Beer here or there?

A cluster analysis of restaurants and breweries in the twin cities

Introduction

In this project we aim to identify one or more optimal locations to open a new brewery in the twin cities, Minneapolis and St. Paul, Minnesota. This report will be targeted to stakeholders looking to open a brewery/taproom in the twin cities area. As there already exists a vibrant community of small, independent breweries in the area, we will look for locations that do not already have breweries nearby. Additionally, we think it is advantageous for breweries to be in close proximity to restaurants, as a possible destination for diners to meet before or after a meal. Hence we will also analyze restaurant density in proximity to the brewery locations and attempt to identify areas with few breweries but high restaurant density. Our conclusions will be based primarily on proximity to restaurants and existing breweries. We will seek to identify areas distant from the nearest breweries, with many restaurants nearby. Our approach and results will be data driven, and we will conclude with suggestions for the best possible neighborhoods to open a new brewery or taproom in the twin cities.

Data

We need brewery and restaurant location data. Additionally, we will need geographic information about the different neighborhoods of Minneapolis and St. Paul.

Data Sources

It should be noted that Minneapolis is divided into smaller **neighborhoods**, and larger **communities** which generally contain several neighborhoods. We will use the **neighborhood** boundaries in our study. On the other hand, St. Paul is divided into neighborhoods only. The areas of the St. Paul neighborhoods are comparable to the area of the Minneapolis communities.

There are 17 neighborhoods of St. Paul. On the other hand, there are 87 neighborhoods in Minneapolis. However, the neighborhoods Camden Industrial, Humboldt Industrial Area, and Mid-City Industrial are primarily industrial neighborhoods. Humboldt Industrial Area has no restaurants or breweries. Camden Industrial appears to have a single restaurant. However Mid-City Industrial has several breweries and restaurants.

The neighborhood boundary data for Minneapolis is extracted from here as a json file. The neighborhood boundary data for St. Paul is extracted from here as a json file. We will use the Fourquare API to extract data about brewery and restaurant locations in the twin cities.

Data Cleaning

From the neighborhood geojson files, we select the name and coordinate features, and create a dataframe with these features, **Neighborhood, Latitude,** and **Longitude**. To acquire the latitude and longitude data for each neighborhood (a single point), we use a Geopy. Some of the neighborhoods were not found using Geopy. These neighborhoods were noted and then manually inserted using a Google search. This process was done for each of the Minneapolis and St. Paul neighborhoods.

We also constructed geodataframe objects which incorporate the boundary data as a `geometry` data type for both the Minneapolis and St. Paul neighborhoods. These geodataframes were used to build maps that highlight the restaurant and brewery locations and densities.

In gathering data about **restaurants** in the twin cities, we did two initial calls per category, one centered in downtown Minneapolis and the other centered in downtown St. Paul. A maximum of 100 results was returned per call, which was far too few. As a way of gathering more results, we decreased the search radius of the Foursquare calls, and did an API call for each neighborhood. For Minneapolis, a search radius of 2 km was used, and for St. Paul the search radius was increased to 3.5 km.

Cycling through each of the neighborhoods, the results were added to a dataframe, after which the duplicate results were removed. This left us with 1201 restaurants in the twin cities region.

In a similar fashion, 112 breweries were found in the twin cities area using the Foursquare API. Because the number of breweries is dramatically less, a constant search radius of 3 km was used. Additionally, upon inspection some venues were listed as breweries that were instead a coffee shop or pizzeria, for example. These were removed from the dataset, reducing the number to 108.

Feature Selection

The Foursquare API call returned restaurants and brewery information with a number of features. The following features were deleted from the venues that were returned.

```
['referralId', 'reasons.count', 'reasons.items', 'venue.location.labeledLatLngs',
'venue.photos.count', 'venue.photos.groups', 'venue.location.postalCode',
'venue.location.cc', 'venue.location.city', 'venue.location.state',
```

```
'venue.location.country', 'venue.location.formattedAddress', 'venue.venuePage.id',
'venue.delivery.id', 'venue.delivery.url', 'venue.delivery.provider.name',
'venue.delivery.provider.icon.prefix', 'venue.delivery.provider.icon.sizes',
'venue.delivery.provider.icon.name', 'venue.location.neighborhood',
'venue.categories', 'venue.location.distance', 'venue.location.crossStreet']
```

The features that were kept from the Foursquare call were

```
['venue.id', 'venue.name', 'venue.location.address', 'venue.location.lat',
'venue.location.lng']
```

Additionally, a `'type'` feature was added, with value 0 for restaurants and value 1 for breweries.

The restaurants and breweries were combined into a single dataframe, and several other features were computed and added. These extra features were

`'x'`	Converted longitude coordinate
`'y'`	Converted latitude coordinate
`'nearest_brew'`	Distance to nearest brewery
`'nearest_brew_name'`	Name of nearest brewery
`'brew_within_r'`	Number of breweries within radius *r*
`'food_within_r'`	Number of restaurants within radius *r*
`'brew_food_ratio'`	Ratio of breweries within *r* to restaurants within *r*

For the analysis below, we fixed *r* to equal 1.25 in the latter three features above. This is measured in kilometers.

Methodology

First, we used geopandas to plot the restaurants and breweries on a map of the neighborhoods in the twin cities.

