The Brewery Project

Mo Shukla

Abstract

The US domestic beer market is undergoing a rapid transition with a significant growth in the craft brewery section. American brewers are experimenting and diversifying brewing styles, and these are being well received by the general public. This project seeks to understand the qualities that distinguish breweries producing some of the best-rated beers. This information can guide the decision making of potential investors looking to expand into the growing beer market.

Keywords

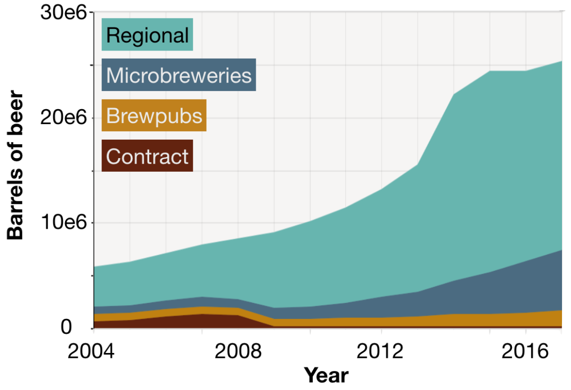
Craft beer — Brewery — Machine Learning — FourSquare — BeautifulSoup

Contents

1. Introduction and Background

The American beer landscape is changing. According to the Brewers Assocation[1], craft beers now account for nearly a quarter of the US beer market. Figure 1 shows the growth of *craft breweries* – small and independent brewing companies – since 2007[2].

The increase in the number of craft breweries is marked by a corresponding rise in general interest in craft brews. Figure 2 shows the rise in Google searches for "craft beer" from 2004 to 2018[3]. It would be safe to say that this is the decade of the Craft Beer Revolution.



Source: https://www.brewersassociation.org/statistics/national-beer-sales-production-data/

Figure 1. Growth of craft breweries in the US.

Overall, the beer market in the US shows a decline, as measured by the number of barrels of beer produced annually. Figure 2 shows the taxable production of beer in the US between 2007 and 2017, as reported to the US Treasury Department's Alcohol & Tobacco Tax & Trade Bureau[4]. Nevertheless, craft breweries continue to grow; current figures on Brewers Association indicates a 5% increase in craft beer production, despite an overall -1.2% change in overall beer volume.

Taken together this data suggests that there is a new Craft Beer trend sweeping across the US, and that this is a great time to be brewing craft beer in the country.

2. The Problem

Imagine an investor looking to dive into the craft beer sector and open a new brewery. As with any business endeavor, there are a host of unknowns, including the obtaining of raw material, finding experienced brewers, installation of the industrial

[1] https://www.brewersassociation.org/statistics/national-beer-sales-production-data/

[2] Figure adapted from National beer sales and production data, Brewers Association

[3] Data gathered from Google Trends.

[4] https://www.ttb.gov/beer/beer-stats.shtml

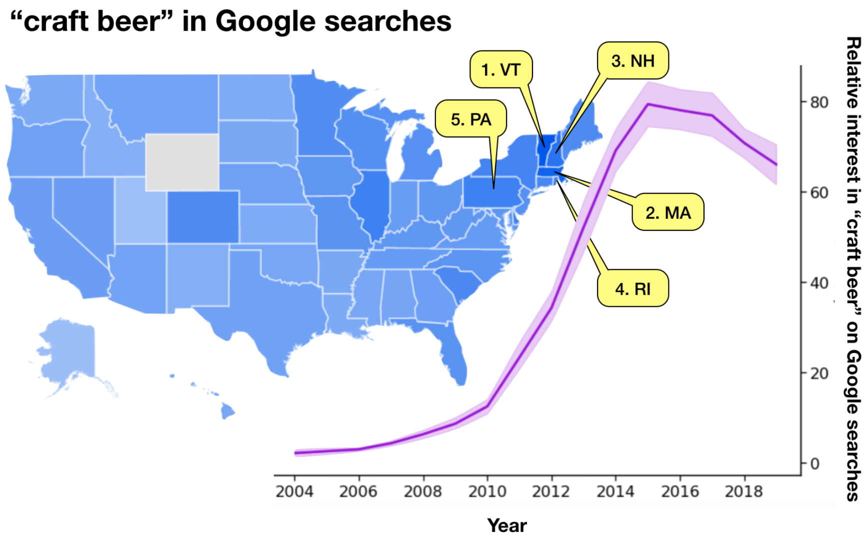


Figure 2. Google searches for “craft beer” between 2004 and 2008. The graph shows the *interest*: number of searches relative to the peak within the period, for each year, aggregated by all months in that year. States are color coded by the relative *interest* in ”craft beer” for this period, with the top 5 states highlighted.

