francisco A Mena', 'Hider Abamecha', 'Daniel Magu',
       'Kona Pacific Farmers Cooperative', 'ITDP International',
       'Jacques Pereira Carneiro', 'Jungle Estate',
       'Great Lakes Coffee Uganda', 'LUSSO LAB', 'AFCA',
       'Juan Luis Alvarado Romero', 'Kawacom Uganda LTD',
       'Exportadora de Cafe Condor S.A', 'Gonzalo Hernandez',
       'Ibrahim Hussien Speciality Coffee Producer &Export',
       'SEID DAMTEW COFFEE PLANATAION', 'Dane Loraas',
       'Colbran Coffeelands, Ltd.', 'Atlantic Specialty Coffee',
       'Assefa Belay Coffee Producer', 'Kyagalanyi Ltd',
       'RASHID MOLEDINA & CO. (MSA) LTD.', 'Ibero Kenya Limited',
       'Compañia Colombiana Agroindustrial S.A',
       'Nomura Trading Co., Ltd.', 'CARCAFE LTDA CI', 'Steven Kil',
       'Eileen Koyanagi', 'Kyagalanyi Coffee Ltd', 'Racafe & Cia
S.C.A',
       'Troy Quimby', 'El Equimite, Cafetal Biodinámico', 'SIMON
MAHINDA',
       'Young Kim', 'Carl Walker', 'Taylor Winch (T) Ltd',
       'ARTEMIO ZAPATA TEJEDA', 'Brian Speckman', 'Philip Schluter',
       '松澤宏樹Koju Matsuzawa', 'Lydiah Mwangi', 'CADEXSA',
       'Consejo Salvadoreño del Café', 'SanJava Coffee', 'Rodrigo
Soto',
       'Fabian Calderon Mora', 'Eric Thormaehlen', 'Rob Tuttle',
       'CQI Taiwan ICP CQI台灣合作夥伴', 'Dream Together',
       'ORGANIZACIONES DE PRODUCTORES DE CAFE COLIMENSE',
       'Benjamin Schmerler', 'Taylor Winch (Coffee) Ltd.', 'Max
Gurdian',
       'ECOM Japan Limited', 'Federacion Nacional de Cafeteros',
       'Eric Wu', 'MARIA IMELDA USCANGA MARTINEZ', 'ALFREDO BOJALIL',
       'Daniel Friedlander', 'Alexandra Katona-Carroll',
       'Aulia Arif Syahri', 'Kao Ming Lee',
       'MARIA AMALIA GUADALUPE TORIELLO ELORZA', 'Raúl Vargas',
       'VICTOR HUGO MELCHOR CORDOVA', 'Tembo Coffee Company Ltd',
       'JESUS SALAZAR VELASCO', 'MANUEL HERRERA JUAREZ', 'Wayner
Jimenez',
       'COOPERATIVA EL GORRION R.L', 'Cafebras', 'CECA,S.A.',
```

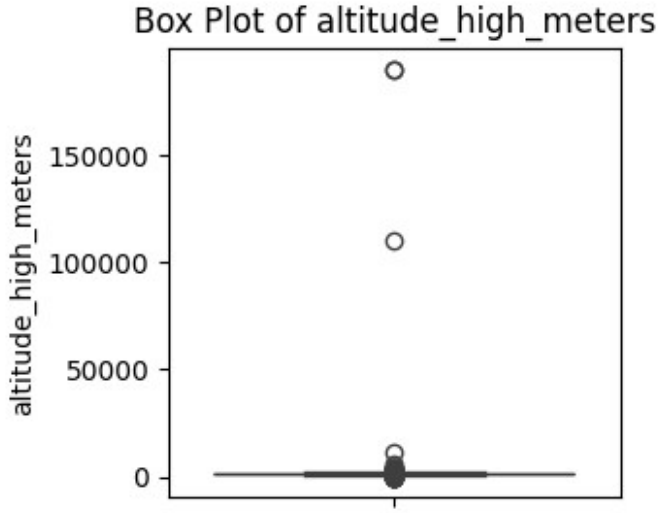
```
       'Asefa Dukamo Keroma', 'Selian Coffee Estate',
       'Olam Agro Colombia', 'Chris Finch', 'ITOCHU Corporation',
       'Owen Carver', 'PT.ROYAL PACIFIC INDAH INTERNATIONAL',
       'ANDRES MARTINEZ LEON', 'Amanda Powers', 'Ipanema Coffees',
       'Doi Tung Development Project', 'CAFES TOMARI SA DE CV',
       'Sarawut Premjit', 'ALMACAFE', 'OSCAR ORTEGA CARBALLO',
       'CECA, S.A.', 'yasmin Cofffee Plantation Plc', 'Garet Alban',
       'FILEMON MENDOZA CAMPOS', 'Doi Chaang Coffee Company',
       'Kennedy Macharia', 'Nile Highland Arabica Coffee Farmers',
       'German Negron', 'SAUL M. HERNANDEZ RAMIREZ',
       'COMERCIAL INTERNACIONAL EXPORTADORA, S.A.', 'Rob Stephen',
       'JUAN LUIS ORTEGA CARBALLO', 'EKAI International Company Ltd.',
       'ANDREAS KUSSMAUL', 'Bulamburi coffee farmers association',
       'Damari Absalome', 'Debesa Agro Industry Plc', nan,
       'MIGUEL CORTES MORENO', 'GABRIEL BERNARDO RIVAS ROSS',
       'Felipe Isaza', 'Specialty Coffee Association of Indonesia',
       'Bugisu Cooperative Union', 'BOURBON SPECIALTY COFFEES',
       'Ngila Estate Ltd', 'Federación Nacional de Cafeteros',
       'J.ANDRADE', 'ITIAH COFFEE LLC',
       'CAFE DE DON BALBINO S.C. DE R.L. DE C.V.',
       'PRODUCTOS Y SERVICIOS CHILINDRON S.A. DE C.V.',
       'CALIXTO GUILLEN VAZQUEZ', 'ERNESTO RODRIGUEZ LUNA',
       'MODESTO LANDEROS FLORES', 'ANDREA BERNAL', 'Sunvirtue Co.,
Ltd.',
       'Tutunze Kahawa Ltd', 'Cafe Politico', 'Mayra Yessenia Torres',
       'Balam Hinyula', 'NESTOR MENDEZ GOMEZ',
       'FERNANDO MENDOZA APARICIO',
       'MARIA LUISA DEL CARMEN ROJAS NARVAEZ', 'UCFA',
       'Irene Alves Santos', 'Star Cafe Ltd',
       'ROSA AURORA FALCON FERNANDEZ', 'SANTIAGO SOLIS AYERDI',
       'Renee A. Perrine', 'Zarah Zamora Perez', 'Andrew Bowman',
       'Expocaccer Coop dos Cafeic do Cerrado Ltda',
       'Nyapea coffee farmers association', 'MARIA GUADALUPE GOMEZ
ANZO',
       'Royal Base Corporation', 'VERONICA LOPEZ CASTILLEJOS',
       'Samuel Muhirwa', 'Joshua Marsceau', 'Coffeebythebag.com ,
INC',
       'Edwin Agasso', 'ARMANDO LUIS POHLENZ MARTINEZ', 'Coffee
Export',
       'SERGIO DE LA VEQUIA BERNARDI', 'ROMULO BELLO FLORES',
       'Rachel Peterson', 'José Luis Rojas Yeo', 'Nitin Coffee
Estate',
       'Adam Kline', 'MONTEGRANDE',
       'GRUPO CAFETALERO LOS BRUJOS SPR DE RL', 'George A. Fernandez',
       'Gabriel Barbara', 'Andry Simarmata', 'Brent Hall',
       'GUILLERMO ROJAS SALDANA', 'Elsy Reyes', 'Shah Plantations',
       'Amkeni Gourmet Coffee Group', 'ENRIQUE MITRE LOPEZ',
       'Enrique Eduardo Lopez Aguilar', 'Brian Beck',
       'Gladness Obed Pallangyo', 'DARIO CESAR GALEANA SANCHEZ',
```

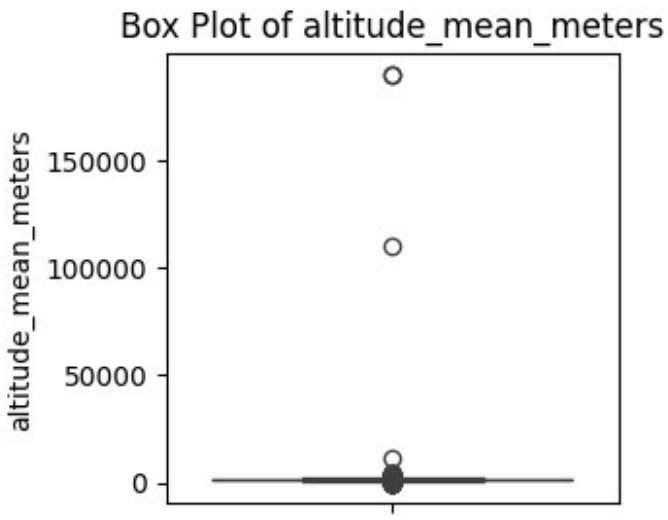
```
       'JOSE DANIEL COBILT CASTRO', 'ALVARO QUIROS PEREZ',
       'OLIVIA HERNANDEZ VIRVES', 'FINCA LAS NIEVES',
       'Pedro Santos e Silva', 'Michael Gavina', 'KlemOrganics',
       'JESUS CARLOS CARDENAS VALDIVIA', 'BENCAFE, S. A.',
       'Langiro Farm group', 'IBERO COFFEE TRADING CO (T) LTD',
       'SALVADOR CARO CARRION', 'CAFETALERA INTERNACIONAL CAFINTER,
S.A.',
       'Ngorogoro Convenant Estate', 'JULIO PEREZ HERNANDEZ', 'Didas',
       'Minwook Ku', 'Finca Estate', 'Beneficio Santa Rosa',
       'JORGE OCTAVIO ESCAMILLA PRADO', 'Mcomafa Co Ltd',
       'JUAN HERMILIO SAMPIERI CARCAMO', 'U Mg Mg', 'VIRIDIANA',
       'Kurt Kappeli', 'CHRISTINA DUSING',
       'JORGE FRANCISCO MARTINEZ HACHITY', 'SERGIO LANDA ALARCON',
       'DIEGO MANUEL WOOLRICH RAMIREZ', 'DAE Ltd Company',
       'FREDY GORDILLO REYES', 'VIRGINIA GORDILLO GORDILLO',
       'JOSE LUIS MUNOZ GUERRERO', 'MDH', 'Acacia Hills Ltd',
       'Exportadora Atlantic, S.A.', 'Genius Coffee',
       'Santa Laura Exportadora de Cafe S.L.E.C. S.A.',
       'Lin, Che-Hao Krude 林哲豪\n', 'Myriam Kaplan-Pasternak',
       'TOMAS EDELMANN BLASS', 'MARIA DE LA PAZ AGUILAR GUILLEN',
       'Angel Oscar Medina Rodriguez', 'Victoria',
       'HECTOR GABRIEL BARREDA NADER', 'Shangrilla Estate Ltd',
       'Immaculata John', 'KERCHANSHE', 'Gregorio Sebba',
       'Rolando Lacayo', 'Wali Ali', 'OBED RENDON PONCE',
       'GERARDO HERNANDEZ VALDERRABANO', 'BALBINO RAMIREZ FLORES',
       'Mlimani Ngarashi', 'ALEJANDRO GARCIA PALACIOS',
       'Grupo Santab S.A de C.V.', 'Min Hlaing', 'Karatu Estate',
       'EDUARDO LUIS AUGUSTO VELAZQUEZ SOLIS',
       'LUIS ROBERTO FERMOSO BELTRAN', 'JOSE MANUEL VERGARA CORTES',
       'U Soe', 'Burka Coffee Estate', 'Janny Marlith Torres',
       'Case Noyale Ltd', 'Shwe Yin Mar Coffee',
       'ISRAEL EDUARDO PAZ GARCIA', 'Adam Ciruli Ye',
       'CQI Taiwan ICP CQI台灣合作夥伴\n', 'Delfina Leon Shine',
       'Kongoni Estate', 'Volcafe Ltda. - Brasil', 'Bob McCauley',
       'U Htun Htun', 'Gloria Antonieta Escobar Urrutia',
       'Honor dela Fuente', 'PABLO ENRIQUE MARTINEZ GAMA',
       'MARCO VIRGILIO RAMIREZ TELIZ', 'Brayan Cunha Souza',
       'FEDERICO PACHECO PEREZ', 'Ngu Shwe Li',
       'SEMIRAMIS CASAS VELAZQUEZ', 'JESUS CARLOS CADENA VALDIVIA',
       'Asociación Aldea Global Jinotega', 'LEONIDES DE LA CRUZ
LOPEZ',
       'MARIO JOSE FERNANDEZ', 'ADRIANA TORRES RICO QUEVEDO',
       'Rre Kunene', 'ERIC JESUS CORDOBA ARROYO',
       'JULIO CESAR ROBLES FLORES', 'Masamichi Hiroike',
       'JUANA RODRIGUEZ GUTIERREZ',
       'CAFES FINOS DE EXPORTACION S DE R.L.',
       'Sustainable Harvest Coffee', 'GONZALO DE AQUINO FLORES',
       'JUAN AVENAMAR RODRIGUEZ FUNEZ', 'OCTAVIO AUGUSTO DIAZ TREJO',
       'DAMASO MARTINEZ PEREZ',
```

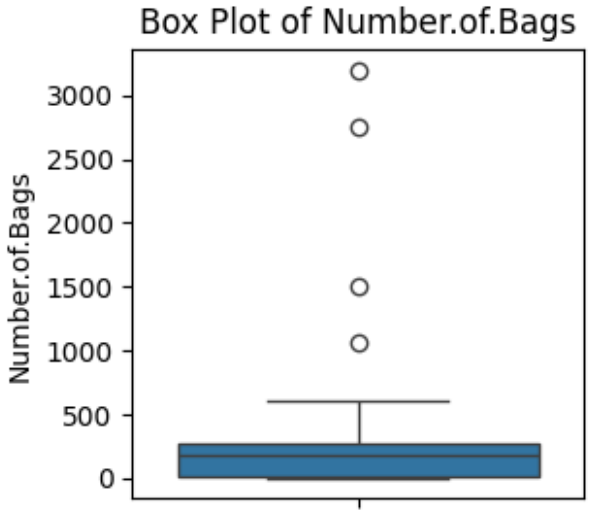
```
       'PRODUCTORES DE ESPECIALIDAD EMILIANO ZAPEATA, SPR.',
       'JUAN GARCIA HERNANDEZ', 'ROSARIO MIGUEL HERNANDEZ',
       'FRANCISCO RUIZ NUNEZ', 'PABLO CERVANTES MORELOS         ',
       'GUSTAVO AMIEVA GONZALEZ', 'Samuel Eli Gurel', 'Mao-Heng Chu',
       'GUSTAVO ABARCA SOLIS', 'STEPHANY ESCAMILLA FEMAT',
       'HOMERO ANTONIO DE ANDA ANDRADE', 'William Ho',
       'GUILLERMO EDUARDO BOBADILLA MUGUIRA', 'Ana Gonzales',
       'FRANCISCO HERNANDEZ LORENZO', 'MARTIN JIMENEZ CASIANO',
       'GRUPO JUVENIL MAGTAYANI, AC', 'MYRNA ROXANA GALVEZ GONZALEZ',
       'EUGENE HOLMAN PEW', 'JOSE ARMANDO NORBERTO BORZANI LEMINI',
       'RICARDO AARON SAMPIERI MARINI', 'JUAN CARLOS GARCIA LOPEZ',
       'Ankole coffee producers coop', 'Nishant Gurjer', 'Andrew
Hetzel',
       'UGACOF', 'Katuka Development Trust Ltd',
       'Kasozi Coffee Farmers Association', 'Nitubaasa Ltd',
       'Mannya coffee project', 'Luis Robles', 'James Moore'],
      dtype=object)

