Beer Challenge 2022

Jatin Krishna Sai Kodali

Dataset

We have a beer review dataset of over 528k reviews from 22.8k users and from initial analysis the data appears to be very clean with not a lot of missing data.

Each row in the dataset was a review and it had data about the timestamp of the review, profileName, review text, beerId, brewerId, ABV of the beer, beer name and style, and numerical rating of aroma, appearance, palette, taste and overall.

Questions to be Answered

1. Rank top 3 Breweries which produce the strongest beers?
2. Which year did beers enjoy the highest ratings?
3. Based on the user’s ratings which factors are important among taste, aroma, appearance, and palette?
4. If you were to recommend 3 beers to your friends based on this data which ones will you recommend?
5. Which Beer style seems to be the favorite based on reviews written by users?
6. How does written review compare to overall review score for the beer styles?
7. How do find similar beer drinkers by using written reviews only?

Q1: Rank top 3 Breweries which produce the strongest beers

The strength of the beer is given by ABV(Alcohol by Volume), so we take the data of just the brewerId, beerId, and ABV and filter out the duplicates and Nan values.

Now for every brewer we calculate the average ABV of their 15 strongest beers (15 largest ABV values), sort them on this average value and the top 3 brewerIds are the breweries with the strongest beers.

With the given data the top 3 brewerIds come out to be **6513, 16866, 35** in that order.

Q2: Which year did beers enjoy the highest ratings

Since the distribution of number of reviews over the years is not uniform, simple mean of ratings over years will not give a quality comparison. So, we need to rank the years using a statistical metric that considers both the number of ratings and the numerical value of the ratings. For this we use bayesian_rank which uses a beta posterior distribution to get the probabilities. This bayseian_rank is our go-to ranking metric for the project.

Since review_time column in the data is in Unix timestamp, we convert it into datetime and we extract the year. Now using bayesian ranking on the overall ratings, we see that **2010** is the year with beer getting highest quality ratings, with **2009** coming at a close second.

Q2: Which year did beers enjoy the highest ratings

Equations for bayesian_rank taken from the book look like

$$\frac{a}{a+b} - 1.65\sqrt{\frac{ab}{(a+b)^2(a+b+1)}}$$

where

$$a = 1 + S \tag{16}$$

$$\tag{17}$$

$$b = 1 + N - S \tag{18}$$

$$\tag{19}$$

where N is the number of users who rated, and S is the sum of all the ratings, under the equivalence scheme mentioned above.

Q2: Which year did beers enjoy the highest ratings

Top 10 years ranked on their overall ratings look like

	review_year	N	S	bayesian_rank
12	2010	93810	72536.5	0.770966
11	2009	83578	64595.6	0.770480
10	2008	69080	52969.7	0.764125
13	2011	110836	84858.1	0.763514
7	2005	29433	22557.7	0.762321
9	2007	46514	35439.5	0.758641
3	2001	602	472.9	0.757025
8	2006	43083	32727.5	0.756230
6	2004	22905	17383.2	0.754240
14	2012	3180	2435.7	0.753390

Q3: Based on the user's ratings which factors are important among taste, aroma, appearance, and palette?

We can frame this as a feature importance study by looking at how important taste rating, aroma rating, appearance rating, and palette rating are to predict overall rating (target) of the beer. Since these features are independent of each other, a univariate analysis should give us their importance.

We use Predictive Power Score (PPS) which uses a non-linear model fit (unlike correlation which is linear) to see how the target is dependent on the feature. Using this we find that **Aroma, Taste, Palette and Appearance is the order** of importance.

Q4: If you were to recommend 3 beers to your friends based on this data which ones will you recommend?

If I were to recommend beers to a friend in general, I would use the overall ratings of the beers as the indicating factor. So, for this question we rank beers using bayesian_rank which considers both the number of ratings and the numerical value of ratings for each beerId. Using bayesian_rank we get **Heady Topper, Founders CBS Imperial Stout and Citra DIPA** as our top 3 recommendations.

This problem can be handled better with Machine Learning if we know the taste (preferences) of the friends before hand so we can use a collaborative filtering method to give more personalized recommendations.

Q4: If you were to recommend 3 beers to your friends based on this data which ones will you recommend?