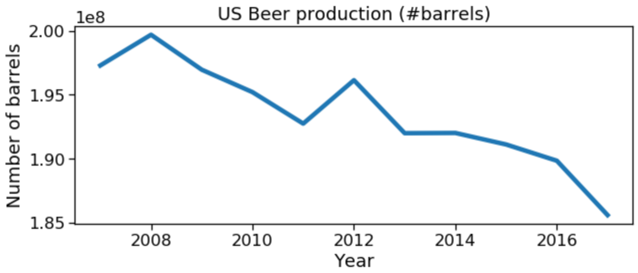


Figure 3. US Beer production (number of taxable barrels) by year. Data gathered from https://www.ttb.gov/beer/beer-stats.shtml

machinery that constitutes a modern brewery, local regulatory constraints, and so forth.

However, above all these issues is the question of the product itself, the brew. Knowing what makes a good brew can be used as a tool to guide the decision for establishing a new brewery – the brewery should share characteristics of other breweries that have been making beer that is consistently appreciated by the drinkers.

What makes a good beer? The very definition of beer and brewing is changing, particularly in the US, where there has not been a long and deep tradition of brewing. For example, the German brewery Weihenstephan, founded in 1040, is nearly a 1000 years old[5], while the first American brewery opened nearly 600 years later[6]. In contrast, American brewers are not saddled with long traditions, and have experimented and expanded the range of beers and beer-making techniques, from the development of American hops like Citra or Centennial or the Randall, designed to infuse beer. In these experimentation and development, can we identify what makes a good brew?

2.1 Towards a solution

In the project, I will use a Data Science approach to answer the question, what makes a high quality beer? I will operationalize the overall quality of a beer as the BeerAdvocate weighted-rank score[7]. As is attributed to the American entrepreneur Seth Godin, “Don’t find customers for your products, find products for your customers.” In this spirit I am looking for those specific products, beers, that are highly valued by the customers.

This allows me to then determine the qualities that define breweries that are brewing these highly rated beers, and use that information to make recommendations to guide the setting up breweries and the brewing of its beers.

3. Data and Approach

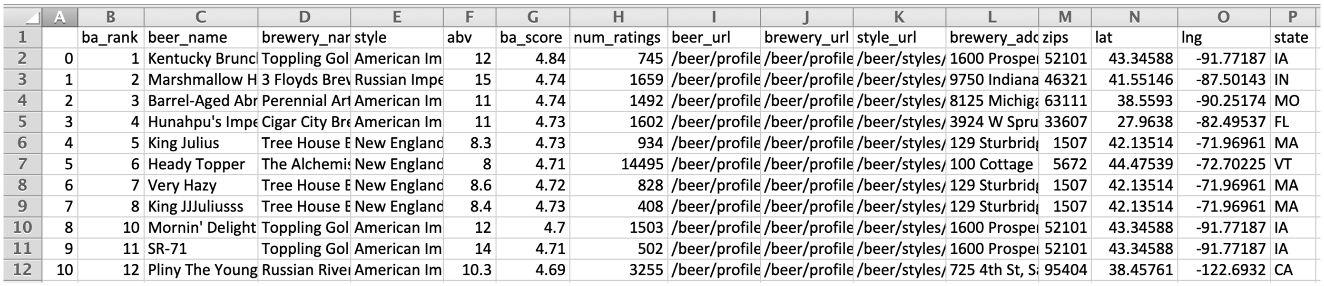
The precise goal of this project is to use drinkers’ ratings of individual beers to determine the characteristics of the breweries making these highly-rated beers. These characteristics would then form the basis for recommending the introduction of a new brewery in the market.

Therefore, I will be taking a beer-centric approach. I will first identify the highly-rated beers, and find the various characteristics of these beers including their styles and contents, as well as characteristics of the breweries, operationalized as the neighborhoods and geographical areas where these are located.

