Springboard Data Science Career Intensive Capstone Project

Bottoms UP – Beer Rating Analysis

By: Ajay Sampath

Mentored By: David Yakobovitch

Jan 17, 2018

1. Introduction

Do you like beer? If so, are you curious about which country produces the best beer? Have you tasted the best rated beer in the world? This report analyzes 220,000 beers from all over the world to answer some of these questions and more.

Beer is one of the most preferred drinks in the world, especially in the United States and in quite a few countries in Europe. However, the question arises as to why would anyone care about the ratings for a beer. That question is answered by the following key points:

- Beer connoisseurs are always interested in finding out more about beers – qualitative data or analytics
- Imagine you are starting out as a brewer. You would want to take a bet on the safest winning combination that would get you beer rally high ratings. For example, you need to know the optimal alcohol content, the style of beer (say lager, stouts or IPA's?) and more importantly the local culture. Do people drink beer in your country or state?
- Restaurants can also make decisions on what type of beer they should stock and advertise if they indeed have the best rated beers in the world.
- If you are not a beer enthusiast, brewer, or a restaurant owner, you can still use insights from the analysis for an interesting cocktail conversation!

2. Data

Data for the analysis was obtained from the ratebeer.com website. The methods used for obtaining the data, cleaning and getting the data suitable for analysis are discussed in the subsequent sections.

2.1. RateBeer Website

RateBeer.com is widely recognized as the most in-depth, accurate, and one of the most-visited source for beer information. RateBeer is a world site for craft beer enthusiasts and is dedicated to serving the entire craft beer community through beer education, promotion and outreach. Established and maintained by dedicated volunteers, RateBeer has become the premier resource for consumer-driven beer ratings, features on beer culture and industry events, weekly beer-related editorials, and an internationally recognized, annual RateBeer Best competition. A vibrant

community of hundreds of thousands of members from more than 100 countries have rated hundreds of thousands of different beers around the world.

2.2. Data Acquisition

The ratebeer website provides an API key for beer enthusiasts who are interested in developing beer apps in partnership with the website and for students who are interested in general data analysis research. For this project, an API key was obtained under the academia agreement with RateBeer.

The API documentation is provided at this link: https://www.ratebeer.com/api-documentation.asp.The data can be downloaded using the API key as json files. The data can be downloaded by ratings, 100 beers a time. A total of 5,000 calls are allowed per month. A for loop was written in python to make 100 calls at a time to avoid overloading the server. The details for data acquisition are provided in this ipython notebook.

2.3. Data Cleaning

The dataset contains 220,000 beers from around the world. Each beer has 18 variables associated with it such as id, description, abv, name, style, style score, overall score, brewery name, brewery type, rating count, average rating, street, city, state, country, continent, twitter link and Facebook link.

The cleaning process involved primarily inspecting and filling the missing values in the required columns. The columns such as beer description, street and continent were dropped during the cleaning process. The 'beer_description' has been considered in the second phase of the project for a preliminary NLP analysis. The street columns were dropped since the analysis does not require that level of granularity. Binary columns for social media were created to analyze the influence on average ratings.

The missing values in the numerical columns were coded as zero values and were filled appropriately using standard statistical methods. A detailed description of the data cleaning process is available in this ipython notebook.

3. Exploratory Data Analysis

This section takes an in-depth look at some of the variables in the dataset. The following are some of the questions the analysis will try to explore:

- Which countries have the most number of beers?
- Which countries have the best beers? A beer will be considered 'best' if it has an average rating greater than 3.5.
- Does the country of origin control the ratings?
- What are some of the key parameters that affect the average ratings? ABV? Overall Score? Beer Style?

The analysis will provide insights into the variables both qualitatively and quantitatively. The average rating is the target variable and the effect of the following variables will be analyzed in detail:

1. Country
2. Alcohol by Volume %
3. Rating Count
4. Social Media
5. Beer Style
6. Brewery Type

3.1 Country

The dataset contains beers from 209 countries. There are some very interesting and not so commonly heard countries in the list such as Abkhazia, Togo and Turkmenistan. There are a few beers from North Korea too! The most interesting is Transnistria, which has its own government but is still not recognized as an independent country by the UN. There are 13 beers in the list from here.

Figure 1 shows the distribution of beers by country in the dataset. The 15 countries shown in the plot account for 87% of the beers in the dataset and will be used for further analysis. There are countries with just 1 beer. It will be hard to get any meaningful statistics for comparison.

