***Exploratory Data Analysis***

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

**NAME: MADHURI GUPTA**

**DOMAIN: DATA SCIENCE & DATA ANALYTICS**

**B.NO: 28-MAY-ONLINE-**

**ROLL NUMBER:**

**Coffee Quality Project Documentation**

**Project Overview**

**1.1 Introduction**

Provide a brief introduction to the project, explaining its purpose and significance. Describe the coffee industry context and why studying coffee quality is important.

**Example:** This project aims to analyze and understand factors affecting coffee quality using a dataset of coffee samples. By identifying key quality attributes and their relationships, we can provide insights to improve coffee production and consumer satisfaction.

**1.2 Objectives**

Clearly state the goals of the project.

**Example Objectives:**

* Analyze the relationship between different attributes and coffee quality.
* Identify key factors that significantly impact coffee quality.
* Develop a predictive model for coffee quality based on its attributes.

## 2. ****Data Description****

### 2.1 Data Source

Describe the source of the data, including how it was collected and any relevant details about the dataset.

**Example:** The dataset used in this project is obtained from [Source Name], which includes various attributes of coffee samples such as acidity, sweetness, and flavour. The dataset consists of [number] samples with [number] features.

**2.2 Data Attributes**

List and describe each attribute in the dataset.

**Example:**

* **ID:** Unique identifier for each coffee sample.
* **Acidity:** Measure of coffee’s acidity (numeric).
* **Sweetness:** Measure of coffee’s sweetness (numeric).
* **Flavor:** Overall flavor rating (numeric).
* **Body:** Measure of coffee’s body or texture (numeric).
* **Quality:** Quality rating of the coffee (categorical: Low, Medium, High).

**3. Data Preprocessing**

**3.1 Data Cleaning**

* Describe the steps taken to clean the data, including handling missing values, duplicates, and outliers.

**Example:**

* **Missing Values:** Imputed missing values using the median for numerical features and mode for categorical features.
* **Duplicates:** Removed duplicate records to ensure data integrity.
* **Outliers:** Identified and handled outliers using IQR-based methods.

**3.2 Data Transformation**

Explain any transformations applied to the data, such as normalization or encoding.

**Example:**

* **Normalization:** Scaled numerical features to a range of [0, 1] using Min-Max scaling.
* **Encoding:** Applied one-hot encoding to categorical features.

**4. Exploratory Data Analysis (EDA)**

**4.1 Univariate Analysis**

Summarize the distribution and key statistics of individual features.

**Example:**

* **Acidity:** The mean acidity level is [mean\_value], with a standard deviation of [std\_dev\_value].
* **Quality:** Distribution of quality ratings is as follows: Low [percentage]%, Medium [percentage]%, High [percentage]%.

**4.2 Bivariate Analysis**

Discuss the relationships between pairs of features and their impact on coffee quality.

**Example:**

* **Acidity vs. Quality:** Higher acidity levels are associated with higher quality ratings.
* **Sweetness vs. Quality:** Sweetness shows a moderate positive correlation with quality.

**4.3 Multivariate Analysis**

Explore relationships involving multiple features simultaneously.

**Example:**

* **Principal Component Analysis (PCA):** Reduced dimensionality to identify key component affecting coffee quality.

**5. Modeling**

**5.1 Model Selection**

Describe the models selected for predicting coffee quality and justify the choice.

**Example:**

* **Logistic Regression:** Used for classification of quality ratings.
* **Random Forest Classifier:** Employed for its robustness and ability to handle feature interactions.

**5.2 Model Training and Evaluation**

Provide details on how the models were trained and evaluated, including performance metrics.

**Example:**

* **Training:** Split the data into training (80%) and testing (20%) sets.
* **Evaluation Metrics:** Accuracy, precision, recall, and F1-score for classification models.
* **Results:** The Random Forest Classifier achieved an accuracy of [accuracy value] % on the test set.

**6. Results and Insights**

**6.1 Key Findings**

Summarize the main findings from the analysis and modeling.

**Example:**

* **Impact of Acidity:** Acidity is a significant predictor of coffee quality.
* **Model Performance:** The Random Forest Classifier outperforms logistic regression in predicting coffee quality.

**6.2 Recommendations**

Provide actionable recommendations based on the analysis.