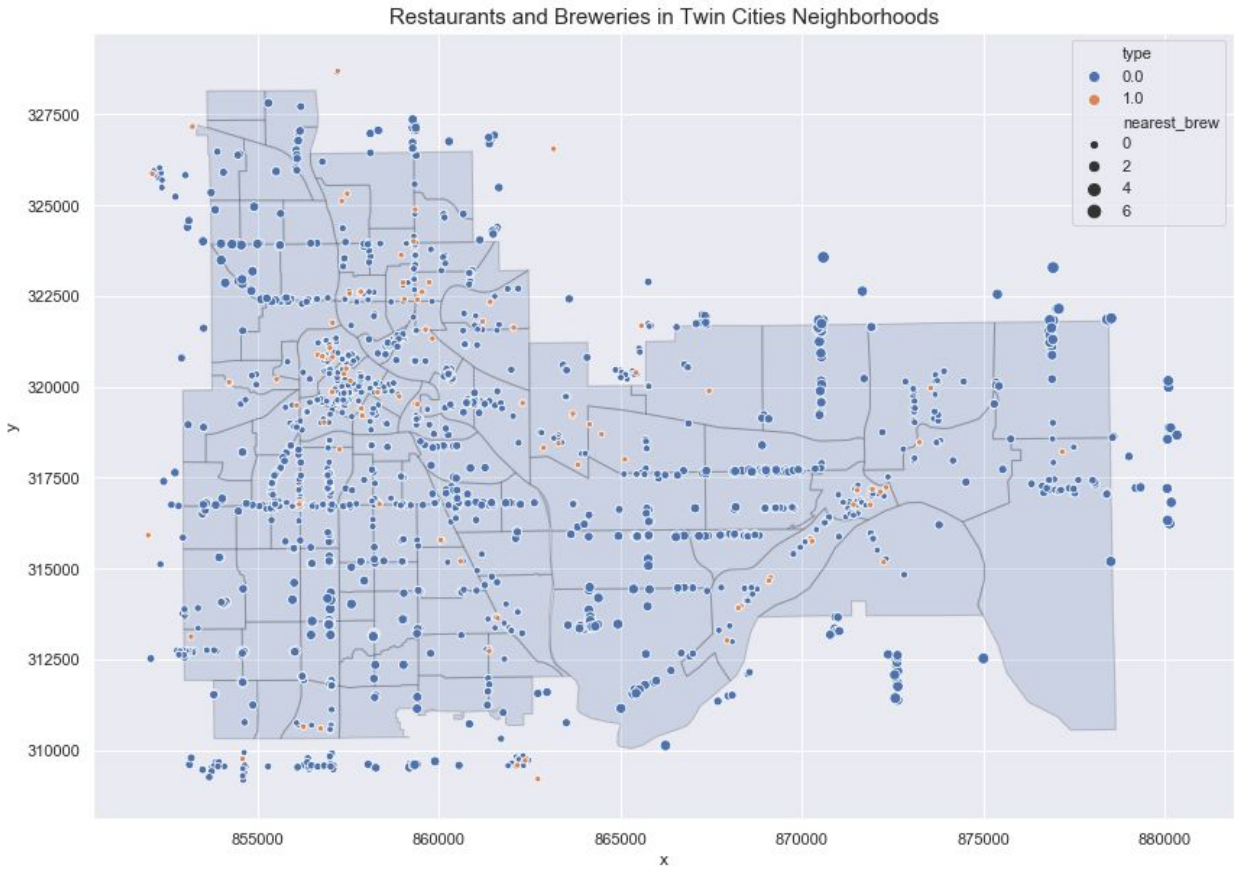

Figure 1

Next, in Figures 2 and 3 we created choropleth maps of the breweries and restaurants in the twin cities. This gives us another method by which to see the neighborhoods with high and low densities of the two venue types.

