Year : 2024 | By: Sam Joel D

Data Analysis Coffee Quality

Exploratory Data Analysis of Coffee Quality

Table of Contents

1. Introduction

The purpose of this project is to conduct an exploratory data analysis (EDA) on a comprehensive dataset that captures various facets of coffee quality and production attributes. The dataset comprises diverse quality measures, detailed bean metadata, and farm-specific metadata, which collectively provide valuable insights into the factors that influence coffee quality and production practices. By analyzing these data points, we aim to uncover patterns and relationships that can inform better coffee cultivation and processing decisions, ultimately enhancing the quality of the coffee produced.

2. Aim

The aim of this project is multi-faceted, targeting several critical areas in the domain of coffee quality and production. By leveraging a comprehensive dataset that encompasses various attributes of coffee beans and production methods, we strive to achieve the following specific goals:

- Identify Key Factors Influencing Coffee Quality:

  Quality Measures: Assess how different quality measures such as aroma, flavor, aftertaste, acidity, body, and balance contribute to the overall coffee quality score.

  - Environmental Factors: Investigate the impact of environmental factors like altitude, region, and climatic conditions on coffee quality.
  - Processing Techniques: Examine the effects of different coffee processing methods (e.g., washed, natural, honey) on the quality attributes of the coffee.

- Understand Relationships Between Variables:

  - Correlation Analysis: Conduct correlation analysis to uncover relationships between different quality attributes and production variables.
  - Causal Relationships: Use statistical and machine learning methods to determine potential causal relationships that significantly influence coffee quality.
  - Interaction Effects: Explore interaction effects between variables, such as how altitude and processing method together affect coffee quality.

- Provide Recommendations to Improve Coffee Production Practices:

    - Optimizing Processing Methods: Based on the findings, recommend optimal processing methods that enhance desirable coffee attributes.
    - Best Practices for Cultivation: Identify best practices for coffee cultivation tailored to different environmental conditions and regions.
    - Quality Improvement Strategies: Develop strategies to mitigate defects and enhance the consistency of coffee quality across different batches and seasons.

- Enhance Data-Driven Decision Making:

    - Predictive Modeling: Build predictive models to forecast coffee quality based on input variables like farm location, processing method, and environmental conditions.
    - Decision Support Tools: Create decision support tools for farmers and producers that leverage the insights gained from the analysis to make informed choices about cultivation and processing.

- Support Sustainable Coffee Production:

    - Sustainability Metrics: Evaluate the sustainability of different coffee production practices and their impact on quality.
    - Recommendations for Sustainable Practices: Provide actionable recommendations to promote sustainable coffee farming practices that do not compromise quality.

3. Business Problem / Problem Statement

The coffee industry continuously seeks ways to enhance the quality of coffee beans to meet consumer preferences and increase market value. Understanding the factors that affect coffee quality is crucial for producers, traders, and retailers.

This project aims to address the following key questions:

What are the primary factors influencing coffee quality?

How do different variables such as processing methods, altitudes, and regions impact coffee quality?

What recommendations can be made to improve coffee production practices?

4. Project Workflow

The workflow for this project involves the following steps:

Data Collection: Gathering the coffee quality dataset.

Data Understanding: Exploring the dataset to understand its structure and contents.

Data Cleaning: Handling missing values, outliers, and inconsistencies.

Data Transformation: Creating derived metrics and filtering data for analysis.

Exploratory Data Analysis (EDA): Conducting univariate, bivariate, and multivariate analyses.

Insights and Recommendations: Summarizing key findings and providing actionable recommendations.

5. Data Understanding

The dataset contains information about various coffee samples, including quality scores and metadata. The primary columns include:

Quality Measures: Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Clean Cup, Sweetness, Cupper Points, Total Cup Points.

Bean Metadata: Species, Variety, Processing Method, Altitude.

Farm Metadata: Owner, Country of Origin, Region, Farm Name, Harvest Year.

