COFFEE REVIEWS

ERWIN EDUARDO GUERRERO CAZARES

July, 2021.

CONTENTS

LIST OF FIGURES

ABSTRACT

This document presents a data analysis about the dataset in question complemented with a model of **Multiple linear regression**. For this we will take the dependent variable that, according to experts, is the most important thing to assess when we talk about coffee. On the other hand, the data analysis begins with understanding the dataset. In this first part, an attempt is made to answer the questions posed previously. Finally, we move on to the second part where we build the previously mentioned model.

1 INTRODUCTION

When we talk about coffee it is impossible not to talk about caffeine. This leads us to talk about the famous addiction to coffee or caffeine. Although one cannot speak precisely of an addiction, caffeine is definitely a psychoactive drug. According to the United States National Library of Medicine, caffeine is a bitter substance found in coffee, tea, soft drinks, chocolate, kola nuts, and certain medicines. It has many effects on the body's metabolism, including stimulating the central nervous system.

On the other hand, coffee has become of vital importance in the world. Interestingly, it also ranks second after oil in terms of international trade figures. In addition, it is estimated that 3 billion cups of coffee are drunk daily, in the world. And we could spend hours talking about coffee, however, our responsibility is to carry out an analysis on the reviews that some experts wrote.

The purpose of this data analysis is to synthesize the information and obtain faster and more concise conclusions. An approach will be taken to the characteristics defined in the dataset forgetting other variables. Obviously, in this case the analysis is, to a certain extent, subjective, since the graphics, the analysis approach and even the design of this document itself are free in the indications. So it can be done in many ways.

Now, in the case of the chosen model, the purpose has to do with the question, **what is the most important characteristic of coffee to conclude that it is a great coffee?** For this we will take flavor as a dependent variable. This is because it is what people stand out the most when they talk about a good coffee. Actually, all the other characteristics are taken into account by a coffee expert, and in this specific case we are going to assume that we do not know anything about coffee and we want to answer the previous question. Although, it is recommended that the model be implemented taking the average of the scores as the dependent variable

2 THE DATASET

The data to be used contains bean, farm, and quality data from reviews of arabica coffee beans were extracted from the Coffee Quality Institute's trained reviewers in January 2018. According to its website, *Coffee Quality Institute (CQI) is a non-profit organization working internationally to improve the quality of coffee and the lives of people who produce it*.

You can download it by entering the following link dataset.

2.1 Data description

The data consists of 43 variables, 24 of which are categorical and 19 numerical. Of the latter, 7 variables describe the quality of the coffee. In the case of categorical, several stand out. However, many columns have multiple null values and many different values. What makes the analysis a bit difficult in general, both to show graphs and to make conclusions.

2.1.1 Numerical variables

Name	Description
Aroma	Aroma value
Flavor	Flavor value
Altitude	Altitude of the Farm
Aftertaste	Aftertaste value
Acidity	Acidity value
Body	Body value
Balance	Balance value
Uniformity	Uniformity
Clean.Cup	Clean cup
Sweetness	Sweetness value
Cupper.Points	Cupper points value
Total.Cup.Points	Total cup points value
Moisture	Moisture value
Category.One.Defects	Category one defects value
Quakers	Quakers value
Category.Two.Defects	Category two defects value
altitude low meters	Altitude low meters farm
altitude high meters	Altitude high meters farm
altitude mean meters	Altitude mean meters farm

Here is a brief explanation of the characteristics of coffee that are present in the dataset.

Aroma. The aroma of coffees is the product of a complex mixture of volatile compounds in the infusion. About 800 compounds have been identified that can affect aroma, including sulfur compounds.

Flavor. It is a general evaluation of coffee and is often measured with reference to a flavor table or what is known as the "coffee flavor wheel”. The flavor is a term that includes all the parameters of the coffee infusion.

Acidity. Acidity is the clear, dry flavor that brings a cup of coffee to life. What is perceived as acidity does not necessarily correspond to the pH of a coffee, but it is considered to be the result of the acids present in coffee

Aftertaste. The aftertaste is the sensation you feel after swallowing the coffee. The tasters evaluate the permanence of the aftertaste, that is, the time it takes from the initial aromatic sensation inside the throat until that sensation ceases.

Body. The body is the weight of the coffee that can be felt by letting the coffee rest on the tongue and then drizzling the tongue towards the palate.

Balance.This is known as the balance of the different aspects of the flavor, aftertaste, acidity and body of the coffee in its complementation.

These values are scored from 0 to 10, with 10 being the best score.

2.1.2 *Categorical variables*

Name	Description
Species	Coffe bean type
Owner	Farm owner
Country.of.Origin	Country of origin
Farm.Name	Farm name
Lot.Number	Lot number
Mill	Mill
ICO.Number	ICO Number
Company	Company name
Region	Region name
Producer	Producer name
Bag.Weight	Bag weight
In.Country.Partner	Country Partner
Harvest.Year	Harvest year
Owner.1	Company Owner
Variety	Coffee variety
Processing.Method	Processing method
Color	Bean color
Expiration	Harvest expiration
Certification.Body	Certification body
Certification.Address	Certification address
Certification.Contact	Certification contact
unit of measurement	Measure

It is important to note that in the following analysis not all the variables involved will be used, since some of them do not present valuable information.

3 ANALYSIS PLAN

This is a problem that involves a multiple linear regression model. For this, we will define flavor as a dependent variable. With this, it is proposed to define the most important factors that affect the taste of coffee (previously defined in the previous section).
On the other hand, in the data analysis we will start with the categorical variables and end with the numerical ones. In addition to that, we will solve the following two questions:

1. Which country produces the most coffee?
2. Taking into account the parameters defined in the data, who produces the best coffee?

4 EXPLORATORY DATA ANALYSIS

4.1 Categorical analysis

Now, we start with the data analysis by evaluating the dataset.

```python
# Import modules needed exploratory analysis
import pandas as pd
from pandas_profiling import ProfileReport # create data dashboards
from pandas_summary import DataFrameSummary
import seaborn as sns
import matplotlib.pyplot as plt
pd.set_option('display.max_rows',50)
```

Since it is a dataset with many columns (43) it is impossible to show a good preview, therefore, if you want to see the data quickly, enter kaggle.

```python
data=pd.read_csv("coffee.csv")
data.dtypes
```

```
Species                   object
Owner                     object
Country.of.Origin         object
Farm.Name                 object
Lot.Number                object
Mill                      object
ICO.Number                object
Company                   object
Altitude                  object
Region                    object
Producer                  object
Number.of.Bags             int64
Bag.Weight                object
In.Country.Partner        object
Harvest.Year              object
Grading.Date              object
Owner.1                   object
Variety                   object
Processing.Method         object
Aroma                    float64
Flavor                   float64
Aftertaste               float64
Acidity                  float64
Body                     float64
Balance                  float64
Uniformity               float64
Clean.Cup                float64
Sweetness                float64
Cupper.Points            float64
Total.Cup.Points         float64
Moisture                 float64
Category.One.Defects       int64
Quakers                  float64
Color                     object
Category.Two.Defects       int64
Expiration                object
Certification.Body        object
Certification.Address     object
Certification.Contact     object
unit_of_measurement       object
altitude_low_meters      float64
altitude_high_meters     float64
altitude_mean_meters     float64
dtype: object
```

Only three numeric variables are of type int, where it is interesting to note that two of them (Category.One.Defects and Category.Two.Defects) represent the type of defect present in the harvest:

1. Full black or sour bean, pod/cherry, and large or medium sticks or stones.
2. Parchment, hull/husk, broken/chipped, insect damage, partial black or sour, shell, small sticks or stones, water damage.

