JUNE 2022

PREDICTING A HIGHLY RATED BEER

Springboard Capstone 2

Report by:
Lindsey Robertson
Data Science Fellow - Springboard
Data Science Career Track

BIG BUSINESS

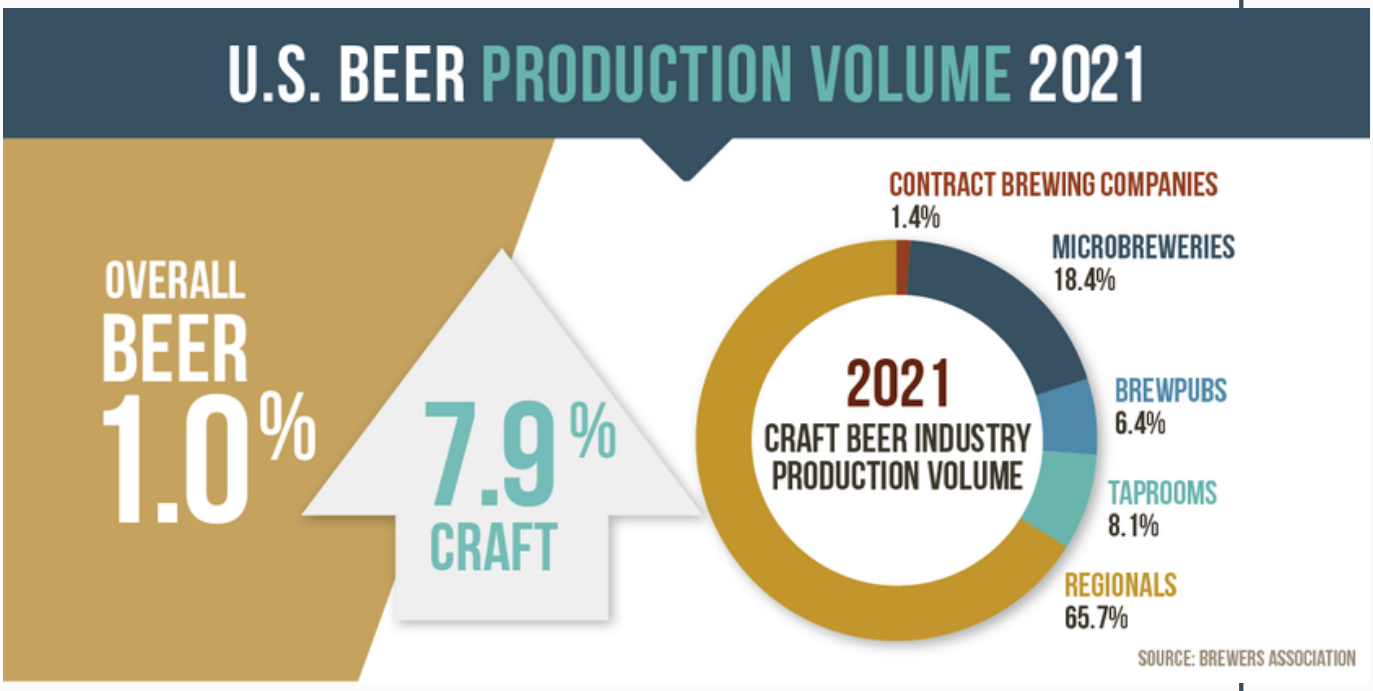
Crafting a winning brew can help tap into a booming market.

North America is expected to be a lucrative region, accumulating 36% market revenue.

Global craft beer market is expected to experience almost 2x growth from 2022 to 2032

GROWING COMPETITION

Craft beer production is up and there are over 4% more new breweries in the last year.

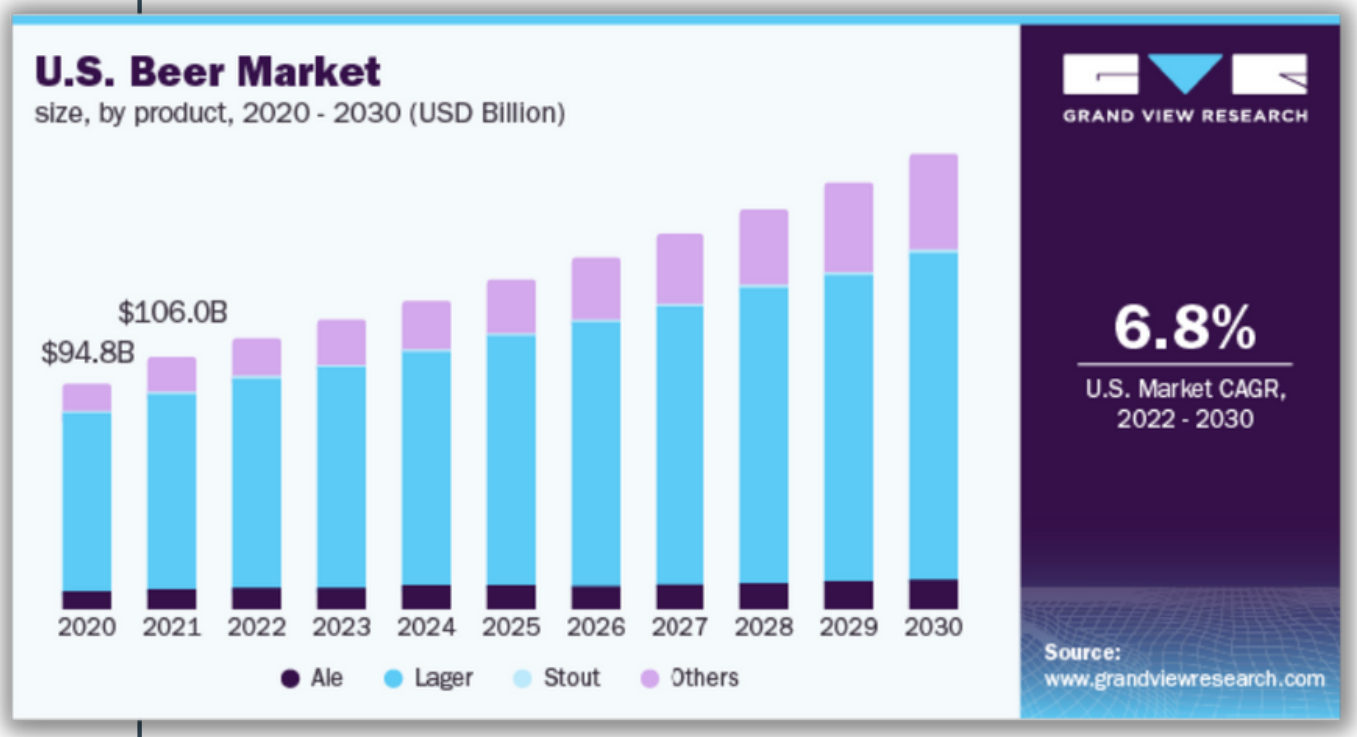


Recent U.S. Brewery Count

	2015	2016	2017	2018	2019	2020	2021	2020 to 2021 % Change
Craft	**4,803**	**5,713**	**6,661**	**7,618**	**8,419**	**8,905**	**9,118**	**4.4%**
Regional Craft Breweries	178	186	202	230	240	220	223	1.4%
Microbreweries	2,684	3,319	3,956	4,518	1,917	1,898	1,886	-0.6%
Taprooms					3,091	3,471	3,708	6.2%
Brewpubs	1,941	2,208	2,503	2,870	3,171	3,302	3,307	0.2%
Large/Non-Craft	44	67	106	104	111	120	129	7.5%
Total U.S. Breweries	**4,847**	**5,780**	**6,767**	**7,722**	**8,530**	**9,025**	**9,247**	**2.5%**

OPPORTUNITY ABOUNDS

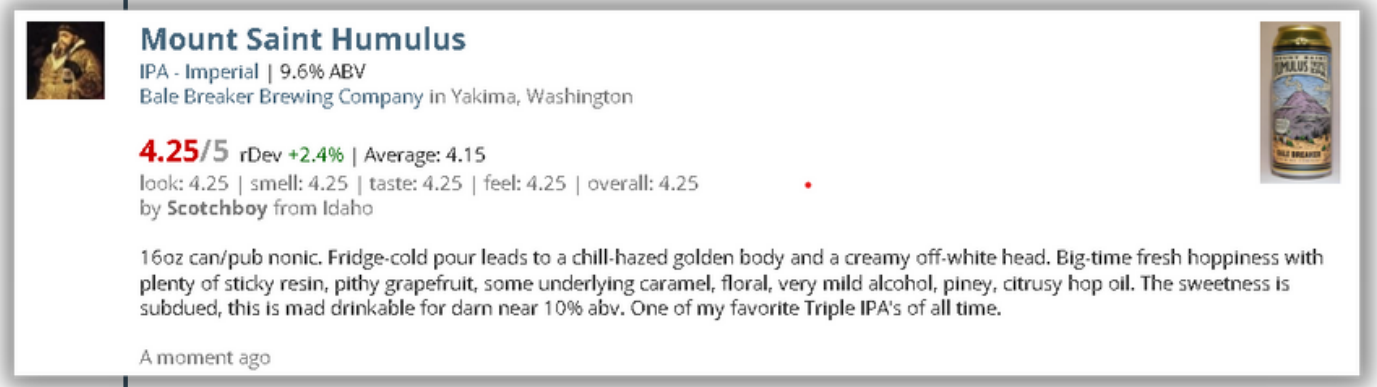
What do sales forecasts say about what do consumers want?