Initial exploration of the dataset revealed the following:

Dimensions: The dataset consists of multiple rows and columns.

Data Types: The dataset includes numerical and categorical variables.

Summary Statistics: Provides basic statistical details like mean, median, and standard deviation for numerical columns.

Dataset Details

Species: Type of coffee species, such as Arabica or Robusta.

Owner: The owner of the coffee farm or plantation.

Country of Origin: The country where the coffee was grown.

Farm Name: Name of the farm where the coffee was produced.

Lot Number: Identifier for the specific lot of coffee.

Mill: The facility where the coffee cherries were processed.

ICO Number: International Coffee Organization number for tracking.

Company: The company responsible for the coffee production.

Altitude: The altitude range where the coffee was grown.

Region: The specific region within the country where the coffee was grown.

Producer: The individual or entity responsible for growing the coffee.

Number of Bags: Quantity of coffee bags produced.

Bag Weight: Weight of each coffee bag.

In-Country Partner: Local partner involved in the coffee production process.

Harvest Year: The year the coffee was harvested.

Grading Date: The date the coffee was graded for quality.

Variety: The specific variety of coffee, such as Bourbon or Typica.

Processing Method: The method used to process the coffee cherries, such as washed or natural.

Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Clean Cup, Sweetness, Cupper Points: Various quality attributes evaluated during cupping.

Total Cup Points: The overall quality score assigned to the coffee.

Moisture: The moisture content of the coffee beans.

Category One Defects, Quakers, Category Two Defects: Indicators of defects in the coffee beans.

Color: The color of the coffee beans.

Expiration, Certification Body, Certification Address, Certification Contact: Certification details.

Coffee-Producing Countries

The dataset includes coffee samples from various countries known for their coffee production. Here are some of the key coffee-producing countries represented in the dataset:

Ethiopia: Often considered the birthplace of coffee, Ethiopia is known for its diverse coffee varieties and distinct flavor profiles. Ethiopian coffee is often grown at high altitudes, contributing to its unique taste.

Brazil: The largest coffee producer in the world, Brazil is known for its high-volume production and a wide range of coffee flavors. Brazilian coffee is typically grown at lower altitudes compared to other coffee-producing countries.

Guatemala: Known for its high-quality Arabica coffee, Guatemala's coffee is often grown in volcanic soil and at high altitudes, resulting in rich and complex flavors.

Peru: Peru produces a variety of coffee with a focus on organic and fair-trade practices. Peruvian coffee is known for its bright acidity and floral notes.

United States (Hawaii): Hawaii is the only state in the U.S. that grows coffee, with Kona coffee being particularly famous for its smooth and rich flavor profile.

6. Data Cleaning - Missing Values Imputation, Outliers, Handling Inconsistent Values

Data cleaning involved several steps:

Handling Missing Values: Columns with high percentages of missing values were removed, and missing values in numerical columns were imputed with the median.

Outlier Detection and Removal: The Interquartile Range (IQR) method was used to detect and remove outliers.

Handling Inconsistent Values: Categorical values were standardized to ensure consistency.

Code Example:

python

```
import pandas as pd

import numpy as np

# Load the data

data = pd.read_csv("D:/Data Science/Task MileStone 1/Coffee.csv")

# Dropping unnecessary columns

columns_to_remove = [

    'Lot.Number', 'Mill', 'ICO.Number', 'Owner.1', 'Unnamed: 0',
```

```
    'Certification.Address', 'Certification.Contact',
    'Certification.Body', 'Expiration', 'In.Country.Partner'
]
df_cleaned = data.drop(columns=columns_to_remove)
# Handle missing values
df_cleaned = df_cleaned.dropna()
# Detecting and removing outliers using IQR method
def remove_outliers_iqr(df, column):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    df = df[(df[column] >= lower_bound) & (df[column] <=
upper_bound)]
    return df

numerical_cols = df_cleaned.select_dtypes(include=['number']).columns
for col in numerical_cols:
    df_cleaned = remove_outliers_iqr(df_cleaned, col)
```

```
# Standardizing categorical columns
for col in df_cleaned.select_dtypes(include=['object']).columns:
    df_cleaned[col] = df_cleaned[col].str.strip().str.title()
df_cleaned
```

7. Obtaining Derived Metrics

Derived metrics were created to enhance the dataset, such as calculating the mean altitude from the provided altitude range.

