Coffee, the fuel of the modern industry in Moline, IL

By Aaron Balk

1. Introduction

The business problem I will explore will be the ideal location to build a new modern drive-through coffee shop in my hometown of Moline, Illinois (https://en.wikipedia.org/wiki/Moline,_Illinois). While coffee shops are by no means a new thing in Moline, most coffee shops are located in newer developed retail areas outside of the city of Moline proper. If a new coffee shop could be set up more conveniently close to the Moline residential areas, it could give that business an edge for people who get their coffee on their way to work downtown (the primary employers in Moline are John Deere and the Rock Island Arsenal, both of which have a majority of their factories downtown along the Mississippi river).

In addition, locally-owned coffee shops are becoming more and more popular over many of the national retail chains, like Starbucks and McDonalds. Residents take pride in supporting their local businesses, and the local shops can often take advantage of the "cool factor" to distinguish themselves from national chains. Some other locally-owned coffee shops that have done exceptionally well are Milltown Coffee (http://milltowncoffeeqc.com/) and Atomic Coffee Bar (https://www.atomiccoffeebar.com/)

The problem is, how can data science and location data from Foursquare help us decide on the perfect location for our new modern drive-through coffee shop? People interested in this analysis would be entrepreneurs, coffee lovers, potential investors, or anyone else who would care about creating a new modern coffee shot in their hometown. Other interested stakeholders could be Moline locals like me who would simply find the analysis of their hometown interesting.

2. Data

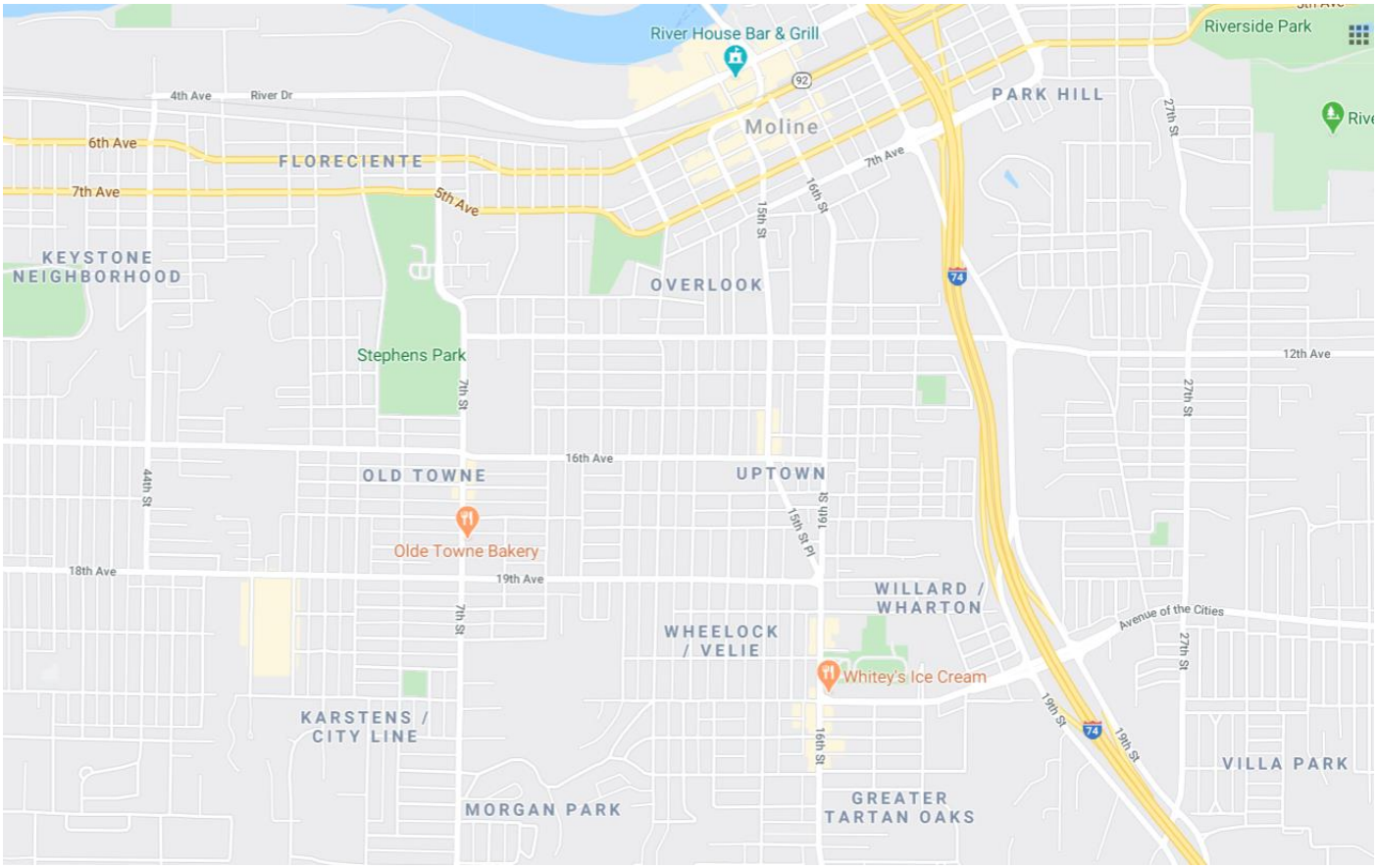
To help determine the perfect location, we will need location data. Moline is made up of many historic neighborhoods (https://en.wikipedia.org/wiki/Moline,_Illinois#Neighborhoods), Unfortunately, there may not be many easily accessible data sources that list or detail these neighborhoods, like there was with the previous Manhattan and Toronto assignments. Once we have a list of neighborhoods and their locations, we can connect to Foursquare and collect data on the most common types of venues in and near these neighborhoods. We will want this data to make sure that we are not building our coffee shop too close to an area that already has lots of coffee options (like other cafés or coffee shops, and to a lesser extent donut shops, diners, and gas stations). To avoid having to go through every single neighborhood in detail, we will use K means cluster analysis to cluster the neighborhoods into similar groups based on location and the most common venues.

