Project Documentation: Exploratory Data Analysis using Python :

|  |  |
| --- | --- |
| **TITLE** | Exploratory Affecting Factors of Coffee Quality |
| **NAME** | PARTHASARTHI A |
| **COURSE** | DA-DS |
| **BATCH** | MAY |

**TABLE OF CONTENT**

|  |  |
| --- | --- |
| **S.NO** | **CONTENT** |
| 1 | Introduction |
| 2 | Aim |
| 3 | Business Problem / Problem Statement |
| 4 | Project Workflow |
| 5 | Data Understanding |
| 6 | Data Cleaning   * Missing Values Imputation * Outlier Treatment * Handling Inconsistent Values |
| 7 | Obtaining Derived Metrics |
| 8 | Filtering Data for Analysis |
| 9 | Statistical Analysis   * Descriptive analysis * Test statistics and hypothesis testing |
|  | Exploratory Data Analysis (EDA) |
| 10 | Univariate Analysis |
| 11 | Bivariate Analysis |
| 12 | Multivariate Analysis |
| 13 | Overall Insights from Analysis |
| 14 | Conclusion |

|  |  |
| --- | --- |
| 1. | INTRODUCTION |

The global coffee industry depends heavily on understanding and maintaining quality standards. Coffee quality is influenced by multiple factors, including species, processing methods, farm attributes, and sensory evaluations such as Aroma, Flavor, Aftertaste, and Body. Evaluating these factors systematically can help producers, exporters, and researchers improve quality control and make data-driven decisions.

This project uses a comprehensive coffee quality dataset containing sensory scores, bean characteristics, and farm metadata. The analysis aims to uncover patterns, relationships, and statistical significance in coffee quality attributes, helping identify the most influential factors.

The project workflow involves:

* **Data cleaning** to handle missing, invalid, and inconsistent values.
* **Deriving new metrics** for better insight, such as altitude ranges.
* **Exploratory data analysis (EDA)** using univariate, bivariate, and multivariate techniques.
* **Hypothesis testing** (T-test, ANOVA, Chi-square) to validate observed differences and relationships.
* **Visualization** to communicate findings effectively.

|  |  |
| --- | --- |
| 2. | AIM |

The aim of this project is to carefully study and understand the different factors that affect the quality of coffee beans by using the **Coffee Quality Dataset**. Coffee quality is judged based on many sensory features such as **aroma, flavor, aftertaste, acidity, body, balance, and total cup points**, along with other details like **species, altitude, bag weight, and country of origin**.

Our main goal is to find out **which factors have the biggest impact on overall coffee quality** and how they differ across various countries and species. We also want to discover patterns that can help **coffee growers, exporters, and buyers** make better decisions about production and quality control.

To achieve this, we will perform **data cleaning, statistical analysis, and visualizations**. These steps will help turn raw data into meaningful insights, making it easier to understand what makes good-quality coffee.

|  |  |
| --- | --- |
| 3. | STATEMENT OF PROMBLEM |

Coffee is enjoyed by millions of people every day, yet the quality of each cup can differ widely. Factors such as the **species of the bean, altitude where it’s grown, country of origin, and even moisture content** all play a role in shaping the final taste, aroma, and overall quality. For farmers, exporters, and buyers, understanding these factors is often challenging. Many struggle to identify what truly makes a coffee bean “high quality” and how to achieve consistency in every harvest.

When this knowledge is missing, decisions about **cultivation methods, quality control, and pricing** may be based on guesswork. This can lead to poor market performance, financial losses, and unsatisfied customers. Many coffee producers also find it difficult to compete globally because they don’t fully grasp the link between bean characteristics and quality scores.

This project addresses these issues by analyzing the **Coffee Quality Dataset** to uncover the **key factors that influence coffee quality**. The goal is to give stakeholders clear, data-driven insights that help them improve production, maintain quality, and make smarter business decisions

|  |  |
| --- | --- |
| 4. | PROJECT WORK FLOW |

1. **Data cleaning** to handle missing, invalid, and inconsistent values.
2. **Deriving new metrics** for better insight, such as altitude ranges.
3. **Exploratory data analysis (EDA)** using univariate, bivariate, and multivariate techniques.
4. **Hypothesis testing** (T-test, ANOVA, Chi-square) to validate observed differences and relationships.
5. **Visualization** to communicate findings effectively.
6. **Find insights** – understand what affects coffee quality.
7. **Make conclusions** – share the important findings

|  |  |
| --- | --- |
| 5. | DATA UNDERSTANDING |

**Quality Measures :-**

Aroma, Flavour, Aftertaste, Acidity, Body, Balance, Uniformity, Cup Cleanliness, Sweetness, Moistur, Defects .

