**Project Information**

* Title: Exploratory Data Analysis using Python
* Name: A Manikandan
* Course: DA/DS
* Batch Number: November 2024
* Online/Offline: Online
* Roll Number: 61024OL002

**Introduction**

This project focuses on analyzing a comprehensive dataset related to global coffee quality using Python in Jupyter Notebook. The contents of the dataset are provided below with the explanation below.

1. Species – Type of coffee bean, such as Arabica or Robusta.
2. Owner – The person or organization owning the coffee farm.
3. Country of Origin – The country where the coffee was grown.
4. Farm Name – The specific name of the coffee-producing farm.
5. Lot Number – Identifier assigned to a particular batch of coffee.
6. Mill – The processing facility where the coffee was prepared.
7. ICO Number – Unique identifier assigned by the International Coffee Organization.
8. Company – The company associated with the coffee sample.
9. Altitude – The elevation at which the coffee was grown.
10. Region – Geographic region within the country of origin.
11. Producer – Individual or group responsible for producing the coffee.
12. Number of Bags – The quantity of coffee bags in the lot.
13. Bag Weight – The weight of each coffee bag.
14. In Country Partner – Local partner involved in coffee production or distribution.
15. Harvest Year – The year the coffee was harvested.
16. Grading Date – The date when the coffee was evaluated or scored.
17. Owner 1 – Secondary or additional ownership information.
18. Variety – The specific varietal or strain of the coffee plant.
19. Processing Method – The method used to process coffee beans (e.g., natural, washed).
20. Aroma – Sensory score for the smell of the coffee.
21. Flavour – Sensory score for the taste of the coffee.
22. Aftertaste – Sensory score for the lingering taste after sipping.
23. Acidity – Sensory evaluation of coffee's brightness or sharpness.
24. Body – Score representing the texture or mouthfeel of the coffee.
25. Balance – Overall harmony between flavor attributes.
26. Uniformity – Consistency across multiple cups in the same batch.
27. Clean Cup – Degree of clarity and absence of defects in taste.
28. Sweetness – Perceived sweetness in the coffee profile.
29. Cupper Points – Score given by the professional cupper.
30. Total Cup Points – Overall quality score combining sensory elements.
31. Moisture – Moisture content in the coffee bean sample.
32. Category One Defects – Number of serious physical defects in beans.
33. Quakers – Underripe or poorly roasted beans in the batch.
34. Color – Color grade of the beans, typically indicating roast level or defects.
35. Category Two Defects – Minor physical defects in the beans.
36. Expiration – Expiry date of the certification or product.
37. Certification Body – Organization responsible for certification of the coffee.
38. Certification Address – Address of the certification organization.
39. Certification Contact – Contact information for certification body.
40. Unit of Measurement – Measurement units used for weight, volume, or other values.
41. Altitude Low (meters) – Minimum elevation for coffee growth.
42. Altitude High (meters) – Maximum elevation for coffee growth.
43. Altitude Mean (meters) – Average elevation for coffee cultivation.

**Aim**

The aim of this project is to analyze a coffee quality dataset using Python in Jupyter Notebook. The project involves importing data from Excel, performing thorough data cleaning, preprocessing, and exploratory data analysis (EDA) to uncover meaningful patterns and insights. Key tasks include handling missing values and outliers, standardizing numerical and categorical data, and conducting both descriptive and inferential statistical analysis. Visualizations were created to support interpretation and storytelling, and the insights were compiled into a presentation. The ultimate goal is to derive actionable recommendations that can enhance coffee quality assessment, focusing on important factors such as aroma, flavor, aftertaste, and altitude.

**Business Problem/Problem Statement**

This project tackles the challenge of maintaining consistent coffee quality in a competitive global market. Variability in factors like altitude, processing methods, and sensory attributes often leads to inconsistent product standards. Such inconsistencies can affect consumer satisfaction, pricing strategies, and brand reputation. By analysing detailed coffee quality data, the project identifies key drivers that influence taste and overall quality. These insights help producers and stakeholders make data-driven decisions to improve consistency and product excellence.