3. Methodology

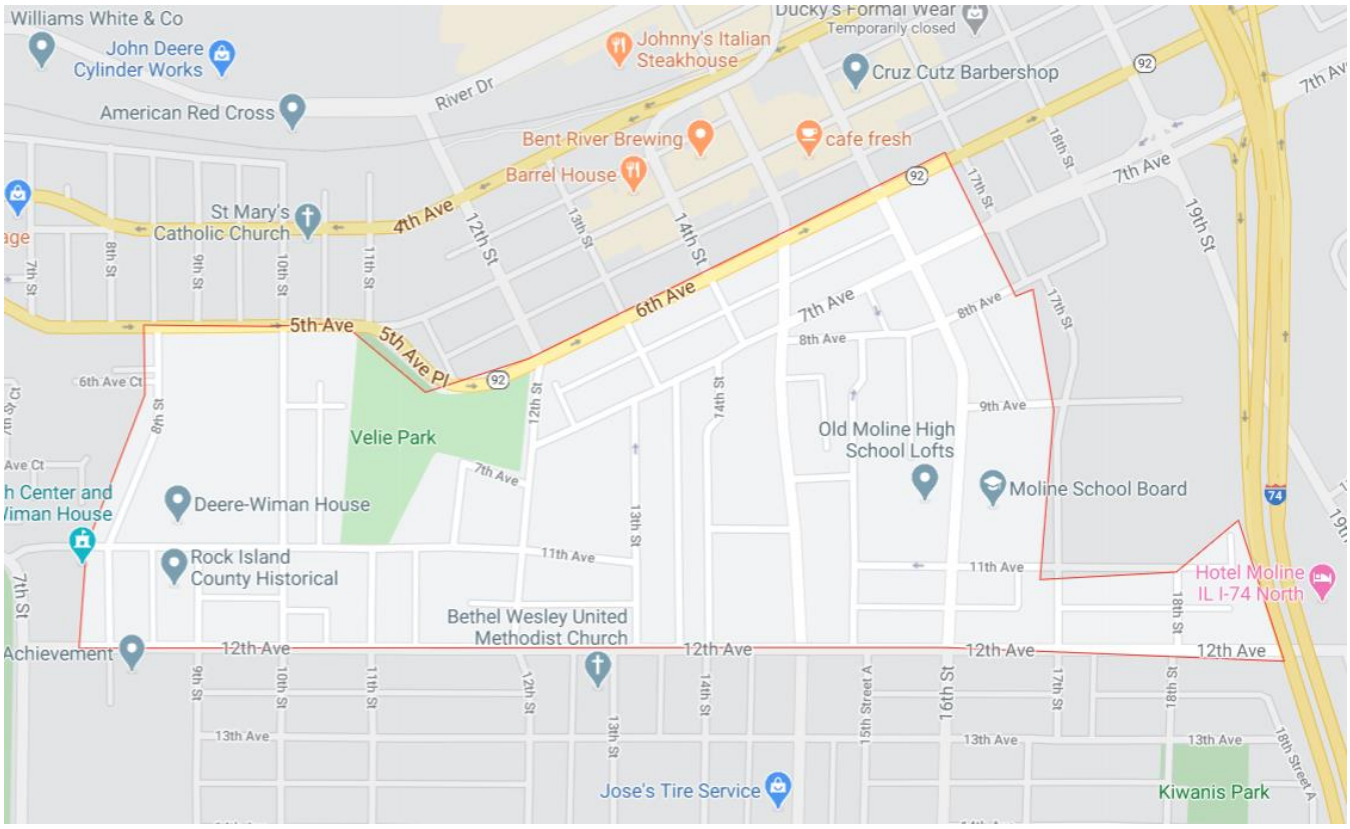
Since robust data on the neighborhoods in Moline did not exist, I ended up creating it myself. First, through Wikipedia and some additional internet searching, I was able to compile a list of 26 historic neighborhoods for Moline:

