**Using the Untappd API & Data Wrangling**

*Date: April 20, 2017*

Objective

Create a short document (1-2 pages) in your GitHub describing the data wrangling steps that you undertook to clean your capstone project data set. What kind of cleaning steps did you perform? How did you deal with missing values, if any? Were there outliers, and how did you decide to handle them? This document will eventually become part of your milestone report.

The Data

For my initial capstone project (“A Guide to Selecting Your Next Craft Beer”), the primary data source is the Untappd application. According to [their website](https://untappd.com/about) “Untappd is a new way to socially share and explore the world of beer with your friends and the world.” Practically speaking, Untappd is an application that allows its users to keep a log and rate the beer that they drink. Ratings are based on a 0-5 numeric scale allowable at 0.25 increments, with 5 being the highest rating. In addition to user ratings, there appear to be 30+ additional factors for each beer that may be used for analysis and prediction.

My project will leverage both my personal dataset of beers tasted as well as a dataset of untasted beers.

Data Acquisition Methodology

1. Acquire access to the Untappd API. Authenticate and explore the API and JSON formats.
2. Access and acquire a dataset (~200+ distinct beers) of my personal ratings and history.
   1. Explore, analyze, and clean dataset.
3. Access and acquire a dataset (~10,000+ distinct beers) of untasted “new” beer to forecast ratings.
   1. Explore, analyze, and clean dataset.

Challenges

* 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 primary challenge with acquiring the dataset is that Untappd limits the number of API calls allowed to 100 per hour**.
  + I plan to circumvent this by building scripts to acquire and load the data into CSV files (and eventually pandas data frames) incrementally.
* Factor analysis and engineering, given that each beer has ~50 factors.

Analyzing the “My Personal Beers” Dataset

Data shape: XX columns, YY rows

Qualitative Fields:

Null Values:

Insufficient Information:

Outliers:



Analyzing the “New Beers” Dataset

Initial Data shape (as received in JSON): XX columns, YY rows

Numeric factors:

Int64Index: 342 entries, 0 to 383

Data columns (total 23 columns):

beer\_abv 342 non-null float64

beer\_ibu 342 non-null int64

beer\_name 342 non-null object

beer\_style 342 non-null object

bid 342 non-null int64

brewery.brewery\_id 342 non-null int64

brewery.brewery\_name 342 non-null object

brewery.brewery\_type 342 non-null object

brewery.country\_name 342 non-null object

brewery.location.brewery\_city 342 non-null object

brewery.location.brewery\_state 342 non-null object

brewery.location.lat 342 non-null float64

brewery.location.lng 342 non-null float64

created\_at 342 non-null object

is\_homebrew 342 non-null int64

is\_in\_production 342 non-null int64

rating\_count 342 non-null int64

rating\_score 342 non-null float64

stats.monthly\_count 342 non-null int64

stats.total\_count 342 non-null int64

stats.total\_user\_count 342 non-null int64

weighted\_rating\_score 342 non-null float64

wish\_list 342 non-null bool

dtypes: bool(1), float64(5), int64(9), object(8)

memory usage: 61.8+ KB

Qualitative Fields: such as “beer description”, “brewery Facebook page”, “twitter name”, etc. were identified and removed. This accounted for 27 factors (that were removed).

Null Values: Information regarding some of the breweries location, specifically some records for the city and state names were missing. I decided to replace these with empty strings ‘’, as we do not need to drop them at the moment.

Bad Information: one of my (initially planned) key factors turned out to not have sufficient information to be relevant. The factor “beer\_ibu” which is a numeric factor that is a measure of the bitterness of a beer was found to be missing for XX%+ of the beers. Due to this, I removed this factor from the dataset.

Outliers:

Removed records if rating\_score was 0.0 or negative.

 

Left: Seaborn “distplot” of the initial data, some beers have ratings at 0.0 or even negative ratings.   
Right: Seaborn “distplot” of the filtered data. Removed beer records with 0.0 and negative ratings.

Next Steps