**Project Workflow**

The project started by importing the coffee quality dataset from Excel into Jupyter Notebook. Data cleaning was performed to handle missing values and outliers using imputation and correction techniques. Preprocessing included standardizing data types and formatting for analysis. Exploratory Data Analysis (EDA) helped uncover trends and anomalies through visualizations. Statistical tests like t-tests were applied to compare attributes across coffee species. Insights were visualized and summarized in a presentation for actionable recommendations.

**Data Understanding**

Data columns (total 46 columns):

# Column Non-Null Count Dtype

0 ID 1339 non-null int64

1 Species 1339 non-null object

2 Owner 1339 non-null object

3 Country.of.Origin 1339 non-null object

4 Farm.Name 1339 non-null object

5 Lot.Number 1339 non-null object

6 Mill 1339 non-null object

7 ICO.Number 1339 non-null object

8 Company 1339 non-null object

9 Altitude 1339 non-null object

10 Region 1339 non-null object

11 Producer 1339 non-null object

12 Number.of.Bags 1339 non-null float64

13 Bag.Weight 1339 non-null object

14 In.Country.Partner 1339 non-null object

15 Harvest.Year 1339 non-null object

16 Grading.Date 1339 non-null object

17 Owner.1 1339 non-null object

18 Variety 1339 non-null object

19 Processing.Method 1339 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 float64

33 Quakers 1339 non-null float64

34 Color 1339 non-null object

35 Category.Two.Defects 1339 non-null float64

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 1339 non-null float64

42 altitude\_high\_meters 1339 non-null float64

43 altitude\_mean\_meters 1339 non-null float64

44 Bag.Weight.Numeric 1339 non-null float64

45 Grading\_Year 1339 non-null int32

46 Expiration\_Year 1339 non-null int32

The initial exploratory analysis revealed that most coffee samples had high sensory scores, with average values above 7 for aroma, flavor, and aftertaste—indicating overall good quality. Descriptive statistics showed minimal variation, but outliers were present, especially in flavor and aroma, suggesting possible inconsistencies in processing or storage. The majority of samples were grown at altitudes between 1000–1500 meters, with a few extreme high-altitude entries. Moisture levels were generally consistent, though some outliers indicated potential storage issues. These early insights highlighted patterns in quality and helped shape deeper statistical and visual analyses.

**Data Cleaning - Missing Values Imputation, Outliers, Handling Inconsistent Values**

**1. Overview**

In this report, I will explain the methods used to clean and preprocess the coffee dataset. The dataset contains various attributes related to coffee, including farm details, coffee variety, altitude, and quality assessments. The cleaning steps addressed missing values, outliers, and inconsistent data points to ensure the data's reliability for further analysis. The detailed steps and the codes used are provided below for reference.

**2. Handling Missing Values**

One of the primary tasks in data cleaning was dealing with missing values. The dataset had several columns with null or missing entries. Below are the methods used to impute missing values:

* Filling Missing Values with Default Values: Several categorical columns had missing values that were replaced with default labels:

Columns such as Variety, Processing.Method, Farm.Name, and others were filled with the default label 'Other' where the value was missing. For example:

df\_coffee['Variety'] = df\_coffee['Variety'].fillna('Other')

df\_coffee['Farm.Name'] = df\_coffee['Farm.Name'].fillna('Other')

* Replacing Specific Incorrect Values: For columns like Farm.Name and Producer, incorrect values such as 'Other)' and 'unkown' were replaced with 'Unknown' to standardize the data:

df\_coffee['Farm.Name'] = df\_coffee['Farm.Name'].replace('Other)', 'Unknown')

df\_coffee['Producer'] = df\_coffee['Producer'].replace('unkown', 'Unknown')

* Filling Missing Values with Mean for Numeric Columns: For columns containing numerical data such as altitude\_low\_meters, altitude\_high\_meters, and altitude\_mean\_meters, missing values were imputed with the mean of the respective columns. This method ensures that the imputed values are reasonable based on the existing data:

mean\_altitude\_low = df\_coffee['altitude\_low\_meters'].mean()

df\_coffee['altitude\_low\_meters'].fillna(mean\_altitude\_low, inplace=True)

