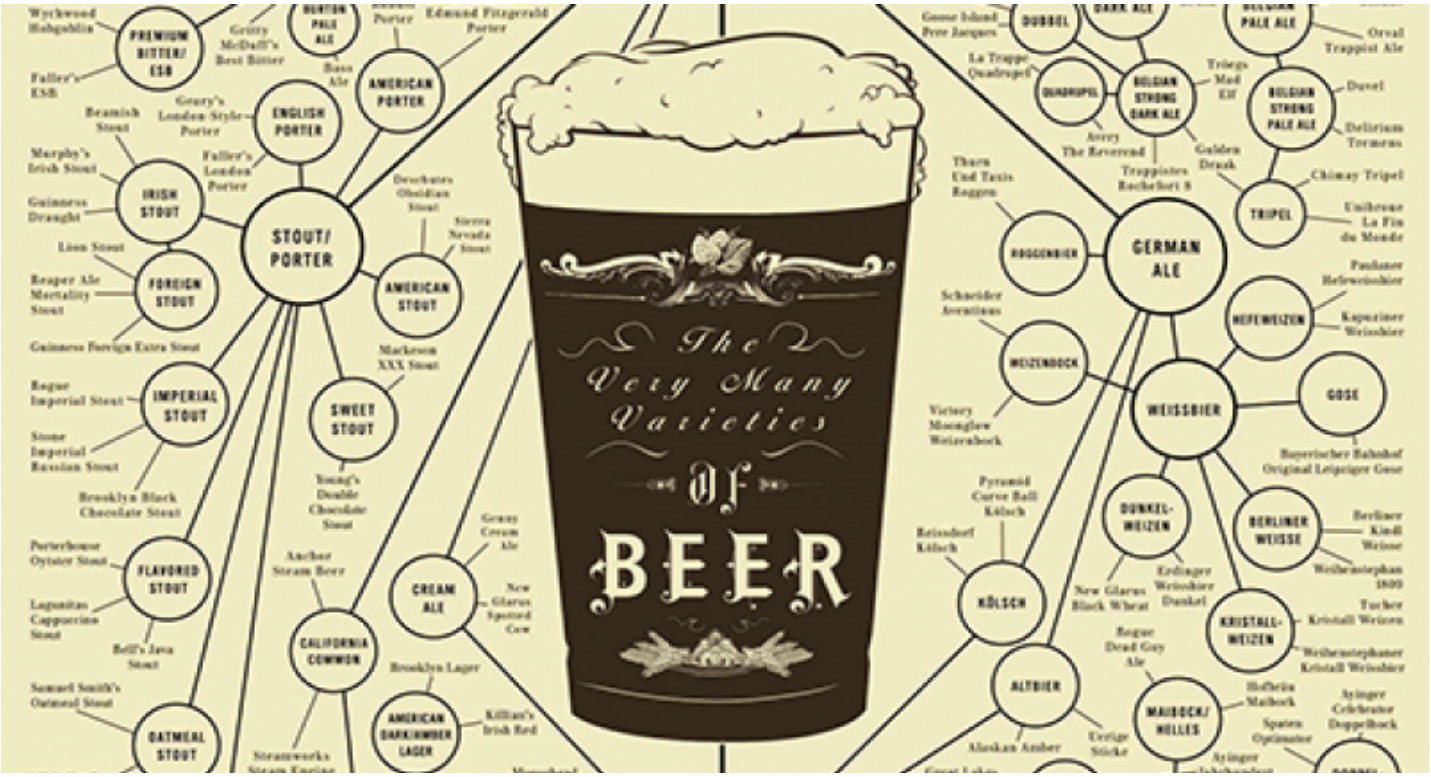
Data Scientist Exercise

Jeff Whitmire 3/18/19

This dataset consists of approximately 1.5 million beer reviews from Beer Advocate.



1) Which brewery produces the beer with the strongest ABV%?
2) If you had to pick 3 beers to recommend using only this data, which 3 would you pick?
3) Which of the factors (aroma, taste, appearance, palate) are most important in determining the overall quality of the beer?
4) If I typically enjoy a beer due to its aroma and appearance, which beer style should I try?

1) Which brewery produces the beer with the strongest ABV%?

