The Best Part About Waking Up Is Clusters In Your Cup

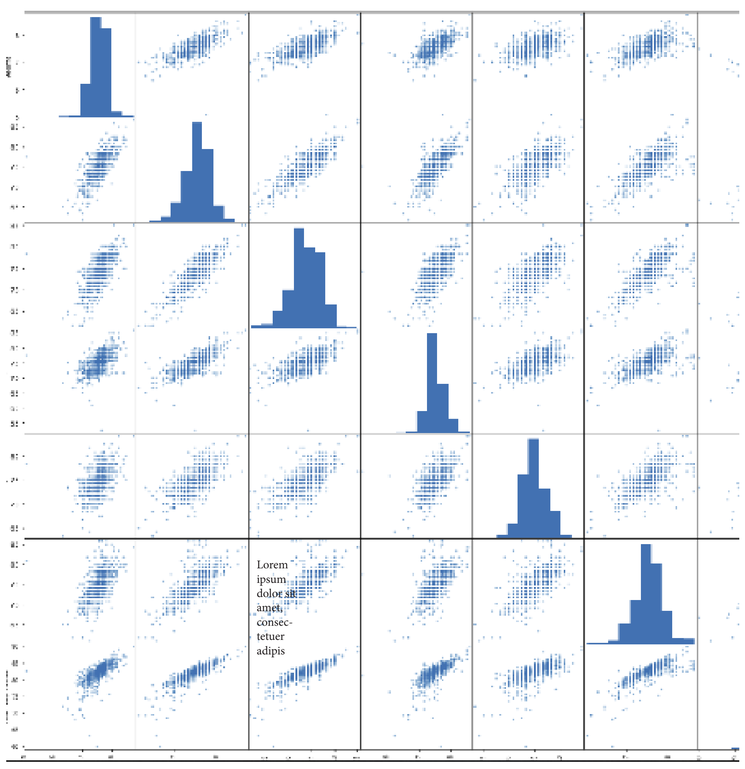
Adam Foster
Jake Kickbush
Team 4

Origins of Our Data and Intro

The specific motivation of our study was to find out whether factors like processing method, country of origin, variety, and color, actually related to the overall score of a specific coffee bean. We decided to drop all rows that had any empty cells, resulting in 407 dropped rows and 0 dropped columns. Our specific dataset includes the following information.
https://github.com/jldbc/coffee-quality-database

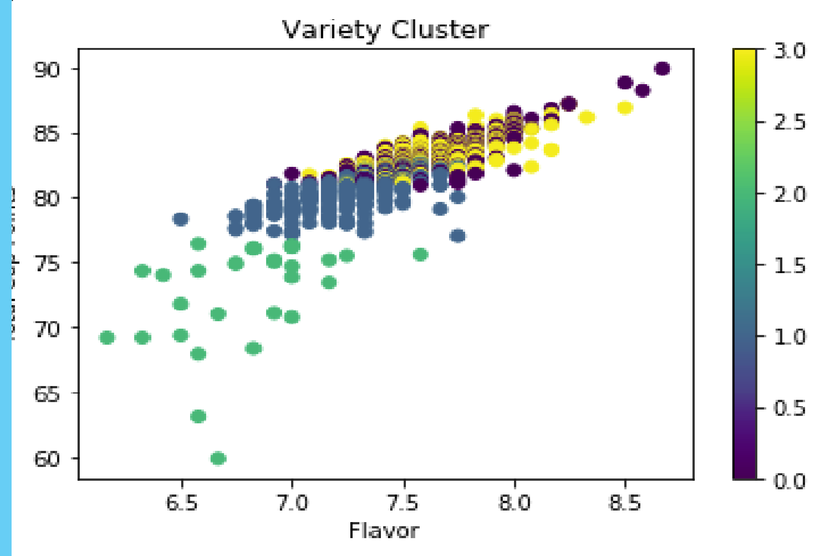
- Reviews of 905 arabica coffee beans from the Coffee Quality Institute.
- Quality - Aroma, Flavor, Aftertaste, Acidity, Body, Balance- Numeric Representation on a 1-10 scale as floats
- About - Processing Method, Color, Variety, Country of Origin as objects
- Rows - 905 Columns - 12