Pale Ale is said to account for 25% of total revenue.

ASK THE CONSUMER

Reviews reveal insight on what makes beer drinkers happy.



Mount Saint Humulus
IPA - Imperial | 9.6% ABV
Bale Breaker Brewing Company in Yakima, Washington

4.25/5 rDev +2.4% | Average: 4.15
look: 4.25 | smell: 4.25 | taste: 4.25 | feel: 4.25 | overall: 4.25
by **Scotchboy** from Idaho

16oz can/pub nonic. Fridge-cold pour leads to a chill-hazed golden body and a creamy off-white head. Big-time fresh hoppiness with plenty of sticky resin, pithy grapefruit, some underlying caramel, floral, very mild alcohol, piney, citrusy hop oil. The sweetness is subdued, this is mad drinkable for darn near 10% abv. One of my favorite Triple IPA's of all time.

A moment ago

Old Speckled Hen
Pale Ale - English | 5% ABV
Greene King / Morland Brewery in Suffolk, England

3.72/5 rDev +7.5% | Average: 3.46
look: 4.25 | smell: 4 | taste: 3.5 | feel: 3.5 | overall: 3.75
by **DrOfGolf** from Delaware

This one was poured from bottle into a beer mug. Dark amber look with a thick and foamy white head. Caramel, malt and hops upon opening. Overall, one I will definitely have again.

8 minutes ago

Pseudo Sue
Pale Ale - American | 5.8% ABV
Toppling Goliath Brewing Company in Decorah, Iowa

4.62/5 rDev +2.7% | Average: 4.5
look: 4.5 | smell: 4.25 | taste: 4.75 | feel: 4.75 | overall: 4.75
by **TheWaySheGoes** from Illinois

Not sure how this is a pale ale with its extremely hazy nature....that being said it's a damn fine beer. World class beer.

Today at 04:12 AM

THE GOAL

Predict a beer rating before it is brewed.

- Analyze aggregate reviews combined with matched tasting profiles to discover attributes most important to higher ratings on a 5-star scale.
- Model the data to predict the rating based on the tasting profile.
- Design a beer with a tasting profile that expects high ratings.
- Gain advantage over relying on opinions from qualified tasters.

PRODUCT POSITIONING

Gaining market share

Craft brew sales continue to climb, but represent a small percentage of the market and competition in the microbrew industry has never been more saturated.

Gaining consumer hype is paramount to marketing success and repeat customers. The desired clout could be boosted by a highly rated product.

A Brewmaster can **tap into data** from beer reviews to find a crowd-pleaser.

THE DATA

Kaggle data set from Beer Advocate reviews here.

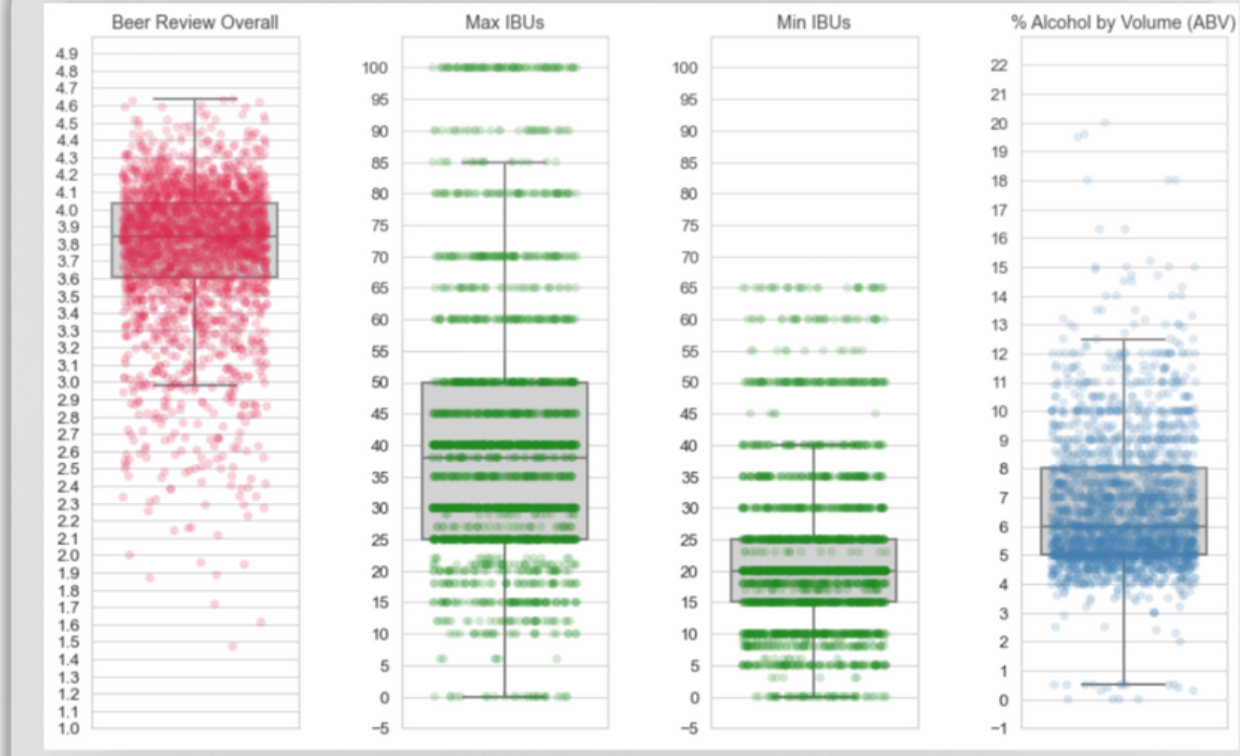
- **1.5 million consumer reviews**
- **3197 unique beers**
- **Analyze rows with 25 or more reviews**
- **934 unique breweries**
- **5 features for descriptions and labels**
- **5 review categories**
- **No duplicate rows**
- **14 beer profile features:**

Makeup	Mouth Feel	Taste	Aroma
ABV	Astringency	Bitter	Fruits
Min IBU	Alcohol	Sweet	Hoppy
Max IBU	Body	Sour	Spices
		Salty	Matly

DATA SUMMARY

	count	mean	std	min	25%	50%	75%	max
ABV	2344.0	6.660913	2.316489	0.000000	5.000000	6.000000	8.000000	28.000000
Min IBU	2344.0	22.329352	13.349367	0.000000	15.000000	20.000000	25.000000	65.000000
Max IBU	2344.0	40.924915	21.257609	0.000000	25.000000	38.000000	50.000000	100.000000
Astringency	2344.0	17.376280	10.157701	0.000000	10.000000	15.000000	22.000000	77.000000
Body	2344.0	48.620734	24.729193	0.000000	31.750000	42.000000	59.000000	175.000000
Alcohol	2344.0	18.309727	17.627883	0.000000	7.000000	12.000000	23.250000	126.000000
Bitter	2344.0	38.854949	25.421842	0.000000	19.000000	33.000000	54.000000	150.000000
Sweet	2344.0	62.349829	32.850397	0.000000	38.000000	57.000000	81.000000	219.000000
Sour	2344.0	35.386519	35.584863	0.000000	12.000000	25.000000	44.000000	241.000000
Salty	2344.0	1.058874	2.219869	0.000000	0.000000	0.000000	1.000000	48.000000
Fruits	2344.0	42.082338	32.107395	0.000000	15.000000	34.000000	64.000000	165.000000
Hoppy	2344.0	44.340870	30.395225	0.000000	21.000000	37.000000	60.000000	172.000000
Spices	2344.0	18.438567	22.408853	0.000000	5.000000	11.000000	24.000000	184.000000
Malty	2344.0	79.365188	37.818441	0.000000	50.000000	77.000000	105.000000	239.000000
review_aroma	2344.0	3.651109	0.495781	1.509615	3.452978	3.739078	3.978759	4.723770
review_appearance	2344.0	3.770150	0.394942	1.827916	3.634602	3.848388	4.026064	4.644231
review_palate	2344.0	3.671633	0.430545	1.682692	3.491868	3.748845	3.955754	4.633255
review_taste	2344.0	3.716694	0.489514	1.355769	3.516896	3.804073	4.042055	4.724590
review_overall	2344.0	3.764302	0.411153	1.471154	3.610140	3.840909	4.032853	4.634615
number_of_reviews	2344.0	314.584044	392.128404	26.000000	72.000000	167.000000	387.000000	3290.000000

