Aidan Sawyer
Thomas Kinsman
CSCI.420 - Principles of Data Mining
2016.12.01

Tapping Into the Data: Final Report

A Pint-by-Pint Analysis of Craft Beer Styles by Characteristics and Region

Introduction

Purpose

As an analysis of craft beer styles and a possible examination of the effectiveness of the most common metrics, I intentioned this project from the outset to be heavily focused on clustering existing data with known datasets. The desire to obtain the predictive ability of a classifier towards identification was much less pronounced and manifested more as afterthought and addendum than a primary goal.

Project Scope

This document is meant primarily as an explication of the strategy, outcome, and discussion of my work on the project over the past two months, as well as a summative description of some of the lessons I learned in the process of its completion. In order to slightly make up for the limited and subpar quality of the Research Paper I turned in in late October, it also covers slightly more about the domain analysis and background information than would have been otherwise required.

References

id	Resource	Description
00	Project Assignment	Assigned/provided requirements and project description (Kinsman)
01	Project Proposal	Initial deliverable outlining the topic and dataset I selected
02	Project Research Paper	Limited description of strategy, data cleaning, prior projects
03	Data Cleaning Code	Code used to clean, filter, parse the data
04	Raw and Cleaned Data	Actual data used in the analysis
05	Graphs and Figures	Screenshots and models output from the Weka analysis
06	Project Wiki	Wiki containing some additional (some duplicate) information

Domain Knowledge

Beer Ingredients

While I expound in depth on some of the finer particularities of the brewing process and the role of certain metrics, ingredients, and considerations which can have a lasting effect on the end product of a brew in the Discussion section of this document, a tacit understanding of what goes into making a beer is required for an adequate understanding of its contents.

Beer is made from four main ingredients: water, yeast, hops, and grains (traditionally malted barley). It is traditionally measured by a variety of measurements, but the ones most commonly displayed proudly on the outside of the can are the International Bitterness Units (IBUs) and ABV (Alcohol by Volume).

It can be argued that IBU translates rather directly to a measure of the (bittering) hops, as it is a measure of dissolved isomerized alpha acids. The ABV only *very* roughly follows from the amount of yeast (different yeast strains withstand varying levels of

Data

Origin

I decided against using more complex and potentially ethically questionable means of getting better and fuller datasets from the bigger players (such as brewersadvocate, ratebeer, and untappd) which provided rate-limited APIs with terms of service against data mining. Instead, I found and downloaded a .csv of around 2400 craft beers from craftcans.com.

While potentially skewed towards strictly *American* beers available in metallic cans, the dataset provides a generally substantive set of data including information on style, IBU, size (12 or 16 oz), brewery, and location (city, state).

Overview

Adjusting the location and style attributes, the cleaned dataset included the following attributes with the following types and ranges:

id	attribute	type	range	description
01	beer	nominal	1000	beer name
02	brewery	nominal	551	brewery name
03	location_state	nominal	50	state of origin
04	location_region	nominal	6	region of origin
05	stlye_12	nominal	12	mid-level, curated specification of beer styles
06	style_05	nominal	5	high-level, curated generic specification
07	size	*nominal*	2	16 oz. or 12 oz. can
08	ABV	rational	2.7-12.8	percentage alcohol by volume
09	IBU	rational	4-138	international bitterness unit
10	style_100	nominal	100	raw style value

Cleaning

Overview

While sufficiently cleaned and generally usable from the existing database at craftcans, the alternative usages of the data for *mining* necessitated small alterations to the data to make it more consistent and less specific. Beyond the tacit cleaning of ids, the plaintext ‘size’ was a simple fix made with the use of a simple matching on the two possible values of 12 and 16.

A slightly more complicated fix was to remove some of the complexity from the “location” attribute by creating the “state” and “region” features. A simple regular expression was used to get the state code from the city, state string. The region was determined by sorting each state into a region based on wikipedia data about regions. [The code for this can be found in the locations.py file].

Style

By far the most demonstrably important task in the cleaning process was determining how I was going to break up the 100 differing styles in the dataset. After consulting a multitude of style guides and finding differing accounts, I decided to manually sort the styles according to my two years of experience with craft beer and discarding the classifications I was unsure about.

With the strategy I explain below, I knew that simply having one level of classification would severely limit my flexibility and effect and decided upon 12 and 5 as the effective and practical limit from my own experience and from what I was seeing in the guides.