1. Hamilton Heights
2. Wildwood
3. Prospect Park
4. Park Hill
5. Forest Hill
6. Highland
7. Villa Park
8. Green Acres
9. Molette
10. Rockview Estates
11. Homewood
12. Heritage
13. Stewartville
14. Deerview
15. Walton Hills
16. Olde Towne
17. Uptown
18. Willard/Wharton
19. Karstens/City Line
20. Morgan Park
21. Greater Tartan Oaks
22. Overlook
23. Floreciente
24. Keystone Neighborhood
25. Wheelock/Velie
26. Downtown

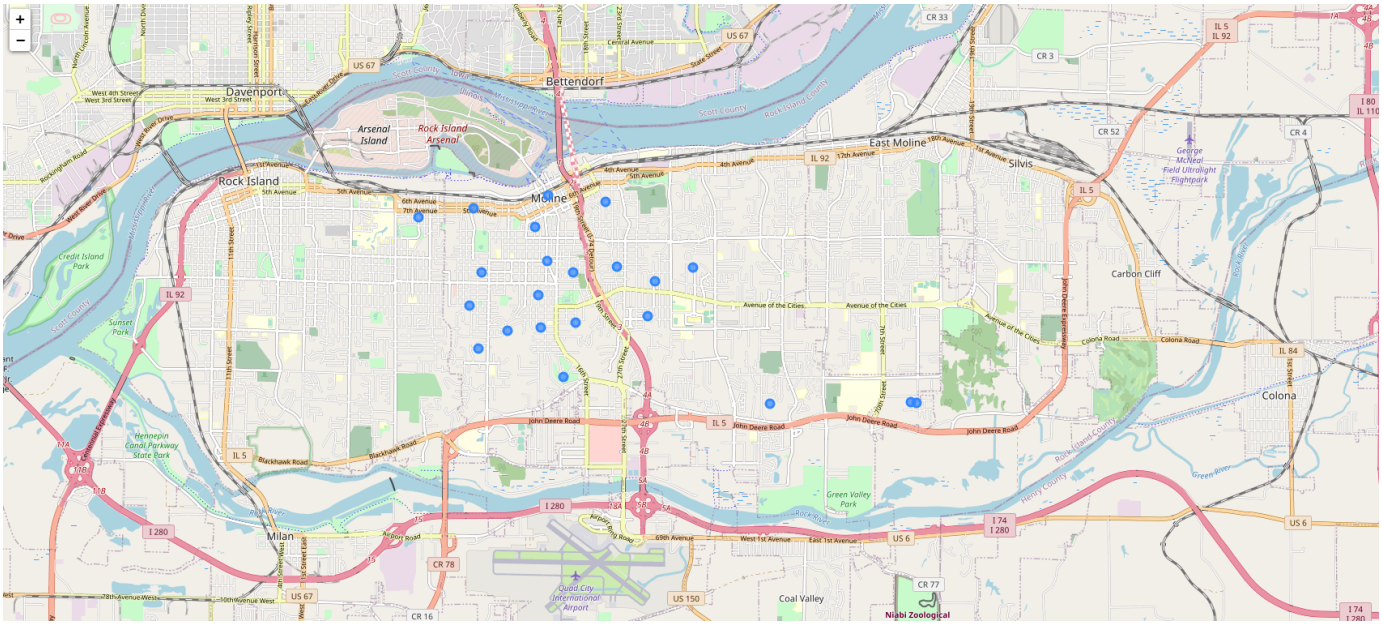
After compiling the list, I needed to get estimates for the longitude and latitude of the center of these neighborhoods. This process was rather manual. Luckily, many of the neighborhoods were already labeled on Google maps, as seen below:



Selecting the neighborhoods that were already labeled in Google maps created a nice outline of the neighborhood (as seen in the screenshot below for the Overlook neighborhood), making it easy to select a location roughly in the center and documenting its latitude and longitude.

