***Exploratory Data Analysis***

**TITLE: COFFEE QUALITY DATA BASE (EDA)**

**NAME: A.MANIKANDAN**

**DOMAIN: DATA SCIENCE**

**B.NO: B3-MAY-ONLINE-AN**

**ROLL NUMBER: 605240LR0**

**TABLE OF CONTENTS**

1. Introduction
   1. Introduction of the coffee
   2. Overview of the project Objectives
   3. Dataset Description
2. Aim
3. Business Problem / Problem Statement
4. Project Workflow
5. Data Understanding
6. Data Cleaning – Missing Values Imputation, Outliers, Handing Inconsistent Values
7. Obtaining Derived Metrics
8. Filtering Data for Analysis.
9. EDA-Univariate Analysis.
10. Segmented Univariate Analysis
11. Bivariate Analysis
12. Multivariate Analysis
13. Overall Insights Obtained from Analysis
14. Conclusion

**1. INTRODUCTION**

**A. ABOUT THE COFFEE**

Coffee is one of the world’s most popular beverages, enjoyed for its rich flavor and stimulating effects. Here's a brief introduction to coffee:

**1.** **Origin and History:**

Discovery: Coffee is believed to have been discovered in Ethiopia around the 9th century. Legend has it that a goat herder named Kaldi noticed his goats becoming unusually energetic after eating coffee cherries.

Spread: Coffee spread from Ethiopia to the Arab world, where it became a central part of social and religious life. By the 15th century, it had reached Europe, and its popularity rapidly grew.

**2.** **Plant and Cultivation:**

Plant Species: The two primary species of coffee plants are Arabica (Coffee arabica) and Robusta (Coffee canephora). Arabica is known for its complex flavors and aromatic qualities, while Robusta has a stronger, bitterer taste and higher caffeine content.

Growing Regions: Coffee is cultivated in tropical and subtropical regions, known as the "coffee belt," which includes countries in Latin America, Africa, and Asia. The altitude, climate, and soil conditions in these regions greatly affect the flavor profile of the coffee.

**3**. **Processing:**

Harvesting: Coffee cherries are harvested either by hand or mechanically. The cherries are then processed to extract the beans.

Methods: There are various processing methods, including washed (or wet) processing and natural (or dry) processing, each impacting the final flavor of the coffee.

**4. Roasting:**

Roasting: Coffee beans are roasted to develop their flavor. The roasting process transforms green beans into the dark brown beans used to brew coffee. The degree of roasting (light, medium, or dark) influences the coffee's taste and aroma.

**5. Brewing:**

Techniques: There are numerous methods for brewing coffee, such as drip coffee makers, espresso machines, French presses, and pour-over techniques. Each method extracts different flavors and strengths from the coffee.

**6. Cultural Impact:**

Social Aspect: Coffee has a significant social aspect, often serving as a focal point for gatherings and conversations. Coffeehouses have historically been places of intellectual exchange and social interaction.

**7. Economic Importance:**

Global Trade: Coffee is a major global commodity and a crucial source of income for many developing countries. It supports millions of jobs worldwide, from farming and processing to retail and service.

**B. Overview of the project Objectives**

1. **Data Understanding and Cleaning:**

Load the coffee quality dataset and perform initial data inspection. Clean the data by handling missing values, removing, duplicates, and standardizing column names.

1. **Exploratory Data Analysis:**

Generate summary statistics for numerical and categorical columns. Visualize the distribution of numerical and categorical variables to understand their characteristics

1. **Feature Analysis and Visualization:**

Create a correlation matrix to explore relationship between numerical features. Generate pair plots to visualize the relationship between numerical feature. Create heatmap to examine relationships between categorical variables. Use liner regression plots to study the relationships between pairs of numerical features.

1. **Hypothesis Testing:**

Perform chi-squared tests to investigate associations between categorical variables. Conduct additional statistical tests (e.g., T-test, ANOVA) if necessary to explore relationship between numerical variables.