Output

The spread of the data favored beers from the West and Midwest [Figure 01] and pale ale and strong ale (particularly IPA) styles [Figure 02, 03], as would be expected from most surveys of the industry.

Name: Region
Missing: 0 (0%)
Distinct: 4
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	Midwest	254	254.0
2	South	200	200.0
3	West	386	386.0
4	Northeast	111	111.0

Figure 01 - Region Overview

Name: Style_05
Missing: 0 (0%)
Distinct: 5
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	Strong Ale	322	322.0
2	Dark	123	123.0
3	Ale	403	403.0
4	Lager	80	80.0
5	Other	23	23.0

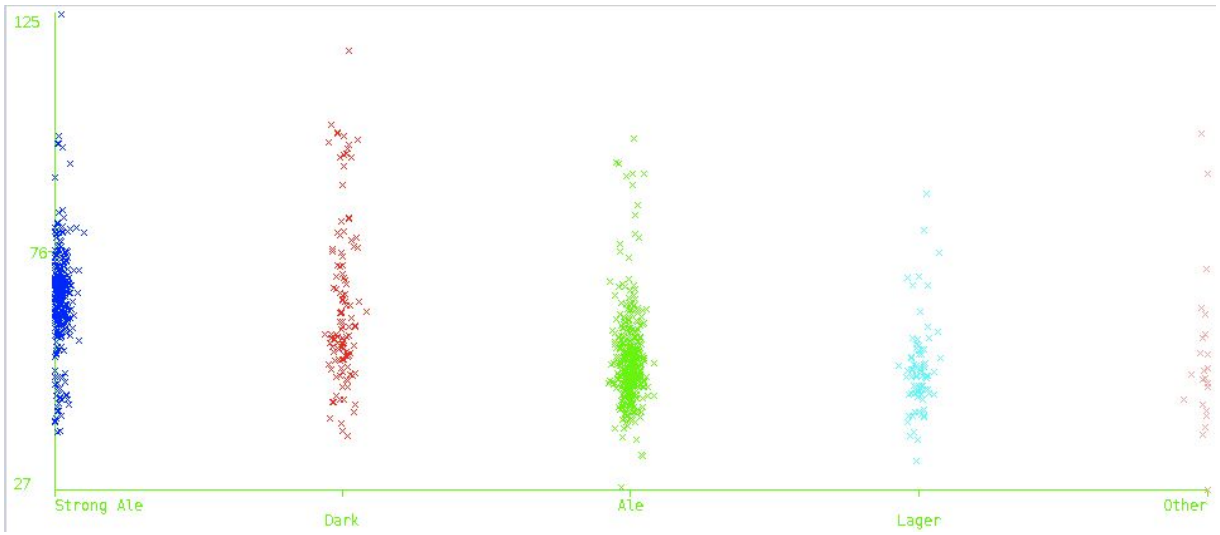
Figure 02 - Style 05 Overview

Name: Style_12 Type: Nominal
Missing: 0 (0%) Distinct: 12 Unique: 0 (0%)

No.	Label	Count	Weight
1	IPA	291	291.0
2	Dark Ale	81	81.0
3	Ale - Other	11	11.0
4	Strong Ale	31	31.0
5	Pale Ale	294	294.0
6	German Lager	36	36.0
7	German Ale	59	59.0
8	Stout - Porter	42	42.0
9	Lager	44	44.0
10	Other	12	12.0
11	Belgian	39	39.0
12	Sour	11	11.0

Figure 03 - Style 12 Overview

A quick glance at the data showed that the data obtained from craftcans has a wide degree of variance with both ABV and IBU, but that distinctions can still be made from the median of each between certain extremes, with particular regard to IBU between the Ale/Pale Ale and Strong Ale/IPA.