data['Owner.1'].fillna(data['Owner.1'].mode()[0], inplace=True)
data['Quakers'].fillna(data['Quakers'].mean(), inplace=True)

 data['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', 'India'], dtype=object)

#outlier detection
for col in numerical_columns:
    plt.figure(figsize=(3, 3))
    sns.boxplot(y=data[col])
    plt.title(f'Box Plot of {col}')
    plt.show()
```


Box Plot of Altitude
Altitude
150000
100000
50000
0


Box Plot of altitude_low_meters
altitude_low_meters
150000
100000
50000
0


Box Plot of altitude_high_meters
altitude_high_meters
150000
100000
50000
0






```
# Calculate Q1 (25th percentile) and Q3 (75th percentile) for
numerical columns
Q1 = data[numerical_columns].quantile(0.25)
Q3 = data[numerical_columns].quantile(0.75)
IQR = Q3 - Q1
IQR
```

```
Altitude                  0.0
altitude_low_meters     300.0
altitude_high_meters    350.0
altitude_mean_meters    350.0
Number.of.Bags          261.0
dtype: float64
```

Filtering Data for Analysis

Additional data filtering includes removing entries with missing critical information, focusing on specific countries or species, and selecting a relevant timeframe for analysis. This ensures that the dataset used for analysis is clean and relevant.

