BeerGression

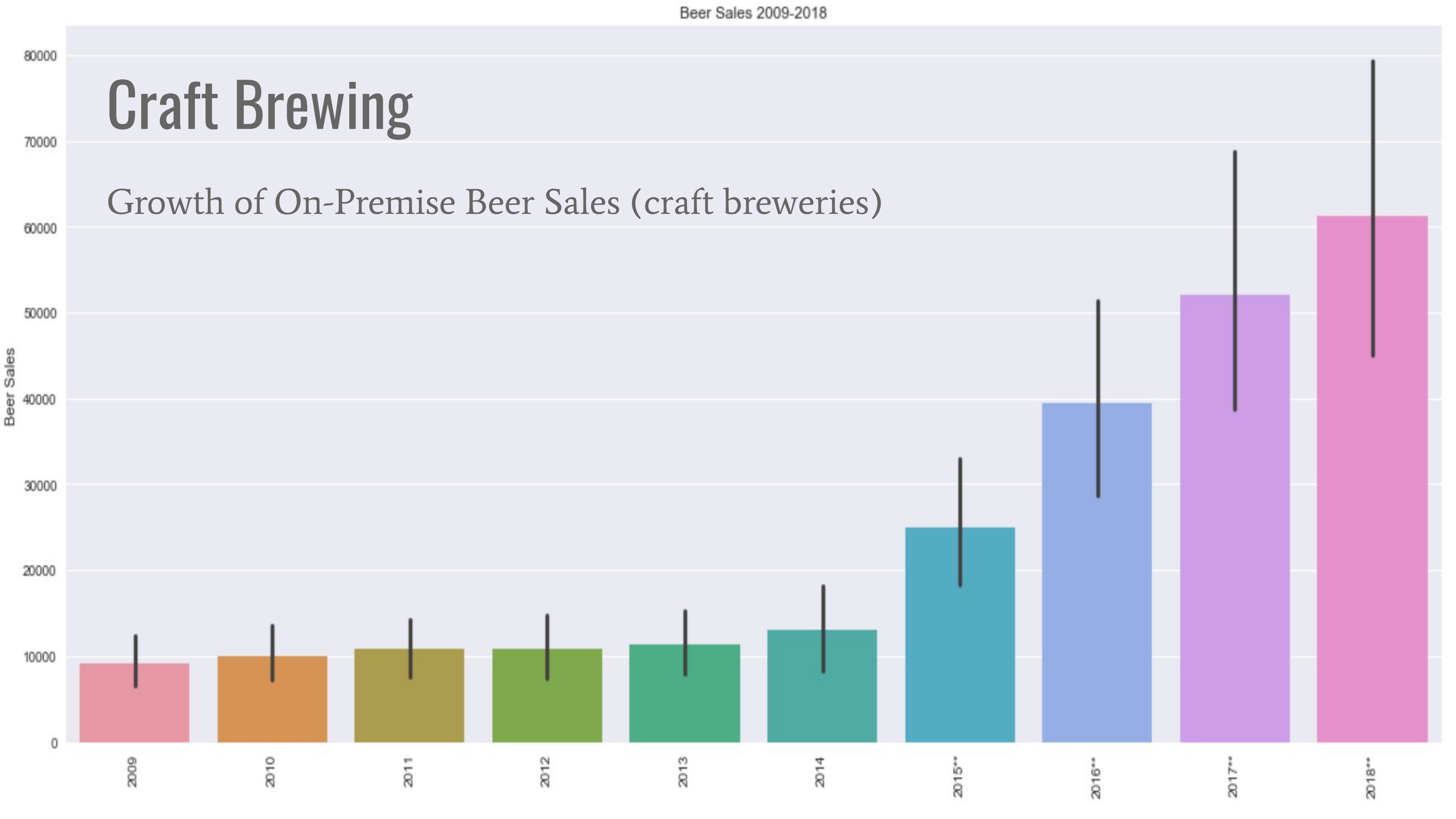
Sarah Smith & Matt Oliver

Big Data Energy

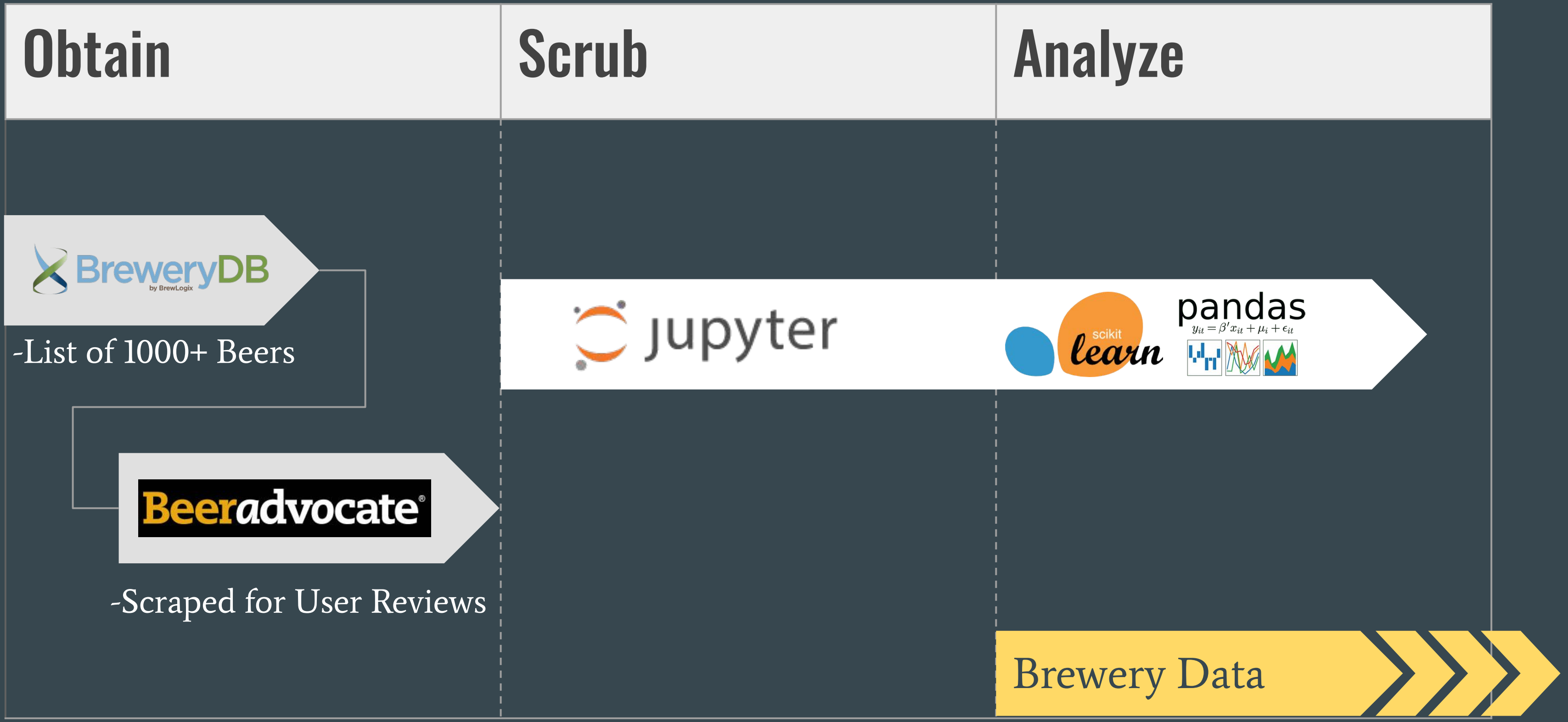
Image:
Allagash Brewing

Craft Brewing

Growth of On-Premise Beer Sales (craft breweries)



Data Process



Getting to know our data:

Outcome:

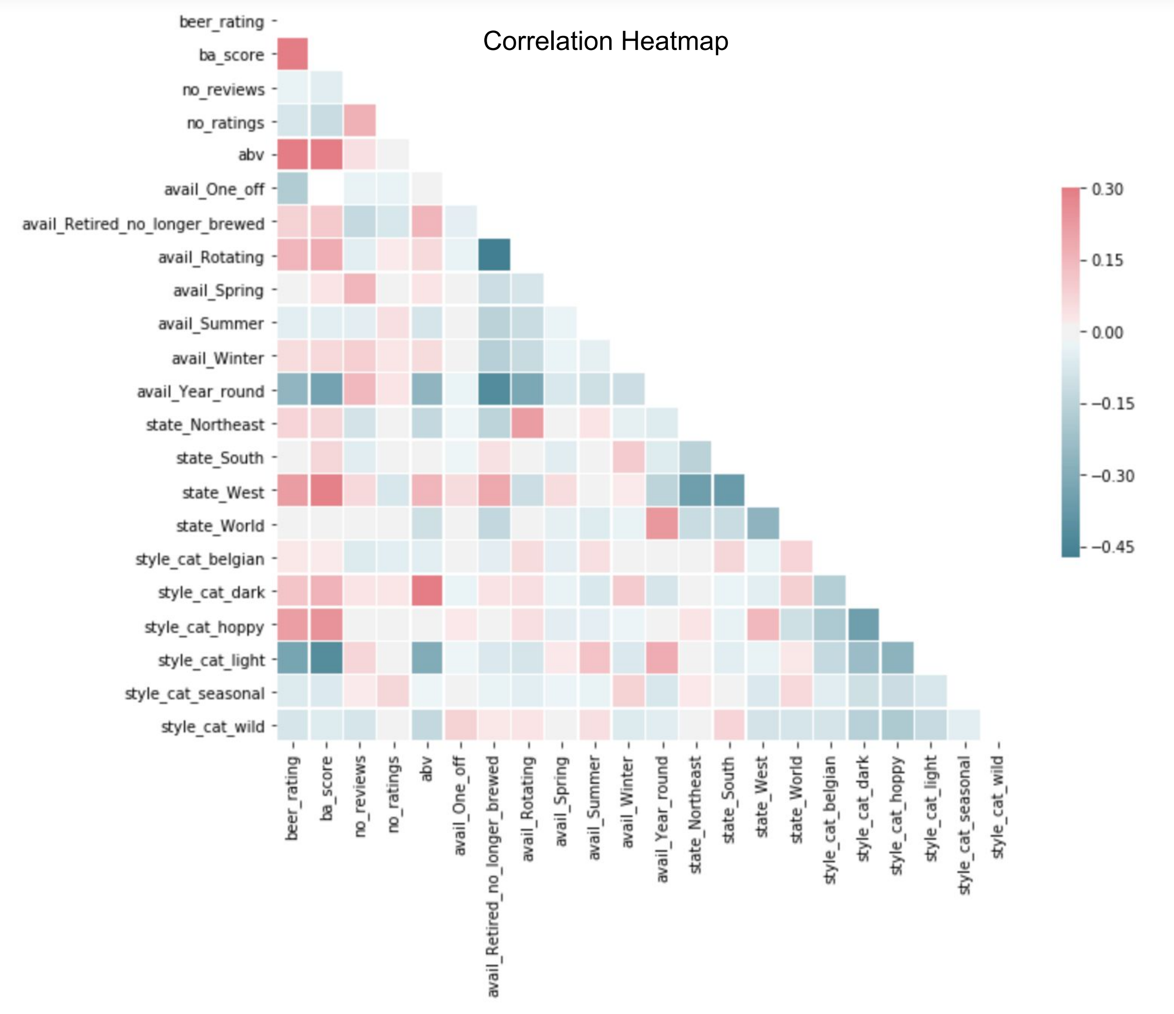
BA Score - int out of 100

Potential Predictors:

