# Capstone Project: Exploring Coffee Quality Data with Power BI

ODIN School – Capstone Project.

#### Developed by – Ramkumar Sembaiyan [S9034]

#### Date submitted :6 - Aug -2024

Catalog

[Capstone Project: Exploring Coffee Quality Data with Power BI 1](#_Toc9200)

[Executive Summary 3](#_Toc28692)

[Project Objectives 3](#_Toc14460)

[Data Overview 3](#_Toc19828)

[Data Collection and Cleaning 3](#_Toc9709)

[1. CofffeeFactTable 4](#_Toc28816)

[2. DimCountry 4](#_Toc1520)

[3. DimVariety 4](#_Toc20907)

[4. DimProcessing 4](#_Toc18394)

[5. DimRegion 4](#_Toc8793)

[6. DateTable 4](#_Toc16418)

[Data Exploration and Analysis 5](#_Toc29098)

[Reports Development 6](#_Toc17502)

[Key Findings 8](#_Toc16012)

[2. Grading Date Trends 8](#_Toc16530)

[3. Defect Analysis 9](#_Toc21124)

[4. Processing Methods and Total Cup Scores 9](#_Toc22417)

[5. Variety Distribution 9](#_Toc20206)

[6. Production Volume 9](#_Toc28602)

[7. Quality Score by Origin 9](#_Toc26971)

[8. Influence to Total Cup Points 10](#_Toc22112)

[Recommendations 10](#_Toc30665)

[Conclusion 10](#_Toc14228)

[Next Steps 10](#_Toc13264)

## Executive Summary

This report presents a comprehensive analysis of global coffee quality, leveraging data from the Coffee Quality Institute (CQI). Using Power BI Desktop, I have explored various aspects of coffee production, processing, and quality evaluation to provide insights for stakeholders in the coffee industry.

## Project Objectives

1. Identify key determinants of coffee quality through sensory attribute analysis.
2. Explore correlations between processing methods, origin regions, and quality scores.
3. Analyze trends in defect occurrences and their impact on overall coffee quality.
4. Investigate the interaction of variables influencing Total Cup Points.

## Data Overview

The [dataset](https://s3.amazonaws.com/grey_campus/production/system/OdinSchool/2023/Edmingle/DataSet/df_arabica_clean.zip) from CQI includes:

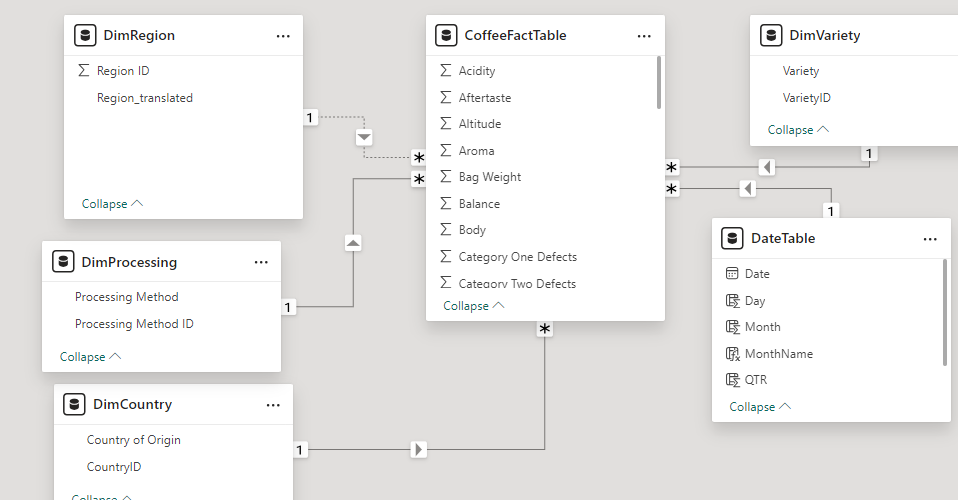
* Sensory evaluations (aroma, flavor, aftertaste, acidity, body, balance, uniformity, clean cup, sweetness)
* Defect counts (Category One and Two. Quakers)
* Processing methods
* Origin information (country, harvest year, variety)

