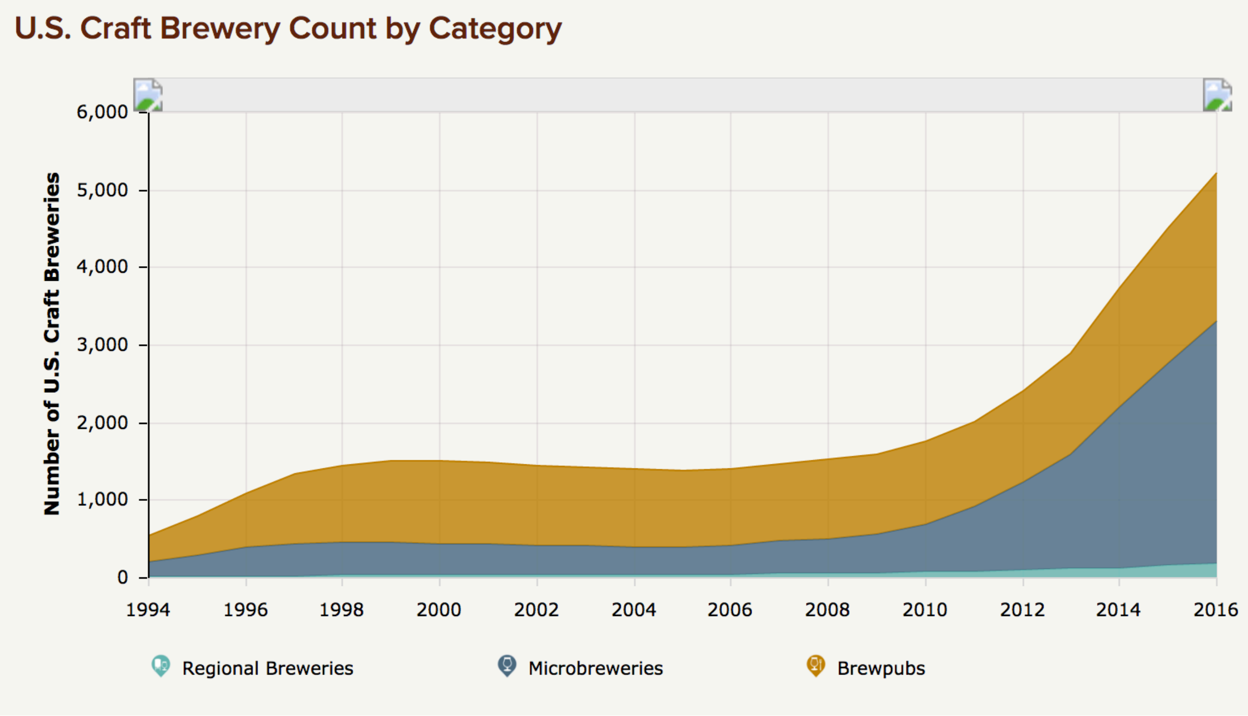
Capstone Project #1: A Guide to Selecting Your Next Craft Beer

Milestone Report #1

Date: May 22, 2017

Introduction and Objective

Since the early 1990s the U.S. craft beer scene has been on the rise. With almost 200% new craft breweries since 2010 and over 5,000 unique craft breweries in the U.S. today, beer is one of the fastest growing industries in America. Given the substantial rise in the number of craft breweries, deciding which beer to try next can be a confusing and daunting decision.



Source: https://www.brewersassociation.org/statistics/number-of-breweries/

The application “Untappd” was created in late 2010 as a platform that allows its users to “check-in” beers as they drink them, and share these check-ins, along with their individual ratings, with their friends. As an avid user of the application, it is great to be able to maintain a log of personally tasted and rated beers.

One key feature that is missing in the application, however, is the ability to leverage that data to help its users decide which beer they should buy next.

The primary objective of my initial capstone project will be to create a recommender system for Untappd users to help them decide which beer to try next given their unique tastes and historical ratings.

Dataset Acquisition

All of the data was acquired via the Untappd API. The critical limitation of the Untappd API is that **its developers are limited to 100 API calls per hour**. This required some creativity to properly acquire the desired datasets.

My approach to acquiring the necessary data was as follows:

- Leverage Python's rich libraries to pull the data via API (in JSON format) calls for analysis.
- Continue to make API calls until all data has been acquired for the two desired datasets.