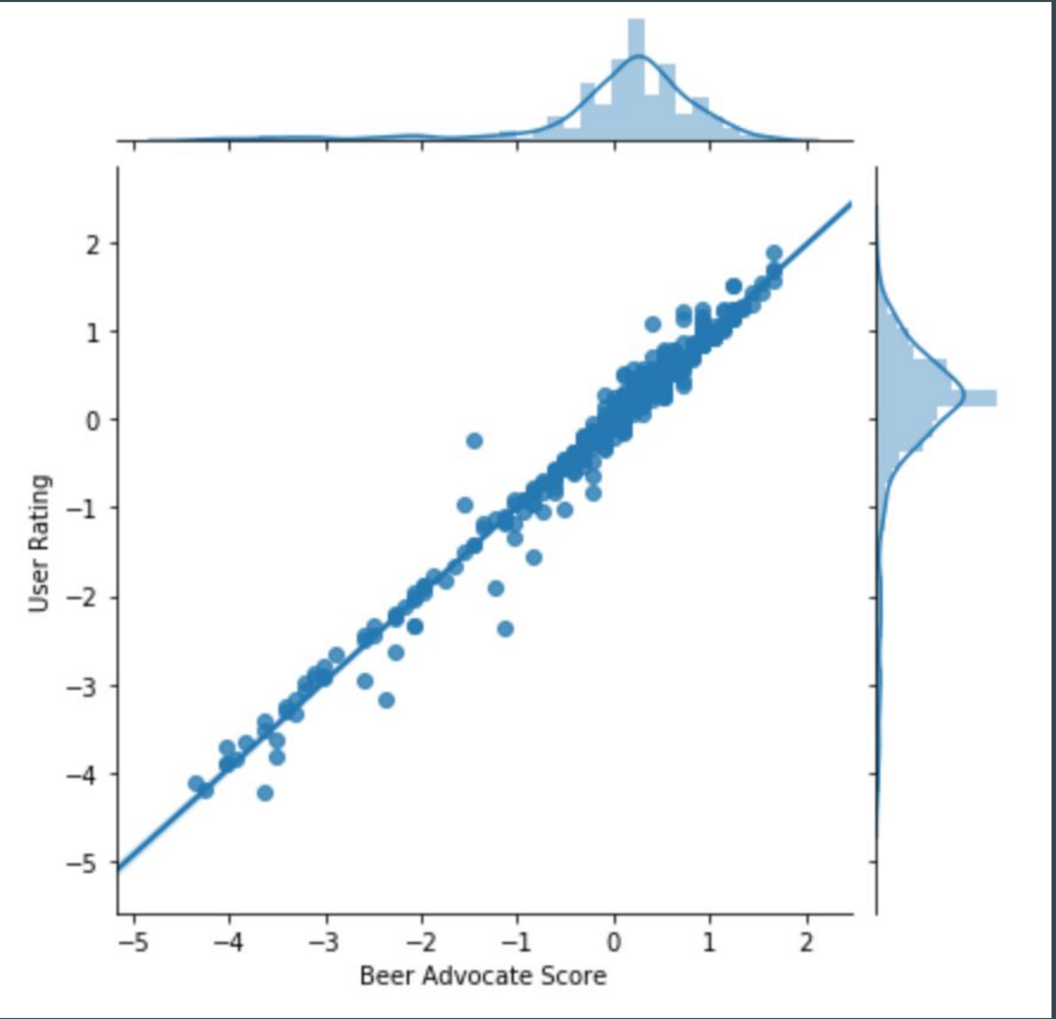
- User Rating
- ABV
- Number of Reviews
- Availability
- Style
- Brewery Location

Data Scrubbing:

- Drop
  - Non-scored*
- Categorize
  - States (Region
  - Beer Style



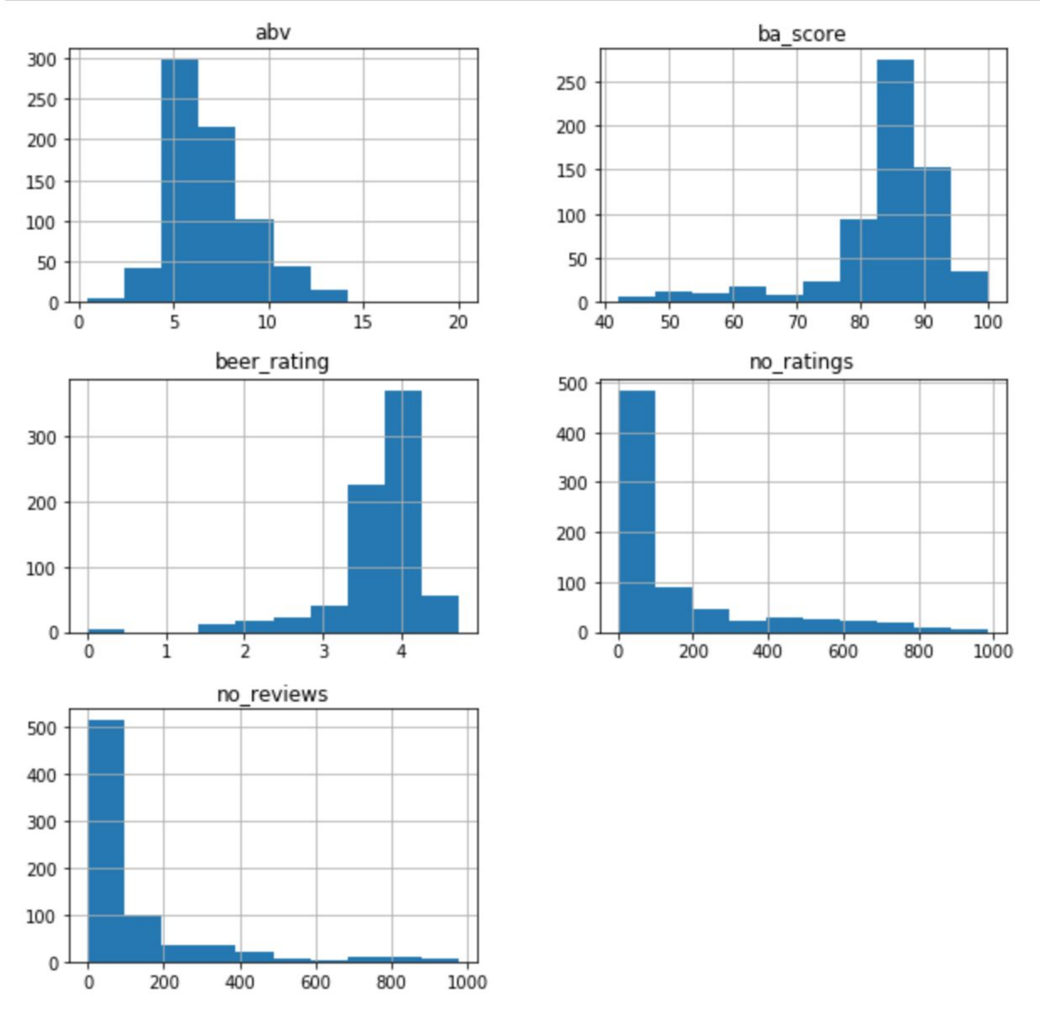
Now that our data is cleaned and categorical data has been binned we took a look at the heatmap for to check for potential correlation