## Data Collection and Cleaning

1. The CSV file "[df\_arabica\_clean.csv](https://s3.amazonaws.com/grey_campus/production/system/OdinSchool/2023/Edmingle/DataSet/df_arabica_clean.zip)" was imported into Power BI desktop using ‘Get Data’ option. **Reference : CapstoneProject\_Coffee\_Odin\_PowerBI\_S9034.pbix**
2. Explored and prepared the data in Power Query Editor using ‘Transform Data Button’
   1. Data types validation and Fix
      1. Extracted the numerical value from ‘Number of Bags’, changed to whole number
      2. Split the Harvest column and standardized the values. Changed to whole number
      3. Split the altitude column and standardized. Changed to whole number
      4. Changed the Grading data and Expiration data columns to Date Type [by removing ‘th,rd’ in the values]
      5. Blank values in Variety,Region columns are replaced by ‘Unknown’ value
      6. Verify there are no errors , null and duplicates in the columns.[View->Column quality]
   2. Creation of Fact and Dim Table[ This is not a mandatory step for this data set, however to explore Power BI and data these were created]
      1. CofffeeFactTable
      2. DimCountry
      3. DimVariety
      4. DimProcessing
      5. DimRegion
      6. DateTable

Steps Involved :

* These Dim tables were created in PQEditor by right clicking the Query name ,
* Removing unwanted columns In new Query/Renamed to Dim table
* Added index columns as ref IDs in Dim table after removing the duplicates.
* Later refIDs are merged/mapped back to CofffeeFactTable .
* The detail columns for these refIDs are removed from the Fact table
* Apply and Close the Power Query Editor. This will load the data into power BI desktop
  1. Data Modeling [I didn’t see an advantage in doing this for this dataset. Explored steps]
     1. Go to Table View, verify all the data [207 rows] are loaded in Fact table and verify no errors and duplicates in Dim tables
     2. Go to Model view and confirm 1 to many relation ship is created between Dim and Fact tables.



Data are now ready to explore the objectives.

## Data Exploration and Analysis

#### New Measures and Columns :

Following new objects were created to explore

1. ***Total Defects*** *= 'CoffeeFactTable'[Category One Defects] + 'CoffeeFactTable'[Category Two Defects]+'CoffeeFactTable'[Quakers]*
2. ***Total Bags*** *= SUM('CoffeeFactTable'[Number of Bags])*
3. ***Total Sample Weight*** *= SUMX('CoffeeFactTable',*

*[Number of Bags] \* [Bag Weight])*

1. ***Coffee Quality Score*** *=*

*VAR TotalWeight = SUMX('CoffeeFactTable', 'CoffeeFactTable'[Number of Bags] \* 'CoffeeFactTable'[Bag Weight])*

*RETURN*

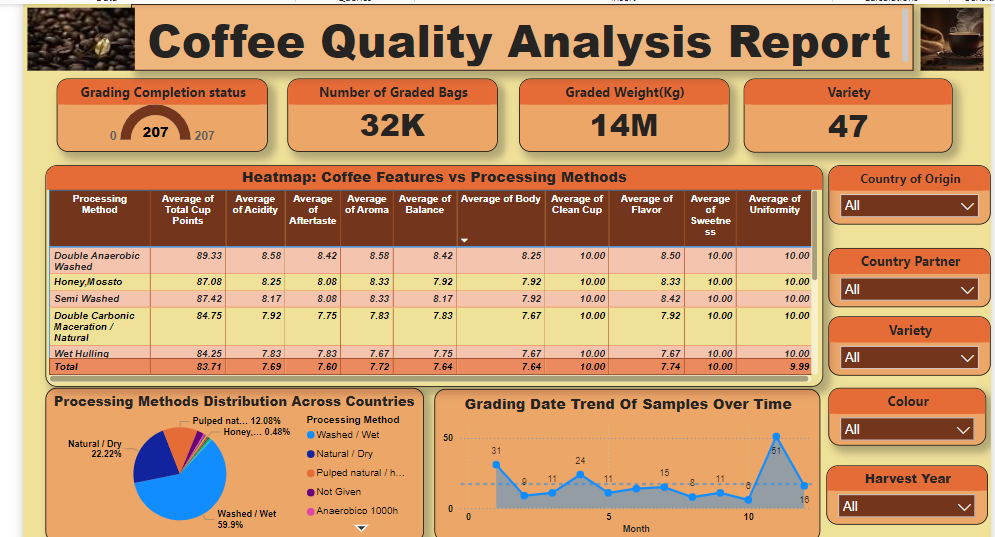
*SUMX('CoffeeFactTable', (('CoffeeFactTable'[Total Cup Points] -'CoffeeFactTable'[Total Defects{Cat 1+2+Q}]) \* 'CoffeeFactTable'[Number of Bags] \* 'CoffeeFactTable'[Bag Weight])) / TotalWeight*