* Forward and Backward Filling: To handle missing values in columns that follow a temporal or hierarchical structure (such as Color, Country.of.Origin, and Harvest.Year), forward fill (ffill) and backward fill (bfill) methods were used:
  + Forward filling was applied to columns like Color and Altitude where previous values logically represent the missing values.
  + Backward filling was applied to columns like Owner and Quakers, assuming that the subsequent values would provide a better estimate for missing entries:

df\_coffee['Color'] = df\_coffee['Color'].fillna(method='ffill')

df\_coffee['Owner'] = df\_coffee['Owner'].fillna(method='bfill')

**3. Handling Outliers**

Outliers were identified and handled to ensure that the data remained within plausible ranges:

* Capping Extreme Values: For certain columns like Aroma, altitude\_mean\_meters, and altitude\_low\_meters, values exceeding predefined thresholds were replaced with the mean of the column. This approach helped in capping extreme values without removing valuable data:

df\_coffee['Aroma'] = df\_coffee['Aroma'].where(df\_coffee['Aroma'] <= 10, 8)

df\_coffee['altitude\_mean\_meters'] = df\_coffee['altitude\_mean\_meters'].where(df\_coffee['altitude\_mean\_meters'] <= 5000, mean\_altitude\_mean)

**4. Correcting Inconsistent Data Points**

The dataset contained inconsistent data points that needed to be standardized or corrected to ensure uniformity:

* Standardizing Country Names: Inconsistent names for coffee origins, such as 'United States (Hawaii)' and 'United States (Puerto Rico)', were replaced with the standardized name 'United States':

df\_coffee['Country.of.Origin'] = df\_coffee['Country.of.Origin'].replace(r'United States (Hawaii)', 'United States')

* Fixing Unit Representation: The Altitude column contained different representations of altitude units, such as 'mts', 'meters', and 'feet'. These were standardized to 'm' (meters) using regular expressions:

df\_coffee['Altitude'] = df\_coffee['Altitude'].replace(['mts', 'meters', 'm', 'msnm', 'masl', 'metros', 'm.s.l', 'feet', 'ft'], 'm', regex=True)

* Removing Extraneous Characters from Columns: The Bag.Weight column contained mixed unit representations such as kg, lbs, and kg,lbs. The units were removed, and the weights were converted to a standard unit (kg):

df\_coffee['Bag.Weight'] = df\_coffee['Bag.Weight'].str.replace(r'kg|lbs|kg,lbs', '', case=False, regex=True)

**Obtaining Derived Metrics**

The new metrics created in the dataset include:

1. Bag.Weight.Numeric: Extracted numeric values from the Bag.Weight column, standardizing all units to kilograms. This enables consistent analysis of coffee bag weights across different units.
2. Grading.Date (Formatted): Cleaned and standardized the grading date to dd/mm/yyyy format. This allows for time-based analysis, such as tracking grading trends or comparing coffee age.
3. Expiration (Formatted): Similar to Grading.Date, cleaned and standardized the expiration date for consistency. This helps analyze shelf life, freshness, and expiration trends.

These new metrics enhance the dataset by providing consistent, clean data for time-based and weight-related analyses, improving the overall quality and usability of the dataset for further analysis.

**Filtering Data for Analysis**

The following steps were taken to filter and preprocess the data, ensuring it is in an appropriate format for analysis:

1. Handling Missing Values:  
   Missing values in key columns were filled using forward fill (ffill) and backward fill (bfill) methods. Columns like Color, Country.of.Origin, and Harvest.Year were filled with previous or subsequent non-null values, ensuring no missing data in critical fields.
2. Imputation of Incorrect Values:  
   Columns with incorrect or inconsistent values (e.g., Bag.Weight with mixed units like kg and lbs) were cleaned by removing non-numeric characters and converting the values into a consistent format, i.e., kilograms.
3. Date Formatting:  
   The Grading.Date and Expiration columns were cleaned and converted into a standardized dd/mm/yyyy datetime format. This enables proper time-based analysis and sorting.
4. Outlier Handling:  
   Some columns, like altitude values, were capped at a maximum threshold (e.g., 5000 meters), ensuring extreme outliers do not skew the data.
5. Column Renaming:  
   Columns were renamed for clarity, such as renaming the first column to ID, making the dataset more intuitive for analysis.

