Capstone Project #1: A Guide to Selecting Your Next Craft Beer
Statistical Inference & Analysis

Date: May 18, 2017

Analyses of Interest:

1. Do any of the characteristics of the beer, e.g. alcohol by volume (ABV), tend to influence my ratings?
2. Is there a significant difference between my individual set of beer ratings when compared to the average user? If yes/no, what does this mean for future prediction methodologies?

Question 1: Investigating beer features and characteristics

Key beer features of interest: alcohol by volume (ABV), IBU (bitterness metric), brewery, and style. In the visuals below I am using Seaborn to create "jointplots" that help relay the correlation between the desired factor and my personal rating.

Chart 1 below shows that alcohol content (ABV) has a slight positive correlation to my ratings, however, it is weak at R-squared: 0.40. Chart 2 is the same plot, this time with bitterness metric (IBU). I noticed that the IBU metric may have some data integrity issues, given that many of the beers have values of 0.0. I likely will disregard IBU as a predictor.

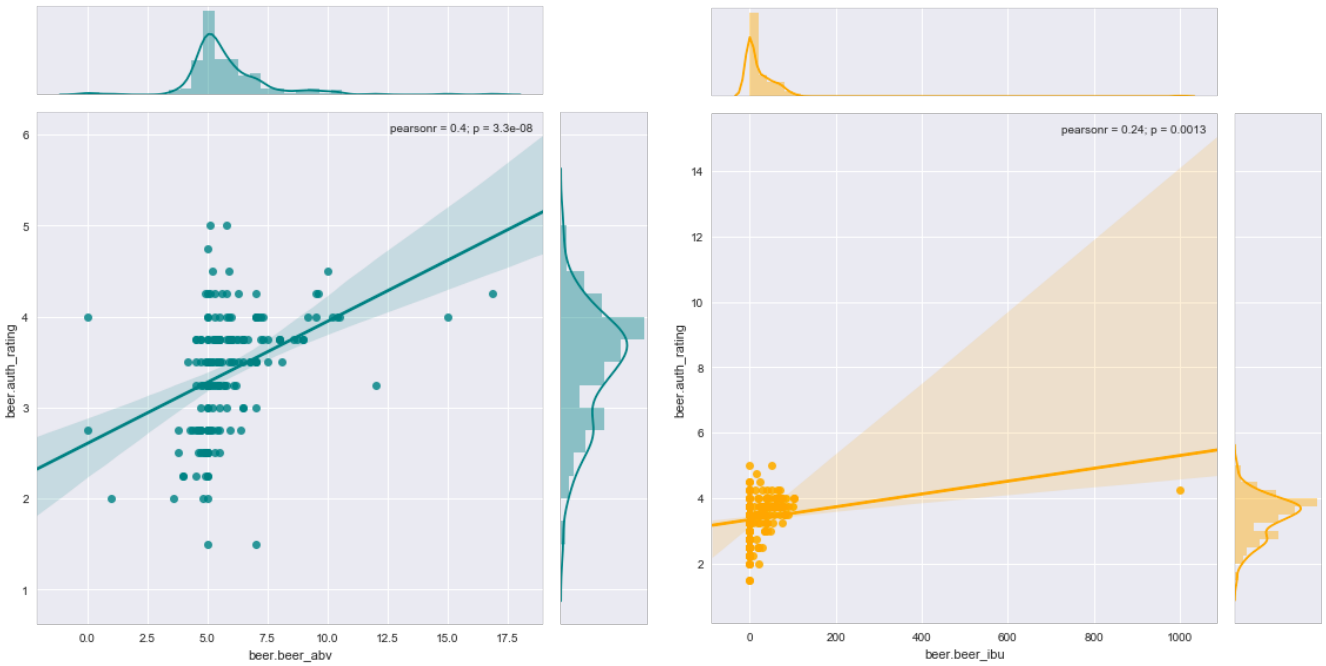

Chart 1 (left): Jointplot of my beer ratings vs alcohol by volume (ABV.
Chart 2 (right):Jointplot of my beer ratings vs bitterness metric (IBU).

Chart 3 below shows that my ratings vary wildly, even by brewery. This makes sense as beer styles and quality vary by brewery, especially given a particular user's preferences and tastes. I may enjoy the Pale Ale, but not the Lager.