As we can see from the dimension, this is a medium or moderately large dataset with 1311 observations. This can limit the number of graphs to display in the analysis.

```
data.shape
```

```
(1311, 43)
```

However, we have a lot of null values in the datasert that suggests a failure when filling the database. Although the most important values (listed from Aroma to Moisture) do not have any null value.

```
data.isna().sum()
```

```
Species                   0
Owner                     7
Country.of.Origin         1
Farm.Name               356
Lot.Number             1041
Mill                    310
ICO.Number              146
Company                 209
Altitude                223
Region                   57
Producer                230
Number.of.Bags            0
Bag.Weight                0
In.Country.Partner        0
Harvest.Year             47
Grading.Date              0
Owner.1                   7
Variety                 201
Processing.Method       152
Aroma                     0
Flavor                    0
Aftertaste                0
Acidity                   0
Body                      0
Balance                   0
Uniformity                0
Clean.Cup                 0
Sweetness                 0
Cupper.Points             0
Total.Cup.Points          0
Moisture                  0
Category.One.Defects      0
Quakers                   1
Color                   216
Category.Two.Defects      0
Expiration                0
Certification.Body        0
Certification.Address     0
Certification.Contact     0
unit_of_measurement       0
altitude_low_meters     227
altitude_high_meters    227
altitude_mean_meters    227
```

On the other hand, **Caturra is the variant of Arabica coffee that most occurs in harvests**. In this sense, among the great advantages that Caturro coffee provides are its adaptability, especially to high altitude terrain, and the intense and pronounced aroma of its grain. Also, its caramel-tinged flavor has made it a consumer favorite. This is shown in the graph below.

```
fig,ax=plt.subplots(figsize=(11,8))

sns.countplot(y="Variety",data=data,capsize=0.1,palette="husl",
order=data.Variety.value_counts().index)

ax.set(ylabel="Variety",xlabel="Count")
plt.title("Variety Count",color='w')
ax.xaxis.label.set_color("w")
ax.yaxis.label.set_color("w")
ax.spines['left'].set_color('black')
ax.spines['bottom'].set_color('black')
ax.tick_params(colors='white',which='both')
for p,label in zip(ax.patches,data.Variety.value_counts()):
    ax.annotate(label,(p.get_width()+1,p.get_y()+p.get_height()/2+0.2),fontsize=11)
```

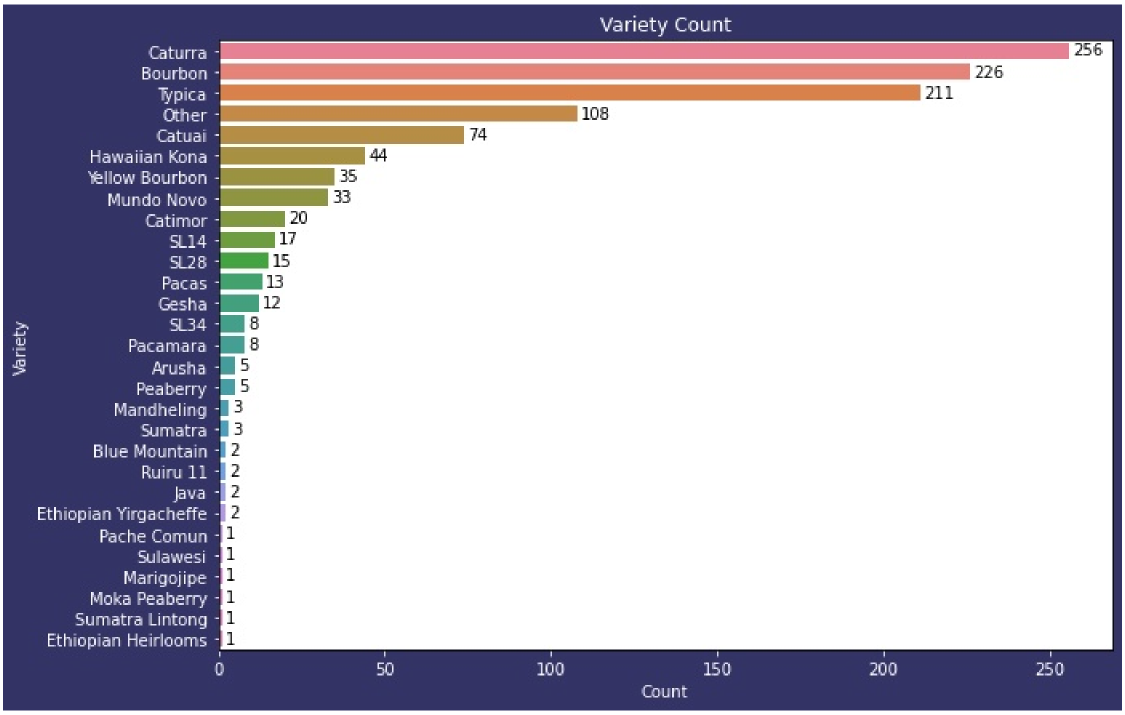


Figure 1: Variety value counts.

This can have some impact on the aroma of the coffee, as shown in the graph below. Although it can be seen how the two most common variants have more outliers than, for example, the Typica variant which has no outliers.

```
fig,ax=plt.subplots(figsize=(10,6))
datos=data[data.Variety.isin(['Caturra','Bourbon','Typica','Other','Catuai'])]
sns.boxplot(x="Variety", y="Aroma", data=datos)
```

```
ax.set(xlabel="Variety",ylabel="Aroma")

plt.title("Top 5 variety by aroma",color='w')
ax.xaxis.label.set_color("w")
ax.yaxis.label.set_color("w")
ax.spines['left'].set_color('black')
ax.spines['bottom'].set_color('black')
ax.tick_params(colors='white',which='both')
```

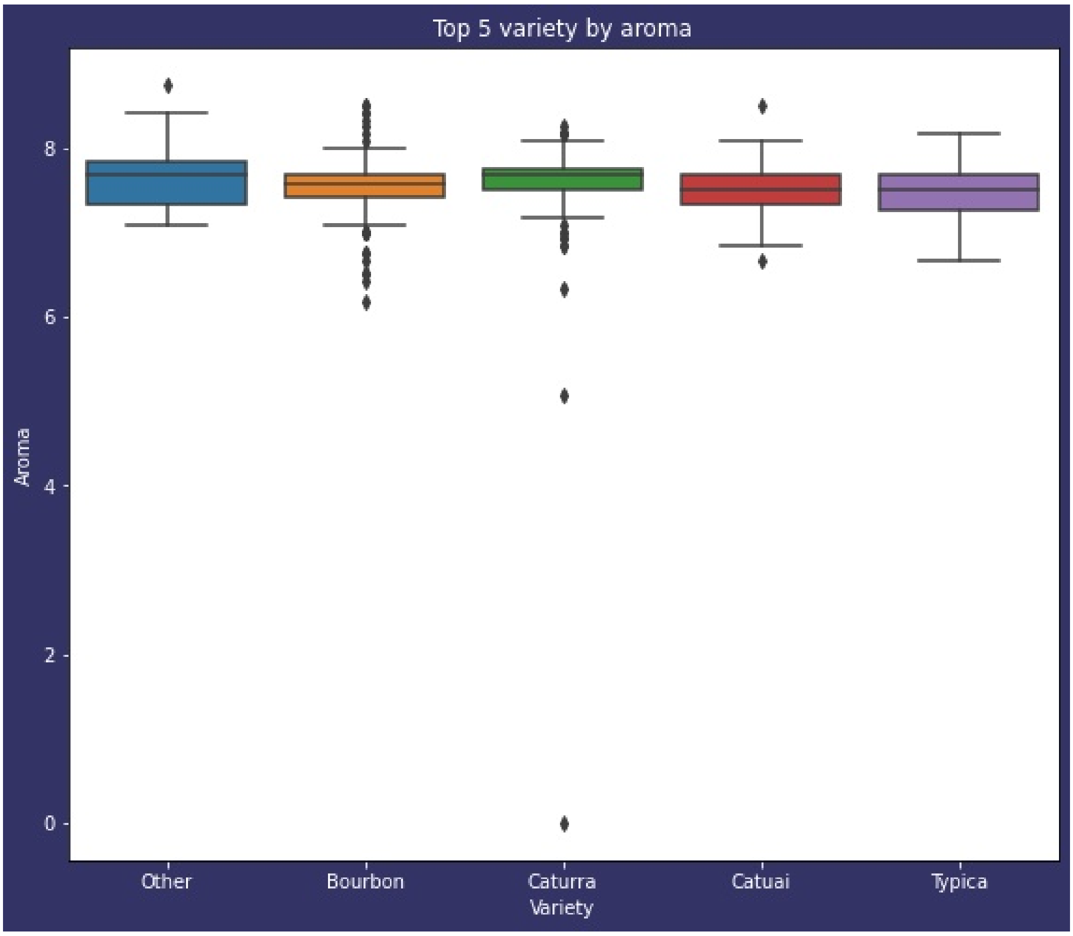


Figure 2: Boxplot of top 5 variety by aroma.