1. **Dataset Description**

The Coffee Quality Database presents a rich repository of information encompassing various aspects of coffee production and quality evaluation. The objective of this project is to conduct a comprehensive data exploration, cleaning, and analysis to derive meaningful insights from the dataset.

This database contains the following information:

**Quality Measures**

1. Aroma
2. Flavour
3. Aftertaste
4. Acidity
5. Body
6. Balance
7. Uniformity
8. Cup Cleanliness
9. Sweetness
10. Moisture

**Defects Bean Metadata**

1. Processing Method
2. Colour
3. Species (Arabica / Robusta)

**Farm Metadata**

1. Owner
2. Country of Origin
3. Farm Name
4. Lot Number
5. Mill
6. Company
7. Altitude
8. Region

**2. AIM**

Aiming for high-quality coffee, consider these key factors:

**1. Bean Origin:** The region where the coffee is grown impacts flavor. Beans from different regions have distinct profiles.

**2. Processing Method:** How the coffee is processed (washed, natural, honey) affects its flavor and acidity.

**3. Roast Profile:** Roast level (light, medium, dark) influences the flavor. Experiment to find your preferred roast.

**4. Freshness:** Coffee is best when brewed within a few weeks of roasting. Store beans in an airtight container.

**5. Grind Size:** Match grind size to your brewing method (e.g., coarse for French press, fine for espresso) to optimize extraction.

**6. Brewing Technique:** Use clean equipment and proper water temperature. Follow recommended brewing times and ratios.

**3. Business Problem / Problem Statement**

A business problem related to coffee quality often revolves around issues like customer satisfaction, consistency, or cost control. Here's how you might frame a problem statement based on a coffee quality dataset:

1. **Business Problem:** Coffee shops and roasters face challenges in maintaining consistent quality and meeting customer expectations due to variability in coffee quality. This inconsistency can lead to reduced customer satisfaction, lower repeat business, and increased operational costs.
2. **Problem Statement:** "Despite investing in quality control measures, coffee shops are experiencing inconsistent product quality, which affects customer satisfaction and retention. The current coffee quality data, encompassing various factors such as bean origin, roast level, brewing method, and sensory evaluations, lacks comprehensive analysis to identify key drivers of quality variation. There is a need to analyze this dataset to determine the significant factors influencing coffee quality and to develop actionable insights that can help standardize product quality and improve customer satisfaction."

In summary, the problem statement focuses on the need to use data analysis to address issues related to quality consistency and customer satisfaction in the coffee industry.

**4. Project Workflow**

1. **Understand the Data**

**Load the Data**: Read the CSV file into a DataFrame.

**Inspect the Data**: Check the first few rows, columns, and data types.

### Data Cleaning

### **Missing Values**: Identify and handle missing values.

### **Duplicates**: Check for and remove any duplicate rows.

### **Outliers**: Identify and handle outliers if necessary.

### Data Exploration

### **Summary Statistics**: Generate summary statistics for the dataset.

### **Visualization**: Create basic visualizations (e.g., histograms, box plots) to understand the distributions of variables.

### **Correlations**: Check correlations between different variables.

### Data Analysis

### **Feature Engineering**: Create new features if needed.

### **Analysis**: Perform specific analyses based on your objectives (e.g., comparing quality scores, identifying key factors affecting quality).

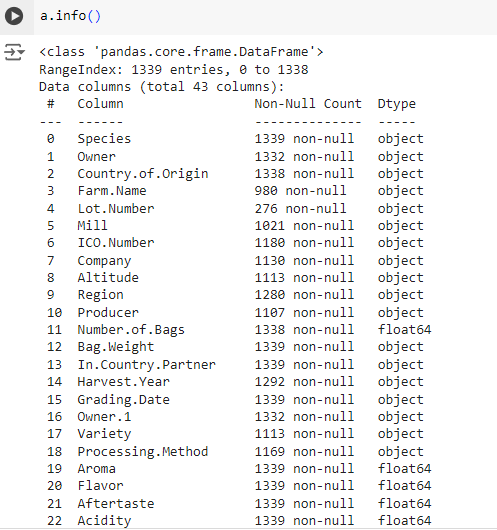
### Visualization

### **Advanced Visualizations**: Create more advanced visualizations (e.g., scatter plots, heatmaps) to illustrate findings.