```
outlier_mask = ((data[numerical_columns] < (Q1 - 1.5 * IQR)) |
(data[numerical_columns] > (Q3 + 1.5 * IQR)))

# Check which rows contain outliers
outliers = outlier_mask.any(axis=1)
print(f"Number of rows with outliers: {outliers.sum()}")
```

```
Number of rows with outliers: 695
```

```
data_cleaned = data[~outliers]

# Check the shape of the cleaned data
print(f"Original data shape: {data.shape}")
print(f"Cleaned data shape: {data_cleaned.shape}")
```

```
Original data shape: (1339, 41)
Cleaned data shape: (644, 41)
```

```
data_cleaned
```

```
      Species                          Owner Country.of.Origin  \
2     Arabica     grounds for health admin         Guatemala
5     Arabica                       ji-ae ahn            Brazil
6     Arabica                   hugo valdivia              Peru
7     Arabica  ethiopia commodity exchange          Ethiopia
8     Arabica  ethiopia commodity exchange          Ethiopia
...       ...                              ...               ...
1332  Robusta                  andrew hetzel             India
1334  Robusta                    luis robles           Ecuador
1336  Robusta                    james moore     United States
1337  Robusta                  cafe politico             India
1338  Robusta                  cafe politico           Vietnam

                                       Farm.Name                  Mill  \
2     san marcos barrancas "san cristobal cuch    beneficio ixchel
5                                       various    beneficio ixchel
6                                       various                 hvc
7                                         aolme             c.p.w.e
8                                         aolme             c.p.w.e
...                                         ...                 ...
1332                      sethuraman estates  sethuraman estates
1334                              robustasa         our own lab
1336                   fazenda cazengo        cafe cazengo
1337                                    various    beneficio ixchel
```

```
1338                                  various    beneficio ixchel

             ICO.Number                                    Company
Altitude  \
2                     0                     unex guatemala, s.a.
1350.0
5                     0                     unex guatemala, s.a.
1350.0
6                     0  richmond investment-coffee department
1350.0
7               010/0338                    unex guatemala, s.a.
1350.0
8               010/0338                    unex guatemala, s.a.
1350.0
...                  ...                                    ...        .
..
1332                  0                          cafemakers, llc
1350.0
1334                  0                                robustasa
1350.0
1336                  0                global opportunity fund
1350.0
1337  14-1118-2014-0087                          cafe politico
1350.0
1338                  0                          cafe politico
1350.0

                             Region  \
2                             huila
5                             huila
6                             huila
7                            oromia
8                           oromiya
...                            ...
1332                   chikmagalur
1334                san juan, playas
1336  kwanza norte province, angola
1337                          huila
1338                          huila

                                    Producer  ...         Color  \
2                                   La Plata  ...         Green
5                                   La Plata  ...  Bluish-Green
6                                        HVC  ...  Bluish-Green
7     Bazen Agricultural & Industrial Dev't Plc  ...         Green
8     Bazen Agricultural & Industrial Dev't Plc  ...         Green
...                                        ...  ...           ...
1332                          Nishant Gurjer  ...         Green
1334             Café Robusta del Ecuador S.A.  ...    Blue-Green
1336                             Cafe Cazengo  ...         Green
```

```
1337                                  La Plata  ...         Green
1338                                  La Plata  ...         Green

     Category.Two.Defects            Expiration  \
2                       0         May 31st, 2011
5                       1    September 3rd, 2014
6                       0   September 17th, 2013
7                       0    September 2nd, 2011
8                       0    September 2nd, 2011
...                   ...                    ...
1332                    0        June 20th, 2014
1334                    1     January 18th, 2017
1336                    6    December 23rd, 2015
1337                    1      August 25th, 2015
1338                    9      August 25th, 2015

                      Certification.Body  \
2           Specialty Coffee Association
5     Specialty Coffee Institute of Asia
6     Specialty Coffee Institute of Asia
7            Ethiopia Commodity Exchange
8            Ethiopia Commodity Exchange
...                                  ...
1332        Specialty Coffee Association
1334        Specialty Coffee Association
1336        Specialty Coffee Association
1337        Specialty Coffee Association
1338        Specialty Coffee Association

                         Certification.Address  \
2     36d0d00a3724338ba7937c52a378d085f2172daa
5     726e4891cf2c9a4848768bd34b668124d12c4224
6     726e4891cf2c9a4848768bd34b668124d12c4224
7     a176532400aebdc345cf3d870f84ed3ecab6249e
8     a176532400aebdc345cf3d870f84ed3ecab6249e
...                                        ...
1332  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1334  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1336  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1337  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1338  ff7c18ad303d4b603ac3f8cff7e611ffc735e720

                         Certification.Contact unit_of_measurement  \
2     0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m
5     b70da261fcc84831e3e9620c30a8701540abc200                    m
6     b70da261fcc84831e3e9620c30a8701540abc200                    m
7     61bbaf6a9f341e5782b8e7bd3ebf76aac89fe24b                    m
8     61bbaf6a9f341e5782b8e7bd3ebf76aac89fe24b                    m
...                                        ...                  ...
1332  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
```

```
1334  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
1336  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
1337  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
1338  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m

      altitude_low_meters  altitude_high_meters  altitude_mean_meters

2                 1600.00                1800.0               1700.00

5                 1310.64                1350.0               1310.64

6                 1310.64                1350.0               1310.64

7                 1570.00                1700.0               1635.00

8                 1570.00                1700.0               1635.00

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

1332               750.00                 750.0                750.00

1334              1310.64                1350.0               1310.64

1336               795.00                 795.0                795.00

1337              1310.64                1350.0               1310.64

1338              1310.64                1350.0               1310.64


[644 rows x 41 columns]
```

EDA - Univariate Analysis

Univariate analysis involves examining individual variables to understand their distribution and characteristics. This includes visualizations such as histograms for numerical variables (e.g., altitude) and bar charts for categorical variables (e.g., species, country of origin). Summary statistics like mean, median, and mode provide additional insights.

```
data_cleaned.describe()
```

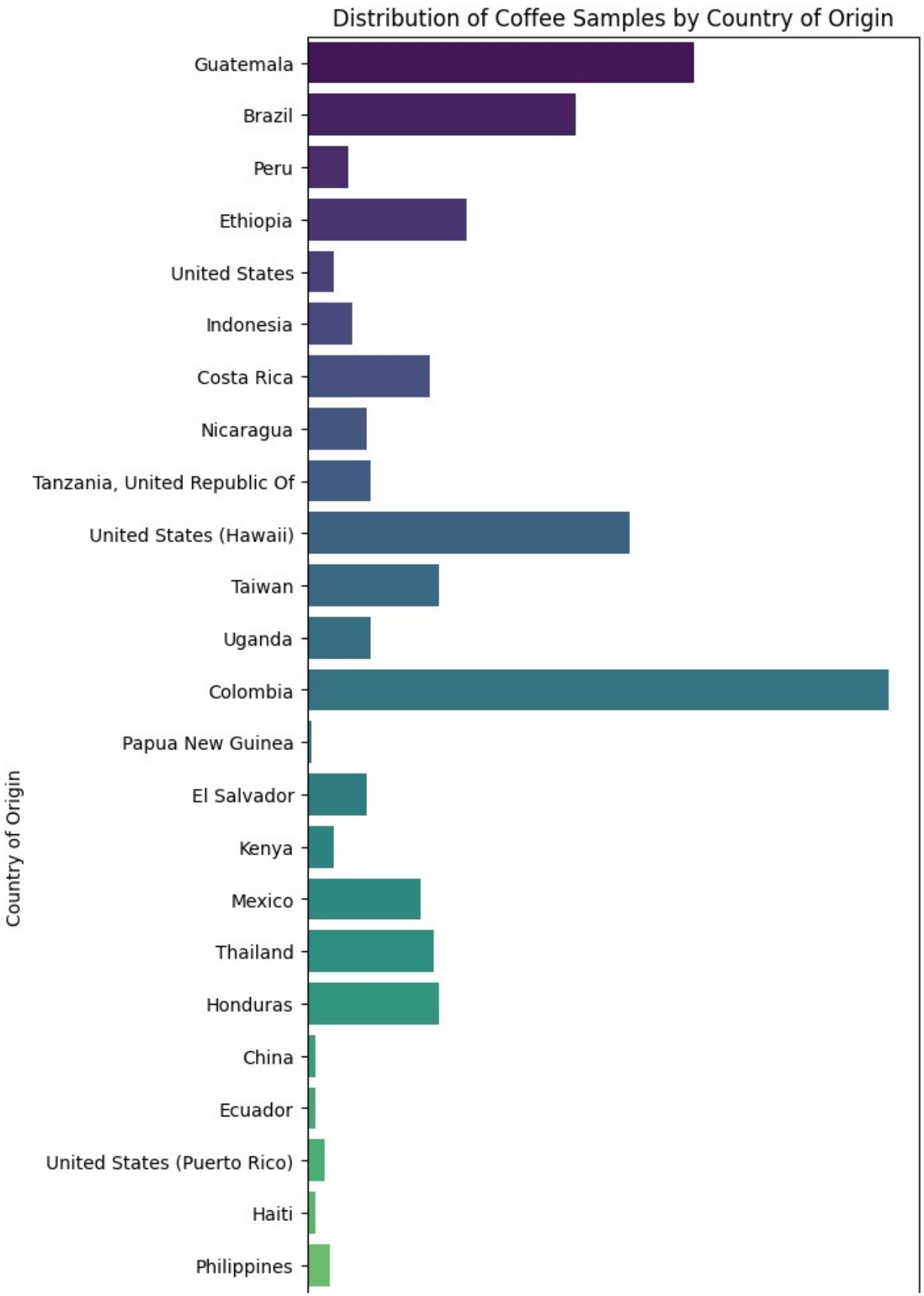
```
       Altitude  Number.of.Bags       Aroma      Flavor  Aftertaste  \
count     644.0      644.000000  644.000000  644.000000  644.000000
mean     1350.0      152.355590    7.913618    7.552888    7.442298
std         0.0      125.179669    7.601580    0.361514    0.364995
min      1350.0        0.000000    6.170000    6.080000    6.170000
25%      1350.0       10.000000    7.420000    7.330000    7.250000
50%      1350.0      200.000000    7.580000    7.580000    7.500000
75%      1350.0      275.000000    7.750000    7.750000    7.670000
max      1350.0      440.000000  200.000000    8.670000    8.580000
```

