Andrew Fridman

EDLD 610

2/6/19

**Final Project Proposal**

**Data source (**https://github.com/jldbc/coffee-quality-database):

I plan to use a dataset housed on GitHub, which includes 1340 reviews of  arabica and robusta coffee beans conducted by the Coffee Quality Institute (CQI). The CQI is a non-profit organization that focuses  on improving not only the quality of coffee, but also all the individuals and communities involved in the farming process across the world. As a result, the organization has compiled numerous pieces of background information, spanning from altitude of the original farm to processing method, as well as the total score awarded to a single cup of coffee produced from the beans. However, I plan to primarily conduct analyses involving the country of origin, harvest year, quality metrics (e.g., acidity, body, aroma, sweetness, etc.), and overall score of the coffee beans.

**Preliminary ideas for visualizations:**

Since this is dataset is such a rich source of international ratings, I would like to display various ratings (e.g., sweetness) by country of origin and world region. For example, I could use a scatterplot and color the points by region, which could help identify clusters of a particular coffee quality (e.g., more acidic coffees tend to come from African countries). I believe that this particular visualization would be a great resource for novices and individuals more familiar with region reputations to identify countries with coffee profiles most similar to their preferences. Additionally, this plot could be modified to distinguish between the two bean species (i.e., arabica and robusta) with different data point shapes.

Similarly, a second visualization could look at the overall ratings by country or world region. I think that this bar graph would be most easily digestible by the layperson, and the data could be stylized as coffee beans to improve appeal. This visualization could also identify a quality coffee powerhouse, and really showcase regional differences.

A third visualization could be a simple species average of quality metrics. For example, I could have a line plot comparison between arabica and robusta across all the measures, and total score (perhaps this is a difference?). While the two bean varieties often are thought to have distinct taste profiles, I am very curious to see if the CQI reviews are consistent. For example, arabicas are thought to be sweeter and more acidic coffees, whereas robustas are considered to have a stronger, harsher taste. This third visualization would once again have layperson appeal and provide a standardized set of ratings for two common coffee bean varities.