**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: | |
| --- | --- |
| Missing values: | Columns types: |
|

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') |
| --- |
| 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:



**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>

**Milestone 2 Submission: Acquire and Understand the Data**

**Data Skewness**

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. 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”).

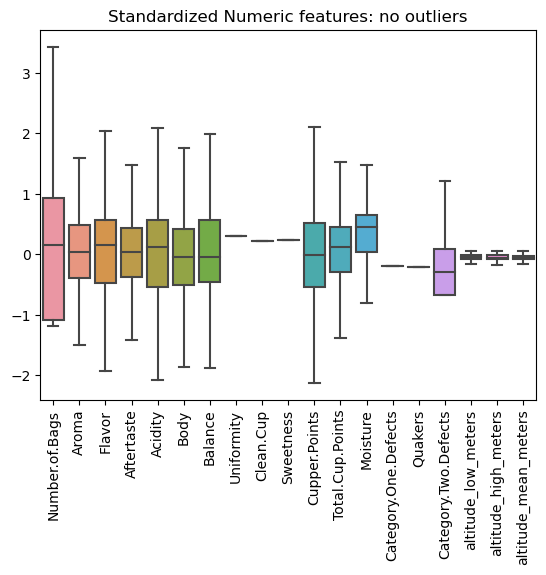
**Collinearity**

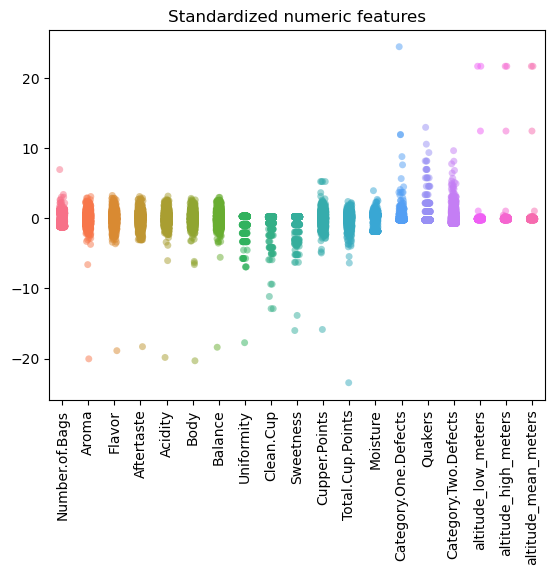
Because we are primarily interested in some of the taste-related features to predict coffee score, we created a pairplot between these features, to gain an understanding of any interactions between the predictors. Upon visual inspection, the ‘Aroma’, ‘Flavor’, ‘Aftertaste’, ‘Body’, ’Acidity’ and ‘Balance’ 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.

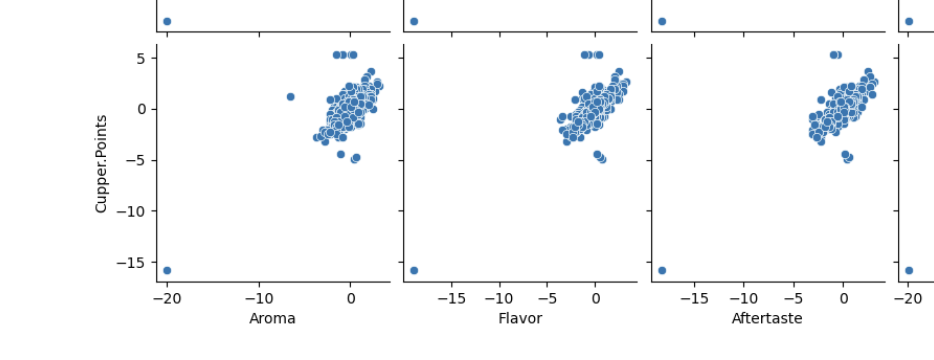
**Data interpretation:**

* Quality descriptions: <https://database.coffeeinstitute.org/coffee/654175/grade>
* Defects descriptions: <https://database.coffeeinstitute.org/coffee/654175/green>
* Owner.1 and Owner
* Lot number will be dropped
* Input error
* There is a test example to drop “ICO.Number” == “TEST”