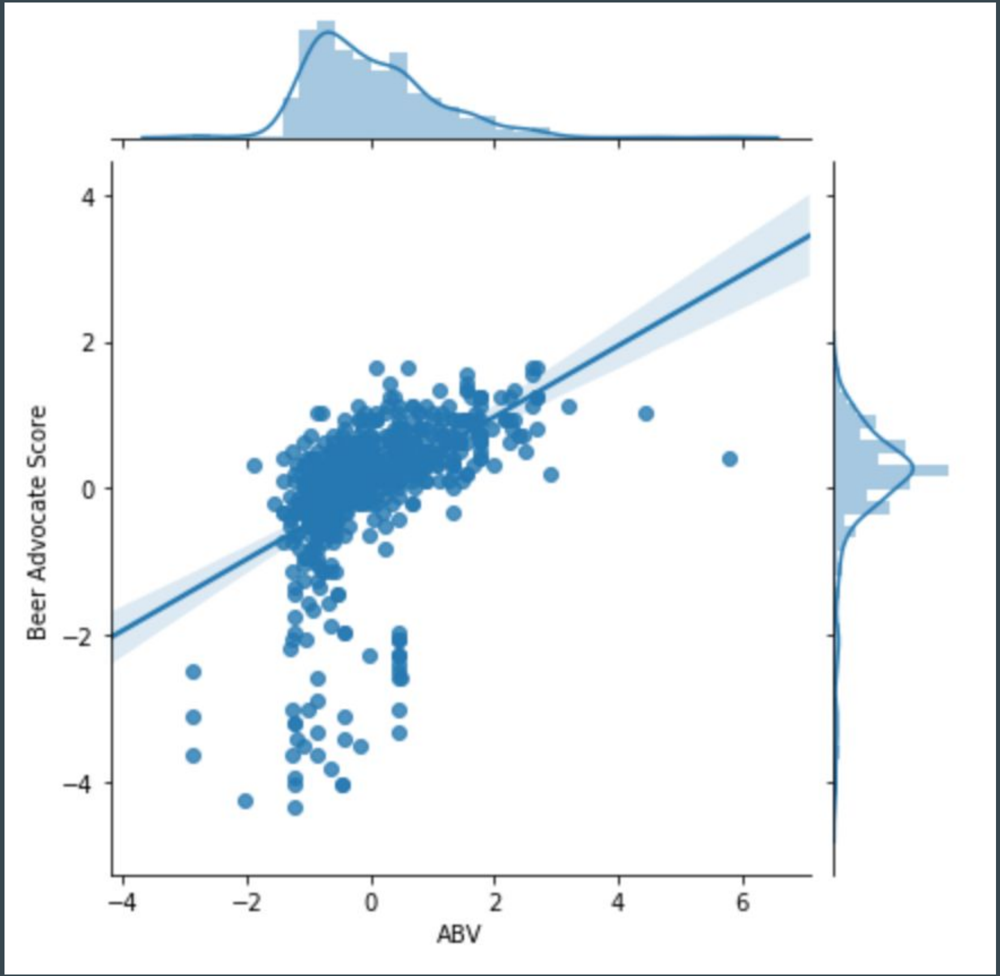
R^2=.978

No Bias in Beer Advocate Score?

