1. Introduction

1.1 Background

Whenever people move to any other place, they explore and try to get as much information as possible about it. While there are many different hobbies and interests which appeal to people, food is a common connection for all groups of people. Often we seek comfort food in times of stress, share food when meeting new people, and celebrate important events with food. Few other things are universally common and celebrated as food. It is also a difficult but profitable business if the time is taken to maximize the chances of success. A great location can help increase the chance of success for a new business. Focusing on the needs of a restaurateur, the question of where in a new market to open a restaurant is critical to the success of the business.

1.2 Problem

The question of location is critical to the success of any business, especially a restaurant. Optimal conditions would be an area well-traveled and affluent, with little or no competition from other food venues. However, such a location is typically not available. While ethnic restaurants have an advantage, the unique experience offered with novel flavors, competition from other vendors can have a significant impact on the business. Given these facts, the ideal location would be one with significant foot or vehicle traffic in an affluent area, which has a minimal number of food vendors and food diversity.

1.3 Interest

Opening a new business is extremely difficult. The majority of new businesses, including restaurants, fail within the first year. Any tool or advantage that can be given to the perspective business owner will give a higher chance of success. However, movement of an established business is typically difficult. New comers to an area may not be familiar enough with the area to determine an optimal location, or fail to account for the influence of all variables. The goal of this project is to provide that data so that the business owner can make an informed and optimal decision.

2. Data acquisition

2.1 Data Sources

- The various cities and the corresponding zip codes for each area of the greater Saint Louis metropolitan area will be web scraped from the following link: https://www.bestplaces.net/find/zip.aspx?st=mo&msa=41180
  The goal of this is merely pull the designated communities and zip codes for later work.
- Geospatial coordinates of the center of each zip code will be determined using the Geopy library with ArcGis API, which will return the center of each zip code in cartesian coordinates.

- Venues in each zip code area will be determined through the Foursquare API. The Foursquare API allows access to a database of more than 105 million places. The majority of the data for this project will come from that databased. It is already being used to get-tag photos, explore new areas, or just to determine a great place to get dinner. This API provides the ability to perform location search, location sharing and details about a business. From the venues pulled for each location, we will be able to determine the number and type of food venues as well as the location of each venue. With this information we will then be able to make a solid recommendation for where the best area or areas are to setup a new restaurant. Due to the size of area we are analyzing the radius of the queries will be set to 5 miles with the return of the top 100 venues for each zip code.

2.2 Packages and Dependencies

To perform the analysis and graphically represent the data we will utilize a number of prebuilt python libraries. All code will be executed in a Jupyter Notebook.

- Pandas - Library for Data Analysis
- NumPy – Library to handle data in a vectorized manner
- beautifulsoup – Library to web scraping activities
- Geopy – To retrieve Location Data
- Requests – Library to handle http requests
- Matplotlib – Python Plotting Module
- Sklearn – Python machine learning Library
- Folium – Map rendering Library