[5] https://www.weihenstephaner.de/en/our-brewery/

[6] http://www.beerhistory.com/library/holdings/chronology.shtml

[7] See https://www.beeradvocate.com/community/threads/top-rated-beers-explained.587593/ for a description



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		ba_rank	beer_name	brewery_nar	style	abv	ba_score	num_ratings	beer_url	brewery_url	style_url	brewery_adc	zips	lat	lng	state
2	0	1	Kentucky Brunc	Toppling Gol	American Im	12	4.84	745	/beer/profile	/beer/profile	/beer/styles/	1600 Prospe	52101	43.34588	-91.77187	IA
3	1	2	Marshmallow H	3 Floyds Brev	Russian Impe	15	4.74	1659	/beer/profile	/beer/profile	/beer/styles/	9750 Indiana	46321	41.55146	-87.50143	IN
4	2	3	Barrel-Aged Abr	Perennial Art	American Im	11	4.74	1492	/beer/profile	/beer/profile	/beer/styles/	8125 Michiga	63111	38.5593	-90.25174	MO
5	3	4	Hunahpu's Impe	Cigar City Bre	American Im	11	4.73	1602	/beer/profile	/beer/profile	/beer/styles/	3924 W Spru	33607	27.9638	-82.49537	FL
6	4	5	King Julius	Tree House E	New England	8.3	4.73	934	/beer/profile	/beer/profile	/beer/styles/	129 Sturbrid	1507	42.13514	-71.96961	MA
7	5	6	Heady Topper	The Alchemis	New England	8	4.71	14495	/beer/profile	/beer/profile	/beer/styles/	100 Cottage	5672	44.47539	-72.70225	VT
8	6	7	Very Hazy	Tree House E	New England	8.6	4.72	828	/beer/profile	/beer/profile	/beer/styles/	129 Sturbrid	1507	42.13514	-71.96961	MA
9	7	8	King JJJuliusss	Tree House E	New England	8.4	4.73	408	/beer/profile	/beer/profile	/beer/styles/	129 Sturbrid	1507	42.13514	-71.96961	MA
10	8	10	Mornin' Delight	Toppling Gol	American Im	12	4.7	1503	/beer/profile	/beer/profile	/beer/styles/	1600 Prospe	52101	43.34588	-91.77187	IA
11	9	11	SR-71	Toppling Gol	American Im	14	4.71	502	/beer/profile	/beer/profile	/beer/styles/	1600 Prospe	52101	43.34588	-91.77187	IA
12	10	12	Pliny The Young	Russian Rive	American Im	10.3	4.69	3255	/beer/profile	/beer/profile	/beer/styles/	725 4th St, S	95404	38.45761	-122.6932	CA

Figure 4. First 10 rows of the CSV file containing the clean dataframe with beers from the Top 250 list on BeerAdvocate.com

For this project I will use the following primary data sources:

3.0.1 BeerAdvocate.com

Beer Advocate is an independent Boston-based website devoted to beer. I will use BEAUTIFULSOUP to extract the top-rated beers and the BA scores for these beers, along with related information, such as the the ABV (% alcohol), and the brewery. I will then use the name of the brewery to also retrieve the brewery location from the website.

3.0.2 Foursquare Developers

Foursquare is a location technology company that gives access to developers to their location data. Location data includes details of "venues" – businesses, restaurants, parks, etc., in relation to a specified geolocation. I will use this information to build neighborhood profiles of breweries, and use these in predictive models to understand if and how such profiles contribute towards predicting high beer scores.

3.0.3 Brewers Association

Brewers Association compiles a list of over 7,000 US breweries. I will again use BEAUTIFULSOUP to extract the names, addresses, and types of all the US breweries they list. I will use machine learning to derive geographical clusters that would be one of the key characteristics included in models (along with the other variables described above) to predict beer scores.

4. Methodology

As described in the previous section, data will be gathered from multiple sources, collated, and used for analyses. All the scripts and analyses were carried out in JupyterLab notebooks instantiated within an Anaconda environment on a Macbook Pro, using Python 3.

4.1 Retrieving the Raw Data

4.1.1 Beer Advocate

The Beer Advocate website utilizes a Bayesian, weighted-rank score gathered from its users who drink and rate beers across the country. They list their top 250 beers in a single webpage[8]. Using the Python REQUESTS and BEAUTIFULSOUP libraries, I retrieved their top 250 beer list. For each of the beers, I extracted and saved the following:

1. Beer name
2. Beer rank
3. Beer score
4. Beer style
5. Beer ABV[9]
6. Brewery name
7. BeerAdvocate URLs of the beer and brewery

This information was saved to a PANDAS dataframe.