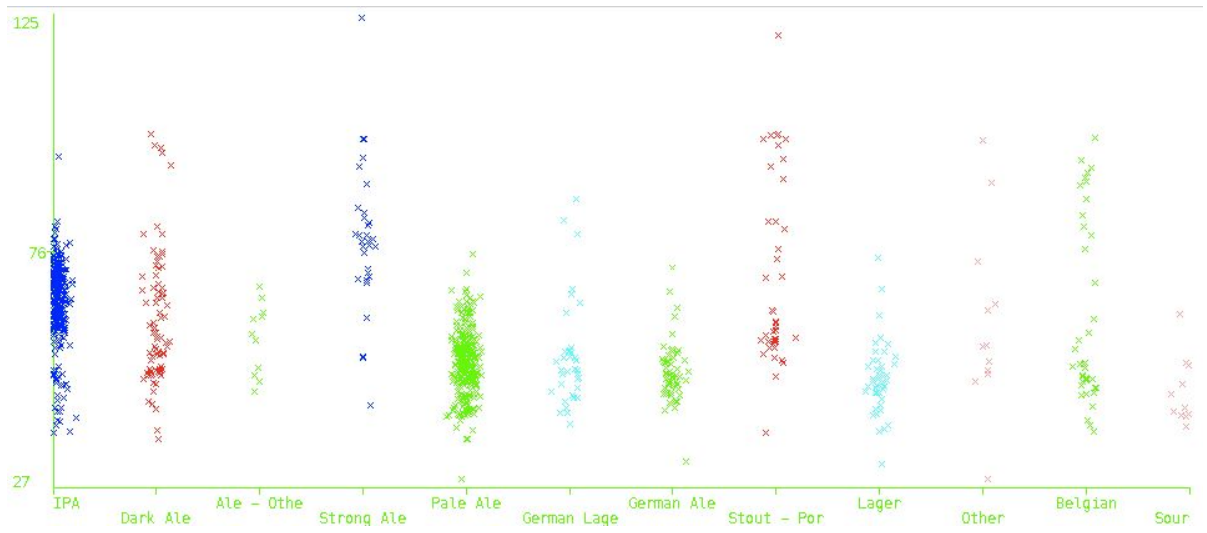
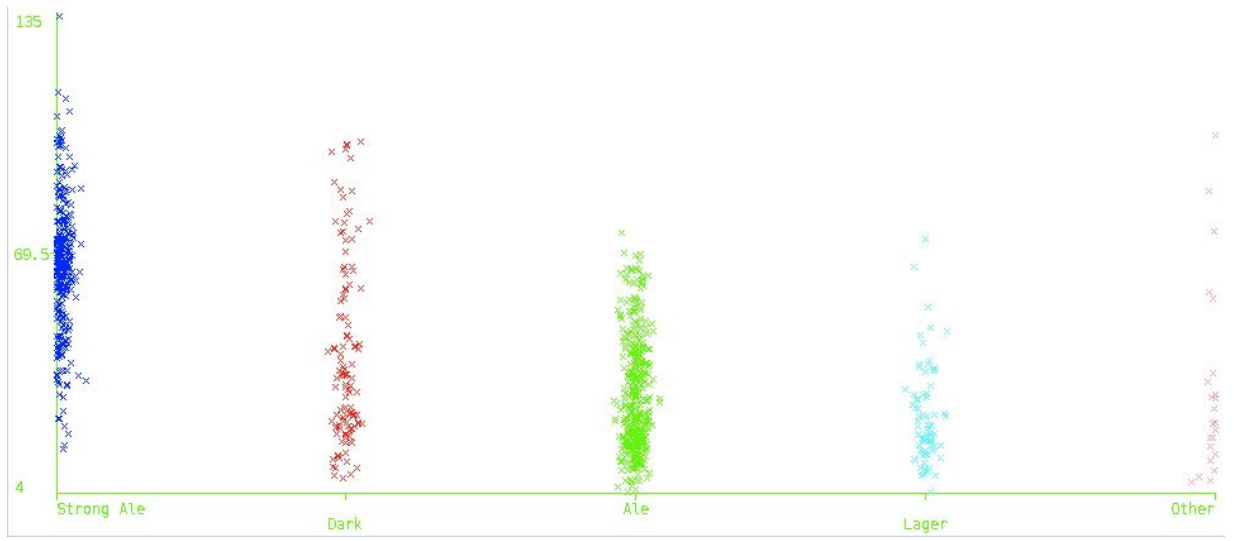