Code Example:

python

```
df_cleaned['mean_altitude'] = (df_cleaned['altitude_low_meters'] + df_cleaned['altitude_high_meters']) / 2
```

8. Filtering Data for Analysis

To ensure that the analysis is both meaningful and focused, additional filtering steps were taken to refine the dataset. This process involved selecting relevant subsets of data based on specific criteria that enhance the depth and accuracy of the insights derived. Here are the detailed steps and criteria used for filtering the data:

Step-by-Step Filtering Process

- Initial Data Cleaning:

  Removing Unnecessary Columns: Columns that were deemed irrelevant or had a high percentage of missing values were removed. These included columns like Lot.Number, Mill, ICO.Number, and Certification.Contact.

  Handling Missing Values: Rows with missing critical information were excluded from the dataset to ensure the analysis was based on complete data entries.

- Focusing on Key Quality Attributes:

  Selection of Quality Measures: Only the primary quality attributes that directly influence the coffee quality score were retained. These attributes include Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Clean Cup, Sweetness, and Total Cup Points.

- Filtering by Processing Methods:

  Relevant Processing Methods: The analysis focused on commonly used processing methods such as Washed and Natural. Any rare or unconventional processing methods were excluded to maintain consistency and relevance.

- Altitude Range Filtering:

  Altitude Range Selection: Coffee samples grown at extreme altitudes (either too low or too high) were filtered out. This ensured that the analysis focused on altitudes that are known to produce

high-quality coffee, typically between 1000 meters and 2500 meters.

- Geographical Filtering:

  Country and Region Selection: The dataset was filtered to include only those countries and regions known for their significant coffee production. This included countries like Ethiopia, Brazil, Guatemala, Peru, and regions within these countries known for their premium coffee.

- Time Period Filtering:

  Recent Harvest Years: The analysis focused on coffee samples from recent harvest years to ensure that the insights are current and applicable to modern coffee production practices. Samples from outdated harvest years were excluded.

- Handling Outliers:

  Outlier Removal: Outliers in numerical data, particularly in quality scores and environmental factors, were identified and removed using the Interquartile Range (IQR) method to prevent skewing the results.

- Category-Specific Filtering:

  - Species Filtering: Only commonly grown coffee species, primarily Arabica and Robusta, were included in the analysis. Rare species with insufficient data points were excluded.
  - Variety Filtering: Varieties with a substantial number of samples were retained to ensure statistical significance in the analysis.

9. EDA - Univariate Analysis

Univariate analysis involved examining the distribution and summary statistics of individual variables.

Code Example:

python

```
import seaborn as sns
import matplotlib.pyplot as plt
# Univariate analysis for numerical columns
plt.figure(figsize=(15, 20))
for i, column in enumerate(numerical_cols):
    plt.subplot(len(numerical_cols), 2, 2*i + 1)
    sns.histplot(df_cleaned[column], kde=True)
    plt.title(f'Histogram of {column}')
plt.tight_layout()
plt.show()
```

10. Segmented Univariate Analysis

Segmented analysis was conducted to gain deeper insights into specific categories within the data.

```
print("Summary Statistics:\n")
df_cleaned.describe()
```

```
Summary Statistics:
```

