Coffee Quality Dataset Analysis
Using Python

1.Context

Coffee is one of the most widely consumed beverages in the world, and its quality plays a key role in global trade, pricing, and customer satisfaction. Factors such as origin, species, processing methods, and flavour attributes all influence the final quality of a coffee sample. This project uses real-world data to explore how these factors affect coffee quality.

2.Objective

The main objective of this project is to:

- Understand what influences high coffee quality scores
- Identify top-performing countries, farms, and species
- Analyse how different processing methods and attributes relate to cup quality
- Present clear insights through visual and statistical analysis

3.Scope

- The project focuses on exploratory data analysis (EDA) using Python
- It covers data cleaning, visualization, and correlation analysis
- Key areas include production trends, flavour profiles, defects, and grading
- This project does not include machine learning or predictive modelling

4.Audience

This project is suitable for:

- Data science learners and students building their analytics skills
- Coffee producers and quality teams seeking insights to improve standards
- Researchers and professionals in agriculture or food industries
- Hiring managers and portfolio reviewers evaluating data projects

5.Techniques Used in the Project

1. **Data Cleaning:**
   Removed missing or incorrect values and filtered out irrelevant records to ensure clean, usable data for analysis.

2. **Exploratory Data Analysis (EDA):**
   Explored trends, distributions, and patterns using Pandas, Matplotlib, and Seaborn to understand the dataset at a high level.
3. **Grouping and Aggregation:**
   Used `groupby()` to compute averages, totals, and comparisons across countries, species, farms, and processing methods.
4. **Correlation Analysis:**
   Generated correlation matrices to study relationships between numerical features such as aroma, flavor, defects, and cup points.
5. **Data Visualization:**
   Created bar charts, scatter plots, and heatmaps to present insights visually for better understanding and communication.
6. **Hypothesis Testing:**
   Applied statistical tests (e.g., t-test, ANOVA) to validate assumptions such as:
   - Whether Arabica scores significantly higher than Robusta
   - Whether processing methods affect average cup scores
   - Whether defects significantly reduce coffee quality

6. **Dataset Overview – Coffee Quality Dataset**

General Information

- The dataset originally contained 13,340 rows and 44 columns.
- Each row represents a unique coffee sample evaluated by professional graders.
- The data was collected from various coffee-producing countries, farms, and exporters.

Key Feature Categories

1. **Farm Details**
   Includes Farm.Name, Owner, Mill, Region, and Country.of.Origin.
   This information describes where and by whom the coffee was produced.
2. **Coffee Variety and Processing**
   Contains Variety and Processing.Method.
   Details the type of coffee bean (e.g., Bourbon, Typica) and the method used for processing (e.g., Washed, Natural).

3. **Altitude Information**

   Includes altitude_low_meters, altitude_high_meters, and altitude_mean_meters.

   Represents the elevation at which the coffee was grown, which often affects quality.

4. **Grading Information**

   Contains Grading.Date, Harvest.Year, and Expiration.

   Provides a timeline for harvesting, grading, and expiration of the coffee.

5. **Defect Scores**

   Includes Category.One.Defects, Category.Two.Defects, and Quakers.

   Measures imperfections in the beans that can lower the coffee's quality.

6. **Cup Quality Attributes**

   Includes Aroma, Flavour, Aftertaste, Acidity, Body, Balance, Uniformity, Sweetness, Clean.Cup, Overall, and Total.Cup.Points.

   These scores reflect the sensory evaluation of the coffee by professionals, typically rated on a scale from 0 to 10.

7. **Certification Info**

   Contains Certification.Body, Certification.Address, and Certification.Contact.

   Identifies the organization responsible for certifying the coffee sample.

8. **Export Details**

   Includes Number.of.Bags, Bag.Weight, and Unit.of.measurements.

   Describes the quantity of coffee exported and the packaging used.

7. Data Cleaning Summary

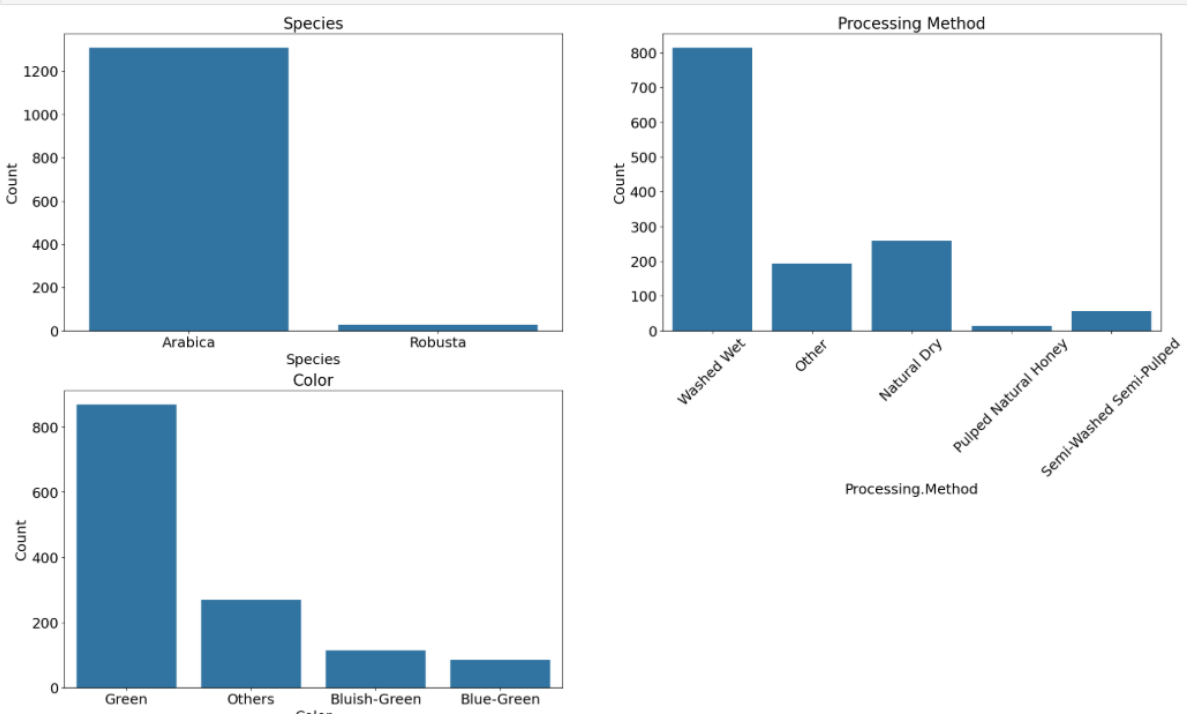
- Standardized spelling and fixed inconsistent values in categorical columns.
- Dropped irrelevant or duplicate columns such as:
- Certification Body, Address, Contact, Lot Number, ICO Number, Altitude Range, Owner.1
- Removed rows with over 50% missing values and all numerical fields as zero.
- Handled missing values:
- Replaced nulls in Color, Variety, and Processing.Method with "Other".
- Filled missing Owner values with the most common (mode) value per country.
- Filled Quakers with 0 if one defect category was already 0.
- Filled remaining categorical nulls with "Unknown".

