Milestone 3

Context

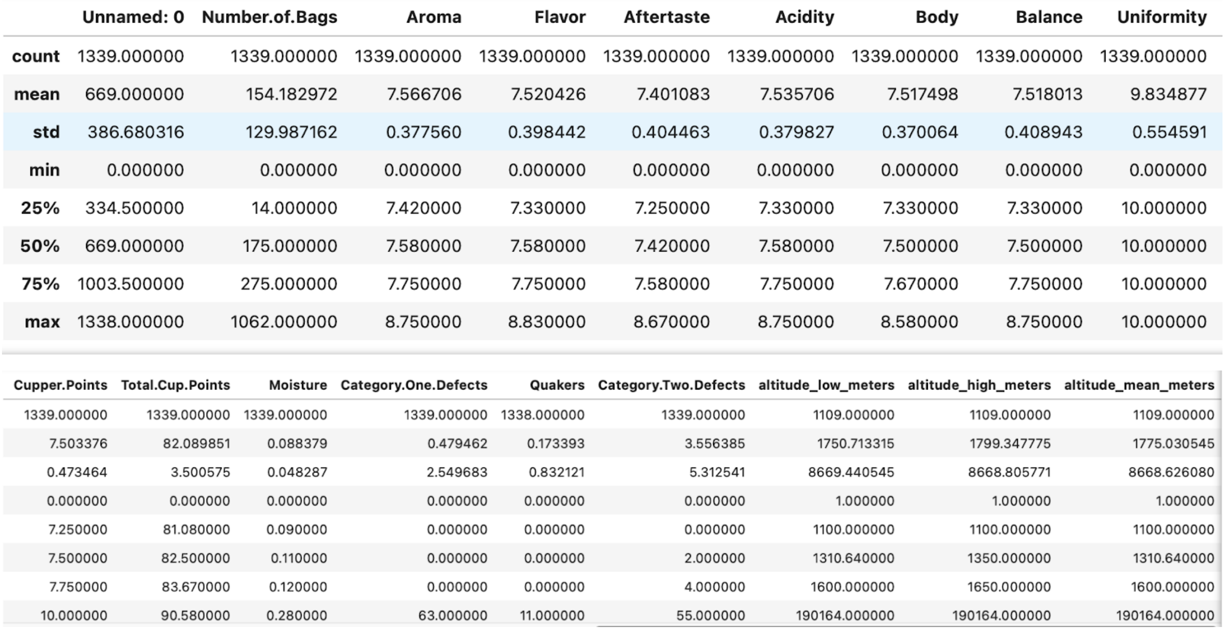
Undoubtedly, coffee consumption is prevalent in adults as sleep deprivation continues to be a concern for the unemployed graduate student. For the five of us, our reliance on coffee extends beyond the mere milligram intake of caffeine, and into the satisfaction gained through sensory input (on our taste buds). In order to assert dominance, our collaborative team of five motivated Master's students aim to start our own coffee chain (sometime in the near foreseeable future). However, we understand the coffee beans themselves are the soul to a good cup of coffee. But what else contributes to the elixir that wakes us up in the morning? This is our goal, to gain a thorough understanding of coffee beans and the features that are the best predictors of quality.

Summary of the Data

We obtained our dataset from Kaggle which contains 3 csv files: one for Arabica coffee, one for Robusta coffee and a merged csv file containing the rows from the previous ones. We are going to use this merged csv file for our study. The merged_dataset has 11339 rows and 44 columns.

Below are tables we used to get a quick summary of the dataset:

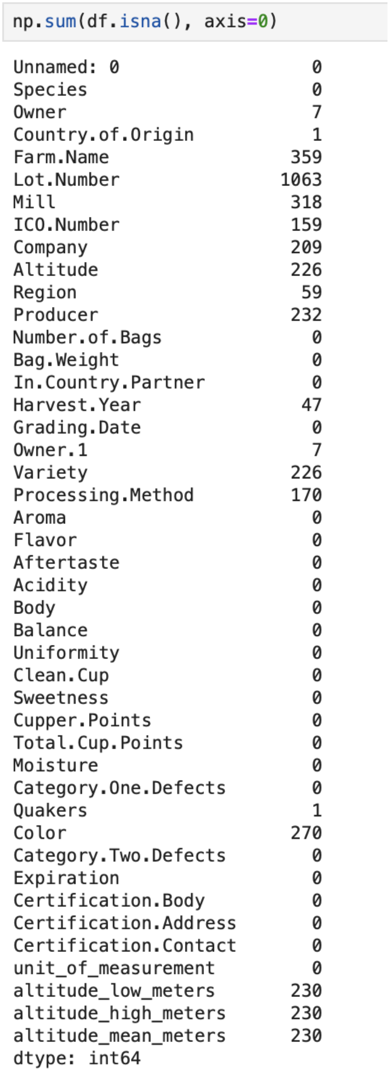
Describe int/float columns:



	Unnamed: 0	Number.of.Bags	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Uniformity
count	1339.000000	1339.000000	1339.000000	1339.000000	1339.000000	1339.000000	1339.000000	1339.000000	1339.000000
mean	669.000000	154.182972	7.566706	7.520426	7.401083	7.535706	7.517498	7.518013	9.834877
std	386.680316	129.987162	0.377560	0.398442	0.404463	0.379827	0.370064	0.408943	0.554591
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	334.500000	14.000000	7.420000	7.330000	7.250000	7.330000	7.330000	7.330000	10.000000
50%	669.000000	175.000000	7.580000	7.580000	7.420000	7.580000	7.500000	7.500000	10.000000
75%	1003.500000	275.000000	7.750000	7.750000	7.580000	7.750000	7.670000	7.750000	10.000000
max	1338.000000	1062.000000	8.750000	8.830000	8.670000	8.750000	8.580000	8.750000	10.000000

Cupper.Points	Total.Cup.Points	Moisture	Category.One.Defects	Quakers	Category.Two.Defects	altitude_low_meters	altitude_high_meters	altitude_mean_meters
1339.000000	1339.000000	1339.000000	1339.000000	1338.000000	1339.000000	1109.000000	1109.000000	1109.000000
7.503376	82.089851	0.088379	0.479462	0.173393	3.556385	1750.713315	1799.347775	1775.030545
0.473464	3.500575	0.048287	2.549683	0.832121	5.312541	8669.440545	8668.805771	8668.626080
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000
7.250000	81.080000	0.090000	0.000000	0.000000	0.000000	1100.000000	1100.000000	1100.000000
7.500000	82.500000	0.110000	0.000000	0.000000	2.000000	1310.640000	1350.000000	1310.640000
7.750000	83.670000	0.120000	0.000000	0.000000	4.000000	1600.000000	1650.000000	1600.000000
10.000000	90.580000	0.280000	63.000000	11.000000	55.000000	190164.000000	190164.000000	190164.000000

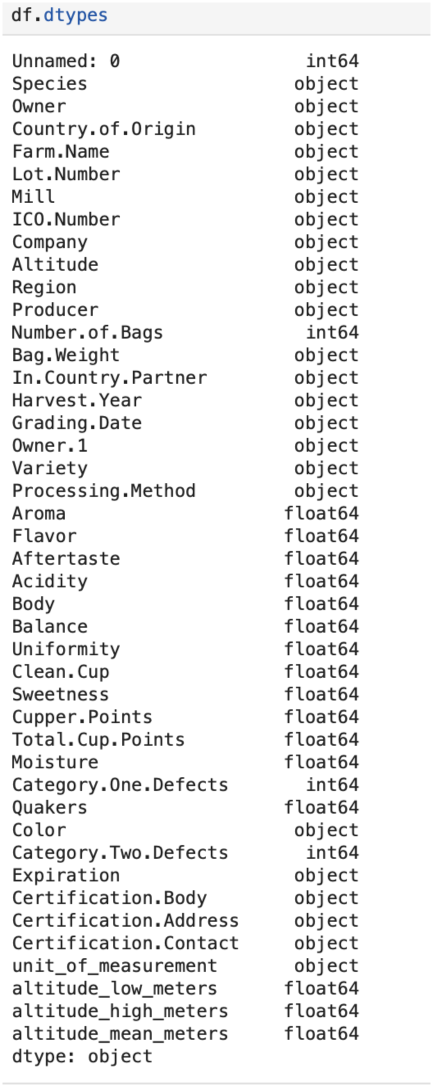
Missing values:



```
np.sum(df.isna(), axis=0)
```

```
Unnamed: 0                 0
Species                    0
Owner                      7
Country.of.Origin          1
Farm.Name                359
Lot.Number              1063
Mill                     318
ICO.Number               159
Company                  209
Altitude                 226
Region                    59
Producer                 232
Number.of.Bags             0
Bag.Weight                 0
In.Country.Partner         0
Harvest.Year              47
Grading.Date               0
Owner.1                    7
Variety                  226
Processing.Method        170
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                    270
Category.Two.Defects       0
Expiration                 0
Certification.Body         0
Certification.Address      0
Certification.Contact      0
unit_of_measurement        0
altitude_low_meters      230
altitude_high_meters     230
altitude_mean_meters     230
dtype: int64
```