## 

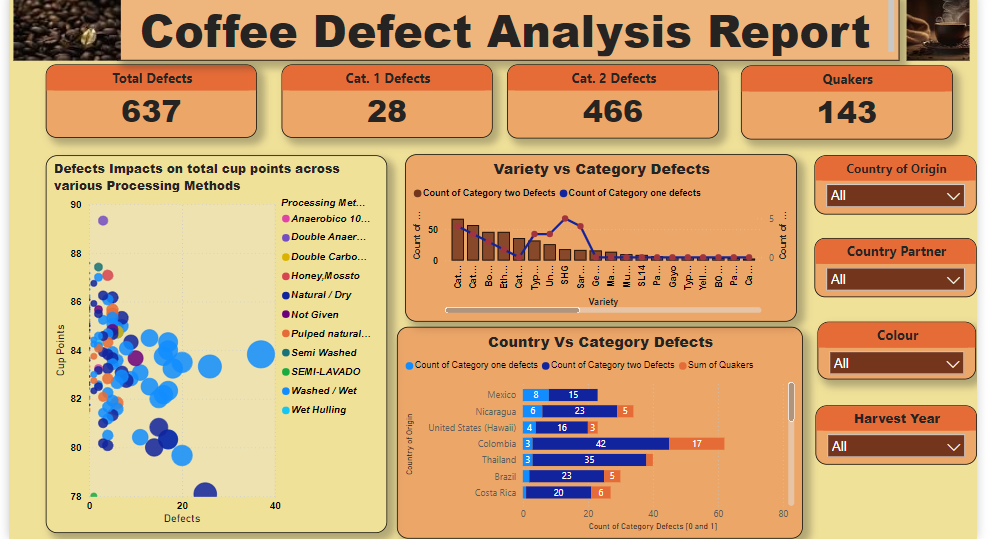
## Reports Development

Following Reports were created in Power BI desktop to explore and analyse this dataset.

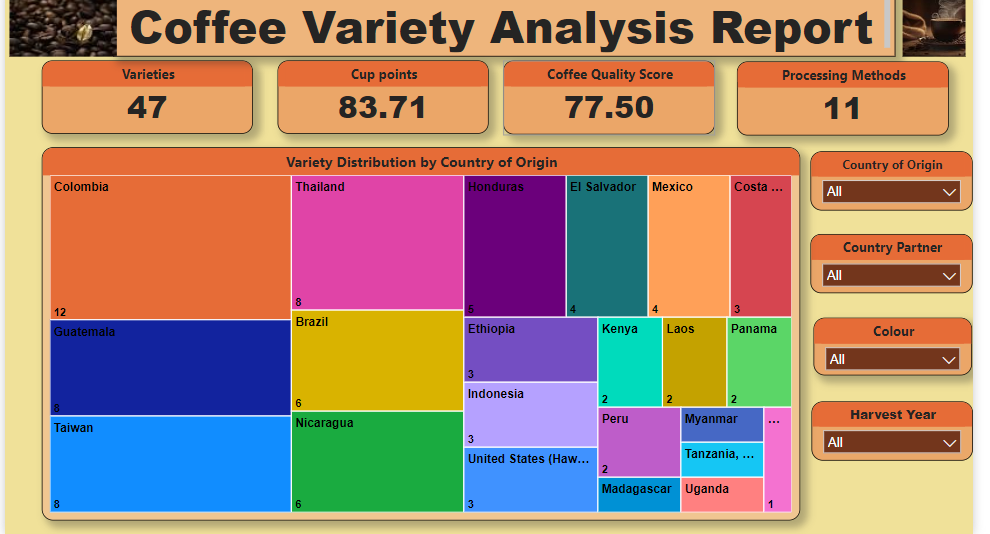
#### Report 1: Heatmap: sensory features vs Processing Methods vs Cup points.