It is possible to observe how the distribution of the varieties is similar as well as the main statistics (median, percentiles, maximum and minimum not atypical). Something totally expected because the interval is small.

Coffee processing is one of the most important stages that coffee goes through before being used in any product. In addition to this, **the process of Washed / Wet was the most used in the data collected**.

```
fig,ax=plt.subplots(figsize=(8,7))

sns.countplot(x="Processing.Method",data=data,capsize=0.1,palette="husl",
order=data["Processing.Method"].value_counts().index)
ax.set(xlabel="Processing Method",ylabel="Count")
```

```
plt.title("Processing Method count",color='w')
ax.xaxis.label.set_color("w")
ax.yaxis.label.set_color("w")
ax.spines['left'].set_color('black')
ax.spines['bottom'].set_color('black')
ax.tick_params(colors='white',which='both')
for p,label in zip(ax.patches,data["Processing.Method"].value_counts()):
    ax.annotate(label,(p.get_x()+0.30,p.get_height()+6),fontsize=10)
```

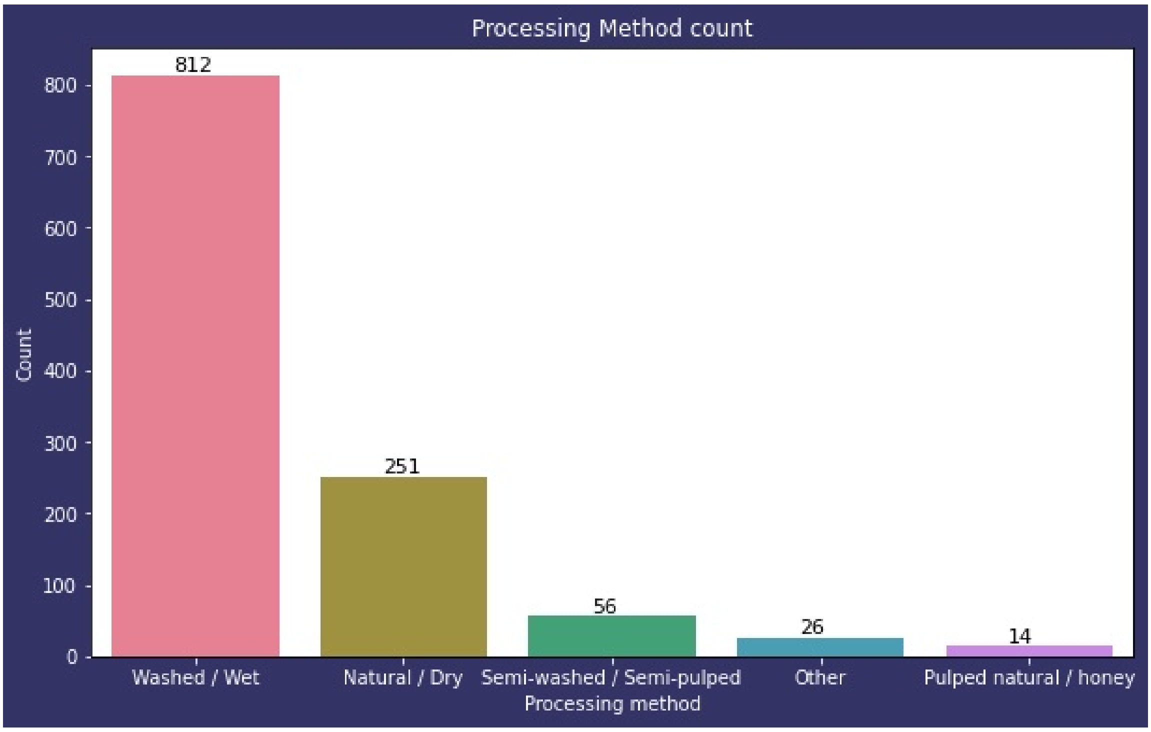


Figure 3: Processing Method value counts plot.

Furthermore, it is possible to observe that the green color predominates in all processed grains.

```
fig,ax=plt.subplots(figsize=(10,6))

sns.countplot(x="Processing.Method",data=data,capsize=0.1,palette="dark",
order=data["Processing.Method"].value_counts().index,hue="Color")
ax.set(xlabel="Processing.Method",ylabel="Count")
plt.title("Processing method by color",color='w')
ax.xaxis.label.set_color("w")
ax.yaxis.label.set_color("w")
ax.spines['left'].set_color('black')
ax.spines['bottom'].set_color('black')
ax.tick_params(colors='white',which='both')
```

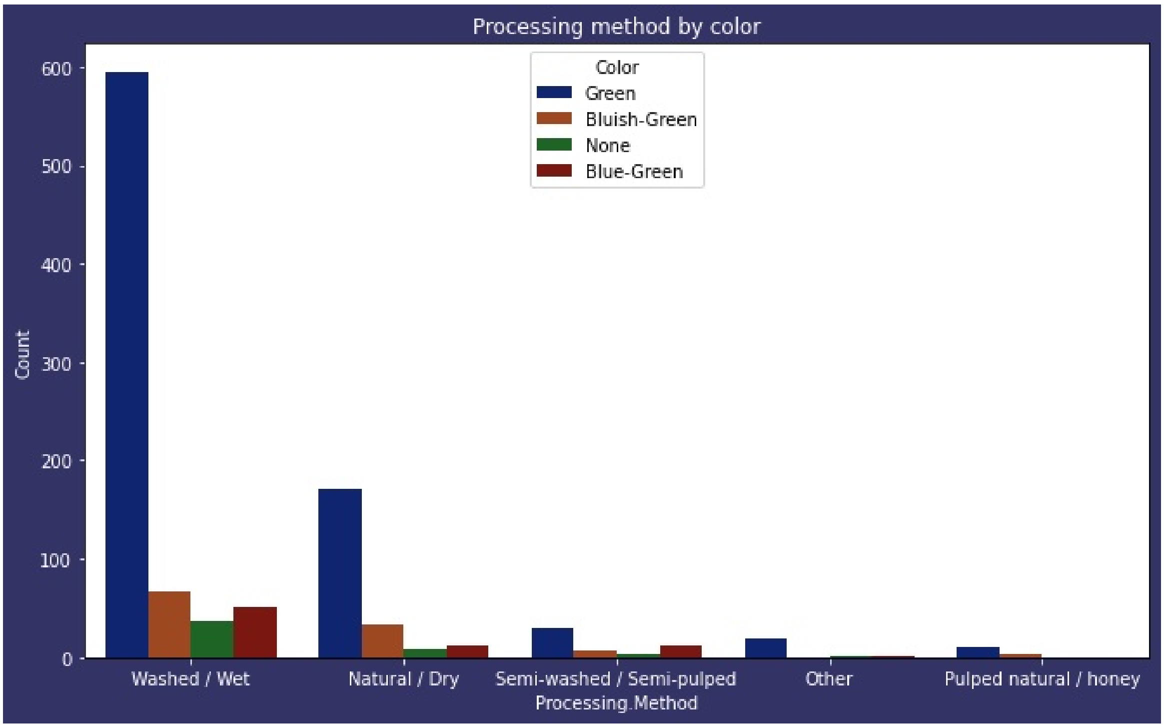


Figure 4: Processing Method graph with color hue.