Columns types:



```
df.dtypes
```

```
Unnamed: 0                 int64
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
```

After a quick glance, we noticed some preliminary data issues in the merged dataset. Such issues include missing values, particularly with regard to geographic location of the bean source. Robusta beans consistently have missing values for “altitude_high_meters”, “altitude_low_meters”, and “altitude_mean_meters”. Since these columns are so similar, we decided that we will drop these columns and use “Altitude” instead. Also, we noticed that the column “Lot.Number” mostly contains missing values and decided to drop it. In addition, we noticed that there are some farms with “altitude_high_meters” greater than 9,000 meters (i.e. greater than Everest). Further analysis revealed that it was most likely due to input error as the same farms have an altitude that is scaled down by 100. We also noticed some duplicated columns : “Owner” and “Owner.1”. Columns “Owner” is the column without capitalization and “Owner.1” is with capitalization. However, owners having many farms are not common and farms very rarely change owners : we then decided to drop this variable that was unlikely to help find trends in the dataset.

Looking beyond just one column at a time, we proceeded to explore the relationships between different predictors. We visualized the relationship between variables by plotting them against each other using seaborn's pairplot function (Fig 1 and Fig 2). We also looked at the overall distribution of each qualitative predictor variable by plotting a box plot with scaled data (Fig 3).

Deeper Understanding of Data:

Data Skewness

We realized that there were a lot of predictors that had different distributions (looking at the tables from above), therefore we decided to standardize the qualitative predictor variables. After standardizing the data and visualizing the distributions of the numeric features, we notice that many of the features have a few dramatic outliers that are more than 15 standard deviations from the mean. These outliers have skewed initial plots that we have made to further visualize the data, and a visualization of the dataset's second principal component reveals most of the data clumped together with only a few points demonstrating the variance of the PC.

Collinearity

It is important to note that we are primarily interested in some of the taste-related features to predict coffee score. Because of this, we created a pairplot between these features, to gain an understanding of any interactions between the predictors. Upon visual inspection, the 'Aroma', 'Flavor', 'Aftertaste', 'Body', 'Acidity' and 'Balance' appear to have a strong positive correlation with each other. This correlation may violate some statistical assumptions of our model, so we will have to consider strategies to address this such as reducing dimensionality using PCA or dropping correlated features.

Data interpretation:

- Quality descriptions: https://database.coffeeinstitute.org/coffee/654175/grade
- Defects descriptions: https://database.coffeeinstitute.org/coffee/654175/green

The data comes from the Coffee Quality Institute, which supports a community of coffee experts that both educates individuals in coffee evaluation and scores individual cups of coffee with "a transparent, verifiable report", according to their website. They invite coffee-growers to submit their coffee for evaluation as a metric that they can provide to coffee buyers. Three certified coffee experts, known as “Q Graders” independently evaluate the coffee and their scores are averaged to produce the published metric. They consider coffee with a minimum Total Cup Score of 80 to be “specialty”. The feature of interest to us, “Cupper.Points”, is said on their website to “reflect the holistically integrated rating of the sample as perceived by the individual panelist”.