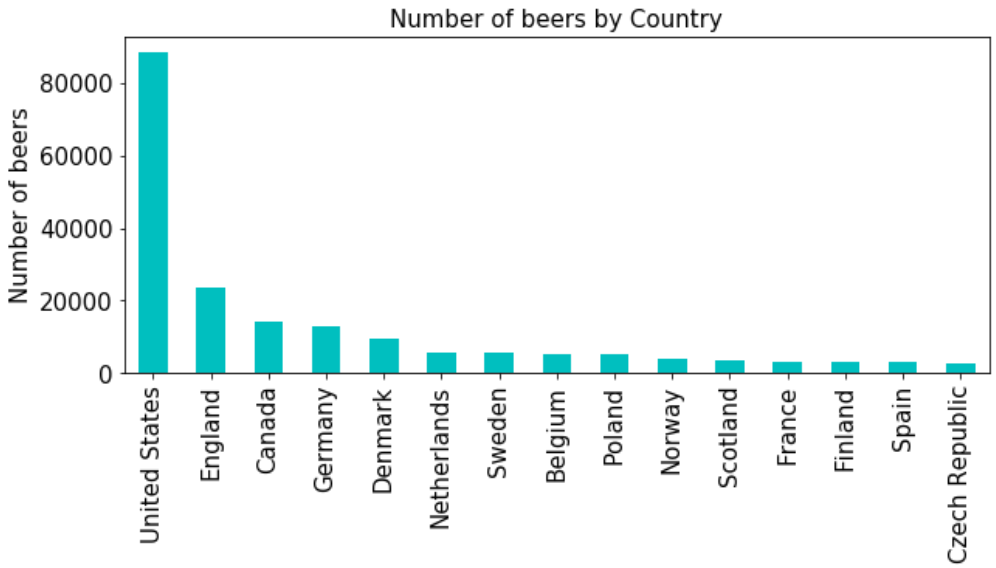

Figure 1. Distribution of beers in the dataset by country

There are 88,228 beers from the United States (approximately 40% of the dataset). It is surprising, that countries like Australia do not produce enough beers.

The data was obtained from a US based website, which may consequently lead to more beers from the US in the site. However, thorough inspection of the ratebeer website shows that there are 'many' reviewers from all over the world, which shows that the website is quite popular in other countries too. It may just be that beer is the most preferred drink in the US, resulting in more number of producers.

Figure 2 shows the percentage of beers with an average rating above 3.5 (a beer can be assumed good if it gets a rating above 3.5).

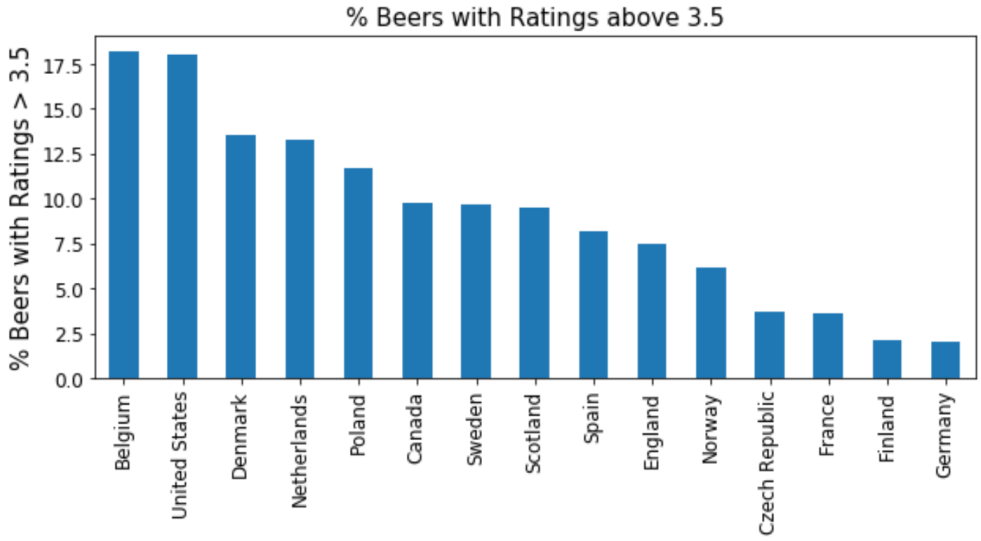

Figure 2. % Beers with Rating > 3.5 by Country

Belgium and United States have a high percentage of beers rated higher than 3.5 and may indeed produce the best beers in the world. The ratings for beers from Germany are surprising. It has one of the lowest percentage of beers (approximately 2%) rated above 3.5. It is a well-known fact that Germans are prolific beer drinkers and host the popular Oktoberfest. There may be some bias in reviewer rating! It is interesting to note countries like Poland have a higher percentage of beers rated above 3.5 than Germany. Other factors such as ABV, rating counts, overall scores etc. may have a higher influence and need to be analyzed to better understand the ratings.

The mean value of the average rating for each country is shown in Figure 3.

