Tapping Into the Data

A **Pint**-by-**Pint** Analysis of Craft Beer Styles, Regions, and Metrics

Aidan Sawyer
Github - atla5
Untappd - NewEden

Outline

- Background / Domain Analysis
  - Beer
  - ABV
  - IBU
- Data
  - Origin
  - Graph
  - Cleaning
    - General
    - Region
    - Style
  - Visualizations
- Strategy
  - Clustering
  - Classification
- Implementation
  -
- Results
  - Strong Ales
  - Pale Ale
  - Belgian Beers
- Tools
  - Weka, Python, R
- Lessons Learned
  - Watch out for latent variables
  - Retain Data Context
  - Data-Driven Strategy
  - Power of Naive Bayes
- Discussion
  - Inter vs. Intra Family Clustering
  - Additional Data

Domain Knowledge

- Domain Research
  - Most enjoyable of any project to date
  - Podcasts, blogs, brewers, firsthand experience
- Beer Ingredients
  - Water, Yeast, Malted Grain (Barley), Hops
- Beer Metrics
  - Alcohol by Volume (ABV)
  - International Bitterness Unit (IBU)
  - Standard Reference Model (SRM)
  - Original and Final Gravity (OG, FG)
- Beer Styles
  - Ale and Lagers
  - Country of Origin
  - Gradients

Data - Introduction

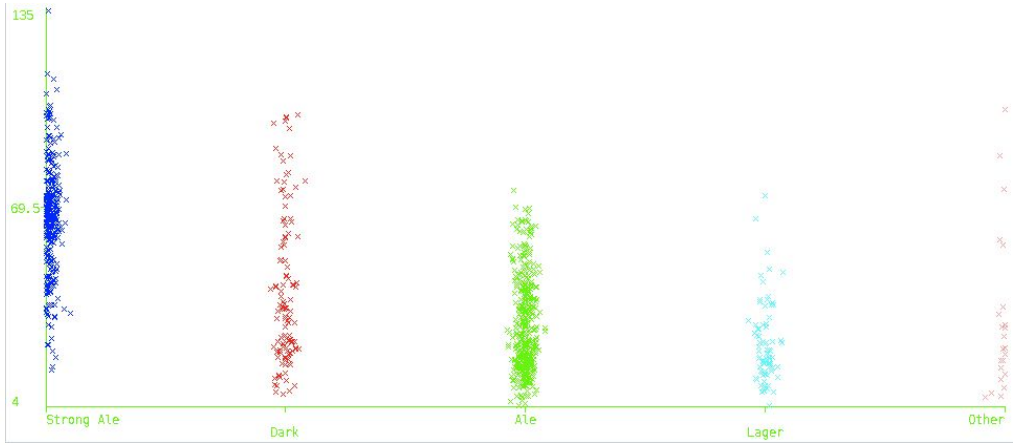
- **Craftcans** - American Beer available in cans
- EntryID, Beer, Brewery, Location, Size, Style_100, ABV, IBU
- Cleaning
  - 2,300 (all) -> 952 (full) -> 726 (strong, ale)
- Overview
  - Mostly IPAs and Pale Ales
  - Limited data from other styles
  - Mostly from the West and Midwest
  - Impressive variance in ABV and IBU