Dataset #1: my personal beer set of ~200 distinct beers. This was easily acquired via 4-5 API calls since the structure of this call returned 50 distinct records per call.

Dataset #2: new (untasted) beer set of ~20,000+ distinct beers. This dataset was **extremely time consuming to acquire** given that I can only make 100 API calls per hour and each call returned 1 distinct beer record. In order to acquiring this dataset, I had to use the following approach:

- Build Python script that initiated and ceased API calls given the parameters of my account (I was limited to 100 API calls per hour).
  - To keep things simple, API call #1 returned results for beer ID (BID) #1. I then increased BID by 1 per call (e.g. API call #2 returned results for BID #2). This was repeated 20,000+ times.
- Leverage application "LaunchControl" that executed the Python script every hour until the desired record count was acquired.
  - Estimated time required: 200 hours (8+ days)
  - Actual time: 400 hours (16+ days)

Untappd API Challenges

As mentioned above, the actual time to acquire dataset #2 was over 16 days of executing the hourly script! Why did it take so long? Here were the primary reasons:

- Acquiring API access proved to be difficult as I had to exchange a few emails with the Untappd technical team and explain the purpose of my capstone project.
- The Untappd dataset has its own challenges, especially with the unique beer identifier (BID). There were times when 100 API calls would return no results due to the BID not existing.

Data Wrangling

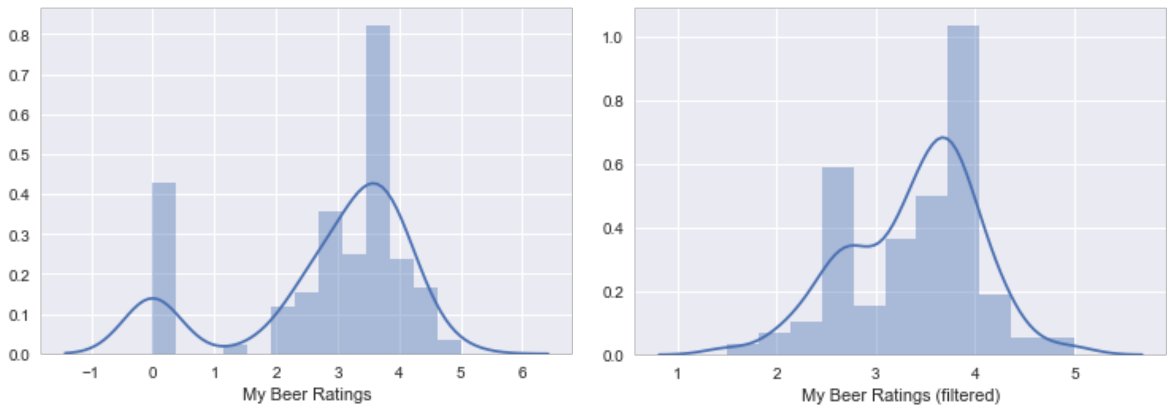
In addition to the Untappd API and data acquisition challenges, the returned results from the API calls also required some data wrangling.

Analyzing the “My Personal Beers” Dataset (Dataset #1)

Initial Data shape: 218 rows (unique beer records), 35 columns (beer features)
Initial Data Types: bool(1), float64(6), int64(9), object(19)

Null Values: Yes, some of the qualitative fields have null values (e.g. beer description, brewery city name). Fill the null values with empty strings (for now).

Outliers: Removed records if my rating was 0.0 or negative (meaning I likely forgot to rate the beer).



Left: Seaborn “distplot” of the initial data, some beers have ratings of 0.0.
Right: Seaborn “distplot” of the filtered data. Removed beer records with 0.0 (since it was user error).

Initial Count:	**35 columns, 218 records**
Null values removed:	*none*
Outliers removed	*37 records*
Final Data shape:	**35 columns, 181 records**

Analyzing the “New Beers” Dataset (Dataset #2)

Initial Data shape (as received via API calls in JSON): 29175 rows (beer records), 53 columns (beer factors)
Initial Data Types: bool(1), float64(15), int64(11), object(26)

Qualitative Fields and Null Values: such as “beer description”, “brewery Facebook page”, “twitter name”, etc. were identified and removed. This accounted for 30 factors (that were removed).