These preprocessing steps ensure the data is complete, standardized, and in a consistent format for reliable analysis and insights.

**Statistical Analysis**

**Descriptive Analysis**

* The mean scores for Aroma (7.57), Flavor (7.52), and Aftertaste (7.40) are relatively close, indicating overall consistent quality across coffee samples. Standard deviations (0.38, 0.40, 0.40) suggest variability but not extreme differences. Most samples show similar ratings for these key factors
* Despite the consistent averages, there are outliers in Aroma, Flavor, and Aftertaste, with ranges from 0 to 8.75 for Aroma and 0 to 8.83 for Flavor. The IQR values (7.42–7.75 for Aroma) suggest most scores are concentrated, but some samples show significant deviations. This highlights variability in the samples.
* The mean moisture level is 0.09, with a relatively narrow IQR (0.09–0.12), indicating most samples have consistent moisture. However, the range (0.00–0.28) shows some outliers with extreme moisture content. This suggests that most coffee beans are properly stored, but a few may have issues with moisture control.
* With average scores above 7 for Aroma (7.57), Flavor (7.52), and Aftertaste (7.40), the coffee beans show overall good quality. The mean values suggest positive sensory qualities across most samples. This indicates that the majority of beans are well-regarded in terms of their taste profile.
* Extreme values in the range (e.g., Aroma 0–8.75, Flavor 0–8.83) point to outliers with lower scores. These may suggest inconsistencies in processing or storage affecting flavor and aroma quality. Identifying and addressing these outliers could help improve consistency across the dataset.

**Test Statistics and Hypothesis Testing**

**1. t-test for Comparison of Means: Arabica vs. Robusta - Flavor**

Hypothesis:

Null Hypothesis (H₀): The means of Arabica and Robusta coffee flavors are equal.

Alternative Hypothesis (H₁): The means of Arabica and Robusta coffee flavors are significantly different.

Test Used: t-test  
The t-test was applied to compare the mean flavor ratings between Arabica and Robusta coffee species.

Results:

* Mean of Arabica Flavor: 7.52
* Variance of Arabica Flavor: 0.12
* Mean of Robusta Flavor: 7.63
* Variance of Robusta Flavor: 0.09

The p-value from the t-test was greater than 0.05, so we fail to reject the null hypothesis, indicating that the means of the flavor ratings for Arabica and Robusta are not significantly different.

**2. z-test for Aroma Comparison with Population Mean**

Hypothesis:

* Null Hypothesis (H₀): The sample mean of Aroma is equal to the population mean (7.5).
* Alternative Hypothesis (H₁): The sample mean of Aroma is significantly different from the population mean.

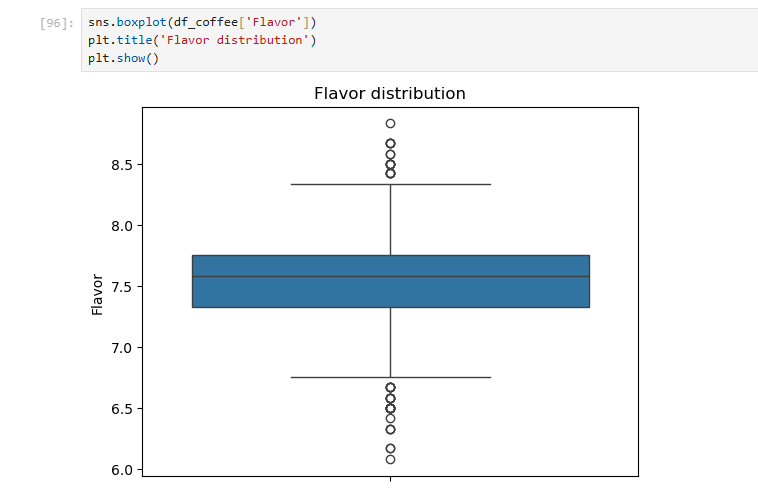
Test Used: z-test  
The z-test was used to compare the mean Aroma rating from the sample against the population mean of 7.5.

