Beer Preferences Across America

Data Mining Project: How Data Can Reveal Our Favorite Beer Styles?

Jessica Carpenter
jeha6343@colorado.edu

Thomas Chavez
thch6537@colorado.edu

Guanbo Bian
gubi2340@colorado.edu

1. Abstract

Beer preferences can vary depending on a number of factors, and in this project, we hope to determine the key influences that determine a style of beer's popularity. By analyzing the data set from the popular beer ranking site "Beer Advocate", we made associations between beer styles and characteristics and how they can be mapped to trends.

Upon thorough data analysis, we can discern the crème de la crème among beer styles. American Wild Ale and Gueuze stand tall, boasting an average rating exceeding 4.0. Renowned for their intricate and tart flavors, these styles often mature in barrels for an extended duration. If you're seeking an adventurous escapade into uncharted brews, these options might just be tailor-made for your palate.

Conversely, at the opposite end of the spectrum, Light Lagers and Low Alcohol Beers sport the lowest average ratings, hovering below 3.0. Characterized by their delicate body and low alcohol content, these styles commonly serve as refreshing quenchers rather than taste-intensive experiences. However, individual preferences vary greatly! Should you yearn for a light and uncomplicated libation, these styles could be the perfect companions.

One of the most fascinating aspects of beer lies in its astonishing array of styles and flavors. Truly, there exists a beer to cater to every taste! With diverse average ratings across various beer styles, the profound diversity and complexity of the beer universe come into focus. So don't hesitate to tread the path of novelty – you might just stumble upon something unexpectedly delightful.

It's worth noting that certain popular beer styles like American Pale Ale and IPA exhibit relatively modest average ratings compared to their counterparts. This phenomenon could stem from their widespread production and consumption, increasing the likelihood of encountering subpar or middling examples of these styles. Thus, passing judgment solely based on a style's popularity might not do justice – the true gems might be the ones hidden in plain sight.

In summation, the data unequivocally reveals an expansive spectrum of beer styles with varying degrees of popularity and excellence. Whether your inclinations lean toward the bold and intricate or the crisp and invigorating, a beer awaits you. So, without hesitation, embark on a journey of experimentation – you just might uncover your next beloved brew.

2. Introduction

Beer, a globally cherished libation, garners widespread affection, with extensive volumes of data delving into beer proclivities. The impetus propelling this undertaking is the aspiration to comprehensively apprehend the foremost factors underpinning consumers' beer selection. Furthermore, our pursuit

aims to fathom whether geographical locale yields discernible disparities. The fruition of such insights might empower the spotting of nascent sales territories, the conceptualization of novel, crowd-pleasing products, and the refinement of the brewing craft through the curation of delectable and enticing recipes – whether pursued in the domestic realm as an adept "homebrewer" or within the commercial sphere as an entrepreneurial enterprise.

Within the realms of our endeavor, we aspire to provide elucidation to inquiries akin to the ensuing:

1. Which breweries command unparalleled popularity?
2. How does predilection ebb and flow across different geographic pockets?
3. Do trends manifest in facets such as fragrance, appearance, or flavor?
4. Can we proffer brewery recommendations predicated on identified trends?

In culmination, this venture stands poised to uncork a wealth of insights into the intricate realm of beer preferences. As the effervescent concoction unites palates around the globe, our exploration promises to shed light on the factors that tantalize taste buds and the interplay of regional nuances. These findings bear the potential to embolden both the industry and the avid homebrewer alike, igniting innovation, refining recipes, and fostering a greater appreciation for the nectarous libation. Through this endeavor, we raise a collective toast to the multifaceted world of beer and the revelations it holds.

3. Related Work

Amidst the burgeoning appeal of microbrews and homebrewers, a plethora of investigations have delved into the realm of beer preferences. A 2013 study conducted by P.A.S. Romano et al. [1] scrutinized the correlations between alcoholic beverage preferences and dietary inclinations. The findings unveiled a proclivity among beer enthusiasts for specific comestibles, painting a nexus between palate preferences and ideal flavor profiles. Another exploration, authored by M. Calvo-Poral et al. [2], endeavored to gauge consumer predilections for craft beer attributes. The research highlighted taste, fermentation process, and color as pivotal attributes for consumers. Moreover, it surfaced that traditional ingredients and methods were prime considerations when selecting beers. J.W. Finley [3] conducted a 2017 study dissecting beer consumption trends to engender novel beer creations. The study's focus encompassed mapping popular beer styles across diverse nations and tracking beer consumption patterns, thereby catalyzing the emergence of novel product offerings.