Meaningful Insights

Several aspects of our EDA have directed our project strategy. When visualizing the summary statistics of the scaled numeric features in boxplots (without outliers), we notice that there are a few features with almost no variance (“Uniformity”, “Clean.Cup”, “Sweetness”, “Category.One.Defects”, and “Quakers”). Although these predictors have no variance, it would be interesting as to whether or not these outlier points provide any significant information when classifying. However, if it is true that the low variance in these predictors truly provides no information to the classifier, we may consider dropping these predictors. Furthermore, after visualizing the distribution of the quality predictors (Fig 1 and Fig 2), it is clear that we have outliers in our data. This is because removing the outliers (Fig 1) reveals a better holistic view of the data. In total, we have discovered from our EDA that there are a total of 134 outliers in our dataset. We already knew that the number of Robusta and Arabica beans are not equally represented in our dataset, but EDA revealed that the “Country.of.Origin” predictor is also underrepresented.

Clean and Labeled Visualization

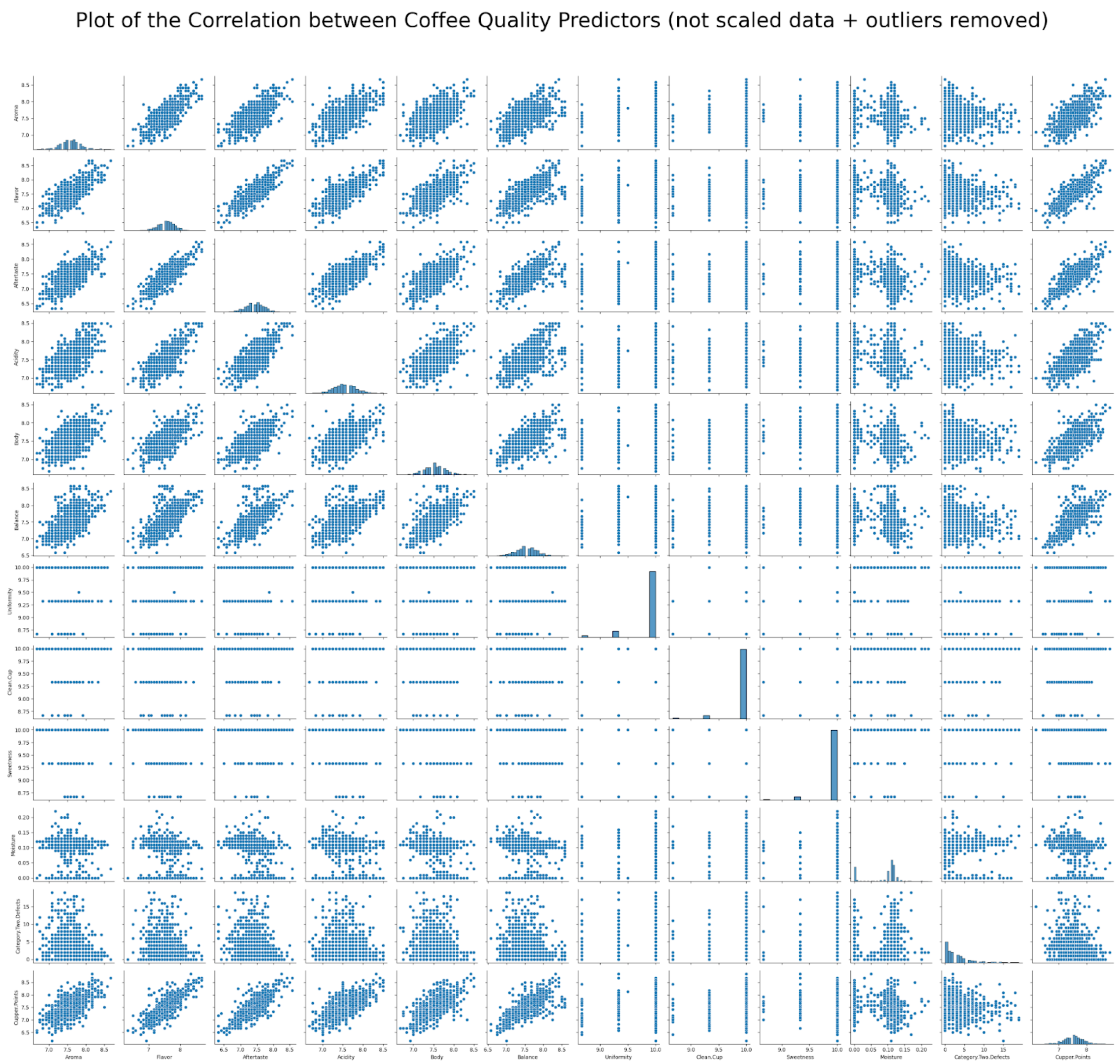
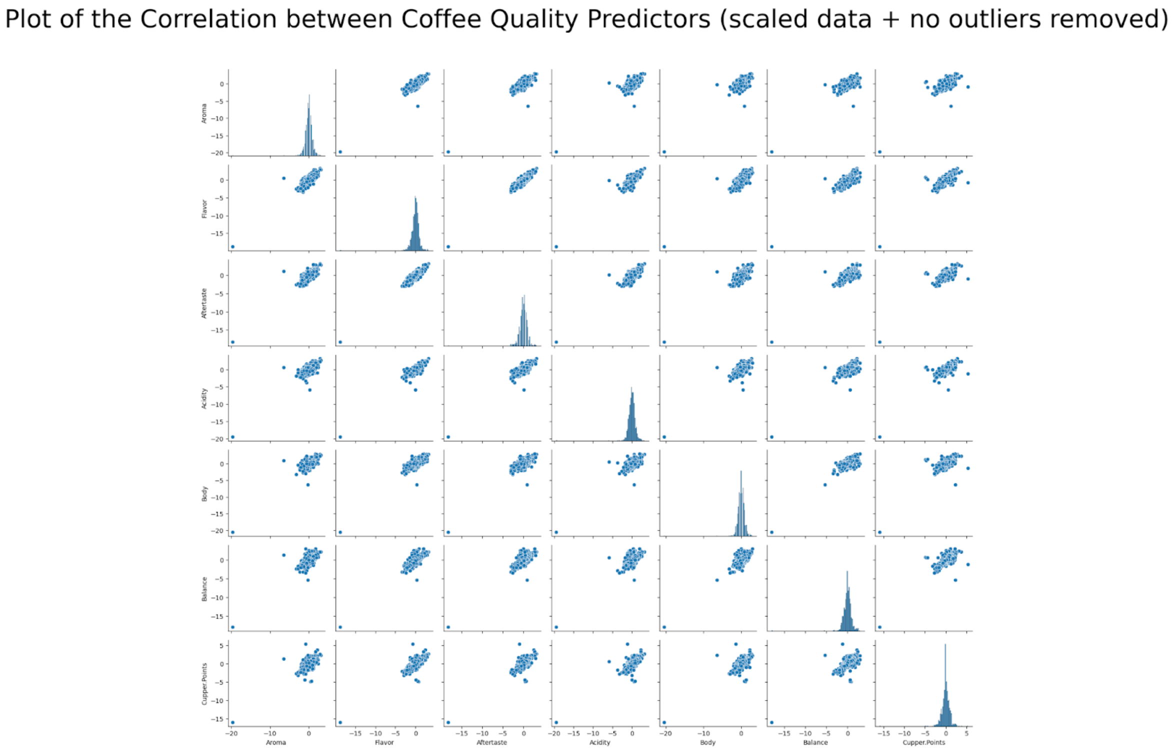