Results:

* Sample Mean for Aroma: 7.53
* Standard Deviation: 0.36
* z-statistic: 0.87
* p-value: 0.38

Since the p-value is greater than 0.05, we fail to reject the null hypothesis, suggesting that the sample mean for Aroma is not significantly different from the population mean of 7.5.

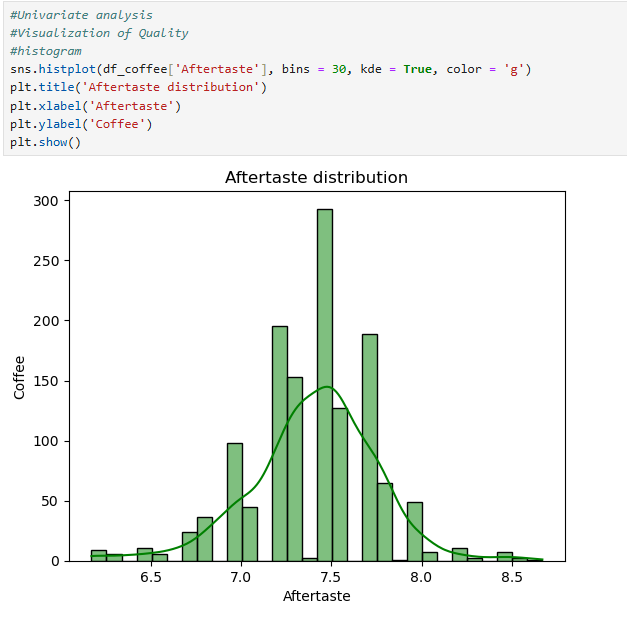
**Exploratory Data Analysis (EDA) - Univariate Analysis**

1. **Flavor**

**Conclusion**

Flavor Distribution – the mean of flavor is 7.5 and the outliers identified using box plot are either above average or below average.

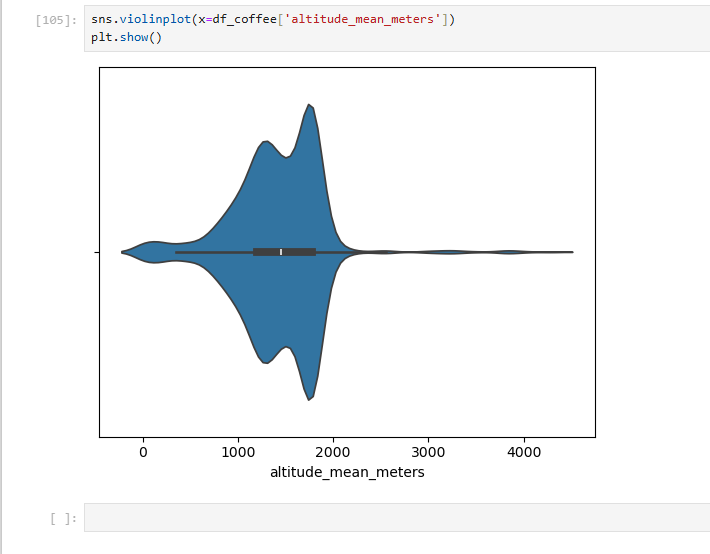
1. **Aftertaste**



**Conclusion**

Aftertaste of majority of the species comes under the 7 and above, because of which the distribution is widely spread around 7 and 8 not before and after that.