While these existent works do offer captivating insights into comparable data, our endeavor aspires to transcend prior research by harnessing Beer Advocate rankings. These rankings, culled from user-generated reviews, brewery appraisals, and social media inputs, bestow an added layer of richness to our analysis. By incorporating these supplementary dimensions, we intend to leverage machine learning algorithms to underpin our analysis, culminating in tailored recommendations for both homebrewing aficionados and commercial brewers alike.

4. Data Set

The data set used for this project is from a set of reviews on the popular beer ranking website and online community, Beer Advocate. Beer Advocate was founded in 1996 and provides a variety of ways for users to rate beers, write reviews, and post to forums. The website also posts style guides and news about beer festivals, education, and trends. The data provided by Beer Advocate is neatly provided in a csv file on Kaggle for use in data science projects. The data set is populated with: brewery name, brewery ID, review

time, overall review, aroma, appearance, name, palate, and taste.

The data set can be accessed at the following address: https://www.kaggle.com/datasets/rdoume/beerreviews

Below is the explanation of some key attributes in the dataset.

- beer_ABV : Alcohol by volume content of a beer
- beer_beerId : Unique ID for beer identification
- beer_brewerId : Unique ID identifying the brewer
- beer_name : Name of the beer
- beer_style : Beer Category
- review_appearance: Rating based on how the beer looks [Range : 1-5]
- review_palatte : Rating based on how the beer interacts with the palate [Range : 1-5]
- review_overall : Overall experience of the beer is combined in this rating [Range : 1-5]
- review_taste : Rating based on how the beer actually tastes [Range : 1-5]
- review_profileName: Reviewer's profile name / user ID
- review_aroma : Rating based on how the beer smells [Range : 1-5]
- review_text : Review comments/observations in text format
- review_time : Time in UNIX format when review was recorded

The table below is the first 10 lines of wine dataset.

Table 1. The first 10 rows of wine dataset

	brewery_id	brewery_name	review_time	review_overall	review_aroma	review_appearance	review_profilename	beer_style	review_palate	review_taste	beer_name	beer_abv	beer_beerid
0	10325	Vecchio Birraio	1234817823	1.5	2	2.5	stcules	Hefeweizen	1.5	1.5	Sausa Weizen	5	47986
1	10325	Vecchio Birraio	1235915097	3	2.5	3	stcules	English Strong Ale	3	3	Red Moon	6.2	48213
2	10325	Vecchio Birraio	1235916604	3	2.5	3	stcules	Foreign / Export Stout	3	3	Black Horse Black Beer	6.5	48215
3	10325	Vecchio Birraio	1234725145	3	3	3.5	stcules	German Pilsener	2.5	3	Sausa Pils	5	47969
4	1075	Caldera Brewing Company	1293735206	4	4.5	4	johnmichaelsen	American Double / Imperial IPA	4	4.5	Cauldron DIPA	7.7	64883
5	1075	Caldera Brewing Company	1325504859	3	3.5	3.5	oline73	Herbed / Spiced Beer	3	3.5	Caldera Ginger Beer	4.7	52159
6	1075	Caldera Brewing Company	1318991115	3.5	3.5	3.5	Reidrover	Herbed / Spiced Beer	4	4	Caldera Ginger Beer	4.7	52159
7	1075	Caldera Brewing Company	1306278018	3	2.5	3.5	alpinebryant	Herbed / Spiced Beer	2	3.5	Caldera Ginger Beer	4.7	52159
8	1075	Caldera Brewing Company	1290454503	4	3	3.5	LordAdmMeleon	Herbed / Spiced Beer	3.5	4	Caldera Ginger Beer	4.7	52159
9	1075	Caldera Brewing Company	1286032824	4.5	3.5	5	augustgarage	Herbed / Spiced Beer	4	4	Caldera Ginger Beer	4.7	52159

5. Main Techniques Applied

- Data clean and preprocessing

First we'd like to find the missing values. Table 2 below is the attribute information. We have 13 columns and the column type includes numeric and strings.