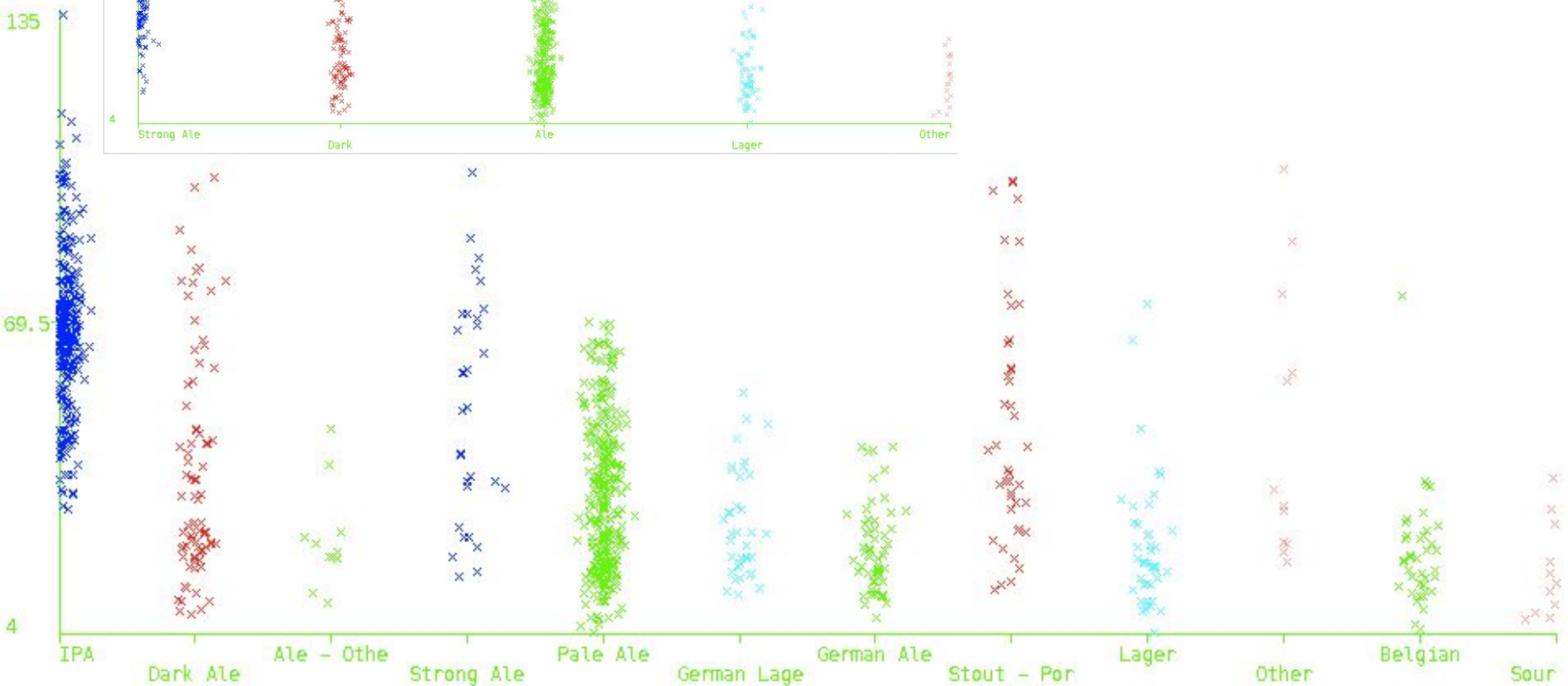
https://twitter.com/craftcans

Data - By Attribute

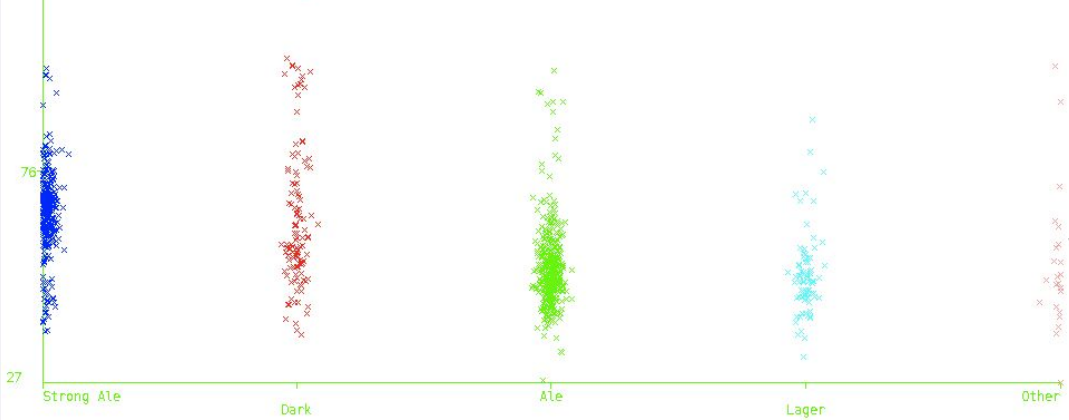
id	attribute	type	range	description
01	beer	nominal	1000	beer name
02	brewery	nominal	551	brewery name
03	location_state	nominal	50	state of origin
04	location_region	nominal	6	region of origin
05	stlye_12	nominal	12	mid-level, curated specification of beer styles
06	style_05	nominal	5	high-level, curated 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