Looping over all the brewery URLs, I captured each brewery page and extracted the address of each brewery. This information was added to the previously saved PANDAS dataframe.

In this process, I found a set of breweries that were outside the US, and I dropped these. A few of the breweries had incorrect or missing zip codes; these were manually entered.

Next, I downloaded a CSV file of US zip codes with their latitude/longitude coordinates from OpenDataSoft[10], and imported this as a PANDAS dataframe. I then cross-referenced each zip code against this database and added latitude, longitude coordinates to each beer in the Beer Advocate dataframe.

During analysis (see below) it was observed that zip code information was not appropriate for the Foursquare analysis. Therefore, I used the brewery addresses of the best beers to look up their latitude/longitude coordinates using the Mapquest API.

After cleaning up missing values, I was left with a dataframe containing 219 beers from the Beer Advocate Top 250 list. This I saved to a CSV and a pickle file. The top 10 rows of the CSV file are shown in Figure 4.

4.1.2 Foursquare data

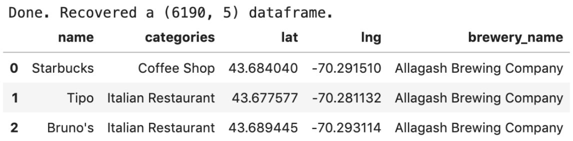
I used the latitude, longitude coordinates for the breweries to recover characteristics of the neighborhoods where all the breweries are located. For each unique brewery in the Best Beers dataframe, I used the Foursquare API to get all venues in a 10Km radius centered on that brewery's coordinates, retrieving up to a maximum of 100 venues per brewery. The large radius was necessary because the scale of brewery locations is the scale of continental US.

The resulting dataframe from the retrieved Foursquare data is shown in Figure 5.

[8] https://www.beeradvocate.com/lists/top/

[9] ABV: Alcohol by Volume, expressed as a %

[10] https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/export/



```
Done. Recovered a (6190, 5) dataframe.
```

	name	categories	lat	lng	brewery_name
0	Starbucks	Coffee Shop	43.684040	-70.291510	Allagash Brewing Company
1	Tipo	Italian Restaurant	43.677577	-70.281132	Allagash Brewing Company
2	Bruno's	Italian Restaurant	43.689445	-70.293114	Allagash Brewing Company

Figure 5. First 3 rows of the dataframe containing venue data retrieved from Foursquare.

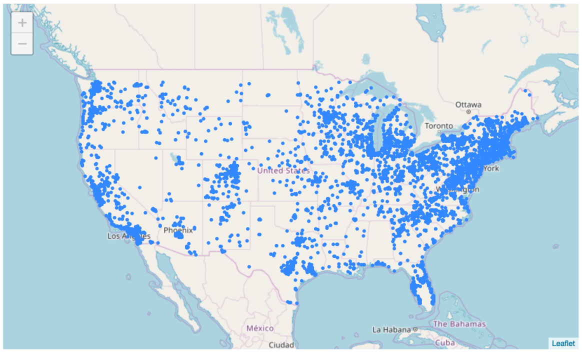


Figure 6. Map showing location of 6,912 breweries in mainland US. Data gathered from https://www.brewersassociation.org/

4.1.3 Brewers Association data

The Brewers Association maintains a fairly exhaustive list of US breweries. I first downloaded the webpage with all US breweries. I then used BEAUTIFULSOUP to extract the brewery names, types, and addresses. I then used REGEX to search for "CC NNNNN" strings, corresponding to 2-character US state and 5-digit US zip code strings. From these I separated out the state and zip code and saved these as separate variables. I then retrieved all state and zip codes, and the latitude/longitude of each brewery.

For some of the entries, the latitude/longitude information was missing. I manually looked up the two most frequent, and used GEOPY to get as many of the rest as I could. The remaining rows with missing latitude/longitude data were dropped. This led to a final sample of 6912 breweries. A FOLIUM maps with these breweries marked is shown in Figure 6.

4.2 Machine Learning

Various analytic tools were used to gain insight into the features that predict high beer scores. The overall strategy was to derive three kinds of predictors from the data gathered as described above:

1. Geographical-scale predictors.
2. Local (neighborhood) predictor.
3. Beer characteristics.

These are described in detail below.