Table 2. The first 10 rows of wine dataset

```
 #   Column              Non-Null Count    Dtype
---  ------              --------------    -----
 0   brewery_id          1586614 non-null  int64
 1   brewery_name        1586599 non-null  object
 2   review_time         1586614 non-null  int64
 3   review_overall      1586614 non-null  float64
 4   review_aroma        1586614 non-null  float64
 5   review_appearance   1586614 non-null  float64
 6   review_profilename  1586266 non-null  object
 7   beer_style          1586614 non-null  object
 8   review_palate       1586614 non-null  float64
 9   review_taste        1586614 non-null  float64
 10  beer_name           1586614 non-null  object
 11  beer_abv            1518829 non-null  float64
 12  beer_beerid         1586614 non-null  int64
dtypes: float64(6), int64(3), object(4)
```

Table 3 shows the number of NA in each column. We can see that the two columns which have most missing values are review_profilename and beer_abv columns.

Table 3. Number of NA in each column

```
brewery_id              0
brewery_name           15
review_time             0
review_overall          0
review_aroma            0
review_appearance       0
review_profilename    348
beer_style              0
review_palate           0
review_taste            0
beer_name               0
beer_abv            67785
beer_beerid             0
dtype: int64
```

We removed the following four columns since they are not quite related to our work: "brewery_id", "review_time", "review_profilename", "beer_beerid" so we don't need to worry NA in profilename. For NA in beer_abv, we can either drop the NA observations or use mean to fill in the NA. Since there are 67785

observations and we don't want to make the result biased, we choose to use the mean of this attribute to fill in the NA.

After these steps and we rerun to check NA, the result is shown in table4.

Table 4. Number of NA in each column

brewery_name	0
review_overall	0
review_aroma	0
review_appearance	0
beer_style	0
review_palate	0
review_taste	0
beer_name	0
beer_abv	0

Since this dataset is also highly based on review scores, we deleted the data observations with abnormal values. For example, we delete all outliers with review score less than 0 or greater than 5.

- Tools we used for data cleaning, processing and analytics

Our project is executed using a variety of tools and libraries. Python will be the main programming language for analysis because it contains multiple libraries and methods for data science projects, and it is easy to use and understand. The Python libraries we will be utilizing are:

Pandas: Pandas can be used for a variety of tasks: to clean data by transforming and cleaning data for preprocessing.

NumPy: NumPy is scientific computation library that can be used to design algorithms for statistics and machine learning.

Matplotlib: Matplotlib is a visualization library for creating charts and plots to display data in an easily readable form.

Seaborn: Seaborn is a visualization library that builds on Matplotlib to create visualizations.

Scikit-learn: Scikit-learn is a machine learning library that provides a range of machine learning algorithms for supervised learning.

- Data classification method

We first use aggregation method in python to get review scores for different beer type and beer brand and give recommendations.

Linear regression method was also used to check the factors which most impact overall rating.

K-means method was used to clustering different breweries. In the example, we selected overall review and review_aroma as the two features and K-means method in sklearn.

- Model evaluation method

Evaluation 1: K-fold Cross-Validation: This is an extension of cross-validation where the dataset is divided into K equal-sized folds. The model is trained K times, each time using K-1 folds as the training set and one fold as the validation set. The results are averaged to obtain an overall performance measure.

Evaluation 2: Leave-One-Out Cross-Validation: This is a special case of K-fold cross-validation where K is equal to the number of instances in the dataset. Each instance is used as the validation set once, and the model is trained on the remaining instances. This method is useful when working with small datasets.

Evaluation 3: Performance Metrics: Various performance metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC) can be used to evaluate the model's performance based on its predictions compared to the ground truth.

6. Key result

In this project we are trying to make beer recommendations based on the taste, overall ranking etc. In the results below we illustrate some recommendations based on different criteria.

First, we would like to recommend breweries based on different beers. Here we list the top 20 beers.