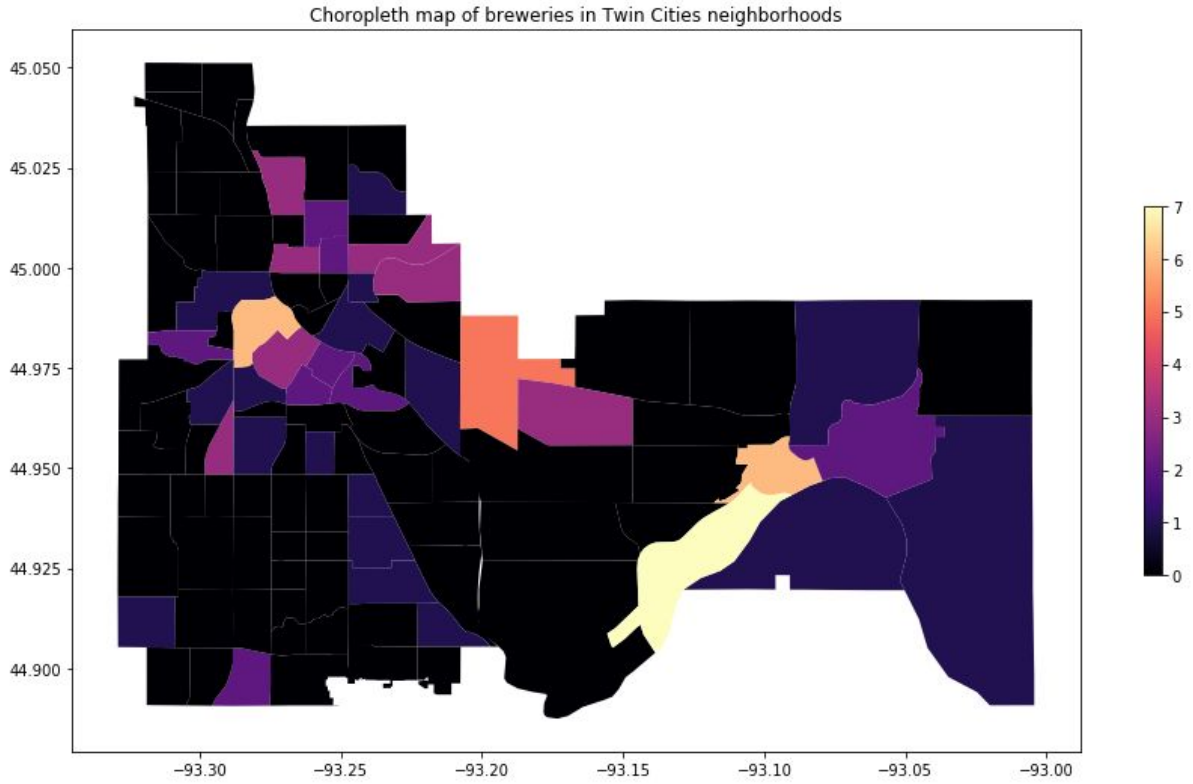

Figure 2

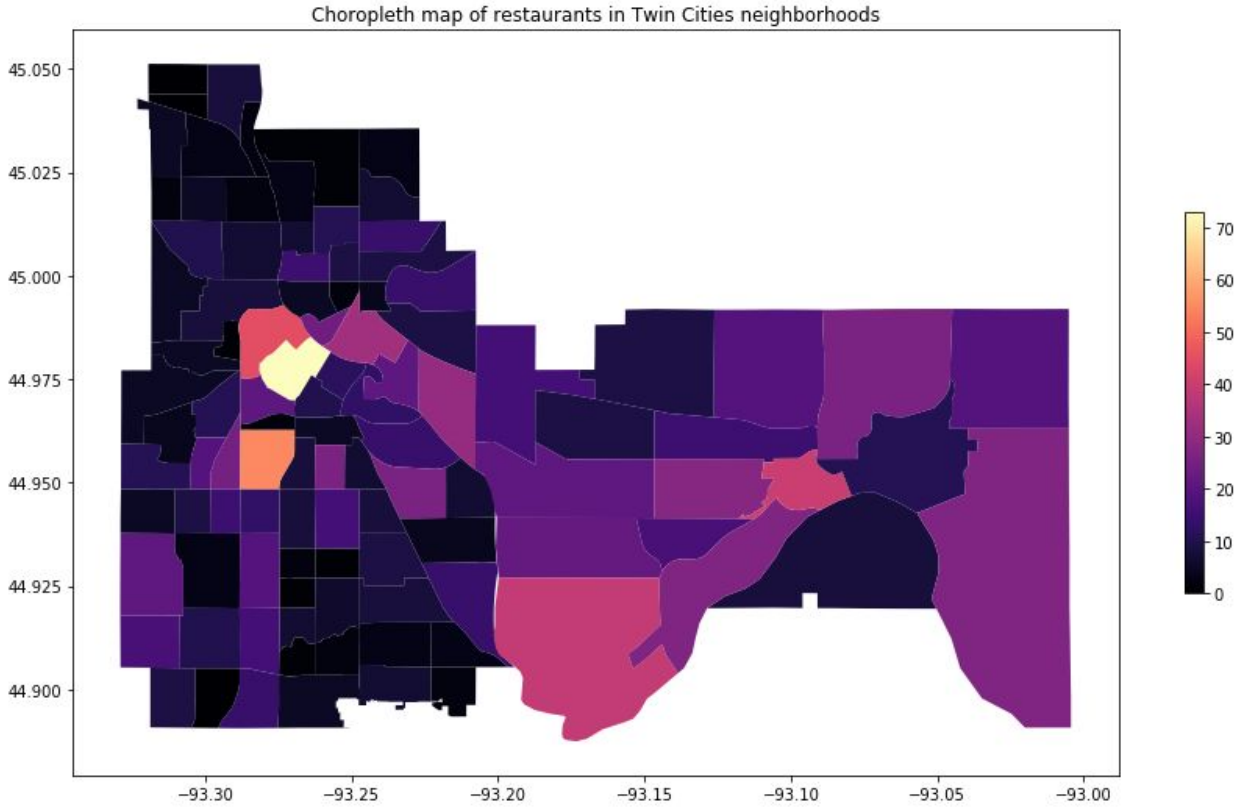

Figure 3

Clustering

We used two unsupervised clustering algorithms to group the restaurants and breweries, **k-means** and **DBSCAN**. Each method has advantages and disadvantages. K-means is an iterative algorithm that puts each venue (observation) into a cluster. In this algorithm, each observation is a part of a cluster; there are no outliers. The number of clusters is determined *before* running the algorithm. On the other hand, DBSCAN (density-based spatial clustering of applications with noise) looks for density-based clusters; i.e. clusters where observations within the cluster are 'close' to one another with regards to some metric. This algorithm does not cluster all observations; some are left as outliers. Additionally, the number of clusters is an output of the algorithm. Two key parameters are specified before running the algorithm, `eps` and `min_samples`. The `eps` parameter represents a radius centered about each venue, and `min_samples` represents the minimum number of observations that must be contained within the `epsilon` ball in order for that observation to be considered a core point. Neighboring core points and their neighbors are then grouped as clusters.

Clustering using KMeans

To determine the ideal number of clusters to use, we use the elbow method. The features that went into the algorithm were `'x', 'y', 'nearest_brew', 'brew_within_r',` and `'brew_food_ratio'`. The elbow method indicates that 3 clusters is the ideal number.

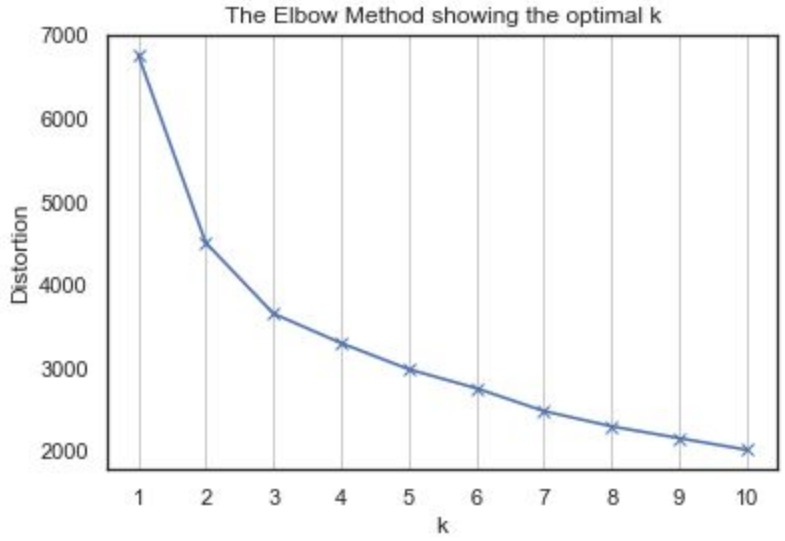

Figure 4

However, this clustering is too broad for our purposes. The three cluster regions are the north and south halves of Minneapolis and all of St. Paul, Figure 5.

