Brewery Site Evaluation

for Montgomery County, PA

Erik Costello
(not a Data Scientist)

July 24, 2019

Introduction:

I like beer - I like drinking it, talking about it, and making it. In an alternate universe, I would start a business as a brewery somewhere near where I live, but in this universe, I will attempt to identify good locations for a new brewery in the philadelphia suburbs, centered on Montgomery county so my evil-universe twin can follow his dream.

I will identify existing breweries and data on their rating as an indication of public perception of the quality of the brewery but not necessarily of the beer. Using this quality rating, I will then segment the breweries into groups based on their ratings and gather venues in the locality so I can do a clustering analysis to determine if there are commonalities among the rating categories.

Finally, If there are common traits for the better breweries, I will attempt to rate other locations in the target area that fit the positive while avoiding the negative criteria. The target areas will be based on geographic distance from other breweries. These locations will be considered as opportunities for a new brew-pub that can be ranked based on the data found.

Data:

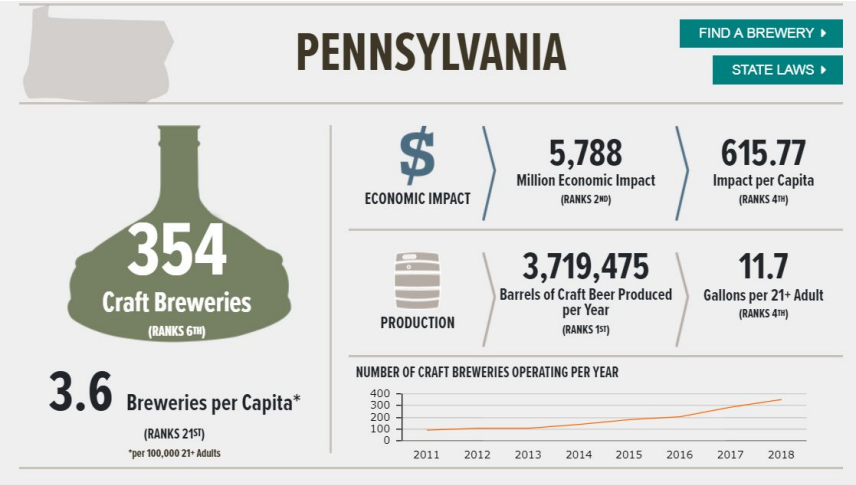
I initially decided I could use the below sources as noted:

- Google maps and search
    - To set a baseline, I used google search within maps to identify breweries. This was helpful to validate Foursquare.
- Location data from Foursquare
    - **Brewery venues to be identified based on search**. Initially I found that the ‘query’ based results from Foursquare were very poor and missed many of the breweries that I expected to find. In addition the venues that were returned often were bars, supply stores or completely missed the category. After reading through the API documentation, I found the category search which had a value for ‘Brewery’. This search resulted in more comprehensive results with few false positives. However, I found that it was pervasive for people to add their home locations or a firepit in their backyard into Foursquare data as a brewery such as 'Nitt Witt Brewery', ‘Basement Bar’, or 'Brass Ass Brewing' for example. Most of these locations were identifiable by missing data fields which I could drop with a bulk drop statement. Other locations had to be validated with Google maps.
    - **Venue ratings and number of ratings**. I planned to use the ratings and the number of votes to segment the breweries into tiers. Unfortunately many

