CLT Brew

Clustering Charlotte, NC Breweries

Dillon DeArmond
IBM Data Science Professional - Capstone Project
March 18, 2021

1. Introduction

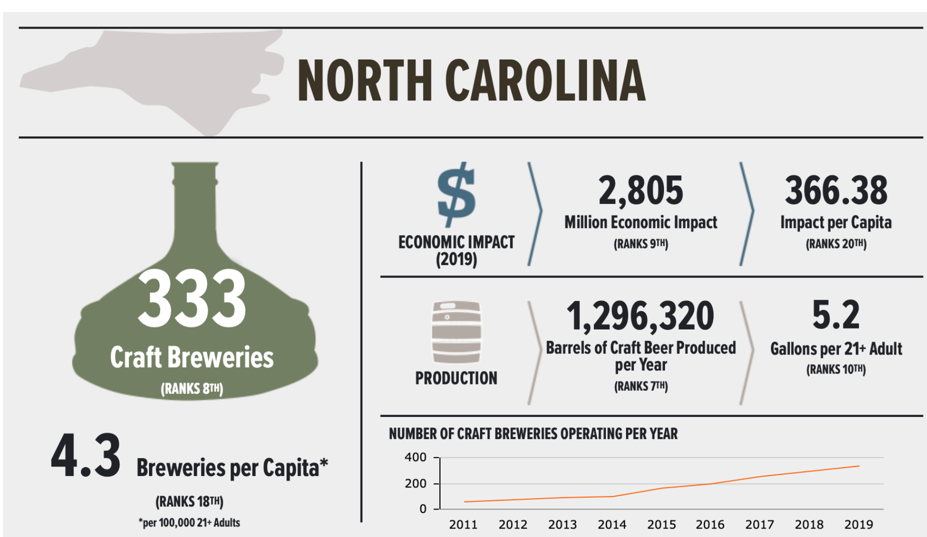
1.1. Background

Craft Brewers have been a growing trend in the United States since 2015. Over the course of just four years, the number of craft brewers increased a whopping 72% going from 4803 to 8275 total craft brewers. The Brewers Association defines an American craft brewer as *"a small and independent brewer"*.

- Small - annual production of 6 million barrels of beer or less (approximately 3 percent of U.S. annual sales)
- Independent - less than 25% of the craft brewery is owned or controlled by a beverage alcohol industry member that is not itself a craft brewer

While being the technical definition of a craft brewer, this is far from truly describing what they are and what they represent. Innovation is at the heart of craft beer and craft brewers. Craft brewers rewrite the script on age-old styles using distinct twists and create new styles that have no point of reference. Craft brewers also have unique, individualistic ways to connect with their customers and tend to be very involved with their local communities. In fact, the majority of Americans live within 10 miles of a craft brewer.

My home of Charlotte, North Carolina is no exception to this surging trend. North Carolina has seen explosive growth (106%) in the Craft Brewer space since 2015. Expanding from 161 Craft Brewers in 2015 to 333 in 2019.



1.2. Problem

With the increasing number of craft breweries in the Charlotte area, it has become harder to distinguish them apart without online research, word of mouth, or visiting them first hand. This project aims to collect and categorize sets of Charlotte craft breweries into groups that are “sufficiently similar” or “close” to one another.

1.3. Interest

Obviously, Charlotte craft beer lovers would be very interested in knowing which breweries are similar in nature to the ones they already like or to discover new spots to visit. Charlotte craft brewers would also be interested in this market analysis that may help them identify potential differentiators without going too far beyond the pale (ale!).

2. Data acquisition and cleaning

2.1. Data sources

I first needed to identify which breweries to use for my analysis. To accomplish this, I leveraged Foursquare’s ‘Places API’. This API allows you to search for and return a list of ‘venues’ around a given location. By passing in Charlotte’s geo-coordinates (longitude/latitude) and filtering the request by FourSquare’s Brewery Category ID, I was able to obtain a list of the Charlotte craft breweries. This list contained the city/state, city latitude/longitude, brewery name, brewery longitude, brewery latitude and venue category.

```
CLT_Brew.head(10)
```

	City	City Latitude	City Longitude	Brewery	Brewery Latitude	Brewery Longitude	Venue Category
0	Charlotte, North Carolina	35.2271	-80.8431	Toucan Louie's Gold District	35.222789	-80.857390	Brewery
1	Charlotte, North Carolina	35.2271	-80.8431	Heist Brewery and Barrel Arts	35.249877	-80.833097	Brewery
2	Charlotte, North Carolina	35.2271	-80.8431	Salty Parrot Brewing Company	35.222871	-80.857052	Brewery
3	Charlotte, North Carolina	35.2271	-80.8431	Pilot Brewing	35.221104	-80.815449	Brewery
4	Charlotte, North Carolina	35.2271	-80.8431	Protagonist	35.186377	-80.881213	Brewery
5	Charlotte, North Carolina	35.2271	-80.8431	Devil's Logic Brewing	35.213875	-80.829262	Brewery
6	Charlotte, North Carolina	35.2271	-80.8431	Fonta Flora Brewery - Optimist Hall	35.234026	-80.827418	Beer Bar
7	Charlotte, North Carolina	35.2271	-80.8431	Petty Thieves Brewing Company	35.238885	-80.835184	Brewery
8	Charlotte, North Carolina	35.2271	-80.8431	Edge City Brewery	35.177465	-80.756283	Brewery
9	Charlotte, North Carolina	35.2271	-80.8431	Summit Seltzery	35.234245	-80.873646	Brewery

Next, I had to figure out what to use for getting additional details about each brewery. To track this information down, I ventured over to “Untappd.com”. Untappd is a geo-social networking service and mobile app which provides a platform for users to rate the beer they are consuming and review brewery beer lists. Unfortunately, Untappd’s API is not free to use and users must submit a request to be approved.