Figure 04: Style/s vs ABV



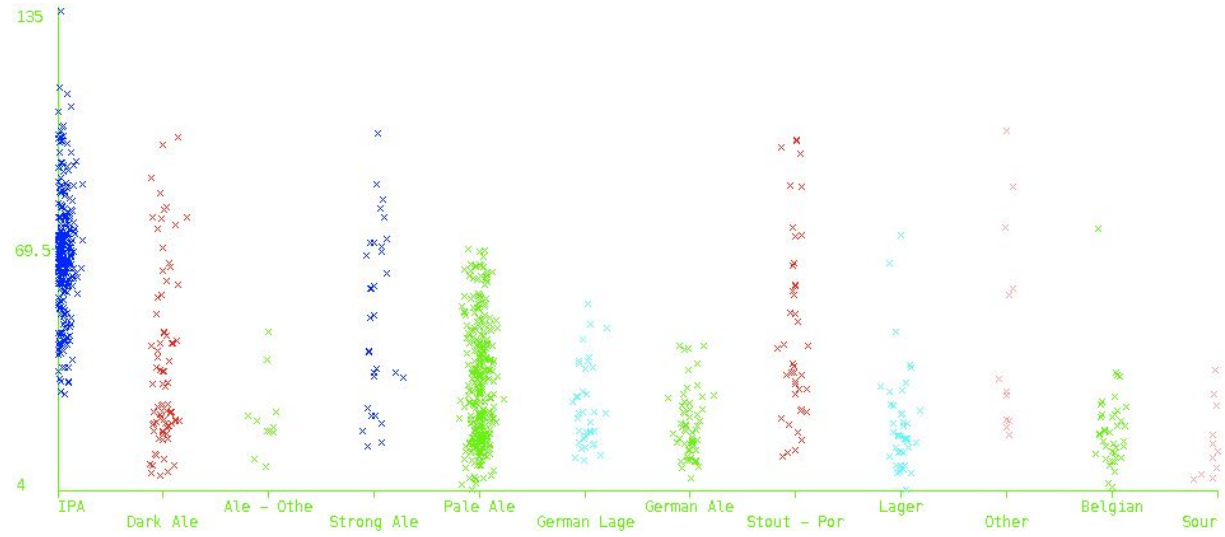


Figure 05: Styles vs IBU

Strategy

Clustering

Agglomerative Clustering

My initial strategy was to utilize the ABV and IBU measurements (perhaps including a feature for “region”) and run agglomerative clustering on the data to compare it against the list of styles that I grouped together from 100 to 12 to 5. Comparing the derived clusters to the known, I’d hoped, would provide me with some sort of basis for an evaluating the degree to which the style could be inferred from the dataset.

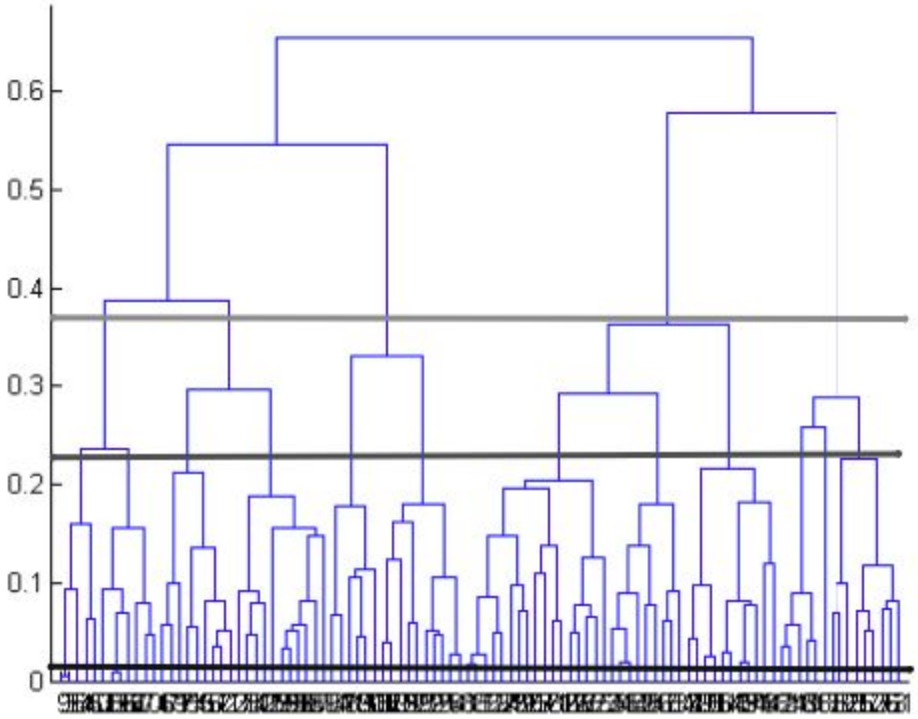


Figure 06: Sample Dendrogram with cutoffs -
[via https://www.mathworks.com/help/stats/dendrogram.html]

K-Means Clustering

Further utilizing these 5 and 12 numbers from my manual clustering of the styles based on my domain knowledge, I would then compare the runs of randomized and manually seeded (determined by the style guide) runs of the K-Means algorithm in order to better understand the variability and function of the K-Means algorithms, and to see how much better the N-Fold Cross Validation would compare the randomized run with the run seeded with known values.

