Data analysis of arabica coffee beans

1.Background

1.1 The market for coffee beans

According to market data research, the market size of China's coffee industry will reach 381.7 billion yuan in 2021, 485.6 billion yuan in 2022, and is expected to reach 617.8 billion yuan in 2023. The Chinese coffee market is entering a stage of rapid development. With the rise of coffee sales, the quantity of coffee beans imported and exported by China has also risen sharply. What kind of coffee beans to choose, which dimensions to evaluate the quality of coffee beans, and which origin of coffee beans to choose all directly affect consumers' choice and satisfaction.

1.2 Introduction of coffee beans

In terms of varieties, coffee beans can be divided into three types: Arabica, Robusta and Liberica. Among them, Arabica and Robusta are the main coffee varieties consumed by people in the world, while Liberica often has less yield and poor quality, so it is often not taken as the scope of research.

Arabica coffee beans account for 70-75% of the world's coffee production, which fully demonstrates the importance of Arabica coffee beans in the coffee industry. It is produced in South Africa, Africa, Asian countries and other places, but because this variety is less resistant to diseases and insect pests, highland areas are more suitable for cultivation, especially the quality of Arabica coffee beans produced in highlands above 1500 meters is the best. The good

quality produced by such efforts, with a balanced flavor and aroma, can be certified as high-grade coffee beans, which are mainly used by people to make single-origin coffee or specialty coffee.

1.3 The significance of analyzing Arabica coffee beans

People describe the taste of Arabica coffee beans, usually from Aroma, Flavor, Acidity, Aftertaste, Balance, Sweetness and other aspects to measure. The study selected the Arabica coffee bean dataset and divided Arabica into smaller subcategories for data analysis. Through the information on the origin, growth conditions, varieties, taste and other aspects of coffee beans, we can better understand the characteristics and advantages and disadvantages of this coffee bean. And through the comprehensive scoring of coffee beans, the coffee beans are rated to provide consumers with various scoring indicators of different grades of coffee beans, and provide consumers with purchasing opinions. In addition, we also discovered the influence of different factors on the quality and taste of Arabica coffee beans, so as to better control the production and processing of coffee beans and improve the quality and taste of coffee beans.

2.Dataset

Our dataset comes from Kaggle.

(https://www.kaggle.com/datasets/volpatto/coffee-quality-database-from-cqi?resource=download)

```
RangeIndex: 1310 entries, 0 to 1309
Data columns (total 44 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Unnamed: 0              1310 non-null   int64
 1   Species                 1310 non-null   object
 2   Owner                   1303 non-null   object
 3   Country.of.Origin       1310 non-null   object
 4   Farm.Name               955 non-null    object
 5   Lot.Number              270 non-null    object
 6   Mill                    1001 non-null   object
 7   ICO.Number              1164 non-null   object
 8   Company                 1102 non-null   object
 9   Altitude                1088 non-null   object
 10  Region                  1254 non-null   object
 11  Producer                1081 non-null   object
 12  Number.of.Bags          1310 non-null   int64
 13  Bag.Weight              1310 non-null   object
 14  In.Country.Partner      1310 non-null   object
 15  Harvest.Year            1264 non-null   object
 16  Grading.Date            1310 non-null   object
 17  Owner.1                 1303 non-null   object
 18  Variety                 1110 non-null   object
 19  Processing.Method       1159 non-null   object
 20  Aroma                   1310 non-null   float64
 21  Flavor                  1310 non-null   float64
 22  Aftertaste              1310 non-null   float64
 23  Acidity                 1310 non-null   float64
 24  Body                    1310 non-null   float64
 25  Balance                 1310 non-null   float64
 26  Uniformity              1310 non-null   float64
 27  Clean.Cup               1310 non-null   float64
 28  Sweetness               1310 non-null   float64
 29  Cupper.Points           1310 non-null   float64
 30  Total.Cup.Points        1310 non-null   float64
 31  Moisture                1310 non-null   float64
 32  Category.One.Defects    1310 non-null   int64
 33  Quakers                 1309 non-null   float64
 34  Color                   1095 non-null   object
 35  Category.Two.Defects    1310 non-null   int64
 36  Expiration              1310 non-null   object
 37  Certification.Body      1310 non-null   object
 38  Certification.Address   1310 non-null   object
 39  Certification.Contact   1310 non-null   object
 40  unit_of_measurement     1310 non-null   object
 41  altitude_low_meters     1084 non-null   float64
 42  altitude_high_meters    1084 non-null   float64
 43  altitude_mean_meters    1084 non-null   float64
dtypes: float64(16), int64(4), object(24)
memory usage: 450.4+ KB
```

The original data set has a total of 1310 records and 44 columns. Columns 0 to 19 are some basic information of coffee beans, such as country of origin, harvest year, processing method, type, etc. Columns 20 to 31 are where the grading agencies rate the quality of the coffee beans. Scores are made from the aroma released after the coffee beans are ground, and the taste, taste, acidity, and sweetness after brewing. Among them, body refers to the thickness, consistency and texture of coffee in taste; balance refers to the balance and coordination among various taste elements (such as sweetness, acidity, bitterness, etc.) in coffee; Clean Cup refers to The degree to which the coffee is free from impurities or off-flavors. Columns 32 to 36 are the grades of unground coffee beans; columns 37 to 39 are the information of the rating agencies; and the last 4 columns are the altitudes where the coffee beans are grown.