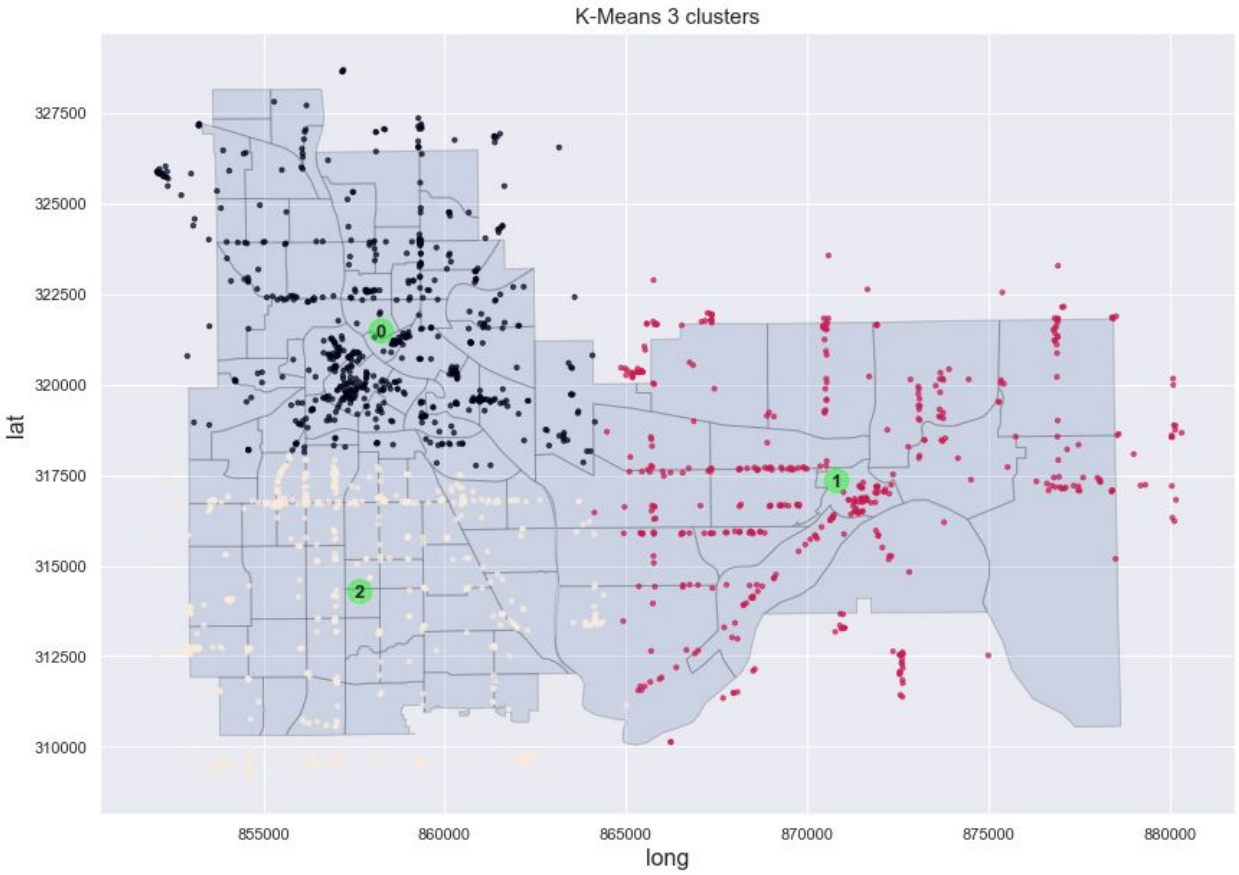

Figure 5

Seeking a more refined clustering, we notice a slight elbow at 7 clusters. The clusters are shown in the following figure.

