**Applied Data Science Capstone Project - The Battle of Neighborhoods**

**Analyzing Racial Segregation in St Louis by Profiling the Shops, Businesses and Cultural Attractions in Each Neighborhood**

**John C. Craig**

Illustration 1: The Gateway Arch in St. Louis credit: Wikipedia, Creative Commons

# Introduction

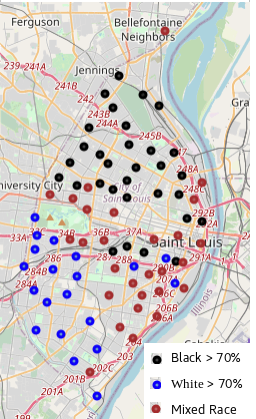
St Louis, Missouri is an important commercial and transportation hub situated at the confluence of the Mississippi and Missouri rivers. The St. Louis metropolitan area is home to about 2.8 million people, although the City of St Louis has under 310,000 residents. St. Louis has a history of racial segregation, which continues today with each of its 79 neighborhoods varying greatly in the proportions of White, Black and other racial groups.

This study looks at the characteristics of each neighborhood as reflected in the types of businesses and other venues in it, and compares those with the geographical location and racial makeup of the neighborhoods. The purpose of the study is to better understand the unique characteristics of the different areas of St. Louis, and racial disparities that exist there. The results show strong effects of racial segregation in the city, and adjacent areas in St. Louis County.

Among the results that we attempt to find is evidence of “food desert” characteristics in neighborhoods and areas. A food desert is defined as an area where there are no supermarkets or fully stocked groceries within a certain distance of a particular area. This is of particular concern for low income people who rely on public transportation and have difficulty going a long distance to buy groceries. They will often rely on convenience stores which have higher prices, less selection, more processed foods and and fewer healthy foods such as fruits and vegetables.

# Data

The Data for this study comes from two main sources, the Wikipedia article “List of neighborhoods of St. Louis”, and information about businesses other venues provided by geographical queries to the Foursquare.com database. Geolocoding of the neighborhoods was done using the arcgis geocoder.

  
Illustration 2: Neighborhoods which are predominantly Black, predominantly White, and of Mixed Race.

The list of neighborhoods was compiled from US Census 2000 data, and includes for each neighborhood the population, percentages of White, Black, Hispanic, Native American, Asian, and Mixed Race groups.

The data from Foursquare.com comes from millions of users who provide information about businesses and other venues they visit. These venues are categorized into 227 categories, such as particular types of restaurants, shops, parks, theaters, etc. The relative numbers and types of venues associated with each neighborhood are used to develop characteristic profiles the the neighborhoods. For this study, we used both a search radius of 1000m, 1500m and 2000m and a maximum of 100 venues were returned in each case. A 1000 meter search radius was chosen because the neighborhoods are approximately that size, or smaller, and would therefore include all of the of the venues which are listed in Foursquare.com within the neighborhood. The search was repeated with a search distance of 1500m and 2000m, since residents of a particular neighborhood are likely to use and benefit from facilities that are up to a few kilometers from their home, not just the immediate vicinity. This is particularly true for those who have cars or easy access to public. This had a significant effect on the clustering of neighborhoods, as can be seen in the results.

# Methodology

This study makes use of Python and a variety of Python modules for data extraction, transformation and analysis. The python modules imported include numpy for numerical calculations, pandas to load and manipulate data frames or tables, json as returned data from Foursquare is in that format, geocoder to determine the geographical coordinates of each neighborhood, matplotlib to plot graphs,, sklearn for data clustering and analysis and folium for creating maps. Jupyter Notebooks are used to to code, document and display results. Maps are used extensively as the spatial context and inherent geography of the data is of greatest interest.

## 1 Obtain and prepare data

The St. Louis Neighborhood data was read from the website and into a pandas dataframe. The neighborhood data was geocoded and then checked by displaying on a map. It was clear that the geocoder erred with one particular neighborhood (Academy), which was placed many kilometers outside city limits, so the coordinates of that neighborhood were obtained from GoogleEarth, and the dataframe was updated with the correction. The neighborhood points were plotted in a folium map to show the correct results.