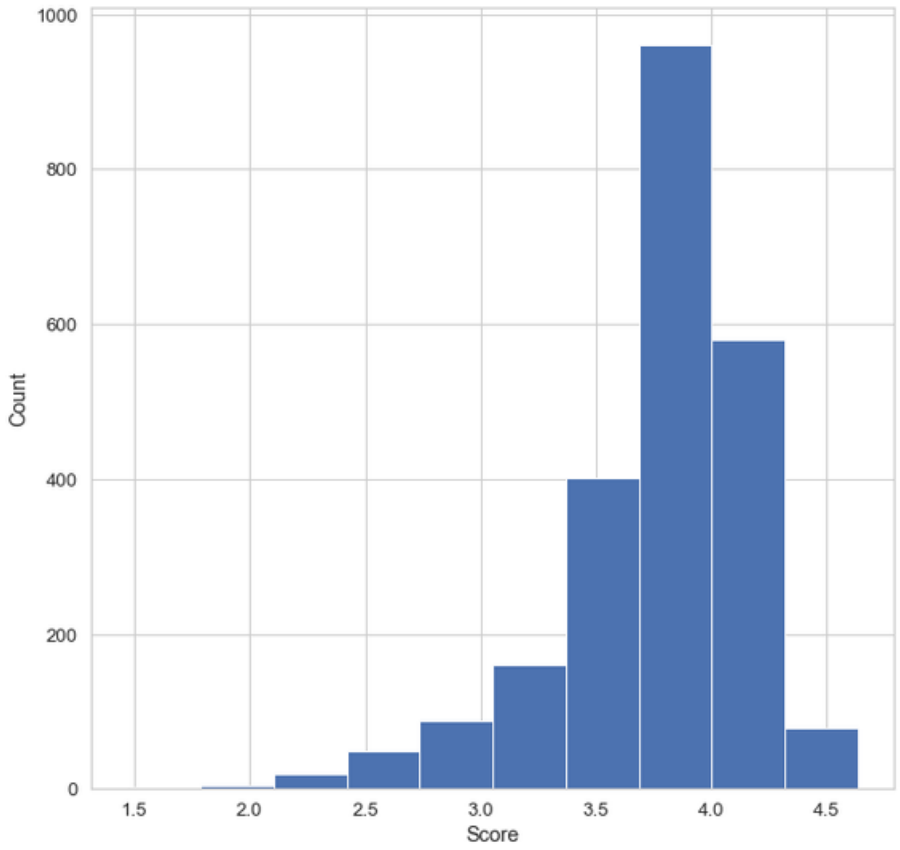
Heavy outlier action!



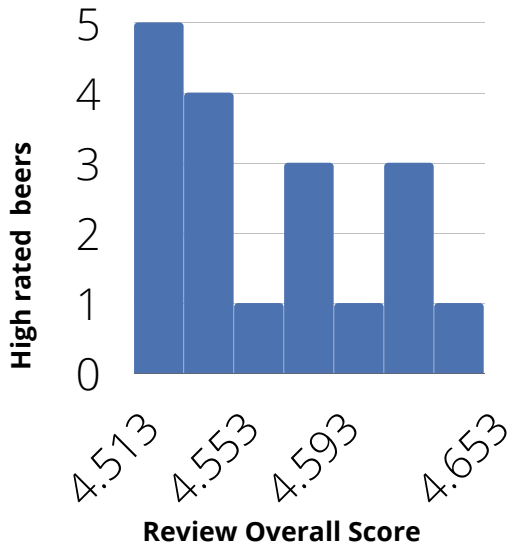
THE SCORES

Highest ratings are extremely hard to come by! Less than 1% make the cut.

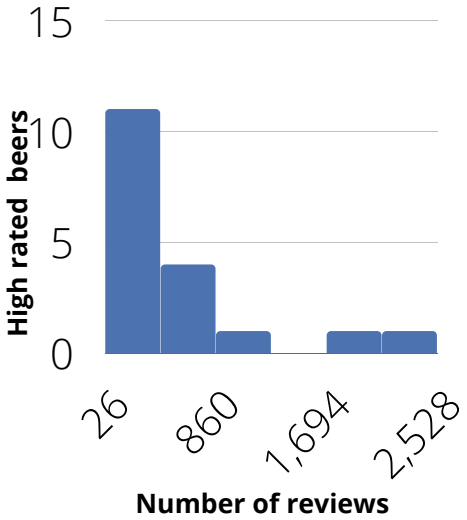
Review Overall Scores on 5 star scale



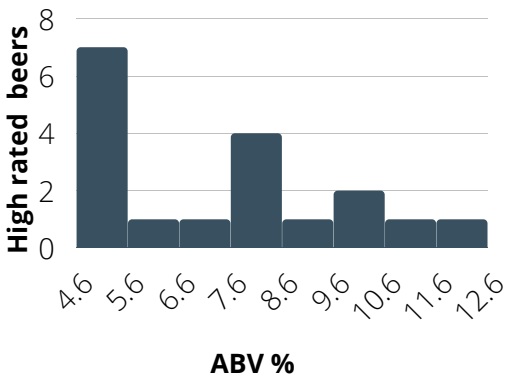
Beers with ratings greater than 4.5

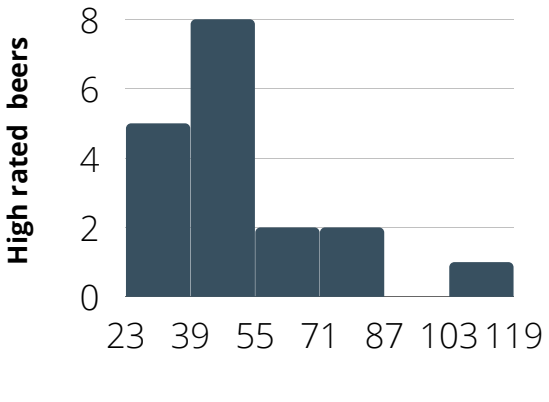


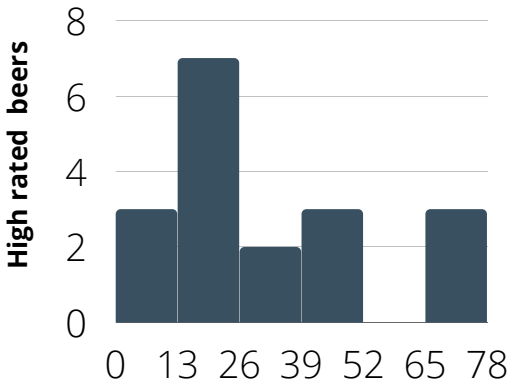
How many reviews did the highest-rated beers have?

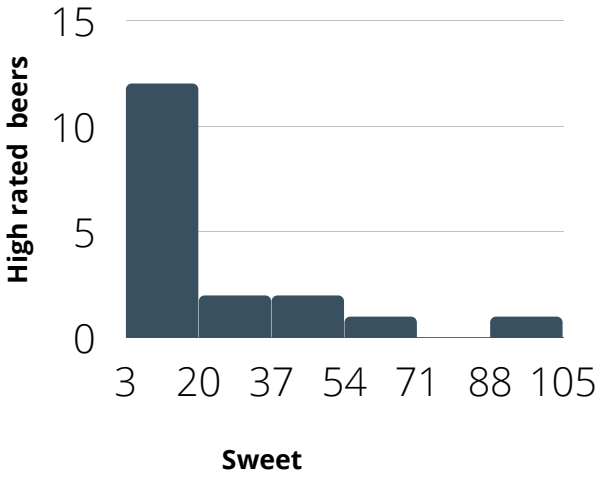


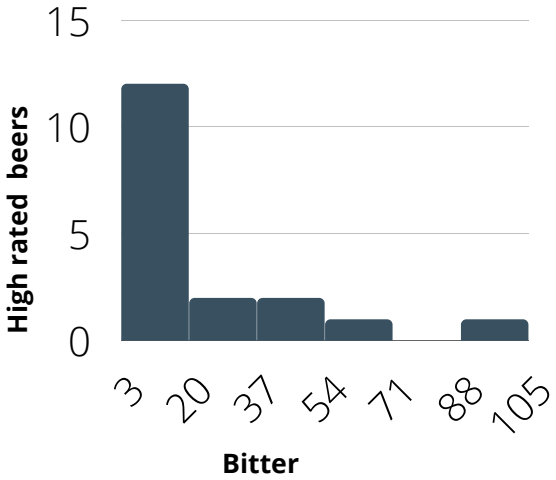
WHAT IS HIGHLY RATED?