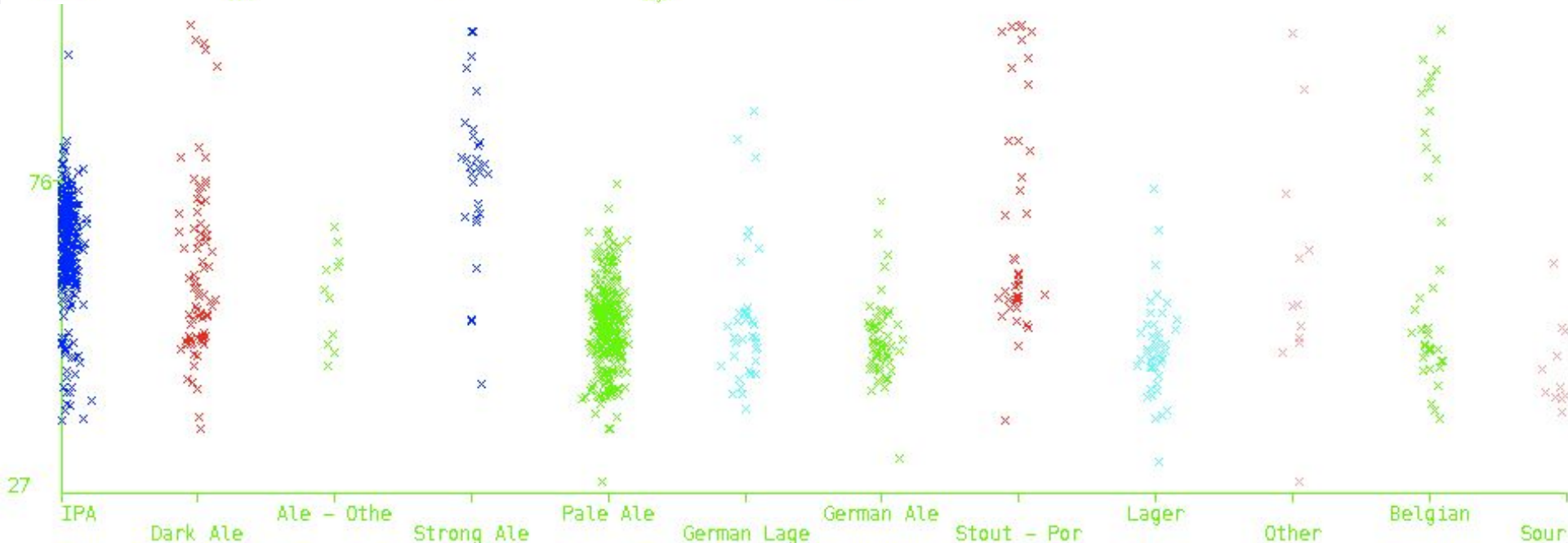
Style vs. IBUs



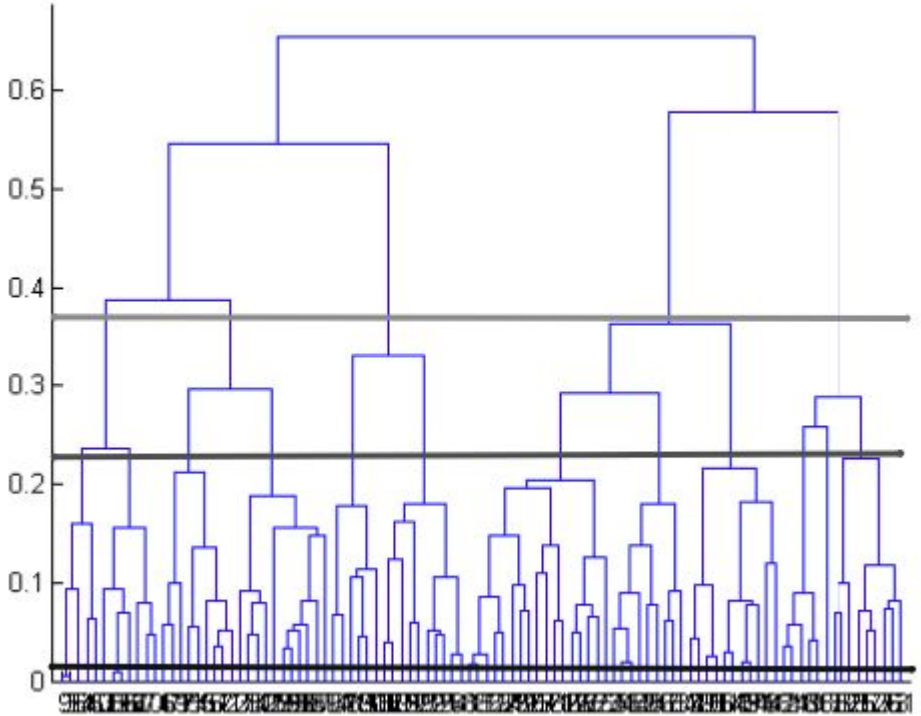


Style vs. ABV


76
27
Strong Ale
Dark
Ale
Lager
Other


76
27
IPA
Dark Ale
Ale – Othe
Strong Ale
Pale Ale
German Lage
German Ale
Stout – Por
Lager
Other
Belgian
Sour