Table 5. overall rating by beer name

	beer_name	review_overall
0	Taeberry Boch	5.0
1	Wasatch Irish Stout	5.0
2	Pale Ale S.C.A.G. (Simcoe, Columbus, Amarillo ...	5.0
3	Distorter Porter	5.0
4	Louwaege's Stout	5.0
5	Belgian Country Pale Ale	5.0
6	Cappy's Cherry Wheat	5.0
7	C. Brown's Pumpkin Ale	5.0
8	John's Mild Bitter	5.0
9	Raspberry Sparkle	5.0
10	Brett Reverend	5.0
11	Belgian Century	5.0
12	Divided Sky Rye IPA	5.0
13	Skull And Bones Foxy	5.0
14	La Sambresse Blonde	5.0
15	Raspberry Bourbon County Brand Coffee Stout	5.0
16	Naked Oat Stout	5.0
17	Mango Maddness	5.0
18	Frostbite Ice	5.0
19	O'Junior's Nitro Irish Ale	5.0

Next, we explored the overall rating by beer type.

Table 6. overall rating by beer type

	beer_style	review_overall
0	American Wild Ale	4.093262
1	Gueuze	4.086287
2	Quadrupel (Quad)	4.071630
3	Lambic - Unblended	4.048923
4	American Double / Imperial Stout	4.029820
5	Russian Imperial Stout	4.023084
6	Weizenbock	4.007969
7	American Double / Imperial IPA	3.998017
8	Flanders Red Ale	3.992722
9	Rye Beer	3.981737
10	Keller Bier / Zwickel Bier	3.981088
11	Eisbock	3.977094
12	American IPA	3.965221
13	Gose	3.965015
14	Saison / Farmhouse Ale	3.962564
15	Belgian IPA	3.958704
16	Baltic Porter	3.955410
17	Roggenbier	3.948498
18	Oatmeal Stout	3.941692
19	American Black Ale	3.934475

Table 7 shows a beer recommendation based on beer aroma, appearance and taste.

Table 7. Beer aroma, appearance and taste

	beer_name	review_aroma	review_appearance	review_taste
0	Sausa Weizen	2.0	2.5	1.5
1	Red Moon	2.5	3.0	3.0
2	Black Horse Black Beer	2.5	3.0	3.0
3	Sausa Pils	3.0	3.5	3.0
4	Cauldron DIPA	4.5	4.0	4.5
5	Caldera Ginger Beer	3.5	3.5	3.5
6	Caldera Ginger Beer	3.5	3.5	4.0
7	Caldera Ginger Beer	2.5	3.5	3.5
8	Caldera Ginger Beer	3.0	3.5	4.0
9	Caldera Ginger Beer	3.5	5.0	4.0
10	Amstel Light	2.0	3.0	2.5
11	Caldera Ginger Beer	5.0	4.0	4.0
12	Caldera Ginger Beer	4.0	4.0	4.0
13	Caldera Ginger Beer	4.5	3.0	3.0
14	Caldera Ginger Beer	4.0	3.0	4.0
15	Caldera Oatmeal Stout	3.0	2.5	3.0
16	Caldera Oatmeal Stout	1.5	2.5	2.0
17	Caldera OBF 15	3.0	4.0	4.0
18	Amstel Light	3.0	3.0	2.0
19	Rauch Ür Bock	4.5	3.0	4.5

Table 8 we showed the beer ranking based on both overall rating and review rate. The beer data was grouped by both brewery name and beer name.

Table 8. beer ranking based on both overall rating and review rate

		review_overall	review_taste
		mean	mean
brewery_name	beer_name		
't Hofbrouwerijke	Blondelle	4.000000	4.000000
	Bosprotter	3.722222	3.777778
	Hof Korvatunturi	3.750000	3.750000
	Hofblues	3.812500	3.593750
	Hofdraak	3.357143	3.428571
	Hofnar	3.000000	3.500000
	Hoftrol	4.500000	3.500000
(512) Brewing Company	(512) Alt	4.090909	3.909091
	(512) Black IPA	4.000000	4.038462
	(512) Brandy Barrel Aged ONE	3.750000	4.062500
	(512) Bruin	3.968750	4.281250
	(512) Cascabel Cream Stout	3.642857	3.571429
	(512) IPA	4.259259	4.185185
	(512) ONE	3.730769	3.807692
	(512) Pale	3.678571	3.678571
	(512) Pecan Porter	4.185185	4.166667
	(512) TWO	3.823529	3.852941
	(512) Three	4.071429	4.142857
	(512) Whiskey Barrel Aged Double Pecan Porter	4.067568	4.135135
	(512) Wit	4.305556	3.972222

Table 9 illustrates the brewery and beers which contain beer ABV greater than 30%.