Duplicate information: **the primary challenge with this dataset is the amount of duplicate records returned in the API calls**. I believe that this is due to the fact that any user may create a record for a beer, regardless if it already exists. This is clearly not ideal, especially in a production environment. This will be discussed in further detail in the capstone final report. **9308 records (32% of the acquired records) were dropped** since they were duplicates.

Null Values: Yes, some of the descriptive records (brewery city, state) were null. Fill with empty strings for now.

Outliers: Yes, 29 records had no beer weighted rating score which was falsely impacting our mean so we will remove these records.

```
count    19867.000000                          count    19838.000000
mean         3.524786                          mean         3.529939
std          0.356703                          std          0.330505
min          0.000000                          min          1.081710
25%          3.491960                          25%          3.492635
50%          3.589430                          50%          3.589645
75%          3.649190                          75%          3.649280
max          4.749380                          max          4.74938
```

Left: Without outliers removed we can see 0.00 values for the weighted beer ratings.
Right: Only a slight change, but we have removed the 0.00 values for the weighted beer ratings.

Initial Count:	**53 columns, 29175 records**
Features removed:	*30 columns*
Duplicates removed:	*9308 records*
Null values removed:	*none*
Outliers removed	*29 records*
Final Data shape:	**23 columns, 19838 records**

Preliminary Analysis & Data Exploration

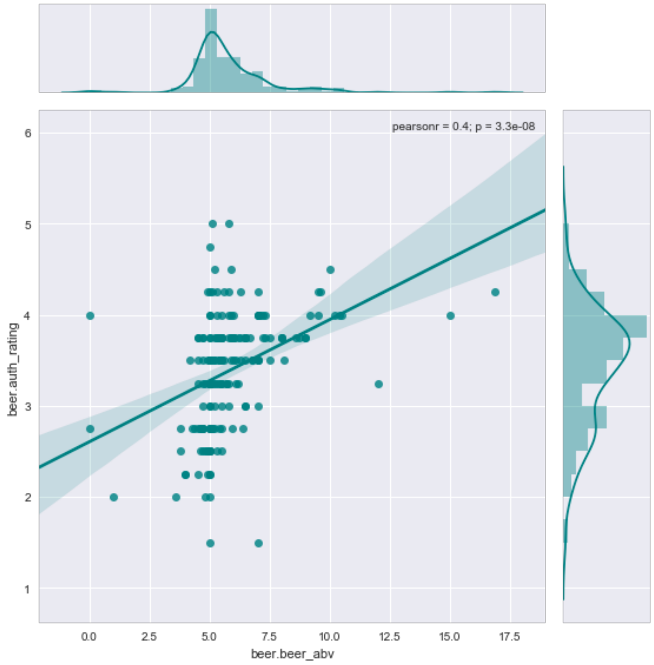
Key Analyses of Interest:

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.



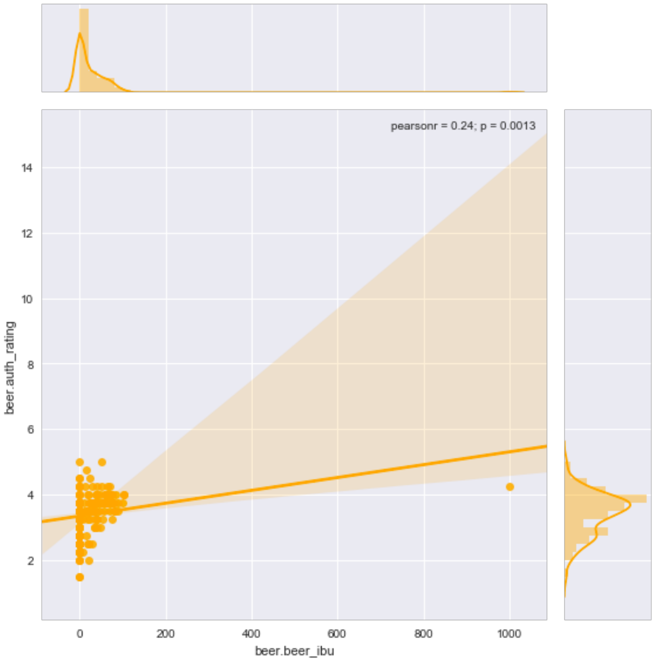


Chart 1 (left): Jointplot of my beer ratings vs alcohol by volume (ABV.
Chart 2 (right):Jointplot of my beer ratings vs bitterness metric (IBU).