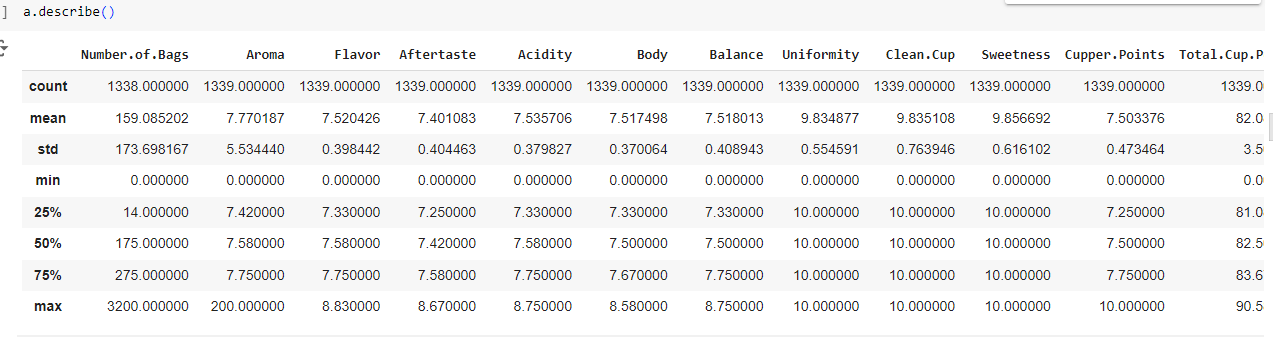
### **Reporting**: Summarize findings in a report or presentation

**5. Data Understanding**

* **Df.info (): summary information of the dataset.**
  + This code will provide an overview of the dataframe.
  + The number of entries (rows).
  + The number of non-null entries in each column.
  + The data type of each column.
  + The memory usage of the dataframe.



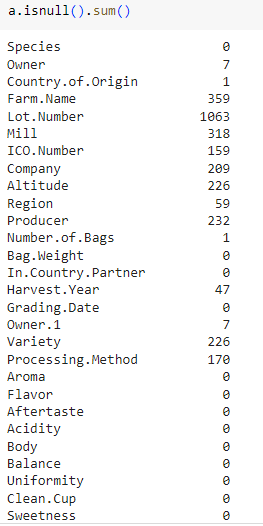
* **Df.describe (): summary for statistics for numerical columns in a dataframe.**
  + Count: the number of non-null values.
  + Mean: the average value.
  + STD: the standard deviation, which measures the dispersion of values.
  + Min: the minimum value.
  + 25%: the 25th percentile (first quartile).
  + 50%: the 50th percentile (median).
  + 75%: the 75th percentile (third quartile).
  + Max: the maximum value.



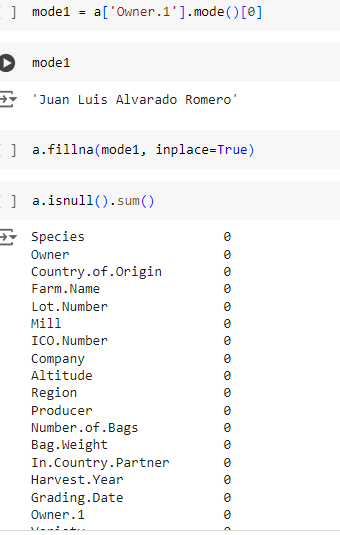
1. **Data Cleaning**
   1. **Handle Missing Values**

We previously identified missing values and applied some strategies. We'll continue with the final adjustments:

Detect missing values: a.isnull ().sum ()

