# Data Source: coffee quality Data

*Data Sourcing & Summary*

This is an external data source that has been provided by [Fatih B. on Kaggle](https://www.kaggle.com/datasets/fatihb/coffee-quality-data-cqi) and has been sourced from the [Coffee Quality Institute](https://database.coffeeinstitute.org/) (CQI). The mission of the CQI is to promote coffee quality through a range of activities that include research, training, and certification programs. The organization works with coffee growers, processors, roasters, and other stakeholders to improve coffee quality standards, promote sustainability, and support the development of the specialty coffee industry. The data was scraped from the CQI list of arabica coffees. It contains 200+ arabica coffees across several countries and regions and includes information about the farms, altitude, sensory evaluations, total coffee quality score, and more.

I have chosen this data because I am interested in and passionate about coffee. I very much enjoy brewing espresso and discovering different types of coffees so I’d love to learn more about what types of coffees there are as well as what quality they hold.

*Data Collection*

The data is collected by the CQI which maintains a web database that serves as a resource for coffee professionals and enthusiasts who are interested in learning about coffee quality and sustainability. The database includes a range of information on coffee production, processing, and sensory evaluation. It also contains data on coffee genetics, soil types, and other factors that can affect coffee quality.

*Data Contents*

This data contains one table, the actual data table.

**df\_arabica\_clean.csv**

|  |  |
| --- | --- |
| **Field** | **Description** |
| Index | Index Number |
| ID | Listing ID |
| Country of Origin | Country where this coffee originated from |
| Farm Name | The name of the farm that grew the coffee |
| Lot Number | The lot number |
| Mill | The mill where the coffee was processed |
| ICO Number | International Coffee Organization number |
| Company | The company name |
| Altitude | Altitude of the farm |
| Region | Region of the farm |
| Producer | Who the producer is |
| # of Bags | # of bags from the harvest |
| Bag Weight | Weight of the bag |
| In-Country Partner | CQI partner organization for cupping |
| Harvest Year | The year the coffee was harvested |
| Grading Date | When the coffee was grated |
| Owner | Owner of the coffee lot |
| Variety | The type of plant |
| Status | Whether the processing was completed or not |
| Processing Method | The processing method used on the coffee |
| Aroma | Refers to the scent or fragrance of the coffee. |
| Flavor | The flavor of coffee is evaluated based on the taste, including any sweetness, bitterness, acidity, and other flavor notes. |
| Aftertaste | Refers to the lingering taste that remains in the mouth after swallowing the coffee. |
| Acidity | Acidity in coffee refers to the brightness or liveliness of the taste. |
| Body | The body of coffee refers to the thickness or viscosity of the coffee in the mouth. |
| Balance | Balance refers to how well the different flavor components of the coffee work together. |
| Uniformity | Uniformity refers to the consistency of the coffee from cup to cup. |
| Clean Cup | A clean cup refers to a coffee that is free of any off-flavors or defects, such as sourness, mustiness, or staleness. |
| Sweetness | It can be described as caramel-like, fruity, or floral, and is a desirable quality in coffee. |
| Overall | Overall rating of the coffee |
| Defects | Defects are undesirable qualities that can occur in coffee beans during processing or storage. |
| Total Cup Points | The total points from Aroma through Overall with defects subtracted |
| Moisture Percentage | The percentage of the moisture |
| Category One Defects | Category One defects are primary defects that can be perceived through visual inspection of the coffee beans. These defects include Black beans, sour beans, insect-damaged beans, fungus-damaged beans, etc. |
| Quakers | Quakers are coffee beans that don't turn dark brown when roasted. They are defective beans that remain light-colored. |
| Color | The color of the coffee |
| Category Two Defects | Category Two defects are secondary defects that are more subtle and can only be detected through tasting. These defects include Over-fermentation, staleness, rancidness, chemical taste, etc. |
| Expiration | When the coffee will expire |
| Certification Body | The certification body name that certified the coffee |
| Certification Address | Address of the certification body |
| Certification Contact | Contact for the certification body |

*Data Limitations*

Since this data is scraped directly from the CQI’s website, there are minimal limitations. Since coffees are only included that have been graded by a Q grader, some may be missing that have not been graded yet. However, we are only interested in those that have been graded as this is how we will complete our analysis. Another limitation may exist in the actual grading process. Different graders may grade in different ways, so gradings may not be entirely consistent. With that being said, great care is put into training graders and grading certificates expire every three years and require re-training to maintain your certificate. All this leads us to believe that there exists some consistency between grades, even if done by different graders.

*Data Ethics*

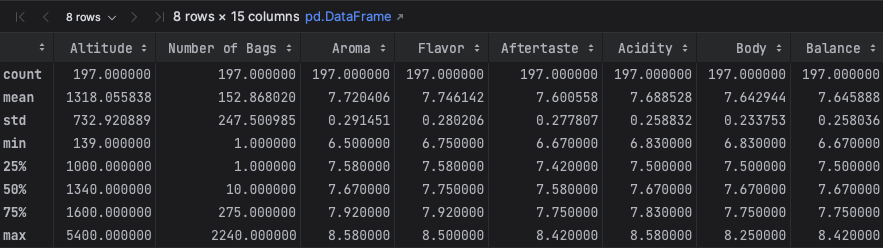
No private information is being used in the data. All the information here is public knowledge on the CQI website.

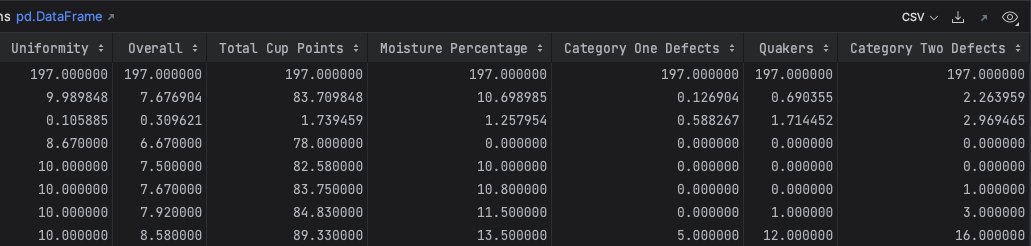
# Data cleaning & Profiling

**Data Cleaning, Wrangling, Consistency Checks**

|  |  |  |
| --- | --- | --- |
| Column | Type of Inconsistency | Action |
| Region |  | Changed the type to string |
| ID, Lot Number, Mill, ICO Number, In-Country Partner, Owner, Status, Certification Address, Certification Body, Certification Contact, Clean Cup, Sweetness, Defects |  | Dropped the column as it does not aid in our analysis |
| Farm Name, Variety, Processing Method | missing values | Since the missing values in these columns made up less than 5% of our data set, we deleted them. We deleted a total of 10 rows. |

**Data Profiling**





# Key Questions

* Does a specific country have better quality coffee than others? What about region?
* Does altitude have any impact on quality?
* Does any particular processing method lead to higher quality coffee?
* Does any particular variety lead to a higher quality coffee?