**Coffee Quality Dataset Analysis**

**Milestone-3**

# **1. Context**

Coffee quality is a key factor in global beverage markets. This project investigates how origin, species, processing methods, and flavor-related metrics impact the sensory quality of coffee beans. Real-world data is used to uncover patterns, trends, and correlations that define high-quality coffee.

# **2. Objective**

• Identify key features that influence cup scores

• Analyze the impact of species, country, and processing methods

• Visualize quality distribution across attributes and regions

• Provide insights for farmers, producers, and data enthusiasts

# **3. Scope**

• Exploratory Data Analysis (EDA) using Python

• Focus on cup attributes, species comparison, and processing trends

• No predictive modeling or machine learning included

# **4. Audience**

This project is ideal for:

• Data science students/analysts improving their EDA skills

• Coffee producers seeking data-driven improvements

• Agricultural researchers

• Recruiters reviewing portfolio work

# **5. Techniques Used**

**1. Data Cleaning:**

Dropped null-heavy or irrelevant fields

Handled missing values in species, processing, and defects

Standardized categorical columns

**2. EDA (Exploratory Data Analysis):**

Distribution plots, boxplots, and bar charts

Grouping by country, species, and processing method

**3. Correlation Analysis:**

Heatmaps to identify relationships

Numerical correlation matrices

**4. Grouping and Aggregation:**

Used group by () to compute averages, totals, and comparisons across countries,

species, farms, and processing methods

**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

**7. Feature Engineering:**

• Created HighQuality label based on score threshold

• Altitude categories added for trend analysis

**8. Reporting:**

• Summary document for stakeholders

# **6. Dataset Overview**

• Total Rows: 1,339

• Columns: 44

• Data Sources: Global coffee farms evaluated by professional graders

**Categories:**

1. **Farm Metadata:** Country, Region, Owner, Farm Name

2. **Bean Properties:** Species, Processing Method, Color

3. **Defects:** Category 1/2 Defects, Quakers

4. **Sensory Attributes:** Aroma, Flavor, Aftertaste, Balance, etc.

5. **Quantities:** Number of Bags, Bag Weight

6. **Scoring:** Total Cup Points, Overall Score

# **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,

• 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.

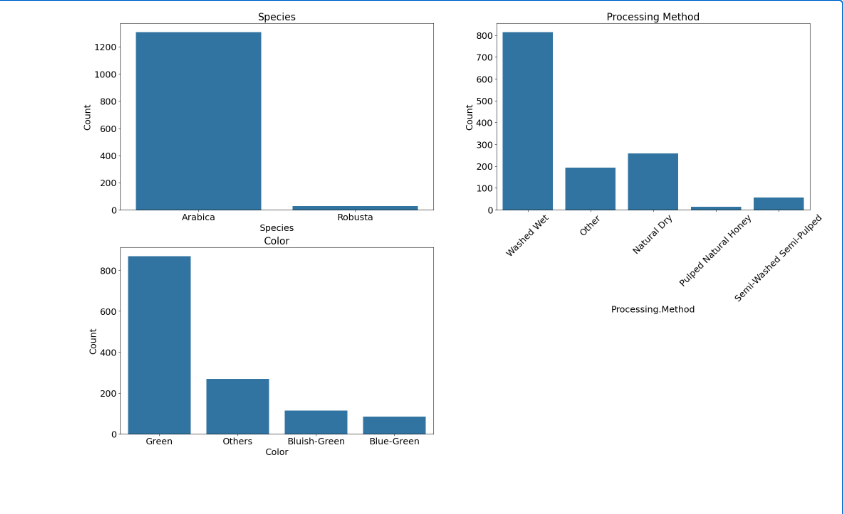
• Replaced outliers with median or other statistical methods where needed.

