Beer Recommendation System

Zixiang Wen
Department of Plant, Soil and Microbial Sciences
Michigan State University

1. Introduction

➢ Objective:

- 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.

➢ Target audience:

- Beer enthusiast
- Breweries operating in the US

1.1 Data source information

- **Download address:**
https://www.kaggle.com/rdoume/beerreviews

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	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)
Re_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

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 average value
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

2. Data Exploration

- **2.1 Describe statistics**

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

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%.

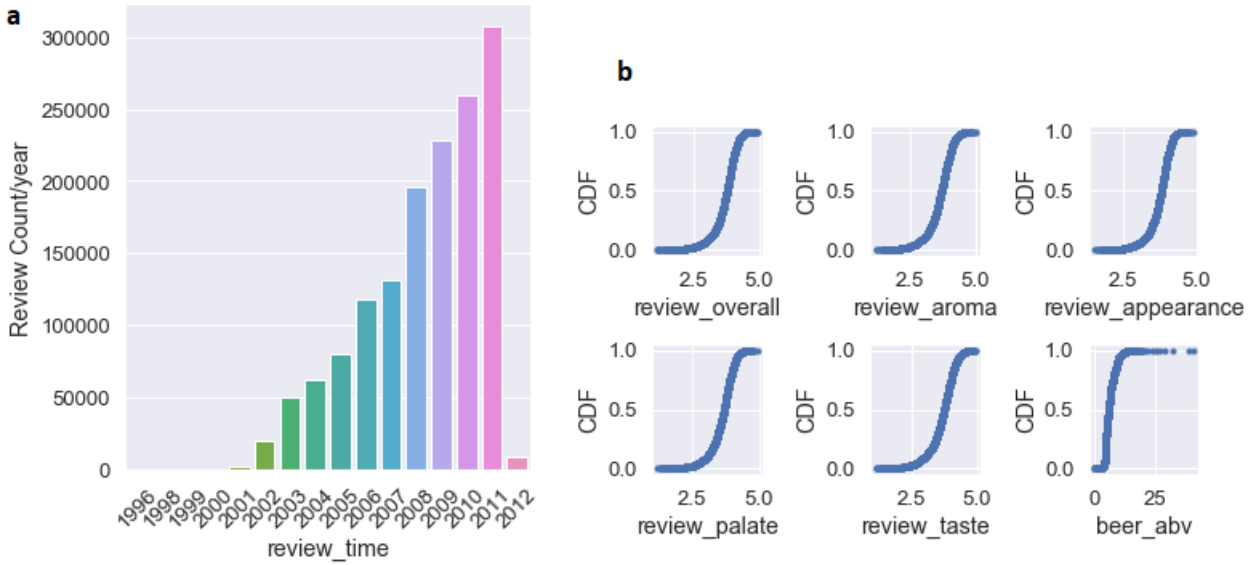

Figure 1 Review time (a) distribution and review indexes (b) cumulative distribution

2.2 Ranking and sorting beer brand, style and brewery

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

Table 5 Top ten the most popular beer style

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

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.