2.1 Data Preprocessing

Due to the large number of data columns, some attributes are irrelevant to the research issues of this report, so we delete them and only keep useful attributes. From Figure 1 we can see that there are many missing values in the original dataset. In order to prevent errors during the analysis, we performed the following padding on the data before starting to analyze the data:

- The minimum/maximum/average altitude of the coffee bean plantation we choose to use

the average value for filling

- Coffee bean processing method, coffee bean harvest year and Quakers take mode to fill
- The color of coffee beans is filled with 'None'
- The type of coffee beans we fill with 'Other'
- The Region of coffee beans is filled with 'Unknown'

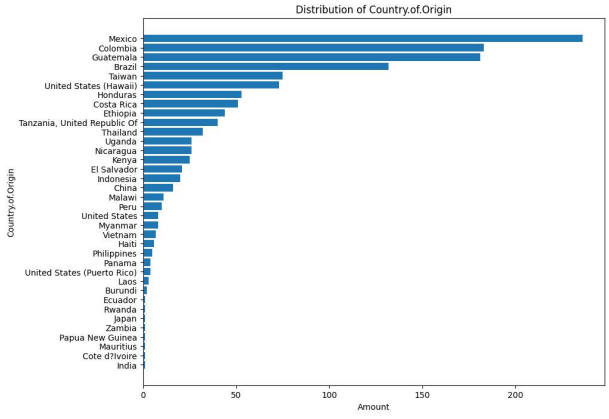
After the data is populated, we start data analysis and visualization.

3.Data Analysis

Next, we conduct data analysis from five aspects, namely, the basic information of coffee beans, such as place of origin, type and color; information related to coffee bean origin and altitude; coffee bean processing methods; rating levels.

3.1 Basic information about coffee beans

3.1.1 National coffee production

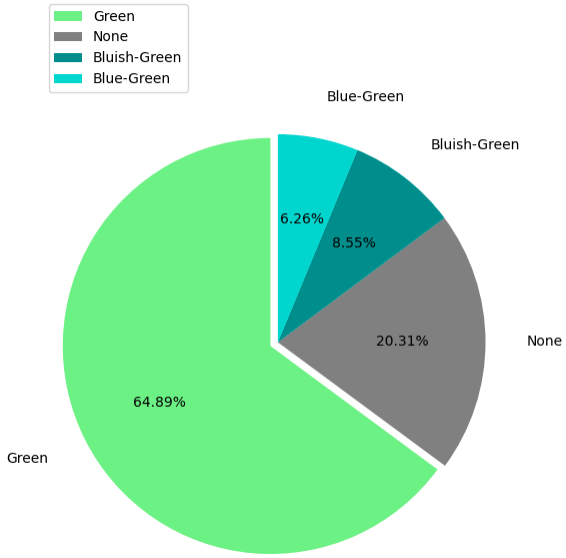


The coffee beans in the dataset come from 36 countries. Among them, 236 pieces of coffee bean data come from Mexico, which accounts for the largest proportion. The number of coffee beans from Colombia and Guatemala ranked second and third, respectively 183 and 181.

This is mainly because Arabica coffee beans like a shaded environment, and most coffee in Mexico is grown in the shade of trees. At the same time, most of the coffee producing areas in Mexico are located at an altitude of about 900 meters. The climate here is relatively cool, which

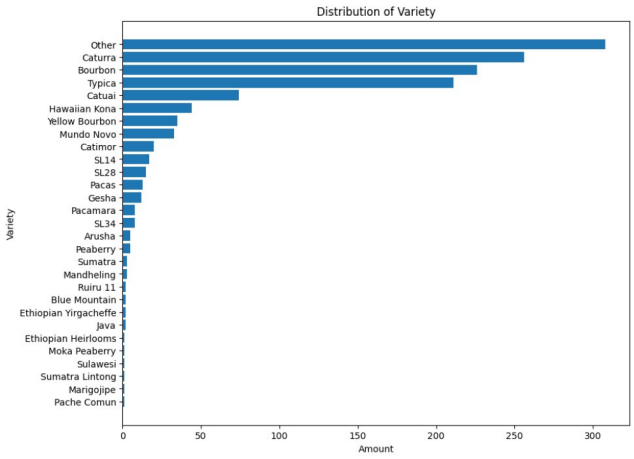
is conducive to improving the quality of coffee beans. Second, Mexico's coffee production costs are lower and its quality is higher, favored by foreign roasters.

