Restaurant Location Selection
Macon, Georgia

Tim Brady

February 17, 2019

1. Introduction

1.1 Background

Macon is a city located in central Georgia (USA) with a metro population of around 229,000 residents. Of the 229,000 residents, approximately 30,000 are college students[1]. Among the colleges and universities that call Macon home is Mercer University. With colleges bringing in people from other states and countries, you expect to see a large variety of tastes and preferences. This drives demand for services that depend on that variety, such as restaurants.

1.2 Problem

To entrepreneurs looking to get their start, restaurants are attractive because they offer an opportunity to provide a repeatable service to many people. The same features that make them attractive as a potential business also make the competition among restaurants significantly higher. One of the mantras for many brick-and-mortar businesses is "location, location, location". So how can we, within a region, be more strategic and data driven about both the type of restaurant we wish to establish, but also the location? This is the answer I hope to uncover for Macon, GA.

1.3 Interest

This process does not only apply to Macon, GA. It can be replicated and improved upon for other cities across the world. One key requirement, which I learned in the early phase, is that in order to replicate this process exactly, there must be Foursquare data available for your city. Even some larger cities (such as Warner Robins, GA – where I originally scoped the project) are missing data through Foursquare. Please see the Future Directions section of this paper for additional information that could be applied to situations such as that.

2. Data Acquisition and Cleaning

2.1 Data sources

This project relies entirely on Foursquare data. I limited my restrictions regarding restaurant type by using only the top-level category for "Food" within the Foursquare data. The starting point for data was Mercer University. The scope was left intentionally broad in an effort to avoid adding my own bias into the project.

2.2 Feature selection

There are only 4 fields within the results which we need to start our analysis. The name of the venue, the category of the venue, and the latitude and longitude (location) of the venue. The location is critical because we will be doing k-means clustering based on the locations of each restaurant. Other aspects of the Foursquare data we pull will be culled for this project. I discuss some other potential options in the Future Directions section of this report, but that remain out of scope at this time.

2.3 Data cleaning

We make API calls to collect Foursquare data and get the results in JSON format. These results have a lot of meta-data which needs to be removed before we can effectively use the data. Additionally, we want to get that information into a dataframe to leverage different python packages in the analysis.

When working with JSON data, the biggest challenge I faced was isolating the desired data into a dataframe with four columns. My first step was to create a variable and assign the relevant pieces of the response to a variable. At this point I flattened the remaining data to a dataframe to get all the data into columns. Getting the data into a dataframe allows us to only keep the columns discussed in the previous section.

The name, latitude, and longitude come through relatively clean. There is still a lot of work that needs done with category though, as there's a significant amount of meta information still within that field. I created a function to isolate the category type from that column and ran that column of the dataframe through the function to get a clean view of categories.

The final result of this data cleaning process is a dataframe with the four desired columns ready for to explore.

3. Exploratory Data Analysis

3.1 Understand the data

Since I approached this project with an intentionally broad scope regarding the outcome, I found it very valuable to visualize the data as I went through the process. Additionally, I collected some different statistics regarding the data. I spent a lot of time slicing and viewing this data in an effort to 'gut-check' each step as I went.

3.1.1 Can the data give us the answer we are looking for

At a high level, I was able to see that the area around Mercer University has 100 different restaurants that fell into 28 categories. This was an important insight because it allowed me to see that there would be some good groupings among the restaurants. If it was 100 restaurants with 100 categories, this project approach would not have been successful because we would not see any overlap in restaurant type between whatever clusters we come up with. If it was the opposite (100 restaurants in 1 category) it would have proven equally difficult because we would see complete overlap between our clusters and therefore identifying opportunity would be difficult or impossible.

3.1.2 Visualize the data

The next step in my exploration was to visualize the 100 restaurants on a map. I performed this step to see if clustering still felt like the right approach. Though the data is pretty spread out, as I reviewed the map, I was able to see that there were definite areas where I could see the data clustering. In Figure 1, the red circle is the location of Mercer University which we used as the starting point. The blue circles are the locations of each restaurant returned by Foursquare.