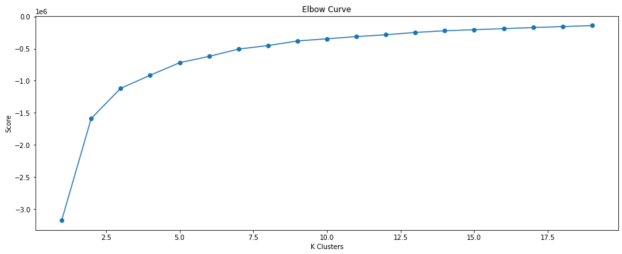
Table 9. Brewery and beer which have beer abv greater than 30%

	brewery_name	beer_name	beer_abv
12919	Schorschbräu	Schorschbräu Schorschbock 57%	57.70
12939	Schorschbräu	Schorschbräu Schorschbock 43%	43.00
12940	Schorschbräu	Schorschbräu Schorschbock 43%	43.00
746387	BrewDog	Sink The Bismarck!	41.00
746396	BrewDog	Sink The Bismarck!	41.00
...	...	...	...
748360	BrewDog	Tactical Nuclear Penguin	32.00
748359	BrewDog	Tactical Nuclear Penguin	32.00
748358	BrewDog	Tactical Nuclear Penguin	32.00
748357	BrewDog	Tactical Nuclear Penguin	32.00
12918	Schorschbräu	Schorschbräu Schorschbock 31%	30.86

We also checked the linear relationship between beer overall rating and other ratings. It turns out that the overall rating is highly related to beer taste and palate.

K-means method was used to classify the beer by overall review score. Figure 0 was the k-means score result. From the plot we can see that after 10 clusters, the error is negligible.

Fig. 0 The appearance and taste ranking by breweries



7. Application

In this project, different beer and brewery are analyzed and recommendations are provided based on different criteria. This research can be applied to beer sales strategy for liquor store and beer factories.

There are many applications of this project. Breweries can leverage beer reviews to gain insights into customer preferences, identify popular beer styles, and make informed decisions about recipe adjustments or new product development.

Also analyzing reviews can help breweries assess the quality of their products. Identifying consistent negative feedback could indicate potential issues with taste, aroma, or other attributes that require attention.

Beer reviews provide valuable data for market analysis. Breweries can identify trends, understand consumer preferences, and tailor their offerings to meet market demands. Beer enthusiasts often rely on reviews to make purchasing decisions. By accessing aggregated reviews, consumers can learn about the taste, aroma, appearance, and overall experience of a particular beer.

E-commerce platforms can use reviews to recommend beers to customers based on their past preferences and similarities to other highly rated products. For example: Gueuze, a Belgian style of beer crafted by blending young and old lambics, boasts the highest mean rating among beer styles, showcasing its popularity. The pinnacle-rated beer, "Armand'4 Oude Geuze Lente (Spring)" by Brouwerij Drie Fonteinen, exemplifies this trend as a Gueuze beer. The elite roster of the top 10 best-rated beers predominantly features American Double / Imperial IPAs and American Wild Ales, known for their daring and intricate flavor profiles. A standout entry, "Trappist Westvleteren 12" from Brouwerij Westvleteren, a Quadrupel beer, not only ranks among the highest-rated brews but also boasts an impressively substantial number of ratings (1272). Goose Island Beer Co., a prominent craft brewery, earns acclaim with "King Henry," an English Barleywine, listed among the finest-rated beers.

Researchers can use beer review datasets to study trends, sentiment analysis, and correlations between beer attributes and consumer preferences. This can lead to valuable insights into the beer industry and consumer behavior.

Natural language processing techniques can be applied to analyze sentiment in beer reviews, providing breweries with an understanding of how customers perceive their products.

Machine learning models can be trained to predict beer ratings or preferences based on attributes such as ingredients, style, and brewing methods. This can aid in optimizing recipes and marketing strategies.

Collaborative filtering and content-based recommendation systems can use beer reviews to suggest similar beers to users, enhancing their exploration of new flavors.

Beer reviews contribute to the larger culture around beer appreciation. They facilitate discussions, sharing of experiences, and the formation of online communities among beer enthusiasts.

Breweries can compare their products' reviews with those of competitors, identifying strengths and weaknesses and informing strategies to gain a competitive edge. Beer reviews can serve as educational resources for those interested in learning about different beer styles, brewing techniques, and flavor profiles.