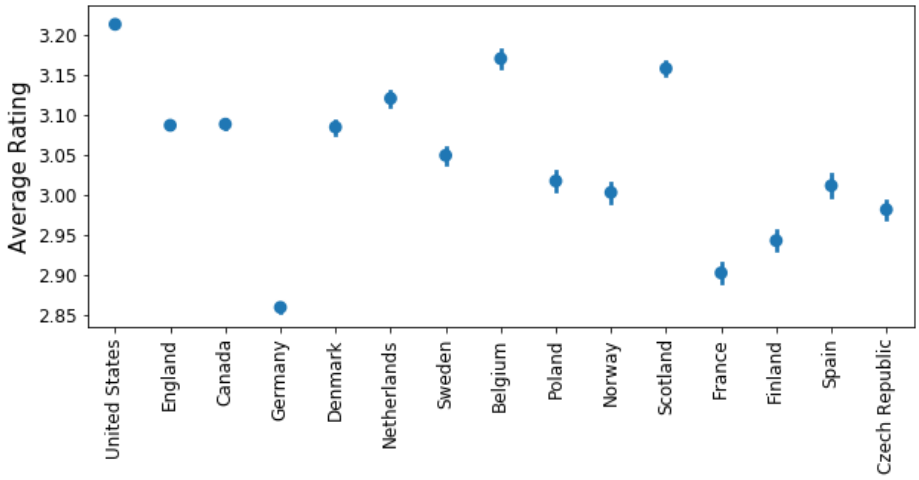

Figure 3. % Mean beer rating by country

United States and Belgium have the best mean average ratings for beers. Germany has the worst average rating among the top 15 countries selected for the analysis. The average rating is likely affected by the number of ratings each beer receives on the website. The combined overall total number of ratings for beers from each county is shown in Figure 4.

The total count for overall ratings is significantly larger for the United States compared to the other countries. This, combined with the highest mean average rating indicates that the United States produces some of the best beer in the world.

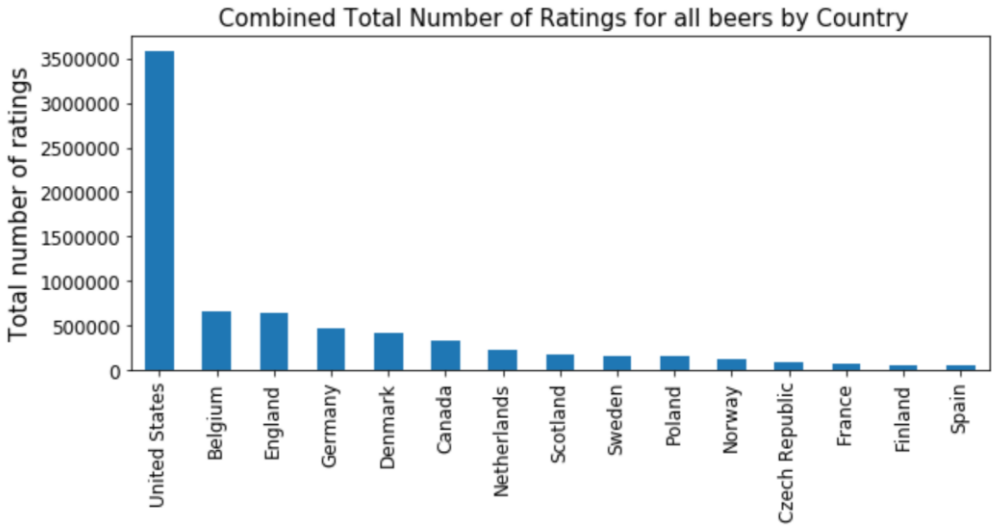

Figure 4. Combined total rating count by country

The beers from the United States were analyzed separately to get insights on distribution and ratings across the different states. Figure 5 shows the beer distribution by states.