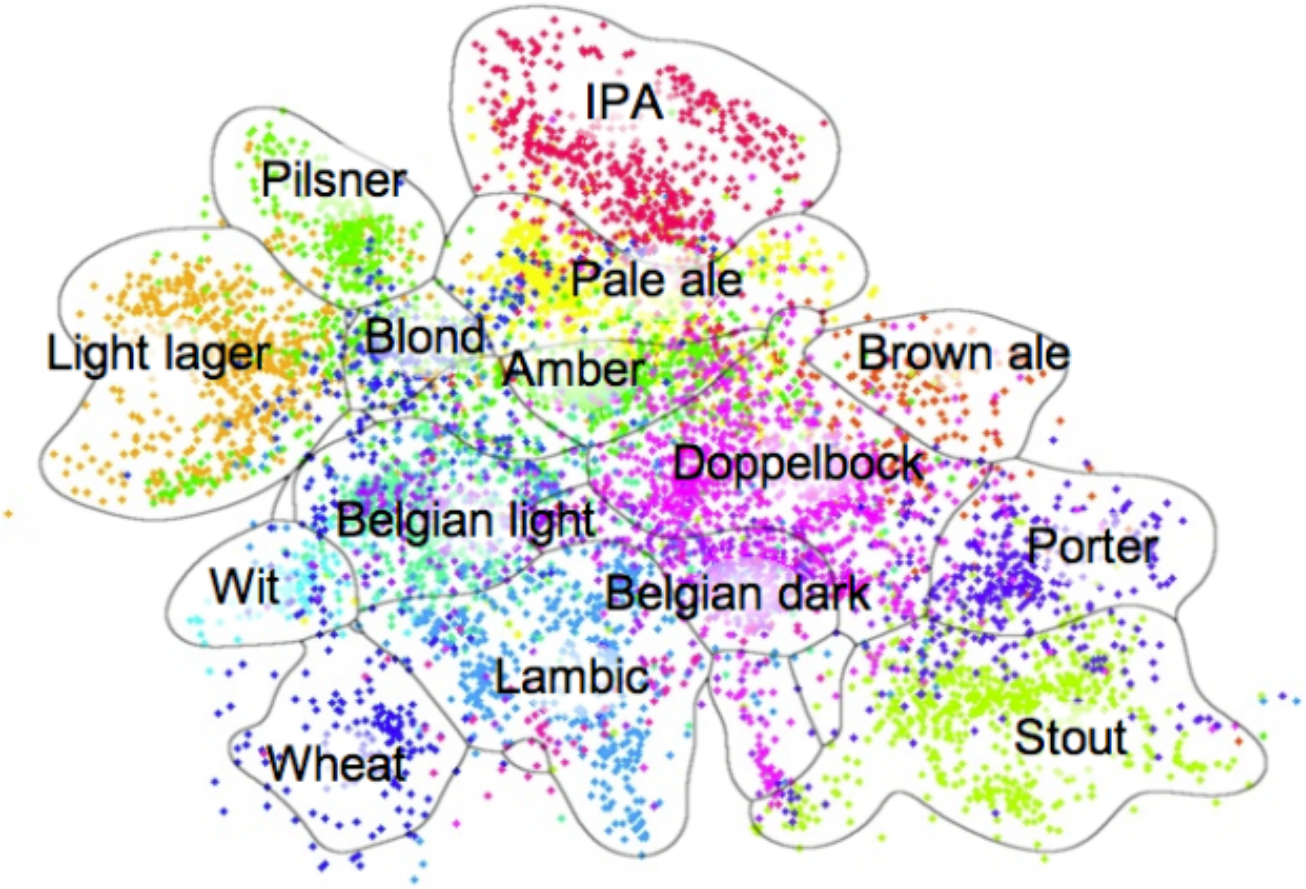
brewery_name	brewery_id	review_overall	review_aroma	review_appearance	review_palate	review_taste	beer_abv	beer_beerid
Schorschbräu	6513.0	3.411765	3.529412	3.558824	3.470588	3.514706	19.228824	34235.676471
Shoes Brewery	14060.0	3.000000	3.000000	3.750000	3.500000	3.250000	15.200000	32949.000000
Rome Brewing Company	2873.0	4.100000	3.600000	3.800000	3.900000	4.400000	13.840000	14293.000000

brewery_name	review_time	review_overall	review_aroma	review_appearance	review_profilename	beer_style	review_palate	review_taste	beer_name	beer_abv
Schorschbräu	1316780901	4.0	4.0	4.0	kappldav123	Eisbock	4.0	3.5	Schorschbräu Schorschbock 57%	57.7
Schorschbräu	1309974178	4.0	4.0	3.5	Sunnanek	Eisbock	4.0	4.0	Schorschbräu Schorschbock 43%	43.0
Schorschbräu	1274469798	3.5	4.0	4.0	kappldav123	Eisbock	4.0	4.5	Schorschbräu Schorschbock 43%	43.0

Schorschbrau brewery produces the single beer with the most alcohol by volume. In fact they produce the top 3 in this category. They also produce the highest mean alcohol content across all beers produced

Schorschbrau

Data Scientist Exercise



2) If you had to pick 3 beers to recommend using only this data, which would you pick?

I tend to prefer an ale over most other beer types. I queried the data frame to build a new dataset that only contains Ale's, then group the data by the beer name, calculate the mean reviews across all categories and sort the information based on the returned data.

After the manipulation is complete, we only end up with 3 beers with a total_review rating of 5 across all categories.

These are the beers I would recommend trying:

Great Lakes Truth Justice And The American AleEdsten Tripl_Wit and Engelbert Moonbeam.