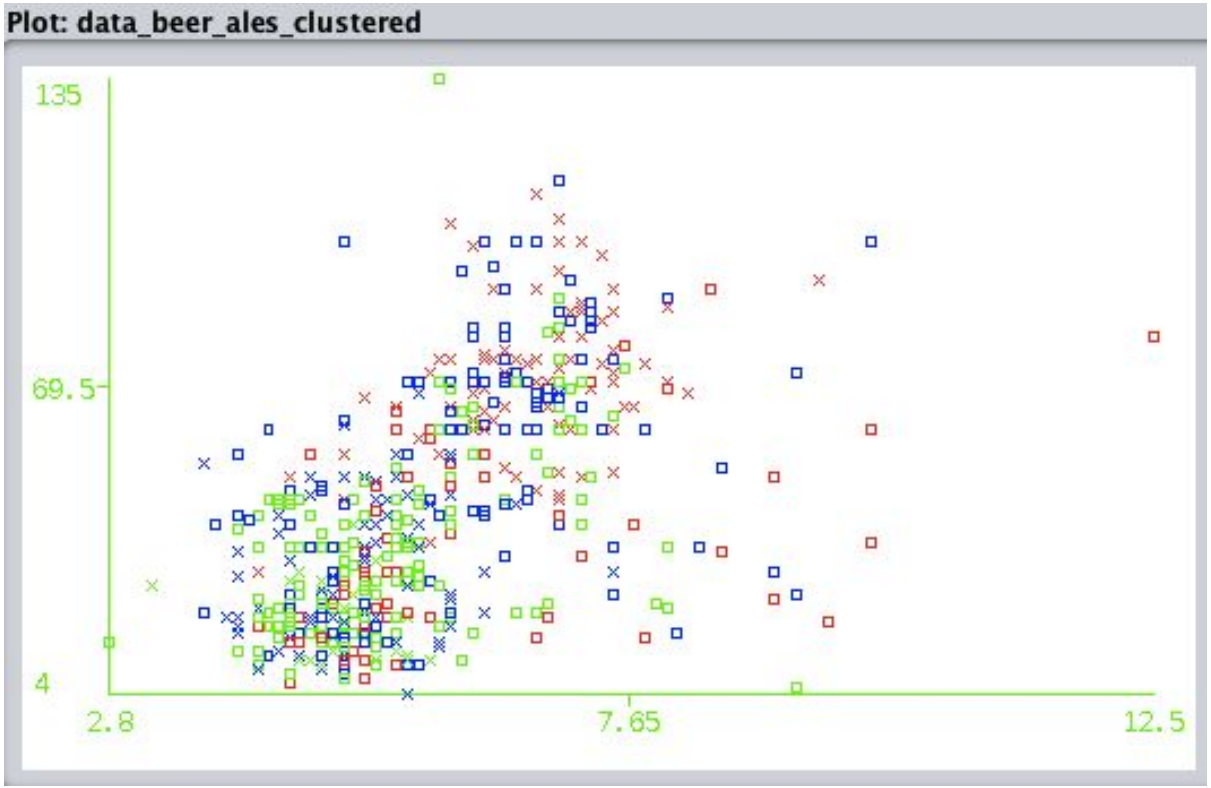
Strategy - Clustering - Agglomerative



- Run agglomerative clustering based on IBU and ABV ranges (separately and together)
- Review and Compare data at 5 and 12

