Analysis of Beer Ratings and Reviews

Joe Vogelpohl & Daniel Haver
SIADS 591 & 592
University of Michigan School of Information

Table of Contents

All data used for the project can be found at kaggle.com/ehallmar/beers-breweries-and-beer-reviews and is based on user review and rating information collected from BeerAdvocate.com.

All code used for the data analysis has been provided in a separate attachment.

Motivation

Along with the ongoing popularity of craft beer, there has been a steady increase in the amount of data collected from community-driven beer rating sites. With millions of user reviews, consumers, retailers, and brewers are interested in leveraging insights hidden within the data. Results can be used to help consumers find beers that better match their tastes. By understanding consumer preferences, brewers and retailers can use the information to make changes to production schedules, modify delivery plans, or focus efforts on the most preferable beers.

Our project will examine approximately nine million reviews of craft beer published over the last twenty years from BeerAdvocate.com, a website that collects user-based beer reviews and ratings. Ratings data includes scores for taste, smell, look, and other categories.

By examining the data, our project intends to explore the following questions:

- What are the most popular beers and their common review attributes?
- What characteristics of a review will determine if a beer is good or great?
- What changes have been seen in beer reviews over time?
- How does the origin or availability of a beer impact its reviews?

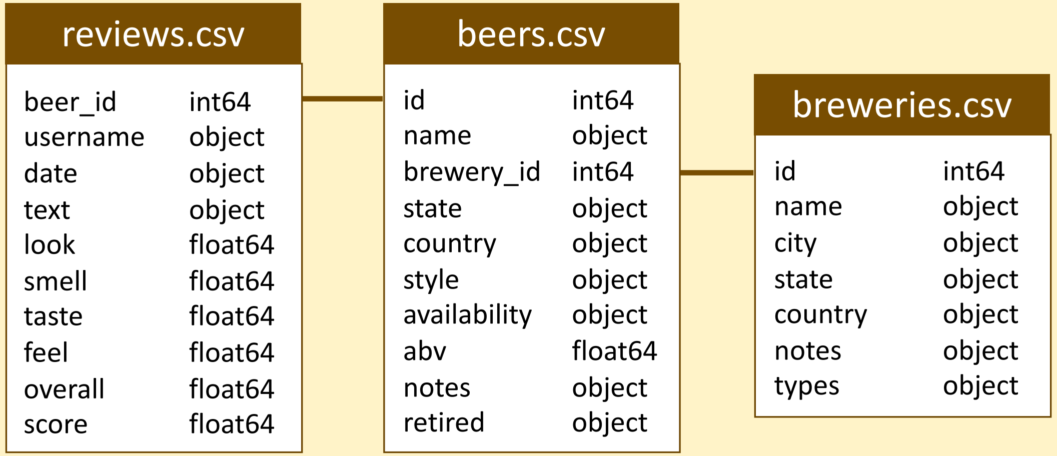
Descriptions of our data manipulation methods and analysis are included in the following pages.

Data Sources

BeerAdvocate was founded in 1996 and is one of the largest community-based beer review sites. Our datasets contain reviews from 1996 to 2018 and are divided into three comma separated value files that were published on kaggle.com. Full links are included in the references. The primary dataset (reviews.csv) includes over nine million user reviews. Each review includes sub-ratings for look, smell, taste, and feel along with an overall rating score. Approximately three million reviews include textual feedback on the rated beers. Usernames, dates, and beer identification codes are also included in the dataset.

There are two secondary datasets. The first (beers.csv) links to the beer identification code from the primary dataset and contains around 360,000 records. It includes beer names, beer identification codes, types of beer, location, availability, and alcohol content. It also provides a brewery identification code used to link to another secondary file with brewery information (breweries.csv).

By linking this file, we were able to get information such as the brewery name, location, and type of brewery. The breweries file includes over 50,000 listings. After some initial data manipulation, the consolidated dataset consisted of over 7.5 million rows with 26 columns.



reviews.csv	
beer_id	int64
username	object
date	object
text	object
look	float64
smell	float64
taste	float64
feel	float64
overall	float64
score	float64

beers.csv	
id	int64
name	object
brewery_id	int64
state	object
country	object
style	object
availability	object
abv	float64
notes	object
retired	object

breweries.csv	
id	int64
name	object
city	object
state	object
country	object
notes	object
types	object

Data Manipulation

Our analysis started by combining the beers and brewery files using a left-join on ‘brewery_id’ in the beers dataset. Each file contained columns for ‘state’ and ‘country’ so, based on an evaluation of missing values, the fields from the breweries files were used (others were dropped). A ‘notes” column from the breweries dataset was examined and was also dropped as it did not contain useful information. The ‘availability’ field had similar categories that needed to be combined due to extra spaces in some of the fields and then a missing data point in the ‘style’ column was added. The cleaned dataset was then merged with the reviews file on ‘beer_id’. The resulting dataset included 7.5 million rows and 26 columns.