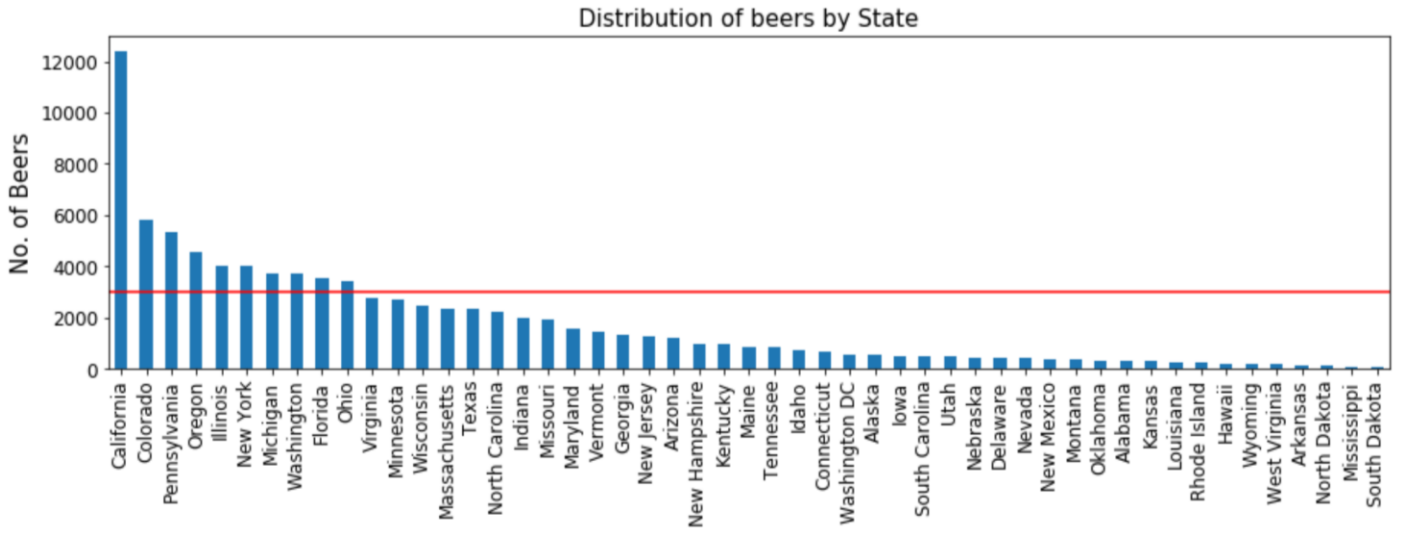

Figure 5. Beer distribution by states in the US

California has the most number of beers, which is not very surprising. The ABV, rating counts and the average ratings are compared in Figure 6 for the states with more than 3,000 beers.

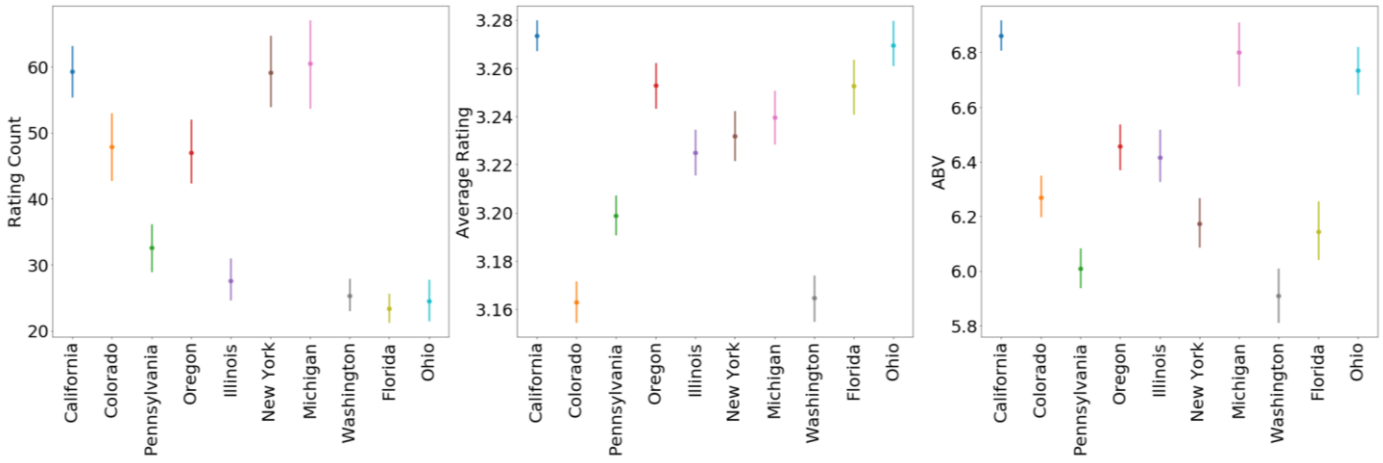

Figure 6. Parameters comparison for beers from top US states

Beers from California and Michigan have a high mean ABV, average rating and rating count. However, the parameters are not correlated for beers from other states. It is likely that the average rating depends on all the parameters combined.

3.2 Alcohol by Volume % (ABV)

Distribution plots for the ABV showed significant kurtosis and quite a few outliers. Some beers have an alcohol content as high as 70%! A scatter plot for the ABV (no outliers) and Average Rating is shown in Figure 7. The figure shows no significant relationship between the two parameters. A spearman test showed a weak positive correlation coefficient of 0.44 that was statistically significant. Similar results were observed for beers only from the United States and California.