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.

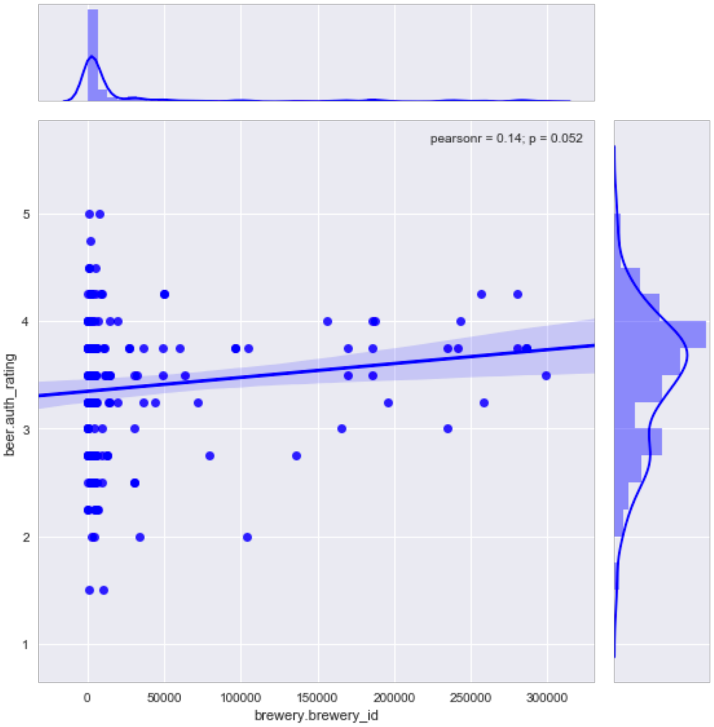
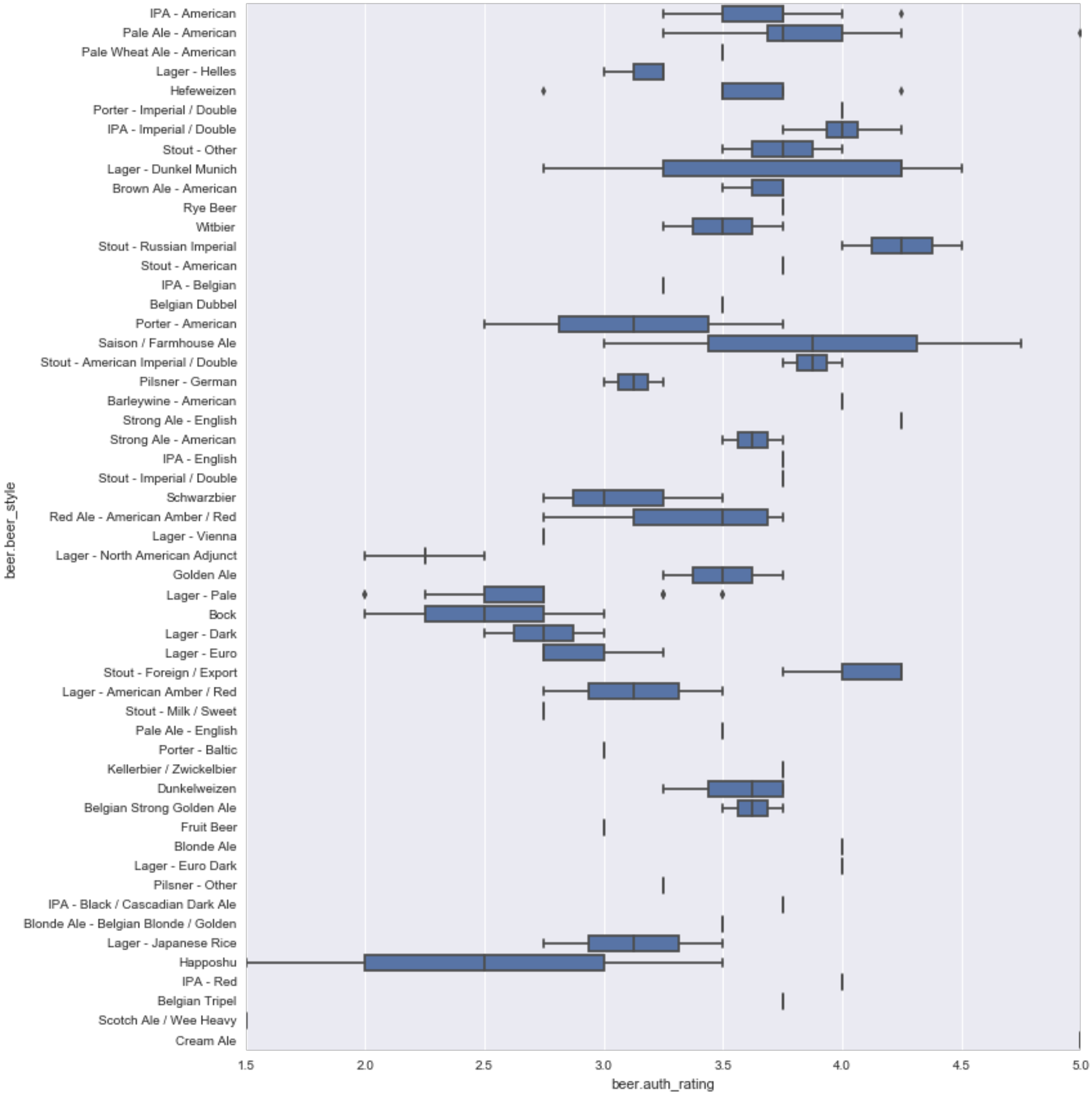