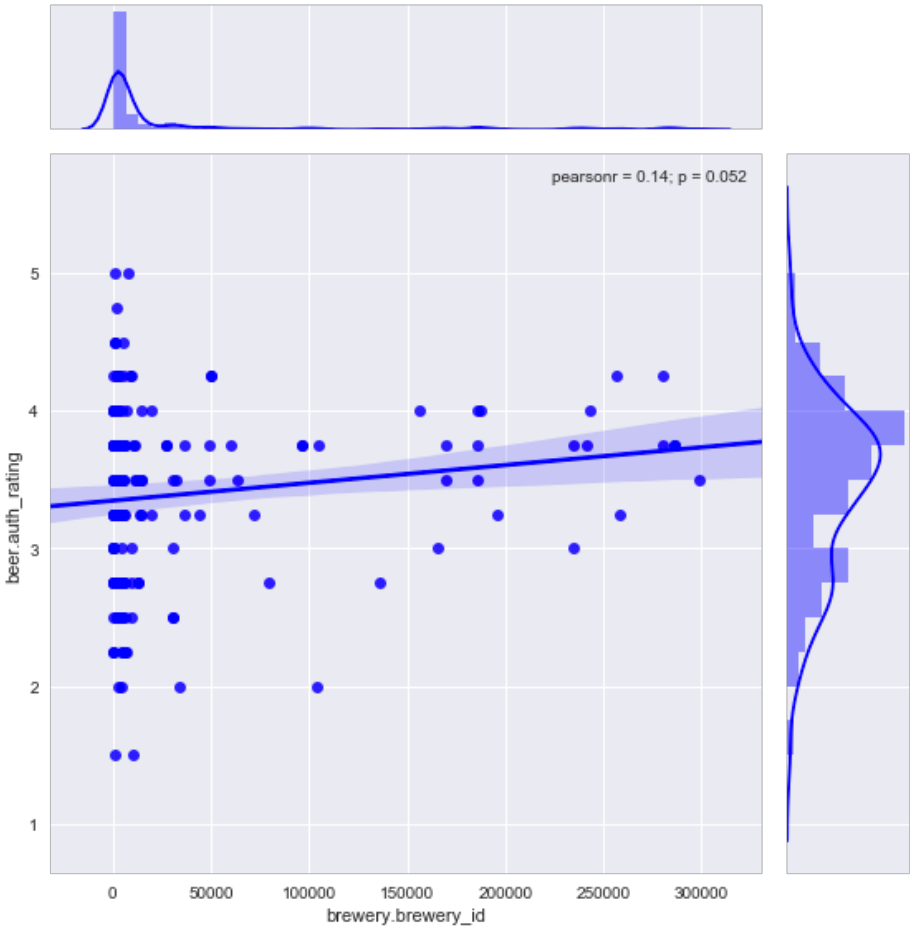

Chart 3: Jointplot of my beer ratings vs brewery.

Finally, I want to compare my beer ratings to the various styles. Given there are many, many styles in the analysis I will use ordinary-least squares models to determine relevance.

Model 1: My Beer Ratings ~ Beer Style

There appears to be the strongest correlation of any variable so far. R-squared of 0.699 shows a strong positive correlation, however, Adj. R-squared is only at 0.570.

OLS Regression Results – Model 1

Dep. Variable:	y	R-squared:	0.699
Model:	OLS	Adj. R-squared:	0.570
Method:	Least Squares	F-statistic:	5.419
Date:	Thu, 18 May 2017	Prob (F-statistic):	3.18e-15
Time:	15:25:26	Log-Likelihood:	-67.707
No. Observations:	181	AIC:	245.4
Df Residuals:	126	BIC:	421.3
Df Model:	54		
Covariance Type:	nonrobust		

Model 2: My Beer Ratings ~ Beer Style + Brewery

Interesting to note that when brewery is included, the R-squared jumps to 0.971!!! Not to get too excited, because our Adj. R-squared is actually worse, so I have crossed into *over-fitting* territory.