As a workaround, I scraped the Untappd website to pull relevant beer data for each Charlotte craft brewery. This resulted in the following beer data attributes for each brewery: Beer name, Beer Style, Beer Alcohol By Volume (ABV), Beer International Bitterness Unit (IBU), Beer Rating, Beer Total Ratings, and Beer Added Date.

Table 2.1 - Beer Table

Data Attribute	**Definition**	**Example**
Brewery	The name of the brewery	Armored Cow Brewing Co.
Beer Name	The name of the beer	Sleepless in Seattle
Beer Style	The style type of the beer	Stout - Imperial / Double
Beer ABV	Alcohol by volume, is the standard measurement used to access the strength of a beer	12%
Beer IBU	International bittering unit, measures the bitterness levels of a beer	50
Beer Rating	Calculated by a Pure Average of the Beer’s raring. Users are allowed to rate a beer from .25 to 5 stars. = Sum of Beer Ratings / Number of Beer Ratings Note: A beer needs at least 10 ratings before a Beer Rating will show on Untappd	4.11
Beer Total Ratings	The total number of ratings that a beer as	359
Beer Added Date	The month, day, & year in which the beer was added to Untappd	01/07/20

2.2. Data cleaning

Brewery and location data obtained from the Foursquare API returned some unexpected results that required cleanup. I utilized a local Charlotte blog/news website, Axios Charlotte, and Untappd to analyze the results and bucket them into one of the following categories:

1. Valid Breweries - breweries that are currently in business and listed on Untappd
2. Duplicate Breweries - breweries that have multiple or satellite locations
3. Invalid Breweries - breweries that did not exist, are no longer in business, or ones that did not serve beer.

Table 2.2- Brewery Cleaning

Category	**Examples**	**Decision**
Valid Breweries	Town Brewing Company, Birdsong Brewing Co.	Kept
Duplicate Breweries	Legion Brewing Southpark, Free Range Brewing Camp North End	Removed
Invalid Breweries	Doug's Desk, Bold Missy Brewery, Goodroad CiderWorks	Removed

Brewery Beer details that were scraped were combined into a single table. A common pitfall when web scraping is that the values can be returned with many unwanted characters (e.g. whitespace, symbols, blank values).

```
beer_df_raw.head(10)
```

	Brewery	Beer Name	Beer Style	Beer ABV	Beer IBU	Beer Rating	Beer Total Ratings	Beer Added Date
0	Armored Cow Brewing Co.	Bang-Bang IPA	IPA - New England	\n7.3% ABV	\n40 IBU	(4.02)	\n714 Ratings	\nAdded 05/04/20
1	Armored Cow Brewing Co.	Hell Yeah!	Gluten-Free	\n6.5% ABV	\nN/A IBU	(3.66)	\n392 Ratings	\nAdded 05/22/19
2	Armored Cow Brewing Co.	Bitchin Betty	Brown Ale - Other	\n5.4% ABV	\nN/A IBU	(3.68)	\n350 Ratings	\nAdded 05/24/19
3	Armored Cow Brewing Co.	Sleepless In Seattle	Stout - Imperial / Double	\n12% ABV	\n50 IBU	(4.11)	\n359 Ratings	\nAdded 01/07/20
4	Armored Cow Brewing Co.	Strawberry 6&20	Blonde Ale	\n4.5% ABV	\nN/A IBU	(3.67)	\n345 Ratings	\nAdded 05/17/19
5	Armored Cow Brewing Co.	Reveille Stout	Stout - Oatmeal	\n5.6% ABV	\nN/A IBU	(3.77)	\n279 Ratings	\nAdded 08/03/19
6	Armored Cow Brewing Co.	Shake That Axe	Lager - American Light	\n4.8% ABV	\nN/A IBU	(3.65)	\n240 Ratings	\nAdded 09/27/19
7	Armored Cow Brewing Co.	Cherry Bomb	Sour - Berliner Weisse	\n4.5% ABV	\nN/A IBU	(3.72)	\n244 Ratings	\nAdded 06/08/19
8	Armored Cow Brewing Co.	Fun51weizen	Hefeweizen	\n5.2% ABV	\nN/A IBU	(3.72)	\n240 Ratings	\nAdded 06/01/19
9	Armored Cow Brewing Co.	The 49er	Kölsch	\n4.8% ABV	\nN/A IBU	(3.57)	\n219 Ratings	\nAdded 05/17/19

First, I used regular expression to replace unwanted characters in the Beer ABV, Beer IBU, Beer Rating, Beer Total Ratings, and Beer Added Date columns. I also trimmed leading/trailing whitespace and replaced blank values with 'NaN' across all columns.