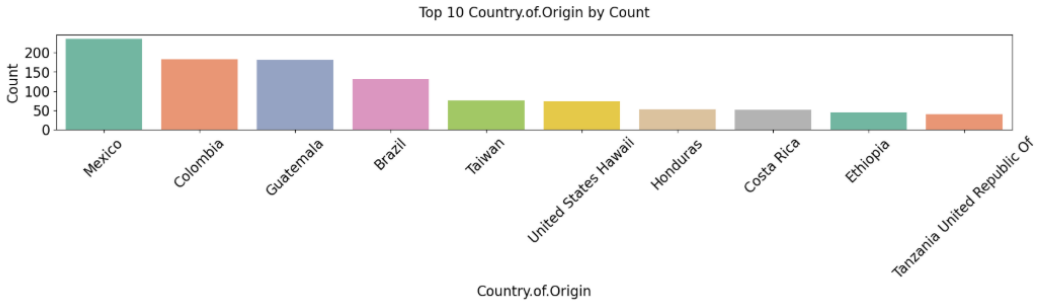
- Reformatted Grading.Date and Expiration using regular expressions and converted them to datetime.
- Estimated missing Harvest.Year from the difference between grading and expiration dates.
- Visualized distributions using boxplots and histograms.
- Replaced outliers with median or other statistical methods where needed.
- Dropped In.Country.Partner due to redundancy with Certification.Body.
- Dropped altitude_high_meters and altitude_low_meters to avoid duplication with altitude_mean_meters.

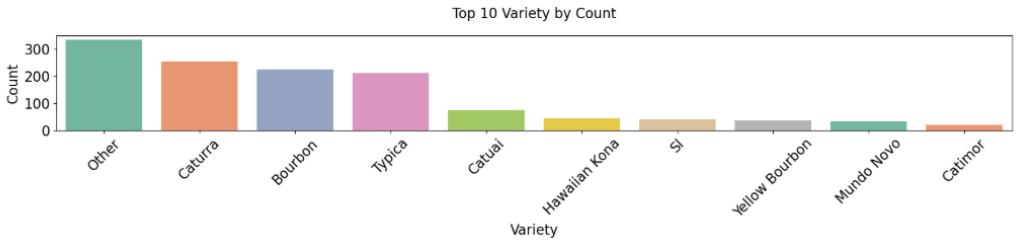
8. Exploratory Data Analysis (EDA)

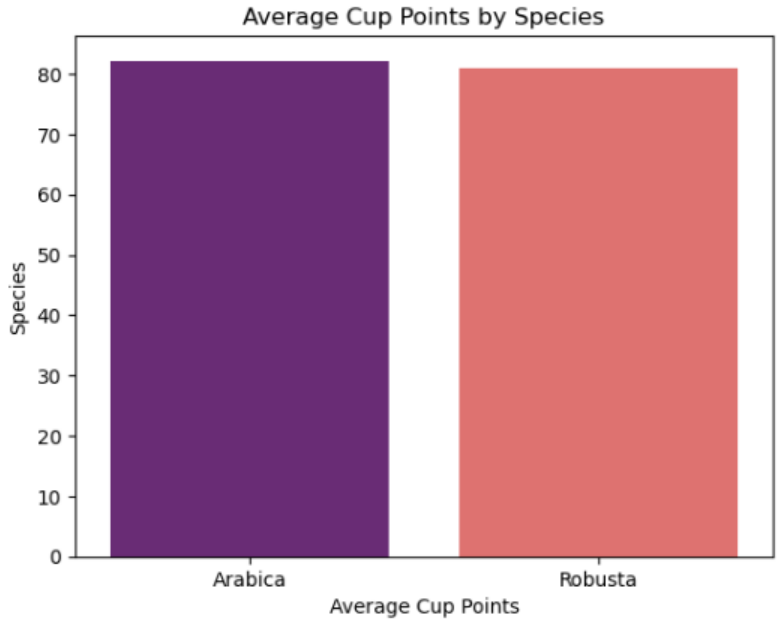
1. Univariate Analysis

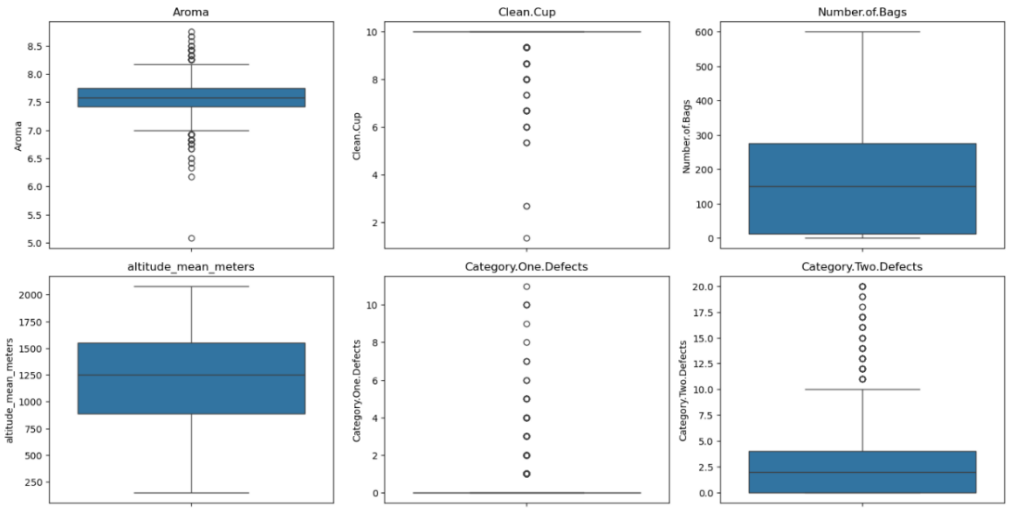
- Examined individual feature distributions using histograms and boxplots.
- Observed that most cup quality attributes (like Aroma, Flavor, Acidity) followed a normal distribution, centered around high scores.
- Found that Arabica was the dominant species in the dataset, making up the vast majority of samples.




Top 10 Country.of.Origin by Count
Count
200
150
100
50
0
Mexico
Colombia
Guatemala
Brazil
Taiwan
United States Hawaii
Honduras
Costa Rica
Ethiopia
Tanzania United Republic Of
Country.of.Origin