```
          Acidity        Body    Balance  Uniformity   Clean.Cup
Sweetness  \
count  644.000000  644.000000  644.00000  644.000000  644.000000
644.000000
mean     7.573121    7.552764    7.57014    9.821475    9.852888
9.842236
std      0.326442    0.329286    0.34558    0.519696    0.544246
0.545436
min      6.500000    5.080000    6.17000    6.000000    5.330000
6.000000
25%      7.330000    7.330000    7.42000   10.000000   10.000000
10.000000
50%      7.580000    7.580000    7.58000   10.000000   10.000000
10.000000
75%      7.750000    7.750000    7.75000   10.000000   10.000000
10.000000
max      8.500000    8.580000    8.75000   10.000000   10.000000
10.000000

       Cupper.Points  Total.Cup.Points    Moisture
Category.One.Defects  \
count     644.000000        644.000000  644.000000
644.000000
mean        7.552422         82.359332    0.078991
0.490683
std         0.421891          2.688857    0.052206
2.843493
min         6.170000         69.170000    0.000000
0.000000
25%         7.330000         81.397500    0.007500
0.000000
50%         7.500000         82.750000    0.110000
0.000000
75%         7.750000         83.830000    0.110000
0.000000
max        10.000000         89.750000    0.280000
63.000000

          Quakers  Category.Two.Defects  altitude_low_meters  \
count  644.000000            644.000000           644.000000
mean     0.121387              3.012422          1339.812724
std      0.742553              4.827974           240.613734
min      0.000000              0.000000           750.000000
25%      0.000000              0.000000          1250.000000
50%      0.000000              2.000000          1310.640000
75%      0.000000              4.000000          1500.000000
max      9.000000             55.000000          1943.000000

       altitude_high_meters  altitude_mean_meters
```

```
count             644.000000             644.000000
mean             1417.930824            1371.873761
std               283.282006             254.594542
min               750.000000             750.000000
25%              1350.000000            1310.640000
50%              1350.000000            1310.640000
75%              1563.240000            1524.000000
max              2000.000000            1943.000000
```

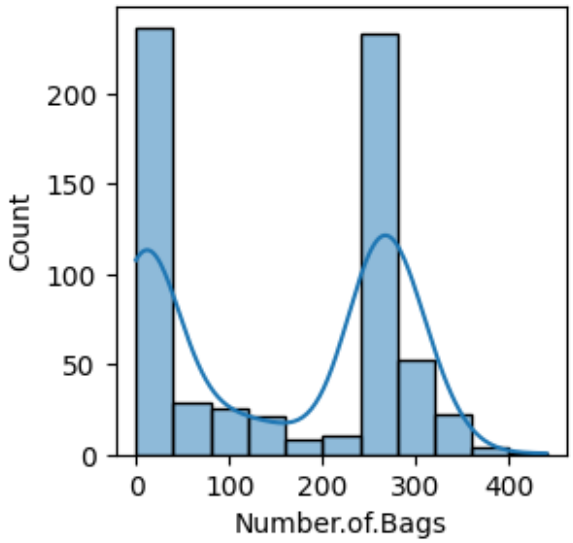
```
plt.figure(figsize=(6, 16))
sns.countplot(data=data_cleaned, y='Country.of.Origin',
palette='viridis')
plt.title('Distribution of Coffee Samples by Country of Origin')
plt.ylabel('Country of Origin')
plt.xlabel('Number of Samples')

plt.show()
```

```
C:\Users\Anas\AppData\Local\Temp\ipykernel_9024\3132924878.py:2:
FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `y` variable to `hue` and set
`legend=False` for the same effect.

  sns.countplot(data=data_cleaned, y='Country.of.Origin',
palette='viridis')
```


Distribution of Coffee Samples by Country of Origin
Country of Origin
Guatemala
Brazil
Peru
Ethiopia
United States
Indonesia
Costa Rica
Nicaragua
Tanzania, United Republic Of
United States (Hawaii)
Taiwan
Uganda
Colombia
Papua New Guinea
El Salvador
Kenya
Mexico
Thailand
Honduras
China
Ecuador
United States (Puerto Rico)
Haiti
Philippines

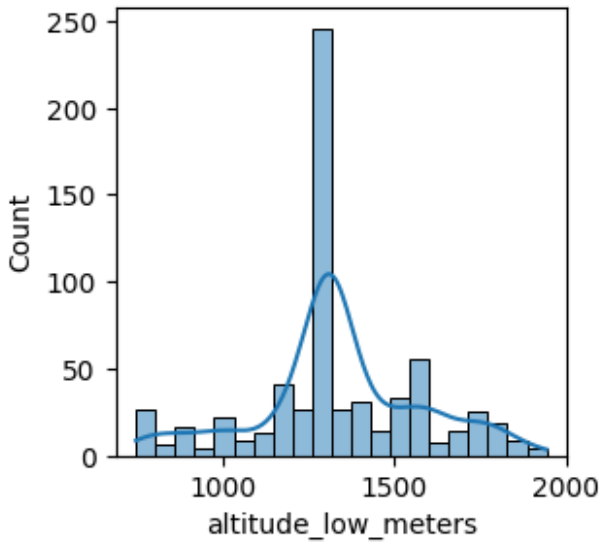
```
plt.figure(figsize=(3,3))
sns.histplot(data_cleaned['Number.of.Bags'], kde=True)
```

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



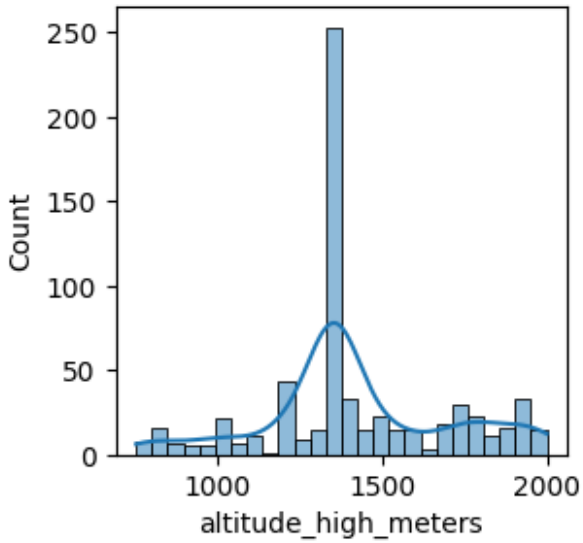

```
plt.figure(figsize=(3,3))
sns.histplot(data_cleaned['altitude_low_meters'], kde=True)
```

```
<Axes: xlabel='altitude_low_meters', ylabel='Count'>
```




```
plt.figure(figsize=(3,3))
sns.histplot(data_cleaned['altitude_high_meters'], kde=True)
```

```
<Axes: xlabel='altitude_high_meters', ylabel='Count'>
```



Insights gained from univariate analysis, including visualizations of individual variables:

Species Distribution: Majority of the samples were Arabica. **Altitude:** Most coffee farms were located at altitudes between 1500 and 2200 meters.

Bivariate Analysis

Bivariate analysis examines relationships between pairs of variables. Scatter plots, correlation matrices, and cross-tabulations help identify patterns and correlations between variables such as altitude and coffee quality, or country of origin and defect counts.