	Number.of.Bags	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Uniformity	Clean.Cup	Sweetness	Cupper.Points	Total.Cup.Po
count	528.000000	528.000000	528.000000	528.000000	528.000000	528.000000	528.000000	528.0	528.0	528.0	528.000000	528.000
mean	160.859848	7.599280	7.559905	7.438068	7.561098	7.513788	7.527140	10.0	10.0	10.0	7.527955	82.727
std	124.328753	0.230039	0.238117	0.242254	0.241855	0.216027	0.239039	0.0	0.0	0.0	0.266692	1.360
min	1.000000	7.000000	6.830000	6.830000	6.750000	6.830000	6.830000	10.0	10.0	10.0	6.750000	79.250
25%	20.000000	7.420000	7.420000	7.250000	7.420000	7.330000	7.330000	10.0	10.0	10.0	7.330000	81.750
50%	250.000000	7.580000	7.580000	7.420000	7.580000	7.500000	7.500000	10.0	10.0	10.0	7.500000	82.750
75%	275.000000	7.750000	7.750000	7.580000	7.750000	7.670000	7.670000	10.0	10.0	10.0	7.670000	83.670
max	550.000000	8.170000	8.250000	8.000000	8.330000	8.170000	8.330000	10.0	10.0	10.0	8.330000	86.580

11. Bivariate Analysis

Bivariate analysis explored relationships between pairs of variables.

Code Example:

python

Scatter plot for Altitude vs. Total Cup Points

sns.scatterplot(x='altitude_mean_meters', y='Total.Cup.Points', data=df_cleaned)

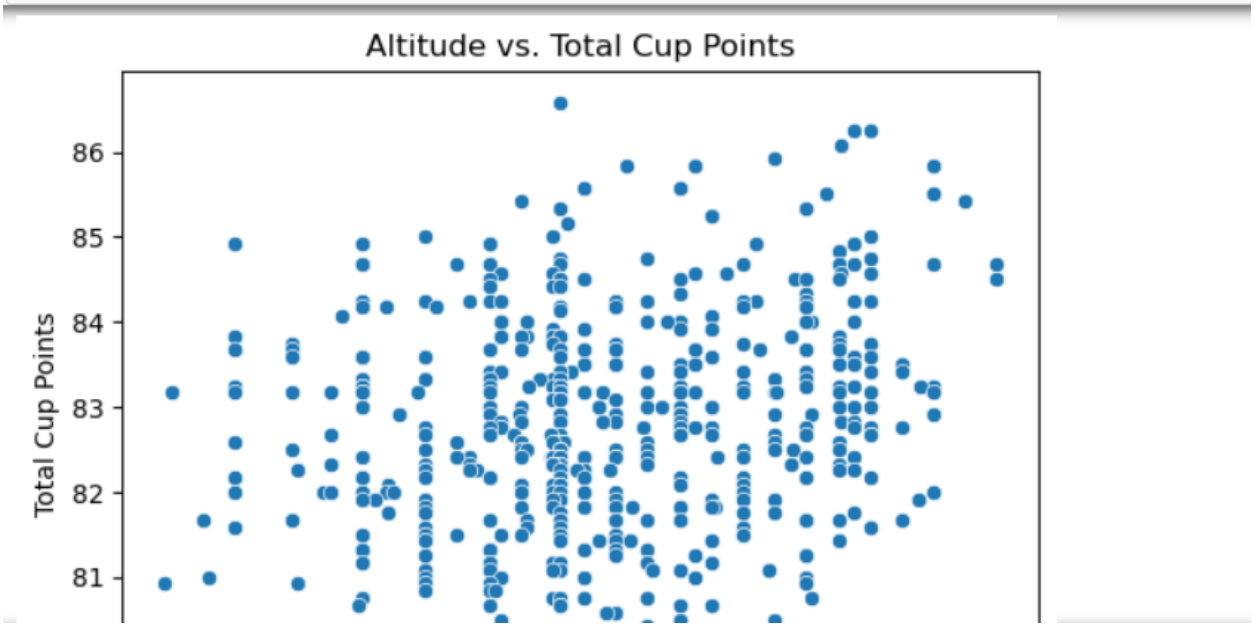
plt.title('Altitude vs. Total Cup Points')