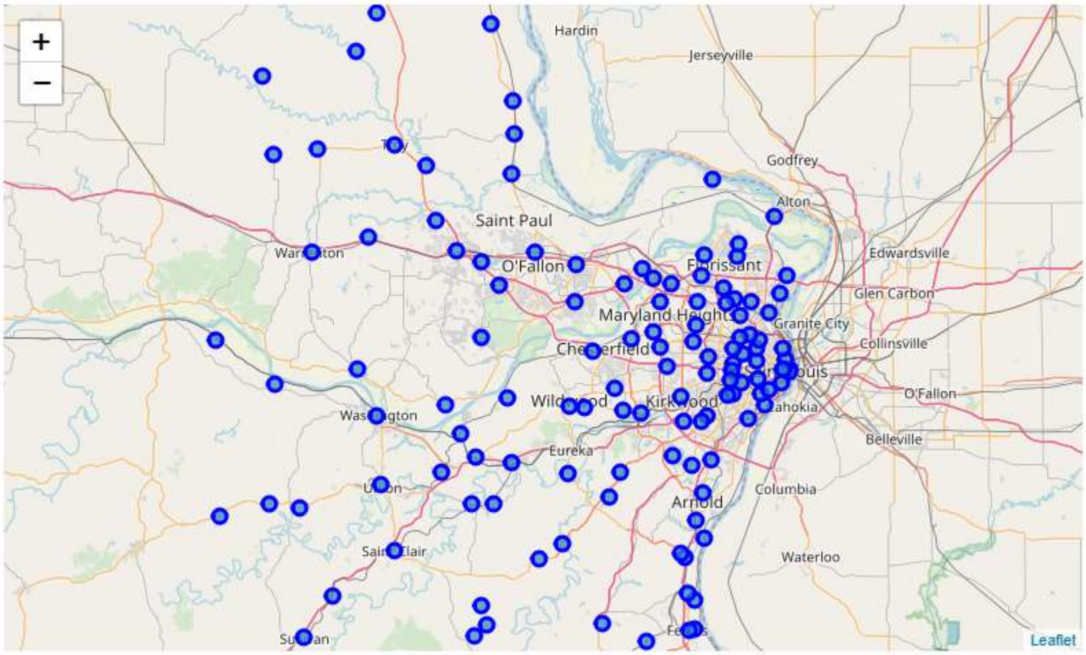
3. Methodology

3.1 Web Scraping

Initial data acquisition begins with the determination of what the greater Saint Louis metropolitan area encompasses. Web scraping of a list of zip codes and corresponding area names was performed using the Beautiful Soup library. This was then saved as a separate CSV file. Please see the following notebook: webscrapSTL.ipynb.

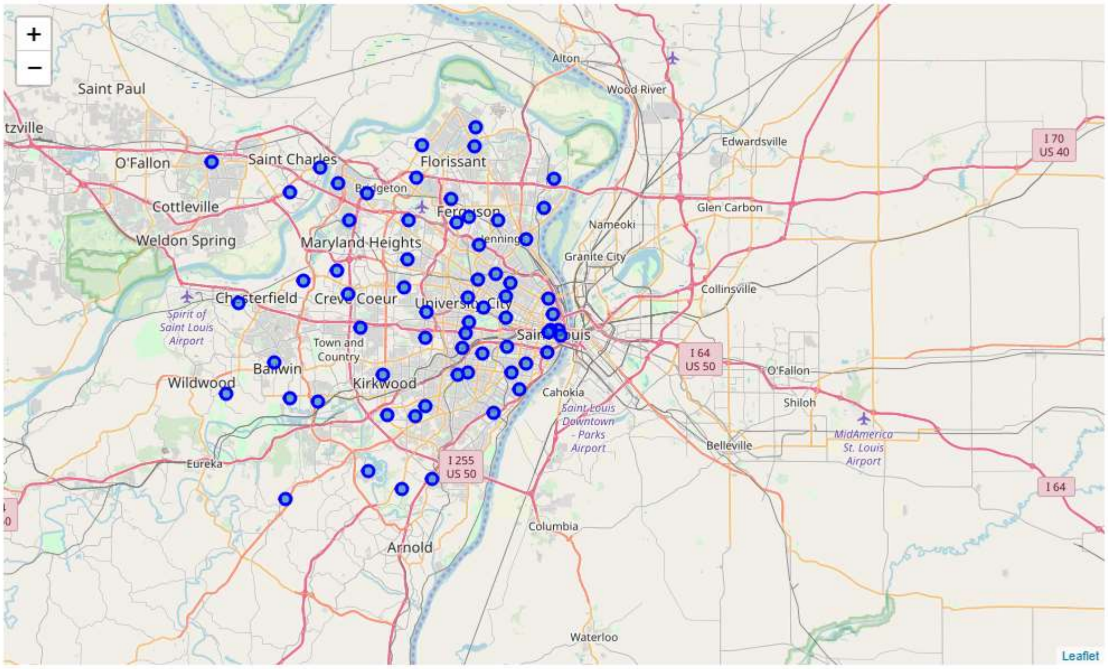
3.2 Geospatial Coordinates

The scraped zip codes and area names were then passed to a new notebook to determine the geospatial coordinates of the center of each zip code. This was performed using the Geopy library with the geocoder command with the ArcGis API option. The data returned a set of latitude and longitude coordinates for the center of each zip code. Visualizing the data, the following map was produced:



Examining the data, I believed the originally determined area was too large and could lead to errors in the analysis. As a result, the area of interest was defined as between Latitude 38.436944 and 38.867190. The longitudinal area of interest was the area between -90.619140 and the river. Zip codes outside of this area were excluded from the analysis. Please see following notebook: STL_Lat&Long.ipynb