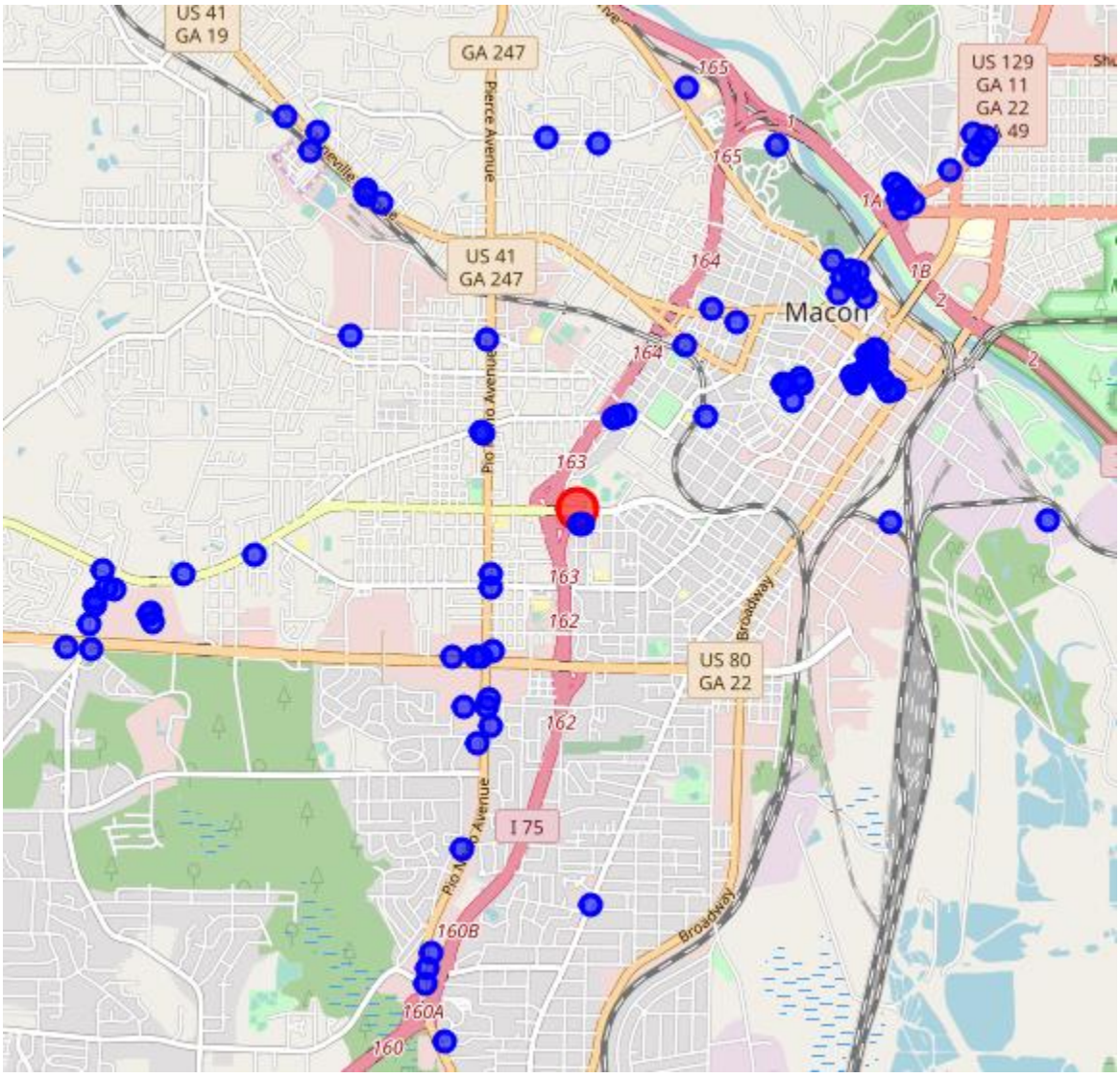
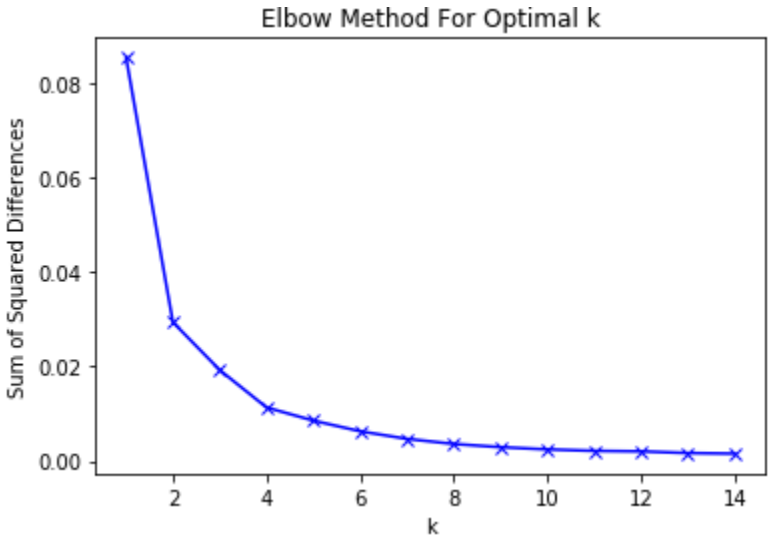