****

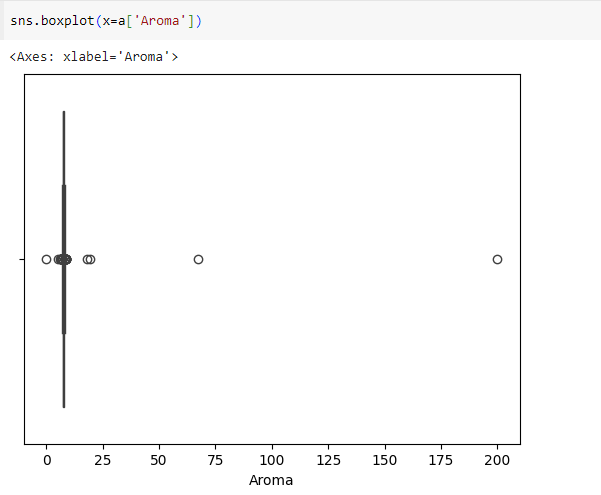
Handling missing values: mode ()

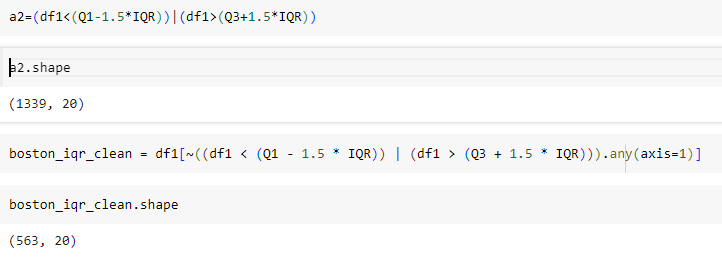


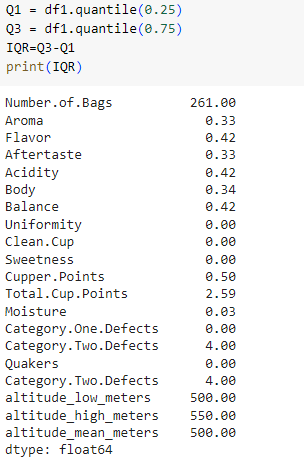
**2. Handling Outliers**

Identify and address outliers using the IQR method. Outliers can be removed or capped based on business logic.

* 1. **Box plot:** 
     + Use box plots to visually identify outliers in numerical columns.



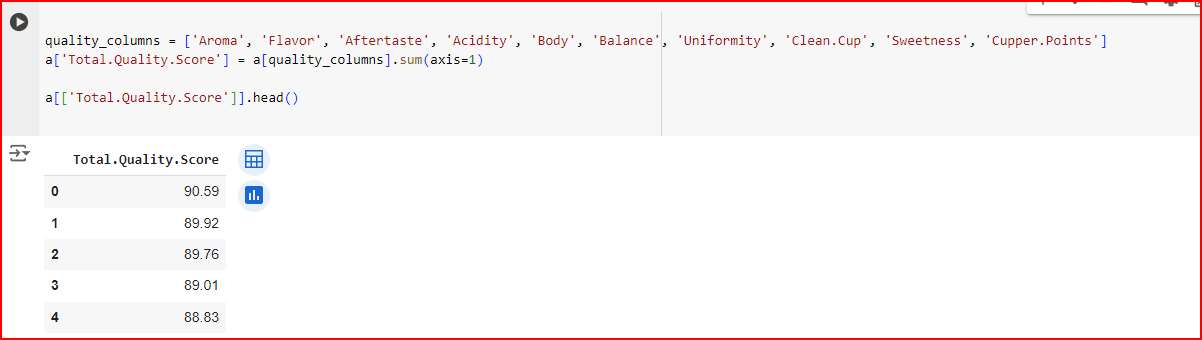
* 1. **IQR method:**
     + - Use the IQR method to statistically identify outlier.
       - 