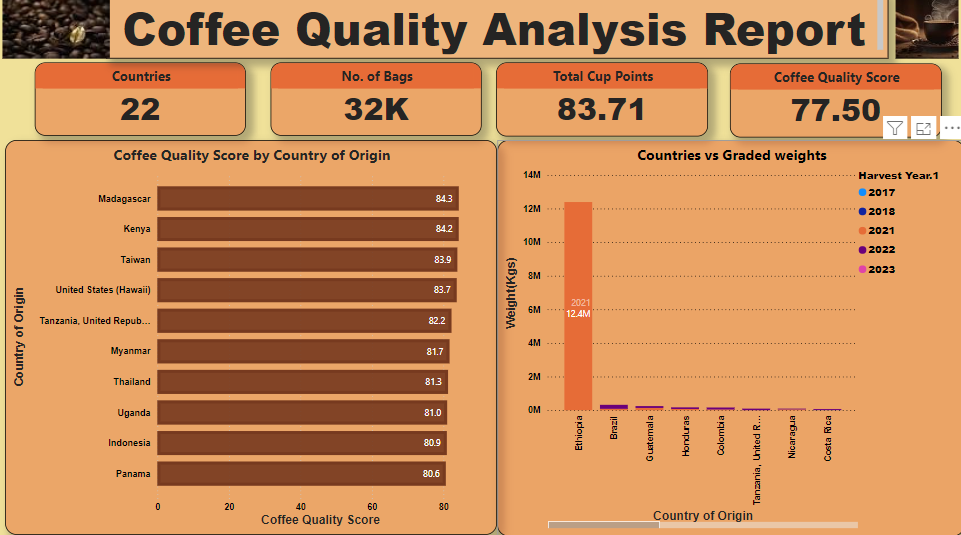
**Report 2: Defect Analysis**



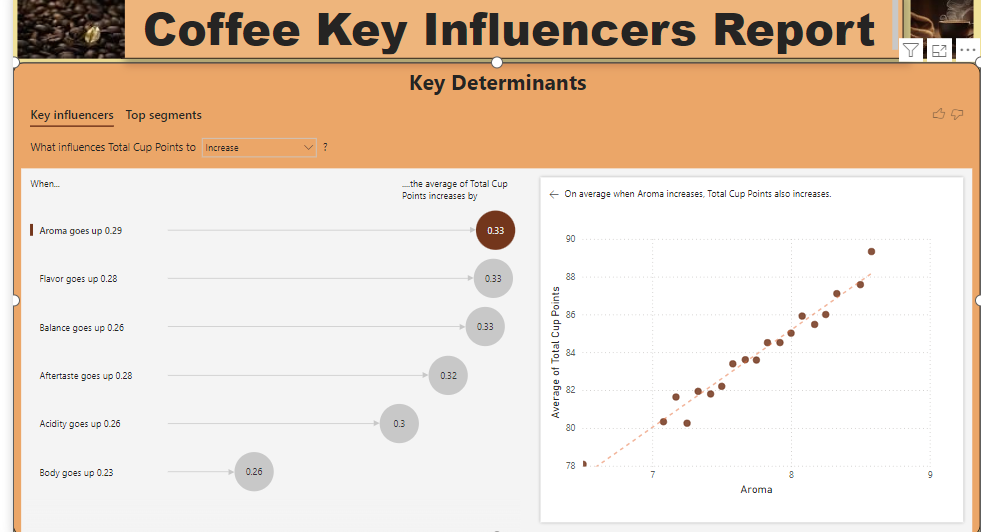
**Report 3: Variety Analysis**



**Report 4: Coffee Quality Score**



**Report 5: Key Determinants**



## Key Findings

#### 1. Processing Methods Distribution

* Dominant method: Washed/Wet (59.9%)
* Second most common: Natural/Dry (22.22%)
* Third most common: Pulped Natural/Honey (12.08%)