Figure 1 – Restaurants around Mercer University (red).

3.2 How many clusters

I've established that the data seems suitable for clustering, but one challenge with k-means clustering is determining how many clusters to use. One approach here is called the "Elbow Method". In this method, I visualize the sum of squared differences within the latitude and longitude over different values of k. The resulting chart should look like a bent arm and the location of the 'elbow' would represent the optimal k. The principle here is that the steeper

portion of the graph has significant reductions in the sum of squared differences, but after the 'elbow' the improvements in reductions becomes less and less for each additional k. The best result is a very distinct elbow, but this was not the case in my data. (This can happen with data that does not have very distinct clusters.) Figure 2 shows the results from performing the Elbow Method.



Figured 2 – Elbow Method to determine optimal k.

I was not satisfied with a distinct break within this chart. I see a break at k=2, but I also still see pretty significant reductions in the sum of squared differences between k = 2 and 4. So rather than looking at the raw data, I looked at the rate of change and determined that after k=4 there was a lot less improvement, and so I settled on using k=4 for further analysis.

4. Predictive Modeling

4.1 Classification models

4.1.1 Setting up k-means clustering

Given the optimal k, I used the scikitlearn python package to perform 15 iterations of k-means clustering in an attempt to find the optimal centroids for each of my four clusters.

To perform this step, I isolated only the latitude and longitude of each restaurant and ran k-means clustering with k=4 and n_init = 15. The array that is returned by this process labels each of the restaurants with a category 0 through 3. From there, I add these labels back to my original dataframe to see which cluster each restaurant belongs in. Figure 3 shows my current dataframe with this information added.

	name	categories	lat	lng	Cluster
0	Bears Den	Southern / Soul Food Restaurant	32.832927	-83.643262	3
1	Margaritas Mercer Village	Mexican Restaurant	32.832739	-83.651474	0
2	Roly Poly	Sandwich Place	32.835343	-83.634710	3
3	Francar's Buffalo Wings	Wings Joint	32.833043	-83.650336	0
4	The Rookery	American Restaurant	32.836280	-83.628172	3

Figure 3 – nearby_venues dataframe with the addition of "Cluster"

From here, I need to identify the centroids themselves. This process sets the centroid for each cluster at the center of each cluster, so the easiest way to identify the location of the centroids is just to look at the mean latitude and longitude values for each of the clusters. Figure 4 shows the results of the means of each cluster.