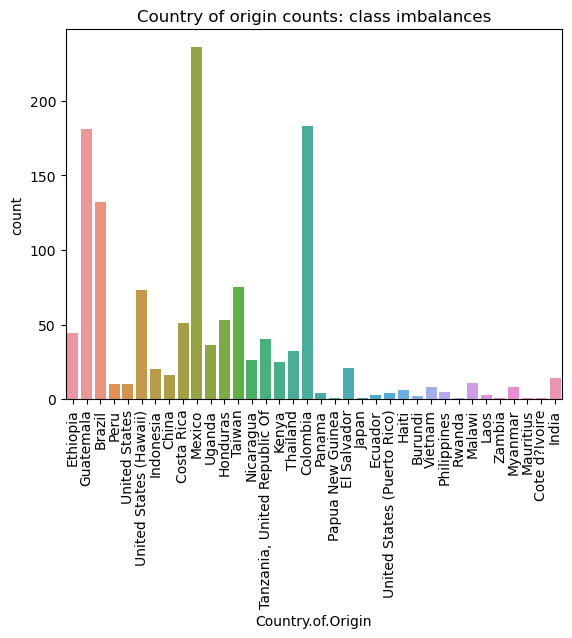
PLOTS :D



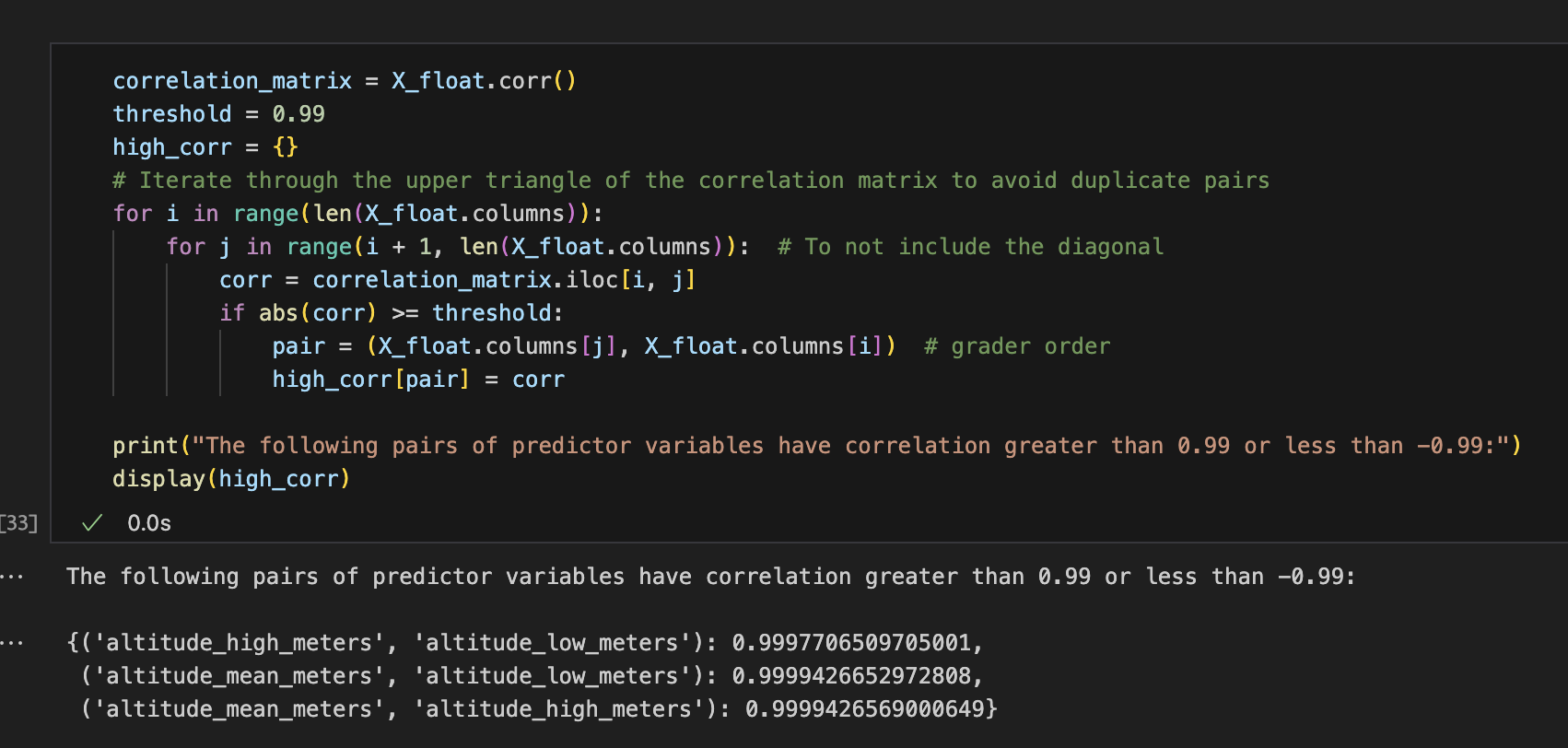


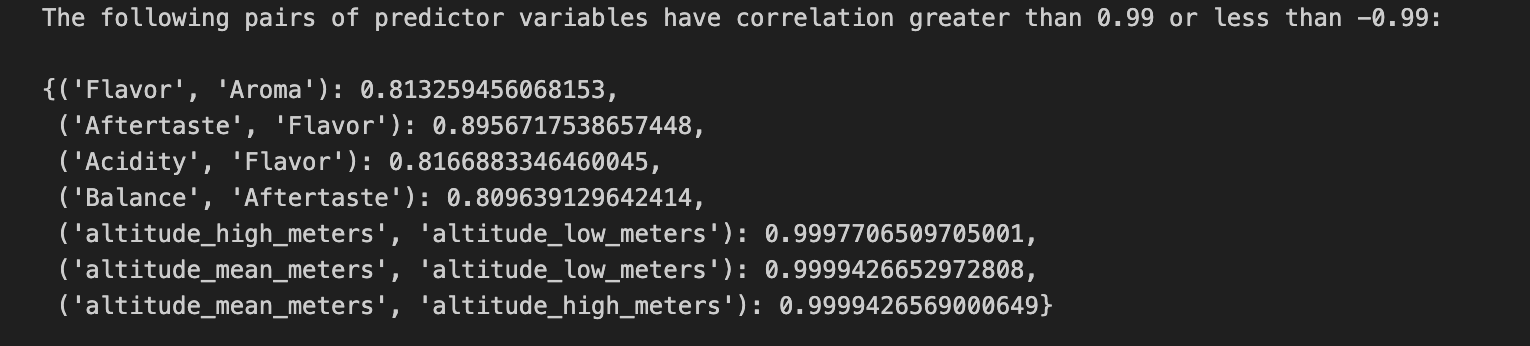


^^^The full plot is too much… main