When analyzing popular beers and a comparison of good and great beers, dataframe functions such as describe() and value_counts() were used to understand data metrics. Conditionals were also used to create smaller filtered dataframes. During the analysis, natural language libraries were utilized. The libraries tokenized columns of text into a single list of words that were then converted to a frequency distribution. Where needed, non-word characters such as ‘!’ or ‘%’ were removed. Additional libraries were used to classify ‘parts of speech’ to identify descriptive words (adjectives).

The data we used for our time series analysis was first modified by converting the date from an object to a date/time type. We grouped by date to find the change in score overtime but changed to ‘year’ to filter out the noise and allow a better visual representation. With this data we analyzed score, taste, smell, feel and look. For our study of beer style popularity over time we joined beers.csv and reviews.csv using a left-join on ‘beer_id’. We dropped any null values and filtered by the style we were studying. We grouped by year to review the score change overtime. For our study on availability’s effect on review score, we grouped by availability, removed any extra empty spaces, and aggregated over score by mean.

Analysis

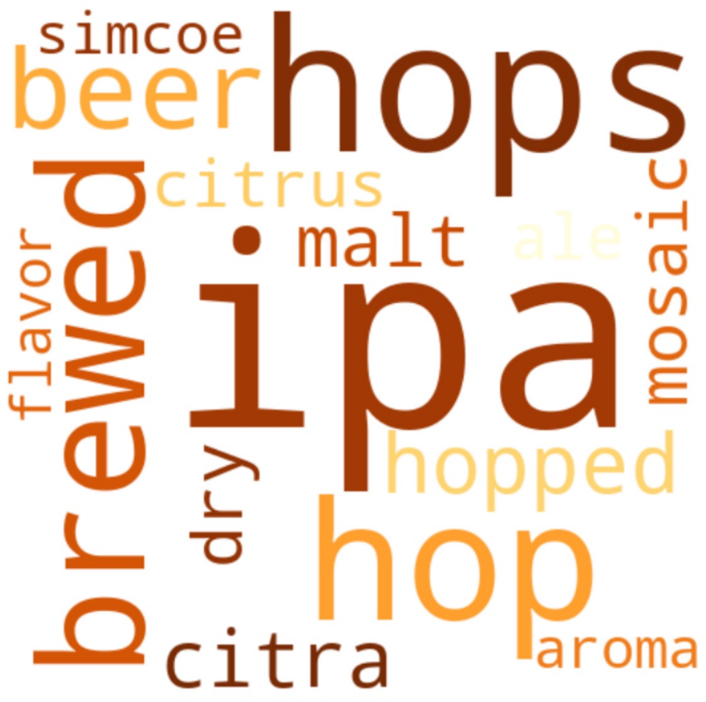
Most Popular Beer Attributes

What are the most popular beers and their common review attributes?

Based on the total number of beers listed, the most popular beer style is the American IPA with 13.5% of all listings. The highest percentage originate from California (24.6%) followed by Michigan (6.4%) and New York (6.2%). 54.2% of American IPAs are available year-round while 28.5% are offered periodically during the year.

Compared with all other beer styles, American IPAs have slightly higher rankings across all categories including look, smell, taste, and feel. The overall mean score is 3.96 which is 0.037 points above the mean for all other styles. On average, the alcohol content for American IPAs is 1% less than all other beers.

Many of the beer listings include descriptive notes. The notes for American IPA were combined and analyzed using natural language libraries. As expected, the word used most in these descriptions is 'IPA'. The American IPAs beer style is known for it's hop flavor, so it was not surprising that both 'hops' and 'hop' are the second and third most used words. The top ten words used in the American IPA beer descriptions are listed in the word cloud. The word usage amount is represented by the relative size of the word.