**Insight**: Processing methods significantly influence coffee flavor profiles and quality. Regional preferences for specific methods may indicate opportunities for quality improvement or specialty coffee development.

### 

### **2. Grading Date Trends**

* Peak grading month: November (51 samples)
* Secondary peaks: January (31 samples), April (24 samples)

**Insight**: Grading activity shows seasonality, likely linked to harvest cycles and market demands. This information can help optimize supply chain management and market timing.

### **3. Defect Analysis**

* Category One defects: 28 instances
* Category Two defects: 466 instances
* Quakers: 143 instances
* Countries with higher defect counts: ElSavador, Ethiopia
* Washed/Wet processing method contributes higher defect counts
* Variety Caturaa and catuai contributes higher defect counts

**Insight**: The prevalence of Category Two defects suggests a need for improved quality control measures, particularly in the identified countries.

### **4. Processing Methods and Total Cup Scores**

* Top performing methods: "Double Anaerobic Washed" and "Honey, Mossto"
* These methods achieve higher scores across all sensory attributes

**Insight**: Exploring and potentially adopting these processing methods in other regions could enhance overall coffee quality.

### **5. Variety Distribution**

* Highest variety diversity: Colombia, Guatemala, Taiwan, Thailand

**Insight**: The wide range of coffee varieties in these countries contributes to their rich flavor profiles and high-quality scores.

### **6. Production Volume**

* Top producers: Ethopia

**Insight**: Understanding production volumes helps in analyzing supply dynamics in the global coffee market. - Sample shows Ethopia contributed 19200 kg of 320 bags-. Verification with CQI on this data correctness is pending

### **7. Quality Score by Origin**

* Highest average quality scores: Madagascar and Kenya

**Insight**: These countries demonstrate superior coffee quality, potentially due to optimized production and processing methods.

### **Influence to Total Cup Points**

* When Aroma goes up 0.29 Total cup points increases by 0.33
* When Flavour goes up 0.29 Total cup points increases by 0.33
* When Balance goes up 0.29 Total cup points increases by 0.33
* When AfterTaste goes up 0.29 Total cup points increases by 0.32
* When Acidity goes up 0.29 Total cup points increases by 0.3
* When Body goes up 0.29 Total cup points increases by 0.26

**Insight**: Using the key influencer visual in power BI in Visual view,we are able to identify above impacts.

## Recommendations

1. **Process Optimization**: Encourage adoption of high-performing processing methods like "Double Anaerobic Washed" in suitable regions.
2. **Defect Reduction**: Implement targeted quality control measures in countries with high defect counts.
3. **Varietal Diversity**: Promote coffee variety diversification to enhance flavor profiles and meet diverse consumer preferences.
4. **Seasonal Planning**: Align production cycles with peak grading periods to ensure timely evaluation and market readiness.
5. **Best Practices Exchange**: Facilitate knowledge sharing between high-scoring origins (e.g., Madagascar , Kenya) and other producing regions.

## Conclusion

This analysis provides valuable insights into the factors influencing coffee quality globally. By leveraging these findings, stakeholders can make informed decisions to enhance coffee production, processing, and quality evaluation practices, ultimately contributing to the advancement of the global coffee industry.

## Next Steps

1. Conduct in-depth analysis of high-performing processing methods to identify key success factors.
2. Develop predictive models for coffee quality based on origin, processing method, and variety.
3. Create interactive dashboards for real-time monitoring of coffee quality trends and defect occurrences.

## Appendix

To explore Power BI service, I was able to publish to Power BI service portal and able to view the reports there.