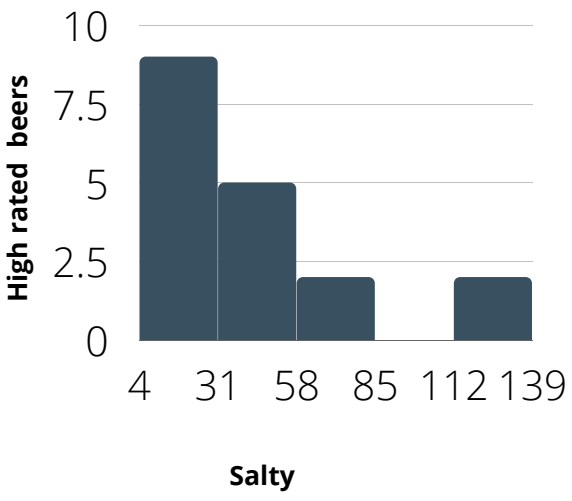


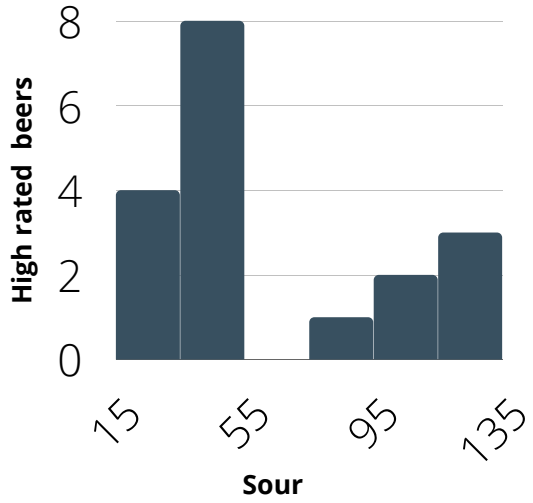












WHAT IS HIGHLY RATED?

Brauerei Zehendner GmbH Mönchsambacher Lager

4.64

26 reviews

Brasserie Cantillon Cantillon Blåbær Lambik

4.63

158 reviews

The Alchemist Heady Topper

4.63

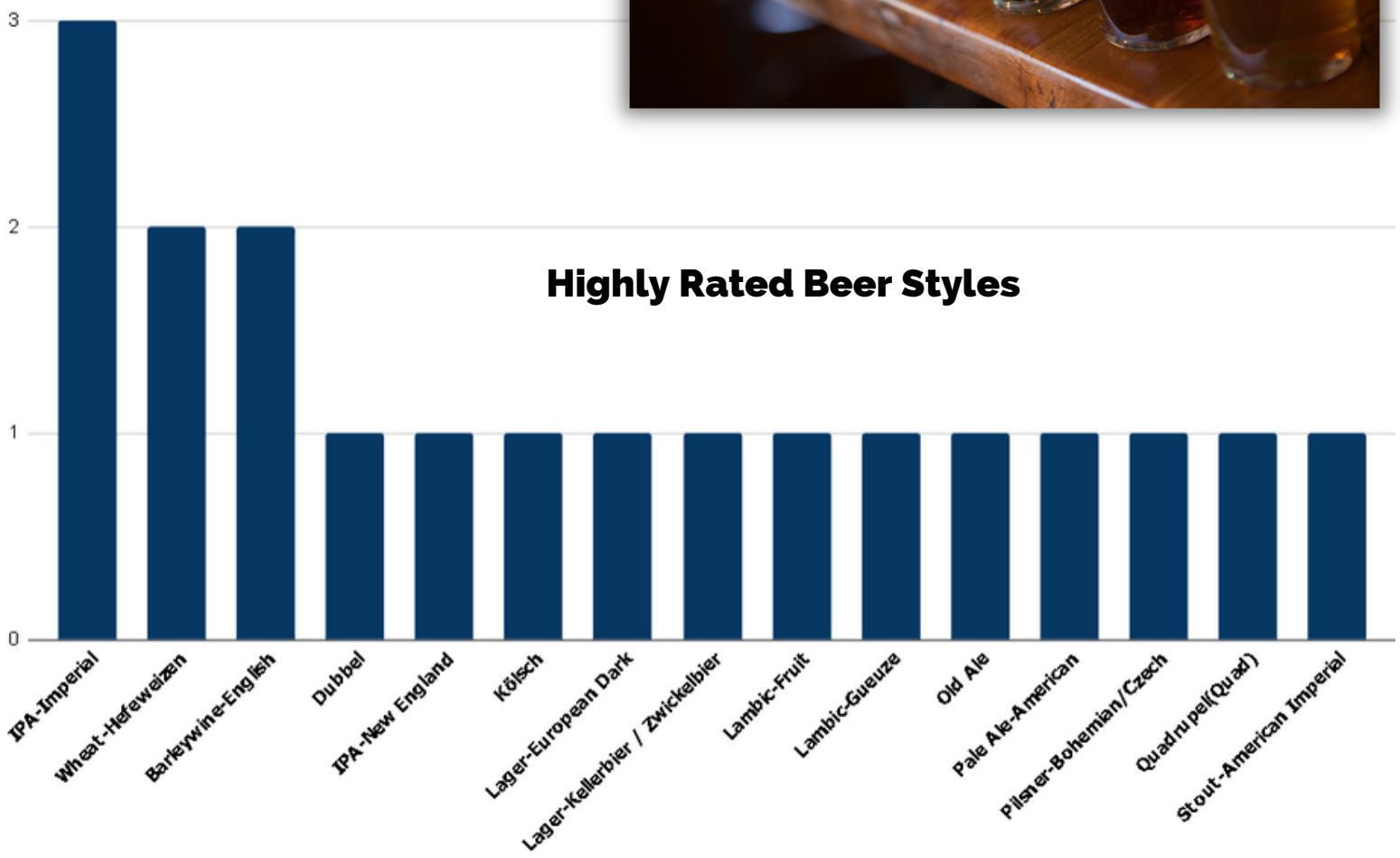
469 reviews

Brouwerij Westvleteren (Sint-Sixtusabdij

4.62

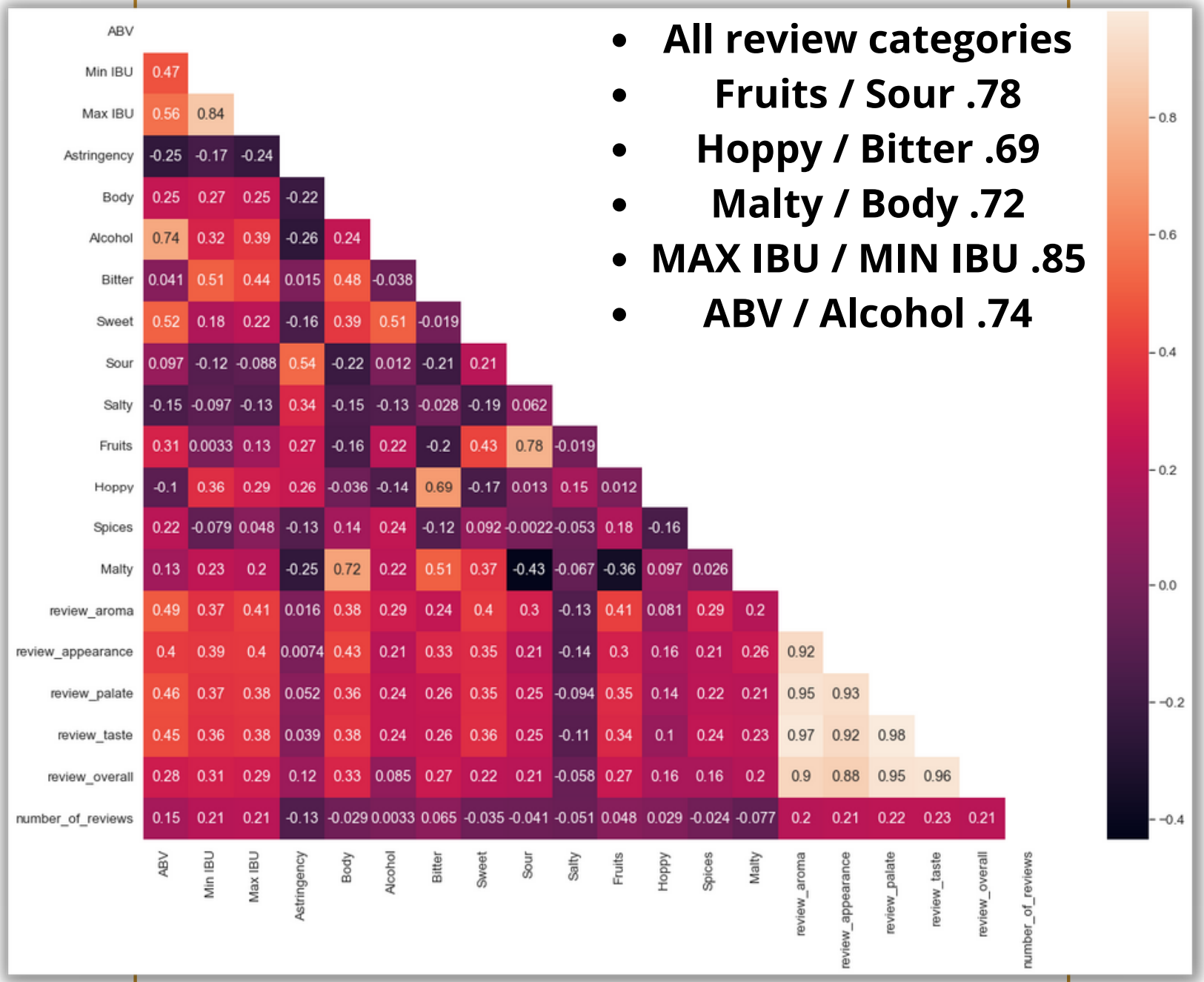
1272 reviews

WHAT IS HIGHLY RATED?