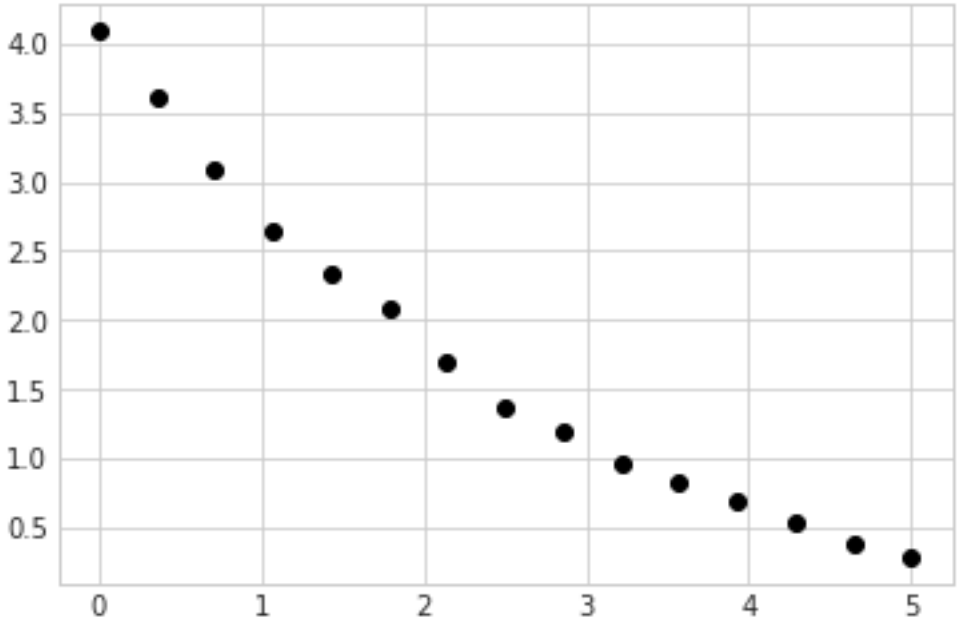


For locations that were not labeled on google maps, I had to do extra searching on the internet to figure out the rough location of those neighborhoods. Real estate websites were usually the most helpful for finding houses in a given neighborhood, and then I would use Google Maps to find the latitude and longitude of the rough center of that neighborhood. Finally, I was able to create my own csv file that listed all 26 historic Moline neighborhoods, as well as their corresponding latitude and longitude data. To avoid having to go through every single neighborhood in detail, we will use K means cluster analysis to cluster the neighborhoods into similar groups based on location and the most common venues. Using Folium, I was able to create the following map detailing the location of the neighborhoods:



Once I loaded my csv file into my Jupyter notebook, I used Foursquare to get data on venues nearby to the locations of each neighborhood. Then I categorized each venue, and found that I pulled back 75 unique categories of venues (a variety which I am happy with). Next, I performed on hot encoding on my data to prepare it for K means clustering. I then created a dataframe that displayed the top 10 venues for each neighborhood.

Finally, I was prepared to perform my K means clustering. However, I needed to decide how many clusters I should create (the value of K). in order to do so, I computed the total within-cluster sum of square for a variety of values of K (see plot below):



Since there is not a clear “elbow” in the data, I just decided to assign k the value of 5.