Chart 3: Jointplot of my beer ratings vs brewery.

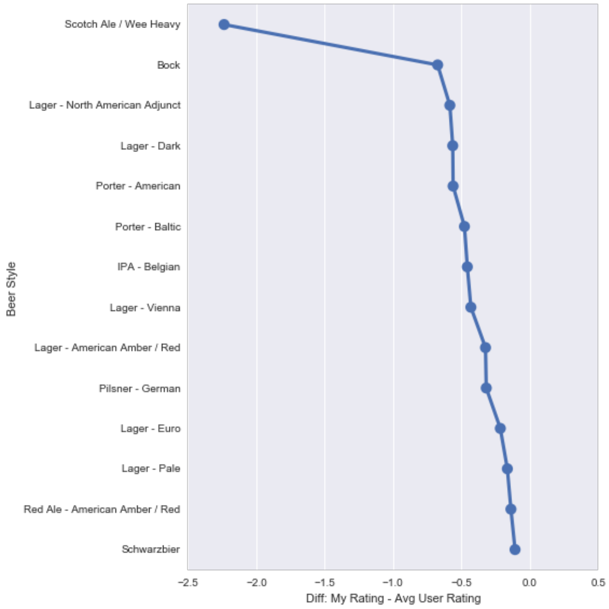
Finally, I want to compare my beer ratings to the various styles. Let's take a look at my ratings by beer style to see the mean comparison as well as the disparity of my ratings by style.

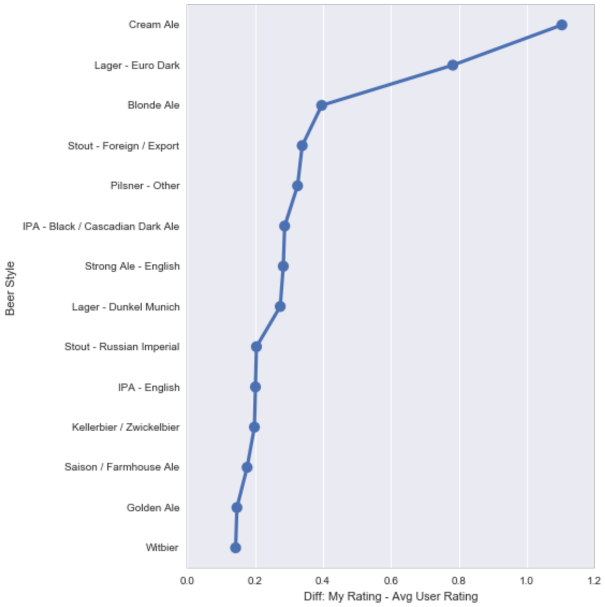


Three major callouts here:

1. The mean ratings vary quite a bit by style. This is important to note because it means that not all beer is created equal.
2. By style, we see pretty wide intervals, even some outliers by style. This means that even by style, not all beer is created equal.
3. It appears that I enjoy the "Ale" and "Stout" beer styles the most, but how does this compare to the average user rating? I will explore this next.