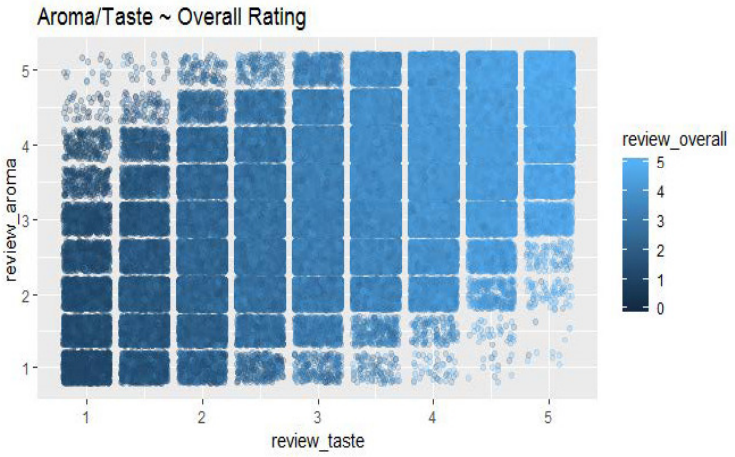
beer_name	brewery_id	review_overall	review_aroma	review_appearance	review_palate	review_taste	beer_abv	beer_beerid	tot_review
Great Lakes Truth Justice And The American Ale	73.0	5.000000	5.000000	5.000000	5.000000	5.000000	4.9	75829.0	5.000000
Edsten Triple-Wit	387.0	5.000000	5.000000	5.000000	5.000000	5.000000	10.0	1734.0	5.000000
Engelbert Moonbeam	642.0	5.000000	5.000000	5.000000	5.000000	5.000000	10.0	1890.0	5.000000
Lips Of Faith - Eric's Ale (Bourbon Barrel Aged)	192.0	5.000000	4.750000	4.750000	5.000000	5.000000	9.0	68665.0	4.900000
Opus Altar Boy	30.0	5.000000	5.000000	5.000000	5.000000	4.500000	10.0	52522.0	4.900000
Dry Hopped Abominable Ale	16353.0	5.000000	5.000000	5.000000	4.500000	5.000000	7.3	76150.0	4.900000

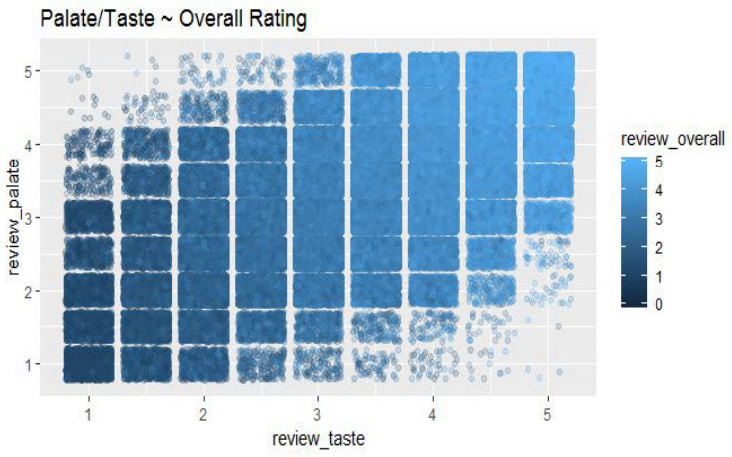
3) Which of the factors (aroma, taste, appearance, palate) are most important in determining the overall quality of the beer?

	review_overall	review_aroma	review_taste	review_appearance	review_palate
review_overall	1	0.612669	0.787111	0.498401	0.698925
review_aroma	0.612669	1	0.714677	0.558925	0.614781
review_taste	0.787111	0.714677	1	0.544432	0.73211
review_appearance	0.498401	0.558925	0.544432	1	0.564407
review_palate	0.698925	0.614781	0.73211	0.564407	1

This heatmap shows the relationship of the variables to each other. From shades of blue to red, we see all these variables positively correlate with review_overall. Taste has the highest coefficient at .79. Palate also has a high coefficient.

I will explore the interaction of a few of the variables with R to determine those relationships.





Using a histrogram in R to compare the variables and coloring that comparison by the overall review I can see there is clearly a positive releationship. Seeing the lighter colors of blue in the upper right quadrant of these two charts not only proves the results of the correlation testing, it further underscores the importance of the interaction of the variables. It can be seen in the chart on the right that if palate reviews are low, overall ratings suffer somewhat regardless of the taste.

Taste, palate and aroma are the 3 most important variables, in that order. However, the interaction between the 3 cannot be dismissed.

Lastly, if I typically enjoy a beer due to its aroma and appearance, which beer style should I try?

I created a new dataset that contains the beer style aroma, appearance and overall rating. After finding the mean rating by style and sorting the list in descending order.

American Wild Ale, Gueuze and Quadruple(Quad) would be the 3 recommended beer types based on ratings.

I would however caution making a decision on only two variables as we have seen the interaction of the variables play a critical roll in the overall rating as well.

beer_style	review_aroma	review_appearance	review_overall
American Wild Ale	4.134354	4.010932	4.100018
Gueuze	4.116157	4.037312	4.087034
Quadrupel (Quad)	4.133515	4.119829	4.073141
Lambic - Unblended	4.126564	3.918191	4.060635

**** All code and markup can be seen in attached files