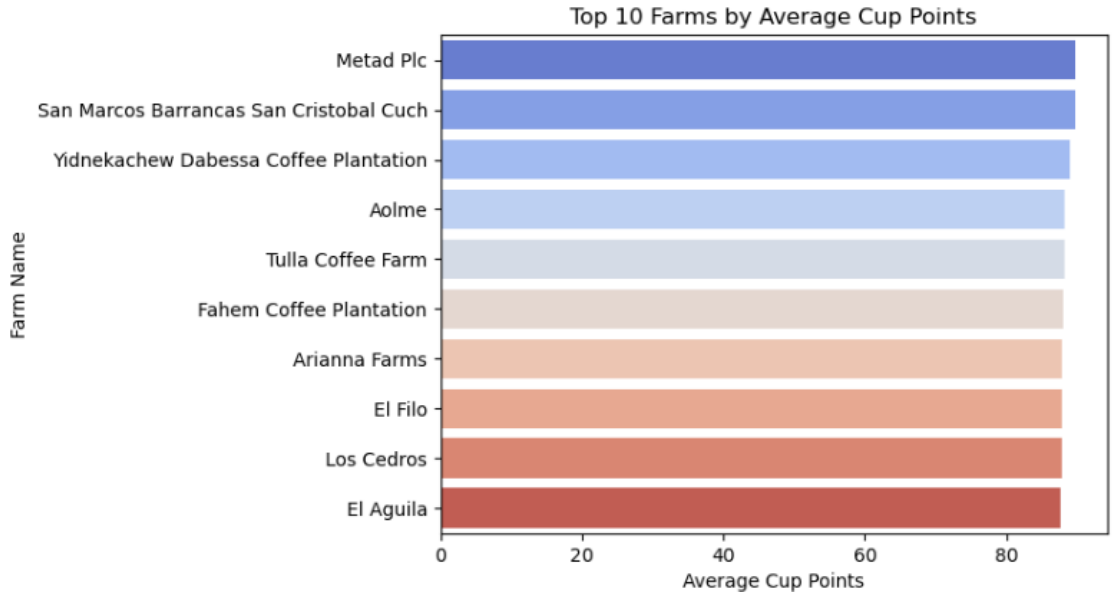

Top 10 Variety by Count
Count
300
200
100
0
Other
Caturra
Bourbon
Typica
Catuai
Hawaiian Kona
SI
Yellow Bourbon
Mundo Novo
Catimor
Variety

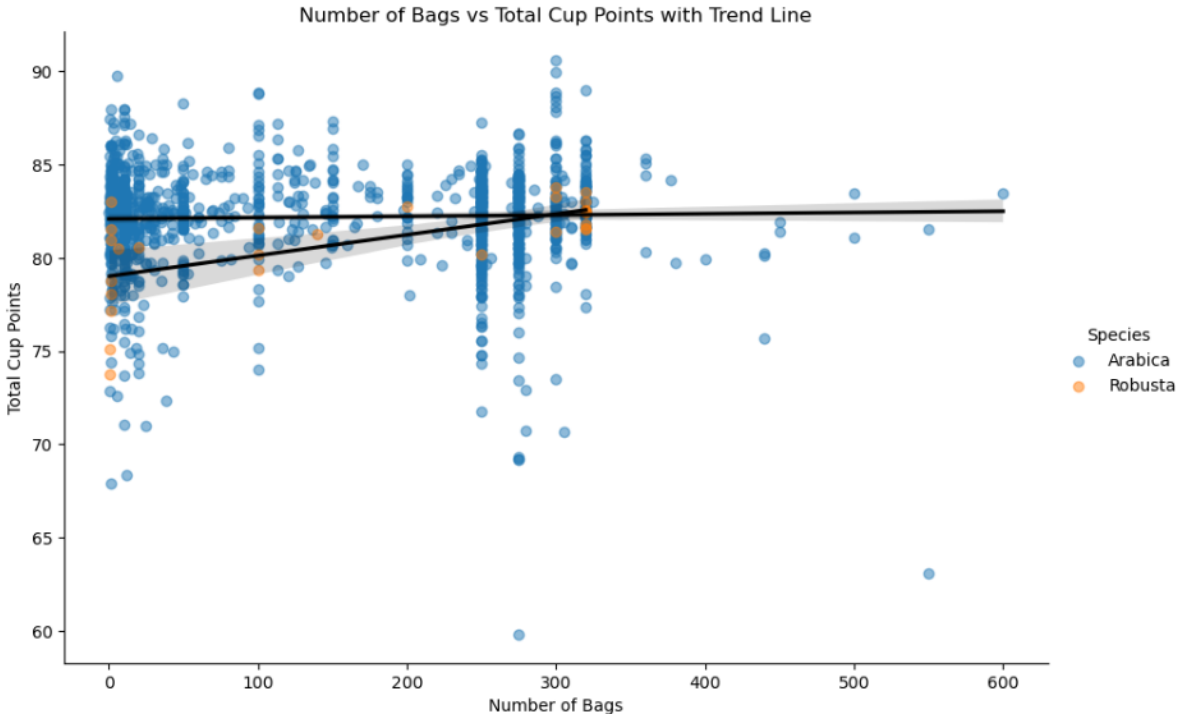

Average Cup Points by Species
Species
80
70
60
50
40
30
20
10
0
Arabica
Robusta
Average Cup Points

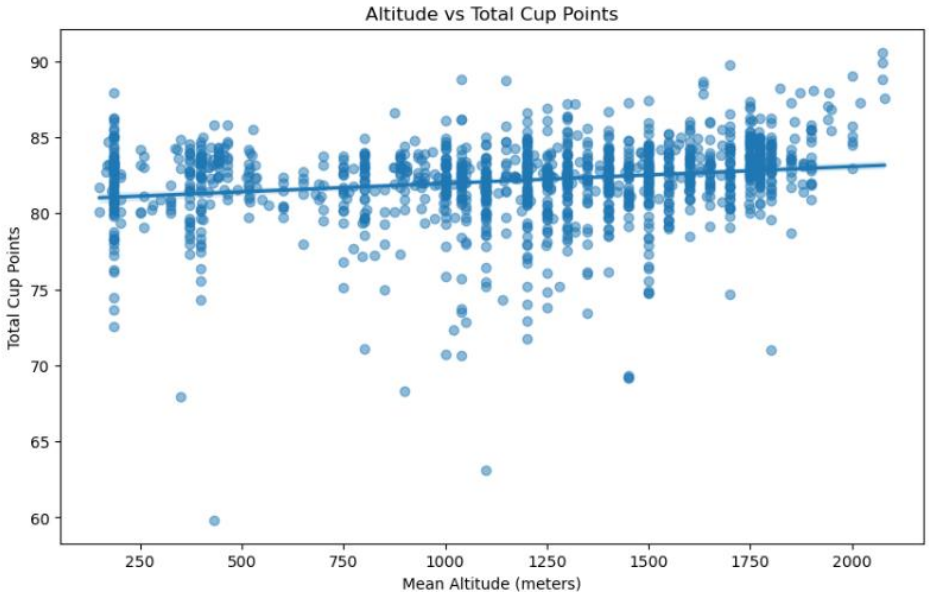

Aroma
Clean.Cup
Number.of.Bags
altitude_mean_meters
Category.One.Defects
Category.Two.Defects