• 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**

**1. Univariate Analysis**

• Most sensory scores (e.g., Aroma, Flavor) are normally distributed

• Arabica beans dominate dataset (~90%)

• Total Cup Points generally between 80–85

• Very few samples are classified as Robusta

**2. Bivariate Analysis**

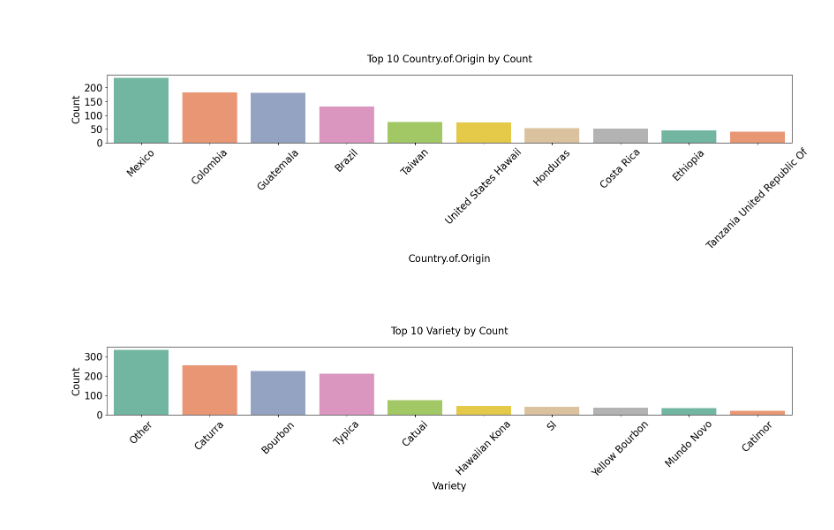
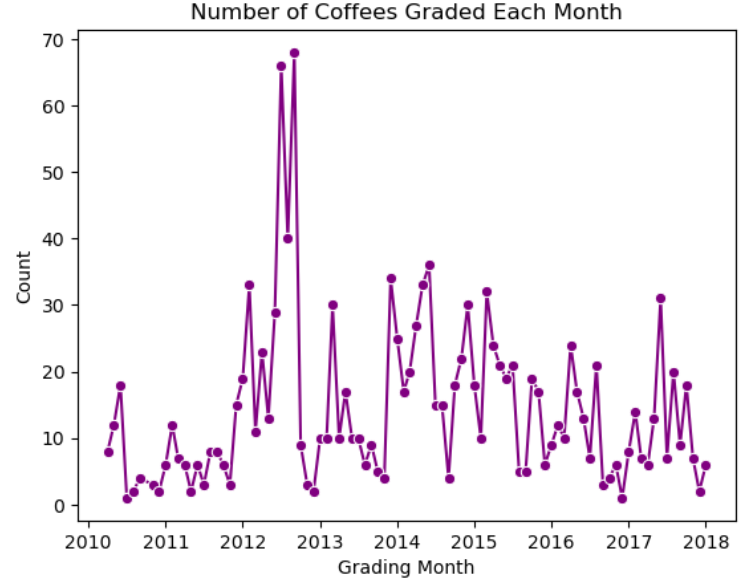
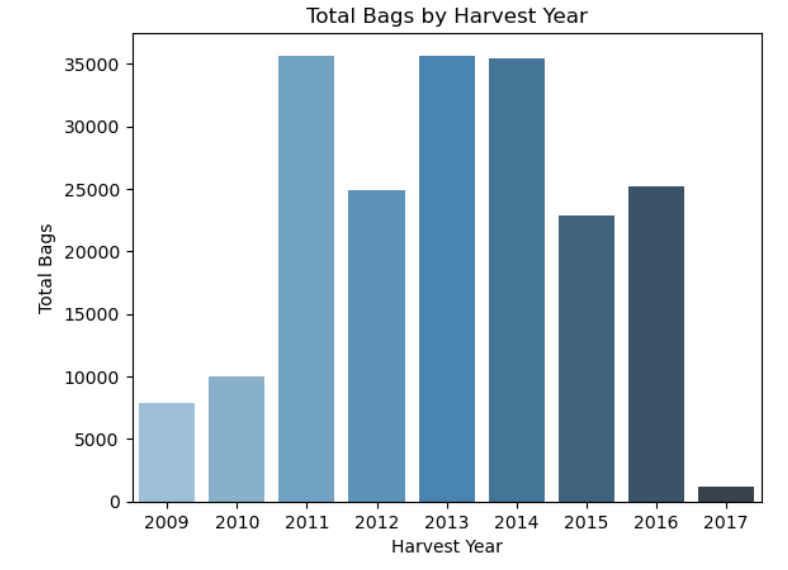
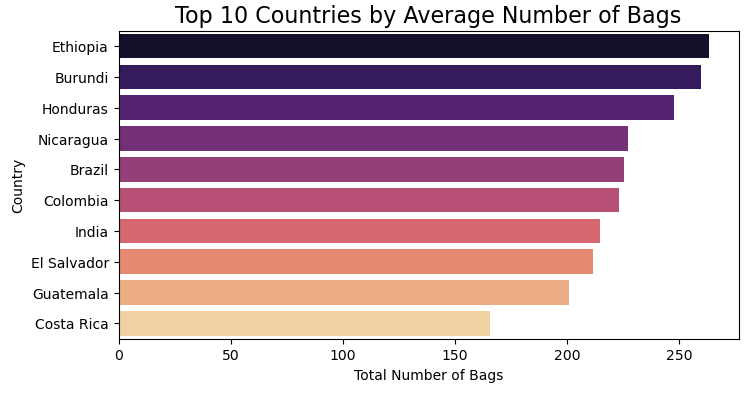
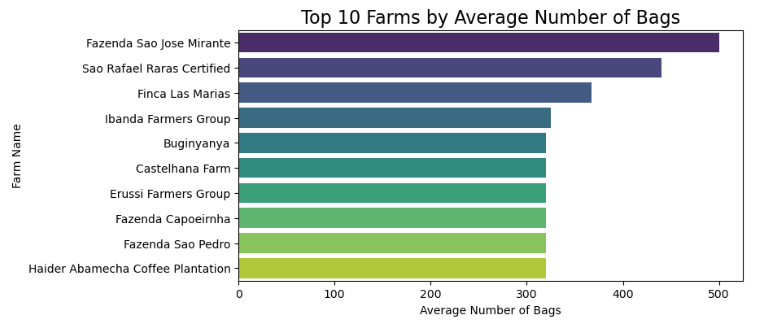