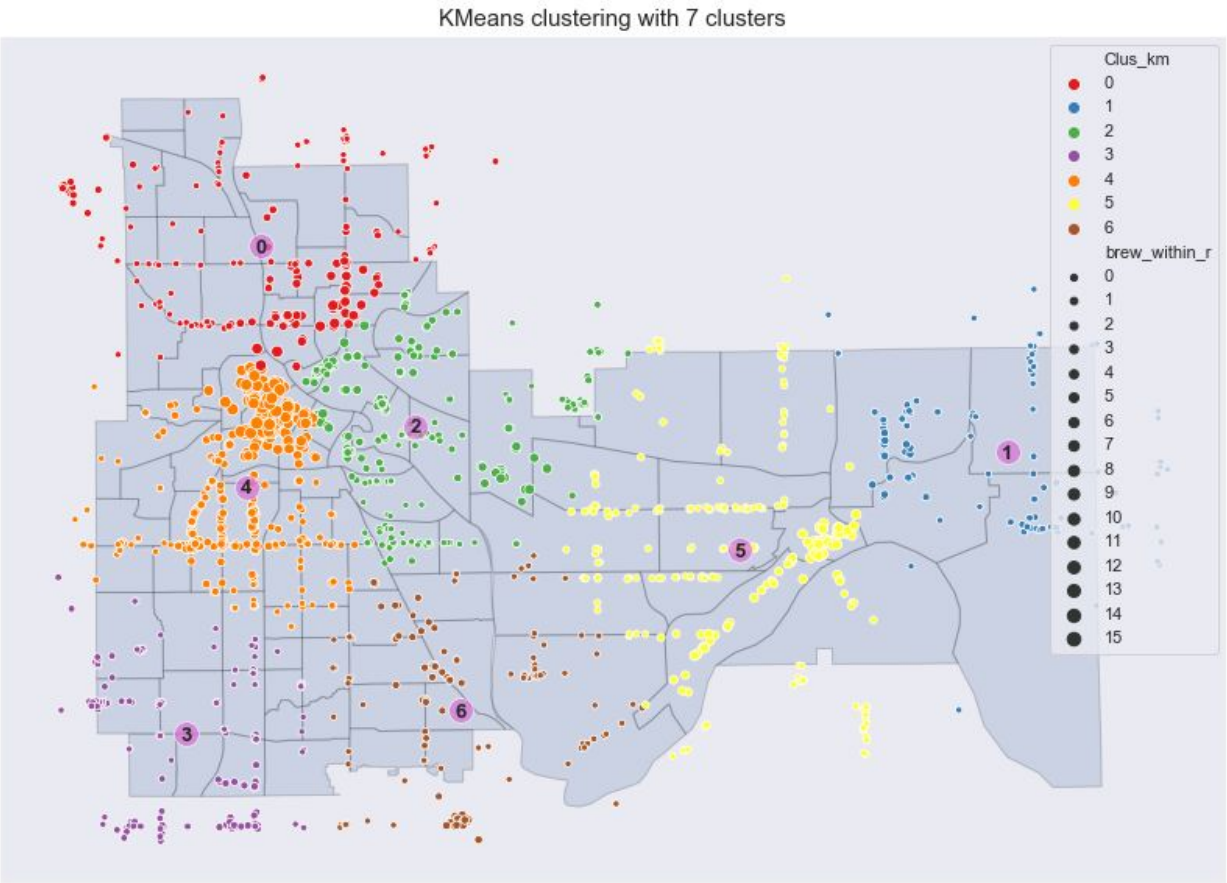

Figure 6

K-means cluster analysis

We have seven clusters using our k-means clustering algorithm. To narrow our search for preferred regions to place a new brewery, we first look at a boxplot of the restaurant and brewery counts by cluster. We note that clusters 1, 3, and 6 have the fewest breweries. Clusters 0, 2, and 4 have more. Cluster 4 in particular also has many restaurants. Looking at the map, cluster 4 contains downtown Minneapolis, which from Figure 7 we see has many restaurants and breweries.

Looking at the second boxplot in Figure 7 below, we plotted the mean brewery to restaurant ratio by cluster. Recall that the feature we took the mean of is `brew_food_ratio`, which gives the ratio of breweries to restaurants in a 1.25 km radius. A low bar either means that on average, we have few breweries or many restaurants. We should take note that this plot shows that cluster 3 has the lowest average brewery/restaurant ratio. Next, looking at the boxplot in the lower right of Figure 7, we see that cluster 3 has the largest average distance to the nearest brewery. Additionally, in the swarmplot in the lower left of Figure 7, we again see that the venues in cluster 3 in general do not have many breweries within a 1.25 km radius. Also from this plot, we see that cluster 1 and 6 also do not have venues with very many breweries within the 1.25 km radius. However, cluster 3 has more venues that have higher restaurant density.

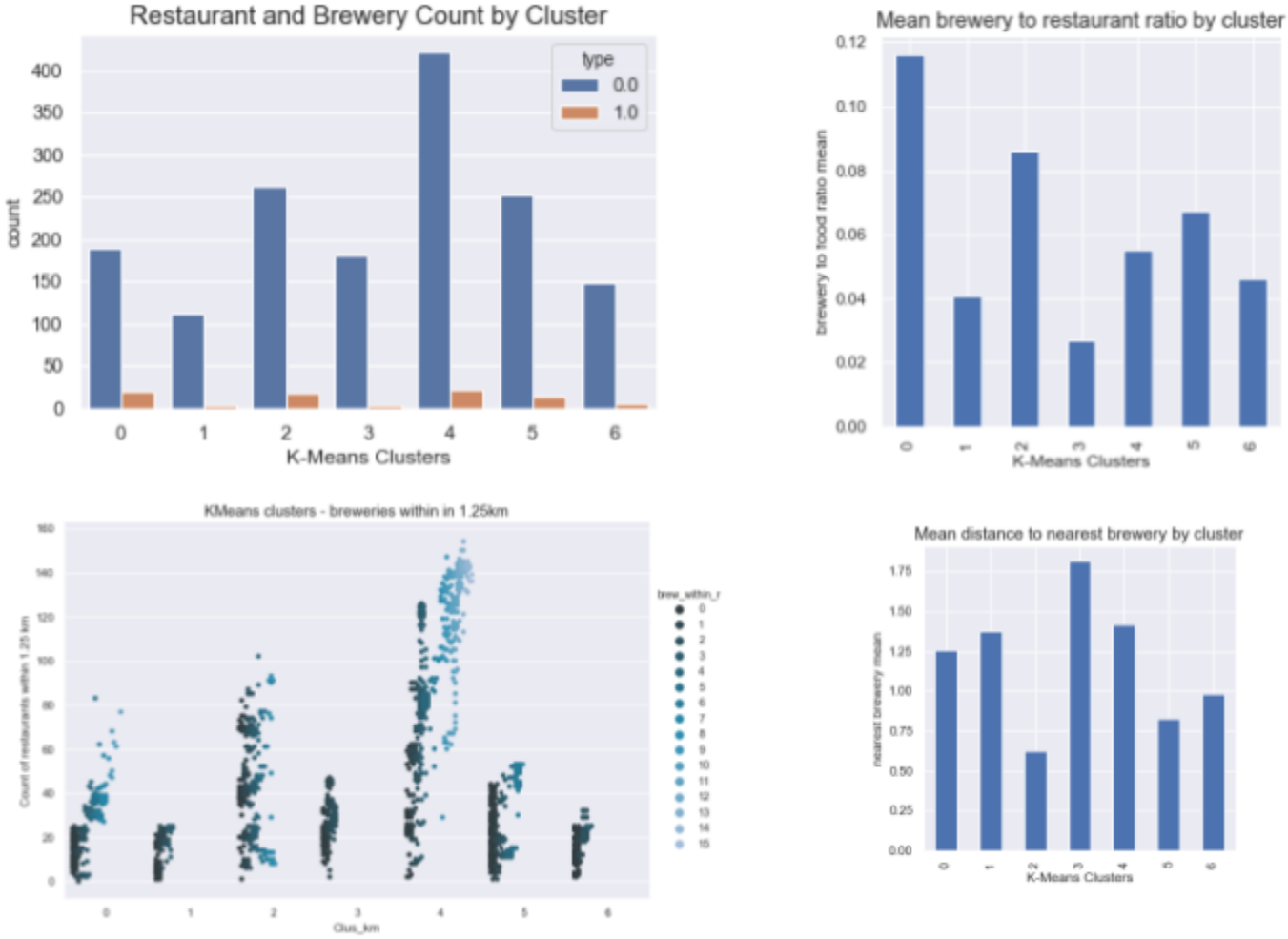

Figure 7

We should also note however, that looking at cluster 3 on the map, there are several groups of venues outside the twin cities boundaries, to the south.