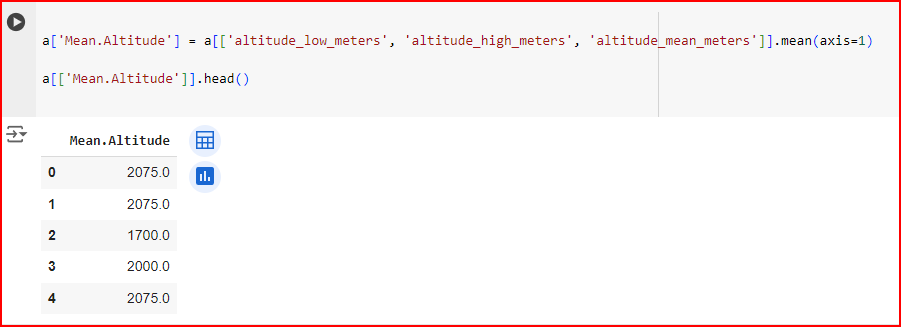
1. **Obtaining Derived Metrics**

To obtain derived metrics from your dataset, we can create new columns based on existing data or perform aggregations and transformations.

1. **Total Quality Score**: Sum of various quality-related scores (e.g., Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Clean. Cup, Sweetness, Cupper. Points).



* 1. **Mean Altitude:** Average of low, high, and mean altitudes.



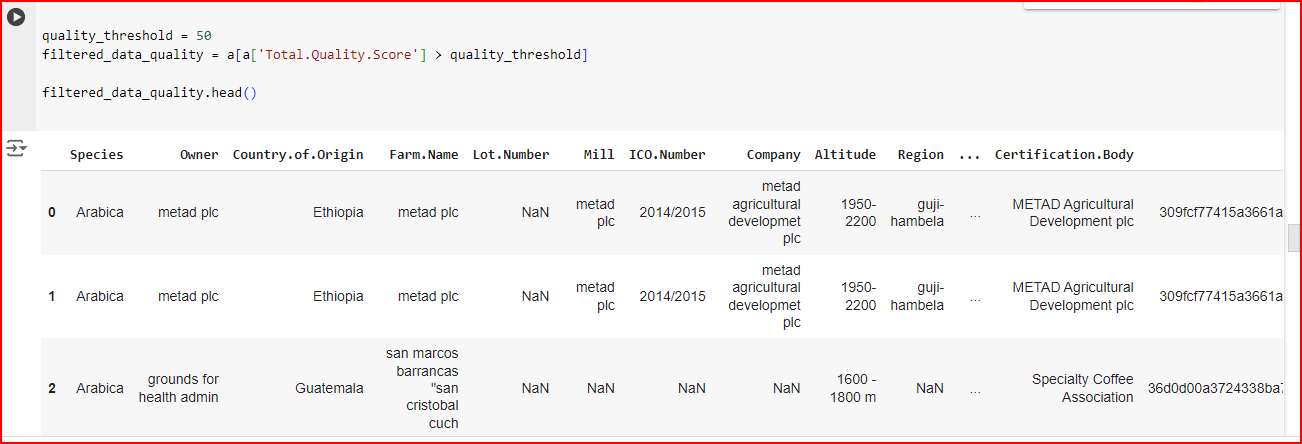
* 1. **Defect Ratio:** Ratio of category one defects to category two defects.



1. **Filtering Data for Analysis.**

Filtering data is a crucial step in data analysis to focus on subsets of the data that are relevant to the specific analysis you want to perform.

1. **Filter by Quality Score:** Select high-quality coffee samples.