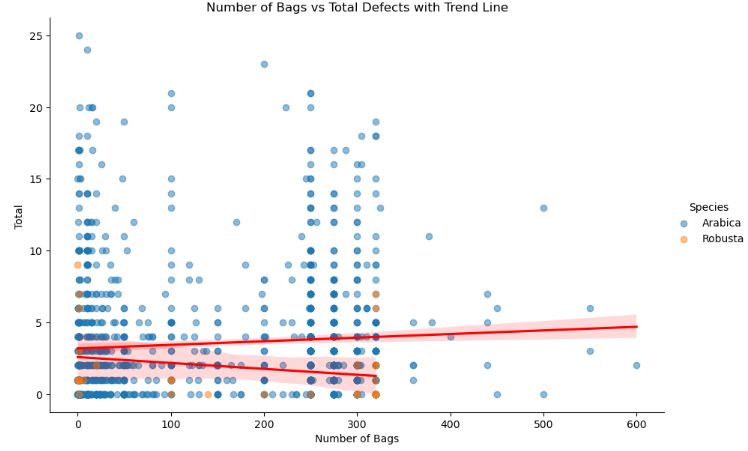
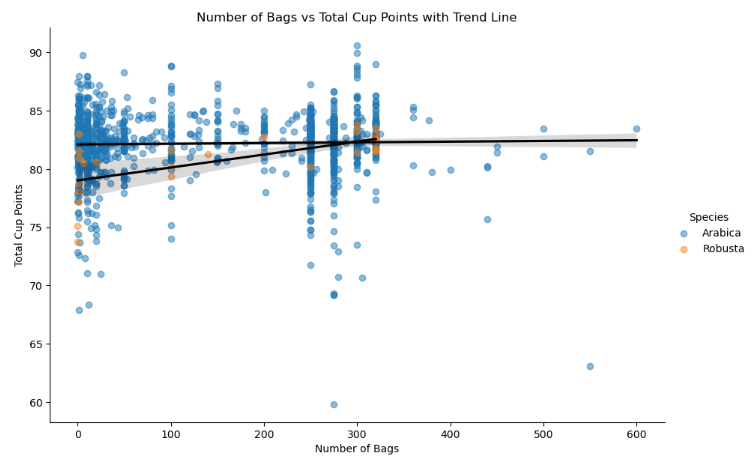
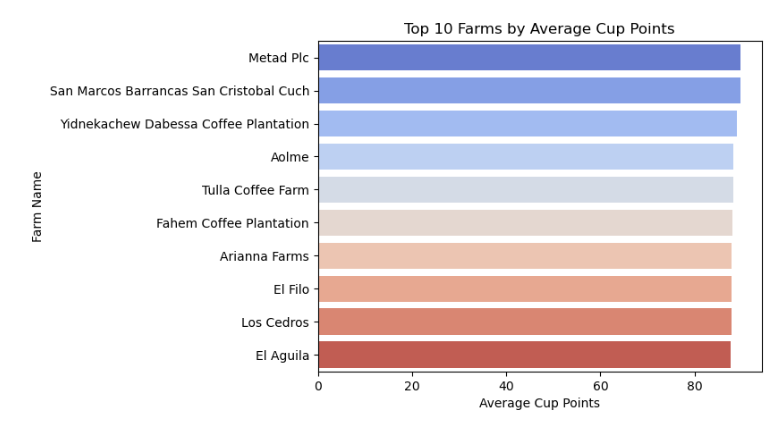
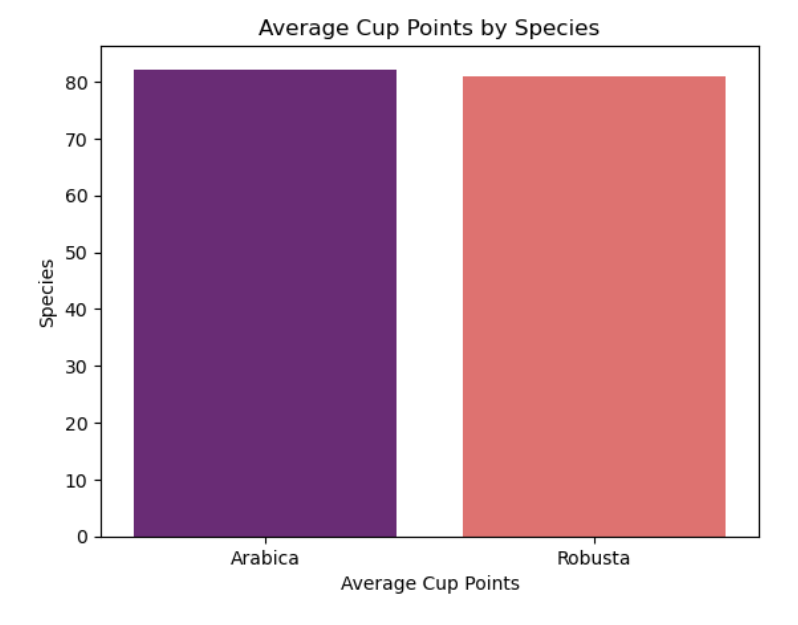
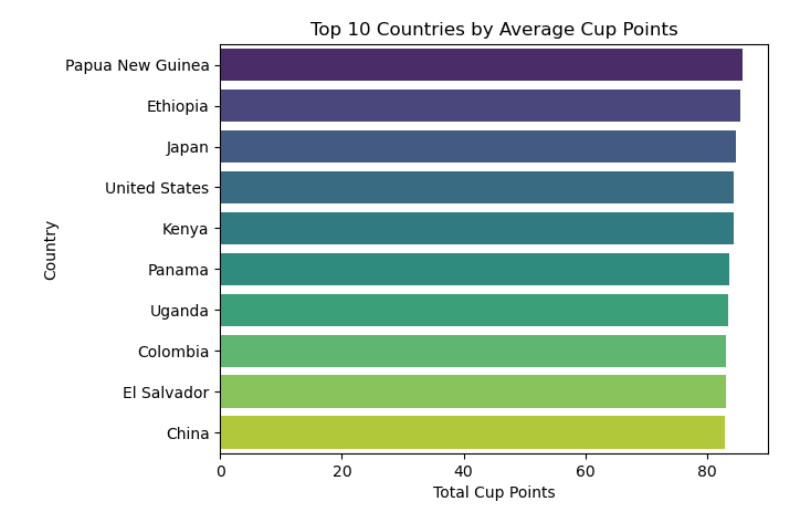
• Country vs Cup Points:

• Ethiopia and Colombia lead in average quality

• Species vs Quality:

• Arabica consistently scores higher

• Processing vs Attributes:

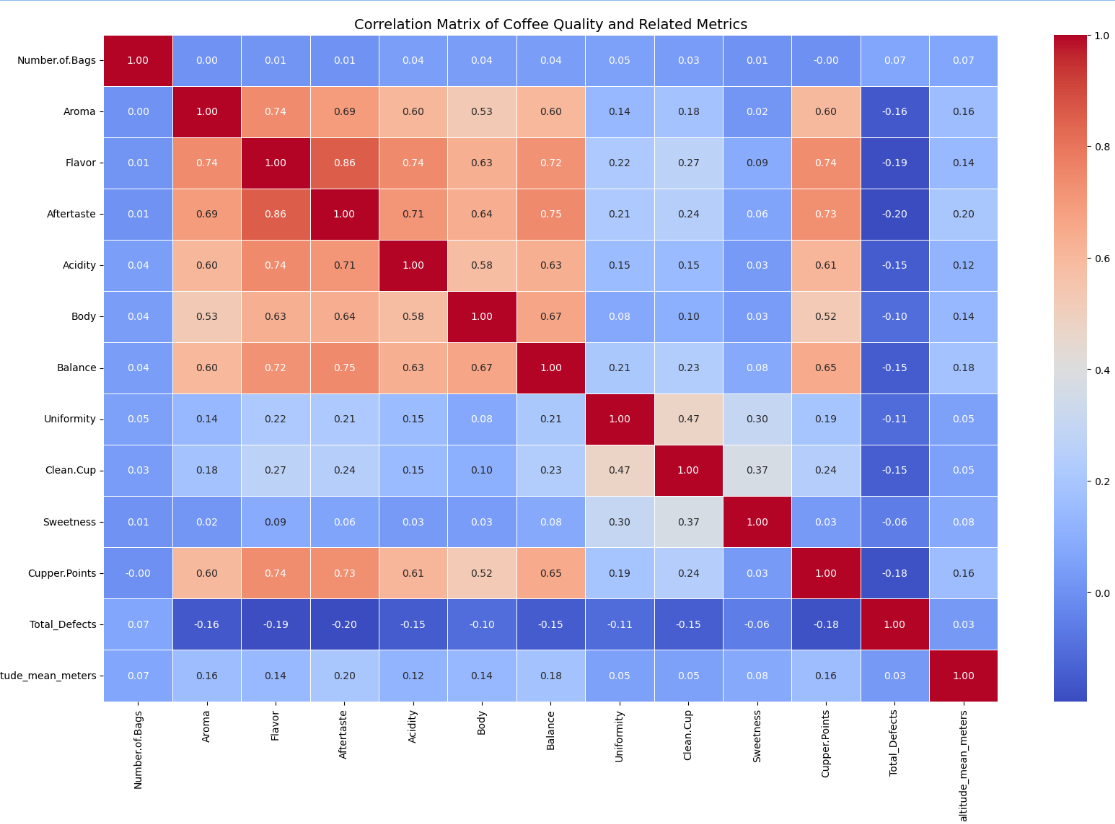