Table 2.2.1 - Beer Table Cleaning

Data Attribute	**Raw Data**	**Clean Data**
Beer ABV	\n7.3% ABV	7.3
Beer IBU	\n40 IBU	40
Beer Rating	(4.02)	4.02
Beer Total Ratings	\n348 Ratings	348
Beer Added Date	\nAdded 05/04/20	05/04/20

I also noticed there were beers which did not have an ABV or IBU value. To fix this, I filled those missing values using the average ABV or IBU of the other beers of that style contained in the table. After this, there were still missing values for ABV and IBU. To fill the remaining missing values, I replaced them with the average ABV or IBU across all of the beers in the table.

Finally, I dropped beers which did not have any ratings at all. Untappd beer ratings will not show up until a beer has at least 10 ratings.

4. Methodology

In order to solve for the problem, I decided to use clustering analysis to group the various breweries. Clustering, in general, is a set of techniques used to partition data into groups, or 'clusters'. Clusters are loosely defined as groups of data objects that are more similar to other objects in their cluster than they are to data objects in other clusters. It can be achieved by various algorithms that differ significantly in their understanding of what constitutes a cluster and how to efficiently find them. Selecting an appropriate clustering algorithm for a dataset is often difficult due to the number of choices available. Some important factors that affect this decision include: the characteristics of the clusters, the features of the dataset, the number of outliers, and the number of data objects.

For this analysis, I used one of the simplest and most popular unsupervised machine learning algorithms called k-means clustering. The objective of k-means clustering is to group similar data points together and discover underlying patterns. To achieve this, k-means looks to find a fixed number of k values (clusters) in a data set.

In the following sections, I will walk through my process of exploratory data analysis, feature selection, and applying k-means clustering on the breweries.

4.1. Exploratory Data Analysis

4.1.1. High Level Stats

Brewery with the most Beers	Catawba Brewery (443)
Top Beer Style	IPA - American (488)
Average Beer ABV	6.33%
Average Beer IBU	34
Average Beer Rating	3.8
Average Total Ratings	454
Max ABV	15%
Min ABV	0.1%

4.1.2. Top Beer Names

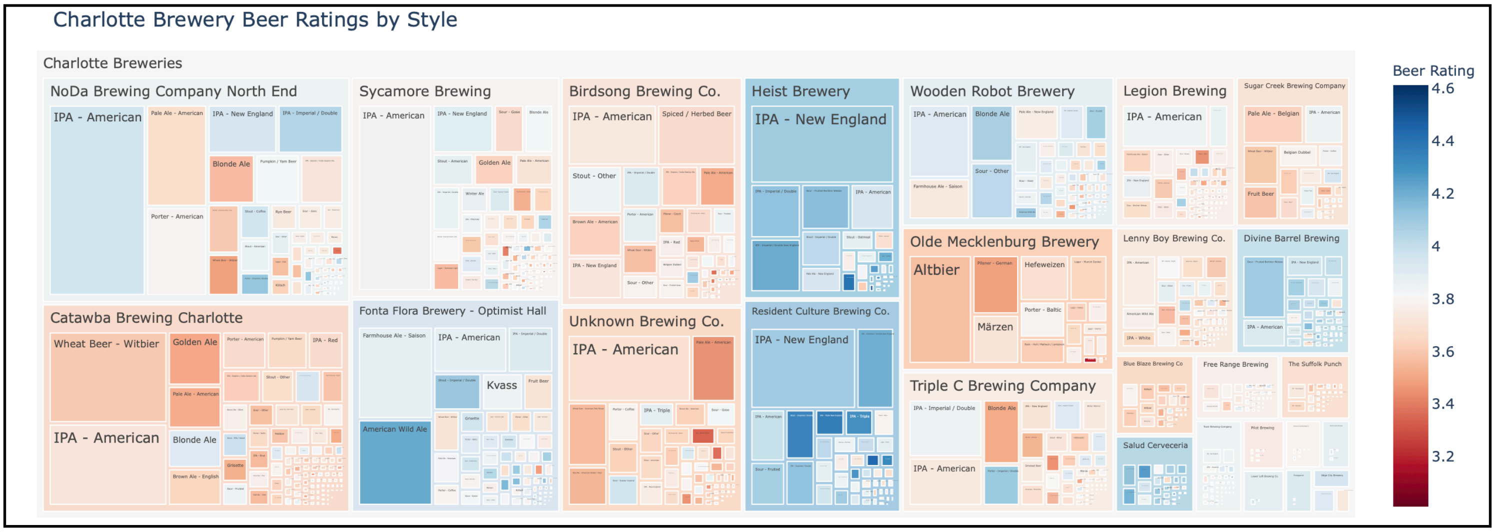
It was interesting to discover that there were two beer names with the same variation that showed up across multiple breweries. Those two beer names were 'Court Shoes Only' (19) and 'Black is Beautiful' (17). With Court Shoes Only having IPA beer style variations and Black is Beautiful having Stout, Pilsner, Porter, and IPA beer style variations.

Court Shoes Only was a Charlotte-area collaborative fundraising beer project. This exclusive release coincided with what would've been the start of the 2021 Queen City Brewers Festivals, which was canceled due to the pandemic. This fundraiser benefits ACEing Autism, a local Charlotte non-profit that connects kids through tennis.