Most used words for American IPAs

Analysis

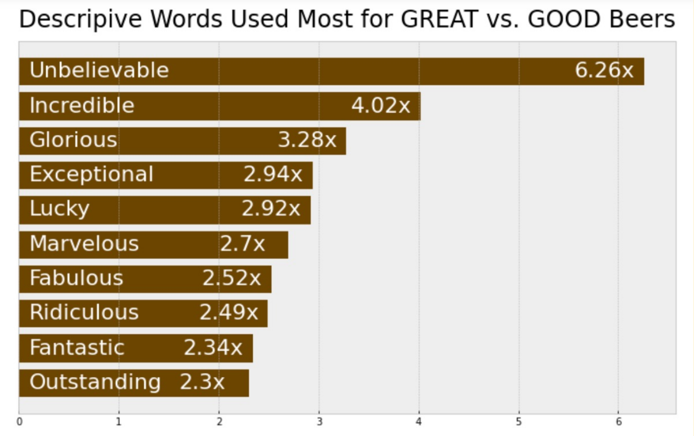
Comparing Good and Great Beers

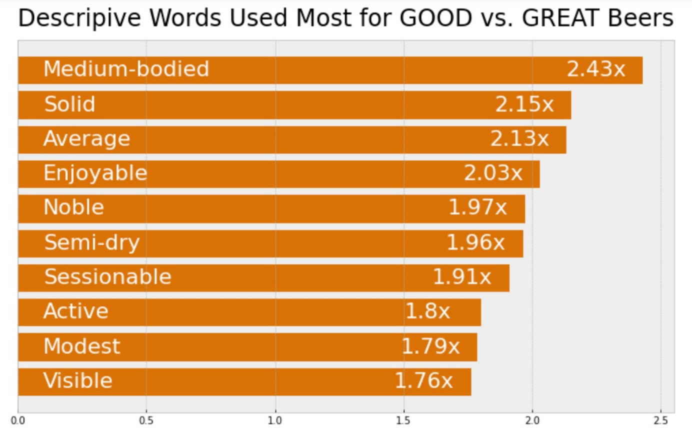
What characteristics of a review will determine if a beer is good or great?

Good and great beers were categorized using a typical grading system. Beers with an A grade (scores in the 90%-100% range) were considered great. Good beers had scores in the 80%-90% range. The first noticeable difference between the two is alcohol content where, on average, great beers are 0.97% higher than good beers. The top three beer styles for great beers were the Imperial IPA, Imperial Stout, and America IPA. Good beers had the same top three styles but in reverse order. All rating metrics were around 0.5 points higher which was expected given the categorization method.

There was little difference in state and availability attributes between the two.

In this analysis, user reviews were investigated using natural language libraries. The frequency counts for all words were calculated and descriptive words (adjectives) were identified. The words were then compared to see which had the highest usage ratio between the two categories. For example, the word ‘unbelievable’ was used 6.26 times more in great user reviews compared with good reviews. Results were calculated for both groups and are included below.





Analysis

Reviews Changes Over Time

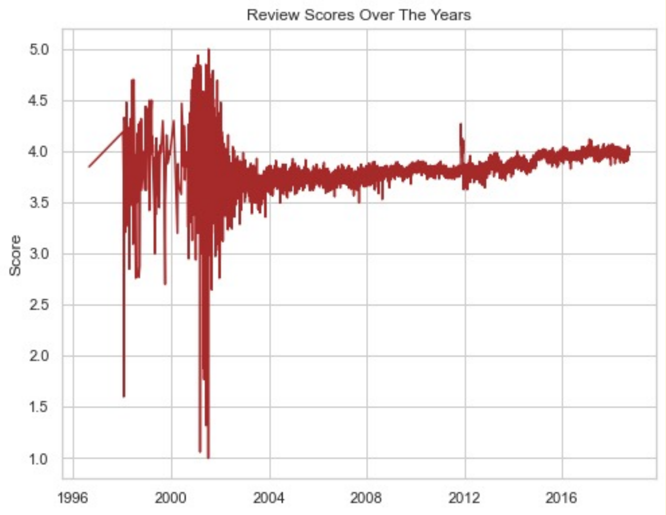
What changes have been seen in beer reviews over time?

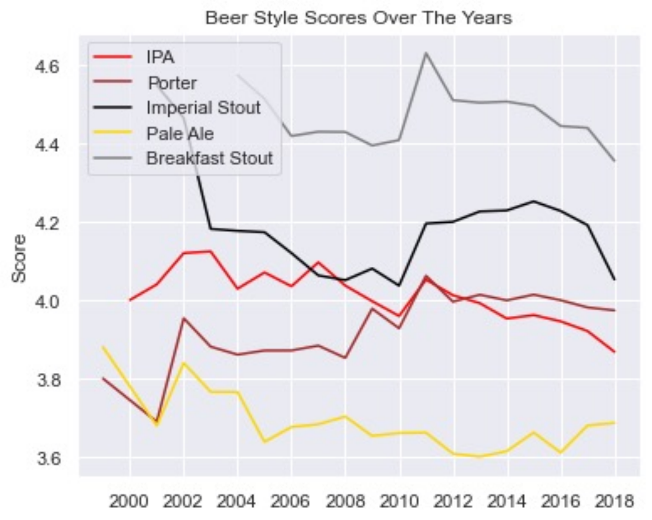
The reviews collected by BeerAdvocate show a much wider range of mean score in the early years of their existence due to a smaller amount of people posting ratings.