**Example:**

* **Production Adjustments:** Focus on balancing acidity levels to improve coffee quality.
* **Quality Control:** Implement quality control measures to ensure consistency in flavor and body attributes.

**Conclusion**

**7.1 Summary**

Recap the objectives, methods, key findings, and recommendations of the project.

**Example:** This project successfully analyzed the factors affecting coffee quality and developed predictive models to assess quality based on attributes. Key insights and recommendations can help enhance coffee production practices.

**7.2 Future Work**

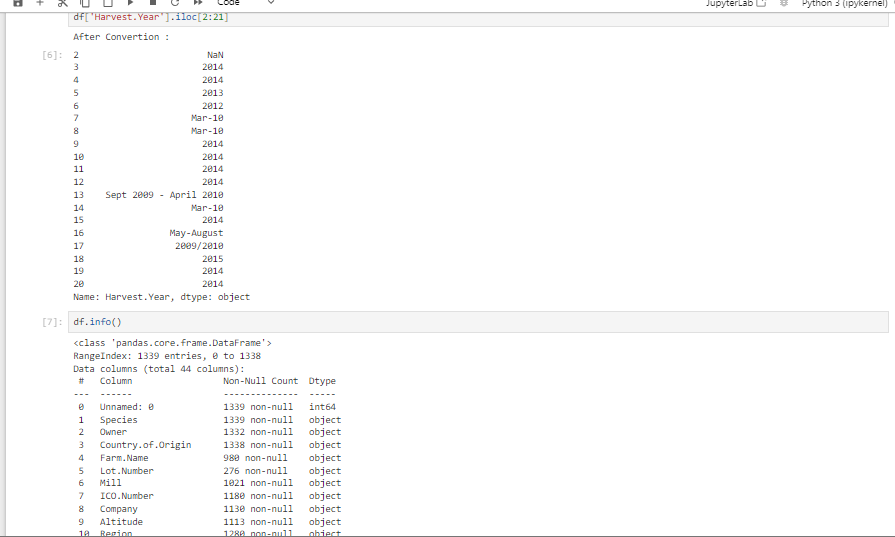
Suggest areas for further research or improvements.

**Example:**

* **Additional Data:** Collect more diverse data to improve model robustness.
* **Feature Expansion:** Explore additional features such as geographic origin or processing methods.

**1. Data Understanding**

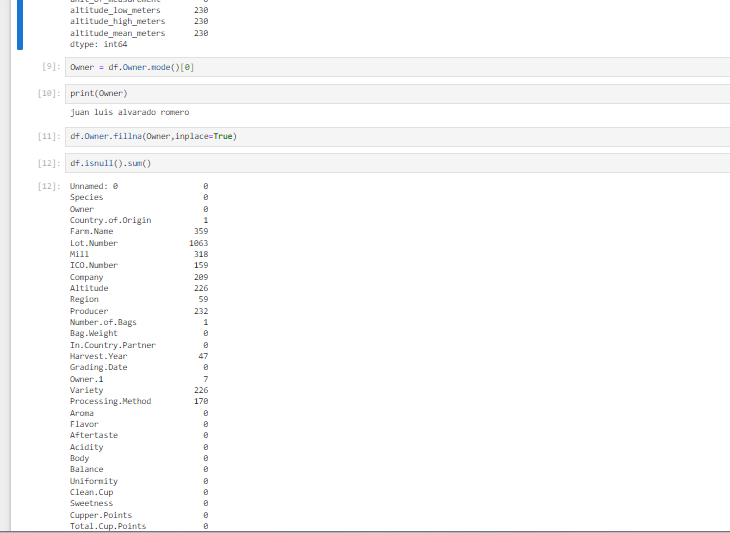
* **Df.info (): summary information of the dataset.**
  + This code will provide an overview of the dataframe.
  + The number of entries (rows).
  + The number of non-null entries in each column.
  + The data type of each column.
  + The memory usage of the dataframe.