Top 10 beer rankings look like

	beer_beerId	N	S	bayesian_rank	beer_name
4279	16814	469	433.9	0.903150	Heady Topper
11986	47658	637	584.9	0.898898	Founders CBS Imperial Stout
14236	56082	252	233.4	0.895261	Citra DIPA
8983	36316	156	144.4	0.884805	Cantillon Blåbær Lambik
1667	6368	662	594.3	0.877049	Masala Mama India Pale Ale
2282	8626	41	39.1	0.870176	Southampton Berliner Weisse
5060	19960	1932	1699.2	0.866883	Founders KBS (Kentucky Breakfast Stout)
4035	15881	1955	1718.7	0.866571	Tröegs Nugget Nectar
3003	11757	2502	2179.0	0.859542	Founders Breakfast Stout
157	645	2170	1883.3	0.855543	Trappistes Rochefort 10

Q5: Which Beer style seems to be the favorite based on reviews written by users?

This is a sentiment analysis problem as we are trying to assign a numerical value (sentiment) to the text review and see its relation to beer style. So, we use the VADER sentiment analyzer which takes both polarity and intensity of the text into account. This analyzer gives 4 scores (neg,neu,pos,compound) which correspond to negative sentiment, neutral sentiment, positive sentiment, and the normalized value of the three.

We run the sentiment analyzer on all the 528k reviews (takes around 8 mins to run, so ran it once and stored it as csv to keep using it for future runs) and then we rank beer style based on the bayesian_rank of the 'compound score' given by the analyzer for the reviews. We see that **Quadrupel(Quad) and American Double/Imperial Stout** are the favorite beer styles based on the reviews.

Q5: Which Beer style seems to be the favorite based on reviews written by users?

Top 20 beer style ranking based on review text looks like,

	beer_style	N	S	bayesian_rank
86	Quadrupel (Quad)	4933	4582.27060	0.922685
11	American Double / Imperial Stout	23352	21582.97050	0.921351
58	Flanders Red Ale	2856	2642.82700	0.916937
38	Dortmunder / Export Lager	1809	1674.03310	0.914707
25	Belgian Strong Dark Ale	15403	14122.05610	0.913112
90	Rye Beer	5179	4739.14050	0.908513
84	Old Ale	4817	4408.55015	0.908408
101	Wheatwine	891	821.96225	0.906734
87	Rauchbier	2607	2385.85675	0.905841
74	Lambic - Fruit	3768	3437.93710	0.904580
20	American Wild Ale	3695	3371.01640	0.904412
4	American Barleywine	10107	9181.79685	0.903644
100	Weizenbock	2235	2041.11175	0.903044
89	Russian Imperial Stout	17183	15553.12120	0.901410
37	Doppelbock	5359	4848.48545	0.897966
27	Berliner Weissbier	933	852.68170	0.897831
83	Oatmeal Stout	6720	6067.65060	0.896843
57	Flanders Oud Bruin	1854	1684.03300	0.896811
98	Tripel	11628	10466.83340	0.895483
94	Scotch Ale / Wee Heavy	6557	5910.06200	0.895135

Q6: How does written review compare to overall review score for the beer styles?

To see this we have ranked beer styles based on overall review scores using bayesian_rank and we look at the top 20 styles again. We observe that there is a lot of intersection in the both the rankings, i.e rankings based on review text and overall rating. We see almost the same beer styles jumbled in positions but in the same ranking vicinity. We also see that **American Double/Imperial Stout** is still the favorite beer style.

This comparison can get better with more advanced sentiment analysis algorithms which use a hybrid of statistical metrics and ML.

Q6: How does written review compare to overall review score for the beer styles?

Top 20 beer styles based on overall ratings and review text(previous question)