The final center points of the zip codes analyzed were visualized below:



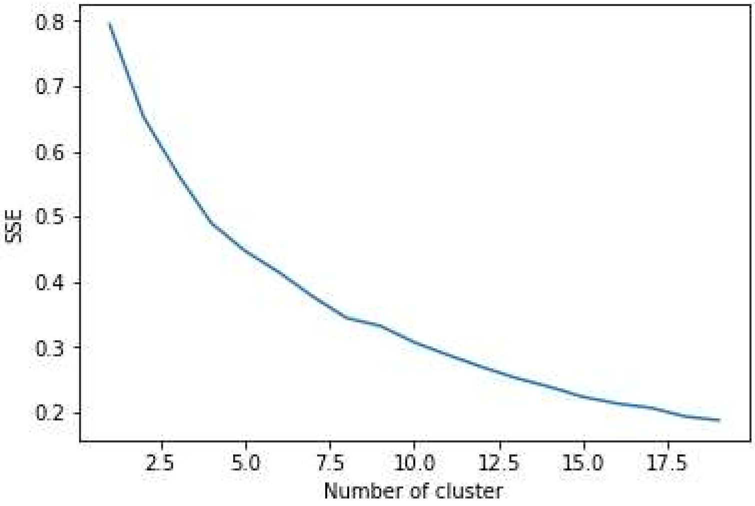
3.3 Venue Data

Venue data for each zip code was gathered using the Foursquare API with the limit of 100 venues and radius of 8046.72 meters, which is approximately 5 miles. For each venue, the venue's latitude, longitude, name, and category was returned. From the query of each zip code, 6395 venues were identified, with 235 categories. As the focus of this project was food venues and the variety of food venues, the data set was then filtered to remove non food offering venues. This reduced the venue count to 3863 venues and 83 venue categories. The food diversity of each zip code was counted as the number of unique venue categories. In addition, the total number of food offering venues was calculated.

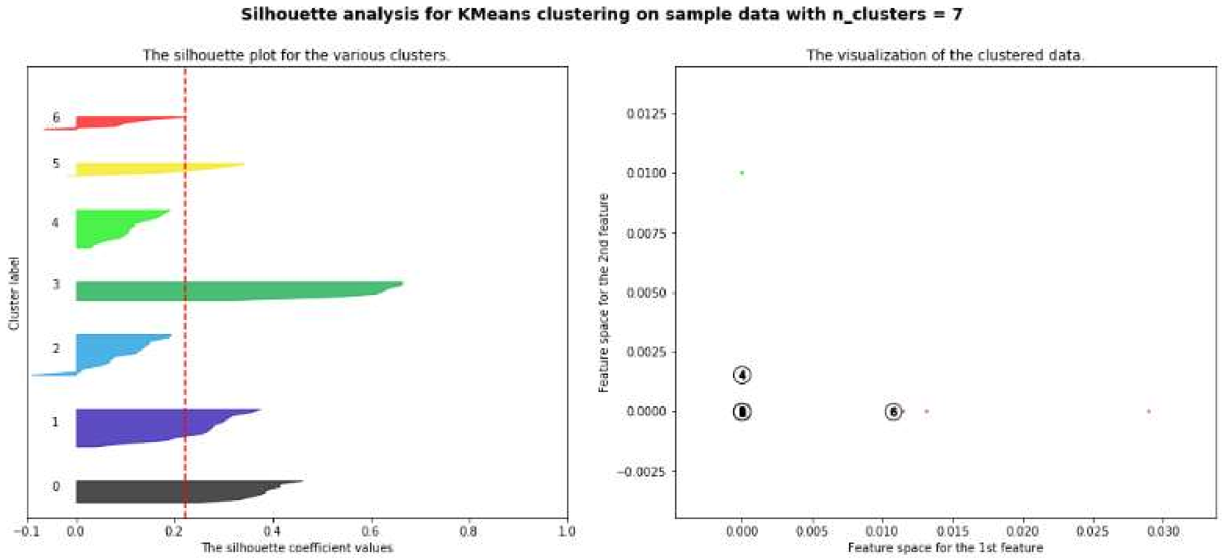
At this step, one hot encoding was performed on the total venue data. Then, the top 10 most common venues for each zip code was determined as a preprocessing step to further analysis.

