Battle of the Neighborhoods

A Coursera Capstone Project

Tim Brady

INTRODUCTION

Macon is located in central Georgia (USA) and home to over 200,000 residents.

A city this size has a lot of restaurants and variety.

PROBLEM

- Success of a restaurant is often heavily influenced by location.
- Building the wrong restaurant in a good location also does not guarantee success.
- How can we determine the right type and location of restaurant to start in Macon, GA.

SOLUTION

- Using Foursquare data, we can determine the current distribution of restaurants in an area to identify under-served restaurant types.
- Segmenting the current restaurants will allow us to identify the best locations for a specific under-served type of restaurant.

DATA ACQUISITION AND CLEANING

The core data source for this project is Foursquare data with “Food” as the top-level category.

Feature Selection

- Name (to identify uniqueness)
- Category
- Location (latitude, longitude)

Data Cleaning

API calls made to Foursquare return JSON data. It’s necessary to strip out much of what is returned to isolate only the features mentioned in Feature Selection.

Tools

Python packages:

- pandas (for dataframes and analysis)
- numpy (to help handle the data)
- scikitlearn (for k-means clustering)
- matplotlib (to create visuals)

The Process

While name and location are easy to isolate in the JSON data, we need to create a function to get an easy to understand category. With this function, we can run through our dataset and clean up category to get a nice dataframe with the four features as columns.

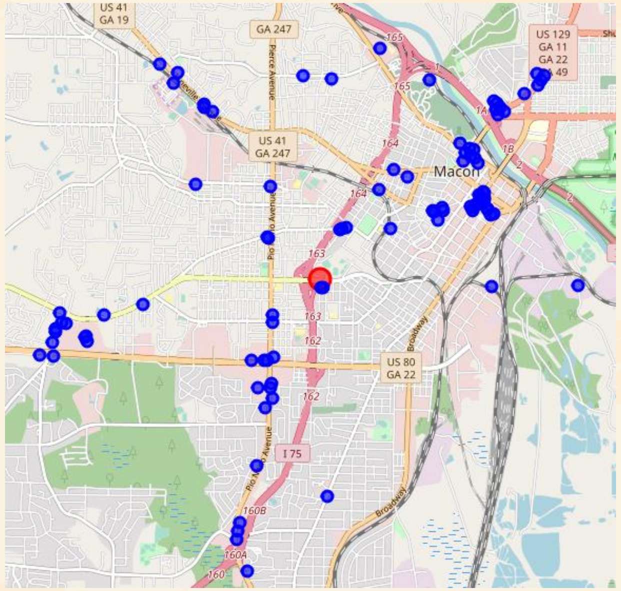
EXPLORATORY DATA ANALYSIS

Starting Location

We pull data from Foursquare based on a radius from a given location.

Mercer University in Macon, GA was chosen.

This location could be changed and the process kept the same for any other location that has Foursquare data available.



Visualize

With all our data into a simple dataframe, what does a map of the current restaurants around Mercer look like?

This is a useful step to get an impression if clustering will be the right approach.

k-MEANS CLUSTERING

An Unsupervised Machine Learning Approach

Why Unsupervised?

I want to limit my input and bias when it comes to finding an optimal business type and location.

This unsupervised approach will provide a way to group the current restaurants without my input. This will provide valuable insight compared to me pre-determining groups.

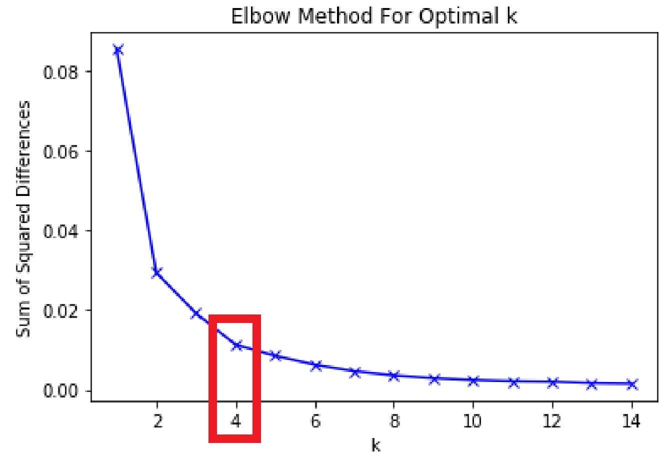
How Many Clusters

I did not want to pick an arbitrary value for *k*, so I utilized the Elbow Method to help determine the optimal *k*.