Clustering using DBSCAN

Using DBSCAN, the two parameters that must be chosen are `eps` and `min_samples`. We fix `min_samples=12` and `eps=.24`. The features we used were `'x', 'y', 'nearest_brew',` and `'brew_food_ratio'`. The inclusion of the latter two features means that these clusters are not purely geographic, but incorporate venues with similar distance to breweries as well as brewery/restaurant density. See Figure 8, below.

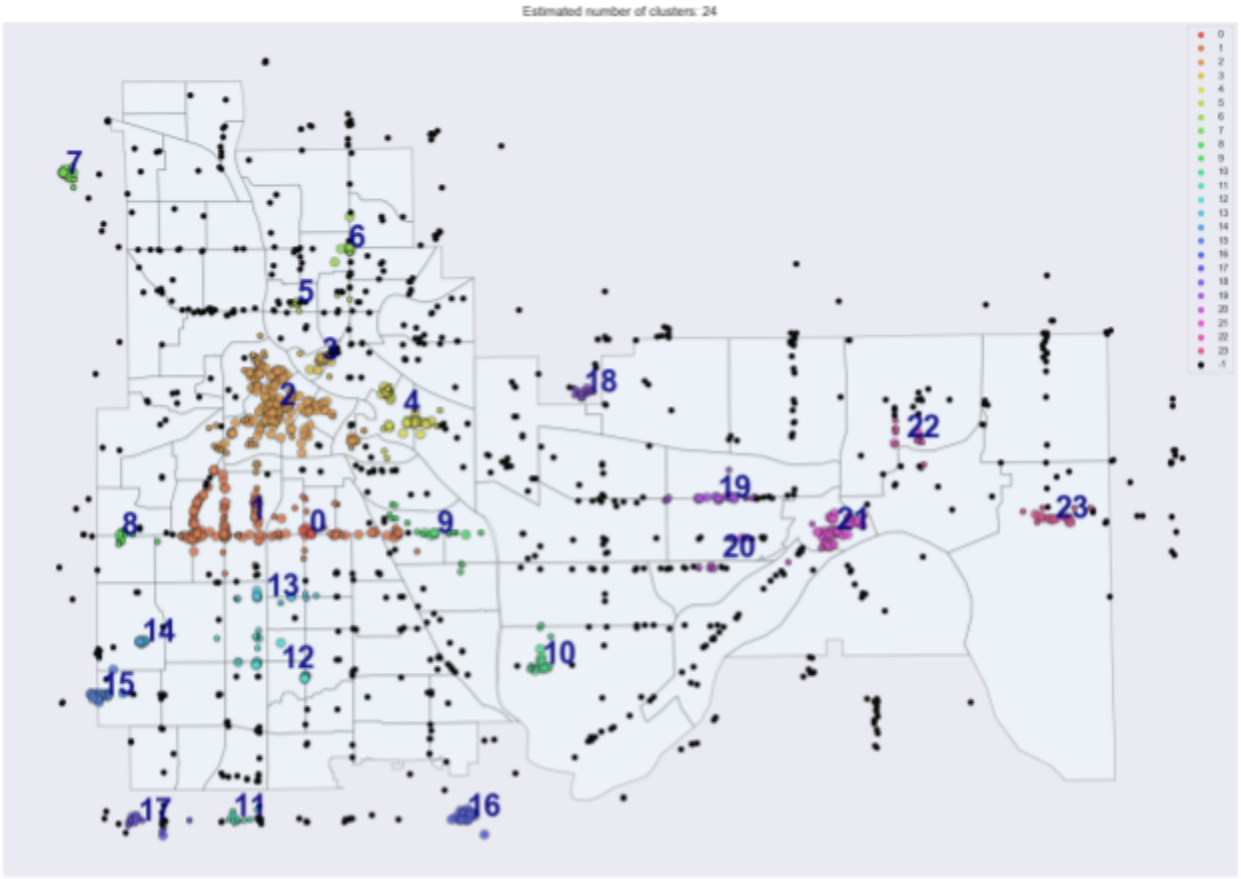

Figure 8

As we can see, these parameters result in 24 clusters, five of which lie outside of the twin cities' boundaries. We will drop these five in the subsequent analysis, looking at the remaining 19. The black dots represent venues that were not clustered in the algorithm. Additionally, the cluster numbers are plotted at the cluster centroid, however this is merely a visual key in order to more easily identify the clusters. Clusters found using DBSCAN in general do not have a meaningful centroid.

DBSCAN cluster analysis

In the following analysis, we exclude clusters 7, 11, 16, 17, and 18 as these lie outside the boundaries of Minneapolis and St. Paul. We will still refer to the remaining 19 clusters by the above cluster numbers. DBSCAN has the advantage of looking at more refined clusters, at the possible expense of omitting many outliers. We will look at the characteristics of these clusters to identify those that would be good candidates to construct a new brewery nearby. First we look at the cluster sizes.

We see here cluster 1 and 2 are clearly the largest, with over 150 venues each. These correspond to the Uptown/Lake Street and Downtown areas of Minneapolis, respectively. Note the next two largest clusters are 4 and 21, corresponding to the University of Minnesota East Bank and Downtown St. Paul areas, respectively.