A larger pool of reviews averaged out gives us a tighter range that is more representative of the entire craft beer consuming crowd.

A slight uptrend in the graph from around 2002 until 2018 illustrates our position that beers have been increasing in overall quality since the early 2000s.

The popularity of craft beers has spawned more micro-breweries, more connoisseurs and stronger competition that is also supported by a larger market. When a company launches a beer that becomes popular, its style receives a positive spike. That specific beer is then replicated by competitors which eventually leads to a dilution of the style's spike as the novelty wears and lesser competitors saturate the market.





Analysis

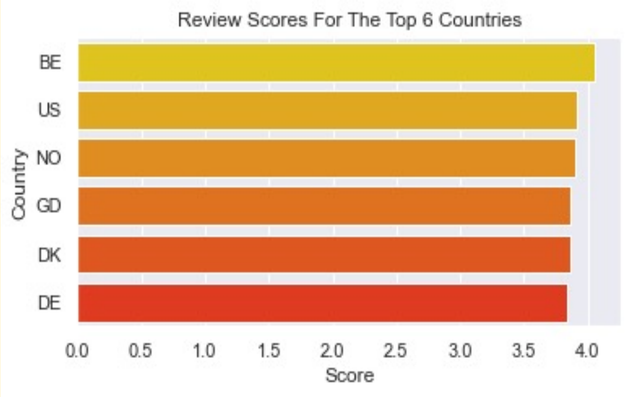
Impact of Origin and Availability

How does the origin or availability of a beer impact it’s reviews?

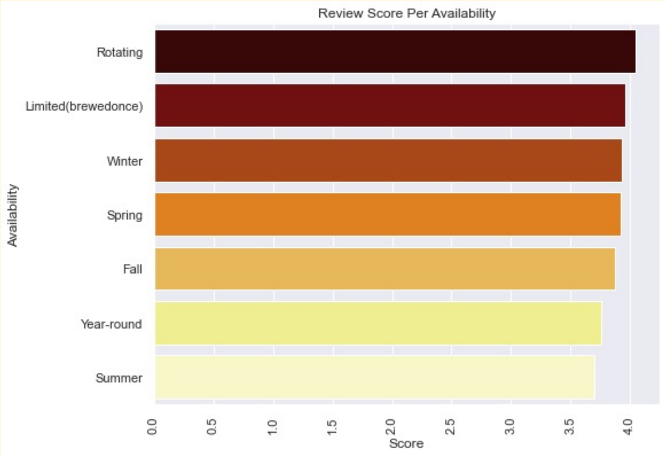
BeerAdvocate has collected reviews from 186 other countries. We looked at the top six countries rated by their average beer score.

As the birthplace of many beers, Europe dominates the list with four countries represented. Belgium places the highest in the list with a 4.06 out of 5, making it the only country with an average score over 4.. Norway (3rd), Denmark (5th), and Germany (6th) round out the remainder of the European countries. The United States and Grenada come in at 2nd and 4th representing the Americas.

Beers which originated in Europe or are still produced in Europe cast the biggest influence amongst reviewers.



With availability we see that Rotating and Limited (brewed once) beers hold the highest ratings. Limiting a beer’s availability to the public increases the ratings score by making it more unique and eventful. Winter beers are heavier with higher alcohol content while Summer beers are light with less alcohol content. This follows the trend of higher alcohol content beers being scored higher.



Statement of Work and References

Statement of Work:

The project was a collaborative effort. We communicated via email and slack and had multiple zoom sessions over the eight weeks. We worked together to build the project proposal and then independently examined the datasets. Feedback from our analysis was combined in order to start from a common starting point. The proposed questions were worked on independently, two for each team member and results were shared. Work for the final proposal was divided into different sections that were completed by each member. The final slides were consolidated and reviewed by each for final structural and grammatical changes.

References:

Data Sources: kaggle.com/ehallmar/beers-breweries-and-beer-reviews
CraftBeer: https://www.craftbeer.com/styles/american-india-pale-ale
BeerAdvocate: https://www.beeradvocate.com/