Overall, beer reviews serve as a rich source of data that can influence decisions within the beer industry, aid in consumer choices, support data analysis endeavors, and contribute to the broader beer culture and community.

8. Visualization

The following figures illustrate the information we can obtain from the dataset.

Of importance is the fact that while there is clearly correlation between review metrics and beer types, the breweries rank differently with respect to taste and

appearance, for example. It cannot be assumed that a given brewery will rank similarly across all metrics. Further analysis will need to be done to determine this effect.

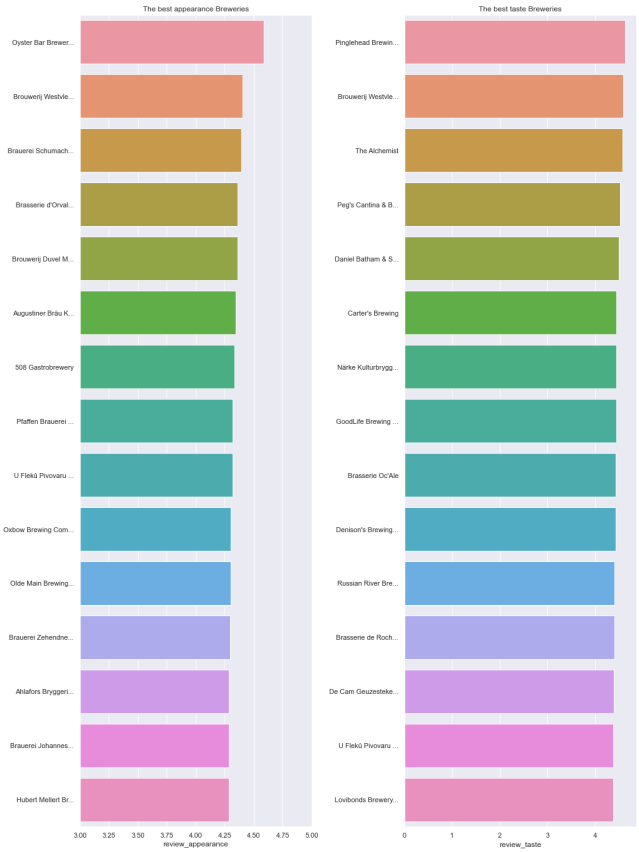


Fig. 1 The appearance and taste ranking by breweries.

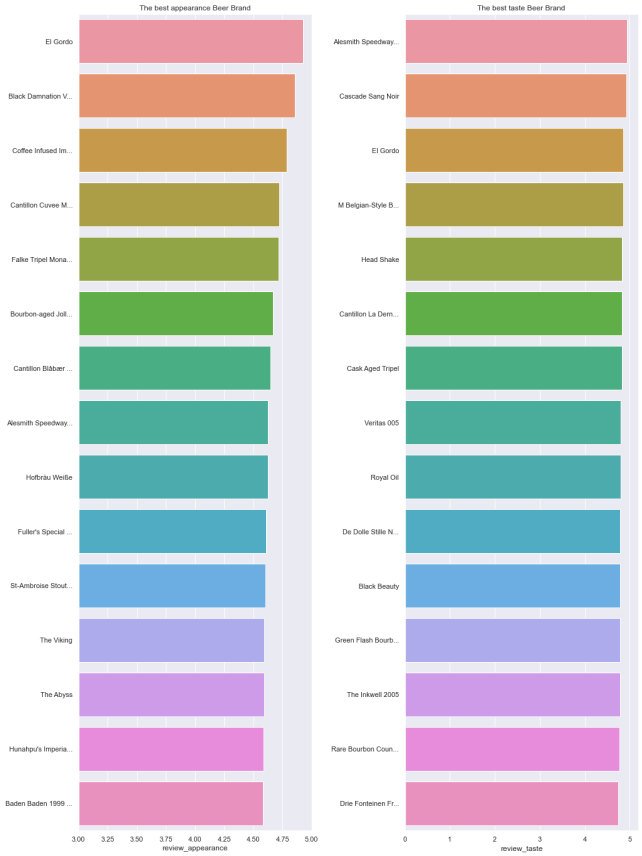


Fig. 2 The appearance and taste ranking by beer brand.