Taking the mean of ratings by style of beer, and then the differences between my ratings and the average user (my rating – average user rating) we can visually look at some of the styles that differ.





Although spread across many styles, the chart on the left shows that I tend to rate "lager" beers lower than the average user. Also, I really do not like "Scotch Ale/Wee Heavy". I would agree with this conclusion. On the right, we see that "Ale" beers appear frequently (IPA, Pale Ale, etc.) which I tend to rate higher than the average user. This is consistent with the fact that "Ales" are my favorite style of beer.

Statistical Inference & Analysis

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.

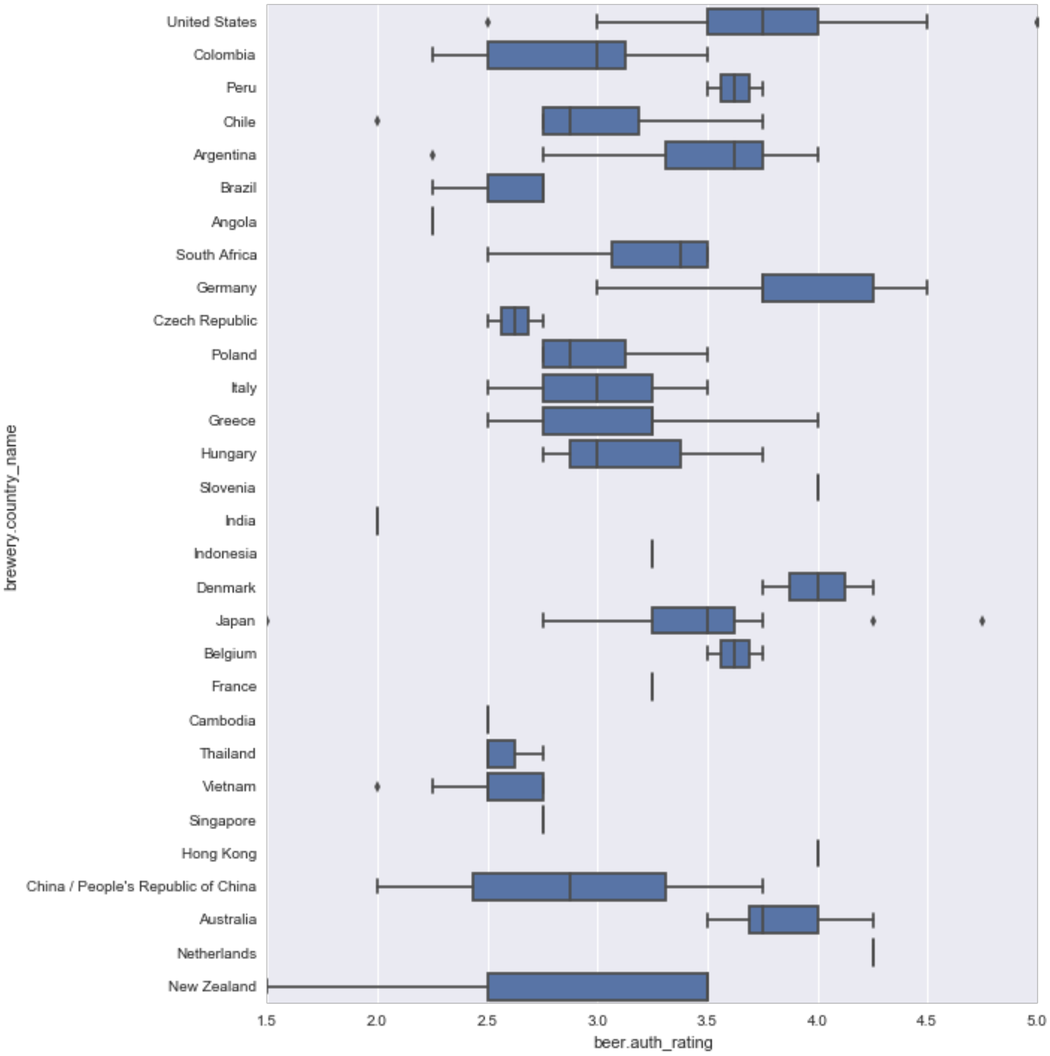
```
                        OLS Regression Results – Model 1
==============================================================================
Dep. Variable:                      y   R-squared:                       0.699
Model:                            OLS   Adj. R-squared:                  0.570
Method:                 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
```

Given that my wife and I have recently returned from traveling the world for 11 months, I want to look if there is a difference in ratings across countries. We can look at boxplot by country of beer production and review the ratings.