Black is Beautiful was a global-wide collaborative effort that was started by Weathered Souls Brewing Co. in San Antonio, TX. Their mission is to bring awareness to the injustices that many People of Color face daily. Over 1200 breweries across 50 states and 22 countries supported this initiative with proceeds from select retailers being donated to the 'Know Your Rights" campaign.

4.1.3. Charlotte Brewery Beer Ratings by Style

We can see that Resident Culture Brewing Co. and Heist Brewery beers have higher than average beer ratings across all of their beer styles.



4.2. Feature Selection

After conducting exploratory analysis, there were 4170 samples and 7 features in the Brewery Beer Details table.

```
beer_df_clean.head(10)
```

	Brewery	Beer Name	Beer Style	Beer ABV	Beer IBU	Beer Rating	Beer Total Ratings	Beer Added Date
0	Armored Cow Brewing Co.	Bang-Bang IPA	IPA - New England	7.3	40.000000	4.02	714	2020-05-04
1	Armored Cow Brewing Co.	Hell Yeah!	Gluten-Free	6.5	27.500000	3.66	392	2019-05-22
2	Armored Cow Brewing Co.	Bitchin Betty	Brown Ale - Other	5.4	19.000000	3.68	350	2019-05-24
3	Armored Cow Brewing Co.	Sleepless In Seattle	Stout - Imperial / Double	12.0	50.000000	4.11	359	2020-01-07
4	Armored Cow Brewing Co.	Strawberry 6&20	Blonde Ale	4.5	20.976190	3.67	345	2019-05-17
5	Armored Cow Brewing Co.	Reveille Stout	Stout - Oatmeal	5.6	30.846154	3.77	279	2019-08-03
6	Armored Cow Brewing Co.	Shake That Axe	Lager - American Light	4.8	15.687500	3.65	240	2019-09-27
7	Armored Cow Brewing Co.	Cherry Bomb	Sour - Berliner Weisse	4.5	9.000000	3.72	244	2019-06-08
8	Armored Cow Brewing Co.	Fun51weizen	Hefeweizen	5.2	16.411765	3.72	240	2019-06-01
9	Armored Cow Brewing Co.	The 49er	Kölsch	4.8	23.222222	3.57	219	2019-05-17

Upon examining the meaning of each feature, it was clear that some would not be relevant for my analysis. The below table shows a breakdown of each feature that was dropped and reasoning for doing so.

Table 4.2 - Dropped Features

Dropped Feature	**Reason for dropping Feature**
Beer Name	Not relevant for clustering purposes
Beer ABV	Assumed that on average, the Beer's ABV would be in a consistent range for each unique beer style
Beer IBU	Assumed that on average, the Beer's IBU would be in a consistent range for each unique beer style
Beer Rating	Beer Rating could be misleading if a beer does not have many ratings and skew the rating higher or lower
Beer Total Ratings	Total Ratings for a specific beer could be skewed by a Brewery which has been around for longer than another
Beer Added Date	Not relevant for clustering purposes

I ultimately decided to focus on Beer Style as the main feature for my analysis. I believe a brewery can be characterized by what type of beer (beer styles) they brew.

4.3. Data Preprocessing

The first step in preparing the dataset is to format it in a way that can be fed into the k-means clustering algorithm. To prepare this dataset with Beer Style as the main feature, I decided to calculate the average frequency of occurrence for each beer style across every brewery's beer list.

Since Beer Style is categorical and the k-means algorithm can only work with numerical data, I needed to perform 'one-hot encoding' for each beer's beer style. This transformed my dataset such that each Beer Style was now represented as its own column and for each beer a new binary variable (0 or 1) was added to denote what the style was.

```
beer_styles_onehot.head()
```

	Brewery	Altbier	American Wild Ale	Barleywine - American	Barleywine - English	Barleywine - Other	Belgian Blonde	Belgian Dubbel	Belgian Quadrupel	Belgian Strong Dark Ale	Belgian Strong Golden Ale	Belgian Tripel	Bière de Champagne / Bière Brut	Blonde Ale	Bock - Doppelbock	Bock - Eisbock (Traditional)
0	Armored Cow Brewing Co.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Armored Cow Brewing Co.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Armored Cow Brewing Co.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Armored Cow Brewing Co.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Armored Cow Brewing Co.	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0

Next, I grouped the rows by Brewery with the average frequency of occurrence for each beer style.

```
beer_styles_freq.head()
```

	Brewery	Altbier	American Wild Ale	Barleywine - American	Barleywine - English	Barleywine - Other	Belgian Blonde	Belgian Dubbel	Belgian Quadrupel	Belgian Strong Dark Ale	Belgian Strong Golden Ale	Belgian Tripel	Bière de Champagne / Bière Brut	Blon
0	Armored Cow Brewing Co.	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.1020
1	Birdsong Brewing Co.	0.003876	0.000000	0.003876	0.003876	0.000000	0.019380	0.023256	0.0	0.003876	0.000000	0.019380	0.0	0.0271
2	Blue Blaze Brewing Co	0.016129	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.1129
3	Catawba Brewing Charlotte	0.000000	0.006865	0.002288	0.002288	0.002288	0.006865	0.006865	0.0	0.006865	0.004577	0.006865	0.0	0.0251
4	Devil's Logic Brewing	0.025641	0.000000	0.000000	0.000000	0.000000	0.000000	0.025641	0.0	0.000000	0.000000	0.000000	0.0	0.0256