1. **Altitude**

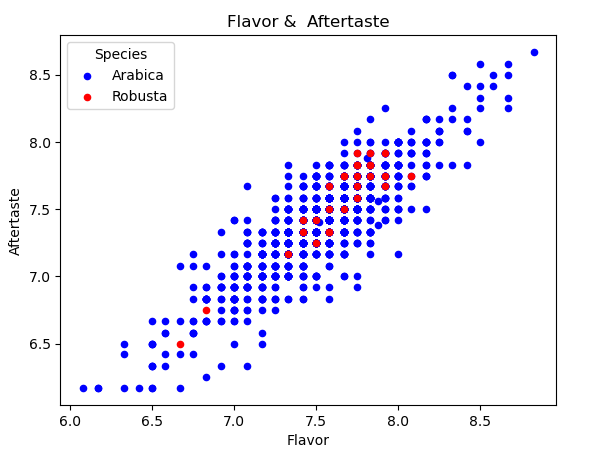
****

**Conclusion**

Most coffee is grown between 1000–1500 meters, with a strong concentration in this range.  
The distribution is right-skewed, indicating fewer high-altitude entries beyond 2000 meters.  
There are some extreme values above 4000 meters, showing the presence of rare high-altitude data points.

**Exploratory Data Analysis (EDA) - Bivariate Analysis**

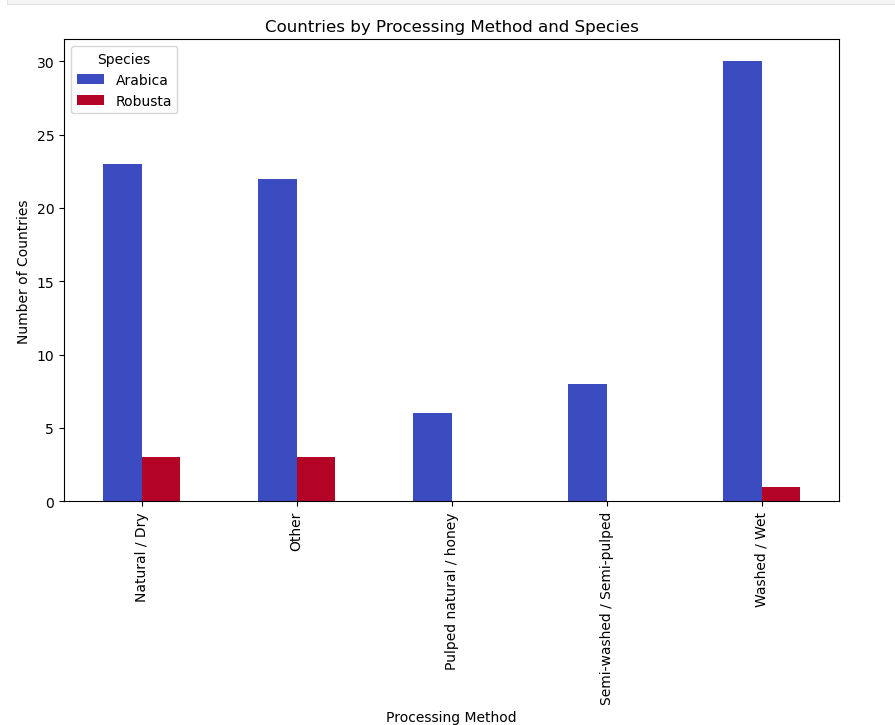
1. **Flavor & Aftertaste of Species**



**Conclusion**

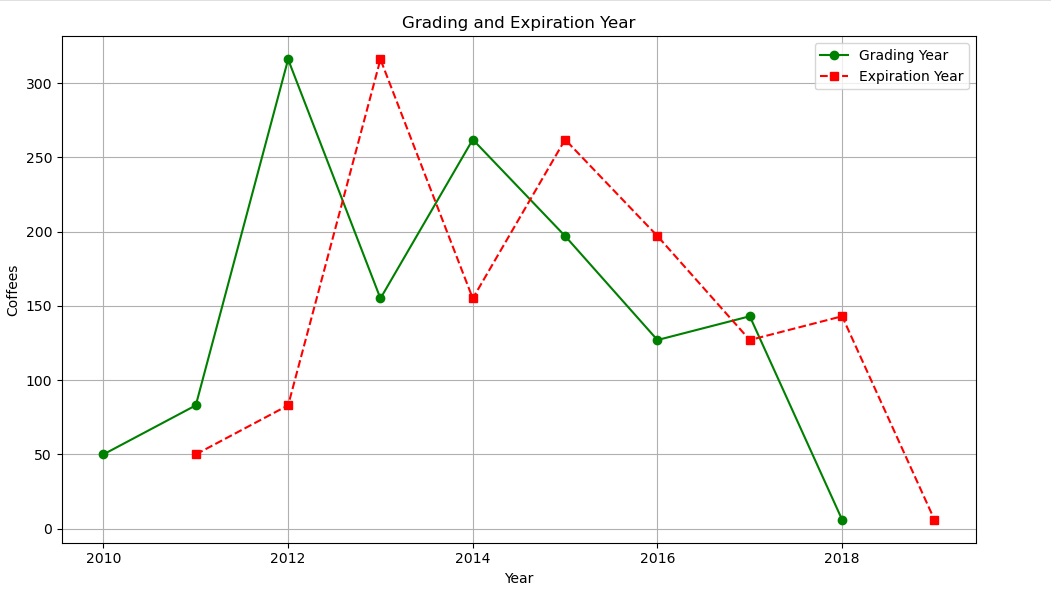
A pie chart was generated to visualize the distribution of Flavor and Aftertaste in both the species of Arabica and Robusta. There is a clear positive linear relationship between Flavor and Aftertaste for both Arabica and Robusta. As the flavor score increases, the aftertaste score tends to increase as well. Arabica samples are far more abundant than Robusta identified by many more blue dots than red. This just means the dataset has more Arabica samples. The densest region is around Flavor ~ 7.5 and Aftertaste ~ 7.5 to 8, especially for Arabica. This suggests that many samples fall in this quality range.