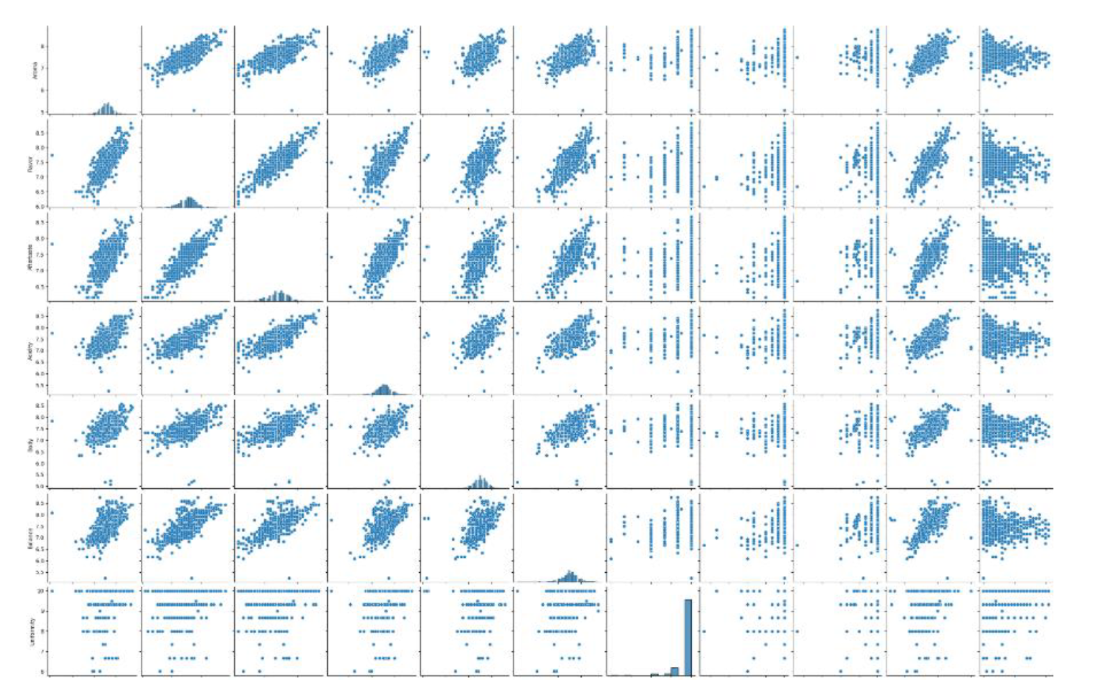
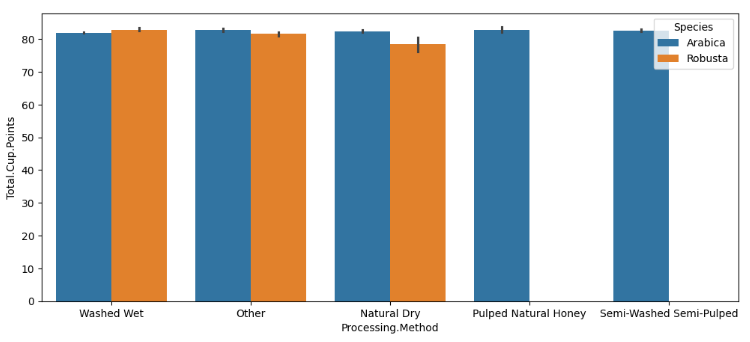
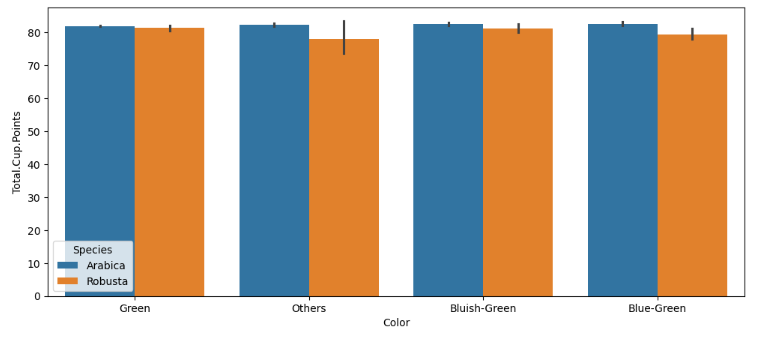
• Washed methods show higher balance and cleanliness 