Now, to answer the first question posed above, we must carefully analyze the dataset, since at first glance it seems that it is enough to do a countplot of all the countries. However, this is wrong.

```
fig,ax=plt.subplots(figsize=(11,8))

sns.countplot(y="Country.of.Origin",data=data,capsize=0.1,palette="hls",
order=data["Country.of.Origin"].value_counts().index)
ax.set(ylabel="Country of Origin",xlabel="Count")
plt.title("Country of Origin Count",color='w')
ax.xaxis.label.set_color("w")
ax.yaxis.label.set_color("w")
ax.spines['left'].set_color('black')
ax.spines['bottom'].set_color('black')
ax.tick_params(colors='white',which='both')
for p,label in zip(ax.patches,data["Country.of.Origin"].value_counts()):
    ax.annotate(label,(p.get_width()+1,p.get_y()+p.get_height()/2+0.25),fontsize=9)
```

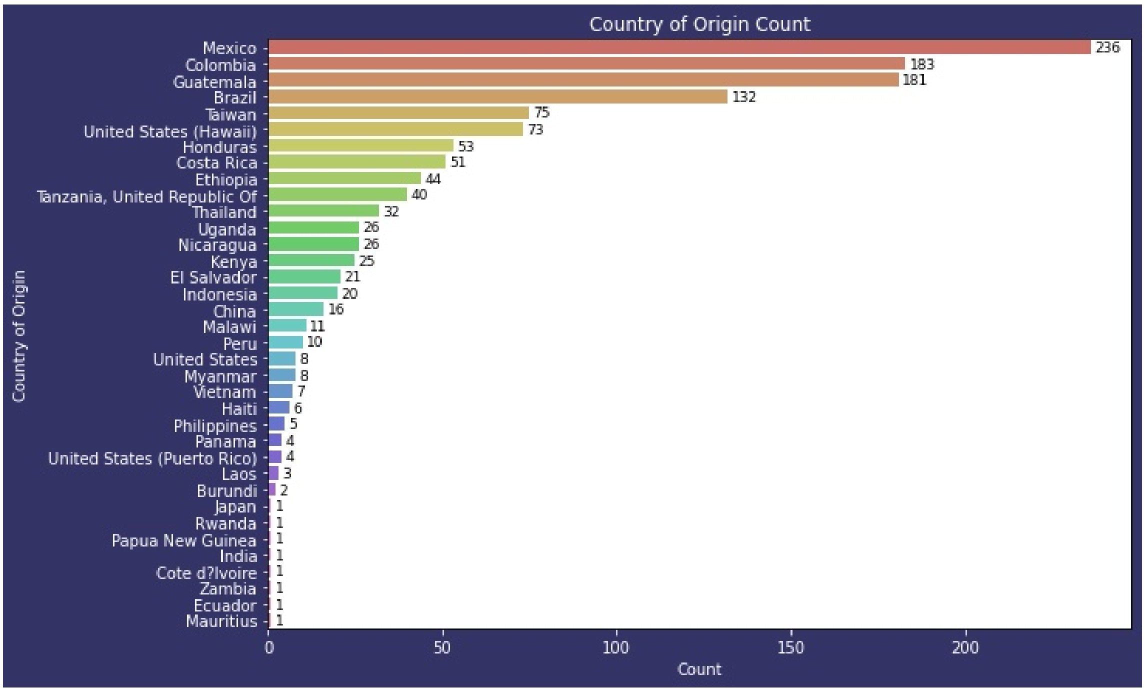


Figure 5: Country of origin count plot.

When we analyze the previous graph, we see that Mexico was the country that generated the most crops, but if we count the number of bags generated by country, we see that Mexico is in fourth place, as can be seen in the following table.

```
num_bags=data[["Country.of.Origin","Number.of.Bags"]].groupby("Country.of.Origin").sum()
num_bags.sort_values(by="Number.of.Bags",ascending=False).head()
```

Country.of.Origin	Number.of.Bags
Colombia	41204
Guatemala	36868
Brazil	30534
Mexico	24140
Honduras	13167

The above does not totally rule out that Mexico is the country that generated the most coffee, since the weight of each bag is not constant. However, as we saw previously, it is an object type data, so, we must convert it to numeric and choose a metric (since lb and kg are used). In this case, we will use kg.

```
#Convert bag weight column to numeric and lb -> kg
def fn(x):
    kg=x["Bag.Weight"].find('kg')
    lbs=x["Bag.Weight"].find('lbs')
    if kg!=-1 and lbs==-1:
            x["Bag.Weight"]=x["Bag.Weight"][:kg]
            return float(x["Bag.Weight"])
```

```
    elif kg==-1 and lbs!=-1:
        x["Bag.Weight"]=x["Bag.Weight"][:lbs]
        return float(x["Bag.Weight"])*0.453592

data["Bag.Weight"]=data.apply(fn,axis=1)

weights=data[["Country.of.Origin","Bag.Weight"]].groupby("Country.of.Origin").sum()
weights.sort_values(by="Bag.Weight",ascending=False).head()
```

Finally, taking the amount in each bag, it is concluded that **Costa Rica is the country that generated the most coffee**.

	Bag.Weight
Country.of.Origin	
Costa Rica	58674.657048
Kenya	40734.000000
Ethiopia	40001.874144
Uganda	23598.000000
Honduras	20658.000000

This suggests that Costa Rica's harvests must be big harvests as it is not even at the top of the number of bags.

In the case of grain color, we see in the following graph that green is the most present in crops. In the world of coffee, green coffee is understood to be one that has not been roasted, that is, the beans obtained after processing the coffee. However, this name is generic and does not mean that the coffee beans, once freed from the cherry and mucilage, are green in color, since their coloration will depend not only on their variety, but also on Also, in this first phase, the amount of moisture they contain and the type of beneficiation to which they are subjected.

```
fig,ax=plt.subplots(figsize=(10,6))

sns.countplot(x="Color",data=data,capsize=0.1,palette="dark",
order=data["Color"].value_counts().index)
ax.set(xlabel="Color",ylabel="Count")

plt.title("Color count",color='w')
ax.xaxis.label.set_color("w")
ax.yaxis.label.set_color("w")
ax.spines['left'].set_color('black')
ax.spines['bottom'].set_color('black')
ax.tick_params(colors='white',which='both')
for p,label in zip(ax.patches,data["Color"].value_counts()):
    ax.annotate(label,(p.get_x()+0.33,p.get_height()+6),fontsize=12)
```

