Proposal for final dataset: The purpose of this study will be to classify coffee into clusters using the Coffee Quality Institute’s data for the purpose of suggesting similar coffees to customers.

Dataset used: <https://github.com/jldbc/coffee-quality-database/tree/master/data>

Dataset description: Pre-cleaned datasets which include traits on 28 Robusta and 1318 Arabica coffee beans as tested by the Coffee Quality Institute with information on the bean’s provenance and flavor profile totaling 44 columns, 33 of which are meta data but eleven are detailing the flavor profile. This data is the result of data crawling in the review pages of the CQI’s website from 2018. There is no data dictionary provided – some assumptions will be made on the interpretation side of things like “acidity” and “sweetness” referring to qualified raters’ opinions rather than some sort of chemical testing.

Number of categorical variables: 2 – Country of Origin and Species

Sample of helpful column names: Species, Country.of.Origin, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Clean.Cup, Sweetness, Cupper.Total, Altitude, Processing.Method

Data Cleansing needed: Not much, there are a number of missing values in Processing.Method – unsure if that variable will be much of use.

Questions to answer:

1. How useful is numerically assessed flavor profile to predicting country of origin? (Maybe there’s a bias in the grading, or maybe coffee from certain countries fundamentally taste differently than others)
2. What useful clusters are there for helping to provide suggestions for other coffees a person may be interested in?

Who would be interested in this information: Any coffee company that wishes to provide suggestions to a customer for coffees either similar or dissimilar to their current coffee, coffee aficionados looking to source coffee for their next cup of coffee, and potentially farmers who want to grow either similar beans or dissimilar beans to provide some natural variation or different taste profiles in their crops.

Modeling techniques used: I anticipate using Random Forest and a hierarchical clustering approach. Random forest will provide me an excellent way to answer question 1, as it will allow me to predict a coffee’s origin country well. Using hierarchical clustering on the second question will allow me to find a coffee that the user has had, and provide suggestions for coffee with similar, or dissimilar, flavor profiles.