Fig 1. Plot of Quality Predictors (not scaled, outliers removed)

Some of the predictors show that they are correlated with each other, However the top correlated predictors (correlation > 0.8) are: ('Flavor', 'Aroma'), ('Aftertaste', 'Flavor'), ('Acidity', 'Flavor'), and ('Balance', 'Aftertaste')



Plot of the Correlation between Coffee Quality Predictors (scaled data + no outliers removed)

Fig 2. Plot of Quality Predictors (scaled, without removing outliers)

This plot reveals that each of the quality predictors have an outlier at the bottom left corner of the plot. We then looked at the new distribution after removing these outliers (Fig 1).

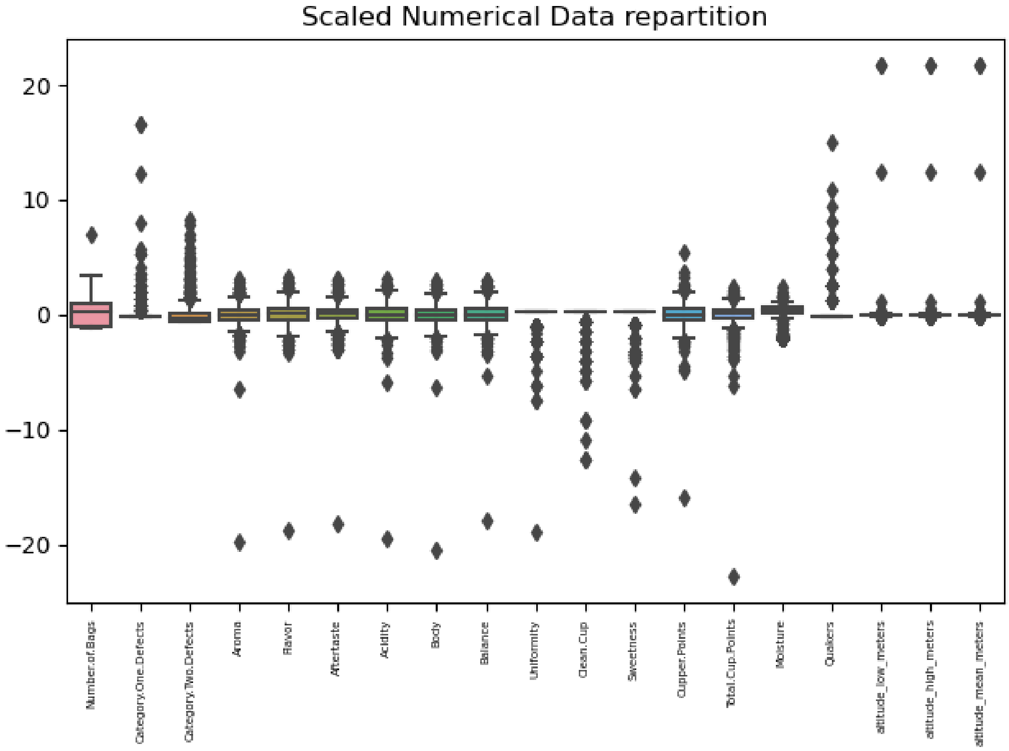


Fig 3. BoxPlot of Numerical Predictors

The distribution of the numerical predictors reveals that most of the predictors have mostly similar distributions, with Number of Bags having the widest distribution. A lot of the predictors have a lot of outliers, which is something we were aware of from Fig 2.