3.1.2 Color Distribution



In this data sample, 20.31% of the coffee beans are not marked with color, and the remaining 79.69% of the coffee beans are divided into three colors, namely green, blue-green and bluish-blue. Among them, green accounts for 64.89%, blue-green accounts for 6.26%, and bluish-blue accounts for 8.55%.

3.1.3 Distribution of coffee beans



In the data set, there are a total of 19 different varieties of Arabica coffee beans. In addition to the 308 varieties of other varieties that have not been counted, the most common varieties are Caturra, 256, Bourbon, 226 and Typica, 211. Caturra has good disease resistance and strong adaptability. It does not need shade trees and can adapt to high-density planting. It can also be vigorous under direct exposure to the sun. And the tree is relatively short, which is convenient for harvesting, and the harvesting cost is low. It is the best choice for coffee farmers. The main reason for the high yield of Typica is that its flavor advantage is

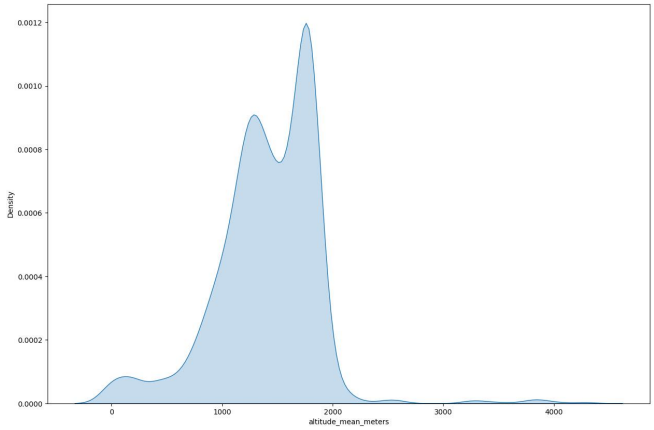
widely sought after by coffee lovers. The flavor of Typica is sweet and clean, with floral, fruity and complex layers.

3.2 Information about coffee bean origin and altitude

3.2.1 World heat map of coffee production

In the world thermodynamic map of coffee, coffee is distributed along the north and south sides of the equator, and most of them are concentrated in the area between north and south latitudes of 25 degrees, Africa, Central America, South America, Asia and some island countries. According to the altitude, climate and soil of each production area, different styles of coffee will be bred.

3.2.2 Altitude Density Map

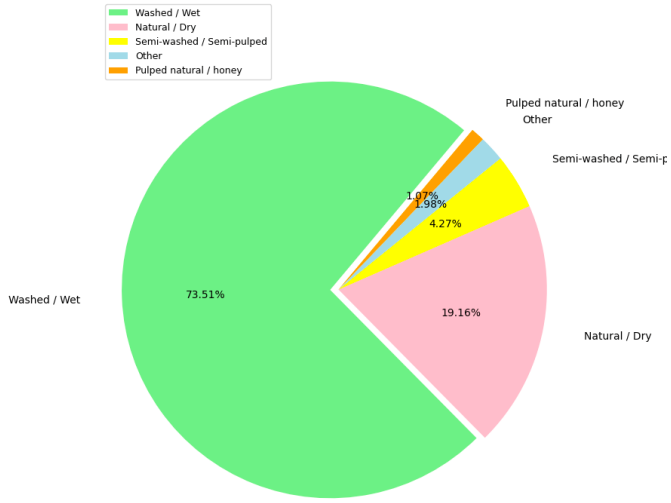


The picture above shows the planting altitudes of all the coffee bean data in the dataset. From the picture, we can easily find that the planting altitudes of most of the data are distributed between 1000 and 2000 meters. Because the temperature in this altitude range is suitable and the sunshine is sufficient, it is conducive to the growth of coffee trees and the maturity of coffee beans. Not

only that, the rainfall in this altitude area is abundant, the soil is suitable, and the soil nutrients are sufficient, which can provide sufficient nutrition for coffee trees and improve the quality of coffee beans.