Classification

K-Nearest Neighbors

Arguably the simplest way to cluster the beers, especially given a known/expected number of groupings/styles, is the K-Nearest Neighbors Algorithm. Given the simplicity of the algorithm I was curious whether the classification could be so easily done and how it would compare to other more established approaches.

Decision Tree

The common method for determining the style, as established by numerous style guide sources from the major players in the field (BJCP Style Comparison, BJCP Style Guide 2015, Beer Advocate, Craftbeer.com) was to set given ranges for each value (ABV, IBU, etc). This stated, it would seem entirely sensible to classify styles according to the sort of boxed-up nature this method of definition affords, but expecting that the diversity of the data would lead to very poorly separated (intermingled) data, I did not expect to get very effective results from this method.

Naive Bayes

I planned to use Naive Bayes with a series of ranges, partially determined by the decision tree output, to attempt to classify and compare styles based on the region, ABV range, and IBU range, but was unable to find the time to do get this implemented. I expected the data gathered from the decision trees ranges, modified or inter-related with the style guides, would afford an effective classification system that might provide a meaningful and semi-reliable classifier given at least region, ABV, and IBU.

Implementation

Looking at the Data

After spending an inordinate amount of time trying to figure out how I was going to get this all to work with inter-style classification/clustering, I had the idea to refine the scope of the calculation back to two main style families that had similar yeast and malt characteristics, allowing me to hone in on the hops and ABV.

This change proved very successful and provided me with data that appears ripe and primed for input into the algorithm, granted, perhaps some more work on normalizing the axes and transmuting the data. The graph output below depicts ABV plotted against IBU with some jitter and the coloration determined by style12:

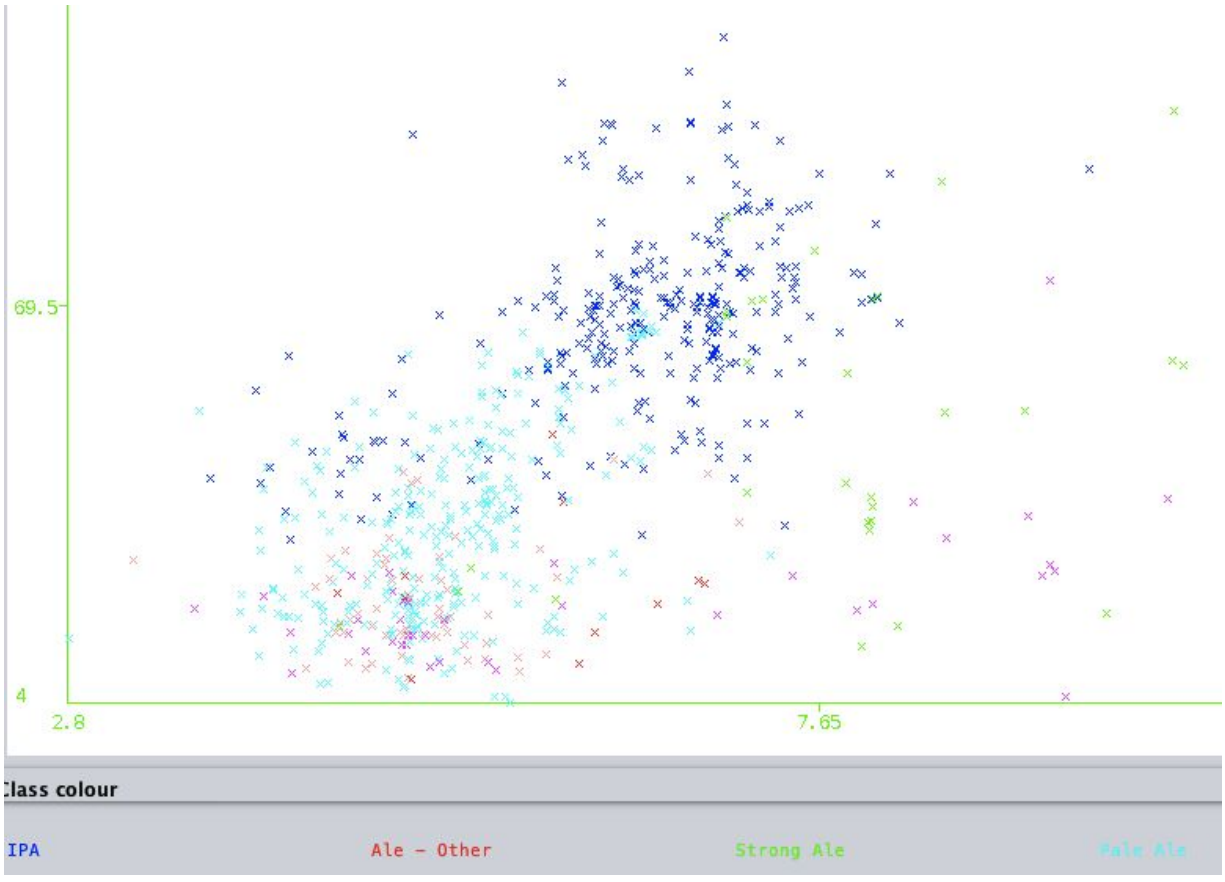


Figure 07: Ales - ABV vs IBU with Style12 Coloration

Closer analysis of the graph suggests that IPAs can be grouped at the top right with the highest *joint* ABV and IBU, while Pale Ales are shown to be the opposite with a grouping on the bottom left. In contrast, Strong Ales are shown to have *variable* IBU, but a strong tendency towards a high ABV, while belgian ales are shown to have a variable ABV, but a reliably low IBU.

Reconsideration

Adjusting my approach to the data I saw, I speculate that the combination of further normalized and pruned results, better optimized to become axially aligned would be the most effective would be the best mode of action towards achieving the most accurate result. Combining these ranges with the naive bayes theorem, I contend, could have granted results that would be feasibly useful.

Results

Complications of Inter-Style/Family Clustering

Inter-Family clustering cannot be effectively achieved with the limitations of the two selected variables. The number of latent variables causing differences that are not accounted for in the dataset prevented any effective analysis or

High ABV Distinguishes Strong Ales

As mentioned by looking at the data in the chart, the IBU is an effectively useless distinguishing factor between Strong and other Ales, as they are almost entirely determined by alcohol content.

Low IBU Distinguishes Belgian-Style Ales

Belgian-style ales are primarily determined by a low IBU content and can have a highly variable alcohol content. This follows from the traditional breakup of Belgians into single, dubbel, triple, and quad styles.

Pale Ales and IPAs Correlate

Pale ales and IPAs seem to be on a continuum in which the IBU and ABV are equally determinate. This suggests that the creation of a feature relating the two may be a useful single feature to distinguish the two from each other.

Challenges

Weka Failure

I found myself categorically unable to get most of the actual classification and clustering functions to work with a limited knowledge of Weka's functionality. I miscalculated the ease with which I could run these analyses functions after miscalculated the importance and ease of implementing them myself.

Learning Curve

The largest challenges, beyond a gap in understanding of Weka, R, and data mining techniques before beginning the project, were faced before the scope refinement and reconsideration. I spent more time than I'd have liked over-fiddling with learning new tools, testing potential inputs, and crafting my own implementations.

I'd started the project with the intention of writing my own versions of each algorithm instead of using the R and Weka presets, and I feel that in doing so I focused far too highly on the very specific

sections of how to translate the complex ideas into methods and data structures and not enough on the theory and tooling that would allow me to conduct the analysis much more quickly.

Had I focused more and earlier on learning one or two tools in depth and focusing on the strategy instead of the systems, I feel as though I would have gotten much more thorough and positive results and would have been able to try more algorithms. Devoting more time to learning R, in particular, would have also allowed for repeatable, and beautifully displayed results, but it wasn't really a possibility given my available time and workload this semester.

Latent Variable Identification

As simple as it sounds, I found much of the difficulty I faced was the result of dropping the context and definition of what my data actually *did* measure and which *latent* measurements I was not taking into account the certain "blind spots" I mention above and how that would affect my results.