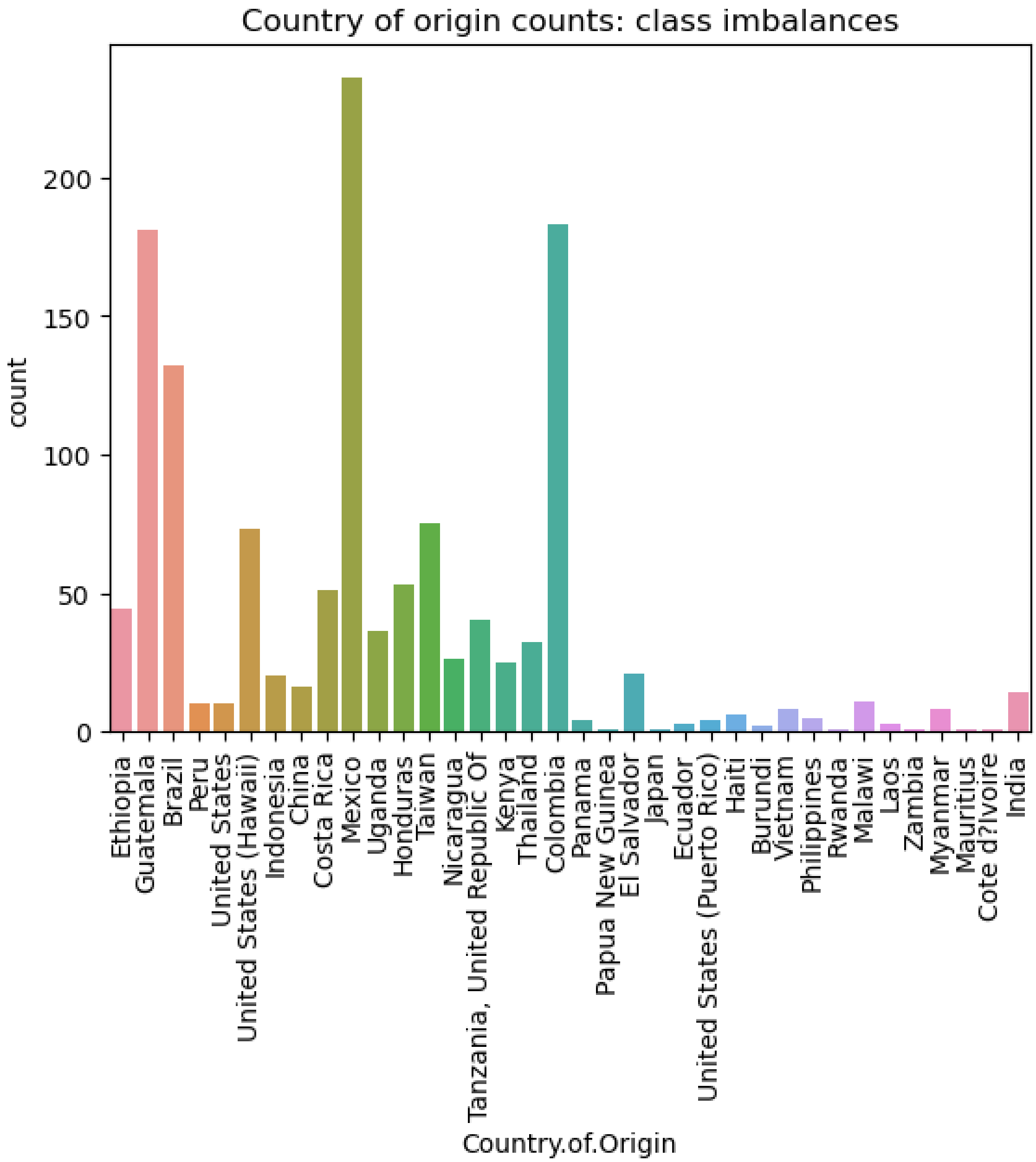


Fig 4. Plot of Class Imbalances for Country of Origin

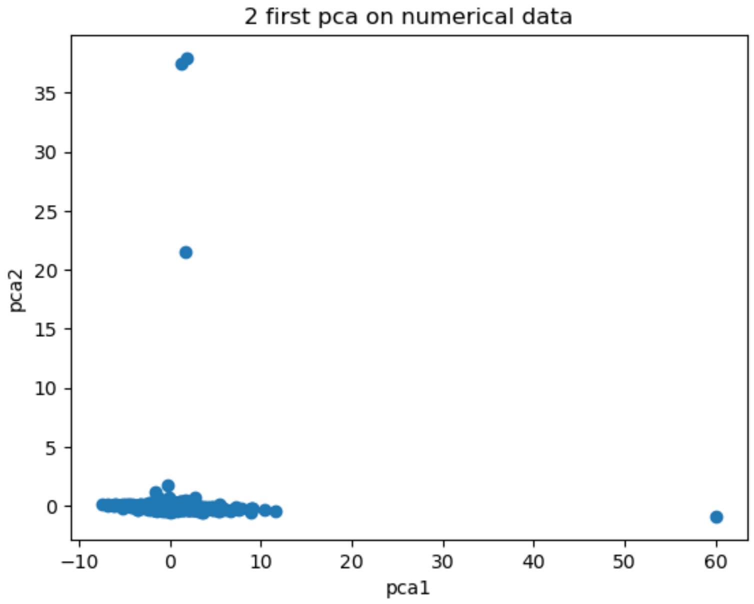


Fig 5. PCA on first two components

Data Description :