4.2.1 Geographic-scale predictors

We can ask if geography matters. At least with wines, *terroir*, the geographical characteristics, define a given wine. Although we do not have a comparable definition or beer, we can ask more generally, does the location of a brewery influence the quality of the beer it produces?

One way to operationalize the geographic location of a brewery is simply to use its latitude/longitude coordinates. But we would prefer a more categorical variable. A second option is the US state that the brewery is located in. However this is not ideal as there are 50 states, so one-hot encoding will add 50 variables.

Therefore I decided to use the latitude/longitude coordinates of breweries to cluster them into fewer groups to reduce the dimensionality of the geographical-scale predictor. I used KMeans clustering from the SKLEARN package in scikit-learn. Using the entire data set, I examined clusters when the number of clusters, K, ranged from 2-19. I used the Silhouette score to estimate the fit for each value of K. The Silhouette score ranges from -1 to 1, with -1 implying mis-classification, 0 implying significant overlap between clusters, and 1 implying perfect clustering. I found that, while the maximum silhouette was for K=13, there was a second, comparable peak at K=4 (see Figure 7). I therefore chose the more parsimonious K=4. Figure 8 shows the result – each brewery is color coded based on the cluster it is assigned.

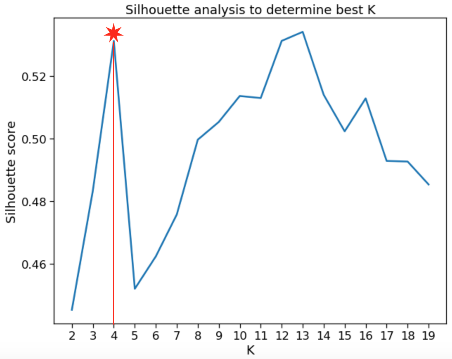


Figure 7. Determining the optimal K in clustering breweries by their latitude/longitude coordinates. The Silhouette score, a measure of cluster definition, is shown as a function of number of clusters, K. An optimal value of K=4 is suggested by a comparatively high silhouette score for comparatively few clusters.

I then assigned clusters to all the beers in the BeerAdvocate-derived Best Beers dataframe, by fitting each row's latitude/longitude coordinates with the KMeans model with K=4.

4.2.2 Segmenting breweries based on Foursquare data

How can we characterize the local "environment" of a brewery to understand how it affects the beer produced there? In this project, I chose to examine the venues returned by Foursquare as a proxy for the local environment of the brewery.

Because Foursquare needs accurate latitude/longitude coordinates, I used Mapquest to retrieve addresses of all breweries producing the best beers. I used this information to retrieve the top 100 venues in a 10km radius around the brewery.

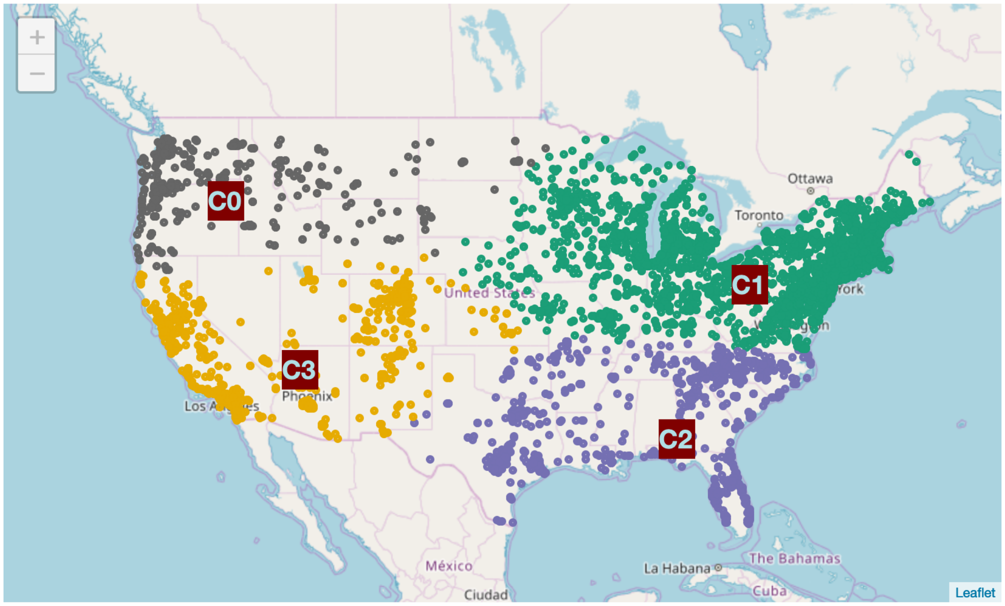


Figure 8. Map showing clustering of the 6,912 breweries in mainland US. Cluster labels (maroon boxes) are positioned at mean latitude, longitude values for that cluster.