CORRELATIONS ARE INTUITIVE

No obvious correlations to overall ratings.

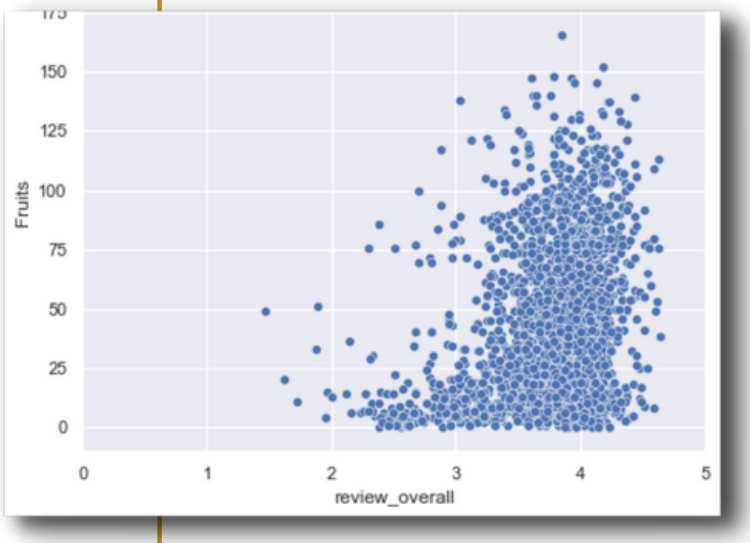


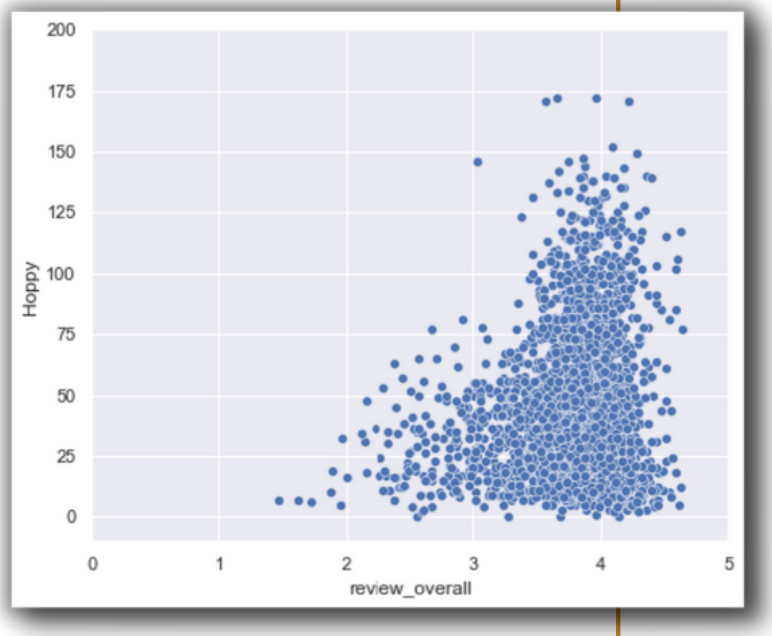
- All review categories
- Fruits / Sour .78
- Hoppy / Bitter .69
- Malty / Body .72
- MAX IBU / MIN IBU .85
- ABV / Alcohol .74

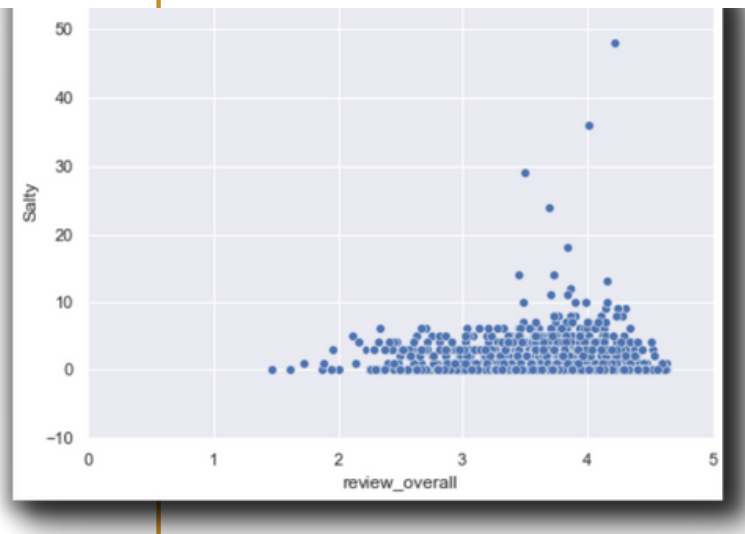
HEAT MAP

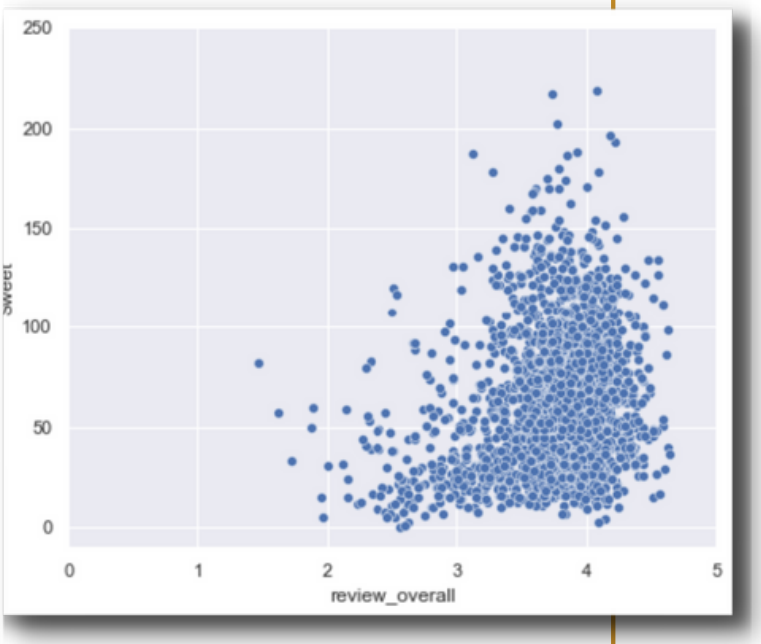
RELATIONSHIPS

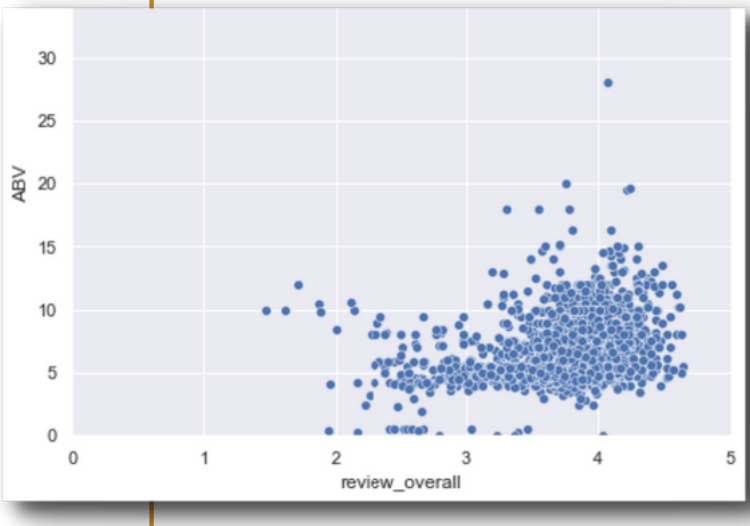
Interesting features plotted with review overall ratings.