I then used it to derive the 10 most common beer styles for each Brewery in order to examine the cluster results later.

```
brewery_beer_style_sorted.head()
```

	Brewery	1st Most Common Beer Style	2nd Most Common Beer Style	3rd Most Common Beer Style	4th Most Common Beer Style	5th Most Common Beer Style	6th Most Common Beer Style	7th Most Common Beer Style	8th Most Common Beer Style	9th Most Common Beer Style	10th Most Common Beer Style
0	Armored Cow Brewing Co.	IPA - New England	Gluten-Free	Blonde Ale	IPA - American	Lager - American Light	Bock - Doppelbock	Red Ale - American Amber / Red	Sour - Berliner Weisse	Cider - Sweet	Spiced / Herbed Beer
1	Birdsong Brewing Co.	IPA - American	Farmhouse Ale - Saison	Sour - Other	Pale Ale - American	Stout - Other	Sour - Fruited	Sour - Gose	IPA - Imperial / Double	Stout - American	Blonde Ale
2	Blue Blaze Brewing Co	IPA - American	Blonde Ale	Stout - Milk / Sweet	Porter - Other	IPA - New England	Fruit Beer	Scotch Ale / Wee Heavy	Hefeweizen	Bock - Doppelbock	Sour - Gose
3	Catawba Brewing Charlotte	IPA - American	Sour - Fruited	Farmhouse Ale - Saison	Sour - Other	Sour - Gose	IPA - New England	Pale Ale - American	Blonde Ale	IPA - Imperial / Double	IPA - Session / India Session Ale
4	Devil's Logic Brewing	IPA - New England	IPA - American	Sour - Fruited Gose	Extra Special / Strong Bitter	Pilsner - German	Pale Ale - American	Stout - Other	Altbier	Kellerbier / Zwickelbier	Festbier

Finally, I needed to convert our features to a numpy array/vector so that it can be fed it into the k-means clustering algorithm.

4.4. K-means clustering

One key thing to note when we observe our data is that it is high dimensional. Dimensionality in statistics refers to how many attributes (aka features) a dataset has. High dimensional data means that the number of dimensions are staggeringly high. So high in fact that calculations become extremely difficult. Our dataset contains 28 observations (breweries) and 183 features (average frequency of beer style occurrence).

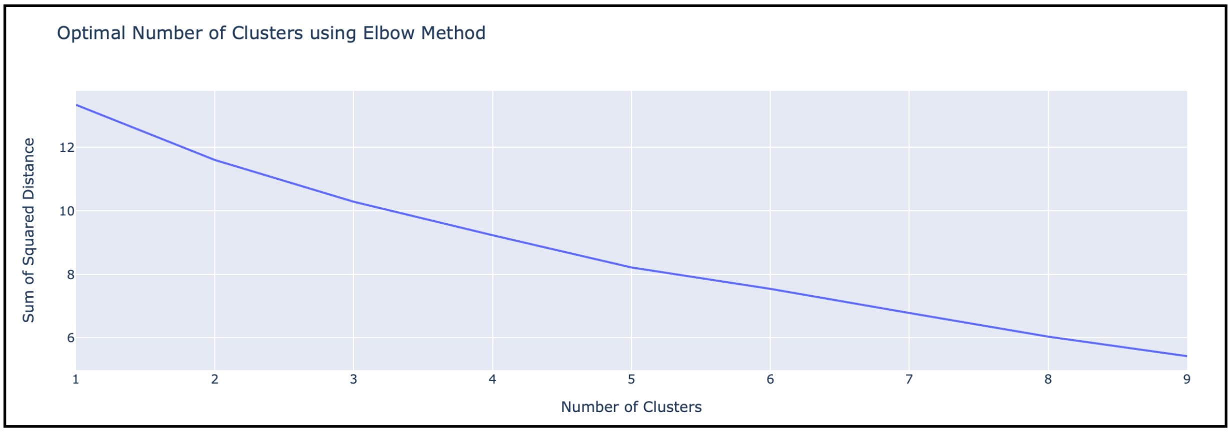
Since k-means typically uses Euclidean distance to calculate the distance, it does not work well with high dimensional datasets due to the curse of dimensionality. This curse, in part, states that Euclidean distances at high dimensionality have very little meaning since they are often very close together. The more dimensions you add to the dataset, the more difficult it becomes to predict certain quantities. You would think that more is better. However, when it comes to adding features, the opposite is true. Each added feature results in an exponential *decrease* in predictive power.

A solution would be to use the Cosine distance which works better in the high dimensional space. Since Cosine distance and Euclidean distance are connected linearly for normalized vectors, we can simply normalize our data. I leveraged the built-in 'preprocessing' module from scikit-learn to take care of this.

With our dataset now formatted and normalized, we are ready to move forward with applying the k-means clustering algorithm. Before doing so, the user must specify the k values (where k = number of clusters). Most of the time this can be done through trial and error, but there are a few techniques that can be leveraged to determine the optimal value of k.

I leveraged the Elbow Method to identify the optimal number of clusters. The Elbow Method consists of plotting the explained variance as a function of the number of clusters, and picking the 'elbow of the curve' as the number of clusters to use.

For each k value ranging from 1-10, I initialized k-means and used the inertia attribute to identify the sum of squared distances of samples to the nearest cluster center. As k increases, the sum of squared distance tends to zero.



