Beer Recommendation System

--Milestone report 1

Outline

1. Introduction

It's common to see someone staring at a wall of beers at local supermarket, contemplating for over 10 minutes before grabbing one. They are not alone. As a beer enthusiast, I'm always looking for something new to try but I also dread being disappointed so I often spend too much time looking up a particular beer over several websites to find some kind of reassurance that I'm making a good choice. So it's necessary to build a recommendation system to leverage users historical ratings to create predictions for un-tasted beers and create a personalized recommendation for which beer to sample next.

My target audience will be the beer enthusiast, believe that drinker will find the results interesting. There are more than 2,748 different breweries operating in the US as of June, 2018 and more than 5,000 unique brands in the U.S. which may cause newfound confusion for the average consumer when shopping for beer. Given the huge rise in the brewery scene in the United States, it can be overwhelming trying to decide which new brewery or beer to try. My recommendation system will help them decide which beer to try next given their unique tastes and historical ratings.

1.1 Data source information

This dataset contains around 1.5 million reviews of Beer from BeerAdvocates. The dataset can be downloaded at https://www.kaggle.com/rdoume/beerreviews. The data (beer_reviews.csv) is in a .csv format consisting of 1,586,614 observations and 13 features: brewery_id, brewery_name, review_time, review_overall, review_aroma, review_appearance, review_profilename, beer_style, review_palate, review_taste, beer_name. The meaning of those features can be found in Table 1.

Table 1 Features information from beer reviews dataset

Feature name	Count	Missing	Type	Note
brewery_id	1586614	0	int64	An identifier for the brewery
brewery_name	1586599	15	object	The name of brewery
review_time	1586614	0	int64	A dict specifying when the review was submitted
review_overall	1586614	0	float64	Rating of the beer overall (1.0 to 5.0)
review_aroma	1586614	0	float64	Rating of the beer's aroma (1.0 to 5.0)
review_appearance	1586614	0	float64	Rating of the beer's appearance (1.0 to 5.0)
review_profilename	1586266	348	object	The profile name of the reviewers
beer_style	1586614	0	object	The style of beer
review_palate	1586614	0	float64	Rating of the beer's palate (1.0 to 5.0)
review_taste	1586614	0	float64	Rating of the beer's taste (1.0 to 5.0)
beer_name	1586614	0	object	Name of the beer
beer_abv	1518829	67785	float64	The alcohol by volume of the beer
beer_beerid	1586614	0	int64	A unique ID indicating the beer reviewed

1.2 Data Cleaning

The data cleaning aimed to deal with missing value, duplicates and outliers among all features. For review time, data were converted to YYYY-MM-DD format. As for the 5 review indexes, any values falling outside of 1 to 5 were dropped. Moreover, review recode with count less than 5 for any specific beer were discarded too. The main cleaning steps were listed in Table 2.

Table 2 Main cleaning steps for the dataset

Feature name	Cleaning steps
brewery_name	▪ Drop missing value. ▪ Keep name started with letter or number (a-zA-Z0-9)
review_time	▪ Convert to YYYY-MM-DD format
review_profilename	▪ Fill 'nan' with 'Unknown'
beer_abv	▪ Fill missing value with the average abv of the style
beer_beerid	▪ Count the number of id. ▪ Discard the id with review count <5
5 review indexes*[*]**	▪ Discard the review small than 1 and larger than 5

Note: 5 review indexes stand for review_overall, review_aroma, review_appearance review_palate, review_taste

2. Data Exploration

2.1 Describe statistics

Describe statistics analysis were conducted for the 5 review indexes and the alcohol content (Table 3): Describe statics showed that there are no big variations existing in features probably due to fixed rating range from 1 to 5. For instance, the overall review(review_overall) varies from 1.14 to 4.9 with a mean of 3.71. However, we observed large variation for the alcohol content (beer_abv), which varies from 0.05% to 41% with a mean of 6.58%.