	Cluster	lat	lng
0	0	32.841573	-83.664249
1	1	32.818784	-83.694931
2	2	32.806457	-83.663854
3	3	32.840681	-83.629549

Figure 4. Centroid locations

4.1.2 Visualizing the clusters

At this point, it's important to validate that the steps I've taken up to this point make sense. The easiest way to do this is to visualize the previous map we produced and color-code by cluster. This allows me to validate that the clusters make sense and that I am not seeing the clusters intermingled. I additionally found it useful to plot the centroids. Figure 5 is the updated map. Mercer University and the centroids are not part of my data from Foursquare and have been coded black and pink to offset them from the rest of the data. Mercer is pink circle, black fill. Centroids are black circle, pink fill. The venues themselves are rainbow colored, assigned dynamically in the plot. Cluster 0 is red, cluster 1 is purple, cluster 2 is teal, and cluster 3 is olive.

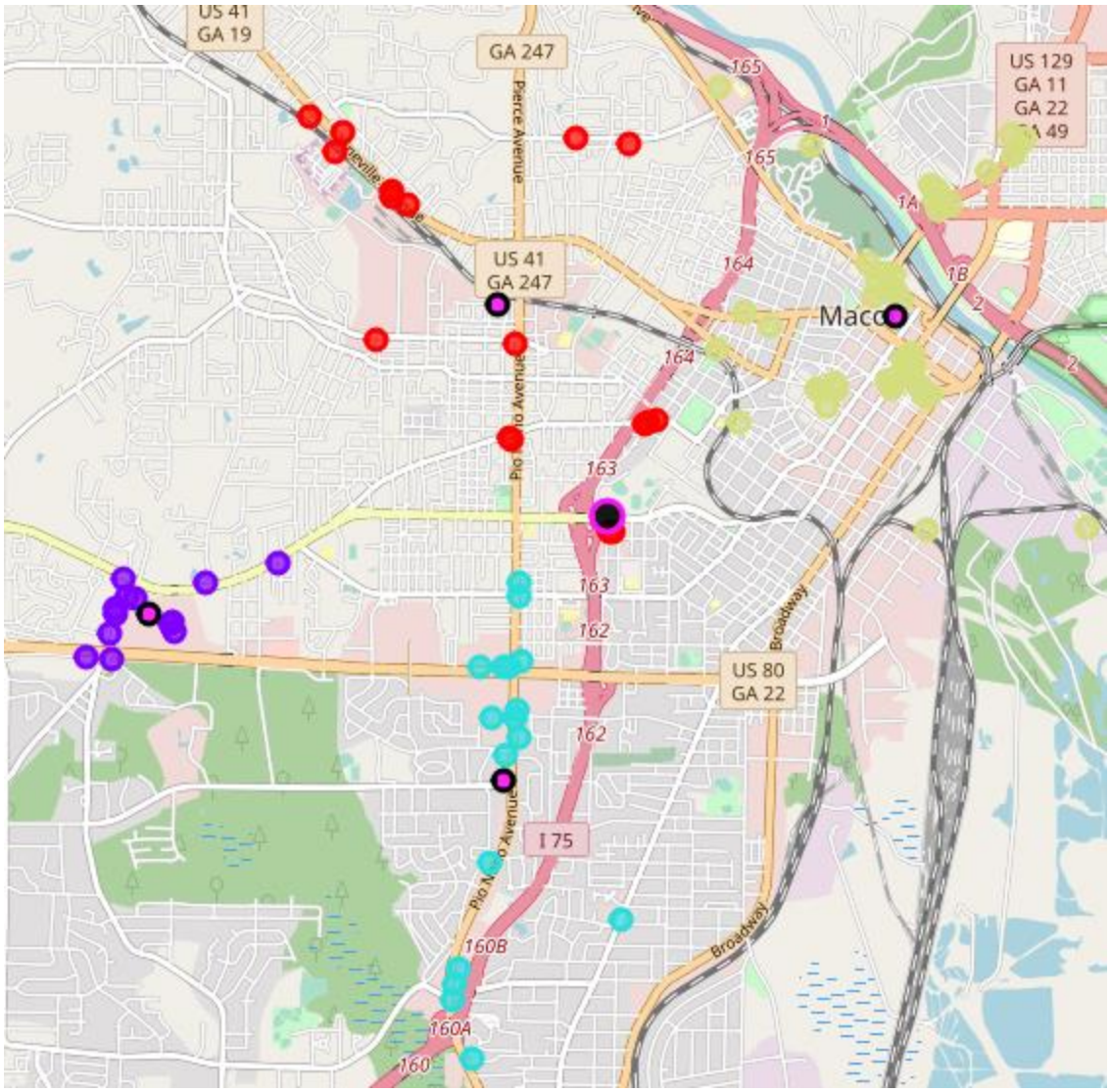


Figure 5 – Four clusters of restaurants in Macon, GA.

4.1.3 Examining the clusters

The overall goal is to determine where a restaurant category is under-represented in one cluster, but highly represented in other clusters. From there, we can establish that X type of restaurant should be located in Y cluster. I need to examine the clusters more closely to properly determine the correct X and Y in this situation.

To do this, I need to 'one-hot' encode all the categories into dummy variables so I can look at the mean representation of each category across each of the clusters. From there, I am able to report on the top 5 categories for each of the clusters.

```
----- Cluster: 0 -----
            categories  freq
0          Pizza Place  0.18
1        Deli / Bodega  0.18
2       Sandwich Place  0.12
3  American Restaurant  0.06
4     Greek Restaurant  0.06

----- Cluster: 1 -----
                          categories  freq
0             Fast Food Restaurant  0.54
1                   Sandwich Place  0.15
2              American Restaurant  0.08
3                           Bakery  0.08
4  Southern / Soul Food Restaurant  0.08

----- Cluster: 2 -----