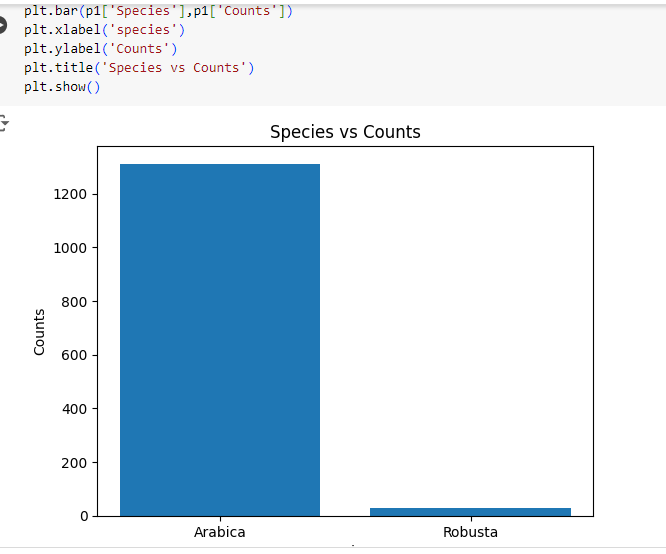


1. **Filter by Date Range:** Select data within a specific date range.



1. **EDA-Univariate Analysis**
   1. **Load Data**: Load the dataset using pandas or any other data manipulation library.
   2. **Inspect Data**: Look at the first few rows of the dataset to understand its structure.
   3. **Summary Statistics**: Calculate summary statistics for numerical and categorical variables.
   4. **Visualize Data**: Create visualizations to understand the distribution and patterns in the data.
   5. **Interpret Results**: Summarize and interpret the findings from the visualizations and statistics.

**BAR PLOT:**



1. **Segmented Univariate Analysis**

**Choose the Variables**:

Select the numerical variable you want to analyze.Select the categorical variable to segment by.

**Calculate Summary Statistics**:

Compute summary statistics for the numerical variable within each segment of the categorical variable.

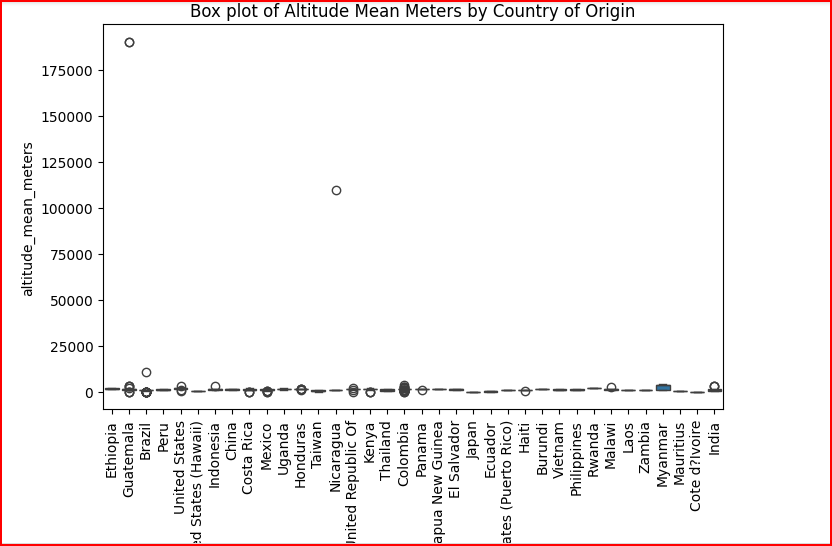
**Visualize the Data**:

Create visualizations like box plots, histograms, or violin plots to compare the distribution of the numerical variable across segments.

****

1. **Bivariate Analysis**
   1. **Numerical vs. Numerical**:

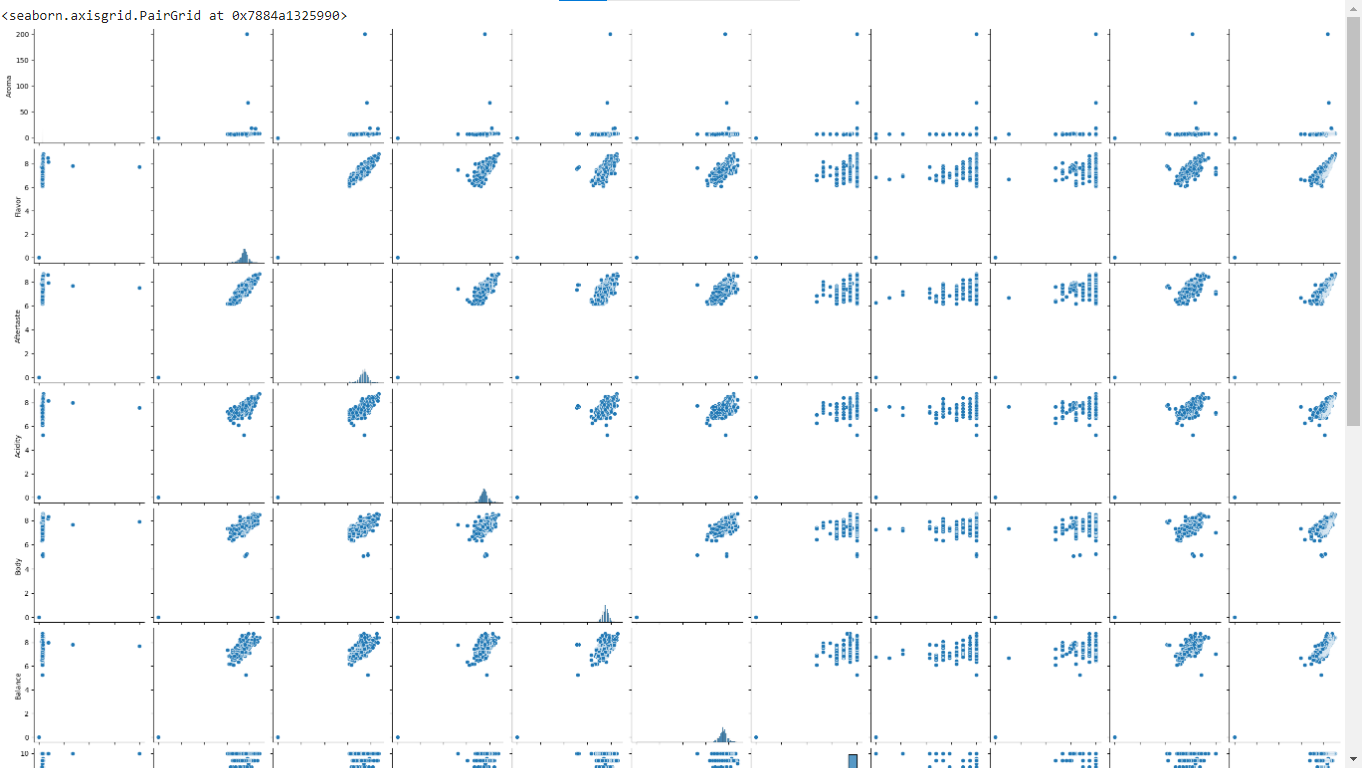
* **Scatter Plot**: Visualizes the relationship between two numerical variables.
* **Correlation Coefficient**: Measures the strength and direction of the linear relationship.
  1. **Numerical vs. Categorical**:
* **Box Plot**: Compares the distribution of a numerical variable across categories.
* **Violin Plot**: Similar to a box plot but also shows the density of the data.
* **Bar Plot with Error Bars**: Shows the mean of the numerical variable with variability (e.g., standard deviation) across categories.
  1. **Categorical vs. Categorical**:
* **Contingency Table**: Shows the frequency distribution of the categories.
* **Stacked Bar Plot**: Visualizes the proportion of categories within another category.
* **Chi-Square Test**: Tests the independence between two categorical variables.