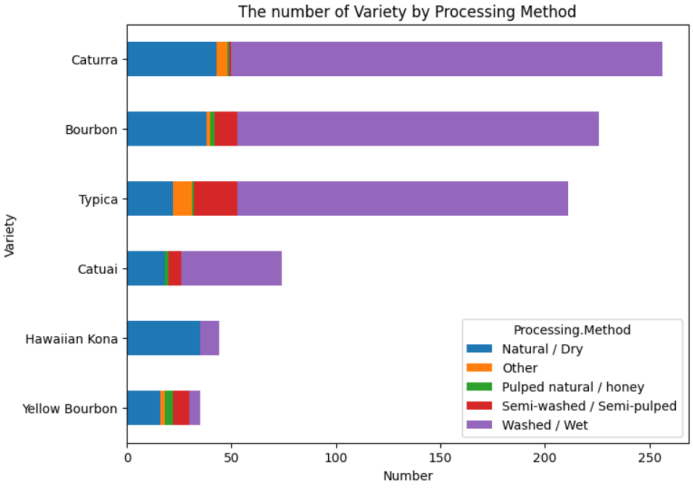
3.3 Processing method

3.3.1 Processing method distribution



There are 5 processing methods for coffee bean samples. Most of the coffee beans are processed by the Washed/Wet method, and the rest are processed by Natural/Dry, Semi-washed/Semi-pulped and Pulped natural/Honey. to process. The coffee beans processed by washing with water have less impurities and a more complete appearance, and since the pulp in the coffee fruit has been removed from the beginning, there is no need to worry about moldy problems, and the overall quality is relatively stable.

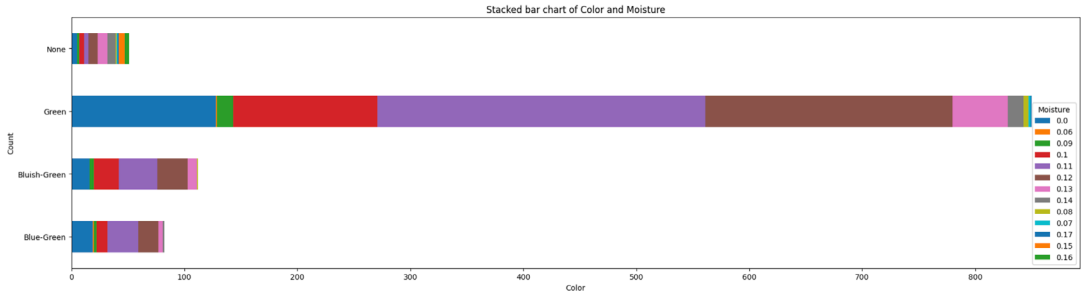
3.3.2 A stacked bar chart showing how different types of coffee beans are processed



As can be seen from the figure, the main processing methods of Caturra, Bourbon, Typica, and Catuai are washed/wet, followed by natural/dry. The person in charge of the Yunnan Pu'er coffee

production area said that there are mainly three processing methods for coffee beans: washed, sun-dried and honey-processed. Only washed is the most able to control the quality, and the other two may have different qualities depending on the farmer. It is difficult to control, which is also quite consistent with the fact that the first four in this chart are mostly washed.

3.3.3 Coffee bean color and humidity stacking chart

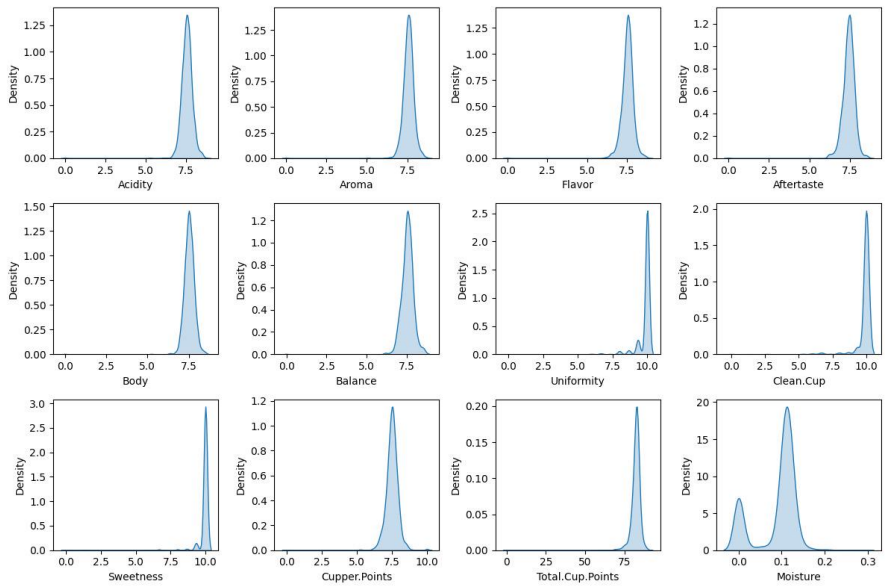


Different colors represent different humidity values, and the stacked bars under each color are stacked according to different humidity levels of the same color, allowing us to compare the quantitative relationship between the humidity levels of different colors. The legend below the graph shows the humidity represented by each color level and its corresponding color.