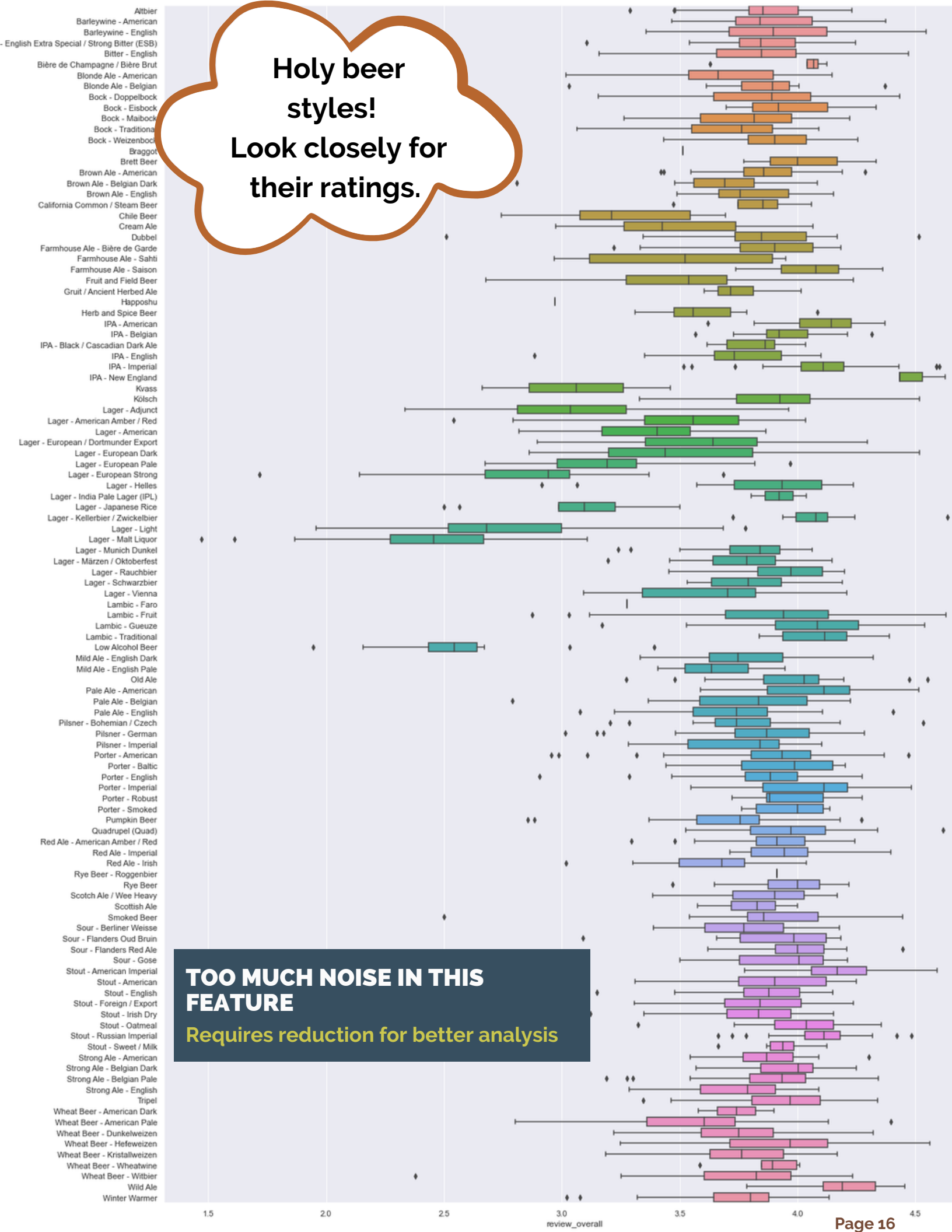




CONSTRAINTS

What are the issues with the data and analysis?

- Response bias of reviews
- Large number of styles
- Styles are categorical features
- No beer color feature
- No analysis of text descriptions


Holy beer styles!
Look closely for their ratings.
Altbier
Barleywine - American
Barleywine - English
English Extra Special / Strong Bitter (ESB)
Bitter - English
Bière de Champagne / Bière Brut
Blonde Ale - American
Blonde Ale - Belgian
Bock - Doppelbock
Bock - Eisbock
Bock - Maibock
Bock - Traditiona
Bock - Weizenboc
Braggot
Brett Beer
Brown Ale - American
Brown Ale - Belgian Dark
Brown Ale - English
California Common / Steam Beer
Chile Beer
Cream Ale
Dubbel
Farmhouse Ale - Bière de Garde
Farmhouse Ale - Sahti
Farmhouse Ale - Saison
Fruit and Field Beer
Gruit / Ancient Herbed Ale
Happoshu
Herb and Spice Beer
IPA - American
IPA - Belgian
IPA - Black / Cascadian Dark Ale
IPA - English
IPA - Imperial
IPA - New England
Kvass
Kölsch
Lager - Adjunct
Lager - American Amber / Red
Lager - American
Lager - European / Dortmunder Export
Lager - European Dark
Lager - European Pale
Lager - European Strong
Lager - Helles
Lager - India Pale Lager (IPL)
Lager - Japanese Rice
Lager - Kellerbier / Zwickelbier
Lager - Light
Lager - Malt Liquor
Lager - Munich Dunkel
Lager - Märzen / Oktoberfest
Lager - Rauchbier
Lager - Schwarzbier
Lager - Vienna
Lambic - Faro
Lambic - Fruit
Lambic - Gueuze
Lambic - Traditional
Low Alcohol Beer
Mild Ale - English Dark
Mild Ale - English Pale
Old Ale
Pale Ale - American
Pale Ale - Belgian
Pale Ale - English
Pilsner - Bohemian / Czech
Pilsner - German
Pilsner - Imperial
Porter - American
Porter - Baltic
Porter - English
Porter - Imperial
Porter - Robust
Porter - Smoked
Pumpkin Beer
Quadrupel (Quad)
Red Ale - American Amber / Red
Red Ale - Imperial
Red Ale - Irish
Rye Beer - Roggenbier
Rye Beer
Scotch Ale / Wee Heavy
Scottish Ale
Smoked Beer
Sour - Berliner Weisse
Sour - Flanders Oud Bruin
Sour - Flanders Red Ale
Sour - Gose
Stout - American Imperial
Stout - American
Stout - English
Stout - Foreign / Export
Stout - Irish Dry
Stout - Oatmeal
Stout - Russian Imperial
Stout - Sweet / Milk
Strong Ale - American
Strong Ale - Belgian Dark
Strong Ale - Belgian Pale
Strong Ale - English
Tripel
Wheat Beer - American Dark
Wheat Beer - American Pale
Wheat Beer - Dunkelweizen
Wheat Beer - Hefeweizen
Wheat Beer - Kristallweizen
Wheat Beer - Wheatwine
Wheat Beer - Witbier
Wild Ale
Winter Warmer
1.5
2.0
2.5
3.0
3.5
4.0
4.5
review_overall
TOO MUCH NOISE IN THIS FEATURE
Requires reduction for better analysis

BEER STYLE CLUSTERING

How do we reduce clusters into meaningful categories?

- With knowledge/opinion
- Industry standards - still too many
- Dropping data - always avoid
- ~~Clustering~~

HIERARCHICAL CLUSTERING

Grouping similar objects

Normalize with whiten()

Handles covariance and variety of ranges in data for the algorithm

Verify clean data has no missing values.

scaled: Max IBU,Min IBU, Alcohol, ABV, Body, Bitter, Sweet, Sour, Fruits, Hoppy, Spices, Malty

Link features for similarity. Ward method to limit variance. Euclidean Distance for the shortest path/ minimize squared differences in observed and estimated values.

Try 6 -10 in an effort to evaluate for most intuitive results, keep cardinality of styles feature low.

Settled on 7

Merge data into the number of clusters desired.

Create new column in dataframe for cluster labels

CLUSTERING RESULTS

what insights do the new clusters reveal?

beer cluster labels counts:

1	338
2	126
3	560
4	264
5	187
6	316
7	553

Cluster 1 - Highest sours & fruits

Cluster 2 - Hihgest malts & sweets

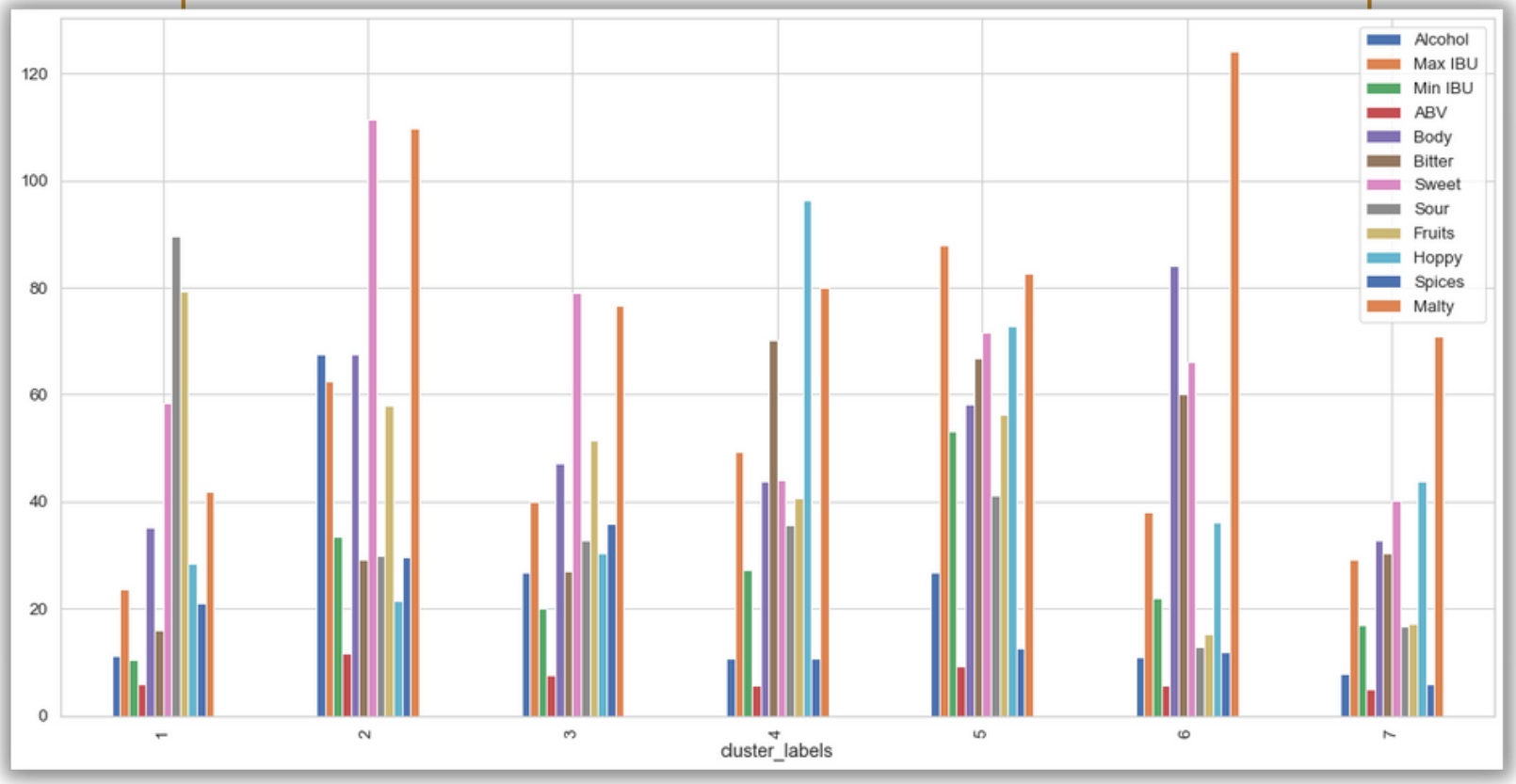
Cluster 3 - Moderate properties, Heavy_moderate, Seasonal