Table 3 Describe statistics of the 5 review indexes and alcohol content for beers

	review _overall	review _aroma	review _appearance	review _palate	review _taste	beer_abv (%)
Count	19209	19209	19209	19209	19209	19209
Mean	3.71	3.62	3.74	3.63	3.66	6.58
Std	0.45	0.48	0.38	0.44	0.50	2.18
Min	1.14	1.17	1.50	1.23	1.15	0.05
25%	3.50	3.40	3.57	3.42	3.42	5.00
50%	3.79	3.69	3.81	3.70	3.75	6.00
75%	4.00	3.94	4.00	3.92	4.00	7.80
Max	4.94	5.00	4.93	4.94	4.94	41.00

The data encapsulated 5230 unique breweries; 4,900 unique beers; approximately 32908 unique user(reviewers) and 22638 of them have rated more than once; approximately 1.5million user-review pairings. It contains over 16 years of data leading up to 2012 (Fig. 1a). The most review number appeared in 2011.Based on the cumulative distribution of the review indexes (Fig. 1b), it is obvious that alcohol content is not normal distributed.

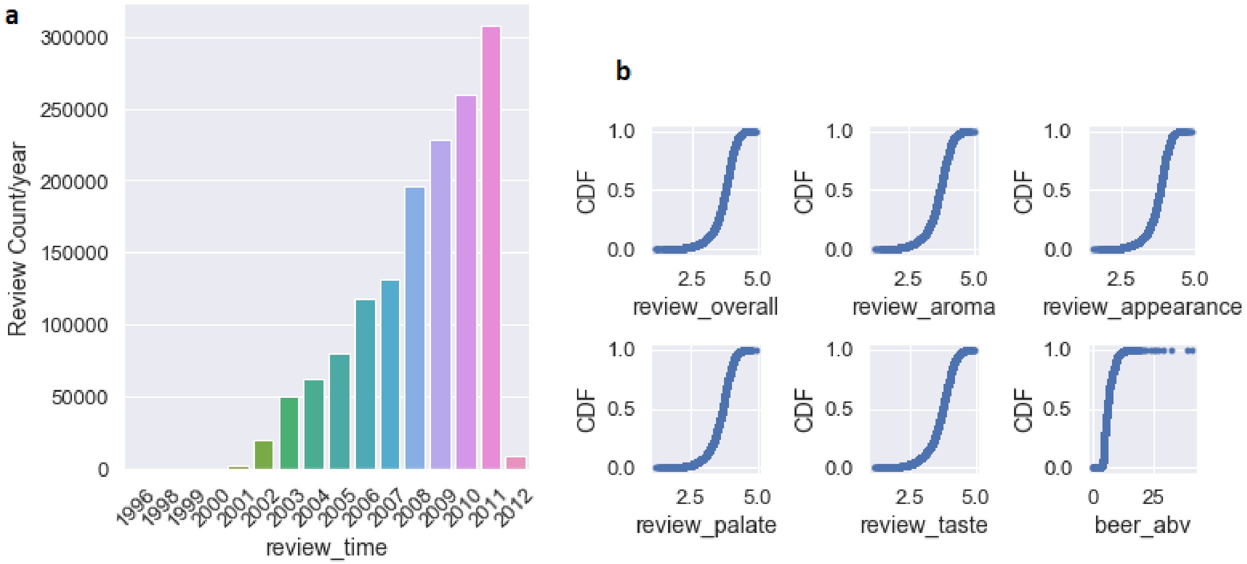

Figure 1 Review time (a) distribution and review indexes (b) cumulative distribution

Kolmogorov-Smirnov test was used to test normal distribution of the review indexes and alcohol content. Since p value of the all tests were close to 0, which suggested that their distribution sufficiently different from the normal distribution, that we can reject the hypothesis that the sample came from the normal distribution.

2.2 Ranking and sorting beer brand, style and brewery

In order to rank the different beer and brewery, mean value of the 5 review indexes were calculated for each beer brand and brewery. The ranking of beer brand, style and brewery were based on the mean value (Table 4,5,6).