In the plot above, the elbow looks to occur at k=5, indicating the optimal k for this dataset.

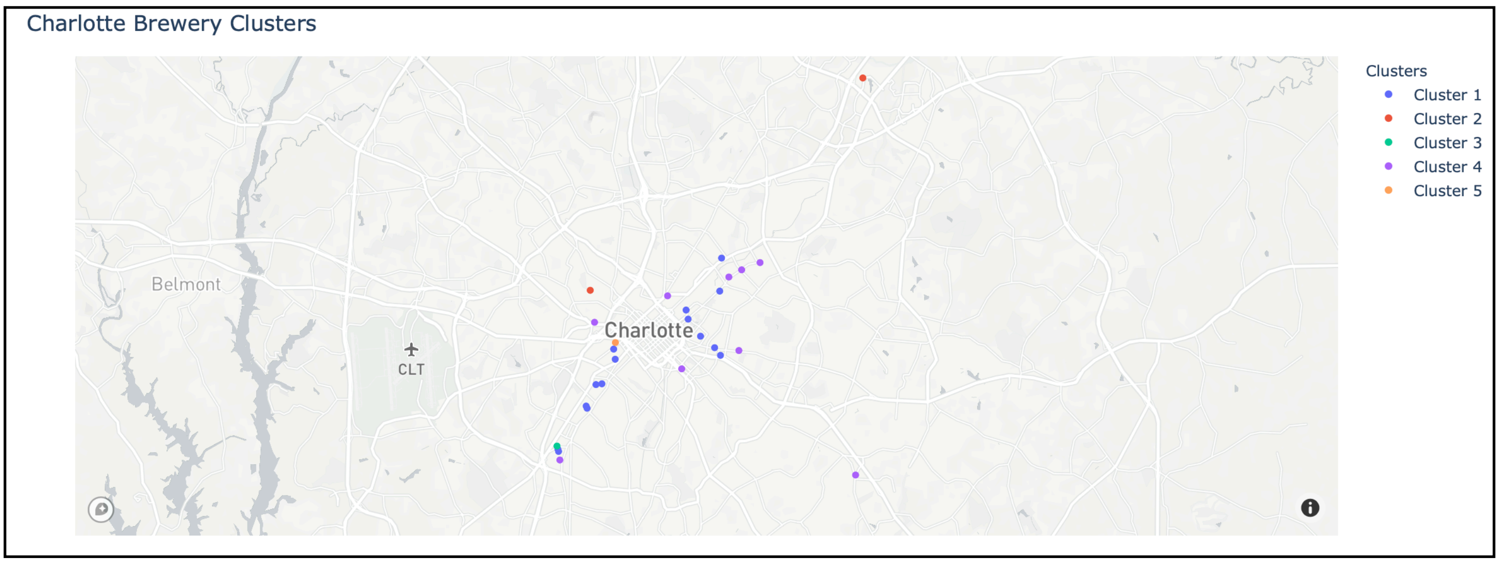
Lastly, we run the k-means clustering algorithm specifying k=5 to give us 5 distinct clusters of breweries.

5. Results

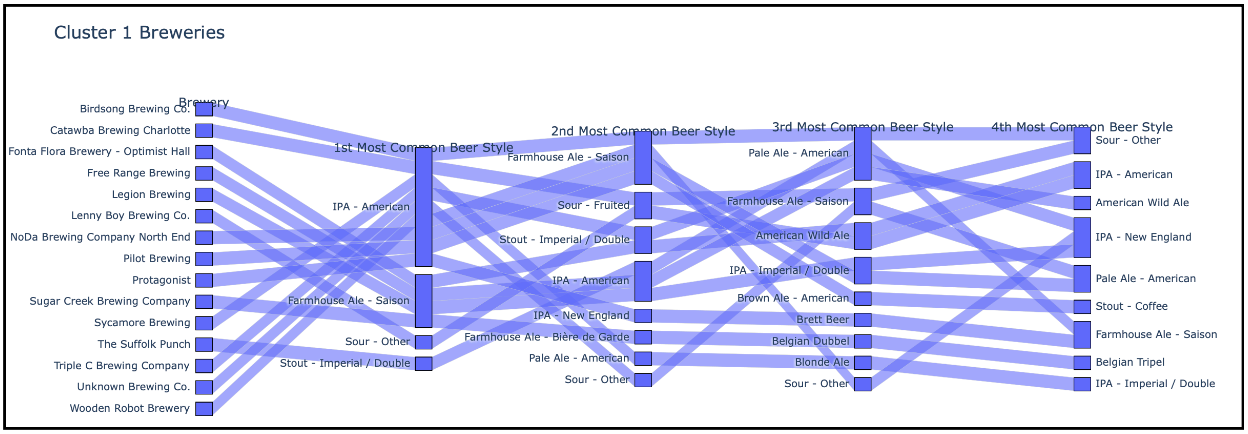
The results from the k-means clustering show that we can categorize the breweries into 5 clusters based on the average frequency of occurrence for each Beer Style.