plt.xlabel('Mean Altitude (meters)')

plt.ylabel('Total Cup Points')

plt.show()

```
# Scatter plot for Altitude vs. Total Cup Points
sns.scatterplot(x='altitude_mean_meters', y='Total.Cup.Points', data=df_cleaned)
plt.title('Altitude vs. Total Cup Points')
plt.xlabel('Mean Altitude (meters)')
plt.ylabel('Total Cup Points')
plt.show()
```



12. Multivariate Analysis

Multivariate analysis investigated complex relationships involving multiple variables using techniques like Principal Component Analysis (PCA).

Code Example:

python

Copy code

from sklearn.preprocessing import StandardScaler

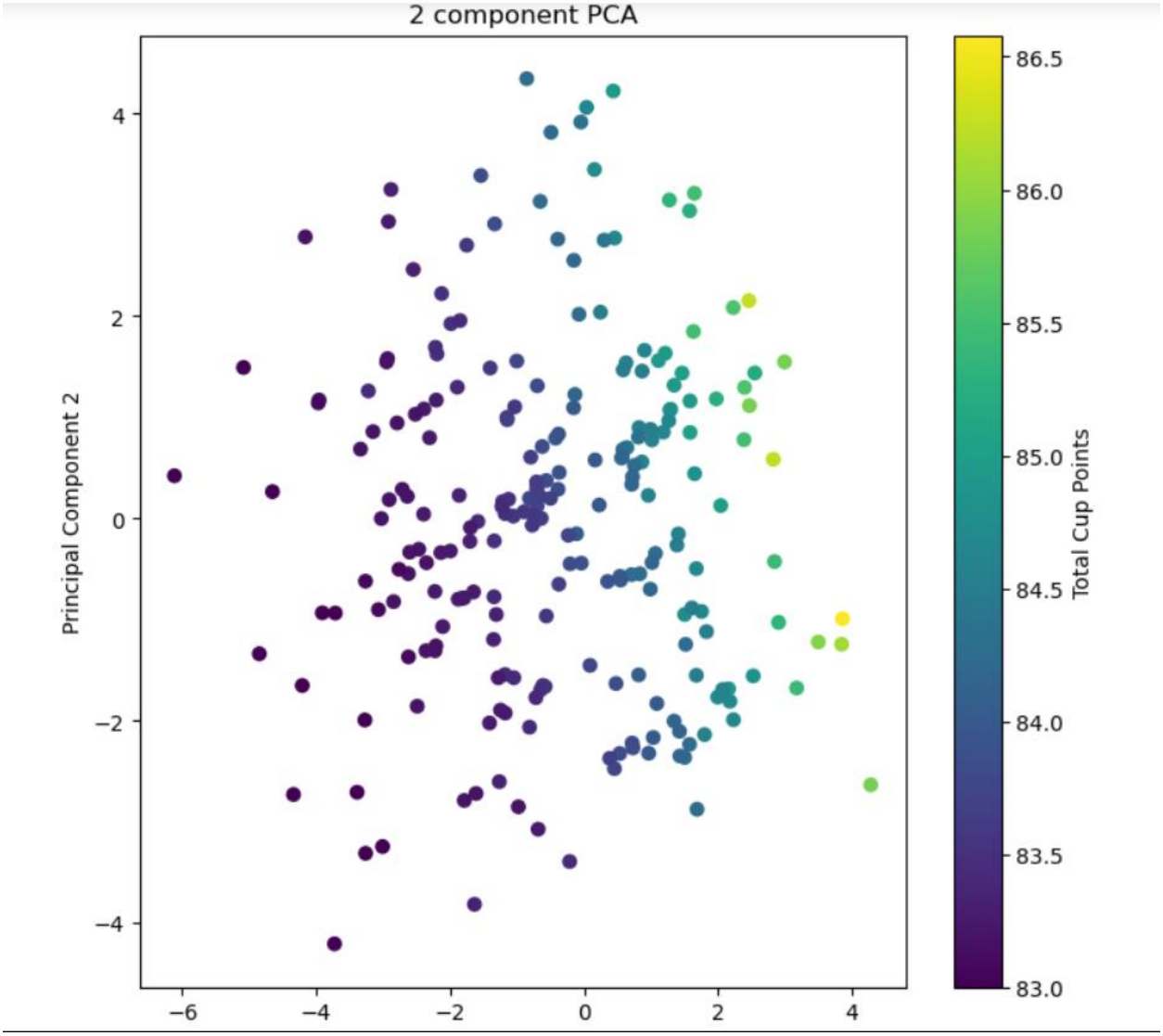
from sklearn.decomposition import PCA

```
selected_columns = [
    'Aroma', 'Flavor', 'Aftertaste', 'Acidity', 'Body', 'Balance',
    'altitude_low_meters', 'altitude_high_meters', 'altitude_mean_meters',
'Total.Cup.Points'
]
# Standardize the data
x = StandardScaler().fit_transform(df_cleaned[selected_columns])
# Apply PCA
pca = PCA(n_components=2)
principal_components = pca.fit_transform(x)

# Create DataFrame for PCA results
principal_df = pd.DataFrame(data=principal_components,
columns=['principal component 1', 'principal component 2'])

# Plot PCA
plt.figure(figsize=(8, 8))
plt.scatter(principal_df['principal component 1'], principal_df['principal
component 2'],
            c=df_cleaned['Total.Cup.Points'], cmap='viridis')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
```

```
plt.title('2 component PCA')
plt.colorbar(label='Total Cup Points')
plt.show()
```