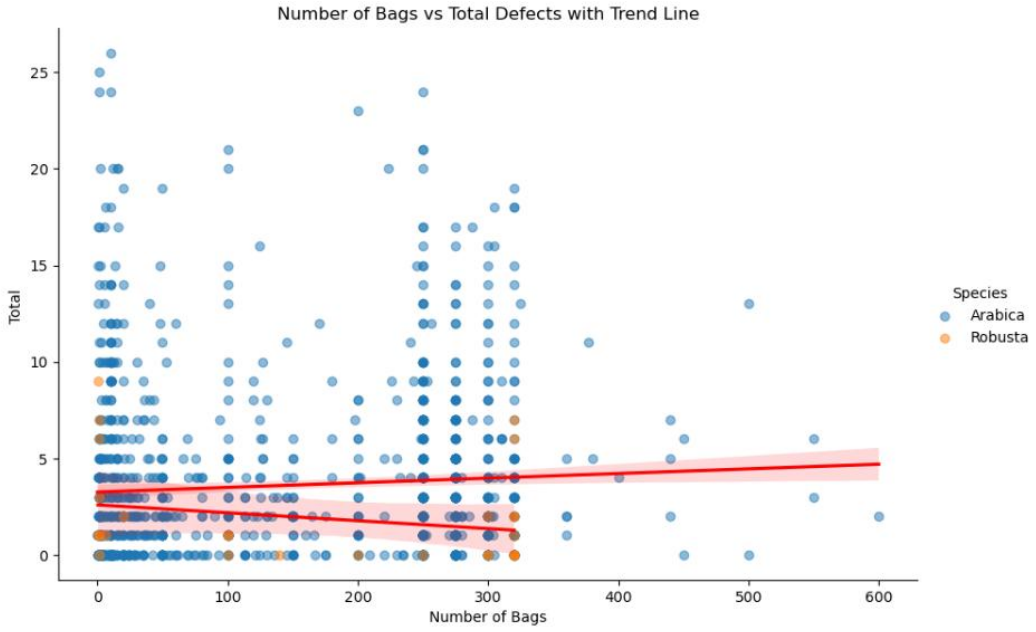
2. Bivariate Analysis

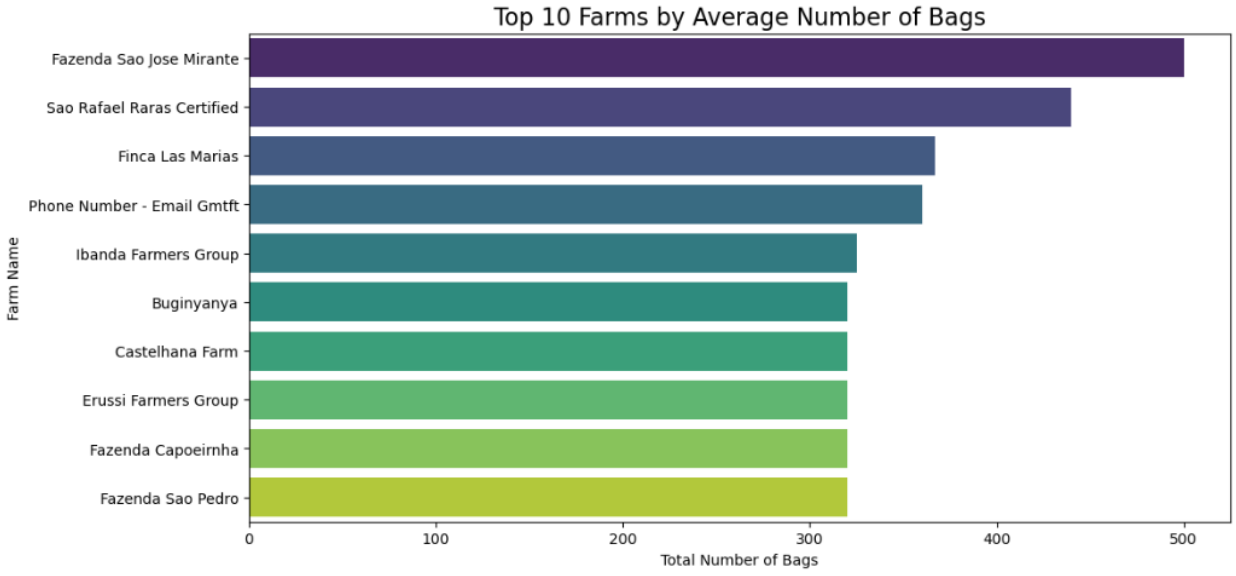
- Country vs Cup Points: Identified top-performing countries such as Papua New Guinea, Ethiopia, and Japan based on average cup scores.
- Species vs Quality: Confirmed that Arabica consistently scores higher than Robusta across attributes.
- Farm vs Bags: Ranked farms based on average number of bags produced.
- Processing Method vs Attributes: Compared average aroma, flavor, and other scores across methods (Washed, Natural, etc.).