3.4 K-Means Clustering Analysis

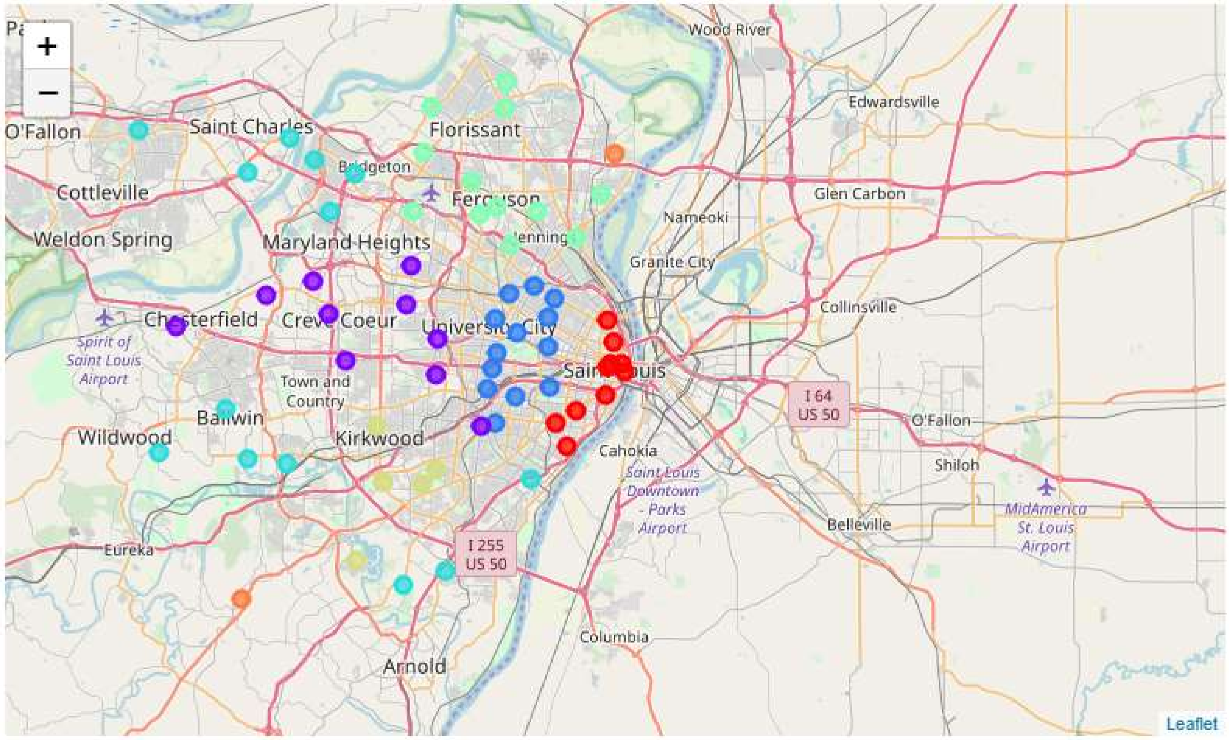
K-means clustering, a type of unsupervised learning, was used to determine the relationships between the different studied zip codes top 10 venues. Essentially, it is way to determine the similarity between different zip codes. However, for K-means clustering the optimal number of clusters must be determined separately. First, an elbow plot was produced:



From an elbow plot the optimal number of clusters is typically taken as the number of clusters after which the plot slope changes. However, the point of inflection was not clear. To better determine the optimal number of clusters, silhouette analysis was performed. A high positive value for a data point in silhouette analysis indicates that the object is well matched in the cluster. A silhouette analysis value of zero can be interpreted as the classification of that data point is questionable, and negative values are usually interpreted as a misclassification of that data point. Based on the silhouette scores, 7 clusters was optimal while avoiding overfitting. The silhouette score plot is included below.