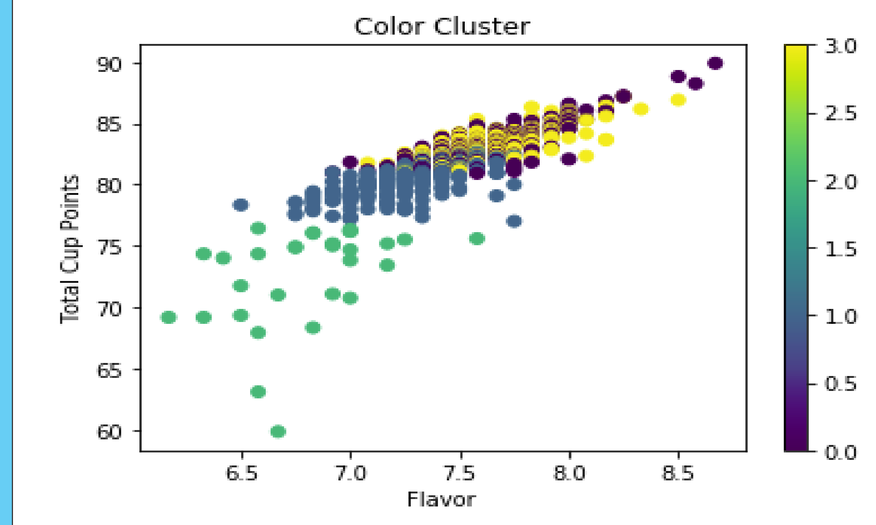
Exploratory Findings

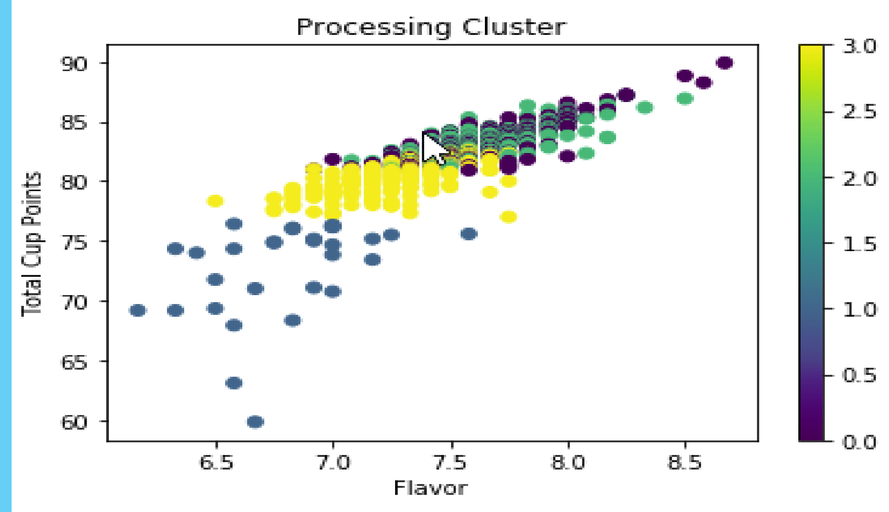


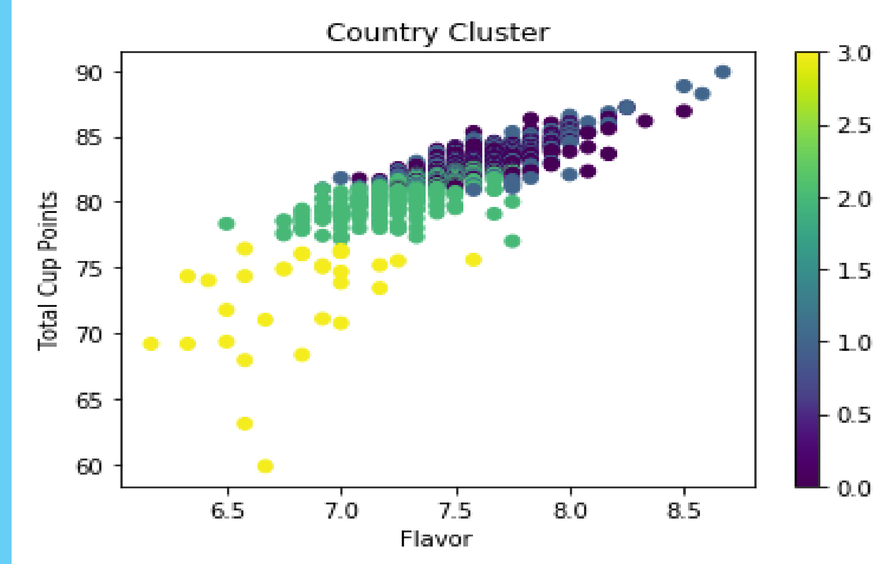
Because we used a scatter matrix to compare the distributions of the different numerical categories, we thought that although slight, the categorical values of the beans must be playing some role in the scoring. This led us to clustering, because we could examine how the categorical data showed a pattern, or lack thereof, in important numerical data, flavor and total cup points.

Findings









Analysis

KMeans Clustering with 4 clusters to see how our different scalar fields(flavor, aroma) correlate with the different categories in our categorical data(Country of Origin, Color)

Results

- Best Variety - Caturra and Other
- Best Color - Green and Other
- Best Processing Method - Semi-washed/Semi-pulped and Washed/Wet
- Best Country - Mexico and Guatemala

As shown in our clusters above, what category the data point belongs to seems to indicate what kind of score the coffee bean will get in any particular scalar variable. For example, We can see that all of the points for individual countries are clustered together, indicating that the country of origin does have an impact on the total cup points and flavor of the bean. Our Other category results are scattered, which reflects our assumption from the data as especially with countries, the Other category includes around 160 different countries

Colorbar Key: 0 - Purple, 1 - Blue, 2 - Green, 3 - Yellow
Variety Key: 0 - Caturra, 1 - Bourbon, 2 - Typica, 3 - Other
Color Key: 0 - Green, 1 - Blueish-Green, 2 - Blue-Green, 3 - Other
Processing Key: 0 - Washed/Wet, 1 - Natural/Day, 2 - Semi-washed/Semi-pulped, 3 - Other
Country Key: 0 - Mexico, 1 - Guatemala, 2 - Colombia, 3 - Other

Limitations and Expansion:

Firstly, our dataset is limited by the fact that we are only looking at one species of bean,
Arabica. There is not data for other species of bean to compare against Arabica, so to further expand on this study we would look to trying to get data on other Coffee bean species to see if the same patterns are reflected there.

Another limitation of this dataset is that it is based on opinion, by the coffee bean reviewers that are creating the data. Opinionated data could skew the legitimacy of the data, especially as we do not know how many different reviewers contributed to the dataset. To further the study in this way, it would help to gain data on the reviewers themselves to prove the legitimacy of the dataset