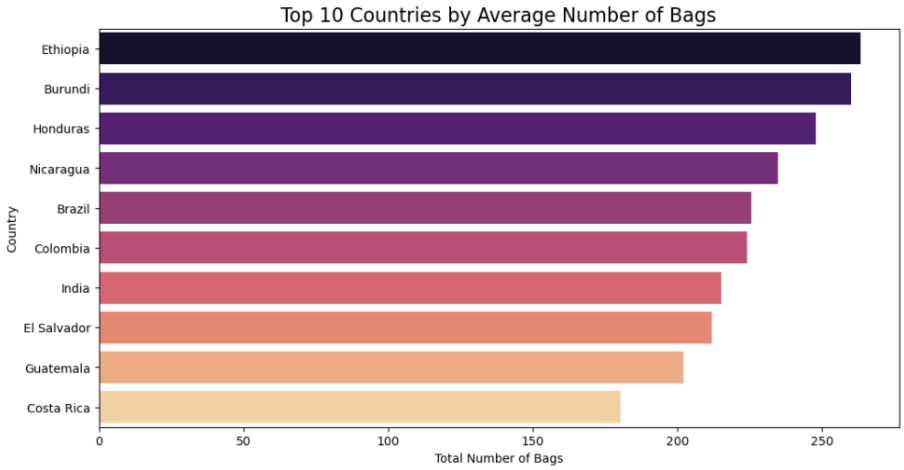


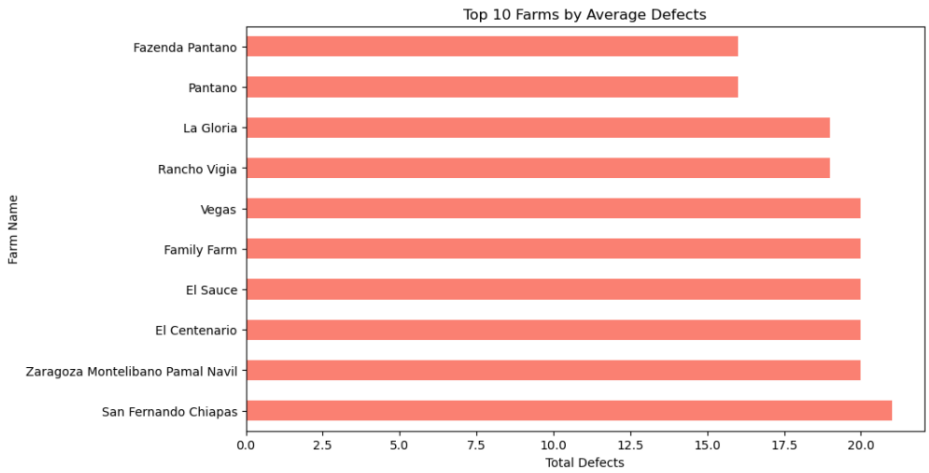


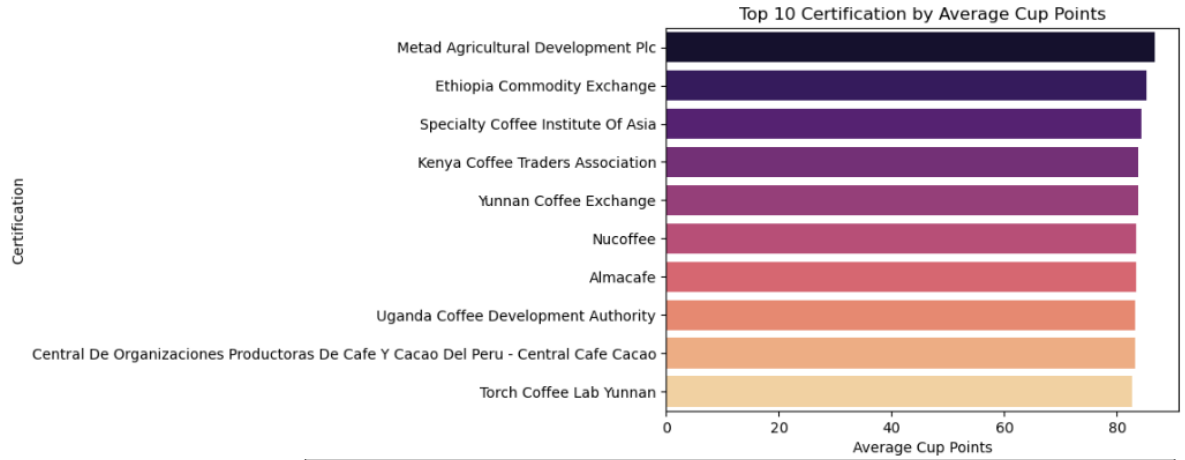

Altitude vs Total Cup Points
Total Cup Points
Mean Altitude (meters)
90
85
80
75
70
65
60
250
500
750
1000
1250
1500
1750
2000

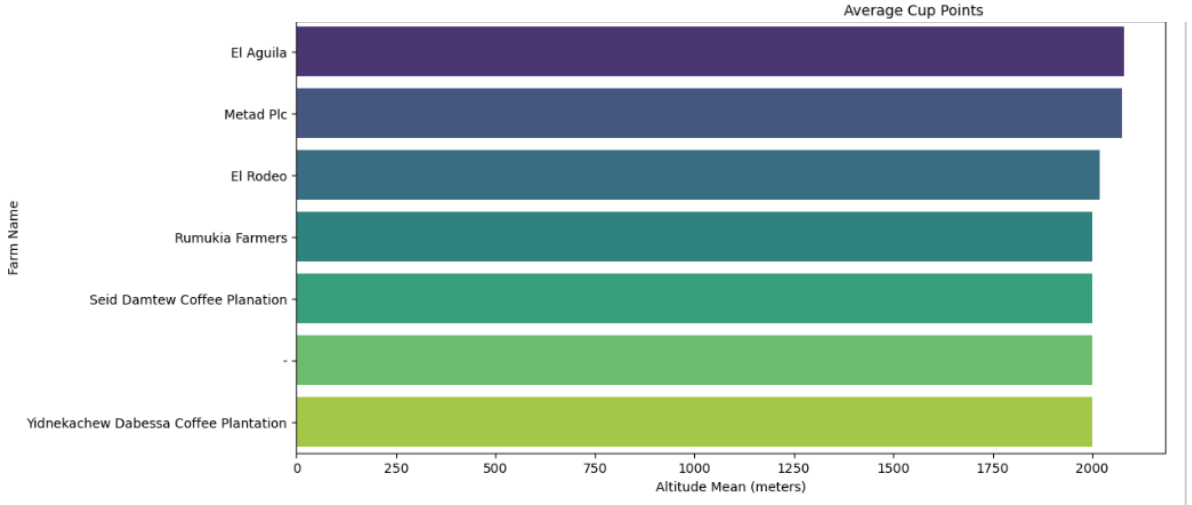

Number of Bags vs Total Defects with Trend Line
Total
Number of Bags
Species
Arabica
Robusta
25
20
15
10
5
0
0
100
200
300
400
500
600

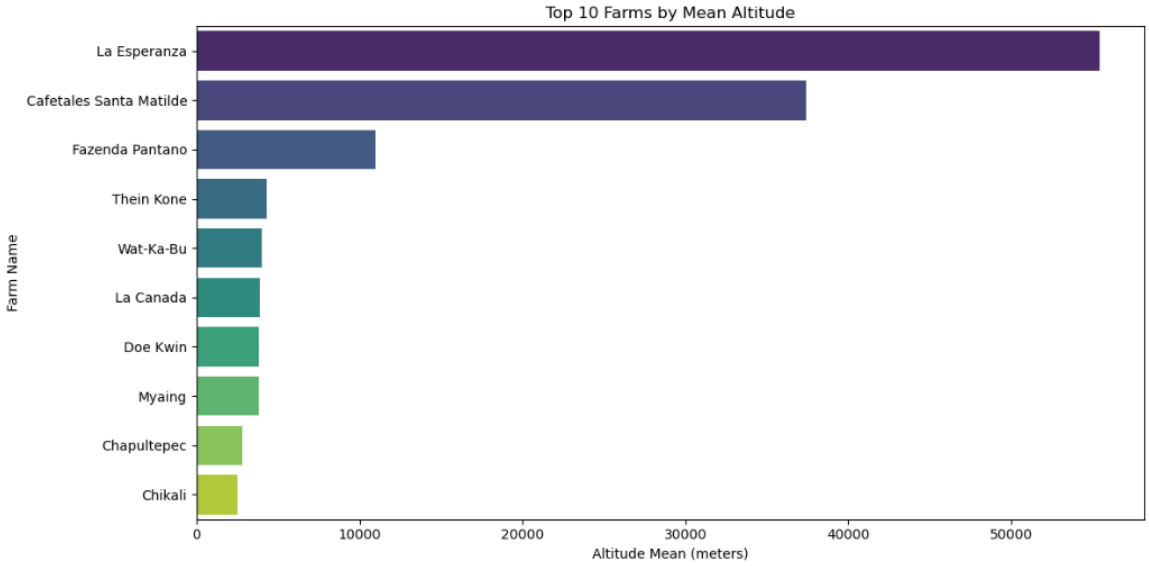

Top 10 Farms by Average Number of Bags
Farm Name
Total Number of Bags
Fazenda Sao Jose Mirante
Sao Rafael Raras Certified
Finca Las Marias
Phone Number - Email Gmtft
Ibanda Farmers Group
Buginyanya
Castelhana Farm
Erussi Farmers Group
Fazenda Capoeirnha
Fazenda Sao Pedro
0
100
200
300
400
500


Top 10 Countries by Average Number of Bags
Ethiopia
Burundi
Honduras
Nicaragua
Brazil
Colombia
India
El Salvador
Guatemala
Costa Rica
Country
0
50
100
150
200
250
Total Number of Bags


Top 10 Farms by Average Defects
Fazenda Pantano
Pantano
La Gloria
Rancho Vigia
Vegas
Family Farm
El Sauce
El Centenario
Zaragoza Montelibano Pamal Navil
San Fernando Chiapas
Farm Name
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Total Defects