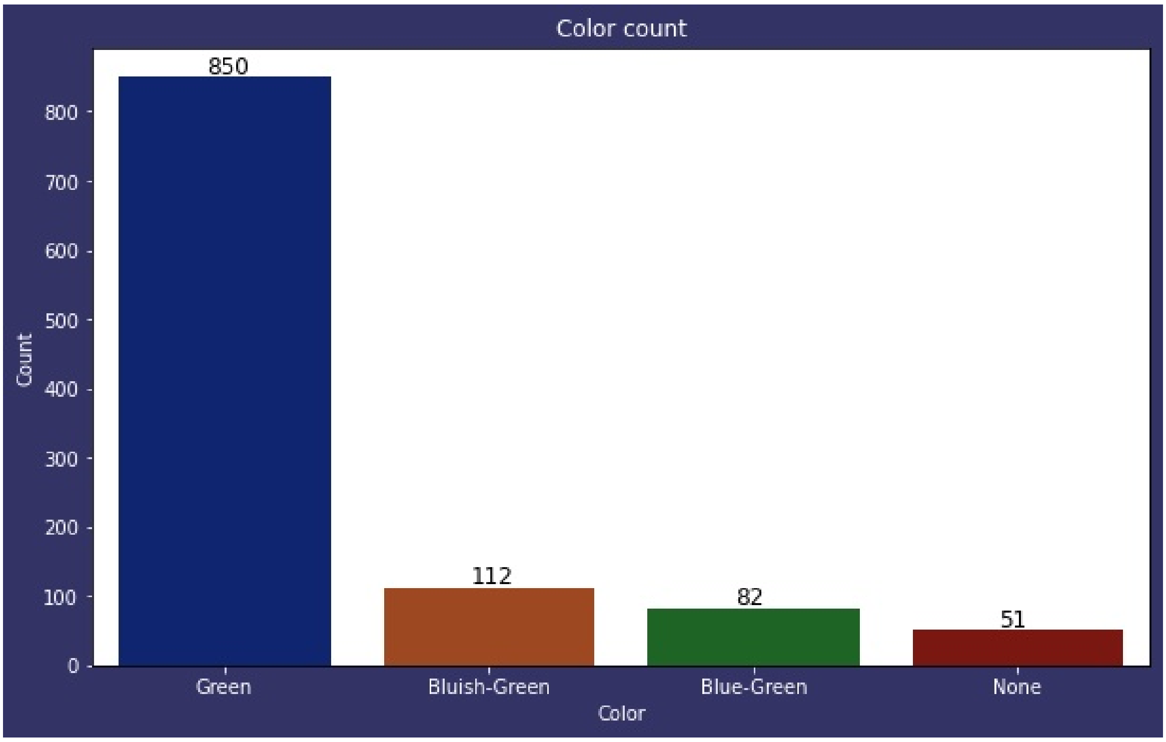


Figure 6: Color count plot.

Color seems to influence when we talk about balance, as seen below. In addition, it is possible to observe that the same behavior that was observed in **figure 2** is repeated, where they have a similar distribution and similar statistics. It could even be intuited that the same would happen with the other variables if the characteristics of coffee were analyzed by color.

```
fig,ax=plt.subplots(figsize=(11,6))

sns.boxenplot(y="Balance",x="Color",data=data,palette='bright')

ax.set(ylabel="Balance",xlabel="Color")
plt.title("Balance by color",color='w')
ax.xaxis.label.set_color("w")
ax.yaxis.label.set_color("w")
ax.spines['left'].set_color('black')
ax.spines['bottom'].set_color('black')
ax.tick_params(colors='white',which='both')
```

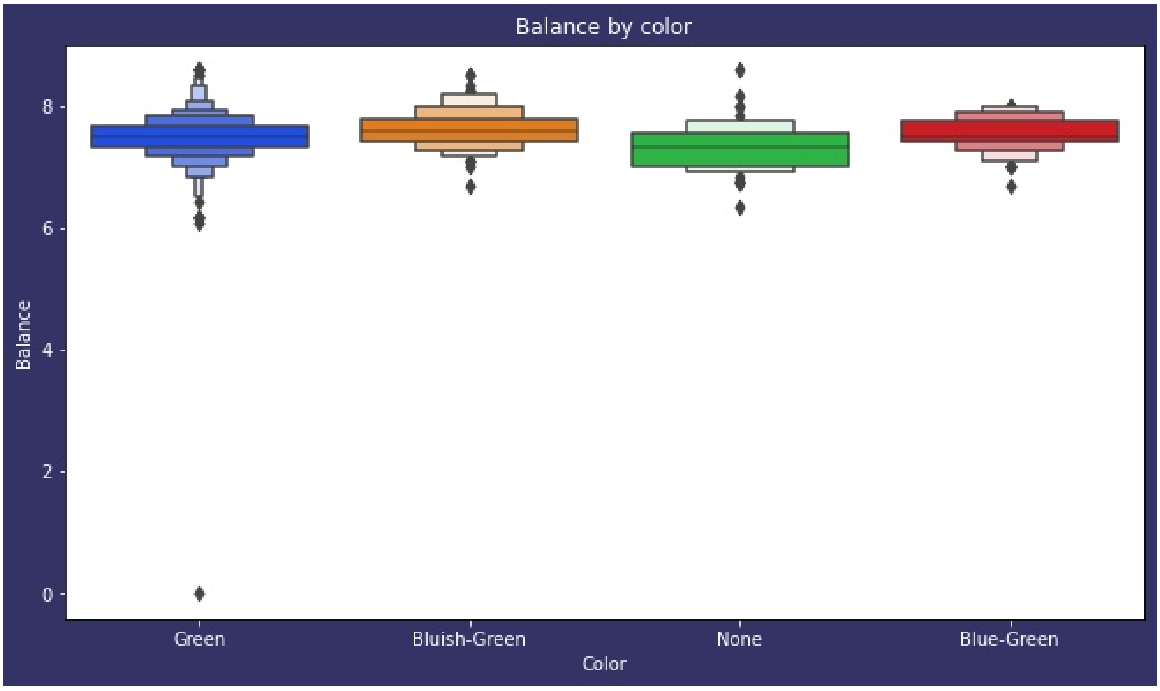


Figure 7: Balance boxen plot by color.

This can be done with each of the different characteristics, that is, color by aroma, Processing Method by flavor, Company by Body and so on. It can even be done with other variables such as the processing method, the country, among others. However, the premise of this analysis is to summarize the most important information. Even so, I made a dashboard in Power BI Coffee Analysis that allows you to view all of the above.

To finish with the analysis of the categorical variables, we will look at the time of harvest, which, unfortunately, it is not possible to perform some time series analysis or even another analysis. This is because *Harvest.year* is an object type column with unusual values as seen in the following table.

```
data[["Harvest.Year","Bag.Weight"]].groupby("Harvest.Year").sum().head(8)
```

Harvest.Year	Bag.Weight
08/09 crop	0.00000
1T/2011	70.00000
1t/2011	70.00000
2009 - 2010	0.00000
2009 / 2010	0.00000
2009-2010	130.00000
2009/2010	150.00000
2010	344.32616

However, if we order in descending order it is possible to see that 2015 was the year where the most coffee was produced.

```
years=data[["Harvest.Year","Bag.Weight"]].groupby("Harvest.Year",as_index=False).sum(). \
sort_values(by="Bag.Weight",ascending=False).head() /
years["Bag.Weight"]=round(years["Bag.Weight"],4)
```

```
fig,ax=plt.subplots(figsize=(10,4))

sns.set_palette("icefire")
sns.barplot(x="Bag.Weight",y="Harvest.Year",data=years,capsize=0.1)

ax.set(ylabel="Year",xlabel="Bag.Weight")
plt.title("Bag weight per year",color='w')
ax.xaxis.label.set_color("w")
ax.yaxis.label.set_color("w")
ax.spines['left'].set_color('black')
ax.spines['bottom'].set_color('black')
ax.tick_params(colors='white',which='both')

for p,label in zip(ax.patches,years["Bag.Weight"]):
    ax.annotate(label,(p.get_width()/2,p.get_y()+p.get_height()/2+0.15),fontsize=10,color='w')
```