3.4 Ratings of coffee beans

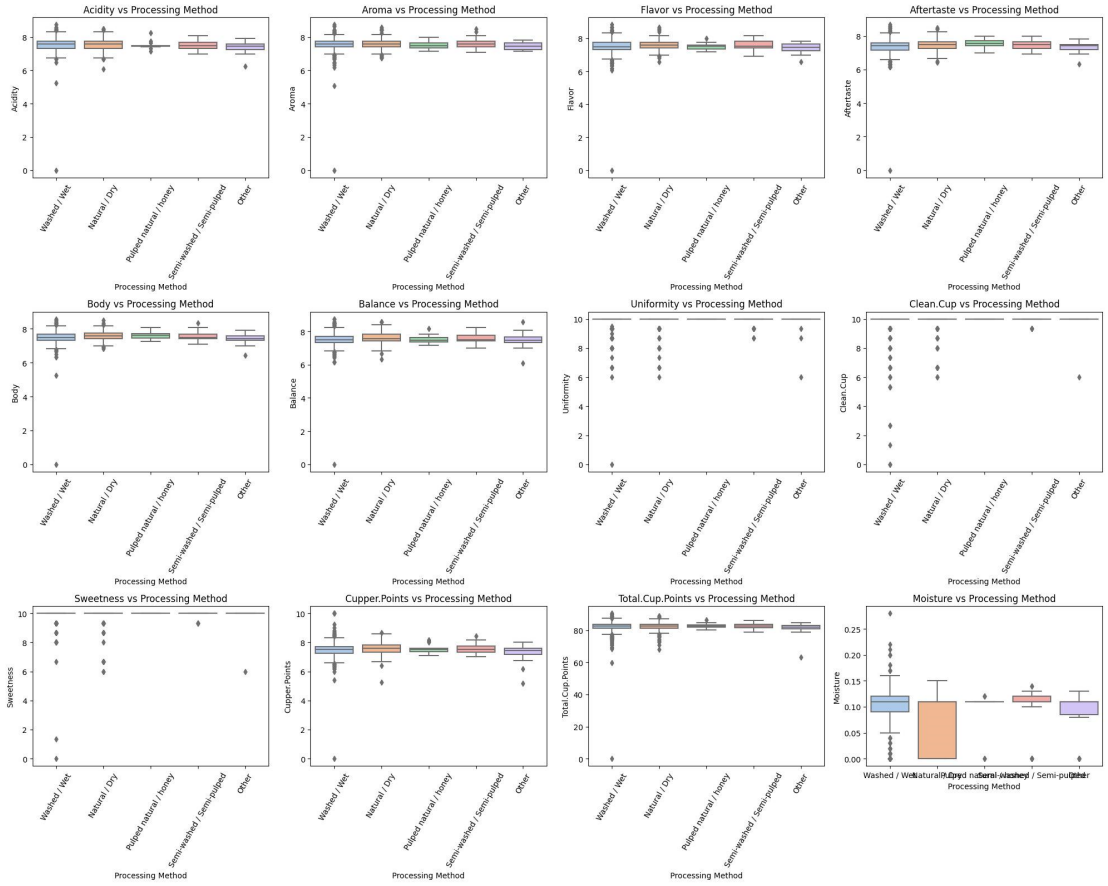
After studying the basic information of coffee beans in the data set, we began to further explore the scoring of coffee beans. In this data, the scoring dimensions of coffee beans have the following 12 dimensions: Acidity, Aroma, Flavor, Aftertaste, Body, Balance, Uniformity, Clean.Cup, Sweetness, Cupper.Points, Total.Cup.Points, Moisture.

3.4.1 Individual rating of coffee beans



The individual score distribution chart of coffee beans is shown in the figure above. We found that the scores of Acidity, Aroma, Flavor, Aftertaste, Body, Balance and Cupper.Points are generally distributed between 6 and 8 points. The scores of Sweetness and Uniformity are around 10 points. It proves that Arabica has a sweet taste, and the sweetness usually comes from the natural sugar and aroma substances in coffee. Uniformity refers to the consistency and stability of a batch of coffee beans in quality. From the data in the above figure, we can see that the taste and quality are relatively stable. In the end, we found that the moisture data of coffee beans is between 0-0.2. By searching the data, we found that the humidity in this range can guarantee the quality and storage time of coffee beans.

3.4.2 The relationship between coffee bean treatment and 12 scoring dimensions



The relationship between the Processing Method of coffee beans and the other 12 dimensions is shown in the box plot above. We use Processing Method as the x-axis variable, and other continuous variables as the y-axis variable respectively, and use the same custom color board (palette) for drawing. As you can see from the figure, the relationship between Processing Method and different dimensions is different.