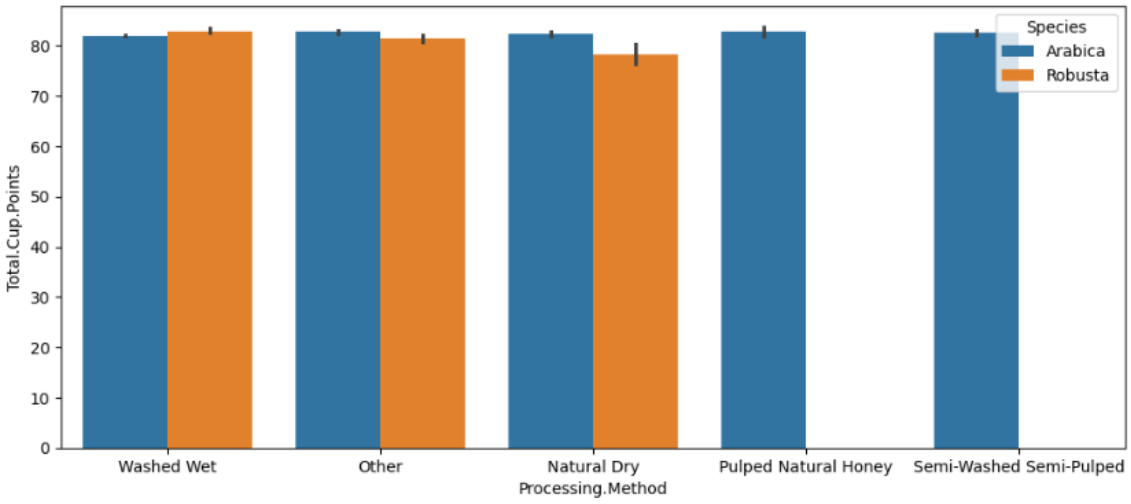

Top 10 Certification by Average Cup Points
Metad Agricultural Development Plc
Ethiopia Commodity Exchange
Specialty Coffee Institute Of Asia
Kenya Coffee Traders Association
Yunnan Coffee Exchange
Nucoffee
Almacafe
Uganda Coffee Development Authority
Central De Organizaciones Productoras De Cafe Y Cacao Del Peru - Central Cafe Cacao
Torch Coffee Lab Yunnan
Certification
0
20
40
60
80
Average Cup Points

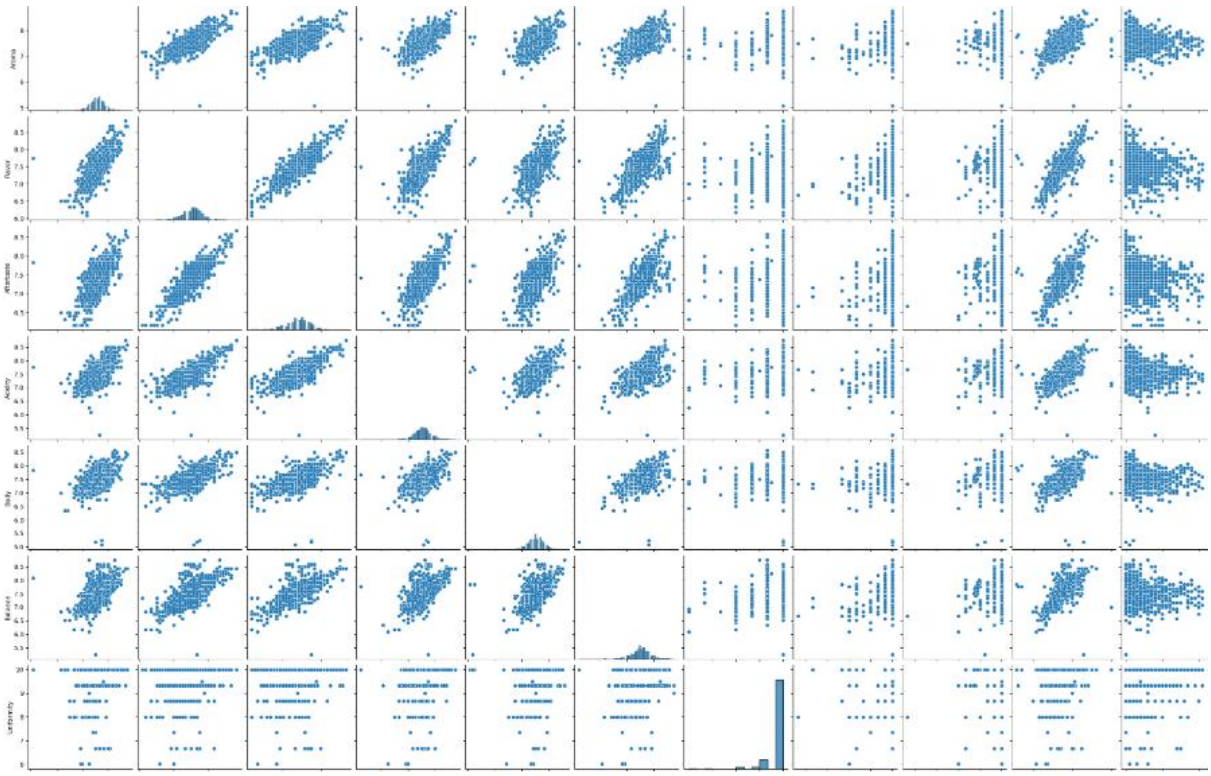

El Aguila
Metad Plc
El Rodeo
Rumukia Farmers
Seid Damtew Coffee Planation
-
Yidnekachew Dabessa Coffee Plantation
Farm Name
0
250
500
750
1000
1250
1500
1750
2000
Altitude Mean (meters)