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

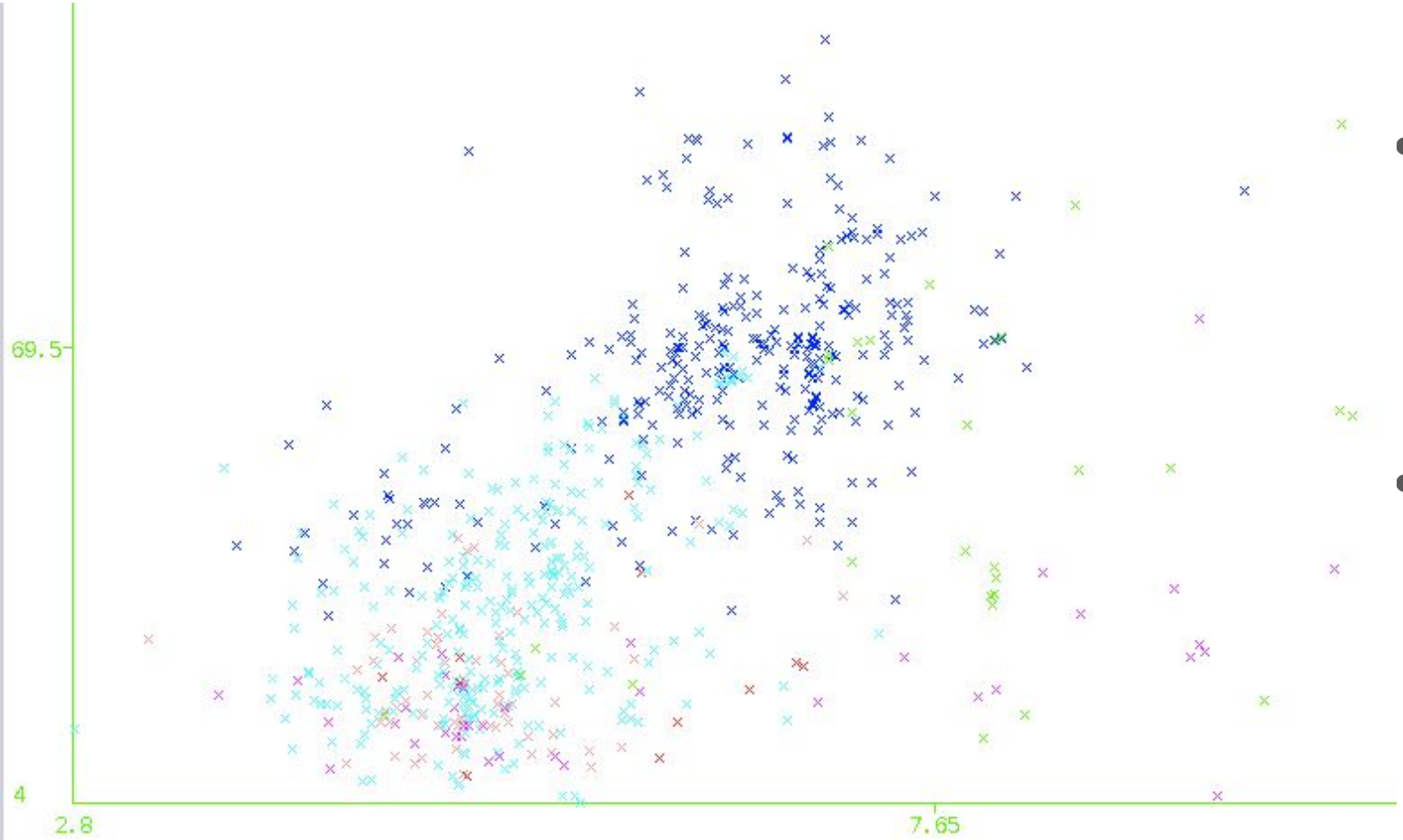
Strategy - Clustering - K-Means



- Give K-Means ‘5’ and ‘12’
- Compare resultant clusters with the known styles

←----------[Didn’t work]

Strategy - Classification - Decision Tree



- Organically create a Decision Tree after axially aligning the data