**Bean Metadata :-**

• Processing Method

• Colour

• Species (arabica / robusta)

**Farm Metadata**

• Owner

• Country of Origin

• Farm Name

• Lot Number

• Mill

• Company

• Altitude

• Region

|  |  |
| --- | --- |
| 6. | DATA CLEANING |

1. **Identifying and Handling Missing Values**

* Checked for nulls in all columns.
* Filled missing values in Number.of.Bags with the **median** to maintain data integrity.
* Left non-critical missing values for removal if imputation wasn’t meaningful.

1. **Standardizing Data Formats**

* Cleaned Bag.Weight by extracting numeric parts and converting to **float**.
* Converted Number.of.Bags to **Int64** for consistency.

1. **Removing Irrelevant or Redundant Columns**

* Dropped unnamed and certification-related columns that didn’t contribute to analysis.

1. **Outlier Detection and Removal**

* Applied **IQR method** on numeric columns to remove extreme outliers and improve result accuracy.

1. **Creating Derived Metrics**

* Added new columns like altitude\_range\_meters to capture valuable insights.

|  |  |
| --- | --- |
| 7. | OBTANING DERIVED METRICS |

1. **Altitude Range (altitude\_range\_meters)**
   * Derived by subtracting altitude\_low\_meters from altitude\_high\_meters.
   * Helps evaluate how farm altitude variations influence coffee quality.
2. **Cleaned Bag Weight**
   * Extracted numeric values from Bag.Weight and converted to float for consistent analysis.
   * Ensured weight data was accurate for comparisons and aggregations.
3. **Imputed Number of Bags**
   * Filled missing values in Number.of.Bags with the **median** and converted the column to Int64.
   * Preserved dataset completeness while minimizing bias.
4. **Dropped Unnecessary Columns**
   * Removed non-essential or redundant metadata columns (e.g., unnamed, certification details) to reduce noise.
5. **Outlier Handling**
   * Identified and removed extreme values in numeric columns using the **IQR method**, improving data reliability.
6. **Outcome:**

* These derived and cleaned metrics provided a solid foundation for descriptive, bivariate, and multivariate analysis, leading to more accurate conclusions.

|  |  |
| --- | --- |
| 8. | FILTERING DATA FOR ANALYSIS |

We removed rows with incomplete or incorrect data and focused on the main columns:

* **Species**, **Country of Origin**, **Total Cup Points**, and altitude details.

|  |  |
| --- | --- |
| 9. | STATISTICAL ANALYSIS |

1. **Descriptive Statistics**

* Calculated **mean, median, standard deviation, min, max, and quartiles** for all sensory attributes (Aroma, Flavor, Aftertaste, etc.).
* Example:
  + **Aroma:** Mean ≈ 7.3, Median ≈ 7.4
  + **Total Cup Points:** Mean ≈ 82, Median ≈ 82.5
* These metrics provided an overview of data distribution and central tendency.

1. **Dataset Overview**

* Used df.info() to inspect data types, non-null counts, and overall structure.
* Most key attributes were numeric; categorical data included Species, Processing Method, and Country of Origin.

1. **T-Test**

* Compared **Total Cup Points** between Arabica and Robusta.
* **p = 0.0621** → Fail to reject H₀; no significant difference between species.

1. **ANOVA**

* Tested if **Processing Methods** impacted Aroma.
* **p = 0.9713** → Fail to reject H₀; no significant differences.

5. **Chi-Square Test**

* Found significant relationships between **Species, Processing Methods, and Country of Origin**, indicating regional dependencies.

|  |  |
| --- | --- |
| 10. | UNIVARIATE ANALYSIS |

1. **Sensory Attributes**

* Key variables analyzed: Aroma, Flavor, Aftertaste, Acidity, Body, Balance, and Total Cup Points.
* Helped understand the central tendency and spread of each attribute.

1. **Distribution Visualizations**

* **Histograms** were used to observe the shape of distributions (e.g., normality, skewness).
* **Boxplots** identified outliers and spread within each attribute.

1. **Categorical Variables**

* Species, Processing Method, and Country of Origin were explored using **bar charts** and **pie charts**.
* Arabica dominated the dataset; certain countries contributed more samples.

1. **Insights**

* Most sensory attributes showed relatively high scores with few extreme outliers.
* Aroma and Flavor had the most significant influence on Total Cup Points.

|  |  |
| --- | --- |
| 11. | BIVARIATE ANALYSIS |

**1.Sensory Attributes vs. Total Cup Points**

* **Scatter plots** showed positive relationships between Aroma, Flavor, Acidity, and Total Cup Points.
* Higher Aroma and Flavor scores were linked to higher overall cup ratings.

**2.Species vs. Total Cup Points**

* **Boxplots** compared cup points between Arabica and Robusta.
* T-Test results: **no significant difference** (p = 0.0621), though Arabica had slightly higher median scores.

**3.Processing Method vs. Aroma**

* **Boxplots** examined Aroma scores across processing methods.
* ANOVA results: **no significant difference** (p = 0.9713).