OLS Regression Results - Model 2

Dep. Variable:	y2	R-squared:	0.971
Model:	OLS	Adj. R-squared:	0.518
Method:	Least Squares	F-statistic:	2.143
Date:	Thu, 18 May 2017	Prob (F-statistic):	0.0783
Time:	15:32:56	Log-Likelihood:	142.58
No. Observations:	181	AIC:	54.84
Df Residuals:	11	BIC:	598.6
Df Model:	169		
Covariance Type:	nonrobust		

Model 3: My Beer Ratings ~ Beer Style + Country

Here I am looking to see if country of production introduces any benefit to the model. Overall the R-squared is lower than Model 2, however, the Adj. R-squared is actually greatest in this version.

OLS Regression Results – Model 3

Dep. Variable:	y3	R-squared:	0.814
Model:	OLS	Adj. R-squared:	0.657
Method:	Least Squares	F-statistic:	5.214
Date:	Thu, 18 May 2017	Prob (F-statistic):	1.92e-14
Time:	15:45:22	Log-Likelihood:	-24.378
No. Observations:	181	AIC:	214.8
Df Residuals:	98	BIC:	480.2
Df Model:	82		
Covariance Type:	nonrobust		

I will continue this type of model exploration once I get to the machine learning section of the course.

Question 2: Statistical analysis between my ratings and the average user

For the purposes of this analysis:

- Beer.auth_rating = My ratings (per beer)
- Beer.rating_score = the Average User Ratings (per beer)

	beer.auth_rating	beer.rating_score
count	181.000000	181.000000
mean	3.392265	3.438840
std	0.642906	0.398313
min	1.500000	2.319000
25%	3.000000	3.168000
50%	3.500000	3.537000
75%	3.750000	3.729000
max	5.000000	4.560000

While the means are very similar, there appears to be noticeable differences in standard deviation of the ratings.

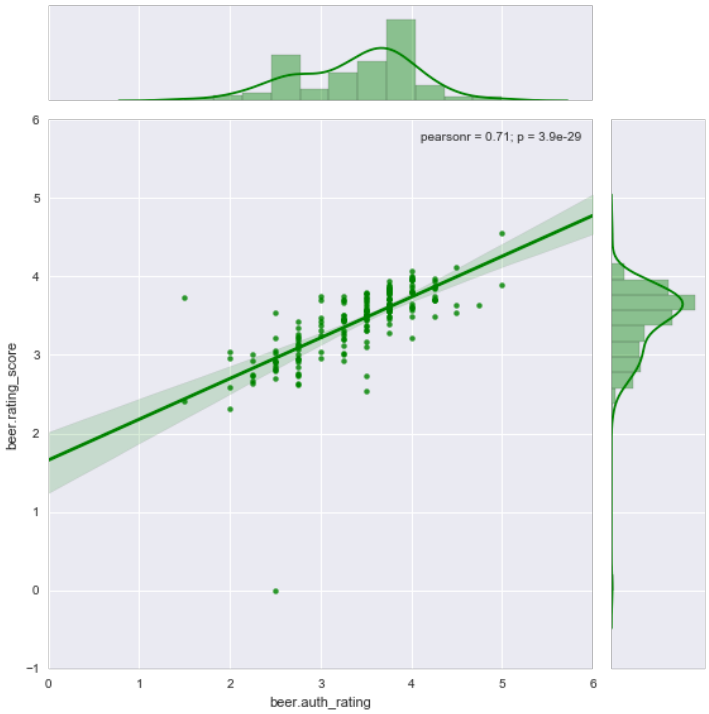

Chart 4: My Beer Ratings vs Avg User Beer Ratings

In Chart 4 I am visually comparing the two ratings, I see that there appears to be a positive correlation (reiterated by the 0.71 pearsonr coefficient), meaning that as my ratings increase so do the average user's ratings.

Two-sided T-test Setup

Null hypothesis: mean of my beer ratings = mean of avg user beer ratings
Alternate hypothesis: mean of my beer ratings != mean of avg user beer ratings

Yields the following results:

```
Ttest_indResult(statistic=-0.82850869202138933, pvalue=0.40793141463231308)
```

Interesting results. Given the high p-value: 0.408 I cannot reject the null hypothesis that the means of my ratings and the average user are different. This leaves me in an interesting situation especially given that the premise of my project was to leverage a *unique user's ratings* to predict future beer ratings. Analyzing my ratings showed that *my average ratings do not statistically differ from the average user's ratings*, however, *this may not be the case for all users*.

Given that beer style was proven to be a strong predictor I still believe that it will be interesting to continue with the project as that alone may yield different predictions by user even if the ratings are similar.