Table 4 Top ten the most popular beer brand

Rank	**Beer Name**	**Review average**
1	Alesmith Speedway Stout - Vanilla And Coconut	4.887500
2	El Gordo	4.828571
3	M Belgian-Style Barleywine	4.735714
4	Capricho Oscuro - Batch 1	4.716667
5	Coffee Infused Imperial Stout Trooper	4.700000
6	Cantillon La Dernière Cuvée Du 89	4.683333
7	Barrel Aged Stout	4.680000
8	Armand'4 Oude Geuze Lente (Spring)	4.673846
9	Kopi Con Leche Stout	4.671429
10	Hitchhiker	4.666667

The number 1 of most popular beer is Alesmith Speedway Stout - Vanilla And Coconut. This brand of beer has chocolate and roasted malts dominate the flavor, supported by notes of dark fruit, toffee, and caramel. A healthy dose of locally-roasted coffee added to each batch brings out the beer's dark chocolate flavors and enhances its drinkability.

The most popular beer style is Quadrupel (Quad). Inspired by the Trappist brewers of Belgium, the Quadrupel is a Belgian-style ale of great strength with even bolder flavor compared to its sister styles Dubbel and Tripel. Typically a dark creation that plays within the deep red and ruby brown end of the spectrum with garnet hues it is a full bodied beer with a rich, malty palate and spicy phenols than are usually kept to a moderate level. Sweet on the palate with a low bitterness yet well perceived alcohol, Quads are well suited to cellaring.

Table 5 Top ten the most popular beer styles

Rank	Beer style	Review AVG*
1	Quadrupel (Quad)	4.135505
2	American Double / Imperial Stout	4.131664
3	Russian Imperial Stout	4.113477
4	American Wild Ale	4.097805
5	Eisbock	4.088967
6	Gueuze	4.084845
7	American Double / Imperial IPA	4.061975
8	Lambic - Unblended	4.038159
9	Weizenbock	4.034948
10	Flanders Red Ale	4.026596

Note: Review AVG stand for the mean value of the 5 review indexes

The best brewery is Brouwerij Westvleteren based on the average reviewing of its beer product. This brewery founded in 1838 at the Trappist Abbey of Saint Sixtus in Vleteren, Belgium. The brewery's three beers have acquired an international reputation for taste and quality; Westvleteren is considered by some to be the best beer in the world. The beers are not brewed to normal commercial demands but are sold in small quantities weekly from the doors of the monastery itself to individual buyers on an advance-order basis.

Table 6 Top ten breweries based the rating of their beers

Rank	Brewery Name	Brewery ID	Review AVG
1	Brouwerij Westvleteren	313	4.51
2	Närke Kulturbryggeri AB	10902	4.48
3	The Alchemist	27039	4.47
4	Micro Cervejaria Falke Bier	10287	4.47
5	Old Chimneys Brewery	6560	4.43
6	Benny Brewing Company	24757	4.42
7	Kane Brewing Company	26676	4.41
8	Denison's Brewing Company & Restaurant	662	4.40
9	Breakwater Brewing	17988	4.38
10	Brauerei Zehendner GmbH	5983	4.38

2.3 Relationship among the features

Third, relationships among the features were analyzed using scatter plots and correlation analysis (Figure 2). The Person correlation coefficient ranged from 0.84 to 0.93. among the 5 review indexes. The results showed that, except for alcohol content, all 5 review indexes showed close relationships with each other.