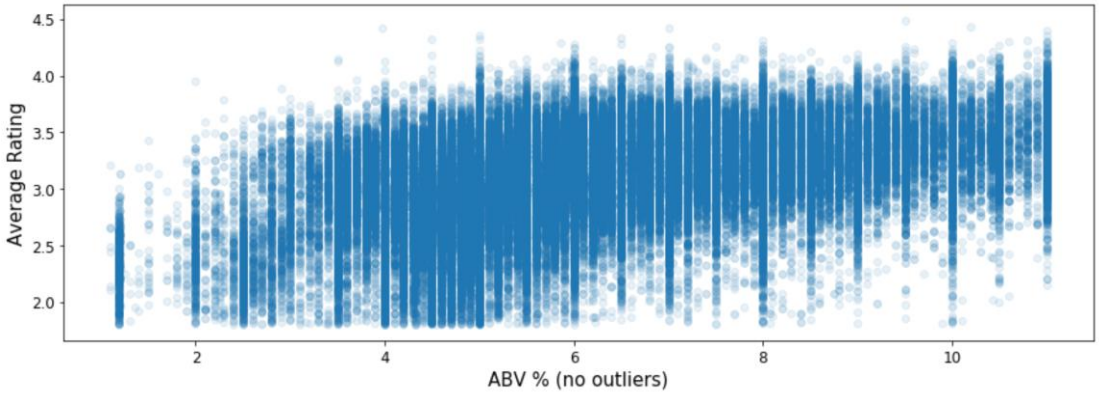

Figure 7. ABV vs Average Rating

3.3 Rating Count

Rating count can be useful to answer questions such as if more people are rating the beer (i.e. drinking the beer), does it lead to better ratings. This can be useful for breweries to make marketing decisions such as investing in social media to get the word out there.

An analysis of the rating count showed a very high variance and standard deviation in the data set. Some beers had rating count as high as 4,600 and the mean was approximately 35 with a standard deviation of 127. However, there were only 4,855 beers out of the 220,000 with a rating count more than 250. Figure 8 shows the rating count (less than 250) vs the average rating.

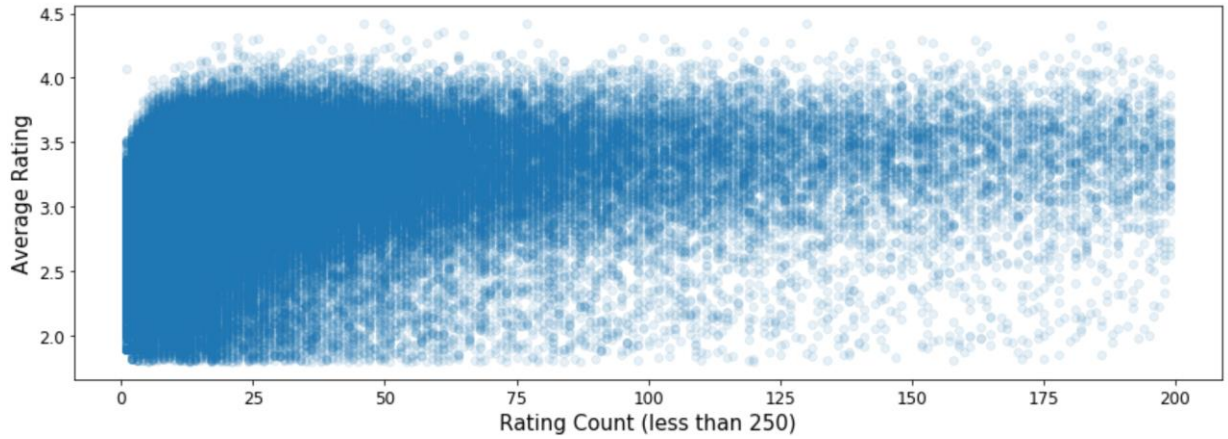

Figure 8. Rating Count (less than 250) vs Average Rating

There is no clear trend in the rating count vs average rating for beers from all countries. A spearman test indicated a weak positive correlation of 0.35 that is statistically significant.

3.4 Social Media

The dataset contains two binary columns that 'has_facebook' and 'has_twitter' that were created during the data cleaning process, which indicates the social media account availability for each beer. Another column called 'SocailMedia' was created that indicates availability of either social media. Figure 9 shows the number of beers with a social media account.

About 74% of the beers have social media accounts. Looks like the breweries are already getting the word out there! Although, more number of breweries prefer a Facebook account to a twitter account. Quite a few of them have both.

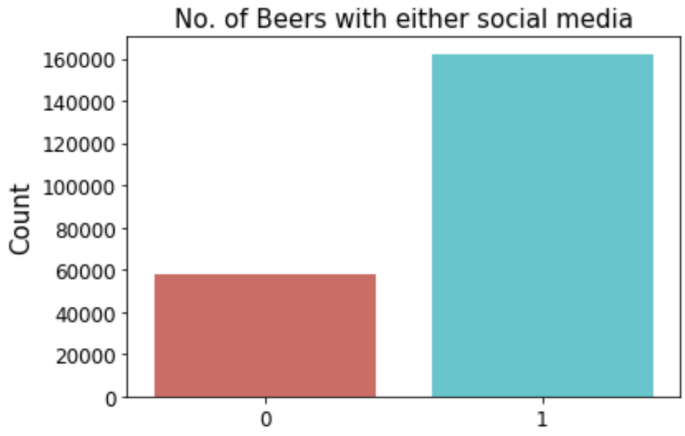

Figure 9. No. of beers with social media

A closer look at the data revealed that the top two beers have neither social media accounts and one of them has 3,332 total ratings. A search on the internet did indicate that the top two beers do not have social media accounts indeed and the data was not missing! In fact, the second ranked beer (Westvleteren) is brewed at a monastery in Belgium and is only available there. This specialty of the beer may have resulted in more people trying the beer, thereby resulting in a high rating count.