point: multicolinearity???



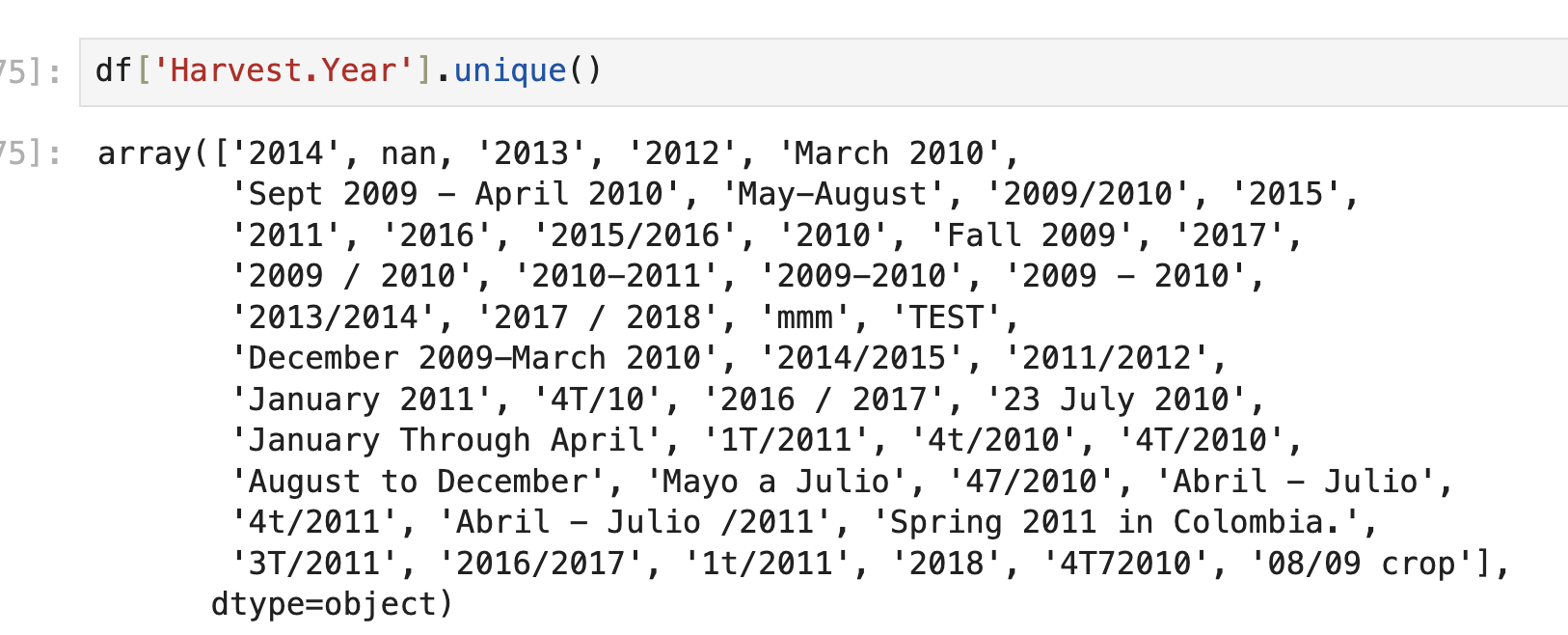
^^ main takeaway here: if we use categorical variables, we have to consider class imbalances