For example, coffee processed with the Honey method generally has a higher Acidity score, while coffee processed with the Natural method generally has a higher Aroma score. A total of 12 boxplots were drawn, each showing the relationship between the Processing Method and different features. The following is a detailed analysis of each picture:

Acidity vs Processing Method: It can be seen that different Processing Methods have a significant impact on the Acidity score. This graph shows that the Natural method coffee beans are the best in terms of Acidity score, rather than the most common Fully Washed method. Therefore, when we choose coffee, the Acidity feature can help us determine which Processing Method is best for our taste.

Aroma vs Processing Method: Similar to Acidity, the impact of different Processing Methods on the Aroma score is also very significant. This graph shows that the coffee beans produced by the Honey method are the best in terms of Aroma score. So when choosing a coffee, the Aroma profile can help us determine the roast treatment we want.

Flavor vs Processing Method: Similar to the previous content, different Processing Methods also have a great impact on the Flavor score. From this figure, it can be seen that the Natural method has the highest score in Flavor. Therefore, when choosing coffee, Flavor characteristics can help us better understand which coffee beans are most suitable for our taste.

Aftertaste vs Processing Method: Different Processing Methods have a great impact on Aftertaste scores. From this figure, it can be seen that the Honey treatment method has the highest score, while the Washed method has the lowest score.

Body vs Processing Method: This picture is different from the previous visualization. The difference in influence between different Processing Methods is not obvious, but the difference between the Honey method and other methods is relatively large. Therefore, when choosing coffee, Body characteristics may not be the main consideration that affects our choice.

Balance vs Processing Method: This graph shows that Honey method coffee beans are the best in terms of Balance score, while other methods generally score relatively low.

Uniformity vs Processing Method: Although the impact of different Processing Methods on Uniformity characteristics is not very significant, the coffee beans of the Honey method have

the highest score in Uniformity. Therefore, Uniformity characteristics may not be the main consideration when choosing a coffee.

Clean.Cup vs Processing Method: The differences between the various Processing Methods are not too obvious. The Natural and Semi-washed methods scored relatively high for Clean.Cup, while the Washed method scored relatively low. Therefore, the Clean.Cup feature may not be the main consideration when choosing a coffee.

Sweetness vs Processing Method: As can be seen, the Natural method generally produces coffee beans with high Sweetness scores. So when we choose a coffee, the Sweetness profile can help us better understand which roasting process is best for our tastes.

Cupper.Points vs Processing Method: As you can see from this graph, there is not much difference between the categories, but beans with a Natural processing method generally score higher. Therefore, when choosing coffee, the Cupper.Points feature may affect our choice of which coffee beans to roast and process.

Total.Cup.Points vs Processing Method: As can be seen from this graph, different Processing Methods have a great impact on the Total.Cup.Points score, and the coffee beans produced by the Natural method usually have the highest score. Therefore, when choosing a coffee, the Total.Cup.Points feature can help us understand which processing method is best for our taste.

Summary: From these visual boxplots, it can be seen that there is a certain relationship between Processing Method and Acidity, Aroma, Flavor, Aftertaste, Body, Sweetness, Cupper.Points and Total.Cup.Points. For example, in the Acidity boxplot, the coffee corresponding to the Natural Processing Method will have a higher Acidity value than other Processing Methods. Therefore, these boxplots clearly show that Processing Method has an impact on these variables.

The common trend among them is that for each y-axis variable, the Natural treatment coffee generally has mostly the highest values, while the Washed treatment coffee generally has the

lowest observed values. Specifically, different Processing Methods lead to differences in the distribution of variables, which need to be taken into account during modeling or analysis. It is worth noting that in some boxplots, such as Cleaning Cup and Moisture, Processing Method does not seem to have a significant effect on these variables.

3.5 Coffee bean rating analysis

3.5.1 Rating standard

```
[ ] # Professional barista tasting: filter for values greater than or equal to 7
    df1= df[df['Cupper.Points'] >= 7.0].dropna(how='any')
    df1
```

```
[ ] # The balance between the various elements of coffee: filter the value greater than or equal to 7
    df2 = df1[df1['Balance'] >= 7.0].dropna(how='any')
    df2
```

```
[ ] # Clean.Cup does not contain impurities, cleanliness and taste: filter values greater than or equal to 9
    df3 = df2[df2['Clean.Cup'] >= 9.0].dropna(how='any')
    df3
```

```
[ ] # Stability of Uniformity coffee: filter for values greater than or equal to 8.5
    df_bestCoffee = df3[df3['Uniformity'] >= 8.5].dropna(how='any')
    df_bestCoffee
```

```
[ ] # select three columns
    df_selected = df_bestCoffee[['Cupper.Points', 'Balance', 'Clean.Cup']]

    # Assign weights respectively
    weights = [0.6, 0.2, 0.2]

    # Calculate the weighted sum and assign it to the new column
    df_bestCoffee['rating'] = (df_selected * weights).sum(axis=1)

    df_bestCoffee = df_bestCoffee.sort_values(by='rating', ascending=False)

    df_bestCoffee
```