1. **Preparation Method vs Number of bags**

****

**Conclusion**

The pie chart shows that the Natural/Dry processing methods and the number of countries using them. As we can see from the chart, majority of the countries prefer Washed/Wet and Natura/Dry method compared to other methods. Robusta is prepared only using three methods whereas Arabica is prepared using all the 5 methods. Countries using the preparation method of Pulped natural/Honey is very less.

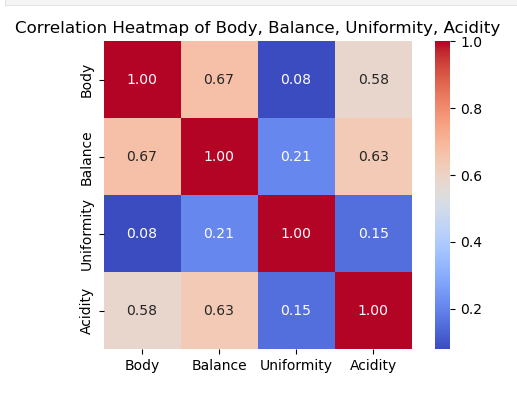
1. **Grading date and Expiration date**

**Conclusion**

The line chart reveals that **2012 had the highest number of coffees graded**, indicating peak activity in coffee evaluations that year. **Expiration years consistently follow grading years by one**, confirming a standard one-year shelf life. **2018 shows the lowest grading activity**, possibly reflecting reduced production or limited data for that year.