**4.Categorical Comparisons**

* **Bar charts and pie charts** highlighted the distribution of species and processing methods across countries of origin.

**5.Insights**

* Coffee quality depends on a combination of sensory attributes, with Aroma and Flavor being the most influential.
* Species and processing method alone were not strong predictors of Total Cup Points

|  |  |
| --- | --- |
| 12. | MULTIVARIATE ANALYSIS |

1.**Correlation Analysis**

* A **heatmap** of sensory attributes revealed strong positive correlations among Aroma, Flavor, Aftertaste, and Balance.
* These factors collectively influence Total Cup Points.

2.**Pairwise Relationships**

* **Pair plots** demonstrated how multiple sensory scores interact and align with higher cup ratings.

3.**Combined Impact on Cup Points**

* Quality depends on the synergy of sensory attributes rather than any single factor alone.

4.**Insights**

* Farms that maintain high Aroma, Flavor, and Balance across multiple conditions tend to achieve consistently high Total Cup Points.

|  |  |
| --- | --- |
| 13. | OVERALL INSIGHTS FROM ANALYSIS |

1. **No Significant Quality Difference by Species:**

* Total Cup Points do not significantly differ between **Arabica** and **Robusta** (p-value > 0.05).
* Both species in this dataset show comparable quality when evaluated by SCA standards.

1. **Processing Method Has No Significant Impact on Aroma:**

* ANOVA results indicate no significant variation in aroma scores across **washed, natural, or honey-processed** coffees (p-value > 0.05).
* Other factors such as **altitude, soil quality, and farm practices** likely play a greater role.
* **Geographical & Processing Patterns:**
* **Arabica** is predominantly linked with washed/honey processing and is common in countries like **Ethiopia** and **Colombia**.
* **Robusta** is often naturally processed, with production concentrated in **Vietnam** and **Uganda**.

1. **Key Quality Drivers:**

* **Aroma, Flavor, and Balance** strongly correlate with higher total cup scores.
* **Moisture** content shows negligible influence on cup quality.

1. **Data Cleaning Impact:**

* Imputing missing values, cleaning Bag.Weight, and removing outliers improved dataset reliability and ensured valid statistical testing.

1. **Statistical Relationships:**

* Chi-square tests confirmed significant associations between **species, processing methods, and countries of origin**, even though these do not directly impact cup scores.

|  |  |
| --- | --- |
| 14. | CONCLUSION |

**Key Findings and Insights**

1. **Species Comparison (T-Test):**
   * The T-Test found **no statistically significant difference** in Total Cup Points between **Arabica** and **Robusta** (p-value > 0.05).
   * Both species performed similarly in terms of overall quality in this dataset.
2. **Processing Method Influence (ANOVA):**
   * **No significant difference** was found in Aroma scores across Processing Methods (p-value > 0.05).
   * Aroma quality likely depends on other factors such as genetics, farm practices, and environmental conditions.
3. **Chi-Square Relationships:**
   * **Species vs Processing Method:** Arabica tends to be associated with washed/honey methods, Robusta with natural processing, though quality outcomes are not significantly different.
   * **Species vs Country of Origin:** Arabica dominates in Ethiopia, Colombia; Robusta in Vietnam, Uganda.
   * **Processing Method vs Country of Origin:** Washed methods are common in Latin America; natural methods in Africa and Asia.
4. **Quality Attributes Correlation:**
   * **Aroma, Flavor, and Balance** remain key contributors to higher cup scores, even though species and processing method show no strong statistical differences.
   * **Moisture** content has minimal influence on cup quality.
5. **Data Cleaning Impact:**
   * Standardizing Bag.Weight, imputing missing values, and removing outliers improved dataset reliability for analysis.

**Recommendations**

1. **Focus on Quality Attributes:**  
   Invest in improving aroma, flavor, and balance through better farming practices rather than relying solely on species or processing method.
2. **Processing Method Strategy:**  
   Since processing method does not significantly affect aroma in this dataset, further studies should explore other quality factors.
3. **Country-Specific Strategies:**  
   Continue optimizing Arabica and Robusta cultivation in their respective dominant regions.
4. **Data Quality Management:**  
   Maintain cleaned and standardized datasets for accurate and actionable insights.
5. **Further Research:**  
   Examine additional variables (altitude, soil, climate) that may have a stronger effect on cup scores.

**Future Work**

* **Predictive Modeling:** Use advanced machine learning to forecast cup scores from farm and bean data.
* **Dataset Expansion:** Include more samples and processing techniques to improve statistical power.
* **Time-Series Tracking:** Study how coffee quality evolves across harvest seasons.
* **Consumer-Centric Analysis:** Map sensory scores to actual consumer preferences.
* **Digital Monitoring:** Develop real-time quality-tracking tools for farmers and suppliers.