Next, The venue data were obtained by querying the Foursquare API. Venues that are within the search radius of more than one neighborhood are counted in the statistics of each of the neighborhoods, and are therefore represented multiple time in the overall data, but the number of unique venues was also calculated. For the 1000m search radius, a total of 2941 venues were returned, of which 1835 are unique. For the 1500m search radius, 4984 venues aere returned of which 1837 were unique. For the 2000m search radius, a total of 6090 venues were returned, of which 1537 were unique. The smaller number of unique venues for the 2000m search radius is likely due to the limit of 100 venues returned per query. The actual number of venues in the Foursquare.com database within a 2000m search radius was likely greater than 100 for more of the neighborhoods, resulting in some venues not being returned. It would have been better to have a higher limit for returned venues for the 2000m search radius, however 100 is the limit for type of Foursquare account used. Despite this, using the 2000m search radius was able to reveal larger scale features, as can be seen in the results section.

## 2 Transform data

A onehot encoding was used to enumerate the venue information, then the venues were grouped by neighborhood. Onehot encoding is a technique whereby for categorical data, a separate column is created for each category, and a 1 or 0 is recorded as the value of each column depending on whether that category was present or not. The grouping used the sum method to obtain the count of all venues of a particular type within each neighborhood. Since we were particularly interested in the availability of groceries, four categories of venues that sell fresh food, ‘Supermarket’, ‘Grocery Store’, ‘Fruit & Vegetable Market’ and ‘Farmers Market’ were aggregated together to create a category of ‘All Grocery’. That category together with the ‘Supermarket’ category were merged together with the Neighborhood census data for analysis of the availability fresh food and ‘food desert’ conditions. The number of venues that are associated with each neighborhood were also calculated for each of the search radii of 1000m and 2000m

## 3 Cluster neighborhoods

The venue data were then sorted for each neighborhood to obtain the 10 most common venues in each neighborhood. This sorted dataset then received unsupervised classification using the KMeans method is SciKitLearn. Five cluster groups were used, and the labels were added to the dataset. This process was repeated for the 2000m search radius dataset. The only criteria that were used for clustering was the frequency of the 10 most common venue types in each cluster. The process was tried with up to 20 most common venues, but it did not change the clustering results. The effects of search radius did significantly change trh clustering results.

## 4 Analyze data

The census data providing the racial makeup of each neighborhood is vital to this study as it helps to explain the results seen in the types and qualities of venues, and is strongly determined by location. The racial statistics are given as percentages of White, Black, Hispanic, Native American, Asian and Mixed Race. The four latter categories together represent less than 10 per cent of the population overall, so the racial makeup of St. Louis is dominated by the Black and White populations. In order to show the geographic distribution, we separate the neighborhoods into groups with greater that 70% White, greater than 70% Black and with less than 70% White or Black, which we call Mixed Race neighborhoods. These categories are also used to analyze the number of venues per neighborhood.

# Results

When the labeled neighborhoods were plotted on a map, geographic areas of similar neighborhoods appear which share a similarity which is reflected in or derived from the types of venues, businesses, parks and other facilities. This effect becomes much more prominent and well-defined when a search radius of 2000m was used. There, 5 well-defined regions appear, which correspond with known regions of the city, which we label as North, South, Central, West and Downtown. These region names are added to the table that presents the detailed racial makeup of each cluster of neighborhoods for the 2000 meter search radius.

The “food desert” that effect can be seen in both the North region and the West region with few or no supermarkets appearing within a 1000m or 2000m search area. The effect of that on the population is very different for each of these regions, as the West region has an affluent population with easy access to transportation, where as the North has a poor population with less ease of transportation.

It may be noted that the largest number of grocery stores appear in mixed race areas, with a higher Hispanic population. Ethnic diversity can lead to demand for more variety of types of food. This is also true with the overall number of venues, which is highest in ethnically diverse neighborhoods.

In order to test the statistical significance of these observations, One-way Analysis of Variance tests were performed. In order to do this, the neighborhoods were organized into three equally sized groups of 25 each, with the predominantly Black group composed of neighborhoods with greater than 84.9% Black, the predominantly White neighborhoods with greater than 58% White, and Mixed neighborhoods with less than 80% Black and 58% White.

Table 1: ANOVA for 'All Grocery' at 1000m, 1500m and 2000m radius

Table 1 shows results for the ‘All Grocery’ category at 1000m, 1500m and 2000m search radius. A significant difference appears between between the Black and Mixed race neighborhoods for the 1000m and 1500m radius data, and between Black and White for the 1500m and 2000m radius data, and less significance for the other differences. Table 2 shows significance results for the ‘Supermarket’ category. Black and White neighborhoods show a significant difference at 1500m search radius, and less significance for the other groups.