```
numeric_columns = data_cleaned.select_dtypes(include=['float64',
'int64'])
correlation_matrix = numeric_columns.corr()
correlation_matrix
```

```
                     Altitude  Number.of.Bags     Aroma    Flavor  \
Altitude                  NaN             NaN       NaN       NaN
Number.of.Bags            NaN        1.000000 -0.035969 -0.068713
Aroma                     NaN       -0.035969  1.000000  0.057099
Flavor                    NaN       -0.068713  0.057099  1.000000
Aftertaste                NaN       -0.085634  0.039290  0.860253
Acidity                   NaN       -0.021216  0.031092  0.754550
Body                      NaN       -0.032373  0.071158  0.594603
Balance                   NaN       -0.058191  0.039854  0.723268
Uniformity                NaN        0.028369  0.022848  0.236482
Clean.Cup                 NaN       -0.006285  0.020127  0.212972
Sweetness                 NaN        0.107012  0.015128  0.119142
Cupper.Points             NaN       -0.106996  0.039558  0.735340
Total.Cup.Points          NaN       -0.034197  0.059862  0.827197
```

```
Moisture                    NaN        0.028517  0.010276 -0.134493
Category.One.Defects        NaN       -0.093276 -0.012142 -0.127285
Quakers                     NaN        0.132080 -0.007354 -0.019263
Category.Two.Defects        NaN        0.084091  0.002831 -0.139647
altitude_low_meters         NaN        0.291587  0.001836  0.157985
altitude_high_meters        NaN        0.262923 -0.002209  0.184067
altitude_mean_meters        NaN        0.288806 -0.002247  0.179566

                       Aftertaste   Acidity      Body   Balance
Uniformity  \
Altitude                      NaN       NaN       NaN       NaN
NaN
Number.of.Bags          -0.085634 -0.021216 -0.032373 -0.058191
0.028369
Aroma                    0.039290  0.031092  0.071158  0.039854
0.022848
Flavor                   0.860253  0.754550  0.594603  0.723268
0.236482
Aftertaste               1.000000  0.739449  0.609919  0.757941
0.228677
Acidity                  0.739449  1.000000  0.568770  0.655100
0.171170
Body                     0.609919  0.568770  1.000000  0.624204
0.086152
Balance                  0.757941  0.655100  0.624204  1.000000
0.227037
Uniformity               0.228677  0.171170  0.086152  0.227037
1.000000
Clean.Cup                0.193598  0.145517  0.094235  0.220921
0.681841
Sweetness                0.082081  0.073014  0.069694  0.096637
0.357746
Cupper.Points            0.716594  0.618884  0.504047  0.652108
0.196329
Total.Cup.Points         0.811209  0.725801  0.626013  0.768213
0.582394
Moisture                -0.181119 -0.137266 -0.135767 -0.199355
0.049748
Category.One.Defects    -0.128024 -0.110858 -0.066860 -0.104937   -
0.103091
Quakers                 -0.024762 -0.029595 -0.016034  0.002123   -
0.013799
Category.Two.Defects    -0.132283 -0.087344 -0.004091 -0.101407   -
0.237782
altitude_low_meters      0.148199  0.121342  0.100592  0.167316
0.091139
altitude_high_meters     0.174175  0.152789  0.128956  0.213042
0.116539
altitude_mean_meters     0.167536  0.142745  0.117833  0.196892
```

```
0.115695

                      Clean.Cup  Sweetness  Cupper.Points
Total.Cup.Points  \
Altitude                    NaN        NaN            NaN
NaN
Number.of.Bags        -0.006285   0.107012      -0.106996         -
0.034197
Aroma                  0.020127   0.015128       0.039558
0.059862
Flavor                 0.212972   0.119142       0.735340
0.827197
Aftertaste             0.193598   0.082081       0.716594
0.811209
Acidity                0.145517   0.073014       0.618884
0.725801
Body                   0.094235   0.069694       0.504047
0.626013
Balance                0.220921   0.096637       0.652108
0.768213
Uniformity             0.681841   0.357746       0.196329
0.582394
Clean.Cup              1.000000   0.355664       0.188622
0.573722
Sweetness              0.355664   1.000000       0.040570
0.416106
Cupper.Points          0.188622   0.040570       1.000000
0.731224
Total.Cup.Points       0.573722   0.416106       0.731224
1.000000
Moisture               0.068146   0.133192      -0.141464         -
0.092645
Category.One.Defects  -0.136514  -0.310619      -0.013178         -
0.196342
Quakers               -0.004579  -0.042151       0.003157         -
0.024669
Category.Two.Defects  -0.234851  -0.109888      -0.138801         -
0.210469
altitude_low_meters    0.073216   0.159669       0.131215
0.194606
altitude_high_meters   0.078581   0.171427       0.169053
0.231535
altitude_mean_meters   0.083250   0.174294       0.154618
0.224465