Table 6 Top ten highly rated brewery

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

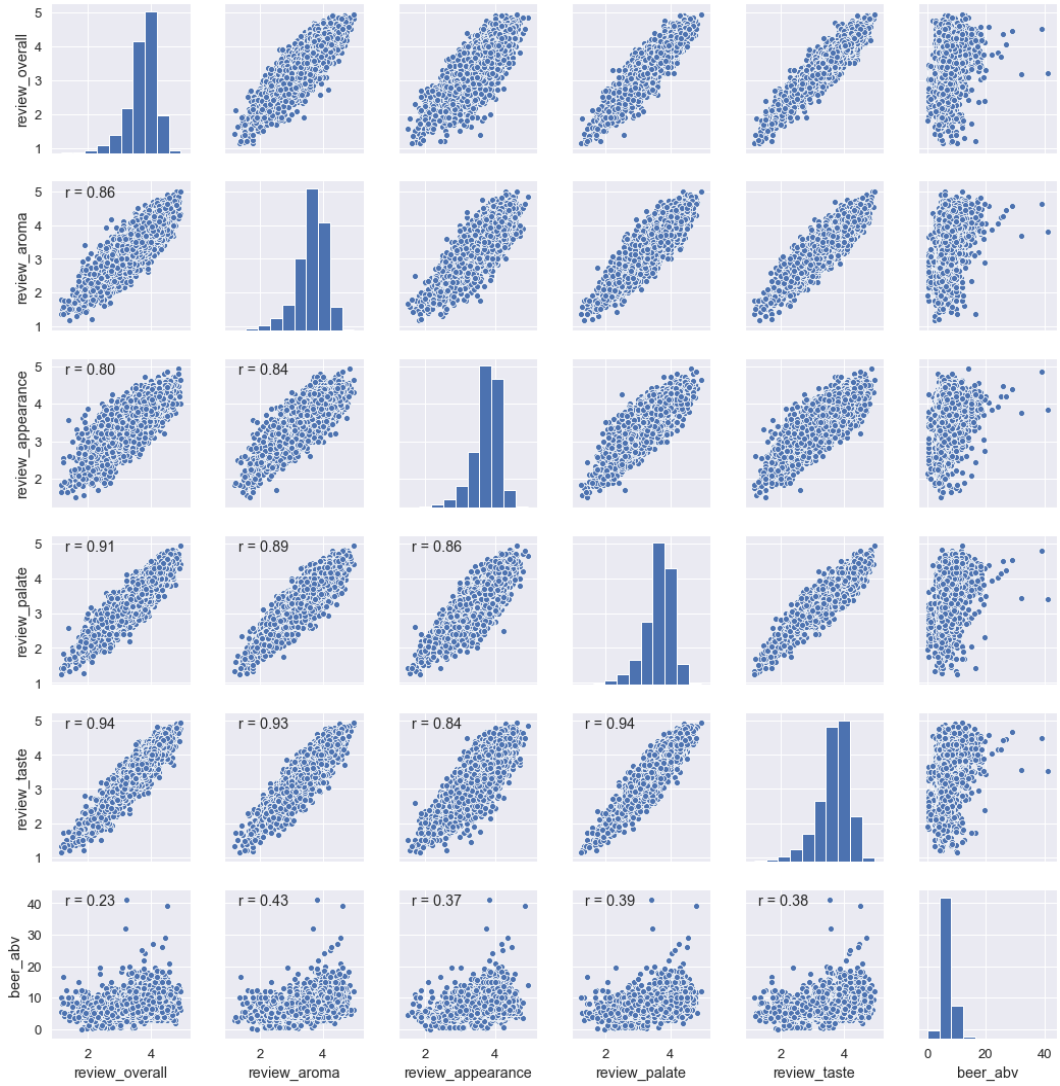

Figure 2 Histogram and scatter plot showing the relationships among the features.