In this method, I plot the sum of squared differences (ssd) and determine the *k* where there is less value in adding *k* based on the reduction in sum of squared differences.

I determined for my data, the appropriate *k* was four. This means, I have 4 distinct groups of restaurants in the Macon area which I will use to compare.

The Elbow Method



Figured 2 – Elbow Method to determine optimal k.

PREDICTIVE MODELING

Setting up *k*-means clustering

4 areas

***k* = 4**

Using scikitlearn, we perform Kmeans clustering using 4 to start segmenting the current restaurant locations into 4 distinct clusters.

15 iterations

n_init = 15

Because we randomly drop in centroids and seek the optimal location, we can get different outcomes if we run the process multiple times.

Running this 15 times helps us truly come to an optimal location for a centroid within each cluster.

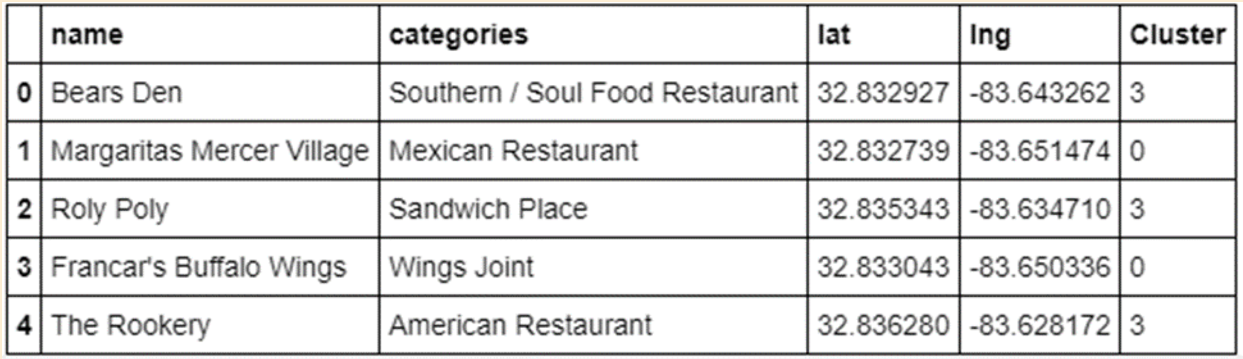
PREDICTIVE MODELING

Results of *k*-means clustering

Each Venue Labeled

- Performing k-means clustering on our dataset will assign each restaurant a label between 0 and 3.
- The label represents their cluster.
- This information is built back into the original dataframe so we have an easy to read table of the location name and the cluster it belongs to.

The Current Dataframe (head)



	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

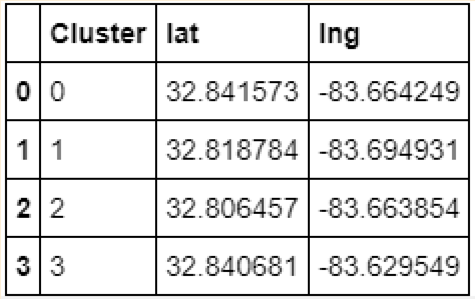
PREDICTIVE MODELING

Results of *k*-means clustering *(cont.)*

Centroid Locations

- Each cluster has an optimal center, as mentioned previously.
- The center is the mean of all the locations within the cluster.
- I created an additional dataframe with the centroids so they can be easily added to the map later as a visual reference.