	beer_style	N	S	bayesian_rank
11	American Double / Imperial Stout	23352	19150.7	0.815913
63	Gueuze	1575	1304.4	0.812091
83	Oatmeal Stout	6720	5484.6	0.808270
27	Berliner Weissbier	933	771.4	0.805655
90	Rye Beer	5179	4215.7	0.804957
86	Quadrupel (Quad)	4933	3994.9	0.800487
89	Russian Imperial Stout	17183	13834.7	0.800118
25	Belgian Strong Dark Ale	15403	12351.2	0.796532
38	Dortmunder / Export Lager	1809	1466.0	0.794845
12	American IPA	43364	34602.0	0.794748
74	Lambic - Fruit	3768	3034.8	0.794612
9	American Double / Imperial IPA	26101	20625.5	0.786038
21	Baltic Porter	4109	3262.2	0.783362
20	American Wild Ale	3695	2934.6	0.783077
58	Flanders Red Ale	2856	2263.0	0.779641
4	American Barleywine	10107	7924.8	0.777281
75	Lambic - Unblended	705	565.5	0.776528
98	Tripel	11628	9098.9	0.776138
100	Weizenbock	2235	1766.2	0.775780
5	American Black Ale	3055	2404.9	0.774797

	beer_style	N	S	bayesian_rank
86	Quadrupel (Quad)	4933	4582.27060	0.922685
11	American Double / Imperial Stout	23352	21582.97050	0.921351
58	Flanders Red Ale	2856	2642.82700	0.916937
38	Dortmunder / Export Lager	1809	1674.03310	0.914707
25	Belgian Strong Dark Ale	15403	14122.05610	0.913112
90	Rye Beer	5179	4739.14050	0.908513
84	Old Ale	4817	4408.55015	0.908408
101	Wheatwine	891	821.96225	0.906734
87	Rauchbier	2607	2385.85675	0.905841
74	Lambic - Fruit	3768	3437.93710	0.904580
20	American Wild Ale	3695	3371.01640	0.904412
4	American Barleywine	10107	9181.79685	0.903644
100	Weizenbock	2235	2041.11175	0.903044
89	Russian Imperial Stout	17183	15553.12120	0.901410
37	Doppelbock	5359	4848.48545	0.897966
27	Berliner Weissbier	933	852.68170	0.897831
83	Oatmeal Stout	6720	6067.65060	0.896843
57	Flanders Oud Bruin	1854	1684.03300	0.896811
98	Tripel	11628	10466.83340	0.895483
94	Scotch Ale / Wee Heavy	6557	5910.06200	0.895135

Q7: How do find similar beer drinkers by using written reviews only?

Since here we can not use any data except review text, the best we can do is look at the similarity in text to get similar beer drinkers. Therefore, we have explored different text similarity metrics in which Cosine Similarity, Path Similarity, Leacock Chordorow (LCH) Similarity, and WuPalmer (WuP) Similarity were meaningful. So, we have made a function get_sim_scores which takes in 2 review texts and returns all the above similarity values.

Since Cosine Similarity needs a vector representation, we used a Gensim pre-trained model for word2vec.

Extension work to this can be a using a clustering algorithm (ML) with these similarities and other metrics (this is definitely required) as features and finding similar users for any new data coming in.

Q7: How do find similar beer drinkers by using written reviews only?

Here reviews at 0 and 1 are given by the same user (based on profileName) and 4 is a different user.

We see that all the similarity scores are higher for (0 and 1) than (0 and 4).

We also show that since these scores are normalized, the closer to 1.0 they are the more similar they are to each other. We show this by getting the scores for identical texts.

```
[17]  get_sim_scores(data['review_text'][0], data['review_text'][1])
      ✓ 5.7s

...   {'Avg.Path Similarity': 0.40724919517378755,
       'Avg.LCH Similarity': 0.7309348617050986,
       'Avg.WUP Similarity': 0.5482606802569114,
       'Cosine Similarity': 0.9701839089393616}

[18]  get_sim_scores(data['review_text'][0], data['review_text'][4])
      ✓ 1.4s

...   {'Avg.Path Similarity': 0.315969504587293,
       'Avg.LCH Similarity': 0.668906750596864,
       'Avg.WUP Similarity': 0.4804201105780893,
       'Cosine Similarity': 0.9309957027435303}

[19]  get_sim_scores(data['review_text'][1], data['review_text'][1])
      ✓ 0.5s

...   {'Avg.Path Similarity': 1.0,
       'Avg.LCH Similarity': 0.9970699303597836,
       'Avg.WUP Similarity': 1.0,
       'Cosine Similarity': 1}
```

Thank You

Thank you for giving me the opportunity to work on this exercise. I was able to achieve good results for the questions with sophisticated statistical analysis methods. Hope the answers mostly align (since it is statistics at the end of the day) with the expected answers. Looking forward to your hearing back from you and your feedback.