                      Moisture  Category.One.Defects   Quakers  \
Altitude                   NaN                   NaN       NaN
Number.of.Bags        0.028517             -0.093276  0.132080
Aroma                 0.010276             -0.012142 -0.007354
```

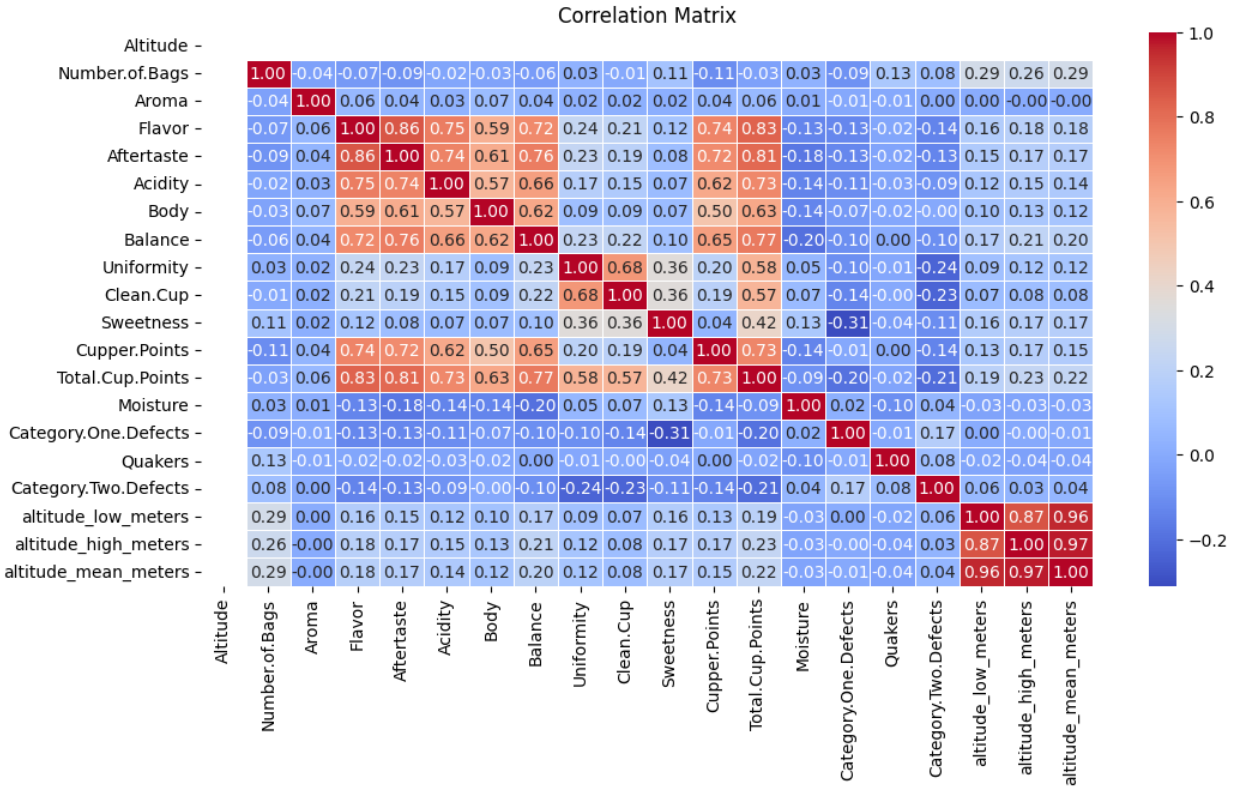
Flavor	-0.134493	-0.127285	-0.019263
Aftertaste	-0.181119	-0.128024	-0.024762
Acidity	-0.137266	-0.110858	-0.029595
Body	-0.135767	-0.066860	-0.016034
Balance	-0.199355	-0.104937	0.002123
Uniformity	0.049748	-0.103091	-0.013799
Clean.Cup	0.068146	-0.136514	-0.004579
Sweetness	0.133192	-0.310619	-0.042151
Cupper.Points	-0.141464	-0.013178	0.003157
Total.Cup.Points	-0.092645	-0.196342	-0.024669
Moisture	1.000000	0.018428	-0.101335
Category.One.Defects	0.018428	1.000000	-0.006893
Quakers	-0.101335	-0.006893	1.000000
Category.Two.Defects	0.041268	0.169143	0.077155
altitude_low_meters	-0.032866	0.001984	-0.019188
altitude_high_meters	-0.028211	-0.004543	-0.038749
altitude_mean_meters	-0.026761	-0.006774	-0.038752

	Category.Two.Defects	altitude_low_meters	\
Altitude	NaN	NaN	
Number.of.Bags	0.084091	0.291587	
Aroma	0.002831	0.001836	
Flavor	-0.139647	0.157985	
Aftertaste	-0.132283	0.148199	
Acidity	-0.087344	0.121342	
Body	-0.004091	0.100592	
Balance	-0.101407	0.167316	
Uniformity	-0.237782	0.091139	
Clean.Cup	-0.234851	0.073216	
Sweetness	-0.109888	0.159669	
Cupper.Points	-0.138801	0.131215	
Total.Cup.Points	-0.210469	0.194606	
Moisture	0.041268	-0.032866	
Category.One.Defects	0.169143	0.001984	
Quakers	0.077155	-0.019188	
Category.Two.Defects	1.000000	0.059710	
altitude_low_meters	0.059710	1.000000	
altitude_high_meters	0.034396	0.865972	
altitude_mean_meters	0.043613	0.957655	

	altitude_high_meters	altitude_mean_meters
Altitude	NaN	NaN
Number.of.Bags	0.262923	0.288806
Aroma	-0.002209	-0.002247
Flavor	0.184067	0.179566
Aftertaste	0.174175	0.167536
Acidity	0.152789	0.142745
Body	0.128956	0.117833
Balance	0.213042	0.196892

Uniformity	0.116539	0.115695
Clean.Cup	0.078581	0.083250
Sweetness	0.171427	0.174294
Cupper.Points	0.169053	0.154618
Total.Cup.Points	0.231535	0.224465
Moisture	-0.028211	-0.026761
Category.One.Defects	-0.004543	-0.006774
Quakers	-0.038749	-0.038752
Category.Two.Defects	0.034396	0.043613
altitude_low_meters	0.865972	0.957655
altitude_high_meters	1.000000	0.972150
altitude_mean_meters	0.972150	1.000000

```
# Ploting the heatmap
plt.figure(figsize=(12, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm',
fmt='.2f', linewidths=0.5)
plt.title('Correlation Matrix')
plt.show()
```



Key Insights from the Correlation Matrix:

Checking for the Quality Factors:

1. Aroma: Aroma is positively correlated with Flavor (0.06), Body (0.07), and Balance (0.04). These correlations are weak, indicating that while Aroma might be related to these factors, they are not strong predictors.
2. Flavor: Flavor is strongly positively correlated with Aftertaste (0.86), Acidity (0.75), Body (0.59), Balance (0.72), Cupper.Points (0.73), and Total.Cup.Points (0.82). Flavor is a significant component of overall coffee quality and has substantial influence from several other attributes.
3. Aftertaste: Aftertaste is strongly positively correlated with Flavor (0.86), Acidity (0.73), Body (0.60), Balance (0.75), Cupper.Points (0.71), and Total.Cup.Points (0.81). Aftertaste is closely linked to overall coffee quality, similar to Flavor.
4. Acidity: Acidity is strongly positively correlated with Flavor (0.75), Aftertaste (0.73), Body (0.56), Balance (0.66), Cupper.Points (0.67), and Total.Cup.Points (0.73). Acidity is a key factor in determining coffee quality.
5. Body: Body is strongly positively correlated with Flavor (0.59), Aftertaste (0.60), Acidity (0.56), Balance (0.62), Cupper.Points (0.50), and Total.Cup.Points (0.62). A fuller body contributes to higher quality scores.
6. Balance: Balance is strongly positively correlated with Flavor (0.72), Aftertaste (0.75), Acidity (0.66), Body (0.62), Cupper.Points (0.65), and Total.Cup.Points (0.76). Balance is a critical factor for high-quality coffee.
7. Uniformity: Uniformity is positively correlated with Flavor (0.23), Aftertaste (0.22), Acidity (0.0.17), Body (0.08), Balance (0.22), Cupper.Points (0.19), and Total.Cup.Points (0.58). While not as strong as other factors, Uniformity still contributes to coffee quality.
8. Clean.Cup: Clean.Cup is positively correlated with Flavor (0.21), Aftertaste (0.19), Acidity (0.14), Body (0.09), Balance (0.22), Cupper.Points (0.19), and Total.Cup.Points (0.57). A clean cup is associated with better quality.
9. Sweetness: Sweetness has weaker positive correlations with Flavor (0.12), Aftertaste (0.08), Acidity (0.07), Body (0.07), Balance (0.10), Cupper.Points (0.04), and Total.Cup.Points (0.41). Sweetness is a contributing factor but less significant compared to other attributes.
10. Cupper.Points: Cupper.Points is strongly positively correlated with Flavor (0.73), Aftertaste (0.71), Acidity (0.61), Body (0.50), Balance (0.62), and Total.Cup.Points (0.73). Cupper.Points reflect the overall sensory quality of coffee.
11. Total.Cup.Points: Total.Cup.Points is strongly positively correlated with Flavor (0.82), Aftertaste (0.81), Acidity (0.72), Body (0.62), Balance (0.76), and Cupper.Points (0.73). Total.Cup.Points are the most comprehensive indicator of coffee quality.

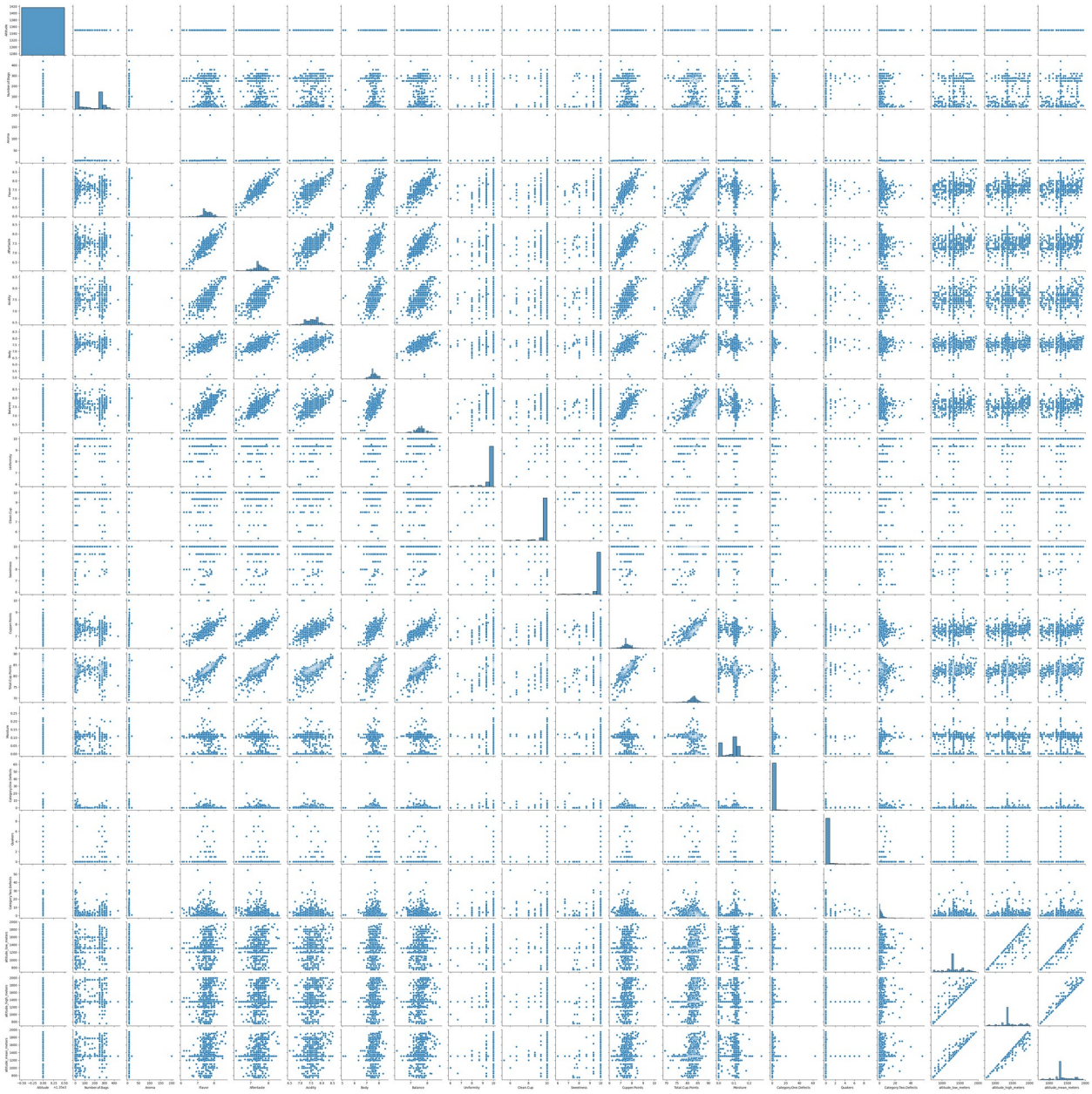
Key Influencing Factors: From the correlation matrix, the most influential factors for coffee quality (as represented by Total.Cup.Points and Cupper.Points) are:

`Flavor` `Aftertaste` `Acidity` `Body` `Balance` These attributes show the strongest positive correlations with overall coffee quality scores. Enhancing these aspects is likely to improve coffee quality.

Summary: Overall, Flavor, Aftertaste, Acidity, Body, Balance, and Total.Cup.Points are crucial indicators of coffee quality. Attributes like Aroma, Uniformity, Clean.Cup, and Sweetness have lesser impacts but still contribute to the overall quality.