Obviously using a predictive model for their scoring system



Next we looked at the distribution and values of our data.
Log transformation did not make them more normal - chose to proceed as is with standardization of the data.



A check for linearity between ABV & Score in our model.

Time for Regression - first look:

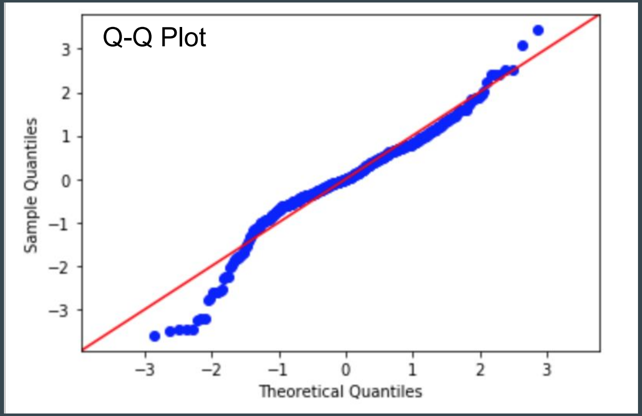
Dep. Variable:	ba_score	**R-squared:**	0.537
Model:	OLS	**Adj. R-squared:**	0.518
Method:	Least Squares	**F-statistic:**	28.63
Date:	Thu, 17 Oct 2019	**Prob (F-statistic):**	6.37e-63
Time:	15:09:34	**Log-Likelihood:**	-471.30
No. Observations:	463	**AIC:**	980.6
Df Residuals:	444	**BIC:**	1059.
Df Model:	18		
Covariance Type:	nonrobust		

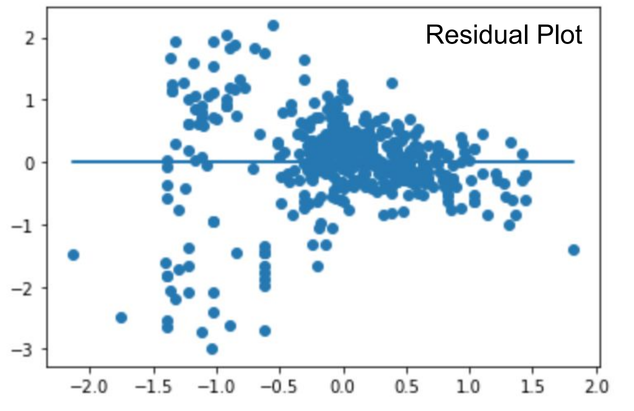
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-0.3233	0.218	-1.480	0.140	-0.753	0.106
no_ratings	-0.0994	0.032	-3.103	0.002	-0.162	-0.036
abv	0.3520	0.038	9.174	0.000	0.277	0.427
avail_Retired_no_longer_brewed	-0.4189	0.206	-2.029	0.043	-0.825	-0.013
avail_Rotating	-0.1744	0.210	-0.832	0.406	-0.586	0.238
avail_Spring	-0.4048	0.327	-1.239	0.216	-1.047	0.237
avail_Summer	-0.2777	0.266	-1.045	0.296	-0.800	0.244
avail_Winter	-0.3071	0.249	-1.234	0.218	-0.796	0.182

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
avail_Year_round	-0.6558	0.214	-3.063	0.002	-1.077	-0.235
state_Northeast	0.9493	0.121	7.844	0.000	0.711	1.187
state_South	0.8923	0.111	8.048	0.000	0.674	1.110
state_West	0.9131	0.086	10.588	0.000	0.744	1.083
state_World	0.9243	0.135	6.857	0.000	0.659	1.189
style_cat_belgian	-0.0155	0.152	-0.102	0.919	-0.313	0.282
style_cat_dark	0.0442	0.116	0.381	0.704	-0.184	0.272
style_cat_hoppy	0.2996	0.109	2.741	0.006	0.085	0.514
style_cat_light	-0.5320	0.123	-4.324	0.000	-0.774	-0.290
style_cat_seasonal	-0.3817	0.199	-1.916	0.056	-0.773	0.010
style_cat_wild	-0.0881	0.144	-0.614	0.540	-0.370	0.194