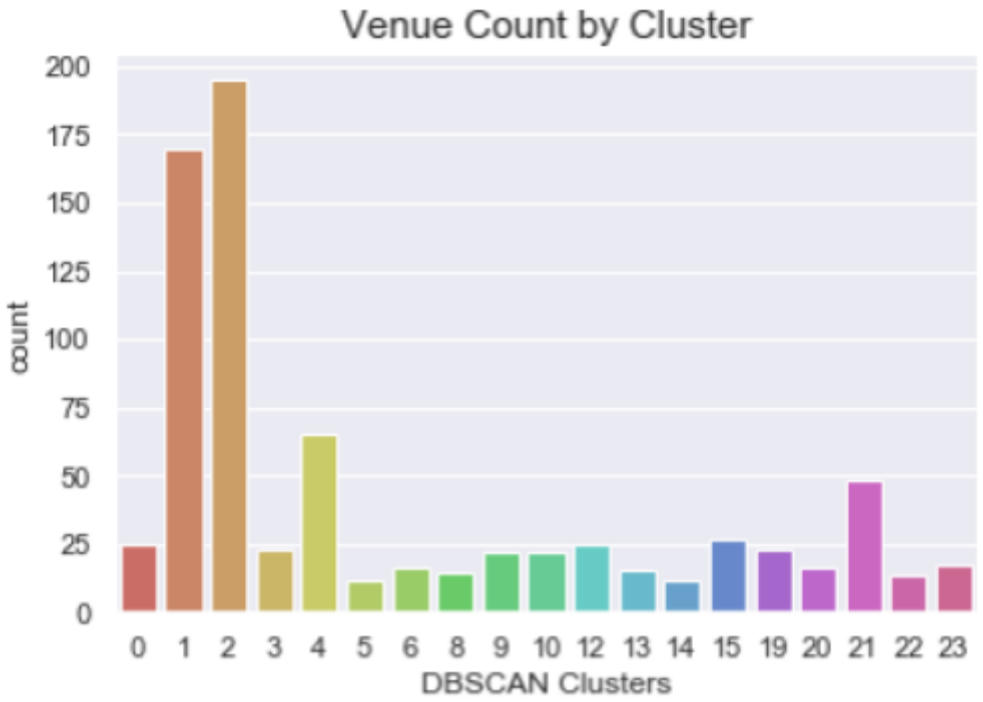

Figure 9

Next we look at the two plots in Figure 10, below. They have the same axis values, but are colored differently. The upper plot is colored by type; red for restaurants and blue for breweries. The lower plot is colored by the number of breweries within the 1.25 km radius of the venue. Darker values indicate fewer breweries within the radius, and lighter values indicate more breweries within the radius. Again we should note that cluster 1 has many venues, but we see now it has few breweries and breweries nearby. Similarly, clusters 3 and 4 have few breweries and few breweries nearby. A couple of other clusters to note are 9, 15, 19, which each have few breweries nearby, and no breweries as part of the cluster. Furthermore, these clusters have relatively higher restaurant density compared to the others listed.

Based on these plots, we now focus our attention clusters 1, 3, 4, 9, 15, and 19. From our analysis, these clusters show good potential to be places to construct a new brewery. In Figure 11 we plot the distance to the nearest brewery vs the number of restaurant venues within 1.25 km. From our perspective, an ideal cluster would have lots of observations in the upper corner of this plot. From this plot, we see both cluster 9 and 19 are both far from breweries, and neither contain any venues within 1.25 km of a brewery. That said, cluster 19 is clearly further removed from the nearest brewery, but has comparable restaurant density as cluster 9. In this sense, cluster 19 is preferred over cluster 9. Looking at cluster 15, we see it has a greater proximity to a brewery, but again with comparable restaurant density.