Cluster 4 - Highest hops &bitterness ,max IBU , Heavy

Cluster 5 = Highest malts & body, lowest sour, Dark

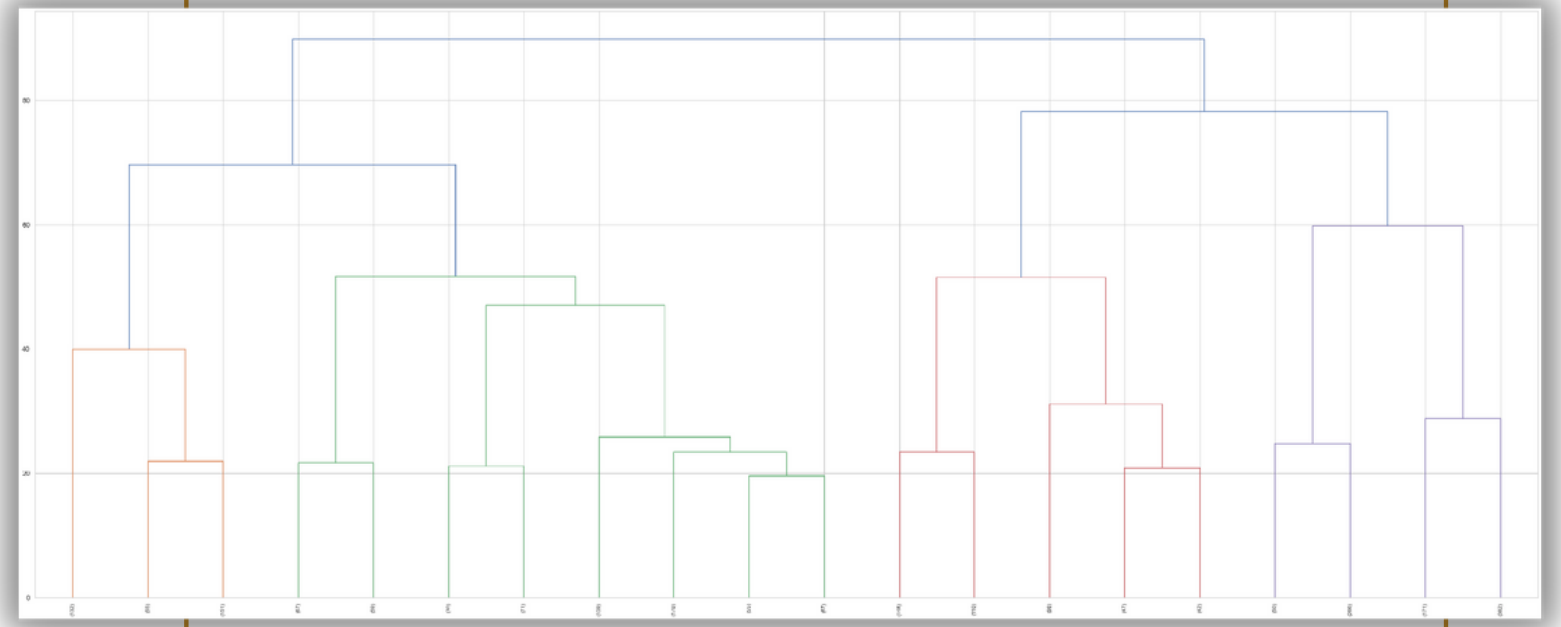
Cluster 6 = Lowest ABV & spices, Light

Cluster 7 = Lowest Alcohol



CLUSTERING RESULTS

What styles are in the clusters?



- Sour / Fruits
- Sweet / Malts
- Moderate / Seasonal
- Ale
- Heavy
- Dark
- Light

BASELINE MODEL

What is best guess performance?

Model	R ^ 2	Mean Absolute Error	Constant (mean predictions)
Dummy Regressor	**0.0017**	**0.29**	**3.76**

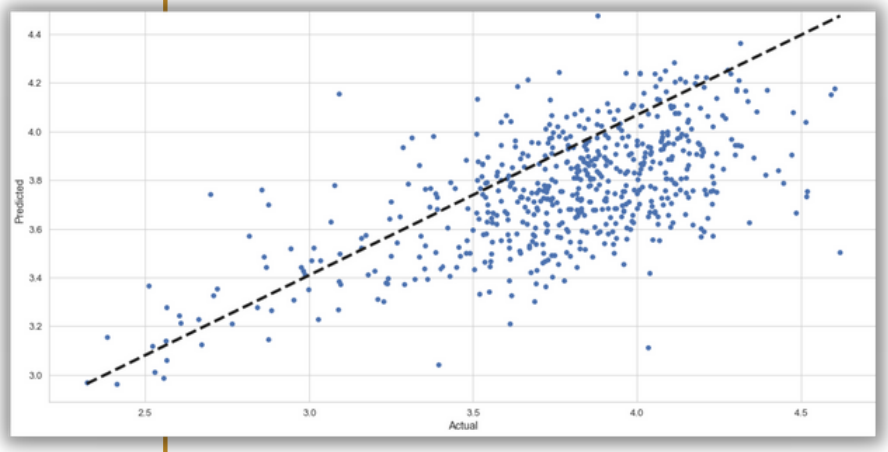
LINEAR REGRESSION MODEL

Supervised learning predictions

Model	R ^ 2	Mean Absolute Error
Linear Regression	**0.39**	**0.23**

OLS Regression Results

Dep. Variable:	review_overall	**R-squared:**	0.369
Model:	OLS	**Adj. R-squared:**	0.362
Method:	Least Squares	**F-statistic:**	50.84
Date:	Fri, 10 Jun 2022	**Prob (F-statistic):**	1.64e-157
Time:	09:31:57	**Log-Likelihood:**	-562.87
No. Observations:	1758	**AIC:**	1168.
Df Residuals:	1737	**BIC:**	1283.
Df Model:	20		
Covariance Type:	nonrobust		



