Beer Analysis

James D. Wilson

February 15, 2017

In the spirit of San Francisco beer week, we will have our first case study be about beer! The data file that we will investigate is located on the GitHub course site. Load it and let's dig into the data.

The data contains reviews from BeerAdvocate.com for many beers. It includes information such as the ABV, brewery, date of review, location of origin, and style of beer. Get into groups of 5 and answer the following questions:

1) How many different beers are reviewed in this data set? Do you recognize any of them? (Hint: the *unique* function may be useful here.)

2) When was the first and last date of reviews here? (Hint: *date* is a new type of class. It is convenient because it has a natural ordering (e.g., Jan, 2011 is less than Jan, 2012). Thus, mathematical functions can be directly applied to *date* classes.)

3) Which beers in this data set were reviewed in California?

4) What is the highest rated beer in the data set? (Think about how you define "highest" in this case). What is the lowest rated beer?

5) Alcohol Content

a) What style of beer has the highest alcohol content? (i.e. ABV?). (Hint: if you want to use an average, applying the *aggregate* function will be useful here.)

b) Which beer has the highest alcohol content?

c) What style of beer has the lowest alcohol content?

d) Which beer has the lowest alcohol content?

6) Of beers with ABV less than 5%, what is the average, maximum, and minimum rating? Of beers with ABV higher than 5%, what is the average, maximum, and minimum rating? Based on your findings, do you conclude that beers with higher alcohol content are rated higher?

7) How many IPAs are reviewed in this data set? (Note, this includes many different styles including for instance "American Double" and "American India Pale Ale.")

8) Identify which beers are a sort of "imperial" style. This is often used to describe a style of beer. Based on features that are common to the "imperial" beers, what would you say this descriptor suggests about a beer?

9) [Open-ended] Devise two (or more) different ways to measure similarity between beers from this study. Then, separate these beers into at least 5 collections of "similar" beers. Discuss what you find. You should play around with this to see if you find any patterns. This is similar to the problem of clustering in machine learning.

10) [Open-ended] Two really well-known beers from the Bay area are "Pliny the Younger" and "Pliny the Elder", each from Russian River Brewing (in Marin county). First, summarize these two beers based on the reviews given. Next, come up with a way to compare these two beers so that you can settle the dispute of which beer is better. (PS: there are statistical ways of doing this, feel free to do this if you know how.)