- Cluster 1: Breweries (15) with mostly IPA - American & Farmhouse Ale - Saison beer styles
- Cluster 2: Breweries (2) with mostly Blonde Ale beer style
- Cluster 3: One brewery with mostly Porter - Baltic & Altbier beer styles
- Cluster 4: Breweries (9) with mostly IPA - New England & IPA - American beer styles
- Cluster 5: One brewery with mostly Other & Porter - American beer styles

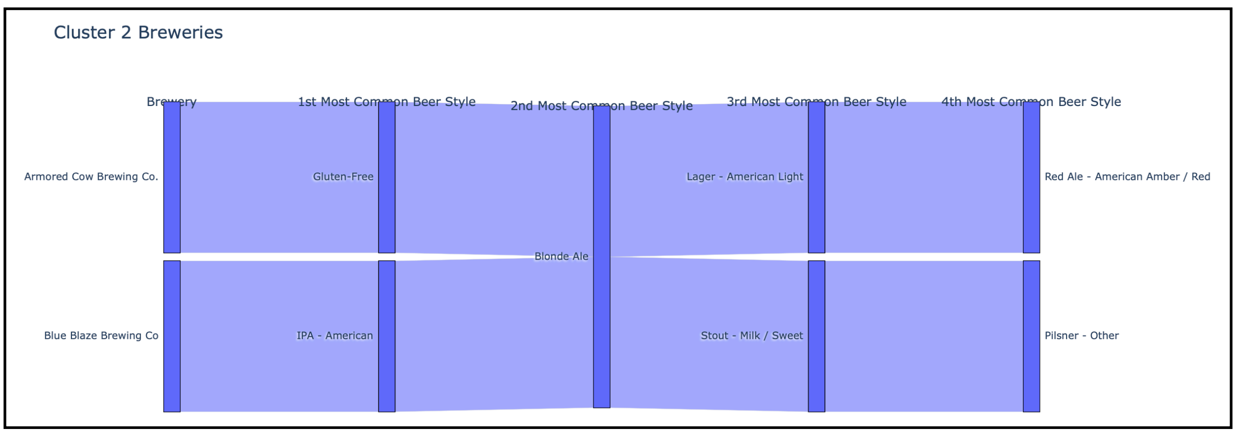
The results of the clustering are visualized in the map below.