Knowing explicitly about the tremendous and increasing variety within and across beer styles given the craft beer explosion, and of the complexity of taste interactions that go into beer, I should have expected a greater difficulty in picking them apart independently. Though I expected this to be eased by my breaking the styles into a smaller and smaller subgroups, I did anticipate a greater correlation than I started to see when I looked at my results for the first time.

Stepping back, going back to the domain research, and reconsidering the data I was working with, I recalled two major latent variables of yeast strain and malted grains that my data had no method of accounting for and that each have enormous effects on how style is determined.

When I, at last, accounted for this difference and focused on just light-colored ales of similar size the roadblock was lifted and I could make feasible progress. That realization just ended up coming a bit too late for me to obtain measurable results.

Tools

I used a variety of tools throughout the course of the project and found that they each had their unique strengths. Python, was by far what I was most familiar with and competent in, and I effectively used the language to import, clean, filter, alter, and export the data, as well as to create features in a flexible and convenient way.

I found RStudio to be very powerful for data exploration and its packages to be powerful. I know that a lot of my peers used it for its packages with some ease after learning it, and regret not using it more heavily in the final phases of gathering and calculating results.

Weka, additionally, proved invaluable for very seamlessly visualizing the relationships between variables in a way that I could not have been able to do otherwise. I did, however, find it very difficult to make sense of its inputs and outputs for the clustering and classification, and wrongfully over-trusted its capacity to automate and understand what it was I was intending it to do for me.

Lessons Learned

Data Cleaning and Selection

While enormously helpful in formating the data in such a way as to make them most sense of the information it holds, it can be a tremendous time sink and distraction. It's best to visualize early and often, and look at creating features to refine and simplify what has already been discovered.

Strategy Should be Data-Driven

I put a large amount of effort into planning out a chain of actions and strategies for stringing along the various algorithms into and amongst each other, and I believe this was in error. A better way to view these algorithms is more like tools in a toolbox, each with particular strengths and weaknesses given the

problem that needs to be solved. In retrospect, I would have spent far more time working on the decision tree and naive bayes classifications and valued the clustering algorithms less intensely.

This further translates in the extent to which I overvalued domain research and knowledge acquisition, planning what I thought the data should conform to as opposed to teasing out what was actually there.

Nominal Values Ground Understanding

One of the most useful features of Weka's visualization tab was the ability to easily select the coloration of the various points according to cluster id or other nominal value. This single feature helped immensely when trying to locate and make sense of the data, and the ability to switch between them readily and visually without having to alter code and rerun was very helpful.

Context and Latent Variables

Thinking back to one of the first days of class it was mentioned to view data from a variety of angles and perspectives ("walls" as containing barriers *and* as paths). This was by far the biggest thing I've learned from the project.

Naive Bayes Can Be Very Powerful

Looking at the examples in class, it never seemed like Naive Bayes could be as highly potent as I now recognize. When particular features have been discovered and are then connected within ranges, I've no doubt that they have an amplifying effect connecting the results of differing pieces of analysis.

Discussion

Larger Dataset

I was initially very pleased with the size of my dataset, coming in at around 2400 entries, but was surprised with how quickly it dwindled down as I cleaned it and selected sub-categories of the data. While I recognize the additional complexity obtained by having an over-supply of data, I did not feel as though that was an important failing point

Intra-Family Clustering Depth

Given a limited time period, restricted dataset, and lack of expertise, the effective scope of my analysis had to be limited to a smaller subset (light-colored ales) of the veritable mosaic of styles that I could have analyzed. Matching, comparing, and inter-informing the outcomes of this larger and more systematic analysis with the official style comparison and style guides of the Beer Judge Certification Program (BJCP) could have profound effect on the judging, analysis, and understanding of what makes beer styles.

Inter-Family Clustering

While I maintain that with a very rich and detailed dataset with a very large number of complete, cleaned entries, and carefully crafted features, classification of style would be *possible*, but I do not think that it would be useful or at all comparable to a human's ability to guess, identify, and determine styles independently.

This is due to a variety of factors including the difficulty of quantifying *perceived* taste and flavor, the tremendous variety and inter-mingling of styles, the interplay between various flavors and parts of the beer.

Alternative Measures

A thorough understanding within *or* between styles and style families can simply not be accurately or meaningfully obtained with only these two metrics (ABV, IBU). The creation of features from combinations of many of the common other variables have already proven themselves to be much more important.