Figure 10 shows the comparison of the average rating for beers with and without social media. The average rating is somewhat higher for beers with a social media.

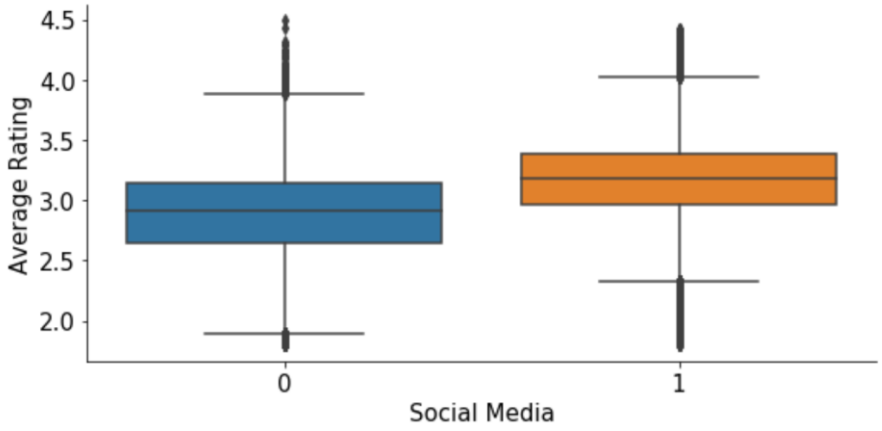

Figure 10. Average Rating vs Social Media Availability

3.5 Beer Style

Beer style is another parameter that can influence the average ratings. There are 94 different beer styles in the dataset. There are sub-styles within a primary style such as

Lambic Style (Faro), Lambic Style (Unblended) and so on. Since the reviewers score on parameters like taste and aroma, the sub-styles have not been combined for analysis. The top 10 styles (total number of beers) are shown in Figure 11.

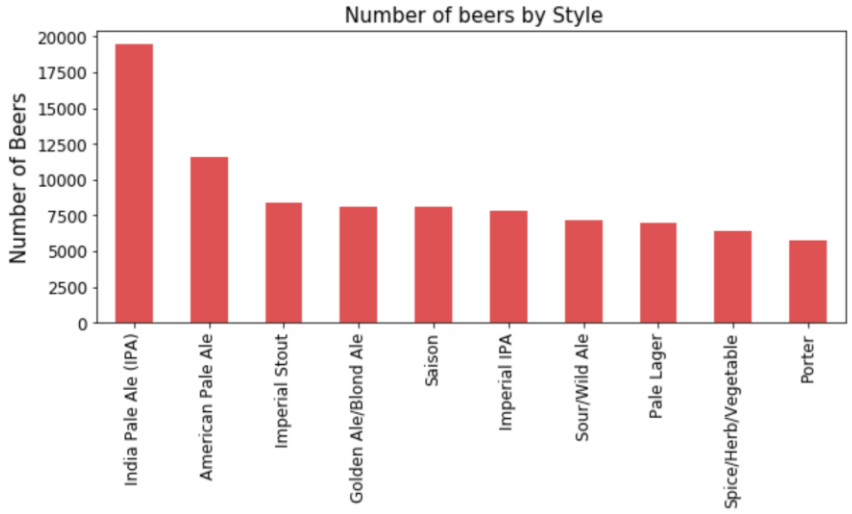

Figure 11. Beer Style Distribution

The pale ales dominate the list followed by the imperial stouts. The dataset also indicated that the pale lagers have the worst mean of the average rating and the imperial stouts have the best ratings. The ABV vs average rating for the styles are compared in Figure 12.