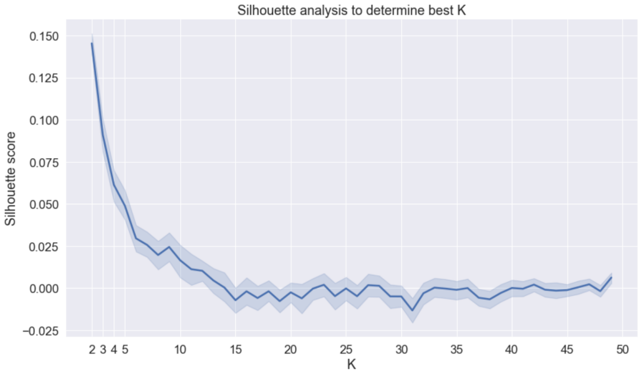


Figure 9. Silhouette scores for different values of K, the number of clusters, in a K-Means algorithm used to determine segmentation of breweries based on nearby venues. The best K was 2, i.e., the minimum number of clusters possible.

To segment this data, I again used the K-Means algorithm. However, I found that there was no structure in the data, as revealed by the fact that the best silhouette values were for 2 clusters, and of these, one contained 69 of the 70 breweries (see Figure 9).

I therefore decided to look only at a specific class of nearby venues – breweries. I therefore counted the number of other breweries next to each of the 70 breweries. The question then becomes, are breweries that are close to other breweries producing better or worse beers?

4.3 Accounting for different beer styles

In order to account for different beer styles, I examined the distribution of number of beers per style (Figure 10).

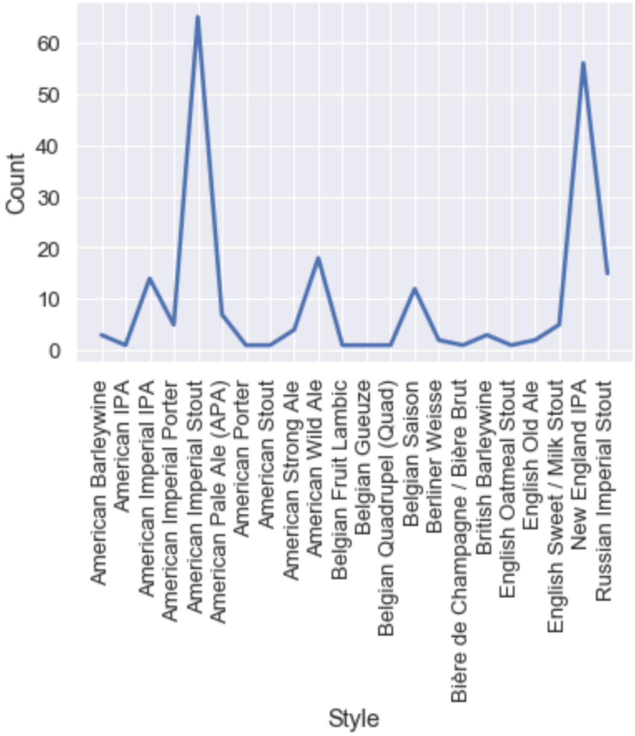


Figure 10. Frequencies of the different styles in the best beers dataset.

Given that many of the styles had low frequencies, I collapsed all styles with less than 10 instances into a separate category called "other."

4.4 A regression model to predict beer scores

We now have all the variables for a linear model. I used the following model:

$$Score \propto ABV + Num\_ratings + Num\_nearby\_breweries + Geo\_cluster + Style$$

where $Score$ is the BeerAdvocate score, ABV is the beer’s ABV, $Num\_ratings$ is the number of ratings received by that beer, $Num\_nearby\_breweries$ is the number of other breweries within a 10Km radius, $Geo\_cluster$ is the geographic cluster (see Figure 8), and $Style$ is the beer style. Note that for “Geo_cluster” and “Style” I used dummy (one-hot) coded variables.

I used several linear regression models, including Ordinary Least Squares (OLS), Ridge regression, and Lasso regression.