                          categories  freq
0             Fast Food Restaurant  0.41
1                   Sandwich Place  0.12
2              Fried Chicken Joint  0.12
3  Southern / Soul Food Restaurant  0.06
4               Seafood Restaurant  0.06

----- Cluster: 3 -----
              categories  freq
0  Fast Food Restaurant  0.15
1   American Restaurant  0.13
2        Sandwich Place  0.13
3           Pizza Place  0.09
4      Asian Restaurant  0.06
```

Reviewing the frequencies of each category between the clusters reveals that Fast Food Restaurants are the most common category among cluster 1, 2, and 3, but do not even make the top 5 in cluster 0. Given this information, I feel comfortable making a recommendation.

5. Conclusions

In this study, I used Foursquare data to determine the restaurant landscape for Macon, GA. With this information I was able to determine that there are less fast food restaurants within a certain part of town and if one wanted to establish a restaurant, that would be a good candidate and the location northwest of Mercer University (cluster 0) would be a good area to begin the search.

6. Future directions

As with any project, there is always room for improvement and often answers bring more questions. In this section I would like to highlight some areas where my analysis could be improved as well as next steps.

My original location was Warner Robins, GA. There was no Foursquare data for that town. That does not mean there is no data available. Scrapping Google locations results could be used for towns not supported by Foursquare.

This approach is simply based on location. There are other factors such as crime rate, population density and demographics that can and should be taken into account for a more in-depth project. They remain out of scope for this project, but could be incorporated into future versions.

Finding an alternative to finding an optimal k in this exercise could be beneficial. I would have liked to see a more distinct break in the elbow, but that method may work well if choosing a different location where the data is clustered more heavily.

References:

1. https://en.wikipedia.org/wiki/Macon,_Georgia