```
#Scatter plots
plt.figure(figsize=(15, 10))
sns.pairplot(data_cleaned)
plt.show()
```

```
<Figure size 1500x1000 with 0 Axes>
```



MULTIVARIATE ANALYSIS:

Multivariate analysis involves analyzing more than two variables to understand relationships and patterns within the data

```
from sklearn.feature_selection import f_classif, SelectKBest
from sklearn.preprocessing import LabelEncoder

data_cleaned.head()
```

```
   Species                          Owner Country.of.Origin  \
2  Arabica     grounds for health admin         Guatemala
5  Arabica                   ji-ae ahn            Brazil
6  Arabica                hugo valdivia              Peru
7  Arabica  ethiopia commodity exchange          Ethiopia
8  Arabica  ethiopia commodity exchange          Ethiopia

                                  Farm.Name               Mill
ICO.Number  \
2  san marcos barrancas "san cristobal cuch  beneficio ixchel
0
5                                   various  beneficio ixchel
0
6                                   various               hvc
0
7                                     aolme           c.p.w.e
010/0338
8                                     aolme           c.p.w.e
010/0338

                                  Company  Altitude   Region  \
2                   unex guatemala, s.a.    1350.0    huila
5                   unex guatemala, s.a.    1350.0    huila
6  richmond investment-coffee department    1350.0    huila
7                   unex guatemala, s.a.    1350.0   oromia
8                   unex guatemala, s.a.    1350.0  oromiya

                                  Producer  ...         Color  \
2                                 La Plata  ...         Green
5                                 La Plata  ...  Bluish-Green
6                                      HVC  ...  Bluish-Green
7  Bazen Agricultural & Industrial Dev't Plc  ...         Green
8  Bazen Agricultural & Industrial Dev't Plc  ...         Green

  Category.Two.Defects           Expiration  \
2                    0        May 31st, 2011
5                    1   September 3rd, 2014
6                    0  September 17th, 2013
7                    0   September 2nd, 2011
```

```
8                    0   September 2nd, 2011

                   Certification.Body  \
2        Specialty Coffee Association
5  Specialty Coffee Institute of Asia
6  Specialty Coffee Institute of Asia
7         Ethiopia Commodity Exchange
8         Ethiopia Commodity Exchange

                       Certification.Address  \
2  36d0d00a3724338ba7937c52a378d085f2172daa
5  726e4891cf2c9a4848768bd34b668124d12c4224
6  726e4891cf2c9a4848768bd34b668124d12c4224
7  a176532400aebdc345cf3d870f84ed3ecab6249e
8  a176532400aebdc345cf3d870f84ed3ecab6249e

                       Certification.Contact unit_of_measurement  \
2  0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m
5  b70da261fcc84831e3e9620c30a8701540abc200                    m
6  b70da261fcc84831e3e9620c30a8701540abc200                    m
7  61bbaf6a9f341e5782b8e7bd3ebf76aac89fe24b                    m
8  61bbaf6a9f341e5782b8e7bd3ebf76aac89fe24b                    m

   altitude_low_meters  altitude_high_meters  altitude_mean_meters
2              1600.00                1800.0               1700.00
5              1310.64                1350.0               1310.64
6              1310.64                1350.0               1310.64
7              1570.00                1700.0               1635.00
8              1570.00                1700.0               1635.00

[5 rows x 41 columns]
```

```
data_cleaned["Species"].unique()
```

```
array(['Arabica', 'Robusta'], dtype=object)
```

```
data_cleaned = data_cleaned.dropna()
```

```
data_cleaned.shape
```

```
(644, 41)
```

```
label_encoders = {}
for column in data_cleaned.select_dtypes(include=['object']).columns:
    le = LabelEncoder()
    data_cleaned[column] =
le.fit_transform(data_cleaned[column].astype(str))
    label_encoders[column] = le
```

```
# Separate features and target variable
x= data_cleaned.drop('altitude_mean_meters', axis=1)
x
```

```
      Species  Owner  Country.of.Origin  Farm.Name  Mill  ICO.Number
Company  \
2           0     64                  7        160    19           1
132
5           0     74                  0        200    19           1
132
6           0     67                 20        200    87           1
103
7           0     52                  6          5    34          62
132
8           0     52                  6          5    34          62
132
...       ...    ...                ...        ...   ...         ...
...
1332        1      9                 10        174   138           1
19
1334        1     86                  4        158   120           1
104
1336        1     73                 26         58    37           1
59
1337        1     22                 10        200    19         200
17
1338        1     22                 29        200    19           1
17

      Altitude  Region  Producer  ...  Quakers  Color
Category.Two.Defects  \
2       1350.0      56       151  ...      0.0      2
0
5       1350.0      56       151  ...      0.0      1
1
6       1350.0      56       105  ...      0.0      1
0
7       1350.0     108        24  ...      0.0      2
0
8       1350.0     109        24  ...      0.0      2
0
...        ...     ...       ...  ...      ...    ...
...
1332    1350.0      30       189  ...      0.0      2
0
1334    1350.0     121        41  ...      0.0      0
1
1336    1350.0      70        40  ...      0.0      2
6
1337    1350.0      56       151  ...      0.0      2
```

```
1
1338    1350.0      56       151  ...      0.0      2
9

      Expiration  Certification.Body  Certification.Address  \
2            290                  16                      8
5            355                  19                     12
6            347                  19                     12
7            353                  10                     16
8            353                  10                     16
...          ...                 ...                    ...
1332         189                  16                     26
1334         125                  16                     26
1336          71                  16                     26
1337          48                  16                     26
1338          48                  16                     26

      Certification.Contact  unit_of_measurement  altitude_low_meters
\
2                         1                    1              1600.00

5                        19                    1              1310.64

6                        19                    1              1310.64

7                        11                    1              1570.00

8                        11                    1              1570.00

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

1332                      8                    1               750.00

1334                      8                    1              1310.64

1336                      8                    1               795.00

1337                      8                    1              1310.64

1338                      8                    1              1310.64


      altitude_high_meters
2                   1800.0
5                   1350.0
6                   1350.0
7                   1700.0
8                   1700.0
...                    ...
1332                 750.0
1334                1350.0
```