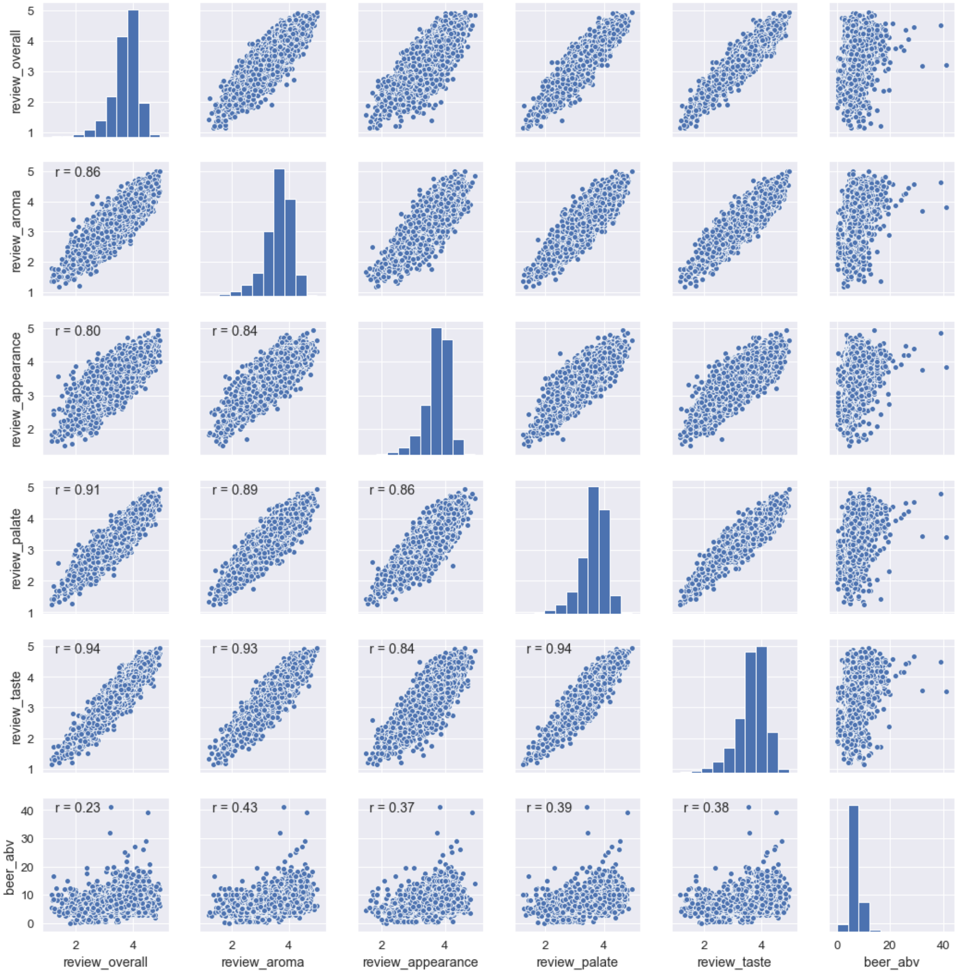

Figure 2 Histogram and scatter plot showing the relationships among the features.

However, the correlation coefficient between alcohol content and the 5 review indexes is relatively low with range of 0.23 to 0.43. The results suggest that alcohol content dose not play important role for beer rating. The histogram of the 5 review indexes (Fig. 2) showed that rating score slightly skewed to right side indicated that there are more people like the beer they rated than who did not.

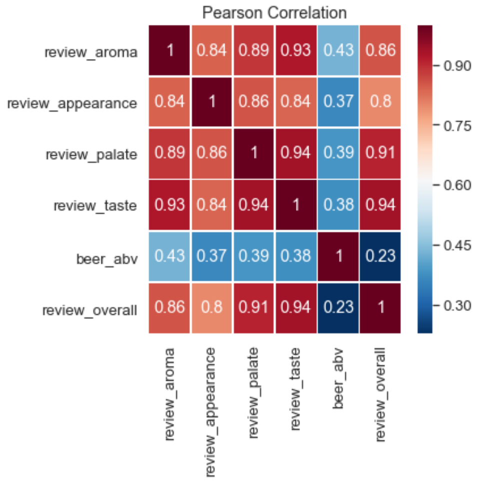

Figure 3 Correlation coefficient among 5 review indexes and alcohol content

The p-value for testing non-correlation were calculated for all pair-wise features (Table 7). The p-value roughly indicates the probability of an uncorrelated system producing datasets that have a Pearson correlation at least as extreme as the one computed from these datasets. The results showed that the p-values are close to 0 or equal to 0., which indicate that all pair-wise features were significantly associated with each other.

Table 7 The p-value for testing non-correlation were calculated for all pair-wise features

	review_overall	**review_aroma**	**review_appearance**	**review_palate**	**review_taste**
review_aroma	0.0				
review_appearance	0.0	0.0			
review_palate	0.0	0.0	0.0		
review_taste	0.0	0.0	0.0	0.0	
beer_abv	9.28e-274	0.0	0.0	0.0	0.0