- Compare resultants with the Style Guide

←------PCA per attribute?

^ ^ objective?

Strategy - Classification - Naive Bayes

REGION	IPA	Not IPA
Midwest		
West		
Neither		

ABV	Strong Ale	Non-Strong
Over 6%		
Under 6%		

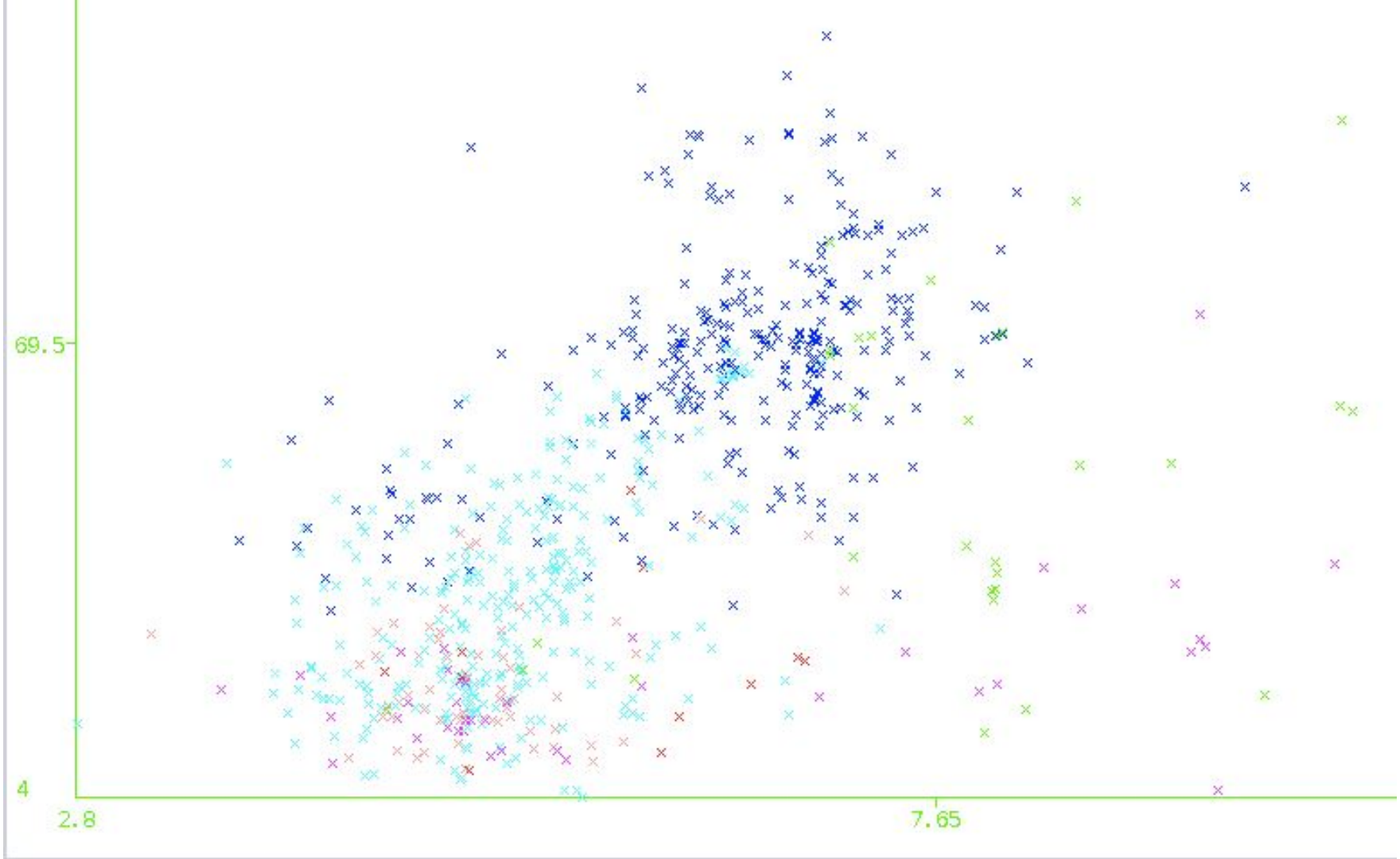
IBU	Belgian	Non-Belgian
Over 30		
Under 30		