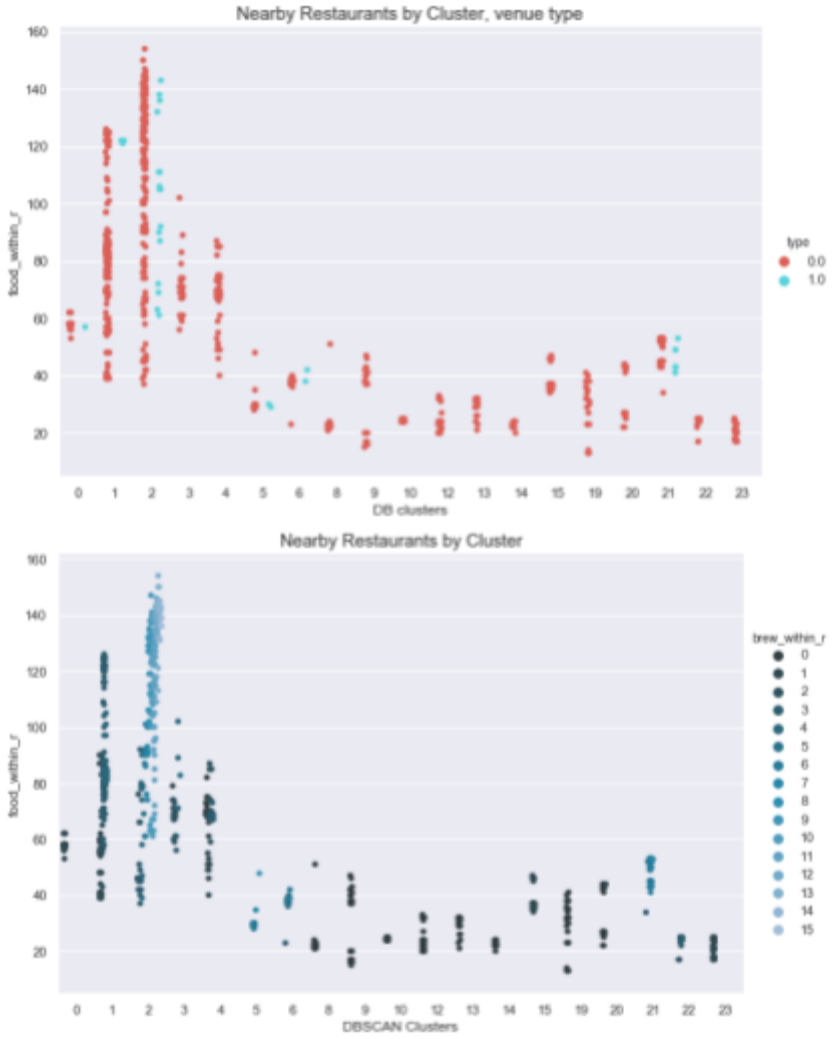

Figure 10

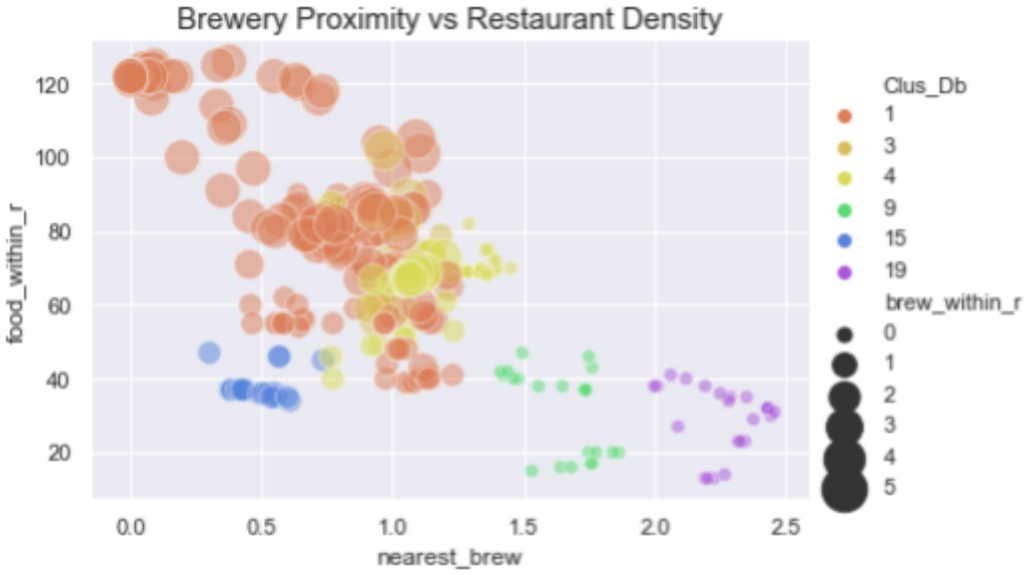

Figure 11

Results

We used k-means and DBSCAN clustering algorithms to cluster restaurants and breweries in the twin cities. From the k-means algorithm, we found that three clusters best clustered the venues in the twin cities, but that this was too broad for our purposes. Using seven with the k-means algorithm provided a better segmentation of the region, but results and interpretation are still vague. On the other hand, those DBSCAN leaves out a lot of outliers, this clustering algorithm does provide a picture of pockets of restaurants and breweries within the twin cities.

From the analysis of the DBSCAN clusters, three strong candidates emerged as possible locations to build a brewery. They corresponded to clusters 1, 3, 4 and 19. Cluster 1 is a large cluster, found in the Uptown region of Minneapolis. The region is home to a lot of restaurants and stores and is home to many young professionals. Cluster 3 is across the Mississippi River from Downtown, and houses several breweries already. Geographically, the cluster does not encompass a large area. Cluster 4 is right near the University of Minnesota East Bank campus and services students and employees of that institution, among others. Lastly, Cluster 19 is along University Avenue, along the border of the Thomas-Dale and Summit-University neighborhoods.

Cluster 1 has three breweries, while clusters 3, 4, and 19 each have none. The restaurant numbers for clusters 1, 3, 4, and 19 are 167, 23, 65, and 23, respectively. Cluster 4 is relatively large, with 65 restaurants and zero breweries, consisting of the dinkytown and Washington Avenue areas, which have high restaurant densities.
On the other hand, we can see from Figure 12 that cluster 19 is the farthest away from the nearest brewery by more than half a kilometer relative to the next furthest venue.

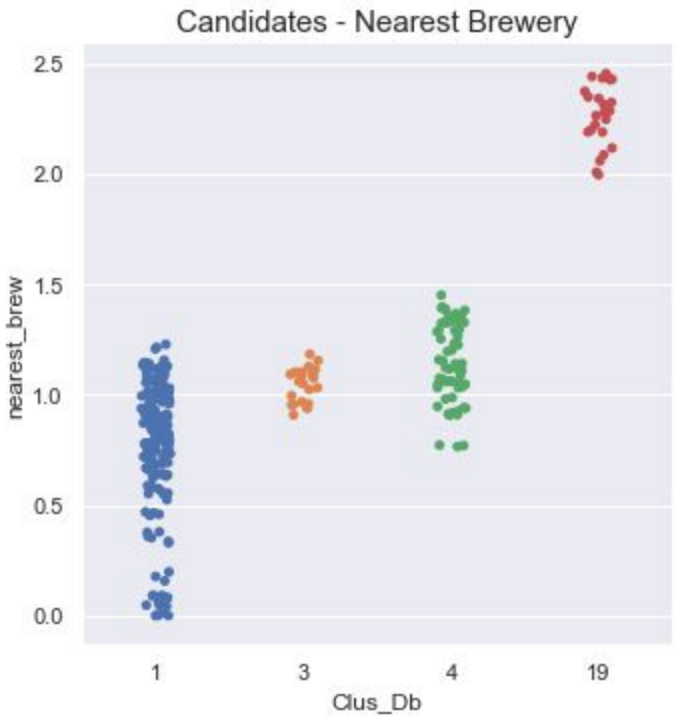

Figure 12

Based upon these observations, we recommend opening a new brewery or taproom near cluster 4, servicing the University of Minnesota neighborhood. There are no immediate breweries currently serving this area. On the other hand there are plentiful restaurants servicing this massive campus with its many students, faculty, and staff. We recommend taking a much closer look at constructing in this area.

Conclusion

We applied k-means and DBSCAN clustering algorithms to our dataset to arrive at several good candidate areas to open a new brewery. Upon close consideration of the data, we recommended that the client build a brewery near the University of Minnesota, labeled cluster 4 according to our DBSCAN output. We believe this would be an excellent area to build. It services a very large population of people associated with the university such as undergraduate and graduate students, faculty and instructors, and administrative staff. The average distance to the nearest brewery is 1.15 km.

There are several closing comments to this report.

1. Our conclusions were based solely on geographic data about restaurants and breweries. Other considerations such as demographics, zoning, traffic and transportation, and crime were not factored. An important followup to this report would be one incorporating some of these other factors.

2. The conclusions were based solely upon geographic proximity to restaurants and breweries. Two primary assumptions of the study are that higher density of restaurants and lower density of breweries are desirable. These principles guided our decision making. However it is possible that the client may hold differing ideas about the relative importance of these principles, which would be addressed in a followup report.

Thank you for giving me the opportunity to work on this project.