The Centroid Dataframe



	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

VISUALIZE THE CLUSTERS

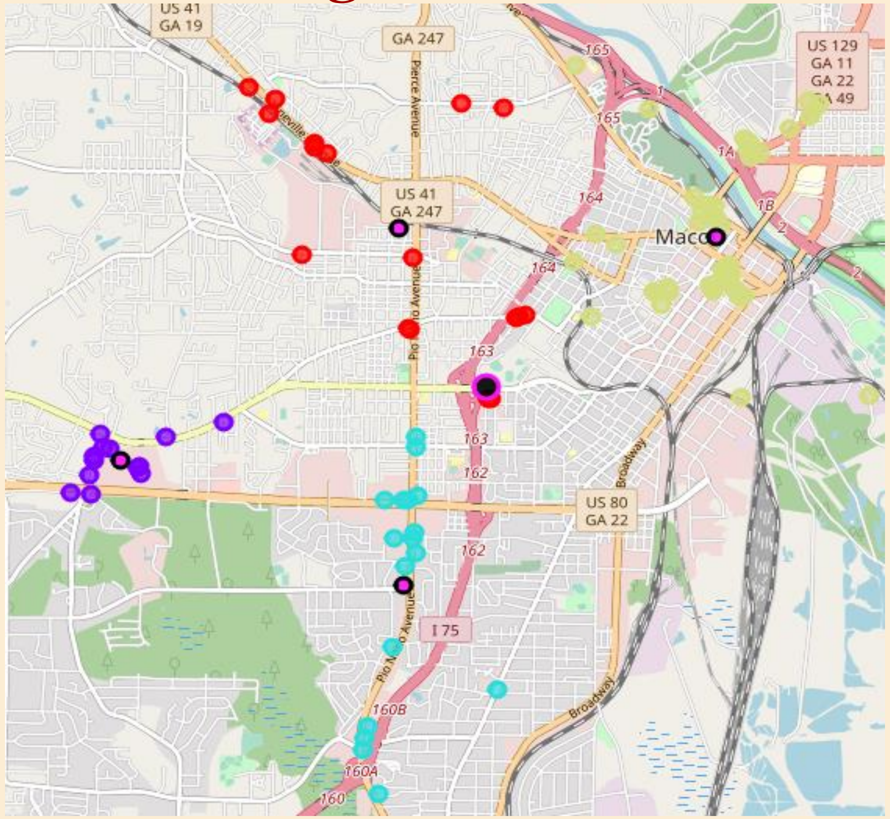
Results of *k*-means clustering *(cont.)*

Cluster Information

- **Mercer University:** pink circle, black fill
- **Cluster Centroids:** black circle, pink fill

- Clusters are color coded dynamically, location relative to Mercer.
  - **Cluster 0** – Red (north-northwest)
  - **Cluster 1** – Purple (west)
  - **Cluster 2** – Teal (south)
  - **Cluster 3** – Olive (northeast)

The Segmented View



EXAMINING THE CLUSTERS

FREQUENCY

Rank the categories

- Need to determine where a category is under-represented.
- Ranking the categories by frequency within each cluster offers a way to see this.
- My approach looks at the top 5 most common types of restaurants within a given cluster.

ONE-HOT ENCODE

Categories to Dummies

- Need dummy variables instead of categorical variables to determine frequency.
- One-hot encoding of the variables effectively translates the categorical variables into dummy variables.
- From here we can look at the mean frequency of a category across the cluster.

GOAL

Under-Represented

- In this situation, under-represented will be determined by the following criteria:

1. Category is represented in the top 5 for at least 2 other clusters.
2. Category is not in the top 5 for a selected cluster.

- If those criteria are met, I will select the most represented category based on criteria 1 and the cluster which applies to criteria 2.

THE RESULTS

3 Clusters

Fast Food Restaurant

Fast Food

This restaurant category is the most common category in cluster 1, 2 and 3 (often by a large margin).

17

Other Restaurants

Cluster 0

Cluster 0 has 17 other restaurants in it as competition, among 12 categories.

Fast Food restaurants don't even make the list in cluster 0. As a reminder, this is the area north-northwest of Mercer University.

cluster

0

Conclusion

Cluster 0 appears to be a very good option for a fast food restaurant, especially given the proximity to a university and the lack of competition for that type of restaurant.

FUTURE DIRECTIONS

Where to go from here?

New locations

New data

Optimal *k*

Thank You

Tim Brady

Repeatable Process

You can easily replicate this process for any city with Foursquare data.

Improvement

Future iterations of this project could include crime data, demographic data, or population densities in the analysis.

Including additional information may improve insight.

Improvement

The Elbow Process described doesn’t always provide a clear answer to the optimal *k*. Other methods could be explored to determine an optimal value.

Recommendations

Please provide descriptive feedback, as I am always looking to improve.

This is especially true if you decide to not offer full credit, as I would like to better understand for the future.

Capstone

This project represents my final deliverable for the Coursera Capstone project.