**3. Multivariate Analysis**

• Aroma, Flavor, Aftertaste, and Overall are strongly correlated

• Total Defects negatively impact Cup Points

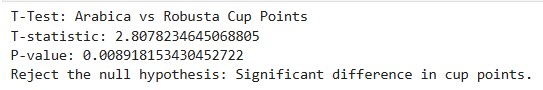
• Number of Bags doesn't correlate with quality



# **9. Hypothesis Testing**

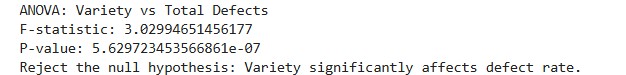
**1. T-Test – Arabica vs Robusta**

* Purpose: To determine whether the type of coffee species (Arabica or Robusta) affects the average cup score.
* Null Hypothesis (H₀): The mean cup points for Arabica and Robusta are equal.
* Alternative Hypothesis (H₁): 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.



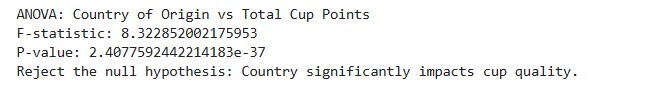
**2. ANOVA – Variety vs Defects**

* Purpose: To test whether the type of coffee variety has an effect on the number of defects found in beans.
* Null Hypothesis (H₀): All varieties have the same average number of defects.
* Alternative Hypothesis (H₁): 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.



**3. ANOVA – Country vs Cup Points**

* Purpose: To examine whether the country where the coffee is produced has an effect on its quality score.
* Null Hypothesis (H₀): All countries have the same mean cup points.
* Alternative Hypothesis (H₁): 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.



# **10. Key Insights**

• Arabica dominates the dataset, making up the majority of all coffee samples evaluated.

• Papua New Guinea, Ethiopia, and Japan topped the charts when it came to average cup scores — these countries are producing some of the finest coffee.

• Where a coffee comes from matters a lot. The data clearly shows that the country of origin has a strong effect on quality.

• Farms like Metad Plc and El Aguila consistently scored high, proving that farm-level practices can make a big difference.

• Surprisingly, high cup scores didn’t always come from big producers. In other words, small farms can produce world-class coffee too.

• The Washed/Wet processing method was the most popular across the board — and for good reason, as it’s linked with cleaner and better-tasting beans.

• That said, we found that different processing methods didn’t drastically change the scores, which was unexpected.

• Coffee varieties like Bourbon and Typica were the most common, but their defect levels varied quite a bit.

• Producing more bags doesn’t mean better quality — we didn’t find a strong link between volume and cup scores.

• In fact, larger farms tended to have fewer defects, probably because they have better tools and processes in place.

• Brazil and Colombia were the biggest producers by volume, but their coffees weren’t always the highest in quality — showing a classic case of quantity vs. quality.

• The color of the coffee beans didn’t seem to influence the scores much at all.

• We didn’t see a clear pattern between altitude and cup scores, but most top-quality beans came from farms located between 1,200 and 2,000 meters.

• Aroma, Flavor, and Aftertaste had the strongest influence on how a coffee scored overall — no surprise there, as those are the core tasting elements.

• Sweetness and Uniformity were solid across most samples, staying consistently high.

• Defects in the beans slightly reduced scores, but the impact wasn’t dramatic.

• One interesting find: coffee grading often happened several months after harvesting — on average, about 400 days later!

• Quality dipped between 2009 and 2018, even though production peaked around 2013. This could be linked to scaling challenges or climate shifts.

# **11. Conclusion**

This analysis confirms industry trends through data. Arabica, Washed methods, and highland origins produce better coffee. Data cleaning, visualization, and hypothesis testing validate key assumptions and provide actionable insights for coffee producers, analysts, and businesses.

# **12. Future Recommendations**

1. Add predictive modeling (e.g., regression for cup score)

2. Segment coffees using clustering (K Means)

3. Incorporate climate and cost data for deeper analysis

4. Build real-time dashboards for producer monitoring

5. Analyze temporal trends by harvest year

**By : Karthick. T**