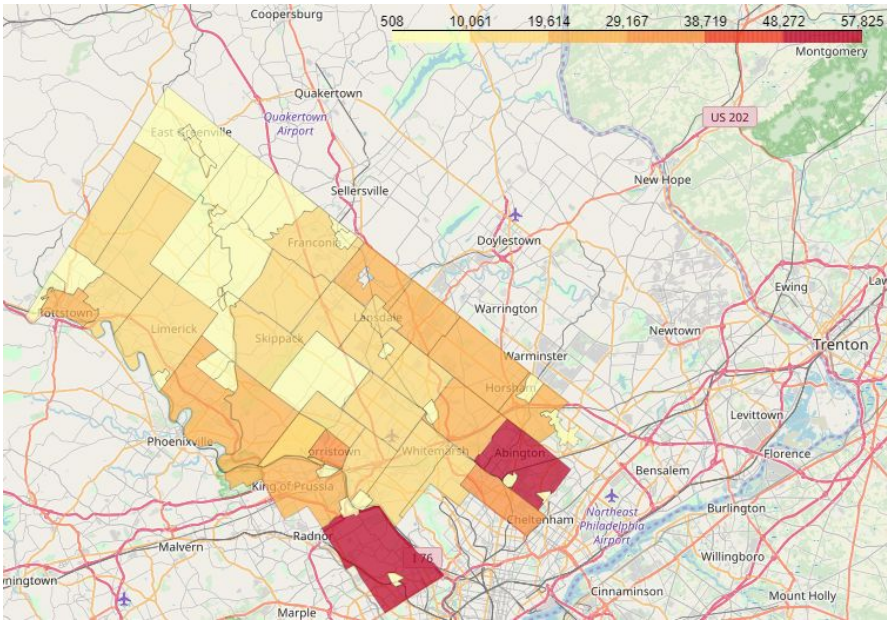
breweries didn't have ratings but did have 'Likes'. I experimented in creating a formula which calculated the difference between the number of ratings and the number of likes to establish people who rated the brewery but didn't bother to 'Like' it, and then normalize that like count versus 'not-like' but the sample data from Foursquare was too few to make this effective so I limited the breweries for the final analysis to those that had ratings. Also due to limits on data retrieval from Foursquare, I was limited in the amount of detail I could gather on breweries and venues.
    - Trending venues near the target breweries which will be used to cluster the different ratings tiers.
    - Venues near the potential new sites for evaluation against the clustered results above.
- Pennsylvania spatial data - https://www.pasda.psu.edu/uci/DataSummary.aspx?dataset=41 Geojson data to segment townships in the target county of Pennsylvania for choropleth plotting. Although this data was available for download, I had to manually extract the county information I needed for the Choropleth maps.
- The Penn State Data Center Municipal level data here, to determine the population of Montgomery county and townships. This data was used to identify potential customer base relative to population, but it was from 2010. Although this data was old, it was fine for showing relative populations for given townships in the target area. I didn't need exact population numbers.
- Nominatim and Geolocator for translation of addresses into latitude and longitude coordinates.
- Kaggle.com - I identified several data sets that I evaluated for brewery and beer ratings. I thought that I could also determine a brewery rating in relation to the beers but this data was poor and required purchase to get meaningful data (such as BeerAdvocate and Untappd datasets). I decided to drop this data from my criteria since the beer quality was not a factor in this assessment.
- www.brewersassociation.org offers limited free data on breweries and beer consumption.

Methodology and Exploratory Analysis

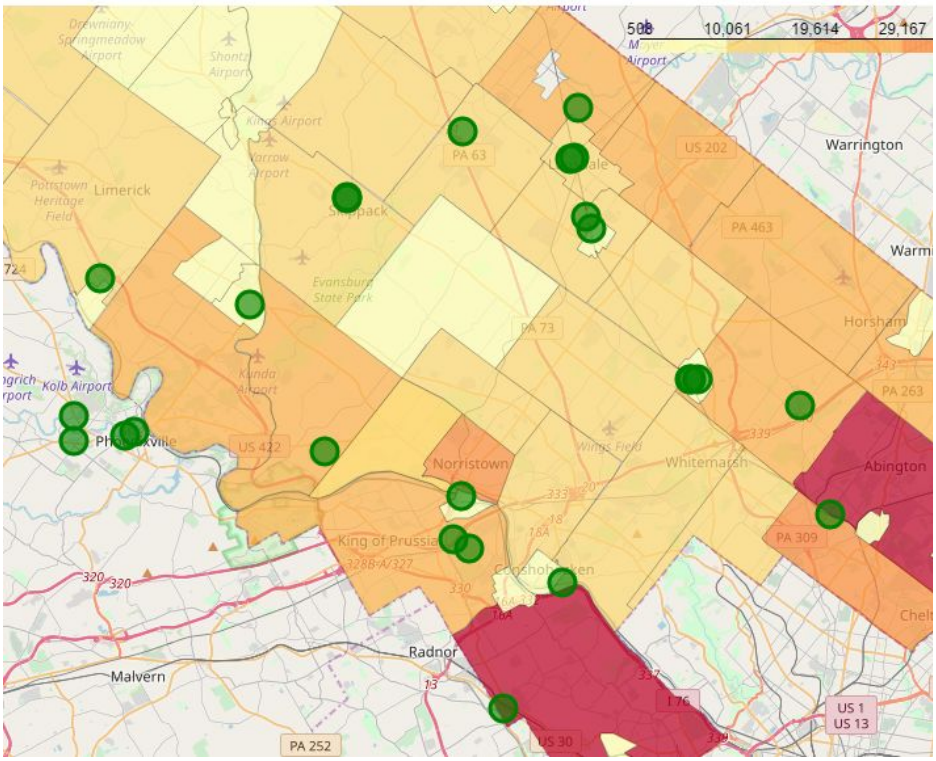
I started my analysis by creating a population map of the target area, which I chose to be the central region of Montgomery County, Pennsylvania. This is one of the more populated counties in the state and sits just above Philadelphia. According to the Brewers Association statistics and data analysis of PA, we rank 21st in Breweries per capita and 4th in rank of gallons consumed per adult.