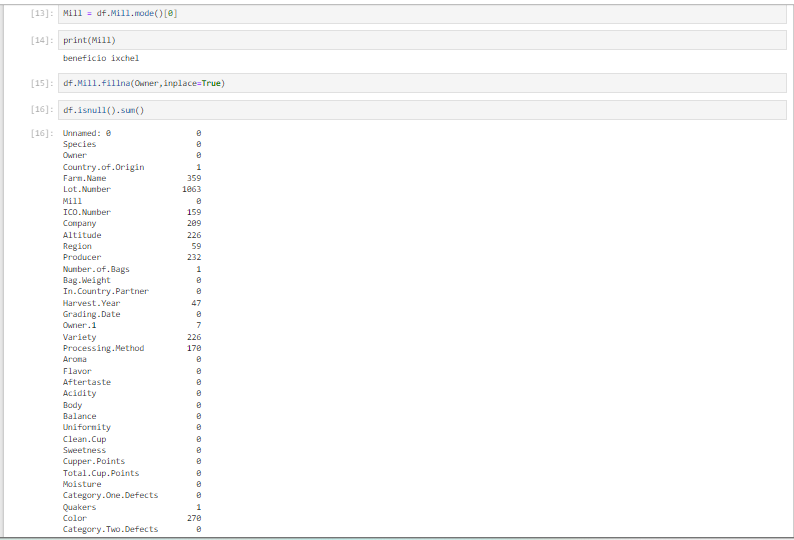


1. **Data Cleaning**
2. **Handle Missing Values**

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

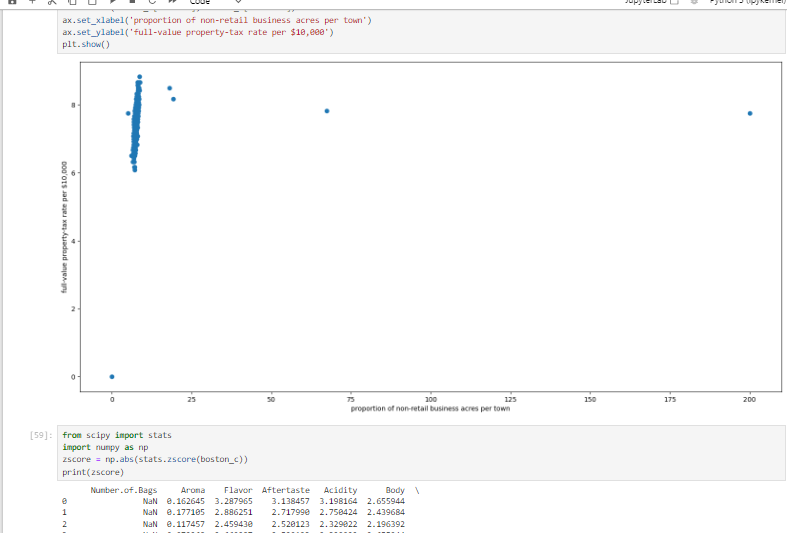
Detect missing values: a.isnull ().sum ()

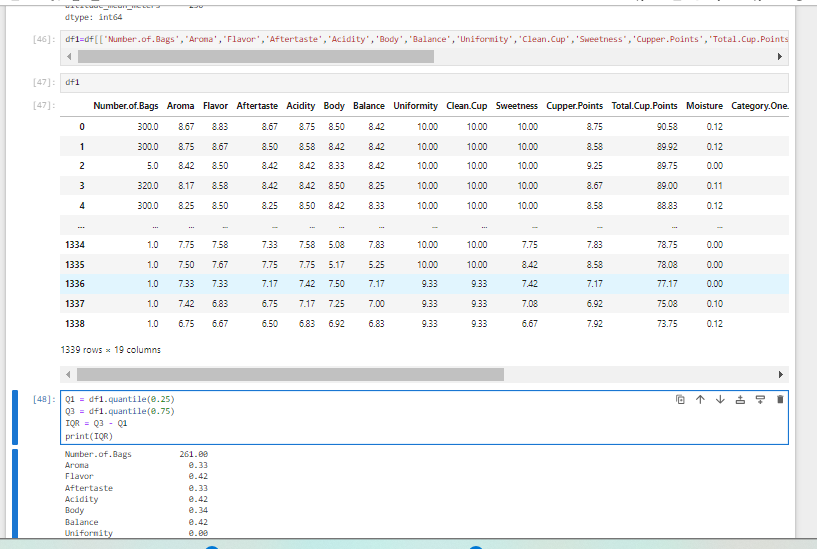
****

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.