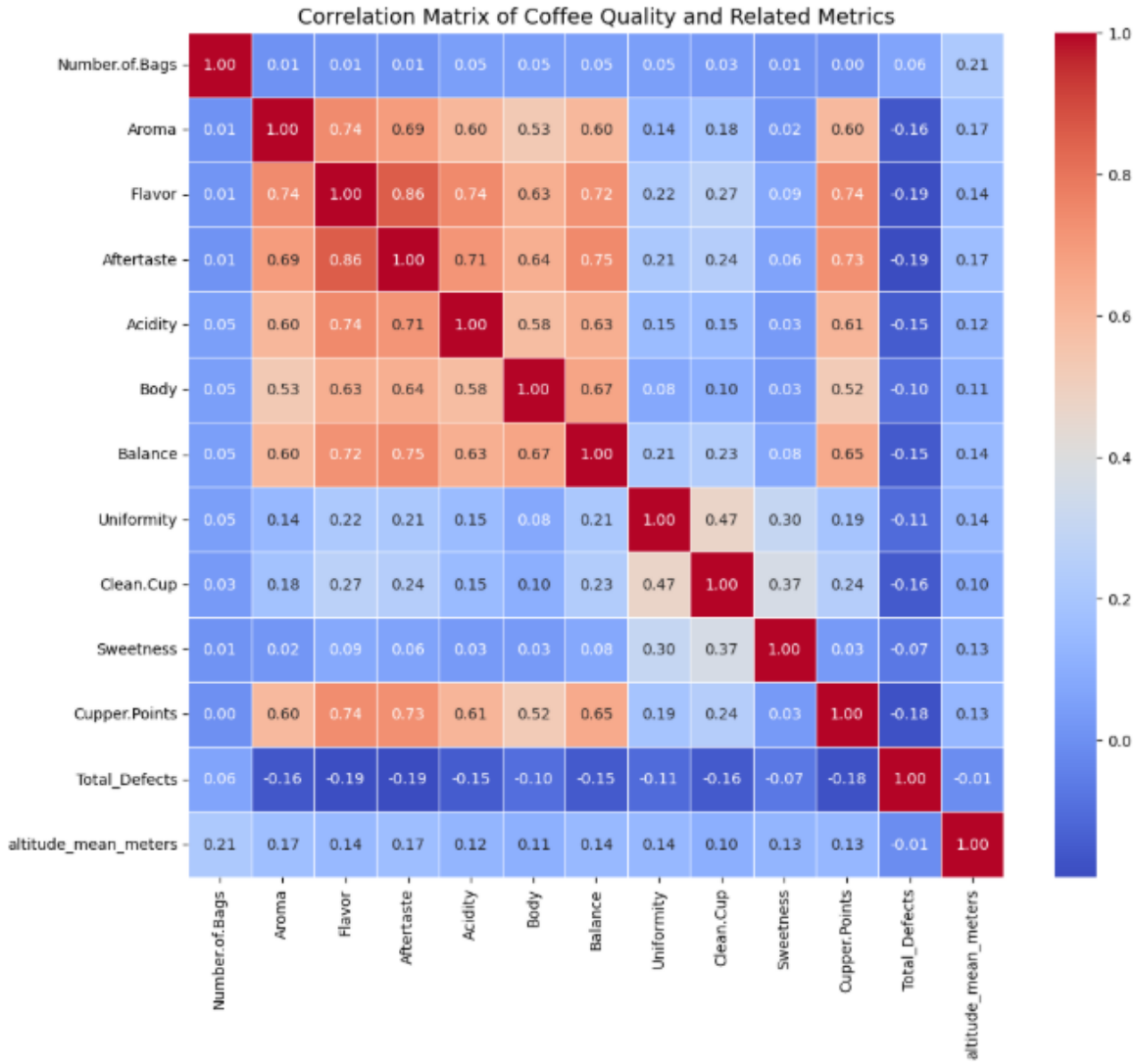


3. Multivariate Analysis

- Created a correlation matrix to explore relationships between attributes.
    - Found strong positive correlations among Aroma, Flavor, Aftertaste, and Cupper.Points.
    - Total_Defects showed a slight negative impact on Cupper.Points.
- Used scatter plots to analyze:
    - Bags vs Cup Points
    - Bags vs Total Defects
- Found that higher bag counts did not always correlate with higher cup scores or more defects.

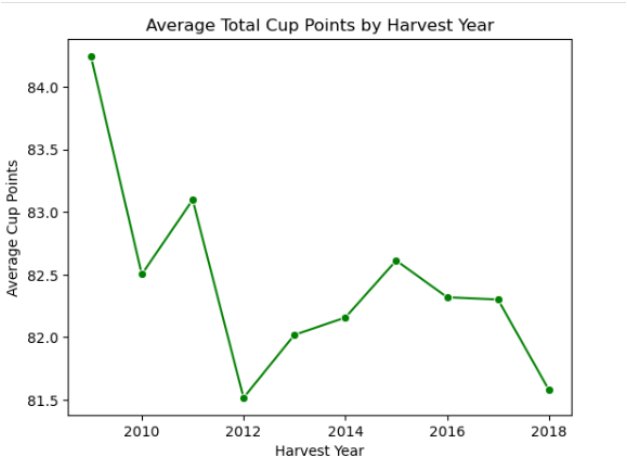


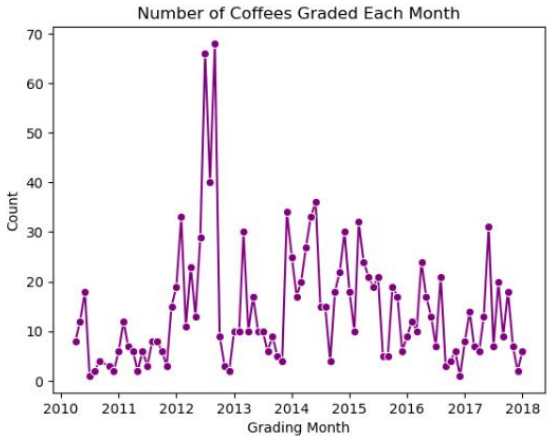


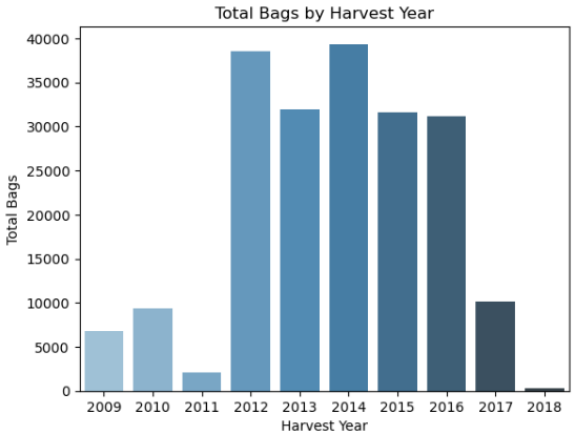

Correlation Matrix of Coffee Quality and Related Metrics
Number.of.Bags 1.00 0.01 0.01 0.01 0.05 0.05 0.05 0.05 0.03 0.01 0.00 0.06 0.21
Aroma 0.01 1.00 0.74 0.69 0.60 0.53 0.60 0.14 0.18 0.02 0.60 -0.16 0.17
Flavor 0.01 0.74 1.00 0.86 0.74 0.63 0.72 0.22 0.27 0.09 0.74 -0.19 0.14
Aftertaste 0.01 0.69 0.86 1.00 0.71 0.64 0.75 0.21 0.24 0.06 0.73 -0.19 0.17
Acidity 0.05 0.60 0.74 0.71 1.00 0.58 0.63 0.15 0.15 0.03 0.61 -0.15 0.12
Body 0.05 0.53 0.63 0.64 0.58 1.00 0.67 0.08 0.10 0.03 0.52 -0.10 0.11
Balance 0.05 0.60 0.72 0.75 0.63 0.67 1.00 0.21 0.23 0.08 0.65 -0.15 0.14
Uniformity 0.05 0.14 0.22 0.21 0.15 0.08 0.21 1.00 0.47 0.30 0.19 -0.11 0.14
Clean.Cup 0.03 0.18 0.27 0.24 0.15 0.10 0.23 0.47 1.00 0.37 0.24 -0.16 0.10
Sweetness 0.01 0.02 0.09 0.06 0.03 0.03 0.08 0.30 0.37 1.00 0.03 -0.07 0.13
Cupper.Points 0.00 0.60 0.74 0.73 0.61 0.52 0.65 0.19 0.24 0.03 1.00 -0.18 0.13
Total_Defects 0.06 -0.16 -0.19 -0.19 -0.15 -0.10 -0.15 -0.11 -0.16 -0.07 -0.18 1.00 -0.01
altitude_mean_meters 0.21 0.17 0.14 0.17 0.12 0.11 0.14 0.14 0.10 0.13 0.13 -0.01 1.00
Number.of.Bags
Aroma
Flavor
Aftertaste
Acidity
Body
Balance
Uniformity
Clean.Cup
Sweetness
Cupper.Points
Total_Defects
altitude_mean_meters
1.0
0.8
0.6
0.4
0.2
0.0

4. Trends Over Time

- Analyzed production and quality trends over the years.
- Found peak production around 2013 and a gradual decline in cup scores after 2009.