The analysis of this topic could be much better and thoroughly conducted with a greater amount of data of more varying type. As mentioned briefly in my domain analysis, there are a great number of factors that I explain below that can have a tremendous effect on the outcome in terms of style, taste, color, and mouthfeel.

SRM

Standard Reference Model (SRM) is the measurement of the coloration (pale to dark) and visual clarity of a beer. Even within the field of styles that I obtained, distinguishing between a black IPA or a regularly IPA, for instance, while within the same style_12, would be nearly impossible without a measurement for this coloration. [Black IPA and What Makes a Style]

Gravity

The final and initial gravity of a beer contribute to the eventual alcohol content as well as the perceived sweetness of the beer. This perceived sweetness serves to counteract the IBU measurement so critical to the actual substance of the analysis I made. As mentioned in "Forget IBU, Think About BU:GU"

Grain Type and Percentage

A large part of what goes into the body, mouthfeel, and coloration of a beer is the variety and roast-amount of the grains that were used at the beginning of the brewing process.

Yeast Type

It is difficult to quantify just how much effect the yeast strain has on the determination of style family, to the point where it very well might be possible that it is the single most effective predictor for style available. This would be an interesting and worthwhile project unto itself.

Hop Varieties

There is a high degree of variability and finesse in selecting different hop varieties and families that can give you a much greater amount of information than the measure of isoalpha bitterness captured by the IBU, even when balanced against gravity and other measurements. C-Strain hops, for instance, (citra, cascade, centennial, columbus, etc) are often used for bittering and contribute to IBUs greatly, while mosiac-style hops, for instance, added later in the brewing process contribute more to taste and aroma.

Addition Times and Temperatures

Craft beer is aptly termed "craft" and brewing qualified as an *art* as well as a science for very good reason. Another hugely important and difficult to quantify-with-a-data-attribute element is the point at which each ingredient is added in the brewing process and the temperature at which the brewing process is completed.

Speaking with the head brewmaster at Crafty Ales and Lagers in Phelps, NY, I found that it is possible to brew two entirely different tasting beers with the same ingredients solely by altering the temperature at which the wort is mixed. He mentioned a particular porter (or stout) called "Curious George" whose nose and taste at the front could be entirely dominated by a banana-type flavor if the brew kettle temperature was altered by just a few degrees.

Adding Rating

As an interesting aside, utilizing user rating data against style would be a very interesting and potentially profitable project idea. Knowledge of trending beer styles by region, state, demographic, season, etc. would prove invaluable to beer distributors, brewers, and even beer loving customers. A particular look at one's personal Untappd data for instance could grant you powerfully predictive recommendations about what beer or beers I might like to try next, or like pandora does with music, help me to identify the characteristics that I care about most in beer (heavily malted, high flavor and strength, sweet and "chewy" beers with a high gravity).

Data Availability

I was very surprised at the ease with which I could find and potentially scrape data from websites as well as aggregate data from API calls online about this topic. And almost equally surprised by the degree to which this practice seemed to be discouraged. Multiple attempts at contacting the owners of this data ended in a dead end, and Untappd actually came back to me to say it was directly against their terms of service to use their API for that purpose. This was an interesting ethical issue that I decided to deal with with caution, leading to my selection of the craftcans data.

Works Cited

- BJCP Beer Style Comparison http://www.bjcp.org/cep/BeerStyleComparison.pdf
- BJCP Beer Official Style Guide
- IBU vs BU:GU http://www.pencilandspoon.com/2012/11/forgot-ibu-think-about-bugu.html
- ABV: https://www.beeradvocate.com/articles/518/
- SRM: https://www.morebeer.com/articles/beercolor
- Black IPA and What Makes a Style: http://craftbeerusa.blogspot.com/2011/03/black-ipa-and-what-makes-beer-style.html
- Mike from Crafty Ales and Lagers
- Craft Beer Anonymous Podcast Episodes 135, 161
- Brewers Association Regional and Economic Statistics https://www.brewersassociation.org/statistics/by-state/

Note: providing references for some of this information is complicated by the fact that much of the knowledge I've used here comes from conversations, beer tastings, and independent research over the past 1.5 years of my life before the project began. While much of this information can be found in the sources I listed, the lack of particularized reference should not be considered an intentional plagiarism. Some of this stuff (podcasts, conversations) are also very difficult to formalize, and since much of the first order expertize on this subject comes from homebrewers and hobby-ists, many sources normally considered disreputable (e.g. blogs) are often the most in depth and authentic references available.