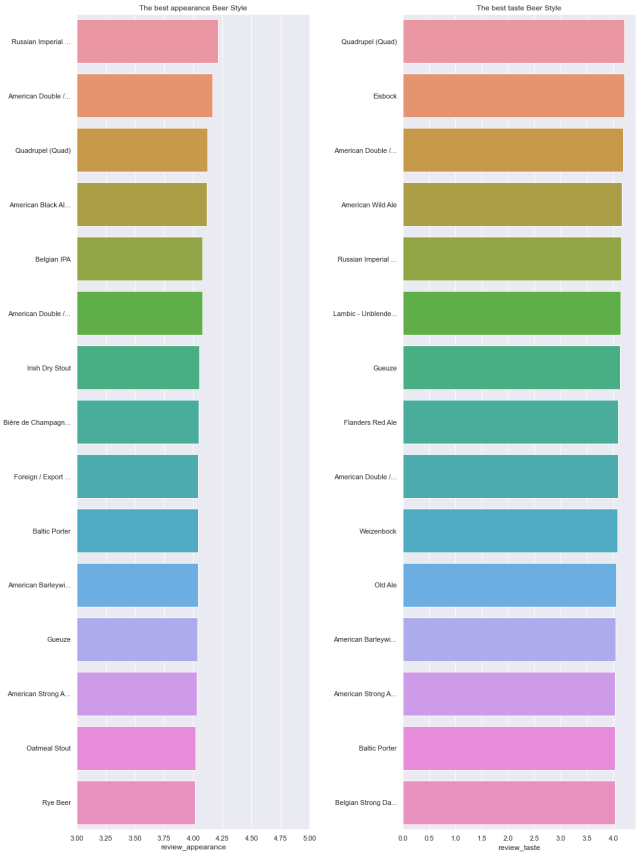


Fig. 3 The appearance and taste ranking by beer style.

Next we explore the histogram of review scores to check what the scores are clustered.

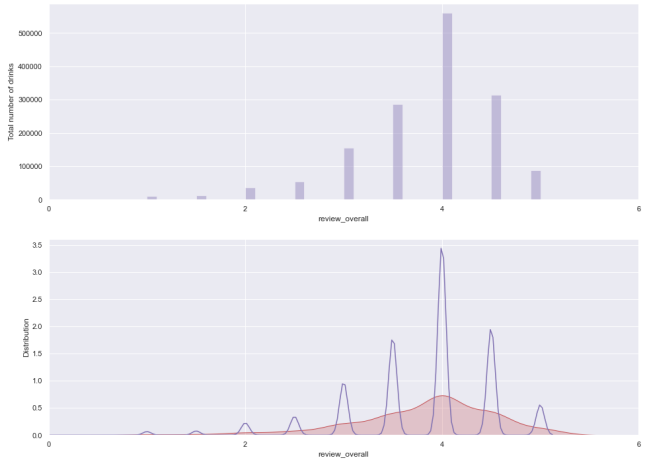


Fig. 4 Histogram of overall review

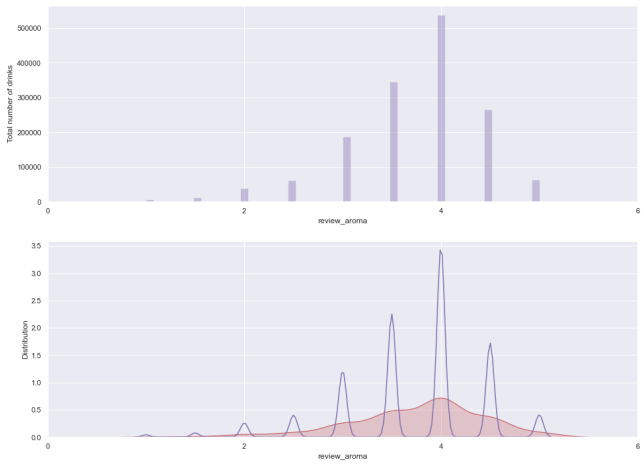


Fig. 5 Histogram of review aroma

We are also interested in the alcohol in brewery, which is indicated as review_abv in the dataset. The plot below is showing the top 20 breweries in terms of alcohol by volume