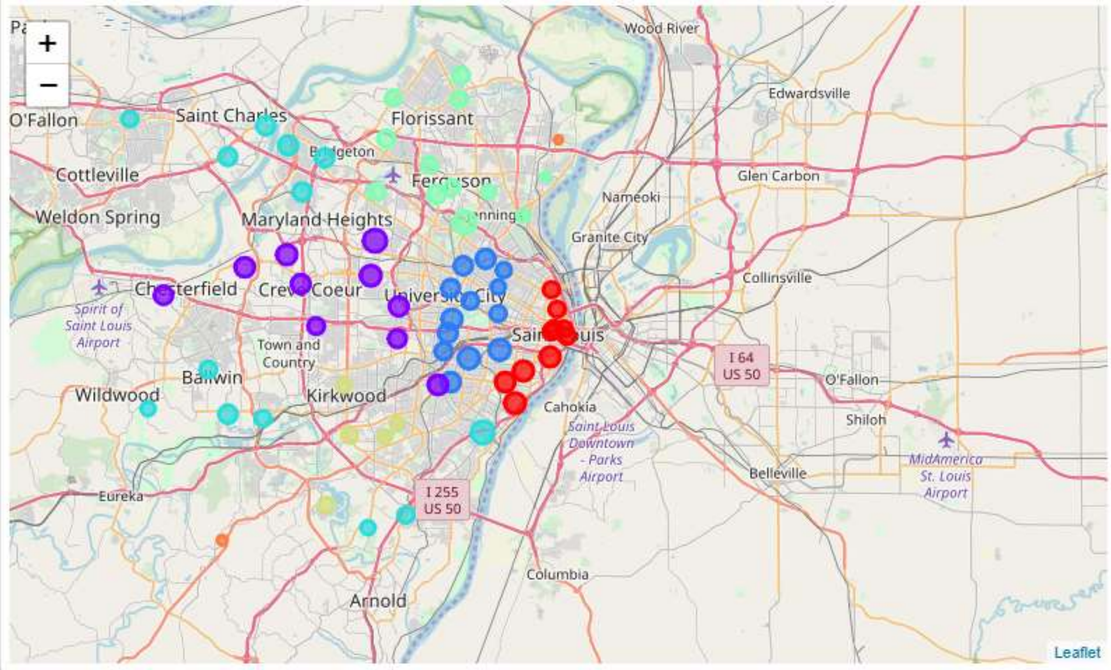


The results of the K-means clustering algorithm are plotted below:

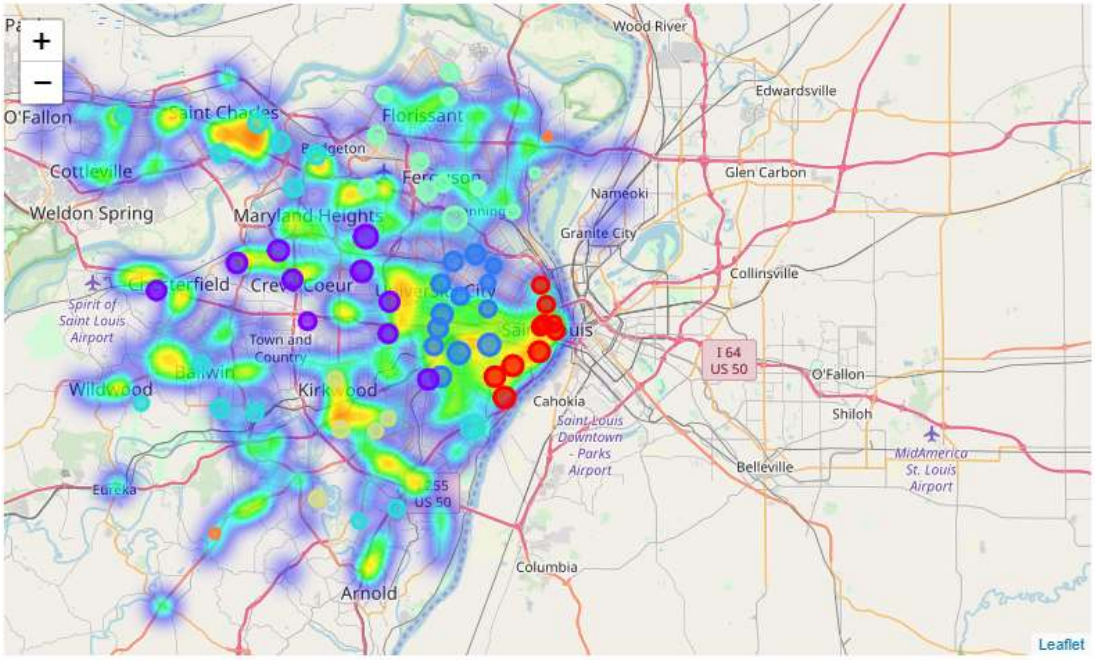


3.5 Heat Mapping and Data integration

To incorporate more data into the plot, making it more functional, the size of the plotted points will be altered to indicative of the relative food venue diversity in that area. See below:



In addition, a heat map of total number of food venues in each zip code can be added to the plot to further enhance the information conveyed.