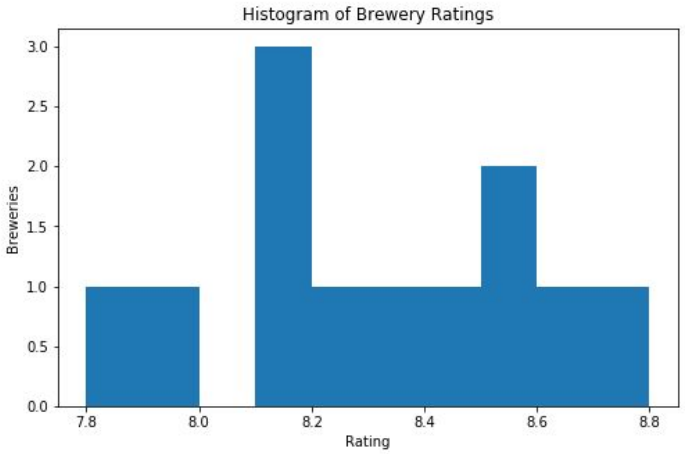
In the below map, we can see that the population is mostly located toward the south-eastern side of the map, which makes sense since that is where Philadelphia is. I would expect that the breweries would fall into the darker orange townships.



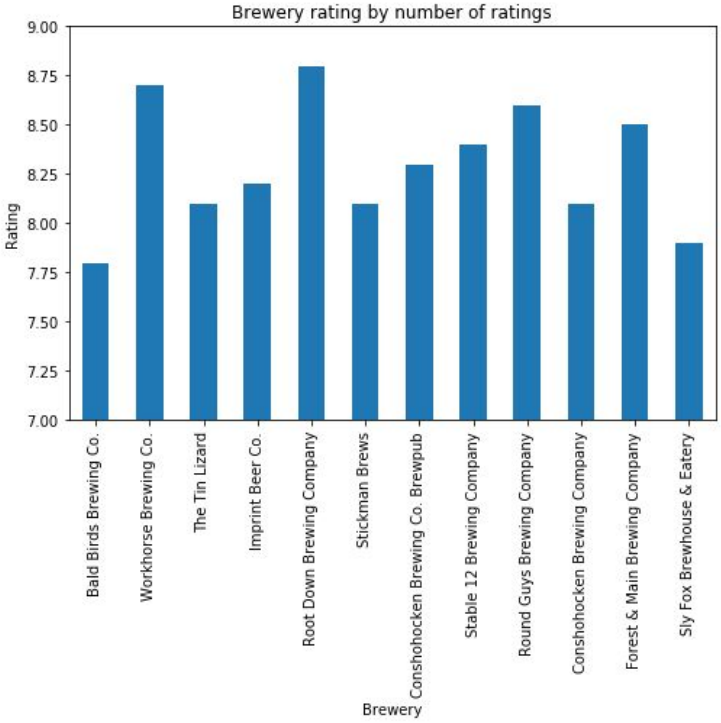
In fact, after getting good results from Foursquare, there were 61 breweries located within a 14 mile range of the centroid of the county. After removing locations with only gps coordinates and no addresses assuming these were not real breweries, removing locations out of the county such as the part of Philadelphia that extends into Montgomery, eliminating the fake breweries and keeping good samples just outside of the county near Phoenixville, I was left with 24 valid sites for evaluation. These 24 breweries did indeed align with my assumption that more breweries would be aligned with population, but not to the extent I expected. Interestingly, there were no breweries in the central townships. This was odd.