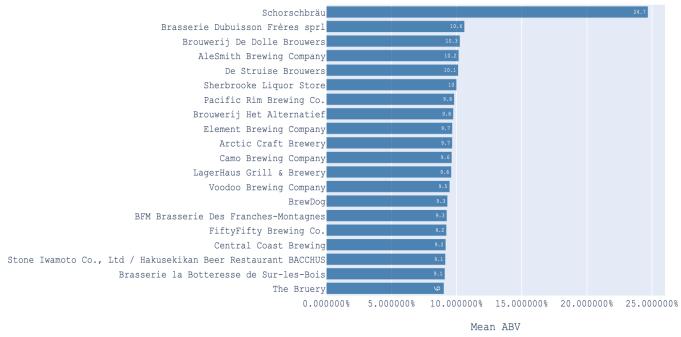


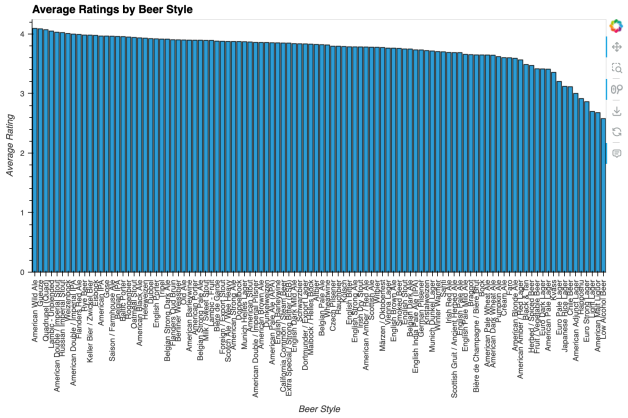
Fig. 6 Top 20 Breweries by Highest Mean ABV

There are five review columns total and we wonder if they are correlated. Correlation matrix is provided and a heat map is plotted too.

	review_appearance	review_aroma	review_palate	review_taste	review_overall	review_average
review_appearance	1.000000	0.559102	0.564549	0.544585	0.498570	0.739205
review_aroma	0.559102	1.000000	0.614935	0.714786	0.612821	0.834470
review_palate	0.564549	0.614935	1.000000	0.732216	0.699036	0.859231
review_taste	0.544585	0.714786	0.732216	1.000000	0.787190	0.905356
review_overall	0.498570	0.612821	0.699036	0.787190	1.000000	0.862499
review_average	0.739205	0.834470	0.859231	0.905356	0.862499	1.000000

Fig. 7 correlation matrix of review dimensions.

Here is the plot of average rating vs beer style.



From the correlation matrix we can see that appearance has no strong correlation with other reviews. Taste has a relatively strong correlation with palate and overall reviews. The overall rating has correlation with almost all the review categories, although with appearance the correlation is small.

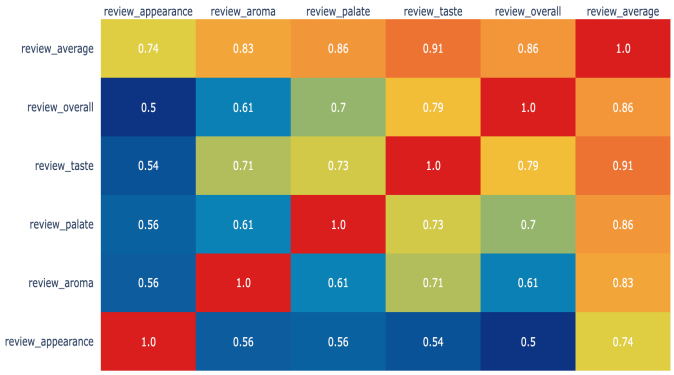


Fig. 7 Heat map between different review categories.

REFERENCES

[1] Romano, P.A.S., et al. "Alcoholic Beverages and Diatary Habits: A Systematic Literature Review." *Appetite* 61, no. 1 (2013): 1-12. dio:10.1016/j.appet.2012.10.001.

[2] Calvo-Porral, M., et al. "Measuring consumers' preferences for craft beer attributes through Best-Worst Scaling." *Agricultural and Food Economics* 8, no.1 (2019): 1-13. doi:10,1186/s40100-019-0138-4.

[3] Finley, J.W. "The Science of Beer: How Data Can Reveal Our Favrotie Styles." *Scientific American* 317, no. 4 (2017): 36-43. Doi:10.1038/scientificamerican0417-36.

[4] https://www.kaggle.com/code/fabiancpl/recommending-beers