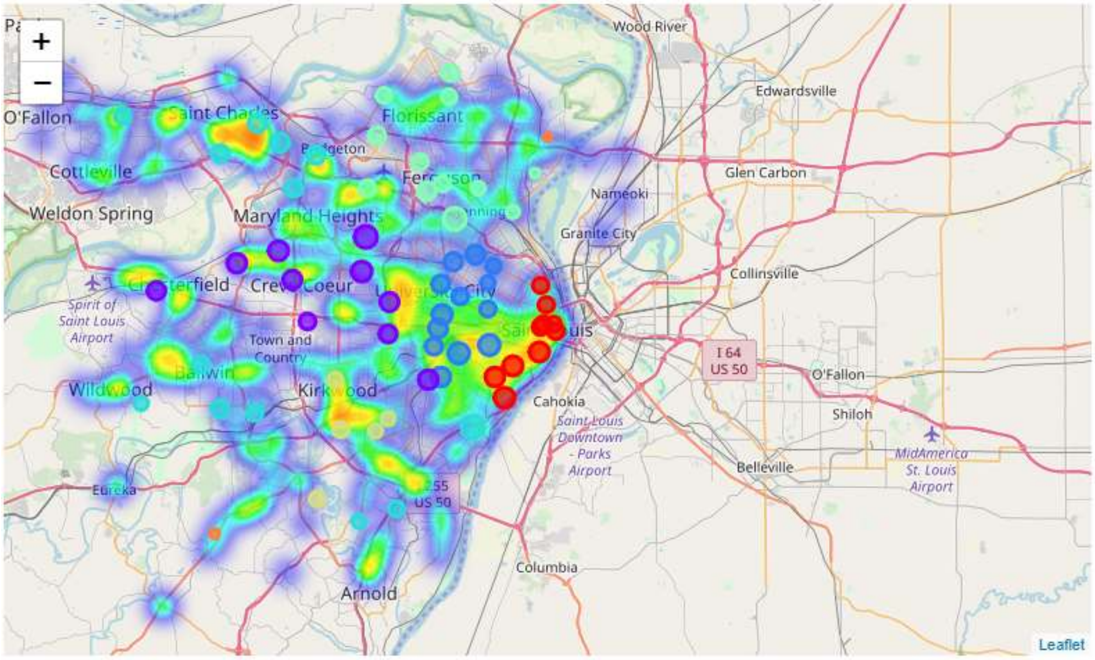
In one simple plot, we have managed to express a tremendous amount of data.

4. Results

The clustering appears to fall along known socioeconomic lines within the city. The light blue cluster, cluster 3, seems to represent the suburban outlying areas which we would expect to be radically different from the rest of the city. The most common venues appear to be pizza places, parks, and fast food restaurants. The light green cluster, cluster 4, seems to represent the northern section of the city which is known to be one of the less affluent parts of the city. The primary venue found was convenience stores. The red cluster, cluster 0, is representative of the down town area with is heavily populated by manufacturing, financial, and associated businesses with top venues being hotels, breweries, and coffee shops. The dark blue cluster, cluster 2, appears to be representative of the area around Forest park, Washington University, and Saint Louis University. The primary venues are the Saint Louis Zoo, bars, restaurants, and parks. The purple cluster, cluster 1, appears to represent the more affluent outlying communities, such as Ladue and Brentwood. The top venues in this cluster appear to be Italian restaurants and grocery stores. The green cluster, cluster 5, seems to represent an area in the southern outlying city with the top venues being primarily restaurants. Finally, cluster 6, in orange, appear to be areas which do not fit well into other clusters with the primary venues being fast food restaurants and parks. The supplied heat map and zip code point size each impart critical information visually and easily for the potential user.