Figure 8: Bag weight bar plot by year

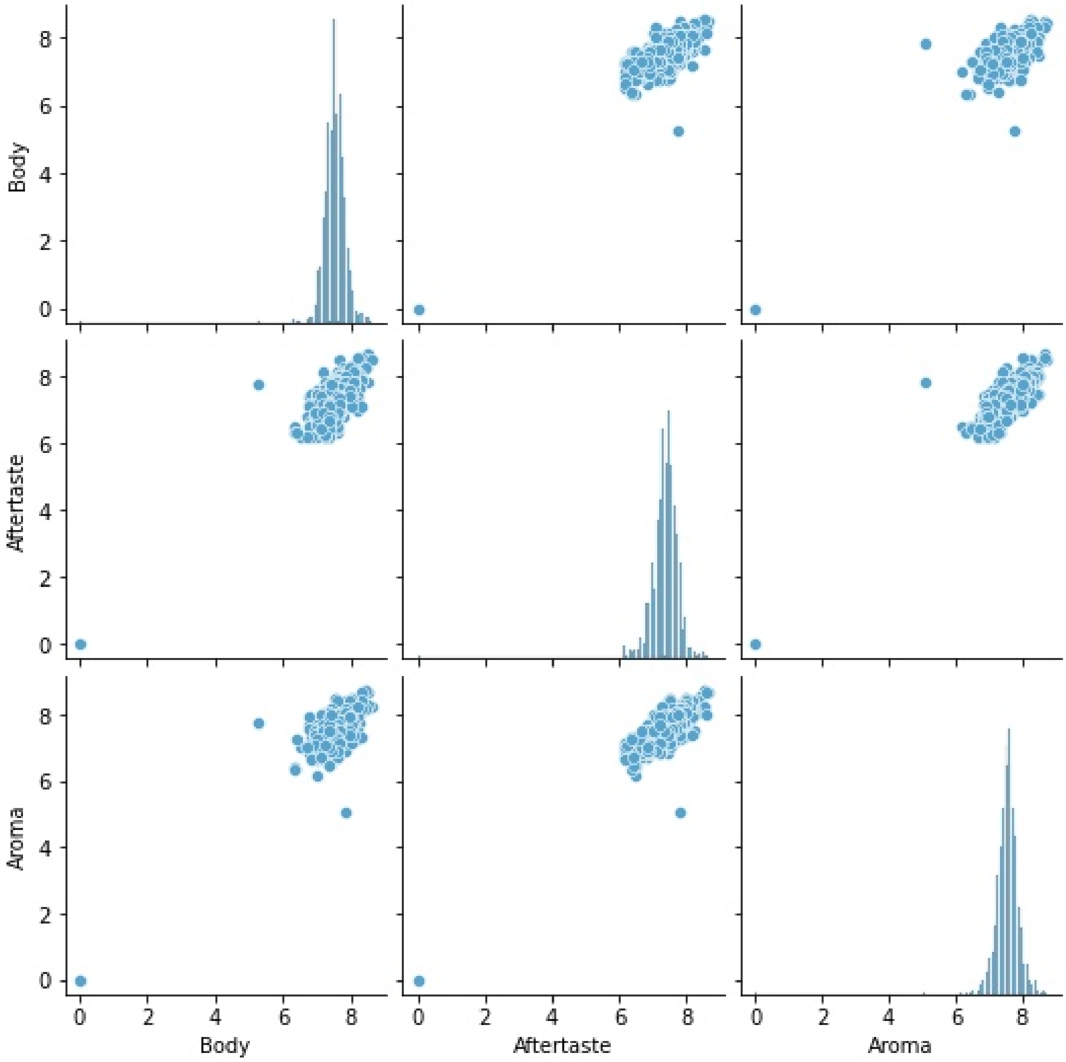
Obviously, this may not be entirely true given how *Harvest.Year* is distributed.

Although it is true that there are 24 variables, the previous analysis covered all the important variables, since the others are considered irrelevant because they do not have valuable information in addition to having many null values.

4.2 Numerical analysis

To begin, we will analyze the distribution of the 6 most important coffee metrics. This can be done with seaborn subplots and histplots, however there is a method called pairplot. With the help of pairplot function it is possible to make a scatter plot between all the variables involved.

```
df=data[["Body","Aftertaste","Aroma"]] # data[["Flavor","Acidity","Balance"]]
sns.pairplot(df)
```



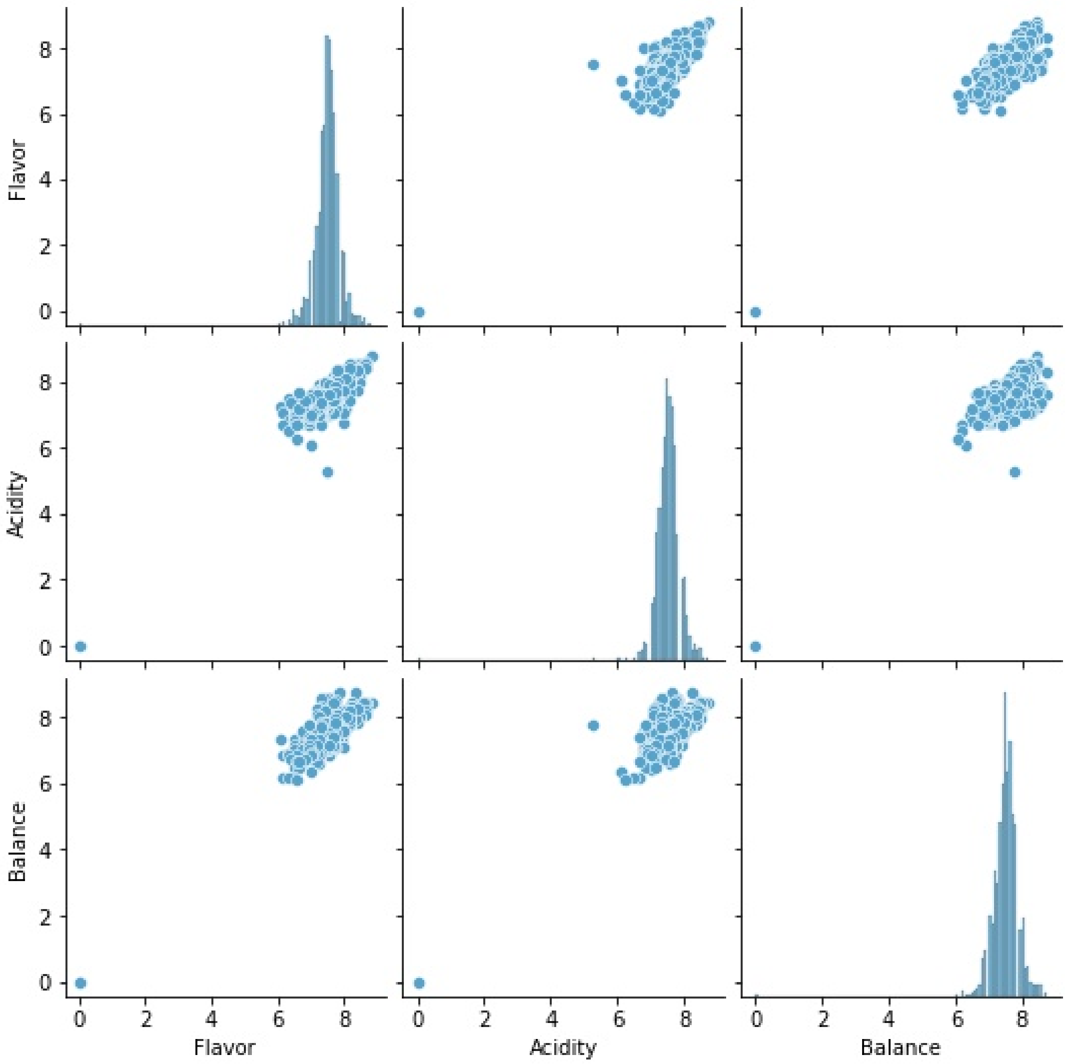


Figure 9: Scatter and histogram plot of each df column.

It is important to see how all the variables involved are between the interval 6 and 10, having some abnormal value outside. This indicates a possible very strong correlation between all the variables. Something that would be expected because in reality there is a very marked relationship between them. Let's not forget that we are talking about the characteristics of the coffee with which it is scored how good or bad it is.