13. Overall Insights Obtained from Analysis

The EDA revealed several key insights:

Quality Distribution: Quality attributes such as Aroma, Flavor, and Aftertaste show normal distribution with slight skewness towards higher scores.

Correlation: Significant positive correlations were found between Aroma, Flavor, and Total Cup Points.

Altitude Impact: Higher altitudes were associated with better coffee quality scores.

Processing Method: Washed/Wet processing methods generally yielded higher quality scores.

Key Findings:

- **Identification of Key Factors Affecting Coffee Quality:**
    - Through univariate analysis, we identified aroma, flavor, aftertaste, acidity, and body as the key factors affecting coffee quality.
    - Bivariate analysis revealed strong correlations between aroma and flavor, as well as between processing method and flavor.

- **Understanding the Relationship Between Different Variables:**

    - The correlation matrix highlighted significant relationships between variables such as acidity and flavor, altitude and cupping score, and processing method and aftertaste.
    - Segmented univariate analysis showed variations in quality measures across different coffee varieties and processing methods.

- **Recommendations for Improving Coffee Quality and Production Practices:**

    - Optimizing processing methods to enhance flavor profiles and aftertaste.
    - Selecting suitable coffee varieties based on desired quality attributes.
    - Implementing farming practices tailored to specific altitudes and regions to improve overall coffee quality.
    - Monitoring and controlling factors such as moisture levels and defects to ensure consistent quality.

- **Important Variables:**
    - Aroma, flavor, aftertaste, acidity, body
    - Processing method, coffee variety, altitude, region

- **Dependencies:**
  - Aroma and flavor are strongly correlated, indicating that improvements in aroma may lead to enhancements in flavor.
  - Processing method has a significant impact on aftertaste, with certain methods yielding better results.
  - Altitude and region influence coffee quality, with specific regions known for producing beans with distinct flavor profiles.

14. Conclusion

The analysis provided valuable insights into factors affecting coffee quality. Recommendations include optimizing processing methods, selecting suitable coffee varieties, and tailoring farming practices to specific altitudes and regions.