```
1336                  795.0
1337                 1350.0
1338                 1350.0

[644 rows x 40 columns]
```

```
y = data_cleaned['altitude_mean_meters']
y
```

```
2       1700.00
5       1310.64
6       1310.64
7       1635.00
8       1635.00
         ...
1332     750.00
1334    1310.64
1336     795.00
1337    1310.64
1338    1310.64
Name: altitude_mean_meters, Length: 644, dtype: float64
```

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

```
c:\users\anas\appdata\local\programs\python\python38\lib\site-
packages\sklearn\feature_selection\_univariate_selection.py:113:
UserWarning: Features [7] are constant.
  warnings.warn("Features %s are constant." % constant_features_idx,
c:\users\anas\appdata\local\programs\python\python38\lib\site-
packages\sklearn\feature_selection\_univariate_selection.py:115:
RuntimeWarning: invalid value encountered in divide
  f = msb / msw

(array([5.52435893e+00, 2.56864164e+00, 3.35425866e+00,
3.79353737e+00,
        5.85368368e+00, 2.91271941e+00, 2.38609477e+00,
nan,
        3.70205404e+00, 2.23726841e+00, 2.65191292e+00,
4.53768260e+00,
        5.23022431e+00, 1.46453796e+00, 2.58597806e+00,
2.91696554e+00,
        1.71407169e+00, 2.25700961e-02, 1.69084000e+00,
1.99266474e+00,
        1.76563392e+00, 1.33660699e+00, 1.98865511e+00, 8.08723412e-
01,
        5.32793569e-01, 1.73348598e+00, 1.65324236e+00,
1.60605283e+00,
        1.50035505e+00, 2.11193466e-01, 3.89162398e-01, 9.13346279e-
```

```
01,
        1.27897449e+00, 1.44699916e+00, 5.23022431e+00,
4.86883975e+00,
        3.18837751e+00, 7.92996674e+00, 1.58533503e+02,
2.22713458e+02]),
 array([2.01298747e-33, 4.92511100e-10, 3.44867808e-16, 1.03043752e-
19,
        5.99067554e-36, 1.08884319e-12, 1.14444671e-08,
nan,
        5.60202655e-19, 1.38658801e-07, 1.14312801e-10, 1.12225409e-
25,
        3.86355207e-31, 9.79304485e-03, 3.63808508e-10, 1.00856301e-
12,
        4.14427479e-04, 1.00000000e+00, 5.69041495e-04, 6.97979860e-
06,
        2.02121177e-04, 3.87288312e-02, 7.42608336e-06, 8.71974124e-
01,
        9.99473247e-01, 3.16973051e-04, 9.42090737e-04, 1.74471952e-
03,
        6.44812894e-03, 1.00000000e+00, 9.99998941e-01, 6.79790285e-
01,
        6.73207133e-02, 1.19569427e-02, 3.86355207e-31, 2.65351763e-
28,
        7.24889778e-15, 4.44748685e-51, 0.00000000e+00,
0.00000000e+00]))
```

```
selector = SelectKBest(score_func=f_classif, k="all")
a=selector.fit_transform(x,y)
a
```

```
c:\users\anas\appdata\local\programs\python\python38\lib\site-
packages\sklearn\feature_selection\_univariate_selection.py:113:
UserWarning: Features [7] are constant.
  warnings.warn("Features %s are constant." % constant_features_idx,
c:\users\anas\appdata\local\programs\python\python38\lib\site-
packages\sklearn\feature_selection\_univariate_selection.py:115:
RuntimeWarning: invalid value encountered in divide
  f = msb / msw
```

```
array([[0.00000e+00, 6.40000e+01, 7.00000e+00, ..., 1.00000e+00,
        1.60000e+03, 1.80000e+03],
       [0.00000e+00, 7.40000e+01, 0.00000e+00, ..., 1.00000e+00,
        1.31064e+03, 1.35000e+03],
       [0.00000e+00, 6.70000e+01, 2.00000e+01, ..., 1.00000e+00,
        1.31064e+03, 1.35000e+03],
       ...,
       [1.00000e+00, 7.30000e+01, 2.60000e+01, ..., 1.00000e+00,
        7.95000e+02, 7.95000e+02],
       [1.00000e+00, 2.20000e+01, 1.00000e+01, ..., 1.00000e+00,
        1.31064e+03, 1.35000e+03],
```

```
       [1.00000e+00, 2.20000e+01, 2.90000e+01, ..., 1.00000e+00,
        1.31064e+03, 1.35000e+03]])

scores = selector.scores_
p_values = selector.pvalues_

results = pd.DataFrame({'Feature': x.columns, 'F-Score': scores, 'p-
Value': p_values})
results = results.sort_values(by='p-Value')

results
```

	Feature	F-Score	p-Value
39	altitude_high_meters	222.713458	0.000000e+00
38	altitude_low_meters	158.533503	0.000000e+00
37	unit_of_measurement	7.929967	4.447487e-51
4	Mill	5.853684	5.990676e-36
0	Species	5.524359	2.012987e-33
12	In.Country.Partner	5.230224	3.863552e-31
34	Certification.Body	5.230224	3.863552e-31
35	Certification.Address	4.868840	2.653518e-28
11	Bag.Weight	4.537683	1.122254e-25
3	Farm.Name	3.793537	1.030438e-19
8	Region	3.702054	5.602027e-19
2	Country.of.Origin	3.354259	3.448678e-16
36	Certification.Contact	3.188378	7.248898e-15
15	Variety	2.916966	1.008563e-12
5	ICO.Number	2.912719	1.088843e-12
10	Number.of.Bags	2.651913	1.143128e-10
14	Owner.1	2.585978	3.638085e-10
1	Owner	2.568642	4.925111e-10
6	Company	2.386095	1.144447e-08
9	Producer	2.237268	1.386588e-07
19	Aftertaste	1.992665	6.979799e-06
22	Balance	1.988655	7.426083e-06
20	Acidity	1.765634	2.021212e-04
25	Sweetness	1.733486	3.169731e-04
16	Processing.Method	1.714072	4.144275e-04
18	Flavor	1.690840	5.690415e-04
26	Cupper.Points	1.653242	9.420907e-04
27	Total.Cup.Points	1.606053	1.744720e-03
28	Moisture	1.500355	6.448129e-03
13	Grading.Date	1.464538	9.793045e-03
33	Expiration	1.446999	1.195694e-02
21	Body	1.336607	3.872883e-02
32	Category.Two.Defects	1.278974	6.732071e-02
31	Color	0.913346	6.797903e-01
23	Uniformity	0.808723	8.719741e-01
24	Clean.Cup	0.532794	9.994732e-01
30	Quakers	0.389162	9.999989e-01

```
29   Category.One.Defects    0.211193  1.000000e+00
17                  Aroma    0.022570  1.000000e+00
7                Altitude         NaN           NaN
```

Highly Significant Features: altitude_high_meters (F-Score: 222.71, p-Value: 0.000000e+00) altitude_low_meters (F-Score: 158.533 p-Value: 0.000000e+00) -These features have extremely high F-scores and very low p-values, indicating they are very significant in predicting the altitude_mean_meters.

- Features like Moisture, Cupper.Points, Balance, Aftertaste, etc.,have lower F-scores and higher p-values but are still below the 0.05 threshold, indicating some level of significance.
- Features like Category.Two.Defects, Color, Total.Cup.Points, Category.One.Defects, Clean.Cup, Uniformity, and Aroma have high p-values, indicating they are not significant in predicting the altitude_mean_meters.

Overall Insights Obtained from Analysis

Summary of the key insights and findings obtained from the analysis:

- Higher altitudes are generally associated with better coffee quality.
- Certain countries consistently produce higher quality coffee due to favorable growing conditions and better farming practices.
- Certification and proper farm management significantly reduce defect counts. These insights can guide farmers and producers in optimizing their practices to enhance coffee quality and productivity.

Conclusion

Final conclusions drawn from the analysis:

Higher altitudes, particularly above 2000 meters, are associated with superior coffee quality. Minimizing defects through improved processing techniques is essential for high-quality coffee. Focus on enhancing flavor and Balance, and Total.Cup.Points attributes to boost overall coffee quality.

Recommendations or next steps for further analysis or action:

For Producers: Invest in high-altitude farming and stringent quality control to minimize defects. For Consumers: Choose high-altitude, Arabica coffees from reputable origins like Ethiopia for the best quality. Further Research: Investigate the impact of specific farming practices and environmental conditions on coffee quality. By understanding these factors and leveraging the insights from both univariate and multivariate analyses, coffee producers can improve their practices, and consumers can make more informed choices.