Talking about the quality of the coffee, The Specialty Coffee Association (SCA) defines two grades of coffee: Specialty Grade and Premium Grade. All beans must be of a specific size, have at least one distinctive attribute in the body, flavor, aroma or acidity and have moisture content between 9 and 13 percent.

```
corr_matrix=data[["Flavor","Aroma","Balance","Acidity","Body","Aftertaste"]].corr(method='pearson')
fig,ax=plt.subplots(figsize=(9,9))
sns.heatmap(corr_matrix,annot=True,cbar=False,annot_kws={"size":8},vmin=-1,vmax=1,
center=0,cmap=sns.diverging_palette(240,10,n=200),square=True,ax=ax)
ax.set_xticklabels(ax.get_xticklabels(),rotation=45,horizontalalignment='right',)
ax.tick_params(labelsize=9)
```

```
ax.xaxis.label.set_color("w")
ax.yaxis.label.set_color("w")
ax.spines['left'].set_color('w')
ax.spines['bottom'].set_color('w')
ax.tick_params(colors='white',which='both')
```

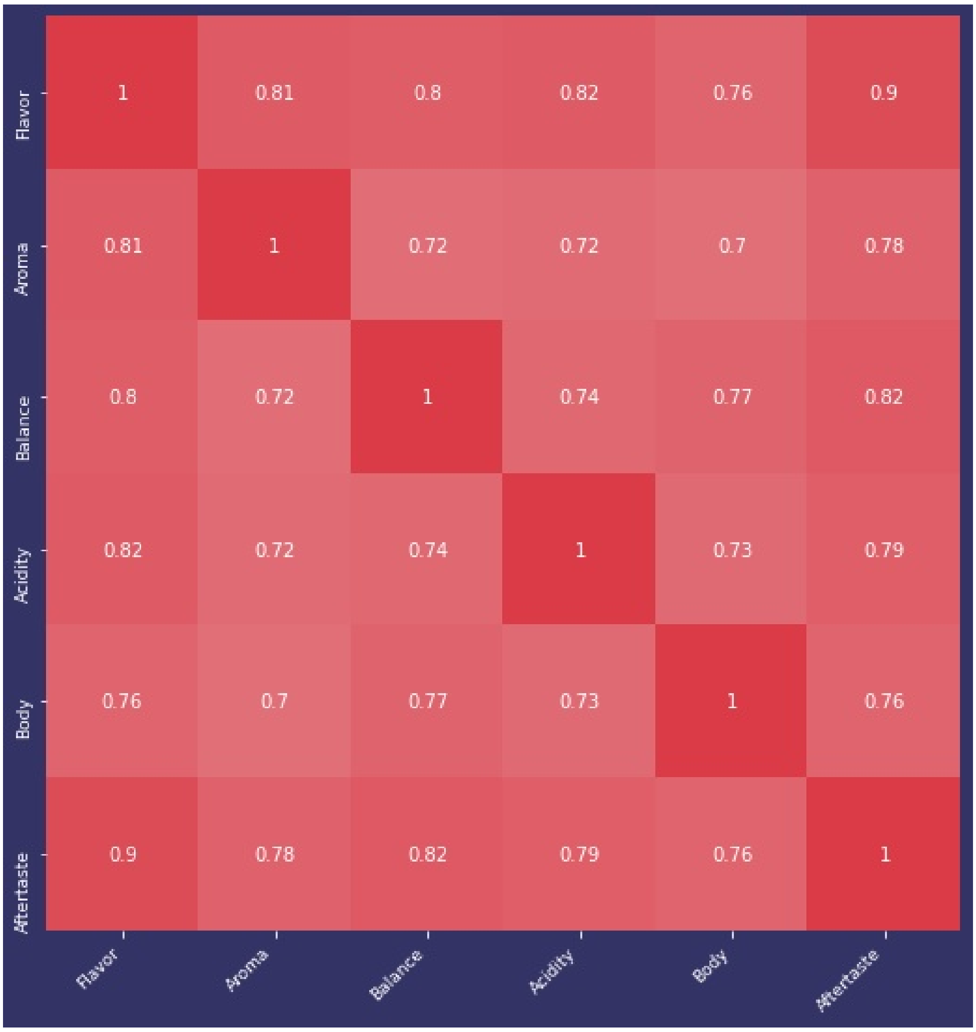


Figure 10: Correlation matrix of each df column.

Note how the aftertaste is the variable that most correlates with the flavor. In addition, the other variables also have a strong correlation with each other, something that, as I said before, was to be expected. However, it is important to mention that **correlation does not imply causation**, that in fact there are many funny examples in tylervirgen where correlations that do not make sense.

Now, let's look at some descriptive statistics.

```
df=data[["Aroma","Balance","Acidity","Body","Aftertaste","Flavor"]]
df.describe()
```

	Aroma	Balance	Acidity	Body	Aftertaste	Flavor
count	1311.000000	1311.000000	1311.000000	1311.000000	1311.000000	1311.000000
mean	7.563806	7.517506	7.533112	7.517727	7.397696	7.518070
std	0.378666	0.406316	0.381599	0.359213	0.405119	0.399979
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	7.420000	7.330000	7.330000	7.330000	7.250000	7.330000
50%	7.580000	7.500000	7.500000	7.500000	7.420000	7.580000
75%	7.750000	7.750000	7.750000	7.670000	7.580000	7.750000
max	8.750000	8.750000	8.750000	8.580000	8.670000	8.830000

Hence, the fact that the mean and standard deviation of almost all of them is very similar. In addition, the aroma, balance and acidity have the same maximum value. This could indicate the close relationship that exists between these three characteristics.

In the case of flavor, we can request more statistics as it is our variable of interest.

```
dfs=DataFrameSummary(df)
dfs['Flavor']
```

mean	7.51807
std	0.399979
variance	0.159983
min	0.0
max	8.83
mode	7.5
5%	6.92
25%	7.33
50%	7.58
75%	7.75
95%	8.0
iqr	0.42
kurtosis	95.172934
skewness	-5.223512
sum	9856.19
mad	0.258962
cv	0.053202
zeros_num	1
zeros_perc	0.08%
deviating_of_mean	1
deviating_of_mean_perc	0.08%
deviating_of_median	17
deviating_of_median_perc	1.30%
top_correlations	Aftertaste: 89.53%, Acidity: 81.67%, Aroma: 81...
counts	1311
uniques	35
missing	0
missing_perc	0%
types	numeric

Name: Flavor, dtype: object

We also see that the mode turns out to be 7.5, which can be interpreted as a regular grade. Which seems to indicate that the reviews were extremely strict. In fact, the minimum value is 0, indicating very bad coffee.In addition to this, due to the skewness and kurtosis we see that the data are very clustered in the mean.

In addition to everything seen so far, it is time to answer the second question and end with this exploratory data analysis.

For this, we create a new row called *score f*.

```
data['Score_f'] = data.Aroma+data.Acidity+data.Body+data.Aftertaste+data.Flavor+data.Balance
```

Now we have the sum of the main characteristics of coffee. We can only group by country of origin.

```
data[['Country.of.Origin','Score_f']].groupby('Country.of.Origin').sum().sort_values(by='Score_f',
                                                    ascending=False).head(6)
```

Country.of.Origin	Score_f
Mexico	10431.71
Colombia	8354.46
Guatemala	8116.54
Brazil	5960.42
Taiwan	3351.31
United States (Hawaii)	3324.20

Note how when adding all the scores, the position of the countries is the same as **Figure 8** and furthermore, we are getting the total sum of the scores by country. This tells us that it is not the correct solution, because as we saw previously, there are countries that repeat themselves with small bags of coffee (less than 2 kg) and other countries that do not with bags greater than 2 kg. So, we will try the average instead of the sum.

```
data['Score_f']=(data.Aroma+data.Acidity+data.Body+data.Aftertaste+data.Flavor+data.Balance)/6

data[['Country.of.Origin','Score_f']].groupby('Country.of.Origin').mean().sort_values(by='Score_f',
                                                    ascending=False).head()
```

Country.of.Origin	Score_f
Papua New Guinea	8.193333
United States	7.993333
Ethiopia	7.956553
Rwanda	7.805000
Kenya	7.779200

This represents the average score per kilo of coffee, that is, it is the average evaluation (in a range of 0 to 10) that the experts give to each kilo of coffee. However, we should not get carried away by the result as it is, because although Papua New Guinea has the best score, only one kilo was reviewed. If we analyze the weights dataframe it is possible to observe how Kenya and Ethiopia are among the 5 largest coffee producers. In this sense, it is possible to observe how Kenya produced 732 kilos more than Ethiopia but obtained a lower average score 0.177353 points less). Therefore, **Ethiopia is the country that produced the best coffee according to the reviews obtained**.