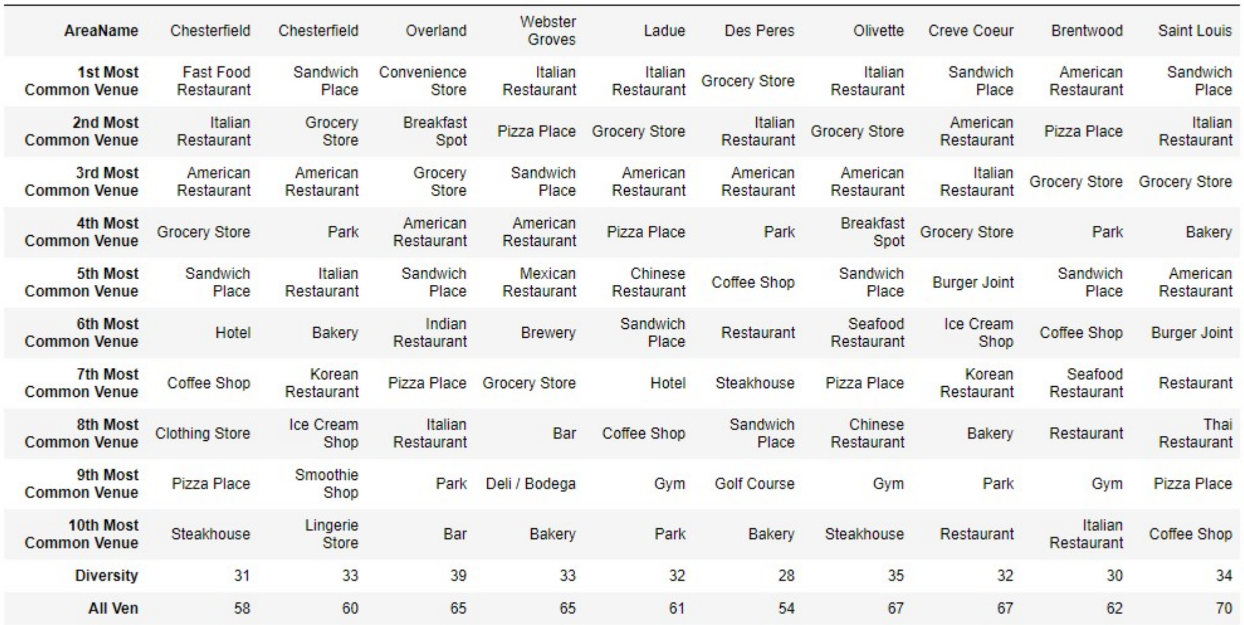
5. Results

Examining the clustered heat map again, some very useful observations can be made.



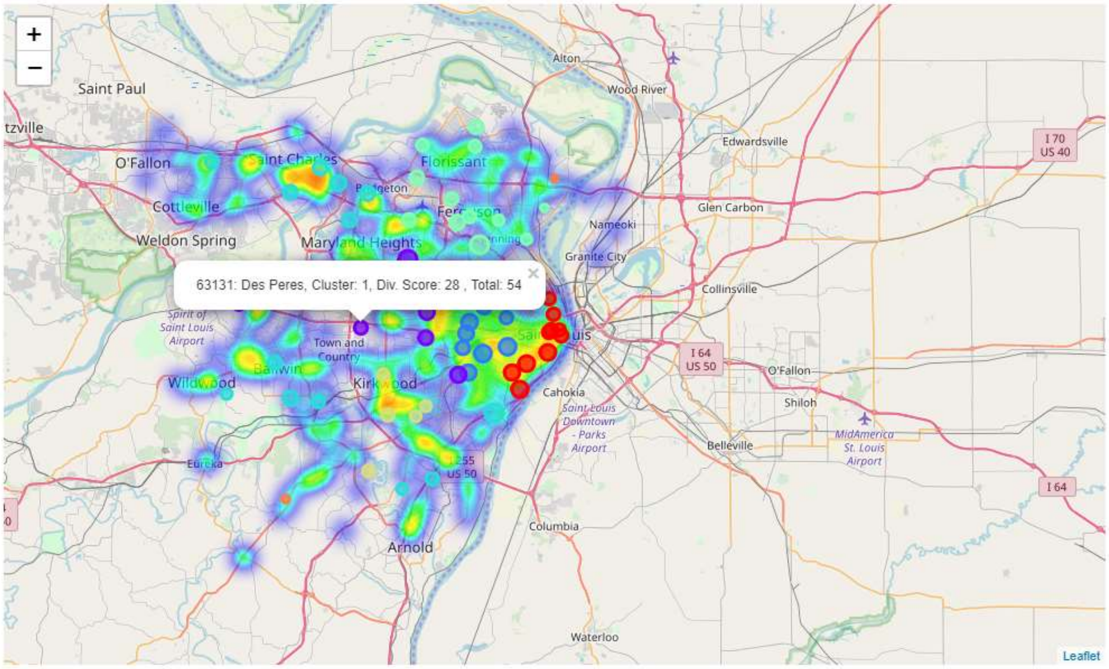
The areas described by cluster zero, cluster 2, and cluster 5 have a high concentration of food venues with a relatively large variety of venues. We can go ahead a rule those areas out of consideration. Parts of the area described by cluster 3 also have a high number of restaurants.