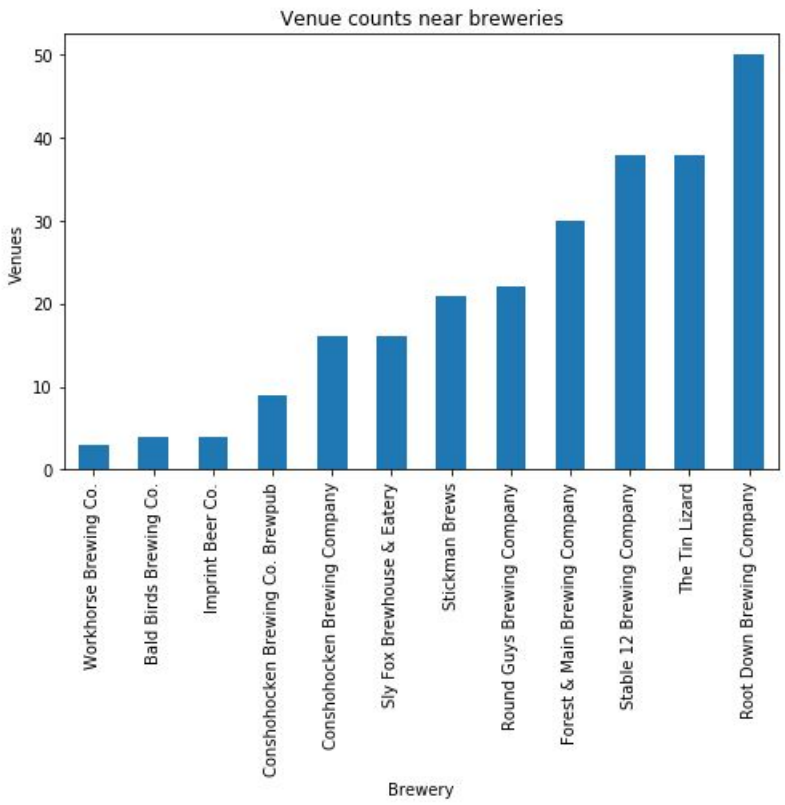
With this new dataset of breweries, I queried the details on the breweries to get ratings, number of ratings and likes and added that to the dataset. As noted in the Data section above, I further reduced the dataset to breweries with ratings and plotted a histogram for their distribution. Unfortunately the small dataset, it is not an ideal gaussian curve but it is indicative of the curve.



It is worth noting that there was no correlation between the rating and the number of ratings. I would have expected the breweries with few ratings to be on the extremes



Next I iterated through each brewery and retrieved the list of venues within 500 meters of the brewery. Breweries ranged from 50 venue locations only 3. I found 251 venues in 91 categories.



With the list of venues grouped by brewery, I one-hot encoded the venues and determined the most common venues by brewery and used K-means clustering to attempt to group the venues into 3 clusters based on High/Medium/Low rating clusters.

Here is the table of those clusters and top 3 venues:

	Cluster Labels	Brewery	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	0	Bald Birds Brewing Co.	Other Great Outdoors	Business Service	Coffee Shop
1	1	Conshohocken Brewing Co. Brewpub	Pizza Place	Gym	Ice Cream Shop
2	1	Conshohocken Brewing Company	Italian Restaurant	Athletics & Sports	Yoga Studio
3	1	Forest & Main Brewing Company	Pizza Place	American Restaurant	Bakery
4	0	Imprint Beer Co.	Pharmacy	General Entertainment	Business Service
5	1	Root Down Brewing Company	American Restaurant	Pub	Pizza Place
6	1	Round Guys Brewing Company	Pizza Place	Ice Cream Shop	Bakery
7	1	Sly Fox Brewhouse & Eatery	Pizza Place	Fast Food Restaurant	Bank
8	1	Stable 12 Brewing Company	American Restaurant	Pub	Pizza Place
9	1	Stickman Brews	Pizza Place	Fast Food Restaurant	Discount Store
10	1	The Tin Lizard	Pizza Place	Indian Restaurant	Café
11	2	Workhorse Brewing Co.	Breakfast Spot	Food Truck	Wine Shop

Not surprisingly, “Pizza Place” was the most common value and was frequent across all the venues, so I excluded that category. When I examine the 3 clusters there was not much correlation to the ratings except the best rated location was in cluster 3 and the worst was in cluster 1 with a mid-to-high rated Imprint brewing.