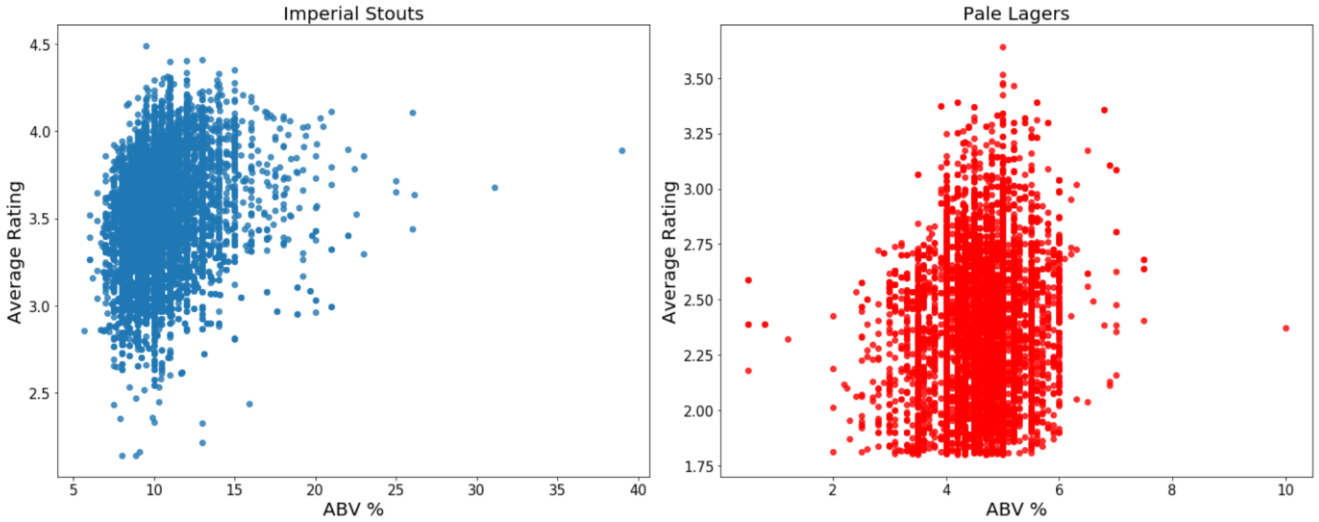

Figure 12. ABV vs Average Rating for beer styles

There is no significant trend within the styles. This indicates that the average rating may not be dependent on any one individual parameter, but a combination of all the variables or may even be completely independent of the all the variables.

3.6 Brewery Type

Figure 13 shows the number of beers for each brewery type in the dataset. More than 50% of the beers in the dataset are from microbreweries. Further analysis indicated that the United States has the most number of microbreweries. This is not surprising since there are a lot of young savvy people starting microbreweries in the United States these days.

The statistical parameters such as mean, standard deviation etc. for the rating count, average rating and abv are comparable for beers from microbreweries and beers from all other breweries. Further analysis has not been done on brewery type during this phase. However, the types have been considered for prediction modeling discussed on the subsequent section.

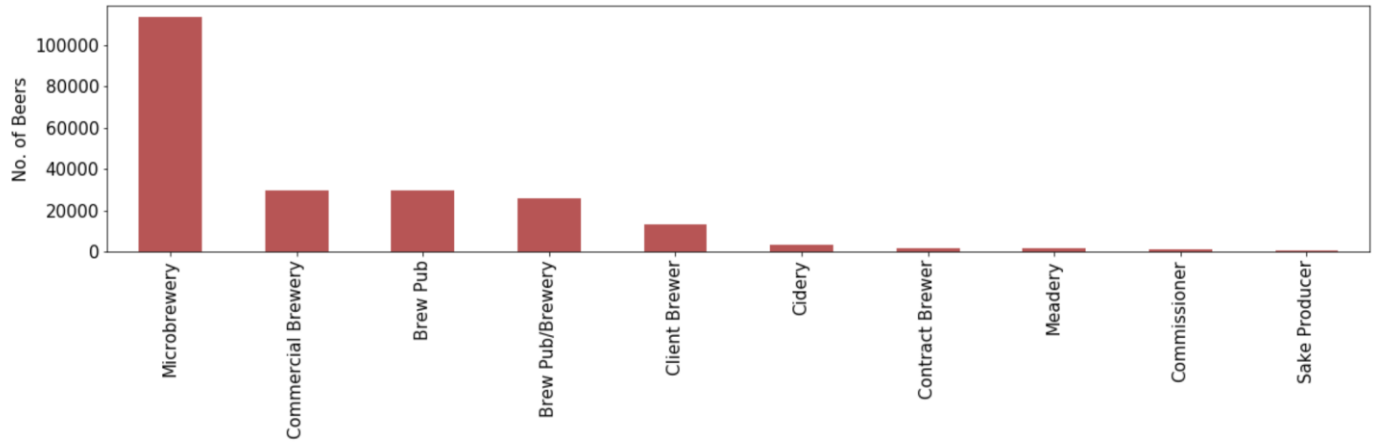

Figure 13. Brewery type vs number of beers

4. Machine Learning Model

The problem has been treated as one that of classification. A beer can be considered good and in the top 200 if it has a rating greater than 4. During the first phase of the modeling, two classification models (k-nearest neighbor and logistic regression) were applied to predict if a beer from the United States will have a rating greater than 4 based on beer styles, abv, rating count and social media. Dummy variables were created for the beer style column as part of pre-processing the data for the model. The data was split in to training and test using the train_test_split module from the sklearn package in python. The training data contained 70% of the entire data set.