←----------objective?

- Combine data from other methods in a chain effect to maximize predictability
- Determine the role of region and state in improving classifier accuracy
- Avoid sequential dependence of the decision tree
- Work well with non-well-separated data

Problems and Challenges

- Maintaining objectivity, not rigging equations (strategizing to fit)
- Traceability to style guides can be obscured by my cleaning
- Chicken and the Egg (derived or classified)
- Latent Variables!
- Scope was too big for the data
- Limited data types and values (SRM, OG/FG)
- Feature development into classes


69.5
4
2.8
7.65

Results (about beer [styles and metrics])

- Beer is a very interesting and rewarding dataset
- I don’t “just dislike lagers” (bocks, altbiers)
- Pale ales and IPAs share a sort of axis
- IPAs have a high variability along this axis.
- German precision (purity laws, “soul of beer”)
- Strong Ales have high variability in IBU, but appear above 6% (style guide)
- Belgian Ales have Low IBU, Varied ABV (Belgian Strong 15-30 style guide)

Tools

- Data Cleaning
  - Python
- Data Exploration
  - R
  - Weka
- Writeups
  - R
- Algorithms
  - Weka

Complexity and Additional Measures

Helpful Measurements

- Standard Reference Model (SRM)
  - Color and Visibility
  - IPA vs Black IPA
- Original/Final Gravity (OG/FG)
  - Residual Sugars/Sweetness
  - Inter-Family Clustering with IBU
- Brew Temperature
  - Lager vs Ale
- Yeast Type
  - Style Determination
- Release Date/Season
  - Winter Warmer vs Saison

Interesting Data

- Hop Varieties
- Grain Type/Percentage
- Addition Times/Temperature
- Temporal Data
- Rating

Conclusions (about data mining)

- Maintain a data-driven strategy to ensure objectivity
- Sanitize, Sanitize, Sanitize (clean your data)
- Maintain the **context** of your data
  - Data as “walls”
  - What the data represents
  - What data you don’t have
  - Latent variables and their potential effects
- Naive Bayes is a very powerful, non-sequential, additive algorithm
- Data mining can be against “terms of use” of APIs, even though it’s displayed openly.

References

- BJCP Beer Style Comparison http://www.bjcp.org/cep/BeerStyleComparison.pdf
- BJCP Beer Official Style Guide
- BeerAdvocate Style Guide https://www.beeradvocate.com/beer/style/
- Craftbeer.com beer style guide https://www.craftbeer.com/beer/beer-styles-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/

Cheers!

questions?