Omnibus:	52.923	**Durbin-Watson:**	2.141
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	115.922
Skew:	-0.624	**Prob(JB):**	6.73e-26
Kurtosis:	5.110	**Cond. No.**	22.7

Final Chosen Model:





Dep. Variable:	ba_score	**R-squared:**	0.425
Model:	OLS	**Adj. R-squared:**	0.419
Method:	Least Squares	**F-statistic:**	67.63
Date:	Thu, 17 Oct 2019	**Prob (F-statistic):**	8.02e-53
Time:	15:10:22	**Log-Likelihood:**	-521.46
No. Observations:	463	**AIC:**	1055.
Df Residuals:	457	**BIC:**	1080.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.8231	0.074	-11.118	0.000	-0.969	-0.678
abv	0.4578	0.035	12.902	0.000	0.388	0.528
state_Northeast	1.2109	0.128	9.447	0.000	0.959	1.463
state_South	1.0065	0.118	8.537	0.000	0.775	1.238
state_West	1.0406	0.091	11.412	0.000	0.861	1.220
state_World	0.9226	0.144	6.394	0.000	0.639	1.206

Omnibus:	76.953	**Durbin-Watson:**	2.150
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	197.414
Skew:	-0.826	**Prob(JB):**	1.36e-43
Kurtosis:	5.740	**Cond. No.**	6.13

Homoscedasticity**

R^2=.425

Data-Qualitative, not causative

Train RMSE: 0.727

Test RMSE: 0.804

Unpenalized Training Error: 80
Ridge Model Training Error: 80
Lasso Model Training Error: 116

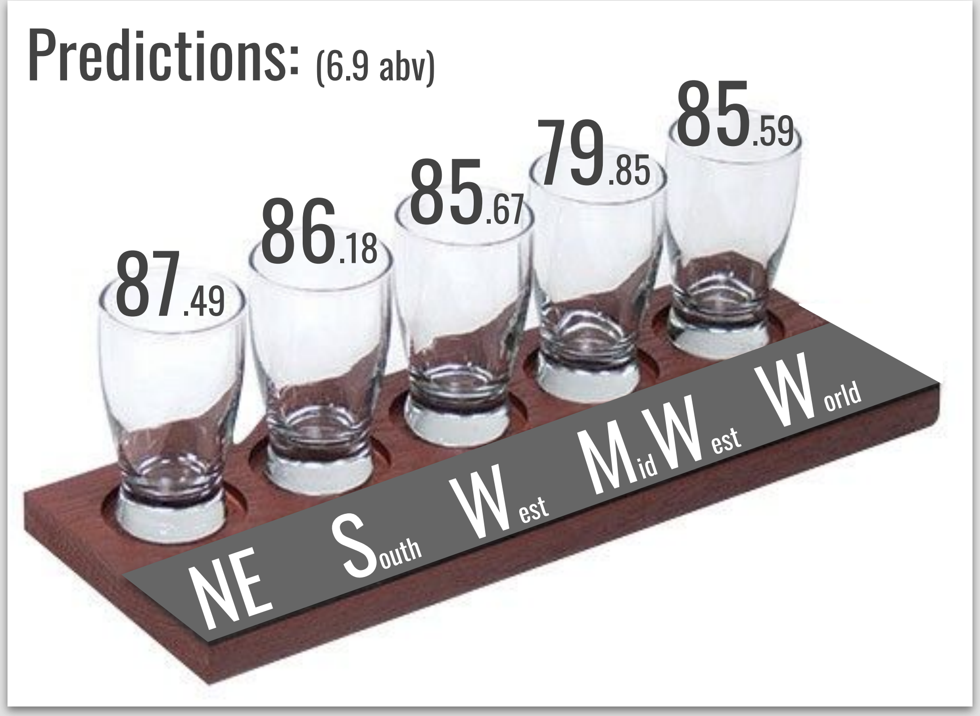
Findings:

Factors that affect score:

- ABV
- Location

Factors that don't:

- # of reviews (Popularity)
- Style
- Availability



Future Research

STONY CREEK BREWERY

BROOKLYN BREWERY

ALLAGASH BREWING COMPANY

Statistical Modeling

Of Brewing/Fermentation data

Using Data from several breweries, a model will be built that can predict outcomes of a future batch.