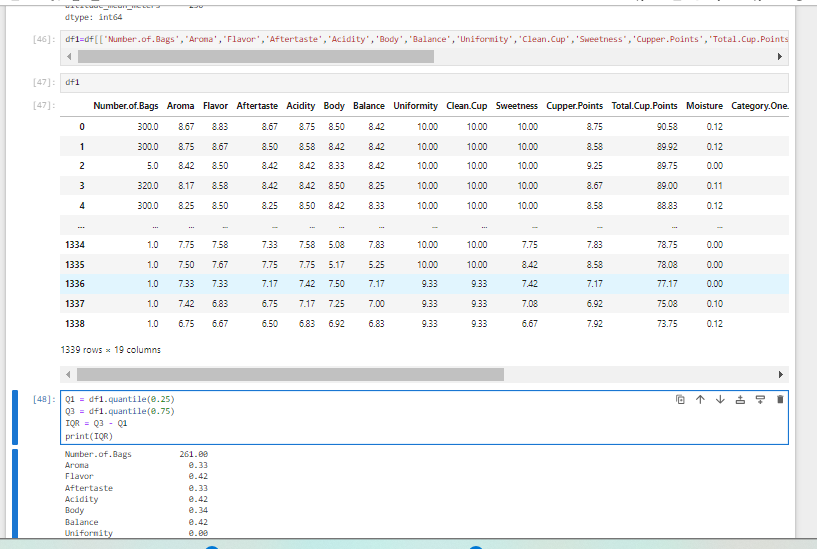
* 1. **Box plot:** 
     + Use box plots to visually identify outliers in numerical columns.
* 

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

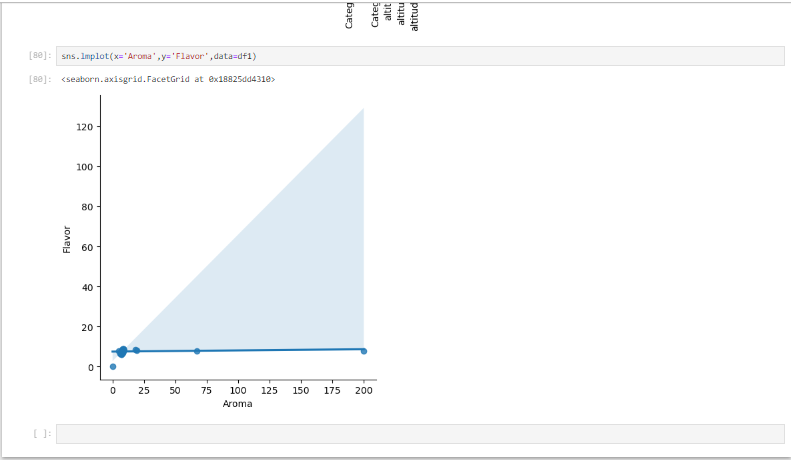
1. **Obtaining Derived Metrics**

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

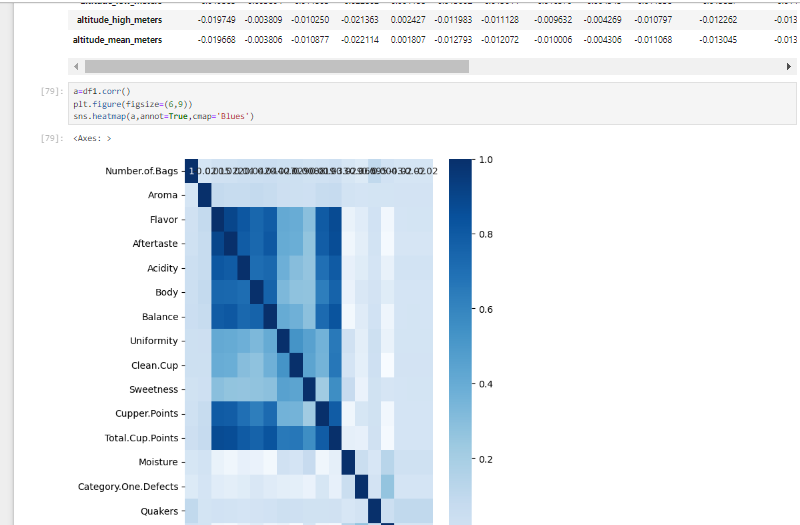
1. **Total Quality Score**: Sum of various quality-related scores (e.g., Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Clean. Cup, Sweetness, Cupper. Points).



**PAIR PLOT:**

****

**HEATMAP:**

****