After processing and visualizing the above data, we found that the data distribution is unbalanced. In order to make the scoring and analysis of coffee beans more scientific and valuable, we weighted and assigned a new attribute called rating. For the filling data of this column, after analyzing and integrating the original dataset, we selected 3 columns of comprehensive scores in the original data, namely 'Cupper.Points', 'Balance' and 'Clean.Cup'. Firstly, we analyzed the importance of these three variables, and decided to use the column 'Cupper. Through the analysis of this column of data, it is decided to filter out the coffee beans

with a score greater than or equal to 7 points, and obtain 1208 pieces of data; secondly, the 'Balance' column representing "the balance between the various elements of coffee" is the first Secondary screening, after analysis, it is decided to screen coffee beans greater than or equal to 7 in this column on the basis of primary screening, and obtain 1198 pieces of data; in the third step, we will represent "the degree of coffee without impurities, cleanliness and taste The 'Clean.Cup' column of "is used as the third level of screening. After verification, the higher the cleanliness of the coffee, the better the taste. After analyzing the data in this column, it was decided to filter the value greater than or equal to 9, and 1172 pieces of data were obtained. ;Finally, after the first three steps of screening, the 'Uniformity' column representing "coffee stability" is used as the final screening. The higher the stability of the coffee, the better the taste. Combined with our data, the selected stability is greater than or equal to 8.5 coffee beans form a new dataframe containing 1164 pieces of data that meet the above conditions.

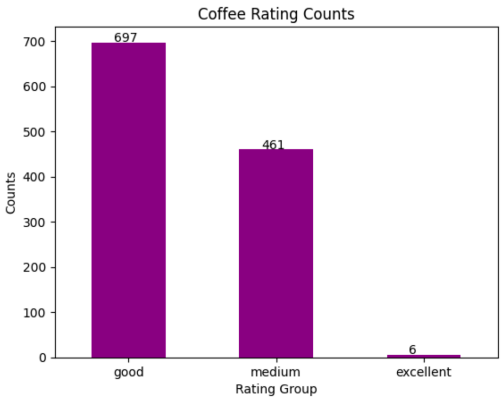
3.5.2 Quantity distribution of coffee beans with different grades

```
bins = [0, 7.999, 8.999, 10]
labels = ['medium', 'good', 'excellent']
df_bestCoffee['rating_group'] = pd.cut(df_bestCoffee['rating'], bins, labels=labels)
df_bestCoffee
```

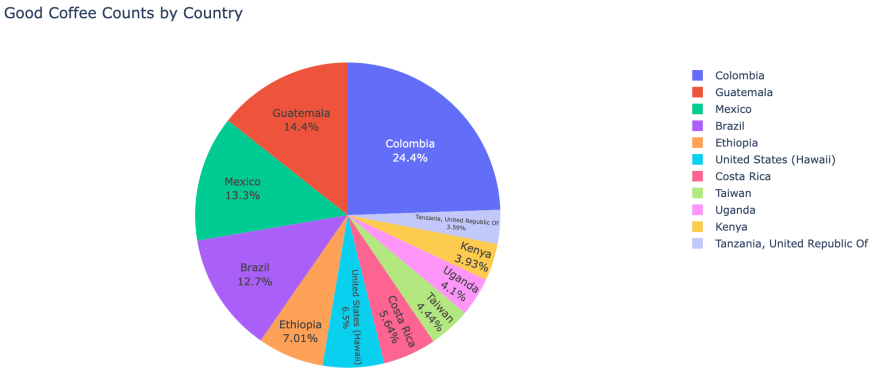
```
import matplotlib.pyplot as plt
%matplotlib inline

# 计算各个分组的数量
rating_counts = df_bestCoffee['rating_group'].value_counts()

# 设置图形
fig, ax = plt.subplots()
ax = rating_counts.plot.bar(rot=0, color='purple')

# 添加标签
for i, v in enumerate(rating_counts):
    ax.text(i - 0.1, v + 1, str(v), color='black', fontsize=10)

# 添加标题和标签
plt.title('Coffee Rating Counts')
plt.xlabel('Rating Group')
plt.ylabel('Counts')
```