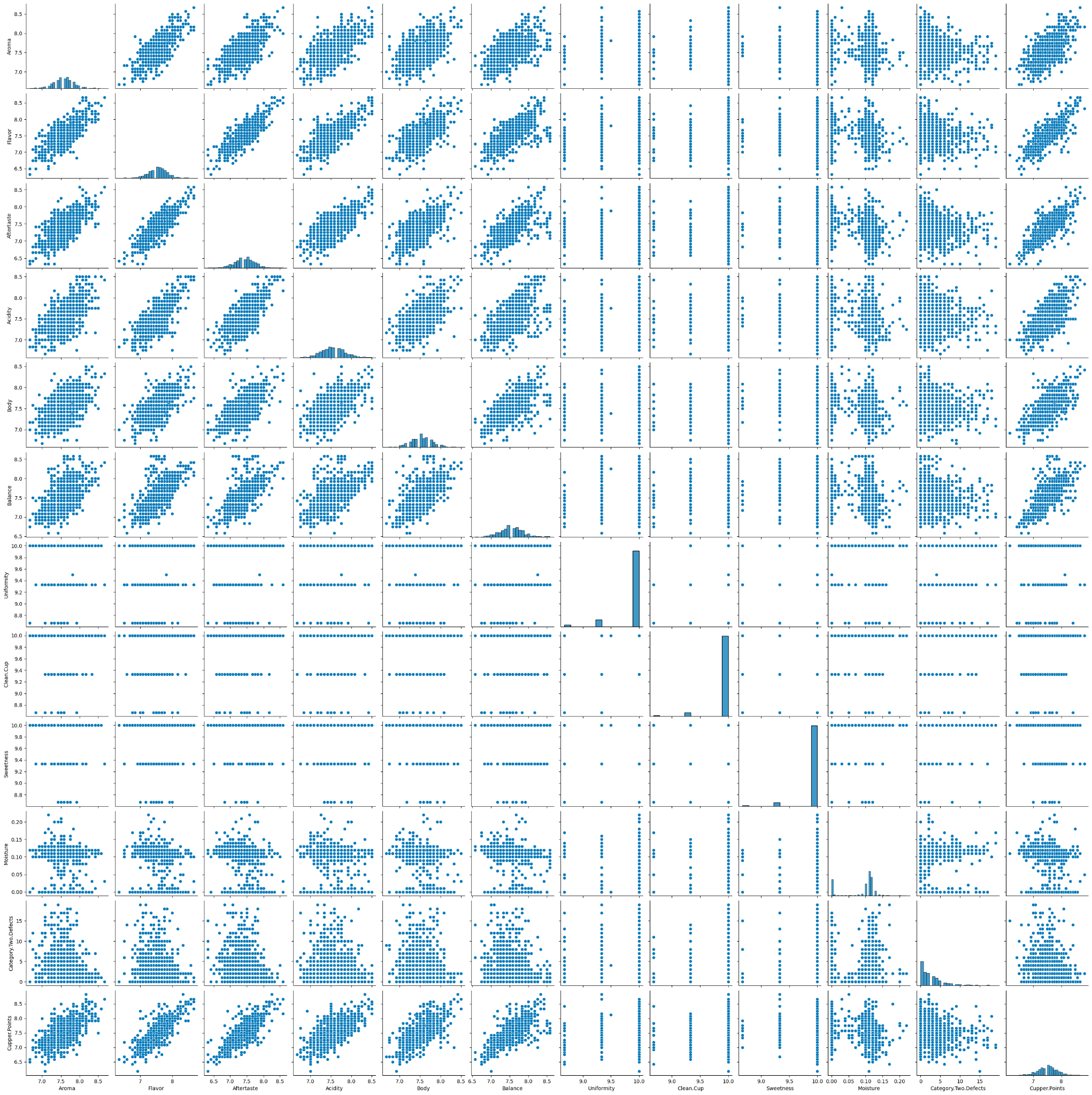


\*\* looking at correlation between numeric predictors → seems like the other ones aren’t *super* correlated?