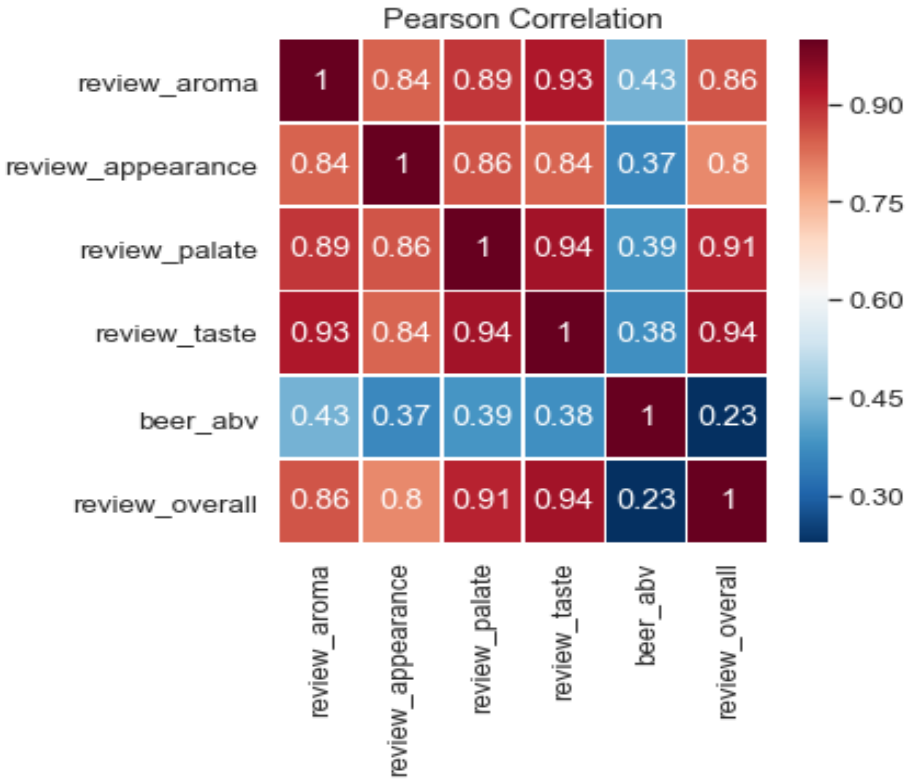

Figure 3 Correlation coefficient among 5 review indexes and alcohol content

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.

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

2.4 Relationship between the beer alcohol content and rating score

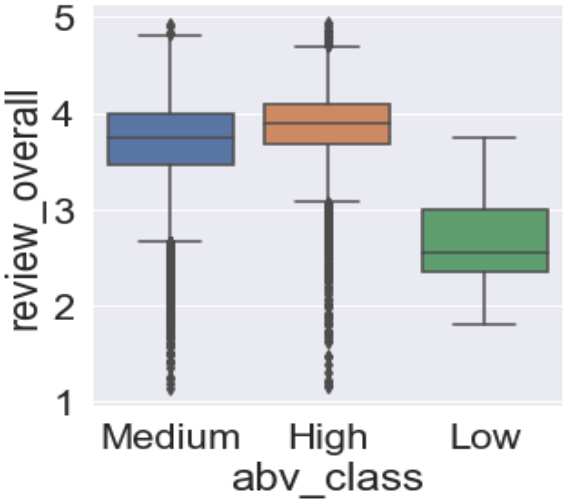

Figure 4 Box plot showing the difference in rating among three ABV classes

Table 8 ANOVA analysis of rating difference among different ABV classes

	Sum_sq	df	F	PR(>F)
abv_class	128.00	2	319.42	3.43E-137
Residual	3848.39	19206	NaN	NaN

2.5 Relationship between the beer style and rating

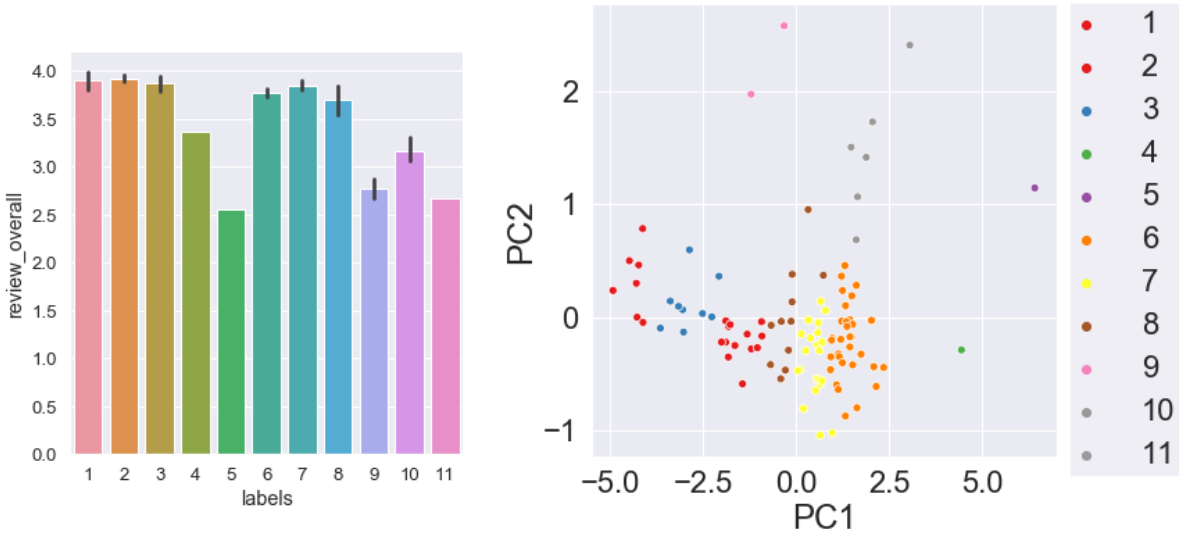

Figure 6 Bar plot and scatter plot showing the difference in rating score among 11 clusters and beer style distribution