So, summarizing the questions:

1. Which country produces the most coffee? R: **Costa Rica.**
2. Taking into account the parameters defined in the data, who produces the best coffee? R: **Ethiopia.**

5 MODEL

One of the big questions that come up when you get into data science is, which model to use? In some cases it is obvious which one can be used. For example, if it is a time series, they quickly emerge as models such as: ARIMA in any of its derivatives, SARIMAX (somewhat more generalized), Holt Winters, GARCH, among others. But nevertheless, When it comes to only numerical data, many more models can be applied. In fact, when this happens, in the end it really becomes a cycle of trial and error, where we seek to obtain the best possible metrics and a prediction that is as faithful as possible to the real values.
Then, in this specific case, a multiple linear regression model will be used, in order to find a linear relationship between the flavor and the other parameters previously studied.

Formal definition of the model

Multiple linear regression tries to fit linear or linearizable models between one dependent variable and more than one independent variable. In this kind of models it is important to test for heteroscedasticity, multicollinearity and specification.
In this part, it is important to mention that a multiple linear regression model follows a series of assumptions, however, these assumptions are better generalized in the Gauss-Markov Theorem. Where, Gauss-Markov theorem is a set of assumptions that an OLS (Ordinary Least Squares) estimator must meet in order for it to be considered ELIO (Optimal Linear Unbiased Estimator). The intersection constant does not make sense in this case, so we can cancel it.

The model looks like this:

$$Y = bX_1 + cX_2 + dX_3 + eX_4 + fX_5 + u \tag{1}$$

With:

- Y = Flavor
- X_1 = Balance
- X_2 = Body

- X_3 = Aftertaste
- X_4 = Acidity
- X_5 = Aroma
- u = Residuals

And b, c, d, e, f are constants to be found.

To find the constants it is possible to do it manually or use Python. In the case of Python it is possible to do it with the sklearn library or the statsmodels library. In this case we will use statsmodels.

```
import statsmodels.api as sm
from sklearn.metrics import mean_squared_error,mean_absolute_error
from permetrics.regression import Metrics

df=data[["Balance","Body","Aftertaste","Acidity","Aroma"]]
reg=sm.OLS(data["Flavor"],df).fit() #OLS means Ordinary Least Squares
reg.summary()
```

OLS Regression Results

Dep. Variable:	Flavor	**R-squared (uncentered):**	1.000
Model:	OLS	**Adj. R-squared (uncentered):**	1.000
Method:	Least Squares	**F-statistic:**	6.445e+05
Date:	Thu, 29 Jul 2021	**Prob (F-statistic):**	0.00
Time:	22:01:30	**Log-Likelihood:**	613.56
No. Observations:	1311	**AIC:**	-1217.
Df Residuals:	1306	**BIC:**	-1191.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Balance	0.0725	0.020	3.621	0.000	0.033	0.112
Body	0.0490	0.019	2.523	0.012	0.011	0.087
Aftertaste	0.4718	0.022	21.175	0.000	0.428	0.516
Acidity	0.1901	0.019	9.791	0.000	0.152	0.228
Aroma	0.2224	0.018	12.215	0.000	0.187	0.258

Omnibus:	51.317	**Durbin-Watson:**	1.986
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	132.872
Skew:	-0.139	**Prob(JB):**	1.40e-29
Kurtosis:	4.535	**Cond. No.**	104.

From the above there are several important conclusions. The first has to do with the p-value, where we can conclude that all variables have a place in the model. The Durbin-Watson statistic tells us that successive error terms are negatively correlated. Obviously we cannot forget the Score, which turns out to be 1, that is, the best possible score. Which could indicate that is the best model, but this should never be concluded without carefully reviewing the results.

So, the final model is:

$$Y = 0.0725X_1 + 0.0490X_2 + 0.4718X_3 + 0.1901X_4 + 0.2224X_5 \quad (2)$$

Comparing the values we see that

```
dfinal={'Predict':reg.predict(),'Real':data['Flavor']}
dfinal=pd.DataFrame(dfinal)
dfinal.head(10)
```

	Predict	Real
0	8.709686	8.83
1	8.611032	8.67
2	8.465054	8.50
3	8.405452	8.58
4	8.360124	8.50
5	8.505407	8.42
6	8.421553	8.50
7	8.470785	8.33

It seems that the model is faithful. To check this we will use four evaluation metrics:

- **MSE**. It is perhaps the simplest and most common metric for regression evaluation, but it is also probably the least useful. MSE basically measures the mean squared error of our predictions.

- R^2. The R-squared is a statistical measure of how close the data is to the fitted regression line. It is also known as coefficient of determination, or coefficient of multiple determination if it is multiple regression.

- **MAE**. In MAE, the error is calculated as an average of absolute differences between the target values and the predictions. The MAE is a linear score, which means that all individual differences are weighted equally on the average. The important thing about this metric is that it penalizes huge errors that are not as bad as MSE does. Therefore, it is not as sensitive to outliers as root mean square error.

- **MAPE**. Mean Absolute Percent Error measures the average error in percentage. It is calculated as the average percentage of the absolute errors.

Calculating the above

```
metrics=Metrics(np.array(dfinal.Real),np.array(dfinal.Predict))
lista=[]
for ele in dir(metrics)[14:21]:
    method=getattr(metrics,ele)
    results=method(decimal=6)
    lista.append([ele,results])
metrics=pd.DataFrame(lista,columns=['Metric','Value'])
metrics=metrics.append({'Metric':'R2','Value':reg.rsquared},ignore_index=True)
metrics
```

	Metric	Value
0	MAE	0.114880
1	MAPE	0.015391
2	MASE	0.690020
3	ME	0.694947
4	MRE	0.015391
5	MSE	0.022962
6	MSLE	0.000418
7	R2	0.999595

The MSE is very small as is the MAPE, which could indicate a model that fits the original data very well. However, all other metrics follow the same "smaller is better" premise (except for R2, obviously).
Therefore, it is possible to conclude that it is a **good model**. Obviously, to determine if it is the best one, the same metrics should be compared but from other models, but the point of this paper is not to show which is the best model that fits the data, but to show a good model that fits the data.

5.1 Interpretation

In this case, we are not interested in analyzing the residuals, given by how we construct the model. So, we are going to analyze the model as such.
The term ceteris paribus is a Latin word used by economists to allow yourself to speculate on what would happen if we isolated the effect of a single variable the remainder remaining constant.
In this sense, if we take a ceteris paribus effect in X_1, that is, we keep the other terms constant, *the increase of one point in the Balance has an impact on the flavor score of 0.0725*. In the same way for X_2, *if the Body increases one point, the flavor increases by 0.0490*. This is repeated with all the other variables, the interesting thing is what happens with X2 and X3, where, in the case of X3, *if the Aftertaste increases one point, then the flavor increases by 0.4718*. **This is the variable that has the greatest effect on flavor according to the model**. **While Body is the variable that has the least effect on flavor**.
Furthermore, in **figure 14**, it is possible to see that Aftertaste is the variable that most correlates with flavor. While the Body is the variable with the least correlation.

So, answering the question posed in the introduction:
What is the most important characteristic of coffee to conclude that it is a great coffee? **Aftertaste**.

6 CONCLUSION

It was interesting to analyze the dataset because I had not worked with so many columns, although I did not really use all of them, since most did not provide relevant information. In this

sense, I think that the dashboard developed in Power Bi is an excellent way to visualize the most important columns among themselves and obtain different conclusions, although, obviously, it was not necessary.
We take a closer look at the dataset where we answered two insteresting questions, where, we conclude that Costa Rica was the country that produced the most coffe during 2009 to 2018. I obtained this interval from exploring the *Harvest.Year* column, since the column has abnormal values. In addition, we also saw that Ethiopia has the best coffe and that Mexico was the most reviewed country.

Using the available dataset, a multiple linear regression model was developed that achieved an r squared of **1.0**, indicating an excellent model. Obviously, it is not possible to conclude that it is the best, since it is very likely that with another model a similar r-squared will also be obtained and other metrics must be looked at. However, it is an excellent model, as I said earlier. Therefore, the purpose of the work was successfully concluded. It may be possible to recommend trying other models and comparing them with each other in order to get the best one. In the case of exploratory analysis I think more questions could be answered using data manipulation.