But if the threshold is reduced down to 0.8, it seems like flavor is correlated with aroma, aftertaste, and acidity which balance and aftertaste are correlated. << good catch



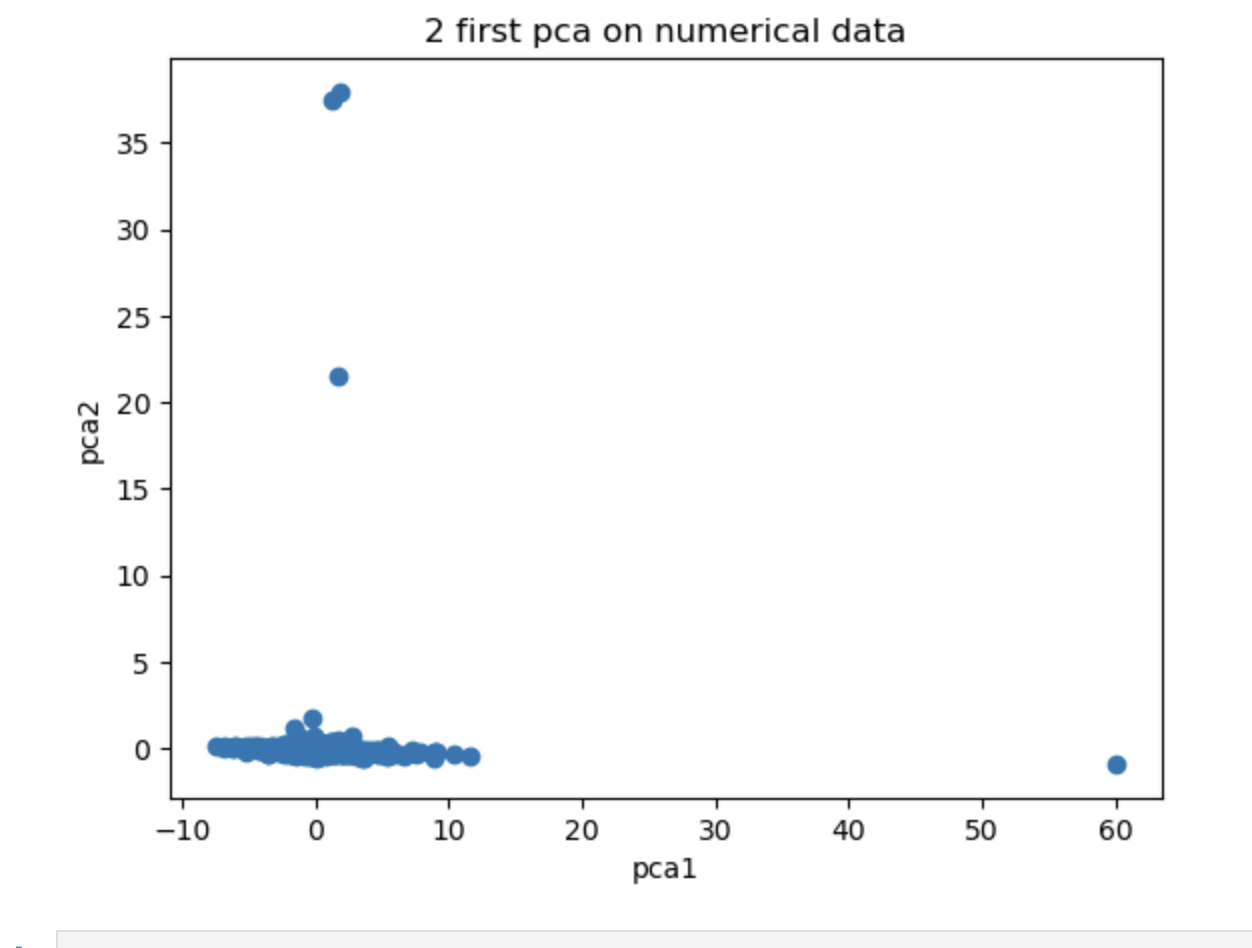
^^ some of the data is still messsyyyyy



Pairplot with the outliers removed → still some correlation, but according to the correlation matrix it doesn’t seem to be terrible imo