4. Results

My results provided me the following clusters of neighborhoods:

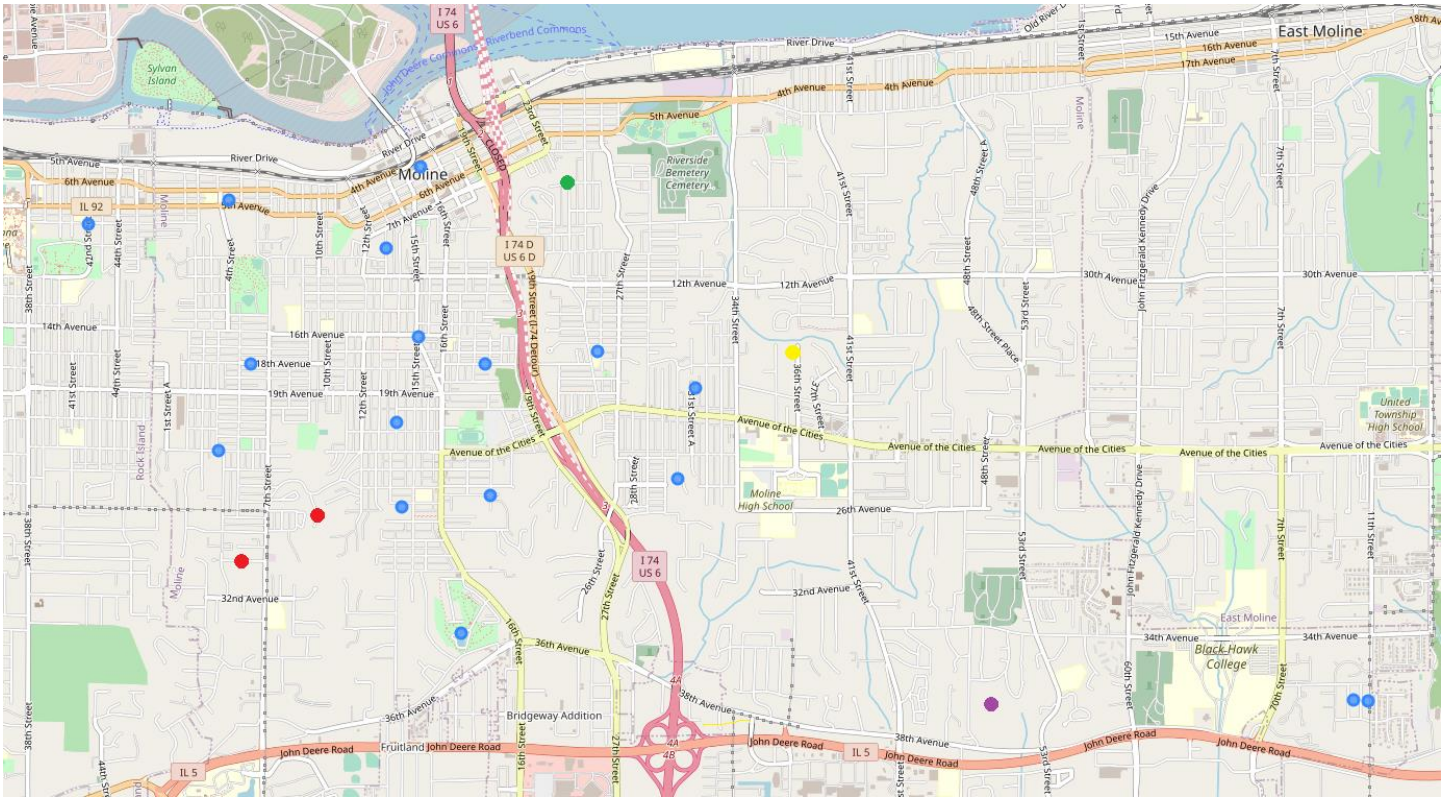
- Cluster 1 = Pizza, ice cream, and restaurants
    1. Prospect Park
    2. Forest Hill
    3. Highland
    4. Villa Park
    5. Stewartville
    6. Deerview
    7. Walton Hills
    8. Olde Towne
    9. Uptown
    10. Willard/Wharton
    11. Karstens/City Line
    12. Greater Tartan Oaks
    13. Overlook
    14. Floreciente

15. Keystone Neighborhood
16. Wheelock/Velie
17. Downtown

- Cluster 2 = Gourmet shops
    1. Hamilton Heights
    2. Morgan Park
- Cluster 3 = Gas stations and electronics
    1. Park Hill
- Cluster 4 = Pizza and donuts
    1. Molette
- Cluster 5 = Movie theater

Please reference my notebook code for thorough details of each cluster and the most common venues associated with each. In my dataframe, I showed the top 10 most common categories of venues so that I could see which clusters showed fewest categories of coffee serving shops.