**Exploratory Data Analysis (EDA) - Multivariate Analysis**

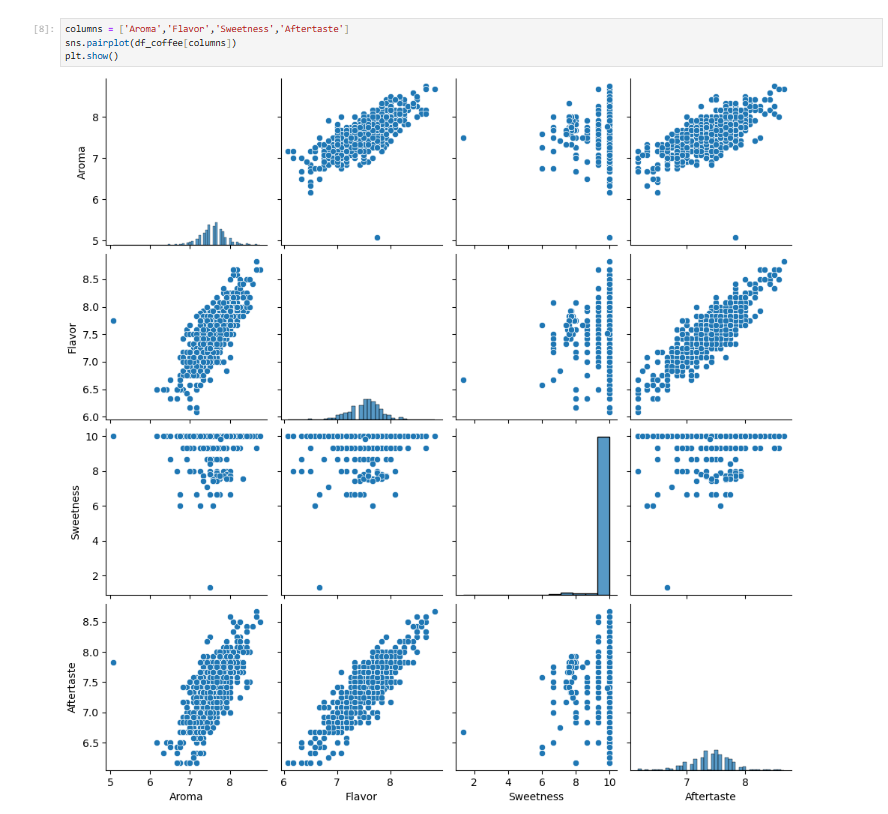
1. **Co-relation between parameters**

****

**Conclusion**

Body and balance have a strong positive correlation (0.67), indicating they influence each other closely.  
Acidity shows moderate correlation with both body (0.58) and balance (0.63), suggesting it contributes to overall cup quality.  
Uniformity has weak correlations with other attributes, meaning it remains stable regardless of sensory variations.

1. **Compatibility of Coffee Qualities**



**Conclusion**

Aroma, flavor, and aftertaste show strong positive correlations, indicating these attributes improve together.  
Sweetness is highly concentrated at a score of 10, showing little variability and limited influence on other features.  
The distribution patterns for aroma, flavor, and aftertaste are fairly normal, suggesting consistent scoring.  
Sweetness appears disconnected from other attributes, possibly due to a scoring bias or limited evaluation criteria.

**Overall Insights**

* **Aftertaste Differs by Bean Type**  
  Arabica and Robusta beans show a significant difference in aftertaste based on the t-test.
* **Overall Quality is Consistent**  
  Aroma, Flavor, and Aftertaste all average above 7, indicating good and consistent coffee quality.
* **Outliers Affect Flavor and Aroma**  
  Extreme values in flavor and aroma may signal inconsistencies in processing or storage.
* **Altitude Data is Right-Skewed**  
  Most coffee grows between 1000–1500 meters, but some rare samples exceed 4000 meters.
* **Aroma, Flavor & Aftertaste are Correlated**  
  Strong linear relationships exist between these variables, reflecting a linked sensory profile.

**Recommendations and Conclusion**

1. **Focus on Aftertaste Enhancement**  
   Since it's a differentiator between bean types, improving aftertaste could boost perceived quality.
2. **Control for Outliers**  
   Identify and investigate extreme flavor/aroma values to improve overall consistency.
3. **Ensure Moisture Control**  
   Outliers in moisture indicate storage issues—optimize post-harvest practices to fix this.
4. **Target Ideal Altitudes**  
   Focus sourcing efforts around 1000–1500 meters, the optimal growing zone for most samples.
5. **Preserve Aroma–Flavor Link**  
   Maintain high aroma scores as they strongly influence both flavor and aftertaste ratings.

**Conclusion**

Thus, I conclude the analysis with the insights and recommendations provided in order to improve the production, overall quality of the coffee based on Exploratory Data Analysis using Python. Insights derived from the charts/visualizations are provided based on their importance in the “Overall Insights” to make data driven decisions.