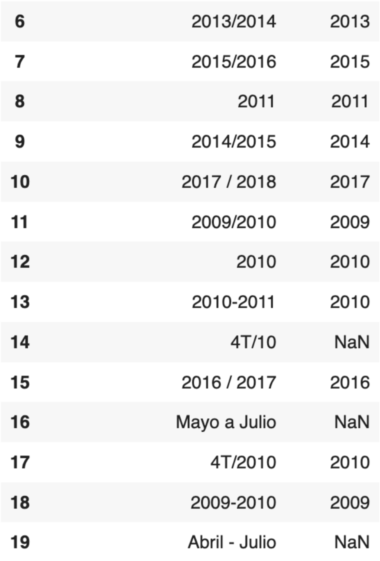
The dataset we are analyzing is the Coffee Quality dataset (https://www.kaggle.com/datasets/volpatto/coffee-quality-database-from-cqi). It contains information on different kinds of coffee (linked to origins, taste, ratings,...). Our goal is to study what makes a good coffee and we plan to use the different taste variables (aroma, balance, acidity,..) in the dataset and the one linked to the origin of coffee. The dataset contains many categorical variables that will require pre-processing and missing values. Data collection :https://github.com/jldbc/coffee-quality-database

The methods we used to explore the data include looking at the individual data types of each predictor (categorical or quantitative data). We noticed that there were 24 categorical predictors and 20 qualitative predictors. We then proceeded to look at whether or not there were missing data points for the predictors and which predictors had the most missing data. Most of the exploration included looking at various plots of the data through EDA (refer to Clean and Labeled Visualization section above).

We followed the following preprocessing steps:

1.remove outliers for numerical variables, where outliers are defined by observations with |z-score| > 3, this keeps 1197 out of 1339 rows.

2.clean Harvest.Year column: we noticed that there are messy values for Harvest.Year column, we extract the first consecutive 4-digit number with format “20xx” as the Harvest.Year (since the Year starts from 2016). If there is no matched number, the Harvest.Year is treated as missing. If the original Harvest.Year is a range (e.g. '2013/2014'), we take the beginning year (2013). The following picture provides an example of the contrast of original and cleaned values:



6	2013/2014	2013
7	2015/2016	2015
8	2011	2011
9	2014/2015	2014
10	2017 / 2018	2017
11	2009/2010	2009
12	2010	2010
13	2010-2011	2010
14	4T/10	NaN
15	2016 / 2017	2016
16	Mayo a Julio	NaN
17	4T/2010	2010
18	2009-2010	2009
19	Abril - Julio	NaN

Noteworthy Findings

When looking at our data we notice several class imbalances: notably, in the ‘Species’ column, there are only 28 ‘Robusta’ observations, while the rest are ‘Arabica’. Another source of class imbalance is in the ‘Country’ column, where there are a few highly represented countries, four with greater than 100 observations, and many other countries with fewer than 20 observations (Fig 4). We can address this by applying stratified sampling during our train-test split; we can also create an “other” category to gather the countries who are the least represented. We noticed a pretty high correlation between some variables too.

Project Question

However, we understand the coffee beans themselves are the soul to a good cup of coffee. This is our goal, to gain a thorough understanding of coffee beans and the features that are the best predictors of quality.

We initially wanted to predict “Total.Cup.Points” and then to analyze the coefficients/trees but we noticed that this column was actually the sum of other variables in the dataset. We now want to predict “Cupper.Points” which reflect the general point of view of a coffee expert on a kind of coffee as detailed here: https://database.coffeeinstitute.org/coffee/654175/grade.

We wish to better understand through this project what features most accurately estimate how coffee connoisseurs would rate certain coffee samples.

Baseline model / implementation plan:

While previous kaggle users have attempted to predict the coffee score, our aim is to focus on interpretable insights from the dataset to learn what features are associated with a good cup of coffee. To start, we will apply multiple linear regression, focusing on our feature selection due to the high dimensionality of the data (we might include Propensity Matching Score as robustness check). We will explore interaction terms and whether polynomial features are necessary. We can analyze feature importance using random forest regression and other decision-tree-based models; a simple decision tree might provide interpretable information as well. Other less interpretable methods we intend to apply are principal component analysis for dimensionality reduction, and potentially a neural network to assess whether there are aspects of the features that contribute to the prediction that were not necessarily captured by the simpler models.

- https://www.kaggle.com/code/devananjelito/ml-coffee-quality-regression#ML-Model