Results

Given the limited dataset I was able to use, the results showed a correlation that I did not predict but is not that surprising. The criteria for the best and the worst did not overlap.

Worst Rated Breweries Proximity to:

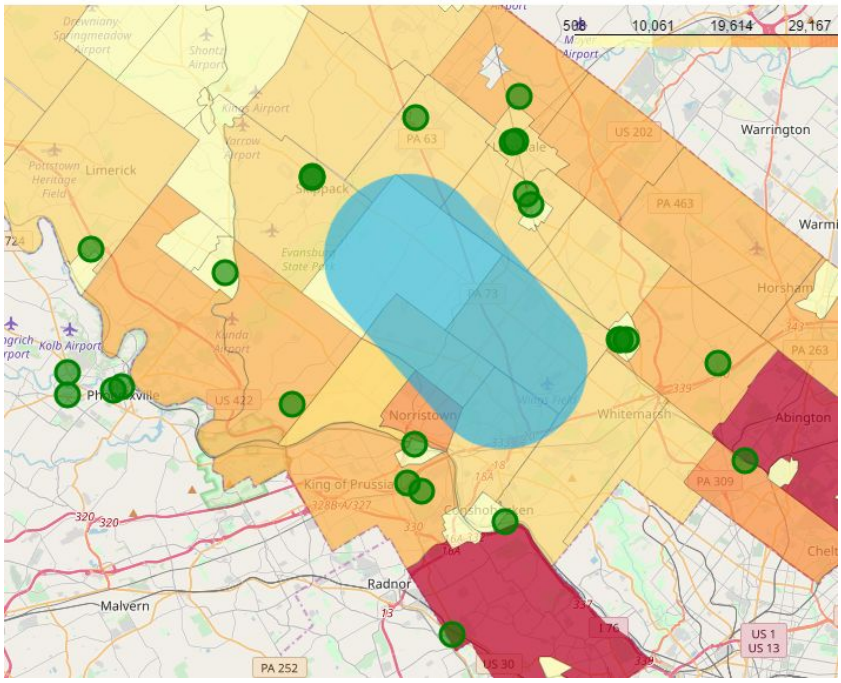
1. Fast Food (Wendy’s, Dairy Queen, Subway, ...)
2. Business Services (Cleaners, corporate offices)

Best Rated Breweries Proximity to:

1. American Restaurants (Black Lab Bistro, Bistro on Bridge, Great American Pub, ...)
2. Pubs (Molly Maguire's, Junction House, The Foodery, PJ Ryan's Pub, ...)

Conclusions

Given the criteria for good and bad above, I marked the large central area in Montgomery County which is empty of any breweries. I noted this target area with a blue oval below.