Both the k-nearest neighbor and logistic regression model performed well in classifying the beers based on a rating greater or less than 4. The k-nearest neighbor method showed a precision of 0.99 and the logistic regression had an AUC score of

0.99. The confusion matrix and the ROC curve for the two models are shown in Table 1 and Figure 14.

	precision	recall	f1-score	support
0	0.99	1.00	1.00	26293
1	0.00	0.00	0.00	176
avg / total	0.99	0.99	0.99	26469

Table 1. Confusion Matrix for k-nearest neighbors

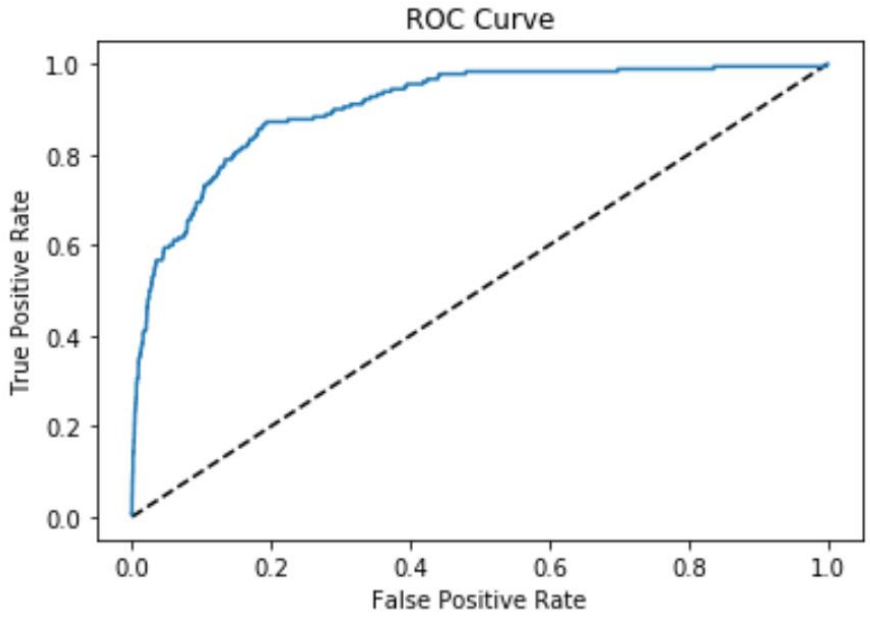

Figure 14. ROC curve for logistic regression – AUC score = 0.90

The logistic regression model was also applied for the entire data set for all beers. The model was applied to predict if a beer will have an average rating greater than 4 based on the rating count, beer style, brewery type, country and availability of social media. Here again the model was split into training and test data sets with 70% of the data in the training data set. Interestingly, the model performed better than the previous model and yielded an AUC score of 0.95.

The next phase of the modelling is ongoing and will involve applying clustering methods to predict country labels. NLP methods will also be applied to get insights on the beer description column of the dataset. The report will be updated with the results for the final deliverable.

5. Conclusions

A dataset with ratings for 220,000 beers was analyzed. We have explored columns such as country, abv, beer style, rating count and social media. Following on some of the key observations:

- United States has the largest number of beers in the dataset and also has the highest percentage of beers rated above 3.5. The biggest surprise was Germany with less than 2% of the beers rated above 3.5
- 2. The beers from the United States were further analyzed. Oklahoma has the highest percentage of beers rated above 3.5. California has the most number of beers and approximately 3,173 beers have been rated above 3.5.
- 3. The alcohol by volume percent has a positive correlation with the average rating. An ABV of 10% typically rated high.
- 4. The rating count also influences the average ratings positively. It is important to get more to people to rate the beers.
- The influence of having a social media account on the rating count was analyzed. Most of the breweries prefer to have a Facebook page compared to twitter. In general, having a social media account gets more ratings, which will increase the average rating.
- The imperial stouts look like the preferred beer style. These also have the most percentage of beers rated higher than 3.5. This is interesting because, the imperial stouts have an average ABV of 10% and most number of ratings.

Based on the analysis it is safe to say that if you are starting out as a brewer it will be best to focus on Imperial Stouts with an ABV of 10%. And of course, it is worthwhile to invest in social media to spread the word about your beer. Facebook or twitter up to you!

6. Future Work

The rate beer website has actual text-based reviews for each beer from reviewers all around the world. It will be worthwhile to get permission from the website to scrape this data since it is not readily available with the api and apply a NLP model to get valuable insights.

There are other sources of data from websites such as BreweryDB and beer advocate that have useful information. It will be good to analyze that data as well.