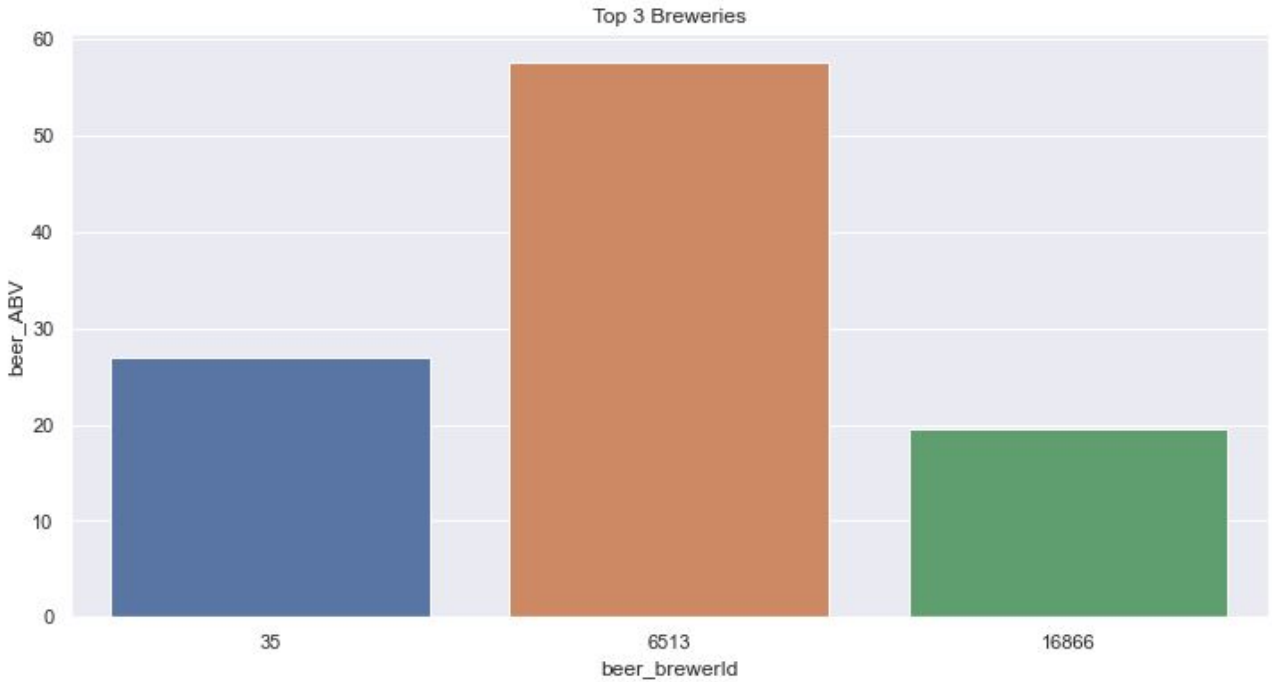
Evolent Health

Beer data analysis

Select top 3 Breweries which produce the strongest beers

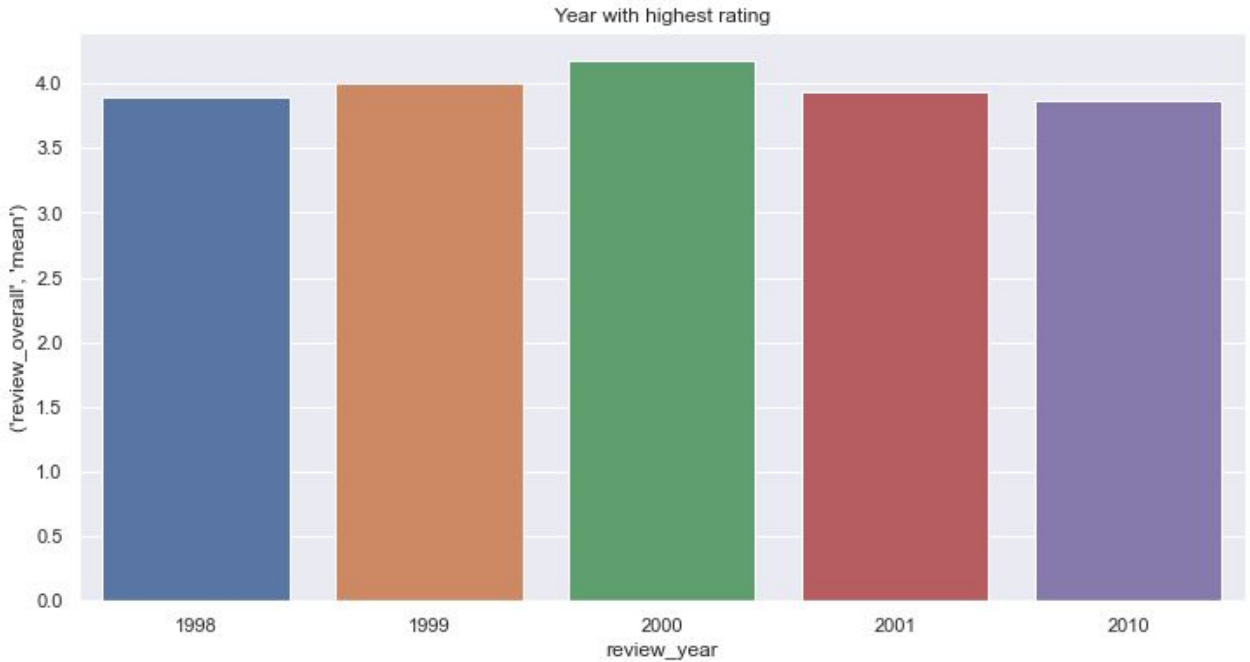
In the given dataset

these are the Breweries which produce strong beers, where brewer with ID 6513 produces 4 quite strong beers with ABV more than 30.