Table 2: ANOVA for 'Supermarket'' at 1000m, 1500m and 2000m radius

Another important result is the total number of venues retrieved from Foursquare.com for the neighborhoods at different search distances. Below are 3D scatter plots showing the total number of venues obtained per neighborhood, based on the percentage of Black and White population in each neighborhood. As the search radius increases, more neighborhoods reach the maximum 100 venues.

|  |  |  |
| --- | --- | --- |
| Illustration 3: Venues per neighborhood by racial composition for 1000m radius | Illustration 4: Venues per neighborhood by racial composition for 1500m radius | Illustration 5: Venues per neighborhood by racial composition for 2000m radius |

The following illustrations show Box plots of the data grouped into 3 categories of predominantly Black, predominantly White and mixed race areas. It can be seen clearly that Black neighborhoods have fewer venues registered on Foursquare.com than White or Mixed neighborhoods. Mixed neighborhoods have the largest number of venues. This is true for each of the search distances, although it is masked for the 2000m distance since most White and Mixed neighborhoods have reached the 100 venue maximum at that distance.

|  |  |  |
| --- | --- | --- |
| Illustration 6: Venues per neighborhood by racial composition for 1000m radius | Illustration 7: Venues per neighborhood by racial composition for 1500m radius | Illustration 8: Venues per neighborhood by racial composition for 2000m radius |

The illustrations below show the venues plotted in red, surrounding the neighborhood points in blue. The additional venues can be seen along the city border, where they extend into neighboring areas. In the interior of the city, many of the venues are duplicated as they appear in more than one neighborhood search.

|  |  |  |
| --- | --- | --- |
| Illustration 9: Venues at 1000m radius. | Illustration 10: Venues at 1500m radius | Illustration 11: Venues at 2000m radius |

Below are the neighborhood clusters which are produced using the k-means classifier, using only the similarity of the ten most common types of venues in or around each neighborhood. When compared with the racial composition of these clusters of neighborhoods, we can see that they tend to correspond with majority White, Majority Black or Mixed race groups. With the increased search radius, the clusters tend to develop into regions. These regions have recognizable characteristics as that will be discussed in the next section. The phenomena that leads to development of spatial structure is know as spatial auto-correlation, which proposed that spatial variables, such as these neighborhood locations, are not independent of one another, but are correlated by virtue of their spatial proximity. A more detailed study of the spatial auto-correlation of the neighborhoods is beyond the scope of this study.

|  |  |  |
| --- | --- | --- |
| Illustration 12: Neighborhood clusters at 1000m radius | Illustration 13: Neighborhood clusters at 1500m radius | Illustration 14: Neighborhood clusters at 2000m radius |
|  |  |  |

# Discussion

The Downtown area is obviously distinct because it is the commercial and cultural center of the city, where people commute to work or enjoy cultural amenities from all over the area. Therefore, the number of venues supported exceeds what would be supported by only the full-time residents. The types of venues are also distinct. As you can see from the table below, common venues include bars, breweries, cafes, restaurants and hotels. Besides being home to a major national brewery, St. Louis features many micro-breweries as well. These are popular with tourists or business travelers, as well as commuters who work in the area. The North of the city as well as adjacent areas of the County is overwhelmingly Black, the West and parts of the South are mostly White, The Central and Southeast areas are of mixed race.

In the North, common venues are fast food, gas stations, discount stores and small groceries, which predominantly serve local residents. In the West, parks and an excellent zoo, as well as restaurants, cafes and breweries highlight an upscale residential area. The Central area also features a variety of restaurants, but lacks proximity to the zoo and many park amenities. The South, has a mixture of restaurants, including ethnic food, groceries, parks and discount stores.

  
Table 3: Most common venue types by Cluster Region

# Conclusion

We have seen in this study that the types and numbers of retail businesses and other venues in a given neighborhood can be used to categorize the neighborhoods and they are correlated with the racial makeup of the neighborhoods in the city of St. Louis, Missouri. St. Louis has a history and current reality of racial segregation, with a large proportion of neighborhood being composed of a significant majority of White or Black residents. It can be seen from the numbers of businesses in each neighborhood that that in general Black neighborhoods are disadvantaged by having fewer venues, and further having significantly fewer grocery stores or supermarkets in some areas, resulting in a “food desert” condition in these areas.

It is further shown that mixed race neighborhoods are in some ways better off than even some of the wealthier White neighborhoods, in terms of the variety of food venues such as grocery stores and restaurants as a more diverse population develops a demand for a greater variety of such restaurants and stores.

It is further in this study that varying the spatial scale of an analysis can reveal different structures and characteristics. If the spatial scale is limited to individual neighborhoods, then it will not show regional structures, but if the scale is increased to include adjacent neighborhood then larger regional characteristics are revealed.