Using Folium, I was able to create a map showing different color markers for each cluster:



5. Discussion

The most obvious thing that jumps out from my results is that the number of neighborhoods in each cluster is very skewed, with most of the neighborhoods grouped into cluster 1. Granted, most of the neighborhoods in cluster 1 are geographically close to each other. Additionally, the Toronto neighborhood clustering assignment had a similar skewed grouping, with most of the neighborhoods falling into one cluster. And if you look at the most common venues near these Moline neighborhoods, there are a lot of similarities. So many similarities in fact, that it makes me wonder if Foursquare really

has robust venue data for the Moline area. Maybe Foursquare data is more complete for bigger cities than it is for smaller metropolitan areas like Moline, and what I am seeing in my results is many of the same exact venues being counted as "close" to multiple neighborhoods. However, it is also possible that since Moline is a smaller city, there are not enough venues to create significant distinctions between nearby neighborhoods. This phenomenon may also be seen in cluster 2, where we have 2 neighborhoods that are very geographically close to one another, and they have nearly identical categories of common venues. This would lead me to believe that the same venues are seen as "close by" for both of these neighborhoods.

Clusters 3, 4, and 5 are all along the outskirts of Moline, and each are only comprised of a single neighborhood. For the most part, geographic location seems to be a the most significant factor that lead to the different cluster groupings.

Now, to decide the best area for our new coffee shop, we must look to see which clusters show no a few signs of venues that serve coffee. One of the first insights that stick out is that venues categorized as "café" and "coffee shop" are almost nowhere to be seen in our data. In fact, Cluster 1 is the only neighborhood where Café is listed as a common venue category (6th most common in the Downtown neighborhood). I believe that this is further evidence that the Foursquare data may be lacking for Moline.

However, not all is lost. I still believe we can get valuable insights into an appropriate place to set up a new coffee shop by looking at other categories of venues that are likely to serve coffee (such as donut shops, gas stations, diners, and convenience stores). Cluster 1 has by far the most nearby venues that serve coffee

Looking at Cluster 1, we see a lot of venues that likely serve coffee:

- Convenience stores nearby in 4 neighborhoods
- A café in the downtown neighborhood
- Donut shops near 3 neighborhoods
- Gas stations in 3 neighborhoods
- Diners in 4 neighborhoods

So it seems clear that cluster 1 is not an ideal location, since there seems to be a decent amount of coffee competition present. Looking at cluster 2, we see very little in terms of places to eat, and the only venues listed that might serve coffee are the diners, but they are listed low on the most common rankings (8th and 9th most common for the two neighborhoods. So, at first glance, cluster 2 looks like a promising area. The most common venue in cluster 3 are gas stations, and gas stations usually serve coffee. Further research may be required to determine if these specific gas stations do much coffee business, and thus gauge the threat they would pose to our new coffee shop. Cluster 4 has donut shops as the 2nd most common venue, which may pose a serious competitive thread. Cluster 4 also has convenience stores and diners in its neighborhood, making it appear unappealing for our new coffee shop. Finally, cluster 5 only lists diners as the 10th most common venue in that neighborhood, so it seems like there is little competition in that area.

From this quick analysis, the ideal location for our new coffee shop is cluster 2 (Hamilton Heights and Morgan Park), and cluster 5 (Homewood). However, we see that Homewood (the purple dot on our map) is geographically far away from most of Moline. Also, given my knowledge of the area, I know

that Homewood is vey close coffee shop options that did not appear to show up in the Foursquare data (such as Panera and Starbucks).

Therefore, cluster 2 appears to be the best location for our new coffee shop. It has no coffee competition in the area, it is near many residential neighborhoods, and it is near 7th Avenue, which is one of the main roads that commuters take to go to work at John Deere or the Rock Island Arsenal.

6. Conclusion

From our data analysis, cluster 2, representing Hamilton Heights and Morgan Park, appears to the be the ideal location for setting up our new modern drive through coffee shop. The cluster analysis was effective at showing that this area is not currently being served by any similar coffee competitors. My personal knowledge of the area confirms this, and I also know that this is a heavily populated residential area and is near a main road that many people take to work every morning on their way to John Deere factories or the Rock Island Arsenal.

However, our data analysis is by no means full proof. While investigating our results, there were several indications that the Foursquare data may not be complete for the city of Moline. One giant red flag was the fact that cafés and coffee shops didn't even show up as common venues, when I know plenty of neighborhoods that have them. If I was not satisfied with the results of my analysis, I could do more comparisons of the Foursquare data against my own knowledge, or against other location data sources. Another possible way to refine my analysis would be to try different numbers of clusters, to see if that significantly changes my results.