3.Building and testing recommender systems with Surprise

- 3.1 Choose the best algorithms

Table 9 Algorithms comparison for the beer dataset

Algorithm	test_rmse	fit_time	test_time
SVDpp	0.598	7313.033	337.3621
BaselineOnly	0.601	3.928861	6.072451
KNNBaseline	0.602	82.24511	232.1964
SVD	0.604	77.26623	6.828465
KNNBasic	0.620	76.56187	209.4991
NMF	0.625	79.14518	5.830669
KNNWithMeans	0.625	77.10334	219.8093
CoClustering	0.666	40.16992	6.11246
NormalPredictor	0.989	2.286215	6.465656

SVDpp algorithms has best performance with the lowest RMSE (0.5975). However, this algorithm is also the most time-consuming one

3.2 Tuning and evaluating the SVDpp model

Hyperparameters tuning:

```
param_grid = {'n_epochs': [5, 10], 'lr_all': [0.001, 0.01],
              'reg_all': [0.1, 0.5]}
gs = GridSearchCV(SVDpp, param_grid, measures=['rmse', 'mae'], cv=3)

gs.fit(data)

# best RMSE score
print(gs.best_score['rmse'])

# combination of parameters that gave the best RMSE score
print(gs.best_params['rmse'])
```

The results showed that best hyperparameters combination is 10 for n_epochs 10, 0.01 for lr_all, and 0.1 for reg_all. The corresponding accuracy measured by RMSE score was 0.603.

3.3 Making recommendations with SVDpp model

Table 10 Top 10 the best predictions for the test dataset

uid	iid	rui	est	Details (was_impossible)	Iu	Ui	err
oteyj	44910	5	5	False	67	28	0
oteyj	51116	5	5	False	67	131	0
FARGO619	21690	5	5	False	36	472	0
FARGO619	7971	5	5	False	36	1907	0
billzav	7971	5	5	False	21	1907	0
paxtonthegreat	7971	5	5	False	10	1907	0
lemasney	7971	5	5	False	78	1907	0
oteyj	731	5	5	False	67	1472	0
kdawg105	6549	5	5	False	53	991	0
hippityead24	7971	5	5	False	23	1907	0

Note: uid — the user ID, for whom we carry out predictions, iid — item ID.(here we treat movies as items), est — estimated rating for an item, as we expect the user to give, U**1** : the set of all users that have rated item i. Iu : the set of all items rated by user u. err, the difference between predicted and true value.

Table 11 The best recommendation for user NaLoGra

Brewery name	Beer name	Beer id	Predicted rating
Southampton Publick House	Southampton Berliner Weisse	8626	4.87
Kern River Brewing Company	Citra DIPA	56082	4.77
Brouwerij Drie Fonteinen	Geuze Cuvée J&J Blauw (Blue)	23413	4.8
Brouwerij Drie Fonteinen	Armand'4 Oude Geuze Zomer (Summer)	70356	4.8
Brouwerij Drie Fonteinen	Armand'4 Oude Geuze Lente (Spring)	68548	4.89
Brouwerij Drie Fonteinen	Geuze Cuvée J&J	23414	4.76
The Lost Abbey	Veritas 004	45957	4.81
The Lost Abbey	Veritas 005	54147	4.79
FiftyFifty Brewing Co.	Imperial Eclipse Stout - Pappy Van Winkle	47668	4.77
Brouwerij Westvleteren	Trappist Westvleteren 12	1545	4.76

4. Discussion and Summary

- A beer recommendation system with SVDpp algorithm was built.
- The age of the data is something that in future analysis can and should be considered.
- Geographical information should be taken into account in the recommendation engine.
- Deep learning, such as Autoencoders and Restricted Boltzmann machines, should be used in the future.

- Thanks for your attention!