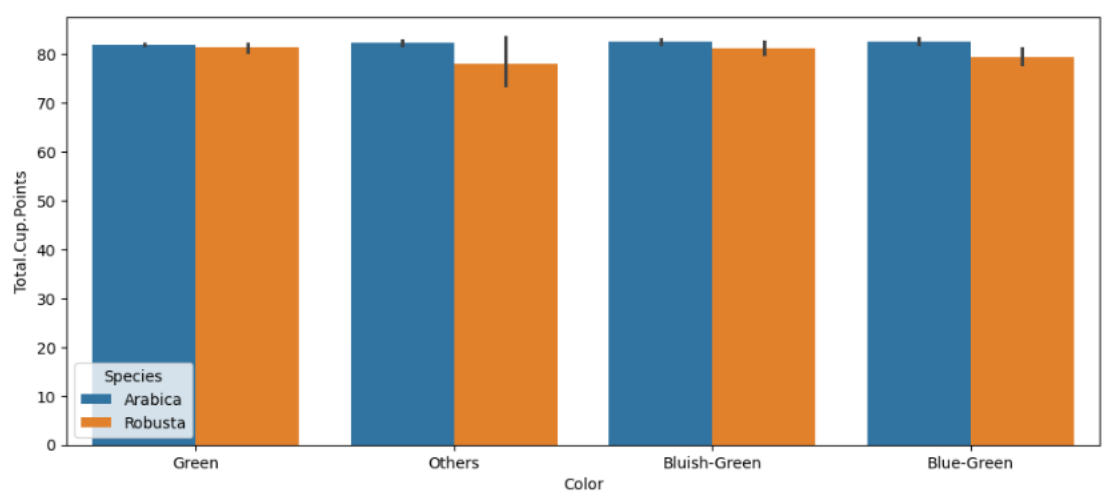






5. Processing & Color Insights

- Washed/Wet emerged as the most preferred method globally.
- Bean color had minimal impact on quality scores.



9.Hypothesis Testing

1. Independent T-Test – Arabica vs Robusta (Cup Points)

- Purpose: To determine whether the type of coffee species (Arabica or Robusta) affects the average cup score.
- Null Hypothesis (H_0): The mean cup points for Arabica and Robusta are equal.
- Alternative Hypothesis (H_1): The mean cup points for Arabica and Robusta are different.
- Test Used: Independent two-sample T-test
- Rationale: Arabica and Robusta are the two main coffee species. This test checks if the commonly held belief that Arabica scores higher than Robusta is statistically supported.
- Result: A statistically significant difference was observed, confirming that species type does influence cup quality.

```
T-Test: Arabica vs Robusta Cup Points
T-statistic: 2.8078234645068805
P-value: 0.008918153430452722
Reject the null hypothesis: Significant difference in cup points.
```

2. One-Way ANOVA – Variety vs Total Defects

- Purpose: To test whether the type of coffee variety has an effect on the number of defects found in beans.
- Null Hypothesis (H_0): All varieties have the same average number of defects.
- Alternative Hypothesis (H_1): At least one variety has a different average number of defects.
- Test Used: One-Way ANOVA
- Rationale: Different coffee varieties (e.g., Bourbon, Typica, Caturra) may have varying resistance to defects during processing.
- Result: The test showed a significant variation in defect levels across some varieties, suggesting that variety may impact bean quality.

```
ANOVA: Variety vs Total Defects
F-statistic: 3.02994651456177
P-value: 5.629723453566861e-07
Reject the null hypothesis: Variety significantly affects defect rate.
```

3. One-Way ANOVA – Country of Origin vs Cup Points

- Purpose: To examine whether the country where the coffee is produced has an effect on its quality score.
- Null Hypothesis (H_0): All countries have the same mean cup points.
- Alternative Hypothesis (H_1): At least one country has a different mean cup score.
- Test Used: One-Way ANOVA
- Rationale: Country of origin includes variations in climate, soil, and processing techniques that could influence cup quality.
- Result: The test revealed statistically significant differences in cup scores across countries, confirming that origin impacts coffee quality.

```
ANOVA: Country of Origin vs Total Cup Points
F-statistic: 8.322852002175953
P-value: 2.4077592442214183e-37
Reject the null hypothesis: Country significantly impacts cup quality.
```

10.Key Insights

- Arabica coffee consistently receives higher cup scores than Robusta.
- It also has fewer defects, and makes up the majority of samples in the dataset, confirming industry expectations.
- Papua New Guinea, Ethiopia, and Japan are the top three countries in terms of average cup points.
- Country of origin significantly influences coffee quality.
- Farms like Metad Plc and El Aguila scored among the highest in cup quality.
- However, high cup points were not always associated with large production volumes.
- Washed/Wet processing was the most preferred method globally across countries.
- Processing methods showed minimal variation in average cup scores.
- Variety types like Bourbon and Typica were dominant but showed varying levels of defects.
- No strong correlation between bag count and cup points — high volume doesn't imply better quality.
- Surprisingly, larger batches had fewer defects, possibly due to better infrastructure at larger farms.

- Countries like Brazil and Colombia had the highest number of bags, reflecting large-scale production.
- Farms with the most bags were not necessarily those with the best quality - highlighting a quantity vs quality trade-off.
- Bean color (e.g., green, others) showed no significant impact on cup points.
- No clear linear relationship was found between altitude and cup score.
- However, most high-quality beans came from altitudes between 1200–2000 meters.
- Strongest correlations found between:
- Aroma, Flavor, Aftertaste, and Cupper Points
- Sweetness and Uniformity remained consistently high
- Total_Defects had a negative but weak correlation with Total.Cup.Points.
- Grading often happened months after harvesting, with an average delay of ~400 days.
- Cup quality declined from 2009 to 2018.
- Production peaked around 2013, followed by fluctuations in volume and quality.

11. Hypothesis Testing Insights

- T-Test (Arabica vs Robusta): Confirmed statistically significant difference in cup scores by species.
- ANOVA (Variety vs Defects): Some varieties are more prone to defects, impacting bean quality.
- ANOVA (Country vs Cup Points): Cup quality varies significantly across countries, confirming the importance of origin.

12.Conclusion

This project successfully explored global coffee quality data using Python to identify key factors affecting cup scores. Insights were gained about species performance, origin quality trends, processing methods, and sensory attributes. Arabica beans, certain countries, and specific farms consistently showed better quality, while defects and overproduction periods affected scores negatively. The analysis validates industry beliefs using data and provides evidence-backed insights for improvement.

13. Future Recommendations

1. Incorporate predictive modeling to forecast cup scores based on production conditions or attributes.
2. Explore farm-level data with more granularity to assess quality consistency over time.
3. Include environmental and economic variables (e.g., rainfall, temperature, market price) for deeper analysis.
4. Track certification impact more precisely to evaluate how much certification adds to coffee value.
5. Apply clustering to group similar coffee profiles and target market segments accordingly.
6. Automate dashboards for ongoing monitoring of quality, trends, and exporter performance.

Date : 04/06/2025 **by,**

Yahavarshini E