Feature Importance

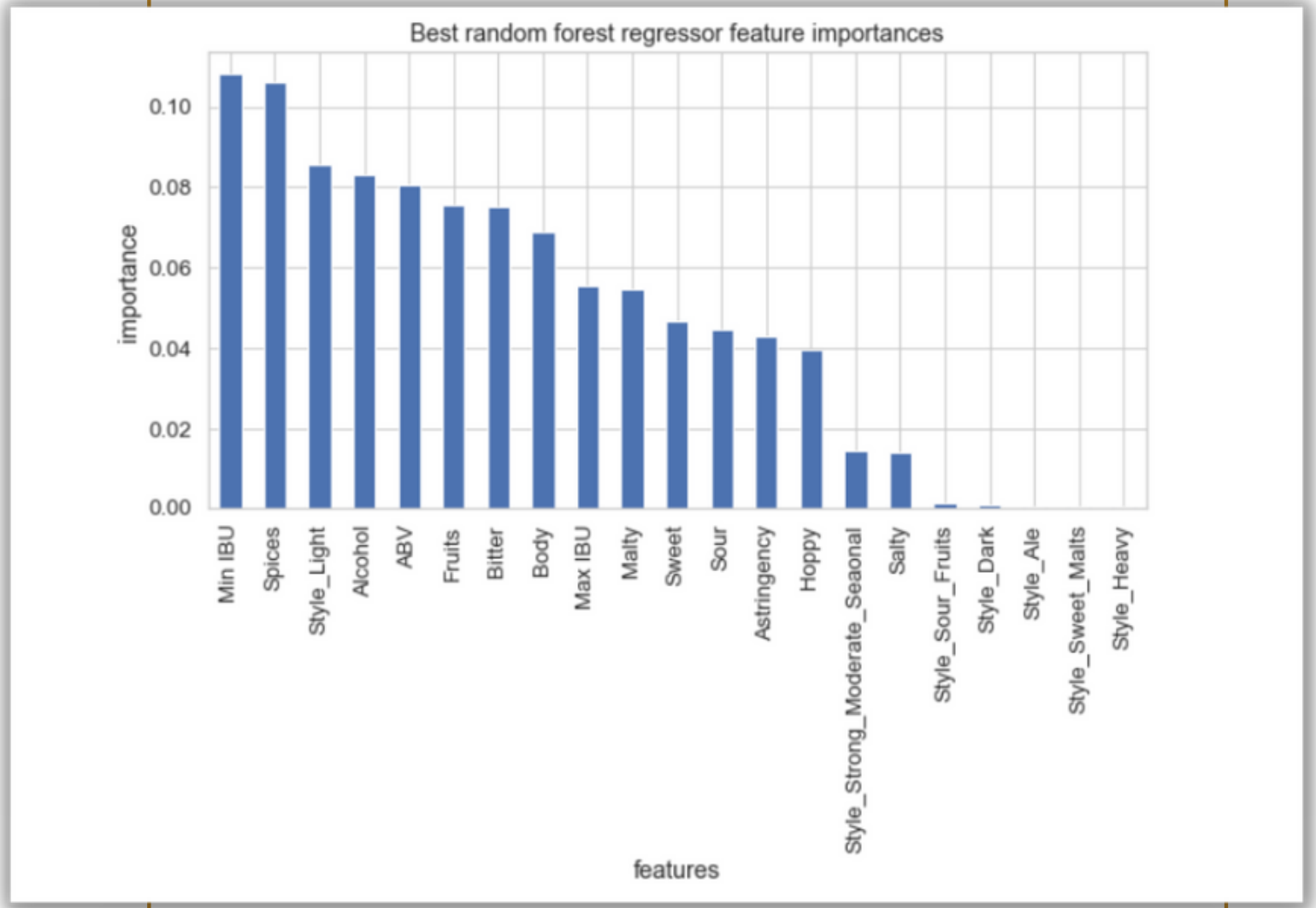
Min IBU	0.145426
ABV	0.115302
Malty	0.095666
Fruits	0.087332
Body	0.082815
Spices	0.059660
Astringency	0.046897
Style_Sweet_Malts	0.045952
Hoppy	0.045648
Sour	0.044658
Style_Sour_Fruits	0.021871
Style_Dark	0.021741
Style_Ale	0.012179
Style_Strong_Moderate_Seaonal	-0.006143
Max IBU	-0.019094
Salty	-0.026480
Style_Light	-0.035986
Style_Heavy	-0.041496
Bitter	-0.049412
Sweet	-0.049517
Alcohol	-0.158543

RANDOM FOREST MODEL

Decision tree bootstrap, supervised learning.

Baseline and Linear Regression models left much room for improvement

Model	R ^ 2	Mean Absolute Error
Random Forest	0.54	0.20



Data was scaled and GridSearch Cross Validation performed on all models.

MODEL COMPARISON

Pycaret for best model selection

We have a winner . . . sort of!

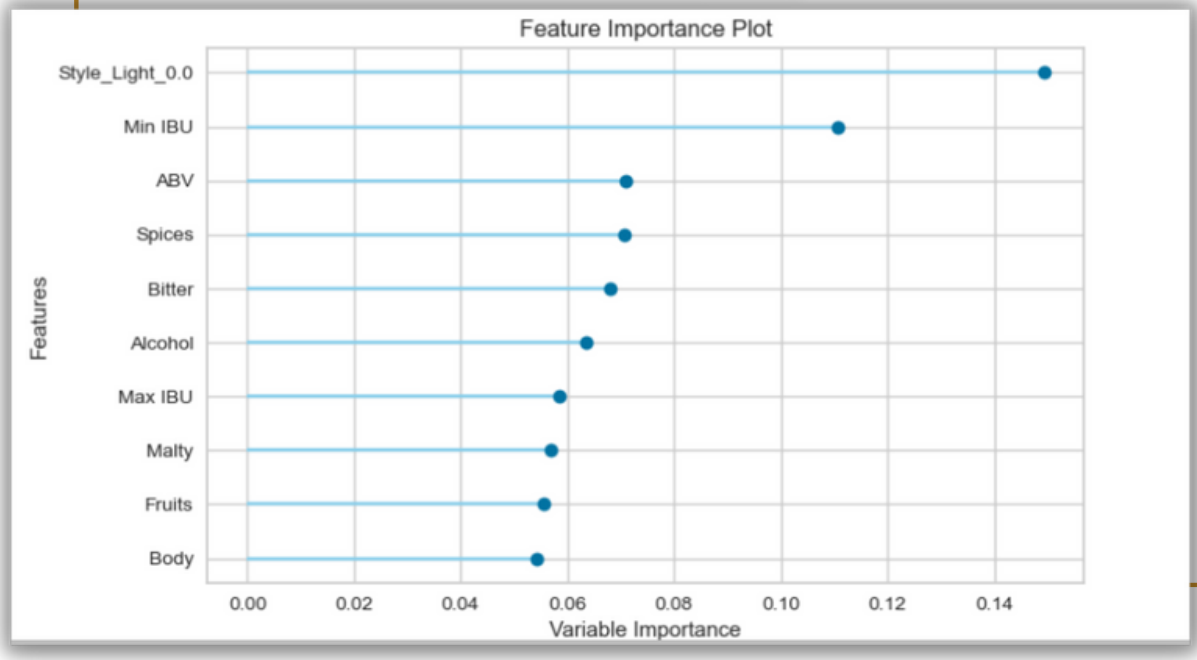
	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
et	Extra Trees Regressor	0.2179	0.0828	0.2872	0.5315	0.0646	0.0618	0.1530
rf	Random Forest Regressor	0.2179	0.0854	0.2915	0.5180	0.0660	0.0622	0.2100
gbr	Gradient Boosting Regressor	0.2176	0.0856	0.2918	0.5168	0.0659	0.0619	0.1150
lightgbm	Light Gradient Boosting Machine	0.2232	0.0880	0.2961	0.5023	0.0666	0.0632	0.1620
knn	K Neighbors Regressor	0.2328	0.0973	0.3113	0.4525	0.0706	0.0667	0.0200
ada	AdaBoost Regressor	0.2438	0.0996	0.3153	0.4365	0.0707	0.0686	0.0710
ridge	Ridge Regression	0.2518	0.1121	0.3342	0.3707	0.0757	0.0723	0.0180
lr	Linear Regression	0.2518	0.1121	0.3342	0.3706	0.0757	0.0723	1.0490
br	Bayesian Ridge	0.2554	0.1142	0.3374	0.3587	0.0767	0.0735	0.0170
en	Elastic Net	0.2666	0.1283	0.3574	0.2834	0.0821	0.0778	0.0060
lasso	Lasso Regression	0.2696	0.1346	0.3661	0.2487	0.0843	0.0792	0.0230
omp	Orthogonal Matching Pursuit	0.2788	0.1438	0.3782	0.1963	0.0868	0.0816	0.0060
lar	Least Angle Regression	0.2950	0.1489	0.3835	0.1558	0.0849	0.0828	0.0150
dt	Decision Tree Regressor	0.2951	0.1589	0.3973	0.1006	0.0902	0.0832	0.0100
llar	Lasso Least Angle Regression	0.3079	0.1798	0.4232	-0.0033	0.0968	0.0910	0.0060
dummy	Dummy Regressor	0.3079	0.1798	0.4232	-0.0033	0.0968	0.0910	0.0070
huber	Huber Regressor	0.3287	0.1891	0.4321	-0.0666	0.0943	0.0906	0.0230
par	Passive Aggressive Regressor	0.4483	0.3396	0.5714	-0.9038	0.1220	0.1214	0.0220

EXTRA TRESS REGRESSOR

Decision tree ensemble, supervised learning

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Extra Trees Regressor	0.2076	0.0776	0.2785	0.5224	0.0614	0.0575

review_overall	Label
3.047059	3.127324
4.354516	4.069642
4.048683	3.860426
3.461538	2.958766
3.781250	3.941975
...	...
4.045872	3.767765
1.612245	2.507699
4.072780	3.834366
3.565657	4.011121
4.025000	3.909742



CONCLUSIONS

Can we predict the beer rating before it is brewed?

- Feature importance revealed light beers with lower IBUs and alcohol tend to be the largest tasting indicators of a rating.

- Beer styles may be slightly arbitrary and not always strongly related to the tasting profile. Perhaps the color factor plays a part, or maybe the wishes of the brewer overrule in labeling a beer with a style. All of which the data was missing.

- High accuracy in this type of prediction will be tough to achieve with the data and type of analysis conducted.

- Given a blind taste test, what would a human's guessing accuracy be? Less than .5 R^2 I am willing to bet!

FUTURE CONSIDERATIONS

- Explore the text descriptions to do sentiment analysis and find text indicators of good ratings.
- Do a marketing analysis of how brands, beer names, and descriptions impact ratings.
- Reframe the problem to predict a good vs. bad beer profile with a rating threshold.