Year with highest beer rating

According to given data year 2000 had highest average rating for beer but, adjacent years of 2000 as well enjoyed highest rating for beer where average ratings were more than 3.5.



Factor importance based on the user's ratings

Below chart represents correlation metrics between different beer factors and with overall review.

Review_aroma, review_taste and review_palette are showing high correlation to overall review so these factors can be considered as important ones with respect to final ratings.

Though this can be further investigated by training a model (eg. random forest) or by using dimensionality reduction technique like PCA.

	review_overall	review_appearance	review_palette	review_taste	review_aroma
review_overall	1.000000	0.486267	0.601750	0.692217	0.783030
review_appearance	0.486267	1.000000	0.547188	0.554188	0.533646
review_palette	0.601750	0.547188	1.000000	0.603794	0.705715
review_taste	0.692217	0.554188	0.603794	1.000000	0.724914
review_aroma	0.783030	0.533646	0.705715	0.724914	1.000000

Recommend 3 Beers

Based on the given data we can recommend these 3 beers.

First table contains 2 beers which are popular and have high ratings.

For the third recommendation we can consider a beer which is less popular but has fare amount of reviews and really good ratings. It might add little surprising element to the recommendation and help user to explore new beer.

	review_overall	
	mean	count
beer_name		
Sierra Nevada Celebration Ale	4.172184	2930
Sierra Nevada Pale Ale	4.252375	2526

	review_overall	
	mean	count
beer_name		
Citra DIPA	4.630081	246

Finding the favorite Beer style

As reviews text don’t have style specific reviews, analysing the whole review seemed good idea.

Did sentiment analysis on all reviews and predicted rating based on the written reviews.

These predicted rating and beer_style’s popularity are considered while finding the favorite beer style i.e. **American IPA**.

Received similar results with overall ratings, which indicate that the sentiment analysis was able to capture the sentiments correctly from written review.

	predicted_rating	
	mean	count
beer_style		
American IPA	3.603272	42664
American Double / Imperial IPA	3.519635	25770

written review compare to overall review score for the beer styles

- Used sentiment analysis based predicted ratings and overall review values to classify review in 3 categories positive, neutral and negative.
- It seems Overall_review has more positive

reviews than the one predicted from written

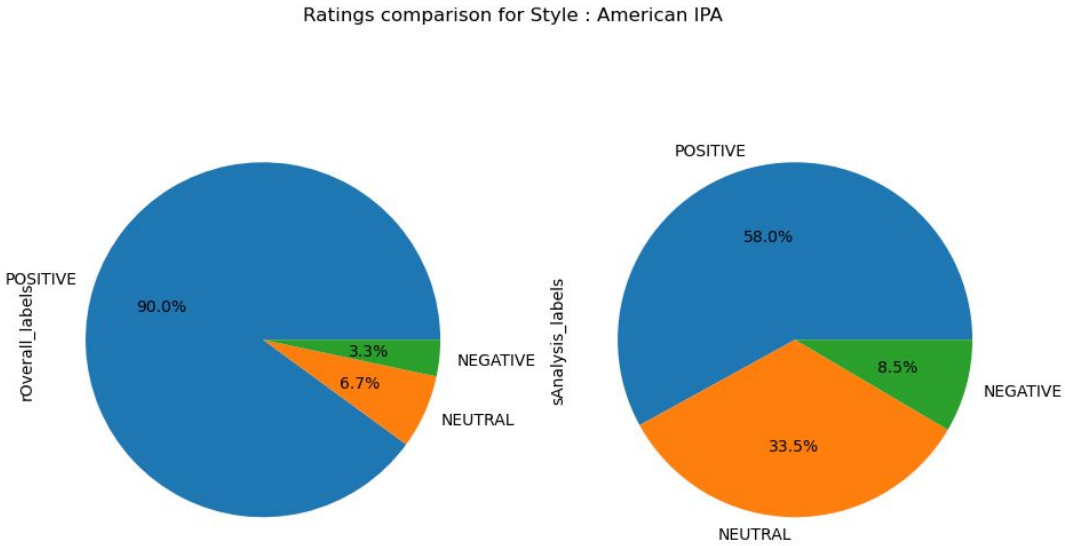
reviews for the American IPA.

- For other styles as well written

reviews are little more inclined toward

neutral sentiment as compare to overcall

score.



Similar beer drinkers based on written reviews

- To find similar beer drinkers here avg w2v, knn(to get right parameter for dbscan) and dbscan clustering algorithm has been used.
- w2v yields vector for each review and as dbscan is density based clustering algorithm so dbscan can leverage it to find the similarity between reviews.
- With similar reviews clustered together we can find out associated beer drinkers who are writing about similar beers.