However, surprisingly several areas described by cluster 1 has area of relatively low food venue concentration as well as low variety. Examining that cluster more closely:



AreaName	Chesterfield	Chesterfield	Overland	Webster Groves	Ladue	Des Peres	Olivette	Creve Coeur	Brentwood	Saint Louis
1st Most Common Venue	Fast Food Restaurant	Sandwich Place	Convenience Store	Italian Restaurant	Italian Restaurant	Grocery Store	Italian Restaurant	Sandwich Place	American Restaurant	Sandwich Place
2nd Most Common Venue	Italian Restaurant	Grocery Store	Breakfast Spot	Pizza Place	Grocery Store	Italian Restaurant	Grocery Store	American Restaurant	Pizza Place	Italian Restaurant
3rd Most Common Venue	American Restaurant	American Restaurant	Grocery Store	Sandwich Place	American Restaurant	American Restaurant	American Restaurant	Italian Restaurant	Grocery Store	Grocery Store
4th Most Common Venue	Grocery Store	Park	American Restaurant	American Restaurant	Pizza Place	Park	Breakfast Spot	Grocery Store	Park	Bakery
5th Most Common Venue	Sandwich Place	Italian Restaurant	Sandwich Place	Mexican Restaurant	Chinese Restaurant	Coffee Shop	Sandwich Place	Burger Joint	Sandwich Place	American Restaurant
6th Most Common Venue	Hotel	Bakery	Indian Restaurant	Brewery	Sandwich Place	Restaurant	Seafood Restaurant	Ice Cream Shop	Coffee Shop	Burger Joint
7th Most Common Venue	Coffee Shop	Korean Restaurant	Pizza Place	Grocery Store	Hotel	Steakhouse	Pizza Place	Korean Restaurant	Seafood Restaurant	Restaurant
8th Most Common Venue	Clothing Store	Ice Cream Shop	Italian Restaurant	Bar	Coffee Shop	Sandwich Place	Chinese Restaurant	Bakery	Restaurant	Thai Restaurant
9th Most Common Venue	Pizza Place	Smoothie Shop	Park	Deli / Bodega	Gym	Golf Course	Gym	Park	Gym	Pizza Place
10th Most Common Venue	Steakhouse	Lingerie Store	Bar	Bakery	Park	Bakery	Steakhouse	Restaurant	Italian Restaurant	Coffee Shop
Diversity	31	33	39	33	32	28	35	32	30	34
All Ven	58	60	65	65	61	54	67	67	62	70

The community of Des Peres appears to be an excellent choice of a community to establish a business. For the given cluster, it has the lowest diversity score and total number of food venues. As indicated in the heat map, large areas of this community have a relatively low density of food venues, making it a low competition area within the central core of the city.



63131: Des Peres, Cluster: 1, Div. Score: 28 , Total: 54

Other areas of low food venue density and variety are found in many of the outlying communities of cluster 3.

6. Conclusion

In this study I was able to successfully pull geolocation data for the city of Saint Louis and model the data into apparently meaningful clusters representing known facts of the city. Using this geolocation data we were also able to measure the relative variety, number, and distribution of food venues. This mapping exercise highlighted that certain areas of the city may be prime locations for the opening of a new food venue, given additional research and preparation.