Three major callouts here:

1. It appears that my ratings are highest for beer produced in United States, Germany, Denmark, and Australia. I would agree with United States and Germany, but Denmark and Australia may have to be adjusted based on sample size.
2. Beer across the southeast Asian countries was not very creative or unique and tended to be lager heavy. I am not surprised that Vietnam, Thailand, etc. are lower on the ratings chart.
3. Although New Zealand has GREAT wine, I was unimpressed with their beer.

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 to explore this type of factor analysis in the machine learning section of the course.

Question 2: Researching significance 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.

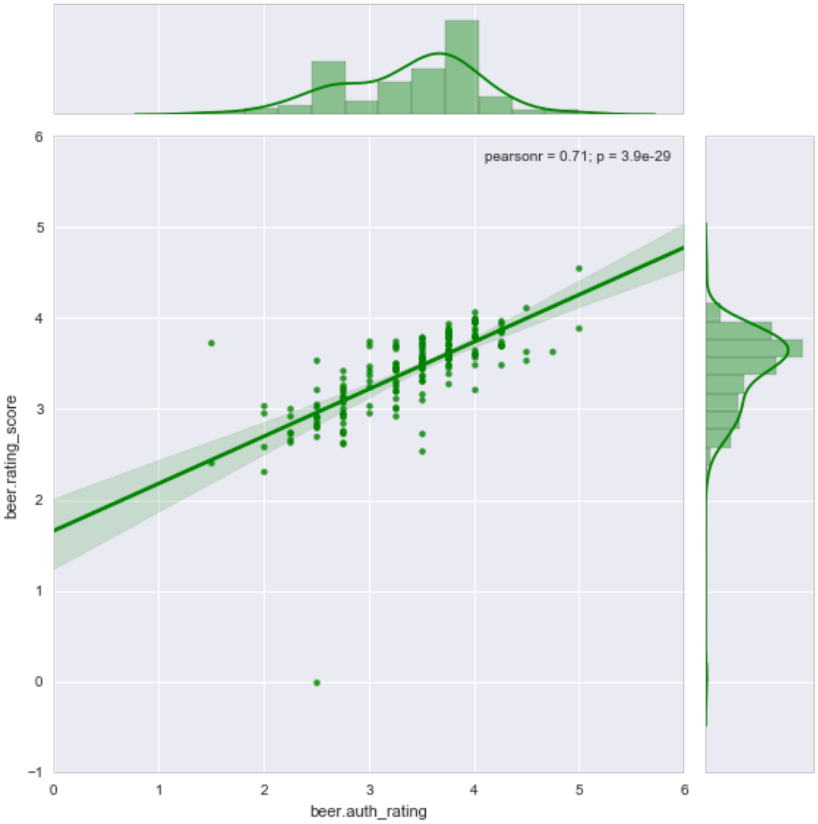


Chart 4: My Beer Ratings vs Avg User Beer Ratings

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.

Conclusion & Next Steps

1. **Not all APIs are created equal**. While I fully expected to have a learning curve associated with learning to access a web API, understand the JSON format and normalize the results, the data returned was a lot “messier” than I expected. For Untappd, this begs a discussion around the production environment setup as well as user access to input new data. It is great that users may create their own beer records (especially for home brewers) , however, there should be tighter constraints on the criteria that users may enter into the system – especially if they are on the verge of creating a duplicate record.

2. **On average the same, by style it appears to be different**. The most surprising result was the fact that on average (across the 218 beers in my personal dataset) my ratings were not statistically different than the average user’s ratings. This was surprising to me given that I believe that each user’s tastes tend to be unique. Digging deeper, we do see a difference in the mean rating by beer style. I will need to explore the significance of the beer style ratings further.

3. **Do my personal results necessarily reflect the mean**? My initial objective was to “*to create a recommender system for Untappd users to help them decide which beer to try next given their unique tastes and historical ratings*.” Given that my unique set of ratings do not statistically differ from the mean, does this change the objective of the project? I do not believe so as some users my have drastically different results. Also, I believe that there is a difference by style of beer that needs to be explored further which may show up in a comparison of different recommender system results.