# Multivariate Project Proposal

**(Group 1)**

## Introduction

The aim of this project is to explore real data using various multivariate analysis techniques. After considering a variety of datasets the group decided to use a dataset of online beer reviews which was made available by the Stanford Network Analysis Project[[1]](#footnote-1). The reviews were collected by BeerAdvocate[[2]](#footnote-2), an independent community of beer enthusiasts dedicated to supporting and promoting beer.

The data set contains approximately 1.5 million online beer reviews which were collected between 1998 and 2011. A large variety of international beers were scored on a scale of 1-5 (at intervals of 0.5) for the following categories: *appearance*, *aroma*, *palate*, *taste* and *overall rating*. Individual beers can be identified by their name and brewing company. The dataset also contains information about alcohol percentage of the beer, beer style, the time and date from when the review was registered (as a Unix timestamp), profile name of the reviewerand a free text field containing a personal review from the beer consumer. No other information was collected about the reviewer, such as gender, age, or nationality. Figure 1 displays a screenshot of a test review from BeerAdvocate.com.

As well as being of interesting subject matter and free availability, this dataset lends itself well to multivariate techniques as many of the beer attributes are inherently related to each other. For example, it is well known that aroma and taste are related in some way, as well as taste and palate. Including all of the attributes in our analysis will allow us to explore these complex relationships.

## Research Questions

The following research questions are of interest to the group:

1. What characterizes a beer that has a high or low overall rating? We are interested in the variables that significantly affect the overall rating, and their relationship to both overall rating and each other.
2. How prevalent is bias (both intentional and unintentional) in large scale online review websites? We are interested in the following sources of bias that may affect a beers rating:

* Repeated reviews of a beer from a single user
* Consistently extreme reviews from a single user (very high or very low)
* False reviews (e.g. high ratings on the individual attributes but a low overall rating)

1. Can we identify beers that are similar to each other? This would allow beer companies to successfully target potential customers.
2. Can we identify time changes in beer preference? This may reflect seasonal or longer term changes associated with changing consumer preferences or brewing style.
3. Can we identify a gap in the market for a beer, based on consumer preferences?

## Implications

The results from this analysis would be useful to beer brewers, merchants and consumers. Beer brewers would be able to identify the characteristics that lead to a highly rated beer which may improve their current product. They would also be able to identify close competitors in the market and potential gaps in the market. Beer merchants would benefit as they more effectively manage the stock of preferred beer types and recommend alternative beers that are similar to each other. Thus improving their customer offering and enhancing their revenues. Beer consumers could use the results to identify highly rated beers and find beers that are similar to ones that they enjoy. Finally, interested researchers can benefit from our insights into the magnitude of bias on online review sites and the possible types of bias.

## Data Concerns

As is common with online data collection open to the general public there are initial concerns about the validity of the data. Initial exploration suggests some ratings lie outside the 1 to 5 range (at 0), which may correspond to missing data or to a very low rating. Similarly, alcohol percentage is out of the ‘normal’ range for beer in some reviews and missing in others, however this should be easily resolved by investigating the beer itself. With regards to the beer reviewers, there may be abundant bias concerned with promoting beers they are affiliated with and criticising the opposition. For example there are consistently low reviews on different beers by a reviewer with the profile name “dog” and some reviewers have rated the same beer several times. It must be decided how to deal with these reviewers such as by eliminating their review, taking the most recent or averaging them.

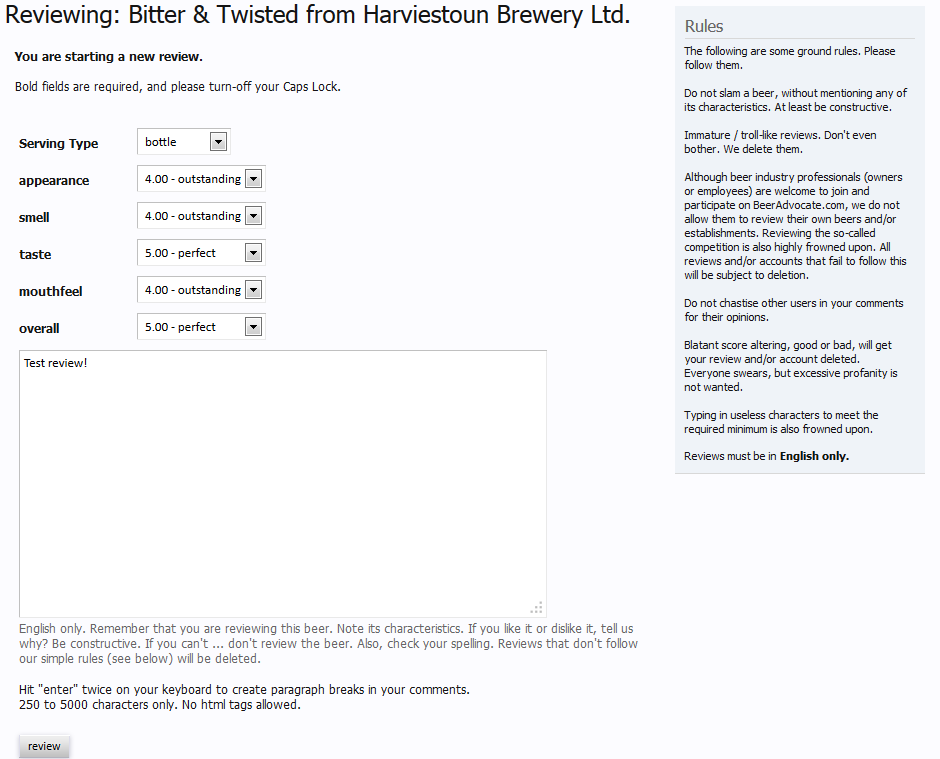


Figure 1: Screenshot of a test review http://www.beeradvocate.com

1. <http://snap.stanford.edu/data/web-BeerAdvocate.html> [↑](#footnote-ref-1)
2. <http://www.beeradvocate.com> [↑](#footnote-ref-2)