**Numerical Data exploration :**

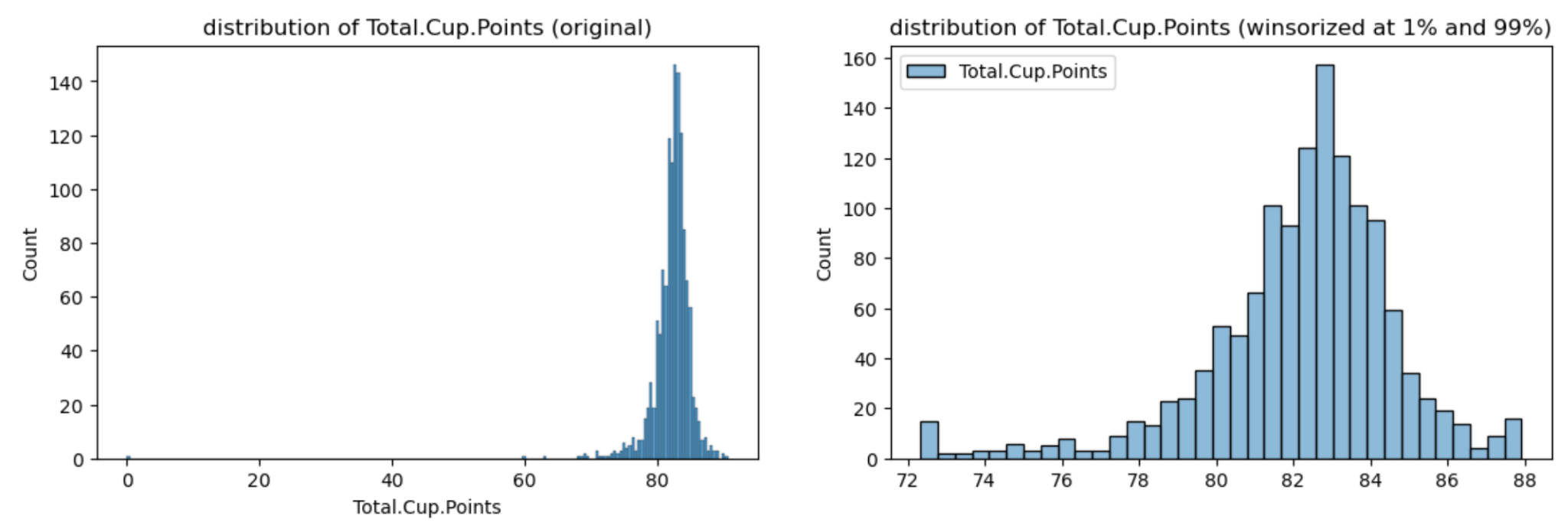
1. When we drop the line with missing values for this df we lose all the robusta beans (robusta beans never have values for the 3 altitude columns).
2. Pca does not show a lot



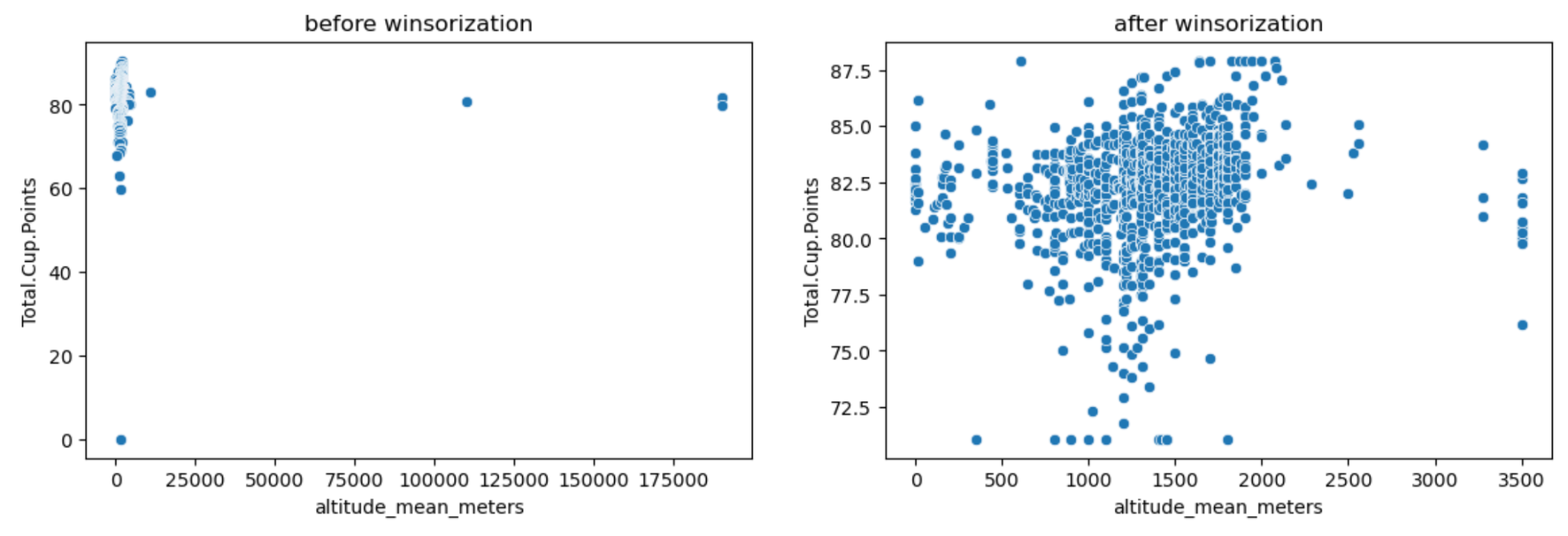
**Data Scaling:**

Some variables in current dataset is highly skewed, the skewness cannot be addressed by standardization, maybe we can consider winsorize them first (for better visualization and prediction).