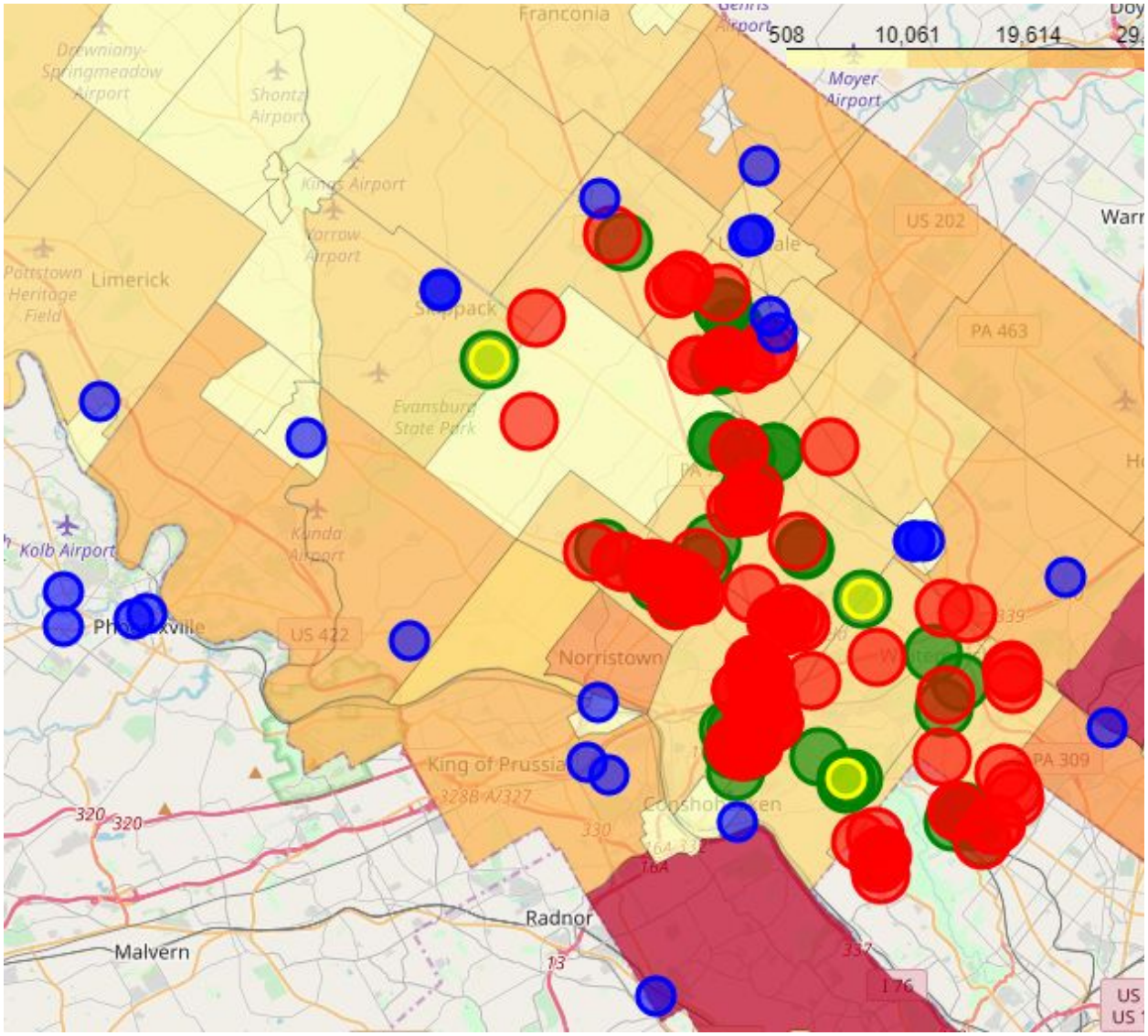


I then ran a series of searches for locations which meet the different criteria:

- Good locations marked with a green circle
- Bad locations marked with a red circle
- Existing breweries with a blue circle.
- Best locations for a new brewery with a yellow circle.

There are 3 locations which fit the criteria:

1. Near the Skippack Golf Club in Evansburg
2. Near the Phil's Tavern in Blue Bell
3. Near Brittingham's Pub in Lafayette Hill.



Discussion

I by no means believe this analysis is sufficient for a real brewery site selection. Given the limited data I had access to for free, I modified the criteria to rating sites based on public opinion of the brewery itself and not on the beer. As I am familiar with most of these breweries, I can vouch for the public opinion of the breweries and generally agree with the ratings (there are exceptions) and it was interesting to see a correlation between public opinion and fast food. Maybe that was just a coincidence but if I had the time and data, I would like to consider the distance from key locations and non-venue related locations. There is a very high correlation between brewery locations and railroad tracks since many of these sites are in warehouses that were near tracks or train stations. Foursquare data is not setup for general location

identification; it is setup for places you would want to go to, so proximity to old age homes, trailer parks, or construction equipment is not an option for consideration in this analysis. Ideally, I would have also been able to gather information on parking, services in the brewery like seating, food, comfort, games, activities, bathrooms, and other features to make better cluster models, but that information was not available through any free services I could find.

I also would have liked to spend more time with clustering and ML algorithms when I gain more experience. I'm not sure it would have helped with this assessment but I might have found a better criteria for ratings.

References

1. Brewers Association
2. Foursquare API
3. Penn State Data Center
4. Google Maps
5. Pennsylvania Spatial Data Access