After screening and processing the data, a more representative rating column is obtained. Through the analysis of the score distribution of the rating column, it is found that the distribution range is between 7.48 and 9.4 points. In order to show the relationship between the coffee data more intuitively, we decided to classify the coffee beans into 'medium', 'good', 'excellent' grades, so as to observe the coffee beans of each grade in different Regional distribution of production and so on. In the classification, we classify 'medium', 'good', and 'excellent' with scores less than or equal to 8 points, greater than 8 points

and less than or equal to 9 points, and scores greater than 9 and less than or equal to 10 points to obtain the 'medium' grade There are 461 pieces of coffee bean data, 697 pieces of 'good' grade coffee bean data, and 6 pieces of 'excellent' grade coffee bean data. The following bar chart statistics are made through this data, and based on this processing, the The following data visualization analysis.



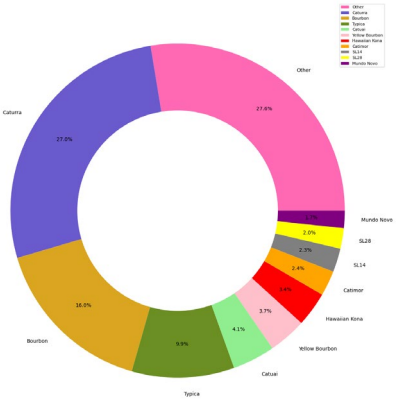
3.5.3 The distribution of high-quality coffee beans in different countries



First of all, after empowerment and marking, our data is unbalanced, and the number of excellent is too small to be referenced, so we use good as a reference standard to check the number of coffee beans with a good label in each country. Finally, the top ten countries with high-quality coffee bean production are selected. Through preliminary data processing, the top ten countries in coffee production are: Colombia, Guatemala, Mexico, Brazil, Ethiopia, and Hawaii in the United States. The coffee industry in these countries or regions is larger in scale

and has stronger production capacity. The production of coffee is also closely related to the country's geographical location, climate and other factors. For example, Arabica coffee grown in the tropics is generally considered to be of higher quality, while Robusta coffee grown in the subtropics has higher caffeine content and a darker, bitter, piney flavor. Therefore, the varieties and qualities of coffee produced under different geographical locations and climatic conditions are also different. On the basis of the above data processing, we further obtain a pie chart about the "good" quality of coffee, to see the proportion of coffee quality in each country (origin) in all high-quality coffee. First, after analysis, we extracted the countries producing more than 20 "good" coffee varieties. After analysis, it is found that Colombia has the largest amount of "good" coffee, accounting for about 24.4%, followed by Guatemala (about 14.4%) and Mexico (about 13.3%); Brazil is the world's largest coffee producer, but from "good" In terms of the quantity of coffee, its proportion is only 12.7%, ranking fourth, indicating that there is a significant difference between the quality of its coffee and the total amount of coffee; in addition, although Hawaii in the United States is only a small island, it has produced about 6.5% for "good" coffee is enough to get our attention; yields have also increased in other countries, including Ethiopia, China, Thailand, etc., but are still low overall.

3.5.4 The proportion of high-quality coffee beans



From the above figure, we found that among the high-quality coffee beans, except for the data that does not specifically indicate the type of coffee beans, the largest type is Caturra, accounting for 27%. Bourbon accounted for 16%, behind Caturra. The remaining 8 varieties accounted for no more than 10%. Caturra

accounts for a large proportion because Caturra coffee trees are highly adaptable to the environment and can grow and mature in different planting environments, which makes the production of Caturra coffee beans more stable and reliable. In addition, the fruit of the Caturra coffee tree matures faster than other varieties of coffee trees, and the Caturra coffee tree can ripen the fruit faster, thereby increasing the yield of coffee beans.

4. Findings

In the process of data processing, real data often has many empty values. How to fill these empty values and in what way to fill these empty values requires rigorous analysis and investigation by the processor to make the final data more realistic. , the obtained data research results can have higher practical value.

Through analysis, we learned that the cultivation of coffee beans should not only take into account the influence of climate and geographical location, but also be related to labor costs, audience groups, and local culture. The flavor of coffee beans is influenced by many factors, including but not limited to region, temperature, altitude, and processing method. The flavor of coffee beans is not only evaluated by whether it tastes good or not, it includes multi-dimensional evaluations such as acidity, aroma, sweetness, balance, etc. Different ratios of beans will find people who like it.

Based on this data set, we hope to provide consumers with more advice and data when choosing coffee beans, and help them choose beans that match their taste. It also hopes to provide data analysis for coffee producers, coffee shops and other related enterprises.

Since the data set focuses on the exploration of coffee bean flavor factors, it is not suitable for predictive models. But we hope that this project can provide more help to coffee lovers as a popular science project..