Eg(1). Total.Cup.Points is very left skewed, the plot below presents the distribution of Total.Cup.Points before and after winsorization

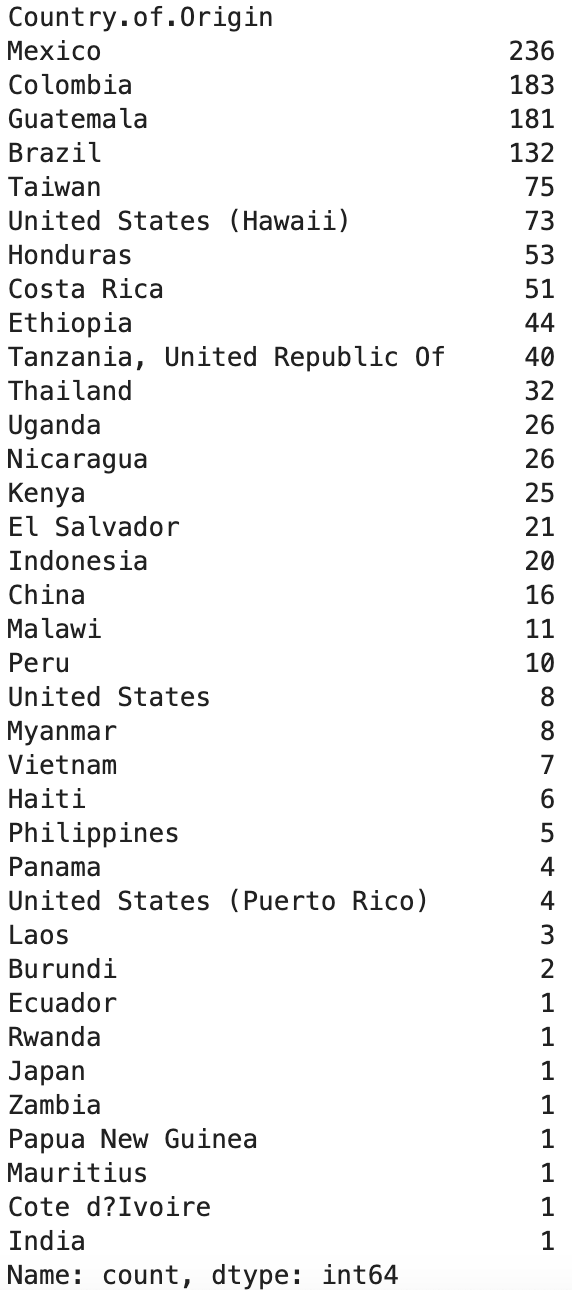


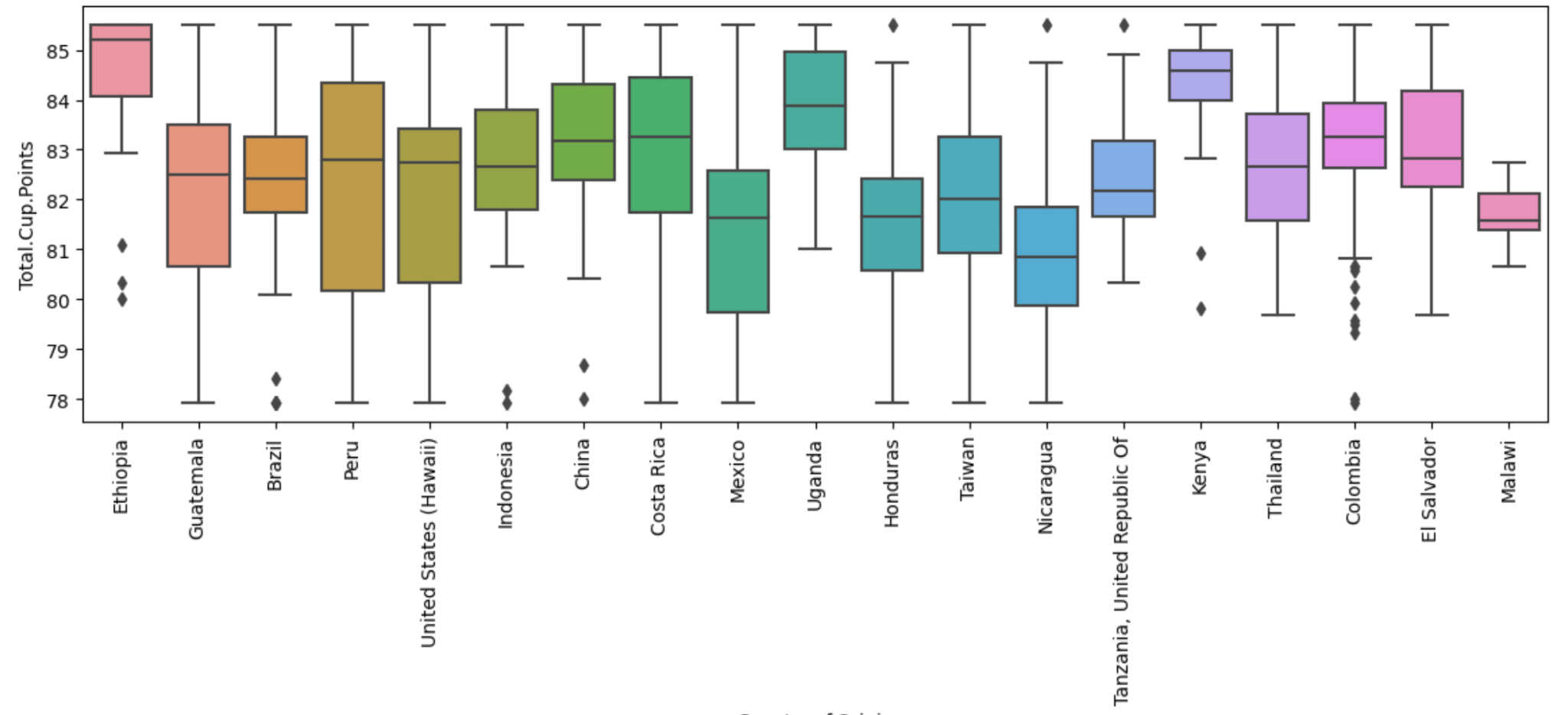
Eg(2) the relationship between Total.Cup.Points and altitude\_mean\_meters before and after winsorization

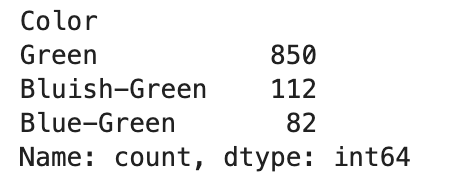


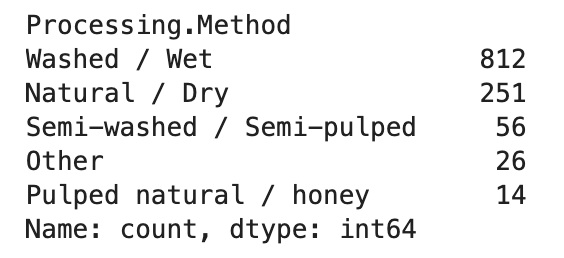
**Data Imbalance:**

1. Countries are highly imbalanced, maybe consider drop countries with rows < 10, or group them together as “Others” (there seems to be some relationship between country and score)
2. Color is also imbalance with most of them are “Green”, but this does not seem to be a big problem, maybe we can just apply one-hot encoding or stratify when split train and test sets.
3. Process method, can consider merge “Pulped natural / honey” and “Semi-washed / Semi-pulped” into “Other”









**Potential Concern:**

If we aim to predict scores using available predictors, the potential predictors are mostly categorical, then we probably will have to include lots of indicator variables in the model, which will be hard to interpret.

**Potential predictors:**

* Categorical: origin of the beans (Country.of.Origin is the most likely one, since other variables such as Farm.Name and Mill have too many categories) and color
* Numerical: altitude, altitude\_mean\_meters, Number.of.Bags, Bag.Weight, years-to-grade-date (Grading.Date - Harvest.Year)
  + I don’t think there are many useful numerical variables, since I cannot think of convincible link btw Number.of.Bags, Bag.Weight, and score. Moreover, there are only 3 years (grading date btw 2014-2016) in the dataset, so calculating years-to-grade-date may not be very helpful. But we can try them nonetheless.

| * Drop Altitude Mean, Altitude Low, and Species * Acidity/Aftertaste/Aroma/Balance/Body/Cupper Points/Flavor/Total.Cup.Points/Uniformity drop the 0 term * Altitude High Meters Convert/Drop to same unit (ft/m)  | 190164 | 2 | 0.2% |  | | --- | --- | --- | --- | | 110000 | 1 | 0.1% |  | | 11000 | 1 |  |  |  * Bag.weight, some in kg, some in lbs, some just numbers * Category.One.Defects interesting to look at * Owner and Owner 1?? Might have to clean and compare |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |



Seems like it might be a web scraping error