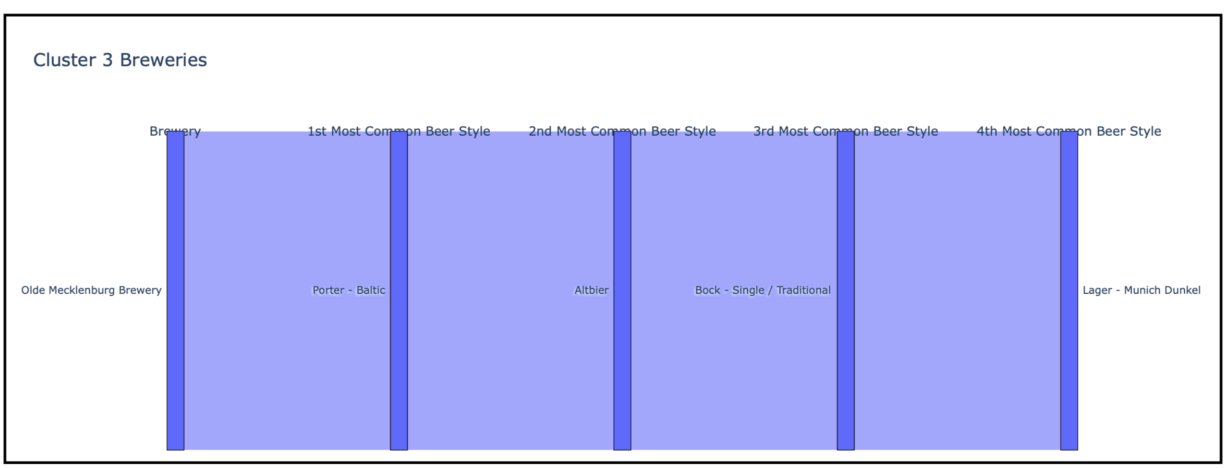
5.1. Cluster 1



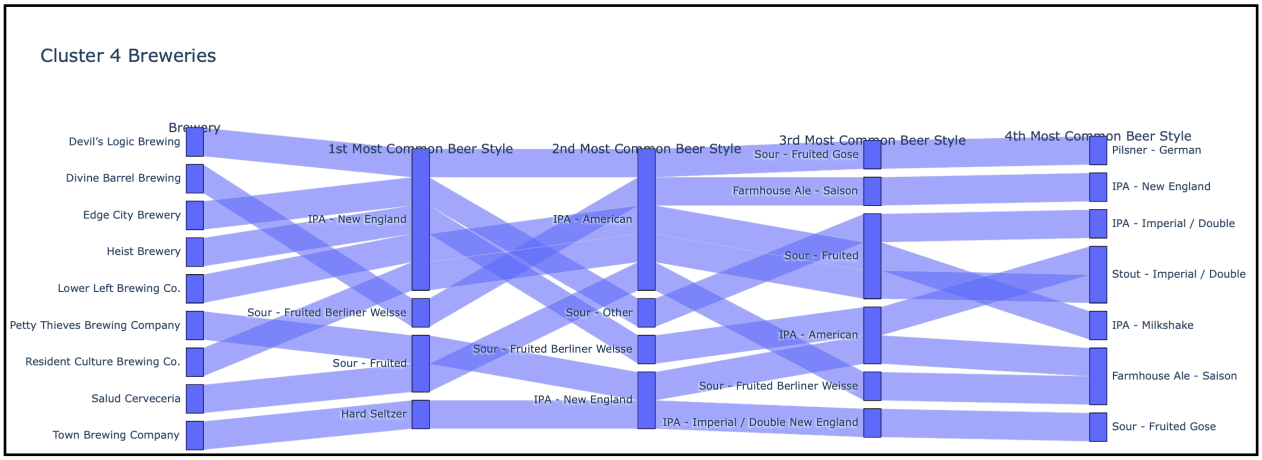
5.2. Cluster 2



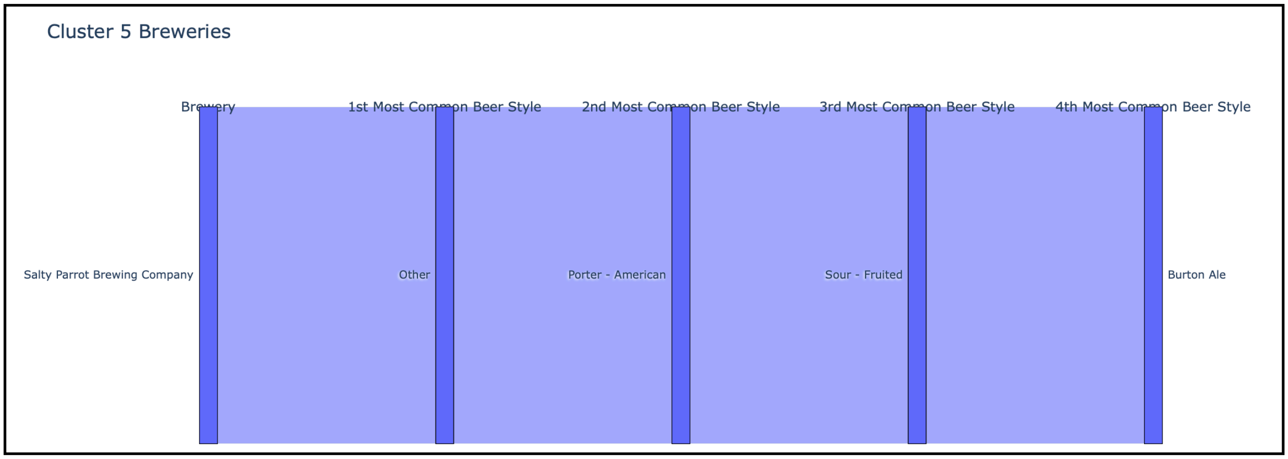
5.3. Cluster 3



5.4. Cluster 4



5.5. Cluster 5



6. Discussion

6.1. Observations

As observations noted from the cluster sections, Cluster 1 is the biggest cluster with 15 of the 28 breweries. I also noticed that the "IPA - American" was one of the most common beer styles across 2 of the 5 clusters (Clusters 1 and 4).

Looking at clusters 3 and 5, we can see that these clusters have only one brewery in each. This is because of the unique beer styles in each of the breweries, hence they couldn't be clustered into similar breweries.

6.2. Recommendations

The aim of this project is to help Charlotte craft beer lovers discover new breweries. For example, if a person is looking to check out a brewery with lots of IPAs we can see that Cluster 4 has IPAs as the top two most common beer styles. If a person is looking for breweries that are very unique when compared to other Charlotte breweries, Clusters 3 and 5 would be good to consider since each of these clusters only have one brewery in them.

7. Conclusion/Future Enhancements

In this project, I only take into consideration one factor: the average frequency of occurrence of beer styles across the breweries. There are in fact many factors that can be used when deciding which brewery to visit. Such as: location/area, atmosphere/ambiance, service, and many more. However, it's tough to quantify a lot of these factors and it would require extensive research.

I also noticed limitations in terms of the beer selections and beer ratings at each brewery. One such limitation is the timing of when this analysis was performed. As breweries are constantly adding new beers (and beer styles) to their menu and users of Untappd are constantly rating beers. This introduces the potential exclusion of certain beer styles that have been added or rated recently.

A second limitation is around what beers each brewery actually has for sale at any given time. For example, in my analysis Birdsong Brewing Co. had 263 beers with ratings listed on Untappd. But when looking at their website, they currently have 13 beers on tap (6- Year Round, 5- Seasonal/Limited Release, & 2- Small Batch) and 13 beers for sale with their 'Beer to go' service in cans and growlers.

Future enhancements for this project could incorporate one or more of these factors and take into account some these limitations mentioned above.

While this single factor is not the end-all be-all for deciding which brewery to visit, we have demonstrated that the results achieved through machine learning can be another resource to help Charlotte craft beer lovers and Charlotte Craft Brewers in their decision making process.

8. References

1. https://www.brewersassociation.org/statistics-and-data/craft-brewer-definition/
2. https://www.brewersassociation.org/statistics-and-data/national-beer-stats/
3. https://www.brewersassociation.org/statistics-and-data/state-craft-beer-stats/?state=NC
4. https://developer.foursquare.com/docs/places-api/endpoints/
5. https://charlotte.axios.com/31429/breweries-in-charlotte/
6. untapped.com