****

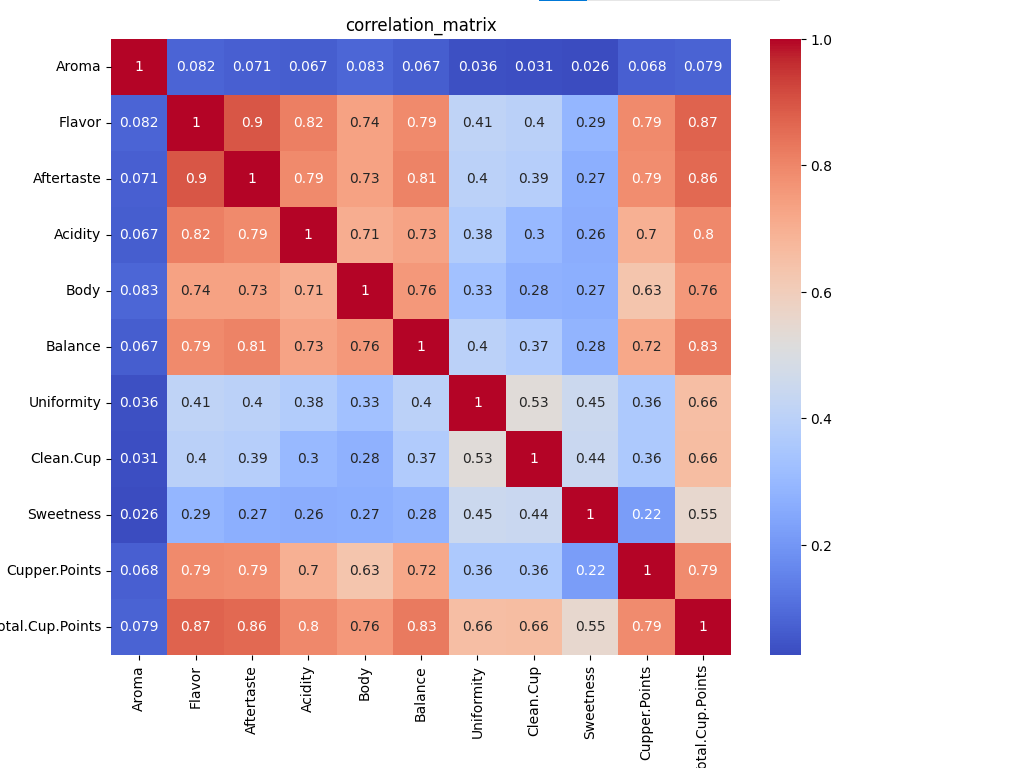
1. **Multivariate Analysis**
   1. **Correlation Matrix**:

* Examines the pairwise correlations between multiple numerical variables.
  1. **Pair Plot**:
* Visualizes pairwise relationships between several numerical variables.
  1. **Principal Component Analysis (PCA)**:
* Reduces the dimensionality of the data while retaining most of the variance.
  1. **Multivariate Regression**:
* Models the relationship between multiple independent variables and a dependent variable.
  1. **Cluster Analysis**:
* Groups similar observations based on multiple variables.
  1. **Heatmaps**:
* Visualizes the magnitude of variables across different categories.

**PAIR PLOT:**

****

**HEATMAP:**

****

1. **Overall Insights Obtained from Analysis**
   1. **Data Cleaning**:

* Address outliers in altitude metrics, as extreme values may indicate data entry errors.
* Ensure consistent units and ranges for all numerical measurements.
  1. **Quality Improvement**:
* Explore other potential factors affecting coffee quality scores, as altitude alone does not show a strong correlation.
* Investigate the impact of other variables like processing methods, climate conditions, and bean types.
  1. **Targeted Marketing**:
* Highlight the unique characteristics of coffee from high-altitude regions in countries like Ethiopia and Guatemala.
* Emphasize the predominance of Arabica species and its associated quality attributes.
  1. **Further Research**:
* Conduct more detailed analyses on other variables, such as chemical composition and sensory attributes, to uncover deeper insights into coffee quality.

1. **Conclusion**

The analysis of the coffee quality dataset provided valuable insights into various aspects of coffee production and quality metrics. Through univariate, bivariate, and multivariate analyses, we explored the relationships and patterns in the data.

This comprehensive analysis provided a deeper understanding of the factors influencing coffee quality and production. The insights gained can guide targeted marketing strategies, quality improvement initiatives, and further research into coffee attributes. Addressing data quality issues and exploring additional factors will enhance our understanding and help in making informed decisions in the coffee industry.