5. Results

5.1 Results of the regression model

Of the linear regression models examined (OLS, Ridge, Lasso), OLS showed the best fit as determined by the largest R^2 value on a train-test split of the data. The result of the final model are shown in Figure 11.

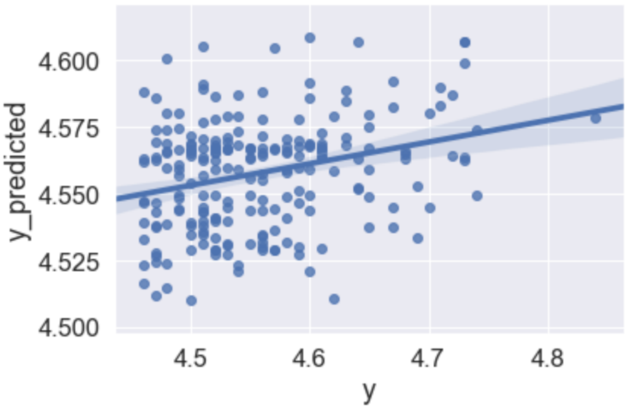


Figure 11. The final regression model output. This plot shows the relation between the actual y values (i.e., BeerAdvocate scores) and the predicted y values (scores).

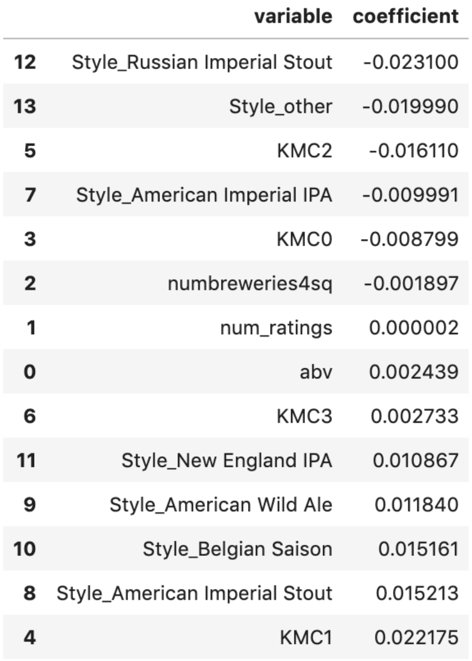
I found that, for this model, the Root Mean Square error was 0.0694. The predicted values were significantly correlated with the actual values; Pearson’s $R = 0.285$, $P = 1.86e-5$.

The coefficients of the model are shown in Figure 12.

6. Discussion

It is satisfying to find that the regression model returns a significant result – implying that there is indeed very likely a relation between the variables extracted and used in this project and beer drinkers’ scores.

Diving into the coefficients of the model, we can look at the comparatively higher absolute values of the coefficients, e.g., $|values| > 0.01$. Doing so, we observe:



	variable	coefficient
12	Style_Russian Imperial Stout	-0.023100
13	Style_other	-0.019990
5	KMC2	-0.016110
7	Style_American Imperial IPA	-0.009991
3	KMC0	-0.008799
2	numbreweries4sq	-0.001897
1	num_ratings	0.000002
0	abv	0.002439
6	KMC3	0.002733
11	Style_New England IPA	0.010867
9	Style_American Wild Ale	0.011840
10	Style_Belgian Saison	0.015161
8	Style_American Imperial Stout	0.015213
4	KMC1	0.022175

Figure 12. Coefficients for the final regression model.

1. The largest positive coefficient is being a member of Cluster 1 (see Figure 8 above). That is, a beer brewed in the North-East of the US, including New England, New York, Pennsylvania, and the Great Lakes.
2. Specific style, American Imperial Stouts, Belgian Saisons, American Wild Ales, and New England IPAs all positively impact ratings (in that order). These are also the very frequent styles we see in the best beers database.
3. The geographical location, the American Southeast, and Russian Imperial Stouts and other infrequent styles result in lower ratings

Interestingly, of the four geographical regional clusters derived from brewery coordinates (Figure 8), the North-Eastern and the South-Western (including almost all of California), contribute positively to higher beer ratings. This finding sits very well with my intuitions as a beer aficionado – these are indeed the regions that the best beers that I have tasted have tended to come from.

7. Conclusions

So, where should a potential investor choose to open a new brewery, and what should they brew?

The answer seems to be that the best would be to brew an American Imperial Stout somewhere in the North Eastern regions of USA.

8. Acknowledgments

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.